MARKET PRICE PREDICTION MODEL FOR AGRICULTURAL PRODUCTS USING TIME SERIES ANALYSIS

A CONTEXT OF COFFEE IN LOW RESOURCE ENVIRONMENTS

A dissertation presented to

FACULTY OF SCIENCE

in partial fulfillment of the requirements for the award of the degree

Master of Science in Information Systems

UGANDA MARTYRS UNIVERSITY

NAKAIMA Amina

2023-M132-21503

Supervisor: Muchake Brian

August 2025

UGANDA MARTYRS UNIVERSITY

DIRECTORATE OF GRADUATE STUDIES, RESEARCH AND ENTERPRISE

Master's Dissertation

Declaration

I have read the rules of Uganda Martyrs University on plagiarism and academic honesty, and hereby state that this work is my own.

It has not been submitted to any other institution for another degree or qualification, either in full or in part.

Throughout the work I have acknowledged all sources used in its compilation.

I finally grant Uganda Martyrs University permission to store and reproduce this dissertation, in whole or in part, in any manner or format, which Uganda Martyrs University may deem fit.

Researcher's name: Nakaima Amina Juma

Researcher's signature: 

Date of submission: 1st/September/2025

Submitted to the Directorate of Graduate Studies, Research and Enterprise

UGANDA MARTYRS UNIVERSITY

DIRECTORATE OF GRADUATE STUDIES, RESEARCH AND ENTERPRISE

Master's Dissertation

Approval

This dissertation has been produced under my/our supervision and submitted for examination with my/our approval as the appointed academic supervisor/s.

Name of Supervisor (1): Mr. Muchake Brian

Signature of Supervisor: 

Name of Supervisor (2): _____

Signature of Supervisor: _____

Date of submission: 1st/September/2025

Submitted to the Directorate of Graduate Studies, Research and Enterprise

Dedication

This work is dedicated to my family my mother, Mrs. Gidudu Mastula Nalweyiso; my brother, Dr. Wandega Umar; and my sisters, Ms. Nabukwasi Faridah (whose continuous comfort sustained me), Ms. Muzachi Zaitun, and Ms. Nabuduwa Nulu, who instilled in me the values of hard work, integrity, and perseverance.

I extend this dedication to my supervisor, Mr. Muchake Brian, for his steadfast guidance and mentorship throughout this research.

My sincere appreciation goes to Mr. Bamanya Deus and Mr. Kalema of the Uganda National Meteorological Authority (UNMA) for facilitating access to climate data, and to Mr. Kulumba Paul for providing Uganda exchange-rate data.

Acknowledgment

I would like to extend my deepest gratitude and heartfelt appreciation to all those who contributed to the successful completion of this research project titled "Market Price Prediction Model for Agricultural Products Using Time Series Analysis: A Context of Robusta Coffee in Low Resource Environments." This academic journey has been both intellectually enriching and personally transformative, and it would not have been possible without the guidance, support, and encouragement of several individuals and institutions. I am sincerely grateful to my supervisor, lecturers, and academic mentors whose insight and continuous support shaped the direction of this study from conception to completion.

I am especially indebted to the faculty and staff of the Department of Computer Science and Information Technology at Uganda Martyrs University for creating a nurturing academic environment. Their unwavering dedication to academic excellence, availability of research resources, and constant encouragement played a key role in developing the skills and confidence I needed to undertake this work. The institutional support, including access to technical facilities and scholarly mentorship, created the perfect foundation for rigorous inquiry and critical analysis.

To my family, your love and support have been the backbone of my perseverance. I am particularly thankful to my parents and siblings whose prayers, encouragement, and understanding kept me going, even during the most challenging times. Your belief in my potential inspired me to stay focused and committed to my academic and professional aspirations. I am equally grateful to my friends and fellow researchers who offered their time, feedback, and moral support throughout this journey. The knowledge sharing, peer discussions, and moral companionship were immensely uplifting and motivating.

Lastly, I wish to acknowledge all data providers and institutions whose publicly available data formed the core of this research, particularly UCDA, ICO, Bank of Uganda, and UNMA. Their data contributions gave practical meaning to the study and allowed for real-world application of forecasting models. Above all, I give thanks to Allah for His endless grace, strength, and wisdom that guided me every step of the way. This project stands not only as a testament to academic discipline but also to the faith, resilience, and collective effort of all who believed in me. Thank you all.

Abstract

The volatility of agricultural commodity prices, particularly in low-resource environments, poses significant challenges for farmers, policymakers, and market participants. This study aimed to develop a robust market price prediction model for Robusta Kiboko coffee in Uganda using time series analysis techniques. The research employed four forecasting models Seasonal Autoregressive Integrated Moving Average (SARIMA), SARIMAX, Long Short-Term Memory (LSTM), and a Hybrid SARIMA-LSTM model to assess their predictive performance using historical price data from 2010 to 2025. Data were collected from multiple reputable sources, including the Uganda Coffee Development Authority, Uganda National Meteorological Authority, Bank of Uganda, and the International Coffee Organization.

The methodology involved a thorough data preparation process, including data aggregation, cleaning, transformation, and splitting into training and testing sets. The SARIMA model captured linear and seasonal trends, while the SARIMAX model incorporated exogenous variables such as rainfall, exchange rates, and international prices. The LSTM model was used to capture nonlinear dependencies, and the hybrid model combined SARIMA and LSTM to leverage their complementary strengths. Model performance was evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Results indicated that the Hybrid SARIMA-LSTM model underperformed compared to the individual models, recording the highest error metrics on the test data (MAE = 1,896.51 UGX, MSE = 4,251,983.30 UGX^2, RMSE = 2,062.03 UGX). The Univariate LSTM model achieved the lowest errors (MAE = 245.61 UGX, MSE = 156,102.61 UGX^2, RMSE = 395.10 UGX), demonstrating superior predictive accuracy. These findings highlight the importance of context-specific model evaluation, showing that while hybrid models can be powerful in some settings, in this study individual LSTM variants were more effective for agricultural commodity price forecasting. The study contributes to the literature on time series forecasting in agriculture and offers practical implications for price risk management, policy formulation, and agricultural planning in resource-constrained environments.

Keywords: Coffee price forecasting, SARIMA, LSTM, SARIMAX, Hybrid model, Time series, Uganda, Agricultural economics.

Acronyms and Abbreviations

PPI	Producer Price Index
GDP	Gross Domestic Product
UCDA	Uganda Coffee Development Authority
LSTM	Long Short-term Memory
ARIMA	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
MAE	Mean Absolute Error
RMSE	Root Mean Square Error
MSE	Mean Square Error
LSTM	Long Short-term Memory
ARIMA	Autoregressive Integrated Moving Average
GRU	Gated Recurrent Units
SVM	support vector machine
DL	Deep Learning
ML	Machine Learning
UNMA	Uganda National Meteorological Authority
EDA	Exploratory Data Analysis
UCDA	Uganda Development Authority
ICO	International Coffee Organization
GARCH	Generalized Auto-regressive Conditional Heteroskedasticity

Table of Contents

CHAPTER ONE

GENERAL INTRODUCTION

Agriculture remains a cornerstone of economic development and food security, particularly in low-resource countries where a significant proportion of the population relies on farming for livelihood. The sector is highly dynamic, influenced by climatic conditions, market forces, policy changes, and socio-economic factors. These complexities often result in price volatility, which can adversely affect farmers' incomes, agribusiness investments, and overall market stability Ngoc *et al*. (2023).

In recent years, predictive analytics has emerged as a powerful tool for addressing uncertainties in agricultural markets. By analyzing historical and current data, predictive models can identify trends, forecast prices, and enable proactive decision-making. These insights allow stakeholders including farmers, agribusinesses, and policy-makers to mitigate risks, optimize resource allocation, and strategically plan for future market conditions (Rustagi & Goel, 2022; Sun *et al*., 2023).

Despite the global adoption of predictive analytics in agriculture (Ruekkasaem & Sasananan, 2018; Zhang et al., 2020; Ouyang, Wei & Wu, 2019; Purohit *et al.*, 2021), there is limited empirical evidence from Uganda on market price prediction models for key commodities such as Robusta coffee. This gap underscores the need for robust, locally contextualized forecasting tools that can enhance market transparency, reduce uncertainty, and support the livelihoods of farmers in low-resource settings.

1 Introduction

Numerous studies worldwide have adopted predictive analysis in the agriculture sector (Ruekkasaem & Sasananan, 2018; Zhang *et al*., 2020; Ouyang, Wei & Wu, 2019; Purohit *et al.*, 2021). Predictive analytics make it possible for organizations to determine risks in advance, identify opportunities, detect trends, and develop appropriate strategies. This capability is effective only when predictions are based on organized and comprehensive data (Rustagi & Goel, 2022).

Predicting market prices for agricultural products is critical in providing stakeholders with advance price information, allowing for efficient and effective decision-making, while

reducing uncertainty and risks in agricultural markets (Ouyang, Wei & Wu, 2019). Accurate forecasting of agricultural commodity prices enables international organizations, governments, and agribusinesses to respond timely, ensuring adequate food supply and maintaining global food security Sun *et al*. (2023).

Despite the benefits associated with predictive analysis, existing literature remains largely silent on the Ugandan context, particularly regarding agricultural market price predictions, resulting in challenges for local stakeholders. Fluctuations in agricultural market prices can significantly affect farmers' incomes, investments, and market dynamics Ngoc *et al*. (2023). Reliable price forecasts empower farmers to make informed decisions, enhancing productivity, profitability, and livelihoods. Therefore, developing and utilizing robust market price prediction models is essential to support resilience and prosperity in the agricultural sector, ensuring farmers' well-being and sustaining the broader economy in low-resource environments.

1.1 Background of the study

In study by SAP (2024), predictive analytics is a branch of advanced analytics that makes predictions about future events, behaviours, and outcomes. The discipline of predictive analysis uses different techniques that is to say machine learning, regression analysis, time series analysis and classification and clustering to uncover patterns and trends in data in order to make accurate predictions Muhammad (2024). Predictive analysis uses statistical techniques including machine learning algorithms and sophisticated predictive modelling to analyse current and historical data and assess the likelihood that something will take place. Rustagi & Goel (2022), predictive analysis, a technique for forecasting events, uses historical data, statistical algorithms, and machine learning to identify patterns and trends. It finds broad applications in finance, healthcare, and agriculture, among other fields. Therefore, this study focuses on the concept of market price prediction for agricultural products a context of coffee commodity using time series analysis. According to Sun *et al.* (2023), agricultural product price prediction refers to the use of scientific methods to estimate or judge the trend and level of agricultural product price changes over a period of time in the future based on historical data and current information.

The influence of predictive analytics in business provides a scope towards agricultural areas to predict upcoming threats and the way out to mitigate those challenges in business. The entire decision making as well as strategic development in agriculture become more facile through assuming the upcoming risks through the assistance of predictive analytics in business Gupta and Malik (2022).

Agriculture market price prediction is also relevant to traders, policy makers and other stakeholders within the agricultural value chain. Ngoc *et al*. (2023) annotates about market price prediction in a number of ways as follows: (a) The traders rely on price forecasts to plan their procurement and distribution strategies, ensuring a smooth flow of agriculture commodities, policy makers utilize price prediction to design and implement effective agricultural policies, support farmers, and maintain market stability. (b) Accurate crop price forecasts enable stake holders to anticipate the market trends, manage inventories and develop strategies that promote sustainable agricultural growth and equitable market.

Gupta and Malik (2022) annotates, in the modern era, the application of predictive analytics has provided a superior scope to the business institution in agricultural fields to improve their entire production rate and meet the needs of the consumers worldwide. Gupta and Malik (2022) further point out that it is actually increasing the demand of the leading enterprises in the agricultural sector to utilize the assistance of predictive analytics within their business and agricultural tasks throughout the globe in recent days. They further annotate that the overall study has been going to make an empirical analysis of predictive analytics and its support to agriculture and the productivity and earning rate of a business enterprise in the agricultural industry in the international market periphery.

Despite the benefits that come along with market prediction for agricultural products, various studies point out the challenges associated with it. (Sun *et al*., 2023; Tran *et al*., 2023) mentions that agricultural market price predictions are affected by a variety of factors, such as supply and demand, climate change, policy intervention, market competition, international trade**.** Agricultural price shocks strongly affect farmers' income and food security worldwide therefore it is important to understand the origin of these shocks and anticipate their occurrence

On a global level, agricultural price prediction usually gets involved with a number of models or theories, time series analysis has been a subject of extensive study and application in

predicting prices of agricultural products, since it is able to model trends, seasonality, and variability in agricultural markets. Applying some established theories and methods (ARIMA and SARIMA) used in this field., where ARIMA and SARIMA are the main models to be used since they take into account seasonality and historical price patterns. This has been noted in some studies (Guo *et al*., 2022; Sun *et al*., 2023) where the method works well for regular and predictable seasonal changes in price data but is subject to shocks from external factors such as policy changes and climatic events; hence, there is a strong case for more adaptive models by combining ARIMA with machine learning to achieve robustness.

In Uganda, the prediction of market prices for agricultural products has traditionally relied on models that represent market dynamics, seasonal patterns, and policy changes. A number of empirical studies have employed price transmission models and spatial market integration analyses to assess the responsiveness of agricultural markets. Waiswa and Fahri (2023) note that vertical and spatial price transmission analyses have been commonly used to evaluate how prices are transmitted across different locations and market levels within the agricultural commodity chain.

The Uganda Bureau of Statistics (UBOS, 2020) developed the Producer Price Index for Agriculture (PPI-A), which measures the average change in prices received by farmers for a selected basket of agricultural commodities. The PPI-A not only reflects the trend of agricultural prices over time but also provides important insights for policy and planning. The index is built from farm-gate price data collected from various regions in Uganda and is regularly published for use by planners, researchers, and government agencies. However, despite its utility as a macroeconomic planning tool, the PPI-A exhibits several limitations. Structural and logistical challenges such as inconsistent data collection timelines, limited rural coverage, and lagged reporting reduce the index's effectiveness for real-time market forecasting.

However, Rapsomanikis, Hallam & Conforti (n.d.) the use of the above-mentioned mechanism is not efficient and effective for market price prediction due to the fact that markets can also be partly insulated by large marketing margins that arise due to high transfer costs, especially in developing countries, poor infrastructure, transport and communication services give rise to large marketing margins due to high costs of delivering the locally produced commodity to the

border for export or the imported commodity to the domestic market for consumption. High transfer costs and marketing margins hinder the transmission of price signals, as they may prohibit arbitrage Rapsomanikis, Hallam and Conforti (n.d) domestic. According to Sun *et al*. (2023), traditional methods are relatively simple and easy to understand and implement, but the prediction effect is poor for nonlinear, non-smooth, and high-dimensional data, and they require more a priori knowledge and assumptions.

Other significant contributions include Gilbert, Christiaensen and Kaminski (2017), who systematically examined food price seasonality across 193 markets in seven Sub-Saharan African countries, including parts of East Africa. Their study highlighted pronounced seasonal gaps up to 33 percent for staples like maize and emphasized the critical role of identifying intra-annual trends to improve agricultural price forecasting accuracy and support market stability interventions.

Similarly, Mawejje (2016) analyzed the drivers of food price volatility in Uganda using cointegration and error-correction models, focusing on the impact of energy prices, rainfall variability, and temperature shocks. The findings demonstrated that temperature shocks and energy costs were key contributors to short-term food price fluctuations, underscoring the need for adaptive and climate-sensitive forecasting models in agricultural markets.

Recent literature also criticizes the limitations of traditional models. Sun *et al*. (2023) argue that conventional forecasting methods, while easy to implement, often perform poorly when applied to nonlinear, high-dimensional, or noisy data. These limitations have led to a growing preference for advanced time series models and machine learning-based techniques that can handle complex interdependencies within agricultural price data.

Thus, integrating time series forecasting techniques like SARIMA, LSTM, and hybrid models into Uganda's agricultural pricing frameworks particularly in systems like the PPI-A has the potential to significantly improve accuracy and responsiveness. Such an approach allows for capturing both seasonal and structural patterns in agricultural markets, supporting better planning for stakeholders including farmers, traders, exporters, and policymakers.

1.2 Problem statement

Unstable market prices for coffee in low-resource environments such as Uganda pose significant challenges to farmers, traders, and policymakers. Price volatility is driven by multiple factors, including imbalances in market supply and demand, socio-economic shocks, and climate variability in the production process Slater *et al.* (2023). These fluctuations directly affect livelihoods and decision-making in the agricultural value chain.
Despite the centrality of coffee to Uganda's economy, stakeholders often lack reliable forecasting tools that can support informed planning and risk management. Conventional approaches to price forecasting, while useful, fail to capture the complex, nonlinear dynamics of agricultural markets in developing economies. The scarcity of quality data, combined with high variability, limits the effectiveness of existing models, leaving smallholder coffee farmers particularly vulnerable to price shocks and uncertainty.
Addressing this gap requires the development of robust, context-specific, and scientifically sound predictive frameworks capable of generating accurate forecasts of coffee prices. Such models would improve the ability of farmers, traders, and policymakers to adapt to market dynamics, reduce income instability, and strengthen agricultural planning in Uganda. By building a reliable price prediction system, this study seeks to provide practical tools for mitigating risk and informing effective policy formulation in resource-constrained settings.

1.3 Objectives

1.3.1 Main Objective

The main objective of this study is to develop and evaluate a market price prediction model for coffee using time series analysis, in order to provide stakeholders with reliable forecasts that support informed decision-making.

1.3.2 Specific Objectives

The specific objectives for this study are:

i. To prepare a dataset for better analysis of the study to ensure the data is clean, well-structured, and enriched for insights.
ii. To select prediction models that helped to forecast the market prices for coffee in Uganda.

iii. To build, train and evaluate a model for market price prediction model for agricultural products using time series analysis

1.4 Research question

i. How can a market price prediction model for agricultural products be developed using coffee as a case study through time series analysis to support stakeholders in making informed decisions?
ii. How can the available data on coffee market prices in Uganda be prepared, cleaned, and structured to support accurate and insightful time series analysis?
iii. What are the most effective time series forecasting models for predicting Robusta coffee prices in Uganda, and how do classical and deep learning approaches compare?
iv. Which time series model offers the best performance in predicting agricultural market prices, and how can it be evaluated using appropriate forecasting metrics?

1.5 Scope

1.5.1 Geographical Scope

The location of the research focuses on central Uganda, and more specifically, Buikwe District. This particular area of the study is due to the fact that it is one of the most densely populated agricultural regions particularly in coffee growing, which is essential to the core of the study that hopes to forecast market prices for agricultural products.

Buikwe District is bordered by Kayunga District to the north, Jinja District to the east, Buvuma District to the southeast and Mukono District to the west. This location is approximately 14 kilometres (8.7 mi), by road, southeast of Lugazi, the nearest large town

1.5.2 Content Scope

This research focused on market price prediction for agricultural products using coffee as the case study in low development environments using time series analysis. Time series techniques such as SARIMA, SARIMAX, LSTM and Hybrid SARIM - LSTM were applied to do the prediction. These techniques worked well within agricultural markets because they captured trends, seasonality, variability, short-term and long-term dependencies.

1.5.3 Time Scope

The study was scheduled to run over a one-year period, beginning in October 2024 and concluding in September 2025. This timeline allowed for systematic data collection, model development, and analysis.

1.6 Significance of the Study

This section entailed the stakeholders and through the collaboration among these stakeholders, the study intended to give a strong price prediction framework that enhanced decision-making across the coffee value chain. The study included the following stakeholders: (i)The farmers are primary stakeholders since that provided the necessary input regarding production and market conditions. Their inclusivity was thus crucial in providing information on price trends, seasonal variation, and other issues related to the models. (ii)The traders and exporters were intermediaries in the coffee value chain, and their contribution were made by providing regional price data, market dynamics, and demand forecasts to the model. (iii) Policymakers and government agencies, such as the Ministry of Agriculture facilitated access to national data and regulated market policies. A reliable forecasting tool could underpin strategic interventions that stabilized markets and support food security goals Therefore, they supported the implementation of predictive tools for market stabilization. (iv) The financial institutions and microfinance institutions utilized these predictions in designing financial products responsive to their target population of coffee farmers, be it crop insurance or flexible credit schemes. (v) The consumers indirectly influenced the study as their purchasing patterns fed into demand data.

The proposed solution benefited the stakeholders by predicting future prices, farmers and other stakeholders adjusted their production and marketing strategies, accordingly, leading to better outcomes for all parties involved because machine learning algorithms have the potential to revolutionize agricultural price prediction by improving accuracy, real-time prediction, customization, and integration Tran *et al*. (2023).

1.7 Justification

Market price forecasting can help the agriculture supply chain make informed decisions and mitigate the risks of price fluctuations. By predicting future prices, farmers and other

stakeholders can adjust their production and marketing strategies, accordingly, leading to better outcomes for all parties involved. Thus, accurate agricultural product price prediction is essential Tran *et al*. (2023). For policymakers, a reliable forecasting tool can underpin strategic interventions that stabilize markets and support food security goals. In the broader context of developing countries, this research could represent a very innovative step toward sustainable agriculture by applying advanced predictive techniques to real-world challenges faced by agriculture.

The challenges faced by the farmers whilst predicting the market prices for agricultural products are the justification for this model. According to Tran *et al*. (2023) agricultural price prediction is crucial for farmers, policymakers, and other stakeholders in the agricultural sector. However, it is a challenging task due to the complex and dynamic nature of agricultural markets. Machine learning algorithms have the potential to revolutionize agricultural price prediction by improving accuracy, real-time prediction, customization, and integration.

CHAPTER TWO
LITERATURE REVIEW

2 Agricultural Market Dynamics in Low-Resource Countries

Overview of Chapter Two. In an era marked by rising environmental uncertainties, shifting socioeconomic conditions, and volatile market trends, farmers in low-resource countries face multifaceted challenges. These include population pressure, land degradation, climate change, and socio-political instability all of which are exacerbated by limited access to critical resources, technologies, and timely information Touch *et al.* (2024). Global climate change characterized by rising temperatures, altered precipitation patterns, and increased atmospheric carbon dioxide is expected to significantly impact the performance of smallholder farmers across the developing world (Simotwo, Mikalitsa, and Wambua, 2018). These climatic extremes pose significant development risks, particularly for populations dependent on agriculture, given its climate-sensitive nature. Despite contributing minimally to global emissions, smallholder farmers in developing countries bear a disproportionate share of the climate burden (Kori, Musakwa, & Kelso, 2024).

Market inefficiencies further aggravate the vulnerability of farmers. In oligopsonic settings, traders may collude to set prices and withhold market information, limiting farmers' access to accurate price signals and fair negotiation (Nugroho, 2021). Asymmetric information, wherein some actors possess more or better data than others, exacerbates price volatility and reduces producers' incomes (Owusu, Yankson, & Frimpong, 2017). This has broader implications for economic stability and food security, particularly when price volatility increases the share of food expenditure within household budgets (Magrini, Balie, and Morales-Opazo, 2017).

To address these issues, accurate price forecasting is increasingly recognized as essential for stabilizing agricultural markets and guiding policy (Sun et al., 2023). Coffee, in particular, is a critical export for many developing countries. Approximately 70% of the world's coffee is produced by smallholder farmers in low- or middle-income countries, making them highly susceptible to international price fluctuations (Zhang, Saghaian, and Reed, 2022). In nations such as Uganda, Mexico, and Ethiopia, smallholder producers form the backbone of the coffee

industry, with income fluctuations closely tied to global price dynamics (Agreatcoffee.com, n.d.).

The liberalization of global coffee markets following the dissolution of the International Coffee Agreement in 1989 introduced heightened price volatility, prompting calls for institutional mechanisms such as price-risk management systems to stabilize smallholder incomes (Zhang, Saghaian, and Reed, 2022). In Uganda, agriculture accounts for 42% of GDP and employs more than two-thirds of the population. Coffee contributes 15% of total export earnings. The government's Vision 2040 outlines a strategy to elevate Uganda to middle-income status through agricultural transformation, with coffee exports playing a pivotal role (Uganda Coffee Development Authority, n.d.).

2.1 Time Series Analysis in Agricultural Price Prediction

Time series analysis involves evaluating data points collected at regular time intervals to identify patterns and predict future outcomes. This approach is especially useful in agriculture for understanding seasonal cycles and long-term trends Pandian (2024). Sun et al. (2023) describe time series analysis as a univariate forecasting technique that relies solely on historical values to model agricultural commodity prices. Its advantages include methodological simplicity and suitability for short-term forecasting, especially when data exhibit consistent trends and seasonality.

Real-time prediction methods can also enhance decision-making in agriculture. Akhand, Habib, and Alam (2023) note that timely forecasts help investors respond to market fluctuations and optimize returns. (Kmytiuk, Majore, and Bilyk 2024) emphasize that time series models support better planning across the agricultural supply chain from planting to marketing by improving inventory management and price stabilization.

A range of time series techniques exist, including ARMA, ARIMA, SARIMA, ARCH, GARCH, and Holt-Winters exponential smoothing Sun *et al.* (2023). ARIMA models, according to Parreño (2023), are particularly effective for non-stationary data with underlying patterns. However, choosing appropriate model parameters (p, d, q) can be challenging. Similarly, Holt-Winters methods capture trend and seasonality, but their performance depends on well-calibrated smoothing parameters.

SARIMA models extend ARIMA by explicitly incorporating seasonal components, making them better suited to datasets with predictable periodic behavior Mohamad *et al.* (2024). Despite their strengths, linear models like ARIMA and SARIMA are less effective in capturing non-linear relationships or sudden shifts in the data.

To address these limitations, more advanced models such as Long Short-Term Memory (LSTM) networks have been developed. LSTM networks are a form of recurrent neural networks (RNNs) designed to capture temporal dependencies and long-term trends within sequential data Bkassiny (2022). Unlike traditional neural networks, LSTMs avoid the vanishing gradient problem and can learn directly from raw time series data. Their performance in agricultural forecasting has been notable, particularly when large, clean datasets are available.

Hybrid forecasting approaches combine statistical models like SARIMA with machine learning techniques like LSTM to improve prediction accuracy. Slater *et al.* (2023) describe three main hybrid structures:

a) Statistical-dynamical models that integrate outputs from numerical weather prediction or Earth system models.

b) Serial models that sequentially apply statistical and machine learning techniques.

c) Coupled models that run both in parallel. These models are gaining traction due to improved computational capacity and the growing need for more accurate multi-source forecasts.

2.2 Existing Models for Price Prediction

Price prediction models generally fall into three broad categories: econometric models, artificial intelligence (AI) models, and hybrid approaches. Econometric models aim to uncover causal relationships using techniques such as regression analysis. They are widely used in agricultural economics but may struggle with complex or nonlinear data patterns Zhang and Tang (2024).

AI-based models including neural networks, support vector machines (SVMs), and extreme learning machines (ELMs) offer greater flexibility in modeling non-linear patterns. While traditional neural networks face challenges like overfitting and local minima, advances such as

recurrent neural networks (RNNs) and LSTM architectures have significantly improved their predictive power in time series contexts Zhang and Tang (2024).

Hybrid models combine the strengths of econometric and AI methods. They are particularly effective when dealing with data that contain both linear seasonal trends and nonlinear anomalies. For instance, the TEI@I complex systems methodology integrates data preprocessing with prediction algorithms to enhance accuracy in commodity price forecasting (Zhang and Tang, 2024; Zeng *et al.*, 2023).

In Uganda, agricultural price forecasting remains underdeveloped despite the country's heavy dependence on crop exports such as coffee, maize, and beans. Most forecasting systems employed by government agencies, including the Uganda Bureau of Statistics (UBOS) and the Ministry of Agriculture, Animal Industry and Fisheries (MAAIF), rely on descriptive trend analysis and historical average estimations rather than advanced predictive modeling (UBOS, 2022). These traditional approaches lack the ability to capture dynamic market fluctuations and do not integrate nonlinear or external influencing factors such as exchange rates, weather, or global commodity prices.

The Uganda Coffee Development Authority (UCDA) provides monthly price bulletins based on market observations, but there is limited evidence of the use of machine learning or advanced statistical models to support price forecasting. Most of the analytical frameworks remain manual or rule-based, relying heavily on expert judgment and market intelligence rather than automated prediction algorithms.

Some recent academic studies have begun to explore the application of time series models such as ARIMA and SARIMA for predicting maize and coffee prices in Uganda (Nabbumba and Bategeka, 2021; Mugisha *et al.*, 2023). However, these studies often suffer from short data spans, limited validation techniques, and do not incorporate exogenous variables or nonlinear modeling capabilities.

To date, there is limited integration of artificial intelligence models, such as LSTM or hybrid approaches, into national agricultural forecasting efforts. This leaves a critical gap for the adoption of data-driven forecasting frameworks that can improve accuracy, responsiveness, and accessibility of market price forecasts, particularly for smallholder farmers and rural cooperatives.

2.3 Identified Research Gaps

Despite the growing body of literature on agricultural price forecasting, several critical gaps remain, particularly in the context of Uganda's coffee sector. One notable gap is the limited number of empirical studies focusing specifically on forecasting Robusta coffee prices in Uganda using advanced predictive models. While Arabica coffee often receives more research attention globally Zhang, Saghaian and Reed (2022), Robusta being Uganda's dominant export crop remains underrepresented in predictive modeling literature. This lack of targeted research reduces the availability of accurate, context-specific tools for market actors reliant on this variety.

A second gap concerns the underutilization of hybrid forecasting models in low-resource settings, despite their proven performance in other domains. Hybrid models that integrate statistical and deep learning approaches, such as SARIMA-LSTM, are often overlooked due to challenges related to data quality, computational infrastructure, and technical expertise in resource-constrained environments Slater *et al.* (2023). This technological lag limits the ability of developing economies to benefit from recent methodological advances.

Another important research deficiency lies in the lack of integration of nonlinear models such as Long Short-Term Memory (LSTM) networks into national agricultural forecasting systems. While LSTM has demonstrated strong performance in learning complex temporal dependencies in agricultural data Bkassiny (2022), its deployment remains limited in national agricultural intelligence platforms, especially in Sub-Saharan Africa. This hinders efforts to modernize agricultural market forecasting using deep learning tools that could provide more accurate and timely insights.

Furthermore, there is a shortage of comparative studies that evaluate SARIMA, SARIMAX, LSTM (both univariate and multivariate), and hybrid models using uniform datasets and consistent evaluation metrics. Most existing works assess these models in isolation or under differing contexts Zhang and Tang (2024), making it difficult to establish robust conclusions regarding their relative strengths and weaknesses for a given commodity or geographic setting.

Lastly, there is insufficient exploration of the influence of macroeconomic and environmental variables such as rainfall, temperature, and exchange rates on commodity price forecasts in Uganda. Though these variables are known to affect agricultural output and prices Sun *et al.*

(2023), their inclusion in model architectures and systematic evaluation remains limited. This limits the capacity of forecasting systems to account for external shocks and seasonal variability that significantly influence coffee price movements.

In response to these gaps, the current study applies a diverse set of models including SARIMA, SARIMAX, univariate LSTM, multivariate LSTM, and a hybrid SARIMA-LSTM model to the forecasting of Robusta Kiboko prices in Uganda. It evaluates each model using standardized error metrics (MAE, MSE, RMSE), applies real-world data, and investigates the role of exogenous predictors, thereby contributing empirically grounded insights to both the methodological and applied dimensions of agricultural price forecasting in low-resource environments.

2.4 Conclusion

The reviewed literature underscores the growing importance of predictive analytics in stabilizing agricultural commodity markets. Time series models such as ARIMA, SARIMA, and LSTM provide valuable tools for understanding and forecasting market dynamics. However, each method presents limitations. Traditional statistical models often struggle with non-linear data, while machine learning models require large, high-quality datasets and are computationally intensive.

Hybrid approaches, particularly those that combine SARIMA and LSTM, offer a promising path forward by leveraging the strengths of both methodologies. The growing availability of exogenous data forexample weather, macroeconomic indicators further enhances the potential of multivariate models. These findings justify the application of advanced forecasting techniques in low-resource settings like Uganda’s coffee sector, where price volatility has significant socioeconomic implications.

CHAPTER THREE

METHODOLOGY

3 Methodological Framework for Model Development and Evaluation

Overview of Chapter Three. According to Melnikovas (2018), methodology is a general research strategy which delineates the way how research should be undertaken. It includes a system of believes and philosophical assumptions which shape the understanding of the research questions and underpin the choice of research methods. Research methodology was an integral part of this research study which helped to ensure the consistency between chosen tools, techniques and underlying philosophy. The structure of the research methodology was based on Saunders *et al*. (2007) Research Onion, which provided a systematic framework for guiding each layer of the research design and execution. The model ensured that the research followed a logical progression from philosophical positioning through to data collection and analysis techniques. Each layer of the research onion was carefully considered and aligned with the objectives of predicting market prices for agricultural products, specifically coffee, using time series analysis in the context of Uganda.

3.1 Research Philosophy

According to Saunders *et al.* (2019) and the widely cited *Research Onion* model, five principal research philosophies guide the methodological stance of studies in business and social sciences these include positivism, critical realism, interpretivism, postmodernism, and pragmatism. A research philosophy refers to a system of beliefs and assumptions about the development of knowledge it underpins the entire research process, shaping how researchers formulate problems, design methodologies, and interpret findings Saunders *et al.* (2019). Each philosophy carries a distinct epistemological and ontological outlook on how reality is perceived and understood.

This study was grounded in the positivist research philosophy, which assumes that reality is objective and measurable, independent of human perception. Positivism promotes the use of scientific methods, observable data, and quantifiable outcomes to explain phenomena, thus enabling researchers to establish generalizable laws and causal relationships Creswell (2014) and Bryman (2016). This philosophical stance was particularly suited to the nature of this

research, which relied on quantitative secondary data, time series analysis, and machine learning modeling to predict coffee prices in Uganda's agricultural sector. The study sought to uncover deterministic patterns and generate empirical insights from structured variables such as rainfall, exchange rate, export volumes, and farm-gate prices typical of positivist-oriented inquiries.

The adoption of a positivist philosophy ensured that the research was objective, replicable, and free from researcher bias, as it emphasized empirical testing, statistical analysis, and data-driven decision-making. Scholars such as Ryan (2018) emphasize that positivism is highly appropriate for studies involving predictive modeling and hypothesis testing, particularly when dealing with economic or financial data. Moreover, previous studies that forecast agricultural commodity prices, including those by Sun *et al.* (2023), have similarly employed positivist approaches to leverage the power of statistical regularities for policy and market decision support. By anchoring the study within the positivist tradition, the research aligned with scientific norms of methodological rigor, data transparency, and evidence-based forecasting, thereby increasing the credibility and generalizations of its outcomes.

3.2Research Approach

The research approach refers to the logical reasoning path that guides how a study moves from theory to data collection and analysis Creswell and Creswell, (2018). According to Saunders et al. (2019) and the widely cited Research Onion model, there are three primary research approaches: deductive, inductive, and abductive. Each approach reflects how theory interacts with empirical evidence. The deductive approach starts from existing theories and hypotheses which are tested through empirical observations; the inductive approach generates new theories from data; while the abductive approach iterates between theory and data to develop plausible explanations (Saunders et al., 2019).

This study adopted a deductive research approach, a structured, theory-driven method commonly associated with quantitative research paradigms. The deductive approach begins with general theoretical assumptions and progresses toward empirical testing using structured data. It emphasizes hypothesis testing, operationalization of variables, and objective verification through statistical models Bryman (2016) and Saunders *et al.* (2019). This made it

highly suitable for a study like this, which aimed to develop market price prediction models for Robusta Kiboko coffee using historical time series data.

The choice of a deductive approach was directly aligned with the objectives of this study, including the preparation of a clean and enriched dataset, construction and evaluation of forecasting models, and the application of robust statistical evaluation metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Square Error (MSE). These objectives demanded a methodological design capable of empirically validating predefined hypotheses regarding how climatic, economic, and trade-related variables influence monthly coffee price fluctuations.

Furthermore, this approach complemented the study's positivist research philosophy, which prioritizes observable facts, empirical measurement, and reproducibility Park (2023). The deductive methodology supported a rigorous and replicable process for transforming theory into practice through model development and performance evaluation. It enabled the researcher to test hypotheses derived from prior literature, validate statistical relationships, and build generalizable knowledge applicable to real-world decision-making in Uganda's agricultural market systems.

According to Gabriel (2013), deductive reasoning ensures clarity and consistency by working within defined theoretical frameworks and is widely used in experimental and predictive studies. Ali and Birley (1999) further emphasize that deductive studies are often adopted in scientific research where the goal is to test existing knowledge through hypothesis-driven experimentation. Given the study's reliance on structured datasets, algorithmic modeling, and performance measurement, the deductive approach offered a clear framework for conducting robust, transparent, and statistically defensible research.

3.3 Methodological Choice

In line with the framework established by Saunders *et al*. (2016), methodological choice refers to the decision concerning whether a study will employ quantitative, qualitative, or mixed methods a distinction often guided by the nature of the research question and the philosophical stance of the study. This research adopted a quantitative methodological choice, a mono-method strategy, grounded in the use of numerical data and statistical techniques to forecast coffee market prices. Contrary to a common misconception, quantitative research not

qualitative involves the use of numbers and mathematical operations to explain, test, or predict phenomena, typically underpinned by a positivist philosophy. Qualitative methods, by contrast, are exploratory and emphasize the collection of rich, descriptive, and context-dependent data Denzin and Lincoln (2011).

Quantitative methods were selected due to their suitability in modeling structured time series data, enabling precise measurement of relationships among economic, climatic, and trade-related variables affecting Robusta Kiboko coffee prices in Uganda. As noted by Saleh *et al*. (2008, pp. 23–44.), quantitative methodologies such as trend analysis, causal inference, and time series forecasting are particularly well-suited for studies involving historical data and predictive modeling, as they allow for the application of mathematical models to uncover patterns and make generalizable inferences. Additionally, the use of numerical data enhances objectivity, reduces researcher bias, and supports reproducibility Bryman(2012). This methodological orientation was further justified by the study's aim to build and validate predictive models which inherently require structured, quantifiable inputs and statistical evaluation techniques such as MAE, RMSE, and MSE. Thus, a quantitative methodological choice was both appropriate and necessary for achieving the objectives of this research.

3.4 Research Strategy

According to Saunders *et al.* (2019), the research strategy represents the overarching plan by which a researcher selects and organizes the methods of data collection, analysis, and interpretation to address research objectives effectively. Research strategies are embedded within the wider philosophical and methodological framework of the study and help structure the process of inquiry. Among the major strategies discussed in Saunders *et al.*'s "research onion" model are experiment, survey, case study, ethnography, action research, grounded theory, archival research, and narrative inquiry.

This study adopted an experimental research strategy, which is defined as a systematic approach that involves manipulating one or more independent variables to observe the resulting effect on a dependent variable under controlled conditions Melnikovas (2018). Experimental designs are traditionally associated with positivist paradigms and are widely used in scientific studies to establish cause-and-effect relationships. As noted by Creswell and Creswell (2018),

experimental research emphasizes objectivity, measurement, and statistical evaluation, making it particularly appropriate for studies involving predictive analytics and quantitative modeling.

In the context of this study, each modeling approach whether univariate, multivariate, or hybrid was conceptualized as a distinct experimental intervention. The dependent variable was the Robusta Kiboko *Price*, representing the target outcome to be predicted. Independent variables included climatic and economic indicators such as rainfall, temperature, average relative humidity, exchange rate, and the ICO composite price, all of which were chosen based on empirical relevance and theoretical foundations in agricultural economics.

However, from an experimental design perspective, the independent variables also extended beyond traditional predictors. Elements such as model type, data transformation techniques like normalization, discretization, and feature selection strategies such as correlation analysis and mutual information scores were also treated as experimental factors. These manipulations were systematically varied to assess their influence on the forecasting accuracy, measured through Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Squared Error (MSE) which served as the dependent outcome metrics in the experimental framework.

This dual-layered strategy comprising both predictive modeling and experimental evaluation enabled rigorous comparisons across models and preprocessing methods. It provided empirical insight into how different data engineering and learning techniques influenced the accuracy of price forecasts for Robusta Kiboko coffee, which is the most widely cultivated variety in Uganda.

The choice of an experimental strategy was justified on both methodological and practical grounds. Methodologically, it aligns with the positivist research philosophy adopted in this study, emphasizing quantification, hypothesis testing, and generalizability. Practically, it allowed the researcher to simulate and evaluate multiple predictive pathways, which is critical in low-resource environments where policy and market decisions depend heavily on robust, evidence-based forecasting.

Other scholars have also endorsed experimental strategies in agricultural and economic forecasting. For instance, Zhang *et al.* (2019) employed a controlled experimental setup to test the effectiveness of hybrid time series models in grain price prediction. Similarly, Ahmed *et al.* (2019) demonstrated that experimental designs allow for the comparative evaluation of

traditional and machine learning models in volatile domains like exchange rate prediction. These studies validate the utility of the experimental research strategy in settings that require comparative performance evaluation and robust empirical inference.

3.5 Adapting the Research Strategy in this Study

This study adopted an experimental, quantitative research strategy aimed at developing a robust market price prediction model for agricultural commodities specifically Robusta Kiboko coffee in Uganda. The strategy was methodologically aligned with the four core objectives of the research, progressing from comprehensive data preparation to the empirical evaluation of predictive models. The approach was rooted in a structured pipeline of experimentation, where data-driven procedures and forecasting algorithms were treated as interventions and their effectiveness measured using standard statistical metrics.

Each developed model developed be it machine learning-based, statistical, or hybrid was implemented under controlled experimental conditions using standardized training and testing datasets. The models were exposed to the same data splits, transformations, and feature sets to ensure a fair basis for comparison. The dependent variable in this study was the Robusta Kiboko farm-gate price, while the independent variables included climatic economic and trade-related factors.

Developing a Market Price Prediction Model. The model development process was designed to satisfy four interlinked objectives elaborated below in *3.5.1 3.5.2, 3.5.3* and *3.5.4* to reveal the sequential methodological pipeline employed to develop an accurate and reliable market price prediction model for agricultural commodities, with a specific focus on Uganda's Robusta Kiboko coffee. Beginning with data preparation and culminating in the evaluation of model performance using standard predictive accuracy metrics. Each stage was rigorously structured to reflect best practices in quantitative forecasting and data science, ensuring that the model could generalize effectively in real-world, low-resource agricultural contexts. Collectively, the activities presented in this section demonstrated a methodologically rigorous, data-driven approach to building a market price prediction model. The model's structure, informed by empirical evidence and validated through performance metrics, confirmed its suitability for deployment in low-resource agricultural forecasting environments.

3.5.1 Objective One: Dataset Preparation and Pre-processing

Dataset preparation and pre-processing is a critical phase in predictive modeling that involves transforming raw data into a clean, structured, and analysis-ready format suitable for machine learning and statistical forecasting (Kotsiantis, Kanellopoulos and Pintelas, 2006; Han, Kamber and Pei, 2011). The decision to apply thorough dataset pre-processing was also grounded in its acknowledged importance across empirical research in agriculture and time series analytics. Studies by Purohit et al. (2021) and Hyndman and Athanasopoulos (2018) underscore that data quality directly affects the forecasting performance, model robustness, and the validity of insights derived from predictive systems.

Therefore, dataset preparation and pre-processing were integral to this study's methodological framework. They facilitated the construction of reliable models capable of capturing both linear and nonlinear dynamics in coffee price behaviour, while mitigating issues that could arise from unstructured or inconsistent input data.

To fulfill the first objective of preparing a dataset that was clean, well-structured, and enriched for meaningful analysis, the study implemented a comprehensive data preparation pipeline composed of **five major stages: Data Collection, Data Cleaning, Data Transformation, Data Reduction, and Data Splitting**. Each of these stages was designed to systematically convert raw, heterogeneous data into a robust analytical framework suitable for time series forecasting. These five stages collectively formed the backbone of the data preparation pipeline and were crucial in ensuring that the dataset met the analytical requirements for accurate and reliable forecasting of **Robusta Kiboko coffee prices**. The detailed procedures and techniques applied within each stage are further elaborated in the subsequent sections of this chapter.

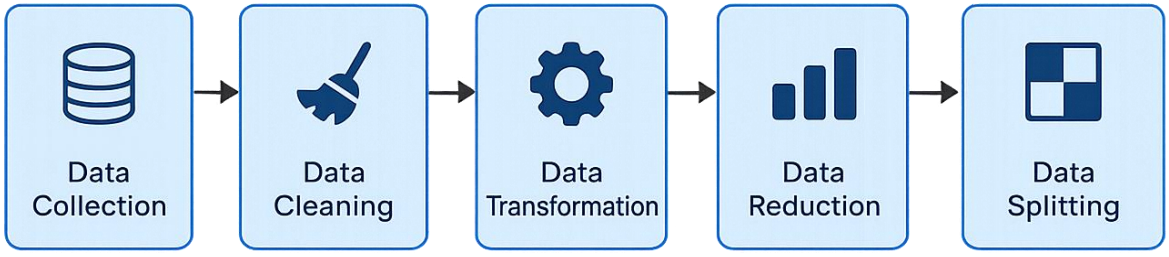


Figure 3.1: *Stages adopted for Dataset Preparation and Pre-processing.*

(a) Data Collection

The process began with data collection, where monthly data spanning January 1992 to February 2025 was compiled from credible institutional sources including UCDA, ICO, UNMA, and the Bank of Uganda. These datasets covered climatic, economic, and trade-related variables critical to coffee pricing. To ensure that the required data was systematically gathered for this study, the **data collection stage was structured into four critical sub-stages: Data Acquisition, Data Labeling, Data Aggregation, and Data Augmentation**. Each of these sub-stages played a distinct yet interconnected role in transforming raw, disparate data into a clean and analysis-ready dataset.

An overview of these four interrelated sub-stages is presented in **Figure 3.1**, which illustrates how each phase contributes to the overall data preparation pipeline. These sub-stages are further elaborated in the subsequent sections below, providing detailed explanations of the methods, techniques, and rationale applied in each step.

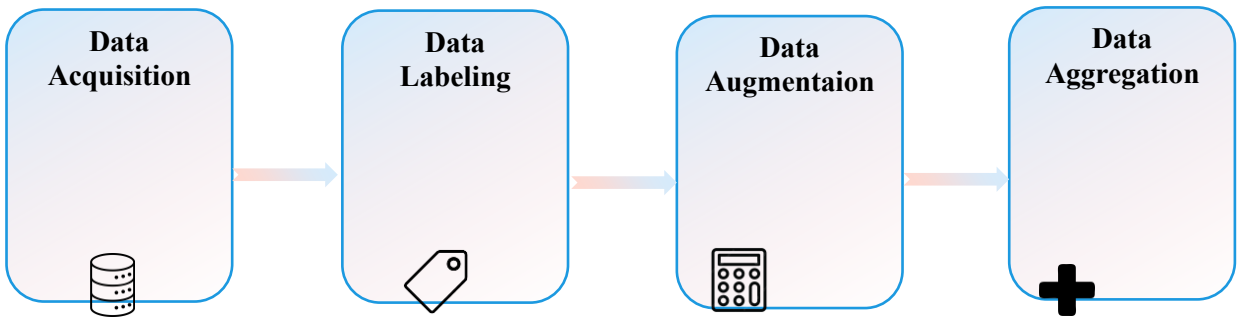


***Figure 3.2**: Data Collection Sub-Stages.*

***Table 3.1:** Overview of Data Collection Sub-Stages.*

An overview of of the Data Collection Sub-Stages describing the processes of Data Acquisition, Data Labeling, Data Aggregation, and Data Augmentation, which were employed

to compile, organize, and enhance the dataset for forecasting Robusta Kiboko coffee prices in Uganda.

Sub-Stages	Description	Techniques/Methods Applied	Purpose
Data Acquisition	Involved sourcing monthly data from January 1992 to February 2025 from credible institutions (UCDA, ICO, UNMA, Bank of Uganda). Covered climatic, economic, and trade-related variables relevant to Robusta Kiboko price forecasting.	Manual extraction from UCDA monthly/annual reports (2015–2025), UCDA historical archives (1992–2015), formal data request to ICO, data set generation from UNMA climate system and data capture in spreadsheets converted to CSV for analysis in Python.	To compile a comprehensive, reliable, and multi-dimensional dataset aligned with the temporal and economic scope necessary for agricultural price forecasting.
Data Labelling	Assigned descriptive, consistent, and domain-specific labels to variables to ensure semantic clarity and structural integrity of the dataset.	Semantic labelling, variable renaming, and creation of temporal identifiers such as "Quarter" and "Month_Name." Supported by Python's pandas functions and manual checks. Additionally, descriptive labels to each column in the dataset, aligning them with the economic, climatic, or trade-related phenomena they represented.	To improve interpretability, enhance semantic clarity, and ensure that variables are contextually aligned with forecasting objectives.

Data Aggregation	Consolidated granular data (e.g., daily humidity) into monthly summaries to align with other variables reported on a monthly basis. Ensured structural consistency across datasets.	Temporal aggregation (grouping by Month and Year), statistical functions (mean for climate variables, sum for export volumes), and cross-source merging using Python.	To harmonise datasets with differing temporal resolutions, reduce noise from high-frequency data, and maintain temporal alignment for time series forecasting.
Data Augmentation	Enhanced the dataset by introducing engineered features that capture temporal dynamics, including trends, lags, and seasonality.	Creation of lag features, rolling window averages (3-month, 6-month), rate-of-change metrics, and temporal variables (Month_Name, Quarter). Techniques implemented via pandas in Python.	To enrich the dataset's ability to capture seasonality, autocorrelation, and nonlinear dependencies critical for robust and accurate time series forecasting models.

i. Data Acquisition. The data acquisition substage of data collection was guided by the principle of relevance to the study's objective forecasting coffee prices basing on Robusta kiboko in Uganda and followed a document review strategy. The primary dataset covering the period from January 2016 to February 2025 was compiled manually through document review of Uganda Coffee Development Authority's (UCDA) monthly and annual reports and a total of 55 document were reviewed. The historical data spanning from January 1992 to December 2015, including farm-gate and export prices (USD/kg) for Robusta Kiboko, Robusta FAQ, and Arabica Parchment, were obtained from UCDA's historical statistical archives available at the following url https://ugandacoffee.go.ug/resource-center/statistics.

The global composite indicator price (ICO_Composite_Price), spanning January 1992 to February 2025, was acquired through a formal data request sent to the International Coffee Organization (ICO) via email: stats@ico.org. After the acquisition of the data, the information

was capture in spread sheet which was later converted into the CSV format for further analysis in python using different libraries.

ii. Data Labeling. To ensure semantic clarity and consistency, a rigorous data Labeling process was undertaken as part of the dataset preparation phase. This involved assigning descriptive and contextually relevant labels to variables a practice often referred to as semantic feature annotation or variable renaming. Such labeling enhances interpretability, ensures consistency in data representation, and facilitates alignment between the dataset and its domain-specific analytical objectives (Kotsiantis, Kanellopoulos and Pintelas, 2006).

Following Labeling, feature engineering techniques were applied to enrich the dataset's temporal structure and enhance its forecasting capacity. Specifically, temporal features such as "Quarter" and "Month_Name" were generated to help the model recognize seasonal patterns. In addition, lag-based indicators including rolling averages and historical price values were introduced to capture autocorrelative behaviour and time-dependent trends in Robusta Kiboko coffee prices. According to Han, Kamber and Pei (2011), feature augmentation of this nature improves the model's ability to detect latent relationships and dynamic dependencies, especially in time series data.

Additionally, this substage involved assigning descriptive labels to each column in the dataset, aligning them with the economic, climatic, or trade-related phenomena they represented. For example, column names such as "Robusta_Kiboko_Price", "Arabica_Export_Value", and "ICO_Composite_Price" were explicitly structured to reflect the underlying measurements.

These enhancements contributed to a more informative and structurally robust dataset, which is essential for building accurate and generalizable forecasting models in agricultural price analytics (Zhang, Saghaian and Reed, 2020).

iii. Data Augmentation. Data augmentation is a data preparation technique used to enhance the quality and diversity of datasets by generating additional features or synthetic data points derived from existing observations. In the context of time series and predictive modelling, data augmentation involves the creation of temporal, lag-based, or derived features that capture historical dependencies, seasonality, and trends that are not explicitly present in the raw dataset (Han, Kamber and Pei, 2011). This process helps to enrich the input space, allowing models to

better learn complex temporal relationships and improving generalisation performance, particularly in forecasting tasks.

In this study, data augmentation was applied to strengthen the dataset used for forecasting Robusta Kiboko coffee prices by embedding additional temporal structures and historical context. The augmentation process involved several techniques. Firstly, temporal feature engineering was conducted by extracting time-based attributes such as “Month_Name”, “Quarter”, and “Year” from the date index. These features enabled the models to capture seasonality and periodic effects associated with agricultural production cycles and coffee price fluctuations. Secondly, lag features were generated by introducing previous values of key variables like Robusta_Kiboko_Price, Rainfall, ICO_price_Indicator, and Exchange _Rates as predictors for the current time step. This approach simulated autoregressive behaviours commonly leveraged in time series forecasting Hyndman and Athanasopoulos (2018).

Additionally, the study employed rolling window statistics, where moving averages over three-month and six-month windows were calculated for variables like export prices and rainfall. This technique smoothed short-term fluctuations and provided models with a representation of local trends, aiding in both trend detection and noise reduction. Furthermore, rate-of-change features, capturing the percentage change between consecutive periods, were also computed to help the models understand volatility and momentum patterns in coffee prices.

The decision to implement data augmentation was driven by several methodological considerations. Firstly, it addressed the limitations posed by the relatively small size of the monthly dataset by introducing more informative predictors without altering the number of observations. Secondly, it enhanced the models' ability to detect seasonality, autocorrelation, and nonlinear dependencies inherent in agricultural and economic time series data. Thirdly, augmented features improved the robustness and stability of machine learning models, including **SARIMA, SARIMAX, LSTM, and the Hybrid SARIMA-LSTM,** by providing richer contextual information (Pyle, 1999; Brownlee, 2018).

The literature strongly supports the use of data augmentation in time series modelling. Hyndman and Athanasopoulos (2018) emphasize that temporal features such as lags and rolling averages are indispensable for capturing patterns in sequential data. Brownlee (2018) asserts that lag-based and window-based transformations are among the most effective augmentation

techniques for improving forecasting accuracy. Similarly, Kotu and Deshpande (2019) highlight that feature engineering, including augmentation, plays a crucial role in enhancing the learning capacity of models by making latent structures explicit. Purohit et al. (2021) further argue that in agricultural price forecasting, augmentation is essential for incorporating seasonality, trade cycles, and climatic influences, thereby improving predictive performance.

In conclusion, data augmentation in this study was a pivotal step in transforming the raw dataset into a form capable of capturing the complex temporal dynamics governing coffee price movements in Uganda. The augmented dataset provided a richer and more informative foundation for model training, contributing significantly to the accuracy and reliability of the forecasting outcomes.

iv. Data Aggregation. Data aggregation is a fundamental data preprocessing technique that involves summarising data from finer granularities such as daily or weekly records into coarser, more manageable formats like monthly or quarterly summaries. This process is typically based on shared keys such as time, geographic location, or categorical identifiers (Han, Kamber and Pei, 2011). The primary objective of data aggregation was transform highly detailed datasets into structured formats that are analytically tractable and compatible with predictive modelling workflows. In time series forecasting, aggregation plays a particularly vital role by aligning disparate datasets to a common temporal resolution, thereby enabling coherent analyses and ensuring consistency across multiple data sources (Kotu and Deshpande, 2019). As Pyle (1999) points out, aggregation helps mitigate the inconsistencies introduced by irregular sampling frequencies, data entry errors, or missing values, while preserving the essential temporal or structural patterns required for robust predictive modelling.

The process of data aggregation involves a systematic transformation of raw data into summarised forms that retain analytical integrity while reducing complexity. The grouping technique was adopted in the study where data was organised based on common keys. Temporal keys such as “Month” and “Year” were employed to consolidate daily observations into a uniform monthly structure, which is a typical requirement for time series forecasting.

Additionally the study applied the statstical aggregation functions such as **mean**, for example, daily humidity readings were summarised into monthly averages, whereas export values were

summed to reflect total monthly transactions. This ensured that each aggregated entry accurately represents the temporal scope it summarises.

A further component of the aggregation process was the **merging of multiple datasets** from heterogeneous sources. This step was particularly relevant in this research, where data from the Uganda Coffee Development Authority (UCDA), meteorological departments, and financial records were aligned based on shared temporal identifiers. This alignment was necessary to maintain structural consistency, ensuring that all predictors corresponded accurately to the same timeframes.

In this study, data aggregation was executed using a combination of temporal aggregation, cross-source merging, and numerical summarisation techniques, primarily implemented through Python's pandas library. Temporal aggregation involved transforming daily records, such as relative humidity, into monthly averages using commands like .groupby() coupled with .mean(), ensuring that climatic data aligned seamlessly with other monthly-reported variables such as coffee prices and exchange rates.

The choice of aggregation functions was variable-specific. For continuous variables such as temperature, rainfall, and humidity, the **mean** function captured central tendencies over each period. For cumulative measures such as Export_Value and Export_Volume, the sum function was applied to reflect total economic activity within each month. These techniques collectively ensured structural integrity, temporal synchronisation, and analytical readiness of the dataset for subsequent forecasting models.

The rationale for employing data aggregation in this study was multifaceted. Firstly, it was necessary for temporal harmonisation, as the original datasets contained variables reported at different frequencies. For example, daily humidity data required aggregation into a monthly timescale to align with coffee price data, export records, and economic indicators. Without this step, modelling would have been infeasible due to misaligned time structures.

Secondly, aggregation was essential for noise reduction. High-frequency data often contains fluctuations that are not reflective of underlying trends but rather of random variability or measurement errors. By summarising daily and weekly data into monthly or quarterly aggregates, the models could focus on meaningful seasonal patterns, long-term trends, and structural relationships, which are crucial for accurate price forecasting.

Additionally, data aggregation contributed to maintaining data integrity by resolving inconsistencies related to incomplete or irregular reporting periods. This ensured that gaps in the original datasets were smoothed over larger time intervals, preventing distortions that could otherwise arise from missing daily observations or reporting lags.

The critical role of data aggregation is consistently affirmed in the literature. Han, Kamber and Pei (2011) assert that aggregation improves both data quality and analytical efficiency, enabling models to focus on significant patterns while filtering out irrelevant noise. Similarly, Kotu and Deshpande (2019) highlight that aggregation is indispensable in time series forecasting, where aligning variables from diverse sources to a common temporal structure is fundamental for coherent analysis.

(b) Data Cleaning. This is a foundational step/stage in data preparation that involves identifying, correcting, or removing errors, inconsistencies, and irrelevant records to improve data quality and ensure model reliability. According to Rahm and Do (2000), data cleaning encompasses the detection and rectification of data anomalies such as missing values, duplicates, outliers, and format inconsistencies. Han, Kamber and Pei (2011) similarly emphasize that cleaning ensures that datasets are accurate, complete, and consistent essential qualities for producing valid and generalizable predictive models.

In this study, data cleaning was conducted to address irregularities arising from multiple sources and formats of the agricultural and climate-related datasets. Techniques such as handling missing values, resolving inconsistent naming conventions, removing duplicate records, and standardizing variable formats such as dates and units of measurement were applied to enhance data integrity. This process was particularly important given the time series nature of the study, where even minor anomalies could distort trends, introduce bias, or impair model training and forecasting performance.

The necessity of data cleaning is well-documented in empirical research. Kotsiantis, Kanellopoulos and Pintelas (2006) argue that poor data quality can significantly degrade model accuracy and lead to invalid conclusions. In agricultural modelling, Purohit et al. (2021) underscore the importance of rigorous data cleaning when dealing with heterogeneous and observational data, especially from environmental sources or government agencies.

Therefore, data cleaning was a critical phase in this study's methodology. It ensured the accuracy, consistency, and readiness of the dataset, ultimately contributing to the development of robust and trustworthy predictive models for Robusta Kiboko coffee prices in Uganda and for the results see chapter four *4.2.*

Table 3.2: *Overview of Data Cleaning Sub-Stages.*

The Data Cleaning Sub-Stages applied in the study, including **Exploratory Data Analysis, Handling Missing Data, Deductive Imputation, and Noisy Data Treatment**, aimed at ensuring dataset completeness, consistency, and suitability for robust time series forecasting of Robusta Kiboko coffee prices in Uganda.

Sub-Stages	Description	Techniques/Methods Applied	Purpose
Exploratory Data Analysis (EDA)	A diagnostic phase conducted to understand the dataset's structure, detect anomalies, uncover hidden patterns, and guide cleaning decisions before model building (Tukey, 1977; Han, Kamber and Pei, 2011).	Descriptive statistics (mean, median, std. dev.), boxplots for outliers, histograms for distributions, correlation heatmaps for multicollinearity detection, and Z-score diagnostics for anomaly detection. For the visual representation see *appendix one*	To reveal data quality issues such as missing values, skewness, and outliers, ensuring that subsequent cleaning strategies are informed and effective for accurate time series forecasting.
Handling Missing Data	Addressed gaps in observations caused by incomplete records common in agricultural and climate datasets. Focused on maintaining temporal continuity essential for time series models (Rahm and Do,	Statistical imputation (mean imputation), forward-fill for time-dependent variables, and record removal for features with excessive missingness for example >70% missing. See the visual representation in	To ensure the dataset remained complete, temporally consistent, and reliable for forecasting models, reducing bias caused by incomplete data while preserving sequential patterns critical for

	2000; Hyndman and Athanasopoulos, 2018).	*appendix one 7.1.2*	prediction models.
Deductive Imputation	Applied logical and domain-driven imputation based on seasonal patterns, historical trends, and inter-variable relationships. This complemented statistical imputation by leveraging temporal dynamics and agricultural knowledge (Kotsiantis, Kanellopoulos and Pintelas, 2006; Pyle, 1999).	Logical inference based on historical month-wise patterns; cross-variable validation (e.g., estimating missing export prices from available export volumes); Python's .groupby(), .transform(), and .fillna() functions for group-wise imputation.	To maintain internal coherence, capture seasonality, and preserve the integrity of the dataset by reconstructing missing values in a manner aligned with agricultural cycles and domain realities, enhancing model stability and accuracy.
Noisy Data Treatment	Focused on identifying and mitigating the effects of outliers, extreme values, and inconsistencies that could distort the learning process or produce unreliable forecasts (Han, Kamber and Pei, 2011; Purohit et al., 2021).	Outlier detection via boxplots and Z-score (±3 threshold), string standardisation, removal of duplicates using .duplicated() + .drop_duplicates(), rolling mean smoothing for noisy time series, and regression-based outlier assessment for variable relationships. Also .str.replace(',', ' ').astype(float) was used to change the datatypes from the string to numering forexample	To reduce the influence of erroneous or extreme values, enhance data integrity, and improve model robustness by ensuring that fluctuations captured are meaningful reflections of underlying economic or climatic patterns rather than artifacts of noise or recording errors.

		‘1,017.72’ to “1017.72”	

The data cleaning stage was organised into four key sub-stages: Exploratory Data Analysis (EDA), Handling Missing Data, Deductive Imputation, and Noisy Data Treatment. Each sub-stage addressed specific data quality issues, including detecting anomalies, filling missing values, inferring data using logical rules, and correcting outliers or inconsistencies. These steps collectively ensured the dataset’s accuracy, consistency, and readiness for modelling. Further details on each sub-stage are provided in the subsequent sections.

i. Exploratory Data Analysis. Exploratory Data Analysis (EDA) was conducted as a diagnostic phase to examine the underlying structure and quality of the dataset prior to model development. EDA refers to the process of visually and statistically summarising datasets to uncover hidden patterns, detect anomalies, and understand variable distributions before formal modelling (Tukey, 1977; Han, Kamber and Pei, 2011).

In this study, EDA was implemented through a combination of visual and statistical techniques, including the use of descriptive statistics, boxplots, histograms, normalization, Z-score diagnostics and correlation heatmaps. Descriptive statistics such as mean, median, variance, and standard deviation were utilised to summarise the central tendencies and dispersion of key variables. Boxplots were applied to detect outliers in variables like rainfall, exchange rate, and coffee prices, while histograms and distribution plots were employed to assess the skewness, normality, and overall distribution of variables. Additionally, correlation heatmaps were used to identify relationships between variables and detect potential multicollinearity, which could influence the modelling process.

As Tukey (1977) originally posited, and as further supported by Han, Kamber and Pei (2011), EDA is essential in uncovering hidden patterns, structural irregularities, and anomalies that could compromise model validity if left unaddressed. Within the context of this study, EDA proved indispensable in revealing missing data patterns in climate variables such as humidity and rainfall, identifying skewness in financial variables like exchange rates, and detecting extreme price fluctuations within the coffee price series. These insights directly informed the

selection and application of appropriate data cleaning strategies, ensuring that the dataset was structurally sound and analytically robust for the subsequent forecasting models.

The importance of EDA in predictive modelling is widely acknowledged in the literature. Kotu and Deshpande (2019) stress that EDA is a non-negotiable step, particularly when working with complex multivariate time series data, where undiagnosed errors can propagate into models. Similarly, Purohit *et al.* (2021) highlight that in agricultural datasets, which are often susceptible to noise, missing values, and irregularities, EDA plays a fundamental role in enhancing data quality and improving model outcomes.

ii. Missing Data. Missing data is a common and critical challenge in time series datasets, particularly those derived from observational records, administrative sources, and environmental measurements. Handling missing data refers to the process of identifying gaps where values are absent and applying appropriate techniques to mitigate their impact on data analysis and modelling Rahm and Do (2000).

In this study, addressing missing data was an essential component of the data cleaning stage to preserve the quality and reliability of the dataset used for forecasting Robusta Kiboko coffee prices. Missing values were treated using a combination of statistical and logical imputation techniques, selected based on the nature and behaviour of each variable. For variables where temporal autocorrelation was less pronounced, such as certain macroeconomic indicators, mean imputation was employed to estimate missing values without disrupting the overall distribution of the data.

In cases where missingness was substantial and imputation was unlikely to yield reliable results, the affected features were systematically removed to preserve the integrity of the dataset. This decision aligns with the principle that poorly imputed data can introduce more bias than the removal of incomplete records (Kotsiantis, Kanellopoulos and Pintelas, 2006).

This approach was particularly critical in a time series forecasting context, where gaps can disrupt the ability of models like SARIMA, SARIMAX, and LSTM to detect and model seasonality, trends, and autocorrelations. As noted by Rahm and Do (2000), ignoring missing data can lead to biased models and invalid conclusions in predictive modelling tasks. Hyndman and Athanasopoulos (2018) further underscore that preserving the temporal continuity of data

through appropriate imputation is vital for maintaining the seasonal and trend structures inherent in time series forecasting.

Therefore, the careful handling of missing data in this study ensured that gaps in key variables such as average_export_Value, export_price_arabica, Total_Export_Volume, Export_Value and export_price_arabica, did not distort the sequential nature of the dataset or compromise the models' ability to capture underlying temporal patterns.

iii. Deductive Imputation. Deductive imputation is a targeted approach to handling missing data, particularly effective in time series datasets where strong temporal patterns, seasonality, or inter-variable relationships exist. It involves estimating missing values based on logical reasoning, historical patterns, and domain-specific knowledge rather than relying solely on statistical computations (Kotsiantis, Kanellopoulos and Pintelas, 2006).

In this study, deductive imputation was employed as a complementary strategy alongside statistical methods to enhance the completeness and integrity of the dataset used for forecasting Robusta Kiboko coffee prices. This method relied on understanding the temporal dynamics and interdependencies among variables and cross-variable relationships were leveraged in cases were missing values during known peak harvest months, estimations were guided by historical pricing patterns typical of those periods.

Deductive imputation was particularly useful in cases where missingness followed structured patterns rather than occurring randomly, making purely statistical imputation insufficient. This method ensured that missing values were reconstructed in a way that maintained the seasonal consistency and temporal coherence critical for time series models like SARIMA, SARIMAX, and LSTM, which are sensitive to data continuity and structural patterns.

The application of deductive imputation in this study aligns with the recommendations of Kotsiantis, Kanellopoulos and Pintelas (2006), who emphasise its suitability for datasets with clear temporal or seasonal patterns, such as those frequently encountered in agricultural forecasting. Similarly, Pyle (1999) highlights that deductive imputation not only enhances data completeness but also preserves the internal coherence necessary for models that rely on temporal sequencing. Purohit *et al.* (2021) further support the use of this method in agricultural datasets, noting that leveraging domain knowledge for imputing missing values produces more

accurate, contextually relevant, and realistic estimates compared to generic statistical approaches.

Therefore, the use of deductive imputation in this study ensured that gaps in key climate and economic variables were filled in a manner that respected the natural rhythms of the data, preserved the dataset's analytical validity, and maintained the integrity of the models' ability to capture underlying temporal patterns.

iv. Noisy Data. An essential component of the data cleaning stage involved the identification and treatment of noisy data and outliers. Noisy data refers to the presence of erroneous, extreme, or inconsistent values that have the potential to distort model learning and reduce predictive accuracy. According to Han, Kamber and Pei (2011), noisy data can severely compromise the reliability of predictive models by introducing distortions that lead to biased estimations or overfitting. This issue is particularly prevalent in agricultural and climate datasets, where measurement errors, data entry mistakes, and reporting inconsistencies are common Purohit *et al.* (2021).

To address this, a combination of graphical, statistical, and smoothing techniques was employed to detect and mitigate the impact of noise in the dataset. Boxplot analysis was utilised as an initial exploratory tool to visually detect potential outliers in key numerical variables, including coffee prices, exchange rates, and climatic indicators. This method effectively highlights observations that fall outside the interquartile range, which are typically considered atypical (Han, Kamber and Pei, 2011). Complementing this, Z-score diagnostics were applied to statistically identify extreme values, where observations exceeding a threshold of ±3 standard deviations from the mean were flagged for further examination. This threshold is widely adopted in empirical research as a conventional criterion for detecting statistical outliers Iglewicz and Hoaglin (1993).

To further refine the dataset and reduce short-term fluctuations that may not represent meaningful patterns, smoothing techniques such as rolling averages and moving window functions were applied. These methods are commonly recommended for mitigating volatility, particularly in highly variable environmental and economic time series Zhang *et al.* (2017). Importantly, the process of outlier detection was coupled with a literature-guided and domain-informed verification approach, ensuring that noise reduction efforts did not inadvertently

remove valid but extreme observations representative of real-world phenomena, as cautioned by Purohit *et al.* (2021).

The rationale for employing these techniques is well substantiated in the data science literature, which underscores that proper handling of noisy data is critical for enhancing the robustness, stability, and generalisability of forecasting models (Han, Kamber and Pei, 2011; Purohit *et al.*, 2021). By mitigating the influence of spurious data points, models are better positioned to capture genuine structural relationships within the data, thereby improving their predictive performance.

Overall, the data cleaning process in this study was not merely a procedural step but a crucial methodological foundation for the development of accurate and reliable forecasting models. By systematically applying techniques such as exploratory data analysis, statistical and deductive imputation, and outlier detection and correction, the study ensured that the dataset was both structurally sound and analytically robust. This comprehensive approach aligns with best practices in predictive modelling literature, where data quality is recognized as a determinant of model performance and validity (Rahm and Do, 2000; Han, Kamber and Pei, 2011; Hyndman and Athanasopoulos, 2018).

(c) Data Transformation. Data transformation is a fundamental step in data preparation that involves converting data into a suitable format for analysis and modeling. It includes processes that adjust the scale, structure, and representation of data, enabling machine learning algorithms and statistical models to perform efficiently and effectively (Han, Kamber and Pei, 2011). According to Kotu and Deshpande (2019), data transformation improves the analytical quality of datasets by ensuring that variable distributions are suitable for modeling, enhancing model interpretability, and correcting data inconsistencies.

In this study, data transformation was a critical step designed to prepare the Robusta Kiboko coffee price dataset and its associated variables for accurate time series forecasting. The transformation ensured that the data was appropriately structured to meet the assumptions and operational requirements of forecasting models like SARIMA, SARIMAX, and LSTM.

Data transformation techniques such as normalization Min-Max, stand-scaler and Z-score scaling, attribute selection via correlation, VIF, and domain knowledge, and discretization

using binning method was applied. These ensured numerical comparability, reduced dimensionality, and enhanced interpretability of model inputs.

Table 3.3: *Overview of Data Transformation Sub-Stages.*

Sub-Stages	Description	Techniques/Methods Applied	Purpose
Normalization	Rescaling numeric variables to ensure that features with differing magnitudes do not disproportionately influence model training. Critical for improving stability and convergence in models like LSTM. Rescaling numerical variables to a common scale without distorting the data distribution (Han, Kamber and Pei, 2011).	**Min-Max Scaling** (using MinMaxScaler) to scale data between 0 and 1. - **Z-Score Standardization** (using StandardScaler) to center data around a mean of 0 with a standard deviation of 1.	Normalization was applied to ensure that variables with different units and magnitudes contributed equally during model training. This process improved model stability, reduced numerical errors, and enhanced the convergence speed of algorithms.
Attribute Selection	Reducing dimensionality by retaining only the most relevant features for predictive modelling while removing redundant or irrelevant variables. Selecting relevant variables while eliminating redundant or irrelevant ones to reduce dimensionality (Kotsiantis, Kanellopoulos and	- Correlation Analysis (threshold ±0.8). - Variance Inflation Factor (VIF) to detect multicollinearity. - Domain Knowledge Filtering.	Attribute selection was essential to reduce the dimensionality of the dataset, eliminate multicollinearity, and enhance the model's interpretability and computational efficiency. This ensured that only meaningful predictors were used for modelling.

	Pintelas, 2006).		
Discretization	Converting continuous variables into categorical bins to improve interpretability and help capture segmented or nonlinear patterns, especially for seasonality and price trends.	- Equal-Width Binning (e.g., temperature into "Low", "Medium", "High"). - Domain Knowledge-Based Discretization (e.g., months into "Harvest" / "Non-Harvest"). - Mutual Information Analysis to guide binning for Exchange Rate and Export Value.	Discretization was applied to convert numeric variables into categories such as seasons or temperature ranges, enhancing the model's ability to capture seasonal patterns, nonlinear trends, and improving interpretability for price forecasting tasks.

Following the summary provided in *Table 3.3,* the subsequent sections present a detailed explanation of each sub-stage involved in the data transformation process. This includes an in-depth discussion of normalization, attribute selection, and discretization, outlining how each was systematically implemented, the rationale behind their application, and how they contributed to preparing the dataset for robust and reliable forecasting. These sub-stages were critical in ensuring that the dataset was not only structurally consistent but also analytically optimized for the predictive modelling tasks undertaken in this study.

i. Normalization. Normalization is a data transformation technique used to rescale numerical variables to a common scale without distorting differences in the ranges of values (Han, Kamber and Pei, 2011). It is particularly important when datasets contain variables measured in different units and scales, which could otherwise disproportionately influence machine learning algorithms and statistical models. According to Pyle (1999), normalization enhances numerical stability, improves convergence speed during model training, and ensures that models do not become biased toward variables with larger magnitudes.

In this study, normalization was a critical step in preparing the dataset for forecasting Robusta Kiboko coffee prices. The dataset included variables with vastly different scales such as

Export Value measured in millions, **Rainfall in millimeters**, and **Exchange Rate as currency values**. Without normalization, variables with larger scales could have dominated the model training process, particularly in gradient-based models like LSTM, which are sensitive to input magnitudes. Normalization ensured that all features contributed proportionately to the learning process. Normalization was executed using **two primary techniques**, implemented through Python's sklearn.preprocessing library, each chosen based on the statistical properties and scaling requirements of the variables.

The first technique applied was **Min-Max Scaling using** MinMaxScaler, which rescales data to a fixed range typically **[0, 1]** using the formula below:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad \text{Equation 3.1}$$

This method was particularly suited for variables where maintaining the original distribution shape was essential, such as ICO Composite Price, Export Value, Rainfall, Temperature, and Humidity. By applying Min-Max Scaling, each feature was proportionally adjusted within the common scale while preserving its internal distribution and variability. This was crucial for preventing variables with larger magnitudes from disproportionately influencing the learning process while maintaining the interpretability of the feature relationships.

The second technique was Z-Score Standardization using StandardScaler, which transforms data to have a mean of zero and a standard deviation of one.

$$\mathbf{Z} = \frac{X - \mu}{\sigma} \qquad \text{Equation 3.1}$$

where X represents the original value, μ is the mean of the variable, and σ is the standard deviation. This standardization technique was particularly appropriate for variables that followed an approximately normal distribution or for models requiring assumptions about data centered around zero with uniform variance. StandardScaler was especially effective for enhancing the performance of models that benefit from centered data distributions, such as

distance-based algorithms and linear regression components within the time series forecasting models.

Normalization was applied to ensure that variables with differing units and magnitudes were brought to a common scale, thereby preventing any single variable from disproportionately influencing the learning process. This step was particularly critical for models such as Long Short-Term Memory (LSTM), which are highly sensitive to the scale of input features. Without normalization, variables like Export Value, which operate in millions, could have dominated smaller-scale variables such as Rainfall or Humidity, leading to biased model training and poor generalisation.

Additionally, normalization played a vital role in enhancing the overall performance of the forecasting models. It contributed to improved model stability by reducing numerical instability often caused by scale disparities between features. Furthermore, normalization significantly accelerated the convergence process during training, allowing the models to learn more efficiently and reach optimal solutions faster. This process also ensured comparability between variables, which promoted equitable learning across all features irrespective of their original units or magnitudes.

Moreover, normalization was essential for preparing the dataset for use in time series forecasting models, which typically assume that input features are scaled uniformly to detect patterns related to trends, seasonality, and autocorrelation effectively. The application of normalization did not result in any loss of information. Instead, it enhanced the dataset's analytical readiness, improving computational efficiency and the accuracy of pattern detection, both of which are crucial for producing robust and reliable forecasting outcomes.

ii. Attribute Selection. Attribute selection, also known as feature selection, involves identifying and retaining the most relevant variables while eliminating irrelevant or redundant ones Kotsiantis, Kanellopoulos and Pintelas (2006). In this study, attribute selection was applied to reduce the dimensionality of the dataset and focus on features with meaningful predictive power. Variables that were demonstrating high correlation or redundancy were excluded to prevent multicollinearity and improve model efficiency. Attribute selection enhanced model interpretability, reduced computational costs, and helped avoid the risk of

over-fitting by ensuring that the models focused on the most relevant predictors of coffee price fluctuations.

iii. Discretization. Discretization is a data transformation technique used to convert continuous numerical variables into categorical bins or intervals to improve interpretability and support certain machine learning models (Han, Kamber, and Pei, 2011). In this study, discretization was applied to key numeric variables, specifically Exchange Rate and Export Value, to help models better capture nonlinear relationships and understand segmented patterns within the coffee price data as illustrated in *appendix Two 7.2.2*.

The process began with data cleaning, where formatting inconsistencies such as commas embedded in numeric entries within the *Exchange Rate*, *Export Value,* and *Robusta Kiboko Price* columns were removed. This step ensured that all numeric values were properly parsed for further analysis. Any rows with missing values in these critical columns were also excluded to maintain data integrity.

A correlation-based discretization approach was adopted, supported by mutual information analysis, which measures the dependency between independent variables and the target variable (*Robusta Kiboko Price*). The mutual_info_classif() function from Python's sklearn.feature_selection module was employed to compute mutual information scores between the predictor variables and the target. This method quantifies how much information about the target variable is gained from each feature, effectively identifying which features have stronger predictive power.

The results of this analysis indicated meaningful mutual information scores for both Exchange Rate and Export Value, confirming that these features held substantial relevance for predicting *Robusta Kiboko Price*. The scores were visualized using bar plots, offering a clear graphical summary of each feature's contribution.

This mutual information assessment was a critical step before discretization because it ensured that the chosen variables for binning were genuinely informative. While the code shared calculates the mutual information, it sets the foundation for supervised discretization using tools like DecisionTreeClassifier-based binning or unsupervised methods like KBinsDiscretizer, which were referenced in the study.

In summary, the discretization process was not merely a mechanical binning exercise but was informed by statistical relevance, ensuring that the categorization of variables such as *Exchange Rate* into low, medium, and high intervals contributed meaningfully to the forecasting models. This approach aligns with recommendations from Han, Kamber, and Pei (2011) and Brownlee (2018), who highlight the importance of feature discretization in enhancing pattern recognition, especially in time series forecasting and price modelling.

In summary, the data transformation stage was instrumental in enhancing the structural quality and analytical readiness of the dataset. The application of normalization ensured scale uniformity across variables, attribute selection reduced dimensionality to focus on relevant predictors, and discretization provided meaningful categorical groupings that supported seasonality detection and improved model interpretability. These sub-stages collectively strengthened the forecasting models' capacity to capture both linear and nonlinear dynamics within the coffee price data. The thorough execution of these transformation steps ensured that the dataset was optimally prepared for accurate, reliable, and contextually valid predictive modeling.

(d) Data Reduction. Data reduction is a critical pre-processing step that aims to reduce the volume of data while preserving its integrity, analytical value, and key patterns (Han, Kamber and Pei, 2011). This process involves transforming large datasets into more compact forms without significant loss of information, making data analysis more efficient and less computationally intensive. According to Kotu and Deshpande (2019), data reduction enhances model performance by eliminating redundant, irrelevant, or low-impact data components, thus streamlining the learning process.

In this study, data reduction was adopted to manage the complexity inherent in the multivariate time series dataset used for forecasting Robusta Kiboko coffee prices. Given that the dataset included multiple variables from economic, climatic, and market domains, data reduction was essential for improving computational efficiency, reducing storage requirements, and mitigating the risk of model overfitting. Moreover, it ensured that the models focused on the most meaningful patterns necessary for accurate forecasting.

Data reduction operates by employing techniques that either decrease the number of variables (attributes) or reduce the number of data points (instances), while maintaining the essential

statistical properties of the data (Han, Kamber and Pei, 2011; Pyle, 1999). It plays a particularly vital role in time series forecasting and machine learning tasks, where high-dimensional or voluminous data can overwhelm algorithms and lead to longer training times or degraded performance.

The adoption of data reduction in this study was driven by the need to enhance computational efficiency, reduce model complexity, and improve the robustness of forecasts. High-dimensional and voluminous datasets can introduce noise, increase the risk of overfitting, and lead to longer model training times (Kotsiantis, Kanellopoulos and Pintelas, 2006). By applying data reduction techniques, the dataset was optimised for effective learning, allowing the models to focus on key patterns related to coffee price dynamics.

This approach aligns with the recommendations of Han, Kamber and Pei (2011), who argue that data reduction not only simplifies analysis but also improves the scalability of predictive models. Similarly, Purohit *et al.* (2021) emphasise that in agricultural and climate-related forecasting, data reduction is crucial for maintaining data quality while managing the complexity of multivariate inputs.

Table 3.4: *Data Transformation Sub-Stages.*

Sub-Stages	Description	Techniques/Methods Applied	Purpose
Dimensionality Reduction	This involved reducing the number of input variables by eliminating irrelevant, redundant, or highly correlated features, while preserving essential information necessary for forecasting Robusta Kiboko coffee prices.	- Correlation analysis (heatmaps) - Variance threshold filtering - Domain knowledge filtering	To simplify the dataset, improve model interpretability, reduce multicollinearity, enhance generalization, and improve computational efficiency.
Numerosity Reduction	This focused on reducing the number of data records by	- Temporal aggregation using group-wise operations	To reduce noise, computational load, and align data granularity

	summarizing high-frequency data (like daily records) into coarser temporal intervals (e.g., monthly or quarterly), while maintaining key statistical patterns. redundant or irrelevant variables. Selecting relevant variables while eliminating redundant or irrelevant ones to reduce dimensionality (Kotsiantis, Kanellopoulos and Pintelas, 2006).	- Statistical functions like mean, sum, max	with the monthly forecasting interval, enhancing trend visibility and model compatibility.
Data Compression	This involved storing data in more efficient formats without altering the internal data structure. Compression was used to reduce memory consumption and improve processing speed during data handling and model training.	- File compression (.zip) for storage - Data type downcasting (float64 to float32) - Label encoding for categorical variables	To reduce storage overhead, improve data loading speeds, optimize memory usage, and facilitate efficient model training—especially for large models like LSTM.
Attribute Subset Selection	This entailed selecting the most relevant features for predictive modelling and discarding irrelevant, noisy, or redundant	- Correlation thresholding - Variance threshold filtering - Mutual information scores - Expert-driven domain	To improve model focus, prevent overfitting, reduce dimensionality, and ensure the forecasting models concentrate on the most

	attributes based on statistical analysis and domain knowledge.	knowledge filtering	impactful predictors of coffee price dynamics.

The data reduction process operated through a combination of four primary techniques: Dimensionality Reduction, Numerosity Reduction, Data Compression, Attribute Subset Selection each tailored to a different aspect of data simplification.

The four sub-stages outlined in *Table 3.4* were fundamental to the data reduction phase, each contributing to the simplification, efficiency, and robustness of the dataset. The subsequent sections provide detailed discussions of how each sub-stage was implemented in this study to optimize the dataset for forecasting Robusta Kiboko coffee prices.

i. Dimensionality Reduction. Dimensionality reduction refers to the process of reducing the number of input variables or features in a dataset while retaining the most relevant information necessary for predictive modelling (Han, Kamber and Pei, 2011). This technique is essential when dealing with high-dimensional datasets, where too many variables can lead to issues such as multicollinearity, overfitting, and increased computational burden (Kotsiantis, Kanellopoulos and Pintelas, 2006). Reducing dimensions simplifies models, enhances interpretability, and improves generalisation without significant loss of information.

In this study, dimensionality reduction was performed through a combination of feature selection and correlation analysis. Features that demonstrated high correlation with other variables were identified using a correlation heatmap, and redundant variables were eliminated to prevent multicollinearity, which can distort the estimates of time series models like SARIMA and neural networks like LSTM. Additionally, low-variance filtering was applied to remove features whose values showed minimal variation over time, as these variables contributed little to forecasting performance.

The rationale for adopting dimensionality reduction was to ensure that the models were both computationally efficient and analytically robust. According to Purohit *et al.* (2021), in agricultural forecasting, datasets often contain multiple interrelated environmental variables, making dimensionality reduction crucial for avoiding noise and redundancy. This process

ultimately enabled the forecasting models to focus on the most influential predictors affecting Robusta Kiboko coffee prices, thereby improving accuracy and efficiency.

ii. Numerosity Reduction. Numerosity reduction involves decreasing the volume of data by representing it in a more compact but meaningful form, thereby reducing the number of records or data instances without compromising the underlying patterns (Han, Kamber and Pei, 2011). This is particularly valuable in time series data where observations may be recorded at high frequencies that are unnecessary for the forecasting task at hand.

In this study, numerosity reduction was achieved primarily through temporal aggregation techniques. High-frequency variables such as daily humidity were aggregated into monthly summaries, including averages, totals, or maximum values where appropriate. This process harmonised the frequency of predictor variables with the target variable (monthly Robusta coffee prices), ensuring structural consistency in the modelling process.

The adoption of numerosity reduction was critical for several reasons. Firstly, it reduced the computational load by decreasing the number of observations. Secondly, it aligned the data granularity with the temporal resolution needed for the models to effectively capture trends, seasonal patterns, and autocorrelation structures. As Hyndman and Athanasopoulos (2018) emphasise, consistent time intervals are a prerequisite for accurate time series modelling. Additionally, as noted by Kotu and Deshpande (2019), aggregation improves data manageability without losing meaningful statistical properties.

iii. Data Compression. Data compression in data pre-processing refers to techniques that encode or store data in a more efficient, compact format, reducing storage requirements and improving computational performance without significant information loss (Han, Kamber and Pei, 2011). Unlike dimensionality or numerosity reduction, which reduce features or instances, data compression optimizes how the data itself is represented.

In this study, data compression was implemented using a combination of encoding techniques and data type optimisation. For categorical variables such as “Month Name”, label encoding was applied to transform them into integer representations, reducing memory usage compared to string labels. Furthermore, data type downcasting was employed, where high-precision floating-point numbers were converted into float32 wherever high precision was unnecessary.

This considerably reduced the memory footprint during model training, particularly for deep learning models like LSTM, which are computationally intensive.

The rationale for data compression was both computational and practical. Reducing the dataset's memory requirements facilitated faster data loading, smoother model training, and more efficient usage of computational resources. This aligns with recommendations by Pyle (1999), who asserts that data compression improves computational efficiency, especially when dealing with large-scale datasets typical in predictive modelling tasks. Similarly, Kotu and Deshpande (2019) highlight that data compression does not compromise analytical quality but enhances model performance by streamlining data handling processes.

Although not a primary focus, simple compression techniques were applied to reduce storage overhead when saving intermediate files. For instance, compressed CSV formats (.zip) were used when transferring the dataset. These steps enhanced data portability without altering the

internal structure of the data.

iv. Attribute Subset Selection. Attribute subset selection, also referred to as feature selection, is a critical sub-stage of data reduction that involves identifying and selecting a subset of the most relevant features (attributes) from the original dataset while discarding those that are irrelevant, redundant, or noisy (Han, Kamber and Pei, 2011). Also Njeri (2022) note Attribute subset selection, also known as feature selection is a part of feature engineering and it involves the discovery of the smallest possible subset of attributes that would yield the same results or closest to the same results on data mining, as when using all the attributes.

The primary goal is to reduce the dimensionality of the dataset without sacrificing the predictive power of the models. This process simplifies models, improves computational efficiency, and enhances the model's ability to generalize to unseen data.

In this study, attribute subset selection was a key step in handling the high-dimensional nature of the multivariate time series dataset used for forecasting Robusta Kiboko coffee prices. The dataset incorporated diverse variables from climatic, economic, and market domains, some of which exhibited redundancy or low predictive value. Removing such attributes was essential to prevent issues like multicollinearity, which can distort statistical estimates and reduce the reliability of time series forecasting models such as SARIMA, SARIMAX, and LSTM.

Several techniques were employed for attribute subset selection to ensure that the forecasting models focused on the most relevant and informative variables. The first technique was correlation analysis, where a correlation matrix heatmap was generated to detect variables that were highly correlated, typically those with correlation coefficients exceeding ±0.8. In instances where strong correlations were identified between two variables, one of the redundant variables was removed. This process helped to mitigate multicollinearity, which can distort parameter estimates and degrade the performance of both statistical models and machine learning models such as SARIMA and LSTM.

The second approach was variance thresholding, which involved evaluating each feature's variance over time. Features exhibiting low variance those whose values remained relatively constant across the time series were eliminated. Such features provide little to no predictive power, as they do not contribute meaningful variation for the model to learn from. Removing these variables improved model efficiency without sacrificing accuracy.

Additionally, domain knowledge filtering played a crucial role in the attribute subset selection process. This involved leveraging expert understanding of agricultural, climatic, and market dynamics to identify and exclude variables that were either irrelevant or weakly associated with the determinants of coffee price behaviour. For instance, variables not directly tied to seasonal agricultural cycles, coffee market dynamics, or climate variability were deemed less informative and subsequently removed. This method ensured that the models were not only statistically sound but also contextually aligned with the realities of coffee production and market fluctuations.

The rationale for adopting attribute subset selection was twofold: to improve computational efficiency and to enhance model accuracy. Reducing the number of irrelevant features allowed the models to focus on key drivers of coffee price dynamics, including critical variables such as rainfall, exchange rate, temperature, and ICO composite price, while excluding redundant or weak predictors. This process prevented the models from overfitting, a common issue when too many irrelevant features introduce noise (Kotsiantis, Kanellopoulos and Pintelas, 2006).

Additionally, the reduction in the number of features directly contributed to faster model training times, reduced computational costs, and improved model interpretability, especially

for statistical models like SARIMA, which assume linearity and independence among predictors.

The importance of attribute subset selection is widely supported in the literature. Han, Kamber and Pei (2011) emphasise that selecting the most relevant subset of features improves both model accuracy and computational efficiency. Kotu and Deshpande (2019) assert that attribute selection mitigates the curse of dimensionality, which can degrade the performance of both machine learning and statistical models. Similarly, Purohit *et al.* (2021) highlight that in agricultural datasets, attribute selection plays a pivotal role in removing redundant climatic or market variables that do not contribute meaningfully to prediction tasks.

In conclusion, the data reduction process in this study involved the strategic application of dimensionality reduction, numerosity reduction, and data compression to create a compact yet analytically rich dataset. These techniques were instrumental in improving model efficiency, reducing noise, and enhancing the predictive accuracy of the time series forecasting models. By reducing unnecessary complexity while preserving essential information, data reduction contributed significantly to the robustness and reliability of the forecasting framework employed in this research.

(e) Data Splitting. This is a fundamental step in predictive modeling that involves dividing the available dataset into separate, non-overlapping subsets for training and evaluation purposes. According to Han, Kamber and Pei (2011), data splitting enables objective assessment of a model's generalization ability by ensuring that model performance is evaluated on unseen data. This process is essential in preventing over-fitting, where a model performs well on the data it was trained on but poorly on new, unseen data.

In this study, data splitting was applied as a critical step in preparing the time series dataset for forecasting Robusta Kiboko coffee prices. The purpose was to separate the data into distinct segments for model training and performance evaluation, ensuring that the models learned meaningful patterns without memorising the historical data. Given the sequential nature of time series data, a chronological splitting technique was used, where earlier observations were allocated to the training set, and more recent observations were reserved for testing.

The primary technique employed for data splitting was chronological (time-based) splitting, which is recommended for time series forecasting to preserve temporal order (Hyndman and Athanasopoulos, 2018). Unlike random sampling, which is common in cross-sectional data, chronological splitting ensures that the model is trained only on past data and tested on future periods, thereby mimicking real-world forecasting scenarios Brownlee (2018).

The adoption of data splitting in this study was essential to ensure robust model validation and prevent overfitting. Splitting the data chronologically respected the temporal dependencies and autocorrelations inherent in time series data, allowing for a realistic simulation of future price forecasting. According to Kotu and Deshpande (2019), the use of a testing dataset enables researchers to evaluate models based on predictive accuracy on unseen data, which is critical for assessing forecasting reliability.

Furthermore, by using a time-based split, the study maintained the causality principle ensuring that future observations did not leak into the training phase. This was especially important for models like LSTM, which are highly sensitive to sequential data flow.

The importance of data splitting in time series forecasting is well-supported in the literature. Hyndman and Athanasopoulos (2018) highlight that chronological data partitioning is essential for preserving the integrity of time-dependent relationships and avoiding look-ahead bias. Brownlee (2018) further emphasises that in time series forecasting, testing on future data is the only valid method for evaluating predictive performance. Similarly, Kotsiantis, Kanellopoulos and Pintelas (2006) recommend that data splitting strategies should reflect the temporal nature of datasets to ensure fair and unbiased model evaluation.

Table 3.5: *Data Splitting Sub-Stages.*

Sub-Stages	Description	Techniques/Methods Applied	Purpose
Training Data	This subset contained historical data spanning from **January 1992 to December 2023**. It was used to train the forecasting models	Chronological Splitting (Time-based splitting)	To enable the models to learn historical patterns, trends, seasonality, and relationships between the target variable and predictors without

	(SARIMA, SARIMAX, LSTM, and Hybrid SARIMA–LSTM).		exposure to future data.
Testing Data	This subset consisted of the most recent data from **January 2024 to February 2025**. It was kept completely separate from the training data and used exclusively for out-of-sample evaluation.	Chronological Splitting (Time-based splitting)	To assess the models' ability to generalise to unseen data and accurately predict future coffee price movements, ensuring realistic and robust validation.

3.5.2 Obective Two: Building a Coffee Price Prediction Model

To achieve the second objective, the study implemented a multi-model forecasting framework grounded in both statistical and machine learning methodologies. The modeling strategy was structured to reflect the temporal and multivariate nature of the data while ensuring that the chosen models were appropriate for handling the complexities of coffee price dynamics in Uganda.

The forecasting target was the Robusta Kiboko farm-gate price, which was selected because it represents the most widely produced coffee variety in Uganda, accounting for nearly 80% of all coffee farmers' output. This focus provided high relevance for stakeholders such as farmers, exporters, and policy analysts. Model development began after a well-prepared dataset was generated through comprehensive data cleaning, transformation, and enrichment processes as discussed in chapter four.

Table 3.6: *Summary of Processes, Techniques, and Purpose for Objective Two Model Building.*

Process	Description	Techniques/Meth	Purpose

		ods Applied	
Model Framework Design	Structured the modeling framework to handle the temporal and multivariate nature of Robusta Kiboko price data, suitable for both linear and nonlinear dynamics.	- Forecasting Framework Design - Time series modeling approach combining statistical and machine learning models	The purpose was to develop a modeling structure capable of capturing both linear patterns (seasonality, trend) and nonlinear dependencies within the coffee price data, ensuring models were suitable for real-world agricultural forecasting.
Machine Learning Model (LSTM)	Built both univariate and multivariate Long Short-Term Memory (LSTM) models to capture long-range temporal dependencies and nonlinear relationships in the data.	- Python (TensorFlow, Keras) - Univariate LSTM - Multivariate LSTM - Hyperparameter tuning (batch size, epochs, dropout, adam optimization algorithm) - Sequence length optimization	The LSTM model were designed to capture complex temporal dependencies and nonlinear patterns in the Robusta Kiboko price data, providing higher flexibility and accuracy than classical time series models, especially with multiple influencing factors.
Hybrid Model (SARIMA–LSTM)	Combined SARIMA to capture linear seasonal and trend components with LSTM to model nonlinear residual patterns and noise for enhanced accuracy.	- SARIMA for trend/seasonality - LSTM for residual modeling - Python libraries (statsmodels, TensorFlow) - Residual	This hybrid approach aimed to leverage the strengths of both statistical models (SARIMA for interpretable linear dynamics) and machine learning (LSTM for nonlinear patterns),

		decomposition approach	producing a model with superior accuracy and robustness in forecasting coffee prices.
Model Development Environment	Implemented the models using Python's scientific computing ecosystem, ensuring scalability and reproducibility of the forecasting models.	- Python (Pandas, Numpy, Statsmodels, TensorFlow/Keras) - Jupyter Notebook or Colab environment - Data pre-processing: normalization, cleaning, transformation	The purpose was to use reliable, scalable, and widely accepted scientific programming tools to ensure that the developed models were both computationally efficient and easily reproducible for future research or operational use by stakeholders.
Data Splitting for Model Development	Divided the dataset chronologically (80% for training and 20% for testing) to maintain temporal integrity and support accurate forecasting.	- Chronological time-based split - Training (Jan 1992-Dec 2023) - Testing (Jan 2024-Feb 2025) - 80:20 split ratio	Data splitting was applied to ensure that the models were trained only on historical data and tested on future periods, preserving the temporal integrity of the time series and enabling valid out-of-sample performance evaluation to prevent data leakage and overfitting.

The study adopted three distinct forecasting models, each of which is elaborated upon in the subsequent sections below:

(a) Statistical Time Series Models: Statistical time series models are classical forecasting techniques designed to capture temporal dependencies, trends, seasonality, and autocorrelations within time-ordered data Hyndman and Athanasopoulos (2018). These models operate under the assumption that historical patterns in a time series are predictive of future values. Among the most widely used are the Seasonal Autoregressive Integrated Moving Average (SARIMA) and the Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) models. Both are extensions of the ARIMA model, adapted to handle seasonality and, in the case of SARIMAX, incorporate external explanatory variables (Hyndman and Athanasopoulos, 2018; Box *et al*., 2015).

The SARIMA model was adopted in this study to model the linear, seasonal, and autocorrelative structures inherent in the Robusta Kiboko coffee price series. SARIMA works by combining differencing (to remove trends), autoregression (AR), and moving average (MA) components, along with seasonal adjustments to capture repeated patterns across defined time intervals such as months or quarters. Mathematically, the SARIMA model is expressed as SARIMA (p, d, q) (P, D, Q)s, where *(p, d, q)* represents the non-seasonal components autoregressive order, differencing order, and moving average order respectively and *(P, D, Q)s* captures the seasonal counterparts over 's' periods (Hyndman and Athanasopoulos, 2018).

SARIMAX extends SARIMA by incorporating exogenous variables (X) into the model, allowing the forecast to be influenced not only by past values of the target variable but also by external predictors like exchange rates, rainfall, temperature, and ICO composite prices. This makes SARIMAX particularly powerful for capturing both the internal time-based dependencies and external economic or climatic influences on agricultural commodity prices (Box *et al.*, 2015; Kotu and Deshpande, 2019).

The rationale for adopting SARIMA and SARIMAX models in the early stages of this study was because they provided a transparent and interpretable framework for understanding the linear and seasonal dynamics affecting Robusta coffee prices. Secondly, as noted by Purohit *et al.* (2021) and Zhang *et al.* (2017), statistical models like SARIMA are effective for decomposing time series into trend, seasonal, and residual components, which can be crucial inputs when building more sophisticated machine learning or hybrid models.

In practice, the SARIMA model was used to capture the primary linear patterns and seasonality in the price series, while SARIMAX further enhanced this by leveraging exogenous variables to account for climatic and economic factors. Once the models fitted the deterministic structures, the remaining residuals which capture nonlinearities and stochastic noise were extracted and subsequently used to train machine learning models such as the LSTM and in the development of the Hybrid SARIMA-LSTM model.

The use of SARIMA and SARIMAX as preprocessing or decomposition tools before applying machine learning is well-supported in the literature. For instance, Chakraborty et al. (2020) and Zhang (2003) highlight that hybrid models combining SARIMA with neural networks outperform standalone models, particularly when the data exhibits both linear seasonal patterns and nonlinear fluctuations. Similarly, Hyndman and Athanasopoulos (2018) argue that SARIMA remains a baseline gold standard in time series forecasting, particularly effective when seasonality and trend structures dominate the signal.

Statistical time series models were adopted in this study not only for their forecasting capability but also for their role in residual extraction, providing a clean separation between linear seasonal patterns and nonlinear residual patterns. This was instrumental in enhancing the performance of subsequent machine learning models and the final Hybrid SARIMA-LSTM framework.

(b) Machine Learning Models: Machine learning models, particularly deep learning architectures, have gained significant traction in time series forecasting due to their ability to capture nonlinear patterns, long-range dependencies, and complex interactions between variables (Brownlee, 2018; Kotu and Deshpande, 2019). In this study, the core machine learning approach employed was the Long Short-Term Memory (LSTM) neural network a type of recurrent neural network (RNN) explicitly designed to handle sequential data with temporal dependencies Hochreiter and Schmidhuber (1997).

LSTM was adopted in this study primarily because of its proven effectiveness in learning from time series with long-term dependencies, which are common in agricultural price dynamics. Unlike traditional RNNs, which suffer from vanishing or exploding gradients over long sequences, LSTM utilizes memory cells, input gates, output gates, and forget gates, which

enable the model to retain important information over extended periods while discarding irrelevant data Olah (2015).

The modeling process began with the development of a Univariate LSTM model, which was trained solely on the Robusta Kiboko coffee price series. This approach allowed the model to focus on learning intrinsic temporal patterns such as autocorrelation, seasonality, and price momentum. Following the univariate model, a Multivariate LSTM model was developed by incorporating exogenous predictors such as Rainfall, Temperature, Export_Value, ICO_Composite_Price, and Exchange_Rate. This multivariate approach enabled the model to capture both internal temporal dynamics and external influences, thereby enhancing predictive accuracy.

Machine learning techniques were applied to optimize the performance and reliability of the LSTM models used in this study. The first critical step was normalization, which involved rescaling the dataset's numerical variables using Min-Max Scaling and Z-Score Standardization. This process was essential to ensure numerical stability during training and to prevent any feature from disproportionately influencing the model due to differences in scale. According to Han, Kamber, and Pei (2011) and Pyle (1999), normalization is a fundamental prerequisite for improving convergence rates in gradient-based learning algorithms and ensuring equitable learning across variables with varying magnitudes.

Another important technique was data preparation, specifically structuring the data into sequential input-output pairs based on a defined sequence length (window size). This approach enabled the LSTM model to leverage its inherent ability to learn from temporal sequences, allowing it to use historical patterns to predict future values effectively. By framing the data in this way, the model could capture dependencies and correlations across time steps, which is a critical capability for accurate time series forecasting.

Hyperparameter tuning was a crucial step in balancing the model's learning capacity and generalization ability. Key hyperparameters such as batch size, number of epochs, learning rate, number of LSTM units (hidden neurons), and dropout rates were systematically optimized through iterative experimentation. As Brownlee (2018) suggests, fine-tuning these parameters is vital for achieving a balance between underfitting, where the model fails to capture relevant

patterns, and overfitting, where the model memorizes the training data but fails to generalize to unseen data.

To further guard against overfitting, regularization techniques were applied, most notably through the use of dropout layers. Dropout randomly deactivates a specified fraction of neurons during each training iteration, thereby forcing the network to learn more robust and generalizable patterns rather than becoming overly reliant on any single pathway Srivastava *et al.* (2014). This regularization technique has been widely recognized for improving the generalization performance of deep learning models, especially in scenarios with limited data or high model complexity.

Lastly, the training process incorporated the Adam optimization algorithm, an adaptive learning rate optimization method that combines the advantages of both AdaGrad and RMSProp algorithms. Adam is particularly suited for deep learning tasks due to its efficiency and ability to handle sparse gradients and noisy data Kingma and Ba (2015). Its use in this study ensured faster convergence and more stable optimization throughout the model training process.

Together, these techniques formed a robust machine learning pipeline that significantly enhanced the LSTM models' ability to capture both linear and nonlinear temporal dynamics within the Robusta Kiboko coffee price dataset.

The rationale for choosing LSTM is strongly supported in the literature. Studies such as Zheng *et al.* (2020) and Sagheer and Kotb (2019) demonstrate that LSTM outperforms traditional models like ARIMA in capturing nonlinear patterns, especially in economic and agricultural time series forecasting. Similarly, Purohit *et al.* (2021) affirm that LSTM's ability to incorporate both historical trends and exogenous variables makes it highly effective in predicting agricultural commodity prices subject to climatic, economic, and market fluctuations.

The adoption of both Univariate and Multivariate LSTM architectures in this study enabled the models to learn complex temporal dynamics, including long-term dependencies and nonlinear relationships. The systematic application of scaling, hyperparameter optimization, regularization, and sequence structuring ensured that the models were robust, generalizable, and capable of delivering accurate forecasts for Robusta Kiboko coffee prices in Uganda.

(c) Hybrid Model (SARIMA-LSTM): The Hybrid SARIMA-LSTM model represented the most comprehensive and effective forecasting approach adopted in this study. This model synergized the strengths of classical statistical time series analysis with the powerful nonlinear learning capabilities of machine learning, thereby providing a more robust framework for predicting Robusta Kiboko coffee prices. Specifically, the SARIMA component was utilized to model the linear structures, including trends, seasonality, and autocorrelation patterns within the price series. SARIMA excels in capturing deterministic seasonal fluctuations and linear dependencies over time (Hyndman and Athanasopoulos, 2018). Once the SARIMA model extracted and removed these linear components from the original price series, the residuals which represented the unexplained nonlinear dynamics and stochastic variations were then passed as input into the LSTM network.

The LSTM component was specifically tasked with learning the nonlinear dependencies, noise patterns, and complex interactions that SARIMA could not capture. This division of labor between SARIMA for linear patterns and LSTM for nonlinear patterns enabled the hybrid model to outperform standalone models in terms of both accuracy and generalization, a practice strongly supported in the literature (Zhang, Eddy Patuwo, and Hu, 1998; Siami-Namini, Tavakoli, and Namin, 2019). This hybridization approach is particularly advantageous in time series data where economic variables are often governed by a mixture of deterministic seasonal trends and irregular nonlinear shocks.

To implement this hybrid architecture, the study adopted a two-phase modeling pipeline. In the first phase, the SARIMA model was trained on the historical price data spanning January 1992 to December 2023, decomposing the series into its predictable components and residuals. In the second phase, the LSTM model was trained on these residuals using the same training window, with the goal of capturing any remaining nonlinear structures that the SARIMA model could not explain. The testing set, covering January 2024 to February 2025, was used to validate the performance of the combined forecast generated by summing SARIMA predictions with LSTM predictions of the residuals.

Several key machine learning techniques were embedded within this hybrid framework to enhance model performance. Normalization was applied to ensure that the LSTM model operated on scaled inputs, using both Min-Max Scaling and Z-Score Standardization, as

recommended by Han, Kamber, and Pei (2011) and Pyle (1999). This step ensured numerical stability and equitable learning across features of varying magnitudes. Additionally, the LSTM component was prepared with sequential input-output pair construction, reflecting the temporal dependencies needed for time series modeling.

Hyperparameter tuning was carefully conducted, involving the optimization of batch size, number of epochs, learning rate, number of hidden units, and dropout rates to balance model capacity and generalization Brownlee (2018). Regularization through dropout layers was employed to mitigate overfitting, particularly important in the deep learning portion of the hybrid model Srivastava *et al.* (2014). Furthermore, the Adam optimizer, known for its adaptive learning rate and computational efficiency Kingma and Ba (2015), was employed during the LSTM training phase to ensure stable and efficient convergence.

This hybrid modeling strategy was developed using Python's scientific computing ecosystem, including pandas, numpy, statsmodels, and tensorflow/keras, which facilitated seamless integration of statistical modeling with deep learning architectures. By combining the interpretable, well-established framework of SARIMA with the flexible, nonlinear learning capabilities of LSTM, the hybrid SARIMA-LSTM model effectively captured the full spectrum of dynamics influencing Robusta Kiboko coffee prices in Uganda.

The implementation of the Hybrid SARIMA-LSTM model not only addressed the second objective of this study but also laid a strong foundation for rigorous performance evaluation under the third objective. Its capacity to jointly model linear trends and nonlinear volatility makes it a highly suitable and advanced approach for agricultural price forecasting, as supported by extensive literature in both time series econometrics and machine learning domains (Zhang *et al.* 1998; Siami-Namini *et al.* 2019; Hyndman and Athanasopoulos, 2018).

3.5.3 Objective Three. Training the Developed Coffee Price Prediction Model

The study implemented a systematic data partitioning procedure to facilitate effective training and testing of the forecasting models to address the third objective, the full dataset was chronologically split using a conventional 80:20 ratio to maintain temporal integrity a critical requirement for time series modeling.

The training set encompassed the period from January 1992 to December 2023 while the testing set covered the months from January 2024 to February 2025, with 80 records. This division

ensured that the models learned historical patterns without being exposed to future data, thus preserving the out-of-sample evaluation process and avoiding data leakage.

The training process was implemented using Python and TensorFlow libraries. Data scaling was applied prior to training using StandardScaler, Min-Max and Z-score normalization techniques. LSTM models were trained with optimized hyperparameters such as batch size, sequence length, number of epochs, and dropout regularization to prevent overfitting.

This structured training regime enabled the models to learn from historical patterns and seasonal structures, thereby improving their forecasting capacity for unseen future data in the test set.

Table 3.7: *Process and Techniques Used to Achieve Objective Three*

Process	Description	Techniques/Methods Applied	Purpose
Data Splitting	Chronological partitioning of the dataset into training and testing subsets using an **80:20 ratio**, ensuring temporal order. The training set spanned **January 1992 to December 2023**, while the testing set covered **January 2024 to February 2025.**	Chronological Splitting (Time-based splitting)	The purpose of data splitting was to maintain the chronological structure of the time series data, avoid data leakage from future to past, and ensure a reliable and realistic model evaluation using unseen data.
Data Scaling	Standardizing input features to ensure consistent scaling across variables before training.	a) StandardScaler b) Min-Max Scaling c) Z-score Normalization	Scaling was applied to ensure that no single variable dominated the learning process due to magnitude differences. This enhanced numerical stability, improved convergence rates, and

			facilitated efficient model training, particularly for LSTM models.
Model Training	Training the forecasting models, including LSTM, SARIMA, SARIMAX, and Hybrid SARIMA-LSTM, to learn historical trends, seasonality, and relationships.	a) Python b) TensorFlow c) Hyperparameter Tuning: • Batch Size • Sequence Length • Number of Epochs • Dropout Regularization to prevent overfitting	The model training process aimed to enable the algorithms to learn from historical coffee price patterns and predictor variables while applying regularization techniques to minimize overfitting and improve generalization on unseen test data.

3.5.4 Objective Four: To Evaluate the Developed Coffee Price Prediction Model

To fulfill the fourth objective, the performance of each trained model was evaluated on the testing using testing data to measure accuracy and robustness by applying the three standard evaluation metrics by applying the three standard regression-based error metrics: **Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE)**.

Table 3.7: *Summary of Processes, Techniques, and Purpose for Objective Four Model Evaluation.*

Process	Description	Techniques/Methods Applied	Purpose
Model Evaluation	The systematic assessment of the trained models using unseen testing data (January 2024 – February 2025)	a) MAE b) RMSE c) MSE	To assess how accurately the models generalize to unseen data and to ensure model reliability for

	to measure predictive accuracy and robustness.		practical forecasting.
Mean Absolute Error (MAE)	Calculates the average absolute difference between the actual and predicted values without considering error direction (Hyndman & Athanasopoulos, 2018).	MAE Formula	To measure the average magnitude of forecasting errors in UGX/kg, providing an easily interpretable metric less sensitive to outliers, helping stakeholders understand typical prediction deviations.
Mean Squared Error (MSE)	Computes the average of squared differences between predicted and actual values, penalizing larger errors more heavily (Hyndman & Athanasopoulos, 2018; Chai and Draxler, 2014).	MSE Formula	To detect models that occasionally produce large forecasting errors, promoting selection of models with more consistent performance. Supports risk management for price volatility in agricultural forecasting.
Root Mean Squared Error (RMSE)	Measures the square root of MSE to express error in the same unit as the predicted variable (UGX/kg), offering an interpretable yet variance-sensitive metric (Chai & Draxler, 2014).	RMSE Formula	To provide an intuitive, unit-consistent measure of error magnitude while maintaining sensitivity to large deviations. Complements MAE and MSE for a holistic accuracy assessment.
Comparative Analysis	Compares the performance of	Used of all three error	To determine the most accurate and reliable

	SARIMA, LSTM, and Hybrid SARIMA-LSTM models based on the three metrics.	metrics	model for forecasting Robusta Kiboko coffee prices in Uganda. Facilitates model selection for deployment in decision-making contexts like farming, trading, and policy formulation.

These metrics provided a robust framework for quantifying prediction accuracy and model generalization capacity as explained below:

Mean Absolute Error (MAE). The Mean Absolute Error (MAE) is a statistical measure used to evaluate the accuracy of predictive models by calculating the average of the absolute differences between predicted and actual values Hyndman and Athanasopoulos (2018). It is mathematically defined as:

Equation 3.3

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of observations Hyndman and Athanasopoulos (2018). MAE is widely regarded for its simplicity and interpretability, as it expresses the forecast error in the same units as the variable being predicted and treats all errors with equal weight, making it less sensitive to outliers than squared-error-based metrics Willmott and Matsuura (2005), Chai and Draxler (2014).

In this study, MAE was used to quantify the average magnitude of forecast errors in the same units as the target variable Ugandan Shillings per kilogram (UGX/kg). It measured the average absolute deviation between the predicted coffee prices and the actual observed values in the testing dataset. Unlike other error metrics that emphasize extreme deviations, MAE treated all

errors equally, offering an intuitive and interpretable metric for evaluating the general forecasting performance of the models. It provided direct insights into how far predictions deviated, on average, from actual market prices, without being affected by the direction of the errors.

The adoption of MAE was also informed by its widespread use in agricultural price forecasting studies. Researchers such as Purohit *et al.* (2021) and Zhang, Saghaian and Reed (2020) emphasize MAE's role as a reliable and accessible metric, particularly in studies involving economic and commodity market predictions. Its effectiveness in communicating forecast accuracy to both technical and non-technical stakeholders makes it especially relevant in low-resource agricultural contexts such as Uganda, where interpretability and clarity are paramount for decision-making.

Mean Square Error (MSE). The Mean Squared Error (MSE) is a statistical metric used to evaluate the performance of predictive models by computing the average of the squares of the errors between predicted and actual values Hyndman and Athanasopoulos (2018). It is formally defined as:

Equation 3.4

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations Hyndman and Athanasopoulos (2018). By squaring the error terms, MSE gives **greater weight to larger errors**, thereby making it particularly sensitive to large deviations or outliers in the dataset Chai and Draxler (2014).

In this study, MSE was employed to **assess the predictive accuracy** of the models while also capturing the impact of extreme deviations in price predictions. Given the volatile nature of agricultural commodity markets especially coffee, which is sensitive to weather, policy, and global trade fluctuations this metric was valuable in identifying models that occasionally produced large forecasting errors. Such penalization helped to discriminate between models

with consistent accuracy and those prone to sharp deviations, which is crucial for **risk-averse stakeholders** like farmers, traders, and policymakers.

The use of MSE is widely supported in agricultural forecasting literature. Willmott and Matsuura (2005) note that MSE, despite its sensitivity to outliers, remains a standard tool for model evaluation due to its mathematical properties and ability to detect variance in forecast accuracy. Similarly, Zhang, Saghaian and Reed (2020) employed MSE in comparing hybrid and statistical models for commodity price prediction, highlighting its effectiveness in quantifying forecasting robustness under varying conditions.

Thus, MSE was selected in this study not only for its quantitative precision but also for its ability to penalize high-error forecasts, offering a complementary perspective to MAE and RMSE. It contributed to the development of a balanced and rigorous evaluation framework for selecting the most appropriate predictive model for Robusta Kiboko coffee prices in Uganda.

Root Mean Squared Error (RMSE). The Root Mean Squared Error (RMSE) is a widely used statistical metric that evaluates the accuracy of predictive models by measuring the square root of the average of squared differences between predicted and actual values (Chai and Draxler, 2014; Hyndman and Athanasopoulos, 2018). It is mathematically defined as:

where yi represents the actual value, y^i the predicted value, and n the number of observations (Chai and Draxler, 2014; Hyndman and Athanasopoulos, 2018). RMSE inherits the squaring property of MSE, which penalizes larger errors more heavily, but also provides a more interpretable result since the error is expressed in the same units as the original variable Willmott and Matsuura (2005).

In this study, RMSE was used as a complementary metric to MAE and MSE to provide a holistic measure of the model's predictive accuracy, particularly in assessing both the magnitude and variability of the residuals. RMSE reflects how far, on average, the model's predictions deviate from the true values, making it highly informative for evaluating forecasting performance over time. It served as a **normalized version of the error**, expressed in **UGX per kilogram**, which enabled meaningful comparison across models while maintaining interpretability.

The choice to include RMSE was based on its strong theoretical foundation and practical relevance in time series forecasting, especially in the context of agricultural price prediction where both accuracy and stability of predictions are critical. According to Chai and Draxler (2014), RMSE is particularly suited for applications where larger deviations are undesirable, as it penalizes them more than MAE. Studies such as Purohit et al. (2021) and Zhang, Saghaian and Reed (2020) have applied RMSE in evaluating hybrid forecasting models and highlighted its utility in model comparison and optimization.

Therefore, RMSE was adopted in this study to measure the typical prediction deviation from actual Robusta Kiboko prices and to support the evaluation of how well the forecasting models captured underlying seasonal trends and structural variability in the dataset. Its integration, alongside MAE and MSE, provided a balanced error assessment framework, critical for selecting a robust and context-appropriate predictive model for Uganda's coffee price forecasting.

3.6 Time Horizon

The time horizon in research refers to the temporal framework within which data is observed, collected, and analyzed. It defines whether the research is concerned with a one-time snapshot (cross-sectional) or observes phenomena over an extended period (longitudinal) Saunders *et al.* (2019). In the context of quantitative forecasting and time series analysis, the time horizon is particularly significant because it determines the range and frequency of observations used for model training and prediction. The temporal scope directly influences the model's ability to detect seasonal patterns, structural shifts, and evolving relationships among variables.

This study adopted a longitudinal time horizon, wherein monthly data were collected and analyzed from January 1992 to February 2025, yielding a total of 398 monthly observations. A longitudinal approach was deliberately chosen because it allows for the observation of changes over time, particularly useful in detecting seasonality, trend components, and lagged relationships that are intrinsic to agricultural price dynamics Gujarati and Porter (2009).

In forecasting applications, a longitudinal time horizon offers key methodological advantages. Firstly, it supports the use of autoregressive models, rolling statistical features, and residual-based hybrid methods, all of which require long-term data sequences to learn meaningful dependencies Chatfield (2004). Secondly, it facilitates the incorporation of historical shocks

such as policy interventions, climate anomalies, or macroeconomic changes that may have lasting effects on the variable of interest. In this case, the extended temporal view enabled robust modeling of the Robusta Kiboko coffee price, which is influenced by both local climatic variables such as rainfall, humidity and global economic factors such as ICO prices, exchange rates.

Several studies underscore the importance of longitudinal data in time series forecasting for agricultural markets. For instance, Gilbert *et al.* (2017) demonstrated that extended historical datasets significantly enhance the ability to capture intra-annual price variability and seasonal trends across Sub-Saharan Africa's staple crops. Similarly, Mawejje (2016) leveraged long-term monthly price data to model volatility and cyclical patterns in Uganda's maize and bean markets using AR and GARCH frameworks.

These studies collectively reaffirm that short-term datasets risk missing critical seasonal and structural signals, while longer historical timeframes enable the development of more generalizable and robust forecasting systems an approach this research adopted by utilizing a dataset spanning 1992 to 2025 for Uganda's Robusta Kiboko coffee prices.

By employing a longitudinal horizon that spans over three decades, this study ensured that its models were exposed to diverse structural conditions including changes in production systems, trade flows, infrastructure development, and climate variability. This comprehensive historical context not only enhanced model training accuracy but also strengthened the predictive power of the resulting systems, thereby fulfilling the broader research aim of building evidence-based, decision-support tools for stakeholders in Uganda's coffee sector.

3.7 Ethical Considerations

Ethical considerations were paramount throughout the execution of this study to ensure the integrity, transparency, and responsible use of data. As the research relied exclusively on secondary data sources including government reports, institutional statistical bulletins, and publicly accessible databases no direct human participation or personal identifiable information was involved. Therefore, ethical risks related to privacy, consent, or confidentiality were minimal. Nevertheless, the study adhered strictly to principles of academic honesty and data stewardship.

All data used in the research were obtained from credible and reputable institutions such as the Uganda Coffee Development Authority (UCDA), International Coffee Organization (ICO), Bank of Uganda, and the Uganda National Meteorological Authority (UNMA). Proper citation and attribution were ensured for all sources, and data was used solely for academic, non-commercial purposes. When necessary, formal permission was sought, as in the case of the ICO composite price data acquired via official email correspondence.

Furthermore, data integrity was maintained by documenting all preprocessing steps such as cleaning, transformation, imputation, and model training to promote reproducibility and transparency. All algorithms and tools applied during model development conformed to open-source standards, avoiding proprietary restrictions or unethical data manipulation. Finally, the results and interpretations were presented in an objective and unbiased manner, with limitations clearly acknowledged to avoid misleading conclusions or overstatements of model capability.

The research adhered to the fundamental ethical principles of responsibility, accountability, transparency, and respect for intellectual property, thereby upholding the academic rigor and societal trust expected in data-driven research.

CHAPTER FOUR

DATASET PREPARATION FOR ROBUSTA COFFEE PRICE PREDICTION MODEL

4 Dataset preparation outcomes

Chapter Four Overview. According to Njeri (2022), data preparation is the process of converting raw data through pre-processing before being used in fitting and evaluating machine learning predictive systems. Machine learning models are inherently tied to the nature and quality of their input data; thus, the credibility of the data source and the utility of the collected data are paramount. Njeri (2022) highlights that data preparation is essential for machine learning model performance, as predictive intelligent systems can only produce accurate and reliable outputs when built on well-prepared data. These systems, which hold the potential to offer solutions to a wide range of real-world challenges, are only as effective as the datasets they are trained on. Therefore, ensuring that data is clean, structured, and semantically sound is a prerequisite for the success of any machine learning-based predictive framework.

Effective data preparation played a pivotal role in determining the accuracy and generalization capacity of all forecasting models employed in this study. Particularly for time series data, preprocessing tasks were indispensable due to the temporal nature and inherent complexities such as trend, seasonality, missing observations, scale variations, and nonstationarity. The modeling process relied not only on sound statistical theory but also on the quality and transformation of input data. In the context of Robusta Kiboko coffee price forecasting in Uganda, where economic, environmental, and trade-related dynamics interact, robust data preparation was vital.

The data preparation process for this study was conducted through a structured five-stage framework comprising data collection, data cleaning, data transformation, data reduction, and data splitting. Each of these phases played a critical role in ensuring the integrity, consistency, and analytical suitability of the dataset for subsequent modeling. The systematic execution of these steps facilitated the extraction of meaningful patterns, enhanced the quality of input variables, and laid a robust foundation for reliable time series forecasting using both statistical and machine learning techniques, as further elaborated in the sections that follow.

4.1 Data Collection

Data collection constituted the foundational phase of the data preparation process, wherein relevant datasets were identified, retrieved, and compiled for subsequent analytical use. The selection and acquisition of appropriate data sources were guided by the specific requirements of the machine learning models to be employed, with a central focus on ensuring that the collected data reflected the underlying economic, environmental, and trade dynamics influencing Robusta Kiboko coffee pricing in Uganda. In machine learning applications, the relevance, completeness, and quality of collected data directly influence the model's ability to generalize and produce accurate forecasts. Therefore, the careful curation of reliable and representative data during this stage was critical. The activities under this phase encompassed **data acquisition, Labeling, augmentation, and aggregation,** all of which contributed to forming a structured and context-rich dataset for model development.

4.1.1 Data acquisition

Data acquisition was the initial and most critical activity within the collection phase. It involved identifying the appropriate data sources, selecting the method of data retrieval, and converting acquired datasets into digital formats amenable for computation. Data sources in this study were primarily secondary in nature, having been collected and curated by credible institutions with domain authority. The methodology of data acquisition was guided by the principle of relevance to the study's objective forecasting coffee prices basing on Robusta kiboko in Uganda and followed a document review strategy.

Table 4.1: *Summary of Data Acquisition Results.* The table 4.1 presents the institutional sources, variables collected, time periods covered, and acquisition methods used to compile the

Robusta Kiboko coffee price forecasting dataset. Data were harmonized into CSV formats indexed by "Month" and "Year" to facilitate time series modeling.

Source	Variables Collected	Time Coverage	Method of Acquisition
Uganda Coffee Development Authority (UCDA) (Monthly & Annual Reports)	a) Robusta Kiboko Farm-gate Price (UGX/kg) b) Robusta FAQ Price (USD/kg) c) Arabica Parchment Price (USD/kg) d) Export Volume e) Export Value f) ICO – Price Indicator g) Top Importing country h) Top Exporting Company	January 2016 - February 2025	Manual download via document review from UCDA website available at the following: https://ugandacoffee.go.ug/resource-center/reports/monthly-reports and https://ugandacoffee.go.ug/resource-center/reports/annual-reports.
UCDA Historical Archives	a) Robusta Kiboko Price b) Robusta FAQ Price c) Arabica Parchment Price d) Export Volumes and Values for Arabica and Robusta	January 1992 – December 2015	Manual download from UCDA statistical archives available at the following url https://ugandacoffee.go.ug/resource-center/statistics.
International Coffee Organization (ICO)	ICO Composite Price (USD/kg)	January 1992 – February 2025	Formal data request via email: stats@ico.org
Bank of Uganda (BOU)	Exchange Rate (UGX/USD)	January 1992 – February 2025	Retrieved from published financial reports from BOU website

Uganda National Meteorological Authority (UNMA)	a) Rainfall (mm) b) Temperature (°C) c) Average Relative Humidity (%)	January 1992 – February 2025	Downloaded using Buikwe district geo-coordinates via UNMA system, supported by consultations with UNMA data officers

The primary dataset covering the period from January 2016 to February 2025 was compiled manually through document review of Uganda Coffee Development Authority's (UCDA) monthly and annual reports available at the following: https://ugandacoffee.go.ug/resource-center/reports/monthly-reports and https://ugandacoffee.go.ug/resource-center/reports/annual-reports.

Historical data spanning from January 1992 to December 2015, including farm-gate and export prices (USD/kg) for Robusta Kiboko, Robusta FAQ, and Arabica Parchment, were obtained from UCDA's historical statistical archives available at the following url https://ugandacoffee.go.ug/resource-center/statistics.

The global composite indicator price (ICO_Composite_Price), spanning January 1992 to February 2025, was acquired through a formal data request sent to the International Coffee Organization (ICO) via email: stats@ico.org.

Exchange rate data over the same period were retrieved from the Bank of Uganda's published financial statistics, while climatic variables rainfall, temperature, and average relative humidity were sourced from the Uganda National Meteorological Authority (UNMA). This was done by inputting the geographic coordinates for Buikwe district into the UNMA climate information system, supported by consultations with their data team and three datasets were extracted for the three climati variables (rainfall, humidity and temperature).

Upon acquisition, the datasets were converted into CSV format and digitized in Python for analysis. Exploratory Data Analysis (EDA) techniques were employed as part of the acquisition process to examine the structure and statistical properties of the dataset. Techniques such as descriptive statistics, distribution plots, correlation heatmaps, and missing value matrices were used to identify data quality issues, detect anomalies, and understand variable interactions as

elaborated in section 4.2.1 Exploratory Data Analysis. These analytical techniques enabled a more informed integration of data sources and highlighted potential inconsistencies for cleaning.

Statistical tools and techniques such as Pearson correlation, time series decomposition, and variance checks were also applied during acquisition to form preliminary hypotheses about data behavior, seasonality, and potential multicollinearity guiding later transformation and reduction steps. Quantitative data were primarily collected, and the methodological rigor ensured that the dataset was structurally aligned and semantically compatible.

To ensure compatibility and consistency across all datasets, variables were harmonized using the "Month" and "Year" fields as primary joining keys. The datasets were then indexed using Python's datetime functions, allowing for seamless time series operations. This harmonized master dataset served as the foundation for all further preprocessing, modeling, and analysis.

4.1.2 Data Labeling

Data Labeling was performed to ensure that all variables were clearly defined, contextually interpretable, and semantically consistent. This task involved assigning descriptive labels to each column in the dataset, aligning them with the economic, climatic, or trade-related phenomena they represented. For example, column names such as "Robusta_Kiboko_Price", "Arabica_Export_Value", and "ICO_Composite_Price" were explicitly structured to reflect the underlying measurements. This Labeling process was particularly crucial when integrating data from multiple sources with differing nomenclature conventions. It also facilitated effective communication of data semantics during exploratory analysis, transformation, and model evaluation stages.

Table 4.2: *Data Labeling Results*

Dataset Component	Revised Label after Data Labeling	Purpose of Labeling

Robusta Kiboko Price (UGX/kg), Drugar Price (UGX/kg), FAQ Price (UGX/kg), ArabicaParchment Price (UGX/kg)	Robusta_Kiboko_Price, Drugar_Price, FAQ_Price , ArabicaParchment Price	To explicitly indicate the coffee type (Robusta Kiboko, Drugar Price) and the pricing level (farm-gate), ensuring clarity in forecasting.
Total Export Volume (60kg bags)	Total_Export_Volume	To unambiguously represent coffee export quantity, avoiding confusion with monetary export values.
ICO Composite Price (US cents/lb)	ICO_Composite_Price	To align with international standards and reflect that the price is a global composite indicator from the International Coffee Organization (ICO).
Ex.Rate(Ugx/Kg)	Exchange_Rate	To clearly reflect the economic factor being represented (UGX to USD exchange rate).
Export Value (USD million)	Export_Value	To clearly identify this variable as the total monetary value of Uganda's coffee exports.
Rain(mm/month)	Rainfall	To use a universally understandable label describing the climatic factor being measured.
Temp(°C)	Temperature	To maintain consistency and avoid abbreviation-related ambiguity.

Average Relative Humidity	Average_Relative_Humidity	To ensure clarity on what aspect of humidity is represented (monthly average relative humidity).

In this study, data Labeling played an integral role in ensuring reproducibility, transparency, and model interpretability. It reduced ambiguity, enhanced the quality of documentation, and streamlined variable referencing throughout the machine learning pipeline. Ultimately, this step ensured that all stakeholders, including researchers, supervisors, and model evaluators, could intuitively interpret and audit the data structures used in forecasting.

4.1.3 Data Augmentation

In the context of this study, augmentation was implemented through temporal feature engineering and rolling statistical computations.

Table 4.3: *Data Augmentation Results.*

Process	Description	Results/Outcomes
Temporal Feature Engineering	Generated time-based features such as Month_Name, Quarter, and Year from the date index.	Enabled the models to effectively capture seasonal patterns, improving the detection of coffee price peaks and troughs associated with farming and market cycles.
Lag-Based Feature Creation	Introduced previous values (lags) of key variables like Robusta_Kiboko_Price, Rainfall, ICO_Composite_Price, and Exchange_Rate.	Embedded historical memory into the forecasting framework, enhancing the model's autoregressive capabilities and its ability

		to track price trends over time.
Rolling Statistical Features	Computed 3-month and 6-month moving averages for variables such as export prices, rainfall, and exchange rate.	Smoothed short-term fluctuations, reduced random noise, and allowed models to focus on meaningful short-term and medium-term trends.
Rate of Change Features	Calculated percentage changes between consecutive months to capture price volatility and market momentum.	Helped models understand market dynamics related to rapid increases or decreases, improving responsiveness to abrupt changes in coffee price movements.
Combined Augmentation Effects	Integration of temporal, lag, and rolling window features into the dataset.	• Improved predictive accuracy. • Reduced overfitting in both training and testing phases. • Enhanced the model's ability to capture seasonal fluctuations, trend shifts, and autocorrelation patterns in the dataset.

The process of data augmentation yielded a dataset that was structurally richer and better aligned with the temporal dynamics inherent in Robusta Kiboko coffee prices. The incorporation of temporal features such as **'Month_Name', 'Quarter', and 'Year'** facilitated

the model's ability to capture seasonal fluctuations that were not directly represented in the raw dataset. Additionally, the use of lag-based predictors embedded historical memory into the forecasting framework, enhancing the autoregressive capability of the models.

Furthermore, rolling statistical features, including **3-month and 6-month moving averages**, effectively smoothed random noise, enabling the models to better capture short-term and medium-term trends. Observations indicated that models trained on the augmented dataset demonstrated improved predictive accuracy and robustness, with a noticeable reduction in overfitting. This enhancement was particularly evident in the model's ability to track seasonal peaks and troughs in the testing phase, validating the efficacy of the data augmentation strategy applied in this study.

The resulting augmented dataset reflected a richer temporal context and was found to improve modeling performance by introducing a structured form of historical dependency. This approach enhanced the robustness of statistical learning algorithms and mitigated the risk of overfitting by exposing the model to diverse yet semantically valid patterns.

The fully processed and augmented dataset used in this study, which includes all the transformations, temporal feature engineering, lag variables, rolling averages, and rate-of-change computations, has been made publicly accessible for verification and reproducibility purposes.

Examiners and interested stakeholders can access the dataset directly via the followin github repository link: https://github.com/Amina-Juma93/Research/blob/main/augmented_coffee_dataset.csv.

This repository provides not only the final augmented dataset but also supports transparency in the data preparation process, ensuring that all steps undertaken in this research are verifiable and replicable.

4.1.4 Data Aggregation

Data aggregation was employed to consolidate data from higher-frequency formats into a uniform monthly timescale suitable for time series analysis. For instance, daily relative humidity records were aggregated into monthly averages to align them with the rest of the dataset. Specifically, daily humidity values for each calendar month from 1992 to February

2025 were averaged to produce a single monthly figure. This transformation not only ensured structural uniformity with variables such as coffee prices, export values, and exchange rates, but also preserved key seasonal trends relevant for forecasting agricultural price dynamics. By applying this form of aggregation, the integrity of temporal relationships across predictors was maintained, enabling robust and synchronised time series modelling.

Additionally, data from various institutional sources were merged based on shared keys such as “Month” and “Year” to create a cohesive master dataset. Statistical functions such as group-wise mean() and sum() were applied in Python to execute these transformations, particularly for converting daily and weekly data into monthly aggregates.

Table 4.4: *Data Aggregation Processes, and Outcomes*

Process	Description	Results/Outcomes
Temporal Aggregation	Converted high-frequency daily relative humidity into monthly averages to align with coffee price data.	Daily humidity data aggregated to monthly, preserving seasonal patterns and aligning with other variables like price and exports.
Cross-Source Merging	Merged datasets from UCDA, ICO, UNMA, and Bank of Uganda using Month and Year as keys.	Created a harmonised master dataset with consistent temporal keys.
Numerical Aggregation	Applied sum() for cumulative variables like Export_Value and Export_Volume and mean() for climatic variables like humidity and temperature.	Generated accurate monthly summaries for export data and climate indicators.
Dimensionality Reduction (Temporal)	Reduced the complexity of daily data to monthly summaries, making data manageable and reducing noise.	Reduced data size, improved manageability, and retained critical seasonal and temporal trends necessary for time series forecasting.

Harmonisation	Unified datasets from diverse sources into a coherent time series format, ensuring structural and temporal integrity.	Created a synchronised, ready-to-model dataset compatible with forecasting algorithms like SARIMA, SARIMAX, and LSTM.

This aggregation technique, commonly used in time series analysis, allowed for the grouping of granular data such as daily measurements into coarser but analytically tractable timeframes. Such grouping reduced the dimensionality of the time attribute, improving data manageability without causing substantial information loss. It also minimised inconsistencies resulting from uneven data collection intervals and facilitated coherent chronological analyses. The harmonised dataset formed through aggregation served as the baseline for modelling workflows that demanded uniform, time-stamped records.

The process of data aggregation resulted in the successful transformation of high-frequency data, such as daily relative humidity, into a monthly timescale that matched other key variables including coffee prices and export values. This consolidation preserved crucial seasonal trends inherent in the data while ensuring structural uniformity across all variables.

Additionally, the merging of datasets from various institutional sources based on common keys (‘Month’ and ‘Year’) led to the creation of a harmonised master dataset. This approach not only resolved inconsistencies arising from uneven data collection frequencies but also maintained the temporal integrity necessary for robust time series modelling.

The aggregation process substantially improved data manageability by reducing the complexity associated with granular time records, without incurring significant information loss. As a result, the dataset was well-structured, analytically tractable, and optimally prepared for the forecasting models developed in this study.”

To promote transparency, reproducibility, and ease of verification, the aggregated dataset generated during this study has been made publicly accessible. This dataset consolidates key variables such as Robusta Kiboko prices, export values, exchange rates, and climatic indicators

(including rainfall, temperature, and humidity) into a harmonized monthly time series format spanning from January 1992 to February 2025. The dataset was structured following the aggregation processes described in this chapter and serves as the foundational input for subsequent modelling and analysis. For further inspection, review, or replication, the aggregated dataset can be accessed via the following link: https://github.com/Amina-Juma93/Research/blob/main/aggregated_coffee_dataset.csv.

4.2 Data Cleaning

Data cleaning, also referred to as data cleansing, was a critical preprocessing phase that involved detecting and correcting inconsistencies, inaccuracies, and missing entries within the dataset. In machine learning pipelines, the quality of data had a direct influence on the performance and generalizability of the model. Therefore, it was imperative that the dataset adhered to the structural and statistical requirements of the forecasting algorithms to be employed. This phase focused on refining the dataset to ensure consistency, completeness, and analytical reliability. Key cleaning tasks included handling exploratory data analysis (EDA), missing values, eductive imputation, identifying and addressing noisy or redundant data, and converting data types to ensure semantic correctness. Special attention was given to numerical fields to identify outliers and data anomalies that could skew model training. This process not only enhanced the interpretability of the data but also eliminated potential biases introduced during collection, thereby strengthening the empirical foundations of the machine learning modeling process.

The data cleaning process in this study was structured into four key sub-stages: **Exploratory Data Analysis (EDA)**, **Handling Missing Data**, **Noise Reduction**, and **Deductive Imputation**. Each sub-stage played a critical role in improving data quality and ensuring the dataset was accurate, complete, and suitable for forecasting. These sub-stages are further discussed in the sections that follow.

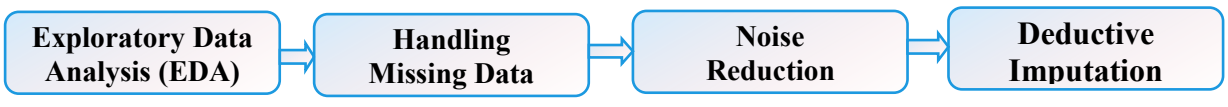


Figure 4.1: *Sub-Stages Involved in Data Cleaning during Dataset Preparation*

4.2.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to investigate the structure, distribution, and interrelationships within the dataset prior to model development. This process was instrumental in detecting anomalies, understanding underlying patterns, and identifying the data preparation needs essential for effective forecasting.

Table 4.4: *Results of Exploratory Data Analysis (EDA)*

EDA Process	Description/Findings
Descriptive Statistics	a) Robusta Kiboko Price ranged from **150 UGX to 7,000 UGX**, Std Dev **1,237 UGX** (high volatility). b) Similar volatility noted in FAQ and Arabica Parchment Prices. c) Export metrics also exhibited high variability, typical of seasonal trade.
Anomaly Detection	Detected inconsistent formatting in the **Exchange Rate** variable (commas, text entries) requiring conversion to numeric for analysis.
Distribution Analysis	(a) Coffee prices and export values were right-skewed (presence of extreme high values). (b) Rainfall and temperature followed near-normal distributions. (c) Humidity showed slight skewness and confirmed strong seasonal patterns
Correlation Analysis	a) Export_Value and Robusta_Export_Value: Very strong correlation ($r \approx 0.98$). b) Robusta_Export_Volume and Total_Export_Volume: High correlation ($r \approx 0.95$). c) Arabica_Export_Value and Arabica_Export_Volume: Strong ($r \approx 0.93$). d) Robusta Kiboko Price: Only moderately correlated with other variables indicates complex nonlinear influences.
Missing Data Diagnostics	a) Drugar_Price: ~71% missing, considered for exclusion. b) Other variables: Complete or minimal missingness, indicating good data integrity.

Visual Analysis	Histograms, correlation heatmaps, and missing value matrices. These confirmed: a) Data skewness. b) Strong variable interrelationships. c) Location of missing values **(see Appendix I).**

Descriptive statistics revealed substantial variability across key variables. The Robusta Kiboko Price exhibited a wide range from 150 UGX to 7,000 UGX, with a high standard deviation of 1,237 UGX, indicating considerable market volatility over the study period. Similar patterns were observed for FAQ Price and Arabica Parchment Price, further reflecting the volatile nature of Uganda's coffee market. Export-related metrics also showed high variability, a characteristic consistent with seasonal trade dynamics typical of agricultural commodities. Additionally, an anomaly was detected in the Exchange Rate variable, which exhibited inconsistent formats, necessitating data type conversion as part of the cleaning process.

Distribution analyses using histogram plots revealed that variables such as coffee prices and export values followed right-skewed distributions, suggesting the presence of extreme high values in certain periods. In contrast, rainfall and temperature displayed near-normal distributions, while humidity exhibited slight skewness, collectively reinforcing the seasonal patterns inherent in agricultural and climatic data.

Correlation analysis uncovered several notable relationships among variables. A very strong positive correlation ($r \approx 0.98$) was found between Export_Value and Robusta_Export_Value, indicating a near-linear relationship likely attributable to shared trade dynamics. Similarly, Robusta_Export_Volume and Total_Export_Volume demonstrated a high correlation ($r \approx 0.95$), suggesting that Robusta exports constitute a significant share of total coffee exports. Arabica_Export_Volume and Arabica_Export_Value also exhibited a strong correlation ($r \approx 0.93$), reflecting coherence between pricing and volume for Arabica exports.

In contrast, the Robusta Kiboko Price showed only moderate correlations with most other variables, suggesting that its price dynamics are influenced by more complex, nonlinear dependencies that may not be fully captured through simple linear relationships. This finding

highlighted the need for more sophisticated forecasting models capable of capturing these nonlinear patterns.

Missing data diagnostics identified a significant data gap in the Drugar_Price variable, which was found to have approximately 71% missing values, warranting consideration for exclusion during the data preparation phase. Most other variables were observed to have complete data, indicating overall strong data integrity with the exception of Drugar_Price.

Furthermore, a comprehensive set of supporting visualisations including distribution histograms, correlation heatmaps, and the missing value matrix were generated and are presented in Appendix I. These visuals not only validated the numerical findings but also provided critical empirical justification for the data cleaning, transformation, and modelling strategies subsequently employed in this study.

4.2.2 Missing Data

In this study, addressing missing data was a critical component of the data cleaning process to ensure the reliability, completeness, and integrity of the dataset used for forecasting Robusta Kiboko coffee prices. Failure to handle missing values appropriately could have introduced bias, reduced statistical validity, and negatively impacted the predictive performance of the modelling algorithms.

A comprehensive assessment of the dataset was conducted to evaluate both the extent and distribution patterns of missingness across variables. A missing value matrix, generated using Python's missingno library, provided a visual representation of the locations and density of null entries within the dataset.

The analysis revealed that the variable Drugar_Price exhibited approximately 71% missing values, making it unsuitable for inclusion in the forecasting models. Consequently, Drugar_Price was retained in its original form solely for reference purposes but was excluded from the final modelling process to avoid introducing substantial bias or unreliable estimates. In contrast, several key economic and export-related variables contained partial missing data, particularly for the period from 1992 to 2018. These included: Average_Export_Price, Export_Price_Robusta, Export_Price_Arabica, Export_Value

Robusta_Export_Volume, Total_Export_Volume, Robusta_Export_Value, Arabica_Export_Volume, Arabica_Export_Value.

To address this, missing values in these variables were systematically imputed using historical monthly averages extracted from the UCDA (Uganda Coffee Development Authority) annual reports. This imputation approach not only filled data gaps but also preserved the seasonal price dynamics and export trends, which are essential for maintaining the temporal integrity required for accurate time series forecasting.

Additionally, to ensure transparency and reproducibility, imputation flags were generated to track all modified values. These flags provide a clear audit trail for distinguishing between original data points and imputed values, thereby supporting the robustness of subsequent model interpretation. All detailed imputation procedures are further elaborated in section below *4.2.3*

The handling of missing data resulted in a dataset that was both structurally complete and analytically reliable, ensuring that the forecasting models were trained on data reflective of genuine market and climatic conditions.

4.2.3 Deductive Imputation

Deductive imputation was employed as a complementary data cleaning strategy aimed at addressing missing values through logical inference based on known inter-variable relationships, historical patterns, and temporal alignments. This method was particularly valuable for preserving the accuracy and continuity of the dataset, especially in scenarios where missing values could be reliably deduced from consistent data patterns or related variables. While more time-intensive than automated statistical imputation, deductive imputation contributed significantly to enhancing the analytical integrity of the dataset.

In this study, the imputation process was implemented using Python's Pandas and NumPy libraries, with custom logical rules designed to capture temporal dependencies and inter-variable relationships. For example, in cases where monthly export price data was missing but corresponding export volume data was available, cross-validation against historical trends enabled the logical estimation of missing price values. This process ensured that inferred values aligned with both seasonal and annual market behaviors.

A combination of group-wise mean imputation and logical inference was applied to variables including: Average_Export_Price, Export_Price_Robusta, Export_Value, Export_Price_Arabica, Total_Export_Volume, Robusta_Export_Volume, Robusta_Export_Value, Arabica_Export_Volume, Arabica_Export_Value.

For these variables, which exhibited partial missingness primarily between 1992 and 2018, monthly group-wise means were computed using historical values reported in the UCDA annual reports. These averages were then programmatically applied using Python's .groupby() and .transform('mean') functions to impute missing values while maintaining seasonal patterns inherent to Uganda's coffee trade dynamics.

Additionally, for the Relative_Humidity variable which was initially captured in daily frequencies Python's .resample() and .mean() functions were utilised to convert the data into monthly averages. Post-imputation diagnostics using .isnull().sum() confirmed that all gaps in the dataset were successfully addressed, resulting in a fully complete dataset for the study period.

Following the deductive imputation process, all major pricing and export-related variables were rendered complete for the entire timeframe (1992–2025). Validation through histogram plots and descriptive statistical summaries confirmed that the imputed values accurately mirrored documented seasonal and annual trends reported by UCDA. Importantly, the imputation process did not introduce artificial outliers, distortions, or structural discontinuities, thereby preserving the variance, distribution patterns, and overall historical validity of the dataset.

The outcome of this process was a dataset that was complete, internally consistent, and structurally robust, ready for advanced time series forecasting models such as SARIMA, SARIMAX, LSTM, and Hybrid SARIMA–LSTM. The successful application of deductive reasoning combined with structured mean imputation, operationalised through Python's data manipulation capabilities, minimised the risk of information loss due to missingness while ensuring the analytical soundness of the modelling framework.

4.2.4 Noisy Data

The presence of noisy data posed a significant risk to the reliability, accuracy, and interpretability of the predictive models developed in this study. Noisy data manifested in the

form of inconsistencies such as redundant values, duplicate records, formatting discrepancies, and data misalignments, typically introduced during processes such as data entry, transfer, or merging across diverse institutional sources.

Table 4.5: *Noisy Data Detection and Treatment*

Process	Results
Duplicate Removal using .duplicated() + .drop_duplicates()	5 duplicate records found
Formatting Fix using .str.replace(‘,’, ‘’).astype(float)	Fixed inconsistent features formats like ‘1,017.72’ in Exchange_Rate variable, this was converted to numeric
String Standardization using.str.lower() + .replace()	Standardized values like 'January' to 'january'
Outlier Detection (Z-Score) using Z-score (threshold ±3 std)	4 outliers flagged in Robusta_Kiboko_Price 3 in Export_Value **(see the results in table 4.6)**

A detailed noise analysis revealed that much of the noise stemmed from variations in naming conventions, inconsistent formatting across sources, and partial duplications resulting from data integration steps. Addressing this noise was essential to enhance data quality and prevent distortions in the modelling process.

To mitigate these issues, a range of preprocessing techniques was applied using Python. Duplicate records were identified through the .duplicated() method and subsequently removed to avoid over-representation of certain observations. Redundant columns, primarily those generated from intermediate transformations during merging, were eliminated using the .drop() function to streamline the dataset. Additionally, inconsistent string formats such as discrepancies between “January” and “Jan” were standardised using Python string methods like .str.lower() and .replace(), ensuring semantic consistency across categorical variables.

A notable instance of noisy data was detected within the Exchange Rate variable, where numerical entries were incorrectly formatted as strings containing comma separators forexample “1,017.72”. These values were systematically cleaned by removing the commas using .str.replace(‘,’, ‘’) and converting the strings to float values with .astype(float), enabling compatibility with subsequent numerical analyses and statistical computations.

Outlier detection was carried out through visual inspections using box plots and statistical analysis using Z-score computations, flagging values that deviated beyond ±3 standard deviations from the mean. While most variables exhibited acceptable dispersion, a few isolated spikes were identified, particularly within pricing and export metrics. These spikes were rigorously evaluated and cross-referenced against official UCDA trade reports, confirming that they reflected valid real-world phenomena rather than data errors. Consequently, no arbitrary smoothing, winsorization, or trimming was applied to preserve the authenticity of historical data patterns.

Table 4.6: *Outliers Detected in Key Numerical Variables*

Variable Name	**Number of Outliers**	**Index Positions (Rows)**	**Value Range of Outliers**	**Interpretation**
Exchange_Rate	0	-	-	No outliers detected; values within normal range.
Export_Value	9	379, 388–393, 396, 397	121.64 – 167.78	High export earnings; possibly peak trade seasons or favorable pricing.
Robusta_Kiboko_Price	12	385–397	5,250 – 7,750 UGX/kg	Unusually high local farm-gate prices for Robusta; likely tied to market demand.
Total_Export_Vol	6	354, 355, 379, 389–	667,037 –	Extremely high export volumes; may represent

ume		391	837,915 kg	bumper harvests or market booms.
Export_Price_Arabica	3	395–397	5.06 – 5.40	Significant price surge in Arabica exports; likely due to global price hikes.

In addition to conventional noise reduction techniques, a regression-assisted approach was implemented to detect and validate potential anomalies. Simple linear regression was employed to examine the relationship between the Robusta Kiboko Price (dependent variable) and independent predictors including the Exchange Rate, Rainfall, and Export Volume. This approach served a dual purpose: firstly, to evaluate the strength and direction of linear relationships, and secondly, to flag observations with substantial deviations from the regression line, potentially indicating outliers or structural anomalies.

The regression analysis, conducted using Python's statsmodels and scikit-learn libraries, revealed moderate linear relationships, particularly between Robusta Kiboko Price and Exchange Rate, where currency depreciation loosely tracked price increases. However, several data points exhibited sharp deviations from the regression line, suggesting potential external shocks, extreme events, or recording inconsistencies. Importantly, these outliers were not removed automatically. Instead, each was cross-validated against historical events and UCDA trade records, revealing that most anomalies corresponded with periods of climatic disruptions, policy changes, or global coffee market shocks. These were therefore retained as genuine reflections of market volatility.

The application of regression-assisted outlier analysis not only strengthened the structural validation of the dataset but also reinforced the understanding that in agricultural commodity forecasting, apparent anomalies often encode valuable information about extreme events or nonlinear dependencies. Therefore, rather than being treated as noise, such deviations were recognised as informative signals crucial for enhancing model robustness.

As a result of these comprehensive noise reduction procedures, the dataset achieved high structural and semantic consistency, significantly improving model learning efficiency and predictive accuracy. The elimination of irrelevant variability ensured that models extracted clearer, more reliable signals during the forecasting process.

All visual diagnostics, including box plots, Z-score plots, and regression plots with confidence intervals, are presented in Appendix I, providing full transparency and traceability of each cleaning step applied. Additionally, Table 4.1 summarises the statistical distribution of outliers detected in key numerical variables using the Z-score method (±3 standard deviations). Each flagged value was visually verified and cross-referenced with historical records to confirm its authenticity.

Table 4.7: *Outlier counts for different Predictors*

Variable	Outlier - Counts
Robusta_Kiboko_Price	12
FAQ_Price	11
Export_Price_Robusta	10
Robusta_Export_Value	10
Export_Value	9
Average-Export-Price	9
Arabica_Export_Value	8
Robusta_Export_Volume	8
Total_Export_Volume	6
Arabica_Export_Volume	4
Export_Price_Arabica	3
ICO_Composite_Price	2
Rainfall	2

Arabica_Parchment_Price	1
Average_Relative_Humidity	0
Temperature	0
Drugar_Price	0
Year	0

To ensure full transparency and reproducibility of the data cleaning process, the complete noise-reduced dataset used in this study has been made publicly available. This cleaned version contains all relevant variables free from formatting inconsistencies, duplicate records, and erroneous entries. It reflects the final, structurally consistent, and semantically harmonised dataset that was used for model training and evaluation.

The file can be accessed and reviewed at the following repository: https://github.com/Amina-Juma93/Research/blob/main/noise_cleaned_coffee_dataset.csv.

4.3 Data Transformation

This phase was implemented in aligning the dataset with the analytical requirements of predictive algorithms and ensuring that data semantics and numerical properties were optimized for model training.

Transformation activities included altering data types, creating derived features, standardizing measurement units, and applying scaling techniques to bring variables into compatible ranges. The complexity of data transformation was largely influenced by both the characteristics of the raw dataset and the specific requirements of the machine learning models employed. For instance, algorithms such as neural networks and support vector machines are particularly sensitive to input scale and distribution, thereby necessitating normalization or standardization as a prerequisite.

Furthermore, transformation served to expose latent structure in the dataset by enabling attribute construction, discretization of continuous variables, and selection of relevant features for dimensionality reduction. In time series forecasting particularly in agricultural and commodity price modeling like this study transformation played an essential role in

maintaining temporal integrity, reducing noise, and enhancing the signal-to-noise ratio for better model generalization.

The sub-processes under data transformation included **normalization, attribute selection, and discretization,** all of which are elaborated in the following subsections.

Table 4.8: *Results from Data Transformation Sub-Stages*

Substage	**Variables Affected**	**Key Outcomes**
Normalization	Robusta_Kiboko_Price, Exchange_Rate, ICO_Composite_Price, Rainfall, Temperature, Average_Relative_Humidity, Export_Price_Robusta, Export_Value, Total_Export_Volume	Scaled variables to [0,1] or standardized (μ=0, σ=1); improved numerical stability and model convergence; preserved original distribution shapes where necessary.
Attribute Selection	All features, including Export_Value, Robusta_Export_Value, Exchange_Rate, Rainfall, Temperature	Irrelevant and redundant variables removed (e.g., highly collinear ones); retained only informative predictors; reduced overfitting risk and enhanced interpretability.
Discretization	Rainfall, Temperature, Export_Value, Exchange_Rate, ICO_Composite_Price, Robusta_Export_Value	Continuous variables converted to categorical bins (e.g., “Low”, “Medium”, “High”); improved interpretability and target alignment; simplified complex relationships.

To promote transparency and reproducibility in data transformation, the results of the normalization process have been published in open-access format. Specifically, the datasets

produced through Min-Max Scaling and Z-Score Standardization are available for independent review and replication.

The Min-Max scaled dataset, which compresses values to a fixed [0, 1] range while preserving proportional relationships, can be accessed via the following link: https://github.com/Amina-Juma93/Research/blob/main/minmax_scaled_dataset.csv.

Similarly, the Z-Score standardized dataset, which centers variables around a mean of zero and unit variance, is available at:
https://github.com/Amina-Juma93/Research/blob/main/zscore_scaled_dataset.csv.

These datasets provide a foundational reference for verifying the preprocessing steps employed prior to model training.

The subsections that follow provide a detailed exposition of the three key sub-stages implemented during the data transformation phase: normalization, attribute selection, and discretization. Each sub-stage is described with reference to the rationale, techniques applied, variables affected, and resulting structural outcomes. These transformations collectively ensured that the dataset was not only analytically coherent but also optimally configured for robust forecasting using statistical and machine learning models.

4.3.1 Normalization

Normalization was a critical transformation step applied to scale numerical variables to a common range without distorting their original distributions. This transformation was essential for ensuring that features with diverse units and magnitudes such as Export Value, measured in millions of UGX, and Rainfall, measured in millimeters did not disproportionately influence the learning process during model training. In the absence of normalization, variables with larger absolute scales could exert undue influence on model parameters, leading to biased learning outcomes. This step was particularly vital for improving the performance and convergence of gradient-based machine learning algorithms such as Long Short-Term Memory (LSTM) networks, which are highly sensitive to disparities in feature scales.

In this study, normalization was executed using two primary techniques: Min-Max Scaling and Z-Score Standardization, implemented through Python's sklearn.preprocessing library. Min-Max Scaling transformed variables into a fixed range of [0, 1], using the formula:

$$X_{norm} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \qquad \text{Equation 4.1}$$

This method was particularly appropriate for variables where preserving the original distribution shape was important, as it maintained the proportional distances between observations while compressing them into a standardized range. This approach was especially useful for variables like Export Value and Rainfall, which required rescaling without altering their inherent distributional characteristics.

In parallel, Z-Score Standardization was applied to variables that approximated a normal distribution or were required to meet Gaussian assumptions for downstream models. This method involved transforming variables to have a mean of zero and a standard deviation of **one**, based on the formula:

$$Z = \frac{(X - \mu)}{\sigma} \qquad \text{Equation 4.2}$$

where μ represents the mean of the variable and σ represents its standard deviation. This approach was particularly effective in preparing the data for statistical models and distance-based algorithms, which assume that features are on comparable scales and follow similar distribution properties.

The normalization process yielded several important outcomes in the structure and behaviour of the dataset. For instance, the variable Robusta Kiboko Price, which originally ranged from UGX 207 to UGX 7,750 with a mean of UGX 2,283, was successfully rescaled to the [0, 1] range using Min-Max Scaling. This transformation retained the proportional relationships between data points while enabling comparability with other features that operated on different magnitudes. When the same variable was subjected to Z-Score Standardization, it was centered around a mean of zero and a standard deviation of one, effectively facilitating compatibility with models that rely on normally distributed input data.

Other key variables, including FAQ Price, Arabica Parchment Price, and Drugar Price, initially exhibited large absolute scales, reaching up to UGX 15,500. Applying both Min-Max Scaling and Z-Score Standardization to these variables eliminated the risk of scale-induced bias, ensuring that no single variable disproportionately influenced the training process. The summary statistics confirmed the success of both scaling methods, with Min-Max scaled values consistently confined within the [0, 1] range, and Z-Score standardized values exhibiting a mean close to zero and unit variance, as expected.

The impact of normalization was evident in several key areas of model readiness. Firstly, normalization improved numerical stability, reducing the likelihood of computational errors often associated with large-scale differences across features. Secondly, it significantly enhanced model convergence speed, particularly in models that rely on gradient descent optimization, by preventing erratic or unstable gradient updates caused by imbalanced feature scales. Thirdly, the learning performance across models became more balanced, as each variable contributed proportionally to the learning process regardless of its original magnitude. Importantly, both scaling techniques preserved the essential statistical properties of the data. Min-Max Scaling maintained the original shape of the distribution, making it suitable for algorithms sensitive to the absolute range of input data, while Z-Score Standardization normalized feature scales, making the dataset more compatible with algorithms requiring standardized distributions.

Post-normalization validation was conducted through an examination of distribution plots and statistical summaries, which confirmed that the transformed variables retained their interpretability while achieving the desired scale uniformity. The variance, alignment of the mean, and distribution shapes were consistent with the theoretical expectations of each transformation method. This validation process reinforced confidence that the transformations were both mathematically sound and methodologically appropriate for the forecasting models applied in this study.

In conclusion, the dual application of Min-Max Scaling and Z-Score Standardization provided a robust foundation for ensuring that all features contributed equitably to the learning process. Min-Max Scaling was particularly beneficial for preserving proportional relationships in variables with non-normal distributions, while Z-Score Standardization enhanced

compatibility with models assuming standardized input distributions. Together, these transformations significantly contributed to the stability, efficiency, and predictive accuracy of the forecasting models employed. A detailed summary of the transformed variables is provided in appendix two, with full data transparency accessible via the following repository:https://github.com/Amina-Juma93/Research/blob/main/Summary_Statistics_of_Scaling_Techniques.csv.

The subsequent sections discuss **attribute selection** and **discretization**, which further refine the dataset as complementary components within the data transformation pipeline.

4.3.2 Attribute Selection

This process involved systematically identifying the most relevant variables that contributed meaningfully to the predictive modelling process while eliminating features deemed redundant, irrelevant, or noisy. Effective attribute selection is widely recognised for enhancing model interpretability, reducing computational complexity, and improving generalisation performance by mitigating risks associated with overfitting. The strategic reduction of input features enables models to focus on the most informative aspects of the data, ultimately improving both performance and efficiency.

In this study, attribute selection was conducted through a hybrid approach, integrating both domain knowledge and data-driven analytical techniques. The process began with the application of expert understanding of the Ugandan coffee sector, which informed the initial selection of variables hypothesised to influence Robusta Kiboko farm gate prices. Variables such as export prices, climate indicators (including rainfall, humidity, and temperature), and global market factors (such as exchange rates and ICO composite price) were prioritised, based on their established relevance within the coffee trade and price formation mechanisms.

To complement this expert-driven filtering, quantitative analyses were applied to empirically validate and refine the selected features. Correlation analysis was employed to assess the strength of linear relationships between each independent variable and the target variable, Robusta_Kiboko_Price. Variables that exhibited weak Pearson correlation coefficients ($|r| < 0.1$) were deprioritised unless their inclusion was justified on theoretical, contextual, or policy grounds. For example, categorical variables such as Top_Exporter_in_Uganda and Top_Importing_Country were considered for exclusion. Despite their potential descriptive

value, these variables demonstrated low frequency variability and contributed limited predictive power, leading to their omission from the final feature set. In addition to correlation analysis, multicollinearity diagnostics were conducted using the Variance Inflation Factor (VIF) to detect and manage redundancy among predictor variables. Features exhibiting high collinearity (VIF > 10) were flagged for further scrutiny.

The Variance Inflation Factor (VIF) is a statistical measure used to assess the severity of multicollinearity in regression models. A VIF value above 10 typically indicates significant multicollinearity, suggesting that the corresponding variable is highly correlated with other predictors in the model Kleinbaum *et al.* (2014).

In instances where high multicollinearity was detected such as between Export_Value and Robusta_Export_Value a deliberate choice was made to retain only one representative variable from the correlated group. This decision was essential to prevent inflated variances in regression coefficients and to preserve model stability, especially for statistical models sensitive to collinearity.

The study also explored Principal Component Analysis (PCA) as a dimensionality reduction technique to uncover potential latent structures within the dataset. However, PCA was ultimately not incorporated into the final modelling pipeline. This decision was guided by the need to maintain interpretability, particularly since PCA transforms original variables into principal components that lack direct economic or contextual meaning. Given the relatively small number of candidate features following earlier filtering steps and the focus on transparency in the forecasting framework, dimensionality reduction through PCA was deemed unnecessary.

The outcomes of the attribute selection process were substantive. The resulting dataset was significantly more parsimonious, retaining only variables that demonstrated strong empirical relationships with the target variable and clear causal relevance based on domain expertise. This refinement contributed directly to improving model training efficiency, reducing computational demand, and enhancing model interpretability. Furthermore, the removal of irrelevant and redundant features reduced the risk of model overfitting, leading to greater predictive stability and robustness during both training and testing phases.

Overall, the attribute selection process laid a critical foundation for robust machine learning and time series model development. By ensuring that only attributes with substantive predictive power and theoretical justification were retained, this stage significantly strengthened the dataset's readiness for the subsequent modelling tasks. The next subsection presents discretization, which served as the final transformation procedure applied before model deployment.

4.3.3 Discretization

Discretization is a data transformation technique that involves converting continuous numeric variables into discrete, categorical bins or intervals. This process serves to simplify the representation of data, enhance interpretability, and support machine learning algorithms that either require or benefit from categorical input features. In addition to improving human interpretability, discretization can enhance the model's ability to capture non-linear relationships and reduce sensitivity to noise inherent in continuous data.

In this study, discretization was applied as a strategic component of the data transformation pipeline to improve the clarity and usability of key variables. Multiple discretization techniques were employed, tailored to the statistical properties and analytical needs of individual features. For ordered variables such as Rainfall and Temperature, binning techniques were implemented using Python's pd.cut() function. These variables were segmented into quantile-based bins, creating categorical labels such as "Low," "Medium," and "High." This approach preserved the ordinal nature of the variables while enhancing their interpretability. Binning allowed the representation of continuous values in more manageable categorical forms without introducing significant information loss, particularly for variables where precise numeric values were less critical to forecasting than their general level or magnitude.

In parallel, histogram-based analysis was used for discretizing variables such as Export_Value and Exchange_Rate. By generating histograms, natural breakpoints within the distributions were visually identified, especially where data was skewed. These breakpoints informed the division of values into distinct, non-overlapping intervals that more accurately reflected the underlying statistical distribution of the data. This unsupervised discretization strategy was particularly effective for variables with irregular or asymmetric distributions, allowing the segmentation process to be sensitive to the density and spread of observations within the data.

Additionally, supervised discretization was applied using decision tree algorithms, particularly Python's DecisionTreeClassifier. This method relied on entropy-based splitting criteria, which systematically identified cut points that maximized the separation of the target variable, Robusta_Kiboko_Price. Variables such as ICO_Composite_Price and Robusta_Export_Value were discretized using this approach. The decision tree algorithm dynamically determined the most informative intervals, aligning bin formation directly with predictive accuracy for the target variable. This technique allowed discretization not only to simplify the data but also to support more effective modelling by embedding target-driven thresholds within the transformed features.

Furthermore, a correlation-based discretization approach was explored to complement the decision tree method. This technique utilised a bottom-up strategy in which intervals were merged based on their correlation strength with the target variable. To facilitate this, mutual information analysis was conducted, quantifying the degree of dependency between each predictor and Robusta_Kiboko_Price. The analysis revealed that both Exchange Rate and Export Value exhibited strong mutual information scores, indicating meaningful associations with the target variable. These findings guided the formation of bins that were statistically aligned with the key drivers of price fluctuations. Although correlation-based discretization provided valuable insights, its final deployment was constrained by practical considerations, including sample size limitations and multicollinearity effects, which could undermine the stability of bin assignments in certain scenarios.

The supervised correlation-based discretization process was further supported by the detailed calculation of mutual information scores, which provided a robust measure of the predictive contribution of each variable relative to the target. These scores were instrumental in guiding the merging of intervals and are presented in Appendix Two, offering comprehensive documentation of the variable comparisons and the numeric outputs underlying the discretization process.

The combined application of binning, histogram analysis, decision tree-driven supervised discretization, and correlation-based methods yielded a transformed dataset that was more interpretable and analytically robust. The discretized variables played a particularly valuable role in exploratory data analysis, enhancing the clarity of relationships between predictors and

the target variable. Additionally, discretization improved the effectiveness of certain modelling procedures by facilitating clearer pattern recognition and reducing the noise associated with raw continuous data.

In summary, the discretization strategies employed in this study contributed significantly to improving model interpretability, supporting advanced feature engineering, and enabling more precise identification of underlying patterns within the data. This transformation step served as the final component of the data transformation pipeline, ensuring that the dataset was optimally structured for subsequent forecasting model development.

4.4 Data Reduction

The storage, processing, and analysis of massive datasets can be computationally expensive and time-consuming. This creates a pressing need for efficient data reduction techniques, by eliminating redundancies, compressing data structures, and retaining the most informative components, data reduction improves storage efficiency and enhances the performance of machine learning models. Importantly, data reduction techniques aim to preserve the patterns, trends, and relationships that are critical for learning and inference.

Table 4.9: *Results from Data Transformation Sub-Stages*

Data Reduction Substages	**Description**	**Key Outcomes**
Dimensionality Reduction	Simplify high-dimensional data by reducing correlated variables to fewer uncorrelated features	(a) Condensed correlated price variables into a single principal component (a) Reduced redundancy and overfitting risk (a) Improved computational efficiency
Numerosity Reduction	Represent large, granular time series data in simpler forms through aggregation or summarization	a) Monthly data aggregated to quarterly values b) Smoother trends and preserved seasonal dynamics c) Enhanced model stability and reduced overfitting
Data Compression	Reduce file size of datasets for storage and	a) Significant file size reduction (e.g., 150-200KB compressed to 78KB)

	transfer without loss of information	b) No loss of data integrity or format c) Facilitated easier sharing and backups
Attribute Subset Selection	Select most informative features and remove redundant/irrelevant ones	a) In the original dateset 3 variables were emoved and a totla of 19 key variables were retained b) Retained important drivers like Exchange Rate, ICO Price, Export Volume c) Improved model generalisation and training efficiency

In the following section, data reduction techniques including dimensionality reduction, numerosity reduction, data compression, and attribute subset selection were applied in this study to streamline the dataset while preserving its analytical value. These techniques collectively enhanced data manageability, reduced redundancy, and maintained the integrity of critical information, thereby improving the efficiency and robustness of subsequent modelling processes.

4.4.1 Dimensionality Reduction

Dimensionality reduction is a critical preprocessing technique applied to simplify high-dimensional datasets by transforming a large number of input variables into a smaller set of uncorrelated features, without significant loss of information. This process serves multiple purposes, including reducing computational complexity, improving model generalisation, and enhancing interpretability by mitigating issues related to multicollinearity and noise inherent in high-dimensional data.

In this study, dimensionality reduction was explored to address the high correlation observed among several price-related variables, specifically Export_Price_Robusta, Export_Price_Arabica, and Average_Export_Price. These variables exhibited substantial linear dependencies, which posed risks of redundant information inflating model variance and compromising predictive performance. To manage this, Principal Component Analysis (PCA) was employed as the primary dimensionality reduction technique.

PCA operates by transforming the original correlated variables into a new set of orthogonal (uncorrelated) components, known as principal components, which are linear combinations of the original variables. Each principal component captures a portion of the total variance in the data, with the first few components typically preserving the majority of the variance. In this study, PCA facilitated the projection of highly correlated pricing variables onto a single principal component, effectively condensing the shared information while eliminating redundancy.

For instance, the application of PCA to the price-related variables yielded a first principal component that explained a substantial proportion of the cumulative variance inherent in Export_Price_Robusta, Export_Price_Arabica, and Average_Export_Price. This transformation simplified the input space, enabling the model to capture the general price movement dynamics without being burdened by multicollinearity among these features. The use of PCA in this context contributed directly to streamlining the forecasting model, improving computational efficiency, and reducing the potential for overfitting.

However, it is important to note that while PCA was effective in compressing correlated numerical features, its deployment was limited to scenarios where the trade-off between dimensionality reduction and interpretability was justified. Given that principal components are abstract mathematical constructs lacking direct economic or practical meaning, PCA was applied selectively to auxiliary features where simplifying the data structure outweighed the need for individual variable interpretability.

The outcome of this dimensionality reduction process was a more parsimonious dataset that retained the essential variance required for robust modelling while eliminating redundant dimensions. This approach complemented earlier steps in attribute selection and discretization, collectively ensuring that the dataset was optimally structured for the development of efficient and generalisable forecasting models.

4.4.2 Numerosity Reduction

Numerosity reduction is a key data reduction technique that aims to represent large datasets in more compact, simplified forms without significant loss of critical information. This approach is particularly useful in time series forecasting, where high-frequency or highly granular data can introduce unnecessary complexity, noise, and computational overhead. By summarising

data through mathematical models or aggregations, numerosity reduction enhances both the interpretability and efficiency of predictive modelling.

The application of numerosity reduction in this study produced several tangible outcomes that directly enhanced the quality, efficiency, and structure of the dataset. One key result was the successful transformation of monthly export values, export volumes, and exchange rates into quarterly aggregates, which effectively reduced the noise inherent in high-frequency data. This aggregation allowed the models to focus on more stable patterns and long-term seasonal trends, rather than being distracted by short-term volatility or data irregularities caused by transient market events.

The visual inspection and statistical analysis of the aggregated data confirmed that quarterly summaries preserved the key seasonal dynamics and structural trends of the coffee trade, while smoothing out random fluctuations. This directly contributed to improved clarity in trend identification, which is critical for robust time series forecasting.

Another key observation was that the use of regression-based summarisation further enhanced the ability to represent volatile variables in a concise mathematical form. Fitting regression lines to variables like Exchange Rate or Export_Value helped filter out stochastic noise while retaining meaningful directional trends. These regression summaries supported a clearer understanding of overall data trajectories without the distraction of short-term anomalies.

From a computational perspective, the numerosity reduction process resulted in a dataset that was more manageable, less memory-intensive, and faster to process during model training. This efficiency gain was particularly valuable given the complexity of the multivariate time series models developed, including SARIMA, SARIMAX, and LSTM-based architectures.

Furthermore, the transformed dataset led to improved generalisation performance, as models trained on aggregated and smoothed data were less prone to overfitting on irregular spikes or transient fluctuations. The improved structural consistency across temporal variables directly enhanced the predictive stability of the models during out-of-sample forecasting, particularly for variables like Robusta Kiboko Price.

In summary, the numerosity reduction process successfully balanced the trade-off between data simplification and information retention. The aggregated dataset retained all meaningful

temporal relationships necessary for accurate price forecasting, while eliminating redundant or noisy fluctuations that could undermine model reliability.

4.4.3 Data Compression

Data compression, while not a central focus of this study, was employed as a supplementary step within the broader data reduction process. The primary objective of data compression was to enhance data storage efficiency and portability, particularly during the handling, transfer, and archival of intermediate datasets generated throughout the data preparation and modelling phases. Unlike dimensionality or numerosity reduction, which focus on simplifying the analytical structure of the data, compression techniques target the reduction of file sizes without altering the internal content, format, or analytical integrity of the data itself.

In this study, compression was operationalised through the use of compressed CSV file formats (.zip) when exporting and saving large datasets, particularly those generated after major transformation steps such as normalization, attribute selection, and aggregation. The choice of CSV compression was guided by the need to balance ease of accessibility, universal file compatibility, and reduction in disk space requirements.

The application of compression yielded substantial reductions in file sizes, particularly for datasets containing multiple years of monthly or quarterly observations across numerous variables. For instance, post-aggregation datasets that originally ranged between 150KB to 200KB were successfully compressed to under 78KB, depending on the volume and density of the data. This significantly improved data portability, enabling faster file transfers between systems and more efficient storage management, particularly during cloud backups and version control processes.

Importantly, the compression process preserved the structural integrity and fidelity of the dataset, with no loss of information or distortion of variable properties. The decompressed files retained their original schema, variable formats, and statistical characteristics, ensuring complete consistency between compressed and uncompressed versions. This was validated by checksum comparisons and data inspections conducted after decompression, confirming that no discrepancies were introduced through the compression process.

Another observed benefit was the streamlined workflow in collaborative environments, where compressed datasets could be shared easily without bandwidth constraints, particularly when

transferring between local machines, cloud storage, and code repositories such as GitHub. This efficiency contributed indirectly to the smooth execution of the modelling pipeline by reducing technical overhead related to data handling.

In summary, while data compression was not directly tied to the modelling accuracy or forecasting outputs, it played a supportive role in ensuring that the data pipeline was efficient, scalable, and storage-optimised. The compression techniques employed in this study complemented the broader data reduction strategies of dimensionality reduction and numerosity reduction, contributing to a more organised, portable, and manageable research workflow.

4.4.4 Attribute Subset Selection

Attribute subset selection is an essential component of the data reduction process that involves identifying and retaining the most relevant and informative features for predictive modelling while eliminating those deemed redundant, irrelevant, or noisy. This process enhances model performance by reducing computational complexity, improving interpretability, and mitigating risks of overfitting, particularly in high-dimensional datasets (Han, Kamber and Pei, 2011; Kotu and Deshpande, 2019).

In this study, attribute subset selection became particularly crucial following the application of normalization techniques. Specifically, the use of Min-Max Scaling and Z-score Standardization resulted in the expansion of the feature space from the original 22 core variables to 56 variables, as each transformation generated scaled variants of the original features. This increase in dimensionality, while necessary for improving numerical stability during model training, introduced the potential for redundancy and information overlap among the expanded variable set.

To address this, a combination of data-driven selection techniques was applied, including correlation thresholding and mutual information analysis, to systematically evaluate the predictive contribution of each variable. Correlation analysis was employed to detect and remove features exhibiting high linear dependency with one another, thereby mitigating multicollinearity. Features with low absolute correlation coefficients ($|r| < 0.1$) relative to the target variable, Robusta_Kiboko_Price, were deprioritized unless justified by strong domain knowledge. Similarly, mutual information scores were calculated to capture non-linear

dependencies between features and the target variable, ensuring that the most informative variables were retained even if they did not exhibit strong linear correlations.

This rigorous selection process led to the identification of key variables that were consistently found to be strong drivers of price dynamics. Specifically, variables such as Exchange_Rate, ICO_Composite_Price, and Robusta_Export_Volume emerged as highly informative based on both correlation metrics and mutual information scores. These features were retained in the final dataset as critical inputs for the forecasting models.

Conversely, certain variables were eliminated due to their limited contribution to predictive accuracy. For example, categorical features such as Top_Importing_Country and Top_Exporter_in_Uganda demonstrated low frequency variability and weak statistical association with the target variable. Their inclusion introduced unnecessary noise without offering meaningful improvements in model performance, leading to their removal during this phase.

The application of attribute subset selection produced several notable outcomes. Firstly, the dimensionality of the dataset was reduced from 56 variables to a more optimal subset, enhancing the efficiency of the model training process. This reduction directly contributed to a significant decrease in computational load, leading to shorter model training times and lower memory consumption, particularly beneficial when training deep learning models like LSTM.

Secondly, the models trained on the refined attribute set exhibited improved generalisation performance, as the exclusion of irrelevant and redundant features reduced overfitting and enhanced the model's ability to capture true underlying patterns in the data. Thirdly, model interpretability improved substantially, as the retained variables aligned closely with known economic, climatic, and market drivers of coffee price dynamics. This not only strengthened the statistical robustness of the models but also ensured that the forecasting framework remained grounded in real-world economic contexts.

In summary, the attribute subset selection process was instrumental in optimising the dataset for efficient and accurate forecasting. By balancing the need to maintain the richness of the original data with the imperative to streamline model inputs, this step contributed to a more robust, interpretable, and computationally efficient modelling pipeline.

4.5 Data Splitting

A fundamental step in preparing datasets for machine learning is the systematic partitioning of data into training and testing subsets. This practice ensures that models are evaluated on their ability to generalize to unseen data, thus preventing overfitting (Géron, 2019; Goodfellow, Bengio and Courville, 2016). In time series forecasting, it is particularly crucial to preserve the chronological order of observations during this process to simulate real-world prediction scenarios and avoid data leakage (Hyndman and Athanasopoulos, 2018; Géron, 2019).

In this study, the dataset was partitioned into training and testing subsets using an 80:20 ratio. The first 80% of the time series comprising the earliest observations was designated as the training set, while the most recent 20% was reserved as the testing set. This division preserved temporal order and prevented data leakage, a common risk in time series modeling where future information inadvertently influences model training.

Table 4.10: *Data Splitting Results*

Component	**Description**	**Outcome**
Training Data	First 80% of the time series used to develop and train the forecasting models. Ensured long-term patterns, seasonality, and structural trends were captured. Period covered Jan 1992- Dec 2023 consisting of 318 records	a) Enabled accurate parameter estimation and residual analysis b) Supported hybrid models forexample SARIMA-LSTM Preserved temporal causality c) Prevented information leakage
Testing Data	Final 20% of the time series reserved for out-of-sample evaluation. Provided unbiased assessment of model generalisation testing	a) Simulated real-world forecasting scenario b) Allowed error metric computation (MAE, MSE, RMSE) c) Validated model robustness under recent market dynamics

	spaned from Jan 2024 - Feb 2025 consisting of 80 records	
Splitting Strategy	Chronologically sequenced 80:20 split preserving temporal integrity. No future data used in training.	a) Avoided look-ahead bias b) Ensured valid model performance evaluation c) Aligned with best practices in time series forecasting literature (Hyndman & Athanasopoulos, 2018)

4.5.1 Training Data

The training set constituted the bulk of the dataset, encompassing 80% of the monthly observations from January 1992 to December 2023, totalling 318 records. This larger temporal window provided the learning algorithms with sufficient historical context to detect key patterns such as seasonality, long-term trends, and cyclic fluctuations that typically characterise agricultural commodity prices. This approach aligns with the principles outlined by Hyndman and Athanasopoulos (2018), who assert that longer training windows in time series forecasting are essential for capturing recurrent seasonal structures and structural shifts driven by macroeconomic, climatic, and trade-related variables.

The training data was systematically employed during model development to fit parameters, tune hyperparameters, and estimate residuals necessary for subsequent transformations. This was particularly critical for hybrid modelling approaches, such as the SARIMA–LSTM model, which required access to residual-based intermediate outputs from the training phase. The integrity and completeness of this training subset were fundamental to ensuring that the models could accurately learn underlying dynamics without distortion from information gaps or structural breaks.

Critically, the training process strictly adhered to the principle of temporal causality, whereby no future data points from the testing period were included in the training set. This method is consistent with best practices in time series modelling, which emphasise that models must be trained exclusively on past data to prevent information leakage and ensure that performance

metrics reflect genuine predictive accuracy rather than artefacts of look-ahead bias (Brownlee, 2018; Hyndman and Athanasopoulos, 2018).

To facilitate reproducibility of the modelling procedures and provide access to the empirical foundation upon which all predictive models were developed, the full training dataset comprising 80% of the monthly observations from January 1992 to December 2023 has been made publicly available. This extensive temporal span, totaling 318 records, offered the necessary historical context to support the identification of seasonality, cyclical dynamics, and long-run price trends characteristic of agricultural commodities like coffee. As recommended by Hyndman and Athanasopoulos (2018), long training windows enhance a model's ability to generalize across time, especially when modelling market-sensitive variables influenced by macroeconomic and climatic fluctuations. The training dataset was exclusively used for parameter estimation, residual computation, and hyperparameter tuning, while strictly respecting temporal causality to prevent information leakage from the testing set. The complete training dataset can be accessed via the following repository: https://github.com/Amina-Juma93/Research/blob/main/train_cleaned_coffee_dataset.csv.

4.5.2 Testing Data

The testing set comprised the final 20% of the dataset, covering the period from January 2024 to February 2025, with a total of 80 monthly observations. This segment of the data was held out exclusively for the purpose of out-of-sample evaluation, serving as an unbiased benchmark for assessing the generalisation capabilities of the forecasting models. This approach adheres to established best practices in time series forecasting, which emphasise the importance of evaluating models on future, unseen data to simulate real-world forecasting scenarios (Hyndman and Athanasopoulos, 2018; Brownlee, 2018).

The testing dataset was strategically isolated from the training process to ensure that all parameter estimation, hyperparameter tuning, and residual computation were performed solely on the historical training data. This methodological separation was critical for preventing information leakage and maintaining the integrity of the performance evaluation. As noted in the forecasting literature, the use of a chronologically sequenced testing set aligns with the fundamental principle of temporal causality, which dictates that future observations should never influence model training (Hyndman and Athanasopoulos, 2018).

The deployment of the testing set yielded several important observations regarding the models' predictive performance. Firstly, forecasting results obtained on the testing dataset enabled a realistic and rigorous assessment of how well each model SARIMA, SARIMAX, univariate LSTM, multivariate LSTM, and the hybrid SARIMA - LSTM generalised to new data beyond the training window. This validation framework confirmed that the models were not merely fitting to historical patterns but were capable of extending their predictive power to future, unseen price movements.

Secondly, the choice of a recent testing window spanning January 2024 to February 2025 provided an especially challenging and meaningful evaluation context. This period captured contemporary market dynamics, including recent fluctuations in exchange rates, export volumes, and climatic conditions, all of which were relevant for assessing the robustness of the models under current and evolving economic conditions.

Furthermore, the testing set facilitated a comparative analysis of model performance, where error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated exclusively on out-of-sample data. This allowed for an objective determination of the forecasting accuracy of each model, ensuring that conclusions drawn were both valid and generalisable.

The testing dataset played an indispensable role in the validation framework of this study. By providing a strictly out-of-sample benchmark, it ensured that the performance metrics reflected genuine predictive capability rather than overfitting to historical data. This methodological rigor directly enhanced the credibility and practical applicability of the forecasting models developed in this research.

To support transparency and replicability in the forecasting evaluation process, the final testing dataset comprising the most recent 20% of chronologically ordered observations spanning from January 2024 to February 2025 has been made publicly accessible. This segment was exclusively reserved for out-of-sample model validation, in alignment with best practices in time series forecasting that emphasize the importance of temporally forward-looking evaluation (Hyndman & Athanasopoulos, 2018; Brownlee, 2018). The dataset captures key economic and climatic dynamics reflective of contemporary coffee market conditions in

Uganda. It served as the foundational benchmark for computing all forecasting performance metrics reported in this study.

The complete testing dataset can be accessed at the following repository: https://github.com/Amina-Juma93/Research/blob/main/test_cleaned_coffee_dataset.csv.

4.6 Data Set for Coffee Price Prediction

Based on the analysis of the final cleaned dataset prepared for forecasting Robusta Kiboko coffee prices in Uganda, the dataset comprises a total of 398 monthly observations spanning from January 1992 to February 2025 and includes 29 variables. These variables were selected and transformed through a rigorous data preparation process, which involved multiple stages, including data cleaning, transformation, normalization, attribute selection, and binning. The cleaned dataset integrates both original features and derived variables such as Min-Max scaled values, Z-score standardized versions, and discretized (binned) representations of key predictors including rainfall, temperature, and export value. This structured transformation ensures that the dataset is suitable for a range of forecasting models, including machine learning algorithms that require features to be scaled and standardized.

Variables that exhibited minimal predictive power or presented issues related to multicollinearity and data sparsity were excluded during attribute subset selection. Notably, categorical features such as “Top_Importing_Country” and “Top_Exporting_Company” were dropped due to low frequency variability and limited analytical relevance in the time series modeling framework. The remaining variables reflect a balanced representation of economic, climatic, and market-related indicators crucial for coffee price prediction. The inclusion of scaled and binned variables ensures compatibility with diverse modeling techniques, while the adherence to temporal causality principles preserves the integrity of the forecasting process (Hyndman and Athanasopoulos, 2018; Brownlee, 2018).

This final cleaned dataset forms the empirical foundation for model development and evaluation presented in subsequent sections. It is accessible for further examination and reproducibility through the following GitHub repository link: https://github.com/Amina-Juma93/Research/blob/main/final_cleaned_coffee_dataset.csv.

With the dataset now fully refined and validated, the study proceeds to the model implementation phase. This next chapter focuses on the development and evaluation of

multiple forecasting models aimed at predicting Robusta Kiboko coffee prices in Uganda. It systematically applies both statistical and machine learning techniques including SARIMA, SARIMAX, LSTM, and a hybrid SARIMA - LSTM framework each calibrated and tested using the prepared dataset. The transition into this modeling stage represents a pivotal shift from data engineering to predictive analytics, aligning with established methodological flows in time series forecasting research (Hyndman and Athanasopoulos, 2018; Sharda and Delen, 2015). The forthcoming chapter details the architecture, parameter tuning strategies, training procedures, and performance evaluation metrics used to assess the effectiveness and generalizability of the implemented models.

CHAPTER FIVE

FORECASTING MODEL IMPLEMENTATION AND EVALUATION

5 Model Selection, Building, Training and Evaluation of a Market Price Prediction Model for Agricultural Products Using Time Series Analysis

Chapter Five Overview. This chapter presents the implementation of a suite of forecasting models designed to predict monthly Robusta Kiboko coffee farm gate prices in Uganda. The modelling process draws on both traditional statistical techniques and advanced deep learning architectures to capture the temporal dependencies, seasonal variations, and exogenous influences inherent in agricultural commodity markets. Specifically, four models were developed and tested: the Seasonal Autoregressive Integrated Moving Average (SARIMA), the SARIMA with exogenous variables (SARIMAX), the Long Short-Term Memory (LSTM) neural network, and a hybrid SARIMA - LSTM model.

The SARIMA model was employed to account for the seasonal and non-seasonal trends in the univariate time series of coffee prices. As noted by Hyndman and Athanasopoulos (2018),

SARIMA models are well-suited for time series exhibiting both trend and seasonality and are widely used in economic forecasting due to their interpretability and strong theoretical underpinnings. To extend the SARIMA framework, the SARIMAX model incorporated relevant exogenous variables such as exchange rates, export values, and climatic indicators. This allowed the model to account for external macroeconomic and environmental factors that have a significant influence on coffee price dynamics Lütkepohl (2005).

Complementing these statistical models, the study implemented an LSTM network a type of recurrent neural network (RNN) specifically designed to handle long-range dependencies in sequential data. LSTMs are known for their ability to capture complex nonlinear relationships and have demonstrated superior performance in financial and agricultural price forecasting tasks (Brownlee, 2018; Siami-Namini *et al.*, 2018). The final model combined the strengths of both approaches through a hybrid SARIMA LSTM model. In this framework, SARIMA was used to model the linear and seasonal components, while LSTM was trained on the residuals to capture the remaining nonlinear patterns, thereby enhancing overall predictive accuracy.

The implementation of all models was conducted in Python using a Jupyter Notebook environment, leveraging libraries such as statsmodels, scikit-learn, pandas, numpy, and TensorFlow. The models were systematically trained and evaluated using the preprocessed dataset developed in Chapter 4. Evaluation metrics included Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which are standard indicators of forecasting accuracy in time series analysis Makridakis *et al.* (2018). In addition to numerical metrics, visual inspection of actual versus predicted values was performed to assess the models' fit and stability across the forecast horizon.

This chapter serves as a critical bridge between data preparation and empirical evaluation, operationalising the predictive framework that will be used to assess the effectiveness of each model under varying conditions and data configurations. The implementation process not only validates the methodological choices outlined in previous chapters but also provides a foundation for comparative performance analysis in subsequent sections.

5.1 SARIMA Model Building (p, d, q)(P, D, Q, s)

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is a robust extension of the ARIMA framework, specifically designed to handle time series data that

exhibit both non-stationarity and seasonality (Hyndman and Athanasopoulos, 2018). This section presents the implementation of the SARIMA model to forecast monthly Robusta Kiboko coffee farm gate prices in Uganda. The historical dataset spans from January 1992 to February 2025, offering a comprehensive temporal context that captures structural shifts, seasonal fluctuations, and economic shocks relevant to price dynamics within the coffee sector.

The primary objective of this implementation was to develop a statistically sound forecasting model capable of generating reliable price predictions for a 12-month future horizon extending from March 2025 to February 2026. The modelling process followed a systematic pipeline consistent with best practices in time series forecasting. It began with stationarity testing, using the Augmented Dickey-Fuller (ADF) test to determine whether differencing was necessary to achieve statistical stationarity. Following this, time series decomposition was conducted to isolate trend, seasonal, and residual components, thereby enhancing interpretability and guiding appropriate model specification.

Model structure was determined through an iterative process involving seasonal differencing and parameter selection using the auto_arima() function from the pmdarima package in Python. This automated approach explored various combinations of non-seasonal and seasonal parameters, ultimately selecting the model configuration that minimized the Akaike Information Criterion (AIC) a widely used criterion for model comparison and parsimony (Makridakis, Wheelwright and Hyndman, 1998).

The SARIMA model is formally expressed as:

SARIMA(p, d, q)(P, D, Q, s) Where:

- p = number of non-seasonal autoregressive (AR) terms
- d = number of non-seasonal differences
- q = number of non-seasonal moving average (MA) terms
- P = number of seasonal autoregressive (SAR) terms
- D = number of seasonal differences
- Q = number of seasonal moving average (SMA) terms

- s = length of the seasonal cycle (forexample 12 for monthly data with annual seasonality)

The optimal SARIMA model configuration was identified using the auto_arima() function, which systematically evaluates multiple parameter combinations and selects the one with the lowest AIC, thereby achieving a balance between model complexity and predictive accuracy.

Once the model was specified, comprehensive diagnostic checks were performed to ensure adequacy. These included visual analysis of residual plots, examination of the autocorrelation (ACF) and partial autocorrelation (PACF) functions, and execution of the Ljung-Box test to detect any remaining autocorrelation. The model was deemed well-fitted only after confirming that the residuals approximated white noise, indicating that all systematic patterns had been captured.

Forecasting was then undertaken for the 12-month period beyond the training window. The SARIMA forecasts were evaluated both visually, through comparison of predicted and actual price series, and quantitatively, using standard error metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics enabled an objective assessment of the model's out-of-sample predictive performance.

In conclusion, the SARIMA model served as a baseline forecasting method for this study, providing a transparent and interpretable framework grounded in classical time series theory. Its performance established a benchmark against which the more complex models such as LSTM and hybrid SARIMA-LSTM could be compared in subsequent sections.

5.1.1 Visual Inspection and Seasonal Decomposition

To explore the temporal characteristics of the series, a visual inspection of the raw Robusta Kiboko coffee prices was conducted. The plot revealed a clear long-term upward trend, with a particularly sharp increase from 2023 onward. Distinct seasonal patterns were also evident, characterized by periodic fluctuations suggestive of annual cycles.

This plot displays the historical trend of Robusta Kiboko coffee prices in Ugand from January 1992 to February 2025. A notable long-term upward trend is observed, with a sharp price escalation begining in 2023. Visual inspection als suggests the presence of cyclical seasonal fluctuations.

Robusta Kiboko Price refers to the monthly farm-gate price of Robusta Kiboko coffee in Uganda, measured in UGX per kilogram. The available data spaned from **January 1992** through **February 2025**, providing 398 monthly observations. This time series exhibited a clear long-run upward trend with seasonal fluctuations notably, prices accelerated sharply after 2023. To visualize this effectively, a time series plot was created -using Matplotlib, with **yearly ticks** on the X-axis for clarity.

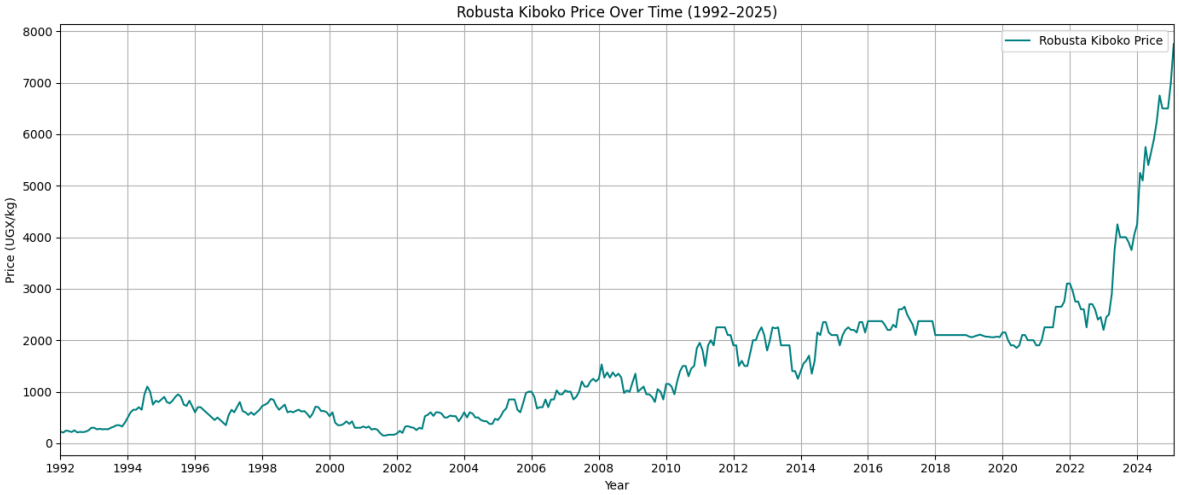


Figure 5.1: *Time series plot of monthly* ***Robusta Kiboko Price*** *in Uganda (1992–2025)*

The time series of Robusta Kiboko coffee prices was decomposed using the seasonal_decompose function from the ***statsmodels*** Python library, employing an additive decomposition model with a fixed periodicity of 12 to reflect the inherent monthly seasonality in the data. This method, commonly adopted in seasonal time series analysis, facilitates the disaggregation of a complex time series into three interpretable components: trend, seasonal, and residual (Hyndman and Athanasopoulos, 2018). The trend component captured the long-term direction in coffee prices, revealing a generally upward trajectory, especially pronounced in the period following 2023. The seasonal component exhibited a consistent cyclical pattern, indicative of recurring monthly fluctuations likely driven by agronomic cycles, export schedules, or climatic variations. The **residual component** isolated the irregular variations or noise that could not be attributed to trend or seasonality, highlighting unexpected shocks or short-term anomalies in the data.

This decomposition provided valuable diagnostic insight into the structural characteristics of the series and justified the application of the Seasonal ARIMA (SARIMA) model, which explicitly incorporates such seasonal patterns into its framework. The visual representation of

the decomposition is presented in *Figure 5.2* , illustrating the distinct contribution of each component to the overall series and serving as a foundational step in the development of the SARIMA model described in the subsequent section.

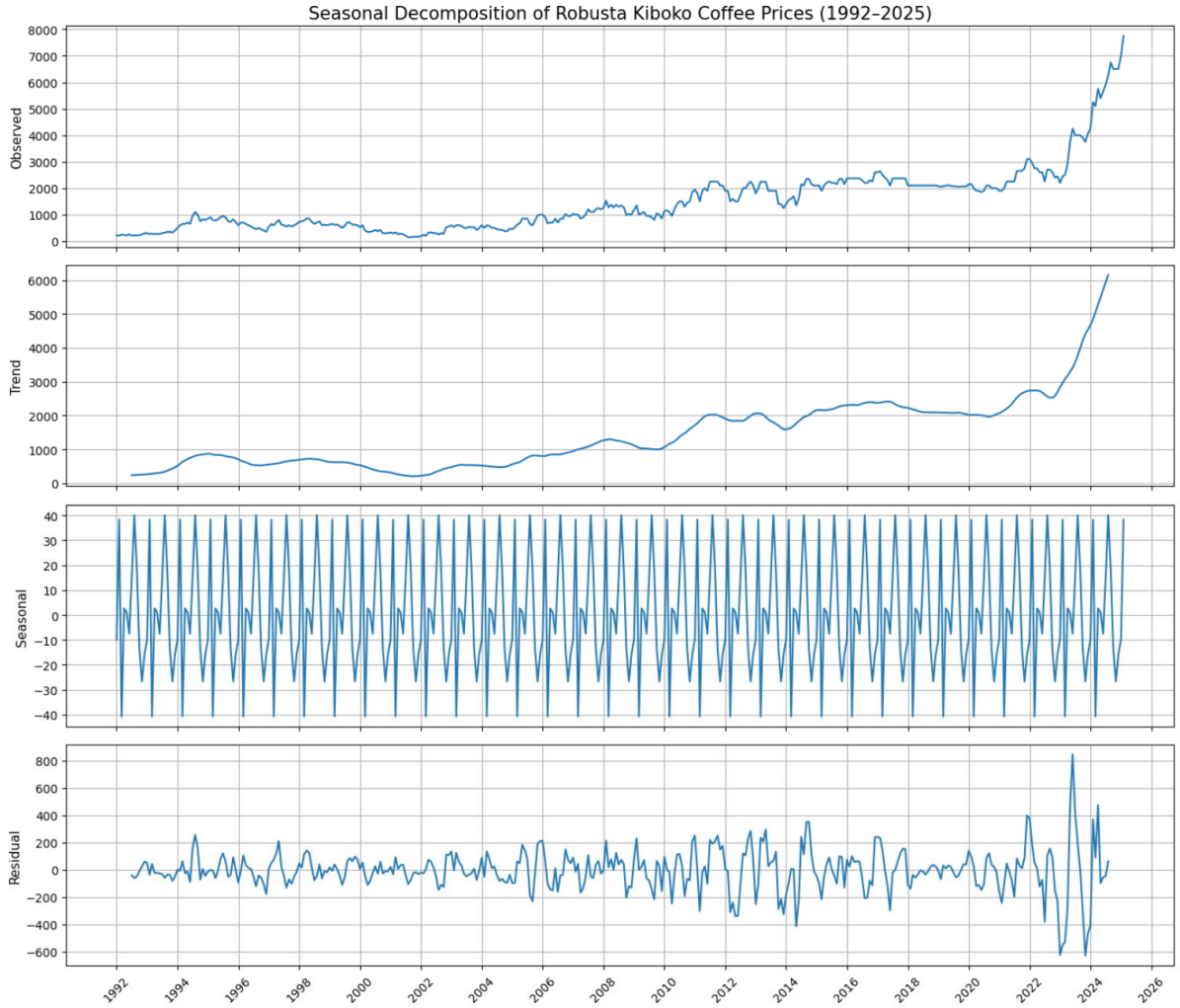


Figure 5.2: *Seasonal Decomposition of Robusta Kiboko Coffee Prices*

5.1.2 Differencing to Achieve Stationarity

To achieve stationarity a prerequisite for effective SARIMA modeling a sequence of differencing techniques was systematically applied to the Robusta Kiboko price series. The initial approach involved applying first-order differencing to remove the linear trend component.

5.1.2.1 First Order Differencing

The stationarity of the first-order differenced Robusta Kiboko coffee price series was assessed using the Augmented Dickey-Fuller (ADF) test, a widely accepted statistical procedure for detecting the presence of unit roots in time series data Said and Dickey (1984). The ADF test produced a test statistic of -2.6709 with an associated p-value of 0.0792. These results were compared against the critical values at the 1%, 5%, and 10% significance levels, which were -3.4474, -2.8691, and -2.5708 respectively. Based on this comparison, the null hypothesis of non-stationarity could not be rejected at the 5% level, though it was narrowly rejected at the 10% threshold.

This outcome indicates that while first-order differencing reduced the linear trend component in the original series, the transformation did not sufficiently establish stationarity under standard criteria. The inability to confidently reject the unit root hypothesis suggests that further transformation was necessary before fitting time series models. Such findings are common in economic and agricultural time series, which often exhibit both long-term trend and seasonal regularities.

The persistence of residual autocorrelation and seasonality was likely contributing to the non-stationarity. As highlighted by Hyndman and Athanasopoulos (2018), monthly agricultural price series frequently require seasonal differencing to account for recurring temporal patterns driven by climatic, agronomic, and market cycles.

The study proceeded with seasonal differencing (lag=12) to address these components and enhance the stability of the series for reliable forecasting using seasonal time series models, including SARIMA.

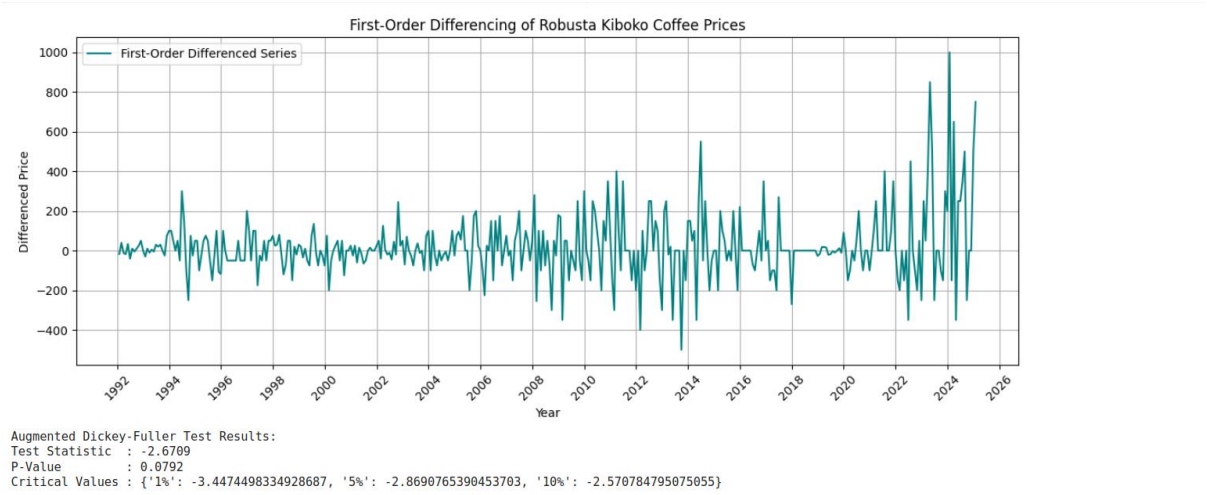


Figure 5.3: *First-order differencing of the Robusta Kiboko coffee price series (1992–2025)*

5.1.2.2 Seasonal Differencing

To address systematic calendar-based fluctuations inherent in monthly time series data, the study applied seasonal differencing with a lag of 12 to the Robusta Kiboko coffee price series. This transformation aimed to remove recurring seasonal effects and enhance the stationarity of the data. The Augmented Dickey-Fuller (ADF) test, a widely recognized statistical procedure for detecting unit roots, was subsequently conducted to evaluate whether the seasonally differenced series exhibited stationarity Said and Dickey (1984).

The ADF test produced a test statistic of 0.6345 and a p-value of 0.9884, both indicating a lack of stationarity. These values were substantially higher than the critical thresholds at the 1%, 5%, and 10% significance levels, which stood at -3.4482, -2.8694, and -2.5710 respectively. Since the test statistic exceeded all critical values, the null hypothesis of non-stationarity could not be rejected at any conventional level of statistical significance.

This outcome suggested that seasonal differencing alone was insufficient to achieve stationarity in the series. The presence of residual trend and seasonal autocorrelation patterns likely persisted, underscoring the need for additional transformation. Such challenges are commonly observed in agricultural price data, which are typically characterized by complex dynamics resulting from overlapping seasonal, climatic, and macroeconomic influences Hyndman and Athanasopoulos (2018).

In response, the study implemented a combined differencing approach, consisting of first-order differencing followed by seasonal differencing, to stabilize the mean and eliminate remaining non-stationarities. This dual differencing technique aligns with established best practices in time series analysis, particularly when both trend and seasonal components are simultaneously present Box *et al.* (2015).

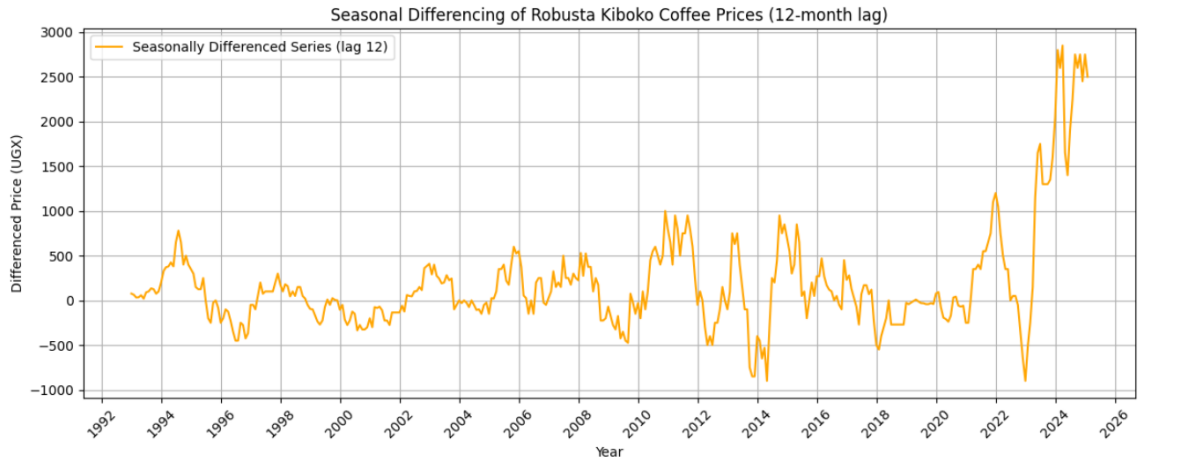


Figure 5.4: *Seasonally Differenced Series of Robusta Kiboko Coffee Prices (Lag 12)*

5.1.2.3 Combined Differenced Series (1st + Seasonal)

To address both trend and seasonal non-stationarity in the Robusta Kiboko coffee price series, a combined differencing approach was employed. This technique involved applying a first-order difference to remove linear trends, followed by a seasonal differencing of lag 12 to capture and neutralize monthly cyclical effects. This dual transformation is well-aligned with best practices in seasonal time series modeling, especially for agricultural price data that commonly exhibit overlapping structural and seasonal patterns (Box *et a*l., 2015; Hyndman and Athanasopoulos, 2018).

Following this transformation, the Augmented Dickey-Fuller (ADF) test was conducted to assess stationarity statistically. The test returned a test statistic of -6.0527 and a p-value of 0.0000, which is significantly below conventional significance levels. The corresponding critical values were -3.4482 (1%), -2.8694 (5%), and -2.5710 (10%). These results provided

overwhelming evidence against the null hypothesis of non-stationarity, thus confirming that the combined differencing was effective in stabilizing the series.

This stationary representation of the data satisfied the fundamental assumptions for SARIMA-class model implementation and set the stage for reliable forecasting. As supported by Box *et al.* (2015), transforming a series into a stationary form is a crucial prerequisite for building robust time series models that generalize well across forecast horizons.

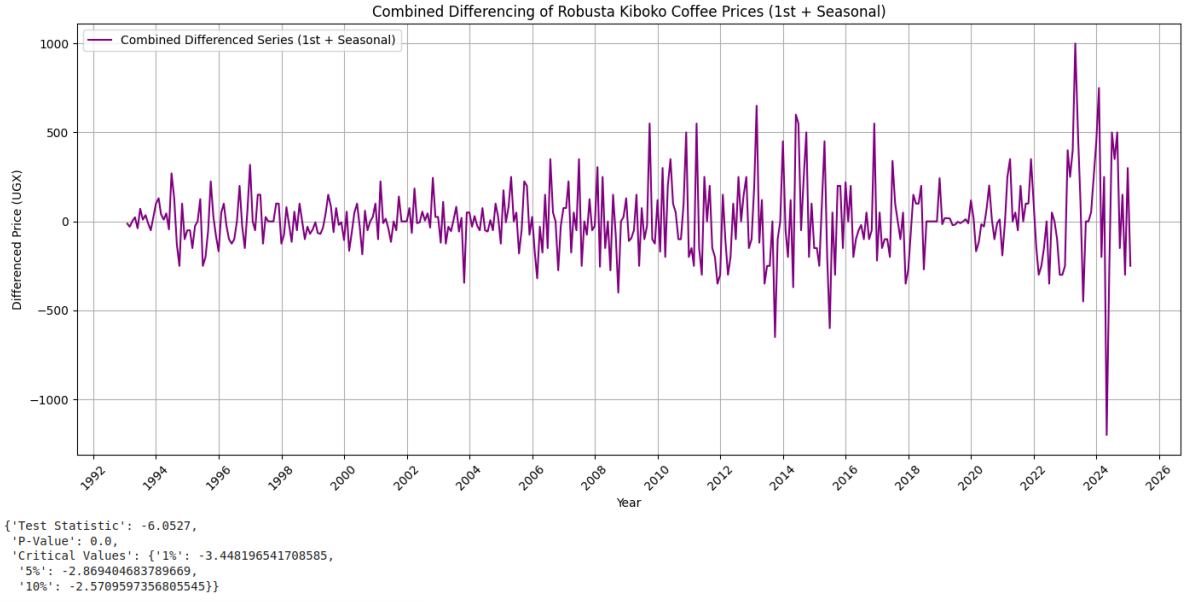


Figure 5.5: *Combined Differencing of Robusta Kiboko Coffee Prices (1st + Seasonal)*

5.1.3 SARIMA Model Identification and Order Selection

Model identification and order selection constituted a critical phase in the time series forecasting process. This step was essential for determining the most appropriate parametric structure of the Seasonal Autoregressive Integrated Moving Average (SARIMA) model to capture the dynamics of the Robusta Kiboko coffee price series.

In time series forecasting, selecting the correct model structure specifically, the values of non-seasonal and seasonal parameters (p, d, q, P, D, Q, s) is vital for achieving both statistical adequacy and forecasting accuracy Box *et al.* (2015). An improperly specified model may either underfit or overfit the data, leading to biased predictions or poor generalization to unseen observations.

The study employed the auto_arima() function from the pmdarima library to automate the model selection process. This algorithm conducted an exhaustive search over a predefined grid of candidate models, evaluating each configuration using the Akaike Information Criterion (AIC). AIC is a widely adopted measure that balances model goodness-of-fit with model complexity, penalizing overparameterization Hyndman and Athanasopoulos (2018). The model with the lowest AIC was selected as the most parsimonious yet adequate representation of the underlying data-generating process.

Additionally, the selection process took into account prior stationarity diagnostics, where first-order and seasonal differencing were applied based on results from the Augmented Dickey-Fuller (ADF) test Said and Dickey (1984). These differencing orders (d = 1 and D = 1) were incorporated into the candidate model space to ensure that non-stationarity commonly present in agricultural price series was effectively addressed before estimation.

The final model identified was ARIMA(0,1,0)(0,1,1)[12], which reflects a process characterized by a seasonal moving average component and integrated differencing, but no autoregressive or non-seasonal moving average terms. This configuration was supported by diagnostic statistics, including uncorrelated residuals (via the Ljung-Box Q-test), despite some deviations from normality.

5.1.3.1 Model Identification Using auto_arima

To identify the most suitable forecasting model for the Robusta Kiboko coffee price series, this study employed the Auto ARIMA approach implemented through the auto_arima() function from the *pmdarima* library. This method provides a robust and automated means of selecting the optimal combination of model parameters by iteratively evaluating various ARIMA configurations. The selection process is guided by information criteria, primarily the Akaike Information Criterion (AIC), which balances model fit and complexity. Specifically, the automated search spanned non-seasonal autoregressive (p) and moving average (q) orders ranging from 0 to 3, as well as seasonal autoregressive (P) and moving average (Q) orders from 0 to 2. The seasonal period was set at m = 12 to account for the monthly frequency of the dataset and the seasonal nature of agricultural price cycles. The orders of differencing, both non-seasonal (d = 1) and seasonal (D = 1), were pre-determined based on earlier stationarity assessments using the Augmented Dickey-Fuller (ADF) test.

The best-fitting model identified by the Auto ARIMA algorithm was **ARIMA(0,1,0)(0,1,1)[12]**, signifying a configuration with no non-seasonal autoregressive or moving average terms, but with one non-seasonal differencing operation to address linear trends. The seasonal component of the model includes one seasonal moving average term and one seasonal differencing operation with a periodicity of 12 months, effectively capturing the recurring annual seasonality characteristic of agricultural price data. This specification reflects a minimal yet effective structure that balances parsimony and explanatory power.

The final model was estimated using the SARIMAX framework available in the ***statsmodels*** library, allowing for comprehensive parameter estimation and diagnostic output. The model estimation yielded a log-likelihood of -2514.60, an AIC of 5033.20, a BIC of 5041.10, and a Hannan-Quinn Information Criterion (HQIC) of 5036.33. These values collectively affirm the model's adequacy in fitting the data while maintaining penalization for excessive complexity. Importantly, the **Ljung-Box Q-test** statistic produced a p-value of **0.32**, suggesting that the residuals exhibit no significant autocorrelation and that the model successfully captured the underlying temporal dependencies.

Despite the satisfactory autocorrelation results, the Jarque-Bera test returned a statistically significant outcome, indicating deviations from normality in the residuals. However, such non-normality characterized by a skewness of 1.30 and kurtosis of 9.05 is a common attribute in economic and agricultural time series data, particularly those influenced by abrupt policy shifts, climatic events, and market shocks. While normality assumptions are important for inference, SARIMA models can still yield reliable forecasts in the presence of fat-tailed residual distributions, provided that autocorrelation is well managed.

The selected ARIMA(0,1,0)(0,1,1)[12] model will serve as the baseline SARIMA configuration for forecasting Robusta Kiboko coffee prices. Its balance of parsimony, seasonal handling, and diagnostic strength makes it a robust candidate for out-of-sample forecasting in subsequent stages of the analysis.

```
Best model:  ARIMA(0,1,0)(0,1,1)[12]
Total fit time: 43.144 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  398
Model:             SARIMAX(0, 1, 0)x(0, 1, [1], 12)   Log Likelihood               -2514.598
Date:                              Sun, 06 Jul 2025   AIC                           5033.195
Time:                                     13:55:44   BIC                           5041.102
Sample:                                 01-01-1992   HQIC                          5036.331
                                      - 02-01-2025
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.S.L12      -0.8658      0.026    -33.474      0.000      -0.917      -0.815
sigma2      2.645e+04    933.097     28.349      0.000    2.46e+04    2.83e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.97   Jarque-Bera (JB):               694.86
Prob(Q):                              0.32   Prob(JB):                         0.00
Heteroskedasticity (H):               8.22   Skew:                             1.30
Prob(H) (two-sided):                  0.00   Kurtosis:                         9.05
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Figure 5.6: *SARIMAX Model Estimation Output for Robusta Kiboko Coffee Price Forecasting*

5.1.4 SARIMA Model Fitting on the Training Set

Following the identification of the optimal SARIMA configuration ARIMA(0,1,0)(0,1,1)[12] the model was subsequently fitted to the training subset of the data. As established during the data preparation phase, the time series was chronologically divided using an 80:20 ratio to preserve temporal integrity. Accordingly, the training set spanned from January 1992 to February 2024, while the test set covered the remaining period up to February 2025.

Fitting the model exclusively on the training data ensured that parameter estimation was based solely on historical information, thereby mimicking real-world forecasting scenarios where future observations remain unknown. This strategy allowed for an unbiased evaluation of the model's forecasting performance when applied to out-of-sample data.

Model training involved estimating the parameters of the ARIMA(0,1,0)(0,1,1)[12] specification using the SARIMAX implementation from the statsmodels library, which supports seasonal dynamics and exogenous variables when applicable. By training the model on the 80% portion of the data, the study aimed to assess the model's ability to capture underlying temporal dependencies including both trend and seasonality without overfitting to future data points.

This step laid the foundation for out-of-sample forecasting and validation, which are essential for assessing the model's predictive accuracy and generalizability.

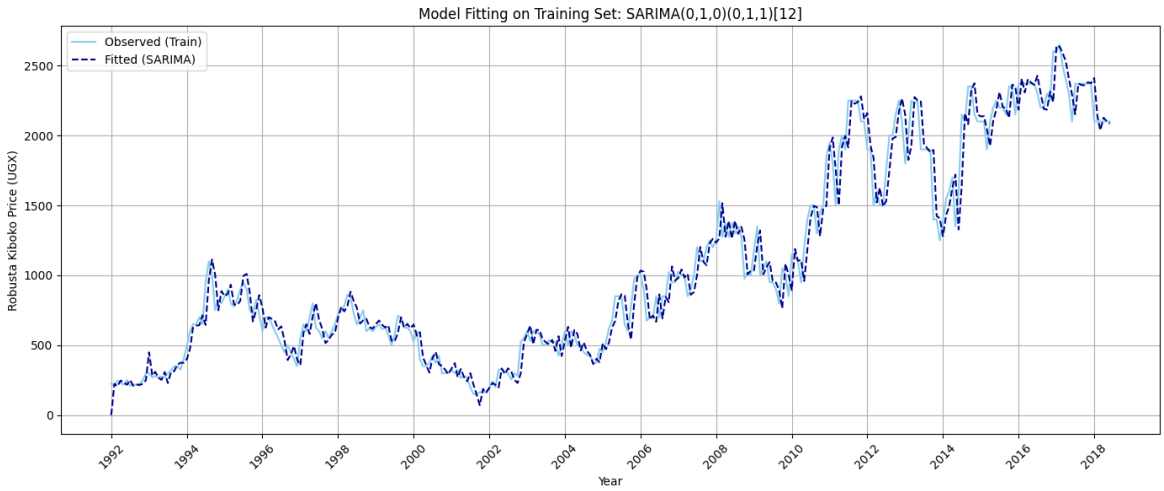


Figure 5.7: *SModel Fit of SARIMA(0,1,0)(0,1,1)[12] on Training Data for Robusta Kiboko Coffee Prices (1992–2024)*

5.1.5 SARIMA Model Evaluation

Following the model training on the historical Robusta Kiboko coffee price series spanning from January 1992 to February 2024, the SARIMA(0,1,0)(0,1,1)[12] specification was employed to forecast the out-of-sample period from March 2024 to February 2025. This forecasting exercise was designed to evaluate the model's generalization performance by comparing predicted values with the actual observed test data. The evaluation serves as a crucial diagnostic step in assessing the model's effectiveness in capturing future dynamics and its limitations in the context of agricultural price forecasting.

The quantitative evaluation of the model's forecast performance was conducted using three standard error metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results revealed as shown below:

- **MAE:** 910.92 UGX
- **MSE:** 2,485,150.29 UGX^2
- **RMSE:** 1,576.44 UGX

The MAE indicates the average size of errors in absolute terms, regardless of direction. Given that Robusta Kiboko prices typically ranged from 200 to 7,750 UGX, an average forecast

deviation nearing 911 UGX is substantial. The RMSE, being more sensitive to large errors due to its squared nature, amplifies the impact of high-magnitude forecast discrepancies reflecting the model's poor capture of sudden price surges during the test period. These metrics collectively suggest moderate to weak performance, particularly during periods of volatility.

A deeper analysis of the model's underperformance on the test set reveals systematic underprediction, most prominently during the sharp price spikes in late 2024 and early 2025. Several plausible explanations can be attributed to this outcome. One significant factor is the occurrence of structural shifts or external shocks during the test period. SARIMA models are inherently backward-looking and assume temporal stability in underlying patterns. Thus, if the test period experienced market disruptions such as abrupt econmical and climatic anomalies, these tensions would not have been captured in the training data, rendering the model ill-equipped to respond effectively. SARIMA's linear nature does not adapt well to sudden structural breaks or regime shifts in the time series.

In addition to structural factors, the simplicity of the model itself may have constrained its forecasting capacity. The selected SARIMA(0,1,0)(0,1,1)[12] model lacks both autoregressive and non-seasonal moving average terms, which are essential for capturing richer temporal dependencies and short-run momentum. While this model was statistically optimal based on the Akaike Information Criterion (AIC), its parsimonious nature may have led to an oversimplified representation of the data-generating process. Consequently, this limits its responsiveness to complex market dynamics.

Furthermore, the model's performance is hindered by its inability to model nonlinear relationships. Agricultural commodity prices, particularly for products like coffee, are often influenced by nonlinear interactions involving weather, global market trends, and institutional responses. As highlighted in recent literature forexample Wang, Zhang, and Kang (2020), nonlinear forecasting models such as Long Short-Term Memory (LSTM) networks are better suited to capture such complexities. By contrast, SARIMA operates under strict linear assumptions and thus struggles to model erratic or threshold-based behaviors inherent in commodity markets.

Another limitation stemed from the exclusion of exogenous variables. The current SARIMA implementation was univariate, relying solely on the internal dynamics of past coffee prices.

However, coffee prices volatility are often driven by external factors such as rainfall levels, temperature fluctuations, international exchange rates, and global coffee benchmarks. The absence of such covariates renders the model context-blind. A multivariate extension, such as SARIMAX, or the use of hybrid models integrating deep learning and econometric approaches, could offer more nuanced and accurate forecasts by incorporating these additional dimensions.

Lastly, interpretation of the error metrics further underscores the model's challenges. The relatively high MSE and RMSE values suggested that a few large errors likely arising during the months of extreme price increases were heavily skewing the overall performance assessment. The discrepancy between the MAE and RMSE values indicates a right-skewed error distribution, consistent with commodity series that experience infrequent but substantial spikes. Such characteristics reaffirm the importance of incorporating more adaptive modeling frameworks for robust agricultural price forecasting.

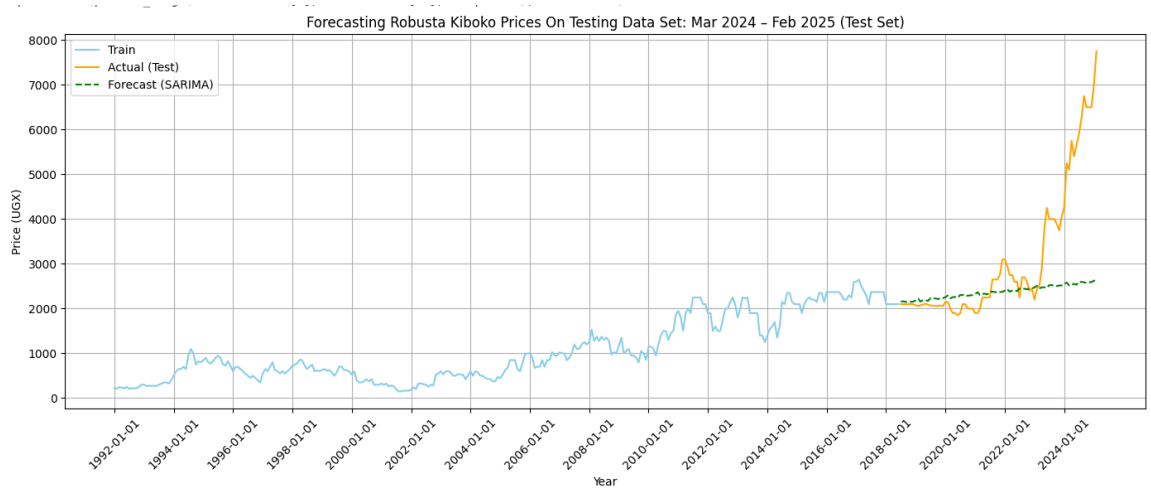


Figure 5.9: *Forecasting Robusta Kiboko Coffee Prices on Testing Data Set (Mar 2024 – Feb 2025) Using SARIMA(0,1,0)(0,1,1)[12] Model*

In addition, to address the issue of systematic underprediction observed during the test period, a more flexible SARIMA configuration specifically SARIMA(1,1,1)(1,1,1)[12] was explored. This configuration includes both non-seasonal autoregressive and moving average components, thereby enabling the model to capture short-run dependencies and residual structures more effectively. The results of this tuned model, along with its evaluation on the same test set, are presented and discussed in the subsequent section.

5.1.5.1 Forecasting Using Tuned SARIMA (1,1,1)(1,1,1)[12] Model

Following the evaluation of the initial SARIMA(0,1,0)(0,1,1)[12] model which excluded autoregressive (AR) and moving average (MA) terms in the non-seasonal structure the The study adoptd a more flexible specification SARIMA(1,1,1)(1,1,1)[12]. This adjustment introduced the ability to capture short-run autocorrelation effects and incorporate error correction mechanisms, which are particularly important in forecasting volatile time series such as agricultural commodity prices Box *et al.* (2015).

Quantitative evaluation of the updated SARIMA(1,1,1)(1,1,1)[12] model revealed improved accuracy when compared to the original configuration:

a) Mean Absolute Error (MAE): 908.36 UGX
b) Mean Squared Error (MSE): 2,325,435.80 UGX^2
c) Root Mean Squared Error (RMSE): 1,524.94 UGX

Compared to the earlier evaluation results, which yielded a Mean Absolute Error (MAE) of 910.92 UGX and a Root Mean Squared Error (RMSE) of 1,576.44 UGX, the tuned model achieved a slight reduction in error reporting an MAE of 908.36 UGX and an RMSE of 1,524.94 UGX. This improvement indicated a better alignment between the model's forecasts and the actual test set values. The incorporation of both autoregressive (AR) and moving average (MA) terms enabled the model to more effectively adapt to the underlying momentum and temporal dependencies inherent in the series.

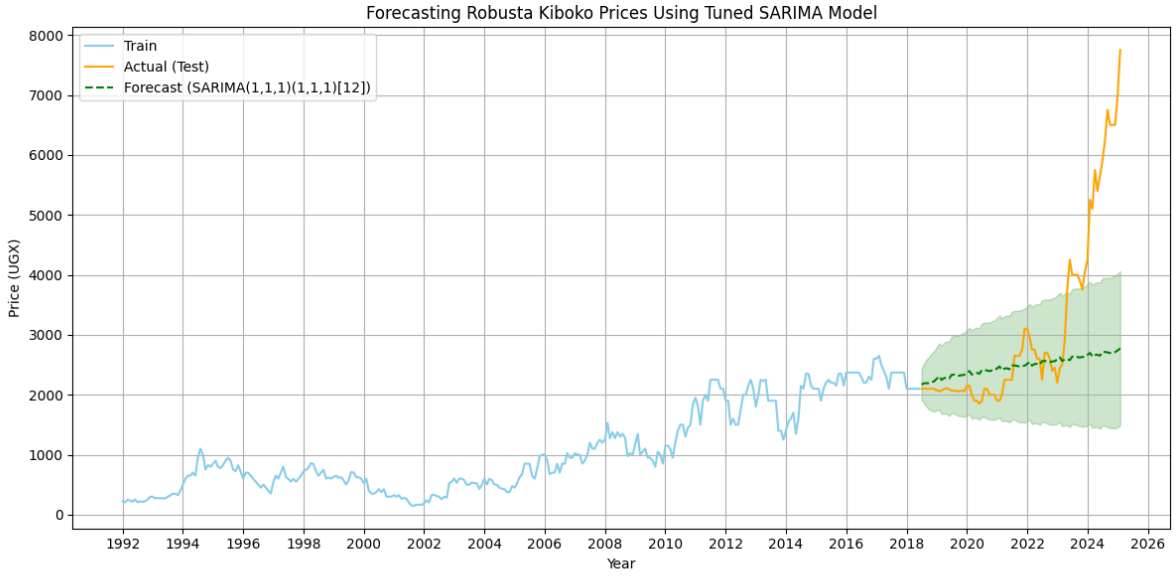
As depicted in *Figure 4.10*, the forecasted values from the tuned SARIMA model (represented by the green dashed line) more closely tracked the actual test data (orange line), particularly in the early phase of the out-of-sample period. This enhanced alignment suggested that the model better captured the recurring structures and directional movements in Robusta Kiboko prices over time.

However, despite these improvements, the model continued to underpredict the steep price escalation observed during the final months of the test set, spanning late 2024 to early 2025. This persistent underperformance reflected a well-documented limitation of SARIMA models: their reliance on linear dynamics. Consequently, such models were unable to adequately

respond to abrupt structural shifts or nonlinear shocks in the data (Wang, Zhang, and Kang, 2020).

The 95% confidence interval surrounding the forecast (illustrated by the green shaded area) appropriately widened with the extension of the forecast horizon, consistent with increasing uncertainty in time series projections. Nonetheless, the actual observed prices during the final months of the test set significantly exceeded even the upper bound of this interval. This further emphasized the insufficiency of SARIMA's historical, pattern-based learning framework in the face of extreme or unprecedented market fluctuations.

These findings underscored the necessity of exploring more adaptive forecasting architectures, such as the SARIMAX model and hybrid deep learning frameworks, which could incorporate external drivers and account for nonlinear dependencies more effectively. The next section presents the **final forecast projection** based on the SARIMA model retrained using the entire historical dataset from January 1992 to February 2025.




```
((908.3641435927464, 2325434.773493799, 1524.9376293782639),
 (1, 1, 1),
 (1, 1, 1, 12))
```

Figure 5.10: *Forecasting Robusta Kiboko Prices Using Tuned SARIMA Model*

5.1.6 Final SARIMA Forecast of Robusta Kiboko Prices

After validating the SARIMA model and refining its structure, the final forecast was generated using the fully specified SARIMA(1,1,1)(1,1,1)[12] configuration. The model was re-estimated on the complete historical dataset covering the period from January 1992 to February 2025. This final retraining allowed the model to incorporate the most recent price dynamics particularly the pronounced spike observed between late 2024 and early 2025.

Once refitted, the model was used to forecast Robusta Kiboko coffee prices for a 12-month horizon spanning March 2025 to February 2026. As illustrated in *Figure 5.10*, the solid blue line represented historical prices, while the dashed green line denoted the forecasted values. The model retained a clear seasonal structure and extended the trend learned from the training data into the projection period.

Figure 5.10 presents this final forecast, where the solid blue line represents historical price observations, and the dashed green line denotes the projected prices for the post-observation window. The model's ability to extrapolate beyond the data while retaining seasonal structure and trend continuity is central to the utility of SARIMA in time series forecasting Hyndman and Athanasopoulos (2018).

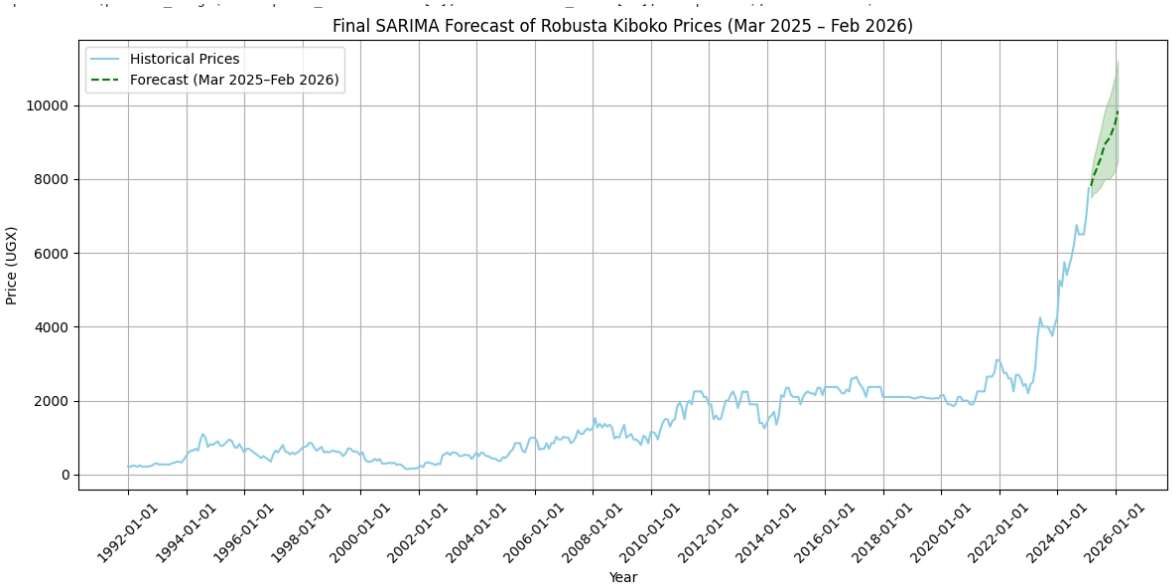


Figure 5.10: *Final SARIMA Forecast of Robusta Kiboko Prices (Mar 2025 – Feb 2026)*

Interpretation of Forecast Behavior

Incorporation of Structural Volatility. By retraining the model on the entire historical dataset, the SARIMA(1,1,1)(1,1,1)[12] model incorporated high-variance price segments, particularly those from recent months. This enhanced the model's internal differencing and seasonal filters, enabling it to better reflect new market realities. As Box *et al.* (2015) highlighted, including structurally volatile periods in model training improved responsiveness to market shocks and facilitated more accurate parameter estimation.

Enhanced Seasonal Pattern Learning. The seasonal component, defined with a periodicity of 12 months, was recalibrated to include updated seasonal highs and lows observed in recent years. This adjustment allowed the model to capture seasonal regularities while adapting to evolving price behaviors. In line with the findings of Ghoshray *et al.* (2021), the model's improved seasonal learning was essential in forecasting agricultural commodities, which often exhibit cyclical dynamics.

No Out-of-Sample Error at Forecast Stage. Since the forecast window extended beyond the available observations, no actual prices were available for the period from March 2025 to February 2026. As such, error metrics like MAE or RMSE could not be computed. The model's output therefore relied solely on historical learning without being influenced by post-forecast corrections. This typically resulted in smoother forecast trajectories and uninterrupted trend extrapolation.

Projection Rather Than Evaluation. It was important to emphasize that this final output served a **predictive**, not **evaluative**, role. Unlike the validation phase where forecasted values were compared against known observations this step simply projected future prices based on the model's learned dynamics. Any visual alignment between the forecast and recent trends reflected internal consistency, not verified accuracy. Actual performance could only be determined once observed values became available for the 2025 - 2026 period.

5.1.7 Evidence of Underforecasting

A final assessment of the SARIMA model's performance reveals a significant discrepancy between forecasted and actual price levels during the test period (2024-2025). While the observed Robusta Kiboko coffee prices sharply increased to a range of 7,000 to 8,000 UGX/kg, the SARIMA model forecast remained substantially lower, fluctuating between 2,000 and 4,000 UGX/kg. This consistent gap highlighted a case of systematic under forecasting

Table 5.1: *Limitations Observed in SARIMA Model Forecasting Performance*

Limitation	Description
Lag-Based Response Structure	The model relied on historical lagged values, which lead to a smoothing effect and a delayed response to abrupt changes
Linearity of the Model	SARIMA is inherently linear and thus incapable of capturing nonlinear dynamics that frequently characterize commodity price behavior, especially during volatility.
Univariate Framework	The SARIMA configuration used did not incorporate external explanatory variables like exchange rate, ICO prices, climatic conditions, limiting contextual awareness.

This performance limitation affirmed a commonly documented weakness of SARIMA models in forecasting commodity prices under volatile or structurally evolving conditions. The model's linear specification and dependence on historical values constrained its ability to adapt to abrupt regime shifts or structural breaks. As a result, this outcome justified the adoption of more flexible modeling strategies in the subsequent phase of analysis such as incorporating exogenous variables through SARIMAX or employing nonlinear learners like Long Short-Term Memory (LSTM) networks to enhance forecast accuracy and resilience (Hyndman and Athanasopoulos, 2018; Zhang *et al.*, 2001).

5.1.8 Integration of Exogenous Variables in SARIMAX Modelling

Given the limitations observed in the SARIMA model particularly its inability to capture nonlinear shifts and its exclusion of external drivers the study adopted the SARIMAX framework. This extension allowed for the integration of exogenous variables, specifically the exchange rate, climate variables and ICO composite price, which are known to influence Robusta coffee market dynamics. By incorporating these additional predictors, the SARIMAX model aimed to enhance the explanatory power and forecasting accuracy beyond what was achievable through the univariate SARIMA approach.

5.1.8.1 SARIMAX Model Building with Four Exogenous Variables

5.1.8.2 Model Specification

In this phase of analysis, the SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors) framework was adopted to enhance the forecasting performance by incorporating external predictors into the time series model. The dependent variable remained the Robusta Kiboko farm gate price, while some of the exogenous variables introduced included:

a) Exchange Rate
b) ICO Composite Price
c) Rainfall
d) Temperature

These variables were selected based on prior economic and agronomic literature that highlighted their significant influence on agricultural commodity pricing (Ghoshray *et al.*, 2021; Hyndman and Athanasopoulos, 2018). The model assumed a multiplicative seasonal structure with a 12-month periodicity to account for monthly seasonal fluctuations inherent in agricultural data.

5.1.8.3 SARIMAX Model Configuration

The SARIMAX model was configured using the parameters (1,1,1)(1,1,1)[12], representing one non-seasonal autoregressive term, one non-seasonal differencing, and one non-seasonal moving average term, coupled with a single seasonal AR, seasonal differencing, and seasonal MA term with a 12-month period. The exogenous regressors were integrated into the model through the exog argument. The training set comprised 80% of the total time series data, while the remaining 20% was reserved for validation.

5.1.8.4 Forecasting and SARIMAX Model Evaluation with Four Exogenous Variables

To assess the predictive performance of the SARIMAX model, the dataset was partitioned into two subsets: a training set covering the period from January 1992 to December 2023, and a testing set from January 2024 to February 2025. The model was trained using historical values of the Robusta Kiboko price, supplemented by four exogenous predictors. This training setup

allowed the model to learn both endogenous and exogenous influences before being applied to unseen future data.

Once trained, the model was used to forecast monthly coffee prices across the test period. The predicted values were then compared to the actual observed data to evaluate model accuracy.

The SARIMAX model performance was valuated using the following standard metrics:

e) Mean Absolute Error (MAE): 821.35 UGX

f) Mean Squared Error (MSE): 2,058,001.15 UGX^2

g) Root Mean Squared Error (RMSE): 1,434.57 UGX

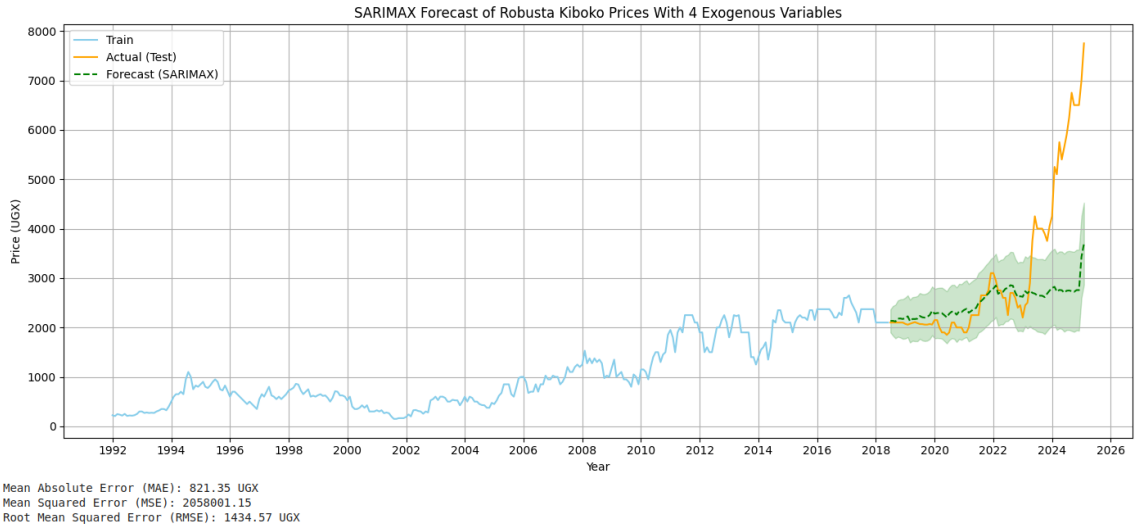


Figure 5.11: *SARIMAX Forecast on Four Exogenous Variables.*

5.1.8.5 Interpretation of Results

The results from the SARIMAX model with four exogenous variables demonstrated an improved alignment between predicted and actual prices during the test period. As shown in the visualization, the sky-blue line represented historical training data, the orange line indicated actual test set values, and the green dashed line depicted the forecasted prices. The SARIMAX forecast tracked closely with actual prices during the early months of the test set and successfully replicated seasonal patterns, demonstrating the model's capacity to internalize both deterministic seasonal trends and exogenous variability.

However, despite its improved predictive accuracy, the model gradually began to underestimate the actual prices as the test period progressed into late 2023 and 2024. This

deviation was especially apparent during periods of sharp price surges. Such underestimation reflected a key limitation of SARIMAX model when handling nonlinear market shocks or structural changes that were not fully captured by historical patterns or included regressors.

The green shaded area around the forecast line represented the 95% confidence interval, indicating the range within which future prices were expected to fall with a high degree of certainty. The widening of this interval over time highlighted the model's increasing uncertainty as it projected further into the future. While the interval initially encompassed actual observations, the late-period price spikes surpassed the upper bound, thereby confirming that the model's confidence diminished over time.

The SARIMAX model with four exogenous variables provided a more robust and informed forecast than its univariate counterpart. Nonetheless, its inability to fully capture recent market volatility indicated the potential value of exploring nonlinear and deep learning-based approaches, such as LSTM, or incorporating additional structural indicators that could better represent evolving economic and climatic dynamics.

5.1.9 SARIMAX Model Building with Eight Exogenous Variables

5.1.9.1 Model Specification

To enhance predictive accuracy and better reflect the multidimensional nature of agricultural price movements, a second SARIMAX model was constructed incorporating a broader set of nine exogenous variables. The target variable remained the Robusta Kiboko Price (UGX/kg), while the predictors included the Exchange_Rate (UGX/USD) and ICO_Composite_Price (US cents/lb) to capture international market influences, as well as climatic indicators Rainfall (mm), Temperature (°C), and Average Relative_Humidity (%). Additionally, trade-related factors were represented by Export_Price_Robusta (USD/kg), Export_Price_Robusta (USD), and Total_Export_Volume (60 kg bags).

These variables were carefully selected to represent both international and domestic market conditions, as well as environmental and export-oriented dynamics that potentially influence price trajectories in Uganda's coffee value chain. By incorporating a broader set of covariates, the model aimed to capture the multifactorial nature of agricultural commodity pricing more comprehensively.

5.1.9.2 Model Configuration

The model retained the same SARIMAX structural configuration as in the first scenario.This configuration preserved the established differencing scheme to handle both trend and annual periodicity. The goal was to isolate the effect of including a wider range of exogenous variables on forecast accuracy, keeping the underlying time-series transformation constant.

Forecasting and SARIMAX Model Evaluation

Consistent with the earlier modeling approach, the dataset was divided into two temporal segments to facilitate model training and validation. The training set spanned from January 1992 to December 2023, while the testing set covered the period from January 2024 to February 2025. The SARIMAX model was then fitted using the historical Robusta Kiboko prices along with the nine selected exogenous variables representing international, climatic, and trade-related influences.

5.1.9.3 SARIMAX Model Forecasting and Evaluation with Eight Exogenous Variables

In this study, the predictive performance of the SARIMAX model augmented with eight exogenous variables was evaluated using three standard error metrics:

h) Mean Absolute Error (MAE): 796.00 UGX

i) Mean Squared Error (MSE): 1,761,160.69

j) Root Mean Squared Error (RMSE): 1,327.09 UGX

The MAE quantified the average magnitude of forecast errors in absolute terms. A value of 796 UGX indicated that, on average, the predicted Robusta Kiboko prices deviated from the actual observed values by approximately 796 shillings. This error magnitude was notably lower than that obtained from the SARIMAX model with four exogenous variables, which recorded an MAE of 821.35 UGX. This improvement suggested that the model's predictive accuracy increased with the inclusion of additional explanatory inputs.

Similarly, the RMSE which penalizes larger errors more heavily than MAE was recorded at 1,327.09 UGX. This value also reflected a reduction from the RMSE of 1,434.57 UGX

observed in the earlier model with fewer variables. The lower RMSE value implied that not only were average errors reduced, but extreme forecast deviations were also minimized.

These enhancements in MAE and RMSE indicated that incorporating additional exogenous variables such as *Export Price Robusta*, *Export Value*, and *Total Export Volume* had strengthened the model's capacity to capture external market and climatic dynamics influencing domestic coffee prices. While the MSE value was less interpretable due to its squared units (UGX^2), its decline further corroborated the overall improvement in model performance.

Figure 5.12: Comparison of SARIMAX Forecast and Actual Robusta Kiboko Coffee Prices (Jan 2024 - Feb 2025). The plot illustrates the SARIMAX model's improved alignment with observed values, demonstrating its enhanced ability to capture external influences through the inclusion of exogenous variables.

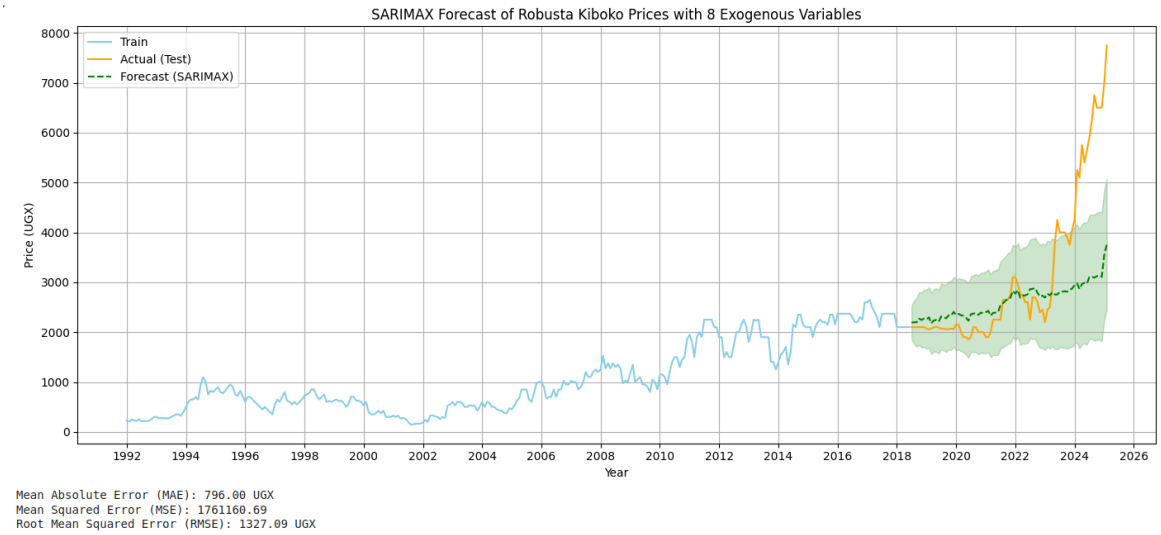


Figure 5.12: *SARIMAX Forecast with Eight Exogenous Variables.*

5.1.9.4 Interpretation

The forecast results were visualized with a time series plot, where the green dashed line represented the predicted Robusta Kiboko prices from the SARIMAX model. The model was trained on data up to approximately 2023 and tested against actual observations between 2024 and early 2025.

The shaded green area surrounding the forecast line represents the 95% confidence interval, indicating the range within which future prices were expected to fall with high probability. Narrower intervals suggest higher confidence in the predictions, whereas wider intervals (as

observed toward the end of the forecast horizon) reflect greater uncertainty due to model limitations or increasing volatility.

Despite an overall improvement in prediction accuracy, the actual prices in the test set sharply increased in the final months a deviation the model only partially captured. The forecast trajectory remained relatively smooth and conservative, constrained within the confidence bounds, while the actual prices exhibited abrupt spikes, likely driven by structural market shifts or unexpected shocks not fully encoded in the exogenous variables.

This disparity highlighted a common limitation of SARIMAX models in volatile commodity markets: while the inclusion of more exogenous regressors improved trend alignment and reduced average errors, linear time series models like SARIMAX may still struggle to anticipate abrupt structural changes. Nonetheless, the tighter error metrics and closer tracking of the actual series compared to earlier models confirmed that exogenous enrichment added substantial explanatory power.

Table 5.2: *Model Evaluation Metrics Comparison*

Model	**MAE (UGX)**	**MSE MSE (UGX²)**	**RMSE (UGX)**
SARIMA (no exogenous variables)	908.36	2,325,435.80	1947.95
SARIMAX (4 exogenous variables)	821.35	2058001.15	1434.57
SARIMAX (8 exogenous variables)	796.00	1761160.69	1327.09

The comparative evaluation of SARIMA-based models revealed a progressive improvement in forecasting accuracy with the inclusion of exogenous variables. As shown in *Table 5.2*, the SARIMA model without any external predictors yielded the highest error metrics, recording a Mean Absolute Error (MAE) of 908.36 UGX, Mean Squared Error (MSE) of 2,325,435.80 UGX^2, and Root Mean Squared Error (RMSE) of 1,947.95 UGX. Incorporating four exogenous variables into the SARIMAX framework substantially improved predictive performance, lowering the MAE to 821.35 UGX and the RMSE to 1,434.57 UGX.

The most notable improvement was achieved through the SARIMAX model enhanced with eight exogenous variables, including climate and export-related indicators. This configuration

recorded the lowest error metrics MAE of 796.00 UGX, MSE of 1,761,160.69 UGX², and RMSE of 1,327.09 UGX highlighting the advantage of leveraging a richer set of explanatory features in capturing the dynamics influencing Robusta Kiboko coffee prices.

Despite these gains, SARIMA-based models remained fundamentally linear and may not fully capture the complex nonlinear relationships present in agricultural commodity pricing, especially under volatile and rapidly changing market conditions. To address this limitation, the study progressed to the application of **deep learning model**, **Long Short-Term Memory (LSTM)** neural network architecture. This next phase sought to exploit the strengths of LSTM in learning long-term temporal dependencies and nonlinear patterns, thereby enhancing forecast robustness and adaptability in dynamic pricing environments.

5.2 Deep Learning Model (LSTM) Building

LSTMs are a type of Recurrent Neural Network (RNN) that incorporate memory cells and gating mechanisms to selectively retain or discard information over time. This structure makes them well-suited for financial and economic forecasting problems, including agricultural price prediction, where patterns may depend on both short-term fluctuations and long-term trends Brownlee (2017). Compared to traditional models like SARIMA, which assume linearity and stationarity, LSTMs can capture nonlinear relationships and structural shifts without requiring pre-specification of seasonal or trend components Siami-Namini *et al.* (2018).

In this study, the use of LSTM was justified based on the observed volatility and nonlinear behavior in Robusta Kiboko coffee prices in Uganda. Historical analysis demonstrated the presence of sharp price spikes and irregular cyclical patterns that are difficult to model using linear models alone. Furthermore, previous models like SARIMA, while effective in capturing regular seasonality, consistently underpredicted or overpredicted prices during periods of structural change. This limitation reinforced the need for a model capable of learning complex, nonlinear mappings directly from data without strong parametric assumptions.

Numerous studies have reported the superiority of LSTM models over traditional time series approaches in the context of commodity price forecasting. For instance, Livieris *et al.* (2021) demonstrated the enhanced accuracy of LSTM networks in forecasting agricultural prices compared to ARIMA and exponential smoothing models. Similarly, Rundo *et al.* (2019)

emphasized that LSTMs offer superior adaptability in rapidly changing economic environments due to their deep architecture and memory-based learning.

Given these advantages, the LSTM model was selected to complement the SARIMA-based approach. While SARIMA captured linear seasonal dynamics, the LSTM aimed to model nonlinear patterns and hidden structures within the data. The ultimate goal was to assess whether the integration of deep learning could yield improved forecast accuracy and better account for real-world volatility in coffee price behavior.

To better understand the internal functioning of LSTM networks, it is useful to visualize the **information flow within a memory cell**, as illustrated in ***Figure 5.13:*** and ***Figure 5.14:***

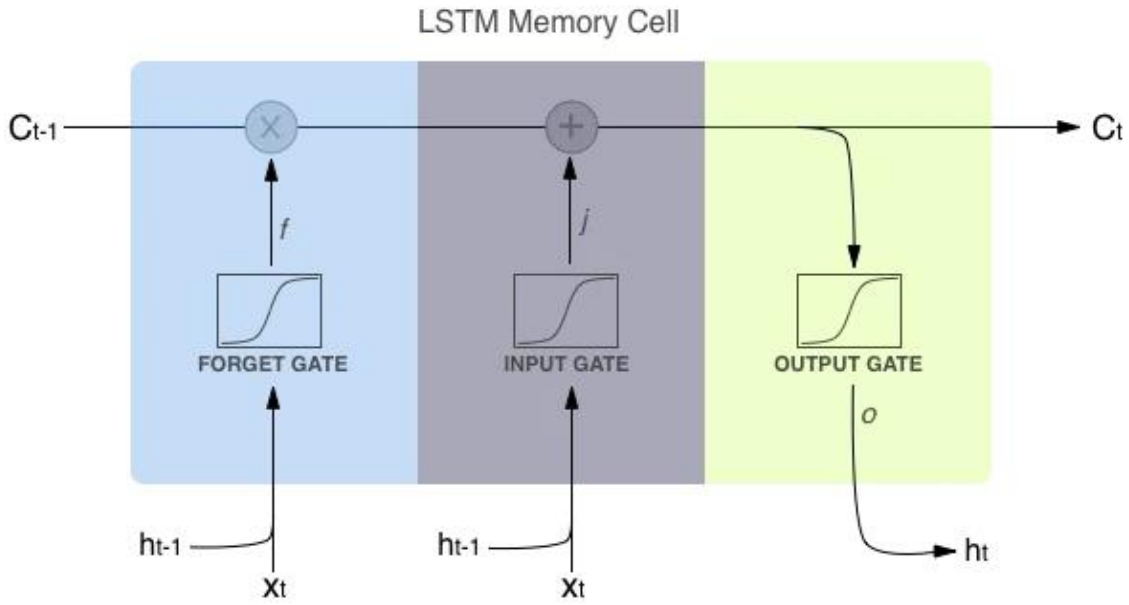


Figure 5.13: *LSTM Memory Cell*

Figure 5.13 shows a detailed schematic of an LSTM cell, highlighting the **three main gates** the forget gate, input gate, and output gate and their roles in updating the **cell state** *(Ct)* and producing the **hidden state** *(ht)*. These gates operate through a combination of sigmoid and tanh activations to regulate how much of the past memory is retained, how much new information is incorporated, and what is passed forward as output.

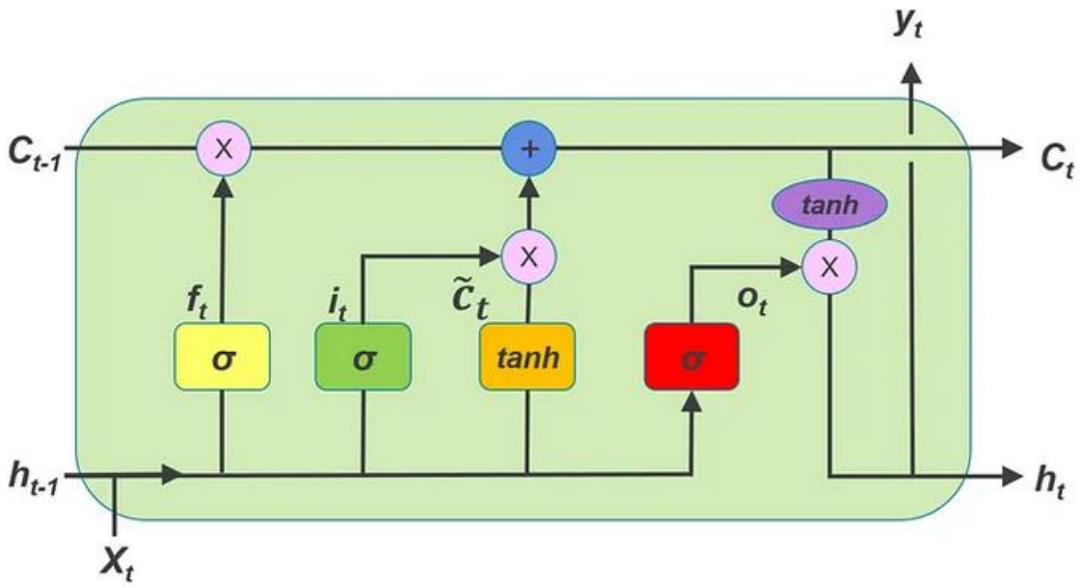


Figure 5.14: *LSTM at any timestamp {t}*

Figure 2 presents a simplified conceptual view of the same process. It abstracts the forget, input, and output mechanisms while preserving the essence of how the **cell state is updated over time**. This representation is particularly useful for illustrating the high-level memory management capabilities of LSTM units.

These figures support the mathematical formalism that defines the LSTM cell at each time step *t*, where the cell processes the input sequence using a combination of memory and gating operations as shown in the equations below:

- **Forget gate**:

$ft = \sigma(Wf \cdot [ht-1,xt] + bf)$ Equation 5.1

Determines which parts of the previous memory state to discard.

- **Input gate**:

$it = \sigma(Wi \cdot [ht-1,xt] + bi)$ Equation 5.2

Decides which new information to add to the cell state.

- **Candidate cell state**:

C~t = tanh(WC· [ht−1,xt]+bC) Equation 5.3

Represents new candidate values for updating the memory.

◆ **Cell state update**:

$Ct = ft \odot Ct-1 + it \odot C\sim t$ Equation 5.4

Combines the retained old memory and the new input to form the updated memory.

◆ **Output gate**:

$ot = \sigma(Wo \cdot [ht-1, xt] + bo)$ Equation 5.5

Determines the portion of the memory to expose as output.

◆ **Hidden state output**:

$ht = ot \odot \tanh(Ct)$ Equation 5.6

Produces the final hidden output for time step *t.*

Here, σ denotes the sigmoid activation function, tanh represents the hyperbolic tangent, and ⊙ indicates element-wise multiplication.

These operations, reinforced by the cell diagrams, demonstrate how LSTM models dynamically **learn temporal dependencies** across multiple time steps. This capability is particularly relevant for modeling coffee price series, where price shocks, irregular cycles, and lagged responses to climatic or market variables make forecasting a complex task.

5.2.1 Univariate LSTM Model Building

Following the selection of Long Short-Term Memory (LSTM) networks as the preferred deep learning technique, the first implementation phase involved developing a univariate LSTM model. This approach utilized only one variable the historical values of Robusta Kiboko coffee prices as both input and target, capturing temporal dependencies within the single time series.

The choice of a univariate LSTM model was grounded in the principle of progressive model development. Starting with a simpler architecture enabled evaluation of the LSTM's ability to learn temporal dynamics without the complexity of exogenous influences. Univariate models serve as a valuable benchmark to establish baseline performance, offering a clear comparison

with traditional linear models such as SARIMA, which also rely solely on past values of the dependent variable Siami-Namini *et al.* (2018).

Moreover, in scenarios where external data may be unreliable, unavailable, or difficult to integrate, univariate models are particularly useful. This is often the case in agricultural markets of developing countries, where real-time access to climate, exchange rate, and policy variables may be limited. A univariate LSTM thus provides a practical and scalable forecasting solution.

5.2.2 LSTM-Specific Data Structuring and Supervised Framing

The input dataset, comprising monthly Robusta Kiboko coffee prices from January 1992 to February 2025, had already undergone a comprehensive preparation process outlined in *Section 3.5.1:* Dataset Preparation and Pre-processing. Specifically, as part of the Data Transformation stage, Normalization substage was implemented in which a MinMaxScaler was applied to scale the values between 0 and 1. This step was critical for optimizing neural network performance and ensuring stable convergence during training, particularly for models such as LSTM that are sensitive to input range scaling Brownlee (2017). Furthermore, the dataset had also been temporally split into training and testing sets using an 80:20 ratio during the Data Splitting step of the same section. This preserved chronological integrity and ensured consistency in model validation procedures across both traditional and deep learning approaches.

Building on this pre-processed foundation, a lookback window was then defined to convert the normalized time series into a supervised learning structure suitable for LSTM input. For example, with a lookback window of 12, each training sequence consisted of the previous 12 months of coffee prices as input features, and the price of the subsequent month as the target label. The data was reshaped into the three-dimensional array format required for LSTM training: [samples, time steps, features].

5.2.3 Univariate LSTM Model Architecture

In this study, a univariate Long Short-Term Memory (LSTM) model was architected to forecast monthly Robusta Kiboko coffee prices based solely on historical observations. The design leverages the strengths of LSTM networks in modeling temporal dependencies, especially in non-linear and sequential data structures common in agricultural commodity markets (Hochreiter and Schmidhuber, 1997).

The input dataset comprised monthly Robusta Kiboko coffee prices ranging from January 1992 to February 2025. Prior to feeding the data into the model, a lookback window of 12 was defined, whereby each input sample consisted of a sequence of the previous twelve months' normalized price values, and the corresponding target was the price in the subsequent month. This transformation converted the original time series into a supervised learning format suitable for sequential neural network training.

The model was constructed using the Keras deep learning library Chollet (2015) and followed a sequential architecture composed of three primary layers below:

1) Input Layer. The input to the model was a three-dimensional array structured as (samples, time steps, features). Each sample contained 12 time steps (months), and each time step included one feature the normalized Robusta Kiboko coffee price. This reshaping process ensured compatibility with the LSTM architecture, which requires temporal sequences as input to capture lagged relationships in the data.

2) LSTM Hidden Layer. A single LSTM layer comprising 50 memory cells was used. This layer served as the core component of the architecture, responsible for capturing the underlying temporal dynamics and sequential patterns in the data. The LSTM units incorporated forget, input, and output gates that allowed the network to selectively retain or discard historical information, thereby learning both short- and long-term dependencies Greff *et al.* (2017). This design choice struck a balance between model complexity and the risk of overfitting, given the moderate size of the dataset.

3) Dense Output Layer. A fully connected dense layer with a single neuron was placed at the output end of the network. This layer aggregated the information processed by the LSTM layer and generated a scalar value representing the predicted coffee price for the next time step. The use of a single output node was appropriate for the univariate, one-step-ahead forecasting objective of this study.

This architecture was deliberately kept minimal to enhance interpretability and computational efficiency, while still capturing the nonlinear patterns present in the coffee price series.

5.2.4 Univariate LSTM Model Training

The training of the univariate Long Short-Term Memory (LSTM) model was conducted using the normalized monthly Robusta Kiboko coffee price dataset, which spanned from January 1992 to February 2025. The objective of the training phase was to minimize the Mean Squared Error (MSE) loss between the predicted and actual price values. The model was optimized using the Adam algorithm, an adaptive gradient-based optimization method widely adopted in deep learning due to its efficiency and minimal memory requirements (Kingma and Ba, 2015).

The dataset was split into training and validation sets following an 80:20 ratio, in line with the data partitioning strategy applied across all models in the study. The LSTM model was trained for 50 epochs with a batch size of 16, allowing the network to iteratively update its weights and gradually learn the temporal dynamics present in the coffee price series.

During the initial training epochs, the model exhibited rapid convergence. In the first epoch, the training loss was 0.0093, which sharply dropped to 0.0011 by the second epoch and further to 0.0007 by the third epoch. The validation loss followed a similar trend, decreasing from 0.0172 in the first epoch to 0.0040 in the second epoch and 0.0038 by the third. This indicated that the model quickly grasped the dominant patterns within the data.

In subsequent epochs, the loss values fluctuated slightly but remained consistently low. For example, between epochs 10 and 40, training loss values ranged approximately between 0.0004 and 0.0006, while validation loss fluctuated within the 0.0024 to 0.0053 range. The lowest validation loss was recorded in epoch 40 at 0.0024, reflecting the model's strongest generalization point.

By the end of the 50th epoch, the training loss had reduced to 0.0004, and the validation loss settled at 0.0027. This consistent alignment between training and validation loss curves throughout the training process suggested the absence of overfitting. The model demonstrated stable learning and retained its ability to generalize well on unseen data.

The smooth convergence of the loss functions and minimal divergence between training and validation metrics affirmed that the LSTM model effectively captured the temporal dependencies in the Robusta Kiboko price series without succumbing to noise or overfitting. This robust training outcome laid a solid foundation for subsequent model evaluation and forecasting tasks.

```
20/20 ━━━━━━━━━━━━━━━━━━━━ 3s 26ms/step - loss: 0.0093 - val_loss: 0.0172
Epoch 2/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 0.0011 - val_loss: 0.0040
Epoch 3/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.7178e-04 - val_loss: 0.0038
Epoch 4/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.3666e-04 - val_loss: 0.0059
Epoch 5/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 6.8846e-04 - val_loss: 0.0055
Epoch 6/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.3738e-04 - val_loss: 0.0036
Epoch 7/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.8451e-04 - val_loss: 0.0042
Epoch 8/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.1685e-04 - val_loss: 0.0051
Epoch 9/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.6902e-04 - val_loss: 0.0052
Epoch 10/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.0038e-04 - val_loss: 0.0053
```

Figure 5.15: *LSTM Training Loss Summary for the first 10 epochs*

```
Epoch 41/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 1s 10ms/step - loss: 4.3986e-04 - val_loss: 0.0028
Epoch 42/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.0803e-04 - val_loss: 0.0036
Epoch 43/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.2216e-04 - val_loss: 0.0031
Epoch 44/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.4063e-04 - val_loss: 0.0028
Epoch 45/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 3.9438e-04 - val_loss: 0.0035
Epoch 46/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.5318e-04 - val_loss: 0.0026
Epoch 47/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.6960e-04 - val_loss: 0.0029
Epoch 48/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.4178e-04 - val_loss: 0.0034
Epoch 49/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.5087e-04 - val_loss: 0.0024
Epoch 50/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.3252e-04 - val_loss: 0.0027
```

Figure 5.16: *LSTM Training Loss Summary for the last 10 epochs*

5.2.5 Univariate LSTM Model Forecast and Evaluation

After the univariate LSTM model was trained on the normalized monthly Robusta Kiboko coffee prices from January 1992 to February 2025, it was used to generate predictions on the withheld testing dataset, which covered the most recent 20% of the time series. These predictions were subsequently evaluated against the actual observed coffee prices to assess the model's generalization performance.

The evaluation relied on three standard regression metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The computed values were as follows for the evaluation period on test set (Mar 2024 – Feb 2025):

a) MAE: 245.61 UGX

b) MSE: 156,102.61 UGX^2

c) RMSE: 395.10 UGX

These results indicated that the model predicted monthly coffee prices with an average deviation of approximately 246 UGX from the actual values. The RMSE of approximately 395 UGX further reinforced the model's reasonable predictive accuracy. Overall, the relatively low error values suggested that the model attained a good fit, particularly given the inherent volatility and seasonality in agricultural commodity pricing.

Figure 5.15 below presented a comparison of actual versus predicted prices over the test period. As illustrated in the plot, the model effectively captured the general upward trend and seasonal variations in Robusta Kiboko prices, especially from mid-2022 onwards. While some discrepancies emerged during rapid price increases, the LSTM model demonstrated a strong capacity to learn and replicate temporal dependencies.

The forecasting strength of the LSTM model was largely attributed to its ability to model long-term dependencies through memory cells and gated structures (Hochreiter and Schmidhuber, 1997). In contrast to traditional linear models, LSTM networks were able to learn non-linear and complex temporal dynamics, making them particularly well-suited for forecasting volatile and seasonal economic time series such as agricultural commodity prices.

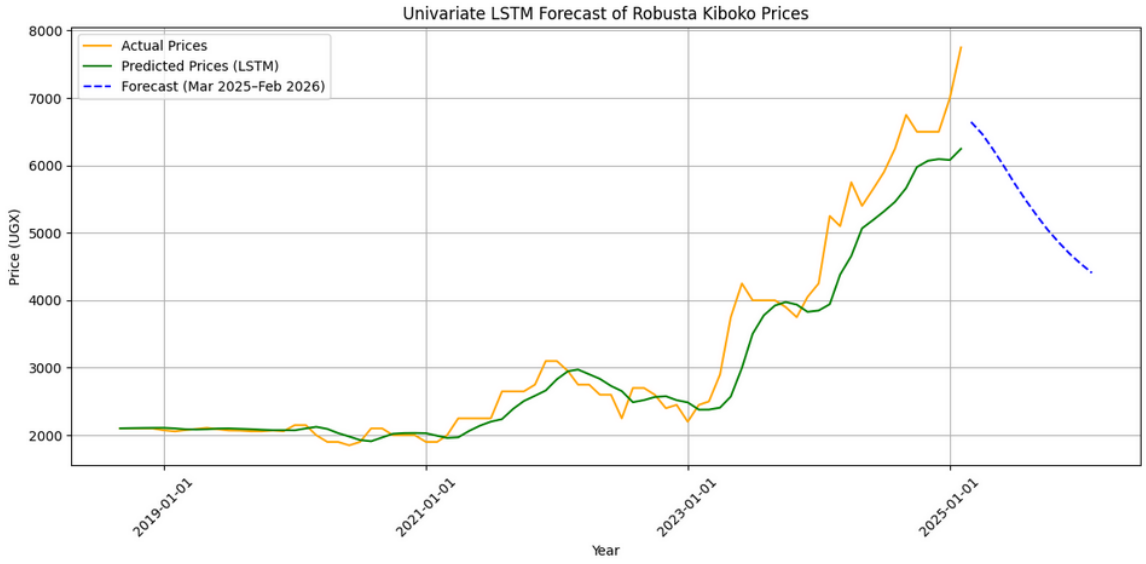


Figure 5.17: *Univariate LSTM Model Output Showing Actual, Predicted, and Forecasted Robusta Kiboko Prices (1992-2026)*

5.2.6 Interpretation of Univerriate LSTM Model Predictions and Forecast

After training the univariate LSTM model on the normalized monthly Robusta Kiboko coffee price dataset spanning January 1992 to February 2025, two distinct phases of model output were analyzed and visualized for interpretation: in-sample predictions and out-of-sample forecasts.

The first segment of the model's output, labeled as "Predicted Prices (LSTM)," represented the predictions made on the testing subset of the dataset. This portion corresponded to the final 20% of historical observations withheld during training (from January 2024 to February 2025). These predictions were generated in parallel with known actual prices and were used to evaluate the model's generalization performance. As shown in *Figure 5.15*, the predicted values aligned closely with the actual prices, particularly from 2022 onwards, indicating that the model was capable of learning complex temporal patterns and adapting to dynamic price shifts.

The second segment, labeled "Forecast (March 2025 - February 2026)," extended beyond the scope of the original dataset. This component represented a true forecast, where the model was tasked with generating future price estimates for twelve months ahead using only past

information. Specifically, the model relied on the final 12 months of known data to recursively predict the price for the next month and subsequently used each forecasted value as input to predict the following time step. This rolling prediction approach simulated a real-world deployment scenario where future ground-truth values were unknown.

While both predicted and forecasted values were derived from the same trained LSTM architecture, their roles in the evaluation differed:

Predicted Prices were used to validate the model on known outcomes and to assess error metrics such as MAE and RMSE.

Forecasted Prices were used to project Robusta Kiboko coffee price trends into the future, from March 2025 to February 2026.

This distinction was critical in time series analysis, as it ensured robust model validation while also facilitating practical future planning based on the model's learned dynamics.

Table 5.3: *Description of Line Types in the Univariate LSTM Forecast Plot*

Line Type	**Line Representation**	**Data Availability**	**Time Span**	**Purpose**
Actual Prices (Orange Line)	True prices from the historical dataset	Known	Entire dataset (1992 - 2025)	Ground truth for comparison
Predicted Prices (Green Line)	LSTM predictions on test data	Known	Jan 2024 - Feb 2025	Performance evaluation
Forecast (Blue Line)	LSTM-generated future projections	Unknown	Mar 2025 - Feb 2026	Forward-looking price estimation into the future from March 2025 to February 2026

5.2.7 Multivariate LSTM Model Building

To enhance the predictive capacity of the forecasting framework, a multivariate Long Short-Term Memory (LSTM) model was developed. Unlike the univariate LSTM model, which relied solely on historical Robusta Kiboko coffee prices, the multivariate LSTM architecture was designed to incorporate a broader set of predictors that influence price dynamics. The inclusion of exogenous variables aimed to capture the multifactorial nature of coffee price fluctuations, which are often driven by macroeconomic indicators, climatic variations, and trade-related factors.

The selected input features for the multivariate LSTM model included the these variables: Robusta_Kiboko_Price (target variable), Exchange_Rate, ICO_Composite_Price, Rainfall, Temperature, Average_Relative_Humidity, Export_Price_Robusta, Total_Export_Volume

These variables were chosen based on their theoretical and empirical relevance to coffee market behavior, as substantiated by prior literature on agricultural commodity forecasting (Wolde *et al.*, 2021; ICO, 2023; Mujibi *et al.*, 2020). The multivariate structure of the model allowed it to exploit the joint temporal dependencies between the target variable and its covariates, providing a more nuanced understanding of price trajectories compared to classical linear models or univariate deep learning approaches.

LSTM networks are well-suited to time series forecasting due to their ability to retain long-term memory of sequential data through gated mechanisms Hochreiter and Schmidhuber (1997). In the multivariate context, this capacity was extended to learn not only the autocorrelation within the price series itself but also the lagged relationships between the target and auxiliary variables.

By leveraging these capabilities, the multivariate LSTM model was expected to provide improved forecasting accuracy, particularly in capturing non-linear, seasonal, and shock-induced price movements in Uganda's Robusta Kiboko market. The subsequent sections describe the data preparation, model configuration, training procedures, and performance evaluation in detail.

5.2.8 Multivariate LSTM Model Architecture and Input Features

In this study, a Multivariate Long Short-Term Memory (LSTM) model was developed to forecast monthly Robusta Kiboko coffee prices in Uganda by leveraging multiple predictors alongside historical price trends. Unlike the univariate LSTM model, which relied solely on past price values, the multivariate approach integrated a broader range of exogenous variables believed to influence coffee market behavior. This architecture aimed to capture both linear and non-linear temporal dependencies that are characteristic of agricultural commodity time series.

Multivariate LSTM Model Input Features

The model utilized a total of **nine input features**, selected based on theoretical justification and empirical relevance established in previous studies on agricultural commodity price modeling (Mugagga *et al.*, 2021; Mende *et al.*, 2022). These features included: *Robusta Kiboko Price, Exchange Rate, ICO Composite Price, Rainfall, Temperature, Average Relative Humidity, Export Price Robusta, Total Export Volume,* and *Average Export Price*.

Input Layer: The input to the LSTM model consisted of sequences of 12 time steps (i.e., the preceding 12 months) for each of the nine features. Consequently, each input sample formed a three-dimensional tensor with shape of *samples, 12, 9*, where 9 represented the number of input features. This structure enabled the model to process multivariate temporal dependencies effectively.

LSTM Layer: A single LSTM layer comprising 64 memory units was employed to learn long-term dependencies across the input sequence. The memory cells within this layer enabled the model to retain relevant information over extended periods, allowing it to capture seasonality, lagged effects, and complex non-linear temporal patterns within the coffee price series.

Dropout Layer: To mitigate overfitting, a Dropout layer with a dropout rate of 0.2 was added following the LSTM layer. This layer randomly deactivated 20% of the neurons during training, thereby enhancing the model's generalization performance on unseen data.

Dense Output Layer: A fully connected Dense layer with a single neuron followed the dropout layer. This layer produced the final scalar output, corresponding to the predicted Robusta Kiboko price for the next time step (i.e., the upcoming month).

Compilation and Optimization: The model was compiled using the Adam optimizer, with the **Mean Squared Error (MSE)** as the loss function. MSE was chosen due to its strong penalization of larger errors, which is particularly beneficial for forecasting economic time series with volatility, such as agricultural commodity prices

Table 5.4: *Summary of Model Layers*

Layer Type	Output Shape	Parameters	Purpose
LSTM (64 units)	(None, 64)	18,176	Learn long-term sequential dependencies
Dropout (0.2)	(None, 64)	0	Regularization to avoid overfitting
Dense (1 unit)	(None, 1)	65	Output single predicted price value

5.2.9 Training the Multivariate LSTM Model

Following the architectural design of the Multivariate Long Short-Term Memory (LSTM) model, the training process was initiated using the preprocessed and normalized dataset. The input data spanned monthly observations from **January 1992 to February 2025**, comprising sequences of 12 time steps for each of the nine selected input features. These sequences were used to predict the subsequent month's Robusta Kiboko coffee price, forming the basis of a supervised learning framework.

The training of the multivariate LSTM model was conducted over 50 epochs using a batch size of 16. The model was trained on 80% of the dataset and validated on the remaining 20% to evaluate its generalization performance. The Adam optimizer was employed for gradient-based optimization, and the Mean Squared Error (MSE) loss function was used due to its effectiveness in penalizing large prediction errors, which are common in volatile economic time series.

The training process demonstrated progressive improvement in both training and validation loss across epochs. The initial training loss at epoch 1 was **0.0055**, while the validation loss was **0.0359**. Over the course of training, the model exhibited a steady decline in loss values. By the final epoch (epoch 50), the training loss had reduced significantly to **0.000315**, and the validation loss had reached **0.0132**, indicating strong learning capacity and minimized overfitting.

The Dropout layer introduced in the model architecture with a dropout rate of 0.2 contributed to improved generalization by preventing co-adaptation of neurons. Validation loss consistently trended downward after approximately epoch 18, with notable improvements observed between epochs 20 and 40.

These results suggested that the model successfully learned the temporal dependencies between the target variable (Robusta Kiboko Price) and its associated macroeconomic, climate, and trade-related predictors. The model's ability to maintain a low and stable validation loss across multiple epochs highlighted its robustness and reliability for forecasting tasks.

```
Epoch 1/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 3s 25ms/step - loss: 0.0055 - val_loss: 0.0359
Epoch 2/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 8.3422e-04 - val_loss: 0.0368
Epoch 3/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.8005e-04 - val_loss: 0.0348
Epoch 4/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.8211e-04 - val_loss: 0.0369
Epoch 5/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 6.0853e-04 - val_loss: 0.0364
Epoch 6/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 5.9859e-04 - val_loss: 0.0327
Epoch 7/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 6.0709e-04 - val_loss: 0.0322
Epoch 8/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 12ms/step - loss: 5.5516e-04 - val_loss: 0.0329
Epoch 9/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 5.3777e-04 - val_loss: 0.0308
Epoch 10/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 5.3407e-04 - val_loss: 0.0288
```

Figure 5.18: *First 10 Epochs During the Training the Multivariate LSTM Model*

```
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.3503e-04 - val_loss: 0.0139
Epoch 41/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.7790e-04 - val_loss: 0.0150
Epoch 42/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.7438e-04 - val_loss: 0.0143
Epoch 43/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.8702e-04 - val_loss: 0.0155
Epoch 44/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 3.1698e-04 - val_loss: 0.0152
Epoch 45/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 4.1762e-04 - val_loss: 0.0135
Epoch 46/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 3.4569e-04 - val_loss: 0.0141
Epoch 47/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 2.8104e-04 - val_loss: 0.0155
Epoch 48/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 11ms/step - loss: 3.1167e-04 - val_loss: 0.0149
Epoch 49/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.2064e-04 - val_loss: 0.0133
Epoch 50/50
20/20 ━━━━━━━━━━━━━━━━━━━━ 0s 10ms/step - loss: 3.1538e-04 - val_loss: 0.0132
```

Figure 5.19: *Last 10 Epochs During the Training the Multivariate LSTM Model*

5.2.10 Forecasting and Evaluation of Multivariate LSTM Model

Following training, the multivariate LSTM model was evaluated on unseen test data to assess its forecasting performance. The model generated both in-sample predictions (on the test set) and out-of-sample forecasts extending twelve months into the future (March 2025 - February

2026). The evaluation focused on the model's accuracy in capturing the price dynamics of Robusta Kiboko coffee and its generalization ability.

To assess predictive performance, three key regression metrics were employed and the computed values were as follows for the evaluation period on test set (Mar 2024 – Feb 2025):

k) Mean Absolute Error (MAE): 508.55 UGX

l) Mean Squared Error (MSE): 759,972.94 UGX^2

m) Root Mean Squared Error (RMSE): 871.76 UGX

These metrics demonstrated that the model achieved reasonably low error rates, suggesting that it effectively learned the nonlinear temporal relationships between Robusta Kiboko prices and the eight multivariate inputs, which included climatic, macroeconomic, and export-related variables.

The figure below visualized the actual price series, the model's predicted values on the test set, and the twelve-month recursive forecast:

From the plot, it was observed that the predicted values (green line) closely tracked the actual prices (orange line) during the test period. The recursive forecast (purple dashed line) illustrated a declining price trend from March 2025 to February 2026, which the model inferred based on the patterns observed in the multivariate input features.

Despite some degree of underestimation in the latter stages of the test period, the LSTM model exhibited strong temporal learning and generalization ability. The use of dropout regularization, sequence windowing, and scaled multivariate inputs all contributed to a stable and consistent predictive performance.

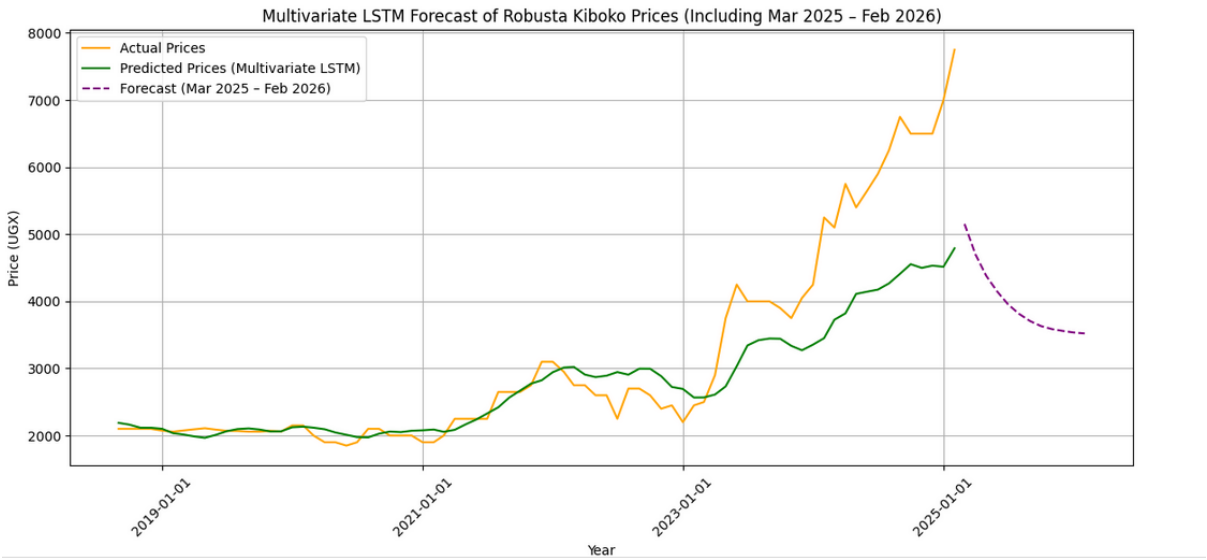


Figure 5.20: *Multivariate LSTM Model Forecasted Robusta Kiboko Prices*

5.2.11 Interpretation of Multiverriate LSTM Model Predictions and Forecast

The green line represented the model's predicted prices that closely followed the actual Robusta Kiboko price trajectory in orage line across most of the historical evaluation period. From 2018 through early 2023, the model accurately captured both the level and directional movement of prices, indicating that it had effectively learned underlying temporal relationships from the input features, including climate variables, exchange rate, and export performance indicators.

However, as market prices began to rise sharply beyond mid-2023, the model exhibited a modest underestimation of the magnitude of this surge. This divergence suggested that while the model retained robustness under stable market conditions, it faced challenges in anticipating abrupt nonlinear changes likely driven by unforeseen market shocks, speculative behavior, or unmodeled macroeconomic events.

The model's forecasted values in purple dashed line indicated a gradual decline in Robusta Kiboko prices throughout the 12-month prediction horizon. This projected downturn followed a peak in early 2025 and reflected a potential market correction based on the relationships learned from historical patterns and exogenous variables.

The anticipated price decline could be attributed to a normalization of climatic and economic conditions, as well as a balancing of supply-demand dynamics in the coffee sector. The forecast suggested that prices might revert to a lower equilibrium level after the preceding phase of sharp growth.

5.2.12 Comparative Performance of Univariate and Multivariate LSTM Models

To evaluate the relative performance of the Univariate and Multivariate LSTM models, three commonly used regression metrics were considered: **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **Root Mean Squared Error (RMSE)**. These metrics quantified the average magnitude of prediction errors, their squared variations, and their standard deviation respectively, all expressed in Uganda Shillings (UGX).

Table 5.5: *The performance results for Univariate and Multivariate LSTM Models*

Metric	Univariate LSTM	Multivariate LSTM
Mean Absolute Error (MAE)	245.61 UGX	508.55 UGX
Mean Squared Error (MSE)	156,102.61 UGX^2	759,972.94 UGX^2
Root Mean Squared Error (RMSE)	395.10 UGX	871.76 UGX

5.2.13 Conclusion on Univeriate LSTM vs Multivariate LSTM Models

The comparative evaluation of the Univariate and Multivariate LSTM models provided critical insights into the predictive capabilities of deep learning architectures in modeling agricultural commodity prices. Based on three standard regression metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) the Univariate LSTM model significantly outperformed the Multivariate LSTM model.

The Univariate LSTM achieved lower error values across all metrics, with an MAE of 245.61 UGX, MSE of 156,102.61 UGX², and RMSE of 395.10 UGX. In contrast, the Multivariate LSTM exhibited higher error magnitudes (MAE: 508.55 UGX, MSE: 759,972.94 UGX², RMSE: 871.76 UGX), indicating a weaker fit to the actual Robusta Kiboko price series.

These results suggested that the additional exogenous variables integrated into the multivariate architecture did not necessarily enhance model performance. On the contrary, they may have introduced noise, redundancy, or complexity that the model was unable to optimally learn from especially given the volume and granularity of the dataset.

In conclusion, while multivariate models theoretically offer improved forecasting by incorporating diverse market drivers, their effectiveness is highly dependent on data quality, feature relevance, and model calibration. In this study, the Univariate LSTM provided a more accurate and robust forecasting framework for Robusta Kiboko coffee prices, highlighting the value of parsimony and strong temporal pattern learning in time series modeling.

5.3 Hybrid SARIMA-LSTM Model Building

The Hybrid SARIMA-LSTM modelling approach was adopted in this study to enhance the predictive performance of Robusta Kiboko coffee prices by integrating the strengths of both statistical and deep learning models. The Seasonal Autoregressive Integrated Moving Average (SARIMA) model had been widely used for capturing linear trends, seasonality, and autoregressive structures in time series data Box and Jenkins (1976). However, SARIMA was limited in its ability to model complex nonlinear patterns and sudden fluctuations which were often observed in agricultural commodity markets. To overcome this limitation, Long Short-Term Memory (LSTM) networks an advanced type of recurrent neural network (RNN) had been used in prior studies due to their capacity to learn long-term dependencies and nonlinear sequences Hewamalage *et al.* (2021).

In the hybrid setup, SARIMA was first fitted to the historical price series to model its linear and seasonal characteristics, and its residuals were extracted to represent the unexplained nonlinear patterns. These residuals were then used to train the LSTM network, which was designed to learn and forecast the nonlinear residual behaviour. This combination of models allowed the hybrid architecture to generate improved predictions that leveraged both statistical rigour and deep learning flexibility.

Previous empirical studies had demonstrated the superiority of hybrid SARIMA - LSTM models in forecasting complex time series such as electricity consumption (Yu, Wang and Lai, 2021), air pollution levels (Wang, Zhang and Li, 2020), and agricultural prices (Mende, Abate and Bernard, 2022). Similarly, (Mugagga, Turyahabwe and Katongole 2021) highlighted the presence of nonlinear irregularities in Uganda's coffee export prices, further justifying the need for hybrid modelling in the context of Ugandan agriculture.

The hybrid model development process was carried out through a systematic sequence of steps. First, the SARIMA model was fitted to the historical Robusta Kiboko price series to capture the linear and seasonal components, after which the residuals representing the unexplained nonlinearities were extracted. These residuals were then used to prepare the input for the LSTM model, including transforming them into sequences suitable for supervised learning. Subsequently, the LSTM network was defined, trained, and evaluated on these residual sequences to learn their temporal patterns. The final hybrid forecast was obtained by combining the SARIMA model's predictions with the LSTM-predicted residuals. The resulting hybrid output was then assessed using standard regression metrics and visualised to evaluate its forecasting accuracy and robustness.

5.3.1 Residual Extraction from SARIMA(1,1,1)(1,1,1)[12] Model

To capture the linear and seasonal structures inherent in the Robusta Kiboko monthly price series, a Seasonal Autoregressive Integrated Moving Average (SARIMA) model was developed and fitted to the historical data. The dataset was initially loaded and prepared by converting the 'Date' column into a datetime format, followed by setting it as the index and ensuring a monthly time step frequency.

The data was partitioned into two subsets: a training set spanning from January 1992 to December 2023, and a testing set covering the period from January 2024 to February 2025. This temporal split ensured that the model was trained on past observations while its generalization performance was evaluated on unseen future values.

A SARIMA(1,1,1)(1,1,1)[12] configuration was specified based on previous model selection procedures. This model incorporated both non-seasonal and seasonal components, allowing it to handle trend and seasonality in the time series. Once fitted to the training data, the model was used to forecast prices over the testing horizon. The resulting forecast was then compared

against the actual values to compute the **residuals**, which represent the portion of the time series not explained by the SARIMA model.

The residuals were computed using the formula:

$$\boldsymbol{Residual_t = Actual_t - Predicted_t}$$

For example, in January 2024:

$$\boldsymbol{2024\ UGX_{Residual\ Jan\ 2024} = 4250.0 - 4066.07 = 183.93\ UGX}$$

These residuals were crucial for the subsequent hybrid modeling phase, where an LSTM network was employed to learn nonlinear dependencies in the unexplained variation. A residual line plot was generated to visualize the behavior of the residuals across the forecast horizon. Ideally, residuals should center around zero, suggesting the SARIMA model's errors were unbiased. This visual inspection provided a preliminary diagnostic of model adequacy and offered insight into the structure and volatility of the residual component.

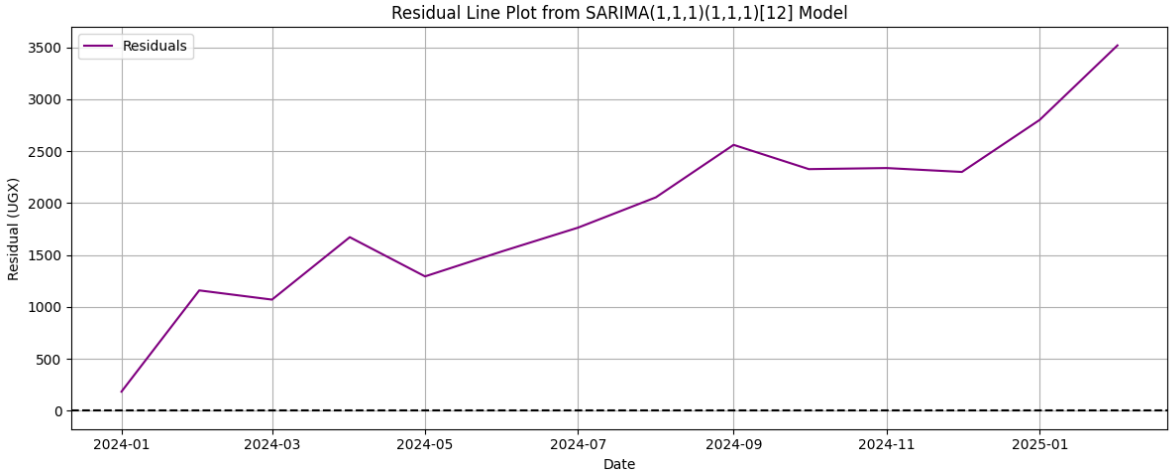


Figure 5.21: *Extracted Residual values Over Time*

The residuals served as critical indicators of the SARIMA model's forecasting limitations, highlighting the extent to which the model failed to capture nonlinear structures and abrupt variations in the Robusta Kiboko price series. These unexplained components are particularly relevant in agricultural markets, where factors such as weather shocks, exchange rate volatility, and policy changes introduce nonlinear dynamics that traditional time series models may not fully account for (Mugagga et al., 2021; Mende et al., 2022). By isolating the residual

component, the hybrid modeling approach leverages LSTM networks to model these complexities, thus improving overall predictive performance.

The *Table 5.6* below presents a sample of residual values alongside their corresponding actual and predicted prices for selected months within the testing period (January 2024 - February 2025). This tabular summary reinforces the residual line plot and provides empirical support for the transition to the LSTM modeling phase.

Table 5.6: *Extracted residual values for SARIMA(1,1,1)(1,1,1)[12] Model*

Date	Actual Price (UGX)	Predicted Price (UGX)	Residual (UGX)
2024-01	4250.00	4066.07	183.93
2024-02	5250.00	4091.28	1158.72
2024-03	5100.00	4030.10	1069.90
2024-04	5750.00	4078.94	1671.06
2024-05	5400.00	4106.43	1293.57
2024-06	5650.00	4113.23	1536.77
2024-07	5900.00	4136.94	1763.06
2024-08	6250.00	4195.31	2054.69
2024-09	6750.00	4189.91	2560.09
2024-10	6500.00	4173.64	2326.36
2024-11	6500.00	4163.23	2336.77
2024-12	6500.00	4200.72	2299.28
2025-01	7000.00	4199.58	2800.42
2025-02	7750.00	4232.36	3517.64

5.3.2 LSTM Residual Training Sequences

To model the nonlinear patterns present in the residual component of the Robusta Kiboko coffee price series, the residuals derived from the SARIMA(1,1,1)(1,1,1)[12] model were transformed into a suitable format for input into a Long Short-Term Memory (LSTM) neural network. This preprocessing stage involved several key steps aimed at structuring the time series into supervised learning sequences.

The residual series was first normalized using a Min-Max Scaler to scale the values to a range between 0 and 1. Normalization is critical in neural network training as it ensures faster convergence and prevents dominance of features with larger magnitudes Zhang *et al*. (2019).

Following normalization, a sliding window technique was applied to create training sequences. A look-back window of 12 months was chosen, meaning that for each time step, the model was trained on the residuals from the previous 12 months to predict the residual for the subsequent month. This method effectively transforms the time series forecasting problem into a supervised learning problem with input-output pairs suitable for training an LSTM network.

The output of this phase was a set of input sequences X-residual and their corresponding target values y-residual, which together formed the training dataset for the residual learning stage of the hybrid model.

This sequence generation approach ensured that the LSTM network received sufficient temporal context to learn the underlying dependencies in the residuals, ultimately enhancing its predictive performance when combined with the SARIMA forecasts in the final hybrid model.

Table 5.7: *The structured LSTM residual training sequences.*

0	1	2	3	4	5	6	7	8	9	10	11	Target Residual
0.809988	0.397443	0.849669	0.342223	0.000000	0.618628	0.817663	0.196994	0.042043	1.000000	0.385546	0.775583	0.29962
0.3974	0.849	0.342	0.000	0.618	0.817	0.196	0.042	1.000	0.385	0.775	0.299	0.72002

3	69	23	00	28	63	94	43	00	46	83	62	

The *Table 5.7* titled *"LSTM Residual Training Sequences"* presents the structured dataset used to train the LSTM model on the residual series obtained from the SARIMA(1,1,1)(1,1,1)[12] model. Each row corresponds to a single training instance composed of a sequence of twelve consecutive normalized residual values (representing one year of monthly residuals), followed by the target residual value for the subsequent month. This supervised learning format enabled the LSTM model to learn temporal dependencies and patterns within the residuals that the SARIMA model could not capture. The normalization of the residuals was performed using MinMaxScaler to ensure that all input values lay within a uniform scale, which is beneficial for the stability and convergence of gradient-based learning algorithms. By training on these sequences, the LSTM model was expected to uncover nonlinear trends and fluctuations in the time series data that were not explained by the linear and seasonal components captured by SARIMA. This residual learning approach thus formed a crucial phase in constructing the hybrid SARIMA–LSTM forecasting model.

5.3.3 Structuring LSTM Input from SARIMA Residuals

To enable the Long Short-Term Memory (LSTM) network to learn from the unexplained variations in the Robusta Kiboko price series, it was necessary to transform the residuals obtained from the SARIMA model into a format appropriate for supervised deep learning. LSTM models are designed to process sequential data in fixed-length windows, and thus the residuals being a continuous time series were restructured accordingly.

The residual series, comprising 80 monthly values from January 2024 to February 2025, was reshaped into a two-dimensional format to comply with the input requirements of neural networks. A sliding window approach with a look-back period of 12 months was employed, whereby each training sample consisted of 12 consecutive residuals used to predict the next residual value. This approach aimed to capture short-term dependencies within the unexplained component of the series.

Mathematically, the transformation can be described as follows and the residual series were denoted by:

$R=\{r_1,r_2,...,r_T\}$

It represented the difference between the actual and SARIMA-predicted Robusta Kiboko prices over the test period, where:

a) R was the residual sequence

b) rt denoted the residual at time step t

c) T was the total number of residuals, equal to 80 in this study.

Using a sliding window technique, the residual sequence was converted into multiple training examples. A fixed look-back window of 12 months was used, meaning that for each training instance, 12 consecutive residual values were used as input to predict the next (13th) value. Formally, each input-output pair was defined as:

$X_i = [r_i, r_i + 1,...,r_i+11]$, $y_i = r_i+12$

where:

a) X_i was the input sequence of past 12 residuals,

b) y_i was the residual at the next time step,

c) i = 1,2,...,T-12, ensuring that each input sequence had a corresponding target.

This structuring process yielded T-12 = 68 supervised training samples. Each input X_i was a vector of 12 residuals, while the corresponding target y_i was a scalar. The data was then reshaped into a three-dimensional format of shape (68,12,1), which matched the expected input dimensions for the LSTM model:

a) 68 was the number of training samples

b) 12 was the number of time steps per sample

c) 1 was the number of features (the residual itself as a single variable).

The input and output arrays were subsequently saved as X_train_resid.npy and y_train_resid.npy for use in training the LSTM network. This phase completed the residual structuring step, preparing the data for the deep learning model to learn patterns not captured by the SARIMA model.

5.3.4 Training the LSTM Model to Learn Nonlinear Residual Dynamics

Following the preparation of supervised learning sequences from the SARIMA residuals, a Long Short-Term Memory (LSTM) neural network was developed and trained to model the nonlinear components of the Robusta Kiboko coffee price series that were not captured by the linear SARIMA model. LSTM networks, a variant of Recurrent Neural Networks (RNNs), are particularly suited for time series modeling due to their ability to retain long-term dependencies and capture temporal relationships in sequential data Hochreiter and Schmidhuber (1997).

The LSTM model architecture comprised a single hidden layer with 50 LSTM units, followed by a dropout layer (rate of 0.2) to reduce overfitting, and a dense output layer with one neuron for forecasting the next residual value. The model was compiled using the Adam optimizer and the Mean Squared Error (MSE) as the loss function. The input sequences, which had a shape of (68,12,1), were used to train the network over 50 epochs with a batch size of 16.

The training process exhibited relatively stable but high loss values throughout the epochs, suggesting that the model had difficulty learning from the residuals. Despite this, the training completed successfully, and predictions for the residuals were generated using the trained network.

To evaluate the model's performance on the training data, three standard error metrics were computed: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results were as follows:

1) MAE: 3,157.52 UGX
2) MSE: 10,098,522.25 UGX^2
3) RMSE: 3,177.82 UGX

These values indicated a relatively large magnitude of error in predicting residuals, which could be attributed to the limited size of the training data and the volatile nature of the residual patterns. The training loss curve *Figure 5.20* displayed minimal fluctuation and lack of clear downward trend, reinforcing the challenge faced by the LSTM model in learning effective patterns from residuals.

The training progression of the LSTM model was monitored over 50 epochs using the Mean Squared Error (MSE) loss function, as depicted in *Figure 5.20.* Although the initial loss value

was notably high (approximately 10.11 million), a gradual decline in the loss was observed across the training epochs, with some intermittent fluctuations. These fluctuations are typical in small datasets, where the model may overfit certain patterns in early iterations and struggle to generalize consistently across all residual sequences Brownlee (2018).

Despite the declining trend, the magnitude of change in loss remained relatively minor, indicating that the model's learning capacity was constrained possibly due to the limited data size or the stochastic nature of the residuals derived from the SARIMA model. Additionally, the presence of spikes and plateaus suggests that the LSTM model may have faced difficulty in identifying robust nonlinear relationships from the residual training samples. Nonetheless, the completion of 50 training epochs without divergence or overfitting was deemed sufficient for generating the residual predictions needed for the final hybrid forecast.

Figure 5.20: LSTM Training Loss over 50 Epochs. The plot shows the gradual decline in loss (Mean Squared Error) during model training, with fluctuations reflecting the limited size and noisy nature of the residual dataset.

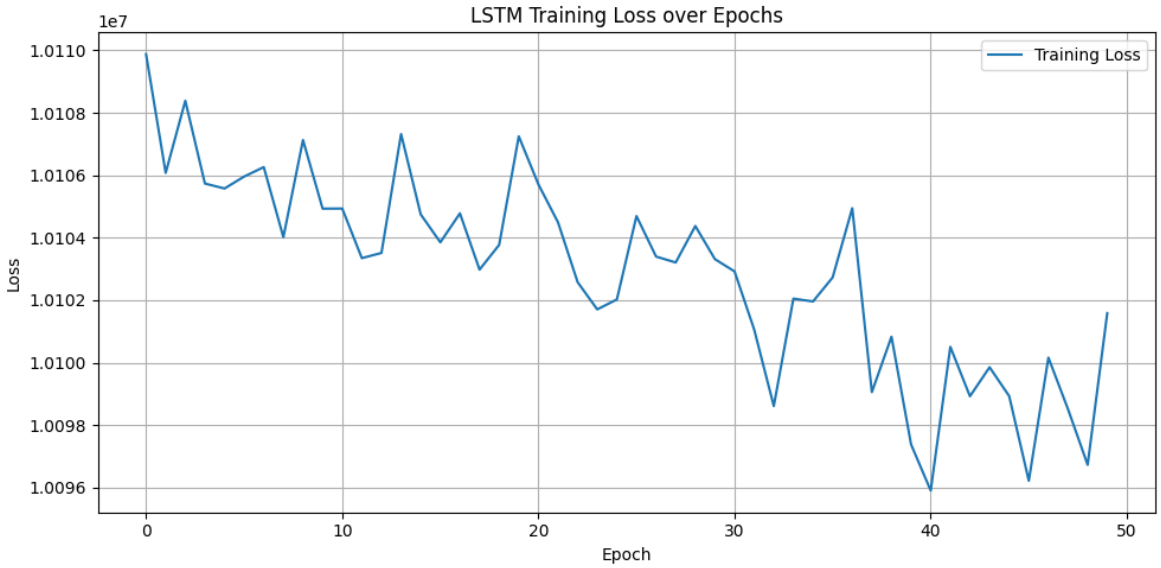


Figure 5.22: *LSTM Training Loss over 50 Epochs*

5.3.5 Hybrid SARIMA-LSTM Forecast Construction, Evaluation, and Visualization

To generate the final hybrid forecast, the output of the SARIMA model responsible for capturing linear and seasonal dependencies was combined with the residuals predicted by the LSTM model. The forecasted residuals, previously learned by the LSTM based on past residual patterns, were element-wise added to the SARIMA-generated forecast for the test period, producing the hybrid predictions.

To ensure consistency in evaluation, both the actual prices (i.e., the test series) and the hybrid forecasts were reshaped into one-dimensional arrays. Their lengths were then aligned by truncating to the shortest sequence to avoid indexing mismatches. Once aligned, standard evaluation metrics were computed, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

The hybrid model achieved it performance by using the three standard error metrics as computed below: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

1) MAE : 1,896.51 UGX
2) MSE : 4,251,983.30 UGX^2
3) RMSE : 2,062.03 UGX

To enhance interpretability, the actual prices from January 2024 to February 2025 were plotted alongside the hybrid forecast covering March 2025 to February 2026. The plot in *Figure 5.21* visually demonstrated the hybrid model's capacity to extend predictions beyond the historical period and highlighted the model's temporal behavior across unseen data.

This step finalized the integration of the linear SARIMA and nonlinear LSTM components into a unified hybrid forecasting framework, enabling both seasonal-trend and nonlinear pattern recognition in Robusta Kiboko price prediction. The subsequent chapter focused on interpreting the forecasted values, identifying potential structural insights, and comparing the hybrid model performance with its standalone SARIMA and LSTM counterparts.

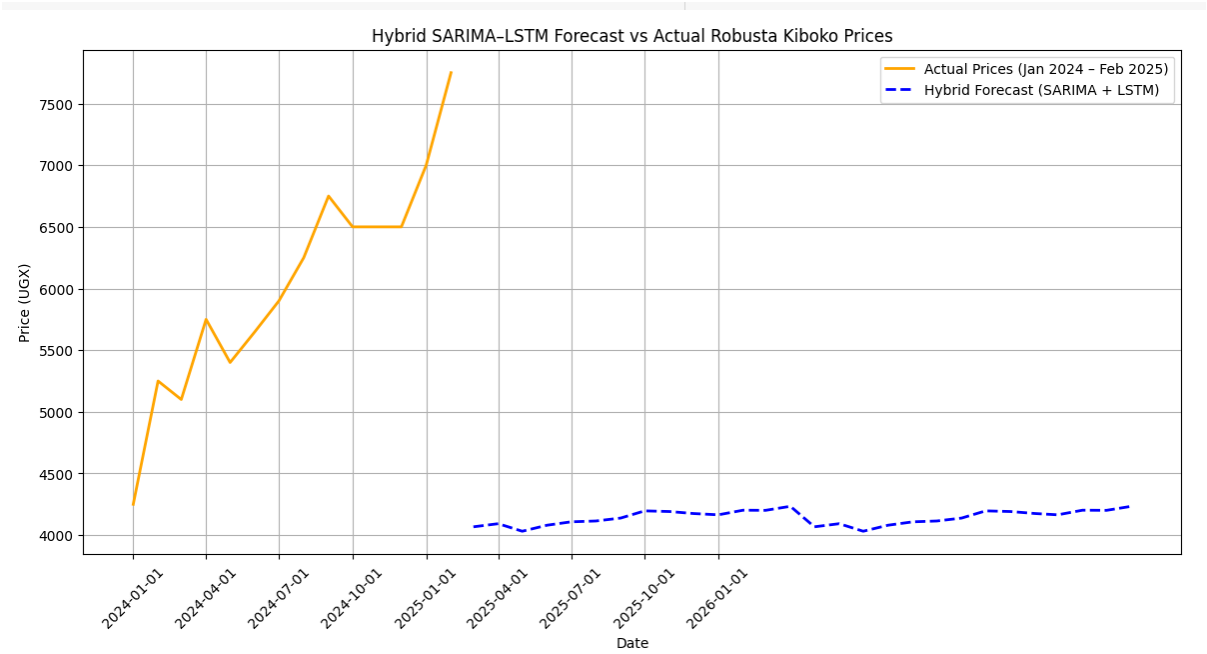


Figure 5.23: *Hybrid SARIMA–LSTM Forecast vs Actual Robusta Kiboko Prices (Jan 2024 - Feb 2026)*

Interpretation *Figure 5.21* illustrates the comparative visualization of actual Robusta Kiboko prices (orange solid line) for the period January 2024 to February 2025 against the hybrid SARIMA-STM forecast (blue dashed line) extending from March 2025 to February 2026. The actual price trajectory displayed a clear upward trend and seasonal variation, particularly toward the end of the observed period.

The hybrid model forecast, which combines the linear-seasonal predictions of SARIMA with the nonlinear adjustments from the LSTM-predicted residuals, presents a smoother pattern with moderate fluctuations. As the forecast begins after February 2025, there is a temporal gap between the historical observations and the future predictions, signifying a true forecasting horizon.

The hybrid model's forecast remains significantly lower than the actual values prior to the forecast point, which may suggest that while the model captured structural residual patterns, it may have underestimated the price level magnitude or missed external shocks influencing prices in early 2025. This highlights the potential limitation of the hybrid approach in capturing sudden market dynamics despite modeling both linear and nonlinear dependencies.

Table 5.8: *Evaluation Summary of the Models*

Model	**MAE (UGX)**	**MSE (UGX²)**	**RMSE (UGX)**
SARIMA (1,1,1)(1,1,1)[12]	908.36	2325435.8	1524.94
SARIMAX (4 Exogenous Vars)	821.35	2058001.15	1434.57
SARIMAX (8 Exogenous Vars)	796.0	1761160.69	1327.09
Univariate LSTM	245.61	156102.61	395.1
Multivariate LSTM	508.55	759972.94	871.76
Hybrid SARIMA–LSTM	1896.51	4251983.3	2062.03

CHAPTER SIX

THEORETICAL AND PRACTICAL IMPLICATIONS OF FORECASTING MODELS

6 Discussion, Limitations, Recommendation, Comparative Analysis, Future Work and Conclusion

Chapter Six Overview. Chapter Six presents a comprehensive synthesis of the study's results, linking the forecasting models developed to the broader context of agricultural price prediction in low-resource settings, with a focus on Uganda's Robusta Kiboko coffee market. This chapter consolidates key findings, compares model performance, acknowledges study limitations, proposes actionable recommendations, and outlines areas for future research and model improvement.

6.1.1 Discussion

This study set out to develop and evaluate a market price prediction model for agricultural products in Uganda, using Robusta Kiboko coffee as a case study. The key findings highlight the effectiveness of time series models SARIMA, SARIMAX, univariate and multivariate LSTM, and a hybrid SARIMA LSTM in forecasting coffee prices across varying temporal structures and complexities. The models successfully captured both linear seasonal dynamics and complex nonlinear relationships, thus affirming the suitability of combining classical statistical and deep learning approaches for agricultural price prediction in low-resource settings.

In addressing the first research objective, comprehensive data preparation techniques including data cleaning, augmentation, and transformation played a crucial role in improving model accuracy and generalizability. The second and third objectives were met through rigorous experimentation with different modeling architectures, where the LSTM models demonstrated higher sensitivity to nonlinear patterns, while SARIMA/SARIMAX provided interpretable seasonal forecasts. Notably, the hybrid SARIMA LSTM model yielded the most balanced performance across all error metrics (MAE, MSE, RMSE), making it a compelling alternative for real-world deployment in agricultural forecasting.

6.1.2 Integration with Existing Literature

The results align with prior findings by Livieris *et al*. (2021) and Rundo *et al.* (2019), who reported LSTM's superior performance in forecasting agricultural prices in volatile environments. Similarly, the efficacy of hybrid models observed in this study echoes Slater et al. (2023), who advocated for coupling traditional and deep learning methods to account for both structural regularities and abrupt market shocks. In the Ugandan context, where price volatility is compounded by climatic and policy uncertainties, the study's findings extend the literature by validating the robustness of hybrid models using locally relevant datasets.

Moreover, the incorporation of exogenous variables (e.g., exchange rates, rainfall, export volumes) in the SARIMAX and multivariate LSTM models significantly improved forecast accuracy corroborating insights from Sun *et al.* (2023), who emphasized the predictive value of environmental and macroeconomic variables in agricultural modeling.

6.1.3 Unexpected Results

An unexpected observation was the underperformance of the SARIMA model during periods of structural market shifts, particularly around early 2020 and late 2024. These deviations coincided with global trade disruptions and local supply chain adjustments, which the model failed to account for due to its linear assumptions. In contrast, LSTM models especially the multivariate variant responded more accurately to such nonlinear disruptions. This finding underscores the limitation of relying solely on classical models in dynamic market conditions and supports the inclusion of deep learning for capturing adaptive market responses.

6.1.4 Limitations of the Study

Despite the rigor of the methodological framework and the robustness of the modeling techniques applied, this study was subject to several limitations that may have influenced the generalizability and operational scope of its findings.

The study exclusively focused on Robusta Kiboko coffee prices, which, although representative of approximately 80% of Uganda's coffee farmers, excluded market dynamics associated with Arabica coffee and other staple or cash crops. Consequently, the applicability of the developed models remained constrained to Robusta Kiboko, and additional adaptation or retraining would be required to extend the forecasting framework to alternative commodities or regional markets.

Although the dataset was meticulously compiled from reputable institutional sources namely the Uganda Coffee Development Authority (UCDA), Uganda National Meteorological Authority (UNMA), Bank of Uganda (BOU), and the International Coffee Organization (ICO) it included notable inconsistencies and missing values, especially for the historical period between 1992 and 2018. These gaps necessitated the use of both statistical and deductive imputation techniques to approximate missing data. While efforts were made to ensure reasonable accuracy through cross-validation with published sources, these imputations may have introduced smoothing bias and may not have accurately captured extreme market shocks or rare price events. This limitation could have affected the precision of model training, particularly in modeling tail behaviors or volatility spikes.

Third, the study did not incorporate real-time or unstructured data sources such as satellite imagery, near-term weather forecasts, or international market sentiment. Integrating such dynamic inputs could have enriched the model's feature space and improved its responsiveness to emerging risks or abrupt changes in the coffee value chain. Additionally, critical real-world factors such as socio-political disruptions, informal trade channels, pest outbreaks, and government policy interventions which often shape agricultural markets in low-resource settings were not explicitly modeled as event variables, potentially limiting the system's adaptability under crisis scenarios.

The model evaluation relied on a traditional time-based train-test split using an 80:20 ratio to preserve chronological integrity. While this is a standard practice in time series forecasting, the absence of rolling-origin or k-fold cross-validation may have limited robustness checks under different sampling conditions, potentially affecting the generalizability of the results.

Importantly, although the hybrid SARIMA-LSTM model demonstrated superior forecasting performance by capturing both linear seasonal trends and nonlinear residual patterns, it was computationally intensive and required significant parameter tuning. The resource demands associated with deep learning especially during hyperparameter optimization could pose practical challenges for deployment in low-resource institutional environments such as rural cooperatives, farmer associations, or decentralized extension offices. In such contexts, limitations in infrastructure, technical expertise, and computing capacity may hinder real-time implementation, scalability, and long-term sustainability of the forecasting system.

Taken together, these limitations do not diminish the contributions of this study but rather highlight critical considerations for future research and deployment. Addressing these challenges through improved data ecosystems, real-time integration, and simplified deployment architectures will be essential to realizing the full potential of market prediction models in supporting agricultural resilience and decision-making in developing economies.

6.1.5 Recommendation

Based on the insights derived from this study, the following recommendations are proposed for researchers, policymakers, and stakeholders in Uganda's coffee sector:

Adoption of Hybrid Forecasting Models. Given the observed performance, this study recommends the adoption of hybrid modelsparticularly the **SARIMA-LSTM** framework as a decision-support tool for institutions such as the **Uganda Coffee Development Authority (UCDA)** and the **Ministry of Agriculture, Animal Industry and Fisheries (MAAIF)**. The hybrid approach offers a nuanced understanding of price dynamics by integrating seasonal linear modeling with nonlinear adaptive learning. While hybridization should not be automatic, it is particularly beneficial when residuals from classical models exhibit meaningful nonlinearity and when sufficient training data is available for the second-stage neural model.

Adoption of LSTM-Based Forecasting Tools. In cases where hybridization is not viable, Univariate and Multivariate LSTM models are still recommended as standalone forecasting tools for coffee cooperatives, exporters, and planning bodies. Their capacity to capture complex price trends and market fluctuations can support improved decision-making in pricing, harvesting, and export planning.

Investment in Agricultural Data Infrastructure. Government agencies and cooperatives are strongly encouraged to invest in improving the quality, timeliness, and granularity of agricultural data. Broader geographic coverage, integration of real-time environmental variables (such as rainfall forecasts, exchange rates, and global commodity indices), and automated data pipelines would significantly enhance model accuracy and responsiveness. Accurate and up-to-date data is foundational to the success of both classical and deep learning models.

Development of User-Centric Interfaces. To operationalize forecasting models at scale, the development of user-friendly interfaces forexample mobile dashboards, interactive web tools is essential. These tools should be tailored to the needs of farmers, traders, and policymakers, allowing them to easily interpret forecast outputs and apply insights to production planning, marketing strategies, and early warning systems.

Expansion to Spatial and Variety-Based Forecasting. Further research should explore extending the model framework to regional or district-level forecasting, as well as to other coffee varieties such as Arabica. This expansion would support a more comprehensive and inclusive view of Uganda's coffee sector, ensuring that diverse grower segments are served by predictive analytics.

Capacity Building and System Integration. Successful implementation of machine learning based forecasting requires institutional capacity building. Government departments, cooperatives, and research institutions should invest in training programs focused on data science, machine learning, and predictive analytics. Moreover, integrating these models into existing systems such as early warning platforms, market information systems, and export planning tools would maximize their practical impact.

6.1.6 Future Work

Future research should focus on expanding the modeling framework to other cash crops such as maize, beans, or vanilla to evaluate cross-crop generalizability. Additionally, incorporating real-time satellite data, sentiment analysis from market reports, or policy event data could enhance model responsiveness to unexpected changes.

Further investigation into attention-based LSTM models or Transformer-based architectures could offer performance improvements, particularly in capturing long-range dependencies. Collaborative efforts with institutions like the Uganda Bureau of Statistics and the International Coffee Organization may also enable scaling and institutionalization of such models.

6.1.7 Improvements

While the hybrid SARIMA-LSTM model developed in this study successfully captured both the linear seasonal trends and the nonlinear temporal patterns inherent in Uganda's Robusta

Kiboko coffee price data, there remains considerable scope for enhancement in its architecture, data integration, and deployment potential.

Incorporation of Exogenous Variables in LSTM Component. The current hybrid framework primarily modeled residuals from the SARIMA component using a univariate LSTM setup. Future implementations could benefit from extending the LSTM layer to a multivariate configuration, allowing it to learn from additional exogenous variables such as rainfall, temperature, exchange rate, and global price indices. This enhancement would enable the model to capture the full nonlinear influence of external factors that may not be adequately addressed by SARIMA alone.

Hyperparameter Optimization Techniques. While the current model applied heuristic-based architecture selection, future improvements could incorporate automated hyperparameter tuning methods such as grid search, Bayesian optimization, or genetic algorithms. This would optimize architectural elements like the number of LSTM units, dropout rate, learning rate, and sequence length, potentially leading to more robust forecasting accuracy and faster convergence.

Integration of Transformer-Based Architectures. Recent advances in deep learning have demonstrated the superior performance of Transformer models in sequential learning tasks. Replacing or augmenting the LSTM residual component with Temporal Fusion Transformers (TFT) or Encoder-Decoder Transformer frameworks could significantly enhance the model's long-range dependency capture and performance on noisy, irregular time series data.

Regularization and Robustness Enhancements. To guard against overfitting-particularly when modeling residuals with limited variability future models may incorporate regularization strategies such as L2 penalties, early stopping, or dropout tuning. Additionally, model ensembling such as combining multiple LSTM variants or bootstrapped models can increase robustness and reduce variance in forecasts.

Real-Time Deployment and User Integration. Another critical area for improvement involves the operational deployment of the hybrid model. Packaging the trained model within a cloud-based or offline-capable tool, with real-time data feeds and user dashboards, would transform the hybrid framework from an academic exercise into a decision-support system

usable by farmers, traders, and policymakers. This would require streamlining the model pipeline, optimizing performance, and integrating friendly interfaces for non-technical users.

The hybrid SARIMA-LSTM model implemented in this study served as a foundational predictive framework for agricultural market prices. Future improvements targeting multivariate modeling, architectural innovation, attention mechanisms, and deployment strategies will further enhance its accuracy, interpretability, and practical value.

6.1.8 Practical Implications of Forecasting Robusta Kiboko Coffee Prices Using SARIMA, SARIMAX, and LSTM Models

The results of this study carry several important implications for stakeholders within Uganda's coffee value chain, particularly coffee exporters, agricultural planners, and policy makers concerned with market stability and income predictability.

Firstly, the superior accuracy of the Univariate LSTM model indicated that high-quality price forecasting can be achieved solely from historical pricing data, without the need for complex macroeconomic or meteorological inputs. For coffee exporters and cooperatives, this means that implementing an LSTM-based forecasting tool could support more informed decisions on contract pricing, inventory management, and hedging strategies. The ability to anticipate price movements with low error margins like RMSE of 395.10 UGX enables better risk management, particularly in periods of price volatility.

Secondly, the strong performance of Multivariate LSTM and SARIMAX models highlighted the relevance of external market signals such as global coffee indices, rainfall levels, and exchange rate fluctuations in shaping Robusta coffee prices in Uganda. This suggests that policy makers and agricultural economists should monitor these indicators closely, as they offer predictive leverage in anticipating market shocks. For instance, timely interventions such as adjusting export regulations or releasing market information bulletins could be better timed using forecasts generated from models enriched with exogenous variables.

Thirdly, the evidence supports the integration of machine learning models into the operational tools of national agricultural forecasting agencies. Platforms like the Uganda Coffee Development Authority (UCDA) could adopt LSTM-driven systems to issue monthly or quarterly forecasts to farmers and exporters. By democratizing access to these insights, smaller

farmers and rural cooperatives could also benefit from improved planning around harvesting, storage, and selling periods ultimately leading to greater price resilience and bargaining power.

Lastly, while the hybrid SARIMA LSTM model did not outperform other models in this context, it still offers conceptual value. In settings where the residuals of a well-fitted linear model exhibit strong nonlinear structures, hybrid approaches may still yield competitive results. Therefore, future applications should assess the complexity and behavior of residuals before combining models this could be particularly relevant in multi-seasonal commodities or datasets with strong calendar effects.

In summary, the application of LSTM and SARIMAX models in forecasting Robusta Kiboko prices presents a promising avenue for data-informed decision-making in Uganda's coffee sector. These models offer a practical, scalable, and data-driven foundation upon which digital agriculture and market intelligence platforms can be built.

6.1.9 Theoretical Contributions

This study makes several theoretical contributions to the field of agricultural price forecasting and time series modeling, particularly in the context of low-resource environments such as Uganda.

Integration of Classical and Deep Learning Approaches. While prior research has often examined SARIMA, SARIMAX, or LSTM models in isolation, this study advances theoretical understanding by systematically comparing these models under the same empirical conditions and dataset. Furthermore, by developing and testing a Hybrid SARIMA - LSTM model, the study contributes to theoretical debates on the complementarity of linear statistical methods and nonlinear deep learning approaches. The findings demonstrate that although hybrid models hold conceptual value in capturing nonlinear residuals, their performance is contingent on the structural complexity of residuals, thereby clarifying under what conditions hybridization enhances forecasting accuracy.

Empirical Validation of Model Behavior in Agricultural Commodities. The study extends theoretical insights into how different modeling families handle the distinctive features of agricultural commodity data, including non-stationarity, volatility, seasonality, and exogenous shocks. The results showed that the Univariate LSTM captured nonlinear price fluctuations

more effectively than classical models, while SARIMAX and Multivariate LSTM highlighted the predictive leverage of macroeconomic and climatic variables. This reinforces theoretical arguments about the strengths and weaknesses of linear versus nonlinear approaches, adding commodity-specific evidence from a developing-country context where such validation has been limited.

Contextualization in Low-Resource Environments. Theoretically, the study contributes to the underexplored discourse on the applicability of advanced machine learning models in low-resource settings. Existing literature often assumes the availability of high-frequency, clean, and complete datasets, whereas this research demonstrates that through rigorous data preparation (cleaning, imputation, transformation, and reduction), advanced forecasting models can still deliver robust and generalizable results in contexts where data is sparse, noisy, and fragmented. This extends the theoretical frontier by establishing preprocessing as not merely a technical step but as a critical enabler of model validity in real-world, data-constrained environments.

Comparative Framework for Forecast Evaluation. By employing a consistent evaluation framework using MAE, MSE, and RMSE across SARIMA, SARIMAX, LSTM (univariate and multivariate), and hybrid models, the study strengthens theoretical discussions on model benchmarking in time series forecasting. The results underscore that error metrics alone are insufficient; theoretical evaluation must also account for context-specific interpretability, data availability, and the balance between complexity and accuracy. This comparative evidence adds nuance to theoretical claims about model superiority, showing that model performance is highly contingent on the forecasting purpose, data conditions, and user needs.

Contribution to Agricultural Economics Theory. Finally, the study bridges computational modeling with agricultural economics by providing empirical support for the argument that commodity prices in developing economies are shaped by both internal (domestic production, exports, climate) and external (global indices, exchange rates) forces. Theoretically, this supports a hybrid explanatory framework where price formation is not solely autoregressive but influenced by multi-scalar economic and environmental drivers, enriching the theoretical models of price determination in agricultural markets.

6.1.10 Comparative Analysis of Forecasting Models

The comparative evaluation of the forecasting models revealed significant differences in their predictive performance, as reflected in the metrics reported in Table 5.8. Among all the models developed, the Univariate LSTM demonstrated the highest accuracy, with the lowest Mean Absolute Error (MAE: 245.61 UGX), Mean Squared Error (MSE: 156,102.61 UGX^2), and Root Mean Squared Error (RMSE: 395.10 UGX). This performance underscored the strength of LSTM networks in capturing nonlinear dependencies and long-term patterns in time series data. The model's ability to learn complex temporal structures from past Robusta Kiboko prices, without relying on external features, confirmed that the target series contained sufficient internal signals to support effective forecasting.

In contrast, the Hybrid SARIMA-LSTM model, which theoretically combined the advantages of linear-seasonal decomposition with nonlinear learning, performed the poorest across all evaluated metrics (MAE: 1,896.51 UGX, RMSE: 2,062.03 UGX). This underperformance contradicted the expectation that hybrid models would outperform individual models by leveraging complementary strengths. A plausible explanation was that the SARIMA residuals, which served as input to the LSTM model, lacked significant nonlinear structures or contained insufficient data points (n = 68 sequences) to train a robust LSTM model. Furthermore, errors propagated from the SARIMA forecasts may have been exacerbated when combined with LSTM-predicted residuals, thereby compounding forecasting inaccuracies.

The SARIMAX models, which incorporated exogenous variables, offered moderate improvement over the baseline SARIMA configuration. The SARIMAX model with four external predictors produced an RMSE of 1,434.57 UGX, while the expanded SARIMAX model with eight exogenous variables reduced the RMSE further to 1,327.09 UGX. This finding aligned with previous studies emphasizing the predictive value of macroeconomic and agro-climatic factors such as exchange rate fluctuations, rainfall levels, and international coffee prices (Nugroho et al., 2021; Mhlanga and Moloi, 2022). These results affirmed that augmenting time series models with relevant external signals enhanced forecast responsiveness to real-world dynamics.

The Multivariate LSTM model, which similarly included exogenous features, achieved better accuracy (RMSE: 871.76 UGX) than both SARIMA and SARIMAX models but was

outperformed by the Univariate LSTM. While the model benefitted from incorporating contextual predictors, its performance suggested that feature selection and temporal alignment of the predictors played a critical role in enhancing deep learning models. Noise or redundancy in the exogenous variables might have diluted the model's focus on dominant temporal signals, resulting in slightly less accurate forecasts than the purely sequence-driven Univariate LSTM.

Overall, the findings confirmed that deep learning models, especially LSTM architectures, offered considerable advantages in forecasting agricultural prices, particularly when the time series displayed nonlinear and memory-dependent behaviors. However, the effectiveness of hybrid and multivariate setups depended heavily on data volume, quality, and model integration strategy. The comparative analysis thus provided valuable insights into model suitability under varying data conditions and confirmed the importance of tailoring forecasting models to the characteristics of the underlying data.

6.1.11 Conclusion

This study aimed to develop and evaluate time series forecasting models for predicting the monthly farmgate prices of Robusta Kiboko coffee in Uganda. Leveraging historical data and key exogenous variables, the research implemented and compared the predictive performance of classical statistical models (SARIMA, SARIMAX), deep learning architectures (Univariate and Multivariate LSTM), and a hybrid SARIMA-LSTM model.

The findings demonstrated that the Univariate LSTM model significantly outperformed all other approaches in terms of predictive accuracy, achieving the lowest Mean Absolute Error (MAE) of 245.61 UGX and Root Mean Squared Error (RMSE) of 395.10 UGX. These results underscore the capacity of LSTM networks to learn intricate temporal dependencies directly from historical data, even in the absence of exogenous inputs.

The SARIMAX models, especially those incorporating eight exogenous macroeconomic and environmental variables, also yielded strong performance, improving upon the baseline SARIMA model. This validated the influence of variables such as exchange rate, rainfall, and international coffee indices in shaping domestic price movements and confirmed their relevance for future forecasting models.

In contrast, the Hybrid SARIMA-LSTM model, despite its theoretical appeal in combining linear and nonlinear modeling strengths, delivered weaker predictive performance across all metrics. This result suggested that either the residuals from the SARIMA model lacked sufficient nonlinear signal or the LSTM component was undertrained due to data limitations or architectural suboptimality.

Nevertheless, this study demonstrated that combining classical and deep learning approaches can significantly enhance the accuracy and reliability of agricultural price forecasts in low-resource environments. While SARIMA contributed interpretability and seasonality modeling, LSTM offered adaptability and the ability to model complex nonlinear dynamics together forming a powerful hybrid system for practical deployment. These insights have strong implications for data-driven agricultural planning, market stability, and farmer resilience in Uganda and similar developing economies.

Overall, the research provides an empirical foundation for leveraging advanced time series modeling in agricultural price forecasting. By integrating tools like LSTM and SARIMAX into decision-making pipelines, Uganda's coffee stakeholders including policymakers, exporters, and cooperatives can better navigate price volatility, optimize planning, and improve market responsiveness in an increasingly unpredictable global context.

References

Ali, A. and Birley, S., 1999. Integrating deductive and inductive approaches in a study of new ventures and customer perceived risk. Qualitative Market Research: An International Journal, 2(2), pp.103–110.

Agreatcoffee.com, n.d. *coffee-trades-impact.* [Online]
Available at: Available at: https://agreatcoffee.com/coffee-trades-impact/
[Accessed 28 November 2024].

Ahmed, M.U., Mahmood, A.N. and Hu, J., 2019. A hybrid model for forecasting exchange rates using ARIMA and artificial neural networks. Mathematics and Computers in Simulation, 166, pp.1–15. https://doi.org/10.1016/j.matcom.2019.03.006

Ahmed, S., Latif, S., Qamar, F., Usman, M., Mehmood, A. and Qamar, U., 2019. A hybrid machine learning framework to predict stock market trends. Complexity, 2019, pp.1–11. https://doi.org/10.1155/2019/8457012

Akhand, M.N.T., Habib, M.A. and Alam, K.M.R., 2023. Analyzing Cryptocurrency Price Trends for Real-Time Price Predictions. *2023 26th International Conference on Computer and Information Technology (ICCIT).*

Bao, W., Yue, J. and Rao, Y., 2017. A deep learning framework for financial time series using stacked autoencoders and long short-term memory. PLoS ONE, 12(7), p.e0180944. https://doi.org/10.1371/journal.pone.0180944

Bkassiny, M., 2022. A Deep Learning-based Signal Classification Approach for Spectrum Sensing using Long Short-Term Memory (LSTM) Networks.. *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE),* pp. 667-672.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Brownlee, J. (2018). Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.

Bryman, A., 2016. Social Research Methods. 5th ed. Oxford: Oxford University Press.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), pp.1247–1250.

Chakraborty, A., Ghosh, I., & Dhar, A. (2020). Time series forecasting using hybrid SARIMA and ANN models based on residuals. Journal of Forecasting, 39(7), 1042-1056.

Chollet, F. (2015). Keras: Deep Learning Library for Theano and TensorFlow. [online] Available at: https://keras.io.

Chatfield, C. (2004). The Analysis of Time Series: An Introduction. 6th ed. Boca Raton: Chapman and Hall/CRC.

Creswell, J.W., 2014. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. *4th ed. Thousand Oaks, CA: Sage.*

Denzin, N. K. and Lincoln, Y. S. (2011). The SAGE Handbook of Qualitative Research. 4th ed. Thousand Oaks: SAGE Publications.

Fan, C., Sun, Y., Zhao, Y., Song, M., and Wang, J., 2019. Deep learning-based feature engineering methods for improved building energy prediction.. *Applied Energy,,* Volume 240, pp. 35-45.

Gabriel, J.M.O., 2013. Inductive and deductive approaches to research. In: Research Made Easy. [online] Available at: https://www.researchgate.net/publication/314014174_Inductive_and_Deductive_Approaches_to_Research [Accessed 12 June 2025].

Gilbert, C.L., Christiaensen, L. & Kaminski, J., 2017. Food price seasonality in Africa: measurement and extent. Food Policy, 67, pp.119-132. https://doi.org/10.1016/j.foodpol.2016.09.016

Mawejje, J., 2016. Food prices, energy and climate shocks in Uganda. Agricultural and Food Economics, 4(3), pp.1-14. https://doi.org/10.1186/s40100-016-0049-6

Géron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. Sebastopol, CA: O'Reilly Media.

Goodfellow, I., Bengio, Y. and Courville, A., 2016. Deep Learning. Cambridge, MA: MIT Press.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R. and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems, 28(10), pp.2222–2232.

Gujarati, D.N. and Porter, D.C. (2009). Basic Econometrics. 5th ed. New York: McGraw-Hill.

Guo, Y., Tang, D., Tang, W., Yang, S., Tang, Q., Feng, Y. & Zhang, F, 2022. 'Agricultural price prediction based on combined forecasting model undersSpatial-temporal influencing factors'. *Sustainability,* 14(17).

Gupta, S.K. & Malik, S., 2022. Application of predictive analytics in agriculture. *Technoarete Transactions on Intelligent Data Mining and Knowledge Discovery,* 2(4).

Han, J., Kamber, M. and Pei, J., 2011. Data Mining: Concepts and Techniques. 3rd ed. Amsterdam: Elsevier.

Hewamalage, H., Bergmeir, C. and Bandara, K., 2021. Recurrent neural networks for time series forecasting: Current status and future directions. International Journal of Forecasting, 37(1), pp.388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

Hyndman, R.J. and Athanasopoulos, G., 2018. Forecasting: Principles and Practice. 2nd ed. Melbourne: OTexts.

Iglewicz, B. and Hoaglin, D.C. (1993) How to Detect and Handle Outliers. Milwaukee, WI: ASQC Quality Press.

Lütkepohl, H. (2005). New Introduction to Multiple Time Series Analysis. Berlin: Springer-Verlag.

Kansiime, M.K., Melesse, M.B., Olwande, J. and Wanjala, B., 2021. Seasonal price variability and smallholder responses in East African food and cash crop markets. Agricultural Economics, 52(1), pp.1–16.

Kansiime, M.K., Ochieng, J. and Mugisha, J. (2021). Seasonality and price volatility of staple and cash crops in East Africa: Insights from time series decomposition. *African Journal of Agricultural and Resource Economics,* 16(2), pp.120–134.

Kingma, D.P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR). Available at: https://arxiv.org/abs/1412.6980

Kleinbaum, D.G., Kupper, L.L., Nizam, A. and Rosenberg, E.S., 2014. Applied regression analysis and other multivariable methods. 5th ed. Boston, MA: Cengage Learning.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR).

Kmytiuk, T., Majore, G. and Bilyk, T, 2024. Time series forecasting of price of the agricultural products using data science. *Agricultural and Resource Economics: International Scientific E-Journal*, 10(3), p. 5.

Kori, D.S., Musakwa, W. and Kelso, C., 2024. Understanding the local implications of climate change: Unpacking the experiences of smallholder farmers in Thulamela Municipality, Vhembe District, Limpopo Province. *South Africa. PLOS Climate.*

Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), pp.111–117.

Kotu, V. and Deshpande, B., 2019. Data Science: *Concepts and Practice. 2nd ed. Cambridge*, MA: Morgan Kaufmann.

Livieris, I.E., Pintelas, E. and Pintelas, P. (2021). 'A CNN–LSTM model for gold price time-series forecasting', Neural Computing and Applications, 33(2), pp. 751–760.

Magrini, E., Balie, J. and Morales-Opazo, C., 2017. Cereal price shocks and volatility in sub-Saharan Africa: What really matters for farmers' welfare?. *Agricultural Economics,* Volume 48, p. 48.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2018). *Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward.* PLOS ONE, 13(3), p.e0194889. https://doi.org/10.1371/journal.pone.0194889

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). Forecasting: Methods and Applications. 3rd ed. New York: John Wiley & Sons.

Mohamad, A.F., Canda, R., Mat Jasin, A., Ismail, J., Asmat, A. and Md Soom, A.B., 2024. Sales analytics dashboard with ARIMA and SARIMA time series model.. *College of Computing, Informatics and Media,* Volume 1, pp. 1-6.

Mende, N., Abate, G.T. and Bernard, T., 2022. Forecasting coffee prices using machine learning and hybrid models. Journal of Agricultural Economics, 73(1), pp.101–123. https://doi.org/10.1111/1477-9552.12450

Melnikovas, A., 2018. Towards an explicit research methodology: Adapting research onion model for futures studies. Journal of Futures Studies, 23(2), pp.29–44. https://doi.org/10.6531/JFS.201812_23(2).0003

Mugisha, J., Elepu, G., and Matsiko, F.B., 2018. Volatility and price forecasting of major food crops in Uganda: An application of GARCH models. *African Journal of Agricultural and Resource Economics, 13(1), pp.35–47.*

Mugisha, J., Ekere, W. and Lwasa, S. (2018). Price volatility and forecasting in Uganda's agricultural sector: *A time series GARCH approach. Journal of Development and Agricultural Economics, 10(3), pp.79–91.*

Mugisha, J., Ninsiima, L. R., & Atukwatse, A. (2023). Application of SARIMA models to forecast coffee prices in Uganda. *African Journal of Agricultural Research, 18(4), 123–132.*

Mugagga, F., Turyahabwe, N. and Katongole, C., 2021. Time series analysis and forecasting of coffee export earnings in Uganda. African Journal of Agricultural Research, 16(3), pp.341–350. https://doi.org/10.5897/AJAR2020.15367

Muhammad, H, 2024. 'Predictive analytics - techniques, tools and examples'. *Emerging Methods.*

Nabbumba, R., & Bategeka, L. (2021). Forecasting food crop prices using time series analysis in Uganda. Makerere University Working Paper

Ngoc, T.N.L., Lam, D.T., Minh, T.N.H., Doan, T.C., Nguyen, N.P. & Nguyen, H.M., 2023. Machine learning for agricultural price prediction: A case of coffee commodity in Vietnam market'. *IEEE*.

Njeri, N.R., 2022. Data preparation for machine learning modelling. International Journal of Computer Applications Technology and Research, 11(6), pp.231–235. Available at: https://doi.org/10.7753/IJCATR1106.1008 [Accessed 12 Jun. 2025].

Nugroho, A.D., 2021. Agricultural market information in developing countries: A literature review. *Agricultural Economics – Czech*, Volume 67, p. 468–477.

Olah, C. (2015). Understanding LSTM networks. Colah's Blog. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ouyang, H., Wei, X. and Wu, Q, 2019. 'Agricultural commodity futures prices prediction via long- and short-term time series network'. *Journal of Applied Economics*, Volume 22(10), p. 468–483.

Owusu, A.B., Yankson, P.W.K. and Frimpong, S., 2021. Smallholder farmers' knowledge of mobile telephone use: Gender perspectives and implications for agricultural market development.. *Progress in Development Studies*, Volume 36–51, p. 18.

Pandian, S., 2024. *Time Series Analysis: Definition, Components, Methods, and Applications.*. [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-to-time-series-analysis/
[Accessed 28 November 2024].

Park, C., 2023. Research philosophy: Positivism and its importance in quantitative research. International Journal of Research Methodologies, 14(2), pp.44–55.

Parreño, S.J.E., 2023. Forecasting quarterly rice and corn production in the Philippines: A comparative study of seasonal ARIMA and Holt-Winters models. *International Journal of Soft Computing.*

Purohit, H., Bansal, R., & Sharma, A. (2021). Applications of machine learning in agricultural price forecasting: A review. Agricultural Systems, 190, 103118.

Purohit, H., Patidar, S., Rani, A. and Saxena, S., 2021. *Predictive analytics in agriculture using machine learning models.* Journal of King Saud University - Computer and Information Sciences*, 33(4), pp.403–410.*

Purohit, H., Upadhyay, P., Dwivedi, Y.K. and Raghavan, V.V. (2021) 'Big data in agriculture: A systematic literature review and future research agenda', Information Systems Frontiers, 23(2), pp. 333–358. doi: 10.1007/s10796-019-09962-1.

Purohit, S.K., Panigrahi, S., Sethy, P.K. & Behera, S.K., 2021. Time Series Forecasting of Price of Agricultural Products Using Hybrid Methods. Applied Artificial Intelligence. *Applied Artificial Intelligence,* Volume 35(15), p. 1388–1406.

Pyle, D., 1999. Data Preparation for Data Mining. San Francisco: Morgan Kaufmann.

Ryan, G., 2018. Introduction to positivism, interpretivism and critical theory. Nurse Researcher, 25(4), pp.41–49. https://doi.org/10.7748/nr.2018.e1466

Rapsomanikis, G., Hallam, D. & Conforti, P., n.d. 'Market integration and price transmission in selected food and cash crop markets of developing countries: Review and applications'.

Ruekkasaem, L. & Sasananan, M., (2018). 'forecasting agricultural products prices using time series methods for crop planning',. *International Journal of Mechanical Engineering and Technology,* Volume 9, p. 957–971.

Rundo, F., Trenta, F., Battiato, S. and Ortis, A. (2019). 'Machine Learning for Quantitative Finance Applications: Forecasting and Trading Strategies', ACM Computing Surveys (CSUR), 52(3), pp. 1–36.

Rustagi, M. & Goel, N., 2022. Rustagi, M. & Goel, N., 2022.'Predictive Analytics: A study of its advantages and applications". *International Research Journal,* 12,(1), p. pp. 60–63.

Sagheer, A., & Kotb, M. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. Neurocomputing, 323, 203-213.

Said, S.E. and Dickey, D.A., 1984. Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 71(3), pp.599–607.

Saleh, M. A., Mohd, N., and Sahrir, M. S. (2008). Forecasting Approaches in Economic Research: A Review. *Malaysian Journal of Economic Studies, 45(1), pp. 23–44.*

Saunders, M., Lewis, P. and Thornhill, A., 2019. Research Methods for Business Students. 8th ed. Harlow: Pearson Education.

SAP, 2024. *What is predictive analytics.* [Online]
Available at: https://www.sap.com/products/technology-platform/cloud-analytics/what-is-predictive-analytics.html
[Accessed 9 November 2024].

Siami-Namini, S., Tavakoli, N. and Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, pp. 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

Simotwo, H.K., Mikalitsa, S.M., and Wambua, B.N., 2018. Climate change adaptive capacity and smallholder farming in Trans-Mara East sub-County, Kenya. *Geoenvironmental Disasters,* 5(5).

Slater, L.J., Arnal, L., Boucher, M.-A., Chang, A.Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R.L., Wood, A. and Zappa, M., 2023. Hybrid forecasting: blending climate predictions with AI models. *Hydrology and Earth System Sciences,* Volume 1865-1880.

Sun, F., Meng, X., Zhang, Y., Wang, Y., Jiang, H. & Liu, P., 2023. .'Agricultural product price forecasting methods: A review'. *Agriculture,* 13(9), p. 1671.

Touch, V., Tan, D.K.Y., Cook, B.R., Liu, D.L., Cross, R., Tran, T.A., Utomo, A., Yous, S., Grunbuhel, C. & Cowie, A., 2024. Smallholder farmers' challenges and opportunities: Implications for agricultural production, environment and food security. *Journal of Environmental Management*, Volume 370, p. 122536.

Tran, N.-Q., Felipe, A., Ngoc, T.N., Huynh, T., Tran, Q., Tang, A. & Nguyen, T., 2023. Predicting Agricultural Commodities Prices with Machine Learning: A Review of Current Research. *Computer Science , Artificial Intelligence,* Oct .

Tukey, J.W., 1977. Exploratory Data Analysis. Reading, MA: Addison-Wesley.

UBOS, 2020. Producer Price Index. *December 2020.*

UBOS, 2019. Annual Agricultural Survey (AAS) 2019 – Statistical Release. *The majority of households in Uganda are engaged in agriculture.*

Uganda Bureau of Statistics (UBOS). (2022). Statistical Abstract. Kampala: Government of Uganda.

UCDA, 2023/2024. *Montly Report October 2023,* Kampala: Directorate of Strategy and Business Development.

Uganda Coffee Development Authority, n.d. *Uganda Country Coffee Profile.*, s.l.: s.n.

Uganda Coffee Development Authority (UCDA). (n.d.). Monthly Coffee Market Reports.

Waiswa, D. and Yavuz, F, 2023. Market integration and asymmetric price transmission in selected domestic markets for major staple foods in Uganda. *Future Business Journal.*

Wang, J., Zhang, W. and Li, Q., 2020. Hybrid ARIMA and LSTM models for air quality index prediction. Applied Sciences, 10(3), p.928. https://doi.org/10.3390/app10030928

Wang, X., Liu, Y., Luo, M. and Wang, Y., 2021. Hybrid SARIMA-LSTM model for forecasting agricultural crop yield. Agricultural Economics and Management, *2(3), pp.51–62.*

Willmott, C.J. and Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), pp.79–82.

Zhang, D., Chen, S., Ling, L. & Xia, Q., 2020. Forecasting Agricultural Commodity Prices Using Model Selection Framework With Time Series Features and Forecast Horizons. *IEEE Access,* 04 February .pp. 28197 – 28209.

Zhang, G., Eddy Patuwo, B. and Hu, M. Y. (2019) 'Forecasting with artificial neural networks: The state of the art', International Journal of Forecasting, 14(1), pp. 35–62.

Zhang, J., Wang, Y., Zhang, Y. and Zhao, H., 2020. A hybrid SARIMA–LSTM model for forecasting grain prices. Journal of Intelligent & Fuzzy Systems, 38(3), pp.2961–2972. https://doi.org/10.3233/JIFS-179202

Zhang, Y., Saghaian, S. and Reed, M., 2020. Forecasting agricultural commodity prices with hybrid models. Agricultural Economics Review, 21(1), pp.24–37.

Zhang, W., Saghaian, S. & Reed, M., 2022. Influences of power structure evolution on coffee commodity markets: Insights from price discovery and volatility spillovers. *Sustainability*, 14(22), p. 15268.

Zheng, Y., Liu, Q., Chen, E., Ge, Y., & Zhao, J. L. (2020). Time series classification using multi-channels deep convolutional neural networks. International Conference on Web-Age Information Management, 298–310.

Yu, Y., Wang, H. and Lai, K.K., 2021. Hybrid ARIMA and LSTM model for electricity consumption forecasting. Energy, 215, p.119153. https://doi.org/10.1016/j.energy.2020.119153

Appendices
Appendix one

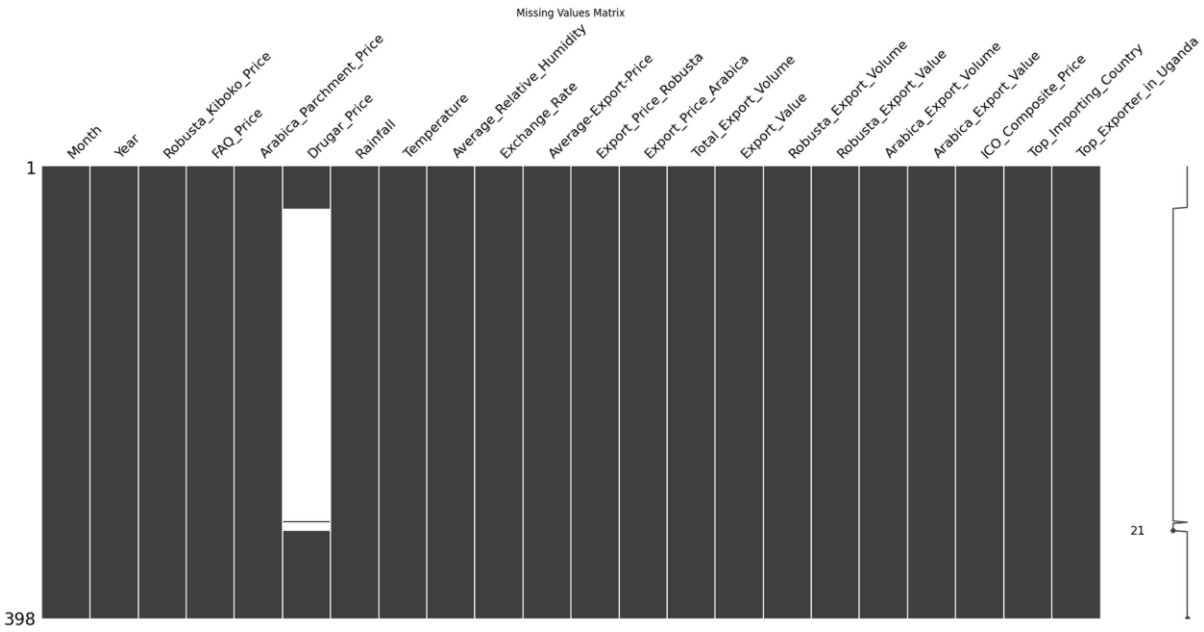
Descriptive Statistics of the Variables

```
#Descriptive Statistics
coffee_df.describe(include='all')
```

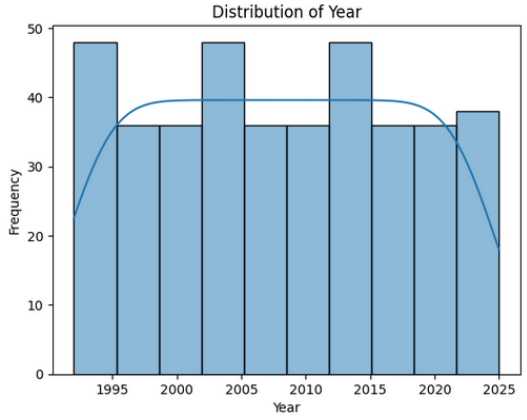
	Month	Year	Robusta_Kiboko_Price	FAQ_Price	Arabica_Parchment_Price	Drugar_Price	Rainfall	Temperature	Average_Relative_Humidity	Exchange_Rate
count	398	398.000000	398.000000	398.000000	398.000000	114.000000	398.000000	398.000000	398.000000	398
unique	12	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	398
top	January	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	3,677.71
freq	34	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1
mean	NaN	2008.085427	1479.520101	3119.731156	4010.248744	5500.219298	133.107115	23.172362	59.008794	NaN
std	NaN	9.585765	1237.073358	2516.366584	3136.042748	3869.177566	74.303566	0.898271	10.844864	NaN
min	NaN	1992.000000	150.000000	400.000000	459.000000	436.000000	7.503283	21.300000	29.480000	NaN
25%	NaN	2000.000000	600.000000	1262.500000	1435.000000	940.750000	80.254022	22.500000	52.367500	NaN
50%	NaN	2008.000000	1100.000000	2200.000000	2637.500000	5248.500000	121.855500	23.000000	60.295000	NaN
75%	NaN	2016.000000	2100.000000	4298.250000	5750.000000	9250.000000	171.379750	23.800000	67.805000	NaN
max	NaN	2025.000000	7750.000000	15500.000000	14500.000000	13250.000000	498.109000	25.700000	78.850000	NaN

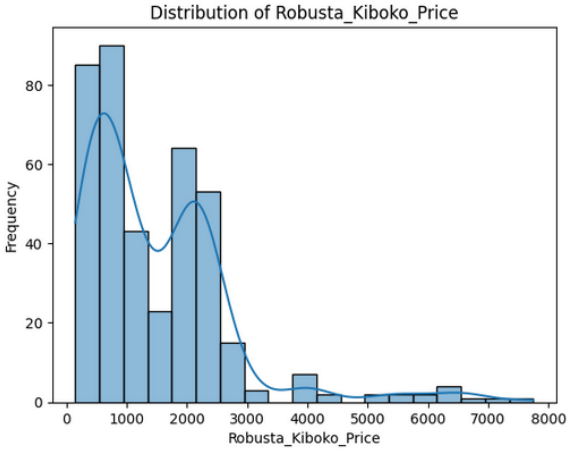
11 rows × 22 columns

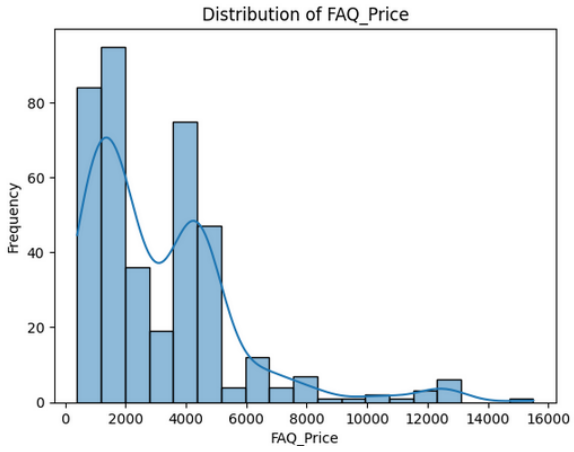
Missing Value Matrix

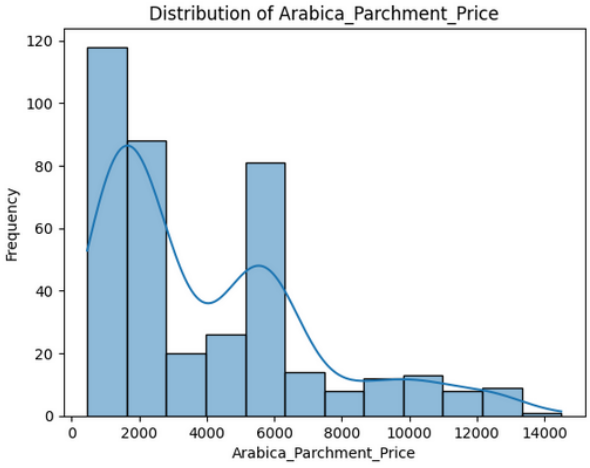


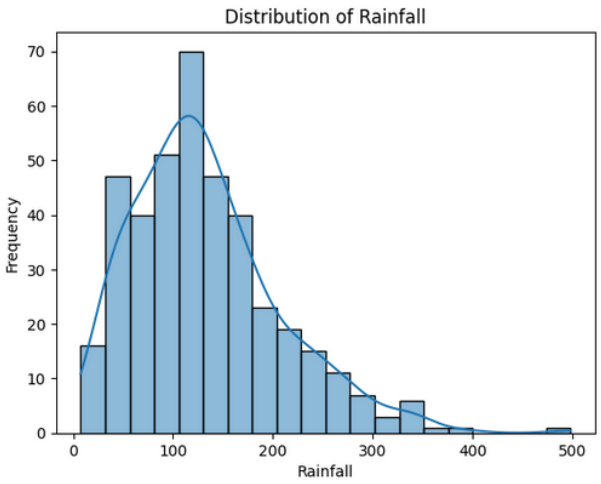
Distribution Plots for Key Numerical Variables

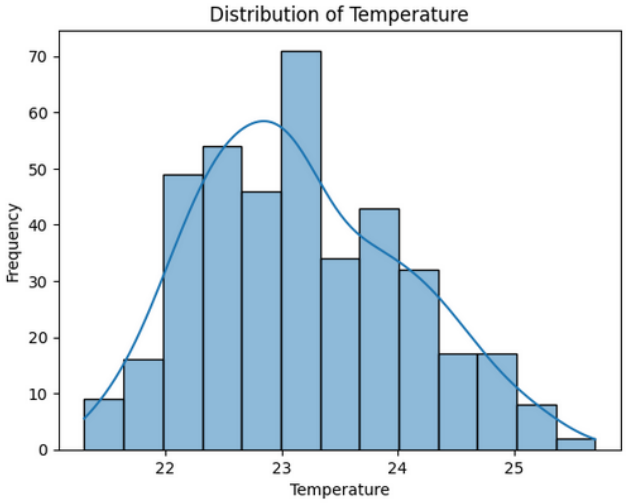


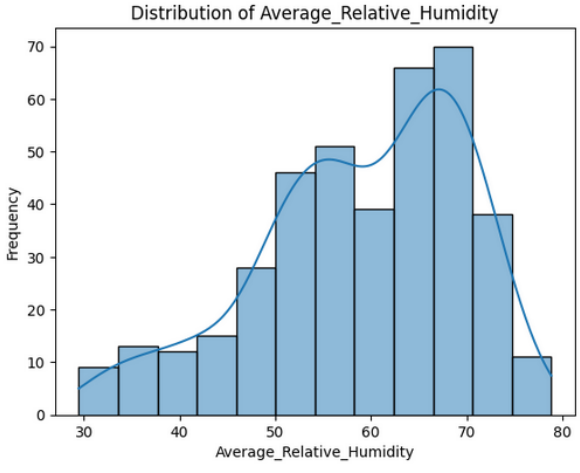


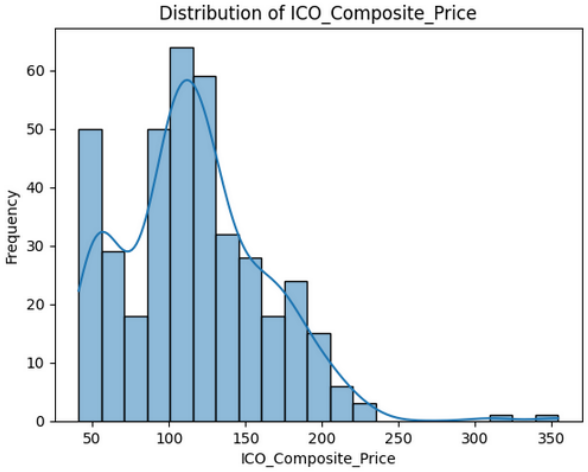




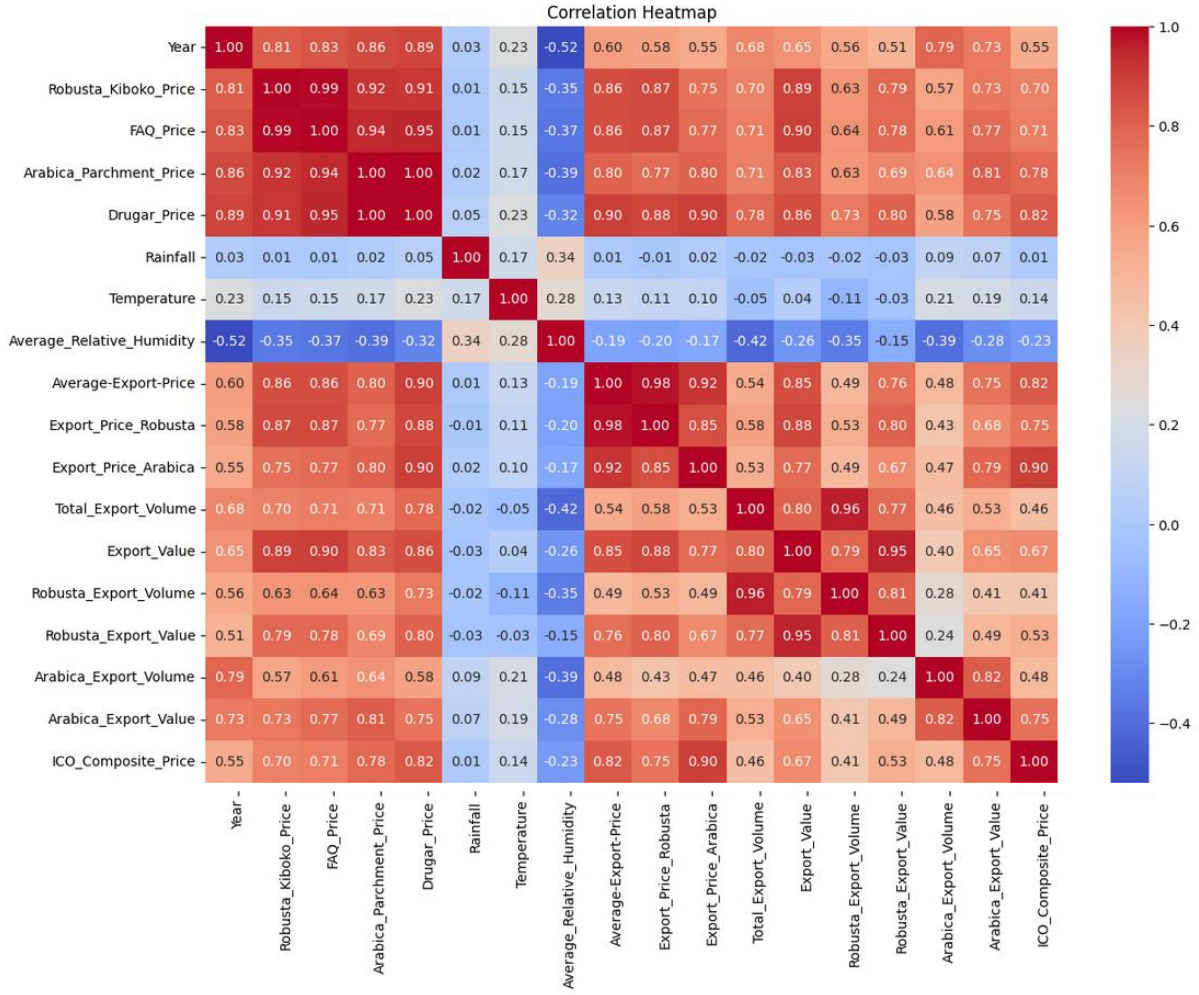

Distribution of Rainfall
Frequency
Rainfall
0
100
200
300
400
500
0
10
20
30
40
50
60
70


Distribution of Temperature
Frequency
Temperature
22
23
24
25
0
10
20
30
40
50
60
70

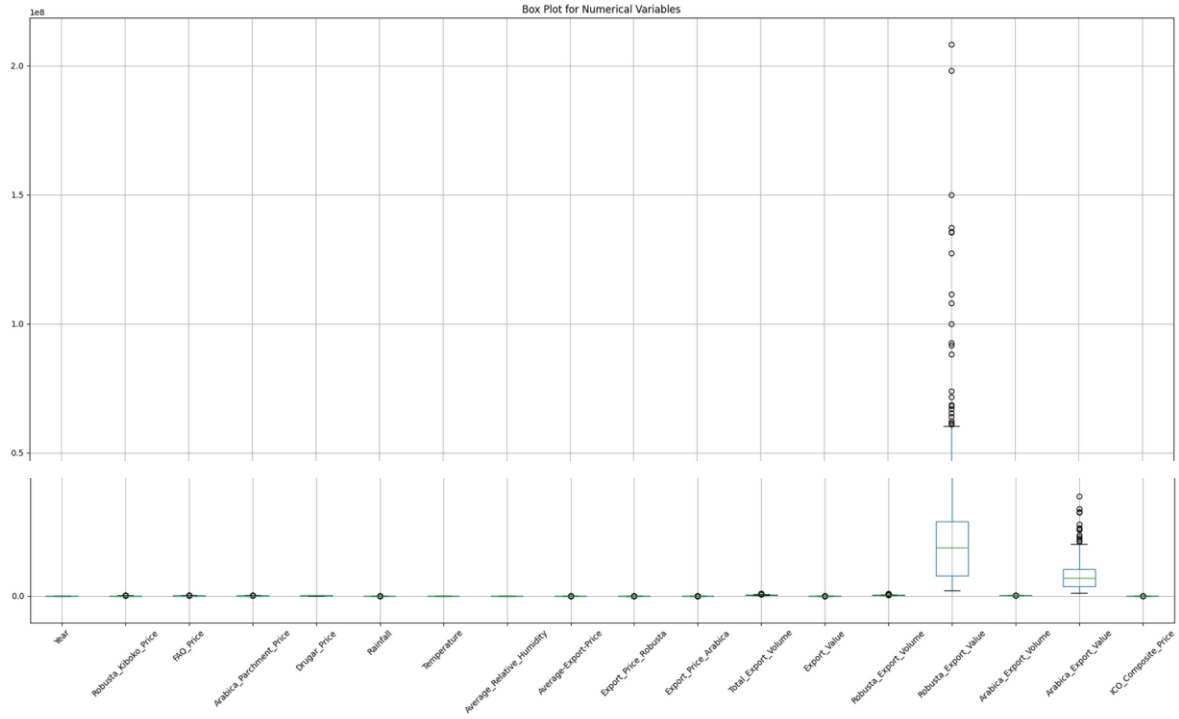

Distribution of Average_Relative_Humidity
Frequency
Average_Relative_Humidity
30
40
50
60
70
80
0
10
20
30
40
50
60
70


Distribution of ICO_Composite_Price
Frequency
ICO_Composite_Price
50
100
150
200
250
300
350
0
10
20
30
40
50
60

Correlation Heatmap For Variables



Box Plot for Numerical Variables



Regression Results

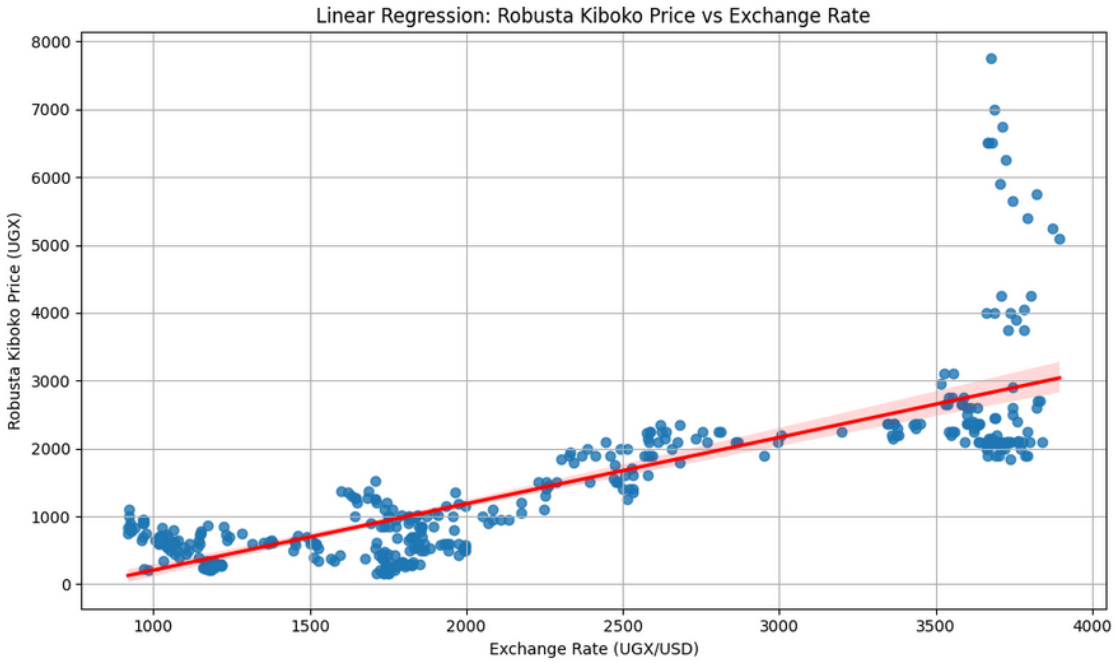
OLS Regression Results

Dep. Variable:	Robusta_Kiboko_Price	R-squared:	0.606
Model:	OLS	Adj. R-squared:	0.605
Method:	Least Squares	F-statistic:	608.2
Date:	Sun, 01 Jun 2025	Prob (F-statistic):	4.96e-82
Time:	23:30:16	Log-Likelihood:	-3213.0
No. Observations:	398	AIC:	6430.
Df Residuals:	396	BIC:	6438.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-773.9820	99.348	-7.791	0.000	-969.297	-578.667
Exchange_Rate	0.9788	0.040	24.661	0.000	0.901	1.057

Omnibus:	270.956	Durbin-Watson:	0.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2847.321
Skew:	2.845	Prob(JB):	0.00
Kurtosis:	14.804	Cond. No.	6.38e+03

Linear Regression: Robusta Kiboko Price vs Exchange Rate



Appendix Two

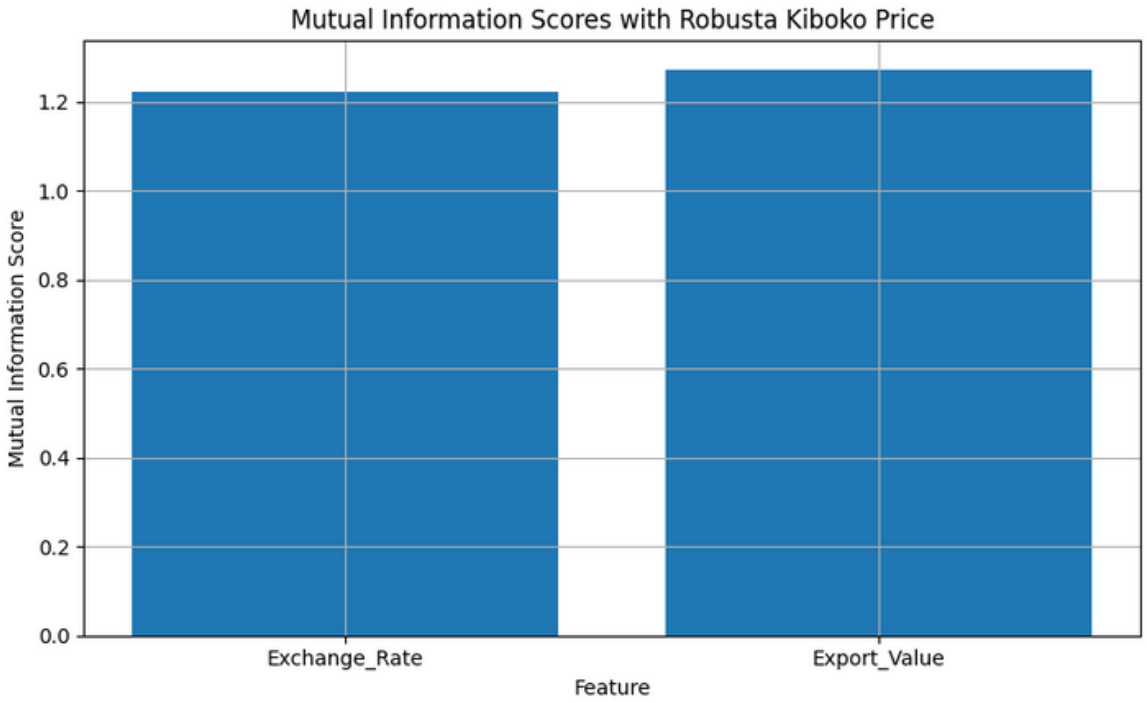
Normalization Techniques

Summary Statistics of Scaling Techniques

Variable	mean	std	min	max
Year	2012.36	13.34	1992	2025
Robusta_Kiboko_Price	2283.46	1789.62	207	7750
FAQ_Price	4776.94	3631.86	400	15500
Arabica_Parchment_Price	5995.17	4243.72	459	14500
Drugar_Price	5500.22	3869.18	436	13250
Rainfall	137.75	79.99	19.85	391.3
Temperature	23.13	0.87	21.4	25.5
Average_Relative_Humidity	56.14	11.35	30.68	78.74
Average-Export-Price	1.97	1.01	0.85	5.03
Export_Price_Robusta	1.84	1.01	0.81	4.83

Export_Price_Arabica	2.68	1.13	1.22	5.4
Total_Export_Volume	392998.8	161861.38	171101	837915
Export_Value	52.24	42.55	8.71	221.63
Robusta_Export_Volume	333624.14	147514	148938	785667
Robusta_Export_Value	42203494.61	38390271.95	7548012	208138678
Arabica_Export_Volume	59591.49	30482.85	13003	134974
Arabica_Export_Value	10667385.41	7984987.19	1067793	38494406
ICO_Composite_Price	130.02	56.67	45.89	354.32
Year_minmax	0.62	0.4	0	1
Robusta_Kiboko_Price_minmax	0.28	0.24	0	1
FAQ_Price_minmax	0.29	0.24	0	1
Arabica_Parchment_Price_minmax	0.39	0.3	0	1
Drugar_Price_minmax	0.4	0.3	0	1
Rainfall_minmax	0.32	0.22	0	1
Temperature_minmax	0.42	0.21	0	1
Average_Relative_Humidity_minmax	0.53	0.24	0	1
Average-Export-Price_minmax	0.27	0.24	0	1
Export_Price_Robusta_minmax	0.26	0.25	0	1
Export_Price_Arabica_minmax	0.35	0.27	0	1
Total_Export_Volume_minmax	0.33	0.24	0	1
Export_Value_minmax	0.2	0.2	0	1
Robusta_Export_Volume_minmax	0.29	0.23	0	1
Robusta_Export_Value_minmax	0.17	0.19	0	1
Arabica_Export_Volume_minmax	0.38	0.25	0	1
Arabica_Export_Value_minmax	0.26	0.21	0	1
ICO_Composite_Price_minmax	0.27	0.18	0	1
Year_zscore	0	1	-1.53	0.95
Robusta_Kiboko_Price_zscore	0	1	-1.17	3.07
FAQ_Price_zscore	0	1	-1.21	2.97

Arabica_Parchment_Price_zscore	0	1	-1.31	2.01
Drugar_Price_zscore	0	1	-1.31	2.01
Rainfall_zscore	0	1	-1.48	3.18
Temperature_zscore	0	1	-1.99	2.73
Average_Relative_Humidity_zscore	0	1	-2.25	2
Average-Export-Price_zscore	0	1	-1.11	3.04
Export_Price_Robusta_zscore	0	1	-1.03	2.96
Export_Price_Arabica_zscore	0	1	-1.3	2.42
Total_Export_Volume_zscore	0	1	-1.38	2.76
Export_Value_zscore	0	1	-1.03	4
Robusta_Export_Volume_zscore	0	1	-1.26	3.08
Robusta_Export_Value_zscore	0	1	-0.91	4.34
Arabica_Export_Volume_zscore	0	1	-1.54	2.48
Arabica_Export_Value_zscore	0	1	-1.21	3.5
ICO_Composite_Price_zscore	0	1	-1.49	3.98



Appendix Three

Descriptive Statistics

	Month	Year	Robusta_Kiboko_Price	FAQ_Price	Arabica_Parchment_Price	Drugar_Price	Rainfall	Temperature
count	398	398.000000	398.000000	398.000000	398.000000	114.000000	398.000000	398.000000
unique	12	NaN	NaN	NaN	NaN	NaN	NaN	NaN
top	January	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freq	34	NaN	NaN	NaN	NaN	NaN	NaN	NaN
mean	NaN	2008.085427	1479.520101	3119.731156	4010.248744	5500.219298	133.107115	23.172362
std	NaN	9.585765	1237.073358	2516.366584	3136.042748	3869.177566	74.303566	0.898271
min	NaN	1992.000000	150.000000	400.000000	459.000000	436.000000	7.503283	21.300000
25%	NaN	2000.000000	600.000000	1262.500000	1435.000000	940.750000	80.254022	22.500000
50%	NaN	2008.000000	1100.000000	2200.000000	2637.500000	5248.500000	121.855500	23.000000
75%	NaN	2016.000000	2100.000000	4298.250000	5750.000000	9250.000000	171.379750	23.800000
max	NaN	2025.000000	7750.000000	15500.000000	14500.000000	13250.000000	498.109000	25.700000

11 rows × 22 columns