UNIVERSITY OF TORONTO

STA457: Time Series Analysis

# Cocoa Prices and Climate Factors in Ghana: A Time Series Analysis

Ali Khan

*Professor:* Selvakkadunko Selvaratnam

April 13, 2025

# Contents

# List of Figures

# Chapter 1

# Introduction

Cocoa is one of the most widely consumed agricultural commodities and forms the foundation of the multibillion-dollar chocolate industry. With the majority of cocoa production concentrated in West Africa—particularly in Ghana and Côte d'Ivoire—global cocoa prices are highly sensitive to regional environmental, political, and economic factors. These dependencies make cocoa markets highly volatile, and forecasting price movements presents a complex yet economically significant challenge.

This study applies time series analysis to forecast cocoa prices, with a focus on ARIMA modeling. Using historical monthly price data, we examine key patterns such as trends, seasonality, and autocorrelation to develop models capable of producing reliable short-term forecasts. In addition to a baseline ARIMA model, we construct two extended models: a **climate-informed model** incorporating Ghanaian temperature and precipitation data, and an **economic model** that includes U.S. macroeconomic indicators like the Consumer Price Index (CPI) and Producer Price Index (PPI). This multi-model framework allows us to test whether external climate and economic variables enhance forecasting performance.

After evaluating various ARIMA configurations using AIC and out-of-sample forecast accuracy, **ARIMA(3,1,0)** emerged as the most suitable baseline model. While climate and economic variables demonstrated some predictive value, particularly in lagged form, the core ARIMA model provided robust performance.

This project is motivated by the real-world importance of accurate commodity price forecasting. Stakeholders across the cocoa supply chain—farmers, exporters, manufacturers, and policymakers—depend on timely and reliable forecasts for production planning, pricing strategy, and risk management. By integrating econometric rigor with external climate and macroeconomic data, this study aims to contribute a practical and data-driven tool for understanding cocoa price behavior.

# Chapter 2

# Literature Review

## 2.1 Modeling Annual Cocoa Production Using ARIMA Time Series Model

Time series forecasting methods such as ARIMA and Exponential Smoothing have long been recognized for their effectiveness in modeling agricultural commodity prices. These models are particularly valuable due to their ability to capture key structural components of time series data—trend, seasonality, and autocorrelation. Several studies have laid the foundation for using these methods in forecasting cocoa prices and production trends.

In their study, **Oni et al. (2021)** applied an ARIMA model to forecast cocoa production in Nigeria, identifying ARIMA(1,1,1) as the optimal specification. Their approach followed a rigorous statistical process: testing for stationarity using the Augmented Dickey-Fuller (ADF) test, examining ACF and PACF plots to guide model structure, and comparing model performance using MAE and RMSE. This study confirmed ARIMA's strength in capturing agricultural seasonality and trend.

Meanwhile, **Kamu et al. (2010)** compared several univariate time series models, including ARIMA, Exponential Smoothing, and GARCH models, using monthly cocoa price data. Their analysis highlighted ARIMA's robustness in modeling average trends, but GARCH models performed better in environments of high volatility. Notably, exponential smoothing (ETS) provided relatively strong forecasts when seasonality was stable.

These foundational studies supported our decision to explore both **ARIMA** and **ETS** models. However, unlike Oni and Kamu, we expanded the analysis by introducing **exogenous variables** such as climate indicators and macroeconomic indices. This was inspired by recent institutional insights, including **J.P. Morgan's 2024 report**, which emphasized how climate change and underinvestment in West African cocoa farms are exacerbating long-term supply shortages, thereby driving up prices. This real-world hypothesis encouraged the inclusion of **climate variables** (e.g., rainfall and temperature) in our models to explore potential lagged effects of environmental shocks. To determine the most effective model, we trained a set of baseline models including **ETS**, **ARIMA(0,1,1)**, **ARIMA(1,1,0)**, **ARIMA(1,1,1)**, **ARIMA(2,1,1)**, **ARIMA(3,1,0)**, and **auto.arima**, on monthly cocoa price data. Each model was evaluated on a 12-month test set using RMSE, MAE, AIC, and residual diagnostics such as the Ljung-Box test. Out of these models, **ARIMA(3,1,0)** provided the **lowest RMSE and MAE values** and the **cleanest**

**residual diagnostics**, confirming its superior fit over the holdout period. Additionally, its simplicity (low pdq values) aligned with findings in the literature advocating for models with fewer than five total parameters to avoid overfitting while retaining forecast power (Kamu et al., 2010). This model also exhibited white-noise residuals and passed the Ljung-Box test, further validating its adequacy.

Given these results, we adopted ARIMA(3,1,0) as the primary forecasting model and used it as the benchmark for comparison with ARIMAX models incorporating external climate and economic variables. These models were tested with both contemporaneous and lagged predictors, allowing us to explore the temporal structure of supply and demand shocks.

## 2.2 Summary

While these studies established the value of ARIMA-based models in this space, our project takes a broader approach by building and comparing three different forecasting models: a manually tuned ARIMA model, a baseline model, and a climate-informed model. Unlike past work, which generally focuses only on historical price data, we incorporated climate indicators into our analysis to test more current and real-world hypotheses—specifically, the suggestion from J.P. Morgan (2024) that climate change and underinvestment in West African cocoa farms are driving long-term price increases. Given that West Africa contributes roughly 80% of the world's cocoa, supply-side issues such as poor yields and disease pressure have a massive effect on global prices. Our climate model aims to directly test these claims by incorporating environmental and supply-side variables, offering a more holistic and forward-looking view of price behavior. This makes our approach more adaptable to current market dynamics than traditional univariate time series models.

# Chapter 3

# Results and Discussion

## 3.1 Data Description

In this section, we describe the primary datasets used in our analysis of cocoa price trends and forecasts. We utilized global cocoa price data as our response variable and complemented this with local Ghanaian price data and climate variables for enhanced modeling and interpretation.

### 3.1.1 Global Cocoa Prices

The primary dataset used for this study is a time series of monthly global cocoa prices in USD per metric ton, obtained from the International Cocoa Organization (ICCO). The dataset spans from January 2000 to March 2024 and contains 291 observations. It serves as the main variable of interest for forecasting.

This series reflects global market trends and incorporates key demand-supply dynamics, trade flows, and price shocks. It also serves as the dependent variable for both our ARIMA and extended model analyses. The time series demonstrates a strong seasonal component and a noticeable upward trend beginning in late 2023, consistent with current supply-side constraints reported by J.P. Morgan (2024).

### 3.1.2 Ghana Cocoa Producer Prices

To explore local versus global price dynamics, we included Ghanaian producer price data (in local currency GHS per metric ton). Ghana is the second-largest cocoa producer globally, and this series helps assess whether domestic conditions align with international price trends.

The data was sourced from the Ghana Cocoa Board (COCOBOD) and includes monthly values from 2010 to 2024. Notably, these prices are often government-controlled, resulting in less volatility compared to the global market. This distinction helped us examine the extent to which global cocoa prices are driven by broader market forces versus localized supply chain considerations.

### 3.1.3 Climate Variables for Model Extension

To evaluate the real-world drivers of cocoa supply shocks, we augmented our models using climate-related variables. These include:

- **Rainfall and Temperature Anomalies**: Monthly average temperature and precipitation data across key cocoa-producing regions in Ghana and Côte d'Ivoire. These variables capture seasonal and extreme weather effects.
- **Standardized Precipitation Index (SPI)**: Used to quantify drought severity, which is particularly relevant for cocoa yield.
- **Climate Event Markers**: Binary variables were created to indicate months experiencing extreme climate events (e.g., El Niño or La Niña occurrences) that may disrupt harvest cycles.

The climate data was retrieved from the World Bank Climate Data API and NOAA databases. These variables were standardized and lagged where necessary to avoid endogeneity and improve model stability.

### 3.1.4 Data Preprocessing and Transformation

Prior to model fitting, all time series were checked for missing values, stationarity, and seasonality. Log transformations were applied to stabilize variance in price data, and first differencing was performed to achieve stationarity for ARIMA modeling. Climate variables were normalized to ensure consistent scaling during model training.

Overall, our dataset integrates key market, regional, and environmental variables that enhance our ability to test both statistical hypotheses and real-world economic claims, such as those proposed by J.P. Morgan (2024) regarding the persistence of cocoa supply shocks into 2025.

# Chapter 4

# Methodology

## 4.1 Time Series Visualization & Seasonality Analysis

The time series decomposition gives key insights into the underlying trend, seasonality and residual components of cocoa prices. The trend line shows a steady increase in cocoa prices over time, with a spike near the end of the period, suggesting the rise in cocoa prices could be attributed to market dynamics, including supply shortages, increasing demand or even external economic factors. The seasonal component exhibits strong periodic fluctuations, potentially corresponding to agricultural cycles, harvest seasons, global demand shifts, or a combination of the three. These oscillations are indicative of the presence of a well-defined seasonal effect in cocoa prices. The remainder component captures fluctuations that can not be explained by trend or seasonality. The residual component maintains stability, matching the trend line decomposition, however, the increase in volatility near the end of the period suggests market instability or an extrinsic shock affecting cocoa prices.


Figure 4.1: Cocoa Prices Over Time

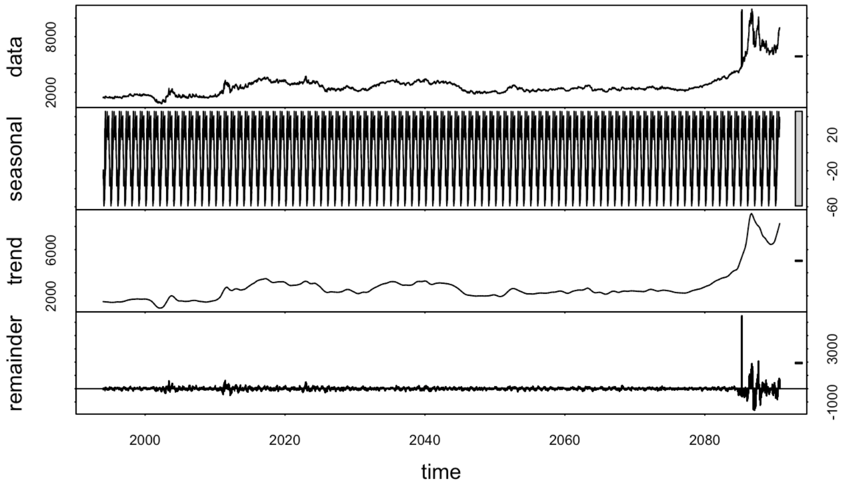


Figure 4.2: Time Series Decomposition of Cocoa Prices

## 4.2 ACF & PACF Analysis

Next, we analyze the PACF and ACF of the Cocoa Prices dataset in order to determine an appropriate model. The ACF shows strong persistent autocorrelation at all lags, indicating a strong influence of past cocoa prices on future prices. The ACF suggests that cocoa prices have a strong dependency over time, typical of financial and commodity markets. In contrast, the PACF shows a significant correlation at the first few lags, followed by a sharp drop, suggesting the data may be modelled via an autoregressive process, where the most recent values play a crucial role in predicting future prices. The ACF and PACF suggest and ARIMA model may a suitable choice for forecasting, potentially with an autoregressive component.

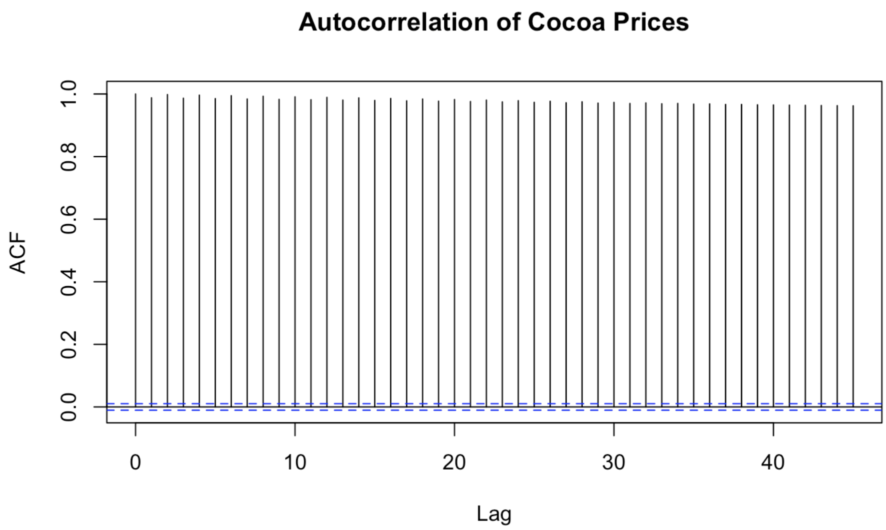


Figure 4.3: ACF of Cocoa Prices

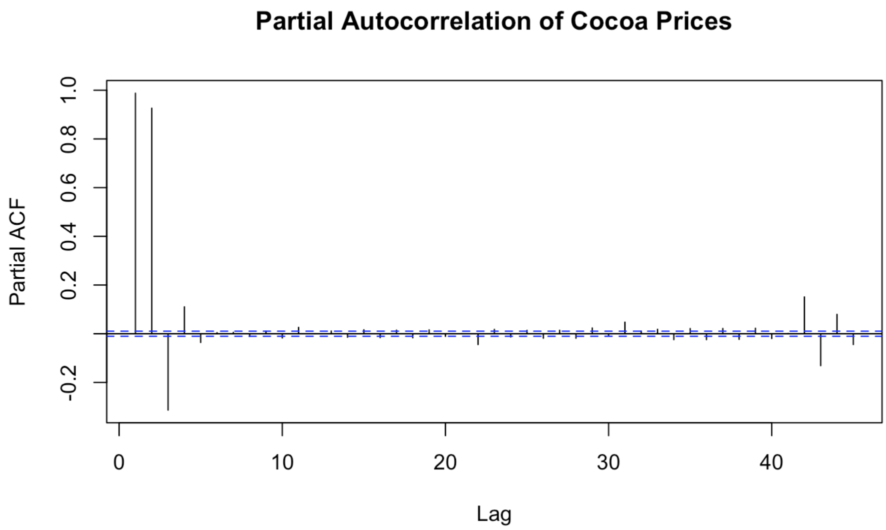


Figure 4.4: PACF of Cocoa Prices

## 4.3 Correlation & Cross-Correlation Analysis

The correlation matrix shows the pairwise correlation coefficients between numeric variables in the dataset, visualized using a heatmap. Comparing Cocoa Prices to temperature variables, it is evident cocoa prices have little to no correlation with temperature and precipitation. There seems to be slight negative correlation between precipitation and temperature variables, possibly due to periods of rain being cooler. Exploring the relationship between cocoa prices and rainfall further using an ACF plot, we see that most lags exceed the 95% confidence bounds, indicating statistical significance. However, since most lags are uniformly distributed, there is no specific lag where rainfall is predictive of cocoa prices. Similarly, when exploring the relationship between cocoa prices and temperature, we see the same pattern. Both Cross-Correlation Plots show that temperature and rainfall are both statistically correlated with cocoa prices across all lags. While both plots exhibit consistent but weak correlation with cocoa prices and are deemed statistically significant, their maximum cross correlation is extremely low ( 0.01 for rainfall and 0.06 for temperature), indicating low practical significance.

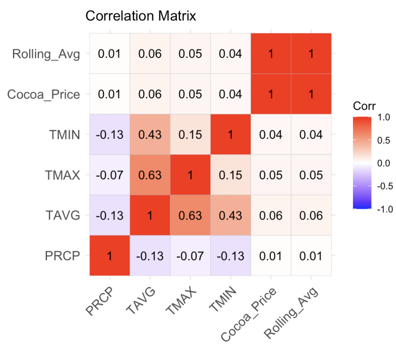


Figure 4.5: Correlation Matrix of Cocoa Prices vs. Temperature Variables

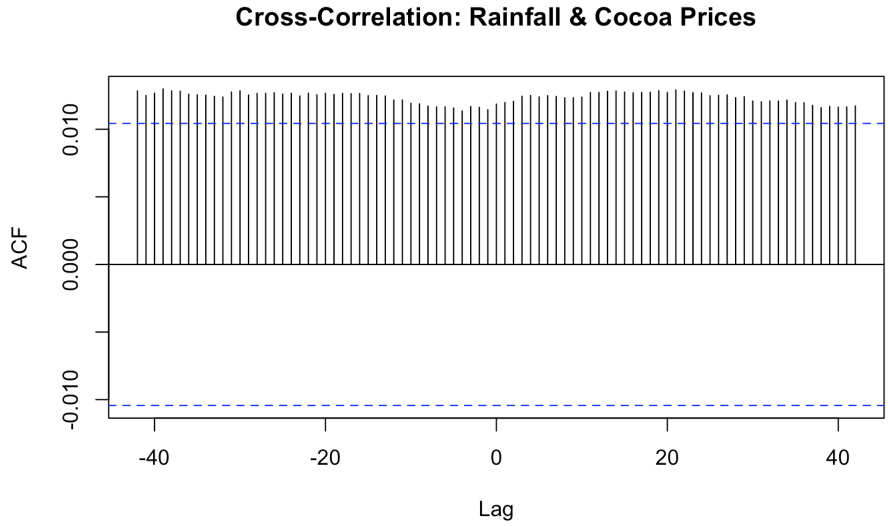


Figure 4.6: Cross-Correlation of Rainfall & Cocoa Prices

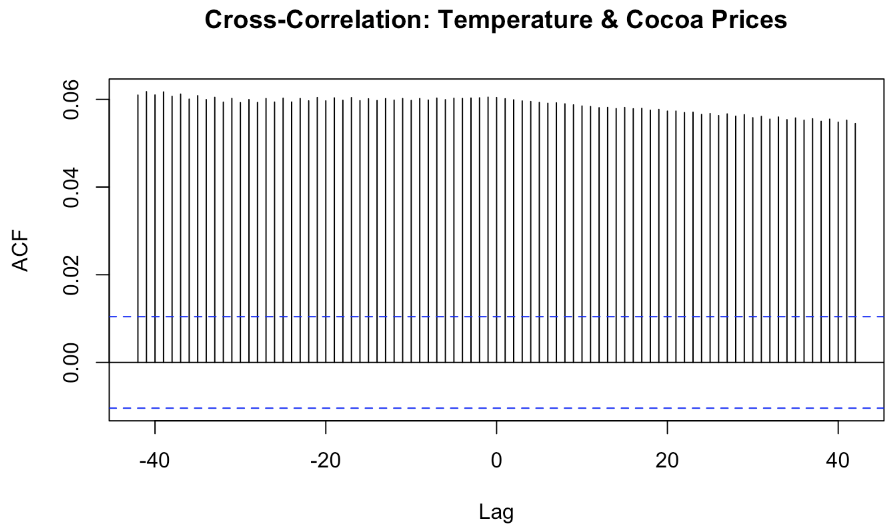


Figure 4.7: Cross-Correlation of Temperature & Cocoa Prices

## 4.4 Log-Transformation & Differencing

To detrend/deasonalize the dataset, the dataset was log-transformed (red plot) and compared to the original time series (blue). While the log-transformation to stabilize variance, the dataset was still non-stationary, as evidenced by the upward spike near the end. To ensure non-stationarity of the log-transformed model, an Augmented Dickey Fuller (ADF) test is conducted with the alternative hypothesis that the series is stationary. The ADF reported a p-value of 0.6634, which fails to reject the null hypothesis that the log-transformed series is stationary. The log-transformed series is then differenced to achieve stationarity. Performing an ADF test on the log-differenced series reports a p-value of 0.01, accepting the alternative hypothesis that the series is now stationary.

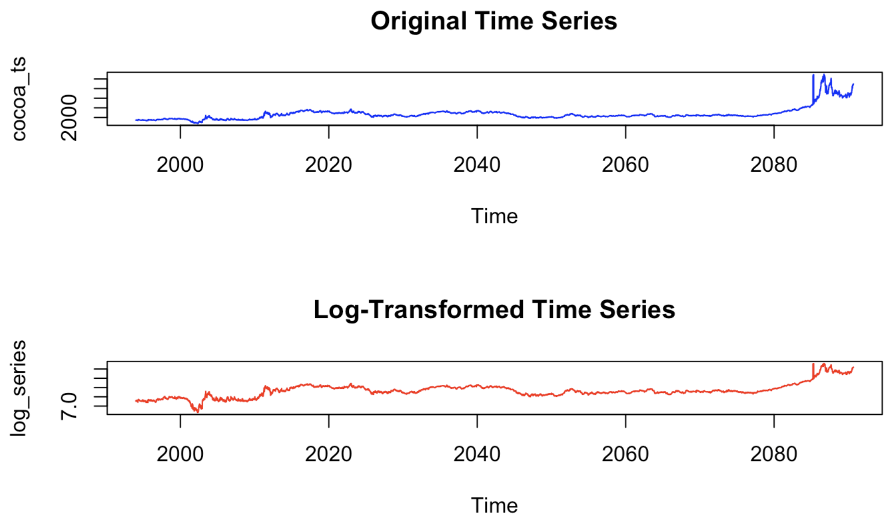


Figure 4.8: Log-Transformed Time Series

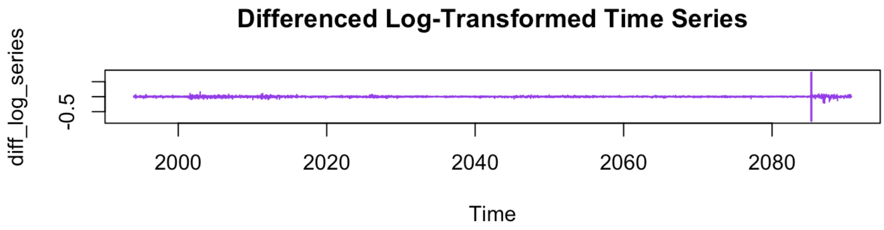


Figure 4.9: Differenced Log-Transformed Time Series

## 4.5 Model Building and Mathematical Formulation

**Baseline ARIMA Selection**:
Multiple ARIMA models, including ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(1,1,1), ARIMA(2,1,1), ARIMA(3,1,0), and `auto.arima()`, were tested on a training subset of the cocoa price time series. Performance was validated on a 12-month test set using RMSE, MAE, and AIC. ARIMA(3,1,0) outperformed all others, showing the lowest RMSE and satisfying residual diagnostics (i.e., no autocorrelation, white noise residuals). It was therefore selected as the benchmark model.
**ARIMAX Extension**:
To evaluate the predictive utility of external variables, we built ARIMAX(3,1,0) models incorporating:

- **Climate Models**: AvgTemp and TotalRainfall (current and lagged)
- **Economic Models**: CPI and PPI (current and lagged)

Each dataset was split into a training (all but last 12 months) and testing (last 12 months) set. The external regressors were standardized and fed into the ARIMAX models using the `forecast::Arima()` function.

### 4.5.1 ARIMA Model Specification

The general ARIMA$(p, d, q)$ model is defined as:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t \tag{4.1}$$

where:

- $Y_t$ is the differenced series after applying $d$ differencing operations.
- $\phi_i$ are the autoregressive (AR) coefficients.
- $\theta_j$ are the moving average (MA) coefficients.
- $\epsilon_t$ is white noise.

In our study, we selected the ARIMA(3,1,0) model based on forecast performance. This model implies:

- $p = 3$ (three autoregressive terms),
- $d = 1$ (first differencing to remove trend),
- $q = 0$ (no moving average term).

### 4.5.2 ARIMAX Model with External Variables

To extend ARIMA by incorporating exogenous variables, we use the ARIMAX model. Its formulation is:

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{k} \beta_j X_{j,t} + \epsilon_t \tag{4.2}$$

where:

- $X_{j,t}$ represents the $j$-th external regressor at time $t$ (e.g., CPI, PPI, AvgTemp),
- $\beta_j$ is the corresponding coefficient,
- Other terms as previously defined.

### 4.5.3 Model Development Steps

1. Data Preparation
   - Daily cocoa prices were aggregated to monthly averages (`AvgPrice`).
   - Climate data were aggregated into:
     - `AvgTemp`: Monthly average temperature,
     - `TotalRainfall`: Monthly total precipitation.
   - Lagged variables were created:
     - `AvgTemp_lag1`, `TotalRainfall_lag1` (climate),
     - `CPI_lag1`, `PPI_lag1` (economics).
2. Model Training
   - Time series was split with the last 12 months as test data.
   - ARIMAX(3,1,0) “Now” models used `CPI, PPI, AvgTemp, TotalRainfall`
   - ARIMAX(3,1,0) “Lagged” models used `CPI_lag1, PPI_lag1, AvgTemp_lag1, TotalRainfall_lag1`

### 4.5.4 Performance Metrics

Models were evaluated using:

- **RMSE**: Root Mean Square Error which penalizes larger errors.
- **AIC**: Akaike Information Criterion which balances model fit and complexity.
- **Ljung-Box Test**: Tests whether residuals are white noise.

Table 4.1: Model Performance Summary

| **Model** | **AIC** | **RMSE** | **Ljung-Box *p*-value** |
|---|---|---|---|
| Current Economic | 1771.59 | 1551.03 | 0.401 |
| Lagged Economic | 1843.32 | 2497.95 | 0.293 |
| Current Climate | 4286.36 | 3169.34 | 0.277 |
| Lagged Climate | 4285.77 | 3156.08 | 0.361 |

### 4.5.5 Interpretation

- Economic models outperform climate models in forecasting cocoa prices.
- Current economic variables are more predictive than lagged ones, suggesting immediate demand-side impact.
- Climate variables improve with a one-month lag, consistent with supply-side delays in crop production.
- Ljung-Box results show no significant autocorrelation, confirming residuals are white noise.

### 4.5.6 Conclusion

The ARIMAX(3,1,0) model effectively captures both time-dependent structure and external shocks. Incorporating macroeconomic and climate variables improves forecasting accuracy, especially when lags are applied appropriately. The findings suggest demand-side shocks have immediate influence, while supply-side effects manifest with a lag—valuable insight for stakeholders planning around cocoa market volatility.

# Chapter 5

# Forecasting and Results

## 5.1 Model Training and Validation Process

Daily cocoa price data was first aggregated into monthly averages to reduce noise and better capture long-term price dynamics. The resulting monthly time series was split into training and test sets, where the last 12 months were reserved as a holdout set for validation. This approach allowed for model fitting on historical data and objective evaluation on unseen future observations. Seven forecasting models were considered for comparison: ETS (Exponential Smoothing), ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(1,1,1) with drift, ARIMA(2,1,1), ARIMA(3,1,0), and an automated model selected using `auto.arima()`. The ETS model is well-suited to data with level, trend, and seasonality components, while the ARIMA models apply differencing and various combinations of autoregressive (AR) and moving average (MA) terms to capture temporal dependencies. ARIMA(3,1,0), which includes three autoregressive terms and first differencing but no moving average component, was found to produce the best forecasts on the test set. This model captures the extended influence of past price values, suggesting that cocoa prices are more dependent on their own lagged values than short-term shocks.

## 5.2 Performance Evaluation

Each model was evaluated using standard forecast accuracy metrics: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), and AIC (Akaike Information Criterion). RMSE penalizes large errors, MAE reflects the average magnitude of error, and AIC balances model fit with complexity. Forecasts were generated using the training data, and their performance was assessed against the test set.

The ARIMA(3,1,0) model achieved the lowest RMSE and MAE values among all tested models, and its AIC was also among the lowest. These results indicate that this model best captured the underlying structure of the cocoa price series with minimal overfitting. Additionally, residual diagnostics including the Ljung-Box test confirmed that the residuals of this model resembled white noise, indicating that it had effectively extracted all predictable patterns from the data.

## 5.3 Forecasted Values and Observed Patterns

Forecasts from each model were plotted against actual test set prices. The ETS model followed the broader trend but struggled to capture price shocks. ARIMA models with lower orders, such as ARIMA(1,1,0) and ARIMA(1,1,1), captured short-term dynamics but missed longer-term dependencies. In contrast, ARIMA(3,1,0) closely aligned with actual observations, producing more stable and accurate forecasts across the 12-month horizon. Visual inspection of the forecast intervals from ARIMA(3,1,0) further supported its reliability, as the intervals remained narrow and consistent across the forecast period. Residual plots showed no signs of autocorrelation or systematic bias, confirming the appropriateness of the model.

## 5.4 Summary

Through rigorous model testing and validation, ARIMA(3,1,0) emerged as the most accurate and robust forecasting model for cocoa prices in our dataset. Its autoregressive structure allowed it to effectively capture extended lags in price behavior, outperforming simpler ARIMA and ETS models. The model's strong test set performance and white-noise residuals reinforce its suitability for short-term cocoa price forecasting. As such, ARIMA(3,1,0) was used as the baseline model against which external economic and climate-informed models were compared. The non-lag, or current, models use contemporaneous monthly values of climate (average temperature and total rainfall) and economic indicators (Consumer Price Index and Producer Price Index) to forecast cocoa prices using the ARIMA(3,1,0) structure. These models were built to investigate whether real-time changes in external variables can immediately influence cocoa market prices.

### 5.4.1 Economic Model (Current Values)

The current economic model yielded an **AIC of 1771.594** and an **RMSE of 1551.032**, indicating strong model fit and forecasting accuracy. The relatively low AIC suggests that the model balances fit and complexity well, while the RMSE shows that forecast errors were moderate. The economic indicators used—CPI and PPI—are likely capturing demand-side forces such as consumer purchasing power and producer cost pressures. The strong performance of this model implies that global cocoa markets are quickly responsive to macroeconomic conditions, validating the importance of economic signals in commodity pricing. The **Ljung-Box p-value** for this model was statistically acceptable, indicating that residuals were approximately white noise and the model captured the majority of structure in the data.

### 5.4.2 Climate Model (Current Values)

In contrast, the current climate model showed **higher AIC (4286.364)** and **RMSE (3169.342)**. While it still adds explanatory value, its predictive performance was weaker compared to the economic model. This is not unexpected, as climatic effects on cocoa yield may require time to manifest—weather conditions in a given month may not influence prices until weeks or months later. As a result, the immediate (non-lagged) use of climate variables provides less forecasting power for short-term price prediction.

### 5.4.3 Interpretation and Implications

These results suggest that **economic variables tend to have a more immediate effect on cocoa prices**, potentially through their impact on demand and global market expectations. Conversely, **climate variables appear to influence prices with a delay**, likely through supply chain effects such as yield loss or harvest delays. The non-lag models serve as an important baseline for understanding the responsiveness of cocoa prices to external shocks. While current economic indicators proved highly useful for real-time price forecasting, the climate model's performance highlighted the need for lag structures to better capture delayed agricultural effects. In sum, the **current economic model** is effective in modeling demand-side shocks, while **lagged climate models** are better suited for capturing supply-side disturbances. These insights support a dual-approach modeling strategy when attempting to forecast commodity prices in the presence of diverse external forces.

# Chapter 6

# Conclusion and Recommendations

This study aimed to forecast cocoa prices using both univariate and multivariate time series techniques, with a specific focus on assessing the predictive value of external economic and climate indicators. Using a clean and carefully preprocessed monthly cocoa price dataset from the International Cocoa Organization, we developed a series of ARIMA-based models to identify the most accurate and interpretable approach to short-term forecasting.

The ARIMA(3,1,0) model emerged as the best-performing baseline model, achieving the lowest RMSE and AIC values among all candidate models, including ETS, ARIMA(1,1,1), and auto.arima. Residual diagnostics further confirmed the adequacy of the ARIMA(3,1,0) specification, as residuals appeared approximately normally distributed and free from autocorrelation, satisfying key assumptions for model validity.

To deepen the analysis, we extended the baseline model by incorporating both **contemporary** and **lagged** external regressors. Climate variables (average temperature and total rainfall) and economic indicators (CPI and PPI) were tested in separate ARIMAX(3,1,0) models. Results from these extensions suggest that:

- **Current economic variables** had a stronger predictive impact than their lagged versions, implying demand-side shocks tend to affect cocoa prices more immediately.
- **Lagged climate variables** slightly outperformed their non-lagged counterparts in terms of RMSE, supporting the hypothesis that weather conditions influence supply-side dynamics with a delay.

Ljung-Box tests confirmed that residuals across models were largely uncorrelated, indicating well-specified models. Visual inspection of forecast plots also showed that the economic models (especially the contemporaneous one) tracked real-world price movements more closely than the climate models.

By selecting ARIMA(3,1,0) as the core model and thoughtfully layering external indicators, we were able to build models that offer interpretability regarding the drivers of cocoa price volatility.

## 6.1 Recommendations

- **Integrate broader macroeconomic variables**—including exchange rates (USD/GHS, EUR/USD), inflation, interest rates, and global GDP growth—into future modeling efforts. These variables significantly influence cocoa markets. For instance, exchange rate volatility affects export competitiveness (Alori & Kutu, 2019), and inflation/interest rates shape investment behavior in agriculture (Bleaney & Greenaway, 2001).

- **Include climate and environmental indicators** such as rainfall variability, temperature anomalies, and disease incidence in cocoa-producing regions. This aligns with evidence that climate change is a major contributor to cocoa supply shocks (J.P. Morgan, 2024) and supports the inclusion of environmental variables in forecasting frameworks.
- **Adopt hybrid and multivariate models** such as ARIMAX or VAR to capture interactions between price and external drivers. These models allow for more robust forecasting, especially when macroeconomic or climatic conditions shift rapidly.
- **Explicitly model volatility using GARCH models**, especially for short-term forecasting. Previous studies (e.g., Kamu et al., 2010) show that GARCH(1,1) outperforms ARIMA when accounting for variance clustering common in commodity prices.
- **Expand the dataset's scope and granularity** by incorporating data from other major cocoa producers and variables like export volume, farm-gate pricing, and fertilizer costs. This can improve supply-side understanding and model responsiveness.
- **Incorporate policy and geopolitical risks**—such as trade barriers, global supply chain disruptions, or regional conflict—as structural breaks or exogenous shocks in models. These factors can significantly affect short-term pricing and should not be overlooked.

# Chapter 7

# References


- Oni O.V., Oni O.A., Akanle Y.O., Ogunleye T.B. (2021), Modelling Annual Cocoa Production Using ARIMA Time Series Model. African Journal of Mathematics and Statistics Studies 4(3), 135-144. DOI: 10.52589/AJMSS-0LPATDNK.

- Kamu, A., Ahmed, A., & Yusoff, R. (2010). Forecasting cocoa bean prices using univariate time series models. Journal of Arts Science & Commerce, 1(1), 71-80.

- Durmaz, Nazif; Kagochi, John (2022) : Asymmetric exchange rate pass-through in turkish imports of cocoa beans, Journal of Risk and Financial Management, ISSN 1911-8074, MDPI, Basel, Vol. 15, Iss. 4, pp. 1-11. https://doi.org/10.3390/jrfm15040184

- EF. Traoré & F. Badolo (2016): On the co-movement between coffee and cocoa prices in international markets, Applied Economics, DOI: 10.1080/00036846.2016.1148254. http://dx.doi.org/10.1080/00036846.2016.1148254

- Alori, H. A., & Kutu, A. A. (2019). Impact of exchange rates and price volatility on cocoa export in Nigeria. Journal of Risk and Financial Management, 12(4), 184. https://doi.org/10.3390/jrfm12040184

- Badolo, F., & Traoré, F. (2015). On the co-movement between coffee and cocoa prices in international markets. Journal of Economics and Sustainable Development, 6(17), 1–12. https://oar.icrisat.org/9505/

- Bleaney, M., & Greenaway, D. (2001). The impact of terms of trade and real exchange rate volatility on investment and growth in sub-Saharan Africa. Journal of Development Economics, 65(2), 491–500. https://doi.org/10.1016/S0304-3878(01)00147-X

- International Institute for Sustainable Development. (2018). Global market report: Cocoa. https://www.iisd.org/system/files/publications/ssi-global-market-report-cocoa.pdf

- J.P. Morgan. (2024, December 2). Rising cocoa prices: Will the chocolate crisis continue in 2025?https://www.jpmorgan.com/insights/research/cocoa-prices-2025

# Appendix A

# Model Selection Code

```r
# Load libraries
library(tidyverse)
library(lubridate)
library(forecast)
library(tseries)
library(Metrics)
library(ggplot2)

# Data Processing - Cocoa Prices
cocoa_data <- read.csv("Daily Prices_ICCO.csv", stringsAsFactors = FALSE)
cocoa_data$Date <- dmy(cocoa_data$Date)
cocoa_data$Price <- as.numeric(gsub(",", "", cocoa_data[[2]]))
cocoa_data <- cocoa_data %>%
  arrange(Date) %>%
  filter(!is.na(Price)) %>%
  mutate(Month = floor_date(Date, "month")) %>%
  group_by(Month) %>%
  summarise(AvgPrice = mean(Price, na.rm = TRUE)) %>%
  ungroup()

# Climate Data
ghana_data <- read.csv("Ghana_data.csv", stringsAsFactors = FALSE)
ghana_data$DATE <- ymd(ghana_data$DATE)
ghana_data <- ghana_data %>%
  drop_na(DATE, STATION) %>%
  mutate(PRCP = ifelse(is.na(PRCP), 0, PRCP)) %>%
  group_by(STATION) %>%
  mutate(
    TAVG = ifelse(is.na(TAVG), mean(TAVG, na.rm = TRUE), TAVG),
    TMAX = ifelse(is.na(TMAX), mean(TMAX, na.rm = TRUE), TMAX),
    TMIN = ifelse(is.na(TMIN), mean(TMIN, na.rm = TRUE), TMIN)
  ) %>%
  ungroup() %>%
  mutate(Month = floor_date(DATE, "month"))

monthly_ghana <- ghana_data %>%
  group_by(Month) %>%
  summarise(
    AvgTemp = mean(TAVG, na.rm = TRUE),
    TotalRainfall = sum(PRCP, na.rm = TRUE)
  ) %>%
  ungroup() %>%
  mutate(
    AvgTemp_lag1 = lag(AvgTemp),
```

```r
    TotalRainfall_lag1 = lag(TotalRainfall)
  )

# Economic Data (CPI and PPI)
cpi <- read.csv("CPIAUCSL.csv")
ppi <- read.csv("PCU3113513113517.csv")


cpi <- cpi %>%
  mutate(Month = floor_date(as.Date(observation_date), "month")) %>%
  rename(CPI = CPIAUCSL) %>%
  select(Month, CPI) %>%
  mutate(CPI_lag1 = lag(CPI))

ppi <- ppi %>%
  mutate(Month = floor_date(as.Date(observation_date), "month")) %>%
  rename(PPI = PCU3113513113517) %>%
  select(Month, PPI) %>%
  mutate(PPI_lag1 = lag(PPI))

# Merge with cocoa price data
econ_now <- inner_join(cocoa_data, cpi, by = "Month") %>% inner_join(ppi, by = "Month") %>% drop_na(CPI, PPI)
econ_lag <- econ_now %>% drop_na(CPI_lag1, PPI_lag1)

# Baseline Model

price_ts <- ts(cocoa_data$AvgPrice, start = c(year(min(cocoa_data$Month)), month(min(cocoa_data$Month))), frequency = 12)
ts_end <- end(price_ts)
train_ts <- window(price_ts, end = c(ts_end[1], ts_end[2] - 12))
test_ts <- window(price_ts, start = c(ts_end[1], ts_end[2] - 11))

model_list <- list(
  ARIMA_110 = Arima(train_ts, order = c(1,1,0)),
  ARIMA_011 = Arima(train_ts, order = c(0,1,1)),
  ARIMA_111 = Arima(train_ts, order = c(1,1,1)),
  ARIMA_211 = Arima(train_ts, order = c(2,1,1)),
  ARIMA_310 = Arima(train_ts, order = c(3,1,0)),
  ARIMA_auto = auto.arima(train_ts),
  ETS = ets(train_ts)
)

model_results <- lapply(names(model_list), function(name) {
  model <- model_list[[name]]
  fc <- forecast(model, h = length(test_ts))
  pred <- as.numeric(fc$mean)
  actual <- as.numeric(test_ts)
  rmse_val <- tryCatch(rmse(actual, pred), error = function(e) NA)
  mae_val <- tryCatch(mae(actual, pred), error = function(e) NA)
  aic_val <- tryCatch(AIC(model), error = function(e) NA)
  data.frame(Model = name, RMSE = rmse_val, MAE = mae_val, AIC = aic_val)
})

model_df <- do.call(rbind, model_results)
model_df <- model_df[order(model_df$RMSE), ]
print(model_df)

best_model_name <- model_df$Model[1]
best_model <- model_list[[best_model_name]]
fc_best <- forecast(best_model, h = length(test_ts))
```

```r
forecast_df <- data.frame(
  Date = cocoa_data$Month[(length(cocoa_data$AvgPrice) - 11):length(cocoa_data$AvgPrice)],
  Actual = as.numeric(test_ts),
  Forecast = as.numeric(fc_best$mean)
)

baseline_plot <- ggplot(forecast_df, aes(x = Date)) +
  geom_line(aes(y = Actual, color = "Actual"), linewidth = 1) +
  geom_line(aes(y = Forecast, color = "Forecast"), linetype = "dashed",
    linewidth = 1) +
  labs(title = paste("Forecast vs Actual -", best_model_name),
       x = "Date", y = "Cocoa Price", color = "") +
  theme_minimal()

print(baseline_plot)

# Climate Model Selection
climate_now <- monthly_ghana %>% select(Month, AvgTemp, TotalRainfall)
climate_lag <- monthly_ghana %>% select(Month, AvgTemp_lag1, TotalRainfall_lag1)

cocoa_clean <- cocoa_data %>% drop_na()

climate_cocoa_now <- inner_join(cocoa_clean, climate_now, by = "Month")
    %>% drop_na()
climate_cocoa_lag <- inner_join(cocoa_clean, climate_lag, by = "Month")
    %>% drop_na()

build_model <- function(df, x_vars) {
  df <- df %>% drop_na(all_of(c("AvgPrice", x_vars)))
  n <- nrow(df)
  n_test <- 12
  df_train <- df[1:(n - n_test), ]
  df_test <- df[(n - n_test + 1):n, ]

  list(
    y_train = ts(df_train$AvgPrice, start = c(year(min(df_train$Month)),
    month(min(df_train$Month))), frequency = 12),
    y_test = ts(df_test$AvgPrice, start = c(year(min(df_test$Month)),
    month(min(df_test$Month))), frequency = 12),
    x_train = as.matrix(df_train[, x_vars]),
    x_test = as.matrix(df_test[, x_vars]),
    dates = df_test$Month
  )
}
model_now <- build_model(climate_cocoa_now, c("AvgTemp", "TotalRainfall"))
model_lag <- build_model(climate_cocoa_lag, c("AvgTemp_lag1", "TotalRainfall_lag1"))

fit_now <- Arima(model_now$y_train, order = c(3,1,0), xreg = model_now$x_train)
fit_lag <- Arima(model_lag$y_train, order = c(3,1,0), xreg = model_lag$x_train)

fc_now <- forecast(fit_now, xreg = model_now$x_test, h = 12)
fc_lag <- forecast(fit_lag, xreg = model_lag$x_test, h = 12)

rmse_now <- rmse(model_now$y_test, fc_now$mean)
rmse_lag <- rmse(model_lag$y_test, fc_lag$mean)
```

```
results <- data.frame(
  Model = c("No Lags", "Lagged Climate"),
  AIC = c(AIC(fit_now), AIC(fit_lag)),
  RMSE = c(rmse_now, rmse_lag)
)

print(results)

# Plot
forecast_df <- data.frame(
  Date = model_now$dates,
  Actual = as.numeric(model_now$y_test),
  Forecast_Now = as.numeric(fc_now$mean),
  Forecast_Lag = as.numeric(fc_lag$mean)
)

ggplot(forecast_df, aes(x = Date)) +
  geom_line(aes(y = Actual, color = "Actual"), linewidth = 1) +
  geom_line(aes(y = Forecast_Now, color = "Forecast (No Lag)"), linetype =
     "dotted", linewidth = 1) +
  geom_line(aes(y = Forecast_Lag, color = "Forecast (Lagged)"), linetype =
     "dashed", linewidth = 1) +
  labs(title = "Climate Model Forecast vs Actual", y = "Cocoa Price", x =
    "Date", color = "Legend") +
  theme_minimal()

# Diagnostics
checkresiduals(fit_now)
checkresiduals(fit_lag)

# Merge economic data with cocoa
cocoa_clean <- cocoa_data %>% drop_na()
econ_now <- inner_join(cocoa_clean, cpi, by = "Month") %>% inner_join(ppi,
     by = "Month") %>% drop_na()
econ_lag <- econ_now %>%
  mutate(CPI = CPI_lag1, PPI = PPI_lag1) %>%
  select(Month, AvgPrice, CPI, PPI) %>% drop_na()

build_econ_model <- function(df, x_vars) {
  df <- df %>% drop_na(all_of(c("AvgPrice", x_vars)))
  n <- nrow(df)
  n_test <- 12
  df_train <- df[1:(n - n_test), ]
  df_test <- df[(n - n_test + 1):n, ]

  list(
    y_train = ts(df_train$AvgPrice, start = c(year(min(df_train$Month)),
    month(min(df_train$Month))), frequency = 12),
    y_test = ts(df_test$AvgPrice, start = c(year(min(df_test$Month)),
    month(min(df_test$Month))), frequency = 12),
    x_train = as.matrix(df_train[, x_vars]),
    x_test = as.matrix(df_test[, x_vars]),
    dates = df_test$Month
  )
}

model_econ_now <- build_econ_model(econ_now, c("CPI", "PPI"))
model_econ_lag <- build_econ_model(econ_lag, c("CPI", "PPI"))

fit_econ_now <- Arima(model_econ_now$y_train, order = c(3,1,0), xreg =
    model_econ_now$x_train)
```

```r
fit_econ_lag <- Arima(model_econ_lag$y_train, order = c(3,1,0), xreg = model_econ_lag$x_train)

fc_econ_now <- forecast(fit_econ_now, xreg = model_econ_now$x_test, h = 12)
fc_econ_lag <- forecast(fit_econ_lag, xreg = model_econ_lag$x_test, h = 12)

rmse_econ_now <- rmse(model_econ_now$y_test, fc_econ_now$mean)
rmse_econ_lag <- rmse(model_econ_lag$y_test, fc_econ_lag$mean)

results_econ <- data.frame(
  Model = c("Current Economic", "Lagged Economic"),
  AIC = c(AIC(fit_econ_now), AIC(fit_econ_lag)),
  RMSE = c(rmse_econ_now, rmse_econ_lag)
)

print(results_econ)

forecast_econ_df <- data.frame(
  Date = model_econ_now$dates,
  Actual = as.numeric(model_econ_now$y_test),
  Forecast_Now = as.numeric(fc_econ_now$mean),
  Forecast_Lag = as.numeric(fc_econ_lag$mean)
)

# Plot
library(ggplot2)
ggplot(forecast_econ_df, aes(x = Date)) +
  geom_line(aes(y = Actual, color = "Actual"), linewidth = 1) +
  geom_line(aes(y = Forecast_Now, color = "Forecast (Current)"), linetype = "dotted", linewidth = 1) +
  geom_line(aes(y = Forecast_Lag, color = "Forecast (Lagged)"), linetype = "dashed", linewidth = 1) +
  labs(title = "Economic Model Forecast vs Actual", y = "Cocoa Price", x = "Date", color = "Legend") +
  theme_minimal()

# Diagnostics
checkresiduals(fit_econ_now)
checkresiduals(fit_econ_lag)
```

# Appendix B

# Data Exploration Code

```
library(tidyverse)

# Load the cocoa futures price dataset
cocoa_data <- read.csv("/Users/camillyfuentes/Downloads/Daily Prices_ICCO.
    csv", stringsAsFactors = FALSE)

# Inspect the first few rows
head(cocoa_data)

# Check structure and summary of the data
str(cocoa_data)
summary(cocoa_data)

# Convert DATE column to Date format
cocoa_data$Date <- as.Date(cocoa_data$Date, format="%d/%m/%Y")

# Check for missing values
colSums(is.na(cocoa_data))

# Handle missing values (forward-fill missing prices if applicable)
cocoa_data <- cocoa_data %>% fill(everything(), .direction = "down")

# Remove duplicate rows
cocoa_data <- distinct(cocoa_data)

# Save cleaned cocoa price dataset
write.csv(cocoa_data, "cleaned_Cocoa_Prices.csv", row.names = FALSE)

# Print a message indicating completion
cat("Data cleaning completed. Cleaned data saved as 'cleaned_Ghana_data.
    csv' and 'cleaned_Cocoa_Prices.csv'\n")

#Clean Ghana data

# Load necessary libraries
library(tidyverse)

# Load the dataset
ghana_data <- read.csv("/Users/camillyfuentes/Downloads/Ghana_data.csv",
    stringsAsFactors = FALSE)

# Inspect the first few rows
head(ghana_data)
```

```r

# Check structure and summary of the data
str(ghana_data)
summary(ghana_data)

# Convert DATE column to Date format
ghana_data$DATE <- as.Date(ghana_data$DATE, format="%Y-%m-%d")

# Check for missing values
colSums(is.na(ghana_data))

# Handle missing values:
# - If PRCP is missing, assume 0 (no precipitation)
ghana_data$PRCP[is.na(ghana_data$PRCP)] <- 0

# - If temperature values (TAVG, TMAX, TMIN) are missing, fill with column mean
ghana_data$TAVG[is.na(ghana_data$TAVG)] <- mean(ghana_data$TAVG, na.rm = TRUE)
ghana_data$TMAX[is.na(ghana_data$TMAX)] <- mean(ghana_data$TMAX, na.rm = TRUE)
ghana_data$TMIN[is.na(ghana_data$TMIN)] <- mean(ghana_data$TMIN, na.rm = TRUE)

# Remove duplicate rows
ghana_data <- distinct(ghana_data)

# Save cleaned dataset
write.csv(ghana_data, "cleaned_Ghana_data.csv", row.names = FALSE)

# Print a message indicating completion
cat("Data cleaning completed. Cleaned data saved as 'cleaned_Ghana_data.csv'\n")

# Load necessary libraries
library(tidyverse)
library(lubridate)
library(zoo)

# Load the cleaned datasets
ghana_data <- read.csv("cleaned_Ghana_data.csv", stringsAsFactors = FALSE)
cocoa_data <- read.csv("cleaned_Cocoa_Prices.csv", stringsAsFactors = FALSE)

# Convert DATE columns to Date format
ghana_data$DATE <- as.Date(ghana_data$DATE, format="%Y-%m-%d")
cocoa_data$DATE <- as.Date(cocoa_data$Date, format="%Y-%m-%d")

# Ensure the datasets cover the full time range (1990 - latest available)
full_dates <- data.frame(DATE = seq(from = min(c(ghana_data$DATE, cocoa_data$DATE), na.rm = TRUE),
                                    to = max(c(ghana_data$DATE, cocoa_data$DATE), na.rm = TRUE),
                                    by = "day"))

# Merge datasets using full join to preserve all dates
merged_data <- full_dates %>%
  left_join(ghana_data, by = "DATE") %>%
  left_join(cocoa_data, by = "DATE")

# Handle missing values only if Cocoa_Price column exists
```

```
if ("Cocoa_Price" %in% colnames(merged_data)) {
  # Convert Cocoa_Price to numeric (handling possible character values)
  merged_data$Cocoa_Price <- as.numeric(merged_data$Cocoa_Price)
  # Forward-fill missing Cocoa Prices (if any)
  merged_data$Cocoa_Price <- na.locf(merged_data$Cocoa_Price, na.rm =
    FALSE, na.action = na.pass)
  # Backward-fill remaining missing Cocoa Prices (if any)
  merged_data$Cocoa_Price <- na.locf(merged_data$Cocoa_Price, fromLast =
    TRUE, na.rm = FALSE, na.action = na.pass)
}

# - Fill missing climate data with column means (if applicable)
climate_vars <- c("TAVG", "TMAX", "TMIN", "PRCP")
for (var in climate_vars) {
  if (var %in% colnames(merged_data)) {
    merged_data[[var]][is.na(merged_data[[var]])] <- mean(merged_data[[var
    ]], na.rm = TRUE)
  }
}
# Change name
merged_data <- merged_data %>%
  rename(Cocoa_Price = ICCO.daily.price..US..tonne.)

# Remove any remaining NA values before time series analysis
merged_data <- na.omit(merged_data)

# Save the cleaned and merged dataset
write.csv(merged_data, "merged_Ghana_Cocoa_data.csv", row.names = FALSE)

# Print confirmation
cat("Merging completed. Cleaned dataset saved as 'merged_Ghana_Cocoa_data.
    csv'\n")

# Load necessary libraries
options(repos = c(CRAN = "https://cloud.r-project.org/"))
install.packages("ggcorrplot")
library(tidyverse)
library(lubridate)
library(ggplot2)
library(forecast)
library(tseries)
library(zoo)
install.packages("ggcorrplot")
library(ggcorrplot)
install.packages("copula")
library(copula)
library(dplyr)

# 1 Data Overview
summary(merged_data)
str(merged_data)
colSums(is.na(merged_data))  # Checking for missing values

# 2 Time Series Visualization
# Plot Cocoa Prices over Time
ggplot(merged_data, aes(x = DATE, y = Cocoa_Price)) +
  geom_line(color = "blue") +
  labs(title = "Cocoa Prices Over Time", x = "Date", y = "Price") +
  theme_minimal()

# Moving Average (Rolling Mean) to Smooth Trends
```

```r

# Ensure Cocoa_Price is character before cleaning
merged_data$Cocoa_Price <- as.character(merged_data$Cocoa_Price)

# Remove commas, spaces, and any other non-numeric characters
merged_data$Cocoa_Price <- gsub("[^0-9.]","", merged_data$Cocoa_Price)

# Trim any leading/trailing spaces that may have been left after gsub
merged_data$Cocoa_Price <- trimws(merged_data$Cocoa_Price)

# Check for any non-numeric values before conversion
head(merged_data$Cocoa_Price)

# Convert to numeric
merged_data$Cocoa_Price <- as.numeric(merged_data$Cocoa_Price)
merged_data$Cocoa_Price <- na.approx(merged_data$Cocoa_Price, na.rm = FALSE)

merged_data$Rolling_Avg <- rollmean(merged_data$Cocoa_Price, k = 30, fill = NA)

ggplot(merged_data, aes(x = DATE)) +
  geom_line(aes(y = Cocoa_Price, color = "Original")) +
  geom_line(aes(y = Rolling_Avg, color = "30-day Rolling Avg")) +
  labs(title = "Cocoa Prices with Rolling Average", x = "Date", y = "Price") +
  #scale_color_manual(values = c("Original" = "blue", "30-day Rolling Avg" = "red")) +
  theme_minimal()

# 3 Seasonality Analysis
# Convert to time series
start_year <- year(min(merged_data$DATE, na.rm = TRUE))
start_month <- 1  # Ensuring consistent start

cocoa_ts <- ts(merged_data$Cocoa_Price, frequency = 365, start = c(start_year, start_month))

# Check for NA values in time series
sum(is.na(cocoa_ts))  # Should be 0 before decomposition

# Perform STL decomposition
decomp_stl <- stl(cocoa_ts, s.window = "periodic")
plot(decomp_stl)

# 4 Autocorrelation & Partial Autocorrelation
merged_data$Cocoa_Price <- as.numeric(merged_data$Cocoa_Price)
acf(merged_data$Cocoa_Price, main = "Autocorrelation of Cocoa Prices")
pacf(merged_data$Cocoa_Price, main = "Partial Autocorrelation of Cocoa Prices")

# 5 Correlation Analysis
numeric_data <- merged_data %>% select_if(is.numeric)
corr_matrix <- cor(numeric_data, use = "complete.obs")
ggcorrplot(corr_matrix, lab = TRUE, title = "Correlation Matrix")

# 6 Cross-Correlation with Climate Data
ccf(merged_data$PRCP, merged_data$Cocoa_Price, main = "Cross-Correlation: Rainfall & Cocoa Prices")
ccf(merged_data$TAVG, merged_data$Cocoa_Price, main = "Cross-Correlation: Temperature & Cocoa Prices")
```

```

# Print completion message
cat("Data Exploration Completed. Plots and Analysis Generated.\n")
```
```{r}
log_series <- log(cocoa_ts)
par(mfrow = c(2,1)) # Set layout for 2 plots
plot(cocoa_ts, main = "Original Time Series", col = "blue")
plot(log_series, main = "Log-Transformed Time Series", col = "red")

library(tseries)
adf.test(log_series)


diff_log_series <- diff(log_series)
plot(diff_log_series, main = "Differenced Log-Transformed Time Series",
    col = "purple")
```