Time Series Models for Forecasting Coffee Prices

Clarissa dos Santos Cardoso

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

May 2024

Supervisor:

Table of Contents

[Table of Figures 3](#_Toc165916774)

[Table of Tables 3](#_Toc165916775)

[Abstract 3](#_Toc165916776)

[Chapter 1: Introduction 3](#_Toc165916777)

[Chapter 2: Research Design 4](#_Toc165916778)

[Research Objectives 4](#_Toc165916779)

[Primary Data 5](#_Toc165916780)

[Model Selection 6](#_Toc165916781)

[Problem Identification and Clarification 6](#_Toc165916782)

[Ethical considerations: 6](#_Toc165916783)

[Chapter 3: Literature Review [extract from last semester proposal > need to be adjusted!!] 7](#_Toc165916784)

[Time Series and Commodity price prediction 10](#_Toc165916785)

[Time Series Components 10](#_Toc165916786)

[Benchmarck models 11](#_Toc165916787)

[Machine Learning (*baseline algorithms/static ML?)* 12](#_Toc165916788)

[Autoregressors (ARMA, ARIMA, SARIMA) 12](#_Toc165916789)

[Long-Term Short Memory 12](#_Toc165916790)

[Chapter 4: Methodology 13](#_Toc165916791)

[Data Collection - Secondary Data 13](#_Toc165916792)

[Data Description 15](#_Toc165916793)

[Data pre-processing and EDA 16](#_Toc165916794)

[Feature Selection: I-CIP 21](#_Toc165916795)

[Data imputation 22](#_Toc165916796)

[Statistical Tests 24](#_Toc165916797)

[Chapter 6: Results 24](#_Toc165916798)

[References 26](#_Toc165916799)

[Appendix 29](#_Toc165916800)

[Appendix 1: 29](#_Toc165916801)

# Table of Figures

[Figure 1 LSTM Architecture [available at https://www.cs.toronto.edu/~graves/preprint.pdf] pg 32. 13](#_Toc165918029)

[Figure 2 Boxplot of each variable for visualise distribution (elaborated by author) 20](#_Toc165918030)

[Figure 3 Price fluctuations from Feb 2023 to Feb 2024 20](#_Toc165918031)

[Figure 4 Monthly distributions of values 21](#_Toc165918032)

[Figure 5 Histogram of distribution (elaborated by author) 22](#_Toc165918033)

[Figure 6 Data interpolation comparison 23](#_Toc165918034)

# Table of Tables

[Table 1 Groups of Coffee of I-CIP 9](#_Toc165918011)

[Table 2 Share of global market used for indicator 10](#_Toc165918012)

[Table 3 Heading of Feb23 dataset (elaborated by author) 16](#_Toc165918013)

[Table 4 Heading of June23 data (elaborated by author) 17](#_Toc165918014)

[Table 5 Data type transformation from objects to floats (elaborated by author) 18](#_Toc165918015)

[Table 7 Null values after first merging of monthly datasets 31](#_Toc165918016)

# Abstract

Chapter 1: Introduction

# Chapter 2: Research Design

In the first section, it will display an overview of the data collection procedures followed by the identification and clarification of the problem that will guide the project to validate its relevance for the data science community and stakeholders involved. The research objectives and expected outcomes will be presented in the third section, followed by the ethical considerations that were observed during the process. *Among the goals of this research are: design, implement, and analyse different algorithms in which were found to be the most effective ones for this case study, based on the data available and its nature.*

## Research Objectives

The core objective is to evaluate various time series models to determine which algorithm most effectively predicts future coffee prices. This endeavour not only seeks to contribute to academic knowledge but also aims to provide insights that could benefit economists, traders, and policy makers in the coffee industry in the short term. The research questions guiding this study include: What is the most effective time series model for predicting I-CIP coffee prices? And *do neural networks provide a superior forecasting capability compared to other supervised machine learning models, in terms of accuracy and implementation?*

The research objectives for this study are as follows:

* Utilize data cleaning techniques to prepare the dataset for exploratory data analysis. This pre-processing will include handling missing values, outliers, and any inconsistencies to ensure data quality.
* employ exploratory data analysis to uncover underlying patterns, such as trends and seasonality, and assess statistical properties, including normality and stationarity.
* Implement feature engineering to select and manipulate the raw data to enrich the dataset and enable the application of predictive machine learning algorithms. This will involve creation of new features, such as lagged values of indicator prices, to capture the temporal dependencies and patterns; data transformation by normalising values to smaller scales or differentiation methods that could improve the models' ability to capture the dynamics in the data. With these techniques in place, apply time series machine learning algorithms to forecast future indicator values for 1 day, 1 week and 1 month.
* Explore the effect of parameter variations on the forecast accuracy across 3 different modelling approaches: a) supervised learning regressive algorithms, such as Linear Regression and Random Forest Regressor, b) autoregressive statistical models (SARIMA), and c) neural networks (LSTM). Visualize the performance of each model in training and testing sets to compare how traditional statistical methods perform against deep learning approaches in terms of forecasting accuracy for I-CIP prices. Using MSE, MAE and RMSE as metrics to compare the models.

## Primary Data

The primary data for this study will be obtained from the experimentation through application of several supervised time series models to historical values of coffee traded in the global commodity stock market based on the International Composite Indicator Price (I-CIP). The experimentation method was selected for representing a standard practice in research to investigate the relationship between variables by implementing systematic processes to evaluate the performance of different machine learning models. The creation of new data will originate from planning, conducting, and analysing the experiments’ outcomes (Fontana et al., 2023). As stated by Saunders et al (2009), the purpose of doing an experiment is to study cause and effect on a selected sample, for example if by changing one independent variable it could generate changes in another dependent variable (Saunders et al., 2009).

Authors also suggest that when conducted in closed laboratories rather than open filed or real life work applications have the tendency to provide a greater control over the research process such as sample selection and the context within which the experiment occurs. To study the dynamics and to attempt a prediction of future movements of the prices in the market, it was originally intended to access data from a period of 1 to 5 years of daily prices, however the research is reliable to the extent of free raw data available at the time of data collection.

### Model Selection

Several time series forecasting models will be applied to the historical data to study the dynamics and predict future price movements. The models selected for comparison will include:

* Supervised machine learning as a benchmark for comparison (Linear Regression, Random Forest and Support Vector Machine)
* ARIMA (Autoregressive Integrated Moving Average): For its efficacy in modelling time-dependent structures.
* SARIMA (Seasonal ARIMA): To account for seasonal variations in the data.
* LSTM (Long Short-Term Memory networks): A deep learning approach suitable for capturing long-term dependencies in sequence prediction.

The comparative experiment intends to analyse the effectiveness of each model measured by the following statistical metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics will help quantify the accuracy of predictions against actual observed values.

Furthermore, model validity will be tested through cross-validation, rolling average of forecasting results to mimic real-world application and assess performance stability over time following the methods similar studies on commodity price prediction have applied based experimentation to answer the hypothesis if neural networks present superior results in this field (Erasmus Kabu Aduteye et al., 2023; Fontana et al., 2023; Hwase and Fofanah, 2021). The expected results should contain forecasted prices for 1 day, 1 week and 1 month future values.

### Problem Identification and Clarification

The motivation for this research originates from a specific interest in comparing the efficacy of different machine learning algorithms in forecasting. The core question driving this investigation is whether neural networks (LSTM) provide a superior forecasting capability compared to other machine learning models, particularly in terms of accuracy and reliability across varying market conditions?

## Ethical considerations:

Given that the data used in this study are publicly available and aggregated without any personal identifiers, the main ethical consideration is ensuring the accuracy and reliability of the information presented. There is the acknowledgement that any misinterpretation of data, overstatements or even assumptions regarding the models' predictive competences should be interpreted as a academic exercise, and therefore not be used as a real-world application. It is recommended that the simulations presented in this study may be considered in similar studies on predictive analysis for comparison misguided decisions or impact on the future studies that could derive from the insights presented here (Saunders et al., 2009).

In addition, the study is committed to maintain fairness to the data collection process taken by assigning the rightful credits to the source of secondary data. With a dataset gathered from the database provided by the International Coffee Organisation (ICO) based on their pledge to distribute accurate and reliable data (ICO, 2024c). All data used for this project was publicly available at the moment of its collection and manipulation and does not compromise the confidentiality or privacy of any entity. Direct contact was made to the source to ensure the fairness and rightful application of data.

### - Validity type <?>

# Chapter 3: Literature Review [extract from last semester proposal > need to be adjusted!!]

*Most of the literature review will consist of journal articles/internet publications (blog posts?) of related fields of forecasting models in stock market from a 5 years range, with additional theoretical books and articles to provide a higher level of background theories to related to the research. The selection of articles published in less than 5 years because they provide an updated perspective of how the theory is being applied in current days, considering the high volatility in the commodity stock market and the constant advances in data science combined with new technologies that are launched and updated constantly. These are efforts to create a study that is relevant and up to date to what is published and avoid outdated information. The models that will be the focus of the experimentation are: Random Forest, SARIMA and LSTM, which are mentioned in this section after a brief explanation of the indicator price used as the main variable for the modelling.*

ICO Composite Indicator Price (I-CIP)

A composite indicator is a measurement that combines multiple individual indicators into a single index, usually to capture an overall concept or the performance of a sector (ICO, 2021; Kumari and Swarnkar, 2020). The I-CIP for green coffee is calculated by the International Coffee Organization (ICO)[[1]](#footnote-1) through the arithmetical mean of the weighted average of daily prices of different coffee groups: “Colombian Milds”, “Other Milds”, “Brazilian Naturals”, and “Robustas”. These groups were established under the Coffee Agreement of 1962 based on the two main types of coffee traded worldwide of coffee beans (arabica and robusta). As stated by Faith (2021), Hwase and Fofanah (2021) coffee is not a homogenous product, due to its genetic diversity, growing conditions, processing methods and even the various harvesting techniques used to its manufacturing can significantly affect quality and taste, and by consequence resulting in different trading rates. The classification based on distinct manufacturing processes and country of origin are displayed in table 1 (ICO, 2021). The table also shows the weights each group in the overall composition of the index based on the volume they are traded in the international market (Fatih, 2021; ICO, 2024b).

The robusta is a type of coffee produced from a tree of the botanical species *Coffea canephora* mainly farmed and exported by African countries, representing 37% of the composite, while the arabica type is divided into three other subgroups according to the type of arabica coffee produced and the type of manufacturing process they are subjected (ICO, 2024b). The distinctive attributes of each subgroup of arabica is due to the processing method used, either wet or dry, defined by the ICO(ICO, 2021). The first is also named “washed” process where farmers obtain the green coffee by washing the harvested fruits and according to ICO (2024), Mild Arabicas, including Colombian Arabicas, are produced by this method. Whereas the dry method, which also can be called “unwashed” or “natural process” are produced by drying coffee cherries without the intervention of water or machines to remove remains fruit from the beans (ICO, 2024b).

Table Groups of Coffee of I-CIP

|  |  |  |  |
| --- | --- | --- | --- |
| **Coffee Group** | **Types of coffee** | **Country of Origin** | **Weight in Composite Indicator calculation** |
| Colombian Milds | Arabica | Colombia, Kenya, Tanzania | 12% |
| Other Milds | Arabica | Bolivia, Burundi, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, India, Jamaica, Malawi, Mexico, Nicaragua, Panama, Papua New Guinea, Peru, Rwanda, Venezuela, Zambia, Zimbabwe | 21% |
| Brazilian Naturals | Arabica | Brazil  Ethiopia  Paraguay | 30% |
| Robustas | Robusta | Angola, Congo, D.R. of, Ghana, Guinea, Indonesia, Liberia, Nigeria,  OAMCAF (Benin, Cameroon, Central African Rep., Congo, Cote d'Ivoire, Equatorial Guinea, Gabon,  Madagascar, Togo), Philippines, Sierra Leone, Sri Lanka, Thailand, Trinidad and Tobago, Uganda, Vietnam | 37% |

The ICO uses the composite index to provide a broad perspective on global coffee market trends where coffee authorities can establish prices for payments to farmers in producer countries, and can also be used by other specialists and governments worldwide to elaborate statistical reports on this subject (Babu and Muniyappa, 2021),(ICO, 2024b), (ICO, 2024c)*.* For reference, the I-CIP is calculated based on an aggregate value of production gathered from the US and EU (FRA, GER )markets of the four groups, as illustrated in table 2 based on ICO’s rules for price calculation (ICO, 2021).

Table Share of global market used for indicator

|  |  |  |
| --- | --- | --- |
| **Coffee Group** | **Share of USA market** | **Share of EU market** |
| Colombian Milds | 57% | 43% |
| Other Milds | 39% | 61% |
| Brazilian Naturals | 27% | 73% |
| Robustas (mostly derived from African countries) | 18% | 82% |

## Time Series and Commodity price prediction

One of the underlying goals of time series forecasting as defined by Guido and Müller (2016) is to “learn from the past and predict for the future”, emphasizing the importance of historical data to shape future expectations. Authors express this type of data analysis requires data points to be organised on a sequential time-based manner (such as days, months or years) since that characteristic is what drives the identification of specific behaviours and underlying patterns to draw insights that help making predictions (Guido and Müller, 2016; Petropoulos et al., 2022).

Such insights are essential for selecting and fitting an appropriate time-series model, however this task is usually a challenge for many data scientists, as pointed by Bajaj (2022), Brunsdon and Smith (1998) due to the differences between sequential data to static data. In the first, each data point is related to its previous points on a time-dependent basis and the second consists of independent data points not influenced by a temporal sequence (Bajaj, 2022a; Brunsdon and Smith, 1998). Algorithms designed specifically for time series forecasting are able to capture the components of sequential data such as trends, cycles, or seasonal fluctuations better than static machine learning algorithms, since they lack the ability to manage this dynamic behaviours (Bajaj, 2022).

### Time Series Components

To assist interpretation of data behaviour in timeseries analysis, Hyndman and Athanasopoulos (2018) declare there are key components involved: trend, seasonality and residuals. The trend component indicates long-term behavior, where the data may show a positive or negative trend depending on whether values increase or decrease over time (Bajaj, 2022a).

Seasonality refers to predictable and periodic patterns that repeat at consistent intervals After accounting for trend and seasonality, the remainder—also known as residual—consists of the data left over, which is useful for detecting anomalies in the time-series. Cycles represent fluctuations occurring at irregular intervals, distinguishing them from seasonal patterns which have fixed periodicity. Lastly, a time-series is considered stationary if its statistical properties such as mean and variance remain constant over time, and its covariance does not depend on the time at which it is measured. These components together facilitate a comprehensive analysis, allowing for a deeper understanding of the underlying dynamics and aiding in accurate forecasting and anomaly detection.(Bajaj, 2022a; Brownlee, 2017; Dokumentov and Hyndman, 2022)

A common techique to extract these components is the seasonal-trend decomposition (STL) using LOESS[[2]](#footnote-2) based on the frequency specified. For seasonal component the default is 7, representing 7 days or 1 week, which should be adjusted accoding to the context of each data (Josef Perktold et al., 2024; Petropoulos et al., 2022).

The commodity market is a classic subject of reseach in the domain of forecasting due to its high instability and volatility that require complex methods to achieve accurate predictions, and as noted by Erasmus Kabu Aduteye et al. (2023), Hwase and Fofanah (2021), one of the challenges is to maintain a centered approach for selection and manipulation of the models to achieve more accurate and reliable results (Erasmus Kabu Aduteye et al., 2023; Fatih, 2021; Fianu, 2022; Hwase and Fofanah, 2021). The next subsections will provide more details on the

### Benchmarck models

*<elaborate on more theory on each model used, >*

As proposed by Hwase and Fofanah (2021), the simplicity of traditional machine learning algoriths such as Linear Regression, despite their easier understanding and less time consumning to implement, are considered to be a good start for modeling selection to calculate the best fitting line of regression and infer basic predictions. However authors also suggest that the model has a tendency to underfit, as in adapting most of the train data for creating unseen data by assuming the linear relationship between variables, which is not usually what happens in complex data, like commodity prices (Hwase and Fofanah, 2021). This was also observed by Fianu (2022) when experimentation between benchmark models proved have inferior performance when compared to decomposition-time based models such as ARIMA and SARIMA (Fianu, 2022). Considering a supervised learning approach, Kumari and Swarnkar (2021) have implemented Support Vector Machines (SVM) for their classification problem and obtained less accurate results when comparing with ANN accuracy (Kumari and Swarnkar, 2021). Although, in terms of regression tasks, SVMs are not the most suitable for application but as tested by (Beniwal et al., 2023) is the regressor version (support vector regressor) that can be used for both linear data and nonlinear.

Unlike the models that assume a direct linearity between data points, tree based models, such as random forest, have a different approach as detailed by Guido and Müller (2016), since they can usually better process non-linearity and robust outliers. However authors also suggest it can be computationally intensive and less interpretable than linear regression (Guido and Müller, 2016).

### 

### Autoregressors (ARIMA, SARIMA)

### Long-Term Short Memory

The choice of ANN is advised by several articles in literature which have shown that Neural Network models outperform some traditional statistical models in modelling meteorological data (Nwokike et al., 2020), energy usage (Fianu, 2022) and is widely used across economic, financial and stock market forecasting (Beniwal et al., 2023; Kumari and Swarnkar, 2021; Zhu, 2022). One type of neural network is the LSTM, which is defined by Hwase and Fofanah(2021) as a “type of time recurrent neural network, suitable for processing and predicting the events of interval and a long delay in time series”. In terms of commodity price prediction focused on coffee, the highlight is on the experiments that compare the application of LSTM to traditional machine learning models and statistical models (Babu and Muniyappa, 2021; Erasmus Kabu Aduteye et al., 2023; Fatih, 2021; Hwase and Fofanah, 2021).

One of the reasons why LSTM is useful for prediction problems is that it can store important past information and forget information that do not add value, Graves (2012) and Singh (2021) determine in figure 1 the architecture of LSTMs as composed by three gates or layers:

• Input gate: adds information to the cell state

• Hidden Layer: removes the information no longer required by the model

• Output gate: selects the information to be shown as output

Figure LSTM Architecture [available at https://www.cs.toronto.edu/~graves/preprint.pdf] pg 32.

A diagram of a network

Description automatically generated

# Chapter 4: Methodology

## Data Collection - Secondary Data

One of the restrictions from accessing the data from this channel was that the ICO only makes public the data from the current month up until the same period from the previous year. Considering the time constraints present in undertaking the research, the secondary data was collected from the period of ‘February of 2023’ to ‘February of 2024’ due to the availably of free data in ICO’s web page (ICO, 2024a). This resulted in a smaller dataset than what was expected at the initial stages of research design, which was considering at least 3 years’ worth of data to have a more comprehensive pattern over years.

#### Data Collection Issues and Limitations

An attempt to gather data from a longer period of at least 2-3 years to further investigate the patterns and cyclical behaviours to increase training dataset was not possible due to the time sensitivity between making the request for additional data, having the risk being denied. As pointed by Svolba (2022), the length of historical data can influence the quality of forecasting future values, on average, the longer history available, the better predictions the models can return. However, on the simulations performed by his study, it was also noted that having too many cycles could negatively influence the predictions by generating biased predictions. This situation was also observed by Hyndman and Athanasopoulos (2018), when using very long time series (Svolba, 2022) This also added to a new risk of a possible delay to conclude the study, between collecting, processing the data and performing the experimentation artifact and the added computational costs a longer dataset could increase processing and training time. This adjustments were kept to ensure the practical boundaries to execute the project.

To obtain further historical data, the ICO has a private plan for subscribing to their entire database which was not judged necessary at this stage due to financial constraints. Direct contact was made to enquire about access for academic purposes, which was met with a limitation that the data provided would not be kept at another’s institution database for data protection, ethical concerns and the cost of storing the original data. This is why it was stablished one year of data would be used for the final experiment collected from open source channels. Under the International Coffee Agreement 2007 for all data collected, transmitted, calculated and publicised by Statistics Committee (ICO, 2021).[[3]](#footnote-3) The committee confirmed via email that any members of the public are free to use the data published on the “Public Market Section” as long as the organisation is accredited in the study as the original source (ICO, 2024a).

## Data Description

This dataset includes daily values of four main coffee beans groups that are used as standard to calculate the final indicator: ‘Colombian Milds’, ‘Other Milds’, ‘Brazilian Naturals’, and ‘Robustas’ as well as the composite indicator calculated through different weights based on patterns of trade. Each price is expressed in US cents per pound of green coffee (ICO, 2021). The period of data ranges from February 2023 to February 2024 however, as stated on research design chapter, it would have been preferable to have a longer range of data, in order to observe more than one cycle to see with more detail the ups and downs of the prices over the years. A more extensive dataset could enable an in-depth analysis of market trends, seasonality, and price volatility over the years as a general rule of thumb for time series is that the more historical data used, the more a forecast can be improved, as stated by Svolba (2022) and Petropoulos et al. (2022).

However, in terms of fast changing and dynamic environments, such as the commodities market, having to many years back in historical inputs could also make pose as a challenge for accurate forecasting because it could mask the recent history fluctuations, by relying on older data as illustrates Svolba(2022), great economic recessions from past decades could influence the predictions by downgrading the overall average of prices ) (Svolba, 2022). As pointed by Hyndman and Athanasopoulos (2018), there is not a specific guideline on how exactly how many data points are required for a timeseries modelling, but they also say that by increasing the data points, the noise may also increase and that overall the theoretical limit is having more observations than parameters for a forecasting model (Hyndman and Athanasopoulos, 2018). The authors suggest that a shorter total of observations (around 200) often result in reasonable models as they offer an approximation of the generated data to compare to real data. This idea is complemented by Zhang et al (2023), due to inequality of the actual impact of past trends to future prices. They suggest that predictions of future prices could be more significantly accurate when analysing more recent price changes over focusing on previous past observations (Zhang et al., 2023).

## Data pre-processing and EDA

Once the secondary data was collected, each csv file contained daily values of i-cip prices divided by month, in total there was 13 files imported to a Jupyter notebook to build the experiment using Python language and its available open source libraries. The first data pre-processing technique used was to combine all csv files into the same dataframe in order to advance the analysis, which had to be sorted in chronological order to respect the daily sequence needed to perform any type of time series modelling. As stated by Brownlee (2017) and Petropoulos et al (2022) respecting the temporal order of observations is the premises of time series to ensure the models are able capture the relationship between past values and the characteristic patterns in the data, since they rely on historical observations to make accurate predictions about future values (Brownlee, 2017; Petropoulos et al., 2022). Disrupting the sequential order of data could lead to incorrect interpretations of patterns and relationships, potentially compromising the integrity and reliability of the analysis. Furthermore, the sequential nature of the data prevents leakage of future information during training and validation sets (Erasmus Kabu Aduteye et al., 2023).

The datasets were scanned to display basic features from heading, to observe names of columns, how they are shaped, data types and identifying missing values. Each one of the 13 datasets have data distributed across 6 variables:

* Date
* ICO Composite
* Colombian Milds
* Other Milds
* Brazilian Natural
* Robusta

By initially merging all monthly data into a single dataframe, It was observed that, despite having the same source, not all files had the same data structure, which caused a misplacing of values that had to be addressed. After analysing the heading for all files separately, it was discovered that where the indicator values had two different names: “I-CIP” and “ICO Composite” for different months as displayed in tables 3 and 4.

From Feb23 to May 23, the values were stored in the following features:

Table Heading of Feb23 dataset (elaborated by author)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Unnamed: 0** | **Unnamed: 1** | **Colombian** | **Unnamed: 3** | **Brazilian** | **Unnamed: 5** |
| **0** | NaN | NaN | NaN | Other Milds | NaN | Robustas |
| **1** | NaN | Indicator (I-CIP) | Milds | NaN | Naturals | NaN |
| **2** | 01-Feb | 171.43 | 235.92 | 223.22 | 191.65 | 102.31 |
| **3** | 02-Feb | 172.50 | 237.34 | 226.26 | 192.86 | 102.00 |

And the remaining months from June23 to Feb24 had a different heading, also containing typos in in the column’s creating errors as table x shows:

Table Heading of June23 data (elaborated by author)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Unnamed: 0** | **Unnamed: 1** | **Colombian** | **Unnamed: 3** | **Brazilian** | **Unnamed: 5** |
| **0** | NaN | I-CIP | NaN | Other Milds | NaN | Robustas |
| **1** | NaN | NaN | Milds | NaN | Naturals | NaN |
| **2** | 01-Jun | 173.56 | 220.40 | 215.27 | 181.78 | 126.54 |

This resulted in a combined dataset with overlapping values and created null values in the “I-CIP” column from the first four months and the “ICO Composite” showed null values for the remaining period. A visual representation of the different column names and null values is displayed on appendix 1. As stated by Hyndman and Athanasopoulos (2018), this type of error is originated from the assumption that all datasets followed the same structured which proved to be wrong and the incompatible column names resulted in empty cells (Hyndman and Athanasopoulos, 2018)

The issue was fixed by first renaming the first four months of the combined dataframe using pandas ‘rename’ function. Then, the *NaN* values from the first two rows were removed as that were as they represented non-essential data, and were found to not impact the experiment. These rows composed extra, empty data rather than missing values for any of the categories, fitting the scenario explained by Harrison (2023) where data can be dropped simply because it is missing by chance, and that does not correlate with observed or unobserved variables. This approach aligns with best practices for data management present in Guido and Müller's (2016) works where they say this type of data removal is acceptable if it represents a small percentage and its absence does not result in further errors or bias the outcomes. Authors also suggest the dropna as a method for row management based on the characteristics of missing data, but in specific cases where columns are considered to be dropped, this practice should be performed after basic exploratory data analysis to ensure the values are not adding value to the analysis (Guido and Müller, 2016; Harrison, 2023; Pandas, 2023). From the output of pandas’ “info” function to see the dataset is arranged by six columns fully populated with non-null entries across 279 observations. However only the first column containing the dates had the expected datatype (datetime), as illustrated in table x. The other columns represent the prices of the indicator and each of the four groups of coffee beans that compose the index are recorded as objects. Data transformation had to be done to ensure the correct numeric data is allocated across the entire dataset to enable appropriate modelling. Most statistical and machine learning models require numeric data inputs for making forecasts or predictions, as well as to analyse the correlation between variables. Numeric data also supports efficient processing and visualization, allowing for more accurate and insightful results between datapoints, including the understanding of the statistical nature of the dataset as displayed in table 5 (Guido and Müller, 2016).

Table Data type transformation from objects to floats (elaborated by author)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **#** | **Column** | **Non-Null Count** | **Original Datatype** | **Datatype after tranformation** |
| 0 | *date* | 279 non-null | datetime64[ns] | datetime64[ns] |
| 1 | *I-CIP* | 279 non-null | object | float64 |
| 2 | *colombian\_milds* | 279 non-null | object | float64 |
| 3 | *other\_milds* | 279 non-null | object | float64 |
| 4 | *brazilian\_nat* | 279 non-null | object | float64 |
| 5 | *robustas* | 279 non-null | object | float64 |

From the descriptive statistics displayed in table x, the i-cip has an mean value of 167.80 usd. The prices range from as low as 145.99usd to as high as 187.29usd showing some fluctuation over the 1 year of data collected so far. Most of the prices (from the 25th to the 75th percentile, or interquartile range) are between 158.50 usd and 176.95 usd, with the median of 170 usd. This suggests that while the majority of prices can be clustered around the average, there are some data points (Brunsdon and Smith, 1998) .

Figure Descriptive analysis of I-CIP

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **I-CIP** | **colombian\_milds** | **other\_milds** | **brazilian\_nat** | **robustas** |
| **count** | 279 | 279 | 279 | 279 | 279 |
| **mean** | 167.806344 | 208.039391 | 205.646452 | 175.466631 | 125.725233 |
| **std** | 10.987927 | 18.839311 | 16.566665 | 15.712895 | 14.122864 |
| **min** | 145.99 | 178.82 | 174.97 | 147.66 | 101.52 |
| **25%** | 158.515 | 191.51 | 192.21 | 160.72 | 117.625 |
| **50%** | 170.11 | 206.31 | 204.1 | 179.98 | 124.47 |
| **75%** | 176.955 | 225.185 | 220.645 | 188.37 | 131.7 |
| **max** | 187.29 | 249.04 | 242.71 | 207.45 | 158.78 |

Further exploratory data analysis was executed at the early stage through basic visualisations to verify the presence of outliers and to better understand how the data is distributed. As defined by Bajaj (2022), outliers are observations that diverge so much from other observations and raise doubts whether the anomaly is a valid data or if was generated by an external mechanism (Bajaj, 2022b). They can impact the increase of noise and damage the interpretation of results or further decision making during and after experimentation phase (Garza, 2023). This allows to observe if there are any consistent patterns, trends, relationships between variables. An additional practice used for anomaly detection in timeseries analysis, as suggested by Bajaj (2022b) and Brownlee (2017) is to separate the dataset into their components of trend and seasonality which will be discussed in further sections.

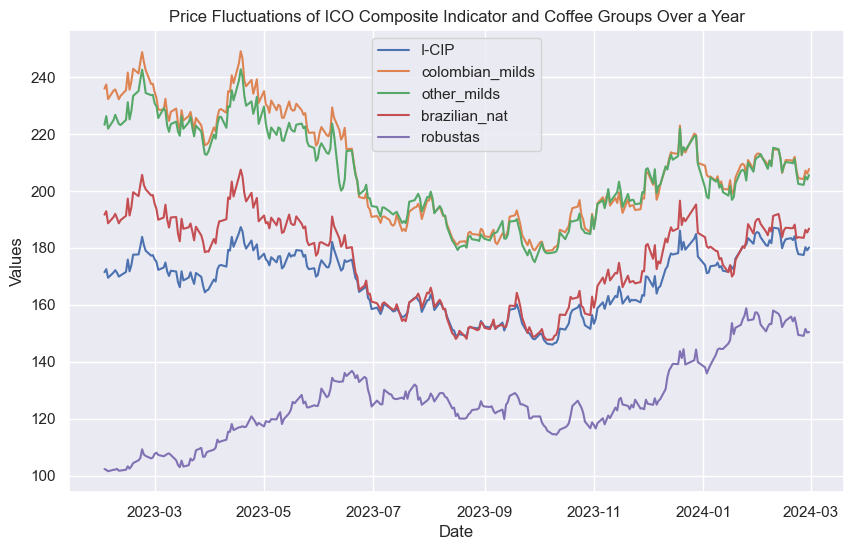
As mentioned by Garza (2023), the exploratory stage is vital to knowing which methods to implement and evidencing those methods' results. From visual representation of boxplots of the descriptive statistics, the 'I-CIP' feature has the narrowest interquartile range, indicating less variability in its values compared to the other categories. 'Robustas' not only have visible outliers but also has the lowest median value and appears to have the most variability. The Brazilian naturals has their median below the Colombian and Other Milds, but above 'I-CIP' and 'robustas'.

Figure Boxplot of each variable for visualise distribution (elaborated by author)

A graph with colored squares

Description automatically generated with medium confidence

Figure Price fluctuations from Feb 2023 to Feb 2024



### Feature Selection: I-CIP

To simplify the EDA and modelling stages, using a univariate timeseries data was defined as appropriate

When comparing across months, the distribution of prices in the later months (like November and December) differs from the earlier months (such as February and March), which could imply changes in market conditions or supply and demand dynamics throughout the year.

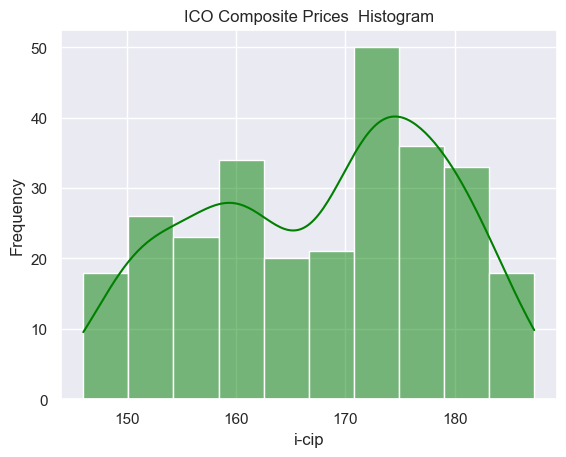
* Price Volatility: The varying heights of the boxes indicate that price volatility changes throughout the year. Some months, like February, March, and April, show greater variability in I-CIP prices, while others, such as August and September, appear more stable.
* Outliers: There are a few outliers, particularly in months like September and November. These could indicate unusual market events or data anomalies.
* The medians of the boxes show a cyclical pattern, suggesting that I-CIP prices may have a seasonal component, with certain months typically having higher or lower prices.

Normalisation/standardisation of data was made to ensure that the different scales would not affect or mislead the comparative analysis as well as minimising the effects of the variance/volatility. In addition, a separate dataset was created with the standardised i-cip prices to be compared with the original values (Al Shalabi and Shaaban, 2006). Authors also indicate it is often beneficial to experiment with multiple approaches to determine which one produces appropriate results.

Based on Al Shalabi and Shaaban (2006) study on normalisation, from the three different methods applied for their experiment, the one that showed better results was the min-max normalisation trough higher accuracy and less time of processing (Al Shalabi and Shaaban, 2006). Based on his studies combined with the experiments from Kumari and Swarnkar (2021), this was the technique selected for scaling the data based on the premisis to maintain its natural distribution.

Garza (2023) also points that scaling the data will maintain the outliers, if any are identified to reshape the data and remove the non-normality differentiation required. To provide a better comparison between these methods, plots were created to see how each method reshapes the distribution (Garza, 2023).

Figure Histogram of distribution (elaborated by author)



### Data imputation

As mentioned earlier, the dataset did not present any missing values at the first scanning (from the 279 observations no null values were identified), however to follow the time series principles, besides having the correct datatypes and respecting a temporal sequence as pointed by Brownlee (2016), there cannot be any missing dates despite the frequency of each case. For the ICO’s data, it shows a daily frequency based on business days, with data published from Monday to Friday, meaning weekends and holidays values are not included in the original calculations (ICO, 2021). By comparing the weekdays and business days present in the range between feb23 and feb24 and the ones expected to have in the dataset, three dates were perceived as missing from the desired sequence to respect the modelling requirements, which were added via data imputation after comparing three different techniques: forward fill (uses previous data to fill null values), backward fill (fill missing value with the next datapoint) and linear interpolation (gets the average between 2 points adjacent to the missing value).

Linear interpolation was the one elected to fill the null values from the new dates added, despite all methods displaying a similar curve as seen in figure **x**, the linear interpolation has a straightforward approach and helps to maintain the overall trend, and it’s use is also indicated when the missing values are in the middle of the dataset instead of the extremities to avoid bias (Koech, 2022).

Figure Data interpolation comparison

A graph of different colored lines

Description automatically generated with medium confidence

Statistical Tests

* Once identified the non-stationarity present via statistical tests (adf/knss), data transformation was applied as attempt to make the dataset stationary and enable a more precise modelling. So differentiation was applied to stabilize the variance and mean of the time series, and also compared to the original patterns found in the i-cip prices prior to applying seasonal decomposition of data.

-Adf/knsss for stationarity combined with visuals of seasonal decomposition made possible to understand the original data is not stationary and presents both trend and seasonality

- from Shapiro-wilk test for normality combined with visualisations of distribution it was identified a slightly skewed pattern.

*-granger causality test (Babu and Muniyappa, 2021) on their work they specifically focus on the relationship between each of the variables used to calculate the indicator based on the granger causality test with minimum 2 lags and max 4.(this test is not suitable for univariate data! ref{Citation})*

# Chapter 6: Results

<…>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Type** | **Parameters** | **MSE** | **MAE** | **RMSE** |
| **Decision Tree** | default | 26.3187 | 3.5375 | 5.1301 |
| **Decision Tree** | average values with cross-validation | - | 6.949420426 | 8.763510409 |
| **Gradient Boosting Regressor** | default | 14.1454 | 2.77807 | 3.76103 |
| **Gradient Boosting Regressor** | average values with cross-validation | - | 7.635333732 | 7.82329619 |
| **Linear Regression** |  | 80.9173 | 7.6027 | 8.9954 |
| **Linear Regression** | average values with cross-validation | - | 8.756623621 | 10.39989642 |
| **Random Forest** |  | 17.9827 | 3.0252 | 4.2406 |
| **Random Forest** | average values with cross-validation | - | 6.868833033 | 8.268929365 |
| **Random Forest** | with Hyperparameter Tuning | 15.041225 | 2.897152056 | 3.878301819 |
| **Random Forest** | [price\_diff] |  | 2.376309753 | 2.875611367 |
| **SVR** |  | 44.6678 | 5.82109 | 6.6834 |
| **SVR** | average values with cross-validation | - | 7.777276199 | 8.911549385 |
| **HOLT-WINTERS ES** | seasonal\_periods=12 | 60.44927895 | 6.721322783 | 7.774913437 |
| **HOLT-WINTERS ES** | seasonal\_periods=5 | 60.44927895 | 6.721322783 | 7.774913437 |

<add arima results>

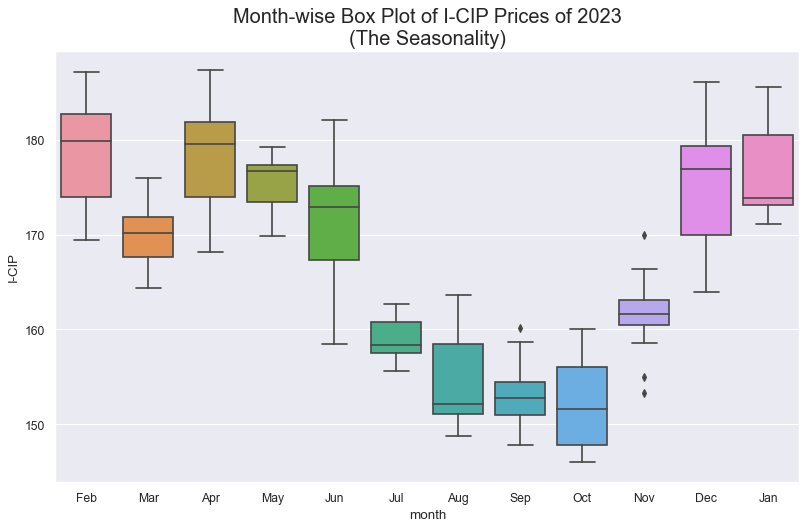
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Parameters** | **MSE** | **MAE** | **RMSE** | **AIC** | **Running Time** |
| **ARIMA** |  |  |  |  |  |  |
| **ARIMA** |  |  |  |  |  |  |
| **ARIMA** |  |  |  |  |  |  |
| **ARIMA** |  |  |  |  |  |  |
| **SARIMA** | (1,1,1)(1,1,1)[5] | 26.20829864 | 4.448954784 | 5.119404129 | 1175.899 |  |
| **SARIMA** | (1,1,1)(1,1,1)[21] | 82.25886395 | 8.086898924 | 9.069667246 | 1048.64 |  |
| **SARIMA** | (1, 1, 1)x(1, 1, 1, 63) | 11.97985146 | 2.81321932 | 3.461192202 | 8740.178 |  |
| **SARIMA** | AUTOARIMA (0,1,1)(1,1,1)[21] | 46.22230851 | 5.987563299 | 6.798699031 | 1164.33 | 59.511 seconds |
| **SARIMA** | GridSearch (1, 0, 0) (0, 1, 1, 21) | 6.158257019 | 2.059985242 | 2.481583571 |  |  |
| **SARIMA** | (1, 0, 1) (1,0,0,21) | 20.68290137 | 4.547845794 | 3.966293546 |  |  |
| **SARIMA** | (1, 0, 1) (1,1,1,63) | 851.059707 | 27.26155745 | 29.17292764 |  |  |
| **SARIMA** | price\_diff = (1,0,1) (1,0,0,21), | 4.661885899 | 1.666590523 | 2.159140083 |  |  |
| **SARIMA** | AUTO ARIMA(0,1,1)(1,1,1)[21] | 46.22230851 | 5.987563299 | 6.798699031 | 1164.33 | 86.119 seconds |
| **SARIMA** | [price\_diff] = AUTO ARIMA(3,1,0)(2,1,0)[5] (wrong scale) | 33632.17915 | 183.3633871 | 183.3907826 | 1315.647 | 8.195 seconds |
| **SARIMA [price\_diff]** | AUTO ARIMA(3,1,0)(2,1,0)[5] | 31.00963126 | 4.766618926 | 5.568629208 | 1315.647 | 6.394 |

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **LSTM** | **LSTM with tuner** |
| **Model nodes** | 2 LSTM nodes | 2-3 layers |
| **Epoch** | 100 | 50 |
| **Batch size** | 32 | print best model in jupyter |
| **Train data** |  |  |
| **Validation data** | 10% |  |
| **Test data** | 20% |  |
| **Optimizer** | ADAM |  |
| **Loss** | 3.10E-04 | 0.006255572 |
| **Learning rate** | ?? |  |
| **Dense Layer** | 1 | 1 |
| **MSE** | 0.000337122 | 10.670067 |
| **RMSE** | 0.018360889 | 3.266506849 |
| **MAE** | 0.014993943 | 2.746224598 |
| **Training time** | 5.10 seconds |  |

# Chapter 7: Evaluation

In the context of this study, the decision to focus only on the I-CIP prices was driven by a need for a rationalized analysis due to time constraints. This approach allows for a more straightforward analysis, making it practical within the limited timeframe and resources available for this research. The feature exhibits relatively low variability as indicated by its standard deviation. Compared to the other coffee groups, icip has the smaller standard deviation (10.98), which is noticed in figure x. , suggesting that it might offer a clearer signal in predictive modelling compared to other categories which have higher fluctuations and might require more complex strategies to accurately predict.

Figure Monthly distributions of values



Additionally, moving average models can drive better results and be easier to perform using univariate timeseries as defended by Brownlee (2018), especially when using SARIMA, which is an extension of ARIMA, that “explicitly supports univariate time series data with a seasonal component”. (Brownlee, 2018; Singh, 2021).

<…….>

- One of the reason why linear regression and the other static machine learning models were applied in the early stage is also to highlight the superior capacities or the auto regressors and neural networks and display their ability to better capture the fluctuations in the data and as consequence, result in more accurate predictions.

- More complex models like auto regressors and neural networks, despite being more time consuming to implement and demand more computational efforts, have the tendency to have more capacity to manage variability and incorporate the additional components intrinsic to time series.

- “*The Support Vector Machine as well as Artificial Neural Network (ANN) has been utilized for stock forecasting*”(Kumari and Swarnkar, 2021)

- On Faith’s (2021) work, for example it was identified the decomposition-based models showed a superior performance when compared to the benchmark static models.

# Chapter 8 Conclusion

<…>

### Further research

It is acknowledged, however, that each of the groups of coffee used in the indicator, has its unique market dynamics and could potentially benefit from a deeper, individualized analysis. Further research could explore and forecast these categories in more detail, especially to understand the factors driving the significant price differences and fluctuations noted in the summary statistics. This detailed analyses are recommended for future studies where time and resource constraints are less restrictive.

# References

Al Shalabi, L., Shaaban, Z., 2006. Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, in: 2006 International Conference on Dependability of Computer Systems. Presented at the 2006 International Conference on Dependability of Computer Systems, IEEE, Szklarska Poreba, pp. 207–214. https://doi.org/10.1109/DEPCOS-RELCOMEX.2006.38

Babu, B.N.P., Muniyappa, A., 2021. The interdependence of coffee futures and spot markets - An econometric analysis. JPC 56–66. https://doi.org/10.25081/jpc.2021.v49.i1.7061

Bajaj, A., 2022a. Time Series Prediction: How Is It Different From Other Machine Learning? [ML Engineer Explains] [WWW Document]. neptune.ai. URL https://neptune.ai/blog/time-series-prediction-vs-machine-learning (accessed 5.3.24).

Bajaj, A., 2022b. Anomaly Detection in Time Series [WWW Document]. neptune.ai. URL https://neptune.ai/blog/anomaly-detection-in-time-series (accessed 5.6.24).

Beniwal, M., Singh, A., Kumar, N., 2023. Forecasting long-term stock prices of global indices: A forward-validating Genetic Algorithm optimization approach for Support Vector Regression. Applied Soft Computing 145, 110566. https://doi.org/10.1016/j.asoc.2023.110566

Brownlee, J., 2017. How to Decompose Time Series Data into Trend and Seasonality. MachineLearningMastery.com. URL https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/ (accessed 3.27.24).

Brownlee, J., 2016. How To Backtest Machine Learning Models for Time Series Forecasting. MachineLearningMastery.com. URL https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/ (accessed 4.28.24).

Brunsdon, T.M., Smith, T.M.F., 1998. The Time Series Analysis of Compositional Data. Journal of Oficial Statistics Vol. 14, 237±253.

Dokumentov, A., Hyndman, R.J., 2022. STR: Seasonal-Trend Decomposition Using Regression. INFORMS Journal on Data Science 1, 50–62. https://doi.org/10.1287/ijds.2021.0004

Erasmus Kabu Aduteye, Tseganesh Tegegne Sete, Yeong Nain Chi, 2023. TIME SERIES ANALYSIS OF GLOBAL PRICES OF COFFEE: INSIGHTS INTO A COMPLEX MARKET. ijbe 8, 138–151. https://doi.org/10.58885/ijbe.v08i2.138.ea

Fatih, C., 2021. Estimating and forecasting Volatility in Coffee Prices: Evidence of ARIMA-GARCH models. https://doi.org/10.13140/RG.2.2.26036.17281

Fianu, E.S., 2022. Analyzing and Forecasting Multi-Commodity Prices Using Variants of Mode Decomposition-Based Extreme Learning Machine Hybridization Approach. Forecasting 4, 538–564. https://doi.org/10.3390/forecast4020030

Fontana, R., Molena, A., Pegoraro, L., Salmaso, L., 2023. Design of experiments and machine learning with application to industrial experiments. Stat Papers 64, 1251–1274. https://doi.org/10.1007/s00362-023-01437-w

Garza, J., 2023. Reinforcement Learning for Stock Option Trading. ICT.

Guido, S., Müller, A.C., 2016. Introduction to Machine Learning with Python, 1st ed. O’Reilly Media, Inc., United States of America.

Harrison, M., 2023. Professional Pandas: Handling Missing Data With Pandas Dropna [WWW Document]. Ponder. URL https://ponder.io/professional-pandas-handling-missing-data-with-pandas-dropna/ (accessed 5.4.24).

Hwase, T.K., Fofanah, A.J., 2021. Machine Learning Model Approaches for Price Prediction in Coffee Market using Linear Regression, XGB, and LSTM Techniques. IJSRST 10–48. https://doi.org/10.32628/IJSRST218583

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: Principles and Practice (2nd ed), 2nd edition. ed.

ICO, 2024a. Public Market Information | International Coffee Organization [WWW Document]. URL https://icocoffee.org/resources/public-market-information/ (accessed 12.3.23).

ICO, 2024b. International Coffee Organization - Glossary Of Terms Used/glossary.asp [WWW Document]. ICO Glossary. URL https://www.ico.org/glossary.asp (accessed 5.3.24).

ICO, 2024c. About Economics & Statistics | International Coffee Organization [WWW Document]. URL https://icocoffee.org/resources/about-economics-and-statistics/ (accessed 4.24.24).

ICO, 2021. RULES ON STATISTICS INDICATOR PRICES.

Josef Perktold, Skipper Seabold, Kevin Sheppard, ChadFulton, Kerby Shedden, jbrockmendel, j-grana6, Peter Quackenbush, Vincent Arel-Bundock, Wes McKinney, Ian Langmore, Bart Baker, Ralf Gommers, yogabonito, s-scherrer, Yauhen Zhurko, Matthew Brett, Enrico Giampieri, yl565, Jarrod Millman, Paul Hobson, Vincent, Pamphile Roy, Tom Augspurger, tvanzyl, alexbrc, Tyler Hartley, Fernando Perez, Yuji Tamiya, Yaroslav Halchenko, 2024. statsmodels/statsmodels: Release 0.14.2. https://doi.org/10.5281/ZENODO.593847

Koech, D.K., 2022. A Complete Guide on How to Impute Missing Values in Time Series in Python [WWW Document]. Engineering Education (EngEd) Program | Section. URL https://www.section.io/engineering-education/missing-values-in-time-series/ (accessed 11.19.23).

Kumari, B., Swarnkar, T., 2021. Stock movement prediction using hybrid normalization technique and artificial neural network. International Journal of Advanced Technology and Engineering Exploration 8. https://doi.org/10.19101/IJATEE.2021.874387

Kumari, B., Swarnkar, T., 2020. Importance of Data Standardization Methods on Stock Indices Prediction Accuracy, in: Pati, B., Panigrahi, C.R., Buyya, R., Li, K.-C. (Eds.), Advanced Computing and Intelligent Engineering. Springer, Singapore, pp. 309–318. https://doi.org/10.1007/978-981-15-1081-6\_26

Nwokike, C.C., Offorha, B.C., Obubu, M., Ugoala, C.B., Ukomah, H.I., 2020. Comparing SANN and SARIMA for forecasting frequency of monthly rainfall in Umuahia. Scientific African 10, e00621. https://doi.org/10.1016/j.sciaf.2020.e00621

Pandas, 2023. How to handle time series data with ease — pandas 2.1.3 documentation [WWW Document]. URL https://pandas.pydata.org/docs/getting\_started/intro\_tutorials/09\_timeseries.html (accessed 11.20.23).

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Ben Taieb, S., Bergmeir, C., Bessa, R.J., Bijak, J., Boylan, J.E., Browell, J., Carnevale, C., Castle, J.L., Cirillo, P., Clements, M.P., Cordeiro, C., Cyrino Oliveira, F.L., De Baets, S., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P.H., Frazier, D.T., Gilliland, M., Gönül, M.S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, Mariangela, Guidolin, Massimo, Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D.F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V.R.R., Kang, Y., Koehler, A.B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G.M., Martinez, A.B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J.M., Pedio, M., Pedregal, D.J., Pinson, P., Ramos, P., Rapach, D.E., Reade, J.J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H.L., Spiliotis, E., Syntetos, A.A., Talagala, P.D., Talagala, T.S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Trapero Arenas, J.R., Wang, X., Winkler, R.L., Yusupova, A., Ziel, F., 2022. Forecasting: theory and practice. International Journal of Forecasting 38, 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001

Saunders, M., Lewis, P., Thornhill, A., 2009. Research methods for business students, 5. ed. ed. Financial Times Prentice Hall, Harlow.

Singh, R., 2021. Time Series forecasting using LSTM/ARIMA/Moving Average use case(Single/Multi-variate) with code. Medium. URL https://ranasinghiitkgp.medium.com/time-series-forecasting-using-lstm-arima-moving-average-use-case-single-multi-variate-with-code-5dd41e32d1fc (accessed 5.4.24).

Svolba, G., 2022. Determining the best length of the history of your timeseries data for timeseries forecasting. MLearning.ai. URL https://medium.com/mlearning-ai/determining-the-best-length-of-the-history-of-your-timeseries-data-for-timeseries-forecasting-f8600a3c086 (accessed 3.15.24).

Zhang, L., Wang, R., Li, Z., Li, J., Ge, Y., Wa, S., Huang, S., Lv, C., 2023. Time-Series Neural Network: A High-Accuracy Time-Series Forecasting Method Based on Kernel Filter and Time Attention. Information 14, 500. https://doi.org/10.3390/info14090500

Zhu, Y., 2022. Stock price prediction using the RNN model, in: 2022 International Conference on Applied Physics and Computing (ICAPC). Presented at the 2022 International Conference on Applied Physics and Computing (ICAPC), IEEE, Ottawa, ON, Canada. https://doi.org/10.1088/1742-6596/1650/3/032103

# Appendix

### Appendix 1:

Table Null values after first merging of monthly datasets

A screenshot of a computer

Description automatically generated

1. The ICO was established in 1963 under the aegis of the United Nations after the first International Coffee Agreement in 1962 was approved. The ICO is the only intergovernmental organization for coffee, bringing together exporting and importing Governments. Nowadays it represents over 93% of world coffee production and 63% of world consumption.(ICO, 2024a) [↑](#footnote-ref-1)
2. LOESS: locally estimated scatterplot smoothing [↑](#footnote-ref-2)
3. Under the International Coffee Agreement of 2007 for all data collected, transmitted, calculated and publicised by Statistics Committee and approved by the International Coffee Council at its 129th Session in April 2021 and implemented on 1 May 2021. This Session also employed a new system of collection of prices across the US and EU markets (ICO, 2021). [↑](#footnote-ref-3)