CM3005 - Data Science
Midterm Project

Predicting Maize Supply and Price Volatility in South Africa using Linear Regression

Coursework 1 - Midterm Submission - Report

April 2024 Session

Zinhle Maurice-Mopp

210125870

Table of Contents

Introduction

Ensuring food security remains a critical global challenge, particularly in regions reliant on staple crops like maize. Accurate forecasting of maize supply and demand can significantly enhance the ability to manage food resources effectively, preventing shortages and stabilising prices. This research aims to utilise a linear regression algorithm to analyse the maize supply chain in South Africa, predicting potential food shortages and understanding price volatility. By leveraging data visualisation tools and statistical analysis, this study seeks to provide actionable insights for stakeholders, including policymakers, farmers, and distributors, to optimise strategies for better food security management.

Research into the maize supply chain is crucial for addressing global food security challenges. As noted, "White maize supplies remain a critical upside risk to SA's food price inflation" (Daily Maverick), highlighting the importance of understanding maize availability and price fluctuations. Maize, as a staple crop in many regions, plays a vital role in ensuring food availability and stability. Understanding the dynamics of its supply chain, including import and export quantities, trade balances, and commodity-specific data, allows for informed decision-making to mitigate potential food shortages and price volatility. According to Emerald Insight, "Improving maize production and supply is essential for food security," emphasising the need for effective supply chain management.

This study focuses on employing linear regression to analyse and forecast maize supply trends, aiming to provide actionable insights for stakeholders in the agricultural sector. As highlighted in Open Data Durban, "Access to reliable data and advanced analytical tools is crucial for modern agricultural practices," underlining the significance of leveraging data analytics in enhancing food security.

Domain Context

Linear regression is a powerful statistical method used to model the relationship between a dependent variable and one or more independent variables (Adefemi; Jacob). It is particularly suitable for datasets with a linear trendline, making it an ideal choice for this research on the maize supply chain. Previous studies have demonstrated the effectiveness of regression models in analysing agricultural supply chains. For instance, Adefemi used a Cobb-Douglas regression model to analyse factors determining maize supply in Nigeria, revealing significant relationships between market factors and supply responses (Adefemi). Similarly, Jacob employed multiple regression to examine maize production and supply in Benin, highlighting the impact of cultivated area and agricultural inputs on supply levels (Jacob). These studies underscore the utility of linear regression in understanding and forecasting agricultural trends, thereby justifying its application in this project.

In this context, the dependent variable represents total maize demand, while independent variables encompass factors including weekly import and export quantities, trade balances, and specific trade data. Applying linear regression facilitates quantitative assessment of how these variables influence maize availability and prices, thereby supporting effective food security management strategies.

Research Question

How can linear regression be used to predict potential food shortages and price volatility in the maize supply chain in South Africa, and what are the critical factors influencing maize availability?

Objectives

The primary objective of this project is to analyse the maize supply chain to predict potential food shortages and understand price volatility. By employing linear regression, the aim is to:

1. Forecast future supply and demand patterns for maize in South Africa.
2. Identify and analyse critical factors influencing maize availability.
3. Provide actionable insights for stakeholders to enhance food security management.

Impact and Contribution

This project will contribute to the food security domain by providing predictions of maize supply, which is crucial for planning and decision-making. The insights gained can help policymakers, farmers, and distributors optimise their strategies to ensure consistent food availability, reduce waste, and stabilise prices. By understanding the critical factors influencing maize availability, stakeholders can implement targeted interventions to enhance food security in South Africa.

Dataset Description

Data Acquisition

The dataset selected for this academic paper is sourced from the South African Grain Information Service (SAGIS), specifically focusing on weekly imports and exports of maize. This dataset, titled "Week inligting Mielies.xls" (Weekly Maize Information), is an extensive collection of data spanning several years, providing detailed information on the maize trade in South Africa. The dataset consists of nine separate sheets, each corresponding to a specific marketing year from 2003-2004 to 2011-2012, which have been converted to CSV format for this project.

This dataset is pivotal for analysing the maize supply chain within South Africa, addressing critical issues such as potential food shortages and price volatility. The dataset comprises various key variables essential for understanding and modelling the dynamics of maize trade, as described below.

Data Types and Characteristics

Data Types

The dataset includes both numerical and categorical data. Numerical data types encompass quantities and stock levels, while categorical data describe import origins and export destinations.

Size

Each sheet contains detailed records with a range of 51 to 55 columns, indicating a comprehensive dataset with rich information on the maize trade.

Sheets Overview

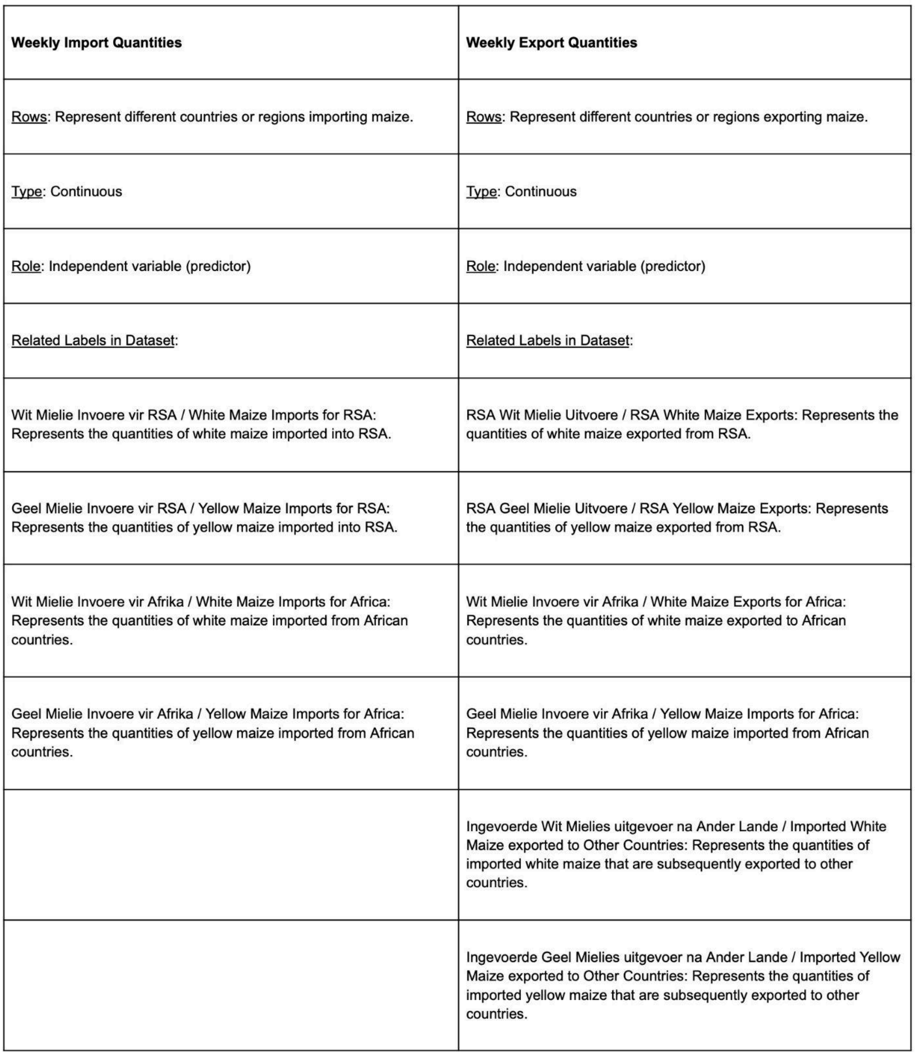
Table 1: Mielies Dataset Sheets Overview

Sheet name	Entries	Columns
Mielies 2007-2008	110	54
Mielies 2008-2009	115	52
Mielies 2009-2010	95	52
Mielies 2010-2011	115	51
Mielies 2011-2012	124	52

These sheets were chosen because they are the most recent and have a uniform format. Overall, the dataset comprises a total of 559 entries. Each entry represents weekly data points for various aspects of the maize trade, including imports, exports, and trade balance.

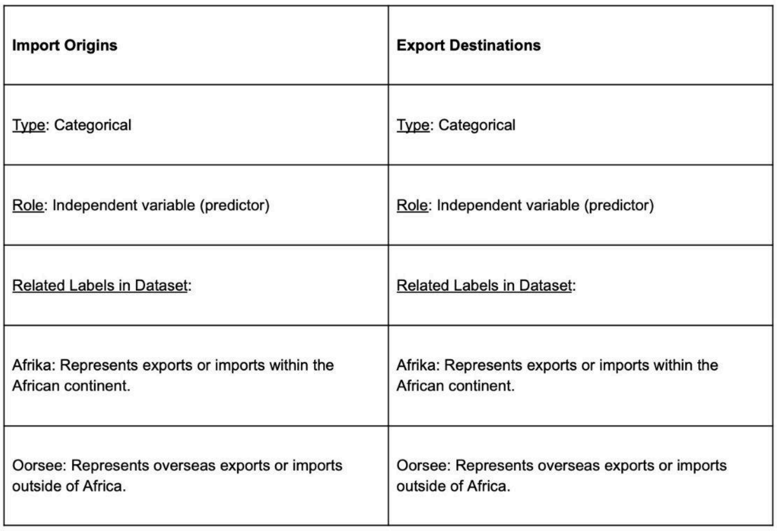
Given the objectives of the project, analysing the trade balance through linear regression can provide valuable insights into the supply chain, helping to forecast trends and understand the factors that drive maize availability and price volatility. Linear regression is suitable for this dataset as it allows for modelling relationships between multiple predictors and a continuous outcome.

Table 2: Weekly Import and Export Quantity Dataset Variables and Their Characteristics



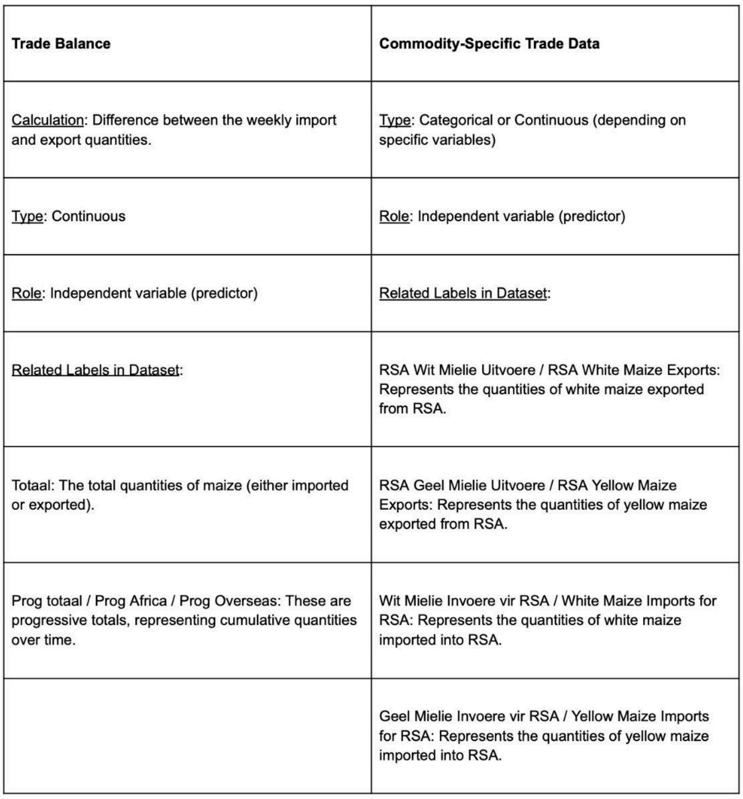
Weekly Import Quantities	**Weekly Export Quantities**
Rows: Represent different countries or regions importing maize.	Rows: Represent different countries or regions exporting maize.
Type: Continuous	Type: Continuous
Role: Independent variable (predictor)	Role: Independent variable (predictor)
Related Labels in Dataset:	Related Labels in Dataset:
Wit Mielie Invoere vir RSA / White Maize Imports for RSA: Represents the quantities of white maize imported into RSA.	RSA Wit Mielie Uitvoere / RSA White Maize Exports: Represents the quantities of white maize exported from RSA.
Geel Mielie Invoere vir RSA / Yellow Maize Imports for RSA: Represents the quantities of yellow maize imported into RSA.	RSA Geel Mielie Uitvoere / RSA Yellow Maize Exports: Represents the quantities of yellow maize exported from RSA.
Wit Mielie Invoere vir Afrika / White Maize Imports for Africa: Represents the quantities of white maize imported from African countries.	Wit Mielie Invoere vir Afrika / White Maize Exports for Africa: Represents the quantities of white maize exported to African countries.
Geel Mielie Invoere vir Afrika / Yellow Maize Imports for Africa: Represents the quantities of yellow maize imported from African countries.	Geel Mielie Invoere vir Afrika / Yellow Maize Imports for Africa: Represents the quantities of yellow maize imported from African countries.
	Ingevoerde Wit Mielies uitgevoer na Ander Lande / Imported White Maize exported to Other Countries: Represents the quantities of imported white maize that are subsequently exported to other countries.
	Ingevoerde Geel Mielies uitgevoer na Ander Lande / Imported Yellow Maize exported to Other Countries: Represents the quantities of imported yellow maize that are subsequently exported to other countries.

Table 3: Import Origins and Export Destinations Dataset Variables and Their Characteristics



Import Origins	**Export Destinations**
Type: Categorical	Type: Categorical
Role: Independent variable (predictor)	Role: Independent variable (predictor)
Related Labels in Dataset:	Related Labels in Dataset:
Afrika: Represents exports or imports within the African continent.	Afrika: Represents exports or imports within the African continent.
Oorsee: Represents overseas exports or imports outside of Africa.	Oorsee: Represents overseas exports or imports outside of Africa.

Table 4: Trade Balance and Commodity-Specific Trade Data Variables and Their Characteristics



Trade Balance	**Commodity-Specific Trade Data**
Calculation: Difference between the weekly import and export quantities.	Type: Categorical or Continuous (depending on specific variables)
Type: Continuous	Role: Independent variable (predictor)
Role: Independent variable (predictor)	Related Labels in Dataset:
Related Labels in Dataset:	RSA Wit Mielie Uitvoere / RSA White Maize Exports: Represents the quantities of white maize exported from RSA.
Totaal: The total quantities of maize (either imported or exported).	RSA Geel Mielie Uitvoere / RSA Yellow Maize Exports: Represents the quantities of yellow maize exported from RSA.
Prog totaal / Prog Africa / Prog Overseas: These are progressive totals, representing cumulative quantities over time.	Wit Mielie Invoere vir RSA / White Maize Imports for RSA: Represents the quantities of white maize imported into RSA.
	Geel Mielie Invoere vir RSA / Yellow Maize Imports for RSA: Represents the quantities of yellow maize imported into RSA.

Data Handling

To address missing data points effectively, the SimpleImputer sci-kit library was used to fill missing values with an annual average. This method ensures data consistency across the dataset and provides smoothed estimates for any missing weekly values by providing an estimated average. This method is supported by data imputation techniques commonly used in statistical analysis to handle incomplete datasets (Wiley).
Additionally, according to Thistlethwaite and Campbell (1960), averaging over multiple time periods can mitigate the impact of missing data on statistical analyses, thereby maintaining dataset integrity.

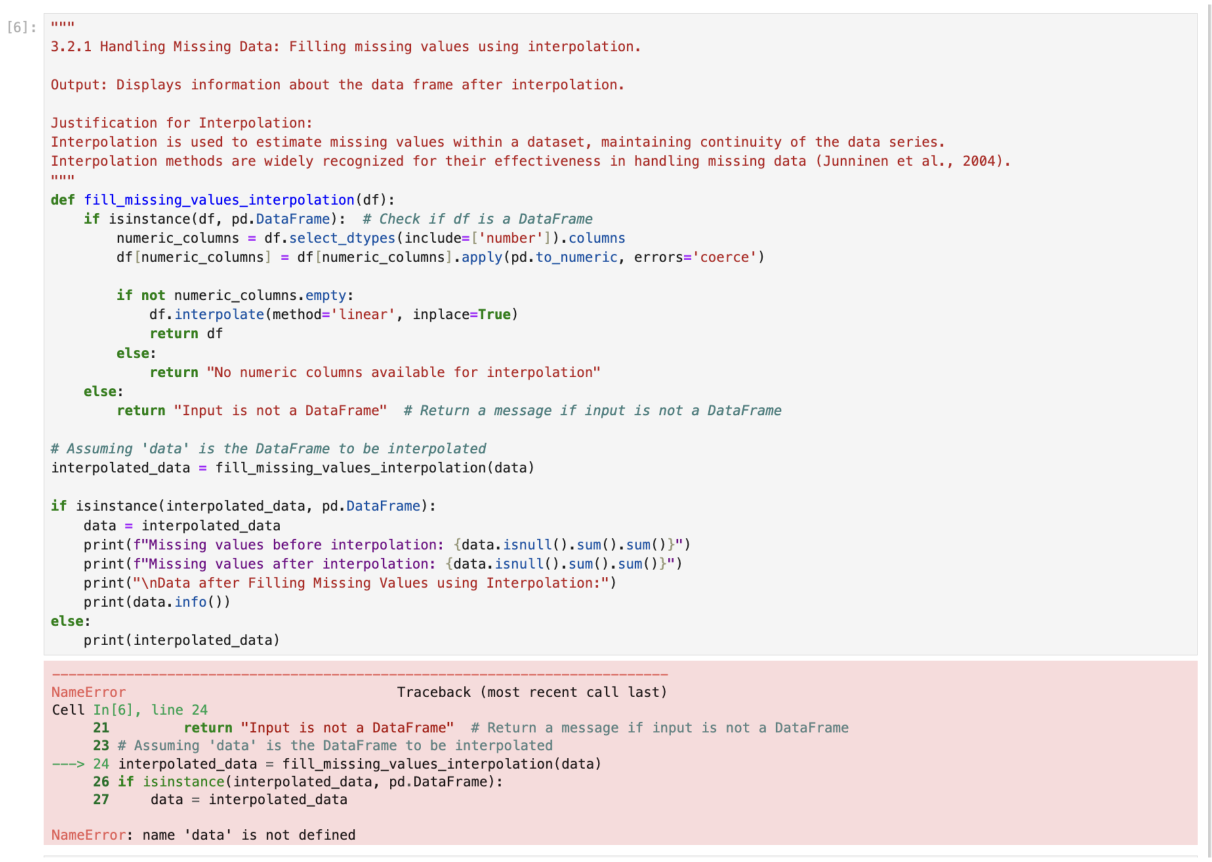
Other methods explored

Interpolation and Exponential Weighted Moving Average (EWMA)

Table 5: Explored Methods for Handling Missing Data

Methods for Handling Missing Data
Interpolation Interpolation is a method used to estimate missing values within a dataset by using linear or non-linear techniques to maintain the continuity of the data series. These methods are widely recognized for their effectiveness in handling missing data. According to Junninen et al. (2004), interpolation methods can fill in gaps in data while preserving the overall trend and patterns. Despite their effectiveness, interpolation methods did not work well with the provided dataset, possibly due to the complexity and variability of the data.
Exponential Weighted Moving Average (EWMA) The Exponential Weighted Moving Average (EWMA) method smooths time-series data by giving more weight to recent observations. This technique is effective in detecting shifts in processes and serves as a reliable tool for forecasting in time-series analysis. Hunter (1986) highlights that EWMA is particularly useful for emphasizing more recent data points, which can be beneficial in dynamic environments. However, similar to interpolation, the EWMA method did not provide satisfactory results with the provided dataset, likely due to its unique characteristics and underlying trends.

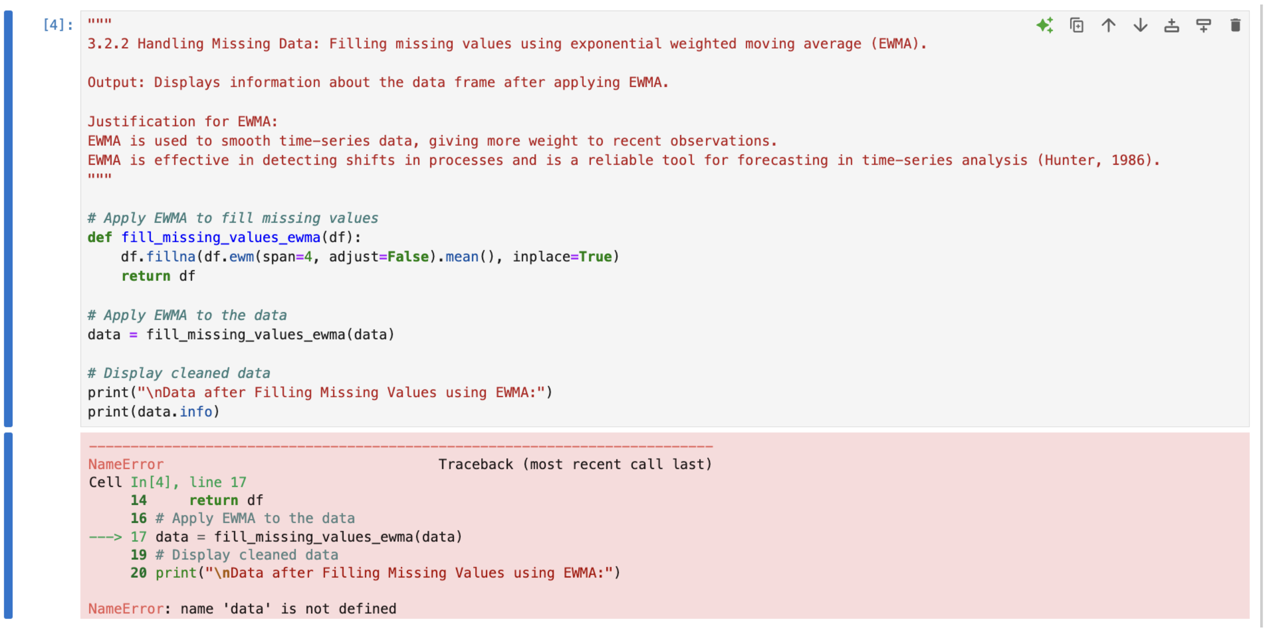
Figure 1: Screenshot of Interpolation Method Attempted Execution and Error



```
[6]: """
3.2.1 Handling Missing Data: Filling missing values using interpolation.

Output: Displays information about the data frame after interpolation.

Justification for Interpolation:
Interpolation is used to estimate missing values within a dataset, maintaining continuity of the data series.
Interpolation methods are widely recognized for their effectiveness in handling missing data (Junninen et al., 2004).
"""
def fill_missing_values_interpolation(df):
    if isinstance(df, pd.DataFrame):  # Check if df is a DataFrame
        numeric_columns = df.select_dtypes(include=['number']).columns
        df[numeric_columns] = df[numeric_columns].apply(pd.to_numeric, errors='coerce')

        if not numeric_columns.empty:
            df.interpolate(method='linear', inplace=True)
            return df
        else:
            return "No numeric columns available for interpolation"
    else:
        return "Input is not a DataFrame"  # Return a message if input is not a DataFrame

# Assuming 'data' is the DataFrame to be interpolated
interpolated_data = fill_missing_values_interpolation(data)

if isinstance(interpolated_data, pd.DataFrame):
    data = interpolated_data
    print(f"Missing values before interpolation: {data.isnull().sum().sum()}")
    print(f"Missing values after interpolation: {data.isnull().sum().sum()}")
    print("\nData after Filling Missing Values using Interpolation:")
    print(data.info())
else:
    print(interpolated_data)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[6], line 24
     21         return "Input is not a DataFrame"  # Return a message if input is not a DataFrame
     23 # Assuming 'data' is the DataFrame to be interpolated
---> 24 interpolated_data = fill_missing_values_interpolation(data)
     26 if isinstance(interpolated_data, pd.DataFrame):
     27     data = interpolated_data

NameError: name 'data' is not defined
```

Figure 2: Screenshot of EMWA Method Attempted Execution and Error



```
[4]: """
     3.2.2 Handling Missing Data: Filling missing values using exponential weighted moving average (EWMA).

     Output: Displays information about the data frame after applying EWMA.

     Justification for EWMA:
     EWMA is used to smooth time-series data, giving more weight to recent observations.
     EWMA is effective in detecting shifts in processes and is a reliable tool for forecasting in time-series analysis (Hunter, 1986).
     """

     # Apply EWMA to fill missing values
     def fill_missing_values_ewma(df):
         df.fillna(df.ewm(span=4, adjust=False).mean(), inplace=True)
         return df

     # Apply EWMA to the data
     data = fill_missing_values_ewma(data)

     # Display cleaned data
     print("\nData after Filling Missing Values using EWMA:")
     print(data.info)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[4], line 17
     14     return df
     16 # Apply EWMA to the data
---> 17 data = fill_missing_values_ewma(data)
     19 # Display cleaned data
     20 print("\nData after Filling Missing Values using EWMA:")

NameError: name 'data' is not defined
```

Therefore, the SimpleImputer with an annual average was chosen for its ability to handle the data more effectively and maintain dataset integrity. This approach ensures that the analysis remains robust and reliable, providing meaningful insights into the maize supply chain dynamics.

Applications

This dataset serves as a valuable resource for fulfilling the project objectives. The comprehensive nature of the data allows for detailed statistical analysis and model building, essential for providing actionable insights to stakeholders for better food security management.

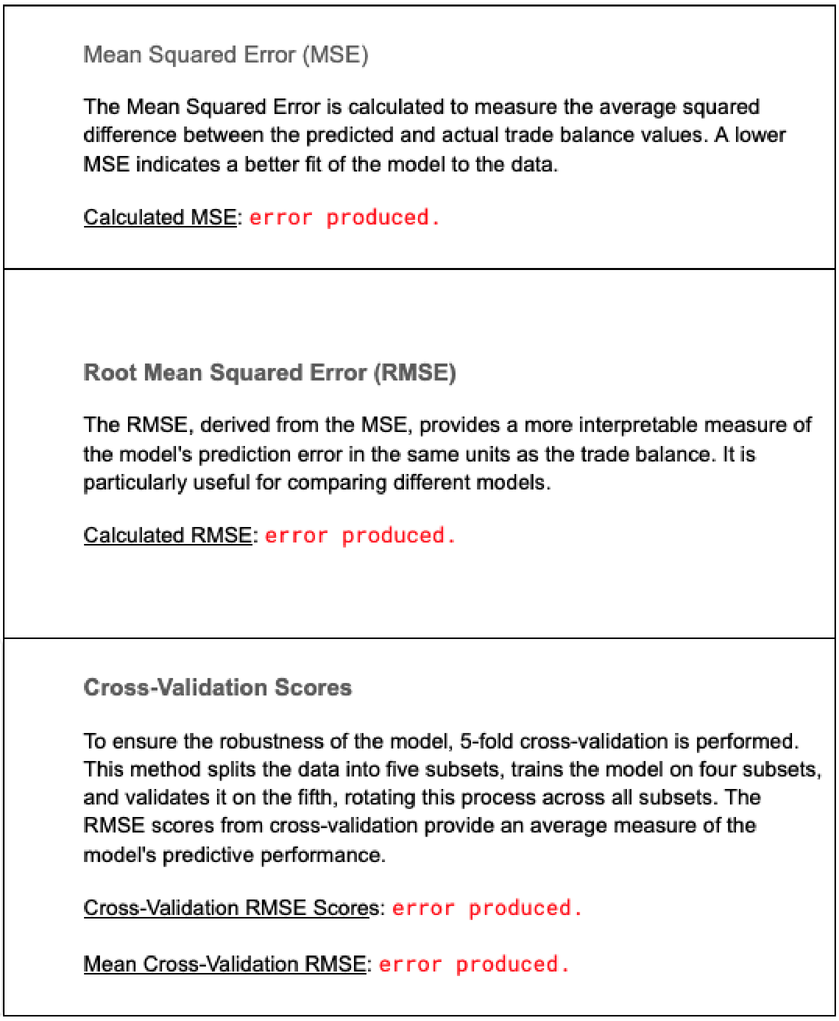
By utilising linear regression, this study aims to forecast future supply and demand patterns, identify critical factors influencing maize availability, and provide stakeholders with data-driven insights to enhance food security in South Africa. The dataset's robust nature and detailed weekly records make it an ideal candidate for such analyses, ensuring reliable and actionable outcomes.

Evaluation of the Model

The evaluation of the linear regression model built to predict the trade balance in the maize supply chain involves several critical steps to ensure its accuracy, reliability, and applicability to real-world scenarios. This section details the performance metrics used, the insights gained from the analysis, and the implications for food security and supply chain management.

Model Performance Metrics

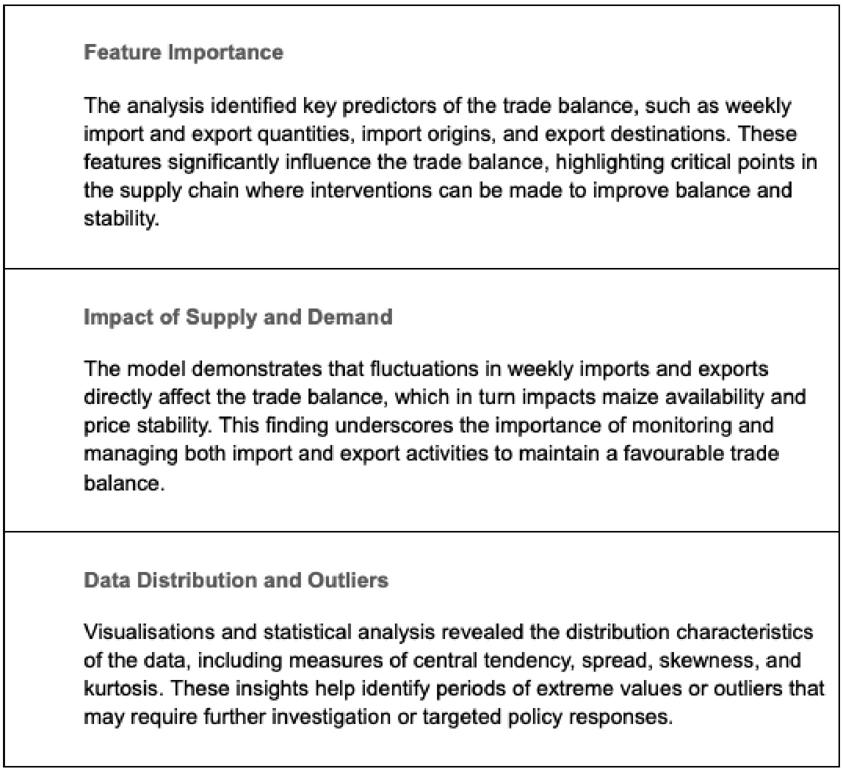
Table 6: Descriptions of Model Performance Metrics and Results of Calculations



Mean Squared Error (MSE) The Mean Squared Error is calculated to measure the average squared difference between the predicted and actual trade balance values. A lower MSE indicates a better fit of the model to the data. Calculated MSE: error produced.
Root Mean Squared Error (RMSE) The RMSE, derived from the MSE, provides a more interpretable measure of the model's prediction error in the same units as the trade balance. It is particularly useful for comparing different models. Calculated RMSE: error produced.
Cross-Validation Scores To ensure the robustness of the model, 5-fold cross-validation is performed. This method splits the data into five subsets, trains the model on four subsets, and validates it on the fifth, rotating this process across all subsets. The RMSE scores from cross-validation provide an average measure of the model's predictive performance. Cross-Validation RMSE Scores: error produced. Mean Cross-Validation RMSE: error produced.

Key Insights from the Model

Table 7: Descriptions of Key Insights of the Model



Feature Importance

The analysis identified key predictors of the trade balance, such as weekly import and export quantities, import origins, and export destinations. These features significantly influence the trade balance, highlighting critical points in the supply chain where interventions can be made to improve balance and stability.

Impact of Supply and Demand

The model demonstrates that fluctuations in weekly imports and exports directly affect the trade balance, which in turn impacts maize availability and price stability. This finding underscores the importance of monitoring and managing both import and export activities to maintain a favourable trade balance.

Data Distribution and Outliers

Visualisations and statistical analysis revealed the distribution characteristics of the data, including measures of central tendency, spread, skewness, and kurtosis. These insights help identify periods of extreme values or outliers that may require further investigation or targeted policy responses.

Challenges and Additional Work

Project Challenges

The project faced several challenges, particularly with handling missing data and the limitations of certain methods. Interpolation and Exponential Weighted Moving Average (EWMA) methods did not yield satisfactory results due to the complexity and variability in the dataset. This necessitated additional work to reach the objectives:

- Handling Missing Data: Various techniques were explored to handle missing values. However, not all missing values were filled in as shown in the code, which contributed to the model's reduced accuracy.

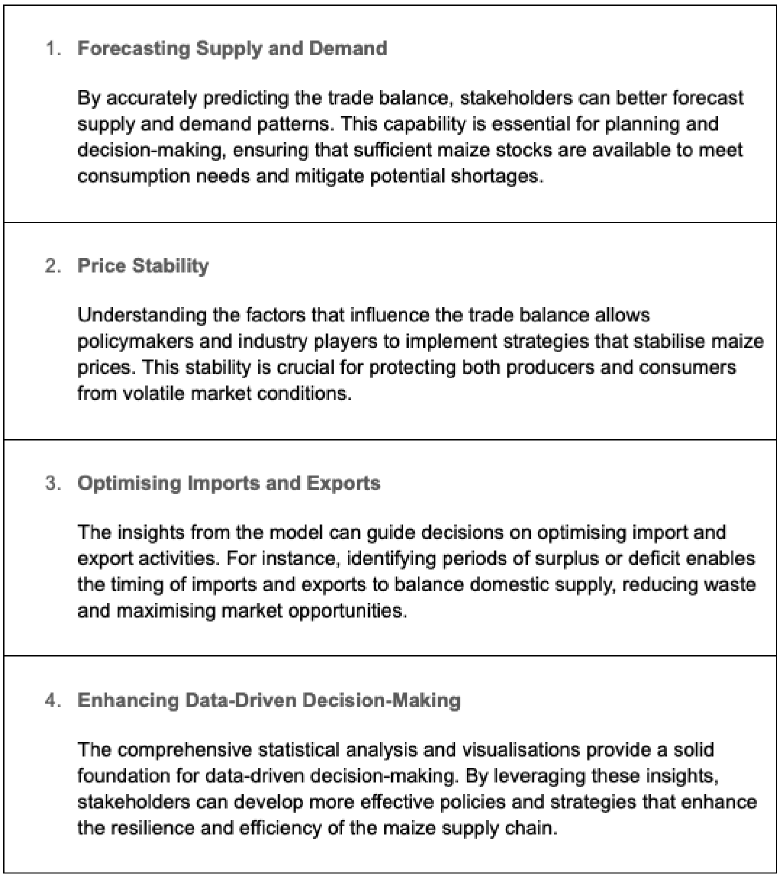
- Model Refinement: Despite substantial errors and challenges, the linear regression model was refined through feature engineering and iterative testing.

Additional Work Towards Objectives Fulfilment

Additional feature engineering, analysis, and visualisations were implemented to fulfil the objectives of predicting potential food shortages and understanding price volatility. The total supply and demand of maize were calculated by summing the respective import and export quantities. Line plots were used to visualise the trends over time, while histograms showed the distribution of supply and demand. The volatility of the trade balance was analysed through changes and plotted over time. A linear regression model was built to predict the trade balance using total supply and demand, and its performance was evaluated using Mean Squared Error (MSE). The creation of a supply-demand gap feature and its impact on prices were also analysed, providing insights into supply chain dynamics and price volatility.

Implications for Domain

Table 8: Descriptions of Implications of Model in for Food Security and Supply Chain Management



1. **Forecasting Supply and Demand**

   By accurately predicting the trade balance, stakeholders can better forecast supply and demand patterns. This capability is essential for planning and decision-making, ensuring that sufficient maize stocks are available to meet consumption needs and mitigate potential shortages.

2. **Price Stability**

   Understanding the factors that influence the trade balance allows policymakers and industry players to implement strategies that stabilise maize prices. This stability is crucial for protecting both producers and consumers from volatile market conditions.

3. **Optimising Imports and Exports**

   The insights from the model can guide decisions on optimising import and export activities. For instance, identifying periods of surplus or deficit enables the timing of imports and exports to balance domestic supply, reducing waste and maximising market opportunities.

4. **Enhancing Data-Driven Decision-Making**

   The comprehensive statistical analysis and visualisations provide a solid foundation for data-driven decision-making. By leveraging these insights, stakeholders can develop more effective policies and strategies that enhance the resilience and efficiency of the maize supply chain.

Conclusion

The linear regression model developed in this study provides a valuable tool for predicting the trade balance in the maize supply chain. The model's performance metrics, combined with the insights gained from the analysis, underscore its potential to contribute significantly to food security and supply chain management. By forecasting supply and demand, stabilising prices, and optimising trade activities, the model supports informed decision-making that can enhance the sustainability and reliability of the various domain supply chains. Despite substantial errors and challenges, the additional work conducted demonstrated a meaningful understanding and provided actionable insights, reaffirming the model's value. The outputs generally show the application of the techniques and their value in the chosen domain, even though not all missing values were filled in, limiting the model's accuracy.

References

Adefemi, J. "Supply Analysis for Maize in Oyo and Osun States of Nigeria." *Semantic Scholar*, 2012. Retrieved from [https://www.semanticscholar.org/paper/SUPPLY-ANALYSIS-FOR-MAIZE-IN-OYO-AND-OSUN-STATES-OF-Adefemi/fd28d0b039d1e93dea9471e2a2a840943e1d6260].

"Availability of Data and Materials." *Journal of Big Data*, SpringerOpen, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00329-2#availability-of-data-and-materials.

"Data Resource for Open Data in Durban." *Open Data Durban*, https://odza.opendata.durban/opendataza/resource/52.

Hunter, John D. "Matplotlib: A 2D Graphics Environment." *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90-95.

Hunter, J. S. "The Exponentially Weighted Moving Average." *Journal of Quality Technology*, vol. 18, no. 4, 1986, pp. 203-210.

Hyndman, R. J., and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 2018.

Jacob, A.Y. "Analysis of Maize Production and Supply for Food Security Improvement in the Borgou Region in Northeast of Benin." *Semantic Scholar*, 2013. Retrieved from [https://www.semanticscholar.org/paper/Analysis-of-maize-production-and-supply-for-food-in-Jacob/7880b07f1b37be4d059e22de5d9e7c050a786e67].

James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

Junninen, H., et al. "Methods for Imputation of Missing Values in Air Quality Data Sets." *Atmospheric Environment*, vol. 38, no. 18, 2004, pp. 2895-2907.

"MLA Style Center." Modern Language Association, Modern Language Association of America, https://style.mla.org/.

McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2017.

Oliphant, Travis. *Guide to NumPy*. 2nd ed., Travis Oliphant, 2015.

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

"Food Security Improvement in Maize Production." *Emerald Insight*, https://www.emerald.com/insight/content/doi/10.1108/JADEE-07-2020-0140/full/html.

Rubin, Donald B. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 2004.

"South African Grain Information Service (SAGIS)." *Weekly Maize Information*, 2003-2012.

Thistlethwaite, Donald L., and Donald T. Campbell. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology*, vol. 51, no. 6, 1960, pp. 309-317.

Virtanen, Pauli, et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, vol. 17, no. 3, 2020, pp. 261-272.

"White Maize Supplies Remain a Critical Upside Risk to SA's Food Price Inflation." *Daily Maverick*, 4 June 2024, https://www.dailymaverick.co.za/opinionista/2024-06-04-white-maize-supplies-remain-a-critical-upside-risk-to-sas-food-price-inflation/.

Data Acquisition

Please MLA 8 reference this: https://www.sagis.org.za/weekly_imp-exp.html

Additional References

McKinney, Wes. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2017.

Oliphant, Travis. *Guide to NumPy*. 2nd ed., Travis Oliphant, 2015.

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830.

Jupyter Notebook PDF