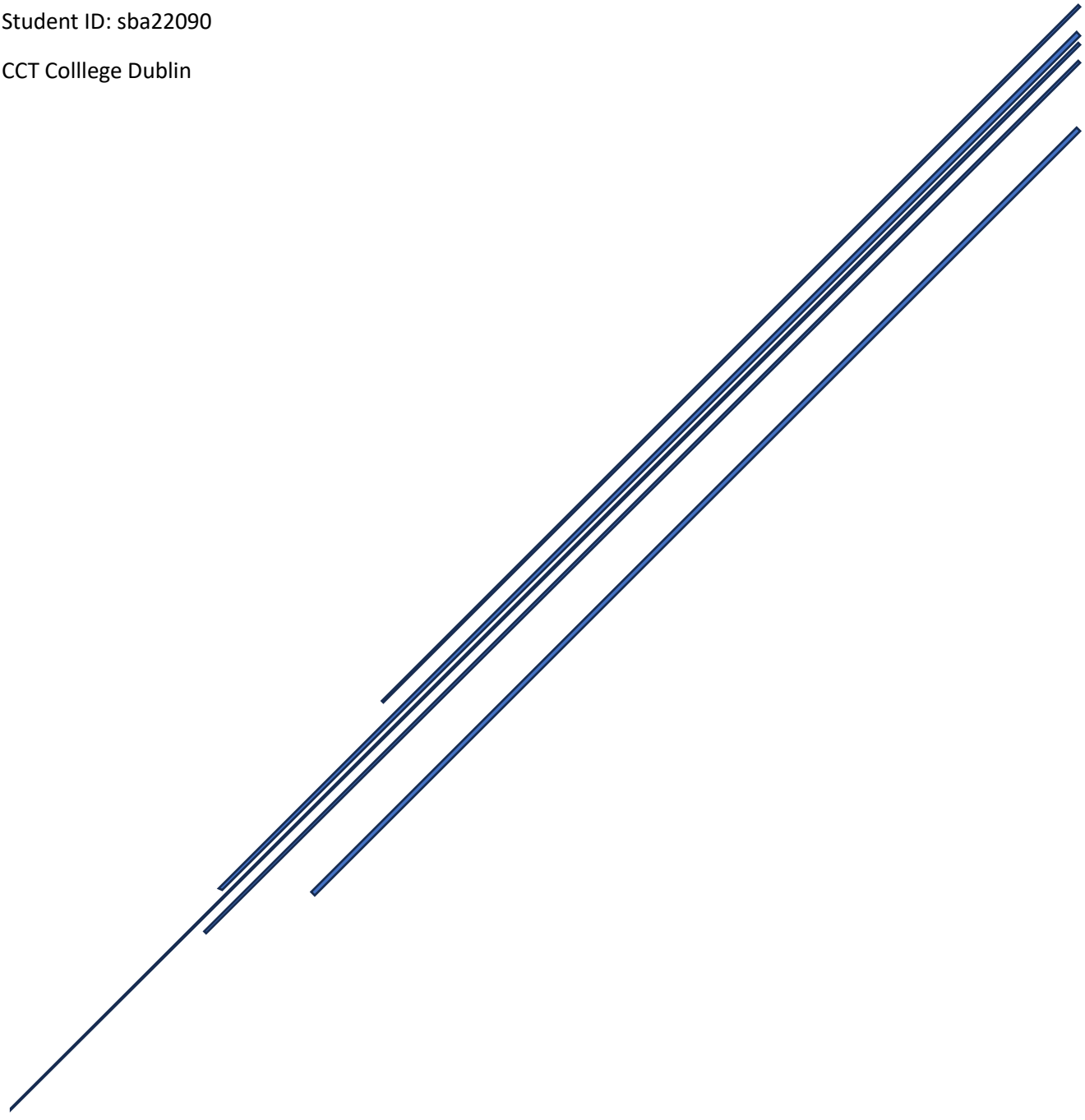
MSc in Data Analytics

Author: Federico Ariton

e-mail: sba22090@student.cct.ie

Student ID: sba22090

CCT Colllege Dublin



Index

Introduction

Agriculture is a vital component of Ireland's economy, playing a significant role in exports, employment, and rural development. With the rising global demand for high-quality agricultural products and the increasing adoption of data-driven strategies, understanding the dynamics of production and trade is more critical than ever. Ireland's agricultural sector, operating within the European Union's Common Agricultural Policy (CAP), is uniquely positioned to compete in global markets, but achieving sustainable growth requires actionable insights derived from data analysis.

This project aims to analyze Ireland's agricultural sector in comparison with other countries worldwide, focusing on production trends, export performance, and trade balance. By employing statistical techniques, machine learning models, and interactive data visualizations, the study seeks to identify key drivers of export value and forecast future trends. Additionally, sentiment analysis offers valuable insights into producers' and consumers' perspectives on agricultural topics, adding a qualitative dimension to the research.

Methodology and Structure

To ensure a structured and comprehensive approach, the project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a widely recognized framework for data analysis projects. The methodology comprises six phases—business understanding, data understanding, data

preparation, modeling, evaluation, and deployment—and has been adapted to the unique needs of this analysis.

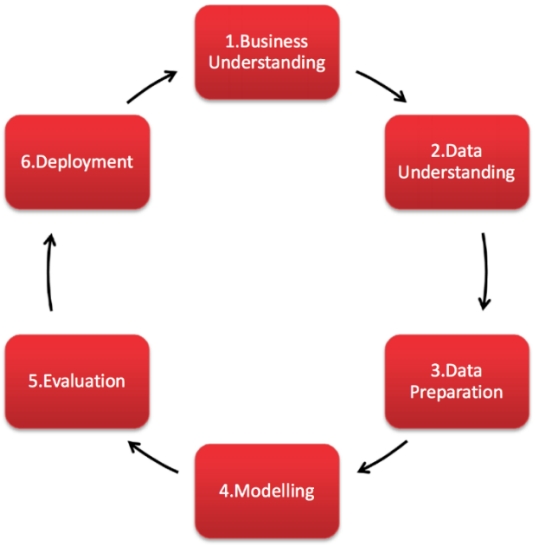


Figure 1 CRISP-DM Diagram

Structure of the Project:

This project is structured into five custom phases, each representing a key step in the analysis process while adhering to the principles of the CRISP-DM methodology (Smart Vision Europe, 2017). To ensure clarity, modularity, and traceability, each phase is implemented in a dedicated Jupyter Notebook file, reflecting a tailored approach to achieving the project's objectives effectively.

Data Preparation: Cleaning and preprocessing raw datasets to ensure data quality and reliability.

Merging Datasets: Combining multiple datasets into a unified structure for analysis.

Statistical Analysis and Exploratory Data Analysis (EDA): Identifying trends, patterns, and relationships within the data using descriptive and inferential statistical techniques.

Modeling and Sentiment Analysis: Developing predictive models and conducting sentiment analysis to uncover actionable insights.

Interactive Dashboard: Creating an intuitive dashboard to visualize key findings and enable users to interact with the data, make predictions, and forecast trends.

The deliverables include an interactive dashboard that enables stakeholders to explore historical data, compare trade balance trends, visualize production and export metrics, and forecast export values for up to 20 years. This comprehensive analysis provides a foundation for data-driven decision-making, with the ultimate goal of enhancing Ireland's agricultural sector's global competitiveness and long-term sustainability.

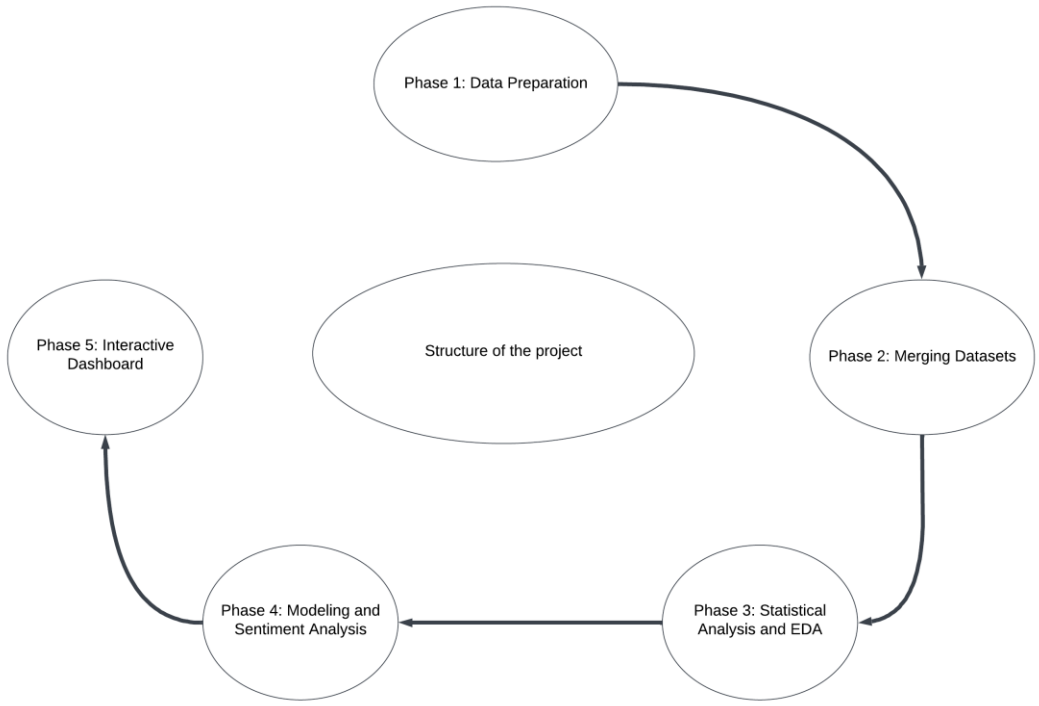


Figure 2 Structure of the project diagram

Implementation of Libraries and Code Quality

The project was programmatically explored using Python tools and libraries within a Jupyter Notebook, ensuring modularity and reusability of code. Key libraries included Pandas for data manipulation, NumPy for numerical operations, and Plotly for interactive visualizations. Statsmodels was utilized for advanced statistical analysis, including time series forecasting, joblib facilitated saving and deploying machine learning models within the dashboard. Code quality standards, such as adhering to PEP 8 (Python, 2023)and ensuring proper commenting, were maintained. Functions were designed to encapsulate repetitive tasks, enhancing code readability and scalability. Each programming decision was documented with a rationale, such as using Pandas for its high performance in handling tabular data.

Business Understanding Objectives

The objective of this analysis is to comprehensively understand Ireland's agricultural sector by identifying key factors influencing production and export performance, comparing its trade balance and production trends with other countries, and forecasting future export values through predictive models. Additionally, the project incorporates sentiment analysis to gain insights into producers' and consumers' perspectives on agricultural topics. The findings will be presented through an interactive dashboard, enabling stakeholders to make data-driven decisions and promoting Ireland's global competitiveness and sustainable growth in the agricultural sector.

Data Understanding

The datasets used for this analysis provide a comprehensive view of Ireland's agricultural sector. Key datasets include the Value of Agricultural Production, detailing production values by category (e.g., crops, livestock); Trade - Crops and Livestock, capturing export/import volumes, trade values, and trade balances; Production - Crops and Livestock, offering insights into crop yields, harvested areas, livestock counts, and production volumes; and Price Indices, reflecting producer price trends over time. Together, these datasets enable a detailed examination of production efficiency, trade performance, and pricing dynamics, forming a robust foundation for the analysis.

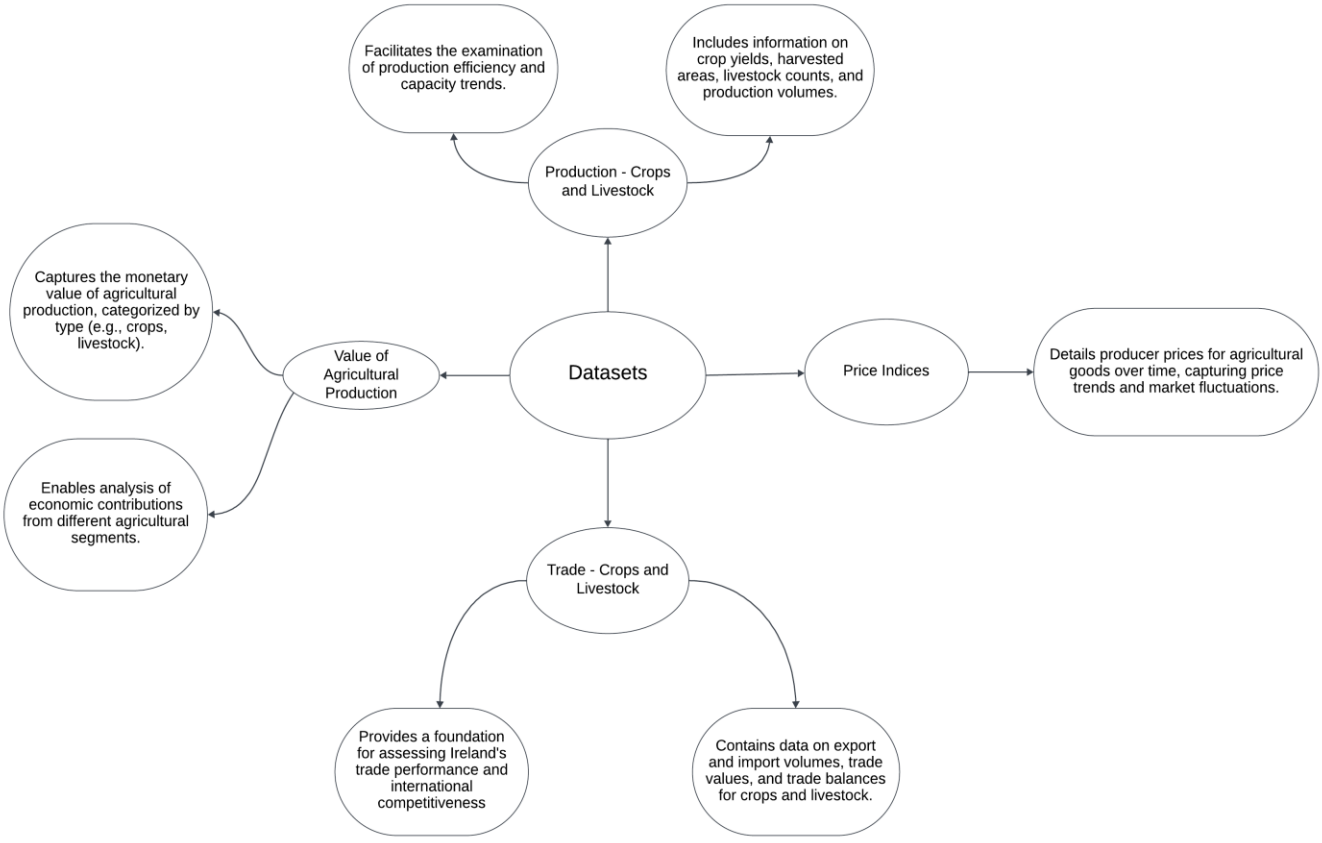


Figure 3 Data Understanding diagram

Phase 1:Data Preparation

Link of the Dataset: https://www.fao.org/faostat/en/#data

The data acquisition process involved sourcing raw data from the FAO (Food and Agriculture Organization) database, publicly accessible under a permissive license for research and education. The raw datasets and metadata are provided in the folder of datasets, ensuring compliance with licensing and proper attribution for ethical use. Challenges included extensive filtering and preprocessing to align the data with the project's scope. Key steps included reshaping the datasets from a wide format to a long format using pd.melt() to facilitate time-series analysis and merging. (pandas.pydata.org, n.d.) The Year column was cleaned for easier analysis, and missing values were addressed to ensure a clean, organized dataset ready for exploratory data analysis (EDA), statistical analysis, and modeling.

Reshaping the datasets and Filtering the Data

To focus on recent trends, data was filtered to include records from 2000 onward, ensuring relevance and actionable insights. Geographic filtering was also applied: for regional analysis, only European countries were retained to compare Ireland's agricultural sector with neighboring regions, for global benchmarking, top agricultural producers were selected based on production and export significance. These steps streamlined the dataset, reduced complexity, and ensured alignment with the project's objectives.

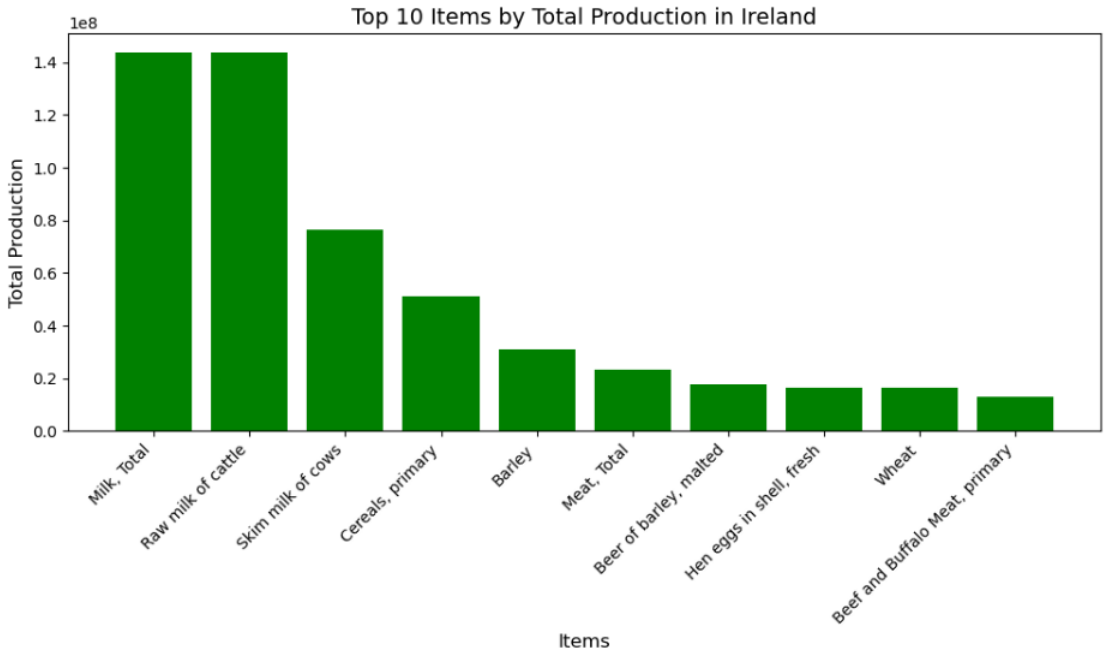


Figure 4 Bar char of the Top 10 Items production of Ireland

After visualizing Ireland's top 10 agricultural items, seven key products were selected for detailed analysis: Barley, Cereals (primary), Hen eggs (fresh), Meat (total), Milk (total), Raw milk of cattle, and Wheat. These items were chosen for their significant contribution to the country's agricultural output, streamlining the analysis for a focused exploration of Ireland's most impactful agricultural sectors. To ensure consistency across datasets, uncategorized items were manually grouped, and sub-items were aggregated using average values, aligning all datasets for accurate comparisons and insights.

Handling missing values on the dataset and Merging

The data preparation process addressed missing values and ensured datasets were ready for integration. The Production and Trade datasets were complete after reshaping and filtering, while the Price dataset initially contained significant gaps. Missing values in the Price dataset were reduced by sourcing additional data from the FAO database, and the Value dataset was completed using median imputation for numeric columns to handle outliers effectively. (Abulkhair, 2023) All processed datasets Production, Trade, Value, and Prices were organized in a folder named 'Preparation for Merge,' ensuring a standardized and modular approach for integration into a unified master dataset. This master dataset supports exploratory data analysis, statistical modeling, and effective traceability throughout the

workflow.

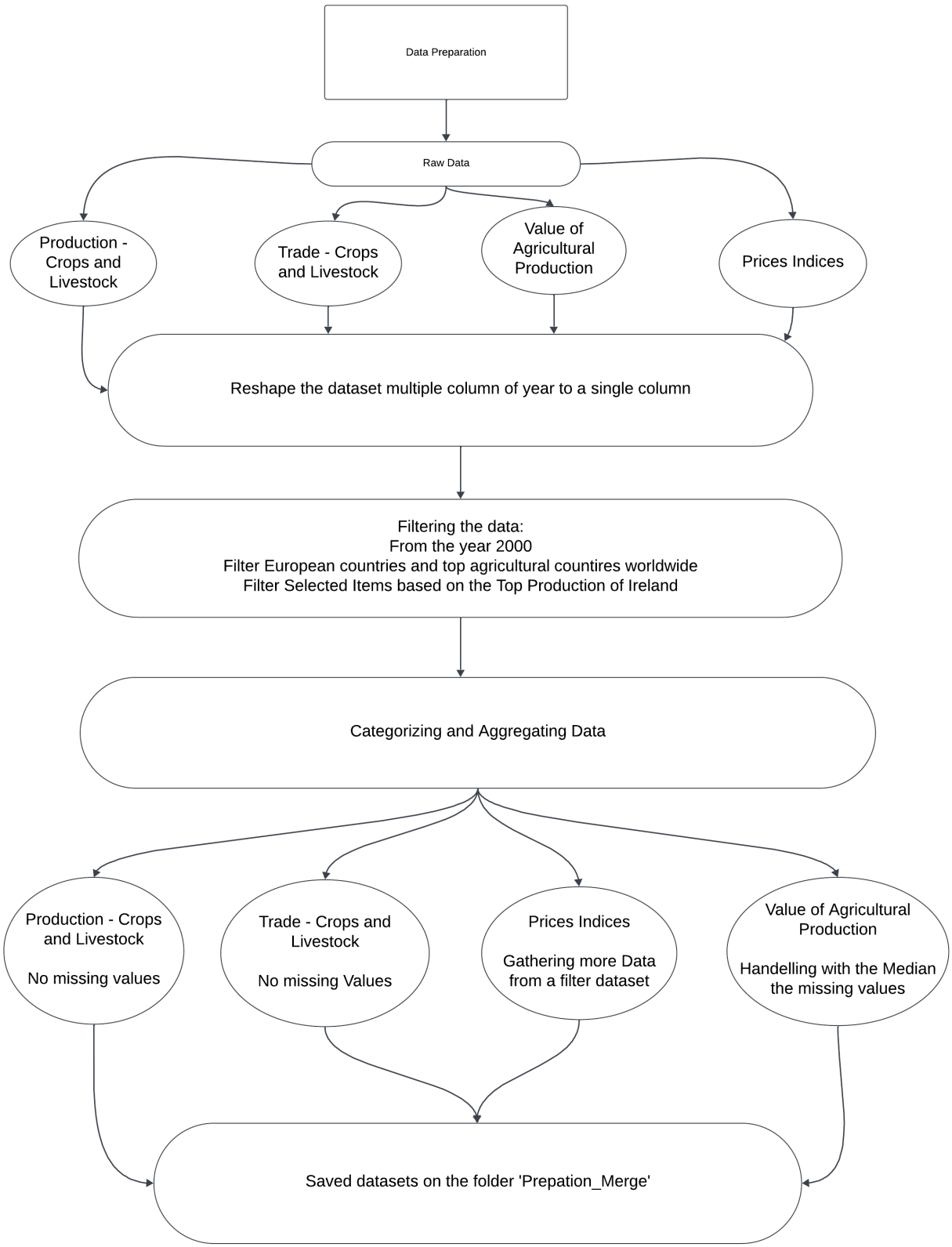


Figure 5 Data Processing diagram

Phase 2:Merging Dataset

The merging phase integrated the cleaned datasets, including Production, Trade, Value, and Prices, into a unified master dataset. This process involved sequential merging using shared keys such as Area, Item, and Year, ensuring data alignment across dimensions like production values, trade metrics, and pricing information. Each merge step was carefully executed to maintain consistency and capture relationships among variables, creating a comprehensive dataset for analysis.

After merging, missing values were identified and addressed. Numeric features with gaps were imputed using the median to preserve data reliability and mitigate the impact of outliers. The resulting master dataset, named Agriculture_data, was saved in the 'Analysis' folder as Agriculture_data.csv. This fully merged dataset provides a complete, clean, and structured foundation for subsequent exploratory data analysis (EDA), statistical analysis, and modeling.

Phase 3 : Statistical Analysis and Exploratory Data Analysis (EDA):

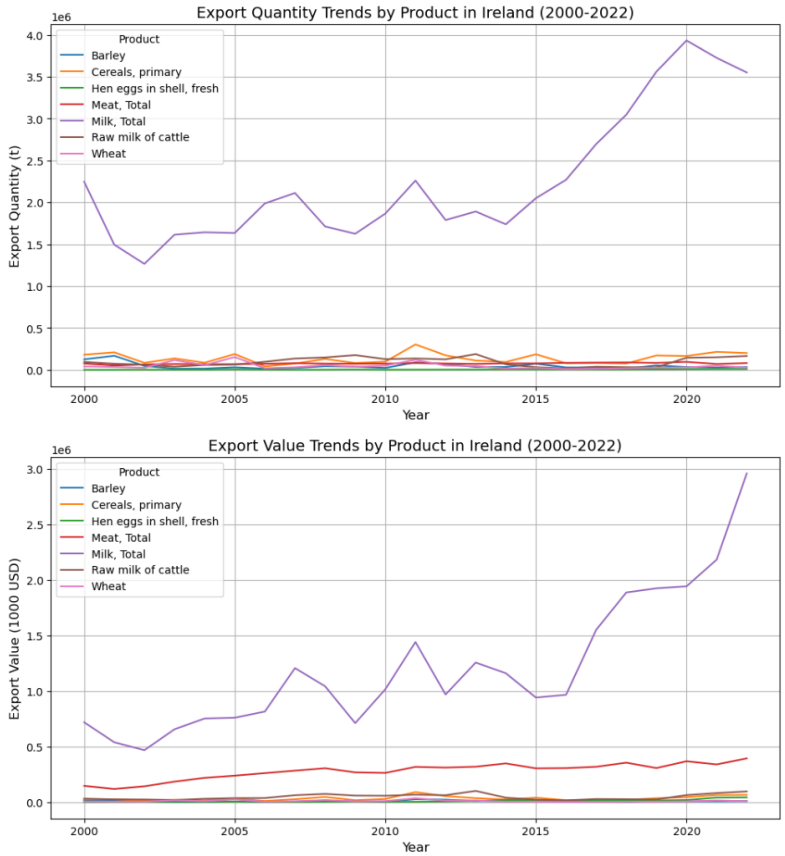


Figure 6 Export Quantity and Export value Ireland

The exploratory data analysis (EDA) examined export trends of key agricultural products in Ireland from 2000 to 2022, focusing on export quantities and values. Milk, Total consistently ranked as the most significant product in terms of both export volume and value, underscoring its pivotal role in Ireland's agricultural economy. Other products, including Meat, Total, Cereals, primary, and Raw milk of cattle, displayed stable or modest fluctuations in export quantities over the period.

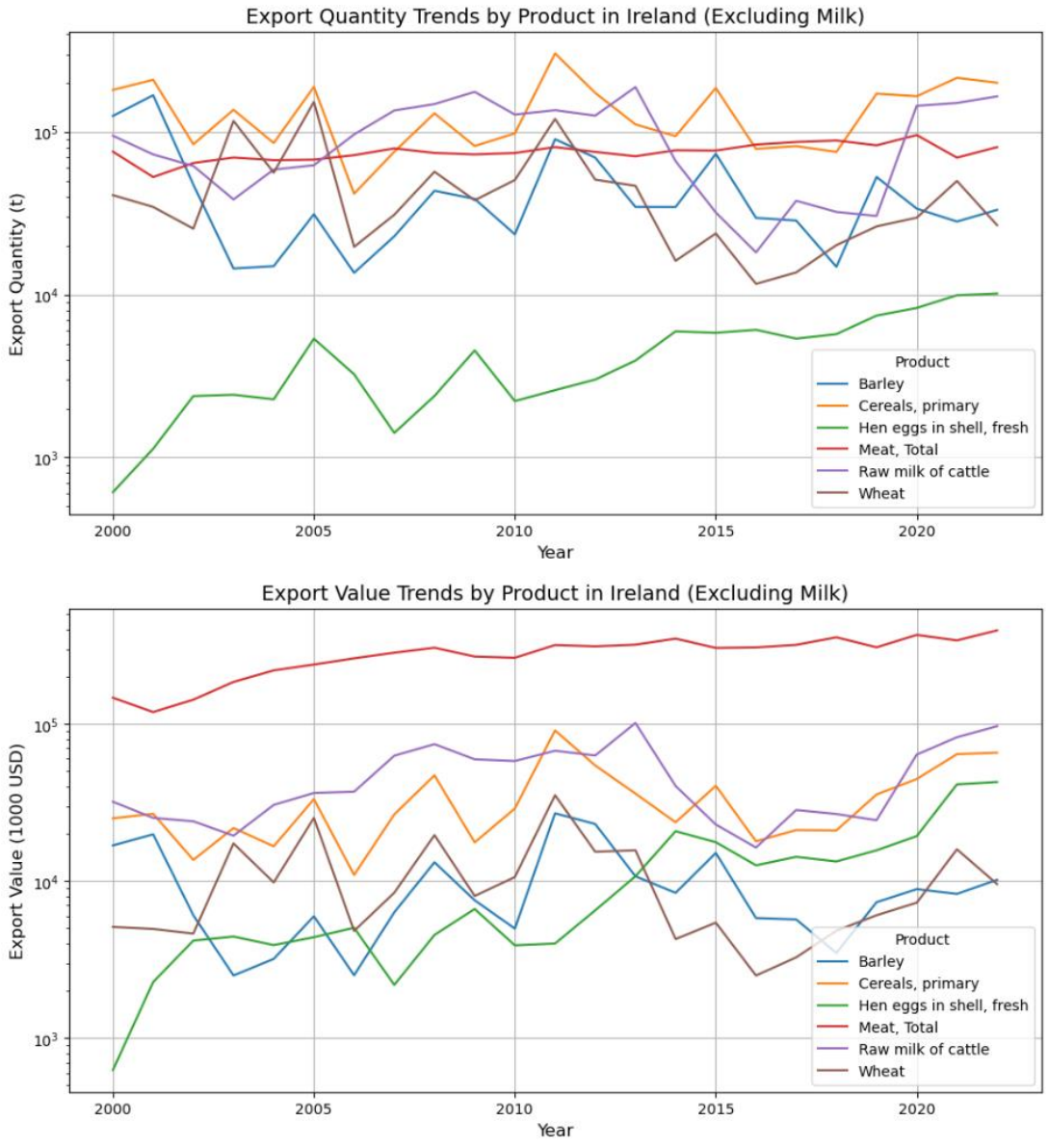


Figure 7 Export Quantity and Export value Ireland

This additional analysis examined export trends for agricultural products in Ireland, excluding Milk, Total to highlight other significant contributions. Adjusting the y-axis scale revealed variations across export quantities and values. Among export quantities, Cereals, primary and Wheat demonstrated moderate

growth, while Hen eggs in shell, fresh and Raw milk of cattle maintained stable volumes. Meat, Total emerged as a key contributor, consistently exhibiting strong export quantities over the years.

In export values, Meat, Total stood out as the most economically significant product after milk, dominating other categories. Raw milk of cattle and Cereals, primary also showed steady value increases, reflecting their importance to Ireland's agricultural economy. While Barley and Hen eggs in shell, fresh displayed stable trends, their overall contributions to export values were less significant. These findings emphasize Ireland's diverse agricultural exports and suggest potential growth opportunities for Meat, Total and Cereals, primary as key drivers of export diversification.

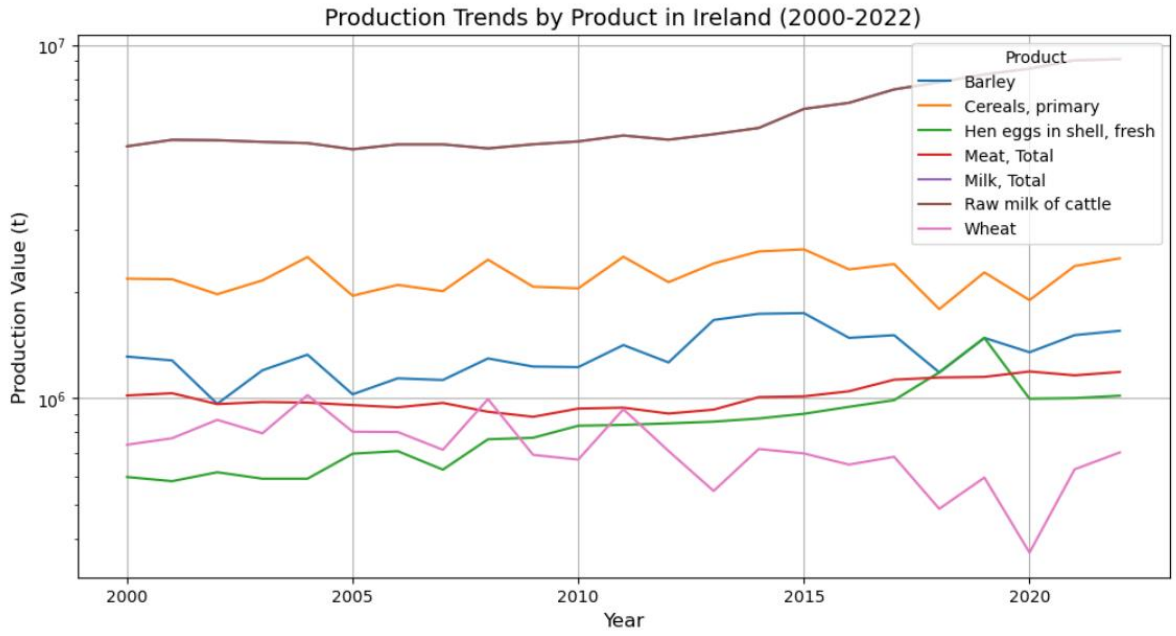


Figure 8 Production Trend of Ireland

The production trends analysis highlights the performance of Ireland's key agricultural products, including Barley, Cereals, primary, Hen eggs in shell, fresh, Meat, Total, Milk, Total, Raw milk of cattle, and Wheat, from 2000 to 2022. Milk, Total consistently recorded the highest production volumes, reinforcing its significance in Ireland's agricultural sector. Other products such as Cereals, primary and Meat, Total displayed stable production levels, with smaller fluctuations over the years. Products like Hen eggs in shell, fresh and Wheat showed greater variability but contributed smaller quantities to the overall production landscape.

Challenges Faced During Analysis

Zero values in key metrics, such as production and export quantities, were addressed by imputing the median value grouped by Year and Area, ensuring consistent and reliable trend analysis. For example, missing data for Hen eggs in shell, fresh during 2018 and 2019 was resolved to maintain data integrity. Geospatial analysis from 2000 to 2022 highlighted dominant producers like the United States, China, and India, while major exporters, including Brazil, the United States, and Germany, showcased significant contributions to global trade. Import trends revealed high demand from China and the United States,

driven by population and food diversity needs. Ireland's consistent role as an exporter emphasizes its competitive position in the global agricultural market, contributing to the understanding of evolving trade dynamics and opportunities.

Statisctical Analysis Hypothesis

The hypotheses and statistical tests were thoroughly documented in the Jupyter Notebook to avoid lengthening the report

Hypothesis 1: Export and Import Values

Paired T-Test: Initially indicated a significant difference (T-Statistic: 3.093, P-Value: 0.0023), but the normality assumption was violated (Shapiro-Wilk P-Value: 0.0000).

Wilcoxon Signed-Rank Test: A non-parametric alternative revealed no statistically significant difference (P-Value: 0.7645).

Conclusion

Based on the Wilcoxon test, which is more appropriate given the non-normal distribution of data, we conclude that there is no significant difference between Ireland's export and import values. This highlights the balanced trade dynamics of Ireland's agricultural sector.

Hypothesis 2 : Trade Balance Trends Over Time

Both the ANOVA test (F-Statistic: 0.802, P-Value: 0.7265) and the Kruskal-Wallis (Peters, 2023) test (H-Statistic: 12.999, P-Value: 0.9332) indicate no significant changes in Ireland's trade balance (export - import values) over the years. These results suggest that the trade balance has remained relatively stable, with no major fluctuations or trends observed, highlighting consistency in Ireland's agricultural trade dynamics over time.

The ANOVA test was selected to evaluate whether there were significant differences in Ireland's trade balance (export - import values) across multiple years, as it is a parametric test suitable for comparing means when normality and homogeneity of variance assumptions are met. To account for potential violations of these assumptions, the Kruskal-Wallis test, a non-parametric alternative, was also applied. This dual approach ensured a robust analysis by validating results under both parametric and non-parametric frameworks, confirming the stability of trade balance trends over time.

Hypothesis 3 : Production vs. Export Values

The analysis of the relationship between production and export values was conducted using Pearson and Spearman correlation tests. The Pearson correlation coefficient was 0.590 with a P-Value of 0.0000, indicating a significant moderate positive linear relationship between production and export values. Similarly, the Spearman correlation coefficient was 0.600 with a P-Value of 0.0000, confirming a significant monotonic correlation. These results suggest that higher production volumes are associated with increased export values, highlighting the critical role of production in driving trade performance.

Hypothesis 4: Producer Price Variations

The analysis of producer price variations across items was conducted using One-Way ANOVA and Kruskal-Wallis tests. Both tests confirmed significant differences in producer prices, with ANOVA yielding

an F-Statistic of 993.308 and a P-Value of 0.0000, and the Kruskal-Wallis test producing an H-Statistic of 2486.716 with a P-Value of 0.0000. These results highlight that the mean producer prices differ significantly among items such as Hen eggs in shell, fresh, Meat, Total, Cereals, primary, and Barley. The findings suggest that economic factors such as production costs, market demand, and supply chain dynamics contribute to these disparities.

Figure 9 Box Plot of Producer Price Ireland

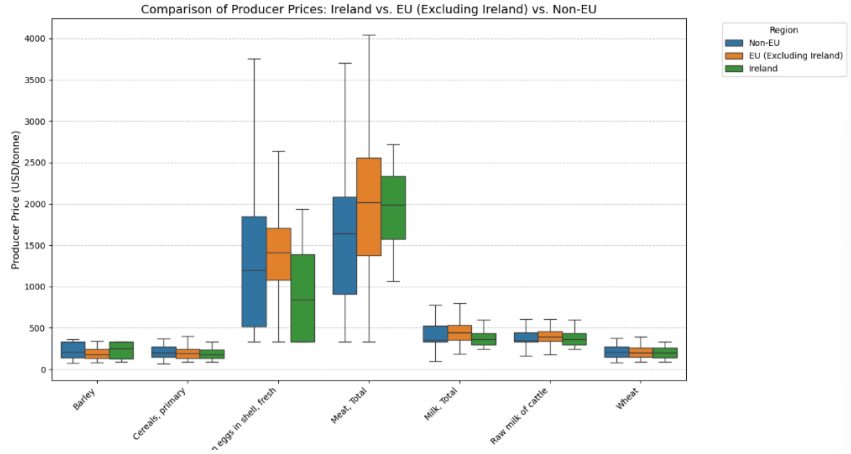


Figure 10 Box Plot of Comparison Producer Price Ireland vs Eu vs Non-EU

Box Plot and Bar Chart Insights

The visual analysis of producer price variations using box plots and bar charts revealed significant differences across items and regions. Items like Meat, Total and Hen eggs in shell, fresh showed higher price variability and averages, particularly in non-EU countries, while items like Barley and Cereals, primary exhibited more consistent pricing across regions. The bar chart further emphasized these trends, offering an accessible comparison of average prices and highlighting the regional differences, with non-EU countries generally showing higher producer prices. These insights provide a clear understanding of the factors influencing pricing variability and competitiveness across items and regions.

Hypothesis 5: Country-Level Performance in Exports

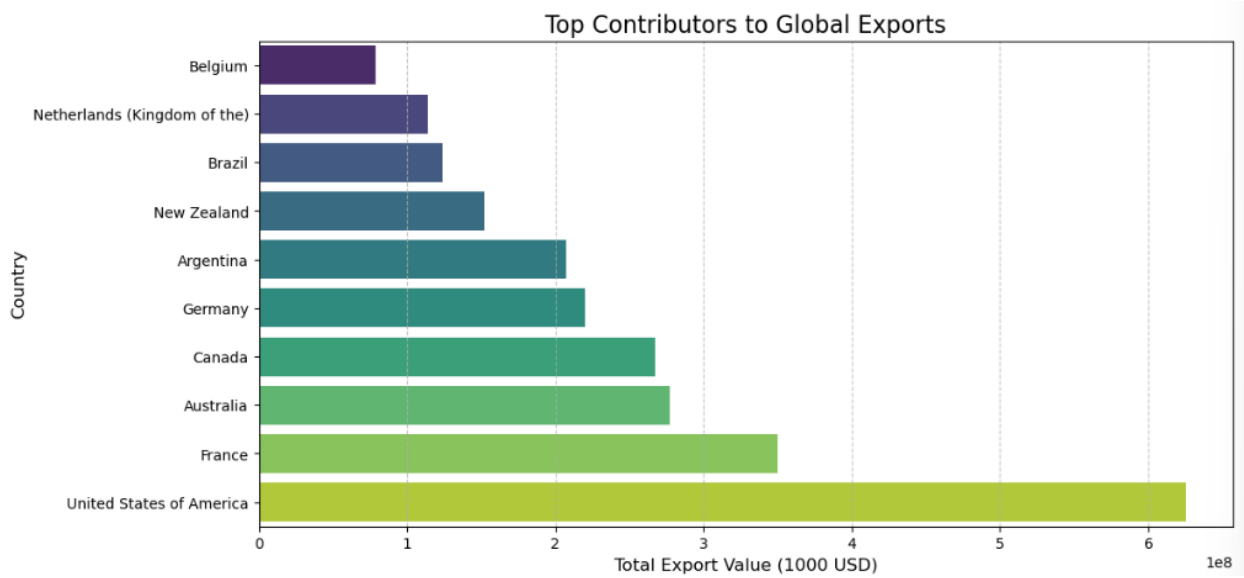


Figure 11 Bar Chart of Top Countries exports

The Chi-Squared Goodness-of-Fit Test results strongly rejected the null hypothesis, with a Chi-Squared Statistic of 1280312968.734 and a p-value of 0.0000, indicating that certain countries contribute disproportionately to the global export value. This finding indicates that certain countries contribute disproportionately to global export values. Key contributors include nations such as the United States, France, and Australia, which dominate due to factors such as higher production capacities, competitive pricing strategies, and favorable trade agreements.

These results highlight an imbalance in global export contributions, where a small group of countries wields significant influence. This underscores the importance of strategic policymaking to address disparities by supporting underperforming regions to enhance their export capabilities. Additionally, leveraging the strengths of top-performing countries can further solidify their competitive edge in global markets while ensuring a more balanced and inclusive global trade environment.

Hypothesis 6: Export Growth Over Years

The analysis tested whether the mean export value changes significantly over the years for a given country. The findings showed no evidence of significant changes in export values over time, suggesting that exports have remained relatively stable across the observed years.

This stability highlights the potential role of consistent trade policies and market conditions in maintaining steady export performance. However, further investigation could explore specific sectors or commodities to identify opportunities for growth or address any fluctuations within individual categories.

Hypothesis 7: GPV vs. GPV_Const

The analysis compared the Gross Production Value (GPV) and its inflation-adjusted counterpart (GPV_Const) to determine if significant differences exist between them. Both the Paired T-Test and the Wilcoxon Signed-Rank Test were applied. The results indicated no significant difference between GPV and GPV_Const, as the null hypothesis could not be rejected. This suggests that inflation adjustments did not significantly alter the observed values in the dataset.

These findings imply that inflation has had a minimal impact on the comparative trends of GPV and GPV_Const over time. However, the consistent trends across countries indicate that inflation-adjusted GPV remains a reliable metric for assessing production values, supporting its use in comparative analyses. Further research could explore whether inflation impacts specific sectors or regions differently.

Evidence-Based Recommendations

Ireland's agricultural trade system has proven to be stable and reliable, ensuring balanced export and import dynamics and consistent trade balances over time. This stability reflects effective governance and trade policies, making Ireland a dependable player in the global agricultural market. To further enhance performance, Ireland should focus on increasing production efficiency and capacity, as higher production volumes are directly linked to export growth. Additionally, addressing price disparities across various products can improve competitiveness and profitability in international markets.

Leveraging the strengths of high-performing export countries while supporting underperforming regions can create a more balanced and equitable trade environment. The stability in export trends provides a strong foundation, but there is an opportunity to explore untapped markets and diversify product offerings to drive further growth. Using inflation-adjusted metrics, such as GPV_Const, ensures a clear understanding of production values and facilitates informed decision-making for future strategies.

Phase 4: Modeling

For predictive modeling, several machine learning libraries were employed to ensure robust performance and accurate results. LabelEncoder from sklearn.preprocessing was used for encoding categorical features, while train_test_split from sklearn.model_selection facilitated splitting datasets into training and testing subsets. LinearRegression and Ridge from sklearn.linear_model were used for linear modeling, with metrics like mean_squared_error, mean_absolute_error, and r2_score from sklearn.metrics applied for evaluation. RandomForestRegressor was implemented for ensemble modeling, and RandomizedSearchCV was used to optimize hyperparameters. Additionally, kurtosis from scipy.stats provided insights into the distribution of target variables, ensuring suitability for modeling.

The objective of this modeling exercise was to develop a predictive model for export value in the agricultural sector. The goal was to understand the relationships between key features and predict export value effectively, leveraging log-transformed data for improved stability and model performance.

Categorical features Area and Item were label-encoded for compatibility with models. Label encoding assigns numeric codes to each unique category in the data, replacing textual or categorical values with

corresponding integers. This method was chosen over one-hot encoding to maintain simplicity and avoid introducing high-dimensional sparse data, especially as the models used can interpret ordinal relationships effectively.

```
({'Argentina': 0,
  'Australia': 1,
  'Austria': 2,
  'Belgium': 3,
  'Brazil': 4,
  'Canada': 5,
  'China': 6,
  'Denmark': 7,
  'Finland': 8,
  'France': 9,
  'Germany': 10,
  'Hungary': 11,
  'India': 12,
  'Ireland': 13,
  'Italy': 14,
  'Netherlands (Kingdom of the)': 15,
  'New Zealand': 16,
  'Poland': 17,
  'Romania': 18,
  'Spain': 19,
  'Sweden': 20,
  'United States of America': 21},
 {'Barley': 0,
  'Cereals, primary': 1,
  'Hen eggs in shell, fresh': 2,
  'Meat, Total': 3,
  'Milk, Total': 4,
  'Raw milk of cattle': 5,
  'Wheat': 6})
```

Figure 12 Mapping of Countries and Items

Kurtosis Analysis (Sharma, 2020)

Kurtosis was computed for numerical features before and after log transformation. High kurtosis indicates heavy tails and outliers, which can negatively affect linear models. Log transformation reduced kurtosis, creating a more normal distribution and enhancing the reliability of regression models.

	Original Kurtosis	Log-Transformed Kurtosis
Year	-1.20	-1.20
Area	-1.20	-1.20
Item	-1.25	-1.25
Production Value (t)	34.24	-0.03
Export Quantity (t)	50.36	0.68
Export Value (1000 USD)	53.49	1.35
Import Quantity (t)	102.11	0.37
Import Value (1000 USD)	123.65	0.85
GPV (1000 USD)	69.56	-0.11
GPV_Const (1000 USD)	43.26	-0.36
Producer Price (USD/tonne)	3.08	-0.47

Figure 13 Kurtosis Visualization

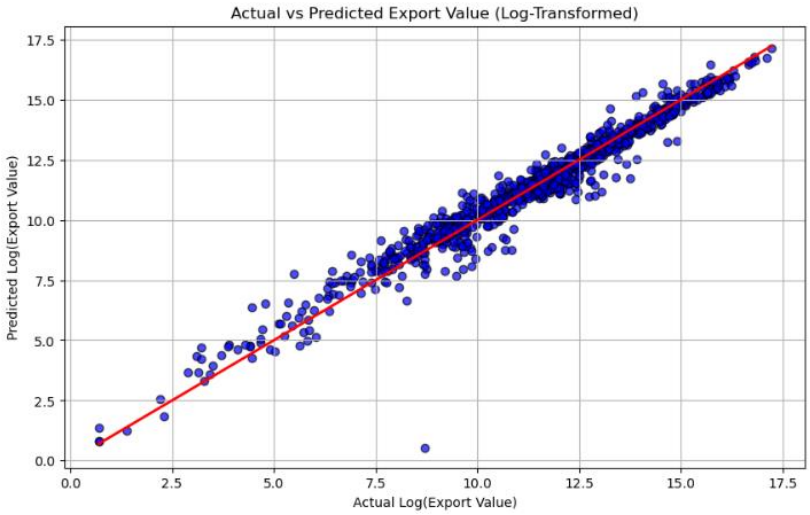
Train-Test Split:

The dataset was split into training (70%) and testing (30%) sets to evaluate model performance on unseen data. Features were selected for their relevance: Year for temporal trends, Area and Item for

regional and product-specific variations, Production Value and Export Quantity for their direct impact on export capacity, and Producer Price for its influence on revenue

Linear Regression

The first model trained was Linear Regression, which achieved a high accuracy with an R^2 score of 0.96 on the test set. To validate the robustness of the model, we implemented 5-fold cross-validation, resulting in an average R^2 score of 0.936 with a standard deviation of 0.022. These results confirmed that the model performed consistently well across different subsets of the data, demonstrating its ability to generalize effectively to unseen data while maintaining stable and reliable predictions.



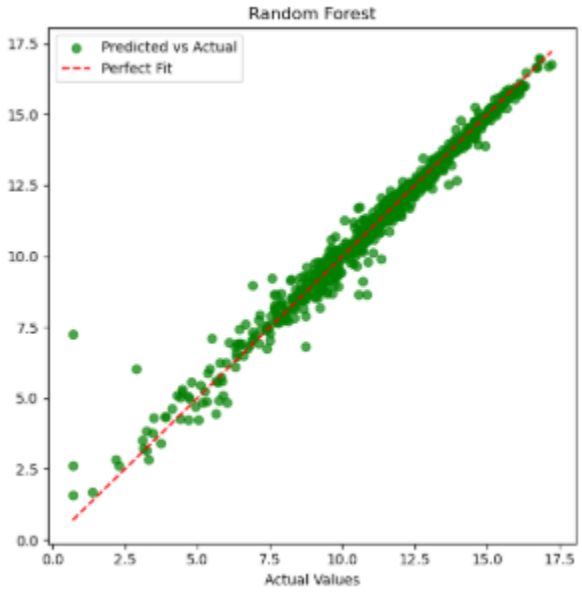
	Feature	Coefficient
4	Export Quantity (t)	0.86
5	Producer Price (USD/tonne)	0.83
3	Production Value (t)	0.07
0	Year	0.01
1	Area	-0.01
2	Item	-0.01,

-33.03873657357071)

Figure 14 Plot of Linear Regression and visualization of Feature Coefficients

The coefficients from the Linear Regression model highlight the relative impact of each feature on export value. Export Quantity (t) is the most influential factor, with a coefficient of 0.86, indicating that increasing export quantity significantly raises export value. Similarly, Producer Price (USD/tonne) strongly contributes to export value with a coefficient of 0.83, reflecting the importance of pricing

Random Forest with hyperparameters



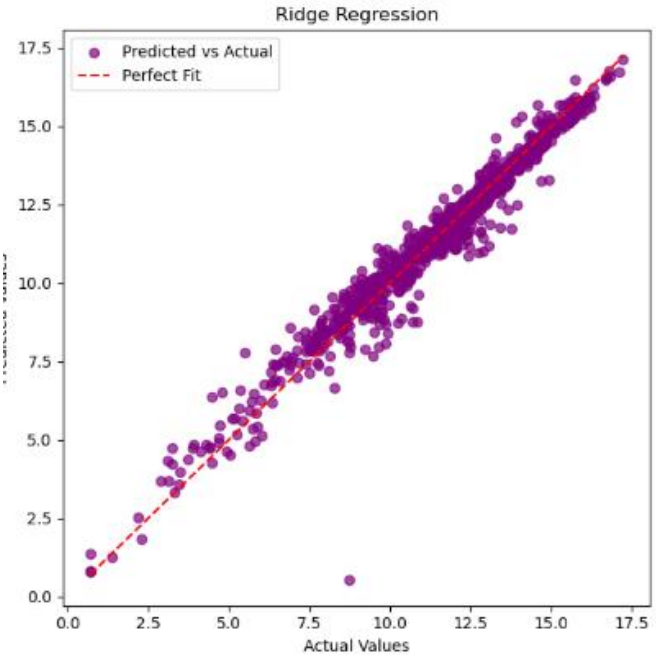
	Feature	Importance
4	Export Quantity (t)	0.89
5	Producer Price (USD/tonne)	0.07
3	Production Value (t)	0.02
2	Item	0.01
0	Year	0.01
1	Area	0.00

Figure 15 Plot of Random Forest and visualization of Feature Coefficients

The Random Forest model, optimized with hyperparameters including n_estimators = 400 and min_samples_split = 15, demonstrated excellent predictive performance. It achieved a low Mean Squared Error (MSE) of 0.1515 and a Mean Absolute Error (MAE) of 0.2052, indicating minimal prediction errors. Additionally, the model achieved a high R-squared (R^2) value of 0.9783, explaining 97.83% of the variance in export value. These results highlight the effectiveness of Random Forest in capturing complex relationships and its robustness after hyperparameter tuning, making it a reliable choice for accurate predictions in this context.

Despite its strong accuracy, the Random Forest model demonstrates a limitation in feature interpretation. Export Quantity (t) is identified as the most influential feature (importance score: 0.89), while Producer Price (USD/tonne), which theoretically should have a significant impact on export value, is assigned a much lower importance score (0.07). This limitation arises because Random Forest models prioritize capturing non-linear relationships and may assign disproportionate importance to correlated features, such as Export Quantity and Producer Price. While Random Forest excels in predictive power, it lacks the interpretability provided by linear models, which can offer clearer insights into feature contributions.

Ridge Regression



	Feature	Coefficient
4	Export Quantity (t)	0.86
5	Producer Price (USD/tonne)	0.82
3	Production Value (t)	0.07
0	Year	0.01
1	Area	-0.01
2	Item	-0.01

Figure 16 Figure 16 Plot of Ridge Regression and visualization of Feature Coefficients

The Ridge Regression model was trained to mitigate overfitting by penalizing large coefficients, ensuring that all features contribute meaningfully to the predictions. After hyperparameter tuning, the optimal regularization parameter was determined to be alpha = 0.01. The model achieved a Mean Squared Error (MSE) of 0.31 and a Mean Absolute Error (MAE) of 0.36, with an R-squared (R^2) value of 0.96, indicating that it explains 96% of the variance in export value.

Ridge Regression effectively captured the relationship between features and the target while maintaining interpretability through its coefficients. For instance, Export Quantity (t) and Producer Price (USD/tonne) were identified as the most impactful features, with coefficients of 0.86 and 0.83, respectively, demonstrating their strong influence on export value. Compared to Random Forest, Ridge Regression offers the advantage of interpretability, clearly quantifying the contribution of each feature. However, its predictive power is slightly lower, as it may not capture complex, non-linear relationships as effectively as ensemble methods.

Comparison of the model

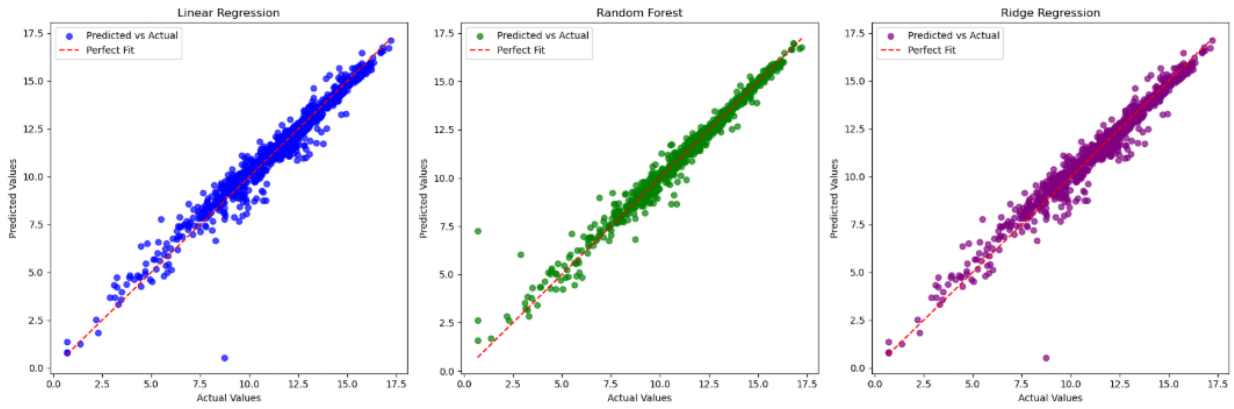


Figure 17 Comparison of the Three models

After evaluating the performance and characteristics of the models, Linear Regression was chosen as the final model for this analysis. Despite Ridge Regression's ability to penalize overfitting and Random Forest's exceptional predictive accuracy, Linear Regression provides a balance between accuracy and interpretability (Fortmann-Roe, 2012)

An interactive dashboard was created to test export value predictions, allowing users to input features like Country, Item, Export Quantity, and more. The dashboard validates inputs, applies necessary transformations, and provides real-time predictions, making it a practical tool for testing and exploring scenarios.

Forecast Model

We implemented a forecasting pipeline for predicting the export values of "Barley" in "Ireland" using historical data and user-provided inputs. The dataset was filtered based on the user's selections, and the data was preprocessed to include the latest user-provided export value. A SARIMA model was then tuned by performing a manual grid search over non-seasonal (p, d, q) and seasonal (P, D, Q, m) parameters, optimizing for the lowest Akaike Information Criterion (AIC). The optimal parameters identified were (0, 1, 0) for non-seasonal order and (1, 1, 0, 12) for seasonal order, effectively capturing annual seasonality. (Artley, 2022)

Using the optimized SARIMA model, we forecasted export values for the next 20 years. A connected visualization was created to seamlessly integrate historical data and forecasts, using pd.concat to handle recent changes in pandas functionality. The final plot displayed historical export data with a continuous transition into forecasted values, clearly differentiating them using distinct line styles. This pipeline dynamically integrates user inputs and provides an accurate, visually intuitive forecasting solution.

Sentimental Analysis

Data from Reddit subreddits on agriculture and farming was analyzed using sentiment analysis to uncover challenges in the Milk and Meat sectors, focusing on negative sentiments.

Milk

Negative views on milk relate to concerns over raw milk safety, low farmer prices, market inefficiencies, and declining demand due to plant-based alternatives. Solutions include better safety guidelines, fairer pricing for farmers, and product diversification to meet evolving preferences.

Meat

Concerns about meat involve unhealthy farming practices, production declines, and competition from lab-grown or plant-based alternatives. Addressing these requires greater transparency, supportive policies for small farmers, and improved product quality or alternatives.

Interactive Dashboard

A dashboard was created to analyze agricultural data, predict export values, and forecast future trends, enabling data-driven decision-making in the agricultural sector. The tool allows users to select a specific country and commodity, input or adjust production metrics (e.g., export quantity, producer price, production value), and explore historical data through a dynamic, downloadable table. Using a pre-trained linear regression model, the dashboard predicts export values based on user inputs and employs a SARIMA model to generate 20-year forecasts, presenting trends through interactive Plotly visualizations. Additionally, the dashboard includes an Interactive Visualizations section, where users can explore production and export trends across countries, trade balances, and other aggregated metrics, further enhancing the tool's ability to provide actionable insights. (jun, 2024)

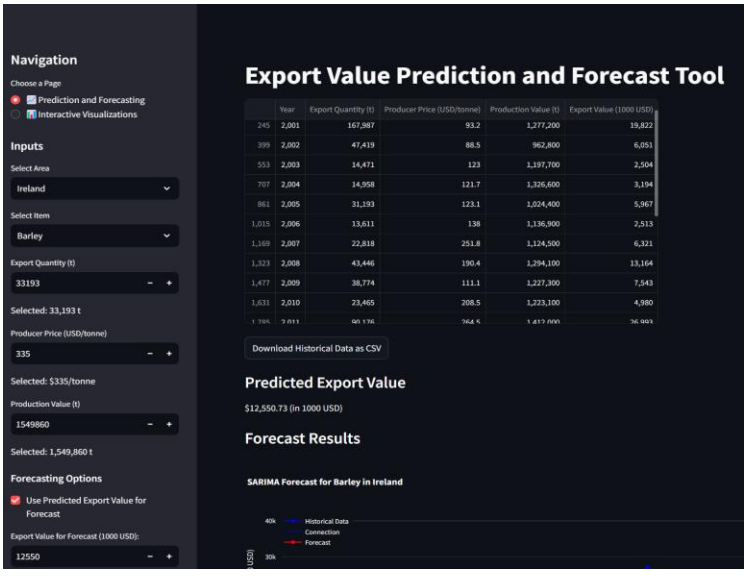


Figure 18 Interactive Dashboard Export value table and prediction

Users can view the historical data table, copy the latest year's values, input them into the fields on the left, instantly generate a predicted export value, and download the table as a CSV file for personal analysis.

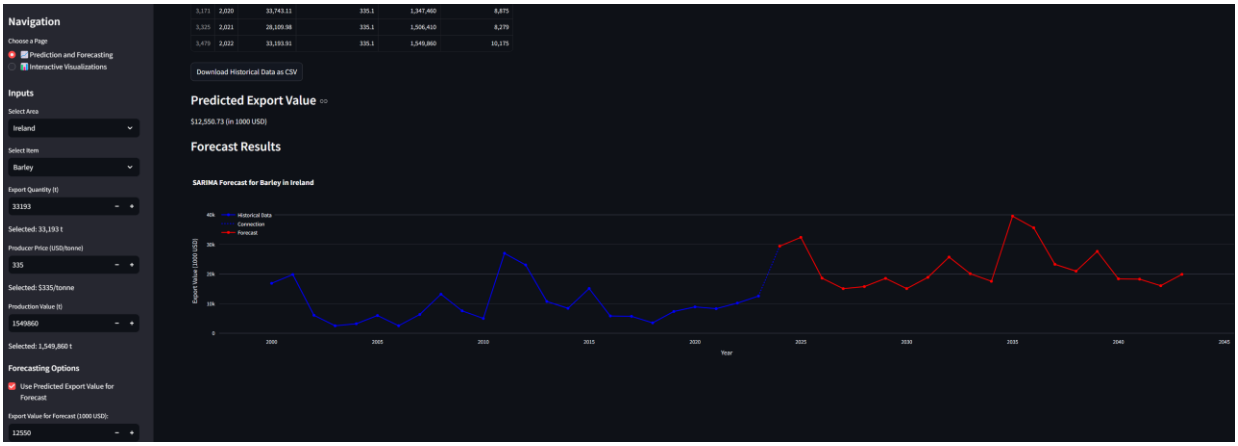


Figure 19 Forecasting the result of the Predicted Export value

The dashboard automatically inputs the predicted export value into the forecasting section, allowing directly to click the "Generate Forecast" button to create a 20-year forecast.

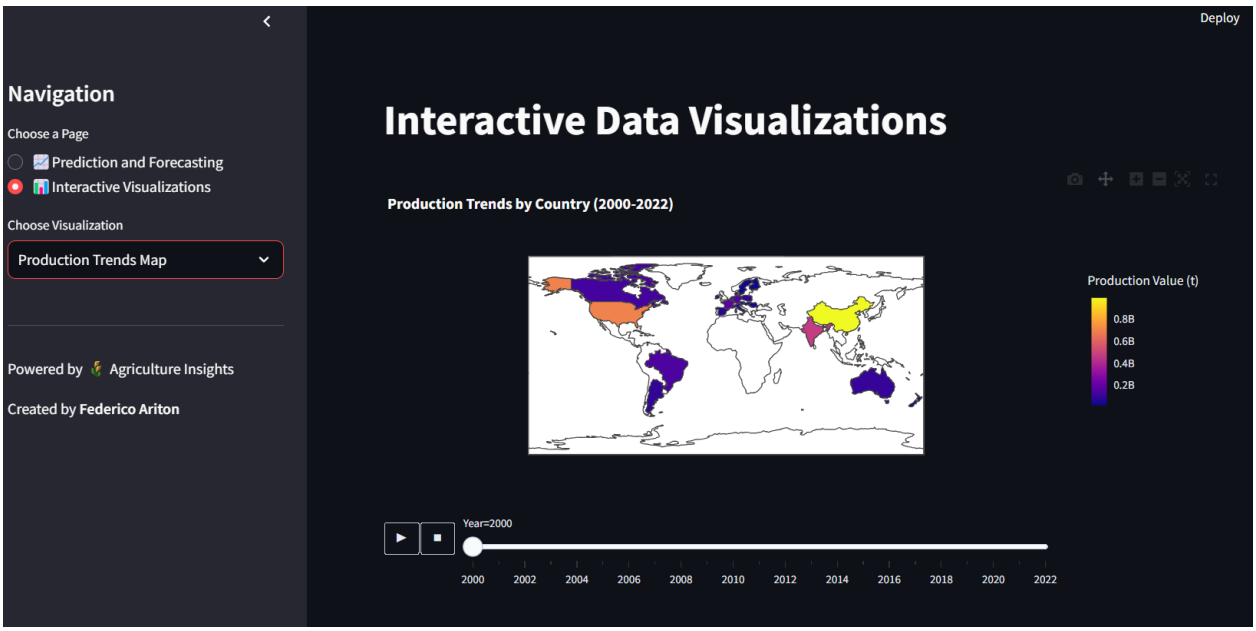


Figure 20 Interactive section showing the map for interaction

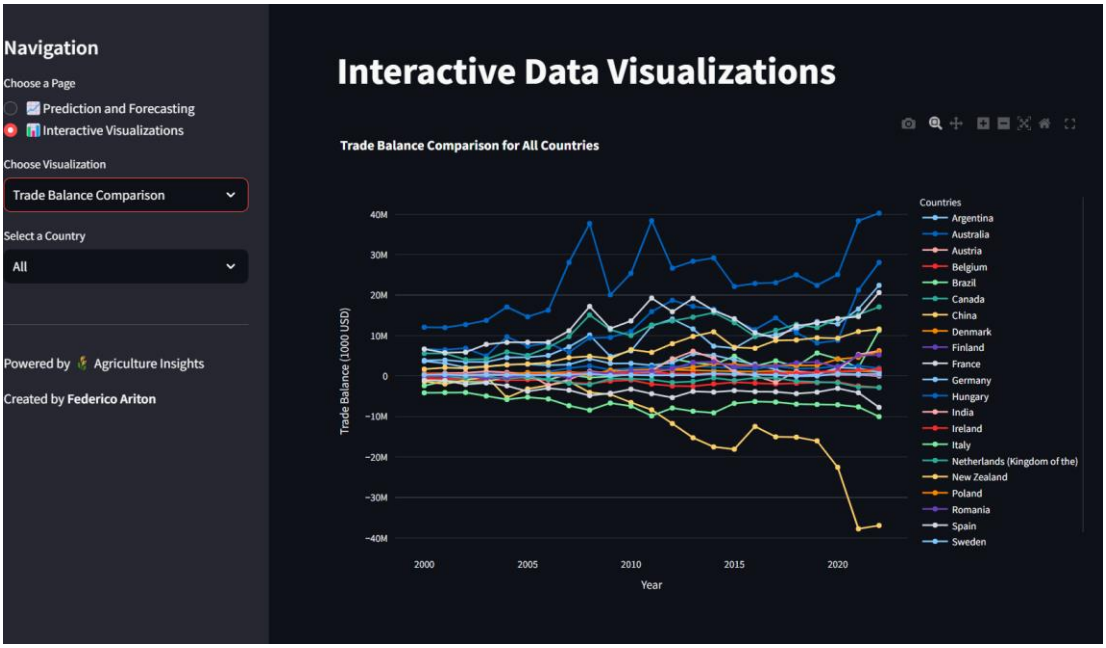


Figure 21 Traba balance comparison

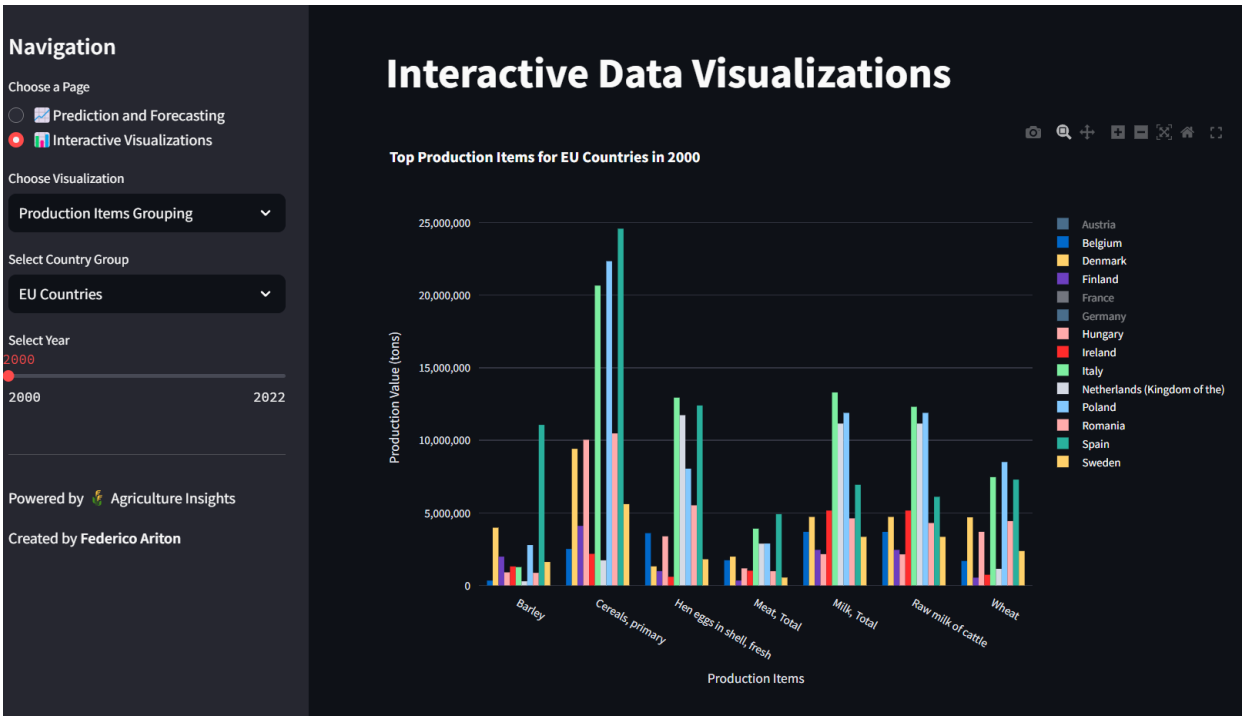


Figure 22 Top production items comparison

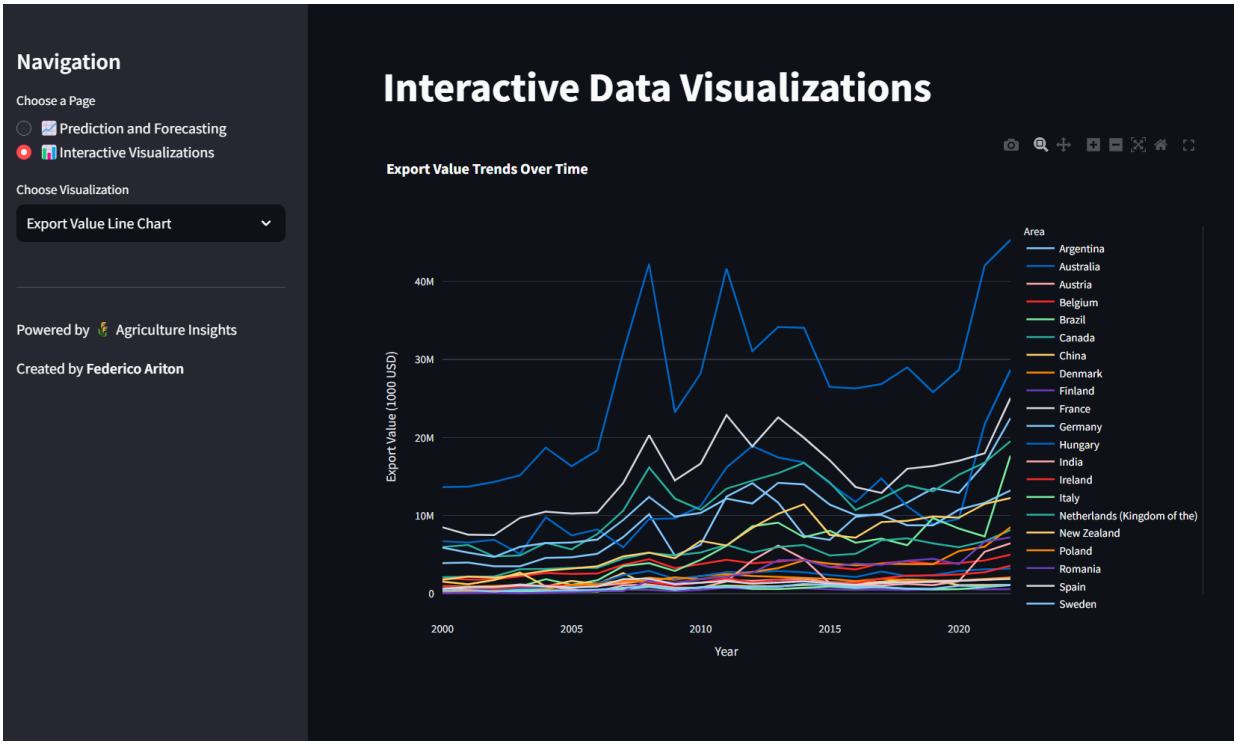


Figure 23 Trends over time Comparison

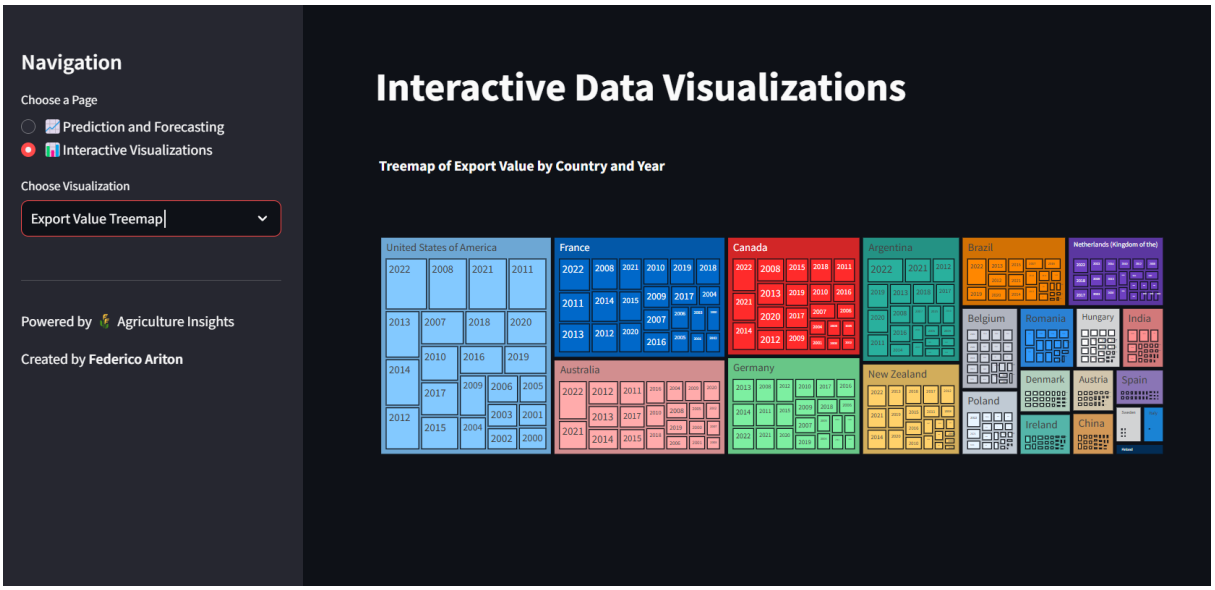


Figure 24 Interactive Treemap of Export value

Optimization

To make the dashboard faster and more efficient, several improvements were applied. Caching was used to avoid repeating the same calculations, making the app respond more quickly. Vectorized operations in Pandas sped up data processing by handling large amounts of data all at once. To save memory, we reduced the size of the data types used, and tools like cProfile helped find and fix slow parts of the code. For forecasting, the SARIMA model was fine-tuned to ensure accurate predictions without taking too

much time to run. By balancing speed and memory use, the dashboard can now handle bigger datasets without slowing down or using excessive resources.

Conclusion

This project comprehensively explored the data analytics pipeline, including data preparation, merging datasets, and conducting exploratory and statistical analyses. Multiple hypotheses were formulated and tested to gain deeper insights into the data and understand underlying relationships. We implemented a Linear Regression model for prediction, a SARIMA model for trend forecasting, and Sentiment Analysis to capture public perception. Finally, an interactive dashboard was created to present key insights, enabling stakeholders to make informed, data-driven decisions effectively.

Reference

Smart Vision Europe (2017). *What is the CRISP-DM methodology?* [online] Smart Vision - Europe. Available at: https://www.sv-europe.com/crisp-dm-methodology/.

Python, R. (2023). *How to Write Beautiful Python Code With PEP 8 – Real Python*. [online] realpython.com. Available at: https://realpython.com/python-pep8/.

Abulkhair, A. (2023). *Data Imputation Demystified | Time Series Data*. [online] Medium. Available at: https://medium.com/@aaabulkhair/data-imputation-demystified-time-series-data-69bc9c798cb7.

pandas.pydata.org. (n.d.). *Reshaping and pivot tables — pandas 2.2.2 documentation*. [online] Available at: https://pandas.pydata.org/docs/user_guide/reshaping.html.

Peters, M. (2023). *Understanding the Kruskal-Wallis Test - Mirko Peters — Data & Analytics Blog*. [online] Medium. Available at: https://medium.com/data-analytics-magazine/understanding-the-kruskal-wallis-test-a4c73e7c6668 [Accessed 29 Dec. 2024].

Sharma, A. (2020). *Skewness & Kurtosis Simplified - Towards Data Science*. [online] Medium. Available at: https://medium.com/towards-data-science/skewness-kurtosis-simplified-1338e094fc85 [Accessed 29 Dec. 2024].

Fortmann-Roe, S. (2012). *Understanding the Bias-Variance Tradeoff*. [online] Fortmann-roe.com. Available at: http://scott.fortmann-roe.com/docs/BiasVariance.html.

Artley, B. (2022). *Time Series Forecasting with ARIMA , SARIMA and SARIMAX*. [online] Medium. Available at: https://medium.com/towards-data-science/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6 [Accessed 29 Dec. 2024].

jun, S. (2024). *Mastering Dashboard Design: From Good to Unmissable Data Visualizations*. [online] Medium. Available at: https://medium.com/@tjdus92422/mastering-dashboard-design-from-good-to-unmissable-data-visualizations-e3a1b5ee108a.