**Comprehensive Analysis of Ireland's Agriculture and Global Comparisons**

**Overview**

Agriculture has played a foundational role in the development of human civilizations, serving as the bedrock of societies and economies worldwide. Over time, the sector has evolved significantly, integrating advanced technologies such as data analytics, machine learning, and artificial intelligence to enhance productivity, sustainability, and resilience. These technological innovations enable deeper insights into production trends, trade dynamics, and livestock statistics, paving the way for more efficient and informed decision-making.

This exercise undertakes a comprehensive analysis of Ireland’s agricultural sector, focusing on key crops—maize and potatoes—and livestock—cattle, chickens, and sheep. Using data sourced from the Food and Agriculture Organization’s FAOSTAT database, this research evaluates historical and current production trends, trade balances, and livestock statistics. To provide broader context, this exercise benchmarks Ireland’s performance against selected European countries: Belgium, Germany, Portugal, Spain, and Poland. By leveraging forecasting methodologies, this research aims to identify actionable insights and offer recommendations for optimizing Ireland’s agricultural sector to meet economic, social, and ecological objectives.

**Objective**

The primary objective of this exercise is to analyze Ireland’s agricultural performance in the context of global trends and offer evidence-based recommendations for its enhancement. The exercise concentrates on specific crops—maize and potatoes—and key livestock—cattle, chickens, and sheep—to achieve the following goals:

1. **Analyze Agricultural Production Trends:**
   * Investigate historical and current production data for the selected crops and livestock.
   * Identify patterns, anomalies, and growth trends to gain a comprehensive understanding of Ireland’s agricultural landscape.
2. **Compare Ireland’s Agricultural Sector Globally:**
   * Benchmark Ireland’s agricultural performance against other countries using critical metrics such as yield, production efficiency, and trade volumes.
3. **Conduct Forecasting and Predictive Analysis:**
   * Employ machine learning models to predict future production trends and highlight potential risks and opportunities.
4. **Provide Evidence-Based Recommendations:**
   * Develop actionable recommendations targeted at policymakers, farmers, and stakeholders to enhance productivity, trade balance, and economic outcomes in the sector.

This ‘research’ not only seeks to highlight Ireland’s strengths but also identifies areas for improvement, aiming to foster a more sustainable, competitive, and resilient agricultural sector.

**Data Source**

This research relies on robust datasets from FAOSTAT, the Food and Agriculture Organization’s comprehensive statistical database. The primary dataset used for this analysis was obtained from [FAOSTAT’s database](https://www.fao.org/faostat/en/#data/QCL) and encompasses agricultural data spanning from 1961 to 2023 for all countries, identified through the M49 coding system. This dataset forms the backbone of this exercise, providing insights into key agricultural domains such as livestock statistics and crop production. The focus areas for this research include cattle, chickens, sheep, maize, and potatoes, offering a holistic view of these pivotal agricultural products.

To enhance clarity and ensure consistent interpretation, an additional dataset, titled "Unit Definitions," was sourced from FAOSTAT’s Definitions section ([FAOSTAT Definitions](https://www.fao.org/faostat/en/#definitions)). This complementary dataset includes metadata such as unit names and descriptions, ensuring uniformity in the analysis. For instance, units such as “ha/cap” (hectares per capita) provide nuanced understanding during comparative evaluations.

**Key Variables**

The primary dataset incorporates several critical variables that are instrumental for the analysis:

* **Domain Code/Domain:** Specifies the data category, such as "QCL" (Crops and Livestock Products).
* **Area Code (M49)/Area:** Identifies countries using the M49 coding system alongside their corresponding names.
* **Element Code/Element:** Indicates the type of measurement or data recorded, such as "Stocks" for livestock.
* **Item Code (CPC)/Item:** Refers to specific agricultural products, including cattle, maize, and potatoes.
* **Year Code/Year:** Denotes the year of data collection.
* **Unit:** Describes the unit of measurement, e.g., "An" (Animal numbers) or metric tons.
* **Value:** Records the numerical value corresponding to the year, item, and element.
* **Flag/Flag Description:** Offers additional metadata, such as whether the data is official (“A”) or estimated (“E”).
* **Note:** Contains supplementary remarks or observations about the data entry.

**License Information:** FAOSTAT data is provided under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) license. This license permits:

* **Sharing**: copying and redistributing the material in any medium or format
* **Adaptation**: remixing, transforming, and building upon the material

Under the following terms:

* **Attribution**: Appropriate credit must be given to FAO as the data source
* **NonCommercial**: The data cannot be used for commercial purposes
* **ShareAlike**: If the data is remixed, transformed, or built upon, the modified material must be distributed under the same license

**Programming for DA Tasks**

**Programming**

The analysis utilized a diverse range of Python tools and libraries to ensure a robust evaluation of Ireland’s agricultural data. Core libraries such as Pandas and NumPy enabled efficient manipulation and aggregation of large datasets. Visualization tools, including Matplotlib and Seaborn, facilitated the creation of insightful graphs and charts, allowing for clear trend analysis and comparisons.

For predictive modeling and machine learning, advanced libraries like Scikit-learn, LightGBM, XGBoost, and pmdarima were employed. Scikit-learn provided a foundation for implementing both linear models and ensemble methods, such as random forests, while LightGBM and XGBoost demonstrated their strengths in handling structured data. The inclusion of pmdarima allowed for automated and efficient time-series forecasting, offering a significant advantage in evaluating temporal patterns within the agricultural data.

Key justifications for these choices include the scalability and versatility of the tools, as well as their alignment with statistical and machine learning requirements. The use of Jupyter Notebook facilitated interactive and iterative analysis, ensuring a structured and dynamic workflow. Code quality was upheld through practices such as consistent commenting, descriptive variable naming, and modular programming, enhancing readability and maintainability.

**Data from Diverse Sources**

The primary dataset was obtained from FAOSTAT’s extensive agricultural database, covering data from 1961 to 2023. This was complemented by an additional dataset containing metadata and unit definitions, which enhanced the interpretability of the main dataset. These sources were processed and integrated programmatically to ensure consistency and usability.

To merge the datasets, the Unit column from the main dataset was matched with Unit Name from the supplementary dataset. This ensured clarity in measurement units and facilitated accurate comparisons. The process involved identifying redundant columns and dropping them post-merge to streamline the data structure. For instance, columns like “Flag Description” and “Note” were removed due to their minimal contribution to the analysis. This integration ensured the data’s completeness and reliability for subsequent analysis.

**Data Manipulation**

The data manipulation phase involved critical aggregation methods to process and analyse the datasets effectively. Pivot tables were extensively used to reshape the data for exploratory and inferential analysis. For instance, livestock data was aggregated by year and livestock type to visualize trends in cattle, sheep, and chickens over time. Similarly, production data for crops was aggregated to enable cross-country comparisons.

Datatypes were corrected where necessary, such as converting year codes to string formats to facilitate merging and filtering. The cleaned and processed data provided a robust foundation for analysis, enabling accurate insights and comparisons.

**Optimization**

Optimization strategies were implemented to enhance computational efficiency and resource utilization. Vectorized operations in Pandas were prioritized over iterative loops to expedite data manipulation tasks. For example, filtering and aggregation tasks were optimized using group-by functions and conditional indexing.

Time-series forecasting tasks were optimized by selecting parsimonious ARIMA models through automated hyperparameter tuning, reducing execution time without compromising predictive accuracy.

Trade-offs were carefully considered, such as prioritizing execution speed during data cleaning at the expense of additional memory usage. Overall, the optimization strategies ensured that the analysis made efficient use of system resources, enabling scalability and responsiveness throughout the project.

**Descriptive Statistics and Visualisations**

The dataset was analyzed using descriptive statistics to summarize the central tendencies and variability in Ireland's agricultural data. The average values highlight significant activity in potato production, with a mean yield of 29,463.24 kilograms per hectare, while maize production was negligible. Livestock analysis showed a high average stock of cattle (6,160,883.78) and sheep (4,324,525.87). Median values confirmed these trends, with potatoes showing a median production of 618,000 tonnes and cattle stocks centred at 6,263,900. Variability was measured using range, variance, and standard deviation, revealing the highest fluctuations in sheep stocks and potato production. Additionally, skewness and kurtosis metrics were used to assess the distribution shapes. Visualizations, such as bar plots and line graphs, were used to clearly illustrate these findings, supporting model selection and justifications.

**Inferential Statistics and Variable Analysis**

Inferential statistics were applied to analyse Ireland’s agricultural data and compare it with selected European countries: Belgium, Germany, Portugal, Spain, and Poland. The analysis focused on cattle stocks and utilized the following statistical tests:

1. **Shapiro-Wilk Test for Normality**:
   * **Null Hypothesis (H0):** The data for Ireland and the comparison country are normally distributed.
   * **Alternative Hypothesis (H1):** The data for Ireland and/or the comparison country are not normally distributed.
   * **Findings:** Ireland’s cattle data and most comparison countries (Poland, Germany, and Spain) were not normally distributed (p-value < 0.05). However, Belgium (p = 0.227) and Portugal (p = 0.127) showed normal distributions, making non-parametric tests more appropriate.
2. **Levene’s Test for Homogeneity of Variances**:
   * **H0:** The variances of the cattle data for Ireland and the comparison country are equal.
   * **H1:** The variances of the cattle data for Ireland and the comparison country are not equal.
   * **Findings:** All comparisons indicated unequal variances (p < 0.05), with the strongest violations observed for Germany (p = 8.19e-23) and Poland (p = 4.87e-22).
3. **T-test for Mean Differences (Parametric)**:
   * **H0:** There is no significant difference in the mean cattle data between Ireland and the comparison country.
   * **H1:** There is a significant difference in the mean cattle data between Ireland and the comparison country.
   * **Findings:** While the t-test indicated significant differences, the violations of normality and homogeneity of variances reduced its reliability.
4. **Mann-Whitney U Test (Non-parametric)**:
   * **Findings:** This test showed significant differences between Ireland and all countries, with the strongest differences observed with Germany and Portugal (both p ≈ 3.66e-22) and relatively weaker differences with Spain (p ≈ 6.97e-07).
5. **Kruskal-Wallis Test (Non-parametric)**:
   * **Findings:** Confirmed the Mann-Whitney U results, highlighting Germany and Portugal as the most distinct from Ireland (statistic ≈ 93.76), while Spain showed smaller but still significant differences (statistic ≈ 24.65).
6. **Kolmogorov-Smirnov Test**:
   * **Findings:** Demonstrated the magnitude of distribution differences, with Germany and Portugal showing the greatest separation from Ireland (statistic = 1.0), while Spain exhibited partial overlap (statistic ≈ 0.40).
7. **Wilcoxon Signed-Rank Test**: While showing significant differences (p < 0.05) for all comparisons, this test is not suited for our analysis as it's designed for paired/dependent samples, which isn't appropriate for this cross-country comparison.

**Hypotheses and Conclusions:**

* Normality and homogeneity tests revealed that most data violated the assumptions for parametric tests, justifying the choice of non-parametric methods.
* Significant differences were consistently identified across all tests, emphasizing the distinctiveness of Germany’s and Portugal’s cattle farming systems compared to Ireland’s.
* Spain and Poland were moderately different, suggesting less pronounced variations in agricultural practices.

**Using Analysis Outcomes to Deepen Research and Address Challenges**

The outcomes of the inferential analysis provided valuable insights into the disparities between Ireland and other European countries. For instance, Germany’s and Portugal’s cattle systems exhibited highly industrialized structures, which could inform policy recommendations for Ireland. Challenges encountered included selecting appropriate statistical tests and interpreting results from non-parametric methods. These were overcome by researching statistical concepts through reliable resources like Wikipedia and applying systematic validation of test assumptions.

**Time Series Forecasting Analysis**

**Data Preparation and Decomposition**

The dataset utilized for this analysis focuses on cattle stocks in Ireland, organized by year, with the 'Value' column representing the number of cattle recorded annually. The data spans multiple decades, providing a robust historical context for trend identification and forecasting. Preprocessing steps involved filtering data exclusively for Ireland and focusing on the 'Cattle' category to ensure specificity and consistency in the analysis.

**Splitting the Dataset**

The dataset was split into training and test sets to evaluate model performance effectively.

* **Training Set**: Includes data from the earliest recorded year up to 2018, covering the majority of historical trends and patterns.
* **Test Set**: Comprises data from 2019 onward, reserved for validating the forecasting accuracy of the model.

The split ensures that the model is trained on past data while being tested on unseen data points to simulate real-world forecasting scenarios.

**Seasonal Decomposition**

To better understand the data, seasonal decomposition was performed, breaking the time series into trend, seasonal, and residual components. This process uncovered the following insights:

* **Trend Analysis**: The data exhibited a general upward trajectory over the years, suggesting long-term growth in cattle stocks. However, subtle fluctuations were observed, particularly around the late 2000s, where a slight levelling off was evident before resuming growth.
* **Seasonality Patterns**: Repeating cycles indicated annual variations likely tied to agricultural activities, such as breeding and market demand cycles. These patterns emphasize the importance of capturing seasonality for accurate forecasting.
* **Residuals and Noise**: Analysis of residuals highlighted randomness, with some deviations from expected patterns. While the residuals appeared mostly stable, some variations suggested the presence of anomalies or external influences that could be explored further.

**Model Selection and Implementation**

Given the dataset characteristics, the Auto ARIMA model was selected due to its proven ability to handle time series data and automatically determine optimal parameters (p, d, q). This feature minimizes manual tuning and ensures the best-fit model. Since initial decomposition indicated weak seasonal effects, the seasonal parameter 'm' was set to 1 to prioritize non-seasonal behavior.

**Key Model Evaluation Metrics:**

* **Ljung-Box Test for Autocorrelation**: Results indicated no significant autocorrelation in residuals, affirming the validity of model assumptions and residual independence.
* **Jarque-Bera Test for Normality**: Residuals displayed non-normality, hinting at potential outliers or skewed distributions. This finding suggests that further investigation into preprocessing or model refinements may be beneficial.
* **Heteroskedasticity Test**: Variance was found to be consistent across residuals, reducing concerns about heteroskedasticity and supporting model reliability.

**Model Performance and Results**

The Auto ARIMA model achieved an RMSE value of **170,105.86**, reflecting moderate forecasting accuracy. Given that the dataset values are measured in millions, this error margin is considered reasonable, especially for capturing broader trends rather than granular fluctuations.

The model’s forecast extended up to 2038, projecting a steady rise in cattle stocks from **6.63 million in 2024** to **7.20 million in 2038**. This forecast suggests sustained growth in the sector, aligning with historical patterns and reinforcing expectations of gradual expansion.

**Insights from Visualization**

1. **Historical Trends and Patterns**:
   * Analysis of historical data revealed consistent growth, punctuated by stabilization periods, particularly around economic or policy-driven events.
   * Cyclical variations emphasized the role of seasonal factors, underscoring the importance of incorporating such dynamics into predictive modeling.
2. **Forecast Visualization**:
   * The forecast plot depicted a smooth continuation of previous trends without abrupt deviations, suggesting reliability in predictions.
   * Confidence intervals around the forecasts were relatively narrow, enhancing trust in projected outcomes while acknowledging minor uncertainties.
   * Additional variables, such as economic indicators, or policy changes, could improve predictions and refine forecast intervals.

**Recommendations**

* **Model Enhancements**:
  + Introduce external regressors, such as livestock feed costs and agricultural policies to capture broader influences on cattle stocks.
  + Investigate non-linear models like LSTMs or neural networks, which may outperform linear models in detecting complex patterns.
* **Residual Diagnostics**:
  + Conduct further tests to address non-normality in residuals. Applying transformations (e.g., Box-Cox) or identifying and treating outliers could improve residual behavior and accuracy.
* **Comparative Modeling**:
  + Implement alternative forecasting models, such as Facebook’s Prophet and exponential smoothing methods, to cross-validate results and potentially enhance performance.
  + Perform ensemble modeling to combine strengths from multiple approaches for more robust predictions.
* **Expanded Data Analysis**:
  + Incorporate socioeconomic and climatic data to enrich the dataset, providing a multidimensional perspective for forecasting.
  + Analyze subcategories, such as age distribution or breed-specific data, to tailor predictions to specific segments of the cattle industry.

The Auto ARIMA model serves as a solid starting point for forecasting cattle stocks in Ireland, delivering consistent trends and reasonable error margins. While the results are promising, opportunities for improvement remain, particularly in integrating external variables, refining residuals, and exploring advanced machine learning techniques. Moving forward, deeper analyses and enhancements can provide a more nuanced understanding of factors influencing cattle stocks, aiding decision-making and policy planning in the agricultural sector.

**Forecasting with Machine Learning Models**

To address the challenge of forecasting cattle values, six supervised machine learning models were evaluated: Ridge, Elastic Net, K-Nearest Neighbors (KNN), Random Forest, LightGBM, and XGBoost. The selection of these models was informed by their suitability for handling diverse data characteristics, ensuring both flexibility and robustness in capturing trends and relationships. The models were tuned using **GridSearchCV**, ensuring optimal hyperparameter configurations for each method.

**Feature and Target Selection**

* **Feature (Year)**: The dataset presented a clear temporal structure, making "Year" a natural feature to forecast cattle values. Time serves as a critical determinant of trends in agricultural metrics, such as cattle value, often reflecting economic, environmental, and market influences.
* **Target (Cattle Value)**: The target variable was value (amount of cattle), a key metric for agricultural forecasting and decision-making. This aligns with the broader scenario of agricultural insights in Ireland and beyond.

**Key Observations**

1. **R² Scores**:
   * Random Forest achieved the highest R² score (0.931), indicating its superior ability to explain variance in the target variable compared to the other models.
   * XGBoost and KNN also performed well, with R² scores of 0.904 and 0.917, respectively.
   * Ridge and Elastic Net had moderate R² scores of approximately 0.70, while LightGBM demonstrated the lowest performance in terms of variance explanation (0.530).
2. **RMSE**:
   * Random Forest outperformed all other models with the lowest RMSE (175,543), suggesting high accuracy in its predictions.
   * XGBoost and KNN followed with RMSE values of 218,501 and 213,221, respectively, which are slightly higher but still indicative of robust predictive capabilities.
   * Ridge and Elastic Net exhibited similar RMSE values (~411,200), signifying moderate prediction error.
   * LightGBM, however, had the highest RMSE (538,189), reflecting significant error in its predictions.

**Model Insights and Justification**

* **Ridge and Elastic Net**: These linear models captured the trend but lacked the flexibility to model complex non-linear relationships in the dataset, which is evident from their moderate R² scores and higher RMSE.
* **KNN**: As a non-parametric model, KNN performed relatively well, benefiting from its ability to adapt to the local data structure. However, its performance is influenced by the choice of neighbors and data density.
* **Random Forest**: Its ensemble approach leveraging multiple decision trees made it particularly effective at capturing complex patterns and interactions in the dataset, resulting in the best overall performance.
* **XGBoost**: This gradient boosting algorithm closely followed Random Forest in performance, combining the strengths of boosting and regularization for robust predictions.
* **LightGBM**: Despite its theoretical advantages in speed and scalability, LightGBM underperformed on this dataset, possibly due to the small dataset size or insufficient tuning of hyperparameters.

**Sophisticated Use of GridSearchCV**

To ensure fairness and optimize model performance, GridSearchCV was employed to fine-tune critical hyperparameters:

* **Random Forest**: GridSearchCV was used to tune the number of estimators, maximum depth, and minimum samples for splits, resulting in a finely tuned ensemble model with high predictive power.
* **XGBoost**: Hyperparameter tuning focused on the learning rate, maximum depth, and subsample parameters, balancing model complexity and overfitting risks.
* **LightGBM**: While hyperparameters such as boosting type, learning rate, and max depth were tuned, the model struggled due to dataset-specific characteristics, highlighting the importance of domain-aligned tuning.
* **KNN**: The number of neighbors was optimized to balance underfitting and overfitting, improving local adaptability.

Random Forest emerged as the most reliable model for forecasting cattle values, balancing high R² scores and low RMSE. XGBoost also showed promise as a close competitor. On the other hand, LightGBM struggled to generalize effectively for this dataset, highlighting the importance of aligning model choice with data characteristics.

This evaluation emphasizes the importance of experimenting with diverse machine learning models and leveraging appropriate scoring metrics to derive actionable insights in forecasting tasks.

This research leverages high-quality datasets from **FAOSTAT**, the Food and Agriculture Organization’s (FAO) comprehensive statistical database, which serves as the backbone of the analysis.

**Overview of Primary Dataset**

The primary dataset, sourced from **FAOSTAT**, encompasses agricultural data spanning **1961 to 2023** for all countries, systematically identified using the **M49 coding system**. This dataset offers rich, longitudinal insights across key agricultural domains such as livestock statistics and crop production.

The primary focus for this research includes:

* **Livestock**: Cattle, chickens, and sheep.
* **Crop Production**: Maize and potatoes.

These focus areas provide a holistic view of pivotal agricultural products, facilitating a nuanced understanding of production trends, economic impacts, and regional variations.

**Supplemental Dataset for Uniformity**

To enhance clarity and ensure consistent interpretation, an additional dataset titled **"Unit Definitions"** was sourced from FAOSTAT’s **Definitions section**. This complementary dataset contains metadata such as unit names and descriptions, ensuring uniformity in the analysis.

For example:

* **Unit Descriptions**:
  + “ha/cap” (hectares per capita) provides critical context during comparative evaluations of land use.
  + “Head” (livestock count) enables precise comparisons across regions.

By integrating this supplemental metadata, the research ensures that all units and measures are clearly defined and interpretable, avoiding misalignment in cross-country or intertemporal comparisons.

**Positives of the Datasets**

1. **Global Relevance**: Covering data from **1961 to 2023**, the dataset captures long-term trends, seasonal patterns, and cross-country variations in agricultural output.
2. **Granular Coverage**: Detailed statistics on livestock and crop production, disaggregated by product type, provide deep insights into economic and environmental dynamics.
3. **Standardized Coding**: The use of the **M49 system** ensures compatibility with other international datasets, facilitating seamless integration and cross-referencing.
4. **Accessibility and Licensing**: FAOSTAT datasets are openly accessible under **FAO’s licensing terms**, which support non-commercial academic use, derivative works, and dissemination with proper attribution.

**Negatives of the Datasets**

1. **Data Gaps**: Some country-level records exhibit missing values or inconsistencies, particularly in earlier years. For example, there is not enough data for maize and chicken in some entries for Ireland.
2. **Granularity Trade-offs**: While global coverage is robust, subnational or localized data may be unavailable, limiting precision for regional analyses.
3. **Unit Interpretation**: Without the supplemental **Unit Definitions dataset**, interpretation of units like “ha/cap” or “head” might lead to confusion in cross-country comparisons.

**Licensing and Permissions**

The datasets are provided under **FAO’s Terms and Conditions**, which explicitly allow for:

* **Non-commercial usage**: Academic and research-oriented applications are supported.
* **Derivative works**: Users may transform and analyze data, provided proper attribution is maintained.
* **Open access**: The FAO promotes data dissemination to enhance global agricultural insights.

Strict adherence to these licensing terms ensures the ethical and compliant use of data throughout this report. Proper attribution to **FAOSTAT** is included in all outputs and publications.

The integration of FAOSTAT’s robust datasets, coupled with supplemental metadata ensures a comprehensive and insightful foundation for this research. While challenges such as data gaps and unit interpretation were encountered, these were systematically addressed through supplemental resources and methodical analysis. This approach exemplifies best practices in data acquisition, balancing rigor, relevance, and ethical compliance.

**References**

Bobbitt, Z. (2021) *Chi-Square Test of Independence: Definition, formula, and example*. <https://www.statology.org/chi-square-test-of-independence/>

Kosaka, M. (2022) 'Efficient Time-Series analysis using Python’s PMDArima Library,' *Medium*, 12 October. <https://towardsdatascience.com/efficient-time-series-using-pythons-pmdarima-library-f6825407b7f0>

Libretexts (2024) *8.3: the Independent Samples T-Test formula*. <https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Statistics%3A_Open_for_Everyone_(Peter)/08%3A_Independent_Samples_t-Tests/8.03%3A_The_Independent_Samples_t-Test_Formula>

Peterka, P. (2024) *Levene’s Test: Your Guide to Understanding, Performing, and Interpreting Homogeneity of Variance*. <https://www.6sigma.us/six-sigma-in-focus/levenes-test/>.

*SciPy* (no date). <https://scipy.org/>.

Wikipedia contributors (2024a) *Kruskal–Wallis test*. <https://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_test>

Wikipedia contributors (2024b) *Mann–Whitney U test*. <https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test>

Wikipedia contributors (2024c) *Shapiro–Wilk test*. <https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test>

Wikipedia contributors (2024d) *Wilcoxon signed-rank test*. <https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test>.

**Appendix**

The Python code below demonstrates the core methodologies employed for the analysis, focusing on data manipulation, exploratory data analysis, statistical analysis, and machine learning tasks. Only the most critical functions and workflows are included for brevity.

**Importing Libraries**

The necessary libraries for data processing, visualization, and machine learning were imported.

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import (

skew, kurtosis, levene, ttest\_ind, mannwhitneyu, wilcoxon, ks\_2samp, kruskal, shapiro

)

from sklearn.model\_selection import train\_test\_split, cross\_val\_score, KFold

from sklearn.linear\_model import LinearRegression, Ridge, ElasticNet

from sklearn.ensemble import RandomForestRegressor

from sklearn.neighbors import KNeighborsRegressor

import lightgbm as lgb

import xgboost as xgb

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.tsa.seasonal import seasonal\_decompose

from statsmodels.graphics.tsaplots import plot\_acf, plot\_pacf

import pmdarima as pm

**Data Loading and Preliminary Exploration**

FAOSTAT datasets were loaded and explored for initial insights.

df = pd.read\_csv('FAOSTAT\_data\_en\_12-24-2024.csv')

unit\_df = pd.read\_csv('FAOSTAT\_data\_units\_12-24-2024.csv')

# Merging datasets

merged\_df = df.merge(unit\_df, left\_on='Unit', right\_on='Unit Name', how='left')

# Cleaning and preprocessing

merged\_df.drop(columns=['Domain Code', 'Area Code (M49)', 'Item Code (CPC)', 'Year Code', 'Unit Name'], inplace=True)

cleaned\_df = merged\_df.dropna()

**Exploratory Data Analysis**

Trends in agricultural production and livestock stocks were visualized.

**Example: Livestock Heatmap**

livestock\_items = ['Cattle', 'Sheep', 'Chickens']

livestock\_data = cleaned\_df[

(cleaned\_df['Item'].isin(livestock\_items)) & (cleaned\_df['Element'] == 'Stocks') & (cleaned\_df['Area'] == 'Ireland')

]

heatmap\_data = livestock\_data.pivot\_table(index='Year', columns='Item', values='Value', aggfunc='sum')

sns.heatmap(

heatmap\_data.tail(20), annot=True, fmt=".0f", cmap='YlOrRd', cbar\_kws={'label': 'Stock Count (in thousands)'}

)

plt.title('Livestock in Ireland Over the Past 20 Years')

plt.show()

**Statistical Analysis**

Inferential statistical tests were applied to compare Ireland with other countries.

**Example: Mann-Whitney U Test**

stat, p\_value = mannwhitneyu(ireland\_data, comparison\_data)

print(f"Mann-Whitney U Test -> Statistic: {stat}, p-value: {p\_value}")

**5. Time Series Forecasting**

A time series analysis was performed to forecast future cattle stocks using the Auto ARIMA model.

model\_auto\_arima = pm.auto\_arima(

train['Value'], m=1, seasonal=False

)

forecast\_auto\_arima = model\_auto\_arima.predict(n\_periods=len(test))

future\_forecast = model\_auto\_arima.predict(n\_periods=15)

**Visualization**

plt.plot(data\_ireland.index, data\_ireland['Value'], label='Historical Data')

plt.plot(future\_forecast.index, future\_forecast.values, label='Forecast', color='orange')

plt.legend()

plt.show()

**6. Supervised Machine Learning**

Multiple machine learning models were tested for predictive analysis.

models = {

"Random Forest": RandomForestRegressor(random\_state=42),

"XGBoost": xgb.XGBRegressor(random\_state=42),

"LightGBM": lgb.LGBMRegressor(random\_state=42)

}

for model\_name, model in models.items():

model.fit(X\_train, y\_train)

predictions = model.predict(X\_test)

print(f"{model\_name}: RMSE = {np.sqrt(mean\_squared\_error(y\_test, predictions))}")