Lorraine Drumm

[Company name]  [Company address]

An investigation into the irish agriculture sector

**Contents**

**1 Background and Introduction 3**

1.1 Abstract…………………………………………...........3

1.2 Background………………………………………........3

1.3 Data Selection………………………………………....3

**2 Data Preparation and Visualization 4**

2.1 Data Cleaning and preparation……………………….4

2.2 Preparing the Data for Machine Learning………..….5

2.3 Data Visualization…………………………………..….6

2.4 Tufte’s Principles………………………………….........7

**3 Statistics 7**

3.1 Descriptive Analysis……………………………..….....7

3.2 Discrete Distributions………………………………….11

3.3 Normal Distributions…………………………………..14

**4 Machine Learning 15**

4.1 Project Management Frameworks…………………..15

4.2 My Machine Learning Models………………..………16

4.2.1 Predictive machine modelling…………….………..16

4.2.1 Classification machine modelling……….…………18

**5 Programming 21**

5.1 Programming Paradigms…………………………..….21

**6 Conclusion 22**

**7 References 23**

*List of Figures*

1. Line plot to fill missing values
2. Heatmap of Agricultural Value over time
3. Line plot of agricultural value
4. Line plot after inflation error corrected
5. Line plot of total agriculture value
6. Subplot of agricultural value per Irish Region
7. Boxplot of agricultural value before outliers removed
8. Boxplot of agricultural value after outliers removed
9. Mean Vs Median by Agricultural Sector
10. Q-Q Plot for distribution
11. Shapefile map showing distribution of Agricultural Value in Ireland
12. Heatmap of Agricultural Value by Irish Region
13. Error bars for Distribution of Irish Agriculture Value
14. Bar chart of Livestock Population by EU Member State
15. Pie chart of EU Livestock Population
16. Distribution of Livestock in Netherlands Visual
17. Distribution of Livestock in Ireland Visual
18. Crop Production Ireland vs Netherlands
19. Linear regression for total agricultural value 2024
20. Decision Tree regression for total agricultural value 2024
21. Random Forest regression for total agricultural value 2024
22. Confusion matrix for Classification model
23. Scattered boxplot showing classification results
24. Sentiment category count
25. Sentiment distribution producer
26. Sentiment distribution consumer
27. Sentiment count producer
28. Sentiment count consumer

*List of Tables*

1. *Percentage increase 2008 to 2009*
2. Mean and median for each category
3. *Normal distribution results*
4. *Total aggregated agriculture value by region*
5. *Confidence interval results*
6. *Confidence interval results* by crop
7. *Ireland vs Netherlands*
8. *Normal distribution results*
9. *Mann – Whitney U test*
10. ANOVA
11. Wilcoxon test results
12. *No. farms Ireland vs Netherlands*
13. *ANOVA*
14. *T test results*
15. *Wilcoxon Signed – Rank Test*
16. *Mann-Whitney U Test*
17. *Regression results*
18. Classification model results
19. Classification model accuracy and F1 scores
20. *Sentiment analysis model 1*
21. *Sentiment analysis model 2*

**An Investigation into the Irish Agriculture Sector and how it Compares to Other EU Countries with a Specific Focus on Livestock and Crops**

Lorraine Drumm

1. **Background and Introduction**
   1. *Abstract*

*This report analysis the Irish agriculture sector with a particular focus on the value of livestock and crop production. The aim of this study was to understand the historical trends, the geographical distribution of and the statistical properties associated with farming in Ireland. This was then compared to other EU countries with a particular focus on the Netherlands. The study found that the value of agricultural to the irish economy has been steadily increasing and that there is a correlation between the value of agriculture in a specific region and the number of farms in that region.*

*1.2 Introduction*

As part of my analysis I wanted to answer 6 questions:

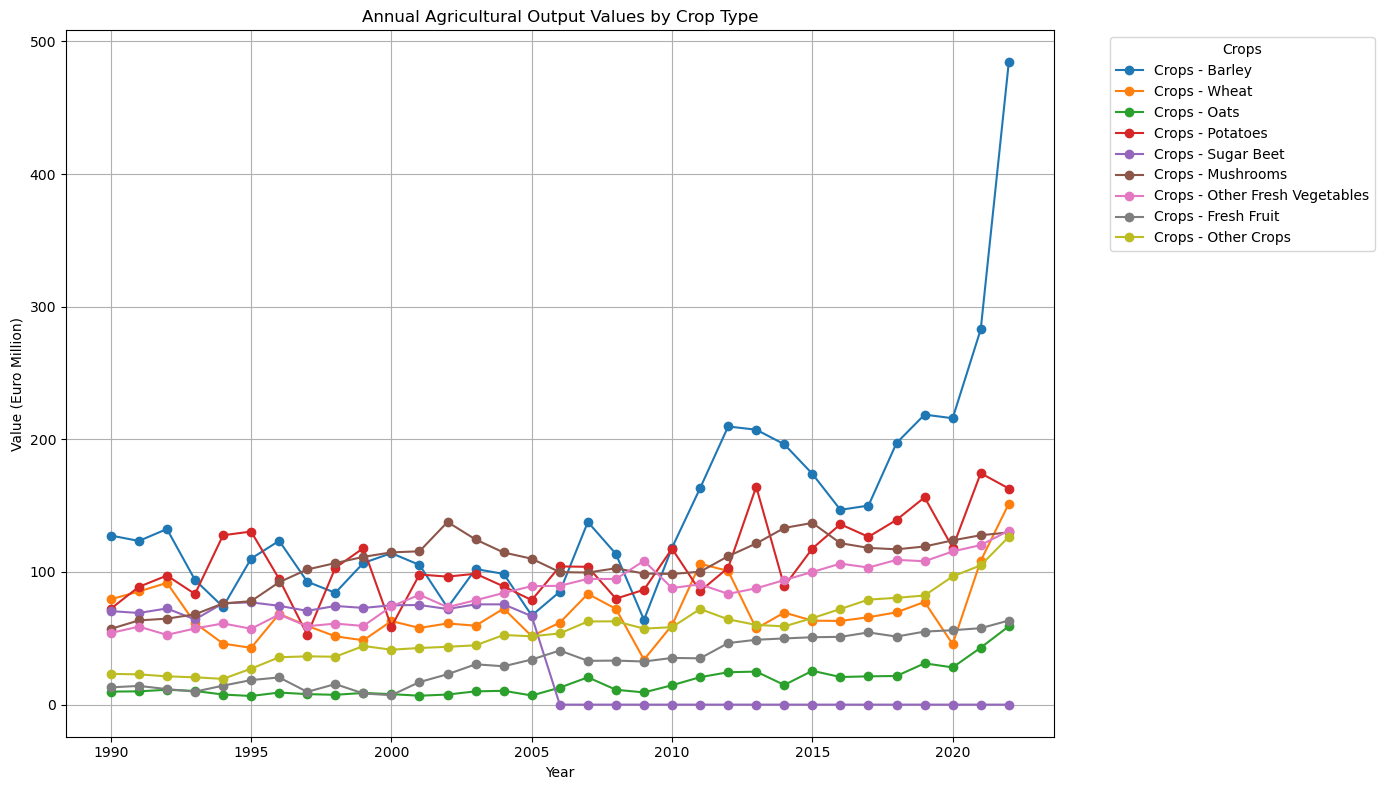
* What are the most valuable sectors of Irish Agriculture?
* How has farming changed over the years in Ireland?
* How does Ireland’s biggest Agriculture sector (Livestock) compare to other EU countries?
* What impact did the 2009 economic recession have on Irish agriculture?
* What is the predicted total value of Irish agriculture in 2024?
* Is there a correlation between regional districts and the type of agriculture in that region?

**2 Data Preparation and Visualisation:**

2.1 Data Acquisition

Datasets were sourced from: the CSO’s Open Data Platform for Irish agricultural data, Eurostat for EU member states, and CBS for the Netherlands. Issues included aligning time periods across datasets, leading to the exclusion of certain years, and switching focus from France to the Netherlands due to data availability and language barriers. CSV imports from CBS initially caused indentation errors, resolved by using a delimiter. Eurostat data had letter flags converting values to NaN, manual imputation addressed this. All csv datasets used are licensed under the “Creative Commons Attribution 4.0 International (CC BY 4.0) license. This allows reuse of data as long as source is stated and changes made are flagged (Creative Commons, 2016).

2.2 Exploratory Data Analysis

EDA was conducted. Basic commands were run to get a surface level understanding of the data such as column names, time range, df size and the number of unique variables in columns. Next the Pandas.info() command revealed there was 136 null values in “VALUE” column. Before dealing with the missing values the dataframe was filtered to only include Livestock and crop outputs to clarify the focus of the analysis. The remaining 18 null values were identified and their respective categories were visualised to help understand the best way to deal with these missing values.

*Figure 1: Lineplot to fill missing values*

The relationship between variables was explored using heatmaps and tables which highlighted max/min values (McQuaid, 2024b).

A screen shot of a graph

Description automatically generated

*Figure 2: Heatmap of Agricultural Value over time*

This gave a clear overview of the most valuable sectors. The top sectors were filtered and observed using a lineplot to help understand how “VALUE” changed over time.

A graph of different colored lines

Description automatically generated  *Figure 3: Line plot of agricultural value*

Visualizing the data showed that inflation had not been considered in the dataset. CPI values obtained from the CSO’s website were used to account for this and ensure they were comparable and representative. This was essential in preparing the data for regression machine learning models but also in being able to interpret the obtained data. After the data was plotted again and it appeared to be better distributed.

A graph of different colored lines

Description automatically generated

*Figure 4: Line plot after inflation error corrected*

A graph with blue lines and numbers

Description automatically generatedThe data was displayed to show trends in “Total Agricultural Value” over time:

*Figure 5: Line plot of total agricultural value*

This showed a large decline in 2009 which could be explained by the economic recession. This initial investigation indicates it would be beneficial to check for outlier in the data.

The regional distribution of this agricultural value was visualised using a subplot and a geoploth to give a general overview of spreadThe subplot highlighted that data was being duplicated by being included in multiple categories. For example, “State” was just a sum of all categories. After research was conducted all categories were removed that didn’t classify as NUTS3 regions.

A group of colorful bars

Description automatically generated with medium confidence

*Figure 2: Subplot of agricultural value per Irish Region*

These regions are the basic building blocks which were fed into some larger regions in the dataset (Central Statistics Office, 2016). This filtered data was displayed using a shapefile below. Issues encountered included mismatched region names for merging but this was fixed using .replace() command.

EU Livestock data was combined and displayed using a choropleth to directly compare change over time by country. This initial exploration paired with research allowed for an informed decision to select the Netherlands to compare more closely with Ireland.

Dataframes were melted and pivoted throughout my programme to allow for

dataframes to be merged. This can be seen when working with Ireland and the Netherlands crop yields in the jupyter notebook: “How Ireland Compares to Other Countries”.

Columns and variables within columns were renamed to allow for merging and concatenation of datasets. This was carried out when using the Irish shapefile and when creating a combined dataframe to compare Ireland and the Netherlands.

2.3 Data Preparation:

The null values in agri\_df that were normally distributed were replaced with mean values for that category (McQuaid, 2024). The line plot identified that Sugar Beet had a value of 0.0 from 2006 to current. After conducting research, it was confirmed that production had ceased in Ireland so this NaN value was replaced with 0.0 (agriculture.ec.europa.eu, 2022). The graph also showed a spike in Barley value in 2022. To calculate the value of barley in 2023 a simple calculation was done using accredited sources (CSO statistical publication, 2024). The result indicated the value was still unusually high and as a result a mean imputation would not have been sufficient.

The dataset from EUROSTAT’s Open Data Platform contained letter flags which caused some values to convert to NaN values while trying to convert from a string to a float. This was identified when trying to plot data. Initially code was ran to remove these special characters then convert to a integer however this didn’t work so the values were manually imputed. The aggregated livestock count over the years was visualised and the breakdown by country to directly compare Ireland to its EU competitors as can be seen in Figure.

Carrying out descriptive statistics on the data allowed for the identification of extreme outliers in 2022 and this year was removed. This improved my ML models(Inuwa, 2022).

For the classification model EU states were classified into 4 categories. They were encoded as the integers 0,1 and 2 so that they could be processed by the model. LabelEncoder() is used and fit\_transform.

Data was split into test and train groups. Theratio was altered for individual models to obtain the best test mean squared error result. Optimum parameters were selected using GridSearchCV and RandomisedSearchCV.

“Adjusted Value” was scaled to improve the ML result.

1. **Statistics**

3.1 Descriptive Statistics

Descriptive Analysis was used to summarize the datasets. Areas of focus included identify annual trends for different agricultural sectors in Ireland and investigating which regions contributed most to agricultural value.

It is evident in figure 2 and 3 there is a large variation in the value of agricultural sectors in Ireland. The most valuable sectors are: Dairy, cattle livestock, crops, forage plants and pig livestock. Boxplots suggested that data was positively skewed and after calculating the IQR for the aggregated data it was found that 2022 was an extreme positive outlier.

A screenshot of a graph

Description automatically generated

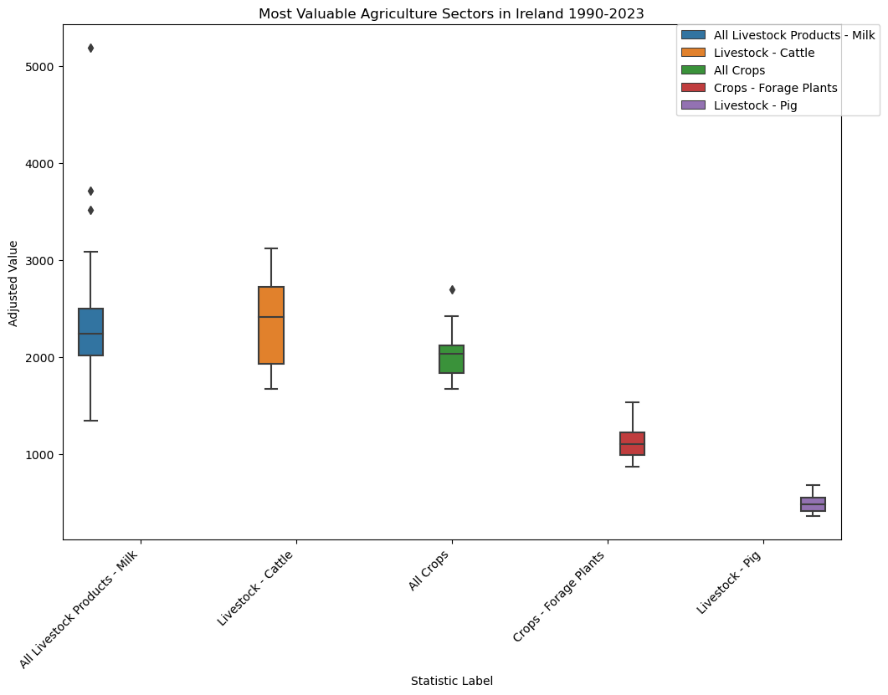
*Figure 7: Boxplot of agricultural value after outliers removed*

|  |  |
| --- | --- |
| **Statistic Label** | % **Increase** |
| *Crops - Wheat* | 205.60 |
| *All Cereals* | 120.62 |
| *Crops - Barley* | 106.21 |
| *Crops - Oats* | 93.54 |
| *All Livestock Products - Milk* | 67.85 |
| *All Livestock Products* | 66.41 |
| *Livestock - Horses* | 29.07 |
| *Crops - Potatoes* | 26.63 |
| *All Crops* | 24.20 |
| *Livestock - Cattle* | 21.52 |
| *Crops - Other Crops* | 20.32 |
| *All Livestock Products Other Product)* | 16.04 |
| *All Livestock* | 15.83 |
| *Livestock - Sheep* | 14.40 |
| *Crops - Forage Plants* | 5.76 |
| *Crops - Other Fresh Vegetables* | 4.41 |
| *Crops - Fresh Fruit* | 3.92 |
| *Livestock - Poultry* | 0.43 |
| *Crops - Mushrooms* | -3.59 |
| *Livestock - Pig* | -5.24 |

**>100% increase**

**>10% increase**

*Table 1: Percentage increase 2008 to 2009*



*Figure 8: Boxplot of agricultural value after outliers removed*

The mean and median of the value of each agricultural sector over the years (after

removing outliers) is as follows:

|  |  |  |
| --- | --- | --- |
| Statistic Label | Mean Value | Median Value |
| All Cereals | 291.22 | 275.04 |
| All Crops | 2002.68 | 2028.82 |
| All Livestock | 3546.08 | 3601.65 |
| All Livestock Products | 2404.41 | 2281.29 |
| All Livestock Products - Milk | 2335.42 | 2224.06 |
| All Livestock Products Other Products (excluding Milk) | 68.98 | 66.08 |
| Crops - Barley | 185.49 | 171.21 |
| Crops - Forage Plants | 1115.23 | 1087.45 |
| Crops - Fresh Fruit | 39.59 | 39.81 |
| Crops - Mushrooms | 139.65 | 136.15 |
| Crops - Oats | 19.25 | 17.27 |
| Crops - Other Crops | 67.70 | 68.23 |
| Crops - Other Fresh Vegetables | 109.38 | 109.57 |
| Crops - Potatoes | 141.62 | 137.66 |
| Crops - Sugar Beet | 55.30 | 0.00 |
| Crops - Wheat | 90.66 | 82.41 |
| Livestock - Cattle | 2333.27 | 2399.19 |
| Livestock - Horses | 235.61 | 244.28 |
| Livestock - Pig | 484.96 | 466.51 |
| Livestock - Poultry | 189.95 | 192.11 |
| Livestock - Sheep | 302.33 | 298.02 |
| Total aggregated agriculture | 16449.62 | 16664.68 |

*Table 2: Mean and median by sector*

As can be seen in the above table after removing outliers there is not a distinct difference between mean and median for each category and a positive or negative skew is not suggested. When we aggregate all categories the median is slightly bigger however only by 1.3% which suggests symmetric distribution over time.

Figure A graph of blue and orange bars

Description automatically generated9: Mean Vs Median by Agricultural Sector

A shapio-wilk test confirmed the normal distribution on the categories and the visualised results are shown below. A graph with blue dots

Description automatically generated

*Figure 10: Q-Q Plot for distribution*

|  |  |
| --- | --- |
| Shapiro-wilk Statistic | 0.964 |
| p-value | 0.336 |
| Result | Normally distributed |

*Table 3: Normal distribution results*

Focusing on regions showed that there is a correlation between geographical location and agricultural value in Ireland. Results are as follows:

|  |  |
| --- | --- |
| **Region** | **Adjusted Value** |
| Border | 109899.644565 |
| Mid-East | 110584.675976 |
| Mid-West | 142403.819768 |
| Midlands | 80213.704686 |
| South-East | 132996.439978 |
| South-West | 181178.655031 |
| West | 87917.343066 |

*Table 4: Total aggregated agriculture value by region*

Southerly regions have higher agricultural value historically.

A map of ireland with different colored areas

Description automatically generated

*Figure 11: Shapefile map showing distribution of Agricultural Value* *in Ireland*

*Figure 12:* A chart with different colors

Description automatically generated*Heatmap of Agricultural Value by Irish Region*

Valuable regions are higher in Southern regions. Less valuable sectors such as sheep is the most popular in the West.

3.2 Inferential Statistics (1 Population)

Confidence interval for the population proportion of crop value out of all agricultural value:

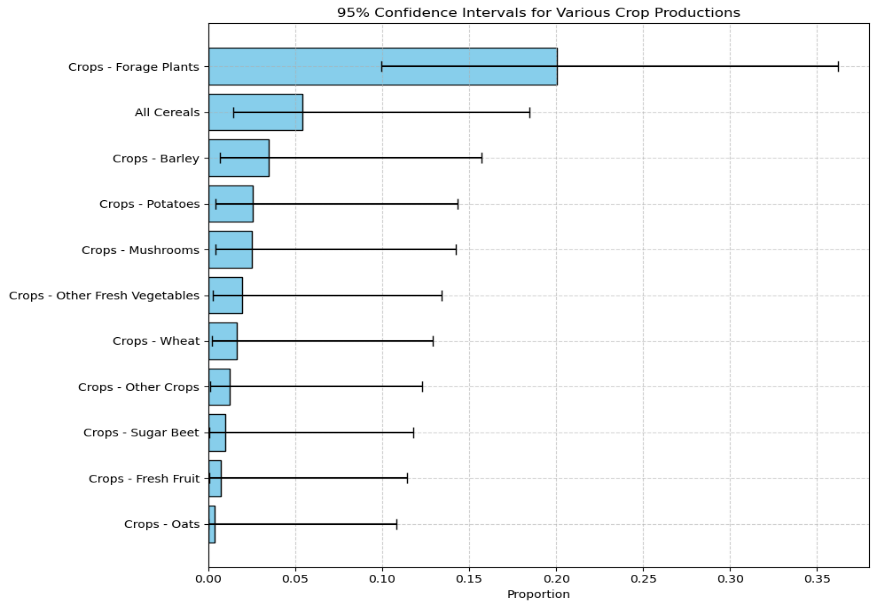
|  |  |
| --- | --- |
| Upper Bound | Lower Bound |
| 0.1996 | 0.5225 |
| 19.96% | 52.25% |

*Table 5: Confidence interval results*

There is a 95% chance that the probability of agricultural value originating from the crop sector is between 1.26%-23.33%. A wide confidence interval indicates high variability or small sample sizes, leading to less precise estimates of the population parameter (Cumming & Finch, 2005). The sector appears unpredictable and can be easily affected by external factors. Breaking this down further to look at specific crops:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic Label | Proportion | Lower Bound | Upper Bound | Interval Width |
| Crops - Oats | 0.0012 | 0.0000 | 0.1037 | 0.1037 |
| Crops - Fresh Fruit | 0.0025 | 0.0001 | 0.1059 | 0.1058 |
| Crops - Sugar Beet | 0.0033 | 0.0001 | 0.1073 | 0.1072 |
| Crops - Other Crops | 0.0042 | 0.0001 | 0.1090 | 0.1088 |
| Crops - Wheat | 0.0056 | 0.0003 | 0.1114 | 0.1111 |
| Crops - Other Fresh Vegetables | 0.0067 | 0.0004 | 0.1132 | 0.1128 |
| Crops - Mushrooms | 0.0085 | 0.0006 | 0.1162 | 0.1156 |
| Crops - Potatoes | 0.0087 | 0.0006 | 0.1165 | 0.1159 |
| Crops - Barley | 0.0118 | 0.0010 | 0.1218 | 0.1207 |
| All Cereals | 0.0185 | 0.0023 | 0.1324 | 0.1301 |
| Crops - Forage Plants | 0.0682 | 0.0205 | 0.2036 | 0.1830 |

*Table 6: Confidence interval results by crop*

All results indicate high level of uncertainty among all crop sectors in Ireland suggesting it is an unstable sector for farmers. This is evident from the larger error bars in the bar chart.

*Figure 13: Error bars for Distribution of Irish Agriculture Value*

3.3 Inferential Statistics (2 populations)

Ireland was compared to EU member state as they all benefit from the Common Agricultural Policy. Ireland’s biggest agricultural sector: Livestock population was compared. The results are as follows:

A graph of value in euro millions by country

Description automatically generated

*Figure 14: Barchart of Livestock Population by EU* *Member State*

The piechart below clearly highlights countries which could be considered for a deeper analysis:

A pie chart with different colored circles

Description automatically generated

*Figure 15: Pie chart of EU Livestock Population*

It was decided to focus on the Netherlands as after conducting additional research it was established both countries have several similarities (Ministerie van Landbouw, 2022):

|  |  |  |
| --- | --- | --- |
| Factor | Netherlands | Ireland |
| Climate and Geography | Maritime climate, fertile soils | Maritime climate, fertile soils |
| Agricultural Intensity | Highly developed, innovative practices | Highly developed, strong dairy and beef |
| Economic Importance | Significant role in the economy, top exporter | Crucial part of the rural economy, key export market |
| Land Use and Farm Structure | Small farm sizes, high productivity | Small farm sizes, high productivity |

*Table 7: Ireland vs Netherlands*

Irish livestock population is increasing while count in the Netherlands decreases. Livestock has grown by since 2010 in Ireland while it has changed by in the Netherlands in that same time period.

The 2 countries livestock populations were checked for normality.

A comparison of graphs and charts

Description automatically generated with medium confidenceA graph with numbers and lines

Description automatically generated

*Figure 16: Distribution of Livestock in Netherlands Visual*

A graph and a chart

Description automatically generated

*Figure 17: Distribution of Livestock in Ireland Visual*

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Statistic | p-value | Normality |
| Ireland | 0.848755 | 0.021363 | Not Normally Distributed |
| Netherlands | 0.88492 | 0.06833 | Normally Distributed |

*Table 8: Normal distribution results*

Ireland is not normally distributed which means non-parametric tests must be used to compare livestock population.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Hypothesis | Statistic | p-value | Result |
| Mann-Whitney U test | H0: The distributions of animal counts in Ireland and the Netherlands are the same. | U-statistic: 196.0 | p-value: 7.4678435688707155e-06 | Reject the null hypothesis (H0) |
|  | H1: The distributions of animal counts in Ireland and the Netherlands are different. |  |  |  |

*Table 9: Mann – Whitney U test*

As the the p value obtained is significantly less than 0.05 the null hypothesis is rejected. There is clear evidence that the distribution of livestock count is significantly different between both countries.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Test | Null Hypothesis (H0) | Alternative Hypothesis (Ha) | F-statistic | p-value | Decision |
| ANOVA | The means of the values for Ireland and NLD are equal. | At least one group mean is different. | 760.100655 | 8.9225795e-21 | Reject H0: There is a significant difference in means between Ireland and Netherlands. |

*Table 10: ANOVA*

The mean values of livestock count among both countries is significantly different as p < 0.05. The F-statistic is extremely high, indicating a substantial difference in the means of the groups relative to the variability within the groups. This is further supported by the results of my Wilcoxon test below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test | Hypothesis | Statistic | p-value | Result |
| Wilcoxon Test | H0: The distributions of animal counts in Ireland and the Netherlands are the same. | U -statistic:  0.0 | p-value:  0.0001220703125 | Reject the null hypothesis (H0) |
|  | H1: The distributions of animal counts in Ireland and the Netherlands are different. |  |  |  |

*Table 11: Wilcoxon test*

It is clear from the results Ireland and the Netherlands are not comparable in Livestock population despite only a 3% difference in their contribution to the EU population.

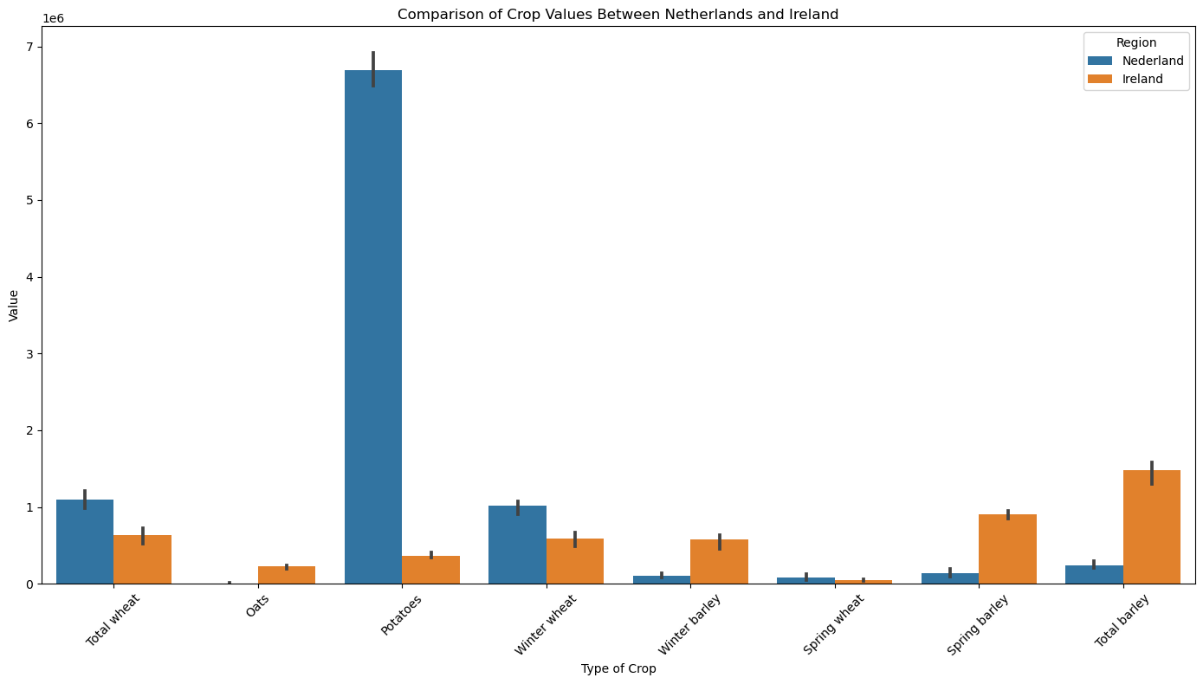
It was decided to focus on crop production and no. of farms in each country.

|  |  |  |  |
| --- | --- | --- | --- |
| Country | Year | Number of Farms | % Change |
| Netherlands | 2013 | 67,481 | - |
| Netherlands | 2016 | 55,688 | -17.47 |
| Ireland | 2013 | 139,600 | - |
| Ireland | 2016 | 137,500 | -1.50 |

*Table 12: No. farms Ireland vs Netherlands*

Irish farm count is significantly higher and declining at a slower rate. How does this impact crop yield?

A datatset was constructed. This dataset directly compares 8 crop categories in both countries under 3 areas: Area under cultivation (ha), Gross yield per ha (1000 kg), Gross yield, total (1000 kg) from 2021 to 2023. Total gross yield over the entire period can be seen in Figure:



*Figure 18: Crop Production Ireland vs Netherlands*

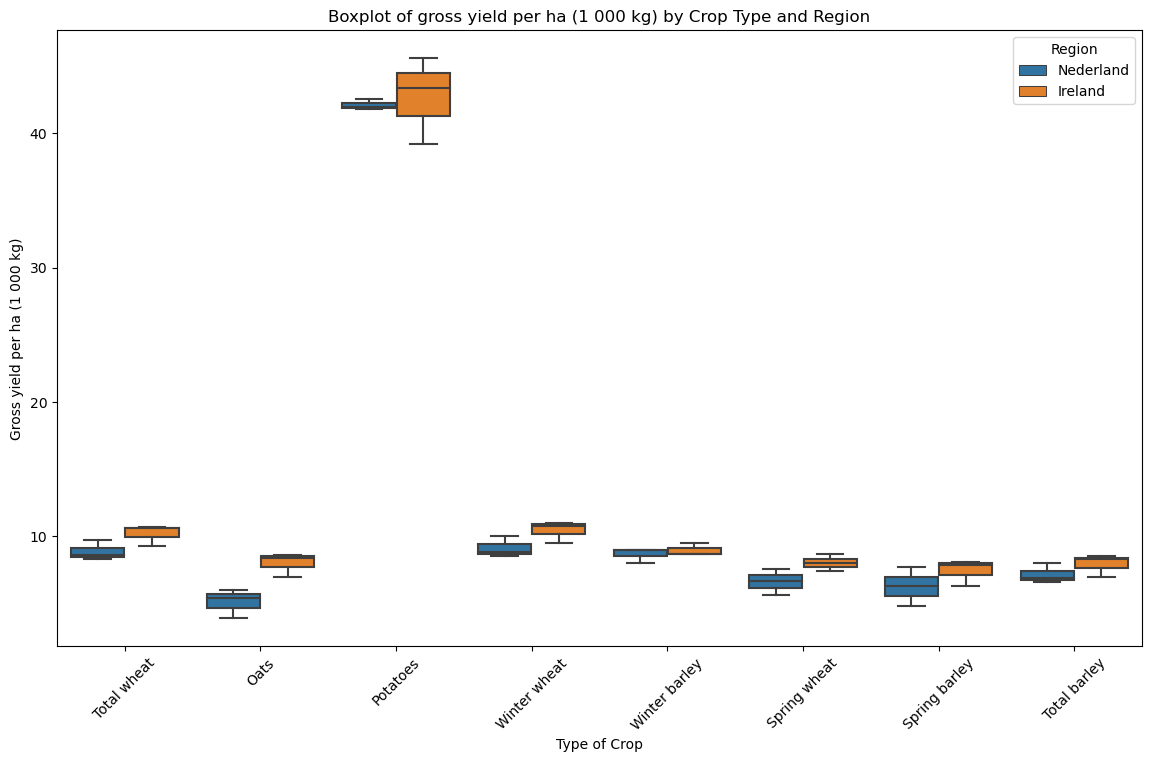
The Netherlands produces more potatoes and wheat while Ireland outperforms in the barley sector. As both countries have different sector sizes and varying numbers of farms it was decided to compare both countries using gross yield per hectare. The results are as follows:

ANOVA Test:

An ANOVA test revealed there is no significant difference in means of Gross yield per ha (1 000 kg) between Ireland and the Netherlands as we obtained a value of 0.7189. Also, The F-statistic value indicates that the variation between the group means is small relative to the variation between the countries.

|  |  |
| --- | --- |
| Hypothesis | ANOVA Results |
| H0: The means of Gross yield per ha (1 000 kg) for Ireland and the Netherlands are equal. | F-statistic = 0.1311832429888559, p-value = 0.718867048487364 |
| Ha: At least one group mean is different. | Fail to reject H0: There is no significant difference in means of Gross yield per ha (1 000 kg) between Ireland and the Netherlands. |

*Table 13: ANOVA*



*Figure 19: ANOVA test results*

Figure 19 above further highlights that gross yield per hectare is relatively similar among the same crops in both countries.

The Mann-Whitney U Test found that the distribution of gross yield per hectare are different between the Netherlands and Ireland. However the p value is only 0.0012 away from my 0.05 marker which suggests that with more data the result may have varied.

|  |  |
| --- | --- |
| Mann-Whitney U Test Results |  |
| Null Hypothesis (H0) | The distributions of gross yield per hectare are the same across the Netherlands and Ireland. |
| Alternative Hypothesis (H1) | The distributions of gross yield per hectare are different between the Netherlands and Ireland. |
| Result | Reject H0 |
| p-value | 0.0488 |

*Table 16: Mann-Whitney U Test*

Conclusion:

Ireland and the Netherlands are 2 countries with highly productive agricultural sectors. Ireland leads the way in livestock production while the Netherlands has a high crop value especially in potato production. They both use the land productively and an ANOVA test revealed that gross yield per hectare is relatively similar in both countries. The number of farms in both countries is declining however but at a more rapid rate in the Netherlands.

Some tests that I ran in mu jupyter notebook contradicted other data, as a result it was decided to not include them in the report but the code can still be viewed.

*4.1 Machine Learning Models:*

Machine Learning models were used to evaluate: What is the expected value of Irish agriculture in 2024 and can the EU member states be categorised based off their livestock population?

*4.1.1 Predictive Machine Modelling*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Predicted Value | Test RMSE | R squared value | Best score (MSE) |
| Linear Regression | 17713.63 |  |  |  |
| Decision Tree Regression | 21252.47 (Euro million) | 1238.51 | 0.82 | 15339606.46 |
| Random Forest Regression | 20397.34  (Euro million) | 1417.89 | 0.76 | 2010415.96 |

*Table 17: Regression results*

The decision tree regression model has the lowest RMSE and the highest R-squared value this indicates it is more accurate than the random forest model. Both MSE scores appear very high however agricultural value is measured in millions and this score is proportionally large. As MSE squares the errors larger errors are more heavily impacted by this (Stephen Allwright, 2022). MSE score for Decision tree is better performing.

In conclusion, based on the above, the decision tree regression is the most accurate for predicting the value of agricultural value in 2024, with the lowest RMSE, highest R-squared value, and lowest MSE (Iqbal, 2024).. However, it is a more simple model which means it could be overfitted (Hegelich, 2016).

The linear regression does not perform as well as it’s output is significantly lower.

A graph with blue dots and a red line

Description automatically generated

*Figure 19: Linear regression for total agricultural value 2024*

*A chart with blue dots and green line

Description automatically generated*

*Figure 20: Decision Tree regression for total agricultural value 2024*

A graph with blue lines and numbers

Description automatically generated

*Figure 21: Random Forest regression for total agricultural value 202*

*4.1.2 Classification Machine Modelling*

A random forest classification model was used to classify the EU countries into 4 categories based on their livestock population. This was carried out by creating 4 quantiles: very low, low, medium and high. The results are as follows:

|  |
| --- |
|  |
| Very Low | **Low** | **Medium** | **High** |
| Cyprus | Bulgaria | Austria | Germany |
| Estonia | Czech Republic | Belgium | Spain |
| Croatia | Finland | Switzerland | France |
| Iceland | Greece | Denmark | Ireland |
| Luxembourg | Hungary | Netherlands | Italy |
| Latvia | Lithuania | Portugal | Poland |
| Malta | Slovenia | Romania |  |
| Slovakia |  | Sweden |  |

*Table 18: Classification Model Result*

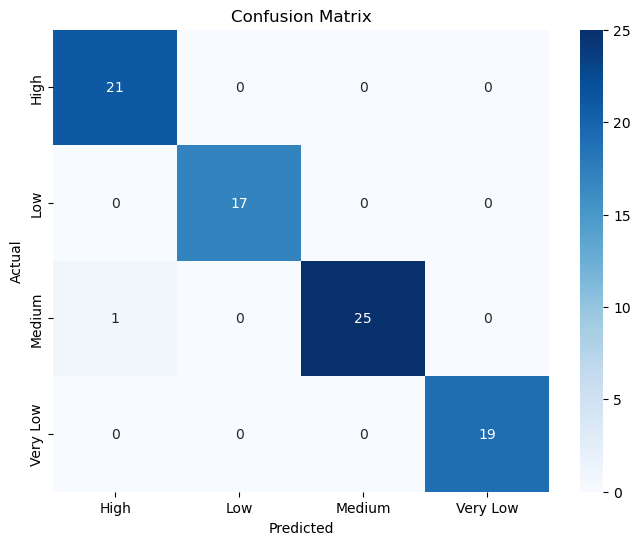
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| High | 0.95 | 1.00 | 0.98 | 21 |
| Medium | 1.00 | 0.96 | 0.98 | 26 |
| Low | 1.00 | 1.00 | 1.00 | 17 |
| Very Low | 1.00 | 1.00 | 1.00 | 19 |
| Accuracy |  |  | 0.99 | 83 |
| Macro Avg | 0.99 | 0.99 | 0.99 | 83 |
| Weighted Avg | 0.99 | 0.99 | 0.99 | 83 |

*Table 19: Classification model accuracy and F1 scores*

The model perfectly predicts “Low” and “Very Low” categories perfectly giving them f1 scores of 1.00 respectively.

One “Medium” state was classified as “High” which gave both categories an f1 score of 0.98 each as this score encapsulates precision and recall to give a total score.

The model is 0.99 accurate which indicate sit was correct 99% of the time. Overall, the model is very strong. This is evident in the confusion matrix below:

****

*Figure 22: Confusion matrix for Classification model*

A graph with numbers and symbols

Description automatically generated with medium confidence

*Figure 23: Scattered boxplot showing classification results*

**4.3 Sentiment Analysis**

2 sentiment analysis were performed, their task was as follows:

Model 1: Analyse the overall sentiment for the increase in crop prices in 2022 in Ireland.

Model 2: How do consumers and producers feel about crop prices?

Results:

|  |  |
| --- | --- |
| Sentiment Summary for Crop Price Spike in Ireland |  |
| Total posts analyzed | 2112 |
| Positive posts | 1320 |
| Negative posts | 489 |
| Neutral posts | 303 |
| Average sentiment | 0.08 |

*Table 20: Sentiment analysis model 1*

Out of 2112 post analysed, 62.5% were positive which indicates the increase in crop value was an overall positive event. 23.2% of posts were negative which indicates some individuals were not happy with the price increase. This could stem from increased costs for consumers. The overall sentiment is only 0.08 which indicates although public opinion appears to be positive there is mixed feelings (Selvaraj, 2020).

A graph of positive and negative

Description automatically generated

*Figure 24: Sentiment category count*

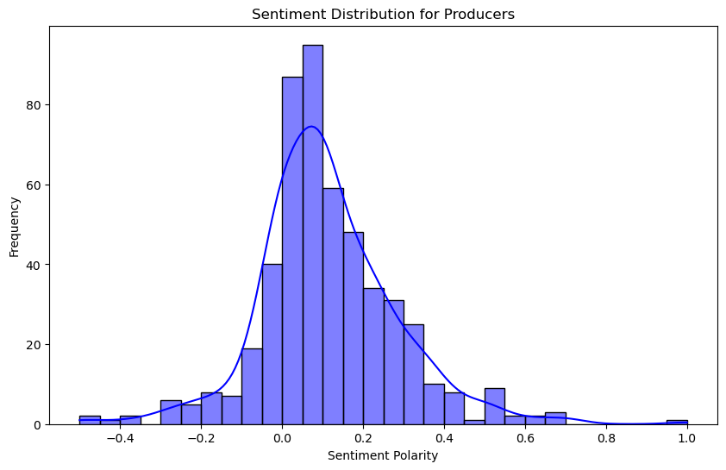
It was decided to break this down further by trying different key words and splitting the posts into “consumer” and “producer”.

|  |  |  |
| --- | --- | --- |
| Metric | Producers | Consumers |
| Total posts | 505 | 186 |
| Positive posts | 370 | 132 |
| Negative posts | 90 | 38 |
| Neutral posts | 45 | 16 |
| Average sentiment | 0.1131 | 0.0534 |

*Table 21: Sentiment analysis model 2*

The results support the initial hypothesis that the value increase is more favourable for producers with 73.3% of posts highlighting positive discussion. Although there is more negative posts in this category there was a much larger sample collected. This larger sample suggests producers are more impacted by the topic as it impacts their livelihood.

A graph of a line graph

Description automatically generated with medium confidence*Figure 25: Sentiment distribution producer Figure 26: Sentiment distribution consumer*

A graph of post and negative

Description automatically generated with medium confidenceA graph of post numbers

Description automatically generated with medium confidence

*Figure 27: sentiment count producer Figure 28: sentiment count consumer*

**5 Interactive Dashboard**

After carrying out extensive research on the Irish Agricultural sector it was decided to focus on Livestock and Crop value when creating a dashboard. Both sectors are unstable and fluctuate over the years, they are more profitable when they are focused in certain regions and it would be important that farmers can make an informed decision as to whether to move into a specific sector, increase their output or relocate to a different region based off past trends and by obtaining a prediction for future value.

The dashboard directly compares 2 agriculture sectors and informs the user which one is currently performing better and a future prediction based off of current trends.

In the dashboard **the representation of numerical data is proportionate.** All barcharts and lineplots are directly proportional to each other. All data is labelled using legends to avoid confusion. There is little writing which could cloud visuals. These design choices were made in alignment with Edward Tufte’s Principles (Tufte, 1983).

**5 Programming:**

5.2 Data Sources

Flat files (CSV, TSV, JSON) can be managed effectively using libraries such as “pandas” within the notebook directly. “Pandas” can be used for smaller SQL databases while SQLAIchemy would be more effective for large databases. NoSQL databases like MongoDB can be handled using libraries such as “pymongo”. Libraries such as “requests” and “praw” can be used to interact with APIs. The choice of library depends on the tasks required and the size of the data.

5.3 Data Manipulation

Throughout this programme data was manipulated and aggregated to gain meaningful insights. Merging and concatenation was used to join data frames allowing for direct comparisons and effective visualisation. Dataframes were melted and pivoted to ensure they were in the correct format for aggregation. Data was grouped by characteristics such as “Year” or summed to understand total populations. This is effective as it allows us to see the bigger picture in our data but in the process we sometimes loose category specification and may oversimplify our data (Wen, 2020).

5.5 Testing

“GridSearchCV” was used to ensure the ML models were performing optimally and flagged if parameters needed to be tuned. The model was fitted to training data and evaluated on test data. Metrics such as R^2 , RMSE and MSE scores provided accuracy and error rate insights. Cross validation requires more time but ensures optimum results. When carrying out sentiment analysis debugging and filtering were incorporated to ensure quality data. Tokenization and stemming were incorporated and debugging checkpoints were added to check the number of posts after each filtering steps. This significantly reduced the data but ensured it was relevant.

5.6 Optimisation:

Throughout the programme several steps were taken to ensure optimisation:

* To reduce memory usage choropleth maps, were saved as HTML files using Plotly (McQuaid, 2024c).
* Geometry was simplified in the .shp file to reduce memory usage
* A search limit was included when conducting sentiment analysis to ensure effective run time.

These strategies ensured efficient use of CPU, RAM, and time, enhancing overall performance and scalability. It also ensure my notebooks were a suitable storage size.

**Project Management Framework:**

This project followed a CRISP-DM framework. It consists of 6 stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment (Hotz, 2023). I had no understanding of the agriculture sector so I had to spend time understanding the domain. Next I began collecting data and through EDA explored interesting subsets and tried to understand what I was working with. The datasets weren’t perfect and extensive data preparation was conducted, this took up most of the project work load. Next I applied statistical and machine learning models to gain interesting insights. Finally I deployed my findings in an interactive dashboard.

**References:**

CSO statistical publication (2024). Area, Yield and Production of Crops 2023 - Central Statistics Office. [online] www.cso.ie.

Creative Commons (2016). Creative Commons — Attribution 4.0 International — CC BY 4.0. [online] Creativecommons.org. Available at: <https://creativecommons.org/licenses/by/4.0/>.

EUROSTAT (2024). [online] Europa.eu.

Inuwa, M. (2022). *Outliers and Overfitting when Machine Learning Models can’t Reason -*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2022/07/outliers-and-overfitting-when-machine-learning-models-cant-reason/.

Iqbal, M. (2024). *Tutorial\_2*. [Lecture].

Hegelich, S. (2016). Decision Trees and Random Forests: Machine Learning Techniques to Classify Rare Events. *European Policy Analysis*, 2(1). doi:https://doi.org/10.18278/epa.2.1.7.

Hotz, N. (2023). *What is CRISP DM?* [online] Data Science Project Management. Available at: https://www.datascience-pm.com/crisp-dm-2/.

McElwain, A. (2019). *Whatever happened to Irish sugar?* [online] The Irish Times. Available at: https://www.irishtimes.com/life-and-style/food-and-drink/whatever-happened-to-irish-sugar-1.3886285.

Ministerie van Landbouw, N. en V. (2022). *Ireland looks more and more to the Netherlands for solutions and technology to achieve its agricultural goals - Nieuwsbericht - Agroberichten Buitenland*. [online] www.agroberichtenbuitenland.nl.

McQuaid, D. (2024a). *How to Deal with Missing Data in Python*.

McQuaid, D. (2024b). *What Visualisations should I use*.

McQuaid, D. (2024c). *Interactive Visualization of Geographical Data.*

Ministerie van Landbouw, N. en V. (2022). *Ireland looks more and more to the Netherlands for solutions and technology to achieve its agricultural goals - Nieuwsbericht - Agroberichten Buitenland*. [online] www.agroberichtenbuitenland.nl

Wen, T. (2020). Data Aggregation. *Encyclopaedia of Big Data*, pp.1–4. doi:https://doi.org/10.1007/978-3-319-32001-4\_296-1.

Selvaraj, N. (2020). *A Beginner’s Guide to Sentiment Analysis with Python*. [online] Medium. Available at: https://towardsdatascience.com/a-beginners-guide-to-sentiment-analysis-in-python-95e354ea84f6.

Stephen Allwright. (2022). *How to interpret MSE (simply explained)*. [online] Available at: <https://stephenallwright.com/interpret-mse/>.

Tufte, E.R. (1983). *The Visual display of quantitative information*. Cheshire, Conn.: Graphics Press.

**Note: issues with github commits, tried to fix but was unsuccessful, I have attached screenshots from computer, sorry about this.**

**Word count: 2976**