**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_CA2* |
| **Lecturer Name:** | *Marina Iantorno/John O’Sullivan*  *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | *Naomi Tunstead* |
| **Student Number:** | *sba22222* |
| **Assessment Due Date:** | *6th January 2023* |
| **Date of Submission:** | *6th January 2023* |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Predictive Modelling on Cattle Prices in Ireland and France**

**MSc in Data Analytics**

Author: N. Tunstead

e-mail: [sba22222@student.cct.ie](mailto:sba22222@student.cct.ie)

Student ID: sba22222

**Abstract**

Accurate prediction of cattle prices can significantly impact the profitability of farmers and the cost of meat for consumers. In this project, we develop machine learning models to predict the price of cattle carcasses in Ireland and France. We gather and analyse data on various factors that may influence cattle prices, such as year, week, feed price, and cattle category, and use multiple machine learning models to identify the most effective model for predicting cattle prices. Our study incorporates sentiment analysis of grain prices to provide a comprehensive understanding of market conditions that impact cattle prices. The results of this project provide valuable insights and tools for farmers to optimize their operations and maximize profits.

**Keywords:**

Cattle prices, farmers, machine learning, sentiment analysis

Table of Contents

[Introduction 8](#_Toc123941307)

[Motivation 8](#_Toc123941308)

[Background 8](#_Toc123941309)

[Machine Learning 8](#_Toc123941310)

[CRISP-DM Framework 9](#_Toc123941311)

[Data Processing & Analysis 10](#_Toc123941312)

[Business Understanding 10](#_Toc123941313)

[Determine the Business Objectives 10](#_Toc123941314)

[Research Questions 10](#_Toc123941315)

[Project Plan 10](#_Toc123941316)

[Version Control 11](#_Toc123941317)

[Data Understanding 11](#_Toc123941318)

[Collecting the Data 11](#_Toc123941319)

[Describe Data 12](#_Toc123941320)

[Data Preparation 15](#_Toc123941321)

[Data Cleaning 15](#_Toc123941322)

[Missing Values 15](#_Toc123941323)

[Statistics 17](#_Toc123941324)

[Modeling 22](#_Toc123941325)

[Decision Tree 22](#_Toc123941326)

[Random Forest 22](#_Toc123941327)

[Linear Regression 22](#_Toc123941328)

[Ridge Regression 22](#_Toc123941329)

[Lasso Regression 23](#_Toc123941330)

[Gradient Boosting 23](#_Toc123941331)

[GridSearch CV Hyperparameter 23](#_Toc123941332)

[Evaluation 23](#_Toc123941333)

[Deployment 25](#_Toc123941334)

[Deployment plan 25](#_Toc123941335)

[Monitor 25](#_Toc123941336)

[Sentiment Analysis 26](#_Toc123941337)

[Dashboard 27](#_Toc123941338)

[References 29](#_Toc123941339)

[Appendices 30](#_Toc123941340)

[Appendix A: CRISP-DM Framework 30](#_Toc123941341)

[Appendix B: GitHub Link 30](#_Toc123941342)

[Appendix C: European Commission License/Permission 30](#_Toc123941343)

[Appendix: Word Count 30](#_Toc123941344)

List of Figures

Figure1: Simple Machine Learning Workflow 9

Figure2: CRISP-DM Framework 9

Figure 3: Gantt Chart of Project Plan 11

Figure 4: Violin plot of Distribution of target attribute Cattle Price 14

Figure 5: Histogram of Distribution of Feed Price before missing values were back filled 16

Figure 6: Histogram of Distribution of Feed Price after missing values were back filled 16

Figure 9: Pair plot of Attributes in Ireland 17

Figure 10: Pair plot of Attributes in France 18

Figure 11: Correlation Heatmap of Attributes in Ireland 18

Figure 12: Correlation Heatmap of Attributes in France 19

Figure 7: Density and ox Plot of Cattle Prices in Ireland and France 20

Figure 8: Shapiro-Wilk Test and Q-Q Plot of Cattle Prices in Ireland and France 20

Figure 9: Most popular words used in sentiment analysis on grain prices 26

List of Tables

[Table 1: Descriptive Statistics on df\_cattle 13](#_Toc123941390)

[Table 2: Descriptive Statistics on df\_feed 13](#_Toc123941391)

[Table 3: Results Table for Each Model Before GridSearch CV was Applied 25](#_Toc123941392)

[Table 4: Results Table for Each Model After GridSearch CV was Applied 25](#_Toc123941393)

# Introduction

## Motivation

The price of cattle carcasses is an important factor in the livestock industry, as it affects the profitability of farmers and the cost of meat for consumers. Accurate prediction of cattle prices can help farmers make informed decisions about their operations and enable meat processors to plan their production and pricing strategies. In this project, we aim to develop machine learning models to predict the price of cattle carcasses in Ireland and France. We will gather and analyse data on various factors that may influence cattle prices, such as year, week, feed price and cattle category. By training and evaluating multiple machine learning models on this data, we hope to identify the most effective model for predicting cattle prices in these two countries. The results of this study may be of interest to a wide range of stakeholders in the livestock industry, including farmers, meat processors, and policymakers.

In addition to predicting the price of cattle carcasses, this model will also incorporate sentiment analysis of grain prices in order to provide a more comprehensive understanding of the market conditions that impact the price of cattle. By analysing the sentiment of grain prices, we can better understand how trends in these markets may influence the demand for cattle and, ultimately, the price farmers can expect to receive for their cattle carcasses.

Overall, this project aims to provide valuable insights and tools for modern farmers to optimize their operations and maximize their profits.

# Background

## Machine Learning

A subset of artificial intelligence (AI) called machine learning enables computers to solve a variety of problems more frequently and accurately than a human being. For instance, algorithms have been designed to detect tumors in patients more precisely than doctors. The objective of machine learning is to create software that can identify patterns and rules in data and use them in unexpected contexts. This enables machines to infer solutions, doing away with the necessity for explicit programming. Many of the issues facing the modern world are complicated, making it difficult to quickly encode them into a computer and fix them in a binary manner. Machines may handle problems in a variety of ways, and these ways are always being improved through research and analysis. The three divisions of machine learning are supervised learning, unsupervised learning, and reinforcement learning. The main focus of this project is the use of supervised learning techniques. The supervised learning techniques are the main emphasis of this project.

A unified algorithmic framework called machine learning was developed to discover computational models that, with little to no human involvement, accurately describe empirical data and the phenomena that underlie it. Figure 1 shows a typical machine learning framework and highlights the key phases in their respective blocks. In order to understand and build applications using a machine-learning framework, we employ the techniques provided in the blocks. A dataset needs to be obtained, cleaned, and separated into training and testing datasets before anything else. The data must subsequently be modified or normalized according to the task at hand. The predictive model is then fitted using the training data. The appropriate model parameters for the method are determined using the training data. Following that, statistical tests are used to evaluate the prediction model's results to see if they are statistically significant and if they are compatible with the hypothesis. The prediction model can then be used and improved with new information. To establish a more cohesive structure, this is depicted in Figure 2. For more detail, please see Appendix A.

Diagram

Description automatically generated

#### Figure1: Simple Machine Learning Workflow

## CRISP-DM Framework

Diagram

Description automatically generated

#### Figure2: CRISP-DM Framework

# Data Processing & Analysis

## Business Understanding

### Determine the Business Objectives

In this project, our aim is to predict the price of cattle carcasses in Ireland and France using machine learning models. Accurate prediction of cattle prices is of interest to a variety of stakeholders in the livestock industry, including farmers, meat processors, and policymakers.

For farmers, understanding the expected price of their cattle carcasses can help them make informed decisions about their operations, such as which breeds to raise and when to sell their animals. Meat processors also have a vested interest in accurate cattle price prediction, as it can help them plan their production and pricing strategies. Policymakers may also find it useful to have accurate cattle price predictions in order to track trends and make informed decisions about regulations and policies related to the livestock industry.

In order to achieve our goal of predicting cattle prices, we will gather data on various factors that may influence cattle prices, such as breed, age, weight, and grade. We will then use machine learning techniques to build predictive models that can take this data as input and output a predicted cattle price. By evaluating the performance of these models, we hope to identify the most effective model for predicting cattle prices in Ireland and France.

### Research Questions

This research will answer the following research questions (RQs):

RQ1 - What are the data requirements for predicting cattle prices for Ireland and France?

RQ2 - How should machine learning be applied to best predict cattle prices?

RQ3 - How can visualisation aid in delivering insight about cattle prices from this predictive model?

### Project Plan

To ensure a smooth and efficient project process, the team will divide the work into the following phases:

Phase 1: Research the current situation and analyse the structure of the data (dataset examination)

Phase 2: Implement code for data representation (data exploration and visualization)

Phase 3: Prepare the data for modeling (selection, cleaning, formatting, and any other necessary actions)

Phase 4: Select modeling techniques and build the model

Phase 5: Analyse the results from the previous phase. Repeat phases 4 and 5 as needed.

Phase 6: Create a report with the results.



#### Figure 3: Gantt Chart of Project Plan

### Version Control

GitHub is a web-based platform that provides version control and collaboration tools for software development. Git Bash was used to pull Jupyter notebooks up to GitHub online where each version of the code was stored. The link to my GitHub can be found in Appendix B

## Data Understanding

Any machine learning project must include the data understanding phase. In order to comprehend a machine learning model's traits and properties, it requires investigating and analysing the data that will be used to train it. This stage is crucial since the accuracy and usefulness of the data can greatly affect how well the machine learning model performs.

It's critical to spot any potential issues or difficulties with the data during the data understanding step, such as missing numbers, outliers, or unbalanced classes. Identifying the links and patterns in the data is also crucial since it can help in choosing the right machine learning algorithms and evaluation measures. Gaining knowledge of the domain or issue under study can also help direct the creation of the machine learning model and facilitate the interpretation of the findings.

Overall, the data understanding step is critical for making sure the machine learning model is constructed on a firm foundation of high-quality and relevant data, which is necessary for obtaining good performance and delivering trustworthy results.

### Collecting the Data

The process of acquiring raw data for my research project was both rewarding and challenging. On the positive side, I was able to access a wide range of data sources that helped me to understand my research question more deeply. For example, I was able to obtain data from public databases such as Eurostat (*Database - Eurostat*, no date), FAO (*FAOSTAT*, no date) and European Commission(Development, no date). This research provided me with valuable insights into my topic even though I didn’t end up using some of them.

However, there were also some negative aspects to my data acquisition process. One challenge I faced was the time it took to locate and access the data I needed. Some data sources were more difficult to find than others, and some required me to go through a lengthy application process before I could access the data. Additionally, I encountered some technical difficulties when downloading and organizing the data, which required me to spend additional time troubleshooting and cleaning the data. For these datasets I could not investigate them any further for this reason which wasted valuable time.

Overall, the process of acquiring raw data for my research project was a valuable experience. While it had its challenges, the data I was able to obtain helped me to make significant progress on my research and contributed to the success of my project.

The data for the predictive model on cattle prices was collected from the European Commission Agri-Food Markets. (Development, no date) The licence/permission can be found in Appendix C For the sentiment analysis it was obtained from twitter. I created a twitter developer account was granted an API key in order to interact with twitter data. (*Developer Agreement and Policy – Twitter Developers*, no date)

### Describe Data

The df\_cattle dataset contained 6150 rows and 8 columns. The columns included Year (2022, 2021), Week (1 to 52), Begin Date (date of the Monday of every week), End Date (date of the Sunday of every week), Member State (Ireland, France), Category (Young cattle, Young bulls, Steers, Heifers, Cows, Calves slaughtered <8M, Bulls), Product (letter and number grade given to cattle based on quality of cattle) and Price (per cattle)

The df\_feed dataset contained 240 rows and 7 columns. The columns included Marketing Year (2022/2023, 2021/2022, 2020/2021), Reference Period (date of the Thursday of every week), Member State (Ireland, France), Product Name (Feed Barley), Market Name (Rouen, Dublin/North East/Midlands), Stage Name, Price (€/Tonne)

In Table 1, we can see the df\_cattle descriptive statistics. Important statistic to take note of are the mean and the median. For the ‘Year’, ‘Week’ and ‘Price’ attributes we can see both the mean and the median are roughly equal. Therefore, the data in the ‘Year’, ‘Week’ and ‘Price’ columns are symmetrical meaning our data is normally distributed. However, if we look at Table 2 for the df\_feed data frame ‘Price (€/Tonne)’ attribute the mean is 261 and the median is 240, which are not equal. Therefore, the data is skewed in the direction of the mean. This is important to note as this will impact the statistical inference.

The range of the data in each attribute is also an important statistic. For df\_cattle the ‘Year’ and ‘Week’ attributes have a relatively small range of 1 and 51, respectively. However, for both ‘Price’ in df\_cattle and ‘Price (€/Tonne)’ in df\_feed both have a large range of 549 and 258. This tells us the mean may be heavily influenced by outliers or extreme values, which could make it less representative of the data. Therefore, from now on we may need to use the median value for our measure of central tendency.

The standard deviation is a measure of the dispersion or spread of a dataset. It is calculated as the square root of the variance, which is the average of the squared differences between the values and the mean. Here we can see for the ‘Year’ and ‘Week’ attribute in df\_cattle the standard deviation is 0.5 and 14.7 which are small. Indicating that the values are more concentrated around the mean. However, for the ‘Price’ attribute in df\_cattle and ‘Price (€/Tonne)’ attribute in df\_feed have significantly higher standard deviations of 71 and 70, indicating that the values are more spread out. It's worth noting that the standard deviation is sensitive to outliers, or values that are significantly larger or smaller than the rest of the data. If there are outliers in the dataset, the standard deviation may be artificially inflated and may not accurately reflect the dispersion of the data. In such cases, other measures of dispersion, such as the interquartile range, may be more appropriate.

The interquartile range (IQR) is a measure of dispersion or spread in a dataset. It is calculated as the difference between the 75th percentile and the 25th percentile and is a robust measure that is less sensitive to outliers than the range or standard deviation. The IQR tells us the range of the middle 50% of the values in a dataset. For the for the ‘Year’ and ‘Week’ attribute in df\_cattle the IQR is 1 and 25 this is small meaning the middle 50% of the values are concentrated around the median, indicating that the data is more homogeneous. However, for the ‘Price’ attribute in df\_cattle and ‘Price (€/Tonne)’ attribute in df\_feed have an IQR of 99 and 103, indicating that there is a significant amount of variability in the data.

Our date in our target attribute ‘Price’ in df\_cattle is continuous and so is ‘Price (€/Tonne)’ in df\_feed. While all the other columns which are numerical contain discrete variables. Therefore we used a regression model instead of a classification model during our modelling phase.

Text

Description automatically generated

##### Table 1: Descriptive Statistics on df\_cattle

Text, table

Description automatically generated

##### Table 2: Descriptive Statistics on df\_feed

In Figure 4, we can see that our target attribute cattle prices for both Ireland and France not only have different distributions with France having outliners and Ireland being very compact and more normally distributed. There is also an interesting significant difference between the cattle category produced in each country. Not only does France produce ‘Calves slaughtered <8M’(purple), aka veal, and Ireland doesn’t, it also Frances largest producing category. Likewise, Ireland produces ‘Young cattle’ (orange) whereas France does not. Finally, the other interesting contracts between both countries is that France produces less carcass from ‘Steers’(pink) than Ireland. Likewise, Ireland produces less carcass from ‘Cows’ (brown) than France. This category of carcass is poorest quality and would naturally get the lowest grading.

A picture containing scatter chart

Description automatically generated

#### Figure 4: Violin plot of Distribution of target attribute Cattle Price

## Data Preparation

The data preparation phase is an essential step in any machine learning project. It involves the process of collecting, cleaning, and pre-processing the data that will be used to train and evaluate the machine learning model. This phase can be time-consuming and labour-intensive, but it is crucial for ensuring that the model is able to learn from high-quality data and generalize well to new, unseen data.

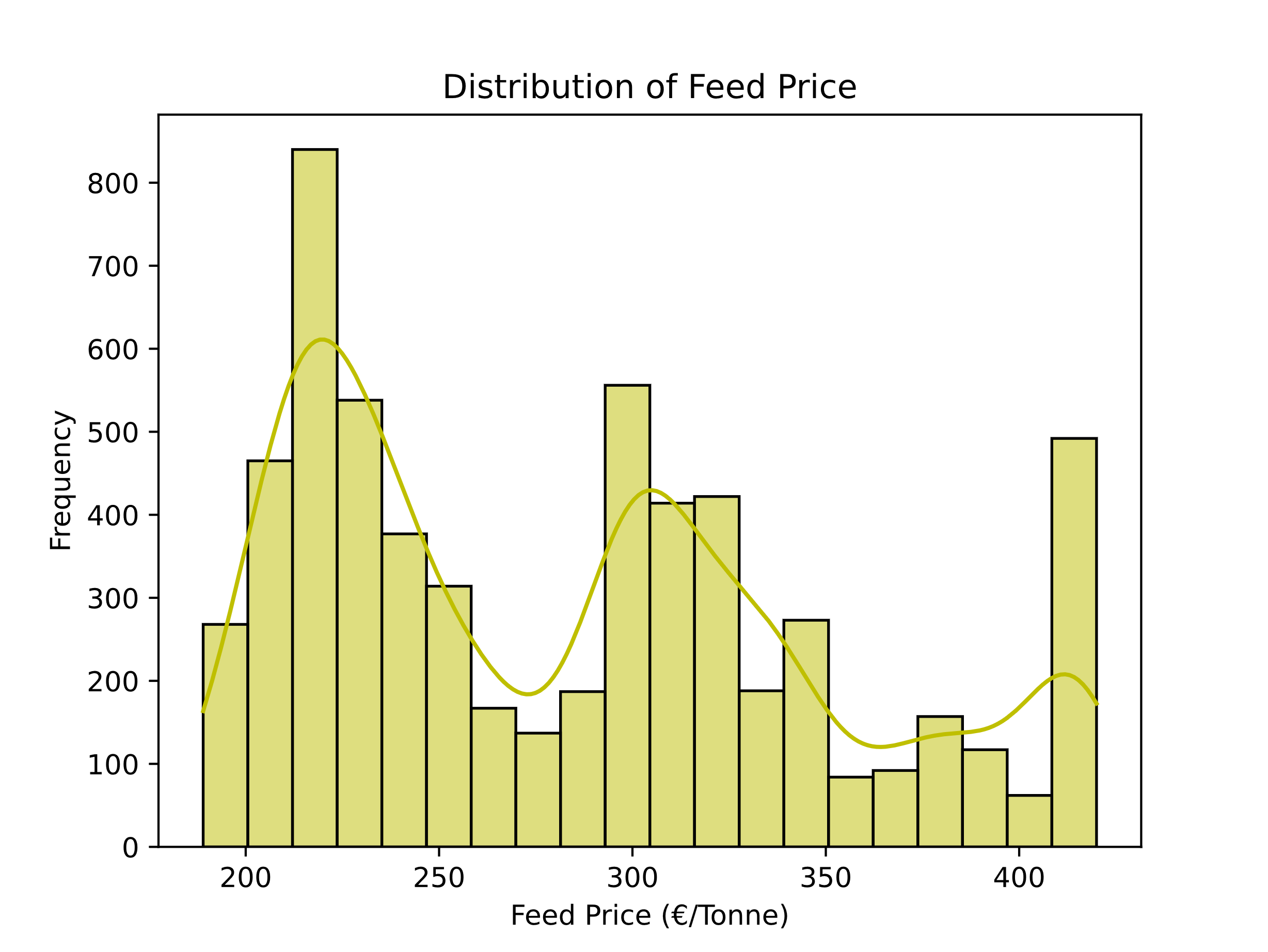
In this project, we will go through the various steps of data preparation, including collecting the necessary data, identifying and handling any missing or corrupted values, and selecting and transforming the features that will be used as inputs to the model. We will also discuss strategies for splitting the data into training, validation, and test sets. By the end of this phase, we will have a clean and properly formatted dataset ready for model training and evaluation.

### Data Cleaning

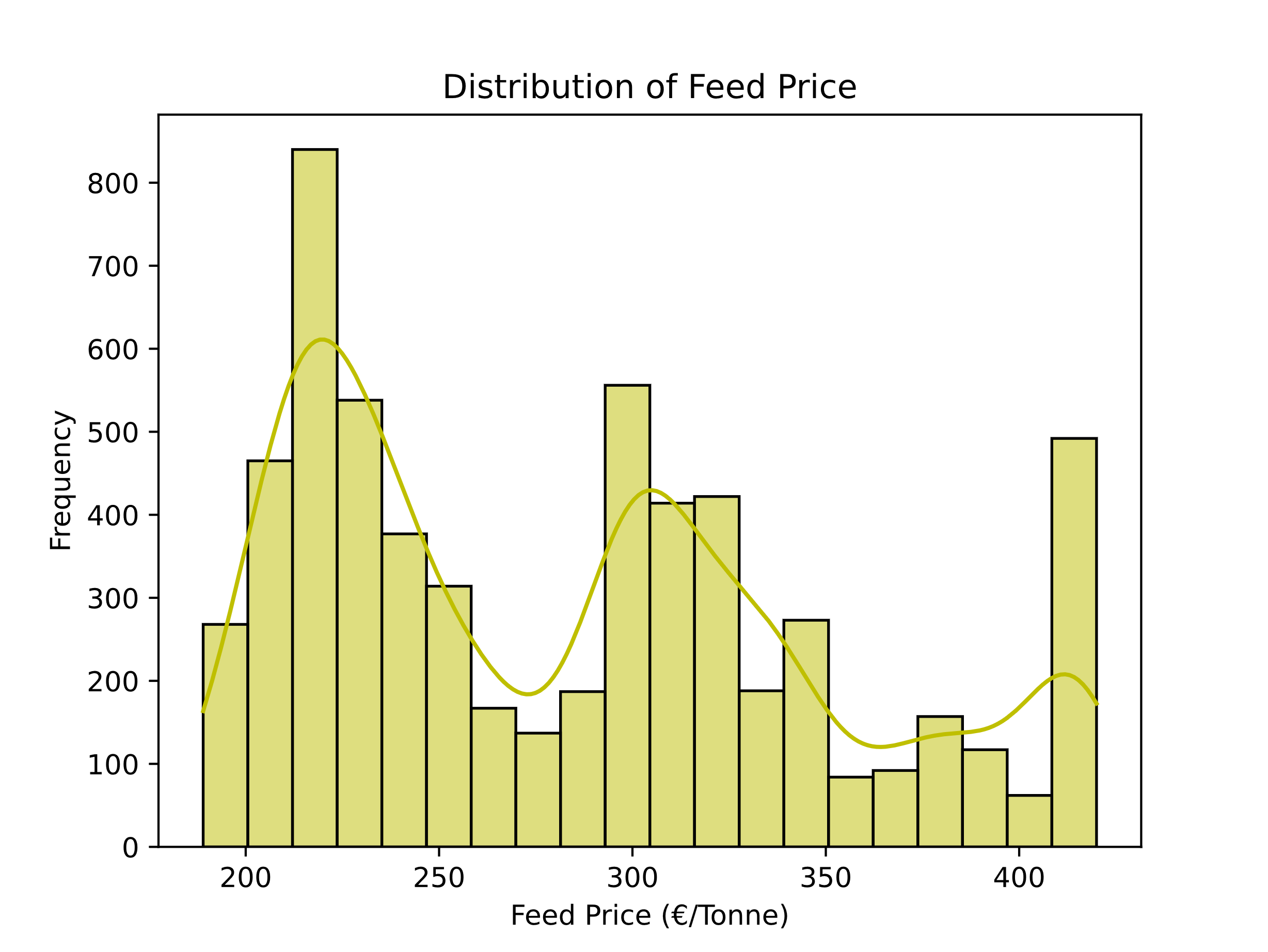
We observed that there were several columns we did not need. Such as in df\_cattle there were ‘End Date’ and ‘Product’, and in df\_feed there were ‘Year’, ‘Product Name’, ‘Market Name’ and ‘Stage Name’. None of these columns added any value to our analysis and therefore were dropped. We also noticed that the dates for both datasets did not match. We needed to resolve this in order for us to merge the two datasets. Using a lambda function we moved all the ‘Reference Period’ dates back three days, so they are no longer on a Thursday but on a Monday so that the dates in both data frames matched. Note this may affect the accuracy of our analysis. We then renamed the ‘Begin Date’ column in df\_cattle to ‘Reference Period’ to match the column in df\_feed. This allowed us to complete a left join. Not only did this join both out datasets together it removed any rows in the df\_feed dataset we did not need. However, this left us with 317 missing values for ‘Feed Prices’.

### Missing Values

There are several ways you can resolve this issue. Impute the missing values with the median, mean, backward filling, forward filling or even using machine learning such as MissForest. We also could have deleted the missing value rows from our dataset however we wanted to keep as much data as possible as this attribute is not our target variable and we wanted to keep as much data as possible. We choose to go with the simplest and most effective way, backward filling. We choose this method because it preserved the overall trend of the time series by replacing missing values with the most recent valid observation and kept the seasonality in the data. This can be seen in Figure 5 and Figure 6.



#### Figure 5: Histogram of Distribution of Feed Price before missing values were back filled



#### Figure 6: Histogram of Distribution of Feed Price after missing values were back filled

### Statistics

Inferential statistics is a branch of statistics that deals with making predictions or inferences about a population based on a sample. It is an important part of machine learning, as it helps to evaluate the performance of a machine learning model and to understand the underlying relationships between variables.

Below are pair plots in figures 9 and 10 depicting the scatterplots of all pairs of variables in a dataset for Ireland and France. This was used to identify relationships and patterns between the variables. Some key insights to take from these plots are that the variable in the scatter plot for Ireland and France are roughly the same. Both show a weak positive relationship between cattle prices and feed prices as we can see with the vertical block of data points with the slight tilt to the right. It is also interesting to note that both show how few data points feed prices has in comparison to cattle price when compared to the week attribute as this depicts the amount of imputation of null values was needed for feed price.

In figure 11 and 12, we see the Pearson correlation between all pairs of columns in both Ireland and France. In both plots we see a strong positive correlation between feed price and cattle price with the year attribute. For Ireland its 0.79 and 0.65, and for France its 0.84 and 0.66. Likewise, for both Ireland and France cattle price and feed price have a positive correlation. France is stronger than Ireland with a value of 0.71 compared with Ireland at 0.58. This means that these variables have a strong linear relationship. In contrast, feed price and cattle price compared to the week variable for both Ireland and France is a weak positive correlation. This means they have a weak linear relationship.

Chart

Description automatically generated

#### Figure 9: Pair plot of Attributes in Ireland

Diagram

Description automatically generated

#### Figure 10: Pair plot of Attributes in France

Chart

Description automatically generated with low confidence

#### Figure 11: Correlation Heatmap of Attributes in Ireland

A picture containing background pattern

Description automatically generated

#### Figure 12: Correlation Heatmap of Attributes in France

In this project, ANOVA testing was used to compare the means of Ireland and France to see if there are any significant differences between the groups. ANOVA helped determine if these differences are due to random chance or if they are statistically significant.

In order to preform ANOVA testing four conditions must be meet first.

1. The variance of the distributions are equal
2. The samples are normally distributed
3. The samples are independent
4. The samples have equal sample sizes

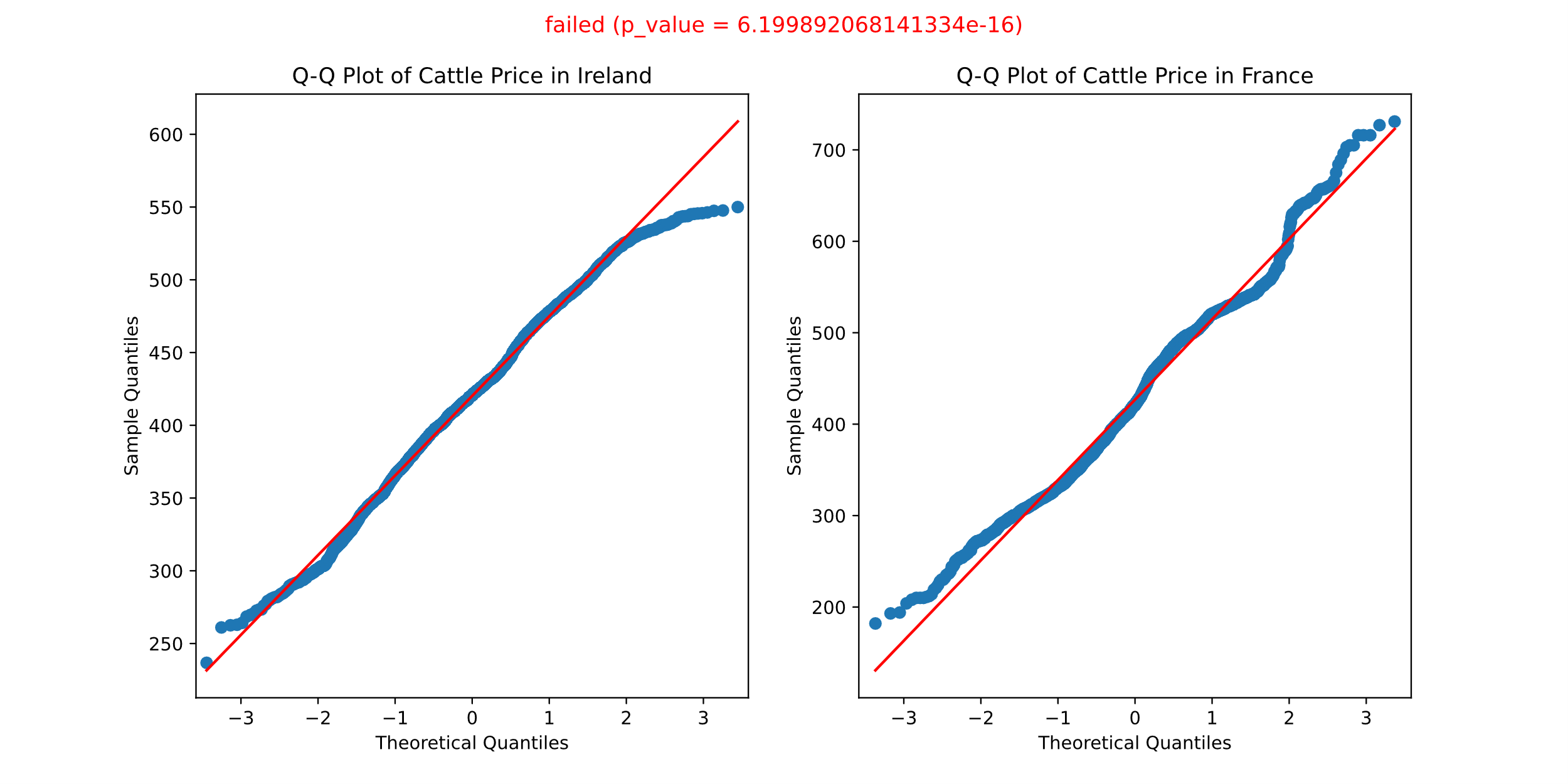
Firstly, we tested if the variance between the groups were equal. We used the Levenes’s test. The Levene's test is a statistical test used to assess the equality of variances between two or more groups. The null hypothesis of the Levene's test is that the variances of the groups are equal. The alternative hypothesis is that the variances are not equal. The test was used to determine whether the variance of a dependent variable was equal across different levels of an independent variable. The Levene's test statistic was 845.76 and the p-value was 2.48e-174. Based on these results, we rejected the null hypothesis that the variances of the groups are equal. The large value of the test statistic and the very small p-value indicated that there was a significant difference in the variances between the groups. This can also be seen in figure 7 where the density plot and the box plot of Ireland versus France show that the spread of the distributions is not equal.

Chart, box and whisker chart

Description automatically generated

#### Figure 7: Density and ox Plot of Cattle Prices in Ireland and France

Secondly, we used the Shapiro-Wilk test as a statistical test to assess the normality of the samples. The null hypothesis of the Shapiro-Wilk test was that the samples came from a normally distributed population. The alternative hypothesis was that the sample did not come from a normally distributed population. The p-value of the Shapiro-Wilk test was 6.2e-16. This indicated that there was strong evidence against the null hypothesis that the sample came from a normally distributed population. The very small p-value suggests that it was unlikely that the sample came from a normally distributed population by chance. Therefore, we rejected the null hypothesis and concluded that the sample is not normally distributed. This is not made clear on the Q-Q plots in figure 8 as the points on both plots for Ireland and France are roughly formed in a straight line which would indicate that the data is normally distributed. Likewise, in the density plot and the box plot. This make because the sample size is too small for the test. Alternatively, the data may not be independent as the Shapiro-Wilk test is based on the assumption that the data are independently and identically distributed (IID).



#### Figure 8: Shapiro-Wilk Test and Q-Q Plot of Cattle Prices in Ireland and France

Thirdly, since our data was continuous, we could not use the chi-squared test to test for independence. Therefore, we used the t-test and the Mann-Whitney U test to test the mean and the median of the two samples.

The null hypothesis for the t-test was that the means of the two groups are equal. The alternative hypothesis was that the means of the two groups are not equal. The t-statistic was -3.63, which indicates that the means of the two groups differ significantly. The p-value was 0.00029, which was less than 0.05. This means that the difference between the means was statistically significant, and the null hypothesis can be rejected. Therefore, we can conclude that there was a significant difference between the means of the two groups.

The Mann-Whitney U test was a non-parametric statistical test used to compare the medians of two groups. The null hypothesis for the Mann-Whitney U test was that the medians of the two groups are equal. The alternative hypothesis was that the medians of the two groups are not equal. The U-statistic was 4453211.0, which indicates that the medians of the two groups differ significantly. The p-value was 0.00869, which was less than 0.05. This means that the difference between the medians was statistically significant, and the null hypothesis was rejected. Therefore, we concluded that there was a significant difference between the medians of the two groups.

Finally, since we did not meet all the conditions of ANOVA, we could not conduct that test because our results would not be accurate or reliable. Also, since our data was non-parametric, we conducted a similar test to ANOVA called Kruskal-Wallis H test. It was a non-parametric statistical test used to compare the mean ranks of two or more samples. The null hypothesis for the Kruskal-Wallis H test was that there was no difference in the medians of the groups being compared. The alternative hypothesis was that there was a difference in the medians of the groups, indicating that the groups came from different populations. The H statistic was 6.885 and the p-value was 0.00869. This suggested that there was a significant difference between the groups being compared, as the p-value was less than 0.05. This meant that it was unlikely that the observed difference between the groups was due to chance and suggested that the null hypothesis should be rejected in favour of the alternative hypothesis.

## Modeling

In this section, we will delve into the process of building predictive models for our dataset. Modeling is a crucial step in the machine learning process, as it involves training algorithms on the data and using the resulting model to make predictions or decisions.

There are many different approaches to modeling, and the specific method that we choose will depend on the nature of the problem we are trying to solve and the characteristics of the data. The techniques we used are decision tree, random forest, linear regression, ridge regression, lasso regression and gradient boosting.

### Decision Tree

In this project, we used a decision tree regressor as our model of choice for a couple of reasons. Our dataset contained both continuous and categorical features, and decision tree models are well-suited to handling such datasets. The tree structure of decision trees allows them to split the data based on either the value of a continuous feature or the category of a categorical feature.

Secondly, decision tree models are relatively fast to train and make predictions, which was important for this project as we were working with a short space of time and needed to make timely predictions.

### Random Forest

We used a random forest regressor as one of our models because, firstly, random forest models are known for their good performance on a wide range of tasks, including regression. They can capture non-linear relationships in the data and handle high-dimensional datasets effectively.

Additionally, random forest models are relatively robust to overfitting, as they build multiple decision trees and combine their predictions to make a final prediction, rather than relying on a single tree. Like decision trees random forest models are relatively fast to train and make predictions.

### Linear Regression

Linear regression is a simple and straightforward model that is easy to understand and interpret. It makes predictions by fitting a linear equation to the data, which can be helpful if the relationship between the features and the target variable is approximately linear. Again, linear regression is relatively fast to train and make predictions.

Finally, linear regression is well-suited to problems with a small number of features, as it can be difficult to accurately estimate the coefficients of a linear equation when there are many features. In this project, we had a relatively small number of features, which made linear regression a good choice.

### Ridge Regression

Ridge regression is a type of linear regression that includes a regularization term to prevent overfitting. Similar to the above models ridge regression is relatively fast to train and make predictions. It is mostly used for larger dataset then the one we have but it still can be used.

Finally, ridge regression is able to handle multicollinearity in the data, which can occur when two or more features are highly correlated. Multicollinearity can cause problems for linear regression models, but ridge regression is able to mitigate these issues by adding the regularization term.

### Lasso Regression

Lasso regression is a type of linear regression that includes a regularization term to prevent overfitting. Like ridge regression, lasso regression can be useful if we have a large number of features and are concerned about the model overfitting to the training data.

However, unlike ridge regression, which includes all the features in the model, lasso regression can also perform feature selection by setting some of the feature coefficients to zero. This can be useful if we have a large number of features and are interested in identifying the most important features for the model. It is also relatively fast to train and make predictions

### Gradient Boosting

Gradient boosting is a powerful and flexible ensemble method that can achieve good performance on a wide range of tasks, including regression. It works by building a series of weak models, such as decision trees, and combining them to make a final prediction.

Additionally, gradient boosting is able to handle a variety of data types and is resistant to overfitting, as the weak models are combined in a way that reduces their individual complexity. Finally, like all the rest of the model selected for this project gradient boosting is relatively fast to train and make predictions.

### GridSearch CV Hyperparameter

Grid search cross-validation (GridSearchCV) is a method used to tune the hyperparameters of a model in order to achieve the best performance. It does this by training the model with a range of different hyperparameter combinations and evaluating each one using cross-validation. This allows you to identify the set of hyperparameters that produces the best performing model. GridSearchCV is particularly useful when you have a limited understanding of the model's hyperparameters and their possible values, or when you have many hyperparameters to tune and you need to find the optimal combination. It is also a useful tool for reproducing results, as it ensures that the best set of hyperparameters is identified in a systematic and objective manner.

## Evaluation

The evaluation stage is a crucial step in any machine learning project, as it involves assessing the performance of the model and determining how well it can solve the problem it was designed for. There are several different evaluation metrics that can be used to measure the performance of a model.

Since this project is using regression models on continuous data the metrics, we will be using to evaluate the performance of are models are MAE, MSE, RMSE, Train Score, Test Score and R2.

The mean absolute error (MAE) is a measure of the average magnitude of the errors in the model's predictions. It is calculated as the average absolute difference between the predicted values and the true values. A smaller MAE score indicates that the model's predictions are closer to the true values on average.

The mean squared error (MSE) is a measure of the average squared difference between the predicted values and the true values. It is generally preferred over the MAE because it penalizes larger errors more heavily. A smaller MSE score indicates that the model's predictions are closer to the true values on average.

The root mean squared error (RMSE) is the square root of the MSE and is expressed in the same units as the target variable. It is a more interpretable measure of error, as it is expressed in the same units as the target variable. Like the MSE, a smaller RMSE score indicates that the model's predictions are closer to the true values on average.

The training score is the performance of the model on the training data, and it is used to evaluate how well the model can fit the training data. A high training score of 1.0 indicates that the model is able to accurately capture the patterns in the training data, while a low training score of 0.0 indicates suggests that the model is not fitting the data well.

The test score is the performance of the model on the test data, and it is used to evaluate how well the model can generalize to new, unseen data. A high-test score of 1.0 indicates that the model is able to accurately make predictions on the test data, while a low-test score of 0.0 indicates that the model is not generalizing well to new data.

The R2 score, also known as the coefficient of determination, is a measure of how well the model fits the data. It is defined as the proportion of the variance in the target variable that is explained by the model. An R2 score of 1.0 indicates that the model perfectly fits the data, while an R2 score of 0.0 indicates that the model does not explain any of the variance in the target variable.

As you can see in Table 3, we preformed the evaluation test on our models before using the hyperparameter GridSearch CV. For Decision tree, Random Forest, linear regression and ridge regression we get perfect scores for all our tests meaning any of these models are perfect for our dataset because the MAE is zero, so they have no errors, MSE and RMSE is zero meaning the model predictions are the same as the true values and the R2 is 1 meaning the model predicts perfectly to the data. However this could all indicate to us that the model is over fitting but because our train and test scores are also 1 this tells us that it is not overfitting and that these models perfectly predict our data.

Similarly, gradient boosting is a good model for our data all be it not as good as our first 4 models. MAE, MSE, RMSE are very low, 0.04, 0.01 and 0.08 meaning there are very few error and it predicts are model very close to the true values. Our train, test and R2 score again confirm that there are no over fitting and that this model fits our data very accurately.

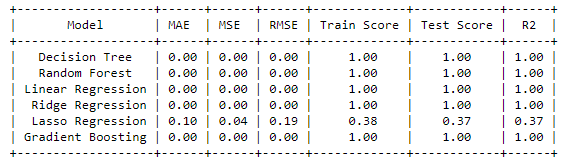
In contrast lasso regression and gradient boosting are not a good fit for our model. All of Lasso scores are very low. Its MAE score is 0.13 meaning it has very few errors, MSE and RMSE are 0.06 and 0.25 meaning model predictions are very similar to the true values which is what we are looking for but its R2, train and test scores are -0.00. Therefore, this indicates that the model does not explain any of the variance in the target variable and is not a good model for our data.

Table

Description automatically generated

##### Table 3: Results Table for Each Model Before GridSearch CV was Applied

In table 4, we preformed the evaluation test on our models after using the hyperparameter GridSearch CV. This makes no difference to the first four models but it does improve out last two. Gradient boosting has now been improved to the same standard as the first four and perfectly predicts out data. While lasso regression has improved but not to a significant enough degree to consider this model as a good fit for out data.



##### Table 4: Results Table for Each Model After GridSearch CV was Applied

## Deployment

### Deployment plan

A web application is created and hosted in a PaaS (Platform as a Service), in this case, the cloud platform service provided by Google, in order to be able to implement the artifact in the real world. It is essential to have access to real data provided by factories in order to use the program. The facters can see a prediction of the price of cattle carcasses.

### Monitor

The subsequent process is established as a monitoring and maintenance strategy. Storing and extracting data from the factories on a week-by-week basis (spreadsheet format). The gathered data should be stored in a cloud platform. Leveraging software as a service to automate processes. Re-evaluation of the factors used to forecast price of cattle carcasses on monthly basis. Monthly report including graphics, information, and deductions about cattle carcasses prices. To enhance the farmers experience, this report will be distributed to the necessary parties.

### Sentiment Analysis

In this project, we did sentiment analysis on grain prices using Twitter API. This involved collecting tweets related to grain prices from around the world and using natural language processing techniques to determine the overall sentiment of the tweets. This was done by training a machine learning model on a large dataset of labelled tweets, where the labels represent the sentiment of the tweet (positive, negative, or neutral). The model was then used to classify new tweets as they are collected from the Twitter API. By analysing the sentiment of tweets about grain prices in different regions of the world.

We found that 52.2% of tweets were positive, 16.7% were neutral and 31.1% of tweets were negative.

Text

Description automatically generated

#### Figure 9: Most popular words used in sentiment analysis on grain prices

### Dashboard

In this project, we created an interactive dashboard that has been developed for beef farmers. This dashboard provides an easy-to-use platform in which farmers can view and interact with data related to the price of cattle carcasses and cereal in Ireland and France. The goal of the dashboard is to provide farmers with the information they need to make informed decisions about their operations, and to give them the tools to analyse the data in a user-friendly environment. Whether you are a small farmer looking to optimize your profits or a large-scale operation seeking to stay competitive in the global market, this dashboard is designed to help you succeed.

We used bar plots for cattle and cereal prices. We choose to use these plots to compare the difference in prices depending on the year, week, country or category of cattle carcass. Using the interactive filters users can choose multiple variables of their choosing and customise their graph to need. This visualisation was the best way to show change in price over time and the distribution of the data.

We also created a map to visualise and create context to the data by showing the prices fluctuation once you press play on the widget. This allows the farmer to explore the data on their own, by zooming in and out, panning, and hovering over different countries to see more information. This makes the dashboard more engaging and effective way of presenting data than a static map.

Finally, we graphed the dataset in a table beside the map. We did this to allow the farmer to interact with the data if they want to see specific values in the dataset, rather than just a summary or visualization of the data. By including a table in a dashboard, we could allow users to interact with the data by filtering and sorting the table to focus on specific subsets of the data. This can be useful for exploring the dataset and finding patterns or trends. A table can be used in conjunction with the other visualizations, such as the bar charts, to provide additional context and detail for the data.

# References

*Database - Eurostat* (no date). Available at: https://ec.europa.eu/eurostat/web/main/data/database (Accessed: 5 January 2023).

*Developer Agreement and Policy – Twitter Developers* (no date). Available at: https://developer.twitter.com/en/developer-terms/agreement-and-policy (Accessed: 5 January 2023).

Development, E.C.-D.-G. for A. and R. (no date) *Agri-food data portal | Agricultural markets*. Available at: https://agridata.ec.europa.eu/extensions/DataPortal/agricultural\_markets.html (Accessed: 5 January 2023).

*FAOSTAT* (no date). Available at: https://www.fao.org/faostat/en/#data (Accessed: 5 January 2023).

*Legal notice* (no date). Available at: https://commission.europa.eu/legal-notice\_en (Accessed: 5 January 2023).

# Appendices

## Appendix A: CRISP-DM Framework

The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is a widely used method for structuring the process of data mining and machine learning. It consists of six main phases:

Business Understanding: This phase involves defining the problem or opportunity, understanding the context in which the problem exists, and identifying the goals and objectives for the project.

Data Understanding: In this phase, the data that will be used for the project is collected and explored. This includes understanding the structure of the data, identifying any issues or problems with the data, and generating some initial hypotheses about the relationships that may exist in the data.

Data Preparation: In this phase, the data is cleaned, transformed, and prepared for modeling. This may involve handling missing or incomplete data, normalizing data values, and creating new features or attributes.

Modeling: This phase involves selecting and applying appropriate machine learning algorithms to the data in order to build predictive models.

Evaluation: In this phase, the performance of the model is assessed and evaluated using appropriate metrics and methods.

Deployment: This phase involves implementing the model in a production environment and monitoring its performance over time.

The CRISP-DM framework provides a structured approach for organizing the data mining process and helps ensure that all necessary steps are taken in order to build accurate and effective predictive models.

## Appendix B: GitHub Link

https://github.com/sba22222

Appendix C: European Commission License/Permission  
“Unless otherwise indicated (e.g. in individual copyright notices), content owned by the EU on this website is licensed under the [license](http://creativecommons.org/licenses/by/4.0/). This means that reuse is allowed, provided appropriate credit is given and changes are indicated.”(*Legal notice*, no date)

## Appendix: Word Count

3,300 number of words without taking code, code comments, titles, references or citations into account