**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:** | Stephen Somerville |
| **Student Number:** | Sba22232 |
| **Assessment Due Date:** | 06/01/2023 |
| **Date of Submission:** | 11/01/2023 |

**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. |

**Introduction**

In this report, I will analyse the Irish agricultural sector and compare it with other countries in Europe, with a focus on key metrics such as production, exports, imports, and trade balance. The analysis will be focused on dairy products across current EU members in order to ensure consistency in reporting, standards, and practices.

The goal of this report is to provide insights and recommendations to stakeholders in the Irish agricultural sector, including farmers, policymakers, and consumers. To achieve this, we will follow the CRISP-DM methodology and cover the six major steps of the process: business understanding, data understanding, data preparation, modelling, evaluation, and deployment.

Throughout the report, we will draw on data and analysis techniques to understand the current state of the Irish agricultural sector and identify trends and opportunities for improvement. By demonstrating the tools, steps, and processes that we would follow for actual Agri stakeholders, we aim to provide a comprehensive and actionable report that will inform decision-making and drive the growth and sustainability of the Irish agricultural sector.

Section 1 - Business understanding

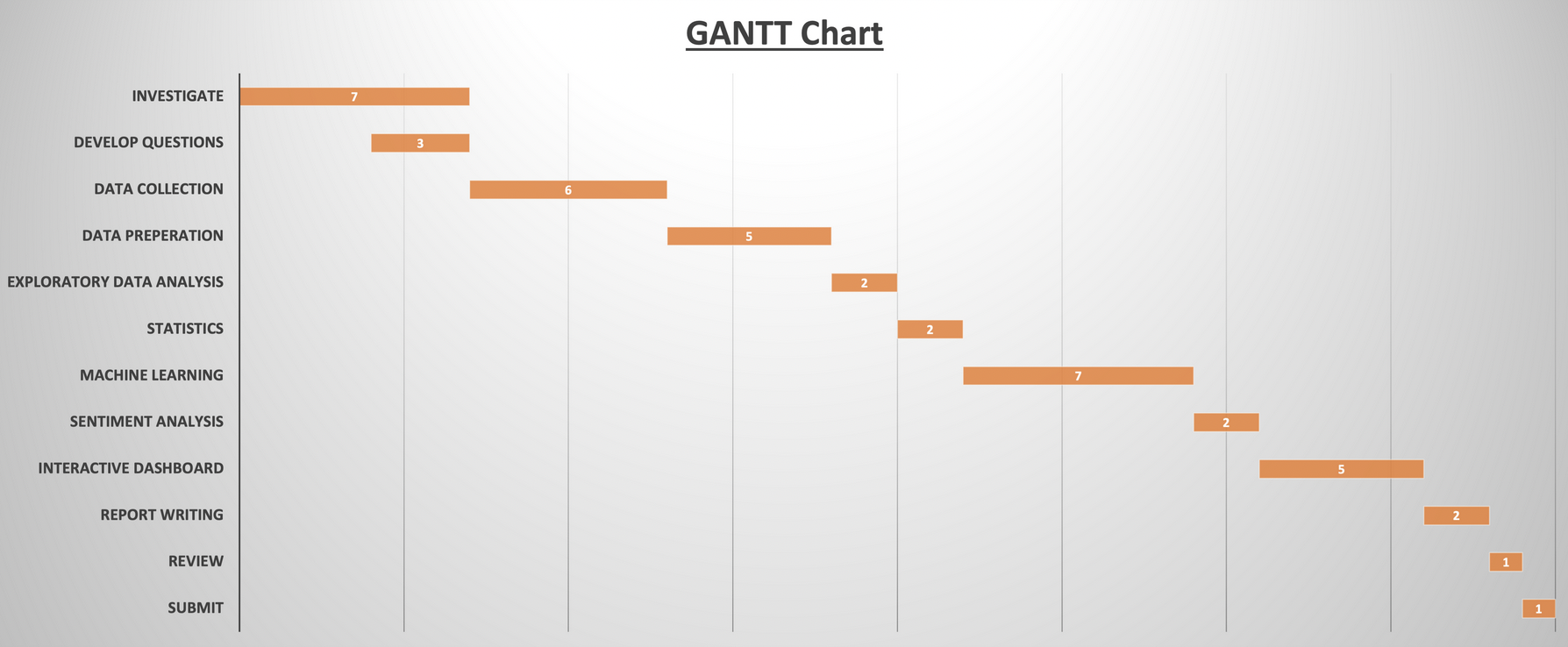
Typically in this phase, we would identify the stakeholders and their needs. This could include farmers, policymakers, consumers, and other groups that are affected by or have an interest in the Irish agricultural sector. I would describe their interests and goals of each stakeholder group and how they relate to the business problem or opportunity. However this time I will not be completing this aspect, as it is outside the needs of the assignment.

As I have limited experience in the Agri Sector, my recommendations will be limited and therefore I will presenting the report as an analyse of the facts available to me. This will add certain constraints such as finding available, appropriate and actionable data plus time constraints as this will be completed outside of my normal working week. I will be completing research throughout the course of the assignment to understand the sector while also research additional techniques I can use to collate, assess and present the data at the end of the report. The combination of these factors could play a factor and may impact my analysis as a result.

As part of the Crisp DM method, a SMART goal or number of them are created as this stage to guide the progress of the project or in this case assignment. My SMART goal is as follows -

* Specific: Analyse the Irish agricultural sector and compare it with other countries worldwide, with a focus on key metrics such as production, exports, imports, and trade balance.
* Measurable: Collect and analyse data on the Irish agricultural sector and other countries, using statistical and visualization tools to identify trends and patterns.
* Achievable: Use data analysis and forecasting techniques to identify opportunities for improvement and make evidence-based recommendations for the sector's development.
* Relevant: Focus on dairy products across current EU members to ensure consistency in reporting, standards, and practices.
* Time-bound: Complete the analysis and report within the allotted time frame, with a deadline for submitting the final report.

As part of my SMART goal I have created the following GANTT chart to monitor my progress over time.



Section 2 - Data Understanding

The process of acquiring raw data for this assignment involved identifying the data I needed, finding a suitable source, and obtaining permission to access and use the data. There were many potential sources of data available online, but navigating the various formats and features offered by these sites, as well as the licensing and permission requirements, proved to be a challenging task. Given this, I decided to approach the main government agency websites for the agriculture industry, such as [www.fao.org](http://www.fao.org) and [www.ifa.ie](http://www.ifa.ie), as these are generally reliable sources of data and typically have clear guidelines for obtaining permission to use their data.

After evaluating several potential sources of data, I ultimately decided to use the data from [www.fao.org](http://www.fao.org) for this project. This website is maintained by the Food and Agriculture Organization (FAO) of the United Nations, and the data is collected and processed according to the standard International Merchandise Trade Statistics (IMTS) Methodology. The data is primarily provided by the United Nations Statistics Division (UNSD) and Eurostat, as well as other national authorities as needed. While the data and content on the FAO website are protected by copyright, the organization is committed to making its content freely available and encourages the use, reproduction, and dissemination of its text, multimedia, and data. The database I used is covered by the Open Data Licensing Policy and governed by the Statistical Databases Terms of Use.

Section 3 – Data Preparation

In the data preparation phase of the CRISP-DM process, the focus is on cleaning, transforming, and organizing the data in a way that makes it suitable for analysis. This can involve tasks such as filling in missing values, handling outliers, encoding categorical variables, and creating derived variables.

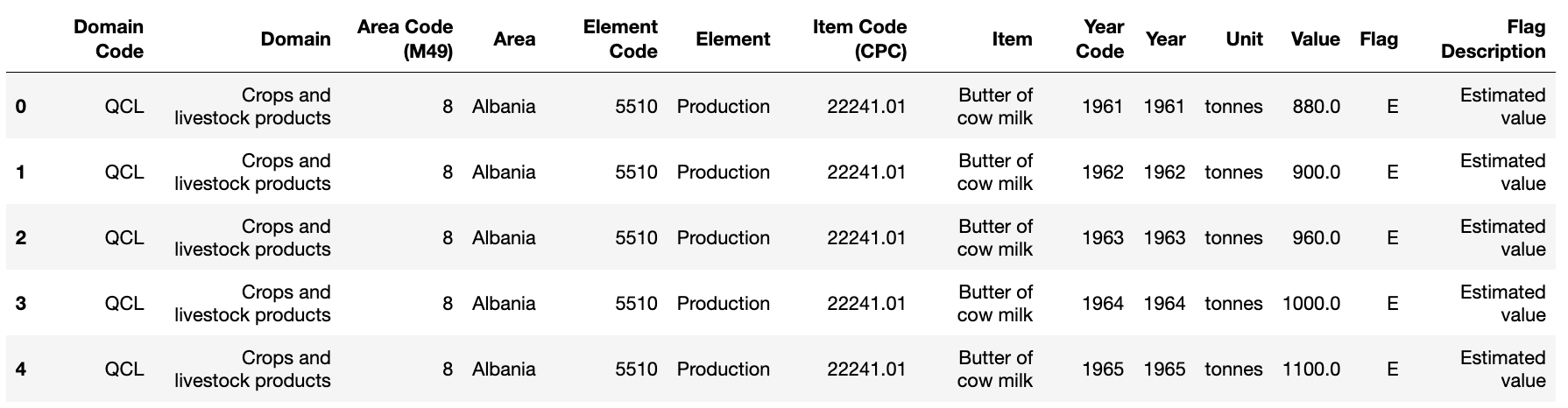
Here is the process and justification for my data preparation I followed during this phase in the assignment:

One of the strangest aspects to the Data is how iterative Exploratory Data Analysis (EDA) is and how much different functions need to be run several times ensure my data was changing and improving.

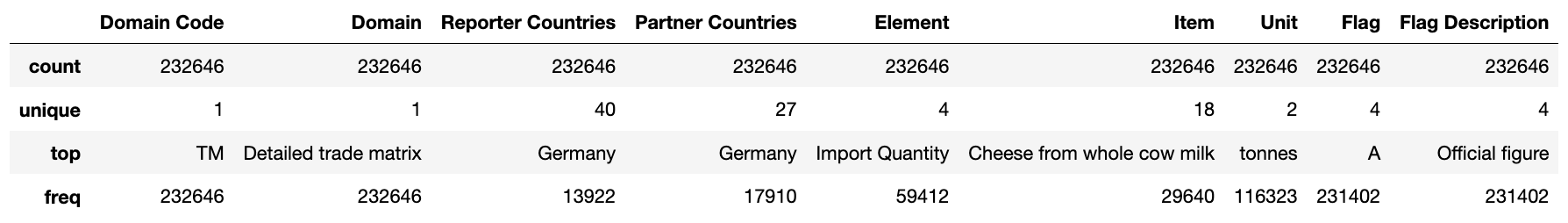
My overall structure is simple, I followed 3 steps that are referenced in my notebook –

Structural investigation: Here I explore the general shape of a dataset and the data types of its features. I examine the number of rows and columns in the dataset, identifying any missing or duplicate data, checking for outliers and other unusual values, and evaluate the distribution of values for different features. By completing these tasks, I gain a deeper understanding of the dataset, and was able identify any potential issues or challenges that may need to be addressed in the subsequent stages of the analysis.

Example of the top 5 rows in the Dataset –



Non numeric data –

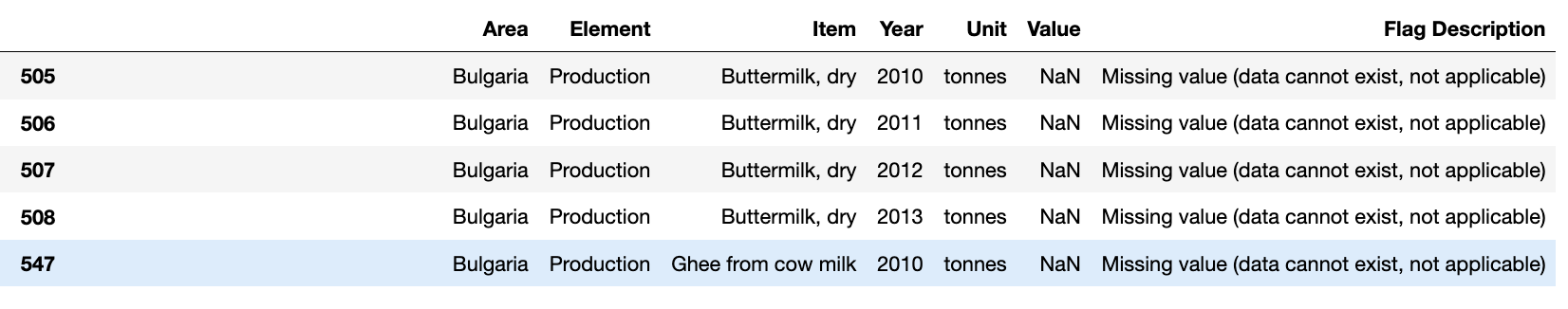


I also reviewed the data type so if any changes were needed at this stage I could process them.

Qualitative investigation: Understand the general quality of the dataset, with regards to duplicates, unwanted entries and missing values.

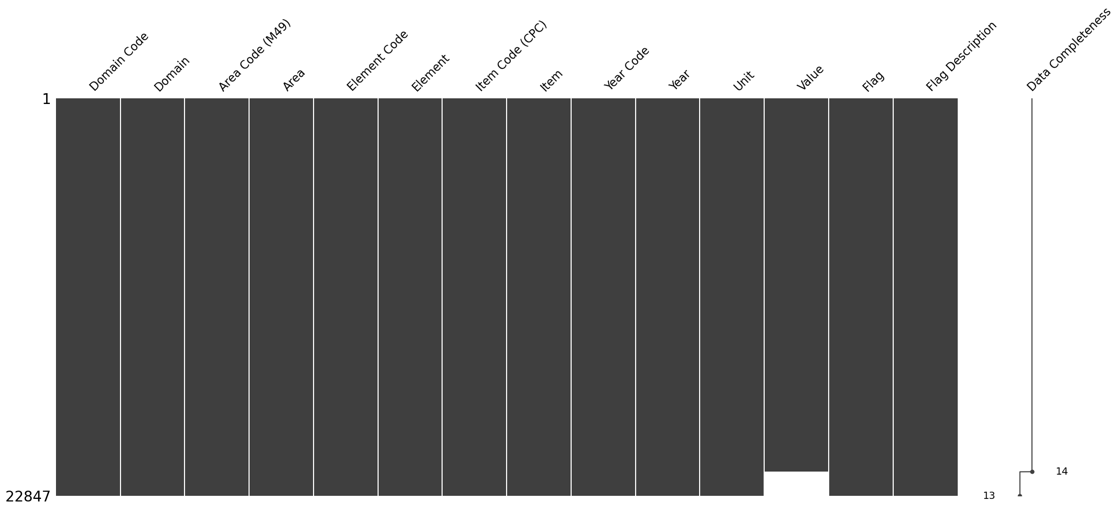
Here I checked for –

* Duplicates – Data didn’t contain any.
* Unique Values – there was mix across the data and contained similar amounts of unique values. This prompted me to review and remove columns where the unique value were duplicates or provided similar information in another column, e.g. Area and Area Code (M49) – both give data on the country.
* Unwanted entries – I found examples of NaN or Null values in the data where the country didn’t have a record of the product or value for a given year or product –



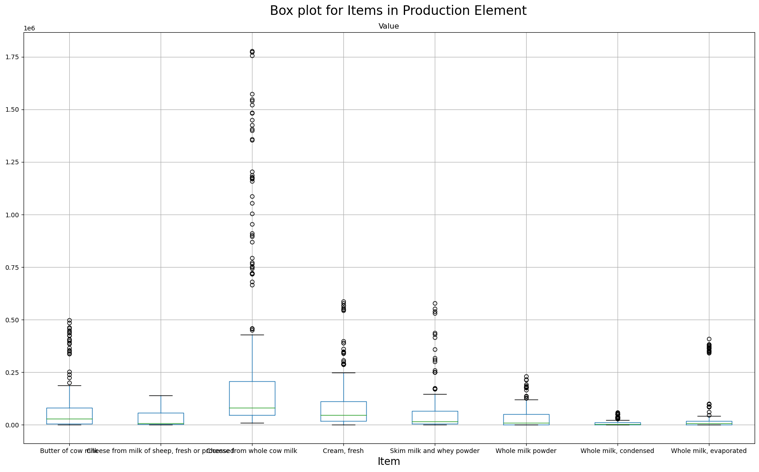
I also removed products and countries from the data that didn’t fit with the dairy theme e.g. Raw Silk.

* Missing values – I used missingno to visual the missing values data and the volume of missing data. Example –

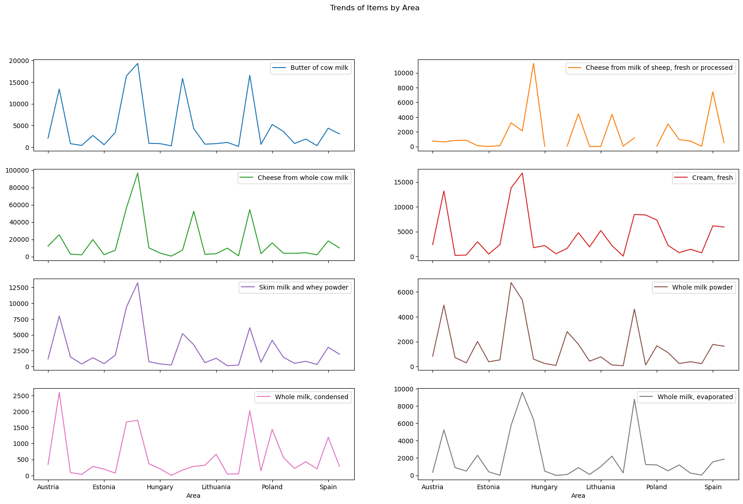


Content investigation: Once the structure and quality of the dataset is understood, we can go ahead and perform a more in-depth exploration on the features values and look at how different features relate to each other. This involved looking at the distribution of values for individual features, as well as examining the relationships between different features. I used different techniques to perform this exploration, such as visualizing the data using plots and charts, calculating summary statistics, and fitting models to the data. This gave me further insights into the data, and helped identify patterns that were useful when building a machine learning model.

Summary Box plot of the product element of the products –



Trends of the Products by Country



With data cleaned and new columns added to give extra features e.g. “Origin” to split Ireland from the rest of EU members on the list. I split the data into two tables – One for the productive elements were the values were in Tonnes and a second one for the Dollar value entries.

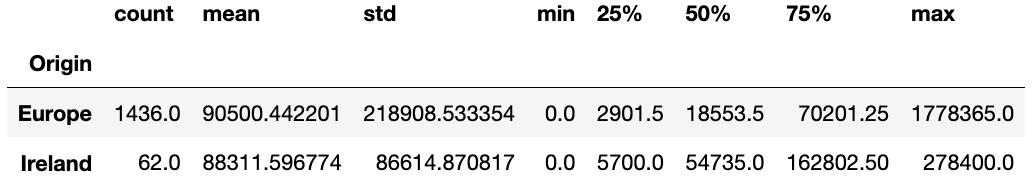
Another aspect to the content investigation is it was clear how much of data was non numerical so I had to use One hot encoding and GridSearchCV to select and handle the features I wanted to use in my Machine Learning models later.

During the data preparation process, I faced some challenges with the dataset I had chosen. One of the main challenges was the format of the data, it required a significant amount of work to make it usable. The dataset was heavily non-numeric which made it difficult to use many of the tests and techniques I had planned to use. Despite these challenges, I learned a lot about how to handle non-numeric data and gain insights from it. For example, I wanted to create a dataframe using the pivot\_table method to create more numeric columns and add context to the data. However, I wasn't able to fully implement this because of the limitations imposed by knowledge and time constraints. This is was classic example of the trade-off between time, effort and the desired result.

In summary, despite facing some challenges in working with the dataset, I learned a lot about how to handle non-numeric data and gain insights from it, but was not able to fully implement all the techniques I had learned due to time constraints.

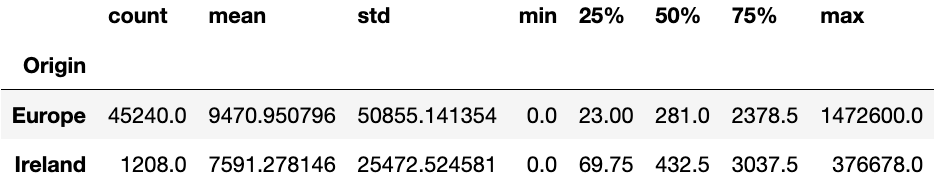
I am happy with the final datasets I will be using for further analysis, they are made up of 8 total columns – 2 numerical and 6 Non numerical. The transaction dataset is significant larger as the monetary set. Both are structured as follows –

Product –





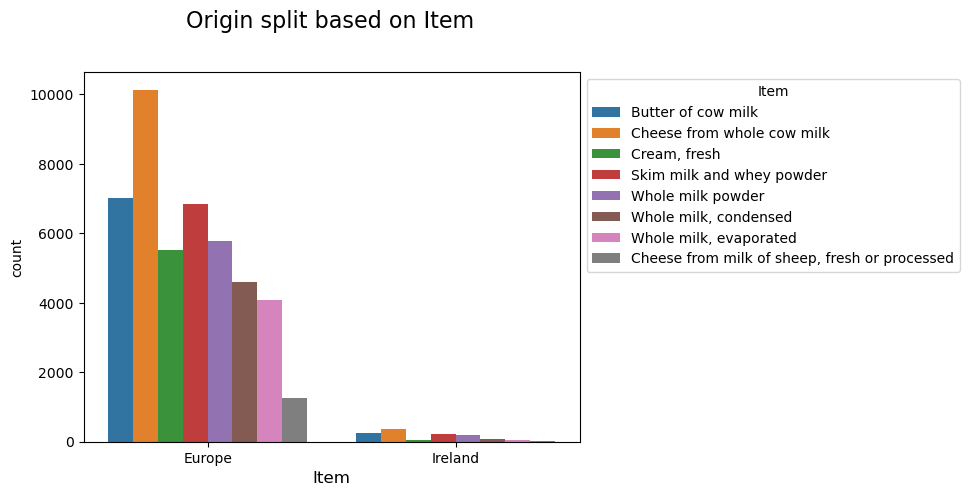
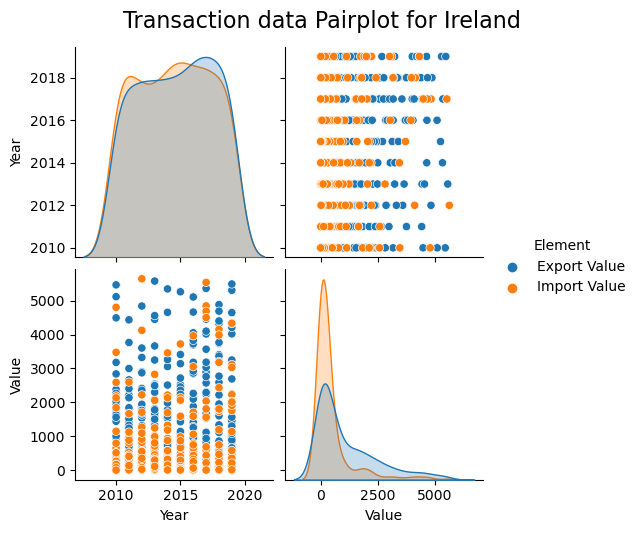
Transaction –





Some interesting points I found during my descriptive analysis – Ireland is ranked 7th and placed in the top 25% of the EU members in the dataset, based on

* mean value of import and export transaction and also the mean value of production volumes across the catalogue.
* Stand out products – All EU and Ireland was Cheese from whole cow milk is highest with butter from Cows Milk second.



Statistical models –

* t-test: A t-test is used to compare the means of two groups. It can be used to determine whether the means of two groups are significantly different from each other.
* ANOVA: Analysis of variance (ANOVA) is used to compare the means of three or more groups. It can be used to determine whether the means of the groups are significantly different from each other.
* Mann-Whitney U test: This is a non-parametric test used to compare two independent samples of data. It's used as an alternative to the independent t-test when the data is not normally distributed
* Chi-square test: A chi-square test is used to compare observed data to expected data. It is often used to determine whether there is a significant relationship between two categorical variables.
* Linear regression: Linear regression is used to predict the value of a continuous outcome variable based on the value of one or more predictor variables. It can be used to determine the strength and direction of the relationship between the predictor and outcome variables.

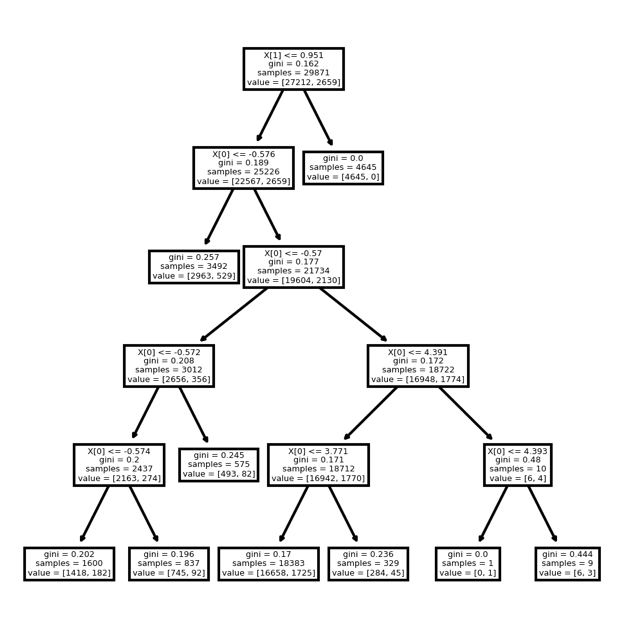
Section 4 - Modelling

For my Machine Learning models I chose to use - Decision Tree & KNN Nearest neighbour.

After trying two different feature engineering methods - GridsearchCV and One Hot Encoding on my data I was able to obtain following scores using the One Hot Encoding method –

* Decision tree = Accuracy: 0.9120305282185178
* KNN Nearest neighbour = Accuracy: 0.9752927551151718

Decision Trees because it is easy to understand and interpret (for both me and the stakeholders), they can handle both categorical and numerical data, and they can be used for both classification and regression tasks. In a decision tree, the data is split at each internal node based on a certain feature, with the goal of minimizing the impurity of the resulting groups of data. The final output is a tree-like model, where each leaf node represents a class label.



KNN, on the other hand, is a non-parametric algorithm, which means it does not make any assumptions about the underlying distribution of the data. It's a lazy-learning algorithm as it doesn't learn any model, but it only memorizes the training data. It works by finding the k-nearest training examples for a new data point, and classifying the new point based on the majority class among its k-nearest neighbours. It's simple to implement and can work well for small datasets.

I used the second neighbour as the parameter.

