The increasing Cost of Butter: A comparison study between Ireland and France

Author: M. Moran

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

Student ID: SBA22220

Continuous assignment No.2 submitted as part of Module 1 assessment for the Degree of Master of Science in Data Analytics

List of Contents

List of Tables

List of Figures

Abstract

1: Introduction

2 Project Management Framework

3.Data Understanding

* 3.1 Collect Initial Data
* 3.2 Describe Data
* 3.3 Explore Data
* 3.4Verify Data Quality

4. Data Preparation Phase

* 4.1 Integrate Data
* 4.2 Select Data – Inclusion/Exclusion
* 4.3 Construct Data/Format Data
* 4.4 Integrate Data
* 4.5 Data Visualisation
* 4.6 Outlier Determination and Removal

5.Statistics

* 5.1 Descriptive Statistics
* 5.2 Central Tendency: Mean/Median/Mode
* 5.3 Testing for Normal Distribution

6- Data Pre-processing - Using only Irish Butter Prices data

##### 6.1 Comparison of Standardization and Normalization of Data

##### 6.2 The Test and Train Split[¶](http://localhost:8888/lab#4.1-Comparison-of-Standardization-and-Normalization-of-Data-,-which-to-use-?)

7 Machine Learning Algorithm

* 7.1 SciKit-Learn Algorithm – Linear Regression (Supervised)
* 7.2 SciKit-Learn Algorithm – KNN (Supervised)
* 7.3 SciKit-Learn Algorithm – RandomForestClassifier
* 7.4 Sentiment Analysis
* 7.5 Dashboard Creation

8 Conclusions

9 References

10. Appendices

**List of Tables**

Table 1 Figure 1 Crisp- DM in detail

Table 2 Summary of Inferential Statistics Tests reviewed

Table 3 Table 3 – Summary Machine learning model results

**List of Figures**

Figure 1: Crisp-DM Lifecycle

Figure 2: EuropeanButterPrices. top 5 rows – columns highlighted

Figure 3: dtypes info

Figure 4: Algorithm Results

Abstract [1,2]

*A data analysis of the Agricultural Topic of Butter and its prices for 2015 – 2022 was chosen. From Article in the Irish Times in May 2022, Conor Pope wrote about the rising cost of butter in Ireland, that in the last 6 months, the price of butter has increased 36% which is 6 times the inflation. I will be looking at these prices but over a few years. All the butter on the Irish market is produced locally and comes from Irish cows’ milk and this has always been the case.*

*The original dataset which is published by Eurostat. Eurostat is the statistical office of the European union. They provide high-quality statistics and data on Europe. These statistics and data are gathered from EU Member states through their own National Statistical Institutes, so in the case of Ireland, our institute is the Central Statistics Office (CSO) and further agricultural data from the government department of agricultural.*

*This "European Butter Prices 10 yearsCodes.csv file contains Butter prices (euro/100kg) for each of the European Member states that contributed data to the Eurostat’s ESS (European Statistical System). A chloropleth visualisation was created and it showed the member states that have the highest butter prices alongside Ireland. I chose France as the state for comparison.*

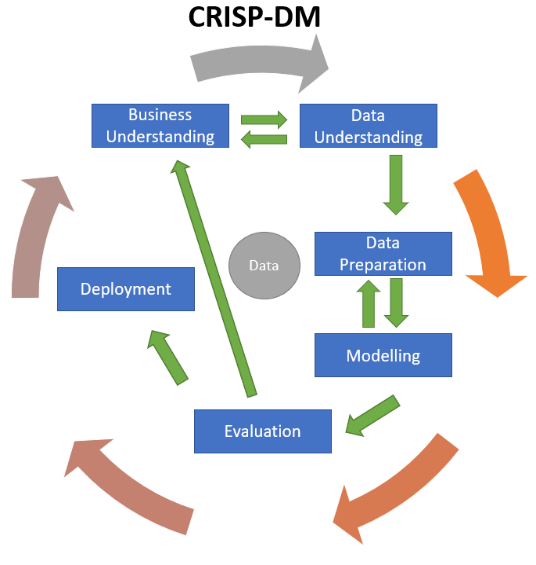
*This data was collating, processed, analysed, and interpreted to show an analysis of the butter prices over time in Ireland and compare to France*

## Introduction [1,2,3]

* The purpose of this assignment is to choose an Agri sector topic to analyse Irelands Agricultural data and use this data to compare to another Country.
* Chosen Topic: Analysis of the Price of Butter in Ireland and France

## 2. Project Management Framework [4,5]

Methodology used: **Crisp-DM** with programming language of **Python**



**Figure 1: Crisp-DM Lifecycle [3]**

Crisp-DM is the cross-industry standard process for data mining and is an open standard process framework model. See Figure 1 for details

|  |  |
| --- | --- |
| **Crisp-DM Phases** | **Project: Analysis of Butter prices in Ireland** |
| 1. Business/Research Understanding Phase – Changing it to Project Understanding | * Project Definition – Chose an Agricultural Topic - Butter * Explore the European and Irish Dataset– through Eurostat and CSO |
| 1. Data Understanding Phase | * Collect Data – created a CSV file of Ireland and Frances butter prices for the years 2015-2022 * Created two separate dataframes for each country. Completed Exploratory analysis on the Irish dataset |
| 1. Data Preparation Phase | * The data was cleaned i.e checked for duplicate values, outliers, null values * No data transformation was required * Split the data into train and test sets for modelling |
| 1. Modeling Phase | * Choose Machine Modelling Techniques |
| 1. Evaluation Phase | * Evaluate the machine learning through RSME |
| 1. Deployment Phase | * Completion of the assignment for CCT |

**Table 1 Crisp- DM in detail**

1. **Data Understanding – Collecting and exploring the input dataset**

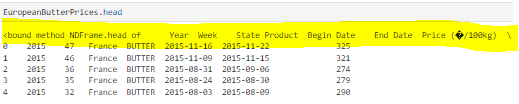
***Reference Jupyter Notebook – Data Prep and Vis.ipynb***

* **3.1 Collect Initial Data [1, 6,7,8,9,10,11,12,13]**

Initial Data was taken from the following Eurostat Data Table: Years 2015- 2022, Countries: Ireland and France and butter as the food.

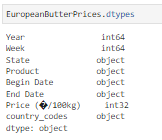
[Dairy prices (europa.eu)](https://agridata.ec.europa.eu/extensions/DashboardDairy/DairyPrices.html)

This Data was saved as a CSV File called - "European Butter Prices 2015-2022.csv" and imported into the notebook as *EuropeanButterPrices, fig 2*



* **Figure 2 EuropeanButterPrices. top 5 rows – columns highlighted**
* **3.2 Describe Data**
* The data was examined under by dtype**:** 
  + **Our target variable:** Price (�/100kg)- described as a float64

This was changed to an integer to ensure efficiency of future calculations

* + 
* **Figure 3 dtype info**
* **3.3 Explore Data – (Section 2.2 in notebook)**
  + Checking for duplication - No Duplicated values noted
  + Checking for Null values – No null (Nan) values
  + Unique values checked
  + Check for 0 values – none
* **3.4 Verify Data Quality**

This dataset is complete with no duplicated results and no missing data. It has been correctly copied from the Eurostat and saved as a csv file. This has correctly been imported to a notebook in jupyter

**4 Data Preparation Phase [6,7,8,9,10,11,12,13]**

**4.1 Integrate Data**

### Splitting the data frame into 2 data frames for countries Ireland and France – FrenchButterPrices2 and IrishButterPrices2

**NB: The following was completed on both data frames**

**4.2 Select Data – Inclusion/Exclusion**

**Columns:** State and Product were deemed as irrelevant and dropped

**4.3 Construct Data/Format Data**

Begin Date column was of type Object – and to pull the month details out we converted to pandas.datetime

Begin date column was set as index and the prices were grouped per monthly and sum

**4.4 Integrate Data: n/a**

* 1. **Data Visualisation–** Notebook reference: 3.0
     + **Implots** – show multiple linear regression.
       - Looking at years 2015-2022 Irish and France Butter price data monthly
       - Graphical representation too dense – too much data, not clear
     + **-Heatmap** 
       - There is poor correlation between the month and Price (�/100kg). There is neither a strong negative nor positive correlation
       - For Years and Price (�/100kg data – the correlation is 0.49 and 0.65 for Ireland and France respectively which is not particularly close to 1 but perhaps as the years increase, so might the price of butter in Ireland based the previous year’s data
     + **Pairplot** 
       - There is a right skewed (distribution Symmetry) histogram plot for the Price (�/100kg) but no other pair relationships apparent. It seems that most of the price data is shown on the left side with decreasing data on the right.
     + **Histogram Plot of all year’s data - frequency distribution of prices**
       - This shows that prices in Ireland were on the frequency for years 2017 and 2018 but seemed to dip for the following 2 years (possibly due to covid) and returned to high prices for 2022.
       - For French data – the histogram shows a more normal frequency distribution
     + **Boxplots – Visual method for Outlier’s determination**
       - These plots show the distribution of the data points by dividing them into different quartiles – lower, median and upper quartiles in the form of a box and whiskers. Any data points outside these are deemed as an outlier/anomaly.
       - For Ireland data - years 2019 and 2022, have outliers and France has for 2016, 2018, 2020 and 2022
     + **Outlier detection - The isolation forest algorithm**
       - This is another method using an algorithm called isolation forest. It is part of the ensemble model class and returns the anomaly score of each instance to measure abnormality
       - It gave details of the outliers present in both countries dataframes**.**

**4.6 Outlier Determination and Removal [19,20]–** Notebook reference: 3.3

To detect outliers, Tukey’s rule also known as the IQR rule. It involves determining boundaries for outliers.

The upper and lower quartiles are determined using the describe panda’s function. describe. IQR is the upper boundary less the lower boundary

The Begin date - timeline columns were removed as the IQR calculation wouldn’t work on the datatime dtype of the date columns.

The outliers that are more than or less than the upper and lower quartiles calculated for all butter prices were removed using the Tukey equation and the isolation algorithm (and the. Df.shape code) was used to check if all the outliers were removed. It was successful for the Ireland data but not for the France data. The outliers remained in the dataset. I chose to leave this as is, as the machine learning modelling with be conducted on the Irish data only. The French dataset will only be used for statistical comparison

1. **Statistics [16,17]**

Reference: Statistic.ipynb Notebook

**Relevant libraries and the dataframes were imported:**

* **French\_prices\_df**
* **Irish\_prices\_df**
* **European\_prices\_df**

**5.1 Descriptive Statistics**

NB: For each of the individual country’s dataset - The Years column data was transposed using the month column as index, so we could complete the descriptive statistic analysis on the price data per year for each country dataset

**5.2 Central Tendency: Mean/Median/Mode**

NB : For years 2015 and 2016 , there are a number missing values ( 0) for the French Butter Prices data frame , so in the interest of a fair comparison , we will just at years 2017 – 2020 for both Countries

The Mean value of butter prices /100kg is higher for each of the Years in Ireland compared to France. The changes in prices over these years are similar in both countries. Again in 2019 and 2020, there is a decrease in the butter prices, but it increases after this.

The Median values in comparison to the mean value for both countries over the years are very similar indicating that there is very little variation between the datasets monthly results.

As the mean and median are similar for both countries, it can be said that there is a rough symmetric distribution and no appearance of skewness

With regards the mode – there are no values the same in both countries’ dataset

For the standard deviation – its consistent for both countries butter prices over the years, indicating that there is very little variation between prices over the months and the overall average.

The variance was calculated on the data frames as a whole, there a bigger variance in the French butter prices compared to the Irish however it took into account years that have no data entry.

**5.3 Testing for Normal Distribution**

* **Q Q or Quantile-Quantile Plot visually check for a normal distribution:**

Both countries data look normal with an even distribution of data along a straight line

### **Testing for an Inferential Statistics in the European\_prices\_df Data frame: Ireland and France combined**

**Hypotheses -to reject or accept - Inferential statistics**

1. The null hypothesis (H0) that the butter prices for both Ireland and France are the same.
2. The alternative hypothesis that the butter prices for both Ireland and France are not the same

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Inferential Statistics** | **Country** | **F stat value** | **P-Value** | **Reject/ Accept H0** |
| **Shapiro Wilks Test** | Ireland | 0.8985 | 5.7919 | Accept - normality observed |
|  | France | 0.9167 | 1.636 | Accept - normality observed |
| **Levene Test** |  | 1.688 | 0.194 | Accept – variation observed |
| **ANOVA One -Way** (parametric) |  | 35.33819 | 4.268814e-09 | Accept |
| **ANOVA two -Way**  (parametric) | State | 35.715780 | 3.550049e-09 | Accept |
|  | Month | 8.906915 | 2.934610e-03 | Accept |
| **T-Test, one populations** | #H0 : u = 500  #H1 : u =!500 | 21.295 | 2.513 x e-37 | Accept the Ho that the average irish butter prices are not 500 euro /100kg |
| **T-Test, two populations** | H0 that mu1 = mu2 (France prices = Ireland Prices) | -5.944 | 4.268x e-09 | Its says accept but we know this should be reject |
|  | H0 that mu1 not equal mu2 (France prices not equal Ireland Prices) | -5.912 | 5.310 x e-09 | Accept |
| **Kruskal-Wallis H-test** ( NON Parametric) | n/a | 47.227 | 6.321 | Reject Ho |
| **Friedman chi square** ( NON Parametric) | Unequal sample size – need 3 variables | | | |
| **Wilcoxon Test** ( NON Parametric) | Cannot use this test as samples must be equal in length | | | |

**Table 2 – Summary of Inferential Statistics Tests reviewed**

The conclusion that can be drawn from this testing is that the mean and median values of both countries are not equal. Therefore, a change in prices in both countries are not relative.

Anova and t-test statistical testing was conducted as the data were deemed to be normal from the Shapiro and QQ plotting

The other tests reviewed are based on the median.

A learning from this statistical study, is another variable and perhaps more data would be useful in giving a fuller statistical picture. Looking at the effect of other variables on the prices of butter such as milk prices and production stock would give more information.

I do believe that the sampling was random and unbiased. Also, that the data is normally distribution and met the requirements for the parametric testing.

***6- Data Preprocessing [21] - Using only Irish Butter Prices data***

##### **6.1 Comparison of Standardization and Normalization of Data :?**[**¶**](http://localhost:8888/lab#4.1-Comparison-of-Standardization-and-Normalization-of-Data-,-which-to-use-?)

The pandas describe function shows us the min and max range of footfall results for each of the years butter prices.

We determined the results are of normal distribution. I conducted the linear regression using both scaled data and unscaled data, there was very little impact with the scaled data. So decided to not use it.

##### **6.2 The Test and Train Split [22,23]**

Split the data into training and test sets in Python using scikit-learn’s built-in train\_test\_split ():

Test size of 0.2 and the random state selected as 0 to ensure same results each time the code is ran

**7.0 Machine Learning Algorithm**

Reference notebook – Machine Learning.ipynb

**Aim: To predict the Irish butter price based on previous years price trends**

I concluded that there a small correlation between Irish butter prices and Year, therefore one can’t expect a straightforward modelling process.

**Measurement of Accuracy and evaluation of models**

RMSE as a measure of the average error and to evaluate trained models for usefulness / accuracy of the predicted footfall for both algorithms

* 1. **SciKit-Learn Algorithm – Linear Regression (Supervised) [.26.27,28]**

The following assumptions were checked prior to linear-regression analysis for prediction of Irish Butter Prices:

1. The Irish Butter Price variable is measured at a continuous level.
2. In EDA, results were plotted and a linear relationship between those two variables was observed
3. The observations are independent of each other (that is, there should be no dependency).
4. Your data should have no significant outliers. – removed
5. Check for homoscedasticity — a statistical concept in which the variances along the best-fit linear-regression line remain similar all through that line.
6. The residuals (errors) of the best-fit regression line follow normal distribution determined after regression

**7.1.1 Results: RSME and R2**

The RSME results for both the training and test data were similar, the test data was only slightly highly, this indicating that the model is evaluated as good for making predictions. There was no indicating of overfitting or underfitting

Other linear models were reviewed and compared: No one model stands out. The R2 is low, indicating variability in data.

Cross validation was analysis but the R2 decreased with the increased in subplot data.

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **RSME Train** | **RSME Test** | **R2** |
| Linear Regression | 308.23468 | 302.54281 | 0.49857 |
| Ridge | 312.2170 | 306.38525 | 0.4855 |
| Lasso | 308.2894 | 301.46399 | 0.4983 |
| BayessianRidge | 308.5977 | 303.2477 | 0.49739 |

**Table 3 – Summary Machine learning model result**

* 1. **SciKit-Learn Algorithm – KNN (Supervised) [24,25]**

**Reasons to use this Algorithm**:

* Simple to implement and intuitive to understand
* Can learn non-linear decision boundaries when used for classification and regression.
* The KNN algorithm has no explicit training step, and all the work happens during prediction
* adding new data to the dataset, the prediction is adjusted without having to retrain a new model.
* Single Hyperparameters:

The goal is to predict the price of an Irish Butter based on previous years pricing data. This is a regression problem because the output variable (butter prices) is continuous.

Predicting a result using 4 nearest neighbours gives us RSME of 209.944, and the test RSME is 251.17 which is higher than the training data indicating underfitting of data. To incorporate a hyperparameter of bagging to improve the result, the testing result increased to 263,39, which is worst. These results indicate that KNN is not the model for the prediction of butter prices, however neither is using previous years prices to predict future pricing.

* 1. **SciKit-Learn Algorithm – RandomForestClassifier [30]**

**Reasons to use this Algorithm:**

The random forest algorithm is more accurate than other non-linear classifiers.

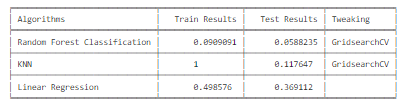
It is robust because it uses multiple decision trees to arrive at its result and overfitting isn’t a issue

You can use this algorithm for both regression and classification problems,

Random forests don’t let missing values cause an issue. They can use median values to replace the continuous variables or calculate the proximity-weighted average of the missing values to solve this problem.

I used my training and testing data in this model and predicted the accuracy. The result was less than 10 % accuracy for both sets.

**Overall Machine Learning Accuracy Results**



* **Figure 4 Algorithm Results**
  1. **Sentiment Analysis [31,32,33]**

**Reference notebook – Sentiment Analysis Twitter Folder**

**Notebook: SentimentAnalysisTwitter.ipynb**

I performed Sentiment analysis using Twitter API. I created my own twitter account and thus gained access to API tokens

I completed this analysis using personal laptop due to issues downloading the libraries require to run sentiment analysis. Eventually I was able to transfer all to this to the current work laptop which I am using for this module.

Unfortunately, when running my notebook referenced above , I came across the following error which was not present previously. All code has been checked.



All files in the sentiment analysis folder have been working previously – I used my own Twitter access codes and along with example code from lectures, I imputed the phrase: Cost of Butter and asked it checked 100 times in Twitter to see what trending. I completed some data cleaning and processing in the notebook and parsed the file to an csv file (reference: response\_python.csv. Here it divided up the sentiments into neg, neu and pos.

The information from the tweets were very varied and not suitable to use for any possible modelling

* 1. **Dashboard Creation [34]**

**Reference Folder – Butter Prices Dashboard Folder**

**Notebook: DashboardTrial 1 and DashboardTrail 2. ipynb**

**Aim:** To create interactive plots with Jupyter Notebook, turn them into a standalone dashboard, and then deploy it on the cloud with Heroku so that other people can see your dashboard.!

But before the cloud, to access it locally using Voila in jupyter notebook

I tried using PANEL and PLOTLY, to create an interactive visualisation. See reference notebooks.

However, I ran into problems when selecting voila – a JavaScript error, please see attached screenshot of error message in appendices section. Also, with regards Panel version – couldn’t create a virtual environment.

So, the creation of the interactive plots and tables with widgets was completed but unfortunately unable to display on a dashboard.

I used the full European dataset for the last 10 years to give a full picture of butter prices in Europe. Results from the actual machine learning were poor and would have been of no value to any possible interested party from the agricultural sector.

## 8.0 Conclusions

The main aim of the work was to create a model that could predict prices of butter in Ireland. I tested three models using the [Scikit Learn](http://scikit-learn.org/) python library to see which was the best. Initially through the EDA analysis, I found there was poor correlation between the variables, concluding that a machine learning model analysis wouldn’t be easy. I looked at prediction through linear regression.

In conclusion, more variables such as external conditions i.e., milk production, fertiliser prices etc, need to be analysed as well to give a better picture of the effect on butter prices

Further work on getting the dashboard up and running using the European data, solving the software issues

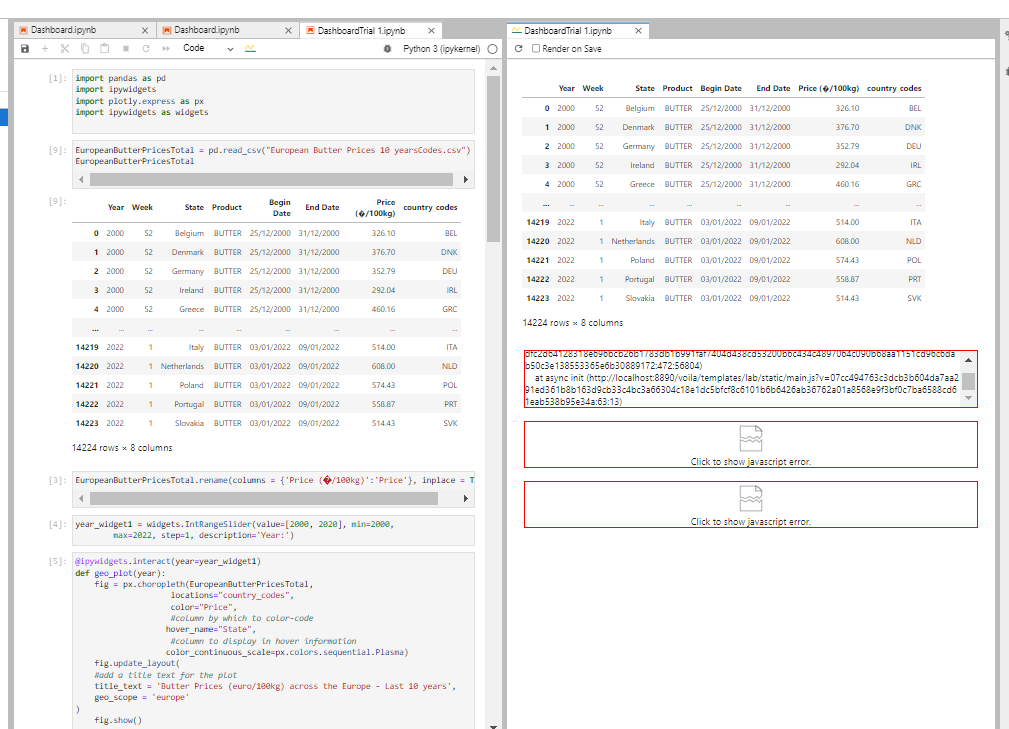
Also, with the sentiment analysis – looking at other API instead of Twitter would be more beneficial to the data gathering process.

## 9.0 References

1. Eurostat background information -[European Statistical System (ESS) - European Statistical System (ESS) - Eurostat (europa.eu)](https://onedrive.coca-cola.com/personal/michelle_moran_coca-cola_com/Documents/Masters%202022/Assignment%202%202022/Assignment/European%20Statistical%20System%20(ESS)%20-%20European%20Statistical%20System%20(ESS)%20-%20Eurostat%20(europa.eu))
2. [Why is the price of Irish butter soaring? – The Irish Times](https://www.irishtimes.com/ireland/2022/05/28/why-is-the-price-of-irish-butter-soaring/) – Conor Pope, Sat May 28 2022 - 07:00
3. Hansen,B.G, 2020, *Forecasting prices of dairy commodities- a comparison of linear and nonlinear models* , vol. 59 , No.1, pp.98-112, Irish journal of Agricultural and Food Research.
4. Crisp-DM Flow Chart, Fig. 1, https://medium.com/towards-data-science/crisp-dm-methodology-for-your-first-data-science-project-769f35e0346c
5. Data Mining, available at <https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb>
6. Outlier Detection available at [Treat Outliers in the Dataset | Outlier Treatment for Data Science (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2021/07/how-to-treat-outliers-in-a-data-set/)
7. Burkov A. 2019, *The Hundred-Page Machine Learning Book - Draft*, 2nd Edition, Smauer Freeman.
8. Cormen T.H, Leiserson C.E, Rivest R.L 2000, *Introduction to Algorithms* , 3rd Edition, The MIT Press.
9. Feature Importance ,available at <https://medium.com/towards-data-science/understanding-feature-importance-and-how-to-implement-it-in-python-ff0287b20285>
10. Exploratory Data Analysis , available at ,<https://medium.com/towards-data-science/exploratory-data-analysis-8fc1cb20fd15>
11. Exploratory Data Analysis , available at , <https://towardsdatascience.com/comprehensive-guide-to-exploratory-data-analysis-of-habermans-survival-data-set-b33f0373c83a>
12. Pearson K.Ronald, 2018, Exploratory Data Analysis using R, CRC Press
13. Chen C, Härdle, W, Unwin A, 2008, Handbook of Data Visualisation, Springer-Verlag
14. Missing Data , available at , <https://www.datasciencemadesimple.com/check-count-missing-values-pandas-python-2/>
15. dtypes conversion, available at ,<https://flexiple.com/python/float-to-int-python/>
16. Weiss. A,Neil, 2017,Introductory Statistics, Pearson
17. Normal Distribution, available at , <https://www.askpython.com/python/normal-distribution>
18. Data visualisation ,available at , <https://www.analyticsvidhya.com/blog/2021/02/an-intuitive-guide-to-visualization-in-python/>
19. Outliers ,available at ,<https://towardsdatascience.com/outlier-why-is-it-important-af58adbefecc>
20. Removing Outliers , available at <https://www.askpython.com/python/examples/detection-removal-outliers-in-python>
21. Normalisation vs Standardization , available at <https://medium.com/swlh/difference-between-standardization-normalization-99be0320c1b1>
22. Splitting the data , available at [, https://towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c](https://onedrive.coca-cola.com/personal/michelle_moran_coca-cola_com/Documents/Masters%202022/Assignment%202%202022/Assignment/,%20https:/towardsdatascience.com/how-to-split-data-into-three-sets-train-validation-and-test-and-why-e50d22d3e54c)
23. Splitting the data , available at <https://towardsdatascience.com/how-to-split-a-dataset-into-training-and-testing-sets-b146b1649830>
24. KNN Algorithm, available at [, https://realpython.com/knn-python/, by Joos Korstanje Apr 07, 2021](https://onedrive.coca-cola.com/personal/michelle_moran_coca-cola_com/Documents/Masters%202022/Assignment%202%202022/Assignment/,%20https:/realpython.com/knn-python/,%20by%20Joos%20Korstanje%20%20Apr%2007,%202021)
25. KNN Algorithm ,available at ,  [[K-Nearest Neighbors Algorithm | KNN Regression Python (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/), 2018](https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python,%20Aishwarya%20Singh%20—%20August%2022,%202018)
26. Linear Regression available at , [https://realpython.com/linear-regression-in-python/#simple-linear-regression-with-scikit-learn, by Mirko Stojiljković Apr 15, 2019](https://realpython.com/linear-regression-in-python/%23simple-linear-regression-with-scikit-learn,%20by%20Mirko%20Stojiljković%20%20Apr%2015,%202019)
27. Linear Regression available at <https://statisticsbyjim.com/regression/interpret-r-squared-regression>
28. Linear Regression available at <https://en.wikipedia.org/wiki/Ridge_regression>
29. Linear Regression available at [https://en.wikipedia.org/wiki/Lasso\_(statistics)#:~:](https://en.wikipedia.org/wiki/Lasso_(statistics)%23:~:%20) Linear Regression available at ,<https://en.wikipedia.org/wiki/Bayesian_inference>
30. Random-forest-regression available at <https://towardsdatascience.com/random-forest-regression-5f605132d19d>
31. Sentiment Analysis available at <https://medium.com/nerd-for-tech/step-by-step-guide-to-twitter-sentiment-analysis-bc250caf3a3c>
32. Sentiment Analysis available https://towardsdatascience.com/how-to-do-a-sentiment-analysis-in-realtime-using-the-jupyter-notebook-kafka-and-nltk-4470aa8c3c30
33. Sentiment Analysis available [What is Sentiment Analysis? - GeeksforGeeks](https://www.geeksforgeeks.org/what-is-sentiment-analysis/)
34. Dashboard creation available at https://towardsdatascience.com/creating-an-interactive-dashboard-from-jupyter-notebook-with-voila-b64918b4d15a

**10.0 Appendices** :

Dashboard : voila javascript error



Appendices

Appendix A

Proof of permission from employment sanctioning the use of data