**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation  Statistics for Data Analysis  Machine Learning for Data Analysis  Programming for Data Analytics |
| **Assessment Title:** | CA2- Integrated Assessment |
| **Lecturer Name:** | David McQuaid  Marina Iantorno  Dr. Muhammad Iqbal  Sam Weiss |
| **Student Full Name:** | Kristine Fae Surat |
| **Student Number:** | sba22208 |
| **Assessment Due Date:** | 07 January 2023 |
| **Date of Submission:** | 07 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. |

Table of Contents

[1.0 Abstract 3](#_Toc123906380)

[2.0 Introduction 3](#_Toc123906381)

[3.0 Data Preparation 3](#_Toc123906382)

[4.0 Statistics 5](#_Toc123906383)

[5.0 Machine Learning 11](#_Toc123906384)

[6.0 Learnings 18](#_Toc123906385)

[7.0 Conclusion 18](#_Toc123906386)

# Abstract

*This paper examines the milk production and utilisation as well as the number of dairy cows available in Ireland and in other EU countries. This paper also looks at the milk products butter and cheese for sentiment analysis.*

*This paper demonstrates that prediction models give a different prediction results and that Linear regression is the best in predicting values and TF-IDF is a great model in sentiment analysis.*

# Introduction

The purpose of this report is to compare milk production and utilisation in Ireland and in other EU countries. Number of dairy cows from each countries will also be looked at as well. Since butter and cheese are products that can be derived from milk, sentimental analysis will be performed from reviews and tweets related to butter and cheese.

For the statistics part, I wanted to compare Ireland and Netherlands for the reason being they have a comparable area as compared to bigger land mass of other EU countries. Another reason is that I have not visited Netherlands yet so I have no idea how their milk or dairy products taste like, or how much their dairy products are when you buy from the shops. To answer my curiosity, I will be using data of Netherlands to compare to the data of Ireland.

# Data Preparation

**3.1. Acquiring the raw data**

One source of raw data collection is from Trading Economics website. More historical data were gathered by installing Trading Economics Excel Add-In Documentation and searched from excel. The add-in connects excel to Trading Economics API which provides direct access to around 300,000 topics and download millions of historical data rows, however, most of the data are locked with no permission to access the dataset, hence, only available data associated with dairy are collected and collated.

Another source is from Eurostat Statistics Explained website. The data gathered from this source compared dairy information from other EU countries as well as some Non-EU countries. There were no licenses needed to access the data. The data gathered here are milk production and utilisation from the countries measured in Tonnes, and number of dairy cows from these countries. The units of the numeric values from the raw data were 1000 Tonnes. Data were extracted on the 3rd of November 2022.

I have also used Central Statistics Office website to get data specific to Ireland.

**3.2. Cleaning the Data**

The data is cleaned by removing unnecessary columns and unimportant rows. Replacing of texts is also done to rename long texts to shorter names. This is done to have more consistent data, prevent errors and cleaner when applied to graphs. Some examples are 'Germany (until 1990 former territory of the FRG)' is renamed to 'Germany' only so the text will be shorter in graphs and for uniformity, Preserved Dutch Butter is renamed to Dutch, and Milk Produced to Milk. I have learned that keeping the labels and indexes’ number of words are kept low, preferably kept to just one word, to prevent errors when dealing with the data in the Jupyter notebook.

Some cells of the raw data are not available. To fill in the missing data, I have calculated the mean values of the previous years and used these values. I have chosen mean to fill in the missing values. This is called mean imputation. I have chosen mean because according to malicksarr.com, choosing mean to fill in the missing values allows missing data replacement with a plausible value that can improve the accuracy of the analysis and reduce bias by restricting the effects of extreme outliers.

After cleaning the separate datasets from the raw data I have gathered, I have saved new files and merged those to create a new dataset that was used for the Statistics and Machine Learning python programming. References 1, 2, and 3- Data Preparation Jupyter notebooks.

# Statistics

**4.1 Descriptive Statistics**

Table 1. Milk Production in Ireland.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | January | February | March | April | May | June | July | August | September | October | November | December | mean\_rows | median\_rows | variance\_rows |
| 2020 | 176.2 | 331.7 | 725.7 | 982.7 | 1115.3 | 1031.0 | 985.0 | 867.6 | 725.4 | 646.8 | 449.8 | 258.8 | 793.538462 | 759.619231 | 198028.920288 |
| 2021 | 182.2 | 352.4 | 829.7 | 1060.5 | 1181.0 | 1067.3 | 1017.3 | 917.4 | 776.7 | 649.6 | 465.9 | 258.5 | 829.192308 | 829.446154 | 202581.997812 |
| 2022 | 183.3 | 367.5 | 808.6 | 1054.6 | 1166.0 | 1057.6 | 1022.8 | 919.0 | 785.6 | 698.5 | 465.9 | 258.5 | 831.530769 | 820.065385 | 199444.411362 |

Chart, pie chart

Description automatically generatedTable 1 explains the volume of milk produced in Ireland (in 1000 Tonnes). Table 1 implies that the mean of milk production increased in the past three years, also displayed in Figure 1. Year 2021 has the highest median among the three years.

*Figure 1. Mean Values of Milk Produced in Ireland for Year 2020 to 2022*

Chart, scatter chart

Description automatically generated

*Figure 2. Milk Produced in Ireland per Month*

Figure 2 shows that milk produced the highest during the month of May and lowest during January. Data is consistent for the three years. Figure 3 shows that milk produced in 2021 and 2022 are closer that with 2020.

A picture containing window

Description automatically generated

*Figure 3. Histogram of Milk Produced in Ireland for 2020 to 2022*

Chart, histogram

Description automatically generated

*Figure 4. Average milk produced in EU for 2011 to 2021.*

It can be inferred from Figure 4 that Germany produced milk the most for the past 10 years, followed by France.

Chart

Description automatically generated

*Figure 5. Bar graph of Milk produced in Ireland from 2011 to 2021*

*Chart, line chart

Description automatically generated*

*Figure 6. Line graph of Milk produced in Ireland from 2011 to 2021*

Figure 5 and Figure 6 show that milk produced in Ireland continued to increase for the past 10 years.

**4.2. Inferential Statistics**

**4.2.1 T-Test, One Population**

I was interested to test the average Milk produced in Tonnes in Ireland using T-Test. The mean of the amount of milk produced in Ireland for 2011 to 2021 was 7003000 Tonnes from Eurostat data *(*Reference *6- Statistics (Milk Production).ipynb).* I wanted to compare and verify if the mean of the dataset from Table 1 will give the same value. At a 5% significance level, I wanted to test mean values from year 2021.

The Hypothesis are:

*Null Hypothesis: The mean value of milk produced in Ireland per year is equal to7003000 Tonnes*

*Alternative Hypothesis: The mean value of milk produced in Ireland per year is not equal 7003000 Tonnes*

The p-value I have gathered was 1.90 e-15 which is less than 0.05, therefore, the null hypothesis is rejected and the alternative hypothesis is accepted. There is enough evidence to say that the mean amount of milk produced in Ireland per year is not equal to 70030000 Tonnes. (Reference 5-Statistics of Ireland Milk, In [28].

Now, I wanted to test if the mean values per year is greater than or less than 7003000 Tonnes of milk produced.

*Null Hypothesis: The mean value of milk produced in Ireland per year is less than 7003000 Tonnes*

*Alternative Hypothesis: The mean value of milk produced in Ireland per year is greater than 7003000 Tonnes*

*The p-value I got is more than 0.05, therefore, I accept the null hypothesis. There is enough evidence to say that the mean amount of milk produced and utilised in Ireland per year is less than 7003000 Tonnes.* (Reference *5-Statistics of Ireland Milk.ipynb*, Out [29].

**4.2.2. T-Test, Two-Population**

I was interested to compare the mean milk production of Ireland and Netherlands, if they are similar, at a 5% significance level.

*Null Hypothesis: The mean value of milk produced in Ireland is equal to the mean value of milk produced in the Netherlands*

*Alternative Hypothesis: The mean value of milk produced in Ireland is not equal to the mean value of milk produced in the Netherlands*

Considering that the variances of the mean values of each country are equal, the p-value I got was 5.95e-11. The p-value is less than the threshold (alpha=0.05), then there is evidence against the null hypothesis of equal population means. Ireland sample mean is not equal to Netherlands sample mean. (Reference *6- Statistics(Milk Production).ipynb*, Out [41].

**4.2.3. ANOVA Test**

I have used the same variables for my ANOVA test. Still, comparing Ireland and Netherlands.

I have used Levene test to determine if the variances of mean values of Ireland and Netherlands can be assumed equal. I got p-value of 0.53 which is greater than 0.05, therefore, the variances are assumed to be equal. (Reference *6- Statistics(Milk Production).ipynb*, Out [46].

*Null Hypothesis: The mean value of milk produced in Ireland is statistically equal to the mean value of milk produced in the Netherlands*

*Alternative Hypothesis: The mean value of milk produced in Ireland is statistically different to the mean value of milk produced in the Netherlands*

I got a p-value of 5.94e-11 which is less than 0.05, therefore, the null hypothesis is rejected. There is enough evidence to say that the mean value of milk produced in Ireland is statistically different to the mean value of milk produced in the Netherlands

**4.2.4.  Kruskal-Wallis H test**

Again, I wanted to compare the median of milk production in Ireland against Netherlands.

*Null Hypothesis: The median of milk produced in Ireland is equal to the median of milk produced in the Netherlands*

*Alternative Hypothesis: The median of milk produced in Ireland is not equal to the median of milk produced in the Netherlands*

The p-value I got was 7.11e-05 which is less than 0.05, therefore, the null hypothesis is rejected. There is enough evidence to say that the medians are statistically not equal.

**4.2.5.  Linear Regression**

Calendar

Description automatically generated with medium confidence

*Figure 7. Heat map of countries and years*

The correlations between the years and the values are positively correlated, which means as one increases the other will increase too. The relationships are strong. It can be inferred that there is a big probability that the amount of milk produced will be increasing every year.

**4.2.5.1  Linear Regression for Ireland**

R-squared of the model in training set is: 0.9771931462509852

-----Test set statistics-----

R-squared of the model in test set is: 0.9269312717920057

The R-squared of the Linear regression model of Ireland dataset is very close to 1.This means that the response variable can be perfectly explained without error by the predictor variable.

**4.2.5.2  Linear Regression for Netherlands**

R-squared of the model in training set is: 0.7692487070827245

-----Test set statistics-----

R-squared of the model in test set is: 0.8895436730256527

The R-squared of the Linear regression model of Ireland is greater than that of the Netherlands. Ireland dataset fits better in regression model.

Linear regression model is a good model to be used when predicting the volume of milk in an area. We can see from the results that R-squared of test sets are close to 1 which are very good values. This means that the independent variables explain the variation in the dependent variable well.

I have chosen to use the 5 models above because I was confident enough that the results I will be getting from Python are easy to calculate and interpret. Also, I just learned the theory behind these 5 models recently, so I know I will be confident enough in interpreting the results.

# Machine Learning

**5.1 Model Comparisons**

Chart, scatter chart

Description automatically generated

*Figure 8. Plot of Milk produced vs Number of Dairy Cows in EU in 2011 to 2021*

Graphical user interface, text, whiteboard

Description automatically generated

*Figure 9. Polynomial Regression graph of Milk produced vs Number of Dairy Cows in EU in 2011 to 2021*

Shape, rectangle

Description automatically generated

*Figure 10. Hyperparameter Tuning Using Grid Search Cross Validation graph*

Chart, line chart

Description automatically generated

*Figure 11. Ridge Regression Graph of Milk produced vs Number of Dairy Cows in EU in 2011 to 2021*

Table 2. Comparison of Models used in Machine Learning

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | | R2 in Training set | R2 in test set | Root mean squared error of the prediction | Mean absolute percentage error of the prediction |
| Linear Regression | | 0.9540 | 0.9781 | 0.0404 | inf |
| Ridge Regression | | 0.9540 | 0.9769 | 0.0415 | inf |
| Lasso Regression | | 0.9102 | 0.9107 | 0.0817 | inf |
| Polynomial Regression | Polynomial degree 1 | 0.97 | 0.95 |  |  |
| Polynomial degree 2 | 0.97 | 0.97 |  |  |
| Polynomial degree 3 | 0.97 | 0.97 |  |  |
| Polynomial degree 6 | 0.98 | 0.97 |  |  |
| Polynomial degree 10 | 0.98 | 0.98 |  |  |
| Polynomial degree 20 | 0.99 | 0.99 |  |  |

The models I used for my Machine Learning are Regression models. Figure 8 and Figure 9 show that the volume of milk produced is directly related to the number of dairy cows producing milk. This proves that as the number of cows producing milk increases, the amount of milk produced also increases.

Figure 10 shows that using Hyperparameter Tuning using grid search CV gave a good result when comparing test and train data. Figure 10 shows that test and train results are very close to each other (0.9510 vs 0.9545).

In can be seen from Figure 11 that train and test scores are very similar when using the ridge regression.

Table 2 is the summary of the models used. Comparing the Linear, Ridge and Lasso regressions, Linear gave the best training and test scores, with the least root mean squared error of the prediction. It can be inferred that the best model for prediction to be used is Linear. Looking at the Polynomial summary, it looks like the higher the polynomial degree, the better the prediction is. As to the Table 2, polynomial degree 20 gave the best result.

**5.4 Sentiment Analysis**

For my sentiment Analysis, I have used two different sources. One source is from Twitter API to get tweets about cheese and butter in Ireland. The tweets are saved as a json file and was then converted to a csv file. I was not very happy with the tweets I have gathered for cheese and butter in EU especially Netherlands and I could not find a good source for dairy or milk related API. I have found *Getting Started with the API | Milk Documentation (milkcms.com)* but I found the site was not working for me. The website asked for registration which I did but logging in was asking for email verification but I did not received any emails. Because I found it hard to get a readily available data for the sentiment analysis, I just searched for reviews for famous brands of Irish butter and Dutch butter. First, I have searched from the internet which brands are well known. Based on Saveur.com, Kerrygold from Ireland and H. J. Wijsman & Zonen Preserved Dutch Butter from the Netherlands are great butters. Second, I looked up for Kerrygold and Presereved Dutch Butter reviews. I have used Amazon US review page as a source instead of gathering reviews from their respective websites to remove bias. Third, I have manually compiled the comments in an excel file. This file is used for my sentiment analysis.

To prepare my data, I have removed stop words, most common used words and rare words as their presence will not of any use in classification of my text data and to remove noise.

**Naïve Bayes**

Multinomial Naive Bayes and Bernoulli Naive Bayes are the models that are used to test my sentimental analysis data. I wanted to use Naive Bayes because it is a powerful algorithm that is used for analysing data for a reason that each feature being classified is not related to any other feature. Naïve Bayes has higher success rate than other algorithms. This means that one feature is not dependent to the other, which indicates I will have less chances of getting errors when analysing my data. (upgrad.com)

Multinomial Naive Bayes calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output. It considers a feature vector where a given term represents the number of times it appears. Bernoulli Naive Bayes, on the other hand, uses binary algorithm when a feature is present or not. (mygreatlearning.com)

Table 3 shows the summary of the probability results of tweets from Twitter API. For the purpose of comparison, I am only analysing the first 10 of the results from Python. (References *9- Sentiment Analysis (part 2)-c and b.ipynb, Out[21] and Out[23]).*

Table 3. Probability of Tests using Tweets form Twitter API

|  |  |  |  |
| --- | --- | --- | --- |
| **Multinomial Naive Bayes** | | **Bernoulli Naive Bayes** | |
| ***Butter*** | ***Cheese*** | ***Butter*** | ***Cheese*** |
| 9.773841e-01 | 2.261593e-02 | 9.999878e-01 | 1.218710e-05 |
| 1.000000e+00 | 2.360300e-09 | 1.000000e+00 | 7.858329e-10 |
| 9.999999e-01 | 1.305410e-07 | 1.000000e+00 | 1.094515e-10 |
| 9.930135e-01 | 6.986550e-03 | 9.999944e-01 | 5.564357e-06 |
| 9.999999e-01 | 1.305410e-07 | 1.000000e+00 | 1.094515e-10 |
| 9.999999e-01 | 7.260163e-08 | 1.000000e+00 | 6.840720e-11 |
| 1.000000e+00 | 9.517416e-11 | 1.000000e+00 | 1.050237e-09 |
| 9.999991e-01 | 9.476786e-07 | 9.999998e-01 | 2.288453e-07 |
| 5.459321e-01 | 4.540679e-01 | 8.274811e-01 | 1.725189e-01 |
| 1.000000e+00 | 6.022613e-09 | 1.000000e+00 | 5.979131e-10 |

Table . Probability of Tests using a Train set from Irish and Dutch cheese review.

|  |  |  |  |
| --- | --- | --- | --- |
| **Multinomial Naive Bayes** | | **Bernoulli Naive Bayes** | |
| ***Butter*** | ***Cheese*** | ***Butter*** | ***Cheese*** |
| 0.229883 | 0.770117 | 0.872567 | 0.127433 |
| 0.185510 | 0.814490 | 0.910943 | 0.089057 |
| 0.039628 | 0.960372 | 0.587789 | 0.412211 |
| 0.319536 | 0.680464 | 0.812716 | 0.187284 |
| 0.039628 | 0.960372 | 0.587789 | 0.412211 |
| 0.039628 | 0.960372 | 0.587789 | 0.412211 |
| 0.185964 | 0.814036 | 0.894024 | 0.105976 |
| 0.504983 | 0.495017 | 0.948885 | 0.051115 |
| 0.141946 | 0.858054 | 0.863520 | 0.136480 |
| 0.023975 | 0.976025 | 0.432079 | 0.567921 |

It can be inferred from Table 3 and Table 4 that probability results for Butter is better using Multinomial Naïve Bayes but for Cheese, Bernoulli Naïve Bayes is better. This could mean that Multinomial Naive Bayes can be used to predict comments about Butter and Bernoulli Naive Bayes for Cheese.

**Table

Description automatically generated**

*Figure 12. Sentiment Analysis. (Reference 11- Text\_Processing\_Butter\_Cheese\_review.ipynb [In 40] , [In 41])*

Graphical user interface

Description automatically generated with medium confidence

*Figure 13. Polarity Rating. (Reference 11- Text\_Processing\_Butter\_Cheese\_review.ipynb [In 48] , [In 49])*

Table 5. Summary of sentiment analysis of review of Irish and Dutch brand Butters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Brand | Positive | Neutral | Negative | Sample |
| Irish | 62% | 24% | 14% | 29 |
| Dutch | 58% | 26% | 16% | 19 |

Figure 12 and Figure 13 shows a part of the sentiment analysis of the Butter reviews analysed in Python and Table 5 is a summary of the reviews. It can be inferred that both brands have a majority of positive feedback. Both brands got negative feedback of overprice and positive feedback that they taste great. Because the data are manually pulled, and my sample numbers are not equal, I am not very confident in analysing which of the two brands are better than the other. I can infer, however, that the reviewers were looking at the price not just the taste of the product.

**Term Frequency-Inverse Document Frequency (TF-IDF)**

TF-IDF model is used to determine the weight of the words based on relevance. The most relevant words are the most important. (capitalone.com)

Table

Description automatically generated

*Figure 14. TF-IDF model result. (Reference 11- Text\_Processing\_Butter\_Cheese\_review.ipynb [In 33])*

It can be seen from Figure 14 that words ‘many’ and ‘dairy’ were used more. ‘Disappointing’ word is used more than the word ‘better’, and ‘Ireland’ is used more than the word ‘Europe’.

Text

Description automatically generated

*Figure 15. TF-IDF model fitting. (Reference 11- Text\_Processing\_Butter\_Cheese\_review.ipynb [In 54] [In 55])*

*Table

Description automatically generated*

*Figure 16. TF-IDF model fitting. (Reference 11- Text\_Processing\_Butter\_Cheese\_review.ipynb [In 56])*

It can be seen from Figure 15 that TF-IDF model gave a test accuracy score of 80% which is a good value. Thus, this model is a good tool in predicting reviews when they have been translated as positive, negative or neutral reviews. Figure 16 shows predicted values of the reviews’ most relevant words. The prediction of the words is fairly high.

# Learnings

Personally, I need more practice in programming part of this course. It took so much of my time trying to learn the codes and what caused the errors. To name a few of my learnings, even a simple addition of [] can affect my coding and will give error, and additional column included in the code can give error. Also, I need to read more articles with statistics as it was really challenging for me to analyse my results.

# Dashboards

Interactive Visualization of Geographical Data of Milk Production in EU from 2011 to 2021 can be found in 12*- Dashboard- Interactive Visualisation of Geographical Data.ipynb [In 5].*

Interactive Visualization of Geographical Data of Bovine Population in EU from 2011 to 2021 can be found in 12*- Dashboard- Interactive Visualisation of Geographical Data.ipynb [In 8].*

Milk vs Cow Graph graphs of EU countries from 2011 to 2021 can be found in *13- Dashboard-Milk\_Bovine\_Interactive.ipynb [In 29].*

For the first dashboard, I used choropleth because I wanted to show visualisations for each country so I used interactive visualisations for each geographical data. Second reason was because I was already familiar using the codes during lectures.

For the seconds dashboard, I used plotly express because I wanted to plot the milk production vs bovine population of the EU countries from 2011 to 2021. I was happy with the result I got especially that the plots are moving when you hit play button, but I know that I need more practice in making a better, more interactive result.

# Conclusion

Milk in Netherlands are produced and utilised more than in Ireland. One of the reasons is because there are more dairy cows available to produce milk in Netherlands than in Ireland. All statistics models use in this paper proved that the mean values of milk production and bovine population are higher in Netherlands than in Ireland.

Linear regression is the best model in predicting milk production in Ireland, and one factor is the number of dairy cows available in producing the milk.

TF-IDF model is great for sentimental analysis for dairy products because it gives a good prediction scores.