**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:** | Continuous Assessment 2 |
| **Lecturer Name:** | Marina Iantorno/Vladimir Milosavljevic  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Henrique Noronha |
| **Student Number:** | Sbs22102 |
| **Assessment Due Date:** | 20th May 2022 |
| **Date of Submission:** |  |

**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. |

1. **Introduction**

It is widely known that meat is a rich source of nutrition for all people. It is a commonly sought-after source of protein in supermarkets and is a highly versatile meat for cooking. Figure 1 below shows how, over the last 50 years, global meat production has expanded dramatically, more than quadrupled since 1961 (Ritchie and Roser, 2017). To attend to this continuously increasing demand of meat in general, it is vital that producers and distributors plan accordingly to increase efficiency of production and reduce waste as to avoid environmental impacts.

Chart, line chart

Description automatically generated

Figure 1

Understanding how different types of meat impact each other in terms of demand can be useful information for farmers and companies within the agricultural sector, as this can lead to more efficient planning for raising livestock and production within slaughterhouses. Previous studies in Turkey have shown that there is a direct correlation between the price of beef and other meats, such as chicken. With this finding, it was possible to predict the price of beef using regression models, with an accuracy rate of 91.4%. (AKIN et al., 2019).

This study aims to understand how the production of adult cattle is impacted by the production of other types of meat and determine an estimation rate of slaughtering (in tonnes) by analysing monthly data of pig meat, lamb, and chicken, using machine learning regression algorithms. Geographically, this study focuses on analysing the Irish market and comparing its findings with findings from the Italian market.

1. **Experimental setup**
   1. *Dataset*

The dataset used for this study can be found on the [Eurostat website via this link](https://ec.europa.eu/eurostat/databrowser/view/APRO_MT_PWGTM__custom_2616459/default/table?lang=en). Eurostat is the official statistical office of the European Union with the mission to provide high quality statistical data on European countries (Eurostat, n.d.). The data consists of monthly slaughtering in slaughterhouses measured in thousands of tonnes for adult cattle, pig meat, lamb, and chicken. The scope of the data covers Ireland and Italy, from January 2004 until February 2022. For a preview of the dataset, refer to Table 1 in the appendix.

* 1. *Exploratory data analysis and preparation*

As the first step, a profile report of the dataset was created using the DataPrep python library. DataPrep is a python open-source library that allows you to prepare and analyse your data with a few lines of code (Dey, 2021). The profile report is a method of this library, which generates a report within the Jupyter notebook cell with a variation of useful information for each of the features of a Pandas DataFrame such as data types, unique values, missing values, quantile statistics, descriptive statistics, and correlations. All of this is generated with visualisations and text (DataPrep, 2020).

After initial analysis, redundant features were excluded – ***DATAFLOW***, ***LAST UPDATE***, ***freq***, ***meatitem***, and ***unit***. Although these features held some details for the observations, each of them had only one unique value, therefore being repeated throughout all observations in the dataset (refer to Annex 1, these features are labelled as constant).

The next step was to transpose the ***meat*** feature from observations to features (see Figure 2 below). This was achieved by using the “pandas.DataFrame.pivot” method, which returns a reshaped DataFrame organized by the given index or column values (The Pandas Development Team, n.d.). The ***meat*** feature contains the different types of meat, identified by codes from the Eurostat database, which in next steps they will be converted to the actual meat names, such as chicken, pork, etc. The purpose of this reshaping is to setup the DataFrame in a way that each meat type is a feature. As it will be presented in further sections of this study, this reshaping is necessary to apply a regression model which uses independent and dependant variables (or features). In this case, the dependent variable is B1200 (adult cattle) and the independent variables are B3100 (pork), B4110 (lamb), B7100 (chicken), and B7200 (duck).

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2

Given that the codes and corresponding type of meat names are know, it is logical to rename the codes to actual meat names, this is more intuitive when calling each feature by its name to perform analysis using Python, renaming will also be done for the other feature names. See Figure 3 below for a preview of the dataset.

A picture containing text, scoreboard, outdoor, black

Description automatically generated

Figure 3

* 1. *Missing values*

It was observed that there are many missing values present in the dataset (refer to Annex 2 in the appendix for a graphical representation). To analyse the missing data further, a plot was created for each of the meat features, showing the year on X axis and the sum of tonnes on the Y axis. The purpose of this graph is to show which years have data available or not. By analysing Annex 3 in the appendix, it can be clearly seen that there is no data for ***duck*** from 1970 to 2002 for both countries, also there is no data available for Ireland from 2009 until 2022. Therefore the ***duck*** feature was excluded from the dataset, due to the lack of data availability.

The ***chicken*** feature, as can be seen in Annex 4, only has data from 2004 onwards. Lamb, pork, and adult cattle (Annex 5, 6, and 7 respectively) have data available through the majority of the years. After this analysis, it was decided to use the ***adult cattle***, ***pork***, ***lamb***, and ***chicken*** features, from 2004 until 2022.

Also, it was noticed that there were missing values in the ***chicken*** feature (see Figure 4 below). To fill in these observations the interpolate method from the Pandas library was used (The Pandas Development Team, n.d.). This technique computes and fills in the missing values in the same ascending sequence as the previous values. The reason why this method was chosen was due to its common use within a time-series dataset, such as the dataset used in this study (S, 2021).

A picture containing text, scoreboard

Description automatically generated

Figure 4 – chicken feature missing values

* 1. *Statistical analysis*
     1. *Central tendency measures*

An analysis of central tendency measures was performed on the ***adult cattle***, ***pork***, ***lamb***, and ***chicken*** features of the Irish dataset(see figure 5 below). For the central tendency measures of the Italian dataset, refer to the Appendix, table 2. Measures of centre are used to indicate where the most typical values of the dataset are, also often called averages (Weiss and Weiss, 2017, p. 116).

Graphical user interface, text

Description automatically generated

Figure 5 – central tendency measures, Irish dataset

Now that the distance of each parking meter to each of the car parks was obtained, it can be used to plot a histogram showing the Poisson distribution (Figure 6), which is a probability distribution used to model how often an event occurs over a specified period (Weiss and Weiss, 2017, p. 273).

Now that the central tendency measure for the Irish dataset are known, it is possible to affirm that the number of ***adult cattle*** meat produced in Ireland has a normal distribution with an average of 47 tonnes and a standard deviation of 6 points. Using these variables, it is possible to answer questions such as: what is the probability of choosing one month within one year that has a production of more than or equal to 50 tonnes? The answer is 31%, as demonstrated in figure 6 below, using Python’s Scipy library.

Text

Description automatically generated

Figure 6

* + 1. *Confidence intervals*

A sample of 20% of the observations of the Irish dataset was extracted, the same was performed on the Italian dataset. The central tendency measures were taken from these samples (see figure 7 and 8 below). As this sample is small (less than 50 observations), the Scipy library can be used to calculate the confidence interval for the sample’s mean. Please note all figures of production are in thousands of tonnes.

Text

Description automatically generated

Figure 7 - central tendency measures, Ireland adult cattle sample

Text

Description automatically generated

Figure 8 - central tendency measures, Italy adult cattle sample

It was found that the variance of adult cattle production in Ireland is of 42.56 and the 90% confidence interval for the mean of production in thousands of tonnes for adult cattle in Ireland is 47.85. For the Italian dataset the variance of adult cattle production is 155.85 and the 90% confidence interval is 71.64.

* + 1. *Hypothesis testing*

To perform the hypothesis test, it was first analysed how the total between Ireland and Italy differ. To accomplish this, a boxplot was used as can be seen in figure 9 below.

Chart, box and whisker chart

Description automatically generated

Figure 9

As it can be observed, the total production and variance of both countries differ quite significantly. Although the graph and initial analysis may lead us to believe that the countries have different patterns of production for each product (adult cattle, pork, and lamb), it is needed to prove this through statistical tests. An experiment was conducted using the Shapiro test to produce statistical evidence to the hypothesis to be defined.

* + 1. *Shapiro test - Ireland*

Firstly, set 3 hypotheses tests to verify the distribution of the Irish subset for each type of meat. The hypothesis are to define if it is possible to accept or reject the hypothesis that the distributions of adult cattle are normal or not, in other words, if the p-value is less or greater than 0.05.

**The hypothesis are as follows:**

H0: Sample of adult cattle is a normal distributions (p>0.05).

H1: Sample of adult cattle is not a normal distributions.

H0: Sample of pork is a normal distributions (p>0.05).

H1: Sample of pork is not a normal distributions.

H0: Sample of lamb is a normal distributions (p>0.05).

H1: Sample of lamb is not a normal distributions.

**The results of the Shapiro tests for Ireland are as follows (extracted from Python code):**

Adult cattle: ShapiroResult(statistic=0.9598056077957153, pvalue=0.1756906658411026)

Pork: ShapiroResult(statistic=0.9293535947799683, pvalue=0.004235260654240847)

Lamb: ShapiroResult(statistic=0.9693857431411743, pvalue=0.3440837562084198)

**For the hypothesis defined above, it is possible to conclude the following:**

H0: Sample of adult cattle is a normal distributions (p>0.05). **ACCEPTED**

H1: Sample of adult cattle is not a normal distributions. **REJECTED**

H0: Sample of pork is a normal distributions (p>0.05). **REJECTED**

H1: Sample of pork is not a normal distributions. **ACCEPTED**

H0: Sample of lamb is a normal distributions (p>0.05). **ACCEPTED**

H1: Sample of lamb is not a normal distributions. **REJECTED**

Chart, line chart

Description automatically generated

Figure 10

* + 1. *Shapiro test - Italy*

The hypothesis set out for Italy are the same which were set for Ireland.

**The results of the Shapiro tests for Italy are as follows (extracted from Python code):**

Adult: ShapiroResult(statistic=0.9233238697052002, pvalue=0.00872541218996048)

B3100: ShapiroResult(statistic=0.9870803356170654, pvalue=0.8632708191871643)

B4110: ShapiroResult(statistic=0.7738800048828125, pvalue=1.609310629646643e-06)

**For the hypothesis defined, the following can be concluded:**

H0: Sample of adult cattle is a normal distributions (p>0.05). **REJECTED**

H1: Sample of adult cattle is not a normal distributions. **ACCEPTED**

H0: Sample of pork is a normal distributions (p>0.05). **ACCEPTED**

H1: Sample of pork is not a normal distributions. **REJECTED**

H0: Sample of lamb is a normal distributions (p>0.05). **REJECTED**

H1: Sample of lamb is not a normal distributions. **ACCEPTED**

Chart, line chart

Description automatically generated

Figure 11

* + 1. *Kruskal-Wallis test - Ireland*

As the Shapiro test above for both Ireland and Italy identified that not all groups of products have normal distributions, a further analysis to perform is the Kruskal-Wallis test, which can determine if there are significant variances between the median of two or more independent samples.

**The hypothesis is as follows:**

H0: The median is equal across all types of meat.(p>0.05)

H1: The median is not equal across all types of meat.(p<0.05)

**The results of the Kruskal-Wallis test for Ireland is as follows (extracted from Python code):**

KruskalResult(statistic=114.49404563309356, pvalue=1.3738328143010203e-25)

**For the hypothesis defined, the following can be concluded:**

H0: The median is equal across all types of meat.(p>0.05) **REJECTED**

H1: The median is not equal across all types of meat.(p<0.05) **ACCEPTED**

Chart, box and whisker chart

Description automatically generated

* + 1. *Kruskal-Wallis test - Italy*

**The results of the Kruskal-Wallis test for Italy is as follows (extracted from Python code):**

KruskalResult(statistic=115.17281856337199, pvalue=9.784535569041728e-26)

**Results of the hypothesis:**

H0: The median is equal across all types of meat.(p>0.05) **REJECTED**

H1: The median is not equal across all types of meat.(p<0.05) **ACCEPTED**

Chart, box and whisker chart

Description automatically generated

Figure 12

* + 1. *Levene’s test – Ireland and Italy*

Here we will use Levene’s test to check whether the variance across both samples of adult cattle from Ireland and Italy are the same. The Levene test is a good alternative to the Bartlett’s test for the samples presented, this can be confirmed because one of the samples does not have a normal distribution (adult cattle in the Italian dataset) so the Levene is a better test to perform as Bartlett’s test is sensitive to skewness in the normality of the sample.

**Defining the hypothesis:**

H0: The adult cattle groups have equal variance between both countries.(p>0.05)

H1: The adult cattle groups have different variances when comparing both countries. (p<0.05)

**The results of the Levene’s test for is as follows (extracted from Python code):**

LeveneResult(statistic=29.818075167975234, pvalue=5.458941325858164e-07)

A

A

A

* + 1. *Correlation matrix heatmaps*

Before applying the machine learning regression models to the datasets for Ireland and Italy, an analysis of the correlation between the features were performed. As demonstrated in Figure 6 below, there is no significant correlation between adult cattle and the other features. This is an important discovery as the low correlation metrics will likely result in poor prediction outcomes from the machine learning regression models.

A picture containing application

Description automatically generated

Figure 13 – Ireland dataset features correlations

Similarly, the correlation between features in the Italian dataset were not significant. There was an interesting discovery regarding the correlation between chicken and adult cattle, which turned out to be -0.69, which means there is an inverse correlation between the two features. Again, these poor correlation scores may turn out to have a strong determination in the predictions of the machine learning models.

Background pattern

Description automatically generated

* 1. *Machine learning algorithm selection*

As it was stated in the introduction, this study aims to understand how the production of adult cattle is impacted by the production of other types of meat and determine an estimation rate of slaughtering (in tonnes), using machine learning regression algorithms.

The reason why linear regression models were chosen for this study is simply because of the nature of the question this study aims to answer. The linear regression model receives an input vector and uses it to predict an output (Hastie, Tibshirani and Friedman, 2009). Therefore, before applying the machine learning models in this study, the features were separated into independent variables, which are denominated X, the input vector (chicken, pork, and lamb), and a dependant variable, denominated y, which is the target feature to be predicted (adult cattle).

* + 1. *Multiple linear regression*

Linear regression is a simple closed-form solution because it uses the simplest model form. The most straightforward machine learning method is linear regression, and it can easily be applied to small data sets and may be read intuitively (Jiang, 2021).

Since the dataset used in this analysis have multiple independent features, it is necessary to perform a multiple linear regression. To accomplish this, a linear regression can be readily extended to many dimensions using the Scikit-learn package (José Unpingco, 2016). Refer to Equation 1 below for the mathematical reasoning of the multiple linear regression.



Equation 1 – multiple linear regression

* + 1. *Polynomial regression*

Linear regression models tend to be limiting and inflexible, a good way to add flexibility to the model is by adding polynomials to the input features. By using these polynomial features along with a linear regression, results in the polynomial regression. Refer to Equation 2 below for a mathematical representation of the polynomial regression.



Equation 2 – polynomial regression

* + 1. *Elastic net regression*

Elastic net is an extension of the linear regression which adds regularization penalties to the loss function during the training of the data, by using the L1 and L2 penalty functions. The hyperparameter "alpha" is used to determine how much weight each of the L1 and L2 penalties is given. The L1 penalty's contribution is weighted by one minus the alpha value, while the L2 penalty is weighted by one minus the alpha value (Brownlee, 2020).

* + 1. *Ridge regression*

Ridge regression is a regularization method that applies a straightforward L2 norm-minimization to the linear regression formula. For ridge regression, a simple closed-form solution can be found, and it can also help with overfitting and other estimation issues in linear regression with a high number of parameters (Jiang, 2021). Refer to Equation 3 below for the ridge regression mathematical equation.

Diagram

Description automatically generated

Equation 3 – ridge regression

* + 1. *Random forest regression*

The bootstrapping Random Forest approach combines ensemble learning methods with the decision tree framework to construct many randomly drawn decision trees from data, then average the results to get a new result that frequently leads to solid predictions (Beheshti, 2022).

* 1. *Model performance comparison*

As mentioned in section 2.4.1, the correlation heatmaps for both Ireland and Italy showed that there were no significant correlations between the features in the datasets. It was also stated that the lack of correlation could have an impact in the results of the regression scores and errors, impacting them negatively.

As it can be seen in table 1 below, there is a record of each model’s R2 score and mean absolute errors (or MAE) applied to each of the countries’ datasets. The mean absolute error calculates the average of the absolute values of the differences occurred between each of the predicted values and the true values in the test dataset. For example, if the true value in the dataset was 5 and the predicted value was 2, there is an absolute error of 3. The MAE calculates the average of all the absolute errors in the test dataset. The R-Squared (R2) is a statistical metric that determines how much of the variance in the dependent variable can be explained by the independent variable, in other words, indicates how well the data fits the regression model.

As is it was expected, the models’ performances were quite low, none scoring an R2 above 40% for Ireland, and none scoring above 62% for Italy. The best performing models for Ireland were the multiple linear regression and ridge regression models. On the Italian dataset, the best performing models were polynomial regression and random forest regression.

Several attempts were made to tune hyperparameters of some of the models in the hopes of increasing prediction performance. For the elastic net regression models for both Italy and Ireland, GridSearchCV was used to find and define the best hyperparameters for the model. It was determined that the best parameters was an alpha of 0.1 and L1 ratio of 0.7. Cross-validation was also applied during the GridSearchCV deployment. The reason for using cross-validation is because, instead of utilizing a single split into a training and validation set, we may utilize cross-validation to examine the performance of each parameter combination for a better estimate of generalization performance (Müller and Guido, 2017).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Ireland** | | **Italy** | |
| **Scoring metrics** | **R2** | **MAE** | **R2** | **MAE** |
| Multiple linear regression | 39.14% | 3.82 | 60.04% | 6.35 |
| Polynomial regression | 36.84% | 3.82 | 61.43% | 6.28 |
| Elastic net regression | 35.45% | 3.86 | 58.83% | 6.63 |
| Ridge regression | 39.13% | 3.82 | 60.08% | 6.34 |
| Random forest regression | 37.04% | 3.80 | 62.15% | 5.79 |

Table 1 – machine learning regression models comparisons

1. **Conclusion**
2. **References**
3. **Appendix**
   1. *Annexes*

*Table

Description automatically generated with medium confidence*

Annex 1

Chart, bar chart

Description automatically generated

Annex 2

Chart, bar chart

Description automatically generated

Annex 3 – duck (tonnes) by year

Graphical user interface, chart, bar chart

Description automatically generated with medium confidence

Annex 4 – chicken (tonnes) by year

*Chart, bar chart

Description automatically generated*

Annex 5 – lamb (tonnes) by year

A picture containing chime, measuring stick

Description automatically generated

Annex 6 – pork (tonnes) by year

Chart, bar chart

Description automatically generated

Annex 7 – adult cattle (tonnes) by year

* 1. *Tables*

Graphical user interface

Description automatically generated

Table 2 – first 5 rows

Graphical user interface, text, application

Description automatically generated

Table 3 – central tendency measures, Italian dataset