**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*

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?

* 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.

* 1. *Model performance comparison*

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 1 – first 5 rows

Graphical user interface, text, application

Description automatically generated

Table 2 – central tendency measures, Italian dataset