Text

Description automatically generated

Jader Leonardo Gomes da Rosa – **sba22199**

**sba22199@student.cct.ie**

**Production and Population**

**Continuous Assessment 2**

**Dublin**

**December 30,2022**

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| INTRODUCTION | 3 |
| Data Manipulation | 4 |
| Exploratory Data Analysis | 5 |
| Statistics and appropriate plots | 9 |
| Boxplot payment per year | 9 |
| Add a calculated column on Data Frame | 11 |
| Created new Data Frame | 12 |
| Report the payment | 12 |
| Boxplot number of spaces per year | 15 |
| Number of spaces per amount in 2006 | 17 |
| Distribution (Binomial) | 18 |
| Decision Trees and Random Forest | 19 |
| Decision TreeRandom forest | 20 |
| Explain about SEMMA | 23 |
| Explain about CRISP-DM | 23 |
| Problems with Testing Machine Learning Models | 25 |
| Principles & Best Practices | 25 |
| Reproducibility | 26 |
| How to Test Machine Learning Models | 26 |
| Invariance Test | 27 |
| Matplotlib pyplot: Pair plot | 28 |
| Columns description and details | 31 |
| CONCLUSION | 35 |
| REFERENCES | 31 |

**INTRODUCTION**

This project is based on two types of databases, the amount of population of the European Union countries between 2015 and 2020 and production of Barley, Meat Cattle, Meat Pig, and Raw Milk for the same period and countries.

Data bases were collected from Food and Agriculture Organization of the United Nations (FAO).

All information contained in this document was based on instructions obtained in the classroom and all complements are correct referenced.

All additions and changes were made by code in the Jupyter notebook and are added in this document.

All of the documents and creations made in this document are on GitHub and saved directly in the CA2\_Project repository

FAO allows filters and extraction of report in CSV format, the following parameters were applied to population, filtered by:

Population data bases: European Union countries, total female population, total male population, total rural population, and total urban population for the years 2015 to 2020.

Production data bases: European Union countries, total female population, total male population, total rural population, and total urban population for the years 2015 to 2020.

after the data were extracted, manipulations and collisions were made.

**Data Manipulation**

Jupyter Notebook, Population\_Eurpean:

1 – Import 4 Population csv databases. (1. Population\_Female, 2. Population\_Male, 3. Population\_Rural and 4. Population\_Urban).

2 – Renamed value columns according to respective reference (Female\_K, Male\_K, Rural\_K and Urban\_K).

3 – Created a Join and Merge for databases numbered as 5 and 6, 7 and 8, their respective consolidated merger generating a data frame called “df\_pop”

4 – The unnecessary/duplicated columns were removed.

5 – Saved new consolidated table called “Consolidated Populating.csv”

Jupyter Notebook, Production Eurpean:

1 – Import 4 production csv databases. (1. Barley, 2. Meat Cattle, 3. Meat Pig and 4. Raw Milk).

2 – Renamed value columns according to respective reference (Barley, Meat Cattle, Meat Pig and Raw Milk).

3 – Created a Join and Merge for databases numbered as 1 and 2, 3 and 4, their respective consolidated merger generating a data frame called “df\_prod”

4 – The unnecessary/duplicated columns were removed.

5 – Saved new consolidated table called “Consolidated Production.csv”

**Exploratory Data Analysis**

Some libraries were imported to execute the codes. During the project’s process, I had needed to add more libraries emerged, which were made according to it needed and in order of need. The main ones were loaded in the first line of Jupyter Notebook

Exploratory data Analysis (EDA) is a technique composed of a range of code that allows an analysis of the consistency and possible issues in the database.

The following commands were applied to EDA.

Full Print results in a summary of the first and last rows of the data frame and show the number of rows and columns.

Table

Description automatically generated

With this command we can identify that there is following column order by “Area”,” Year”,”Female\_K”, “Male\_K”, “Rural\_K” and “Urban\_K” with 162 rows and 6 columns.

Graphical user interface, table

Description automatically generated with medium confidence

With this command we can identify that there is following column order by “Area”,” Year”, ”Barley”, “Meat Cattle”, “Meat Pig” and “Raw Milk” with 122 rows and 6 columns.

With that, I was able to improve the database and insert a small adjustment that made my database more comprehensive to be able to generate other models.

Basic information, this informative command shows how many columns there are in the database, what is the name of each column, number of items filled in for each “non-Null” column, the type of data contained in each column and the size of file.

Graphical user interface, text

Description automatically generated

describe()

This function returns the count, mean, standard deviation, minimum and maximum values and the quantile values for the given series.

count()

This function returns the number of items in a series, excluding NaN values.

mean()

This function returns the mean of the values for the requested axis.

std()

This function returns the standard deviation of the values for the requested axis.

min()

This function returns the minimum value of the given series.

max()

This function returns the maximum value of the given series.

quantile()

This function returns values at the given quantile.

Graphical user interface, text, application

Description automatically generated

The “shape” command returns quantitative information for the numbers/quantity of columns and rows are there in the Data Frame (DF)

Graphical user interface, text, application, chat or text message, email

Description automatically generated

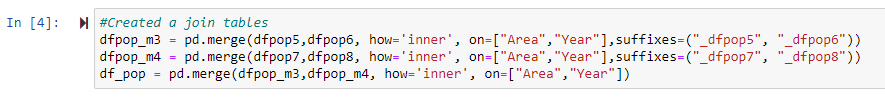
Renaming of columns to Population data base and Production data base  
  
Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

The INNER JOINs were made for Population data bases and Production data bases



Text

Description automatically generated with medium confidence

Removal of duplicate/unnecessary columns and save as a new data frame for Population and Production.

Text

Description automatically generated

Text

Description automatically generated with medium confidence

**T-Test, one population**

A one-sample t-test is a type of inferential statistical test which is used to determine whether a sample of observations could have been generated by a process with a specific mean. It is most used when the population standard deviation is unknown. The test is applied to the sample mean and is based on Student's t-distribution. The one-sample t-test is used to compare the mean of a sample to a known value, called the null hypothesis. If the sample means are significantly different from the null hypothesis, then the null hypothesis is rejected. It is commonly used to determine whether a process or treatment has influenced a population.

We are analyzing population applied for "Rural" from Ireland compared to other European union countries in average 4202k between 2015 to 2020.

Graphical user interface, text, application, email

Description automatically generated

We reject H0 ---> We accept H1, so there is enough evidence to say that the average of population is not 4202k in the European countries between 2015 to 2020.

We are analyzing population applied for "Rural\_K" from Ireland compared to Croatia in average 1785.48k.

Graphical user interface, text, application

Description automatically generated

We reject H0 ---> We accept H1, so there is enough evidence to say that the average of population is 1785.48k in Ireland and Croatia between 2015 to 2020.

**Statistics and Appropriate plots**

**Anova – Variable population**

A one-way ANOVA is a statistical test that is used to compare the means of two or more independent groups. The purpose of the test is to determine if there is a significant difference between the means of the groups. This test is also known as an Analysis of Variance (ANOVA) test. It is used in fields such as psychology, economics, and education to compare groups of data.

Two new columns were added in the population data frame containing the variation between urban and rural population and the total "sum of the two columns"

Chart

Description automatically generated

Filter was made to relate only the year 2020 to the data frame of the population

Table

Description automatically generated

"Shapiro wilk test" was carried out for the following countries (Ireland, Croatia, Denmark, Luxembourg, Malta, Portugal and Germany).

A picture containing graphical user interface

Description automatically generated

**CONCLUSION**

**REFERENCES**

<https://github.com/jaderleonardo/CA2_Project>

https://www.fao.org/faostat/en/#data/OA