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”

GitHub:

What is Git and GitHub?

Git is a version control system used for tracking changes in computer files and coordinating work on those files among multiple people. It is primarily used for source code management in software development, but it can be used to keep track of changes in any set of files.

GitHub is a web-based hosting service for version control using Git. It is mostly used for computer code. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features. It provides access control and several collaboration features such as bug tracking, feature requests, task management, and wikis for every project.

How do I Use Git and GitHub?

Git and GitHub can be used together to manage and store projects, such as websites, computer code, and documents.

1. Sign up for a GitHub account.

2. Download and install the Git command line tools.

3. Create a new repository or project on GitHub.

4. Create a local copy of the repository by cloning it to your computer.

5. Make changes to the project files on your computer.

6. Add the changes to the local repository using the "git add" command.

7. Commit the changes to the repository using the “git commit” command.

8. Push the changes to the remote repository on GitHub using the “git push” command.

9. Pull changes from the remote repository on GitHub using the “git pull” command.

10. Resolve any conflicts between the local and remote repositories.

**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

Graphical user interface

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

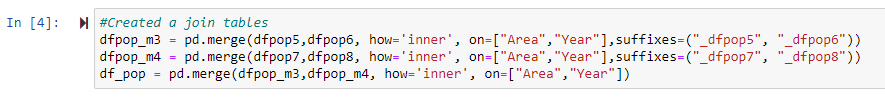
Renaming of columns to Population data base and Production data base  
  
Graphical user interface, text, application

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**

**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

Create a new data frame For Population with Ireland and some other European countries for 2020.

Table

Description automatically generated

Create a new data frame For Population with Ireland and some other European countries

Text

Description automatically generated

Create a new data frame for Production with Ireland and some other European countries and rename columns with “\_” between names.

Text

Description automatically generated

Organize New data frame by country.

The standard deviation of the columns can be found as follows:

>>> df.std()

age 18.786076

height 0.237417

A picture containing graphical user interface

Description automatically generated

Levene's test is a statistical test used to assess the homogeneity of variance of a dataset. It tests the null hypothesis that the variance of the data is equal across all groups. The test is named after statistician William G. Levene, who developed it in the 1950s. The test is based on the absolute deviation from the median and is more robust to outliers than other tests such as the F-test. It is often used in ANOVA tests to ensure that the data is suitable for analysis.

Graphical user interface, text, application, email

Description automatically generated

**Population data analyst**

In this next graph, we can analyze the data frame with a sample of some European countries (Including Ireland). and we can identify that most of the variation between the Urban and Rural Population is less than 10 million, except for Germany that its variation exceeds 40 million in 2020 and Ireland has a variation of only 1.335 million

Overall, we can observe that most European countries have a higher percentage of people living in urban areas compared to rural areas. This trend is expected to continue in the coming years as urbanization continues to increase.

Chart

Description automatically generated

Chart

Description automatically generated

Table

Description automatically generated

With the next table we can describe several variations and differences between the countries filtered for analyses, such as that in Luxembourg and Malta that have the highest number the male population than the female one, differently from what happens in other countries.

however, the focus of the analysis for the next graphs is only on the variation and the total population in 2020 for these countries

The greatest representativeness of the population variation between Urban and Rural continues to be in Germany 81.05%, followed by Denmark 7.90%, Portugal 5.96% and Ireland 2.39%, those are the top 5 countries of the sample.

Chart, pie chart

Description automatically generated

Analyzing the total population of the sample, we can identify a small change and a deviation in the representativeness of the total sample compared to the population variation.

Obs.: Germany has a large population compared to other European countries in part because it is the most populous country in Europe, with a population of over 82 million people. Additionally, Germany has seen relatively high levels of immigration in recent decades, with many people coming to the country in search of economic opportunities. Additionally, Germany has one of the lowest fertility rates in Europe, which has contributed to its high population.

In 2020 the total population based on the selected sample Germany has 76.01%, followed by Portugal 9.41%, Denmark 5.34%, and Ireland 4.50%.

Chart, pie chart

Description automatically generated

The next graph shows absolute values ​​for the representativeness of the percentage data previously presented.

Chart, waterfall chart

Description automatically generated

Chart

Description automatically generated

**Production data analyst**

For production analysis data frame was used the same countries were used on Population data analyst, but Croatia and Portugal did not have data for the production items in our sample and Malta has less information.  
  
A picture containing table

Description automatically generated

Barley Production:

We see that Barley production is mostly distributed between 3 countries out of the 5 selected for our sample, with Germany producing at 61.05%, Denmark at 23.19%, and Ireland at 15.54% the other 2 countries representing less than 1% of Barley production.

Chart

Description automatically generated

The graph shows absolute values ​​for the representativeness of the percentage data previously presented.

Chart

Description automatically generated

Meat Cattle Production:

The Meat Cattle production is mostly distributed only between 2 countries out of the 5 selected for our sample, with Germany and Ireland producing at 56.77% and 37.09%, Denmark represents only 5.55% the other 2 countries representing less than 1% of Meat Cattle production.

Chart

Description automatically generated with medium confidence

The graph shows absolute values ​​for the representativeness of the percentage data previously presented.

Chart, bar chart

Description automatically generated

note: in the next two charts we can identify a market dominance by Germany compared to the other 4 countries in our sample, with Germany representing more than 70% in the next two items.

Meat Pig Production:

The sample is distributed in the following percentages, Germany 71.96%, Denmark 21.13%, Ireland 6.68%, and the other 2 countries continue with less than 1% of representativeness.

Chart, pie chart

Description automatically generated

The graph shows absolute values ​​for the representativeness of the percentage data previously presented.

Chart

Description automatically generated

Raw Milk Production:

The sample is distributed in the following percentages, Germany 70.24%, Denmark 13.22%, Ireland 15.60%, and the other 2 countries now with almost 1% representation.

Chart, pie chart

Description automatically generated

The graph shows absolute values ​​for the representativeness of the percentage data previously presented.

Chart

Description automatically generated

**PCA and Clustering**

PCA (Principal Component Analysis) is a dimensionality reduction method used to reduce the number of features or variables in a dataset. It works by transforming the data into a new set of uncorrelated variables, known as principal components. They goal of PCA is to identify patterns in the data and represent the data in a simplified form.

Clustering is an unsupervised learning method used to identify groups or clusters of similar data points in a dataset. It is based on the concept of similarity and can be used to identify patterns and relationships within the data. Clustering can also be used to reduce the dimensionality of data by creating clusters of related features. Clustering can group customers, and similar group documents into categories, or group images into different classes.Table

Description automatically generated with low confidence

Graphical user interface, text, application, email

Description automatically generated

In this PCA we are relating and dealing with the values ​​of each production item for the year 2020 with all the countries of the European Union, the value of 0.811 and 0.166 represents the Explained Variance ratio of these items and country.

**PCA for Machine Learning Model**

PCA (Principal Component Analysis) is a dimensionality reduction technique used in machine learning models. It takes a set of data points with n-dimensions and creates a new set of data points with fewer dimensions. The new set of data points is called principal components.

PCA helps to reduce the complexity of the data and helps to identify patterns, trends, and outliers from the data. It also helps to reduce the noise from the data, making it easier for the model to learn from the data. PCA is also used to reduce the number of features in a dataset, making the model more efficient and easier to interpret.

Graphical user interface, text, application

Description automatically generated

With the next graph, we can identify that most of the data is related to point 0 for the Second Principal Component and -1 for the First Principal Component, showing that there is no big discrepancy between these data and also shows us some isolated points like the one next to -1.5 in the Second and 4 in First, 1.5 in Second and 2.5 in First, 0.5 second and 5 in First.

Chart, scatter chart, bubble chart

Description automatically generated

**Heatmap**

A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colours. It is used to visualize patterns and relationships in data. Heatmaps can be used to display information about a variety of topics including population density, economic activity, crime rates, and so on. They are also used in a variety of fields, such as marketing, finance, and medicine.

Population:

Heatmap shows a great relationship between the data, we can identify a great relationship between Female\_K with Male\_K and Total\_K and their presence of 0.94 in Rural\_K and 0.99 in Urban\_K.

Chart

Description automatically generated with low confidence

Production:

Heatmap shows a great relationship between the data, in this case we have no absolute relationship, but what is least related are Meat Cattle and Meat Pig with a value of 0.66.

Chart

Description automatically generated

A “histogram” is a representation of the distribution of data.

This function shows on each series in the Data Frame, resulting in one histogram per column.

Chart

Description automatically generated

**Appropriate plots**

A boxplot is a type of graph that can be used to display a set of data. It is especially useful for showing the distribution of the data, including its median, quartiles, range, and outliers. Boxplots are often used to compare two or more sets of data.

Some examples of appropriate uses of boxplots include:

-Comparing the distribution of scores in a test for two different classes.

-Comparing the average monthly expenses for different types of households.

-Comparing the range of temperatures in two different cities.

-Comparing the distribution of weights of two different types of animals.

-Comparing the median salary of two different occupations.

Barley:

Boxplot for Barley, we can say that his data are concentrated between 1.5 and a little more than 2.

Chart, box and whisker chart

Description automatically generated

Meat Cattle:

Boxplot for Meat Cattle, we can say that your data are more distributed and concentrated between 2.5 and 3, 4 and above 6.

Chart, box and whisker chart

Description automatically generated

Meat Pig:

Boxplot for Meat Pig, we can say that your data are more distributed and concentrated only between 0.2 and 0.4 containing consistent data greater than these concentrates.

Chart, box and whisker chart

Description automatically generated

Raw Milk:

Boxplot for Raw Milk, we can say that its data are more concentrated between 0.8 and a little above 1.0, the other data are concentrated above 1.2.

Chart, box and whisker chart

Description automatically generated

**Matplotlib pyplot: Pair plot**

The pair plot is a type of seaborn plot that is used to explore the relationships between multiple variables in a dataset. A pair plot is a two-dimensional graph that plots each variable against every other variable in a dataset. It is a great way to quickly visualize the relationships between the variables and get an idea of which variables might be useful for further analysis. It is also useful for identifying outliers or clusters in the data.

Production:

The Production Pairplot shows the data and their relationships with other data, the closer to zero (data point) the closer the relationship between the data.

In the case of production, we have a relationship close to the point 0.

Chart, scatter chart

Description automatically generated

we are presenting another format that shows the same relationship.



Chart, scatter chart

Description automatically generated

Population:

In the case of the population, the data are ascending with a linear growth and with little variation in this alignment.



Chart, scatter chart

Description automatically generated

we are presenting another format that shows the same relationship.



Chart, scatter chart

Description automatically generated

**CONCLUSION**

**REFERENCES**

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

https://devmountain.com/blog/what-is-github-and-how-do-you-use-it/

https://www.youtube.com/watch?v=pBy1zgt0XPc&ab\_channel=GitHub

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

<https://www.europelanguagejobs.com/blog/germany_the_most_popular_european_country.php>