# Group ID - MSc in Data Analytics

Author: Luiz Lopes

e-mail: [sba22211@student.cct.ie](mailto:sba22211@student.cct.ie)

Student ID: sba22211

Contents

[Group ID - MSc in Data Analytics 1](#_Toc123717816)

[Abstract 3](#_Toc123717817)

[1. Introduction 4](#_Toc123717818)

[1.1. Version Control 4](#_Toc123717819)

[1.2. Project Framework 4](#_Toc123717820)

[1.3. Project Background 5](#_Toc123717821)

[1.4. Data collection 5](#_Toc123717822)

[2. Data Extraction and Data Cleaning 5](#_Toc123717823)

[2.1. Data Extraction 5](#_Toc123717824)

[2.2. Data Cleaning 7](#_Toc123717825)

[3. Exploratory Data Analysis 9](#_Toc123717826)

[4. Statistical Analysis 15](#_Toc123717827)

[4.1. Descriptive Statistics 15](#_Toc123717828)

[4.2. Inferential Statistical Techniques 17](#_Toc123717829)

[4.2.1. T-Test one Way 17](#_Toc123717830)

[4.2.2. T-Test Two Way 18](#_Toc123717831)

[4.2.3. ANOVA One Way 18](#_Toc123717832)

[4.2.4. ANOVA Two Ways 20](#_Toc123717833)

[4.2.5. U-Mann Whitney 20](#_Toc123717834)

[5. Machine Learning 20](#_Toc123717835)

[6. Conclusion 21](#_Toc123717836)

[7. References 21](#_Toc123717837)

# Abstract

*It is a useful and common practice to put the abstract in Times New Roman 12-point italics. Throughout this document the styles used reflect the styles we suggest you use in your scientific report*.

# Introduction

Mushrooms are the biggest export sector within Irish horticulture, with over 80% of Irish mushroom production destined for the UK market. According to Bord Bia It currently employs over 3,500 people directly involved on the production of Mushrooms. Button/Closed cup mushrooms hold the largest share of the market followed by Flat/Breakfast mushrooms, value mushrooms and then the sliced and exotic mushrooms

(Bord Bia, 2019).

## Version Control

I have used GitHub for Version control. I have set up a new account under the college email and set up the repository to Public. I have also installed the desktop version and all pull requests where managed through the desktop version.

Project could be found at: <https://github.com/LuizLopes-cct/CA2>

Some of the initial research performed prior to my final choice of data set is stored in a folder called OLD. There are previous downloaded data sets from different sources. After creating the data frames and perform a quick EDA on these data, it was clear that the data was not suited for the project.

The files on this folder are not relevant for this project and can be ignored.

## Project Framework

For the project framework, I have chosen to use CRISP-DM. not just because my personal familiarity with this framework but it because the Agile like approach proposed by CRISP-DM with an Adaptive life cycle.

At the beginning of the project, I was not sure what data to use, CRISP-DM gave me the flexibility to test different datasets. Also the objectives were not well defined at the beginning, meaning that as I progressed with the Data Understanding I could update the Business Understanding phase accordingly.

This assignment paper is formatted to fit the subjects, but we could clearly perceive the CRISP-DM phases through it.

Data Understanding and Data Preparation phases are divided between chapters 2- Data Cleaning, 3- EDA and chapter 4- Statistics.

The Modeling stage is divided between chapters 2-Data Cleaning, 3-EDA and chapter 5-Machine Learning.

The remaining phases, Evaluation and Deployment concentrated on the Machine Learning chapter

## Project Background

The project focus on the Export and Import Values for Mushrooms and truffles in Ireland and compare with Its closest competitors in Europe (The Netherlands and Belgium).

I have analysed the Export/Import Quantity (Tonne) and Total Export/Import Values (1000 US Dollars) per year since 2000. I have also added the average price of 1 Tonne of mushrooms and truffles and analysed its variance through the years. The Idea is to predict if the price of the Tonne tends to go up or down on the following years based on the current trend.

## Data collection

The data used for the analysis were extracted from FAOSTAT website:

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

Selection for the data set were:

Area = Special group -European Union (27)> list

Elements = Select all

Items = Mushrooms and Truffles

Years = Select all

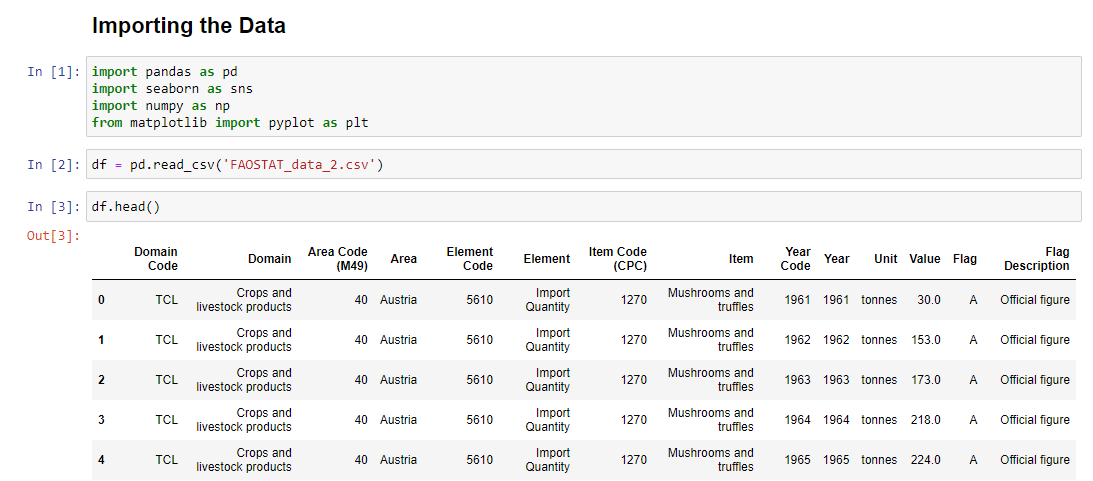
I have downloaded the data as a CSV and that’s the original source provided with the project. (FAOSTAT\_data\_2.csv)

# Data Extraction and Data Cleaning

## Data Extraction

Data Extraction is presented at the Jupyter Notebook called CA2\_Data\_Analytics.

As usual I have started by importing the relevant Libraries and Loading the file with the data. The next important step is to check the Table.



As we could see using the function head(), the majority of the columns are not relevant to the analyses and the models later.

Below is a reasoning for removing the columns from future analysis:

|  |  |  |
| --- | --- | --- |
| **Column Header** | **Removed** | **Reason** |
| Domain Code | Yes | Code for Domain, not necessary for the calculations |
| Domain | Yes | Unique Value reference of the data downloaded |
| Area Code (M49) | Yes | Area description will be used instead |
| Area | No | Not removed |
| Element Code | Yes | Element Description will be used instead |
| Element | No | Not Removed |
| Item Code (CPC) | Yes | Item description will be used instead |
| Item | No | Not Removed |
| Year Code | yes | Same as year |
| Year | No | Not Removed |
| Unit | Yes | Units are either Tonnes or 1000USD this is specified in charts or described on assignment paper |
| Value | No | Not Removed |
| Flag | Yes | Code for Flag Description |
| Flag Description | Yes | The flag is irrelevant for the calculations. |

The Flag Description could be Official Value or in case the official value is missing, 'Unofficial figure' or 'Imputed value' was used. I have used the value provided regardless if it is official or unofficial value. Later when I had the data set for Analysis I went back to check and all values used on the main dataset were Official Values.

The Item column could have been removed as all the values are the same (Mushrooms and truffles), but I have decided to leave it for now.

Table

Description automatically generated

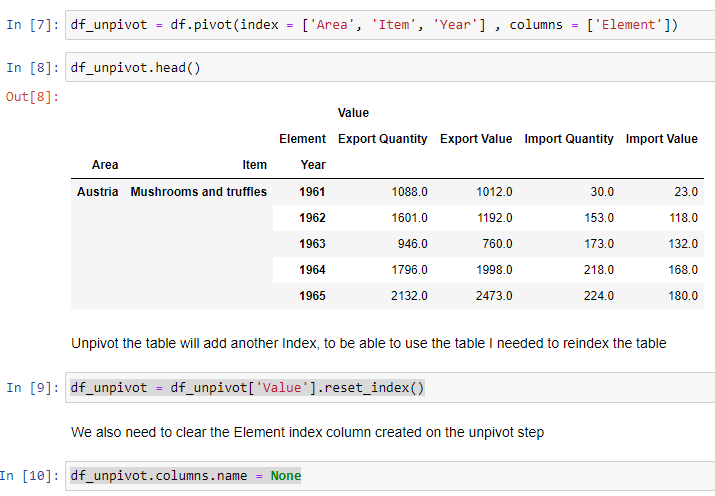
After removing the unnecessary columns, we ended up with the following table:

Graphical user interface, text, application, email

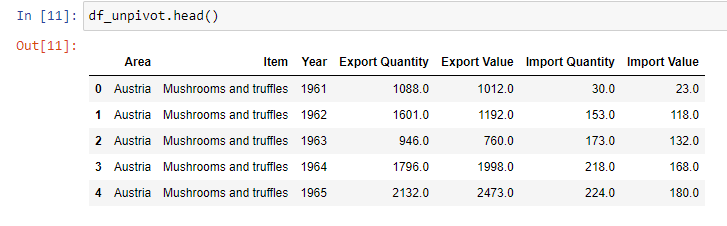
Description automatically generated

The element column contains for values: 'Import Quantity', 'Import Value', 'Export Quantity', 'Export Value', we want to unpivot those values to create 4 columns. The Idea is to have one row for every Area (Country)-Year combination.

Unpivoting the table will leave with a Value and Element indexes that will need to be removed.



The following table was used to clean the data:



## Data Cleaning

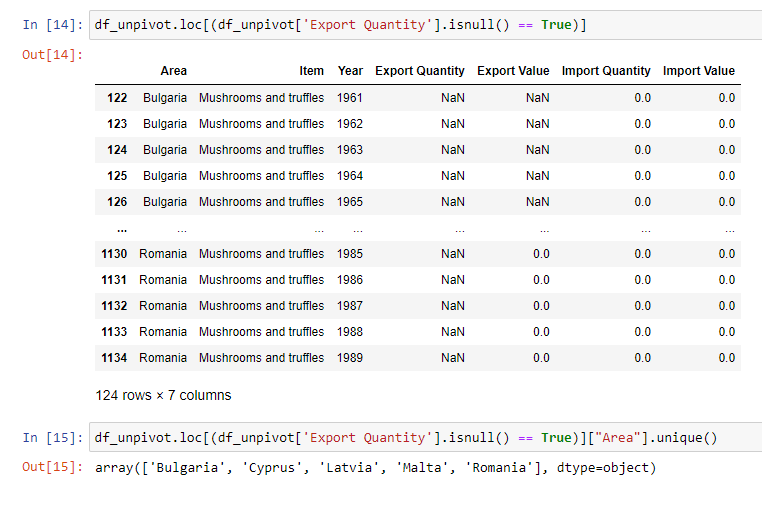
For data cleaning, the first step is to look for null values. Null values happen when the data is missing or is not available. For this assignment I have checked the null values and see if they could be removed or if I need to treat them differently.

Text

Description automatically generated

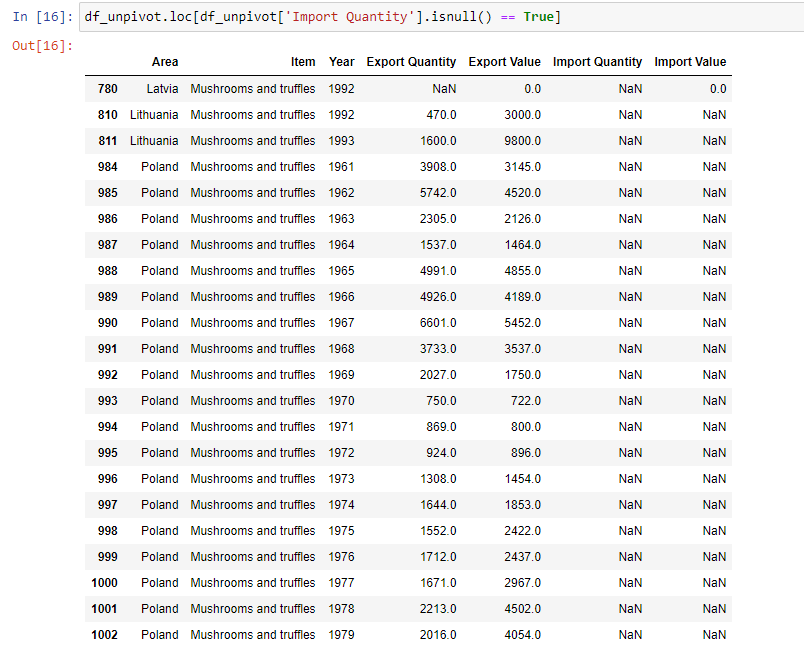
The number of null rows is relatively small when compared with the total data. But I wanted to investigate further to see if removing this data would cause some issues with my analysis later.

First I have analysed Export Quantities:

None of the countries 

Only a 124 row in this category, also majority of them have no Import Quantities available. Also, the countries in the selection are not the top producers of Mushrooms and Truffles. I considered it safe to remove these rows from the data frame.

Next I have Checked for the Null values on the import Quantity column.



Although in small numbers, in this selection I have Poland, one of the biggest producers of Mushrooms and positive values on the Export Quantities column.

Upon checking the Export Quantity values on the selection and compare it with the total values, I have found that these values represent about 1% only of the total quantities, as per code below:

Graphical user interface, text, application, email

Description automatically generated

Based on the above Analyses, it is safe to remove the null values with minimum to zero impact to the calculations.

There is no need to remove or change zero values as they are valid values for the calculation.

As expected, no Duplicated Rows were found:

Graphical user interface, text, application, email

Description automatically generated

# Exploratory Data Analysis

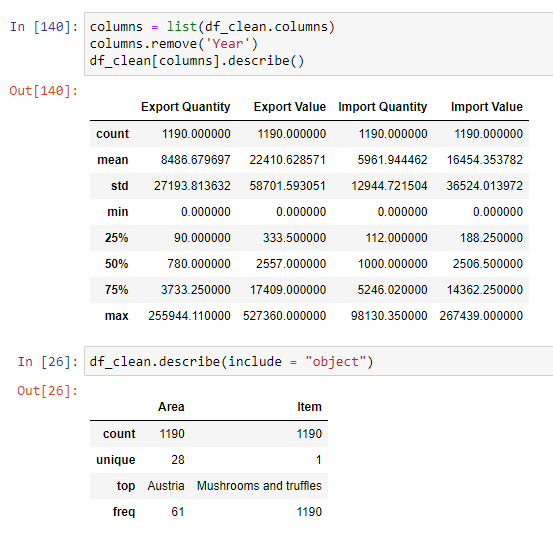
I have started the EDA checking the number of rows and columns of the Dataframe and the data types.

Graphical user interface, text, table

Description automatically generated

The datatypes are as expected, and no changes are necessary.

I have used the function Describe() to have a better understand of the data:



Based on the above, the difference between the Median and the mean are very big, meaning the data is not normally distributed, so the distribution is skewed.

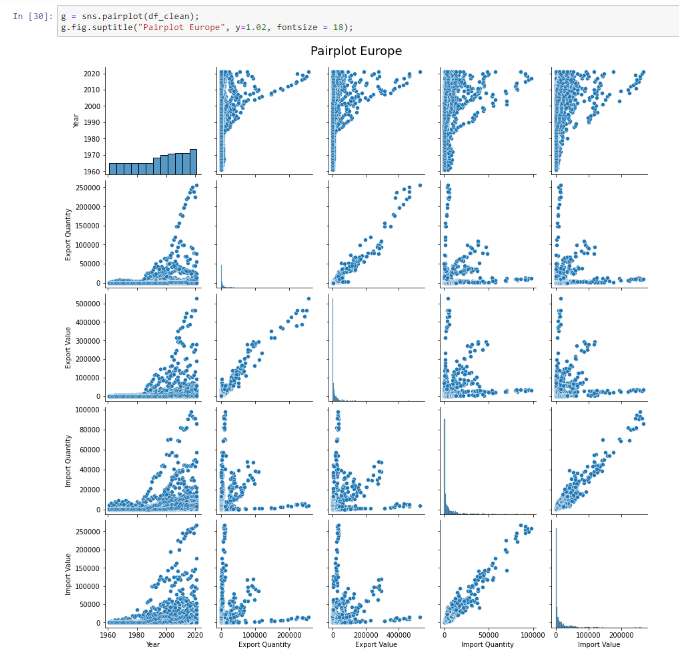
Next step on the EDA is to check the correlation between the columns:

Graphical user interface

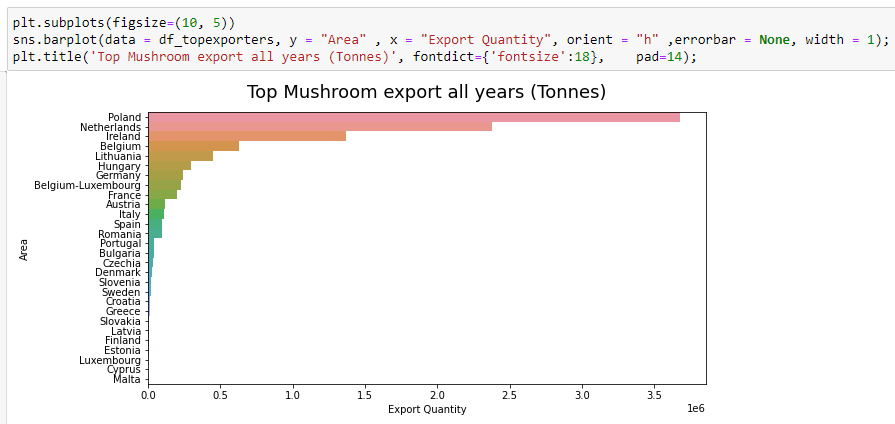
Description automatically generated

As expected, we have a strong correlation between the Export Quantity and the Export Value and between the Import Quantity and the Import Value.

We can visually see the correlation using a Pair plot visualization:



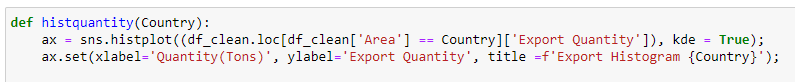
I want to drill down the data into individual countries, so I can compare with Ireland production. I will select the top three exporters of Mushrooms and Truffles from the dataframe.



Based on the Analyses of total export per country, Poland is way ahead than the others and it would distort the analyses. So I have picked the next three countries to conduct our Analysis; Netherlands, Belgium and Ireland

After deciding which countries, we will be comparing to Ireland I will perform individual analysis to each set.

I started with the Histogram, For that I have created a function which will generate the histogram for each desired country, this will save me time when creating the histograms.



Ireland Export Histogram

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Using a Line chart to see the Export Quantity per Year per Country:

Chart, line chart, histogram

Description automatically generated

Based on the Line chart above, we only have data for Belgium from 2000 onwards so I have limited the dataframe to data from 2000 only.

Chart, line chart

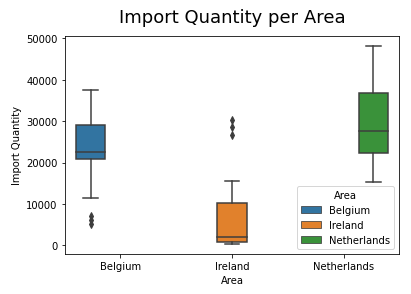
Description automatically generated

Using Boxplot I could analyse the distribution and outliers per country and then per Element:

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

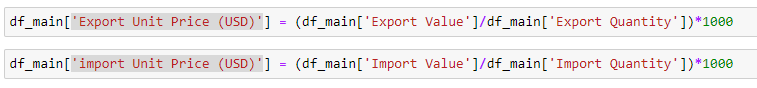
Description automatically generated

Chart, box and whisker chart

Description automatically generated

The Export Data for Belgium is much more concentrated with less outliers when compared to Ireland and Netherlands

The Last step on our EDA is to add the price per unit (1 Tonne), I have added the calculation for Export and Import.

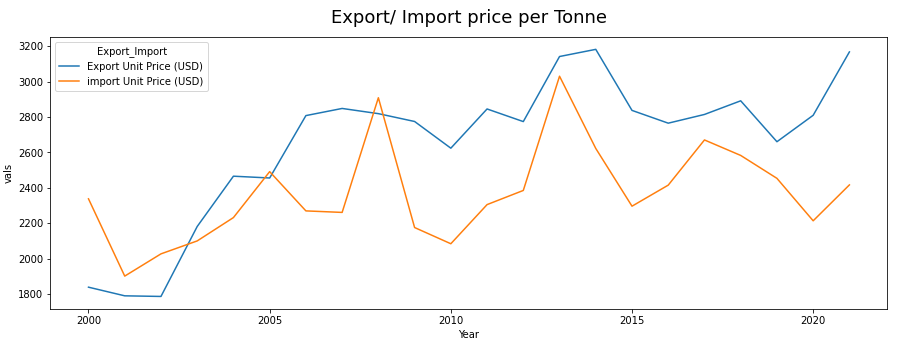


With the addition to the two new columns, we have our new dataframe ready for Machine Learning:

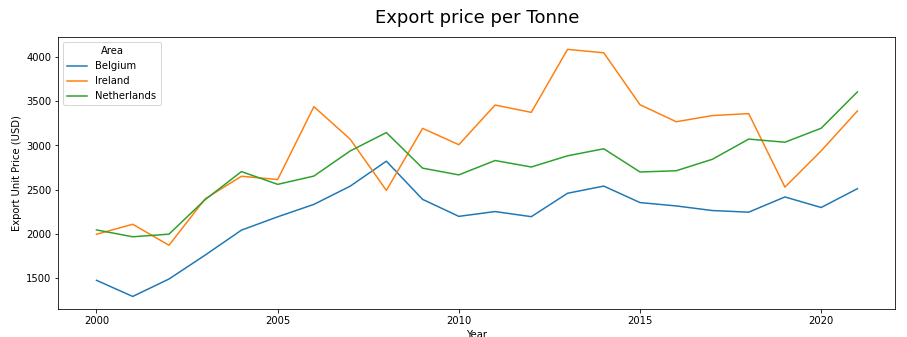
Graphical user interface

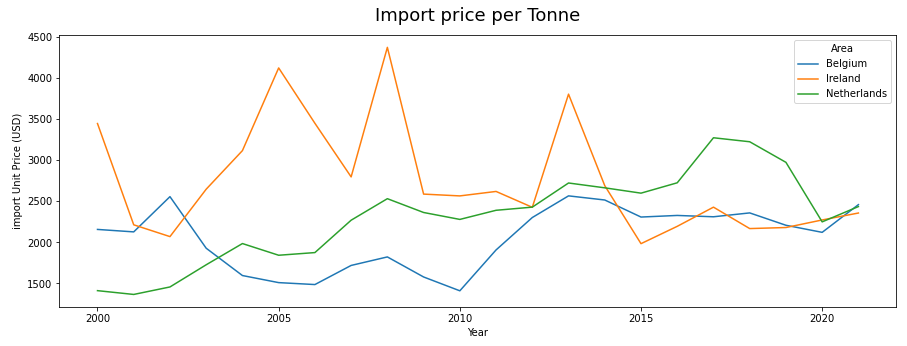
Description automatically generated

To check the relation between the Export/Import unit price I have put the distribution in a line chart:



I have also compared an Import/Export unit price graph per country





Although all the countries on the scope are part of European Union and follow the same regulations the export price per Tonne could vary considerably. It could be different factors as the Type of Mushrooms and Truffles produced per country or other factors not available from the data sources.

# Statistical Analysis

Some of the tables and charts regarding the Descriptive Statistics are already presented and explained at the EDA part of the Assignment. I have chosen to repeat them in the Jupyter Notebook CA2\_Statistics.ipynb, so they can be used as a reference to the calculations in the file and to facilitate the readability of the document.

## Descriptive Statistics

As explained during the data cleaning and EDA we have selected a sample of our data containing the Export and Import figures for the past 20 years for Ireland Belgium and Netherlands.

First I have started with the Describe function to summarize the central tendency, dispersion and shape of a dataset's distribution.

Table

Description automatically generated

The describe function provides most of the descriptive statistics measures, it provides the Total count, the mean the standard variation, min-max values, the median (50%), and the upper and lower percentile.

Although the describe function would provide most of the measures I will need, I also wanted to add the mode and the Variance. I have created my own table with the Central Tendency Measures and the Variation Measures.

Table

Description automatically generated

Upon this analysis I could verify that for the Export Unit Price (USD), the Mean and the Median, this could be a good indicator of a normal distribution.

Next step, I have run an histogram for all the columns to verify the data distribution for them.

Chart, histogram

Description automatically generated

After the Export Unit Price Histogram, I have decided to drill down the distribution per country.

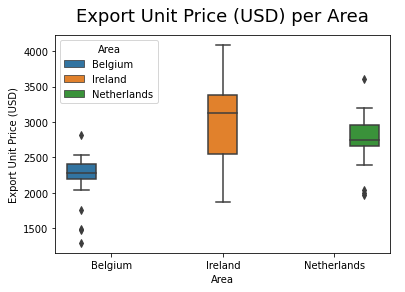
Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

I have also created a boxplot chart to Analyse the outliers.



Base on my analysis the Export Unit Price is the variable closest to a normal distribution. And it provides good points for comparison between countries.

After choosing my variable for the models I have performed some Inferential statistics and determined my Confidence Interval.

I have analysed the central tendencies per country:

Table

Description automatically generated

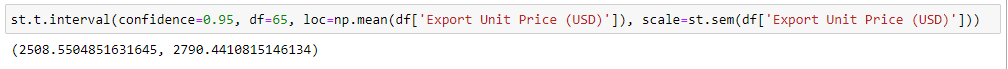
I have checked the probability plot to verify the normality of my distribution:

Chart, line chart, scatter chart

Description automatically generated

Closest to the red Line the better, The data is very close to normality and it will allow me to perform most of the parametrical tests. For the non-parametrical tests, they are more flexible and do not require the distribution to be normal.

Next, I have checked my Confidence Interval based on a 95% confidence.



The values for 95% confidence are between 2508.55 and 2790.44.

## Inferential Statistical Techniques

We have used a combination of Parametric and Non-Parametric Inferential Statistical Techniques.

### T-Test one Way

Null Hypothesis (h0) = The average unit price is equal 2600 USD

Alternative Hypothesis (h1) = The average unit price is different 2600 USD

5% significance level.

Graphical user interface, text, application, email

Description automatically generated

The P-Value is bigger than 0.05 which means I can accept the Null Hypothesis.

In other words, the average unit Price of $ 2600 falls in my 95% confidence interval.

### T-Test Two Way

Null Hypothesis (h0) = Average Export Unit Price for Ireland and Netherlands has no significant difference.

Alternative Hypothesis (h1) = Average Export Unit Price for Ireland and Netherlands has a significant difference.

5% significance level.

Graphical user interface, text, application, email

Description automatically generated

With the P-Value at 0.10 we can consider the Average Export Unit Price for these two countries similar within our stipulated parameters.

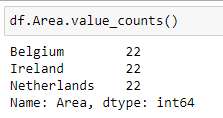
### ANOVA One Way

To be able to use the ANOVA technique we need to follow these 4 requirements:

1. The samples must be independent.

As per our EDA performer earlier we can guarantee that the samples are independent.

1. The groups must have the same sample size.



1. The populations from which the samples were obtained must be normally or approximately normally distributed.

For this we need to perform a Shapiro Wilk test:

P-Value bigger than 0.05 means the distribution is normal or is close to a normal distribution.

Text

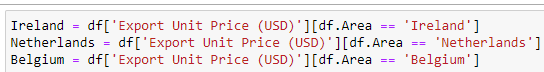
Description automatically generated

As we have analysed before, the distribution for Belgium is not close to a normal distribution, therefore I will proceed using the values for Ireland and the Netherlands.

1. The variances of the populations must be equal.

To test the homogeneity of the variance I have used Levene method.

P-Value bigger than 0.05 means equal Variances:



I have performed a quick analysis before running the Levene Method.

Graphical user interface, text

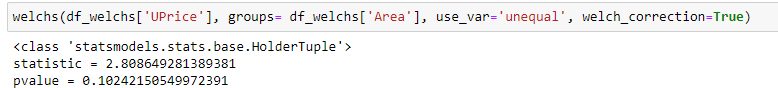
Description automatically generated

The Levene has just confirmed that the only acceptable combination was between the Netherlands and Belgium. But the Belgium distribution failed the normality test, therefore I could not use this combination.

Graphical user interface, text, email

Description automatically generated

The distribution has not passed the requirements, when this happens its recommended to use the ANOVA Welchs. The ANOVA Welchs is an alternative to the traditional ANOVA and it is recommended when the assumption of equal variances is violated.



The p-Value result is bigger than 0.05 which means that the means are equal or without a substantial difference.

### ANOVA Two Ways

We tried to run the Two Way ANOVA to confirm that the result is not satisfactory.

Graphical user interface, application

Description automatically generated

As we can see we could not get good results as we didn’t meet all the requirements for the ANOVA, therefore we will try a non-parametric test.

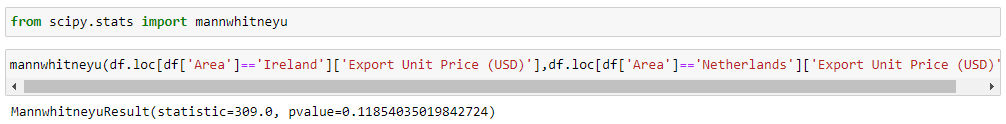
### U-Mann Whitney

The Mann-Whitney U test is a nonparametric statistical significance test for determining whether two independent samples were drawn from a population with the same distribution

Null Hypothesis (h0) = Sample distributions are equal.

Alternative Hypothesis (h1) = Sample distributions are not equal.

5% significance level.



I do not reject h0 because 0.11 is bigger than 0.05. We don’t have sufficient evidence to conclude that the Export unit price distribution between the two countries are different, especially because we are analysing the Export Unit Price and not the Export Quantity.

### Outcome and Findings

Although some of the hypotheses were rejected, the chosen data sample represents a good reference for analysis. Using the top 3 exporters (Excluding Poland), we have covered more than 50% of the total Export Quantity for European Countries. This will be extremely important on the next chapter when I will use prediction models for the Export Unit Price.

Regarding difficulties or challenges, the end dataset was small and I could not find a data sample that would pass the requirement of all the tests. Also, it requires some practice to be comfortable in choosing the right test.

# Machine Learning

# Conclusion

# References

Knisely, K. 2005, *A Student Handbook for Writing in Biology*, 2nd Edition, Smauer Freeman.

Pechenik, J. A. 1997, *A Short Guide to Writing about Biology*, 3rd Edition, Addison-Wesley.

Silyn-Roberts, H. 1996, *A Practical Handbook for Science, Engineering and Technology Students,* Longman.

Harvard Style *UseIt,* Available at <http://www.library.uq.edu.au/training/citation/harvard.html>

Bord Bia, Mushroom Promotion 2019 Available at

[*https://www.bordbia.ie/industry/news/food-alerts/mushroom-promotion-2019/*](https://www.bordbia.ie/industry/news/food-alerts/mushroom-promotion-2019/)

Anove Welch

[*https://www.statsmodels.org/dev/generated/statsmodels.stats.oneway.anova\_oneway.html*](https://www.statsmodels.org/dev/generated/statsmodels.stats.oneway.anova_oneway.html)