# 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](#_Toc123478086)

[Abstract 3](#_Toc123478087)

[1. Introduction 4](#_Toc123478088)

[1.1. GitHub 4](#_Toc123478089)

[1.2. Project Framework 4](#_Toc123478090)

[1.3. Project Background 4](#_Toc123478091)

[1.4. Data collection 4](#_Toc123478092)

[2. Data Extraction and Data Cleaning 4](#_Toc123478093)

[3. Exploratory Data Analysis 4](#_Toc123478094)

[4. Statistical Analysis 4](#_Toc123478095)

[5. Machine Learning 4](#_Toc123478096)

[6. Conclusion 4](#_Toc123478097)

[7. References 4](#_Toc123478098)

# 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

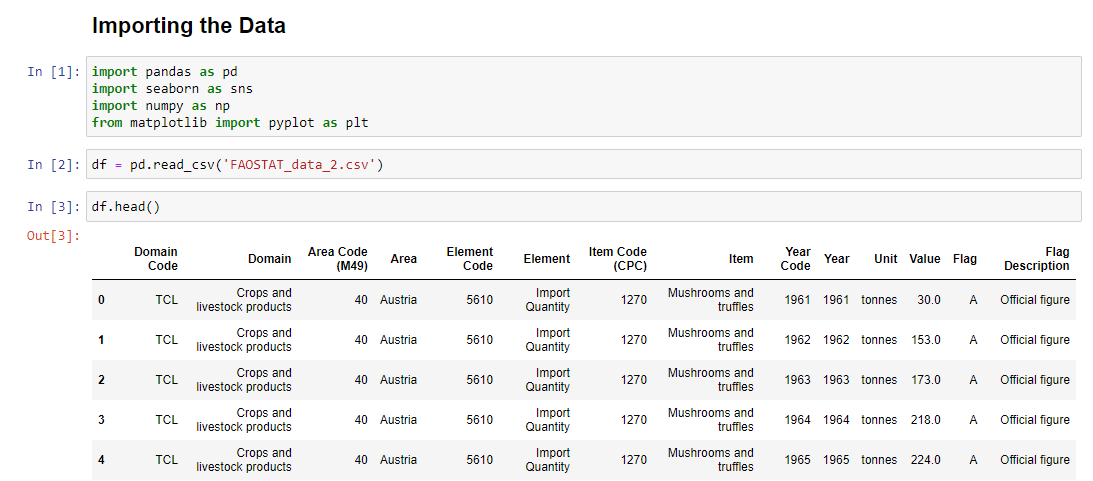
I have downloaded the data as a CSV and that’s the original source provided with the project. (FAOSTAT\_data\_2.csv)

# 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

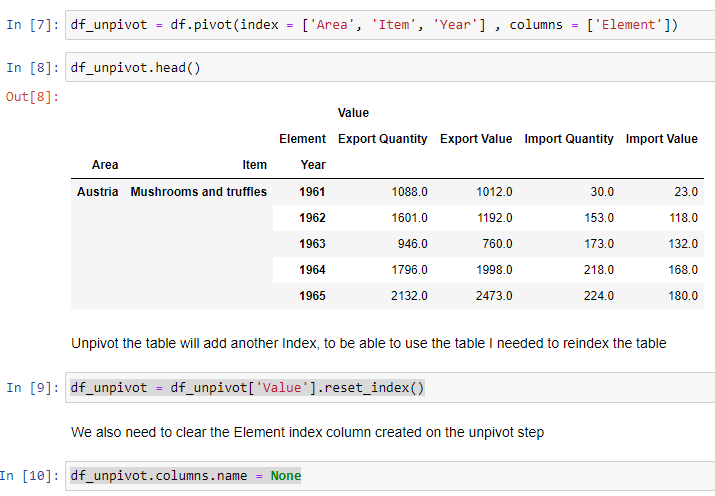
After removing the unnecessary columns, we ended up with the following table:

Graphical user interface, text, application, email

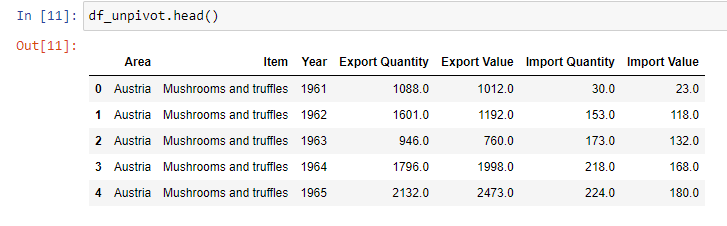
Description automatically generated

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

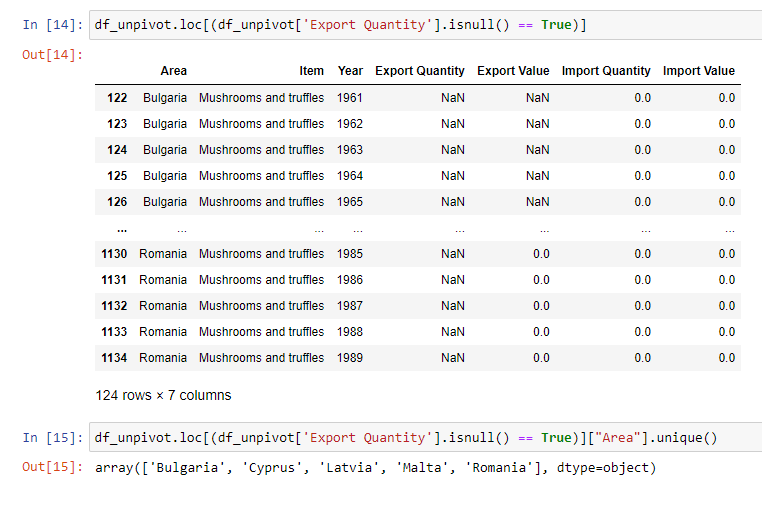
For data cleaning, the first step is to look for null values. Null values happen when the data is missing or is not available. For this assignment I have checked the null values and see if they could be removed or if I need to treat them differently.

Text

Description automatically generated

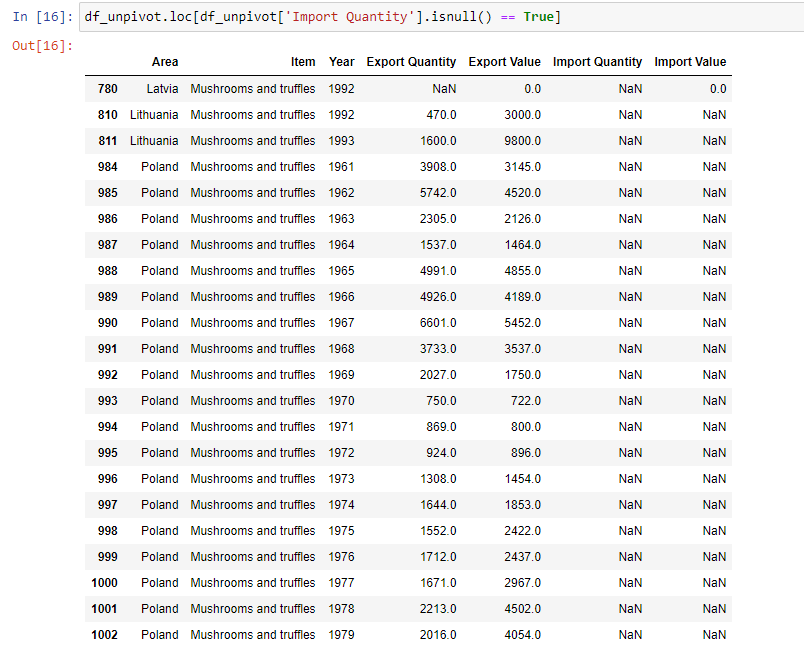
The number of null rows is relatively small when compared with the total data. But I wanted to investigate further to see if removing this data would cause some issues with my analysis later.

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

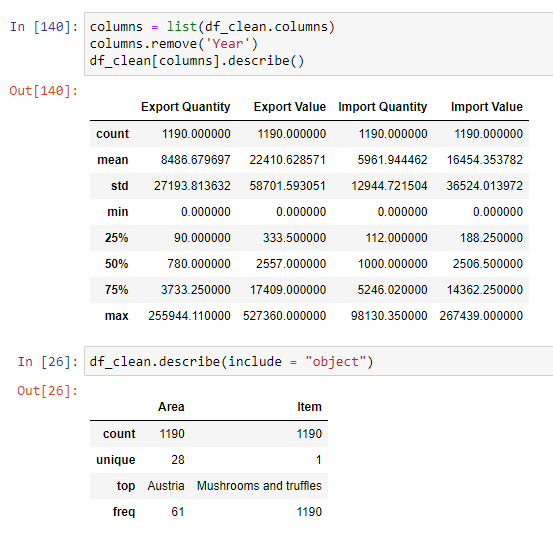
I have started the EDA checking the number of rows and columns of the Dataframe and the data types.

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.

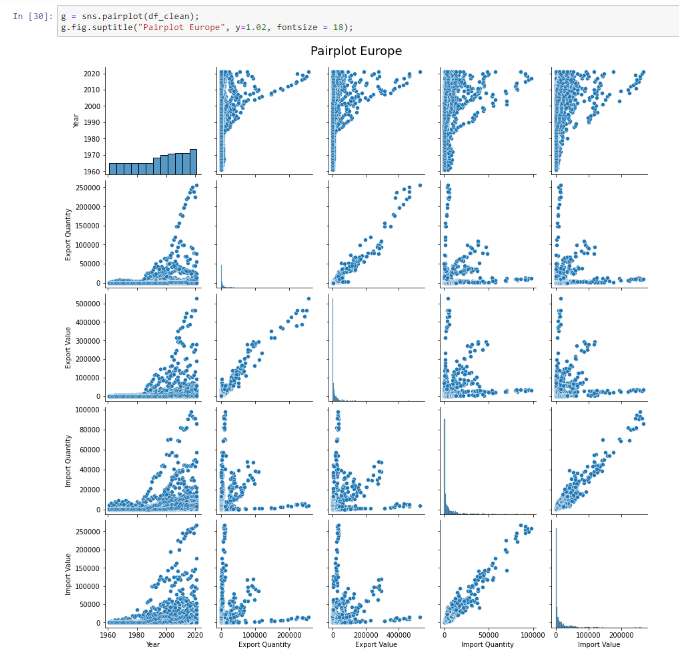
Next step on the EDA is to check the correlation between the columns:

Graphical user interface

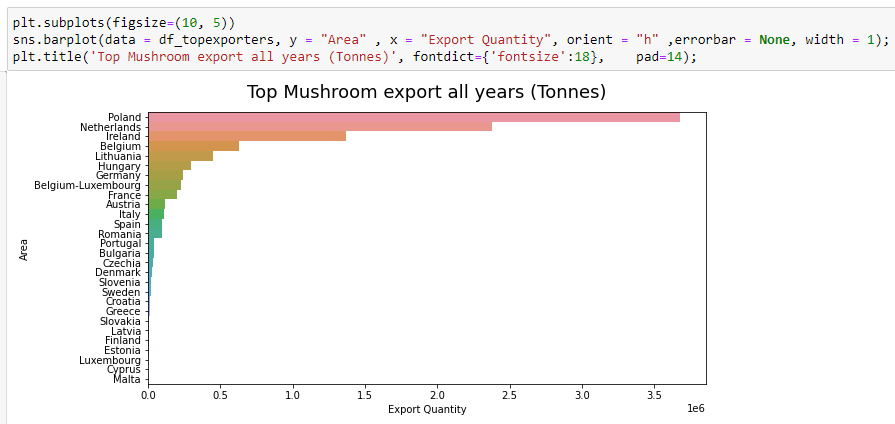
Description automatically generated

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:



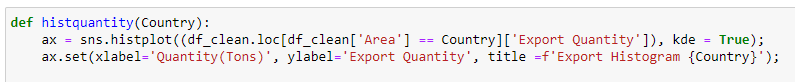
I want to drill down the data into individual countries, so I can compare with Ireland production. I will select the top three exporters of Mushrooms and Truffles from the dataframe.



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

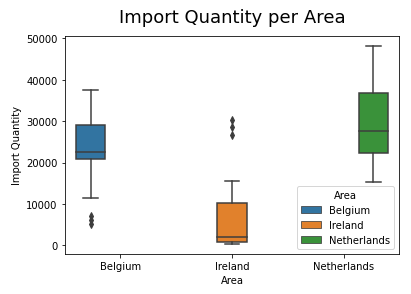
Description automatically generated

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

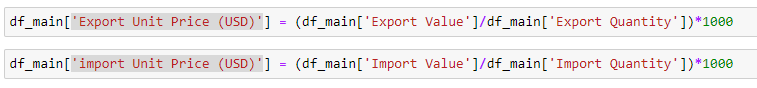
Description automatically generated

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.

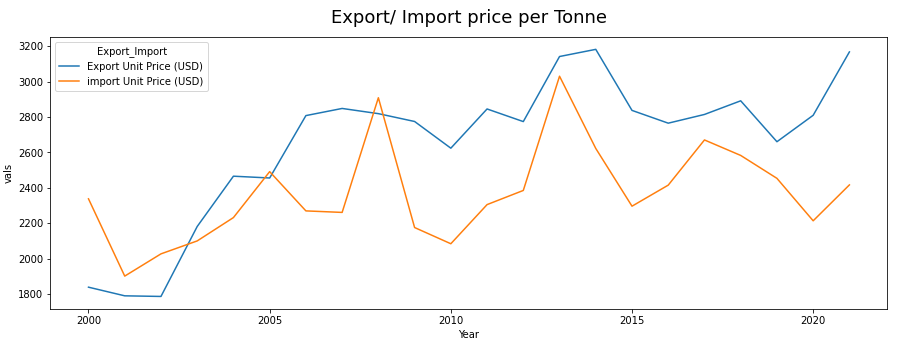


With the addition to the two new columns, we have our new dataframe ready for Machine Learning:

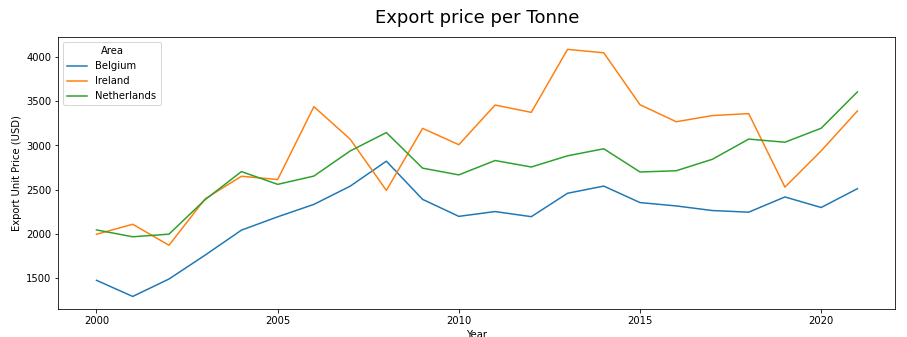
Graphical user interface

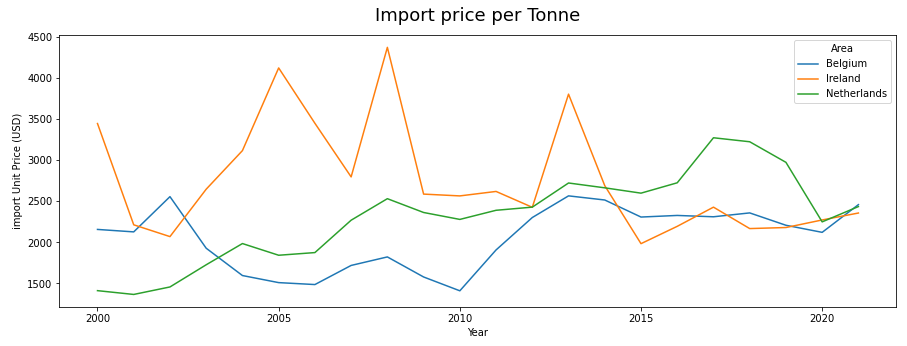
Description automatically generated

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

# 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/)