# Group ID - MSc in Data Analytics

**Author: Joao Ribeiro**

**e-mail:** [**sba22203@student.cct.ie**](mailto:sba22203@student.cct.ie)

**Student ID: sba22203**

**Dublin, 06 January 2022**



## **Acknowledgments**

To my teachers from the Master in Data Analytics at CCT college:

**David McQuaid**

Sam Weiss

Dr Muhammad Iqbal

Marina Iantorno

Words count: 3245

Github: https://github.com/sba22203

**---- Report ----**

***About EU Agriculture***

EU rural population has on average four million people, but also a huge range between minimum of Malta with 22 thousand and Germany with a maximum of 18 million.

In 2020, Ireland had 1.7 million people living of agriculture compare to other countries like Greece of 2.2 million people, or Germany 18.610 million people

EU agric. Production in the last 10 years is on average 12,634 million USD, with a range of minimum of 104 million USD, in Malta and maximum in France 74,030 million USD, that can be explained less or more productive countries.

In 2020, Ireland had a production of 8,454 million USD, compared to similar countries like Greece (15,888 million USD), or Portugal 6,392 million USD.

EU agric. prod. per capita is on average 4,327 USD per person, with a minimum in Malta of 629 USD per person, and maximum in Germany 34,630 USD per person.

In 2020, Ireland productivity was 4758 Usd, compared to Belgium with 29,582 Usd, or Denmark with 13,8666 Usd.

***Relation between values in Countries***

There is a strong relation between Rural population and Agricultural production in the EU, meaning the number of people working in Agriculture will influence the production, which is quite normal.

Using inferential tests, I could confirm strong relationship in values per capita (GDP and agriculture production), meaning that agriculture productivity have a positive impact on GDP and EU population well-being overall.

Especially comparing Ireland with France and Finland, I could constate strong relations.

***Evolution of EU agriculture production***

I’ve compared Agriculture production in Ireland to other countries like France, Germany and Spain, and the all seems very linear, with some up and downs.

Ireland has a very straight line, meaning not too my change in agriculture in the last 10 years, and need to change.

***Suggestion***

Regarding those numbers, especially productivity, Ireland needs to improve its capacity to compete with other markets, maybe better machines and/or improve production processes, and maybe investing in people and in their skills.

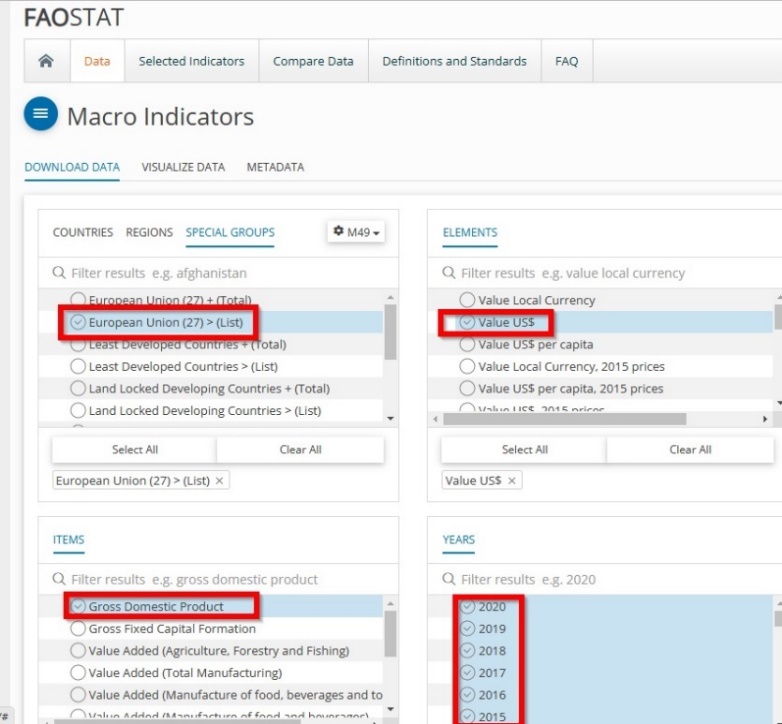
France has a good production overall and Ireland should learn from their production processes, in order to improve Irish Productivity that seems not growing over the last years.

**METHODOLOGY**

# PROCESS OF COLLECTING DATA

I've selected FAO site (<https://www.fao.org/faostat/en/#data>,), because it has lots of information on agriculture and allow us to query data on different indicators like production, population, imports/exports etc and export queries on a csv file and also it’s an open free source.

I've selected different indicators for my study on FAO site, by querying the data I wanted (fig. 1.1)



*Fig. 1.1 Faostat*

I've selected one item to create dataset, in order to have numerical value columns by item, and also selected *on all datasets*, all 27 European countries (not UK), for years 2010 to 2020 (I consider enough), and added:

for g.d.p.csv file

* elements: value in Million USD (FAO doesn't have EURO)
* Items: Gross Domestic Product (GDP)

forurban\_pop.csv file

* elements: Urban population
* Items: Population - Est. & Proj.

forrural\_pop.csv file

* elements: Rural population
* Items: Population - Est. & Proj.

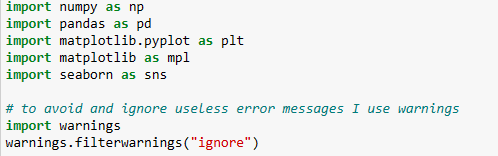
for production.csv file

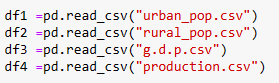
* elements: Gross Production Value (current thousand US$)
* Items: Agriculture

# DATA PREPROCESSING

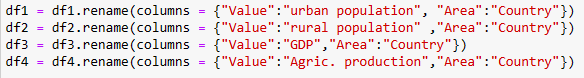
At this stage, I need to pre-process data in order to make it useful for analyse.

Now that I have the dataset, I import libraries to allow me to import, use, transform, and plot data start my dataset analyses





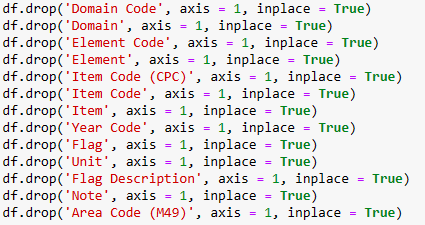
I changed names so I can have item on the columns with numbers instead of value, and also country on area.



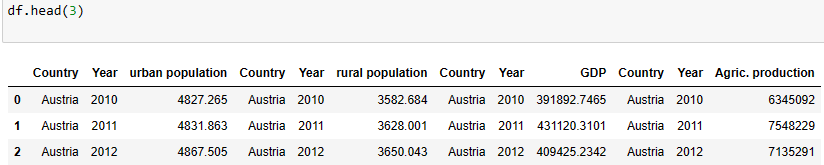
Now merge datasets into a consolidated one, using concatenate, that will join datasets by identical rows (year, country).



I clean the dataset, by deleting useless categorical values in.

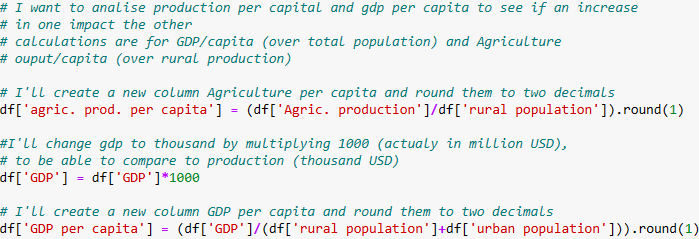


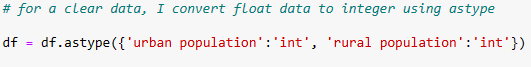
check first 3 lines of the dataset with head, to identify any error and correct it



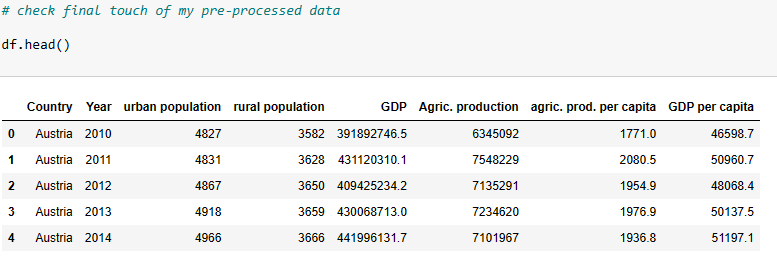
it looks like country and year are duplicated; to solve this issue, I remove duplicate columns





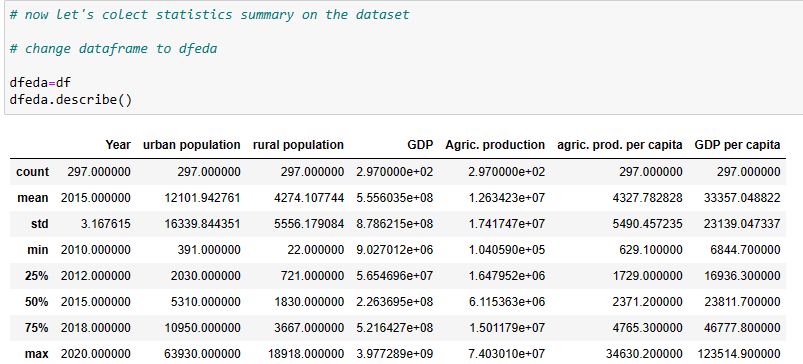


And voila,



# EXPLORATORY DATA ANALYSIS

* 1. Now let's collect descriptive statistics on the dataset



*Urban population*: On average 12 million people. The range shows a huge difference between min (391 thousand) and max (63 million), that can be explained by number of people in small and big countries (for ex: Malta and Germany). For quartiles, I have 2030 in Q1 (under 25% data) or 5310 in Q2 (median and under 50% of the data) and 10950 in Q3 (under 75% of the data).

*Rural population*: shows on average 4 million people, but also a huge range between min (22 thousand) and max (18 million), that can be explained by number of people living in rural area in small and big countries (for ex: Malta and Germany). For quartiles, I have 721 in Q1 (under 25% data) or 1830 in Q2 (median and under 50% of the data) and 3667 in Q3 (under 75% of the data).

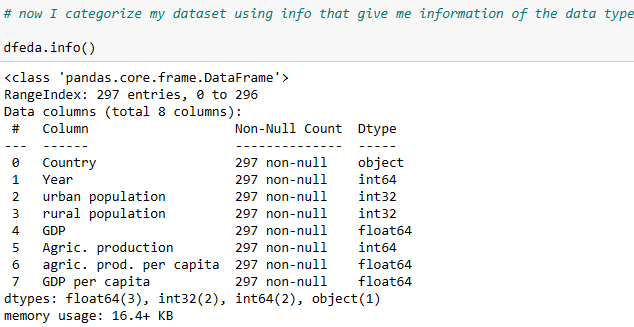
*GDP*: on average 555 million USD, with a range of min (9,027 million USD) and max (3,977 billion USD), that can be explained richer or poorer countries (for ex. poorest Malta and richest Germany). For quartiles, I have 5.6 in Q1 (under 25% data) or 2.2 in Q2 (median and under 50% of the data) and 5.2 in Q3 (under 75% of the data).

*Agric. production*: on average 12,634 million USD, with a range of min (104 million USD) and max 74,030 million USD, that can be explained less or more agric. productive countries (for ex. less productive Malta and Germany, as more productive). For quartiles, I have 1.64 in Q1 (under 25% data) or 6.11 in Q2 (median and under 50% of the data) and 1.5 in Q3 (under 75% of the data).

*GDP per capita*: on average 33 thousand USD per person, with a range of min (7 thousand usd per person) and Max (123 thousand USD per person), and the same as GDP can be explained between richer or poorer countries (for ex. poorest Malta and richest Germany). For quartiles, I have 16936 in Q1 (under 25% data) or 23811 in Q2 (median and under 50% of the data) and 46777 in Q3 (under 75% of the data), and it has also due to poorer or richer countries like Malta or Germany).

*Agric. prod. per capita*: on average 4,327 USD per person, with a min (629 USD per person), and max (34,630 USD per person). For quartiles, I have 1729 in Q1 (under 25% data) or 2371 in Q2 (median and under 50% of the data) and 4765 in Q3 (under 75% of the data), and it has also due to poorer or richer countries like Malta or Germany)

Standard deviation is high for all measures, except Agric. Production (1.74) and GDP (8.89), meaning the values are dispersed from the mean.

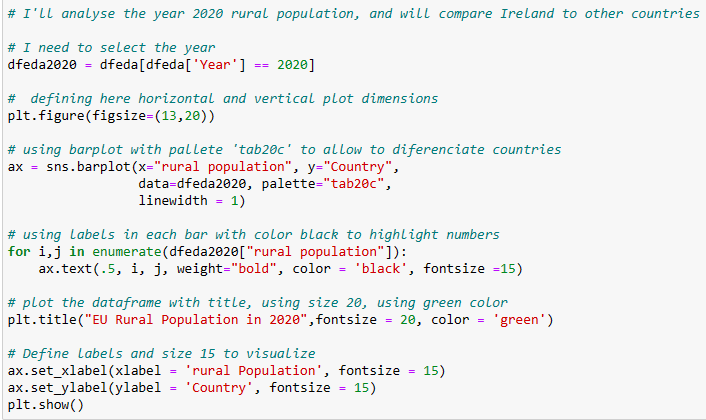


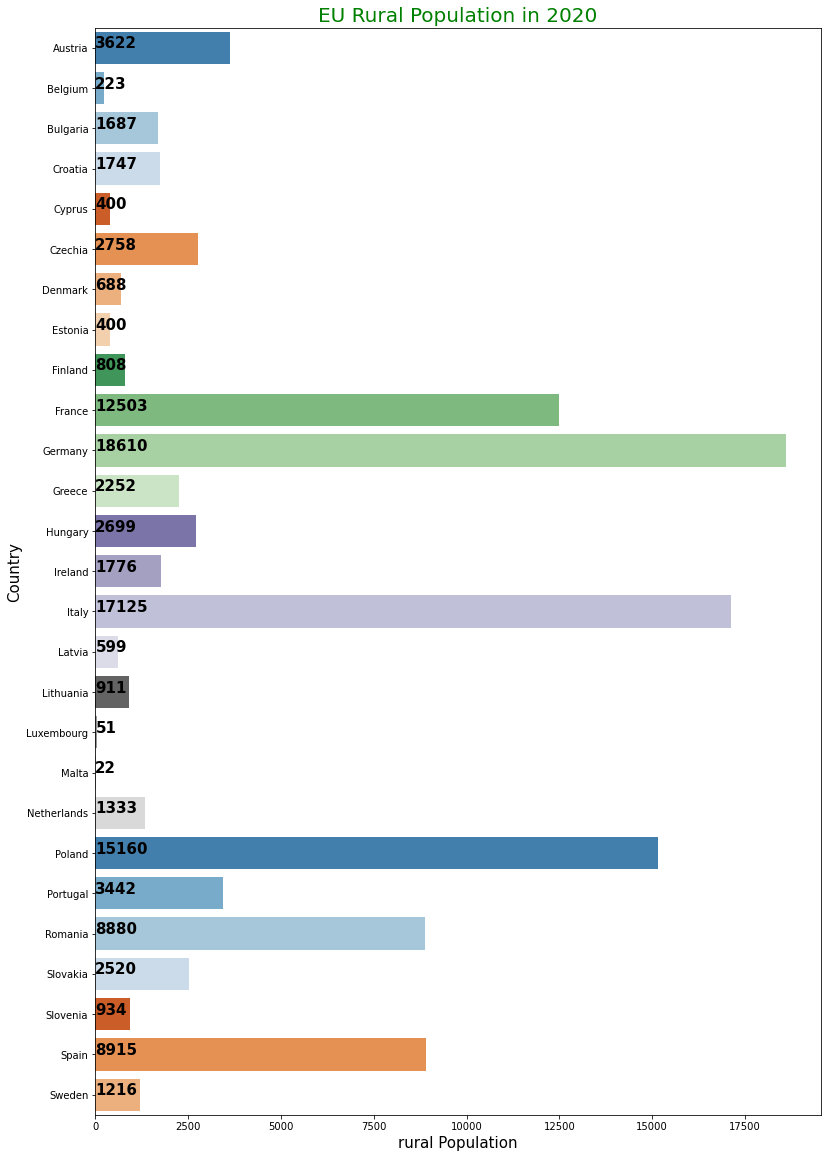
So, my dataset is defined:

* there are 297 rows in all columns, with no null values,
* 4 integer type,
* 3 decimals numbers
* 1 one object (country)

3.2 EDA Univariate Analysis

In this Analysis, data visualization is realized on one variable at the time, and I need to decide what charts to plot to better understand the data.

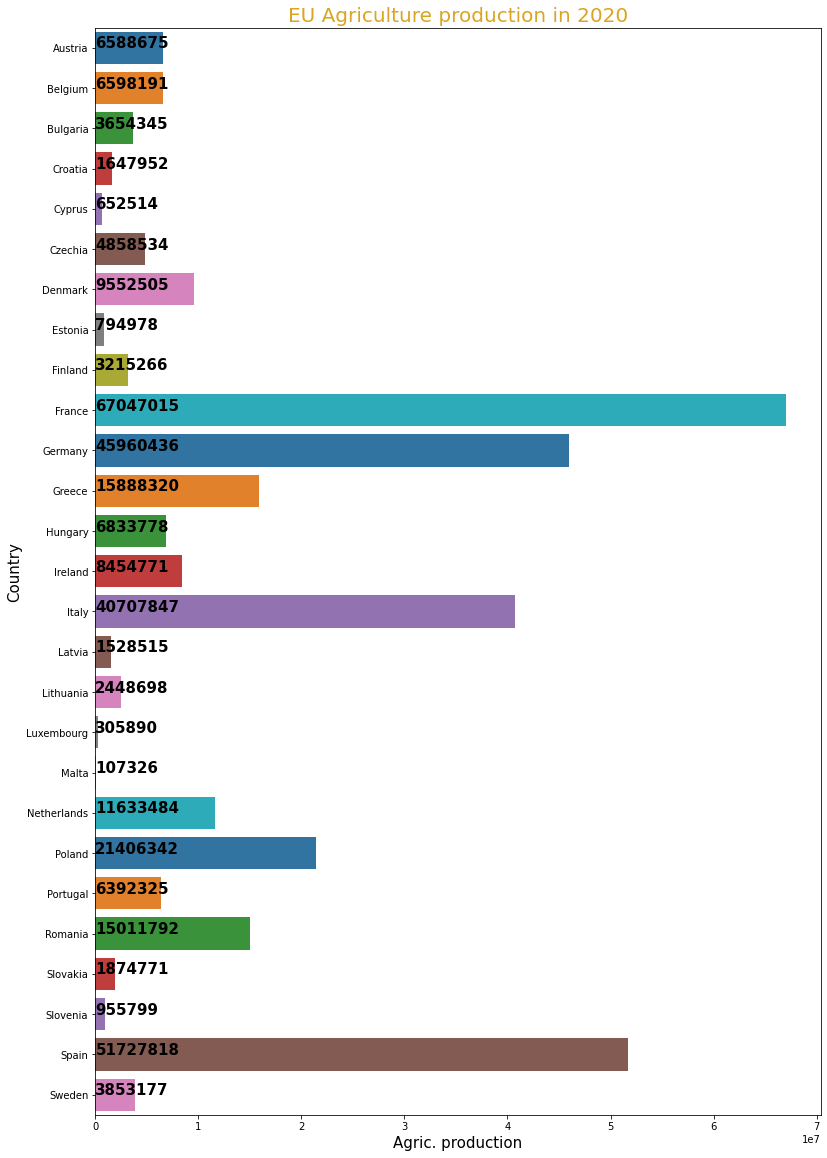




*Fig. 3.1 EU rural Population in 2020 (thousand people)*

as we can see on fig. 3.1, there are more people living on agriculture in Germany (18,610), followed by Italy (17,125) compared to Ireland (1,776), because of the size of the country.

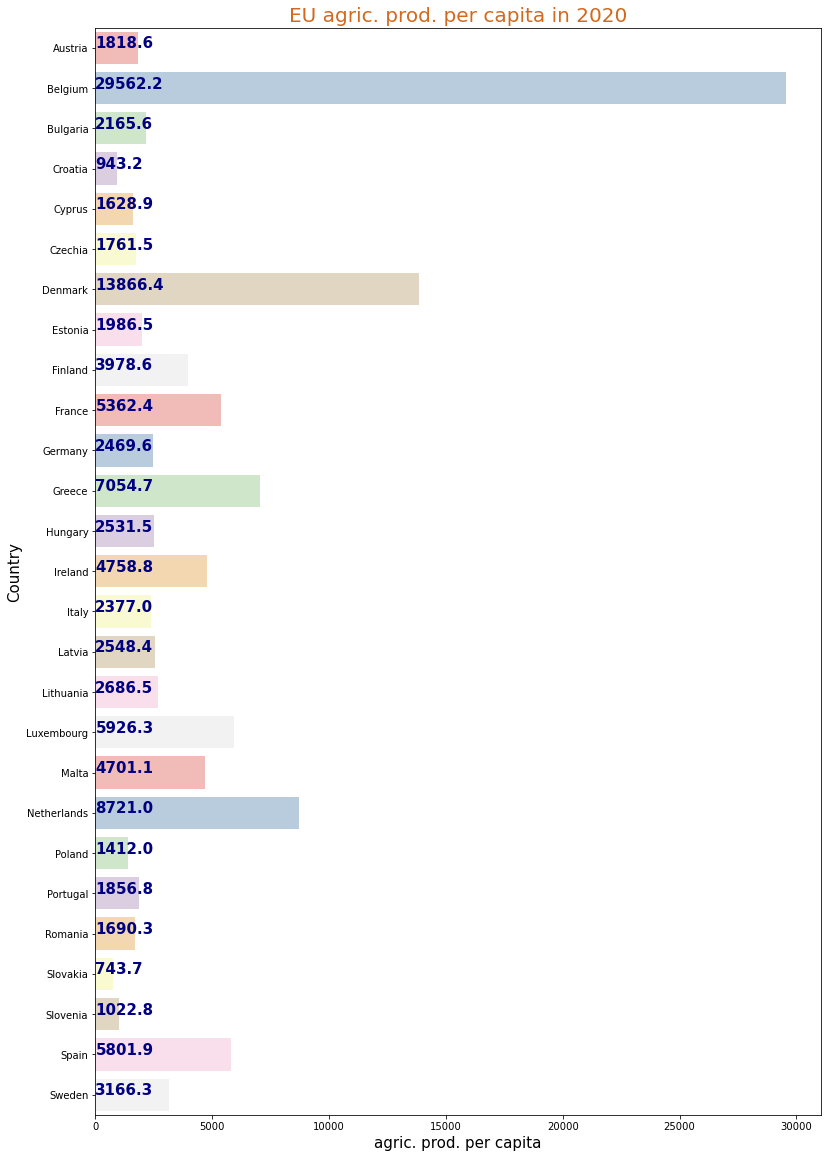




*Fig. 3.2 EU Agriculture production in 2020 Thousand USD*

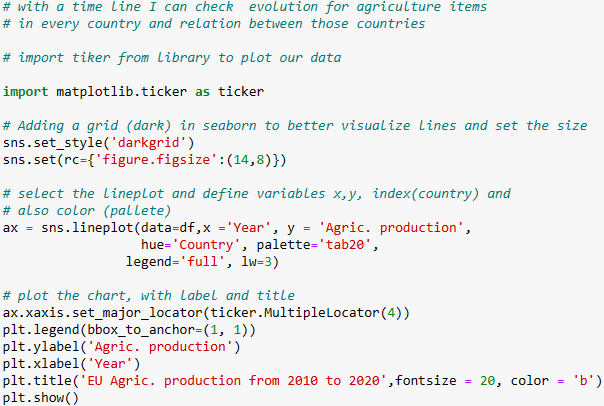
we can see on fig. 3.2, that France a good production, in the first place before Germany, with less people working in agriculture. Ireland has a smaller production considering the size of the country.

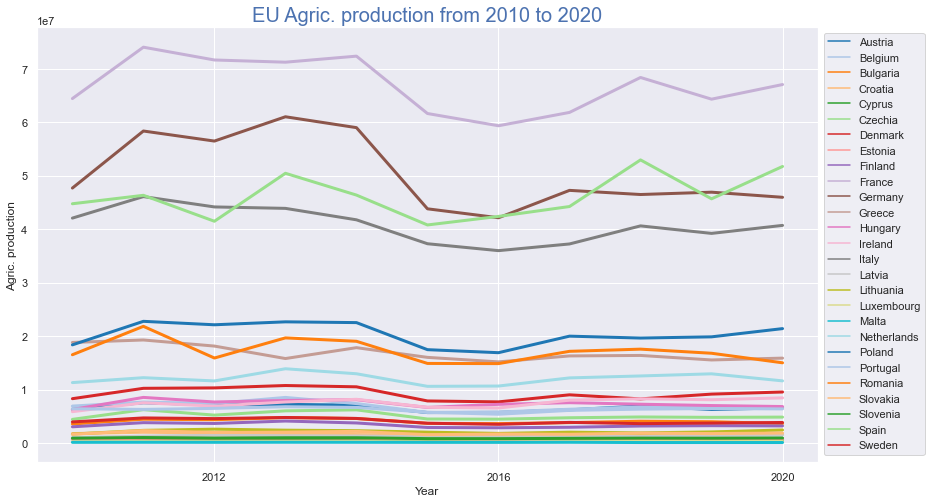
**



*Fig. 3.3 EU rural productivity in 2020*

As we can see on fig. 3.3, rural productivity at the highest in Belgium (29,582 usd), followed by Denmark (13,866 usd) compared to Ireland (4,758 Usd), showing that Ireland needs to improve its capacity to compete with other markets, maybe better machines and/or improve production processes (people, machines).

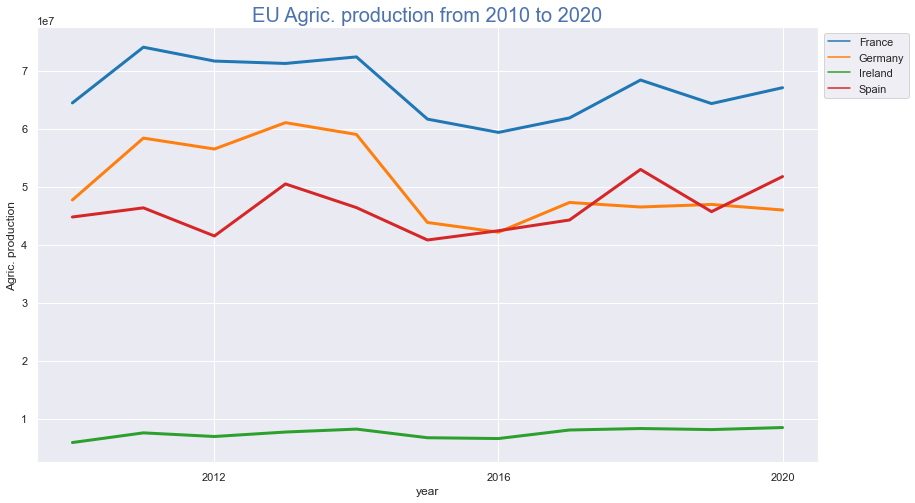




*Fig. 3.4 EU Agric. production from 2010 to 2020*

The fig. 3.4 is quite confusing with all those countries, and I'll only compare Ireland to top countries

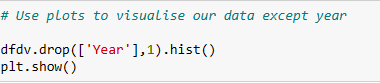
**

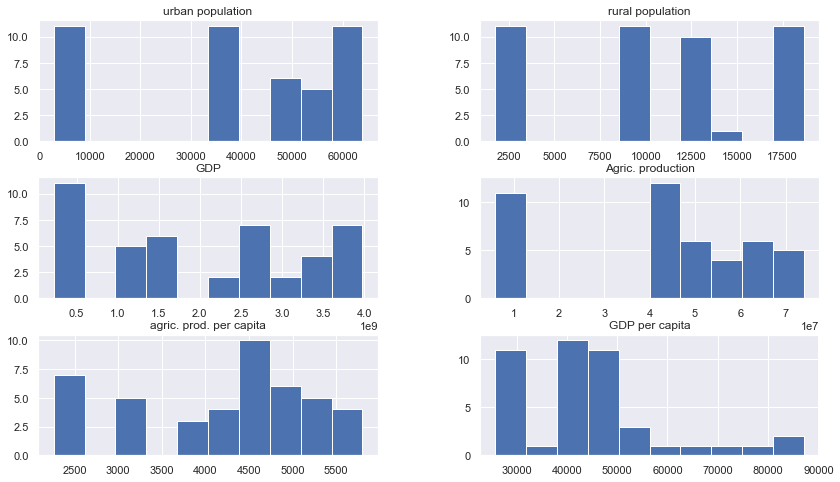


*Fig. 3.5 EU countries Agric. production from 2010 to 2020*

On fig. 3.5, France (blue), followed by Spain (red) and Germany (orange) is have a production very consistent with some up and downs.

Ireland production is below those countries, Ireland Production is also linear, not improving a lot.





*Fig. 3.6 EU countries in 2020, by measures*

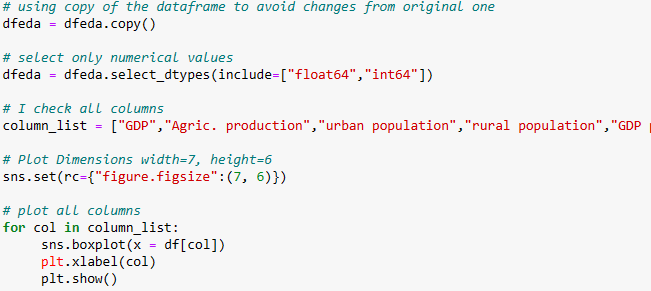
On fig. 3.6, some measures are concentrated on some areas of the histograms:

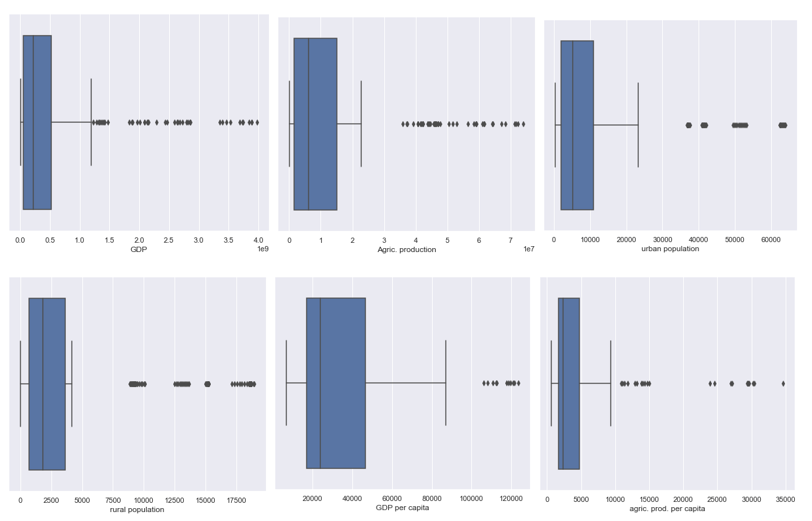
* + urban population 50000-60000.
  + GDP 2 to 4,
  + Agric. Production above 4 to 7,
  + Rural is dispersed,
  + Agric. Production per capita above 3750,
  + GDP per capita above 40000-50000,

**OUTLIERS**

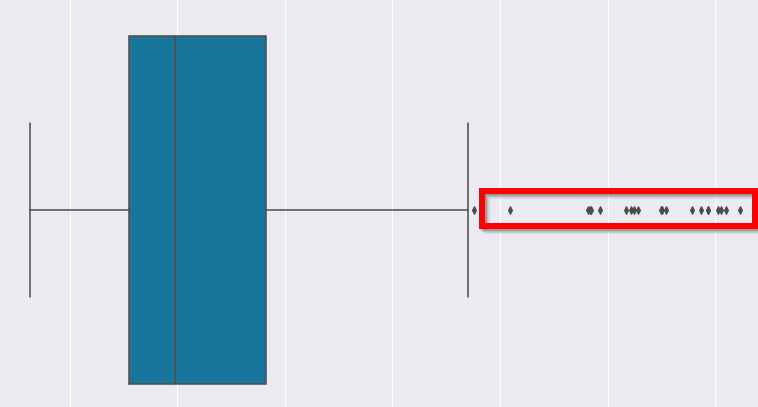
now I check for outliers using Inter Quartile Range and plots, it will allow to see values outside normal distribution and correct dataset if necessary

# BOX PLOTS





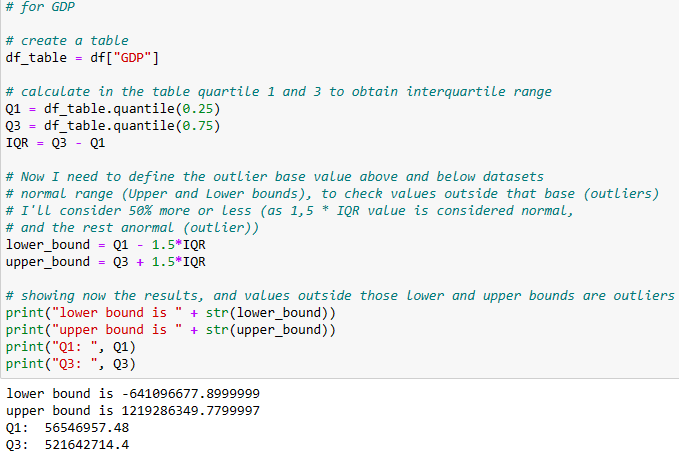
*Fig. 3.7 Boxplot (2020)*



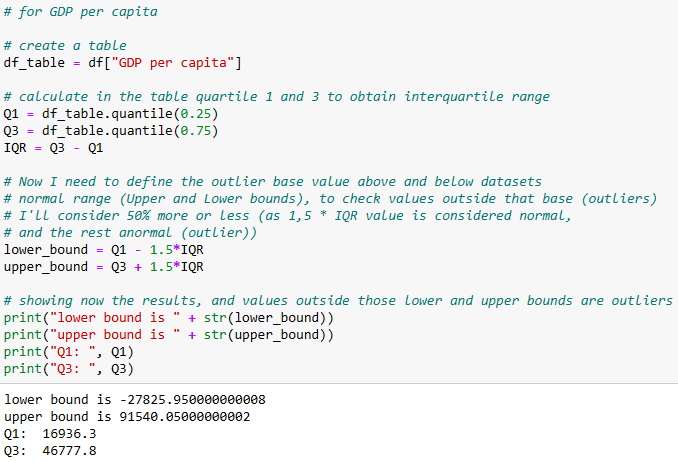
*Fig. 3.8 Boxplot (2020)*

On fig. 3.7 and 3.8, I can see that all columns have outliers represented by plots outside the boxplot, all above (upper bound)

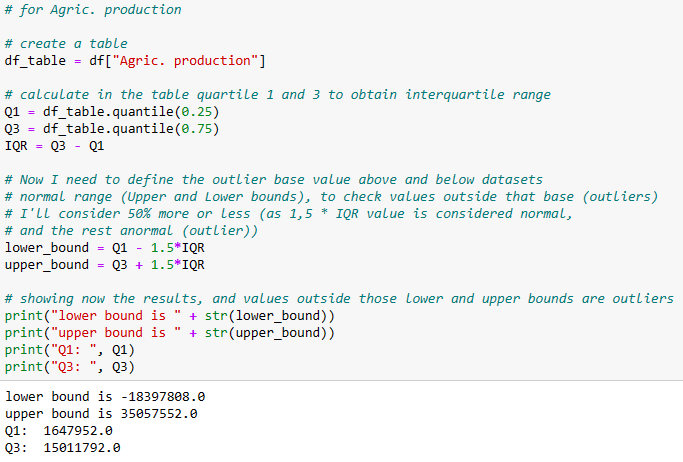
# INTER QUARTILE RANGE



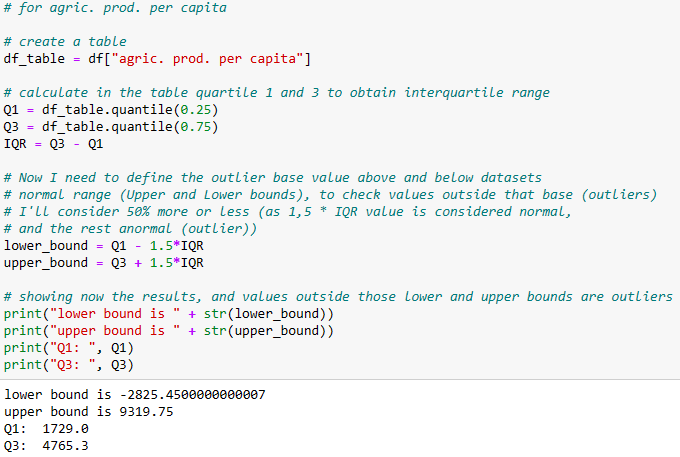
Outliers are above upper bound of 1219286349.0.



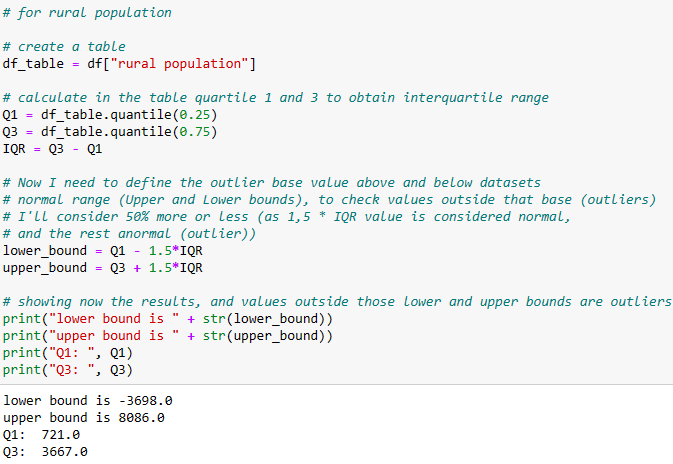
Outliers are above upper bound of 91540.0.

**

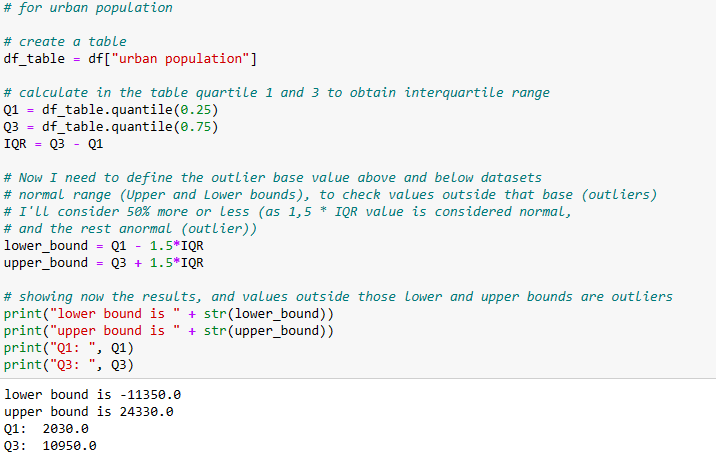
outliers are above upper bound of 3505755.



Outliers are above upper bound of 9319.75, because in the boxplot, we can see clearly there are above the plot.



Outliers are above upper bound of 8086.0.



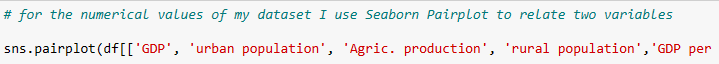
Outliers are above upper bound of 24330.0.

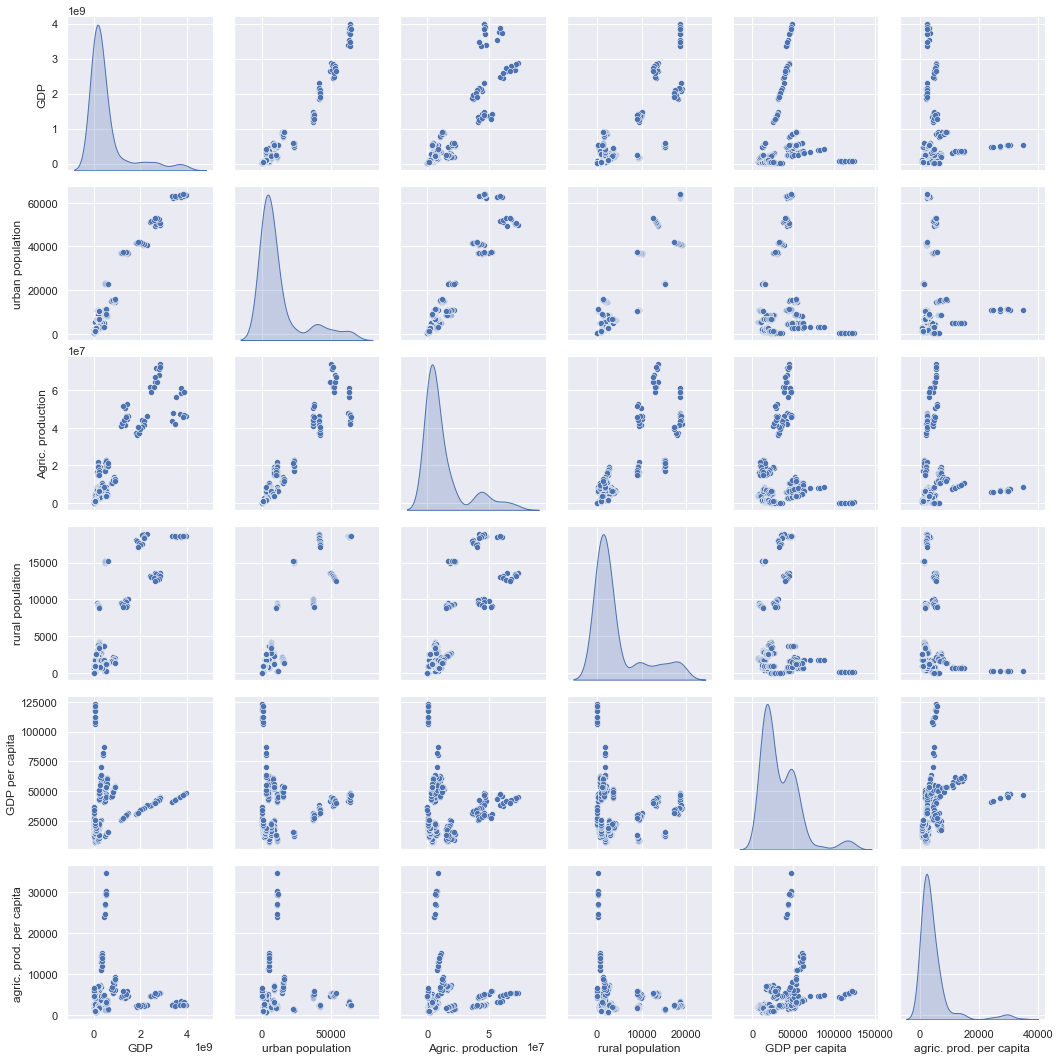
**DEALING WITH OULIERS**

Now that I know where are the outliers, I can delete them using IQR and function drop, however when I did a lot of information was deleted, my plots comparing Ireland with other EU countries were wrong and inconsistent (some countries not showing values during several years), so I decided to not delete outliers for that reason.

# 3.3 EDA BIVARIATE ANALYSIS

Through this analysis, I'll understand how variables are related to each other in the dataset.





*Fig. 3.9 Relations analyses*

In fig. 3.9, between GDP and other variables don't have a very linear relationship, most linear to urban population

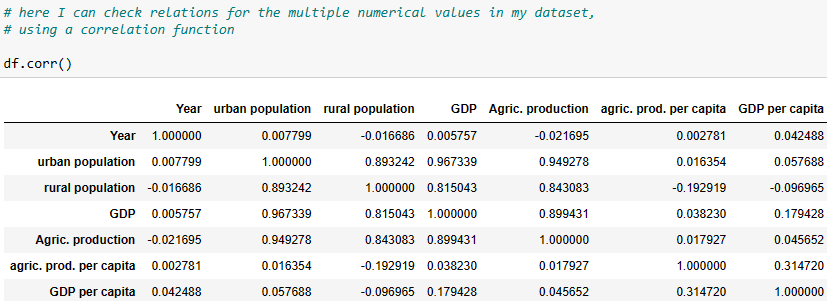
Between Urban population and the other variables don't have a very good linear relationship, as the values seem very disperse in the plot, however with more concentrated values in some areas.

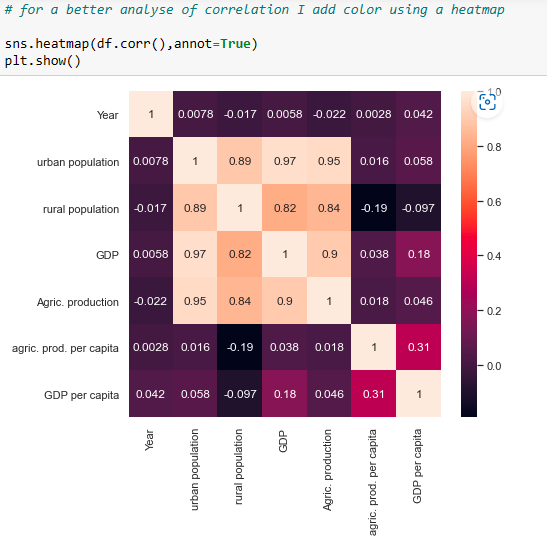
Between Agric. production, rural population, gpd per capita, agric. prod. per capita and the other variables don't have a very good linear relationship, as the values seem very disperse in the plot. however, the two variables per capital seems to have an area of relationship

In all the plots, the data seems disperse, having some few areas with concentrated values and relationships

# 3.4 EDA Multivariate Analysis

with this analyse I look at more than two variables, to determine relationships and analyse patterns for any dataset

**



*Fig. 3.10 heatmap*

In fig. 3.10 the higher the correlation, the better relation between two variables. we can see that most correlation are positive, so the variables relate in the same way.

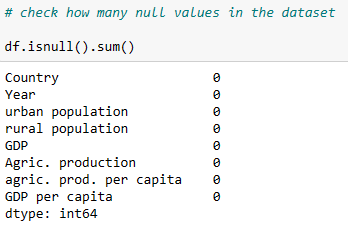
The urban population and Agri. prod. is 0.95 and also GDP (0.97), meaning positive and strong correlation between them.

Rural population and Agri. prod is 0.84 and also urban population 0.89 are a positive and medium relationship.

GDP and urban population are strongly related 0.97 or with Agric. prod. (0.90).

# 3.5 EDA - missing values

I need to check if there any missing values (NaN or blank), and if so, to impute missing values to my dataset, by their Mean, median for example

**

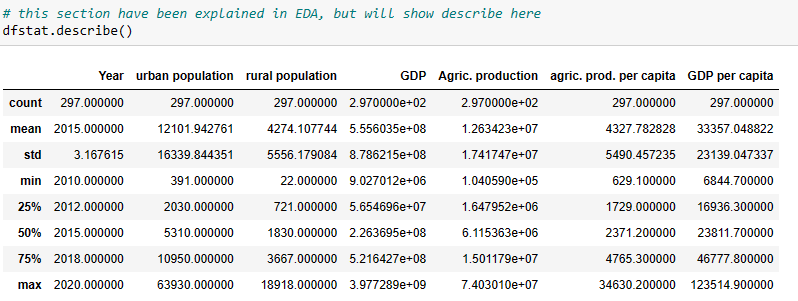
There're no null values.

# STATISTICS

****

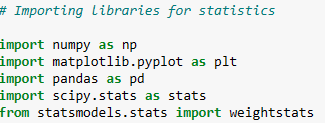
**4.1 Descriptive Statistics**

some have been explained in paragraph 3, like describe function, Histograms, bar plots, line plots, and box plots.



**4.2 Inferential statistics and visualisations**

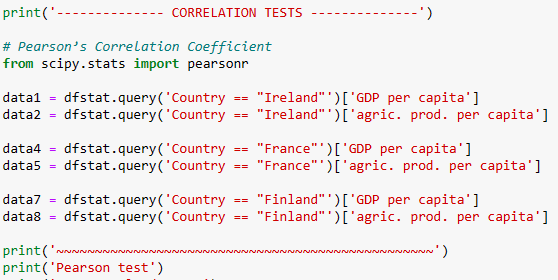
**4.2.1 Comparing countries using several statistical tests**

**

in the plot (*Fig. 3.3 EU rural productivity in 2020),* I’ve compared some countries and some productivity in countries similar to Ireland.

* Ireland 4758
* Finland 3978
* France 5362

I now compare them with inferential tests:



## I used correlation Tests to check how *GDP per capita* and *agric. prod. per capita* per country (samples) are related.

## For p value, I need to consider null hypothesis:

## The p-value (under 0.05) is statistically significant

## p-value higher than 0.05 is not statistically significant and indicates strong evidence for the null hypothesis.

### ***Pearson’s Correlation Coefficient***



The samples have a high correlation (prob. Dependent).

the strongest is France (0.791), followed by Finland (0.791) compared to Ireland (0.642). The p-value tell us the sample is statistically significant.

### ***Spearman’s Rank Correlation***



To test my samples’ monotonic relationship, that have a high correlation (prob. Dependent).

the strongest is Finland (0.809), followed by France (0.800) compared to Ireland (0.682). The p-value tell us the sample is statistically significant.

### ***Kendall’s Rank Correlation***

### 

To test my samples’ monotonic relationship, that have a high correlation (prob. Dependent). the strongest is France (1), followed by both Finland and Ireland (0.527). The p-value tell us the sample is statistically significant.

### ***Chi-Squared Test***



To test my samples’ monotonic relationship, that have a high correlation (prob. Dependent). the strongest is Ireland (1271.3), followed by Finland (325.7) and Ireland (80.2). The p-value tell us the sample is statistically significant.

### ***Student’s t-test***

### 

Now I test if the means of two independent samples are significantly different which is confirmed here (probably different distribution).

### ***Paired Student’s t-test***

### 

Now I test if the means of two independent samples are significantly different which is confirmed here (probably different distribution).

### ***Analysis of Variance Test (ANOVA)***

### 

Now I test if the means of two independent samples are significantly different which is confirmed here (probably different distribution).

## **Nonparametric Statistical Hypothesis Tests**

### *Mann-Whitney U Test*

### 

I test if the distributions of two independent samples are equal or not, and the result is independent confirmed here (probably different distribution).

### *Wilcoxon Signed-Rank Test*

### 

Tests whether the distributions of two paired samples are equal or not, and the result is independent confirmed here (probably different distribution).

### *Kruskal-Wallis H Test*



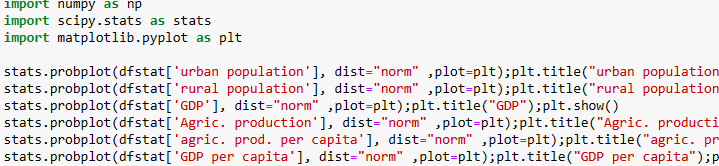
Tests whether the distributions of two or more independent samples are equal or not and the result is independent confirmed here (probably different distribution).

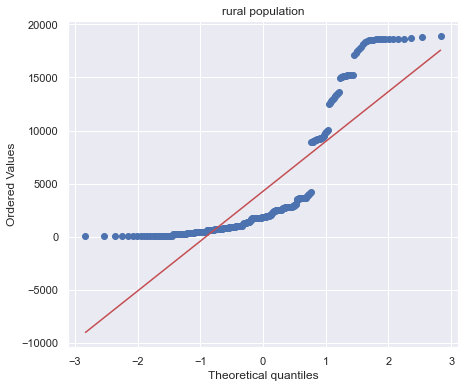
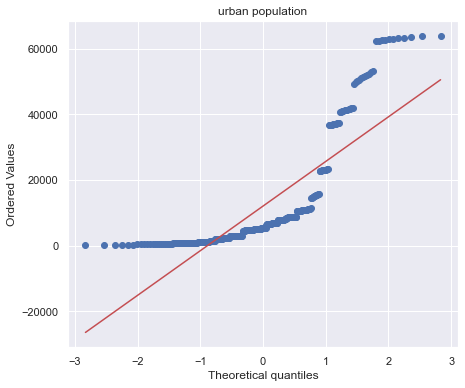
### ***Friedman Test***

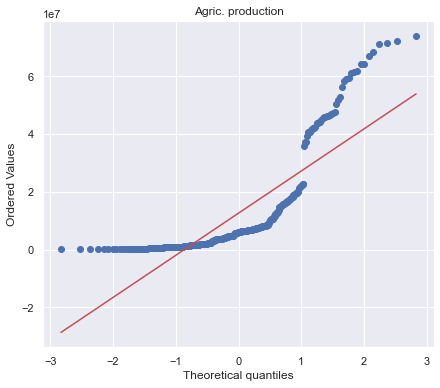
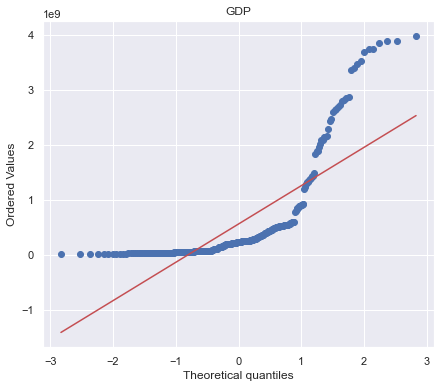


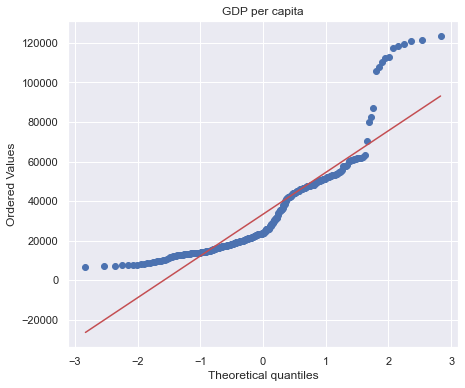
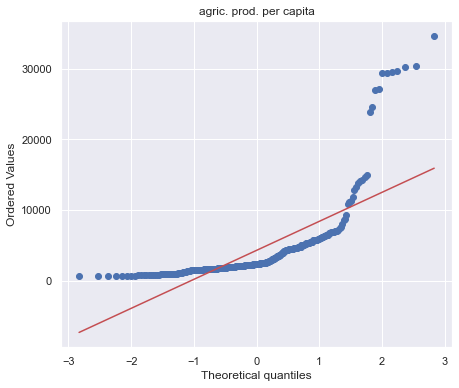
Tests whether the distributions of two or more paired samples are equal or not, and the result is independent confirmed here (probably different distribution).

# 4.2.2 ANOVA Test

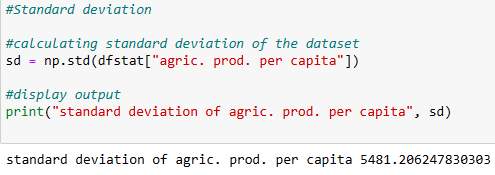




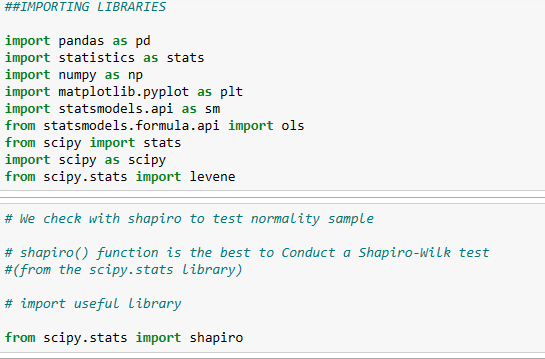




*Fig. 3.11 Anova Subplots*

**

Standard deviation for Agriculture is high (5481.2), and we can see the values are dispersed from the average of 4328 (vide parag. 3, for mean in describe function).

**



With the Shapiro-Wilk test, I will check samples’ normality, determining if the sample comes from the normal distribution or not, and also assume a significance level at 0.05 (i.e. 95% confidence interval).

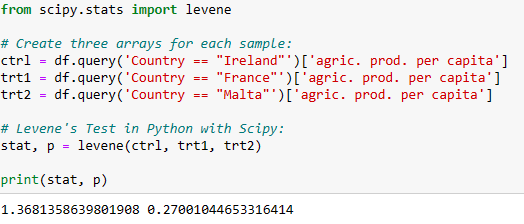
For Ireland (0.21) and Finland (0.26), both p-value> 0.05, then I fail to reject the null hypothesis, meaning I do not sufficient evidence to say that sample does not come from a normal distribution.

For Malta (0.03) and France (0.03), both p-value < 0.05, then we reject the null hypothesis meaning we have sufficient evidence to say that sample does not come from a normal distribution.



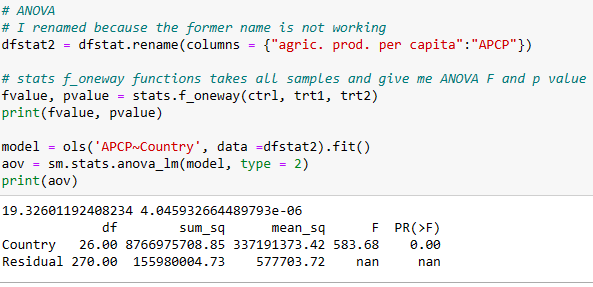
And now I’ll use Bartlett’s test t**o check if the variances between several groups are equal or not, in this case using samples “Agri. Prod. Per capita by Country”**

P-value for the group (0.048) < 0.05, then we can reject the null hypothesis and conclude that not all groups have the same variance.

****

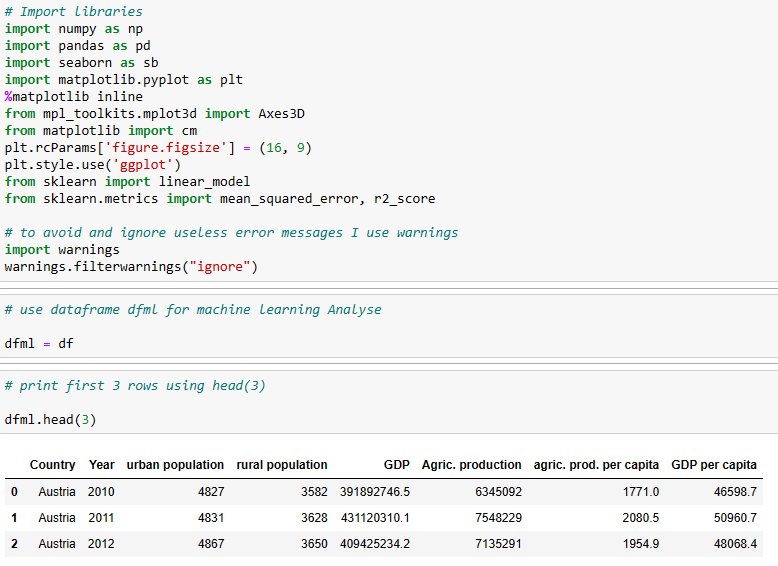
I’ll use Levene’s test by checking samples to have or not same variances.

P-value for the group (0.27) > 0.05, then I fail to reject the null hypothesis, meaning I have sufficient evidence to say that samples variances are not equal.

****

I can see that p value is 4.045932664489793e-06, meaning 0.000004045932664489793 (p < 0.05), and therefore, then we can reject the null hypothesis and conclude that there are significant differences among samples.

# MACHINE LEARNING



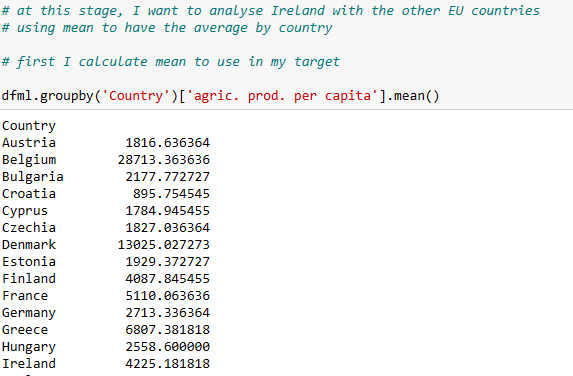


My dataset is made of 297 rows and 9 columns

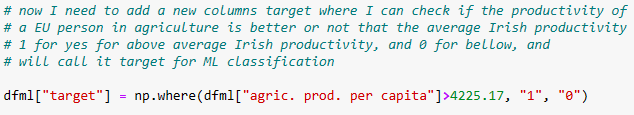
**5.1 MACHINE LEARNING ALGORITHMNS COMPARISON**

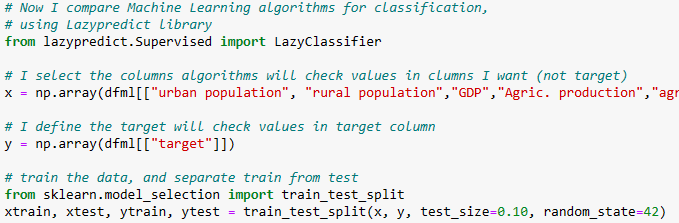
# 5.1.1 Classification

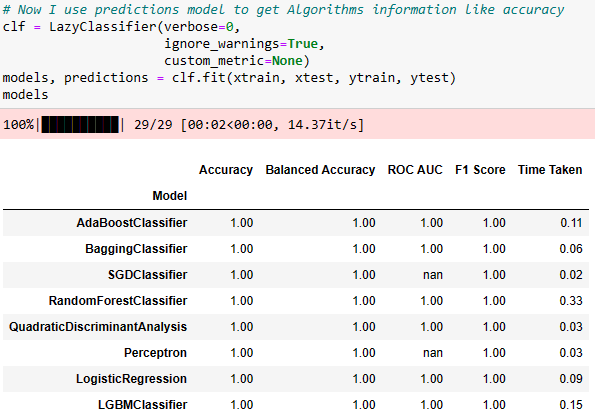
at this stage, I want to analyse Ireland with the other EU countries



For Ireland, the average for is 4225.17 and I want to compare it to other EU countries to see if they're above or below Irish Agric. Productivity.

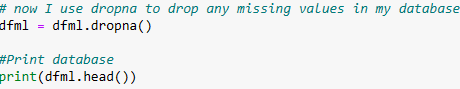


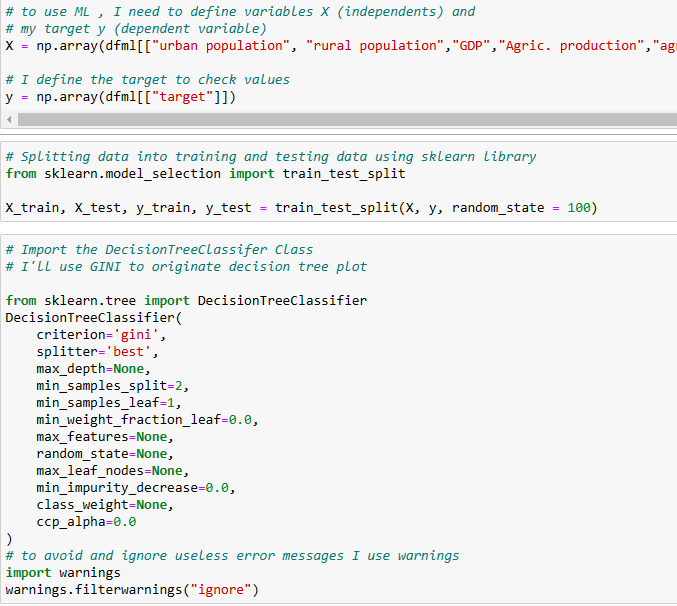


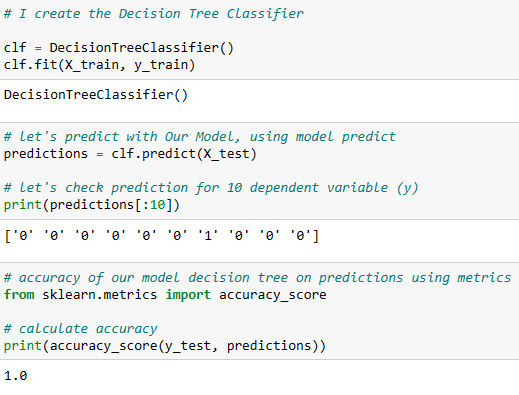


Lazypredict has sorted all algorithms accuracy and:

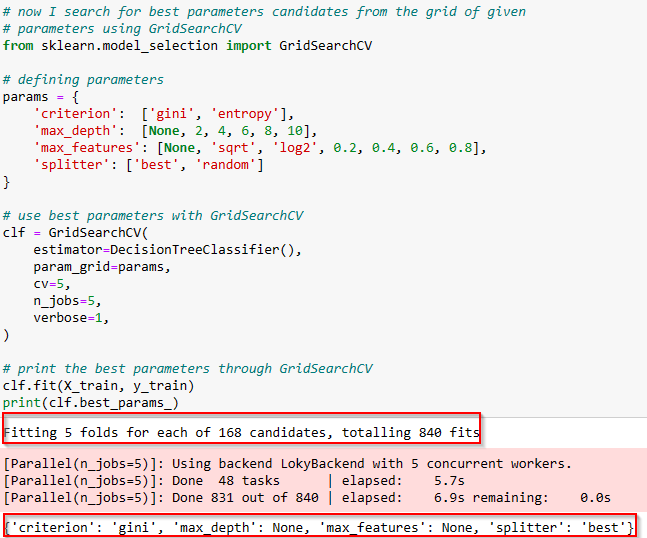
* perfect accuracy (1 or 100%) on the 13 from AdaBoostClassifier to ExtraTreeClassifier
* almost perfect accuracy (0.97 or 97%) on the 9 following from GaussianNB to PassiveAggressiveClassifier



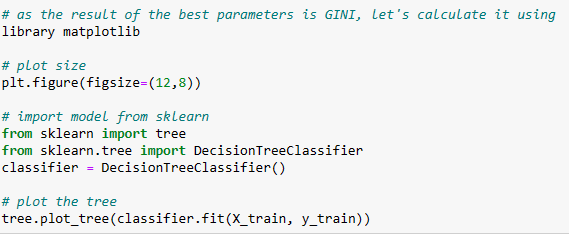


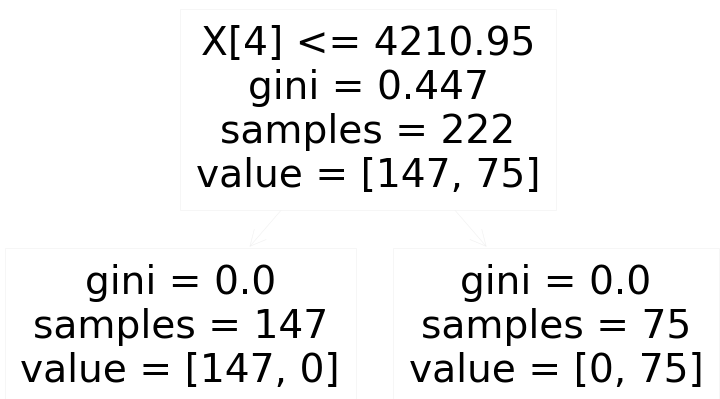


I reconfirm now that decision tree is accurate 100% is the same as lazypredict

**

The best parameter is GINI using gridsearchCV, so I will use it.

**

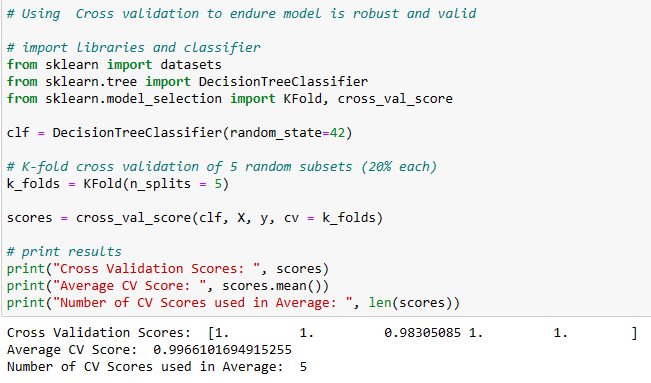
**

Gini index in decision tree provide the amount of probability of a specific feature that is classified incorrectly when selected randomly, and in our case

when Gini = '0', it is a leaf node with only one class (pure), and we can see that on that here on both leaves on the bottom of the results

We can see a total of 222 samples (147+75), with a Gini impurity 0.447 (almost the maximum of 0.5)

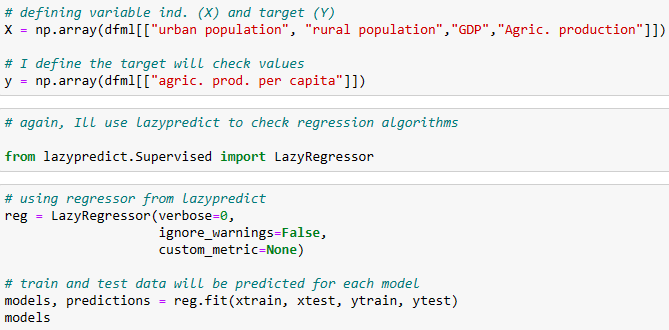
We can see the root node as skewness of the wavelet, of <=4210.95, we have 147 samples, the 75 are above that value; So, for condition <=4210.95 TRUE we have calculated GINI = 0.447 (impure)

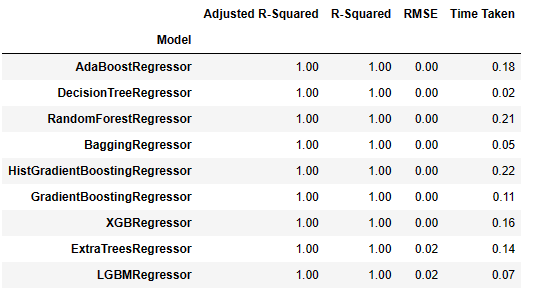


for the 5 subsets, almost all of them have 100% validation (one is 98.3%), and on average 99.66% meaning an excellent score of validating my tree decision model

**5.1.2 Regression**

After classification I will do some regression for Machine Learning, as my dataset is ideal for forecasting



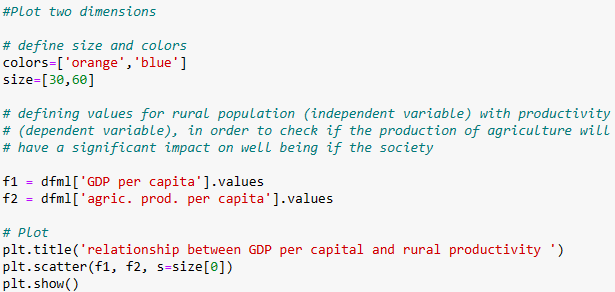


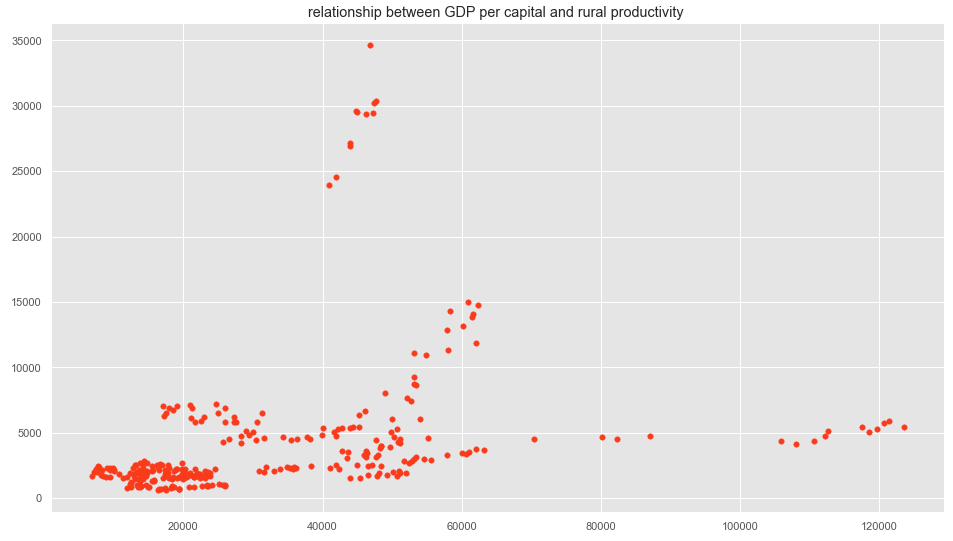
Lazypredict has sorted all algorithms accuracy and:

* perfect accuracy (1 or 100%) on the 9 from AdaBoostClassifier to LGBMRegressor
* almost perfect accuracy (0.83 or 83%) on the SVR and nuSVR

for my study I will use linear regression, because it has a high accuracy 73%

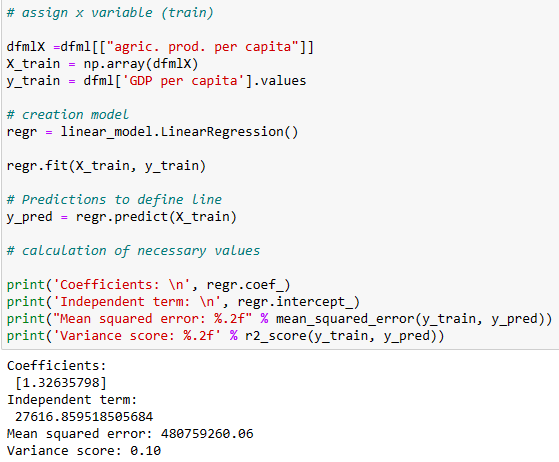
# forecasting

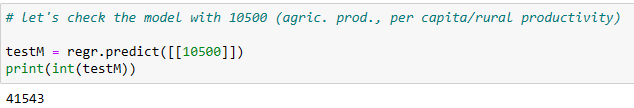




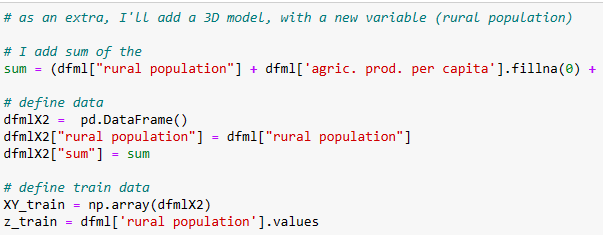
*Fid. 5.1 relationship between GDP per capita and rural Productivity*

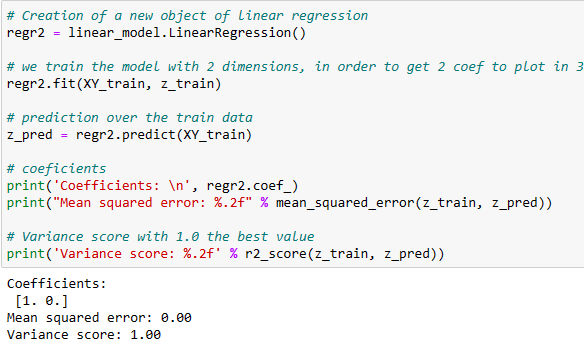
the area from 0 to 60000 (GDP per capita) and 0 to 10000 (agri. proc. per capita) is where the values are concentrated. some values are outside this area (outliers).

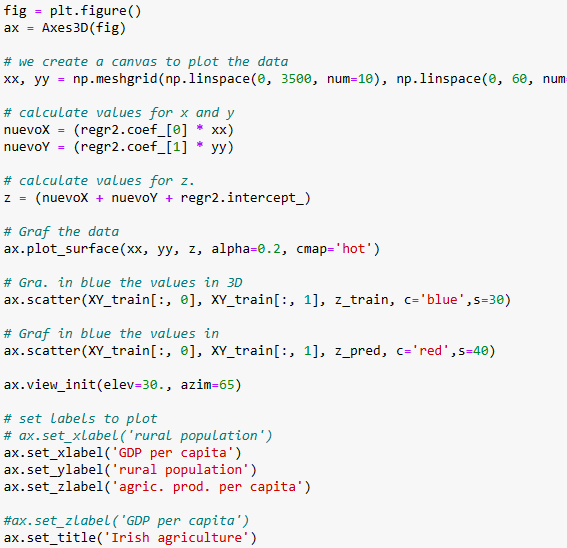
******

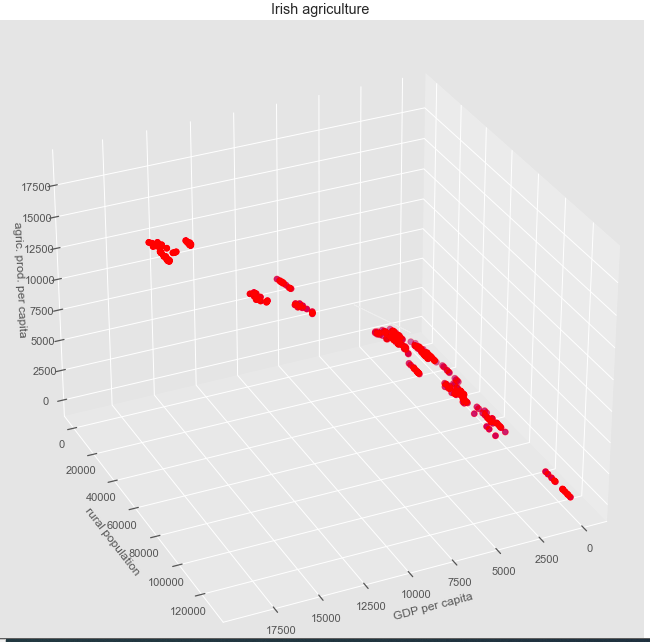
****

Now I’ve tested my model with a value of 10500, and it worked, meaning values can be forecast by the model.

**

****





*Fig. 5.2 - 3D Irish agriculture*

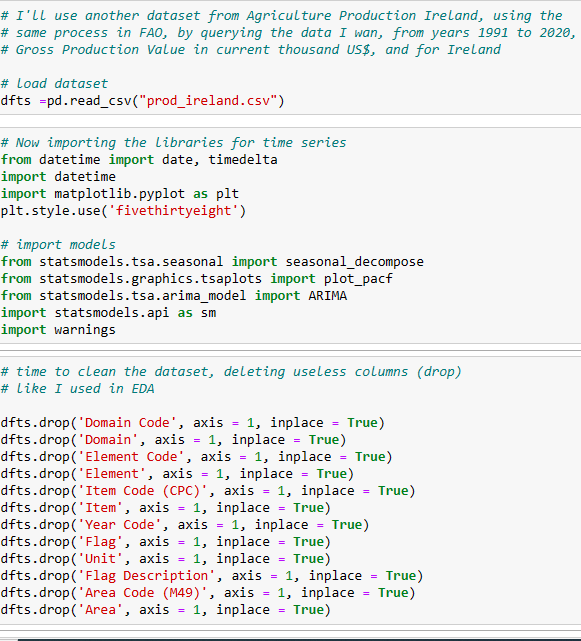
For there some relationship between the three variables, because see a line with those values.

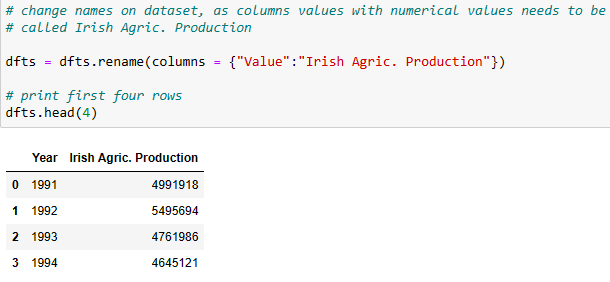
# 5.3TIMES SERIES

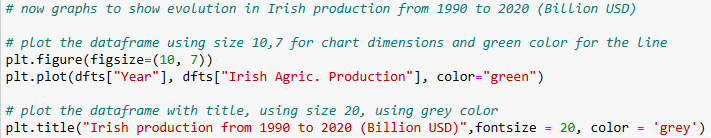
As my data frame is using data information about Irish production, on good model to use is time series, because this model has several benefits from cleaning data, understanding it and also provide forecast on future data points.

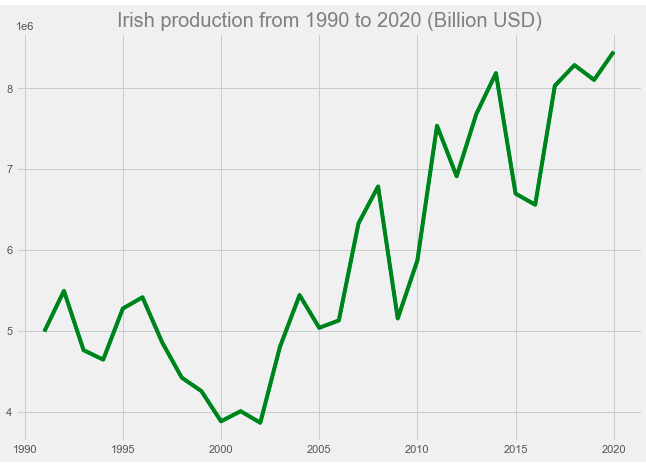
I’ll run through an example of Irish Production forecasting using a SARIMA is a model that supports univariate time series data with a seasonal component.

From 1990 to 2020, I’ll gather Irish Agriculture Production data and create a model to predict production after 2020, and its effectiveness will be measured using mean absolute percentage error.





**

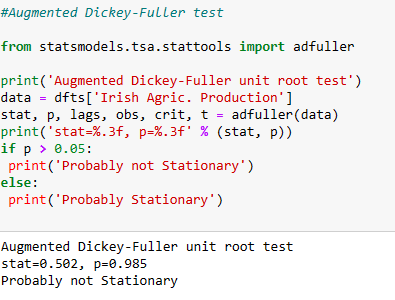


*Fig. 5.3 Irish production 1990 to 2020*

On fig. 5.3, I can see that Irish Agric. production has been up and down from 1990 to 2020. The lowest production year was 2002 (3,864 Musd), and the highest is 2020 (8,454 Musd) from 2002 to 2020, Agriculture production is growing but some decrease in 2005, 2009, 2016.

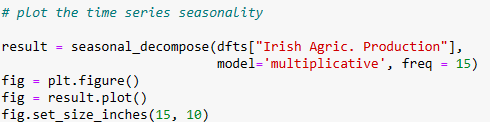
When analysing a time series, I need to identify at least three insightful aspects of the data like autocorrelation, seasonality, and stationarity

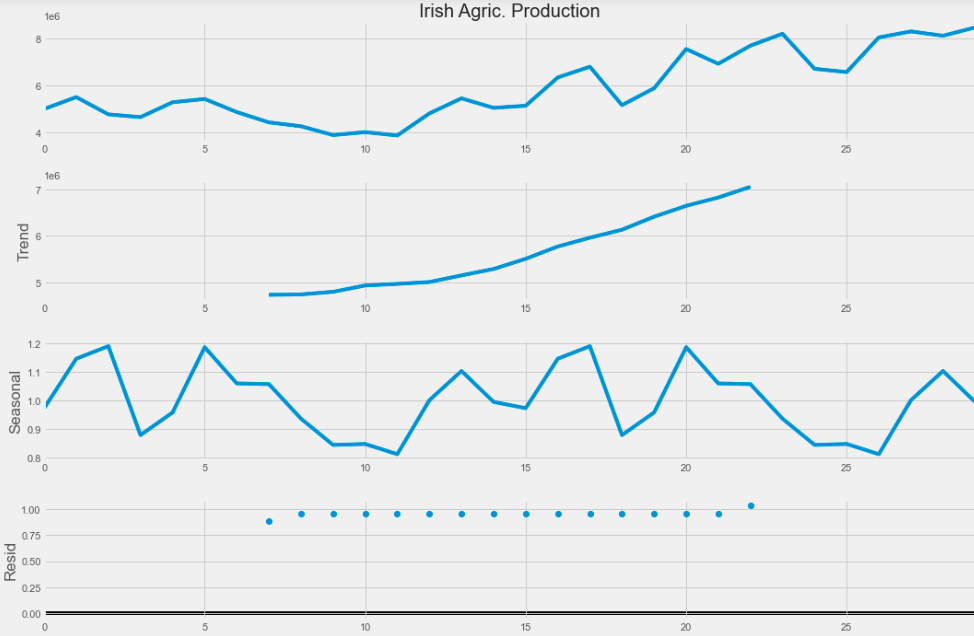
stationarity to check how little a time series’ mean and variance changes over time, and for that I'll use the augmented Dickey-Fuller test



The null hypothesis (p-value > 0.05) for this test is that the data is not stationary. Based on our output from the code above, the p-value was calculated to be ~ 0, so we can reject the null hypothesis and say that the data is in fact stationary. On to modelling

I now check seasonality to see if observations and patterns repeat themselves at regular intervals.



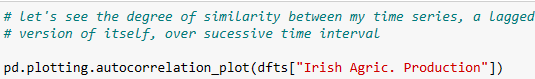


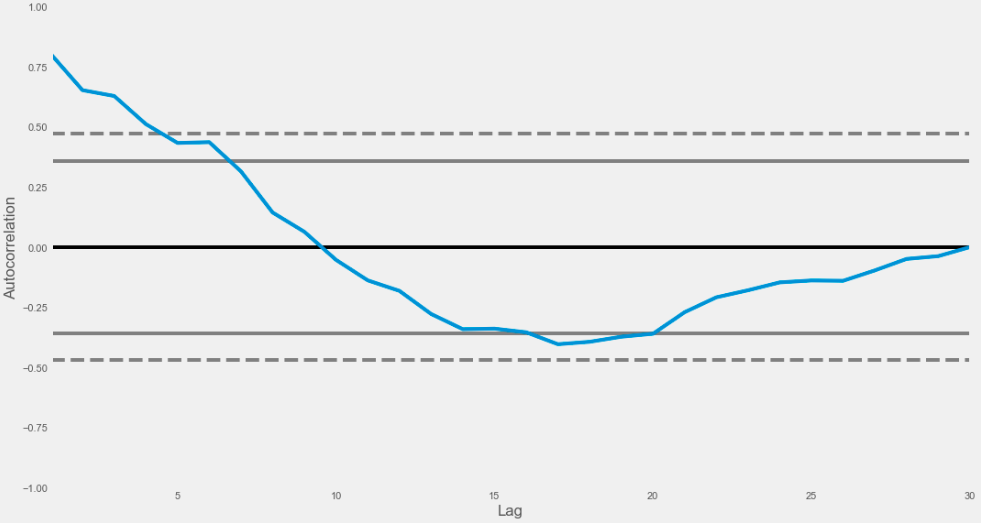
*Fig. 5.4 Irish Agr. production seasonality*

On Fig. 5.4, the seasonal line is not linear, having ups and downs for Irish Agric. production, showing some seasonality. this seasonality was also confirmed with evolution plot above, with not liner production during years analysed.

The trend range is between 4 and 7.

**Autocorrelation**

****

****

*Fig. 5.5 Autocorrelation*



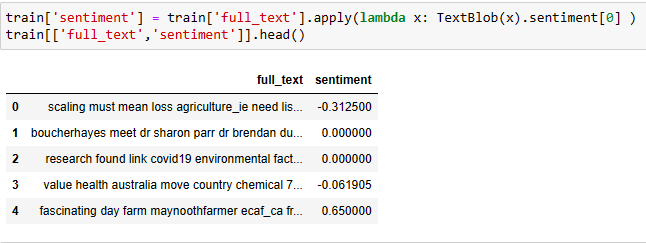
On fig. 5.5, The observation of the plot is a positive autocorrelation, with a smooth curve, meaning that a time series will be linear.

So now I need to forecast my data using predictions and my historical data. I need a model to forecast and for time series SARIMAX will help, as it includes seasonal effects and exogenous factors with the autoregressive and moving average component in time series model

# 

**5.4 Sentimental analyse:**

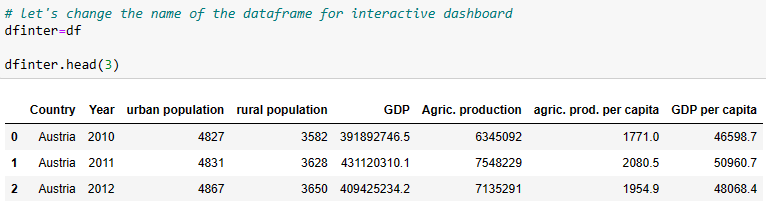
I’ve used Twitter URL scapper on apify, to get some data (EU agriculture) from the web,and will get the final result here.

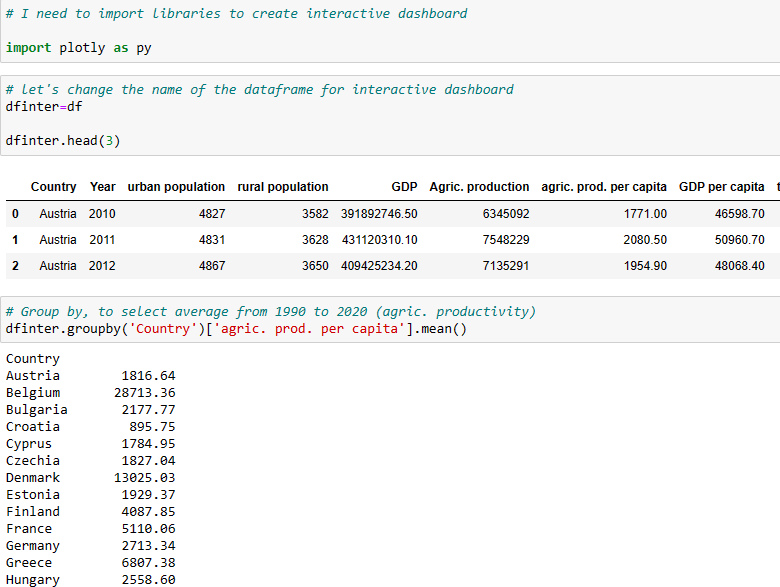
****

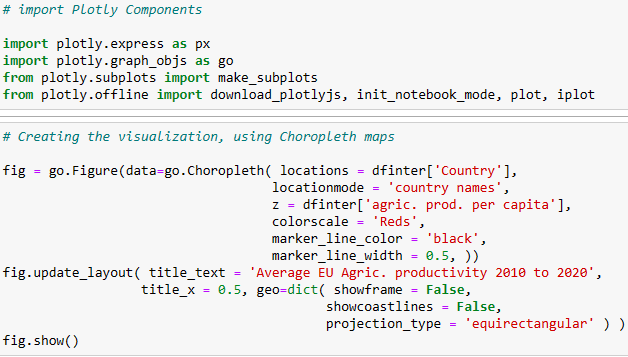
Now need to filter all the tweets resulted from my sentimental analyses, to found what the problems that originated the negative sentiment.

Sentiment Analysis allows you to get an overview of how things are going, to check the negative sentiments and get the benefits in understanding and overcoming them.

1. **Interactive DASHBOARD**



**



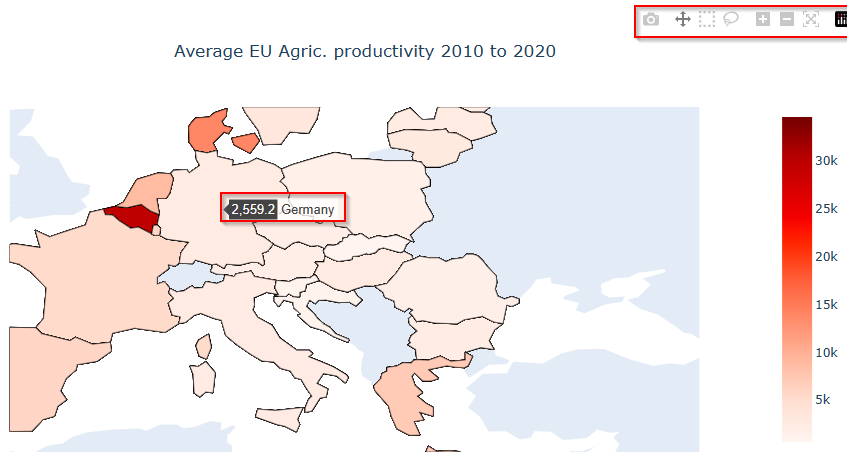
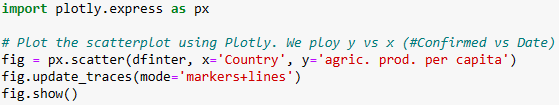


Fig. 6.1 interactive map Average Productivity Europe

On 6.1, this interactive map on notebook allows us to select any country in Europe and have the average productivity in this country, above option to size, download or pan the plot



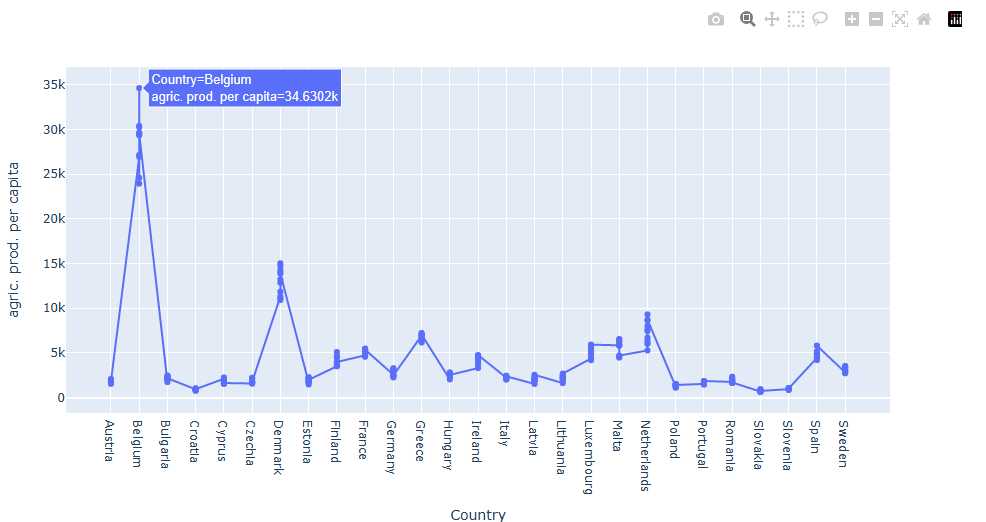


Fig. 6.2 interactive map Average Productivity Europe

On Fig. 6.2, this interactive map on notebook allows us to select any country in Europe and have the average productivity in this country, above option to size, download or pan the plot.

The advantage is comparing productivity every country easily, as we see now Belgium has the highest productivity

**Important**: at this point I used DASH, PANEL, and several other ways of doing dashboard, but technical issues implementing on python (I couldn’t solve), so I leave those interactive graphs like this.

**Sources: opensource from FAO**