# Data Visualization Project

Kalmanidis Theodoros
ID:218580

Language Used: Python 3

**The dataset**

The dataset we are going to analyze in this project is called states.csv and it contains general information for every country that is member or candidate to join European Union. It consists 38 countries including Turkey (the european provinces) , Iceland, Norway, Switzerland and Liechtenstein which are not either members or candidates of European Union, however they participate in different economic organizations of the EU.

**The objective**

The objective of this dataset is to find different correlations and patterns for all the countries that are members of EU or trade with eurozone. For instance what country is the largest py population or surface area? Which country has the highest Gpa? Is a statistically difference between north and south Europe countrie's Gpa?

We can see the **variables** of our dataset below:

Country: the name of the country

European Union: the status of the country, member/candidate or neither of them

Accesion Year: the year a member country entered EU

Council Votes: the number of votes the country has on Counsil of Europe

European Parliament Seats: the number of parliament seats the country has on Counsil of Europe

European Free Trade Agreement: the status for non EU members if a country has signed free trade agreement with the eurozone.

European Single Market: : the status of the country, member/or not member in the European single market(a market which allows free movement of goods and capital between countries)

European Monetary Union: : the status of the country, member/candidate/not member/applicable for a country with the European monetary union (a union with purpose to progressively bring closer a country to an economic integration with EU and adopt euro )

Currency: the currency of the particular country

Currency Code: the global code of the currency a country has

Language: the language/es are spoken in the country

Population:number of population of the country

Area (kmÂ²): the area of the country in $km^2$

Population Density:number of people live per $km^2$

GDP (â,¬, millions):GDP is a measure describing the value of a country's economy by calculating the overall goods and services produced annually. That variable describes the gdp of a country in euros (millions) but we are not gonna use it.

GDP ($, millions): gdp in dollar (millions), provided by World Bank

GDP per capita ($, millions): is the gdp of a country in dollars (millions) divided by the population, it provides better explanation of the living standards of the country

**Preprocessing and Correlation Matrix**

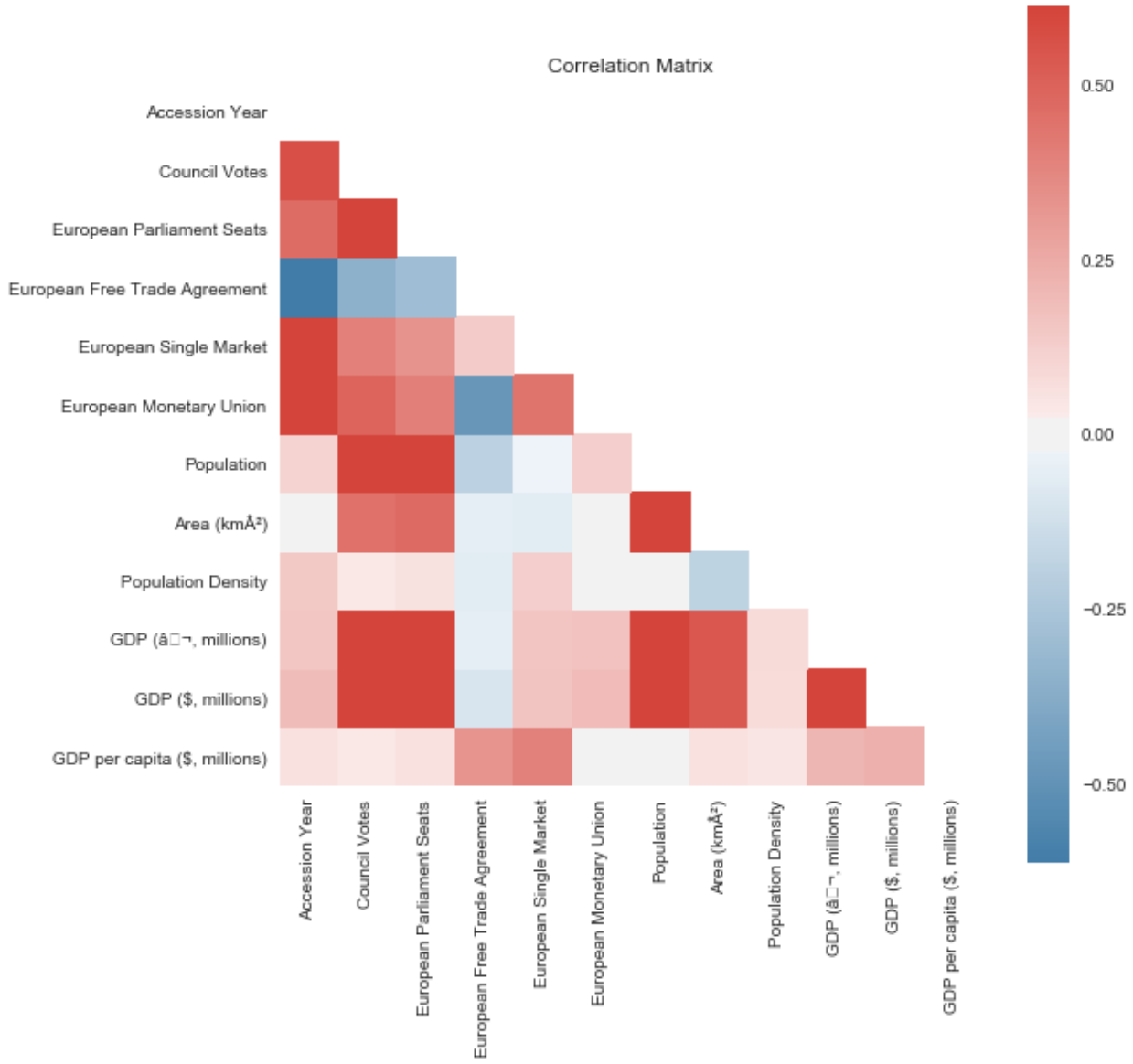
First of all we need to check if we have any missing values in our dataset.

```
++++++++++++++++++++++++++
Country                        False
European Union                  True
Accession Year                  True
Council Votes                   True
European Parliament Seats       True
European Free Trade Agreement   True
European Single Market          True
European Monetary Union         True
Currency                       False
Currency Code                  False
Language                       False
Population                     False
Area (kmÂ²)                    False
Population Density             False
GDP (â‚¬, millions)               True
GDP ($, millions)               True
GDP per capita ($, millions)    True
dtype: bool
++++++++++++++++++++++++++
```

The variables that have the word “True” next to them contain missing values. That may have occurred because of lack of information or rejection purposes( for example countries who haven’t signed free trade agreement contain blank cell instead of “Not Member”

Secondly  we have to map our categorical data in order to find various correlations, for example Not Member to 0 and Member to 1.

Below we can see the produced Correlation Matrix, it is scaled with shaded of blue for negative correlations and red for positive correlations. It is interesting to notice that the correlation matrix appears as a triangle and not the typical square sized that python produces, that change of shape helps the viewer not to get lost with too many squares and colours.




At first glance, we can see a pretty strong negative correlation for free trade agreement and strong positive correlations for Council Votes,Parliament Seats, European Market and Monetary Union, Population and GDP. In order to further investigate the correlation with number we pass the produced python correlation dataframe to excel and we scale also with shaded of blue for strong negative correlations, yellow orange and green for correlation between -0.5 to .5 and red for strong positive correlations. Because the size of the matrix is too big we split it in four:

| | Accession Year | Council Votes | European Parliament Seats |
|---|---|---|---|
| Accession Year | 1 | 0.568638499 | 0.466274836 |
| Council Votes | 0.568638499 | 1 | 0.96718545 |
| European Parliament Seats | 0.466274836 | 0.96718545 | 1 |
| European Free Trade Agreement | -0.61397215 | -0.353882442 | -0.291907402 |
| European Single Market | 0.697084874 | 0.401787112 | 0.331422581 |
| European Monetary Union | 0.703290106 | 0.498767725 | 0.404446845 |
| Population | 0.112008753 | 0.741007389 | 0.810414234 |
| Area (kmÂ²) | -0.019274375 | 0.452987953 | 0.473059213 |
| Population Density | 0.145218978 | 0.036756177 | 0.059215949 |
| GDP ($, millions) | 0.188595299 | 0.783282467 | 0.891644546 |
| GDP per capita ($, millions) | 0.066400295 | 0.034745667 | 0.064815326 |

In the first excel correlation matrix we see strong correlations for the following variable matching:

- Accession Year with European Free Trade Agreement,European Single Market & Monetary Union (we will explain the correlation later)
- Gdp($,millions) with council votes and European parliament seats (we will explain the correlation later)
- Population with council votes and European parliament seats (we will explain the correlation later)
- Council Votes with European parliament seats which is the most obvious, the more seat a country has on European council the more the votes and vice versa.

| | European Free Trade Agreement | European Single Market | European Monetary Union |
|---|---|---|---|
| Accession Year | -0.61397215 | 0.697084874 | 0.703290106 |
| Council Votes | -0.353882442 | 0.401787112 | 0.498767725 |
| European Parliament Seats | -0.291907402 | 0.331422581 | 0.404446845 |
| European Free Trade Agreement | 1 | 0.137620471 | -0.479928976 |
| European Single Market | 0.137620471 | 1 | 0.444090586 |
| European Monetary Union | -0.479928976 | 0.444090586 | 1 |
| Population | -0.195521033 | -0.025728764 | 0.128813416 |
| Area (kmÂ²) | -0.048999005 | -0.064108066 | -0.003056739 |
| Population Density | -0.063134756 | 0.127790761 | 0.020714522 |
| GDP ($, millions) | -0.096037989 | 0.164118233 | 0.193945309 |
| GDP per capita ($, millions) | 0.329572874 | 0.395209502 | -0.006195277 |

In the second excel correlation matrix we see strong correlations for Accession Year with European Free Trade agreement, European single market and European monetary union. That means two things. First if a country has a value in the accession year cell(so that country joined EU ) the higher the chance to also join European single market and monetary union. Second, the strong but negative correlation between accession year and European free trade agreement , means that if a country has zero value in the accession year (never joined EU) it is more likely to have signed free trade agreement.

| | Population | Area (kmÂ²) | Population Density |
|---|---|---|---|
| Accession Year | 0.112008753 | -0.019274375 | 0.145218978 |
| Council Votes | 0.741007389 | 0.452987953 | 0.036756177 |
| European Parliament Seats | 0.810414234 | 0.473059213 | 0.059215949 |
| European Free Trade Agreement | -0.195521033 | -0.048999005 | -0.063134756 |
| European Single Market | -0.025728764 | -0.064108066 | 0.127790761 |
| European Monetary Union | 0.128813416 | -0.003056739 | 0.020714522 |
| Population | 1 | 0.76330878 | 0.020926911 |
| Area (kmÂ²) | 0.76330878 | 1 | -0.186999589 |

| | | | |
|---|---|---|---|
| Population Density | 0.020926911 | -0.186999589 | 1 |
| GDP ($, millions) | 0.873882381 | 0.536398751 | 0.079653794 |
| GDP per capita ($, millions) | 0.008934491 | 0.062663643 | 0.05098147 |

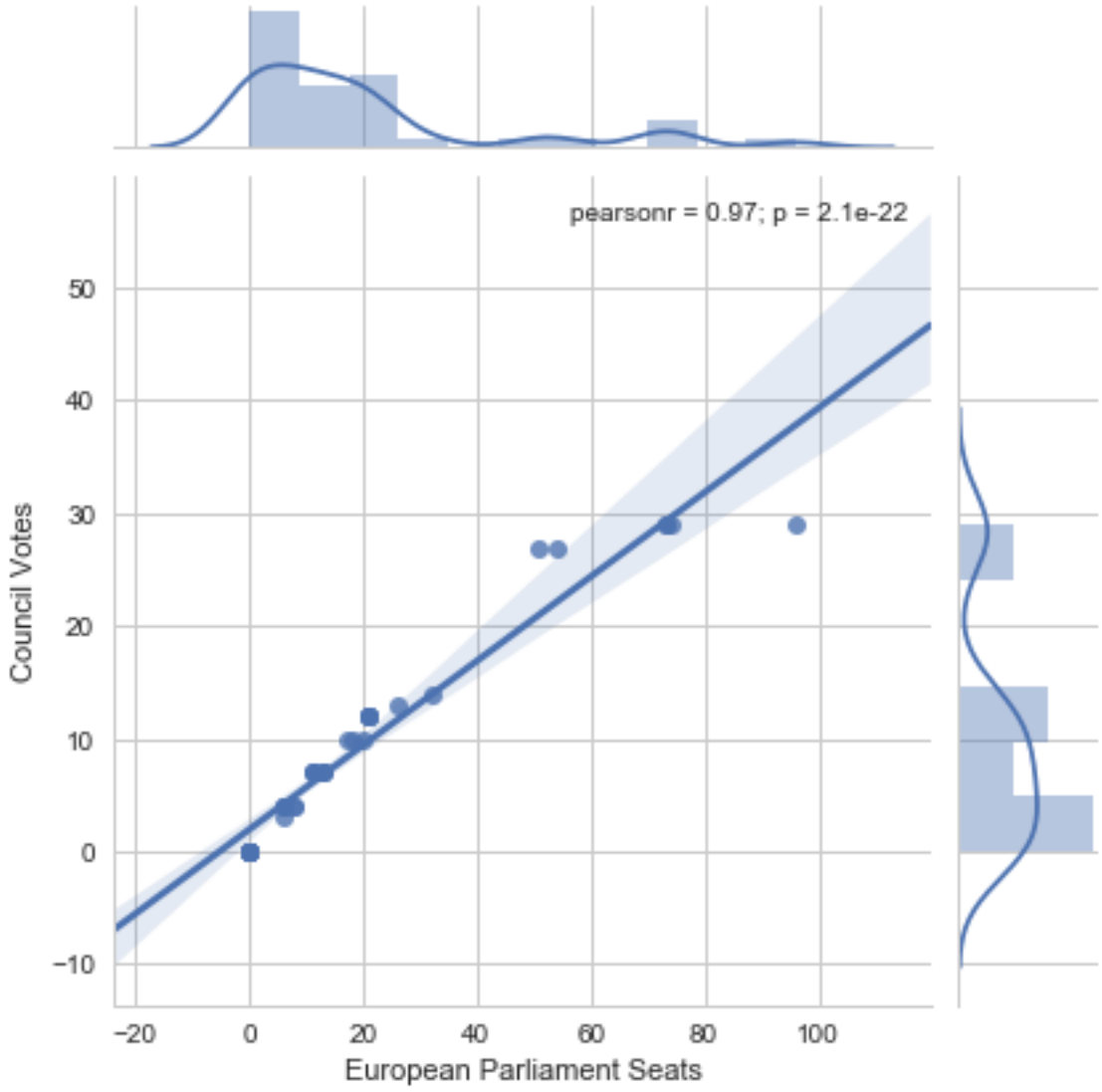
In the third matrix we can see two main groups of correlations, the first is area with population, which means the bigger the area a country rules the higher the population that lives in the country. The second is Population with council votes,parliament seats and GDP($,millions)

| | GDP ($, millions) | GDP per capita ($, millions) |
|---|---|---|
| Accession Year | 0.188595299 | 0.066400295 |
| Council Votes | 0.783282467 | 0.034745667 |
| European Parliament Seats | 0.891644546 | 0.064815326 |
| European Free Trade Agreement | -0.096037989 | 0.329572874 |
| European Single Market | 0.164118233 | 0.395209502 |
| European Monetary Union | 0.193945309 | -0.006195277 |
| Population | 0.873882381 | 0.008934491 |
| Area (kmÂ²) | 0.536398751 | 0.062663643 |
| Population Density | 0.079653794 | 0.05098147 |
| GDP ($, millions) | 1 | 0.233755093 |
| GDP per capita ($, millions) | 0.233755093 | 1 |

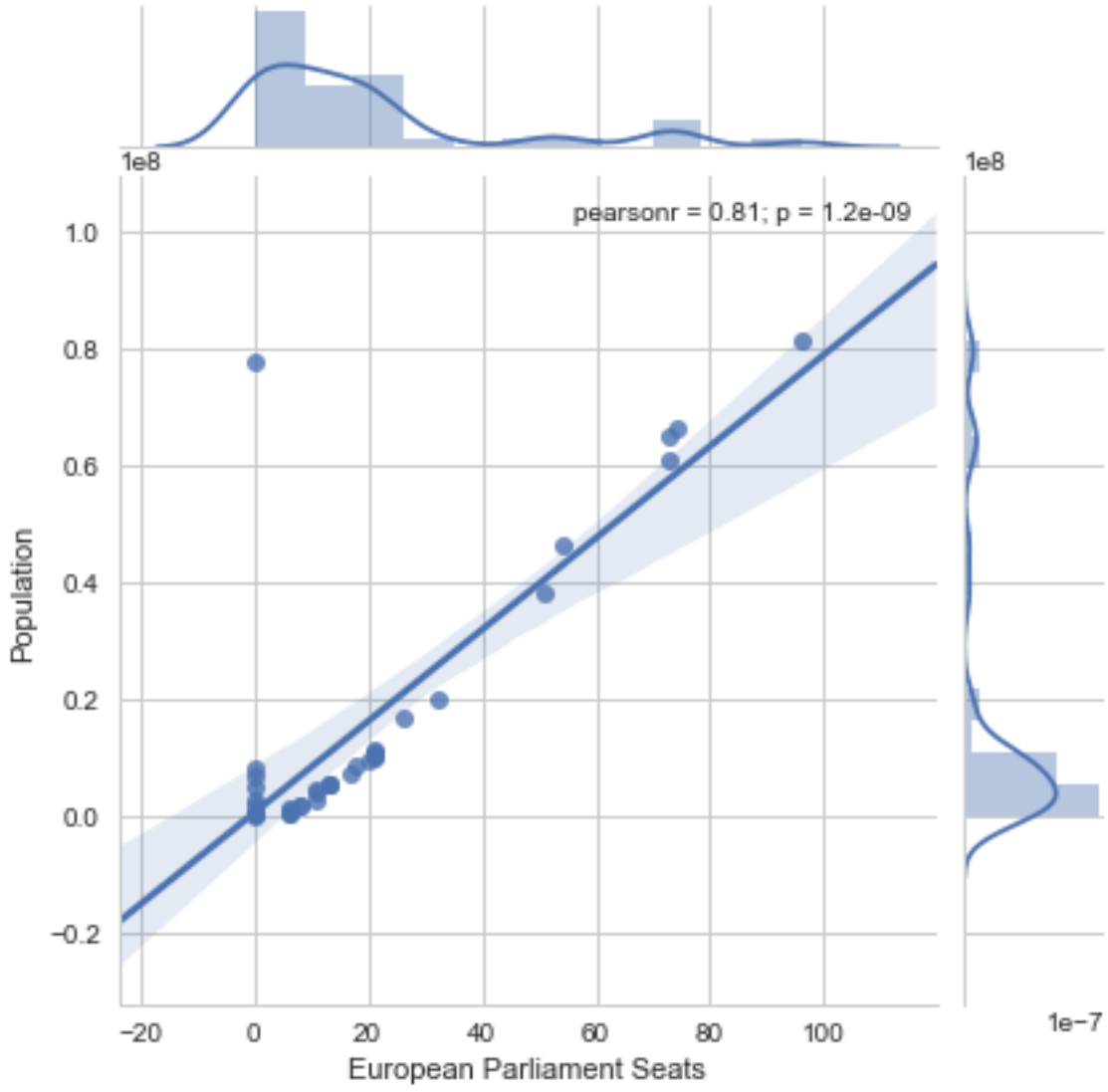
In the last correlation matrix we can see strong correlation for GDP($,millions) with Council Votes, Population and Parliament Seats. That's an indicator that strong economically countries have more seats on European council and more votes, plus a big population.

Now let see some joinplots for the highiest correlated variables (>0.75 or <-0.75),which are:

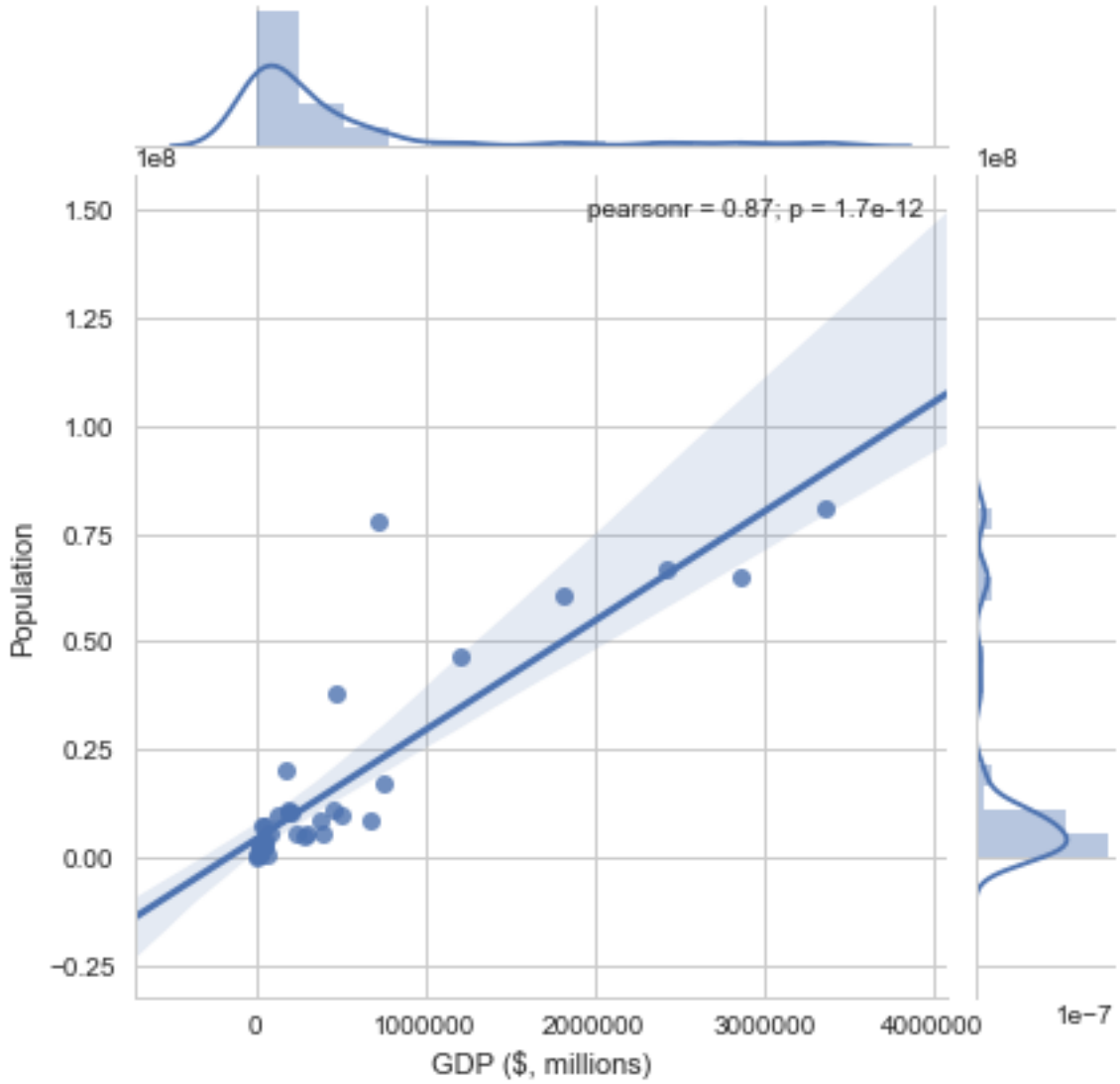
- Council Votes-European Parliament Seats (0.96 corr)



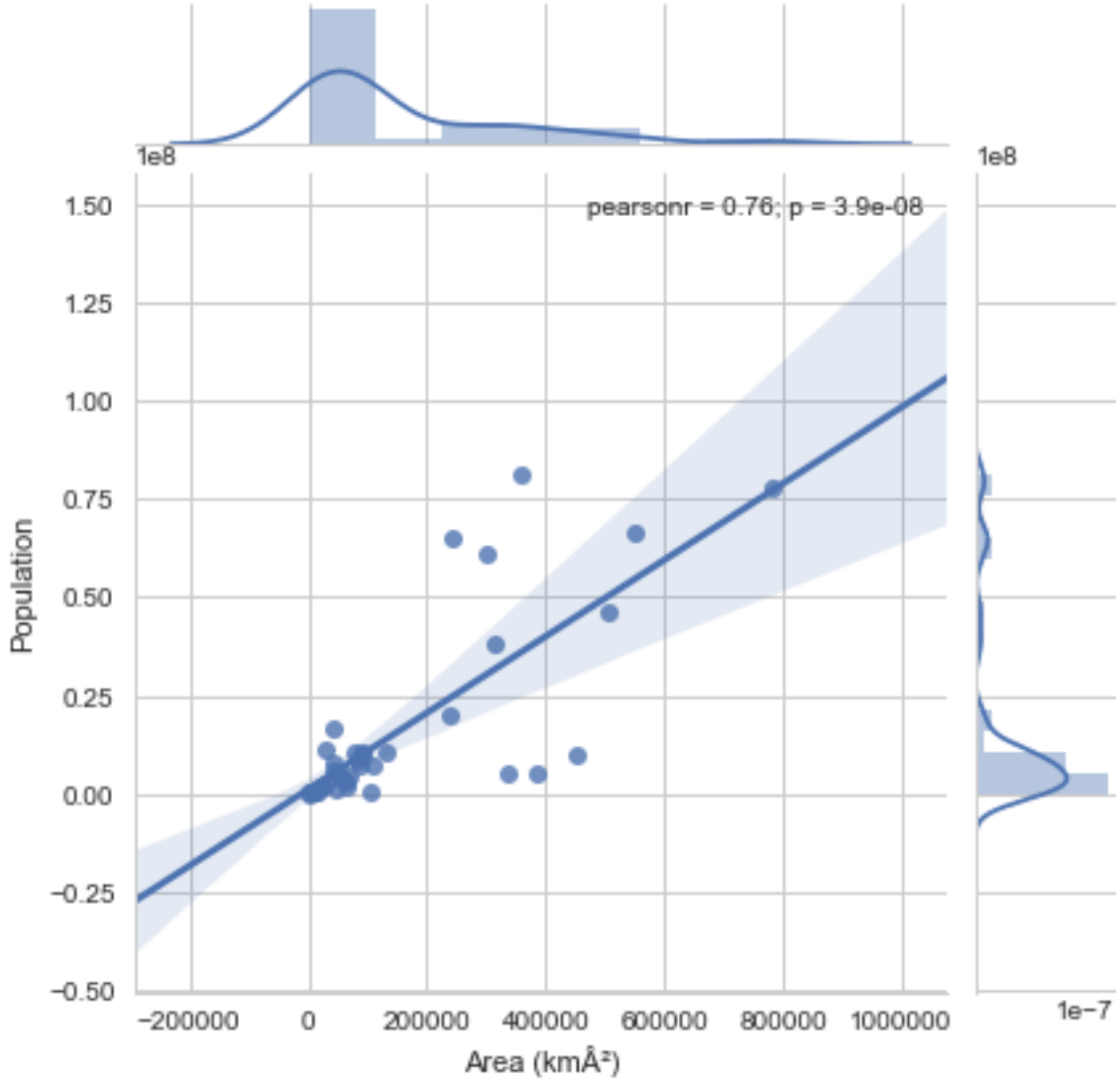
- Population-European Parliament Seats (0.81 corr)



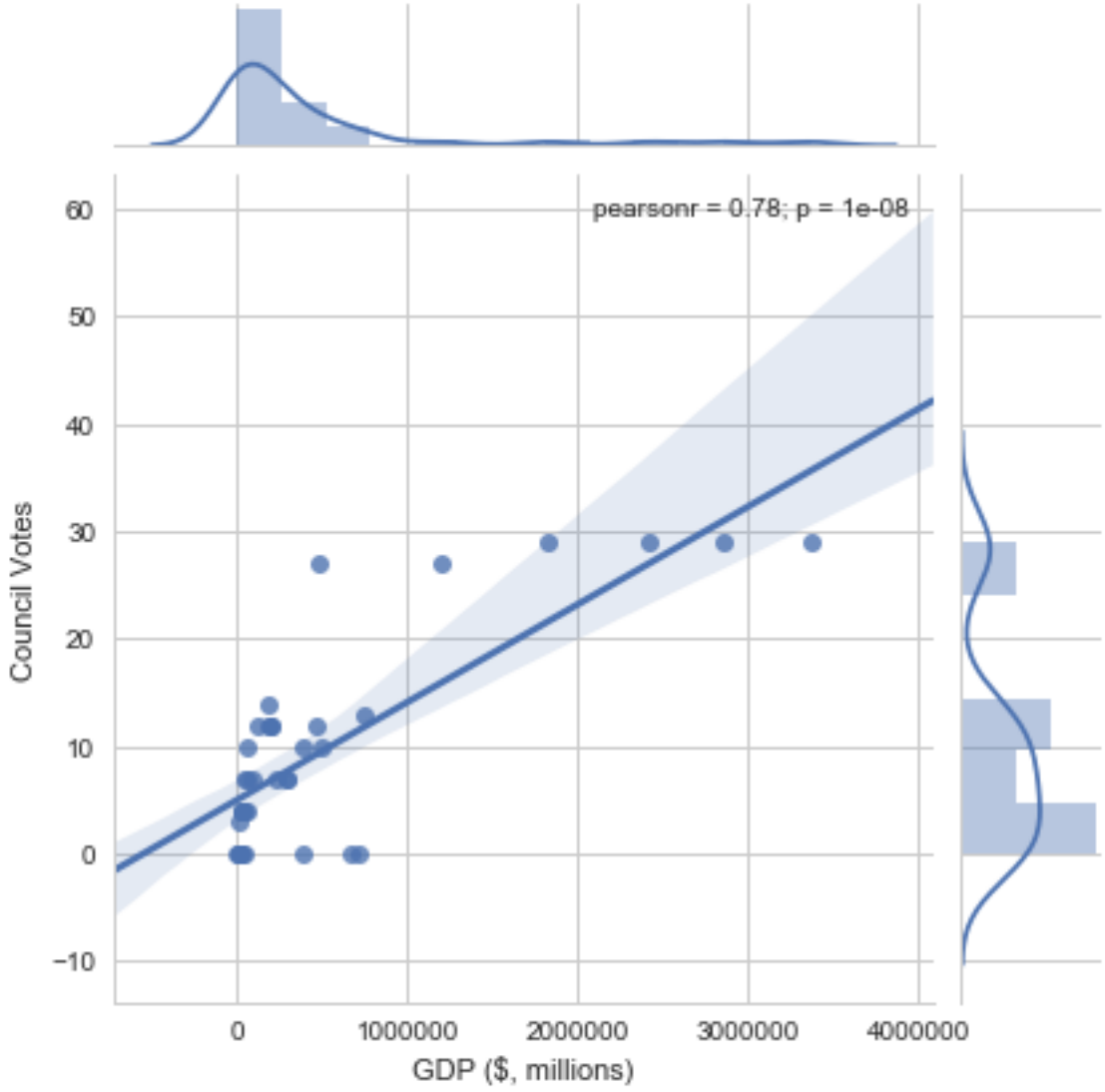
- Population-Gdp ($,millions) (0.87 corr )



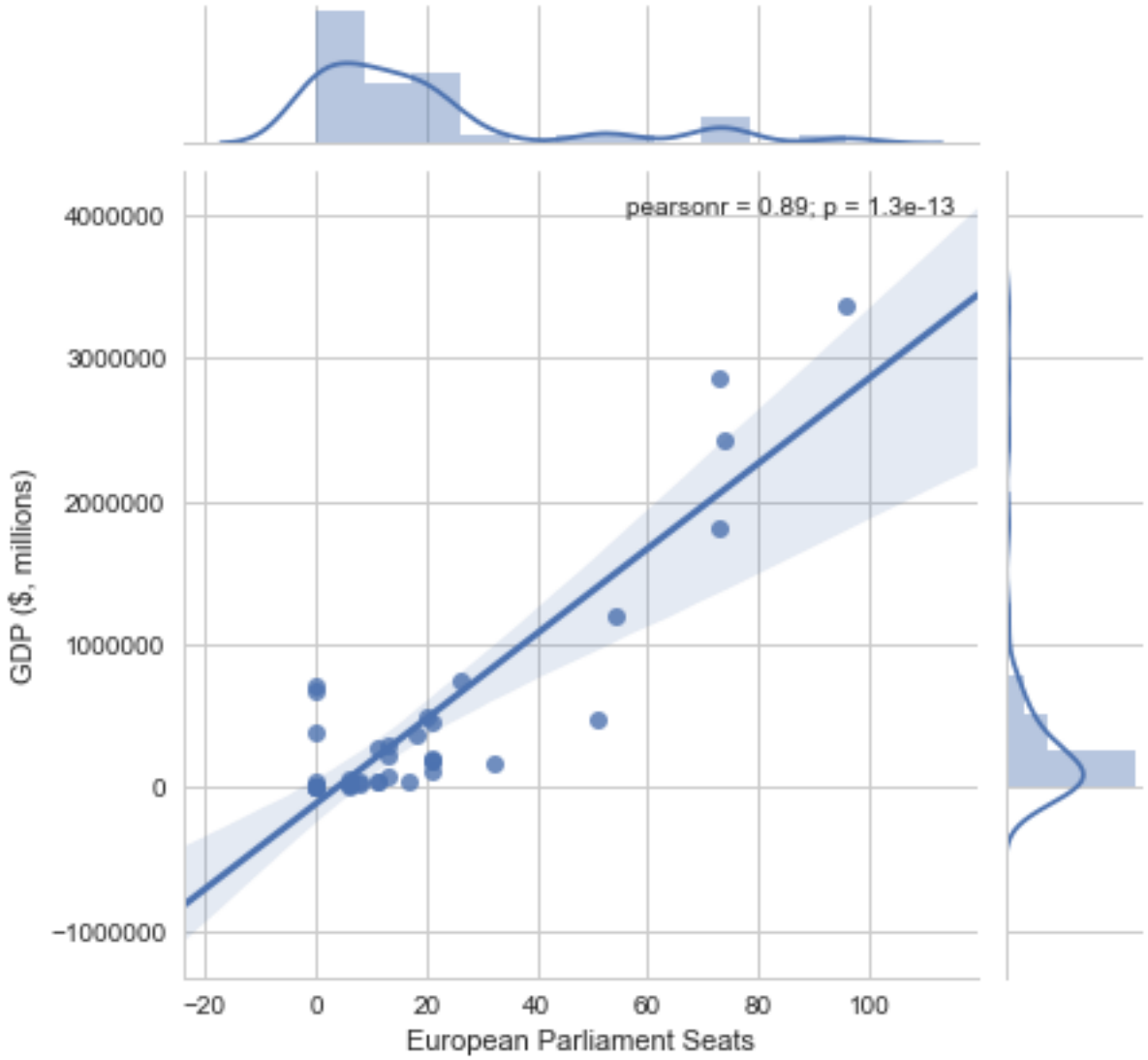
- Population-Area (0.76 corr)



- Council votes-Gdp ($,millions) (0.78 corr)



- EuropeanParliament Seats -Gdp ($,millions) (0.89 corr)




In order to do some statistical tests we split our dataset into four regions as the Wikipedia region map depicts below:


Note thay British Isles will be included with Western Europe.

The goal is to compare gdp and gdp per capital between regions and see if there is statistically difference.

Statistical Tests

From python's  mstats library we run a normality test to see if our dataframes follow the normal distribution. We name north_gpa  the dataframe for north's regions gdp and north_capital for north's region gdp per capital, same naming for the other regions respectively.

The results of the normality test that the gdp dataframe of eastern Europe doesn't follow the normal distribution

```
north_gpa is normal
western_gpa is normal
eastern_gpa is not normal
south_gpa is normal
north_capital is normal
western_capital is normal
eastern_capital is normal
south_capital is normal
```

So we try a Shapiro test for normality to see if the first test was wrong.

```
shapiro gdp
(0.8506752848625183, 0.09678085893392563)
(0.7989734411239624, 0.014087950810790062)
(0.7227372527122498, 0.0014078984968364239)
(0.8013042211532593, 0.029560184106230736)
shapiro gdp capital
(0.8828698992729187, 0.20056480169296265)
(0.9033446311950684, 0.23832771182060242)
(0.9339947700500488, 0.4243824779987335)
(0.9367749691009521, 0.5796536803245544)
```

On the right of every parentheces we can see the p value, for the first set the third row p value equals 0.01 <0.05 so we are 95% that again eastern's Europe gdp doen't follow normal distribution. On the other hand the last 4 rows (aka gdp capital) have p values greater than 0.05 so we are pretty sure that we can run parametric tests for gdp datasets.

Since gdp per capital dataframes fullfill normality requirements, we need to check them if they also fulfill the equality of variances requirements.

So we run levene test for every combination of gdp per capital dataframe, north europe's gdp capital with western's etc.

```
LeveneResult(statistic=array([ 0.0030043]), pvalue=array([ 0.95696722]))
LeveneResult(statistic=array([ 9.98742648]), pvalue=array([ 0.00541473]))
LeveneResult(statistic=array([ 4.31267156]), pvalue=array([ 0.05672274]))
LeveneResult(statistic=array([ 4.05588727]), pvalue=array([ 0.05766758]))
LeveneResult(statistic=array([ 9.98742648]), pvalue=array([ 0.00541473]))
LeveneResult(statistic=array([ 1.0013161]), pvalue=array([ 0.33025532]))
LeveneResult(statistic=array([ 1.84437831]), pvalue=array([ 0.19328234]))
LeveneResult(statistic=array([ 1.0013161]), pvalue=array([ 0.33025532]))
LeveneResult(statistic=array([ 4.31267156]), pvalue=array([ 0.05672274]))
```

For all the possible combinations we see p values greater than 0.05 so the gdp per capital dataframes fulfill and the equality of variances criterion.

To summarize what we know so far:

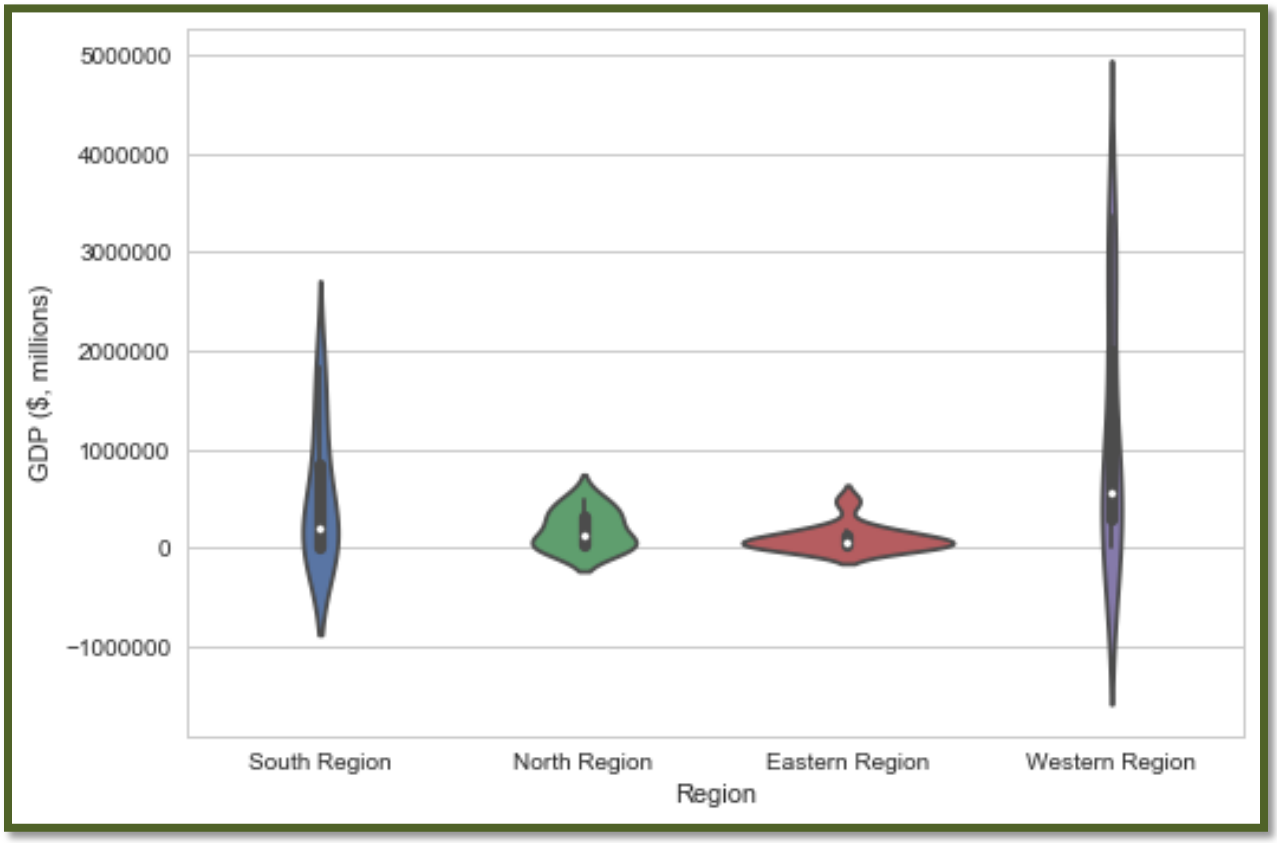
Gdp 4 dataframes

- One of four doesn't follow normal distribution with both  normality tests tried

So we can run only a Kruskal Wallis test:

```
kruskal
KruskalResult(statistic=7.7310247501276761, pvalue=0.051910324429234468)
```

And we can observe that there is no statistically significant difference between gdp of all regions because p=0.0519>0.05




In the image above we can see a violin plot for our four region's gdp.

If we look carefully we can observe the white dots inside each violin plot which correspond to the mean value. Except from the western region gdp which looks enough higher the other three look the same, but that's logical because from the non parametric test we run We saw that the p value was so close to the limits of statistical difference. From the widths of each violin plot we can understand the number of instances that have a certain value for example Eastern region is so width a little bit above zero which means it has many low gdp countries, on the other hand western region violin plot has big height but small width which help us to understand that we have fewer countries compared to other regions but with high gdp.

Gdp per Capital 4 dataframes

- ✓ All dataframes follow normal distribution
- ✓ All dataframe have equal variances

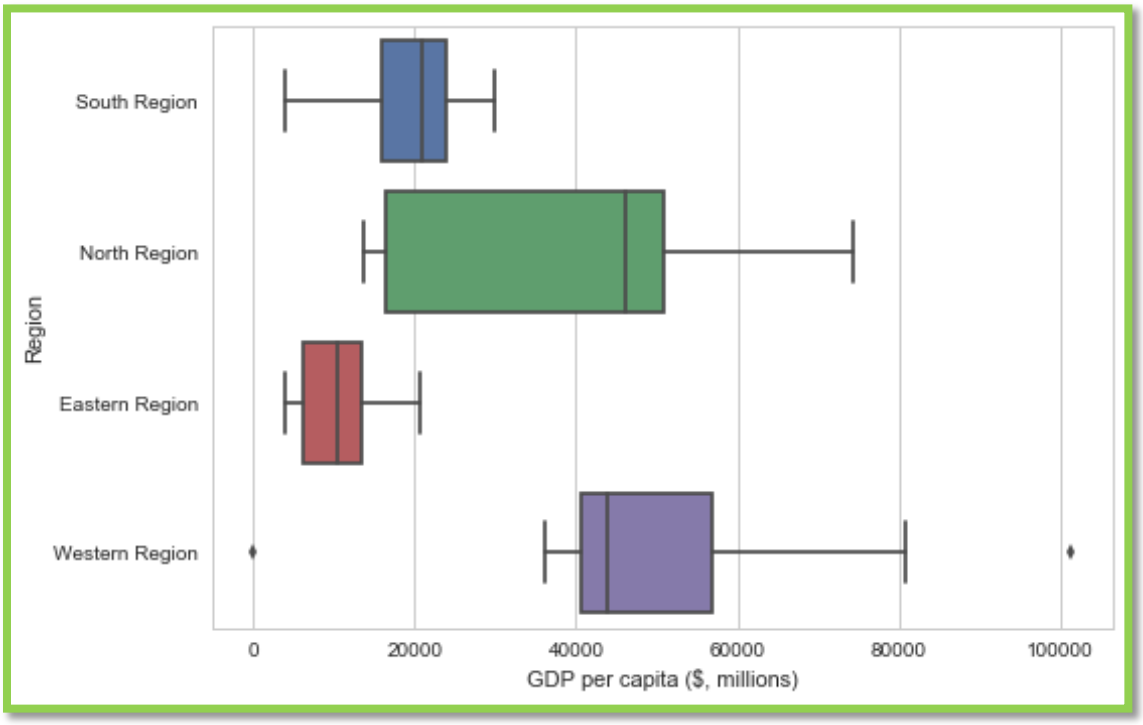
So we can run a one way anova test:

```
anova
F_onewayResult(statistic=array([ 10.16205476]), pvalue=array([  6.33139591e-05]))
Ttest_indResult(statistic=array([ 2.41679764]), pvalue=array([ 0.02989087]))
Ttest_indResult(statistic=array([-0.84119839]), pvalue=array([ 0.41263285]))
Ttest_indResult(statistic=array([ 4.35063659]), pvalue=array([ 0.00038534]))
Ttest_indResult(statistic=array([-2.67289143]), pvalue=array([ 0.01551877]))
Ttest_indResult(statistic=array([ 4.82796279]), pvalue=array([ 0.00010218]))
Ttest_indResult(statistic=array([ 3.01221604]), pvalue=array([ 0.00826609]))
```

The p value we get is almost 0<0.05 so there is a statisticall difference between gdp per capital of all regions.

We run a t test for all permutations and we see also statistically significant difference between:

- North –South Gdp per capital
- North-Western Gdp per capital
- Western-Eastern Gdp per capital
- Western-South Gdp per capital




In the image above we can see the four boxplots of our regions.

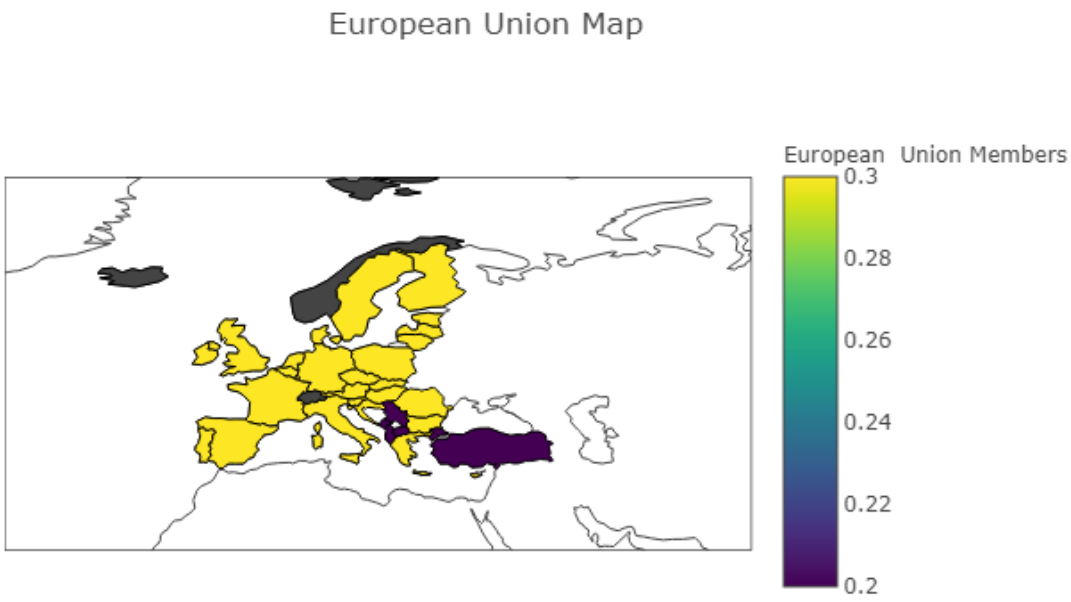
Since we know that there is significant difference in this case,

the boxoplots help us to see the differences more clearly. As we can see Eastern Europe

has the most “poor” countries where as North and Western have the more “rich” course

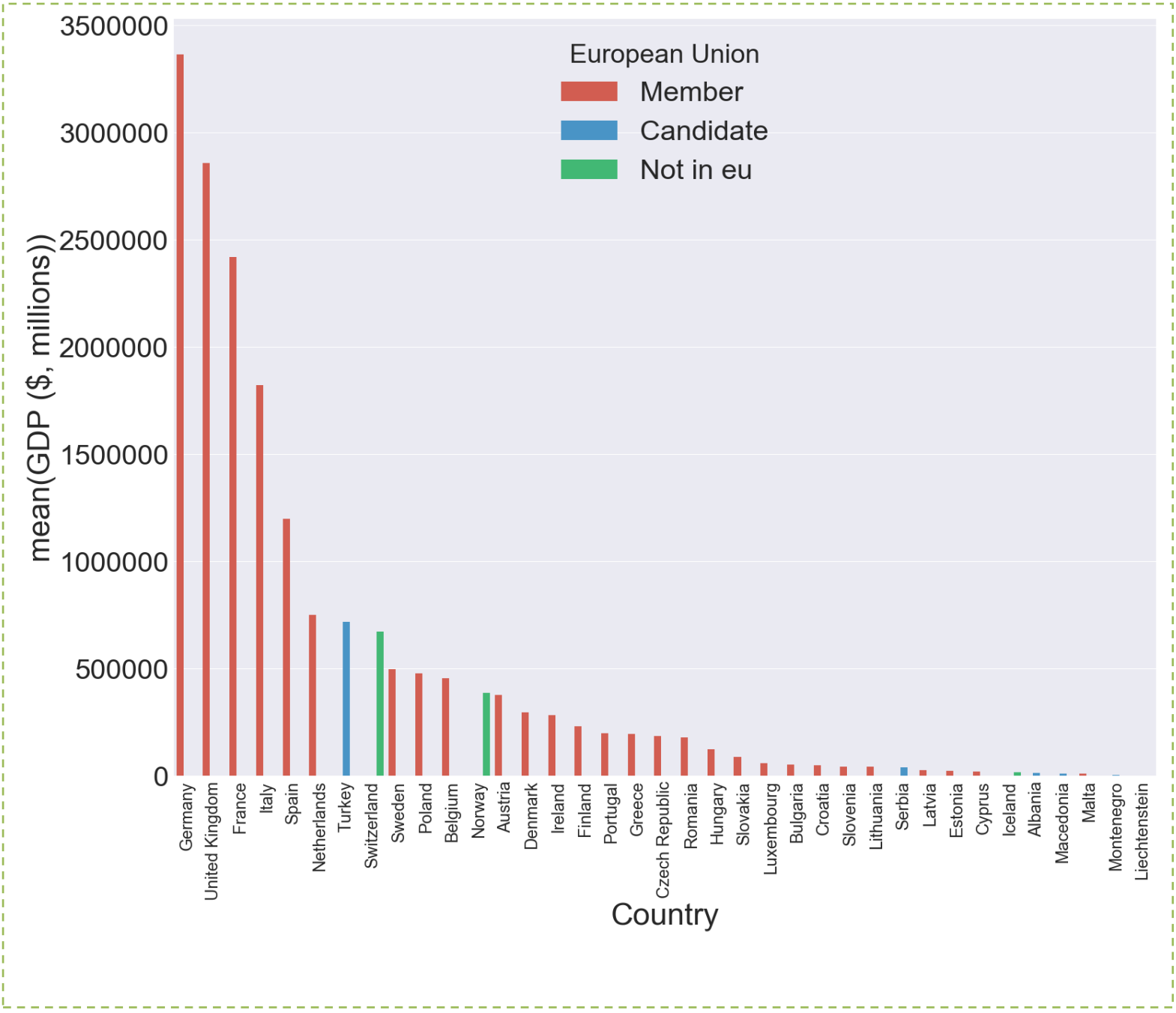
## Data Visualizations

First of all let see how many countries are members of the European Union. In the choropleth map below we can see with yellow the countries that are members of the European union with purple the candidate countries to join EU and with grey countries that are not members.

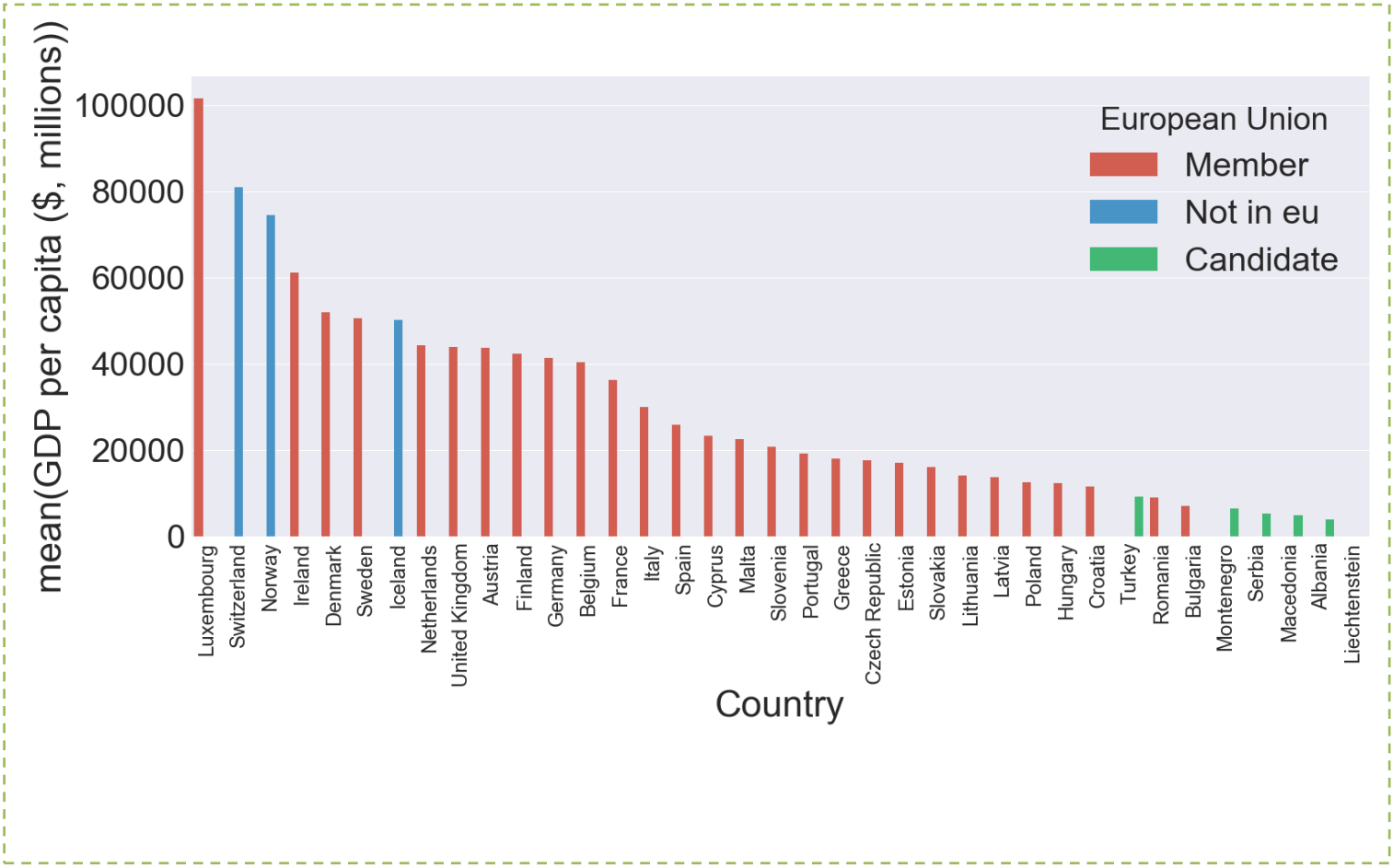



So we see that the majority of European countries have or are willing to join EU. For this map we use plotly, note we use decimal numbers for each colour because plotly accept scaling from zero to one so we have to adjust our categorical values (candidate,member etc) to a specific decimal in that range.

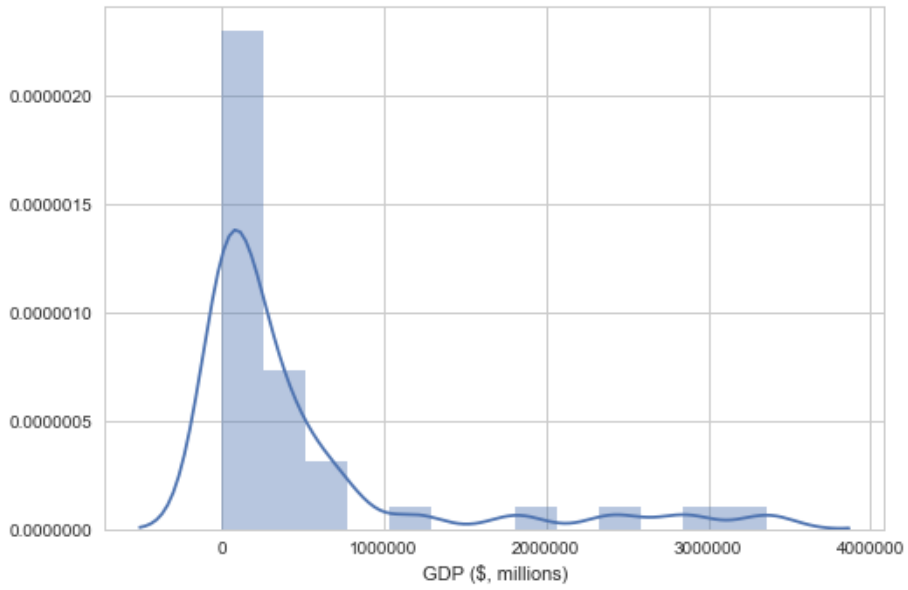
Now let's investigate our dataset  a bit further,below we can see two barplots for Gdp($,millions), Gdp per Capital in order to find the countries with highest and lowest gdp and gdp per capital.



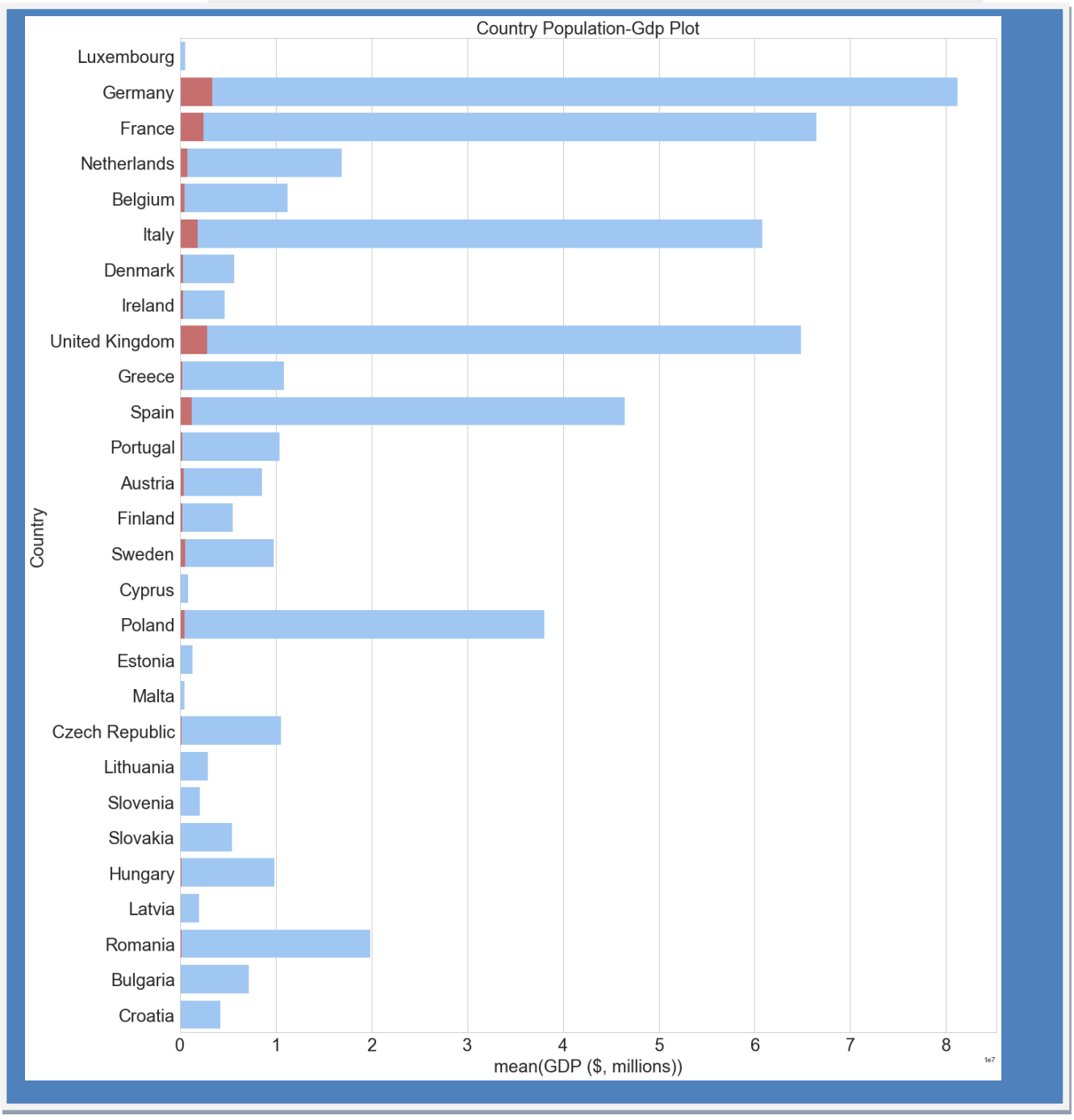
One interesting thing to notice is that we use not only sorted barplot so that we can easily spot the top 5 rich and poor countries but also we use color notation for the 3 status a country can have (member,candidate, not in eu). That can help us to compare different countries



We can also see the Gdp distribution of all europe's countries looks like a Gaussian distribution

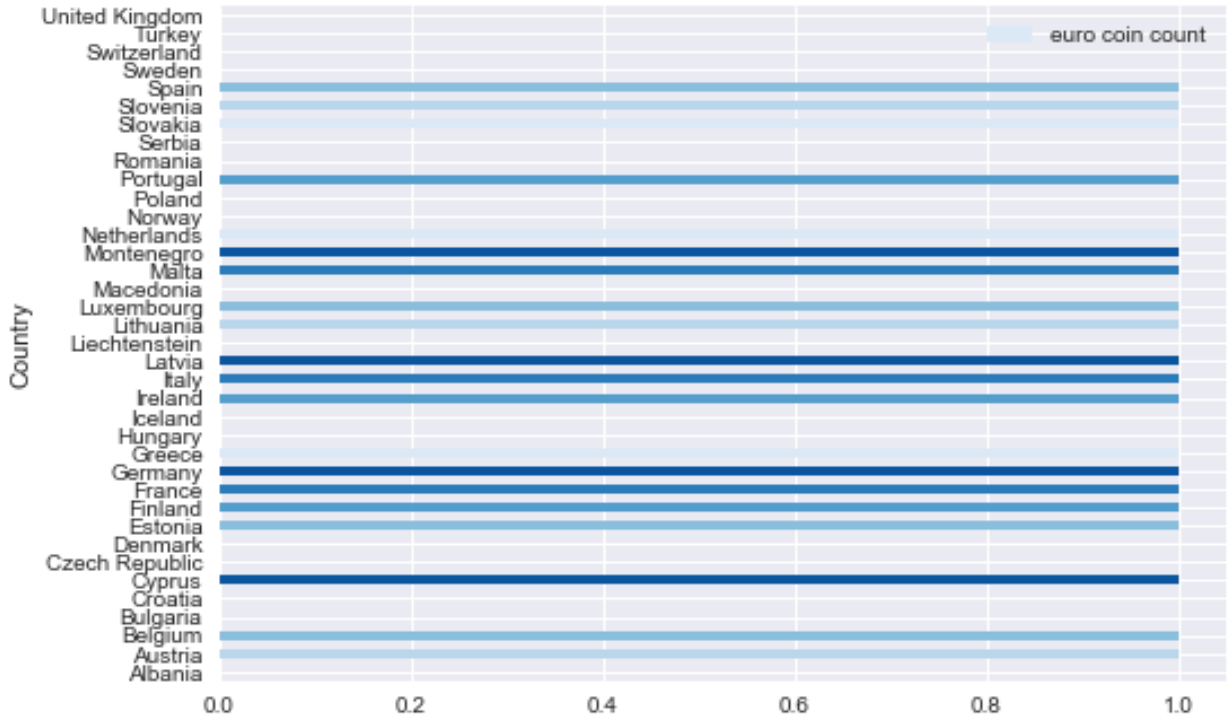


Since we saw the Gdp, the Gdp per capital, lets see what coins are popular in our dataset and also if the population plays a significant role with the overall economy:

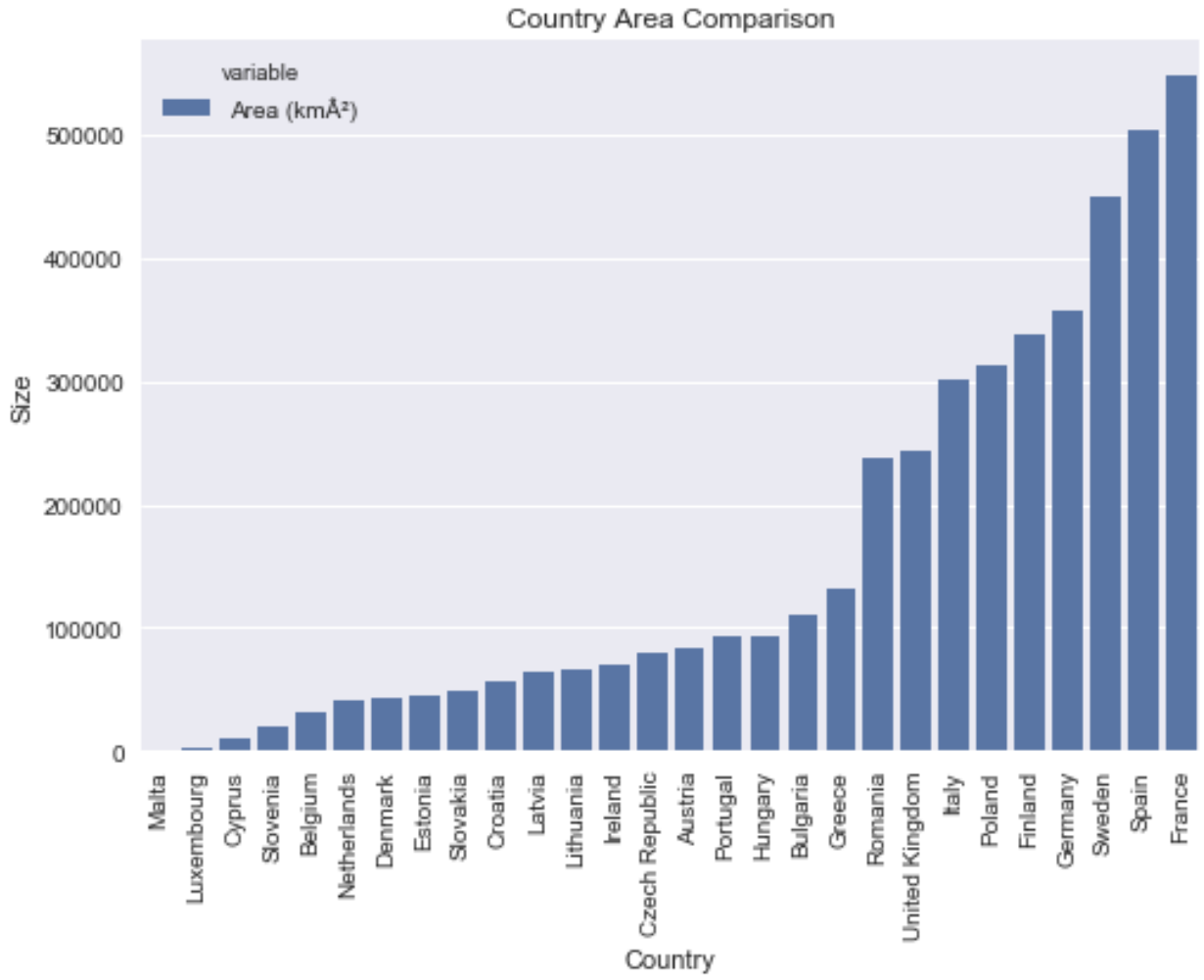


Looking at the horizontal barplots we see that bigger in population countries have also strong economy. For this comparison we used horizontal barplots with two opposite colours inside the same bar for each variable for even more precise comparison.

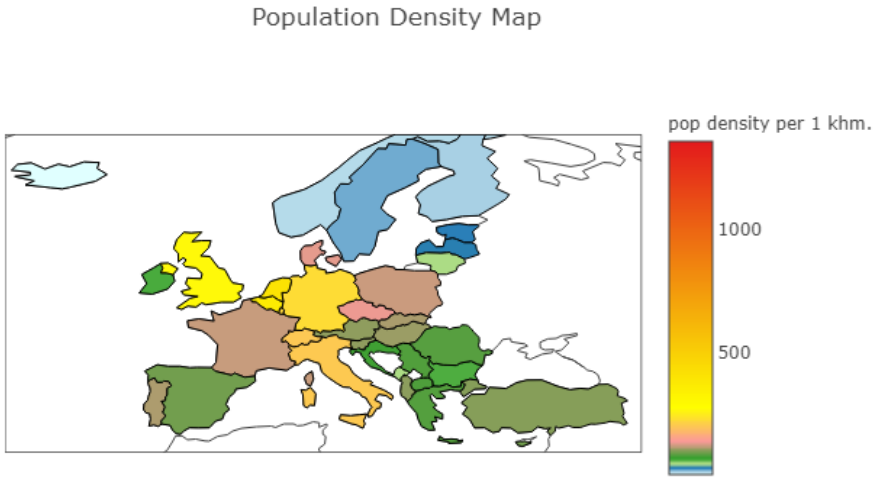
In the following barplot we see the same teqnique of horizontal barplots to plot countries that have euro coin with shades of blue:



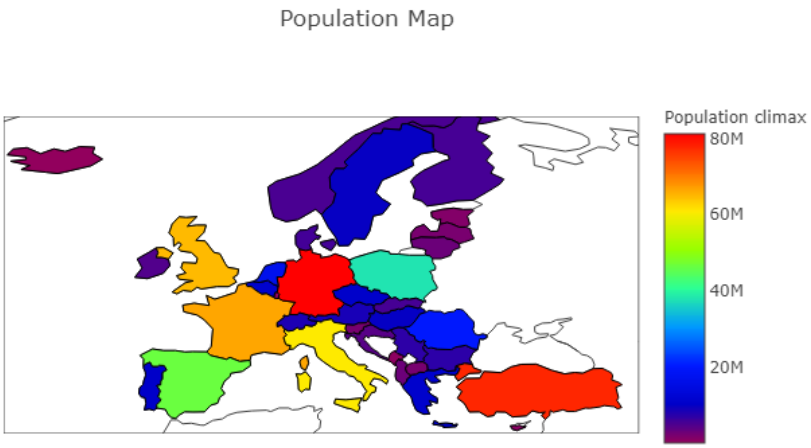
Apart from the economic point of view we can analyze countries in terms of demographics, and much more. In the barplot below we can see the countries compared based on their area size:



We can also see the density and the population of Europe using maps, we scale the continous variables with different colours:

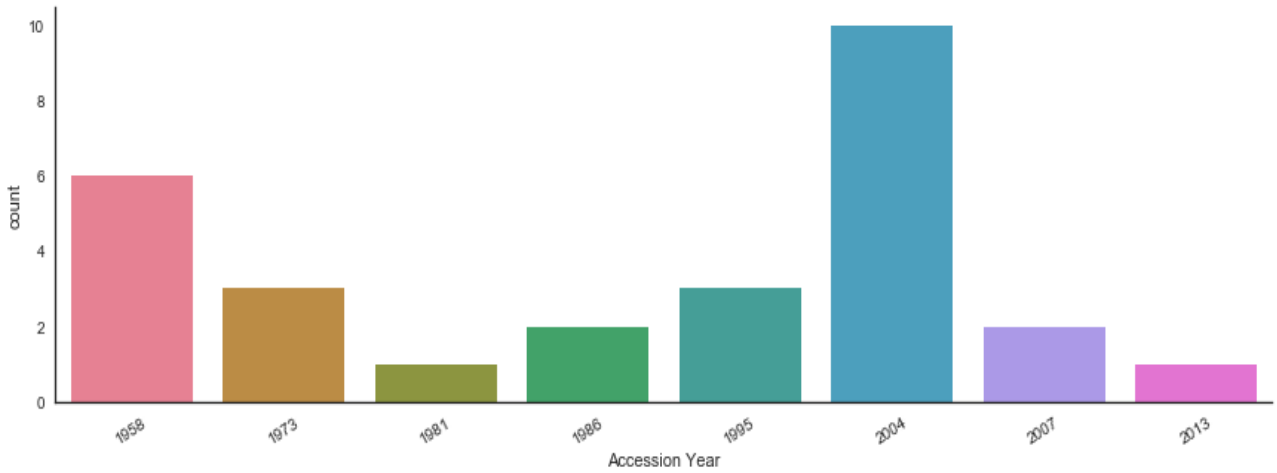



The central Europe has more density which is logical since there are big-center cities and not much villages.



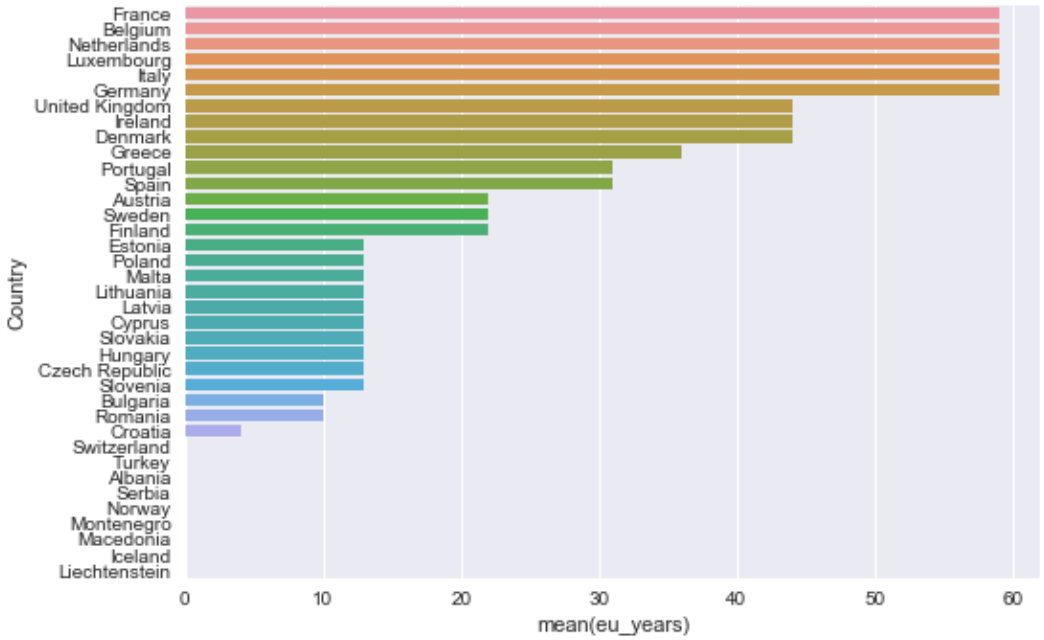

In the population map we see that turkey and german have the bigger population in Europe where as Ireland , Albania and countries of old ussr have the less.

Furthemore we see can which year we had the most entry members:

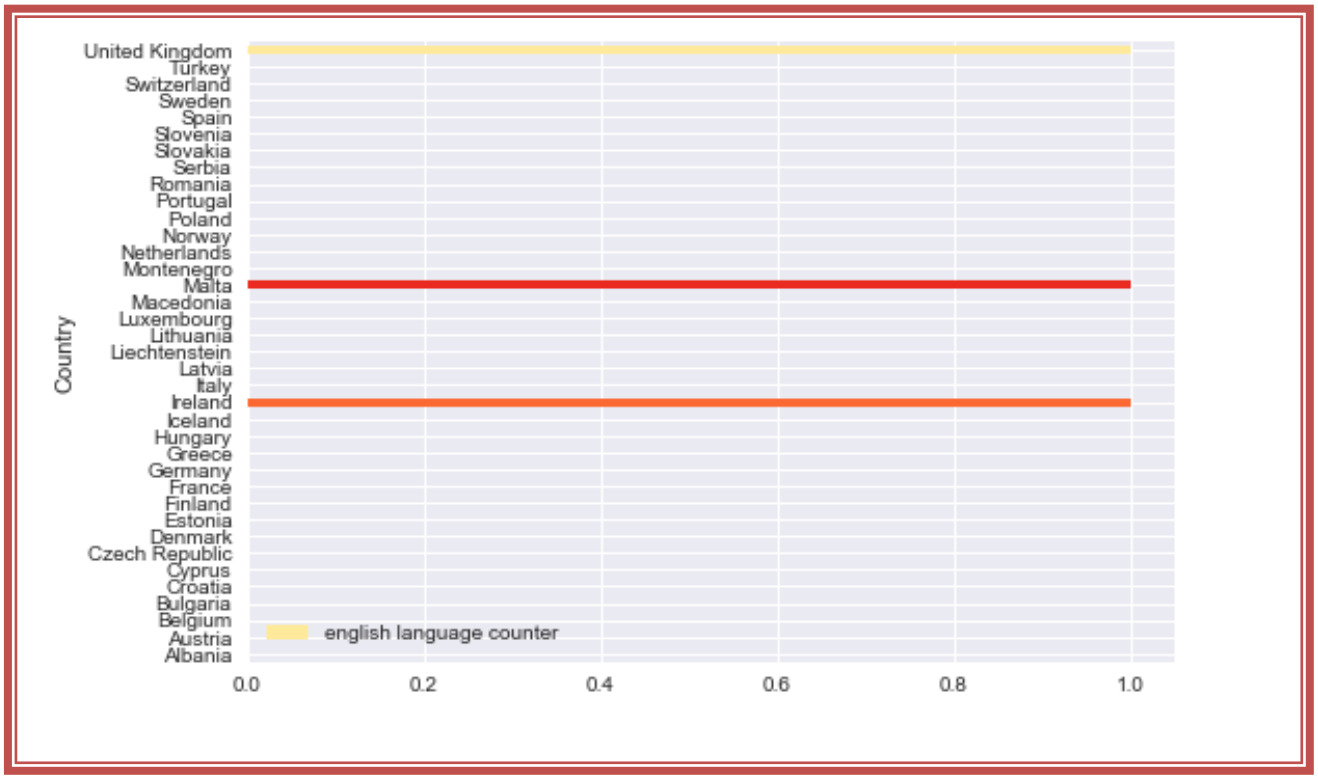


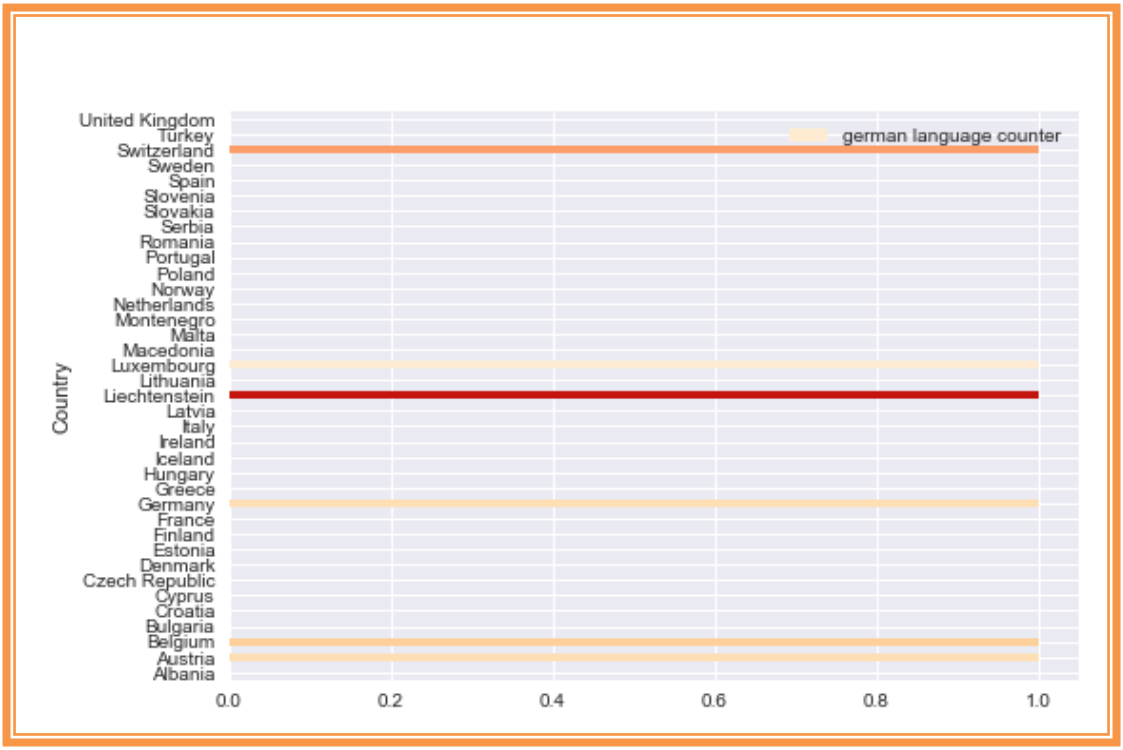

In the barplot above we can see that in year 2004 we had the most entries, note we used different colours for better column separation.

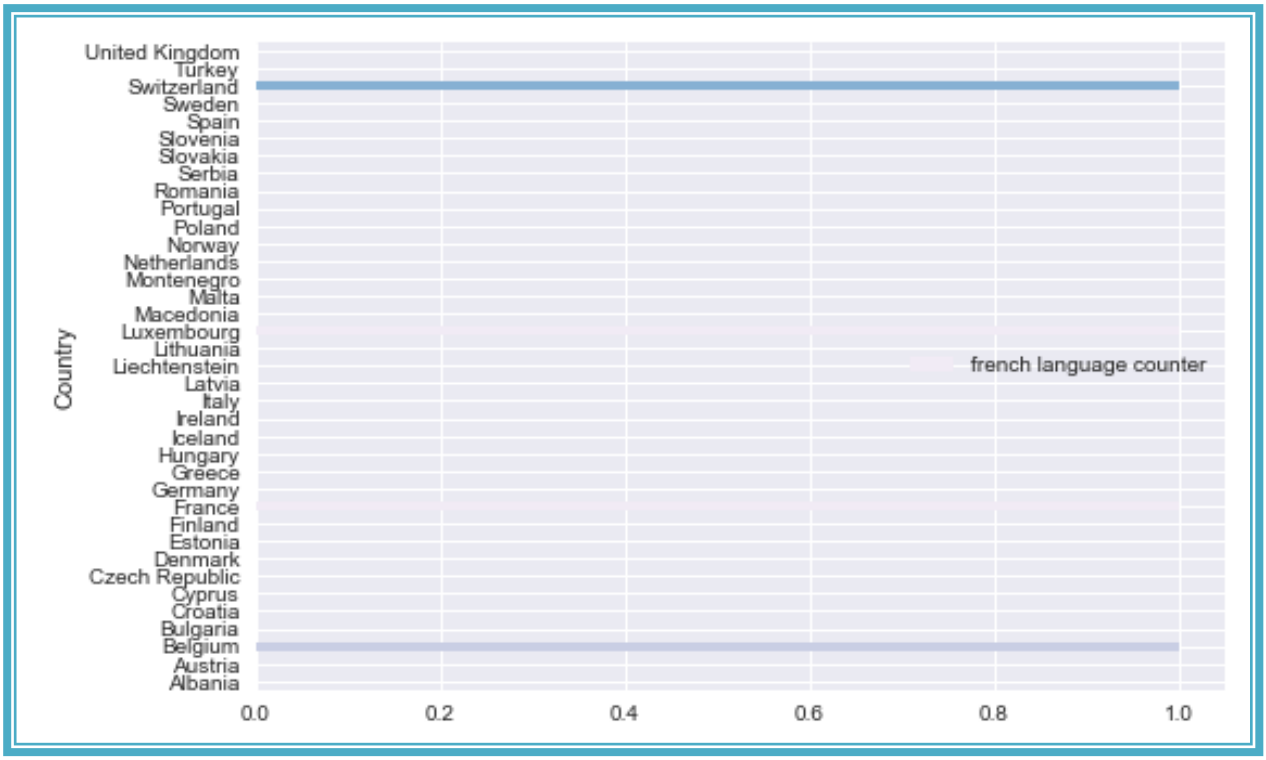
And also which countries are the oldest in the union using a colourful horizontal bar for better comparison:

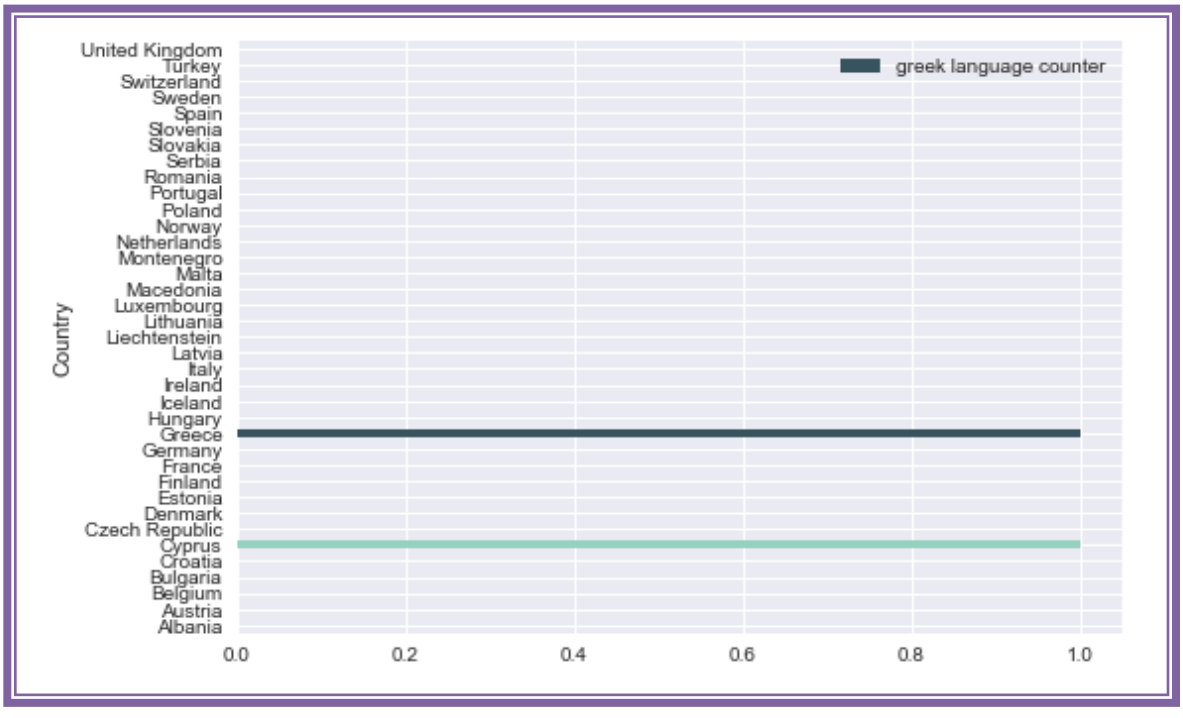


We could also check how many countries speak English,German,French and Greek:

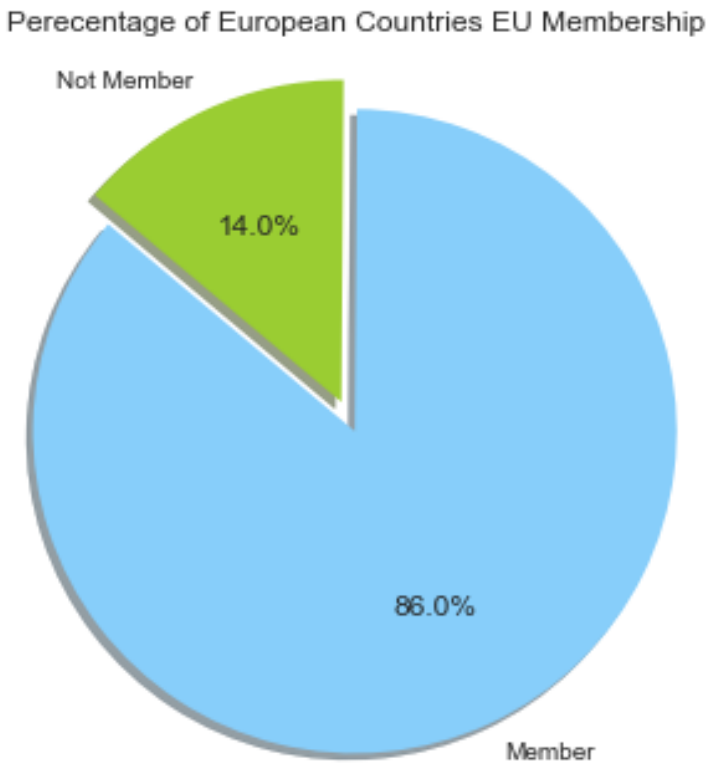


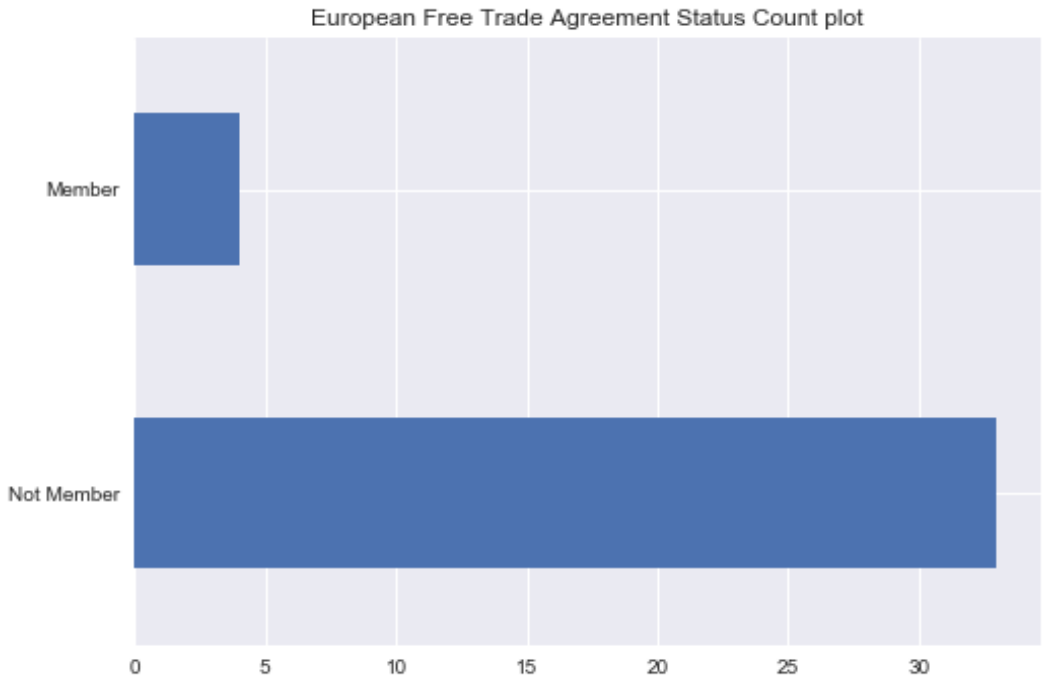

United Kingdom
Turkey
Switzerland
Sweden
Spain
Slovenia
Slovakia
Serbia
Romania
Portugal
Poland
Norway
Netherlands
Montenegro
Malta
Macedonia
Luxembourg
Lithuania
Liechtenstein
Latvia
Italy
Ireland
Iceland
Hungary
Greece
Germany
France
Finland
Estonia
Denmark
Czech Republic
Cyprus
Croatia
Bulgaria
Belgium
Austria
Albania
Country
french language counter
0.0
0.2
0.4
0.6
0.8
1.0


United Kingdom
Turkey
Switzerland
Sweden
Spain
Slovenia
Slovakia
Serbia
Romania
Portugal
Poland
Norway
Netherlands
Montenegro
Malta
Macedonia
Luxembourg
Lithuania
Liechtenstein
Latvia
Italy
Ireland
Iceland
Hungary
Greece
Germany
France
Finland
Estonia
Denmark
Czech Republic
Cyprus
Croatia
Bulgaria
Belgium
Austria
Albania
Country
greek language counter
0.0
0.2
0.4
0.6
0.8
1.0

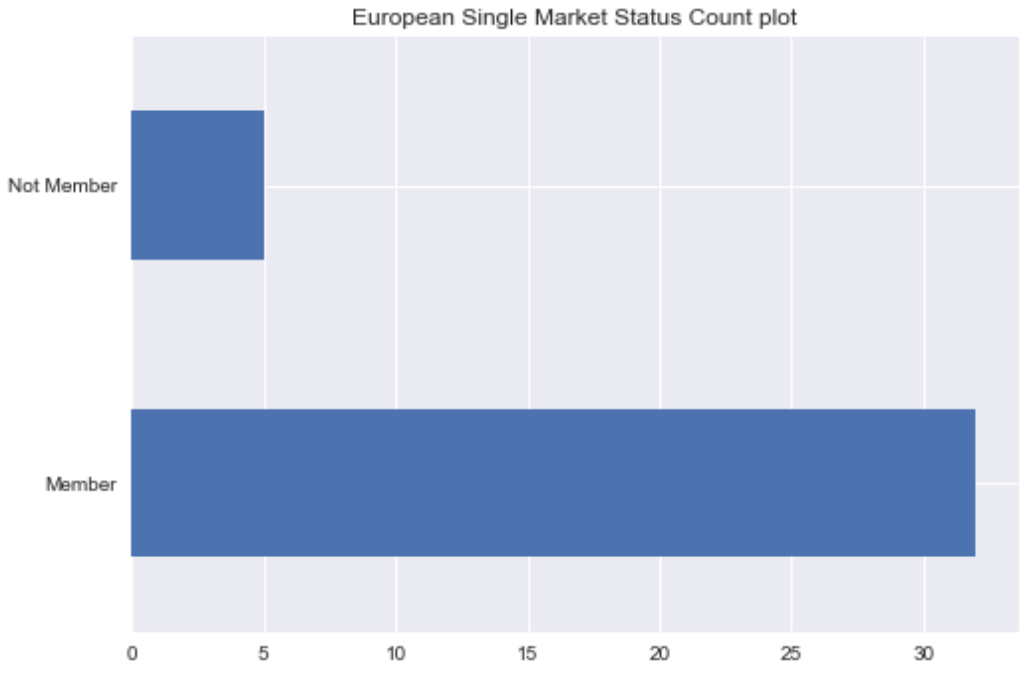
It is also interesting to see the status of the different European organizations like European Single Market, European Monetary Union  and European Free Trade Agreement

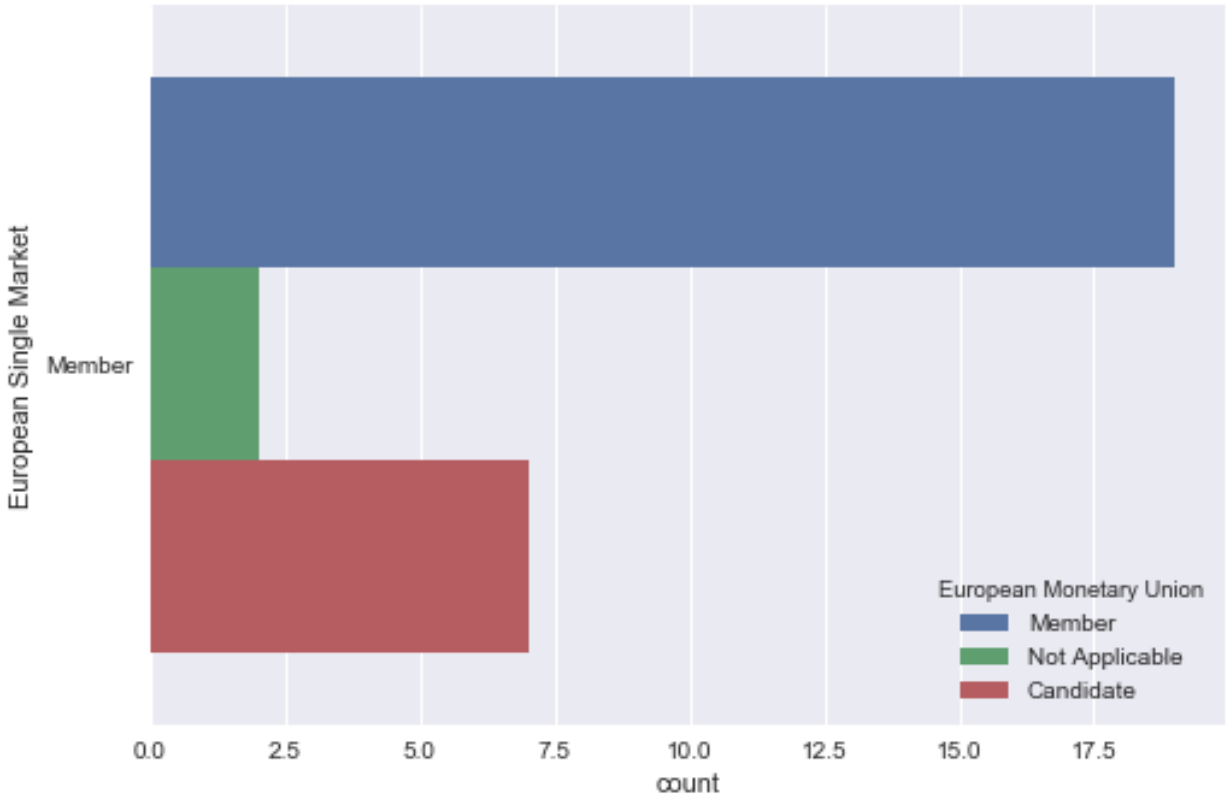


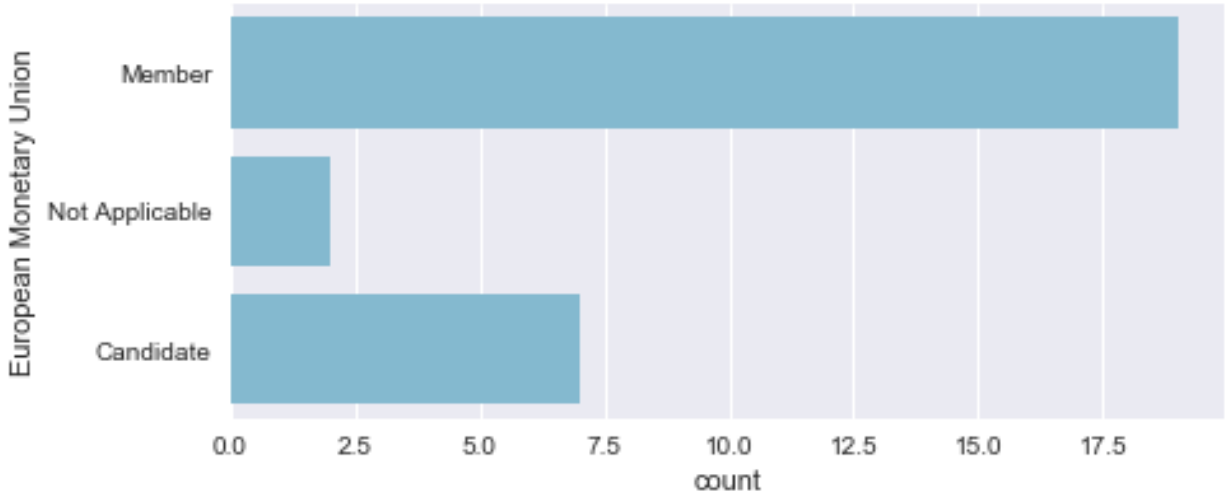

As we see in the piechart 86% of Europe's countries are EU members. In the next  two page we can see different horizontal barplots for the different EU organizations

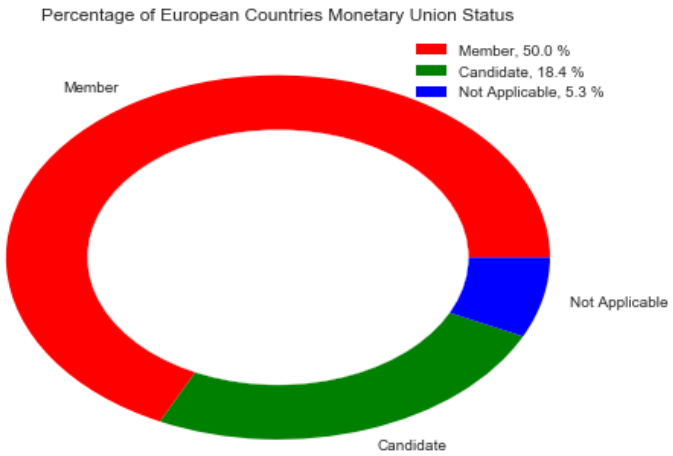



We see most of the European countries have not signed European free trade agreements, that's logical since they are members of eu monetary union ( a country can't be at both)

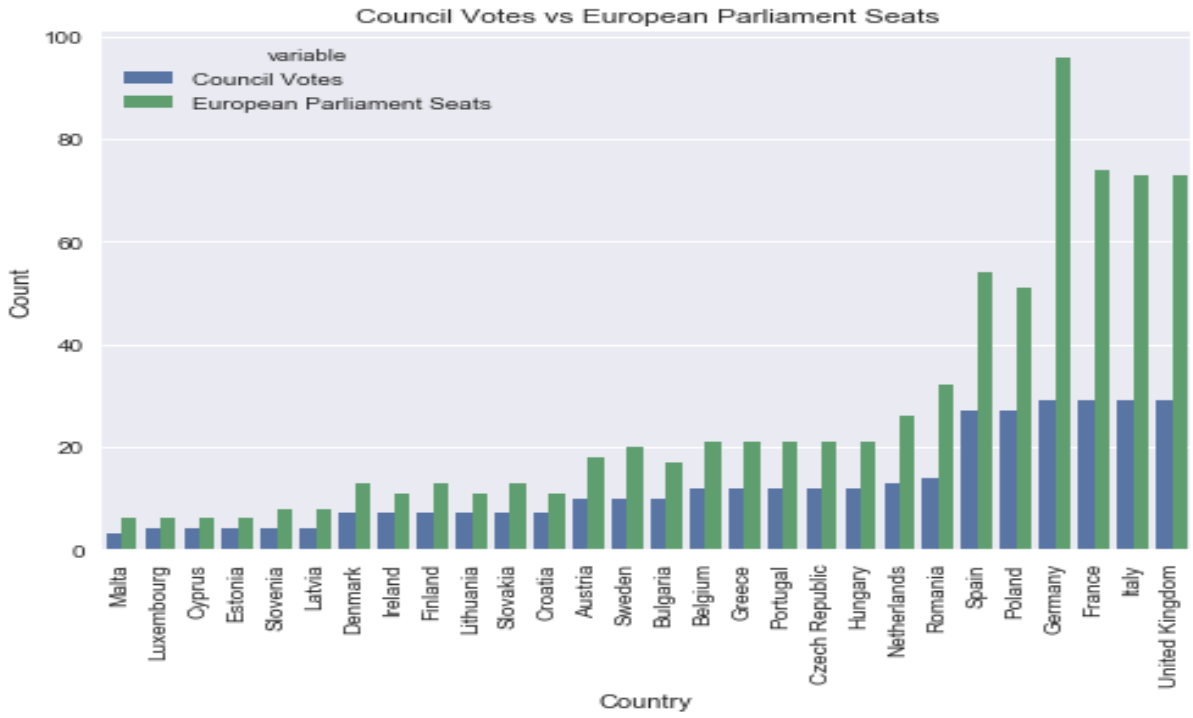



European Single Market
Member
European Monetary Union
Member
Not Applicable
Candidate
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
count


European Monetary Union
Member
Not Applicable
Candidate
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
count

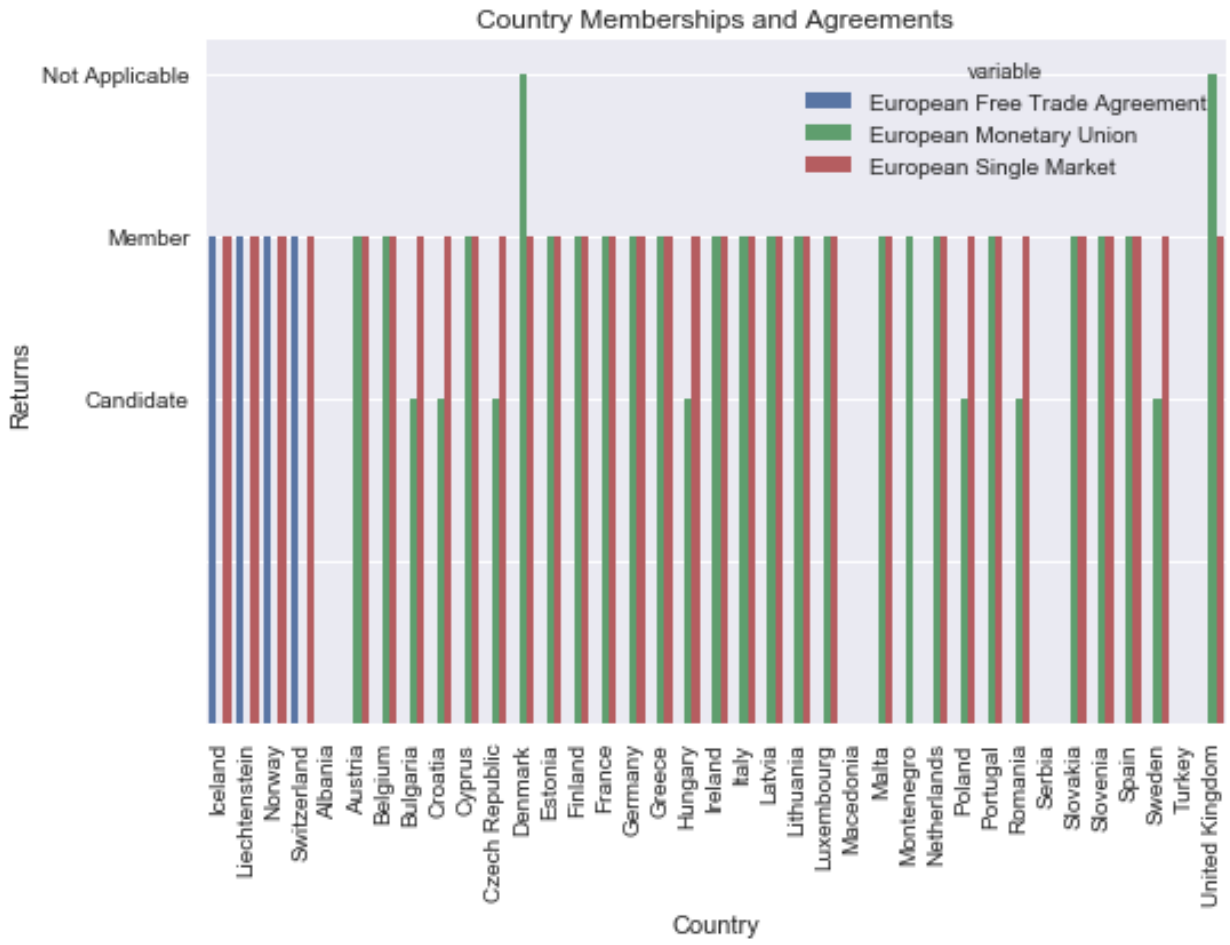



As we see in the previous barplots and the piechart most of the European countries are either members or candidates for  EU organizations.




In addition to our eu organizations plots we can see above a European parliament seats vs council votes countplot for each country

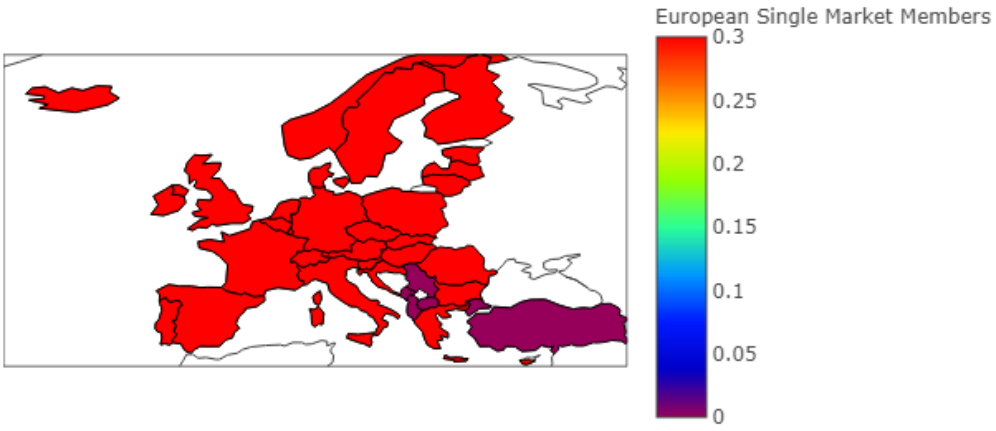
The most “complete” barplot as we can say is the following:




It shows each EU organizations with different colour for each country. The height of each bar shows the status of each country for that particular organization(zero height means not applicable).

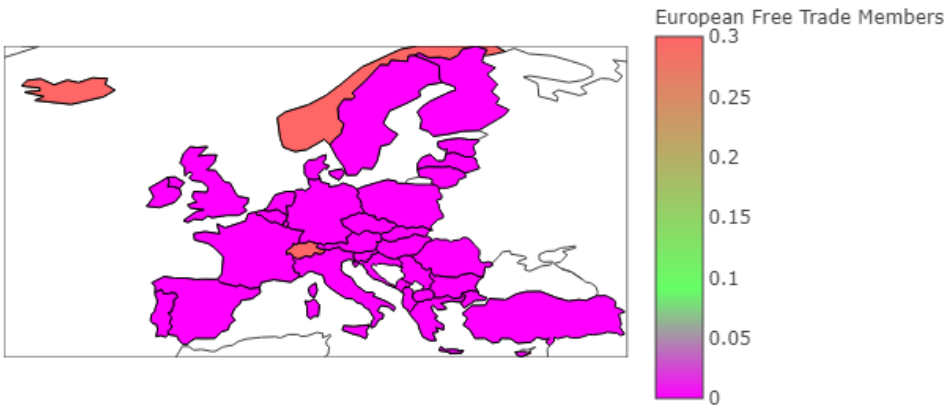
The same information of course could split in three different choropleth maps as we can see below:

European Single Market Members Map




With purple are depicted the countries that are not member of the European Single Market and with red the countries that are members.

European Free Trade Members Map




The same two colour map comparison was applied  for the European Free Trade Members  map. With red we see the members and with purple the countries that are not members.

For the European Monetary Union map we use green for candidates,yellow for members,purple for not members and red for not applicable (because of different coin).