Characterizing Spanish Provinces

Alejandro González Casal

March 07, 2021

1. Introduction

1.1. Background

Nowadays, it is very common that people want to move to another country. It might be due to work requirements or simply because they like the country.

In the latter case, it is usual that people don't know to what part of the country move. However, even if it is the same country the differences could be quite big.

When that situation comes, a lot of people find it very difficult due to the huge amount of options within the same country. To solve that would be very useful have some kind of region classification that reduces the number of places to check. With that tool, people only would have to review each group.

Then almost every region within the group would be very similar, so the decision has less importance.

Figure 1: Spain's map with autonumous communities and provinces

1.2. Problem

Taking into account the situation described in the previous paragraph, Foursquare location data combined with other data might be used to cluster the provinces across the country. The aim of this project is to create this clusters and then make it easier to answer the question "Which region of the country choose?". Precisely, in that project, the country of application will be Spain but it could also be applied to any other country.

1.3. Interest

As for the interest of the project, it is, as noted earlier, to make easier the region selection to people that want to move to another country but don't have a precise idea to which part of it.

2. Data

2.1. Data sources

In order to perform the clustering some data is needed about the provinces to be analysed:

The main sources of this data are two datasets scraped from Wikipedia (https://es.wikipedia.org/wiki/Anexo:Provincias_de_Espa%C3%B1a_por_PIB & https://es.wikipedia.org/wiki/Anexo:Provincias_y_ciudades_aut%C3%B3nomas_de_Espa%C3%B1a). The first dataset included information about GDP and GDP per capita while the second one information about population and area of the province.

Additionally using Foursquare API the data about the number of Industrial States and Power Plants in each province will be extracted. To do so, previously is necessary to retrieve data about the geographical coordinates using Nominatim package from Geopy library.

2.2. Data preparation

Firstly, both datasets were scraped from Wikipedia using Beautiful Soup library. In both cases only some columns from datasets were needed so only those were retrieved to the Dataframe.

In order to achieve the right format several changes had to be done. Line break tokens, punctuation marks and currency symbols were deleted. In the population-

area datasets additional changes had to be performed due to the uncommon format of data.

Next step was merging both datasets using province name as key. In order to do it, the name of “Islas Baleares” needed to be changed to “Baleares” in population-area dataset to match the name of that province in GDP dataset.

As already stated, then, using Nominatim package the geographical coordinates of each province were obtained and then inserted on a new dataframe column.

In that point, two missing values were detected in the area column. As they were few and easily findable, they were manually replaced with their actual values searched on the internet.

After that the type of numeric column was changed from string to float in order to perform mathematical operation among them.

As Foursquare only allows search within a radius some approximations were necessary. To simulate a search inside each province the searching radius was calculated as the equivalent radius, using the circle’s area formula and isolating the radius.

Following this approximation, two types of venues were retrieved from each province’s “area: Industrial States and Power Plants. Using other venues would have been impossible due to the amount of them that would be within a hole province (remember the 100 venues limitation of each Foursquare call).

The result of both searches was stored in two independent datasets for future uses and also using the count function the number of results of each province for each venue was inserted in the main dataset. The value displayed in provinces without any venue of one of the types was NaN and, to solve it, those values were replaced with 0. The type selected for those columns was integer.

Once the data was prepared it was saved as csv toward avoiding repeat this complete procedure each time that analyses be made on this data.

3. Exploratory Data Analysis

This section was focused in get a better data understanding in terms of its relationships and geographical distribution.

3.1. Geographical Distribution

In this subsection some geospatial plots were displayed. The main features studied here were GDP per capita and the number of industrial states and power plants.

3.1.1 GDP per capita

In that case, two plots were proposed. The first one was a bubble representation over the folium map, here the result:

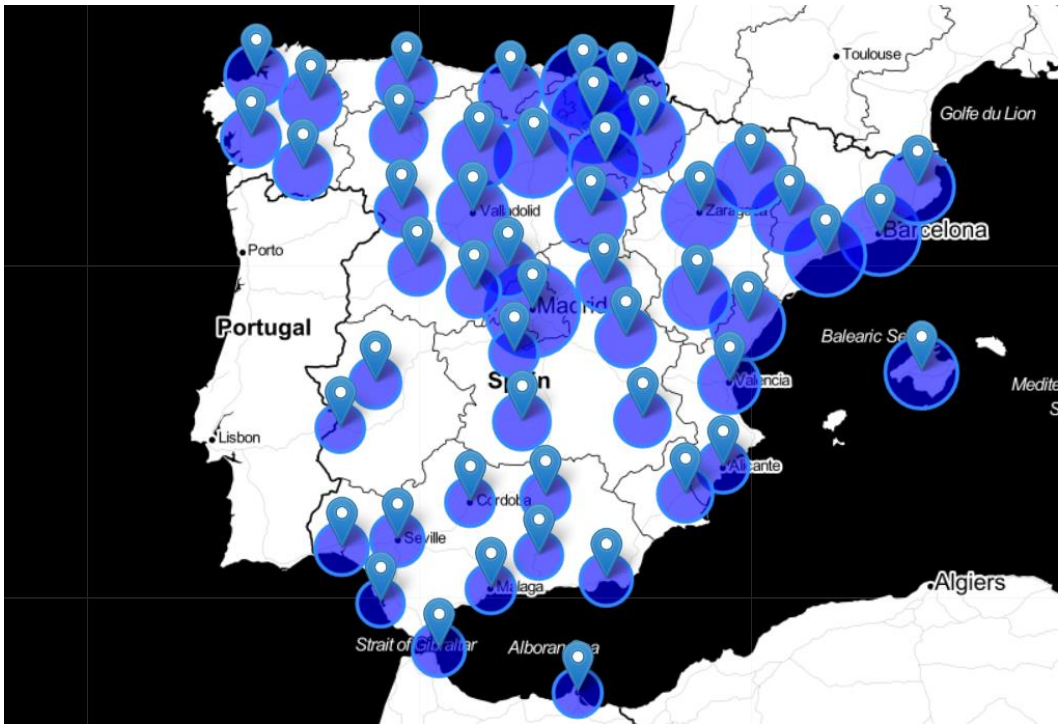

Figure 2: GDP per capita geographical distribution I

The alternative was a choropleth map with the same information. To perform that visualization a geojson file with the Spanish provinces was needed. It was downloaded from the following repository:

https://geographica.carto.com/u/alasarr/tables/spain_provinces/public

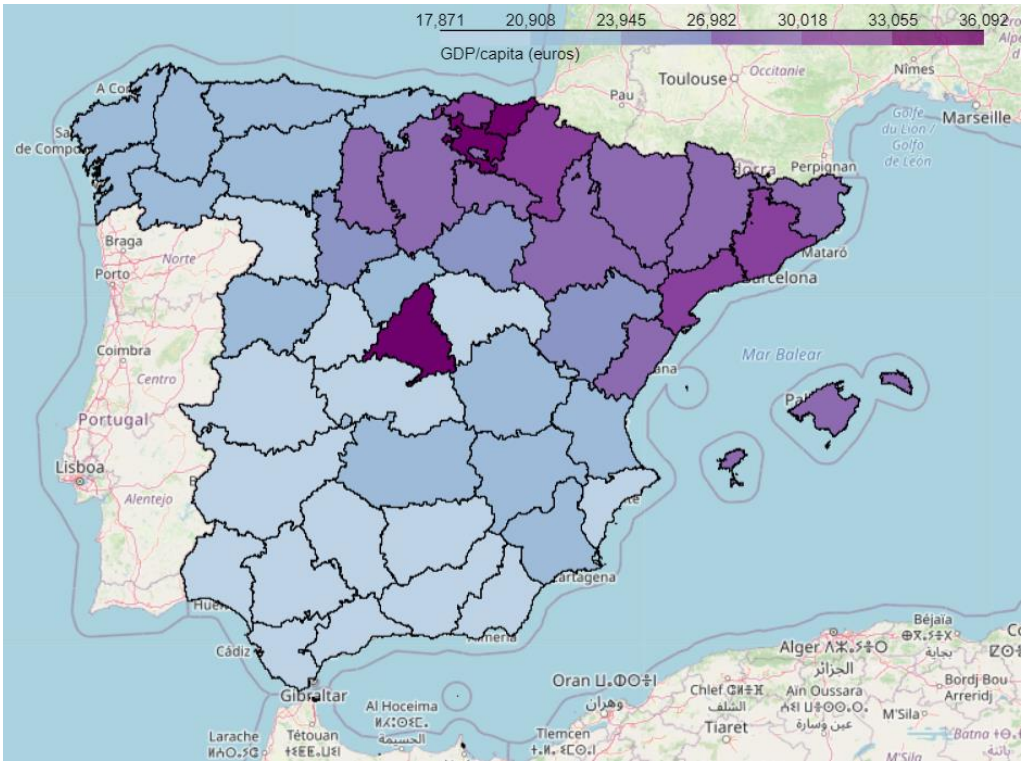

Figure 3: GDP per capita geographical distribution II

In both maps may be seen that the higher PIB per capita is grouped in the northeast corner. An exception to this pattern is Madrid, which holds the higher GDP per capita (35913 €). On the other side southern provinces hold the lower values.

3.1.1 Industrial States and Power Plants location

For this representation two plots were made, one showing industrial states (blue markers) and other showing power plants (red markers).

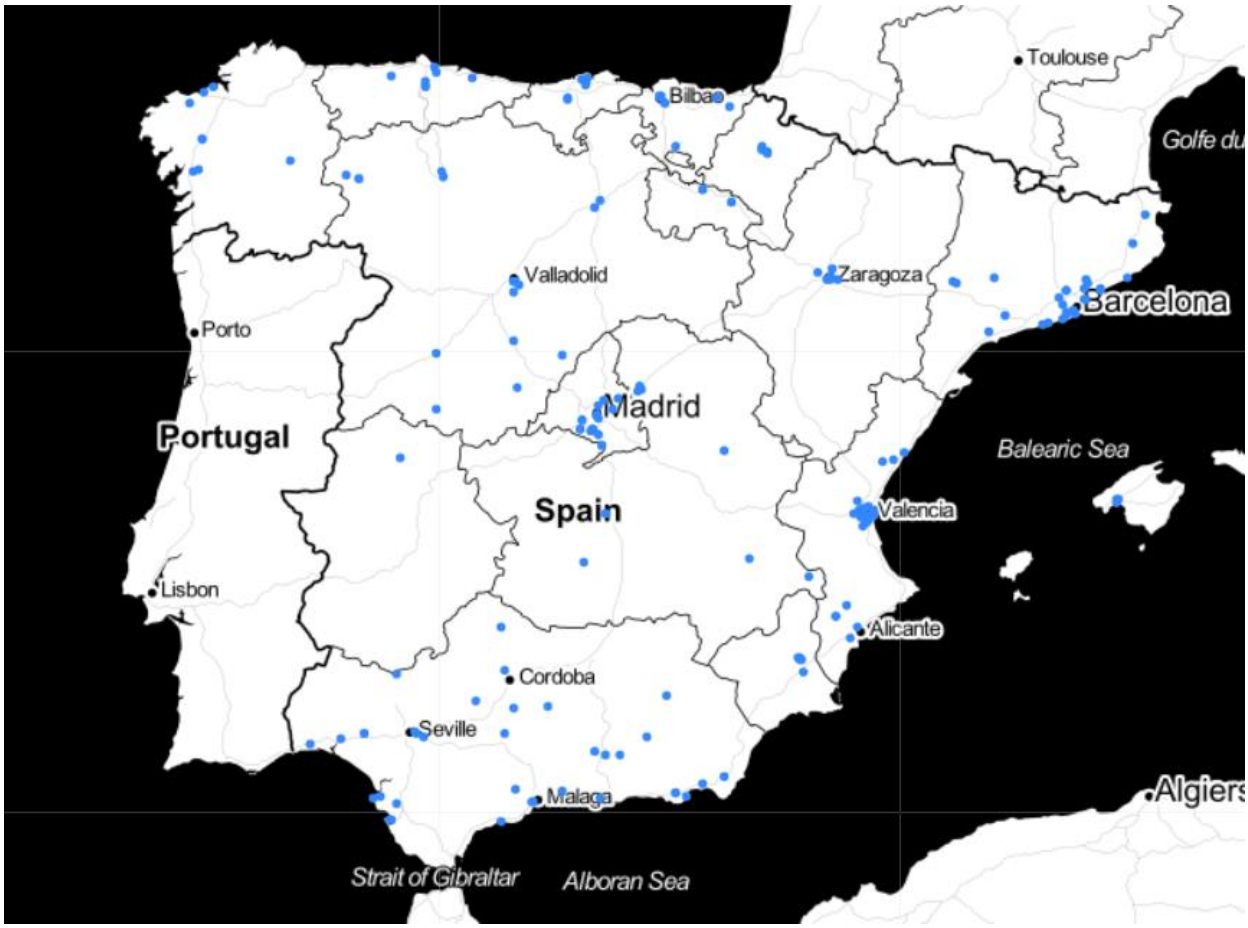

Figure 4: Indusrial State's distribution

In relation to the industrial states situation, they are grouped around the most important cities of each autonomous community. In this respect, it is worth mentioning Madrid, Zaragoza, Barcelona and Valencia; all these cities are within the provinces with its same name.

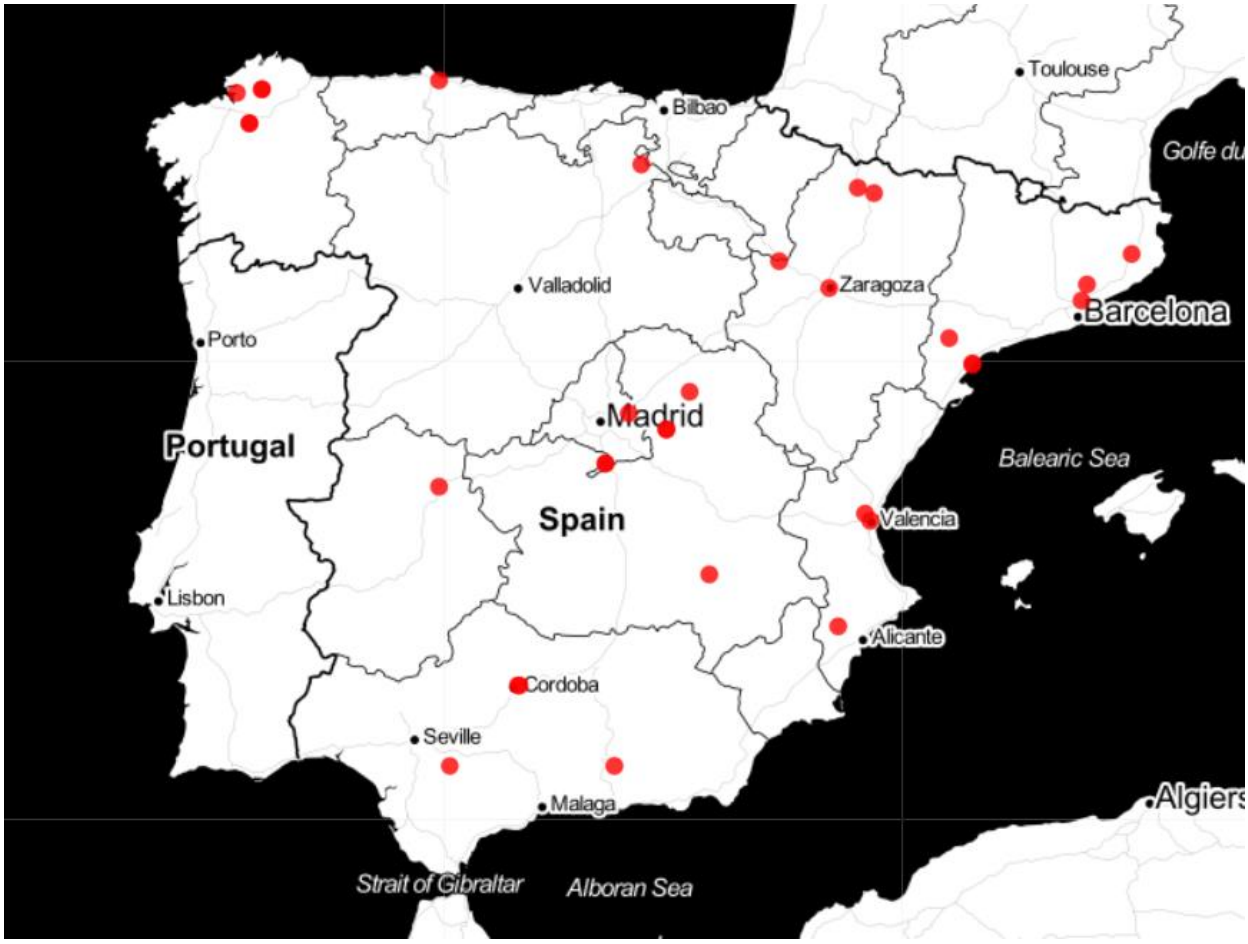

Figure 5: Power Plant's distribution

A similar but less clear pattern is visible in the case of power plants. As can be seen every city previously mentioned has at least one power plant and some more near. The most remarkable difference is A Coruña with three power plants.

3.2. Regression analysis

In order to have an overall picture about the correlation between the selected features several regression plots were done.

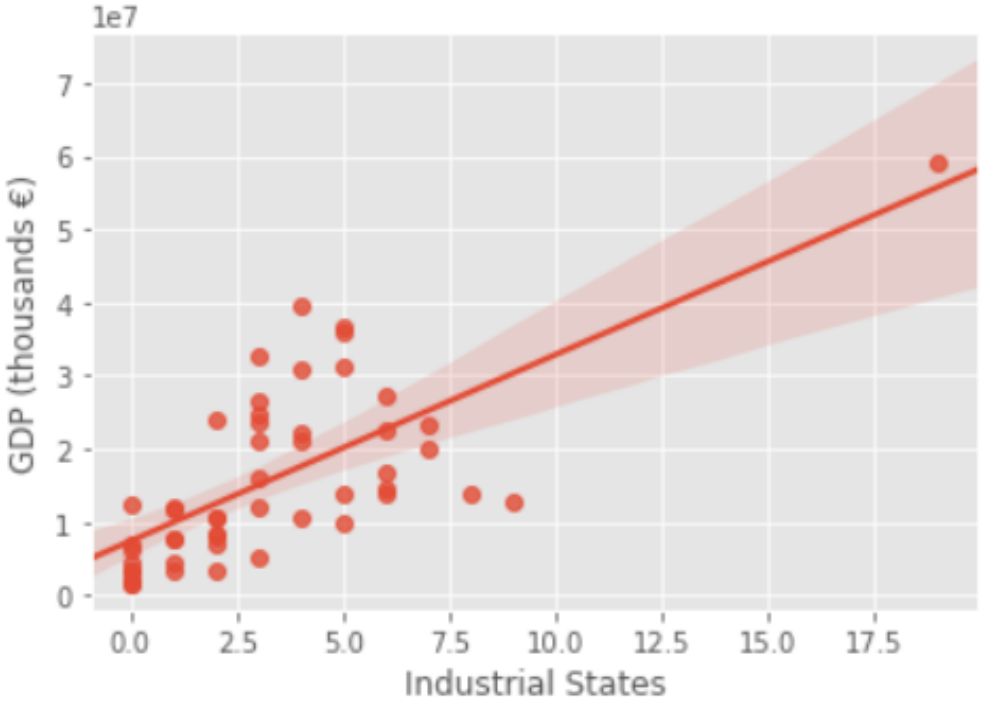

Figure 6: Global GDP vs Industrial States

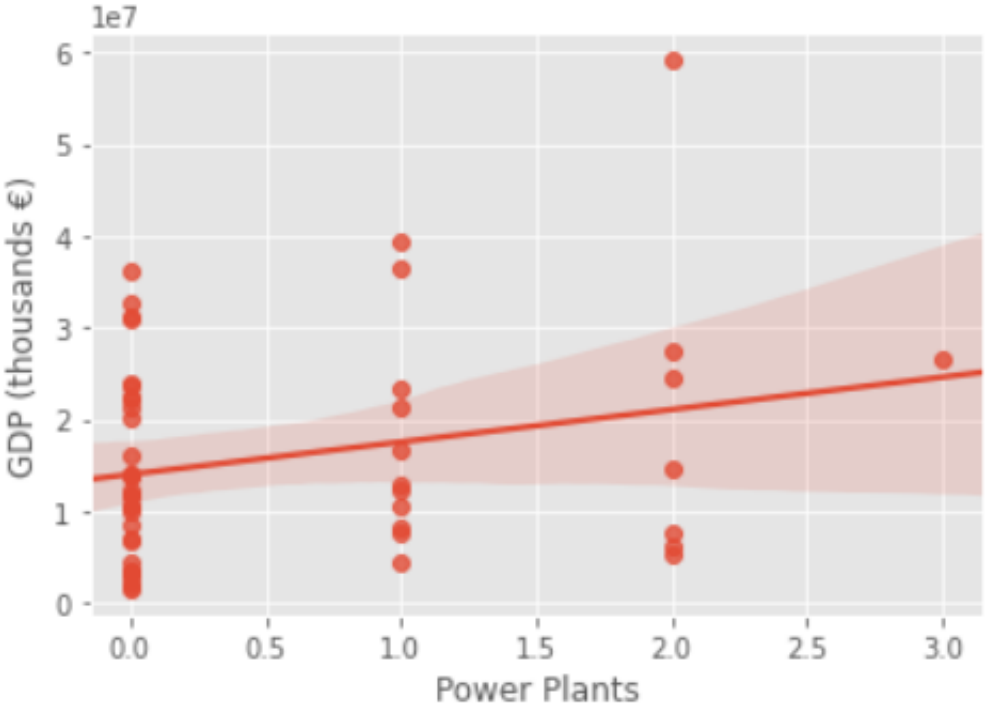

Figure 7: Global GDP vs Power Plants

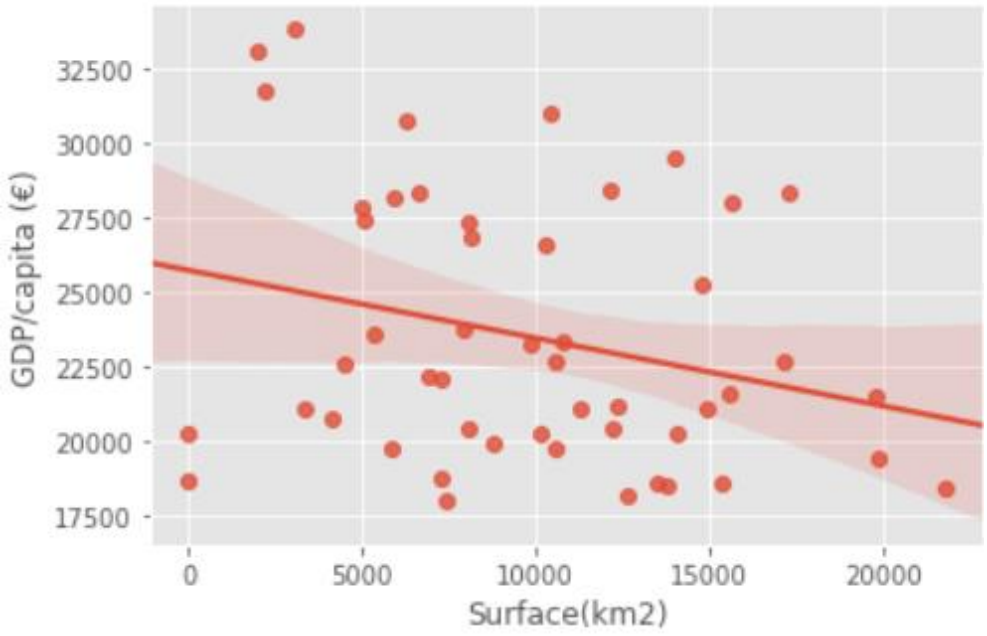

Figure 8: GDP per capita vs Area

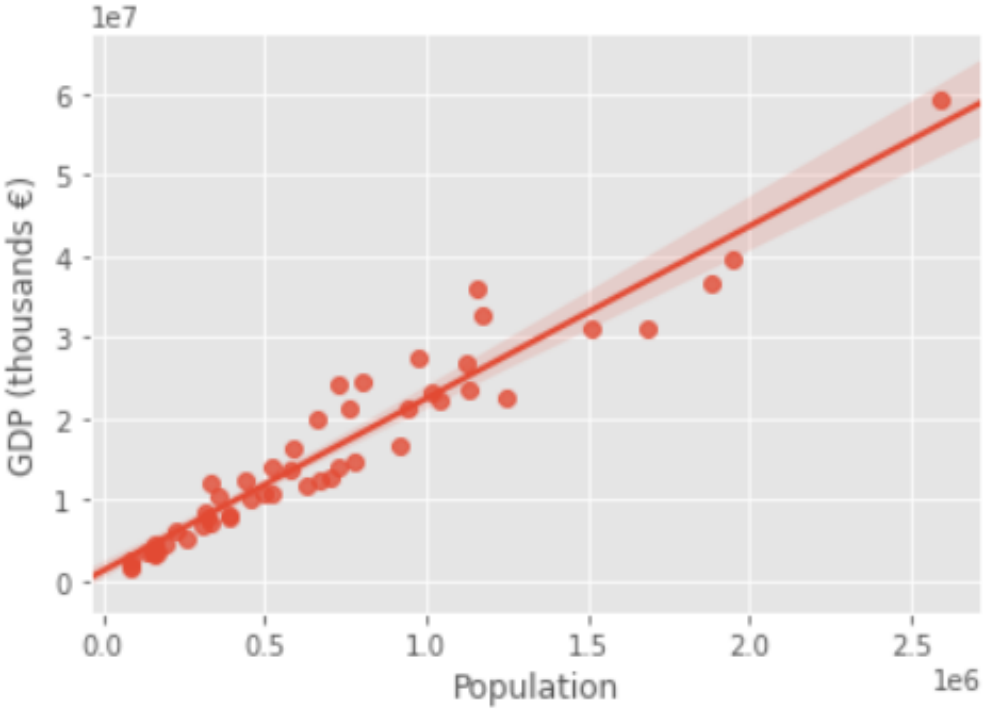

Figure 9: Global GDP vs Population

Watching these results several statement can be done. Firstly, there is a clear relationship between the Population and the global GDP (Figure 9). In the case of the number of industrial states that relation with GDP is weaker (Figure 6) and near inestimable for the power plants (Figure 7).

In the case of GDP per capita it only is clearly related with one feature, the area of the province. As the latter increases the GDP decreases (Figure 8).

4. Province clustering

In this section a clustering technique named Kmeans was applied to provinces' data for the purpose of group them in some types that share common characteristics.

4.1. Feature selection

The first step to apply that technique was choose which of the available features were suitable for the characterization. The following features were selected:

- GDP per capita
- GDP province
- Population
- Area
- Nº Industrial States
- Nº Power Plants

4.2. Feature Scaling

Secondly, the features were scaled. That avoid the predominance of some of them over others. As kmeans is based in distances between each register, a feature with higher value range would have more influence in the global value.

4.3. Clusters quantity selection

For this choice several tests, searching the most suitable cluster distribution. In some of this test even features were dismissed but finally the full previous selection was used.

After those samples, the cluster number was set at four. Once the model was trained the predicted labels were introduced into the dataframe.

5. Discussion - Cluster Analysis

With each province grouped the next step is to study such groups and try to infer its characteristics. To achieve that two representations were used: a choropleth map with the clusters and some box and scatter plots with each feature.

5.1. Choropleth map

To carry out this representation some changes were needed in the default folium.choropleth() function. There was no problem with colour assignment, even if the colour selection is a continuous function it provides clear colours for each cluster. The real problem was the legend which displayed a continuous and not very clear colour-value relation. To handle it the default legend was erased and a new categorical one was created using a function pulled from Stack Overflow (https://stackoverflow.com/questions/65042654/how-to-add-categorical-legend-to-python-folium-map).

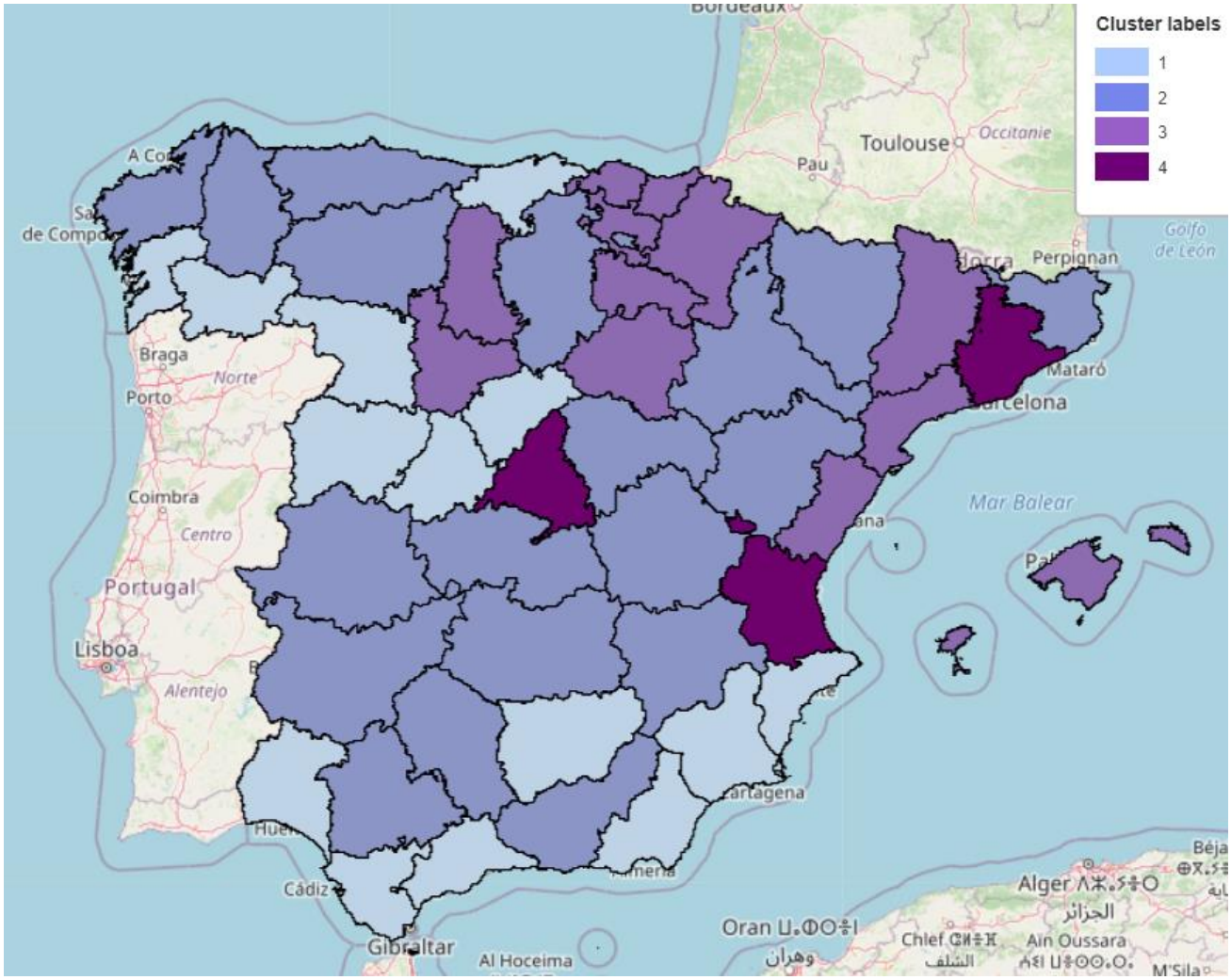

Figure 10: Spain's provinces cluster representation

Using this map several statement can be done:

- Cluster 4 is composed of three provinces (Madrid, Barcelona and Valencia) which were previously noted in the exploratory Data Analysis, when talking about Industrial States location.
- Cluster 3 includes País Vasco's provinces, Navarra, La Rioja, Baleares and some provinces from Cataluña, Valencia and Castilla y León. This province almost match with the provinces with the highest GDP per capita excluding, admittedly, the one includes in cluster 4.
- Cluster 1 and 2 englobe the remaining provinces, although later the difference between these two clusters will be shown. Islas Canarias doesn't appear in the above map, due to its distant location, so is necessary to clarify that it belongs to cluster 1.

5.2. Cluster-related feature distribution

Finally, in order to make a clearer differentiation between each cluster, box and scatter plots were made for each feature. Scatter plots have alphas < 1 for the purpose of make clear the values with higher number of entries.

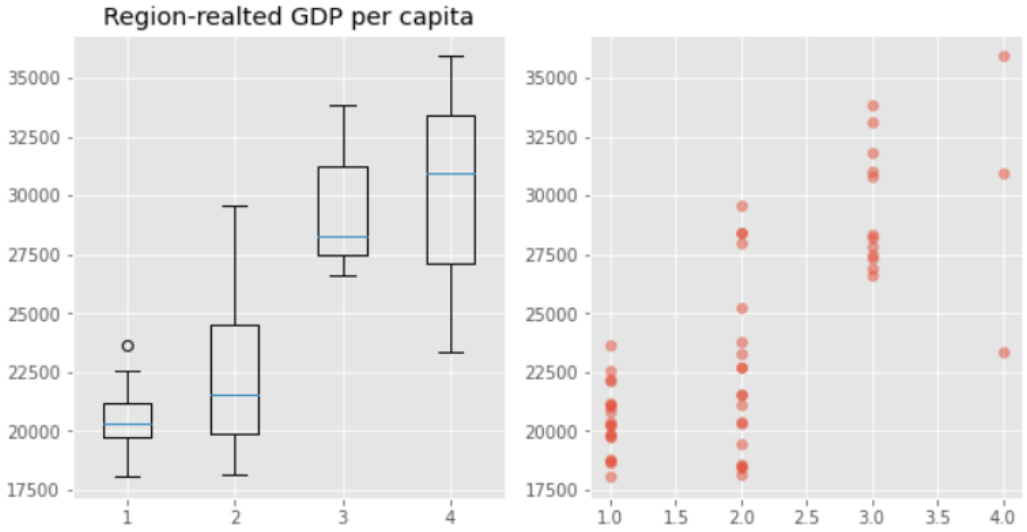

Figure 11: Box and Scatter plots - GDP per capita by cluster

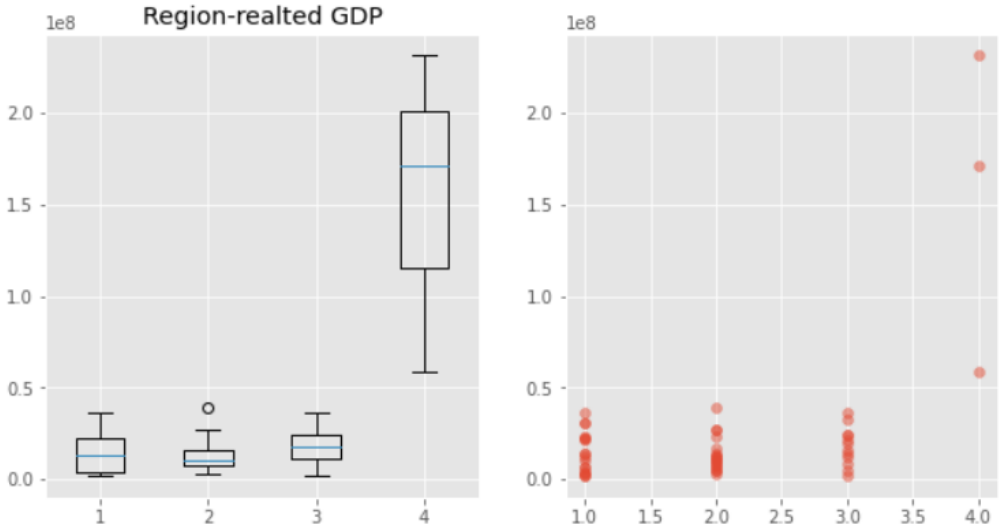

Figure 12: Box and Scatter plots - Global GDP by cluster

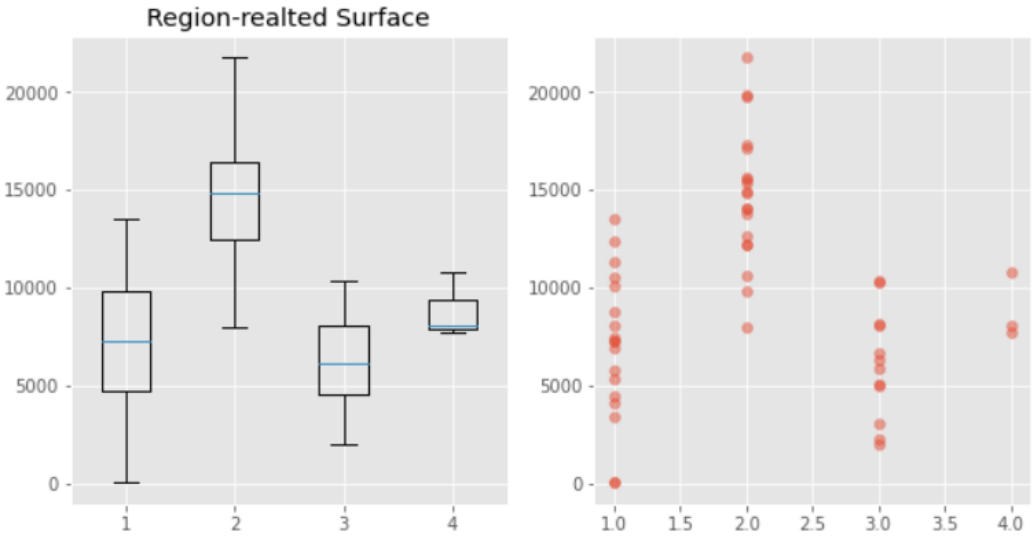

Figure 13: Box and Scatter plots - Area by cluster

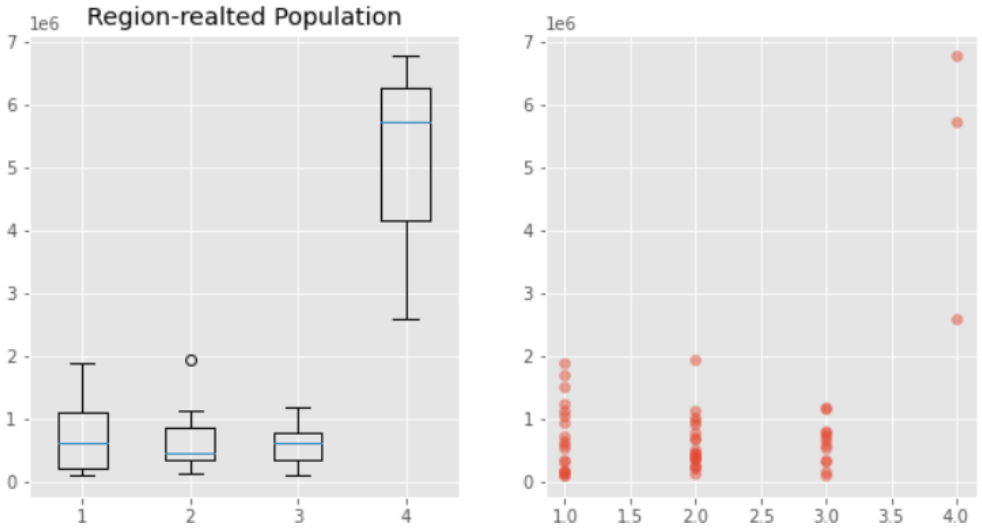

Figure 14: Box and Scatter plots - Population by cluster

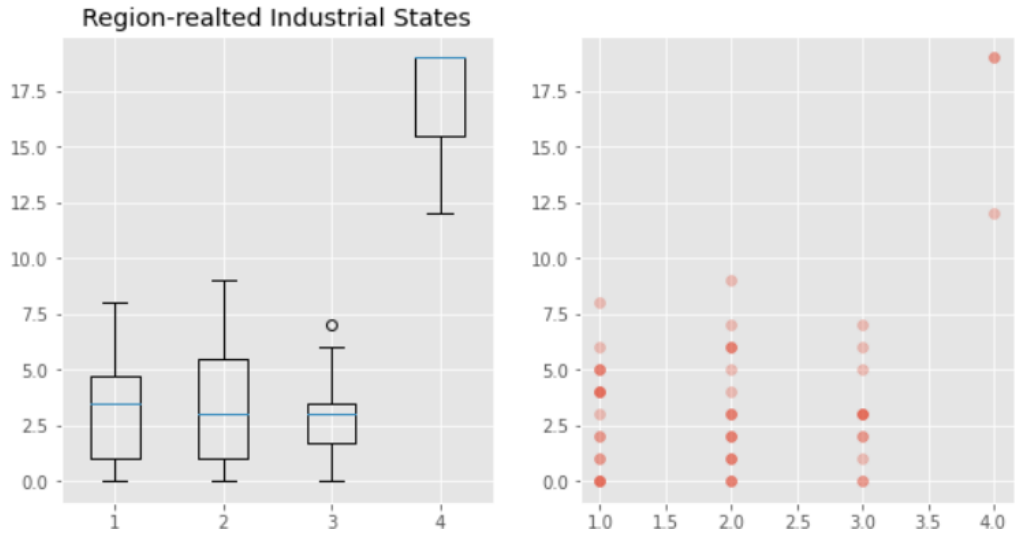

Figure 15: Box and Scatter plots - Industrial States by cluster

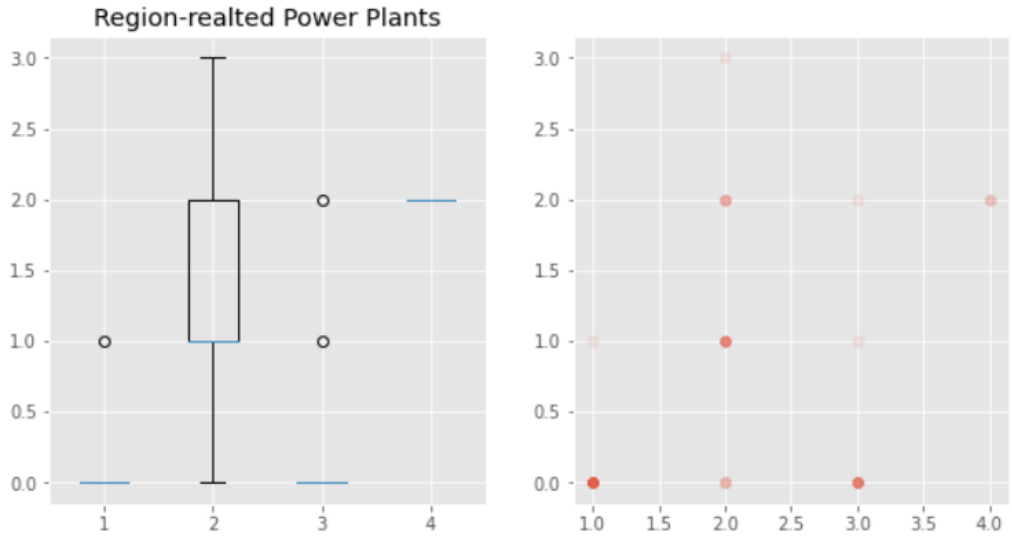

Figure 16: Box and Scatter plots - Power plants by cluster

As was previously noted these graphs show that cluster 4 and 3 have clearly the highest values of GDP per capita, the difference between this two clusters is the global GDP where cluster 4 overtake the other clusters significantly. A similar pattern appears in terms of population and industrial states number.

As for cluster 1 and 2, again is shown their lower values in terms of GDP per capita. They also have similar values to cluster 3 in population, global GDP and industrial states numbers. The difference between these two similar clusters is in terms of area and power plants where cluster 2 has higher values.

6. Conclusion

In this study, similitudes and differences between Spain's provinces were studied. Also, the relationship between some of their features like GDP or industrial states. By doing so, each province was grouped with similar ones. Using a kmeans non-supervised classifier the clustering was done and 4 groups were generated. In simple terms the clusters can be described as follows:

1 - Low GDP provinces with low area and few power plants.

2 - Low GDP provinces with high area and several power plants.

3 - High GDP per capita provinces with low global GDP, population and few industrial states.

4 - High global and per capita GDP with high population and much industrial states.

Knowing these specifications, it will be easier to people to choose a province within Spain without having to check every single province. Achieving that the scope of this project is completed and the subsequent problem made easier.

It must be noted that this analysis made for Spain could be done to other countries or other kind of places if having the same information of them. Also, the analysis could be extended to more features, be they other venue data from foursquare or extracted from alternative dataset. The limitation of venue type was due to the limitation of 100 venues of Foursquare in addition with the large search radius needed. This limitation might be solved using another API.