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.

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 surface 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-surface 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-surface 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 surface 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.