SOCIO-ECONOMICAL ANALYSIS OF MADRID'S NEIGHBORHOODS

INTRODUCTION

As the financial capital in Southern Europe, Madrid is a safe and stable environment for companies to grow, and the chosen headquarter location of 2000 companies. Driving the young talent, Madrid offers extensive opportunities for further education, with 17 universities and over 30 research centers. With over 75 million tourists visiting Spain every year, the country occupies a significant worldwide economic position. It is the fourth metropolis in the EU by Gross Domestic Product thanks to leading employers such as Telefónica, Iberia, and BBVA. Sharing frontiers with 8 countries by land and sea, Spain is a logical destination for international trade, facilitated by the country's high-speed rail system, the second-longest network in the world.

The city is divided into 21 districts and 131 neighborhoods. At first, we planned to use the Foursquare API, but we realized that there isn't enough data for some neighborhoods. Fortunately, Madrid has an excellent City Council Open Data Website, that provides all kinds of socioeconomic data for the city. Madrid also has an Statistics Portal with so many useful data that is hard to choose just a few indicators. We will use datasets provided for both websites to perform a socio-economic analysis of Madrid and classify their neighborhoods based in their most common venues, the size and age of their population, their middle income and information provided by the land registry.

We consider that this is information can be useful for companies that want to invest in the city. For example, if someone wants to open a toy store, it will be able to choose the best part of the city to do it thanks to this analysis.

Data

As I said earlier, the City Council Open Data website has lots of information, for this project we are going to use:

-Registry of venues and the activity that is performed in them. This is a huge dataset with more than 163000 rows and 46 columns. The most important columns are: venue id, district, neighborhood, venue situation ('open', 'closed',…) and description of the activity ('hairdresser service', 'bar', 'restaurant',...). It needs a lot of data cleaning and preprocessing to extract the relevant information, all the open venues in each neighborhood, and their activities. This information will be used to cluster the neighborhoods. Link

-Demographic indicators by neighborhood in 2016. It shows the total population, the mean age, the percentage of the population that it's 65 years old and older... We consider that population and mean age are relevant data that can help us to classify the neighborhoods. Link

-The middle income per person and neighborhood. We will use it as well to classify the neighborhoods. For example, a company that sells luxury products needs a customer base with a high middle income. Link

-Land registry data. It offers data of the land uses for each neighborhood (commercial, offices, hospitality,..) along with the average land registry value for each kind of use. We want to add this information to the analysis because it gives an approximate idea of the cost of a commercial or office venue. Link

-Table with the coordinates of each neighborhood. This information was obtained using geopy and Nominatim and will be use to create a map of the neighborhoods and its clusters.

-Geojson file of Madrid that will be used in choropleth maps. Link

After cleaning and preprocessing all the datasets, we will combine them in a pandas dataframe to perform cluster analysis of the neighborhoods. We will represent the clusters in a map, alone and combined with choropleth maps.