Coursera Capstone

IBM Applied Data Science Capstone

Identifying real estate opportunities in Madrid, Spain

By Juan Carlos Martinez
May 2020

Introduction

For the average family buying a house is a life-change decision and in most of the cases a one time event. With the increase of population among the cities, and therefore the demand and prices of houses, this decision is getting even tougher.

All this factors are not new, but the recent world pandemic due to the COVID virus is going to affect our lives in a way that we are not even certain yet, including the way of investing in a new house, changing the rules of the game.

Prior to this emergency, we could assume that the closer a house is to the city center the higher the price, as most of the business are in this area and the main venues of the city, but now companies understand that it is indeed possible for a high jobs percentage to work from home, changing the way the current real estate market works.

Business Problem

The purpose of this capstone project is to analyze the different Madrid neighborhoods in order to find similitudes, and to compare the price of the houses, so we can understand which areas are underpriced.

Target Audience of this project

This project is useful for real estate investors as well to individuals who are looking to buy a new home. It has been developed now under the assumption that work conditions are going to drastically change, making possible to live further away from our place of work. This combined to the high rents of the capital of Spain, should be a big enough stimulus to consider moving to other neighborhoods in the city.

Data

To solve this problem we will need the following data:

- List of neighborhoods in Madrid. This defines the scope of this project which is confined to the city of Madrid, the capital city of the country of Spain.
- Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and also to get the venue data.
- Venue data for the different neighborhoods, in order to see the similitude of the different areas.
- Housing prices for the different neighborhoods. We will use the price per square meter to compare economic situation of the real estate sector in the different areas.
- Neighborhoods geometry and position, to be able to divide the city in its different parts.

Sources of data and methods to extract them

1.-We will use Wikipedia (https://es.wikipedia.org/wiki/Anexo:Barrios_administrativos_de_Madrid) to get the list of Madrid city neighborhoods.We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and wikipedia package.

2.-We will get the house prices per neighborhood from idealista (https://www.idealista.com/sala-de-prensa/informes-precio-vivienda/venta/madrid-comunidad/madrid-provincia/madrid/). Idealista is the principal real estate portal in Spain.

3.- We will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

4.-We will use the Foursquare API to get the venue data for those neighbourhoods. Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers.

5.- We will use the geojson given by the Madrid statistics web (https://datos.madrid.es/portal/site/egob/menuitem.9e1e2f6404558187cf35cf3584f1a5a0/?vgnextoid=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default)

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.