**Data Science Capstone Project**

**Housing Prices and Social Venues of Buenos Aires City**

***By Matias Rapagna***

***June 22 2020***

# Index

[1. Index 2](#_Toc43744275)

[2. Introduction 3](#_Toc43744276)

[2.1 Description & Disscusion of the Background 3](#_Toc43744277)

[2.2 Data Description 3](#_Toc43744278)

[3. Methodology 5](#_Toc43744279)

[4. Results 11](#_Toc43744280)

[5. Discussion 12](#_Toc43744281)

[6. Conclusion 13](#_Toc43744282)

[7. References: 14](#_Toc43744283)

# 

# Introduction

## Description & Disscusion of the Background

Greater Buenos Aires, also known as the Buenos Aires Metropolitan Area, refers to the urban agglomeration comprising the autonomous city of Buenos Aires and the adjacent 24 partidos (districts) in the Province of Buenos Aires. Thus, it does not constitute a single administrative unit. The conurbation spreads south, west and north of Buenos Aires city. To the east, the River Plate serves as a natural boundary.

Urban sprawl, especially between 1945 and 1980, created a vast conurbation of 9,910,282 inhabitants in the 24 conurbated partidos, as of 2018, and a total of 14,967,000 including the City of Buenos Aires, a third of the total population of Argentina and generating more than half of the country's GDP. This makes Buenos Aires, the 13th largest city in the world.

To try to simplify the exercise, I will focus only in the City of Buenos Aires, the capital of Argentina. It is divided in 48 districts, has a 203 square kilometers of area, and 3,075,646 inhabitants, with a population density of 15,151 hab/m2.

As you can see from the figures, Buenos Aires is a city with a high population and population density. When making a decision to purchase a house or apartment in the city, investors would want to choose a district with low property value, but with hi social venues density. Today, there is no consolidated information regarding this topic.

When we consider these problems, we can create a map and information chart where the real estate index is placed on Buenos Aires and each district is clustered according to the venue density.

## Data Description

To consider the problem we can list the datas as below:

* I have found Geo Referential Data regarding districts in Buenos Aires, from a National Government site <https://datosgobar.github.io/georef-ar-api/download/> where you can obtain json files to use in your code. I obtained also the information regarding each district to use as a base data frame. I will obtain 2 files, one is a GeoJson to be able to draw Choropleth map of the districts in Buenos Aires based on the prices of Sq M2. The other file contains latitude and longitude data for each of the districts.
* I had to create an Excel data with the average price per square meter in each of the districts. This was based on information obtained by <https://data.buenosaires.gob.ar/>, from the City of Buenos Aires government. This file will be merged with the Location information of each of the districts and the top 10 venues.
* I used **Forsquare API** to get the most common venues of given each of the districts in Buenos Aires.
* I also used Wikipedia information regarding general data of Buenos Aires City, Greater Buenos Aires, and other useful data.

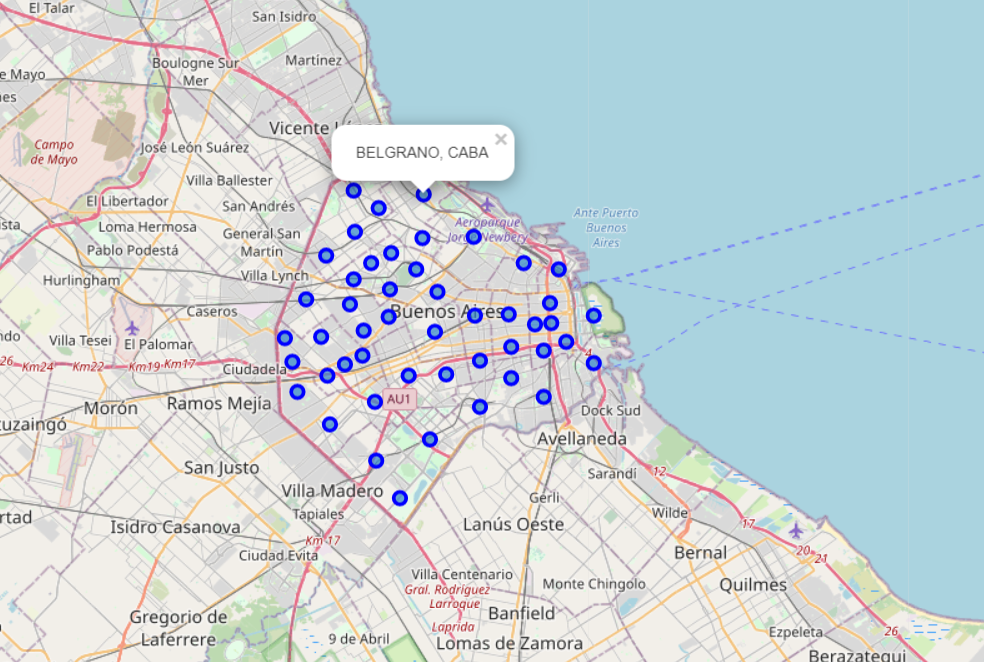
# Methodology

My master data which has the main components *District, Latitude* and *Longitude, and average prices per square meter.*



Localidad is the column that contains District name.

I used python **folium** library to visualize geographic details of Buenos Aires and its districts. Based on the information obtained from national government data, I was able to draw the map, with each of the districts, with their correspondent labels.



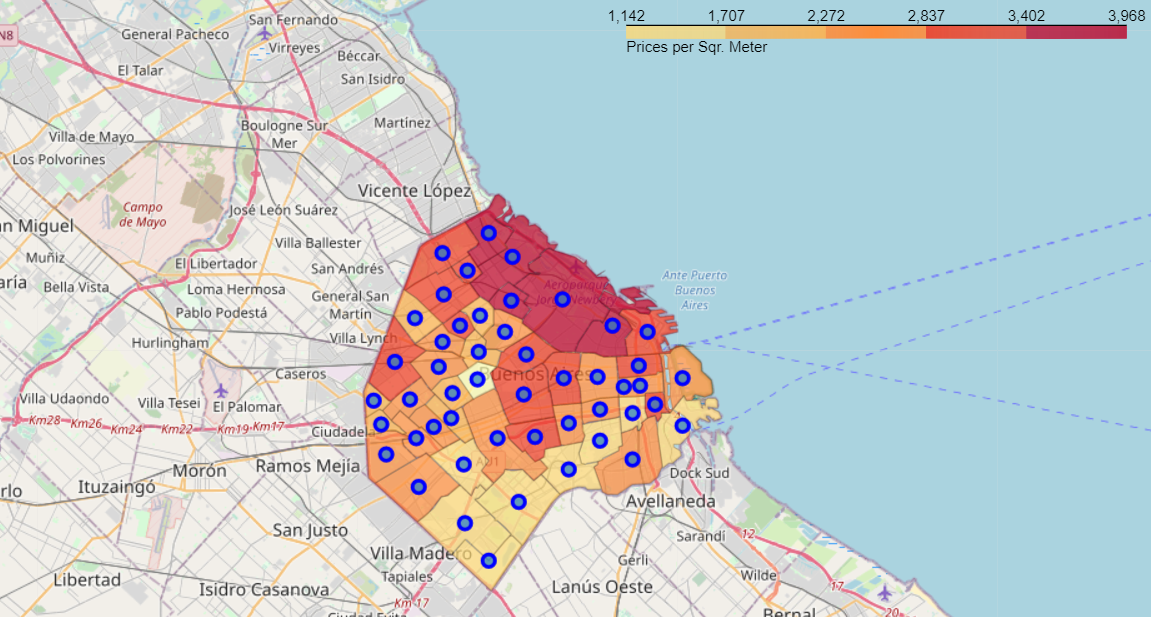
Using the information of the average price per M2, I uploaded the data:



And I Merged this information, with the geo data of each district.

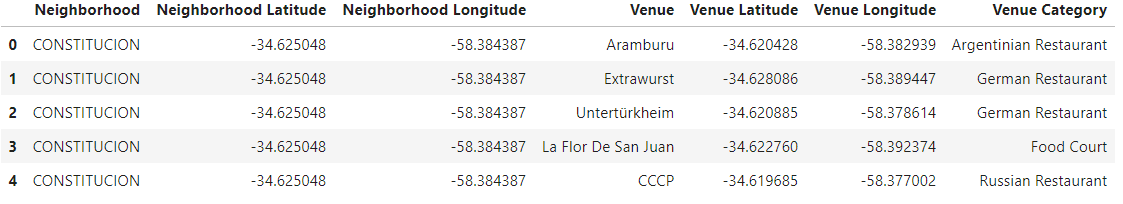


This information allowed me to create a Choropleth Map, with price information:

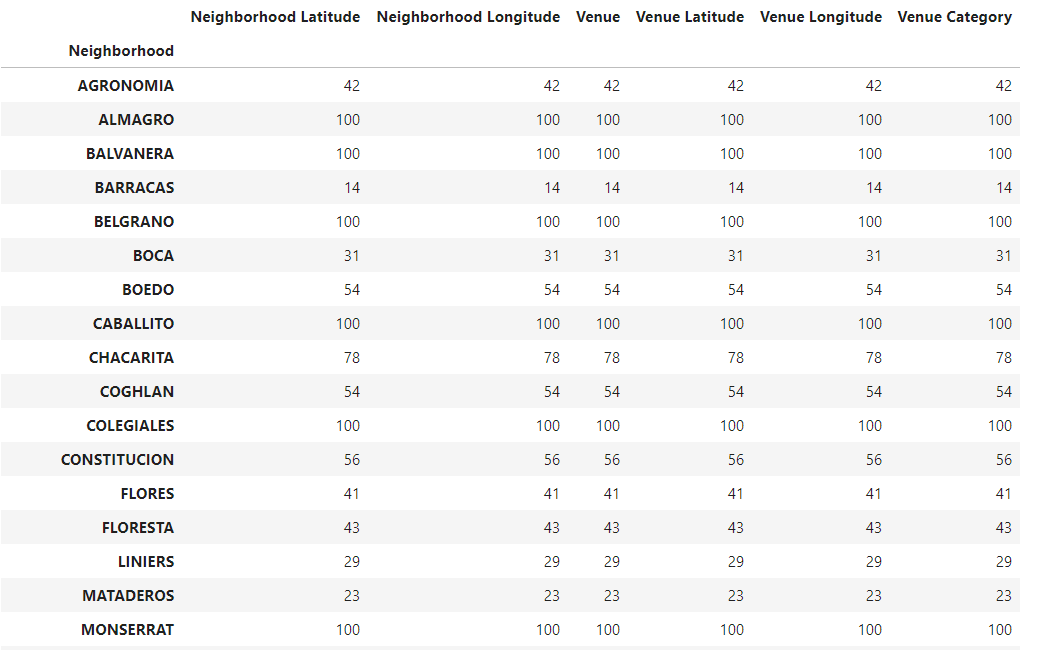


With this done, we have half of the needed data to perform our analysis. We need to complete it with the social venues for each district. To complete this, I used the Foursquare API to explore each of the districts in the city, and obtain venues in a radius of 1000 meters from the coordinates of the district.

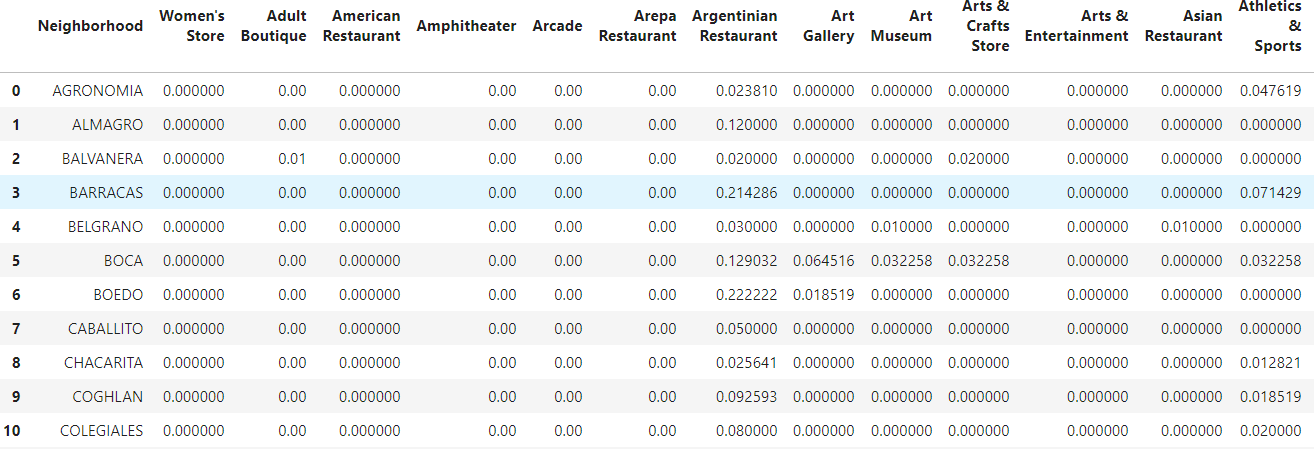
I was able to retrieve 2669 venues for the 48 districts.



Even though I established a limit of 100 venues per district, we can see that not all districts have that many venues, which can be an important piece of information, as could indicate districts that do not have many social venues nearby.



I obtained  **251** unique categories from Foursquare, and I was able to establish the frequence of each one in all the districts:

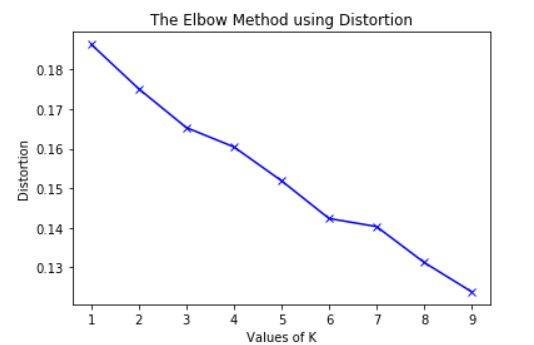


Once I was able to do this, I was able to establish the 10 most common Venues of each district:



We have common venue categories in the districts. In this reason I used unsupervised learning **K-means algorithm** to cluster the boroughs.

To be able to do so, I needed to select the best K for the K-Means algorithm, so I used the Elbow method:



Unfortunately, there was no significant curve in the graph, so I established 6 as my K for the K-means, as it was the point where the line started to have a less steeper curve.

I established the Cluster Label for each district:



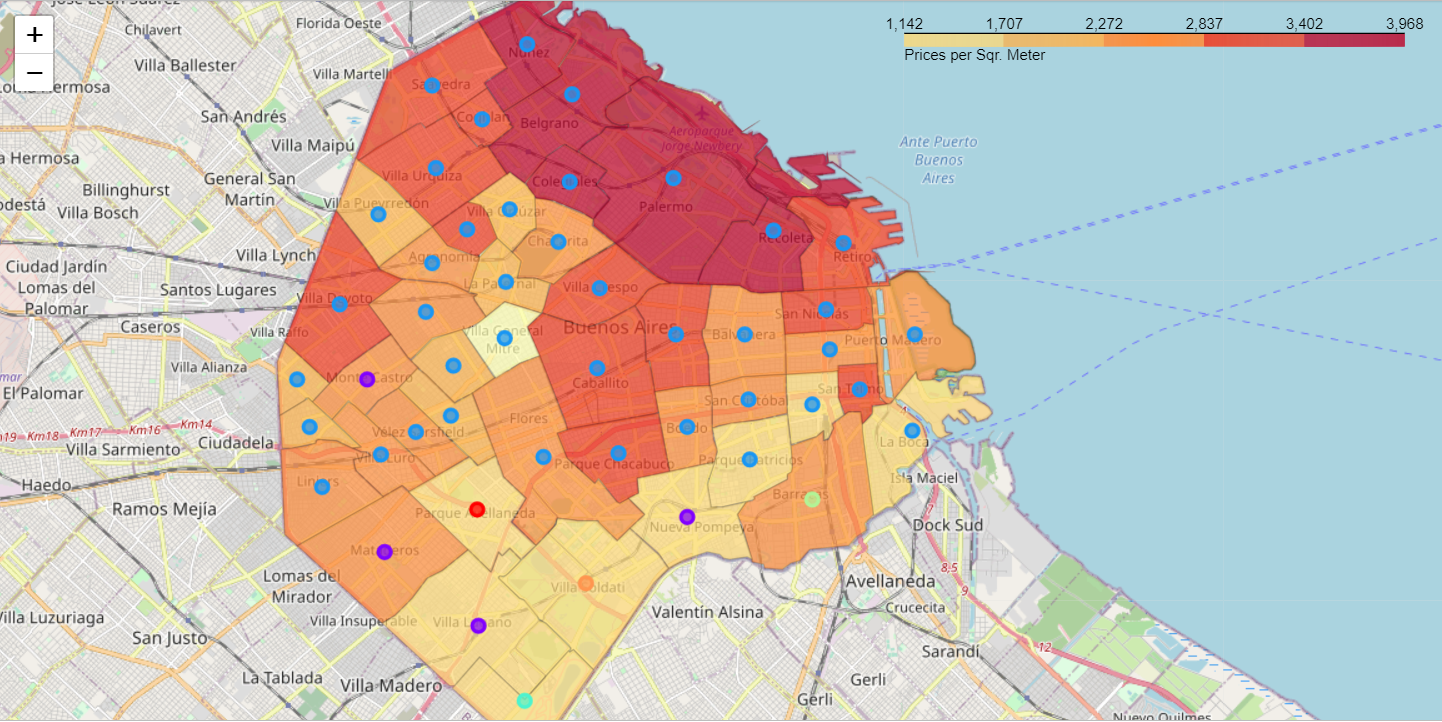
# Results

Once I had all the information merged in one Data Frame, and generated the 6 clusters of districts, I was able to draw the map, with the clusters, and the Choropleth information with the price of per M2 for all the districts.

To do this, I obtained the Geo Data for all the districts in Buenos Aires city, from the National Government data service.

I created choropleth map which also has the below informations for each borough:

* District Name
* Cluster name,
* Average Price per m2



# Discussion

Buenos Aires is a big city, with a lot of social venues, between all of its 48 districts.

Having created 6 clusters using K-means, with an optimal K based on an Elbow method, we can see that most of the districts were in the cluster number 2. Being from Buenos Aires, allows me to see the similarities in the districts that are in this cluster, but also I can see the difference in price ranges within the cluster, allowing me to choose a district with a lower price, but with similar venues.

I chose to perform this analysis at a district level, but the data set can be expanded and the details of the neighborhood or street can also be drilled.

Also, in the future, I would create a list of excluded venues, as Foursquare brings some venues that do not add value to the decision-making process of buying a property.

# Conclusion

With the information at hand, potential buyers should be able to choose the district that will have the most attractive venues nearby, and also invest less money than going to the most typical expensive districts.

# References:

* Neigborhood data from Buenos Aires City. <https://data.buenosaires.gob.ar/dataset/barrios/archivo/1c3d185b-fdc9-474b-b41b-9bd960a3806e>
* Georeferential data from Buenos Aires City Neighborhood <https://datosgobar.github.io/georef-ar-api/download/>
* [Forsquare API](https://developer.foursquare.com/)
* Housing Prices by square meter by neighborhood <https://www.estadisticaciudad.gob.ar/eyc/?p=27739>
* Projection of habitants in Buenos Aires City from 2010 to 2040 <https://www.estadisticaciudad.gob.ar/eyc/?p=29329>