**IBM Data Science Capstone Project**

**Buenos Aires Territorial Pricing Analysis**

**Introduction**

The IBM Data Science Professional certificate course on Coursera concludes with a Capstone Project. This project consists in using data science toolset on a real-life problem and demonstrating the creation of value by applying the learned skills. The analysis was performed in Python.

**Description and Background**

Buenos Aires is the capital of Argentina and is a city with a high population.

For this project, first, we have scraped data from various data about Buenos Aires Territorial price. Secondly, we used Foursquare API to get the common venues of the neighborhoods.

After that, we clustered Buenos Aires Neighborhoods using K-means clustering on the basis of the common venues.

This project is mainly made for territorial investor. As from territorial investor point of view, we want to invest in such places were the territorial prices are low and the facilities (shops, restaurants, parks...etc) and social venues are nearby.

This project is targeted towards:

\* Territorial Investor who wants to know the territorial price in different neighborhoods in Buenos Aires.

\* People interested in buying territory in Buenos Aires for living.

**Data Collection**

For the project we used the following data:

\* An excel file provided by the Buenos Aires government. ('http://cdn.buenosaires.gob.ar/datosabiertos/datasets/terrenos-valor-de-oferta/precio-de-terrenos-2018.xlsx') This excel file contained the list of territorial prices for different neighborhoods in Buenos Aires.

\* Foursquare API to get the most common venues for each of the selected neighborhoods. https://foursquare.com/developers/apps.

\* The geopy.geocoders library to get the coordinates for each neighborhood in order to map it using the folium library, the coordinates were also used for the Foursquare API.

\* We designed the limit as 100 venue and the radius 1400 meters for each neighborhood from their given longitude and latitude information.

**Data Preprocessing**

First of all, the data scraped from the Buenos Aires government website has to be clean.

On one hand, all the hyperlinks numbers were removed and there are more than one Postal codes for some locations so we kept only one Postal Code. On the other hand, all the null values and the columns which contains information not needed for this report were removed. Once cleaning this table, we considered that the column of Average Price need to be create since this is useful for the comparison.

We utilized the Foursquare API to explore the neighborhoods and segment them. The limit as 100 venue and the radius 1400 meters for each neighborhood was designed from their given longitude and latitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API. Finally, by using the Foursquare API in the conjunction with the created datasets, a table of most common visited venues in Buenos Aires neighbors is generated.

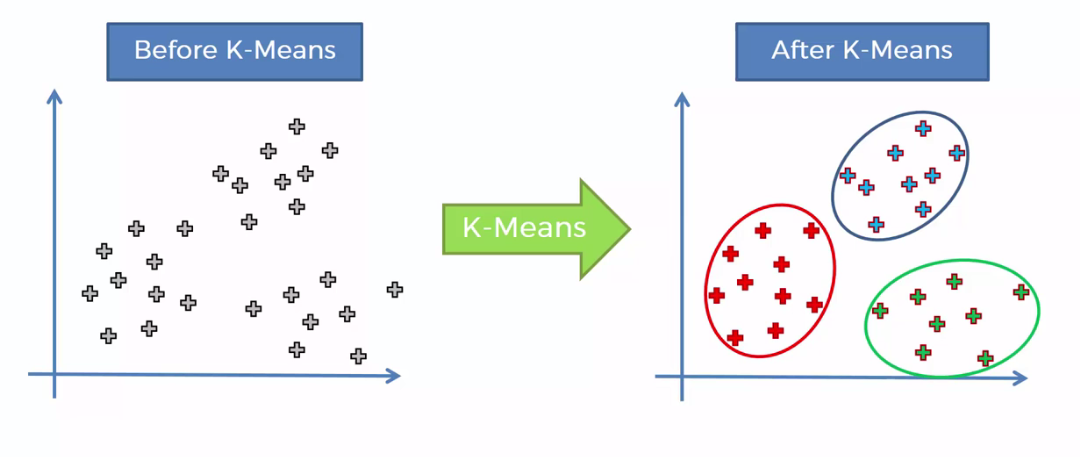
**Machine Learning**

Due to the reason that we have some common venue categories in boroughs, we used unsupervised learning K-means algorithm to cluster the boroughs. K-means algorithm is one of the most common cluster method of unsupervised learning.

**K-Means:**

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the muldimensional features.

In the following figure there is a graphical example of how a K-Means algorithm works. As it is possible to see, dispersion is minimized by representing all clustered data into one group or cluster.

[](https://camo.githubusercontent.com/e4b8f933aa4a222d5cfd55f4ed25046f61f5dc8d/68747470733a2f2f63646e2d696d616765732d312e6d656469756d2e636f6d2f6d61782f313630302f312a745761615a5837356f756d5677424d634b4e2d6548412e706e67)

It is necessary for this algorithm to have a prior idea about the number of clusters since it is considered an input of this algorithm. For this reason, the elbow method is implemented. A chart that compares error vs number of cluster is done and the elbow is selected. Then, further analysis of each cluster is done.

First, I ran K-means to cluster the boroughs into six clusters because when I analyze the K-means with elbow method it ensured me the 5 degrees for optimum k for the K-means. After that, I merged table with cluster labels for each neighbor.

Once examining each cluster, I labeled each cluster as the following:

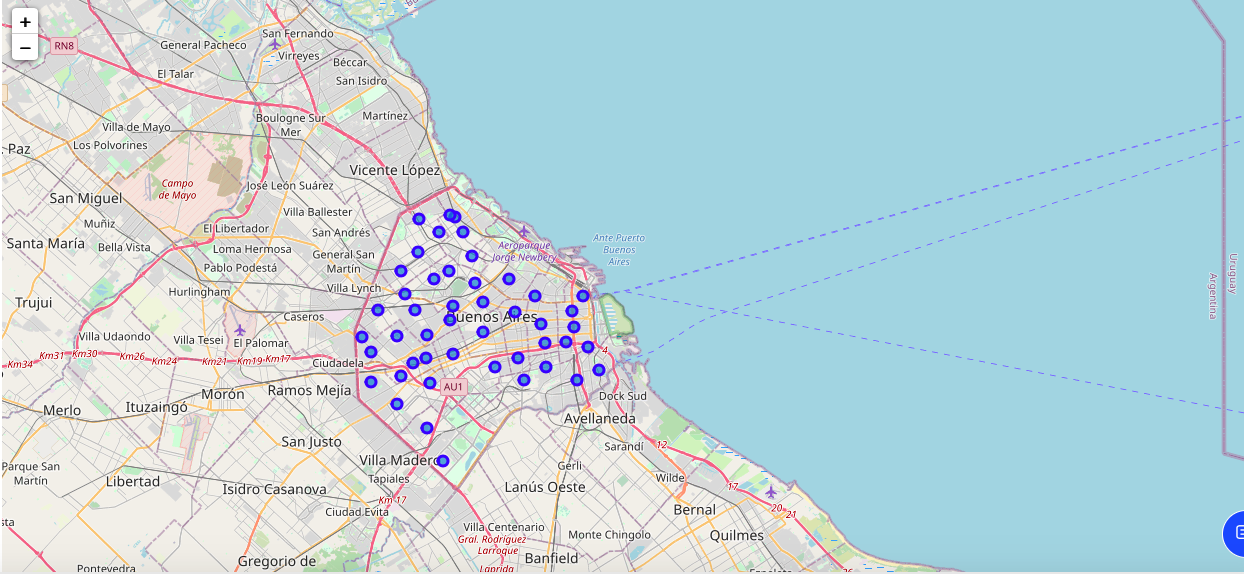
1. Mixed Social Venues
2. Hotels and Social Venues
3. Stores and seafood restaurants
4. Pubs and Historical places
5. Sports and Athletics
6. Restaurants and Bars

After examining Average Price, I labeled each price as follows.

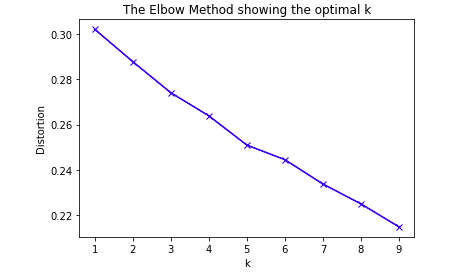
1. <75000: “Low Level 1”
2. 75000 – 125000: “Low Level 2”
3. 125000 – 175000: “Average Level 1”
4. 175000 – 225000: “Average Level 2”
5. 225000 – 275000: “High Level 1”
6. >275000: “High Level 2”

**Result**

Firstly, data is plotted in a geographical map to get a notion of the world location. The following image shows the neighborhoods in Buenos Aires.

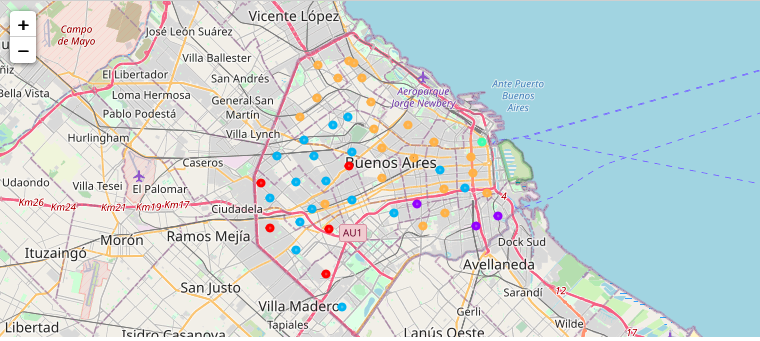


Secondly, the cluster algorithm is implemented. For this purpose, it is necessary to have a prior idea about the number of clusters. Therefore, the mean squared error (MSE) is plotted vs the number of clusters. The number of clusters start with a value of 1 increasing until a value of 9. This chart is shown in the image below.

****

As it is expected, the MSE decreases over the number of clusters. The elbow method here is implemented in order to select the appropriate number of groups. In this case, it is possible to see that the elbow is found more or less around 5. The MSE found below this number shows little changes rather than big ones. Finally, once the number of clusters is fixed, the clustering algorithm is repeated through samples and each neighborhood is labeled according to the clusters found.

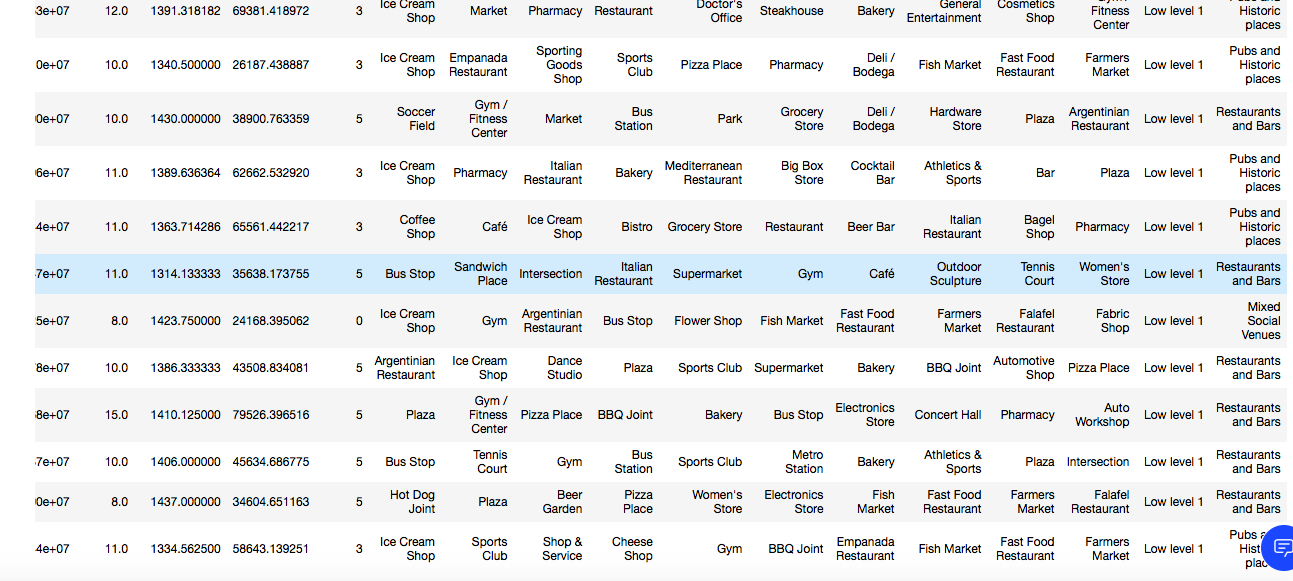
For visualization purposes, the geographical data is again plotted but with different colors. Each color represents the cluster for which that neighborhood belongs. This image is shown below.



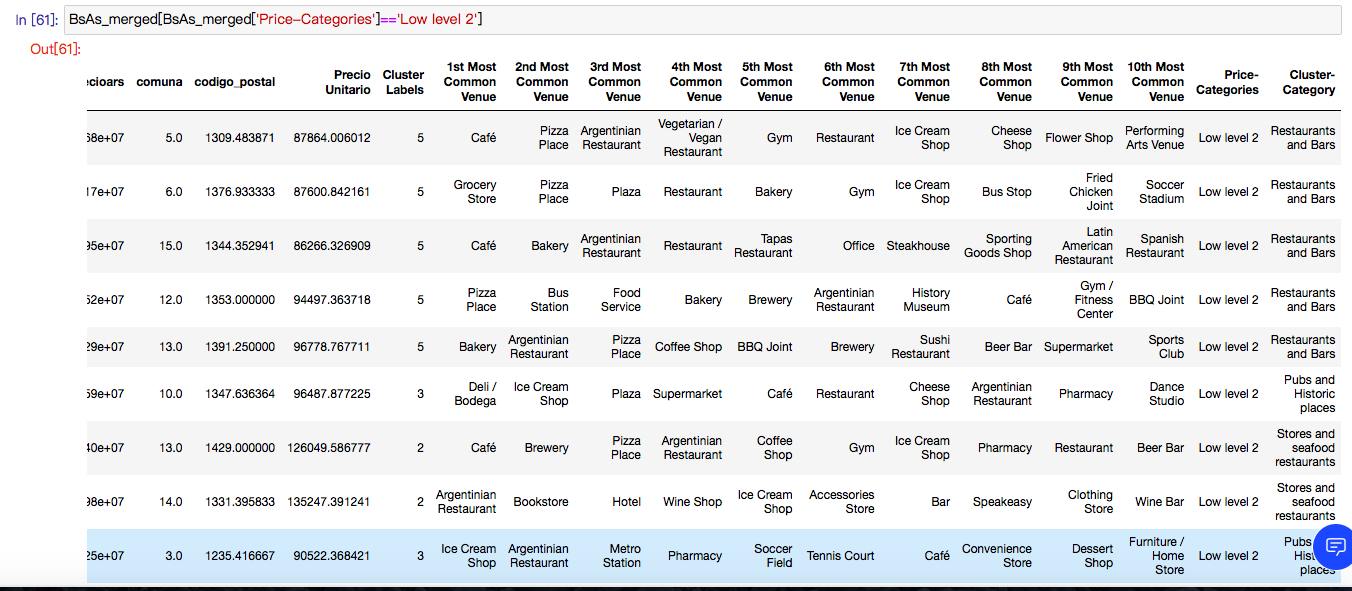
Once clustering the neighborhoods by the categories mentioned above, these neighborhoods are also classified by price label mentioned.

The neighborhoods and the cluster categories for the price label “Low Level 1”:





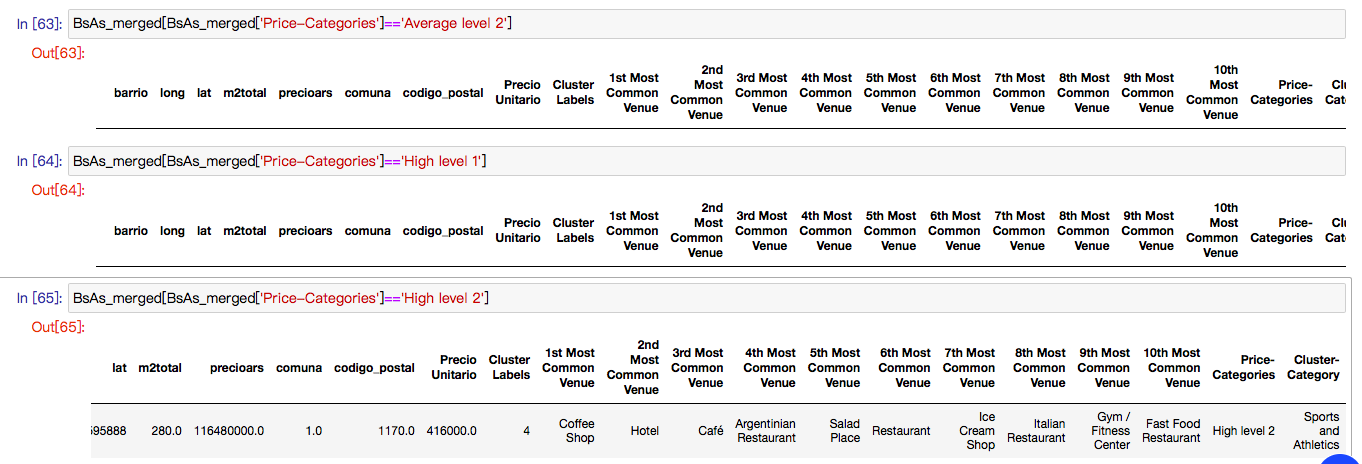
The neighborhoods and the cluster categories for the price label “Low Level 2”:



The neighborhoods and the cluster categories for the price label “Average Level 1”:



As result, there are no neighborhoods belong to the price categories “Average Level 2” and “High Level 1”. There is only one borough named “Rertiro” classified as “High Level 2”.



**Conclusion**

In this project, Buenos Aires is segmented into different clusters according to the facilities nearby, and also classified by the territorial prices. This work is not only useful for the territorial investors but also for the people who are planning to move to Buenos Aires due to their work or business since from their point of view, they would like to live or invest in such places where the territorial prices are low and the facilities (Shops, restaurants, parks, bars…etc) and social venues are nearby.

After reviewing the results, we could observe that most of the neighborhoods are classified as Low Level 1 and Low Level 2. In addition to this, we could also conclude that almost all the low-price neighborhoods are close to restaurants, bars, pubs, etc. The boroughs with higher territorial price are mostly nearby stores, seafood restaurants, and also sport facilities.