Clustering neighbourhoods in Buenos Aires City

Coursera Capstone project

IBM Data Science professional Certificate

By Tomas Leake

Introduction:

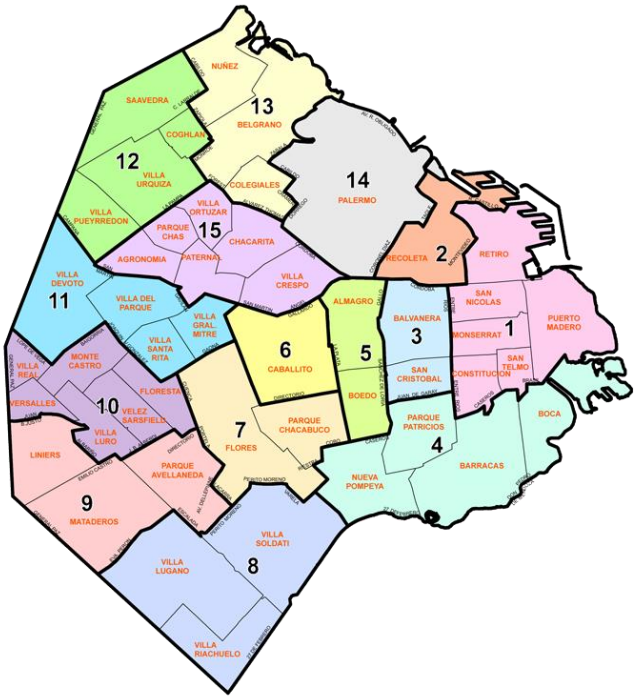


Figure 1 - Barrios and Comunas of Buenos Aires City

Buenos Aires is the capital of Argentina, and has a population of 2.89 million people (according to 2010 national census). Dispite being one of the main concentrations of wealth in Argentina and Latin Amerrica, economic inequality is high. However, to a visiting tourist, it ls likely they will not be exposed to the poorer areas of the city doe to the distribution of the neighbourhoods relative to the main attractions the city haas to offer.

The aim of this study is to perform a brief overview of the neighbourhoods of Buenos Aires, by analyzing publicly available geographic and economic data, and clustering neighbourhoods based on their most numerous type of venues. (note, the scope of this study only includes neighbourhoods within the Autonomous city of Buenos Aires, not the adjacent surrounding cities of the Province of Buenos Aires)

This study is aimed as a starting point for anybody interested in investing or starting business in Buenos Aires, in order to gain a comprehensive overview of social and economic distribution within the city.

For this study i will be using Exploratory Data Analysis (EDA) to uncover hidden properties in the data, with an aim to give the reader a basic overview about the economy, characteristics and distribution of Buenos Aires 48 neighbourhoods.

Before reading, the reader should note i kept the local terminology for Neighbourhoods and Boroughs: "Barrio" and "Comuna" respectively. The different Comunas of the city of Buenos Aires are identified by numbers, from 1 to 15.

Data

The Government of the city of Buenos Aires hosts public data on the following website:

https://data.buenosaires.gob.ar/dataset

In this case the main datasets we are interested are:

- Barrios and Comunas of Buenos Aires city
- Average household income per Comuna
- Geographic outline for each Barrio (GeoJSON)

Also, in order to search for venues within the vicinity of each barrio, its convenient to work with specific location points for each one and explore within a certain radius. For this we shall geocode the latitude and longitude of each Barrio through Nominatim.

The nearby venues will be obtained through the FourSquare API.

Buenos Aires neighbourhoods (barrios) and boroughs (comunas)

First of all, information for Barrios and corresponding Comunas was downloaded from the following link:

http://cdn.buenosaires.gob.ar/datosabiertos/datasets/barrios/barrios.csv

The downloaded table consists of data for 48 Barrios, with the following Features:

- WKT (Polygon coordenates outlinine of each Barrio)
- Barrio
- Comuna
- Surface
- Perimeter

For the purpose of this study, only the Barrio and Comuna columns were kept. As for the geographic data, it was directly downloaded in GeoJSON format to use later when plotting neighbourhoods on the map.

Average household income per Comuna

For convenience, I have downloaded the average Income per capita dataset I worked with in .xlsx format. It can be found at:

https://www.estadisticaciudad.gob.ar/eyc/wp-content/uploads/2018/05/MT_eah_2417.xlsx

This dataset simply consists of each Comuna and its measured Average Household income in Argentine Pesos (AR$).
Note: this is the computed average from 2019. Inflation in Argentina that year was 53.8%, and at the time of writing continues to rise. For the sake of simplicity we shall consider these values as current. It is also unlikely that relative differences in income between boroughs has changed significantly since then, which is more important for the purposes of this study.

Barrio locations

Although the geographic outline of each Barrio was supplied on the barrios.csv file, for the purpose of this study it was necessary to count with a specific point in each neighbourhood, (as representative as possible of its geographical center) in order to xplore venues within a nearby radius.

This was accomplished by iterating through the names of all the Barrios in the dataset and geocoding their location through OpenStreetMap/Nominatim. This allows us to create the following map:

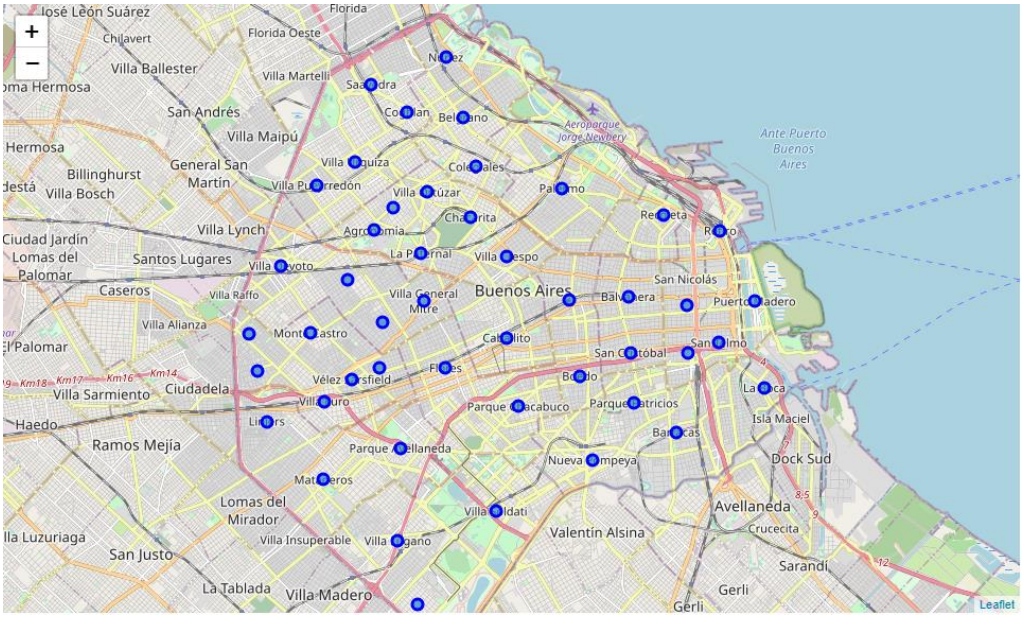


Figure 2 - Map of Barrio Locations within Buenos Aires

Exploring nearby venues with the Foursquare API

Through the FourSquare API, the top 100 venues were explored within a radius of 800m between each neighbourhood. The result was a list of 2068 venues of 228 different categories, and the following information to each one:

Barrio

Venue Latitude and Longitude

Venue category

	Comuna	Barrio	Barrio Latitude	Barrio Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	15	CHACARITA	-34.588011	-58.454156	El Imperio de la Pizza	-34.586890	-58.454967	Pizza Place
1	15	CHACARITA	-34.588011	-58.454156	Sede Central de Whisky	-34.586750	-58.450596	Whisky Bar
2	15	CHACARITA	-34.588011	-58.454156	Fábrica de Churros Olleros	-34.586983	-58.453640	Bakery
3	15	CHACARITA	-34.588011	-58.454156	Santos 4040	-34.588822	-58.449863	Theater
4	15	CHACARITA	-34.588011	-58.454156	Donnet	-34.588333	-58.450745	Vegetarian / Vegan Restaurant

Figure 3 -example of results obtained after exploring nearby venues through FourSquare

In order to make this information more understandable by a ML algorithm, one-hot encoding was used to group the data by Barrio, and count how many venues of each category were present.

The venue count was recomputed to show each venue as a proportion of total venues, in order to sort by most common venues, and keeping the top 10 most common venues for each Barrio. An example of this dataset can be seen in the following figure:

	Barrio	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	CONSTITUCION	Café	Bus Stop	Hotel	Soccer Field	German Restaurant	Pub	Shopping Plaza	Coffee Shop	Russian Restaurant	Cultural Center
1	MONSERRAT	Hotel	Spanish Restaurant	Café	Argentinian Restaurant	Coffee Shop	Hostel	Camera Store	Pizza Place	Italian Restaurant	Sandwich Place
2	PUERTO MADERO	Argentinian Restaurant	Coffee Shop	Hotel Bar	Hotel	Food Truck	Café	Park	Outdoor Sculpture	Gym	Italian Restaurant
3	RETIRO	Hotel	Coffee Shop	Café	Argentinian Restaurant	Italian Restaurant	Plaza	Ice Cream Shop	Sandwich Place	Cocktail Bar	Restaurant
4	SAN NICOLAS	Restaurant	Café	Other Great Outdoors	Comfort Food Restaurant	Supermarket	Park	Coffee Shop	Outdoors & Recreation	Hotel	Argentinian Restaurant

Figure 4 - Example of dataset showing top 10 most common venues in each neighbourhood

Recapitulation

So, now we have a very complete dataset that includes:

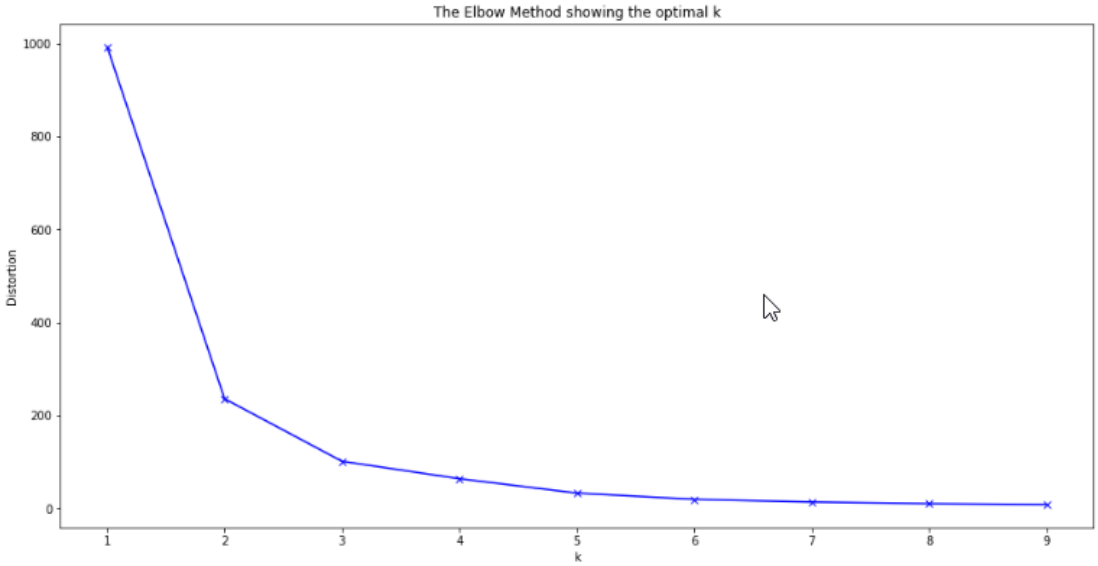
-Barrio/Comuna
-Average household income
-geographic location
-Number of venues in vicinity found on Foursquare
-10 most common venues in vicinity

Methodology

In this section, we cluster the different neighbourhoods based on the venues found through FourSquare, then we visualize and compare it to the data we obtained for average household income and amount of venues found.

K-means clustering

K-Means clustering is used to cluster the different neighbourhoods based on the most common type of venues. We use the Elbow method to study the impact of the number of K on the precision of the model. We cycle through different numbers of clusters and plot their precision (the mean distance between each point and its clusters centroid).



From the graph, we choose 4 as our ideal number of clusters. The model returned a cluster label, a number ranging from 0 to 4, to identify which cluster each Barrio was assigned.

Results and Discussion

Now we visualize, compare and analyse the data we have obtained.

Average Income per Comuna

First of all, we create a choropleth map of Average household income vs Comunas, to get an idea of wealth distribution within the city of Buenos Aires.

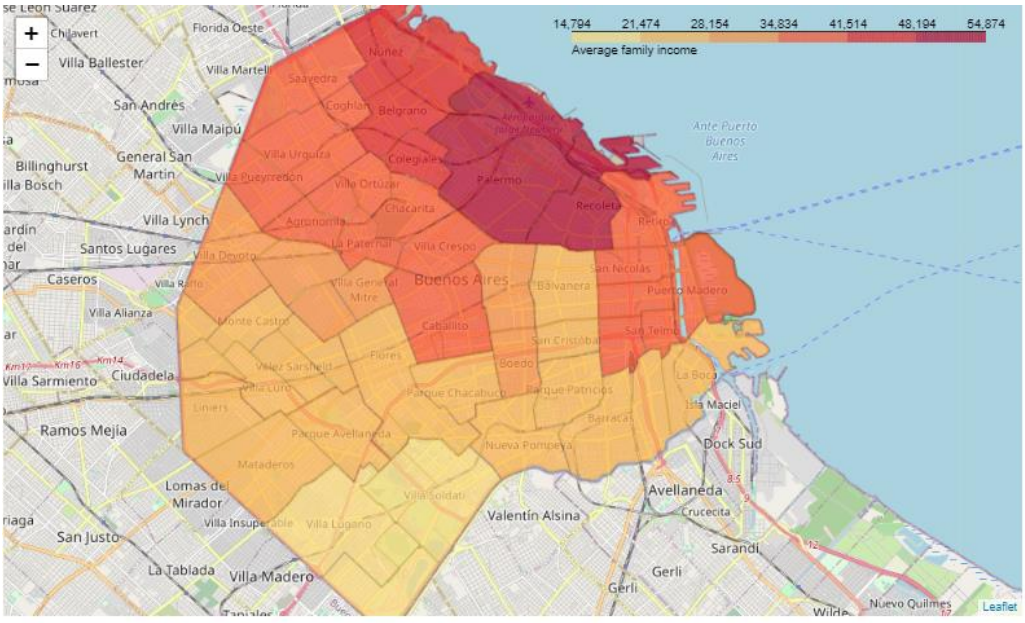


Figure 5 - Average household income per Comuna in Buenos Aires City

We can clearly see from this map the most affluent Barrios are concentrated in the Northeastern side of the city, with the average income decreasing towards the south. The Richest Comunas have an income roughly 3.7 times higher than the poorest ones.

Venues per neighbourhood

We visualize the ammount of venues found by foursquare for each neighbourhood

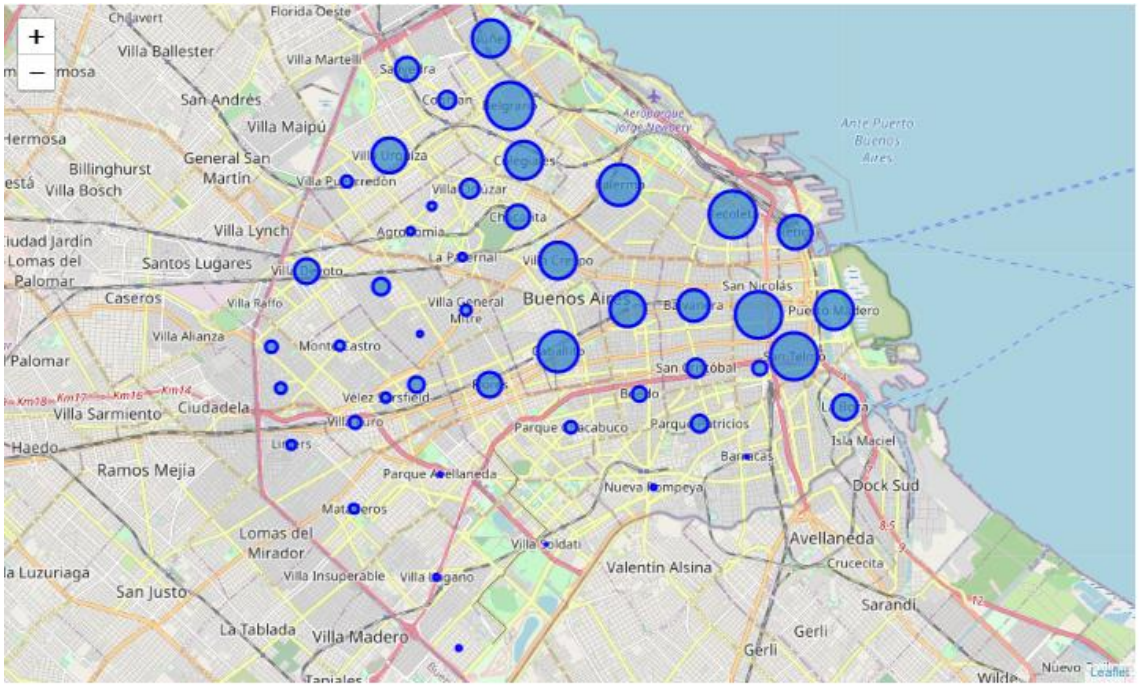


Figure 6 - Visualization of ammount of venues returned for each Barrio within an 800m radius by FourSquare API

Here we can see that more venues were found by FourSquare for the richer neighbourhoods up North than for the poorer neighbourhoods down South. This doesn't necessarily mean there are less venues in the poorer neighbourhoods, but it is more likely that a venue within a richer neighbourhood will show up on FourSquares database.

Barrio Clusters

Now we colour code the different Barrios in the city depending on which clusters they were grouped in by the K-means algorithm.

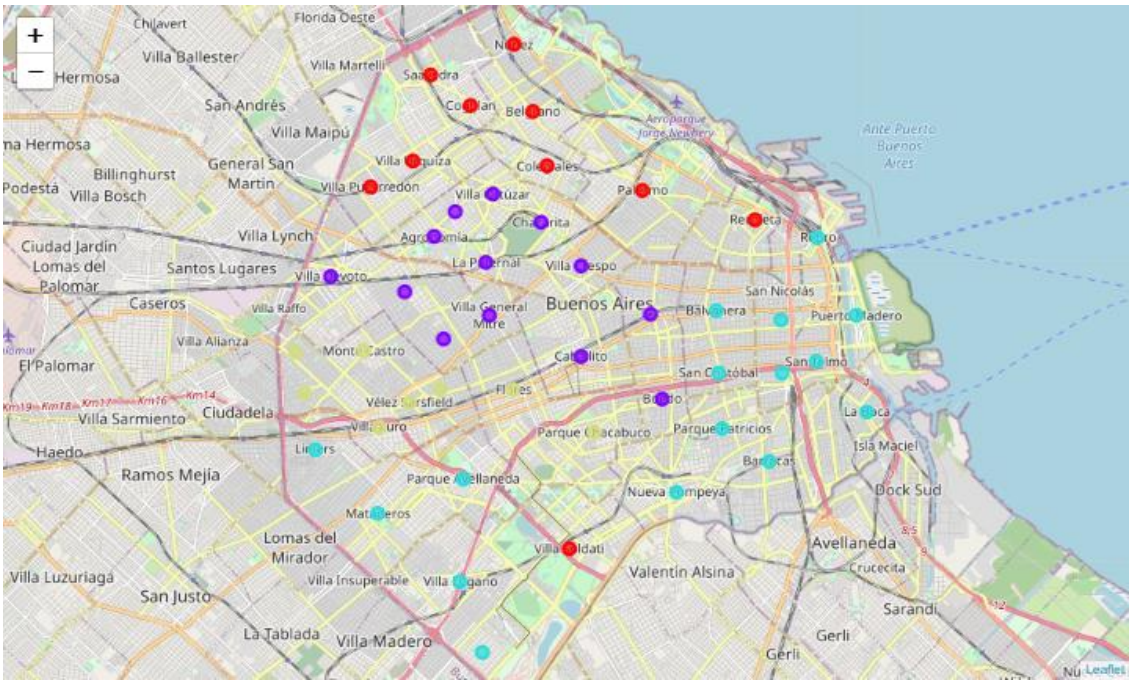


Figure 7 - Clusters of Barrios identified by teh K-means algorythm

Discussion

At first glance, the clusters 0, 1 and 3 (red and purple and green(ish) roughly correspond to different income levels observed in the first map. Cluster 0 does have one outlier, suprisingly in one of the lowest income neighbourhoods of the city, though from the previous map we can see that Villa Soldati only has 4 Venues.

The eastern coast of the city, in an area roughly between the barrios Retiro and Constitucion houses many commercial, government and financial office buildings, and is the destination for many commuters from both the city and province of Buenos Aires, so it makes sense it is largely in a separate cluster (2 - blue). Again, the fact that the 5 neighbourhoods in the poorer southwest of the city have been grouped into the same cluster is probably due to the very low number of venues registered on FourSquare in that area.

Cluster analysis

Cluster 0 is located un north and includes the highest income residential neighbourhoods of the city. Figuring strongly in the top 10 most frequent venues are bakeries, coffee shops and ice cream parlours. Also, Deli/wine shops only figure in neighbourhoods in this cluster. Villa Soldati, the sole outlier in this cluster, only has 4 venues available on FoursQuare in its vicinity, and most likely ended up in this cluster doeto imprecise data.

Cluster 1 groups the relative "middle income" residential neighbourhoods just south of cluster 0. Pharmacies figure frequently in the top 10 venues here, more than the other clusters, along with fewer gyms.

Cluster 2 contains the Microcentro, Buenos Aires busy commercial/political/financial centre. it also contains many of the most popular landmarks and tourist attractions. In this this analysis It stands out mostly due to the strong presence of Parks, hotels, and grocery stores/supermarkts.

It also includes several poorer neighbourhoods in the southwest of the city, but they may have been clustered together due to their low number of venues.

Cluster 3 seems to group a wedge of lower income neighbourhoods that runs from the western side of the city into the centre. Notable observations one can make are the number of Sporting clubs (football/soccer clubs), number of restaurants, particularly Argentinian/bbq places, a high number of Bus stops and no hotels. Also, there are a lower number of cafes and coffee shops compared to the other clusters.

Conclusion

In this study, Geographic and economic data for neighbourhoods and venues in Buenos Aires city was obtained from the official data portal of the city, Nominatim and FourSquare API, in order to cluster similar neighbourhoods.

There is considerable economic contrast between the northern and southern neighbourhoods of Buenos Aires, seen clearly when mapping the average income per borough (comuna). In the case of this study this was reflected in the type of most numerous venues: Ice cream parlours and coffee shops seem to be more common in affluent neighbourhoods, while Argentine/BBQ restaurants and sports clubs seem to be more numerous in lower income areas.

A disparity in available data on venues was observed between the higher income neighbourhoods and the lower income ones, so FourSquare is probably not the best source of data when taking into account the lower income areas of Argentina.