The Battle of Neighborhoods

Applied data science capstone

C. Urrego

July 2020

Introduction

Bogotá is the Colombia's capital, a country in South America. It has approximately 7'743.955 inhabitants, it is the city with the major population in the country. As a capital city, the government has its main headquarters there and also concentrates industry and the economy in general.

The city is divided into boroughs and these, in turn, into neighborhoods distributed throughout its urban and rural areas. The large inhabitants amount, the increase in the migrants arrival, both from other areas of the country and from other countries and the continuous city growth has made the number of people in some localities increase more than in others, but not in the same way the spaces they have for recreation and social gathering.

The intention with this project is to identify the similarity among the areas of the city using the venues they have and setting the relationship that exists between it and the inhabitants number in each boroug and thus identify in which areas and what types of places can be created to encourage the balanced city growth .

This could be used as a guide for people who want to start entertainment sites and determine which areas could be good options to establish their business.

Data Description:

BeautifulSoup Library is used to obtain two datasets from Wikipedia web pages:

- Dataset 1 : https://es.wikipedia.org/wiki/Anexo:Localidades_de_Bogot%C3%A1

  Will use Boroughs information like id, name, surface, population

- Dataset 2 : https://es.wikipedia.org/wiki/Anexo:Barrios_de_Bogot%C3%A1

  Will use Neighborhoods information like id, name, borough, subdivisions

Also is used Geocoder library to get each neighborhood geographic information (latitude, longitude).

Foursquare API to get venues information for each neighborhood in the city.

I'm using the venues information by neighborhood to get the number of venues per district and analyze the relationship to the value of the population and identify how similar each district is to the others.

Data cleaning:

For both datasets extracted from Wikipedia pages, it was necessary to delete some columns that were not going to be used, blank spaces were removed from the numeric fields to be able to convert them into numerical data, spaces and line breaks were also removed at the beginning and at the end of the content of each cell.

	Id_Borough	Borough	PostalCode	Surface	population	Density
0	0	0	0	0	0	0
1	01	Usaquén	110111-110151	65.31	501 999	7 686.4
2	02	Chapinero	110211-110231	38.15	139 701	3 661.88
3	03	Santa Fe	110311-110321	45.17	110 048	2 436.3
4	04	San Cristóbal	110411-110441	49.09	404 697	8 243.98
5	05	Usme	110511-110571	215.06	457 302	2 126.39
6	06	Tunjuelito	110611-110621	9.91	199 430	20 124.11

Table 1Borough data

From the second data set (neighborhoods) was extracted from the borough column the id in a new column to allow the union with other data sets more easily. Also, from the use of the geocoder library, the latitude and longitude columns were created with the geographic coordinates of each of the neighborhoods.

	UPZ_Id	UPZ_Name	Borough	Neighborhood	Borough_Id
1	1	Paseo de los Libertadores	Usaquén	Canaima, La Floresta de La Sabana yTorca.	01
2	9	Verbenal	Usaquén	Altos de Serrezuela, Balcones de Vista Hermosa...	01
3	10	La Uribe	Usaquén	Bosque de San Antonio, Conjunto Camino del Pal...	01
4	11	San Cristóbal Norte	Usaquén	Ainsuca, Altablanca, Barrancas, California, Ce...	01
5	12	Toberín	Usaquén	El Toberín, Babilonia, Darandelos, Estrella de...	01

Table 2 Neighborhood Data

	UPZ_Id	UPZ_Name	Borough	Neighborhood	Borough_Id	Latitude	Longitude
1	1	Paseo de los Libertadores	Usaquén	Canaima, La Floresta de La Sabana yTorca.	01	4.72773	-74.06471
2	9	Verbenal	Usaquén	Altos de Serrezuela, Balcones de Vista Hermosa...	01	4.75180	-74.04315
3	10	La Uribe	Usaquén	Bosque de San Antonio, Conjunto Camino del Pal...	01	4.73953	-74.02239
4	11	San Cristóbal Norte	Usaquén	Ainsuca, Altablanca, Barrancas, California, Ce...	01	4.74315	-74.04198

Table 3Neighborhood Data with Lat/Lng

Data analysis

Using folium on python all neighborhood can be visualized on a Bogota map

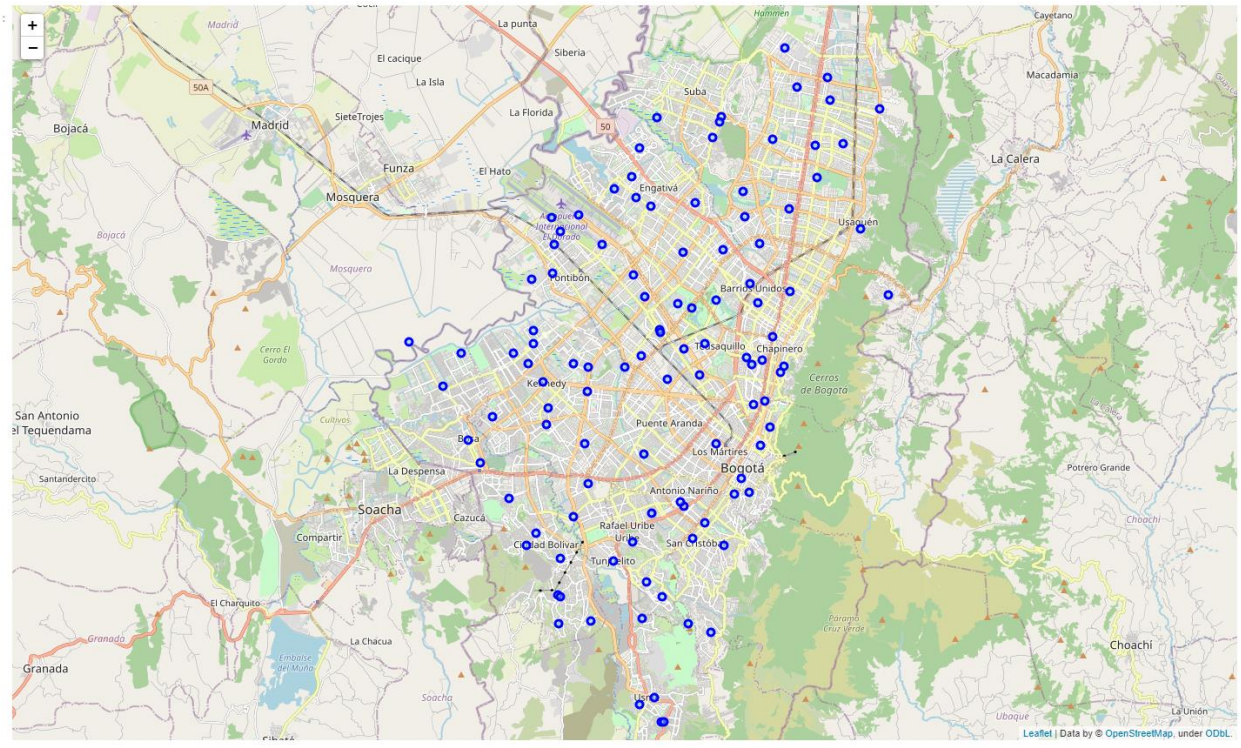

Figure 1 Bogotá Neighborhoods map

From boroughs data were generated two graphics to represent the population and surface in each one.

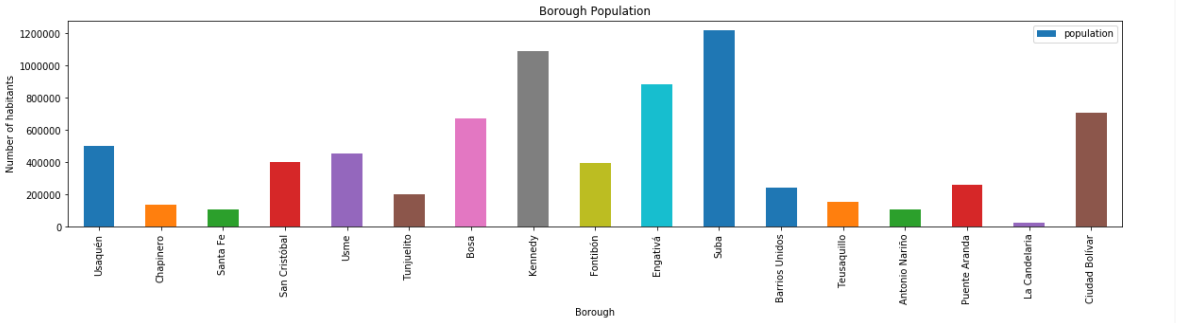

Figure 2 Borough Population

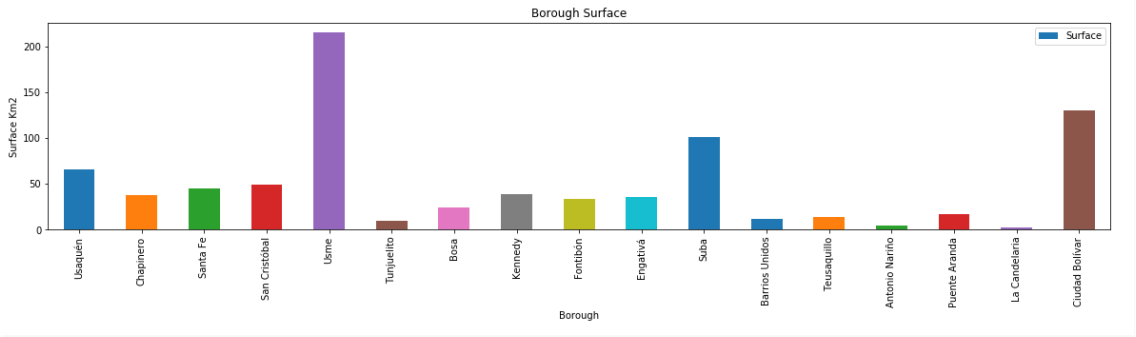

Figure 3 Borough Surface

Comparing each locality it can be seen that not all of them have the same relationship between the inhabitants number and the surface, it is found that for example Kennedy borough having similar surface to other localities has more than double the population, in both graphs the values are very irregular, the difference between the data becomes quite wide.

Foursquare API is used to get Bogotá venues, whit the geographic data of each neighborhood can be obtained the venue name, location and category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Paseo de los Libertadores	4.72773	-74.06471	Juan Valdez Café	4.728148	-74.064754	Café
1	Paseo de los Libertadores	4.72773	-74.06471	WRIBS	4.726935	-74.064903	Sports Bar
2	Paseo de los Libertadores	4.72773	-74.06471	Hondashi	4.727923	-74.064987	Asian Restaurant
3	Paseo de los Libertadores	4.72773	-74.06471	Sagal La Colina	4.726559	-74.061541	Steakhouse
4	Paseo de los Libertadores	4.72773	-74.06471	El Corral Colina Campestre	4.726972	-74.064814	Burger Joint

Table 4 Venues Data

Grouping the information by borough, the venues amount is obtained, it can be seen that the difference between the one with the most (Usaquén) and the least (Tunjuelito) is quite large, and the majority of venues are concentrated in four localities, Usaquén, Chapinero, Suba and Teusaquillo.

	Borough_Id	Borough	Venue
0	01	Usaquén	297
1	02	Chapinero	279
2	03	Santa Fe	121
3	04	San Cristóbal	95
4	05	Usme	83
5	06	Tunjuelito	10
6	07	Bosa	74
7	08	Kennedy	94
8	09	Fontibón	190
9	10	Engativá	112
10	11	Suba	295
11	12	Barrios Unidos	122
12	13	Teusaquillo	221
13	14	Mártires	52
14	15	Antonio Nariño	43
15	16	Puente Aranda	89
16	17	La Candelaria	54
17	18	Rafael Uribe	19
18	19	Ciudad Bolívar	23

Table 5 Venues per Borough

If you compare the graph of the population with the Venues per Borough, you can identify that the proportion is similar but there are three cases that do not behave the same. Usaquén, Chapinero and Teusaquillo have a large number of venues even when their population is not so high, this is attributed to the fact that in these areas are the large financial centers, offices and universities of the city and have a high flow of people per day.

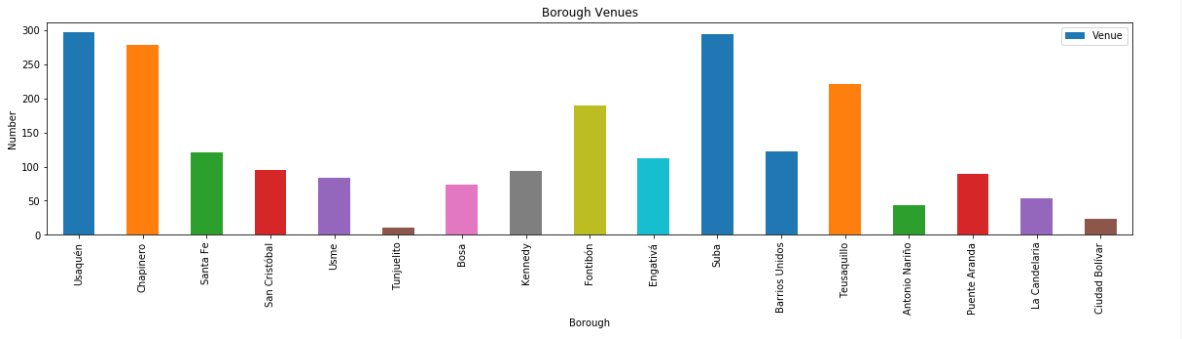

Figure 4 Borough Venues

When graphing the number of venues per neighborhood, it can be seen that there is an even more marked difference because the number is reduced to almost half after the first twenty neighborhoods, this is only a fifth of the total

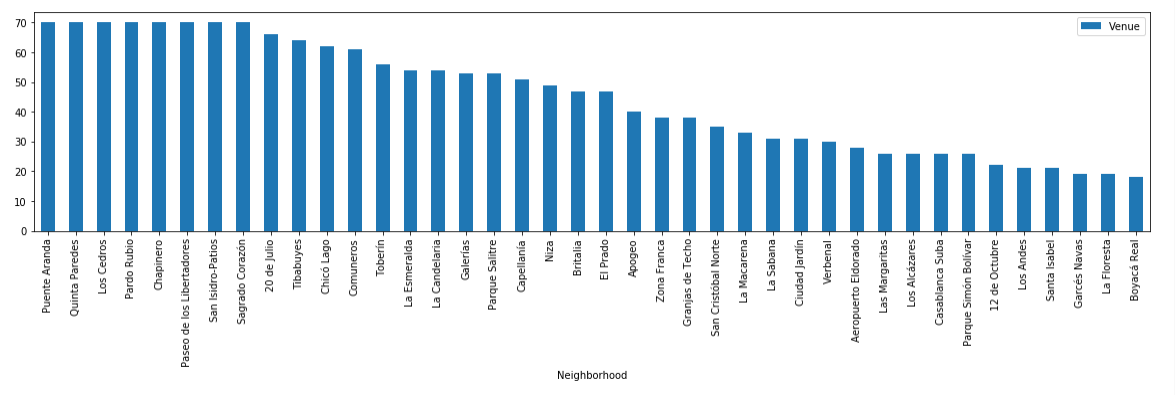

Figure 5 Neighborhood Venues

On the other hand, if we analyze the venues categories, there is a great trend or interest in everything related to food and drinks, different types of restaurants, bakeries, cafes, pubs and bars. An interest in sports venues such as parks and gyms is also identified.

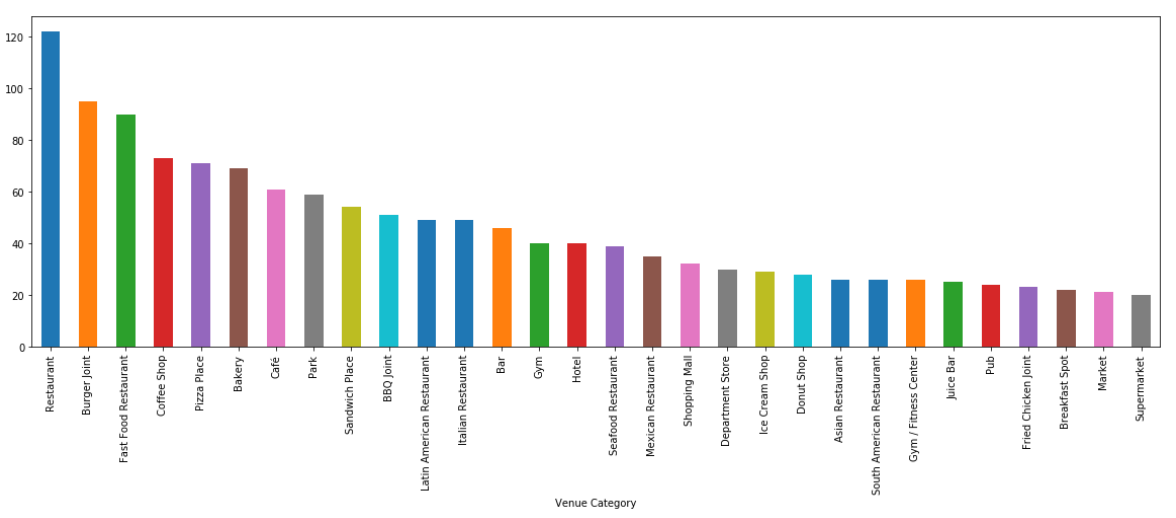

Figure 6 venue categories

Modelling

To analyze how similar the distribution of venue types in the city is, a clustering method is used. With k-means the areas are divided according to the most common venue types.

One hot encoding is done on the area data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the area and the mean of the venues is calculated, finally the 10 common venues are selected for each area. This process is made with boroughs and neighborhoods.

Results

	Cluster Labels	Borough	1st Most Common Venue	2nd Most Common Venue	3
0	3	Antonio Nariño	Restaurant	Pizza Place	
1	1	Barrios Unidos	Fast Food Restaurant	Restaurant	
2	1	Bosa	Coffee Shop	Fast Food Restaurant	
3	1	Chapinero	Restaurant	Bar	
4	4	Ciudad Bolívar	Park	Department Store	
5	1	Engativá	Fast Food Restaurant	Bowling Alley	
6	1	Fontibón	Airport Lounge	Coffee Shop	
7	1	Kennedy	Park	Pizza Place	
8	1	La Candelaria	Nightclub	Restaurant	
9	3	Mártires	Shopping Mall	Restaurant	
10	1	Puente Aranda	Café	Restaurant	
11	0	Rafael Uribe	Supermarket	Coffee Shop	
12	1	San Cristóbal	Fast Food Restaurant	Restaurant	
13	1	Santa Fe	Restaurant	Sandwich Place	
14	1	Suba	Fast Food Restaurant	Burger Joint	
15	1	Teusaquillo	Restaurant	Hotel	
16	2	Tunjuelito	Park	Movie Theater	
17	1	Usaquén	Burger Joint	Pizza Place	
18	1	Usme	Café	Bakery	

Table 6 borough clusters

	UPZ_Name	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
9	Santa Bárbara	4.69151	-74.14861	0.0	Moving Target	Rental Service	Park
19	Lourdes	4.56788	-74.08392	0.0	Construction & Landscaping	Park	Big Box Store
50	Bavaria	4.68637	-74.15100	0.0	Park	Furniture / Home Store	Grocery Store
94	Ciudad Montes	4.57906	-74.14325	0.0	Latin American Restaurant	Women's Store	Fish Market
106	Monteblanco	4.56784	-74.16172	0.0	Construction & Landscaping	Music Store	Park

Table 7 Neigborhoods clusters / cluster 1

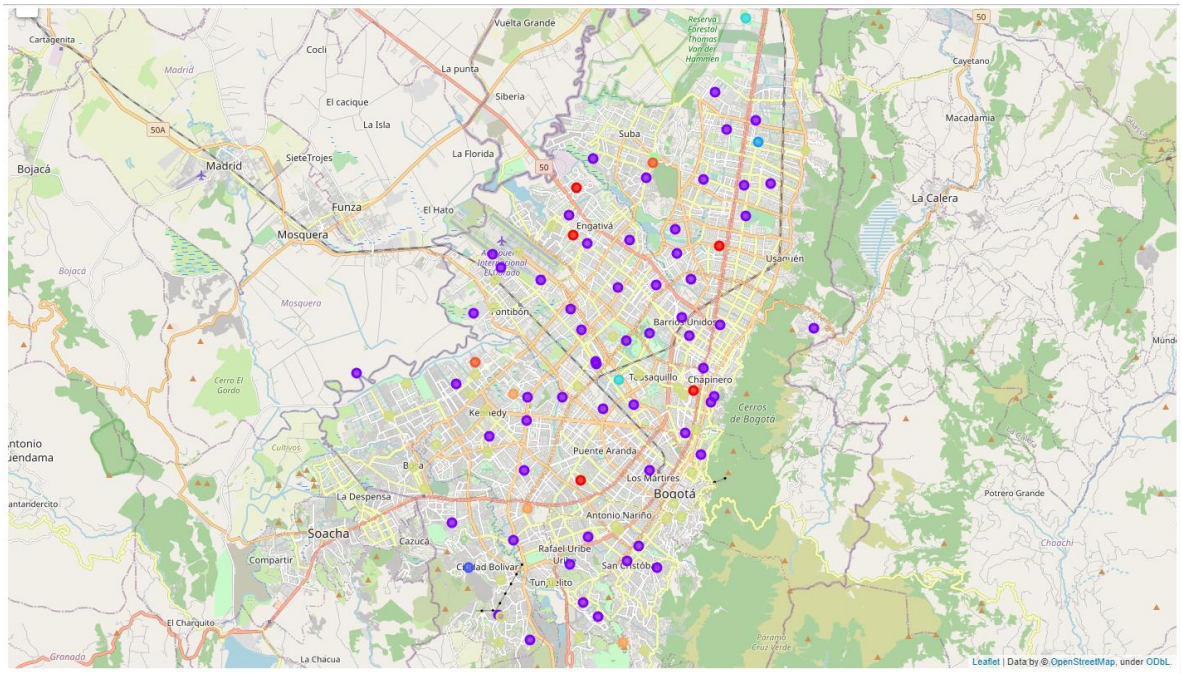

Figure 7 Neigborhood Clusters map

As a result, it is found that most areas fit the same cluster, it is visible in the data obtained that few neighborhoods are different. These results allow to better identify the starting point for the project objective.

small clusters are distributed by different areas without representing any specific pattern, it would be necessary a deeper study of each of these areas to determine if there is any factor that contributes to this division.

The most common categories in each borough and neighborhood are mostly related to food, we can find different kinds of restaurants and pubs.

Conclusion

From the initial data and the data obtained from the analysis and the application of the clustering model, it can be said that the distribution of the venue types is not related to the inhabitants amount, borough as Kennedy with a fairly high population does not have the same proportion from venues. On the other hand, if you take the boroughs that were left alone in a cluster, you can see that they are characterized by being small and having a very low venues quantity.

The neighborhoods classified in small clusters can be a good recommendation for people who want to start some type of place that allows balancing the characteristics of that area. A great interest in restaurants and bars is evident, which becomes an important fact to select a category when starting a business.