Applied Data Science Capstone Project - The Battle of Neighborhoods

Carlos Llano

January of 2021

Table of contents

Introduction/Business Problem

Bogotá is the metropolitan center of Colombia, which is located at 2,600 meters above sea level. This city offers varied artistic expressions such as museum exhibits, dance, theater, music, and splendid cuisine (1). There are also hundreds of places to go to eat and frequent food festivals (2), and consequently, the catering industry could be a very competitive field in this city.

In this project, the data available on restaurants and territorial distribution of Bogotá will be used to analyze and cluster different locations in which a restaurant could be potentially be established. It is also expected that the information analyzed can be useful to obtain a preliminary insight into the most appropriate types of restaurants and the users that are most likely to visit them. The results obtained in this study could be of interest to businessmen or investors that want to open a restaurant in Bogotá, Colombia.

Data acquisition

The necessary data to perform the analysis of the best locations for a restaurant in Bogotá, Colombia will be obtained from the following sources.

- Lists of the boroughs and neighborhoods and their geospatial position will be obtained from www.bogota-laburbano.opendatasoft.com.
- Data about the territorial distribution of Bogotá will also be acquired from www.bogota-laburbano.opendatasoft.com. This web page contains information about tourist, urban and company areas in Bogotá.
- The types and location of each restaurant in all the neighborhoods will be obtained from Foursquare API (www.foursquare.com).

Methodology

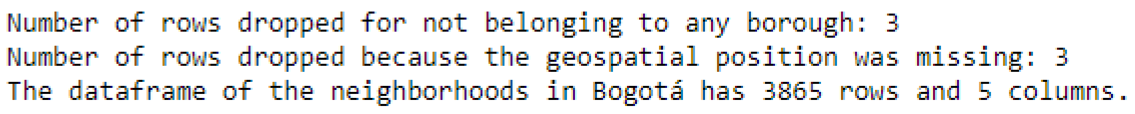
The methodology used in this project to determine the best locations for a restaurant in Bogotá, Colombia, is divided into four main parts:

1. All the data needed about the neighborhoods in Bogotá, Colombia was first searched and collected. Taking as base the data available the next steps were carried out.
2. The data about the boroughs and neighborhoods were cleaned and then analyzed in order to identify the boroughs that could have the most potential customers for the restaurant. Information about tourist places, offices and companies were used to examine the boroughs.
3. Data about the most popular venues in each neighborhood were retrieved from the Foursquare API. Since this project is only focused on restaurants, the venues belonging to other categories were discarded.
4. The neighborhoods of the selected boroughs were clustered using the K-means model. The most common restaurant categories were used as based to perform the neighborhood segmentation.

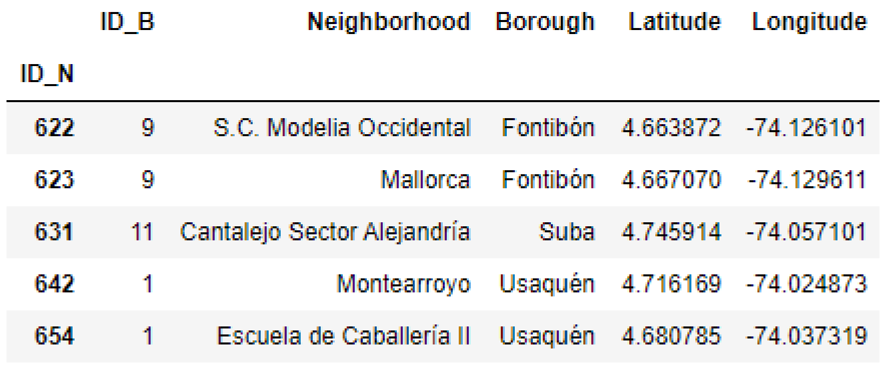
Analysis

Neighbors and boroughs

The data of the neighborhoods and the boroughs with the geospatial position were imported. All the information obtained in dictionary format was organized in a data frame. The rows without information about geospatial position were dropped. Three neighborhoods were also dropped because they were not assigned to any borough. In addition, some neighborhoods do not have any borough names specified. After examining all these rows, all of them corresponded to the same borough ID. Therefore, all of them were filled with the missing Borough name (Suba). The ID of two boroughs (Candelaria and Santa Fe) was corrected since both had the same value in this column. the data frame obtained for the neighborhoods is shown in the Table 1, while in the location of all them in the map of Bogotá is shown in Fig 1.



```
Number of rows dropped for not belonging to any borough: 3
Number of rows dropped because the geospatial position was missing: 3
The dataframe of the neighborhoods in Bogotá has 3865 rows and 5 columns.
```



	ID_B	Neighborhood	Borough	Latitude	Longitude
ID_N					
622	9	S.C. Modelia Occidental	Fontibón	4.663872	-74.126101
623	9	Mallorca	Fontibón	4.667070	-74.129611
631	11	Cantalejo Sector Alejandría	Suba	4.745914	-74.057101
642	1	Montearroyo	Usaquén	4.716169	-74.024873
654	1	Escuela de Caballería II	Usaquén	4.680785	-74.037319

Table 1. Data frame obtained for the neighborhood and boroughs in Bogotá, Colombia (3).

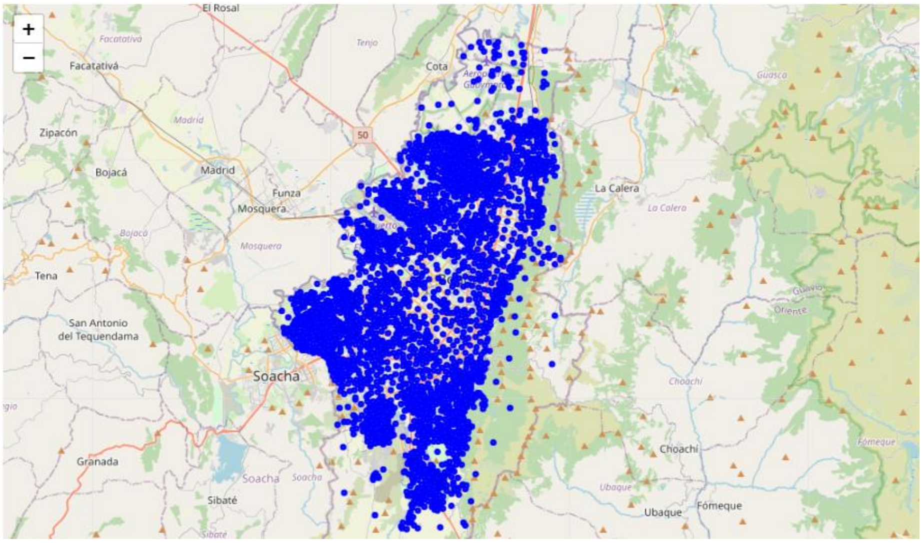
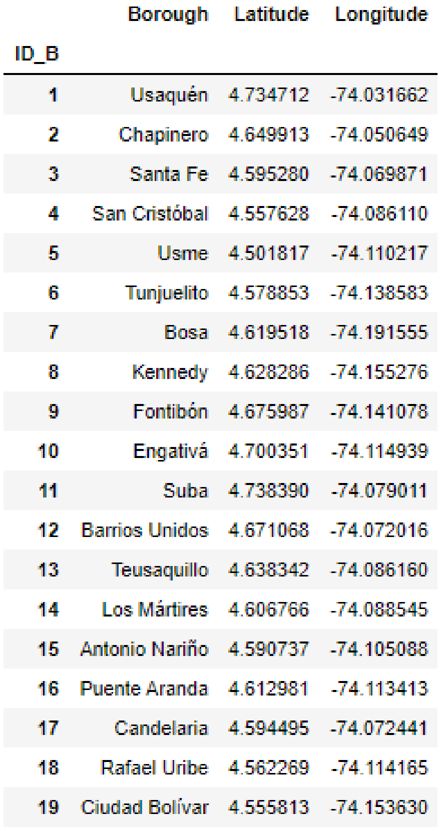


Fig 1. Map of Neighborhoods of Bogotá. The Neighborhoods are represented as blue dots.

Since the amount of neighborhoods obtained is very high (3865). The boroughs are going to be explored first in order to only consider the most suitable. Bogotá is composed of 20 boroughs. The 20th borough (Sumapaz) is mostly rural and is less populated (4). Therefore, this borough was not had into account in the database initially. The boroughs of Bogota are shown in the Table 2.



	Borough	Latitude	Longitude
ID_B			
1	Usaquén	4.734712	-74.031662
2	Chapinero	4.649913	-74.050649
3	Santa Fe	4.595280	-74.069871
4	San Cristóbal	4.557628	-74.086110
5	Usme	4.501817	-74.110217
6	Tunjuelito	4.578853	-74.138583
7	Bosa	4.619518	-74.191555
8	Kennedy	4.628286	-74.155276
9	Fontibón	4.675987	-74.141078
10	Engativá	4.700351	-74.114939
11	Suba	4.738390	-74.079011
12	Barrios Unidos	4.671068	-74.072016
13	Teusaquillo	4.638342	-74.086160
14	Los Mártires	4.606766	-74.088545
15	Antonio Nariño	4.590737	-74.105088
16	Puente Aranda	4.612981	-74.113413
17	Candelaria	4.594495	-74.072441
18	Rafael Uribe	4.562269	-74.114165
19	Ciudad Bolívar	4.555813	-74.153630

Table 2. Data frame obtained for the Boroughs in Bogotá, Colombia (3).

In order to explore the distribution and dynamics of all the boroughs, data about tourist places, office areas, and the number of companies in Bogotá were employed.

Tourist places

A data frame containing information of all the tourist places in Bogotá and their geospatial location was employed. The amount of tourist places in each borough is shown in Fig 2. The borough with more tourist places is Candelaria, followed by Santa Fe and Chapinero. Restaurants located in these places are interesting since could be very attractive for tourists.

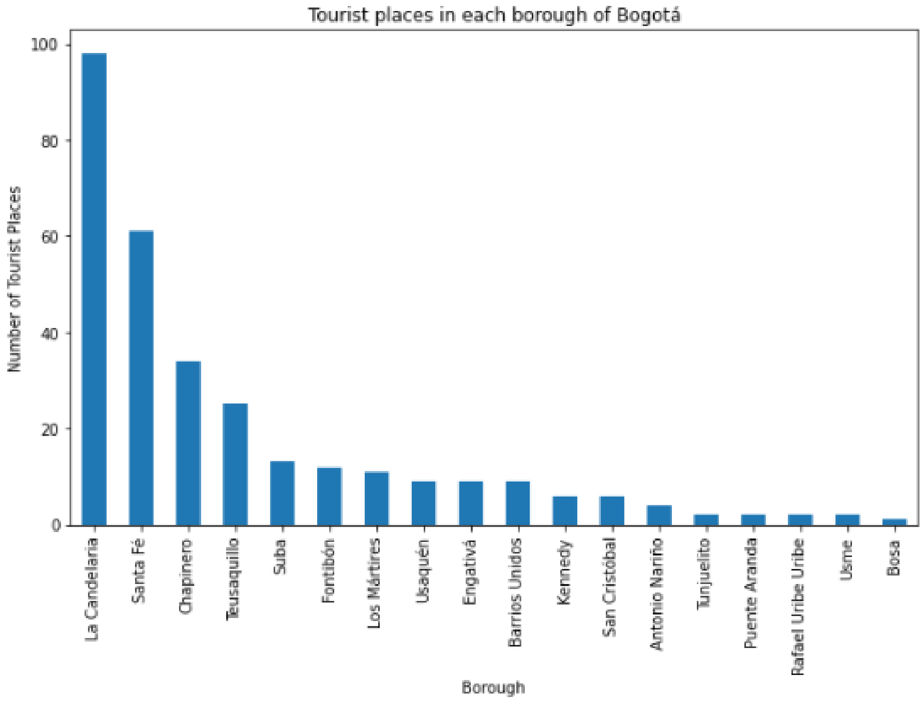


Fig 2. Number of Touristic places in each borough of Bogotá (3).

The map in the Fig 3 shows the locations of the boroughs and the places with tourist attractions. Most of the tourist places are located in the east of the city, where the borough La Candelaria, Santa Fé and Chapinero are located.

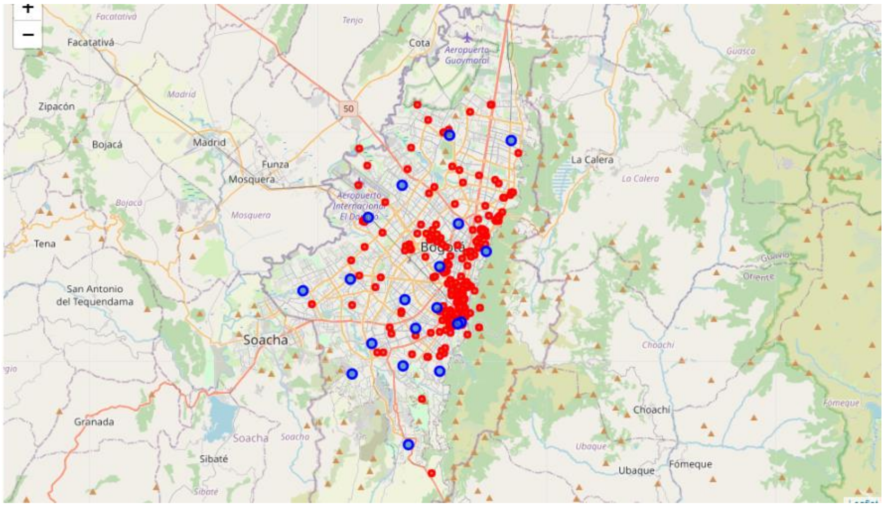


Fig 3. Map of Boroughs and tourist places in Bogotá of Bogotá. The Boroughs are represented as blue dots and the tourist places as red dots.

Offices and companies

Data with information about offices and companies in Bogotá were used to explore each borough. The two dataframes contained data of areas destined to offices (between 2001 and 2015) and companies subscribed to the chamber of Trade and Industry in 2015. This information could be useful to find administrative and corporate zones, where many people work and need a restaurant close for meals.

The areas destined to offices between the years 2001 and 2015 were summed to get an approximation of the total areas that are currently used for offices. These results presented in Fig 4 shows that in the boroughs Chapinero and Usaquén are higher amounts of areas used for offices.

The number of companies by size in each borough is shown in Fig 5. The boroughs Chapinero and Usaquén have again the higher number of Large and small companies.

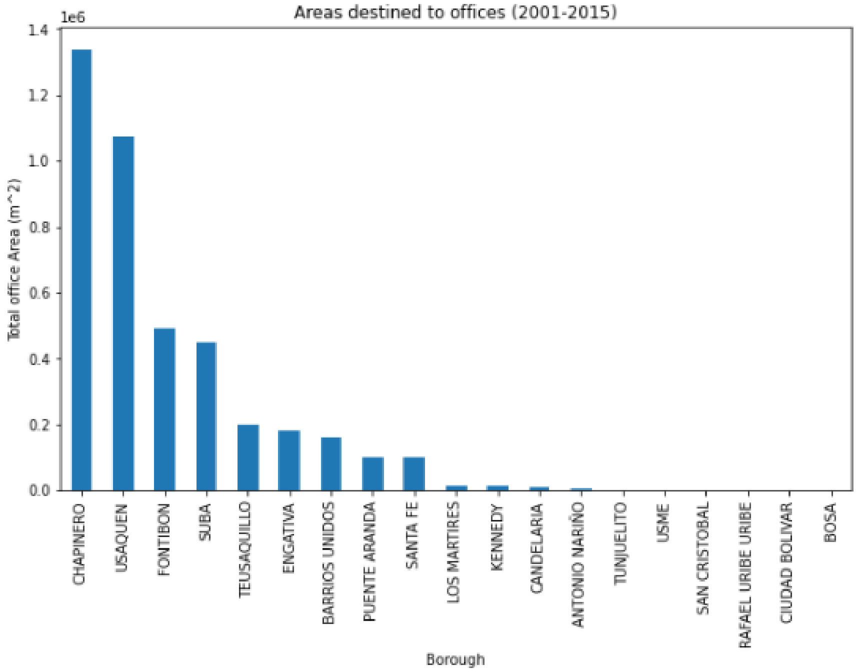


Fig 4. Areas destined to offices in each borough between the years 2001 and 2015.

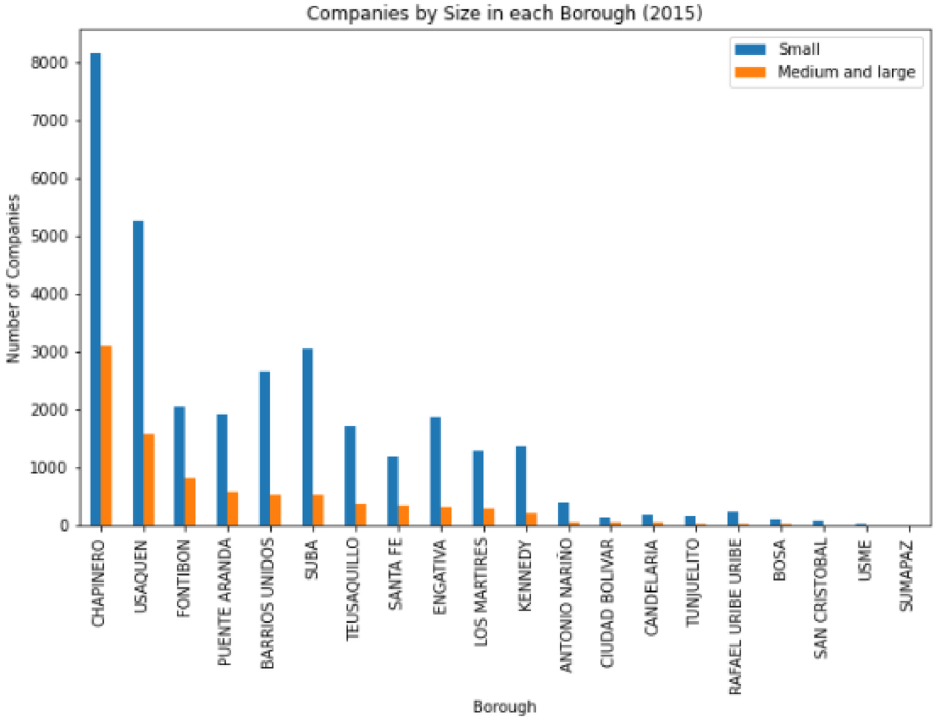
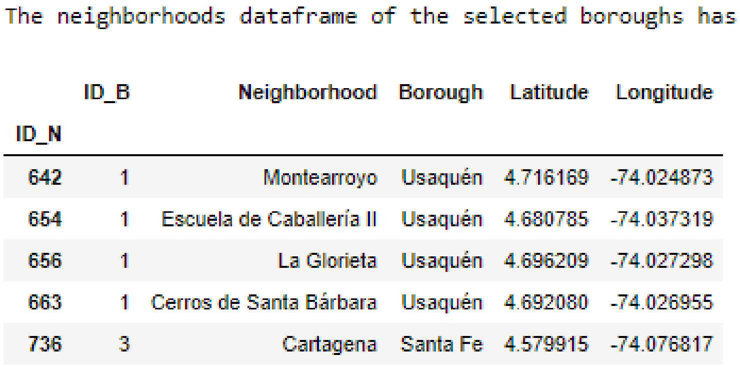


Fig 5. Number of companies by size in each borough in the year 2015.

Based on the previuos analyzis only four bouroghs were having into account: Candelaria, Santa Fé, Chapinero and Usaquén. The new dataframe with only the neighborhoods used is presented in table 3.



```
The neighborhoods dataframe of the selected boroughs has 455 rows and 5 columns.
```

	ID_B	Neighborhood	Borough	Latitude	Longitude
ID_N					
642	1	Montearroyo	Usaquén	4.716169	-74.024873
654	1	Escuela de Caballería II	Usaquén	4.680785	-74.037319
656	1	La Glorieta	Usaquén	4.696209	-74.027298
663	1	Cerros de Santa Bárbara	Usaquén	4.692080	-74.026955
736	3	Cartagena	Santa Fe	4.579915	-74.076817

Table 3. Data frame obtained for the neighborhood with only the selected boroughs (3).

The map in Fig 6 shows the location of the selected neighborhoods. It can be seen that the selected zone corresponds to the east of the city.

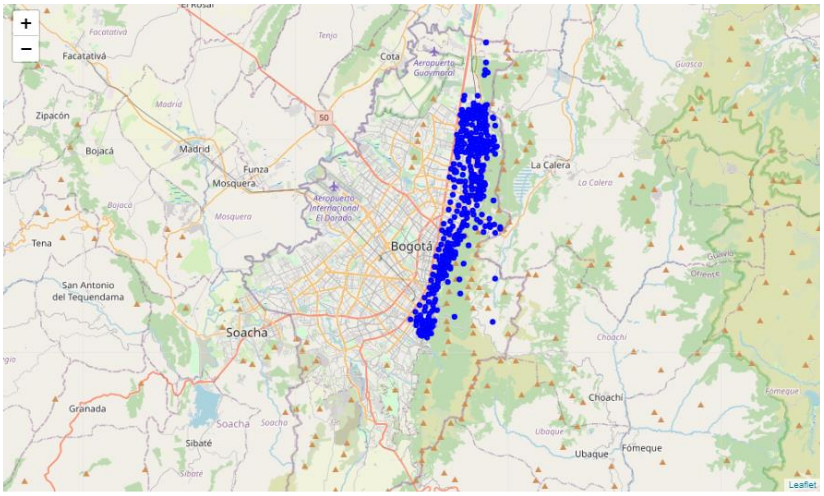
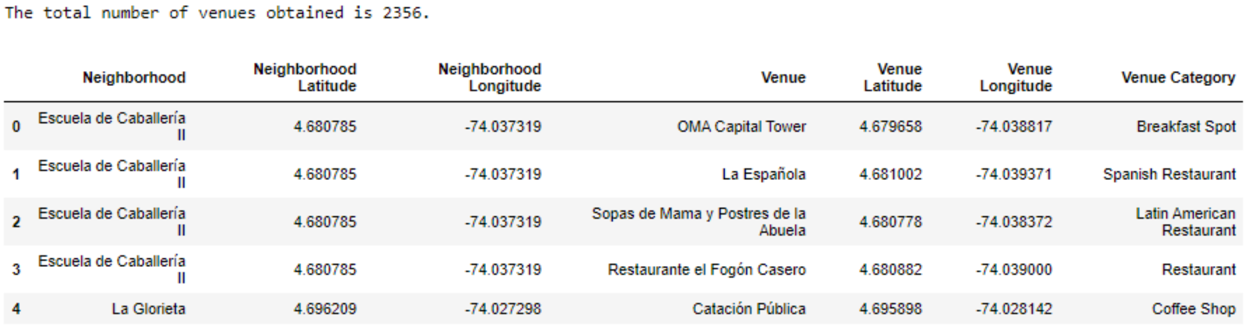


Fig 6. Map of Neighborhoods in the selected boroughs. The Neighborhoods are represented as blue dots.

Importing venues from Foursquare API

The Foursquare API is used to retrieve the most popular venues in each location. A limit of 100 venues in a radius of 250 meters is selected for each neighborhood. Some of the venues obtained are shown in Table 3.



```
The total number of venues obtained is 2356.
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Escuela de Caballería II	4.680785	-74.037319	OMA Capital Tower	4.679658	-74.038817	Breakfast Spot
1	Escuela de Caballería II	4.680785	-74.037319	La Española	4.681002	-74.039371	Spanish Restaurant
2	Escuela de Caballería II	4.680785	-74.037319	Sopas de Mama y Postres de la Abuela	4.680778	-74.038372	Latin American Restaurant
3	Escuela de Caballería II	4.680785	-74.037319	Restaurante el Fogón Casero	4.680882	-74.039000	Restaurant
4	La Glorieta	4.696209	-74.027298	Catación Pública	4.695898	-74.028142	Coffee Shop

Table 3. Venues obtained for the neighborhoods in the borough selected.

Since this project focuses only on restaurants, all the other venues that do not contain the word restaurant in the venue category column are dropped. A total number of 654 restaurant venues is obtained.

The number of venues retrieved for each restaurant category is presented in Fig. 7. Most of the venues are only categorized as restaurants, therefore these do not belong to any specific category. The Italian and fast-food restaurants would be the type of restaurants more common in the chosen boroughs.

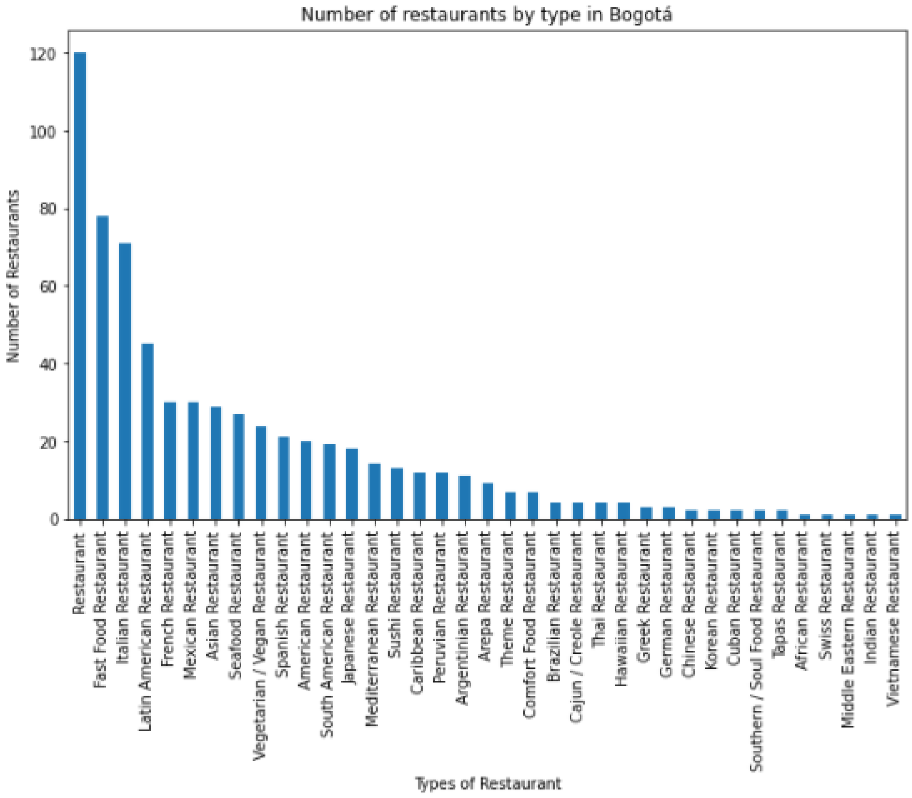


Fig 7. Number of restaurants category in the selected boroughs

The number of restaurants in each borough is shown in Fig. 8. The boroughs Usaquén and Chapinero have a considerable difference in comparison with the two last ones (Santa Fé and La Candelaria).

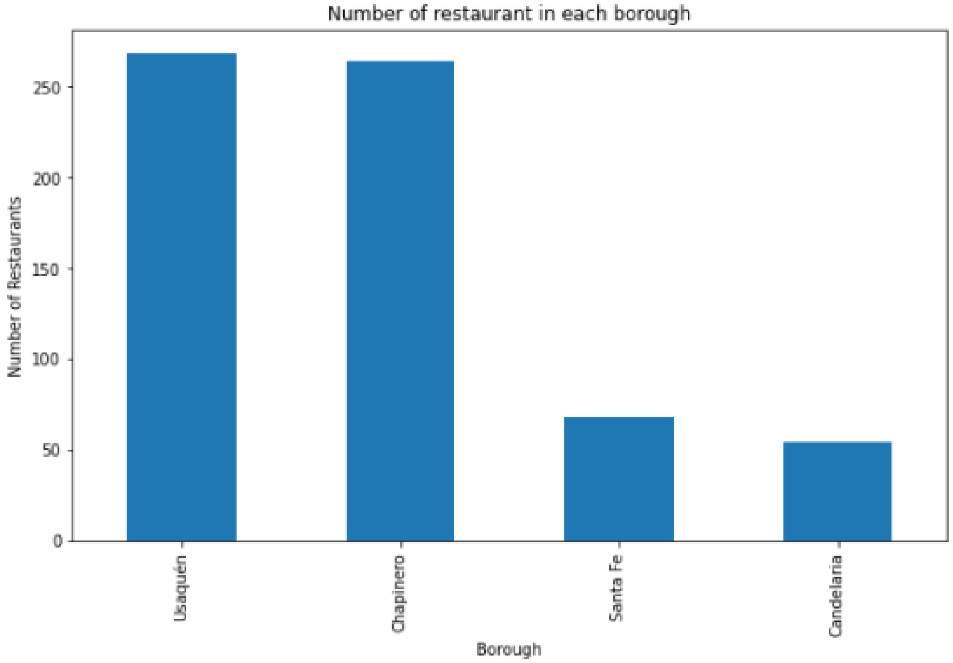
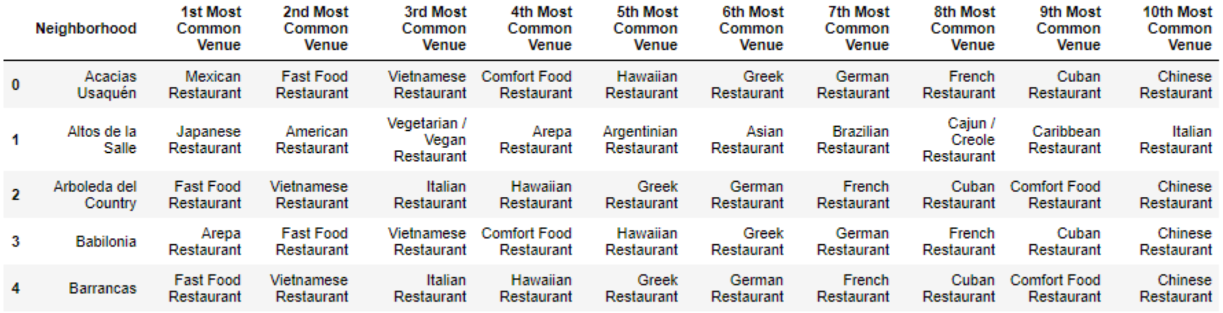


Fig 8. Total Number of restaurant venues in each borough selected

Clustering

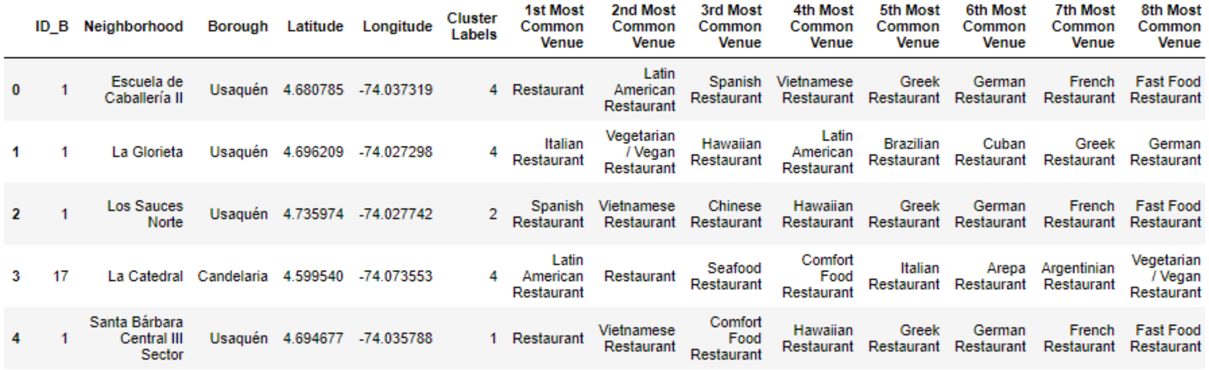
Before segmenting the Neighborhoods in clusters, the data should be arranged. The dummie function was used to create a binary column for each category, the column with the neighborhoods was also inserted to this dataframe. The rows were grouped by neighborhood and the mean frequency is taken for each category. Then, a dataframe was also created to display the 10 most common venues categories in each neighborhood, as shown in Table 4. This dataframe was used to cluster the neighborhoods in 5 groups.



	Neighborhood	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	Acacias Usaquén	Mexican Restaurant	Fast Food Restaurant	Vietnamese Restaurant	Comfort Food Restaurant	Hawaiian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Cuban Restaurant	Chinese Restaurant
1	Altos de la Salle	Japanese Restaurant	American Restaurant	Vegetarian / Vegan Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Brazilian Restaurant	Cajun / Creole Restaurant	Caribbean Restaurant	Italian Restaurant
2	Arboleda del Country	Fast Food Restaurant	Vietnamese Restaurant	Italian Restaurant	Hawaiian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Cuban Restaurant	Comfort Food Restaurant	Chinese Restaurant
3	Babilonia	Arepa Restaurant	Fast Food Restaurant	Vietnamese Restaurant	Comfort Food Restaurant	Hawaiian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Cuban Restaurant	Chinese Restaurant
4	Barrancas	Fast Food Restaurant	Vietnamese Restaurant	Italian Restaurant	Hawaiian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Cuban Restaurant	Comfort Food Restaurant	Chinese Restaurant

Table 4. Most common restaurants categories in each neighborhood

The dataframe with the cluster labels and most common venue category are grouped with the dataframe that contains the geospatial position of each neighborhood. Both dataframes have in common the neighborhood column, which is used to pair the rows. Some neighborhoods had to be dropped since for them, data were not obtained from the Foursquare API. The resulting dataframe is presented in Table 5. The distribution of the clustered neighborhoods in the map of Bogotá is shown in Fig. 9.



	ID_B	Neighborhood	Borough	Latitude	Longitude	Cluster Labels	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
0	1	Escuela de Caballería II	Usaquén	4.680785	-74.037319	4	Restaurant	Latin American Restaurant	Spanish Restaurant	Vietnamese Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Fast Food Restaurant
1	1	La Glorieta	Usaquén	4.696209	-74.027298	4	Italian Restaurant	Vegetarian / Vegan Restaurant	Hawaiian Restaurant	Latin American Restaurant	Brazilian Restaurant	Cuban Restaurant	Greek Restaurant	German Restaurant
2	1	Los Sauces Norte	Usaquén	4.735974	-74.027742	2	Spanish Restaurant	Vietnamese Restaurant	Chinese Restaurant	Hawaiian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Fast Food Restaurant
3	17	La Catedral	Candelaria	4.599540	-74.073553	4	Latin American Restaurant	Restaurant	Seafood Restaurant	Comfort Food Restaurant	Italian Restaurant	Arepa Restaurant	Argentinian Restaurant	Vegetarian / Vegan Restaurant
4	1	Santa Bárbara Central III Sector	Usaquén	4.694677	-74.035788	1	Restaurant	Vietnamese Restaurant	Comfort Food Restaurant	Hawaiian Restaurant	Greek Restaurant	German Restaurant	French Restaurant	Fast Food Restaurant

Table 5. Dataframe obtained after segmenting the neighborhoods in clusters

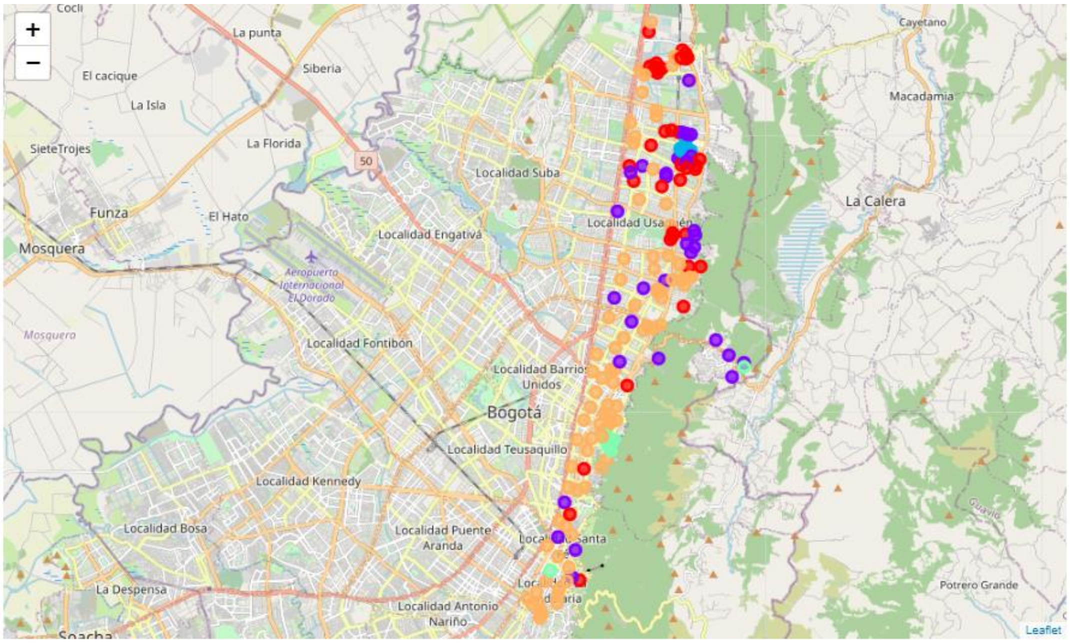


Fig 9. Distribution of the clustered neighborhoods in the map of Bogotá. Where red dots = cluster 0, purple dots = cluster 1, the blue dots = cluster 2, cyan dots = cluster 3 and orange dots = cluster 4.

Results and Discussion

The data obtained for the boroughs were used to select the best possible boroughs for a restaurant. The number of companies, area of offices and number of tourist places in each borough were assesed. This allowed to group the selected boroughs in three types:

- Boroughs mainly tourists as La Candelaria and Santa Fé, which have a higher number of tourist places and a low number of offices and companies.
- an industrial/occupational borough (Usaquén) which higher number of offices and companies and a lower number of tourist places.
- and a borough Mixed (Chapinero), where there is a high number of office areas and companies and a considerable amount of tourist places.

This classification could help to identify the kind of costumers that could be expected in each borough. As it was shown above, turists boroughs as La Candelaria and Santa Fé have a very lower amount of restaurants in comparisson with the other two bouroghs explored. On the other hand, the Italian and fast-food restaurant are the most common type of restaurants in the zomes explored. Interestingly, Latin American food restaurants occupy third place with a considerable difference from the first two restaurant types.

The neighborhoods were clustered based on the most common restaurant category in each one of them. The classification and how each cluster is interpreted is shown below:

Cluster 0: In this cluster, the fast-food restaurants are the most common restaurant types. These neighborhoods are mostly located in Usaquen. Despite this borough has a low number of tourist attractions, has a considerable number of companies and office areas. Therefore, the high number of fast-food restaurants could be due to workers or locals that do not search high price restaurants.

Cluster 1: This cluster corresponds mainly to neighborhoods with restaurants without a defined category assigned and other international food restaurants. The neighborhoods of this cluster are located mostly north in the boroughs of Chapinero and Usaquén.

Cluster 2: The Spanish restaurants are the most common in these areas. With the exception of one, All the neighborhoods belonging to this cluster are located in Usaquén.

Cluster 3: The Italian and other international restaurants are the most common in the neighborhoods belonging to this cluster. The neighborhoods that belong to this cluster are located in the boroughs Chapinero and Santa Fé.

Cluster 4: In this cluster, the most frequent are the Latin American/southamerican restaurants, which also include Argentinian, Brazilian, Peruvian, Mexican and Cuban restaurants. These clusters are more common in all the boroughs explored but mainly in the south where there are more tourist attractions.

Conclusions

In this project the data available of the neighborhoods in Bogotá, Colombia was used to explore and cluster them in order to obtain insights into the distributions of restaurants in the city. Data of tourist places, companies, and office areas were used to explore the boroughs and select the most adequate ones. A deeper exploration is needed in order to obtain a better description that allows determining the boroughs that have the best characteristics or have a higher possible number of customers for a restaurant. In order to do this, more data would be needed, as for example information about urban zones, malls, trading points, etc.

The neighborhoods were clustered in five groups based on the most common restaurant categories, in each one of them. This allowed finding areas where there are similar types of restaurants. Tourist neighborhoods like Santa Fé and La Candelaria were found to have a lower number of restaurants, and most of them Latin-American food is served. In Usaquén, where there are a considerable number of offices and companies, the fast-food restaurants were the most common. The last borough explored Chapinero has both tourist and business areas, and here different international food restaurants were found. It is expected that this analysis could help investors from the catering industry who are planning to start a business in one of Bogotá neighborhoods.

References

(1) Visit Colombia feel the rhythm (2021/01) Tourism in Bogota: a city for experiencing culture. https://colombia.travel/en/blog/tourism-bogota-city-experiencing-culture

(2) Tripadvisor. (2021/01) Explora Bogotá. https://www.tripadvisor.co/Tourism-g294074-Bogota-Vacations.html

(3) Laboratorio Urbano Bogotá´. (2015/01) https://bogota-laburbano.opendatasoft.com/explore.

(4) Alcaldía de Bogotá D.C. (2021/01) https://bogota.gov.co/mi-ciudad/localidades/sumapaz