Using Analytics to Crack Bogota's Gastronomic Tour:
Neighborhood classification in Bogota based on their gastronomic offer

Juan David Rodríguez Lamus
December 12, 2019

Background

Bogota has dramatically increased its gastronomic offer. Dozens of restaurants are opened every day, and hot zones are wider each month, with alternative offerings being opened in completely new places, far away from traditional restaurant zones.

Introduction and business problem

This project intends to cluster neighborhoods in Bogota depending on their restaurants' offering so that Foodies can easily define a Tour route based on their preferences. Two deliverables are presented in this report:

1. Descriptive statistics from Bogotá venues, including the amount of restaurants per category, and a map showing their geographical distribution
2. Clustering of neighborhoods based on offer density and restaurant categories. Following deliverables will be presented:
   a. A description of each cluster
   b. A table with the classification of each neighborhood in Bogota, identifying the cluster it belongs to, amount of restaurants in the area, and distribution of restaurants per category
   c. A map to represent neighborhood clusters in a graphical matter

Deliverables of this project will allow the users to have an accurate map of the restaurant offer in Bogota and explore alternative offers that are not so obvious in the mainstream gastronomic scene in Bogota. Resulting information can also be an input for the commercial zones to increase their visibility for Foodie tourism or for the tourism industry to better structure gastronomic tourism offerings.

Data

Data Source for Bogota Neighborhood Location

Bogota's Urban Laboratory Institute ("Laboratorio Urbano de Bogota") has a comprehensive database that including the name, location and description of each neighborhood in Bogota. I used that database as the main source of external data for the project. Raw data can be downloaded from the following link: https://bogota-laburbano.opendatasoft.com/explore/dataset/barrios_prueba/download?format=xls

Data Uploading and Data Cleaning

I imported the file to the project storage using the "File and Add data" functionality from the Jupyter Notebook application. Once uploaded I had to clean the data (as in any analytics project the source data was not perfect and has to be modified in format and content to ease the following steps in the analysis). Some of the modifications I made to the data set were:

1. I dropped several lines that didn't have a Borough identified ("Localidad"), that were not a legal neighborhood, or that didn't have a name
2. The original table had several columns that were not required for the analysis I intended to perform, original data set included the following columns: "OBJECTID, Codigo Localidad, Localidad, Estado, Nombre, Codigo, SHAPE.AREA, SHAPE.LEN geo_shape geo_point_2d". I only used the columns "OBJECTID Localidad and Nombre geo_point_2d". Column geo_point_2d includes both Latitude Longitude and has to be separated. I then dropped all other columns.
3. Finally, I re-named the columns as follows: "Neighborhood_ID Borough Neighborhood Latitude Longitude"

For Example: Neighborhood "Usaquén" is one of the most important ones for the restaurant industry in Bogota. This neighborhood will definitely have to be included in the data set, final values for the mentioned columns shall be:

1. **Neighborhood_ID**: 657
2. **Borough**: Usaquén
3. **Name**: Usaquén
4. **Latitude**: 4.69474025606
5. **Borough**: -74.030740809

Neighborhood "Ciudadela El Poblado" in line 3000 from the original table on the contrary is not a legal neighborhood and doesn't specify a Borough, it will be dropped accordingly

The following figure shows the structure from the source data:

Barrios de Bogotá 2016

	OBJECTID	Codigo Localidad	Localidad	Estado	Nombre	Codigo	SHAPE.AREA	SHAPE.LEN
1	1.430	19	Ciudad Bolivar	LEGALIZADO	Jerusalén La Pradera	190.083	0.000	0.022
2	1.431	19	Ciudad Bolivar	LEGALIZADO	Gibraltar I	190.076	7.3505201702556e-7	0.004
3	1.434	19	Ciudad Bolivar	LEGALIZADO	San Francisco La Palmera	190.212	7.37351728543469e-7	0.006
4	1.437	19	Ciudad Bolivar	LEGALIZADO	La Florida del Sur	190.180	0.000	0.006
5	1.440	19	Ciudad Bolivar	SIN LEGALIZAR	Urb. Villa del Rio Zona Comercial	190.285	0.000	0.017
6	1.444	19	Ciudad Bolivar	SIN LEGALIZAR	El Edén y Librmani	190.275	0.000	0.010
7	1.448	19	Ciudad Bolivar	LEGALIZADO	Naciones Unidas (Santa Rosa)	190.240	0.000	0.006
8	1.450	19	Ciudad Bolivar	LEGALIZADO	El Tesoro	190.067	0.000	0.024
9	1.452	19	Ciudad Bolivar	SIN LEGALIZAR	S C La Torre	190.181	0.000	0.006
10	1.453	19	Ciudad Bolivar	LEGALIZADO	Villa Jacqui	190.264	0.000	0.006
11	1.459	19	Ciudad Bolivar	LEGALIZADO	Bella Flor	190.016	0.000	0.024

Methodology

I followed the following steps to analyze the data and organize it in a way that allows statistical methods to provide meaningful results.

1. Create Bogota Map and Add Neighborhood Markers

The map provided me density information as well as information on the distribution of neighborhoods per Borough. I was able to identify that even after data cleanup, there is a significant amount of neighborhoods in Bogota, and predict that once the information is filtered with the Foursquare search query, the amount of relevant neighborhoods for the study shall be significantly reduced.

2. Foursquare query

Once Bogota neighborhoods' source information was clean, and the data was in the proper format, I performed a search in Foursquare for restaurant venues for each neighborhood. The output of the query was cleaned and grouped in order to provide meaningful results:

A. The amount of categories obtained from the Foursquare query is extensive, and as a result it will be harder to understand and cluster results

For Example: When I query Usaquen, it will probably show a wide number of categories for different Asian restaurants, including probably: Asian, Filipino, Chineese. I will group them as Asian.

The final categories were defined based on an exploratory analysis and frequency count from the original categories provided by Foursquare. I went from 89 categories originally provided by Foursquare to only 9:

1. American/Fast Food Restaurant
2. Asian/Japanese Restaurant
3. Bar/Winery, Cafe/Bakery
4. Diner/General Restaurant
5. International/European Restaurant
6. Italian/Pizza Restaurant
7. Latin American/Colombian Restaurant
8. Seafood/Fish Restaurant

B. The information was consolidated and merged with the original dataframe

3. Exploratory analysis on Bogota restaurant data

Graphical Representation from the Neighborhoods' Restaurant Offer in Bogota

In order to analyze the results, I have created two graphs:

1. A map with the neighborhoods that have restaurants according to Foursquare search query, and
2. A bar chart with the amount of restaurants per category and per Borough in the entire city.

4. Clustering

Once the final dataframe was properly set, I performed the clustering. Neighborhoods were clustered using the amount of restaurants per grouped category, and the total amount of restaurants as clustering variables. Total amount of restaurants is included in the analysis in order to provide a sense of offer density.

An initial analysis using 5 clusters resulted in one cluster including only one neighborhood. Therefore, a **second analysis was performed using 4 clusters**, resulting in interesting insights.

## 5. Analysis and presentation¶

Finally, I performed an analysis based on the results, categorized each neighborhood and provided a description of the characteristics of each neighborhood.

Results

Exploratory Results for Bogota data
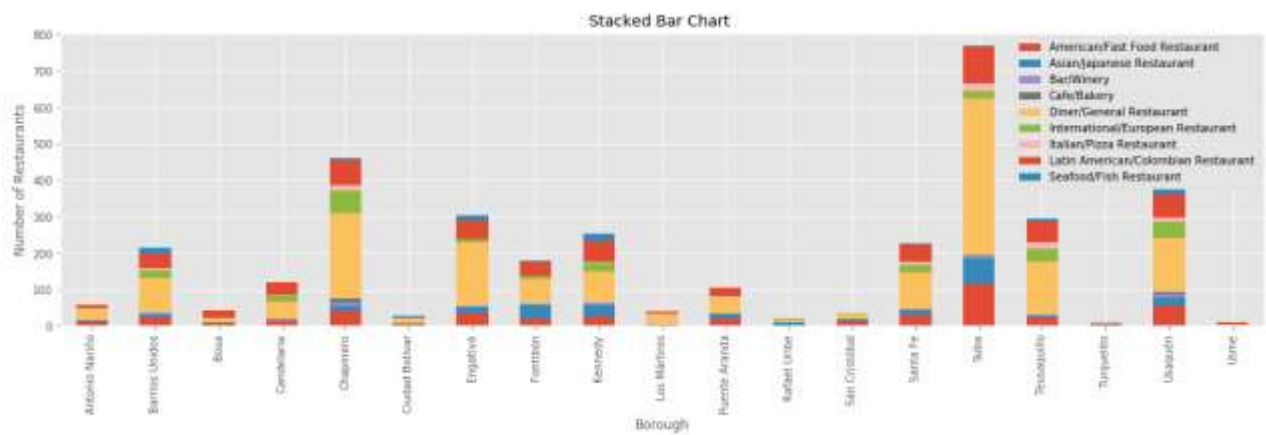
Information was processed and grouped to provide insights on the classification of neighborhoods in Bogota depending on their gastronomic offer.

A first graphical representation in the map showed the complexity of the analysis given the amount and distribution of neighborhoods in Bogota:
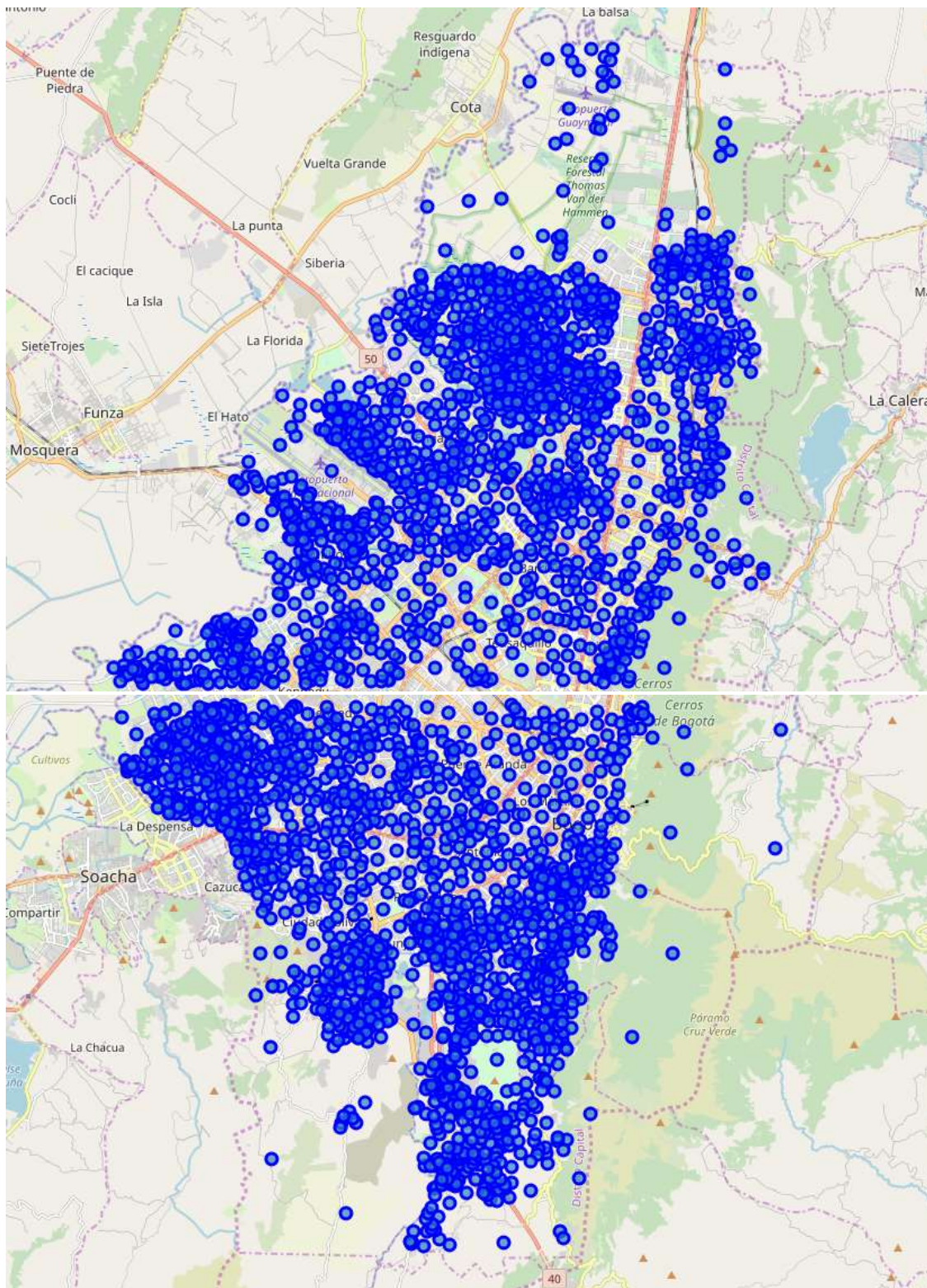
The following table and bar chart shows the distribution of restaurants per category for each Borough:

Borough	American/Fast Food Restaurant	Asian/Japanese Restaurant	Bar/Winery	Cafe/Bakery	Diner/General Restaurant	International/European Restaurant	Italian/Pizza Restaurant	Latin American/Colombian Restaurant	Seafood/Fish Restaurant
Antonio Nariño	11.0	3.0	0.0	2.0	31.0	0.0	2.0	8.0	0.0
Barrios Unidos	21.0	18.0	4.0	2.0	94.0	22.0	4.0	40.0	17.0
Bosa	7.0	3.0	0.0	0.0	11.0	0.0	0.0	20.0	0.0
Candelaria	13.0	1.0	1.0	5.0	45.0	19.0	3.0	34.0	1.0
Chapinero	38.0	16.0	9.0	13.0	233.0	61.0	16.0	68.0	6.0
Ciudad Bolívar	5.0	2.0	0.0	0.0	19.0	0.0	6.0	0.0	0.0
Engativá	34.0	18.0	5.0	0.0	174.0	8.0	1.0	46.0	16.0
Fontibón	20.0	38.0	2.0	1.0	86.0	7.0	1.0	40.0	3.0
Kennedy	25.0	33.0	6.0	0.0	88.0	24.0	3.0	56.0	20.0
Los Mártires	1.0	1.0	3.0	0.0	25.0	1.0	2.0	5.0	1.0
Puente Aranda	20.0	12.0	0.0	2.0	43.0	2.0	1.0	24.0	1.0
Rafael Uribe	2.0	7.0	0.0	0.0	7.0	3.0	2.0	2.0	0.0
San Cristóbal	13.0	5.0	0.0	0.0	14.0	2.0	0.0	0.0	0.0
Santa Fe	27.0	12.0	1.0	6.0	88.0	22.0	8.0	47.0	3.0
Suba	114.0	71.0	6.0	2.0	401.0	20.0	21.0	100.0	2.0
Teusaquillo	21.0	8.0	2.0	1.0	145.0	35.0	17.0	57.0	8.0
Tunjuelito	2.0	2.0	0.0	0.0	4.0	0.0	0.0	1.0	0.0
Usaquén	58.0	23.0	8.0	6.0	148.0	45.0	10.0	64.0	16.0
Usme	0.0	0.0	0.0	0.0	5.0	0.0	0.0	4.0	0.0

Stacked Bar Chart

Legend:
- American/Fast Food Restaurant
- Asian/Japanese Restaurant
- Bar/Winery
- Cafe/Bakery
- Diner/General Restaurant
- International/European Restaurant
- Italian/Pizza Restaurant
- Latin American/Colombian Restaurant
- Seafood/Fish Restaurant

Y-axis: Number of Restaurants
X-axis: Borough

A map was also created including the neighborhoods with restaurants according to the Foursquare search query. The size of the marker was defined by the amount of restaurants in each neighborhood:
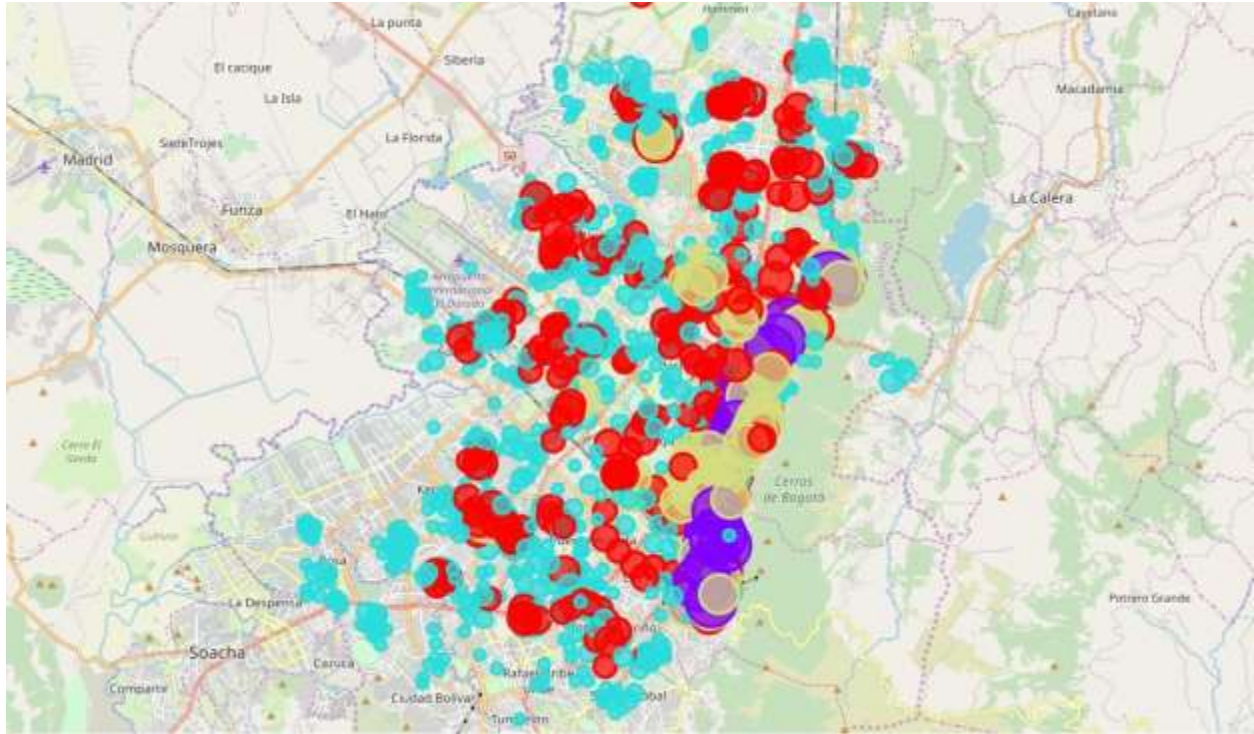

Clustering Results

Initially, the clustering analysis was performed using 5 clusters, obtaining the following results:

As seen in the table above, cluster 4 has only one neighborhood. Therefore I decided to run the analysis again using 4 clusters. Following results were obtained:

The new analysis resulted in four clusters with differentiated characteristics and sufficient amount of neighborhoods per cluster to be relevant.

I then continued to create a map representing every neighborhood by their cluster classification (color of the marker: Cluster 0: Red, Cluster 1: Purple, Cluster 2: Green, Cluster 3: Yellow) and number of restaurants per neighborhood (size of the marker):

Discussion

Results provided insightful data on neighborhood segmentation and restaurant distribution in Bogota:

Cluster Typification and Naming

Cluster 0: Neighborhoods with low density restaurant offering and mostly general restaurants without any specialization

Cluster 0 is characterized by its relatively low restaurant offering, composed mostly by diners or general restaurants without any specialization. They have the highest percentage of Diner/General Restaurant and the lowest for Latin American/Colombian. This is probably the group that would be the least interesting for a Foodie Tour

Cluster 1: Neighborhoods with the highest density restaurant, mainstream places for the restaurant scene in Bogota

Cluster 1 has the highest density with more than 25 restaurants per neighborhood in average and the highest offering of International/European cuisine and Cafes/Bakeries compared to other clusters. It also has good offering of Latin American and Colombian cuisine and Italian restaurants. This is probably the most interesting group for a Foodie tour, it is composed by 19 neighborhoods with a very interesting offering.

This cluster includes 19 neighborhoods including the most famous and traditional in Bogota restaurant scene such as La Candelaria, Macarena, Usaquen, Virrey, Parque 93 or Quinta Camacho

The following table shows the 10 neighborhoods with the highest amount of restaurants in this group.

Cluster Labels	Neighborhood_ID	Borough	Neighborhood	Latitude	Longitude	Total Restaurants	American/Fast Food Restaurant	Asian/Japanese Restaurant	Bar/Winery	Cafe/Bakery	Diner/General Restaurant	International/European Restaurant	Italian/Pizza Restaurant	Latin American/Colombian Restaurant	Seafood/Fish Restaurant	Total Restaurants %
62	1	287.3	Candelaria	La Celedea	4.58054	-74.0738	44.0	7.0	1.0	1.0	3.0	19.0	3.0	1.0	14.6	1.0	0.012084
961	1	3290.8	Santa Fe	Veracruz	4.59325	-74.3723	31.0	1.0	3.0	3.0	3.0	17.0	2.0	0.0	4.5	1.0	0.009784
660	1	3297.0	Santa Fe	San Martin	4.51698	-74.3674	30.0	7.0	1.0	0.0	2.0	19.0	4.0	0.0	0.0	0.0	0.008482
797	1	2520.0	Usaquen	Urb. El Cataleno	4.59911	-74.0333	27.0	2.0	0.0	0.0	0.0	12.0	3.0	2.0	0.0	0.0	0.007634
788	1	2710.8	Chapinero	S.C. Chapinero Central	4.64480	-74.3639	27.0	1.0	3.0	0.0	1.0	16.0	1.0	1.0	4.0	0.0	0.007634
179	1	657.3	Usaquen	Usaquen	4.59474	-74.3107	26.0	2.0	0.0	1.0	0.0	19.0	4.0	1.0	0.0	0.0	0.007361
324	1	2040.0	Chapinero	S.C. Quinta Camacho	4.65470	-74.3066	26.0	4.0	1.0	0.0	1.0	12.0	2.0	1.0	4.0	0.0	0.007361
962	1	3260.0	Santa Fe	San Diego	4.61280	-74.0087	26.0	3.0	1.0	0.0	1.0	12.0	2.0	3.0	0.0	0.0	0.007361
726	1	2961.3	Chapinero	S.C. Lago Gaitan	4.59997	-74.3586	25.0	1.0	0.0	1.0	1.0	11.0	5.0	0.0	3.0	1.0	0.007080
558	1	3295.0	Santa Fe	La Macarena	4.51425	-74.3054	23.0	2.0	0.0	0.0	1.0	5.0	7.0	3.0	0.0	0.0	0.006563

Cluster 2: Neighborhoods with the lowest density restaurant offering with high percentage of fast food restaurants

Cluster 2 is the biggest one of all, with 764 neighborhoods. It has very low density with only 1.3 restaurants per neighborhood in average. It has the highest percentage of fast food restaurant and seafood restaurants. Interestingly enough, it also has a high percentage of Asian restaurants compared to the other clusters.

Just like cluster 0, this cluster is probably not interesting for a Foodie tour, it is likely that most of the restaurants are simple places where locals have lunch regularly

Cluster 3: Neighborhoods with high density restaurant offering and alternative offering to the traditional/famous zones

Cluster 3 includes 71 neighborhoods with an average of 10 restaurants per neighborhood. Although the offer is not as dense as in cluster 1, it could be a very interesting alternative to the mainstream, traditional offering from Bogota. It includes a high percentage of Latin American and Colombian restaurants and is the second one with the highest offering of International/European cuisine.

This group includes very interesting neighborhoods such as Teusaquillo, Chapinero, Pasadena or La Capuchina that are worth exploring to taste alternative and authentic cuisine that is probably not in the regular Tour guides.

The following table shows the 10 neighborhoods with the highest amount of restaurants in this group.

Cluster Labels	Neighborhood_ID	Borough	Neighborhood	Latitude	Longitude	Total Restaurants	American/Fast Food Restaurant	Asian/Japanese Restaurant	Bar/Winery	Cafe/Bakery	Diner/General Restaurant	International/European Restaurant	Italian/Pizza Restaurant	Latin American/Colombian Restaurant	Seafood/Fish Restaurant	Total Restaurants %
787	3	2712.0	Chapinero	S.C. Sucre	4.62983	-74.3654	18.0	4.0	1.0	1.0	0.0	4.0	1.0	1.0	5.0	1.0	0.005080
739	3	3082.0	Chapinero	El Retiro - Villa Amelia	4.64683	-74.0631	17.0	2.0	0.0	2.0	0.0	7.0	0.0	1.0	4.0	1.0	0.004690
727	3	2546.0	Chapinero	Chico Reservado	4.67014	-74.0426	17.0	0.0	0.0	0.0	0.0	8.0	1.0	1.0	5.0	0.0	0.004690
882	3	3083.5	Usaquen	Usaquen - Urb. Portal Al Boro	4.68141	-74.0206	17.0	0.0	0.0	1.0	0.0	8.0	4.0	1.0	3.0	0.0	0.004690
726	3	2546.0	Chapinero	Chico Norte - Urb. Conjunto de la Cien	4.68048	-74.3414	15.0	1.0	0.0	0.0	0.0	9.0	1.0	1.0	4.0	0.0	0.004241
732	3	2634.0	Chapinero	S.C. Porciuncula	4.66983	-74.3678	15.0	4.0	0.0	0.0	0.0	8.0	0.0	0.0	3.0	0.0	0.004241
259	3	675.0	Teusaquillo	Teusaquillo	4.62398	-74.3723	15.0	2.0	0.0	1.0	0.0	8.0	1.0	1.0	2.0	0.0	0.004241
982	1	2972.0	Chapinero	S.C. Granada	4.64047	-74.3981	14.0	2.0	1.0	0.0	0.0	9.0	0.0	0.0	1.0	0.0	0.003966
901	3	3250.0	Chapinero	S.C. La Salle	4.64474	-74.039	13.0	0.0	1.0	0.0	1.0	7.0	1.0	1.0	3.0	0.0	0.003675
779	3	2674.3	Chapinero	S.C. Antiguo Country	4.67199	-74.3913	13.0	2.0	0.0	1.0	1.0	9.0	0.0	0.0	3.0	0.0	0.003675

Conclusion

Interpretation of results and impact on the Bogota restaurant business

Just as expected, Bogota has a wide offer of restaurants and extensive variety of categories. However, the restaurants are not evenly distributed among neighborhoods; analytical tools performed during this project have resulted in a very insightful classification of the neighborhoods depending on their restaurants' offer.

- Cluster 0 and Cluster 2 are mostly neighborhoods with a low amount of restaurants per neighborhood. Their offer is mostly fast food or diners where locals probably have lunch on a daily basis
- Cluster 1 and 3 are the ones more interesting for the restaurant offer: Cluster 1 includes the mainstream areas that are commonly known. Cluster 3 are alternative neighborhoods with high amount of restaurants and wide variety, and probably areas that are starting to take off beyond the traditional offer.

Results have also shown that Latin American food and International food are the categories that differentiate clusters with high density and potential. While fast food and diners are the ones that are mostly common in neighborhoods with low gastronomic offer.

Further analysis on the restaurant business

A further and natural step would be to enhance this analysis using customer recommendations or trending queries. Unfortunately Foursquare has not enough information in Bogota on this point to provide meaningful, specially for non-mainstream venues. This information could be gathered from alternative sources that are more popular in Colombia such as google maps.

Another improvement would be to further analyze/cluster neighborhoods within clusters 1 and 3 from our analysis

Finally, a periodic and sequential analysis using the same methodology and recommendations could result in an interesting Artificial Intelligence project to try to predict the future behaviour of restaurant areas in Bogota.

Analysis on the source data, methodology and analytical tools

Although the source data from Bogota's Urban Laboratory Institute was surprisingly complete and accurate for Bogota, a deep analysis of the data show some inconsistencies that although had no material impact on the conclusions, might have affected the results. Some of these inconsistencies were identified thanks to the visual analysis performed with the maps:

- Incorrect Borough assignment
- Overlapping areas in some of the neighborhoods
- Incorrect names for some of the neighborhoods
- Out-of-date legal status for some of the neighborhoods

Other official sources of information (such as the ones in the governmental records office "catastro") would be definitely more precise and could improve the results.

Methodology was proven to be powerful and capable of providing meaningful insights: Out of the 3700 neighborhoods in the original source data, we were able to differentiate and describe 90 neighborhoods in

Clusters 1 and 3 with high variety of offering and potential, including many neighborhoods that are not traditionally identified as hubs for the gastronomic industry in Bogota.