Exploring Bogota venues across all its neighborhoods using the Foursquare API

IBM Data Science Capstone project

Daniel Andrade

Contents

1. Introduction

In this project an analysis of the Bogota neighborhoods, capital city of Colombia (South America) would be done based in the venues present across all neighborhoods at the present date using the foursquare API. The main idea is to explore most of the foursquare registered venues, identifying its category, location and possibles relations among other venues in the same neighborhood. Bogota it is know as one of the biggest city in Latin America region with almost 7.200.000 inhabitants and it is in the 6th place in population across all the region. Using the theory and techniques reviewed in the course it would be possible to identify if certain neighborhood its mostly an industrial place or a office place or other type of place, in other words identify what venue category correspond to each neighborhood as well as identify the top 10 most common and less common venues for each neighborhood. This information could be used for someone stockholders interested in open a new venue in a certain neighborhood, using the category along with some information related to the neighborhood such number of inhabitants.

Bogota has an area of 1775 km² in which almost 1800 small neighborhoods exists arranged in 20 boroughs each one know as locality. Something particular happens in Bogota as its a constantly growing city, some peripheral areas of the city are invaded for people looking for a place to live. More and more people come to this places until its big enough to be called a neighborhood, but for that it has to be legalized to the correspond local authorities, until that happens those neighborhoods are know as non legalized neighborhoods. Some information about those places are available and it will be included in the present analysis such as compare their location with the legalized neighborhoods, other analysis such a venue analysis will be left for an upcoming work.

Analysis of legalized vs non legalized neighborhoods are also valuable in terms of determining the internal situation of the city and it give the tools to the local autorities to develop strategies in order to improve those situations. It is also valuable as it give us the notion of how the city its organized in terms of its venues and it could be used for the local authorities in the implementation of better territorial arrangement planning, allowing the sustainable development of the city.

2. Data Description

The data to be used in this project is acquired from a local goverment site which contains information and data for most of the local features of the city such as name, number of neighborhood, geospatial coordinates, number of inhabitants among other data. It is a great site to get local data and information and it is part of an effort of the Colombian

government to provide open data of interest, all the data contained in this site could be exported as csv file. In particular the data for this project has the following structure:

OBJECTID	Locality Code	Locality	Legal Status	Neighborhood	Code	Latitude	Longitude

Where:

- **OBJECTID**: Is the identification of each entry in the dataset.
- **Locality Code**: Is the code of the borough or locality and can take values in the range of 1-20 for each defined locality in Bogota.
- **Locality or Borough**: The name of the locality or Borough.
- **Legal Status:** Describe the legal status of the neighborhood. This column can take values of 'LEGALIZED' OR 'NON LEGALIZED'
- **Neighborhood:** Name of the neighborhood.
- **Code:** Neighborhood's code.
- **Latitude and Longitude:** Geospatial coordinates of each Neighborhood.

Having the location of each neighborhood (latitude and longitude) it is possible to acquire all the correspond venues and its metadata (Category, Location, comments, etc) using the Foursquare API. These information is the base of the analysis that will be done in the current project. Once all the venues information is acquired and cleaned and organized, the neighborhoods will be clustered according with the similarities with the venues specifically the category of each one using the k-means machine learning algorithm. According with the results a respective analysis and comments will be done.

3. Methodology

3.1. Data Cleaning

The first step in the project is to clean the data we previously acquired from the local government site. Columns such as OBJECTID, LOCALITY CODE, CODE are not required for the current analysis so those columns are dropped from the dataset. As the dataset is coded in Spanish, after the corresponding translation is necessary to replace the separation comma for a point separation sign , this is done avoid any conflict whit the python libraries and to standardize the dataset.
It is also necessary to transform the geospatial coordinates (latitude and longitude) as in the original data set, are encoded as object type column.

After this is done the data set is separated into two subsets: Once with the information about the legalized neighborhoods and one with the information of the non legalized . The data cleaned should look something like this for both cases: legalized and non legalized dropping the correspond column as is not useful anymore.

Borough	Neighborhood	Latitude	Longitude	Longitude
0	Candelaria	Lourdes_2	4.589327309650732	-74.07284617350001

3.2. Exploratory Data Analysis EDA

Once the data is this stage some exploratory data analysis is done. For example we can determine how many categories are present in the Legal Status Column using the value_counts built in pandas function. This give a lot of information, the number of total neighborhoods and how many are legalized and how many are not.
With the two subsets ready its time to explore the location and distribution of each case using the using the folium package.

3.3. Foursquare API

After this the Foursquare API is used to get the venues and the metadata of each legalized neighborhood limited to some radius . An analysis of the non legalized neighborhood is left for some upcoming project.
With this information it is possible to analyze the distribution of venues per neighborhood, determining for example the most and less frequent venue as well as its category.
Its also possible to determine the frequency of each venue for each neighborhood using dummy variables which would be valuable for implementing the k-mean algorithm.

3.4. K-MEAN Machine Learning algorithm

With all the data manipulation that has been done until now, its is possible to cluster the neighborhoods according the similarities in the venues categories present in each neighborhood this is done using the K-MEANS algorithm.
First it is necessary determine the optimal k that will be used, for this the cost function is calculated for a range of values of k (1-20), a plot is done whit this data and the elbow method is use for determine the best value of k.
With the best value of k, the k-mean algorithm is deployed to cluster the neighborhoods according to th venues categories.

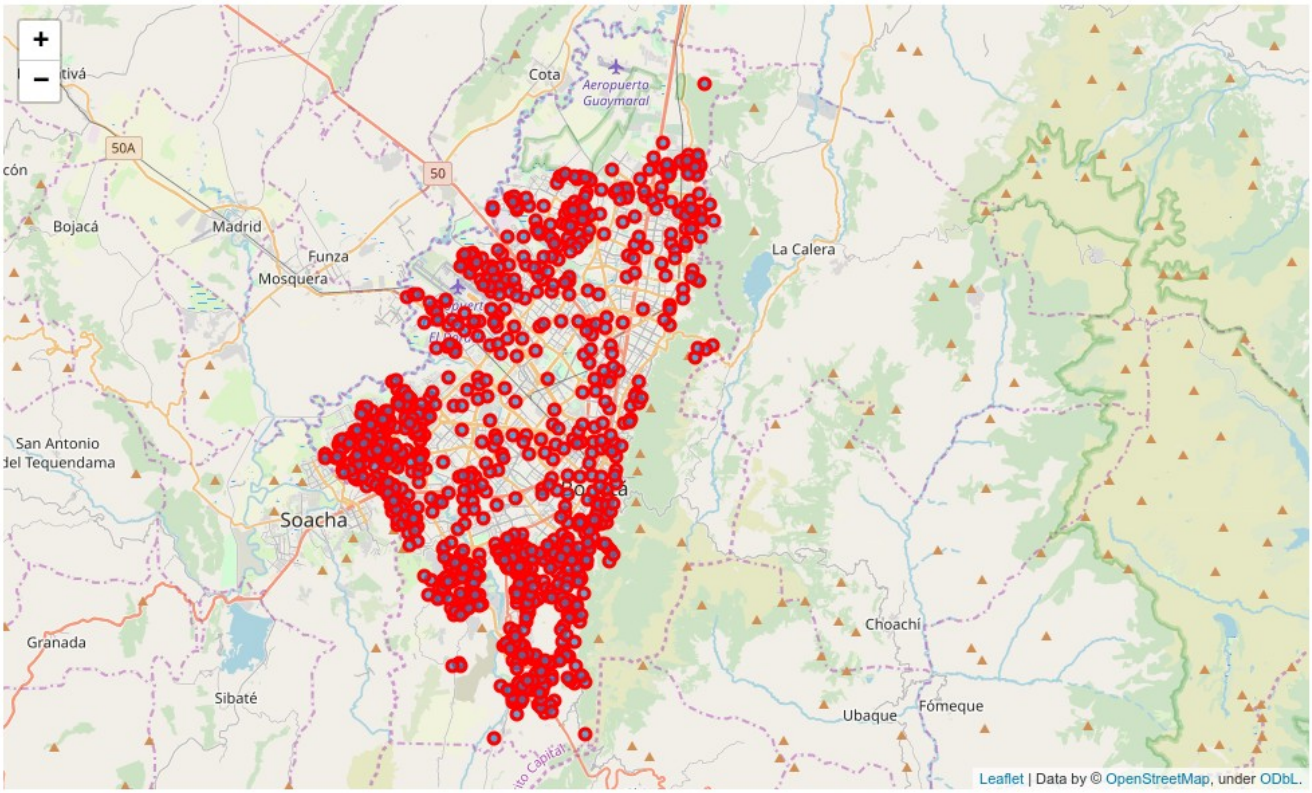
As a first approach to the results it is possible the visually explore the clusters using the folium python package. Then a more extensive analysis is done exploring each cluster formed with the data, determining the top 10 venues most frequent venue as well the respective k cluster.

4. Results and discussion

Legalized vs non legalized

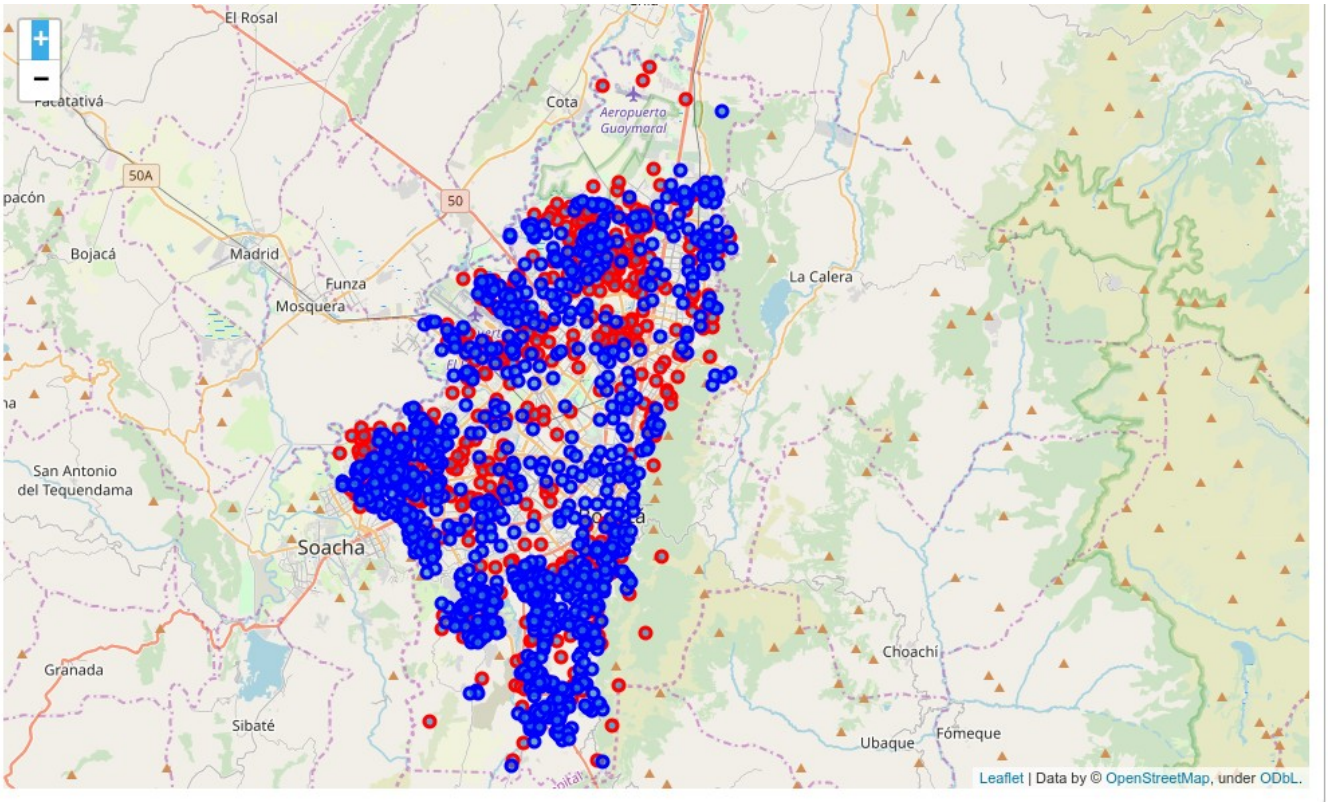
As a first insight of the data its the total number of neighborhoods in Bogota and its classification in terms of its legal status. First its to notice that the total neighborhoods in Bogota according to their local authorizes and the data they posses is 3833 in total of which 2032 are non legalized and 1801 are legalized. It is notice that some of those neighboords are actually new buildings and edifications, not a neighborhood in the common sense of the word.

Lets see how is the geographical distribution of the neighborhoods in Bogota.



It is to notice the existence of a large concentration of neighborhoods in the peripheral regions of the city, those areas correspond in large amount a residential neighborhood meanwhile the central areas of the city corresponds to a office or neighborhood areas.

Now lest take a look at the distribution of some of the legalized neighborhoods vs the non legalized ones:

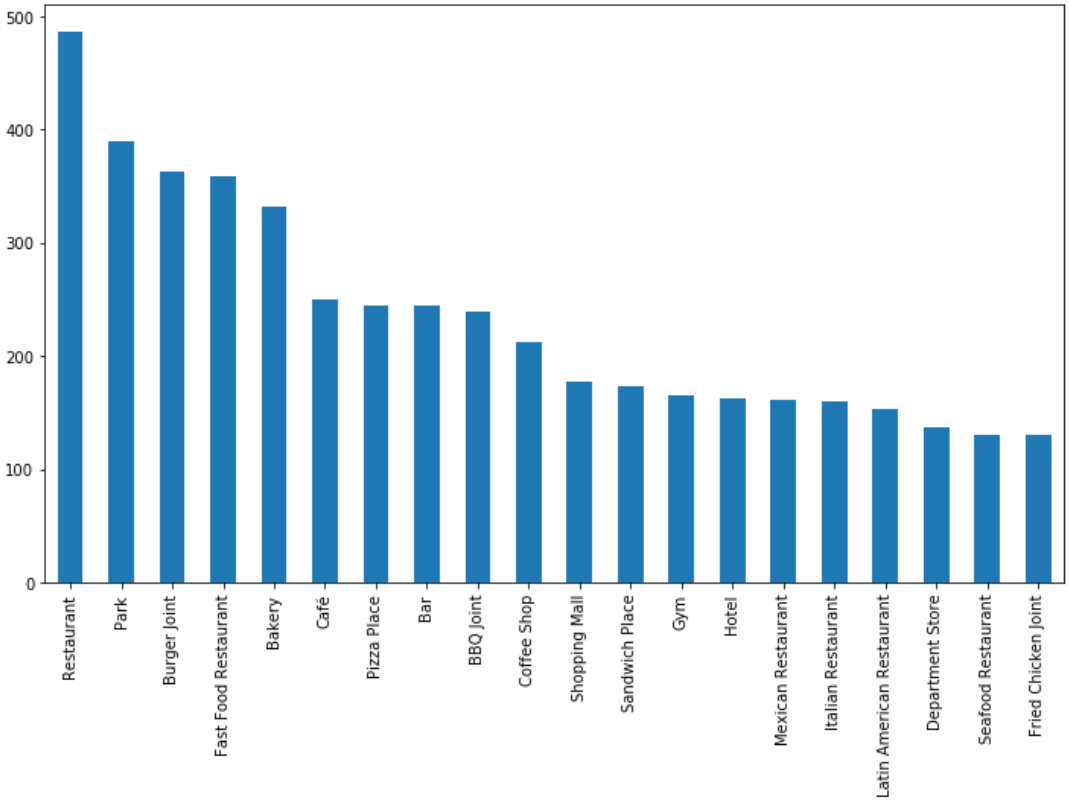


In the previous map the legalized neighborhoods are colored in blue and the non legalized in red. As we can see most of the non legalized neighborhoods are close to the legalized one. Thus some of those are new localizations are constructed in new places near the legalized neighborhoods.

Not all those new neighborhoods are located in the peripherical areas of the city as one might think due to invasion of empty areas of the city that most of the times are located on those areas.

Foursquare analysis.

After the venues of the legalized neighborhoods are extracted using the foursquare API, the first thing to notice is what kind of venues are the most frequent in the city. This is done by counting the number times a categories is present in the dataset.



As we can see the most common venue registered in the foursquare API for Bogota are the restaurants followed by parks, burger joins and so on.

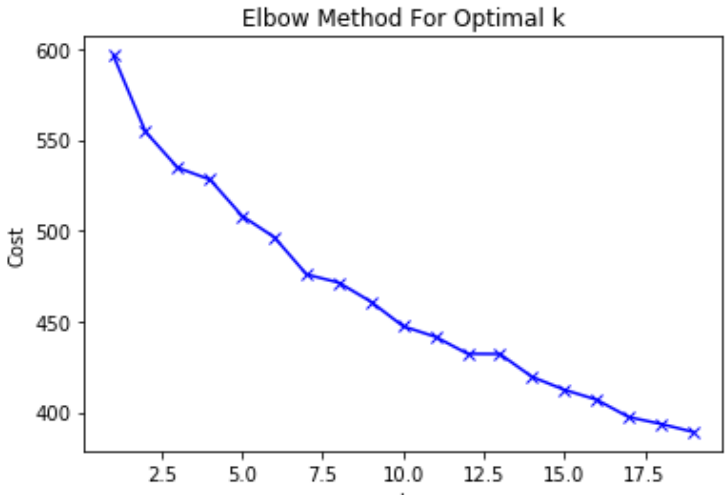
Its is possible to determine what are the most frequent venues for each neighborhood in the city. Some of that information in provided in the next table.

	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	11 de Noviembre	BBQ Joint	Fast Food Restaurant	Latin American Restaurant	Donut Shop	Restaurant	Furniture / Home Store	Farmers Market	Motorcycle Shop	South American Restaurant	Fish & Chips Shop
1	12 de Octubre	BBQ Joint	Fast Food Restaurant	Latin American Restaurant	Donut Shop	Restaurant	Furniture / Home Store	Farmers Market	Movie Theater	South American Restaurant	Fish & Chips Shop
2	8 de Diciembre	Non-Profit	Yoga Studio	Eastern European Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Falafel Restaurant
3	Abraham Lincoln	Gift Shop	Café	Breakfast Spot	Shopping Mall	Yoga Studio	Factory	Event Service	Event Space	Exhibit	Fabric Shop
4	Acacia III Parte Baja	Motorcycle Shop	Eastern European Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Falafel Restaurant	Farmers Market

In this way it is possible to set an study of what kind of venue is available for each neighborhood, and what kind of venues makes sense to open in a specified location.

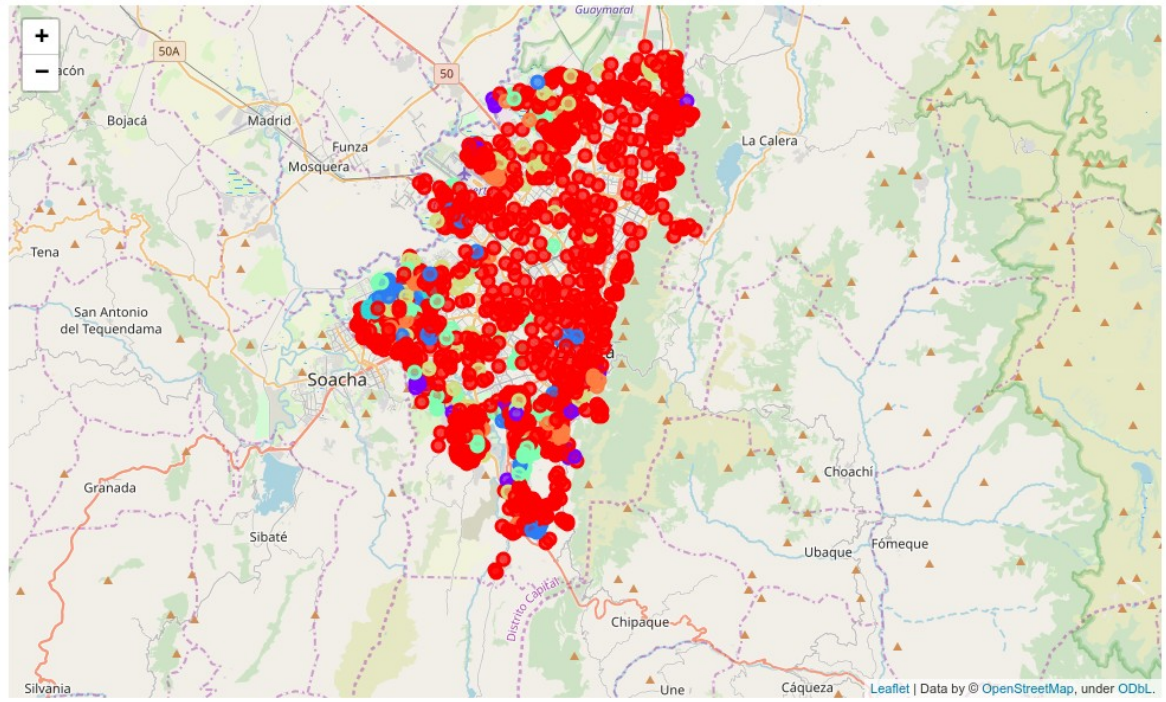
Clustering the neighborhoods using K-mean algorithm

This first step is determine the best value of k to use in the algorithm. For that the cost function is calculated for different values of k. The following plot shows the results.



The elbow methods tell us that k must be select in the inflection point where the cost function tends to decrease in major value. In this case the selected k is 7. So now its possible to implement the clustering algorithm.

A geographical approach with the defined clusters show us how the city of Bogota is distributed in terms of the Foursquare registered venues.



The first thing to notice is the existence of one super cluster (red) which contains most of the row of the registered venues. These cluster is formed by a variety of venues as restuarants, hostels, hotels, bakeries, candy store and so on. This cluster tell us that major part, Bogota is not a segmented city, at least in the major area.

This not implies that there in not cluster of neighborhoods that share similarities in their venues, for example we can see that the cluster with label 1, in purple its conformed by the most part for construction and landscaping venues, as its show in detail when the resulting dataset is classified by cluster label.

	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
34	Sierra Morena	Construction & Landscaping	Coffee Shop	Falafel Restaurant	Electronics Store	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory
62	Urb. Colmena I Los Pinares	Construction & Landscaping	Print Shop	Cosmetics Shop	Coworking Space	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory
78	San Blass II Sector	Construction & Landscaping	Playground	Falafel Restaurant	Electronics Store	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory
82	Granjas y Huertas El Ramajal	Construction & Landscaping	Playground	Falafel Restaurant	Electronics Store	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory
94	Ramírez	Construction & Landscaping	Eastern European Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Falafel Restaurant	Farmers Market
152	Villas de Bolivar	Construction & Landscaping	Electronics Store	Food & Drink Shop	Food	Flower Shop	Flea Market	Fish Market	Fish & Chips Shop	Fast Food Restaurant	Eastern European Restaurant
296	Santa Cecilia I	Construction & Landscaping	Diner	Falafel Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Farmers Market
316	Buenos_Aires	Construction & Landscaping	Museum	Falafel Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Farmers Market
324	La Fortaleza	Construction & Landscaping	Soccer Field	Eastern European Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Falafel Restaurant
374	Marco Fidel Suárez	Construction & Landscaping	Massage Studio	Falafel Restaurant	Empanada Restaurant	Event Service	Event Space	Exhibit	Fabric Shop	Factory	Farmers Market

So, in Bogota in the major part of the city doesn't exist a defined neighborhoods by type of venue. Still some little clusters are formed in some specific areas of the city. This is a sign of the diversity of venues and its locations across the city.

5. Conclusions

One of the great conclusions of this study is that the non legalized neighborhoods of the city of Bogota are not found exclusively in the pernicious areas. These non-legalized neighborhoods are also made up of new construction areas near the traditional neighborhoods of the city.

When an analysis of the city sites is carried out, the most frequent sites that are registered on the foursquare site are sites related to the city's tourist activity such as restaurants, parks and shopping centers.

It is also possible to observe that there are small groupings of neighborhoods that share similarities based on the category of registered venues, but it is not a generality in the city.

This study is useful for those interested in analyzing the distribution of the most frequent sites in the city, based on the foursquare API and can be used to determine in which sites it is more useful to open a new commerce according to more and less sites Frequently analyzed in this project.

6. Perspectives and future work

As saw in the project, the respective analysis of the non-legalized neighborhood is left for future inclusion. This analysis include the study of the venues that exist in those and how are distributed.
Only the information of foursquare is included, to improve the quality of the results some local information should be added. Information of the local government for example. This addition may ended in better and more precisely outcomes. Also a comparative study of those two sets could be done to infer some conclusions.