Data Science Capstone
IBM Data Science Professional Certificate
May 24, 2020

Segmenting the Best Places to Live in Bogota - Colombia

According to an economic, social and life quality approach

1. Introduction

1.1.Background

Bogotá is the capital of Colombia and the administrative and economic center of the country. This city is divided into 20 boroughs, 112 zones planning units, 1169 cadastral areas and 1048 neighborhoods .

Each area of the city is classified depending on the characteristics of the houses, the urban environment of the area, and the urban context. Thus, the city is subdivided into 6 socio-economic sectors, which the number 1 being the lowest and 6 the highest. To identify areas of action and distribute the cost of public services, where the higher sectors subsidize the lower ones and these, in turn, can access education or health benefits given the classification. Consequently, each citizen has to pay the cost of public services according to a sector, the number 6 pays more than 1. All this results in a social classification of rich and poor people. This has allowed the city to quickly identify vulnerable sectors and, among other things, has managed to guarantee free minimum vital water consumption to sectors 1 and 2.

The real estate market encompasses all these factors in order to provide a price and quality of life advantages that should be analyzed when choosing a home.

1.2.Problem

A married couple lives in the suburbs of Bogota and every day they take 2 hours per way to arrive at home or work. Bogota has a big problem with its traffic, there are a lot of cars and

public transportations is not good at all as an incentive to take it. Because of this, they decide to move to Bogota under the following conditions:

- The new place has to be close to their jobs. It is close to Parque de la 93
- Only they can afford up to sector 4.
- They do not want to live in sectors 1 and 2.
- It should be a supermarket, pharmacies, and bus stations close to the new place.

This project aims them to choose the best place to live according to their rules and based on Foursquare and governmental data from Bogota.

1.3.Interests

The married couple is the main interest, but it is also a good example for all citizen who decide to move into Bogota.

2. Data Acquisition and Cleaning

2.1.Data sources

First of all, this project has to be divided into two frameworks:

1. Socioeconomic Data
2. Life quality Data

The first item will be acquired from Colombia's Open Data Portal (*Datos Abiertos del Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia - MinTIC -* www.datos.gov.co). It is an open-source delivery tool to publish in a unified manner, all the dataset produced by the public entities of Colombia, in an open format so that they can be used freely and without restrictions by any person to develop applications or value-added services, make analysis and research, exercise control or for any type of commercial or non-commercial activity.

The second item will be acquired from Foursquare firmographic data using its API. Foursquare is a social networking service available for smartphones. Its purpose is to help to discover and share information about businesses and venues around a defined geographical point. It is important to get the venues around different boroughs so in this project the venue details will be used in a regular endpoint approach.

2.1.Evaluation and Preliminary Recognition

The MinTIC's Open Data Portal has datasets about geographical information of Bogota in diverse formats. A GeoPackage format was chosen because it is a good specification to describe several types of datasets in just one database. A new GeoPackage file was built from the layers: classification sectors, areas, boroughs, blocks, and neighborhoods.

To get the final dataset, a Venn Diagram can summary this merge process

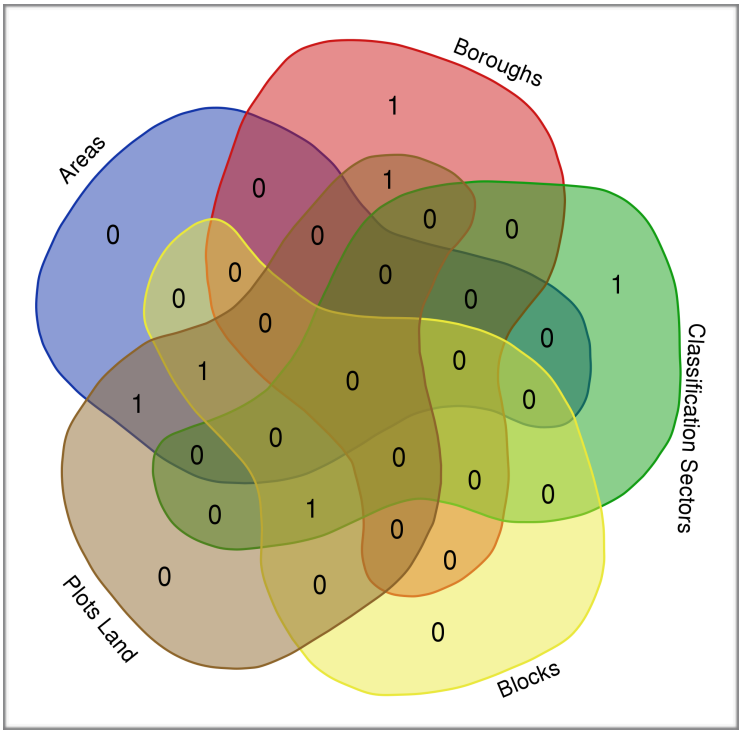

Figure 1. Venn Diagram to create the final dataset

Table 1. Intersection Venn Diagram

Names	Total	Elements
Areas Blocks Plots Land	1	NeighborhoodCode
Blocks Classification Sectors Plots Land	1	BlockCode
Areas Plots Land	1	NeighborhoodName
Boroughs Plots Land	1	LocCode
Boroughs	1	LocName
Classification Sectors	1	ClassSector

To obtain the data group *Plots Lands*, it was necessary to extract the information of the boroughs and the blocks of codes from a code that assigns the district of 30 digits, this allowed to generate the intersection that was necessary according to the Venn Diagram.

The different datasets layers obtained has code fields, decrees and laws, and other information that was not relevant to this project since all that is required is classification sectors, neighborhood, boroughs and geographical location.

2.2.Data Cleaning

All the dataset was defined by:

- Keep codes and names from places
- Keep information geographical to show in a map.

The remaining columns, in turn as mentioned in the previous section, were eliminated. The remaining columns, in turn as mentioned in the previous section, were eliminated. However, when preliminarily plotting the Bogota's map to see the status of the data, (see figure 2) it was found that some points were not in their correct location because it was in another borough, then should be debugged, these outliers were reviewed one by one as it is possible that there was some error in the merging process or the original database is like that.

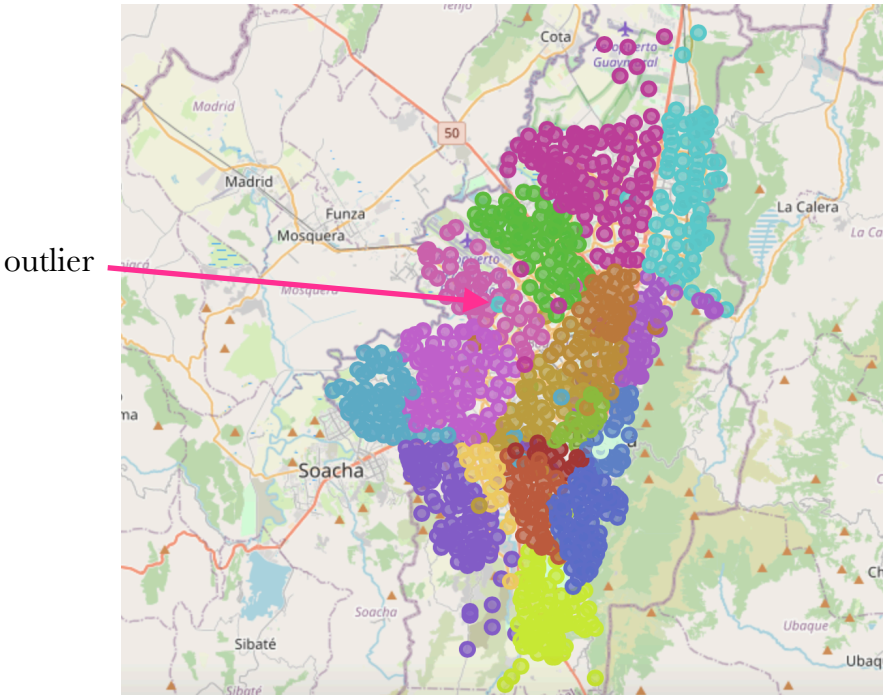

Figure 2. Outlier in the Bogota's map

When these outliers were checked, it was found there several neighborhoods with the same name, although the merge process was with the codes, not the names, when the dissolving process was done the code field was eliminated, so this code had to be kept.

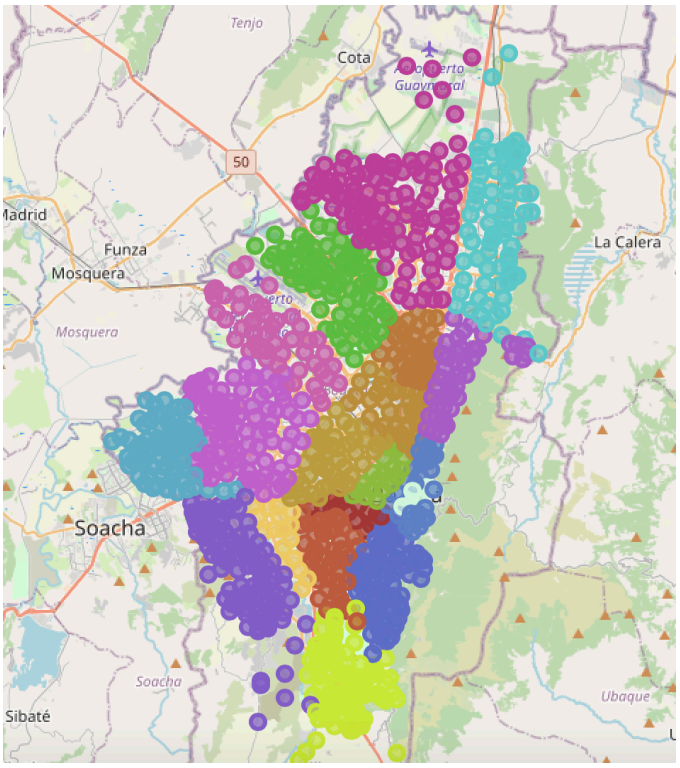

Figure 3. Without outliers in the Bogota's map

Finally, the dataset to use in this project is:

Table 2. Head Final Dataset, rows 43904

index	BlockCode	Neighborhood	Borough	Borough Code	Class Sector	SHAPE Leng	SHAPE Area	geometry
0	8520013	BOSQUE DE PINOS	USAQUEN	1	0	0.001817	9.897130E-08	POLYGON ((-74.02287 4.73709, -74.02288 4.73709...
1	8520014	BOSQUE DE PINOS	USAQUEN	1	0	0.001326	8.577165E-08	POLYGON ((-74.02198 4.73747, -74.02207 4.73732...
2	8520017	BOSQUE DE PINOS	USAQUEN	1	2	0.001332	9.888299E-08	POLYGON ((-74.02196 4.73690, -74.02198 4.73686...
3	8520022	BOSQUE DE PINOS	USAQUEN	1	2	0.001982	1.521141E-07	POLYGON ((-74.02151 4.73493, -74.02155 4.73486...
4	8520026	BOSQUE DE PINOS	USAQUEN	1	2	0.001365	8.470870E-08	POLYGON ((-74.02028 4.73527, -74.02024 4.73526...

The final dataset will be able to use to get the neighborhoods using the dissolve method.

Table 3. Head Neighborhoods, rows 1048

NeighborhoodCode	Neighborhood	Borough	SHAPE Leng	SHAPE Area	Latitude	Longitude
1101	LAS BRISAS	SAN CRISTOBAL	0.008383	2.786418E-06	5	-74.081617
1102	BUENOS AIRES	SAN CRISTOBAL	0.004245	7.403113E-07	5	-74.080557
1103	VITELMA	SAN CRISTOBAL	0.003767	2.825050E-07	5	-74.077010
1104	MOLINOS DE ORIENTE	SAN CRISTOBAL	0.003773	3.784194E-07	5	-74.069386
1106	SAN BLAS	SAN CRISTOBAL	0.002653	2.349225E-07	5	-74.083557

2.3.API Foursquare

For illustration purposes, let's simplify the above data and segment only the neighborhoods in *Usaquen* where the geographical coordinates of *Usaquen* are 4.694969, – 74.0310933.

This is just an example, but in order to choose the borough and know the radius to make the query into Foursquare, it needs to review the point where the couple's work is located and propose the radius of mobilization between the house and the workplace.

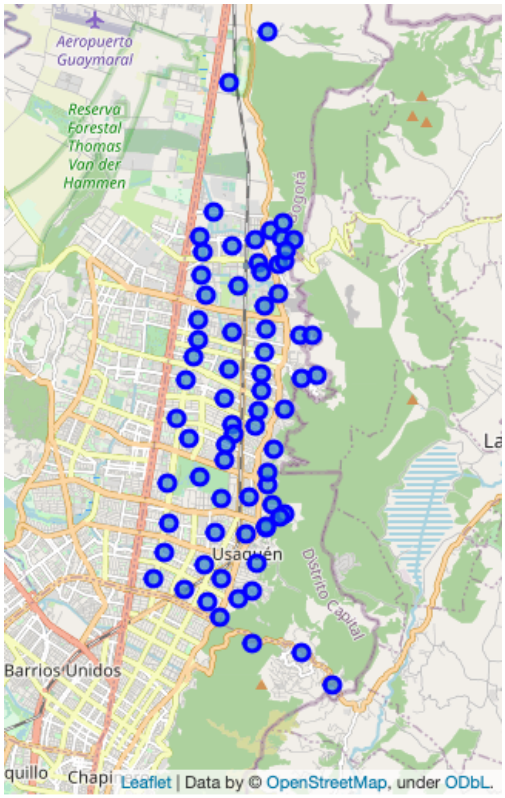

Figure 4. Neighborhoods in USAQUEN

As an example, SANTA ANA OCCIDENTAL will be taken:

Table 4. Foursquare Data

index	name	categories	lat	lng
0	Crepes & Waffles	French Restaurant	4.691003	-74.036620
1	Café de la montaña	Café	4.688961	-74.038279
2	Carulla Santa Ana	Supermarket	4.691209	-74.038176
3	Tiendas Jumbo Santa Ana	Grocery Store	4.690623	-74.037221
4	Parque Santa Ana	Park	4.687358	-74.037176

2.4.Data Selection

The workplace is close to a venue site in Bogota called *Parque de la 93* and will be established as the search centre with a radius of 6km.

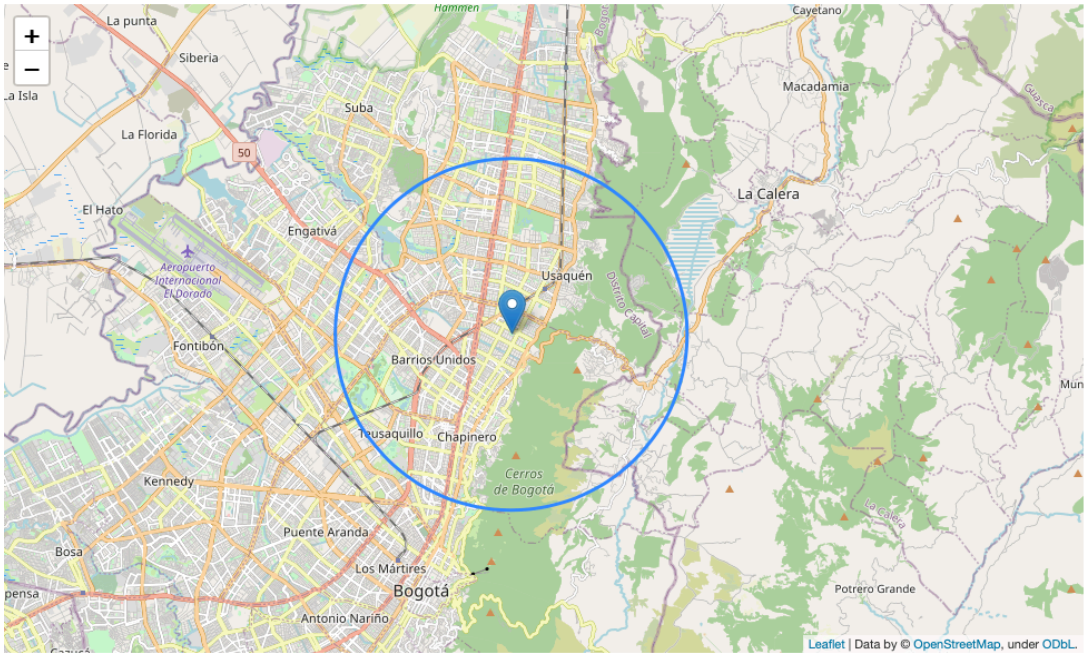

Figure 5. Search perimeter

With this perimeter will make it possible to evaluate the conditions described above as the "rules" to be able to select the neighborhood and then its closest venues. This will be done in the next section of exploratory analysis.

3. Exploratory Data Analysis

3.1.Methodology to find the best

The problem said that a perimeter should be selected where the new home can be selected and that it should be as close as possible to the workplace. Bearing in mind that it is in the Parque de la 93, a radius of 6 km will be selected.

First, using this perimeter, the blocks will be filtered to obtain the points to be studied. Then, another filtering can be done using the other condition: only classification sectors 3 and 4 can be selected.

Before classifying each point, it is important to survey the stakeholders about what type of venues they would like to have near their future home. This will allow the creation of a table of categories to search for at each point.

With this information obtained manually and with the dataset filtered, the Foursquare database will be used to consult all of the venues that each center in each block has within a radius of 500 meters, this is because this distance can be considered adequate for walking.

Finally, a map will be presented highlighting each cluster created with the *k-means* technique to classify each point from its classification sectors and nearby locations as well as a small description of what can be found at each location.

3.2.Datasets Building

From the search perimeter of Figure 5, a set intersection is made in order to limit the amount of data, this will allow to optimize the search of the sites of interest. It is ensured that the resulting distribution contains classification blocks 3 and 4.

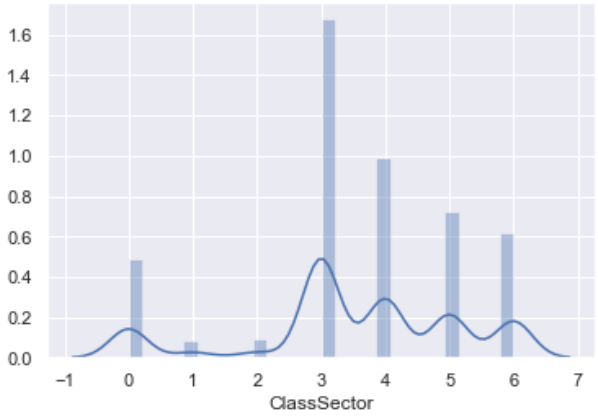

Figure 6. Sectors Classification Distribution

A map can better visualize this filtering process:

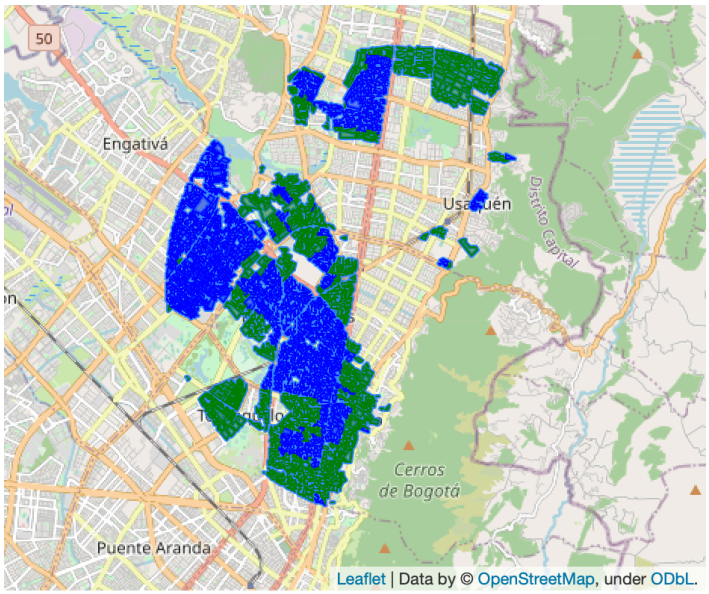

Figure 6. Filtered dataset displayed on the map

This process consists of using the overlay process: "When working with multiple spatial datasets – especially multiple polygon or line datasets – users often wish to create new shapes based on places where those datasets overlap (or don't overlap). These manipulations are often referred using the language of sets – intersections, unions, and differences. These types of operations are made available in the geopandas library through the `overlay` function"[1]. This allowed the dataset to be reduced to 3339 rows, and the Foursquare database can be reviewed for each venue.

3.3.Categories Selection

Foursquare contains a database with a variety of venue categories. This could be counterproductive because the clusters would be overfitted. To do this, skateboarders are surveyed with the categories they would be interested in nearby, for example it is not as important

[1] Official site from GeoPandas: https://geopandas.org/set_operations.html

to have a car store close to a supermarket. The output of this process was a data set of 100 elements, the ID is maintained to make the selection process more optimal.

Table 5. Head Categories Type, 100 rows

Name	Id	Type
General Entertainment	4bf58dd8d48988d1f1931735	Arts Entertainment
American Restaurant	4bf58dd8d48988d14e941735	To Eat
Asian Restaurant	4bf58dd8d48988d142941735	To Eat
BBQ Joint	4bf58dd8d48988d1df931735	To Eat
Bakery	4bf58dd8d48988d16a941735	To Eat

In order not to have the 100 categories, a manual process of grouping is made to finally obtain 9 categories: *Arts Entertainment, Food Drink Shop, Health Services, NightLife, Outdoors, Professional Services, Public Transportation, Shop and Services* and *To Eat.*

3.4.Data exploration

With each geographical point obtained from Section 3.2, a search of the venues is carried out with a radius of 500 meters since this is a distance considered to be covered on foot. It is filtered by category types and a dataset is obtained with the number of sites per category at each point. This query is classified, according to Foursquare, as ordinary. This query should have been done with moderation because there is a limit to the number of hours that can be spent on this type of query and this dataset contains 3339 items. Finally the complete dataset was building to perform the clustering process.

4. Clustering

All the data needs to be grouped according to their common characteristics, such as the sector of classification, how many places to eat in the vicinity, how many grocery stores are in the vicinity, etc. For this purpose, it was considered more convenient to use the k-means algorithm because it allows measuring that similarity between points for the unsupervised data. The items that will be used for the classification are: 'ClassSector', 'Arts_Entertainment', 'Food_Drink_Shop', 'Health_Services', 'NightLife', 'Outdoors', 'Professional_Services', 'Public_Transportation', 'Shop_and_Services' and 'To_Eat'.

A critical point in this project was the choice of the number of clusters because there is no starting point to establish a selection criteria. Several quantities were chosen and it was analyzed how each cluster described its services and try that there were not so many "almost" common characteristics between data. In the end, five groups were selected to also allow the stakeholders to have a choice.

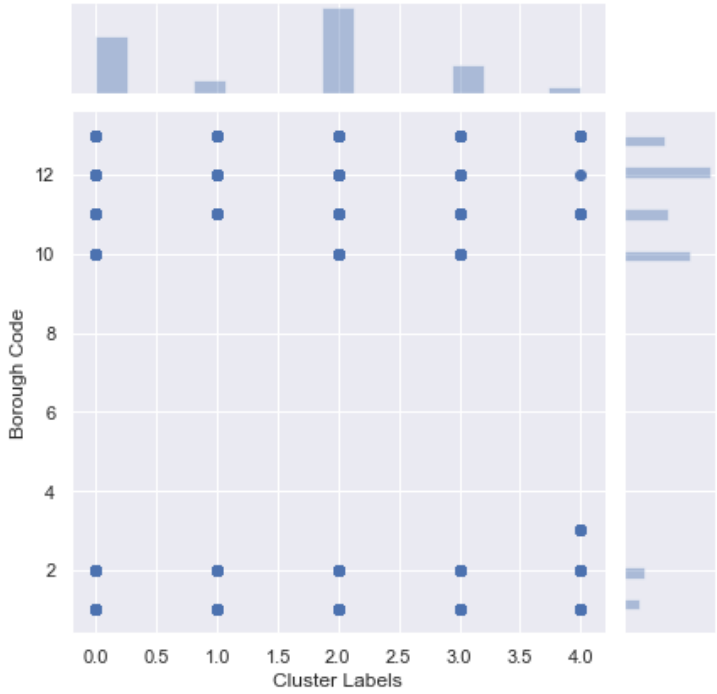

Figure 7. Cluster Labels - Borough Code Distribution

Figure 6 illustrates that cluster 2 has the most blocks and borough 12 has the most variety of clusters.

To better visualize the data, a bar chart will be made but it is necessary to normalize the data using the maximum. Thus, the different services of each cluster can be compared.

Table 5. Normalized mean data of each cluster

Cluster	Arts Entertainment	Food Drink Shop	Health Services	Night Life	Outdoors	Professional Services	Public Transportation	Shop and Services	To Eat
0	0.666470	0.593852	0.0	0.088313	0.383231	0.183426	0.083138	0.066958	0.141254
1	1.000000	0.859930	0.0	0.745323	0.252241	0.686722	0.000000	0.904919	0.615705
2	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
3	0.537693	0.845650	0.0	0.263595	0.490323	0.558619	0.000000	0.417273	0.340280
4	0.649935	1.000000	0.0	1.000000	1.000000	1.000000	0.000000	1.000000	1.000000

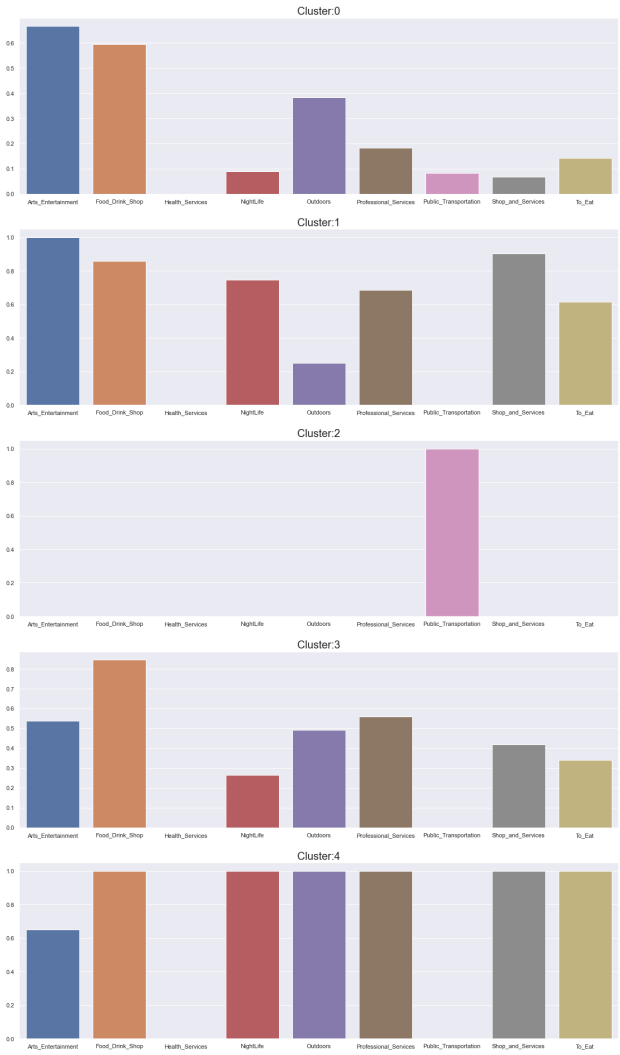

Figure 8. Bar-graphs - Normalized mean data of each cluster

4.1.Geographical Location of Clusters

One of the best ways to present the findings to stakeholders is by highlighting the clusters on the map of Bogotá. They were grouped by type for better visualization.

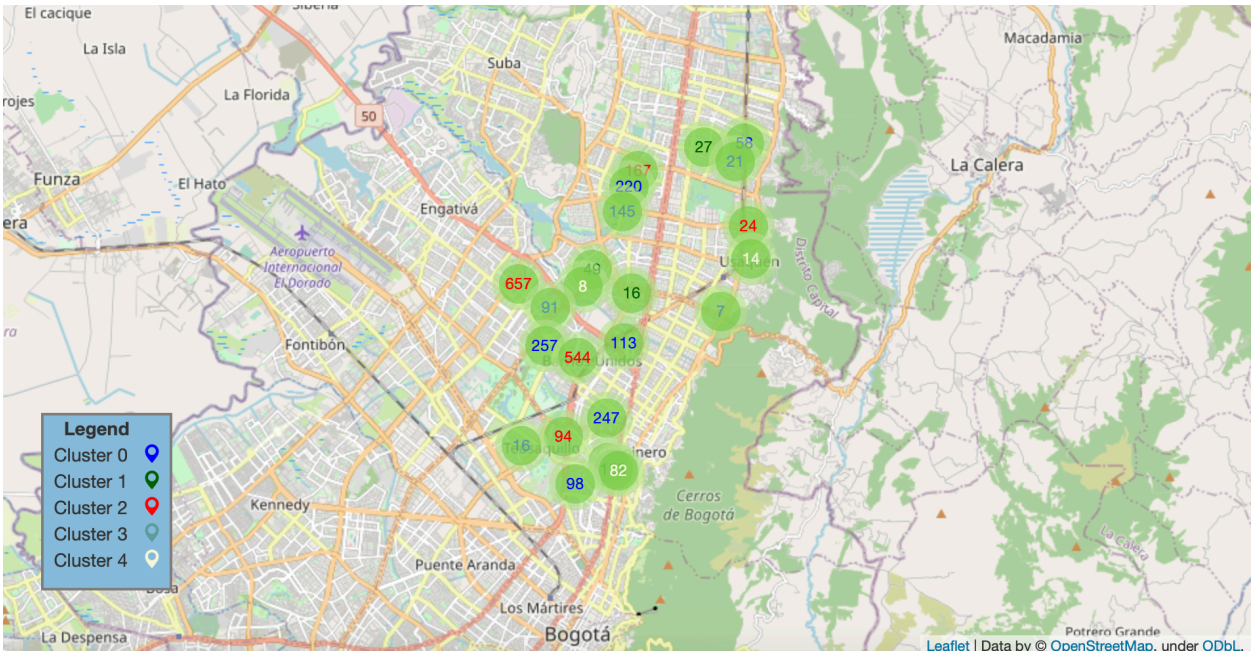

Figure 9. Geographical Location of Clusters

4.2.World Clouds

Although the map gives a global vision of the results and with the support of the bar graphs it allows to generate an idea of the venues of each area, it is still not possible to know which are the neighborhoods in which they could be consulted to rent. When looking for an apartment to rent it is always easier to do so using the neighborhood. The problem is that there are approximately 130 neighborhoods and a two-dimensional graph would not be useful. A wordcloud was used to do this as it allows the names of the neighborhoods to be displayed and highlights which ones have the most points within each cluster.

Figure 10. Neighborhood Wordcloud

4.3.Conclusion and Future directions

Bogotá has countless neighborhoods where each one offers a variety of services, although each one will depend on the purchasing power of each person. This analysis shows that within a radius of 6km with the center of Parque de la 93 there are different points suitable for living according to the rules of the couple. Bogotá offers a better quality of life towards the north of the city, and this can be seen in the results. Usaquén is one of the places with the highest number of venues, which allows for a better quality of life when choosing a home.

On the other hand, it is important to emphasize that none of the selected points do have medical services nearby, this is because the classification of each sector is being filtered.In addition, the proximity to supermarkets, shops and restaurants is sacrificed for the proximity to bus stations, so this could not meet the initial requirements and would have to rethink that initial ideal.

In this analysis, the intention was not to provide the married couple with a place to live since this implies other factors such as availability of rental housing, price, size, among others. However, we would suggest cluster one because it is the one with the most services and also because its nightlife is not so dominant and this allows for a more peaceful life.

For future work we could have a database of prices available for renting, this could give a greater spectrum of analysis of the areas.

The purpose of this project was to identify possible living areas within a radius of 6 km, taking as a center a point close to the couple's workplaces under established rules: sectors classification 3 and 4 and the largest number of venues within 500 meters. Since the socio-economic classification of Bogotá is by blocks and not by neighborhood, this generated a large amount of data to be analyzed. However, by filtering all the blocks in each neighborhood under these conditions, the number of data could be reduced considerably. We checked that the resulting distribution was within the requested limits, finding that the only thing that could not be met was the proximity of the bus stations.

Finally, we used word clouds and bar charts to present the skateboarders with possible solutions to their problem, where they could review considerations if they wanted more night localities nearby because that could be annoying or if it is more important the art sites among others.