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 992 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.
- 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 premium 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 three layers: classification sectors, boroughs, zones planning units, and neighborhoods.

A Venn Diagram can summary this merge process

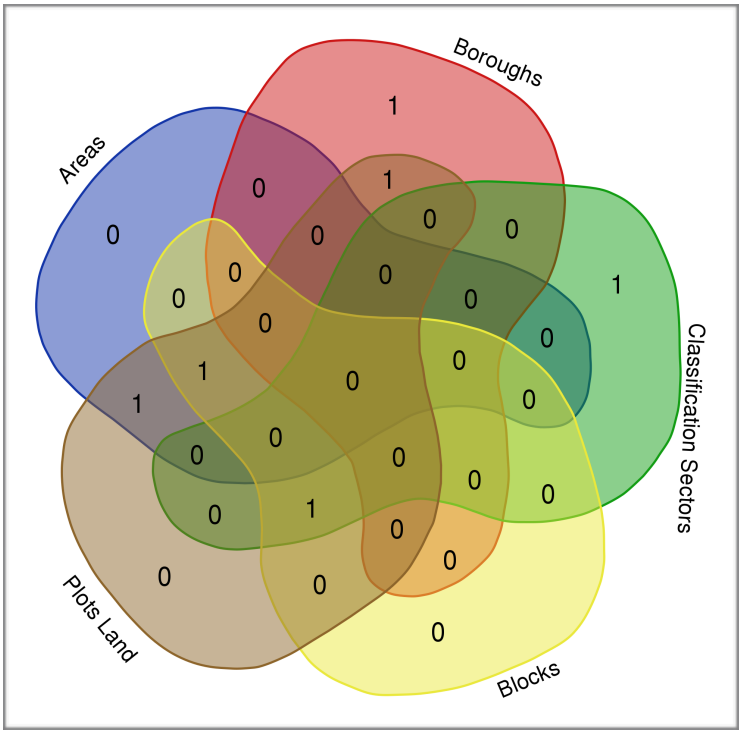


Table 1. Intersection Venn Diagram

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

Two issues were found when the data were first analyzed, the dataset related to the zones planning unis (in Spanish is UPZ), there were 4 regional zones and 2 additional UPZ that are in the process of being recognized by the city authorities. The 4 regional zones can be removed from the dataset as it does not represent valuable information for the analysis of this paper. On the other hand, the dataset of the sector classification is not related by neighborhoods but by blocks. It is important to make a cross-reference of each block with its respective neighborhood in order to classify the problem according to the initial rules.

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

Table 2. Final Dataset, rows 43904

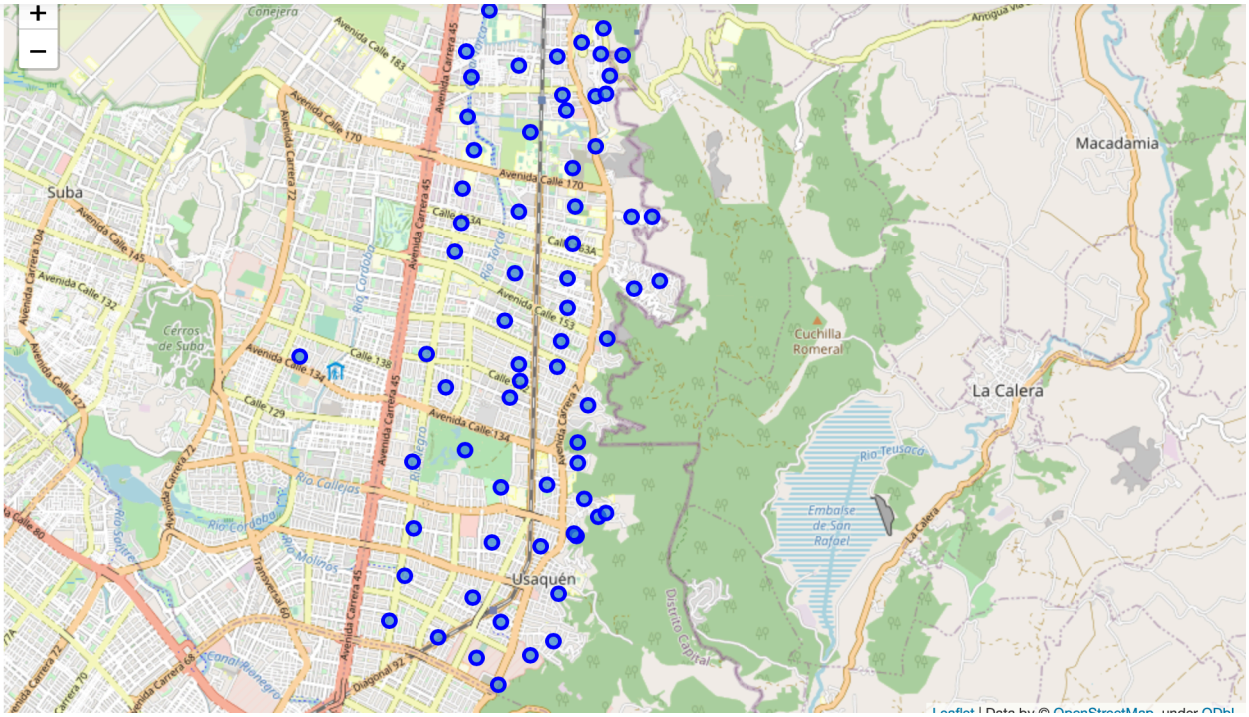
Index	ManCodigo	SCaNombre	LocNombre	LocCodigo	ESTRATO	geometry
0	0	8520013	BOSQUE DE PINOS	USAQUEN	1	POLYGON ((-74.02287 4.73709, -74.02288 4.73709...
1	1	8520014	BOSQUE DE PINOS	USAQUEN	1	POLYGON ((-74.02198 4.73747, -74.02207 4.73732...
2	2	8520017	BOSQUE DE PINOS	USAQUEN	1	POLYGON ((-74.02196 4.73690, -74.02198 4.73686...
3	3	8520022	BOSQUE DE PINOS	USAQUEN	1	POLYGON ((-74.02151 4.73493, -74.02155 4.73486...
4	4	8520026	BOSQUE DE PINOS	USAQUEN	1	POLYGON ((-74.02028 4.73527, -74.02024 4.73526...

Table 3. Neighborhoods Dataset, rows 992

SCaNombre	LocNombre	SHAPE_Leng	SHAPE_Area	Latitude	Longitude
ABRAHAM LINCOLN	TUNJUELITO	0.001989	2.290245E-07	4.557500	-74.123840
ACACIAS USAQUEN	USAQUEN	0.003574	7.450900E-07	4.723630	-74.034378
ACEVEDO TEJADA	TEUSAQUILLO	0.002901	3.985020E-07	4.630761	-74.081533
AEROPUERTO EL DORADO	FONTIBON	0.002593	3.930862E-07	4.698717	-74.142727
AGUAS CLARAS	SAN CRISTOBAL	0.001283	9.113985E-08	4.557141	-74.069099

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.

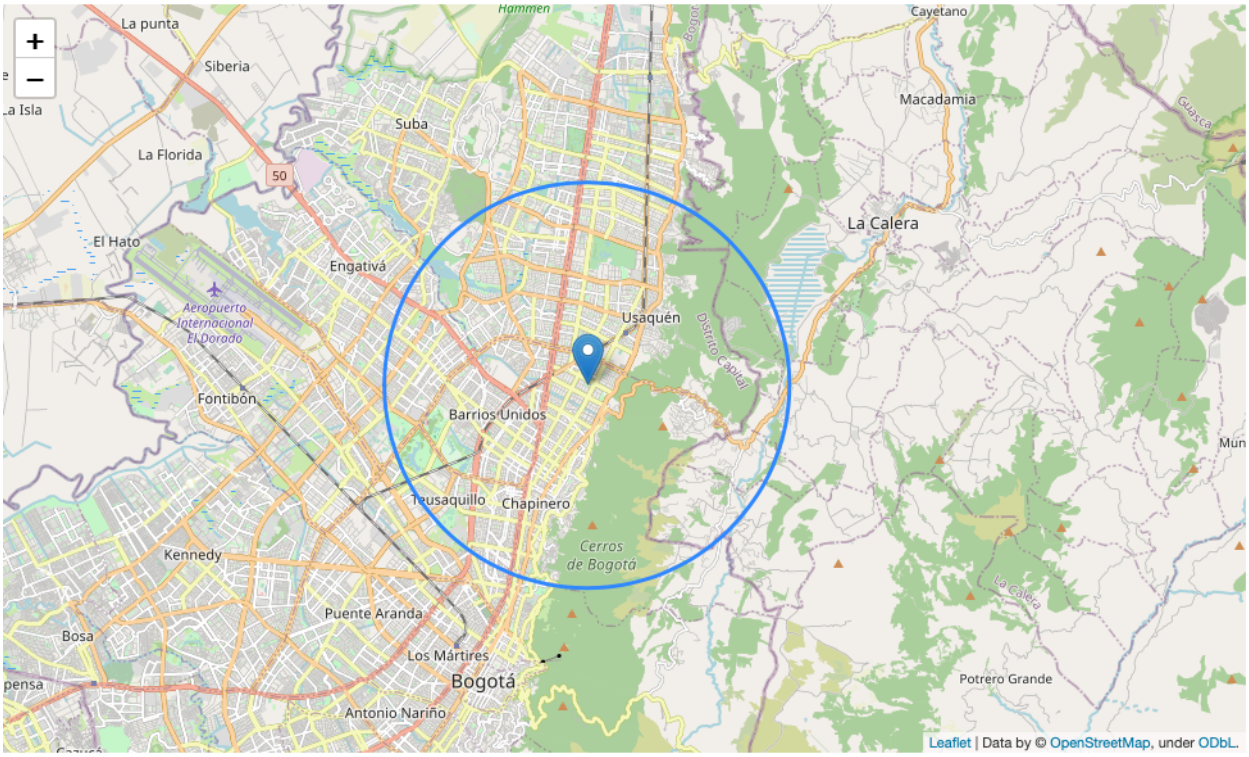
As an example, SANTA ANA OCCIDENTAL will be taken:

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.



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.