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 988 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.

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.

Table 1. Boroughs Dataset, 20 rows

index	LocNombre	LocCodigo	SHAPE_Leng	SHAPE_Area	geometry
0	ANTONIO NARIÑO	15	0.108973	0.000397	POLYGON ((-74.13075 4.59335, -74.12917 4.59327...
1	TUNJUELITO	6	0.210542	0.000807	POLYGON ((-74.13777 4.59489, -74.13165 4.59363...
2	RAFAEL URIBE URIBE	18	0.174513	0.001126	POLYGON ((-74.12803 4.59254, -74.12777 4.59233...
3	CANDELARIA	17	0.067158	0.000168	POLYGON ((-74.06621 4.60317, -74.06620 4.60317...
4	BARRIOS UNIDOS	12	0.121180	0.000969	POLYGON ((-74.05725 4.68684, -74.06249 4.65594...

Table 2. Cadastral areas datasets, 1169 rows

index	SCaTipo	SCaNombre	SHAPE_Leng	SHAPE_Area	geometry
0	0	GINEBRA	0.015404	0.000012	POLYGON ((-74.02441 4.70958, -74.02473 4.70923...
1	0	ATENAS	0.026614	0.000022	POLYGON ((-74.07302 4.73279, -74.07295 4.73278...
2	0	LOS CEDROS ORIENTAL	0.016383	0.000010	POLYGON ((-74.03228 4.72542, -74.03143 4.72505...
3	0	CEDRO NARVAEZ	0.033577	0.000024	POLYGON ((-74.03254 4.72139, -74.03239 4.72134...
4	0	SALAZAR USME	0.006979	0.000002	POLYGON ((-74.11391 4.51906, -74.11387 4.51894...

Table 3. Zone planning lists dataset, 118 rows

index	UPlCodigo	UPlNombre	SHAPE_Leng	SHAPE_Area	geometry
0	UPZ100	GALERIAS	0.060563	0.000193	POLYGON ((-74.06463 4.64965, -74.06619 4.64146...
1	UPZ83	LAS MARGARITAS	0.058093	0.000120	POLYGON ((-74.17077 4.62848, -74.17141 4.62811...
2	UPZ107	QUINTA PAREDES	0.059576	0.000142	POLYGON ((-74.09245 4.63648, -74.09184 4.63569...
3	UPZ101	TEUSAQUILLO	0.063565	0.000192	POLYGON ((-74.06707 4.63644, -74.06840 4.62882...
4	UPZ91	SAGRADO CORAZON	0.064912	0.000120	POLYGON ((-74.06602 4.62604, -74.06570 4.62554...

Table 4. Classification sectors, 45051 rows

index	ESTRATO	SHAPE_Leng	SHAPE_Area	geometry
0	0	2203.905471	267773.513997	POLYGON ((-74.03556 4.82078, -74.03767 4.82167...
1	6	6350.896563	889000.149632	POLYGON ((-74.00643 4.81884, -74.00860 4.81886...
2	0	1488.773919	100924.112922	POLYGON ((-74.04749 4.82353, -74.04731 4.82342...
3	6	745.327665	33775.855324	POLYGON ((-74.05559 4.82178, -74.05562 4.82159...
4	0	1122.296445	70820.836179	POLYGON ((-74.03431 4.82171, -74.03140 4.82143...