IBM Data Science Professional Certification Capstone project

Guatemala shopping mall classification

Develop by: Engr. Javier Navarro

Guatemala, april 10, 2020

Table of contents

1 Introduction

Given the current situation of the Covid-19 health emergency in the Guatemalan territory, problems are experienced regarding the supply of basic supplies, one of these problems is the high crowding of people in supermarkets, even though quarantine is in force, this implies great risks for containment of the outbreak, but not in all supermarkets crowds are observed, that is why as a data scientist I have the objective of making a classification by municipalities (regions) in relation to the provision of supermarkets, this to provide a broad paronama for the Correct distribution of citizens, that is, after carrying out the analysis, the areas with the least amount of supermarkets can be identified, which implies a greater probability of crowding..

2 Data

The data to be used is the distribution of municipalities in the department of Guatemala, information provided by the site: https://www.geodatos.net/coordines/guatemala basically it is the triplet <city, latitude, longitude> this data will be preprocessed for validate its integrity and later it will be transformed in such a way that it can allow obtaining information about the nearby supermarkets using the Foursquare API. An example of data provided for a site is the following:

Coordenadas de ciudades de Guatemala

Ciudad	Coordenadas
Ciudad de Guatemala	14.6407204, -90.5132675
Mixco	14.6333303, -90.6063919
Villa Nueva	14.5269403, -90.5875015
Petapa	14.50278, -90.5516663
San Juan Sacatepéquez	14.7188902, -90.6441727
Quetzaltenango	14.8347197, -91.5180588
Villa Canales	14.48139, -90.5316696
Escuintla	14.3050003, -90.7850037

3 Methodology

The methodology that use is the following:
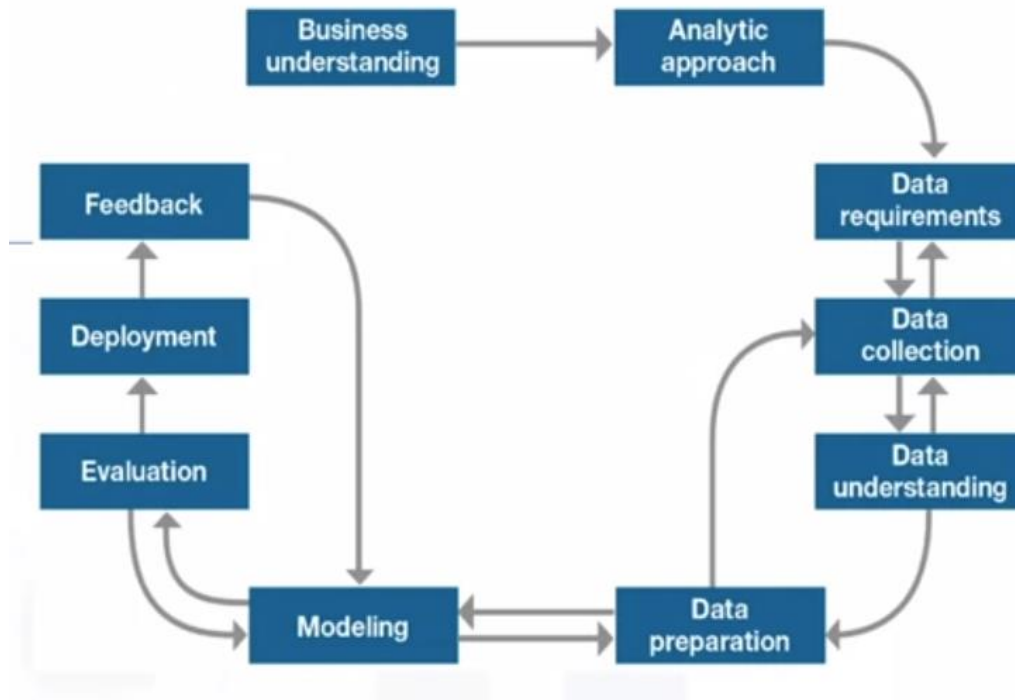
3.1 Business understanding

Getting clarity around the problem to be solved, that is know a neighborhood that is vulnerable, allows you to determine wich data will be used to aswer the core question. In this case, data from https://www.geodatos.net/coordines/guatemala.

3.2 Analytic approach

The question requires a yes/no answer implies use a classification model, k-means is a simpliest choice.
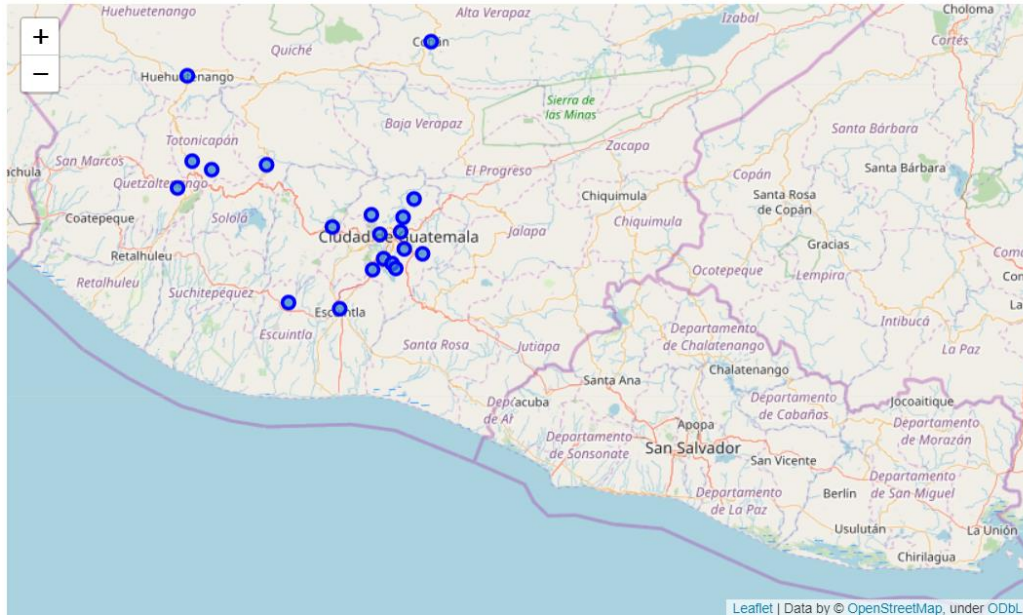
3.3 Data requirements

Data requirements

- Who: users
- What: rating malls or stores
- Where: Guatemala
- How: from your mobile
- Why: share experiences
- When : previusly

3.4 Data collection and understanding data

The content of data is the positions and ratings of users that make possible find nearby places or commonly places:

3.5 Data preparation

Split coordinates column, because atomicity is needed

[21]:		Ciudad	Coordenadas
	0	Ciudad de Guatemala	14.6407204, -90.5132675
	1	Mixco	14.6333303, -90.6063919
	2	Villa Nueva	14.5269403, -90.5875015
	3	Petapa	14.50278, -90.5516663
	4	San Juan Sacatepéquez	14.7188902, -90.6441727

Convert to:

[22]:		Neighborhood	Latitude	Longitude
	0	Ciudad de Guatemala	14.64072	-90.513267
	1	Mixco	14.63333	-90.606392
	2	Villa Nueva	14.52694	-90.587502
	3	Petapa	14.50278	-90.551666
	4	San Juan Sacatepéquez	14.71889	-90.644173

3.6 Modeling

Clustering model for classification, but first:
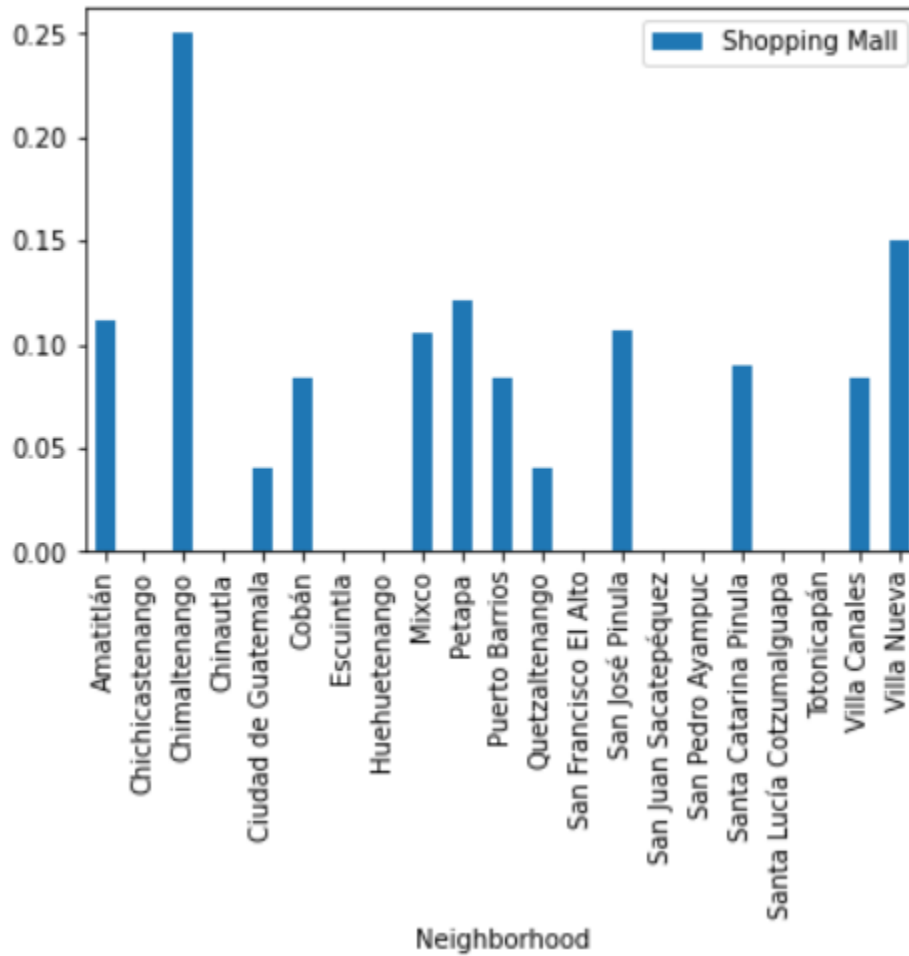
4 Results

Classification neighborhoods to proper cluster using k-means.

	Neighborhood	Latitude	Longitude	Cluster Labels	Shopping Mall
16	Puerto Barrios	15.72778	-88.594437	2	0.083333
17	San Francisco El Alto	14.95000	-91.449997	0	0.000000
18	Cobán	15.47083	-90.370827	2	0.083333
19	San José Pinula	14.54611	-90.411392	2	0.106383
20	San Pedro Ayampuc	14.78667	-90.451111	0	0.000000

Clusters in a map:
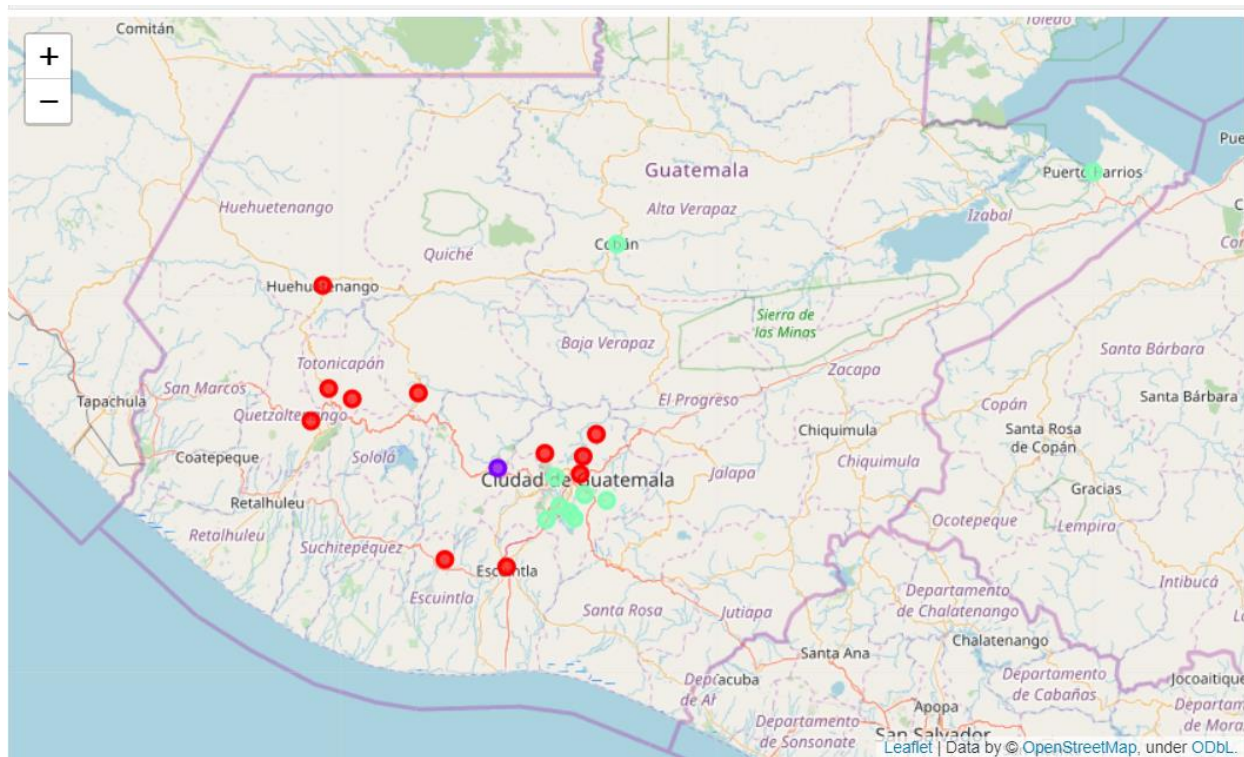
5 Discussion

Before discussing the results obtained, it is important to highlight a couple of considerations. First, it is assumed that the Foursquare API is widely used in my country and that the data obtained provides a broad perspective of various users and is up-to-date. Second, they were only considered shopping centers but they could also be considered supermarkets, as well as neighborhood grocery stores (a small business that is very common in my country).

It is noteworthy that in rural areas, that is, those far from the country's capital, clusters of low number of nearby businesses predominate, so I would recommend implementing government measures to improve the supply system in the countryside.

6 Conclusions

- Three clusters were built that represent the low, intermediate or high prevalence of shopping centers in the most common neighborhoods of the country of Guatemala
- It was determined that the further away from the capital you are, the less there are shopping centers and therefore crowds are induced
- The low clusters identified in the capital correspond to areas marked as "red", a relationship could be established between social status and the agglomeration or ignorance of the rules imposed by the government

7 References

[1] [Forsquare API](#)

[2] [Geodatos](#)