Clustering neighborhoods in Mexico City according to venue type, population density and income level for a CPG retail business

Rogelio Antonio Melo Juarez

Mexico City, Mexico

2019 – 11 – 09

Table of Contents

[1.0 Introduction 3](#_Toc24216583)

[2.0 Data 3](#_Toc24216584)

[2.2 3](#_Toc24216585)

[2.3 4](#_Toc24216586)

[2.4 4](#_Toc24216587)

[2.5 5](#_Toc24216588)

[3.0 Methodology 5](#_Toc24216589)

[3.1 Foursquare data 5](#_Toc24216590)

[3.2 Mexico City's socioeconomic and demographic data 7](#_Toc24216591)

[3.3  Clustering 9](#_Toc24216592)

[4.0 Results: 10](#_Toc24216593)

[5.0 Discussion 14](#_Toc24216594)

[6.0 Conclusion 14](#_Toc24216595)

[7.0 References 14](#_Toc24216596)

## 1.0 Introduction

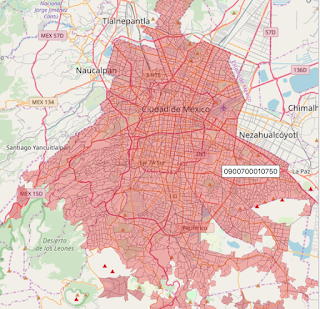
Opening a business in a city with which you're not familiar with can be a tricky decision. However if a product or market is not exploited in this new location, it could represent a profitable opportunity for the people opening it. Leveraging Foursquare data, the objective of this investigation is to cluster neighborhoods in the district of Álvaro Obregón in Mexico City in order to produce a trustworthy guide for people wanting to open a consumer packaged goods retail business in one of Mexico City's richest districts. We will match public geo referenced income data published by the INEGI, which is Mexico's statistics and geography institute, with the information extracted from the Foursquare API to obtain lists of Mexico City's neighborhoods in Álvaro Obregón which are similar to each other.

## 2.0 Data

**2.1** I used Foursquare API's location data for Álvaro Obregón District in Mexico City regarding to the retail venues present in each neighborhood. These could be all types of consumer packaged goods distributors, such as convenience stores, bakeries, butcher's, supermarkets, etc.

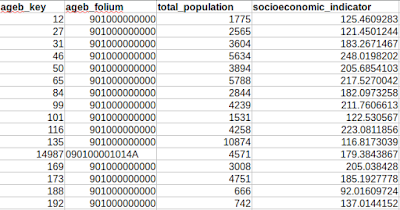
### 2.2

In order to establish a relation with socioeconomic indicators I used a shape file provided by Mexico's Government which segments Mexico City in polygons named AGEBs. Each AGEB has a code which can be matched with several databases containing demographic, economic, social, etc. indicators.



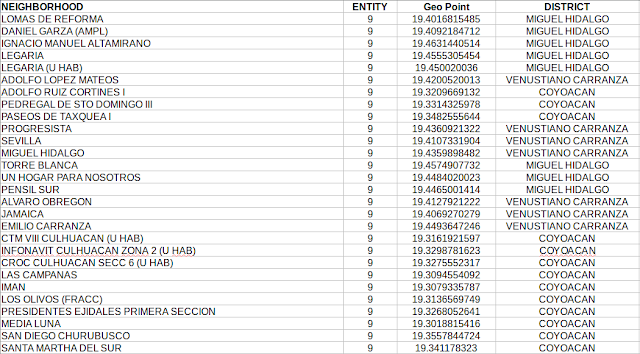
### 2.3

Sample of the data containing each ageb key with its total population and socioeconomic indicator.



### 2.4

I used another database which lists the coordinates of each neighborhood in Mexico City. The next image is a sample of the data containing each of Mexico City's neighborhood with its geo point.



### 2.5

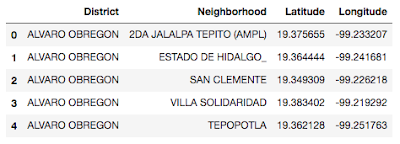
This table contains a merge which includes neighborhood location, venue location and ageb data.



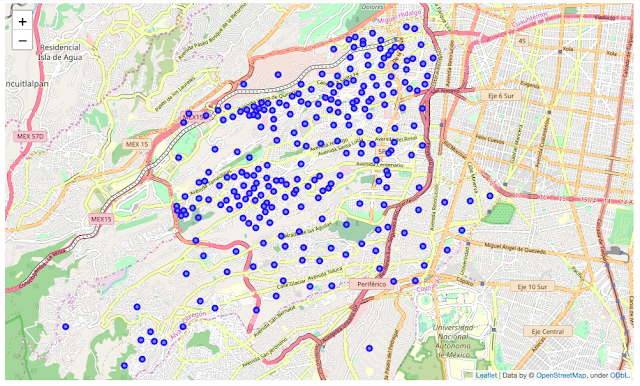
## 3.0 Methodology

### 3.1 Foursquare data

First, I used the Mexico City's neighborhoods table and did a visual analysis of the locations of interest. In my case, I chose to narrow my exercise to the district of Álvaro Obregón. The sample of the table looks as follows.



Afterwards, I used the Folium python library to represent the geo points in the table indicating each neighborhood.



Once I concluded the first visual analysis, I used the Foursquare API to get the venues located in each neighborhood. I defined the limit of results to **100 venues**and at a distance not greater than **500** **meters** of each geo point.

The following dataset containing the venue's latitude and longitude data, the venue category and name and the neighborhood in which it is located.



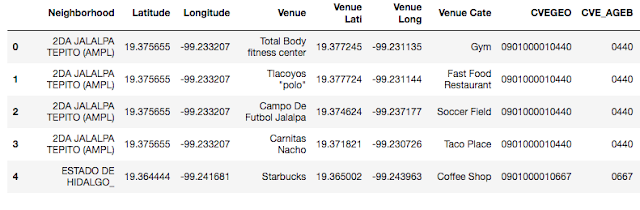
I found that in the whole data there where 245 unique categories found. Next, I created a table which sorts each neighborhood's most common venues.



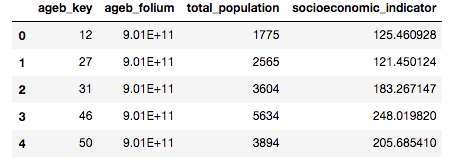
### 3.2 Mexico City's socioeconomic and demographic data

To match the venues and neighborhood's data with the socioeconomic indicators I first used an open source tool called Qgis, which lets you intersect shapefiles with geo points. That's how I obtained a dataset which matches each venue with the AGEB it's located in. When I have the AGEB for every venue I do another merge with the indicators database which has the key of all the AGEBs in Mexico City.

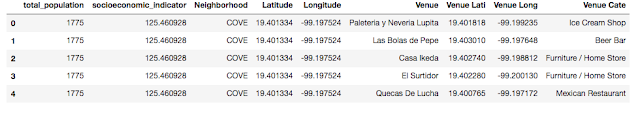
First, I matched each venue with the AGEB in which it is located.



Second, using the socioeconomic indicators database I matched [ CVE\_AGEB = ageb\_key ].

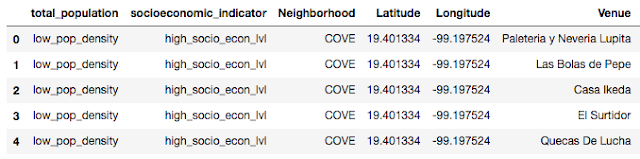


Finally, I obtained a dataset that looks like this:

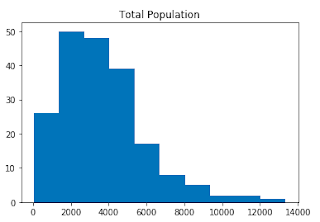


Since I wanted to make my data on income and population more understandable, I transformed it to dummy variables which indicate a certain segment of population density and income level.

The dataset looks something like this:

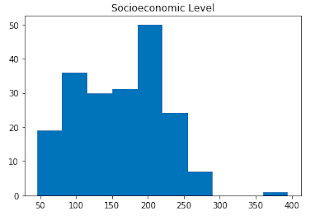


I performed two more visual analysis regarding the distributions of income and population across neighborhoods. I found that the most neighborhoods have a population between 2000 and 5000 inhabitants and that most neighborhoods have a medium socioeconomic level.



To analyze the data with categorical variables, I divided the population quantities into bins:

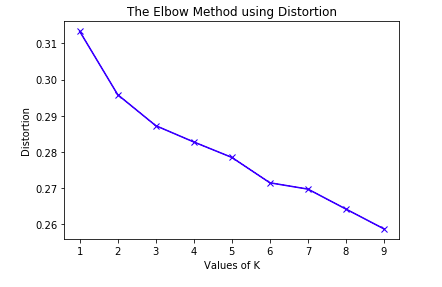
* 0 - 5000 inhabitants: “Low population density"
* 5000–10,000 inhab. : “Medium population density”
* 10,000–15,000 inhab. : “High population density”



* 0 - 150: “High socioeconomic level"
* 150 – 200 : “Medium socioeconomic level"
* 250 – 400 : “Low socioeconomic level"

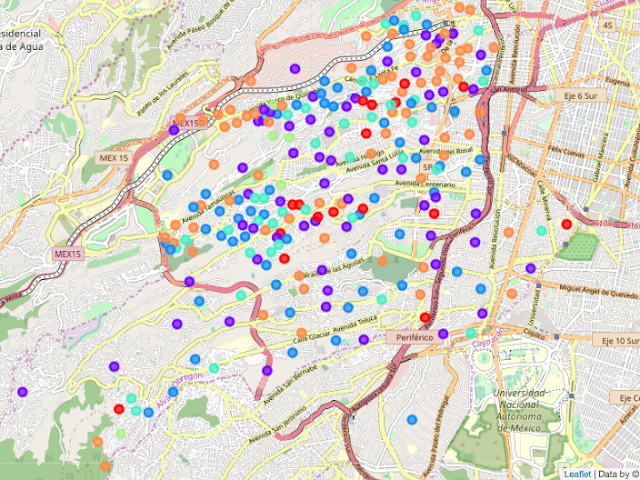
### 3.3  Clustering

Since neighborhoods share venue categories and indicators I decided to use K-means clustering but first I ran a validation to select the optimal K using the elbow method. I decided to use k=6 because it's a point where the slope of the curve tragically flattens, even though it keeps decreasing.



### 4.0 Results:

Using the Folium library, I made an Álvaro Obregón District map showing its neighborhoods grouped by cluster labels.



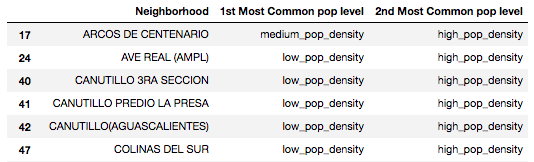
Cluster 0 (Red dots) :

I called cluster 0 the "Italian cluster" since the most common venue in almost all of its neighborhoods is a "Pizza place". Also, it's characterized by a predominantly low population density in the Álvaro Obregón district and by a balanced medium - high socioeconomic level.

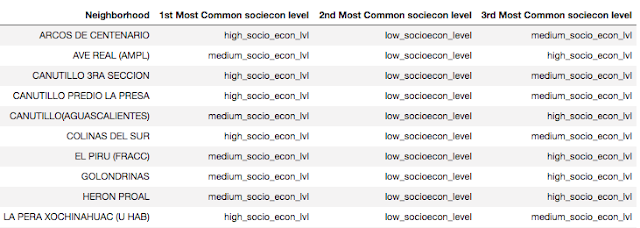
Most common venues:



Most common population density level:



Most common socioeconomic level:



I repeated this tables and obtained the following results:

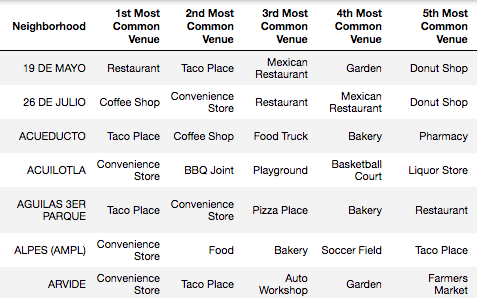
Cluster 1 (Purple dots) :

I called this cluster, the "Recreation Cluster" since it comprises several venue types related to different activities, you can find restaurants and coffee shops but also movie theaters, swimming pools, golf courses and zoos appear in the most common places of the neighborhoods. This cluster also is characterized by low population density and high socioeconomic level.



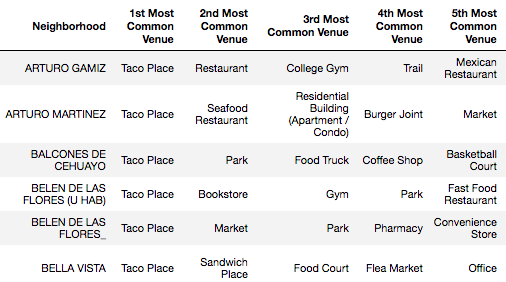
Cluster 2 (Blue dots) :

I called this cluster the "Practical Cluster", it seems that taco place's and convenience stores are very present here. This could be related to people who don't have much time to have a meal or tho go to the supermarket or grocery store. Also, medium socioeconomic level is predominant as well as low population density.



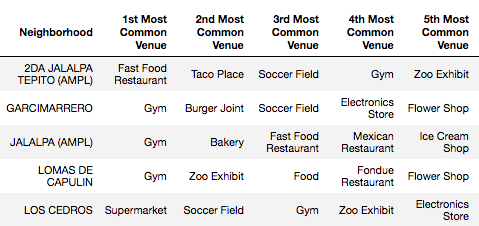
Cluster 3 (Aqua dots) :

By all means this is the "Taco Cluster", the name is self explanatory and we could also add that theres other common venues present but without a doubt, tacos dominate the scene. It's also characterized by low population density and medium income level. You could say this is the average taco lover Joe cluster.



Cluster 4 (Light green dots) :

I called this cluster the "fitness cluster" since it's most common venue is a gym together with some soccer fields. Predominantly you find low population density and medium income.



Cluster 5 (Orange Dots) :

This is also a mexican cluster since it's most common venues are taco places too, but also other mexican food restaurants. You can also observe that medium socioeconomic level and low population density are present. We could say that compared to the "Taco Cluster" this one has more variety regarding to mexican food.



## 5.0 Discussion

Even though the results seem pretty robust in some clusters, there's still significant variation inside others. To make decisions as a CPG business which will be translated into an upfront investment, this analysis might not be enough. Taking this into account, what we can do as next steps is to go deeper inside every cluster. The greatest benefit of this study is that it narrows our scope into which we have to look in order to make an informed decision.

Additionally, if more data is included in the study for example, for all the districts in Mexico City, it's possible that the results of the model will improve.

## 6.0 Conclusion

After running a K-means algorithm and obtaining a clustering analysis for the neighborhoods in Álvaro Obregón district of Mexico City, we found that for the objective of opening a CPG business we might need a deeper analysis but the most important result of this study is to narrow the scope of search which is also needed to start the business.

For example, if we take the "Italian cluster" we might not want to look into there if we want to open a Pizza parlour.

The results are straightforward to understand, but we recommend that further analysis is performed to reach the objective defined in the title of this investigation.

## 7.0 References

Agebs shapefile

https://datos.gob.mx/busca/dataset?tags=AGEBS

Neighborhoods in Mexico City

https://datos.cdmx.gob.mx/explore/dataset/coloniascdmx/information/

Githu repository

https://github.com/rogeliomj/CC\_Final\_Assignment