Capstone Project - The Battle of Neighborhoods – Week 2

Prediction of the appropriate locations for a stationery company – Rio de Janeiro

Samira Habib

April 27, 2021

1. Introduction & Business Problem

1.1. Background

The city of Rio de Janeiro is one of the most popular in Brazil, where it has a large flow of people and, consequently, is a very important shopping center.

Despite being a good business opportunity in this location, there is also a lot of observation by space and clientele. Every analysis for an enterprise must have prior knowledge of the location, the movement of people in the locality and the enterprises similar to the chosen branch. This reduces the risk of failure and returns the profits paid to the investor.

1.2. Problem

This project refers to an evaluation of schools that are located around the Maracanã neighborhood, in the city of Rio de Janeiro. These locations will be chosen so that the concentration of schools is greater in number for investment in a stationery enterprise. Competition and opportunity will be taken into account, and we will find ideal locations for a stationery store, through the study that will be done. Specifically, this report will be directed to those interested in opening a stationery store near Maranã, in Rio de Janeiro.

As main comments through factors such as:

• Proportion of schools by area;

• Number of stationery stores around these schools.

2. Data acquisition and cleaning

City used for data analysis in the project: Rio de Janeiro
We will use the data below to understand the location.

2.1. Data Sources

Data from the geographic coordinates of the city of Rio de Janeiro will be used as input to the Foursquare API, which will be used to provide location information for each neighborhood.

Along with this, a name search will also be carried out which will include 'schools' and 'stationers', which are located within a radius of 2,000 m in the surroundings of Maracanã, RJ.

The Foursquare database provided a JSON file (A lightweight format for exchanging information / data between systems and its meaning is JavaScript Object Notation) through latitudes and longitudes of the analyzed location. The institutions are organized by categories and carry the names of each of them in their information.

Data we will use:

- Name
- Categories
- Address
- Latitude
- Longitude
- Id

Link to Foursquare API:

https://pt.foursquare.com/city-guide

2.2. Cleaning and organizing the data table

In order to better visualize the data and be able to work with it, a cleaning of information that will not be used or that is incomplete in the original table is cleared.

Incomplete or missing data needs to be taken out so that it doesn't get in the way of the analysis. Thus, it is important to think about the main idea of the evaluation so as not to cause unnecessary loss of information about the problem in question.

Initially, I chose the Center of Rio de Janeiro, however, according to the cleanliness of the data and the course of the evaluation, little data was seen to analyze and few schools in the region, so it was necessary at that time to reassess the problem and change the location. Then, a new study and data search was initiated in the region of the state located in Maracanã, in the Brazilian state of Rio de Janeiro. That is why it is important to clean the data and see if its study would be feasible for an occasional solution to the proposed problem.

Data table:

- Schools

	name	categories	address	lat	lng	labeledLatLngs	distance	formattedAddress	id	id_global
0	Colégio Militar do Rio de Janeiro (CMRJ)	School	R. S. Francisco Xavier, 267	-22.916797	-43.227150	[{'label': 'display', 'lat': -22.9167972204570...	660	[R. S. Francisco Xavier, 267, Rio de Janeiro, ...	4dc1528422713750ba79c2ad	1
1	Colégio Batista	None	Rua Visconde de Itamarati, 75	-22.916780	-43.231521	[{'label': 'display', 'lat': -22.9167803740600...	515	[Rua Visconde de Itamarati, 75 (Maracanã), Rio...	4dbfe92c4b2221ec2d640bac	1
2	Colégio Ressurreição	School	Rua Oto de Alencar, 23	-22.915196	-43.225821	[{'label': 'display', 'lat': -22.9151963527452...	644	[Rua Oto de Alencar, 23, Rio de Janeiro, RJ, B...	512b53c1e4b0fd3b1159b204	1
3	Colégio Pedro II	School	R. S. Francisco Xavier, 204/208	-22.916347	-43.225898	[{'label': 'display', 'lat': -22.9163471631107...	714	[R. S. Francisco Xavier, 204/208 (Campus Tijuc...	5071a03ce4b0e1b65ea15489	1
4	Colégio Colégio Nossa Senhora de Lourdes (NSL)	Private School	R. Oito de Dezembro, 328	-22.909630	-43.240261	[{'label': 'display', 'lat': -22.9096297441251...	973	[R. Oito de Dezembro, 328, Rio de Janeiro, RJ,...	4ea5556577c8d0ce5f233bbb	1

- Stationery

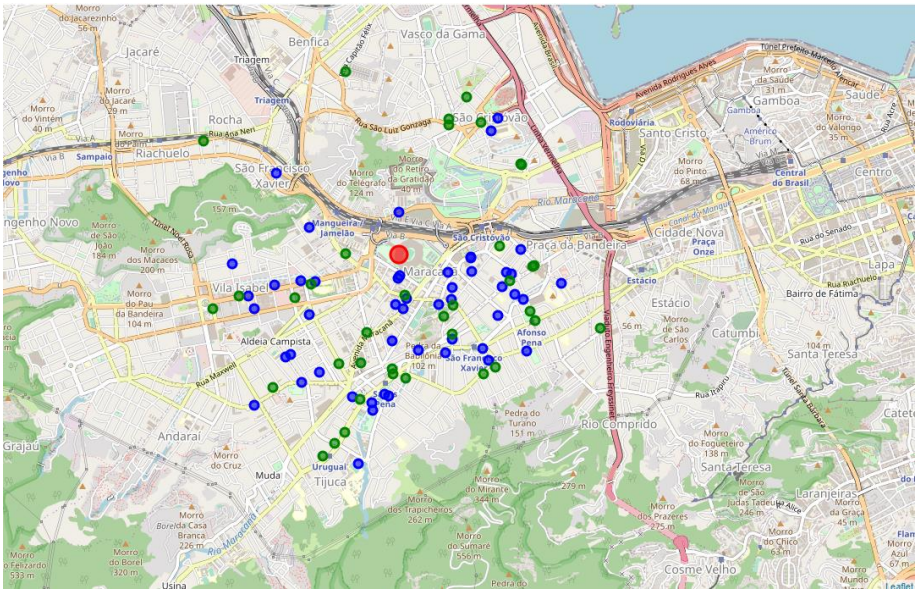
	name	categories	address	lat	lng	labeledLatLngs	distance	formattedAddress	id	id_global
0	Papelaria Porto Seguro	Paper / Office Supplies Store	Rua São Francisco Xavier	-22.915969	-43.230628	[{'label': 'display', 'lat': -22.9159694730426...	427	[Rua São Francisco Xavier, Rio de Janeiro, RJ,...	4e946c658b81ef41a94b7377	0
1	Papelaria Yunes	Bookstore	NaN	-22.912098	-43.236651	[{'label': 'display', 'lat': -22.9120980467721...	561	[Brasil]	4f6258a3e4b0832f00d7d3cd	0
2	papelaria raposo	Locksmith	NaN	-22.919445	-43.234497	[{'label': 'display', 'lat': -22.9194454163645...	879	[Brasil]	53c547c1498ef1a0284198aa	0
3	Papelaria Papel Moderno	Stationery Store	R. Sto. Afonso, 101	-22.922829	-43.231915	[{'label': 'display', 'lat': -22.9228290454082...	1189	[R. Sto. Afonso, 101, Rio de Janeiro, RJ, Brasil]	549da21b498e520a4e30b6b6	0
4	Papelaria Saens Pena	Arts & Crafts Store	Conde de Bonfim, 318	-22.923311	-43.231805	[{'label': 'display', 'lat': -22.92331113, 'ln...	1242	[Conde de Bonfim, 318, Rio de Janeiro, RJ, Bra...	4d318a195017a0939722409b	0

3. Methodology

In this project, we will direct our efforts in the detection of areas with fewer stationery stores and more schools. We will limit our analysis to an area of approximately 2,000 m around the central point of study.

The first step is to check the areas that contain schools and stationery, graphically visualizing the occupation of the area. To do this, we will join the two sets of data, which are schools and stationery, to check the graph for the proximity and quantity of each one.

Graphic map:



The graph shows the distribution of green dots (stationery shops) and blue dots (schools), some of them furthest from the center, also places some stationeries very close to each other, which for this business is its competition. Also evaluating the number of schools in the region, as they will be the target audience of the business.

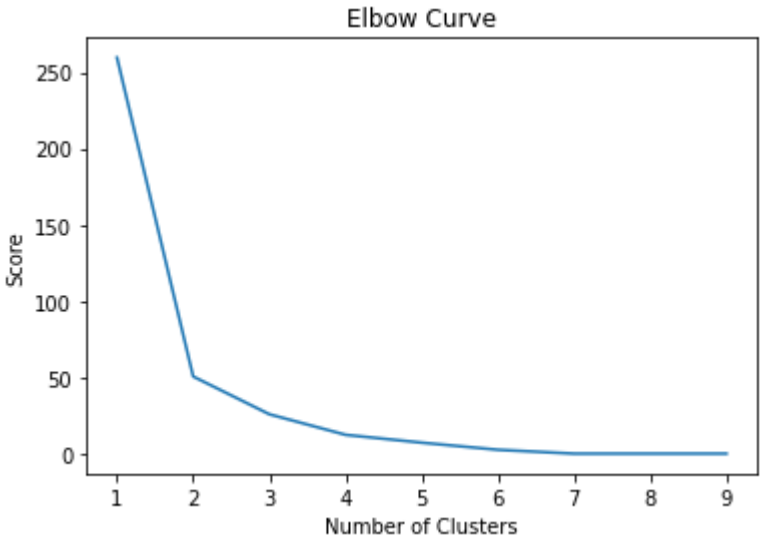
3.1. K-Means

K-means is considered as an unsupervised data mining algorithm that provides a classification of information according to the data itself. It is used to provide a classification of information according to the data itself. This classification is based on analysis and comparisons between numerical values of the data. In this way, the algorithm will automatically provide an automatic classification without the need for any human supervision, that is, without any existing pre-classification.

In the second and last stage, using the unsupervised algorithm, K-Means, we will create clusters of localities that meet some basic requirements established in the discussion with the interested parties: we will take into account locations with the fewest stationery stores or with many schools that need more service, because there is more per occupied area. We will present the map of all these locations, but we will also create clusters of these locations to identify zones / neighborhoods / general addresses that should be a starting point for the final exploration of the 'street level' and search for the ideal location of the site by the interested parties.

Using two K-Means methods, we will do the ‘elbow’ method to evaluate the best ‘k’ value, as well as the ‘silhouette’ method. This will help to know what is the ideal number of clusters to be made by the algorithm to obtain the best performance of the analysis.

Elbow Method



What we find in this graph is:

- From 3 on, the graph starts to stabilize, showing that from that number of ‘k’ we will have good results for clustering.

Silhouette Method

```
N_cluster: 2, score: 0.7122384664844124
N_cluster: 3, score: 0.7035769181418999
N_cluster: 4, score: 0.7388596296711039
N_cluster: 5, score: 0.8660353234275127
N_cluster: 6, score: 0.9490635904029654
N_cluster: 7, score: 0.9945051883541731
```

The numbers above indicate that the best number of clusters to be made is the number ‘7’. So that we get a good view of the clusters in the graph.

4. Results and Discussion

Our analysis shows that, there are many schools that are well served by stationers in the region, there are low stationery supplies in only three locations. The highest concentration of schools was detected in the regions of Maracanã, São Cristóvão and Vila Isabel, so we concentrated on these areas to assess the amount of stationery contained therein.

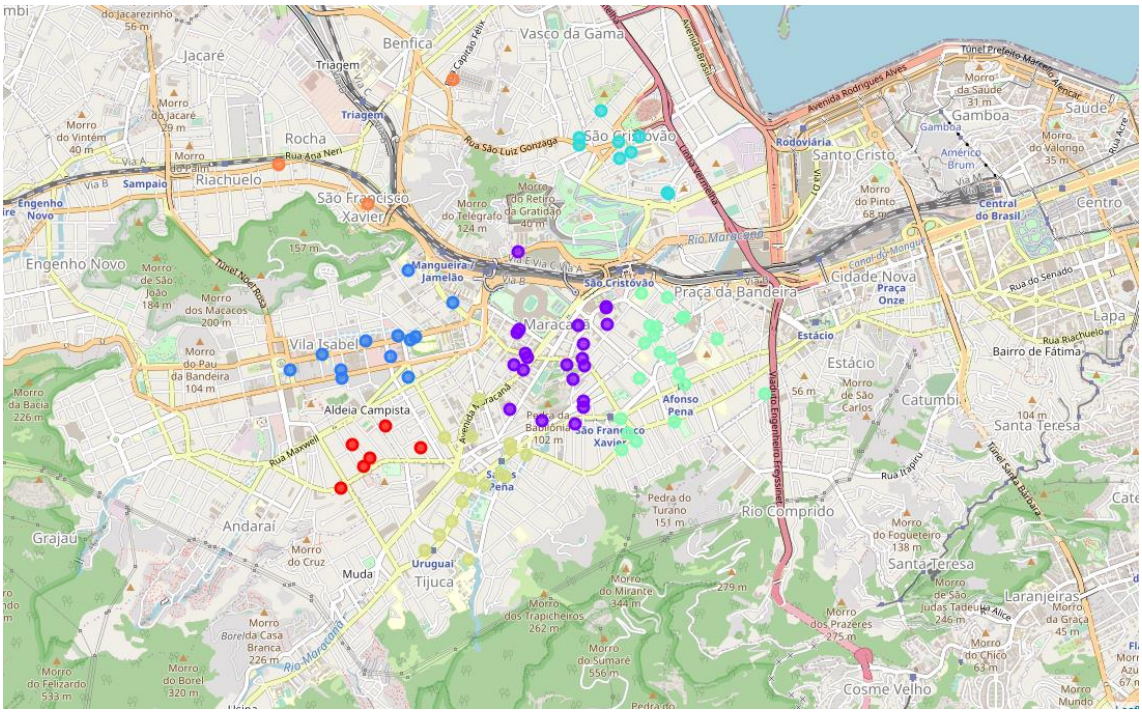
The possibility of doing business is difficult if you think about the number of stationery stores in some regions, which have few schools and many stationeries per region. But there is also the opportunity to have many schools with few stationery stores, in these regions it is necessary to identify the best point.

The result of all of this is 3 zones containing a significant number of schools based on the number of existing locations. This, of course, does not mean that these zones are, in fact, ideal places for a new stationery store! The purpose of this analysis was to provide only information about areas close to Maracanã, but not overflowing with stationery - there is a possibility that there are other reasons for not having so many stationery stores in these regions. The recommended zones should therefore only be considered as a starting point for a more detailed analysis that could eventually result in a location that not only has no competition nearby, but also other factors taken into account and all other relevant conditions met.

Result of clusters by latitude and longitude:

	Cluster Labels	lat	lng
0	1	-22.915969	-43.230628
1	2	-22.912098	-43.236651
2	5	-22.919445	-43.234497
3	5	-22.922829	-43.231915
4	5	-22.923311	-43.231805
...	...	...	...
45	5	-22.927237	-43.235858
46	0	-22.921495	-43.242150
47	0	-22.923894	-43.243463
48	5	-22.925436	-43.235918
49	4	-22.921150	-43.218220

Division of clusters on the map:



Distribution of stationery clusters by number of schools:

Cluster Labels	id_global	
0	0	9
	1	11
1	0	4
	1	17
2	0	5
	1	7
3	0	6
	1	3
4	0	10
	1	6
5	0	1
	1	5
6	0	2
	1	1

- Number '0' refers to stationers in the region;
- Number '1' refers to schools.

5. Conclusion

The objective of this project was to identify the areas of Rio de Janeiro close to Maracanã with a low number of stationery stores, in order to help interested parties to restrict the search for the ideal location for a new stationery store. Calculating the distribution of schools from Foursquare data, we first identify the schools closest to the central point, then we generate another distribution of points where the stationers are grouped in the region. The grouping of these sites was then carried out in order to create the main zones of interest (containing the largest number of potential sites) and the addresses of these zone centers were created to be used as starting points for the final exploration by the interested parties.

The final decision on the ideal location of the stationery will be at the discretion of the interested parties based on the specific characteristics of the locations in each recommended zone, taking into account additional factors such as the attractiveness of each location (proximity to offices or shopping centers), levels of noise / proximity to main roads, real estate availability, prices, social and economic dynamics of each neighborhood etc.

6. Future directions

In the future, the search will be broader, finding clients such as offices and other commercial institutions that use stationery services. As well, look for other competing establishments for a better evaluation.