Clustering Neighborhoods of São Paulo

Vinícius Longo

July 21, 2020

1. Introduction

1.1 Background

São Paulo is the largest city of the southern hemisphere, both economically and in population. With a metropolitan area of around 21 million people, it is known for its strong international influences in commerce, finance, arts and entertainment. For those reasons, many companies and business owners are interested in placing their enterprises in the city.

1.2 Problem and Interest

Imagine we are willing to open a unit of a fast-food burger chain in the city, but we are still deciding where it should be placed. We are targeting middle to upper middle-class customers, serving both in our restaurant space and by deliveries. We also figured that brazilians are not so used to have burgers for lunch, so we should focus on dinners.

1.3 Interest

We must find the neighborhoods that fit best for the business. We are looking for regions quite dense and with an above average income per capita. Also, we'll target regions that people might go eat for dinner, so probably around bars and clubs. Since the city is huge, we'll start the analysis considering the hole city first (with all 31 boroughs), and then go micro for each neighborhood within the boroughs selected.

2. Data acquisition and cleaning

2.1 Data sources

The first data set (extracted from wikipedia) is a list comparing every "Subprefecture" (Boroughs) of the city by HDI (Human Development Index). This help us to have an idea on the average household income of each region. I also used a data set (extracted from São Paulo's city hall website) with the population and demographic density of each borough and what are the neighborhoods located in those boroughs. Besides that, the Foursquare API were implemented to extract information about a bunch of the venues located around those regions of interest.
Links:
- https://pt.wikipedia.org/wiki/Lista_de_subprefeituras_de_S%C3%A3o_Paulo_por_%C3%8Dndice_de_Desenvolvimento_Humano
- https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/subprefeituras/dados_demograficos/index.php?p=12758
- https://developer.foursquare.com/

2.2 Data cleaning

Both tables were scraped using the read_html function from pandas. In the first data set, the information was divided into 3 tables, separating the Boroughs by their HDI levels, so I concatenated them into one. I also dropped unimportant columns and translated the column names to English, I only used 2010 data.

For the second data source, some of the neighborhood names were disconfigured, mainly because of some Portuguese signs and letters that wasn't recognized in the html file (e.g. "~, ´, ç"). So I reconfigured Neighborhood and Borough names and translated column names. Population and Demographic Density were also in the wrong format (e.g. "123.123"), so I had to solve that.

After that, I merged both tables into one. Then, I used the library 'geopy' to track latitude and longitude of each Borough.

	Borough	HDI	Population	Demographic Density	Latitude	Longitude
0	Pinheiros	942	289743	9140	-23.567249	-46.701951
1	Vila Mariana	938	344631	13005	-23.583700	-46.632741
2	Santo Amaro	909	238025	6347	-23.656230	-46.719116
3	Lapa	906	305526	7619	-23.521576	-46.704349
4	Sé	889	431106	16454	-23.550443	-46.633446

Figure 1. Boroughs of São Paulo with the biggest HDI levels.

3. Exploratory Data Analysis

3.1 Finding Neighborhoods within Boroughs of interest

I used the merged data frame to visualize São Paulo's map and see the Borough's locations and how they are distributed. Since we are looking for upper middle class customers, I dropped from the data frame Boroughs with an HDI below 800 and Demographic Density below 5000, so I was left with 12 out of the 31 Boroughs of the city. Then, I searched into the old table from the town hall's website to see what Neighborhoods were situated within the Boroughs of interest. Within the 12 Boroughs, there were 49 Neighborhoods.

	Borough	Latitude	Longitude
Aricanduva	Aricanduva/Formosa/Carrão	-23.578024	-46.511454
Carrão	Aricanduva/Formosa/Carrão	-23.551530	-46.537791
Vila Formosa	Aricanduva/Formosa/Carrão	-23.566876	-46.546323
Butantã	Butantã	-23.569131	-46.721874
Morumbi	Butantã	-23.596499	-46.717845

Figure 2. Neighborhoods assign to their Boroughs and coordinate.

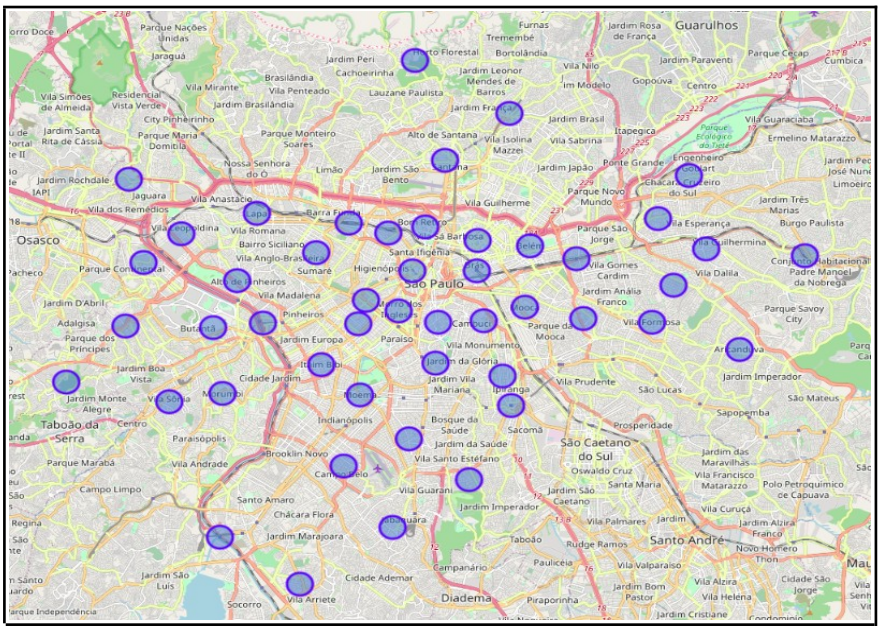

Figure 3. Neighborhoods of interest distributed in São Paulo

3.2 Analyzing most common venues for each Neighborhood

Using Foursquare API and the coordinates of each Neighborhood, I tracked all the venues within a radius of 500 foot and created a data frame with the top 10 most common venues on each of the Neighborhoods. In total, there were 256 unique categories of venues around those points.

4. Statistical Inference and Clustering

Since I was dealing with unlabeled data, i decided to use Kmeans clustering for the project. It was a great tool to divide the city into different clusters and it facilitated my job at finding the ideal neighborhood. After some tests, I realized K=5 was the best fit. I chose neighborhoods labeled as cluster 1, since the neighborhoods presented had a bigger volume of bars and restaurants than the others (the focus of the restaurant is to mainly serve at dinner time). After that, I ended up with 17 neighborhoods within 9 Boroughs. So I decided to study the incidence of possible competitors within those neighborhoods by creating a formula. The formula was based on the number of negative parameters (Burger Joints, BBQ Joints and Sandwich places) in relation to a

positive parameter, the number of Bars.

5. Results

Then I created a data frame with the Neighborhoods and its final result (table in descending order):

Neighborhood	Burger Joint	Sandwich Place	BBQ Joint	Bar	Results
Campo Belo	0.022222	0.000000	0.000000	0.111111	15.556
Mooca	0.092593	0.000000	0.018519	0.222222	14.815
Lapa	0.000000	0.000000	0.000000	0.051282	10.256
Saúde	0.000000	0.047619	0.000000	0.047619	4.762
Liberdade	0.000000	0.028571	0.028571	0.042857	2.857
Tucuruvi	0.010204	0.020408	0.000000	0.030612	1.020

Those were the only neighborhoods that ended up with a positive result. As you can see, the formula was based on the incidence of the venue categories divided by TOTAL. The best results are from the neighborhoods “Campo Belo”, “Mooca” and “Lapa”, all good candidates for the restaurant.

6. Discussion about the fast-food project

There are many other variables that we should take in consideration before opening a new restaurant, like rental prices, average age of local citizens, violence rates, car and people traffic. We should also look at specific spots, not only its neighborhoods. Nevertheless, with this research we can have a good idea on what are the best candidate neighborhoods to place the unit of the fast-food burger chain, giving the analysis taking the target customers and the competition into consideration.

7. Conclusion

There are literally hundreds of thousands of possible spots to place a fast-food restaurant in São Paulo, and it is quite hard to confidently choose the best ones. As shown in the Results section, we came up with 3 out of the 96 neighborhoods of the city, they are “Campo Belo”, “Mooca” and “Lapa”. I can confidently say these have good commercial potential for our Burger Joint, but I am not sure they are the actual best. We would have to take many other variables into consideration to conclude that.

Nevertheless, this was a very interesting project and helped me get a deeper understanding of my birth town and to develop my Data Science understanding. Thank you for reading.