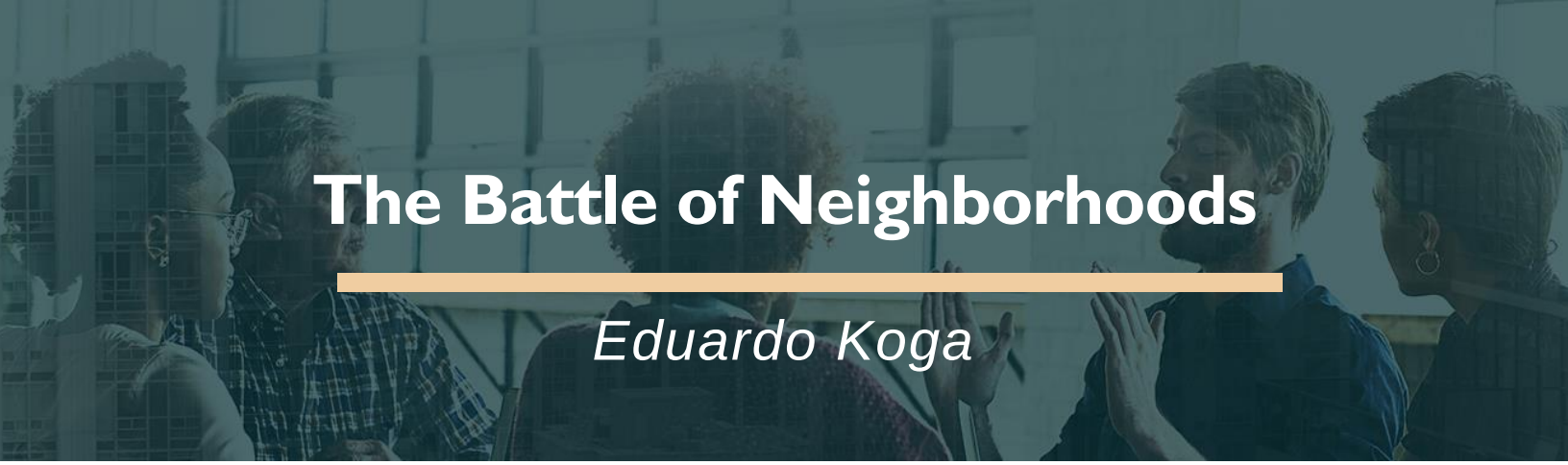


The Battle of Neighborhoods

Eduardo Koga

Introduction

Background

São Paulo is a municipality in the Southeast Region of Brazil. The metropolis is an alpha global city and the most populous city in Brazil, the Americas, the Western Hemisphere and the Southern Hemisphere. Additionally, São Paulo is the largest Portuguese-speaking city in the world. The municipality is also the world's 4th largest city proper by population. The city is the capital of the surrounding state of São Paulo, the most populous and wealthiest state in Brazil. It exerts strong international influences in commerce, finance, arts and entertainment. São Paulo is a cosmopolitan, melting pot city, home to the largest Arab, Italian, Japanese, and Portuguese diasporas, with examples including ethnic neighborhoods of Mercado, Bixiga, and Liberdade respectively. São Paulo is also home to the largest Jewish population in Brazil, with about 75,000 Jews. In 2016, inhabitants of the city were native to over 200 different countries and the city counts with more than 11million inhabitants.

Problem

Understand the characteristics of each district of the city to find out where there could be opportunities to open a new business.

Interest

The result of this work will be interesting for any people who want to undertake any kind of business and do not know how to start the research phase and/or how the city is distributed in terms of businesses across the different neighborhoods.

Data acquisition and cleaning

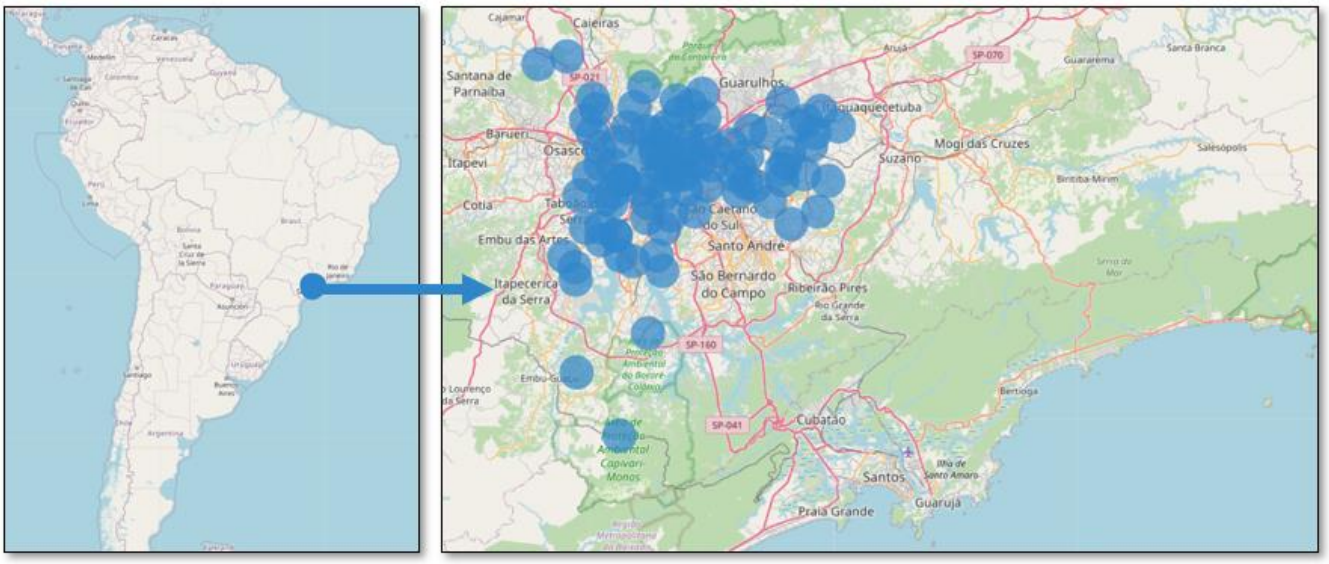
Data sources

- Wikipedia: List of districts of São Paulo city, area, population and HDI (human development index)
- https://pt.wikipedia.org/wiki/Lista_de_subprefeituras_do_munic%C3%ADpio_de_S%C3%A3o_Paulo
- Proprietary data for the geo references covering latitude and longitude by district
- Foursquare data for the different categories of venues for each of the respective neighborhoods

Data cleaning
Some districts do not have foursquare data, so these have been dropped from the data frame.

Methodology

Data acquisition and preparation
The raw data comes from Wikipedia where 97 different districts of São Paulo city were listed complemented with some proprietary information about geographic references (latitude and longitude). By using *folium*, all districts were plotted in São Paulo city map as follows.

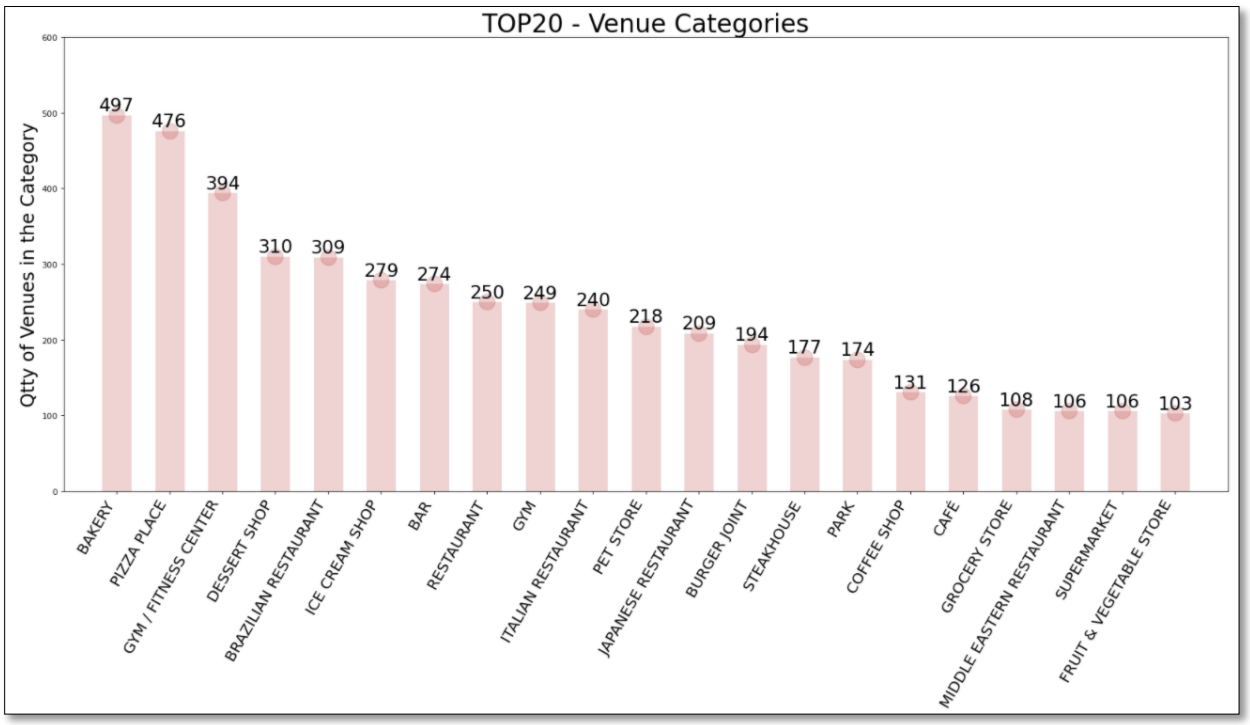


As shown in the map above, it is very difficult to have insights using only the districts placed in different points of the city, so in the next section I start getting more interesting data from Foursquare such as venues and categories of businesses.

Data analysis

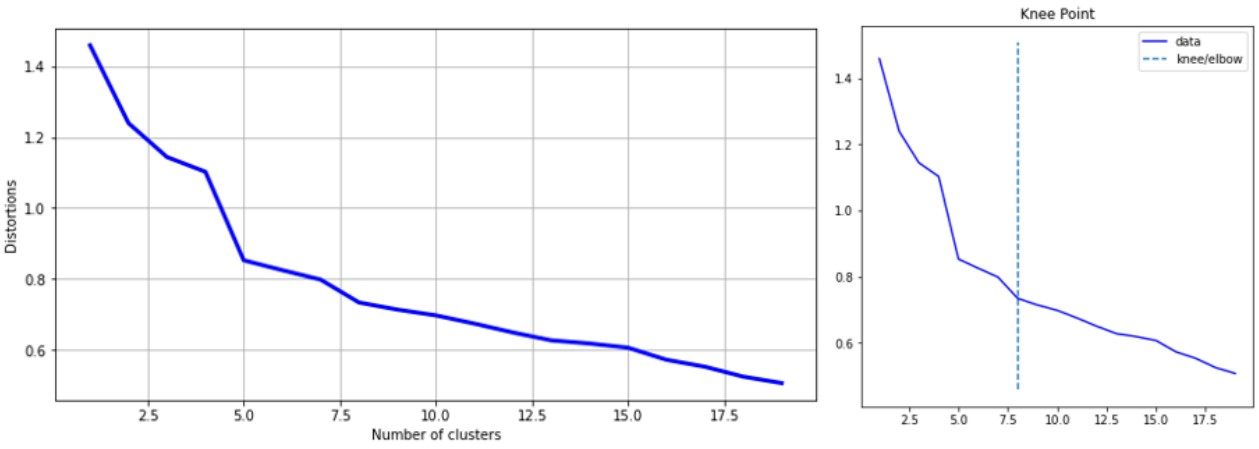
By using the Foursquare data, limiting the analysis to the city's districts with a radius of 5000 meters from each neighborhood, it ends up with a list of 9376 venues organized in 300 different categories.

The data being used in the in the analysis will cover all of these 9376x300 matrix, however, in order to give an idea about what are the main type of businesses found in the city, below is a chart showing the Top 20 venues.

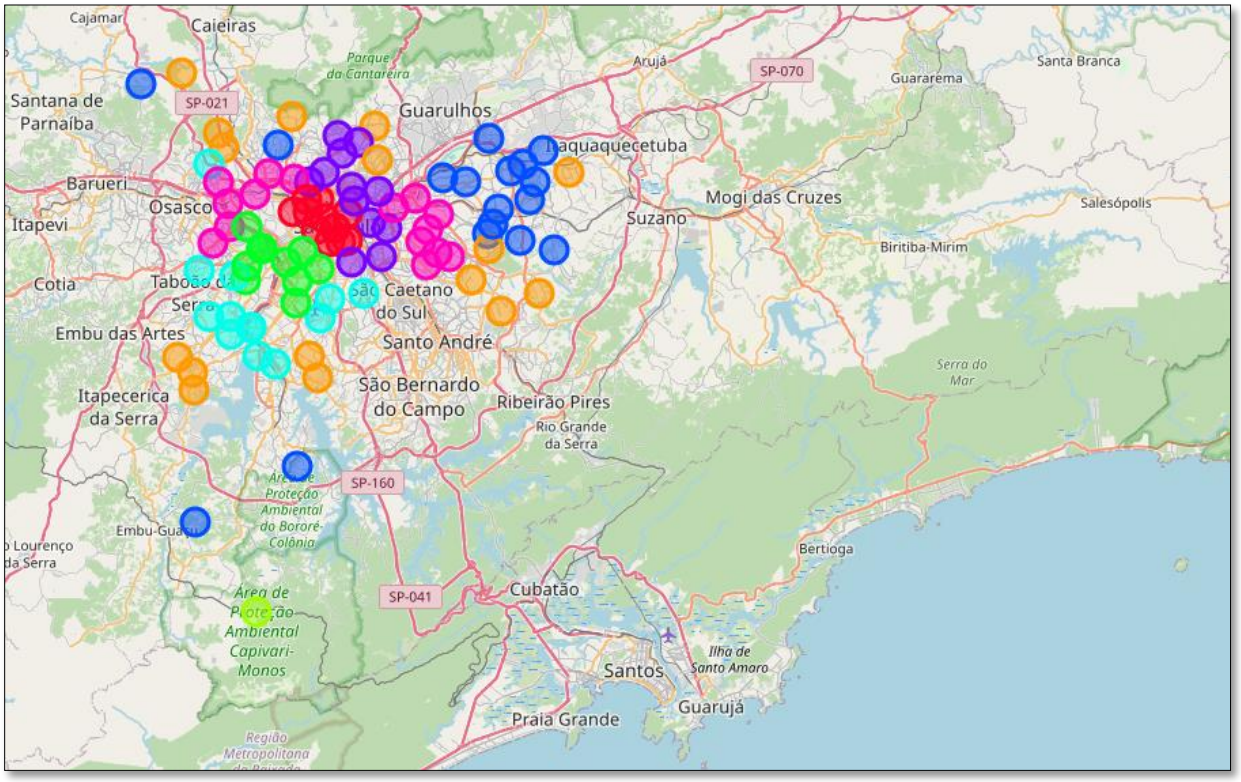


As for the data analysis, I have used the *k-means* method to group the city's districts into different clusters based on their similarities and 3 most common categories of venues.

The result of this analysis is shown in the far left side chart and on the right side the *Knee Point* chart shows that the ideal number of clusters would be 8.

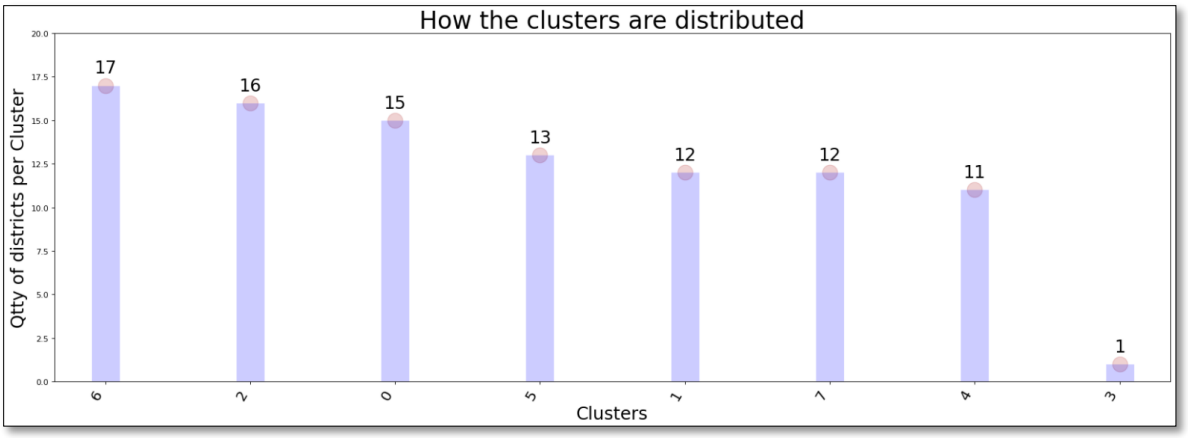


By using 8 different clusters as proposed by the *k-means* method, the city's districts are now organized as follows, i.e. it is much clear to see some pattern based on the different types of venues which will be explored later.



Results section

As explained previously, the clustering methods has grouped the city's districts into 8 different clusters given the similarities on venues and neighborhood. Curiously each cluster presented almost the same quantity of districts, except cluster 3m, which can be considered as an outlier.



When we analyze each of these clusters, we find the following characteristics:

Cluster 6: represents the largest and most common cluster which comprises 17 districts. It is composed of different types venues, but mostly pizza places and bakeries. This cluster is more concentrated in the east zone of the city.

Cluster 2: here we find mostly bakeries and gyms and the cluster covers 16 districts spread mainly in the periphery of the city.

Cluster 0: is primarily composed of dessert shops and is covering 15 districts located in both west and northern zones.

Cluster 5 and 7 here is where we find a large diversity of restaurants, burger places as well as bars. The cluster is composed of 25 districts.

Clusters 1: this is the central districts, where the main companies are, and there we can find mostly brazilian restaurants, Ice Cream shops, theaters and bookstores.

Clusters 4: cover primarily Italian restaurants and ice-cream shops and is located in one of the city's most expensive neighborhoods with 13 districts.

Clusters 3: is an outlier composed of just one district where the main venues are market, other great outdoors and campground.

Discussion

The analysis transformed data into information and it can provide some good reading about how the city is organized in terms of venues as well as where opportunities can be found. The combination of geographic references, maps and clustering methods are quite powerful and with the inclusion of some other information, such as HDI and population per area the analysis can definitely be boosted.

Conclusion

Foursquare seemed to be a great repository of data, however for some regions that are far from downtown, some venues might fail to provide information. Having said that, this kind of analysis would probably be better if paid-information is added to it as well as other variables such as different metrics and methodologies.