**The Battle of Neighborhoods**

**Eduardo Koga**

1. **Introduction**
   1. **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.

* 1. **Problem**

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

* 1. **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 how the city is distributed in terms of businesses across the different neighborhoods.

1. **Data acquisition and cleaning**
   1. **Data sources**
      1. Wikipedia: List of districts of São Paulo city, area, population and HDI (human development index)

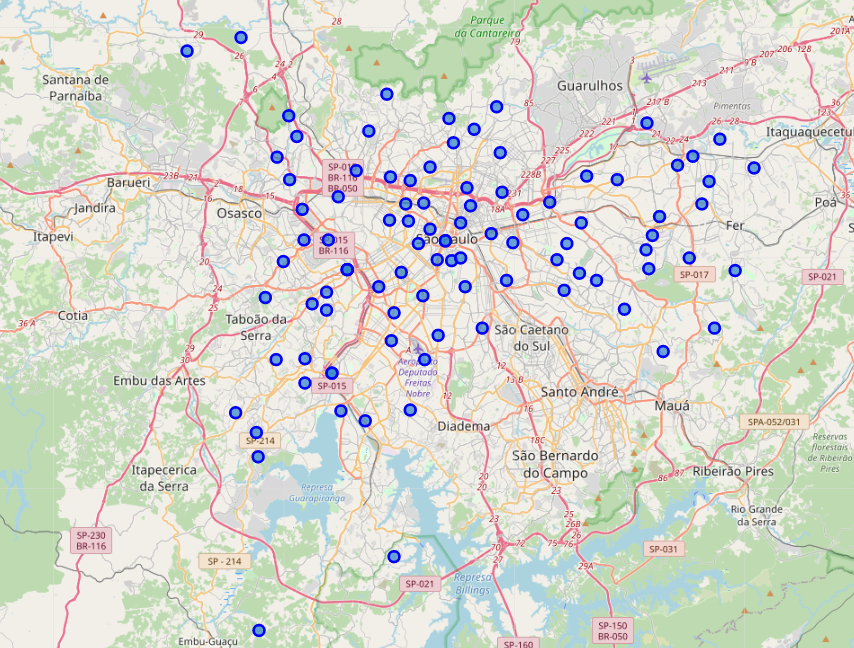
<https://pt.wikipedia.org/wiki/Lista_de_subprefeituras_do_munic%C3%ADpio_de_S%C3%A3o_Paulo>

* + 1. Proprietary data for the geo references covering latitude and longitude by district
    2. Foursquare data for the different categories of venues for each of the respective neighborhoods
  1. **Data cleaning**

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

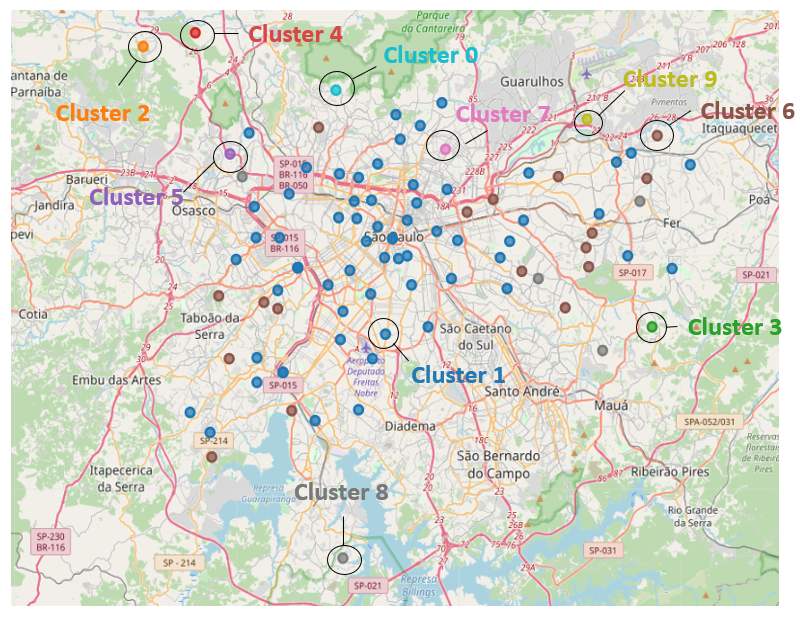
1. **Methodology**
   1. **Data acquisition and preparation**

The raw data comes from Wikipedia where the 97 different districts of São Paulo city were listed with their population, area and HDI (human development index). After that the data was complemented by some proprietary information about geographic references (latitude and longitude) as this information was not found on internet. By using folium, all districts were plotted on the São Paulo city map



* 1. **Data analysis**

As for the data analyzes, I have used the k-means method to group the 96 different districts into 10 different clusters based on their similarities considering the 5 most common categories of venues. The result of this grouping method is plotted on the city’s map as shown below.



1. **Results section**
   1. Cluster 1 represents the largest and most common cluster and comprises **64** **districts**. It is composed of different types venues, mainly restaurants, bars, pizza places, among others. This cluster is also concentrated in a central area of the city, where most of the companies are.
   2. Cluster 6: here we find mostly pizza places, restaurants, gym and fitness covering **17 districts.** The region these districts are found is where the residences are, which explains the quantity of pizza places. São Paulo is one of the biggest pizza consumers around the globe.
   3. Cluster 8 is primarily composed of bakeries and is covering **5 districts**
   4. Clusters 0, 5 and 7 are considered as outliers in this analysis as they have just 1 district each and the categories found are diverse/random. This is because of the low number of venues provided by foursquare
   5. Clusters 2, 3, 4 and 9 are located in the peripheric areas of the city and the most common venues are related to any kind of athletic & sports, so we find tennis courts, soccer fields, yoga studios and farmer market. These clusters have 1 district each.
2. **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.

1. **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.