Analyzing venue availability in São Paulo neighborhoods

Lucas Falcão Monteiro

February 12, 2021

1. Introduction

1. Background

São Paulo is one of the largest cities in the world, and with over 20 million people living in its metropolitan area, it ranks among the top 5 urban areas in the world, being the largest one in the southern hemisphere and tied with Mexico City in the Americas. Despite Rio being the most famous city internationally, São Paulo is Brazil's financial center, and the homonymous state accounts for one fifth of the country population and one third of its GDP — comparable to a country such as Saudi Arabia.

2. Problem

São Paulo reflects Brazil's rampant inequality in that its neighborhoods are also divergent in terms of income levels and access to infrastructure. The central neighborhoods tend to be richer and with multiple options for food and leisure, while neighborhoods on the outskirts — what we call the 'periphery' — are poorer and lack proper employment, access to public transportation and assorted commercial establishments.

3. Interest

Having information about what neighborhoods lack diverse types of venues would be of interest to the government and possibly non-governmental organizations when setting up directions for the development of the city.

2. Data

1. Data acquisition

Neighborhood information for São Paulo is available on the city's website, with the shapefile format. This was imported and read with the PyShp library.

Venue information was obtained from the Foursquare API, utilizing each neighborhood's center as the latitude and longitude values for the search and a radius of 1.6 km, which is a reasonable distance for São Paulo.

2. Data transformation

The information in the shapefile for the districts featured coordinates for the neighborhoods' boundaries, to the central location was obtained by averaging these values. However, those coordinates were given with the WGS-84 format, and this was converted to UTM (the usual format for latitude and longitude) with the pyproj library.

3. Data cleaning and feature selection

The working table (Figure 1) contained codes and names of each neighborhood, their latitudes and longitudes and venues that were close to their center (at most 100). The venues also had descriptions of name, location and category.

[178]:

	Code	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	JBO	JOSE BONIFACIO	-23.568415	-46.429467	Kodomo-no-Sono	-23.566109	-46.428694	Japanese Restaurant
1	JBO	JOSE BONIFACIO	-23.568415	-46.429467	Galpão das Plantas	-23.572315	-46.442596	Flower Shop
2	JBO	JOSE BONIFACIO	-23.568415	-46.429467	Roldão Atacadista	-23.564515	-46.415218	Big Box Store
3	JBO	JOSE BONIFACIO	-23.568415	-46.429467	Batata	-23.554245	-46.428270	Food Truck
4	JBO	JOSE BONIFACIO	-23.568415	-46.429467	Supermercado Ricoy	-23.563833	-46.416634	Grocery Store

Figure 1: Head of the data table

This table was used to construct other tables that display the number of different types of venues in each neighborhood, or that use one-hot encoding to display the types of venue present in each neighborhood. We can then extract information to group neighborhoods according to the diversity of the venues nearby.

3. Methodology

The focus was to group neighborhoods based on the availability and diversity of venues nearby. The data used to perform this task was the geographical coordinates of each neighborhood, number of venues found within a 1.6 km radius and types of venues found in the Foursquare database.

A histogram was constructed to visualize the distribution of types of venues according to neighborhood, and a table containing number of venues and number of types of venues was put together (Figure 2).

[79]:

	Neighborhood	Number of Venues	Types of Venues
0	AGUA RASA	100	59
1	ALTO DE PINHEIROS	100	60
2	ANHANGUERA	15	12
3	ARICANDUVA	53	30
4	ARTUR ALVIM	100	53

Figure 2: Neighborhoods and the number and types of venues found within a 1.6 radius (maximum = 100).

Based on this table, the k-means algorithm was executed using 3 clusters. This machine-learning function from the sklearn package identifies the distance between each data point (e.g. Euclidean) and performs successive iterations to group them into "k" different clusters, where the center of the cluster is the mean of each data point in it. In each round, each point is assigned to the closest cluster center.

The types of venues were also used to cluster neighborhoods based on diversity. A similar table was put together with one-hot encoding to display whether a neighborhood had some type of venue or not (Figure 3).

Similarly, k-means was used with 6 clusters to identify similar neighborhoods in this aspect. 3 neighborhoods were excluded due to being outliers. Finally, the neighborhoods were plotted on a map using the folium package, color coded according to their cluster label.

[161]:

	District	Acai House	Accessories Store	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Aquarium
0	AGUA RASA	0.000	0.000	0.000	0.000	0.000	0.010	0.000
1	ALTO DE PINHEIROS	0.010	0.000	0.000	0.000	0.000	0.000	0.000
2	ANHANGUERA	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	ARICANDUVA	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	ARTUR ALVIM	0.000	0.010	0.000	0.000	0.000	0.000	0.000
...	...	...	...	...	...	...	...	...
90	VILA MARIANA	0.010	0.000	0.000	0.000	0.000	0.000	0.000
91	VILA MATILDE	0.014	0.000	0.000	0.000	0.000	0.000	0.000
92	VILA MEDEIROS	0.000	0.000	0.000	0.000	0.000	0.010	0.000
93	VILA PRUDENTE	0.000	0.000	0.000	0.000	0.000	0.000	0.000
94	VILA SONIA	0.000	0.000	0.000	0.000	0.000	0.000	0.000

95 rows × 355 columns

Figure 3: One-hot encoding of types of venues present in each neighborhood.

4. Results

The histogram (Figure 4) showed that about 50% of the neighborhoods have a good variety of types of venues, while some 40% has an intermediate variety of venues, and about 10% has low variety.

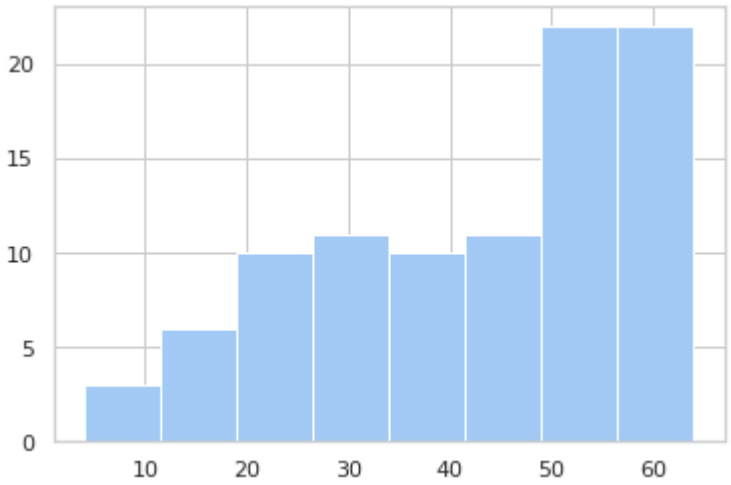

Figure 4: Histogram showing number of different types of venues on the x axis, frequency on the y axis.

Grouping neighborhoods in this aspect with k-means (Figure 5) showed what would be expected: central districts cluster together in regions with high access to different types of venues (red points). A second category shows neighborhoods that have intermediate variety of venues nearby (light green) and a third one shows regions with poor access to venues.

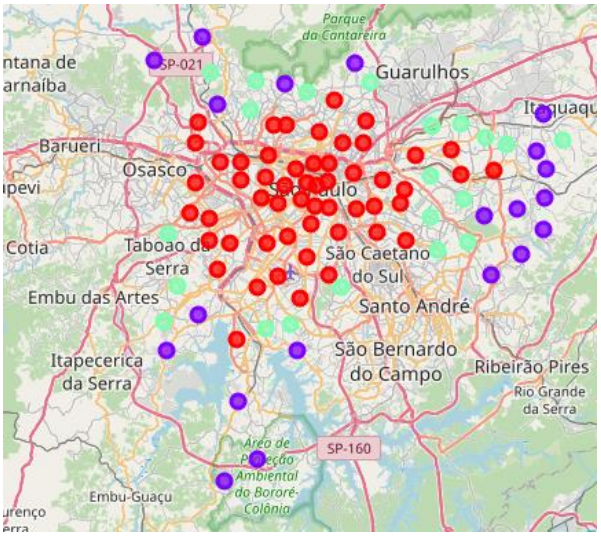

Figure 5: Neighborhoods clustered based on the number and different types of venues found nearby.

Grouping neighborhoods based on the types of venues yielded 3 clusters that were outliers and removing them produced a more in-depth clustering, showing 6 regions that contain similar types of venues (Figure 6).

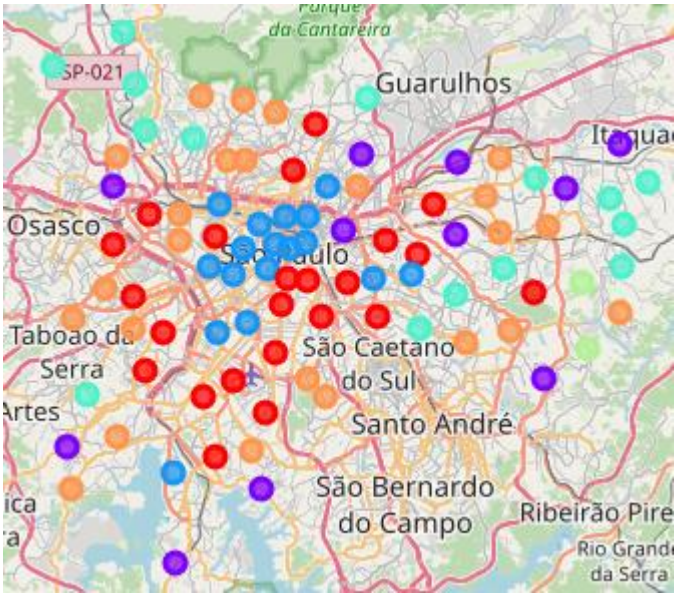

Figure 6: Neighborhoods clustered according to the types of venues present nearby.

5. Discussion

The clustering of neighborhoods based on the number and types of venues nearby was expected: central neighborhoods contain a larger variety, whereas peripheral neighborhoods do not. The analysis of diversity showed an interesting result, however: some regions farther away from the center display a similar level of variety in comparison to central districts.

For example, Parque do Carmo (represented by the red point isolated in the East Side of the city) is not a rich neighborhood, being notable for having a large city park and Japanese immigrants that used to farm land there. Still, it is closest to more central neighborhoods than to districts nearby.

Conversely, Belém, a previously industrial district in the city with lots of Italian immigrants, clusters together with more peripheral districts despite being close to the center.

Generally, there is some spatial connection between the clusters, but the analysis allows us to track and identify which regions do not show great diversity of venues: namely the purple, light blue and light green clusters. These areas could be of focus to companies and the public sector in order to further develop the commerce and options for the population.

A limitation of this analysis is the lack of demographic information about the areas and the imprecise centering of each neighborhood: some regions are sparsely populated and the place where people actually reside may not coincide with the center of the neighborhood as defined by the city. This was not considered in the work.

Another shortcoming is the fact that Foursquare data was limited to 100 venues per district center and increasing this number would allow for finer distinction between neighborhoods. Also, we cannot be sure that the annotation in the platform truly reflects the venues that were nearby. In any case, this could still point us towards a notion of development in the sense that poorer neighborhoods with less access to internet and technology would tend not to register their venues in the app, thus leading to disparities seen in the analyses.

Future work could be done to cross-verify annotation in Foursquare with other platforms such as Google Maps, improving the coordinates of each district based on actual population density and increasing the maximum number of venues retrieved for each location.

6. Conclusion

In this project I could create a list with all districts in São Paulo and their coordinates, clustering them together based on the number diversity of venues, and based on the types of venues themselves.

The results showed that central districts have a larger diversity of venues, while peripheral districts have intermediate or very low diversity. Interestingly, clustering them based on the types of venues shows that some neighborhoods in the center are more similar to others in the outskirts and vice versa.

Despite imperfections in the data, this could still point out to neighborhoods that need improvement in the quantity and quality of establishments nearby, and this could be of use to both the private and public sectors.