Battle of Neighborhoods

Leonardo Carneiro

June 24, 2020

1 Introduction

1.1 Background

Sao Paulo is the most populous city in Brazil with a population of over 12.25 million inhabitants, as of 2019. It exerts strong international influences in commerce, finance, arts and entertainment and is listed as an alpha global city by the Globalization and World Cities Study Group (GaWC). It has the largest economy by Gross Domestic Product (GDP) in Latin America and the Southern Hemisphere, representing 10.7% of all Brazilian GDP and being home to 63% of stablished multinational companies in Brazil.

Sao Paulo is also a cosmopolitan, melting pot and ethnically diverse city, home to the largest Arab, Italian, Japanese, and Portuguese diasporas. It is also home to the largest Jewish population in Brazil, with almost 75,000 Jews. In 2016, inhabitants of Sao Paulo were native to 196 different countries.

Such a diverse culture translates to a diverse cuisine. We can find many different categories of restaurants in Sao Paulo: Italian, Asian, Argentinian, just to name a few. Sao Paulo attracts many to start their businesses in the food industry, either small ones such as mobile food vendors, food truck and fast food joints, or larger ones as restaurants. Before starting to operate, though, they need to find the appropriate location to open. What do they take into account when making this decision?

1.2 Problem

Upon exploring the districts of Sao Paulo, I hope to find whether opening a restaurant in a neighborhood of restaurants plays a role in the success of the business. To simplify our analysis, this report focuses on the Italian cuisine based on the total number of restaurants we can find in Sao Paulo. The methodology used here applies to other cuisines as well.

2 Data

The data used in this report is listed below:

- Districts in Sao Paulo. I scraped the Wikipedia webpage for districts names.
- Geocoder library in Python to get the geographical coordinates from the districts.
- Foursquare API, to extract the most common venues located nearby each district, as well as their respective ratings, likes and tips counts.

2.1 Data Cleaning

The data scraped from the internet, as well as their respective coordinates were combined into one table with three columns (District name, latitude and longitude) and 96 rows (one for each district).

With these data at our disposal, we utilized the Foursquare API to collect all the venues names, IDs and categories located in every district, limiting to a total of 100 in a 500 meter radius. Then, we created a dataframe using the pandas library from the data extracted from Foursquare and filtered the Category column to only display Italian restaurants. I found a total of 49 Italian restaurants in the city of Sao Paulo.

Then, we made another request in the Foursquare API, to collect ratings, as well as likes and tips counts for every Italian restaurant from their IDs. These data were stored in another pandas dataframe, that I used to group all venues by district and calculate the average ratings for the Italian restaurants located there. However, not all restaurants had available information on ratings, likes and tips so I dropped the only venue without these data. The cleaned database consists of 23 districts hosting at least one Italian restaurant.

2.2 Feature Selection

I merged the previous table with the first, created with the districst latitude and longitude. It shows every district where we can find at least one Italian restaurant along with its latitude, longitude and the average rating of all Italian restaurants found in that district. The table is displayed below:

Table 1: Feature Selection

District	Latitude	Longitude	Average Rating
Barra Funda	-23.525462	-46.667513	7.1
Bela Vista	-23.562210	-46.647766	7.9
Campo Belo	-23.626731	-46.669421	9.2
Carrão	-23.551531	-46.537791	8.5
Consolação	-23.557887	-46.660321	8.0
Ipiranga	-23.589273	-46.606162	8.5
Itaim Bibi	-23.584381	-46.678444	8.4
Jaraguá	-23.446658	-46.736213	5.4
Jardim Paulista	-23.567436	-46.663692	7.8
Lapa	-23.521576	-46.704349	7.7
Liberdade	-23.566704	-46.631809	5.6
Moema	-23.597085	-46.662888	7.9
Pari	-23.532976	-46.615849	8.7
Penha	-23.523683	-46.543782	7.5
Perdizes	-23.537929	-46.680671	7.2
República	-23.545335	-46.642257	8.2
Sacomã	-23.601282	-46.602555	5.8
Santana	-23.499321	-46.628933	8.8
Tatuapé	-23.540252	-46.576642	7.9
Tucuruvi	-23.480075	-46.603270	7.2
Vila Leopoldina	-23.530072	-46.734319	8.0
Vila Maria	-23.513184	-46.589156	6.8
Vila Sônia	-23.599935	-46.739162	6.6

In order to find if location and nearby venues play a role in business decision making and success, we will cluster the above districts according to their coordinates (latitude and longitude) and the district average rating.

2.3 Data Preprocessing

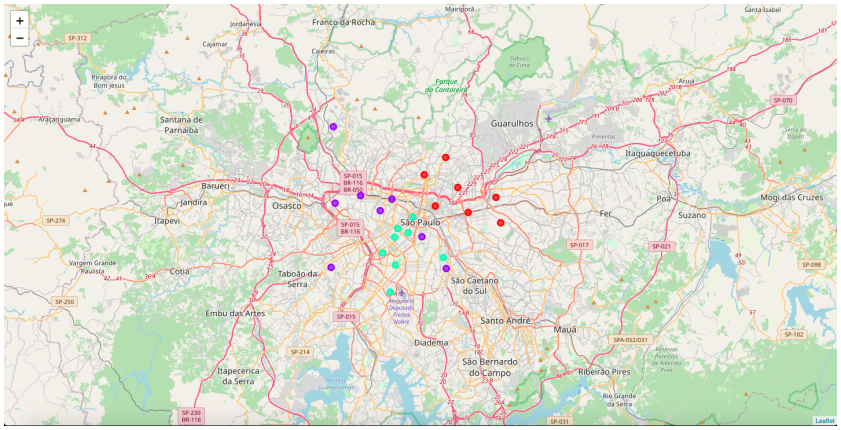
Before we proceed to run our analysis on the Latitude, Longitude and Average Rating columns of Table 1 we need to normalize our features, since they range in different scales. We use the StandardScaler function from the preprocessing module under the scikit-learn library to normalize our features.

3 Methodology

After properly standarding and scaling our features, we can run the k-means clustering algorithm to group all districts using latitude, longitude and average rating as features.

I ran the k-means algorithm with 2, 3, 4 and 5 clusters and stored the inertia scores of each run. According to the elbow method, the optimal number of clusters to use is $k = 3$. Then, I stored the cluster labels estimated for the districts and plotted a map of the city of Sao Paulo showing markers for each district colored by cluster label. The map is displayed below:

Figure 1: Districts of Sao Paulo Clustered by Latitude, Longitude and Average Rating



3.1 One Hot Encoding

In order to further assess our custers, I explored every district's most common venues to chech whether nearby businesses factor in the success of Italian restaurants.

From the first Foursquare API, I collected a total of 1,395 venues, belonging to a total of 222 different cateories. I used one hot encoding in the dataframe collected from the Foursquare API and grouped by district to calculate proportions for all the categories of venues.

Then, with a dataframe consisting of 223 columns (one for each venue category plus the district name) and 23 rows (one for each district), I defined a function to get the most common venues and looped through the dataframe to create a table with the 1st up to the 10th most common kinds of venues located in each district.

4 Results

We begin our analysis of the results by looking separately at every cluster. First, we can see from Table 1 that the average ratings calculated for every district range from 5.4 to 9.2. If we look at cluster 1, shown in Table 2, however, there are 8 districts assigned to it, with a maximum average rating of 8.0 and mean 6.7. It seems like the districts with the worst reviews, on avearge, in our database were grouped in the same cluster.

The presence of many places such as bars, pizza places, bakeries, pastelarias and grocery stores among the most common for these districts might suggest a public preference for lighter meals in these regions. This could be a possible reason for the poor reviews Italian restaurants received, since they tend to attract customers willing to appreciate more complete meals, and spend some time at the restaurant.

Table 3 displays information on the districts listed on cluster 2, which, on the other hand, seems to be consisted of the districts with the best reviews, on average. The mean value calculated from the average ratings is the highest among the clusters (8.2) and the lowest average rating here is 7.8.

Table 2: Cluster 1

District	Average Rating	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Barra Funda	7.1	Restaurant	Music Venue	Brazilian Restaurant	Café	Sandwich Place	Bookstore	Supermarket	Fair	Chocolate Shop	Motel
Jaraguá	5.4	Grocery Store	Bakery	Gym / Fitness Center	Brazilian Restaurant	Pharmacy	Pet Store	Italian Restaurant	Falafel Restaurant	Event Space	Fair
Lapa	7.7	Pastelaria	Brazilian Restaurant	Restaurant	Candy Store	Market	Sporting Goods Shop	Pharmacy	Bar	Bakery	Tea Room
Liberdade	5.6	Pizza Place	Pet Store	Farmers Market	Gym / Fitness Center	Pharmacy	Bar	Bakery	Brazilian Restaurant	Chinese Restaurant	Pastelaria
Perdizes	7.2	Burger Joint	Bar	Café	Restaurant	Gym / Fitness Center	Dessert Shop	Bakery	Pizza Place	Pharmacy	Ice Cream Shop
Sacomã	5.8	Pharmacy	Department Store	Brazilian Restaurant	Cosmetics Shop	Bar	Farmers Market	Chocolate Shop	Restaurant	Café	Martial Arts Dojo
Vila Leopoldina	8.0	Brazilian Restaurant	Italian Restaurant	Bar	Market	Burrito Place	Supermarket	Stadium	Fruit & Vegetable Shop	Spa	Snack Place
Vila Sônia	6.6	Bakery	Japanese Restaurant	Grocery Store	Farmers Market	Burger Joint	Gym / Fitness Center	Pizza Place	Italian Restaurant	Deli / Bodega	Food & Drink Shop

Unlike districts in cluster 1, the districts assigned to cluster 2 seem to host restaurants from a wide variety of cuisines (Chinese, Middle Eastern, Spanish, Japanese, Italian, among others). Also, one of the most common venue categories is Italian restaurants, which suggests that the Italian restaurants located in these regions seem to succeed despite the competition. The district of Itaim Bibi alone, for instance, is home to a total of 11 Italian restaurants and the average rating for this district is 8.4.

Additionally, districts such as Itaim Bibi, Moema, Campo Belo, Ipiranga and Jardim Paulsta belong in the south side of Sao Paulo, which is the wealthiest region of the city. It might be the case that the venues owners had a solid amount of capital to invest in their businesses, providing high quality service to their customers and, in return, got high ratings.

Finally, cluster 0 exhibits the greatest variety of venues categories, ranging from plazas and shopping malls to eletronics and toy/game stores. Also, cluster 0 display the rangiest average ratings, from 6.8 to 8.8. That is typical of central Sao Paulo, where one can find any sort of goods, services and food places. In fact, most districts assigned to cluster 0 are locaeted nearby the central side of the city. Each district here is unique in its own way and that also applies to the public taste. There are Italian restaurants thriving in this environment, just as much as there Italian restaurants not as appealing. We display cluster 0 information in Table 4.

Table 3: Cluster 2

District	Average Rating	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Bela Vista	7.9	Italian Restaurant	Pizza Place	Bar	Nightclub	Restaurant	Café	Hotel	Brazilian Restaurant	Cosmetics Shop	Coffee Shop
Campo Belo	9.2	Bar	Bakery	Pizza Place	Restaurant	Brazilian Restaurant	Pharmacy	Dessert Shop	Pet Store	Fast Food Restaurant	Café
Consolação	8.0	Brazilian Restaurant	Coffee Shop	Ice Cream Shop	Café	Movie Theater	Vegetarian / Vegan Restaurant	Hotel	Gym / Fitness Center	Indie Movie Theater	Lounge
Ipiranga	8.5	Burger Joint	Bakery	Pizza Place	Pet Store	Brazilian Restaurant	Gym	Gastropub	Empanada Restaurant	Beer Store	Eletronics Store
Itaim Bibi	8.4	Italian Restaurant	Japanese Restaurant	Restaurant	Hotel	Brazilian Restaurant	Burger Joint	Bar	Ice Cream Shop	French Restaurant	Dessert Shop
Jardim Paulista	7.8	Italian Restaurant	Gym / Fitness Center	Restaurant	Hotel	Spanish Restaurant	Burger Joint	Coffee Shop	Café	Spa	Bar
Moema	7.9	Dessert Shop	Furniture / Home Store	Restaurant	Japanese Restaurant	Middle Eastern Restaurant	Gym / Fitness Center	Massage Studio	Pharmacy	Sushi Restaurant	Italian Restaurant
República	8.2	Brazilian Restaurant	Bar	Coffee Shop	Record Shop	Café	Pizza Place	Tea Room	Theater	Restaurant	Burger Joint

Table 4: Cluster 0

District	Average Rating	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Carrão	8.5	BBQ Joint	Convenience Store	Pharmacy	Brazilian Restaurant	Pizza Place	Gym / Fitness Center	Yoga Studio	Plaza	Eletronics Store	Steakhouse
Pari	8.7	Clothing Store	Brazilian Restaurant	Middle Eastern Restaurnt	Shopping Mall	Restaurant	Café	Warehouse Store	Women's Store	Falafel Restaurant	Bar
Penha	7.5	Café	Pharmacy	Supermarket	Bakery	Coffee Shop	Chocolate Shop	Japanese Restaurant	Clothing Store	Department Store	Cosmetics Shop
Santana	8.8	Restaurant	Pizza Place	Burger Joint	Middle Eastern Restaurant	Pharmacy	Toy / Game Store	Spa	Bookstore	Brazilian Restaurant	Gym / Fitness Center
Tatuapé	7.9	Ice Cream Shop	Café	Dessert Shop	Coffee Shop	Clothing Store	Pizza Place	Snack Place	Brazilian Restaurant	Restaurant	Burger Joint
Tucuruvi	7.2	Ice Cream Shop	Pizza Place	Fast Food Restaurant	Clothing Store	Chocolate Shop	Department Store	Dessert Shop	Market	Bakery	Snack Place
Vila Maria	6.8	Bar	Pharmacy	Shoe Store	Italian Restaurant	Food Truck	Breakfast Spot	Flower Shop	Brazilian Restaurant	Dessert Shop	Brewery

5 Discussion

In this report, I explored, examined and clustered districts in Sao Paulo according to the Italian restaurants listed on the Foursquare API and their respective ratings. I found that the clusters generated by the k-means algorithm have their own peculiarities, with cluster 2 likely exhibiting a wealthy neighborhood, cluster 0 located nearby the central side and cluster 1 with lower average ratings.

We can not properly assess the success of a business based on average ratings alone, since there are many factors that might play role and are not being considered. However, finding all of them is out of scope of this report.

On the other hand, average ratings might suggest the public perception

of the restaurants and the nearby venues found in the districts gives us an idea of the profile of people living there and passing by. Since the districts grouped in cluster 2 are located in a prime area of Sao Paulo, it would not be at all surprising to find high quality restaurants from any cuisine – Italian inclusive. These restaurants tend to be highly reviewed, pushing the district average ratings up.

6 Conclusion

In this report, we saw the peculiarities of each cluster, how the nearby venues might indicate the public preferences in the districts member to each cluster. Although we could not examine closely the reasons for success of the Italian restaurants, we saw that the geographical location is not an important factor, specially when you have a well defined target audience, as indicated by the clusters and nearby venues. There are top rated restaurants spread all around the city, thriving in either high or low competition.