Capstone Project - The Battle of Neighborhoods - Foursquare data
Countries in São Paulo

Marcio Sakamoto

June 27, 2020

Introduction: Business Problem

This project used Foursquare restaurant data to try to identify regions of São Paulo city with more concentration of specific immigrants. This initial study can be useful for the stakeholders to decide the best neighborhood for a new imported goods store, another type of restaurant of a cuisine, or the region to stay or live in Sao Paulo.

Although some cuisines are more popular and spread around the city like Italian restaurants and Japanese restaurants, others are still more popular among its community. The study focused on identifying these niches by locating clusters of restaurants of the same category's family in an area of the city.

Data

The data for this project was extracted from Foursquare APIs and São Paulo City's administration website.

The study's search area was limited by a radius of 10Km from the city center.

The Foursquare Explore API is limited to 100 results per query. In order to split the requests around the city, it was done by district and the radius of each request was defined by each district area.

The districts info came from a shapefile pack that was opened in a GeoPandas Dataframe.

The geolocation data from the file were in UTM coordinate system that uses metric units. Using this data it was possible to easily calculate the Area and Radius of each District.

The distance of each District to the Center was calculated considering the SE district as the center of the city.

In order to show the district limits on the map, we applied latitude and longitude coordinate projection to match Foursquare data and Folium map.

With the new coordinate data, the Latitude and Longitude of each District was calculated by the centroid of the district's Geometry Data.

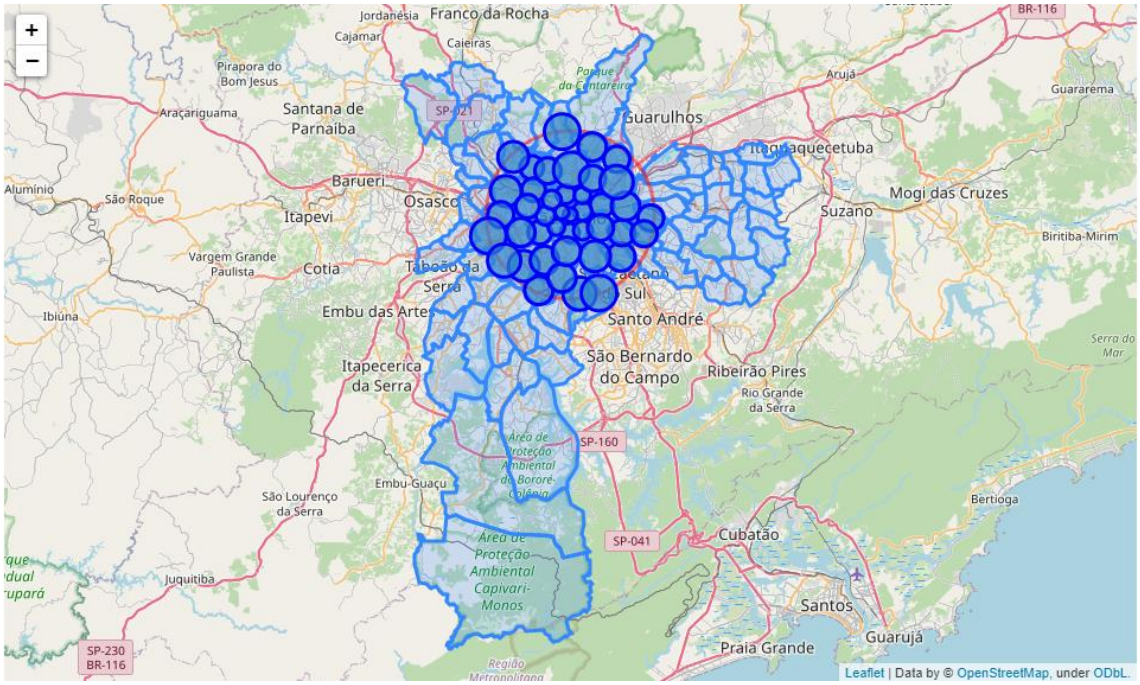
To limit the search area to 10Km, districts out of this range was dropped from the Dataframe.

	District	geometry	Area	Radius	CenterDist	Latitude	Longitude
0	VILA PRUDENTE	POLYGON ((-46.58103 -23.57258, -46.58100 -23.5...	9.583361	1746.561877	7779.427332	-23.592145	-46.572337
1	JD PAULISTA	POLYGON ((-46.66935 -23.58357, -46.66939 -23.5...	6.186069	1403.241631	4163.116518	-23.566106	-46.666312
2	LAPA	POLYGON ((-46.71896 -23.53626, -46.71879 -23.5...	10.281937	1809.099816	8101.003642	-23.522228	-46.705529
3	LIBERDADE	POLYGON ((-46.63739 -23.55583, -46.63710 -23.5...	3.650702	1077.986309	2131.577112	-23.566547	-46.631510
4	LIMAO	POLYGON ((-46.68931 -23.50863, -46.68926 -23.5...	6.456407	1433.575331	7186.137512	-23.497105	-46.675590

Districts DataFrame top rows snapshot

Map of study area

São Paulo districts and search area. Cycle in red is the 10 Km radius. In blue the radius based on each district area.



São Paulo districts and search area.

Foursquare data

Now that we have the districts, coordinates and radius of search, we get the venues of these areas from Foursquare Explore API:

- https://api.foursquare.com/v2/venues/explore

The API call was restricted to 'Food' categoryId ('4d4b7105d754a06374d81259') and limited to the API's maximum range of 100 results per query.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	SE	-23.547305	-46.631001	Piero Pasta & Café	-23.548734	-46.632186	Italian Restaurant
1	SE	-23.547305	-46.631001	Café do Páteo	-23.547905	-46.632732	Café
2	SE	-23.547305	-46.631001	Café da Bolsa BM&FBovespa	-23.545452	-46.634039	Café
3	SE	-23.547305	-46.631001	Café Girondino	-23.544493	-46.634238	Café
4	SE	-23.547305	-46.631001	Bendita Panelinha	-23.546309	-46.634384	Brazilian Restaurant

Venues DataFrame top rows snapshot

Since some categories are sub category of a cuisine, for example ‘Sushi Restaurant’ category is a sub category of ‘Japanese Restaurant’ category, it was necessary to add the Parent Category to each venue and filter it to show only categories related to a cuisine of country.

The Foursquare Categories was extracted from Foursquare Categories API:

- https://api.foursquare.com/v2/venues/categories

Each category was listed with its parent category and then added to Venues DataFrame.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	Parent	id
0	SE	-23.547305	-46.631001	Piero Pasta & Café	-23.548734	-46.632186	Italian Restaurant	Italian Restaurant	4bf58dd8d48988d110941735
1	SE	-23.547305	-46.631001	Café do Páteo	-23.547905	-46.632732	Café	Café	4bf58dd8d48988d16d941735
2	SE	-23.547305	-46.631001	Café da Bolsa BM&FBovespa	-23.545452	-46.634039	Café	Café	4bf58dd8d48988d16d941735
3	SE	-23.547305	-46.631001	Café Girondino	-23.544493	-46.634238	Café	Café	4bf58dd8d48988d16d941735
4	SE	-23.547305	-46.631001	Bendita Panelinha	-23.546309	-46.634384	Brazilian Restaurant	Brazilian Restaurant	4bf58dd8d48988d16b941735

Venues DataFrame with Parent Category

With the Parent Category data, we removed the categories that could not be used to identify a foreign cuisine:

> 'Brazilian Restaurant', 'Pizza Place', 'Bakery', 'Café', 'Cafeteria', 'Food Truck', 'Restaurant', 'Snack Place', 'Diner', 'Burger Joint', 'BBQ Joint', 'Breakfast Spot', 'Bagel Shop', 'Buffet', 'Comfort Food Restaurant', 'Deli / Bodega', 'Donut Shop', 'Fast Food Restaurant', 'Fish & Chips Shop', 'Food Court', 'Food Stand', 'Fried Chicken Joint', 'Gastropub', 'Gluten-free Restaurant', 'Hot Dog Joint', 'Salad Place', 'Sandwich Place', 'Seafood Restaurant', 'Steakhouse', 'Theme Restaurant', 'Wings Joint', 'Vegetarian / Vegan Restaurant', 'Irish Pub', 'Dumpling Restaurant'

and applied a filter to keep only categories that have a nationality in the name.

	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Category
0	SE	-23.547305	-46.631001	Piero Pasta & Café	-23.548734	-46.632186	Italian Restaurant
1	SE	-23.547305	-46.631001	Bar Linguiçaria Di Callani	-23.542001	-46.629535	Italian Restaurant
2	SE	-23.547305	-46.631001	Temakeria Guin	-23.548372	-46.634524	Japanese Restaurant
3	SE	-23.547305	-46.631001	Fenícia Culinária Libanesa	-23.542819	-46.632094	Lebanese Restaurant
4	SE	-23.547305	-46.631001	Restaurante Primeiro Mundo	-23.549688	-46.635063	American Restaurant

Venues DataFrame snapshot after the category filtering

Methodology

With the data collected we will run DBSCAN clustering over the geospatial coordinates and categories of venues.

It will provide clusters based on Restaurant category and localization.

The DBSCAN Clustering method was chosen due to the non-fixed number of clusters and the feature to limit the cluster range and points by its epsilon and min_samples parameter.

Analysis

Ordering the venues by Category, we find that Japanese and Italian restaurants are the main cuisines among the categories analyzed.

Category	Qtd
Japanese Restaurant	228
Italian Restaurant	203
Chinese Restaurant	54
Korean Restaurant	29
Argentinian Restaurant	27

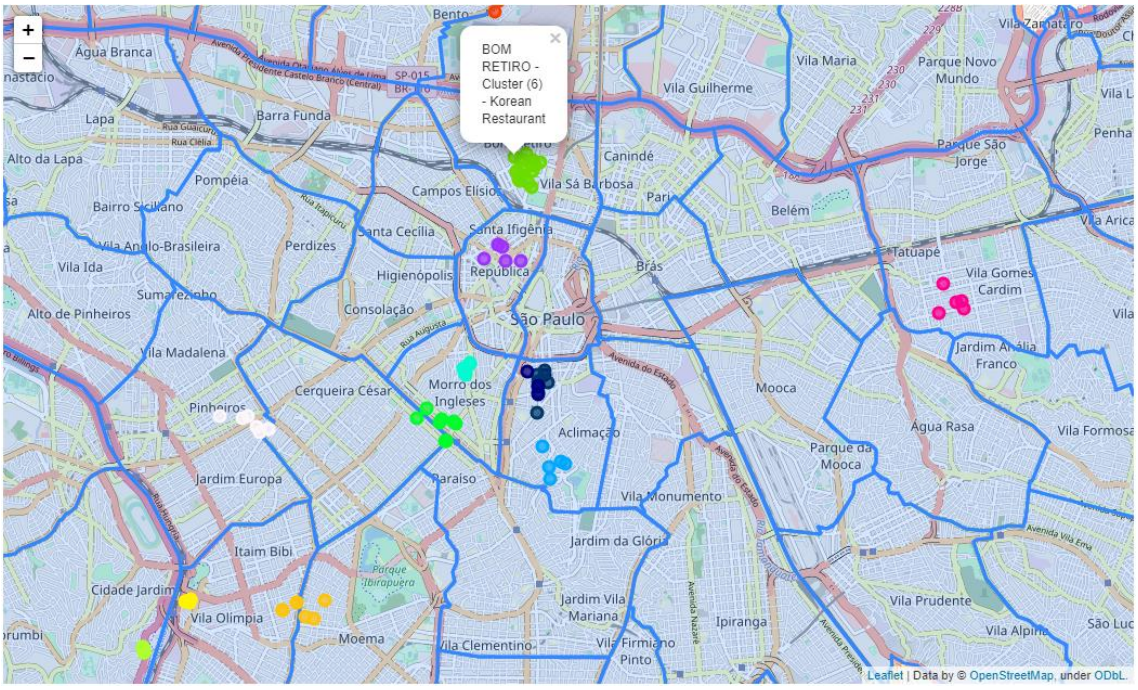
Top 5 Restaurants

And checking it by district, we have the district of Itaim Bibi and Jd. Paulista as the areas with greater amount of restaurants in this study.

District	Qtd
ITAIM BIBI	36
JD PAULISTA	36
MOEMA	34
BOM RETIRO	29
BELA VISTA	28

Top 5 Districts

Clustering the data by geolocation and restaurant category, we find 13 restaurants clusters in 11 Districts.



Restaurant clusters and districts

Results and Discussion

Checking the list of clusters, its position on the map, virtual visit in Google Street View and previous knowledge of the city, we can add the following notes about some of the cluster.

BELA VISTA

Japanese Restaurant cluster. Possible positive cluster of immigrants. Although it's close to Paulista Avenue, a touristic and business center with concentrated number of restaurants, it's also where the Japanese Consulate is located.

Italian Restaurant cluster. Positive Cluster of immigrants. The cluster is located in Bexiga neighborhood that is famous by its traditional Italian restaurants.

BOM RETIRO

Korean Restaurant cluster. Positive cluster of immigrants. This cluster has the greatest number of restaurants. Virtually visiting the region it's possible to find other Korean venues in the neighborhood.

LIBERDADE

Japanese, Chinese and Korean Restaurant cluster. Positive Cluster of immigrants. The neighborhood is famous by its concentration of Japanese, Korean and Chinese stores and restaurants.

REPUBLICA

Peruvian Restaurant cluster. Possible positive cluster of immigrants. Checking the region virtually, it's a very commercial area with several types of business. But there is a concentration of venues targeting Peruvian and other Latin American public.

MORUMBI

Italian Restaurant cluster. False cluster of immigrants. The cluster is located over a shopping mall and the restaurant concentration is due to its food corner.

PINHEIROS

Italian Restaurant cluster. Possible False cluster of immigrants. The cluster is located close to a wealthy region with high concentration of all types of restaurants.

SEARCH AREA

The search area was limited by district area. Few districts in the city had elongated format, creating few uncovered area in its edges. But for a better coverage in a future study in São Paulo or in another city with borough with nonuniform formats, it should be done with a different approach.

FOURSQUARE API

The limitation of Foursquare API to 100 results per query makes it more difficult to scan the whole region at once and it also caps the result in regions with more restaurants concentration.

For a future study, one could run requests in each region per category of restaurant in order to maximize the number of results.

Or, for a specific nationality, the search could be done only for this restaurant category over the entire region.

CONCLUSION

Using Foursquare restaurant data and Clustering techniques to be a coarse identification of neighborhoods with more concentration of specific immigrants seemed to be possible showing some positive identification.