Clustering Neighborhoods from São Paulo based on venues categories to find best region of the city for future investor in the gatronomic business

Luís Galvão

Feb, 2021

1 – Introduction

1.1 – Background

With over 12,500 restaurants serving cuisines from all over the world, São Paulo is today one of the most exciting gastronomic capitals in the world [1]. Duo to its cosmopolitan status and concentration of distinct cultures, São Paulo appears as one of the most relevant gastronomic investiments scenarios of the atuality, been reference in Italian, Japanese and French food in the Latin America.

In this project i'll propose an analysis of the São Paulo neigborhoods based on the categorie of most common venues for each specific neighborhood. That can guide future stakeholders who dosen't know the city landscape and want to install restaurant business in well locate places.

1.2 – Problem

Duo the big size of the city of São Paulo (over 12 million), it's dificult to new investors in the food business to understand where are the best neighborhoods for future instalations.

My propose is to divide the neighborhoods in clusters (using machine learning algorithm Kmeans) based on every location most common venues.

The visualization (using Foursquare API and folium library) of this clusters over the map of the city of São Paulo can serve as a guide for future investors who want to understand where are the best neighborhoods for investiments in restaurants.

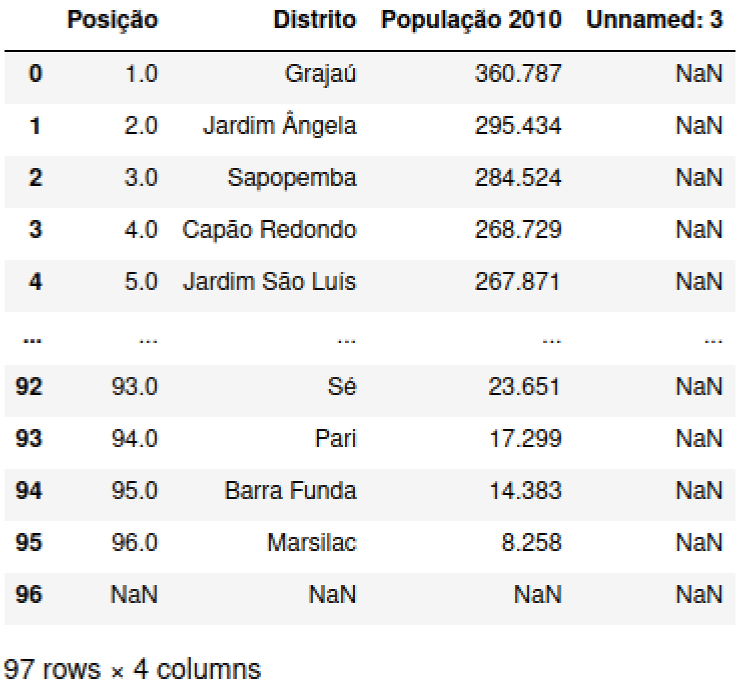
2 – Data

To execute the clustering technique i will need a DataFrame containing the name of each neigborhood in São Paulo, as well as every exact location (latitude and longitude) for each neighborhood – thats how the data can be used to performe the clustering analysis, create maps and retrieve Foursquare requests.

There is no such data avaiable, but i will extract these information from diferent sources, to finally compose the right DataFrame sp_neighborhood.

2.1 – Data Source and Cleaning

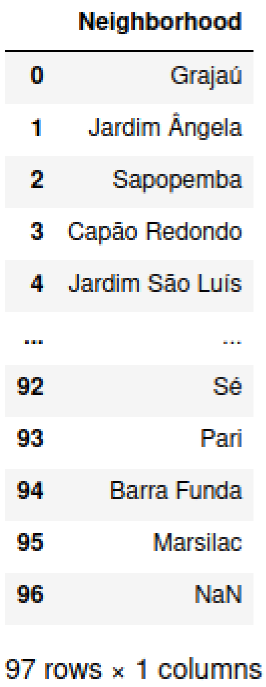
The source of the neighborhoods names is a Wikipedia page listing every neighborhood and its respective population. The extraction was made by using Panda library. The population, as well as NaN values, will be dropped in the cleaning data section, as its not necessary for the analysis.



	Posição	Distrito	População 2010	Unnamed: 3
0	1.0	Grajaú	360.787	NaN
1	2.0	Jardim Ângela	295.434	NaN
2	3.0	Sapopemba	284.524	NaN
3	4.0	Capão Redondo	268.729	NaN
4	5.0	Jardim São Luis	267.871	NaN
...	...	...	...	...
92	93.0	Sé	23.651	NaN
93	94.0	Pari	17.299	NaN
94	95.0	Barra Funda	14.383	NaN
95	96.0	Marsilac	8.258	NaN
96	NaN	NaN	NaN	NaN

97 rows × 4 columns

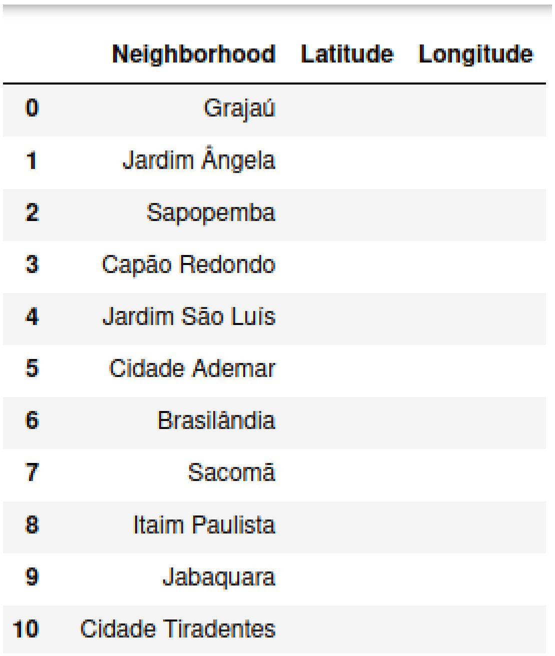
----------->



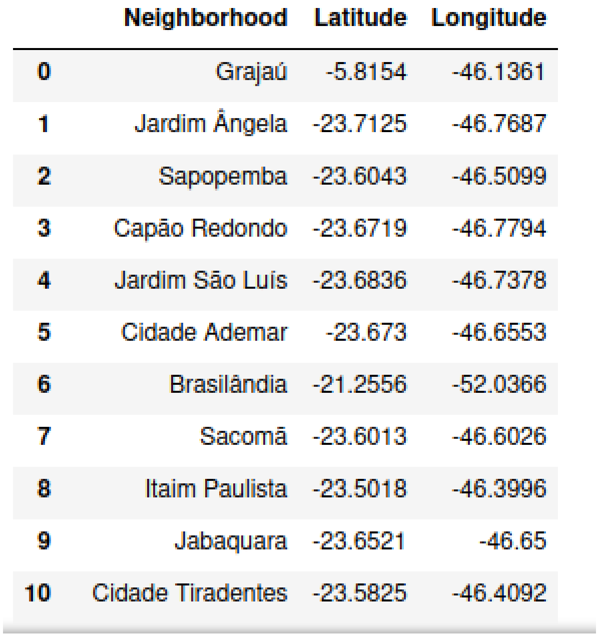
	Neighborhood
0	Grajaú
1	Jardim Ângela
2	Sapopemba
3	Capão Redondo
4	Jardim São Luís
...	...
92	Sé
93	Pari
94	Barra Funda
95	Marsilac
96	NaN

97 rows × 1 columns

After the Wikipedia **extraction and cleaning**, two more columns were added. Then i performed a loop trough all the neighborhood, retriving Coordinates from Geolocator and filling the sp_neighborhood DataFrame with the respective Latitude and Longitude for each neighborhood.



	Neighborhood	Latitude	Longitude
0	Grajaú		
1	Jardim Ângela		
2	Sapopemba		
3	Capão Redondo		
4	Jardim São Luís		
5	Cidade Ademar		
6	Brasilândia		
7	Sacomã		
8	Itaim Paulista		
9	Jabaquara		
10	Cidade Tiradentes		

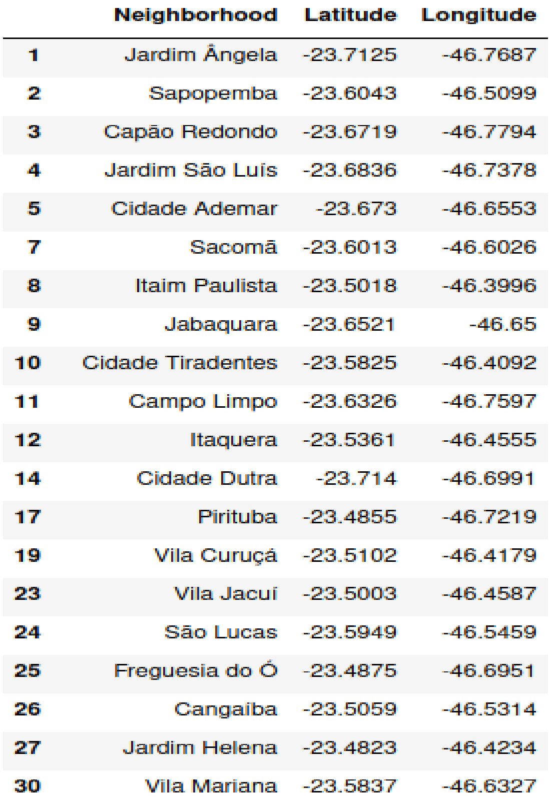


	Neighborhood	Latitude	Longitude
0	Grajaú	-5.8154	-46.1361
1	Jardim Ângela	-23.7125	-46.7687
2	Sapopemba	-23.6043	-46.5099
3	Capão Redondo	-23.6719	-46.7794
4	Jardim São Luís	-23.6836	-46.7378
5	Cidade Ademar	-23.673	-46.6553
6	Brasilândia	-21.2556	-52.0366
7	Sacomã	-23.6013	-46.6026
8	Itaim Paulista	-23.5018	-46.3996
9	Jabaquara	-23.6521	-46.65
10	Cidade Tiradentes	-23.5825	-46.4092

With those coordinates the neighborhoods can be visualized on the map (from folium library) as well as be used to retrieve Foursquare caracteristics of the location (related to venues).

After the geolocator loop, some of the entrys were wrong (pointing to other brazilian locations with the same name), and i filtered the columns Latitude and Longitude to only include locations inside the limits of the territory of the city of São Paulo.

After this, the DataFrame is read to be used for clustering techniques, maps and near venues analysis.



	Neighborhood	Latitude	Longitude
1	Jardim Ângela	-23.7125	-46.7687
2	Sapopemba	-23.6043	-46.5099
3	Capão Redondo	-23.6719	-46.7794
4	Jardim São Luís	-23.6836	-46.7378
5	Cidade Ademar	-23.673	-46.6553
7	Sacomã	-23.6013	-46.6026
8	Itaim Paulista	-23.5018	-46.3996
9	Jabaquara	-23.6521	-46.65
10	Cidade Tiradentes	-23.5825	-46.4092
11	Campo Limpo	-23.6326	-46.7597
12	Itaquera	-23.5361	-46.4555
14	Cidade Dutra	-23.714	-46.6991
17	Pirituba	-23.4855	-46.7219
19	Vila Curuçá	-23.5102	-46.4179
23	Vila Jacuí	-23.5003	-46.4587
24	São Lucas	-23.5949	-46.5459
25	Freguesia do Ó	-23.4875	-46.6951
26	Cangaiba	-23.5059	-46.5314
27	Jardim Helena	-23.4823	-46.4234
30	Vila Mariana	-23.5837	-46.6327

3 – Methodology section

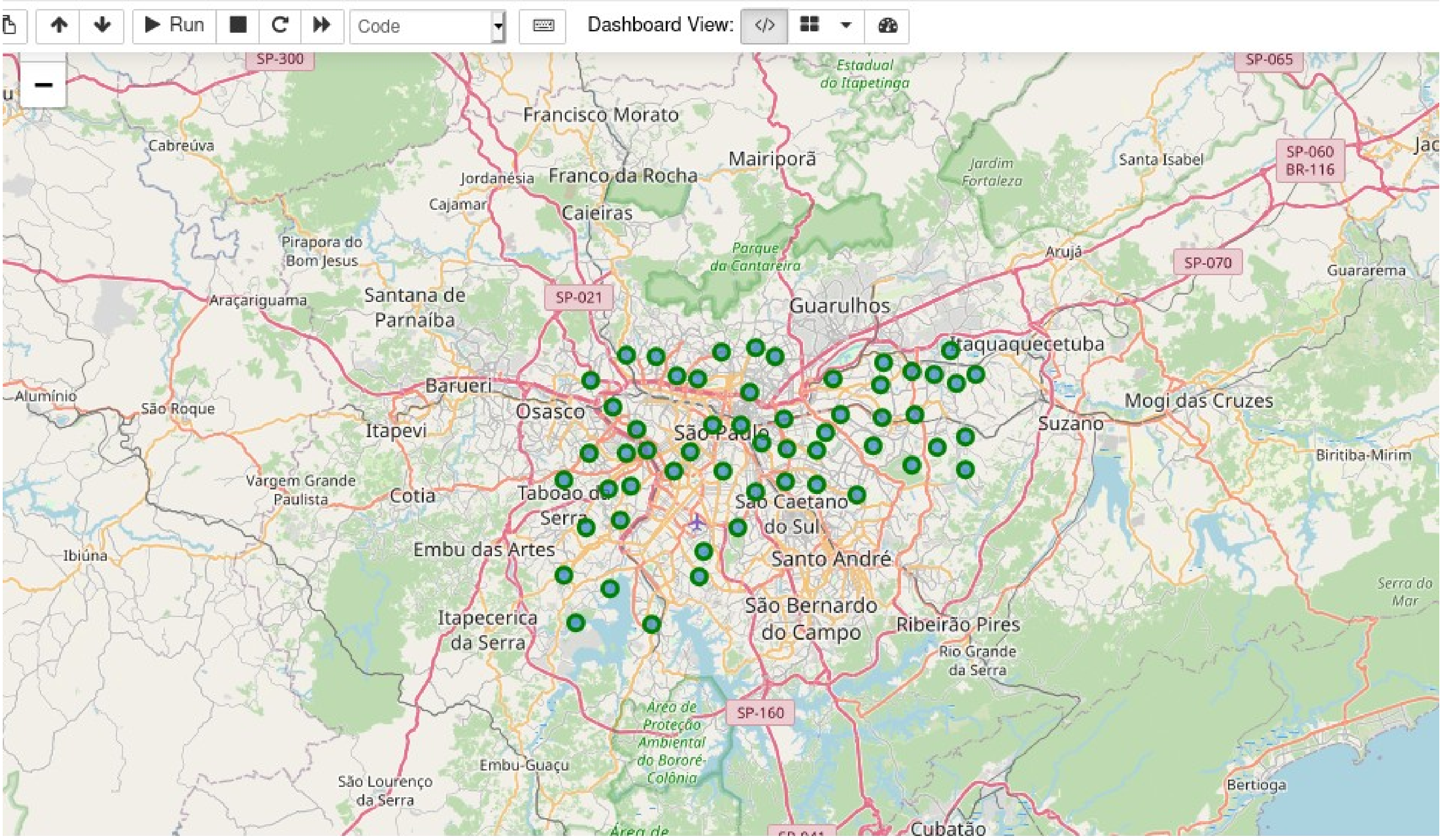
In this section i will use the Folium library to create two maps (with all the neighborhoods coordinates and with different clusters of the neighborhoods).

Initially, the first map will inform about the quality of the data.

Then i will use the Foursquare API to retrieve information about the neighborhoods and segment then based on machine learning clustering algorithm Kmeans, and apply the prediction on the second map.

3.1 – Vizualising data on top of the São Paulo’s map

After cleaning the data, this is the final map of São Paulo with neighborhoods superimposed on top, using Folium and the refined coordinates from Geolocator.



3.2 - Utilizing the Foursquare API to explore the neighborhoods relevant venues

To perform the clustering techinique we need to retrieve the data (venues for every location) from the API and process the DataFrame.

To retrieve all the information for every neighborhood i create a function to repeat the same process to all the neighborhoods in São Paulo. The order of the process is:

- Get the neighborhood's latitude and longitude values (geocoder loop)

- Get the top 100 venues that are in every Neighbourhood within a radius of 500 meters (Foursquare API)

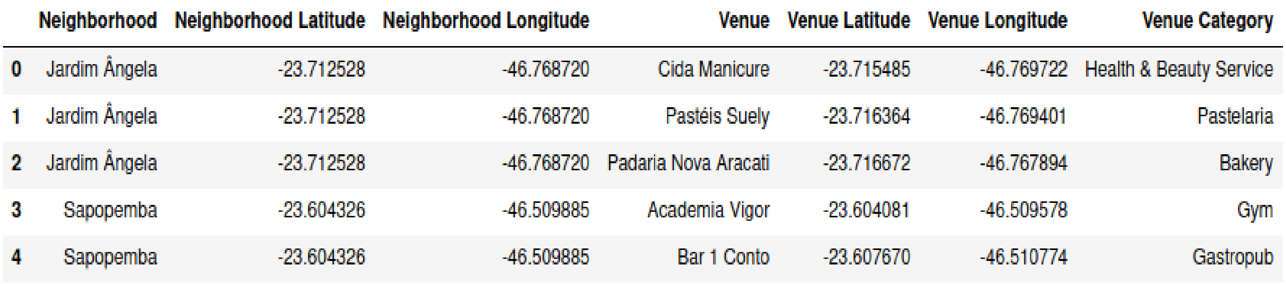
- Create the GET request URL

- Send the GET request and examine the resutls

- Clean the json and structure it into a _pandas_ dataframe

- Analizing data

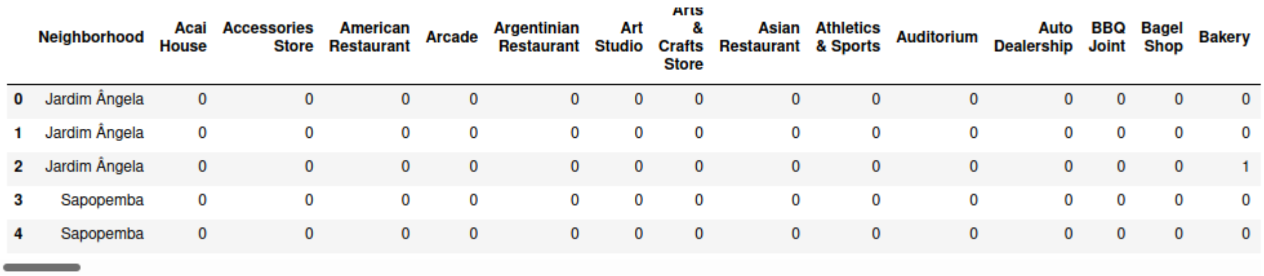
The resulting DataFrame include all the 1019 venues:



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Jardim Ângela	-23.712528	-46.768720	Cida Manicure	-23.715485	-46.769722	Health & Beauty Service
1	Jardim Ângela	-23.712528	-46.768720	Pastéis Suely	-23.716364	-46.769401	Pastelaria
2	Jardim Ângela	-23.712528	-46.768720	Padaria Nova Aracati	-23.716672	-46.767894	Bakery
3	Sapopemba	-23.604326	-46.509885	Academia Vigor	-23.604081	-46.509578	Gym
4	Sapopemba	-23.604326	-46.509885	Bar 1 Conto	-23.607670	-46.510774	Gastropub

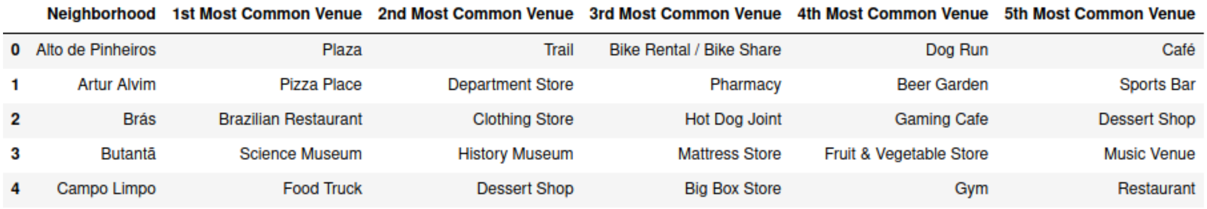
3.3 – Analyzing Each Neighborhood and preparing with one_hot_encoding for clustering the data

After apply the one_hot_encoded method to prepare the data for clustering, the table is:



	Neighborhood	Acai House	Accessories Store	American Restaurant	Arcade	Argentinian Restaurant	Art Studio	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auditorium	Auto Dealership	BBQ Joint	Bagel Shop	Bakery
0	Jardim Ângela	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Jardim Ângela	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Jardim Ângela	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	Sapopemba	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Sapopemba	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Then grouping rows by neighborhood and by taking the mean of the frequency of occurrence of each category, converting into a _pandas_ dataframe and creating dataframe for top 5 venues for each neighbourhood for further analysis. We got:

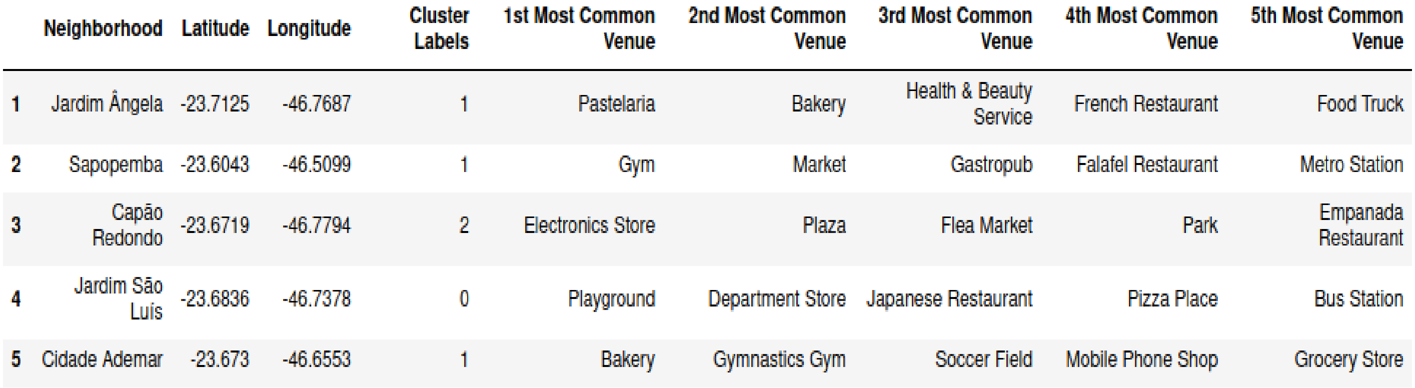


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alto de Pinheiros	Plaza	Trail	Bike Rental / Bike Share	Dog Run	Café
1	Artur Alvim	Pizza Place	Department Store	Pharmacy	Beer Garden	Sports Bar
2	Brás	Brazilian Restaurant	Clothing Store	Hot Dog Joint	Gaming Cafe	Dessert Shop
3	Butantã	Science Museum	History Museum	Mattress Store	Fruit & Vegetable Store	Music Venue
4	Campo Limpo	Food Truck	Dessert Shop	Big Box Store	Gym	Restaurant

3.4 – Applying clustering Techinique (Kmeans Algorithm)

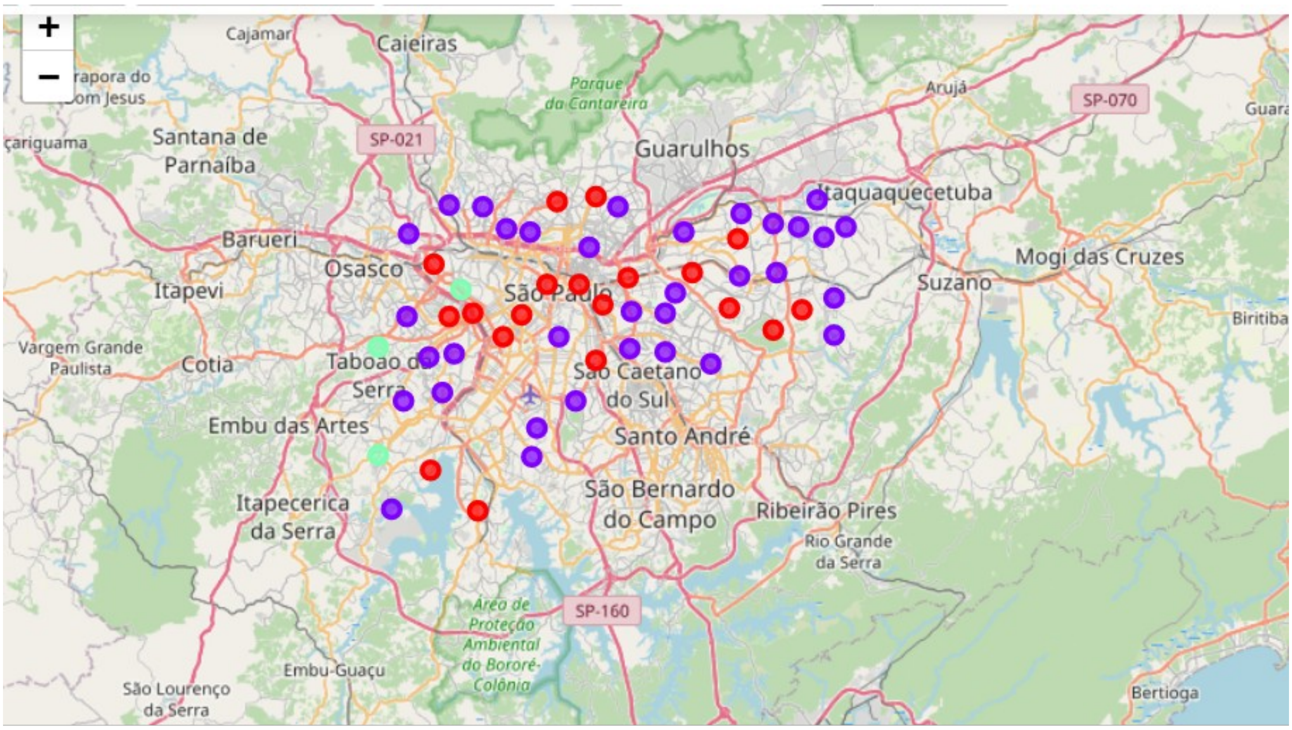
After apply the one_hot_encoding (before creating the dataframe above) we got the right table to perform the Kmeans clustering techinique (machine learning algorithm), and select 3 clusters for the final result.

So, before the map vizualisation, i create a new dataframe that includes the cluster as well as the top 5 venues for each neighborhood, só the clusters can be identified on the top of the map.



	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Jardim Ângela	-23.7125	-46.7687	1	Pastelaria	Bakery	Health & Beauty Service	French Restaurant	Food Truck
2	Sapopemba	-23.6043	-46.5099	1	Gym	Market	Gastropub	Falafel Restaurant	Metro Station
3	Capão Redondo	-23.6719	-46.7794	2	Electronics Store	Plaza	Flea Market	Park	Empanada Restaurant
4	Jardim São Luís	-23.6836	-46.7378	0	Playground	Department Store	Japanese Restaurant	Pizza Place	Bus Station
5	Cidade Ademar	-23.673	-46.6553	1	Bakery	Gymnastics Gym	Soccer Field	Mobile Phone Shop	Grocery Store

3.5 – Visualization of the clusters on top of the map of São Paulo



red – cluster0

purple – cluster1

green – cluster2

3.6 – Examining the cluster to label then accordly

Examining each cluster and determine the discriminating venue categories that distinguish each cluster. Based on the defining categories, we can then assign a properly label to each cluster – and that can solve the problems of restaurant investor and stakeholders with interest in doing business in São Paulo.

- Cluster 0: 1st Most Common Venue: Restaurant; 2nd Most Common: Bar; Venue 3rd Most Common Venue: Restaurant

- Cluster 1: 1st Most Common Venue: Bakery; 2nd Most Common: Bakery; 3rd Most Common Venue: Restaurant

- Cluster 2: 1st Most Common Venue: Plaza; 2nd Most Common: Plaza Venue 3rd Most Common Venue: Flea Market

After analysis of each clusters most common venues, i came up with those label, to help future investor on the gastronomic sector to understand the city distribution.

- BarAndRestaurant_cluster = cluster0
- Bakery_cluster = cluster1
- Hotel_cluster = cluster2

4 - Results section

After the analysis we came up with 3 different clusters that can help future investor to understand the citys panoram.

The BarAndRestaurant cluster is located onto the center of the city, or in the central latitude of the city, and is related to location with high concentration of restaurants and bars, and great opportunites for this kind of investment duo its large number of consumers.

The Bakery_cluster is located around the center of the city, where there is more residential neighborhoods, with more business related to this kind of neighborhood, like bakerys, desert companys and markets. This is a good opportunity for investor in this kind of familiar food business.

The Hotel_cluster reflects an area far from the center of the city, with less comerce related to gastronomic business and more opportunities for hotel business.

5 - Discussion section

I noted that the clustering technique came out with two relevant clusters for gastronomic investiments, but the dierences between this clusters represent the difference between investors in this field.

One cluster is related to Restaurants and bars, the more logic option for one who wants to invest in the gastronomic scenario in one of the most important citys for this.

But, the second cluster also shown an opportunit for investor in the gastronomic area: in a more familiar and small frame business like bakerys and deserts – a great niche as well.

My recomendation is to select some location in the BarAndRestaurant cluster for those who intend to enter the gastronomic business scenario of the city of São Paulo.

6 – Conclusion

My conclusion is that is possible to guide future investors using clustering techniques, map visualization, and the Fourquare API to made possible to understand its applications better.

I got a pretty good description of the São Paulo gastronomic scene, and thats is a real asset for real business stakeholders, mainly those who want to enter the competitive scenario but dosen’t have good knowledge about the city as a whole.

My suggestion for next steps is to add new features for the analysis, considering caracteristics of the neighborhood as HDI and others.