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

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 Luís	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	Cangaíba	-23.5059	-46.5314
27	Jardim Helena	-23.4823	-46.4234
30	Vila Mariana	-23.5837	-46.6327