Coursera_Capstone

The Battle of the Neighborhoods in São Paulo City, Brazil.

Final project of the Coursera IBM Data Science Course

1. Introduction: Business Problem

In this project we will try to find another ways to make a new marketing campains by a segmentation method. This report will be targeted to people interested in suggest marketing contents, produce advertising and etc, to the people tha lives in the Districts of São Paulo's Capital, Brazil.

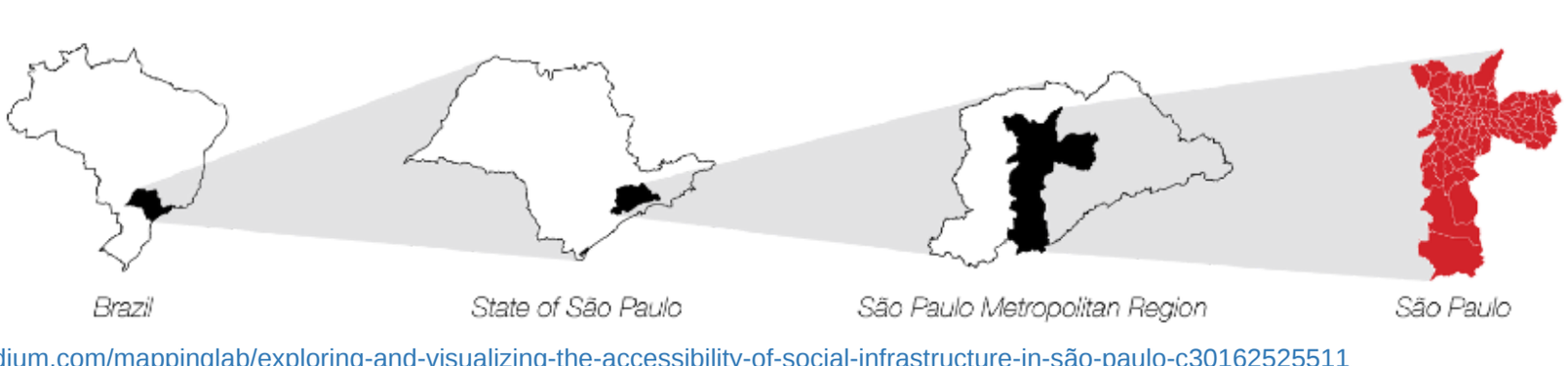
All entrepreneurs looks a way for increase the selling. Therefore, diferent marketing strategies are made of. We are talking about keeping a good presence in the social medias, to optimize the mechanisms of search, investing in custumer service. All of those are so good actions and also bring great results, but We are able to make more.

We know that are lots of venues in the São Paulo City, therefore We will try figure out the business profile of each borough. Assuming that São Paulo is an alpha global city, We must to know what are the subject most reached by our costumers.

And finally, help them to make better buying decisions. We will use all power of data science to get a success we need.

Districts of São Paulo City

We are talking about 96 Districts



https://medium.com/mappinglab/exploring-and-visualizing-the-accessibility-of-social-infrastructure-in-são-paulo-c30162525511

2.Data

The Data that We will use has been found at Wikipedia.com.br, and It exists since 2010.

So We must make a Web Scraping from page following this steps bellow.

1. Grabbing the HTML Content from the URL
2. Parsing the HTML content - BeautifulSoup
3. Extructuring the content into a Data Frame - Pandas
4. Converting and saving the Data like a JSON Archieve
5. Importing that in our Jupyter Notebook

How?

For that We will use some libs in Python language.

Like that:

```
import time
import requests
import pandas as pd
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.firefox.options import Options
from selenium.webdriver.common.by import By
import json
```

URL: https://pt.wikipedia.org/wiki/Lista_dos_distritos_de_São_Paulo_por_população

	Districts	Population
0	Grajaú	360.787
1	Jardim Ângela	295.434
2	Sapopemba	284.524
3	Capão Redondo	268.729
4	Jardim São Luís	267.871
...	...	...
92	Sé	23.651
93	Pari	17.299
94	Barra Funda	14.383
95	Marsilac	8.258
96	None	NaN

97 rows × 2 columns

3. Methodology

This survey is a Clustering problem. In this case We will use a machine learning method.

K- Means machine learning method will be used for acquiring information concerning the venues located close to each area the data platform Foursquare was used. In this project our purpose is to create an appropriate number of clusters based on the Geolocalization and the quantity and the kind of venues located in each District in São Paulo City.

In the end of this project we will be able to see the different types of public that we may consider important according to the kind of business We have. Also, We will figure out what kind of destinations each costumer has an interest. In general we acquire a clean perception of the most visitable venues in each District. From that We Will see what kind of venues we may offer like recommendations to each client.

---

4. Preparing Data

4.1 Manipulation the dataframe to discover the coordinates of the districts. For that we must make a tratament of our Dataset.

The first step is to eliminate the NaN datas. The second one consists on create a new column with the Address. (This is so important to get the coordinates of each Distric of São Paulo)

	Districts	Population	Address
0	Grajaú	360.787	Grajaú São Paulo Brasil
1	Jardim Ângela	295.434	Jardim Ângela São Paulo Brasil
2	Sapopemba	284.524	Sapopemba São Paulo Brasil
3	Capão Redondo	268.729	Capão Redondo São Paulo Brasil
4	Jardim São Luís	267.871	Jardim São Luís São Paulo Brasil
...	...	...	...
91	Jaguara	24.895	Jaguara São Paulo Brasil
92	Sé	23.651	Sé São Paulo Brasil
93	Pari	17.299	Pari São Paulo Brasil
94	Barra Funda	14.383	Barra Funda São Paulo Brasil
95	Marsilac	8.258	Marsilac São Paulo Brasil

96 rows × 3 columns

The Coordinates had been gotten by The Library GeoPy, that fetchs the Latitude and Longitude from the Address, how We said lastly.

from geopy.geocoders import Nominatim

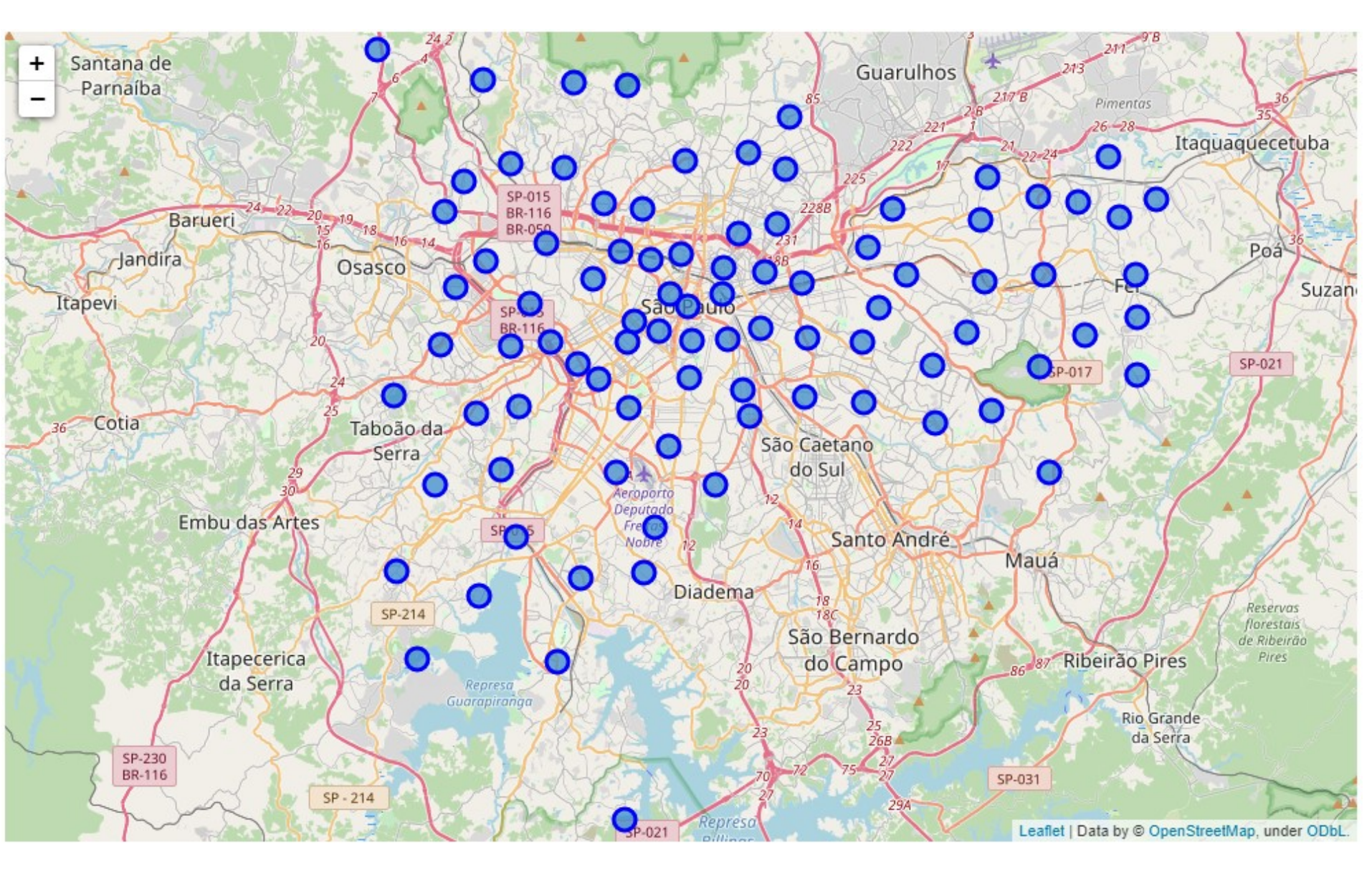
	Districts	Population	Address	Latitude	Longitude
0	Grajaú	360.787	Grajaú São Paulo Brasil	-23.785907	-46.665197
1	Jardim Ângela	295.434	Jardim Ângela São Paulo Brasil	-23.712528	-46.768720
2	Sapopemba	284.524	Sapopemba São Paulo Brasil	-23.604326	-46.509885
3	Capão Redondo	268.729	Capão Redondo São Paulo Brasil	-23.671903	-46.779435
4	Jardim São Luís	267.871	Jardim São Luís São Paulo Brasil	-23.683573	-46.737762
...	...	...	...	...	...
91	Jaguara	24.895	Jaguara São Paulo Brasil	-23.507446	-46.755315
92	Sé	23.651	Sé São Paulo Brasil	-23.550651	-46.633382
93	Pari	17.299	Pari São Paulo Brasil	-23.532976	-46.615849
94	Barra Funda	14.383	Barra Funda São Paulo Brasil	-23.525462	-46.667513
95	Marsilac	8.258	Marsilac São Paulo Brasil	-23.937142	-46.710230

96 rows × 5 columns

5. Visualizing Data

5.1 Now We will utilize the Foursquare API to figure out the Latitude and Longitute. The next step is to plot the Maps through the Folium Library

FOURSQUARE



How We can see, each blue point represents each District of our dataset. But It isn't enough. We will need to figure out all kind of Venues we can find from each reference points.

5.2 It is awesome thinking the Foursquare API is able to return each name and each category of the business places.

	name	categories	lat	lng
0	Ofner	Dessert Shop	-23.669999	-46.711387
1	Parque da Juventude Città Di Maróstica	Skate Park	-23.697099	-46.549651
2	Parque Severo Gomes	Park	-23.646914	-46.718936
3	Jardim Botânico de São Paulo	Botanical Garden	-23.639119	-46.626204
4	Sonda Supermercados	Supermarket	-23.660019	-46.571291
5	Palácio Tangará	Hotel	-23.632152	-46.722919
6	Associação Atlética Banco do Brasil (AABB)	Athletics & Sports	-23.649957	-46.752272
7	Bacio di Latte	Ice Cream Shop	-23.622474	-46.698722
8	Záffari Hipermercado	Big Box Store	-23.632101	-46.737020
9	Agello Cucina	Italian Restaurant	-23.626143	-46.669758

5.3 We are talking about 100 places returned by Foursquare API. There are 290 uniques categories.

	Neighborhood	sp Latitude	sp 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

	Neighborhood	sp Latitude	sp 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

5.4 How we know, To work with Machine Learning, It is necessary to convert the categorical variables into numerical variables.

For that we have used The One Hot Encoding method.

	Neighborhood	Acai House	Accessories Store	African Restaurant	American Restaurant	Antique Shop	Arcade	Argentinian Restaurant	Art Gallery	Art Museum	...	Video Store	Vietnamese Restaurant	Volleyball Court	Warehouse Store
0	Jardim Ângela	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
2	Jardim Ângela	0	0	0	0	0	0	0	0	0	...	0	0	0	0
3	Sapopemba	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
95	Itaim Paulista	0	0	0	0	0	0	0	0	0	...	0	0	0	0
96	Itaim Paulista	0	0	0	0	0	0	0	0	0	...	0	0	0	0
97	Jabaquara	0	0	0	0	0	0	0	0	0	...	0	0	0	0
98	Jabaquara	0	0	0	0	0	0	0	0	0	...	0	0	0	0
99	Jabaquara	0	0	0	0	0	0	0	0	0	...	0	0	0	0

100 rows × 291 columns

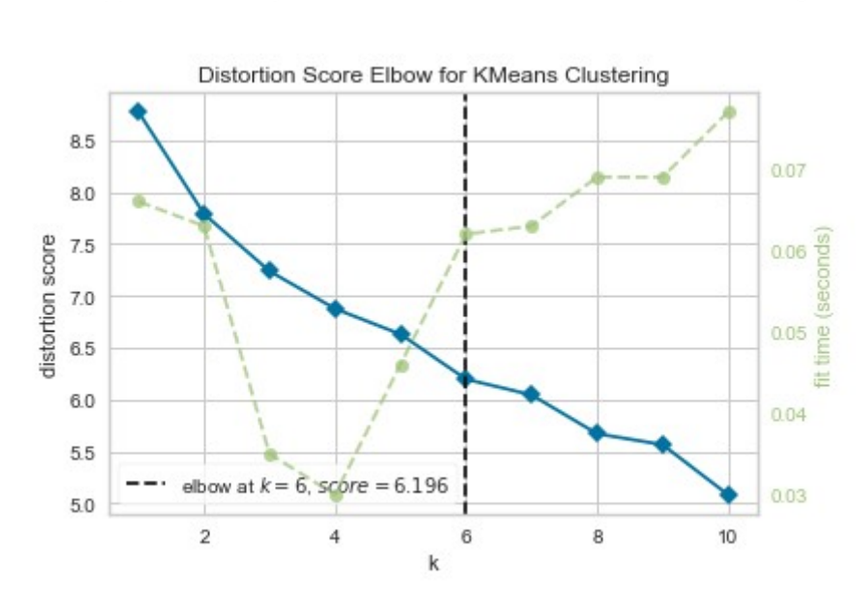
5.5 In case of this business problem We want to discover the most common Venue of São Paulo's District. All of this driven by location.

	Neighborhood	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
0	Alto de Pinheiros	Plaza	Dog Run	Bike Rental / Bike Share	Café	Gym / Fitness Center	Trail	Spa	Department Store	Dessert Shop	French Restaurant
1	Anhanguera	Gym / Fitness Center	Plaza	Ice Cream Shop	Bakery	Pizza Place	Convenience Store	Grocery Store	Pet Store	Fish & Chips Shop	Escape Room
2	Aricanduva	Bakery	Clothing Store	Gym / Fitness Center	Candy Store	Grocery Store	Yoga Studio	Flea Market	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
3	Artur Alvim	Pizza Place	Department Store	Newsstand	Grocery Store	Beer Garden	Bar	Bakery	Gymnastics Gym	Pastelaria	Chinese Restaurant
4	Barra Funda	Music Venue	Restaurant	Brazilian Restaurant	Café	Sandwich Place	Chocolate Shop	Country Dance Club	Nightclub	Office	Museum

6. Modeling

6.1 It's interesting We figure out the Best K number for a Clustering Algorithm. In this case, We are using K-Means algorithm and the number of K represents the best quantities of Clusters may use for.

So, We might have used the Elbow Method. This Approach is very common used with the K-means Clustering. So We have got that K is equal 6.

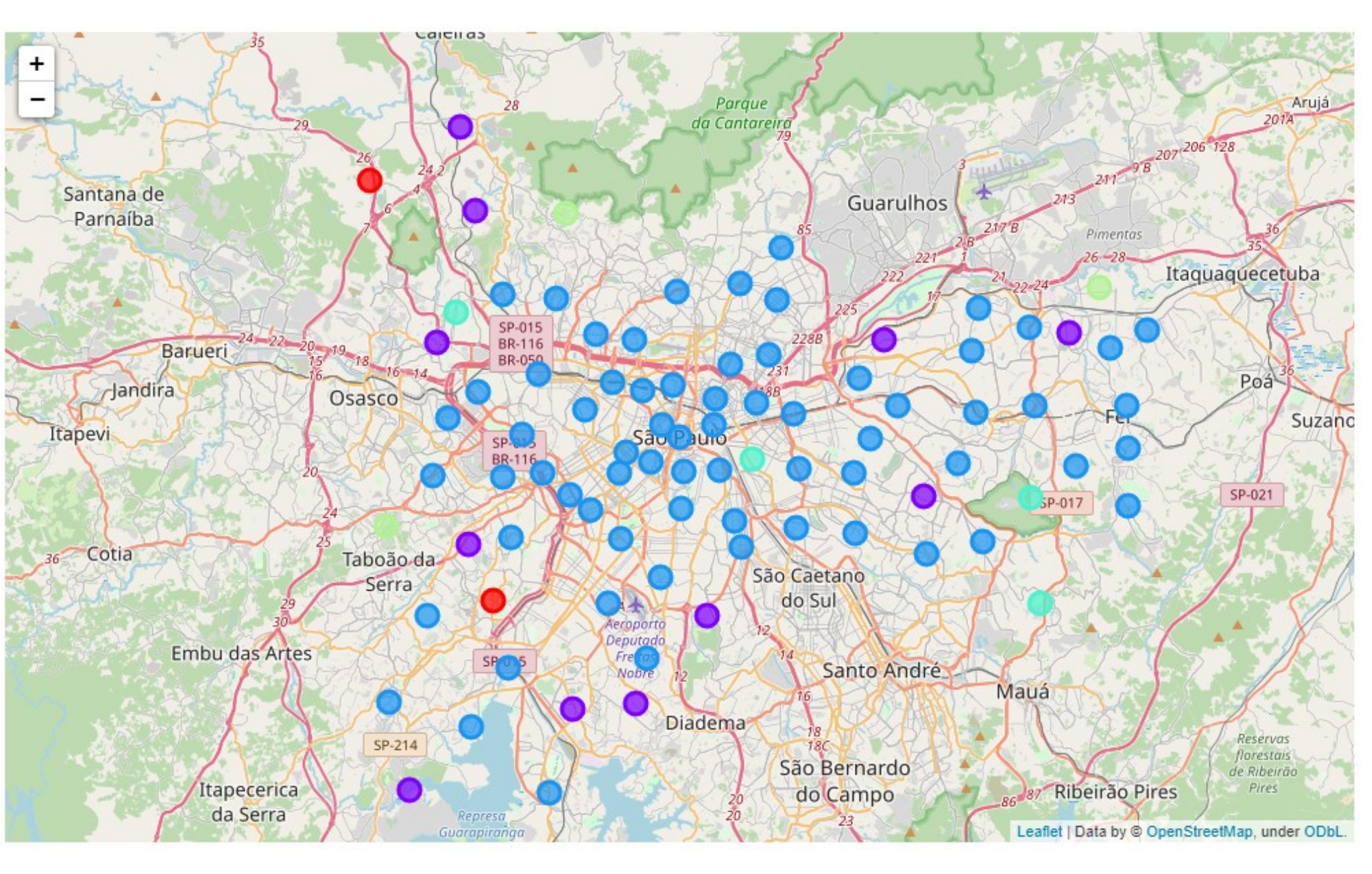



```
In [ ]: # set number of clusters
kclusters = 6

sp_grouped_clustering = sp_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(sp_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

	Districts	Population	Address	Latitude	Longitude	Cluster Labels	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
0	Grajaú	360.787	Grajaú São Paulo Brasil	-23.785907	-46.665197	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Jardim Ângela	295.434	Jardim Ângela São Paulo Brasil	-23.712528	-46.768720	1.0	Bakery	Health & Beauty Service	Pastelaria	Flea Market	Escape Room	Event Space	Falafel Restaurant	Farmers Market
2	Sapopemba	284.524	Sapopemba São Paulo Brasil	-23.604326	-46.509885	2.0	Breakfast Spot	Gastropub	Market	Gym	Grocery Store	Flea Market	Food	Pharmacy
3	Capão Redondo	268.729	Capão Redondo São Paulo Brasil	-23.671903	-46.779435	2.0	Park	Soccer Field	Flea Market	Pharmacy	Food Stand	Food Court	Food & Drink Shop	Food Truck
4	Jardim São Luís	267.871	Jardim São Luís São Paulo Brasil	-23.683573	-46.737762	2.0	Japanese Restaurant	Playground	Pizza Place	Department Store	Bus Station	Yoga Studio	Food	Farmers Market

6.2 After merging the data, We had needed to make a new tratament of the dataset.

6.3 Now We have a new map with the clusters separated by colors. It's simply wonderful! The 6 clusterings separated by Colors!



The Clustering consists in:

Cluster 1

```
sp_merged.loc[sp_merged['Cluster Labels'] == 0, sp_merged.columns[[1] + list(range(5, sp_merged.shape[1]))]]
```

	Population	Cluster Labels	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
34	127.015	0	Gym / Fitness Center	Pet Store	Market	Gym	Residential Building (Apartment / Condo)	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
73	65.859	0	Gym / Fitness Center	Plaza	Ice Cream Shop	Bakery	Pizza Place	Convenience Store	Grocery Store	Pet Store	Fish & Chips Shop	Escape Room

Cluster 2

```
sp_merged.loc[sp_merged['Cluster Labels'] == 1, sp_merged.columns[[1] + list(range(5, sp_merged.shape[1]))]]
```

	Population	Cluster Labels	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
1	295.434	1	Bakery	Health & Beauty Service	Pastelaria	Flea Market	Escape Room	Event Space	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
5	266.681	1	Music Venue	Fried Chicken Joint	Bakery	Gymnastics Gym	Flower Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market
15	184.818	1	Bakery	Convenience Store	Grocery Store	Gym / Fitness Center	Farmers Market	Pet Store	Arts & Crafts Store	Brazilian Restaurant	Italian Restaurant	Diner
26	136.623	1	Bakery	Chocolate Shop	Gym / Fitness Center	Café	Pharmacy	Food Stand	Food Court	Food & Drink Shop	Food	Food Truck
28	131.183	1	Brazilian Restaurant	Nature Preserve	Snack Place	Bakery	French Restaurant	Frame Store	Food Truck	Fried Chicken Joint	Food Stand	Falafel Restaurant
42	109.088	1	Brazilian Restaurant	Bakery	Market	Food & Drink Shop	Farmers Market	Gym	Gym / Fitness Center	Paper / Office Supplies Store	Furniture / Home Store	Pizza Place
43	108.441	1	Bakery	Skate Park	Farmers Market	Grocery Store	Gourmet Shop	Food & Drink Shop	Gym / Fitness Center	Motorcycle Shop	Deli / Bodega	Pizza Place
50	100.713	1	Brazilian Restaurant	Dessert Shop	Gym / Fitness Center	Pharmacy	Convenience Store	Bakery	Gym	Farmers Market	Chocolate Shop	Candy Store
57	92.081	1	Bakery	Fruit & Vegetable Store	Grocery Store	Gym	Japanese Restaurant	Breakfast Spot	Pizza Place	Dessert Shop	Convenience Store	Market
59	89.622	1	Bakery	Clothing Store	Gym / Fitness Center	Candy Store	Grocery Store	Yoga Studio	Flea Market	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
68	80.187	1	Bakery	Convenience Store	Gym / Fitness Center	Gymnastics Gym	Dessert Shop	Chocolate Shop	Plaza	Food Stand	Food Court	Food Truck
91	24.895	1	Bakery	Gym	Soccer Field	Seafood Restaurant	Flower Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market

Cluster 3

```
sp_merged.loc[sp_merged['Cluster Labels'] == 2, sp_merged.columns[[1] + list(range(5, sp_merged.shape[1]))]]
```

	Population	Cluster Labels	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
2	284.524	2	Breakfast Spot	Gastropub	Market	Gym	Grocery Store	Flea Market	Food	Pharmacy	Japanese Restaurant	Falafel Restaurant
3	268.729	2	Park	Soccer Field	Flea Market	Pharmacy	Food Stand	Food Court	Food & Drink Shop	Food Truck	Food	Empanada Restaurant
4	267.871	2	Japanese Restaurant	Playground	Pizza Place	Department Store	Bus Station	Yoga Studio	Food	Farmers Market	Fast Food Restaurant	Fish & Chips Shop
7	247.851	2	Pharmacy	Bar	Brazilian Restaurant	Department Store	Cosmetics Shop	Clothing Store	Chocolate Shop	Farmers Market	Sports Bar	Gym Pool
8	224.074	2	Japanese Restaurant	Bakery	Food Truck	Bowling Alley	Dessert Shop	Gym / Fitness Center	Pizza Place	Brewery	Brazilian Restaurant	Food & Drink Shop
...	...	...	...	...	...	...	...	...	...	...	...	...
89	33.892	2	Korean Restaurant	Brazilian Restaurant	Women's Store	Dessert Shop	Café	Bar	Coffee Shop	Grocery Store	Bakery	Convenience Store
90	29.265	2	Brazilian Restaurant	Clothing Store	Pizza Place	Park	Burger Joint	Furniture / Home Store	Gaming Cafe	Café	Shoe Store	Sandwich Place
92	23.651	2	Brazilian Restaurant	Bookstore	Miscellaneous Shop	Cosmetics Shop	Bakery	Music Venue	Chocolate Shop	Japanese Restaurant	Arts & Crafts Store	Vegetarian / Vegan Restaurant
93	17.299	2	Clothing Store	Brazilian Restaurant	Middle Eastern Restaurant	Restaurant	Shopping Mall	Café	Falafel Restaurant	Warehouse Store	Bar	Hotel
94	14.383	2	Music Venue	Restaurant	Brazilian Restaurant	Café	Sandwich Place	Chocolate Shop	Country Dance Club	Nightclub	Office	Museum

70 rows × 12 columns

Cluster 4

```
sp_merged.loc[sp_merged['Cluster Labels'] == 3, sp_merged.columns[[1] + list(range(5, sp_merged.shape[1]))]]
```

	Population	Cluster Labels	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
21	143.992	3	Bar	Park	Brewery	Food	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Flower Shop	Food & Drink Shop
63	84.843	3	Bar	Chinese Restaurant	Gym / Fitness Center	Burger Joint	Diner	Snack Place	Flea Market	Food	Fast Food Restaurant	Fish & Chips Shop
72	68.258	3	Brazilian Restaurant	Bar	Planetarium	IT Services	Food & Drink Shop	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Flower Shop
76	63.133	3	Bar	Burger Joint	Mexican Restaurant	Japanese Restaurant	Gym / Fitness Center	Acai House	Chinese Restaurant	Middle Eastern Restaurant	Brewery	Snack Place

Cluster 5

```
sp_merged.loc[sp_merged['Cluster Labels'] == 4, sp_merged.columns[[1] + list(range(5, sp_merged.shape[1]))]]
```

	Population	Cluster Labels	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
6	264.918	4	Farmers Market	Food Truck	Big Box Store	Yoga Studio	Food & Drink Shop	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Flower Shop	Food
27	135.043	4	Food Truck	Brazilian Restaurant	Gym	Grocery Store	Restaurant	Big Box Store	Comfort Food Restaurant	Burger Joint	Pizza Place	Food Stand
51	100.164	4	Food Truck	Yoga Studio	Food	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market	Flower Shop	Food & Drink Shop

Cluster 6

```
sp_merged.loc[sp_merged['Cluster Labels'] == 5, sp_merged.columns[[1] + list(range(5, sp_merged.shape[1]))]]
```

	Population	Cluster Labels	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
37	118.797	5	Southern Brazilian Restaurant	Gym	Snack Place	Yoga Studio	Flower Shop	Falafel Restaurant	Farmers Market	Fast Food Restaurant	Fish & Chips Shop	Flea Market

7. Conclusion

7.1 Finally We can see the different categories of the Venues in our dataset. So, How can this study help us?

With those tools We will able to make recommendations to each public in each clustering. We may use the neighborhood, distance, the position in the map, the various combination of informations to suggest the best product or service to our clients.

It is a good way to construct a New marketing campain to your constumers. Dada help us to understand our public and get less mistakes than other market players.