IBM Data Science Final Project

Picking the right location for an Irish Pub in Sao Paulo using unsupervised Machine Learning

Raul Lima Alves
August, 2020

1 - Introduction

The IBM data science certification program, through Coursera, has its final step as a Capstone Project. The main goal is to choose a real-life problem and analyse it using machine learning techniques. An investor, before opening a new venue, should be able to analyze the data, grouping by the customers profile.

This report analyses the Sao Paulo geolocation data and try to define which locations are best for opening an Irish Pub. Unsupervised machine learning techniques are going to be used, aiming to cluster the venues data of the city.

2 - Data

The first data needed is the geolocation of each metro station. The metro station is a good index because it concentrates millions of people that use the public transport daily. The venues nearby indicates the types of costumers of specifics profiles that passes through that specific location. The metro locations were found at a Kagle dataset [1]. It contains the latitude and longitude of the main metro stations of Sao Paulo.

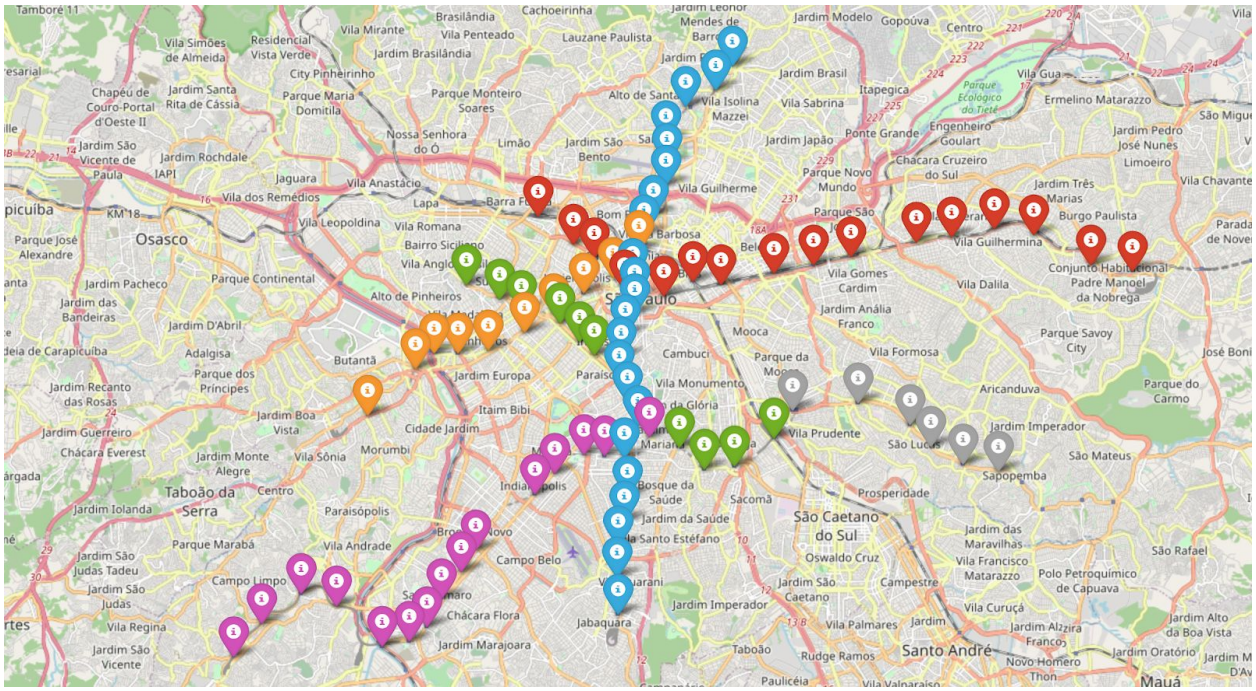
The venues location nearby each metro station were acquired using the Foursquare API.

3 - Methodology

The first step is preprocessing the geolocation data of the metro stations of Sao Paulo. The CSV data, acquired here [1] contains the following columns: Unnamed: 0, name, station, lat, lon, line and neigh. It is needed only the name, lat, lon and line columns. The other ones will be dropped.

The second step is plotting the metro stations in a map, using the Folium python library [2]. Folium is an important visualization library that is going to be used also in the other steps of this project.

The plot of the geolocation of the metro stations of Sao Paulo can be seen at the following image.



Some metro stations belongs to more than one metro line. For visualization purposes, the first line was chosen. We can see in the previous image that there are 6 metro stations present in the dataset used for this project.

The next step is, for all metro stations, acquire the nearby venues and their categories. The Foursquare API explore is used. It is accessed through a get request for each metro location. The general URL is the following: = *https://api.foursquare.com/v2/venues/explore?&client_id={CLIENT_ID}&client_secret={CLIENT_SECRET}&v={VERSION}&ll={LATITUDE},{LONGITUDE}&radius={RADUIS}&limit={LIMIT}*, where

- CLIENT_ID = The ID for the Foursquare account
- CLIENT_SECRET = The secret for the Foursquare account
- VERSION = The version of the API to be used
- LATITUDE = The latitude value of the venues to be acquired
- LONGITUDE = The longitude value of the venues to be acquired
- RADIUS = The radius, in meters, to perform the query

- LIMIT = The maximum number of venues for that location

In this project, the version used is 20180605, the radius is 1000 and the limit is 100 per location.

The get request returns a json containing the venues information among other meta data for the results of the query. The response should be parsed for grouping only the needed information.

For each metro station, given its latitude and longitude data, one different request to the Foursquare API should be performed. After acquiring the response, it should be parsed and count the occurrences of each venue type. For that purpose, the Pandas library [3] was used. Obviously, not all venues categories are present at the metro stations. Hence, the NaN occurrences should be replaced by 0. The following image shows the head of the grouped pandas Dataframe.

	name	lat	lon	line	Art Museum	Arts & Crafts Store	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	...	Irish Pub	Social Club	Swiss Restaurant	Herbs & Spices Store	High School	Bus Stop	Hawaiian Restaurant	Borek Place	Piadineria	Post Office
0	Aacd Servidor	-23.597825	-46.652374	['ilias']	2.0	2.0	2.0	1.0	1.0	2.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Adolfo Pinheiro	-23.650073	-46.704206	['ilias']	0.0	2.0	1.0	0.0	0.0	2.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Alto Da Boa Vista	-23.641625	-46.699434	['ilias']	0.0	1.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Alto Do Ipiranga	-23.602237	-46.612486	['verde']	0.0	1.0	1.0	0.0	1.0	3.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Ana Rosa	-23.581871	-46.638104	['azul', 'verde']	0.0	0.0	0.0	0.0	0.0	4.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

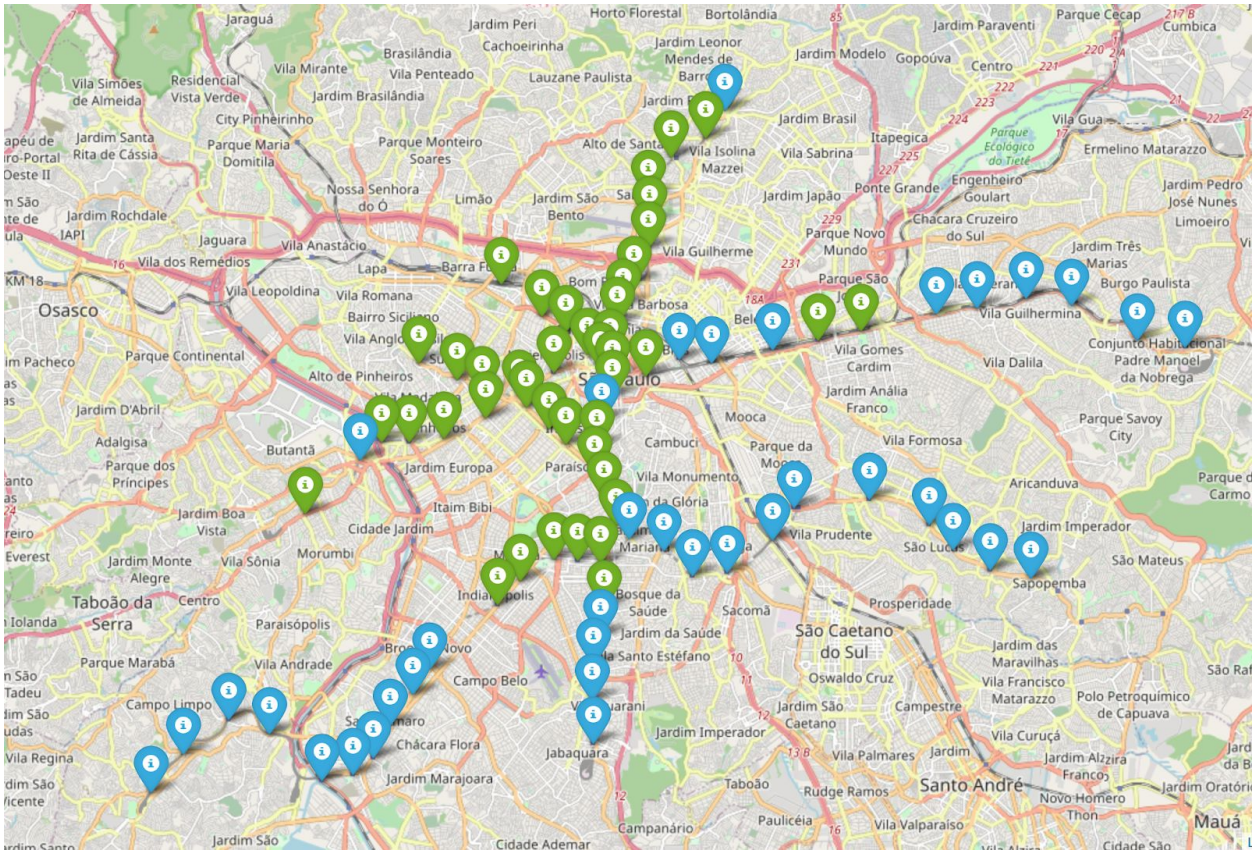
The next step is normalizing the occurrences per each venue category. The MinMax Scaler was used, located at the Scikit python library [4].

The last step is using K-means, which is an unsupervised machine learning algorithm. K-means was used with different cluster sizes, and the results will be shown at the next session.

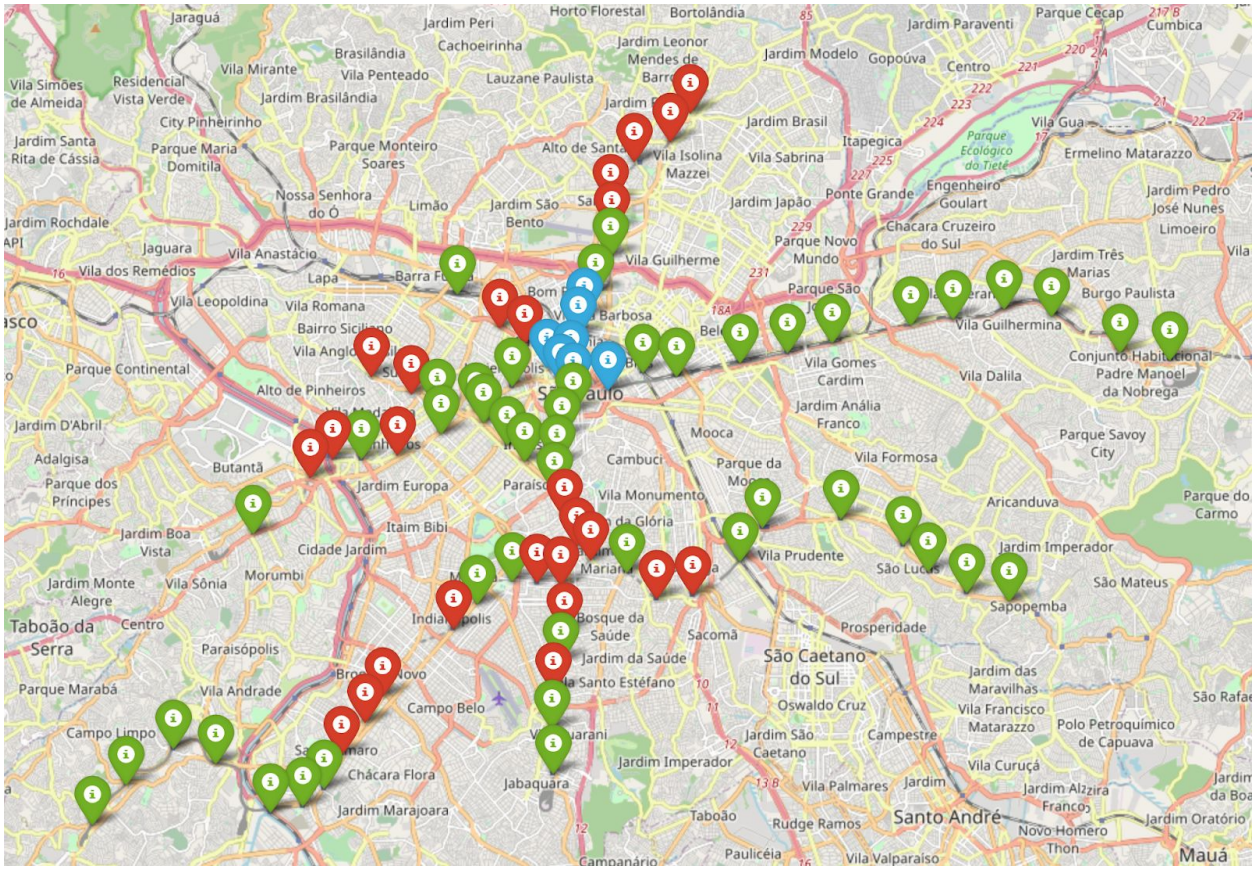
4 - Results

First, K-means was applied to all categories, with different clusters sizes. Following, we can analyse the results.

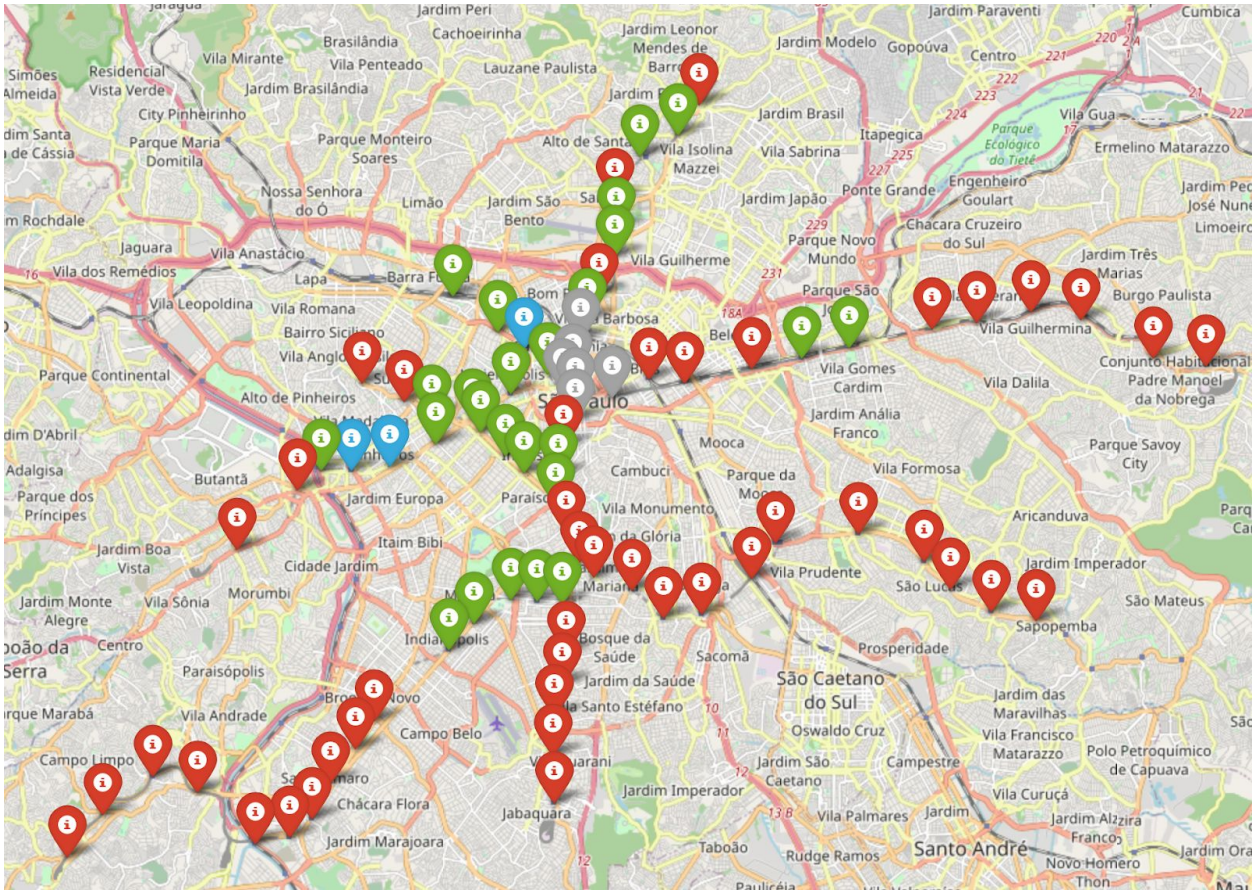
- Cluster size = 2



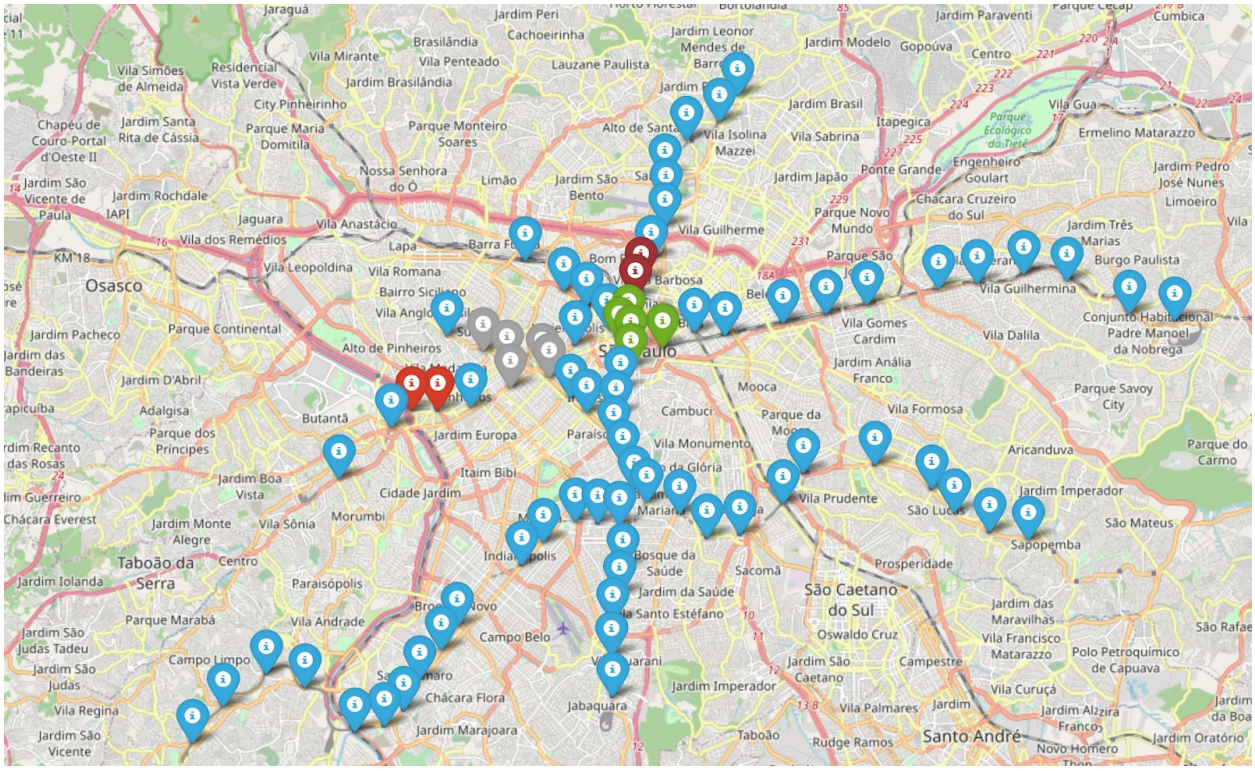
- Cluster size = 3



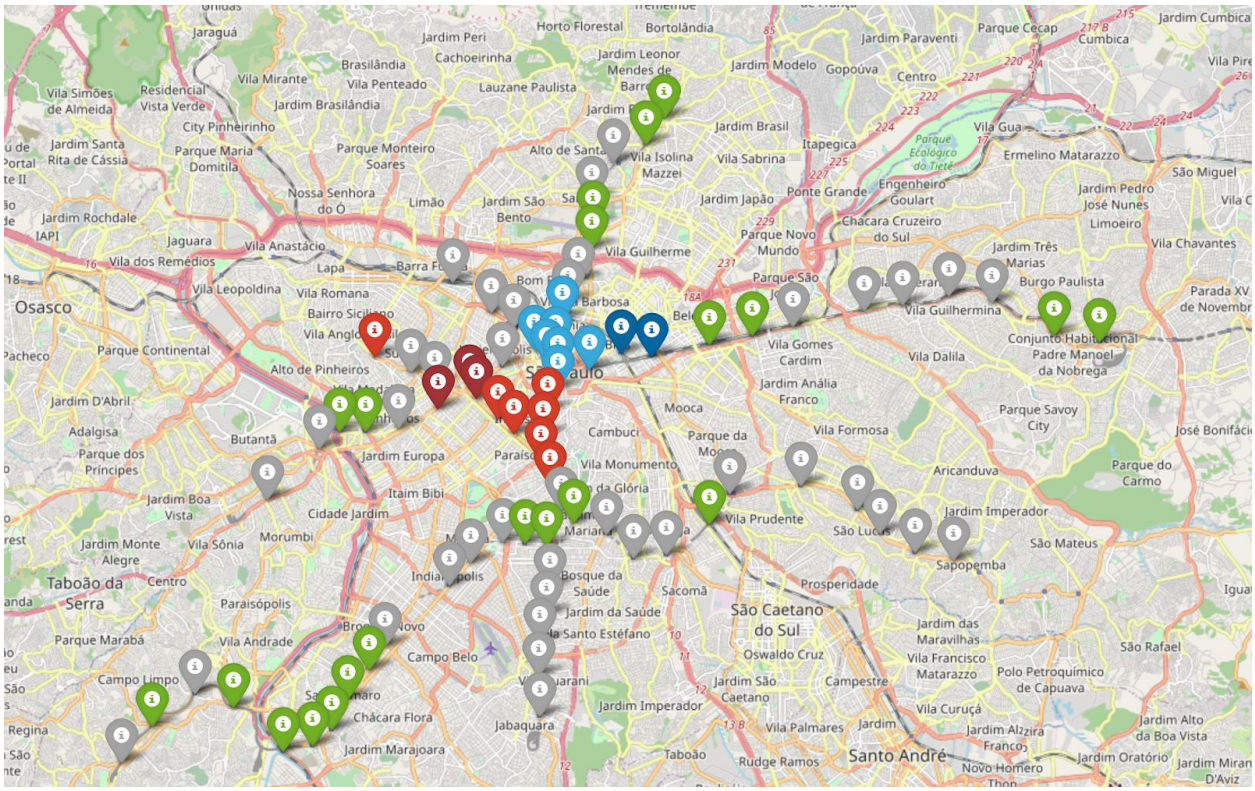
- Cluster size = 4



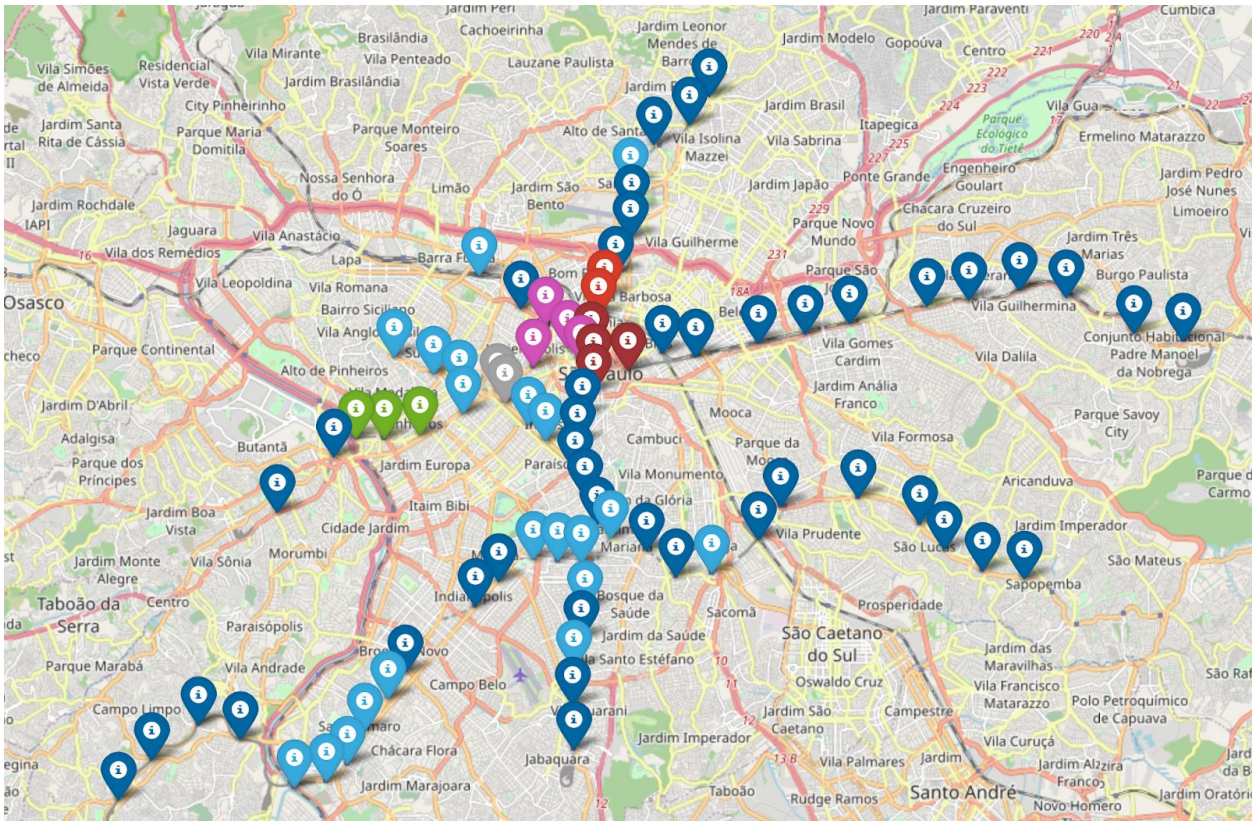
- Cluster size = 5



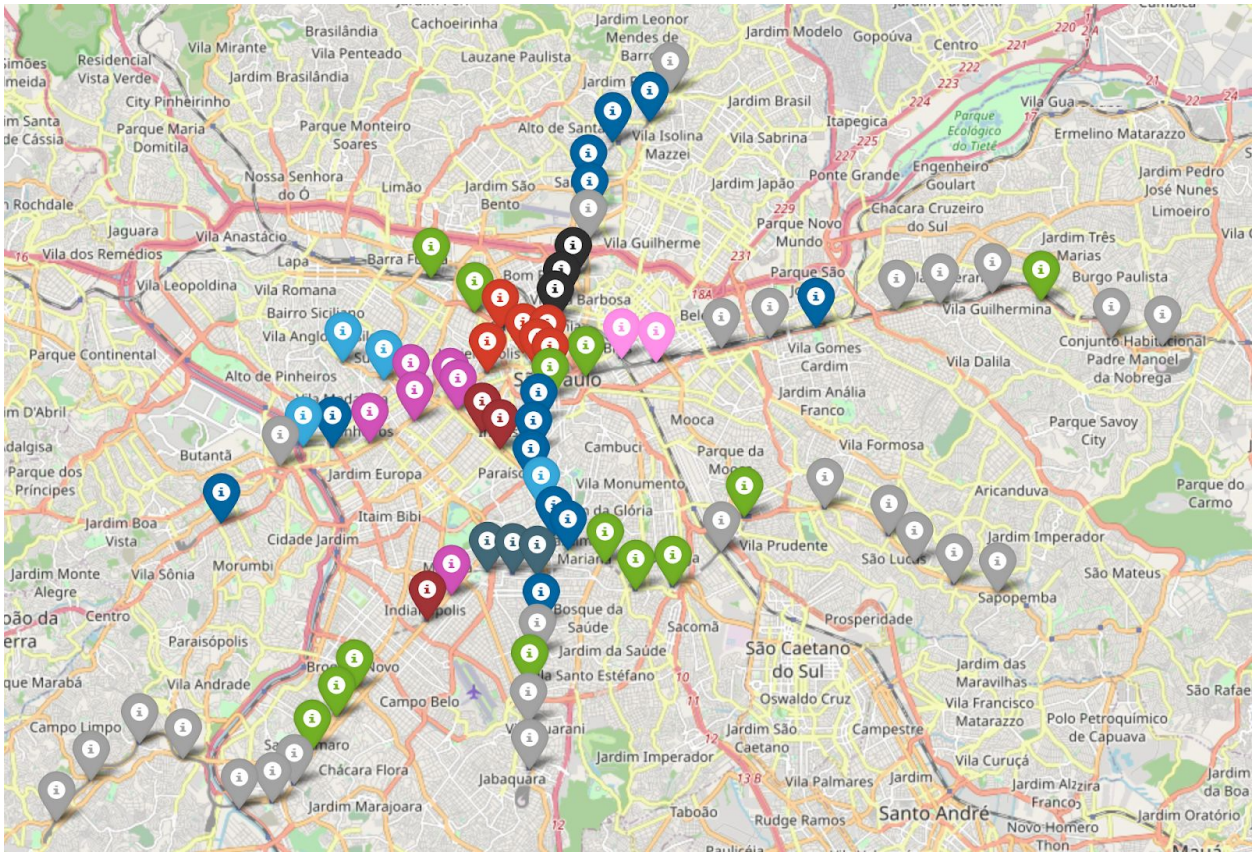
- Cluster size = 6



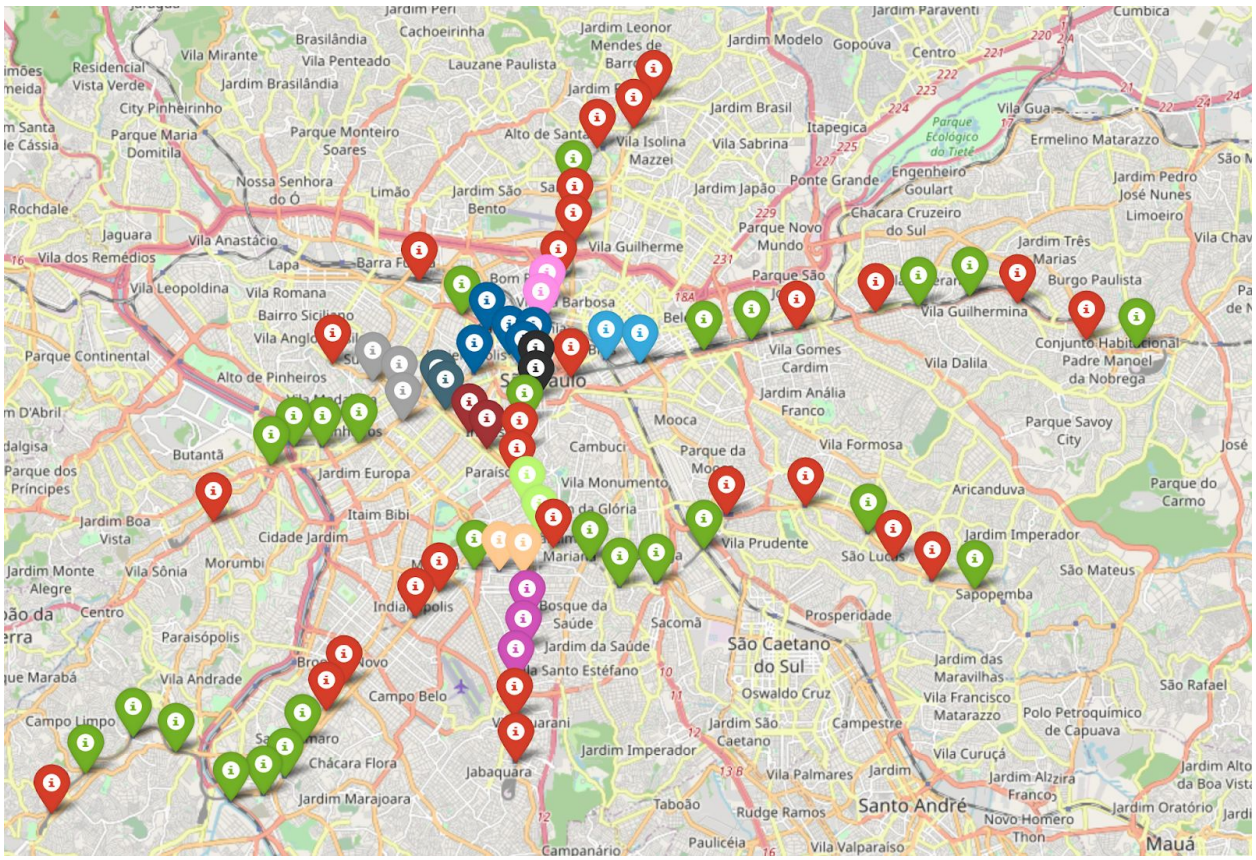
- Cluster size = 7



- Cluster size = 10



- Cluster size = 12

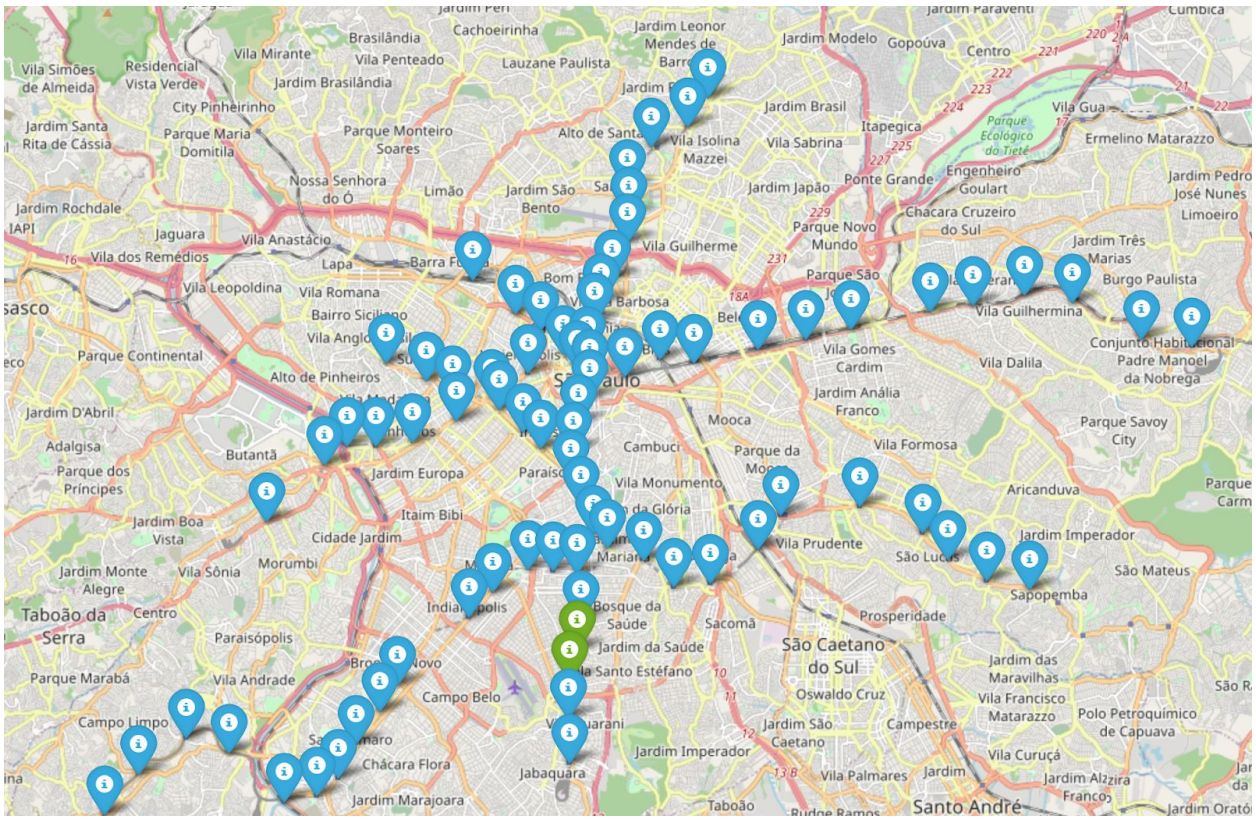


The results above were obtained grouping all existing categories of venues. We can observe that near locations tends to have belong to similar clusters. The analysis can be used to map different customers profiles by each region.

We observe that, even with different cluster sizes, the center of Sao Paulo belongs to the same cluster. Each region has its own particularity and culture.

Those information are necessary for an investor analyse and choose the best location to open a new venue.

When clustering only with the Irish Pub category, with 2 clusters, we observe the following:



In fact, the Irish Pubs in Sao Paulo are concentrated in cluster 1, in green. Hence, the regions of those two metro stations have all irish pubs in town.

5 - Discussion

A new investor should probably check for two factors when choosing the location for its new Irish Pub: concurrency and the probability of the customers of that region want to go to an Irish Pub. If the Irish Pubs in Sao Paulo are concentration in the nearby of those two regions in green, that leds to having more customers inclined to go to an Irish Pub. Hence, the best location would be the two nearby metro station areas in the blue cluster.

The idea of choosing a location of a different cluster where the Irish Pub are located is because, customers of those regions nearby may go to the two metro stations location where the Irish Pubs are located. Also,

opening an Irish Pub in a location where has lower concurrency is obviously better for the business.

6 - Conclusion

This report uses geolocation data of metro stations in Sao Paulo to analyse the customer profile for a new Irish Pub by verifying the nearby venues categories.

The code was developed in Python and can be seen here [5]

References

[1] https://www.kaggle.com/thiagodsd/sao-paulo-metro/data?select=metrosp_stations.csv

[2] https://python-visualization.github.io/folium/

[3] https://pandas.pydata.org/

[4] https://scikit-learn.org/stable/

[5] https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/778b6e15-54b5-4ca1-9c3d-18c05cbc4a92/view?access_token=1fbcf5fb5772a552bfe8dfef27627fdb02a08951564962b4946d5acd75617e41