Clustering of São Paulo Subway Stations using Foursquare API

Felipe Testa

March 28th, 2020

1. Introduction

The São Paulo subway commonly called Metro is one of the urban railways that serve the city of São Paulo, alongside the São Paulo Metropolitan Trains Company (CPTM), forming the largest metropolitan rail transport network of Latin America. The six lines in the metro system operate on 101.1 kilometers (62.8 mi) of the route, serving 72 stations. The metro system carries about 5,300,000 passengers a day.

Undoubtedly, the subway is an important part of any metropolis around the world, especially São Paulo with a population of 12,2 million moving around. The idea of this project is to use an unsupervised machine learning technique — KMeans in order to segment all São Paulo's subway stations according to the venue density by using data from Foursquare API.

A desirable intention is of this project will help to understand the diversity of subway station by leveraging venue data from Foursquare's 'Places API' and 'k-means clustering' unsupervised machine learning algorithm. By analyzing this data we can classify stations by their surrounding. This data can be useful for city planners to determine where from and where people are most likely to travel for work and leisure, plan a further extension of the network and find places for new development. Also after this analysis, it is expected to have an initial project to help politicians to make better public policy for the population and help to understand areas close to Metro in São Paulo.

2. Data

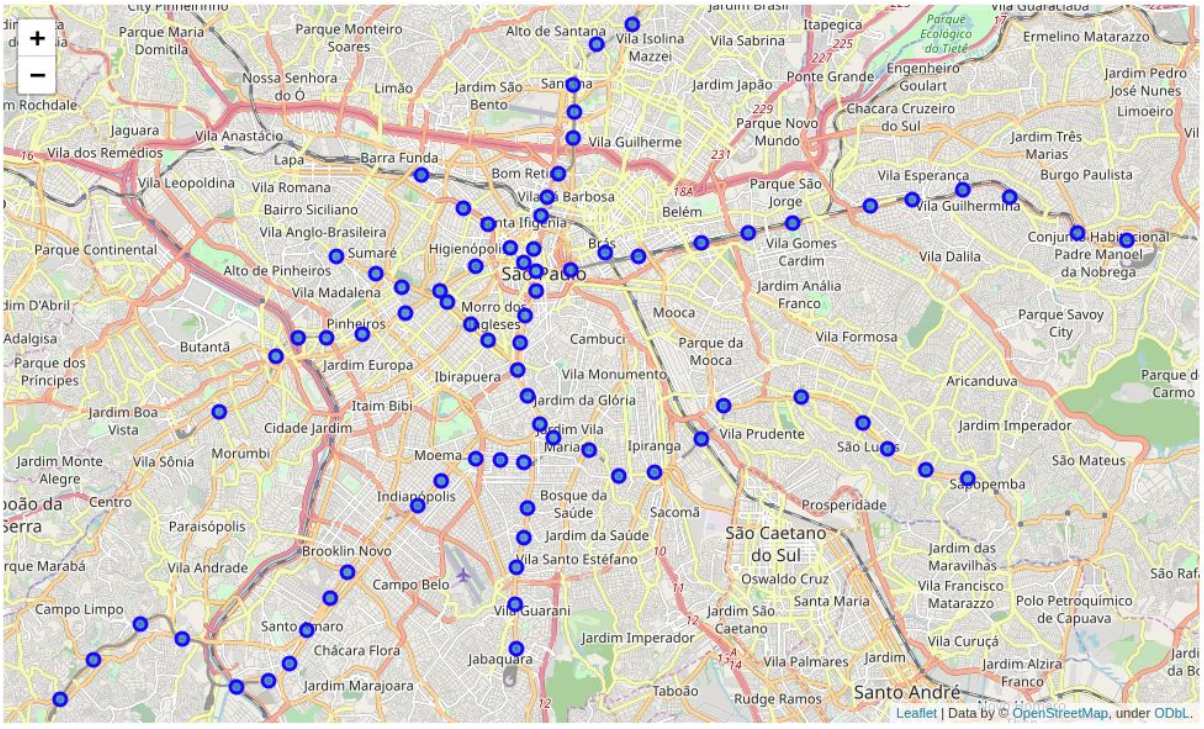
At this project, It will be used the following datasets:

Metro Station Location Data

All data for subway stations (Metro) was downloaded from Kaggle Dataset. Sample data below.

	metro_station	latitude	longitude	metro_line
0	aacd-servidor	-23.597825	-46.652374	['lilas']
1	adolfo-pinheiro	-23.650073	-46.704206	['lilas']
2	alto-da-boa-vista	-23.641625	-46.699434	['lilas']
3	alto-do-ipiranga	-23.602237	-46.612486	['verde']
4	ana-rosa	-23.581871	-46.638104	['azul', 'verde']

Following below plot with all metro stations location overview.



São Paulo Metro Station Overview

Foursquare Venue Data

At this notebook will be used RESTful API calls to retrieve data about venues in different areas. As mentioned Foursquare API is used to explore the metro station surrounding and segment them. To access the API, CLIENT_ID, CLIENT_SECRET, and VERSION are defined in a credentials file, in order to get credentials for your project just sign up on the following link.

Following below an example of a response from the API. This is the link to Foursquare API documentation for more details.


```
{'categories': [{'id': '4d4b7104d754a06370d81259',
'name': 'Arts & Entertainment',
'pluralName': 'Arts & Entertainment',
'shortName': 'Arts & Entertainment',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/default_',
'suffix': '.png'},
'categories': [{'id': '56aa371be4b08b9a8d5734db',
'name': 'Amphitheater',
'pluralName': 'Amphitheaters',
'shortName': 'Amphitheater',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/default_',
'suffix': '.png'},
'categories': []},
{'id': '4fceea171983d5d06c3e9823',
'name': 'Aquarium',
'pluralName': 'Aquariums',
'shortName': 'Aquarium',
'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/arts_entertainment/aquarium_',
'suffix': '.png'},
'categories': []}]
```


There are many endpoints available on Foursquare for various GET requests. But, to explore the subway surrounding, it is required the number of venues per category establish at Foursquare Venue Category Hierarchy.

Get Response From Foursquare API

We'll be querying the number of venues in each category in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance. Following below all categories resulted from the GET response at Foursquare API.

- Arts & Entertainment (4d4b7104d754a06370d81259)
- College & University (4d4b7105d754a06372d81259)
- Event (4d4b7105d754a06373d81259)
- Food (4d4b7105d754a06374d81259)

- Nightlife Spot (4d4b7105d754a06376d81259)
- Outdoors & Recreation (4d4b7105d754a06377d81259)
- Professional & Other Places (4d4b7105d754a06375d81259)
- Residence (4e67e38e036454776db1fb3a)
- Shop & Service (4d4b7105d754a06378d81259)
- Travel & Transport (4d4b7105d754a06379d81259)

3. Methodology

We can use the Foursquare explore API with category_id to query the number of venues of each category in a specific radius. The response contains a results value for the specified coordinates, radius, and category. At this project, all requests were made setting a 1000m radius for each category.

Following below the sample dataset after getting all results per metro station.

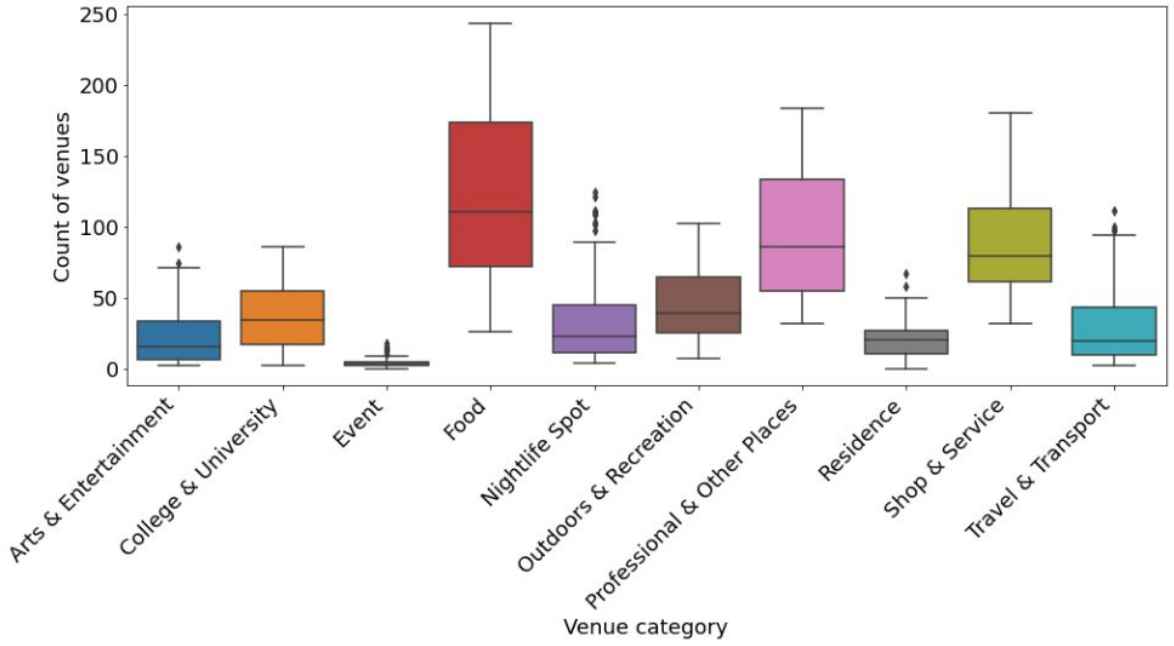
	metro_station	metro_line	latitude	longitude	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	aacd-servidor	['lilas']	-23.597825	-46.652374	13	71	110	23	76	121	11	62	23
1	adolfo-pinheiro	['lilas']	-23.650073	-46.704206	8	42	96	11	33	89	14	89	15
2	alto-da-boa-vista	['lilas']	-23.641625	-46.699434	2	24	71	8	30	90	15	65	9
3	alto-do-ipiranga	['verde']	-23.602237	-46.612486	10	23	68	12	43	47	33	58	5
4	ana-rosa	['azul', 'verde']	-23.581871	-46.638104	32	75	214	55	68	136	29	94	47

Exploratory Data Analysis

Now, let's look deeper into the data. It has been create a descriptive analysis and a box plot in order to get an overview of all metro stations.

category_name	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
count	79.0	79.0	79.0	79.0	79.0	79.0	79.0	79.0	79.0	79.0
mean	22.0	37.0	4.0	125.0	35.0	45.0	94.0	21.0	91.0	31.0
std	21.0	23.0	4.0	63.0	34.0	24.0	45.0	14.0	43.0	30.0
min	2.0	2.0	0.0	26.0	4.0	7.0	32.0	0.0	32.0	2.0
25%	6.0	17.0	2.0	72.0	12.0	25.0	54.0	10.0	61.0	10.0
50%	15.0	34.0	3.0	110.0	23.0	39.0	86.0	20.0	79.0	19.0
75%	33.0	54.0	5.0	174.0	45.0	64.0	133.0	26.0	113.0	44.0
max	86.0	86.0	18.0	243.0	124.0	102.0	183.0	67.0	180.0	111.0

Let's display the number of venues as a boxplot to better visualize the data profile and get better insights.



As we can see, the top 3 venues categories with a higher frequency around the Metro station in São Paulo:

- Food
- Professional & Other Places
- Shop & Service

It means when we're looking at any subway station surrounding in São Paulo is more likely to have a higher number of venues related to those categories than others. Another important fact is that the category Event has fewer venues, therefore, it has been not considered for the further clustering method.

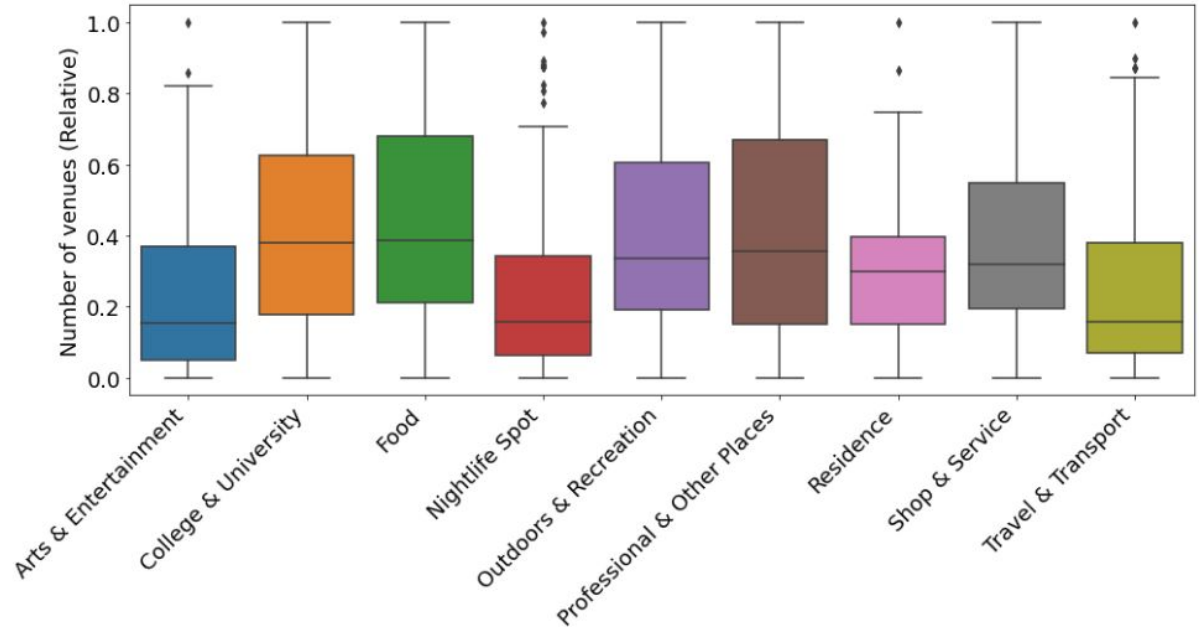
Feature Engineering

The next important step before any actual machine learning model is Feature Engineering. This important step mainly consists of two things:

- Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
- Improving the performance of machine learning models.

The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering.
— Luca Massaron

Keeping this in mind, now it's time to normalize our dataset. For this task, it has been used min-max scaling (scale count of venues from 0 to 1 where 0 is the lowest value in a set and 1 is highest). This both normalizes the data and provides an easy to interpret score at the same time. The scaled boxplot looks like this:



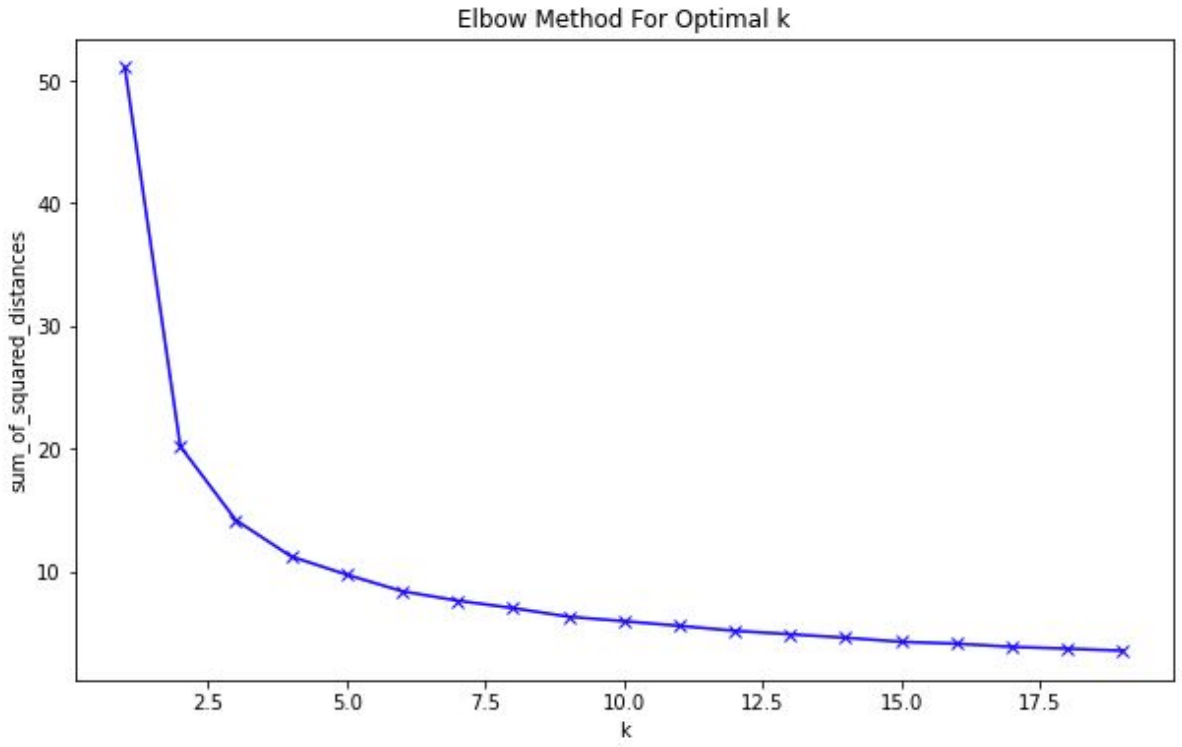
K-Means Clustering

'K-Means' is an unsupervised machine learning algorithm that creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count venues for each cluster label for variable cluster size. To implement this algorithm, it is very important to determine the optimal number of clusters (i.e. k). There are 2 most popular methods for the same, namely 'The Elbow Method' and 'The Silhouette Method', for this project will be used 'The Elbow Method'.

Elbow Method

The Elbow Method calculates the sum of squared distances of samples to their closest cluster center for different values of 'k'. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances. Following is an implementation of this method (with varying number of clusters from 1 to 20):

Sometimes, Elbow method does not give the required result, which did not happen in this case. If there was a gradual decrease in the sum of squared distances, an optimal number of clusters could not be determined. To counter this, another method can be implemented, such as the Silhouette Method.



The Elbow Method determines an optimal number of clusters of Four.

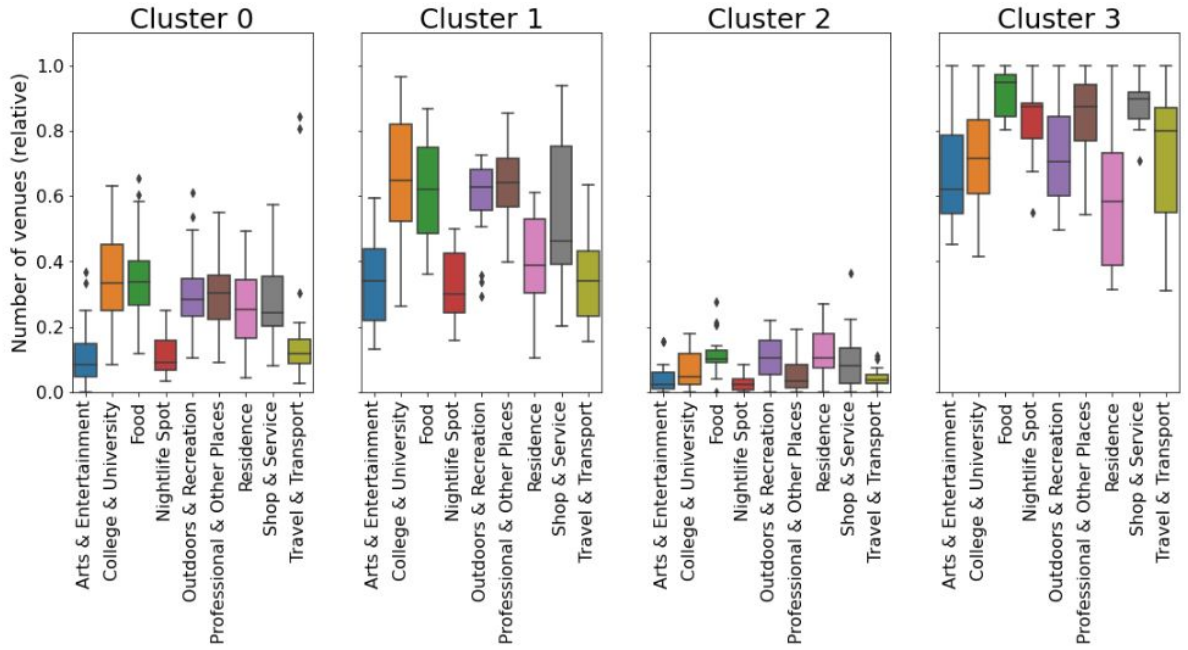
4. Results

K-Means Algorithm

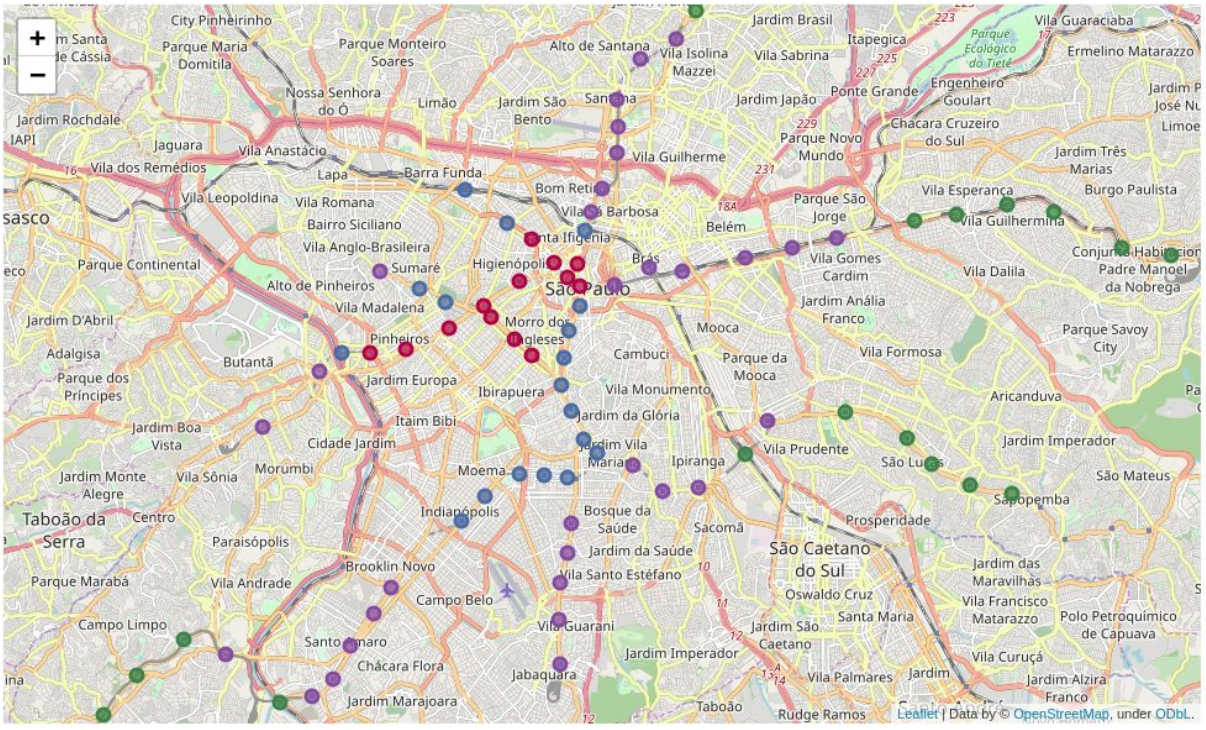
Now, it is the time to run the K-Means algorithm to cluster the dataset.After getting the cluster for all metro stations, let's visualize a sample of it.

	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	cluster	metro_station
0	0.130952	0.821429	0.387097	0.158333	0.726316	0.589404	0.164179	0.202703	0.192661	1	aacd-servidor
1	0.071429	0.476190	0.322581	0.058333	0.273684	0.377483	0.208955	0.385135	0.119266	0	adolfo-pinheiro
2	0.000000	0.261905	0.207373	0.033333	0.242105	0.384106	0.223881	0.222973	0.064220	0	alto-da-boa-vista
3	0.095238	0.250000	0.193548	0.066667	0.378947	0.099338	0.492537	0.175676	0.027523	0	alto-do-ipiranga
4	0.357143	0.869048	0.866359	0.425000	0.642105	0.688742	0.432836	0.418919	0.412844	1	ana-rosa

After getting the result DataFrame, It has been displayed the boxplot below. It's noticed that the major difference between clusters was related to how 'crowded' of venues is the subway surrounding. For example, Cluster 3 has a higher number of venues (relative) medians, when compared to other clusters. Then, we could imply that subway stations from Cluster 3 have more venues density in a 1000m radius than any other metro stations from other clusters.



Let’s visualize all cluster on the map:



On the figure cluster division shows clearly divisions on the map like circles, which we could easily identify 4 circles in Sao Paulo, which one has their own characteristics. More details on clusters can be defined as followed:

- **Cluster 3 (Red) — 13 metro stations**

Metro stations within cluster 3 have a higher frequency of venues and contain Sao Paulo Downtown Neighborhoods (Praca da Se, Republica e Anhangabau) and important streets in Sao Paulo (Av. Paulista, Faria Lima, Reboucas and Oscar Freire). Those streets have headquarters of many financial and cultural institutions, it's known the financial capital of Brazil. Usually, those areas have a higher frequency of Professional, Food, Shop and Service venues.

- **Cluster 1 (Blue) — 18 metro stations**

Metro stations within cluster 1 do not have the highest frequency of venues in Sao Paulo. However, It's close to Downtown and Financial Center of Sao Paulo. Neighborhoods close to those metro stations are also super important in Sao Paulo, many companies have headquarters and important places in Sao Paulo are located in this areas. And it also has great restaurant areas such as Moema and Itaim.

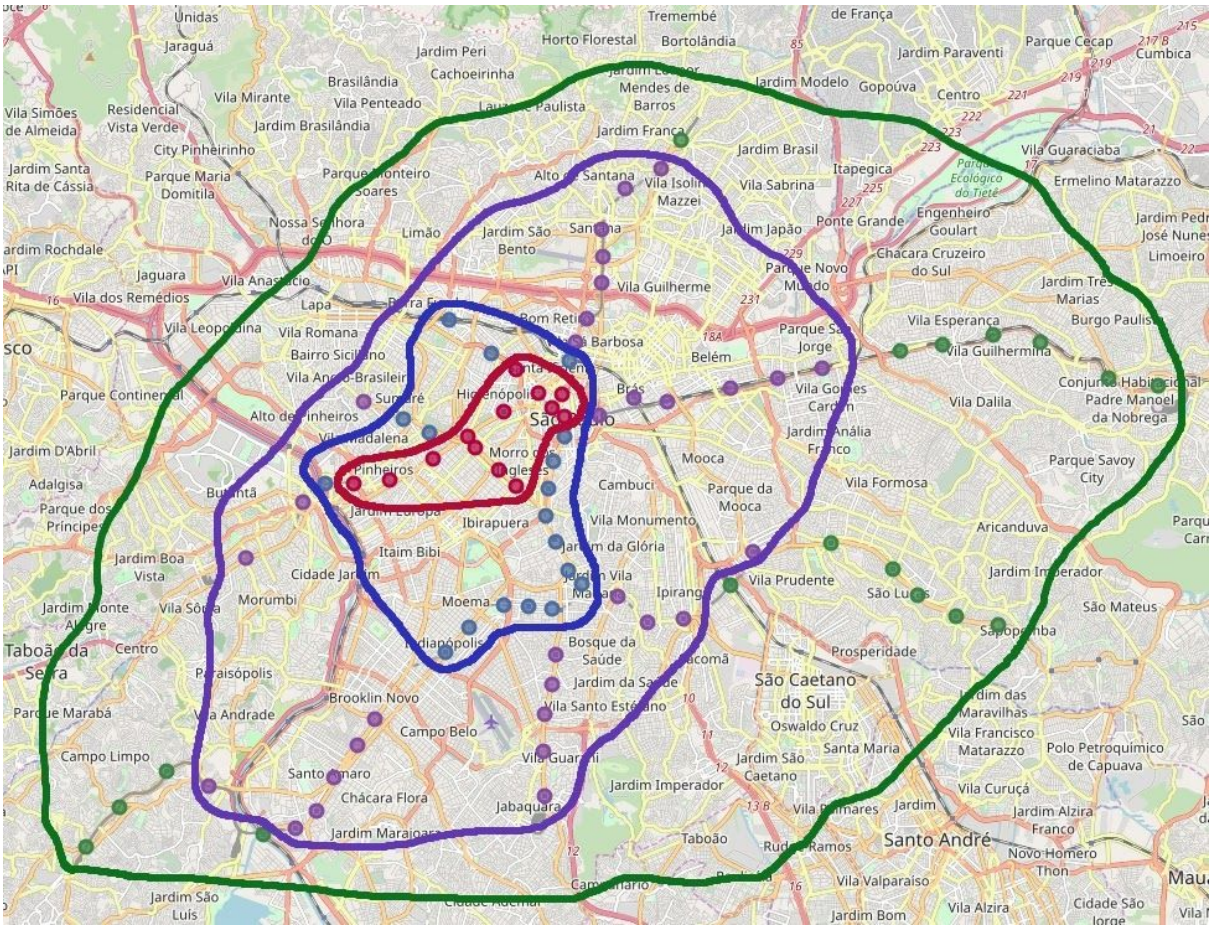
- **Cluster 0 (Purple) — 31 metro stations**

It is the biggest cluster on this analysis with 31 metro stations, this area contains stations that are further from downtown and financial center in São Paulo. It also contains the fewer frequency of venues in every category.

- **Cluster 2 (Green) — 17 metro stations**

Subway stations from this cluster are in the farther area in Sao Paulo and it’s major Residence Venues.

As explained before K-means was able to cluster metro stations by using their surrounding venues, and it has been produced Four different clusters as shown below. Those areas are different from each other mainly due to venue concentration. Stations that are more close to downtown has more venues within 1000m radius than stations further to the center.



5. Discussion

The purpose of this project was to cluster different metro stations in Sao Paulo based on the surrounding areas of every metro station. For that, Foursquare API venue data was used. Foursquare data isn't all-encompassing since data doesn't take into account a venue's size (e.g. a big restaurant attracts a lot more people that a hot dog stand — each of them is still one Foursquare "venue").

Another possible development is to include more data e.g housing prices and criminality and passenger per station it would be interesting to add this kind of information to the analysis. This could potentially be valuable for getting more detailed clusters and a profile of each metro station helping politicians and Metro Company to take better decisions.

6. Conclusion

Four clusters were identified. The main differences between the clusters are the average number of venues per metro station and the most common venues surrounding it are Food, Professional & Other Places and Shop & Service. K-Means clustering method was able to separate stations by a number of venues within a 1000m radius and showed that Sao Paulo has more venues concentration close to downtown than on the city border.