Clustering Shopping Centers in the Metropolitan Region of São Paulo

Wanessa Nunes dos Santos

April 24th, 2021

1. Introduction

São Paulo is the city with the largest number of shopping centers. Shopping centers are projects with a very special commercial vision, embodied in the joint strategy of convergent attractiveness for all business players, that is, investors, tenants, consumers, neighbors, the general public and even the public authorities. .

1.1. The positive externality of Shopping Centers

Externality is understood as the possibility of a shopping mall, since it is inserted and installed within a neighborhood or micro-region, to positively influence a radius of stores in its surroundings. Because the simple presence of this shopping center, generates advantages for other tenants, who take advantage of the visit of consumers to those malls

This characteristic is fully consigned by scholars, who recognize the attractiveness of consumers provided by large shopping centers, through the variety of products and services, incomparable advertising and recognized notoriety of these malls, generating confidence and determination in the purchase, which also reinforces the positive externality.

1.2. Target audience

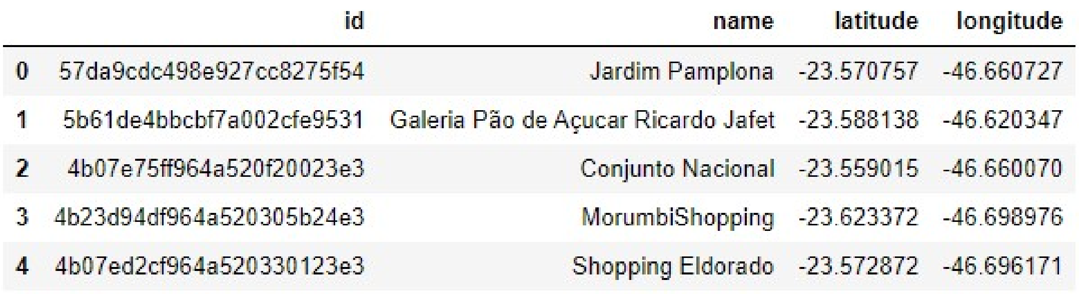
This project will help to understand the diversity of São Paulo Shopping Centers by leveraging venue data from Foursquare's 'Places API' and 'k-means clustering' unsupervised machine learning algorithm. By analyzing this data we can ** classify the malls by the locations that are most frequent around those malls **. This data can be useful for entrepreneurs wishing to invest in some business close to Shopping Centers in São Paulo Capital.

2. Data

All data for Shopping Centers of São Paulo (Metro) was downloaded from RESTful FourSquare API to retrieve data about Shopping Malls, categories and venues in different areas. This is the <a href='https://developer.foursquare.com/docs'>link</a> to Foursquare API documentation for more details.

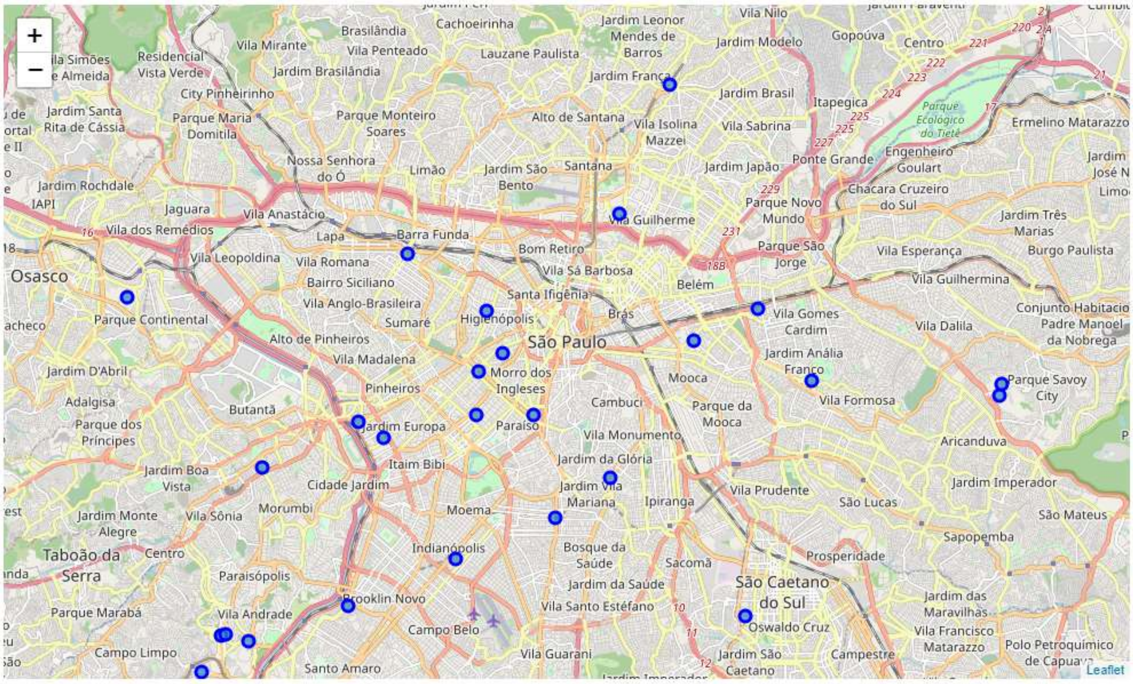
Shoppings Centers location Data

All data for Shopping Centers of São Paulo (Metro) was downloaded from RESTful FourSquare API. Sample data below.



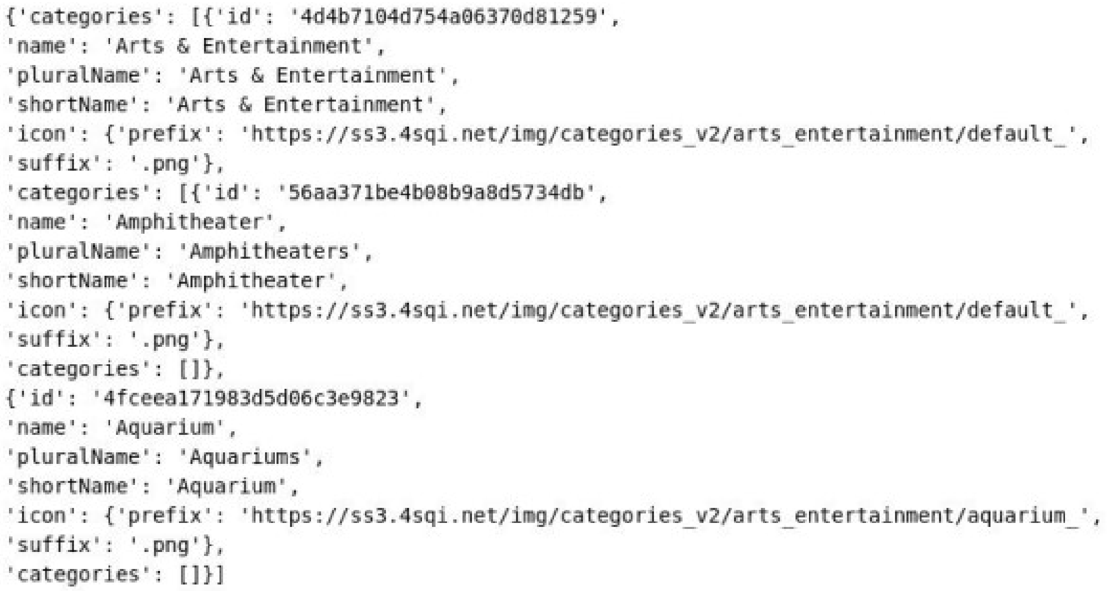
	id	name	latitude	longitude
0	57da9cdc498e927cc8275f54	Jardim Pamplona	-23.570757	-46.660727
1	5b61de4bbcbf7a002cfe9531	Galeria Pão de Açucar Ricardo Jafet	-23.588138	-46.620347
2	4b07e75ff964a520f20023e3	Conjunto Nacional	-23.559015	-46.660070
3	4b23d94df964a520305b24e3	MorumbiShopping	-23.623372	-46.698976
4	4b07ed2cf964a520330123e3	Shopping Eldorado	-23.572872	-46.696171

Following below plot with all shoppings centers location 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 shopping center 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 mall. This radius was chosen because 1000m is a reasonable walking distance. Following below all categories resulted from the GET response at Foursquare API.

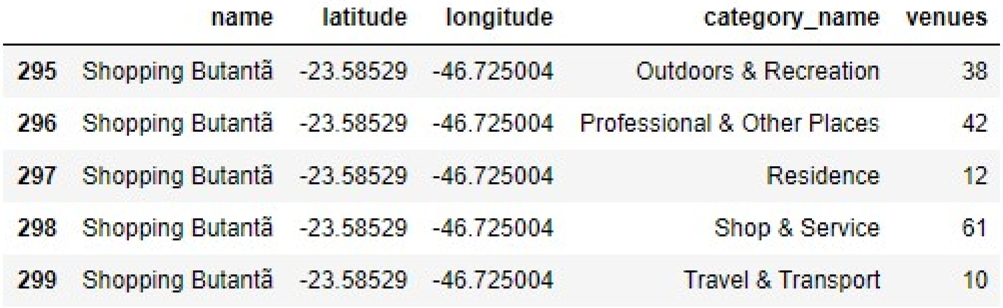
- Arts & Entertainment (4d4b7104d754a06370d81259)
- College & University (4d4b7105d754a06372d81259)
- Event (4d4b7105d754a06373d81259)
- Food (4d4b7105d754a06374d81259)
- Nightlife Spot (4d4b7105d754a06376d81259)

- 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.

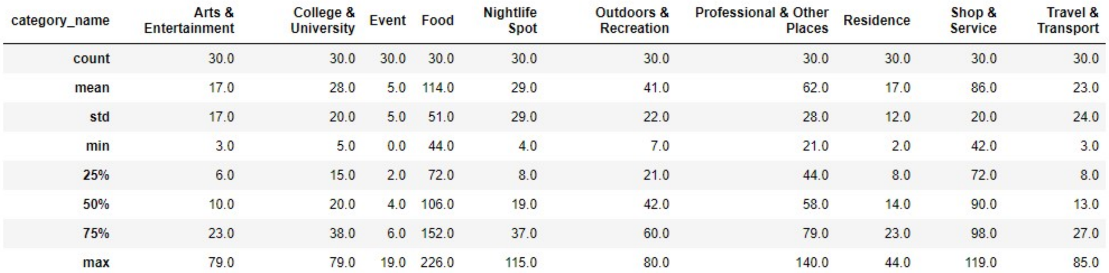
Following below the sample dataset after getting all results per shopping mall.



	name	latitude	longitude	category_name	venues
295	Shopping Butantã	-23.58529	-46.725004	Outdoors & Recreation	38
296	Shopping Butantã	-23.58529	-46.725004	Professional & Other Places	42
297	Shopping Butantã	-23.58529	-46.725004	Residence	12
298	Shopping Butantã	-23.58529	-46.725004	Shop & Service	61
299	Shopping Butantã	-23.58529	-46.725004	Travel & Transport	10

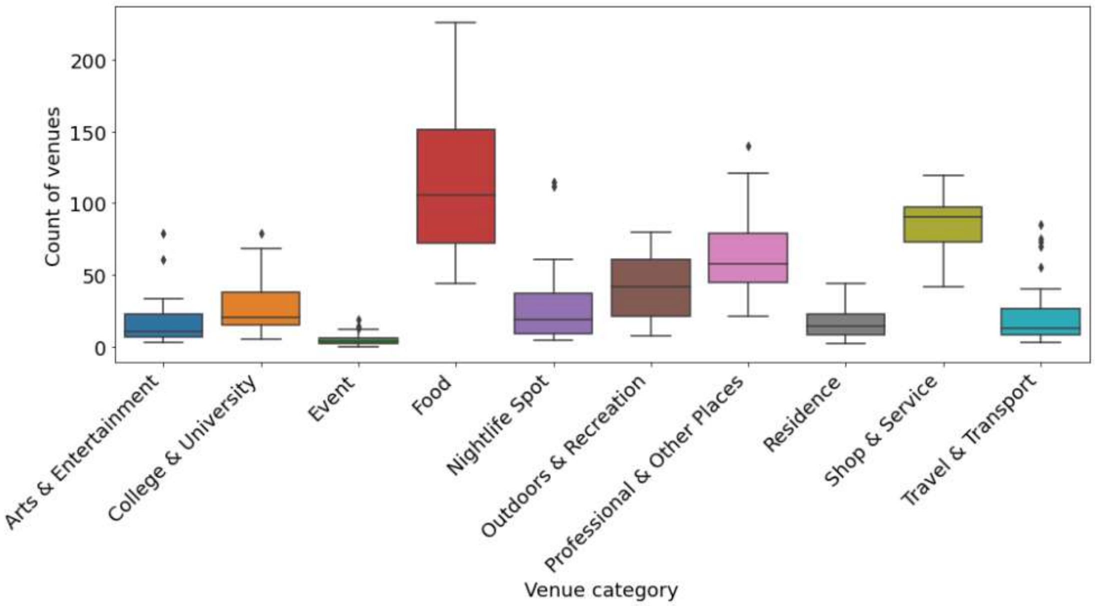
Exploratory Data Analysis

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 shopping centers.



category_name	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
count	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0	30.0
mean	17.0	28.0	5.0	114.0	29.0	41.0	62.0	17.0	86.0	23.0
std	17.0	20.0	5.0	51.0	29.0	22.0	28.0	12.0	20.0	24.0
min	3.0	5.0	0.0	44.0	4.0	7.0	21.0	2.0	42.0	3.0
25%	6.0	15.0	2.0	72.0	8.0	21.0	44.0	8.0	72.0	8.0
50%	10.0	20.0	4.0	106.0	19.0	42.0	58.0	14.0	90.0	13.0
75%	23.0	38.0	6.0	152.0	37.0	60.0	79.0	23.0	98.0	27.0
max	79.0	79.0	19.0	226.0	115.0	80.0	140.0	44.0	119.0	85.0

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:

Shop & Service
Food
Professional & Other Places

It means when we're looking at any shopping mall 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

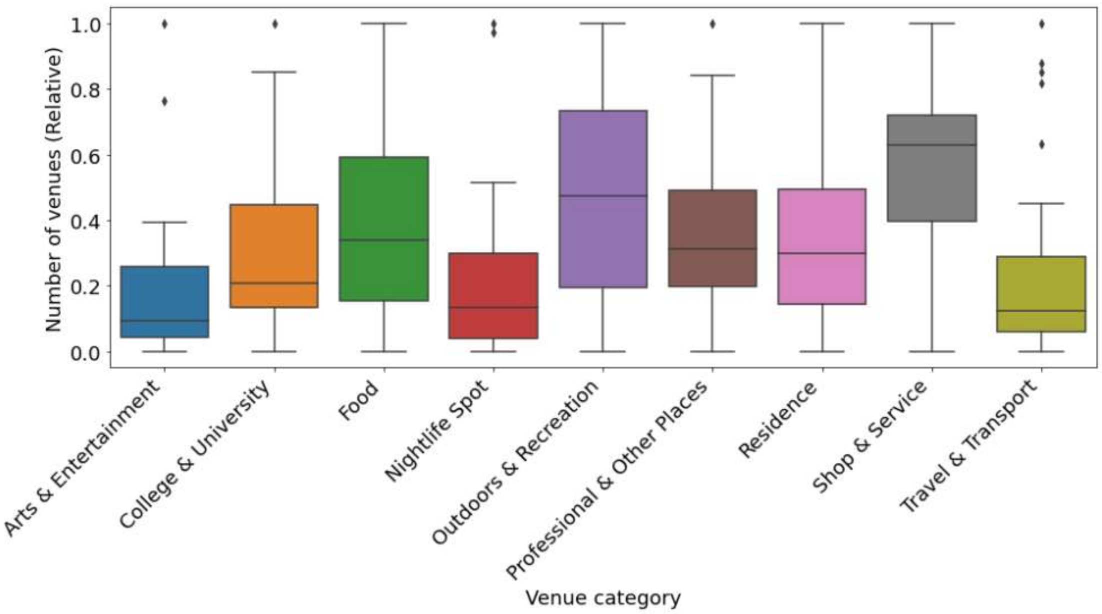
The next important step before any actual machine learning model is Feature Engineering. This important step mainly consists of two things:

- 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:



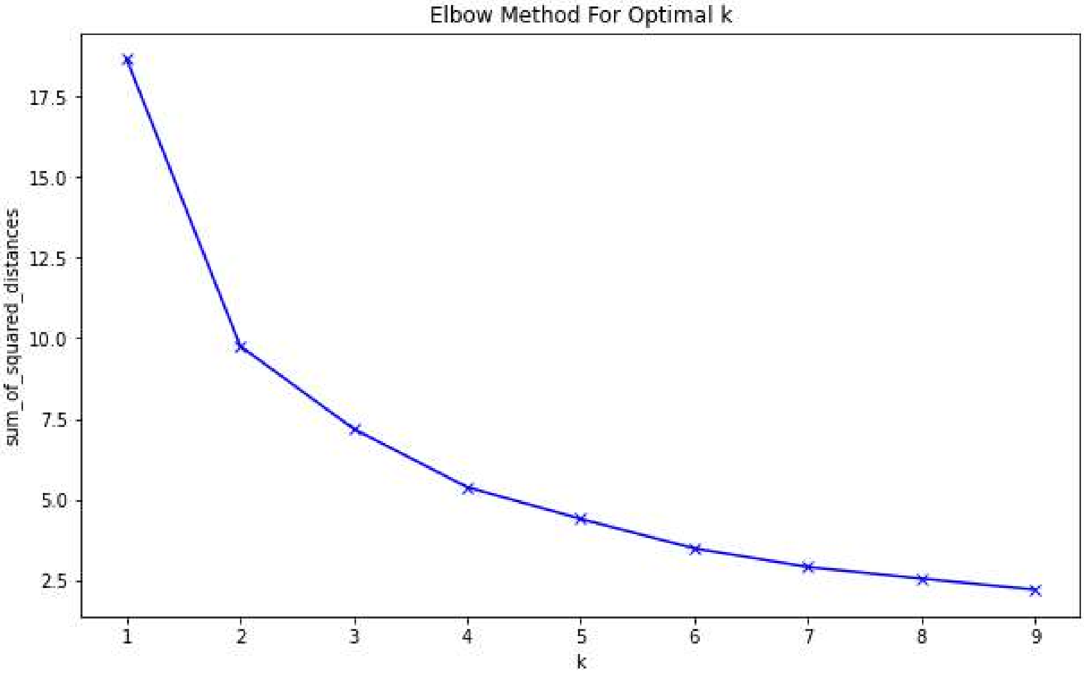
K-Means Clustering

'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'.

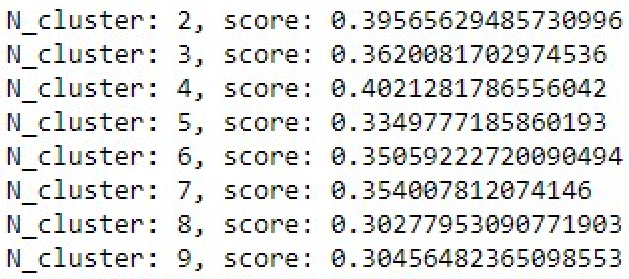
Elbow Method

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 Three.



```
N_cluster: 2, score: 0.39565629485730996
N_cluster: 3, score: 0.3620081702974536
N_cluster: 4, score: 0.4021281786556042
N_cluster: 5, score: 0.3349777185860193
N_cluster: 6, score: 0.35059222720090494
N_cluster: 7, score: 0.354007812074146
N_cluster: 8, score: 0.30277953090771903
N_cluster: 9, score: 0.30456482365098553
```

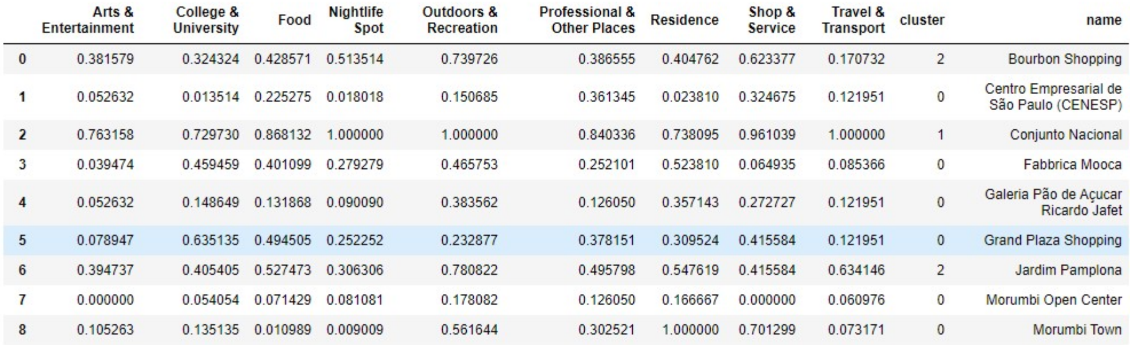
The score method determines an optimal number of clusters of Four.

4. Results

K-Means Algorithm

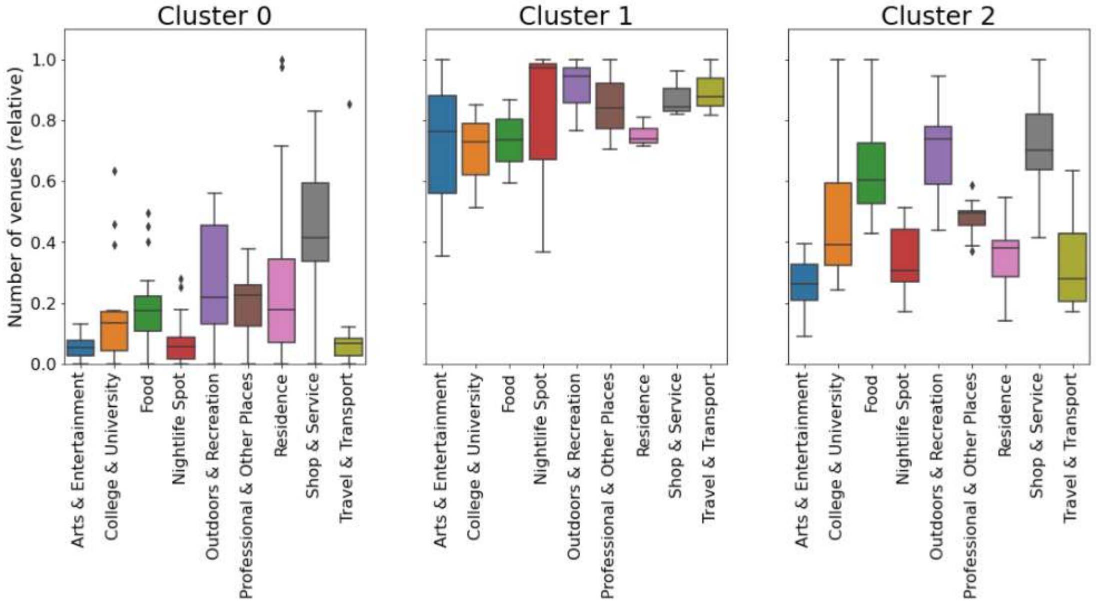
After testing clustering with 4 clusters and 3 clusters, the only difference noted was the creation of a new cluster with only 2 malls, or any apparent motivation. As we do not have many malls to cluster, I will choose to create 3 clusters.

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.

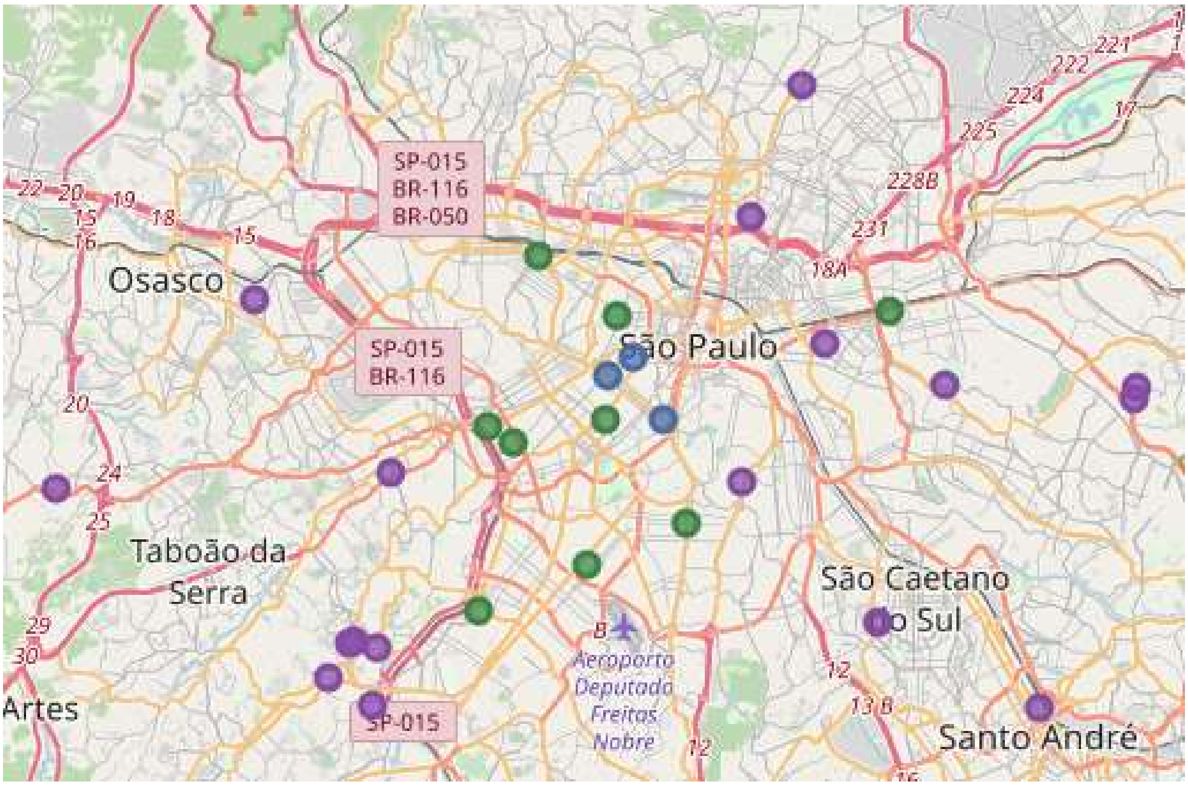


	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	cluster	name
0	0.381579	0.324324	0.428571	0.513514	0.739726	0.386555	0.404762	0.623377	0.170732	2	Bourbon Shopping
1	0.052632	0.013514	0.225275	0.018018	0.150685	0.361345	0.023810	0.324675	0.121951	0	Centro Empresarial de São Paulo (CENESP)
2	0.763158	0.729730	0.868132	1.000000	1.000000	0.840336	0.738095	0.961039	1.000000	1	Conjunto Nacional
3	0.039474	0.459459	0.401099	0.279279	0.465753	0.252101	0.523810	0.064935	0.085366	0	Fabbrica Mooca
4	0.052632	0.148649	0.131868	0.090090	0.383562	0.126050	0.357143	0.272727	0.121951	0	Galeria Pão de Açucar Ricardo Jafet
5	0.078947	0.635135	0.494505	0.252252	0.232877	0.378151	0.309524	0.415584	0.121951	0	Grand Plaza Shopping
6	0.394737	0.405405	0.527473	0.306306	0.780822	0.495798	0.547619	0.415584	0.634146	2	Jardim Pamplona
7	0.000000	0.054054	0.071429	0.081081	0.178082	0.126050	0.166667	0.000000	0.060976	0	Morumbi Open Center
8	0.105263	0.135135	0.010989	0.009009	0.561644	0.302521	1.000000	0.701299	0.073171	0	Morumbi Town

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 Shoppings surrounding. For example, Cluster 3 has a higher number of venues (relative) medians, when compared to other clusters. Then, we could imply that shoppings malls from Cluster 3 have more venues density in a 1000m radius than any other shopping malls from other clusters.



Let’s visualize all cluster on the map:



In the map above the division of clusters we can see that the malls were grouped according to their main characteristics, as explained below:

Cluster 1 (Blue) - within 3 Shoppings Malls

Shoppings within cluster 1 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. Larger even than the establishment of the malls themselves.

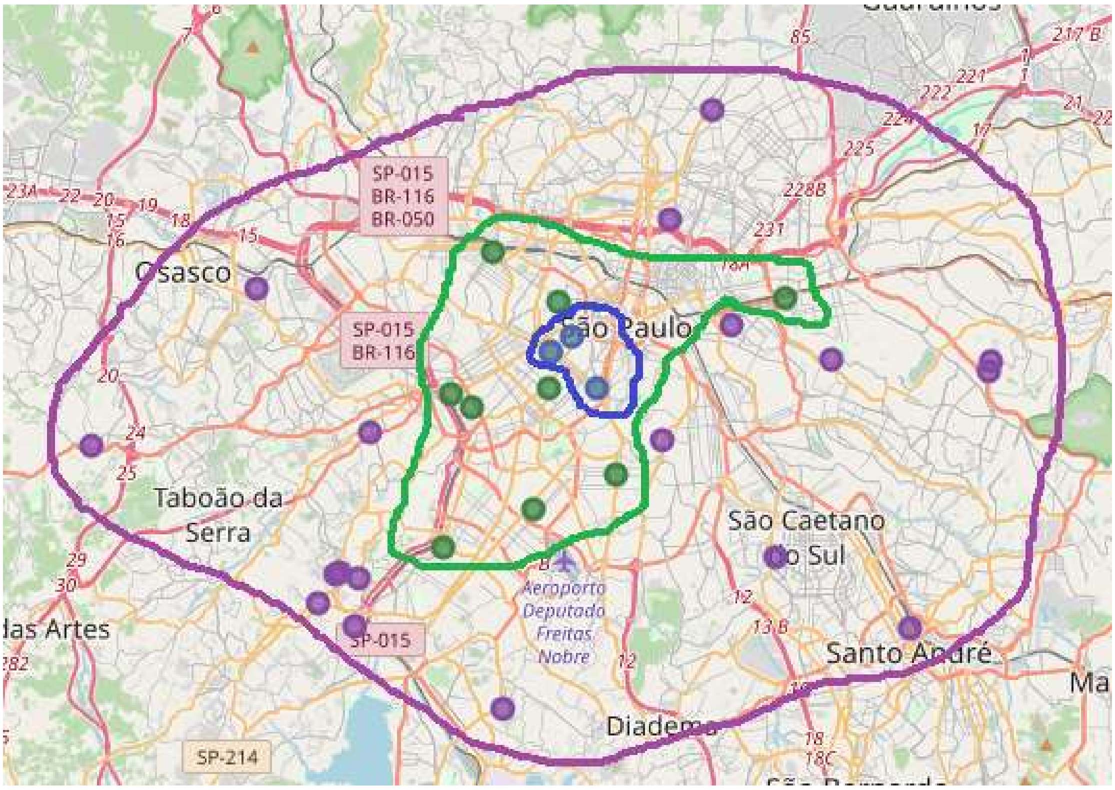
Cluster 2 (Green) - within 9 Shoppings Malls

Shoppings 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 malls 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) - within 18 Shoppigs Malls

It is the biggest cluster on this analysis with 18 Shopping Malls, this area contains malls that are further from downtown and financial center in São Paulo. With a high frequency of 'Shop & Service' and 'Food', which signals malls with little concentration of locations outside the mall.

As explained before K-means was able to cluster shopping malls 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. Malls that are more close to downtown has more venues within 1000m radius than malls further to the center.



5. Discussion

The purpose of this project was to cluster different Shoppings Malls in Sao Paulo based on the surrounding areas of every mall. 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 mall 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 Shopping Mall helping entrepreneurs to take better decisions.

6. Conclusion

Four clusters were identified. The main differences between the clusters are the average number of venues per Shoppings Centers and the most common venues surrounding it are Shop & Service, Food and Professional & Other Places. K-Means clustering method was able to separate the malls by a number of venues within a 1000m radius and showed that Sao Paulo has a group of malls that have few locations around them, but many locations working within the malls, malls that are located in more central areas and therefore with a complete offer of locations both outside and inside the malls and finally those malls that are located in more upscale areas of the city and therefore have an interesting number of establishments taking advantage of the externality propagated by these malls.

As an insight for a deeper analysis I would highlight cluster 2, green in color, 9 malls, as it has an interesting number of locations inside and outside the mall, but even less than the frequency of locations found in cluster 1, and higher external frequency than verified by cluster 0, signaling a good indication of the spread of externality generated by these malls.