Opening an Art Gallery in São Paulo

IBM Data Science Capstone Final Task

Thales A. Castelo Branco e Souza

June 07, 2020

1. Introduction

In this project, we will evaluate the best place to open an Art Gallery in the city of São Paulo, located in Brazil, South America. The city's metropolitan area, the Greater São Paulo, ranks as the most populous in Brazil [1] and the 12th most populous on Earth, with almost 22 million people.

Figure 1: São Paulo, Brazil.

Having the largest economy by GDP in Latin America and the Southern Hemisphere, São Paulo has the 11th largest GDP in the world, representing alone 10.7% of all Brazilian GDP. Besides that, it represents 36% of the production of goods and services in the state of São Paulo, being home to 63% of established multinationals in Brazil. In addition, the city has been responsible for 28% of the national scientific production.

Then, it is obvious that there are many museums and art galleries in every corner of the town. However, as many cities of the developing world, it is an unequal city, having many different borough standards. There are very expensive and very poor neighborhoods. Thereby, our analysis will be based on popularity of neighborhoods: *the boroughs with more places relation to art have more chance to be good locations for our stakeholders invest*.

2. Data Acquisition and Cleaning

The main purpose of the project is to find the places with more venues related to art. Thus, the first step is to select how the places will be organized. Logically, the a data grouping by the city's boroughs is very appropriated, since each borough has its own personality and atmosphere. Thereby, a strategy of selecting the ones with more venues related to art matches perfectly with our goal here.

For this, it was used, in this project, two data sources:

- File provided in the City Hall website, containing all official neighborhoods and its geographical coordinates (latitude and longitude);
- Foursquare data, obtained with API's, containing the venues close, in a 500m radius, to each location provided.

The first one was a CSV file, which had to be converted in a pandas dataframe, with 94 boroughs and for each one, its geographical coordinates.

Table 1: Boroughs coordinates, created with pandas (94 rows in total).

	Borough	Latitude	Longitude
0	Sé Bela Vista	-23.559353	-46.647325
1	Bom Retiro	-23.525558	-46.640975
2	Cambuci	-23.564380	-46.621792
3	Consolação	-23.551929	-46.655807
4	Liberdade	-23.559745	-46.635536
5	República	-23.543598	-46.642705

The second was extracted of Foursquare and it shows all the kinds of venues around (500m radius) every neighborhood of São Paulo. It has 295 types of venues, as shown below, such as Art Gallery, Museums, Stores, Supermarkets, Colleges, Restaurants etc.

Table 2: Boroughs venues (94 rows, 295 columns).

	Neighborhood	Acai House	Accessories Store	African Restaurant	Alternative Healer	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Art Museum	Art Studio	Arts & Crafts Store	Arts & Entertainment
0	Alto de Pinheiros	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Anhanguera	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Aricanduva	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Artur Alvim	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Barra Funda	0	0	0	0	0	0	0	0	0	0	0	0	1	0
5	Belém	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Bom Retiro	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Brasilândia	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Brás	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Butantã	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	Cachoeirinha	0	0	0	0	0	0	0	0	0	0	0	0	0	0

As one can think, it is necessary to filter this table with only art related venues, to be an easier and more precise clusterization. However, as we will filter that table, we should consider that the public that would frequent the art gallery would also like other venues, such as bookstores, music places and cultural centers. Then, if our gallery were close them, we would increase the chances to be seen and grow! Other kinds of important venues, besides arts and cultural places, are public transportation, which is essential in the modern live in big cities.

Thereby, we will filter Table 2, to achieve only the venues that we are interested. The result is shown below, in Table 3, that contains 94 rows and 41 columns.

Table 3: Boroughs art related venues (94 rows, 41 columns).

	Neighborhood	Art Gallery	Art Museum	Art Studio	Arts & Crafts Store	Arts & Entertainment	Auditorium	Antique Shop	Bookstore	Bus Station	Bus Stop	Camera Store	Circus	College Bookstore	College Theater
0	Alto de Pinheiros	0	0	0	0	0	0	0	1	0	0	0	0	0	0
1	Anhanguera	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Aricanduva	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Artur Alvim	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Barra Funda	0	0	0	1	0	0	0	0	0	0	0	0	0	0

It is possible to see that now, in Table 3, that we have a more art-related data, which will result in much better analysis, since we will be able to cluster the neighborhoods based exclusively on art-related venues.

Then, with the data ready, we could start to explore the relation of them in the Exploratory Data Analysis section.

3. Exploratory Data Analysis

With all data gathered, we were able to plot all the boroughs and see its distribution in a real map, to check if they are befitting reality. The map was created with folium library and is displayed bellow.

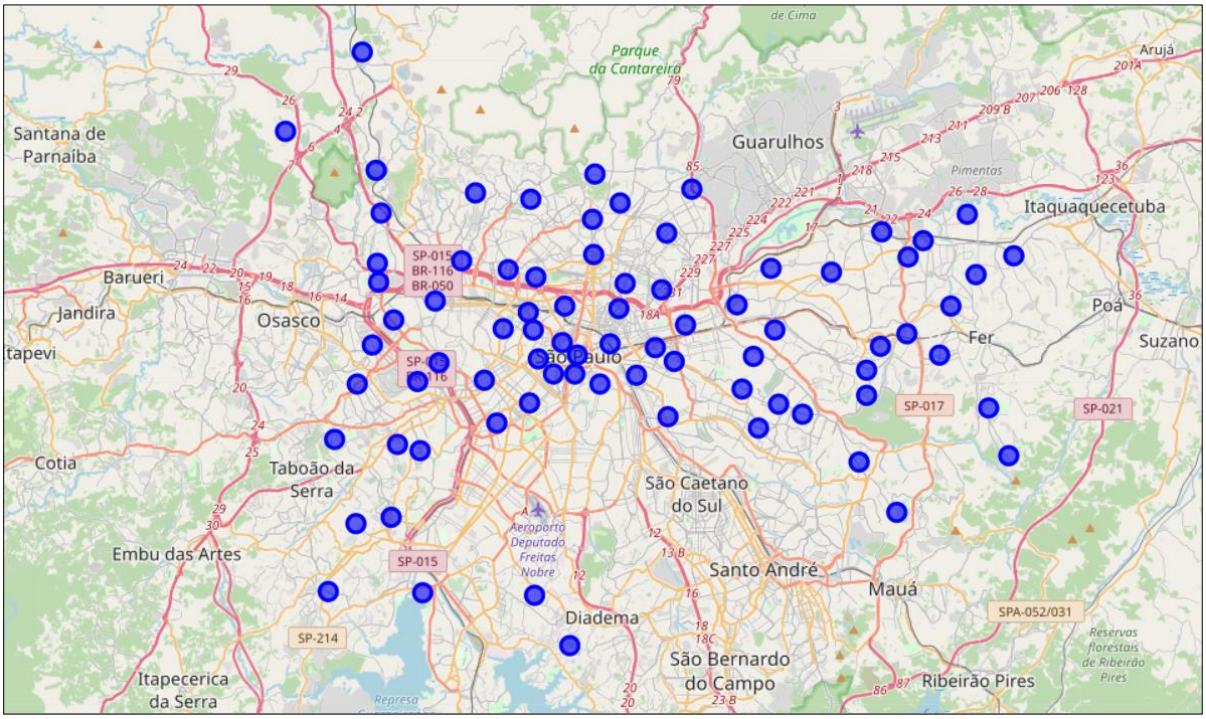

Figure 2: São Paulo neighborhoods and its locations (created with *folium*).

It is possible to see that the neighborhoods are well spread, which is good because it indicates that we will get very different characteristics for each location. Also, as a person that lived in the city, I know that many of the peripheral boroughs are poor and probably won't be appropriate places for the art gallery. The clusters in the next sections should show this relation.

In addition, it would be important if we get a hint of what we should expect in the next section, using a very quick and easy analysis. Thus, we summed all the columns of Table 3, and sorted them, to see the neighborhoods with more art-related venues. This new table is shown below.

Table 4: Quick Analysis Top 10 Neighborhoods with Art-Related Venues.

Neighborhood	Total
Sé	13
Barra Funda	12
República	12
Consolação	11
Sé Bela Vista	8
Moema	7
Vila Mariana	7
Butantã	6
Liberdade	6
Casa Verde	5

It is possible to see that the top 10 neighborhoods are in the central area, which is an older location of the city, with more venues related to art. In the next section, the cluster analysis will be done to refine this idea and for us to check the best neighborhoods to invest, to open an art gallery in São Paulo.

4. Methodology: Clustering Neighborhoods and Selecting the Best Locations

In the Methodology section, we will use Table 3 to cluster the data into five groups. These groups will contain the neighborhoods that have similar venues. The technique we will utilize is called *k-mean clusterization*. It is an unsupervised machine learning algorithm, that aims to group similar data together. A more scientific definition is written below.

> *"k-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. (...) It is popular for cluster analysis in data mining. k-means clustering minimizes within-cluster variances (squared Euclidean distances) (...)."* **Wikipedia** [2].

Thereby, we will cluster the neighborhoods of São Paulo with the art-related venues. The first thing to do is to determine the 10 most common venues for each borough. This analysis is shown in Table 5.

Table 5: 10 Most Common Venues of each Neighborhood (94 rows, 11 columns)

	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	Bookstore	Train Station	Dance Studio	Camera Store	Cultural Center	Concert Hall	Community College	College Theater	College Bookstore	Circus
1	Anhanguera	Train Station	Camera Store	Dance Studio	Cultural Center	Concert Hall	Community College	College Theater	College Bookstore	Circus	Bus Stop
2	Aricanduva	Train Station	Camera Store	Dance Studio	Cultural Center	Concert Hall	Community College	College Theater	College Bookstore	Circus	Bus Stop
3	Artur Alvim	Train Station	Camera Store	Dance Studio	Cultural Center	Concert Hall	Community College	College Theater	College Bookstore	Circus	Bus Stop
4	Barra Funda	Theater	Music Venue	Museum	Bus Stop	Arts & Crafts Store	Public Art	Bookstore	Indie Movie Theater	Cultural Center	Movie Theater

With Table 5 data already processed, we can now cluster the neighborhoods according to their most common art-related venues. This calculus will be done in the next section.

5. Results and Discussion

Then, based on the data processed above, we developed five clusters to analyze the neighborhoods, with *k-means*, generating Table 6 below.

Table 6: Cluster Labels and the most Common Venues of each Neighborhood (94 rows, 14 columns)

	Borough	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	9th Most Common Venue	10th Most Common Venue
0	Sé Bela Vista	-23.559353	-46.647325	2.0	Theater	Arts & Crafts Store	Antique Shop	Bookstore	Jazz Club	Movie Theater	College Bookstore	Cultural Center	Concert Hall	Community College
1	Bom Retiro	-23.525558	-46.640975	1.0	Train Station	Camera Store	Dance Studio	Cultural Center	Concert Hall	Community College	College Theater	College Bookstore	Circus	Bus Stop
2	Cambuci	-23.564380	-46.621792	1.0	Train Station	Camera Store	Dance Studio	Cultural Center	Concert Hall	Community College	College Theater	College Bookstore	Circus	Bus Stop
3	Consolação	-23.551929	-46.655807	3.0	Theater	Dance Studio	Bookstore	College Theater	Movie Theater	Art Gallery	Antique Shop	Arts & Entertainment	Bus Station	Bus Stop
4	Liberdade	-23.559745	-46.635536	0.0	History Museum	Public Art	Bookstore	Music Venue	Dance Studio	Train Station	Camera Store	Concert Hall	Community College	College Theater

A "*Cluster Labels*" column was created and two columns with the coordinates from Table 1, that will be used for visualization purposes. It is possible to see that different neighborhoods can be associated in the same cluster, or in different clusters, depending on their venues characteristics.

Now, that we have all the neighborhoods clustered and properly labeled, we can finally plot them and see everything on a map, to facilitate the results visualization. That is shown in Figure 3 below.

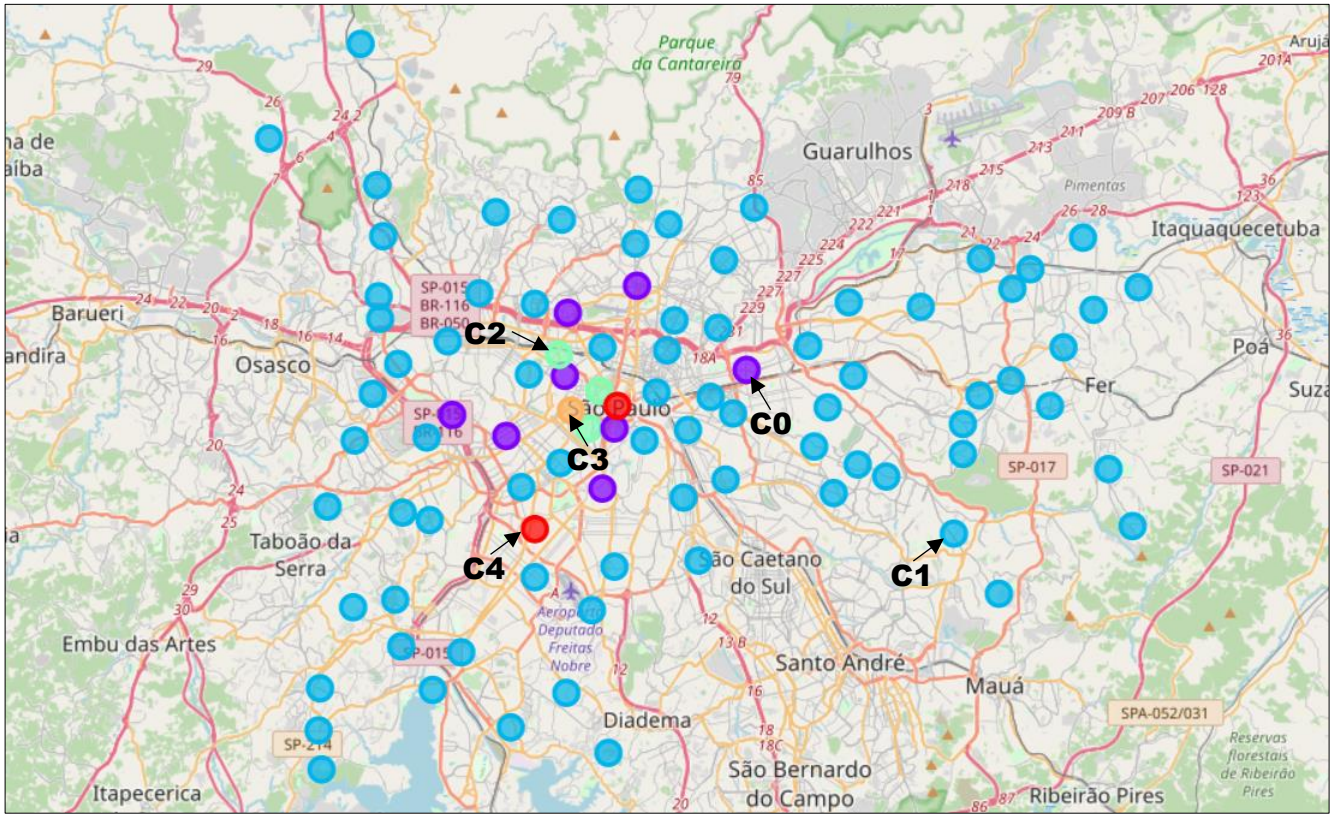

Figure 3: *k-means* Clusterization Results (created with *folium*)

It is now clear that many boroughs would not be appropriate to open an art gallery. Those would be from clusters C0 and C1. That actually makes sense, because the majority of the neighborhoods from C0 and C1 are very far from downtown and they are residential, therefore they do not have art-related venues. Therefore, as we do not seek these characteristics, C0 and C1 are eliminated from our analysis.

However, if we check closely to C2, C3 and C4 we can see that these are promising boroughs, by analyzing the data from the clustering.

Table 7: Cluster 2 (C2)

	Borough	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	Sé Bela Vista	Theater	Arts & Crafts Store	Antique Shop	Bookstore	Jazz Club	Movie Theater	College Bookstore	Cultural Center	Concert Hall	Community College
5	República	Theater	Music School	Arts & Crafts Store	Record Shop	Piano Bar	Bookstore	Music Store	Cultural Center	Jazz Club	Camera Store
64	Barra Funda	Theater	Music Venue	Museum	Bus Stop	Arts & Crafts Store	Public Art	Bookstore	Indie Movie Theater	Cultural Center	Movie Theater

Table 8: Cluster 3 (C3)

	Borough	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
3	Consolação	Theater	Dance Studio	Bookstore	College Theater	Movie Theater	Art Gallery	Antique Shop	Arts & Entertainment	Bus Station	Bus Stop

Table 9: Cluster 4 (C4)

	Borough	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
7	Sé	Art Gallery	Cultural Center	Arts & Crafts Store	Bookstore	Historic Site	College Bookstore	Music Venue	Theater	Art Museum	Bus Station
91	Moema	Arts & Crafts Store	Art Gallery	Music Venue	Dance Studio	Art Studio	Art Museum	Arts & Entertainment	Antique Shop	Bookstore	Bus Station

After a deeper analysis in all the clusters and their characteristics, and without forgetting the total amount of venues calculated in Table 4, we can determine the best neighborhoods to invest in an art gallery in São Paulo. This final result is shown in Table 10 below.

Table 10: Top 5 Best Neighborhoods to Open an Art Gallery in São Paulo

Position	Neighborhood Name	*Cluster*
1st	Sé	*C-4*
2nd	Moema	*C-4*
3rd	Consolação	*C-3*
4th	Bela Vista	*C-2*
5th	República	*C-2*

The results are very appropriate, since all these neighborhoods are very artistic in their own way. *Sé*, *Bela Vista*, *Consolação* and *República* are central and old locations, where there are many museums and art galleries. Moema is a new neighborhood, but it is a very hipster place, composed mainly by young people. Therefore, the recommendation above matches perfectly with the reality of São Paulo.

6. Conclusion

This project started with one CSV file containing data from neighborhoods and its geographical coordinates. In addition, it was obtained from foursquare art-related venues of each neighborhood. After that initial phase, a machine learning technique, called *k-means* clusterization was used to group all the similar neighborhoods together and to check which of them were the best ones to open an Art Gallery in São Paulo.

It was possible to see that more central the boroughs have more chance to have many art-related venues, increasing the chances of a good business in the area. Table 10 summarize all the results, showing that *Sé* is the best location for our art gallery. It makes sense, since it is the neighborhood with more art galleries in the city and it the neighborhood with more general art-related venues.

Therefore, *Sé* would certainly be the recommendation of this report for anyone that seeks to open an art gallery in São Paulo. The code, all the data and analysis utilized in this report is available in GitHub [3].

7. References

[1] In: Website: **São Paulo Information**. Avaiable at: < *https://www.britannica.com/place/Sao-Paulo-Brazil* >. Acess date: 06/07/2020.

[2] In: Wikipedia website: **k-means clustering**. Avaiable at: < *https://en.wikipedia.org/wiki/K-means_clustering#:~:text=k%2Dmeans%20clustering%20is%20a,a%20prototype%20of%20the%20cluster* >. Acess date: 06/07/2020.

[3] In: Website: **Git Hub Coursera Casptone Final Task**. Avaiable at: < *https://github.com/thalescastelo/Coursera_Capstone/blob/master/Final_Task.ipynb* >. Acess date: 06/07/2020.