Where the Music Lives

A cluster analysis of live music venues in São Paulo's districts

—

Priscila Brito

IBM Professional Certificate

Capstone Project

June, 2021

Table of Contents

Introduction

Live music will soon make its comeback after almost two years on a hold due to the Covid-19 pandemic. In preparation for this pause to end, and with expectations of a spike in the demand for this kind of entertainment, having an overview of the current state of the music venues in a lively city such as São Paulo can be insightful for all the professionals who work in the live music business:

- **Investors** who want to profit from the rebound of live music may find opportunities for starting up new venues or investing in current ones;
- **Music venues** that are about to reopen to the public can develop strategies to stand out from the competitors in a moment where all venues will be fighting for the audience's attention;
- **Booking agents and managers** can have a broader overview of all the venues where they can book their artists;
- **Music events companies** can spot the most appropriate venues for their events to take place.

That is why this report is presenting a cluster analysis of all the music venues in São Paulo.

Data

To perform this cluster analysis, I will need the location and the type of music venues in São Paulo. So, I'll be using retrieved data from Foursquare based on their music venues categories.

I'll be also working with a list of all São Paulo's districts, as well as their geographical coordinates, since these data will be needed to make the correspondence with data gathered from Foursquare. The coordinates data are those mapped and made available with a public domain license in Kaggle as a JSON file[1].

[1] Available at https://www.kaggle.com/caiobsilva/sp-district-coordinates

Methodology

The following methodology is a summary of the analysis performed on Python[2].

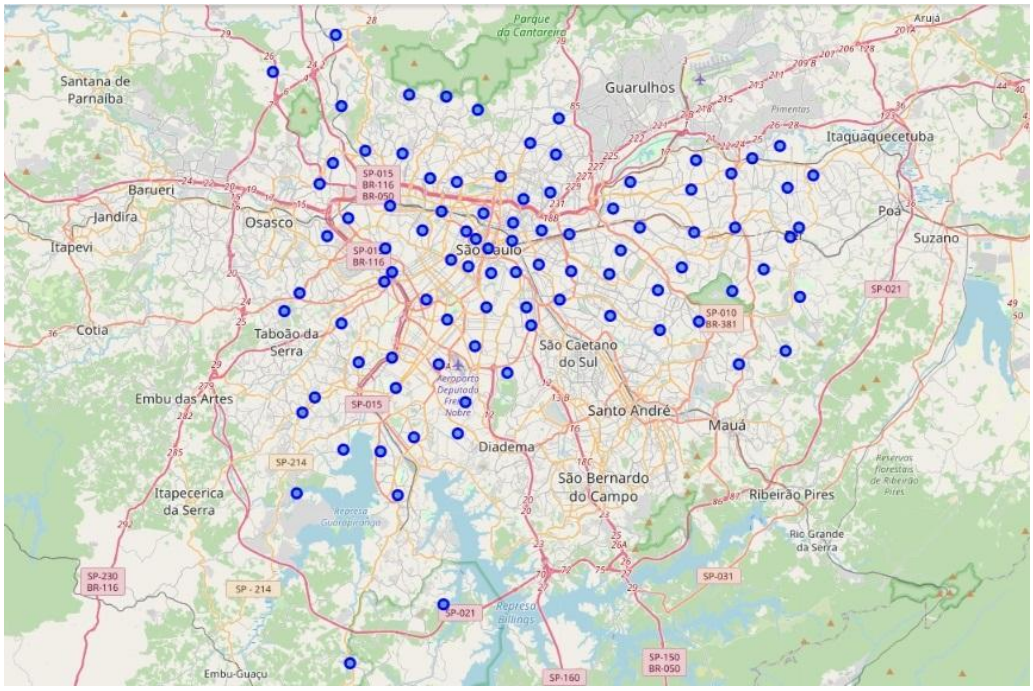
1. Data acquisition

1.1 Coordinates data

After a quick search for datasets on Kaggle, a JSON file with geographical coordinates of São Paulo's districts was found and then converted into a dataframe.

The file contained the geographical coordinates of all 96 districts in São Paulo, plus the corresponding population in each of them. Since the population data was not needed for this analysis, it was eliminated from the file. A simple adjustment to the order of the columns was also applied.

In order to check the accuracy of the data, a map was plotted with markers in each of the 96 districts in São Paulo.



[2] Notebook available at https://github.com/PriscilaBrito/coursera_capstone/blob/master/capstone_project.ipynb

1.2 Music venues data

All the music venues data used in this analysis comes from Foursquare API, which gathers data from several venue types from cities all over the world.

Foursquare divides the venues into categories and subcategories, and music has its special place in this taxonomy. The API lists many music-related categories, such as record stores, music shops, music schools, karaoke bars and so on.

But since this analysis is interested in getting to know São Paulo's live music scene, the scope will be limited to music venues where musicians perform live. Therefore, music venues in São Paulo were retrieved in the following nine categories: Amphitheater, Concert Hall, Music Venue, Jazz Club, Piano Bar, Rock Club, Music Festival, Nightlife Spot and Nightclub.

To perform this task, a function was created to search for the desired venue categories. Another function was defined to match all the venues retrieved to the respective neighborhoods, then the results were put into a dataframe.

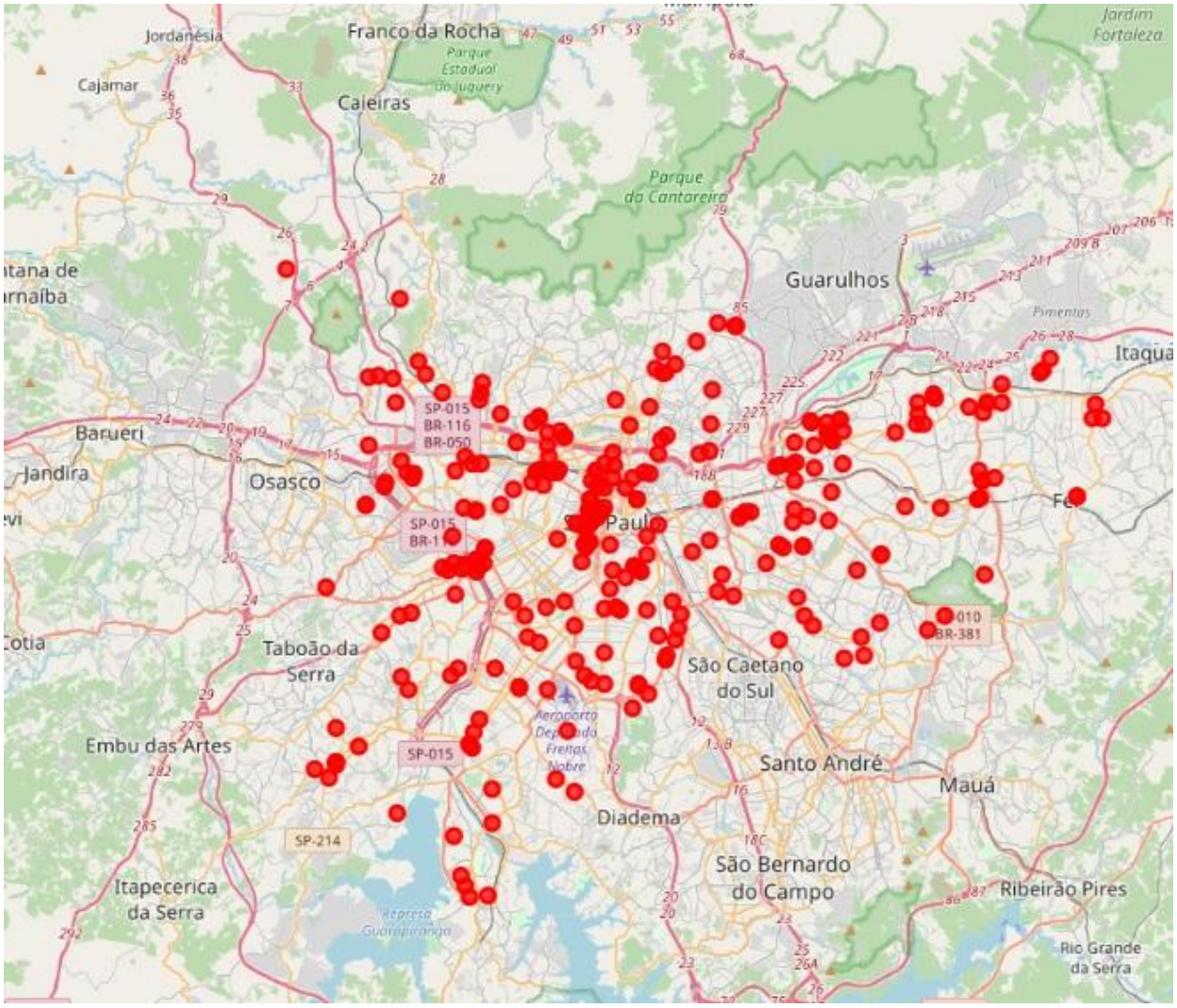
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue Name	Venue ID	Category ID	Venue Category	Venue Latitude	Venue Longitude	Venue City
0	Sapopemba	-23.604327	-46.509885	Bar do Rock	4eace4d86da1a0e3a44720a2	4bf58dd8d48988d116941735	Bar	-23.604877	-46.509026	São Paulo
1	Sapopemba	-23.604327	-46.509885	Willie Dixon	4f2c9904e4b0124ab79ff557	4bf58dd8d48988d116941735	Bar	-23.604973	-46.508946	São Paulo
2	Sapopemba	-23.604327	-46.509885	Amorin's Escola de Musica E Artes	4f720b8be4b09a109634ca37	4bf58dd8d48988d1e5931735	Music Venue	-23.600253	-46.516090	NaN
3	Sapopemba	-23.604327	-46.509885	Retrô Bar e Lanches	577ac990cd108f2f161b6704	4bf58dd8d48988d116941735	Bar	-23.614113	-46.511894	São Paulo
4	Sapopemba	-23.604327	-46.509885	Faive - Boate do Di	5039b27ce4b03fdd88e3f1a8	4bf58dd8d48988d11f941735	Nightclub	-23.604420	-46.514323	NaN

1. Data cleaning

Some data cleaning was performed in order to keep only the necessary data. Unnecessary venues categories non-related to music that were retrieved by Foursquare API were removed. All entries that were not from venues in São Paulo and had missing values were also eliminated from the data.

Is worth mentioning that, although the goal was to retrieve nine music-related venue categories, Foursquare retrieved eight of them and left one out. The non-found category was "Music Festival".

After the cleaning, the data comprised 267 unique music-related venues in São Paulo in eight categories. The map below shows the distribution of the venues across the city.



3. Data pre-processing

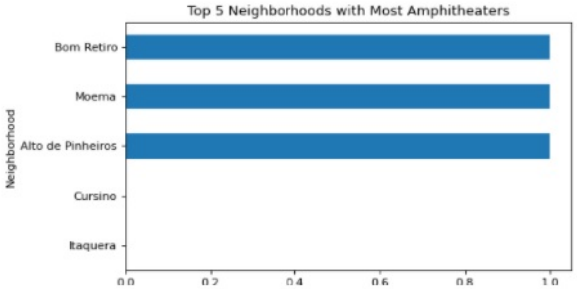
Before moving forward with the K-means analysis, it was important to convert the results into numerical values, so the algorithm would be able to properly read the data. The get_dummies() function was applied for this purpose.

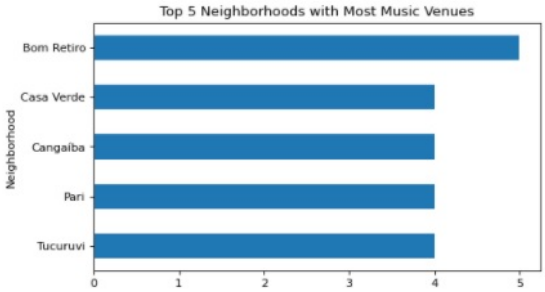
	Neighborhood	Amphitheater	Concert Hall	Jazz Club	Music Venue	Nightclub	Nightlife Spot	Piano Bar	Rock Club
5	Sapopemba	0	0	0	0	1	0	0	0
33	Sapopemba	0	0	0	0	0	0	0	1
34	Sapopemba	0	0	0	1	0	0	0	0
35	Sapopemba	0	0	0	1	0	0	0	0
56	Capão Redondo	0	0	0	1	0	0	0	0

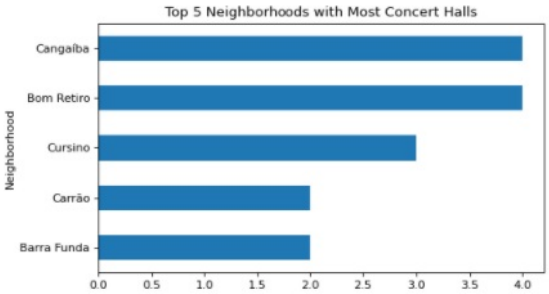
4. Exploratory analysis

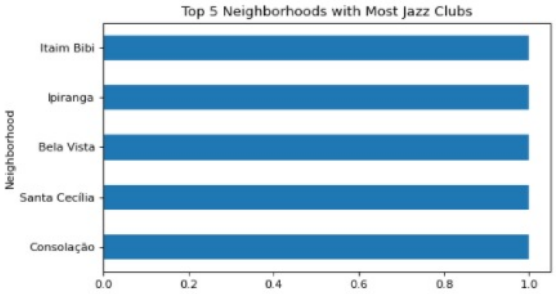
Apart from the K-means analysis that will follow soon in this report, the numerical values obtained in the previous step were useful for conducting an exploratory analysis of the data.

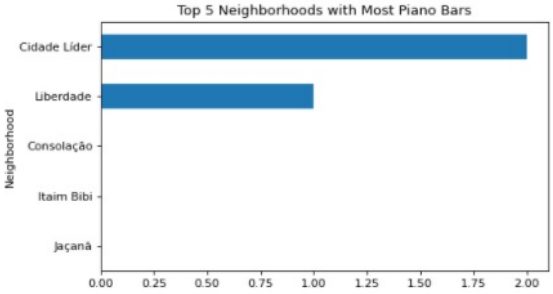
For each music venue category retrieved, a graphic was plotted to identify the top 5 neighborhoods for a given venue category, as the image shows in the next page.

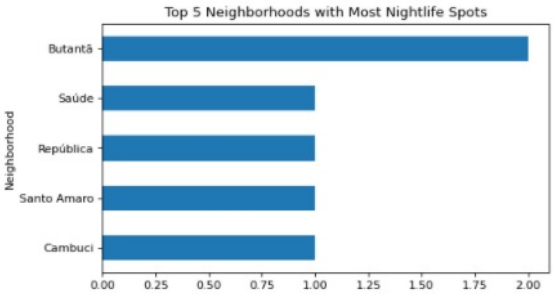

Top 5 Neighborhoods with Most Amphitheaters
Neighborhood
Bom Retiro
Moema
Alto de Pinheiros
Cursino
Itaquera
0.0
0.2
0.4
0.6
0.8
1.0

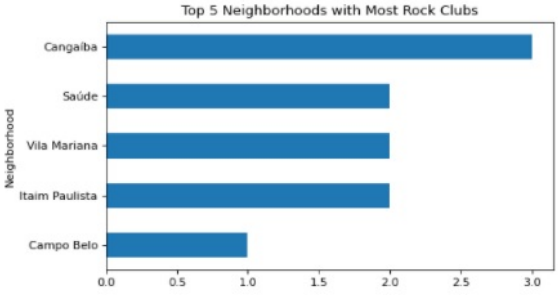

Top 5 Neighborhoods with Most Music Venues
Neighborhood
Bom Retiro
Casa Verde
Cangaíba
Pari
Tucuruvi
0
1
2
3
4
5

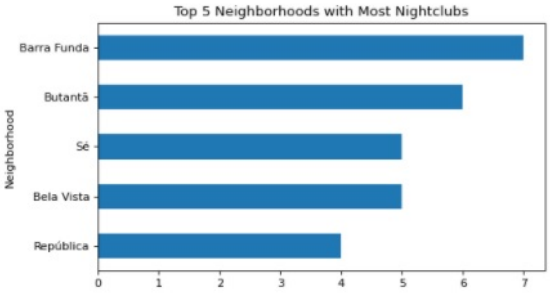

Top 5 Neighborhoods with Most Concert Halls
Neighborhood
Cangaíba
Bom Retiro
Cursino
Carrão
Barra Funda
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Top 5 Neighborhoods with Most Jazz Clubs
Neighborhood
Itaim Bibi
Ipiranga
Bela Vista
Santa Cecília
Consolação
0.0
0.2
0.4
0.6
0.8
1.0

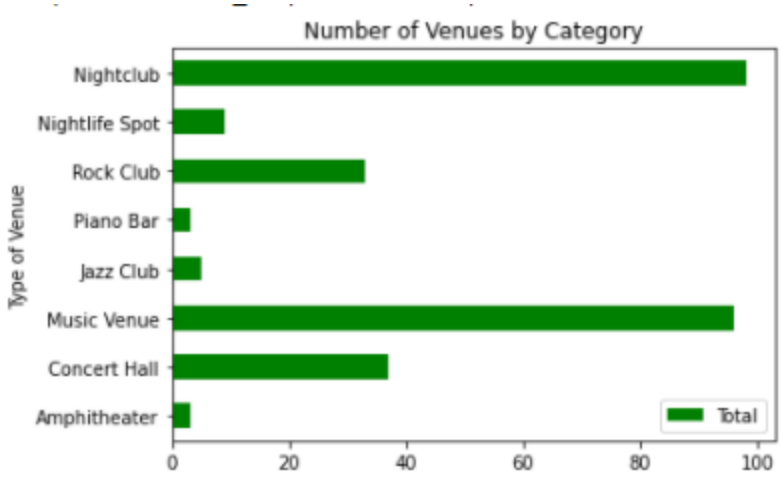

Top 5 Neighborhoods with Most Piano Bars
Neighborhood
Cidade Líder
Liberdade
Consolação
Itaim Bibi
Jaçanã
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Top 5 Neighborhoods with Most Nightlife Spots
Neighborhood
Butantã
Saúde
República
Santo Amaro
Cambuci
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Top 5 Neighborhoods with Most Rock Clubs
Neighborhood
Cangaíba
Saúde
Vila Mariana
Itaim Paulista
Campo Belo
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Top 5 Neighborhoods with Most Nightclubs
Neighborhood
Barra Funda
Butantã
Sé
Bela Vista
República
0
1
2
3
4
5
6
7

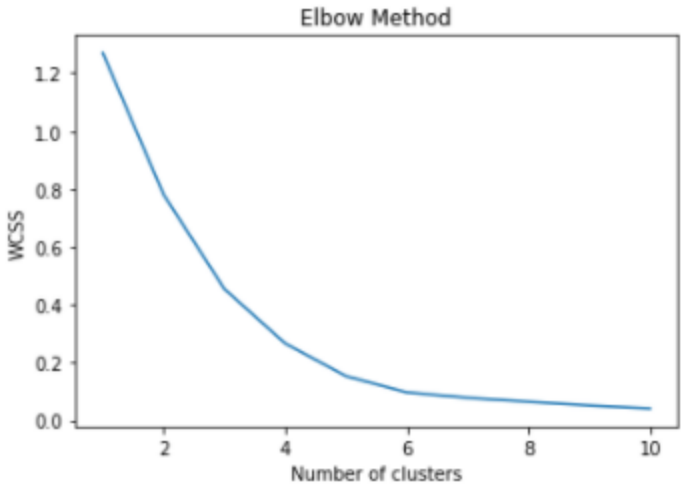
The distribution of all music venue categories across neighborhoods in São Paulo was also plotted.



5. Clustering with K-means

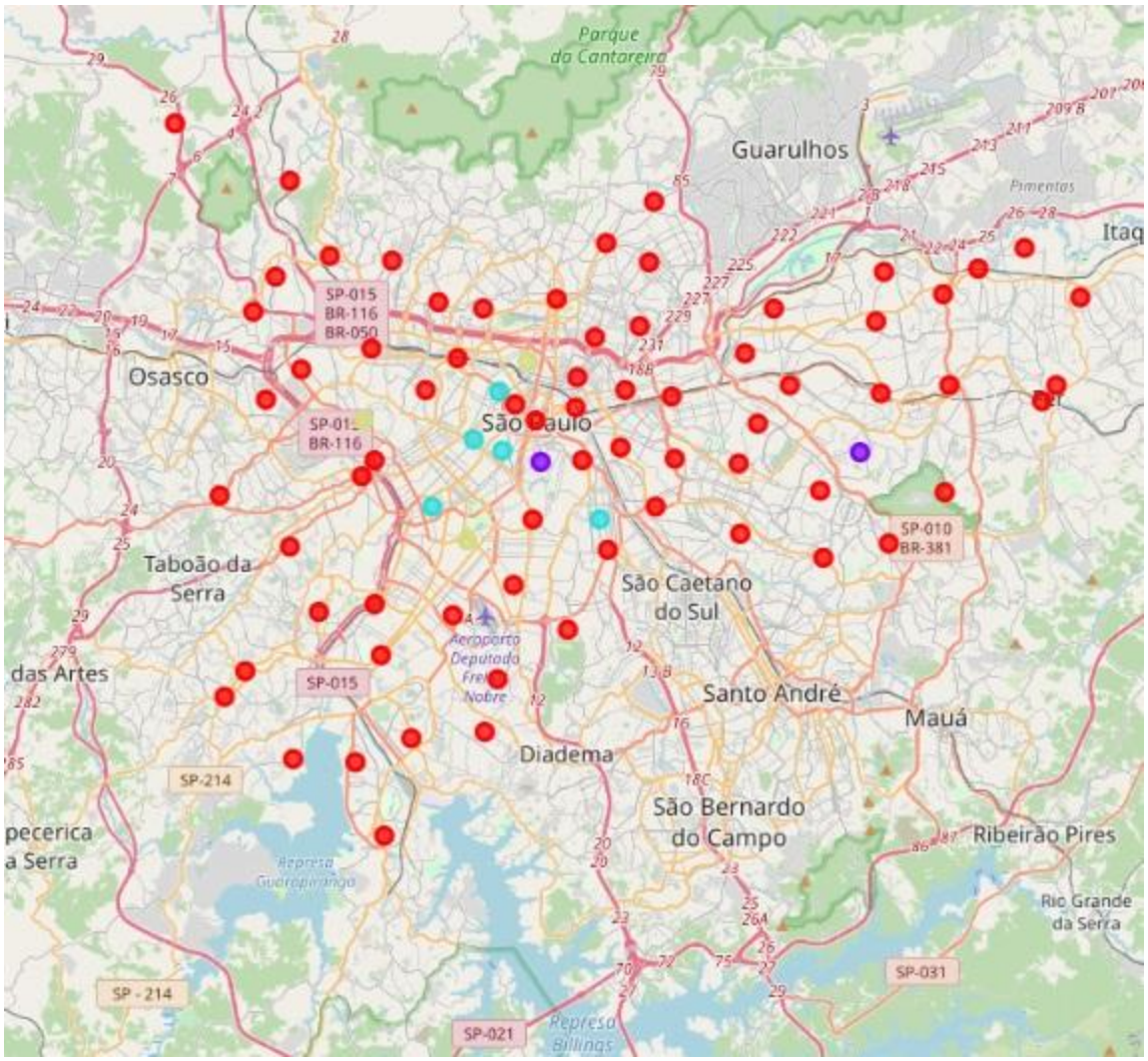
With the numerical values assigned to the music venues category, the final step for the analysis before running the k-means algorithm was to calculate the mean for each category.

Then, the elbow method was applied to the dataset to determine how many clusters should be used for this analysis, and the result was 4.



After running the algorithm and plotting the results into a map, one can see a large cluster spread all over São Paulo and three other small clusters around the central area of the city.

The next section will dive deeper into each of these clusters in order to profile them.



Results

1. The 4 “official” clusters

The k-means algorithm provided the following clusters, as profiled below:

Music Mixing

This cluster is the largest one found in the analysis. It's also the most diverse in terms of types of venues. It represents a suitable overview of the live music scene that one would expect from one of the biggest cities in the world.

Exactly 69 neighborhoods are grouped in this giant cluster and one can note that diversity is the pattern found here. There is no prevalent music venue category: rock clubs, concert halls, nightclubs, nightlife spots, piano bars and general music venues can be spotted in all the neighborhoods, and all these categories fluctuate between the 1st and 5th positions for the most common venues.

	Districts	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Sapopemba	Rock Club	Music Venue	Nightclub	Piano Bar	Nightlife Spot
5	Capão Redondo	Music Venue	Nightclub	Rock Club	Piano Bar	Nightlife Spot
6	Jardim São Luís	Rock Club	Piano Bar	Nightlife Spot	Nightclub	Music Venue
7	Cidade Ademar	Concert Hall	Nightclub	Rock Club	Piano Bar	Nightlife Spot
8	Itaim Paulista	Rock Club	Nightclub	Piano Bar	Nightlife Spot	Music Venue
...	...	...	...	...	...	...
91	Brás	Concert Hall	Music Venue	Rock Club	Piano Bar	Nightlife Spot
92	Jaguara	Nightclub	Rock Club	Piano Bar	Nightlife Spot	Music Venue
93	Sé	Nightlife Spot	Nightclub	Music Venue	Rock Club	Piano Bar
94	Pari	Music Venue	Nightclub	Rock Club	Piano Bar	Nightlife Spot
95	Barra Funda	Nightclub	Concert Hall	Music Venue	Rock Club	Piano Bar

69 rows × 6 columns

A Bit of Piano

This small cluster distinguishes itself from the others by the recurrency of piano bars as the first most common venue.

	Districts	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
30	Cidade Líder	Piano Bar	Rock Club	Nightlife Spot	Nightclub	Music Venue
76	Liberdade	Piano Bar	Concert Hall	Music Venue	Rock Club	Nightlife Spot

All That Jazz

Another small cluster and another distinct feature: this time, what makes this cluster unique is the number of jazz clubs as the most common music venue in all the neighborhoods.

	Districts	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
50	Ipiranga	Jazz Club	Music Venue	Rock Club	Piano Bar	Nightlife Spot
62	Itaim Bibi	Jazz Club	Rock Club	Piano Bar	Nightlife Spot	Nightclub
70	Santa Cecília	Jazz Club	Concert Hall	Nightclub	Rock Club	Piano Bar
77	Bela Vista	Jazz Club	Nightclub	Music Venue	Rock Club	Piano Bar
81	Consolação	Jazz Club	Rock Club	Piano Bar	Nightlife Spot	Nightclub

For the Crowds

The final cluster has also a very clear pattern, with amphitheaters being the first most common venue, concert halls being the second one and regular music venues being the third.

	Districts	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
67	Moema	Amphitheater	Concert Hall	Music Venue	Nightclub	Rock Club
84	Alto de Pinheiros	Amphitheater	Music Venue	Nightclub	Rock Club	Piano Bar
90	Bom Retiro	Amphitheater	Concert Hall	Music Venue	Nightclub	Rock Club

2. The “missing” cluster

When cleaning the data after retrieving data from Foursquare API, only neighborhoods with music venue data were left in the dataset. More precisely, 79 out of 96 neighborhoods were left. The other 16 did not have any music-related venues retrieved from Foursquare.

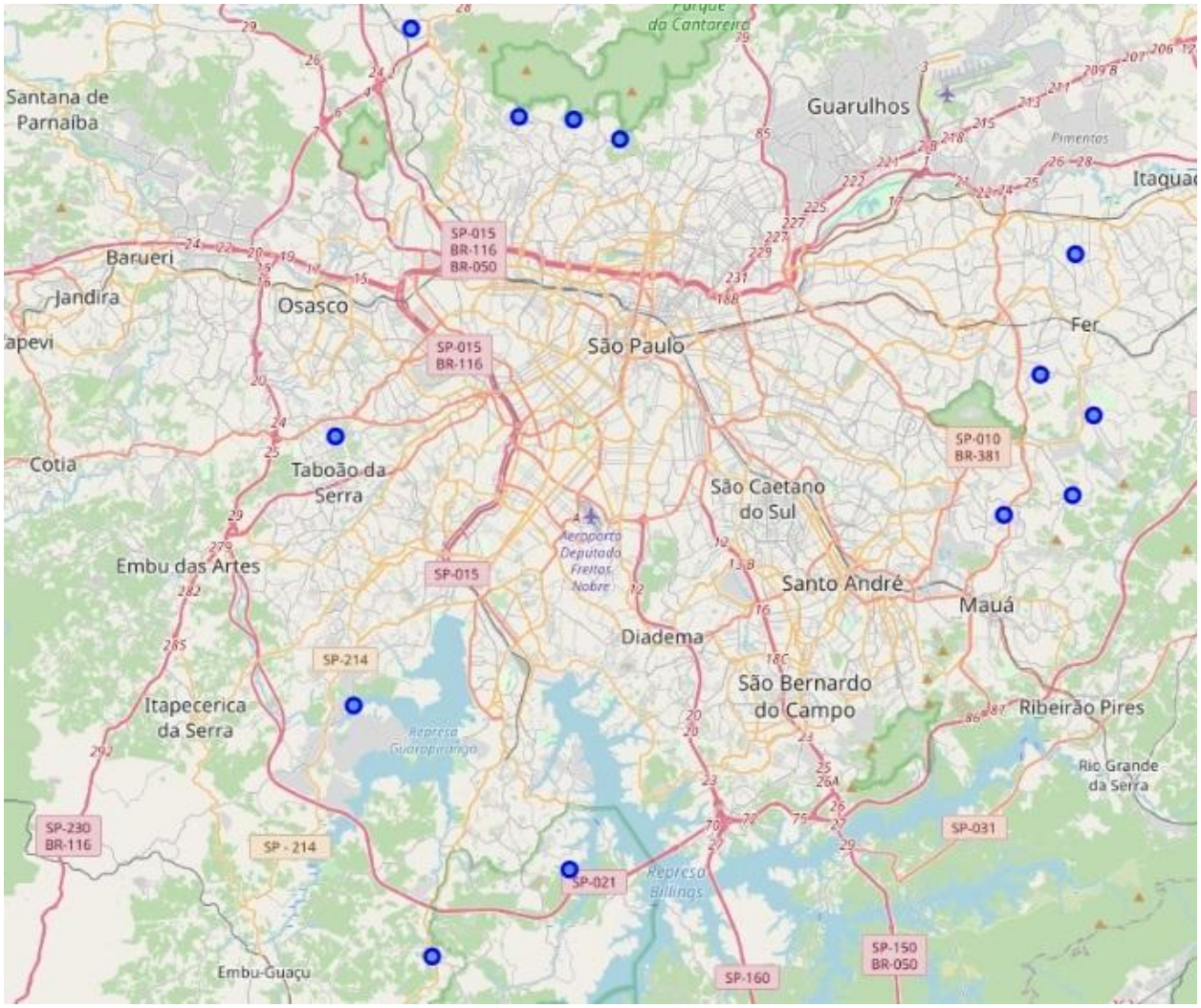
Having no music venues makes this group of neighborhoods homogeneous, and also worth noting for this analysis, even though they are not part of the "official" clustering analysis run by the algorithm.

Since the aim of this report is to bring some insights into the live music venues scene in São Paulo, taking into account places that have no music venues at all might be just as important information as considering the neighborhoods with music venues X or Y. Hence, this “cluster” is also being profiled, apart from the other four.

	Districts	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Grajaú	-23.785798	-46.665575	NaN	NaN	NaN	NaN	NaN	NaN
3	Jardim Ângela	-23.712246	-46.771206	NaN	NaN	NaN	NaN	NaN	NaN
4	Brasilândia	-23.448272	-46.690269	NaN	NaN	NaN	NaN	NaN	NaN
11	Cidade Tiradentes	-23.582497	-46.409207	NaN	NaN	NaN	NaN	NaN	NaN
16	Tremembé	-22.957140	-45.547526	NaN	NaN	NaN	NaN	NaN	NaN
20	Vila Curuçá	-23.510151	-46.417893	NaN	NaN	NaN	NaN	NaN	NaN
21	Pedreira	-22.741347	-46.894846	NaN	NaN	NaN	NaN	NaN	NaN
23	Cachoeirinha	-23.449337	-46.663636	NaN	NaN	NaN	NaN	NaN	NaN
26	São Rafael	-23.627159	-46.453241	NaN	NaN	NaN	NaN	NaN	NaN
27	Parelheiros	-23.824791	-46.733078	NaN	NaN	NaN	NaN	NaN	NaN
32	Iguatemi	-23.618271	-46.419028	NaN	NaN	NaN	NaN	NaN	NaN
40	José Bonifácio	-23.564091	-46.434767	NaN	NaN	NaN	NaN	NaN	NaN
43	Mandaqui	-23.457940	-46.641243	NaN	NaN	NaN	NaN	NaN	NaN
47	Raposo Tavares	-23.591765	-46.780607	NaN	NaN	NaN	NaN	NaN	NaN
60	Perus	-23.408492	-46.743632	NaN	NaN	NaN	NaN	NaN	NaN
64	Jardim Paulista	-22.195211	-46.740504	NaN	NaN	NaN	NaN	NaN	NaN
96	Marsilac	-23.937142	-46.710230	NaN	NaN	NaN	NaN	NaN	NaN

The Sound of Silence

This fifth group of neighborhoods has two features in common: besides having no music venues at all, they are located in the extremes of the city of São Paulo, as shown in the map below.


Santana de Parnaíba
Guarulhos
Barueri
Jandira
Osasco
São Paulo
Cotia
Taboão da Serra
São Caetano do Sul
Embu das Artes
Santo André
Mauá
Diadema
São Bernardo do Campo
Ribeirão Pires
Itapecerica da Serra
Rio Grande da Serra
Embu-Guaçu
SP-015
BR-116
BR-050
SP-010
BR-381
SP-214
SP - 214
SP-230
BR-116
SP-021
SP-031
SP-150
BR-050
SP-160
Aeroporto Deputado Freitas Nobre
Represa Guarapiranga
Represa Billings

Discussion

From the results above, some challenges and opportunities come up.

Challenges

High competition

As pointed out above, the *Music Mixing* cluster is essentially diverse, as it has several types of music venues, and great in number. This may be excellent for the customers, who have many options at hand, but a challenge for the venues.

When you have such a variety of options spread all over the city, competition for the public is high. Standing out in the eyes of the customers is the way out to thrive in such a crowded scene, but this requires strategic efforts from the business and further investigation on what areas should be improved.

Investment in a time of crises

Improvements in any kind of business, as highlighted above, sometimes lead to financial investments (marketing, facilities renovation, special programming), and for music venues this can be critical in a time when many of them have been struggling with little or no revenue at all due to the Covid-19 crisis.

Opportunities

Room for niche venues and experimentation

Both the *A Bit of Piano* and *All That Jazz* clusters rise as niche clusters. The areas where they are located could potentially bring opportunities for experimenting with other niche-oriented venues.

The hotspot

The cluster profiled as *For the Crowds* can potentially be a hotspot in the reopening in the post pandemic, since it distinguishes from the others for the music venues with larger capacities, probably the kind of live music entertainment that many people are more eager to resume.

Equal access to live music

The “missing” cluster, *The Sound of Silence* is the one with no music venues. However, by no means this is equivalent to non-existent demand for music-related venues.

These neighborhoods are more likely to lack music venues due to their locations. They are farthest from the city centre and usually these places are poorer areas dismissed by public policies and commercial ventures.

Therefore, these areas should be seen by investors as a business opportunity. And more than a business opportunity, this could also be a chance to assure equal access to live music for people in the whole city, regardless of where they live.

A final note

The data analysis presented in this report should be considered as exploratory. A further analysis with additional data from Foursquare may provide even more insights. Also, bringing additional data sources could help verify to what extent Foursquare data is complete.

Since the API was not able to find any music festivals in São Paulo, a city that knowingly host the vast majority of music festivals in Brazil and South America, including some international franchises like Lollapalooza, we should assume that the live music scene in the city is not entirely represented by Foursquare API data.

Conclusion

This report made a clustering analysis of music venues in São Paulo in order to have an overview of the current live music scene in the city.

As expectations grow for the reopening of this kind of venue, such analysis can be insightful for people in the live business who will eventually be able to resume their activities.

The results showed that there are some challenges and opportunities in this context. A high competition in a larger portion of the city and the need for investments as a strategy to stand out in the crowd are some of the potential challenges.

On the other hand, some opportunities are presented for niche venues, venues with larger capacities and for investors who may be willing to invest in underprivileged areas.

As insightful as it can be, this analysis is not comprehensive and additional data could be brought up to deepen the findings.