São Paulo Foursquare Data Clustering

Lucas Hosoya

Github: https://github.com/ShigueruHosoya
Contact: shigaaaa@gmail.com

The Data

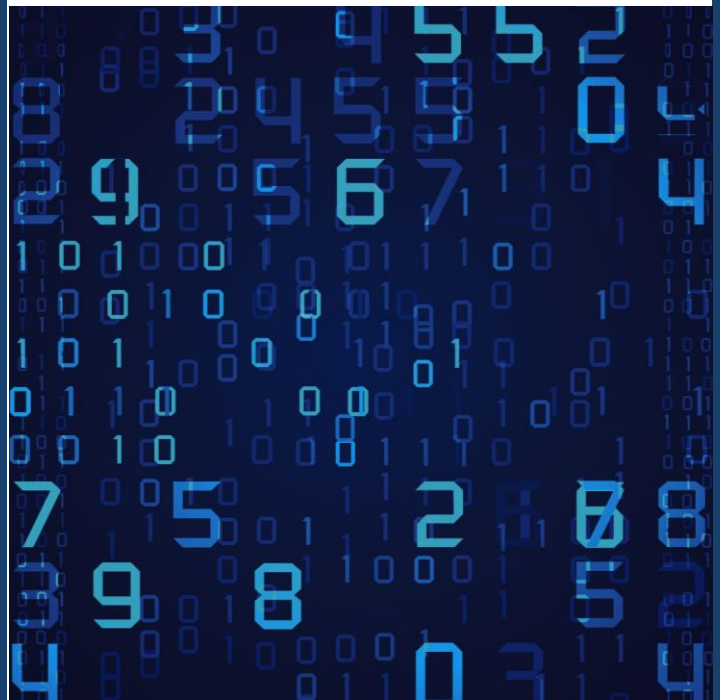
- São Paulo Districts
- São Paulo top locations to work.
- São Paulo top locations to live.

- All the data was gathered using data scraping with BeautifulSoup, Foursquare API explorer and Geopy for geographic data.

Fonts: Please review the links.txt on the github link for the links.

The problem

- Where would be the best Neighborhoods/Districts to live in São Paulo city, located in Brazil?
- Where would be the best Neighborhoods/Districts to work in São Paulo city, located in Brazil?
- From the decisions above, why the named districts?



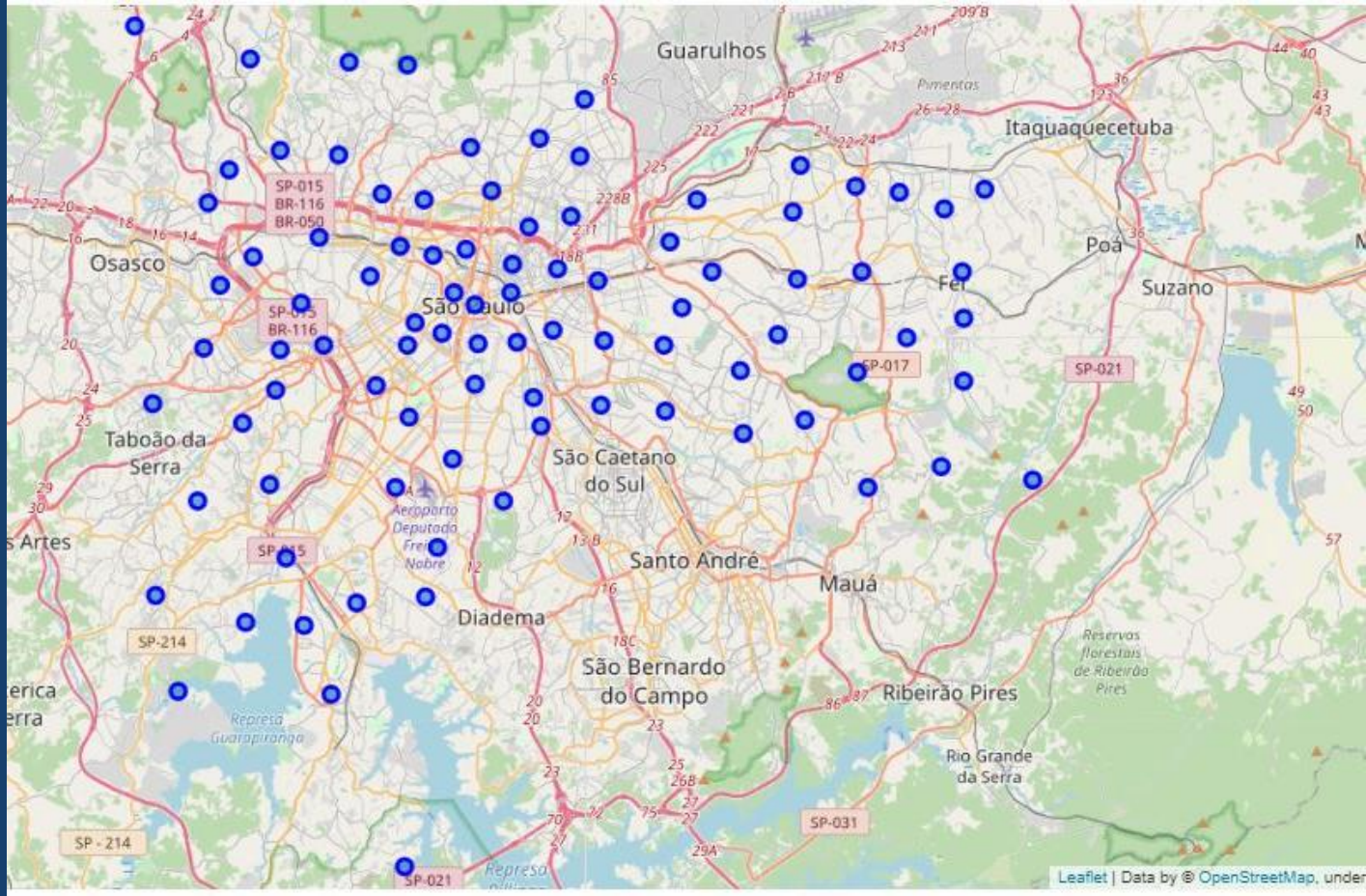
About it

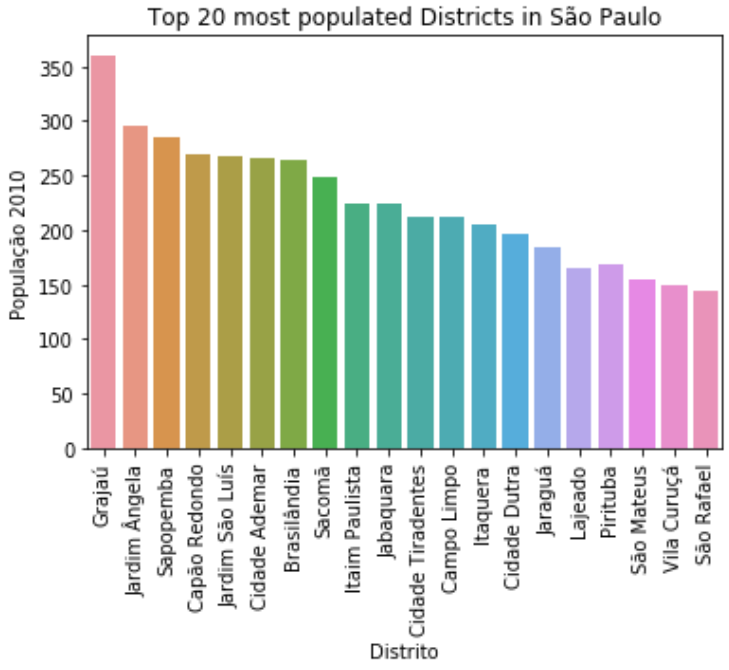
São Paulo is a fast growing city/state in the last decades and many people are looking to change their lives by moving to São Paulo.

Population Analysis

A view of São Paulo's population by District in the image below, the population is shown in thousands. By living in São Paulo we know that most of people live in these areas, but most of the population live by the suburbs.

In the image Above we can view the selected Districts. The most populated districts are in the suburbs located by the Northwest, Southwest and East locations.

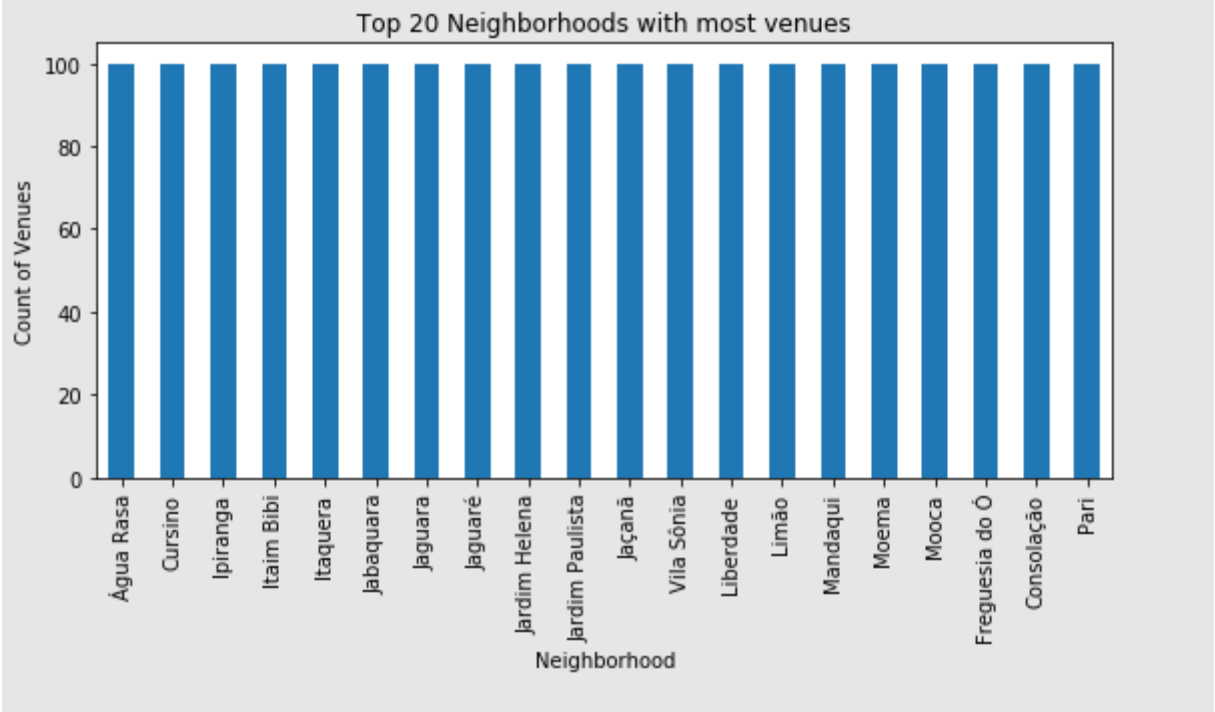


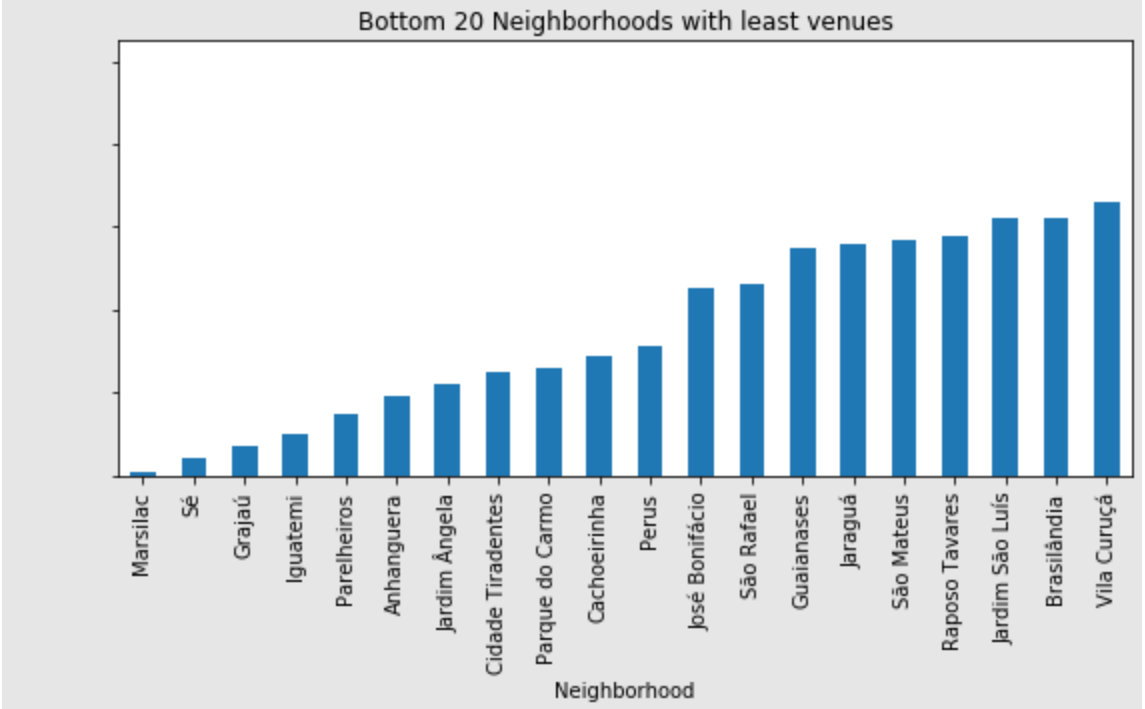


Gathering Data Problems

After gathering data, there's the problem by gathering more data from Foursquare API in which explores the districts, but we cannot have the right precision for these districts due to the count of venues.

As we can see, the bottom 20 neighborhoods does not present many venues, interfering the processes for clustering the neighborhoods.





Data Clustering and Analysis – PREPARATION

Before Clustering, the data is prepared by encoding all the data from Foursquare by venue types and then grouped by the most frequent/common venues represented in the images.

By doing the encoding and then grouping the frequency of the venues, we can see the distribution of venues by a determined neighborhood, in which we are going to cluster by groups later, based on their frequency.

The last image below shows the distribution by ranking the most common venues by neighborhood in which will be utilized for the next analysis and clustering.

```
saopaulo_grouped = saopaulo_onehot.groupby('Neighborhood').mean().reset_index()
saopaulo_grouped
```

	Neighborhood	Zoo Exhibit	Acai House	Accessories Store	Airport	Airport Lounge	Airport Service	American Restaurant	Amphitheater	Aquarium	...	Warehouse Store	Watch Shop	Water Park	Whisky Bar	Wine Bar
0	Alto de Pinheiros	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.00	0.0	0.0	0.0	0.00
1	Anhanguera	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.00	0.0	0.0	0.0	0.00
2	Aricanduva	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.01	0.0	0.0	0.0	0.00
3	Artur Alvim	0.00	0.00	0.01	0.0	0.0	0.0	0.00	0.00	0.0	...	0.00	0.0	0.0	0.0	0.00
4	Barra Funda	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.00	0.0	0.0	0.0	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
89	Vila Matilde	0.01	0.01	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.00	0.0	0.0	0.0	0.00
90	Vila Medeiros	0.00	0.00	0.00	0.0	0.0	0.0	0.01	0.00	0.0	...	0.00	0.0	0.0	0.0	0.00
91	Vila Prudente	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.01	0.0	0.0	0.0	0.00
92	Vila Sônia	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.01	0.0	...	0.00	0.0	0.0	0.0	0.00
93	Água Rasa	0.00	0.00	0.00	0.0	0.0	0.0	0.00	0.00	0.0	...	0.00	0.0	0.0	0.0	0.01

94 rows × 364 columns

```
----Alto de Pinheiros----
                venue  freq
0               Plaza  0.07
1  Gym / Fitness Center  0.04
2          Restaurant  0.04
3   Athletics & Sports  0.04
4             Dog Run  0.04
5      Clothing Store  0.03
6      Cosmetics Shop  0.02
7                 Spa  0.02


----Anhanguera----
                venue  freq
0       Grocery Store  0.16
1             Bakery  0.11
2               Park  0.11
3                Gym  0.05
4         Pizza Place  0.05
5  Gym / Fitness Center  0.05
6   Convenience Store  0.05
7   Food & Drink Shop  0.05
```

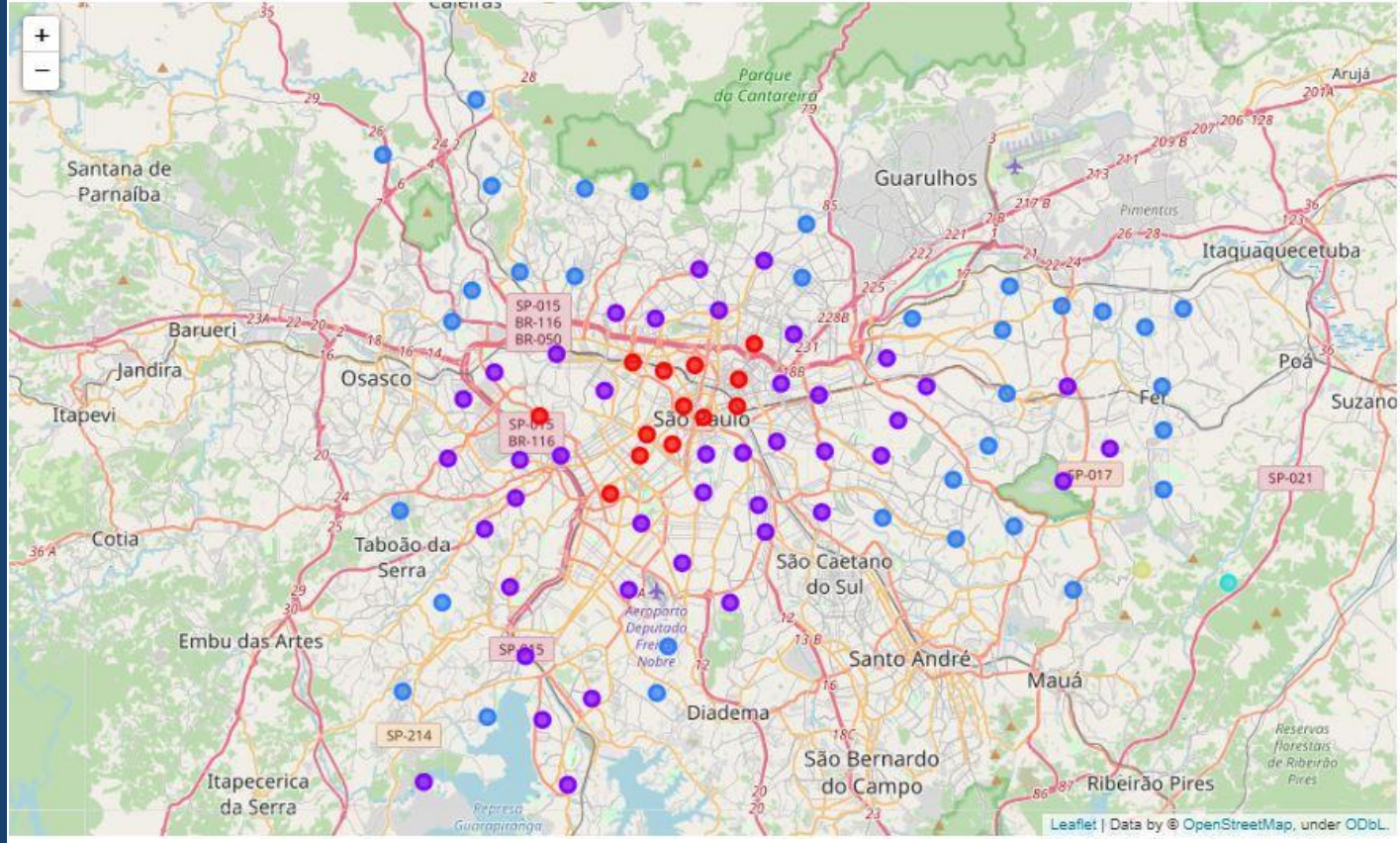
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue
0	Alto de Pinheiros	Plaza	Gym / Fitness Center	Restaurant	Dog Run	Athletics & Sports	Clothing Store	Cosmetics Shop	Track
1	Anhanguera	Grocery Store	Bakery	Park	Restaurant	Gym / Fitness Center	Burger Joint	Gym	Skate Park
2	Aricanduva	Bakery	Pizza Place	Café	Gym / Fitness Center	Food Truck	Supermarket	Gym	Clothing Store
3	Artur Alvim	Bakery	Pizza Place	Gym / Fitness Center	Clothing Store	Chocolate Shop	Pharmacy	Gymnastics Gym	Department Store
4	Barra Funda	Pizza Place	Dessert Shop	Motel	Italian Restaurant	Restaurant	Café	Hotel	Pet Store

Data Clustering and Analysis – Model and Cluster

Since there are many "types" of districts, the decision was to divide the cluster in seven (7) for a better analysis and decision making.

As we can see the image above, the distribution of the cluster by color of the different regions. Even though there are 7 clusters, we can see that we have a lot of common neighborhoods near the center of the map.

The other colors show the suburbs by their different diversity from the center.



Selecting cluster model with Kmeans to get clusters of similar districts

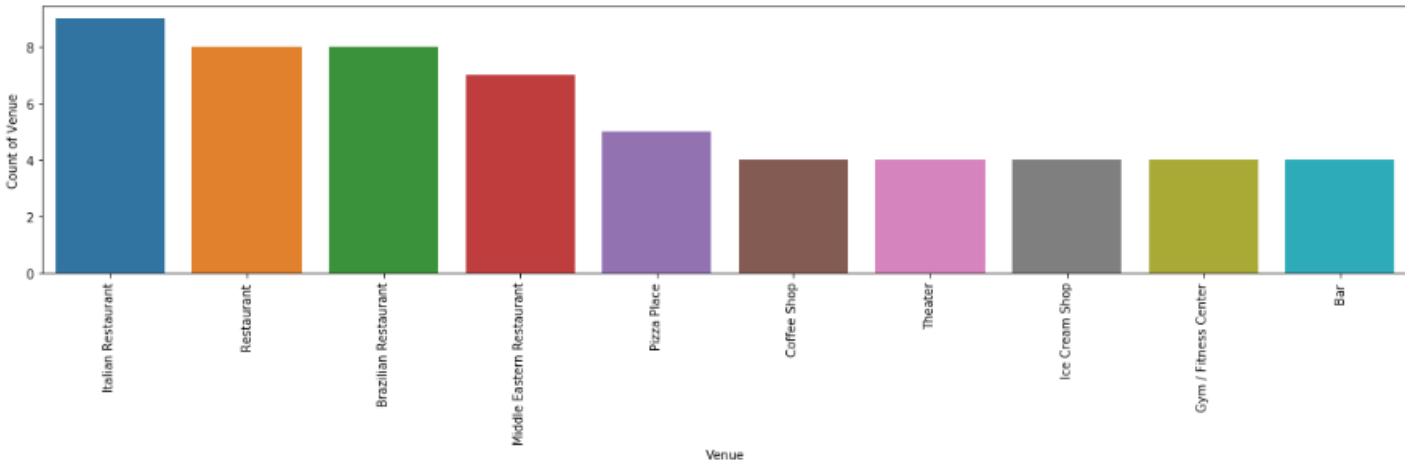
```
from sklearn.cluster import KMeans
# set number of clusters
kclusters = 7

saopaulo_grouped_clustering = saopaulo_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(saopaulo_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

```
array([0, 2, 2, 2, 0, 0, 1, 0, 2, 0])
```

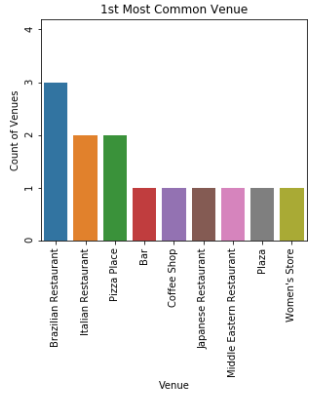
Data Clustering and Analysis – Deep Analysis

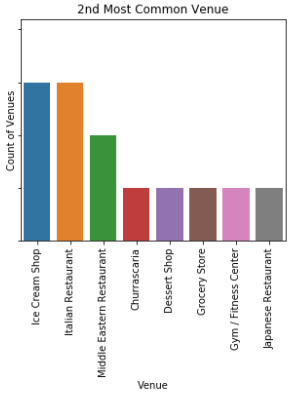
To get into conclusions, first we need to analyze the clustered data in order to know the best places to work/live in São Paulo. Based on that, we'll see the data by three main clusters (Cluster 0, Cluster 1 and Cluster 2), because they have more data/neighborhoods, just like we saw on the previous slide by the most common colors in the map.

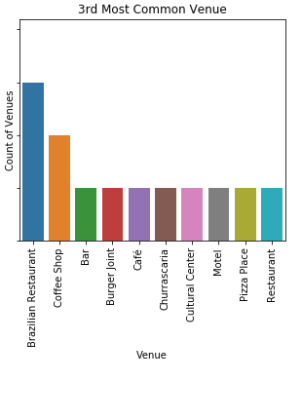
Since the visualizations are pretty difficult to visualize, please visualize via jupyter notebook or we'll put three more slides with bigger images.

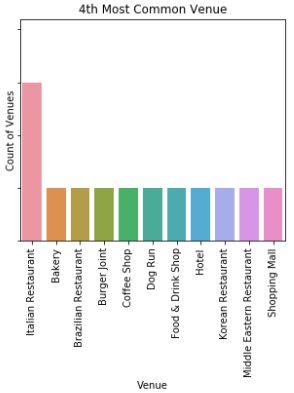


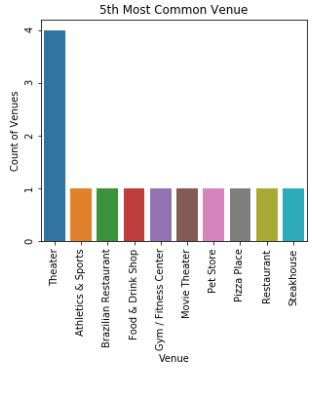
Cluster 0 - Red

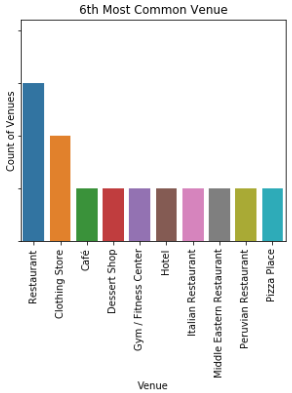


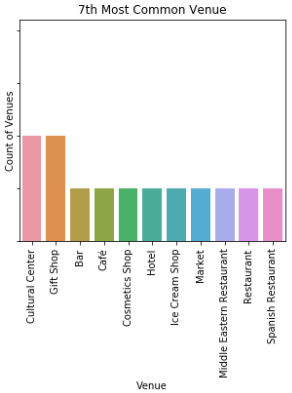


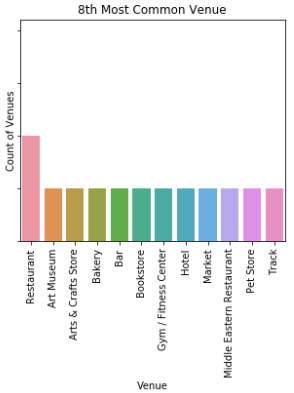


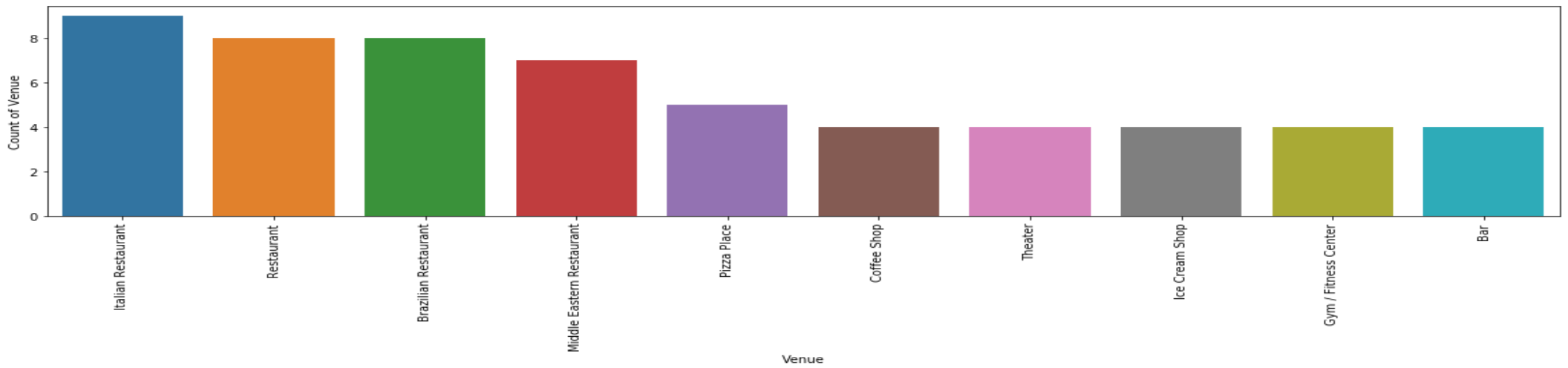




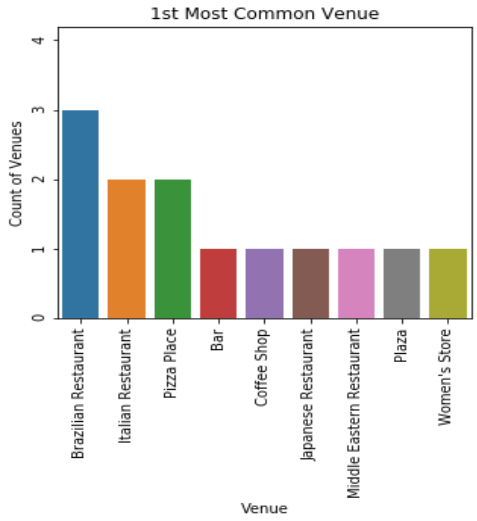


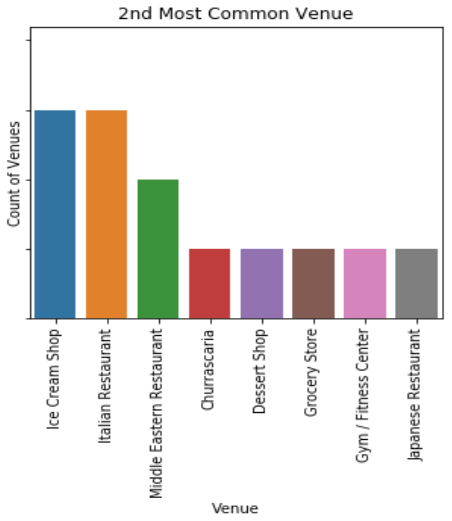


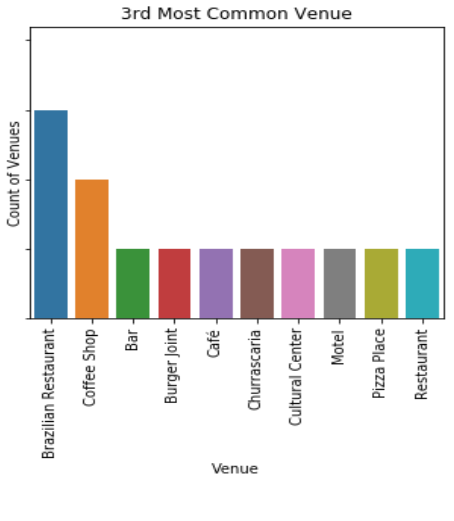


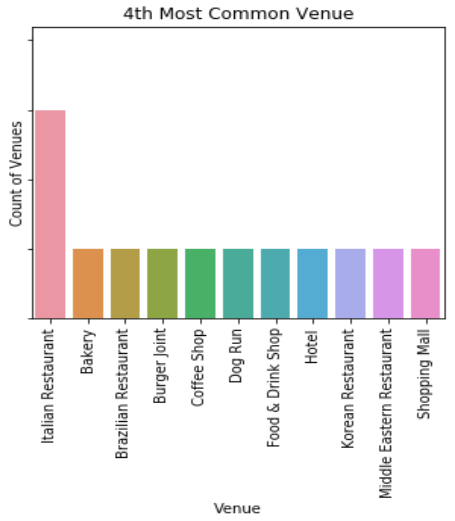


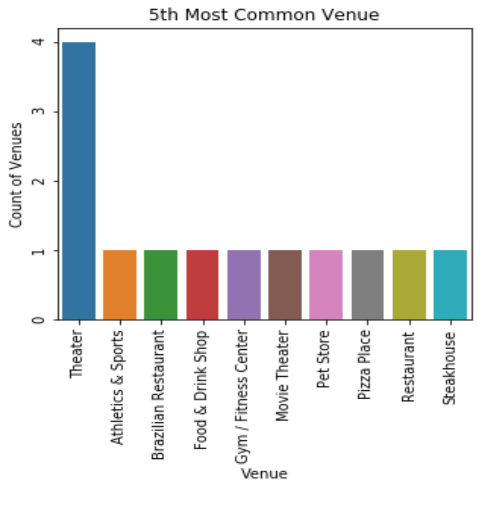
Cluster 0 - Red

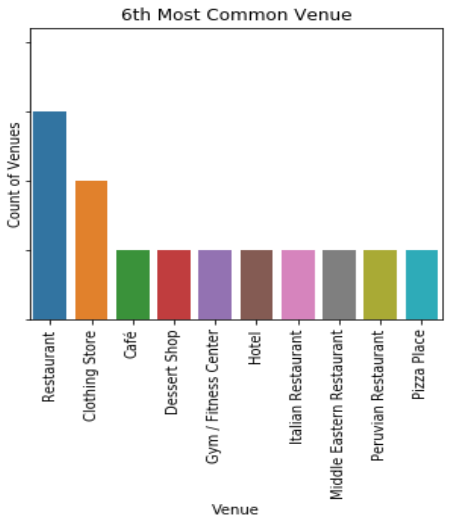


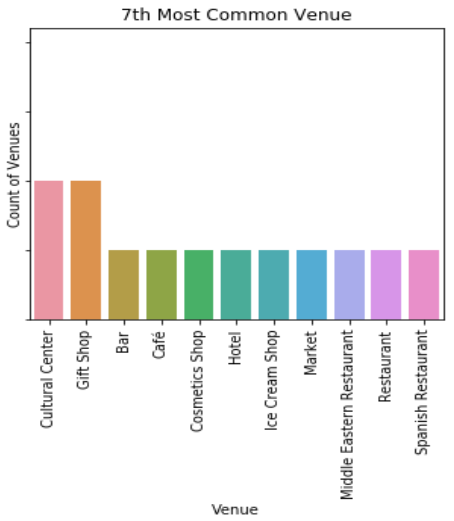


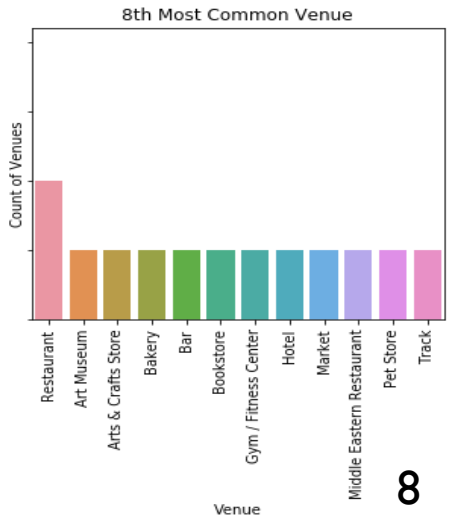


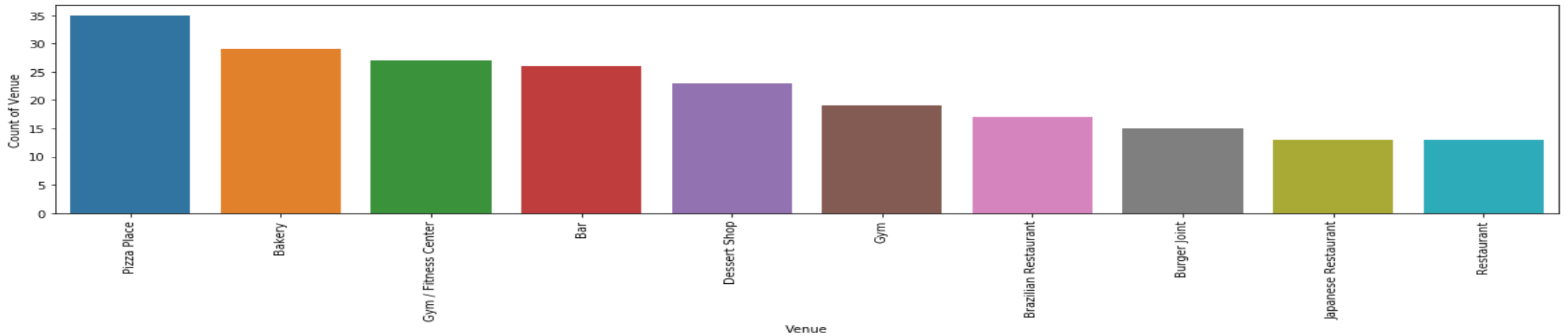




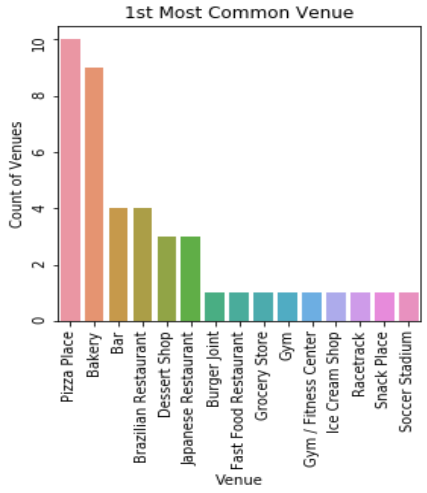


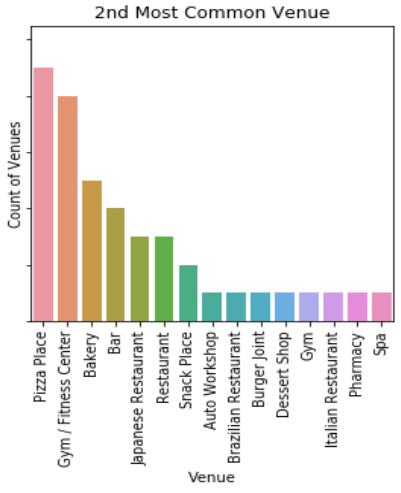


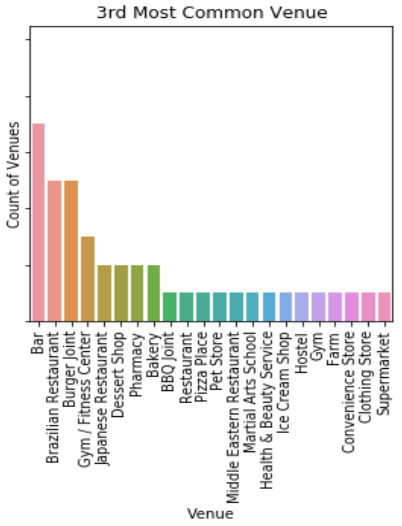


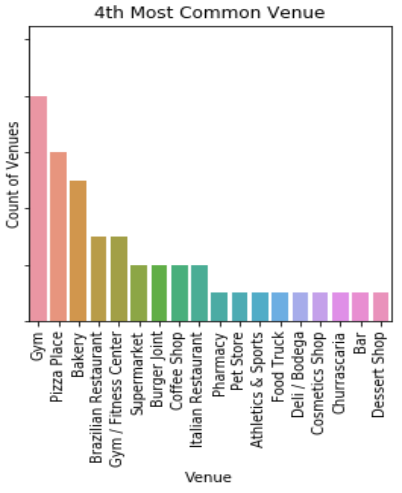


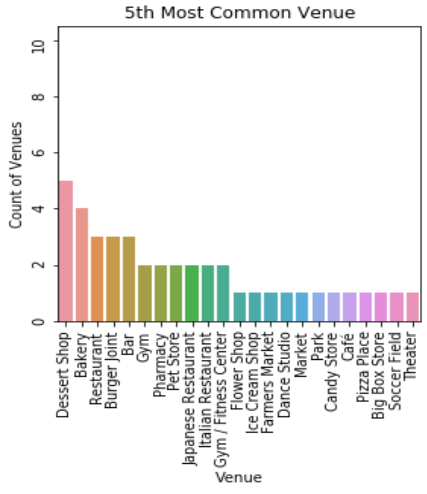
Cluster 1 - Purple

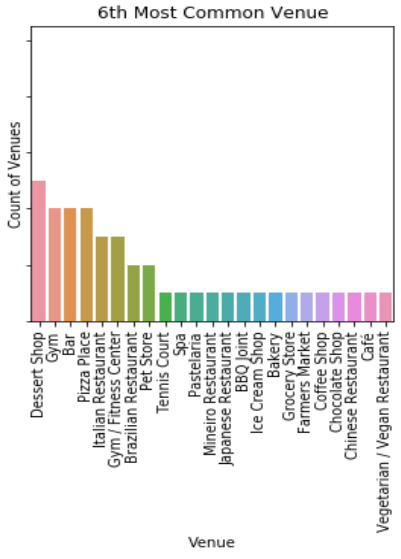


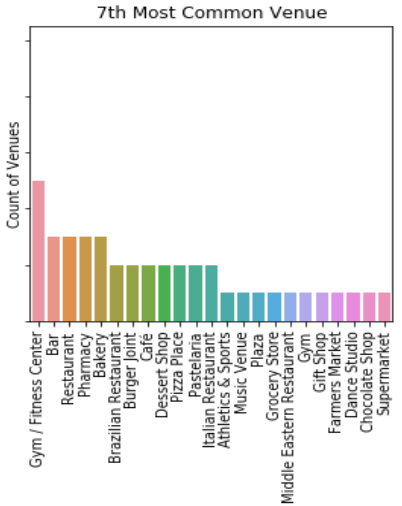


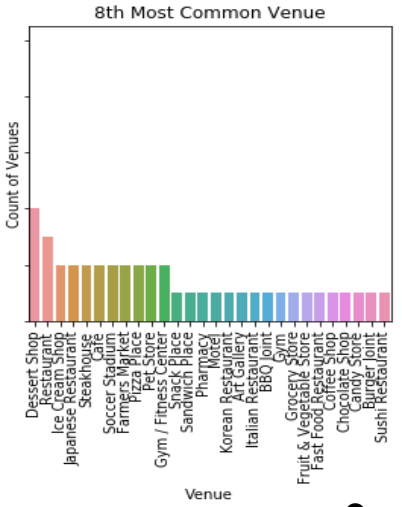


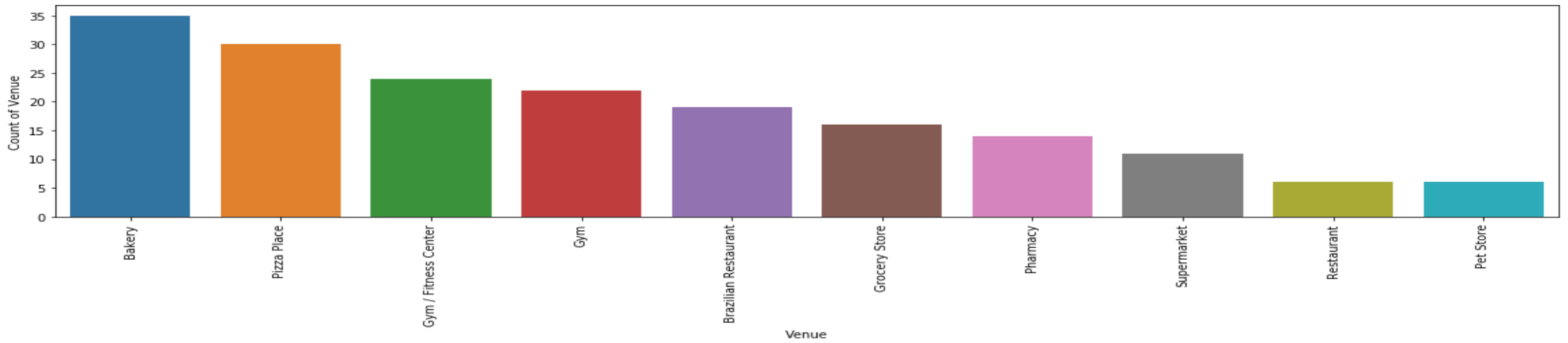




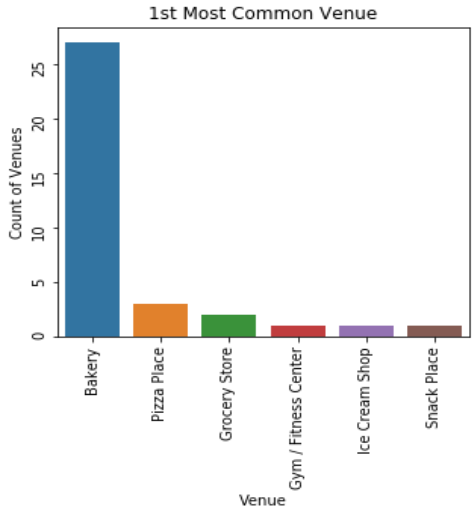


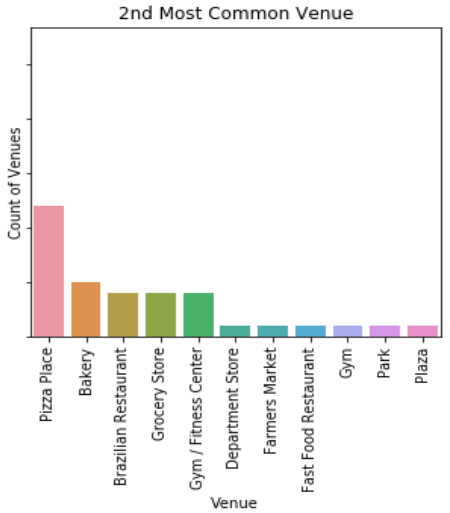


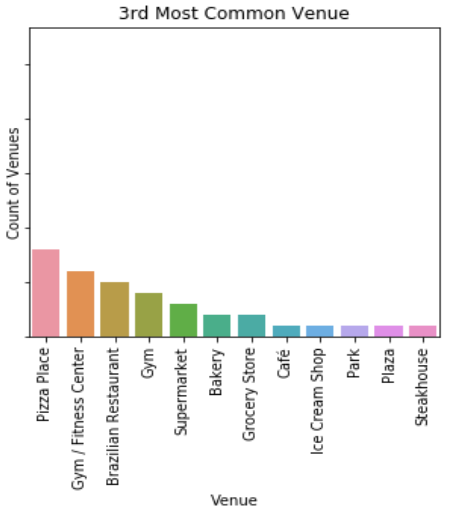


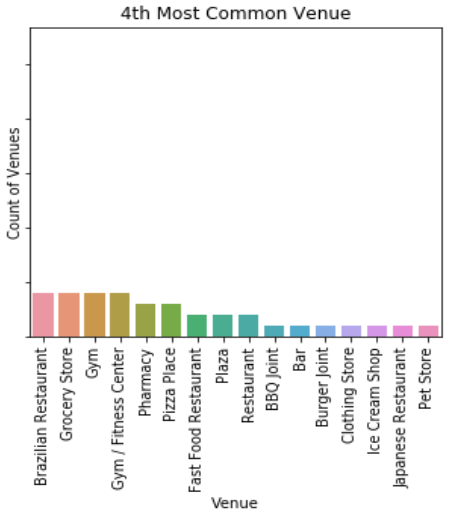


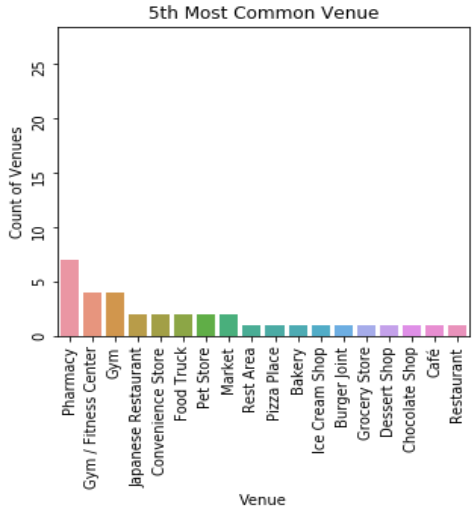
Cluster 2 - Blue

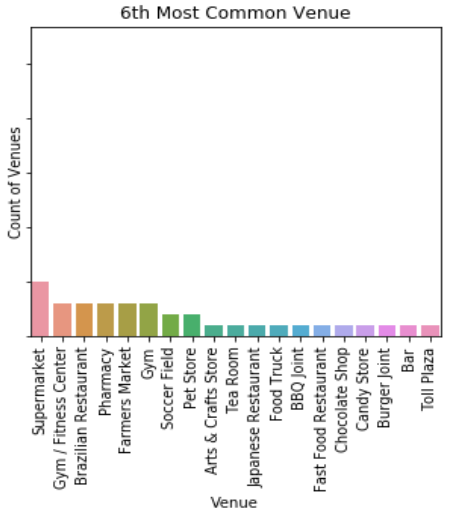


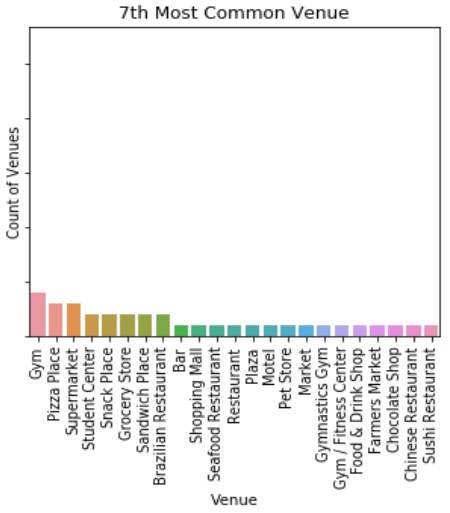


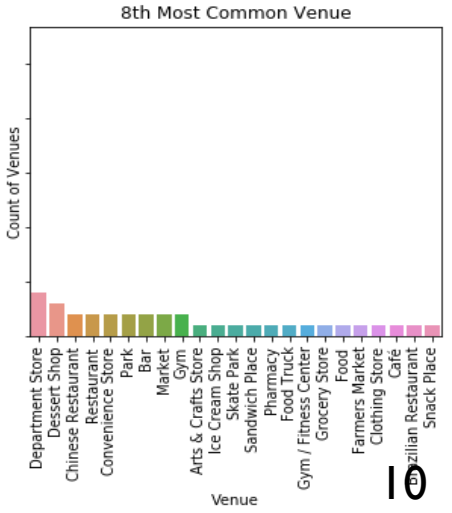










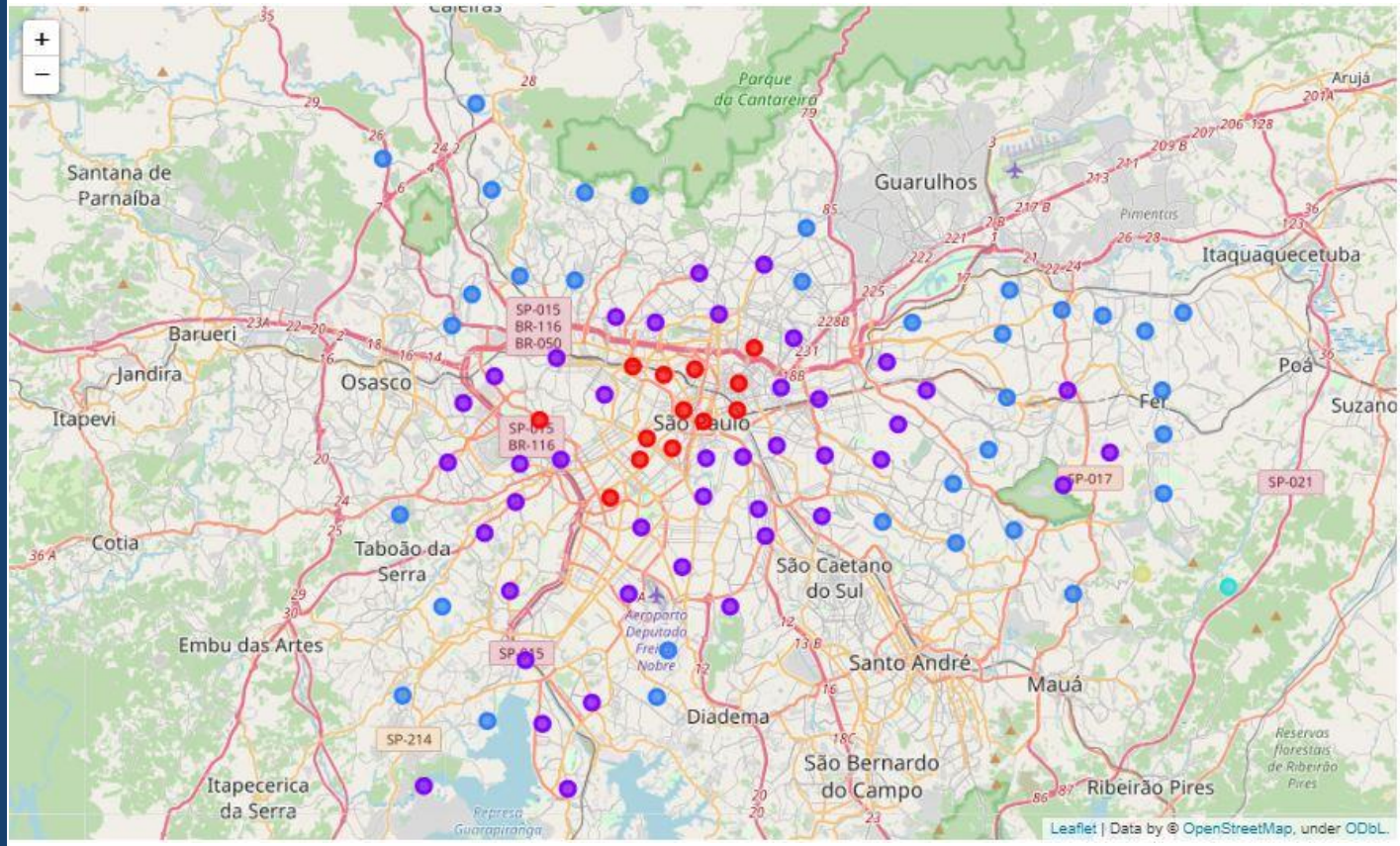


Data Clustering and Analysis – Conclusion I

By analyzing the top 3-5 of each of the three clusters, we can see the relationship between the commercial areas and the residential.

By analyzing the information from the websites about the top neighborhoods to live in São Paulo, we can see how the city is divided by commercial, commercial/residential and residential areas mostly by their venues.

We can visualize that purple and blue regions are pretty similar in terms of venues, but the difference is in the distribution of people and commerce.



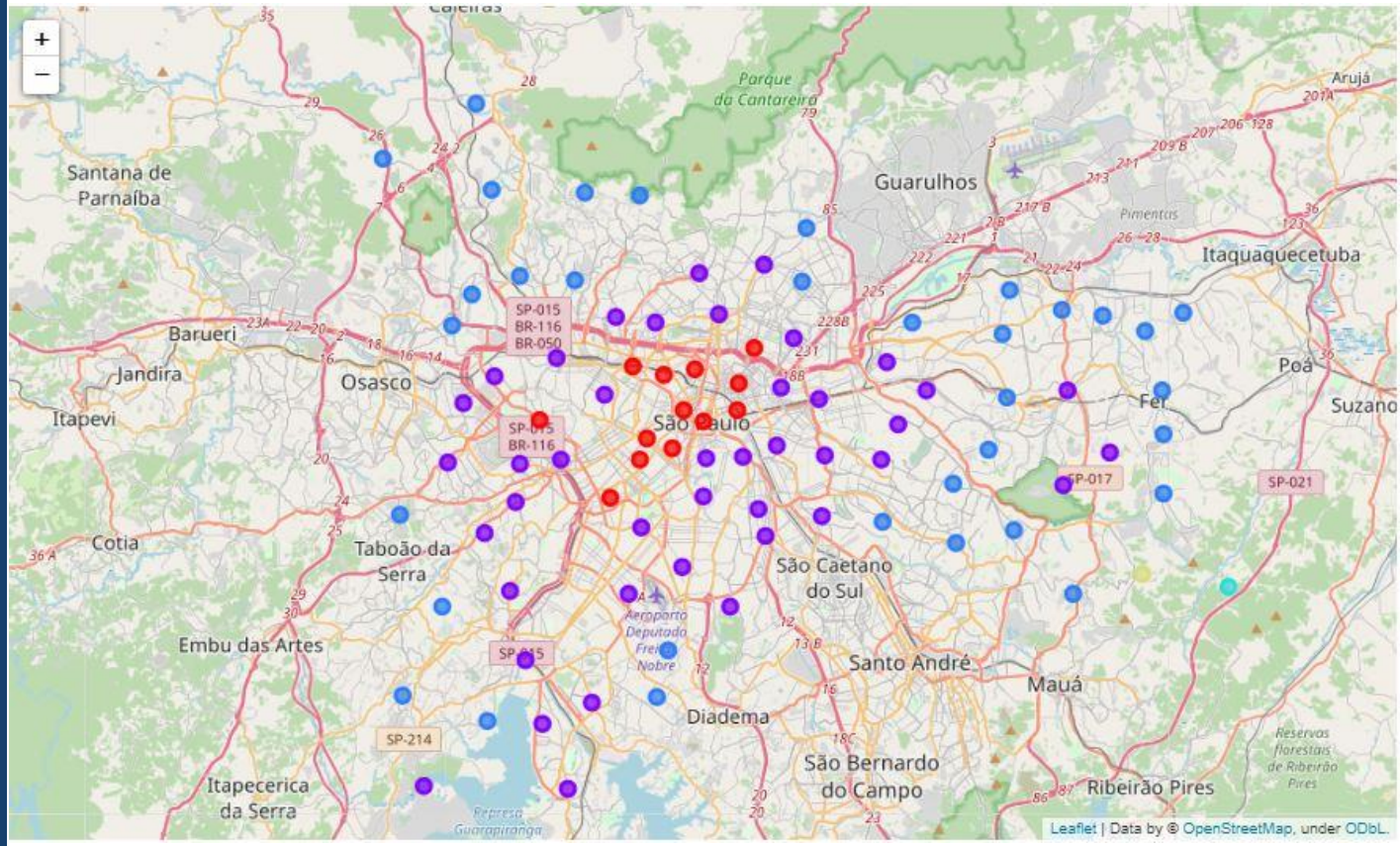
Let's go back to the clustered in which we can see the map. We notice that the inner area (red) is the most commercial area of São Paulo in which has a lot of different restaurants and theaters, made for people that work nearby. For curiosity, this red area is the oldest part of São Paulo.

The purple area is the growing part of São Paulo for high-end buildings for either residential and commercial sectors and we can see that its mainly focused with Pizza places, Bakery, Gyms, Bars and Desert shops, which fits perfectly for a growing, "modern" region.

Now the blue area contains mainly a residential area (suburbs) due to the localization for being far from the center (center = expensive, suburbs = cheaper), and we can see a higher density of Bakery, Pizza places and Gyms, just like the purple area.

Data Clustering and Analysis – Conclusion II

With the previous slide, we saw the relationship between the regions and neighborhoods and now we can conclude which regions are the best for new workplaces and residences based on the venues.



Based on the previous analysis, we can conclude that the purple area is the best fit (for now, it seems) to start a new workplace and resident due to their mix of evolution and venues. But the question is, why the purple?

- The kinds of venues show a more high-end type of venue.
- The new tech industry is located mostly in the purple region.
- Not only the venues are high-end, but the residences and workplaces.

Even though the venues help us see a bit better the difference between regions, there are more relations that can help understanding why the purple area is the best fit for new residence and commerces.

PS: Please check the links.txt for more information, sadly it is in Portuguese, but Google Translator can help

END!

Lucas Hosoya

Github: https://github.com/ShigueruHosoya
Contact: shigaaaa@gmail.com