THE BATTLE OF NEIGHBORHOODS

State of Sao Paulo, Brazil

INTRODUCTION – BUSINESS PROBLEM

- The state of Sao Paulo is considered the most populous of the Country and it is natural to have a high concentration of business in the same region. However, we can notice that often the concentration of companies is in the same activity like markets, hotels and drugstores, which creates opportunities for new businesses not yet explored. So, the challenge is identify these unexplored areas and suggest which type of business the area needs and show the concentration for each one.
- This is the goal of this project: identify new opportunities and suggest to anyone interested.

INTRODUCTION – BUSINESS PROBLEM

- To do it, we will use the service provided for Foursquare to find all the companies and its segment to cross the location with the map of state of Sao Paulo.
  Several techniques will be used along this project to show the analysis, like Folium, K-Means and others.

DATA REQUIREMENTS

- The dataet used was obtained from Anatel in this link: https://www.anatel.gov.br/Portal/verificaDocumentos/documento.asp?numeroPublicacao=285354&assuntoPublicacao=Rela

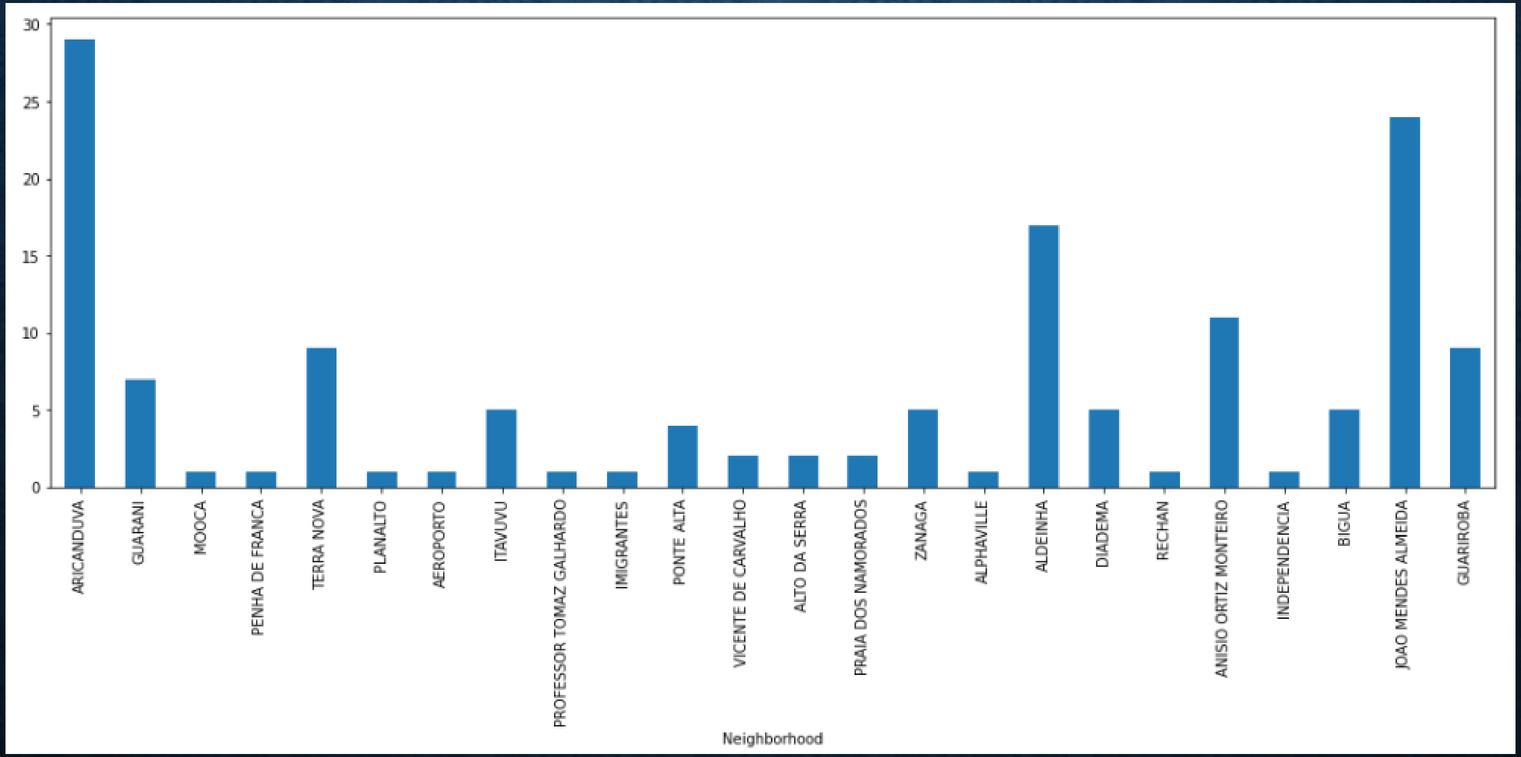
- Note that the file is a PDF and contains several information about the State, like Longitude and Latitude (in degrees). It is important to mention that this dataset was previosly treated on Excel – the coordinates were converted to decimal. Nothing else was needed for the analysis.

METHODOLOGY

- Once we have the neighborhoods data of Sao Paulo and also have the most popular venues in each neighborhood obtained using Foursquare API we can perform one hot encoding on the obtained data set and use it find the 10 most common venue category in each neighborhood.
  Then clustering can be performed on the dataset. Here K - Nearest Neighbor clustering technique have been used. To find the optimal number of clusters silhouette score metric technique is used.
- The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

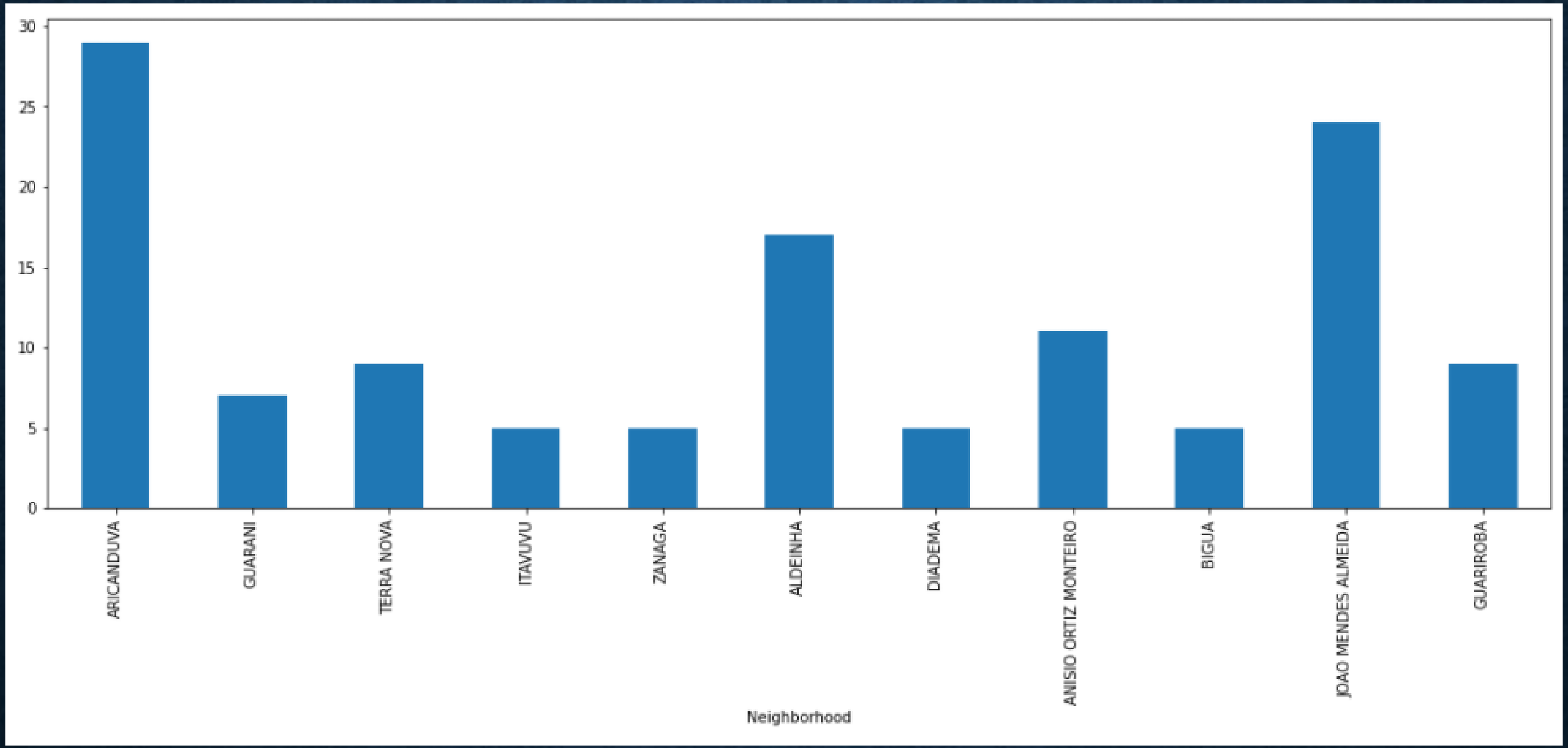
ANALYSIS

- Looking into the dataset we can observe that many neighborhoods has less then 5 venues which we can remove to perform a better analysis and better results.
- The chart bellow shows the concentration of each venues.



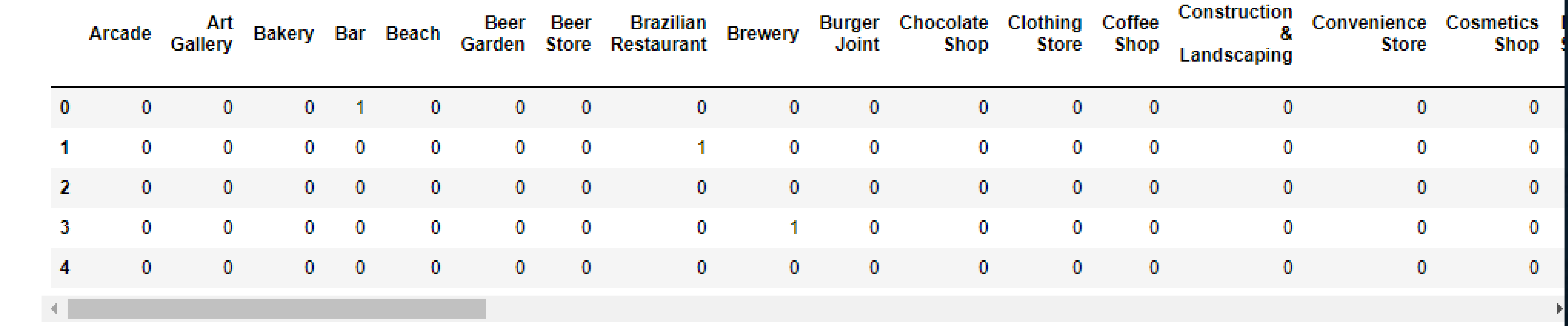
ANALYSIS

- After fill out the venues <= 5 we obtained the following chart.



ANALYSIS

- One hot encoding is performed on the filtered data to obtain the venue categories by transposed into columns.

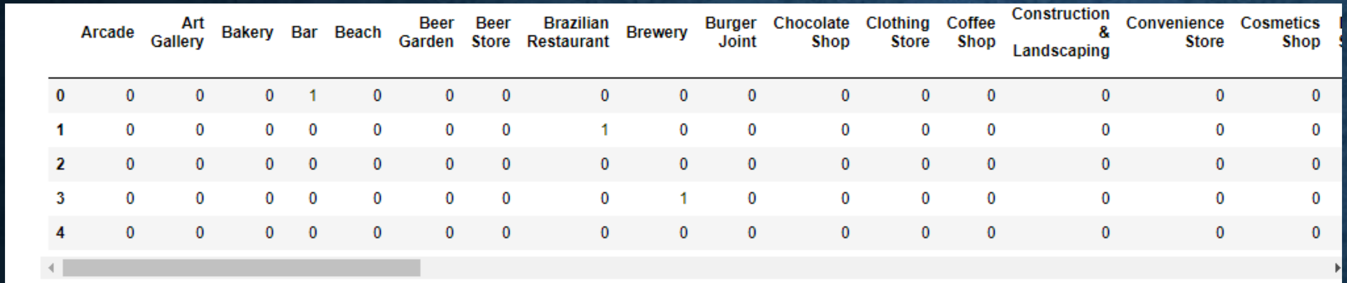


	Arcade	Art Gallery	Bakery	Bar	Beach	Beer Garden	Beer Store	Brazilian Restaurant	Brewery	Burger Joint	Chocolate Shop	Clothing Store	Coffee Shop	Construction & Landscaping	Convenience Store	Cosmetics Shop
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- This is used to obtain the top 10 most common venues in each neighborhood. Then the resultant dataset can be used for the clustering (K-Means).

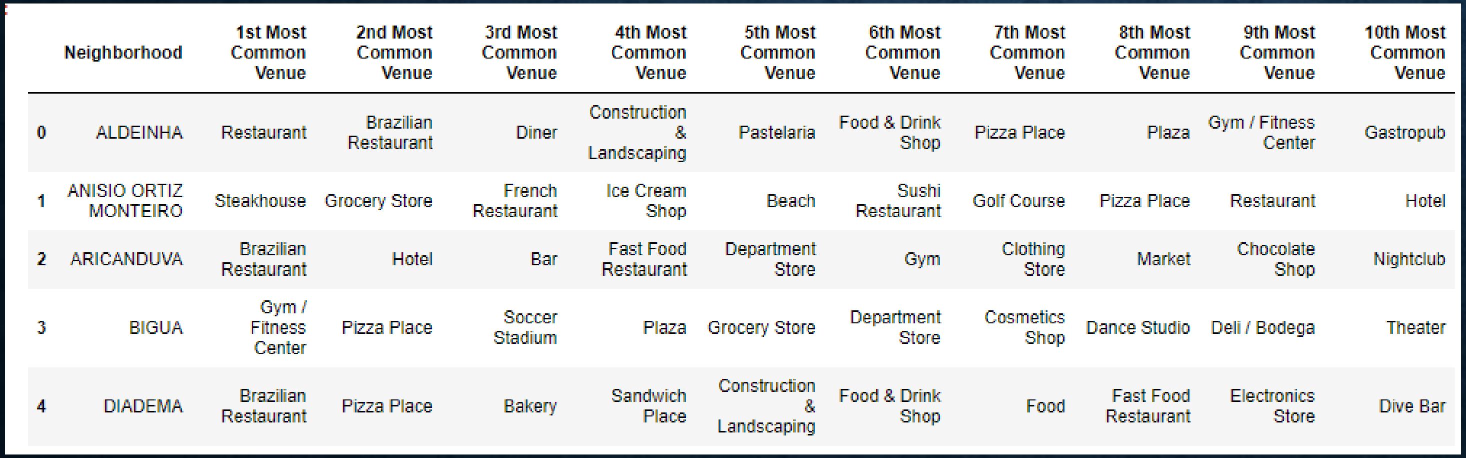
ANALYSIS

- One hot encoding is performed on the filtered data to obtain the venue categories by transposed into columns.



	Arcade	Art Gallery	Bakery	Bar	Beach	Beer Garden	Beer Store	Brazilian Restaurant	Brewery	Burger Joint	Chocolate Shop	Clothing Store	Coffee Shop	Construction & Landscaping	Convenience Store	Cosmetics Shop
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

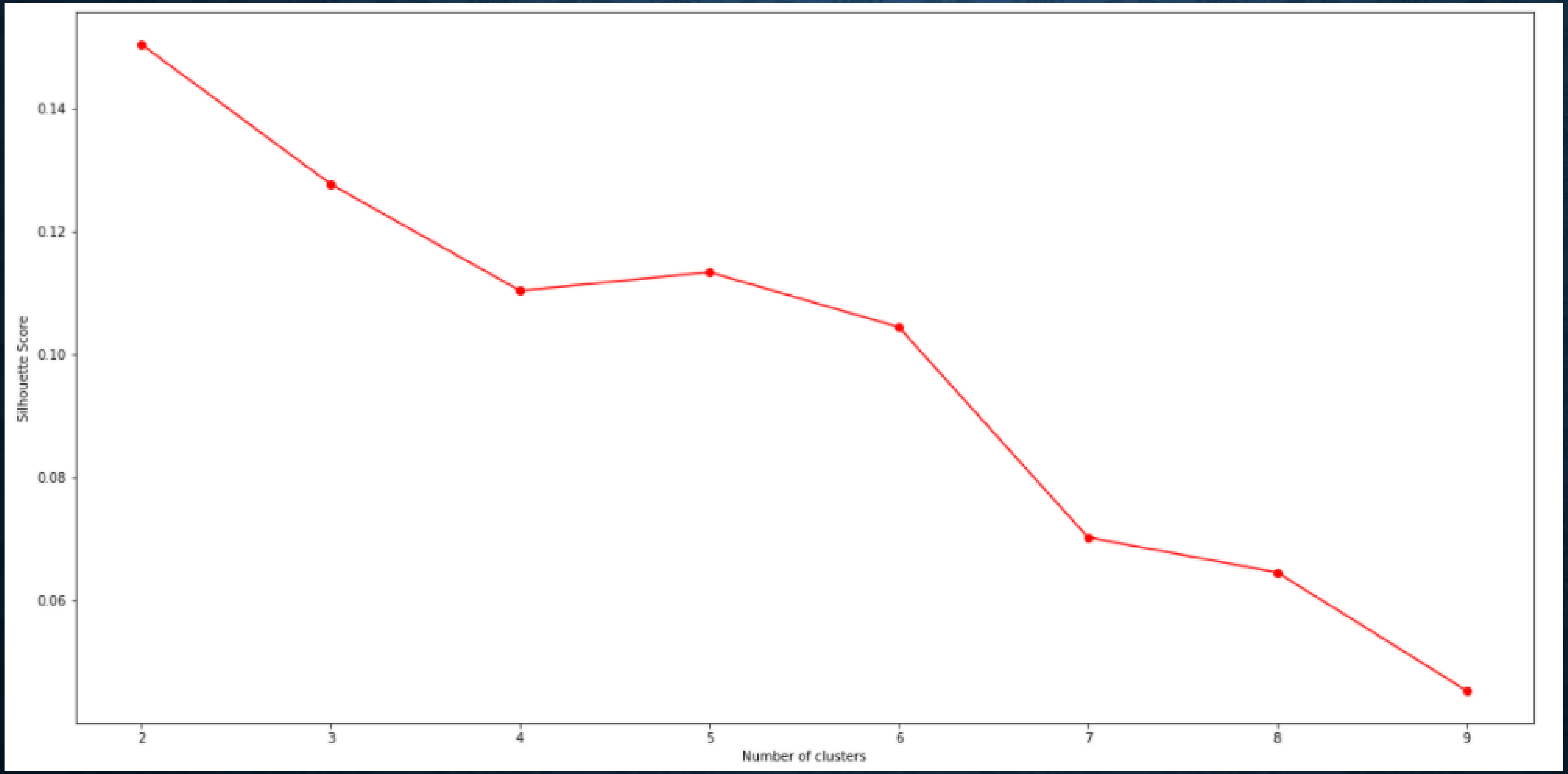
- This is used to obtain the top 10 most common venues in each neighborhood. Then the resultant dataset can be used for the clustering (K-Means).



	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	ALDEINHA	Restaurant	Brazilian Restaurant	Diner	Construction & Landscaping	Pastelaria	Food & Drink Shop	Pizza Place	Plaza	Gym / Fitness Center	Gastropub
1	ANISIO ORTIZ MONTEIRO	Steakhouse	Grocery Store	French Restaurant	Ice Cream Shop	Beach	Sushi Restaurant	Golf Course	Pizza Place	Restaurant	Hotel
2	ARICANDUVA	Brazilian Restaurant	Hotel	Bar	Fast Food Restaurant	Department Store	Gym	Clothing Store	Market	Chocolate Shop	Nightclub
3	BIGUA	Gym / Fitness Center	Pizza Place	Soccer Stadium	Plaza	Grocery Store	Department Store	Cosmetics Shop	Dance Studio	Deli / Bodega	Theater
4	DIADEMA	Brazilian Restaurant	Pizza Place	Bakery	Sandwich Place	Construction & Landscaping	Food & Drink Shop	Food	Fast Food Restaurant	Electronics Store	Dive Bar

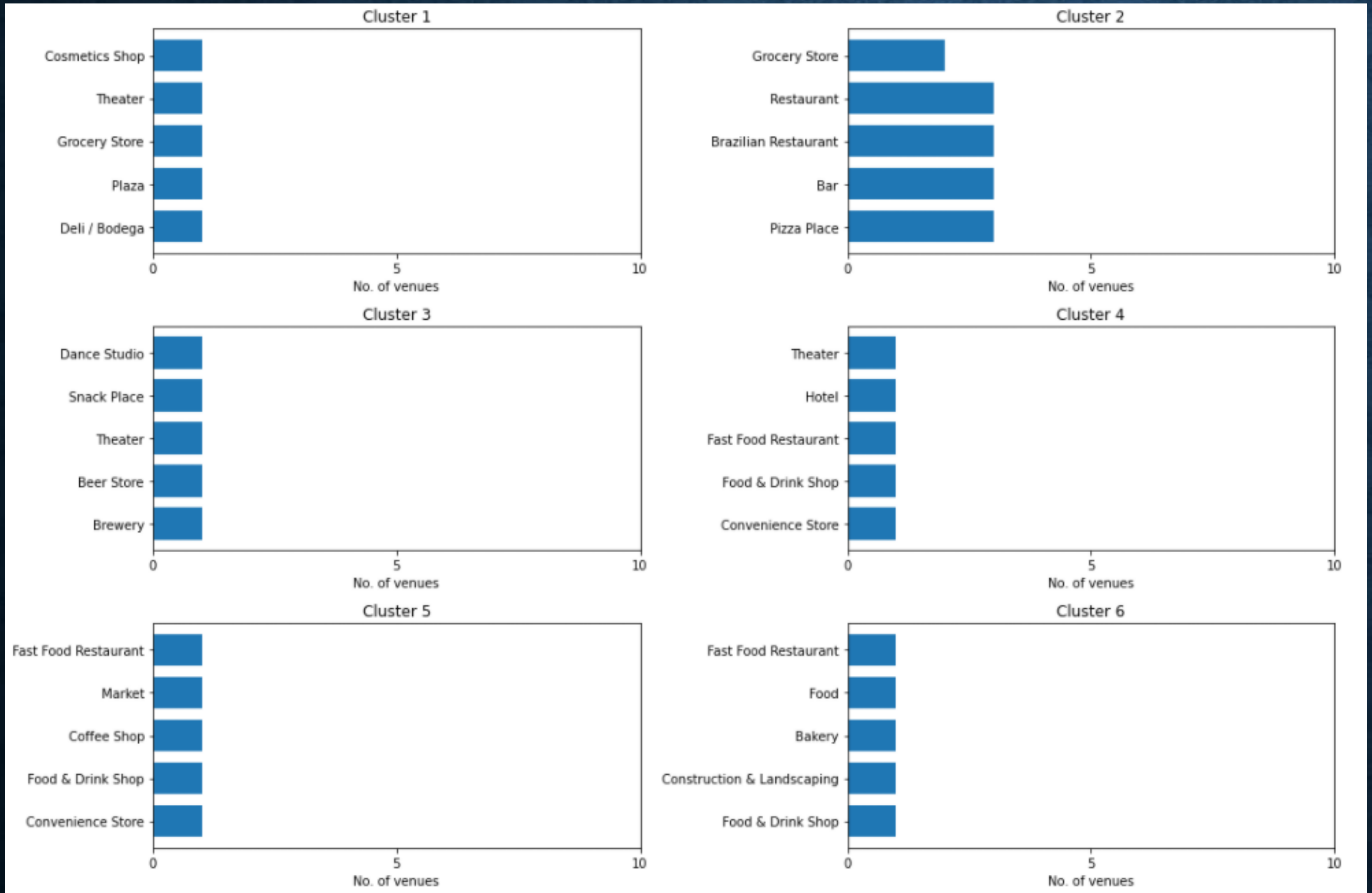
ANALYSIS

- From the plot we choose to use K = 5.



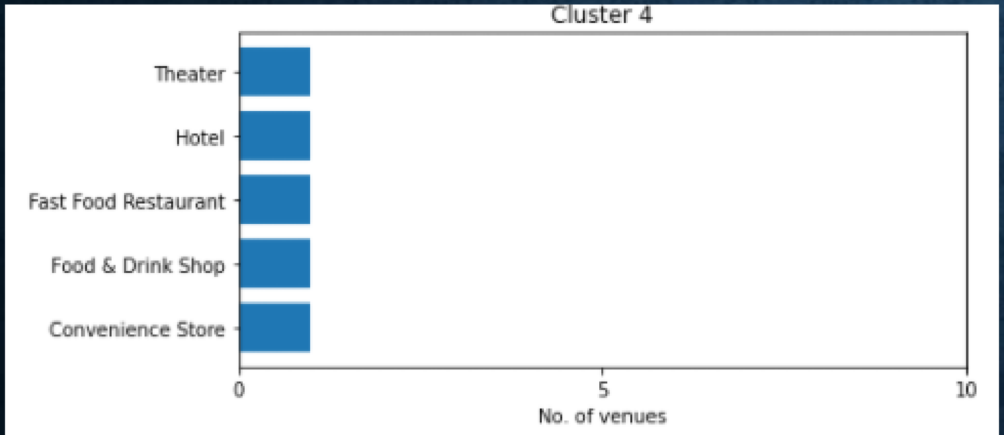
FINAL RESULTS

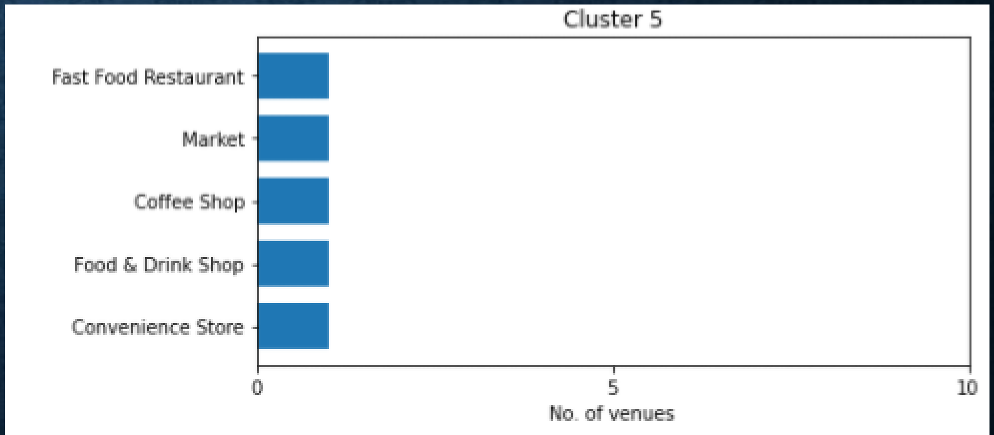
- From the graphs plotted, we are able to concluded several things as the best place do start a new hotel business is on the Cluster 5.



FINAL RESULTS

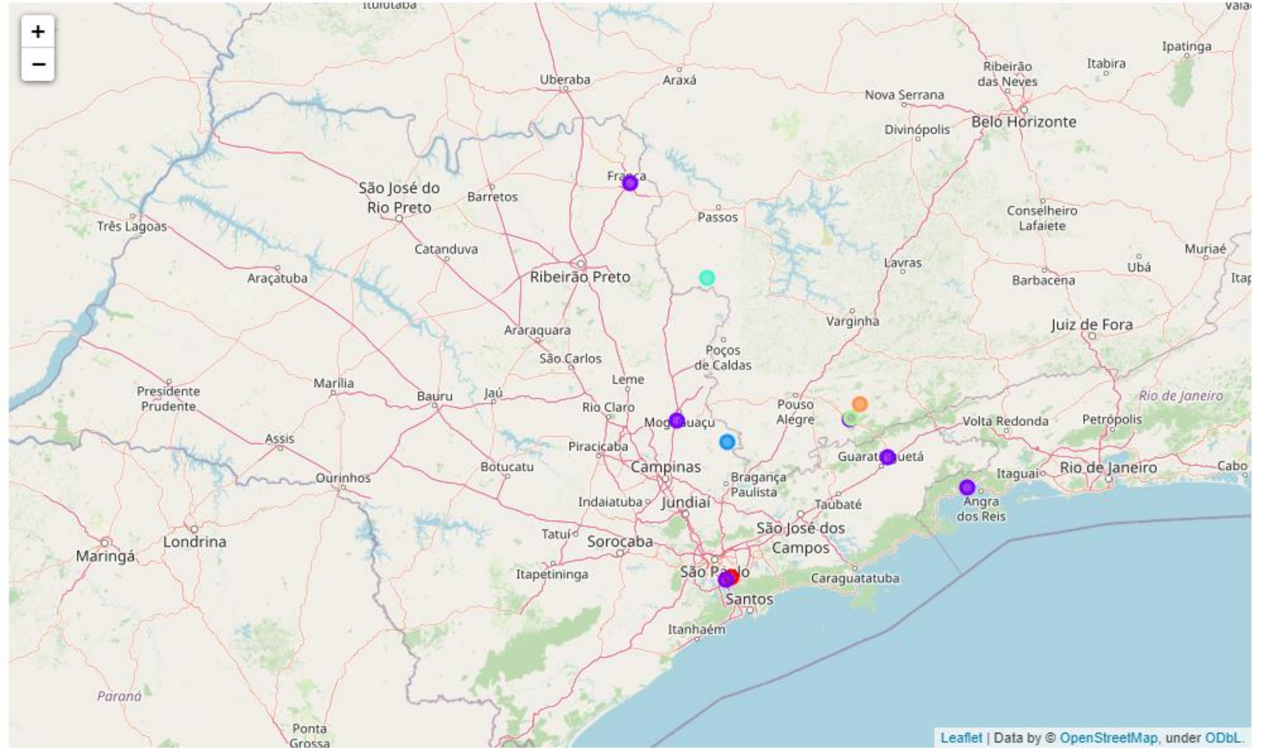
- The reason is simple: As we can see the cluster 4 is where we have more Hotel, so this Cluster is not the best choice to open another one. Otherwise, the others clusters does not have Hotels. It does not mean that all these clusters are good to open one. From my understanding, every Hotel needs to be well located, in other words it means that Hotels needs Restaurants, Coffee Shops and this kind of store around. So, based on that, we can concluded that the Cluster5 its the best option once we have Market, Restaurant, Coffee Shop and others.





FINAL RESULTS

- The following is a map of State of Sao Paulo with the neighborhood clusters superimposed on top of it. This map can also be used to select a vast suggestion area for a particular type of business based on the category.



Leaflet | Data by © OpenStreetMap, under ODbL.

CONCLUSION

- The purpose here in this project (Capstone - The Battle of Neighborhoods) was to analyze the neighborhoods of state of Sao Paulo and create a clustering model to suggest the best location to start a new business. All the data was obtained from an online source (Anatel) and from the Foursquare API that was used to find the major venues in each neighborhood.

  But we found that a considerable numbers of neighborhoods had less than 5 venues returned. So, in order to build a good Data Science model, these locations were all removed. Thus, the remaining locations were used to create the clustering model. The best number of clusters (5) was obtained using the silhouette score. cluster.

  One example was presented (Hotel) and a map showing the clusters have been showed. It is important to mention that both these can be used by stakeholders or investors to decide the location for the new business desired.