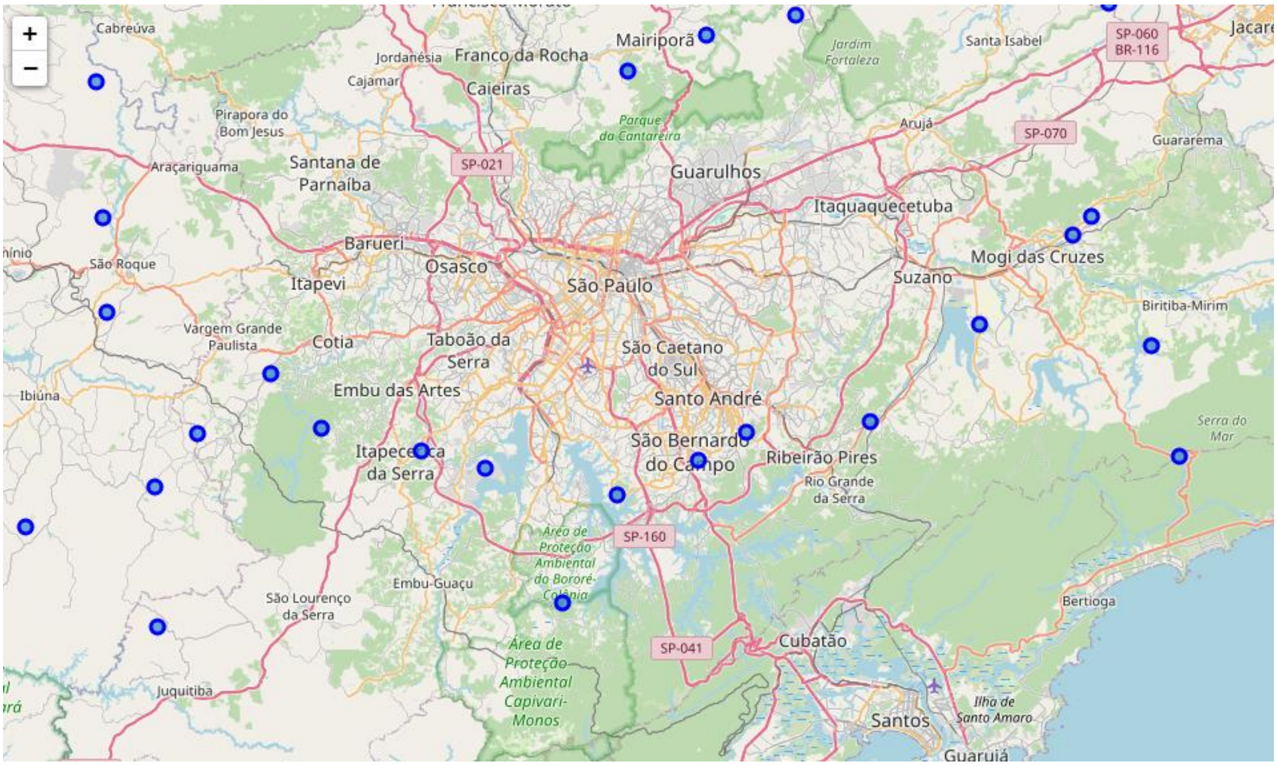
The Battle of Neighborhoods
State of Sao Paulo



Applied Data Science Capstone

William Roesch

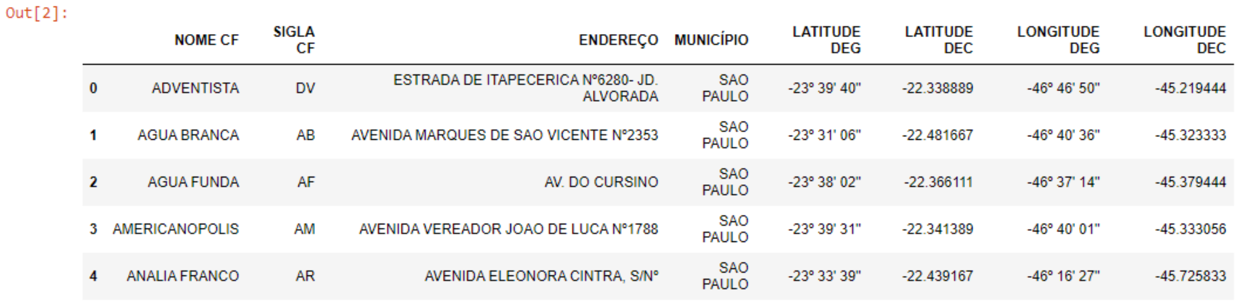
1. INTRODUCTION - BUSINESS PROBLEM

The objective of this project is to find the best location to open a new business-like restaurant, hotels, gym and so in Sao Paulo, Brazil. This report can be used by investors.

The **Foursquare API** is used to access the venues in the neighborhoods. Since, it returns less venues in the neighborhoods, we would be analyzing areas for which countable number of venues are obtained. Then they are clustered based on their venues using **Data Science Techniques**. Here the **k-means clustering algorithm** is used to achieve the task. The optimal number of clusters can be obtained using silhouette score. **Folium visualization library** can be used to visualize the clusters superimposed on the map of Chennai city. These clusters can be analyzed to help small scale business owners select a suitable location for their need such as Hotels, Shopping Malls, Restaurants or even specifically Indian restaurants or Coffee shops.

2. DATA REQUIREMENTS

The state of Sao Paulo has several cities connected as one big city. Considering that, we will use a pdf file provided by Anatel to perform the initial steps – the link can be accessed here.

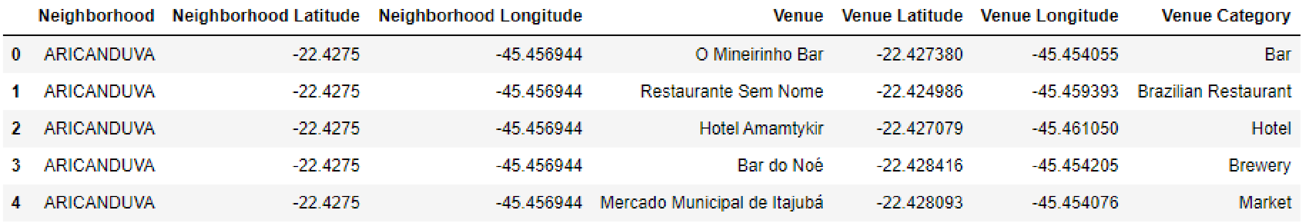


Out[2]:

	NOME CF	SIGLA CF	ENDEREÇO	MUNICÍPIO	LATITUDE DEG	LATITUDE DEC	LONGITUDE DEG	LONGITUDE DEC
0	ADVENTISTA	DV	ESTRADA DE ITAPECERICA Nº6280- JD. ALVORADA	SAO PAULO	-23° 39' 40"	-22.338889	-46° 46' 50"	-45.219444
1	AGUA BRANCA	AB	AVENIDA MARQUES DE SAO VICENTE Nº2353	SAO PAULO	-23° 31' 06"	-22.481667	-46° 40' 36"	-45.323333
2	AGUA FUNDA	AF	AV. DO CURSINO	SAO PAULO	-23° 38' 02"	-22.366111	-46° 37' 14"	-45.379444
3	AMERICANOPOLIS	AM	AVENIDA VEREADOR JOAO DE LUCA Nº1788	SAO PAULO	-23° 39' 31"	-22.341389	-46° 40' 01"	-45.333056
4	ANALIA FRANCO	AR	AVENIDA ELEONORA CINTRA, S/Nº	SAO PAULO	-23° 33' 39"	-22.439167	-46° 16' 27"	-45.725833

Note that there are 2 columns for Latitude and 2 columns for Longitude. The columns with prefix DEG are the originals and the prefix DEC are the columns converted from degrees to decimal – it is needed to perform our analysis with Foursquare.

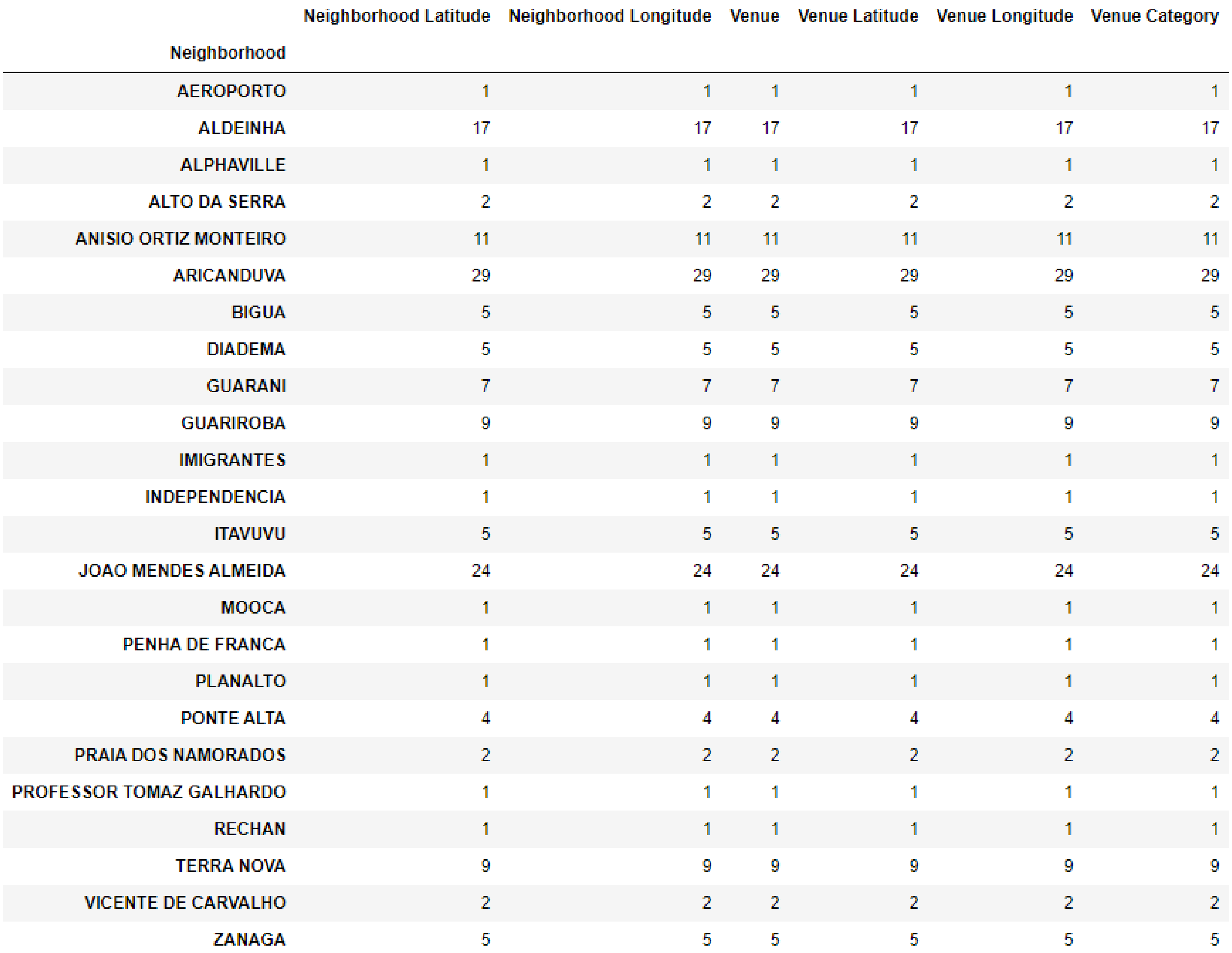
Once we have all the coordinates what we need to do is get all the venues from Foursquare for each NOME CF.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	ARICANDUVA	-22.4275	-45.456944	O Mineirinho Bar	-22.427380	-45.454055	Bar
1	ARICANDUVA	-22.4275	-45.456944	Restaurante Sem Nome	-22.424986	-45.459393	Brazilian Restaurant
2	ARICANDUVA	-22.4275	-45.456944	Hotel Amamtykir	-22.427079	-45.461050	Hotel
3	ARICANDUVA	-22.4275	-45.456944	Bar do Noé	-22.428416	-45.454205	Brewery
4	ARICANDUVA	-22.4275	-45.456944	Mercado Municipal de Itajubá	-22.428093	-45.454076	Market

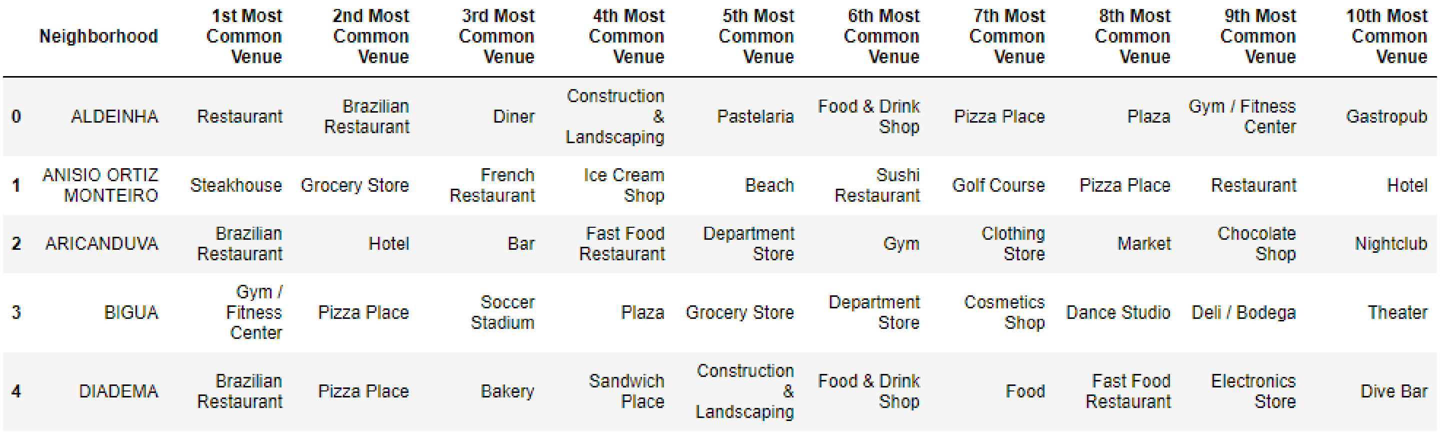
3. METHODOLOGY AND ANALYSIS

Now, we have the neighborhoods data of Sao Paulo and also have the most popular venues in each neighborhood obtained using Foursquare API.



	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
AEROPORTO	1	1	1	1	1	1
ALDEINHA	17	17	17	17	17	17
ALPHAVILLE	1	1	1	1	1	1
ALTO DA SERRA	2	2	2	2	2	2
ANISIO ORTIZ MONTEIRO	11	11	11	11	11	11
ARICANDUVA	29	29	29	29	29	29
BIGUA	5	5	5	5	5	5
DIADEMA	5	5	5	5	5	5
GUARANI	7	7	7	7	7	7
GUARIROBA	9	9	9	9	9	9
IMIGRANTES	1	1	1	1	1	1
INDEPENDENCIA	1	1	1	1	1	1
ITAVUVU	5	5	5	5	5	5
JOAO MENDES ALMEIDA	24	24	24	24	24	24
MOOCA	1	1	1	1	1	1
PENHA DE FRANCA	1	1	1	1	1	1
PLANALTO	1	1	1	1	1	1
PONTE ALTA	4	4	4	4	4	4
PRAIA DOS NAMORADOS	2	2	2	2	2	2
PROFESSOR TOMAZ GALHARDO	1	1	1	1	1	1
RECHAN	1	1	1	1	1	1
TERRA NOVA	9	9	9	9	9	9
VICENTE DE CARVALHO	2	2	2	2	2	2
ZANAGA	5	5	5	5	5	5

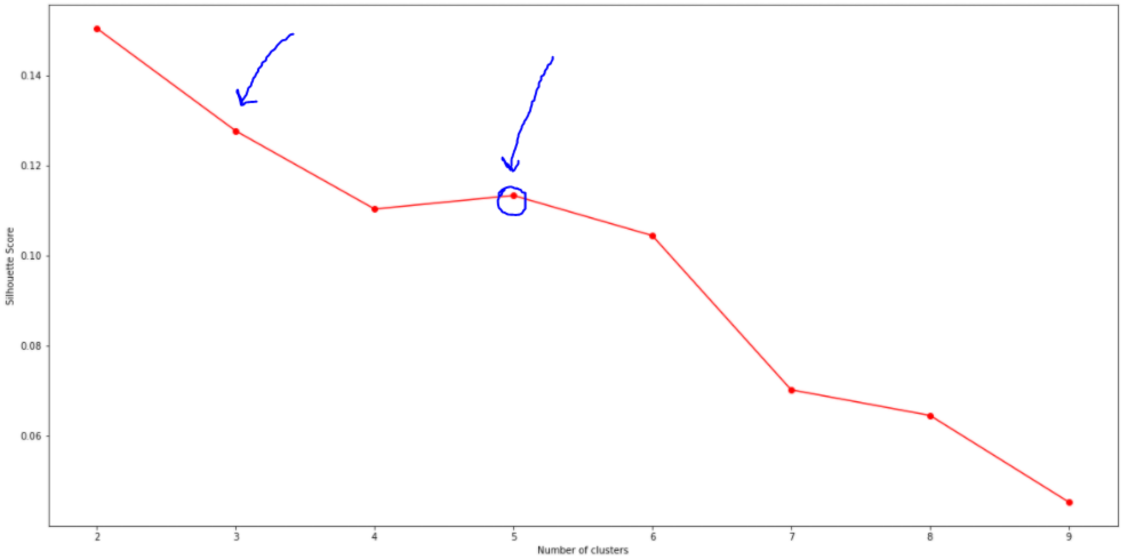
We can perform one hot encoding on the obtained data set and use it find the 10 most common venue category in each neighborhood.



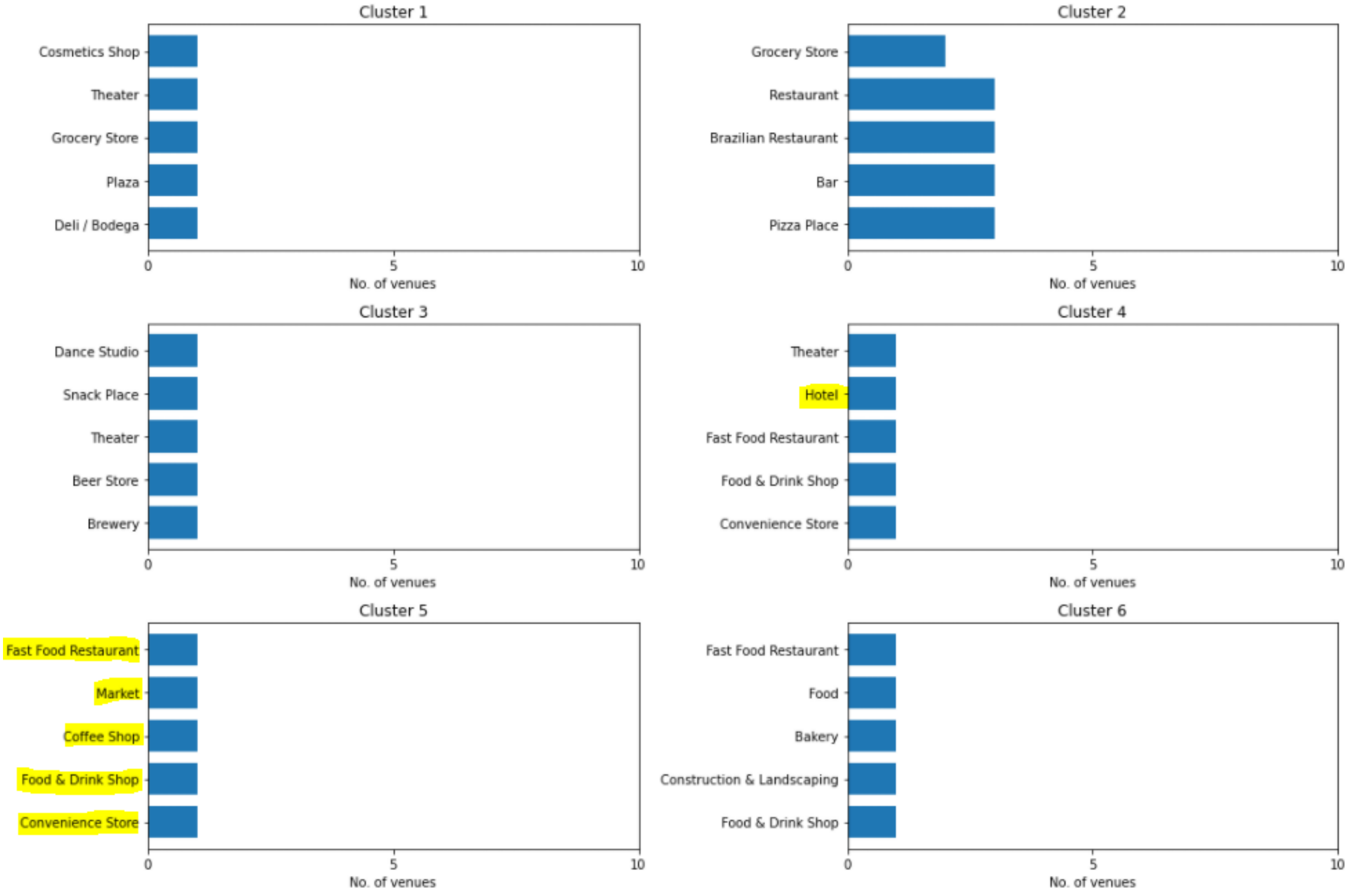
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	ALDEINHA	Restaurant	Brazilian Restaurant	Diner	Construction & Landscaping	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

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.



All plots presented above can be used to suggest valuable information to Business persons. Let's discuss a few examples.

Let’s suppose that the intention is open a new Hotel. 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 do 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 conclude that the Cluster5 it’s the best option once we have Market, Restaurant, Coffee Shop and others.

4. 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 number 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.