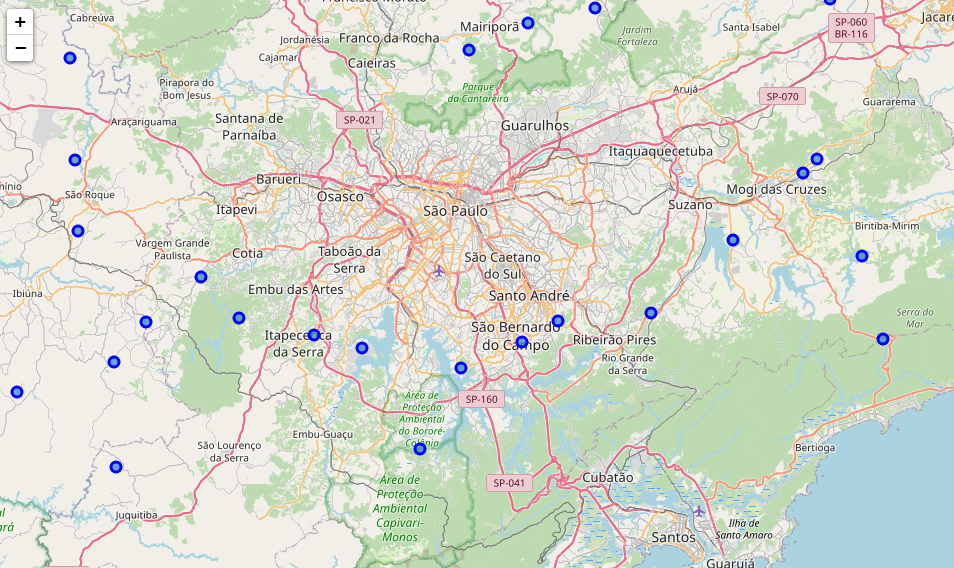
**The Battle of Neighborhoods**

**State of Sao Paulo**

Applied Data Science Capstone

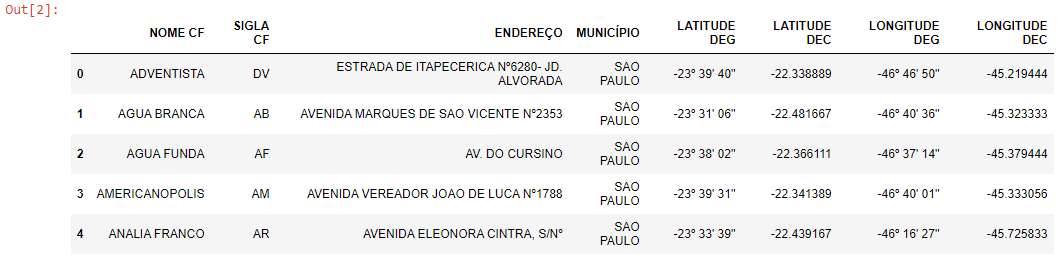
William Roesch

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.

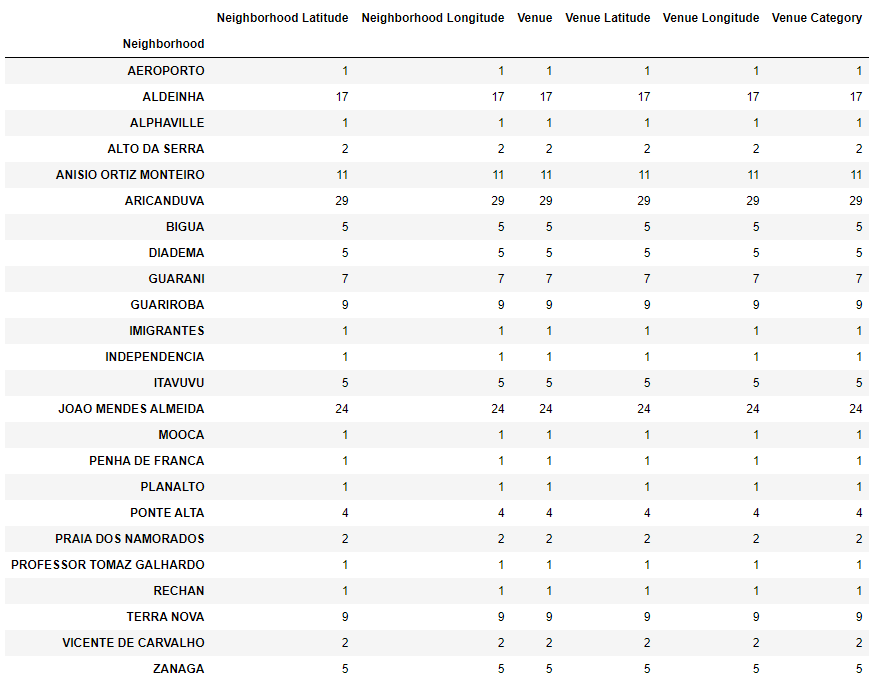
1. **DATA REQUIREMENTS**

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](https://www.anatel.gov.br/Portal/verificaDocumentos/documento.asp?numeroPublicacao=285354&assuntoPublicacao=Rela).

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.

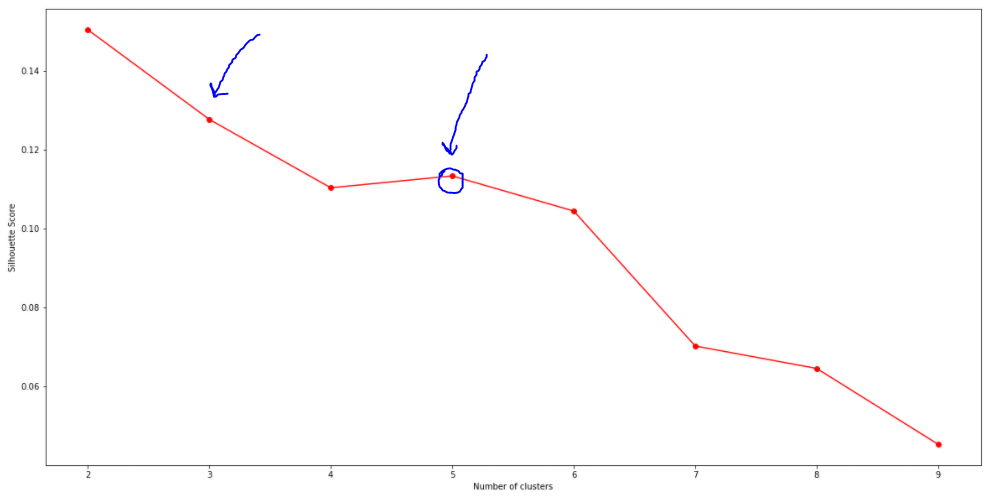
1. **METHODOLOGY AND ANALYSIS**

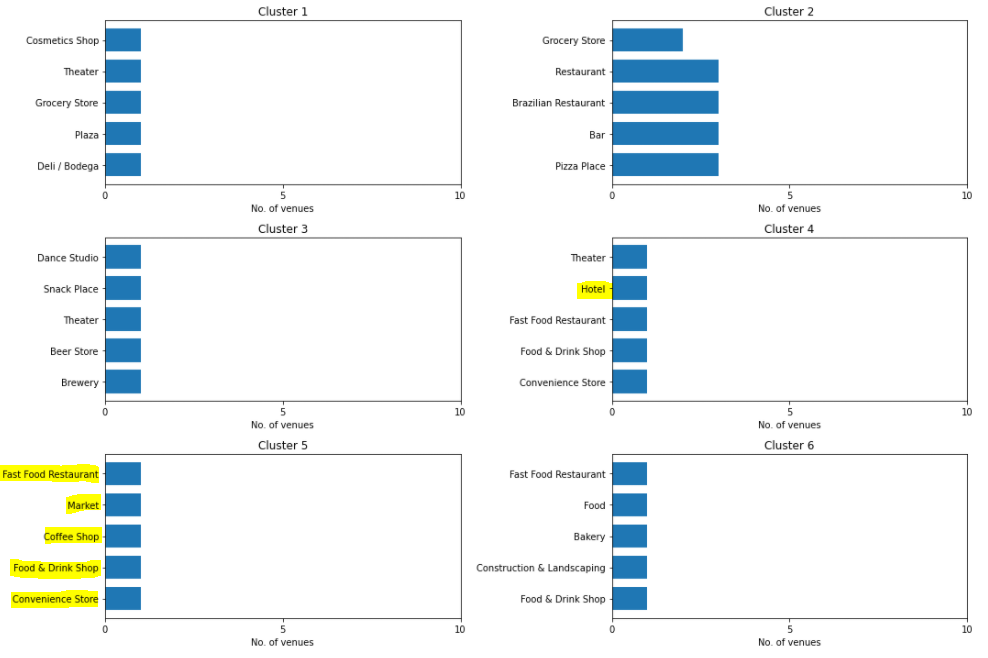
Now, we have the neighborhoods data of Sao Paulo and also have the most popular venues in each neighborhood obtained using Foursquare API.

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.

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.

1. **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.