THE BATTLE OF NEIGHBORHOODS

APPLIED DATA SCIENCE CAPSTONE PROJECT

(Naidugari Manjunath)

Business Problem Section

Background

Many people want to invest their money in the business to make profits. There are different types of business like Restaurants, Shopping Malls, and Departmental Stores etc. It is better to start their business in major cities in the country. In this project, I'd like to study the neighborhoods in Chennai, Tamil Nadu, India.

Business Problem

While starting a business, we need to search for a neighborhood that is suitable to their preferences. In this project, we study neighborhoods and venues to know which locality is good to set up a business.

The major Target Audience would be small-scale business owners and stakeholders planning to start their business at a location in Chennai. This project would help them find the optimal location based on the category of their business such as,

- What is the best location to start a Restaurant in Chennai?
- Which area is best suitable for opening a Shopping Mall and a Departmental Store in Chennai?

Data Requirements

Chennai has multiple neighborhoods. The chennaiiq.com website has a dataset which has the list of neighborhoods in Chennai along with their Latitude and Longitude. Foursquare API is used to obtain the venue details in each neighborhood.

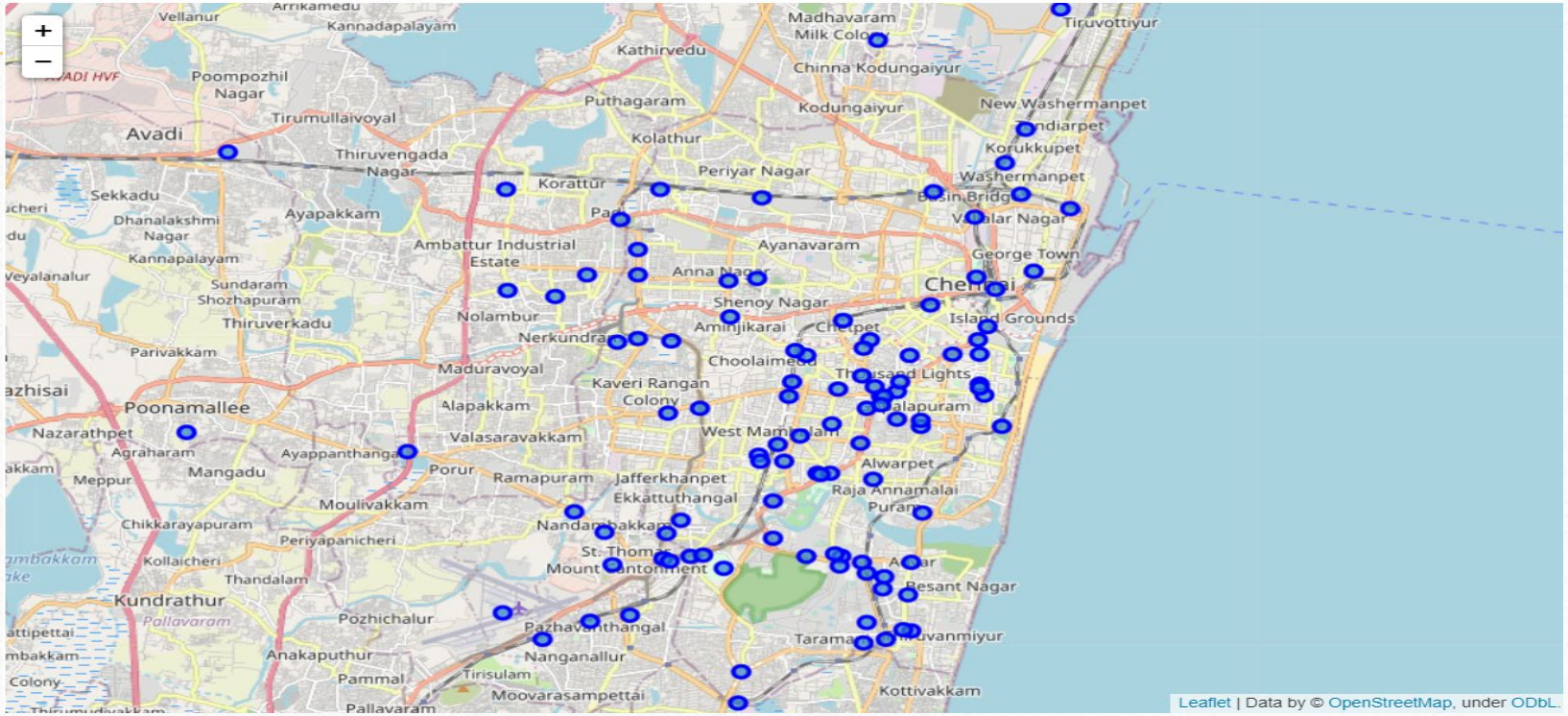
- https://chennaiiq.com/chennai/latitude_longitude_areas.asp
- https://www.foursquare.com

The Foursquare API is used to access the venues in the neighborhoods. Since it returns fewer venues in the neighborhoods, we would be analyzing areas for which a 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 metrics. 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, Departmental Stores and Coffee shops.

Data Requirements

There is a total of 105 neighborhoods as shown in figure. But the Latitude and Longitude data obtained are in Degrees Minute Seconds format which needs to be converted to Decimal Degrees Format.. The following data are obtained from the Foursquare API,

- Venue
- Venue Latitude
- Venue Longitude
- Venue Category data



A total of 1009 venues data have been obtained from Foursquare.

Methodology

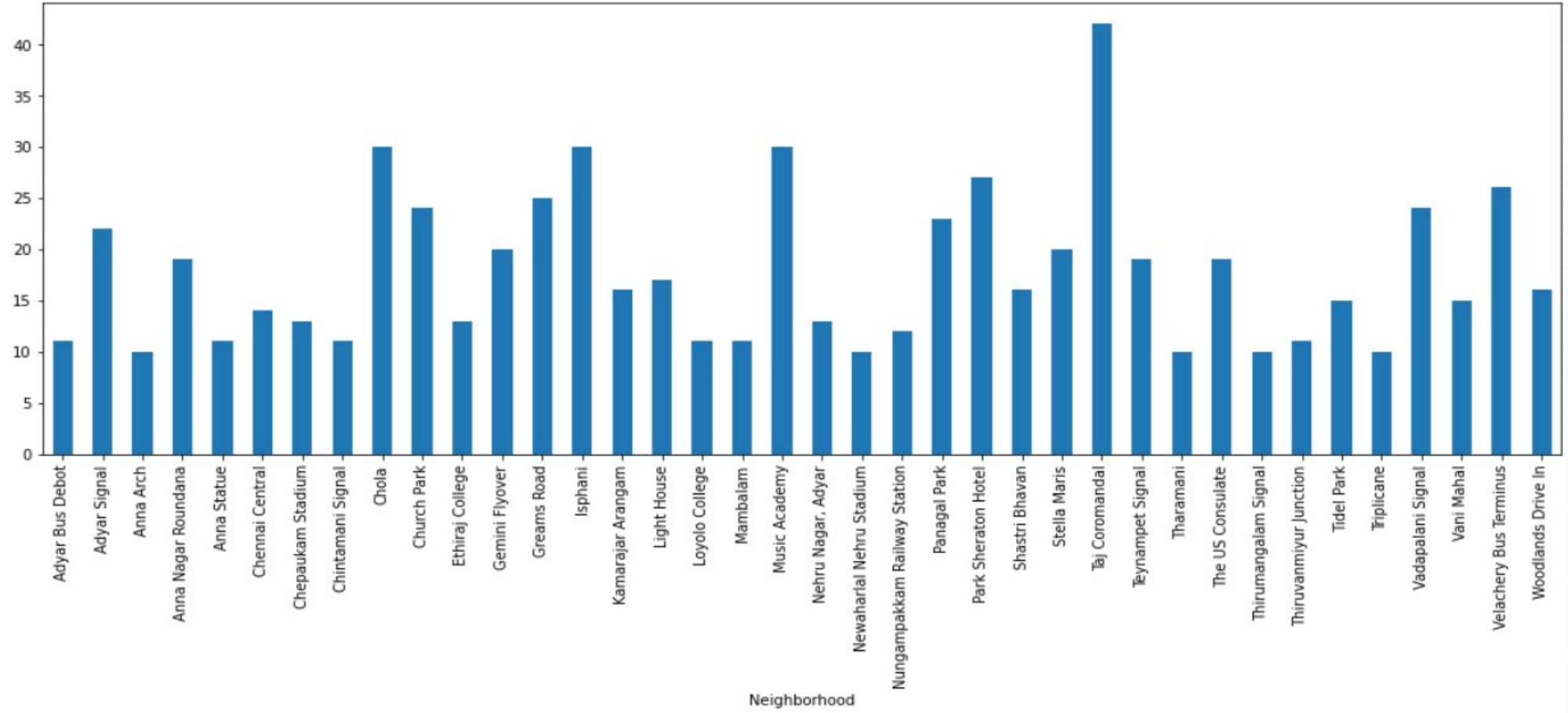
Now, we have the neighborhoods data of Chennai **(105 neighborhoods)**. We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of **1009 venues** have been obtained in the whole city and **136 unique categories**. But as seen we have multiple neighborhoods with less than 10 venues returned. In order to create a good analysis let's consider only the **neighborhoods with more than 10 venues**.

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.

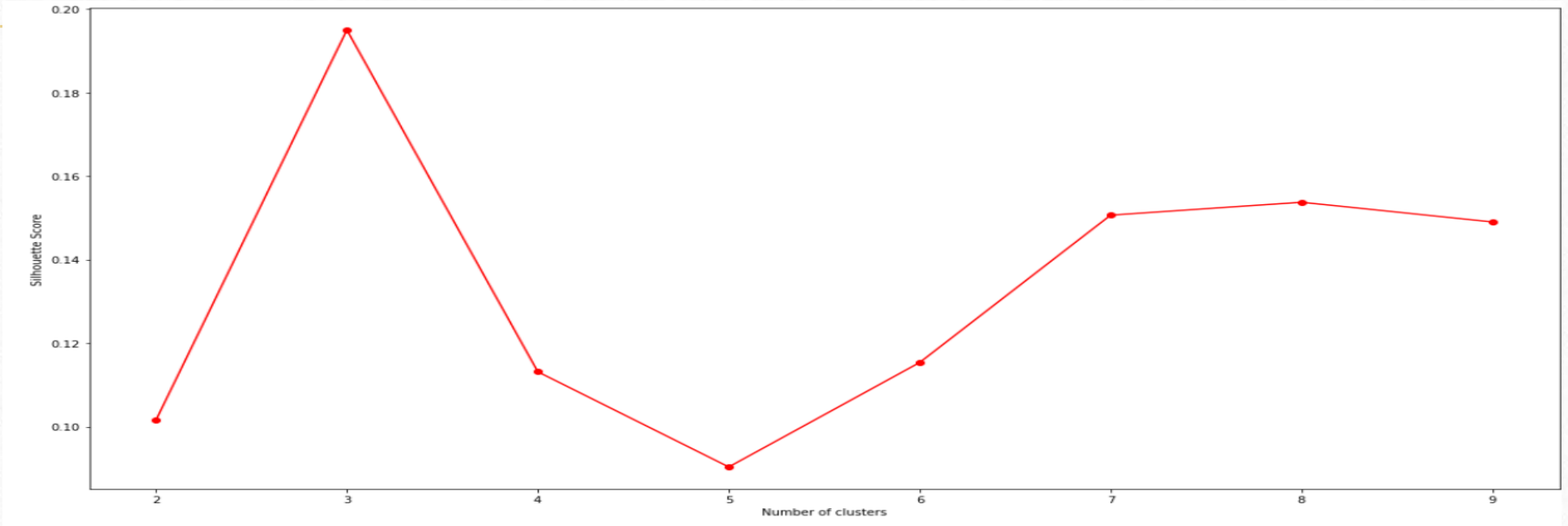
Analysis

- From the dataset we get to know that there are many neighborhoods with less than 10 venues which can be remove before performing the analysis to obtain better results. The following plot shows only the neighborhoods from which 10 or more than 10 venues were obtained.



Now we will **One hot encoding** to categorize the venues in each neighborhood. Then we group the data by neighborhood and take the mean value of frequency of occurrence of each category. This is used to obtain the top 10 most venues which have high frequency of occurrence.

The resultant dataset after encoding is used for clustering. Here **K-Nearest Neighbor(KNN)** clustering algorithm is used. It is an unsupervised Machine Learning that clusters the given data into n-different clusters.



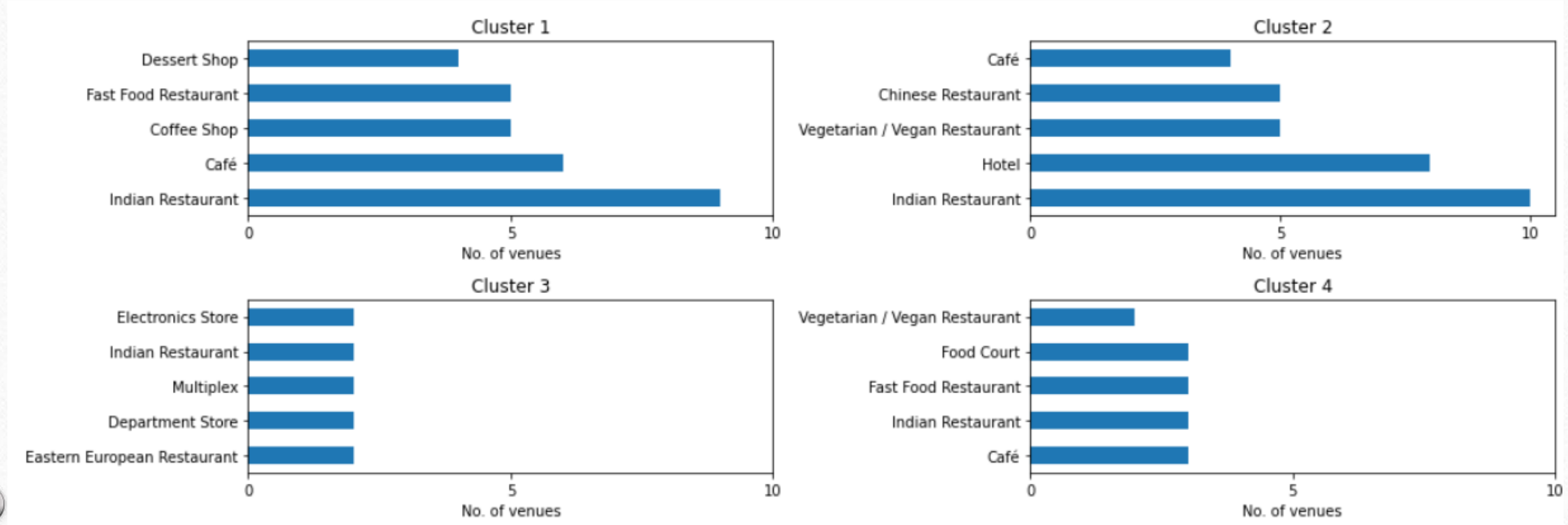
For optimal result we need to select the best value of k. Here, we can use **silhouette score** to find the best value of k.

- A range of clusters 2 to 10 considered in KNN line plot. From the plot we can see that a k values of 2 provides best score . Now, this k value is used for K-Means Clustering Technique and we get data like in the figure.

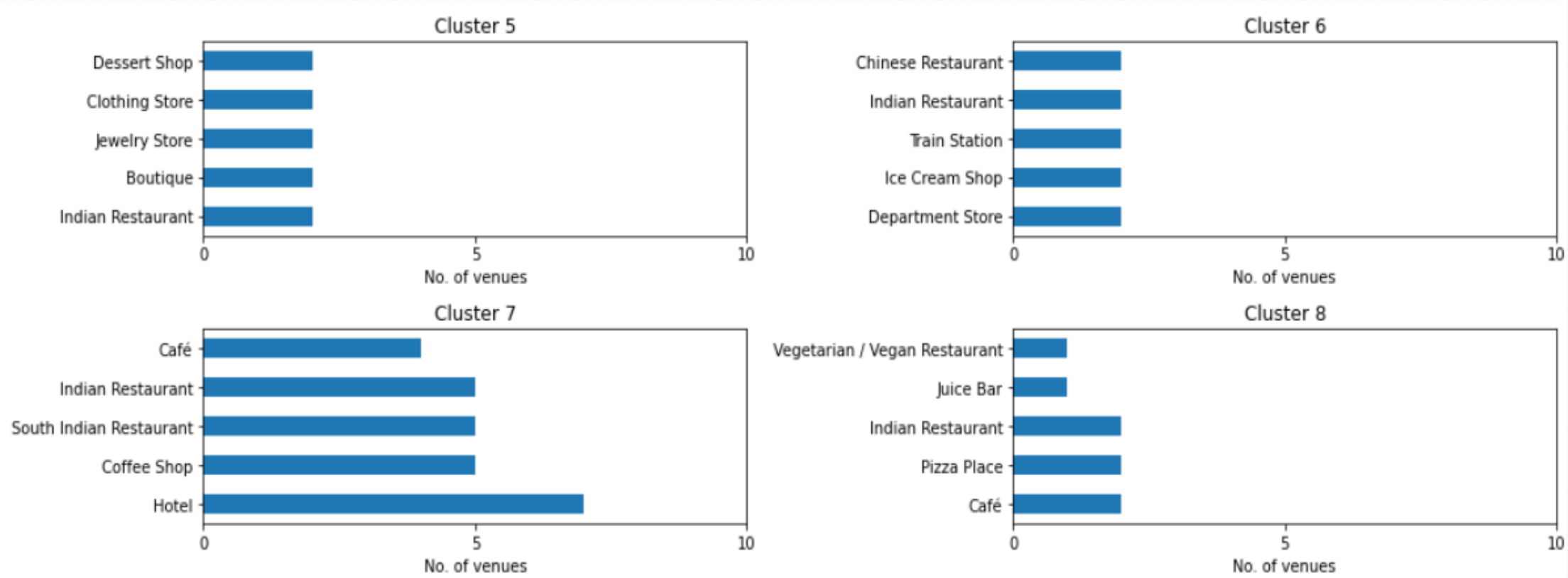
Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
Adyar Bus Debot	12.997222	80.256944	0	Asian Restaurant	Indian Restaurant	Pizza Place	Sandwich Place	BBQ Joint	Fast Food Restaurant	Middle Eastern Restaurant	Café	Breakfast Spot
Adyar Signal	13.006389	80.257500	0	Indian Restaurant	North Indian Restaurant	Bakery	Coffee Shop	Snack Place	Fast Food Restaurant	Electronics Store	Dessert Shop	Rock Club
Anna Arch	13.074444	80.218333	3	Clothing Store	Fast Food Restaurant	Bookstore	Food Court	Electronics Store	Café	Multiplex	Pizza Place	Gym / Fitness Center
Anna Nagar Roundana	13.084444	80.218056	0	Indian Restaurant	Clothing Store	Bookstore	Hotel Bar	Coffee Shop	Electronics Store	Fast Food Restaurant	Café	Middle Eastern Restaurant
Anna Statue	13.068056	80.271944	2	Indian Restaurant	Multiplex	Movie Theater	Dessert Shop	Flea Market	General Entertainment	Department Store	Donut Shop	Eastern European Restaurant

Results and Discussion

- Lets visualize the top 5 most common venue categories in each of the cluster.



Clusters 5 to 8:



- This plot can be used to suggest valuable information to Business persons. Let's discuss a few examples considering they would like to start the following category of Business.

1. Hotel

The neighborhoods in cluster 2 has the greatest number of hotels, hence opening one here is not the best choice. So, is it best to open one at the neighborhoods in cluster 6 or 8? Not likely, since the place has a smaller number of food restaurants. Thus, an optimal place would be one which has less hotels, but also have restaurants and other places to explore. Considering all these facts, the best choice would be Cluster 1 and Cluster 7. such as the Adyar Bus Depot, Gemini Flyover neighborhoods.

---

2. Shopping Mall

By using the same procedure as above, the suitable clusters would be the Cluster 1, Cluster 2 and Cluster 7, since it does not have shopping malls in any of the clusters and also it has many Hotels and Restaurants which gives an advantage.

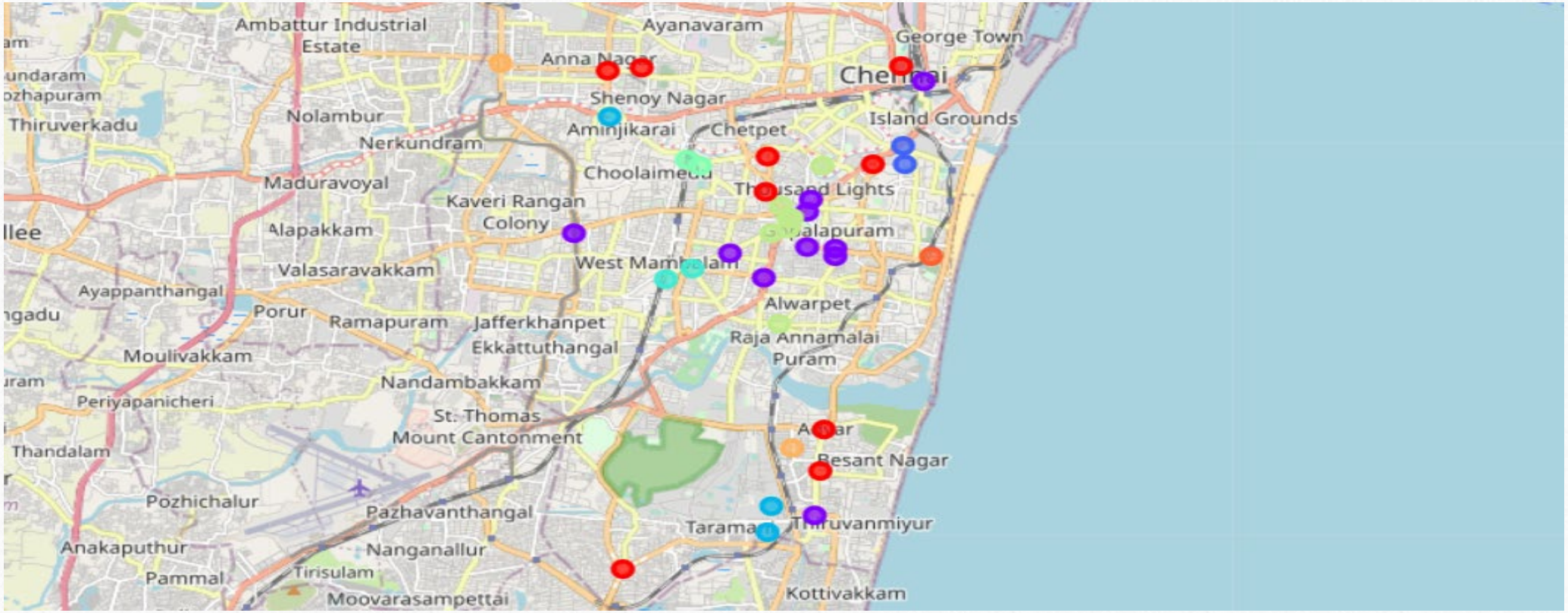
3. Departmental Stores

Repeat the procedure used in hotels. Although cluster 3 have departmental stores but cluster 4 and cluster 7 is also suitable neighborhood for business. Since we don't have enough information from foursquare

- **Map of Chennai with the clusters superimposed on top**
- This map can be used to find a suitable location to start a new business based on the category.

  for example: Red = Cluster 1 , Violet = Cluster 2, Sky Blue = Cluster 4, Light Green = Cluster 7.

  These clusters are suitable to start a business in Chennai.



Conclusion

- Purpose of this project was to analyze the neighborhoods of Chennai and create a clustering model to suggest personals places to start a new business based on the category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 10 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e. 8 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster.
- A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided. Both these can be used by stakeholders to decide the location for the particular type of business. A major drawback of this project was that the Foursquare API returned only few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.