CHENNAI
THE BATTLE OF NEIGHBORHOODS

report by: Balkrishna Mishra

INDEX

01

BACKGROUND: BUSINESS PROBLEM

Introduction: Business Problem

This project deals with discussing the neighborhoods of **Chennai,** Tamilnadu in India. This project would specifically help Business community planning to start **Restaurants, Hotels, Café etc.** in Chennai

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 will be further analyzed to help business owners selecting a potential location to open-up Hotels, Shopping Malls, Restaurants or Coffee shops.

02

DATA GATHERING

Data Gathering

Chennai has multiple potential neighborhoods. Let's use the following dataset with data wrangling and beautification techniques to produce meaningful information.

- Loading Data from various sources

Chennai Neighborhood data with their Latitude and Longitude from below path and then converting html data into a data frame (table).
https://chennaiiq.com/chennai/latitude_longitude_areas.asp

- Exploring Neighborhood with Chennai City Map

Neighborhood venues will be explored within a radius of 500 meters from respective Lat. And Long. from earlier dataset created with the help of Foursquare API.
https://api.foursquare.com

03

PROBLEM SOLVING METHODOLOGY

Problem Solving Methodology

Choosing the correct number of clusters.

https://www.jeremyjordan.me/grouping-data-points-with-k-means-clustering
K - Nearest Neighbor clustering technique have been used and to find the optimal number of clusters silhouette score metric technique is used
A Silhouette coefficient is calculated for observation, which is then averaged to determine the Silhouette score. The coefficient combines the average within-cluster distance with average nearest-cluster distance to assign a value between -1 and 1. A value below zero denotes that the observation is probably in the wrong cluster and a value closer to 1 denotes that the observation is a great fit for the cluster and clearly separated from other clusters. This coefficient essentially measures how close an observation is to neighboring clusters, where it is desirable
to be the maximum distance possible from neighboring clusters.
We can automatically determine the best number of clusters, k, by selecting the model which yields the highest Silhouette score.

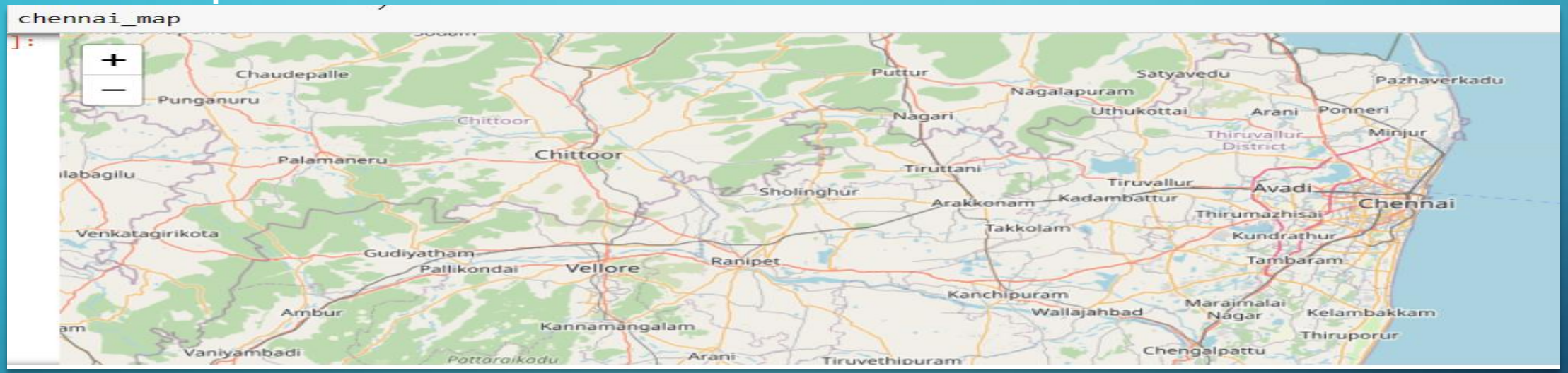
My highest score was 9.

04

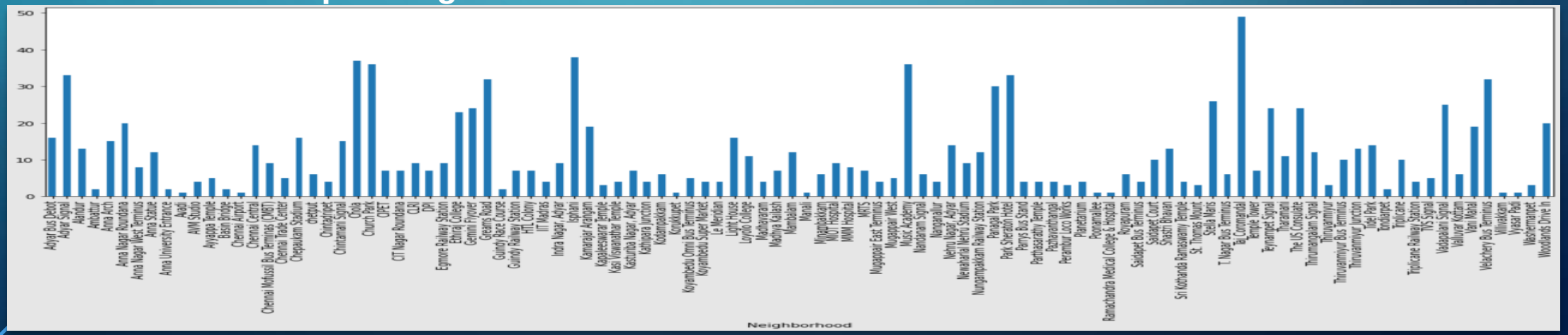
ANALYTICS & RESULTS

Analytics (1/3)

Chennai Map

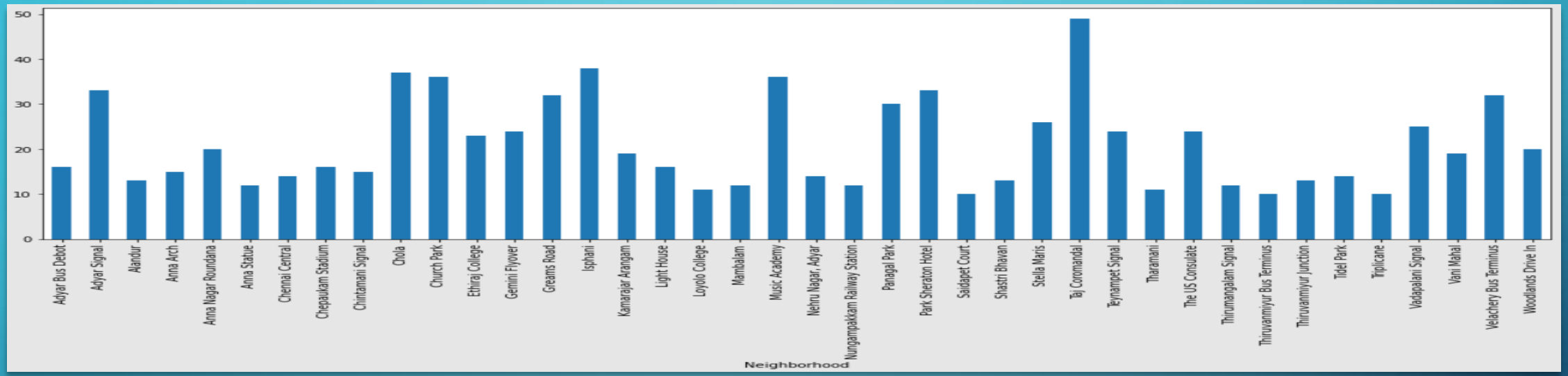


Count # of venues per Neighborhood

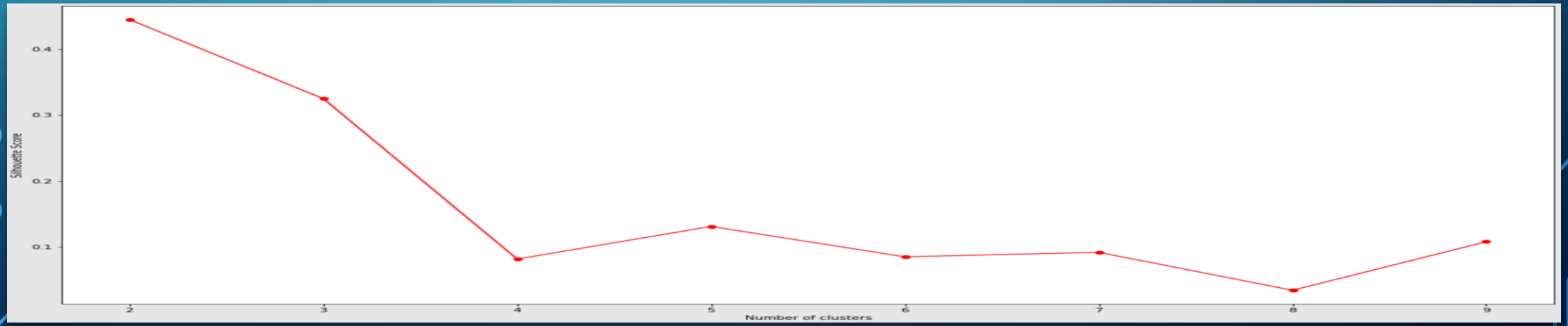


Analytics (2/3)

Removing anomalies (i.e. removing neighborhoods with less than 10 venues)



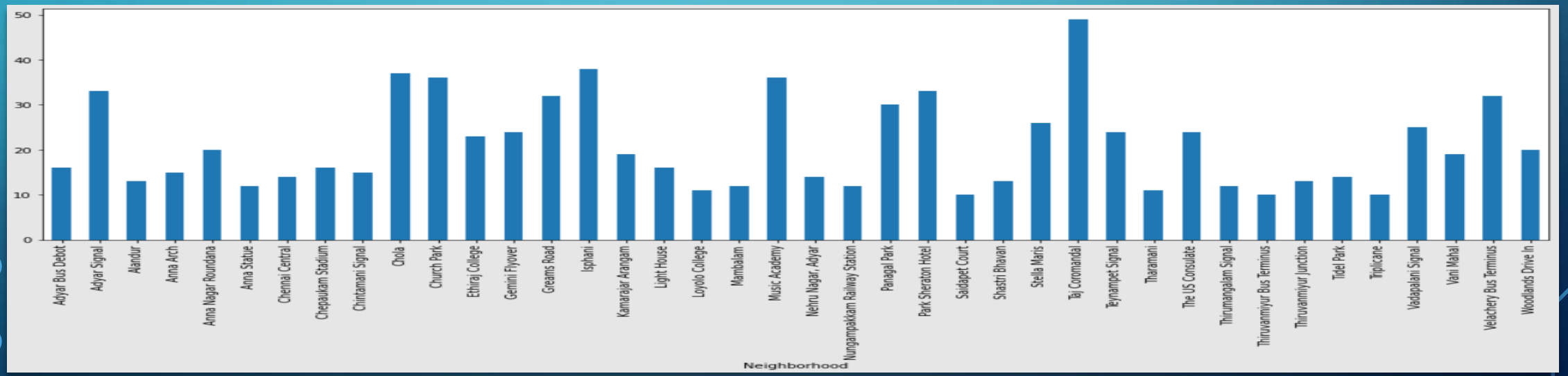
Clustering Neighborhoods



Analytics (2/3)

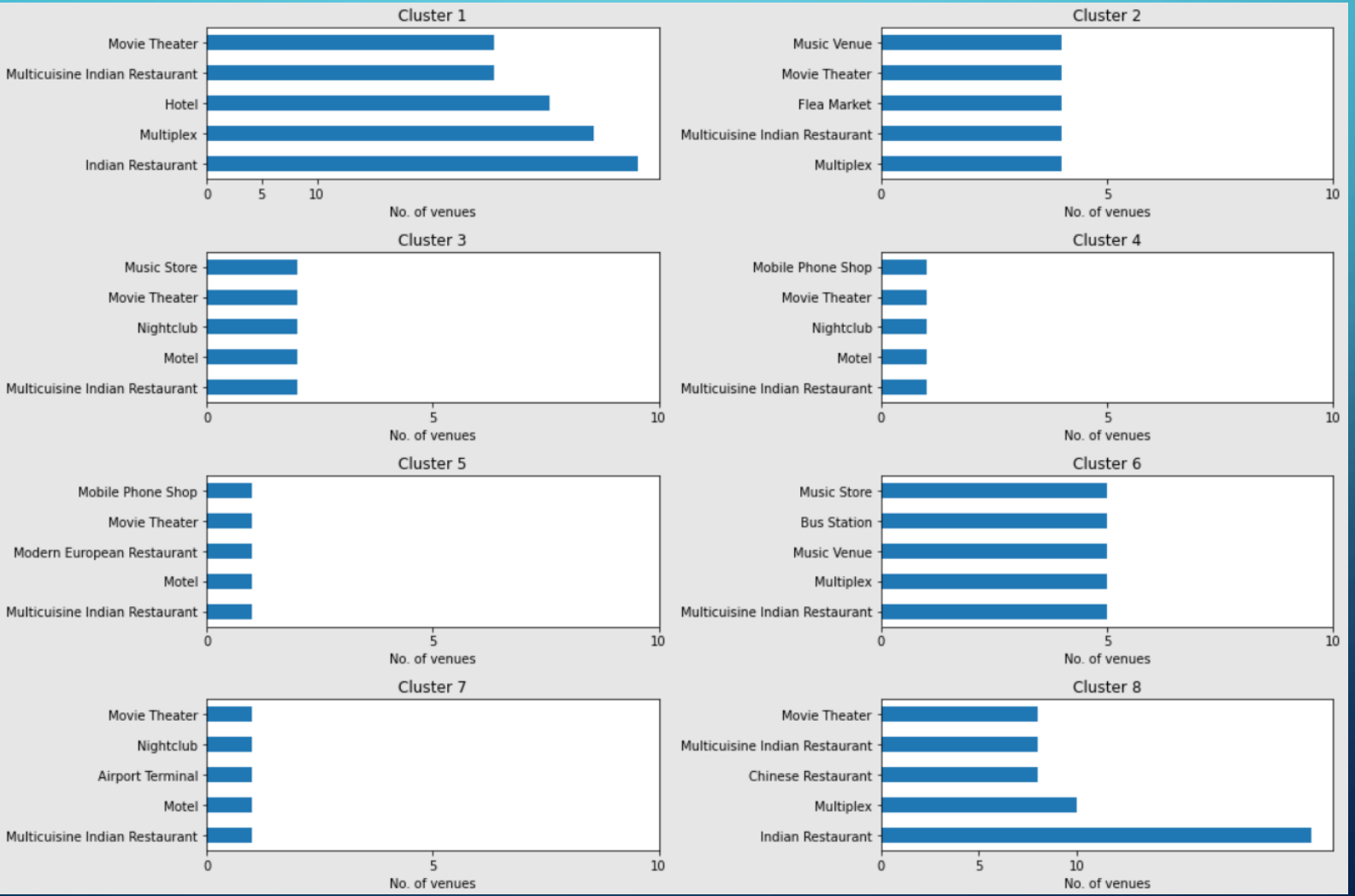
Removing anomalies (i.e. removing neighborhoods with less than 10 venues)

Looking into the dataset we found that there were 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. The resultant dataset consists of **37 neighborhoods**.



Analytics (3/3)

This chart can be used to suggest valuable information to Business persons.
We will now discuss few examples from below outcome as to see a suitable location to start up a business venture basis the diversity and density of existing business environment



05

DISCUSSIONS

Dioscussion

Let us discuss about appropriate business opportunity by Venue Category

5. Discussion

Let us discuss about appropriate business oportunity by Venue Category

A. Hotels

The neighborhoods in **cluster 1** has the greatest number of hotels, 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 6 and Cluster 7 as it has some local cusine restaurant whihc would attract tourists who would love to experience these.

B. Multiplex

The neighborhoods 2 and 5 has notable number of Multiplexes. By using the same procedure as above, the suitable cluster would be the Cluster 1 and since it has not much shopping malls and also it has many Hotels and Restaurants which gives an advantage.

Similarly, based on the requirement suggestions can be provided about the neighborhood that would be best suitable for the business.

06

CONCLUSION

Conclusion

Objective of this project was to analyze the neighborhoods of Chennai and create a clustering model to suggest potential places to start a new business venture based on the available venue category. The neighborhoods data along with latitude and longitude 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 was having less than 10 venues returned. In order to build an appealing recommendation 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 owing to listing efficacy. As a future improvement, better data sources can be used (in place of Foursquare) to obtain more venues in each neighborhood.

However, based on the available data and analytics presented thereupon a decision can be easily draw by the users to decide their entrepreneurial venture.
Superimposed exact location tag for a business set-up under this project at Lat. & Long**. "13.04277 & 80.24602"**