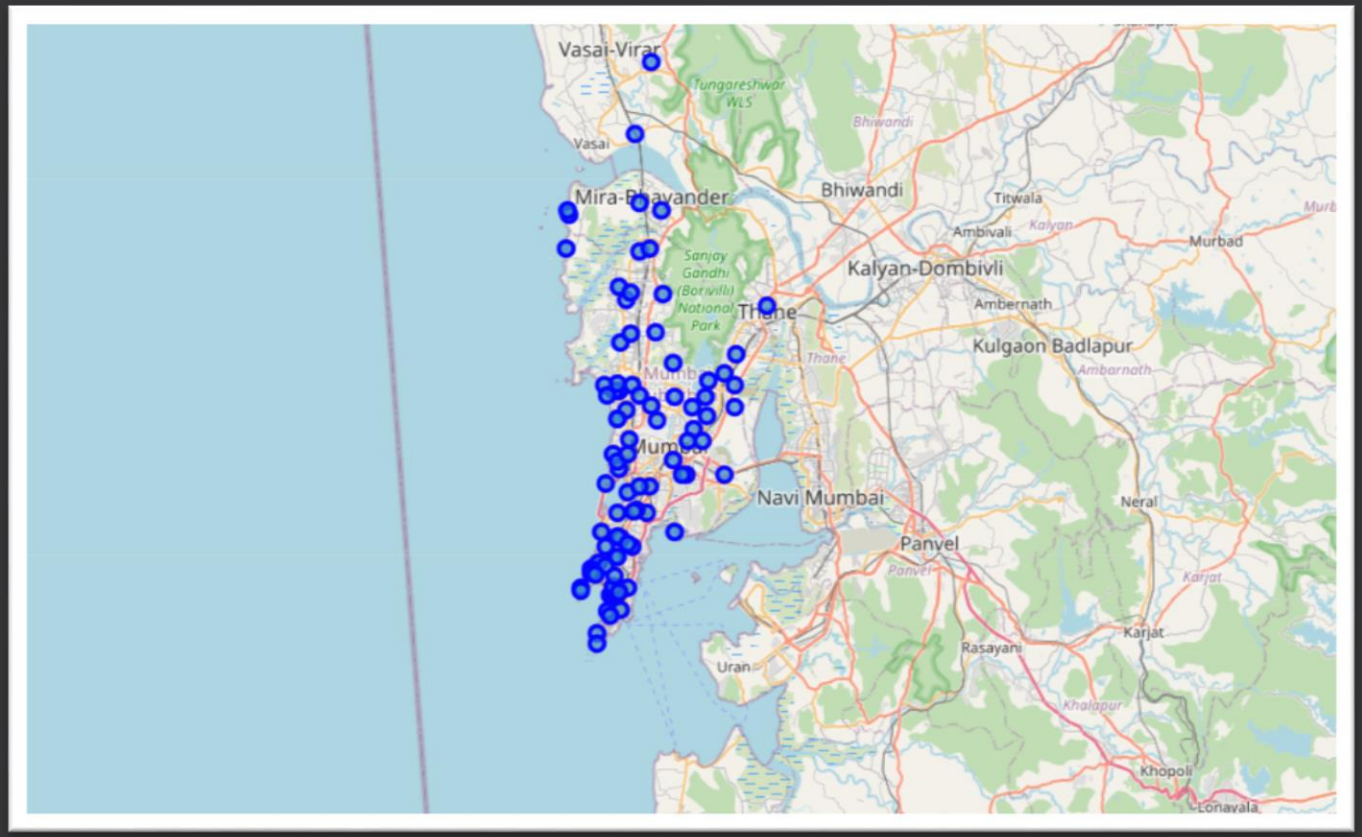
IBM Data Science Capstone Project

The Battle of Neighborhoods – Mumbai

By : Lucky Rathod

January 2021



INTRODUCTION : Business Problem

This project deals with discussing the neighborhoods of **Mumbai**, The Financial Hub of India.This project would specifically help Business people planning to start Restaurants, Hotels, Gym/FitnessCamp/Yoga studio etc.in Mumbai, Maharashtra, India.

Mumbai - City that never Sleeps

The major **Target Audience** would be small-scale business owners and stake holders planning to start their business at a location in Mumbai. 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 new hotel in Mumbai with restaurants around?
- Which area is best suitable for opening a Gym/Yoga/Fitness Camp in Mumbai?

INTRODUCTION : Business Problem

- The Foursquare API is used to access the venues in the neighborhoods. Since, it returns less venues in the neighborhoods, we would be analysing 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 Mumbai 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 or Gym/Fitness camp/Yoga studio.

DATA REQUIREMENTS

- Mumbai has multiple neighborhoods. Lets use the following dataset which has the list of Neighborhoods in Mumbai along with their Latitude and Longitude:

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai#Mumbai_neighbourhood_coordintes

- Next the details of venues in each neighborhood namely **Venue, Venue Latitude, Venue Longitude, Venue Category** data needs to be obtained. Here, Foursquare API is used to obtain this data.

https://foursquare.com/

- A total of 1365 venues data have been obtained from Foursquare

METHODOLOGY

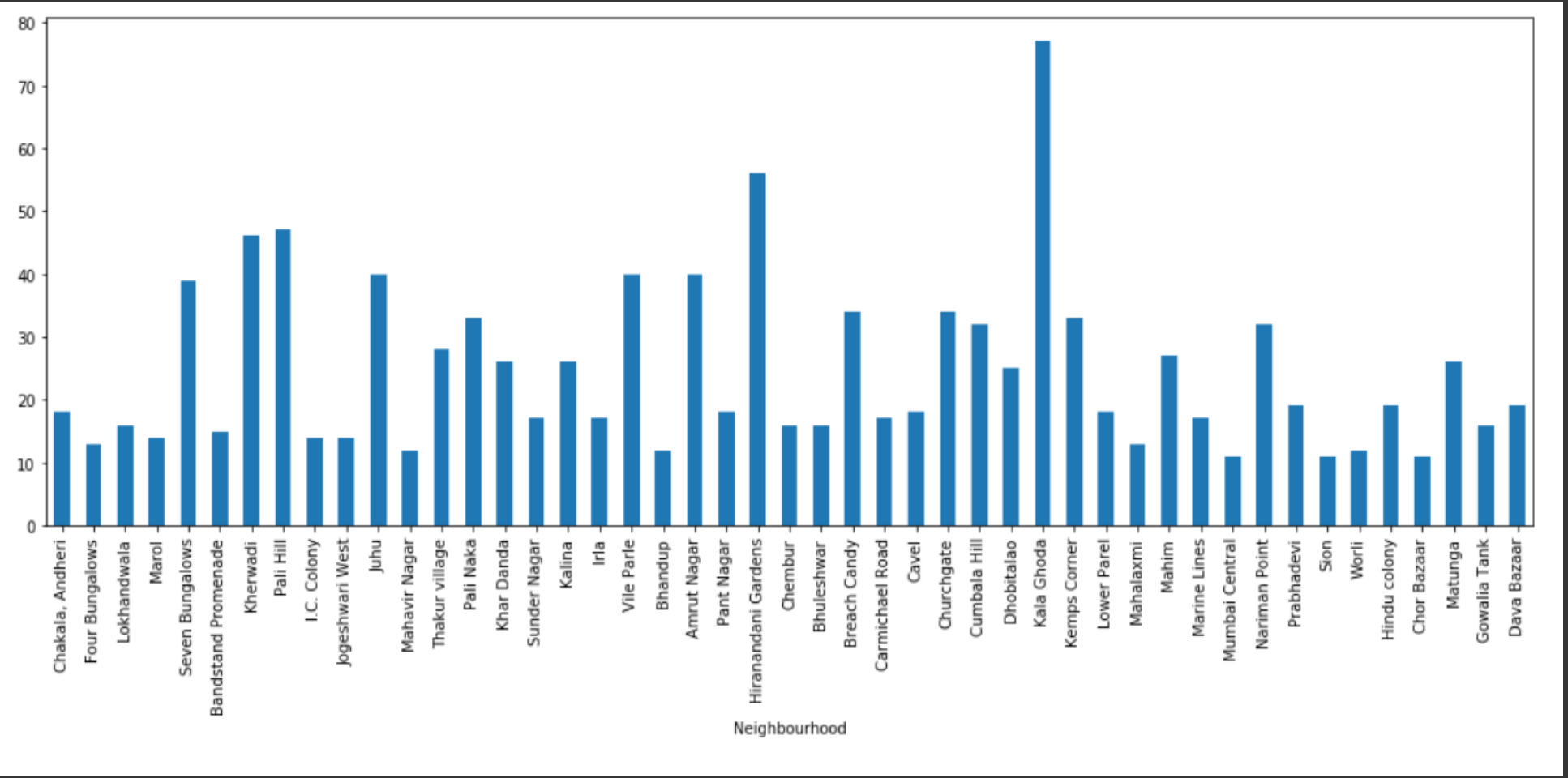
- Now, we have the neighborhoods data of Mumbai **(93 neighborhoods).** We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of **1365 venues** have been obtained in the whole city and **181 unique categories**. But as seen we have multiple neighborhoods with 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

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 **47 neighborhoods.**



Analysis

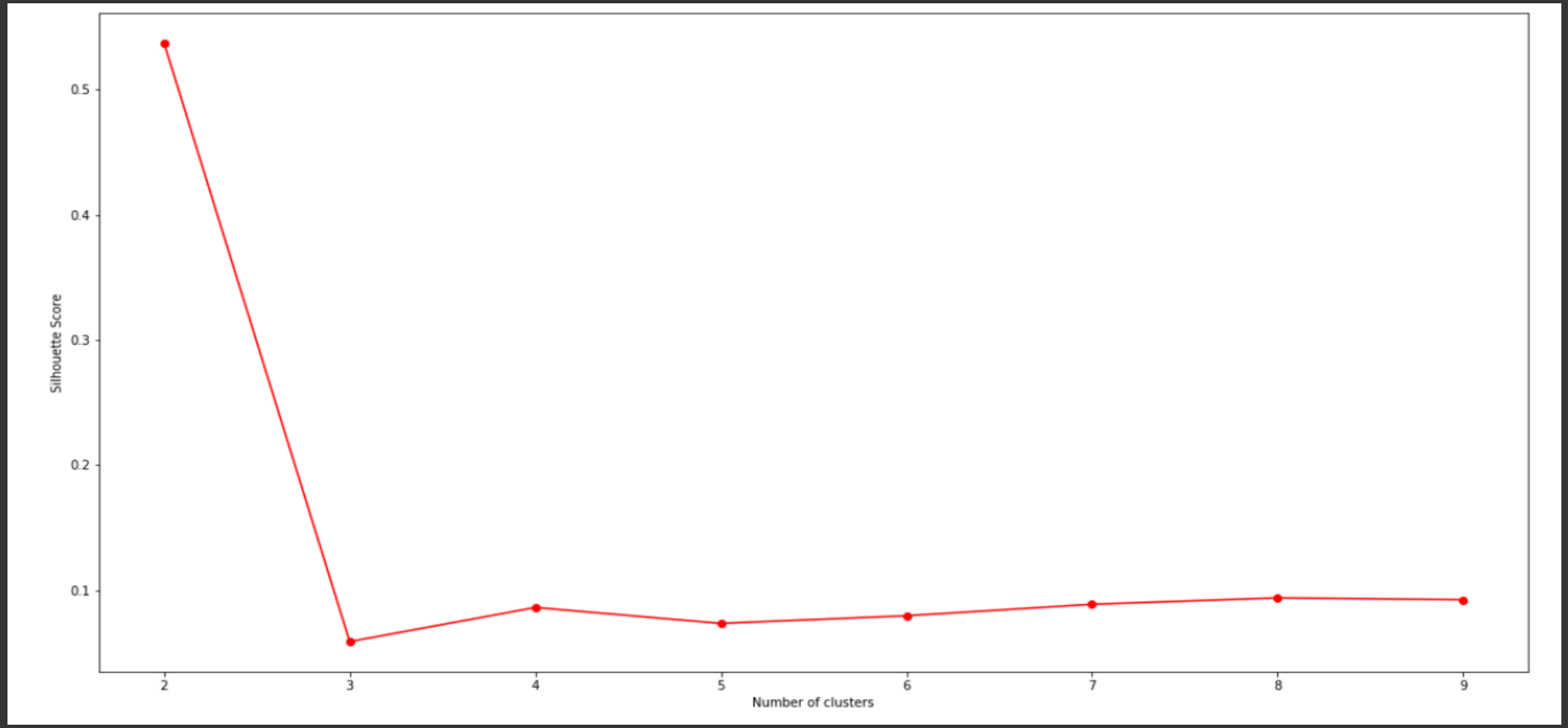
- One hot encoding is performed on the filtered data to obtain the venue categories in each neighborhood. Then group the data by neighborhood and take the mean value of the frequency of occurrence of each category.Sample Output is shown below

- Dataset is used to obtain the top 10 most common venues in each neighborhood i.e. the 10 venues with the highest mean of frequency of occurrence.

- The resultant dataset can be used for the clustering algorithm. Here, the K-Nearest Neighbor (KNN) clustering algorithm is used. It is an unsupervised machine learning technique that clusters the given data into K number of clusters. For optimal result we need to select the best value for K. Here, the silhouette score is used to find the best value for K. A range of values from 2 to 10 was considered, KNN clustering was performed on the dataset and the silhouette score was calculated and plotted on a line plot

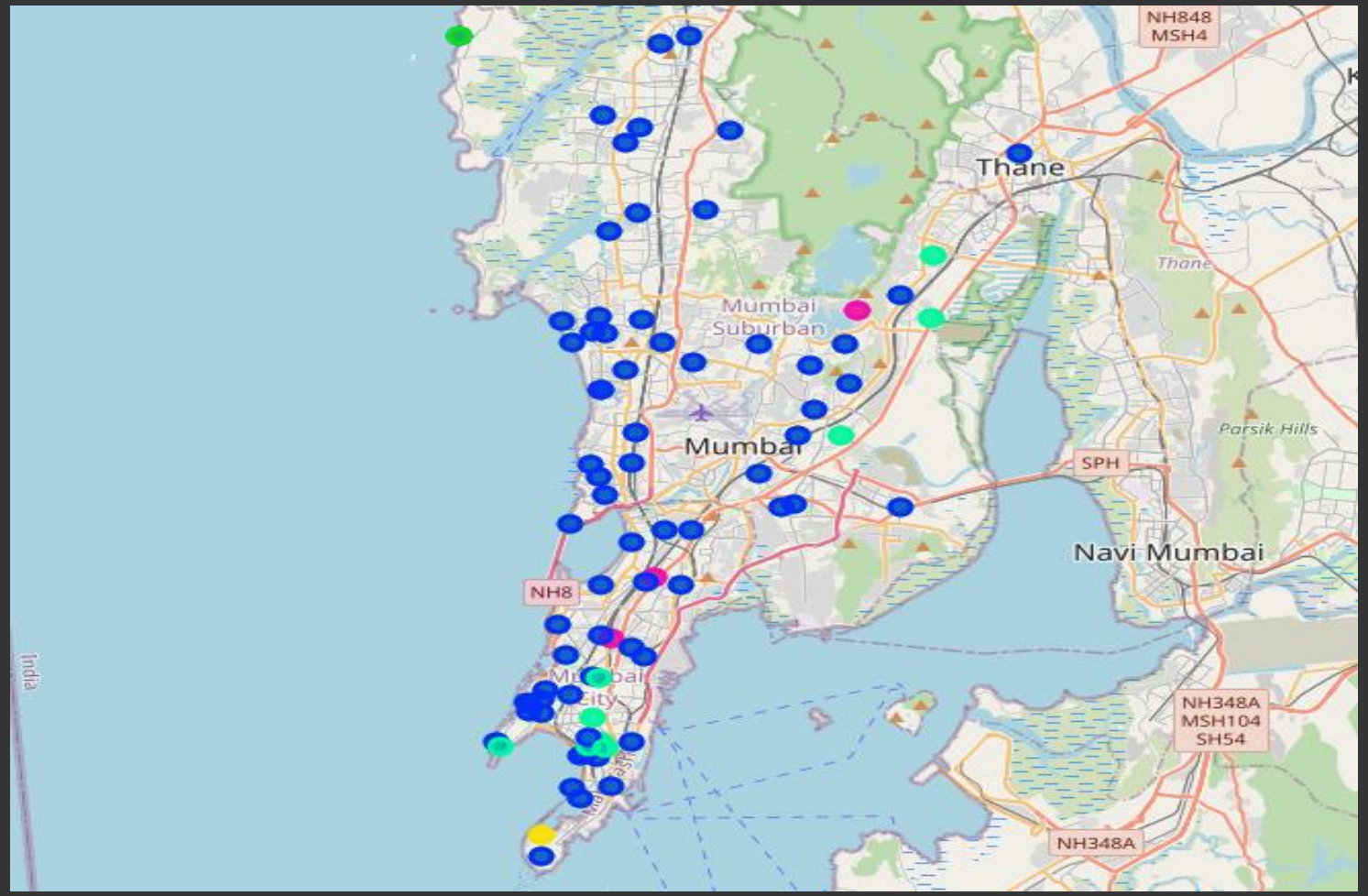
Analysis

From the plot we can see that a K value of 6 provides the best score. This K value is used for the K-Means Clustering Technique. The K-Means labels obtained were included in the top neighborhoods dataset for examining the characteristics of each cluster.



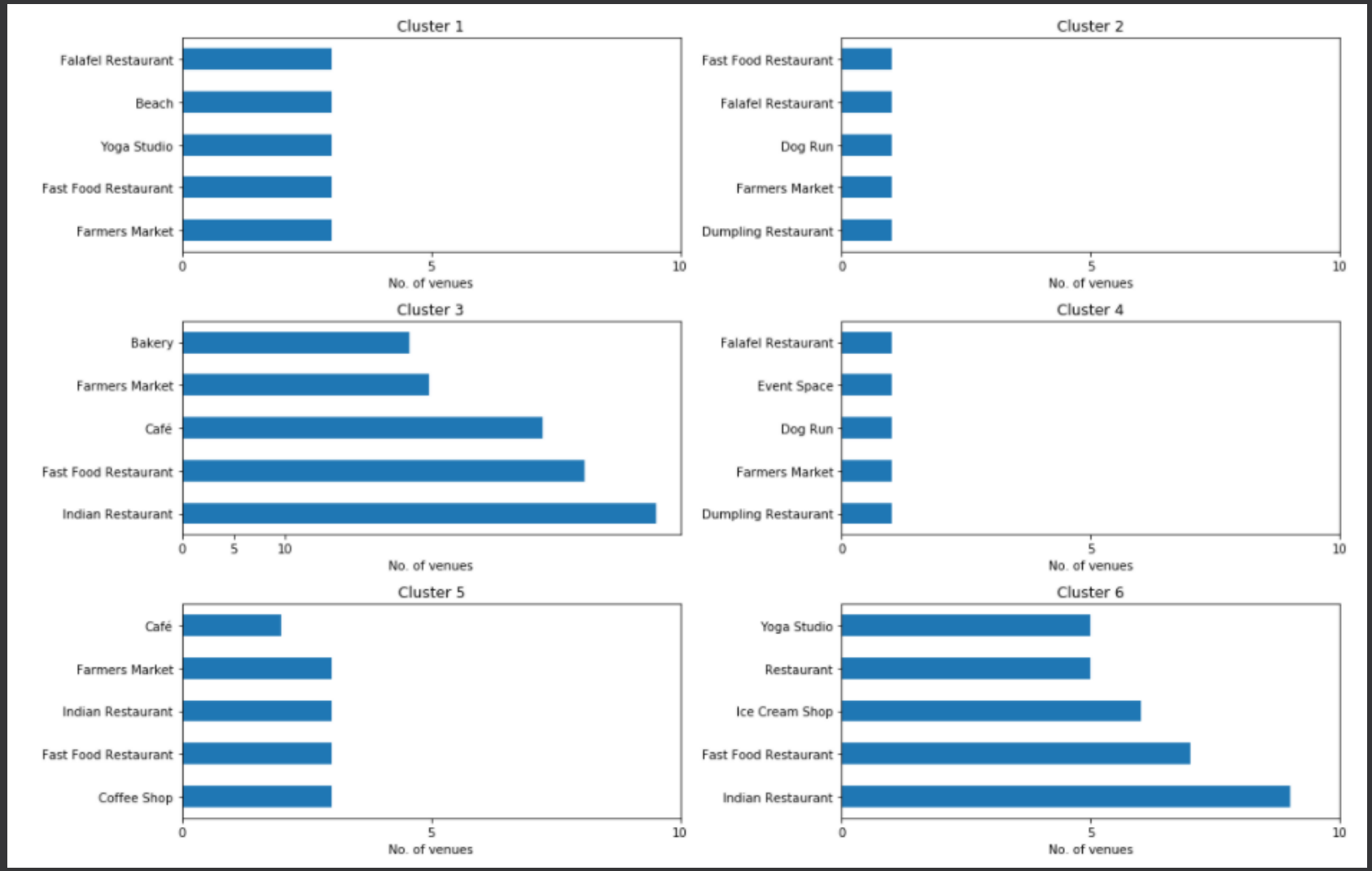
RESULTS And DISCUSSION

Clusters on Mumbai City as per our code



RESULTS And DISCUSSION

We have the clusters and the top venue categories let's visualize the top 5 venue category in each Cluster for comparison.



RESULTS AND DISCUSSION

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 3 and Cluster 6 has the greatest number of Restaurants, hence opening one here is the best choice.We can also open one at the neighborhoods in cluster 2 or 4 because there not many hotels.

2. Gym/Yoga/Fitness Camp

The neighborhoods in cluster 1 and cluster 6 has many food restraurant.Also there are many yoga studios in this neighborhood it is not a great idea.So it will be best to open a yoga/gym/fitness camp in Cluster 5 because it contains cafe and restraurants but there are not many gym/yoga/fitness camp.

Similarly, based on the requirement suggestions can be provided about the neighborhood that would be best suitable for the business . Result map can be used to suggest vast location to start a new business based on the category

RESULTS AND DISCUSSION

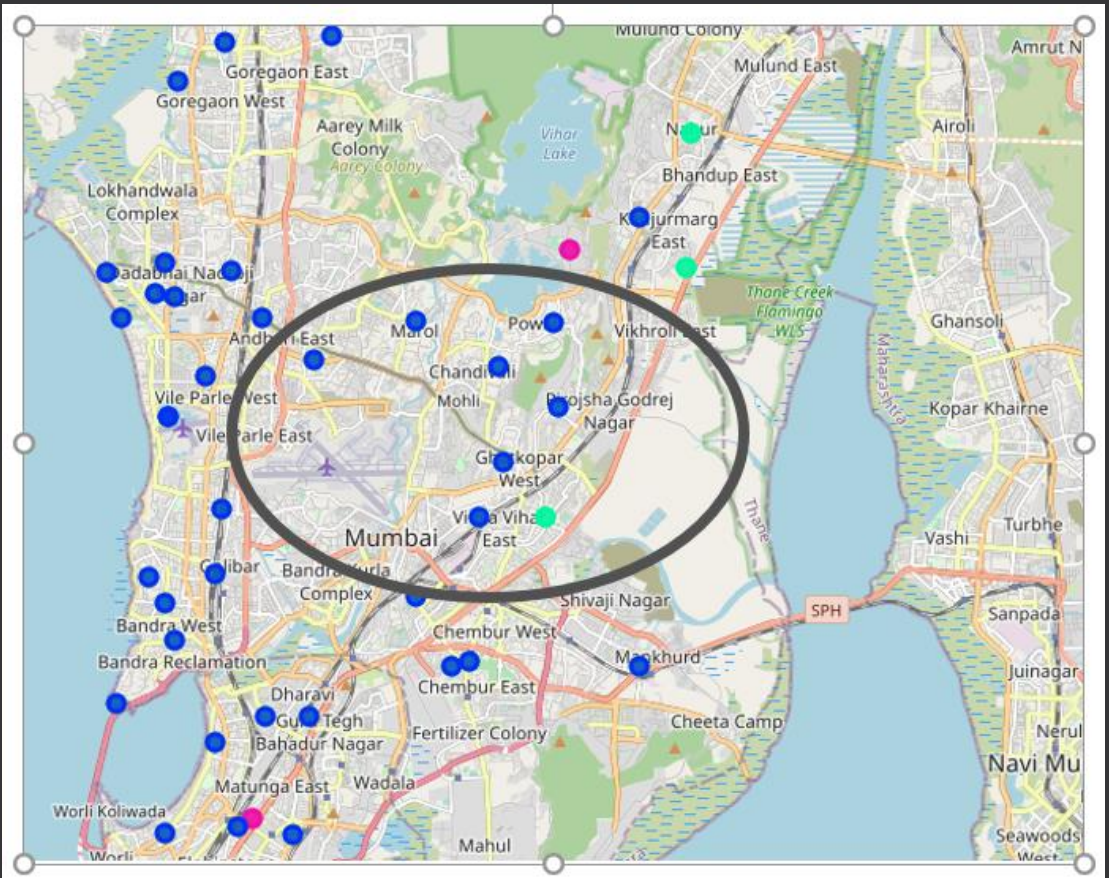


Fig8 : Location Suitable for New Hotel

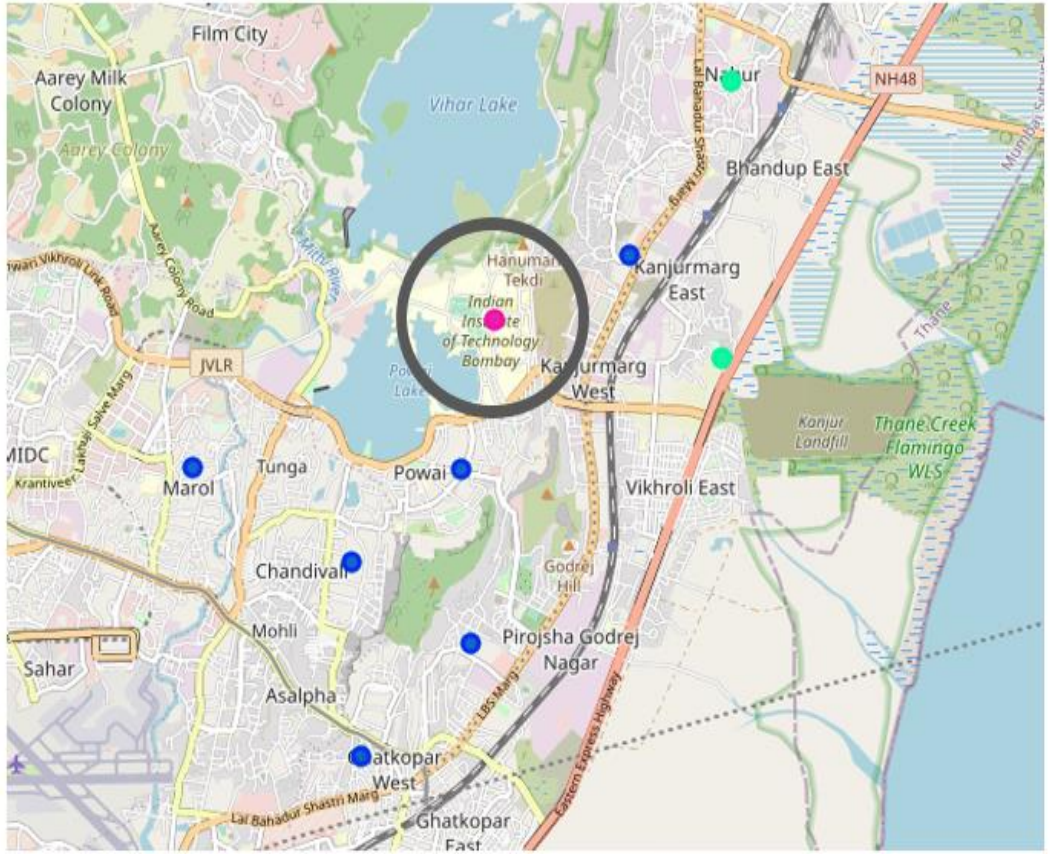


Fig9 : Location Suitable for New Gym/Yoga studio/Fitness camp

DRAWBACKS

- 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 using KNN Clustering Algorithm

CONCLUSION

- Purpose of this project was to analyze the neighborhoods of Mumbai 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. 6 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.

Thank You