**The Battle of Neighborhoods – Coursera / IBM Capstone Project (Week 2)**

**Moyinoluwa Adeniyi**

**EXPLORING MUMBAI NEIGHBORHOODS FOR STARTING A RESTAURANT**

Table of content

1. Introduction
2. Data Collection
3. Methodology
4. Results
5. Discussion
6. Conclusion

**Introduction**

“Life is beautiful.” What really makes life beautiful? To me, the beauty of life is a reflection of the beauty of the people. Human connections, affiliations and relationships are the ingredients cooked up to make the beauty of life. There are many ways people connect, one of which are places such as restaurants. The business of running a restaurant has several benefits which can’t be over emphasized. In this project, I will be exploring Mumbai neighborhood data with the aim of starting a restaurant.

Mumbai, also known as Bombay is the [capital city](https://en.wikipedia.org/wiki/Capital_city) of the [Indian](https://en.wikipedia.org/wiki/India). According to the [United Nations](https://en.wikipedia.org/wiki/United_Nations), as of 2018, Mumbai is the [second-most populous](https://en.wikipedia.org/wiki/List_of_cities_in_India_by_population) city in the country after [Delhi](https://en.wikipedia.org/wiki/Delhi) and the [seventh-most populous](https://en.wikipedia.org/wiki/List_of_largest_cities) city in the world with a population of roughly 20 million. Mumbai is the [financial](https://en.wikipedia.org/wiki/Financial_centre), commercial, and the entertainment capital of India.  It is also one of the world's top ten centres of commerce in terms of global financial flow, generating 6.16% of India's GDP, and accounting for 25% of industrial output, 70% of maritime trade in India ([Mumbai Port Trust](https://en.wikipedia.org/wiki/Mumbai_Port_Trust) and [JNPT](https://en.wikipedia.org/wiki/Jawaharlal_Nehru_Port)), and 70% of capital transactions to [India's economy](https://en.wikipedia.org/wiki/Economy_of_India) (Mahajan Poonam; July, 2014).

Everyone, irrespective or race, religion or background loves food and the people of India are not an exception. Thus, the aim of this project is to study the neighborhoods in Mumbai to determine possible locations for starting a restaurant.

This findings/results from this project can be a powerful resource/tool for business owners and entrepreneurs who are looking to invest in a restaurant in Mumbai. The main objective of this project is to carefully analyze appropriate data and find recommendations for the stakeholders concerning the best areas/neighborhoods to invest in.

**DATA COLLECTION**

The required data for this project include the following:

1) Neighborhood data of Mumbai was scrapped from the webpage below:

https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Mumbai

2) Geographical coordinates of Mumbai and its Neighborhood was obtained using GeoPy library in python.

3) Venue data for neighborhoods in Mumbai was obtained using Foursquare API.

Finally, Folium library was employed in visualizing the obtained data.

**Neighborhoods Data:**

The data of the neighborhoods in Mumbai was scraped from  
https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Mumbai. The data is  
read into a pandas data frame using the read\_html() method. The main reason for  
doing so is that the Wikipedia page provides a comprehensive and detailed table  
of the data which can easily be scraped using the read\_html() method of pandas.  
The top 10 rows of the dataframe are shown in Figure 1.

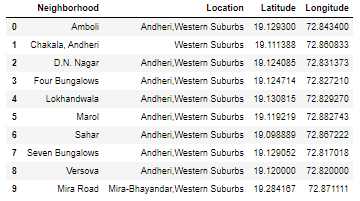


Figure 1: Top 10 rows of Mumbai neighborhoods data scraped from Wikipedia.

**Geographical Coordinates:**

The geographical coordinates for Mumbai has been obtained from the GeoPy  
library in python. This data is relevant for plotting the map of Mumbai using the  
Folium library in python. The code for getting the geographical coordinates of  
Mumbai is shown in Figure 2.

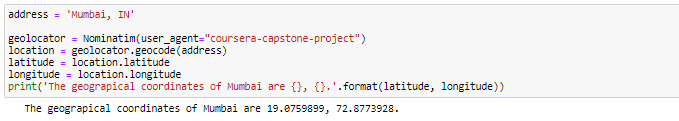


Figure 2: Obtaining geographical coordinates of Mumbai.

The geocoder library in python has been used to obtain latitude and longitude  
data for various neighborhoods in Mumbai. The coordinates of all neighborhoods  
in Mumbai are used to check the accuracy of coordinates given on Wikipedia and  
replace them in our data frame if the absolute difference is more than 0.001. These  
refined coordinates are then further used for plotting neighborhoods using the  
Folium library in python. Figure 3 shows the coordinates of neighborhoods in  
Mumbai obtained from Wikipedia as ‘Latitude’ and ‘Longitude’ and those  
obtained from geocoder as ‘Latitude1’ and ‘Longitude1’. Furthermore, it also  
shows the absolute difference between the two latitude columns and the two  
longitude columns as ‘Latdiff’ and ‘Longdiff’, respectively. Once again only the  
top 10 rows are shown.

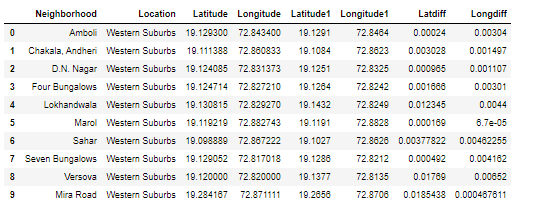


Figure 3: Absolute difference between latitude and longitude values obtained from Wikipedia and Geo coder.

Figure 4 shows the top 10 rows of the final Mumbai neighborhoods data frame  
after replacing the latitude and longitude values as mentioned before and  
dropping unnecessary columns.

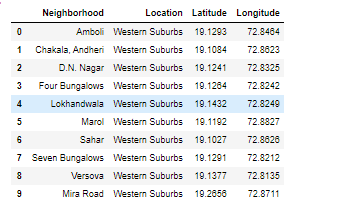


Figure 4: Final Mumbai neighborhoods dataframe

**Venue Data:**

The venue data has been extracted using the Foursquare API. This data contains  
venue recommendations for all neighborhoods in Mumbai and is used to study  
the popular venues of different neighborhoods as well as build the unsupervised  
learning model to cluster neighborhoods. The venue recommendations of all  
neighborhoods were obtained with a limit of 200, that is, maximum of 200 venue  
recommendations per neighborhood and a radius of 1 km around the neighborhood’s geographical coordinates. Figure 5 shows the top 10 rows  
depicting the results obtained after cleaning the data from Foursquare API.

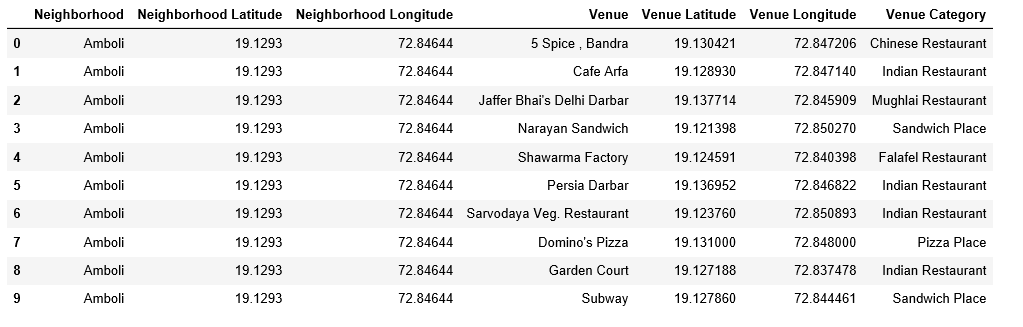


Figure 5: Data obtained from Foursquare API after cleaning.

**Methodology**This section provides details for the methodology used in the project.

**Data Visualization**

In order to understand the data obtained for Mumbai neighborhoods, basic  
visualization was carried out. Figure 6 shows a bar plot depicting the number of  
neighborhoods in each location in Mumbai.

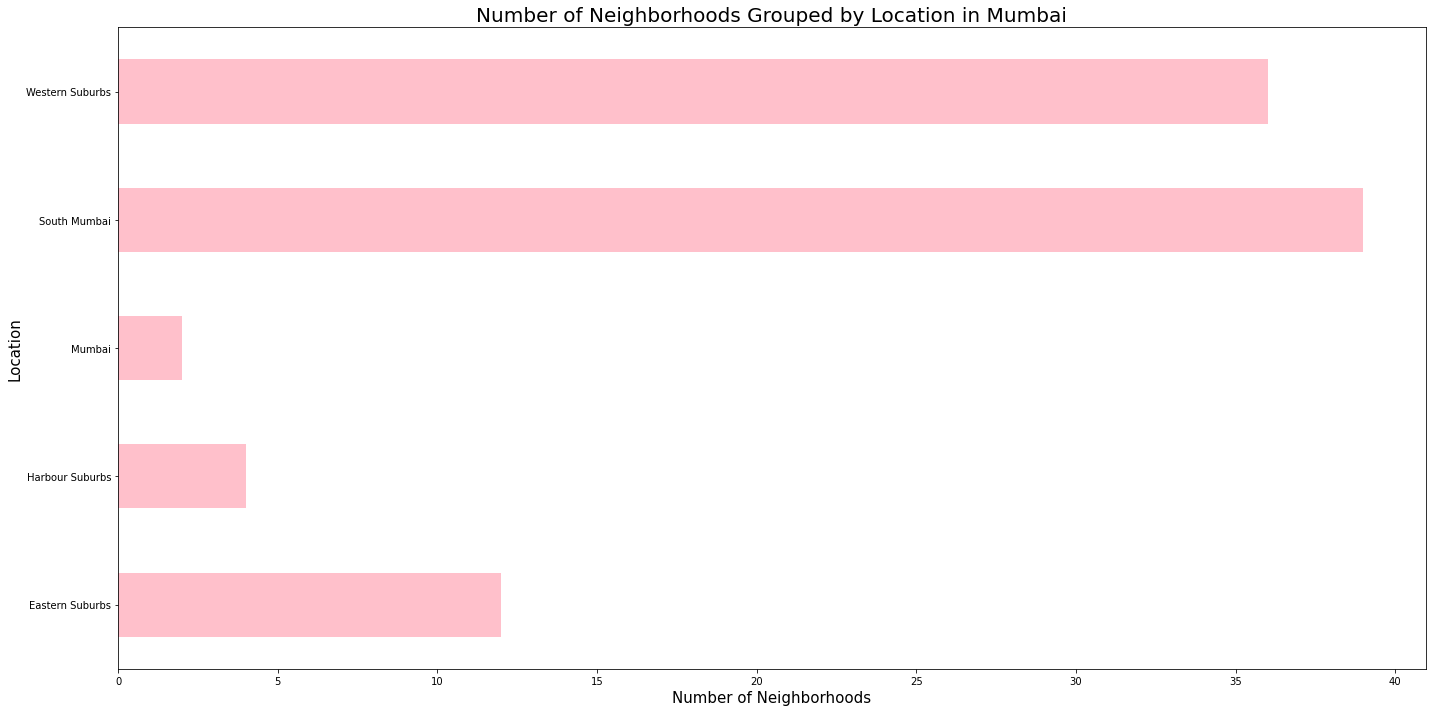
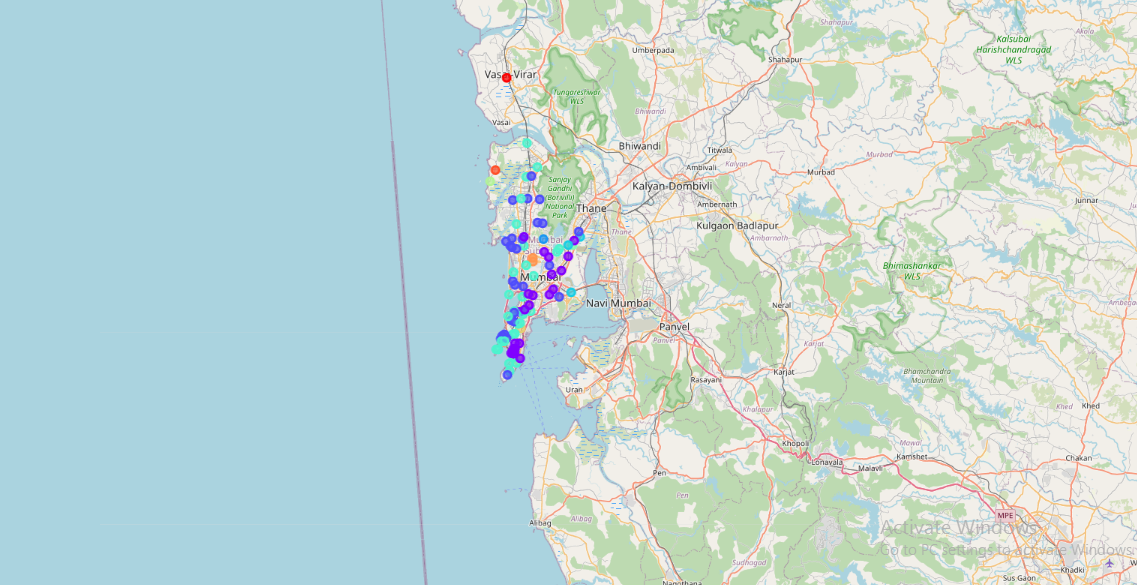


Figure 6: Number of neighborhoods grouped by location.

Using folium, a map was plotted to show how the different neighborhoods are  
spread all across Mumbai. This is shown in Figure 7.

Figure 7: Depicting the neighborhood spread across Mumbai.



**Feature Extraction**

Feature extraction was carried out to obtain features from the Foursquare API  
data (as shown in Figure 5) which was used for building the unsupervised learning  
model. In order to achieve this, the “Venue Category” column had to be converted  
to some form of numeric value to be used for building the model. This was  
achieved by the One-hot Encoding method which takes all the unique categories  
and creates a column for each category. Then, if a neighborhood venue belongs  
to that category, it would get a value of 1 for that row in that specific category  
column and if a neighborhood venue does not belong to the particular category,  
the value would be 0. This process was repeated for all venues in all  
neighborhoods and the result was a sparse matrix containing the neighborhood name and all unique category columns with either 1 or 0 based on whether the  
neighborhood venue belonged to that category or not. This data frame was then  
grouped by the neighborhood name and the average value was taken for all  
categories. The result is shown in Figure 8 which shows only the top 10 rows.

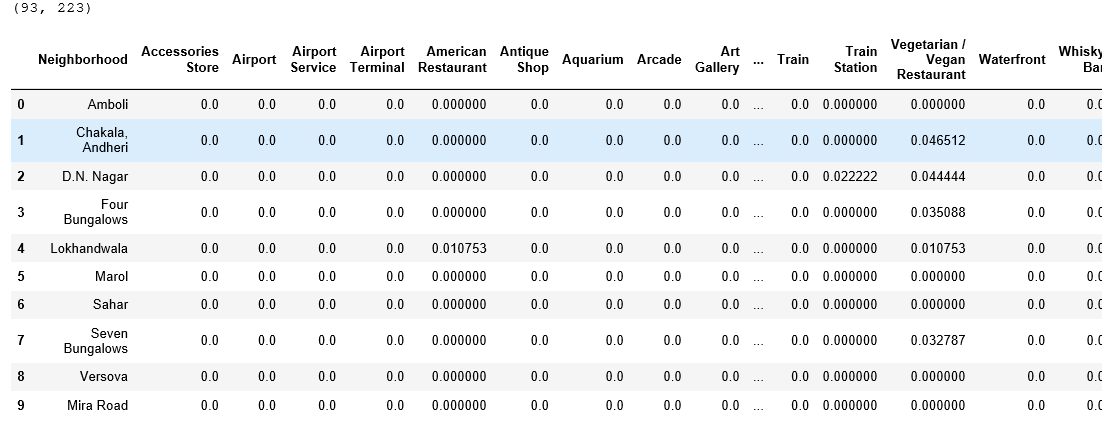


Figure 8: One-hot Encoding resulting data frame.

Notice that most of the values are 0 since there were a large number of unique  
categories and not all neighborhoods had venues belonging to each category.  
This data was used for the unsupervised learning model with the neighborhood  
name dropped. The unsupervised learning model is explained in the next section.  
A data frame was also created which contained the top 10 most common venues  
of all neighborhoods. Though this is not a part of Feature Extraction, it is  
important to provide a glimpse into what this data frame looks like as it will be  
used later to combine the results from the unsupervised learning model. The top  
10 rows of this data frame are shown in Figure 9.



Figure 9: Top 10 most common venues for neighborhoods.

**Unsupervised Learning**

K-means unsupervised learning technique was used to cluster the  
neighborhoods based on the category of venues near the neighbourhoods. One  
important aspect of the k-means model is to determine the number of clusters to  
use in model development. This was determined by the Silhouette score which  
was calculated for a range of clusters from 2 to 15. The resulting number of  
clusters and their respective Silhouette scores are shown in Figure 10.

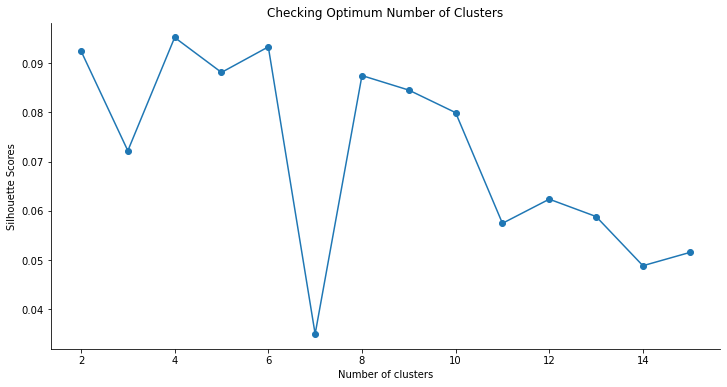


Figure 10: Silhouette scores for different number of clusters.

**Results**  
The clustering model then clusters the neighborhoods in Mumbai and provides  
a label for each neighborhood which is representative of the cluster it belongs  
to. The cluster labels were then added to the data frame in Figure 9 along with  
the Location, Latitude, and Longitude columns to provide a complete summary  
of the clustering. The top 10 rows are shown in Figure 11.

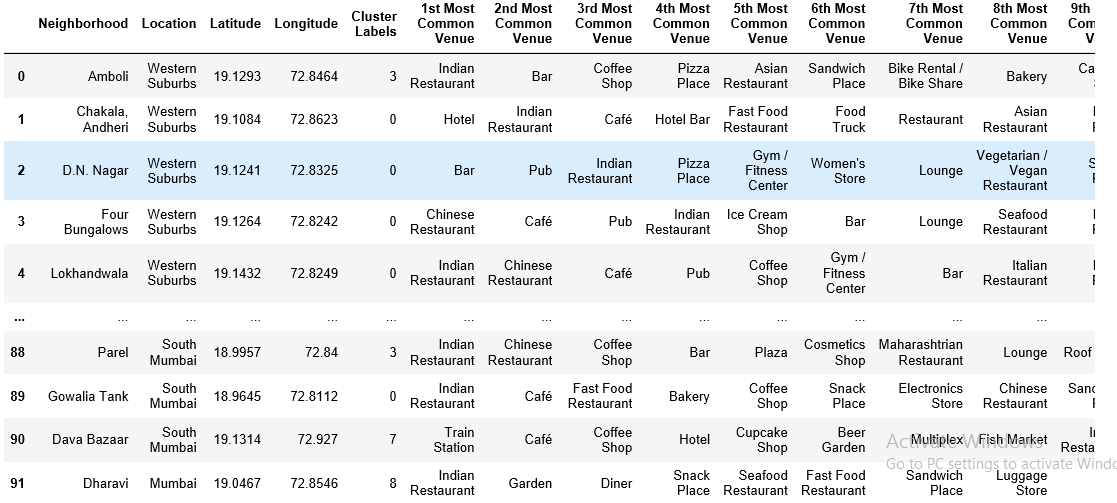


Figure 11: Clustering neighborhoods in Mumbai.

Furthermore, neighborhoods in each individual cluster can be extracted using  
cluster labels and thus the details of specific clusters can be seen. This is done  
below for all clusters with only 10 rows for clusters that contain a high number  
of neighborhoods.

Figure 12: Cluster 1.

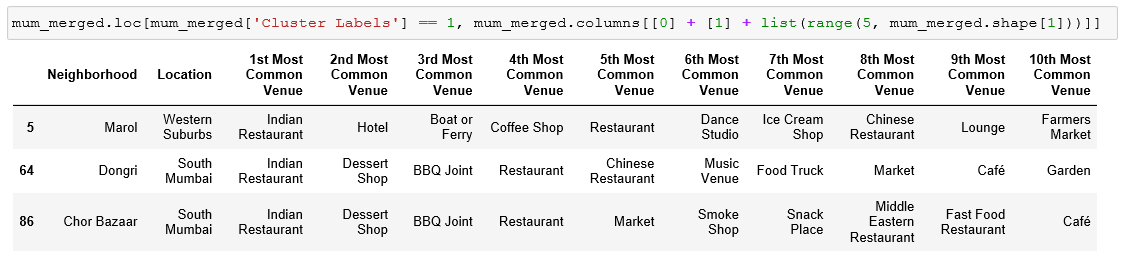
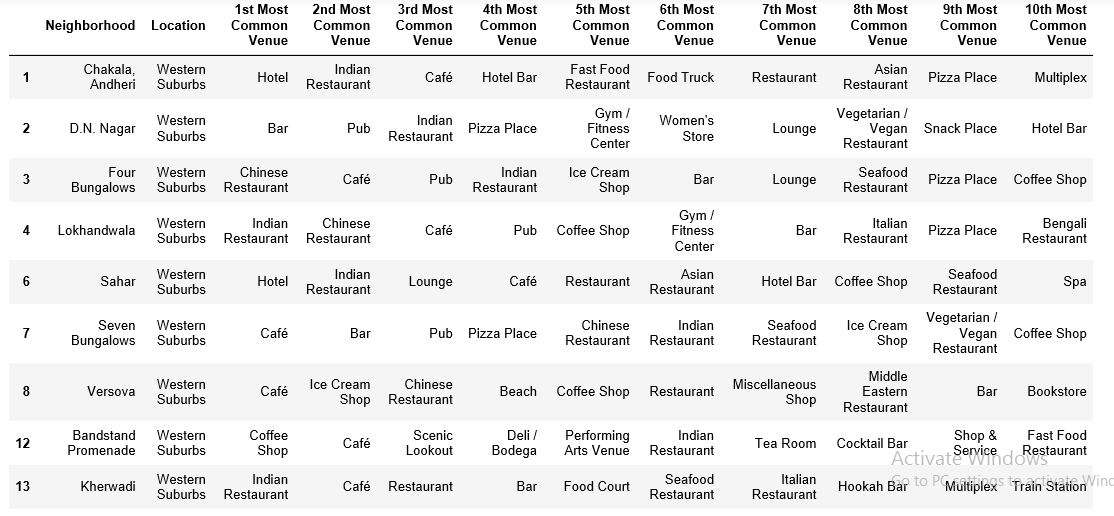


Figure 13: Cluster 2.

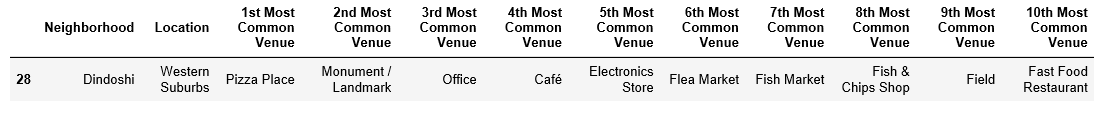


Figure 14: Cluster 3



Figure 15: Cluster 4

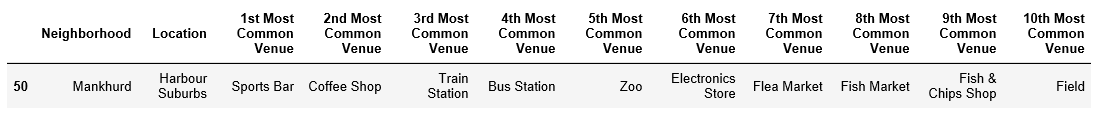
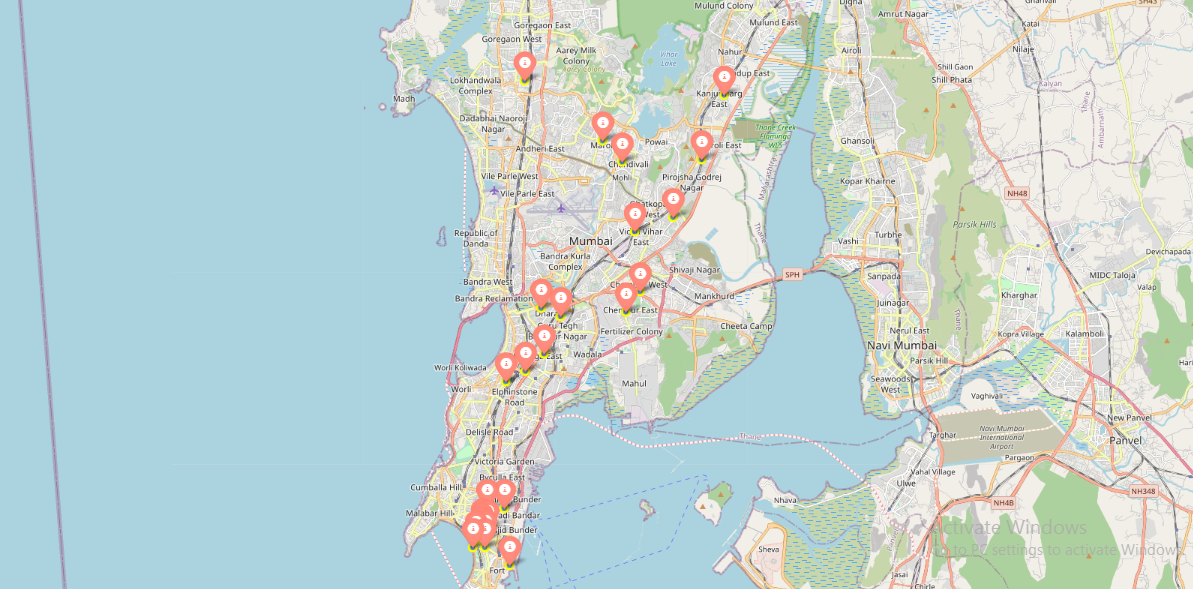


Figure 16: Cluster 5

Based on the clusters shown above, the neighborhoods can once again be plotted  
on a map of Mumbai, however, this time with different colour markers to  
distinguish between different clusters. This is shown in Figure 17.



## **Discussion**

From my analyzing of the five clusters obtained, it is observed that cluster 1 has a high degree of hotels, bars, Chinese restaurants, cafes, Indian restaurants, clothing stores, etc. in its top 10 most common venues. Cluster 2 has mainly Indian restaurants in its first most common venue. Cluster 3 has pizza place in its first most common venue. Cluster 4 has mainly Indian restaurants in its first most common venue. Cluster 5 has sports bar in its first most common venue.

From these analyses, since most neighborhoods in cluster 4 seem to have Indian Restaurant as their top most common venue; it is recommended that the new restaurant can be opened in the neighborhoods belonging to cluster 4 since it looks best suited for the purpose of starting a new restaurant. This neighborhood can be further plotted on a map as shown below.

# Conclusion

I have successfully completed analysis of the neighborhoods in Mumbai, India for determining which would be the best neighborhoods for opening a new restaurant. Based on our analysis, neighborhoods in cluster 4 are recommended as locations for the new restaurant. This has also been plotted in the map above. The stakeholders and investors can further tune this by considering various other factors like transport, legal requirements, and costs associated. These were out of the scope for this project and thus were not considered.