**Analyzing the Neighborhoods in Mumbai for Starting a Restaurant**

Applied Data Science Capstone Project

By: Murugavel C

Date: August 26, 2021

Table of Contents

[Introduction 2](#_Toc47451679)

[Data Collection 2](#_Toc47451680)

[Neighborhoods Data 3](#_Toc47451681)

[Geographical Coordinates 3](#_Toc47451682)

[Venue Data 5](#_Toc47451683)

[Methodology 6](#_Toc47451684)

[Data Visualization 6](#_Toc47451685)

[Feature Extraction 7](#_Toc47451686)

[Unsupervised Learning 9](#_Toc47451687)

[Results 10](#_Toc47451688)

[Discussion 13](#_Toc47451689)

[Conclusion 15](#_Toc47451690)

[Final Comments 15](#_Toc47451691)

# Introduction

Mumbai is the financial capital of India and is one of the most densely populated cities in the world. It lies on the west coast of India and attracts heavy tourism from all over the globe every year. Personally, I have been brought up in Mumbai and have loved the city from the bottom of my heart. It is one of the major hubs of the world and is extremely diverse with people from various ethnicities residing here. The multi-cultural nature of the city of Mumbai has brought along with it numerous cuisines from all over the world. The people of India generally love food and I personally love to try different cuisines and experience different flavors. Thus, the aim of this project is to study the neighborhoods in Mumbai to determine possible locations for starting a restaurant. This project can be useful for business owners and entrepreneurs who are looking to invest and open a restaurant in Mumbai. The main objective of this project is to carefully analyze appropriate data and find recommendations for the stakeholders.

# Data Collection

The following data is required for the project:

1. Neighborhood data of Mumbai
2. Geographical coordinates of Mumbai and all neighborhoods in Mumbai
3. Venue data for neighborhoods in Mumbai

## Neighborhoods Data

The data of the neighborhoods in Mumbai was scraped from [https://en.wikipedia.org/wiki/List\_of\_neighborhoods\_in\_Mumbai](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.

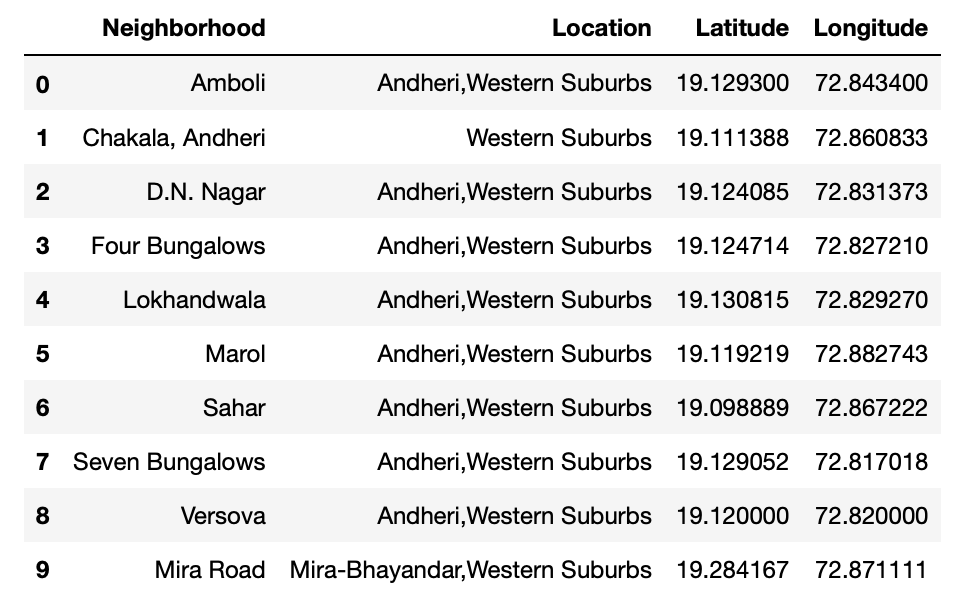


Figure 1: Top 10 rows of Mumbai neighborhoods data scraped from Wikipedia.

## 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.

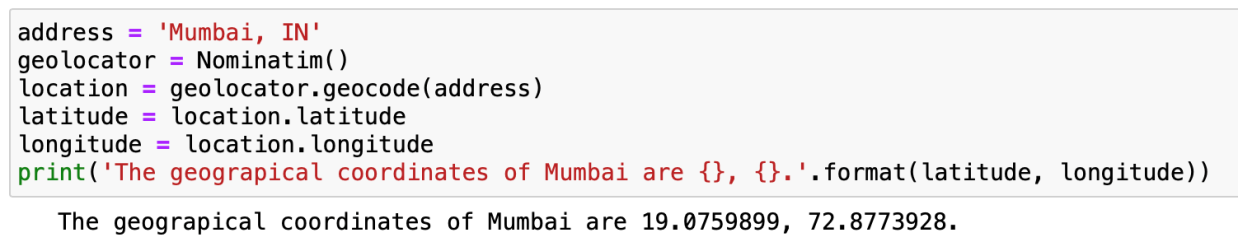


Figure 2: Obtaining geographical coordinates of Mumbai.

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.

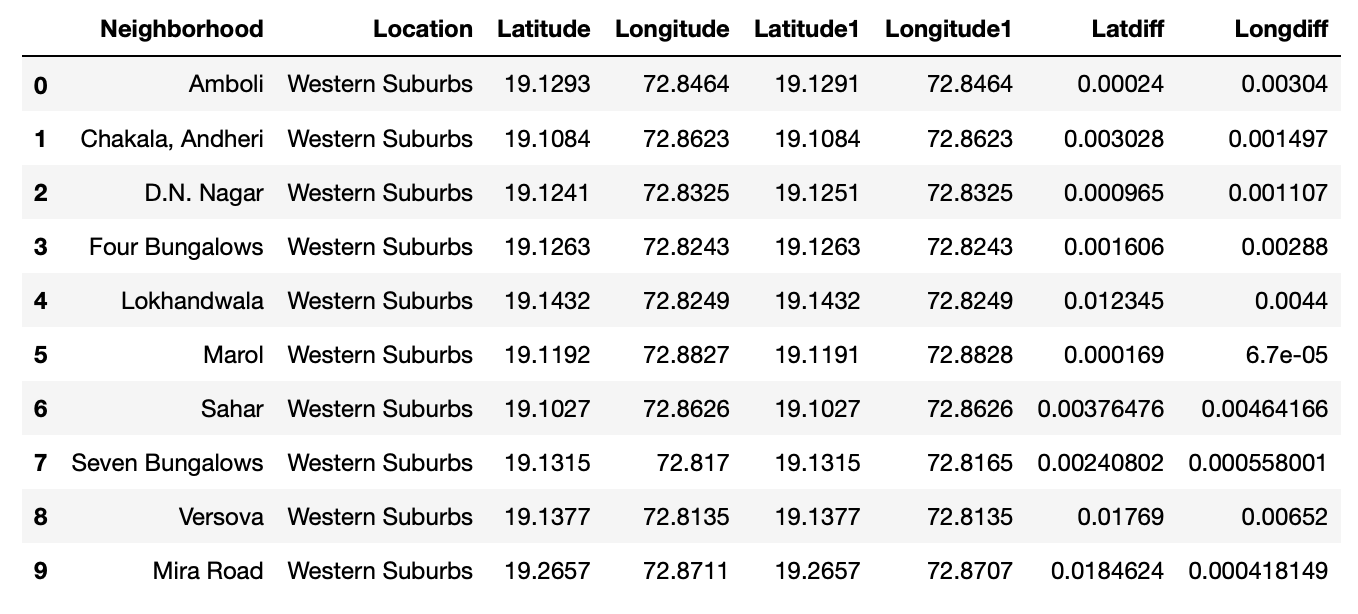


Figure 3: Absolute difference between latitude and longitude values obtained from Wikipedia and Geocoder.

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

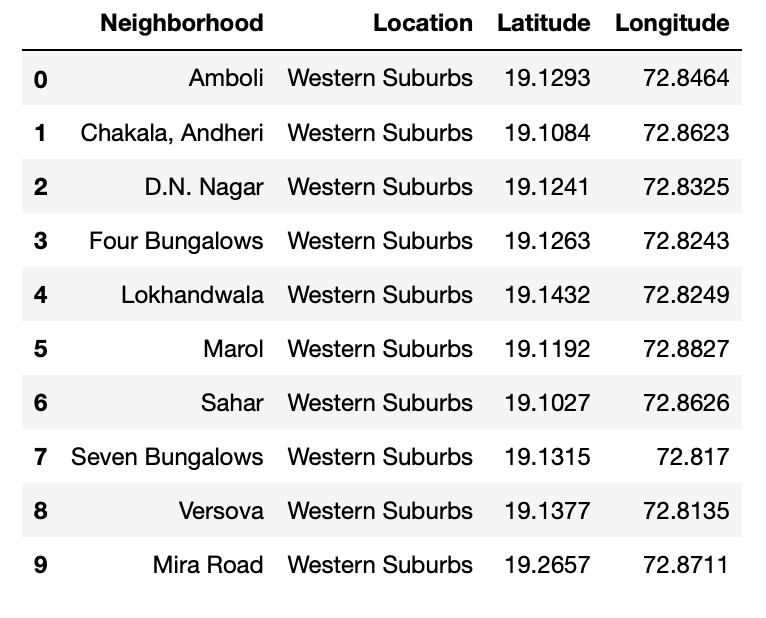


Figure 4: Final Mumbai neighborhoods dataframe.

## 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.

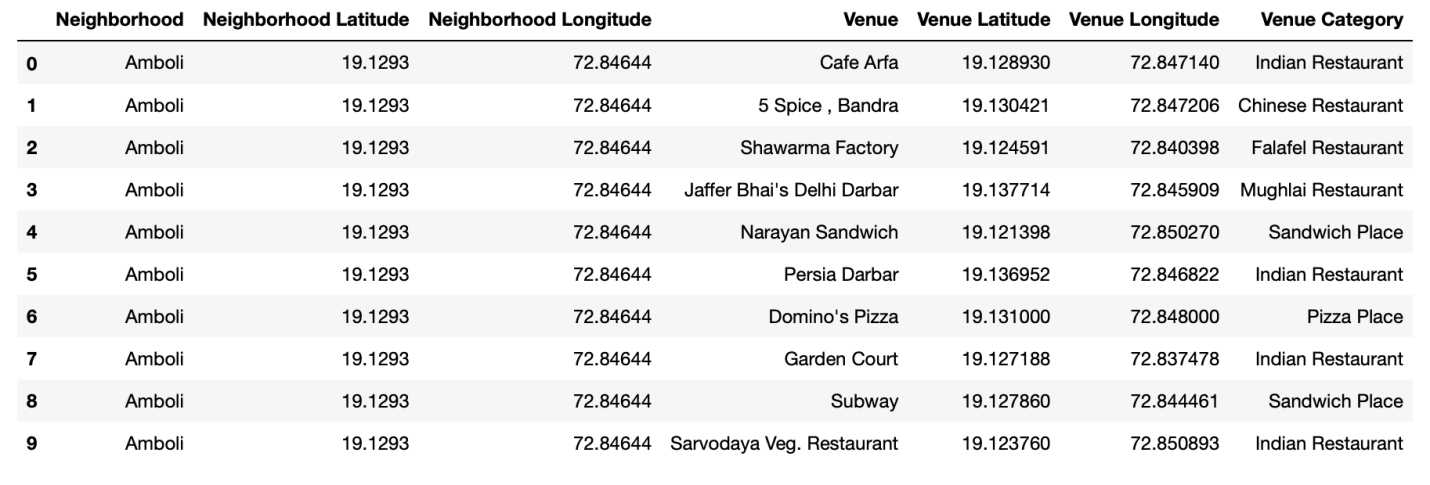


Figure 5: Data obtained from Foursquare API after cleaning.

# 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.

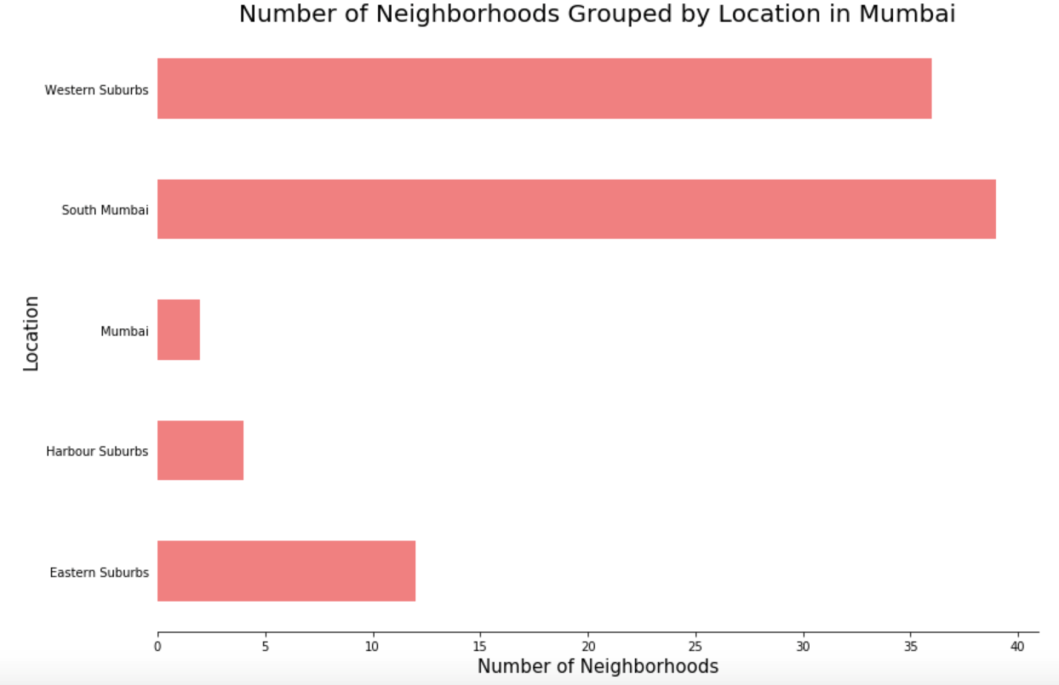


Figure 6: Number of neighborhoods grouped by location.

It is evident from Figure 6 that South Mumbai and Western Suburbs have the most number of neighborhoods. Notice how we see one of the locations as Mumbai itself? This is because the neighborhoods contained in this location are located at the outskirts of the city and thus have been termed as just Mumbai.

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

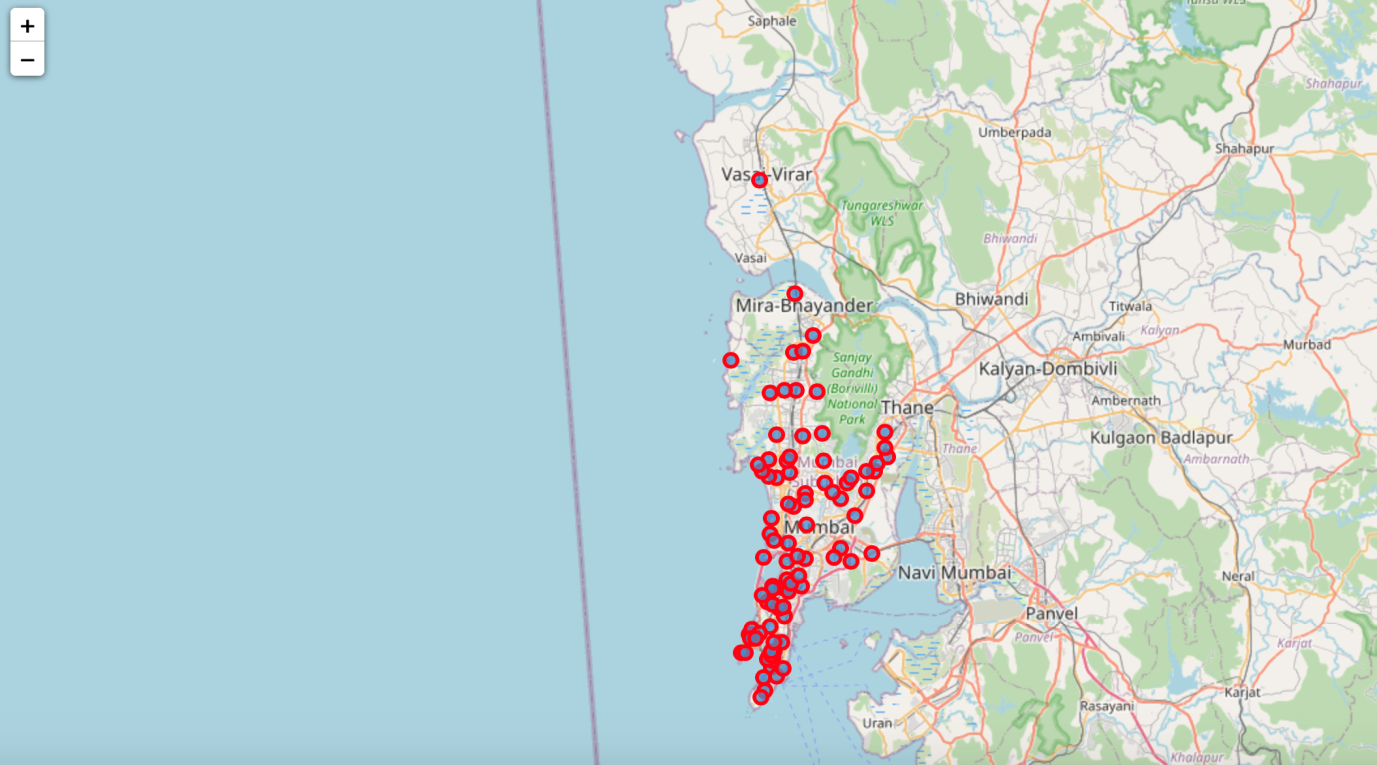


Figure 7: Depicting the neighborhood spread across Mumbai.

## 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 dataframe 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.

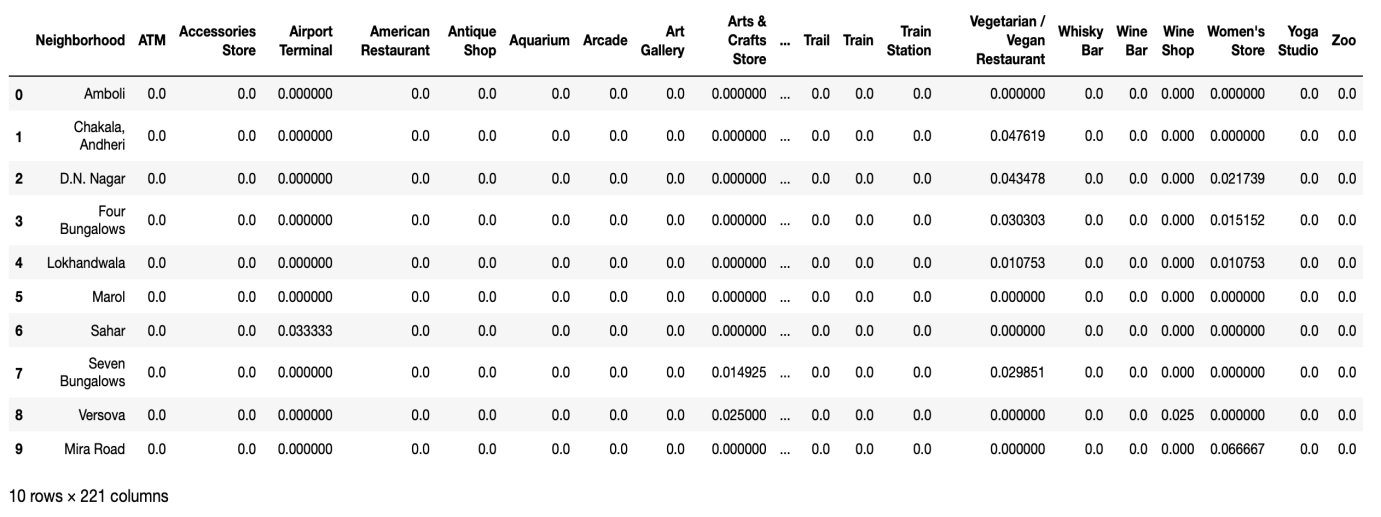


Figure 8: One-hot Encoding resulting dataframe.

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 dataframe 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 dataframe looks like as it will be used later to combine the results from the unsupervised learning model. The top 10 rows of this dataframe 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 neighborhoods. 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.

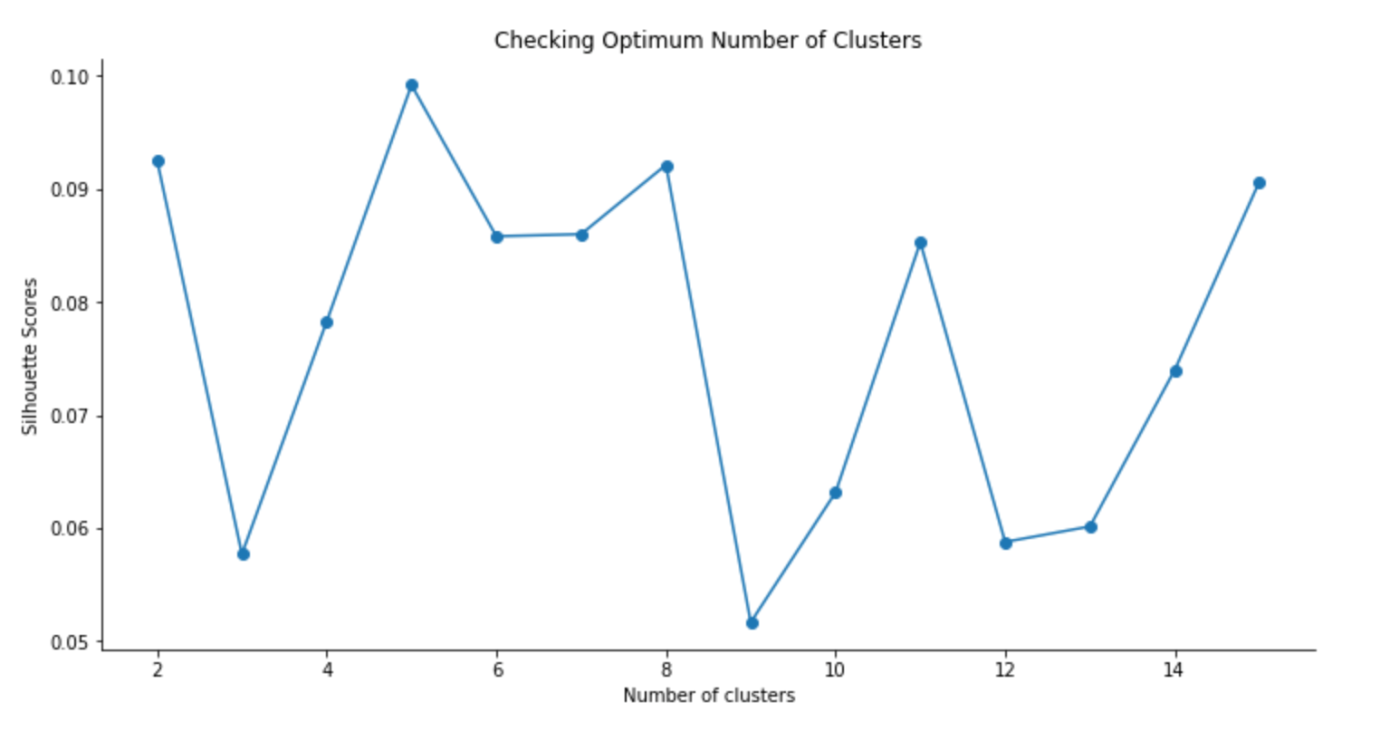


Figure 10: Silhouette scores for different number of clusters.

It is evident that the Silhouette scores are not very high even as the number of clusters increases. This means that the inter-cluster distance is not very high over the range of k-values. Despite this, the data will be clustered to the best possible extent. For this, 5 clusters will be used for the k-means clustering model since it provides the highest silhouette score as seen in Figure 10.

# 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 dataframe 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.

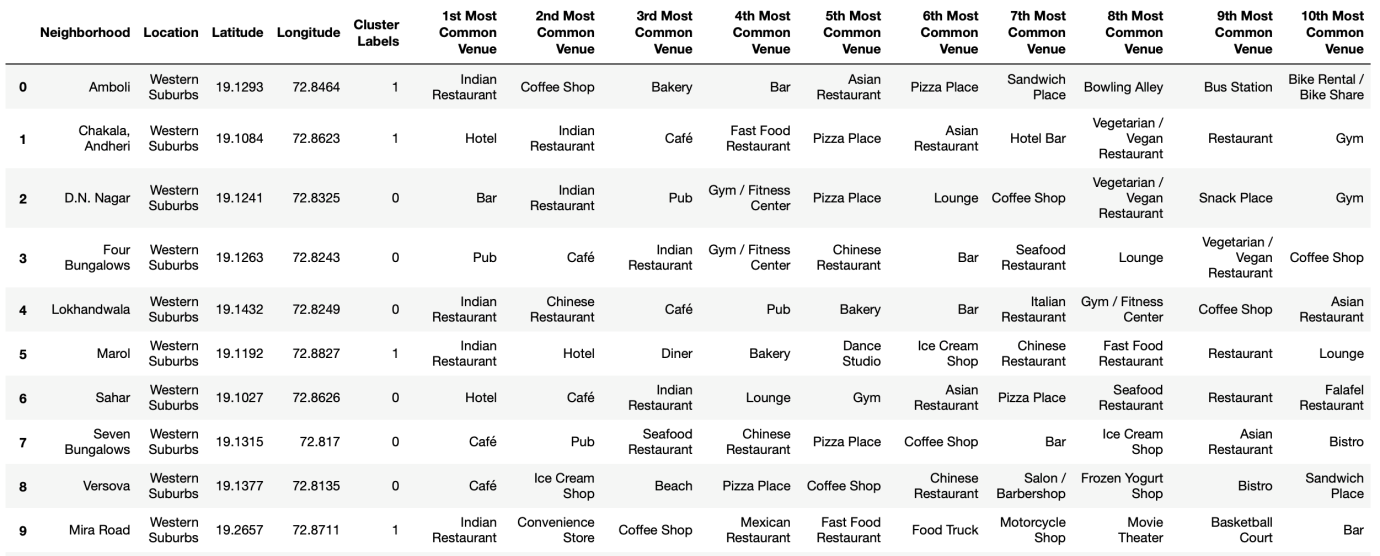


Figure 11: Clustering neighborhoods in Mumbai.

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.

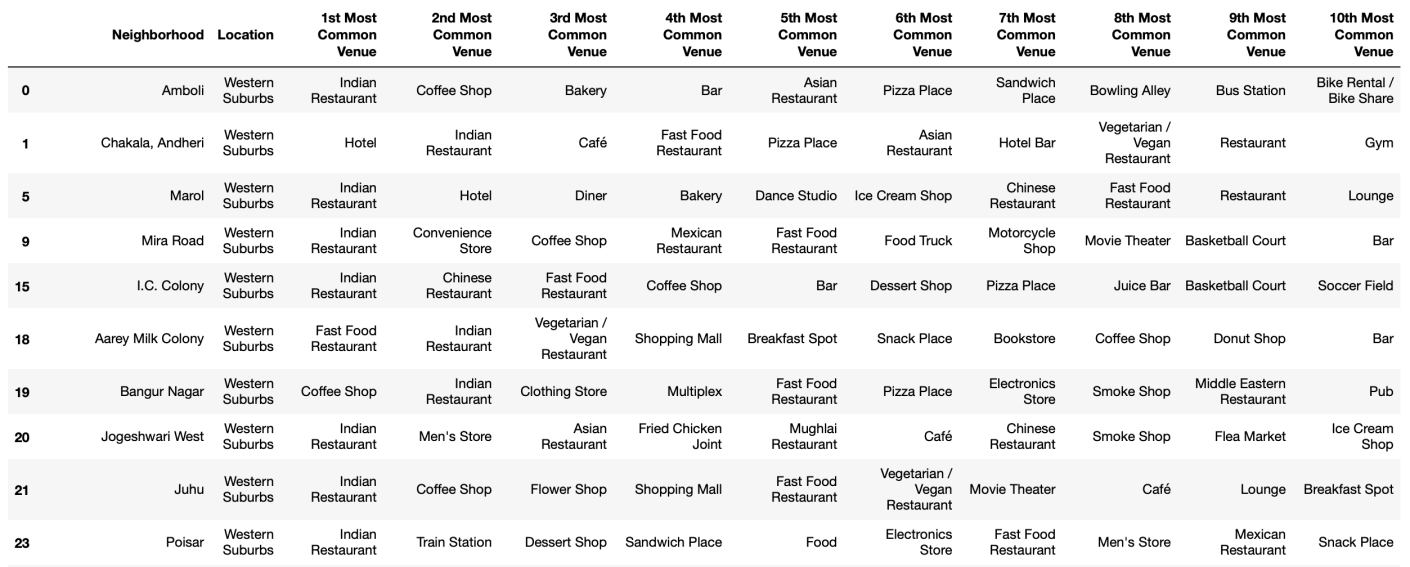


Figure 13: Cluster 2.



Figure 14: Cluster 3.

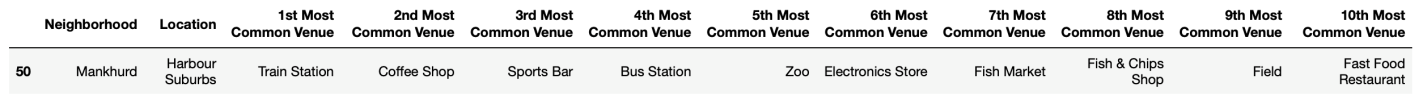


Figure 15: Cluster 4.

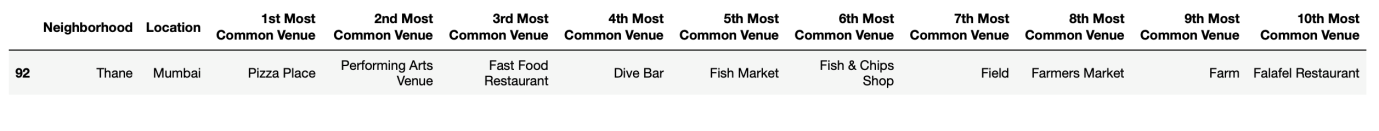


Figure 16: Cluster 5.

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

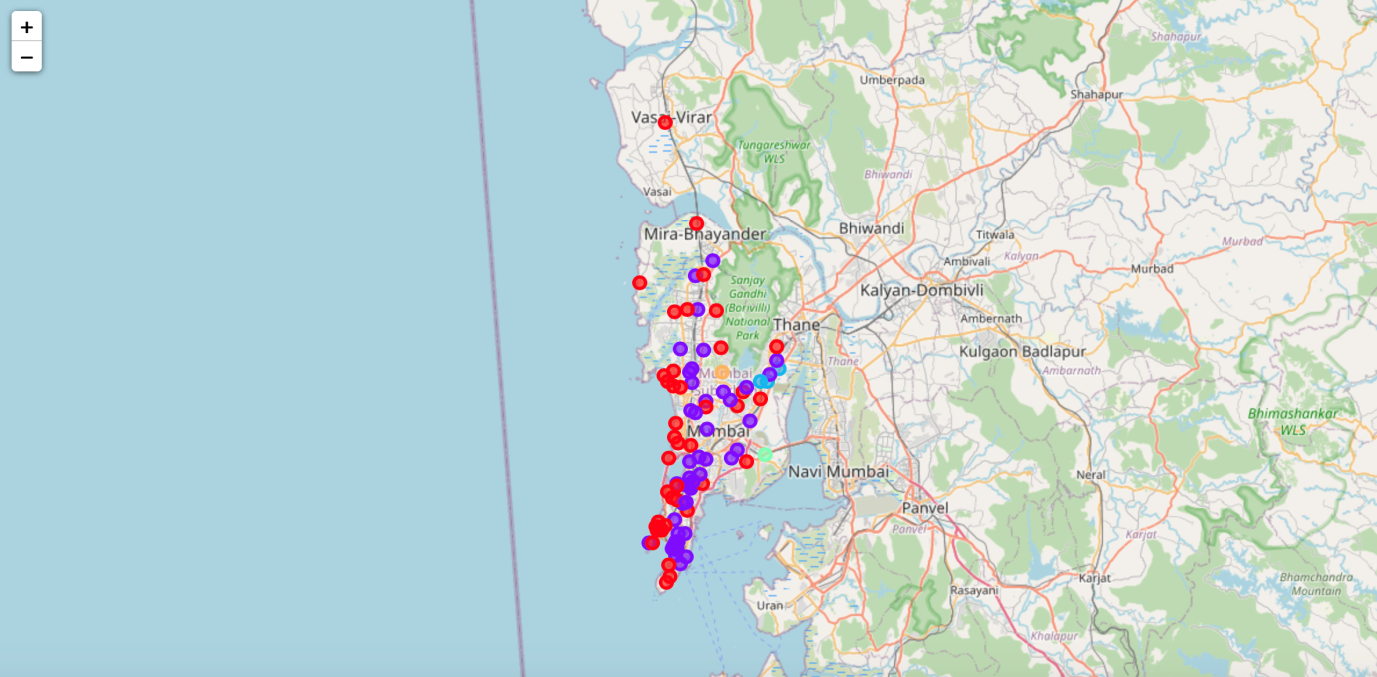


Figure 17: Visualizing the clustering of neighborhoods in Mumbai.

# Discussion

By analyzing the five clusters obtained we can see that some of the clusters are more suited for restaurants and hotels, whereas, other clusters are less suited. Neighborhoods in clusters 3, 4, and 5 contain a small percentage of restaurants, hotels, cafe and pubs in their top 10 common venues. These clusters contain a higher degree of other venues like train station, bus station, fish market, gym, performing arts venue and smoke shop, to name a few. Thus, they are not well suited for opening a new restaurant. On the other hand, neighborhoods in clusters 1 and 2 contain a much higher degree of restaurants, hotels, multiplex, cafes, bars and other food joints. Thus, the neighborhoods in these clusters would be well suited for opening a new restaurant.

Comparing clusters 1 and 2, neighborhoods in cluster 1 seem to be more suited for starting a restaurant since they contains a larger percentage of food joints in the top 10 most common venues than cluster 2. The neighborhoods in cluster 1 contain a variety of food joints like restaurants, tea rooms, bakery, cafe, steakhouse and pubs and also contain very diverse cuisines like Japanese, Indian, Chinese, Italian and seafood restaurants. Most neighborhoods in cluster 2 seem to have Indian Restaurant as their top most common venue; however, on careful analysis we can see that neighborhoods in cluster 2 also contain other venues like soccer field, flea market, smoke shop, gym, train station, dance studio, music store, cosmetics shop and so on. Thus, it is recommended that the new restaurant can be opened in the neighborhoods belonging to cluster 1. This neighborhood can be further plotted on a map as shown below in Figure 18.

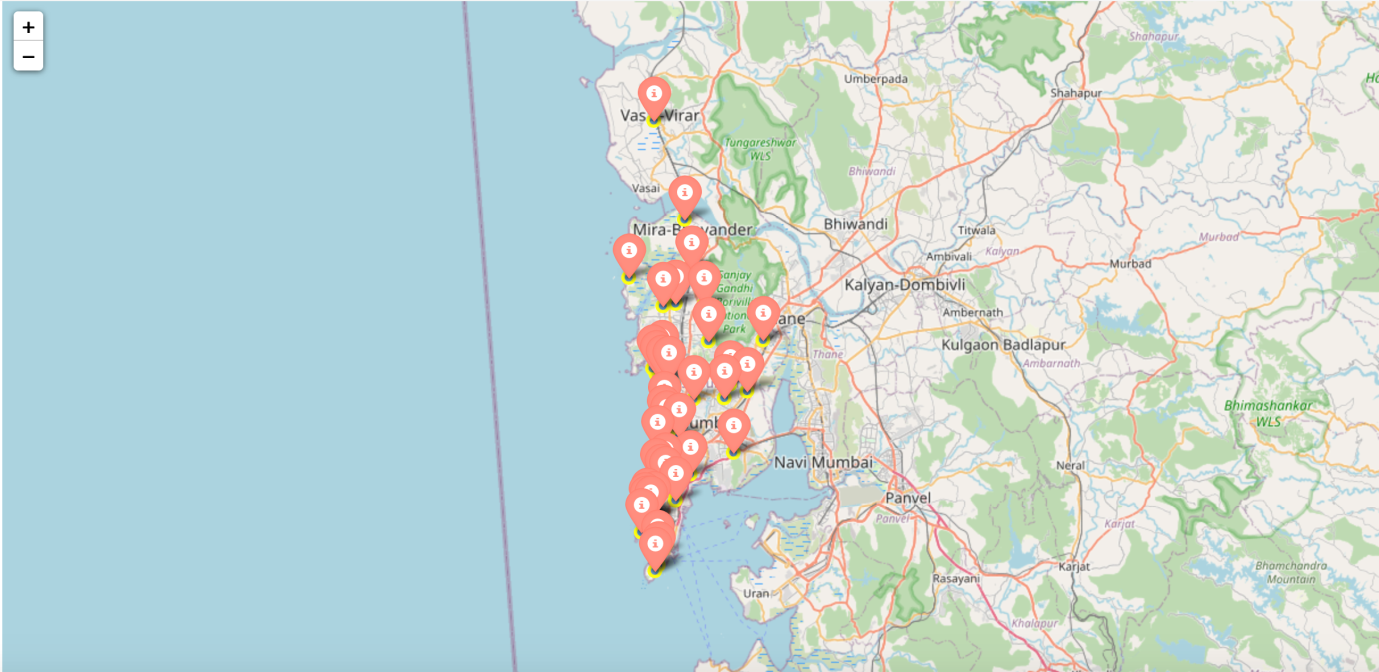


Figure 18: Neighborhoods most suited for starting a new restaurant.

# Conclusion

In this project, the neighborhoods in Mumbai, India have been successfully analyzed for determining which would be the best neighborhoods for opening a new restaurant. Based on the analysis carried out, neighborhoods in cluster 1 are recommended as locations for the new restaurant. This has also been plotted in the map in Figure 18. 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.

# Final Comments

**Note 1:** In order to view the code for this project, kindly refer to the notebook on the github repository at: <https://github.com/raunakbhutoria/Coursera_Capstone>