Capstone Project - Battle of the Neighbourhoods in the city of Mumbai, India

Table of contents:

Introduction: Business Problem Statement

In this project we will try to find an optimal location for opening a restaurant. Specifically, this report will be targeted to stakeholders interested in opening any food-joint/restaurant in Mumbai, Maharashtra, India.

Since there are lots of restaurants in Mumbai, we will try to detect locations that are not already crowded with restaurants. We are also particularly interested in neighbourhoods which are similar to the best neighbourhoods in terms of amenities. We would also prefer locations as close to city centre as possible, assuming that first two conditions are met.

We will use our data science technical expertise to generate a few most promising neighbourhoods based on these criteria. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants in the neighbourhood (any type of restaurant)
- variety of amenities in the neighbourhood, if any
- distance of neighbourhood from city centre

Neighbourhoods have been defined based on names of post offices as given in the form of a table in the website: http://pincode.india-server.com/cities/mumbai/

Following data sources will be needed to extract/generate the required information:

- names of all neighbourhoods (post office names) will be extracted from the above-mentioned website using Pandas package in python
- approximate addresses of centres of these neighbourhoods will be obtained using GeoPy Geocoder package in python
- number of restaurants and their type and location in every neighbourhood will be obtained using Foursquare API

Data Cleaning:

The table present in the website was procured using the Pandas ‘read_html()’ method:

List of all Mumbai post offices with pincode

	S.No.	Post office	Office type	Pincode
0	1.0	A I Staff Colony	S.O	400029
1	2.0	Aareymilk Colony	S.O	400065
2	3.0	Agripada	S.O	400011
3	4.0	Airport	S.O	400099
4	5.0	Ambewadi	S.O	400004
...	...	...	...	...
234	235.0	Worli Colony	S.O	400030
235	236.0	Worli Naka	S.O	400018
236	237.0	Worli Police Camp	S.O	400030
237	238.0	Worli	S.O	400018
238	239.0	Worli Sea Face	S.O	400030

239 rows × 4 columns

DataFrame gathered from website

The DataFrame was then cleaned to only include names of the post offices which are the names of the neighbourhoods:

List of all Mumbai post offices with pincode

	Post office
0	A I Staff Colony
1	Aareymilk Colony
2	Agripada
3	Airport
4	Ambewadi
...	...
234	Worli Colony
235	Worli Naka
236	Worli Police Camp
237	Worli
238	Worli Sea Face

239 rows × 1 columns

DataFrame after dropping unwanted columns

The names of neighbours were then extracted from the above DataFrame and a new dataframe was created consisting of the neighbourhood names and their latitude and longitude.

	name	latitude	longitude
0	Agripada	18.975302	72.824898
1	Airport	21.086220	79.063768
2	Ambewadi	16.715835	74.204431
3	Andheri East	19.115883	72.854202
4	Andheri	19.119698	72.846420
...	...	...	...
167	Wadala	19.026919	72.875934
168	Wadala Truck Terminal	19.034708	72.876617
169	Worli Colony	19.022352	72.832755
170	Worli	19.011696	72.818070
171	Worli Sea Face	19.005470	72.813674

172 rows × 3 columns

Methodology

In the cells below the following format of the project will be followed in order to explore the neighbourhoods in Mumbai and try to analyse and suggest an optimal location for opening a food-joint/restaurant.

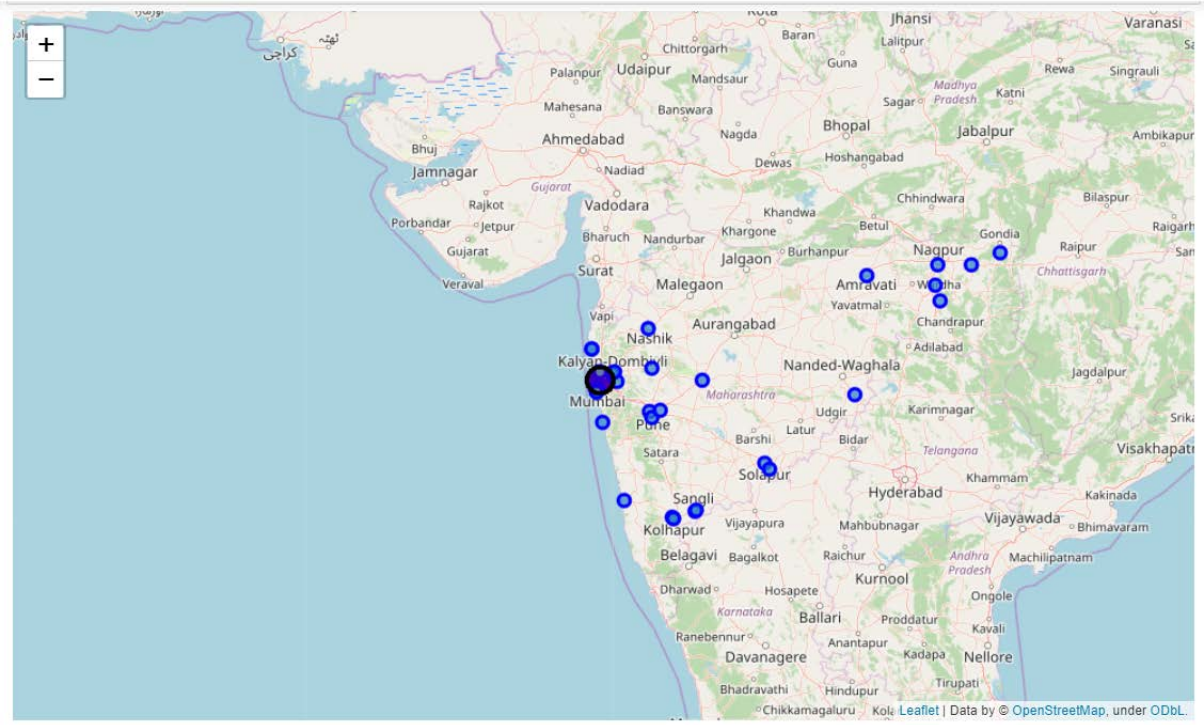
1. Get list of all neighbourhoods from the website listing names of postal offices all over Mumbai.
2. Obtain latitudes and longitudes for all these neighbourhoods.
3. Check for any outliers, i.e. postal office names which aren't within 25Kms from the centre of Mumbai and remove these data points.
4. Send GET requests to Foursquare API to get list of venues (maximum 100) within the vicinity of all neighbours (vicinity is defined within a radius of 1000m from the neighbourhood).
5. Remove all neighbourhoods which do not have more than 20 venues in their vicinities since there isn't enough data for accurate clustering using K-Means Clustering.
6. One-hot encode the data and feed the dataframe to K-Means clustering Algorithm to form 10 clusters from the list of available neighbourhoods.
7. Create visualizations using the cluster number of all neighbourhoods to better understand the data create valuable insights. Here, we try to find out the best neighbourhoods amongst the available list of neighbourhoods and try to identify it's cluster.
8. Again, send GET requests to Foursquare API to get list of venues (maximum 100) of 'section = food' to obtain only venues related to food joints within the vicinity of

all selected neighbourhoods (vicinity is defined within a radius of 1000m from the neighbourhood).

9. Create visualizations to find out which neighbourhoods have maximum food joints and which ones have maximum types of food joints.
10. Suggest optimal locations for opening a restaurant based on all the statistical analysis done.

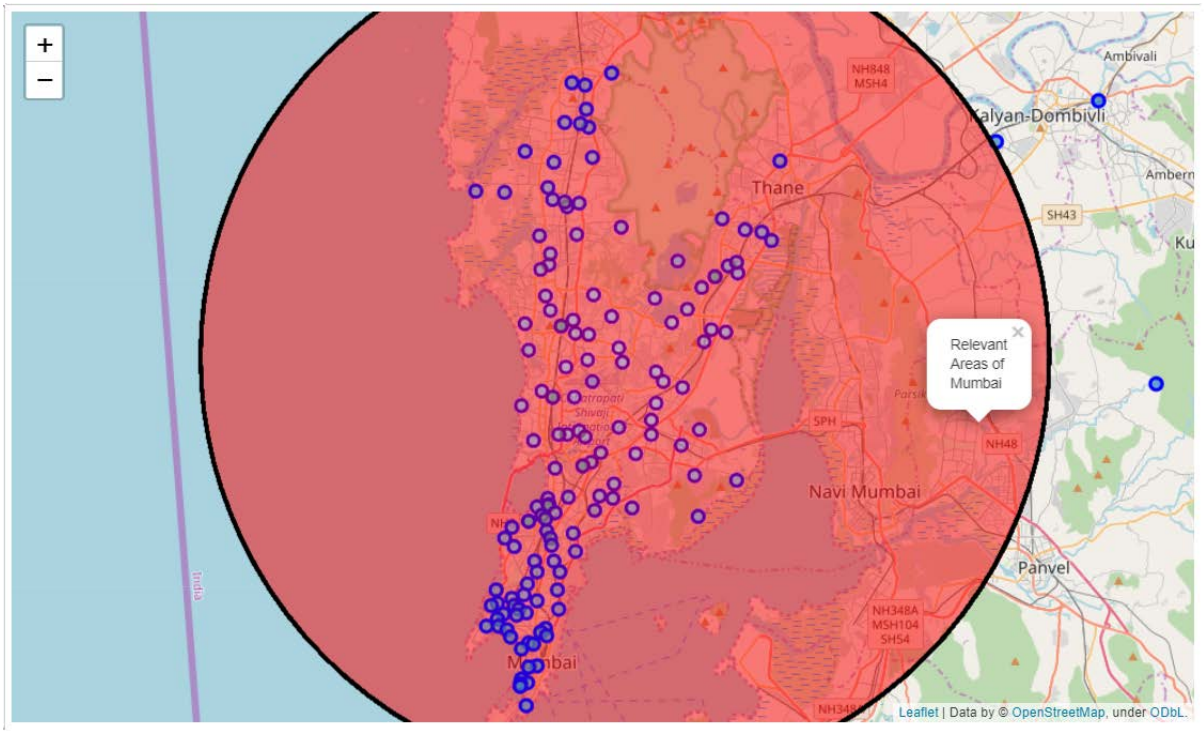
Data Visualization

The obtained latitude and longitude values for all the neighbourhoods are then used to plot all postal offices/neighbourhoods in the map of Maharashtra. It is observed that a lot of these are far away from the centre of the city. The circle with a 25km radius from the centre of the city has been marked for reference.



Map with all postal offices

The previous map has been appropriately zoomed to highlight the relevant neighbourhoods in Mumbai.



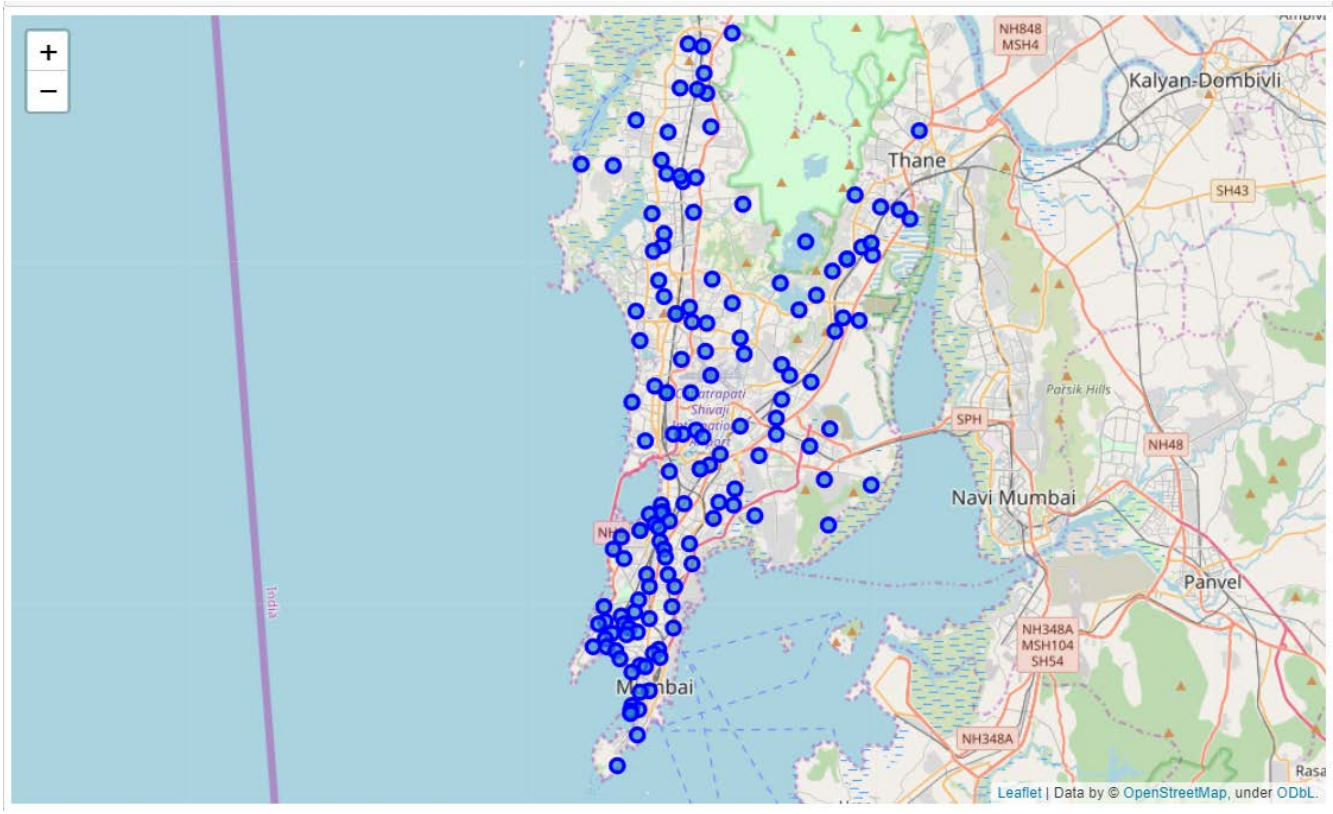
Thus, the dataframe is shortened to only those neighbourhoods which are within 25Kms from the city centre and the final dataframe is as follows:

	name	latitude	longitude	dist_from_centre
0	Sakinaka	19.100090	72.881478	0.379917
1	Marol Naka	19.108156	72.879478	0.628107
2	Barve Nagar	19.095283	72.900178	2.050079
3	B P T Colony	19.101937	72.861599	2.219168
4	International Airport	19.090201	72.863808	2.457877
...	...	...	...	...
140	Mantralaya	18.927662	72.827039	20.308105
141	V K Bhavan	18.926926	72.823050	20.510702
142	Nariman Point	18.925951	72.823208	20.608386
143	Colaba	18.915091	72.825969	21.675108
144	Asvini	18.900689	72.816134	23.499370

145 rows × 4 columns

We can see that after filtering out the 239 initial neighbourhoods, we now have 145 relevant neighbourhoods which can possibly be used to suggest an optimal location for a restaurant.

The map of Mumbai with relevant neighbourhoods is as follows:



The dataframe created after getting the response from the Foursquare API as nearest venues within 1000m from a neighbourhood is as follows:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category	Venue Latitude	Venue Longitude
0	Sakinaka	19.100090	72.881478	JW Marriott Mumbai Sahar	Hotel	19.102502	72.878236
1	Sakinaka	19.100090	72.881478	J W Cafe	Restaurant	19.103212	72.877504
2	Sakinaka	19.100090	72.881478	Starbucks: A Tata Alliance	Coffee Shop	19.099318	72.874339
3	Sakinaka	19.100090	72.881478	The Bar Stock Exchange	Pub	19.105542	72.884159
4	Sakinaka	19.100090	72.881478	Romano's	Italian Restaurant	19.103115	72.877408
...	...	...	...	...	...	...	...
5573	Colaba	18.915091	72.825969	Bentley's Hotel	Hotel	18.919641	72.830861
5574	Colaba	18.915091	72.825969	Di Bella	Coffee Shop	18.918652	72.830413
5575	Colaba	18.915091	72.825969	Village	Indian Restaurant	18.914541	72.818211
5576	Colaba	18.915091	72.825969	Garage Inc. Public House	Cocktail Bar	18.919511	72.830323
5577	Colaba	18.915091	72.825969	Wodehouse Gym Tennis Court	Tennis Court	18.922260	72.827566

5200 rows × 7 columns

The above table is grouped by neighbourhood names and the count of every column is taken to find out the number of venues present in every neighbourhood. Neighbourhoods which have less than 20 venues are discarded since there isn't enough data to cluster these neighbourhoods properly.

The resulting table further filters the 145 possible neighbourhoods obtained above into 103 neighbourhoods and these are the final neighbourhoods which shall be grouped into clusters.

The resulting table is one-hot encoded in order to create all unique ‘Venue Category’ values into columns with a value of 1 or 0. 1 meaning that it falls under that category.

The table created after one-hot encoding is as follows:

```
mumbai_onehot.head()
```

	Neighborhood_Name	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	...	Train	Train Station	Travel & Transport	Vegetarian / Vegan Restaurant	Whisky Bar
0	Agripada	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
1	Agripada	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	Agripada	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	Agripada	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	Agripada	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 234 columns

```
mumbai_onehot.shape
```

```
(5200, 234)
```

This dataframe is then grouped by the neighbourhood name again and the mean of every column is taken as the values for every row after grouping. The resulting table is as follows:

	Neighborhood_Name	Airport	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	...	Train	Train Station	Travel & Transport	Vegetarian / Vegan Restaurant
0	Agripada	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.027778	0.0	0.000000
1	Andheri	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.000000	0.0	0.034483
2	Andheri East	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.000000	0.0	0.050000
3	Andheri Railway Station	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.000000	0.0	0.034483
4	Azad Nagar	0.0	0.0	0.0	0.0	0.029412	0.0	0.0	0.000000	0.000000	...	0.0	0.000000	0.0	0.029412
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
98	Tulsiwadi	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.020408	0.0	0.020408
99	V K Bhavan	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.010000	...	0.0	0.000000	0.0	0.010000
100	Worli	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.045455	0.045455	...	0.0	0.000000	0.0	0.000000
101	Worli Colony	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.017241	0.000000	...	0.0	0.000000	0.0	0.000000
102	Worli Sea Face	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.050000	...	0.0	0.000000	0.0	0.000000

103 rows × 234 columns

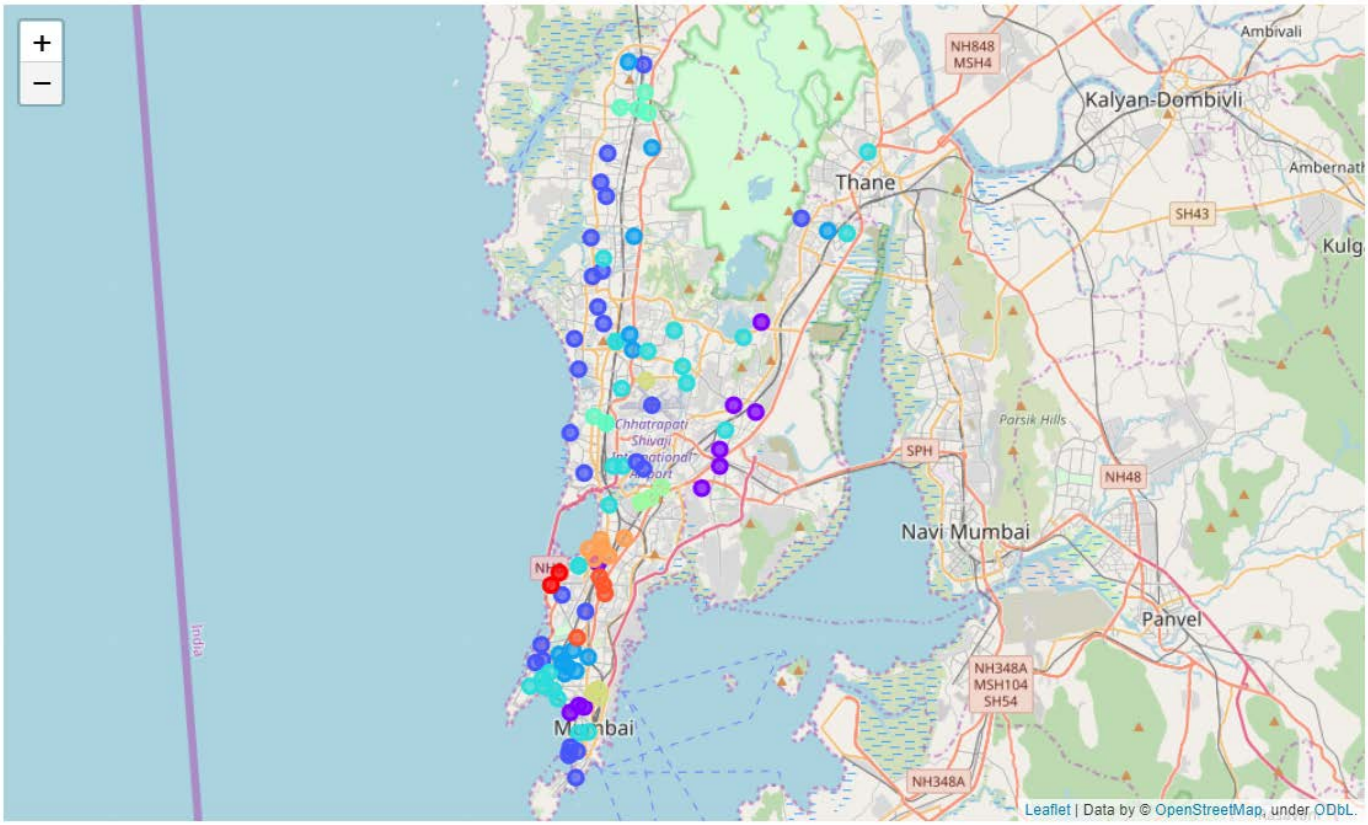
The above dataframe is fed to the K-Means clustering algorithm after dropping the ‘Neighborhood_Name’ column to create clusters amongst these 103 neighbourhoods.

A dataframe is created which consists of the names of the neighbourhoods, their latitudes and longitudes as well as their assigned cluster number. The dataframe is as follows:

	name	latitude	longitude	dist_from_centre	Cluster_Number
0	Agripada	18.975302	72.824898	15.421206	3
1	Andheri	19.119698	72.846420	4.218935	4
2	Andheri East	19.115883	72.854202	3.298365	3
3	Andheri Railway Station	19.119698	72.846420	4.218935	4
4	Azad Nagar	19.128315	72.840038	5.266106	2
...	...	...	...	...	...
98	Tulsiwadi	18.973423	72.818156	15.902341	3
99	V K Bhavan	18.926926	72.823050	20.510702	2
100	Worli	19.011696	72.818070	12.210406	0
101	Worli Colony	19.022352	72.832755	10.389105	8
102	Worli Sea Face	19.005470	72.813674	13.040575	0

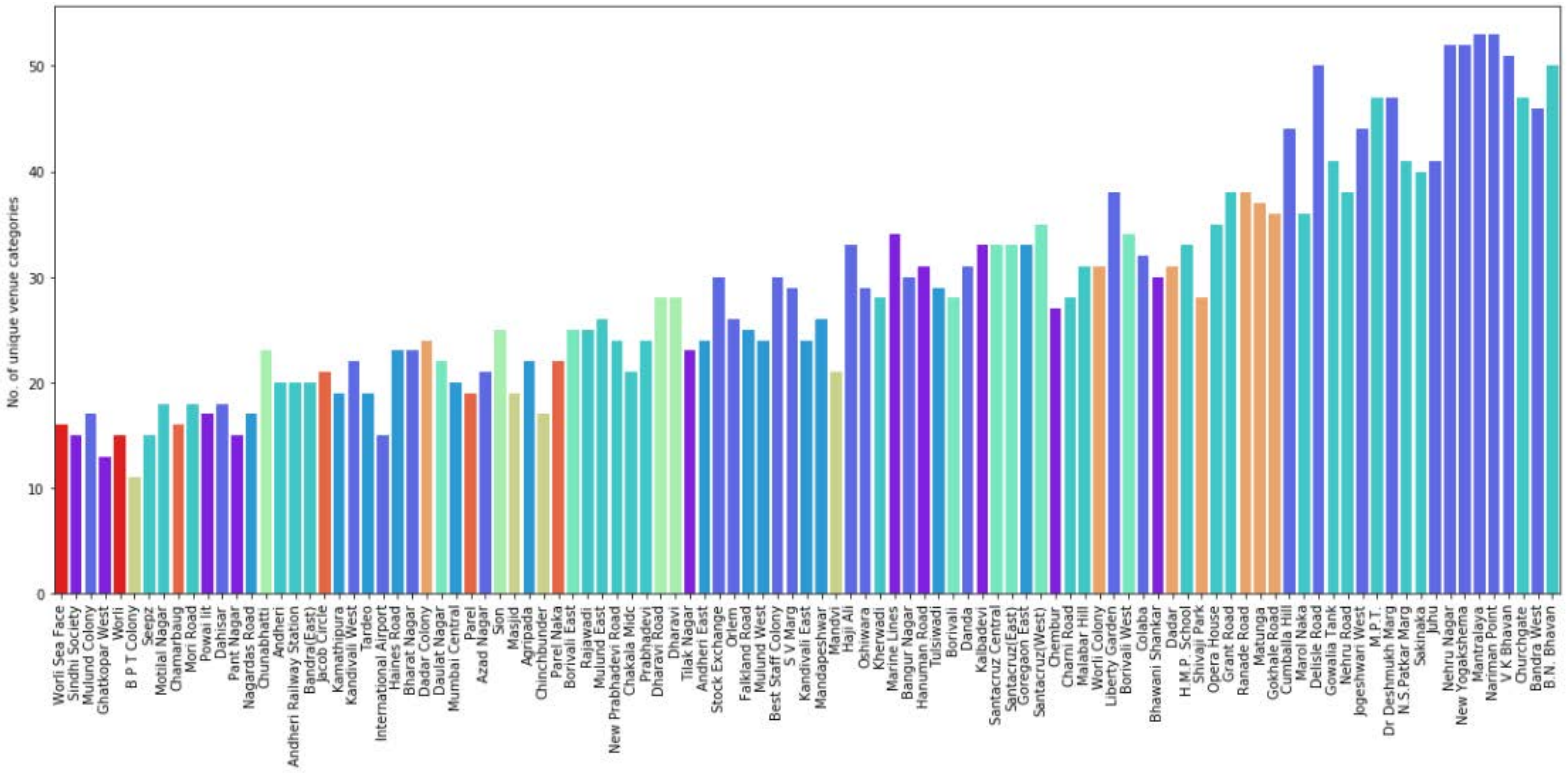
103 rows × 5 columns

The map of Mumbai is plotted with circle markers for every neighbourhood which are colour coded according to the cluster number of every neighbourhood. The map is as follows:



Analysis

A bar graph was plotted using the number of unique venue categories along the Y-axis with the corresponding neighbourhood name on the X-axis. This is done to estimate the quality of a neighbourhood based on the assumption that the best neighbourhood will have the highest diversity of amenities which in this case is the number of unique venue categories. The bar plot has been colour coded such that all bars of the same colour correspond to neighbourhoods in the same cluster. The bar graph is as follows:



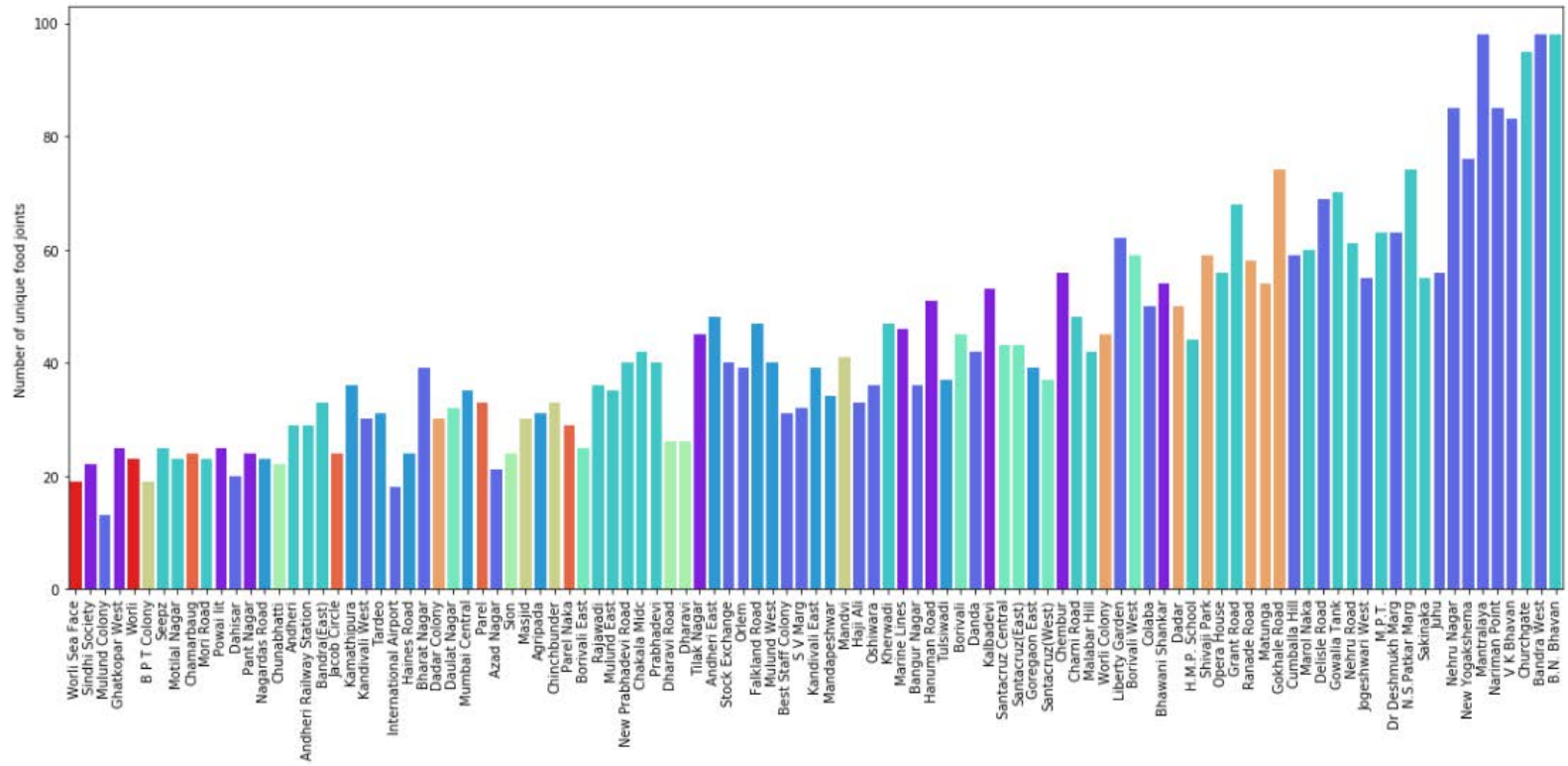
From the above plot, we can conclude that clusters pertaining to Nariman Point and B.N. Bhavan, i.e. cluster number 2 and 4 respectively can definitely be considered the best clusters since they consist of neighbourhoods with the highest diversity of amenities.

Now, in order to judge neighbourhoods on the basis of food-specific venues, the Foursquare API is again used to request for venues pertaining to foods, for all the final candidate neighbourhoods and resulting dataframe is as follows:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agripada	18.975302	72.824898	Celejor	18.975844	72.823679	Bakery
1	Agripada	18.975302	72.824898	Neel	18.980407	72.820403	Indian Restaurant
2	Agripada	18.975302	72.824898	Sigdi Restaurant	18.970523	72.831034	Indian Restaurant
3	Agripada	18.975302	72.824898	Persian Darbar	18.976055	72.833643	Indian Restaurant
4	Agripada	18.975302	72.824898	Gloria Restaurant	18.975485	72.833884	Asian Restaurant
...	...	...	...	...	...	...	...
4618	Worli Sea Face	19.005470	72.813674	Tastee	18.999275	72.816955	Fast Food Restaurant
4619	Worli Sea Face	19.005470	72.813674	Fishland	18.998818	72.817330	Seafood Restaurant
4620	Worli Sea Face	19.005470	72.813674	Banjara	19.006259	72.821605	Restaurant
4621	Worli Sea Face	19.005470	72.813674	Mahindra Holidays Cafeteria	19.005692	72.822309	Snack Place
4622	Worli Sea Face	19.005470	72.813674	Poh	19.005826	72.823116	Chinese Restaurant

4623 rows × 7 columns

Now, the bar graph is plotted with the total number of food-joints/restaurants in a neighbourhood on the Y-axis with the corresponding neighbourhood name on the X-axis. This is done in order to estimate the saturation of a neighbourhood with respect to the number of food-joints present in the neighbourhood since we want to detect locations that are not already crowded with restaurants. The bars are color-coded according to the cluster number here as well. The bar graph is as follows:



Results and Discussion

Based on the bar graph showing the number of unique venue categories for each neighbourhood, we concluded that clusters pertaining to Nariman Point and B.N. Bhavan, i.e. cluster numbers 2 and 4 respectively can definitely be considered the best clusters since they consist of neighbourhoods with the highest variety of amenities.

Thus, prospective stakeholders would like to open up their restaurant in similar neighbourhoods since they have the largest customer base and also have a huge demand. However, it is possible that these neighbourhoods might be saturated with venues. Therefore, we also look at the plot showing the number of unique venues for each and every neighbourhood. Here we see that the while the best neighbourhoods in terms of diversity of amenities such as Churchgate, Nariman Point, Bandra West B.N. Bhavan etc. are saturated with a large number of venues, there exist neighbourhoods which are part of the same cluster yet have lesser number of venues. These neighbourhoods might be optimal for opening up restaurants since they fall to the same cluster and are therefore similar to neighbourhoods with highly diverse amenities yet have a smaller number of venues.

Instead of looking at the total number of venues of a certain neighbourhood to determine the saturation, it would be better to simply look at the total number of food joints / restaurants in that neighbourhood. Thus, we look at the bar graph showing the number of venue categories pertaining to food for each neighbourhood. We can see that our initial assumption is still correct and neighbourhoods in cluster 2 and 4 have the highest variety of food joints are thereby considered the best neighbourhoods.

So, the optimal location to open up restaurants in Mumbai would be neighbourhoods in the **best clusters** i.e. clusters 2 and 4 with the **least number of existing food joints**. The ideal candidate neighbourhoods can be seen from the plot of number of unique venues related to food for every neighbourhood. Here we see that the neighbourhoods with the least number of food joints belonging to cluster 2 and 4 are:

1. Mulund Colony
2. Dahisar
3. Motilal Nagar
4. Mori Road
5. International Airport

Conclusion

The purpose of this project was to identify areas/neighbourhoods in Mumbai close to the city-centre with low number of restaurants in order to aid stakeholders in narrowing down the search for optimal location for opening of a food-joint/restaurant. By calculating venue density distribution from Foursquare data we have first identified general clusters that justify further analysis and created groups of similar neighbourhoods, and then generated extensive visualizations to justify how good a certain cluster is based on characteristics (diversity of amenities) of the neighbourhoods in the clusters.

Then the Foursquare data was again used in order to find venues pertaining to food for all the neighbourhoods to determine the saturation of a neighbourhood with respect to number of food joints. **Optimal locations were identified as those neighbourhoods which are part of the best clusters (having neighbourhoods with most diverse amenities) yet have the least number of food joints.** The optimal locations obtained are:

1. Mulund Colony
2. Dahisar
3. Motilal Nagar
4. Mori Road
5. International Airport

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighbourhoods and their locations, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighbourhood etc.