Battle of Neighborhoods in Mumbai for opening a Shopping Mall

Applied Data Science Capstone Project

By: Saurabh Mahadane

Date: June 6, 2021

Table of Contents

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 Mumbai generally love food & shopping. Thus, the aim of this project is to study the neighborhoods in Mumbai to determine possible locations for starting a Shopping Mall. This project can be useful for business owners, Builders and entrepreneurs who are looking to invest and open a Shopping Mall in Mumbai. The main objective of this project is to carefully analyse 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. 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.

	Neighborhood	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270
5	Marol	Andheri,Western Suburbs	19.119219	72.882743
6	Sahar	Andheri,Western Suburbs	19.098889	72.867222
7	Seven Bungalows	Andheri,Western Suburbs	19.129052	72.817018
8	Versova	Andheri,Western Suburbs	19.120000	72.820000
9	Mira Road	Mira-Bhayandar,Western Suburbs	19.284167	72.871111

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.

```
address = 'Mumbai, IN'
geolocator = Nominatim()
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinates of Mumbai are {}, {}.'.format(latitude, longitude))
The geograpical coordinates of Mumbai are 19.0759899, 72.8773928.
```

Figure 2: 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.

	Neighborhood	Location	Latitude	Longitude	Latitude1	Longitude1	Latdiff	Longdiff
0	Amboli	Western Suburbs	19.129300	72.843400	19.1291	72.8464	0.00024	0.00304
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833	19.1084	72.8623	0.003028	0.001497
2	D.N. Nagar	Western Suburbs	19.124085	72.831373	19.1251	72.8325	0.000965	0.001107
3	Four Bungalows	Western Suburbs	19.124714	72.827210	19.1264	72.8242	0.001666	0.00301
4	Lokhandwala	Western Suburbs	19.130815	72.829270	19.1432	72.8249	0.012345	0.0044
5	Marol	Western Suburbs	19.119219	72.882743	19.1191	72.8828	0.000169	6.7e-05
6	Sahar	Western Suburbs	19.098889	72.867222	19.1027	72.8626	0.00377822	0.00462255
7	Seven Bungalows	Western Suburbs	19.129052	72.817018	19.1286	72.8212	0.000492	0.004162
8	Versova	Western Suburbs	19.120000	72.820000	19.1377	72.8135	0.01769	0.00652
9	Mira Road	Western Suburbs	19.284167	72.871111	19.2656	72.8706	0.0185438	0.000467611

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.

	Neighborhood	Location	Latitude	Longitude
0	Amboli	Western Suburbs	19.1293	72.8464
1	Chakala, Andheri	Western Suburbs	19.1084	72.8623
2	D.N. Nagar	Western Suburbs	19.1241	72.8325
3	Four Bungalows	Western Suburbs	19.1264	72.8242
4	Lokhandwala	Western Suburbs	19.1432	72.8249
5	Marol	Western Suburbs	19.1192	72.8827
6	Sahar	Western Suburbs	19.1027	72.8626
7	Seven Bungalows	Western Suburbs	19.1291	72.8212
8	Versova	Western Suburbs	19.1377	72.8135
9	Mira Road	Western Suburbs	19.2656	72.8711

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 100, that is, maximum of 100 venue recommendations per neighborhood and a radius of 2 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.

(7197, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amboli	19.1293	72.84644	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
1	Amboli	19.1293	72.84644	Jaffer Bhai's Delhi Darbar	19.137714	72.845909	Mughlai Restaurant
2	Amboli	19.1293	72.84644	Merwans Cake shop	19.119300	72.845418	Bakery
3	Amboli	19.1293	72.84644	Narayan Sandwich	19.121398	72.850270	Sandwich Place
4	Amboli	19.1293	72.84644	Cafe Arfa	19.128930	72.847140	Indian Restaurant
5	Amboli	19.1293	72.84644	Hard Rock Cafe Andheri	19.135995	72.835335	American Restaurant
6	Amboli	19.1293	72.84644	Pizza Express	19.131893	72.834668	Pizza Place
7	Amboli	19.1293	72.84644	Mainland China	19.140391	72.838033	Chinese Restaurant
8	Amboli	19.1293	72.84644	Doolally Taproom	19.135917	72.833094	Brewery
9	Amboli	19.1293	72.84644	Joey's Pizza	19.126762	72.830001	Pizza Place

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.

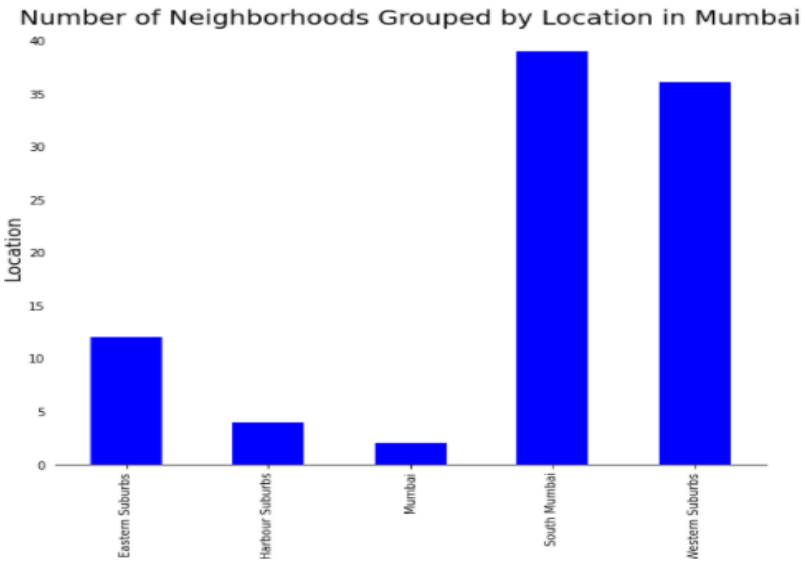

Figure 6: Number of neighborhoods grouped by location.

It is evident from Figure 6 that South Mumbai and Western Suburbs have the most number of neighborhoods.

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

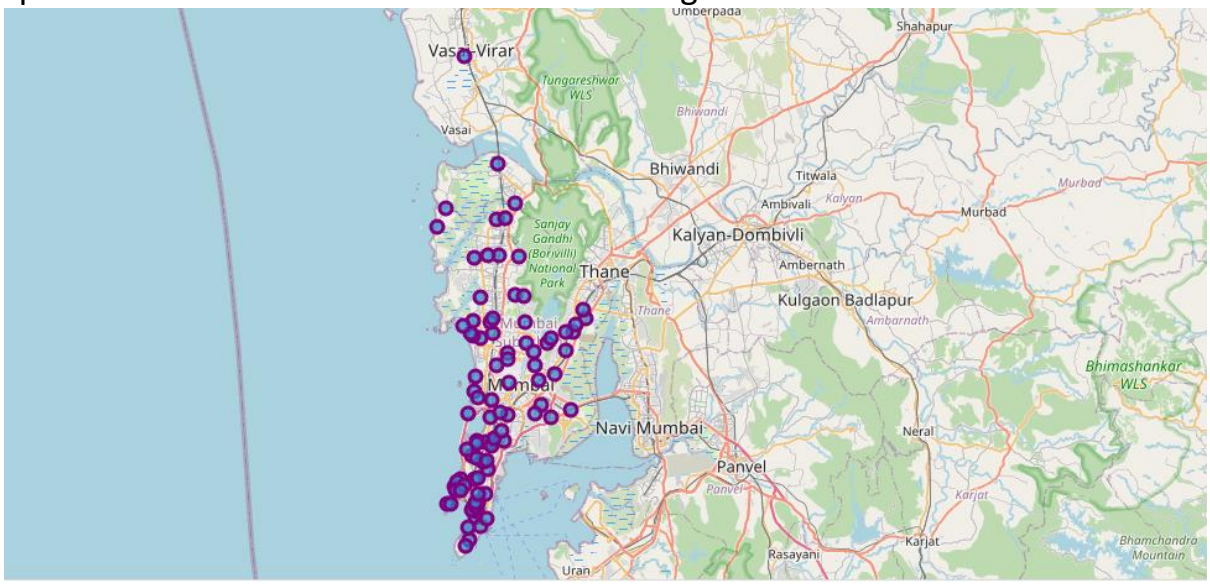

Figure 7: Depicting the neighborhood spread across Mumbai.

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.

	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Arcade	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant
0	Aarey Milk Colony	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.022222
1	Agripada	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.014925	0.000000	0.014925	0.0	0.0	0.000000
2	Altamount Road	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.010101
3	Amboli	0.013514	0.0	0.000000	0.000000	0.000000	0.000000	0.013514	0.000000	0.000000	0.000000	0.0	0.0	0.000000
4	Amrut Nagar	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.011494	0.000000	0.000000	0.000000	0.0	0.0	0.011494
5	Asalfa	0.000000	0.0	0.010526	0.010526	0.010526	0.010526	0.010526	0.000000	0.000000	0.000000	0.0	0.0	0.010526
6	Ballard Estate	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.010000	0.000000	0.000000	0.010000	0.0	0.0	0.020000
7	Bandstand Promenade	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.0	0.022222
8	Bangur Nagar	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.020000	0.000000	0.010000	0.000000	0.0	0.0	0.000000
9	Bhandup	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.045455	0.000000	0.0	0.0	0.000000

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.

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 wcss score which was calculated for a range of clusters from 1 to 11. The resulting number of clusters and their respective WCSS by Elbow method are shown in Figure 10.

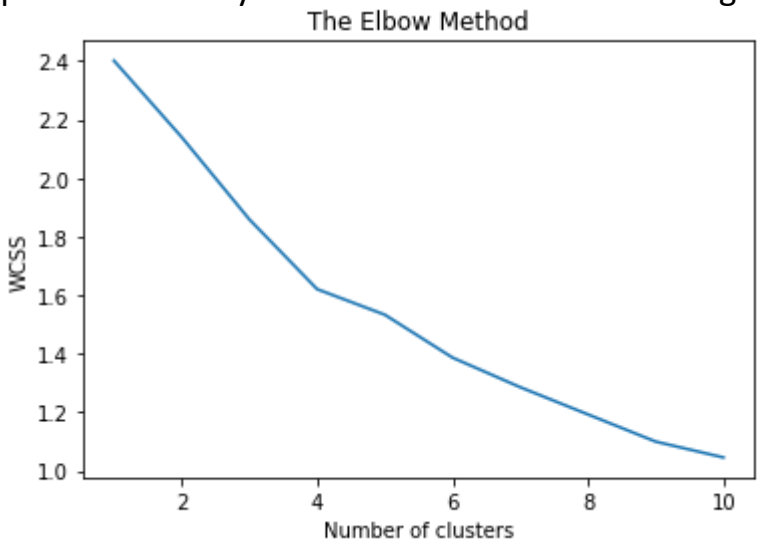

Figure 10: Elbow method for different number of clusters.

It is evident that the WCSS is 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, 4 clusters will be used for the k-means clustering model since it provides the highest WCSS score at Elbow location with in graph as seen in Figure 10.

Results

The clustering model then clusters the neighborhoods in Mumbai to filter out Shopping malls with in all 4 clusters.

(93, 6)

	Neighborhood	Shopping Mall	Cluster Labels	Location	Latitude	Longitude
0	Aarey Milk Colony	0.066667	1	Western Suburbs	19.1703	72.8711
1	Agripada	0.000000	2	South Mumbai	18.9763	72.8262
2	Altamount Road	0.000000	2	South Mumbai	18.9643	72.8078
3	Amboli	0.000000	2	Western Suburbs	19.1293	72.8464
4	Amrut Nagar	0.011236	0	Eastern Suburbs	19.1452	72.8467

Figure 11: Clustering neighborhoods in Mumbai for shopping Malls.

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.

	Neighborhood	Shopping Mall	Cluster Labels	Location	Latitude	Longitude
46	Juhu	0.010000	0	Western Suburbs	19.0149	72.8452
90	Virar	0.012987	0	Western Suburbs	19.0166	72.8585
89	Vile Parle	0.011765	0	Western Suburbs	19.0962	72.8502
3	Amboli	0.013514	0	Western Suburbs	19.1293	72.8464
41	Hiranandani Gardens	0.012987	0	Eastern Suburbs	19.119	72.9068
86	Versova	0.010000	0	Western Suburbs	19.1377	72.8135
50	Kemps Corner	0.010000	0	South Mumbai	18.9647	72.8054
40	Hindu colony	0.010309	0	South Mumbai	19.0197	72.8474
51	Khar Danda	0.010000	0	Western Suburbs	19.0843	72.8269
68	Naigaon	0.011111	0	Western Suburbs	19.0119	72.8453
64	Matunga	0.010989	0	South Mumbai	19.0272	72.8559

Figure 12: Cluster 1.

	Neighborhood	Shopping Mall	Cluster Labels	Location	Latitude	Longitude
49	Kanjurmarg	0.045455	1	Eastern Suburbs	19.1314	72.9357
75	Pant Nagar	0.048780	1	Eastern Suburbs	19.0863	72.915
83	Thakur village	0.057692	1	Western Suburbs	19.2102	72.8754
87	Vidyavihar	0.050505	1	Eastern Suburbs	19.08	72.8973
88	Vikhroli	0.044444	1	Eastern Suburbs	19.1111	72.9278
71	Navy Nagar	0.041667	1	South Mumbai	18.906	72.8155
23	Cotton Green	0.046154	1	South Mumbai	18.9862	72.8412
0	Aarey Milk Colony	0.066667	1	Western Suburbs	19.1703	72.8711
13	C.G.S. colony	0.052632	1	South Mumbai	19.1389	72.9382
9	Bhandup	0.045455	1	Eastern Suburbs	19.1456	72.9486

Figure 13: Cluster 2.

	Neighborhood	Shopping Mall	Cluster Labels	Location	Latitude	Longitude
19	Chembur	0.0	2	Harbour Suburbs	19.054	72.8997
10	Bhayandar	0.0	2	Western Suburbs	19.3074	72.8518
67	Nahur	0.0	2	Eastern Suburbs	19.1537	72.9467
66	Mumbai Central	0.0	2	South Mumbai	18.9697	72.8151
65	Mira Road	0.0	2	Western Suburbs	19.2656	72.8711
11	Bhuleshwar	0.0	2	South Mumbai	18.9512	72.83
63	Marol	0.0	2	Western Suburbs	19.1192	72.8827
62	Marine Lines	0.0	2	South Mumbai	18.9434	72.8232
60	Mankhurd	0.0	2	Harbour Suburbs	19.0485	72.9322
59	Malabar Hill	0.0	2	South Mumbai	18.95	72.795
58	Mahul	0.0	2	Harbour Suburbs	19.0454	72.8932
57	Mahim	0.0	2	South Mumbai	19.0407	72.8431

Figure 14: Cluster 3.

	Neighborhood	Shopping Mall	Cluster Labels	Location	Latitude	Longitude
5	Asalfa	0.031579	3	Eastern Suburbs	19.0953	72.8926
24	Cuffe Parade	0.020619	3	South Mumbai	18.913	72.8205
54	Lower Parel	0.020000	3	South Mumbai	18.9981	72.8281
28	Dagdi Chawl	0.022472	3	South Mumbai	18.9771	72.8291
30	Dava Bazaar	0.022727	3	South Mumbai	19.1314	72.927
77	Poisar	0.022989	3	Western Suburbs	19.2116	72.8527
76	Parel	0.030000	3	South Mumbai	18.9957	72.84
8	Bangur Nagar	0.030000	3	Western Suburbs	19.1674	72.8323
72	Nehru Nagar	0.020000	3	Eastern Suburbs	19.0005	72.8228
26	Currey Road	0.021739	3	South Mumbai	18.9952	72.8346
92	Worli	0.021739	3	South Mumbai	19.0074	72.8169

Figure 15: Cluster 4.

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 16.

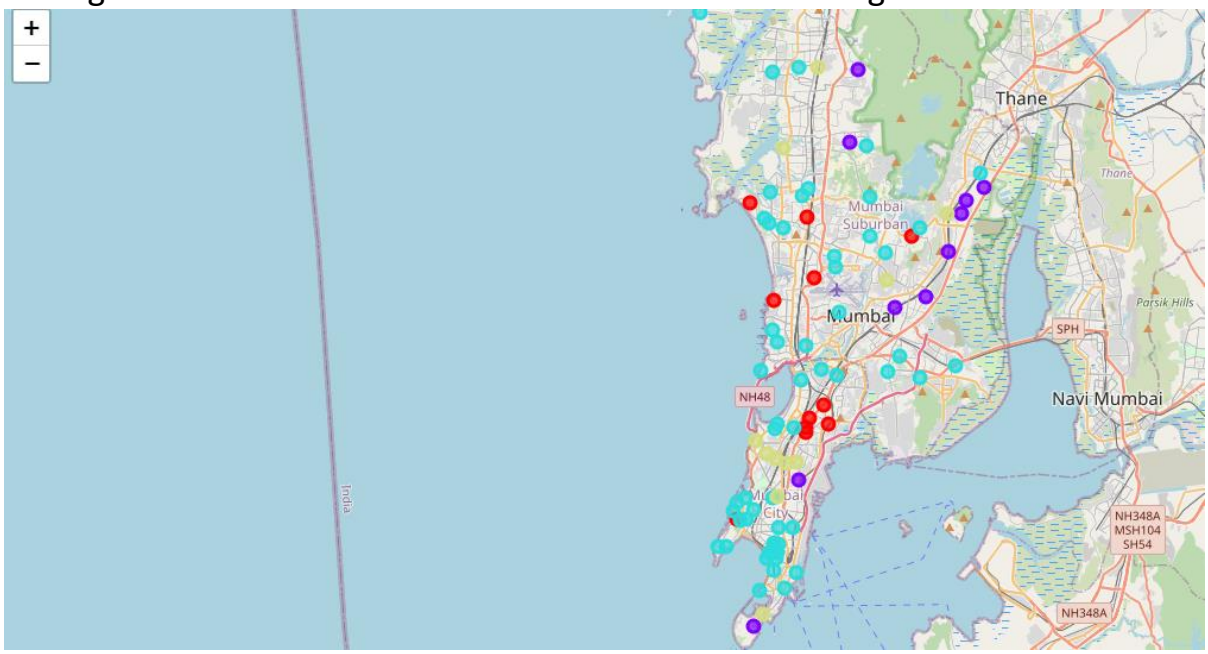

Figure 16: Visualizing the clustering of neighborhoods in Mumbai.

Discussion

By analysing the four clusters obtained we can see that

- Clusters 3 have negligible of shopping Malls in comparison with other clusters.
- Clusters 1 have minimum no of Shopping malls in comparison to Cluster 2&4
- Recommendation: To find the least competition among the other Mall properties as per analysis Cluster 1 areas/locations would be advisable as Best recommendation locations.
- Thus, the most optimal neighborhoods for opening shopping mall are in cluster 1

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 4 clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new shopping mall. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The neighborhoods in **cluster 3** are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.

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.

Notebook Link: