Exploring Neighborhoods of Mumbai for Starting Café

Applied Data Science Capstone Project

By: Abhijeet Mukherjee

Date: March 22, 2021

Table of Contents

Introduction

Mumbai is one of the busiest metropolitan in India, being the financial capital and one of the densely populated City. Due to multiple iconic spots, Mumbai is a major tourist attraction every year. I have been working in Mumbai for a while now and have unique experience with the neighborhoods. It is one of the major IT hub and everyday thousands of professionals prefer business meetups in Café. This is a everyday dilemma faced by IT professional like myself to actually find a place to carry out small tasks like sending mails, making calls without background disturbance or grabbing a coffee between a hectic day. Thus, the aim of this project is to study the neighborhoods in Mumbai to determine possible locations for starting a Café . This project can be useful for business owners and entrepreneurs who are looking to invest in a Café in Mumbai. The main objective of this project is to carefully analyze appropriate data and find recommendations for the stakeholders.

Data Collection

For this project we need the following data:

- Neighborhoods Data of Mumbai City
- Geographical Coordinates of Mumbai and it's Neighborhood
- Venue Data of Neighborhoods in Mumbai

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. Here is a view of top 10 rows:

	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

Geographical Coordinates

The geographical coordinates for Mumbai data 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 coordinates of Mumbai is:

```
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.
```

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 coordinates are then further used for plotting using the Folium library in python. Figure below 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	0.00376476	0.00464166
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	19.2657	72.8707	0.0184624	0.000418149

Next figure shows the final data frame after replacing latitude and longitude data and dropping unnecessary columns out of scope of our analysis.

	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	19.2657	72.8711

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 shows the top 10 rows depicting the results obtained after cleaning the data from Foursquare API.

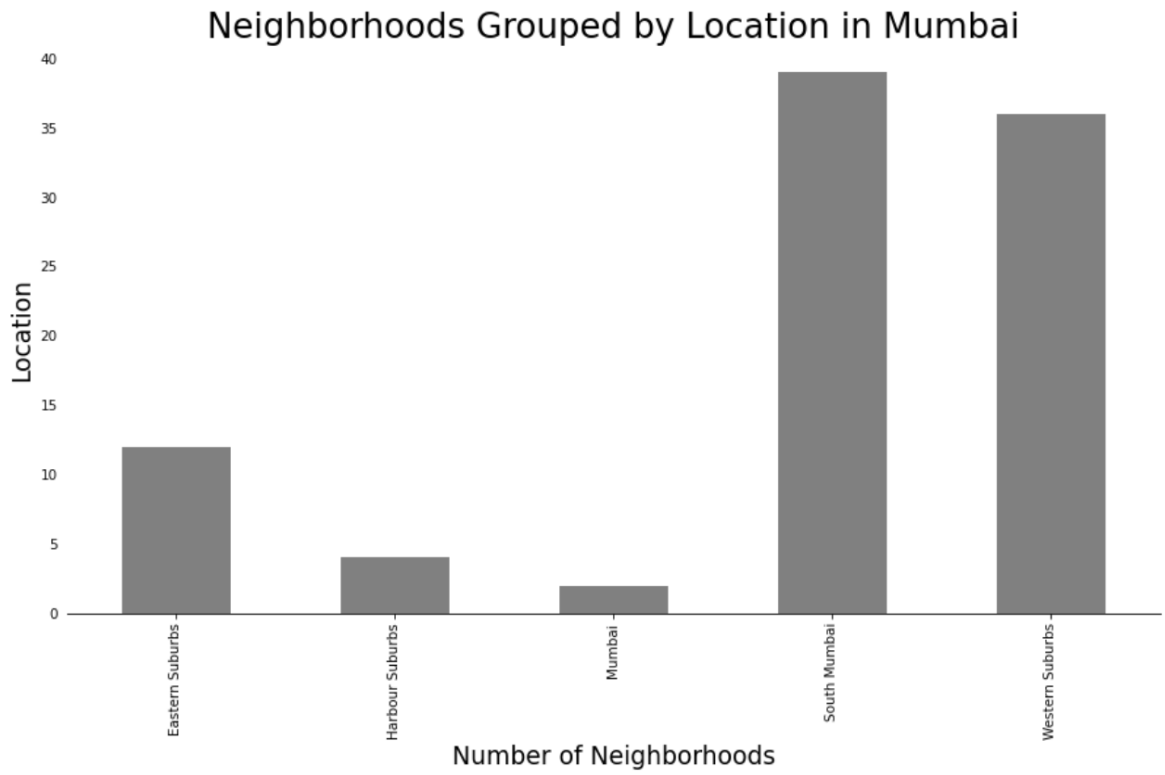
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amboli	19.12930	72.84644	Cafe Arfa	19.128930	72.847140	Indian Restaurant
1	Amboli	19.12930	72.84644	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
2	Amboli	19.12930	72.84644	Jaffer Bhai's Delhi Darbar	19.137714	72.845909	Mughlai Restaurant
3	Amboli	19.12930	72.84644	Narayan Sandwich	19.121398	72.850270	Sandwich Place
4	Amboli	19.12930	72.84644	Shawarma Factory	19.124591	72.840398	Falafel Restaurant
5	Amboli	19.12930	72.84644	Persia Darbar	19.136952	72.846822	Indian Restaurant
6	Amboli	19.12930	72.84644	Domino's Pizza	19.131000	72.848000	Pizza Place
7	Amboli	19.12930	72.84644	Garden Court	19.127188	72.837478	Indian Restaurant
8	Amboli	19.12930	72.84644	Subway	19.127860	72.844461	Sandwich Place
9	Amboli	19.12930	72.84644	Sarvodaya Veg. Restaurant	19.123760	72.850893	Indian Restaurant

Methodology

This section explains the methodology used in the analysis of the project.

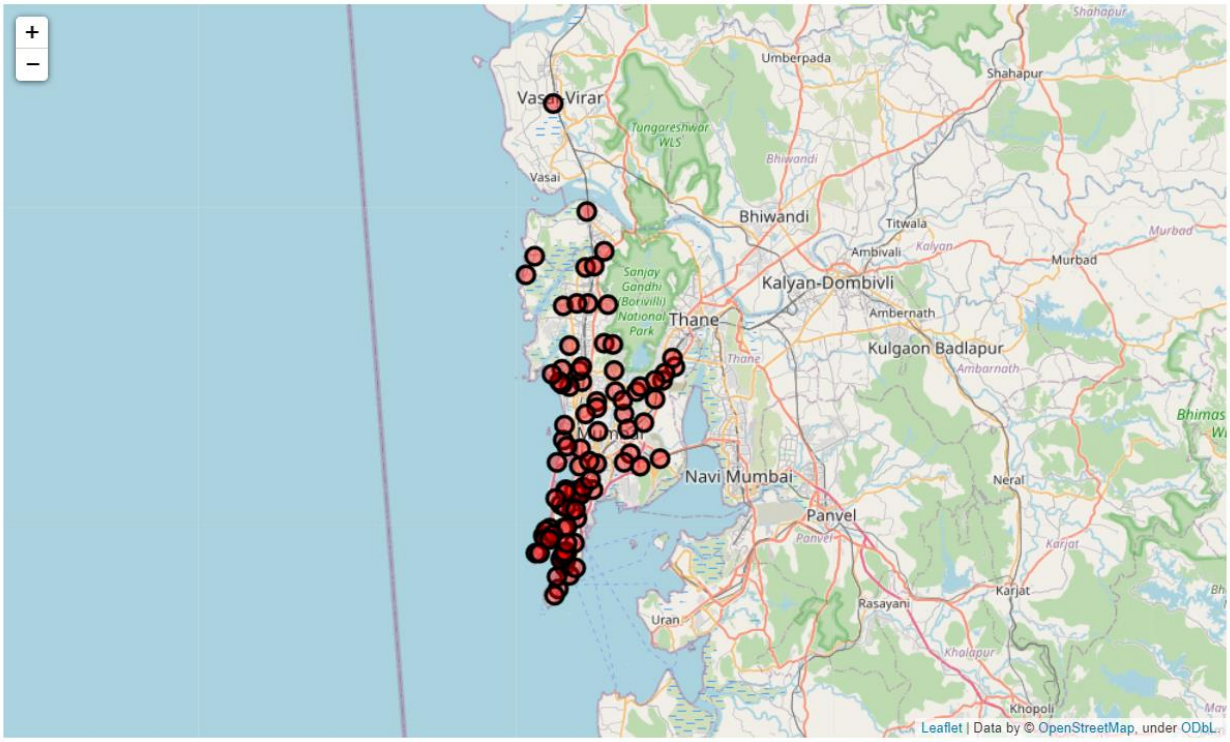
Visualization

For the better understanding of Neighborhood of Mumbai, some visual representation on a basic level has been used. Figure below shows the bar plotting of neighborhood in each location of Mumbai.



It is evident from Figure above that South Mumbai and Western Suburbs have the most number of neighborhoods. But why 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 below.



Feature Extraction

Feature extraction was carried out to obtain features from the Foursquare API data which was used for building the unsupervised learning model. In order to achieve this, the "Venue Category" column had to be converted to 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 which shows only the top 10 rows.

	Neighborhood	Accessories Store	Airport	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Arcade	Art Gallery	Arts & Crafts Store	...	Train	Train Station	Travel & Transport	Vegetarian / Vegan Restaurant	Whisky Bar
0	Amboli	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.000000	0.0
1	Chakala, Andheri	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.048780	0.0
2	D.N. Nagar	0.021739	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.043478	0.0
3	Four Bungalows	0.014925	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.029851	0.0
4	Lokhandwala	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.010870	0.0
5	Marol	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.000000	0.0
6	Sahar	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.000000	0.0
7	Seven Bungalows	0.015873	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.031746	0.0
8	Versova	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.025	...	0.0	0.0	0.0	0.000000	0.0
9	Mira Road	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	0.0	0.0	0.0	0.062500	0.0

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.

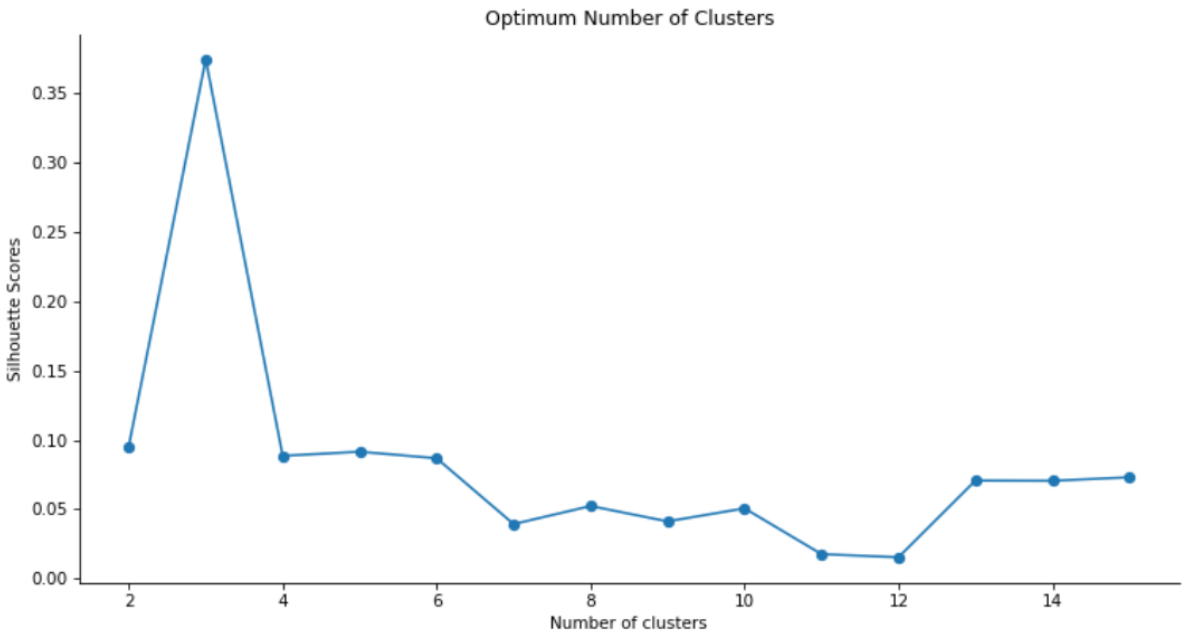
A dataframe was also created which contained the top 10 most common venues of all neighborhoods. Though this is not a part of Feature Extraction, it is important to provide a glimpse into what this dataframe looks like as it will be used later to combine the results from the unsupervised learning model. The data

frame is shown in Figure below.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Amboli	Indian Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Bar	Asian Restaurant	Bakery	Camera Store	Falafel Restaurant	Snack Place
1	Chakala, Andheri	Hotel	Café	Indian Restaurant	Fast Food Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Restaurant	Hotel Bar	Asian Restaurant	Bar
2	D.N. Nagar	Bar	Gym / Fitness Center	Pub	Pizza Place	Indian Restaurant	Snack Place	Women's Store	Vegetarian / Vegan Restaurant	Lounge	Market
3	Four Bungalows	Indian Restaurant	Chinese Restaurant	Café	Pub	Ice Cream Shop	Bar	Seafood Restaurant	Coffee Shop	Vegetarian / Vegan Restaurant	Gym / Fitness Center
4	Lokhandwala	Indian Restaurant	Chinese Restaurant	Café	Pub	Gym / Fitness Center	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Bar	Lounge
...	...	...	...	...	...	...	...	...	...	...	...
88	Parel	Indian Restaurant	Coffee Shop	Chinese Restaurant	Pharmacy	Clothing Store	Playground	Plaza	Multicuisine Indian Restaurant	Restaurant	Roof Deck
89	Gowalia Tank	Indian Restaurant	Café	Bakery	Coffee Shop	Fast Food Restaurant	Snack Place	Chinese Restaurant	Pizza Place	Electronics Store	Salon / Barbershop
90	Dava Bazaar	Train Station	Café	Clothing Store	Food Truck	Beer Garden	Asian Restaurant	Cupcake Shop	Coffee Shop	Fish Market	Falafel Restaurant
91	Dharavi	Indian Restaurant	Fast Food Restaurant	Music Venue	Shoe Store	Seafood Restaurant	Sandwich Place	Lake	Food & Drink Shop	Café	Garden
92	Thane	Pizza Place	Performing Arts Venue	Dessert Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space	Electronics Store	Donut Shop

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.



The data will be clustered to the best possible extent. For this, 3 clusters will be used for the k-means clustering model since it provides the highest silhouette score as seen in Figure above.

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 along with the Location, Latitude, and Longitude columns to provide a complete summary of the clustering. The data frame is shown in Figure below.

	Neighborhood	Location	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th M Com Venu
0	Amboli	Western Suburbs	19.1293	72.8464	1	Indian Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Bar	Asian Restaurant	Bakery	Camera Store	Falaf Resta
1	Chakala, Andheri	Western Suburbs	19.1084	72.8623	1	Hotel	Café	Indian Restaurant	Fast Food Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Restaurant	Hotel Bar	Asian Resta
2	D.N. Nagar	Western Suburbs	19.1241	72.8325	1	Bar	Gym / Fitness Center	Pub	Pizza Place	Indian Restaurant	Snack Place	Women's Store	Vegetarian / Vegan Restaurant	Loun
3	Four Bungalows	Western Suburbs	19.1264	72.8242	1	Indian Restaurant	Chinese Restaurant	Café	Pub	Ice Cream Shop	Bar	Seafood Restaurant	Coffee Shop	Vege / Veg Resta
4	Lokhandwala	Western Suburbs	19.1432	72.8249	1	Indian Restaurant	Chinese Restaurant	Café	Pub	Gym / Fitness Center	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Bar
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
88	Parel	South Mumbai	18.9957	72.84	1	Indian Restaurant	Coffee Shop	Chinese Restaurant	Pharmacy	Clothing Store	Playground	Plaza	Multicuisine Indian Restaurant	Resta
89	Gowalia Tank	South Mumbai	18.9645	72.8112	1	Indian Restaurant	Café	Bakery	Coffee Shop	Fast Food Restaurant	Snack Place	Chinese Restaurant	Pizza Place	Elect Store
90	Dava Bazaar	South Mumbai	19.1314	72.927	0	Train Station	Café	Clothing Store	Food Truck	Beer Garden	Asian Restaurant	Cupcake Shop	Coffee Shop	Fish Mark
91	Dharavi	Mumbai	19.0467	72.8546	1	Indian Restaurant	Fast Food Restaurant	Music Venue	Shoe Store	Seafood Restaurant	Sandwich Place	Lake	Food & Drink Shop	Café
92	Thane	Mumbai	19.1409	72.8826	2	Pizza Place	Performing Arts Venue	Dessert Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space	Elect Store

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.

CLUSTER 1

```
In [66]: mum_merged.loc[mum_merged['Cluster Labels'] == 0, mum_merged.columns[[0] + [1] + list(range(5, mum_merged.shape[1]))]]
```

Out[66]:

	Neighborhood	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
32	Nalasopara	Western Suburbs	Multiplex	Pizza Place	Fast Food Restaurant	Ice Cream Shop	Diner	Department Store	Bus Station	Farmers Market	Falafel Restaurant	Factory
36	Bhandup	Eastern Suburbs	Train Station	Fast Food Restaurant	Indian Restaurant	Asian Restaurant	Zoo	Dhaba	Field	Farmers Market	Falafel Restaurant	Factory
40	Kanjurmarg	Eastern Suburbs	Train Station	Chinese Restaurant	Asian Restaurant	Gym	Food Truck	Cupcake Shop	Multiplex	Gift Shop	Diner	Dive Bar
50	Mankhurd	Harbour Suburbs	Bus Station	Coffee Shop	Sports Bar	Train Station	Zoo	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space
59	Cotton Green	South Mumbai	Plaza	Train Station	Pizza Place	Multiplex	Ice Cream Shop	Fast Food Restaurant	Falafel Restaurant	Factory	Event Space	Electronics Store
90	Dava Bazaar	South Mumbai	Train Station	Café	Clothing Store	Food Truck	Beer Garden	Asian Restaurant	Cupcake Shop	Coffee Shop	Fish Market	Falafel Restaurant

CLUSTER 2

```
[67]: mum_merged.loc[mum_merged['Cluster Labels'] == 1, mum_merged.columns[[0] + [1] + list(range(5, mum_merged.shape[1]))]]
```

:[67]:

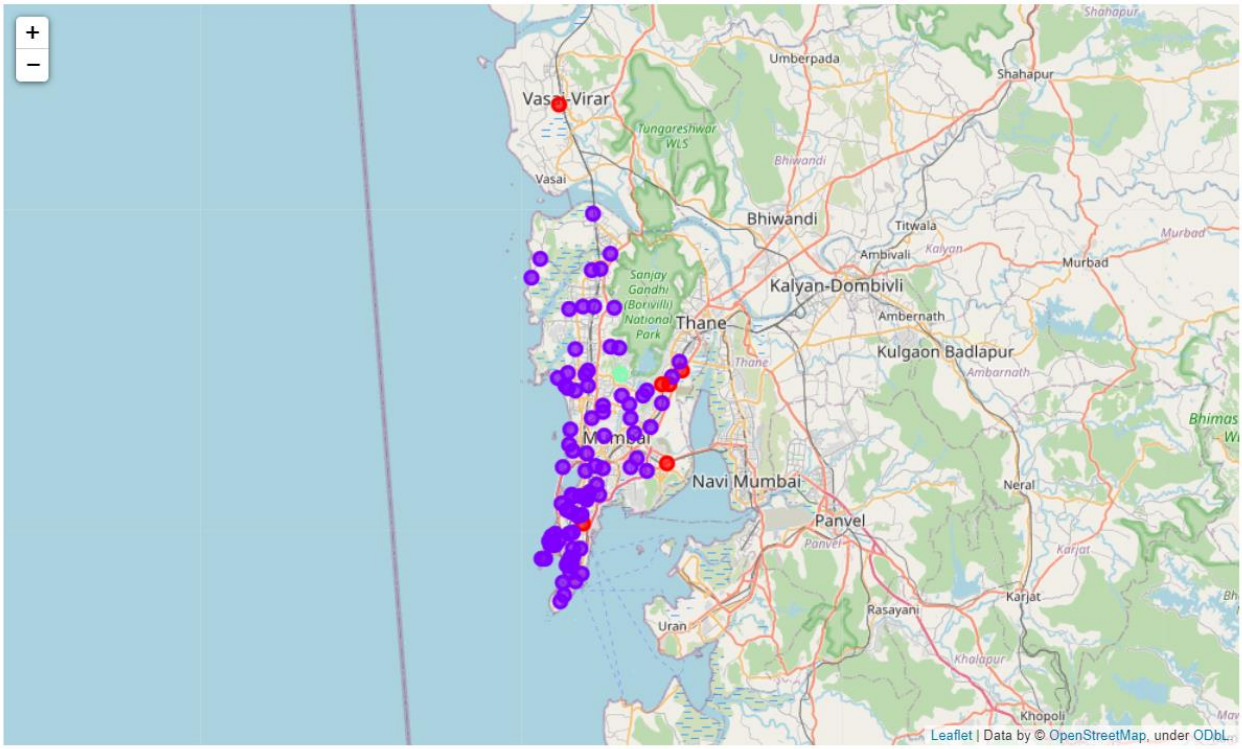
	Neighborhood	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Amboli	Western Suburbs	Indian Restaurant	Pizza Place	Coffee Shop	Sandwich Place	Bar	Asian Restaurant	Bakery	Camera Store	Falafel Restaurant	Snack Place
1	Chakala, Andheri	Western Suburbs	Hotel	Café	Indian Restaurant	Fast Food Restaurant	Pizza Place	Vegetarian / Vegan Restaurant	Restaurant	Hotel Bar	Asian Restaurant	Bar
2	D.N. Nagar	Western Suburbs	Bar	Gym / Fitness Center	Pub	Pizza Place	Indian Restaurant	Snack Place	Women's Store	Vegetarian / Vegan Restaurant	Lounge	Market
3	Four Bungalows	Western Suburbs	Indian Restaurant	Chinese Restaurant	Café	Pub	Ice Cream Shop	Bar	Seafood Restaurant	Coffee Shop	Vegetarian / Vegan Restaurant	Gym / Fitness Center
4	Lokhandwala	Western Suburbs	Indian Restaurant	Chinese Restaurant	Café	Pub	Gym / Fitness Center	Coffee Shop	Fast Food Restaurant	Italian Restaurant	Bar	Lounge
...	...	...	...	...	...	...	...	...	...	...	...	...
86	Chor Bazaar	South Mumbai	Indian Restaurant	Dessert Shop	BBQ Joint	Bus Station	Ice Cream Shop	Rest Area	Restaurant	Food	Middle Eastern Restaurant	Chinese Restaurant
87	Matunga	South Mumbai	Indian Restaurant	Fast Food Restaurant	Café	Snack Place	Ice Cream Shop	Chinese Restaurant	South Indian Restaurant	Vegetarian / Vegan Restaurant	Train Station	Coffee Shop
88	Parel	South Mumbai	Indian Restaurant	Coffee Shop	Chinese Restaurant	Pharmacy	Clothing Store	Playground	Plaza	Multicuisine Indian Restaurant	Restaurant	Roof Deck
89	Gowalia Tank	South Mumbai	Indian Restaurant	Café	Bakery	Coffee Shop	Fast Food Restaurant	Snack Place	Chinese Restaurant	Pizza Place	Electronics Store	Salon / Barbershop
91	Dharavi	Mumbai	Indian Restaurant	Fast Food Restaurant	Music Venue	Shoe Store	Seafood Restaurant	Sandwich Place	Lake	Food & Drink Shop	Café	Garden

CLUSTER 3

```
mum_merged.loc[mum_merged['Cluster Labels'] == 2, mum_merged.columns[[0] + [1] + list(range(5, mum_merged.shape[1]))]]
```

	Neighborhood	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
92	Thane	Mumbai	Pizza Place	Performing Arts Venue	Dessert Shop	Fast Food Restaurant	Farmers Market	Falafel Restaurant	Factory	Event Space	Electronics Store	Donut Shop

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 below

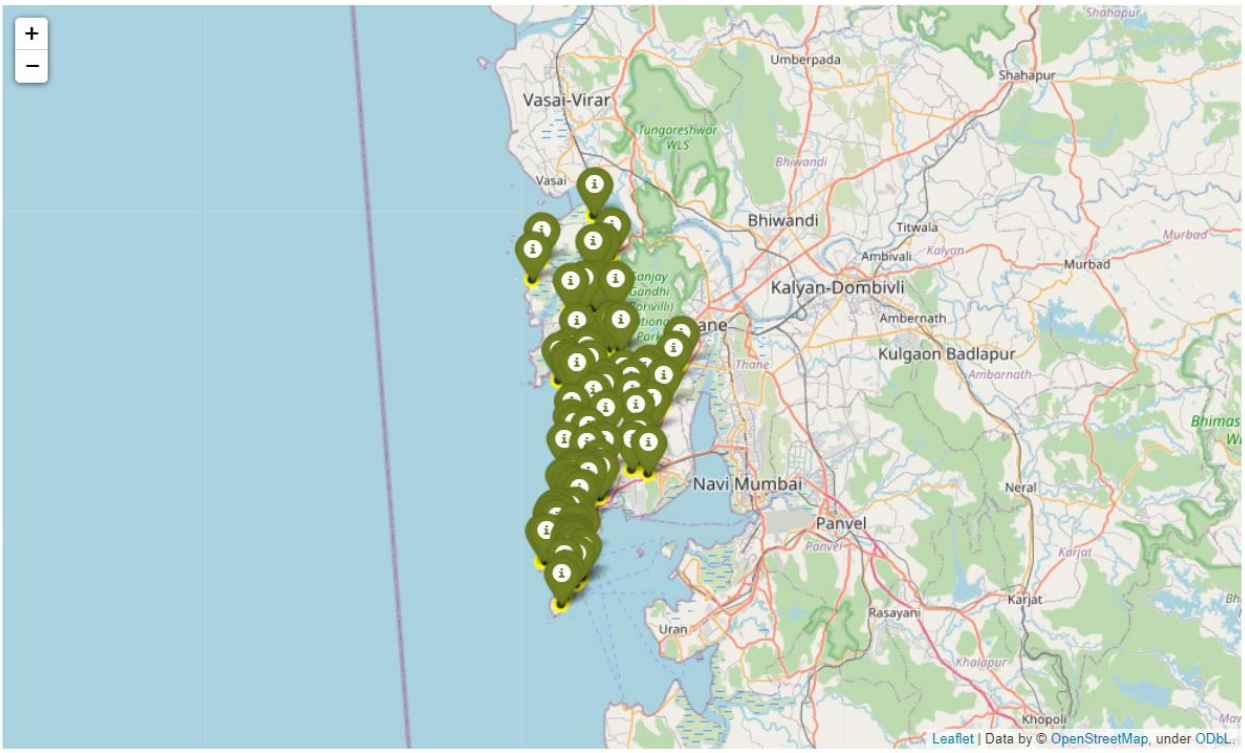


Discussion

By analyzing the three clusters obtained we can see that some of the clusters are more Cafe, whereas, other clusters are less suited. Neighborhoods in clusters 1 and 3 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 cafe. On the other hand, neighborhoods in cluster 2 contain a much higher degree of restaurants, hotels, multiplex, cafes, bars and other food joints. Thus, the neighborhoods in this clusters would be well suited for opening a new restaurant.

Most neighborhoods in cluster 2 seem to have Cafe 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 Cafe can be opened in the neighborhoods belonging to cluster 2. This neighborhood can be further plotted on a map as shown below.

This neighborhood can be further plotted on a map as shown below in Figure below.


Vasai-Virar
Umberpada
Shahapur
Bhiwandi
Titwala
Ambivali
Murbad
Kalyan-Dombivli
Ambernath
Kulgaon Badlapur
Navi Mumbai
Panvel
Neral
Karjat
Rasayani
Uran
Khopoli
Leaflet | Data by © OpenStreetMap, under ODbL.

Conclusion

Successfully analyzed the neighborhoods in Mumbai, India for determining which would be the best neighborhoods for opening a Cafe. Based on our analysis, neighborhoods in cluster 2 are recommended as locations for the new cafe. 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.