Analyzing the Neighborhoods in Mumbai for Best Infrastructure Place

Applied Data Science Capstone Project

By: Hitesh Chhugani

Date: December, 2020

Contents

INTRODUCTION

The purpose of this Project is to help people in exploring better facilities around their neighbourhood. It will help people making smart and efficient decision on selecting great neighbourhood out of numbers of other neighbourhoods in Mumbai, Maharashtra, India.

Mumbai is a financial capital of India so Lots of people are migrating from various states of India to Mumbai for their livelihood Therefore they need lots of research for area with good infrastructure like schools, hospitals, malls, cafe, restaurants and connectivity with the different parts of the city. Even many companies are coming in Mumbai so they can select the areas for their company office to check the infrastructure available and connectivity available in that area with different parts, so it can be very helpful. This project is for those people who are looking for better neighbourhoods.

It will help people to get awareness of the area and neighbourhood before moving to a new city, state, country or place for their work or to start a new fresh life.

Data Collection

The following data is required for the project:

1) Neighbourhood data of Mumbai
2) Geographical coordinates of Mumbai
3) Venue data for neighbourhoods in Mumbai

Neighbourhoods Data

The data of the neighbourhoods in Mumbai was scraped from https://www.nativeplanet.com/india-pin-codes/maharashtra/mumbai/ The data is read into a pandas data frame using the read_html() method. The main reason for doing so is that the Web 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 data frame are shown in Figure 1.

	Post Office	Pincode	City
1	Antop Hill	400037	Mumbai
2	B P T Colony	400037	Mumbai
3	Haffkin Institute	400012	Mumbai
4	Mazgaon Dock	400010	Mumbai
5	Lal Baug	400012	Mumbai
6	Parel Rly Work Shop	400012	Mumbai
7	Princess Dock	400009	Mumbai
8	Reay Road	400033	Mumbai
9	Cotton Exchange	400033	Mumbai
10	Dadar Colony	400014	Mumbai

Figure 1: Top 10 rows of Mumbai neighbourhood's data scraped from Wikipedia.

Geographical Coordinates

The geographical coordinates for Mumbai have 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.

```
!pip install geocoder
```

```
import geocoder
```

```
def get_lati_long(Pincode):
    lati_long_coords = None
    while(lati_long_coords is None):
        g = geocoder.arcgis('{}, Mumbai, India'.format(Pincode))
        lati_long_coords = g.latlng
    return lati_long_coords

get_lati_long('400009')
```

```
[18.95742566200005, 72.83766500000007]
```

```
Pincode = df['Pincode']
coords = [get_lati_long(Pincode) for Pincode in Pincode.tolist()]
```

Figure 2: Obtaining geographical coordinates of Mumbai.

The geocoder library in python has been used to obtain latitude and longitude data for various neighbourhoods in Mumbai.

Figure 3 shows the top 10 rows of the final Mumbai neighbourhoods data frame after replacing the latitude and longitude values as mentioned before and dropping unnecessary columns.

	Post Office	Pincode	City	Latitude	Longitude
0	Antop Hill	400037	Mumbai	19.020313	72.868280
1	B P T Colony	400037	Mumbai	18.997550	72.840608
2	Haffkin Institute	400012	Mumbai	18.971480	72.843874
3	Mazgaon Dock	400010	Mumbai	18.997550	72.840608
4	Lal Baug	400012	Mumbai	18.997550	72.840608
5	Parel Rly Work Shop	400012	Mumbai	18.957426	72.837665
6	Princess Dock	400009	Mumbai	18.986693	72.844630
7	Reay Road	400033	Mumbai	18.986693	72.844630
8	Cotton Exchange	400033	Mumbai	19.015996	72.847255
9	Dadar Colony	400014	Mumbai	18.971480	72.843874

Figure 3: Final Mumbai neighbourhoods data frame.

Venue Data

The venue data has been extracted using the Foursquare API. This data contains venue recommendations for all neighbourhoods in Mumbai and is used to study the popular venues of different neighbourhoods as well as build the unsupervised learning model to cluster neighbourhoods. The venue recommendations of all neighbourhoods were obtained with a limit of 100, that is, maximum of 100 venue recommendations per neighbourhood and a radius of 500 m around the neighbourhood's geographical coordinates. Figure 4 shows the top 10 rows depicting the results obtained after cleaning the data from Foursquare API.

	Post Office	Post Office Latitude	Post Office Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Antop Hill	19.020313	72.868280	Malvani Days	19.018984	72.865363	Diner
1	Antop Hill	19.020313	72.868280	Monginis Cake Shop	19.022737	72.865670	Convenience Store
2	Antop Hill	19.020313	72.868280	Wadala Skywalk	19.018421	72.864422	Trail
3	B P T Colony	18.997550	72.840608	ITC Grand Central	18.998469	72.838433	Hotel
4	B P T Colony	18.997550	72.840608	Kebabs & Kurries	18.997938	72.837639	Indian Restaurant
5	B P T Colony	18.997550	72.840608	Terrace Garden	18.998119	72.838529	Roof Deck
6	B P T Colony	18.997550	72.840608	Hornby's Pavilion	18.998141	72.838419	Restaurant
7	B P T Colony	18.997550	72.840608	The Point Of View	18.998211	72.838543	Lounge
8	B P T Colony	18.997550	72.840608	7 Spice - Chinese Restaurant	18.995603	72.839060	Chinese Restaurant
9	B P T Colony	18.997550	72.840608	Dublin	18.997737	72.837737	Bar

Figure 4: 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 neighbourhoods, basic visualization was carried out.

Using folium, a map was plotted to show how the different neighbourhoods are spread all across Mumbai. This is shown in Figure 5.

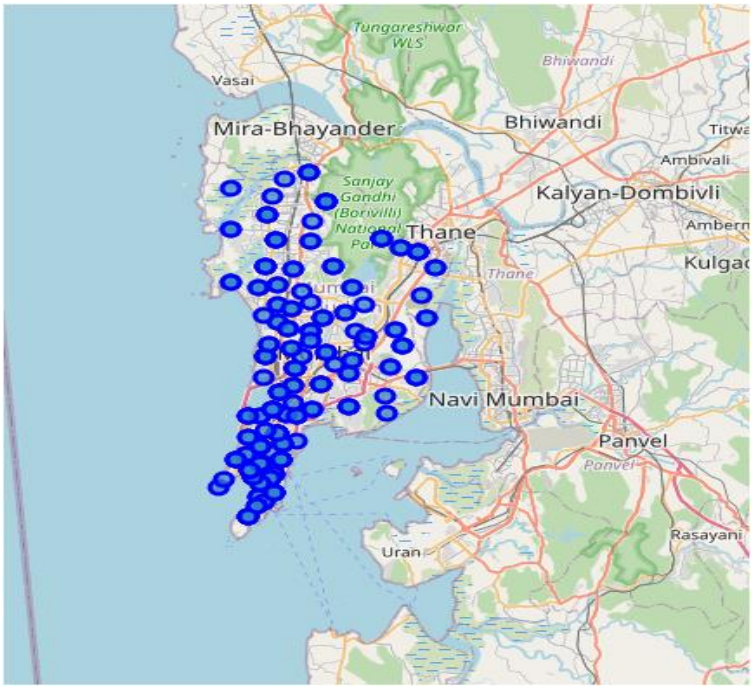

Figure 5: Depicting the neighbourhood spread across Mumbai.

Feature Extraction

Feature extraction was carried out to obtain features from the Foursquare API data (as shown in Figure 4) 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 neighbourhood venue belongs to that category, it would get a value of 1 for that row in that specific category column and if a neighbourhood venue does not belong to the particular category, the value would be 0. This process was repeated for all venues in all neighbourhoods and the result was a sparse matrix containing the neighbourhood name and all unique category columns with either 1 or 0 based on whether the neighbourhood venue belonged to that category or not. This data frame was then grouped by the neighbourhood name and the average value was taken for all categories. The result is shown in Figure 8 which shows only the top 10 rows.

	Post Office	ATM	Airport	Airport Lounge	American Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Workshop	BBQ Joint	Bagel Shop	Bakery	Bank	Bar	Beach	Bed & Breakfast	Beer Garden	Bengali Restaurant	Bistro
0	A I Staff Colony	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.125000	0.0	0.0	0.000000	0.0	0.0
1	Aareymilk Colony	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0
2	Agripada	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0
3	Airport (Mumbai)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0
4	Ambewadi (Mumbai)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.062500	0.0	0.062500	0.0	0.0	0.000000	0.0	0.0
5	Andheri East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.019608	0.000000	0.0	0.0	0.0	0.078431	0.0	0.078431	0.0	0.0	0.019608	0.0	0.0
6	Andheri H.O	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.166667	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0
7	Andheri Railway Station	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.166667	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0
8	Antop Hill	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0
9	Asvini	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.000000	0.0	0.0

Figure 6: 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 neighbourhoods had venues belonging to each category. This data was used for the unsupervised learning model with the neighbourhood 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 neighbourhoods. 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 5 rows of this data frame are shown in Figure 7

	Post Office	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	A I Staff Colony	Indian Restaurant	Pet Store	Bus Station	Vegetarian / Vegan Restaurant	Fast Food Restaurant	Bar	Juice Bar	Falafel Restaurant	Food	Flower Shop
1	Aareymilk Colony	Pizza Place	Indian Restaurant	Fast Food Restaurant	Neighborhood	Restaurant	Zoo	Electronics Store	Flower Shop	Flea Market	Field
2	Agripada	Hotel	Dance Studio	Shoe Store	Fast Food Restaurant	Falafel Restaurant	Food & Drink Shop	Food	Flower Shop	Flea Market	Field
3	Airport (Mumbai)	Spa	Indian Restaurant	Café	Indie Movie Theater	Gym	Italian Restaurant	Fast Food Restaurant	Zoo	Falafel Restaurant	Food
4	Ambewadi (Mumbai)	Italian Restaurant	Indian Restaurant	Snack Place	Food Truck	Coffee Shop	Hotel	Gym	Ice Cream Shop	Shopping Mall	Spa

Figure 7: Top 5 most common venues for neighbourhoods.

Unsupervised Learning

K-means unsupervised learning technique was used to cluster the neighbourhoods 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 8.

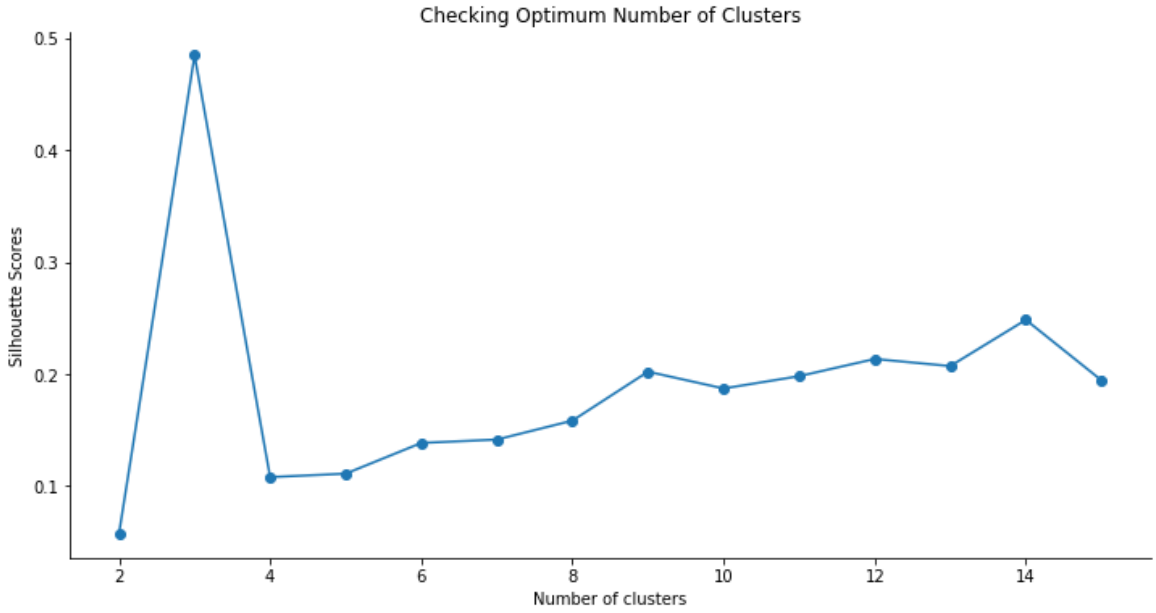

Figure 8: Silhouette scores for different number of clusters.

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, 3 clusters will be used for the k-means clustering model since it provides the highest silhouette score as seen in Figure 8.

Results

To get the best possible area in Mumbai with good infrastructure, the venue results were evaluated. Figure shows some major infrastructure's for which venue results were analysed. It includes 33 major infrastructures.

```
# Quality Infrastructure
search_query= ['Restaurant', 'Hotel', 'Farmers Market', 'Shopping Mall', 'Gym / Fitness Center', 'Pharmacy',
                        'Electronics Store', 'Indie Movie Theater', 'Light Rail Station','Metro Station', 'Train','Train Station', 'Garden',
                         'Theater','ATM', 'Office', 'Bus Station', 'Bank', 'Market' , 'Business Service', 'Monument / Landmark' ,
                         'Resort', 'Hospital', 'Police Station', 'School', 'College', 'Café' , 'Park', 'Playground',
                'Convention Center', 'College Auditorium', 'Government Building', 'Airport Terminal',
                        ]
print(len(search_query))
```

```
33
```

Figure 9: 33 Major of Infrastructure

	Post Office	High Court Building (Mumbai)	Opera House	Stock Exchange
ATM	0	0	0
Bank	1	1	1
Bus Station	0	0	0
Café	7	7	7
College Auditorium	0	0	0
Electronics Store	0	0	0
Farmers Market	0	0	0
Garden	0	0	0
Gym / Fitness Center	1	1	1
Hotel	2	2	2
Indie Movie Theater	0	0	0
Light Rail Station	0	0	0
Market	0	0	0
Monument / Landmark	1	1	1
Office	0	0	0
Park	0	0	0
Pharmacy	0	0	0
Playground	0	0	0
Restaurant	1	1	1
Shopping Mall	0	0	0
Theater	0	0	0
Train Station	0	0	0
Total Infrastructure	26	26	26

Figure 10: Best Infrastructure Place.

After analysing this all the infrastructure were added for each postal code and was displayed in total infrastructure column. After adding all infrastructure, the area with maximum infrastructure were called the best place in Mumbai. Figure 10 shows High Court Building, Opera House and Stock exchange were some areas of best infrastructure in Mumbai.

After analysing the best place of infrastructure in Mumbai, it's also become too important to find out the area with least infrastructure. Below Figure 11 shows some infrastructure with least infrastructure. There are 35 areas with least infrastructure from that list.

	Post Office	Pincode	City	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 most common venue	10th most common venue
0	Antop Hill	400037	Mumbai	19.020313	72.868280	2.0	Convenience Store	Diner	Trail	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field
1	B P T Colony	400037	Mumbai	18.997550	72.840608	0.0	Indian Restaurant	Office	Bar	Shoe Store	Roof Deck	Luggage Store	Lounge	Restaurant	Chinese Restaurant	Coffee Shop
2	Haffkin Institute	400012	Mumbai	18.971480	72.843874	0.0	Ice Cream Shop	Indian Restaurant	Dessert Shop	Flea Market	Café	Chinese Restaurant	Zoo	Farmers Market	Food & Drink Shop	Food
3	Mazgaon Dock	400010	Mumbai	18.997550	72.840608	0.0	Indian Restaurant	Office	Bar	Shoe Store	Roof Deck	Luggage Store	Lounge	Restaurant	Chinese Restaurant	Coffee Shop
4	Lal Baug	400012	Mumbai	18.997550	72.840608	0.0	Indian Restaurant	Office	Bar	Shoe Store	Roof Deck	Luggage Store	Lounge	Restaurant	Chinese Restaurant	Coffee Shop
5	Parel Rly Work Shop	400012	Mumbai	18.957426	72.837665	0.0	Indian Restaurant	Hotel	Furniture / Home Store	Convenience Store	Chinese Restaurant	Café	Sandwich Place	Flea Market	Smoke Shop	Harbor / Marina
6	Princess Dock	400009	Mumbai	18.986693	72.844630	2.0	Whisky Bar	Train Station	Plaza	Bakery	Zoo	Falafel Restaurant	Food & Drink Shop	Food	Flower Shop	Flea Market
7	Reay Road	400033	Mumbai	18.986693	72.844630	2.0	Whisky Bar	Train Station	Plaza	Bakery	Zoo	Falafel Restaurant	Food & Drink Shop	Food	Flower Shop	Flea Market
8	Cotton Exchange	400033	Mumbai	19.015996	72.847255	0.0	Indian Restaurant	Coffee Shop	Café	Hotel	Vegetarian / Vegan Restaurant	Lounge	Restaurant	Juice Bar	Bakery	Farmers Market
9	Dadar Colony	400014	Mumbai	18.971480	72.843874	0.0	Ice Cream Shop	Indian Restaurant	Dessert Shop	Flea Market	Café	Chinese Restaurant	Zoo	Farmers Market	Food & Drink Shop	Food

Figure 11: 35 Least Infrastructure

	Post Office	Total Infrastructure	ATM	Bank	Bus Station	Café	College Auditorium	Electronics Store	Farmers Market	Garden	Gym / Fitness Center	Hotel	Indie Movie Theater	Light Rail Station	Market	Monument / Landmark	Office	Park	Pharmacy	Playground	Restaurant	Shopping Mall	Theater	Train Station
0	A I Staff Colony	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Aareymilk Colony	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
2	Agripada	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
7	Azad Nagar (Mumbai)	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Borivali East	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 12: Clustering neighbourhoods in Mumbai.

The clustering model then clusters the neighbourhoods in Mumbai and provides a label for each neighbourhood 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 12. Furthermore, neighbourhoods 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 5 rows for clusters that contain a high number of neighbourhoods.

	Post Office	Pincode	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 most common venue	10th most common venue
1	B P T Colony	400037	0.0	Indian Restaurant	Office	Bar	Shoe Store	Roof Deck	Luggage Store	Lounge	Restaurant	Chinese Restaurant	Coffee Shop
2	Haffkin Institute	400012	0.0	Ice Cream Shop	Indian Restaurant	Dessert Shop	Flea Market	Café	Chinese Restaurant	Zoo	Farmers Market	Food & Drink Shop	Food
3	Mazgaon Dock	400010	0.0	Indian Restaurant	Office	Bar	Shoe Store	Roof Deck	Luggage Store	Lounge	Restaurant	Chinese Restaurant	Coffee Shop
4	Lal Baug	400012	0.0	Indian Restaurant	Office	Bar	Shoe Store	Roof Deck	Luggage Store	Lounge	Restaurant	Chinese Restaurant	Coffee Shop
5	Parel Rly Work Shop	400012	0.0	Indian Restaurant	Hotel	Furniture / Home Store	Convenience Store	Chinese Restaurant	Café	Sandwich Place	Flea Market	Smoke Shop	Harbor / Marina

Figure 13: Cluster 1.

	Post Office	Pincode	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 most common venue	10th most common venue
16	Noor Baug	400009	1	ATM	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field	Fast Food Restaurant	Farmers Market
115	Trombay	400088	1	ATM	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field	Fast Food Restaurant	Farmers Market
135	Oshiwara	400102	1	ATM	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field	Fast Food Restaurant	Farmers Market
151	Dahisar	400068	1	ATM	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field	Fast Food Restaurant	Farmers Market

Figure 14: Cluster 2.

	Post Office	Pincode	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 most common venue	10th most common venue
0	Antop Hill	400037	2.0	Convenience Store	Diner	Trail	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field
6	Princess Dock	400009	2.0	Whisky Bar	Train Station	Plaza	Bakery	Zoo	Falafel Restaurant	Food & Drink Shop	Food	Flower Shop	Flea Market
7	Reay Road	400033	2.0	Whisky Bar	Train Station	Plaza	Bakery	Zoo	Falafel Restaurant	Food & Drink Shop	Food	Flower Shop	Flea Market
10	V K Bhavan	400010	2.0	Gym	Smoke Shop	Cupcake Shop	Pizza Place	Café	Movie Theater	Falafel Restaurant	Food	Flower Shop	Flea Market
11	Wadala Rs	400031	2.0	Convenience Store	Diner	Trail	Falafel Restaurant	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Field

Figure 15: Cluster 3.

Based on the clusters shown above, the neighbourhoods 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.

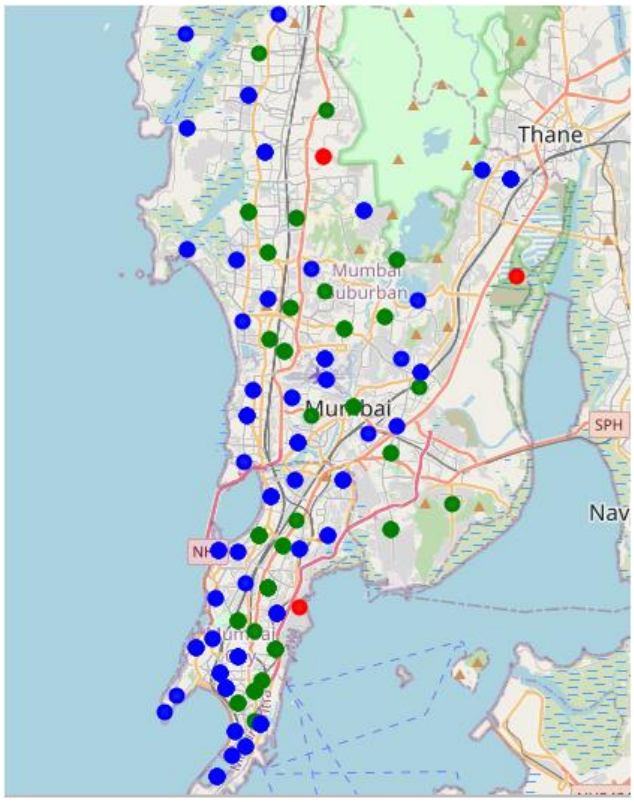

Figure 16: Visualizing the clustering of neighbourhoods in Mumbai.

Discussion

Mumbai has a very good rail connectivity with different parts of the city therefore transportation is not a problem, therefore area near to your job location can be chosen. What Mumbai has a problem of highly expensive houses, so best area can be selected with affordable house price near to the job location along with the facilities like schools, malls, bank, hospital and other basic need requirements.

The basic infrastructure was selected from different venues. This infrastructure is located in different areas in Mumbai. Finally, the area with the best infrastructure is selected. Along with this the area with least infrastructure were located. High Court Building, Opera House and Stock exchange were some areas of best infrastructure in Mumbai. There are 35 areas with least infrastructure from that list.

Clusters were selected with the help of silhoette score. from 2 to 15 no.s of clusters silhouette score is plotted. 3 Cluster are used as it gives maximum value of silhouette score.

Cluster 0 looks like area with multiple office along with other basic needs facility as well as restaurants and other facilities for enjoyment as along with professional life there need some good social life. So, this cluster look good for the people coming to this city and having work location within this cluster.

Cluster 1 is a very small cluster of 4 different areas which include food courts and some farmer's market. So, this cluster doesn't look great to stay in Mumbai from the basic infrastructure point of view.

Cluster 2 is a very big cluster include restaurants, spa, gym, bar, lounge and Hotel. This cluster looks good for new people coming to this city. This cluster also includes hotels, lounge therefore also can be the good area for the tourist in the city, as this area includes airport along with other enjoyment places.

Conclusion

I have successfully analysed the neighbourhoods in Mumbai, India for determining the location for the new people coming to the city. The people coming to the cities can look from the area with good infrastructure like hospitals, schools, gym, malls and proper connectivity with different parts of the city. High Court Building, Opera House and Stock exchange were some areas of best infrastructure in Mumbai.

People can further select the best place according to their job location in the city and there capability of taking the expenses of that area in the city.

Final Comments

Note 1: In order to view the code for this project, kindly refer to the notebook on the github repository at: