DETERMINING THE PLACES FOR ACCOMODATION IN MUMBAI

1. Introduction

1.1 Background

Mumbai is one of the most vibrant cities of India. As of 2018, Mumbai was the most populous city of India. Furthermore, Mumbai ranks seventh in the world among the cities with most population. With around twenty million population, everything gets crowded. Also, Mumbai is the financial, commercial and entertainment capital of India. This leads to the city having numerous corporates settling themselves in the city.

Moreover, Mumbai is situated on a narrow peninsula. This leads to it not being able to expand proportionally to the needs. This results in a much higher population density.

Hence the need for living peacefully and being able to obtain a liveable and healthy environment becomes one of the major priorities when searching for a proper neighbourhood to live in.

1.2 Problem

The problem addressed is the choice to find the correct neighbourhood to live in the city of Mumbai. Since Mumbai is a diverse city, one can easily be misled and find himself/ herself in an area not suitable to live in. The proposed solution is an analysis which identifies the neighbourhoods as one where one can find all the necessary areas in the surroundings of the house. Also, those areas are found where there isn't an environment to live there, rather they are commercial or entertainment areas. These areas would be more polluted, traffic prone and one might struggle to find a peaceful environment.

1.3 Interest

The target audience for this analysis are those who are looking for houses in various neighbourhoods, and want to know the areas which are popular in the neighbourhoods. Also, this analysis can be useful for existing residents, who want to shift or keep their house for sale. People could also use this to set or estimate the prices according to the luxuries and necessities present in the neighbourhood.

2. Data Overview

2.1 Data Source

The data was obtained by scraping the web page of Wikipedia. Link: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

2.2 Data Cleaning

The data was cleaned and preprocessed. All the boroughs were combined. The neighbourhoods were combined, and listed as a single comma separated attribute.

Since there was no latitude or longitude value for the boroughs, and no geo-json file available, the mean of the latitude and longitude values of the neighbourhoods lying in the same borough were taken and listed as the latitude and longitude value of the borough.

2.3 Feature Selection

Feature selection was done by converting the popular areas around a neighbourhood in a 500m radius through one-hot encoding. Then the top ten locations were selected for each borough and used as the features for clustering.

3. Exploratory Data Analysis

3.1 Attribute Analysis

The initial dataset did not have the necessary data to extract the characteristics of each borough from the four square API. Hence, the data set was modified and shortened.

The initial dataset was:

	Area	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

This dataset was modified as shown below and converted so that the areas are well defined and the data can be extracted for each area:

	Location	Area	Latitude	Longitude
0	Andheri,Western Suburbs	Amboli, D.N. Nagar, Four Bungalows, Lokhandwal...	19.122009	72.839780
1	Western Suburbs	Chakala, Andheri, Dahisa, Jogeshwari West, Juh...	19.191909	72.838363
2	Mira-Bhayandar,Western Suburbs	Mira Road, Bhayandar, Uttan	19.284722	72.835370
3	Bandra,Western Suburbs	Bandstand Promenade, Kherwadi, Pali Hill	19.055339	72.825511
4	Borivali (West),Western Suburbs	I.C. Colony, Gorai	19.248548	72.815926
5	Goregaon,Western Suburbs	Aarey Milk Colony, Bangur Nagar	19.157927	72.857004
6	Kandivali West,Western Suburbs	Charkop, Poisar, Mahavir Nagar	19.210671	72.836984
7	Kandivali East,Western Suburbs	Thakur village	19.210206	72.872980
8	Khar,Western Suburbs	Pali Naka, Khar Danda	19.065670	72.834719
9	Malad,Western Suburbs	Dindoshi, Sunder Nagar	19.175691	72.853445
10	Sanctacruz,Western Suburbs	Kalina	19.081667	72.841389

3.2 Data Transformation

The data from the Four Square API is collected and transformed such that the top ten visited places for the area are its attributes for the predictive analysis.

	Location	Area	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
0	Andheri,Western Suburbs	Amboli, D.N. Nagar, Four Bungalows, Lokhandwal...	19.122009	72.839780	0	Vegetarian / Vegan Restaurant	Train Station	Indian Restaurant	College Cafeteria	Athletics & Sports	Chinese Restaurant	Falafel Restaurant	Fish & Chips Shop	Department Store
1	Western Suburbs	Chakala, Andheri, Dahisa, Jogeshwari West, Juh...	19.191909	72.838363	0	Indian Restaurant	Chinese Restaurant	Bar	Fast Food Restaurant	Gym	Restaurant	Grocery Store	Italian Restaurant	Department Store
3	Bandra,Western Suburbs	Bandstand Promenade, Kherwadi, Pali Hill	19.055339	72.825511	0	Bakery	Bar	Café	Indian Restaurant	Event Space	Pizza Place	Wine Shop	German Restaurant	Ice Cream Shop

This data is prepared by using the ONE-HOT encoding technique. The list of all the neighbourhoods and the frequency of visits to the places in the area were obtained and the mean was taken. This was converted to the data-frame.

	Neighborhood	Women's Store	American Restaurant	Antique Shop	Arcade	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bar	Bed & Breakfast	Beer Garden	Bike Rental / Bike Share	Bookstore	Breakfast Spot
0	Andheri,Western Suburbs	0.000000	0.000000	0.0	0.000000	0.000000	0.142857	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000
1	Antop Hill,South Mumbai	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000
2	Bandra,Western Suburbs	0.000000	0.000000	0.0	0.000000	0.031250	0.000000	0.0	0.125000	0.093750	0.000000	0.03125	0.0	0.000000	0.000000
3	Byculla,South Mumbai	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000
4	Eastern Suburbs	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.142857	0.000000	0.000000	0.00000	0.0	0.000000	0.000000
5	Fort,South Mumbai	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000
6	Ghatkopar,Eastern Suburbs	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.125000	0.000000	0.000000	0.00000	0.0	0.000000	0.000000

The frequency analysis for each venue was done as shown below:

```
----Powai,Eastern Suburbs----
                venue  freq
0     Indian Restaurant  0.17
1  Fast Food Restaurant  0.07
2                   Bar  0.05
3                 Café  0.05
4           Restaurant  0.03
5     Department Store  0.03
6    Italian Restaurant  0.03
7         Shopping Mall  0.03
8    Chinese Restaurant  0.03
9                  Park  0.03


----Sanctacruz,Western Suburbs----
                      venue  freq
0          Indian Restaurant  0.13
1              Women's Store  0.09
2                     Market  0.09
3            Sandwich Place  0.04
4             Jewelry Store  0.04
5                    Lounge  0.04
6                       Gym  0.04
7  Middle Eastern Restaurant  0.04
8    Furniture / Home Store  0.04
9                Food Truck  0.04


----South Mumbai----
                        venue  freq
0                         Gym   0.2
```

Along with the above analysis, there were six neighbourhoods identified in which there were no significant venues to be considered among the top ten places. Hence, those areas or boroughs were considered as outliers and not included in the analysis.

4. Predictive Modelling

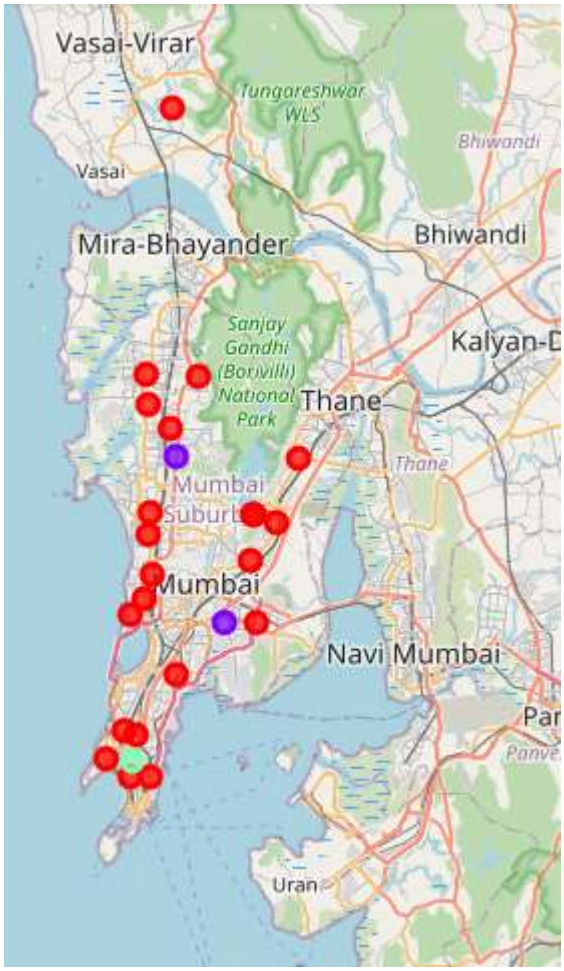
4.1 Clustering

The analysis of the dataset needed an unsupervised classification problem. K-Means Clustering was used to classify the areas as ones to live or buy a house and those to avoid. Since the city is developed, the clustering techniques yielded those selected areas where one should prefer to not buy the house.

K-Means clustering technique is a popular technique to determine data points with common attributes. In our scenario, the clustering of the areas have to be done based on the venues visited most often. The result would be those places which are similar in localities. In such case, those places which are home oriented would have departmental and convenience stores in the top venues visited. On the other hand, clusters which would be a prime entertainment spot or a commercial spot would have restaurants, theatres, bars and other stores as the top visiting venues.

4.2 Plotting

Maps were used to plot the clusters obtained using the folium library.



In the map presented, the three clusters are highlighted with the three dufferent coloured circles.

4.3 Analysis

The clusters obtained were as:

First Cluster:

mumbai_cluster_0

	Location	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	Andheri,Western Suburbs	0	Vegetarian / Vegan Restaurant	Train Station	Indian Restaurant	College Cafeteria	Athletics & Sports	Chinese Restaurant	Falafel Restaurant	Fish & Chips Shop	Department Store	Dessert Shop
1	Western Suburbs	0	Indian Restaurant	Chinese Restaurant	Bar	Fast Food Restaurant	Gym	Restaurant	Grocery Store	Italian Restaurant	Department Store	Coffee Shop
3	Bandra,Western Suburbs	0	Bakery	Bar	Café	Indian Restaurant	Event Space	Pizza Place	Wine Shop	German Restaurant	Ice Cream Shop	Fast Food Restaurant
6	Kandivali West,Western Suburbs	0	Indian Restaurant	Dessert Shop	Chinese Restaurant	Bike Rental / Bike Share	Fast Food Restaurant	Wine Shop	Flea Market	Department Store	Dhaba	Diner
7	Kandivali East,Western Suburbs	0	Ice Cream Shop	Indian Restaurant	Pizza Place	Coffee Shop	Fast Food Restaurant	Juice Bar	Café	Japanese Restaurant	Residential Building (Apartment / Condo)	Restaurant
8	Khar,Western Suburbs	0	Bar	Indian Restaurant	Dessert Shop	Café	Asian Restaurant	Seafood Restaurant	Salad Place	Cocktail Bar	Lounge	Fish & Chips Shop
9	Malad,Western Suburbs	0	Vegetarian / Vegan Restaurant	Pizza Place	Indian Restaurant	Snack Place	Wine Shop	Fish & Chips Shop	Deli / Bodega	Department Store	Dessert Shop	Dhaba

Second Cluster:

mumbai_cluster_1

	Location	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
5	Goregaon,Western Suburbs	1	Indian Restaurant	Hotel	Sandwich Place	Lounge	Fast Food Restaurant	Chinese Restaurant	Café	Gym / Fitness Center	Bakery	Bar
19	Govandi,Harbour Suburbs	1	Paper / Office Supplies Store	Gastropub	Electronics Store	Diner	Pool	Smoke Shop	Coffee Shop	Bar	General Entertainment	Indian Restaurant

Third Cluster:

mumbai_cluster_2

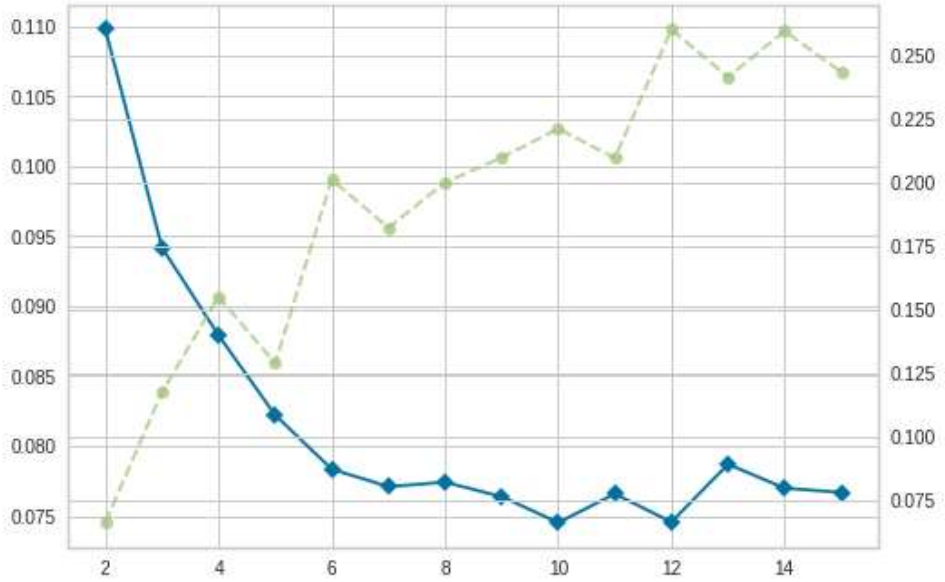
	Location	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
28	Kamathipura,South Mumbai	2	Indian Restaurant	Fried Chicken Joint	Dessert Shop	Restaurant	Breakfast Spot	BBQ Joint	Flea Market	Ice Cream Shop	Antique Shop	Dhaba

The first cluster contained those areas where one can seek to live or purchase a new residential property. The second and third clusters are those which cannot be deemed as the better residential property areas.

5. Results

The K-Means Clustering algorithm implemented was tested at various levels of K and the elbow plot with the 'silhouette' metric was created.

```
/usr/local/lib/python3.6/dist-packages/sklearn/base.py:197: FutureWarning: From versic
  FutureWarning)
KElbowVisualizer(ax=<matplotlib.axes._subplots.AxesSubplot object at 0x7f56a0b45860>,
                 k=None, metric=None, model=None, timings=True)
```



The plot helped determine the value of K at K=3.

6. Conclusion

To conclude our analysis, we can identify the neighbourhoods in which one should opt to live and one should not opt to live in Mumbai.

Cluster 1: It shows the areas where people opt to live. It consists of 25 areas, consisting of multiple neighbourhoods. We could observe the existence of departmental stores, convenience stores, women's store, gym, stations, grocery stores along with the restaurants and cafes.

Cluster 2, 3 and other 6 Neighbourhoods: These localities are mainly commercial and entertainment areas. It would not be ideal for one to live in such areas. These areas are mostly the ones consisting of restaurants, bars, flea market, smoking and bbq joints as well as offices and commercial buildings. Using K-Means clustering algorithm was beneficial for the unsupervised data where one could cluster the localities into those ideal for living and those which are not.

7. Discussion

One can back these results by combining it with the average price of the household prevailing in this area. Also, population density at night and daytime can be compared. These analysis would further extrapolate the accuracy of the analysis and back it up with statistical significance.

8. References

IBM Professional Data Science Certificate – Coursera
StackOverflow