DETERMINING THE PLACES FOR ACCOMODATION IN MUMBAI

Rohan Shukla
As a part of Applied Data Science Capstone-IBM

Introduction

- Mumbai is one of the most vibrant cities of India. As of 2018, Mumbai was the most populous city of India.
- With around twenty million population, everything gets crowded
- 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.

Target Audience

- 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.
- People could also use this to set or estimate the prices according to the luxuries and necessities present in the neighbourhood.

Database Used

- https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai
- 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.
- The data was cleaned and preprocessed. All the boroughs were combined. The neighbourhoods were combined, and listed as a single comma separated attribute.

Data 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.
-

	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

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

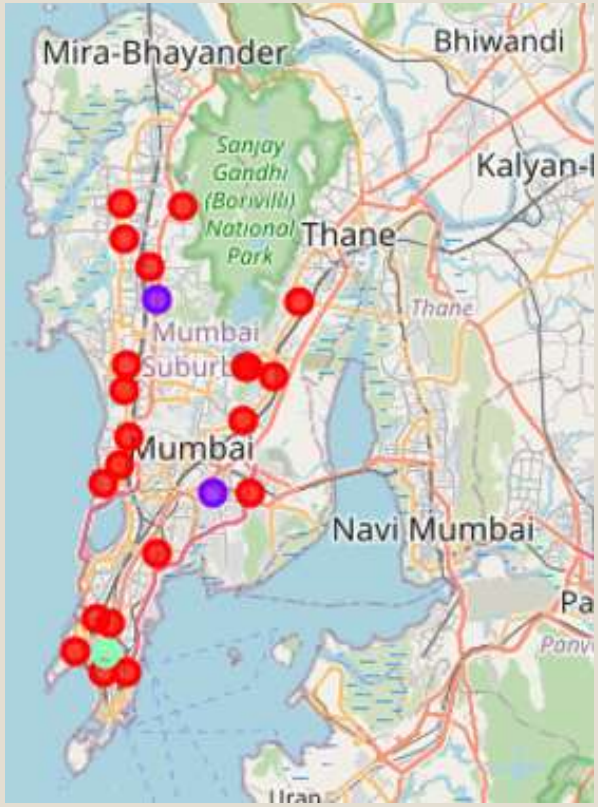
Outliers

- 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

Clustering

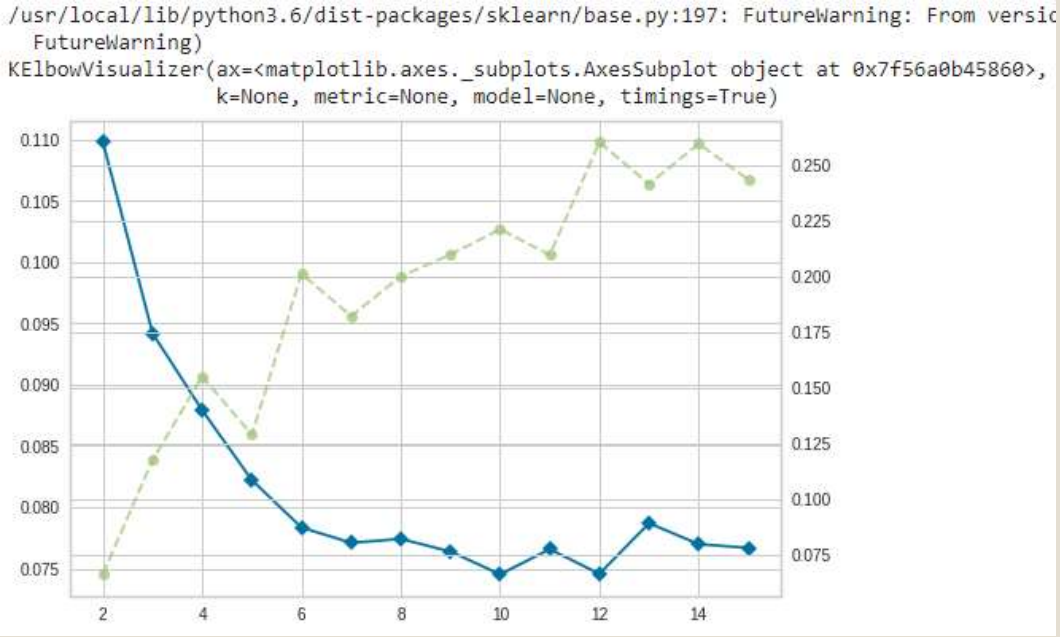
- K-Means Clustering was used to classify the areas as ones to live or buy a house and those to avoid.
- Since the city is developed, the clustering techniques yielded those selected areas where one should prefer to not buy the house.

Plotting the clusters on the Map



Analysis

- The K-Means Clustering algorithm implemented was tested at various levels of K and the elbow plot with the ‘silhouette’ metric was created. Optimal value at K=3



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.

Further Scope

- 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.

THANK YOU