COURSERA-IBM DATA SCIENCE

Capstone Project

by

Nishad Thakur

City Of Dreams Mumbai

Introduction

Mumbai, the city of dreams, is also the financial capital of India. Anyone who has tried looking for houses in the city knows how difficult and tormenting the task is. This is also true for people looking to pursue different entrepreneurial openings in the city's already saturated markets. Our aim for this article is to analyze and cluster the different neighborhoods of Mumbai city based upon a variety of factors.

This would be helpful for anyone and everyone looking for compatible neighborhoods while finding new homes

Problem Statement: Exploring and Clustering the neighborhoods of Mumbai city on the basis of several factors in order to find similarity among them thus helping anyone looking to find homes.

Target Audience: Anyone out in the market looking for homes can be aided through this project.It can also help real estate developers and investors to find facilities lacking in the area which can be improved upon so as to make them more attractive to potential customers.

Data

For a list of neighborhoods along with their rental prices we scrape data from the following website: www.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

This website contains a list of different neighborhoods in Mumbai, along with the Latitude and longitude position of that Area.

Other sources of data include the **Foursquare API** which is used to find out the most common venues in each neighborhood and the **HEREMAPS API** which helps us discover the services and facilities available in the locality.

	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
...	...	...	...	...
88	Parel	South Mumbai	18.990000	72.840000
89	Gowalia Tank	Tardeo,South Mumbai	18.962450	72.809703
90	Dava Bazaar	South Mumbai	18.946882	72.831362
91	Dharavi	Mumbai	19.040208	72.850850
92	Thane	Mumbai	19.200000	72.970000

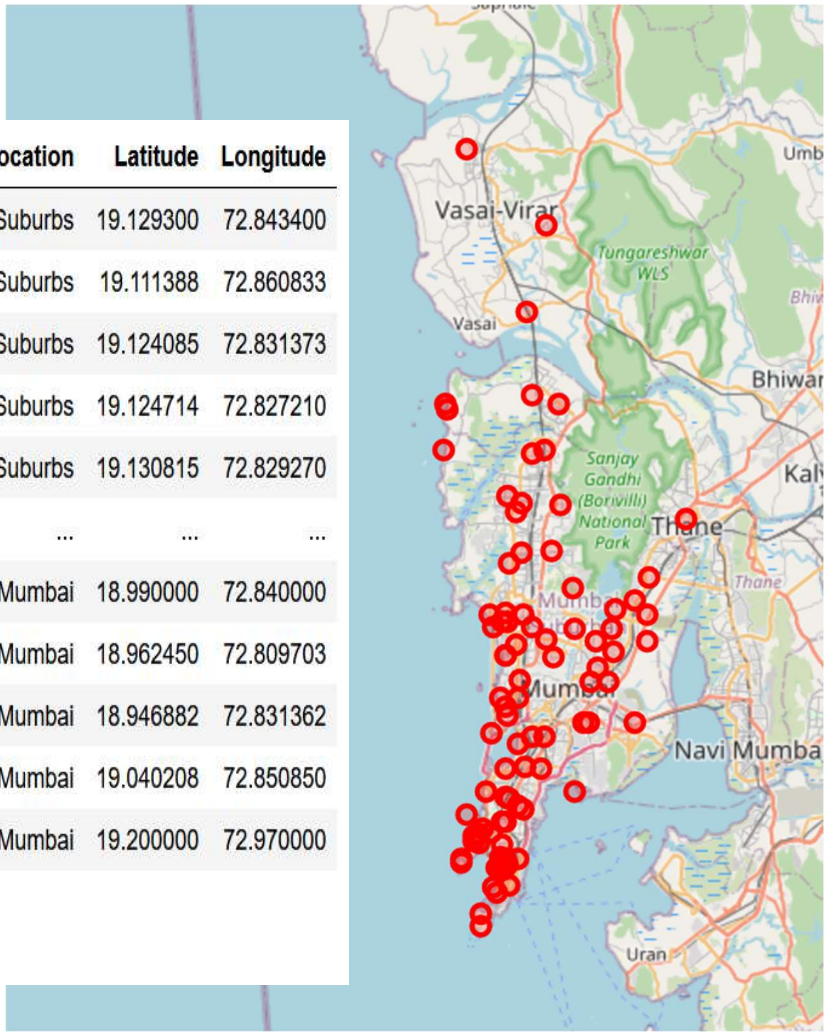
93 rows × 4 columns

Methodology

Web scrapping done using the pandas I first scrape the data from the website mentioned in the Data Section. It has 1 table on different localities of Mumbai and Using a geocoding library (Nominatim), I plotted the dataset on the map

	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
...	...	...	...	...
88	Parel	South Mumbai	18.990000	72.840000
89	Gowalia Tank	Tardeo,South Mumbai	18.962450	72.809703
90	Dava Bazaar	South Mumbai	18.946882	72.831362
91	Dharavi	Mumbai	19.040208	72.850850
92	Thane	Mumbai	19.200000	72.970000

93 rows × 4 columns



With this, we now work towards finding the most common venues around each neighborhood. For this task we employ the Foursquare API. We create the API query URL and make the GET request for each locality, finding venues and their corresponding categories in a 1km radius.

```
venues_in_Mumbai.groupby('Neighborhood').head()
```

:

	Neighborhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Amboli	19.1293	72.8434	Shawarma Factory	Falafel Restaurant
1	Amboli	19.1293	72.8434	Cafe Arfa	Indian Restaurant
2	Amboli	19.1293	72.8434	Jaffer Bhai's Delhi Darbar	Mughlai Restaurant
3	Amboli	19.1293	72.8434	5 Spice , Bandra	Chinese Restaurant
4	Amboli	19.1293	72.8434	Pizza Express	Pizza Place
...	...	...	...	...	...
2089	Thane	19.2000	72.9700	Mad Over Donuts	Donut Shop
2090	Thane	19.2000	72.9700	Starbucks	Coffee Shop
2091	Thane	19.2000	72.9700	Food Court	Food Court
2092	Thane	19.2000	72.9700	Korum Mall	Shopping Mall
2093	Thane	19.2000	72.9700	Café Coffee Day	Coffee Shop

441 rows × 5 columns

One hot encoding is then done to create a new dataframe to show all the unique venue categories and whether they are near the locality or not. The dataframe produced is as follows:

	Neighbor	Afghan Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Arcade	Art Gallery	Asian Restaurant	...	Toy / Game Store	Track	Trail	Train	Train Station	Vegetarian / Vegan Restaurant	Wine Bar	Wine Shop
0	Aarey Milk Colony	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
1	Agripada	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.038462	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
2	Altamount Road	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.033333	0.000000	0.0
3	Amboli	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.066667	...	0.0	0.0	0.0	0.0	0.033333	0.033333	0.000000	0.0
4	Amrut Nagar	0.033333	0.0	0.0	0.0	0.033333	0.0	0.0	0.000000	0.033333	...	0.0	0.0	0.0	0.0	0.000000	0.033333	0.000000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
85	Vikhroli	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
86	Vile Parle	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
87	Virar	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
88	Walkeshwar	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0
89	Worli	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.033333	0.033333	...	0.0	0.0	0.0	0.0	0.000000	0.000000	0.033333	0.0

90 rows × 188 columns

To find out how frequently a venue category comes up for a neighborhood we take the mean value of all columns for each neighborhood. The resultant columns for each neighborhood can be then sorted for finding the top ten common venues in each neighborhood. We can then merge our thus obtained dataframe with original data frame This is shown 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	Aarey Milk Colony	Resort	Café	Hotel	Gym / Fitness Center	Indian Restaurant	Farm	Golf Course	Restaurant	Modern European Restaurant	Monument / Landmark
1	Agripada	Indian Restaurant	Bakery	Gym	Coffee Shop	History Museum	Club House	Tea Room	Platform	Pizza Place	Pharmacy
2	Altamount Road	Bakery	Café	Chinese Restaurant	History Museum	Bookstore	Brewery	Sandwich Place	Salon / Barbershop	Restaurant	Deli / Bodega
3	Amboli	Indian Restaurant	Bar	Pizza Place	Asian Restaurant	Pub	Chinese Restaurant	Bowling Alley	Burger Joint	Mughlai Restaurant	Snack Place
4	Amrut Nagar	Indian Restaurant	Lounge	Clothing Store	Fast Food Restaurant	Diner	Afghan Restaurant	Coffee Shop	Brewery	Shopping Mall	Café

We now begin to find the number of amenities in each neighborhood. For this we use the HERE API which provides a useful free-text query feature. We define a function which contains the API query URL and makes the GET request to find out the different facilities available within a 1km radius of the neighborhood.

The number of amenities thus found are merged with their corresponding neighborhoods. The nearby amenities we search for include: Hospitals, Schools, Leisure facilities, Shopping facilities, Emergency services,Banks and Cinemas. One can give importance to any of these categories based upon their needs.

The obtained dataframe is merged with the venues dataframe. This accounts for all the data we require for our analysis. We check for any columns with unwanted data types and change them accordingly.

The dataframe produced can also be used for future investigations.

	Area	Latitude	Longitude	Hospitals	Schools	Emergency Services	Leisure	Shopping Facilities	Banks	Cinemas
0	Amboli	19.129300	72.843400	30	100	13	58	100	100	15
1	Chakala, Andheri	19.111388	72.860833	16	87	10	39	100	100	10
2	D.N. Nagar	19.124085	72.831373	22	100	9	61	100	100	22
3	Four Bungalows	19.124714	72.827210	17	89	8	60	100	100	19
4	Lokhandwala	19.130815	72.829270	19	96	8	70	100	100	20

click to scroll output; double click to hide

	Neighborhood	Latitude	Longitude	Hospitals	Schools	Emergency Services	Leisure	Shopping Facilities	Banks	Cinemas	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Mc Comm Ven
0	Amboli	19.129300	72.843400	30	100	13	58	100	100	15	Indian Restaurant	Bar	Pizza Place	Asian Restaurant	Pub	Chine Restaura
1	Chakala, Andheri	19.111388	72.860833	16	87	10	39	100	100	10	Hotel	Café	Seafood Restaurant	Indian Restaurant	Chinese Restaurant	Fast Fo Restaura
2	D.N. Nagar	19.124085	72.831373	22	100	9	61	100	100	22	Bar	Pub	Pizza Place	Vegetarian / Vegan Restaurant	Gym / Fitness Center	Ca
3	Four Bungalows	19.124714	72.827210	17	89	8	60	100	100	19	Pub	Café	Vegetarian / Vegan Restaurant	Pizza Place	Coffee Shop	Gyr Fitne Cen
4	Lokhandwala	19.130815	72.829270	19	96	8	70	100	100	20	Pub	Italian Restaurant	Café	Multiplex	Lounge	Juice B

We now begin to mold our dataframe so that we can apply KMeans clustering method on it.

We begin by scaling all our numerical data columns for better results.

For clustering purpose we use the dataframe that was obtained by finding the mean of the one-hot encoded values of the venues and append the scaled numerical columns to it.

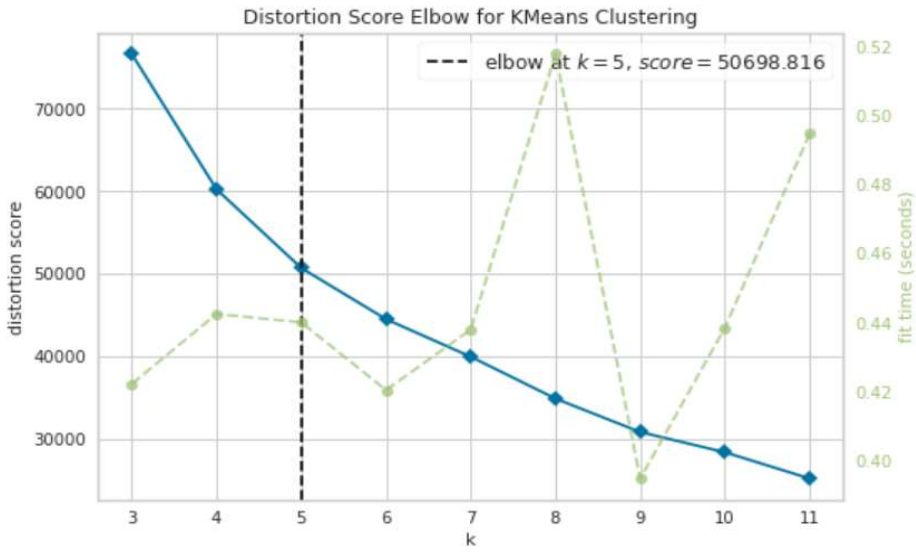
	Afghan Restaurant	Airport	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Arcade	Art Gallery	Asian Restaurant	Athletics & Sports	...	Wine Shop	Women's Store	Zoo	Hospitals	Schools	Emergency Services	Leisure	S I
0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.000000	30	100	13	58	
1	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.038462	0.0	...	0.0	0.0	0.038462	16	87	10	39	
2	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.000000	22	100	9	61	
3	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.066667	0.0	...	0.0	0.0	0.000000	17	89	8	60	
4	0.033333	0.0	0.0	0.0	0.033333	0.0	0.0	0.0	0.033333	0.0	...	0.0	0.0	0.000000	19	96	8	70	

5 rows × 194 columns

The KMeans method takes an argument specifying the number of clusters we want to divide out data into. To find out the ideal number of clusters we employ the Elbow method. The result of the Elbow method:

```
model = KMeans()
visualize = KElbowVisualizer(model, k=(3,12)) # Check for optimal clusters between 3-12

visualize.fit(mumbai_clustering)         # Fit the data
visualize.show()
```

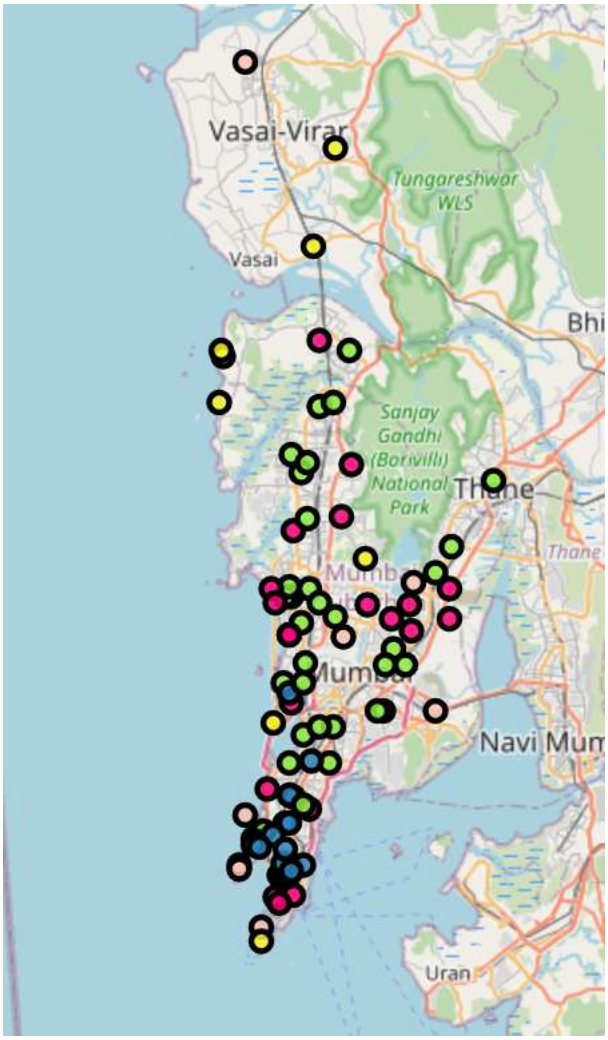


The ideal value is found to be k=5 (k is the number of clusters). Therefore, we apply the KMeans method with k=5 and divide the neighborhoods into seven cluster. The cluster label of each neighborhood is attached to the original dataframe.

	Cluster Labels	Neighborhood	Latitude	Longitude	Hospitals	Schools	Emergency Services	Leisure	Shopping Facilities	Banks	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Com V
0	3	Amboli	19.129300	72.843400	30	100	13	58	100	100	...	Indian Restaurant	Bar	Pizza Place	Asian Restaurant	Pub	Chi Resta
1	3	Chakala, Andheri	19.111388	72.860833	16	87	10	39	100	100	...	Hotel	Café	Seafood Restaurant	Indian Restaurant	Chinese Restaurant	Fast Resta
2	3	D.N. Nagar	19.124085	72.831373	22	100	9	61	100	100	...	Bar	Pub	Pizza Place	Vegetarian / Vegan Restaurant	Gym / Fitness Center	
3	3	Four Bungalows	19.124714	72.827210	17	89	8	60	100	100	...	Pub	Café	Vegetarian / Vegan Restaurant	Pizza Place	Coffee Shop	C Fi C
4	3	Lokhandwala	19.130815	72.829270	19	96	8	70	100	100	...	Pub	Italian Restaurant	Café	Multiplex	Lounge	Juic

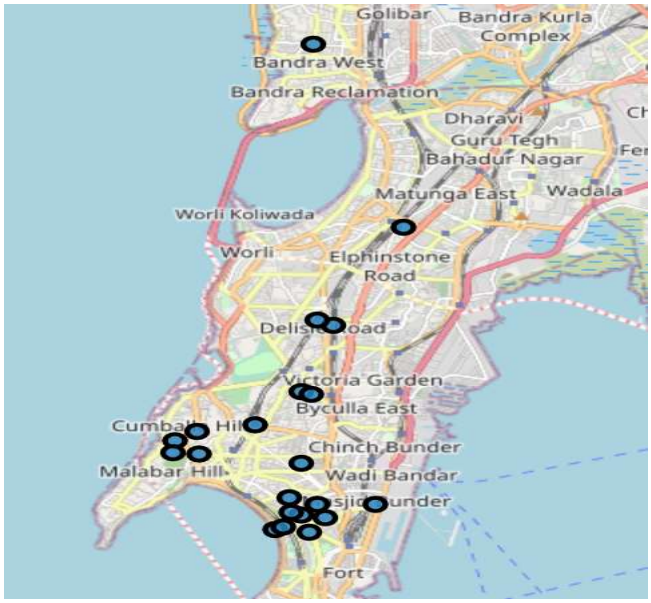
Results

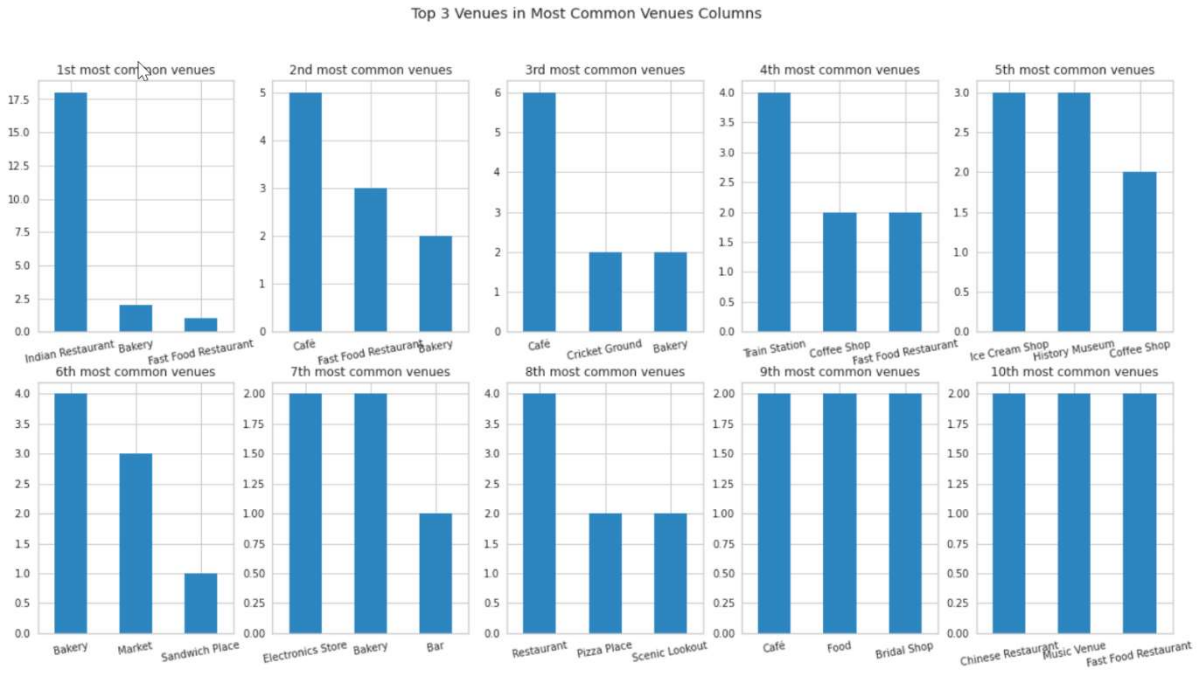
After performing the clustering we obtained the above dataframe. It is then used to create a map of Mumbai with different clusters distinguished by different colors.



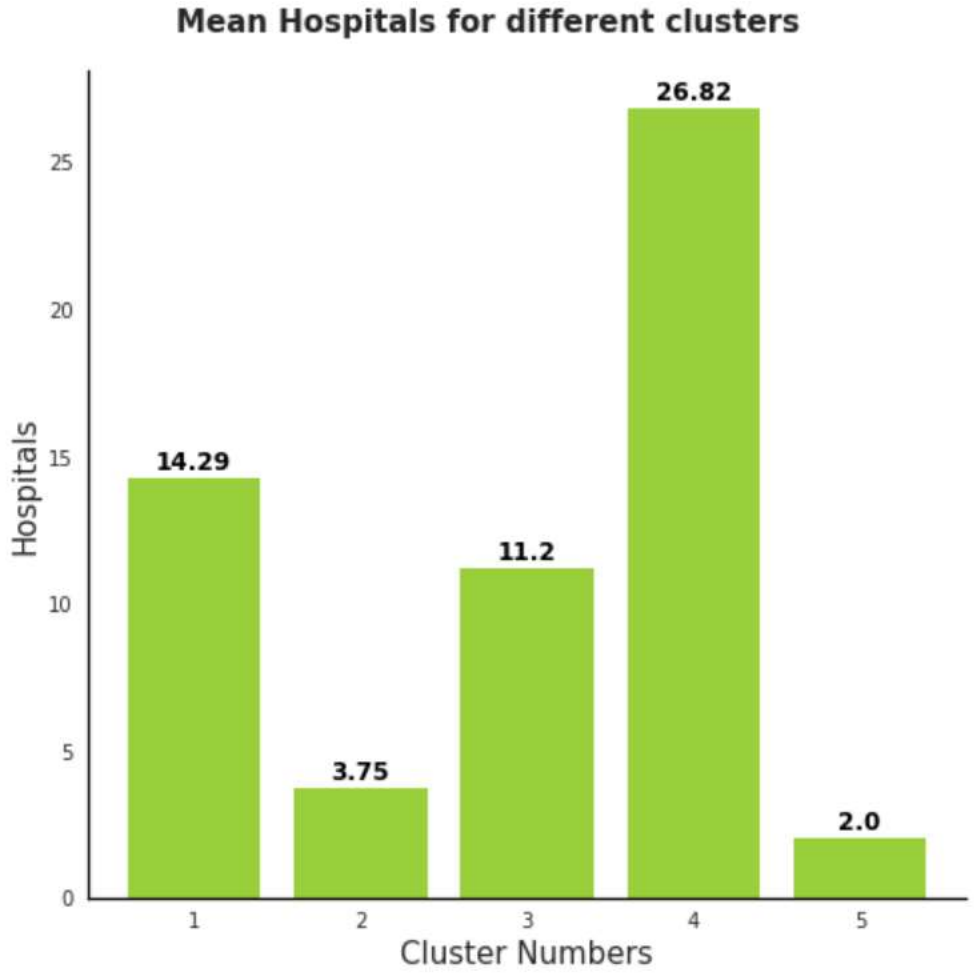
We then analyse each cluster. For each cluster we first print its details then visualize it on the map. We then plot 10 bar graphs, for each venue column, demonstrating the top 3 common venues in each of the most common venues column of the cluster. This helps us understand the what are the popular trends in the neighborhood. This plot is viable for both business owners and resident, helping them understand their surroundings

Results for one such cluster, Cluster 1, is as follows:





Finally, we plot the average value of each amenity of each cluster against each other. This done to compare the different neighborhood clusters and find out their differences and similarities. Example of one such comparison:



Each cluster is analyzed to better understand the factors leading to the resemblance of their constituent neighborhoods. The complete analysis of all the clusters can be obtained from the notebook.

Discussion

All the clusters are completely saturated with shopping facilities except cluster 5. Therefore, it may seem a viable location for up and coming shopping complexes.

A large number of schools and hospitals are located in specific clusters, while there seems to be a lack of them in other neighborhoods.

Cinema Lovers should check the properties in Cluster 1

Indian Restaurants are spread all across Mumbai and are the most common places across all the clusters.

In the end, it all boils down to an individual's preferences and choices. Using the above investigations one can interpret the results to fit their needs and find the best possible solutions for their situation.

Conclusion

Through the above analysis we have found the trends and preferences of different parts of Mumbai city and have successfully divided into 7 clusters. Each cluster is formed on the basis of numerous factors.

Other preferences can also be added to this analysis if desired by an individual.

This exploration would help house-hunters find suitable neighborhoods according to their needs and also help entrepreneurs and owners to find out those neighborhoods and businesses investments in which would yield maximum returns in the future.

References

1. www.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai
2. FoursquareAPI
3. Heremaps API
4. Mapping Your Favorite Coffee Shop using Google Places API and Folium
5. A tale of Two Cities