COURSERA-IBM DATA SCIENCE SPECIALIZATION

Capstone Project by Utkarsh Lath

Analyzing Mumbai Neighborhoods

Introduction

The City of Mumbai, also known as the City of Dreams is the financial capital of India. In recent times, the price of real estate in the city has sky-rocketed while income equality has plummeted. This has made house hunting a significantly greater challenge than it already was.

In this project therefore, we try to ease this process

by comparing the different neighborhood of Mumbai on the basis of a variety of factors.

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.

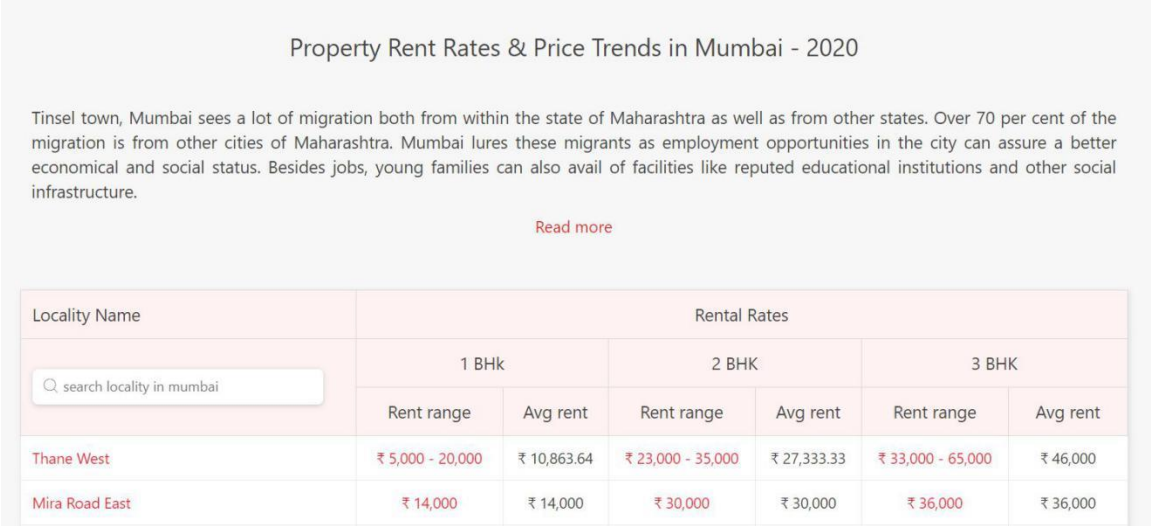
Target Audience: Anyone out in the market looking for homes can be aided through this project. It can also help real estate developers and investors to find facilities lacking in the area which can be

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:

https://www.makaan.com/price-trends/property-rates-for-rent-in-mumbai



Property Rent Rates & Price Trends in Mumbai - 2020

Tinsel town, Mumbai sees a lot of migration both from within the state of Maharashtra as well as from other states. Over 70 per cent of the migration is from other cities of Maharashtra. Mumbai lures these migrants as employment opportunities in the city can assure a better economical and social status. Besides jobs, young families can also avail of facilities like reputed educational institutions and other social infrastructure.

Read more

Locality Name	Rental Rates					
search locality in mumbai	1 BHk		2 BHK		3 BHK	
	Rent range	Avg rent	Rent range	Avg rent	Rent range	Avg rent
Thane West	₹ 5,000 - 20,000	₹ 10,863.64	₹ 23,000 - 35,000	₹ 27,333.33	₹ 33,000 - 65,000	₹ 46,000
Mira Road East	₹ 14,000	₹ 14,000	₹ 30,000	₹ 30,000	₹ 36,000	₹ 36,000

We use the column 'Avg Rent' for 2BHK flats as a feature to build our dataset. This is because 2BHK flats are most commonly rented and possess the maximum data on the website.

The coordinates of the different neighborhoods are found using the Nominatim library. Nominatim library is an Open-source geocoding

Technology which can find the coordinates of a place using the its address and vice versa.

Next, we use two different APIs, namely Foursquare API and HERE API to find out different venues and facilities available near each locality.

The Foursquare API is used to find out the most common venues in the neighborhood. The free-text query search feature available in the HERE API is used to find out the number of different amenities, essential while looking for houses, such as hospitals, schools and colleges, shopping services etc.

Methodology:

For conducting the analysis of different Mumbai neighborhoods, the main difficulty was found while preparing and cleaning the data.

We first scrape the data from the website mentioned in the Data Section. It has 102 web pages of Rent Data in different localities of Mumbai. We choose the 2BHK rent column for our analysis as it the most popular form of housing in Mumbai. Most of the web pages at the end contain no data, they have to be removed from our analysis.

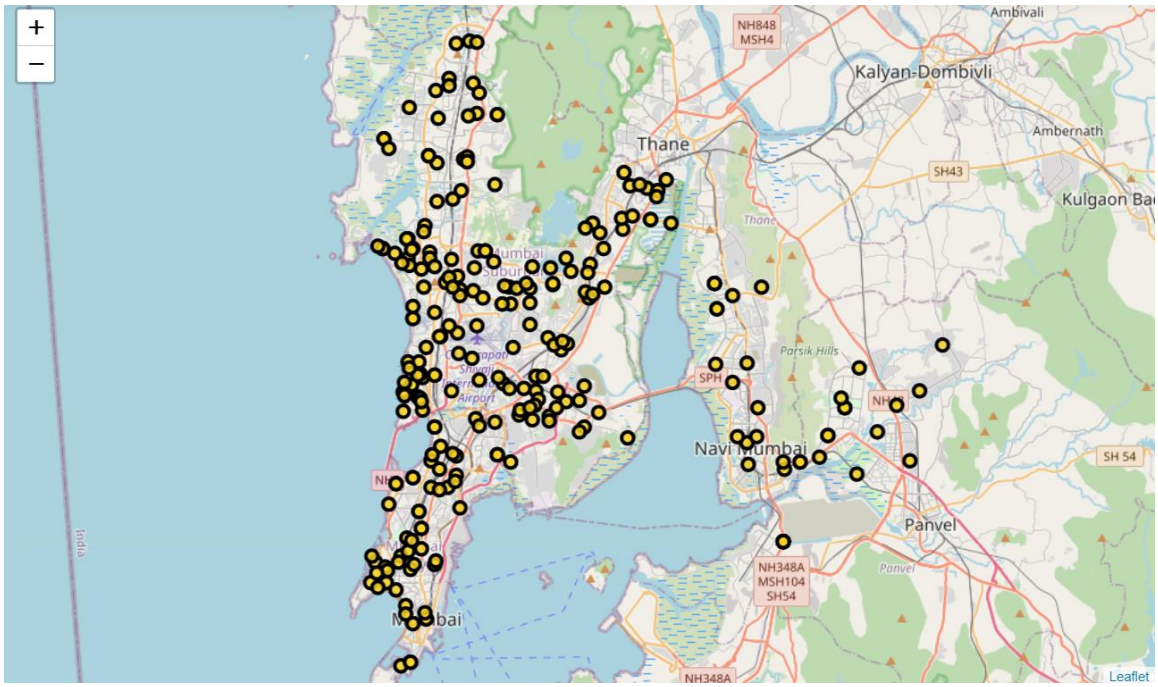
After loading the data and removing the missing values, the following dataframe is obtained:

Neighborhood	Avg 2BHK Rent
CENTRAL PARK	50,000
Chuim Village	40,000
Eastern Express Highway	25,000
PL Lokhande Marg	30,000
Sector-4 Kopar Khairane	21,000

After obtaining this dataframe, we add columns for each neighborhood's latitude and longitude coordinates using the Nominatim library. Those neighborhoods whose coordinates are unavailable are removed. Duplicate neighborhoods were also removed. We now obtain the following:

	Neighborhood	Avg 2BHK Rent	Latitude	Longitude
0	Thane West	27,333.33	19.175020	72.971802
1	Mira Road East	30,000	19.187896	72.836596
2	Kharghar	32,000	19.025773	73.059185
3	Chembur	39,000	19.061213	72.897591
4	Kandivali East	33,694.62	19.210381	72.864084
5	Powai	85,000	19.118720	72.907348
6	Goregaon East	30,000	19.169262	72.855255
8	Ulwe	10,000	18.980436	73.038731
9	Andheri East	22,000	19.115883	72.854202
10	Virar	8,000	19.467682	72.887997

Using this dataframe and the Folium library, we plot the different neighborhoods in the dataset, on a map of Mumbai. The result:



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.

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	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Antique Shop	...	Track	Train	Train Station	Vegetarian / Vegan Restaurant	Whisky Bar	V
0	Thane West	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
1	Thane West	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
2	Thane West	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
3	Thane West	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	
4	Thane West	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 258 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 each neighborhood's Avg Rent. 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	Avg 2BHK Rent
0	114 Goregaon Mulund Link Road	Bus Station	Restaurant	Fast Food Restaurant	Pizza Place	Train Station	Electronics Store	Dim Sum Restaurant	Diner	Donut Shop	Dumpling Restaurant	44,000
1	14TH ROAD	Indian Restaurant	Café	Bar	Italian Restaurant	Bakery	Lounge	Seafood Restaurant	Pub	Asian Restaurant	Dessert Shop	77,750
2	15th Rd	Indian Restaurant	Bar	Bakery	Lounge	Seafood Restaurant	Fast Food Restaurant	Italian Restaurant	Café	Dessert Shop	Asian Restaurant	80,000
3	16th Cross Road	Indian Restaurant	Bakery	Bar	Dessert Shop	Lounge	Italian Restaurant	Café	Seafood Restaurant	Asian Restaurant	Cupcake Shop	95,000
4	16th Rd	Indian Restaurant	Coffee Shop	Italian Restaurant	Lounge	Bakery	Asian Restaurant	Clothing Store	Café	Gym / Fitness Center	Chinese Restaurant	68,000

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 and Spiritual centers. 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.

	Neighborhood	Latitude	Longitude	Hospitals	Schools	Emergency Services	Leisure	Shopping Facilities	Spiritual Centers	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th M Comn Ve
0	Thane West	19.175020	72.971802	4	50	7	27	100	54	Smoke Shop	Indian Restaurant	Dessert Shop	Border Crossing	Fast F Restau
1	Mira Road East	19.187896	72.836596	30	45	4	23	100	63	Coffee Shop	Indian Restaurant	Fast Food Restaurant	Café	I
2	Kharghar	19.025773	73.059185	9	8	0	7	83	10	Café	Fast Food Restaurant	Multiplex	Train Station	Departm St
3	Chembur	19.061213	72.897591	36	57	6	35	100	98	Indian Restaurant	Seafood Restaurant	Café	Ice Cream Shop	Pizza Pl
4	Kandivali East	19.210381	72.864084	30	18	3	19	100	31	Indian Restaurant	Fast Food Restaurant	Restaurant	Pizza Place	Shopp

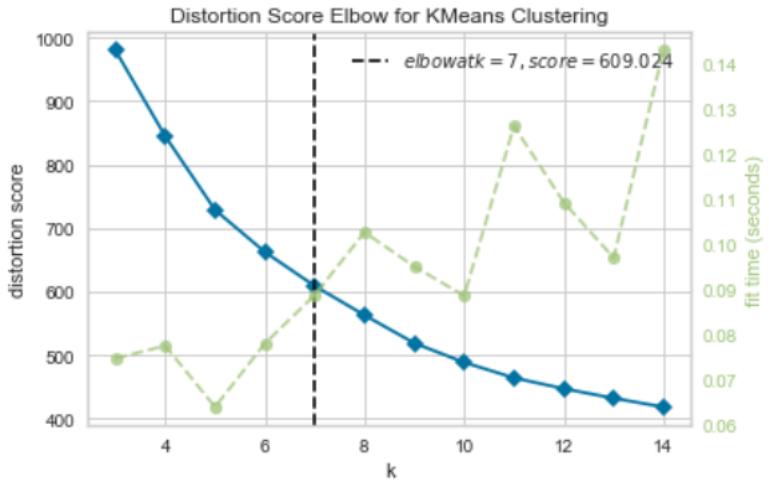
We now begin to mold our dataframe so that we can apply KMeans clustering method on it.

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.

	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	Airport	Airport Lounge	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Women's Store	Yoga Studio	Zoo	Avg 2BHK Rent	Hospitals
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	-0.647429	-0.911006
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	-0.562003	0.581206
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	-0.497932	-0.624042
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.0	0.0	-0.273687	0.925562
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.014706	0.0	0.0	-0.443645	0.581206

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:

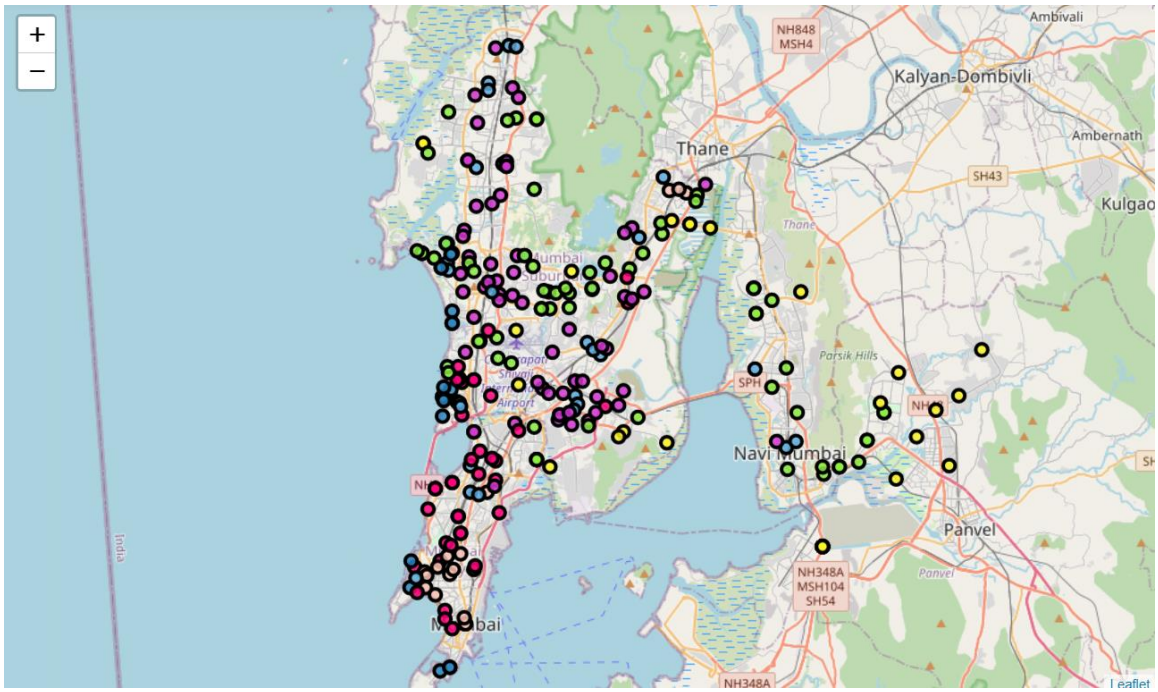


The ideal value is found to be k=7 (k is the number of clusters). Therefore, we apply the KMeans method with k=7 and divide the neighborhoods into seven cluster. The cluster label of each neighborhood is attached to the original dataframe.

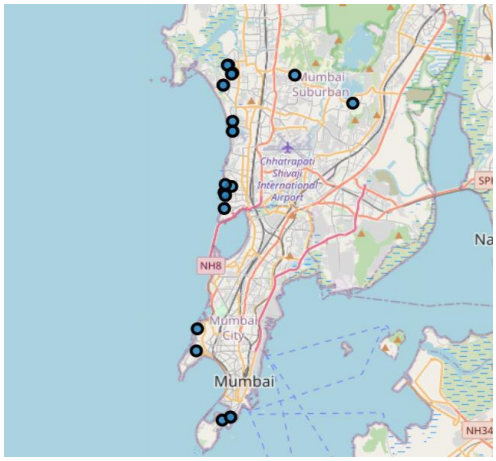
	Cluster Labels	Neighborhood	Latitude	Longitude	Avg 2BHK Rent	Hospitals	Schools	Emergency Services	Leisure	Shopping Facilities	...	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Com V
0	6	Thane West	19.175020	72.971802	27333.33	4	50	7	27	100	...	Smoke Shop	Indian Restaurant	Dessert Shop	B Cro
1	6	Mira Road East	19.187896	72.836596	30000.00	30	45	4	23	100	...	Coffee Shop	Indian Restaurant	Fast Food Restaurant	
2	3	Kharghar	19.025773	73.059185	32000.00	9	8	0	7	83	...	Café	Fast Food Restaurant	Multiplex	St
3	5	Chembur	19.061213	72.897591	39000.00	36	57	6	35	100	...	Indian Restaurant	Seafood Restaurant	Café	C
4	3	Kandivali East	19.210381	72.864084	33694.62	30	18	3	19	100	...	Indian Restaurant	Fast Food Restaurant	Restaurant	

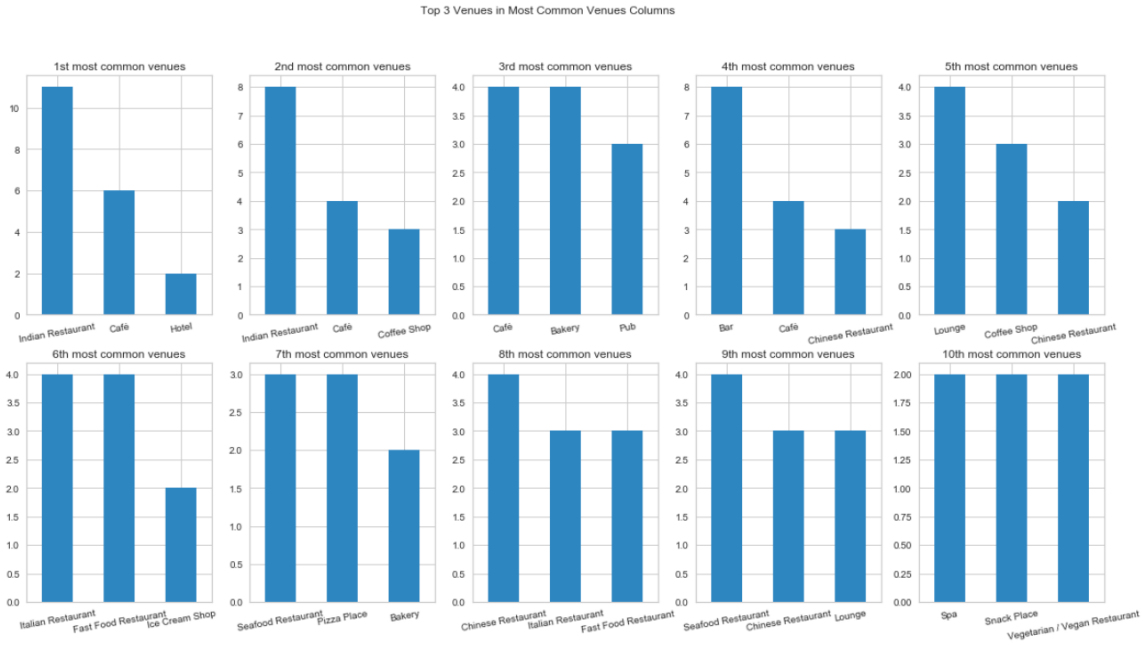
Results:

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.

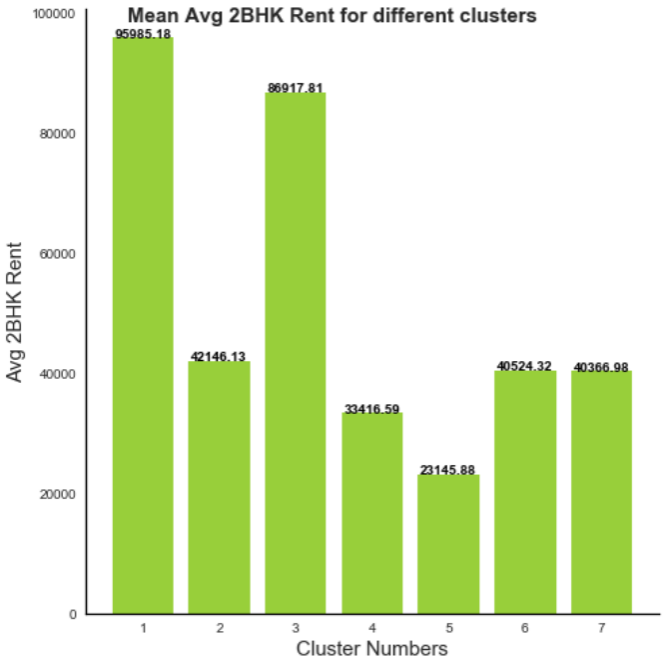


We then analyse each cluster. For each cluster we first print its details then visualize it on the map. We then plot 10 bar graphs, for each venue column, demonstrating the top 3 common venues in each of the most common venues column of the cluster. This helps us understand the what are the popular trends 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:

We find out that a number of different factors influence the distribution of neighborhoods across the city.

One interesting thing found was that more number of amenities didn't necessarily translate into higher rent neighborhood, i.e. neighborhoods with maximum rent did not necessarily have larger number of facilities available.

Indian Restaurants and Cafes were very popular across most of the clusters (except cluster 5) which may lead one to believe that they are profitable across Mumbai and maybe chosen as prospective business opportunities.

Also there are a lot of shopping complexes everywhere in Mumbai which perhaps has caused market saturation. But cluster 5 have shockingly low shopping facilities and it seems like a viable opportunity for business owners in those neighborhoods.

A lot of emergency services facilities are located in cluster 2 while educational services produce a mixed results across the clusters.

Since, cluster 5 seems to perform poorly in almost every department, it also has lower avg rent than other properties.

Several other facts and conclusions maybe drawn from the analysis. It can also be easily be fit around a users specific needs. These

conclusions and exploration can therefore prove to be extremely useful.

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.