**COURSERA-IBM DATA SCIENCE SPECIALIZATION**

**Capstone Project by Hiral K**

**Mumbai Neighborhood Analysis**

**Introduction**

Mumbai is one of the metro cities of India. It is also known as the financial capital of India. Mumbai is one of the most expensive city of the world in terms of real estate both residential and commercial. In addition to the real estate prices, the Mumbai is also famous for its heavy traffic conditions. Travelling in Mumbai from one location to another is difficult task.

This project explores different localities of Mumbai between Andheri to Dahisar locations for availability of restaurants and buy rates of residential properties. This project can help someone who is looking to open restaurant in locality between Andheri to Dahisar based on the number of restaurants in particular locality. Here I have assumed that person also wants to purchase a house in the same or nearby locality to avoid travelling considering the heavy traffic situations in Mumbai.

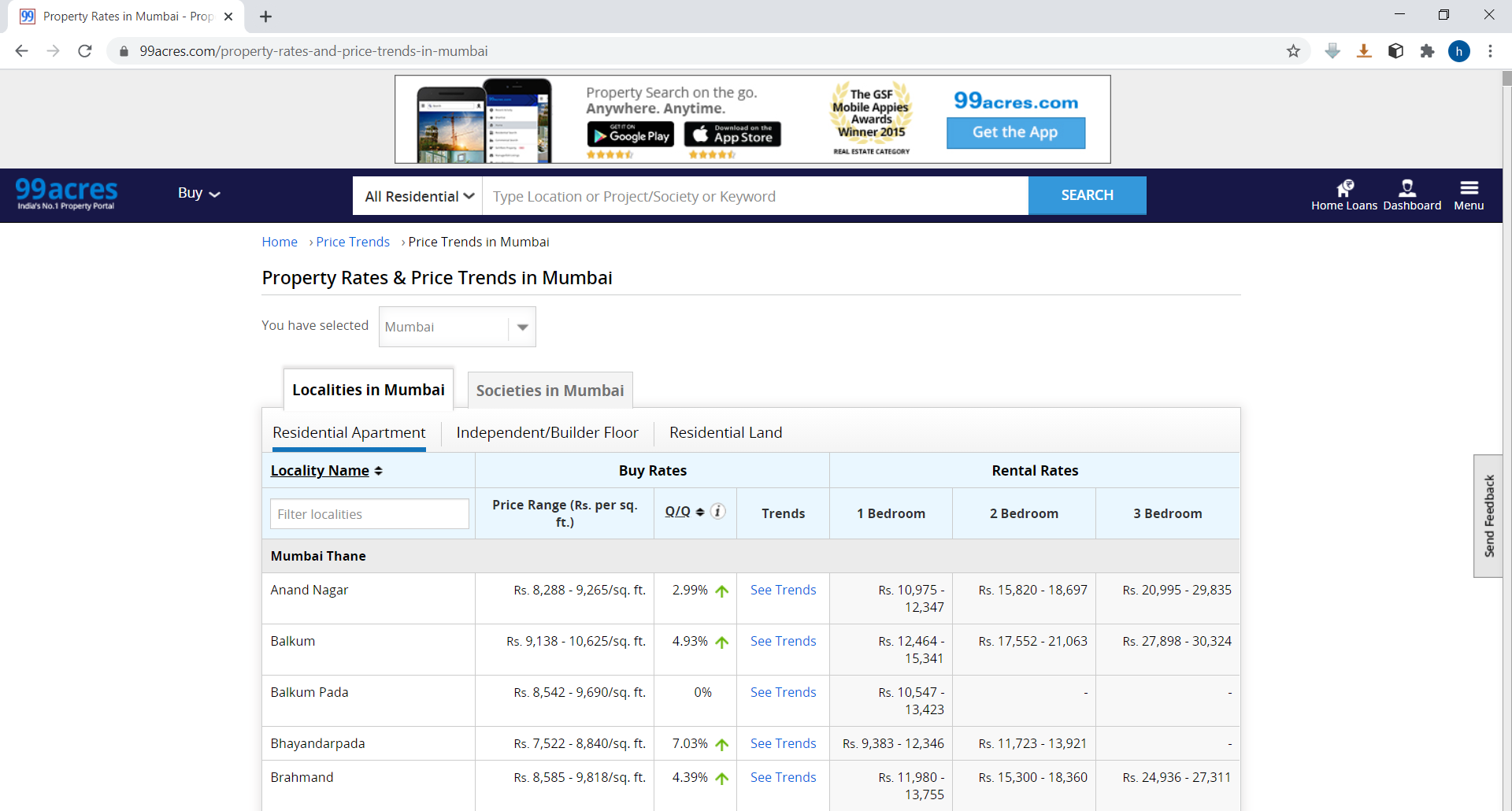
**Problem Statement:** Exploring the localities of the Mumbai from Andheri to Dahisar locations to find suitable locality to open a restaurant. Also exploring buy rates of residential properties in the same localities with the assumption that person wants to purchase house in the same locality where he/she opens a restaurant. Clustering the localities for better observation of number of restaurants and buy rates of houses.

**Target Audience:** Anyone who is looking to open restaurant and/or purchasing house between Andheri to Dahisar locations in Mumbai city. The project can also help real estate developers for new constructions and giving offers to buyers. Real estate agents can also use the system to suggest particular localities to the customers with the same need.

**Data:**

The main data required to build the system is details of localities between Andheri to Dahisar in Mumbai city along with their average buy rates. This data is scraped from 99acres.com website. The URL for the same is:

https://www.99acres.com/property-rates-and-price-trends-in-mumbai



As shown above, the table on the above website contains many features. We use name of localities and buy rates. The table is read into panda’s dataframe using read\_html. Then dataframe is processed to get required features (localities and average buy rates) in required format.

The coordinates of the different localities 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. Then we have used Foursquare API to find out different venues in each locality.

**Methodology:**

The main difficulty in the project was preparation and cleaning of data. The data is scraped from above mentioned website. The data is present for almost all locations of Mumbai. It also contains data for buy rates and rent rates. We are interested in buy rates of residential properties between Andheri and Dahisar locations.

After scraping the data and removing unwanted data from it, we got the following dataframe:

|  |  |  |
| --- | --- | --- |
|  | Locality | Rates |
| 0 | 4 Bunglows | Rs. 20,272 - 25,288/sq. ft. |
| 1 | Aarey Milk Colony | Rs. 7,565 - 8,840/sq. ft. |
| 2 | Alika Nagar | Rs. 13,642 - 14,918/sq. ft. |
| 3 | Amboli | Rs. 17,255 - 19,465/sq. ft. |
| 4 | Andheri (East) | Rs. 15,512 - 17,935/sq. ft. |

As shown in above datframe the rates are in range format that we need to convert in average rates. After performing some string format operations and changing the datatype we get rates in average rate format. Next we need latitude and longitude for all localities. We have used Nominatim library to find the coordinates of all localities. The localities whose coordinates are null are removed. Also duplicates values are removed and we get following dataframe:

|  |  |  |  |
| --- | --- | --- | --- |
| **Locality** | **Rates** | **Latitude** | **Longitude** |
| 4 Bunglows | 22780.0 | 19.044471 | 72.910060 |
| Aarey Milk Colony | 8202.5 | 19.156129 | 72.870722 |
| Amboli | 18360.0 | 19.131992 | 72.849960 |
| Andheri (East) | 16723.5 | 19.115883 | 72.854202 |
| Andheri (West) | 22015.0 | 19.117249 | 72.833968 |

Next we plot Localities of Mumbai on map using above dataframe and Folium library. The next task is to find the venues near each locality. We have used FourSquare API for that. We can create developer account and get credentials for FourSquare API. Using this credentials we have to create request URL and get the response. We will also get the latitude, longitude and category for all venues. After that we get following dataframe:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Locality** | **Rates** | **Latitude** | **Longitude** | **Venue** | **Venue Latitude** | **Venue Longitude** | **Venue Category** |
| 4 Bunglows | 22780.0 | 19.044471 | 72.91006 | n Bar & Grill | 19.046523 | 72.908267 | Lounge |
| 4 Bunglows | 8202.5 | 19.044471 | 72.91006 | Shivaji Chowk | 19.045973 | 72.909093 | Plaza |
| 4 Bunglows | 16723.5 | 19.044471 | 72.91006 | Natraj Lawns | 19.046134 | 72.908068 | Garden |
| 4 Bunglows | 18360.0 | 19.044471 | 72.91006 | The bar stock exchange | 19.046546 | 72.908193 | Lounge |

Next we have performed one hot encoding of above dataframe that will help us to find average occurrences of the venues by locality. Then we take mean values of all columns based on localities. Next we will find the localities based where number of restaurants are above threshold value. And we get the following dataframe:

|  |  |
| --- | --- |
| **Locality** | **Restaurant** |
| Aram Nagar | 0.090909 |
| Asha Nagar | 0.166667 |
| Borivali (West) | 0.117647 |
| Chikuwadi | 0.037037 |
| Chincholi Bunder | 0.045455 |

Next we will use KMeans algorithm to find clusters of localities based on number of restaurants. We will analyse these clusters along with buy rates of properties in the clusters. We have used ElbowVisualizer method to find the best value of K. The best value of K found is 5. So we have used K=5. After applying KMeans algorithm we have found cluster labels for all localities. We got following dataframe with cluster lables:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Locality** | **Restaurant** | **Cluster Labels** | **Rates** | **Latitude** | **Longitude** |
| 4 Bunglows | 0.0 | 0 | 22780.0 | 19.044471 | 72.910060 |
| Aarey Milk Colony | 0.0 | 0 | 8202.5 | 19.156129 | 72.870722 |
| Amboli | 0.0 | 0 | 18360.0 | 19.131992 | 72.849960 |
| Andheri (East) | 0.0 | 0 | 16723.5 | 19.115883 | 72.854202 |

**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 analyse all cluster and print sample output for the same:

Cluster 0:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Locality** | **Restaurant** | **Cluster Labels** | **Rates** | **Latitude** | **Longitude** |
| 4 Bunglows | 0.0 | 0 | 22780.0 | 19.044471 | 72.910060 |
| Aarey Milk Colony | 0.0 | 0 | 8202.5 | 19.156129 | 72.870722 |
| Andheri (East) | 0.0 | 0 | 16723.5 | 19.115883 | 72.854202 |
| Andheri (West) | 0.0 | 0 | 22015.0 | 19.117249 | 72.833968 |
| Amboli | 0.0 | 0 | 18360.0 | 19.131992 | 72.849960 |

Cluster 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Locality** | **Restaurant** | **Cluster Labels** | **Rates** | **Latitude** | **Longitude** |
| Asha Nagar | 0.166667 | 1 | 16405.0 | 19.169198 | 72.944605 |
| Dahanukar Wadi | 0.200000 | 1 | 15682.5 | 19.249450 | 72.859621 |
| Devipada | 0.166667 | 1 | 15618.5 | 19.174020 | 72.869522 |

Cluster 2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Locality** | **Restaurant** | **Cluster Labels** | **Rates** | **Latitude** | **Longitude** |
| DN Nagar | 0.285714 | 2 | 23268.5 | 19.241026 | 72.844074 |
| Dahisar | 0.250000 | 2 | 12133.5 | 19.248693 | 72.864059 |

Cluster 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Locality** | **Restaurant** | **Cluster Labels** | **Rates** | **Latitude** | **Longitude** |
| Aram Nagar | 0.090909 | 3 | 24310.0 | 19.134031 | 72.813052 |
| Borivali (West) | 0.117647 | 3 | 16872.0 | 19.115287 | 72.861808 |
| Gorai 2 | 0.083333 | 3 | 13302.5 | 19.163328 | 72.841200 |

Cluster 4:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Locality** | **Restaurant** | **Cluster Labels** | **Rates** | **Latitude** | **Longitude** |
| Chikuwadi | 0.037037 | 4 | 16299.0 | 19.179251 | 72.838415 |
| Chincholi Bunder | 0.045455 | 4 | 15236.0 | 19.098705 | 72.851599 |
| Dahisar (West) | 0.055556 | 4 | 14683.5 | 19.220890 | 72.853254 |
|  |  |  |  |  |  |

**Discussion:**

By analysing full data of all clusters we find that cluster 2 is good choice for opening a new restaurant. Also localities near Dahisar are good option in terms of buying rates of house as compared to localities towards Andheri.

**Conclusion:**

Hence we can conclude that using Machine Learning approaches we can analyse large amounts of data to take some decisions or do some predictions.