Coursera Capstone Project

IBM Data Science Professional Certification

The Battle of Neighborhoods

Finding a potential venue for opening new Hotel in Mumbai, India

By : Hardik Dave

Introduction

In a metro city like Mumbai, the business tourism has been developing quite rapidly. As the Mumbai city is said to be the financial capital of India, the business tours from foreign and other cities of India happen quite often. Thus hotel business contribute well in the economy also. Hotels near Railway stations and Airport are most likely to be the first choice for business professionals. Hotels are better option for staying during business visits. There are number of hotels in different areas of Mumbai which serve as better staying option for tourism especially business tourism because of the financial activities in Mumbai. For building a new hotel and starting new business in hotel industry requires several attributes to be considered. The main attribute to be considered is the area and surroundings in which we want to open the hotel. The location and venue of the hotel to be opened is the most important feature to be considered for opening new Hotel.

Business Problem

The main motivation of doing this capstone project is to develop a plan to find best suitable venue or location for opening new hotels in Mumbai, India. As discussed in the previous section, we need to focus on the area or surroundings for different neighbourhood and suburbs in Mumbai. We need to come up with concrete information about the distribution of hotels in Mumbai's various areas. Then we will be able to analyse that which area should be considered as the potential area for opening new hotel. The main objective of this project is to suggest a potential area where a person can build and open hotel for maximum benefits.

Target Audience

The target audience for this project is property developers and hotel business professional builders who are looking to invest or venture by opening new hotel. By this project, we will be able to recommend better and potential areas for investment for such personnel. There are excessive amount of hotels in some areas, so we can smartly analyse by the data that which area is the most beneficial for opening new hotel.

Data

The data that we will need to complete this project is :

- List of suburbs in Mumbai, India
- Coordinates for each suburb (Neighborhood) in terms of Longitude and Latitude
- The data related to Hotels (Venue Data)

Data Description and Sources of Data

The data we will be using for collecting suburb information of Mumbai, India is available at this location : https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai

This data contains 42 neighborhoods or suburbs of Mumbai. This data can be extracted by using Web Scrapping method. For that we can use BeautifulSoup package of Python.

We can get coordinates in terms of Longitude and Latitude from Geocoder package of python for each suburbs of Mumbai.

Foursquare API can be used for collecting venue data for the neighborhoods in Mumbai with the help of gathered coordinates. We will extract mainly Hotel venue data for the project.

After getting this hotel venue data for each suburbs of Mumbai, we can do the visualization of this data and clustering the data to gather important insights.

Methodology

In this capstone project, the main objective is to find out potential areas for developing new hotel projects. For this we need to get the appropriate data. This project is specifically concentrated on Mumbai, India.

First, we need to gather the neighbourhoods and suburbs data of Mumbai City. We can gather this data from Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai) and use web scrapping for reading the data from web and storing in pandas dataframe. This data will give us 42 suburbs list of Mumbai.

We need to collect the coordinates of these suburbs which can be done by using Geocoder package of python. The longitude and latitude information of each suburb will be available by using this package. Foursquare API can be used to collect information of surrounding and other information by querying with coordinates. So in order to get the information about various places and surroundings of the suburbs of Mumbai, we need to have coordinates in terms of longitudes and latitudes.

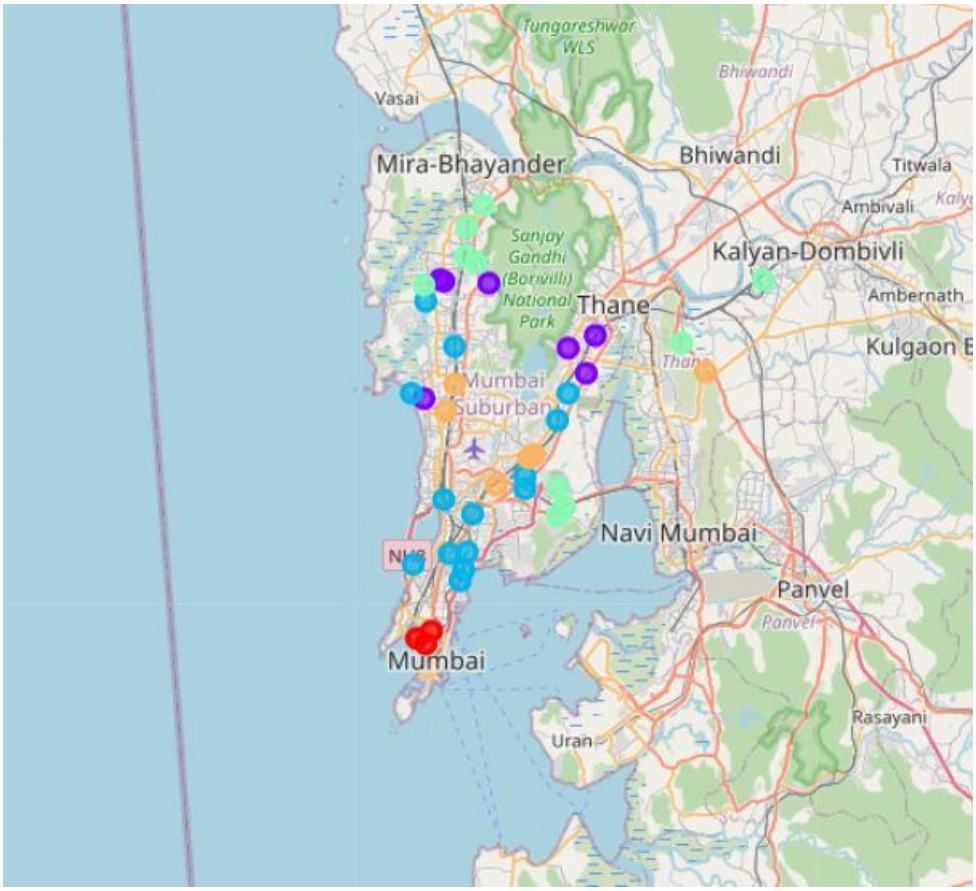
Now we have longitudes and latitudes for each suburb of Mumbai, we can pass this information to Foursquare API calls to get venue data in JSON format. For this project we have used top 100 venues within the radius of 5000 meters. From the collected data, we extract venue name, category, longitude value and latitude value. For using Foursquare API, we need to pass client secret key and ID which can be obtained by creating developer account on Foursquare website. We get 3805 values for different venues associated with different suburbs. There are 159 unique categories found for venues.

Then we extract information which contains “Hotel” as venue category. So we can have new dataframe consisting only nighborhood and hotel information.

We can now use clustering algorithm K-Means Clustering to cluster these information in 5 different clusters. After clustering we can visualize the results. Also we can view details of suburbs falling in each cluster individually to get an idea of hotel density among these suburbs.

Results

The results of clustering are shown in the below map :



This map indicates that there are different density of hotels in different areas. So that we can visualize this in a way to determine potential areas for developing new hotel project. By results of clustering and seeing the suburbs and their respective hotel venue data, we can draw an important insight as follows: Mostly the hotels are scattered in cluster 3 and 5. Because these areas have higher financial capabilities and nearer to Railway Stations and Airport. We can see that cluster 2 and 4 have almost no hotels in the neighborhood. So that we can see these areas as potential areas for opening new hotels.

The observations from different clusters are shown below:

Cluster 1

```
[ ]  mb_merged.loc[mb_merged['Cluster Labels'] == 0]
```

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
17	Kalyan	0.04	0	18.953937	72.820367
14	Grant Road	0.05	0	18.959290	72.831080
24	Matharpacady, Mumbai	0.05	0	18.950694	72.827268

Cluster 2

```
[ ]  mb_merged.loc[mb_merged['Cluster Labels'] == 1]
```

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
22	Mahavir Nagar (Kandivali)	0.01	1	19.210940	72.841370
27	Mulund	0.01	1	19.171830	72.955650
18	Kandivali	0.01	1	19.211982	72.837573
20	Kausa	0.01	1	19.127580	72.825390
34	Thakur village	0.01	1	19.210200	72.875410
4	Bhandup	0.01	1	19.145560	72.948560
33	Sonapur, Bhandup	0.01	1	19.163940	72.935440

Cluster 3

```
[ ]  mb_merged.loc[mb_merged['Cluster Labels'] == 2]
```

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
32	Sion, Mumbai	0.03	2	19.043410	72.863320
29	Pestom sagar	0.03	2	19.070640	72.902170
35	Tilak Nagar (Mumbai)	0.03	2	18.996160	72.852790
38	Vikhroli	0.02	2	19.111090	72.927810
40	Western Suburbs (Mumbai)	0.03	2	19.197010	72.827680
39	Wadala	0.03	2	19.017200	72.858170
19	Kanjurmarg	0.03	2	19.131380	72.935680
30	Seven Bungalows	0.02	2	19.131342	72.816342
41	Worli	0.03	2	19.007440	72.816880
3	Bandra	0.03	2	19.054370	72.840170
13	Goregaon	0.02	2	19.164550	72.849460
7	Chembur	0.02	2	19.062180	72.902410
16	Juhu	0.03	2	19.014920	72.845220
11	Eastern Suburbs (Mumbai)	0.03	2	19.004270	72.855792

Cluster 4

```
[ ]  mb_merged.loc[mb_merged['Cluster Labels'] == 3]
```

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
6	Charkop	0.0	3	19.208660	72.826120
9	Devipada	0.0	3	19.224690	72.866050
28	Mumbra	0.0	3	19.167632	73.021408
8	Dahisar	0.0	3	19.250030	72.859070
26	Mogra Village	0.0	3	24.375900	75.954570
25	Mira Road	0.0	3	19.265674	72.870681
23	Mankhurd	0.0	3	19.048530	72.932220
2	Baiganwadi	0.0	3	19.062940	72.926630
1	Anushakti Nagar	0.0	3	19.042830	72.927340
5	Borivali	0.0	3	19.229360	72.857510
10	Dombivli	0.0	3	19.212750	73.083240

Cluster 5

```
[ ]  mb_merged.loc[mb_merged['Cluster Labels'] == 4]
```

	Neighborhood	Hotel	Cluster Labels	Latitude	Longitude
37	Vashi	0.070000	4	19.084650	72.904810
15	Jogeshwari	0.060000	4	19.137920	72.849410
31	Shil Phata	0.074074	4	19.146580	73.040050
12	Ghatkopar	0.060000	4	19.086523	72.909008
21	Kurla	0.060000	4	19.064980	72.880690
36	Uttan	0.064516	4	26.866340	80.938840
0	Andheri	0.060000	4	19.118459	72.841763

Discussions and Possible Enhancement

From the results and observations, we can say that Cluster-2 and Cluster-4 are most suitable for development of new hotel project. The suburbs in the areas falling under Cluster-2 and Cluster-4 are can be thoughtful suggestions locations for building new hotel. In contrast, the Cluster-1 and Cluster-5 has excessive amount of hotels as compared to other areas. So we can say that in the areas falling under the Cluster-2 and Cluster-4 has great potential and provide great opportunity for building a new hotel project. In the areas where the hotel density is very high, the competition and other aspects are not good for starting new hotel there. We can take smart decision by analysing this data and come to a better location or place. Furthermore this project can be further extended to not only be used for hotel venue, we can analyse other attributed or venues also. We can also consider multiple attributes for suggesting the potential location for starting new hotel. We can take population, nearby Railway station or Airport into consideration for enhancing the decision making process of the system.

Conclusion

From this project, we can conclude that we can find potential areas or locations which can be beneficial for developing and starting new hotel business. We can use data science concepts along with the appropriate data and usage of data to make suggestions which can not be given without looking at and understanding data. We have used Wikipedia data for collecting suburbs information of Mumbai, India. This data is fed to Geocoder package for obtaining coordinates in terms of longitude and latitude. Then we have employed Foursquare API for extracting the venues information based on the longitude and latitude information of each suburb. Then the clusters are created based on the data using K-Means clustering algorithm and from the results we can suggest or recommend potential and beneficial areas for developing new hotel project.

References

1. Suburbs in Mumbai, India.
   Source : https://en.wikipedia.org/wiki/Category:Suburbs_of_Mumbai
2. Foursquare API Docs
   Source : https://developer.foursquare.com/docs/places-api/