Clustering Hotels in Mumbai, India

Niharika Kumari

30/03/2021

1. Introduction

1.1 Background

Tourism in Mumbai is an industry that attracts almost 6 million tourists per year, making it the 30th-most visited location worldwide. According to United Nations, as of 2018, Mumbai was the second most populous city in India after Delhi and the seventh most populous city in the world with a population of 19.98 million. Mumbai offers natural heritage and modern entertainment including leisure spots, beaches, cinemas, studios, holy places, amusement parks and historical monuments.

According to a survey done in 2020, it is the 60th most expensive city for expatriates globally, while it ranks 19th in Asia. Mumbai is the most expensive among all the Indian cities surveyed.

A trip to the city of Mumbai will be both, expensive and enjoyable. A good planned trip means lesser cost and more fun.

1.2 Problem

Good analysis of the city hotels, taking into consideration the cost of stay and access to venues nearby can save a traveler a lot of money. We will be trying to cluster our hotels based on their prices and venues in the radius of 1 km from the hotel. The project aims to help our fellow travelers hoping to visit Mumbai, decide on which hotel to opt for based on their budget and interest.

1.3 Interest

Anyone, planning to visit Mumbai in the near future can use our analysis to choose their place of stay.

2. Data Acquisition and cleaning.

2.1 Data Sources

We will be extracting our data using the TripAdvisor API, Geopy python package and the Foursquare API. TripAdvisor API will provide us with all the hotels in Mumbai sorted by popularity. We will have details like price per night of stay at the hotel, it's locality and its features. One crucial detail that is missing here is the latitude and longitude values of the hotels. We will extract this information using the Geopy library. To get nearby venues in the radius of 1km from the hotel, we will be using the Foursquare API.

	hotel_name	url	locality	reviews	tripadvisor_rating	checkIn	checkOut	price_per_night
0	Sofitel Mumbai BKC	http://www.tripadvisor.com/Hotel_Review-g30455...	Mumbai	6940	NaN	2021/04/10	2021/04/15	$41</td> <td>Tripadvisor</td> <td>11</td> <td>Free Wifi ,Free parking</td> </tr> <tr> <td>1</td> <td>Trident, Bandra Kurla, Mumbai</td> <td>http://www.tripadvisor.com/Hotel_Revie g30455...</td> <td>Mumbai</td> <td>4730</td> <td>NaN</td> <td>2021/04/10</td> <td>2021/04/15</td> <td>$69

Through this method we have data of 210 hotels. While exploring and extracting nearby venues for each hotel, a hard limit of 100 venues per hotel was given. Ultimately, we have 10882 venues for different hotels.

2.2 Data Cleaning

Data downloaded or scraped from multiple sources were combined into one table. We could obtain accurate latitude and longitude values of only 10% of the total number of hotels extracted from TripAdvisor, i.e, 2229. We will be dropping all the hotels for which we do not have the latitude and longitude co-ordinates.

Plotting this data on map highlighted outliers present in our dataset. Not quite surprisingly, due to same name hotels present in different parts of the world, some hotel co-ordinates extracted fell not in Mumbai, not even in India, but in different countries like Malaysia, Mexico, Africa and London. After dropping these values from the dataset we are left with 200 hotels and 10663 venues around these.

2.3 Feature Selection

A lot of information obtained from TripAdvisor is not very useful to us, for eg, the url, tripadvisor rating, checkin, checkout date that we added as our filter. We will be dropping these columns from our dataset.

Another column containing features of each hotel is explored, for 200 hotels it had only 19 unique set of features which are as follows –

['Free parking ,Pool ,Visit hotel website ',
'Free Wifi ,Free parking ,Visit hotel website ',
'Free Wifi ,Free parking ', 'Free parking ,Pool ',
'Free Wifi ,Restaurant ', 'Free Wifi ,Room service ',
'Free Wifi ,Beach ', 'Free parking ,Restaurant ',
'Free Wifi ,Pool ', 'Free Wifi ', 'Restaurant ,Room service ',
'Room service ', 'Free parking ,Room service ', 'Free Internet ',
'Pool ,Restaurant ', 'Free parking ', 'Restaurant ,Bar/Lounge ',
'Bar/Lounge ', 'Restaurant ']

We observed that most of them are similar in nature and do not add much value to our dataset, thus we drop these from our dataset.

2.4 Feature Engineering

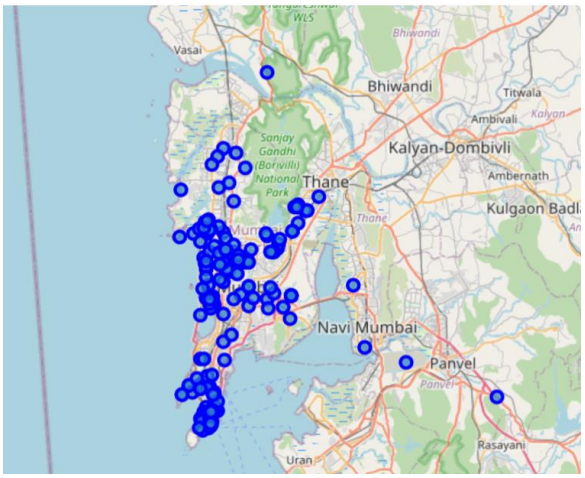
As mentioned in the introduction section, we plan on clustering our hotels into different groups. For clustering algorithm, we need categorical data and not continuous. To achieve this, we perform an operation called one-hot encoding. A one-hot is a group of bits among which the legal combinations of values are only those with a single high bit and all the others low. A similar implementation in which all bits are '1' except one '0' is sometimes called one-cold.

We perform this operation on nearby venues as well as prices of each hotel. Note that the names of the venues will not add a lot of value, instead we go for category of the venue, for eg. Bar, Chinese Restaurant, Spa, Market etc. instead of their board names which can be anything. We have 71 unique prices and 244 unique venue categories.

Our one-hot encoded data is ready with 315 features for 198 hotels.

2.5 Data Visualization

We plot our geographical dataset using the latitude and longitude values on a map of Mumbai using the python Folium library. We can see how our dataset is spread over all of the city.



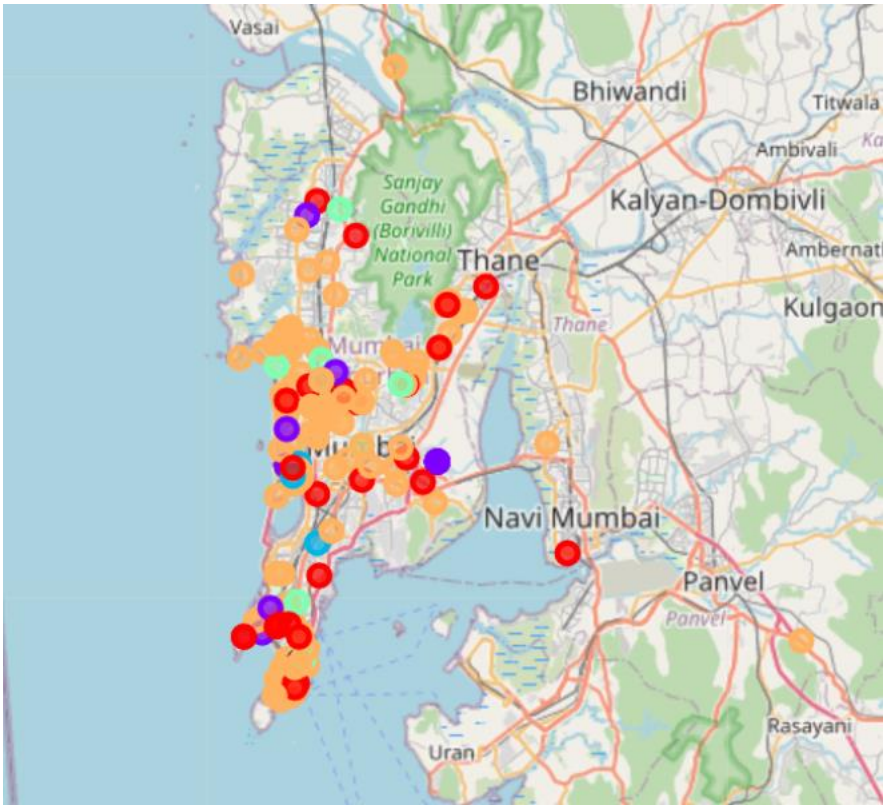
3. Modeling

There are mainly two kinds of modeling in machine learning – Supervised and Unsupervised. Our data and problem fall in the latter, since we don't have a label attached to our features. We just pass these values and ask the model to group them into clusters such that hotels with similar features fall into same category.

There are many models available for clustering, we will be using K-means algorithm. It clusters the dataset by calculating the centroid of groups of points and placing each point in the cluster whose centroid is nearest to it. By default it uses Euclidian distance to calculate the distance between two points.

We will be clustering our dataset into 5 different clusters.

Let us visualize how our points are categorized:



4. Results and Discussion

The city of Mumbai provides us with a lot of options to choose our place of stay with a wide range of prices. It also offers us good quality and a variety of cuisines starting with Indian flavors along with different Asian and fast-food options.

We can see that clusters 3 and 4 are completed categorized on the basis of price per night and both of them offer a good number of Restaurants and cafes and other places nearby, saving the cost of travel.

The 2nd cluster are all mid-range priced hotels and best choice for someone fond of sports. The first cluster is also ranging from low to mid-range hotels and perfect for a foodie. The last cluster more or less comprises of high price hotels and (owing to my background knowledge) are mostly 5 Star hotels.

The best thing that I observe here is that all the clusters are spread throughout the city. So, say we want to stay close to a particular venue, say for eg. Nariman Point or Juhu Beech, we can still choose a hotel from low priced clusters.

5. Conclusion

The purpose of our project was to help our fellow travelers choose their place of stay in the city of Mumbai, India by analyzing cost of booking and availability of venues nearby. By extracting data from different sources and keeping two main features- nearby venues and prices of stay, we clustered our dataset for 198 hotels into 5 clusters using K means algorithm of SK learn library of Python.

Analyzing these clusters gives us a good idea and helps us decide what place to choose to make the most of our trip. Based on our budget we can choose between high priced or low-priced hotels. Based on our interest in sports or bars or cafes or shopping, we can select our place of stay.

6. Future Work

Further work can be done on the project by adding features in the dataset like distance of the hotels from major tourist sites and then prepare a cluster as to which place would be ideal keeping in my mind not only price but accessibility to Airports or famous tourist destinations.