Clustering Hotels in the City Of Mumbai, India

NIHARIKA KUMARI

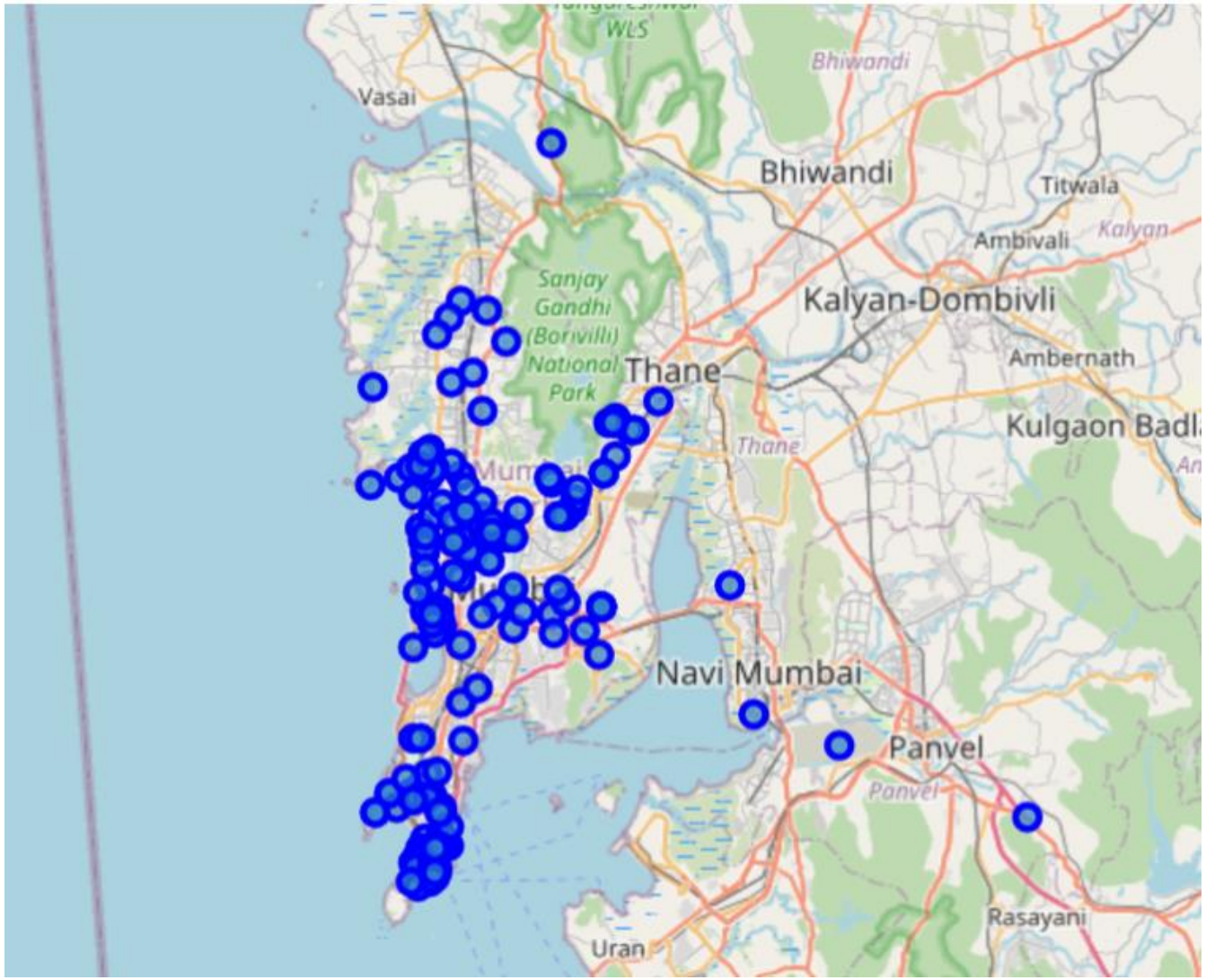
To choose a good place of stay in a new city is the best way to plan our budget and make the most of our trip!

- Tourism in Mumbai is an industry that attracts almost 6 million tourists per year.
- A trip to the city of Mumbai will be both, expensive and enjoyable. A good planned trip means lesser cost and more fun.
- 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.
- Anyone, planning to visit Mumbai in near future can use our analysis to choose their place of stay.

Data acquisition and cleaning

- Extract Hotels data, containing hotel names, price per night of stay, rating and hotel features from TripAdvisor API.
- Extract latitude and longitude values of each hotel using the Geopy library.
- Extract nearby venues within a radius of 1 km from the hotel using Foursquare API.
- Drop all hotels for which we could not extract latitude and longitude values.
- Remove outliers from the data by filtering our dataset with the latitude and longitude of Maharashtra.
- Drop features that do not add value to our dataset, like the url, TripAdvisor rating, checkin, checkout date, no. od deals, and hotel features which are mostly similar in kind for all hotels.
- Cleaned data contains 200 hotels and 8 features.

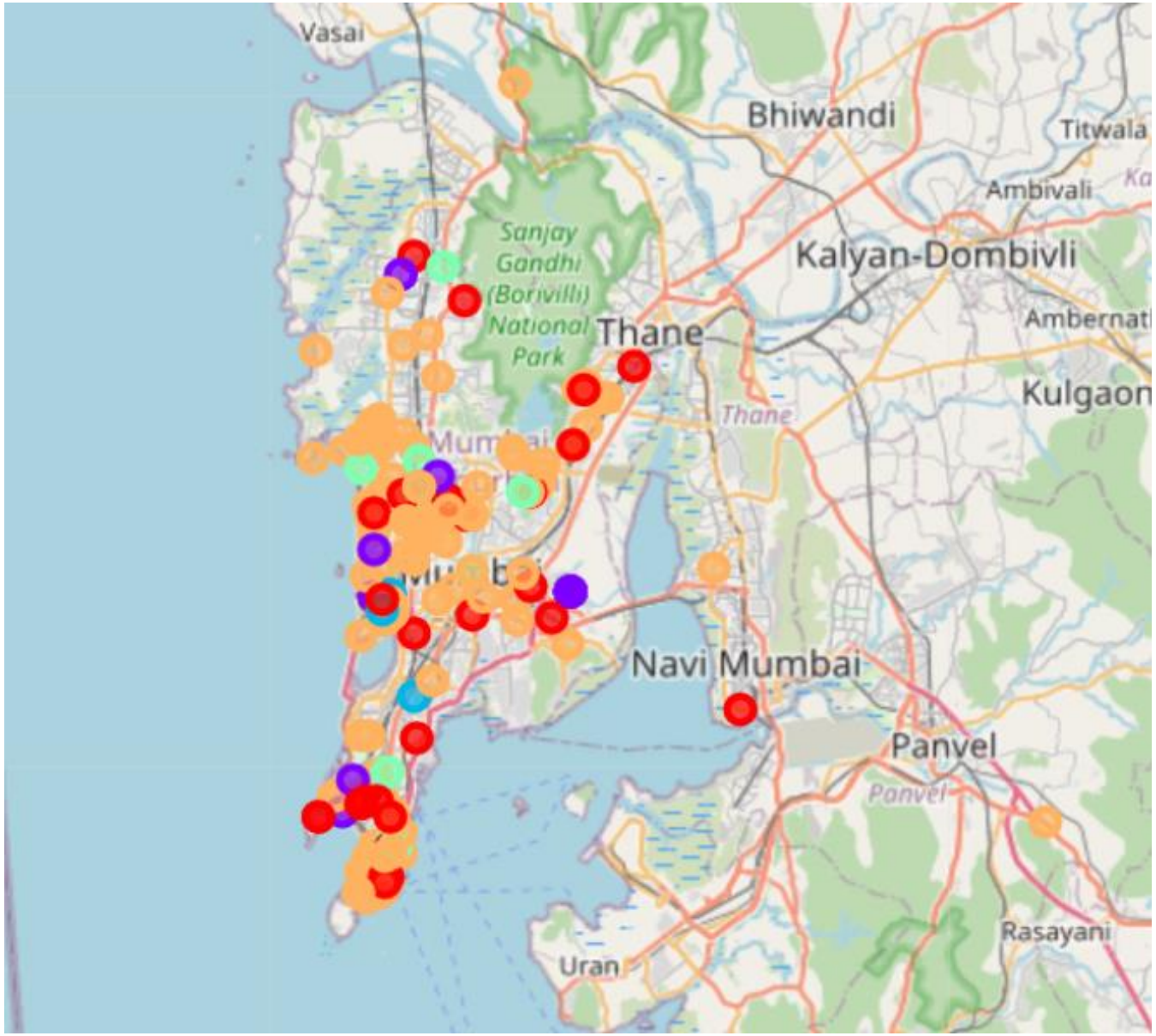
Input data visualization on Map



Modeling

Our Problem is that of unsupervised learning.

We will be using K-means to cluster our dataset into 5 groups. This is how the clusters look on the map.



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.

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.

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.