IBM Applied Data Science Final capstone report

Coursera Capstone Project

IBM Applied Data Science

Optimum location to set up a new club in Mumbai

- Arindam A Baruah

Introduction

As urbanisation and modernisation amongst the masses has taken place, more and more people are seen to love the night life with drinks and dance bars more than ever. Most cities in India have started having enjoyable late nights in clubs or sports bars to enjoy their weekends.

In this report, we will be taking the case study for the city of Mumbai which has seen a great transformation to the night life culture. The reasons on why Mumbai was chosen for our case study is as follows:

Pros

- Mumbai is called the " Indian city that never sleeps". This means the crowd at markets, bars, clubs are omnipresent meaning more business hours and profit.
- Mumbai is the financial capital of India and hence, makes a good choice to cater to the niche market.
- Mumbai is the residence to all the top Bollywood and sports stars of the country. These stars often do visit clubs and it makes the club extremely likeable to the general public.

However, like most things in life, Mumbai also has some drawbacks which are described below.

Cons

- There is a huge economic diversity in the city. While the city is the host to the richest people in the country, it also houses the world's largest slum.
- Land acquisition is an expensive affair in Mumbai.
- Few areas of the city are highly radical in terms of religion.

Hence, judging by the above pros and cons, it is important to leverage some data science and machine learning which can help people choose which areas to target when one plans to invest in a nightclub.

Business problem

In order to open a night club, there is significant investment with respect to land acquisition, construction and obtaining a liquor license. Keeping all the high investments in mind, we need to be able to make a sound decision to turn our investment into a smart investment. We shall use data science and unsupervised machine learning in particular to point out the areas where it'll be profitable to set up a potential night club.

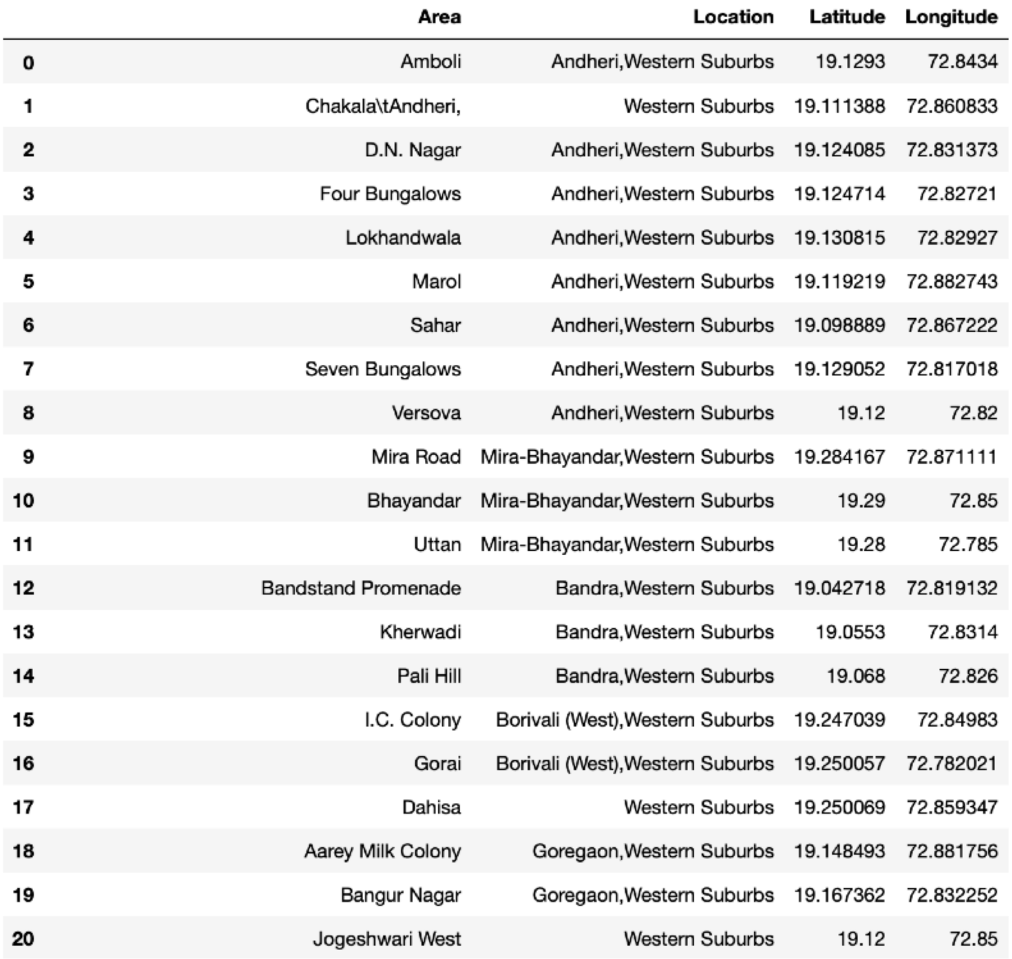
Target audience

The findings of this report will be of particular interest to the property developers and our potential investors who would be interested to have some stake as ownership of the club. As Mumbai is becoming more and more cosmopolitan, it is certain that there will be many potential businessmen who would like to invest in such a property. Considering the fact that nightclubs and bars have a fairly high median profit, such a property would definitely be deemed as a sound investment if the right choices in terms of location of the nightclub are made.

Data

- For the purpose of our study, we require a detailed list of all the areas ,locations along with their latitudes and longitudes of Mumbai city. We require a web scraping tool such as BeautifulSoup which can help us to get the relevant table containing the tables of areas and locations.

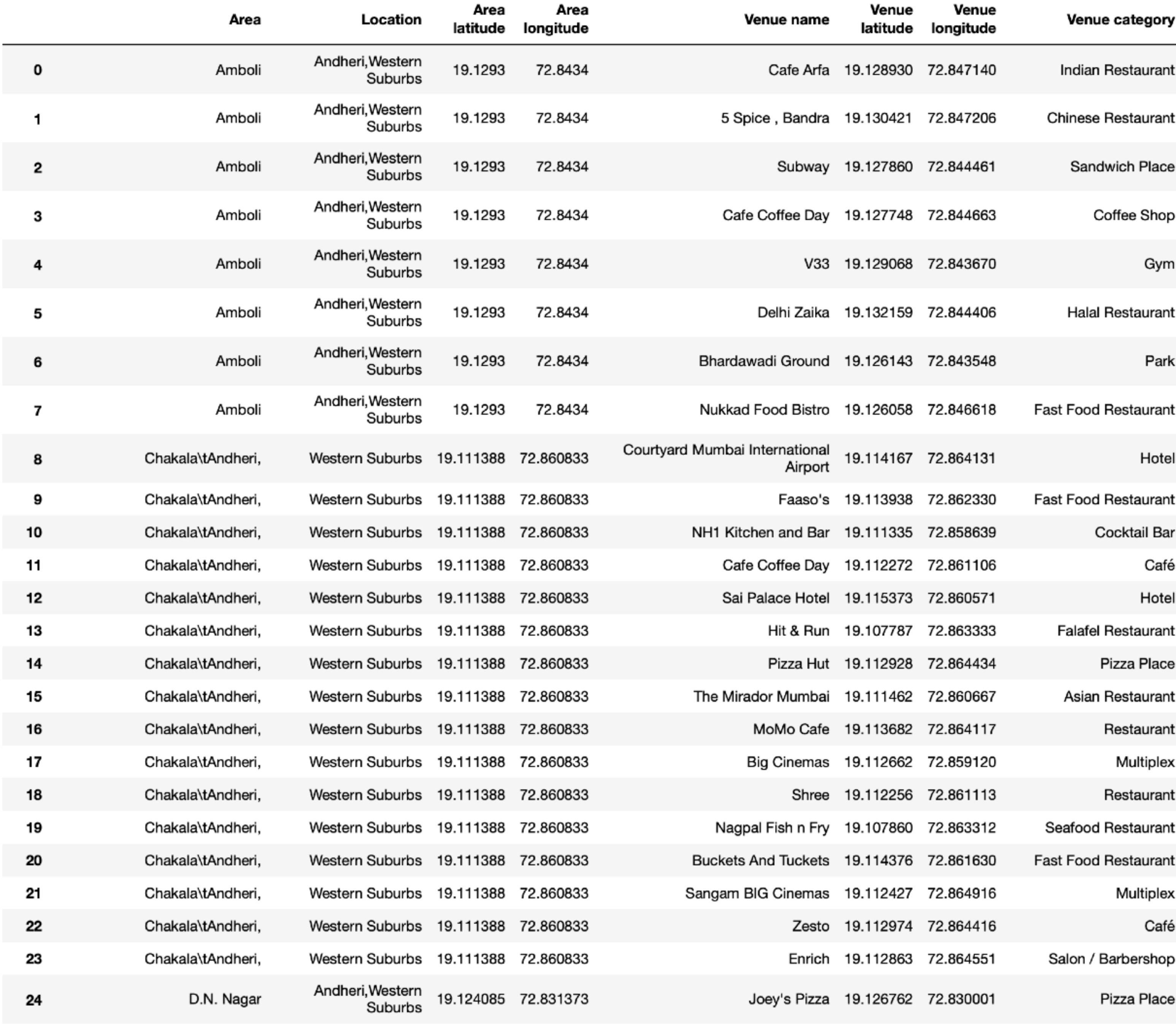
- For the purpose of finding the latitudes and longitudes, we can use the geolocator API. However, this is a tedious process. Thankfully, in our case, web scrapped data already contained the latitudes and longitudes. Hence, they could be scrapped using the BeautifulSoup tool. A screenshot of the web scrapped data when put into a data frame is show below.



	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.1293	72.8434
1	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.82721
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.82927
5	Marol	Andheri,Western Suburbs	19.119219	72.882743
6	Sahar	Andheri,Western Suburbs	19.098889	72.867222
7	Seven Bungalows	Andheri,Western Suburbs	19.129052	72.817018
8	Versova	Andheri,Western Suburbs	19.12	72.82
9	Mira Road	Mira-Bhayandar,Western Suburbs	19.284167	72.871111
10	Bhayandar	Mira-Bhayandar,Western Suburbs	19.29	72.85
11	Uttan	Mira-Bhayandar,Western Suburbs	19.28	72.785
12	Bandstand Promenade	Bandra,Western Suburbs	19.042718	72.819132
13	Kherwadi	Bandra,Western Suburbs	19.0553	72.8314
14	Pali Hill	Bandra,Western Suburbs	19.068	72.826
15	I.C. Colony	Borivali (West),Western Suburbs	19.247039	72.84983
16	Gorai	Borivali (West),Western Suburbs	19.250057	72.782021
17	Dahisa	Western Suburbs	19.250069	72.859347
18	Aarey Milk Colony	Goregaon,Western Suburbs	19.148493	72.881756
19	Bangur Nagar	Goregaon,Western Suburbs	19.167362	72.832252
20	Jogeshwari West	Western Suburbs	19.12	72.85

- For the purpose of web scrapping, we utilised the data from www.wikipedia.com .

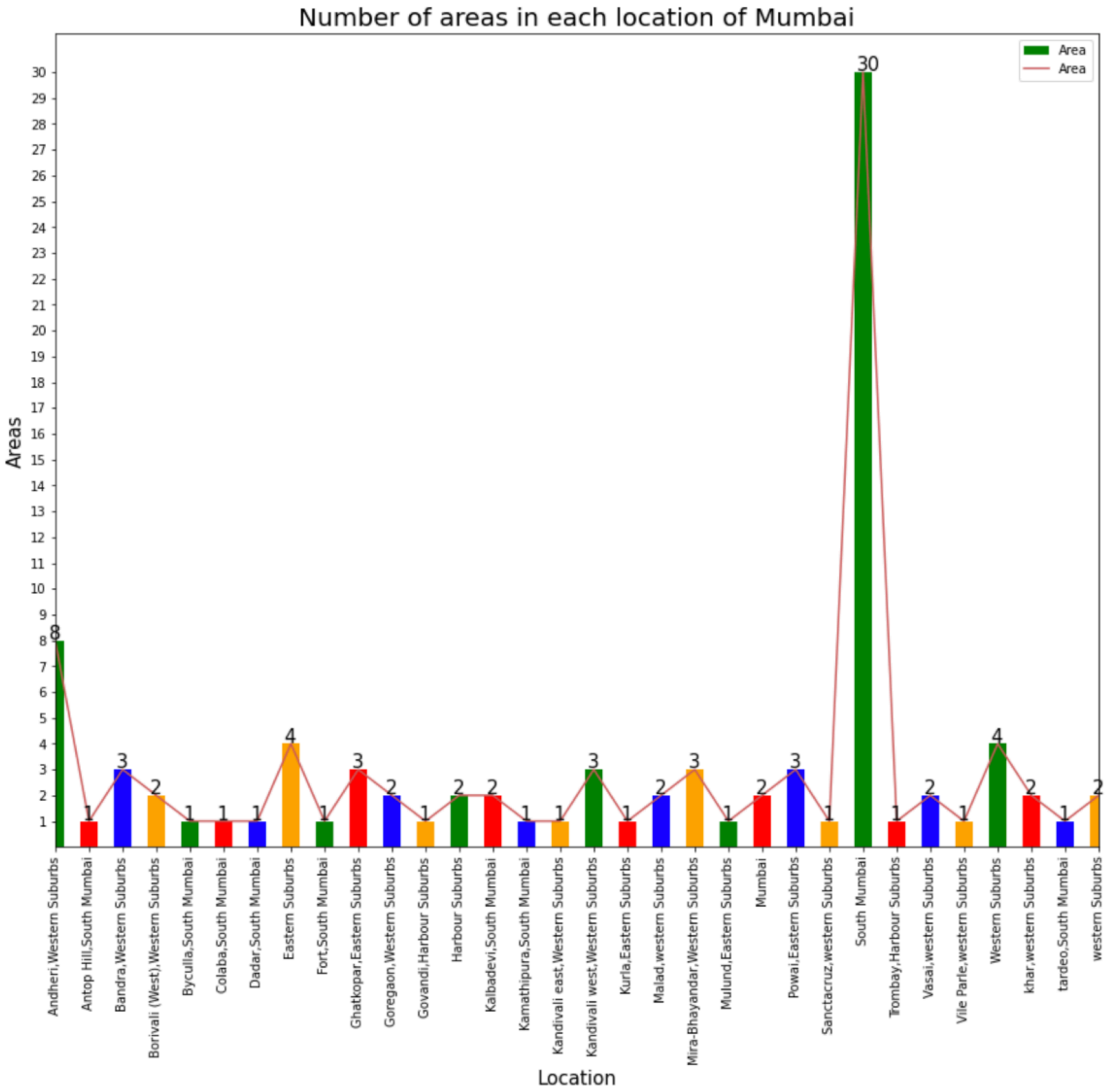
- We use the Foursquare API (www.foursquare.com) to get the various search results using the previously obtained list of latitudes and longitudes of various areas of Mumbai. These search results are stored in pandas data frame for better readability. A screenshot of the data obtained from Foursquare when restructured into a data frame looks as follows.



	Area	Location	Area latitude	Area longitude	Venue name	Venue latitude	Venue longitude	Venue category
0	Amboli	Andheri,Western Suburbs	19.1293	72.8434	Cafe Arfa	19.128930	72.847140	Indian Restaurant
1	Amboli	Andheri,Western Suburbs	19.1293	72.8434	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
2	Amboli	Andheri,Western Suburbs	19.1293	72.8434	Subway	19.127860	72.844461	Sandwich Place
3	Amboli	Andheri,Western Suburbs	19.1293	72.8434	Cafe Coffee Day	19.127748	72.844663	Coffee Shop
4	Amboli	Andheri,Western Suburbs	19.1293	72.8434	V33	19.129068	72.843670	Gym
5	Amboli	Andheri,Western Suburbs	19.1293	72.8434	Delhi Zaika	19.132159	72.844406	Halal Restaurant
6	Amboli	Andheri,Western Suburbs	19.1293	72.8434	Bhardawadi Ground	19.126143	72.843548	Park
7	Amboli	Andheri,Western Suburbs	19.1293	72.8434	Nukkad Food Bistro	19.126058	72.846618	Fast Food Restaurant
8	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Courtyard Mumbai International Airport	19.114167	72.864131	Hotel
9	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Faaso's	19.113938	72.862330	Fast Food Restaurant
10	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	NH1 Kitchen and Bar	19.111335	72.858639	Cocktail Bar
11	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Cafe Coffee Day	19.112272	72.861106	Café
12	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Sai Palace Hotel	19.115373	72.860571	Hotel
13	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Hit & Run	19.107787	72.863333	Falafel Restaurant
14	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Pizza Hut	19.112928	72.864434	Pizza Place
15	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	The Mirador Mumbai	19.111462	72.860667	Asian Restaurant
16	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	MoMo Cafe	19.113682	72.864117	Restaurant
17	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Big Cinemas	19.112662	72.859120	Multiplex
18	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Shree	19.112256	72.861113	Restaurant
19	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Nagpal Fish n Fry	19.107860	72.863312	Seafood Restaurant
20	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Buckets And Tuckets	19.114376	72.861630	Fast Food Restaurant
21	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Sangam BIG Cinemas	19.112427	72.864916	Multiplex
22	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Zesto	19.112974	72.864416	Café
23	Chakala\tAndheri,	Western Suburbs	19.111388	72.860833	Enrich	19.112863	72.864551	Salon / Barbershop
24	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373	Joey's Pizza	19.126762	72.830001	Pizza Place

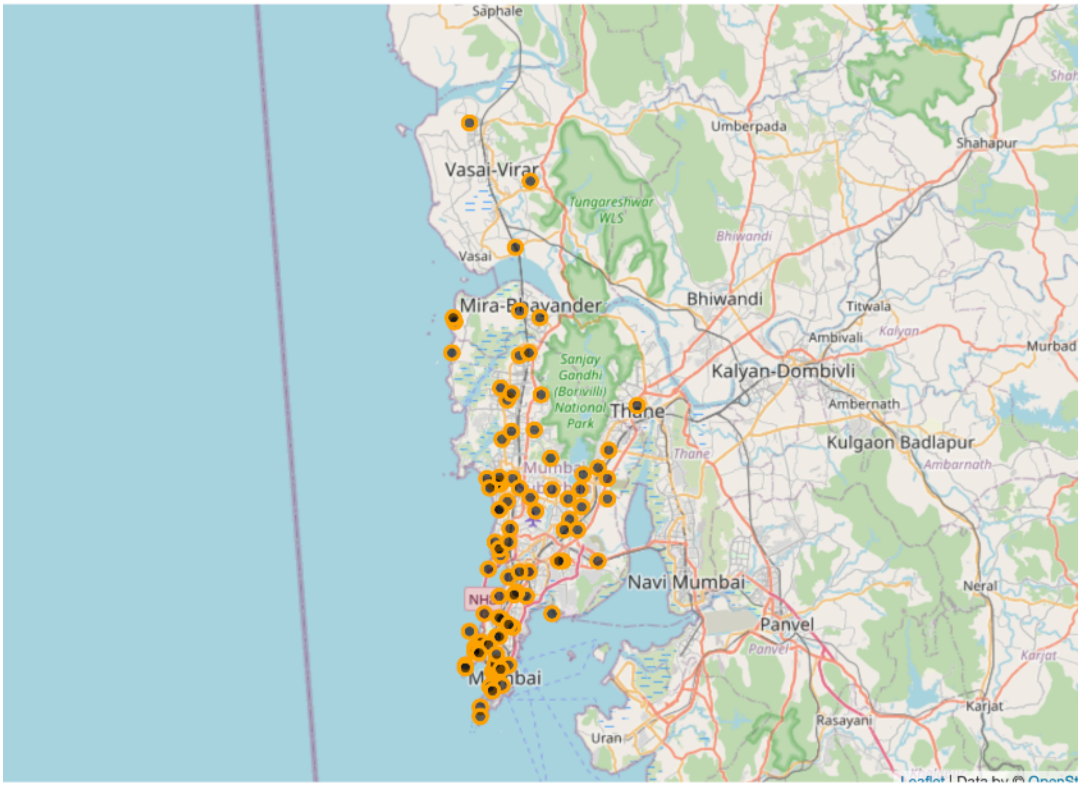
Methodology

- Initially, through the Beautiful Soup web scrapping tool, we extract the list of various areas and locations. One particular location contains multiple areas. Hence, after creation of the data frame, we make a bar plot using the matplotlib.pyplot library and analyse the number of areas present in each location.

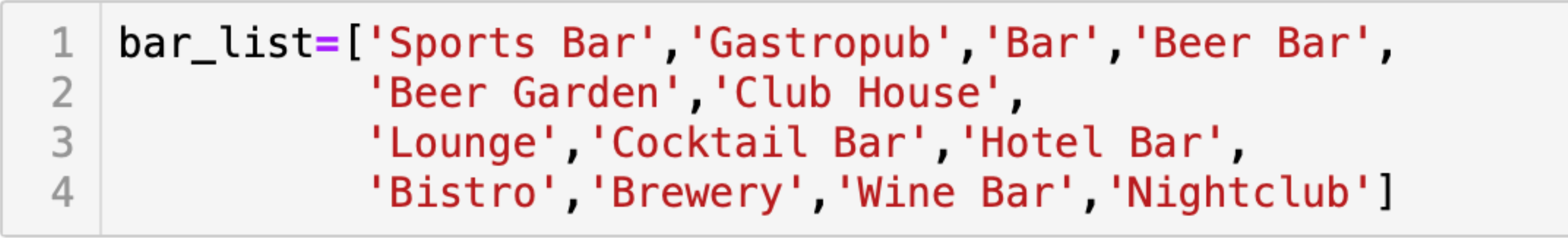


- Once we analyse the number of areas, we can get a general idea of which are the popular locations which maybe of interest to us. Once the name of the areas are obtained, we can use wikipedia or a geolocator to obtain the various coordinates of the areas.

- As the locations are obtained, we plot all these on a folium map of Mumbai to get an understanding of the regions we are dealing with. We apply suitable popup markers to pin the locations on the map for easier readability. The obtained Folium map of the city is shown below.

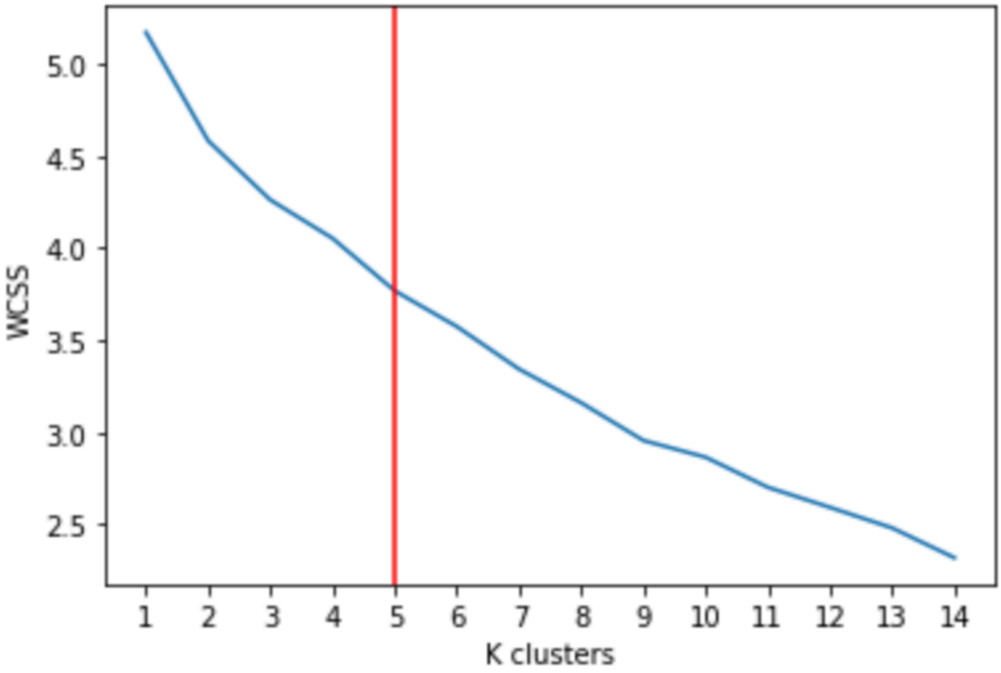


- After going through the above data, it is quite evident that the area density is quite high in the South Bombay region. Part of the reason could be that this region houses the wealthy people of the city. Most of the commercial activities are situated here. Hence, the wealthy population here is high as well. Regions with high population density are often divided into several areas to maintain proper law and order. Hence, South Bombay has higher area divisions than most other locations of Mumbai.
- Once we have a general idea of the locations of Mumbai, we use the Foursquare API to retrieve data of Mumbai. For the purpose of study we use a radius of 500 meters for each location of Mumbai.
- Under the search results, we get data from various categories such as restaurants, bars, shopping malls, grocery shops, etc.
- However, we are mainly catering to the needs such as food, nightclubs and liquor. Hence, we shall filter out the results such that we can get the idea of where the various venues are concentrated. For segregation, we filtered out all results which had their venue categories as follows.



```python
bar_list=['Sports Bar','Gastropub','Bar','Beer Bar',
          'Beer Garden','Club House',
          'Lounge','Cocktail Bar','Hotel Bar',
          'Bistro','Brewery','Wine Bar','Nightclub']
```

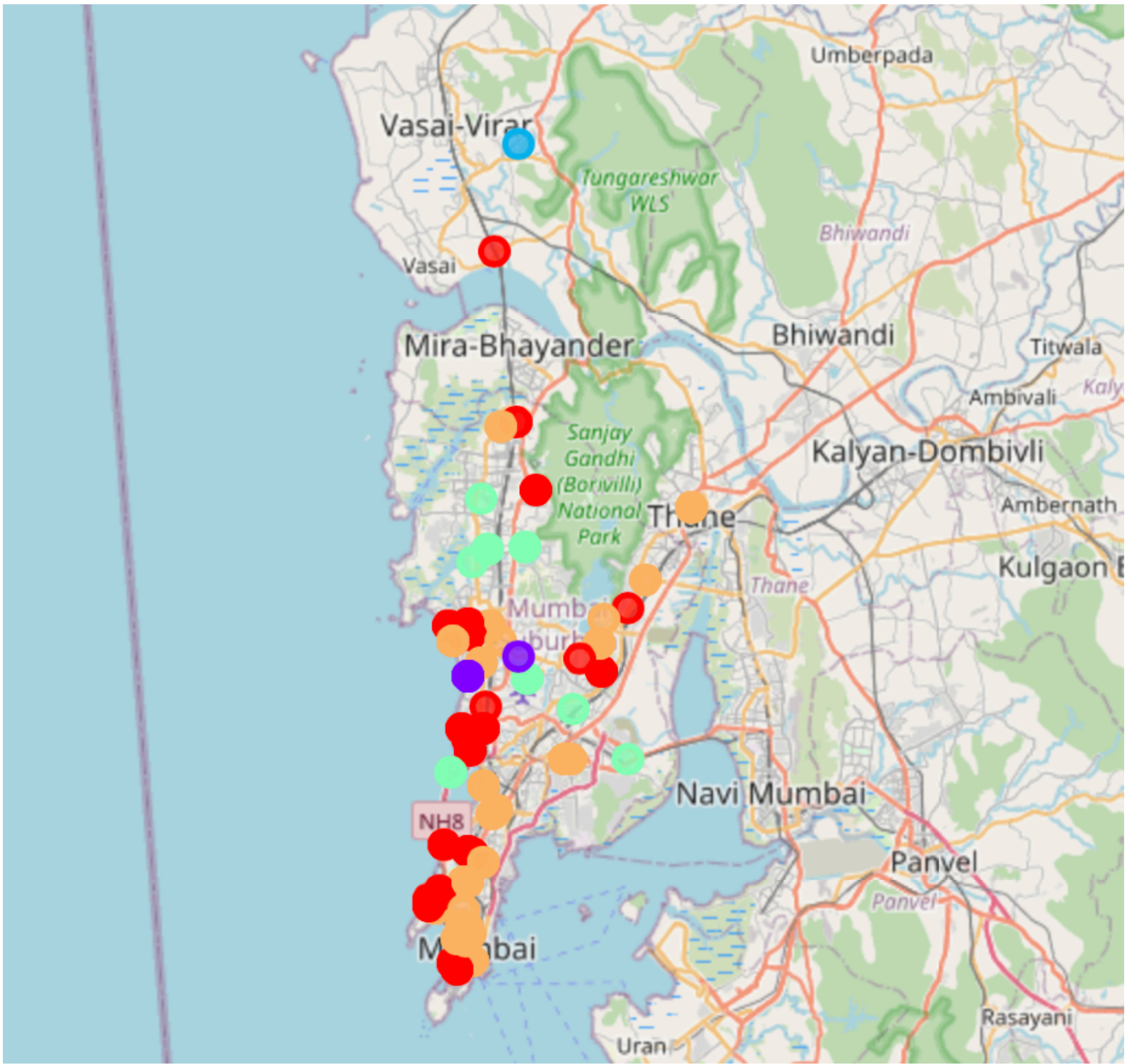
- Once the various places are filtered out from the foursquare search data, we find the location of the areas these places are based in for the purpose of understanding where these places are based in on the map.
- We now use KMeans clustering technique to cluster these various locations.
- For the purpose of using KMeans, we need to decide on the number of clusters that will give us a good result.
- This can be determined using the Within Clusters Sum of Squares parameter (WCSS). This is also called the elbow technique. The elbow graph is shown below.



- From the above elbow figure, it can't be completely determined as to what should be the optimum clusters. At k=5, it gives decent results in our analysis. Hence, we choose k=5 for further study.

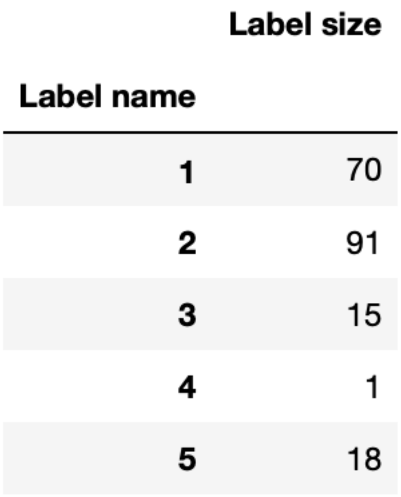
Results

- As previously discussed, we divided our data into 5 clusters. The clusters are colour marked in a Folium map to understand where these venues are located.
- Through the cluster division, we can get an idea of the regions which are highly populated with the venue categories of interest to us.
- As was expected, most of these regions are placed in the wealthy part of the city which is South Bombay.
- Below is the image of the 5 clusters on the map of Mumbai.
    1. Cluster 1: This a highly populated zone of venues in and around the South Bombay and Powai area. These are the wealthy zones.
    2. Cluster 2: This cluster is the most highly populated. It contains various venues from South Bombay, Andheri, Bandra and Khar.
    3. Cluster 3: This cluster is much less populated than other 2 clusters. This contains the Juhu and Ville Parle region which has mostly beaches and some industries.
    4. Cluster 4: This cluster contains only one location which is Nalasopara. This region is in Vasai Vihar which is quite away from the main economic zones of the city.
    5. Cluster 5: This is another relatively less populated cluster containing areas from Andheri,Malad and Bandra



Discussion

- From the observations noted above, it is clear that we have particular regions where the number of venues of interest to us are far higher. This clearly indicates that not all regions are profitable to open a nightclub. As it can be seen in the above map, the South Bombay regions have far higher number of nightclubs, bars, bistros, hotel bars and gastropubs.
- This clearly indicates that the South Bombay region is quite popular for it's parties and clubs. This was already established initially owing to the high wealth of the region. The above clusters clearly indicate our hypothesis as true.
- Both cluster 1 and cluster 2 have very high number of venues in them. This means there will be significant amount of competition if we were to choose this location. The land acquisition costs in these two clusters are higher and the investment on the property needs to be quite significant since these regions are host to the big nightclubs with very strong investors.
- cluster 3 and cluster 5 have much lower number of venues. Moreover, these are regions are not too far away from the very expensive South Bombay region. Hence, the crows will consider these regions as an option for a fun weekend if they want to try out new places which are relatively cheaper than the ones in cluster 1 and 2.
- cluster 4 is quite far away from the main attractive locations of Mumbai. This cluster primarily consists of areas from Vasai like Nalasopara. A little bit of background behind this region indicates that this area is highly religious with lots of Buddhist and Hindu temples spread around.
- Below is an image that shows the number of areas in each cluster.



	Label size
Label name	
1	70
2	91
3	15
4	1
5	18

Scope of further research

- In our study, we have primarily taken distance as our study. In order to further refine our study, we could also take into account the ratings of each of the venues.
- Once the ratings are obtained from Foursquare API, we could choose all the venues having a 4+ rating and map them.
- The region with higher ratings can be studied and clustered.
- We could also choose traffic as a parameter to understand the number of people visiting these places on an average basis. Regions with high average attendance will be of interest to us

Conclusion

- From the preliminary study, it can be concluded that if the initial investments are quite high and the stakeholders are open to tough competition, clusters 1 and 2 which are mainly the rich locations of Mumbai can be used since these regions generally have a very good nightlife and party culture.
- If investments are limited and stakeholders do not want initial competition, clusters 3 and 5 are to be chosen for the property development. These areas are not too far away from the major party locations of the city and also have the advantage of low competition. Moreover, these regions are close the airport and various multinational company offices. Most of these companies have regular corporate parties and our nightclubs will be in a good position to cater to these needs.
- cluster 4 which is situated in a far away location of Vasai should be avoided at all costs. This region is away from the main wealthy zones of the city and hence, wouldn’t be into the party or the nightlife culture. Moreover, the highly religious nature of the area would not be supportive of the urban culture.