Neighbourhood clustering and recommendation system: Mumbai

A. Introduction

A.1. Description & Discussion of the Background

Mumbai is the capital city of the state of Maharashtra in India[2]. Mumbai is also financial capital of India, and it's the most populous city in India. As the 4th most populous city in the world and one of the populous urban regions in the world, Mumbai has a metro population of about 20.41 million in 2020. The most recent census was conducted in India during 2011, which put Mumbai's Urban Agglomeration at 20,748,395, while the city itself was recorded at 12,478,447. The next national census is scheduled for 2021[1].

Population Size and City Size

Mumbai's urban population is estimated to be over 22 million, and the densely populated city is the largest in India in terms of population, trade activity and business. The metropolitan area has experienced an explosion in growth over the past 20 years, a common occurrence with metropolitan areas in India. The rapid population growth is attributed to migration from other regions in the country, with migrants seeking business and employment opportunities.

The population of Mumbai has more than doubled since 1991, when the census showed that there were 9.9 million people living in the area. The rapid population growth is attributed to migration from other regions in the country, with migrants seeking business and employment opportunities. On an average 25000 people come to Mumbai daily for work[2]. The migrating population is often oblivious of the rent and facilities across neighbourhoods in Mumbai. We can use data to build a system to:

1. Analyse property prices and show the property prices in forms of a heat map to understand the distribution of property prices across the city
2. Cluster similar Neighbourhoods and analyse the types of Neighbourhoods in the city

Build a Recommender system to recommend Neighbourhoods to a new-comer in the city based on his/her requirements such as the type of residence, budget, and lifestyle preference

A.2. Data Description

To build the system I have used the below data:

- Property rates are sourced from 99acres.com. 99Acres is an Indian real estate database website founded in 2005.
- *Nominatim API* is used to get coordinates of neighbourhoods under analysis.
- *FourSquare API* is used to get the details and types of venues in the vicinity of a neighbourhood.

B. Methodology

1. Files are hosted on Github at following repository:

https://github.com/girishtere/Applied-Data-Science-Capstone

We started with scrapping the property price data available on 99acres.com. The data was wrangled to get it set up in a desired format and get the *Locality names, the average property price per sq.ft., and the rent for 1 Room, 2 Rooms and 3 Rooms* in the locality respectively.

Locality Name	buy_rate_avg	Rent_1B_avg	Rent_2B_avg	Rent_3B_avg
4 Bunglows	19826.5	27625.0	45900.0	56767.5
Aarey Milk Colony	7543.5	18417.5	22277.0	26614.5
Airoli	10561.5	14259.0	22567.5	32742.0
Ambedkar Nagar	16235.0	25239.5	38675.0	56104.0
Andheri (East)	16809.0	25478.5	39634.5	54910.0
Andheri (West)	21058.5	25712.5	41536.0	66139.0
Asha Nagar	16171.5	18025.5	24593.5	32961.5
Azad Nagar	17871.5	24420.0	40350.0	58378.0
Bandra (West)	37400.0	49608.0	79957.5	151512.5
Belapur	9860.0	11465.0	24952.0	43741.0

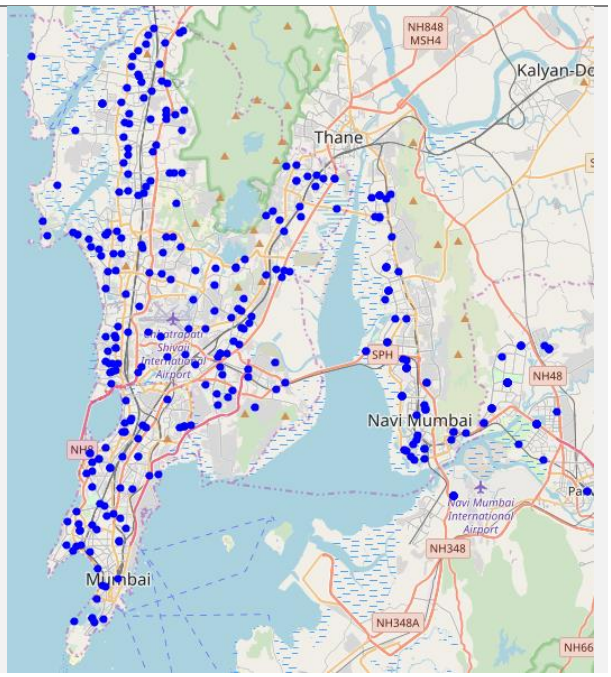
Source: https://www.99acres.com/property-rates-and-price-trends-in-mumbai

2. The geographical coordinates were fetched using **Nominatim open street API**. Coordinates for all the localities could not be fetched. So I plotted the fetched co-ordinates using **Folium** library and found that without the missing co-ordinates, there is still a good distribution of the localities could be gathered.

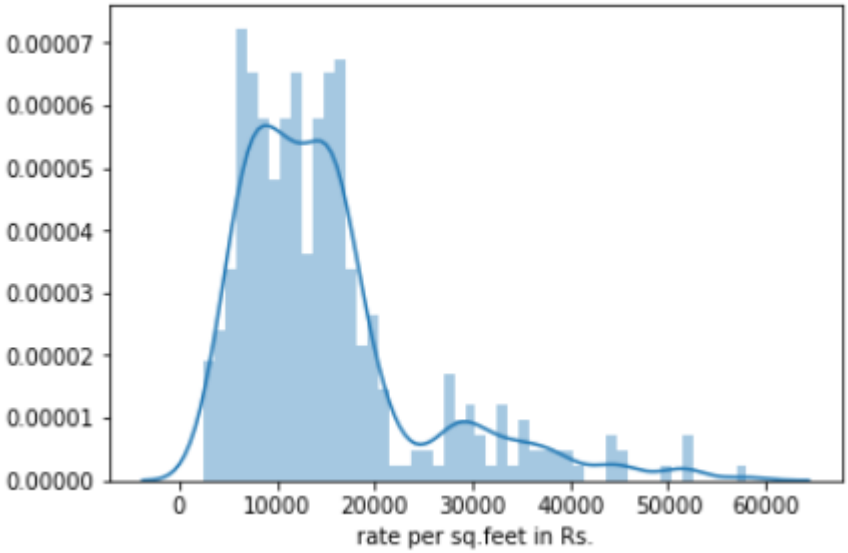
LATITUDES AND LONGITUDES

	Locality Name	Latitude	Longitude
1	Aarey Milk Colony	19.156129	72.870722
2	Abhyudaya Nagar	18.990477	72.844057
6	Airoli	19.158515	72.999402
8	Alika Nagar	19.198397	72.874267
10	Ambedkar Nagar	19.070822	72.828865
16	Amboli	19.132010	72.849864
17	Amrut Nagar	19.100845	72.911820
18	Anand Nagar	18.966523	72.811888
21	Andheri (East)	19.115883	72.854202
22	Andheri (West)	19.117249	72.833968

DISTRIBUTION OF SAMPLE DATA



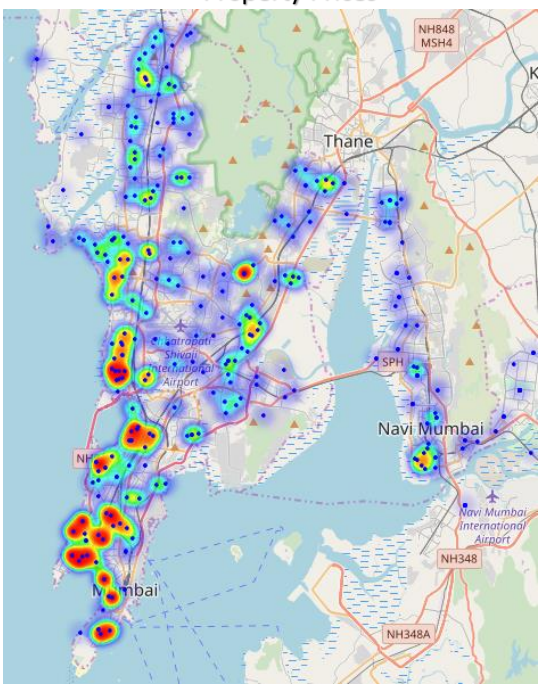
3. Plotting the property prices per square feet on a distribution curve, we can see that there is a positive skew in the data. Which indicates some areas of the cities are insanely overpriced as compared to the median price.



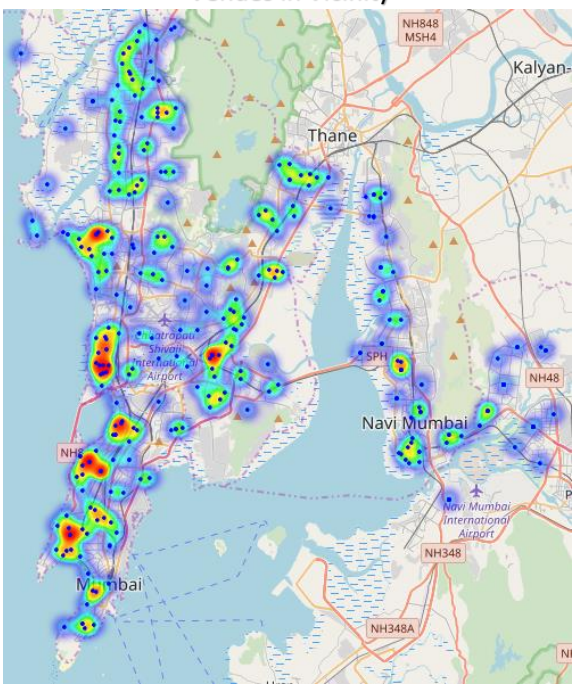
4. ***A heatmap*** plotted on the map based on prices shows that the extreme high property prices are concentrated on certain neighbourhoods. To find the reason of such skewed price distribution, we can cluster the locations according to the available venues in a *3 km radius* of the neighbourhood. Using **FourSquare API**, we can receive the venues and the types of venues nearby a neighbourhood. Comparing the heatmaps of prices and heatmap of number of nearby venues, we find a striking similarity. *Thus, the*

number of venues and their types can be good parameters to cluster the neighbourhoods and study the cluster types.

Property Prices



Venues in Vicinity



5. The total types of venues received from the Four-Square API was 236. These types are often overlapping and similar in terms of their ability to contribute to clustering of Neighbourhoods. It made more sense to converge the types of venues by grouping them.
The following 14 venue types were identified to map each of the 236 venue types.

1	regular_restaurants	8	sports_fitness
2	nature_view	9	café_fastfoods
3	tourist_interest	10	cuisine_restaurants
4	Shopping	11	arts_culture_recreation
5	transport_vicinity	12	bars_nightlife
6	business_hub	13	kids_family_residential
7	stores_daily_conveniences	14	education_colleges

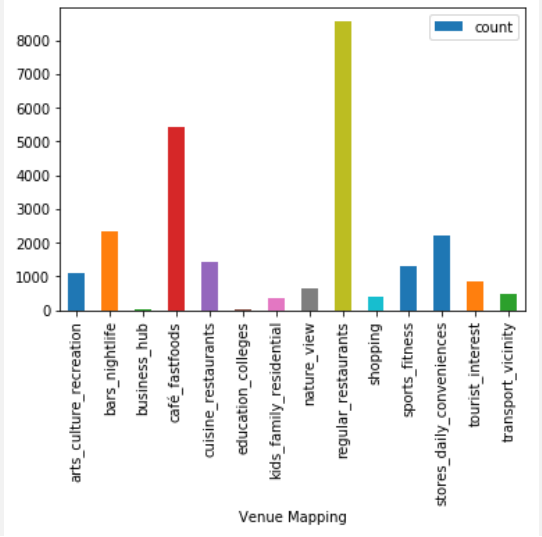
For Example, the following Venue Categories were mapped against sports_fitness.

Venue Category	Venue Type
Stadium	sports_fitness
Gym	sports_fitness
Gym / Fitness Center	sports_fitness
Salad Place	sports_fitness
Arcade	sports_fitness
Pool	sports_fitness
Athletics & Sports	sports_fitness
Basketball Court	sports_fitness
Sports Bar	sports_fitness
Sporting Goods Shop	sports_fitness
Bowling Alley	sports_fitness
Racetrack	sports_fitness
Yoga Studio	sports_fitness
Sports Club	sports_fitness
Gym Pool	sports_fitness
Track	sports_fitness

Moving Target	sports_fitness
Baseball Field	sports_fitness
Dance Studio	sports_fitness
Golf Course	sports_fitness
Soccer Field	sports_fitness
Cricket Ground	sports_fitness
Field	sports_fitness
Hockey Arena	sports_fitness
Tennis Court	sports_fitness
Recreation Center	sports_fitness
Soccer Stadium	sports_fitness
Club House	sports_fitness
Indoor Play Area	sports_fitness
Pool Hall	sports_fitness
Track Stadium	sports_fitness

6. After Mapping the venues, it was found that there is a heavy bias towards regular restaurants and cafes. To counter the effect of this on clustering of neighbourhoods, the frequency factors were reversed to highlight the less frequent venues like education and business hub, so that their effect can be more prominent on clustering.

INITIAL DISTRIBUTION OF VENUE CATEGORIES



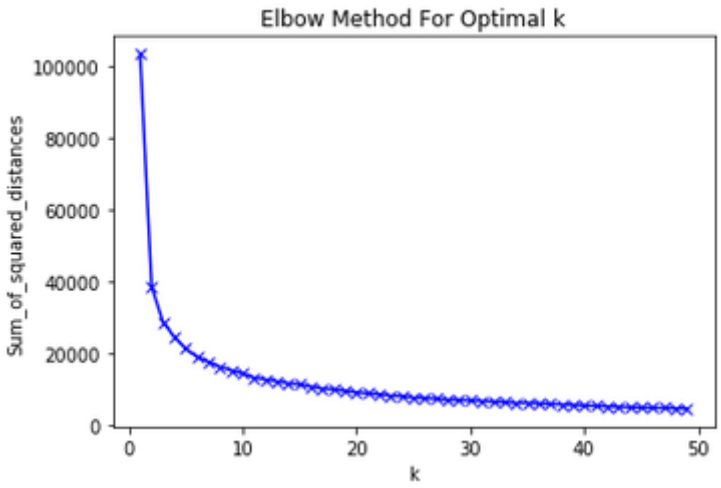
IMPORTANCE IN CLUSTERING

	Venue Mapping	count	importance
0	arts_culture_recreation	1086	0.023020
1	bars_nightlife	2324	0.010757
2	business_hub	28	0.892857
3	café_fastfoods	5406	0.004624
4	cuisine_restaurants	1440	0.017361
5	education_colleges	25	1.000000
6	kids_family_residential	369	0.067751
7	nature_view	630	0.039683
8	regular_restaurants	8542	0.002927
9	shopping	375	0.066667
10	sports_fitness	1310	0.019084
11	stores_daily_conveniences	2199	0.011369
12	tourist_interest	846	0.029551
13	transport_vicinity	480	0.052083

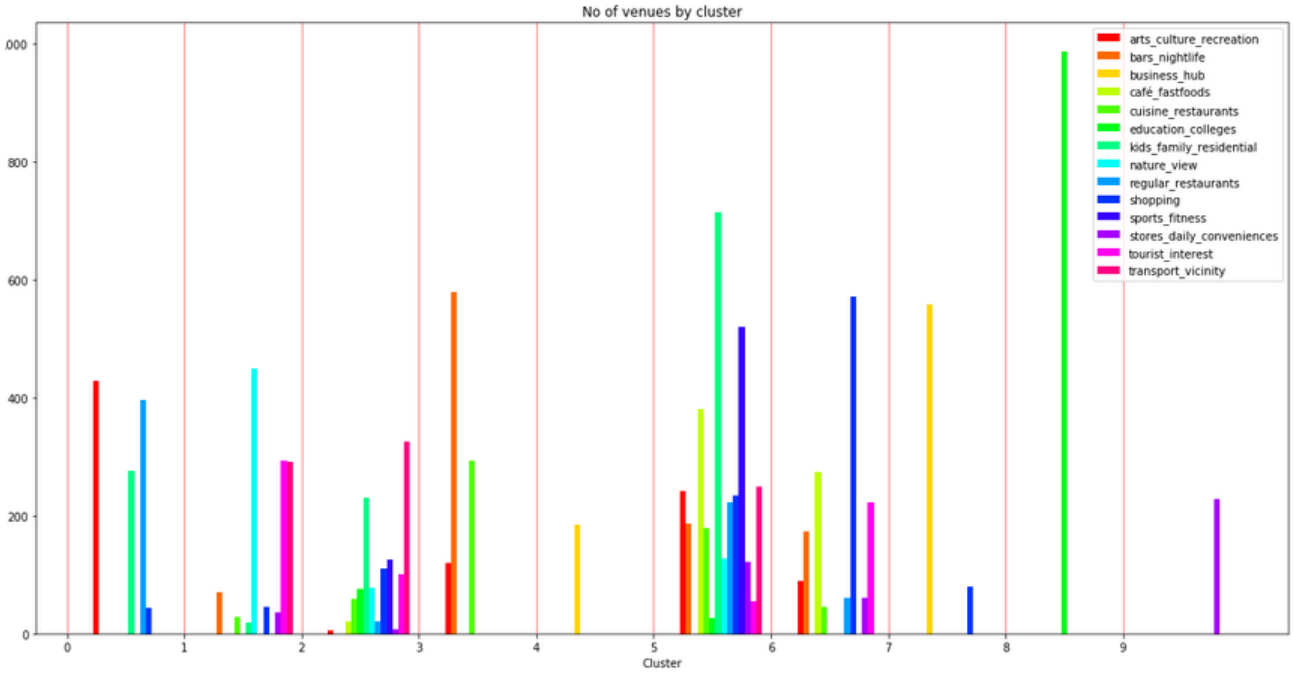
7. *One-hot Encoding* was used to assign dummy variables to each "venue_types". The sum of venues for each locality was multiplied with the "*importance in clustering".* Hence the data was prepared for clustering.

	arts_culture_recreation	bars_nightlife	business_hub	café_fastfoods	cuisine_restaurants	education_colleges	kids_family_residential	nature_view	regular_restaurant
0	0.115101	0.0645439	0.892857	0.0462449	0.0347222	0	0.0677507	0.0396825	0.058534
1	0.138122	0.129088	0.892857	0.0693674	0.121528	0	0	0.0396825	0.087801
2	0	0.0215146	0	0.0231225	0	0	0	0	0.02048
3	0.092081	0.0537866	0.892857	0.0601184	0	0	0.135501	0	0.058534
4	0.0690608	0.150602	0	0.101739	0.208333	1	0.135501	0.0793651	0.052680

8. K-Means Algorithm was used to cluster the neighbourhoods. The algorithm was iterated with cluster numbers 1 to 50, to decide the optimum degree for K-Means. We can see that the graph becomes fairly asymptotic at 10, so we can choose *10 as our number of clusters.*



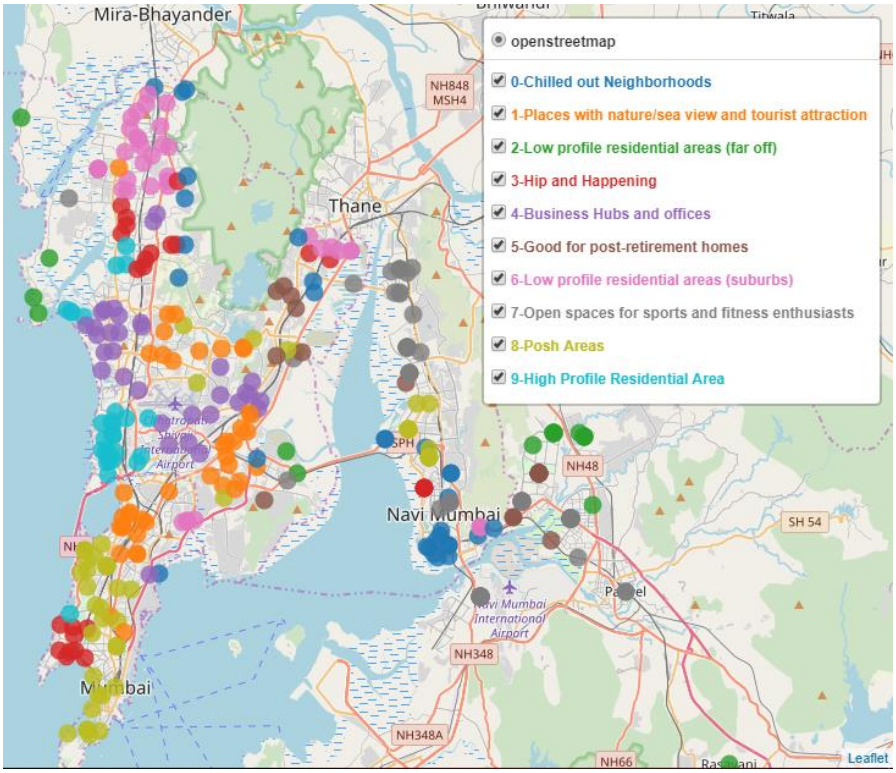
9. We can generate the most common venues in a cluster, which can help us to find a label for each cluster.



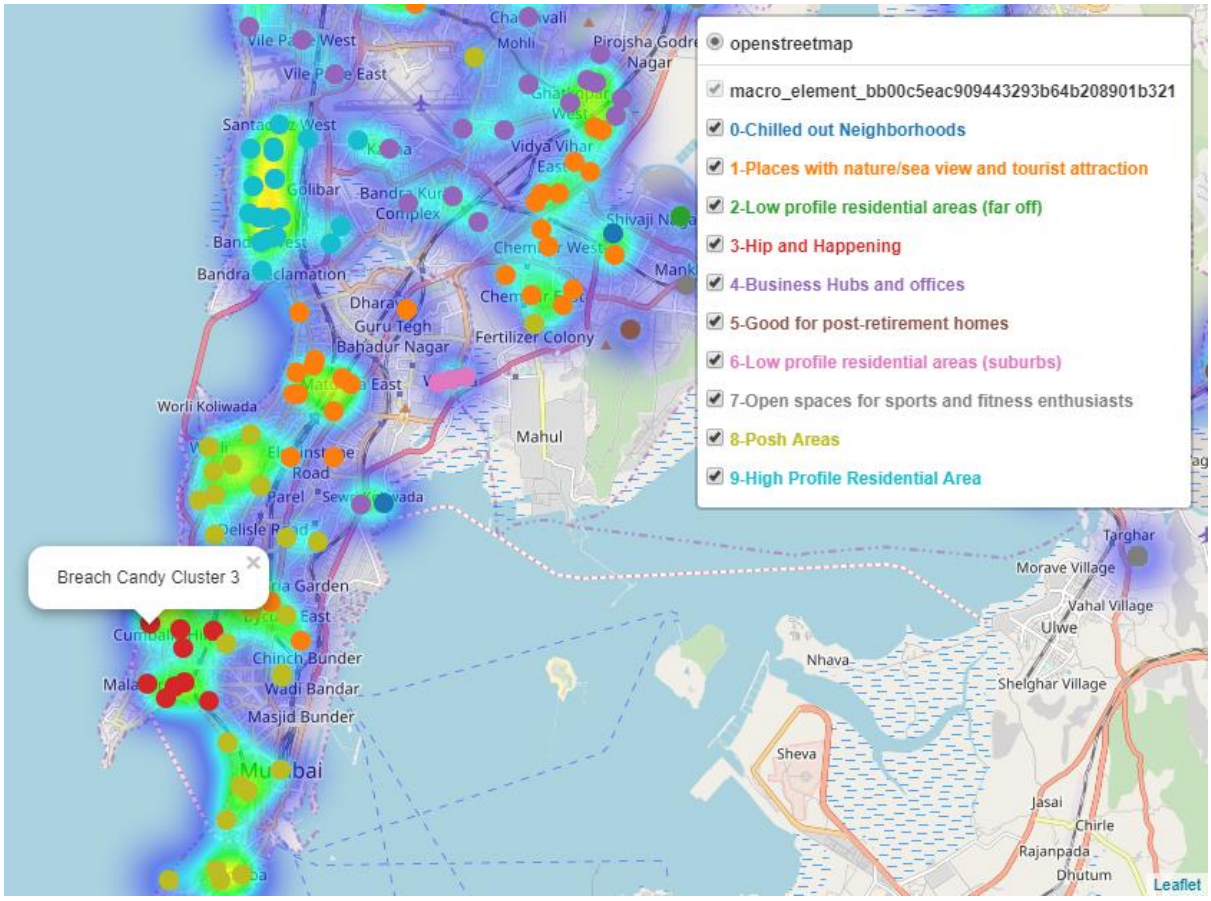
Examining the above graph, we can label each cluster as follows.

0	Good for arts and culture explorers	**5**	Good for post-retirement homes
1	Places with nature/sea view and tourist attraction	**6**	Low profile residential areas
2	Low profile residential areas	**7**	Good for sports and fitness enthusiasts
3	Hip and Happening	**8**	Posh Areas
4	Business Hubs and offices	**9**	High Profile Residential Area

10. We can plot the clusters on map to visualize the distribution of different types of Neighbourhoods in the City

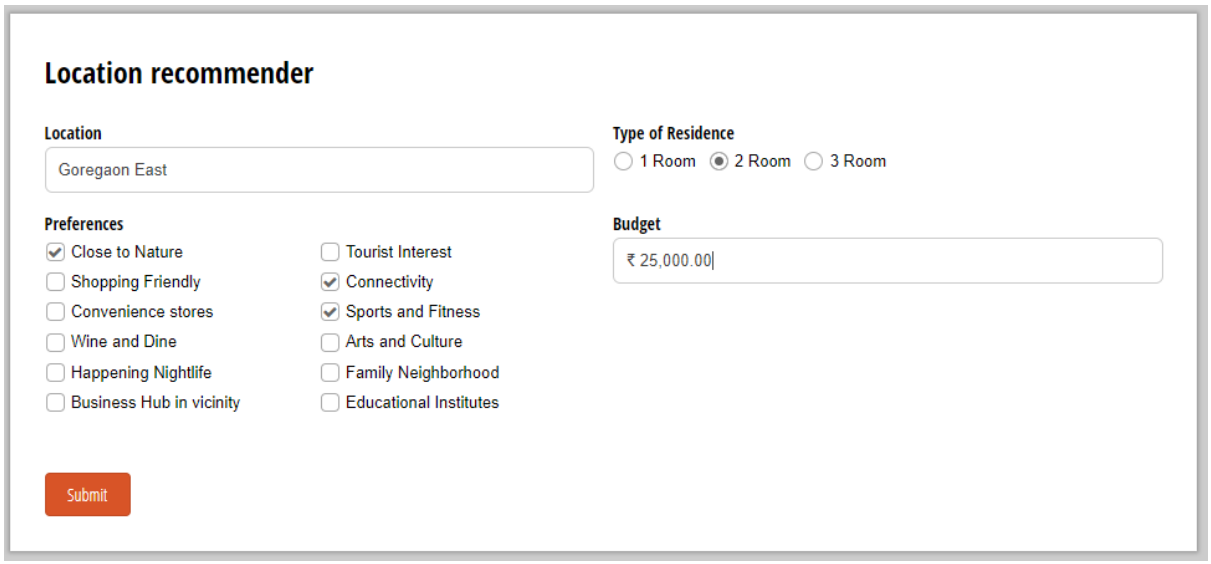


11. From the plot it is evident that similar types of areas are fairly well clustered by the *K-Means algorithm.* We can try to superimpose the price- heatmap with the cluster to visualize how Prices vary in the city with the type of Neighborhood. For example below is a snippet view of the superimposition of both the analysis for south Mumbai area



C. Discussion and Conclusion

People all over the world are turning to big cities to start a business or for work. This model can be used to further build a recommender system which can recommend most favoured locations as per the preferences of a user.



References:

[1] Press Information Bureau, Government of India
[2] Mumbai Population 2020
[3] India Census 2011 - Mumbai City Census 2011 data
[4] Municipal Corporation of Greater Mumbai - Mumbai portal with information and data
[5] Foursquare Developers Access to venue data: https://foursquare.com/