Problem Statement & Background

Mumbai is the largest city in India with more than 20 million residents. Mumbai city attacks talent from all over the country. People can find lot of good employment options in Mumbai. However, Mumbai is also very large city. It is desirable to stay in the locality so that commute to office is easy. if right area is not selected then it may take 4-5 hours just to commute to office one way. Sure, one can forget about the work then. So it is very important for a Mumbaikar to find the right house in the right locality. This project is developed to help fellow properly lookers to identify the locality where they want to search the listings.

There are multiple parameters one will like to evaluate before buying/leasing a property in a particular area. If I am currently staying in the Mumbai, I may want to find the property in the locality which is similar to my current locality. It may be possible; I want to live in the locality which provides venues which are similar to most posh localities of the city. However, Budget/affordability also plays a key role here.

Target Audience - This project will be useful to the people who are looking for buying/renting property in Mumbai. Using this project, one can find similar localities as per their current standards or aspiration standards. I have also put in price as a parameter for comparison so that they can finalize locality based on their budget.

Data Source

We will need 3 sets of data to complete this project

1. Locality data and housing prices data
2. Locality lat long data
3. locality venues data

I am planning to extract housing pricing data available at Magicbricks website. Magicbricks is one of the popular properly listing sites in the India. I am planning to use mapsquare tool for lat long data and I am planning to use Foursquare API to get venue details.

I have compiled the csv file using data available on https://www.magicbricks.com/Property-Rates-Trends/ALL-RESIDENTIAL-rates-in-Mumbai and have added lat long using MapSquare API.

Methodology

First step in the process was to understand data set available and clean the data. The data was not available in API format or a single table format which can be easily integrated in Notebook. (It can be understood as organisations will take some measures to protect their data). So i had to first combine data from multiple table locations on the magic bricks site and clean it using Microsoft excel tool. I also added lat long and created a csv file which I have hosted also on Google drive -

https://drive.google.com/file/d/178NGiwTy-mlibskihi0fOtaOPFe4IXgY/view?usp=sharing

I also hosted it in IBM database and used IBM Watson studio to run my Python notebooks.

I imported csv file into pandas dataframe as displayed in the image below.

In [10]:

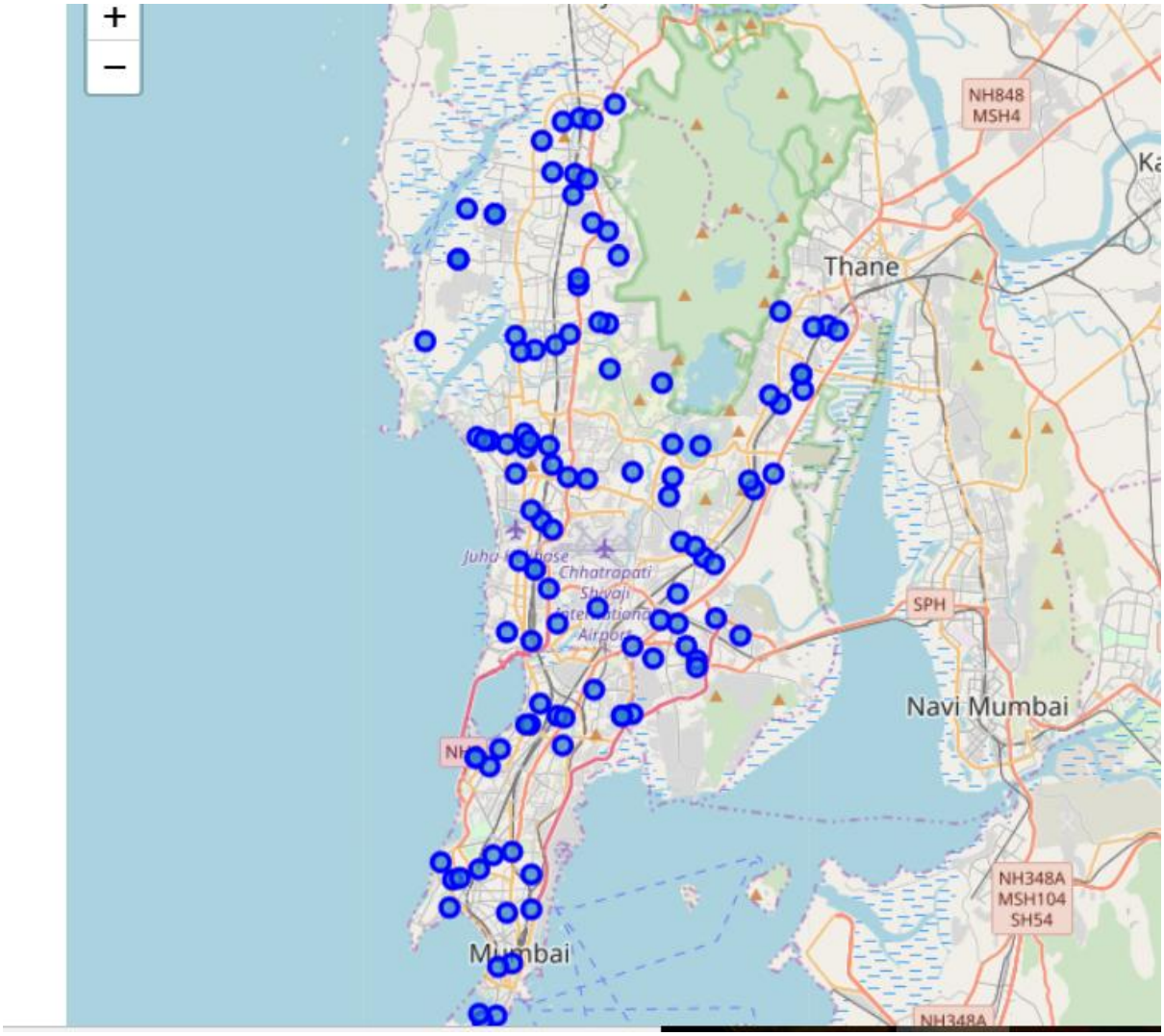
```
import types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

client_d5d9 = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='zHW3vWjFUlaiMClDzlq5SYQeB8AGWp6zG5SiDD-kBeeM',
    ibm_auth_endpoint="https://iam.eu-gb.bluemix.net/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url='https://s3.eu-geo.objectstorage.service.networklayer.com')

body = client_d5d9.get_object(Bucket='capstoneproject-donotdelete-pr-131jtxibkrupec',Key='Book1.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

df_Mum = pd.read_csv(body)
df_Mum.head()
```

Out[10]:

	Locality	Price	Latitude	Longitude
0	Aarey Milk Colony	13504	19.156129	72.870722
1	Agripada	26788	18.975302	72.824897
2	Alika Nagar	14228	19.198397	72.874267
3	Altamount Road	71016	18.966362	72.809148
4	Amboli	20908	19.127587	72.847115

Data for 116 localities within Mumbai area was imported along with avg price for a residential apartment in INR/sqft and lat, long the locality.

Next step, I just tried to visualize imported data to ensure data is properly captured or not.. I used folium library to visualize the data.



Next I used Foursquare API to get venues within 750 m radius from locality centre. I exported data in pandas data frame. Data was imported using explore venues API from Foursquare. Here is how my pandas table looked like.

```
[24]: Mumbai_venues.head(10)
```

Out[24]:

	Locality	Latitude	Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Aarey Milk Colony	19.156129	72.870722	cafe coffee day	19.159053	72.869031	Coffee Shop
1	Aarey Milk Colony	19.156129	72.870722	Panchvati Fast Food Corner	19.157628	72.874506	Fast Food Restaurant
2	Aarey Milk Colony	19.156129	72.870722	Chota Kashmir Boating	19.160905	72.872482	Lake
3	Aarey Milk Colony	19.156129	72.870722	Panchavati Fast Food Corner	19.157709	72.876511	Café
4	Aarey Milk Colony	19.156129	72.870722	Subway	19.158622	72.876653	Sandwich Place
5	Agripada	18.975302	72.824897	Celejor	18.975844	72.823679	Bakery
6	Agripada	18.975302	72.824897	Tote On The Turf	18.980266	72.820294	Nightclub
7	Agripada	18.975302	72.824897	cafe coffee day	18.976988	72.824051	Coffee Shop
8	Agripada	18.975302	72.824897	Neel	18.980407	72.820403	Indian Restaurant
9	Agripada	18.975302	72.824897	Maratha Mandir	18.971266	72.821819	Movie Theater

Then I used get_dummies function of pandas library to understand distribution of different kind of venues within each locality. We have to understand this in the order to define similarity among different localities. Types and Frequency of venues within localities is key parameter in our algorithm to understand similarity between 2 localities.

In [26]:
```
Mumbai_d = pd.get_dummies(Mumbai_venues[['Venue Category']], prefix="", prefix_sep="")

Mumbai_d['zLocality'] = Mumbai_venues['Locality']

fixed_columns = [Mumbai_d.columns[-1]] + list(Mumbai_d.columns[:-1])
Mumbai_d = Mumbai_d[fixed_columns]

Mumbai_d.head()
```

Out[26]:

s r	Sports Club	Stadium	Steakhouse	Supermarket	Tea Room	Tennis Court	Thai Restaurant	Theater	Track	Track Stadium	Trail	Train	Train Station	Travel & Transport	Vegetarian / Vegan Restaurant	Volleyball Court	Wine Shop	Women's Store	Yoga Studio	Zoo
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

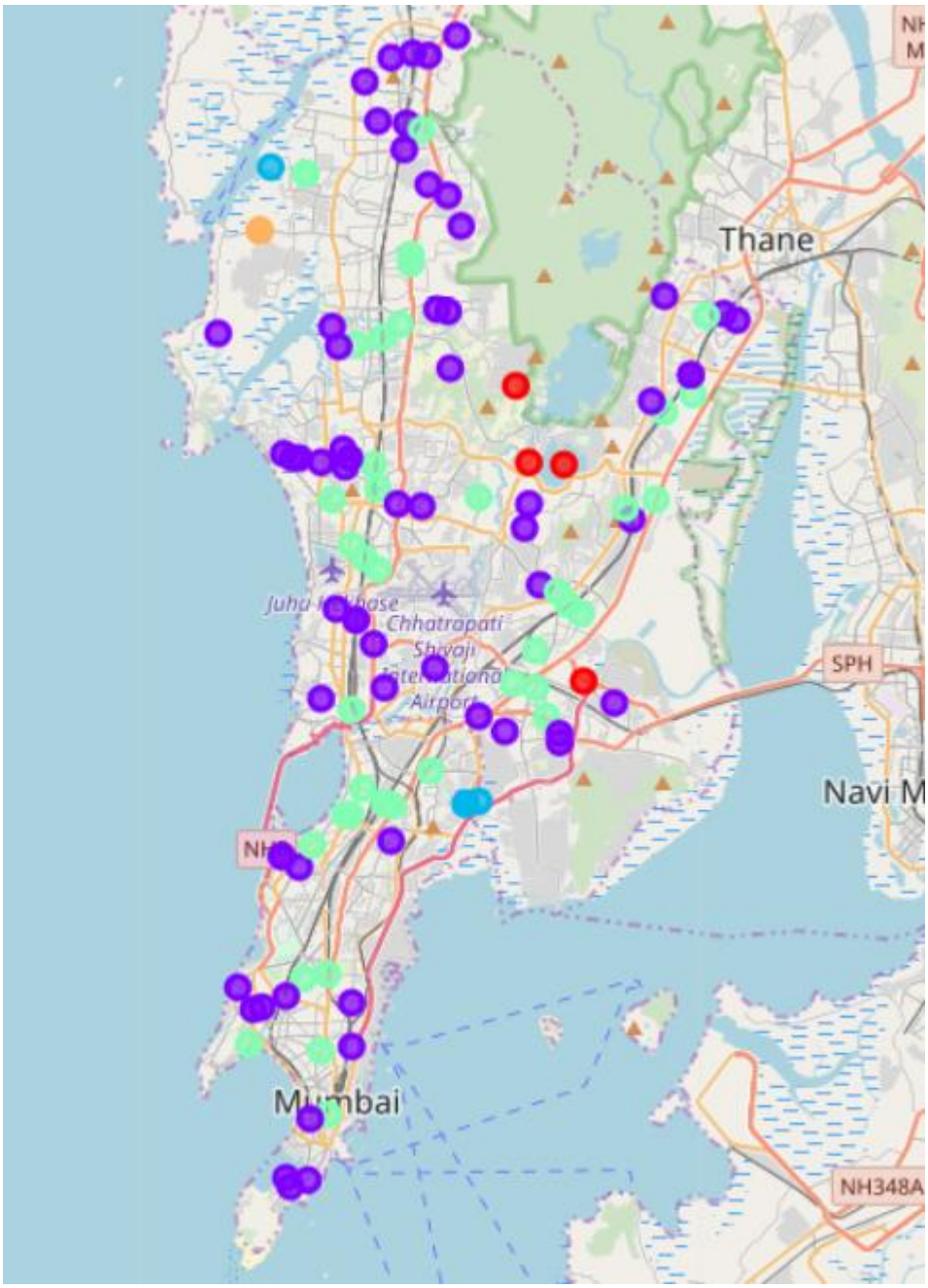
Then I identified 20 most common venues for each locality this will serve as base in comparison of locality using machine learning tool.

```
num_top_venues = 20

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Locality']
for ind in np.arange(num_top_venues):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Locality'] = Mumbai_venue_mean['zLocality']

for ind in np.arange(Mumbai_venue_mean.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = return_most_common_venues(Mumbai_venue_mean.iloc[ind, :], num_top_venues)

neighborhoods_venues_sorted.head()
```

ut[32]:

	Locality	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue
0	Aarey Milk Colony	Sandwich Place	Fast Food Restaurant	Lake	Café	Coffee Shop	Zoo	Factory	Flea Market	Fish Market	Field	Farmers Market	Farm	Falafel Restaurant	Event Space	Food
1	Agripada	Indian Restaurant	Bakery	Gym	Restaurant	Japanese Restaurant	Coffee Shop	Nightclub	Movie Theater	Fast Food Restaurant	Mediterranean Restaurant	Bank	Zoo	Farm	Farmers Market	Factory
2	Alika Nagar	Indian Restaurant	Coffee Shop	Sandwich Place	Shopping Mall	Gym / Fitness Center	Dessert Shop	Arts & Crafts Store	Food Truck	Ice Cream Shop	Chinese Restaurant	Soccer Field	Bakery	Plaza	Farm	Zo
3	Altamount	Café	Bakery	Sandwich	Coffee	Pizza	Salon /	Indian	Hotel	Chinese	Snack Place	Department	Lounge	Brewery	Italian	Bookstore

Once we had created the clusters, I tried to analyze each cluster. I started with visualization tool to check distribution of each cluster.



Then I listed details of each cluster to understand cluster in more detail

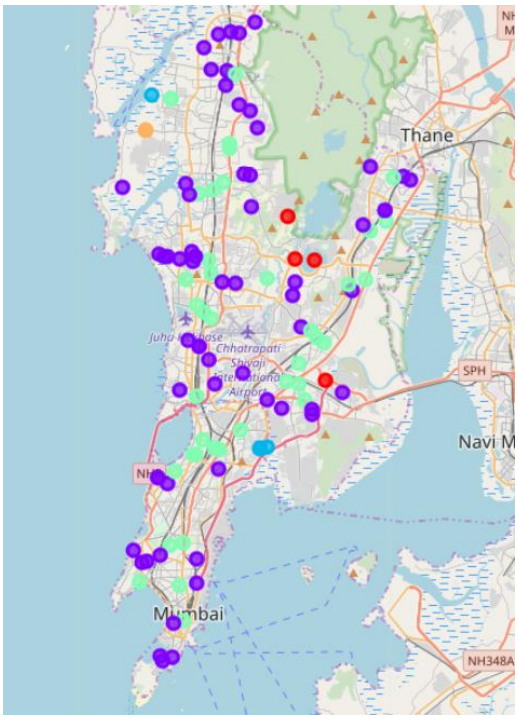
	Locality	Price	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aarey Milk Colony	13504	19.156129	72.870722	1	Sandwich Place	Fast Food Restaurant	Lake	Café	Coffee Shop	Zoo	Factory	Flea Market	Fish Market	Field	Farmers Market
2	Alika Nagar	14228	19.198397	72.874267	1	Indian Restaurant	Coffee Shop	Sandwich Place	Shopping Mall	Gym / Fitness Center	Dessert Shop	Arts & Crafts Store	Food Truck	Ice Cream Shop	Chinese Restaurant	Soccer Field
3	Altamount Road	71016	18.966362	72.809148	1	Café	Bakery	Sandwich Place	Coffee Shop	Pizza Place	Salon / Barbershop	Indian Restaurant	Hotel	Chinese Restaurant	Snack Place	Department Store
5	Anand Nagar	13990	18.966523	72.811888	1	Chinese Restaurant	Café	Pizza Place	Coffee Shop	Bakery	Vegetarian / Vegan Restaurant	Concert Hall	Fast Food Restaurant	Snack Place	Bookstore	Food Truck
7	Andheri East	18033	19.115883	72.854202	1	Indian Restaurant	Hotel	Ice Cream Shop	Chinese Restaurant	Vegetarian / Vegan Restaurant	Sandwich Place	Cocktail Bar	Burger Joint	Food Truck	Pizza Place	Camera Store
10	Asha Nagar	16495	19.210955	72.864322	1	Lounge	Movie Theater	Restaurant	Seafood Restaurant	Sporting Goods Shop	Food & Drink Shop	Snack Place	Skating Rink	Sandwich Place	Burger Joint	Food Truck
11	Ashok Nagar Central Mumbai	14841	19.052995	72.879788	1	Café	Electronics Store	Performing Arts Venue	Soccer Field	Vegetarian / Vegan Restaurant	Food	Hotel	Indian Restaurant	Zoo	Farm	Farmers Market
13	Azad Nagar 2	21730	19.092060	72.899401	1	Café	Food Truck	Factory	Indian Restaurant	Donut Shop	Food	Flower Shop	Flea Market	Fish Market	Field	Fast Food Restaurant
14	Azad Nagar Versova Road	20355	19.126910	72.837648	1	Bar	Indian Restaurant	Pub	Chinese Restaurant	Coffee Shop	Pizza Place	Park	Burger Joint	Sandwich Place	Bowling Alley	Smoke Shop
16	Bandra East	31525	19.061514	72.850093	1	Indian Restaurant	Restaurant	Café	Pizza Place	Fast Food Restaurant	Spa	Bar	French Restaurant	Bistro	Chinese Restaurant	Sports Club

Results

I could divide Mumbai city in 5 clusters with 2 large clusters and 3 small clusters. Mumbai real estate prices range from 5k INR/sqft to 72k INR/Sqft.

When we analyze the distribution of the nodes in graphical format, we can see than both large clusters are spread across the Mumbai.

However smaller nodes are more aligned to geography. Like one smaller node is near to national park, other smaller node is kind of an small island itself.



Discussion

Mumbai city locality analysis highlights that there are broadly 2 sets of locality within Mumbai city. Large two clusters are spread across the Mumbai (purple and green dots) So both the types are also uniformally spread across Mumbai city. So if one is looking to move from one part of the city to any other part of the city, this person will not have to compromise on the lifestyle. This may be one reason for large mobility within Mumbai area.

Conclusion

This project helps users in the first step of home search. User will need to find the right locality before s/he starts looking for the apartments in particular locality. Identification of right locality using machine leaning and based on user preference can save a lot of time and effort for the user to go and explore the locality. It may also be biased and hard to conclude. This project will also help user to filter localities based on user budget and preference.

Users should use this project to identify the ideal locality for them and then start looking for apartments within that locality for best results.