Capstone Project - The Battle of Neighbourhoods

Restaurants in Mumbai Metropolitan Region

1. Description of the problem and a discussion of the background.

With the population of 20.1 million, Mumbai is the largest city in India in terms of area as well as population. Mumbai Metropolitan Region (MMR) consists of Mumbai and its satellite towns of Thane and Navi Mumbai. The region has an area of 6,355 square kilometres and with a population of over 26 million it is among the most populous and ethnically diverse metropolitan areas in the world.

With a natural harbour, it's one of the most economically important locations in the world and is the core of the country's economy. It has always had a history of influx of people , in search of better prospects and opportunities. With Mumbai emerging as a growing hub of commerce and with rapid globalization , the taste spectrum of the inhabitants has broadened have begun preferring wide variety of cuisines. These has opened up great opportunities for multi-cuisine restaurants , adhering to the cravings of its diverse population.

So, this project deals with how we can leverage Foursquare Data and machine learning to help us make decision and find appropriate, suitable neighbourhoods based on various cuisines and economic point of view. Opening a eatery or restaurant here is an lucrative idea for a enterprise or individual who want to extends its business to Asia. They would be very interested in this project. Also, people looking for best places to try out their favoured cuisine can refer this.

2. Data source and how it will be used to solve the problem:

- https://www.kaggle.com/srivpuneet16/zomato-mumbai-restaurant-analysis

  Description: Contains a Zomato Csv file containing database of all active restaurant locations in MMR. The locations of these can be extracted and the coordinates can be found out using GeoPy .This is later accessed to gather Foursquare Data.

• Restaurants in each neighbourhood of Mumbai:

Data source: Foursquare APIs

By using this API we will get all the venues in each neighbourhood. We can filter these venues to get only restaurants.

3. Methodology

3.1 .1 Data Preparation

Firstly I use the Zomato CSV file from Zaggle, that has the database of all active restaurants in the Mumbai Metropolitan Region. I extract the unique locations in a Dataframe using Pandas.

```
df= pd.read_csv("zomato_mum.csv")
```

```
df.head()
```

	Additional_outlet_count	Call	Cost_for_two(Rs.)	Cuisines	Features	Home_Delivery	Operational_hours	Restaurant_Location	Restaurant_Name	Restaur
0	1.0	True	1500	Finger Food, Continental, European, Italian	Food Hygiene Rated Restaurants In Mumbai, Best...	False	12noon – 1am (Mon-Sun)	Kamala Mills Compound	Lord of the Drinks	Loun
1	1.0	True	800	Pizza	Value For Money, Best of Mumbai	False	11am – 12:30AM (Mon-Sun)	Malad West	Joey's Pizza	C
2	NaN	True	2500	Seafood	Super Seafood, Best of Mumbai	False	Closed (Mon),12noon – 3pm, 7pm – 12midnight...	Bandra West	Bastian	Casual
3	NaN	True	1800	Finger Food, Continental	Where's The Party?, Best of Mumbai, Food Hygie...	False	12noon – 1am (Mon-Sun)	Lower Parel	Tamasha	L
4	2.0	True	450	North Indian, Street Food, Fast Food, Chinese	NaN	True	12noon – 4pm, 7pm – 11:45pm (Mon-Sun)	Vashi	Bhagat Tarachand	Cas

5 rows × 22 columns

After extracting the unique Restaurant locations , we get the following Dataframe with the Location column. Next we need the coordinates of each location

	Location
1	Malad West
2	Bandra West
3	Lower Parel
4	Vashi
5	Bandra Kurla Complex
6	Juhu
7	Cuffe Parade
8	Andheri West
9	Powai
10	Dadar East
11	Santacruz West
12	Fort
13	Khar
14	Marol
15	Borivali West
16	Churchgate
17	Byculla
18	Kandivali West
19	Colaba

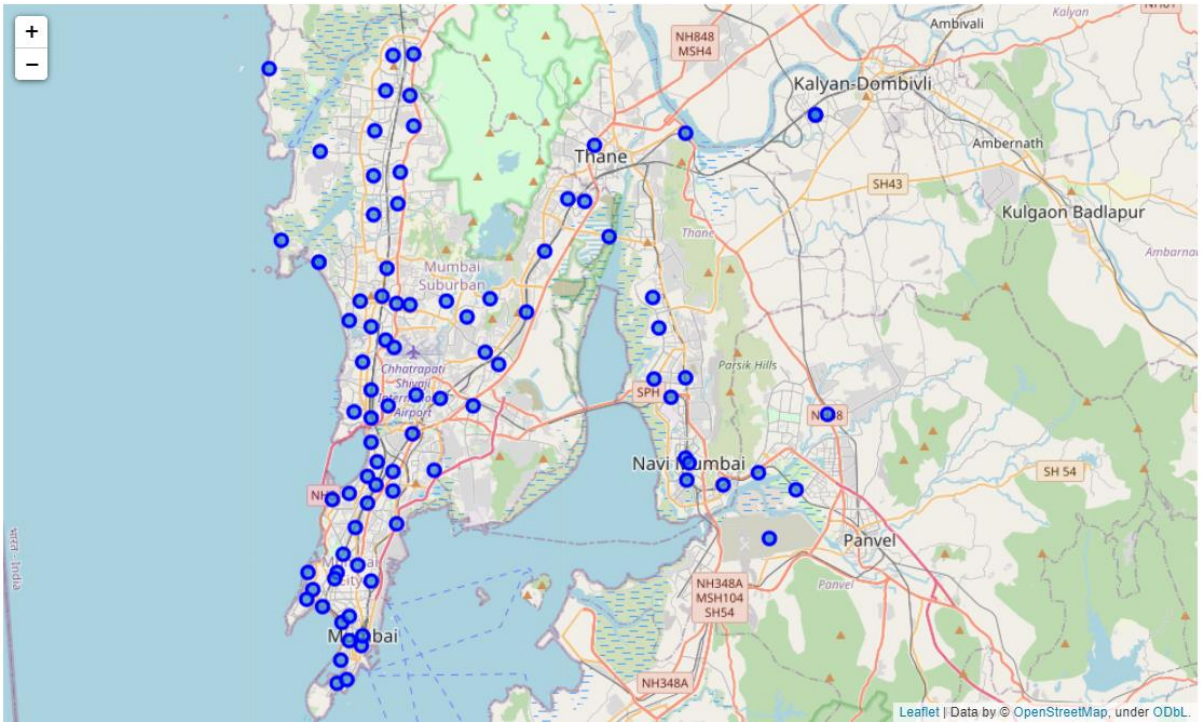
3.1.2 Getting Coordinates of Locations : GeoPy client

Now we get coordinates of 93 locations within Mumbai , using Geocoder from the GeoPy client

```
l1=df_split[0]
l1['coordinates'] = l1['Location'].apply(geolocator.geocode).apply(lambda x: (x.latitude, x.longitude))
l1[['Latitude', 'Longitude']] = l1['coordinates'].apply(pd.Series)
l1.drop(['coordinates'], axis=1, inplace=True)
l1
```

	Location	Latitude	Longitude
1	Malad West,Mumbai	19.184013	72.841216
2	Bandra West,Mumbai	19.058336	72.830267
3	Lower Parel,Mumbai	18.996332	72.830860
4	Vashi,Mumbai	19.075713	73.000354
5	Bandra Kurla Complex,Mumbai	19.067115	72.865724
6	Juhu,Mumbai	19.107021	72.827528
7	Cuffe Parade,Mumbai	18.913641	72.820930
8	Andheri West,Mumbai	19.117249	72.833968
9	Powai,Mumbai	19.118720	72.907348
10	Dadar East,Mumbai	19.016253	72.852227

Now, I used Folium Python library to geographically visualize the parts of Mumbai. The following map was created using Folium, and the latitudes and longitudes of the Locations.



3.2 Exploratory analysis:

Now, I used Exploratory Data analysis to get subtle properties from the Data that can provide useful insights to an investor or any resident.

3.2.1 Using Foursquare Location Data

Now, we make use of Foursquare API to get top venues in let's say ,
Dadar locality in Mumbai within a radius of 500 meters.

We notice that there are 19 unique categories of data returned from Foursquare from Dadar.

```
print ('{} unique categories in Dadar '.format(nearby_venues['categories'].value_counts().shape[0]))
```

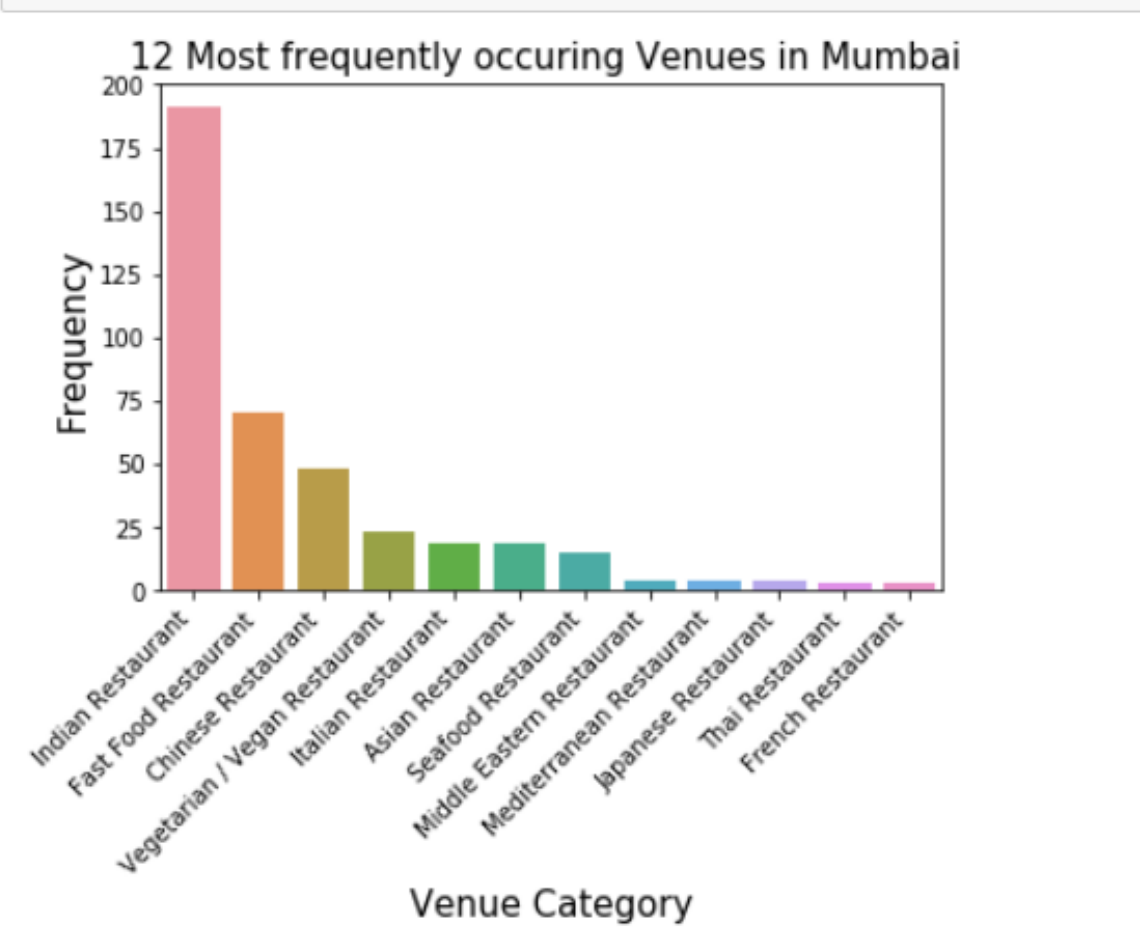
```
19 unique categories in Dadar
```

```
print (nearby_venues['categories'].value_counts()[0:10])
```

```
Indian Restaurant          5
Fast Food Restaurant       3
Coffee Shop                2
Movie Theater              2
Café                       2
Farmers Market             2
Flower Shop                1
Plaza                      1
Maharashtrian Restaurant   1
Women's Store              1
Name: categories, dtype: int64
```

Now, we collect the similar data from all locations together and retrieve venues of type 'Restaurant' from the 93 locations searched.

We find that there are 27 unique categories of restaurants , scattered all over the locations. Indian restaurant tops the charts o-f frequencies of categories as we can see below in the visualization



As we can see Indian restaurants have the highest frequency (for obvious geographical reasons) and also because the Indian cuisine is very extensive and also delicious.

Further we try to get the top 5 restaurants in each location for further analysis.

We proceed as follows:

1) Create Dataframe with pandas Onehot encoding with Venue Categories.

```
Mum_onehot = pd.get_dummies(Mum_Venues_only_restaurant[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
Mum_onehot['Neighborhood'] = Mum_Venues_only_restaurant['Neighborhood']
cols = Mum_onehot.columns.tolist()
cols = cols[-1:] + cols[:-1]
Mum_onehot = Mum_onehot[cols]
Mum_onehot.head()
```

	Neighborhood	American Restaurant	Asian Restaurant	Bengali Restaurant	Chinese Restaurant	Fast Food Restaurant	French Restaurant	German Restaurant	Goan Restaurant	Gujarati Restaurant	...	Mughlai Restaurant	Multicuisine Indian Restaurant	R
1	Malad West	0	0	0	0	1	0	0	0	0	...	0	0	
2	Bandra West	0	0	0	0	0	1	0	0	0	...	0	0	
3	Bandra West	0	0	0	0	0	0	0	0	0	...	0	0	
4	Bandra West	0	0	0	0	0	0	0	0	0	...	0	0	
5	Bandra West	0	0	0	1	0	0	0	0	0	...	0	0	

5 rows × 28 columns

2) We group the Dataframe with Neighbourhood and Calculate the mean of frequency of occurrence of each venue category at all locations.

Grouping rows by neighborhood , by taking the mean of the frequency of occurrence of each category

```
Mum_grouped = Mum_onehot.groupby('Neighborhood').mean().reset_index()
Mum_grouped
```

	Neighborhood	American Restaurant	Asian Restaurant	Bengali Restaurant	Chinese Restaurant	Fast Food Restaurant	French Restaurant	German Restaurant	Goan Restaurant	Gujarati Restaurant	...	Mughlai Restaurant	Multicuisine Indian Restaurant	R
0	Andheri	0.000000	0.000000	0.0	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
1	Andheri East	0.000000	0.000000	0.0	0.166667	0.500000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
2	Andheri West	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
3	Bandra	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
4	Bandra East	0.000000	0.000000	0.0	0.250000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
5	Bandra Kurla Complex	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.142857	...	0.000000	0.000000	
6	Bandra West	0.000000	0.111111	0.0	0.185185	0.074074	0.037037	0.000000	0.000000	0.000000	...	0.000000	0.000000	
7	Bhandup	0.000000	0.000000	0.0	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
8	Borivali East	0.000000	0.000000	0.0	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
9	Borivali West	0.000000	0.111111	0.0	0.333333	0.111111	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	
10	Breach Candy	0.000000	0.000000	0.0	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	

We get top 5 venues for each Location in Mumbai.

Top 5 venues in each Locality

In [279]:
```
num_venues = 5

for hood in Mum_grouped['Neighborhood']:
    print("-*-*-"+hood+"-*-*-")
    temp = Mum_grouped[Mum_grouped['Neighborhood'] == hood].T.reset_index()
    temp.columns = ['venue','freq']
    temp = temp.iloc[1:]
    temp['freq'] = temp['freq'].astype(float)
    temp = temp.round({'freq': 2})
    print(temp.sort_values('freq', ascending=False).reset_index(drop=True).head(num_venues))
    print('\n')
```

```
-*-*-Andheri -*-*-
                       venue  freq
0       Fast Food Restaurant   0.5
1          Indian Restaurant   0.5
2        American Restaurant   0.0
3  Middle Eastern Restaurant   0.0
4            Thai Restaurant   0.0
```

Finally, we use K-Means clustering to cluster all 93 locations into clusters according to similarity in their respective venue categories. We set K value to 5 , in accordance to such high number of localities , needed to be clustered.

Run k-means to cluster the neighborhood into 5 clusters.

In [306]:
```
kclusters = 5

Mum_grouped_clustering = Mum_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(Mum_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

Out[306]:
```
array([0, 0, 1, 1, 1, 1, 4, 0, 0, 4])
```

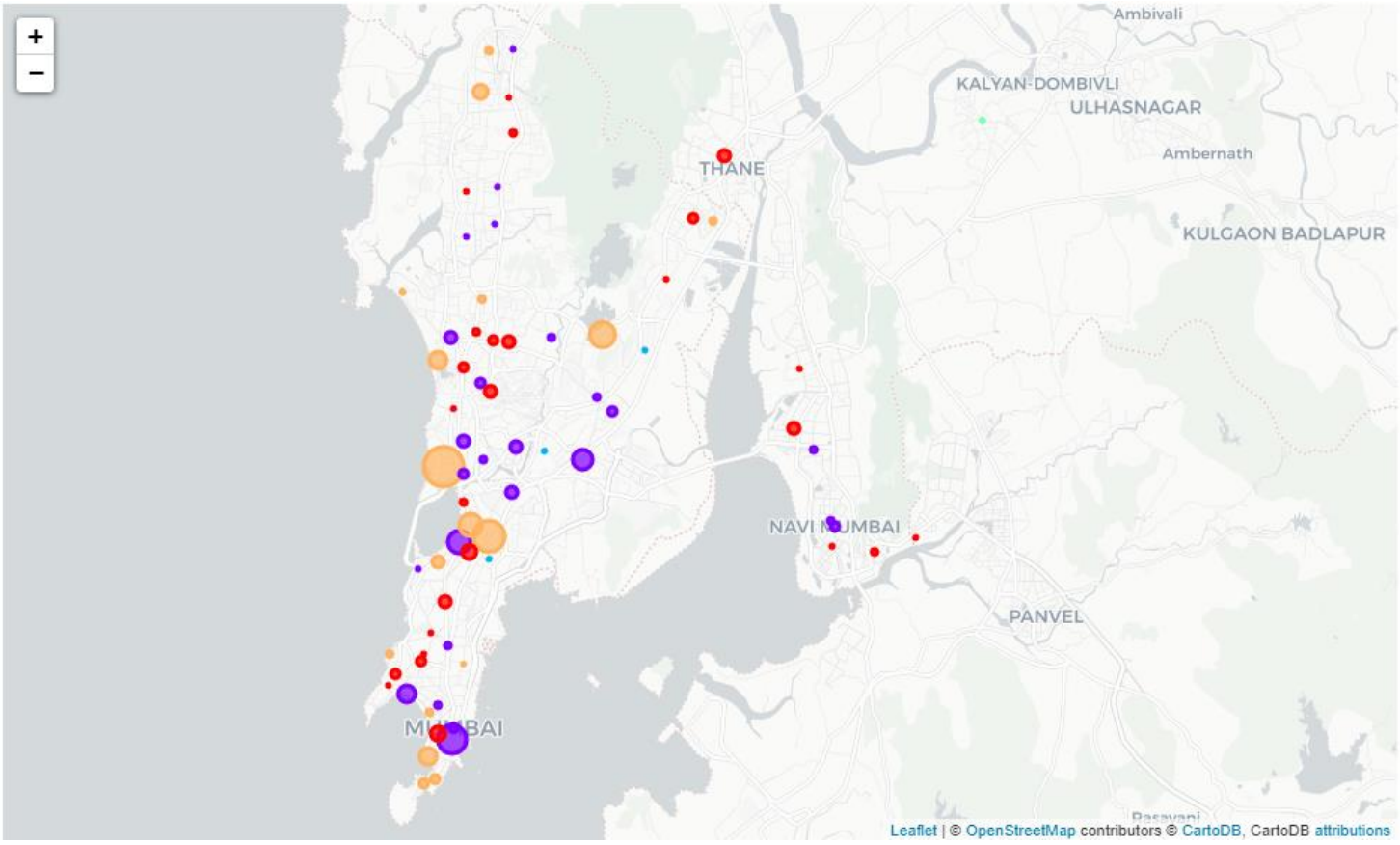
In [307]:
```
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

Mum_merged = df_f

Mum_merged.rename(columns={'Location':'Neighborhood'}, inplace=True)

# merge to add latitude/longitude for each neighborhood
Mum_merged = Mum_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

Mum_merged.head() # check the last columns!
```

Out[307]:

	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Malad West	19.184013	72.841216	0.0	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Maharashtrian Restaurant	Asian Restaurant	Bengali Restaurant
2	Bandra West	19.058336	72.830267	4.0	Indian Restaurant	Chinese Restaurant	Asian Restaurant	Fast Food Restaurant	Vegetarian / Vegan Restaurant
3	Lower Parel	18.996332	72.830860	0.0	Fast Food Restaurant	Vegetarian / Vegan Restaurant	Thai Restaurant	Chinese Restaurant	Indian Restaurant
4	Vashi	19.075713	73.000354	0.0	Fast Food Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant	Middle Eastern Restaurant	Maharashtrian Restaurant
5	Bandra Kurla Complex	19.067115	72.865724	1.0	Indian Restaurant	Gujarati Restaurant	Italian Restaurant	Vegetarian / Vegan Restaurant	Maharashtrian Restaurant

We get a Dataframe containing the cluster each Location belongs to , along with its top 5 venue categories.

Now, We can visualize the 5 clusters on a leaflet map like below:



Cluster 1 : Red

Cluster 2 : Violet

Cluster 3 : Light Blue

Cluster 4 : Green

Cluster 5 : Orange

These 5 clusters are scattered geographically but are grouped according to similarity in Restaurant categories. The cluster results tell a great deal about the general scheme of restaurant localisation in Mumba

4. Result and Discussion:

Through this we got a glimpse of the restaurants in Mumbai and were able to derive some interesting insights which might be useful for tourists, travellers, and investors. Some of the main findings are :

- Indian Restaurants are on the top charts and are followed by Fast food Restaurants. There are plenty of Chinese and Asian restaurants scattered.
- Bandra, Dadar, Matunga and Powai has the highest no of Restaurants. Mainly because these are both residential and working areas.
- Malad, Borilvali , Dahisar, Worli have low Restaurant frequencies.
- Cluster 1 has more Fast food Restaurants than others which may be a result of more malls and educational institutions located here.
- Cluster 5 has more posh, high standard well off areas , with many Corporate locations like South Mumbai and Powai. Hence, we can see more diverse categories of restaurants and a variety of cuisines to choose from .

The Locations and clustering is completely based on the most common venues from Foursquare data . We haven't considered a lot of factors here like the commute of daily travellers , which is a very important factor in Mumbai. Factors like price of restaurants, ratings and so on. Also , Foursquare data may miss out on some Venues and Restaurants which may influence this study. Hence , this just gives one an overview of the Restaurant distribution in MMR .
Also , various other clustering and visualization techniques can be used that may give a different inference.

5. Scope and Limitations:

- All the inferences are based solely on Foursquare Data .
- The Zomato data can be used(which has varies other parameters like ratings, timings etc) along with Foursquare Data to get a more comprehensive analysis of Restaurants in Mumbai.
- Data with Factors like Standard of Living, Income, Population density , rent can be integrated with this to make it a more generalized model.
- Analysis can be done for specific categories of restaurants that help investors .

6. Conclusion:

Any real-life scenarios and problems can be simplified using insights derived from studying data. Here, in this example , I have used Foursquare data to cluster the Restaurants in 93 localities of the Mumbai Metropolitan region , in the view that It could help a tourist or an investor make certain decisions.

I have used Python libraries to extract and clean data. Used Folium to Visualize clusters geographically and used K-means to cluster the 93 localities .

This project is a small depiction of analysis and can have numerous variations. Many other real-life problems that people face in metro cities can be solved using data. Also, there is always some limitation and scope of adding more parameters to the analysis to make is more complete and comprehensive, as mentioned above.

7. References:
   - Kaggle Dataset: https://www.kaggle.com/srivpuneet16/zomato-mumbai-restaurant-analysis
   - Foursquare API : https://developer.foursquare.com/