Coursera Capstone

IBM Applied Data Science Capstone

Battle of Cafes -Mumbai

Introduction:

Coffee being a most consumed beverage in the world. This project is for those who love coffe much and always wanted to explore coffee shops around cities.Basically it will help tourists/coffee to explore nearby coffee shops and also people who want to open a new coffee shop and looking for a less or more crowded place in Mumbai. With the help of data science we can group places or neighborhoods according to number of cafes available also with advanced business analysis we can advice any one who wants to open a new coffee shop

Business Problem:

The objective of this capstone project is to guide tourists so that they can stay nearby most of coffee shops or explore most of the coffee shops and also to analyse a proper location to open new coffee shop with less competition around using the power of data and machine learning

Targeted Audience:

- Tourists
- Coffee Lovers
- Businessmen
- Marketing companies for advertising campaigns

Data Desciption:

I have used Foursquare API As it is mandatory it provides Neighborhood,Neighborhood Latitude,Neighborhood Longitude,Venue,Name of the venue e.g. the name of a store or restaurant,Venue Latitude,Venue Longitude,Venue Category just by passing parameters like client id,client secret and latitude and longitude

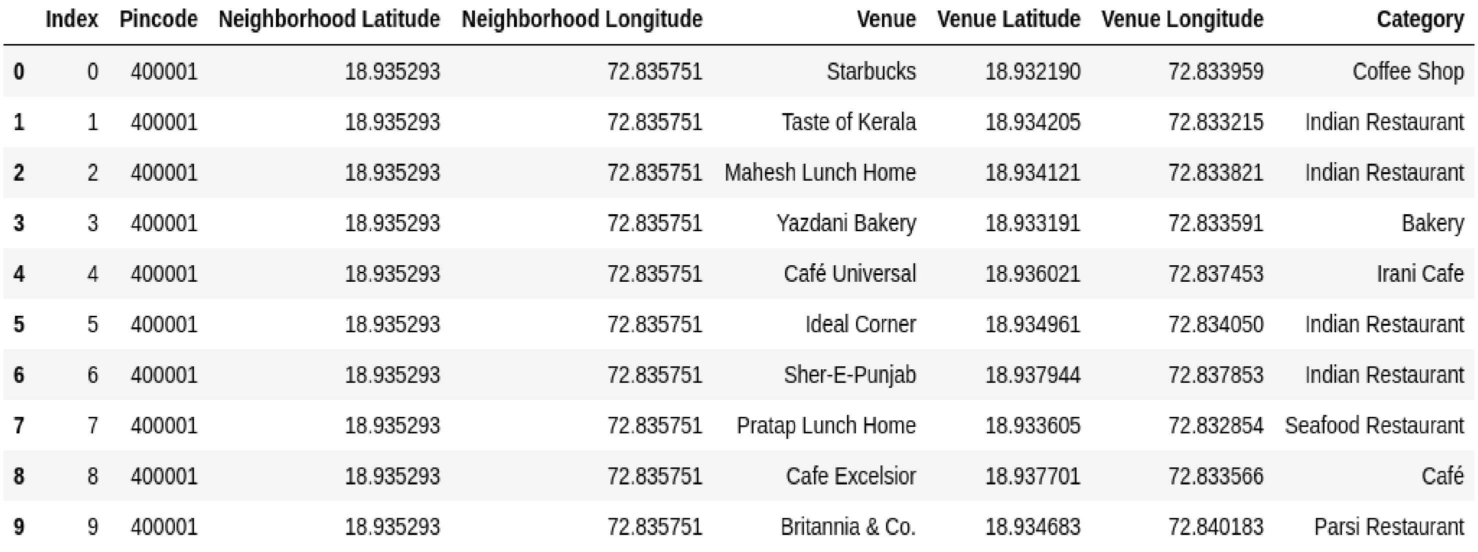
To solve the problem, we will need the following data:

- List of neighbourhoods in Mumbai .This defines the scope of this project which is confined to the city of Mumbai
- Latitude and longitude coordinates of those neighbourhoods.This is required in order to plot the map and also to get the venue data.
- Venue data, particularly data related to venues.

As there is no data available of various areas ,zipcodes/pincodes, coordinates of Mumbai so i have scraped them from website :-

- https://www.mapsofindia.com/pincode/india/maharashtra/mumbai/
- https://geographic.org/streetview/india/maharashtra/konkan/mumbai.html

After Scarping data from these sites and from Foursquare API i get the dataset consiting of 3047 rows and 8 columns
following image shows first 10 rows



	Index	Pincode	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Category
0	0	400001	18.935293	72.835751	Starbucks	18.932190	72.833959	Coffee Shop
1	1	400001	18.935293	72.835751	Taste of Kerala	18.934205	72.833215	Indian Restaurant
2	2	400001	18.935293	72.835751	Mahesh Lunch Home	18.934121	72.833821	Indian Restaurant
3	3	400001	18.935293	72.835751	Yazdani Bakery	18.933191	72.833591	Bakery
4	4	400001	18.935293	72.835751	Café Universal	18.936021	72.837453	Irani Cafe
5	5	400001	18.935293	72.835751	Ideal Corner	18.934961	72.834050	Indian Restaurant
6	6	400001	18.935293	72.835751	Sher-E-Punjab	18.937944	72.837853	Indian Restaurant
7	7	400001	18.935293	72.835751	Pratap Lunch Home	18.933605	72.832854	Seafood Restaurant
8	8	400001	18.935293	72.835751	Cafe Excelsior	18.937701	72.833566	Café
9	9	400001	18.935293	72.835751	Britannia & Co.	18.934683	72.840183	Parsi Restaurant

Methodology:

Exploratory Data Analysis:

Exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. I have perfom some basic EDAs using Seaborn library

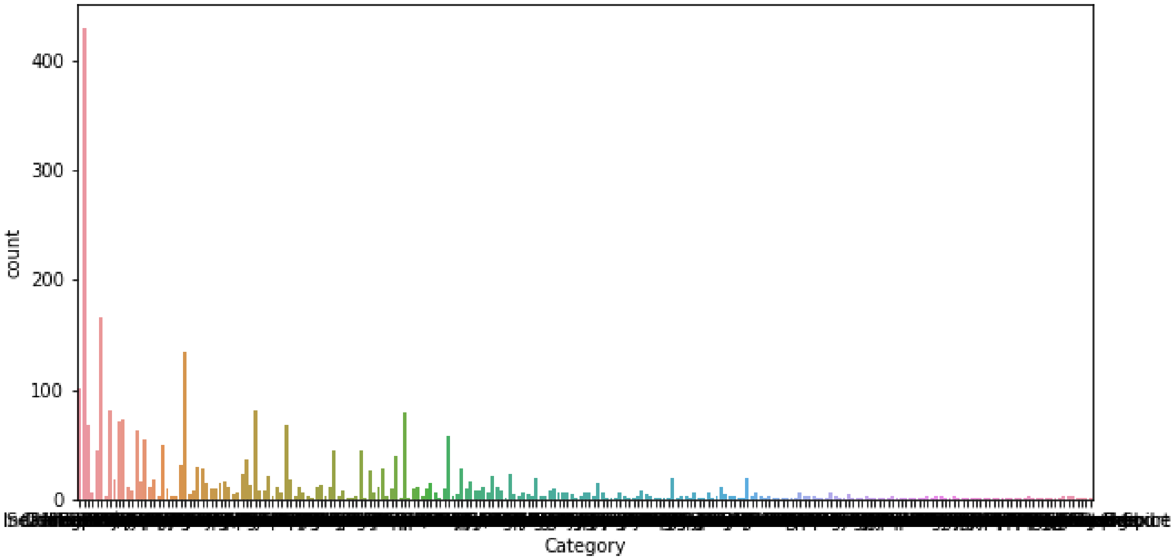


fig 1 .Counting all number of Categories present in Dataset

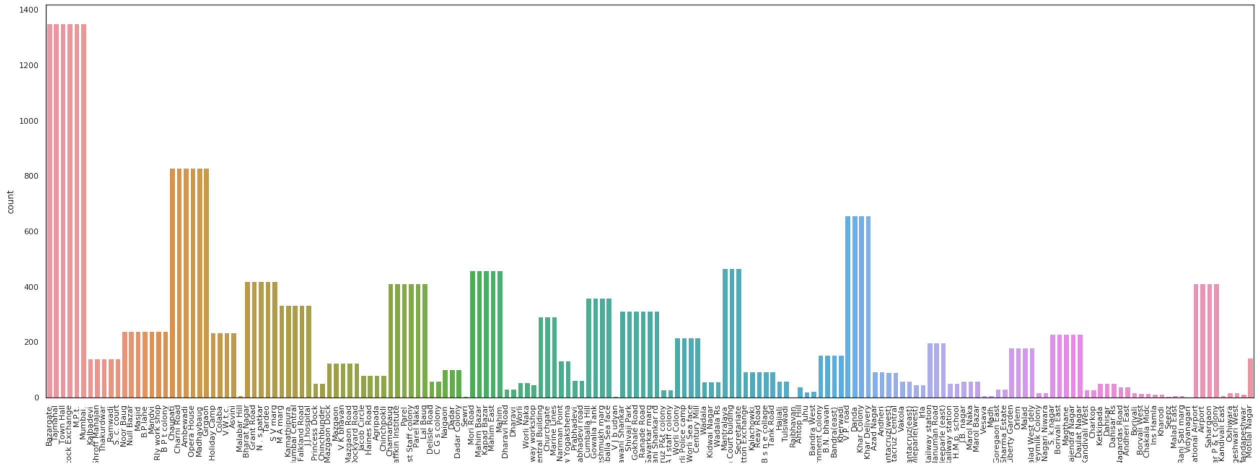


fig 2.Counting locations present in dataset

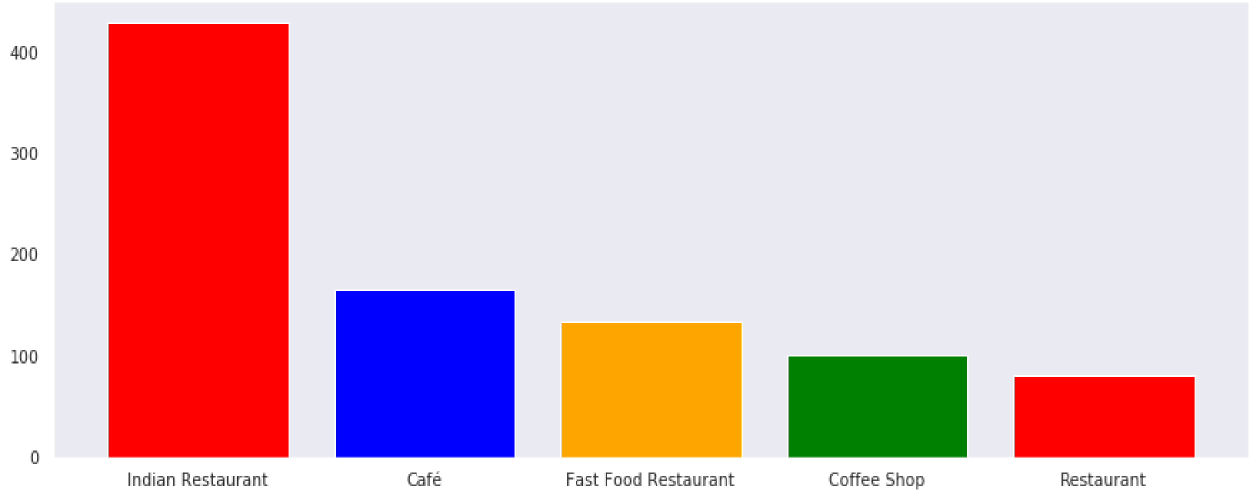


fig 3 .Plotting Highest Number of Categories

The above figure shows Top five categories venues with highest number. As we can see indian restaurant are high in number with more than 400 count and count of cafes are more than 150 . so our aim is to cluster these cafes and neighborhood

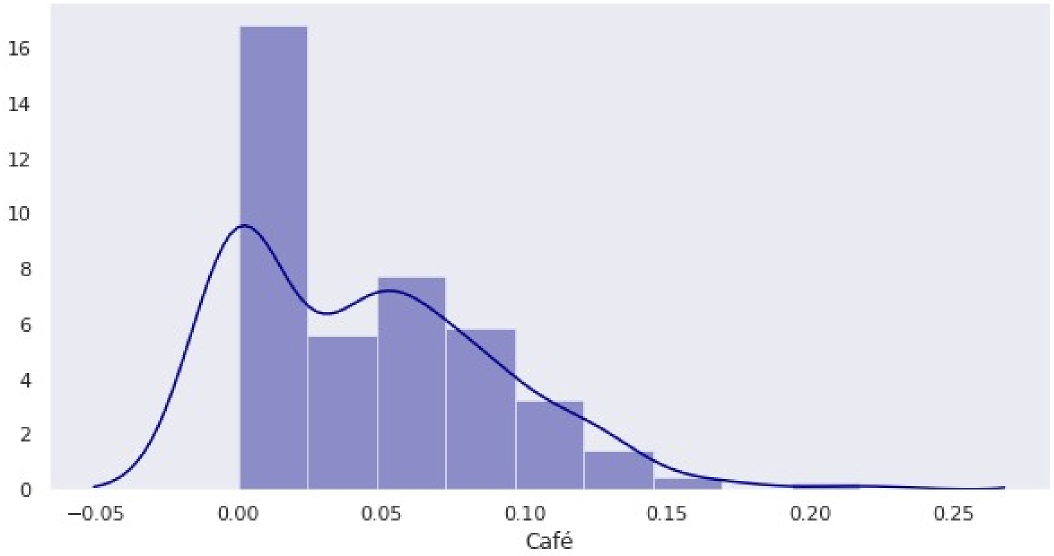
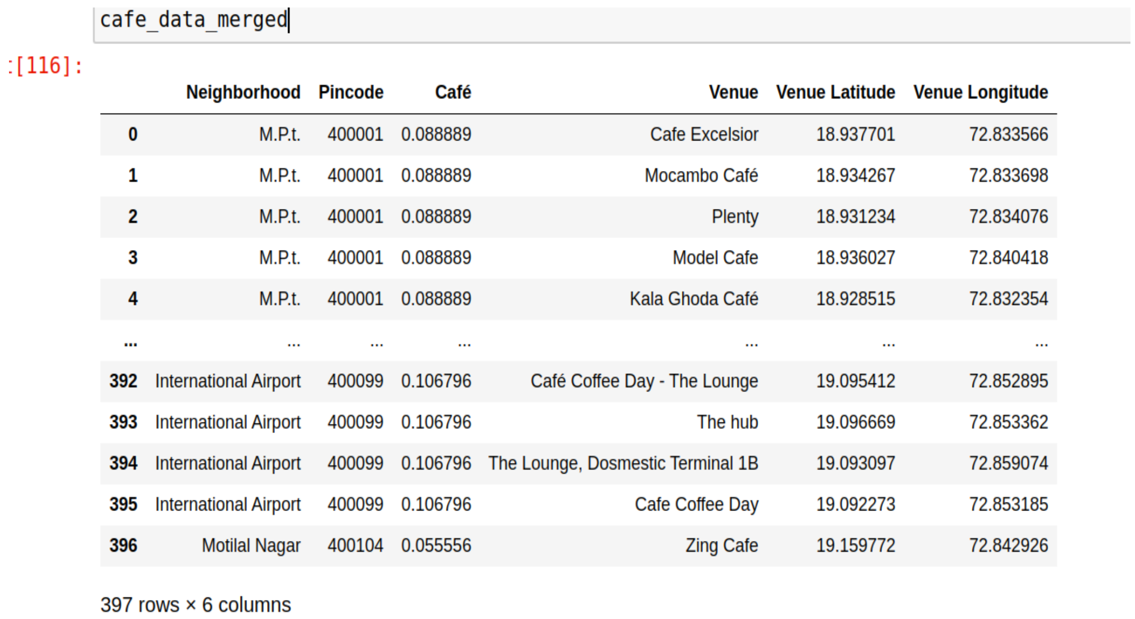


fig 4 . Distribution of freguency of Cafes

Most of neighborhood has frequency of between 0.00 to 0.05

After cleaning data and selecting only Neghborhoods where category is ‘cafe’ we get following dataset where cafe column represents frequency



```
cafe_data_merged
```

:[116]:

	Neighborhood	Pincode	Café	Venue	Venue Latitude	Venue Longitude
0	M.P.t.	400001	0.088889	Cafe Excelsior	18.937701	72.833566
1	M.P.t.	400001	0.088889	Mocambo Café	18.934267	72.833698
2	M.P.t.	400001	0.088889	Plenty	18.931234	72.834076
3	M.P.t.	400001	0.088889	Model Cafe	18.936027	72.840418
4	M.P.t.	400001	0.088889	Kala Ghoda Café	18.928515	72.832354
...	...	...	...	...	...	...
392	International Airport	400099	0.106796	Café Coffee Day - The Lounge	19.095412	72.852895
393	International Airport	400099	0.106796	The hub	19.096669	72.853362
394	International Airport	400099	0.106796	The Lounge, Dosmestic Terminal 1B	19.093097	72.859074
395	International Airport	400099	0.106796	Cafe Coffee Day	19.092273	72.853185
396	Motilal Nagar	400104	0.055556	Zing Cafe	19.159772	72.842926

397 rows × 6 columns

fig 5 . Final dataset

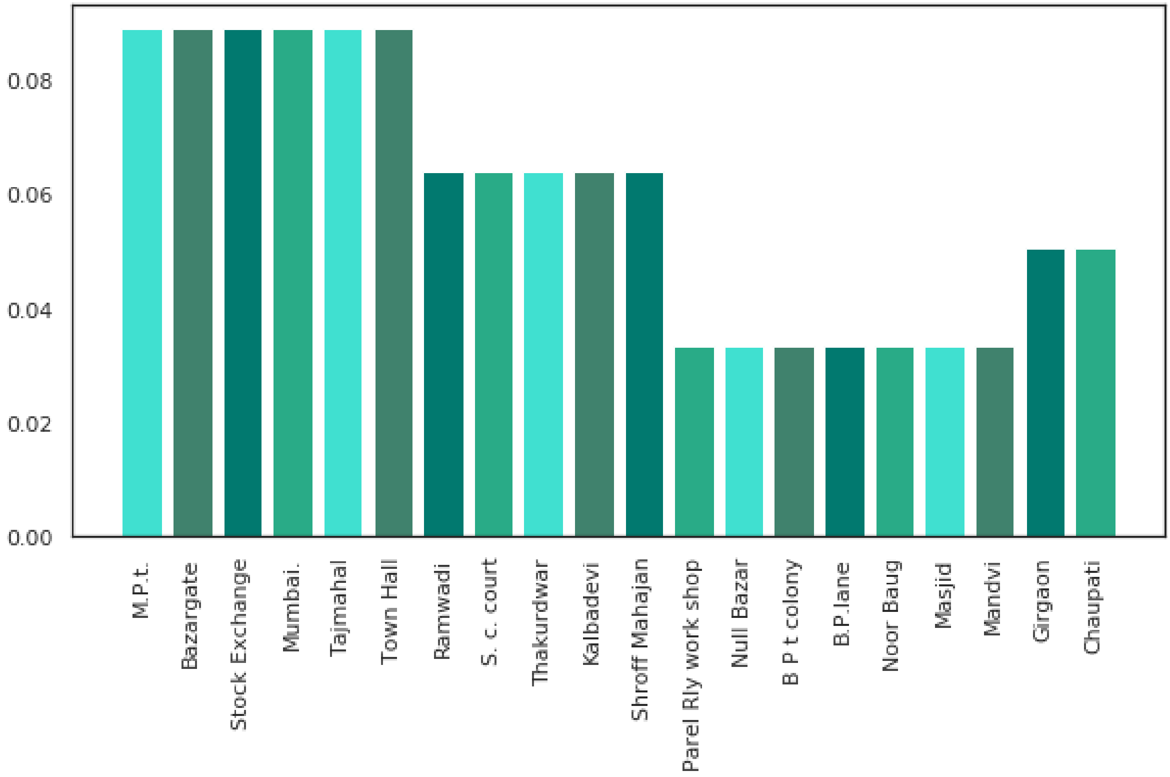


fig .6 Cafe Frequencies vs Neighborhoods

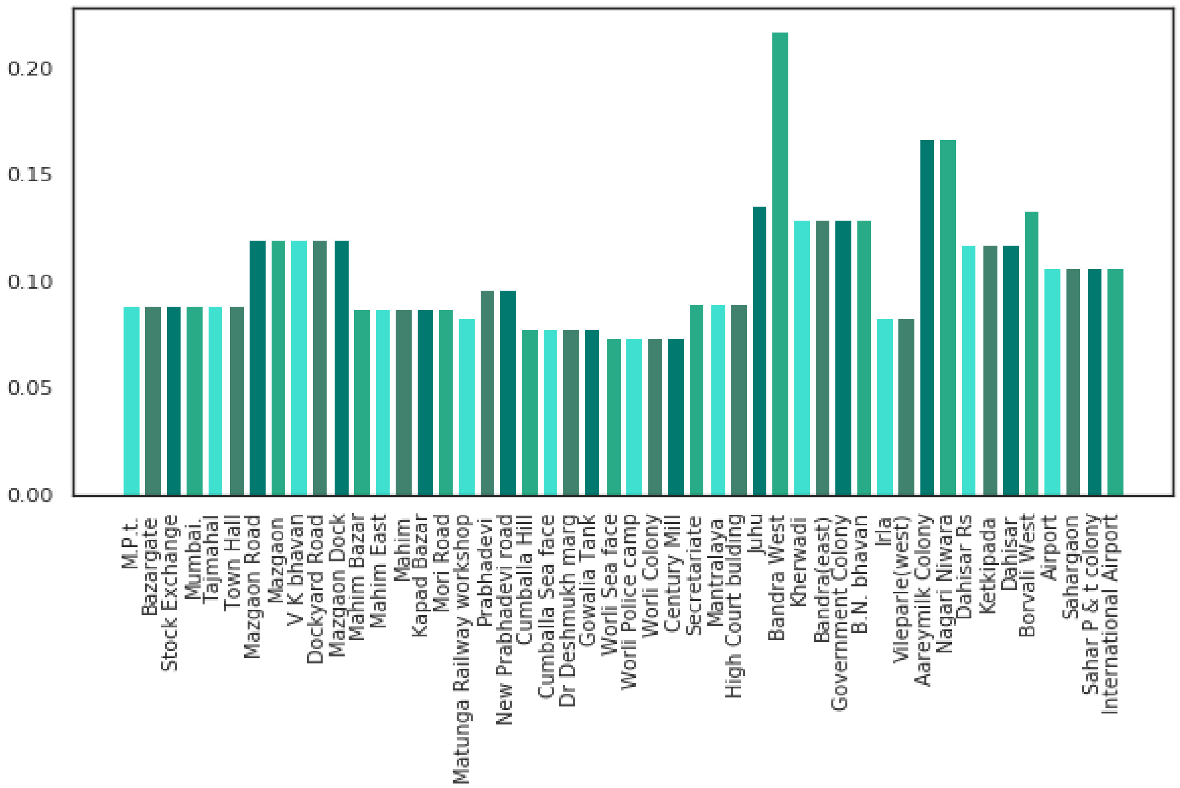


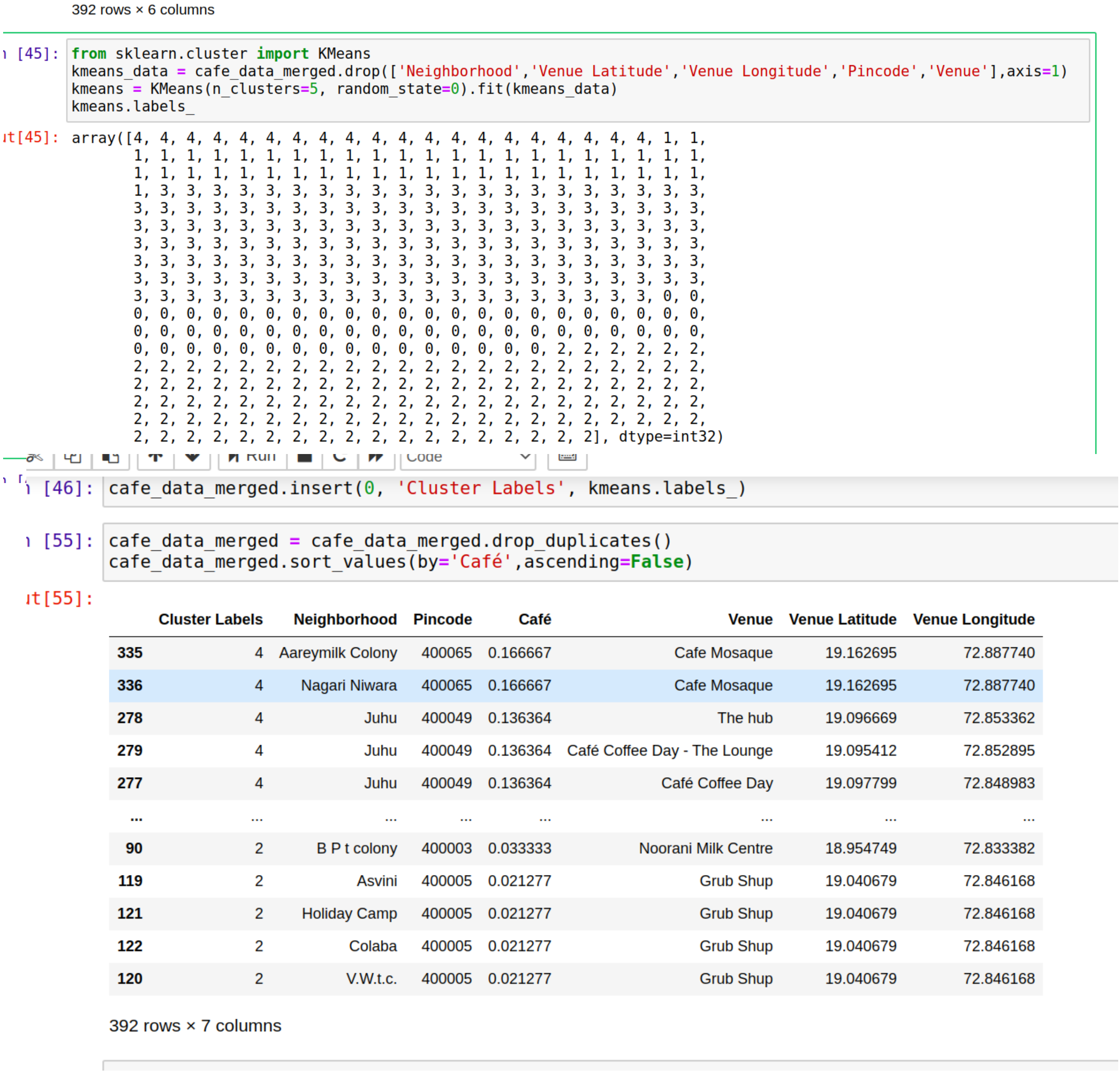
fig 7. Neighborhoods with frequency

from the above figure we can say that Bandra West has highest frequency i.e 0.2 that mean in this neighborhood number of cafes must be more than others

Application of Machine Learning

Kmeans Clustering:

kmeans algorithm is an unsurpurvised alogrithm which means we don’t need any labeled data. Kmeans group datapoints using euclidian distance .Using Kmeans we can cluster neighborhoods based on frequency of cafes or number of cafes present in neighborhoods.



392 rows × 6 columns

```
ı [45]: from sklearn.cluster import KMeans
        kmeans_data = cafe_data_merged.drop(['Neighborhood','Venue Latitude','Venue Longitude','Pincode','Venue'],axis=1)
        kmeans = KMeans(n_clusters=5, random_state=0).fit(kmeans_data)
        kmeans.labels_
```

```
ıt[45]: array([4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 1, 1,
               1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
               1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
               1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
               3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
               3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
               3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
               3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
               3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
               3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 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, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
               2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2], dtype=int32)
```

```
ı [46]: cafe_data_merged.insert(0, 'Cluster Labels', kmeans.labels_)
```

```
ı [55]: cafe_data_merged = cafe_data_merged.drop_duplicates()
        cafe_data_merged.sort_values(by='Café',ascending=False)
```

ıt[55]:

	Cluster Labels	Neighborhood	Pincode	Café	Venue	Venue Latitude	Venue Longitude
335	4	Aareymilk Colony	400065	0.166667	Cafe Mosaque	19.162695	72.887740
336	4	Nagari Niwara	400065	0.166667	Cafe Mosaque	19.162695	72.887740
278	4	Juhu	400049	0.136364	The hub	19.096669	72.853362
279	4	Juhu	400049	0.136364	Café Coffee Day - The Lounge	19.095412	72.852895
277	4	Juhu	400049	0.136364	Café Coffee Day	19.097799	72.848983
...	...	...	...	...	...	...	...
90	2	B P t colony	400003	0.033333	Noorani Milk Centre	18.954749	72.833382
119	2	Asvini	400005	0.021277	Grub Shup	19.040679	72.846168
121	2	Holiday Camp	400005	0.021277	Grub Shup	19.040679	72.846168
122	2	Colaba	400005	0.021277	Grub Shup	19.040679	72.846168
120	2	V.W.t.c.	400005	0.021277	Grub Shup	19.040679	72.846168

392 rows × 7 columns

fig 8. Adding Cluster labels

Results

After using kmeans in the dataset we get 5 clusters i.e 0,1,2,3 and 4 above dataset shows Neighbhorhood and corresponding cluster labels. For example
Neghborhood called Colaba belongs to cluster 3

Observations Noted

- We can see that Neighborhood Juhu is repeated as because there are more cafes and also the cafe called Grub shop is also repeated as it is in radius of more than one neighborhood that is why location is same . Which means that if you are in Juhu you may find many cafes around you and if you are in Grub shop you are near to these Neighborhoods

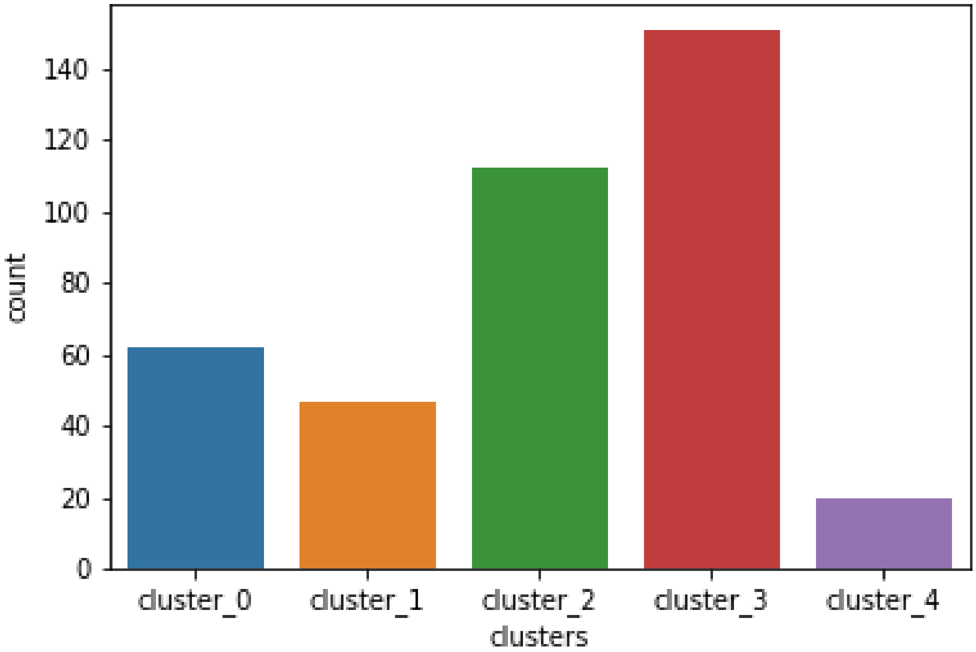
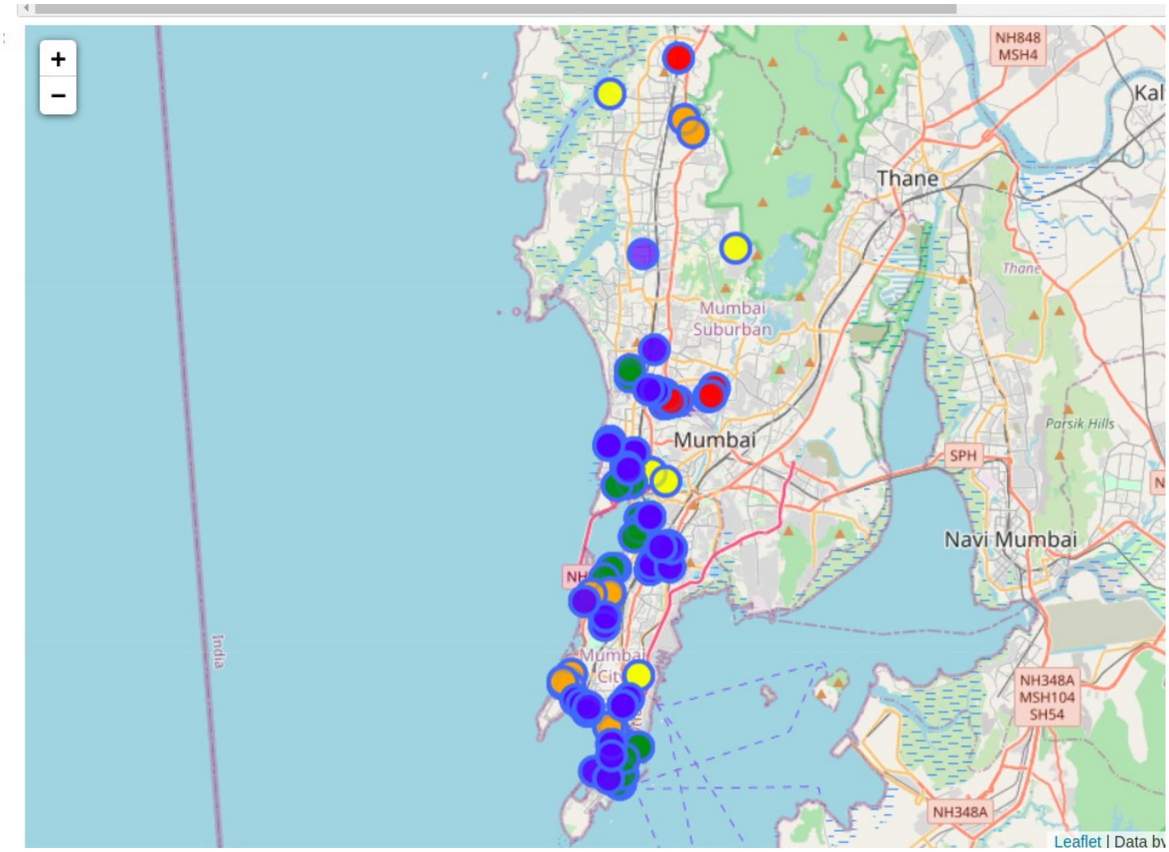


fig 9 . Cluster Count

from the above figure we can say that cluster 3 has more Neighborhood

Conclusion:

- cluster 1 has 0.1 we means that in cluster 1 no of Neighborhood are less but in that small area there are more cafes its concentrated in small area on the other hand Neighborhoods in cluster 3 are widely spread over the region or mumbai



- Green and Blue dots repressents cluster 3 and cluster 2 which is as said widley spread and has more neighborhoods
- yellow dot represents cluster 1 which is concentrated in small area but has high frequency
- so we can conclude that if you want to explore more cafes in less time you should choose neighborhoods belonging to cluster 2 and 3
- also if you want to open a new cafe consider cluster 4 and cluster 0 neighborhoods where compitation seems less