Capstone Project - The Battle of Neighborhoods

Report: IBM Capstone

Coursera IBM Professional Data Science Project

27 – 04 – 2020

Data Science Authored by:

Samir Mulani

(BE-comp.Sci)

Introduction:

A.1:Business Problem:

- **Finding best new location for coffee shop in Mumbai city.**

A.2:Description & Discussion of the Background:

The city of Mumbai consists of a large number of coffee shop, restaurants, but still there is always scope for new ones. The total area of Mumbai is 603.4 km2 (233 sq mi).Of this, the island city spans 67.79 km2 (26 sq mi).and Mumbai was the second most populous city in India after Delhi and the seventh most populous city in the world with a population of 19.98 million. As per Indian government population census of 2011. The most recent census was conducted in India during 2018, which put Mumbai's Urban Agglomeration at 20,748,395, while the city itself was recorded at 12,478,447.

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 person come to Mumbai daily for work.

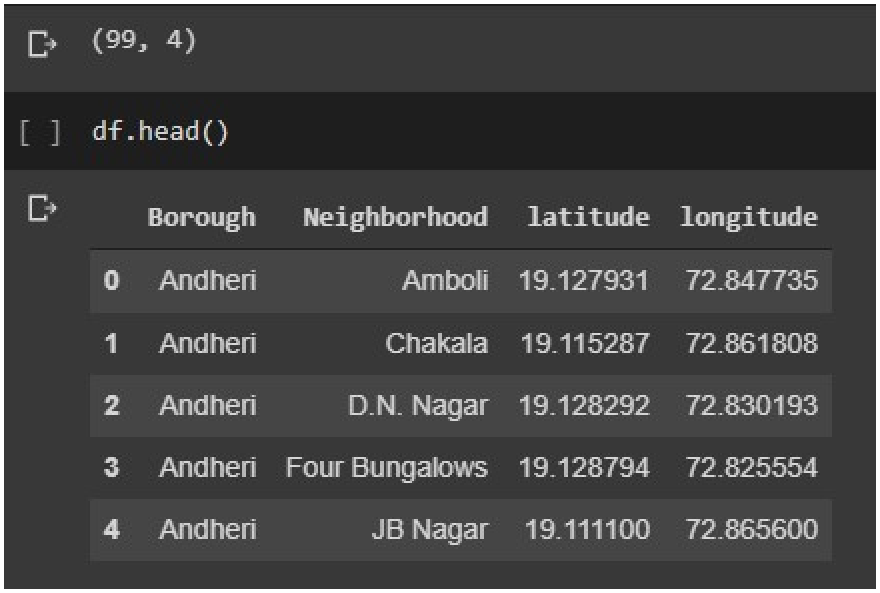
The Goal of this problem is to find a location that suits the below criteria:

1) A location that has many restaurants in the vicinity like (Indian, South Indian).

2) A location that has no or few café coffee, as this will ensure that there very little competition with other competitors.

A.3. Data Description:

The data that will be used in these projects is a csv file having data related to all neighborhoods in the city of Mumbai. File data collected from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai using this URL.



```
(99, 4)
```

```
[ ] df.head()
```

	Borough	Neighborhood	latitude	longitude
0	Andheri	Amboli	19.127931	72.847735
1	Andheri	Chakala	19.115287	72.861808
2	Andheri	D.N. Nagar	19.128292	72.830193
3	Andheri	Four Bungalows	19.128794	72.825554
4	Andheri	JB Nagar	19.111100	72.865600

Data that will be used to solve problem of finding best location using neighborhoods and location.

We explore the neighborhoods using Foursquare API to find the avenues within 500 meters of each neighborhood.

The Foursquare API that will be used to explore the neighborhoods is https://api.foursquare.com/v2/venues/explore. This API returns json response which will be transformed into a Data Frame, taking only the required details into consideration.

A.4: Target Audience

To recommend the correct location, XYZ Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the management which neighborhood of Mumbai city will be best choice to start a Coffee Shop. The Management also expects to understand the rationale of the recommendations made. This would interest anyone who wants to start a new café in Mumbai city.

A.5: Success Criteria:

The success criteria of the project will be a good recommendation of borough/Neighborhood choice to XYZ Company Ltd based on Lack of such café in that location and nearest sources of customer.

B: Methodology:

As version control and hosting of files Github was used. Below is the link to the repository.

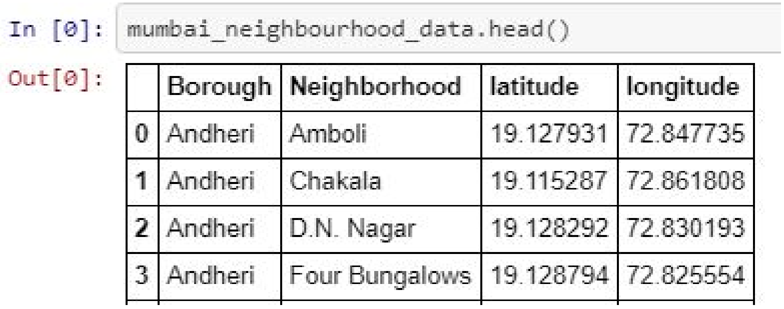
https://github.com/samir-cell/fantastic-datascience-Coursera_Capstone

1.Using the Foursquare API venue information is obtained in nearby vicinity locations in Mumbai.

Started with scrapping the List of neighborhoods in Mumbai . The data was wrangled to get it set up in a desired format and get the neighborhoods from following link:

https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai

2.Using geolocator API scrapping the list of latitude and longitude .

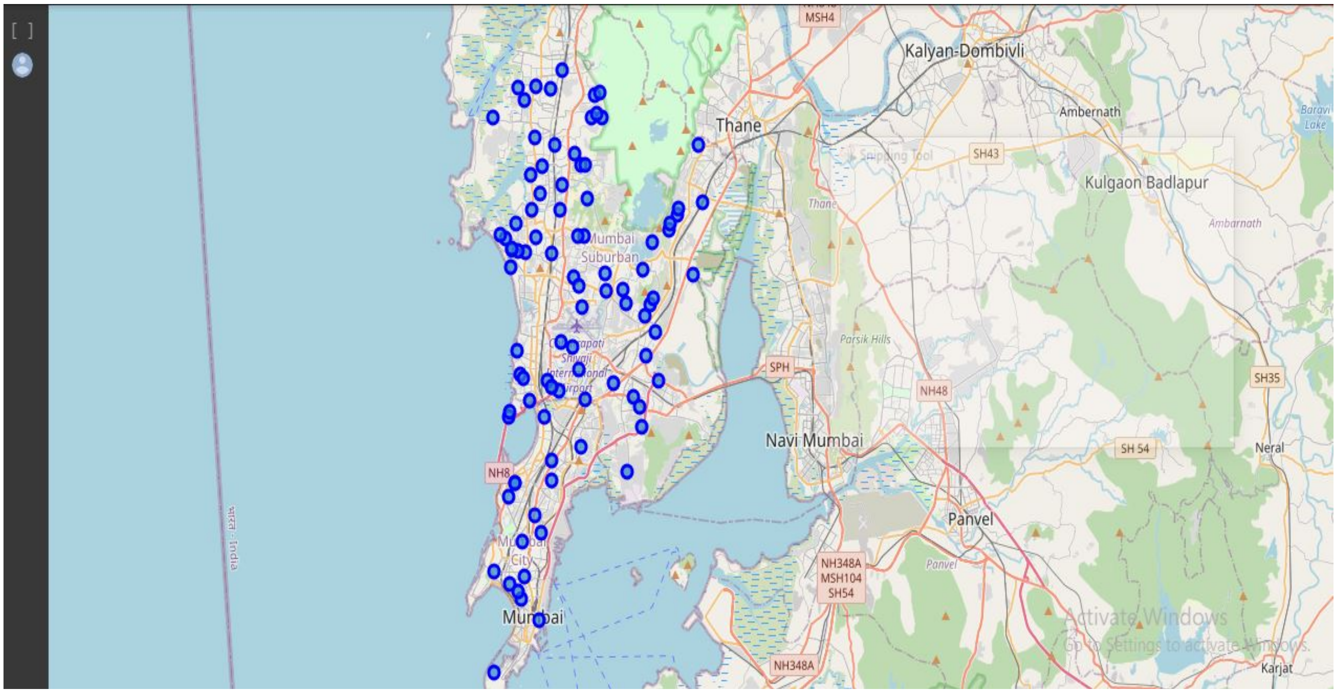


```
In [0]: mumbai_neighbourhood_data.head()
Out[0]:
```

	Borough	Neighborhood	latitude	longitude
0	Andheri	Amboli	19.127931	72.847735
1	Andheri	Chakala	19.115287	72.861808
2	Andheri	D.N. Nagar	19.128292	72.830193
3	Andheri	Four Bungalows	19.128794	72.825554

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.

Distribution of Sample Data

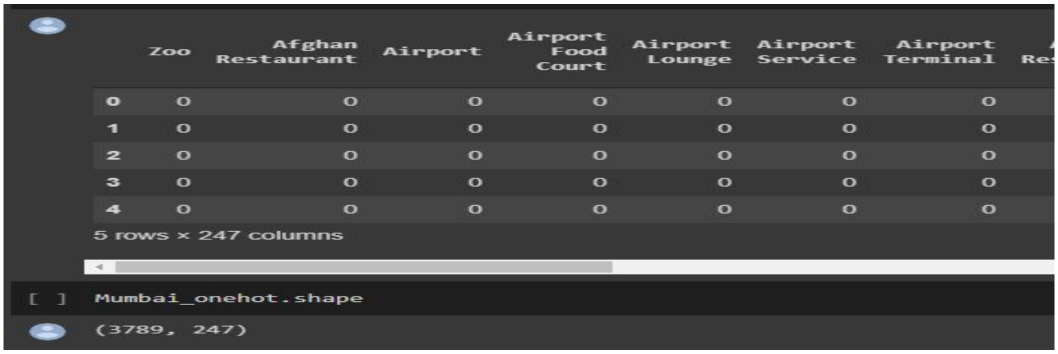


A radius is set to cover large neighborhoods in a particular area in Mumbai

3.Getting Nearby Venues using Four-Square API:

The total types of venues received from the Four-Square API was 3789. 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.
There are 247 unique categories.

4. 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.



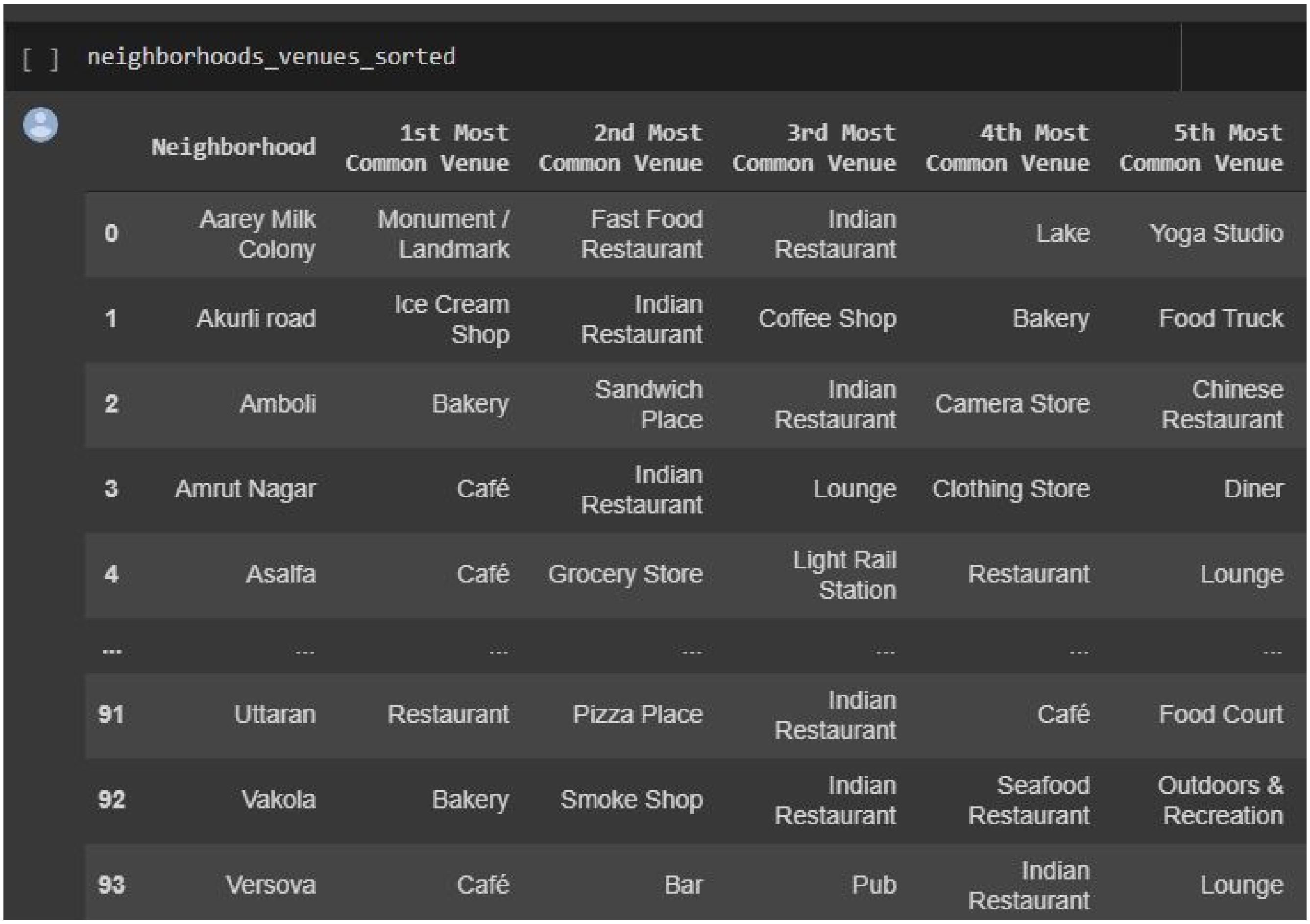
	Zoo	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal
0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0

5 rows × 247 columns

```
[ ] Mumbai_onehot.shape
```

```
(3789, 247)
```

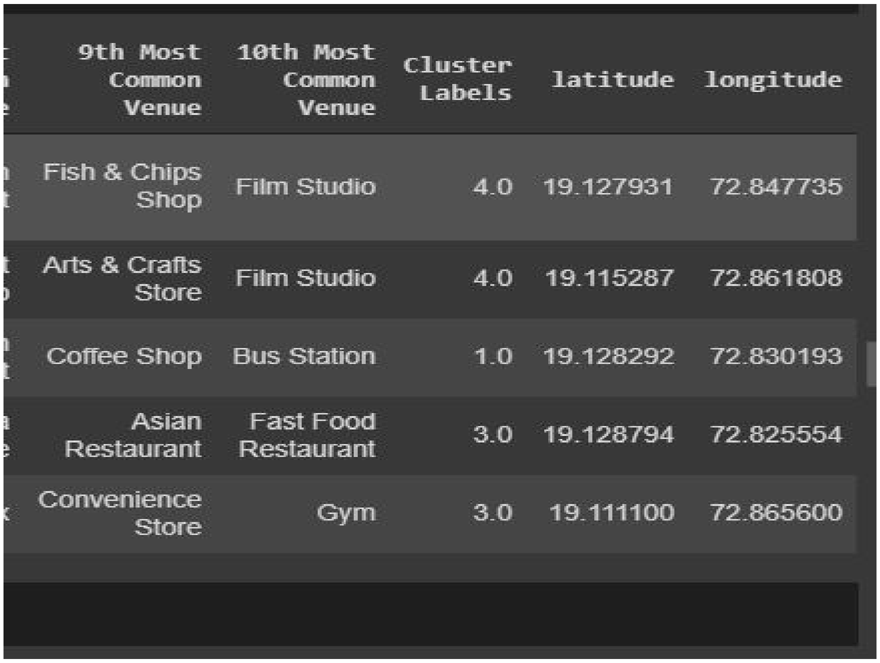
5. we get most common top 10 venues of all neighborhoods.



```
[ ]  neighborhoods_venues_sorted
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Aarey Milk Colony	Monument / Landmark	Fast Food Restaurant	Indian Restaurant	Lake	Yoga Studio
1	Akurli road	Ice Cream Shop	Indian Restaurant	Coffee Shop	Bakery	Food Truck
2	Amboli	Bakery	Sandwich Place	Indian Restaurant	Camera Store	Chinese Restaurant
3	Amrut Nagar	Café	Indian Restaurant	Lounge	Clothing Store	Diner
4	Asalfa	Café	Grocery Store	Light Rail Station	Restaurant	Lounge
...	...	...	...	...	...	...
91	Uttaran	Restaurant	Pizza Place	Indian Restaurant	Café	Food Court
92	Vakola	Bakery	Smoke Shop	Indian Restaurant	Seafood Restaurant	Outdoors & Recreation
93	Versova	Café	Bar	Pub	Indian Restaurant	Lounge

6. K-Means Algorithm was used to cluster the neighborhoods. The algorithm was iterated with cluster.



9th Most Common Venue	10th Most Common Venue	Cluster Labels	latitude	longitude
Fish & Chips Shop	Film Studio	4.0	19.127931	72.847735
Arts & Crafts Store	Film Studio	4.0	19.115287	72.861808
Coffee Shop	Bus Station	1.0	19.128292	72.830193
Asian Restaurant	Fast Food Restaurant	3.0	19.128794	72.825554
Convenience Store	Gym	3.0	19.111100	72.865600

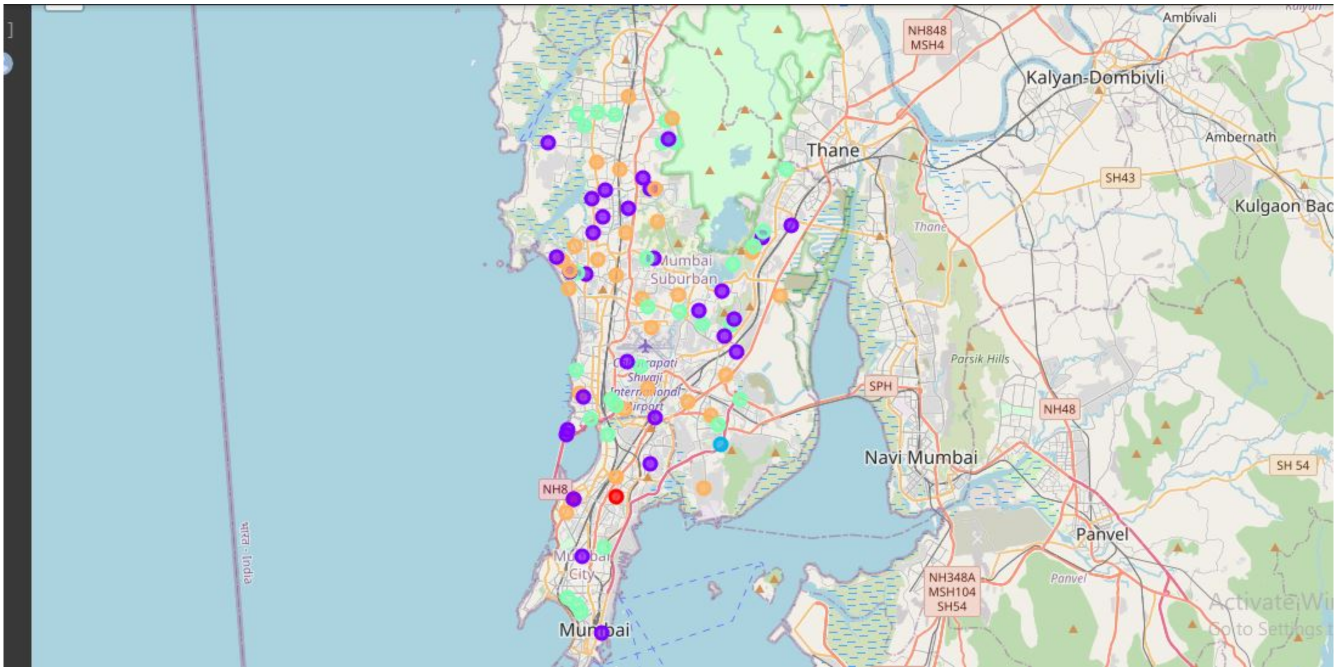
With appending longitude, latitude and cluster label we just created using k- means.

The **K-means clustering** algorithm is **used** to find groups which have not been explicitly labeled in the data. This can be **used** to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

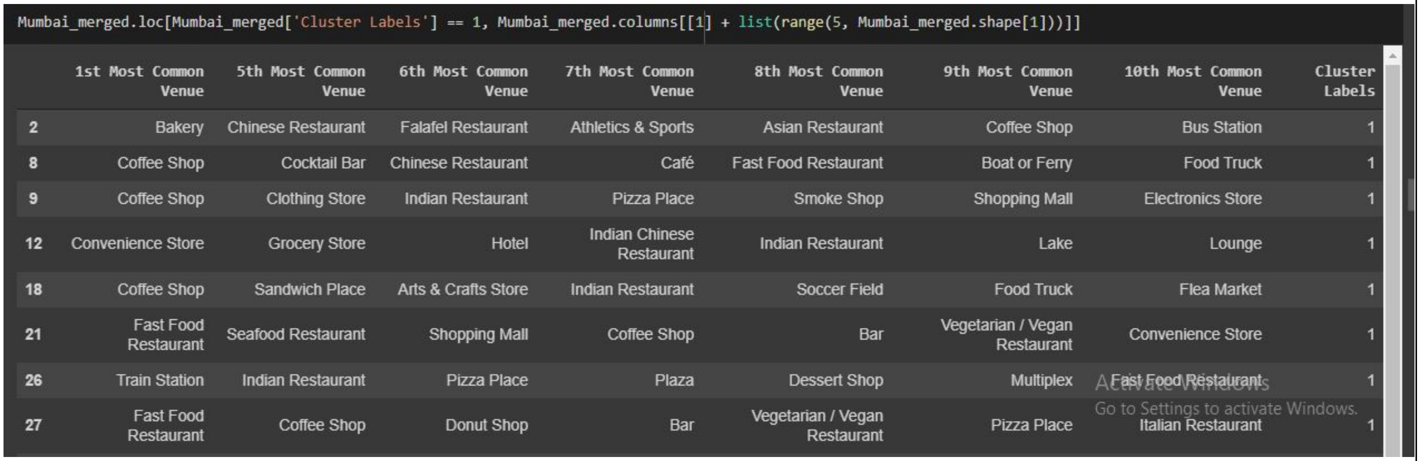
Other clustering algorithms with **better** features tend to be more expensive. In this case, **k-means** becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied. **K-means** is the simplest

C: DATA Analysis:

1.We Create Folium Map of Cluster we just define for better understanding of cluster location and intensity.



1. Analysis the Data of common venue from each cluster



```
Mumbai_merged.loc[Mumbai_merged['Cluster Labels'] == 1, Mumbai_merged.columns[[1] + list(range(5, Mumbai_merged.shape[1]))]]
```

	1st 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	Cluster Labels
2	Bakery	Chinese Restaurant	Falafel Restaurant	Athletics & Sports	Asian Restaurant	Coffee Shop	Bus Station	1
8	Coffee Shop	Cocktail Bar	Chinese Restaurant	Café	Fast Food Restaurant	Boat or Ferry	Food Truck	1
9	Coffee Shop	Clothing Store	Indian Restaurant	Pizza Place	Smoke Shop	Shopping Mall	Electronics Store	1
12	Convenience Store	Grocery Store	Hotel	Indian Chinese Restaurant	Indian Restaurant	Lake	Lounge	1
18	Coffee Shop	Sandwich Place	Arts & Crafts Store	Indian Restaurant	Soccer Field	Food Truck	Flea Market	1
21	Fast Food Restaurant	Seafood Restaurant	Shopping Mall	Coffee Shop	Bar	Vegetarian / Vegan Restaurant	Convenience Store	1
26	Train Station	Indian Restaurant	Pizza Place	Plaza	Dessert Shop	Multiplex	Fast Food Restaurant	1
27	Fast Food Restaurant	Coffee Shop	Donut Shop	Bar	Vegetarian / Vegan Restaurant	Pizza Place	Italian Restaurant	1

Resulting :

1.Indian Restaurant are present in each cluster (1st common venue)

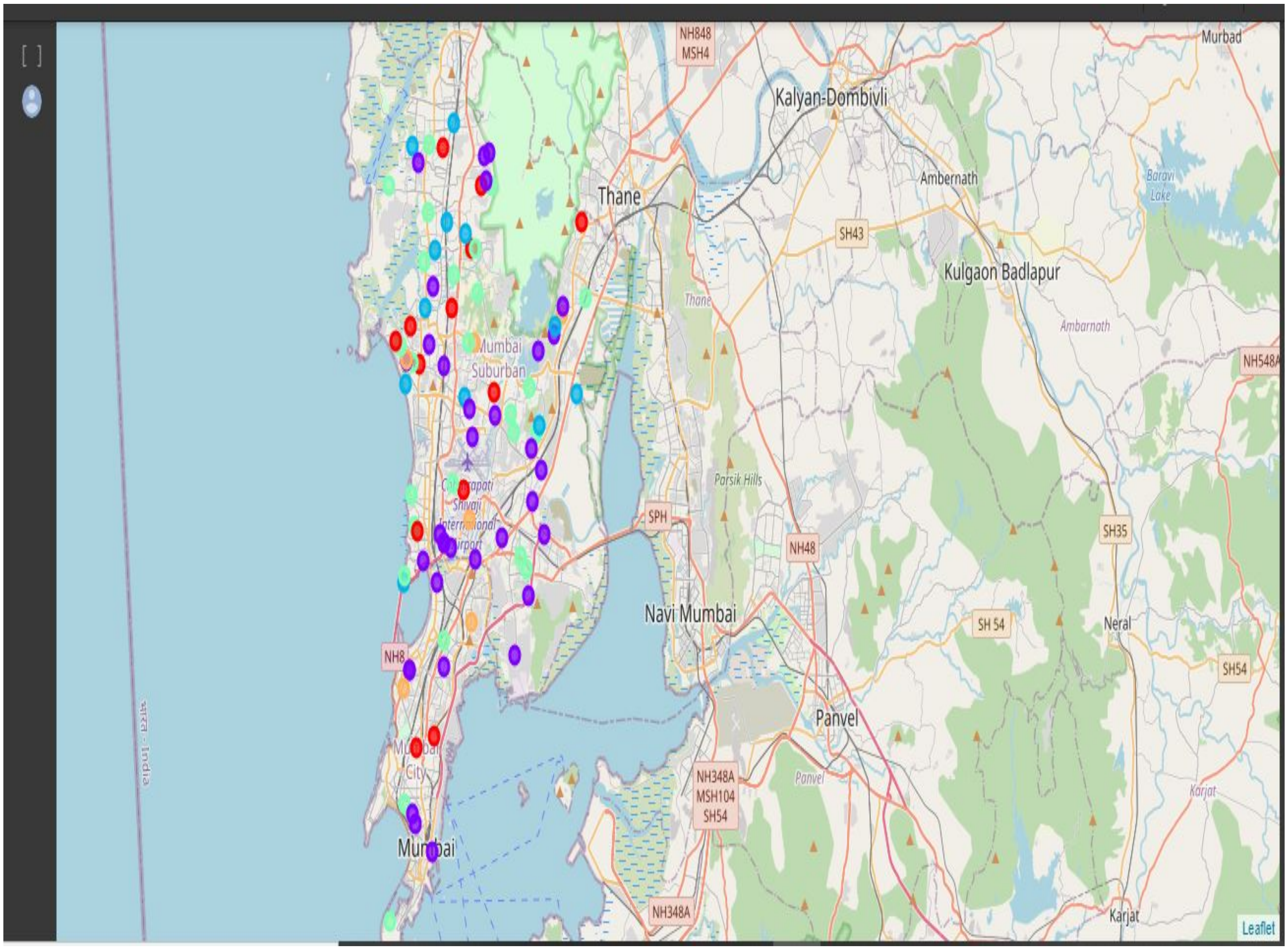
2.coffe shop is popular among Mumbai people. (1st common venue in some neighborhood)

we can locate coffee shop location for opening new coffee shop.

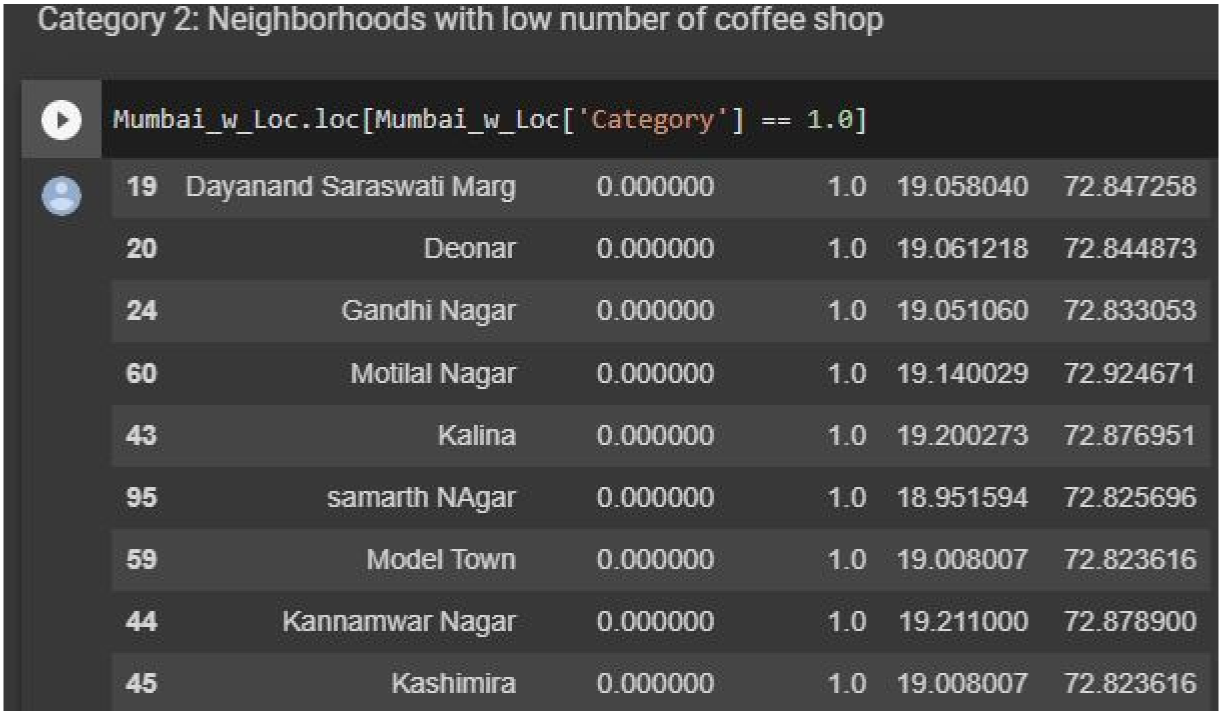
3.Lets Find out the Best location or Neighborhood for new coffee shop:

4.We cluster venue again and Get Mean for each Neighborhoods of coffee shop.

5.And Created Folium Map to analyze cluster just defined. Along with cluster label.



6. Cluster are sorted with its number and Neighborhoods with very low, low, medium high and very high number of Coffee Shop. Like this.



Category 2: Neighborhoods with low number of coffee shop

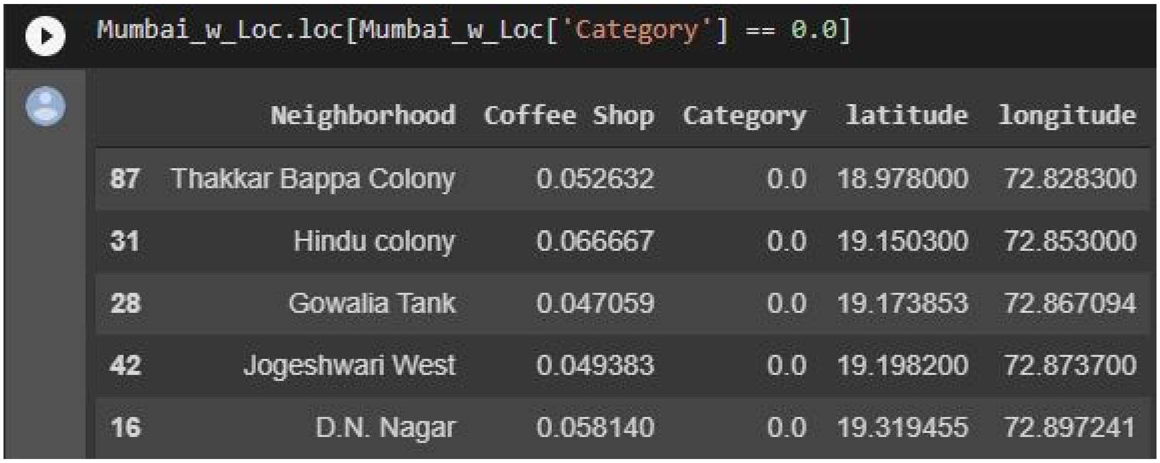
```
Mumbai_w_Loc.loc[Mumbai_w_Loc['Category'] == 1.0]
```

19	Dayanand Saraswati Marg	0.000000	1.0	19.058040	72.847258
20	Deonar	0.000000	1.0	19.061218	72.844873
24	Gandhi Nagar	0.000000	1.0	19.051060	72.833053
60	Motilal Nagar	0.000000	1.0	19.140029	72.924671
43	Kalina	0.000000	1.0	19.200273	72.876951
95	samarth NAgar	0.000000	1.0	18.951594	72.825696
59	Model Town	0.000000	1.0	19.008007	72.823616
44	Kannamwar Nagar	0.000000	1.0	19.211000	72.878900
45	Kashimira	0.000000	1.0	19.008007	72.823616

C. Discussion and Conclusion:

1: Results:

By exploring the requirements we found only two neighborhoods that match the requirements (Many restaurants in the vicinity & only a few coffee shops.) refer above fig.



```
Mumbai_w_Loc.loc[Mumbai_w_Loc['Category'] == 0.0]
```

	Neighborhood	Coffee Shop	Category	latitude	longitude
87	Thakkar Bappa Colony	0.052632	0.0	18.978000	72.828300
31	Hindu colony	0.066667	0.0	19.150300	72.853000
28	Gowalia Tank	0.047059	0.0	19.173853	72.867094
42	Jogeshwari West	0.049383	0.0	19.198200	72.873700
16	D.N. Nagar	0.058140	0.0	19.319455	72.897241

2: Discussion:

According to results we observe that most common venues (Top 10) come out to be restaurants and café, coffee shape and like Snack places, which means any new chef/business man can start a restaurant provided that they need to compete with existing restaurants (Other Types) but only if he provides top class facilities to get to top.

Result was more beneficiary having dataset of rating of shop and population of neighborhood.

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

3: Conclusion:

Based on the both Clusters formed it would be a good idea to open a new coffee shop in Clusters 0 or 1 & 2 since the other clusters already have coffee shop and Restaurants in their vicinities. cluster 0 and "NAN" value having no competition has possibility is there people are not interested at coffee shop.

cluster 1 & 2 have many restaurants in the vicinity (Pizza restaurants, European, Intercontinental, Indian) so one will be able to attract a good crowd.

(Note) These results have limitations - The venue data obtained is of top 10 venues in each neighbourhood, where we might neglect Coffee shop with less frequency or another venue name of same category.

References: Wikipedia pages :

1. https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai
2. https://en.wikipedia.org/wiki/ Mumbai