Capstone Project- The Battle of Neighbourhoods Report

1. Introduction

- This report covers the major aspects of a study done as a part of specialization certification for Data Science provided by IBM.
- Here, we have tried to apply all the knowledge acquired in the courses so far to enhance understanding of the subject matter as well as to gain some real life experience on working on a data science project.
- In this study we will focus on Mumbai city of India and different venues in it. We will then dive deeper to group venues in categories and segregate different areas of the city using data science methodologies.

1.1 Problem Definition

- Mumbai! A city that never sleeps yet dreams, every day. With over 2 and a half Crores of population, the city welcomes thousands of new people every day. In year 2020, what happened to the whole world, also happened to Mumbai. The hottest lines of this country, the Mumbai locals were stopped for first time since they started. A city that never sleeps became sleepless and everything came to halt for uncertain duration of time.
- We all know these stories and have cursed 2020 enough. Let's fall in place now, let's get back to our places. From 'chai tapri' to executive dine-ins, from art galleries to cricket grounds, we have covered everything in this study.
- As we are slowly getting back towards normal, we need to provide a detailed study of different venues and their categories as well as which corners of the city to find them in.

2. Data

- As we are building this project from a sketch, the most basic step is to collect the data regarding neighbourhoods of Mumbai and their respective pin codes.

- We have collected the basic data from here and copied it in to an excel file.

- We will use python's pandas library to read this data and store in to a data frame as shown in the figure below.

```
In [54]: df=pd.read_excel('Mumbai_Data.xlsx')
         df.head()
```

Out[54]:

	Neighborhood	PostalCode
0	August Kranti Marg	400036
1	Aarey Milk Colony	400065
2	Andheri (East)	400069
3	Andheri (West)	400058
4	Antop Hill	400037

Figure 1 Primary Data

- Now to add location data like Longitude and Latitude, we will use geocoder package of Python and prepare a data frame with features like: pin codes, neighbourhoods, longitude and latitude as shown in the figure below.

```
In [63]: df.head(10)
```

Out[63]:

	Neighborhood	PostalCode	Latitude	Longitude
0	August Kranti Marg	400036	18.964005	72.807983
1	Aarey Milk Colony	400065	19.161085	72.884394
2	Andheri (East)	400069	19.119298	72.851100
3	Andheri (West)	400058	19.122935	72.840610
4	Antop Hill	400037	19.020313	72.868280
5	Anu Shakti Nagar	400094	19.033945	72.925200
6	B A R C	400085	19.016345	72.926988
7	Ballard Estate	400038	18.940170	72.834830
8	Bandra (East)	400051	19.060715	72.854564
9	Bandra (West)	400050	19.052259	72.829405

Figure 2 Data with Longitude and Latitude

2.1 Four Square API

- Four Square API is a power full tool to retrieve location data for any location. This includes various venues by their categories, their exact location, user reviews, recommendations and other venue specific details.

- We will use this opportunity and get list of venues with their location and category for each of the neighbourhoods.
- We will then apply exploratory data analysis techniques to ultimately prepare a data frame with each neighbourhood and their top 10 most frequent venues category wise.
- This data frame will be used for carrying out clustering and analysis of different clusters.

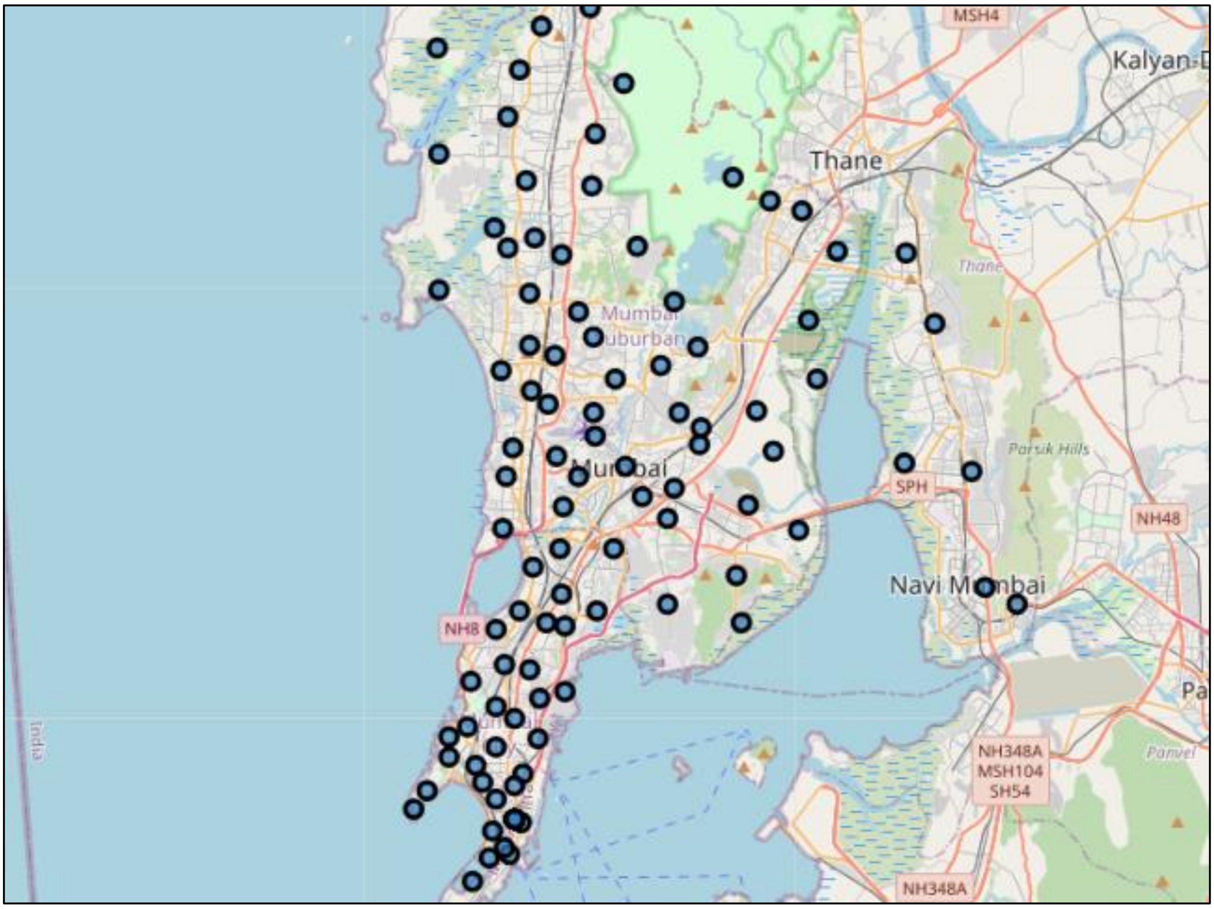

Figure 3 Map of Mumbai

3. Methodology

- The backbone of modern day Data Science is rich with legacy statistical algorithms.
- As we are aiming to segregate and cluster different areas of Mumbai according to their similarities in terms of categories of venues around them, we will use K-means clustering algorithm.
- This algorithm works in a following way:
    - First it selects random K centroids in n-dimensional space where K is an arbitrary number and n is number of features.
    - It then segregates data points and labels them according to the nearest arbitrary centroid.
    - When the labelling is completed for all data points, the centroids are moved to mean of the data points in their respective clusters.

    - The same process is repeated with these newly obtained centroids and stopped when there is no change possible further.
- The core idea behind K-means clustering is, the resultant clusters should have maximum inter-cluster distance and minimum distance among the points within same cluster.
- The term ‘Within Clusters Sum of Squares’ (WCSS) is to depict the sum of sum of squares of distances between data points and their respective centroids.
- As we increase value of K, the WCSS keeps decreasing but after one point, the drop in value of WCSS is not so considerable.
- The optimum value of K should be selected as the value after which increment in K doesn’t effect WCSS much.
- If we plot a graph of values of WCSS against values of K, it shows a sharp decline initially and then the value of WCSS doesn’t decrease much. Therefore it creates an Elbow type graph and hence this method of selecting optimum K is called ‘Elbow method’.
- This is a trial and error method and differs from cases to cases.

4. Results

- We have retrieved venues with 238 unique categories from Four Square API.
- From these categories, we selected 2 higher level categories such as:
    - Outing and recreational places
    - Food venues
- We then selected appropriate categories that fall under these two categories and filtered our venues data frame accordingly.
- We then divided Mumbai city in to 10 clusters for both major categories and obtained following results using folium maps.
- The following figure shows folium map for ‘Outing and recreational places’ category.

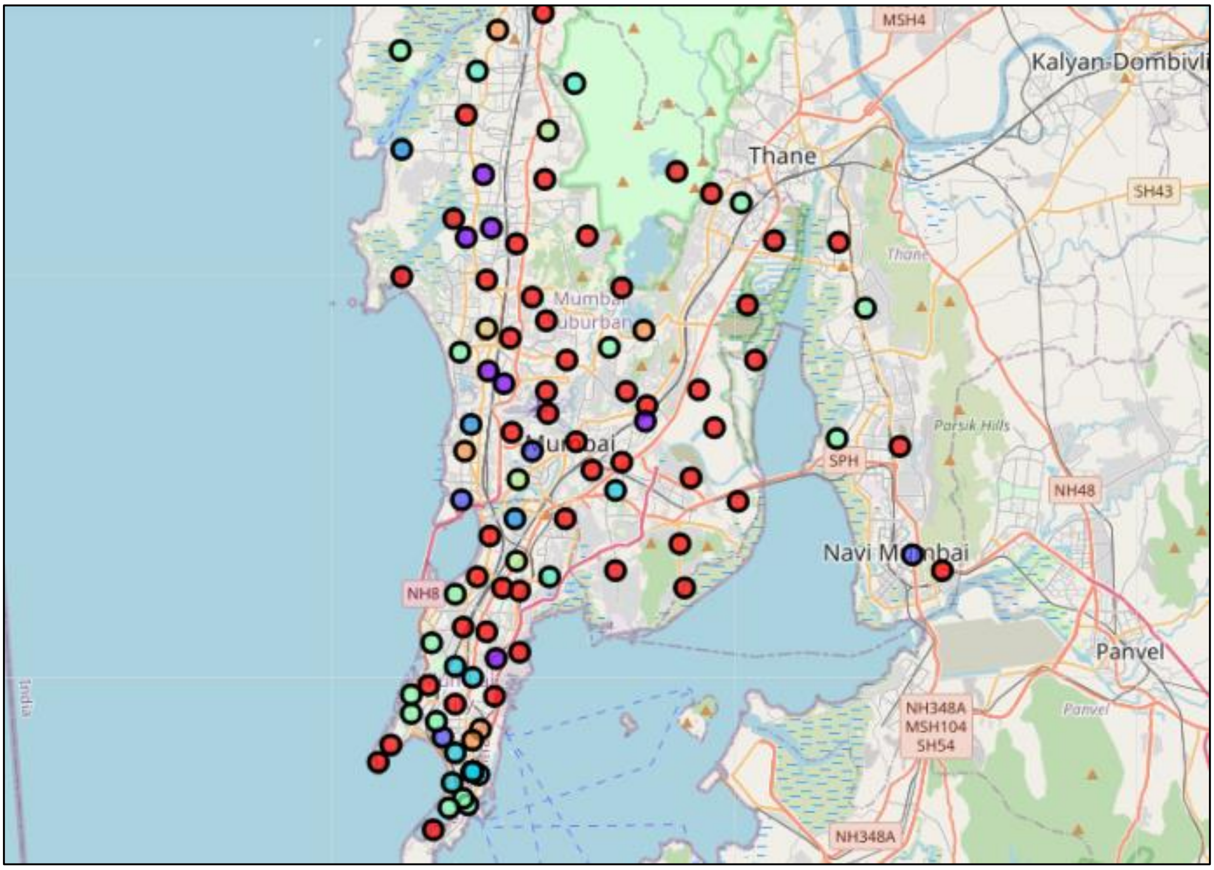

Figure 4 Clustering map for 'Outing and recreational places' category

- The following figure shows folium map for 'Food venues' category.

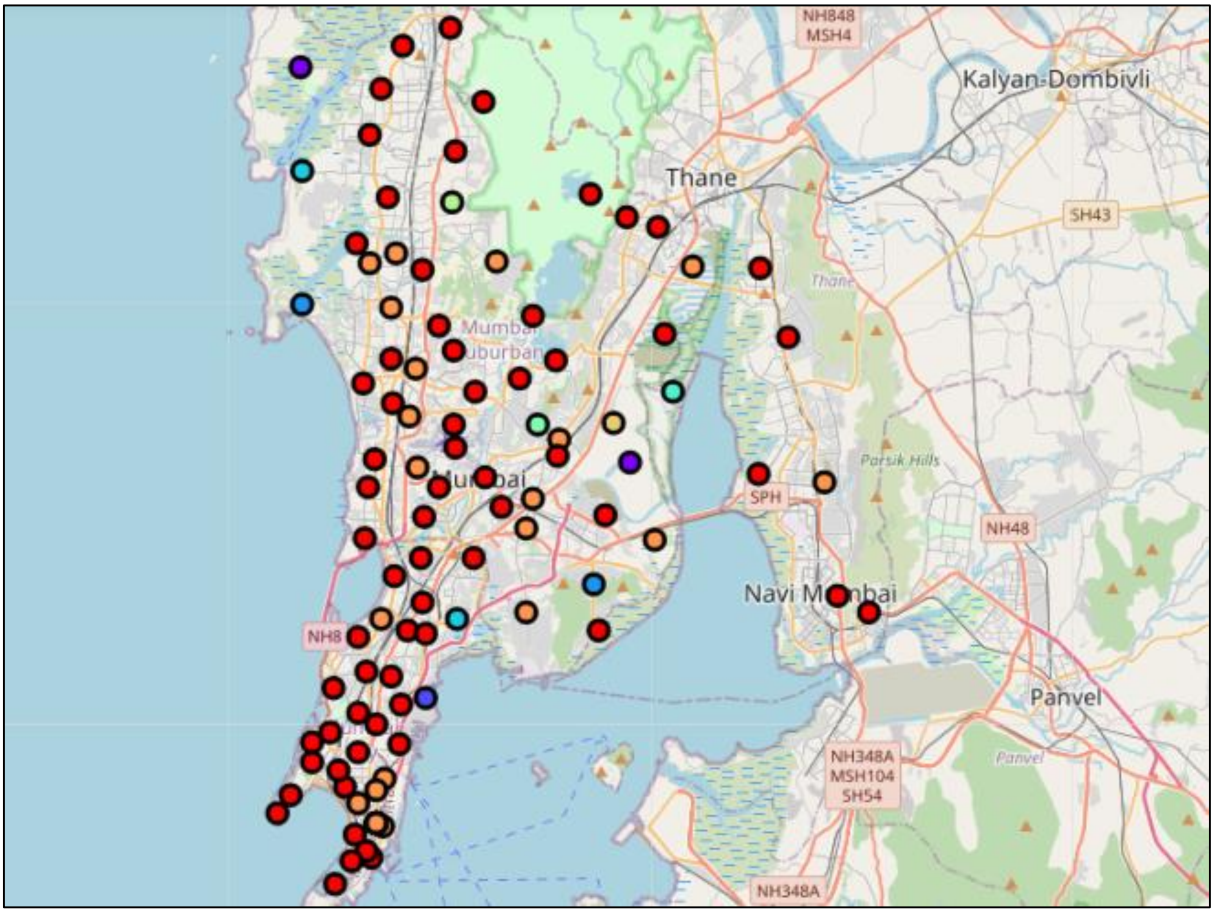

Figure 5 Clustering map for 'Food venues' category

5. Discussion

- In both the cases, what we have observed is, the categories with highest number of venues across the city have major impact on clustering.
- These categories cause imbalanced clustering and we could observe over presence of venues of these categories in clusters with higher number of neighbourhoods.

a) ‘Outing and recreational places’ category:

  - In following figures, we can see that 'Multiplex', 'Cricket Ground' and 'Movie Theater' has the highest frequency and as it can be seen they are everywhere in Cluster 4 and 6 which are Clusters with maximum neighbourhoods.

```
In [215]: mumbai_fun_places_merged['Cluster-Labels'].value_counts()

Out[215]: 11.0    41
          4.0     16
          6.0     14
          0.0      9
          1.0      7
          9.0      5
          2.0      4
          3.0      3
          7.0      3
          5.0      3
          8.0      1
          Name: Cluster-Labels, dtype: int64
```

Figure 6 Clusters with neighbourhood counts - Case 1

```
In [217]: mumbai_fun_places['Venue Category'].value_counts()

Out[217]: Multiplex              46
          Cricket Ground         29
          Movie Theater          22
          History Museum         19
          Indie Movie Theater    17
          Theater                17
          Hockey Arena           11
```

Figure 7 Category frequency - Case 1

```
In [219]: mumbai_fun_places_merged.loc[mumbai_fun_places_merged['Cluster-Labels'] == 4]

Out[219]:
```

	Neighborhood	PostalCode	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
7	Ballard Estate	400038	18.940170	72.834830	4.0	Cricket Ground	Multiplex	Indie Movie Theater	History Museum	Hockey Arena	Zoo	Arcade	Art Gallery
16	Mumbai G P O	400001	18.939031	72.837345	4.0	History Museum	Multiplex	Zoo	Indie Movie Theater	Arcade	Art Gallery	Bowling Alley	Comedy Club
21	Chembur	400071	19.056035	72.897040	4.0	General Entertainment	Performing Arts Venue	Multiplex	Zoo	Indie Movie Theater	Arcade	Art Gallery	Bowling Alley
24	Council Hall	400039	18.940170	72.834830	4.0	Cricket Ground	Multiplex	Indie Movie Theater	History Museum	Hockey Arena	Zoo	Arcade	Art Gallery
39	Jacob Circle	400011	18.983709	72.826845	4.0	Racetrack	History Museum	Multiplex	Zoo	Indie Movie Theater	Arcade	Art Gallery	Bowling Alley

Figure 8 Cluster 4.0 - Case 1

```
In [218]: mumbai_fun_places_merged.loc[mumbai_fun_places_merged['Cluster-Labels'] == 6]
```

Out[218]:

	Neighborhood	PostalCode	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 M Comm Ver
0	August Kranti Marg	400036	18.964005	72.807983	6.0	Theater	Stadium	Concert Hall	History Museum	Zoo	Indie Movie Theater	Arcade	Art Gallery	Bowl A
19	Borivli HO	400091	19.237596	72.803000	6.0	Theme Park Ride / Attraction	Theme Park	Water Park	Zoo	Hockey Arena	Arcade	Art Gallery	Bowling Alley	Com C
25	Cumballa Hill	400026	18.971712	72.807475	6.0	Theater	Stadium	Zoo	Indie Movie Theater	Arcade	Art Gallery	Bowling Alley	Comedy Club	Con I
35	Grant Road	400007	18.960945	72.818459	6.0	Indie Movie Theater	Theater	Opera House	History Museum	Zoo	Arcade	Art Gallery	Bowling Alley	Com C
36	Hutatma	400023	18.926536	72.832575	6.0	History	Movie	Art	Stadium	Indie Movie	Arcade	Bowling	Comedy	Con

Figure 9 Cluster 6.0 - Case 1

b) 'Food venues' category:

- In following figures, we can see that 'Indian Restaurant' and 'Cafe' has the highest frequency and as it can be seen they are everywhere in Cluster 0 and 9 which are Clusters with maximum neighbourhoods.

```
In [237]: mumbai_eats_merged['Cluster-Labels'].value_counts()

Out[237]: 0.0     61
          9.0     29
          11.0     5
          1.0      2
          3.0      2
          4.0      2
          8.0      1
          5.0      1
          2.0      1
          7.0      1
          6.0      1
          Name: Cluster-Labels, dtype: int64
```

Figure 10 Clusters with neighbourhood counts - Case 2

```
In [239]: mumbai_eats['Venue Category'].value_counts()

Out[239]: Indian Restaurant                 573
          Café                              232
          Coffee Shop                       164
          Fast Food Restaurant              158
          Chinese Restaurant                129
```

Figure 11 Category frequency - Case 2

```
In [240]: mumbai_eats_merged.loc[mumbai_eats_merged['Cluster-Labels'] == 0]
```

Out[240]:

	Neighborhood	PostalCode	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	8tl Co
0	August Kranti Marg	400036	18.964005	72.807983	0.0	Indian Restaurant	Bakery	Café	Coffee Shop	Sandwich Place	Pizza Place	Dessert Shop	Res
3	Andheri (West)	400058	19.122935	72.840610	0.0	Indian Restaurant	Bar	Vegetarian / Vegan Restaurant	Coffee Shop	Fast Food Restaurant	Pizza Place	Restaurant	C Res
6	B A R C	400085	19.016345	72.926988	0.0	Ice Cream Shop	Vegetarian / Vegan Restaurant	Food	Dessert Shop	Dhaba	Dim Sum Restaurant	Diner	
8	Bandra (East)	400051	19.060715	72.854564	0.0	Indian Restaurant	Pizza Place	Restaurant	Café	Italian Restaurant	Bar	Diner	Fas Res
9	Bandra (West)	400050	19.052259	72.829405	0.0	Indian Restaurant	Café	Coffee Shop	Chinese Restaurant	Bakery	Bar	Pizza Place	
15	Mumbai Central	400008	18.967725	72.827071	0.0	Indian Restaurant	Fast Food Restaurant	Ice Cream Shop	Dessert Shop	Restaurant	Middle Eastern	Chinese Restaurant	

Figure 12 Cluster 0 - Case 2

```
In [241]: mumbai_eats_merged.loc[mumbai_eats_merged['Cluster-Labels'] == 9]
```

Out[241]:

	Neighborhood	PostalCode	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
1	Aarey Milk Colony	400065	19.161085	72.884394	9.0	Hotel	Café	Indian Restaurant	Restaurant	Vegetarian / Vegan Restaurant	Food	Dhaba	Dim Sum Restaurant
2	Andheri (East)	400069	19.119298	72.851100	9.0	Indian Restaurant	Chinese Restaurant	Fast Food Restaurant	Pizza Place	Hotel	Restaurant	Sandwich Place	Seafood Restaurant
7	Ballard Estate	400038	18.940170	72.834830	9.0	Indian Restaurant	Café	Bakery	Bar	Seafood Restaurant	Coffee Shop	Chinese Restaurant	Hotel
14	Bhavani Shankar Road	400028	19.020358	72.836280	9.0	Indian Restaurant	Chinese Restaurant	Ice Cream Shop	Fast Food Restaurant	Bar	Café	Coffee Shop	Breakfast Spot
16	Mumbai G P O	400001	18.939031	72.837345	9.0	Indian Restaurant	Café	Seafood Restaurant	Hotel	Coffee Shop	Chinese Restaurant	Bar	Irani Cafe
21	Chembur	400071	19.056035	72.897040	9.0	Indian Restaurant	Café	Pizza Place	Seafood Restaurant	Fast Food Restaurant	Bar	Diner	Chinese Restaurant

Figure 13 Cluster 9.0 - Case 2

- An important point to note here is, in both cases we have a cluster named '11.0' for those neighbourhoods with zero venues that fall under selected categories. Ironic that a city like Mumbai has 5 neighbourhoods with zero food places and 41 neighbourhoods with zero fun places!! This might be a result of inefficient data registration or limited access for available licence in Four Square API.

6. Conclusion

- This study has provided credible and considerable amount of hands-on experience for data mining, data analysis, data visualization and machine learning algorithm over a real life problem.
- We can further optimize the clustering by adjusting K value using Elbow method and see how we obtain different results.
- We can carry out same analysis for more such categories like healthcare, banking etc. and see which area is all over the best in Mumbai with everything available 5 steps away.