A TALE OF TWO CITIES:

Looking at Mumbai through the New York lens

Introduction/Business Problem

New York is considered one of the most influential cities on our planet. Being the financial capital of the United States of America, New York carries a certain weight. It is also the home to some of the wealthiest and most prominent people in our world. The success of New York can also be attributed to its urban planning, especially the Commissioner's plan of 1811, that laid the roadmap to its success.

New York consists of five boroughs, namely Bronx, Brooklyn, Queens, Staten Island, and Manhattan. Among the five boroughs, Manhattan is considered the epicenter of the city's growth. Manhattan was the earliest site of development, and the other boroughs were later incorporated in the city. Its iconic museums, skyscrapers, monuments, and parks are visited by millions of people every year. Manhattan is also the central business district of the city, and its iconic skyline, as seen from the Hudson River, gives it unique fame among all cities. Among all the influential cities in our world, such as New York, London, Paris, Tokyo, and Sydney, it is one of the few that carries a certain weight despite not being the seat of the political power of its country.

Fig 1: New York City skyline

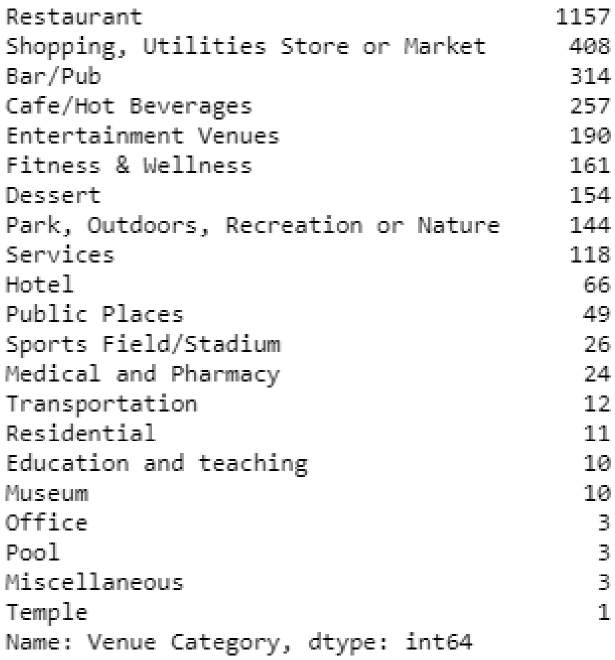
Similar to New York, Mumbai city is considered the financial capital of India and is the most influential city in the country. It is also the home to the wealthiest and most influential people of India. New York and Mumbai have many similarities and yet are different in so many aspects. Despite being the wealthiest city of India, Mumbai's development was not planned, and that has restricted its growth despite its vast potential. It has become significantly overcrowded and expensive. Many people believe that it has become incapable of providing quality living standards to its inhabitants. Mumbai's local transportation, which is considered the lifeline of the city, is running above its capacity. The development of new modes of transportation, such as the Mumbai Metro project, is slow and often faces hurdles in expansion. Mumbai is also the home to the largest slum on the planet.

Fig 2: Mumbai City skyline

The idea behind this project is to consider the Manhattan area of New York as the standard model for a prosperous city. A dataset comprising of all neighborhoods in Manhattan will be used for modeling. The data for all venues in each neighborhood will be accessed through Foursquare location data. A final dataset comprising of each neighborhood and the most common venue categories in them will be prepared. K-means clustering algorithm will be used to segment the neighborhoods in five clusters based one the most common venue categories in each neighborhood. This clustered data will be used to create a model for an ideal city. This model will then be used to map Mumbai city's neighborhoods in accordance with Manhattan's neighborhoods. The project will help characterize different neighborhoods in Mumbai city based on their similarities or differences with Manhattan's neighborhoods. The results can be leveraged by the city administration in planning future projects as well as by investors in their decision making.

Data and Methodology

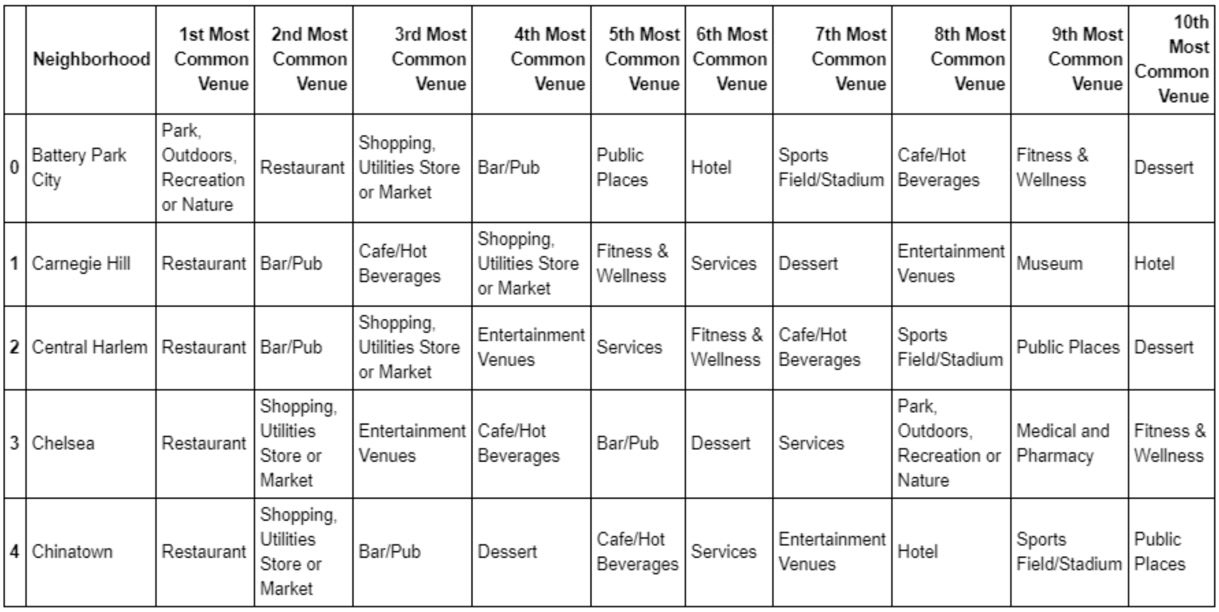
The first part of the project deals with accessing New York City's neighborhoods' data and segmenting them in clusters. The JSON file comprising of New York's data was provided in the Practice Lab in Week 3. For ease of handling data, only neighborhoods falling in the Manhattan borough was considered. Geopy package of python was used to get the coordinates for each neighborhood. The Foursquare location data was used to get the list of venues, venue category, and coordinates for each neighborhood. The list venues were categorized in over 330 categories in the original data. Although for the ease of handling, the venues were re-categorized in 21 broad categories (Table 1). This was done by exporting the CSV file containing information related to venues in each neighborhood and processing in SQL. The edited dataset was then imported back as a pandas data frame and grouped by the frequency of venue categories in each neighborhood.



```
Restaurant                              1157
Shopping, Utilities Store or Market      408
Bar/Pub                                  314
Cafe/Hot Beverages                       257
Entertainment Venues                     190
Fitness & Wellness                       161
Dessert                                  154
Park, Outdoors, Recreation or Nature     144
Services                                 118
Hotel                                     66
Public Places                             49
Sports Field/Stadium                      26
Medical and Pharmacy                      24
Transportation                            12
Residential                               11
Education and teaching                    10
Museum                                    10
Office                                     3
Pool                                       3
Miscellaneous                              3
Temple                                     1
Name: Venue Category, dtype: int64
```

Table 1: Count of venues in different venue categories in Manhattan.

The venue categories were assigned labels using Label Encoder and the label mapping schema was saved for further processing. The dataset was then processed to get the frequency of appearance for all venue categories in each neighborhood. A final dataset comprising of each neighborhood and its top ten most common venue categories were compiled by defining a function to extract such a dataset.

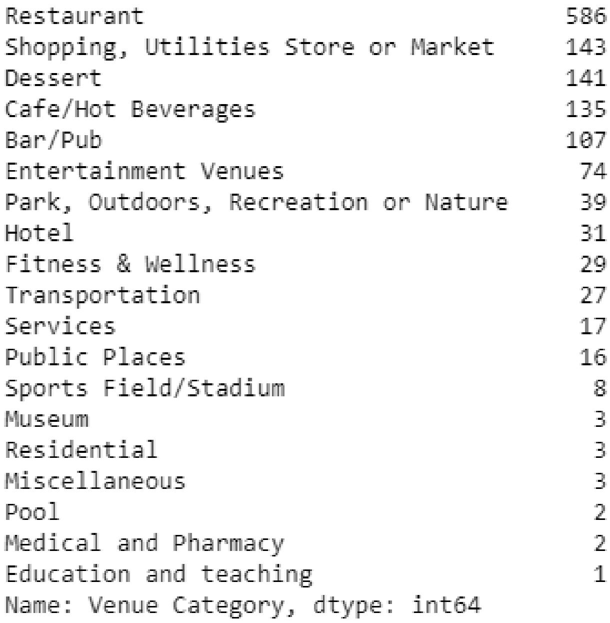


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Park, Outdoors, Recreation or Nature	Restaurant	Shopping, Utilities Store or Market	Bar/Pub	Public Places	Hotel	Sports Field/Stadium	Cafe/Hot Beverages	Fitness & Wellness	Dessert
1	Carnegie Hill	Restaurant	Bar/Pub	Cafe/Hot Beverages	Shopping, Utilities Store or Market	Fitness & Wellness	Services	Dessert	Entertainment Venues	Museum	Hotel
2	Central Harlem	Restaurant	Bar/Pub	Shopping, Utilities Store or Market	Entertainment Venues	Services	Fitness & Wellness	Cafe/Hot Beverages	Sports Field/Stadium	Public Places	Dessert
3	Chelsea	Restaurant	Shopping, Utilities Store or Market	Entertainment Venues	Cafe/Hot Beverages	Bar/Pub	Dessert	Services	Park, Outdoors, Recreation or Nature	Medical and Pharmacy	Fitness & Wellness
4	Chinatown	Restaurant	Shopping, Utilities Store or Market	Bar/Pub	Dessert	Cafe/Hot Beverages	Services	Entertainment Venues	Hotel	Sports Field/Stadium	Public Places

Table 2: Data set containing top 10 most common venue categories in each neighborhood.

The k means clustering algorithm was used to segment this dataset in five clusters. The cluster labels for each neighborhood were inserted as the ‘Cluster Label’ column into the data set. These clusters were then explored by data visualization to understand the segmentation. Further, k nearest neighbors algorithm was used to train a model on the final New York data set, taking the top ten most common venue category columns as independent variables and the assigned cluster labels as the target variable. A crucial point to consider in this work is that there aren’t perfect sharp differences among the five clusters. The schema of clustering cannot be manually controlled and the clustering algorithm automatically assigns labels based on its algorithm. Thus, the accuracy of the classification model to train on the clustered dataset will be reasonably low. In the current work, the best accuracy on the test set was achieved to be 50%.

The second part of the project deals with Mumbai city’s neighborhood data. An exhaustive list of neighborhoods in Mumbai city is available on the city’s Wikipedia page, which was imported as pandas data frame. Similar processing as New York neighborhood dataset were performed for the Mumbai city dataset. Foursquare location services were used to extract the venues in each neighborhood. The venues were categorized in over 170 categories. Thus, this dataset was exported for further processing so that the venues could be categorized in broad categories as done in New York City’s case. The final data set was created by storing the top ten most common venue categories in each neighborhood. This final data set was then fit in the model created earlier, and the predicted labels were stored. An analysis of each cluster of the neighborhoods was done through data visualization.



```
Restaurant                               586
Shopping, Utilities Store or Market      143
Dessert                                  141
Cafe/Hot Beverages                       135
Bar/Pub                                  107
Entertainment Venues                      74
Park, Outdoors, Recreation or Nature      39
Hotel                                     31
Fitness & Wellness                        29
Transportation                            27
Services                                  17
Public Places                             16
Sports Field/Stadium                       8
Museum                                     3
Residential                                3
Miscellaneous                              3
Pool                                       2
Medical and Pharmacy                       2
Education and teaching                     1
Name: Venue Category, dtype: int64
```

Table 3: Count of venues in different venue categories in Manhattan City

Results

This section will deal with analyzing the observations and describing the results. This section has been divided into three parts, for ease of understanding- New York City, Machine Learning Algorithms used and Mumbai City. As only the Manhattan borough of New York has been included in this study, the terms New York and Manhattan have been used interchangeably many times.

1. New York City

A glance at the count of venue categories in both city’s datasets suggests an abundance of restaurants, dessert shops, cafes and shopping places in comparison to other venue categories. While it is natural to expect such venues to be in large number is all neighborhoods, this data is also due to lesser listings of venues such offices, teaching institutions, or religious places on the location provider services in comparison to restaurants and cafes. Location services are frequently used by users to explore new places, read reviews about places online, or to locate them for traveling. People have a lesser tendency to search for reviews of places such as offices and colleges on location service providers. Instead, they prefer other specific websites for better results. Also, Foursquare is a less popular location service in India. Thus, the data extracted from Foursquare is inadequate for better results. Another important intuition drawn from this data is that the overwhelming number of 'Restaurant' venue categories will also be reflected in the most common venue categories data.

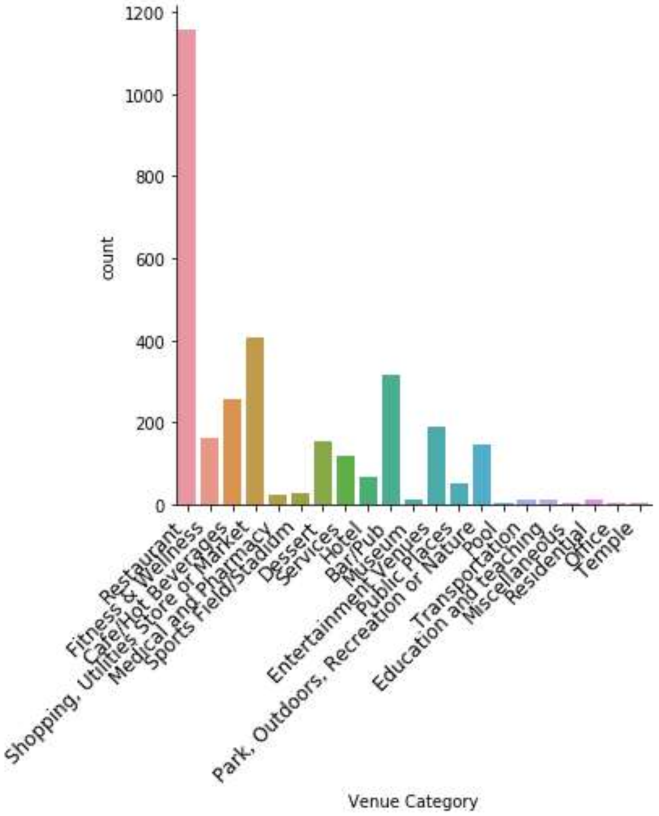
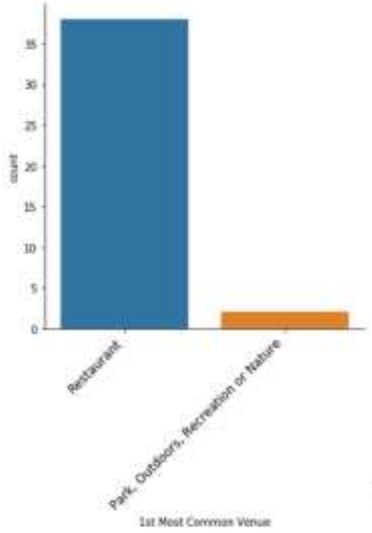
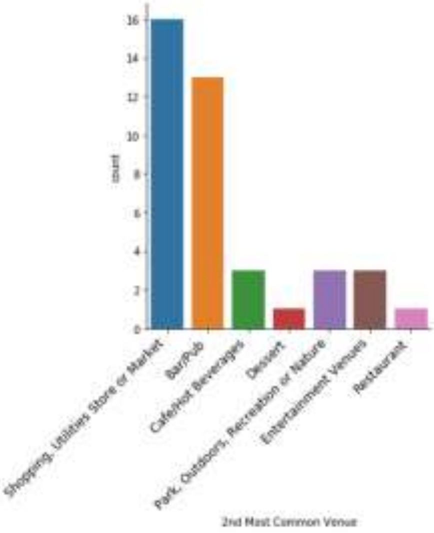


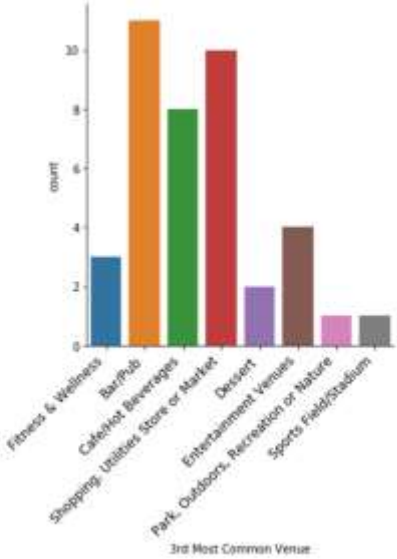
Fig 3: Count of venue categories in New York city data

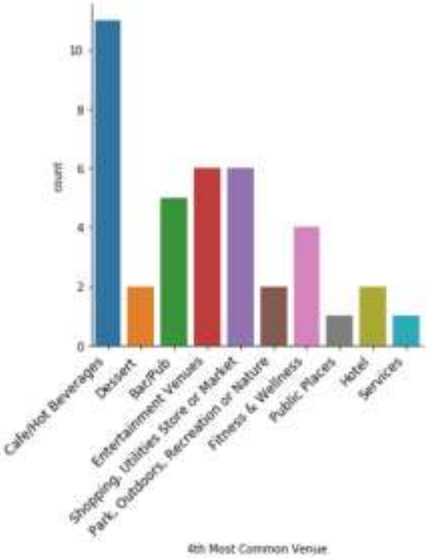
Another important data to consider is the most common venue categories per neighborhood. As the distribution in the count of venue categories was highly uneven, it was easy to interpret that more categories will appear as we move from the most common venue category to 10th most common venue category. In simpler words, 33% of all the venues in the Manhattan area belong to the 'Restaurant' category. Moreover, around 75% of all the venues belong to only top 5 venue categories by count. It is also reflected in the 10 plots below (Fig 4) that show different venue categories in accordance to their count of appearance in most common venue categories column to 10th most common venue category column for every neighborhood. Unsurprisingly among the 39 neighborhoods in Manhattan, 'Restaurants' are the most common kind of venues in 37 neighborhoods; the other two have 'Park, Outdoors, Recreation or Nature' as the most common category.

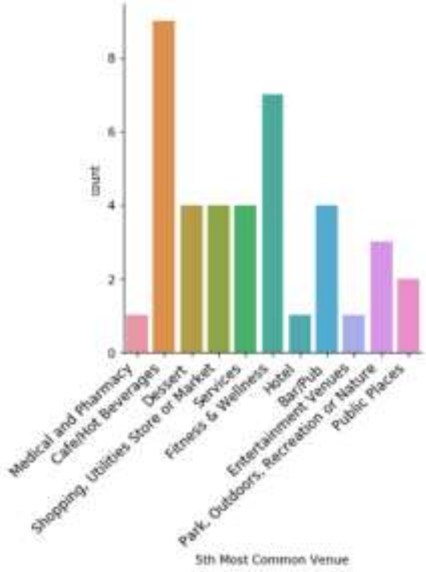
Among the second most common venue types in neighborhoods, there are seven different categories of venues marking their presence. The number of venue categories in the most common venues increases moving towards lesser common venue plot, having 13 different types of venues in the 10th most common venue categories column. Thus, it can be inferred that the neighborhoods become more distinct when we look at the less common venues types. It is not difficult to understand as it is the less frequent but distinct type of venues, that makes a neighborhood unique. Therefore while segmenting neighborhoods in different clusters, the less common venues will play a significant role in marking the distinction.

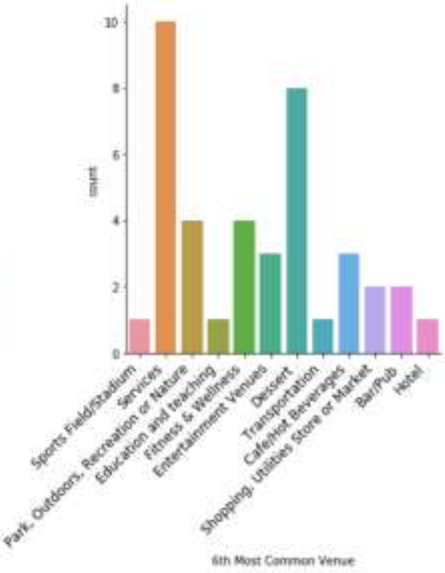












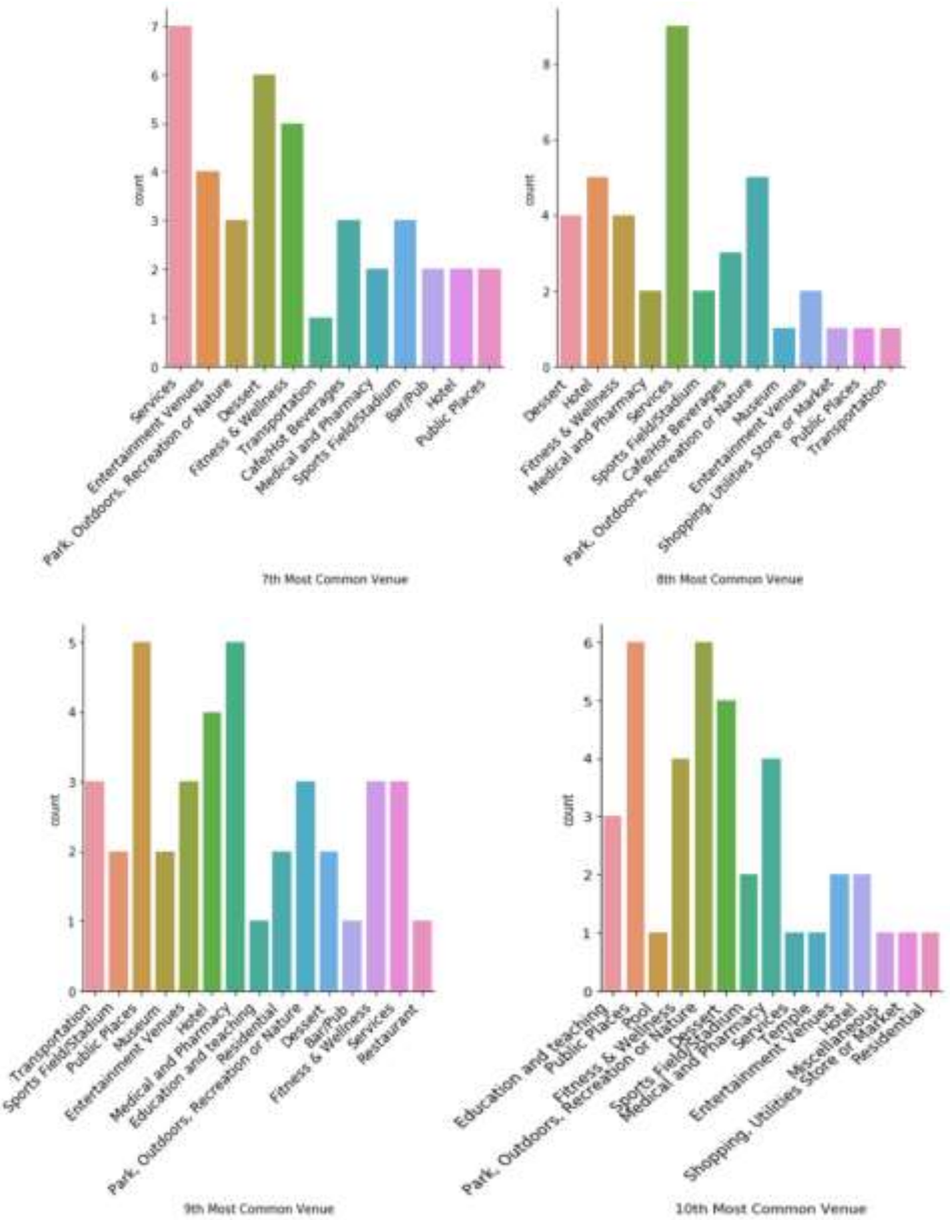


Fig 4: Most common venue categories in the neighborhoods (Manhattan)

The tables mentioned below are the hints to the solution of the mystery – how are the five clusters of Manhattan different from each other. Once this distinction among different clusters is established, the neighborhoods of the Mumbai city can be mapped consequently. As discussed above, the key factor will not be the more common venues. Rather, the less common but unique venues will create a major distinction among the different clusters. The five tables listed below, one for each cluster, contains data related to the different venue categories, their count, percentage of venue categories in a cluster, the overall percentage of that category, and the overall category percentage per neighborhood. Brief description of each column in the tables listed below are as follows:

- The 'Percentage in cluster' column indicates the distribution of all venues in the cluster among the different venue categories. (Eg. for the entry in row 1, (total 'Restaurant' venues in the cluster / total venues in the cluster)*100).
- The 'Overall Percentage' column indicates the distribution of each venue category among the five clusters. (Eg. for the entry in row 1, (total 'Restaurant' venues in the cluster/total 'Restaurant' venues in all clusters)*100).

- The 'Percentage per neighborhood' column simply indicates the 'Overall Percentage' per neighborhood for each category in the cluster. It is an indicator of the impact of the neighborhoods in a cluster with respect to certain venue categories. Thus if the 'Percentage per neighborhood' value of a category is higher in a particular cluster, it shows that the neighborhoods of that cluster have an exclusive/unique distinction in regards to that type of venues in comparison to any other neighborhood. This index has been used for comparison only and has no statistical significance on its own.

The Entries in each table are color-coded:

- Cells highlighted in Yellow color are important features in establishing the dissimilarities among the clusters. The column categories corresponding to such cells have lower 'Percentage in cluster' value but higher 'Overall Percentage' value, which indicates their exclusiveness to the cluster.
- Cells highlighted in orange color are also important, although less than the Yellow cells.
- Cells highlighted in Green color corresponds to the venue categories having higher impact per neighborhood in the cluster. All cells having an impact value higher than 5 are indicated.

<u>Cluster 1</u>

Venue Categories	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	479	44.85018727	41.4	3.18
Bar/Pub	129	12.07865169	41.08	3.16
Shopping, Utilities Store or Market	106	9.925093633	25.98	2
Cafe/Hot Beverages	84	7.865168539	32.68	2.51
Dessert	64	5.992509363	41.56	3.2
Entertainment Venues	47	4.400749064	24.74	1.903076923
Fitness & Wellness	43	4.026217228	26.71	2.054615385
Services	31	2.902621723	26.27	2.020769231
Park, Outdoors, Recreation or Nature	25	2.34082397	17.36	1.335384615
Hotel	23	2.153558052	34.85	2.680769231
Public Places	12	1.123595506	24.49	1.883846154
Medical and Pharmacy	8	0.74906367	33.33	2.563846154
Sports Field/Stadium	5	0.468164794	19.23	1.479230769
Museum	4	0.374531835	40	3.076923077
Education and teaching	3	0.280898876	30	2.307692308
Residential	2	0.187265918	18.18	1.398461538
Transportation	2	0.187265918	16.67	1.282307692
Pool	1	0.093632959	33.33	2.563846154
Total Neighborhoods in Cluster			**13**	
Total Venues in Cluster			**1068**	

<u>Cluster 2</u>

Venue Categories	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Park, Outdoors, Recreation or Nature	25	22.93578	17.36	5.786666667
Restaurant	21	19.26606	1.82	0.606666667
Shopping, Utilities Store or Market	12	11.00917	2.94	0.98
Bar/Pub	7	6.422018	2.23	0.743333333
Public Places	7	6.422018	14.29	4.763333333
Cafe/Hot Beverages	6	5.504587	2.33	0.776666667

Fitness & Wellness	6	5.504587	3.73	1.243333333
Sports Field/Stadium	6	5.504587	23.08	7.693333333
Hotel	4	3.669725	6.06	2.02
Transportation	4	3.669725	33.33	11.11
Services	3	2.752294	2.54	0.846666667
Dessert	2	1.834862	1.3	0.433333333
Education and teaching	2	1.834862	20	6.666666667
Entertainment Venues	2	1.834862	1.05	0.35
Residential	2	1.834862	18.18	6.06
Total Neighborhoods in Cluster			**3**	
Total Venues in Cluster			**109**	

Cluster 3

Venue Categories	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	300	33.97508	25.9291271	2.59291271
Shopping, Utilities Store or Market	173	19.5923	42.4019608	4.24019608
Cafe/Hot Beverages	77	8.720272	29.9610895	2.99610895
Bar/Pub	72	8.15402	22.9299363	2.29299363
Fitness & Wellness	60	6.795017	37.2670807	3.72670807
Dessert	52	5.889015	33.7662338	3.37662338
Services	43	4.869762	36.440678	3.6440678
Entertainment Venues	38	4.303511	20	2
Park, Outdoors, Recreation or Nature	20	2.265006	13.8888889	1.38888889
Hotel	17	1.925255	25.7575758	2.57575758
Public Places	8	0.906002	16.3265306	1.63265306
Medical and Pharmacy	6	0.679502	25	2.5
Museum	5	0.566251	50	5
Sports Field/Stadium	5	0.566251	19.2307692	1.92307692
Education and teaching	2	0.226501	20	2
Miscellaneous	2	0.226501	66.6666667	6.66666667
Office	1	0.11325	33.3333333	3.33333333
Pool	1	0.11325	33.3333333	3.33333333
Transportation	1	0.11325	8.33333333	0.833333333
Total Neighborhoods in Cluster			**10**	
Total Venues in Cluster			**883**	

Cluster 4

Venue Categories	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	134	44.22442	11.58168	2.316336
Park, Outdoors, Recreation or Nature	31	10.23102	21.52778	4.305556
Cafe/Hot Beverages	30	9.90099	11.67315	2.33463
Shopping, Utilities Store or Market	28	9.240924	6.862745	1.372549
Bar/Pub	20	6.60066	6.369427	1.2738854
Services	14	4.620462	11.86441	2.372882
Fitness & Wellness	13	4.290429	8.074534	1.6149068
Public Places	6	1.980198	12.2449	2.44898
Entertainment Venues	5	1.650165	2.631579	0.5263158
Dessert	4	1.320132	2.597403	0.5194806
Medical and Pharmacy	4	1.320132	16.66667	3.333334
Hotel	3	0.990099	4.545455	0.909091
Sports Field/Stadium	3	0.990099	11.53846	2.307692

Transportation	3	0.990099	25	5
Residential	2	0.660066	18.18182	3.636364
Miscellaneous	1	0.330033	33.33333	6.666666
Museum	1	0.330033	10	2
Office	1	0.330033	33.33333	6.666666
Total Neighborhoods in Cluster			**5**	
Total Venues in Cluster			**303**	

Cluster 5

Venue Categories	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	223	29.41952507	19.27398	2.141553333
Entertainment Venues	98	12.92875989	51.57895	5.730994444
Shopping, Utilities Store or Market	89	11.7414248	21.81373	2.423747778
Bar/Pub	86	11.34564644	27.38854	3.043171111
Cafe/Hot Beverages	60	7.915567282	23.3463	2.594033333
Park, Outdoors, Recreation or Nature	43	5.672823219	29.86111	3.317901111
Fitness & Wellness	39	5.145118734	24.2236	2.691511111
Dessert	32	4.221635884	20.77922	2.308802222
Services	27	3.562005277	22.88136	2.542373333
Hotel	19	2.506596306	28.78788	3.198653333
Public Places	16	2.110817942	32.65306	3.628117778
Sports Field/Stadium	7	0.92348285	26.92308	2.991453333
Medical and Pharmacy	6	0.791556728	25	2.777777778
Residential	5	0.659630607	45.45455	5.050505556
Education and teaching	3	0.395778364	30	3.333333333
Transportation	2	0.263852243	16.66667	1.851852222
Office	1	0.131926121	33.33333	3.703703333
Pool	1	0.131926121	33.33333	3.703703333
Temple	1	0.131926121	100	11.11111111
Total Neighborhoods in Cluster			**9**	
Total Venues in Cluster			**758**	

There are certain key points for every cluster that can be used to analyzing the uniqueness of each cluster. The key points are as follows:

- Cluster 1
  - 40% of all museums in Manhattan are situated in cluster 1.
  - Around 35% of all hotels in Manhattan are situated in cluster 1.
  - 33% of all pools in Manhattan are situated in cluster 1.
  - 25% of all Public places in Manhattan are situated in cluster 1.
  - 41.4% of all Restaurants as well as 41.08% of all Bars/Pubs in Manhattan are situated in cluster 1.
- Cluster 2
  - 23% of all Sports Fields/Stadiums with a high impact per neighborhood index.
  - 33.33% of all Transportation facilities with an exceptionally high impact per neighborhood.
  - 17.36% of all Park, Outdoors, Recreation or Nature venues, which also constitute 23% of all venues in Cluster 2.

  - 20% of all venues belonging to Educational and Teaching category in Manhattan lies in Cluster 2. The impact per neighborhood index is also notably high.
  - High number of Residential venues and fairly low number of Entertainment venues.
  - A reasonably good numbers (although least in comparison to other clusters) of Restaurant, Shops, Utilities Stores, and Market in the neighborhoods in this cluster.
- Cluster 3
  - 50% of all Museums and 26% of all hotels lie in Cluster 3.
  - 36% of all stores that offer some kind of Services in Manhattan are in Cluster 3.
  - 19% of all Sports fields/Stadiums, 16% of all Public Places, 14% of all Parks, Outdoors, Recreation and Nature venues as well as 20% of all Entertainment venues.
  - Over 42% of all Shops, Utilities Stores or Markets are in Cluster 3. There are also a fairly good amount of Restaurants, Cafes, Bars and Transportation facilities.
  - Over 37% of all venues belonging to the Fitness and Wellness category are situated here.
  - No residential venue.
- Cluster 4
  - Although only 10% of all venues in the cluster belong to Parks, Outdoors, Recreation or Nature category, they contribute 21% among all such venues in Manhattan.
  - 18% of all residential apartments/houses in Manhattan lie in cluster 4.
  - 17% of all Medicals and Pharmacies lie in Cluster 3.
  - Excellent availability of Transportation facilities.
  - 44% of all venues in cluster 3 are Restaurants. A relatively high number of such restaurants are specialty restaurants serving Asian, Indian, Middle Eastern and Mexican cuisine.
  - A reasonably low number of Entertainment venues and Hotels.
- Cluster 5
  - 52% of all Entertainment Venues
  - A fairly good number of Shops & Markets, Bars, Pub, Cafes, Parks, Outdoors, Recreational venues, and Natural places.
  - 24% of all Fitness & Wellness related venues are in cluster 5. Moreover, most of these venues are Gyms, Sports clubs, or Fitness centers that attract the young crowd.
  - Over 45% of all Residential venues lie in this cluster. It is also backed by good availability of transportation facilities.
  - 30% of all Educational and teaching venues lie in this cluster.
  - High number of Sports venues, Public places, hotels, and stores offering services of different kinds.
  - 29% of all venues in this cluster are restaurants, many of which are economical fast food shops serving all different kinds of food.

The following key points about each neighborhood suggest how each cluster is different from one another. Thus, the five clusters, which are comprised of different neighborhoods, can be identified as follows:

Cluster 1: Decent residential areas having a mix of general characteristics desired in a neighborhood.

Cluster 2: Suitable for people who love the calm and simple life. Suggested for families with young children, older adults and people wanting to get away from busy city life.

Cluster 3: Major tourist and shopping place. Suggested for tourists, shopaholics, and outdoor people.

Cluster 4: Decent residential areas. Suitable for immigrants who tend to live in societies with other immigrants as well as people belonging to poor sections of the society.

Cluster 5: Suitable for youngsters, bachelors, students, and people interested in chaotic yet quality city life.

2. Machine Learning Algorithms

I. K means clustering:

K means clustering is an unsupervised machine learning algorithm that segments the data in different clusters. The algorithm initiates clustering by randomly assigning cluster centroids and associating data points to the clusters by determining the nearest centroid. The algorithm iterates this process by reassigning centroids to minimize the cost function. K means clustering algorithm was used to segment the neighborhoods in five clusters based on the top ten most common venue categories in each neighborhood. The following table gives the number of neighborhoods assigned to each cluster label (marked 0 to 4).

Cluster Label	Number of Neighborhoods
0	13
1	3
2	10
3	5
4	9

II. K nearest neighbors:

This is a supervised machine learning algorithm that is used for classification. To classify a data point to a particular class, this algorithm takes into consideration its k nearest neighbor data points. The class is determined by a simple majority from the classes of its nearest neighbor. The number of nearest neighbors to consider (k) is a critical part of this algorithm. The optimal k value is simply determined by running the algorithm for a series of different values of k, and choosing the one with the highest accuracy.

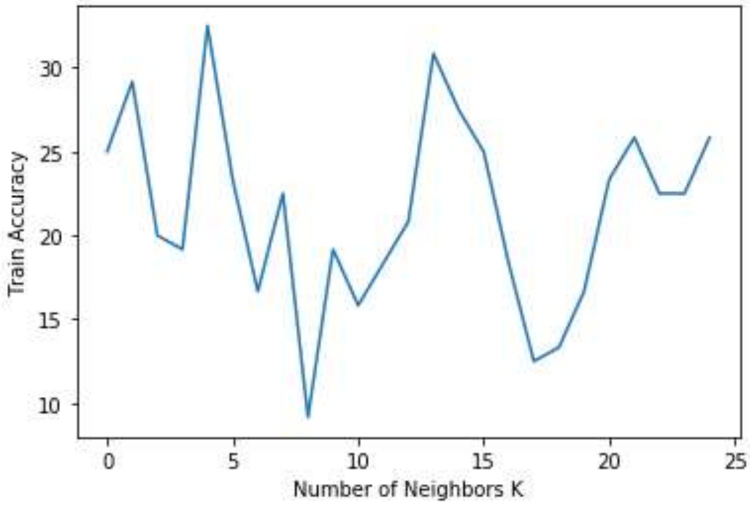
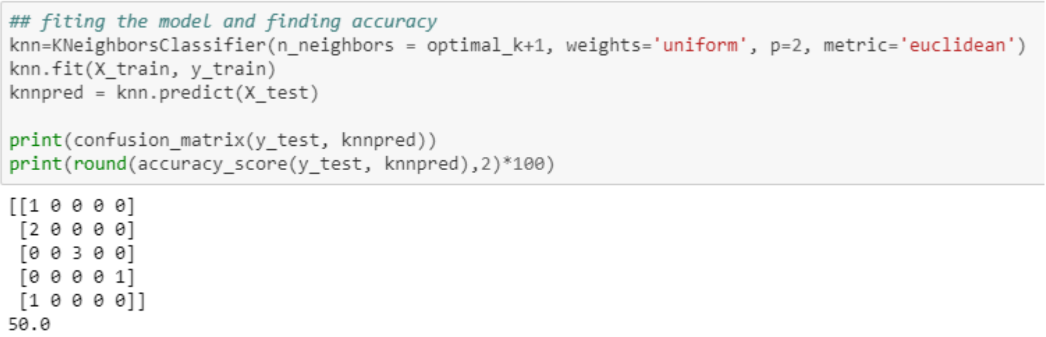


Fig 5: Plot of accuracy for different values of k

In this project, the optimal value of the nearest neighbors was found to be 4. It is also important to understand that this algorithm works by considering the distance of each data point from other data points in the data set. Hence, scaling the independent variables is very important for the sake of better results. The variables were encoded by label encoder beforehand and Standard Scaler function was used for scaling. Knn algorithm with the following parameters was trained on 80% of the dataset, while the rest was used for testing. An accuracy of 50% was achieved on the test set.



```
## fiting the model and finding accuracy
knn=KNeighborsClassifier(n_neighbors = optimal_k+1, weights='uniform', p=2, metric='euclidean')
knn.fit(X_train, y_train)
knnpred = knn.predict(X_test)

print(confusion_matrix(y_test, knnpred))
print(round(accuracy_score(y_test, knnpred),2)*100)
```

```
[[1 0 0 0 0]
 [2 0 0 0 0]
 [0 0 3 0 0]
 [0 0 0 0 1]
 [1 0 0 0 0]]
50.0
```

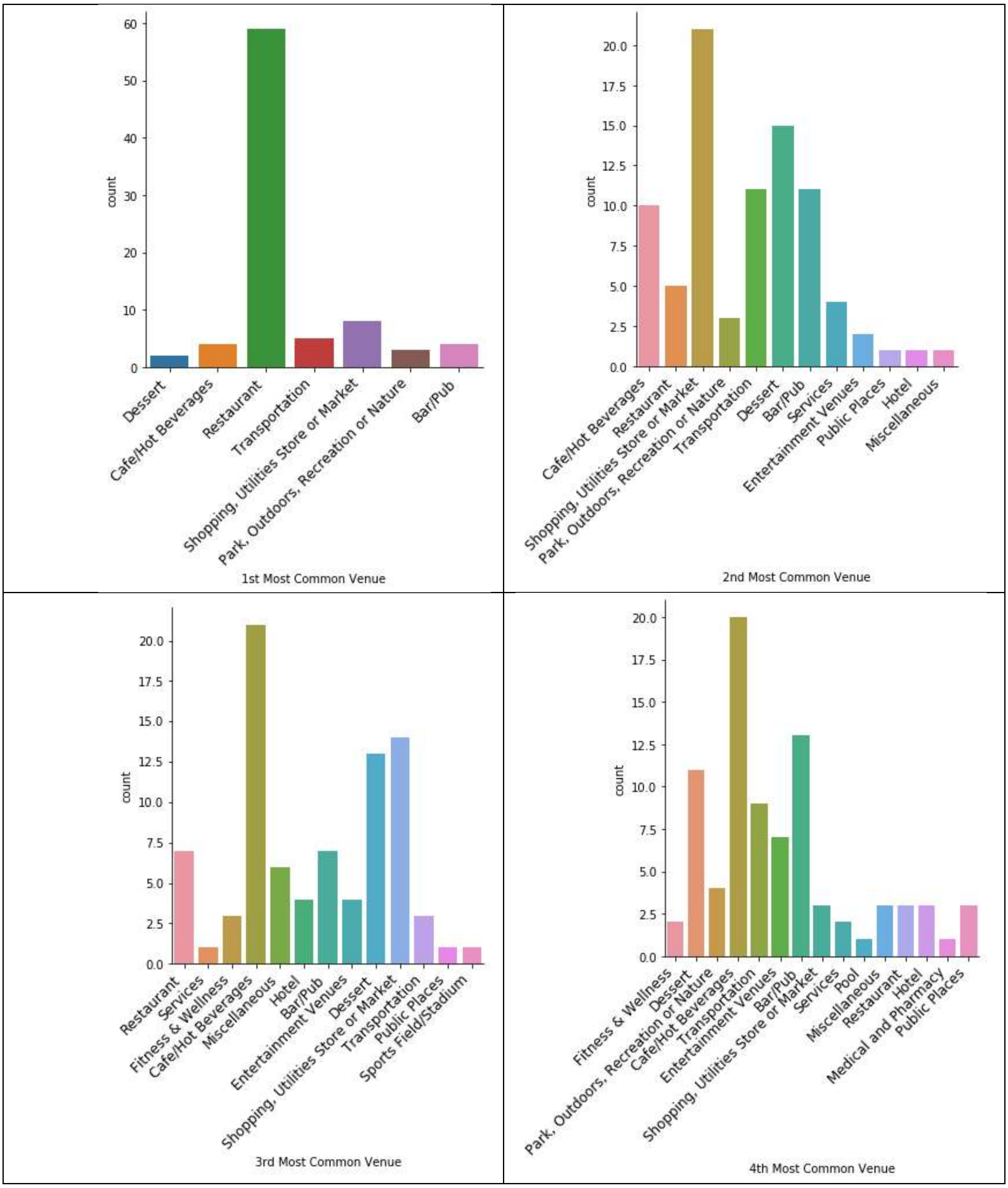
Fig 6: Parameters used in knn algorithm and accuracy of model on test set.

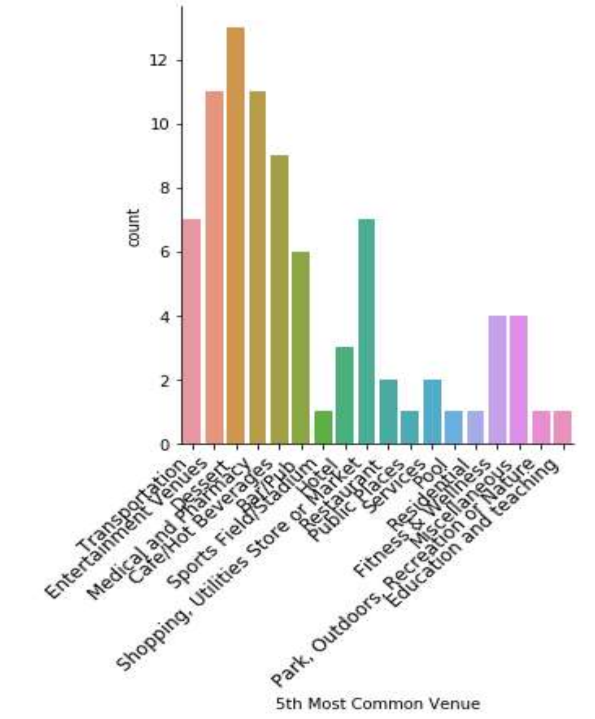
The effectiveness of this model can be argued based on its accuracy. However, there are many factors responsible for this poor performance. Still for exploration and pursuit for the objective of this project, this model was used to classify the neighborhoods of Mumbai city in different clusters. Mumbai City neighborhoods assigned to each cluster are as follows:

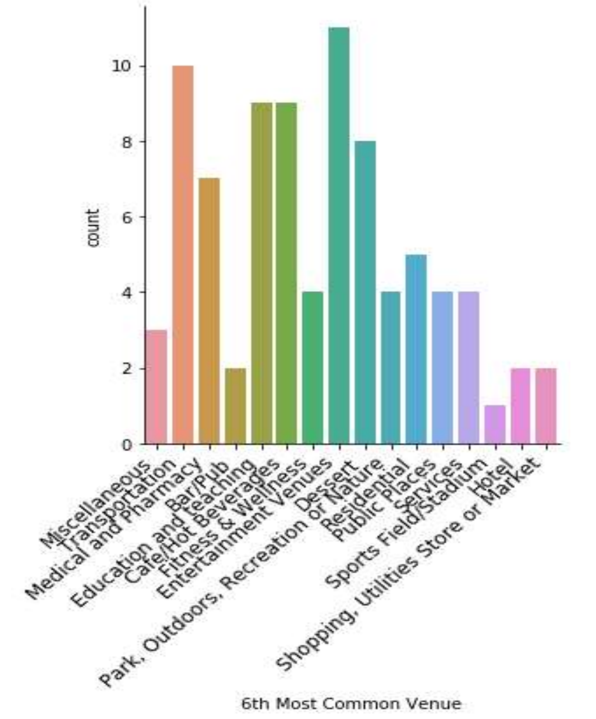
Cluster Label	Number of Neighborhoods
0	22
1	2
2	37
3	12
4	12

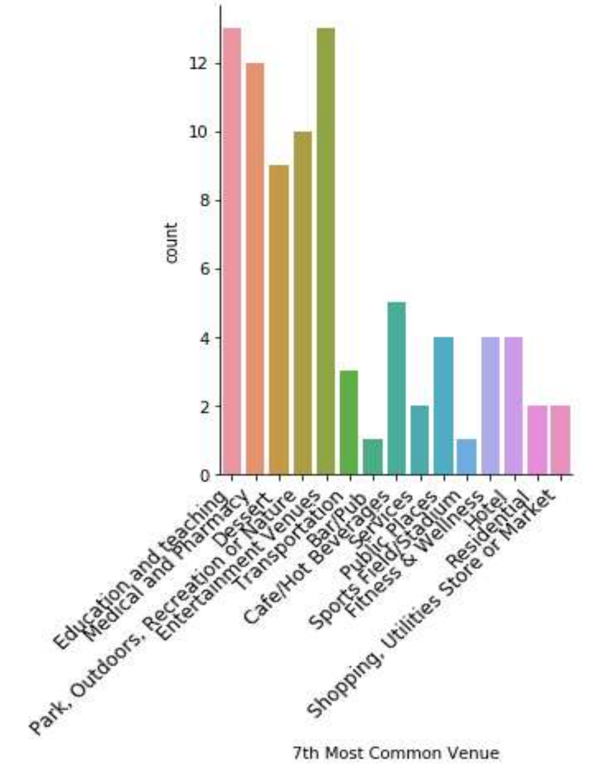
3. Mumbai City

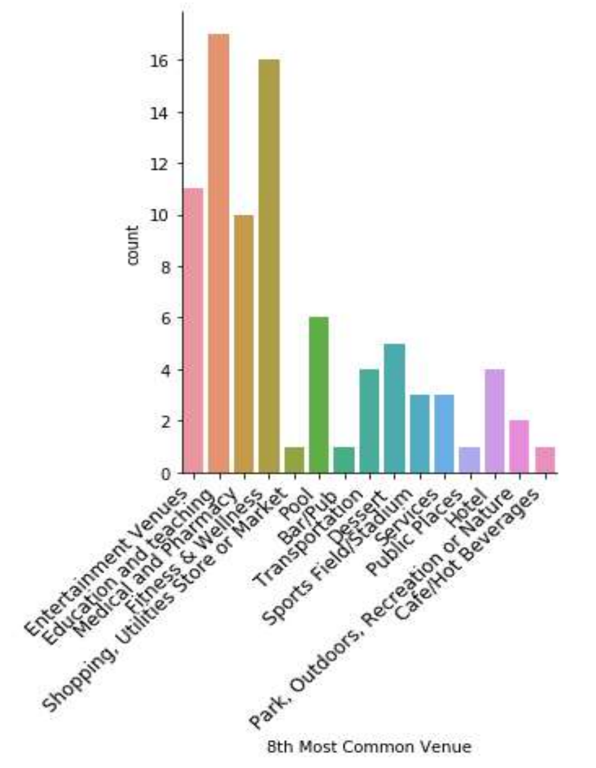
Some of the general observations made for the New York City dataset are also applicable to Mumbai City. Restaurants, Cafes, Bars, Shopping and Utilities store etc. dominate the list of venues. Also more venue categories appear as we move from the list of most common venues to 10th most common venues columns. Although there is clearly more diversity and spread of venue categories in the most common venues data as compared to New York City's case, which can be observed in the following plots showing the count of venue categories by their number of appearances in the top ten most common venue categories columns in Mumbai City dataset. The diversity observed in following plots is a positive sign for a city as it indicates the greater number of options available to its inhabitants. However, in Mumbai city's case, this can also be attributed to the inadequate amount of data gathered from Foursquare location service due to its unpopularity as comparison to other location services in India.


count
60
50
40
30
20
10
0
Dessert
Cafe/Hot Beverages
Restaurant
Transportation
Shopping, Utilities Store or Market
Park, Outdoors, Recreation or Nature
Bar/Pub
1st Most Common Venue
count
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Cafe/Hot Beverages
Restaurant
Shopping, Utilities Store or Market
Park, Outdoors, Recreation or Nature
Transportation
Dessert
Bar/Pub
Services
Entertainment Venues
Public Places
Hotel
Miscellaneous
2nd Most Common Venue
count
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Restaurant
Services
Fitness & Wellness
Cafe/Hot Beverages
Miscellaneous
Hotel
Bar/Pub
Entertainment Venues
Dessert
Shopping, Utilities Store or Market
Transportation
Public Places
Sports Field/Stadium
3rd Most Common Venue
count
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Fitness & Wellness
Dessert
Park, Outdoors, Recreation or Nature
Cafe/Hot Beverages
Transportation
Entertainment Venues
Bar/Pub
Shopping, Utilities Store or Market
Services
Pool
Miscellaneous
Restaurant
Hotel
Medical and Pharmacy
Public Places
4th Most Common Venue


count
0
2
4
6
8
10
12
Transportation
Entertainment Venues
Dessert
Medical and Pharmacy
Cafe/Hot Beverages
Bar/Pub
Sports Field/Stadium
Hotel
Shopping, Utilities Store or Market
Restaurant
Public Places
Services
Pool
Residential
Fitness & Wellness
Miscellaneous
Park, Outdoors, Recreation or Nature
Education and teaching
5th Most Common Venue


count
0
2
4
6
8
10
Miscellaneous
Transportation
Medical and Pharmacy
Bar/Pub
Education and teaching
Cafe/Hot Beverages
Fitness & Wellness
Entertainment Venues
Dessert
Park, Outdoors, Recreation or Nature
Residential
Public Places
Services
Sports Field/Stadium
Hotel
Shopping, Utilities Store or Market
6th Most Common Venue


count
0
2
4
6
8
10
12
Education and teaching
Medical and Pharmacy
Dessert
Park, Outdoors, Recreation or Nature
Entertainment Venues
Transportation
Bar/Pub
Cafe/Hot Beverages
Services
Public Places
Sports Field/Stadium
Fitness & Wellness
Hotel
Residential
Shopping, Utilities Store or Market
7th Most Common Venue


count
0
2
4
6
8
10
12
14
16
Entertainment Venues
Education and teaching
Medical and Pharmacy
Fitness & Wellness
Shopping, Utilities Store or Market
Pool
Bar/Pub
Transportation
Dessert
Sports Field/Stadium
Services
Public Places
Hotel
Park, Outdoors, Recreation or Nature
Cafe/Hot Beverages
8th Most Common Venue

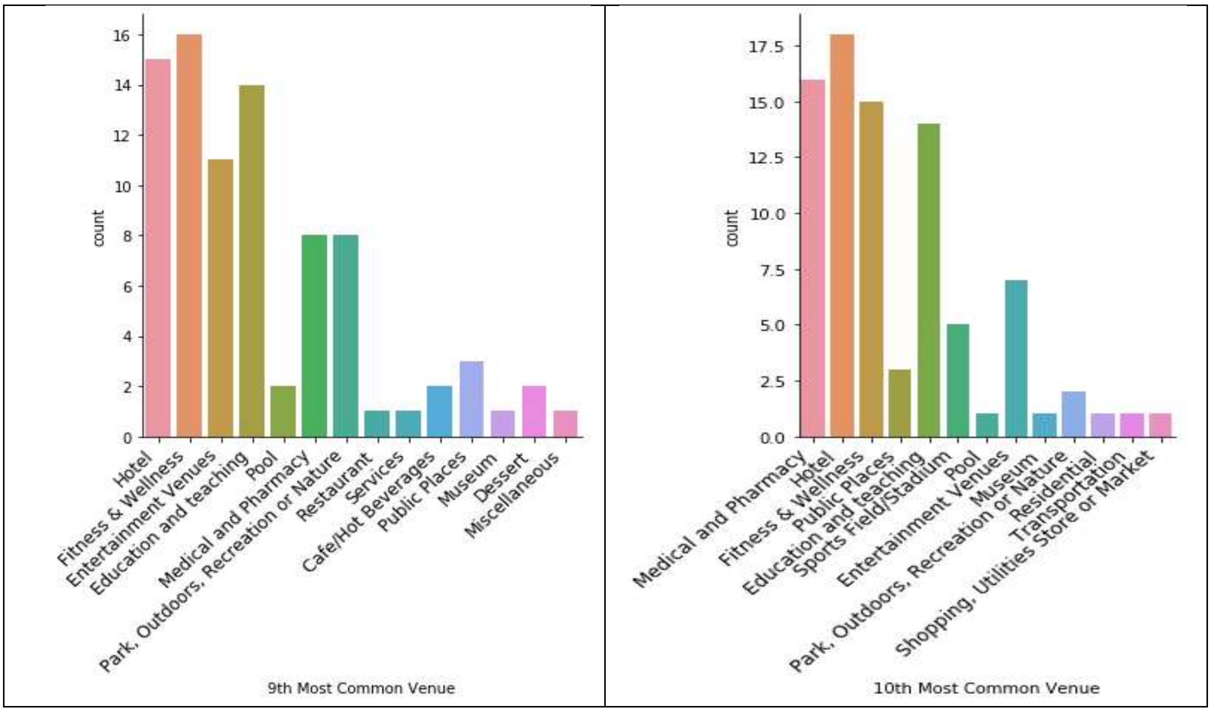


Fig 6: Most common venue categories in Mumbai City neighborhoods

Another critical analysis to be made is observing the clusters of neighborhoods mapped by the classification model. These clusters are mapped according to the resemblance of their characteristics with the clusters in Manhattan. This analysis could be interesting as both cities are very different from one another. While Manhattan is one of the most well-planned business districts in the world, Mumbai is extremely chaotic due to its unplanned mushroom growth. When a city grows without planning, its areas develop randomly in accordance with the demands of its inhabitants and its offering to the private investors. The growth attracts more inhabitants as well as investors, and this cycle runs indefinitely until any drastic steps are taken by the administration in shaping the growth in a particular direction.

Therefore, it is natural to expect most of the neighborhoods to be alike with minimal distinctions. This might result in some clusters occupying the majority of neighborhoods, as has happened in the case of Mumbai. Around 70% of the neighborhoods are assigned to only two clusters. Moreover, one of the clusters got only 2 neighborhoods out of 85. The results would have been far more monotonous if clustering was performed independently on the neighborhoods, and not being under the influence from New York's model. However, as the clusters have been created through New York's model, each cluster must ideally share the same characteristics (more or less) as its counterpart in New York city having the same cluster label.

The following tables, each for one cluster, mentions the characteristics of each cluster. The parameters are same as in New York's datasets, and share a similar color-coding protocol.

Cluster 1

Venue Category	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	168	43.75	28.66894198	1.303133726
Dessert	51	13.28125	36.17021277	1.64410058
Cafe/Hot Beverages	39	10.15625	28.88888889	1.313131313
Shopping, Utilities Store or Market	32	8.333333333	22.37762238	1.017164654
Bar/Pub	29	7.552083333	27.10280374	1.231945624
Entertainment Venues	21	5.46875	28.37837838	1.28992629
Fitness & Wellness	8	2.083333333	27.5862069	1.253918495
Hotel	8	2.083333333	25.80645161	1.173020528
Park, Outdoors, Recreation or Nature	8	2.083333333	20.51282051	0.932400932
Public Places	5	1.302083333	31.25	1.420454545
Transportation	4	1.041666667	14.81481481	0.673400673
Services	3	0.78125	17.64705882	0.802139037
Sports Field/Stadium	3	0.78125	37.5	1.704545455
Pool	2	0.520833333	100	4.545454545
Miscellaneous	1	0.260416667	33.33333333	1.515151515
Museum	1	0.260416667	33.33333333	1.515151515
Residential	1	0.260416667	33.33333333	1.515151515
Total neighborhoods in cluster				**22**
Total venues in cluster				**384**

Cluster 2

Venue Category	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Bar/Pub	9	29.03225806	8.411214953	4.205607477
Restaurant	9	29.03225806	1.535836177	0.767918089
Entertainment Venues	5	16.12903226	6.756756757	3.378378378
Dessert	3	9.677419355	2.127659574	1.063829787
Hotel	2	6.451612903	6.451612903	3.225806452
Shopping, Utilities Store or Market	2	6.451612903	1.398601399	0.699300699
Cafe/Hot Beverages	1	3.225806452	0.740740741	0.37037037
Total neighborhoods in cluster				**2**
Total venues in cluster				**31**

Cluster 3

Venue Category	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	253	44	43.17406143	1.166866525
Shopping, Utilities Store or Market	64	11.13043478	44.75524476	1.20960121
Cafe/Hot Beverages	54	9.391304348	40	1.081081081
Dessert	52	9.043478261	36.87943262	0.996741422
Bar/Pub	37	6.434782609	34.57943925	0.934579439
Entertainment Venues	29	5.043478261	39.18918919	1.059167275
Park, Outdoors, Recreation or Nature	21	3.652173913	53.84615385	1.455301455
Transportation	18	3.130434783	66.66666667	1.801801802
Hotel	13	2.260869565	41.93548387	1.133391456
Services	12	2.086956522	70.58823529	1.907790143

Fitness & Wellness	10	1.739130435	34.48275862	0.931966449
Public Places	5	0.869565217	31.25	0.844594595
Sports Field/Stadium	3	0.52173913	37.5	1.013513514
Miscellaneous	2	0.347826087	66.66666667	1.801801802
Museum	2	0.347826087	66.66666667	1.801801802
Total neighborhoods in cluster				**37**
Total venues in cluster				**575**

Cluster 4

Venue Category	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	90	40.7239819	15.35836177	1.279863481
Cafe/Hot Beverages	23	10.40723982	17.03703704	1.419753086
Dessert	23	10.40723982	16.31205674	1.359338061
Shopping, Utilities Store or Market	20	9.049773756	13.98601399	1.165501166
Bar/Pub	19	8.597285068	17.75700935	1.479750779
Entertainment Venues	11	4.977375566	14.86486486	1.238738739
Fitness & Wellness	8	3.619909502	27.5862069	2.298850575
Park, Outdoors, Recreation or Nature	7	3.167420814	17.94871795	1.495726496
Hotel	6	2.714932127	19.35483871	1.612903226
Public Places	4	1.809954751	25	2.083333333
Transportation	3	1.357466063	11.11111111	0.925925926
Medical and Pharmacy	2	0.904977376	100	8.333333333
Residential	2	0.904977376	66.66666667	5.555555556
Services	2	0.904977376	11.76470588	0.980392157
Sports Field/Stadium	1	0.452488688	12.5	1.041666667
Total neighborhoods in cluster				**12**
Total venues in cluster				**221**

Cluster 5

Venue Category	Count	Percentage in cluster	Overall Percentage	Percentage per neighborhood
Restaurant	66	42.30769231	11.26279863	0.938566553
Shopping, Utilities Store or Market	25	16.02564103	17.48251748	1.456876457
Cafe/Hot Beverages	18	11.53846154	13.33333333	1.111111111
Bar/Pub	13	8.333333333	12.14953271	1.012461059
Dessert	12	7.692307692	8.510638298	0.709219858
Entertainment Venues	8	5.128205128	10.81081081	0.900900901
Fitness & Wellness	3	1.923076923	10.34482759	0.862068966
Park, Outdoors, Recreation or Nature	3	1.923076923	7.692307692	0.641025641
Hotel	2	1.282051282	6.451612903	0.537634409
Public Places	2	1.282051282	12.5	1.041666667
Transportation	2	1.282051282	7.407407407	0.617283951
Education and teaching	1	0.641025641	100	8.333333333
Sports Field/Stadium	1	0.641025641	12.5	1.041666667
Total neighborhoods in cluster				**12**
Total venues in cluster				**156**

The following observations are made about the clusters in Mumbai City:

Cluster 1:

Ideally, the neighborhoods must be decent residential areas having a mix of all general characteristics. In the present case, the neighborhoods have ample amount of restaurants, cafes, markets, public places, parks and other types of venues. Thus they offer a wide array of options and, therefore, can be termed as ‘general’ neighborhoods. As expected, 22 out of 85 neighborhoods of Mumbai are situated in this cluster, making it the second most populated cluster of the city.

Cluster 2:

Ideally, the neighborhoods should be calm and peaceful, attracting people who want to get away from busy city life. Cluster 2 matches the profile most suitably in comparison to any other cluster of Mumbai City. Although its characteristics are still very different from cluster 2 in New York. While the neighborhoods of New York had more open spaces (such as parks, natural venues, sports fields) and fewer restaurants and bars, the case with Mumbai is entirely opposite. Still, the lesser number of Hotels and Cafes, and absence of many other types of venues in the neighborhoods suggest a similar profile. Unsurprisingly in a city like a Mumbai, only 2 neighborhoods are matching such a profile.

Cluster 3:

Ideally, the neighborhoods should be more chaotic and busy with tourists and customers flocking to restaurants, bars, public places, museums, and markets etc. Considering Mumbai, it’s safe to argue that a majority of its neighborhoods may end up in this cluster. Unsurprisingly, 37 of its neighborhoods belong to this cluster. The cluster also has good transportation facilities, serving to the crowd of people traveling to and from these neighborhoods. The cluster also has 54% of the city’s parks, outdoors and recreational venues unlike the case of New York where the number was below 14%, indicating the city’s poor planning. Also there wasn’t any residential venue listed in this cluster in New York, as has been the case with Mumbai. Although, this could be again due to an inadequate amount of data as most busy areas of India’s cities are also crowded residential areas.

Cluster 4:

Ideally, the neighborhoods should be more suitable for either immigrant and poor section of the society, or well off people living in closed societies. The suitability of this cluster for two extremely different types of people may be disappointing. However, the current model cannot make such a distinction as such factors are not included in the process. For example, the current model treats all restaurants as similar, without making any distinction in roadside dhabas, economical fast food joints, or expensive diner. The neighborhoods have less entertainment venues and stores offering services and have major residential areas.

Cluster 5:

Ideally, the neighborhoods should be suitable for students, bachelors and young people interested in experiencing the busy yet quality city life. Although due to inadequate data, especially in regards of Residential venues, a clear picture cannot be obtained about these neighborhoods. Comparing all clusters of Mumbai to the respective clusters of New York, this cluster inarguably differs the most from cluster 5 of New York.

Discussion

All the points mentioned in the result section have been well described there. Still, several key observations can be drawn that have yet not been discussed. The objective of the project was to map Mumbai considering New York City as an ideal model for development. The efforts made in this project have been aimed to fulfill this objective. Yet, many factors poised challenges and can be rectified in future studies:

1. For the ease of processing, only the Manhattan borough of New York City was considered in this project. As there are only 40 neighborhoods in Manhattan (where the model was supposed to be trained) in comparison to 85 neighborhoods in Mumbai (dataset for prediction), the insufficient amount of data for testing hampered the accuracy of the model.
2. There are limitations in making API calls associated with the free developer account on Foursquare, which restricts the usage.
3. There isn't sufficient data available regarding categories like Universities, Schools, Hostels, and Offices, which affect the quality of the analysis.
4. The unpopularity of Foursquare in India makes it unsuitable for such a study. However, the expenses involved in using other location services such as Google Maps makes it the only viable option.
5. There isn't a way to further categorize a venue based on its affordability. If such a thing could happen, the analysis would be have been far more effective.

Based on the clustering performed on New York City's data, a model was created that was further implanted on Mumbai City. The clusters formed by the classification model suggest some severe lapses in the city's planning.

- There are only 39 venues belonging to 'Parks, Outdoors, Recreation or Nature' category in a vast and populous city like Mumbai, in comparison to 144 in just Manhattan area of New York. It is an important indicator of the differences in the quality of life offered by the two cities to its inhabitants.
- As discusses earlier, the 'Percentage per neighborhood' column in the Cluster tables is an indicator of the impact of the neighborhoods in a cluster in regards to specific venues. Thus if the 'Percentage per neighborhood' value of a category is higher in a particular cluster, it shows that the neighborhoods of that cluster have an exclusive/unique distinction in regards to that type of venue in comparison to any other neighborhood. Any cell having value more than 5 is counted as highly impactful and highlighted in green color. There are 13 incidents of a cluster having high impact value in New York's case, as compared to only 3 in Mumbai.
- Only 25% of neighborhoods in New York belong to Cluster 3, which includes neighborhoods having higher than average crowds and chaos. In comparison, 44% of Mumbai's neighborhoods belong to Cluster 3.

There are also certain observations made in the analysis which indicates the errors, and hence need for addressing, in this study.

- The presence of Dagdi Chawl and Hiranandani Gardens, two neighborhoods with residents belonging to opposite ends of economic status, poses a challenge to the effectiveness of this study. Thus some economic parameters could also be included in the clustering process to enhance the quality of the report.
- The neighborhoods belonging to Cluster 2 should be calm and peaceful. Yet somehow, Dharavi, the largest slum in the world, appears in this cluster.

Conclusion

The report aims at looking at Mumbai through the New York lens, and observing the key similarities as well as differences, and identifying the key areas of improvement. Despite the inadequate amount of data and limitations of using Foursquare in such a study, the report has successfully laid down critical observations and discussed some significant issues. Identifying different areas of Mumbai having similarities to New York will help plan infrastructure development and resource management of those areas to achieve their understated potential. It can also indicate the potential of specific neighborhoods that might not have been looked at previously. This study suggests an entirely different way of planning a city’s development.

The critical difference between Mumbai and New York is in the vision and foresight of its authorities. While a large portion of New York was planned well in advance under The Commissioner’s plan of 1811, Mumbai didn’t witness any such attempts. And that has hampered its ability to achieve its true potential.