A TALE OF TWO CITIES:

Looking at Mumbai through the New York lens

A Project Report By:

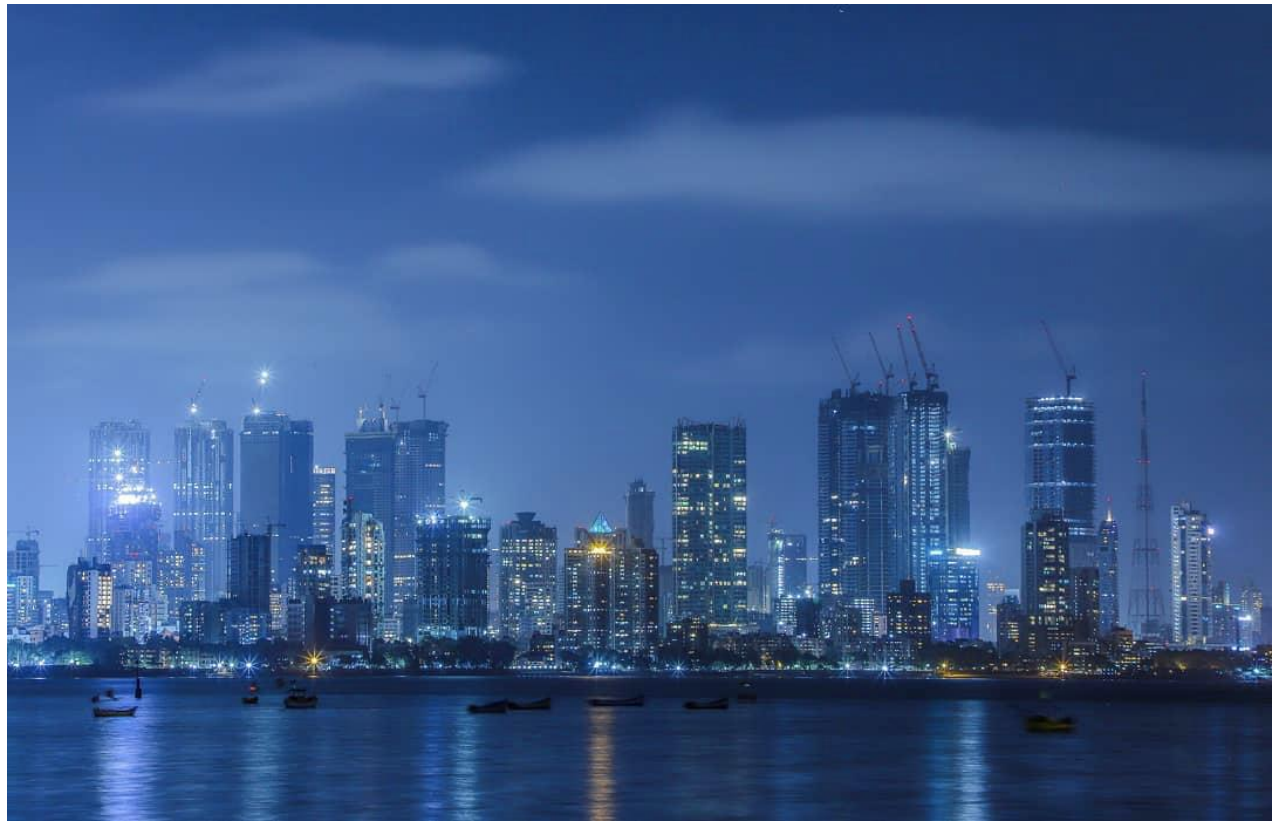
Parth Sharma

New York City

- 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.

Mumbai City

- 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 richest 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.



Data and Methodology

- The first part of the project deals with accessing New York City's neighborhoods' data and segmenting them in clusters. 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

```
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
```

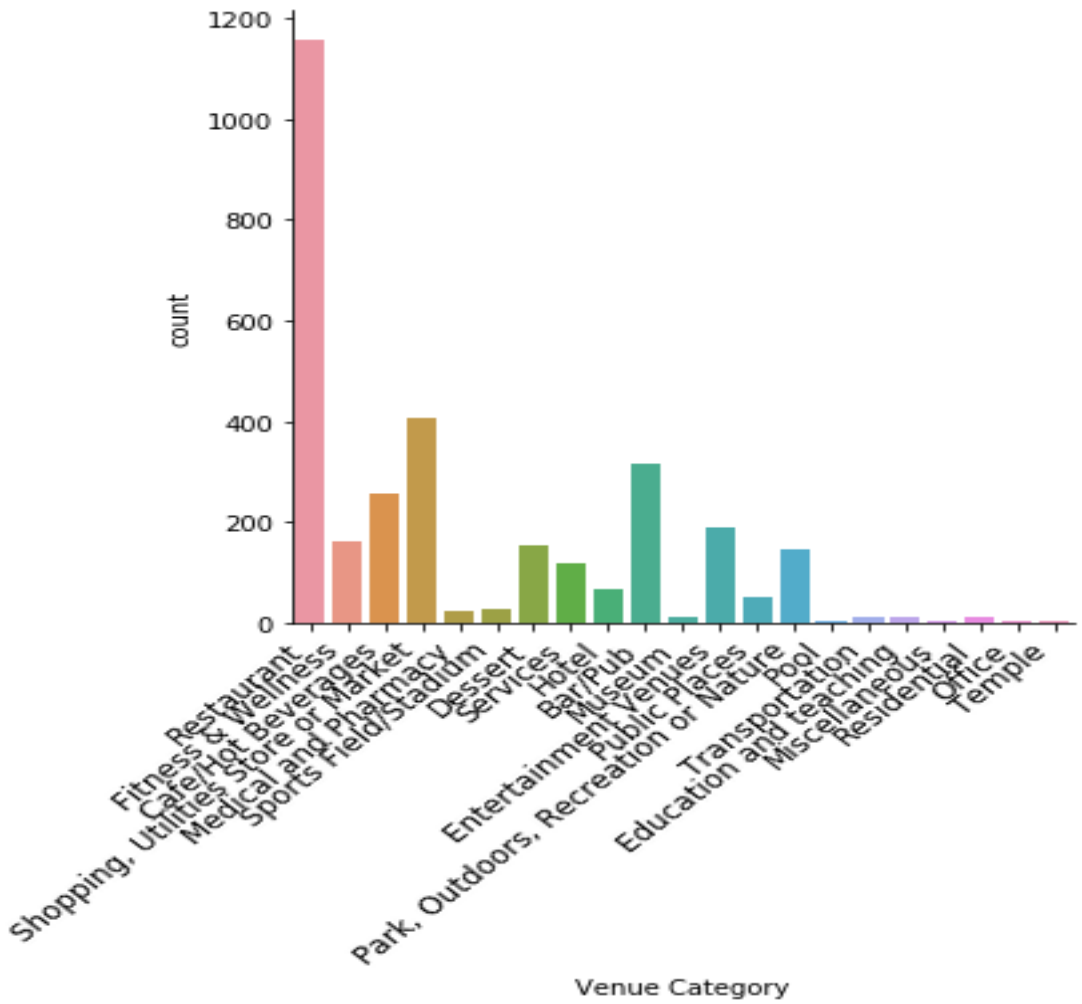
- The dataset was then processed to get frequency of appearance for all venue category in each neighborhood. A final dataset comprising of each neighborhood and its top ten most common venue categories were compiled by defining a function to extract such a dataset.
- Similarly, the Mumbai City data was exported from its Wikipedia page and processed to generate a similar dataset.
- The k means clustering algorithm was used to segment this dataset in five clusters. 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. This model was then used to classify each Mumbai City neighborhood in five clusters

```
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
```

Results and Discussion

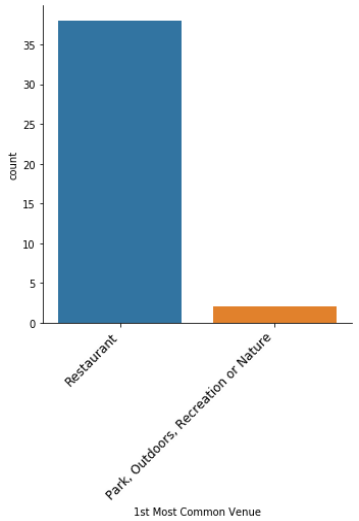
New York City

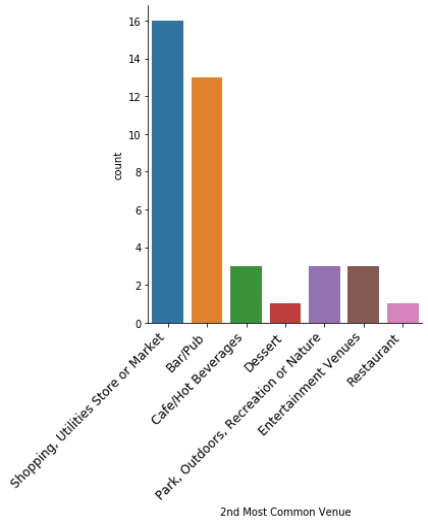
- A quick glance at the count of venue categories in both city's datasets suggest an abundance of restaurants, dessert shops, cafes and shopping places in comparison to other venue categories. While it is natural to expect such venues to be in large number is all neighborhoods, this data is also due to lesser listings of venues such offices, teaching institutions, or religious places on the location provider services in comparison to restaurants and cafes.
- 33% of all the venues in Manhattan area belong to the 'Restaurant' category. Moreover, around 75% of all the venues belong to only top 5 venue categories by count.
- It can be inferred that the neighborhoods become more distinct when we look at the less common venues types. Thus, it is the less frequent but distinct type of venues that make a neighborhood unique.

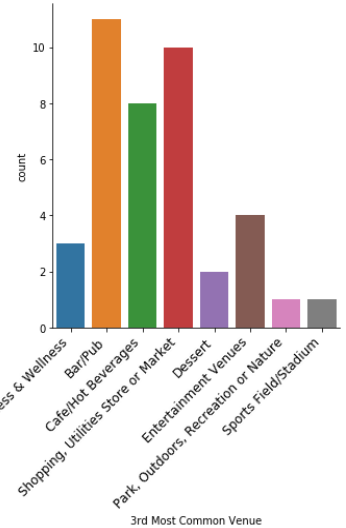


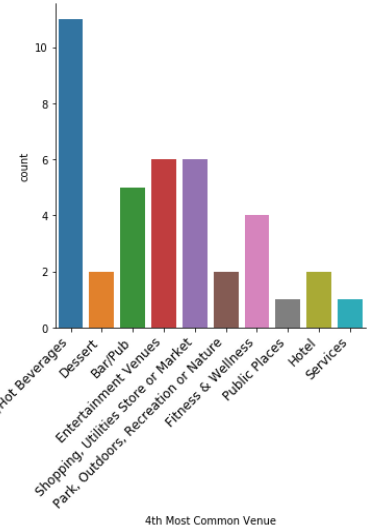
Count of venue categories in New York city data

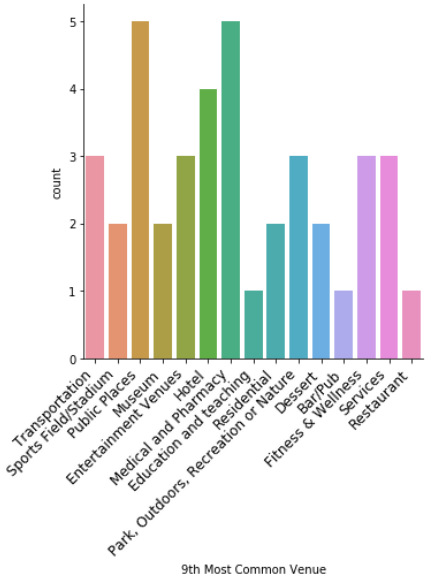
It is also reflected in the 10 plots below that show different venue categories in accordance to their count of appearance in most common venue categories column to 10th most common venue category column for every neighborhood.

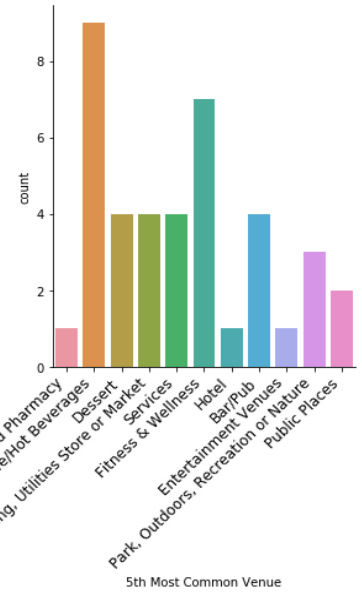


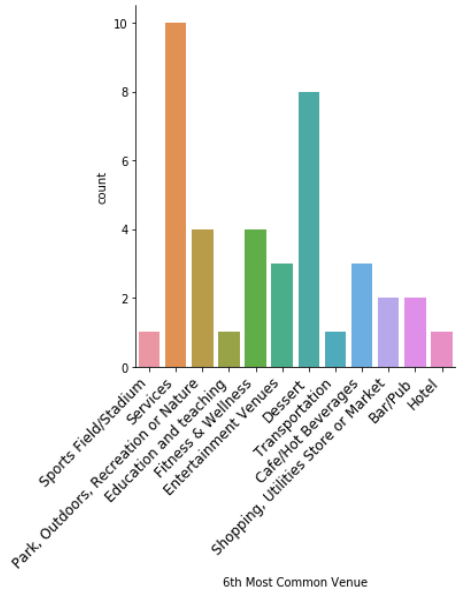


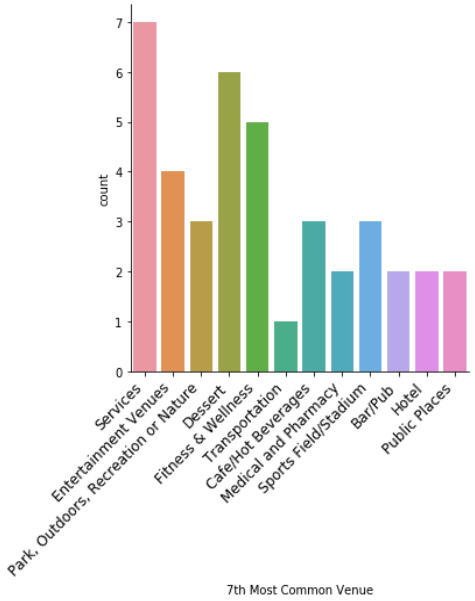


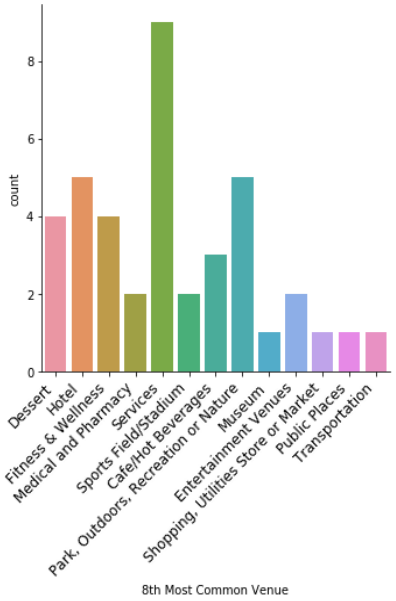


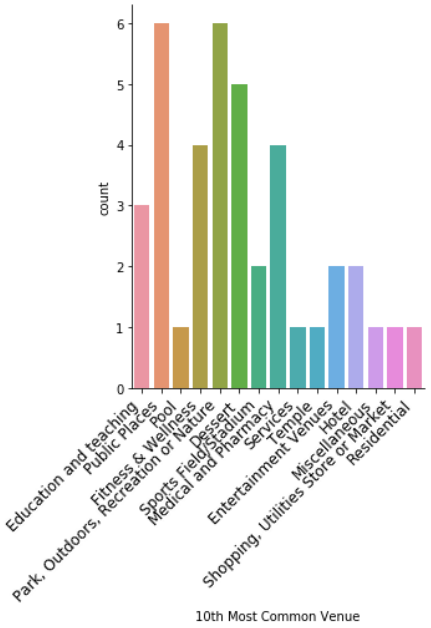












How are the five clusters in New York distinct from each other?

Cluster 1 (NY)

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

Key Points:

- 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 are located in Cluster 1
- 25% of all Public places in Cluster 1
- 41.4% of all Restaurants as well as 41.08% of all Bars/Pubs are situated in this cluster

How are the five clusters in New York distinct from each other?

Cluster 2 (NY)

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

Key Points:

- 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 fairly good numbers (although least in comparison to other clusters) of Restaurant, Shops, Utilities Stores, and Market in the neighborhoods in this Cluster.

How are the five clusters in New York distinct from each other?

Cluster 3 (NY)

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

Key Points:

- 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 fairly good amount of Restaurants, Cafes, Bars and Transportation facilities.
- Over 37% of all venues belonging to Fitness and Wellness category are situated here.
- No residential venue.

How are the five clusters in New York distinct from each other?

Cluster 4 (NY)

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

Key Points:

- 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.
- Very good availability of Transportation facilities.
- 44% of all venues in cluster 3 are Restaurants. A fairly high number of such restaurants are specialty restaurants serving Asian, Indian, Middle Eastern and Mexican cuisine.
- A fairly low number of Entertainment venues and Hotels.

How are the five clusters in New York distinct from each other?

Cluster 5 (NY)

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

Key Points:

52% of all Entertainment Venue.

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 young crowd.

Over 45% of all Residential venues lie in this cluster. It is also backed by good availability transportation facilities.

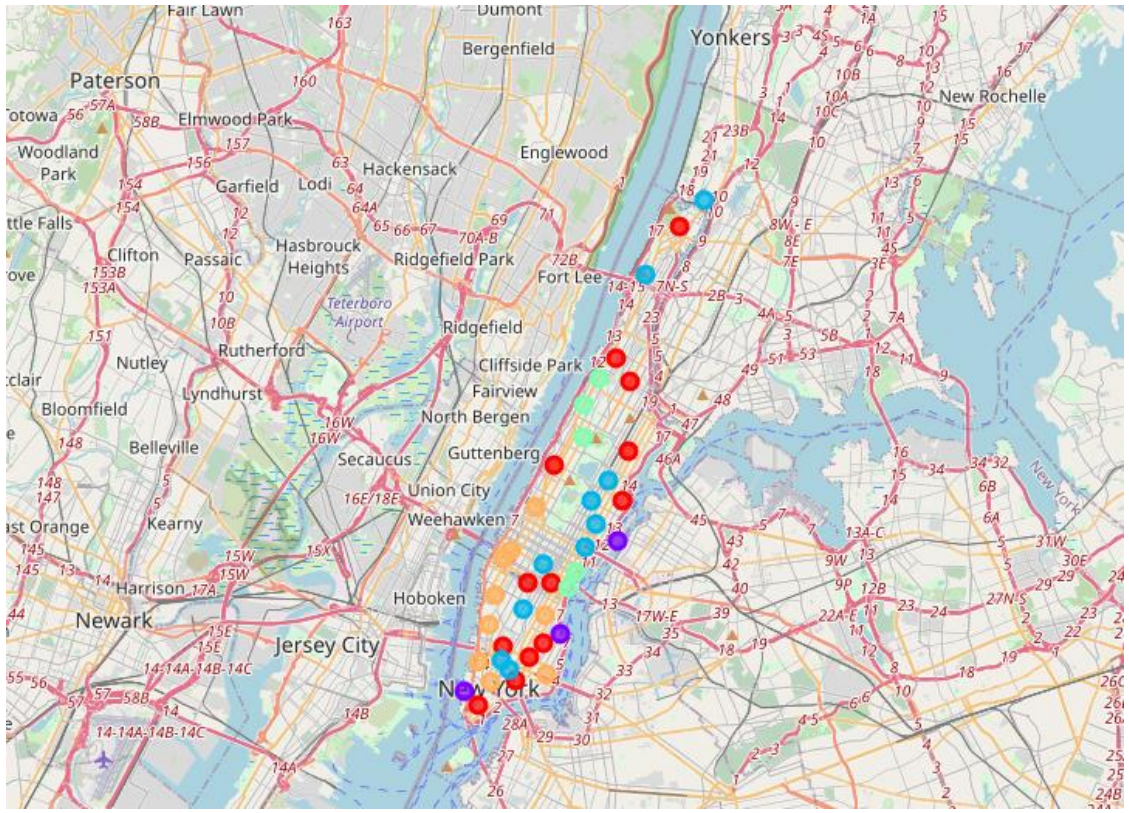
30% of all Educational and teaching venues lie in this cluster.

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 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 calm and simple life. Suggested for families with young children, elderly people 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.



Machine Learning Algorithms

K means Clustering:

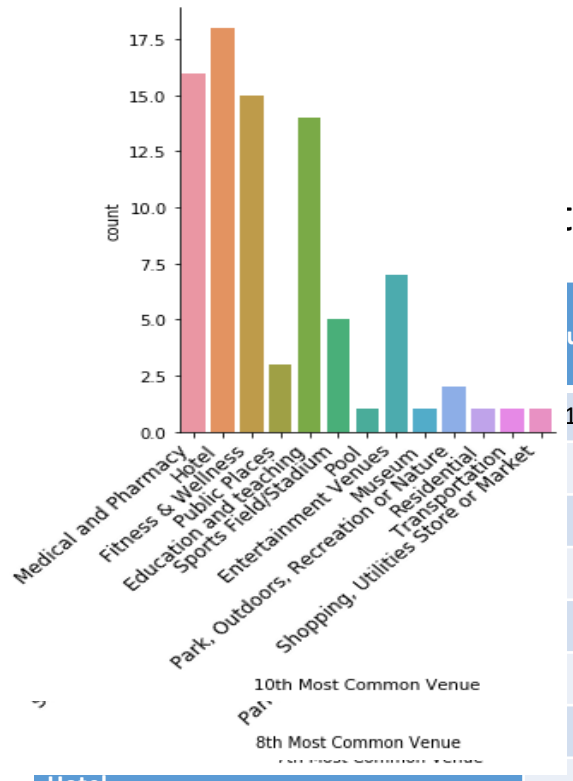
K means clustering algorithm was used to segment the neighborhoods in five clusters based on the top ten most common venue categories in each neighborhood. The following table gives the number of neighborhoods assigned to each cluster label (marked 0 to 4).

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

K nearest neighbors:

In order to classify a data point to a certain class, this algorithm takes into consideration its k nearest neighbor data points. 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. The optimal value of nearest neighbors was found to be 4.

Knn algorithm was trained on 80% of the dataset, while the rest was used for testing. An accuracy of 50% was achieved on the test set.



:er 1 (Mumbai)

	unt	Percentage in cluster	Overall Percentage	Percentage per neighborhood
	168	43.75	28.66894198	1.303133726
	51	13.28125	36.17021277	1.64410058
	39	10.15625	28.88888889	1.313131313
	32	8.333333333	22.37762238	1.017164654
	29	7.552083333	27.10280374	1.231945624
	21	5.46875	28.37837838	1.28992629
	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 (Mumbai)

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 (Mumbai)

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 (Mumbai)

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

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 wants 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 the cluster 2 in New York. While the neighborhoods of New York had more open spaces (such as parks, natural venues, sports fields) and less restaurants and bars, the case with Mumbai is completely 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, there are only 2 neighborhoods matching such 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 travelling 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 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 immigrants and poor section of the society, or well off people living in closed societies. The suitability of this cluster for two extremely different type of people may be disappointing, although the current model cannot make such 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 the 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.

Conclusion

- The project 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 inadequate amount of data and limitations of using Foursquare in such a study, the report has successfully laid down critical observations and discussed some very important issues. Identifying different areas of Mumbai having similarities to New York will be helpful in planning infrastructure development and resource management of those areas to achieve their understated potential. It can also indicate towards the potential of certain neighborhoods that might not have been looked at previously. This study suggests a completely different way of planning a city's development.

- The key 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.