Comparing 2 cities of dreams: Mumbai and New York

Rikki Mohanty

1. Introduction

1.1 Background

Description & Discussion of the Background

I have chosen Mumbai and New York for this project. These cities are called city of dreams for their own charismatic feature.

Let's find about the cities which are part of this project

Mumbai is the second most populous city in India and the seventh most populous city in the world with a population of 19.98 million in 2018. Mumbai is the financial, commercial and entertainment capital of India. Whereas New York City (NYC) is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2). New York City has been described as the cultural, financial, and media capital of the world.

I have decided to use these cities, explore and compare the neighborhoods and find if these cities have any similarities. (As mentioned in the Capstone Project we can either choose a problem or idea.) This idea of mine will help regular people more and make their life easier and more comfortable in these two cities if they are visiting or migrating.

We will go through each step of this project and address them separately. I will first outline the initial data preparation and describe future steps to start the battle of neighborhoods Project.

1.2 Target Audience

What type of clients or a group of people/stakeholders would be interested in this project?

1. People who are visiting these cities can make the best of city experience; also find the similar places for comfort.
2. Business personnel who want to invest. This analysis will give them and an idea of where to invest.
3. People who are migrating to these cities will have better ideas where to settle down, which places have the right resource and others.

2. Data Description

2.1 Data acquisition and cleaning

1. I have taken the dataset of New York from the Wikipedia page and found their respective coordinates.
2. For Mumbai city: the data availability is infrequent and dispersed in many places, so I've manually scraped the list of neighborhoods from this Wikipedia page http://zipcodepincode.com/India/Maharashtra/Mumbai/Mumbai/index.html. For this, I've used requests and Beautifulsoup4 library to create a dataframe with coordinates and pin codes which was manually scrapped from web.
3. I have used the Foursquare API to explore the neighborhoods of both the cities and segmented them.
4. These venues are then clustered using k-means. Found the most common venues (MCV) and finally compared the (MCV) of both cities to look for similarities.

Note: The Wikipedia page which is used here doesn't have all the pin codes of Mumbai. (Cannot find all the data's in one website)

3. Methodology

3.1 Programming Section (Initial Processing: Scraping from the web and getting the coordinates)

1. New York: Importing the libraries and getting the coordinates and refining the data

I have taken the information of New York and its coordinates from Coursera Capstone project

```
neighborhoods.head() # Displaying 5 rows
```

]:

	Neighborhood_NY	Latitude_NY	Longitude_NY
0	Wakefield	40.894705	-73.847201
1	Co-op City	40.874294	-73.829939
2	Eastchester	40.887556	-73.827806
3	Fieldston	40.895437	-73.905643
4	Riverdale	40.890834	-73.912585

2. Mumbai: Importing the libraries and getting the coordinates and refining the data

The Wikipedia page of Mumbai pin codes contains the table of 204 neighborhoods of Mumbai. I have used Beautifulsoup4 and pandas library to create the initial data-frame. Even though not complete but it gives us quite a detailed picture of the corresponding neighborhoods, as later on I have considered top most venues. After this initial preparation, I moved on to the next step to obtain coordinates manually because when I tried to use geopy library for these pin codes, it didn't work.

	Neighborhood_Mum	Latitude	Longitude
0	A I Staff Colony,Santacruz P&t Colony	19.0797	72.8679
1	Agripada,Chinchpokli,Haines Road,Jacob Circle	18.9810	72.8268
2	Airport (Mumbai),International Airport,Sahar P...	19.0929	72.8654
3	Ambewadi (Mumbai),Charni Road,Chaupati,Girgaon...	18.9580	72.8214
4	Andheri,Azad Nagar (Mumbai)	19.1121	72.8611

3.2 Visualization of the Cities

I used python folium library to visualize geographic details of New York and Mumbai, and its neighborhoods and I created a map of New York and Mumbai with Neighborhoods superimposed on top. I used latitude and longitude values to get the visual as below:

3.2.1 Visualization of New York

```
map_newyork
```

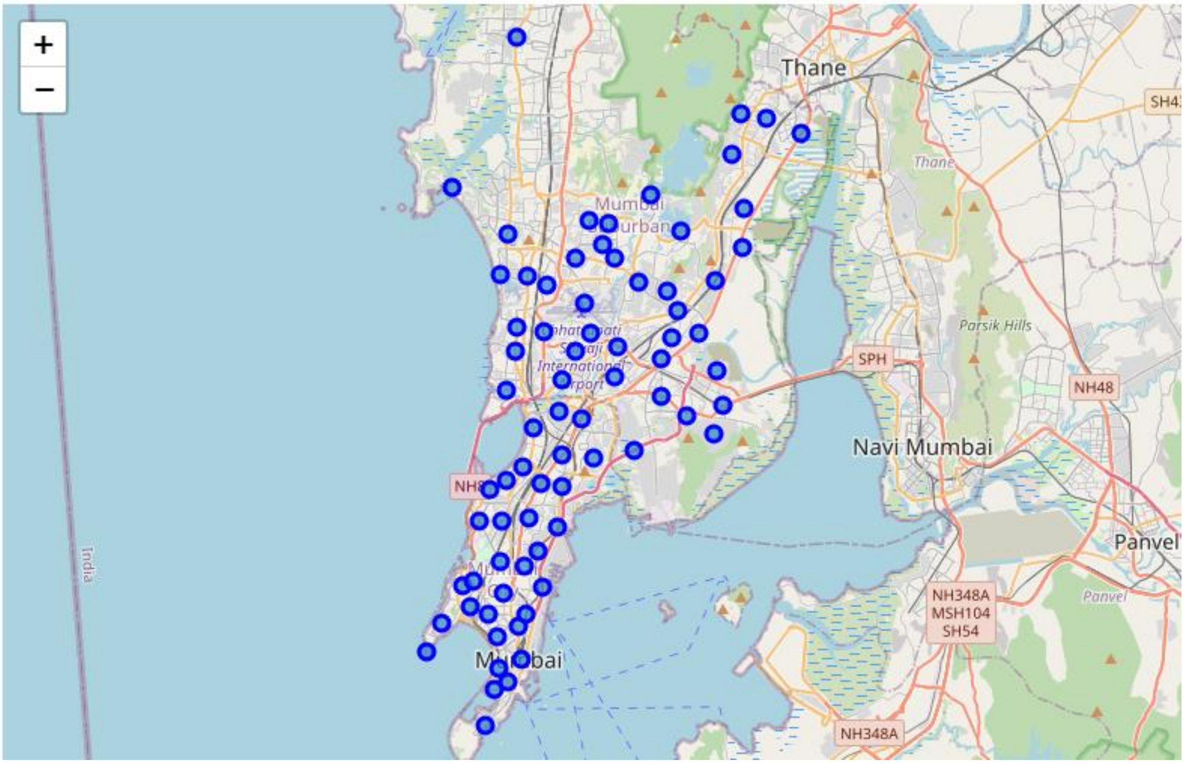
The geograpical coordinate of New York City are 40.7127281, -74.0060152.

Out[16]:

3.2.2 Visualization of Mumbai

```
map_mumbai
```

[17]:



3.3 Define Foursquare Credentials and Version

I have Used Foursquare login and got all the venues, categories and others.

```
CLIENT_ID = '' # your Foursquare ID
CLIENT_SECRET = '' # your Foursquare Secret
VERSION = '20200606' # Foursquare API version

print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
```

```
Your credentails:
CLIENT_ID:
CLIENT_SECRET:
```

3.4 Now, let's get the top 100 venues that are in each Neighborhood within a radius of 500 meters

I utilized the Foursquare API to explore the neighborhoods and segment them. I designed the limit as 100 venues and the radius 500 meter for each neighborhood from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

3.5 Explore Neighborhoods in New York and Mumbai

I have created a function to repeat the same process to all the neighborhoods for New York and Mumbai. And also a function to create a new dataframe for New York and Mumbai called *New_York_venues* and *Mumbai_venues* respectively.

The result doesn't mean that inquiry has run all the possible results in neighborhoods. Actually, it depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each neighborhood. We can increase the possibilities with Neighborhood information with more Latitude and Longitude information.

3.5.1 New York

```
print(New_York_venues.shape)
New_York_venues.head()
```

```
(9972, 7)
```

5]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

3.5.2 Mumbai

```
print(Mumbai_venues.shape)
Mumbai_venues.head()
```

```
(1048, 7)
```

5]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	A I Staff Colony,Santacruz P&t Colony	19.0797	72.8679	King Chilly	19.078382	72.866350	Chinese Restaurant
1	A I Staff Colony,Santacruz P&t Colony	19.0797	72.8679	The Camp	19.077917	72.865643	Asian Restaurant
2	A I Staff Colony,Santacruz P&t Colony	19.0797	72.8679	Tip Top Kebab Corner	19.078340	72.866356	Snack Place
3	A I Staff Colony,Santacruz P&t Colony	19.0797	72.8679	Nilesh Dry Fruits	19.077578	72.864080	Food & Drink Shop
4	Agripada,Chinchpokli,Haines Road,Jacob Circle	18.9810	72.8268	Cafe Coffee Day	18.981954	72.823608	Coffee Shop

3.6 Analyze Each Neighborhood

In summary of this 432 unique categories were returned by Foursquare for New York and for Mumbai it is 169 unique categories, then I created a table which shows list of top 10 venue category for each neighborhood in below table.

3.6.1 New York

	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	Allerton	Pizza Place	Chinese Restaurant	Supermarket	Deli / Bodega	Donut Shop	Spanish Restaurant	Fried Chicken Joint	Bus Station	Gas Station	Fast Food Restaurant
1	Annadale	Pizza Place	Restaurant	Train Station	Diner	Liquor Store	Sports Bar	Bakery	Ethiopian Restaurant	Event Service	Event Space
2	Arden Heights	Pizza Place	Pharmacy	Deli / Bodega	Bus Stop	Coffee Shop	Women's Store	Entertainment Service	Ethiopian Restaurant	Event Service	Event Space
3	Arlington	American Restaurant	Deli / Bodega	Grocery Store	Bus Stop	Coffee Shop	Women's Store	Field	Ethiopian Restaurant	Event Service	Event Space
4	Arrochar	Italian Restaurant	Pizza Place	Deli / Bodega	Bus Stop	Athletics & Sports	Pharmacy	Liquor Store	Bagel Shop	Supermarket	Middle Eastern Restaurant

3.6.2 Mumbai

	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	A I Staff Colony,Santacruz P&t Colony	Snack Place	Chinese Restaurant	Food & Drink Shop	Asian Restaurant	Food Court	Food	Flower Shop	Flea Market	Field	Fast Food Restaurant
1	Agripada,Chinchpokli,Haines Road,Jacob Circle	Indian Restaurant	Racetrack	Gym	Coffee Shop	Restaurant	Bakery	Electronics Store	Food	Flower Shop	Flea Market
2	Airport (Mumbai),International Airport,Sahar P...	Airport	Airport Lounge	Coffee Shop	Jewelry Store	Bakery	Zoo	Fast Food Restaurant	Food Truck	Food Court	Food & Drink Shop
3	Ambewadi (Mumbai),Charni Road,Chaupati,Girgaon...	Indian Restaurant	Snack Place	Vegetarian / Vegan Restaurant	Coffee Shop	Fast Food Restaurant	Electronics Store	Farmers Market	Food Court	Food & Drink Shop	Food
4	Andheri East,Nagardas Road	Diner	Chinese Restaurant	Hotel	Luggage Store	Asian Restaurant	Pub	Restaurant	Zoo	Food	Flower Shop

3.8. Cluster Neighborhoods and examining Neighborhoods

We have some common venue categories in neighborhoods. In this reason I used unsupervised learning K-means algorithm to cluster the neighborhoods. K-Means algorithm is one of the most common cluster methods of unsupervised learning.

First, I will run K-Means to cluster the neighborhoods into 5 clusters. And finally compare the cities for any similarity.

3.8.1. New York

	Neighborhood_NY	Latitude_NY	Longitude_NY	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	
0	Wakefield	40.894705	-73.847201	0.0	Pharmacy	Deli / Bodega	Ice Cream Shop	Laundromat	Donut Shop	Gas Station	Dessert Shop	Sandwich Place	
1	Co-op City	40.874294	-73.829939	0.0	Fried Chicken Joint	Restaurant	Park	Bagel Shop	Grocery Store	Pharmacy	Fast Food Restaurant	Mattress Store	
2	Eastchester	40.887556	-73.827806	0.0	Caribbean Restaurant	Bus Stop	Deli / Bodega	Diner	Donut Shop	Pizza Place	Platform	Bus Station	
3	Fieldston	40.895437	-73.905643	3.0	Plaza	Cosmetics Shop	High School	Bus Station	Field	Entertainment Service	Ethiopian Restaurant	Event Service	E
4	Riverdale	40.890834	-73.912585	4.0	Park	Bus Station	Food Truck	Plaza	Baseball Field	Medical Supply Store	Gym	Home Service	
5	Kingsbridge	40.881687	-73.902818	0.0	Pizza Place	Bar	Sandwich Place	Supermarket	Mexican Restaurant	Latin American	Spanish Restaurant	Donut Shop	

3.8.2. Mumbai

	Neighborhood_Mum	Latitude	Longitude	Cluster Numbers	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	
0	A I Staff Colony,Santacruz P&t Colony	19.0797	72.8679	1.0	Snack Place	Chinese Restaurant	Food & Drink Shop	Asian Restaurant	Food Court	Food	Flower Shop	Flea Market	
1	Agripada,Chinchpokli,Haines Road,Jacob Circle	18.9810	72.8268	2.0	Indian Restaurant	Racetrack	Gym	Coffee Shop	Restaurant	Bakery	Electronics Store	Food	
2	Airport (Mumbai),International Airport,Sahar P...	19.0929	72.8654	1.0	Airport	Airport Lounge	Coffee Shop	Jewelry Store	Bakery	Zoo	Fast Food Restaurant	Food Truck	F
3	Ambewadi (Mumbai),Charni Road,Chaupati,Girgaon...	18.9580	72.8214	2.0	Indian Restaurant	Snack Place	Vegetarian / Vegan Restaurant	Coffee Shop	Fast Food Restaurant	Electronics Store	Farmers Market	Food Court	D
4	Andheri,Azad Nagar (Mumbai)	19.1121	72.8611	1.0	Hotel	Fast Food Restaurant	Restaurant	Multiplex	Pizza Place	Café	Asian Restaurant	Cocktail Bar	

3.9 Comparing the cities

We are using 1st Most Common Venue as reference in each cluster. Thus, we are creating a Master Dataframe which may help us to find proper labels for each cluster.

```
df_Similar_0 = New_York_merged.loc[New_York_merged['Cluster Labels'] == 0, New_York_merged.columns[0:5]].merge(Mumbai_merged.loc[Mumbai_merged['Cluste
df_Similar_0
```

51]:

	Neighborhood_NY	Latitude_NY	Longitude_NY	Cluster Labels	1st Most Common Venue	Neighborhood_Mum	Latitude	Longitude	Cluster Numbers

```
df_Similar_1 = New_York_merged.loc[New_York_merged['Cluster Labels'] == 1, New_York_merged.columns[0:5]].merge(Mumbai_merged.loc[Mumbai_merged['Cluste
df_Similar_1
```

52]:

	Neighborhood_NY	Latitude_NY	Longitude_NY	Cluster Labels	1st Most Common Venue	Neighborhood_Mum	Latitude	Longitude	Cluster Numbers
0	Van Nest	40.843608	-73.866299	1.0	Pizza Place	Chembur,Sindhi Society	19.0521	72.9007	1.0
1	Van Nest	40.843608	-73.866299	1.0	Pizza Place	Kandivali RS	19.2072	72.8348	1.0
2	Van Nest	40.843608	-73.866299	1.0	Pizza Place	Kurla,Kurla North,Netajinagar	19.0741	72.8804	1.0
3	Arden Heights	40.549286	-74.185887	1.0	Pizza Place	Chembur,Sindhi Society	19.0521	72.9007	1.0
4	Arden Heights	40.549286	-74.185887	1.0	Pizza Place	Kandivali RS	19.2072	72.8348	1.0
5	Arden Heights	40.549286	-74.185887	1.0	Pizza Place	Kurla,Kurla North,Netajinagar	19.0741	72.8804	1.0
6	Edgewater Park	40.821986	-73.813885	1.0	Italian Restaurant	Asvini,Colaba,Colaba Bazar,Holiday Camp,V W T C	18.9104	72.8198	1.0
7	Edgewater Park	40.821986	-73.813885	1.0	Italian Restaurant	Nariman Point,New Yogakshema	18.9256	72.8242	1.0

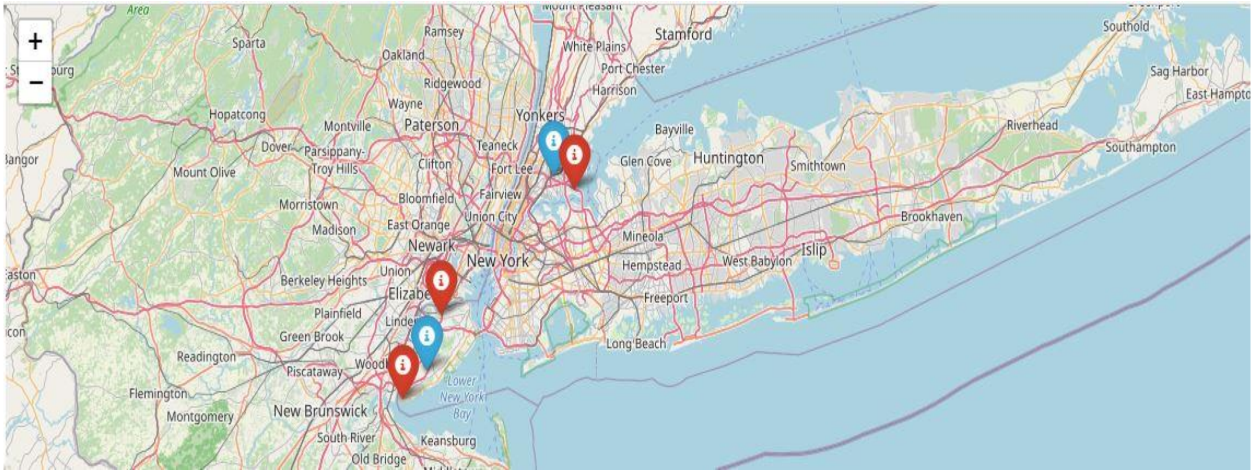
4. Result Section

I have merged those new variables with related cluster information in our main master table. The map shows the Neighborhoods with respect to 1st Most Common Venue column.

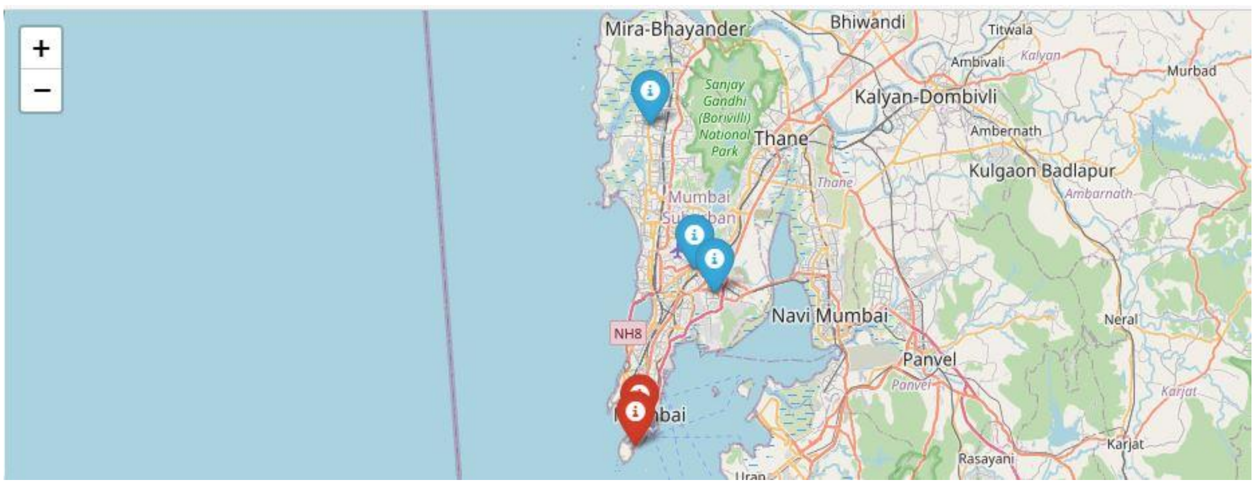
	Neighborhood_NY	Latitude_NY	Longitude_NY	Cluster Labels	1st Most Common Venue	Neighborhood_Mum	Latitude	Longitude	Cluster Numbers
0	Van Nest	40.843608	-73.866299	1.0	Pizza Place	Chembur,Sindhi Society	19.0521	72.9007	1.0
1	Van Nest	40.843608	-73.866299	1.0	Pizza Place	Kandivali RS	19.2072	72.8348	1.0
2	Van Nest	40.843608	-73.866299	1.0	Pizza Place	Kurla,Kurla North,Netajinagar	19.0741	72.8804	1.0
3	Arden Heights	40.549286	-74.185887	1.0	Pizza Place	Chembur,Sindhi Society	19.0521	72.9007	1.0
4	Arden Heights	40.549286	-74.185887	1.0	Pizza Place	Kandivali RS	19.2072	72.8348	1.0
5	Arden Heights	40.549286	-74.185887	1.0	Pizza Place	Kurla,Kurla North,Netajinagar	19.0741	72.8804	1.0
6	Edgewater Park	40.821986	-73.813885	1.0	Italian Restaurant	Asvini,Colaba,Colaba Bazar,Holiday Camp,V W T C	18.9104	72.8198	1.0
7	Edgewater Park	40.821986	-73.813885	1.0	Italian Restaurant	Nariman Point,New Yogakshema	18.9256	72.8242	1.0
8	Mariner's Harbor	40.632546	-74.150085	1.0	Italian Restaurant	Asvini,Colaba,Colaba Bazar,Holiday Camp,V W T C	18.9104	72.8198	1.0
9	Mariner's Harbor	40.632546	-74.150085	1.0	Italian Restaurant	Nariman Point,New Yogakshema	18.9256	72.8242	1.0
10	Tottenville	40.505334	-74.246569	1.0	Italian Restaurant	Asvini,Colaba,Colaba Bazar,Holiday Camp,V W T C	18.9104	72.8198	1.0
11	Tottenville	40.505334	-74.246569	1.0	Italian Restaurant	Nariman Point,New Yogakshema	18.9256	72.8242	1.0

You can now see similar Neighborhoods of Neighborhood_NY, Neighborhood_Mum columns with their Latitude_NY, Longitude_NY and Latitude, Longitude respectively. They are color coordinated with respect to 1st Most Common Venue.

1. New York



2. Mumbai



So this is our result shown below:

1. Van Nest, Arden Heights in New York and (Chembur, Sindhi Society), Kandivali RS, (Kurla, Kurla North, Netajinagar) in Mumbai are similar with the most common venue is **Pizza Restaurant** for both the neighborhoods.
2. Edgewater Park, Mariner's Harbor, Tottenville in New York and (Asvini, Colaba, Colaba Bazar, Holiday Camp, V W T C), (Nariman Point, New Yogakshema) in Mumbai are similar with the most common venue is **Italian Restaurant** for both the neighborhoods.

Note: The latitude and longitude for Mumbai is with respect to Postal Codes, not the Neighborhood_Mum.

5. Discussion Section

In this project we found the similar places with respect to 1st Most Common Venue and found 5 places in New York and 5 places in Mumbai are similar to each other. We could have added more Common places as base and we could have used different methods for clustering and classification studies. However, due to the complexities I have used these methods for simplification purposes.

I have used the K-means algorithm as part of this clustering study. I set the optimum k value to 5. I have used the Capstone project Json file for New York coordinates and for Mumbai I have manually scraped the data from web. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled.

I ended the study by visualizing the data and clustering information on the New York and Mumbai map. In future studies, we can add more common areas as base for more detailed visualization and similarities.

6. Conclusion Section

I would like to conclude by saying how different the culture maybe in different cities we are connected in some way or the other. This project gives us a small idea in visual form. We can create different projects with different cities so make it easier for people to visit these cities without any discomfort and enjoy the neighborhoods.