Battle of the Neighbourhoods

Part 1

Business Problem

The problem I have chosen to analyse involves the decision regarding where to open a gym in Manhattan, New York City , USA. There are several factors which will play a role in determining this optimal location. The gym needs to be centrally located and needs to be easily accessible from all parts of the city.

The gym should also have complementary venues near it. For example, the presence of a spa nearby will encourage customers to purchase gym membership of said gym. The gym should also not exist in a place where several gyms already exist which would generate unnecessary competition. Clusters will be generated to assess each neighbourhood.

I will attempt to generate a list of neighbourhoods most suitable which will result in maximum revenue for the stakeholder wishing to set up the gym.

Data

Based on the factors I discussed above, the following data will be required:

- The distance of each neighbourhood from the centre of the city.
- The venues nearby which are complementary to the gym.
- The number of gyms that exist in each neighbourhood.

The neighbourhoods have been identified from the source: https://cocl.us/new_york_dataset. All the features will be extracted from this data source. The other data will be extracted using the Foursquare API. The location of each neighbourhood will be obtained from the GeoPy GeoCoder package in Python.

Methodology

The steps involved in the conduction of this project and the efficient detection of suitable neighbourhoods in which to establish a gym:

- The first step involves obtaining the data in the appropriate format

- I obtained the **latitude and longitude** of all these neighbourhoods in New York City
- I then focused my attention to Manhattan Borough and plotted its neighbourhoods using Folium.
- I fixed the centre of the city as **Central Park** to maintain a reference point
- The **distance** of each neighbourhood from the centre was calculated using GeoPy.
- Only those neighbourhoods were shortlisted which were **less than 5 Kms from the centre**. This gave me **21 neighbourhoods** to work with.
- Obtained the venues in each neighbourhood's vicinity which gave me approximately **1500 locations**. This was done by sending **get requests to Foursquare API**.
- I obtained the **top 5 most common venues** categories for each neighbourhood.
- I then obtained a data frame which shows the number of gyms within 500m each neighbourhood.
- **KNN algorithm was applied to this dataset** to sort the neighbourhoods into **5 clusters**.
- Each cluster was analysed to assess its characteristics
- I then plotted a **bar chart** using seaborn to showcase the **potential of each neighbourhood** when it comes to establishing a Gym.

Data

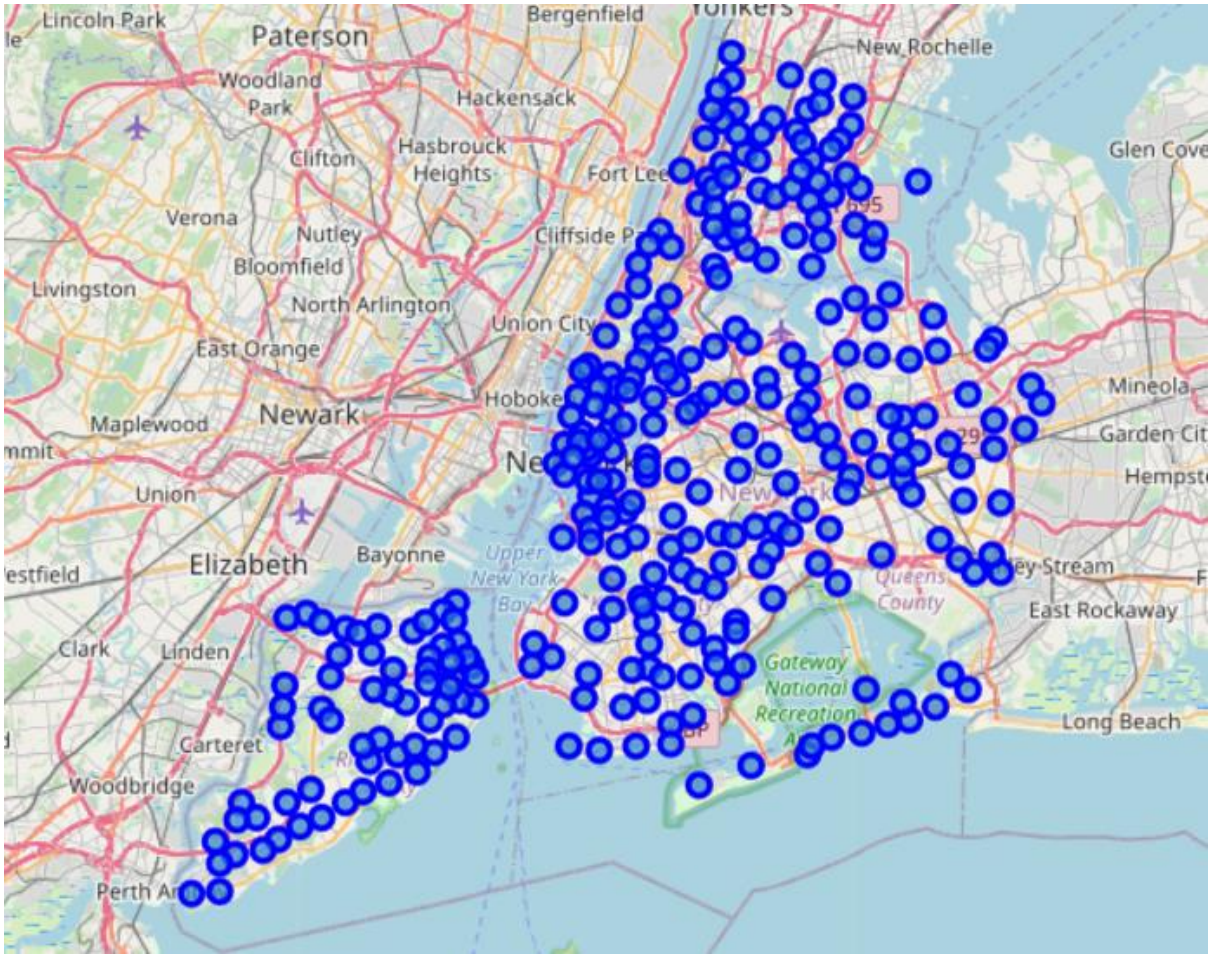
The following data frame was obtained from the JSON file :

[10]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585
...	...	...	...	...
301	Manhattan	Hudson Yards	40.756658	-74.000111
302	Queens	Hammels	40.587338	-73.805530
303	Queens	Bayswater	40.611322	-73.765968
304	Queens	Queensbridge	40.756091	-73.945631
305	Staten Island	Fox Hills	40.617311	-74.081740

306 rows × 4 columns

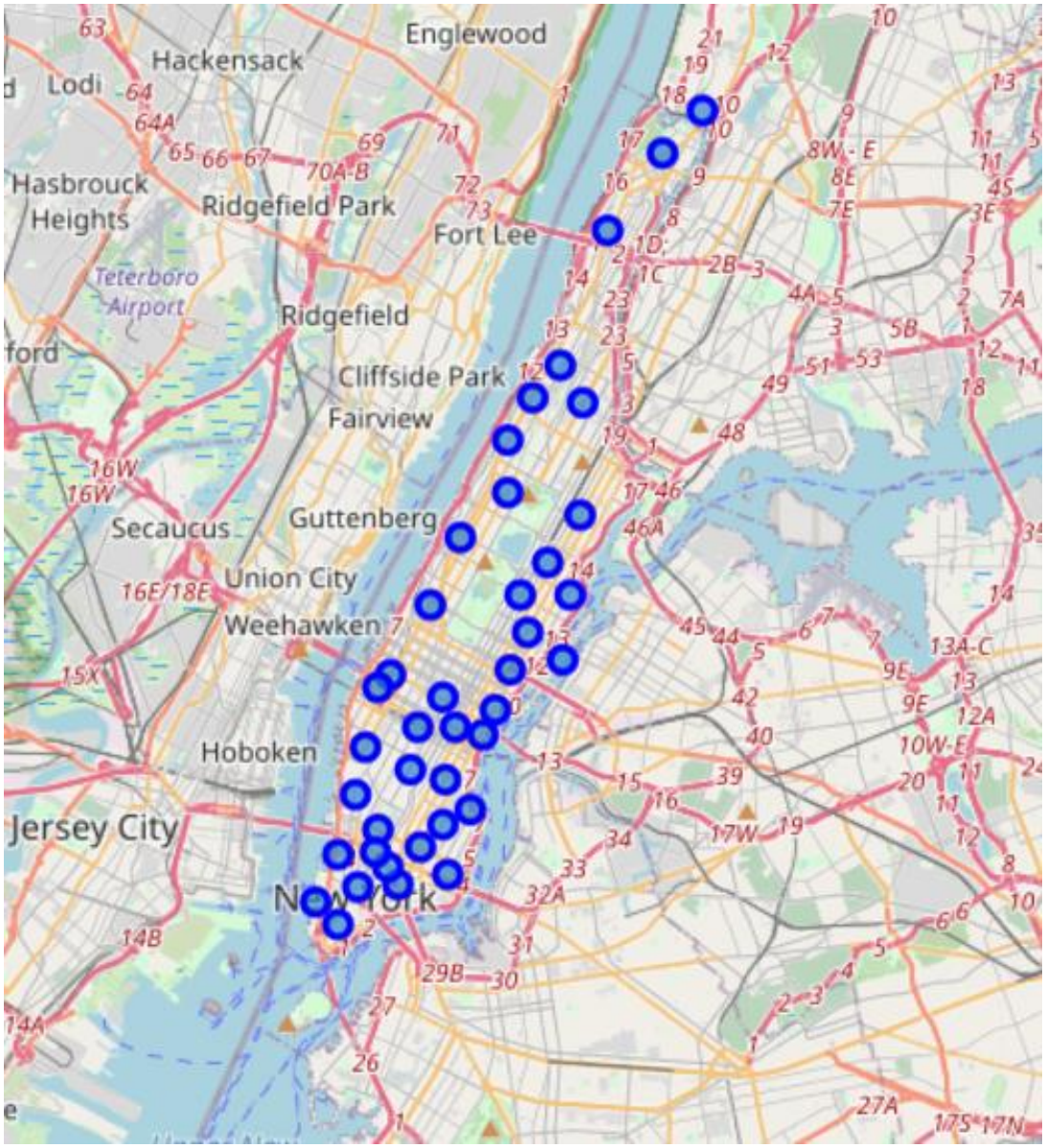
A map of New York with neighbourhoods superimposed on top



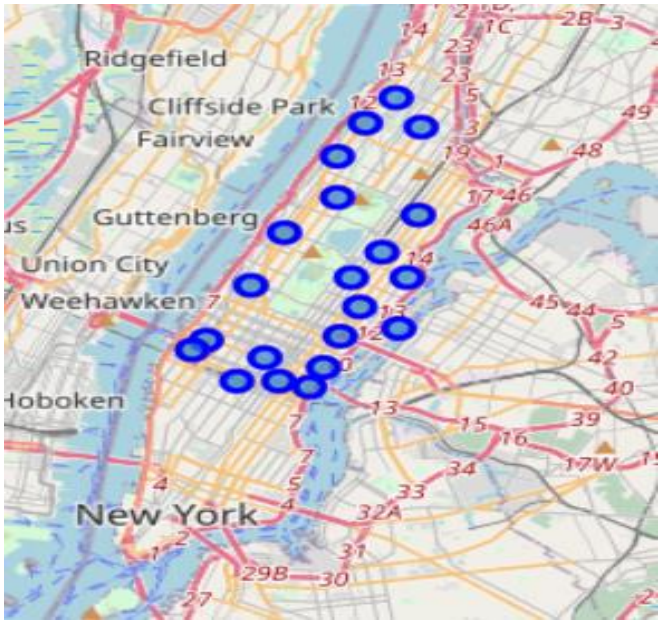
Since this represents the entire NYC region, there are several Boroughs included in this. I need Manhattan only which led me to a smaller data set. The first five lines of this data frame are shown below:

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Let us visualize only the neighbourhoods in Manhattan



We can see that there exist points so far away from the centre of the city, it makes no sense to include them in our analysis. Therefore, I removed the points from the data frame that were lying more than 5 kilometres from the centre of the city. The 21 points remaining generated a visualisation like this:



For obvious reasons, this represents a far better group of data suitable for the analysis we have planned.

Use of FOURSQUARE API

The resulting dataset can now be subjected to further analysis. We now attempt to assess the venues in the vicinity of each neighbourhood. This leads us to the following table.

[38]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Carnegie Hill	Coffee Shop	Yoga Studio	Bakery	Gym / Fitness Center	Gym
1	Central Harlem	Cosmetics Shop	African Restaurant	Seafood Restaurant	Bar	French Restaurant
2	Clinton	Theater	Coffee Shop	Gym / Fitness Center	Wine Shop	Gym
3	East Harlem	Mexican Restaurant	Bakery	Thai Restaurant	Latin American Restaurant	Deli / Bodega
4	Hamilton Heights	Pizza Place	Café	Coffee Shop	Deli / Bodega	Mexican Restaurant
5	Hudson Yards	Hotel	American Restaurant	Gym / Fitness Center	Café	Italian Restaurant
6	Lenox Hill	Italian Restaurant	Pizza Place	Coffee Shop	Café	Cocktail Bar
7	Lincoln Square	Plaza	Café	Concert Hall	Gym / Fitness Center	Performing Arts Venue
8	Manhattan Valley	Yoga Studio	Bar	Pizza Place	Coffee Shop	Mexican Restaurant
9	Manhattanville	Seafood Restaurant	Coffee Shop	Italian Restaurant	Park	Mexican Restaurant
10	Midtown	Coffee Shop	Hotel	Clothing Store	Theater	Spa
11	Midtown South	Korean Restaurant	Hotel	Japanese Restaurant	Burger Joint	Dessert Shop
12	Morningside Heights	Park	American Restaurant	Coffee Shop	Bookstore	Burger Joint
13	Murray Hill	Sandwich Place	Coffee Shop	Hotel	Pizza Place	Gym / Fitness Center
14	Roosevelt Island	Park	Plaza	Japanese Restaurant	Soccer Field	Farmers Market
15	Sutton Place	Italian Restaurant	Gym / Fitness Center	Park	Coffee Shop	Gym
16	Tudor City	Café	Park	Mexican Restaurant	Deli / Bodega	Pizza Place
17	Turtle Bay	Italian Restaurant	Coffee Shop	Park	Deli / Bodega	Wine Bar
18	Upper East Side	Italian Restaurant	Juice Bar	Bakery	Gym / Fitness Center	Exhibit
19	Upper West Side	Italian Restaurant	Bakery	Wine Bar	Coffee Shop	Mediterranean Restaurant
20	Yorkville	Coffee Shop	Italian Restaurant	Gym	Deli / Bodega	Bar

This will be useful in the elimination if neighbours which are not suitable for establishing a gym. They may not suggest the presence of athletic activities or do not have any complementary businesses present which may enhance gym revenues like a spa, park, or a yoga facility.

Next let us try to obtain the number of gyms in the vicinity of each neighbourhood. We used the API for this purpose and got the desired list. I then appended this list to the data frame. We now have a data set which shows the number of gyms near each neighbourhood. This gives us an idea of the competition in each locality.

The data frame is shown below:

0	Upper West Side	Manhattan	40.787658	-73.977059	0.793568	1.0
1	Upper East Side	Manhattan	40.775639	-73.960508	1.238088	0.0
2	Carnegie Hill	Manhattan	40.782683	-73.953256	1.296437	3.0
3	Manhattan Valley	Manhattan	40.797307	-73.964286	1.397942	0.0
4	Lincoln Square	Manhattan	40.773529	-73.985338	1.928930	2.0
5	Lenox Hill	Manhattan	40.768113	-73.958860	2.046453	3.0
6	Yorkville	Manhattan	40.775930	-73.947118	2.056120	5.0
7	East Harlem	Manhattan	40.792249	-73.944182	2.184156	1.0
8	Morningside Heights	Manhattan	40.808000	-73.963896	2.570835	0.0
9	Sutton Place	Manhattan	40.760280	-73.963556	2.783986	2.0
10	Roosevelt Island	Manhattan	40.762160	-73.949168	3.014885	1.0
11	Midtown	Manhattan	40.754691	-73.981669	3.559979	2.0
12	Manhattanville	Manhattan	40.816934	-73.957385	3.653879	0.0
13	Turtle Bay	Manhattan	40.752042	-73.967708	3.670355	0.0
14	Clinton	Manhattan	40.759101	-73.996119	3.721841	4.0
15	Central Harlem	Manhattan	40.815976	-73.943211	4.029970	1.0
16	Hudson Yards	Manhattan	40.756658	-74.000111	4.146005	2.0
17	Murray Hill	Manhattan	40.748303	-73.978332	4.172419	1.0
18	Tudor City	Manhattan	40.746917	-73.971219	4.246387	2.0
19	Midtown South	Manhattan	40.748510	-73.988713	4.413320	0.0
20	Hamilton Heights	Manhattan	40.823604	-73.949688	4.555719	0.0

The 6th column represents the distance between the neighbourhood and Central Park while the last column tells us the number of gyms in each neighbourhood.

KNN Clustering

We can now apply KNN algorithm to form clusters which will help us sort the venues into groups. I have chosen 5 groups for this purpose. The algorithm was run, and cluster labels were obtained. This list of cluster labels was attached to the data set to get the final data set. The clusters were then visualized using folium.

They can be seen in the map below:

The unfilled marker represents Central Park. The details of each cluster are given below:

Cluster 1- It represents the neighbourhoods that are at a greater distance from the centre of Manhattan. Each neighbourhood as at least 1 gym. I would not recommend setting up a gym here as it is far from Central Park. These neighbourhoods are marked in **red**. The members of this cluster are:

	Cluster Labels	Neighborhood	Borough	Latitude	Longitude	Distance from Centre	Gym
10	0	Roosevelt Island	Manhattan	40.762160	-73.949168	3.014885	1.0
11	0	Midtown	Manhattan	40.754691	-73.981669	3.559979	2.0
15	0	Central Harlem	Manhattan	40.815976	-73.943211	4.029970	1.0
16	0	Hudson Yards	Manhattan	40.756658	-74.000111	4.146005	2.0
17	0	Murray Hill	Manhattan	40.748303	-73.978332	4.172419	1.0
18	0	Tudor City	Manhattan	40.746917	-73.971219	4.246387	2.0

Cluster 2 - These are neighbourhoods with plenty of gyms already. There will be fierce competition here and that may impact revenue. These points are marked in **purple**. The members of the cluster are:

	Cluster Labels	Neighborhood	Borough	Latitude	Longitude	Distance from Centre	Gym
6	1	Yorkville	Manhattan	40.775930	-73.947118	2.056120	5.0
14	1	Clinton	Manhattan	40.759101	-73.996119	3.721841	4.0

Cluster 3 - These are the points closest to Central Park and are greatly untapped. These neighbourhoods offer great potential. I would recommend these neighbourhoods. However, one must keep in mind that these neighbourhoods are expensive and there will be significant overhead costs to setting up a gym here. These points are marked in light blue. They include:

	Cluster Labels	Neighborhood	Borough	Latitude	Longitude	Distance from Centre	Gym
0	2	Upper West Side	Manhattan	40.787658	-73.977059	0.793568	1.0
1	2	Upper East Side	Manhattan	40.775639	-73.960508	1.238088	0.0
3	2	Manhattan Valley	Manhattan	40.797307	-73.964286	1.397942	0.0
7	2	East Harlem	Manhattan	40.792249	-73.944182	2.184156	1.0
8	2	Morningside Heights	Manhattan	40.808000	-73.963896	2.570835	0.0

Cluster 4 - These are points that are slightly further away from the centre. However, no gym exists in the neighbourhood of these vicinities so it may result in high revenue. The cost of setting up the gym might also be much lower than cluster 2. They are marked in light green. They include:

	Cluster Labels	Neighborhood	Borough	Latitude	Longitude	Distance from Centre	Gym
12	3	Manhattanville	Manhattan	40.816934	-73.957385	3.653879	0.0
13	3	Turtle Bay	Manhattan	40.752042	-73.967708	3.670355	0.0
19	3	Midtown South	Manhattan	40.748510	-73.988713	4.413320	0.0
20	3	Hamilton Heights	Manhattan	40.823604	-73.949688	4.555719	0.0

Cluster 5- This cluster shows points that are somewhat between the outskirts and centre of Manhattan. There exist gyms already which may offer some competition. However, that number is far lesser than that of cluster 1. They can also be promising as prices will be lower. These are shown in orange. They include:

	Cluster Labels	Neighborhood	Borough	Latitude	Longitude	Distance from Centre	Gym
2	4	Carnegie Hill	Manhattan	40.782683	-73.953256	1.296437	3.0
4	4	Lincoln Square	Manhattan	40.773529	-73.985338	1.928930	2.0
5	4	Lenox Hill	Manhattan	40.768113	-73.958860	2.046453	3.0
9	4	Sutton Place	Manhattan	40.760280	-73.963556	2.783986	2.0

Potential of each neighbourhood

The potential of each neighbourhood will be a function of its distance from the centre and the number of gyms in it. Since we want a lesser value for both these categories, we can say that the potential will be inversely proportional to both. To measure potential, I have used the formula below:

$$P = \frac{5}{(d+1)(n+1)}$$

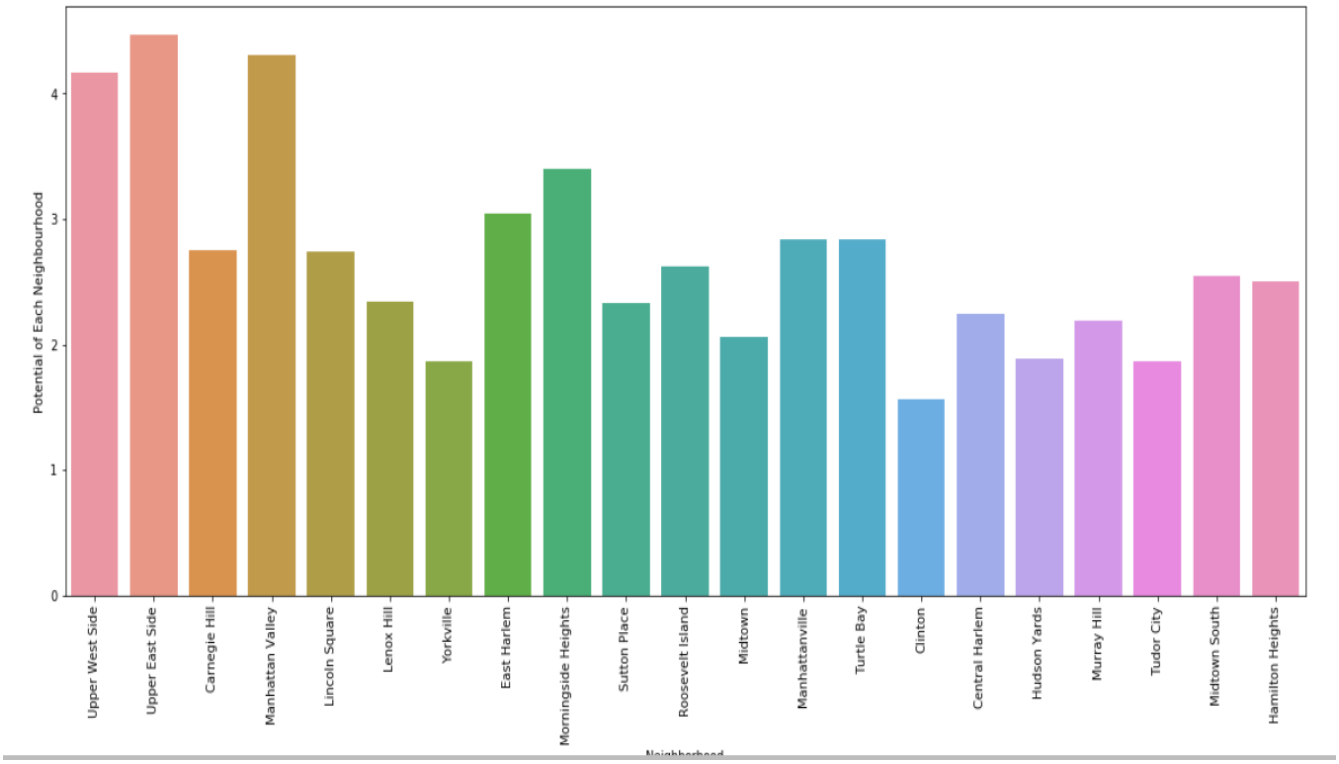
P= Potential of the neighbourhood

d= Normalised distance from the centre

n= Normalised count of the number of gyms in the neighbourhood.

We employ the use of normalised values to generate a fair bar chart without providing either variable with higher preference.

We then employed the use of the seaborn package to generate the bar chart.



Analysis

I believe this bar chart further proves the inferences we made from the clusters obtained. We eliminate all the neighbourhoods in cluster 0 and 1 for low potential. The highest potential is seen in neighbourhoods

- Upper West Side
- Upper East side
- Manhattan Valley
- Morningside Heights
- East Harlem

These points belong to cluster 1 and are the ideal spots to generate maximum revenue. They have almost no pre-existing gyms and are very close to the centre of the city. Upper West Side mainly comprises of restaurants and delis. **Nothing here suggests any athletic activities take place in this locality**. Thus, it can be eliminated. This is the issue with East Harlem as well.

We can divide the remaining prospect neighbourhoods into 2 categories. The first being neighbourhoods that are close to the centre and indicate athletic activities. Little to no gyms exist here. The neighbourhoods are:

- Upper East Side

- Morningside Heights
- Manhattan Valley

One must keep in mind that these areas are expensive. Upper east side and Manhattan Valley are the ideal places to set up the gym as Morningside Heights is a little further away. Personally, I would suggest **Manhattan Valley** as the ideal location as it has Yoga centres as well as a spa. These services nicely complement a gym and will enhance revenue.
The next category is that of gyms that are further away from the centre but will be cheaper to set up. The competition is also higher in this category. The neighbourhoods here are:

- Carnegie Hill
- Lincoln Square
- Manhattanville
- Sutton Place
- Turtle Bay

The bar chart suggests that **Manhattanville and Turtle Bay** both are good prospects. Carnegie Hill and Lincoln square are a tad bit overcrowded with gyms and Sutton Place is too far from the centre of Manhattan. There exist parks in both these regions which will help with customer traffic.

Conclusion

The purpose of this project was to identify a suitable location for stakeholders looking to set up a gym. Through the application of clustering and Foursquare API, I have identified the best neighbourhoods for this project. They are

- Manhattan Valley
- Upper East Side
- Manhattanville
- Turtle Bay

Manhattan Valley would require higher capital but would generate the maximum revenue. It is the closest to the centre and has complementary businesses. Upper East Side is a good prospect as well but will be even more expensive. Manhattanville and Turtle Bay are ideal prospects for stakeholders who would not wish to play the excessive charges of location. The competition in this region is higher as most gym owners would prefer this region.

The final decision would ultimately depend on the stakeholder and the budget he/she has. Personal preference will play a big role as well in determining the best location.