# Battle of the Neighborhoods


# Mukund Sureshkumar


- ## Introduction/ Business Problem

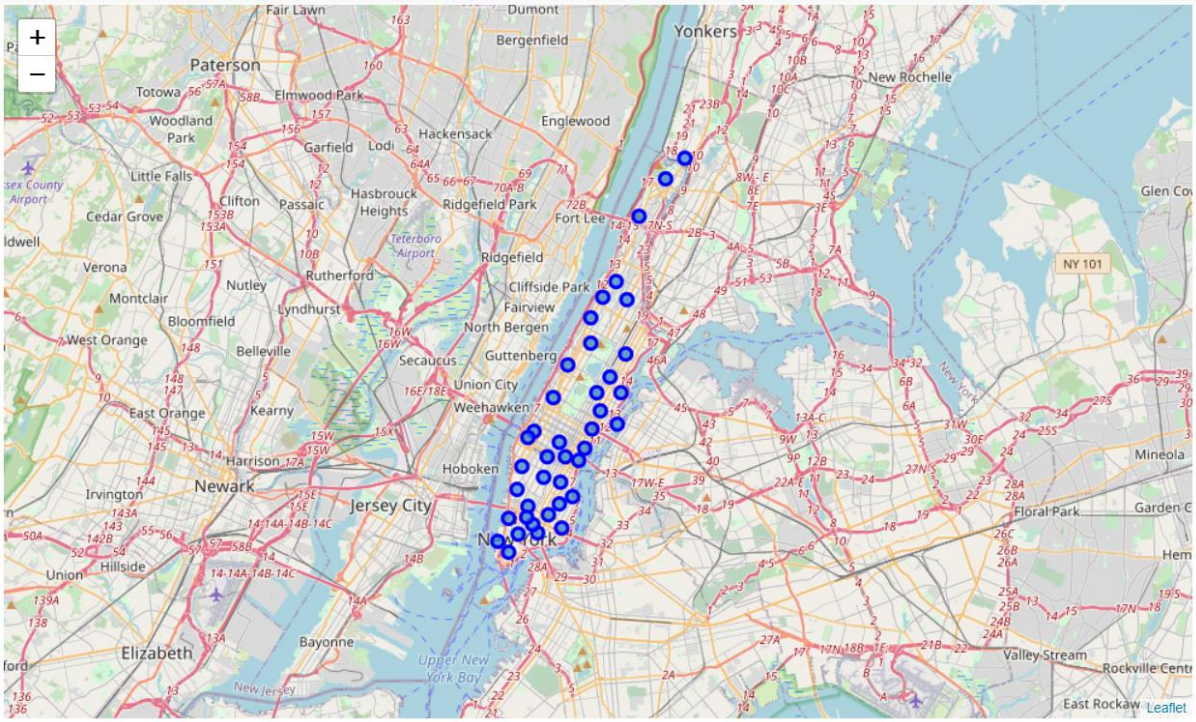
The right location to open a **supplement store** in Manhattan, NYC is crucial to the success of supplement store franchises looking to expand. Location analysis can play a crucial role in determining the right spot for opening up a franchise.

Many sports enthusiasts and gym goers from various ethnicities reside in New York. With a population north of 8 million, this is a prime city for business.

The **right location** to open up a supplement store would be a place that is frequently visited by sports enthusiasts and gym goers. The right spot would be one where the most number of such people are found. Additionally, we'd look for a spot that has no competition so we'll take other supplement stores into account as well.

- ## Data

  Neighborhood data can be found on this JSON file:
  ( https://geo.nyu.edu/catalog/nyu_2451_34572)



The venues in each neighborhood can be found using Foursquare's API.

We'll use the locations of the following venue categories to clearly define the right location to open a supplement store:

- Athletics & Sports Venues
- Baseball Fields
- Basketball Court
- Boxing Gyms
- Climbing Gyms
- Cycle Studios
- Gyms, Gym/Fitness Centers, Gym Pools
- Soccer Fields
- Sports clubs, Sporting Goods Shops
- Tennis court

All venue categories will be grouped by their neighborhood for further inspection.

Density of existing supplement shop densities have to be taken into consideration to avoid competition which would affect revenue.

| | Neighborhood | Supplement Shop | Athletics & Sports | Baseball Field | Basketball Court | Boxing Gym | Gym | Cycle Studio | Gym / Fitness Center | Gym Pool | Soccer Field | Sports Club | Sporting Goods Shop | Tennis Court |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Battery Park City | 0.0 | 0.015873 | 0.0 | 0.0 | 0.0 | 0.047619 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 1 | Carnegie Hill | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.034091 | 0.000000 | 0.034091 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 2 | Central Harlem | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.021739 | 0.021739 | 0.043478 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 3 | Chelsea | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.010000 | 0.020000 | 0.000000 | 0.0 | 0.0 | 0.00 | 0.0 | 0.0 |
| 4 | Chinatown | 0.0 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.01 | 0.0 | 0.0 |

- # Methodology

We'll analyze each neighborhood and take a **mean** of all sports facilities to quantify the **concentration** of each sports facility in each neighborhood. We have ignored all stadiums and other facilities which are used infrequently throughout the year and are reserved for tournaments.

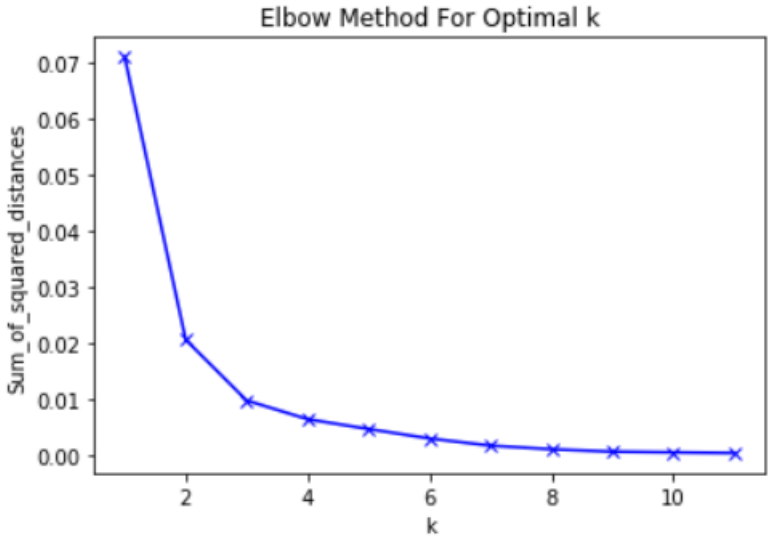
Further, we'll sum all column mean values to get a value on the **total** sports facility concentration in **each** neighborhood. This step will exclude supplement shops.

| Neighborhood | Supplement Shop | Athletics & Sports | Baseball Field | Basketball Court | Boxing Gym | Gym | Cycle Studio | Gym / Fitness Center | Gym Pool | Soccer Field | Sports Club | Sporting Goods Shop | Tennis Court | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flatiron | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.040000 | 0.020000 | 0.070000 | 0.000000 | 0.000000 | 0.01 | 0.020000 | 0.000000 | 0.160000 |
| Hudson Yards | 0.000000 | 0.017544 | 0.000000 | 0.000000 | 0.000000 | 0.035088 | 0.017544 | 0.052632 | 0.000000 | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.122807 |
| Marble Hill | 0.040000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.080000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.120000 |
| Roosevelt Island | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.038462 | 0.000000 | 0.038462 | 0.000000 | 0.038462 | 0.00 | 0.000000 | 0.000000 | 0.115385 |
| Stuyvesant Town | 0.000000 | 0.000000 | 0.055556 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.055556 | 0.000000 | 0.000000 | 0.00 | 0.000000 | 0.000000 | 0.111111 |

Afterwards, we'll use the **KMeans** Machine Learning algorithm to cluster neighborhoods based on the sports facilities and supplement shops present. We'll create a new Dataframe for this purpose. We'll drop all sports venue category columns and retain the neighborhood, sum, and supplement shop columns.

| | Neighborhood | Sum | Supplement Shop |
|---|---|---|---|
| 0 | Battery Park City | 0.063492 | 0.0 |
| 1 | Carnegie Hill | 0.068182 | 0.0 |
| 2 | Central Harlem | 0.086957 | 0.0 |
| 3 | Chelsea | 0.030000 | 0.0 |
| 4 | Chinatown | 0.010000 | 0.0 |

The **elbow method** will be deployed to calculate the optimal k value.



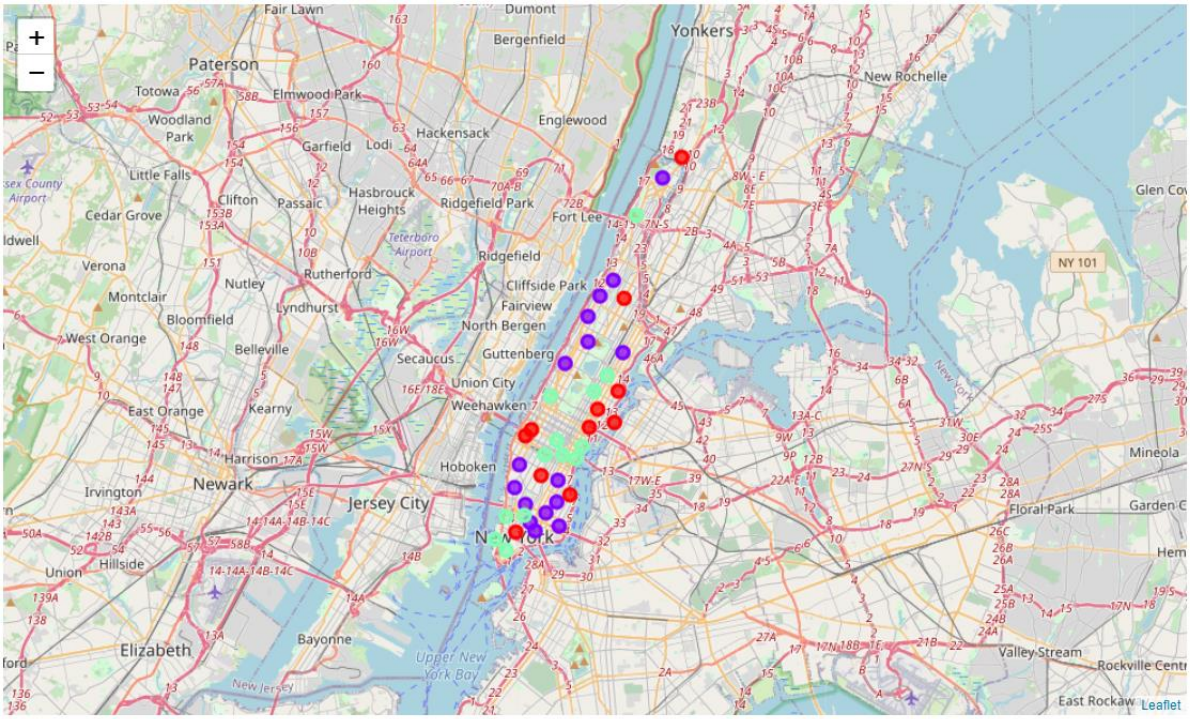

The optimal number of clusters for this dataset is 3.

We'll drop the neighborhood column and train the KMeans algorithm with this dataset and form 3 clusters.

We'll include the cluster labels in a Dataframe that includes the coordinates of each neighborhood.

| | Borough | Neighborhood | Latitude | Longitude | Cluster Labels | Sum | Supplement Shop |
|---|---|---|---|---|---|---|---|
| 0 | Manhattan | Marble Hill | 40.876551 | -73.910660 | 0 | 0.12000 | 0.040000 |
| 1 | Manhattan | Chinatown | 40.715618 | -73.994279 | 1 | 0.01000 | 0.000000 |
| 2 | Manhattan | Washington Heights | 40.851903 | -73.936900 | 2 | 0.05814 | 0.011628 |
| 3 | Manhattan | Inwood | 40.867684 | -73.921210 | 1 | 0.00000 | 0.000000 |
| 4 | Manhattan | Hamilton Heights | 40.823604 | -73.949688 | 1 | 0.00000 | 0.000000 |

We'll use this Dataframe to plot a map with the clusters overlayered on top.




- # Analysis

Analyzing each cluster reveals that clusters 0 and 2 have a significant portion of sports facilities and closer inspection reveals that cluster 0 has a significant chunk of sports facilities.

Meanwhile, cluster 1 has to be avoided like the plague as it has very few sports facilities; therefore, it's in the best interest of stakeholders to avoid cluster 1 and focus on cluster 0.

Cluster 0

| | Neighborhood | Sum | Supplement Shop |
|---|---|---|---|
| 0 | Marble Hill | 0.120000 | 0.04 |
| 6 | Central Harlem | 0.086957 | 0.00 |
| 9 | Yorkville | 0.090000 | 0.00 |
| 10 | Lenox Hill | 0.100000 | 0.00 |
| 11 | Roosevelt Island | 0.115385 | 0.00 |
| 14 | Clinton | 0.100000 | 0.00 |
| 32 | Civic Center | 0.103093 | 0.00 |
| 34 | Sutton Place | 0.100000 | 0.00 |
| 37 | Stuyvesant Town | 0.111111 | 0.00 |
| 38 | Flatiron | 0.160000 | 0.00 |
| 39 | Hudson Yards | 0.122807 | 0.00 |

Cluster 1

| | Neighborhood | Sum | Supplement Shop |
|---|---|---|---|
| 1 | Chinatown | 0.010000 | 0.0 |
| 3 | Inwood | 0.000000 | 0.0 |
| 4 | Hamilton Heights | 0.000000 | 0.0 |
| 5 | Manhattanville | 0.000000 | 0.0 |
| 7 | East Harlem | 0.024390 | 0.0 |
| 12 | Upper West Side | 0.022727 | 0.0 |
| 17 | Chelsea | 0.030000 | 0.0 |
| 18 | Greenwich Village | 0.020000 | 0.0 |
| 19 | East Village | 0.000000 | 0.0 |
| 20 | Lower East Side | 0.020000 | 0.0 |
| 22 | Little Italy | 0.000000 | 0.0 |
| 24 | West Village | 0.020000 | 0.0 |
| 25 | Manhattan Valley | 0.023256 | 0.0 |
| 26 | Morningside Heights | 0.024390 | 0.0 |
| 27 | Gramercy | 0.000000 | 0.0 |
| 31 | Noho | 0.020000 | 0.0 |

Cluster 2

| | Neighborhood | Sum | Supplement Shop |
|---|---|---|---|
| 2 | Washington Heights | 0.058140 | 0.011628 |
| 8 | Upper East Side | 0.057471 | 0.000000 |
| 13 | Lincoln Square | 0.084211 | 0.000000 |
| 15 | Midtown | 0.080000 | 0.000000 |
| 16 | Murray Hill | 0.063158 | 0.000000 |
| 21 | Tribeca | 0.064103 | 0.000000 |
| 23 | Soho | 0.040000 | 0.000000 |
| 28 | Battery Park City | 0.063492 | 0.000000 |
| 29 | Financial District | 0.050000 | 0.000000 |
| 30 | Carnegie Hill | 0.068182 | 0.000000 |
| 33 | Midtown South | 0.050000 | 0.000000 |
| 35 | Turtle Bay | 0.040000 | 0.000000 |
| 36 | Tudor City | 0.066667 | 0.000000 |

- # Conclusion

Closer analysis of cluster 0 reveals that there is a dearth of supplement shops but ample presence of sports facilities that would attract prospective customers.

| | Neighborhood | Sum | Supplement Shop |
|---|---|---|---|
| **38** | Flatiron | 0.160000 | 0.00 |
| **39** | Hudson Yards | 0.122807 | 0.00 |
| **0** | Marble Hill | 0.120000 | 0.04 |
| **11** | Roosevelt Island | 0.115385 | 0.00 |
| **37** | Stuyvesant Town | 0.111111 | 0.00 |
| **32** | Civic Center | 0.103093 | 0.00 |
| **10** | Lenox Hill | 0.100000 | 0.00 |
| **14** | Clinton | 0.100000 | 0.00 |
| **34** | Sutton Place | 0.100000 | 0.00 |
| **9** | Yorkville | 0.090000 | 0.00 |
| **6** | Central Harlem | 0.086957 | 0.00 |

Clearly, Flatiron and Hudson Yards have the highest concentrations of sports facilities and surprisingly, no supplement shops. It would be ideal to focus on the neighborhoods on the list above for further inspection.

This capstone project is a part of the IBM Data Science Professional Course.

The purpose of this project is to extract meaningful data about Manhattan, NYC and apply data science techniques to identify the right locations to open up a shop.

Final decision has to be taken by stakeholders to determine which of the locations described above are ideal depending on the vicinity of the surroundings itself.