Suggesting best location for Sales

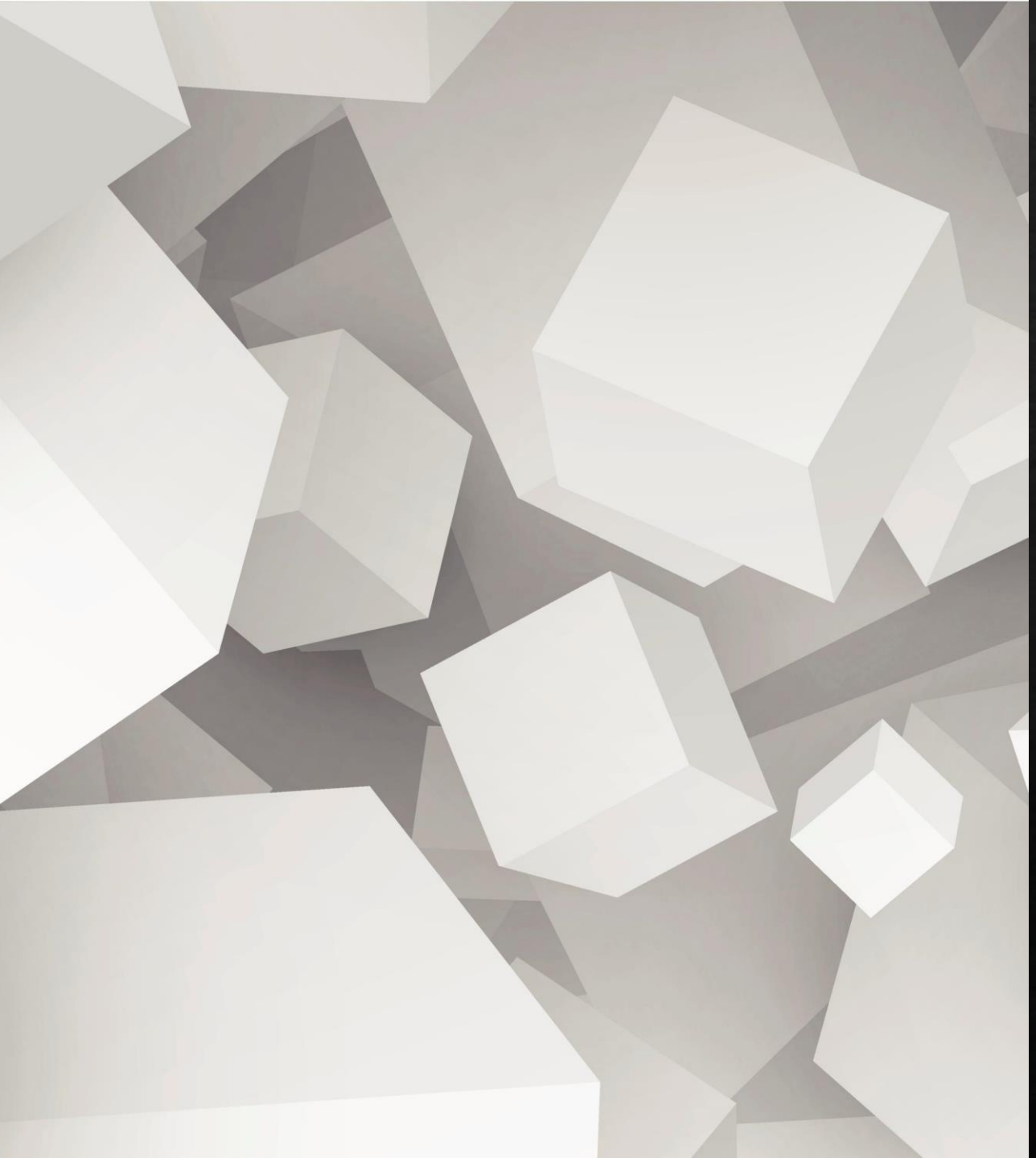
Suhetu Ring

Introduction/Business Problem

- As the Data Scientist of a large Gym/Fitness Corporation, I am held with the responsibility to come up with the best 5 locations in the borough ‘Staten Island’ of New York City such that we encounter maximum registrations. The Gym would also set up its store which would sell the company branded products as well as healthy edibles such as juices and salads. So, ultimately the locations provided by me should prove best in sales for the Gym and the store set up by the Corporation.

Data

- The data would in json format downloaded from 'https://geo.nyu.edu/catalog/nyu_2451_34572'.
- Credit goes to NYU for hosting the useful data being used by many individuals like me for Projects.
- Once I get the json data from the link, I would use the Foursquare API to explore the neighbourhoods and use the data to compare the neighbourhoods to come up with the best locations for the Gym Corporation.
- While dealing with the data in this project I have observed that the data was organised/structured 3 times by me each time in a different form:
  - The initial data is in json format which has been converted to a dataframe which contains all boroughs of New York.
  - Then the data was filtered and cleaned such that only the 'Staten Island' borough data is present in the dataframe.
  - The final major dataframe change occurred when I neglected the useless venue categories which I had retrieved from Foursquare API. That means, I will from this point onwards only focus on exploring neighbourhoods based on these particular venues only.



Methodology

- There are four major parts of this Methodology section to help you understand my project better.

Starting from next slide...

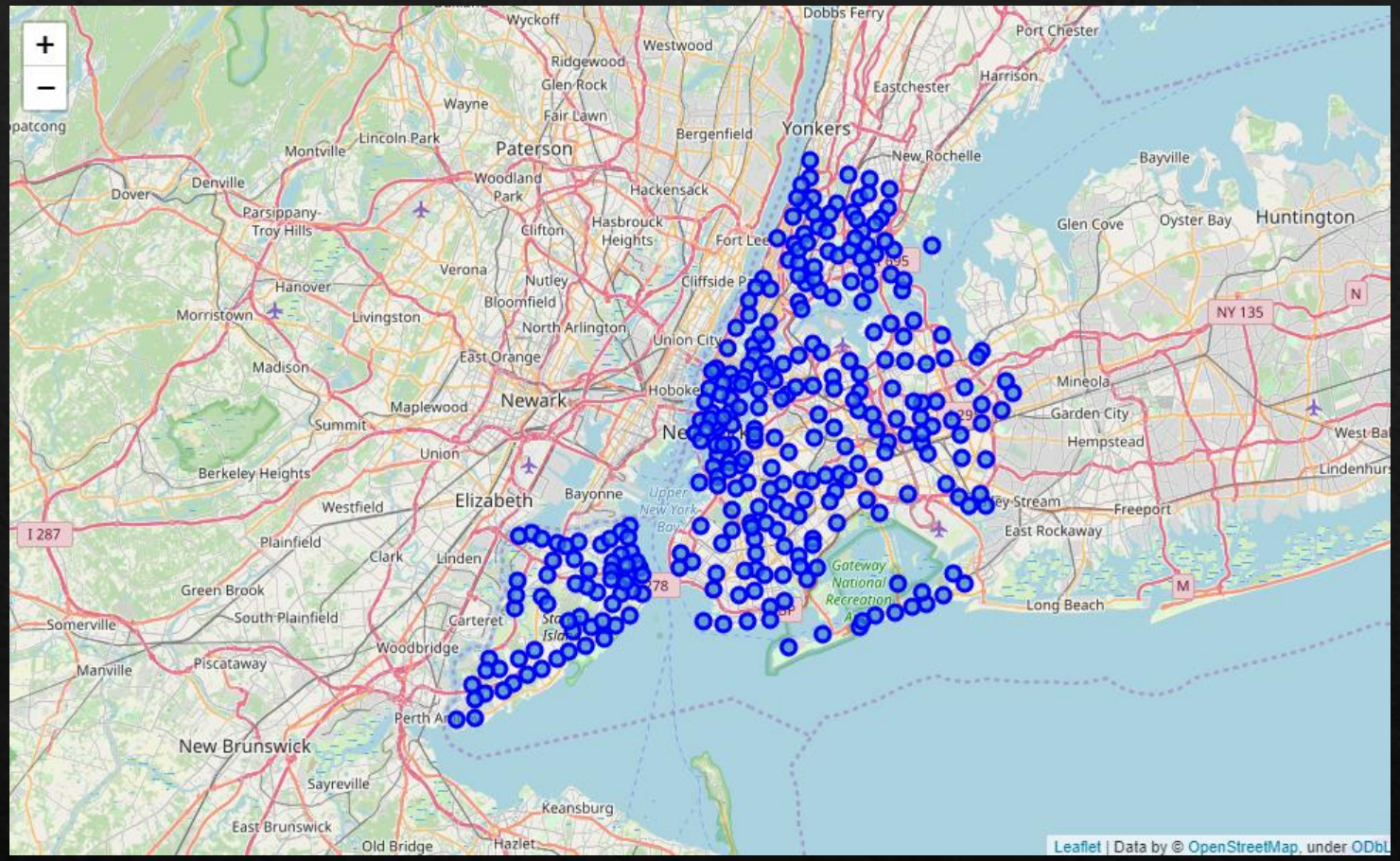
1. Data Cleaning/Structuring

- The initial data was downloaded and imported locally which was in json format. Further the data in json file was converted to a dictionary based on the key which contained the data that I required. This dictionary was then converted to a dataframe which resulted into the creation of our first dataframe.
- Now, since this dataframe contains data of neighbourhoods of all boroughs of New York, we filter this dataframe such that we only have the data of the borough ‘Staten Island’ of New York.
- I have then used Folium library to map the neighbourhoods of Staten Island.

Out[17]:

	Borough	Neighborhood	Latitude	Longitude
0	Staten Island	St. George	40.644982	-74.079353
1	Staten Island	New Brighton	40.640615	-74.087017
2	Staten Island	Stapleton	40.626928	-74.077902
3	Staten Island	Rosebank	40.615305	-74.069805
4	Staten Island	West Brighton	40.631879	-74.107182

Output of dataframe.head()



Neighborhoods of New York

2. Applying Foursquare API

- I have then applied the Foursquare API by loading my credentials.
- I had used the query of ‘exploring’ with the radius of 500 and limit of 100 to all the neighbourhoods of Staten Island.
- This had returned all the venues in each venue category available of that borough.
- Now I had applied a critical step to our analysis.
- I had created a dataframe of venues of each respective neighbourhood which would only contain venue categories that are useful for our analysis. The useful venue categories are:

- Baseball Stadium, Sporting Goods Shop, Breakfast Spot, Athletics & Sports, Spa, Gym, Gym / Fitness Centre, Sports Club, Sports Bar, Golf Course, Basketball Court, Tanning Salon, Skating Rink, Supplement Shop, Yoga Studio, Smoothie Shop, Dance Studio
- Now the neighbourhood analysis will only be done on the basis of these above venue categories.

3. Data Exploration

- One hot encoding of the useful venues have then be done as shown below,

```
In [54]: 1 si_useful
```

Out[54]:

	Neighborhood	Baseball Stadium	Sporting Goods Shop	Breakfast Spot	Athletics & Sports	Spa	Gym	Gym / Fitness Center	Sports Club	Sports Bar	Golf Course	Basketball Court	Tanning Salon	Skating Rink	Supp
0	Annadale	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.083333	0.00	0.000000	0.00	0.000000	0.
1	Arden Heights	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
2	Arlington	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
3	Arrochar	0.000000	0.000000	0.000000	0.055556	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
4	Bay Terrace	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
5	Bloomfield	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
6	Bulls Head	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.022222	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
7	Butler Manor	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.
8	Castleton Corners	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.00	0.000000	0.

- Then I had performed another critical step,
- getting the neighbourhoods which are best in either any of the useful venue categories, which resulted into below dataframe,

```
In [91]: 1 top
```

Out[91]:

	Venue Category	Neighborhood
0	Baseball Stadium	St. George
1	Sporting Goods Shop	St. George
2	Breakfast Spot	Pleasant Plains, West Brighton
3	Athletics & Sports, Gym / Fitness Center	Park Hill
4	Spa	Lighthouse Hill, Richmond Town
5	Gym	Sunnyside
6	Sports Club	Travis
7	Sports Bar, Dance Studio	Annadale
8	Sports Bar, Skating Rink	New Dorp Beach
9	Golf Course	Silver Lake
10	Yoga Studio	Shore Acres
11	Smoothie Shop	Annadale

4. K-Means Clustering

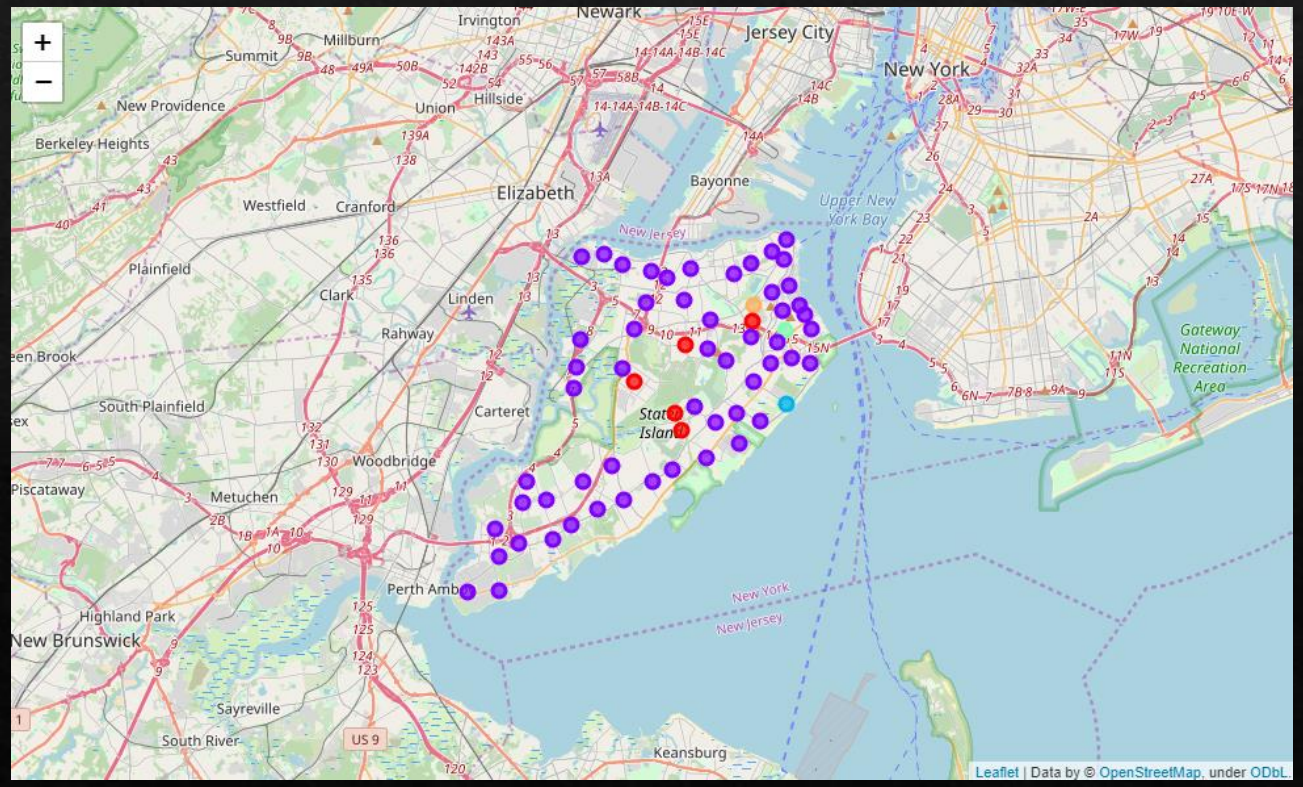
- Once I had got the neighbourhood data with the useful venue categories, I applied K-Means clustering with k=5 because I needed to come up with 5 best locations for the setting up of new Gym/Fitness centre, resulting in:

```
si_merged.head()
```

Out[61]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Staten Island	St. George	40.644982	-74.079353	1	Sporting Goods Shop	Baseball Stadium	Sports Club	Breakfast Spot	Athletics & Sports
1	Staten Island	New Brighton	40.640615	-74.087017	1	Dance Studio	Sports Club	Sporting Goods Shop	Breakfast Spot	Athletics & Sports
2	Staten Island	Stapleton	40.626928	-74.077902	1	Breakfast Spot	Dance Studio	Sports Club	Sporting Goods Shop	Athletics & Sports
3	Staten Island	Rosebank	40.615305	-74.069805	1	Breakfast Spot	Dance Studio	Sports Club	Sporting Goods Shop	Athletics & Sports
4	Staten Island	West Brighton	40.631879	-74.107182	1	Breakfast Spot	Dance Studio	Sports Club	Sporting Goods Shop	Athletics & Sports

- Once I had got the above clusters of data, I decided to map them using Folium

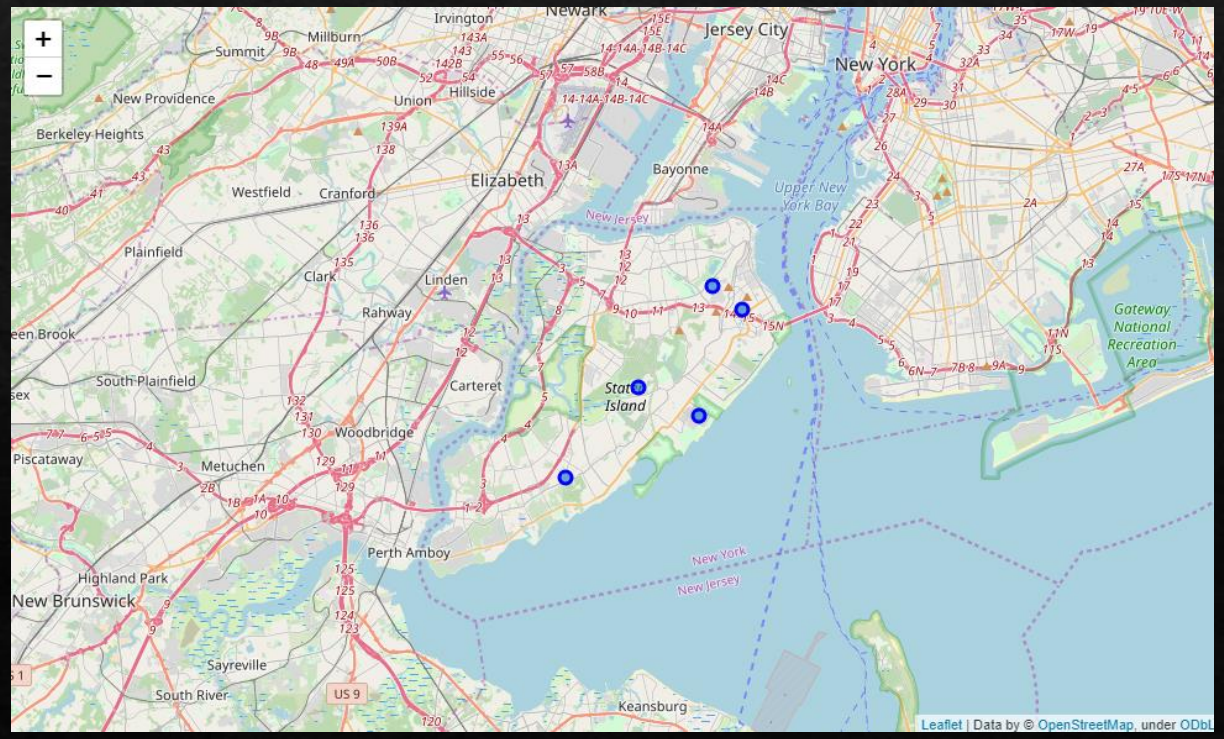


- Next, I examined each of the cluster.
- Now, the final locations have been selected such that:
- One location from each cluster which belongs to top dataframe.
- By following above selecting parameter, we can ensure that the locations we select are famous/known locations of this(fitness/gym) kind of category.
- These locations also would result in an increase in the number of registration and sales.

Results

- The top 5 locations that I have come up with are:
  - Annadale, Stale Island
  - Lighthouse Hill, Stale Island
  - New Dorp Beach, Stale Island
  - Park Hill, Stale Island
  - Silver Lake, Stale Island

- When these above locations are marked on map using Folium, we get the below map:



Discussion

- The major observation that I would like to point out is that it looks like that the Southern part of Staten Island is more commercial than the Northern part.
- This could occur naturally or could be a flaw in the Foursquare data of Staten Island neighbourhoods.
- One more reason of the above observation could be because of the near proximity of other boroughs to the Southern part of Staten Island, or because of the seaside.

Conclusion

- I would like to conclude this report by going over the accomplishment of objectives that I had laid out in the Introduction/Problem Statement section of this report.
- All the objectives have been met and the five locations that I have suggested will hopefully bring maximum sales to the Gym/Fitness Corporation.