**CAPSTONE PROJECT REPORT**

**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:

1. The initial data is in json format which has been converted to a dataframe which contains all boroughs of New York.
2. Then the data was filtered and cleaned such that only the ‘Staten Island’ borough data is present in the dataframe.
3. 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:**

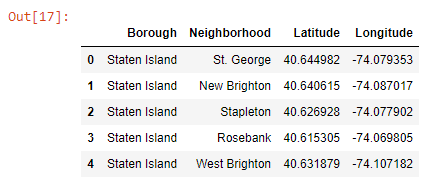
There are four major parts of this Methodology section to help you understand my project better.

1. **Data Cleaning/Structuring-**

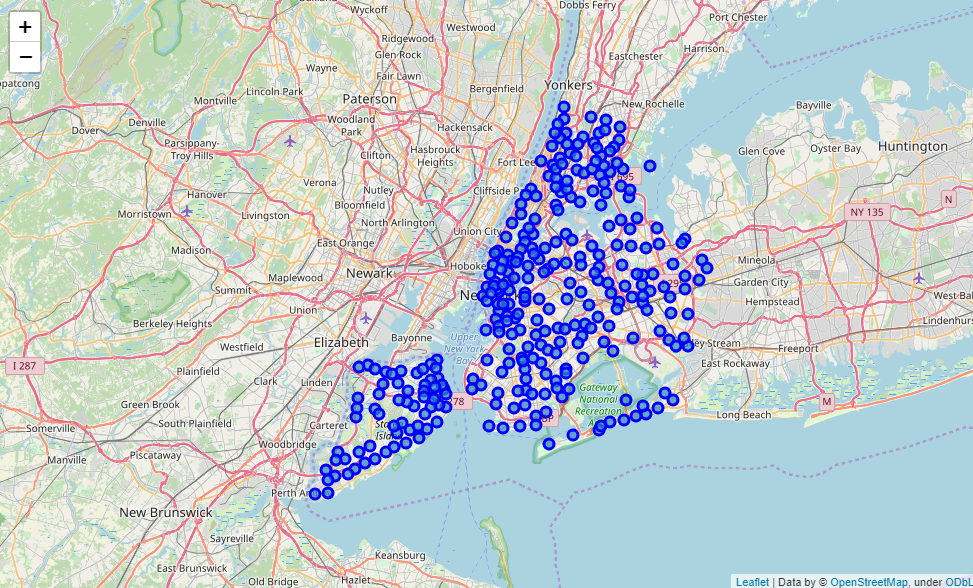
The initial data was downloaded and imported locally which was in json format. Further the data in json file was converted to a dictionary based on the key which contained the data that I required. This dictionary was then converted to a dataframe which resulted into the creation of our first dataframe.

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.

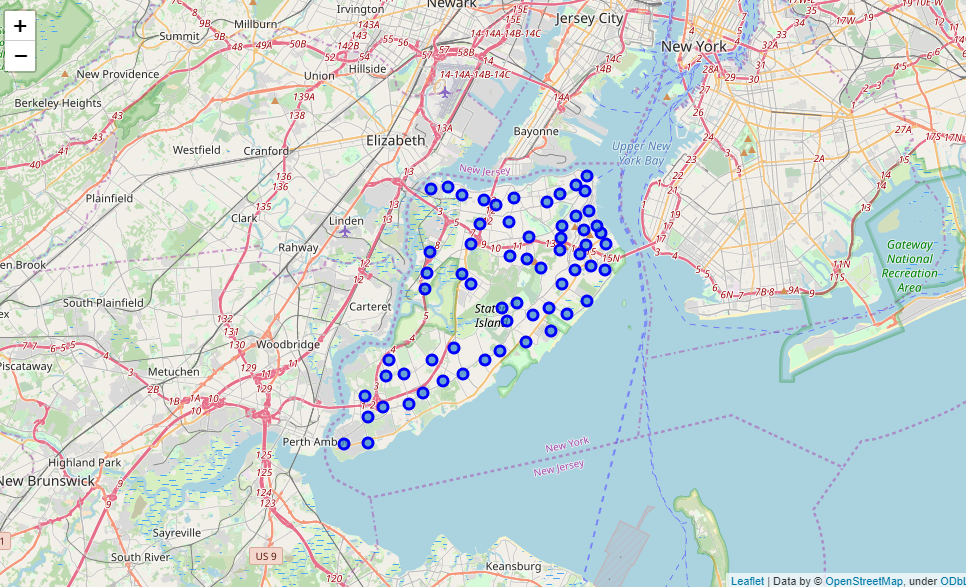
I have then used Folium library to map the neighbourhoods of Staten Island.



(.head() of the Staten Island neighbourhoods)



(Neighbourhoods of New York)



(Neighbourhoods of Staten Island borough)

1. **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.

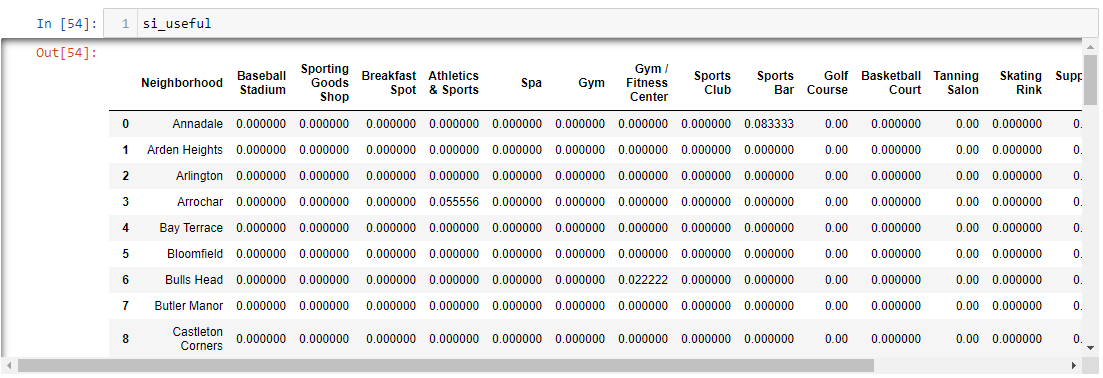
I had created a dataframe of venues of each respective neighbourhood which would only contain venue categories that are useful for our analysis. The useful venue categories are:

* Baseball Stadium
* Sporting Goods Shop
* Breakfast Spot
* Athletics & Sports
* Spa
* Gym
* Gym / Fitness Center
* Sports Club
* Sports Bar
* Golf Course
* Basketball Court
* Tanning Salon
* Skating Rink
* Supplement Shop
* Yoga Studio
* Smoothie Shop

Now the neighbourhood analysis will only be done on the basis of these above venue categories.

1. **Data Exploration-**

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



Then I had performed another critical step,

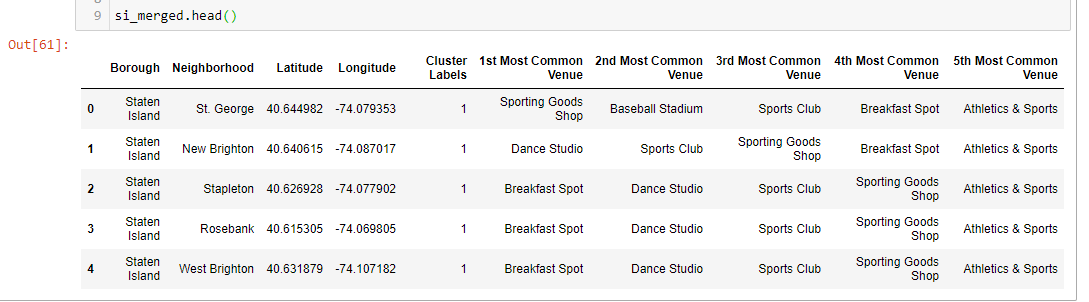
getting the neighbourhoods which are best in either any of the useful venue categories, which resulted into below dataframe,



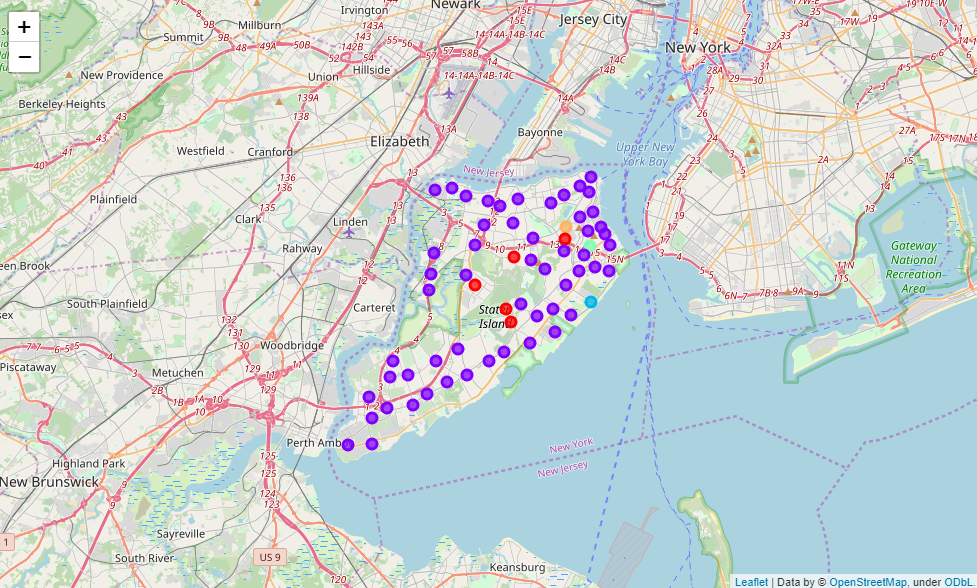
1. **K-Means Clustering-**

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 centres.

The clustering resulted in,

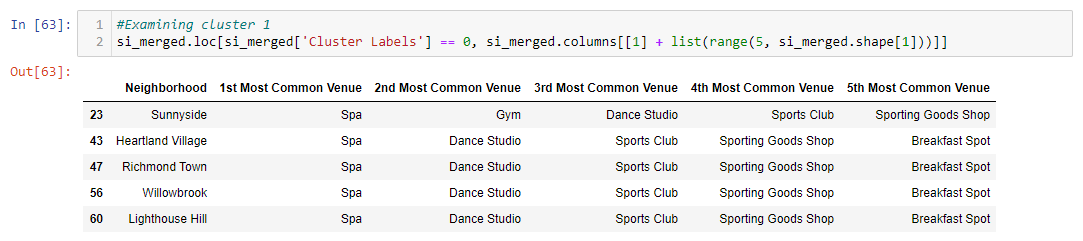


Once I had got the above clusters of data, I decided to map them using Folium

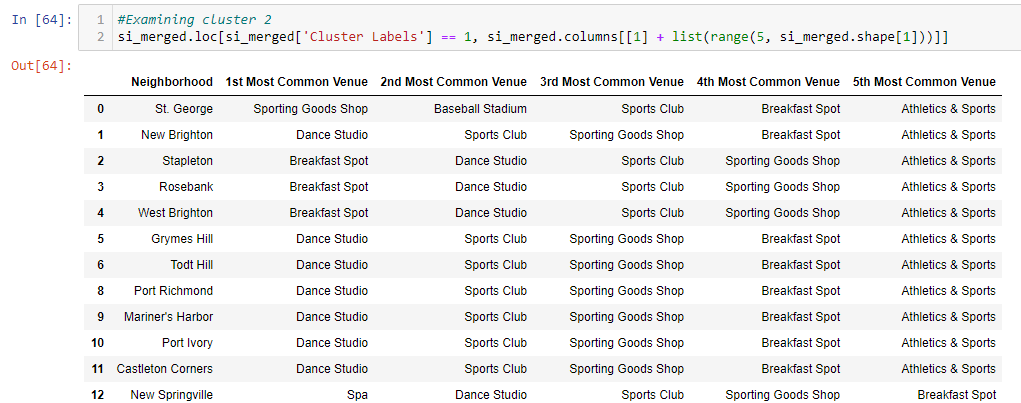


Next, I examined each of the cluster.

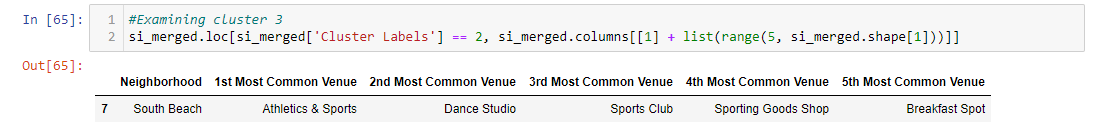
Cluster 1:



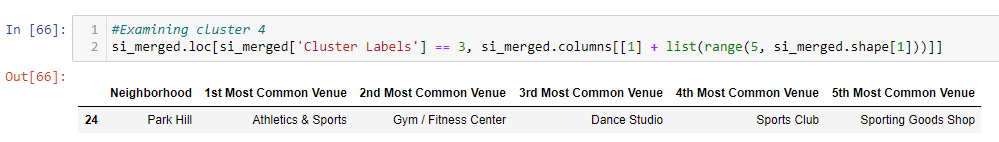
Cluster 2:



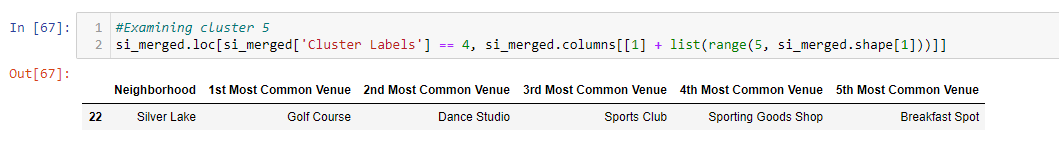
Cluster 3:



Cluster 4:



Cluster 5:



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.

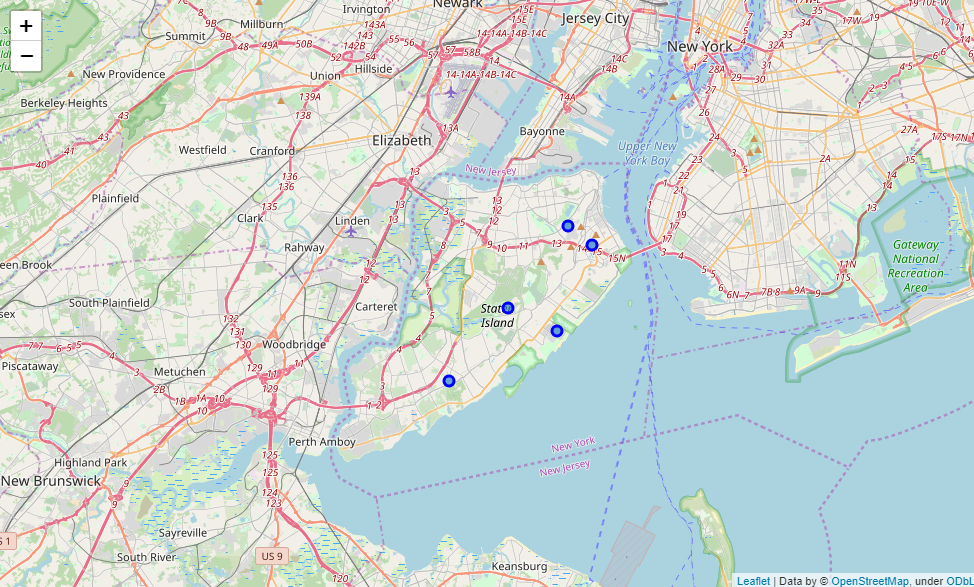
These locations also would result in an increase in the number of registration and sales.

**Results:**

The top 5 locations that I have come up with are:

1. Annadale, Stale Island
2. Lighthouse Hill, Stale Island
3. New Dorp Beach, Stale Island
4. Park Hill, Stale Island
5. 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.