CAPSTONE PROJECT- THE BATTLE OF NEIGHBOURHOODS

- By Sai Vignesh Reddy Cholleti

1. Introduction/Business problem:

- **Background:**
  Safety is a top concern when moving to a new area. If you don't feel safe in your own home, you're not going to be able to enjoy living there.

- **Problem:**
  This project aims to select the safest borough in New York City (NYC) based on total crimes, explore the neighbourhood of that borough to find the 10 most common venues in each neighbourhood and finally cluster the neighbourhoods using k-means clustering.

- **Target Audience:**
  Expats who are considering to relocate to NYC will be interested to identify the safest borough in NYC and explore its neighbourhoods and common venues around each neighbourhood.

2. Data Acquisition and Cleaning:

Data Acquisition: The data acquired for this project is a combination of data from three sources:

- The first data source of the project uses a New York City crime data that shows the complaints registered by police department.

- The second source of data is scraped from a Wikipedia page that contains the Boroughs of New York City. This page contains additional information about the boroughs.

- The third data source is the list of neighbourhoods in each borough which is taken from coursera previous lab session.

Data Cleaning: The data cleaning process for each of the three sources of data are done separately.

- From the NYC crime data, the crimes during the most recent year (2015) are only selected. The major categories of crime are pivoted to get the total crimes per boroughs for each major category.

- The second data is scraped from a Wikipedia page using the BeautifulSoup library in python. Using this library we can extract the data in the tabular form as shown in the website.

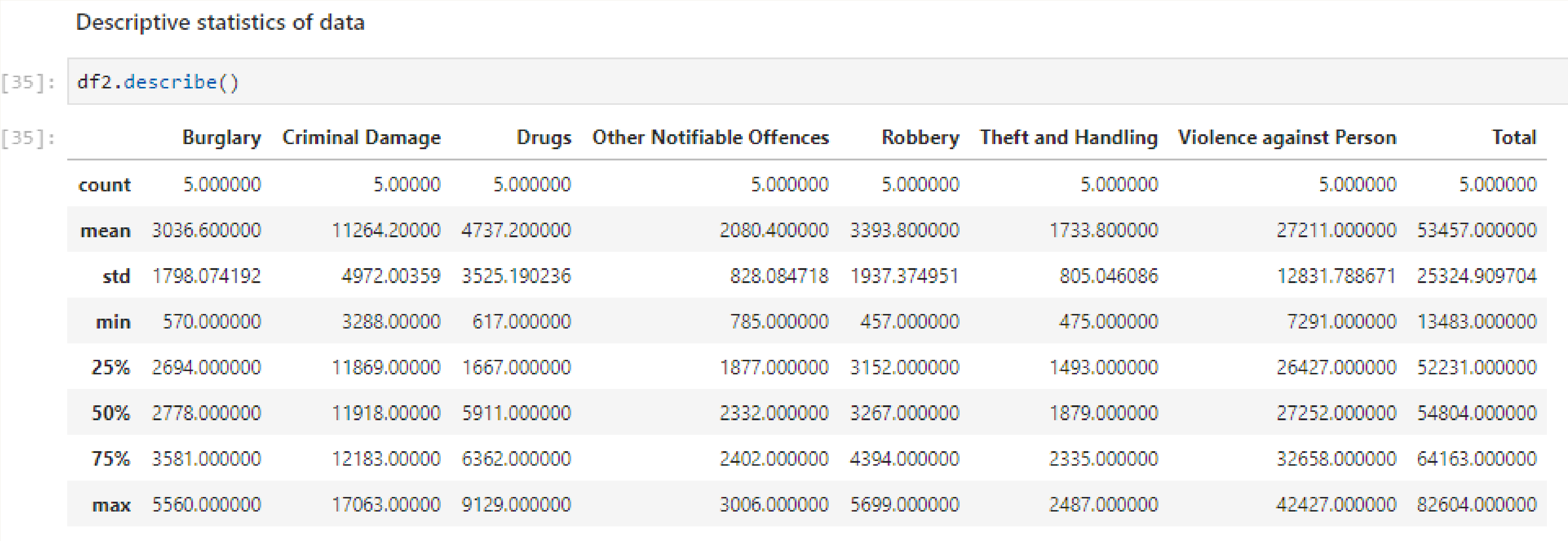
- The two data sets are merged on the Borough names to form a new data set. The purpose of this data set is to visualize the crime rates in each borough and identify the borough with the least crimes recorded during the year 2015.

- After visualizing the crime in each borough we can find the borough with lowest crime rate. The third data set is created, with the names of the neighbourhoods and the name of the borough with the latitude and longitude obtained using python geocoder.

- The new data set is used to generate the 10 most common venues for each neighbourhood using the Foursquare API, finally using k-means clustering algorithm to cluster similar neighbourhoods together.

3. Methodology:

Exploratory Data Analysis



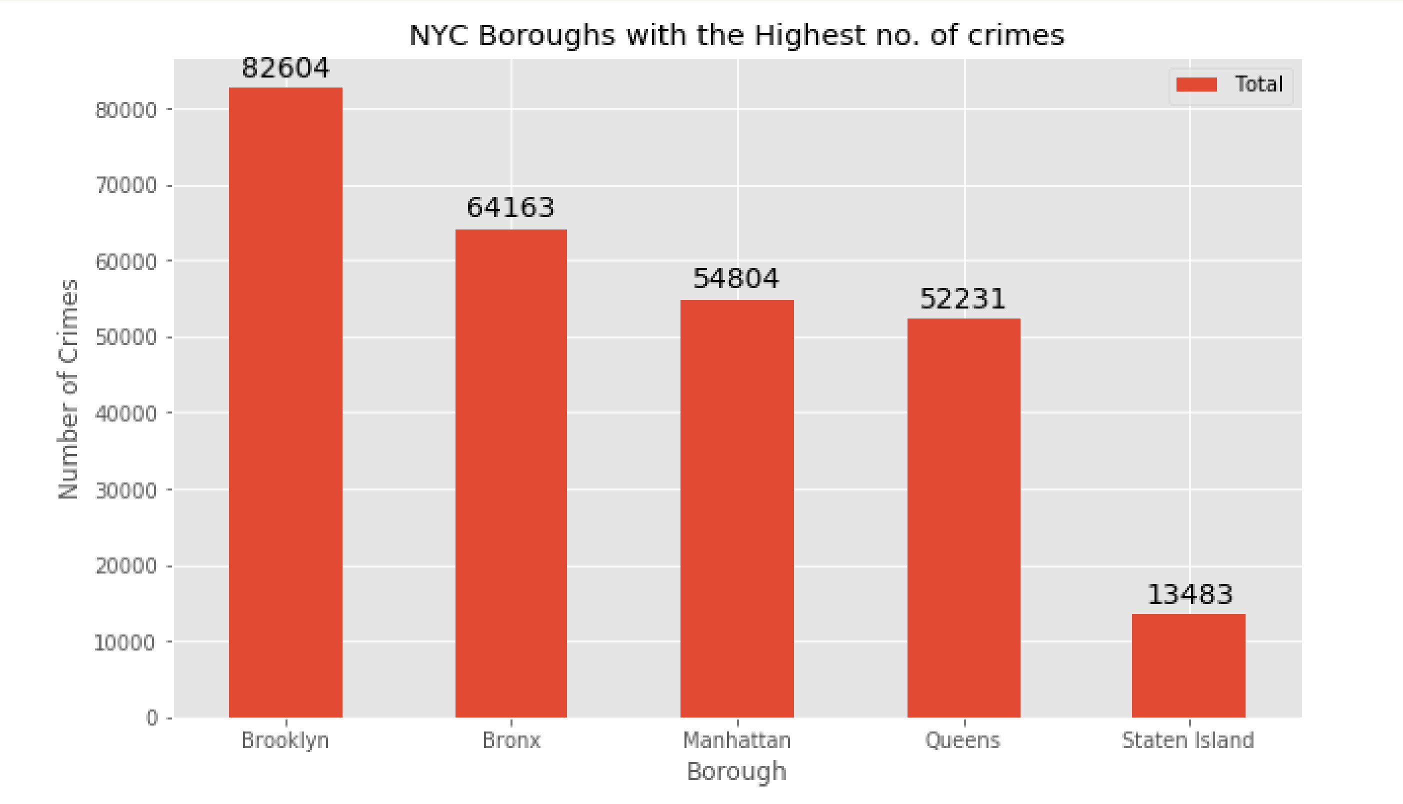
Descriptive statistics of data

```
[35]: df2.describe()
```

```
[35]:
          Burglary  Criminal Damage        Drugs  Other Notifiable Offences      Robbery  Theft and Handling  Violence against Person          Total
count     5.000000          5.00000     5.000000                   5.000000     5.000000            5.000000                 5.000000       5.000000
mean   3036.600000      11264.20000  4737.200000                2080.400000  3393.800000         1733.800000             27211.000000   53457.000000
std    1798.074192       4972.00359  3525.190236                 828.084718  1937.374951          805.046086             12831.788671   25324.909704
min     570.000000       3288.00000   617.000000                 785.000000   457.000000          475.000000              7291.000000   13483.000000
25%    2694.000000      11869.00000  1667.000000                1877.000000  3152.000000         1493.000000             26427.000000   52231.000000
50%    2778.000000      11918.00000  5911.000000                2332.000000  3267.000000         1879.000000             27252.000000   54804.000000
75%    3581.000000      12183.00000  6362.000000                2402.000000  4394.000000         2335.000000             32658.000000   64163.000000
max    5560.000000      17063.00000  9129.000000                3006.000000  5699.000000         2487.000000             42427.000000   82604.000000
```

The count for each of the major categories of crime returns the value 5 which is the number of NYC boroughs. ‘Violence against Person’ is the highest reported crime during the year 2015 followed by ‘Criminal Damage’, ‘Robbery’. The lowest recorded crimes are ‘Theft and Handling’, ‘Other Notifiable Offences’ and ‘Burglary’.

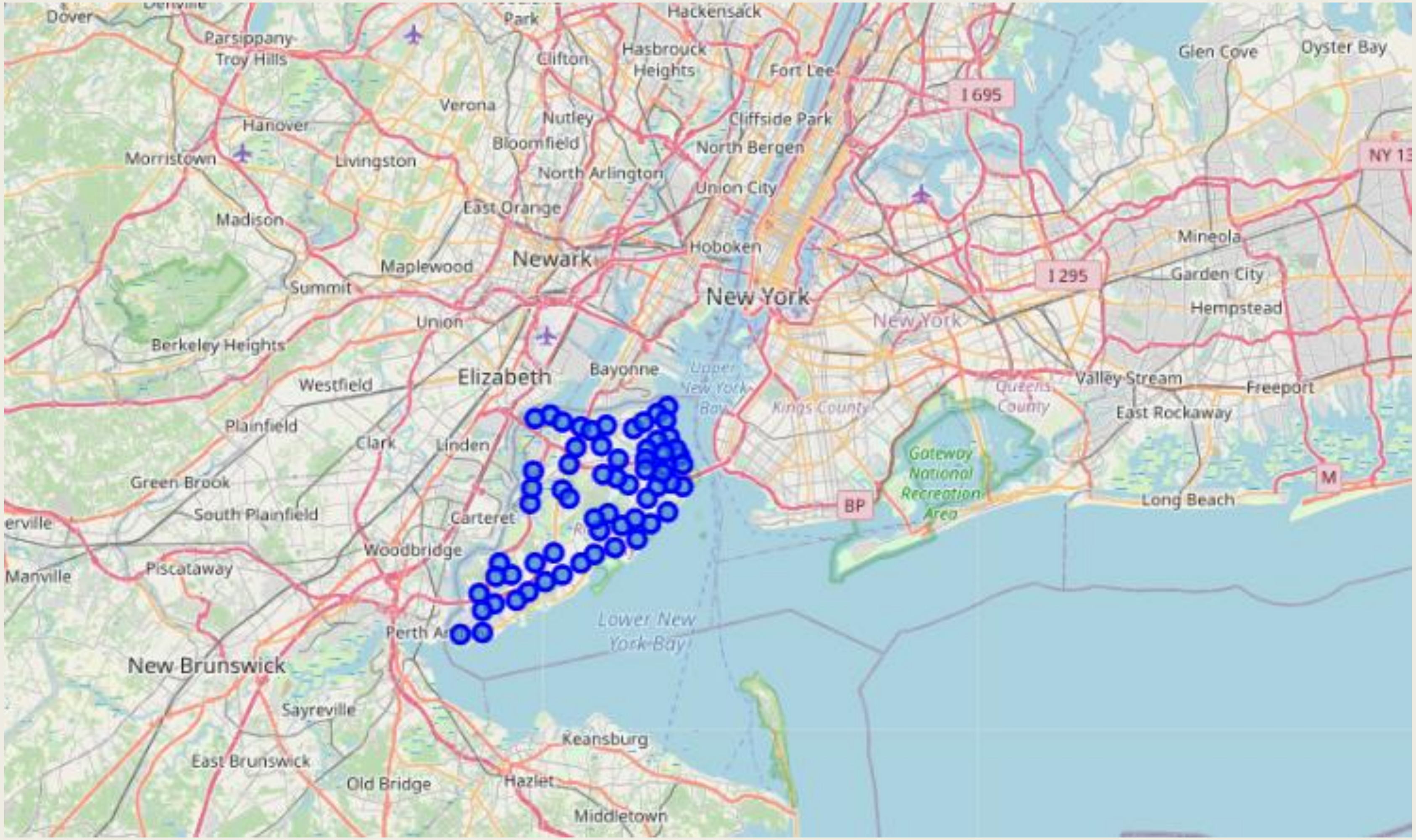
Boroughs with the highest crime rates



- Comparing five boroughs with the highest crime rates during the year 2015, it is evident that Brooklyn has the highest crimes recorded followed by Bronx, Manhattan, Queens and Staten Island. Staten Island has a significantly least crime rate than the other 4 boroughs.

- Therefore, Staten Island which has least crime rate is selected for clustering process.

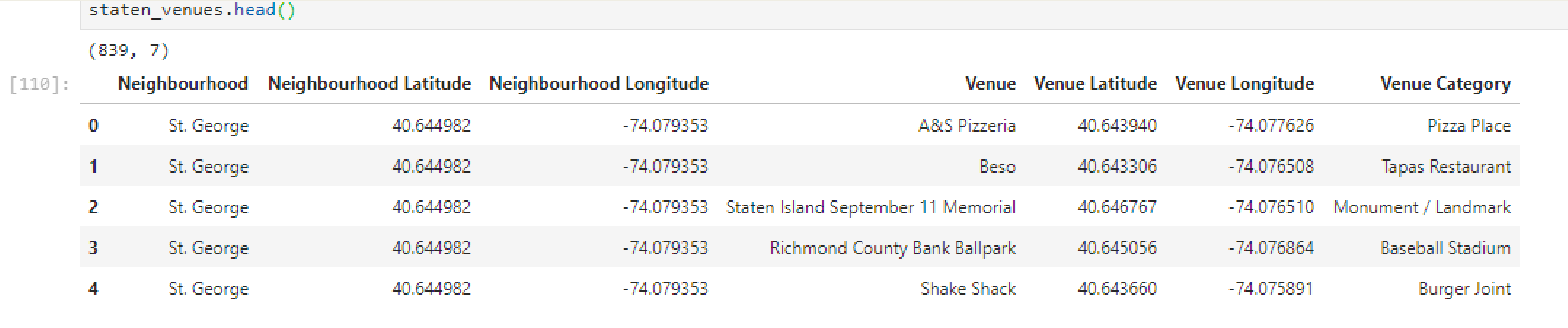
Neighbourhoods in Staten Island



There are 63 neighbourhoods in the borough of Staten Island. They are visualized on a map using folium on python.

Modelling:

- Using the final data set containing the neighbourhoos in Staten Island along with latitude and longitude, we can find all the venues within a 500 meter radius of each neighbourhood by connecting to the Foursquare API



```
staten_venues.head()
```

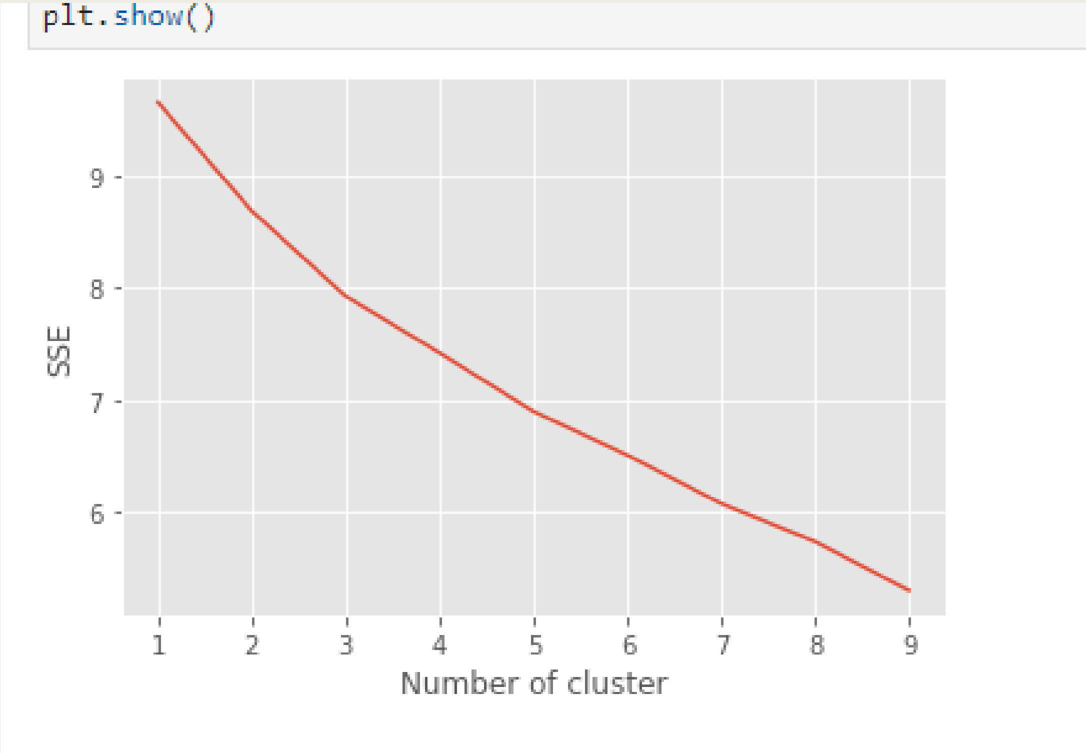
```
(839, 7)
```

[110]:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	St. George	40.644982	-74.079353	A&S Pizzeria	40.643940	-74.077626	Pizza Place
1	St. George	40.644982	-74.079353	Beso	40.643306	-74.076508	Tapas Restaurant
2	St. George	40.644982	-74.079353	Staten Island September 11 Memorial	40.646767	-74.076510	Monument / Landmark
3	St. George	40.644982	-74.079353	Richmond County Bank Ballpark	40.645056	-74.076864	Baseball Stadium
4	St. George	40.644982	-74.079353	Shake Shack	40.643660	-74.075891	Burger Joint

- One hot encoding is done on the venues data. The venues data is then grouped by the neighbourhood and the mean of the venues are calculated, finally the 10 common venues are calculated for each of the neighbourhoods.
- To help people find similar neighbourhoods in the safest borough, we will be clustering similar neighbourhoods using k-means clustering which is a form of unsupervised machine learning algorithm that clusters based on predefined cluster size.
- We will use a cluster size of 3 for this prokect that will cluster the 63 neighbourhoods into 3 clusters. The reason to conduct a k-means clustering is to cluster neighbourhoods with similar venues together so that people can shortlist the area of their interest based on the venues/amentities around each neighbourhood.

Elbow Method:



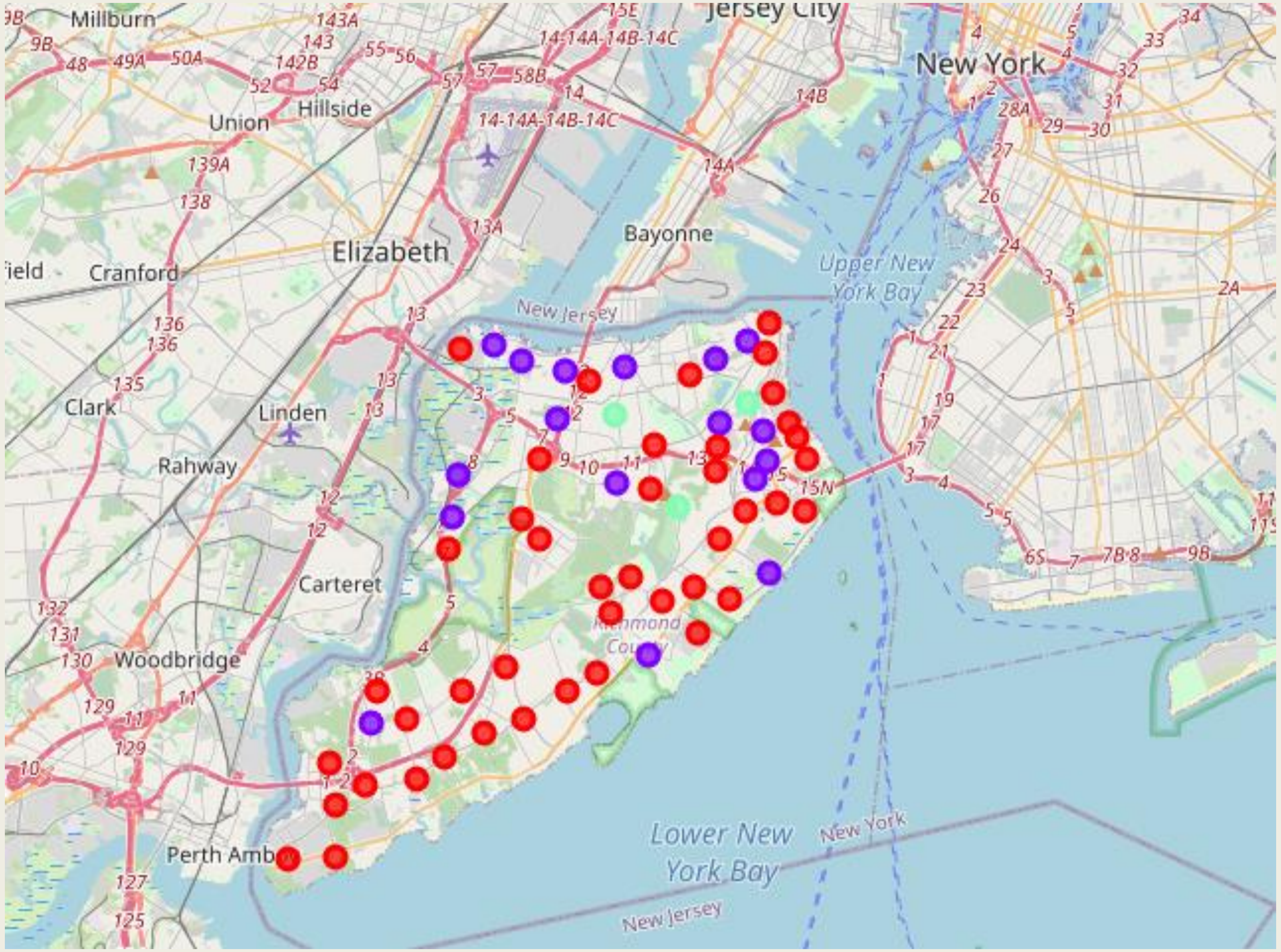
```
plt.show()
```

Elbow method for optimal k value

- A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The **Elbow Method** is one of the most popular methods to determine this optimal value of k.

- To determine the optimal number of clusters, we have to select the value of k at the “elbow” i.e the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is **3**.

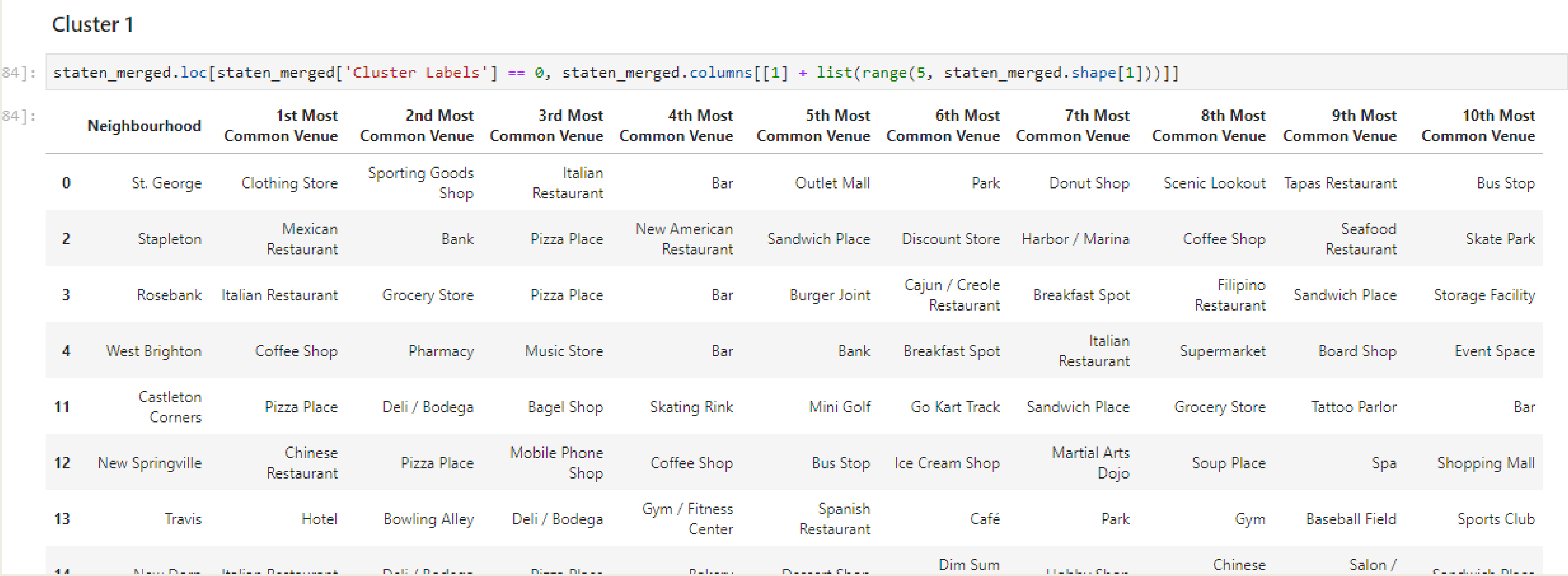
4. Results:



- After running the k-means clustering we can access each cluster created to see which neighbourhoods were assigned to each of the three clusters. Visualizing the clustered neighbourhoods on a map using folium library.

- Each cluster is color coded for the ease of presentation, we can see that majority of the neighbourhood falls in the red cluster which is the first cluster. Three neighbourhoods have their own cluster (Green), and Other cluster which is blue colored is second cluster.

Cluster 1: Looking into the neighbourhoods in the first cluster



Cluster 1

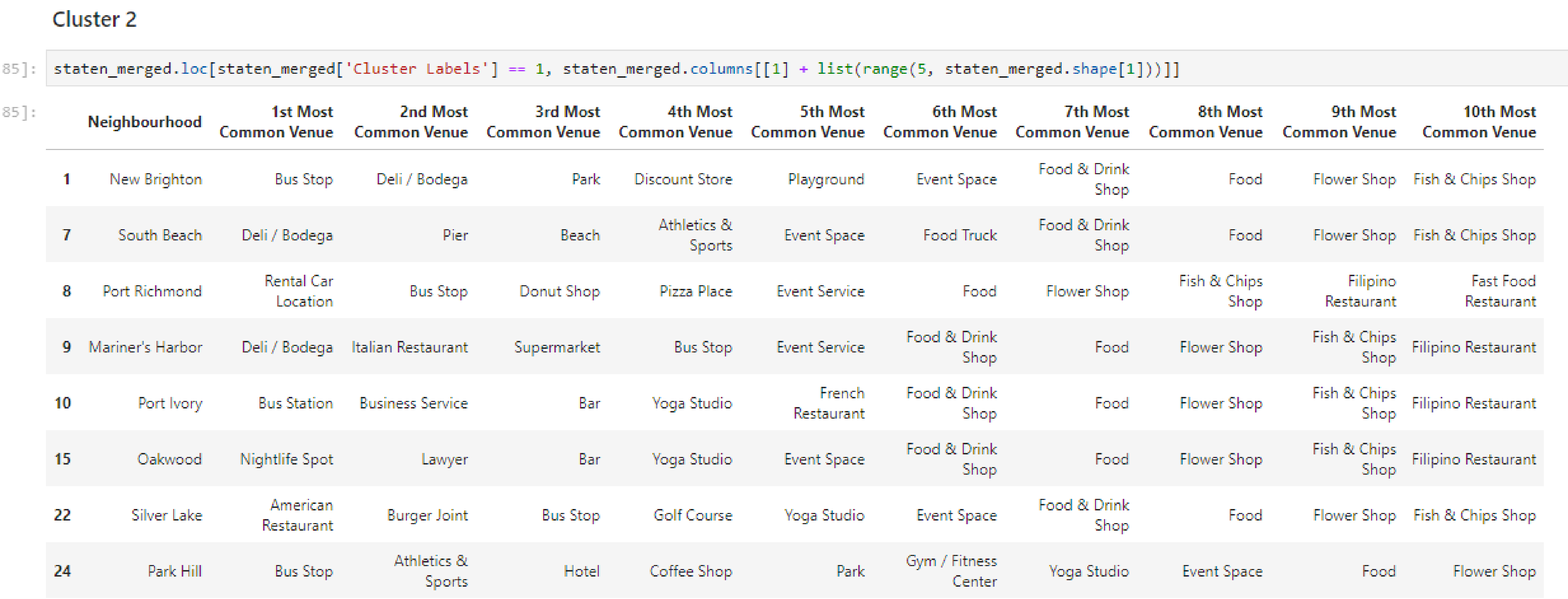
```
84]: staten_merged.loc[staten_merged['Cluster Labels'] == 0, staten_merged.columns[[1] + list(range(5, staten_merged.shape[1]))]]
```

84]:

	Neighbourhood	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	St. George	Clothing Store	Sporting Goods Shop	Italian Restaurant	Bar	Outlet Mall	Park	Donut Shop	Scenic Lookout	Tapas Restaurant	Bus Stop
2	Stapleton	Mexican Restaurant	Bank	Pizza Place	New American Restaurant	Sandwich Place	Discount Store	Harbor / Marina	Coffee Shop	Seafood Restaurant	Skate Park
3	Rosebank	Italian Restaurant	Grocery Store	Pizza Place	Bar	Burger Joint	Cajun / Creole Restaurant	Breakfast Spot	Filipino Restaurant	Sandwich Place	Storage Facility
4	West Brighton	Coffee Shop	Pharmacy	Music Store	Bar	Bank	Breakfast Spot	Italian Restaurant	Supermarket	Board Shop	Event Space
11	Castleton Corners	Pizza Place	Deli / Bodega	Bagel Shop	Skating Rink	Mini Golf	Go Kart Track	Sandwich Place	Grocery Store	Tattoo Parlor	Bar
12	New Springville	Chinese Restaurant	Pizza Place	Mobile Phone Shop	Coffee Shop	Bus Stop	Ice Cream Shop	Martial Arts Dojo	Soup Place	Spa	Shopping Mall
13	Travis	Hotel	Bowling Alley	Deli / Bodega	Gym / Fitness Center	Spanish Restaurant	Café	Park	Gym	Baseball Field	Sports Club
14	New Dorp	Italian Restaurant	Deli / Bodega	Pizza Place	Bakery	Dessert Shop	Dim Sum	Hobby Shop	Chinese	Salon /	Sandwich Place

The cluster one is the biggest cluster with most of the neighbourhoods in the borough Staten Island. Upon closely examining these neighbourhoods, we can see that the most common venues in these neighbourhoods are Restaurants, Pizza places, Cafe, Supermarkets, and stores etc.

Cluster 2: Looking into the neighbourhoods in the second cluster



Cluster 2

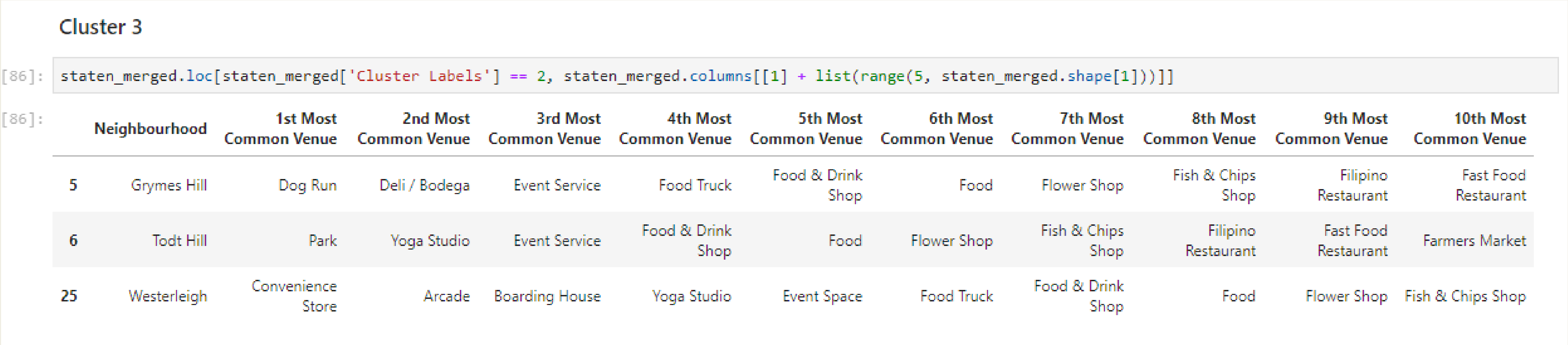
```
85]: staten_merged.loc[staten_merged['Cluster Labels'] == 1, staten_merged.columns[[1] + list(range(5, staten_merged.shape[1]))]]
```

85]:

	Neighbourhood	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
1	New Brighton	Bus Stop	Deli / Bodega	Park	Discount Store	Playground	Event Space	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop
7	South Beach	Deli / Bodega	Pier	Beach	Athletics & Sports	Event Space	Food Truck	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop
8	Port Richmond	Rental Car Location	Bus Stop	Donut Shop	Pizza Place	Event Service	Food	Flower Shop	Fish & Chips Shop	Filipino Restaurant	Fast Food Restaurant
9	Mariner's Harbor	Deli / Bodega	Italian Restaurant	Supermarket	Bus Stop	Event Service	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop	Filipino Restaurant
10	Port Ivory	Bus Station	Business Service	Bar	Yoga Studio	French Restaurant	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop	Filipino Restaurant
15	Oakwood	Nightlife Spot	Lawyer	Bar	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop	Filipino Restaurant
22	Silver Lake	American Restaurant	Burger Joint	Bus Stop	Golf Course	Yoga Studio	Event Space	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop
24	Park Hill	Bus Stop	Athletics & Sports	Hotel	Coffee Shop	Park	Gym / Fitness Center	Yoga Studio	Event Space	Food	Flower Shop

The second cluster has some neighborhoods which consists of venues such as Bus stops, Deli/Bodega, and Restaurants.

Cluster 3: Looking into the neighbourhoods in the third cluster



Cluster 3

```
[86]: staten_merged.loc[staten_merged['Cluster Labels'] == 2, staten_merged.columns[[1] + list(range(5, staten_merged.shape[1]))]]
```

[86]:

	Neighbourhood	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
5	Grymes Hill	Dog Run	Deli / Bodega	Event Service	Food Truck	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop	Filipino Restaurant	Fast Food Restaurant
6	Todt Hill	Park	Yoga Studio	Event Service	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop	Filipino Restaurant	Fast Food Restaurant	Farmers Market
25	Westerleigh	Convenience Store	Arcade	Boarding House	Yoga Studio	Event Space	Food Truck	Food & Drink Shop	Food	Flower Shop	Fish & Chips Shop

The third cluster is the smallest cluster with only three neighborhoods which consists of venues such as Dog Run, Park, and Convenience store.

5. Discussions:

- The aim of this project is to help people who want to relocate to the safest borough in New York City, expats can chose the neighbourhoods to which they want to relocate based on the most common venues in it.

- For example if a person is looking for a neighbourhood with good connectivity and public transportation we can see that Clusters 2 have Train stations and Bus stops as the most common venues.

- If a person is looking for a neighbourhood with stores and restaurants in a close proximity then the neighbourhoods in the first cluster is suitable.

- For a person who enjoys nature, I feel that the neighbourhoods in Cluster 3 are more suitable dues to the common venues in that cluster, these neighbourhoods have common venues such as Dog Run, Park, and Convenience store. Cluster 1 being biggest cluster with most number of neighbourhoods and variety of venues in each neighbourhood, is ideal for any kind of person. The choices of neighbourhoods may vary from person to person.

6. Conclusion:

- This project helps a person get a better understanding of the neighbourhoods with respect to the most common venues in that neighbourhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighbourhood.

- We have just taken safety as a primary concern to shortlist the safest borough of New York City. The future of this project includes taking other factors such as cost of living in the areas into consideration to shortlist the borough, such as filtering areas based on a predefined budget.