Capstone Project – The Battle of Neighbourhoods.

Report by - Sai Vignesh Reddy Cholleti.

Date – 10th June, 2020.

Table of Contents:

1. Introduction

1.1 Background:

The average American moves about eleven times in their lifetime. This brings us to the question: Do people move until they find a place to settle down where they truly feel happy, or do our wants and needs change over time, prompting us to eventually leave a town we once called home for a new area that will bring us satisfaction? Or, do we too often move to a new area without knowing exactly what we're getting into, forcing us to turn tail and run at the first sign of discomfort? To minimize the chances of this happening, we should always do proper research when planning our next move in life. Consider the following factors when picking a new place to live so you don't end up wasting your valuable time and money making a move you'll end up regretting. 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.

1.2 Problem:

The crime statistics dataset of New York city found on Kaggle has crimes in each Boroughs of New York city from 2001 to 2015. The year 2015 being the latest we will be considering the data of that year which is actually old information as of now. The crime rates in each borough may have changed over time. This project aims to select the safest borough in New York City based on the total crimes, explore the neighbourhoods of that borough to find the 10 most common venues in each neighbourhood and finally cluster the neighbourhoods using k-mean clustering.

1.3 Target Audience:

Expats who are considering to relocate to New York City will be interested to identify the safest borough in New York city and explore its neighbourhoods and common venues around each neighbourhood. Most people like international students who could relocate to New York City for their education, career aspects etc. would get benefitted because they might not know anything about the neighbourhood before they move in.

2. Data Acquisition and Cleaning

2.1 Data Acquisition:

The data acquired for this project is a combination of data from two sources. The first data source of the project uses a New York Historical crime data that shows the crime per borough in New York. The dataset contains the following columns:

● CMPLNT_NUM: Randomly generated persistent ID for each complaint.

● CMPLNT_FR_DT: Exact date of occurrence for the reported event (or starting date of occurrence, if CMPLNT_TO_DT exists.

● CMPLNT_FR_TM: Exact time of occurrence for the reported event (or starting time of occurrence, if CMPLNT_TO_TM exists.

● CMPLNT_TO_DT: Ending date of occurrence for the reported event, if exact time of occurrence is unknown

● CMPLNT_TO_TM: Ending time of occurrence for the reported event, if exact time of occurrence is unknown

● RPT_DT: Date event was reported to police

● KY_CD: Three digit offense classification code

● OFNS_DESC: Description of offense corresponding with key code

● PD_CD: Three digit internal classification code (more granular than Key Code)

● PD_DESC: Description of internal classification corresponding with PD code (more granular than Offense Description.

● CRM_ATPT_CPTD_CD: Indicator of whether crime was successfully completed or attempted, but failed or was interrupted prematurely.

● LAW_CAT_CD: Level of offense: felony, misdemeanor, violation

● JURIS_DESC: Jurisdiction responsible for incident. Either internal, like Police, Transit, and Housing; or external like Correction,Port Authority etc.

● BORO_NM: The name of the borough in which the incident occurred

● ADDR_PCT_CD: The precinct in which the incident occurred

● LOC_OF_OCCUR_DESC: Specific location of occurrence in or around the premises; inside, opposite of, front of, rear of

● PREM_TYP_DESC: Specific description of premises; grocery store, residence, street, etc.

● PARKS_NM: Name of NYC park, playground or greenspace of occurrence, if applicable (state parks are not included)

● HADEVELOPT: Name of NYCHA housing development of occurrence, if applicable

● X_COORD_CD: X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)

● Y_COORD_CD: Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)

● Latitude: Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

● Longitude: Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326).

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 following are the columns:

● Borough: The names of the 33 London boroughs.
● County: Categorizing the borough as an Inner London borough or an Outer London Borough.
● Population: The population in the borough recorded during the year 2019.
● GDP_billions: The Gross Domestic Product of the borough in billions.
● GDP_percapita: The Gross Domestic Product of the borough in per-capita.
● Area_sqmiles: Land Area of the borough in square miles.
● Area_sqkms: Land Area of the borough in square kms.
● Density_sqmiles: Density of population of that borough in square miles.
● Density_sqkms: Density of population of that borough in square kms.

The third data source is the 'list of neighbourhoods in New York City boroughs' as found on Coursera previous lab session. This dataset is created from scratch using the list of neighbourhood available on the site (https://geo.nyu.edu/catalog/nyu_2451_34572), the following are columns:

- Neighbourhood: Name of the neighborhood in the Borough.
- Borough: Name of the Borough.
- Latitude: Latitude of the Borough.
- Longitude: Longitude of the Borough.

2.2 Data Cleaning:

The data preparation for each of the three sources of data is done separately. From the New York City 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 the boroughs for each major category (see fig *2.1*).

[54]:

	Borough	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence against Person	Total
0	BRONX	2694	11918	9129	1877	4394	1493	32658	64163
1	BROOKLYN	5560	17063	6362	3006	5699	2487	42427	82604
2	MANHATTAN	2778	11869	5911	2332	3152	2335	26427	54804
3	QUEENS	3581	12183	1667	2402	3267	1879	27252	52231
4	STATEN ISLAND	570	3288	617	785	457	475	7291	13483

Fig 2.1 New York City crime data after preprocessing

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 format as shown in the website. After the web scraping, string manipulation is required to get the names of the boroughs in the correct form (see fig *2.2*). This is important because we will be merging the two datasets together using the Borough names.

```
df
```

72]:

	Borough	County	Population	GDP_billions	GDP_percapita	Area_sqmiles	Area_sqkms	Density_sqmiles	Density_sqkms
0	The Bronx	Bronx	1,418,207	42.695	30,100	42.10	109.04	33,867	13,006
1	Brooklyn	Kings	2,559,903	91.559	35,800	70.82	183.42	36,147	13,957
2	Manhattan	New York	1,628,706	600.244	368,500	22.83	59.13	71,341	27,544
3	Queens	Queens	2,253,858	93.310	41,400	108.53	281.09	20,767	8,018
4	Staten Island	Richmond	476,143	14.514	30,500	58.37	151.18	8,157	3,150

Fig 2.2 List of New York city Boroughs

The two datasets are merged on the Borough names to form a new dataset that combines the necessary information in one dataset (see fig *2.3*). The purpose of this dataset is to visualize the crime rates in each borough and identify the borough with the least crimes recorded during the year 2015.

```
df2
```

[77]:

	Borough	County	Population	GDP_billions	GDP_percapita	Area_sqmiles	Area_sqkms	Density_sqmiles	Density_sqkms	Burglary	Criminal Damage	Drugs	Other Notifiable Offences	Robbery	Theft and Handling	Violence against Person	Total
0	Bronx	Bronx	1,418,207	42.695	30,100	42.10	109.04	33,867	13,006	2694	11918	9129	1877	4394	1493	32658	64163
1	Brooklyn	Kings	2,559,903	91.559	35,800	70.82	183.42	36,147	13,957	5560	17063	6362	3006	5699	2487	42427	82604
2	Manhattan	New York	1,628,706	600.244	368,500	22.83	59.13	71,341	27,544	2778	11869	5911	2332	3152	2335	26427	54804
3	Queens	Queens	2,253,858	93.310	41,400	108.53	281.09	20,767	8,018	3581	12183	1667	2402	3267	1879	27252	52231
4	Staten Island	Richmond	476,143	14.514	30,500	58.37	151.18	8,157	3,150	570	3288	617	785	457	475	7291	13483

Fig 2.3 New York City Borough Crimes

After visualizing the crime in each borough we can find the borough with the lowest crime rate and hence tag that borough as the safest borough. The third source of data is acquired from the list of Neighbourhoods in the safest borough from Coursera previous lab session. This dataset is created from scratch, the pandas data frame is created with the names of the neighbourhoods and the name of the borough with the latitude and longitude.

The coordinates of the neighbourhoods is be obtained using Python Geocoder package to get the final dataset (See Fig *2.4*).

```
neigh
```

[89]:

	Borough	Neighbourhood	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
...	...	...	...	...

Fig 2.4 Neighbourhoods of the safest borough

The new dataset 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

3.1 Exploratory Data Analysis:

Statistical summary of crimes - The describe function in python is used to get statistics of the New York City crime data, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%) for each of the major categories of crime (See *fig 3.1.1*).

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

Fig 3.1.1 Statistical description of the New York City crimes

The count for each of the major categories of crime returns the value 5 which is the number of New York City boroughs. ‘Violence against the person’ is the highest reported crime during the year 2015 followed by ‘Criminal damage’. The lowest recorded crimes are ‘Theft and Handling’, ‘Other Notifiable offenses’ and ‘Burglary’.

Boroughs with the highest crime rates

Comparing five boroughs with the highest crime rate 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 (see *fig 3.1.2*).

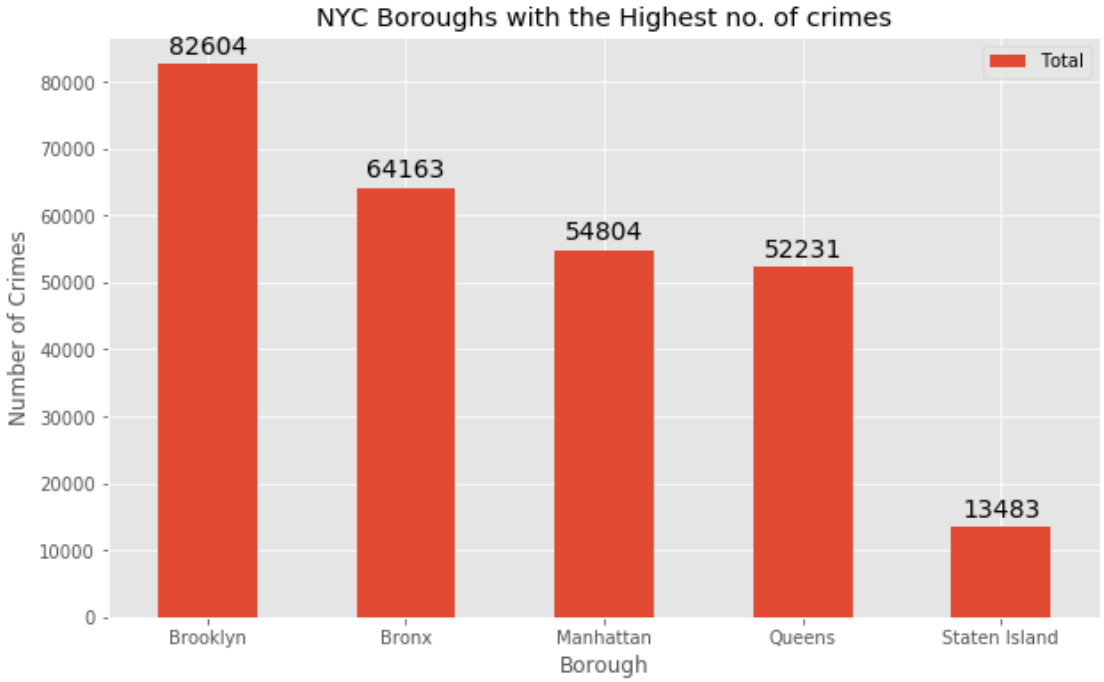

Fig 3.1.2 Boroughs with the highest crime rates

Neighbourhoods in Staten Island:

There are neighbourhoods in the borough of Staten Island, they are visualized on a map using folium on python (see *fig 3.1.3*).

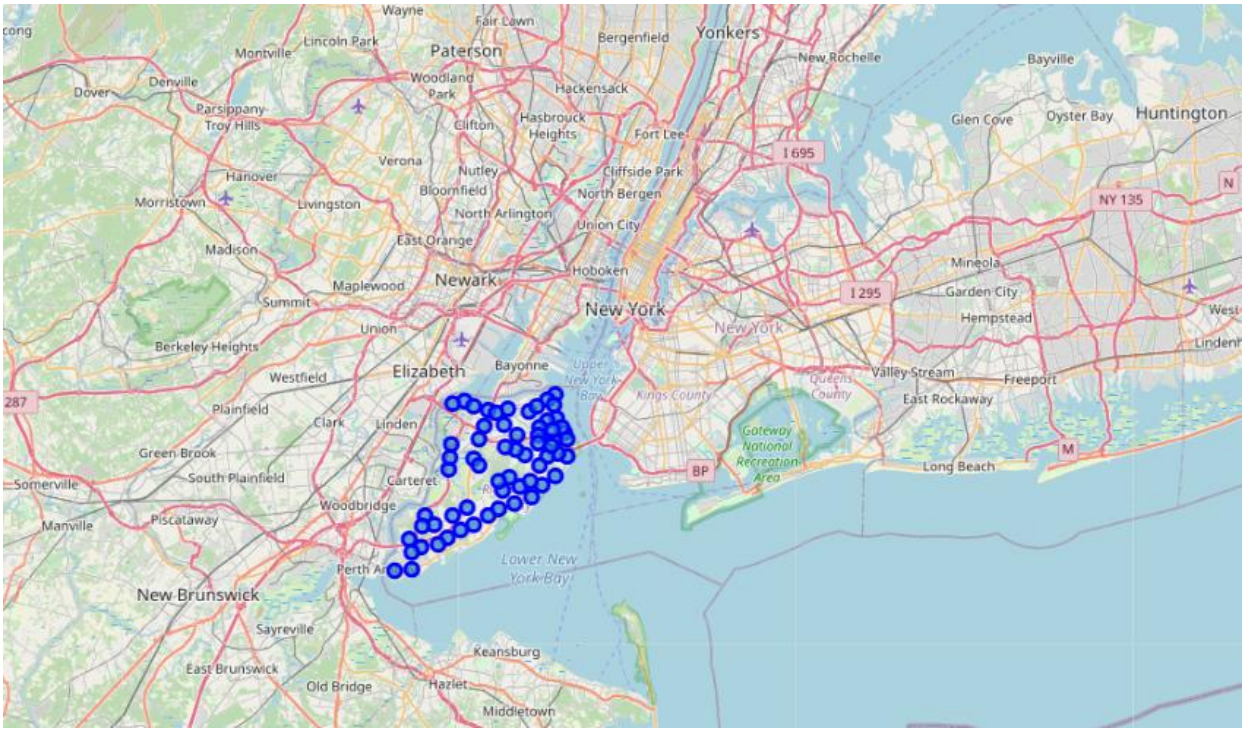

Fig 3.1.3 Neighbourhoods in Staten Island

3.2 Modelling:

Using the final dataset containing the neighbourhoods in Staten Island along with the latitude and longitude, we can find all the venues within a 500 meter radius of each neighborhood by connecting to the Foursquare API. This returns a json file containing all the venues in each neighbourhood which is converted to a pandas data frame. This data frame contains all the venues along with their coordinates and category (see *fig 3.2.1*).

```
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

Fig 3.2.1 Venue details of each Neighbourhood

One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML

algorithms to do a better job in prediction). 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 data based on predefined cluster size. To find the optimal value for k (Number of clusters), we use Elbow method. We will use a cluster size of 3 for this project 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 interests based on the venues/amenities around each neighbourhood.

Elbow Method:

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.

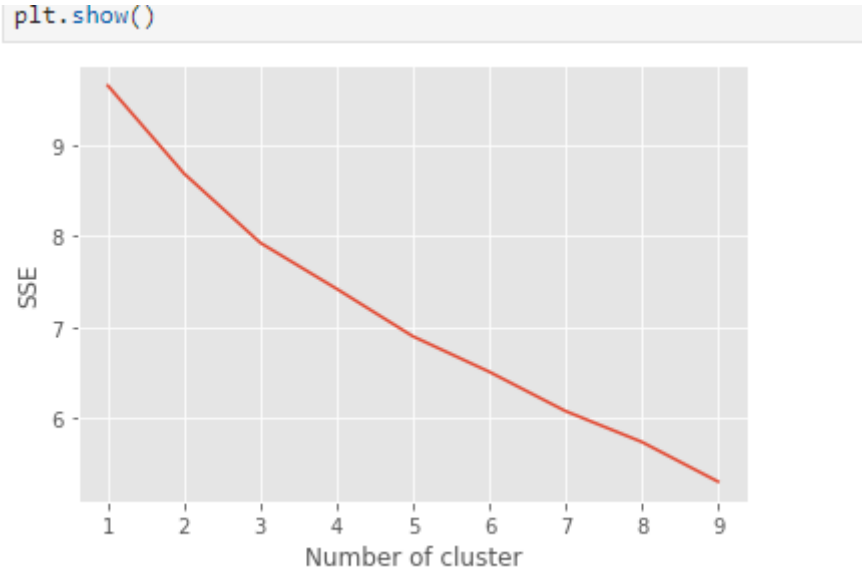

Fig 3.2.2 Elbow method for optimal k value

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**. (*see fig 3.2.2*)

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. Looking into the neighbourhoods in the first cluster (see *fig 4.1*)

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

Fig 4.1 Cluster 1

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

```
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

Fig 4.2 Cluster 2

The second cluster has some neighborhoods which consists of venues such as Bus stops, Deli/Bodega, and Restaurants. (*see fig 4.2*)

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

Fig 4.3 Cluster 3

The third cluster has only three neighborhoods which consists of venues such as Dog Run, Park, and Convenience store. (*see fig 4.3*)

Visualizing the clustered neighbourhoods on a map using the folium library.

Fig 4.4 Clustered neighbourhoods in the Borough of Staten Island

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. (*see fig 4.4*)

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.