Applied Data Science Capstone Project Report

The Battle of Neighborhoods

Coursera - IBM Data Science Professional Certificate

Finding an optimal location for a Supermarket in Brooklyn, New York City

Submitted by:

Somraj Chowdhury

Table of contents

1. Introduction

The New York City (NYC) is the most populous city in the United States and provides a lot of business opportunities. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, advertising, legal services, accountancy, insurance, theater, fashion and the arts in the United States. In the 21st century, New York has emerged as a global node of creativity and entrepreneurship, social tolerance and environment sustainability.

New York City consists of five boroughs, each of which is a separate county of the State of New York.

The five boroughs are:

1. Brooklyn
2. Queens
3. Manhattan
4. The Bronx
5. Staten Island

New York is a highly developed city and so the cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully.

2. Business Problem

2.1 Problem

This project will aim at finding a suitable and optimal location for a Supermarket in Brooklyn, in the City of New York, United States.

2.2 Problem Description:

A supermarket is a self-service shop offering a wide variety of food, beverages and household products, organized into sections and shelves. With the increase in the population, the demand and availability of essential resources like food and other

households increase and supermarkets serve this very purpose of providing the resources under one roof.

2.3 Target Audience

This project is aimed at those individuals and business groups who are interested to invest in opening a supermarket in Brooklyn, New York.

3. Data

In this project we will be exploring the New York City dataset and analyzing the neighborhoods of Brooklyn borough.

3.1 New York City Data

This dataset contains data of the boroughs and the neighborhoods that exist in each borough along with the geographical coordinates of each neighborhood.

This dataset is available for free online.

Link: https://geo.nyu.edu/catalog/nyu_2451_34572

This dataset holds data of a total of 5 boroughs and 306 neighborhoods.

Borough	Number of Neighborhoods
Bronx	52
Brooklyn	70
Manhattan	40
Queens	81
Staten Island	63
Total Neighborhoods	306

The New York dataset has the following features and format.

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

3.2 Supermarket Data from Foursquare API

Information about the existing supermarkets in Brooklyn, New York City like the name of supermarkets in a neighborhood and their location along with geographical coordinates etc. will be obtained from the Foursquare API.

We request the required data by writing an URL along with the client_id, client_secret, version, latitude, longitude, radius, limit and category id for the type of venue you would like to retrieve information in this case it supermarket.

The following screenshot shows a part of the supermarket data retrieved from the Foursquare API.

Supermarket Name	Supermarket Latitude	Supermarket Longitude
Met Fresh Supermarket	40.616528	-74.034003
Metropolitan CityMarket	40.617375	-74.030735
CTown Supermarkets	40.629234	-74.022803
Jmart 新世界超市	40.610080	-74.001221
Foodtown	40.619927	-74.032301
Scaturro Supermarkets	40.629409	-74.005051
Food Dynasty	40.611275	-74.008544

3.2 Population data for each Brooklyn neighborhood

Population dataset for each Brooklyn neighborhood was created by finding the population of each neighborhood individually from the web.

4. Methodology

4.1 Load data and explore NYC neighborhoods

First we download the dataset from the server which is a file with ***.json*** extension.

Then we extract the required data from the ***features key*** of the downloaded json file.

Once all the required NYC neighborhood data is retrieved, transform the json data into a pandas dataframe.

Now our dataframe is created with the following features:

	Borough	Neighborhood	Latitude	Longitude

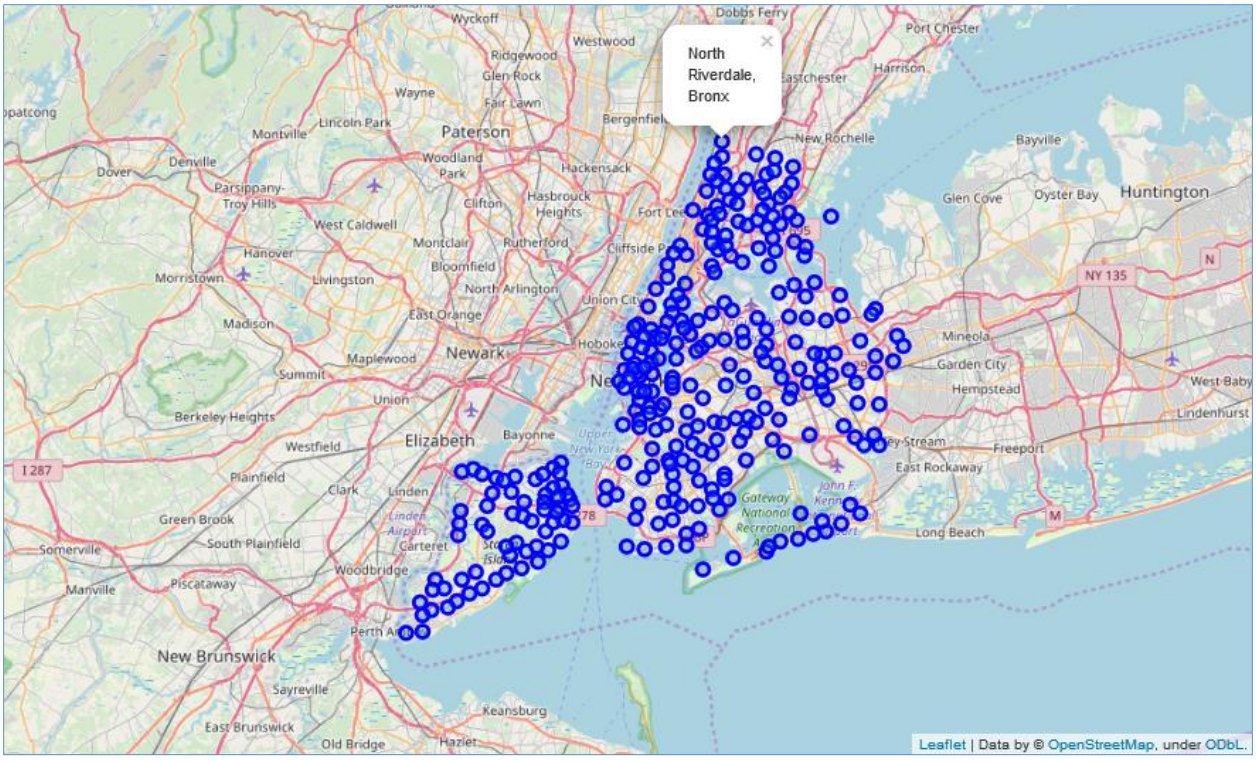
Next we use the ***geopy*** library to locate the coordinates of addresses, cities, countries, and landmarks across the globe using third-party geocoders such as Google Geocoding API (V3) etc. and other data sources.

We obtain the latitude and longitude values of the New York City using the geopy library by giving the address as:

```
address = 'New York, US'
```

We then use the ***folium*** library that helps to visualize data that has been manipulated in Python on an interactive map. It enables both the binding of data to a map as well as passing rich visualizations as markers on the map.

Using the folium library, we then create a map of New York with the ***306*** neighborhoods superimposed on top as markers.



We can click on any of the blue colored circular markers to popup the name of the neighborhood and its respective borough.

Since we want to find a suitable and optimal location for opening a supermarket in the neighborhoods of Brooklyn borough, we create a dataframe that only consists of Brooklyn neighborhoods and all the related data.

4.2 Explore Brooklyn neighborhoods

Firstly, we make a dataframe with only Brooklyn neighborhoods by retrieving only the rows whose borough has the value ***Brooklyn***.

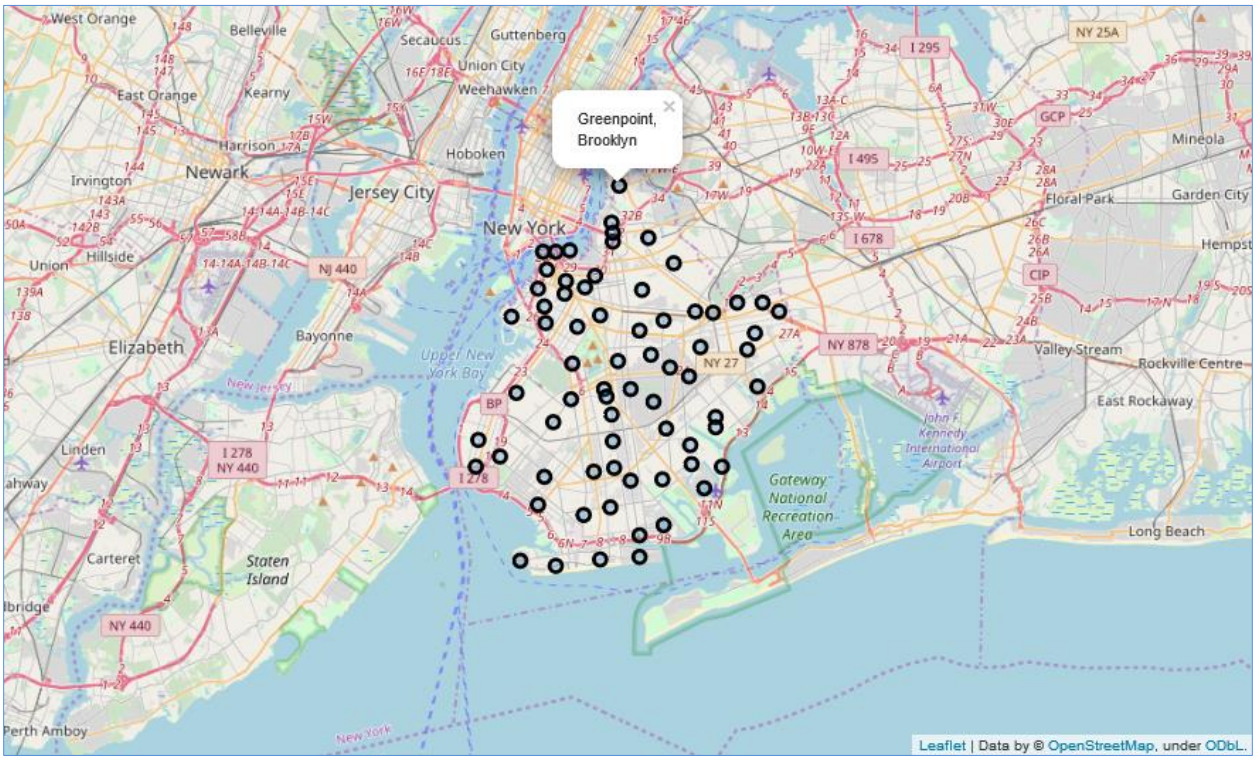
Now the Brooklyn dataframe is created with the following features:

	Borough	Neighborhood	Latitude	Longitude

We obtain the latitude and longitude values of Brooklyn using the geopy library by giving the address as:

```
address = 'Brooklyn, NY'
```

Using the folium library, we then create a map of Brooklyn with the ***70*** neighborhoods superimposed on top.



4.3 Explore supermarkets in Brooklyn using Foursquare API

Once our Brooklyn dataframe is created, we then explore supermarkets located in each of Brooklyn's neighborhoods by leveraging the ***Foursquare API***.

In order to use the Foursquare services, you need to create a free developer account first. Once created, you are provided with unique ***CLIENT_ID*** and ***CLIENT_SECRET*** which will be required to request venue data in your application and very importantly that shouldn't be shared with anyone else.

In order to retrieve venue specific data for example data about restaurants, schools etc. you need to specify a specific ***category ID*** associated with it. In this problem, we want data related to supermarkets in Brooklyn; therefore we select the supermarket category,

All venue categories can be found in the official Foursquare website.

Link: https://developer.foursquare.com/docs/resources/categories

Once you have all the information for retrieving venue data from Foursquare, the next thing you should do is request data from Foursquare API using a ***URL*** specific to your requirements.

The data returned by the Foursquare API will be in ***json*** format and all the important information that we require about the venue (supermarket) is stored in the ***items*** key of the retrieved json data.

In this capstone project, we have retrieved information about all supermarkets within ***4km*** radius of the Brooklyn borough.

The retrieved supermarket data from Foursquare API is then transformed into a pandas dataframe.

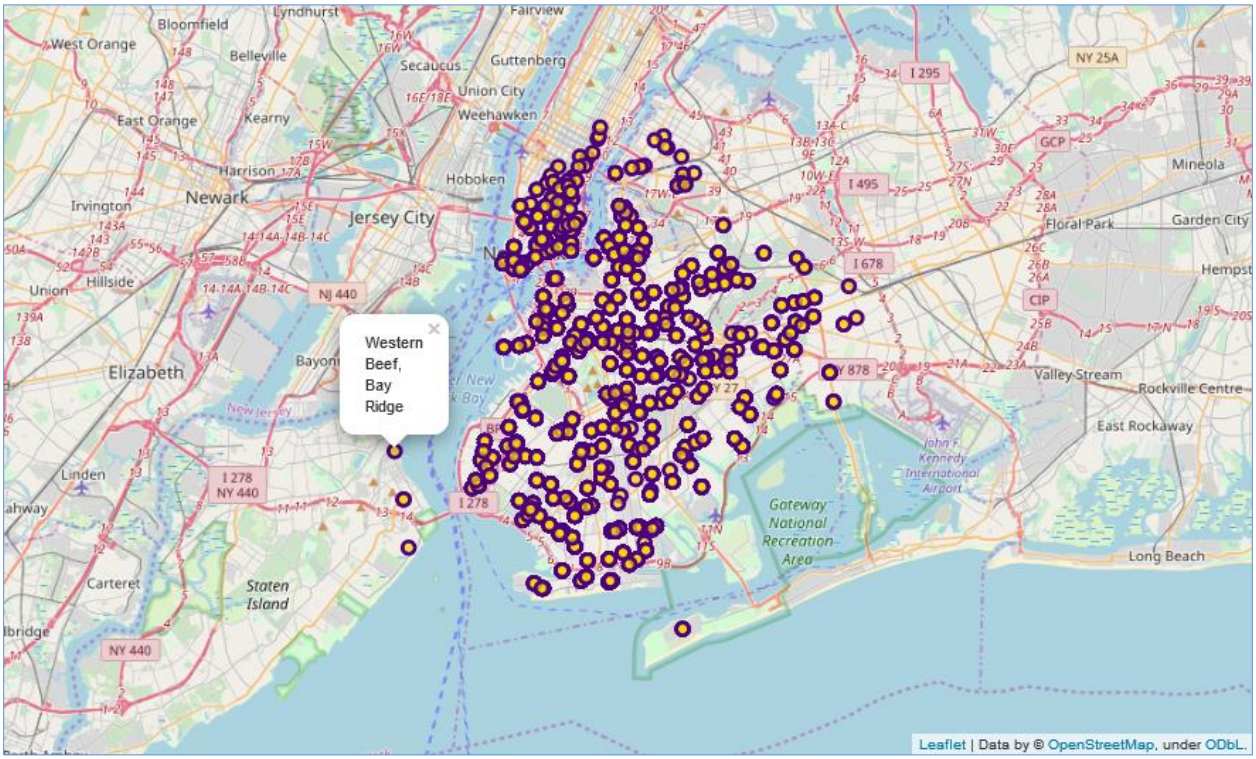
Supermarket Name	Supermarket Latitude	Supermarket Longitude
Met Fresh Supermarket	40.616528	-74.034003
Metropolitan CityMarket	40.617375	-74.030735
CTown Supermarkets	40.629234	-74.022803
Jmart 新世界超市	40.610080	-74.001221
Foodtown	40.619927	-74.032301
Scaturro Supermarkets	40.629409	-74.005051
Food Dynasty	40.611275	-74.008544
Bingo Wholesale	40.629211	-73.999403
Key Food Fresh & Natural	40.628058	-74.029126
Fei Long Market	40.633397	-74.011286

This supermarket data is added to the Brooklyn neighborhood data so as to know which supermarket is in which neighborhood.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Supermarket Name	Supermarket Latitude	Supermarket Longitude
0	Bay Ridge	40.625801	-74.030621	Met Fresh Supermarket	40.616528	-74.034003
1	Bay Ridge	40.625801	-74.030621	Metropolitan CityMarket	40.617375	-74.030735
2	Bay Ridge	40.625801	-74.030621	CTown Supermarkets	40.629234	-74.022803
3	Bay Ridge	40.625801	-74.030621	Jmart 新世界超市	40.610080	-74.001221
4	Bay Ridge	40.625801	-74.030621	Foodtown	40.619927	-74.032301
5	Bay Ridge	40.625801	-74.030621	Scaturro Supermarkets	40.629409	-74.005051
6	Bay Ridge	40.625801	-74.030621	Food Dynasty	40.611275	-74.008544
7	Bay Ridge	40.625801	-74.030621	Bingo Wholesale	40.629211	-73.999403
8	Bay Ridge	40.625801	-74.030621	Key Food Fresh & Natural	40.628058	-74.029126
9	Bay Ridge	40.625801	-74.030621	Fei Long Market	40.633397	-74.011286

At this stage the Brooklyn supermarket data is properly compiled and formatted into one dataframe where each supermarket is provided with information like its latitude and longitude (geographical) coordinates and in which neighborhood it is located.

Using the folium library, we then create a map of Brooklyn with the all supermarkets in the neighborhoods superimposed on top.

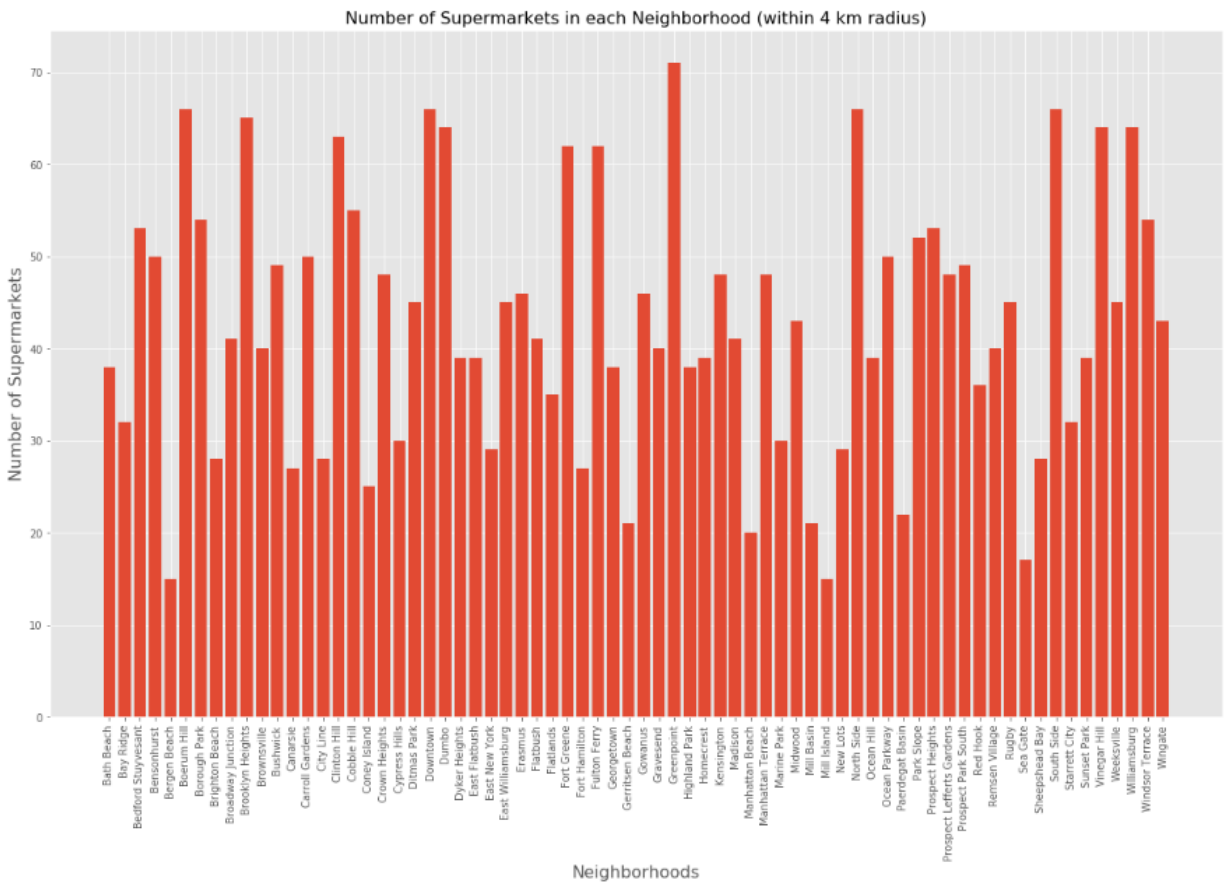


We can click on any of the circular markers to popup the name of the supermarket and the name of neighborhood where it is located.

Then we find the ***total number of supermarkets*** in each neighborhood.

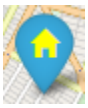
	Neighborhood	Number of Supermarkets
0	Bath Beach	38
1	Bay Ridge	32
2	Bedford Stuyvesant	53
3	Bensonhurst	50
4	Bergen Beach	15
...	...	...

We use the ***matplotlib*** library to visualize the number of supermarkets in each neighborhood.

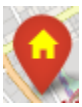


Then we create a visualization using folium where we group the neighborhoods based on the number of supermarkets present in that neighborhood and distinguish each group of neighborhoods using colored folium markers.

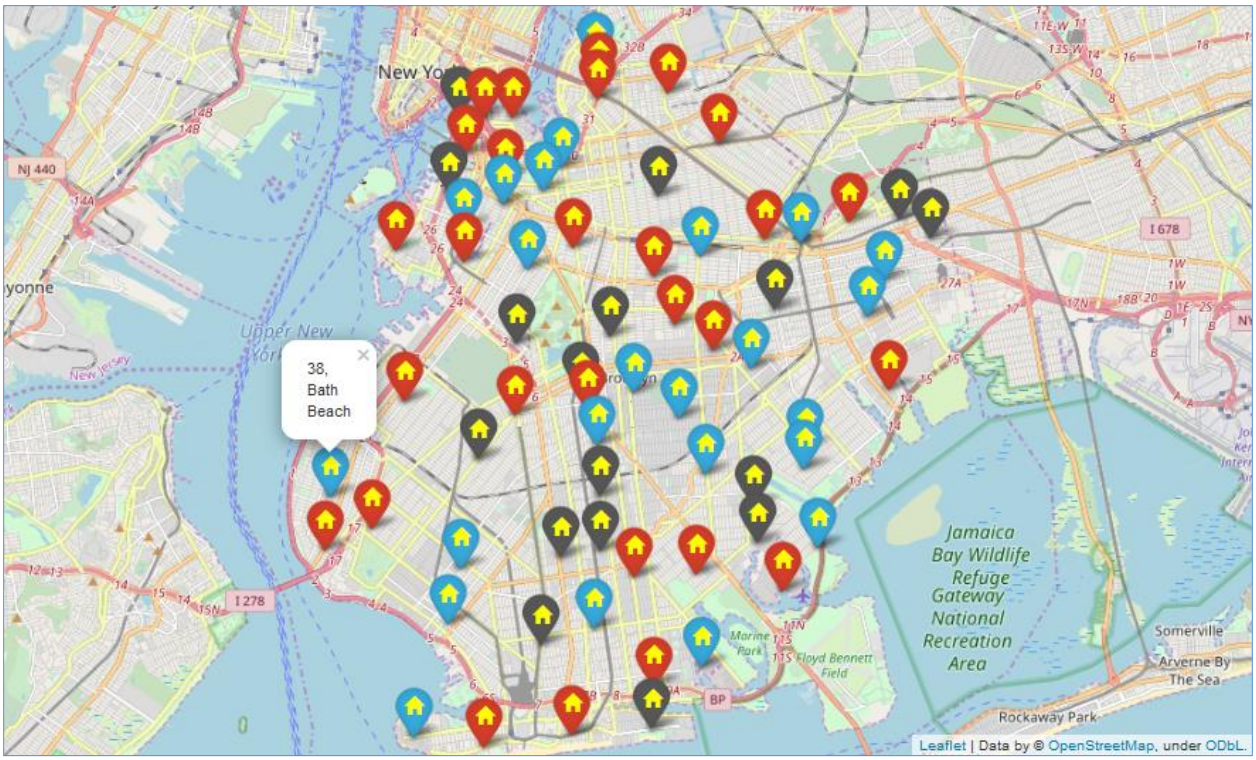
Neighborhoods with ***15 - 30 supermarkets*** are indicated by **gray** markers.



Neighborhoods with ***31 - 45 supermarkets*** are indicated by **blue** markers.



Neighborhoods with ***46 or more supermarkets*** are indicated by **red** markers.



We can click on any of the colored markers to popup the number of supermarkets in a neighborhood and the name of that neighborhood.

4.4 Population and portion data of each Brooklyn neighborhood

After determining number of supermarkets in each neighborhood, we then add the population data for each of the Brooklyn neighborhood along with a feature specifying in which portion of Brooklyn borough the neighborhood resides in.

After adding the all the required data we create a dataframe which looks like this:

	Neighborhood	Portion of Borough	Population	Number of Supermarkets
0	Bath Beach	Southwestern Brooklyn	29931	38
1	Bay Ridge	Southwestern Brooklyn	79371	32
2	Bedford Stuyvesant	Northern Brooklyn	158000	53
3	Bensonhurst	Southwestern Brooklyn	152000	50
4	Bergen Beach	Southern Brooklyn	45231	15

We can divide the neighborhoods in Brooklyn borough into ***6*** unique portions:

1. Central Brooklyn
2. Eastern Brooklyn
3. Northern Brooklyn
4. Northwestern Brooklyn
5. Southern Brooklyn
6. Southwestern Brooklyn

Therefore we create 6 separate dataframes for each portion of the borough and perform data analysis in each individual portion.

5. Results

After partitioning Brooklyn borough into 6 portions, we have made the following observations in terms of the **total number of supermarkets** and the **total population** in each borough portion and a ***score*** that will help in finding an optimal location to open a supermarket.

Portion 1: *Central Brooklyn*

```
Number of neighborhoods in Central Brooklyn: 15
Total number of supermarkets in Central Brooklyn: 684
Total population in Central Brooklyn: 1197391
Score: 1751
```

Portion 2: *Eastern Brooklyn*

```
Number of neighborhoods in Eastern Brooklyn: 8
Total number of supermarkets in Eastern Brooklyn: 253
Total population in Eastern Brooklyn: 450720
Score: 1782
```

Portion 3: *Northern Brooklyn*

```
Number of neighborhoods in Northern Brooklyn: 7
Total number of supermarkets in Northern Brooklyn: 387
Total population in Northern Brooklyn: 706258
Score: 1825
```

Portion 4: *Northwestern Brooklyn*

```
Number of neighborhoods in Northwestern Brooklyn: 15
Total number of supermarkets in Northwestern Brooklyn: 845
Total population in Northwestern Brooklyn: 271829
Score: 322
```

Portion 5: *Southern Brooklyn*

```
Number of neighborhoods in Southern Brooklyn: 17
Total number of supermarkets in Southern Brooklyn: 524
Total population in Southern Brooklyn: 959338
Score: 1831
```

Portion 6: *Southwestern Brooklyn*

```
Number of neighborhoods in Southwestern Brooklyn: 8
Total number of supermarkets in Southwestern Brooklyn: 304
Total population in Southwestern Brooklyn: 589085
Score: 1938
```

6. Discussion

In the results section above, we calculated scores for all 6 portions of the borough. This score is calculated by dividing the ***total population in a portion*** by the ***total number of supermarkets in that portion***.

$$score = \frac{Total\ population}{Total\ number\ of\ supermarkets}$$

What is the significance of this score?

- A lower score indicates that there is sufficient number of supermarkets for the present population.
- A higher score signifies that there is lesser number of supermarkets compared to the population of the portion.

Based on the analysis of each portion, we note that:

- **Portion 1 - Central Brooklyn** has the highest population of 1.2 million (approx.) which means a higher chance of customer expectancy.

- **Portion 4 - Northwestern Brooklyn** has the lowest score which means there is more business and competition among the supermarkets.

- **Portion 6 - Southwestern Brooklyn** has the highest score which means there is a deficiency in the number of supermarkets.

Therefore, the individual or business group interested in investing in opening a supermarket in the Brooklyn borough can invest in the Northwestern Brooklyn where there is much more competition in the supermarket business or can invest in Central Brooklyn where there is the highest population of all the 6 portions.

7. Conclusion

The results of this capstone project are obtained by performing analysis on limited data. The purpose of this project was to find an optimal and suitable location for opening a supermarket in the Brooklyn borough in order to aid stakeholders with best locations that could benefit the business and the growth of the brand. If a neighborhood or location has a high density of supermarkets, then it means that there is a high demand in those locations. Also locations with larger population will need more supermarkets in order to cater their needs. The final decision on selection of an optimal location will be taken by the stakeholders based on characteristics such as popularity of the location and proximity to major roads etc.