A) Introduction

The purpose of this project is to provide a template to identify rental properties in a specific neighborhood, within the budget constraints and preference of the individual(s).

Finding a rental property with one's preferences and constraints can be very time comsuming task, requiring a lot of effort. An individual may be looking for a rental space for different purposes, and, within different budgets like for work or educational purposes. The project can help find tailored options as per these criteria for the individuals, and, help save a lot of time and effort.

Through this template one can not only find the affordable rental properties but also make sure to keep proximity to nearby places-of-interest in mind, to narrow down to few options meeting the requirements.

This project can be used by both individuals well as by real estate agents to skim through the rental properties as per the requirements.

B) Data Description

For this project, multiple datasets are being used to build a holistic view of the rental properties and nearby places. The datasets being used are sourced from different sources:

1. **Rental properties data:**

   Details

   This data has the list of rental properties with additional information in the specific neighborhood.

   Source:

   https://www.trulia.com/rent/

   Data:

   Borough, neighborhood and property specific details like property address (street address), price (in USD), size (in sqft), number of bedrooms and bathrooms.

2. **New York City data:**

   Details

   This data has list of booughs and neighborhood along with the geo-coordinates.

   Source

   https://data.cityofnewyork.us/City-Government/NeighborhoodNames-GIS/99bc-9p23

   Data:

   Borough, neighborhood and coordinates of each neighborhood in terms of latitude and longitude.

3. **Foursquare data:**

Details

This data contains nearby places of interest, in a specific neighborhood, with their distance from the centroid.

Source

https://foursquare.com/

Data:

Neighborhood, venue category, venue latitude, venue longitude and venue name.

C) Data Cleaning

There are different sources of data as mentioned above. So it was difficult to bring all information together and extract the relevant information from each neighborhood.

- **New York City dataset**

Imported the NhoodNameCentroids.csv file using pandas. There were 5 different boroughs in New York City i.e Bronx, Manhattan, Brooklyn, Queens and Staten Island. Filteration process is used to filter the Brooklyn Borough, renamed the name Name column to Neighborhood and dropping **AnnoAngle**, **AnnoLine1**, **AnnoLine2**, **AnnoLine3**, **Stacked**, **OBJECTID**. Next step is to splitting up the **the_geom** column into latitude and longitude which were the coordinates for each neighborhood in Brooklyn Borough.

- **Trulia Real Estate data**

After scrapping the data is saved as brooklyn.csv. After importing this file into pandas I noticed that there were many problems with this dataset like Price, Bedrooms and Bathrooms column is in object format, NaN values in Size column and it also in object format.

1. Dropping the rows that contains the text Studio Bedrooms replace the suffix bd with blank string.
2. Replaces the prefix /mo with blank string in Price column and replaces the prefix $ with blank as well.
3. Replaces the suffix ba in Bathrooms column.
4. Dropping the NaN values from the whole dataset.
5. Merged the this scrapped dataset with the New York dataset on the basis of Borough columns.

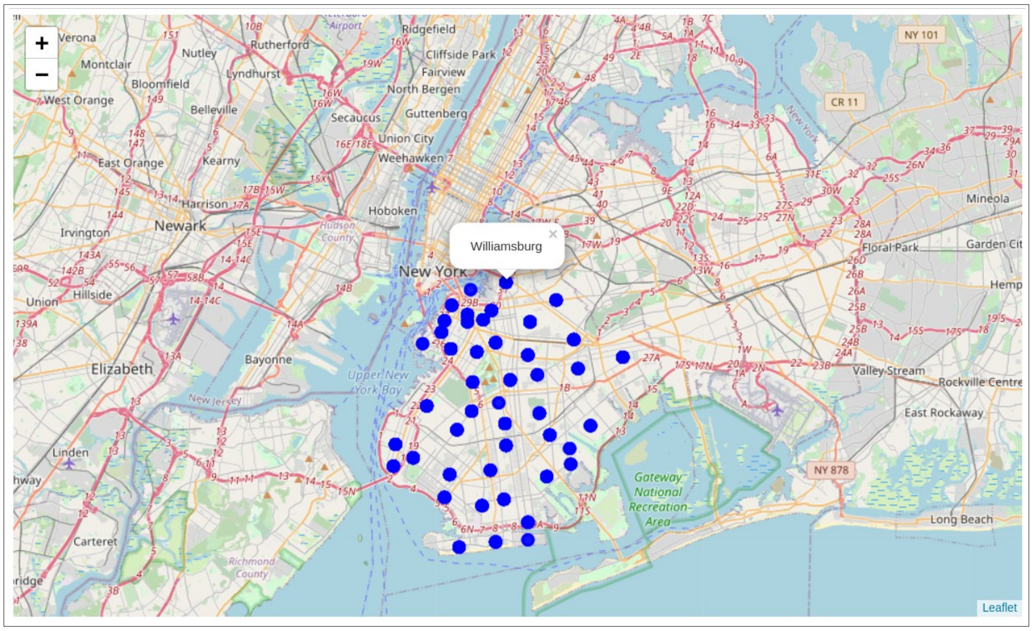
D) Feature Selection

So after cleaning the data there are 1629 observations and 10 attributes.

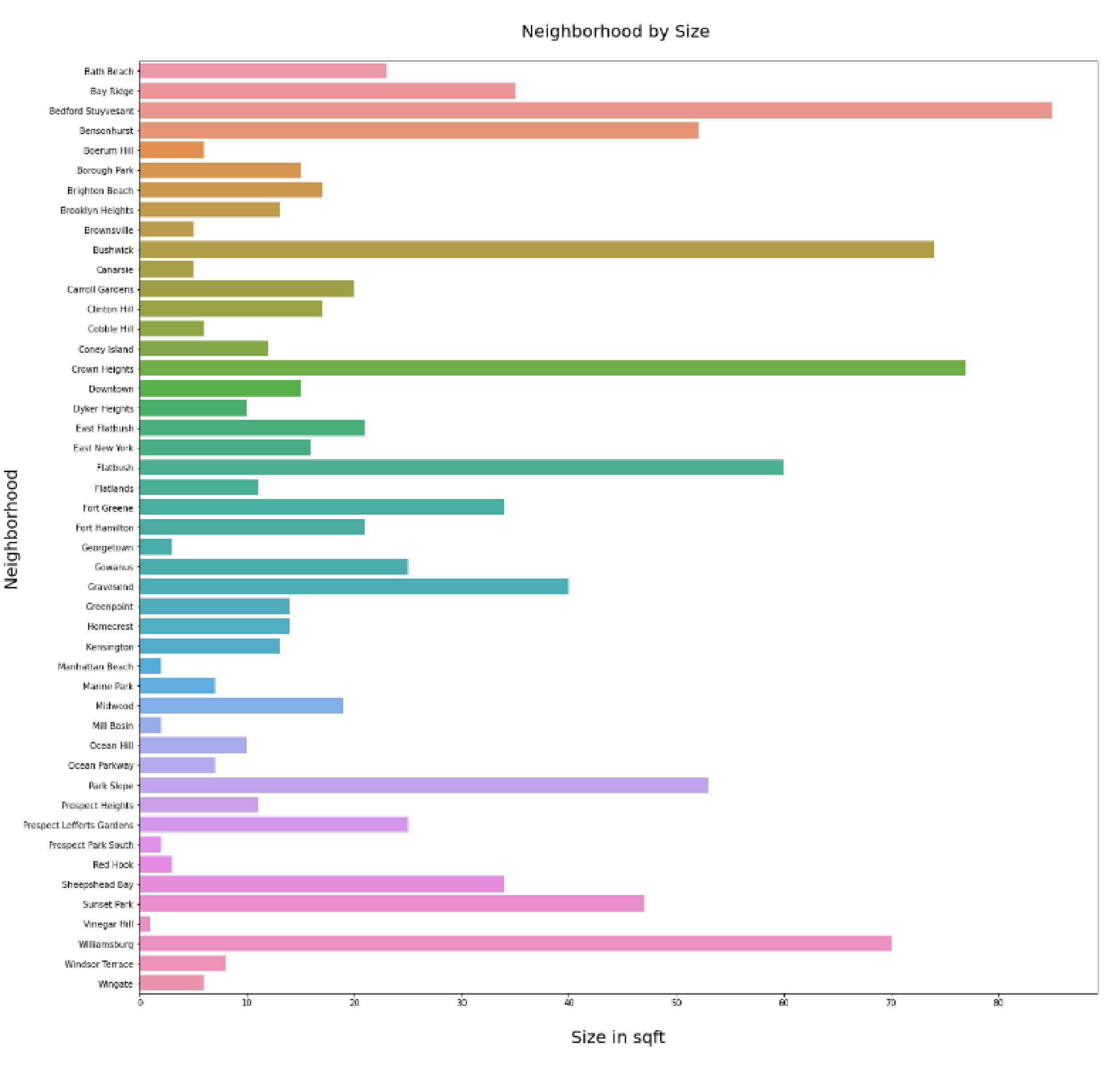
By using groupby function on the Neigborhood column of a scrapped data there will be a new dataset with 47 Unique Neighborhood and their Latitude and Longitude.

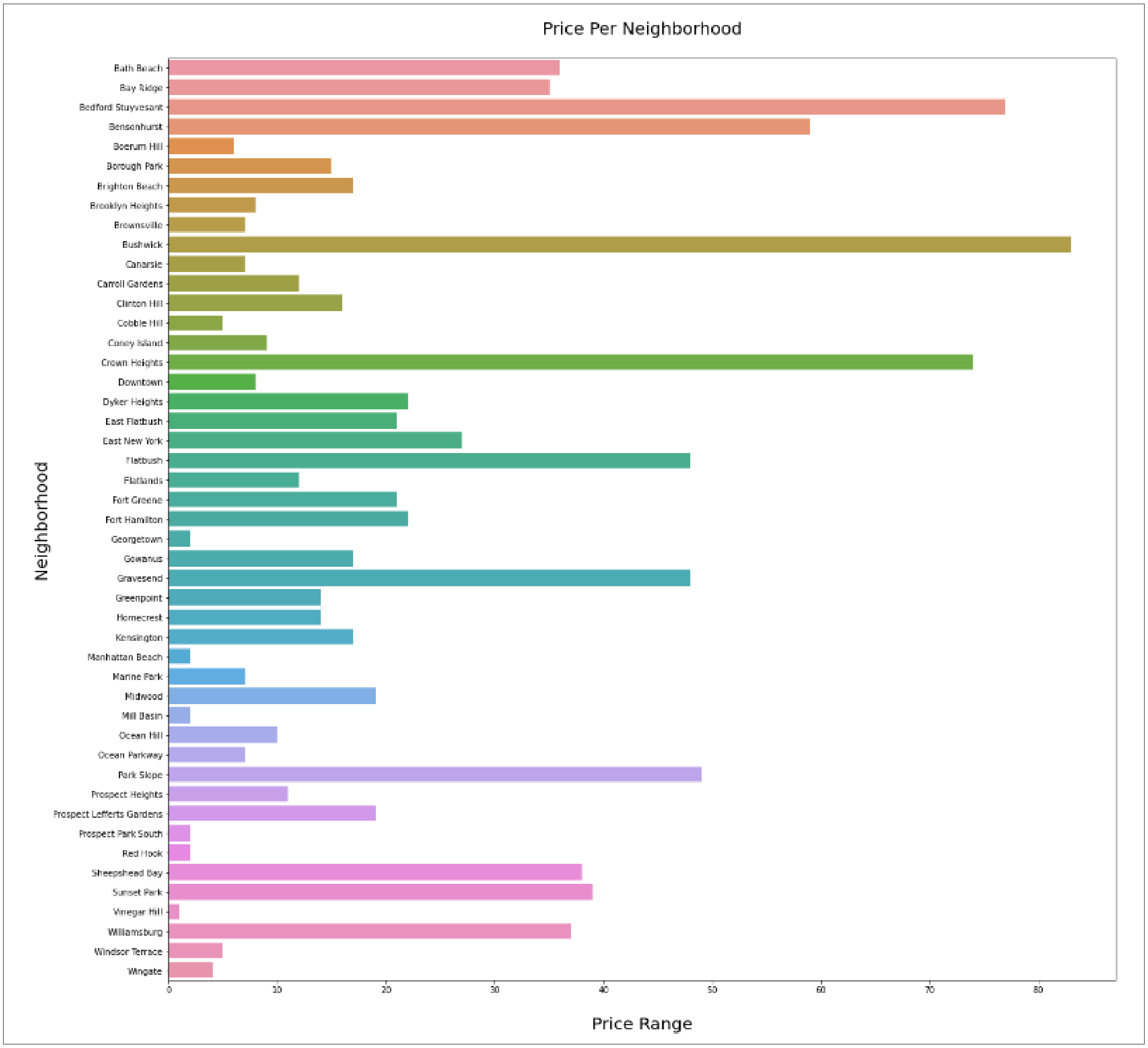
E) Exploratory Data Analysis

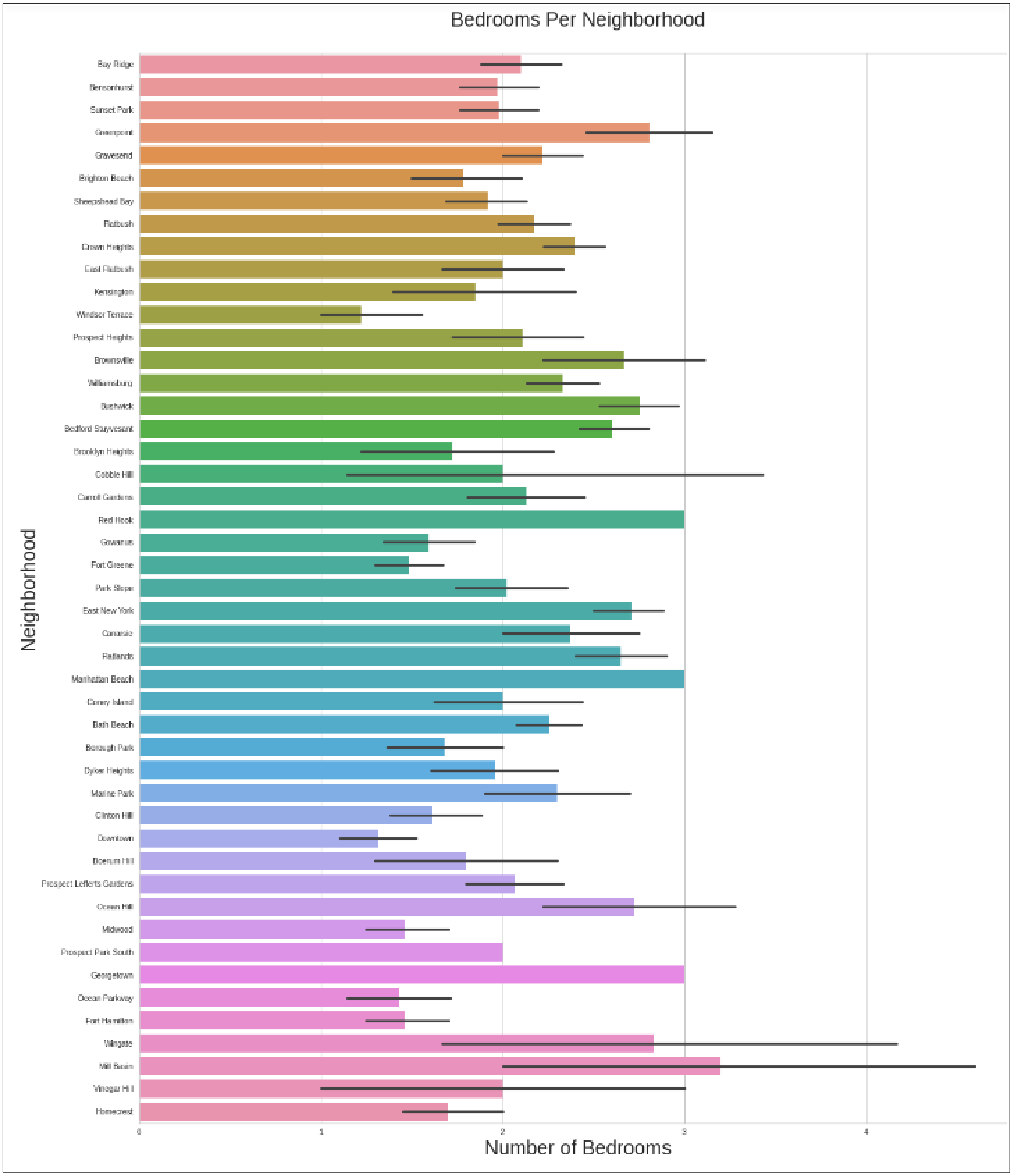
By displaying all 47 unique Neighborhood using folium library, it will looks like:

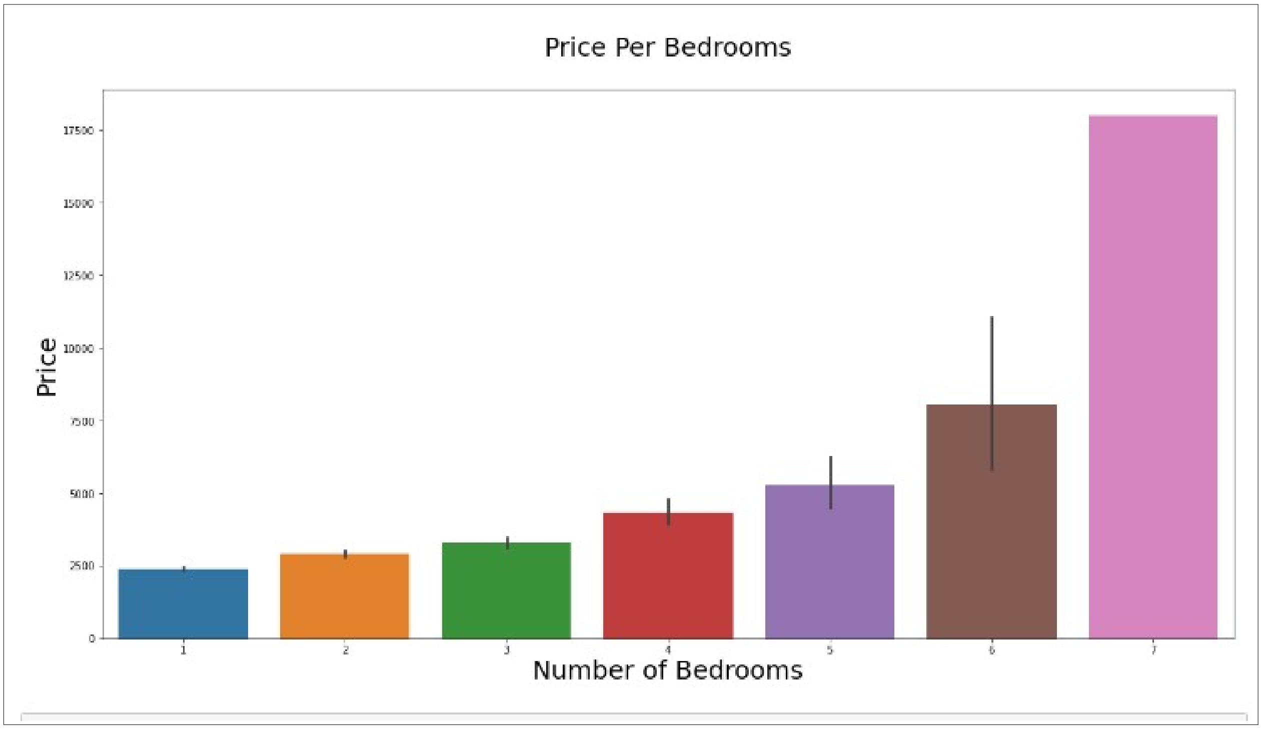


Few data exploration charts are given below:


Neighborhood by Size
Neighborhood
Bath Beach
Bay Ridge
Bedford Stuyvesant
Bensonhurst
Boerum Hill
Borough Park
Brighton Beach
Brooklyn Heights
Brownsville
Bushwick
Canarsie
Carroll Gardens
Clinton Hill
Cobble Hill
Coney Island
Crown Heights
Downtown
Dyker Heights
East Flatbush
East New York
Flatbush
Flatlands
Fort Greene
Fort Hamilton
Georgetown
Gowanus
Gravesend
Greenpoint
Homecrest
Kensington
Manhattan Beach
Marine Park
Midwood
Mill Basin
Ocean Hill
Ocean Parkway
Park Slope
Prospect Heights
Prospect Lefferts Gardens
Prospect Park South
Red Hook
Sheepshead Bay
Sunset Park
Vinegar Hill
Williamsburg
Windsor Terrace
Wingate
0
10
20
30
40
50
60
70
80
Size in sqft


Price Per Neighborhood
Neighborhood
Bath Beach
Bay Ridge
Bedford Stuyvesant
Bensonhurst
Boerum Hill
Borough Park
Brighton Beach
Brooklyn Heights
Brownsville
Bushwick
Canarsie
Carroll Gardens
Clinton Hill
Cobble Hill
Coney Island
Crown Heights
Downtown
Dyker Heights
East Flatbush
East New York
Flatbush
Flatlands
Fort Greene
Fort Hamilton
Georgetown
Gowanus
Gravesend
Greenpoint
Homecrest
Kensington
Manhattan Beach
Marine Park
Midwood
Mill Basin
Ocean Hill
Ocean Parkway
Park Slope
Prospect Heights
Prospect Lefferts Gardens
Prospect Park South
Red Hook
Sheepshead Bay
Sunset Park
Vinegar Hill
Williamsburg
Windsor Terrace
Wingate
0
10
20
30
40
50
60
70
80
Price Range


Bedrooms Per Neighborhood
Neighborhood
Bay Ridge
Bensonhurst
Sunset Park
Greenpoint
Gravesend
Brighton Beach
Sheepshead Bay
Flatbush
Crown Heights
East Flatbush
Kensington
Windsor Terrace
Prospect Heights
Brownsville
Williamsburg
Bushwick
Bedford Stuyvesant
Brooklyn Heights
Cobble Hill
Carroll Gardens
Red Hook
Gowanus
Fort Greene
Park Slope
East New York
Canarsie
Flatlands
Manhattan Beach
Coney Island
Bath Beach
Borough Park
Dyker Heights
Marine Park
Clinton Hill
Downtown
Boerum Hill
Prospect Lefferts Gardens
Ocean Hill
Midwood
Prospect Park South
Georgetown
Ocean Parkway
Fort Hamilton
Wingate
Mill Basin
Vinegar Hill
Homecrest
0
1
2
3
4
Number of Bedrooms



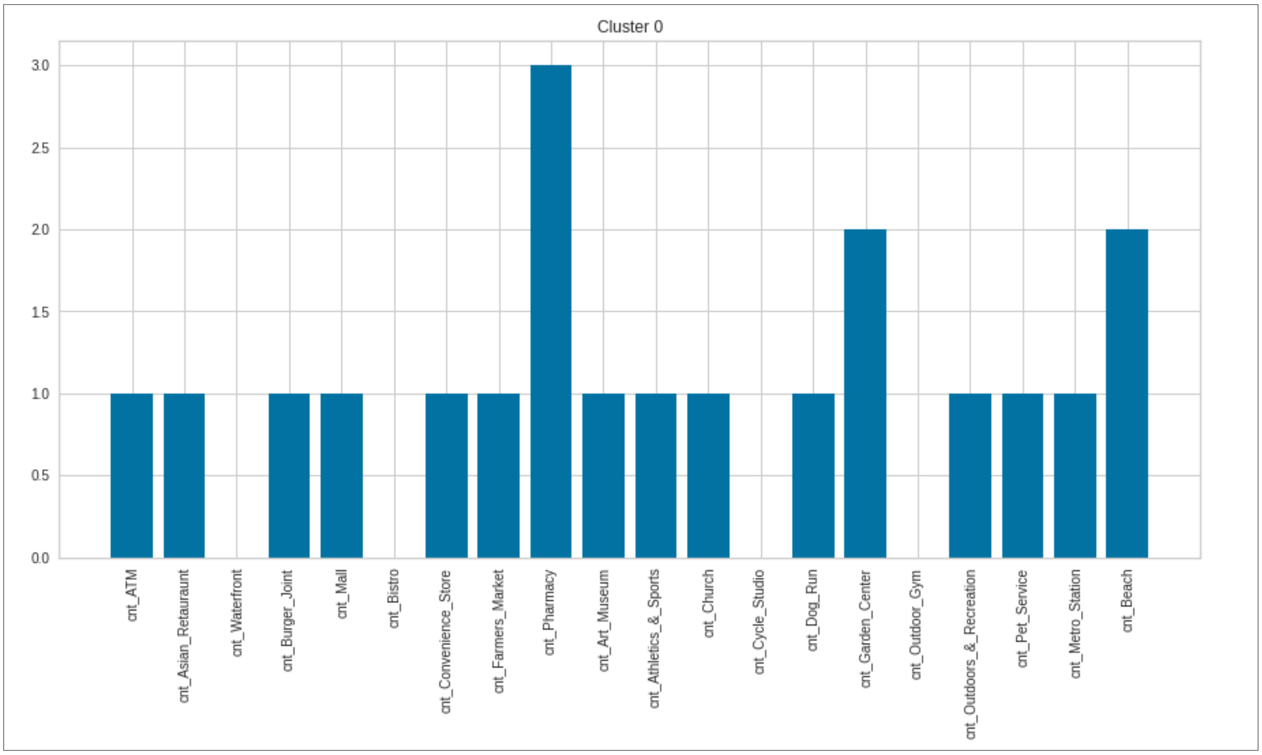
F) Clustering

To group properties of similar nature into a single cluster. For this we use Kmeans clustering algorithm. The algorithm looks at input features like Number of ATMs, Number of Gyms etc. Closer to the properties and , groups properties with similar features.

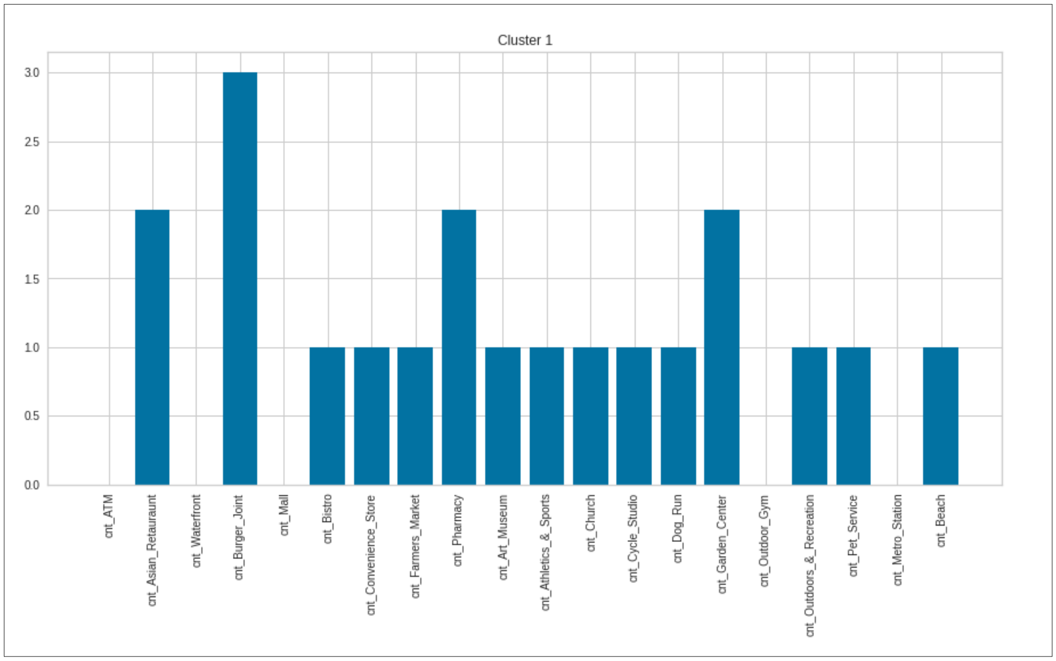
I chose 3 clusters to segment the dataset.

I plotted the graph for all three clusters for some important venues as shown below:

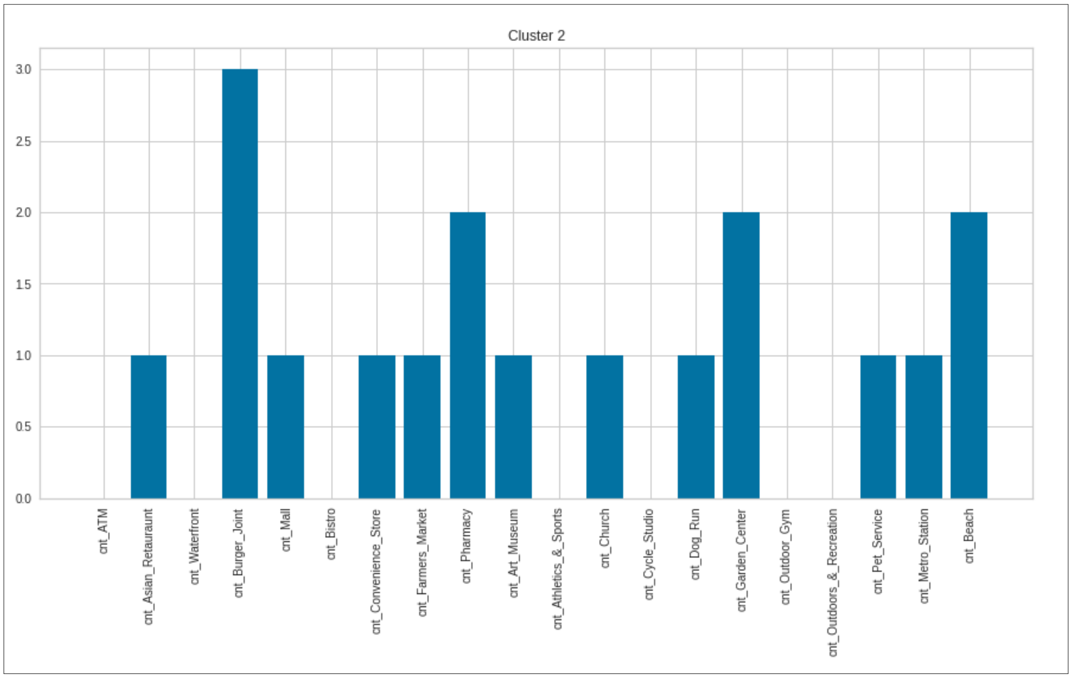
Cluster 0



Cluster 1



Cluster 2

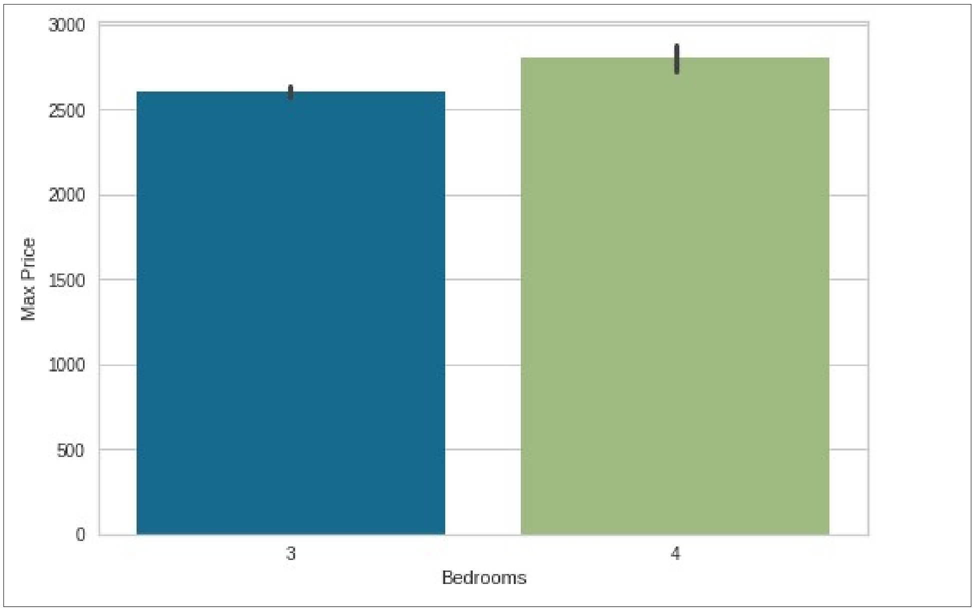


According to above three clusters graphs, I conclude that:

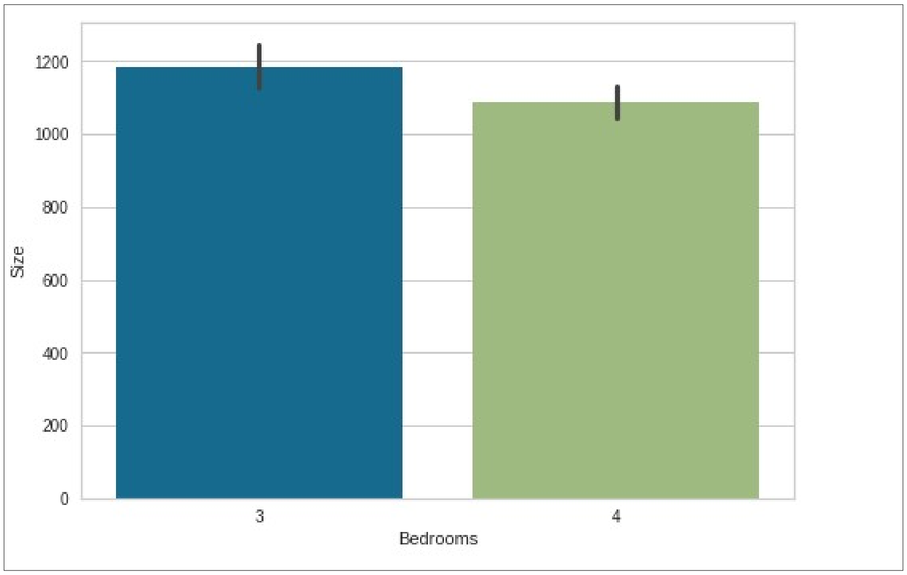
We should only look for properties in **Cluster 0** due to their close proximity to:

1. shopping places
2. metro
3. financial services
4. recreational activities

Further deep-diving into the observations into cluster 0 we got the following insights.

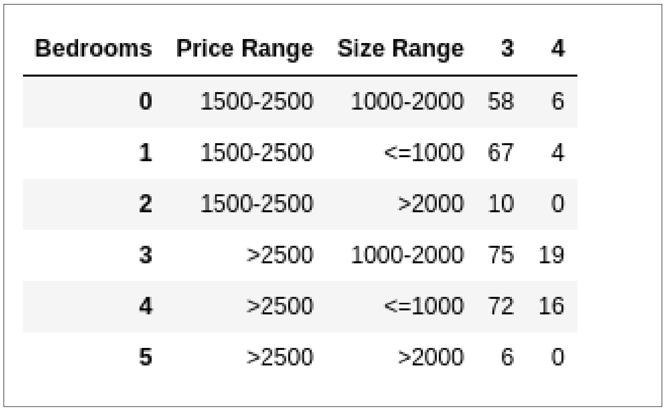


So after analyzing the above graphs, we came to know that there are 3 bedrooms properies within the range of USD 2500 and 4 bedroom properties around USD 3000.



So after analyzing the above graphs 2, we came to know that there are 3 bedrooms properies with 1200 sqft and 4 bedroom properties with 1500 sqft.

On further summarzing the data we get the following view:



Bedrooms	Price Range	Size Range	3	4
0	1500-2500	1000-2000	58	6
1	1500-2500	<=1000	67	4
2	1500-2500	>2000	10	0
3	>2500	1000-2000	75	19
4	>2500	<=1000	72	16
5	>2500	>2000	6	0

1. Based on the properties in Cluster 0, we get properties in specific price and size range with 3 to 4 bedrooms.
2. These properties are in the price range of: USD 1500-2000 and USD 2500-3500.
3. In these price ranges, we have options for both both 3 and 4 bedroom properties.
4. Though the number of bedrooms differ in properties, but there are options within same price range with different number of bedrooms.
5. These options can help the individual to easily narrow down few proeperties as per their preferences/ constraints and prioritize visits to these properties.

G) Key take-away from clustering exercise and comparing the clusters

Based on the clustering exercise, we are able to narrow down to the properties in Cluster 0 because of the following reasons:

- Close proximity to finacial services – **ATMs**
- Quick access to daily need stores - **Convenience stores and Malls**
- Good transporation connectivity - **Metro Station**
- Fitness and Relaxation - Oudoor and Recreational places