ANALYSING NEW YORK APARTMENTS SALES DATASET AND CLUSTERING THE CITIES USING KMEANS CLUSTERING

INTRODUCTION

In this project we will help people who are all looking to buy an apartment in the New York City.

- They can choose to live in residential or commercial areas and can see which borough is best
- which BOROUGH has the highest sales and which has the lowest sales
- which BOROUGH has the highest sale price for the apartment and which is having the lowest sale price for the apartment
- One can able to find the advantages and disadvantages of the Neighborhoods by exploring the dataset.
- One can able to find the appropriate price for the square feet in different cities and boroughs.
- By using kmeans clustering and the Four Squared location data set we can cluster the cities in the Borough.

PROBLEMS

- If somebody wants to buy an apartment in New York, they don't know where to buy that apartment in a cheap and effective manner
- They don't know the price per square feet of the apartment
- They don't know about their neighborhoods and they don't know the advantages and disadvantages of the cities where they are going to buy

DATA ACQUISITION AND CLEANING

Data Sources

NEW YORK CITY NEIGHBOURHOODS DATASET

This city has total of 5 boroughs and 306 Neighborhoods. In order to segment the Neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the Neighborhoods that exists in each borough as well as the latitude and longitude coordinates of each Neighborhood. The dataset used in this project has been acquired from kaggle dataset

This dataset is a record of every building or apartment sold in a New York city property market in 2017 over a period of 12 months.

DATASET:

This contains the location, address, type, sales price and sale date of building units sold

- BOROUGH: A digit code for the borough the property is located in; in order these are Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5).
- BLOCK; LOT: The combination of borough, block, and lot forms a unique key for property in New York City. Commonly called a BBL.
- BUILDING CLASS AT PRESENT and BUILDING CLASS AT TIME OF SALE: The type of building at various points in time. See the glossary linked to below.

Data Cleaning

The dataset consists of 84548 rows and 22 columns. The first step is to check for those attributes which are helpful for our data analysis

Here we are considering only 8 columns out of the 22 columns and the 8 columns are as follows:

- Zip code
- Borough
- Neighborhood
- Building Category

- Residential Unit
- Building Class
- Land Square Feet
- Sales Price

Since the column names have spaces in it for our convenient, we are renaming our column names. They are as follows

- ZIP CODE- ZIP_CODE
- LAND SQUARE FEET – SIZE
- SALE PRICE – PRICE
- BUILDING CLASS CATEGORY - BUILDING_CATEGORY
- RESIDENTIAL UNITS - RESIDENTIAL_UNITS
- BUILDING CLASS AT A TIME OF SALES – BUILDING CLASS

Now we are checking for the types of the attributes using the command dtypes. We have to change all the attributes to the respective data types by using the command .astype()

```
df.dtypes
```

```
ZIPCODE              object
BOROUGH              object
NEIGHBORHOOD         object
BUILDING_CATEGORY    object
RESIDENTIAL_UNITS   float64
BUILDING_CLASS       object
SIZE                float64
PRICE               float64
dtype: object
```

Since we are having zero values in the price column as well as the size column, we have to replace it with NaN and remove those NaN values using the command .dropna

By using the shape command, we are checking the length, now it has been reduced to 48799

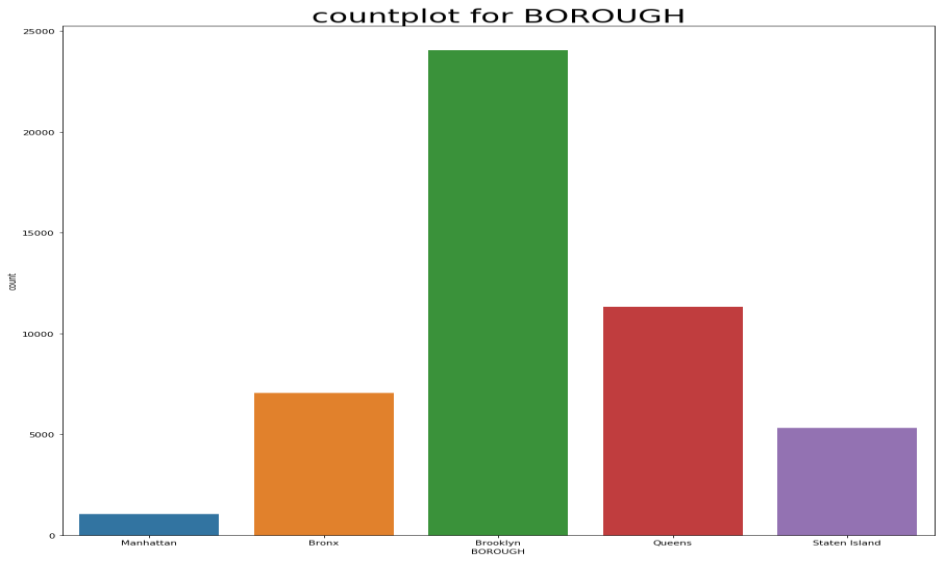
```
In [11]: df['PRICE'] = df['PRICE'].str.replace('-','NaN')
         df['SIZE'] = df['SIZE'].str.replace('-','NaN')

In [18]: df.dropna(subset =['SIZE'],inplace = True)

In [20]: df.dropna(subset =['PRICE'],inplace = True)

In [21]: df.shape

Out[21]: (48799, 8)
```

The borough has the values 1,2,3,4 and 5 representing each borough. We have to change it into their respective borough names i.e. Manhattan, Bronx, Brooklyn, Queens, Staten Island respectively

```
In [23]: df['BOROUGH'] = df['BOROUGH'].replace(1,'Manhattan').astype(object)
         df['BOROUGH'] = df['BOROUGH'].replace(2,'Bronx').astype(object)
         df['BOROUGH'] = df['BOROUGH'].replace(3,'Brooklyn').astype(object)
         df['BOROUGH'] = df['BOROUGH'].replace(4,'Queens').astype(object)
         df['BOROUGH'] = df['BOROUGH'].replace(5,'Staten Island').astype(object)
```

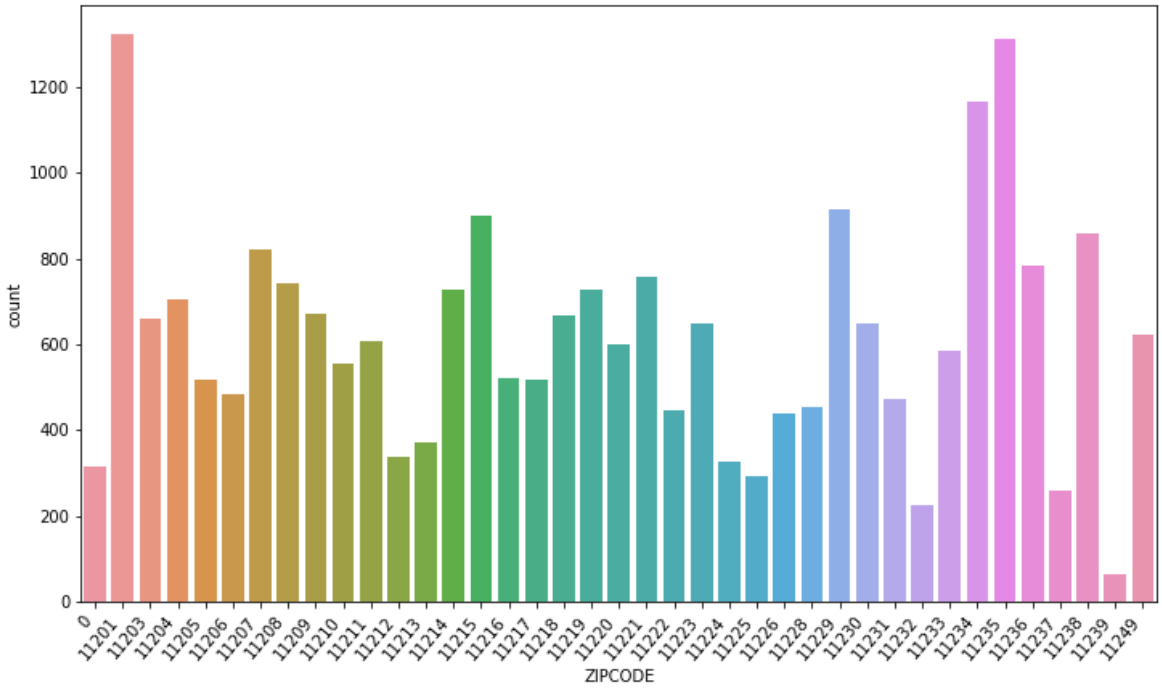
EXPLORATORY DATA ANALYSIS

COUNT PLOT FOR BOROUGH



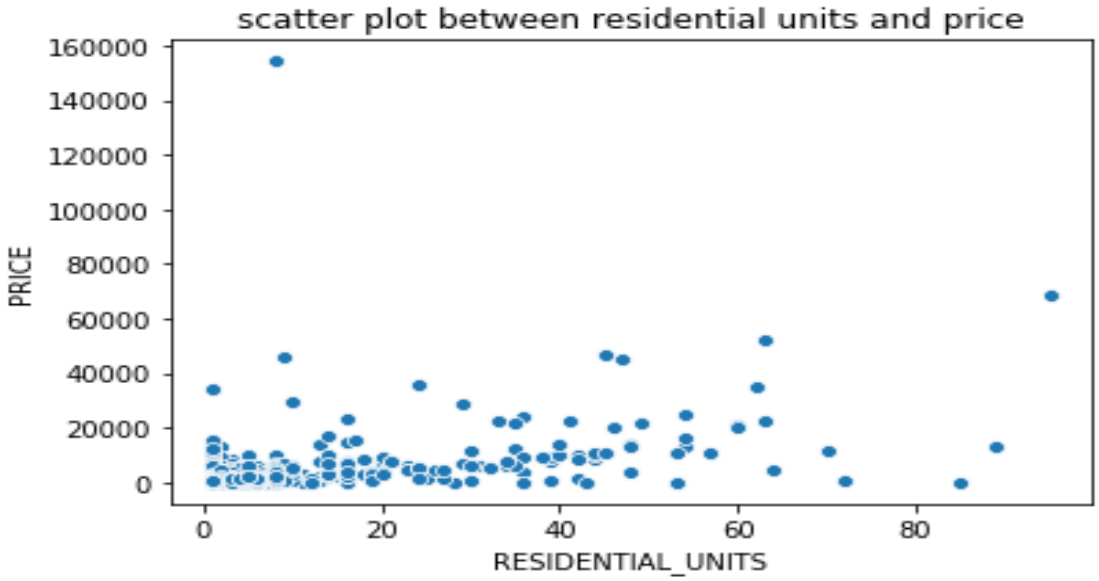
From this graph, we can see that the borough has 5 cities in which Brooklyn is the first highest , Queens has the second highest, Bronx is the third highest, Staten Island is the fourth highest and Manhattan is the last, From this graph we can infer that most of the people are preferring to buy apartments in the Brooklyn Borough.

COUNT PLOT FOR ZIP CODE OF BROOKLYN



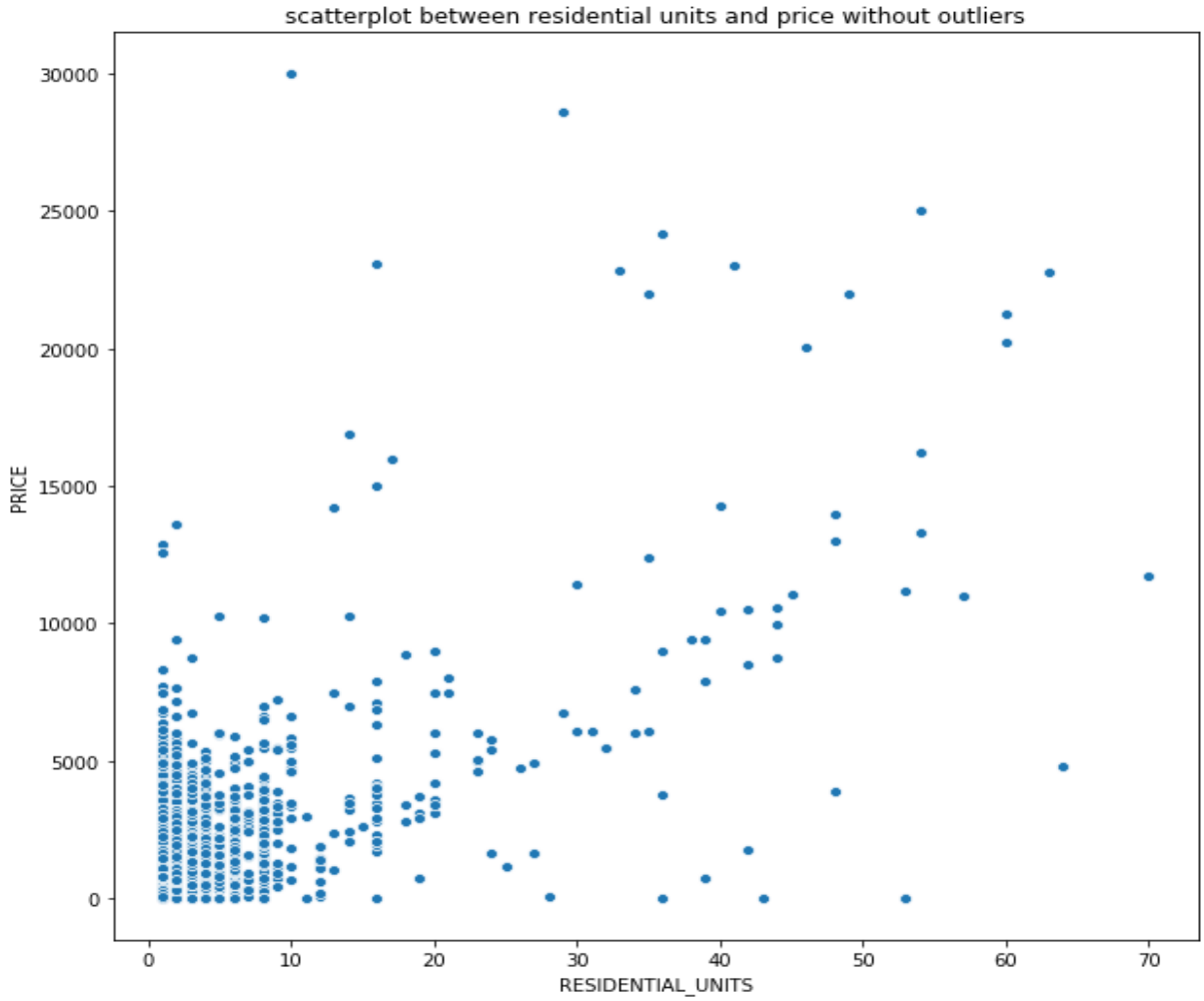
This graph represents each neighborhood in Brooklyn. From this graph, we can see that the zip code 11201 has the highest sales and the count is 1324 and the second highest is zip code is 11235 and its count is 1320. Zip code 11234 is the third highest which has the count of 1190. And the least zip code is 11239. From this we can infer that, people preferably choose the zip code 1120.

SCATTER PLOT BETWEEN RESIDENTIAL UNITS AND PRICE



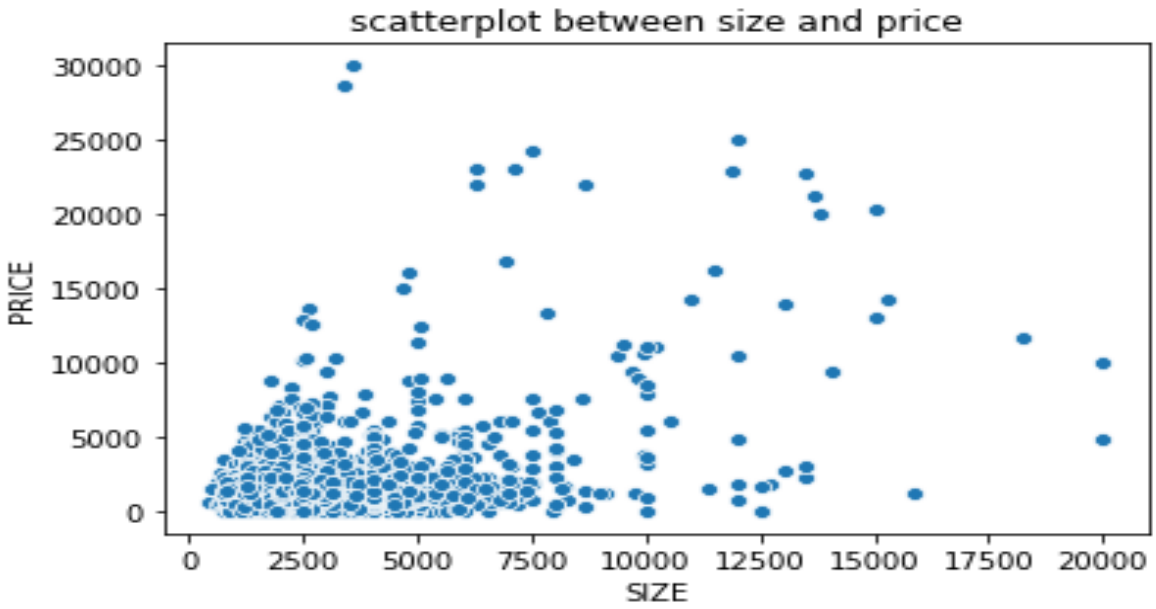
This is the Scatter plot between residential units of the neighborhood and the price of the apartment . This scatter plot consist of outliers on (10,160000).From this graph we can infer that as the Residential units of the Neighborhood increases, the price of the apartment are moderately increases.

Scatter plot between residential units and the price without outliers



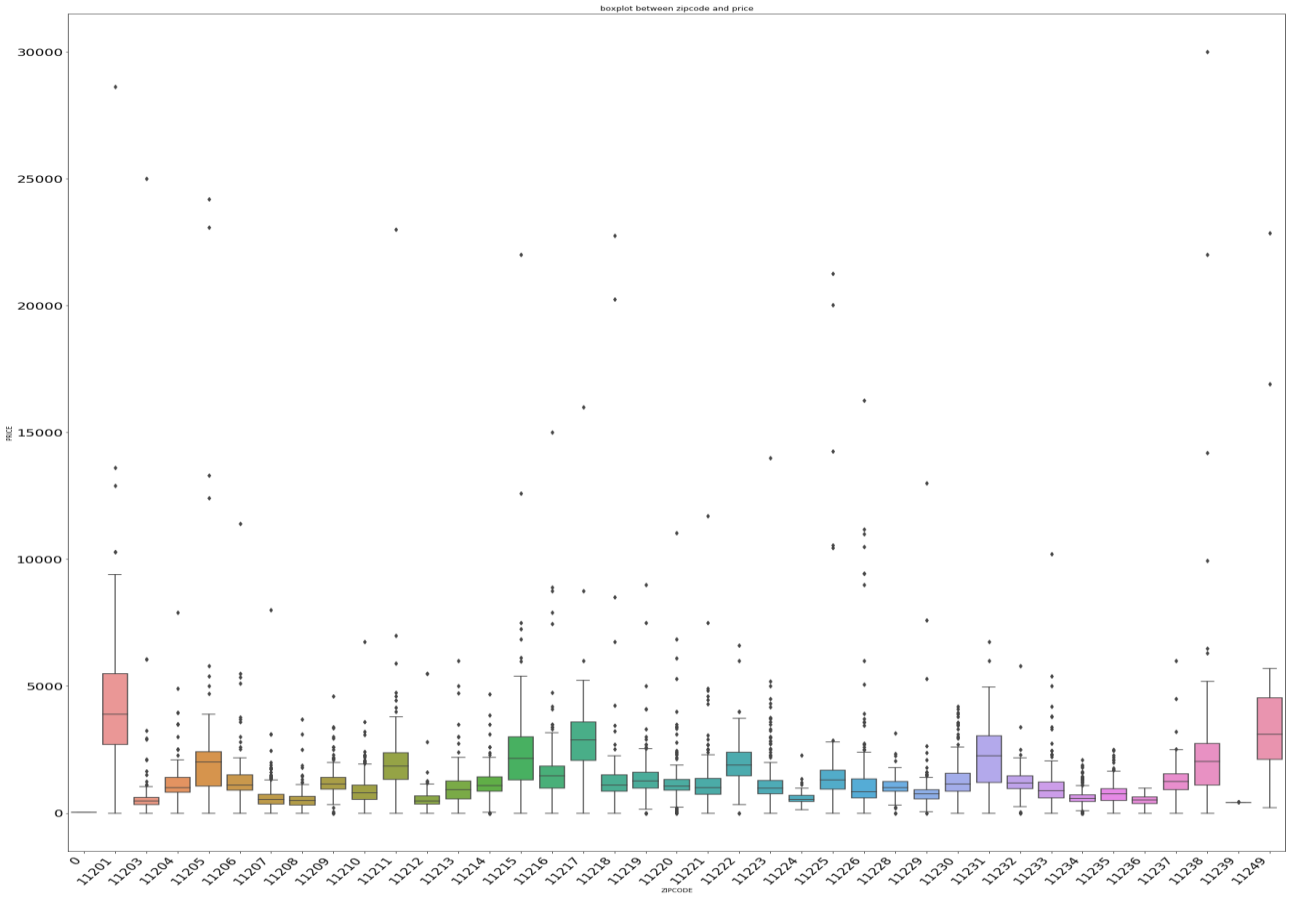
In order to remove the outliers, the price range has been set up from 0 to 30000 and the residential units has been allocated from 0 to 70

SCATTER PLOT BETWEEN SIZE AND PRICE



This scatter plot is drawn between size/m2 and the price of the apartment . From this graph we can infer that as the size of the apartment increases the price of the apartment also increases. In this plot in order to find the error we have normalized the value of price by dividing it with 1000 throughout the dataset.

BOXPLOT FOR ZIPCODE AND PRICE



This box represents each neighborhood in Brooklyn along with its zip code. From this graph, we can see that the zip code 11201 has the highest sales and the count is 1324 and the second highest is zip code is 11235 and its count is 1320. Zip code 11234 is the third highest which has the count of 1190. And the least zip code is 11239. From this we can infer that, people preferably choose the zip code 11201 and those we can see the outliers for all the respective Zip codes accurately with the help of the box plot.

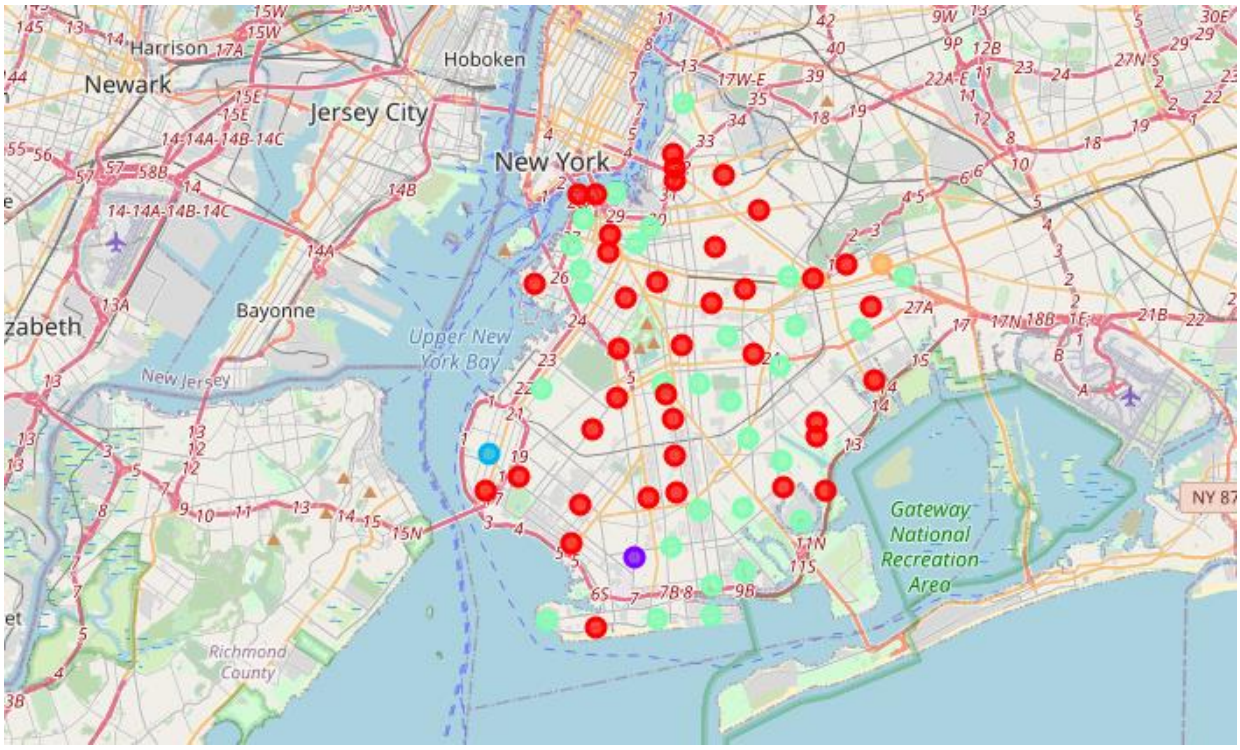
CLUSTERING

NEW YORK CITY NEIGHBOURHOODS DATASET

- This city has total of 5 boroughs and 306 Neighborhoods. In order to segment the Neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the Neighborhoods that exists in each borough as well as the latitude and longitude coordinates of each Neighborhood
- From this dataset, we extracted the Brooklyn Borough data alone and save it as a separate data frame
- Each city in the Brooklyn borough has a separate latitude and longitude, with the help of these data we will be able to extract the venues location data from FOURSQUARED API
- Then we are taking the city at the 7^{th} position in that separated dataset of Brooklyn i.e Ocean Hill

- In order to get the location data of the Ocean Hill neighborhood we are using the Four Squared API with the help of client id and password a dataset has been generated and saved it as a separate dataframe.
- This dataframe consist of details of venues, venues categories of the Neighborhood and the also identifies how many times the venues has been repeated in it.
- Now we have to install the KMEANS CLUSTERING in order to cluster the data according to their similarities. Since it is the unsupervised learning model the dataset has been clustered based upon the similarities and the performance of the venues categories and plotted in the folium map,
- In this project we have given 4 clusters in which the dataset will be classified based upon the 4 clusters. And these classified clusters has been plotted into the folium map along with their latitude and longitude data in order the visualize the data in a better manner and also for the better understanding.

VISUALIZATION OF EACH CLUSTER IN BROOKLYN BOROUGH



RESULTS

From the New York Apartment sales dataset

- Sales in Brooklyn Borough is high when compared to all other Boroughs in the dataset
- From the scatter plot we can come to the conclusion that as Size of the apartment increases Price also increases
- And also as the Residential Unit increases the price of the apartment also increases.

DISCUSSION AND CONCLUSION:

- From the analysis, we came to know that Brooklyn is one of the best Boroughs of the New York Cities and have varieties of Venues to explore.
- In the Brooklyn Boroughs, the cities have the highest sale is cheap, Effective , Eco – friendly, have high Employment opportunity.