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.

DATA ACQUISITION AND CLEANING

DATA ACQUISITION:

- 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.

DATA ACQUISITION AND CLEANING

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.

DATASET USED FOR CLUSTERING NEIGHBORHOOD

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

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

DATA CLEANING

- 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

DATA CLEANING

- 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
```

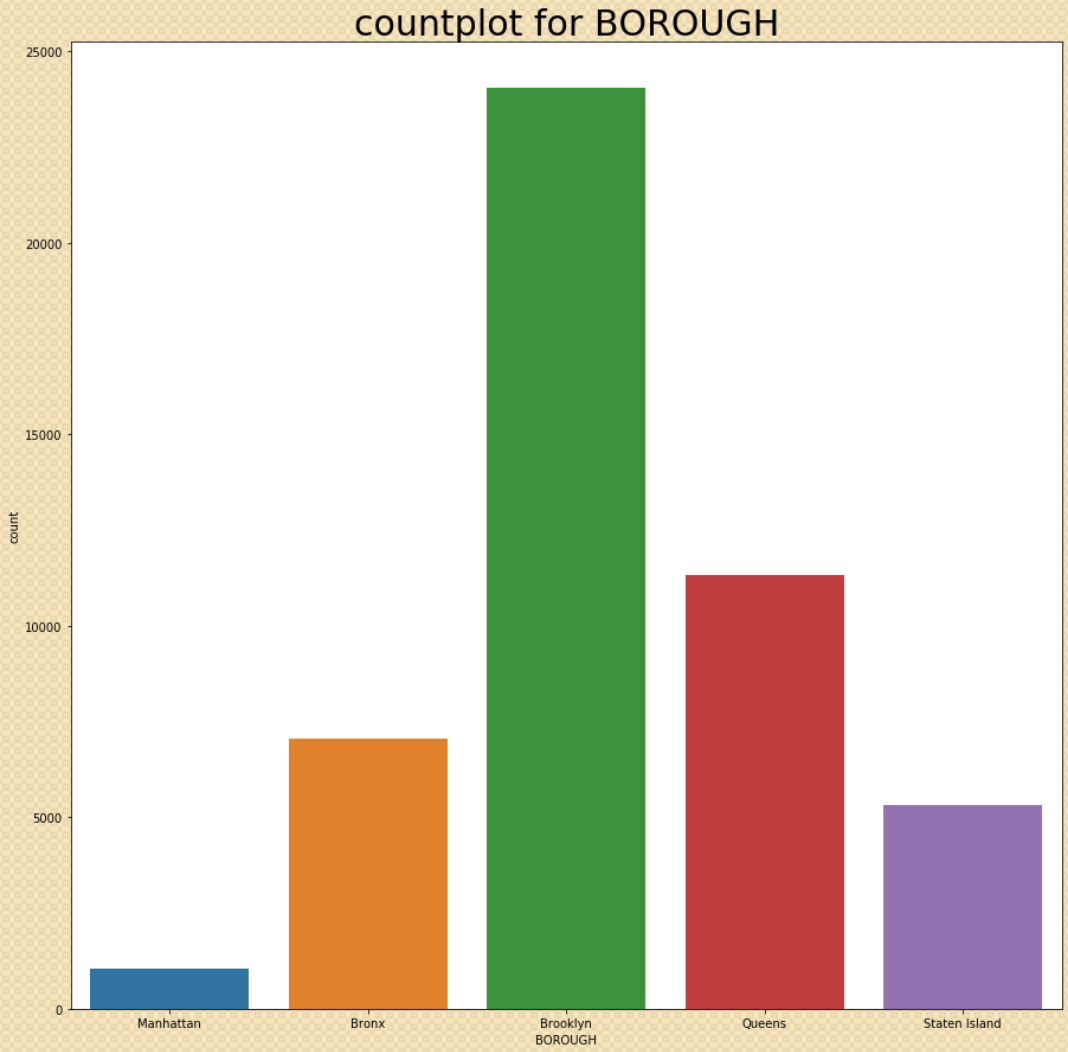
DATA CLEANING

- Since we are having z values in the p column as well as size column, we have replace it with NaN remove those N values using command .dropna
- By using the sh command, we checking the length, i it has been reduced 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)
```

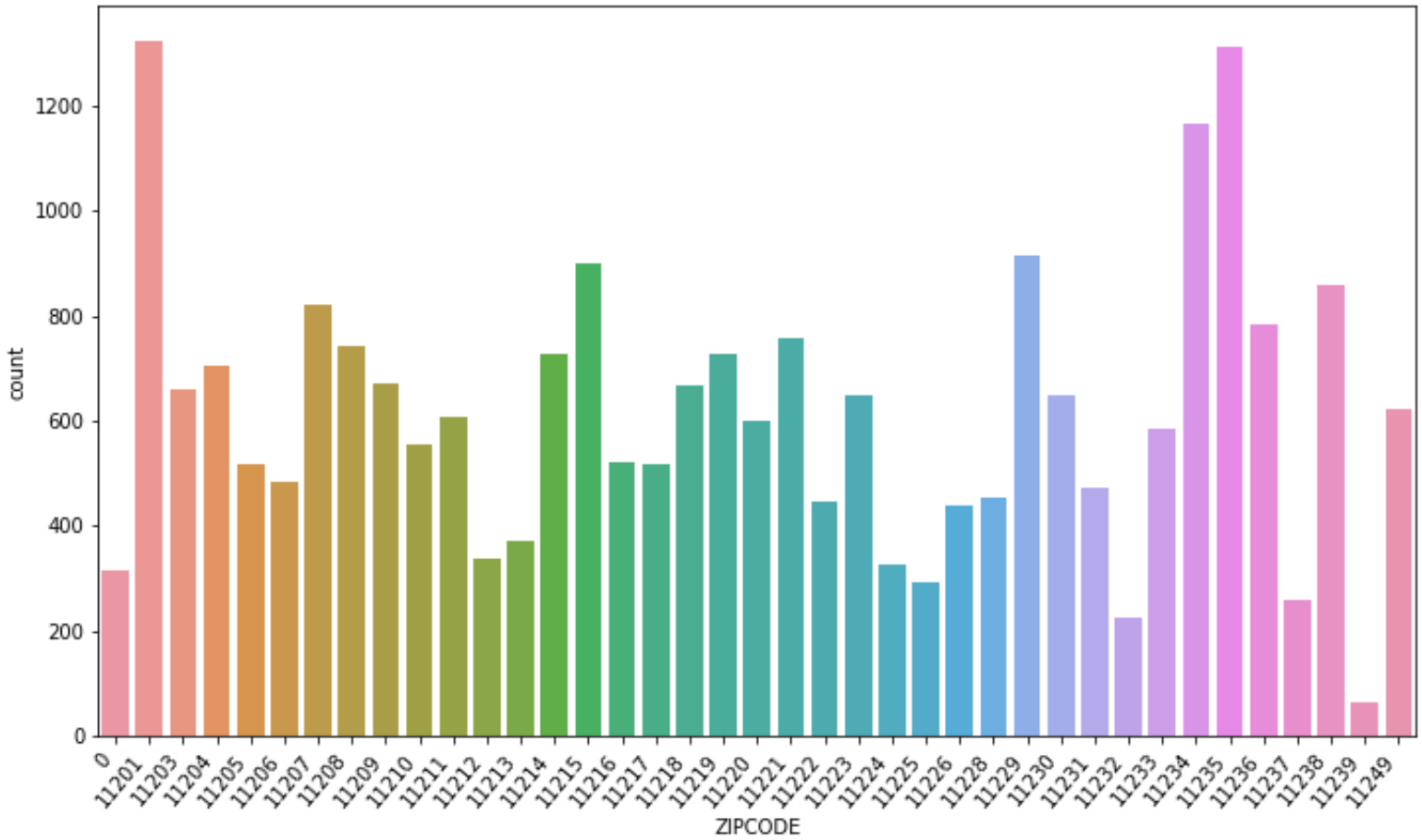
In [23]:
```
df['BOROUGH'] = df['BOROUGH'].replace(1,'Manhattan').astype(object)
df['BOROUGH'] = df['BOROUGH'].replace(2,'Bronx').astype(object)
df['BOROUGH'] = df['BOROUGH'].replace(3,'Brooklyn').astype(object)
df['BOROUGH'] = df['BOROUGH'].replace(4,'Queens').astype(object)
df['BOROUGH'] = df['BOROUGH'].replace(5,'Staten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

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



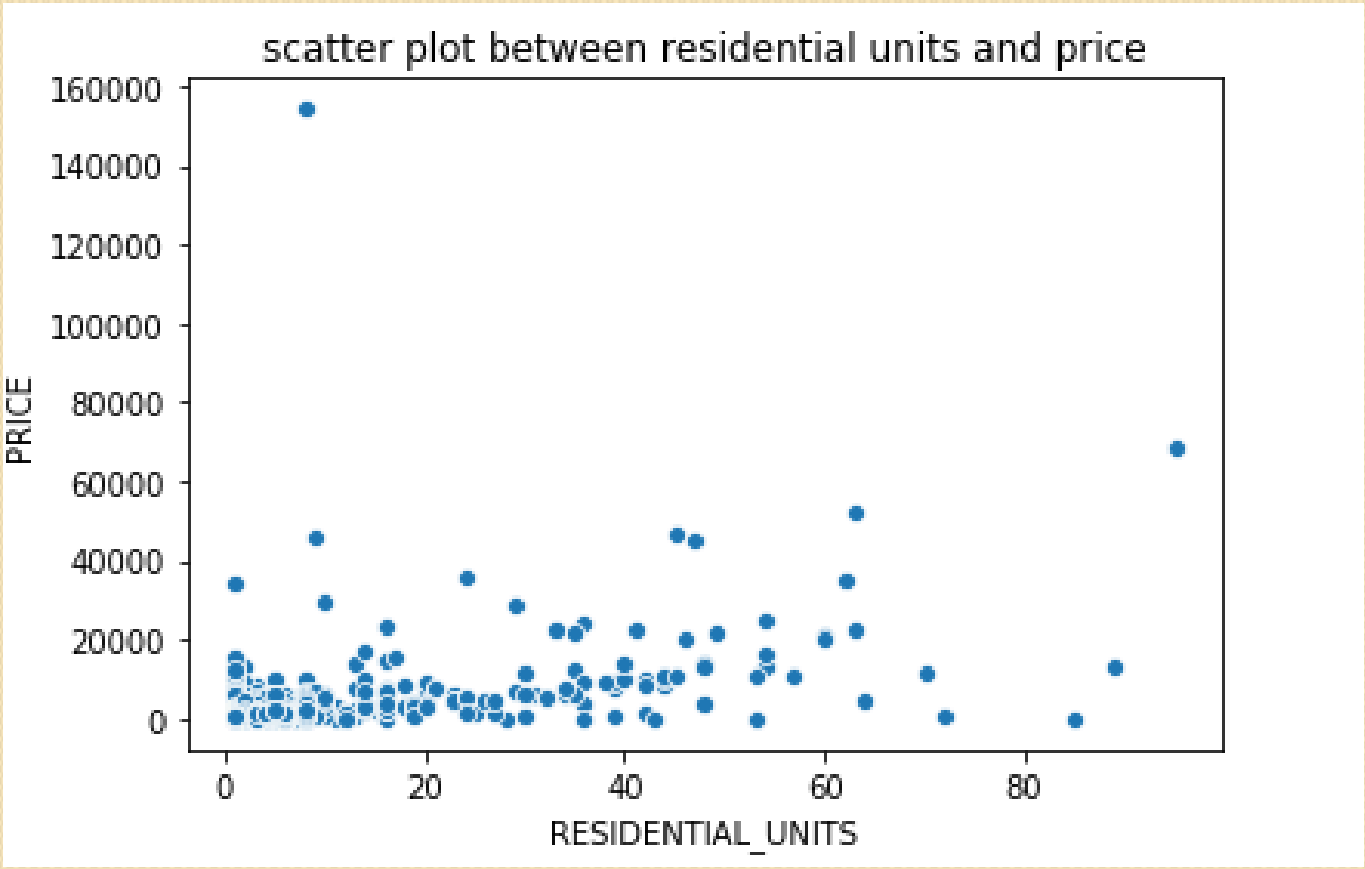
From this graph, we can see that the borough has 5 cities in which Brooklyn has the highest sales

COUNT PLOT FOR ZIP CODE OF BROOKLYN



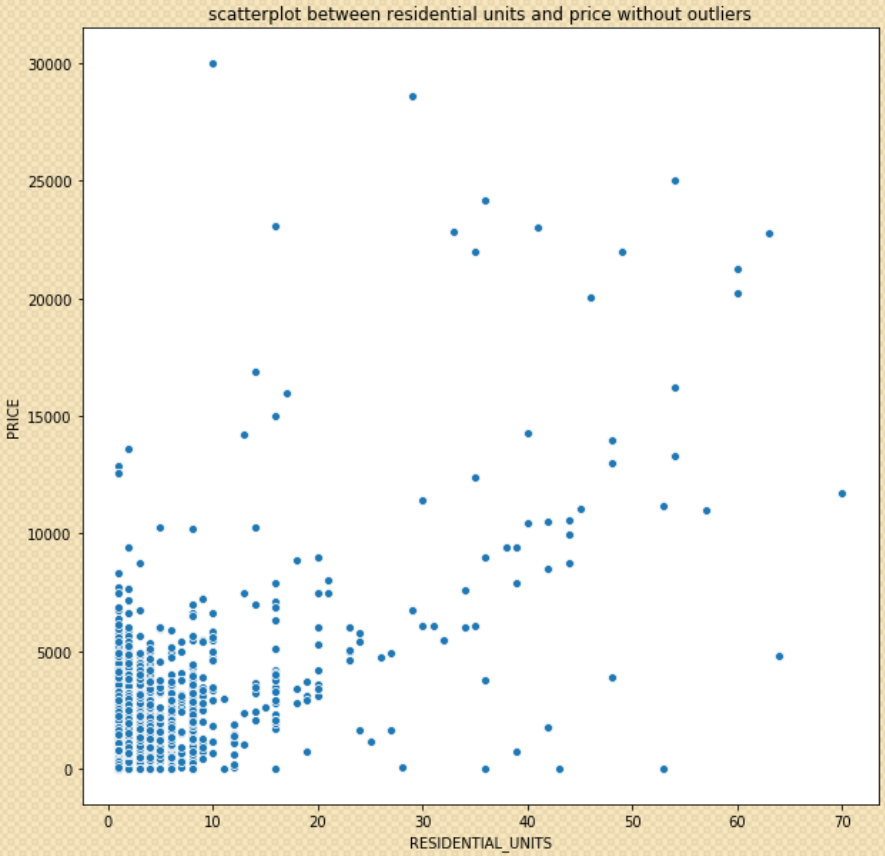
From this graph, we can see that the zip code 11201 has the highest sales and the count is 1324

SCATTER PLOT BETWEEN RESIDENTIAL UNITS AND PRICE



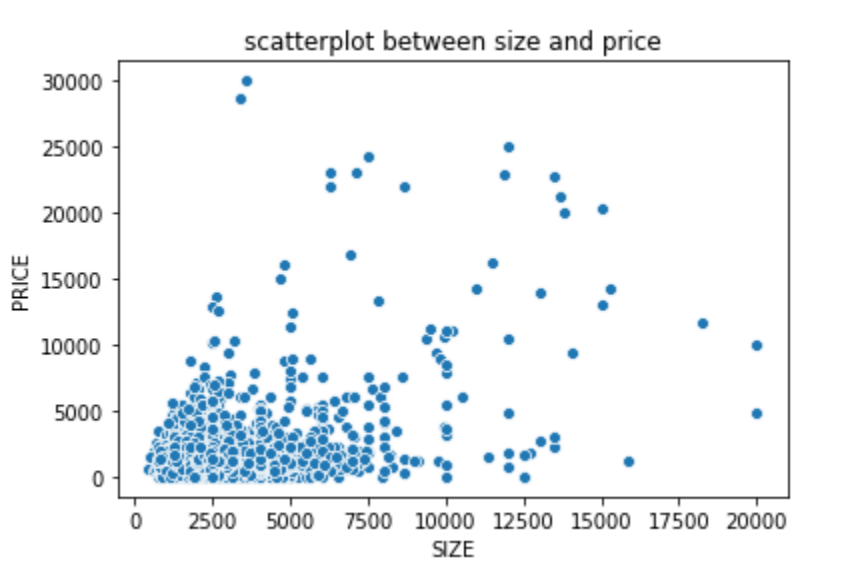
This scatter plot is between residential units and price and it has an outliers

SCATTER PLOT BETWEEN RESIDENTIAL UNITS AND PRICE WITHOUT OULIERS



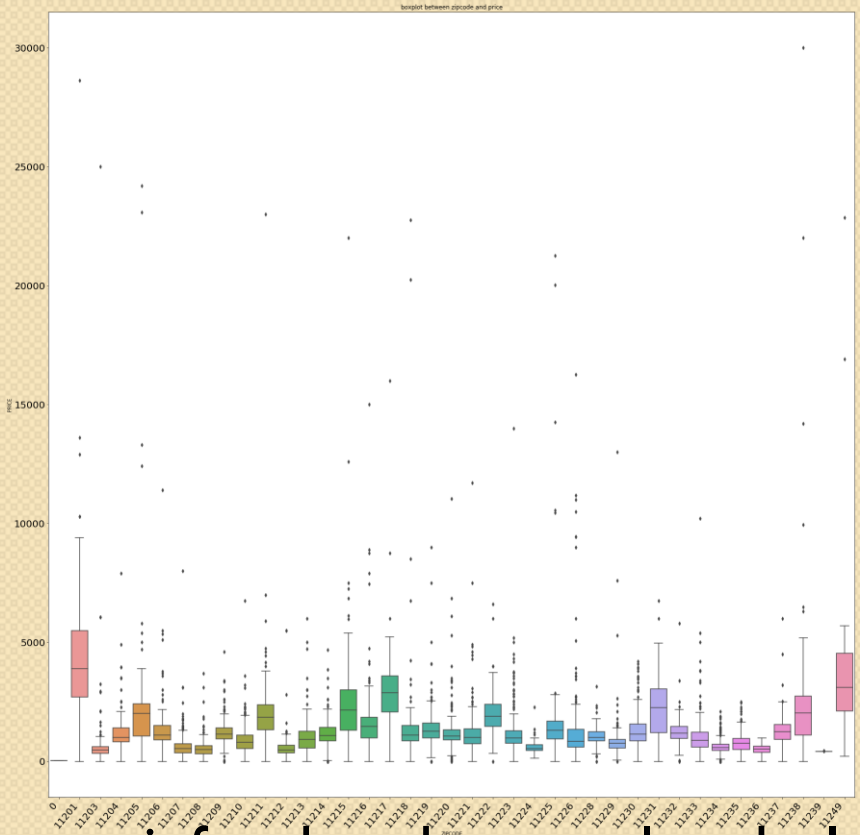
In order to remove the outliers, the price range has been set up from 0 to 30000 and the residential units has been allocated from 0 to 70

SCATTER PLOT BETWEEN SIZE AND PRICE



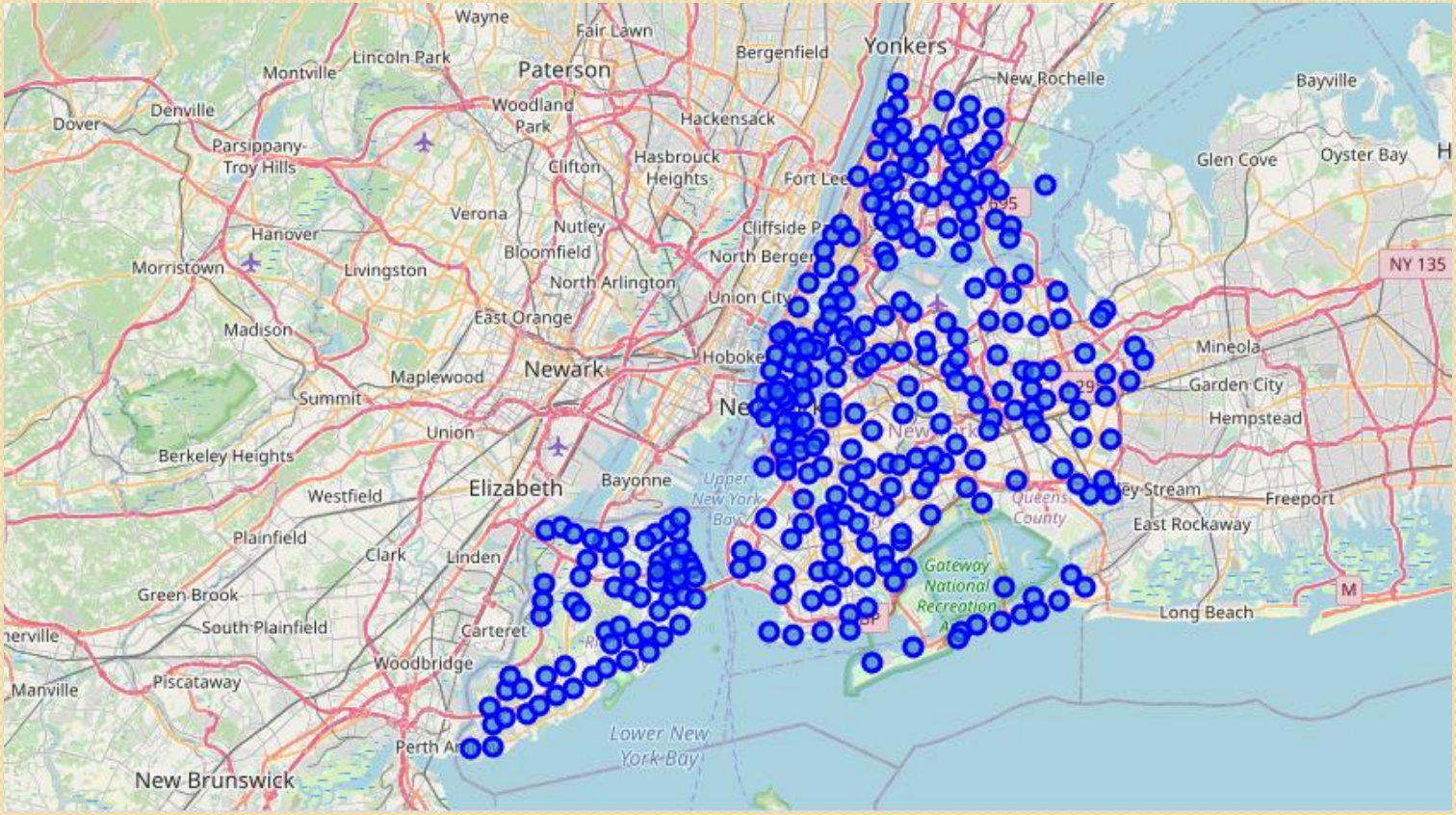
This scatter plot is between size and price

BOXPLOT FOR ZIPCODE AND PRICE



From this , we infer that the postal code 11201 has the highest price range of 3000 with a maximum limit of 9000

VISUALIZATION OF CITIES IN ALL BOROUGHS OF THE NEW YORK CITY



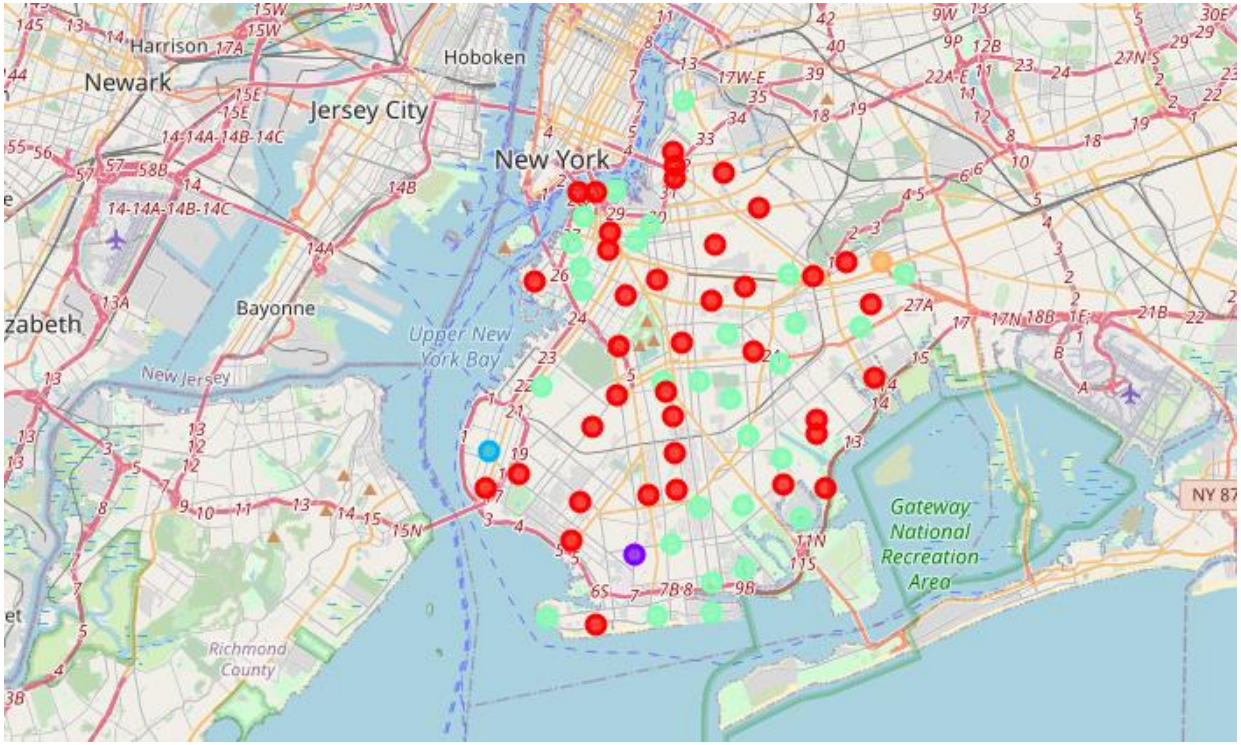
CLUSTERING

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

CLUSTERING

- From this, venue category data can be obtained
- These venues category data is used for the clustering purpose with the help of KMEANS clustering



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.