BATTLE OF

NEIGHBORHOODS

NEW YORK CITY - PROPERTY SALES ANALYSIS TO FIND THE BEST HOME FOR YOU

KARTIK NAIR

JULY 2020

1. INTRODUCTION

1.1 BACKGROUND

For people trying to find the best places to live, it's always a good idea to compare cities and if possible, to compare neighborhoods to see if its suites your taste and fits your budget. The cost of living in the neighbour hood and the amenities with the property and nearby it is a top concern when moving to a new area. For some its a restaurant nearby for some a coffee shop , in all everyone has their requirements which are needeed to be looked out for.

1.2 PROBLEM

The Property Sales dataset of New York has sales details with price of different types of houses sold in each borough of New York over the years 2010-2019. The market price of each place changes with time. This project aims to select the boroughs in NYC based on the highest number of sales, explore the neighborhoods of that borough to find the 10 most common venues in each neighborhood and finally cluster the neighborhoods using k-mean clustering.

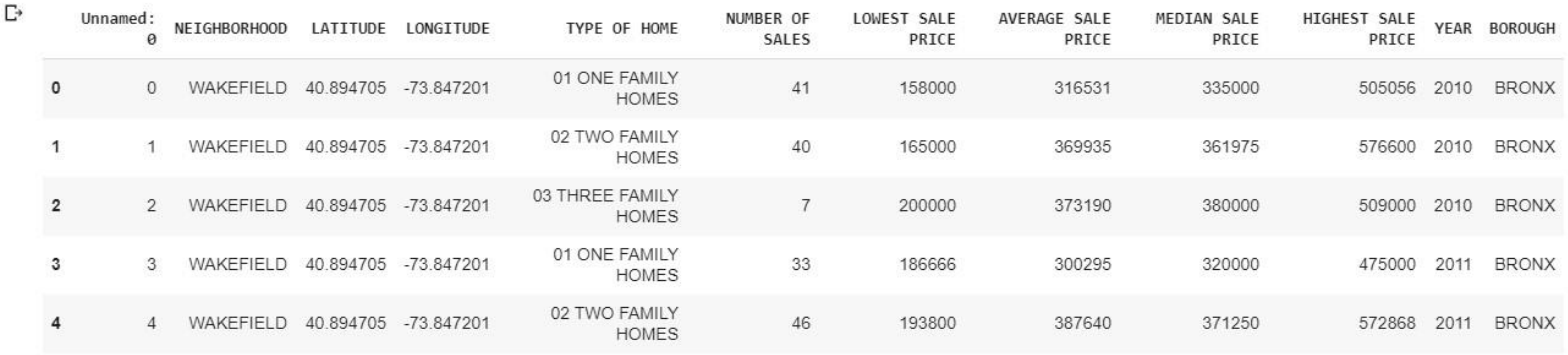
2. Data

There were two main datasets used in this project, one being the New York City Property Sales dataset 2010-2019 and the other dataset contained all the neighbourhoods of NYC with their geographical coordinates. But for Comfort The Geographical coordinates was merged with the Property sales dataset

2.1 Dataset

2.1.1 NYC Property Sales Dataset

Let’s have a quick look at our dataset:-



	Unnamed: 0	NEIGHBORHOOD	LATITUDE	LONGITUDE	TYPE OF HOME	NUMBER OF SALES	LOWEST SALE PRICE	AVERAGE SALE PRICE	MEDIAN SALE PRICE	HIGHEST SALE PRICE	YEAR	BOROUGH
0	0	WAKEFIELD	40.894705	-73.847201	01 ONE FAMILY HOMES	41	158000	316531	335000	505056	2010	BRONX
1	1	WAKEFIELD	40.894705	-73.847201	02 TWO FAMILY HOMES	40	165000	369935	361975	576600	2010	BRONX
2	2	WAKEFIELD	40.894705	-73.847201	03 THREE FAMILY HOMES	7	200000	373190	380000	509000	2010	BRONX
3	3	WAKEFIELD	40.894705	-73.847201	01 ONE FAMILY HOMES	33	186666	300295	320000	475000	2011	BRONX
4	4	WAKEFIELD	40.894705	-73.847201	02 TWO FAMILY HOMES	46	193800	387640	371250	572868	2011	BRONX

- Some Important Columns :

1.NEIGHBORHOOD:- This column has the neighborhood name of each sale

2. NUMBER OF SALES :- This numeric column indicates the total sales done in each neighborhood of a particular type of house.

3. BOROUGH :- This column is very important for our project this column indicates the borough of the sale

And 11 more columns for now , some will be cleaned later and these columns are not relevant to our analysis either.

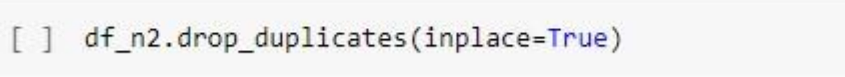
2.2 Data Cleaning

The New York Sales Dataset needs a few minor cleaning,

1. Duplicate rows needed to be removed
2. Uneccessary columns to be dropped
3. And Column Names to be renamed for further processes
4. Total number of sales per Borough needed

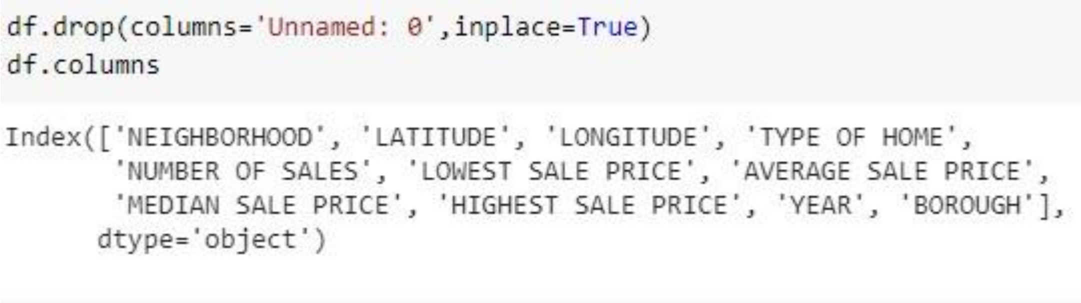
These all processes were done using the Pandas Library as follows :

1.



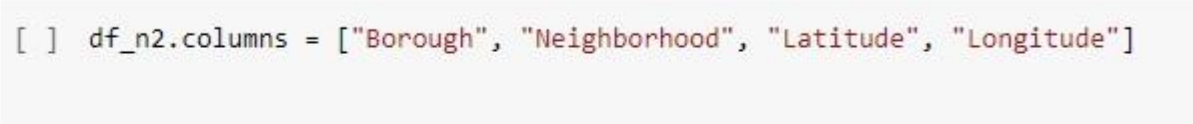
```
[ ]  df_n2.drop_duplicates(inplace=True)
```

2.



```
df.drop(columns='Unnamed: 0',inplace=True)
df.columns
```

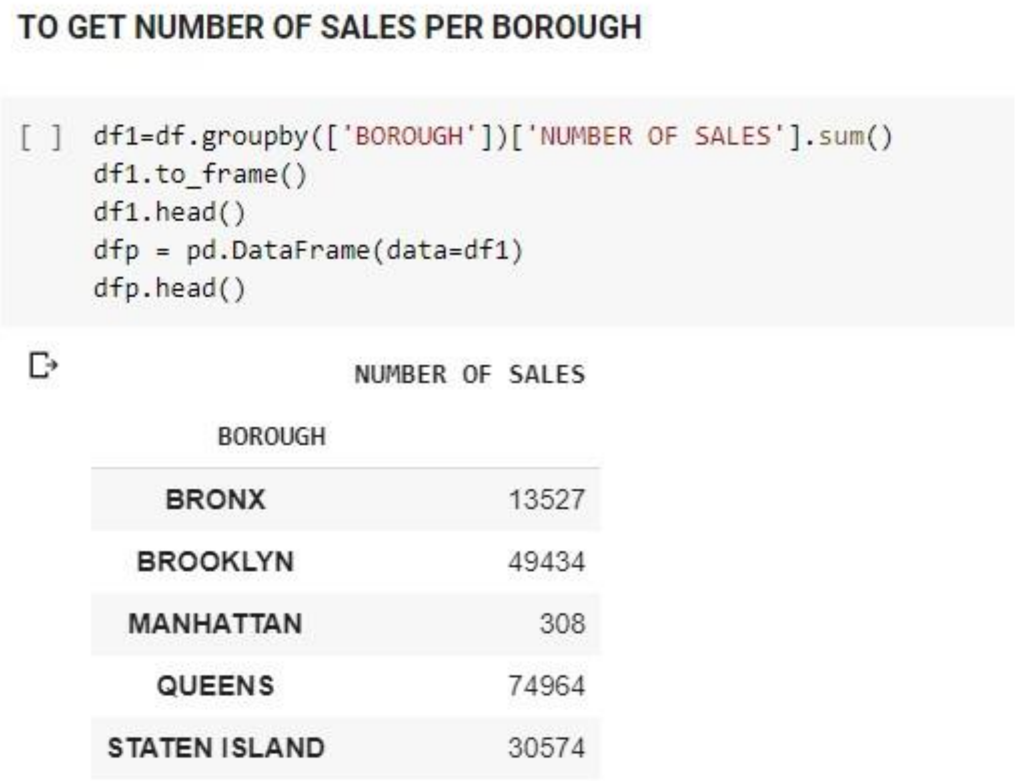
```
Index(['NEIGHBORHOOD', 'LATITUDE', 'LONGITUDE', 'TYPE OF HOME',
       'NUMBER OF SALES', 'LOWEST SALE PRICE', 'AVERAGE SALE PRICE',
       'MEDIAN SALE PRICE', 'HIGHEST SALE PRICE', 'YEAR', 'BOROUGH'],
      dtype='object')
```



```
[ ]  df_n2.columns = ["Borough", "Neighborhood", "Latitude", "Longitude"]
```

3.

4.



TO GET NUMBER OF SALES PER BOROUGH

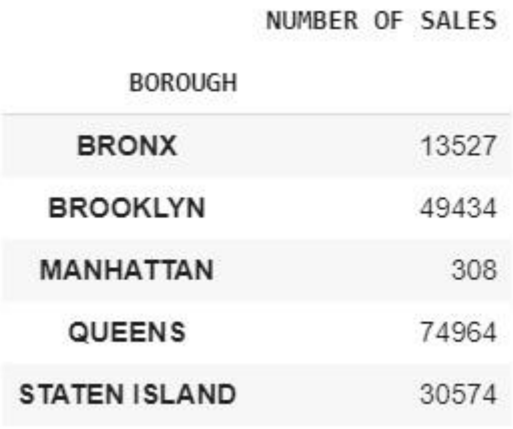
```
[ ]  df1=df.groupby(['BOROUGH'])['NUMBER OF SALES'].sum()
     df1.to_frame()
     df1.head()
     dfp = pd.DataFrame(data=df1)
     dfp.head()
```

```
                          NUMBER OF SALES
          BOROUGH
           BRONX                    13527
         BROOKLYN                   49434
         MANHATTAN                    308
          QUEENS                    74964
       STATEN ISLAND                30574
```

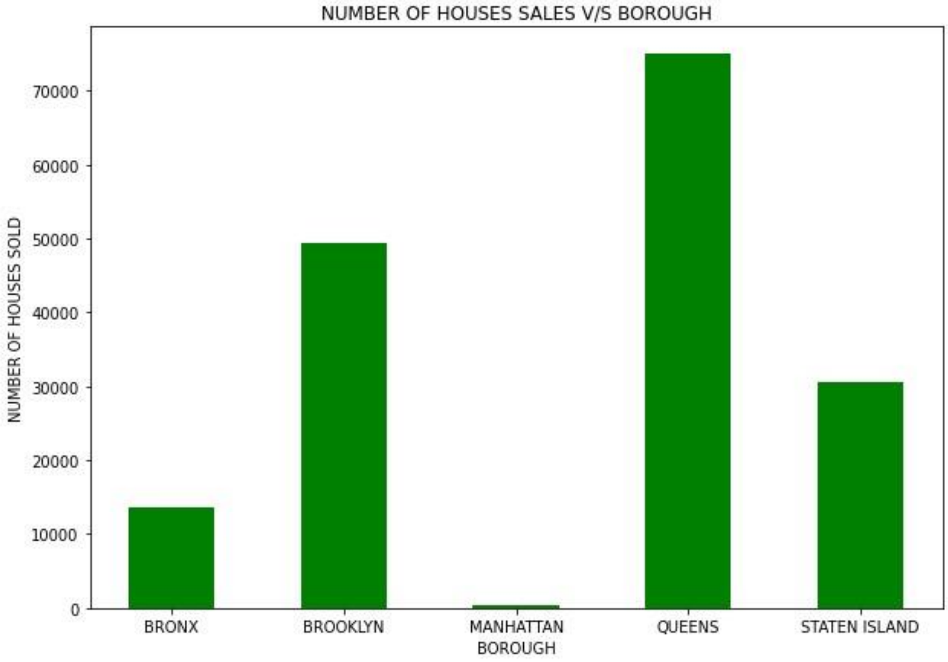
3. METHODOLOGY

3.1 Exploratory Data Analysis

Let’s Visualize The Sales in each borough



	NUMBER OF SALES
BOROUGH	
BRONX	13527
BROOKLYN	49434
MANHATTAN	308
QUEENS	74964
STATEN ISLAND	30574



Comparing five boroughs it is evident that Queens has the highest Property Sales recorded followed by Brooklyn, Staten Island, Bronx and Manhattan

It’s clearly seen that people prefer QUEENS than any other borough so that’s our first preference so we’ll explore this.

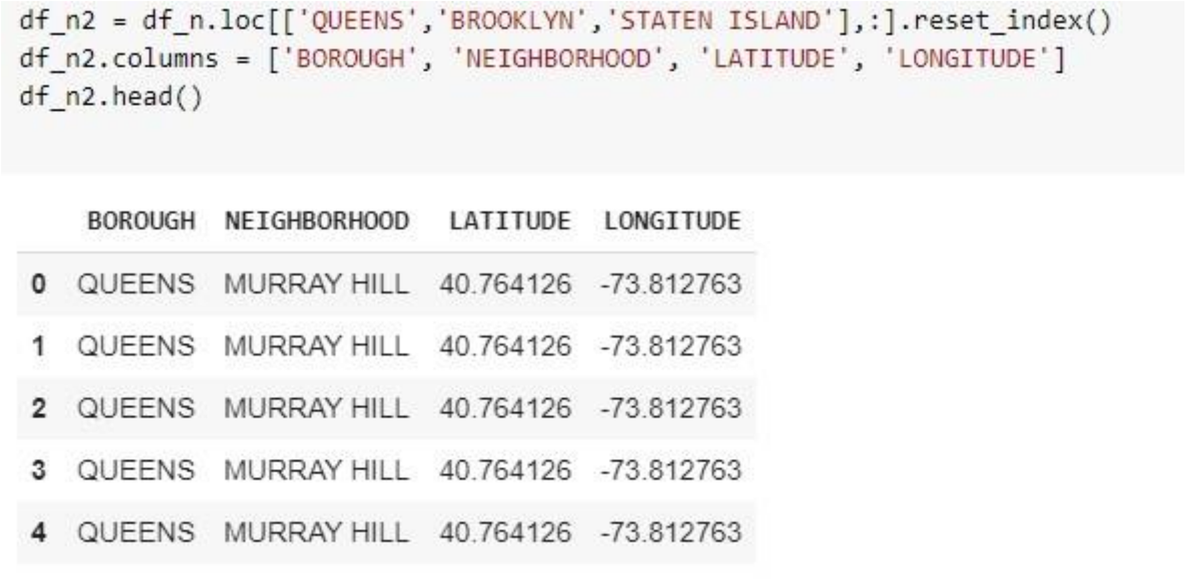
3.1.2 Neighborhoods in QUEENS

Exploring the Queens Borough showed that only a single borough won’t have enough neighbourhoods for a person to choose from . Hence under this case we will take 3 Borough with most property sales which are

- QUEENS
- BROOKLYN
- STATEN ISLAND

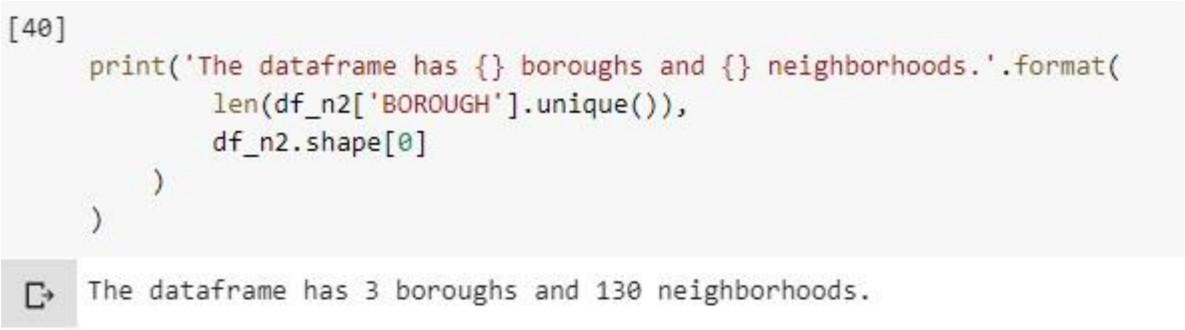
3.1.3 Visualize the neighbourhoods

Firstly, we select the neighbourhoods only from these 3 Boroughs, that is Bronx, Queens and Staten Island, and save them in a new dataframe.



```
df_n2 = df_n.loc[['QUEENS','BROOKLYN','STATEN ISLAND'],:].reset_index()
df_n2.columns = ['BOROUGH', 'NEIGHBORHOOD', 'LATITUDE', 'LONGITUDE']
df_n2.head()
```

	BOROUGH	NEIGHBORHOOD	LATITUDE	LONGITUDE
0	QUEENS	MURRAY HILL	40.764126	-73.812763
1	QUEENS	MURRAY HILL	40.764126	-73.812763
2	QUEENS	MURRAY HILL	40.764126	-73.812763
3	QUEENS	MURRAY HILL	40.764126	-73.812763
4	QUEENS	MURRAY HILL	40.764126	-73.812763

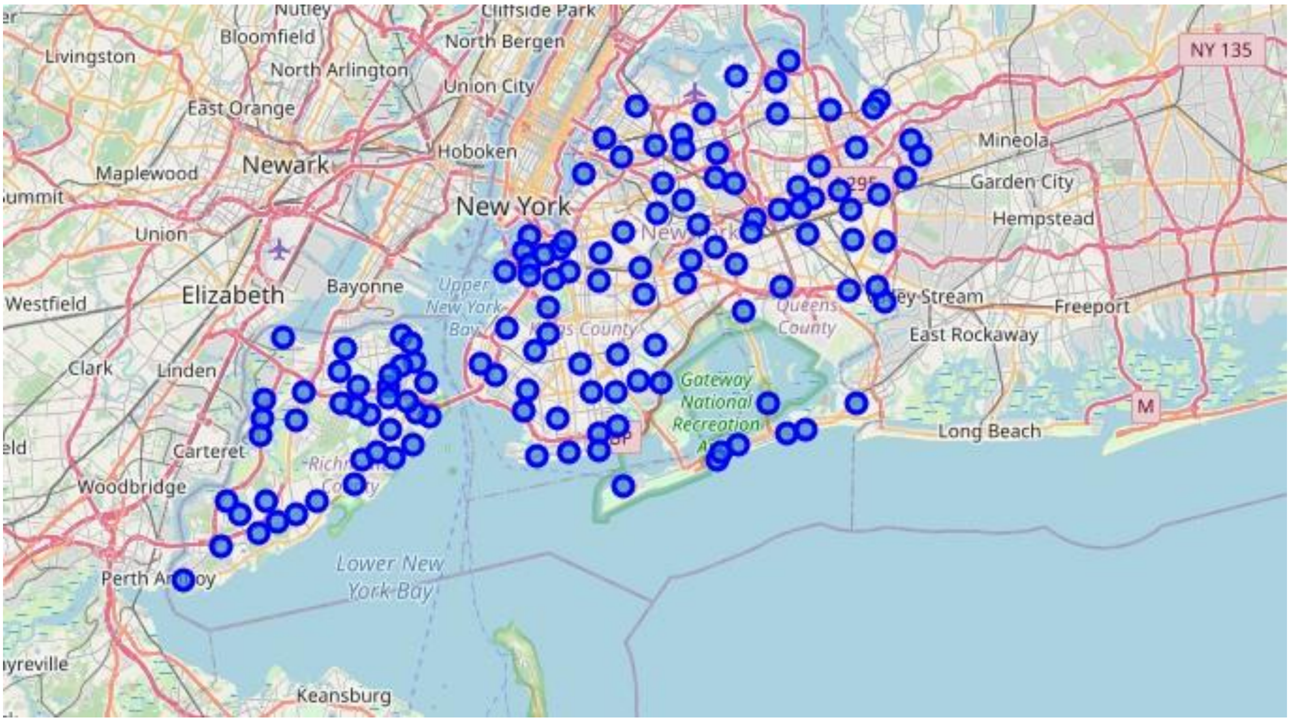


```
[40]
    print('The dataframe has {} boroughs and {} neighborhoods.'.format(
            len(df_n2['BOROUGH'].unique()),
            df_n2.shape[0]
        )
    )
```

```
The dataframe has 3 boroughs and 130 neighborhoods.
```

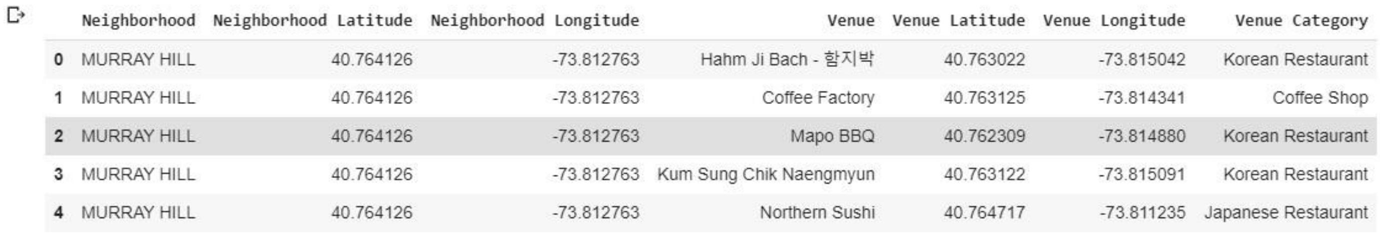
We find that there are 130 neighbourhoods recorded in the 3 Boroughs

To visualize this we use the *Folium* Library



3.2 Modelling

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



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	MURRAY HILL	40.764126	-73.812763	Hahm Ji Bach - 함지박	40.763022	-73.815042	Korean Restaurant
1	MURRAY HILL	40.764126	-73.812763	Coffee Factory	40.763125	-73.814341	Coffee Shop
2	MURRAY HILL	40.764126	-73.812763	Mapo BBQ	40.762309	-73.814880	Korean Restaurant
3	MURRAY HILL	40.764126	-73.812763	Kum Sung Chik Naengmyun	40.763122	-73.815091	Korean Restaurant
4	MURRAY HILL	40.764126	-73.812763	Northern Sushi	40.764717	-73.811235	Japanese Restaurant

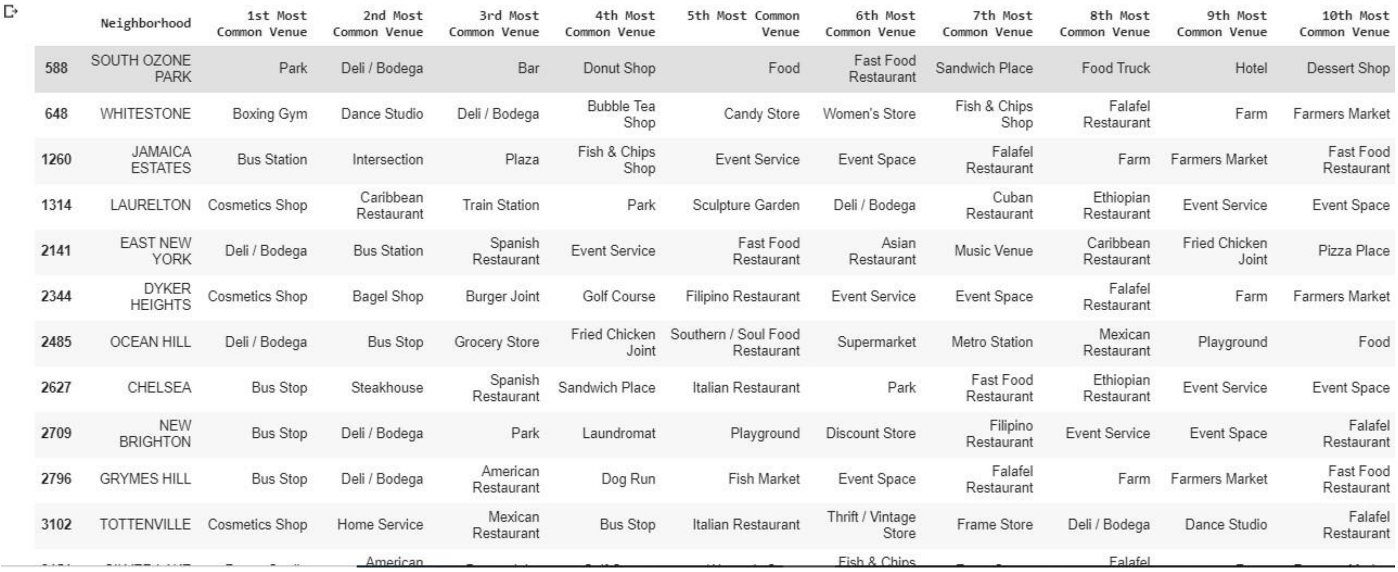
One hot encoding is done on the venues data. (One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction). The Venues data is then grouped by the Neighbourhood and the mean of the venues are calculated, 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. We used the elbow method to find the best cluster size and found 5 clusters to be ideal.

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.

4. Results

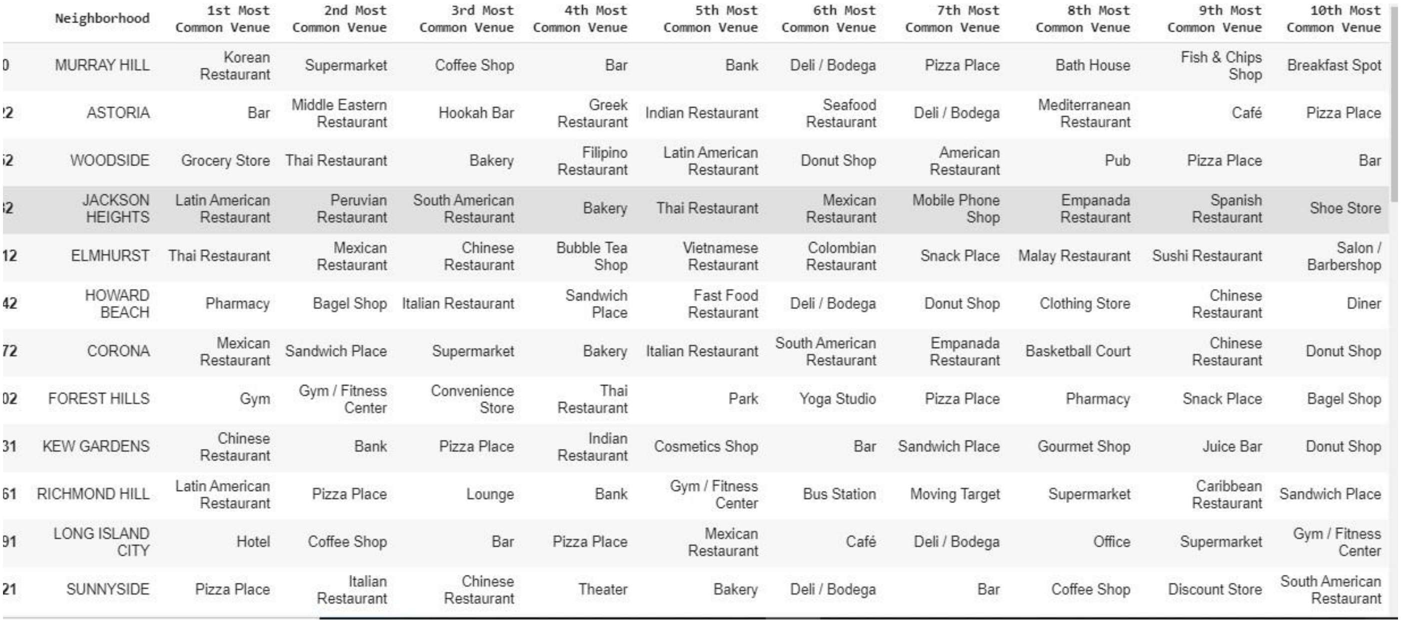
After running the K-means clustering we can access each cluster created to see which neighborhoods were assigned to each of the five clusters. Looking into the neighborhoods in the first cluster



	Neighborhood	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
588	SOUTH OZONE PARK	Park	Deli / Bodega	Bar	Donut Shop	Food	Fast Food Restaurant	Sandwich Place	Food Truck	Hotel	Dessert Shop
648	WHITESTONE	Boxing Gym	Dance Studio	Deli / Bodega	Bubble Tea Shop	Candy Store	Women's Store	Fish & Chips Shop	Falafel Restaurant	Farm	Farmers Market
1260	JAMAICA ESTATES	Bus Station	Intersection	Plaza	Fish & Chips Shop	Event Service	Event Space	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
1314	LAURELTON	Cosmetics Shop	Caribbean Restaurant	Train Station	Park	Sculpture Garden	Deli / Bodega	Cuban Restaurant	Ethiopian Restaurant	Event Service	Event Space
2141	EAST NEW YORK	Deli / Bodega	Bus Station	Spanish Restaurant	Event Service	Fast Food Restaurant	Asian Restaurant	Music Venue	Caribbean Restaurant	Fried Chicken Joint	Pizza Place
2344	DYKER HEIGHTS	Cosmetics Shop	Bagel Shop	Burger Joint	Golf Course	Filipino Restaurant	Event Service	Event Space	Falafel Restaurant	Farm	Farmers Market
2485	OCEAN HILL	Deli / Bodega	Bus Stop	Grocery Store	Fried Chicken Joint	Southern / Soul Food Restaurant	Supermarket	Metro Station	Mexican Restaurant	Playground	Food
2627	CHELSEA	Bus Stop	Steakhouse	Spanish Restaurant	Sandwich Place	Italian Restaurant	Park	Fast Food Restaurant	Ethiopian Restaurant	Event Service	Event Space
2709	NEW BRIGHTON	Bus Stop	Deli / Bodega	Park	Laundromat	Playground	Discount Store	Filipino Restaurant	Event Service	Event Space	Falafel Restaurant
2796	GRYMES HILL	Bus Stop	Deli / Bodega	American Restaurant	Dog Run	Fish Market	Event Space	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant
3102	TOTTENVILLE	Cosmetics Shop	Home Service	Mexican Restaurant	Bus Stop	Italian Restaurant	Thrift / Vintage Store	Frame Store	Deli / Bodega	Dance Studio	Falafel Restaurant
			American				Fish & Chips		Falafel		

Upon closely examining these neighborhoods we can see that the most common venues in these neighborhoods are Bus Stop, Coffee shops and restaurants.

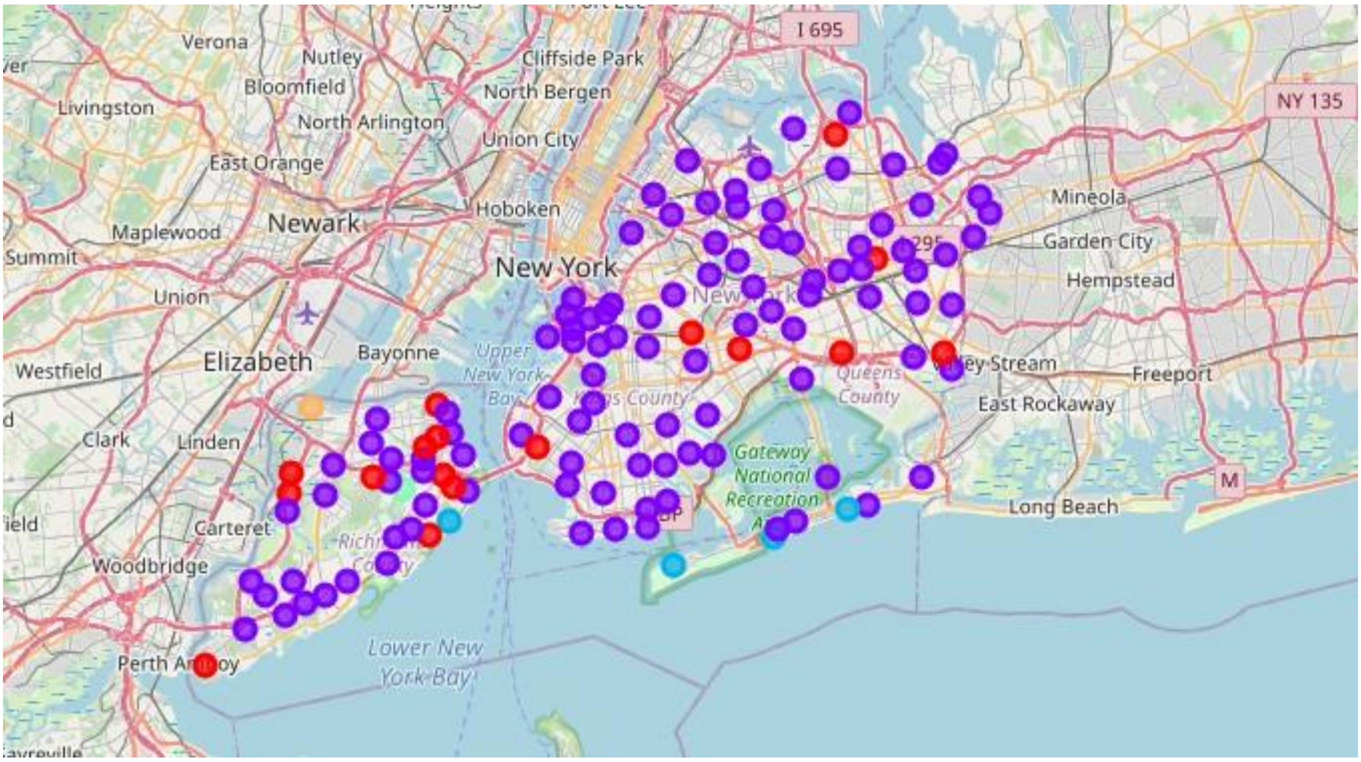
Similarly looking at the 2nd cluster we see that it mainly consists of the neighbourhoods with venues such as be Bar, Pizza Place, Pharmacy.



	Neighborhood	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	MURRAY HILL	Korean Restaurant	Supermarket	Coffee Shop	Bar	Bank	Deli / Bodega	Pizza Place	Bath House	Fish & Chips Shop	Breakfast Spot
2	ASTORIA	Bar	Middle Eastern Restaurant	Hookah Bar	Greek Restaurant	Indian Restaurant	Seafood Restaurant	Deli / Bodega	Mediterranean Restaurant	Café	Pizza Place
2	WOODSIDE	Grocery Store	Thai Restaurant	Bakery	Filipino Restaurant	Latin American Restaurant	Donut Shop	American Restaurant	Pub	Pizza Place	Bar
2	JACKSON HEIGHTS	Latin American Restaurant	Peruvian Restaurant	South American Restaurant	Bakery	Thai Restaurant	Mexican Restaurant	Mobile Phone Shop	Empanada Restaurant	Spanish Restaurant	Shoe Store
12	ELMHURST	Thai Restaurant	Mexican Restaurant	Chinese Restaurant	Bubble Tea Shop	Vietnamese Restaurant	Colombian Restaurant	Snack Place	Malay Restaurant	Sushi Restaurant	Salon / Barbershop
42	HOWARD BEACH	Pharmacy	Bagel Shop	Italian Restaurant	Sandwich Place	Fast Food Restaurant	Deli / Bodega	Donut Shop	Clothing Store	Chinese Restaurant	Diner
72	CORONA	Mexican Restaurant	Sandwich Place	Supermarket	Bakery	Italian Restaurant	South American Restaurant	Empanada Restaurant	Basketball Court	Chinese Restaurant	Donut Shop
02	FOREST HILLS	Gym	Gym / Fitness Center	Convenience Store	Thai Restaurant	Park	Yoga Studio	Pizza Place	Pharmacy	Snack Place	Bagel Shop
31	KEW GARDENS	Chinese Restaurant	Bank	Pizza Place	Indian Restaurant	Cosmetics Shop	Bar	Sandwich Place	Gourmet Shop	Juice Bar	Donut Shop
61	RICHMOND HILL	Latin American Restaurant	Pizza Place	Lounge	Bank	Gym / Fitness Center	Bus Station	Moving Target	Supermarket	Caribbean Restaurant	Sandwich Place
91	LONG ISLAND CITY	Hotel	Coffee Shop	Bar	Pizza Place	Mexican Restaurant	Café	Deli / Bodega	Office	Supermarket	Gym / Fitness Center
21	SUNNYSIDE	Pizza Place	Italian Restaurant	Chinese Restaurant	Theater	Bakery	Deli / Bodega	Bar	Coffee Shop	Discount Store	South American Restaurant

Similarly, we can examine each cluster to find out which neighbourhoods suits our best interest by looking at the most common venues.

Finally lets visualize the clustered neighbourhoods using Folium Library.



5. Discussion

The aim of this project is to help people who want to relocate to the Best borough in New York city, expats can choose 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 Cluster 1 has 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 second cluster is suitable. 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 find out more about places before moving into a neighbourhood.We have just taken Price(Budget) as a primary concern to shortlist the best boroughs in New York city.

The future of this project includes taking other factors such as saefty in the areas into consideration to shortlist the borough, such as filtering areas based on a Number of Crimes Recorded in the Borough.