The Battle of Neighborhoods
Based on Housing Prices

COURSERA CAPSTONE FINAL PROJECT REPORT

GAURAV KUMAR

FEB 2020

INDEX

1.INTRODUCTION

1.1.Background:

Housing is the basic necessity of human beings. Whenever a person or company gains some surplus in income or profits, their first priority is to get abode of their own or improve their dwelling or expand their enterprises. Despite the increase in the housing project, the demand-supply equilibrium cannot be achieved. Everyone has a different set of reservations over their choice in buying their dream house. Some people cannot afford their own house, they are in need of an affordable house and some have concerns over the locality and facilities near their house.

1.2.Problem

The NYC Property Sales dataset is a record of every building or building unit (apartment, office space, condos,etc.) sold in the New York City property market over a 12-month period. This housing sales dataset will provide trends in housing prices and can be useful in predicting sales price.

This project also aims at selecting the houses or apartments in a borough-based on housing prices and explore the neighborhoods of each borough and cluster venues into a group of top 10 venues in each neighborhood using the K Means clustering technique and using FourSquare API to get the venue around a neighborhood.

1.3.Interest

A person who is considering buying an apartment, office space, condos, etc. in real estate based on his choice of location in Newyork will be interested in getting the best at the sale price. He also would be interested in the locality around the house interested in and would explore neighborhoods and venues around the neighborhood.

2.DATASETS

2.1.Data Source:

Dataset[1].NYC Property Sales

This dataset is a record of every building or building unit (apartment, etc.) sold in the New York City property market over a 12-month period. This dataset acquired from the KAGGLE dataset(https://www.kaggle.com/new-york-city/nyc-property-sales). This dataset has the following columns:

BOROUGH, NEIGHBORHOOD, BUILDING CLASS CATEGORY, TAX CLASS AT PRESENT, BLOCK, LOT, EASE-MENT, BUILDING CLASS AT PRESENT, ADDRESS, APARTMENT NUMBER,ZIP CODE,RESIDENTIAL UNITS,COMMERCIAL UNITS,TOTAL UNITS,LAND SQUARE FEET,GROSS SQUARE FEET,YEAR BUILT,TAX CLASS AT TIME OF SALE,BUILDING CLASS AT,TIME OF SALE,**SALE PRICE**,SALE DATE

Columns which are used in this project are:

[1].BOROUGH: Newyork state has 5 boroughs namely Manhattan (1), Bronx (2), Brooklyn (3), Queens (4), and Staten Island (5). The name of the borough in which the corresponding property is located.

[2].NEIGHBORHOOD: Neighborhood name in a borough where the property is located.

[3].BUILDING CLASS CATEGORY: Category of property describing whether it is Condos Apartment or for rental or elevator apartment or loft apartment.

[4].SALE PRICE: Corresponding Sale Price of property in real estate.

Dataset[2]. Neighborhoods in New York City

This dataset is scrapped using BeautifulSoup from the Wikipedia page which contains a list of neighborhoods(https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City). The dataset contains the following columns:

1Community_board, Area(km^2), Population census, Pop./km^2, Neighborhoods

This dataset is merely used to get more neighborhood in a borough and also get population census in each neighborhood

Dataset[3].FourSquare API: Foursquare API is used to get the location of each neighborhood and venue around the neighborhood using requests.

Data Cleaning

The Newyork sales dataset has 24 columns out of which only 5 columns are used for this project. The required data frame is sliced from the dataset and the column's name has been for convenience.

```
[515] newyork.head()
```

	Unnamed: 0	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT	ADDRESS	APARTMENT NUMBER	ZIP CODE	RESIDENTI... UNI...
0	4	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2A	392	6		C2	153 AVENUE B		10009	
1	5	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	26		C7	234 EAST 4TH STREET		10009	
2	6	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2	399	39		C7	197 EAST 3RD STREET		10009	
3	7	1	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	2B	402	21		C4	154 EAST 7TH STREET		10009	
				07 RENTALS -						301			

Newyork sales data before preprocessing

```
NEWYORK.head()
```

	HousingID	Neighborhood	BUILDING CLASS CATEGORY	SALE PRICE	YEAR BUILT
0	4	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	6625000	1900
1	5	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	-	1900
2	6	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	-	1900
3	7	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	3936272	1913
4	8	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	8000000	1900

The sale price column of the NEWYORK data frame has a large amount of Nan value due to the fact that these are not transactional sales(Transfer property). The entire row is dropped which has Nan value as this input would not help in predicting future sales.

Dropping the row with Nan value

```
[433] NEWYORK.head()
```

	HousingID	Neighborhood	BUILDING_CLASS_CATEGORY	SALE_PRICE	YEAR_BUILT
0	4	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	6625000	1900
3	7	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	3936272	1913
4	8	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	8000000	1900
6	10	ALPHABET CITY	07 RENTALS - WALKUP APARTMENTS	3192840	1920
9	13	ALPHABET CITY	08 RENTALS - ELEVATOR APARTMENTS	16232000	1920

```
[442] neighborhoods.head()
```

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	WAKEFIELD	40.894705	-73.847201
1	Bronx	CO-OP CITY	40.874294	-73.829939
2	Bronx	EASTCHESTER	40.887556	-73.827806
3	Bronx	FIELDSTON	40.895437	-73.905643
4	Bronx	RIVERDALE	40.890834	-73.912585

Dataset[2] Newyork neighborhood data with their location

Final Dataframe

```
[449] newyork1.head(10)
```

	HousingID	Neighborhood	BUILDING_CLASS_CATEGORY	SALE_PRICE	YEAR_BUILT	Borough	Latitude	Longitude
188	238	CHELSEA	08 RENTALS - ELEVATOR APARTMENTS	4600000	2014	Manhattan	40.744035	-74.003116
189	238	CHELSEA	08 RENTALS - ELEVATOR APARTMENTS	4600000	2014	Staten Island	40.594726	-74.189560
190	243	CHELSEA	08 RENTALS - ELEVATOR APARTMENTS	2341975	2014	Manhattan	40.744035	-74.003116
191	243	CHELSEA	08 RENTALS - ELEVATOR APARTMENTS	2341975	2014	Staten Island	40.594726	-74.189560
660	499	CHELSEA	13 CONDOS - ELEVATOR APARTMENTS	3210237	2013	Manhattan	40.744035	-74.003116
661	499	CHELSEA	13 CONDOS - ELEVATOR APARTMENTS	3210237	2013	Staten Island	40.594726	-74.189560
662	500	CHELSEA	13 CONDOS - ELEVATOR APARTMENTS	5875000	2013	Manhattan	40.744035	-74.003116
663	500	CHELSEA	13 CONDOS - ELEVATOR APARTMENTS	5875000	2013	Staten Island	40.594726	-74.189560
664	501	CHELSEA	13 CONDOS - ELEVATOR APARTMENTS	44105000	2013	Manhattan	40.744035	-74.003116
665	501	CHELSEA	13 CONDOS - ELEVATOR APARTMENTS	44105000	2013	Staten Island	40.594726	-74.189560

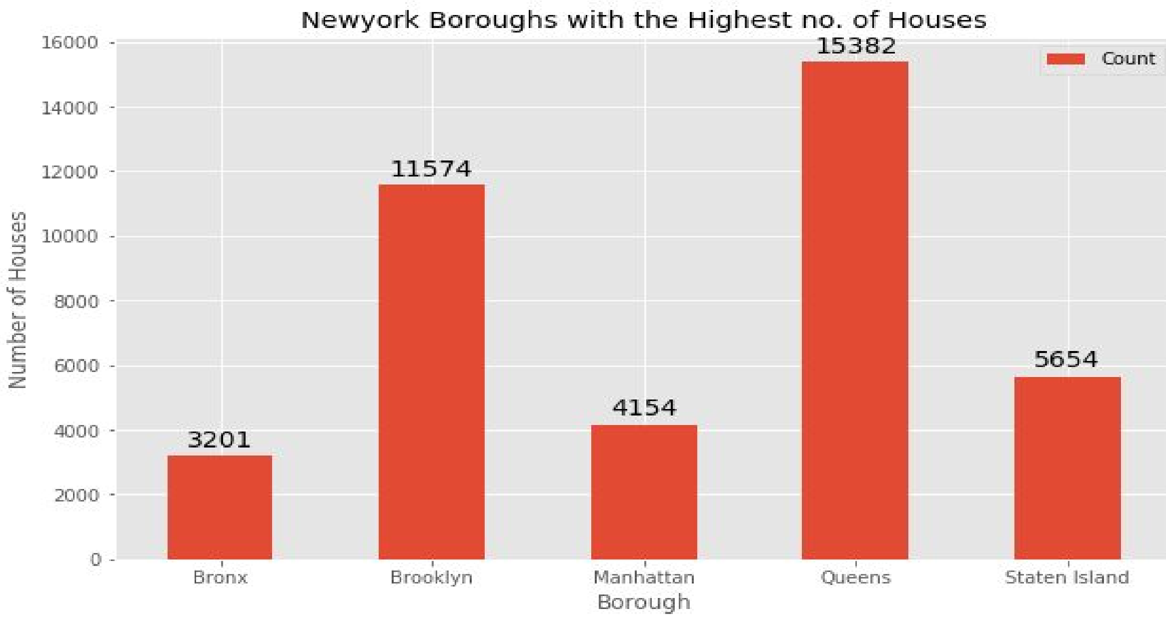
Final merged dataset after preprocessing

The above two datasets are merged into two forms' final dataset which has eight features and each house is now with its location as well as information about which Borough it will be in.

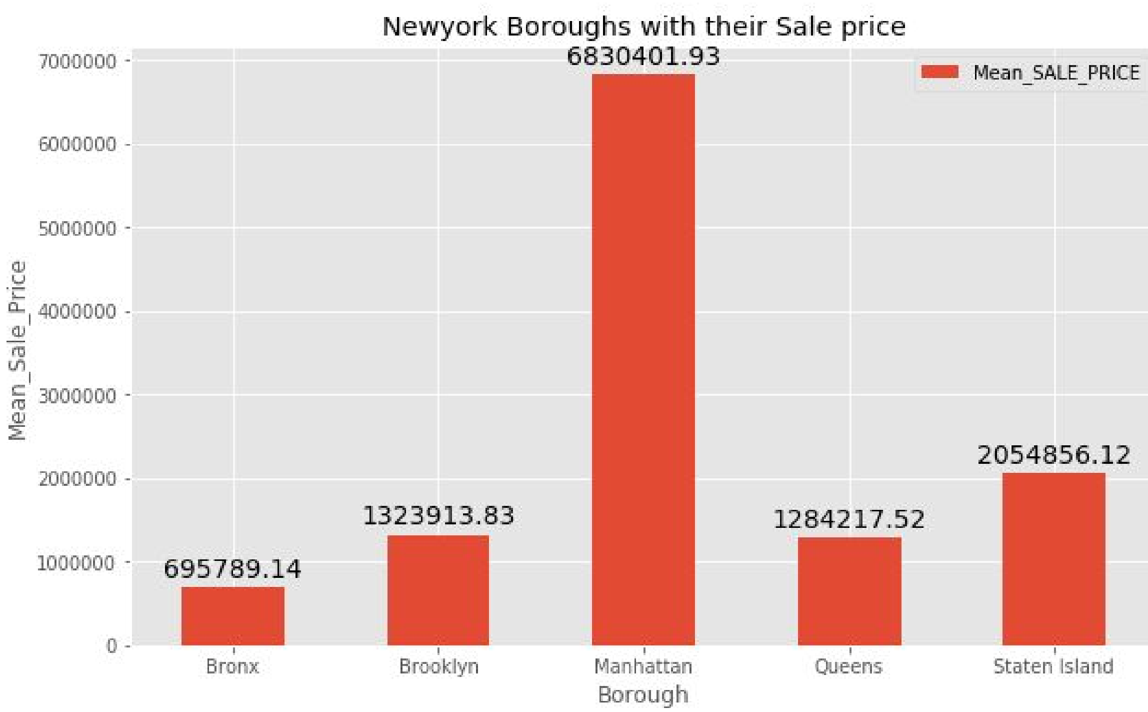
3.METHODOLOGY

3.1.EXPLORATORY DATA ANALYSIS:

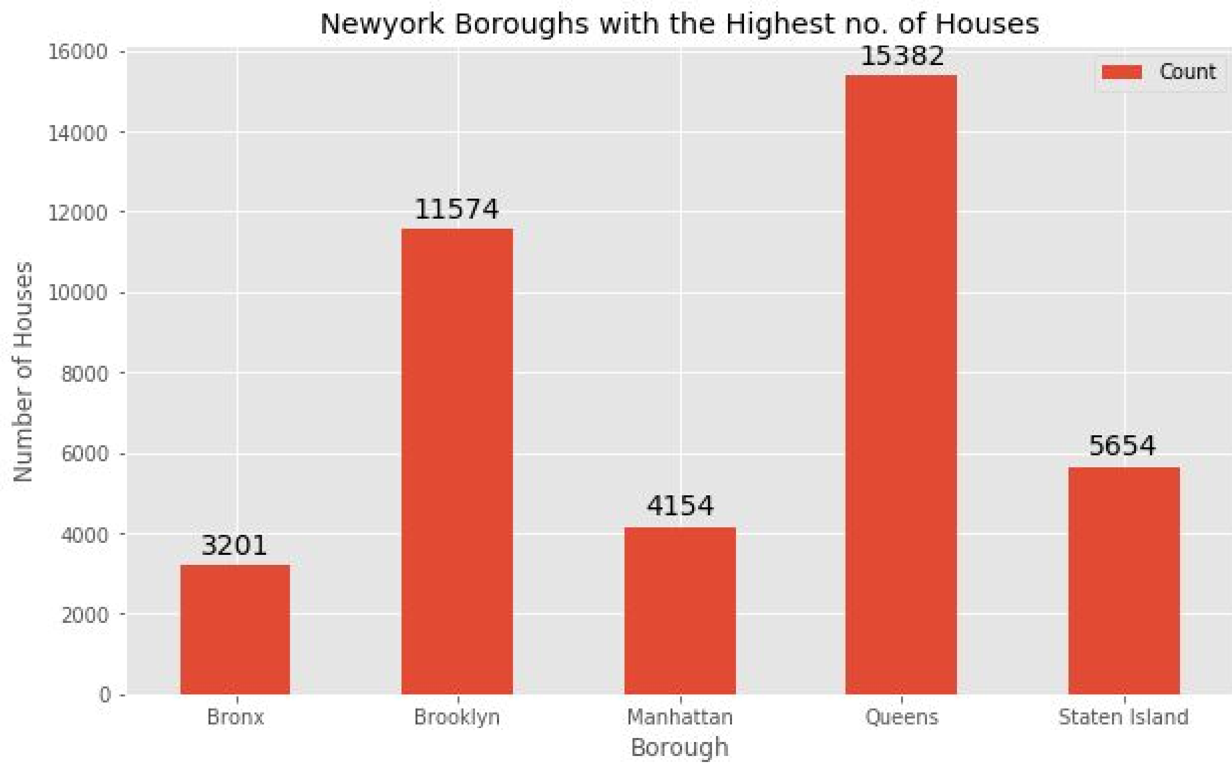
3.1.1.Borough with the Highest Number of Houses sold



This Bar graph compares the number of houses sold in each Borough. This bar graph utilizes Newyork sales dataset in which Houses are built after 2010.



This bar graph shows the mean sale price of houses in each Borough. This bar graph utilizes the same dataset as above,



This bar graph compares the number of houses sold in 2016 in each of Borough. This bar graph utilizes the entire dataset of the NewYork Sales dataset.

4.MODELING

Using the final dataset containing the neighborhoods in Manhattan along with the latitude and longitude, we can find all the venues within a 500-meter radius of each neighborhood by connecting to the Foursquare API. This returns a JSON file containing all the venues in each neighborhood which is converted to a pandas data frame. This data frame contains all the venues along with their coordinates and category, Venue details of each Neighborhood.

One hot encoding is done on the venue's 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 Neighborhood and the mean of the venues is calculated, finally, the 25 common venues are calculated for each of the neighborhoods.

To help people find similar neighborhoods in the borough we will be clustering similar neighborhoods using K - means clustering which is a form of unsupervised machine learning algorithm that clusters data based on predefined cluster size. We will use a cluster size of 10 for this project that will cluster 25 neighborhoods into 10 clusters. The reason to conduct a K- means clustering is to cluster neighborhoods with similar venues together so that people can shortlist the area of their interests based on the venues/amenities around each neighborhood.

5.RESULT

The aim of this project is to help people who want to relocate to the different area of Newyork, any person can choose the neighborhoods to which they want to relocate based on the most common venues in it. For example, if a person is looking for a neighborhood with good connectivity and public transportation we can see that Clusters 3 and 4 have Train stations and Bus stops as the most common venues. If a person is looking for a neighborhood with stores and restaurants in close proximity then the neighborhoods in the first cluster are suitable. For a family I feel that the neighborhoods in Cluster 4 are more suitable due to the common venues in that cluster, these neighborhoods have common venues such as Parks, Gym/Fitness centers, Bus Stops, Restaurants, Electronics Stores and Soccer fields which is ideal for a family.

Cluster 0

Cluster 0

```
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 0, manhattan_merged.columns[[1] + list(range(manhattan_merged.shape[1]))]]
```

	Neighborhood	HousingID	Neighborhood	BUILDING_CLASS_CATEGORY	SALE_PRICE	YEAR_BUILT	Borough	Latitude	Longitude	Cluster Labels	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	Co V
154	SOHO	10702	SOHO	13 CONDOS - ELEVATOR APARTMENTS	26948580	2013	Manhattan	40.722184	-74.000657	0	Women's Store	Clothing Store	Shoe Store	Men's Store	Yoga Studio	Supermarket	Dessert Shop	Dance Studio	C S
155	SOHO	10703	SOHO	13 CONDOS - ELEVATOR APARTMENTS	4200000	2013	Manhattan	40.722184	-74.000657	0	Women's Store	Clothing Store	Shoe Store	Men's Store	Yoga Studio	Supermarket	Dessert Shop	Dance Studio	C S
156	SOHO	10704	SOHO	13 CONDOS - ELEVATOR APARTMENTS	3725000	2013	Manhattan	40.722184	-74.000657	0	Women's Store	Clothing Store	Shoe Store	Men's Store	Yoga Studio	Supermarket	Dessert Shop	Dance Studio	C S
157	SOHO	10705	SOHO	13 CONDOS - ELEVATOR APARTMENTS	2462265	2013	Manhattan	40.722184	-74.000657	0	Women's Store	Clothing Store	Shoe Store	Men's Store	Yoga Studio	Supermarket	Dessert Shop	Dance Studio	C S
158	SOHO	10706	SOHO	13 CONDOS - ELEVATOR APARTMENTS	4350000	2013	Manhattan	40.722184	-74.000657	0	Women's Store	Clothing Store	Shoe Store	Men's Store	Yoga Studio	Supermarket	Dessert Shop	Dance Studio	C S

Cluster 1

Cluster 1

```
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 1, manhattan_merged.columns[[1] + list(range(manhattan_merged.shape[1]))]]
```

	Neighborhood	HousingID	Neighborhood	BUILDING_CLASS_CATEGORY	SALE_PRICE	YEAR_BUILT	Borough	Latitude	Longitude	Cluster Labels	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 M Con Ve
66	EAST VILLAGE	1913	EAST VILLAGE	13 CONDOS - ELEVATOR APARTMENTS	1050000	2012	Manhattan	40.727847	-73.982226	1	Vietnamese Restaurant	Pizza Place	Coffee Shop	American Restaurant	Swiss Restaurant	Pet Café	Park	Des S
67	EAST VILLAGE	1914	EAST VILLAGE	13 CONDOS - ELEVATOR APARTMENTS	1700000	2012	Manhattan	40.727847	-73.982226	1	Vietnamese Restaurant	Pizza Place	Coffee Shop	American Restaurant	Swiss Restaurant	Pet Café	Park	Des S
68	EAST VILLAGE	1915	EAST VILLAGE	13 CONDOS - ELEVATOR APARTMENTS	2600000	2012	Manhattan	40.727847	-73.982226	1	Vietnamese Restaurant	Pizza Place	Coffee Shop	American Restaurant	Swiss Restaurant	Pet Café	Park	Des S
69	EAST VILLAGE	1916	EAST VILLAGE	13 CONDOS - ELEVATOR APARTMENTS	1050000	2012	Manhattan	40.727847	-73.982226	1	Vietnamese Restaurant	Pizza Place	Coffee Shop	American Restaurant	Swiss Restaurant	Pet Café	Park	Des S
70	EAST VILLAGE	1917	EAST VILLAGE	13 CONDOS - ELEVATOR APARTMENTS	2450000	2012	Manhattan	40.727847	-73.982226	1	Vietnamese Restaurant	Pizza Place	Coffee Shop	American Restaurant	Swiss Restaurant	Pet Café	Park	Des S
[illegible]	EAST	[illegible]	EAST	13 CONDOS - ELEVATOR	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Vietnamese	Pizza	Coffee	American	Swiss	Pet	[illegible]	Des

Cluster 2

Cluster 2

```
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 2, manhattan_merged.columns[[1] + list(range(manhattan_merged.shape[1]))]]
```

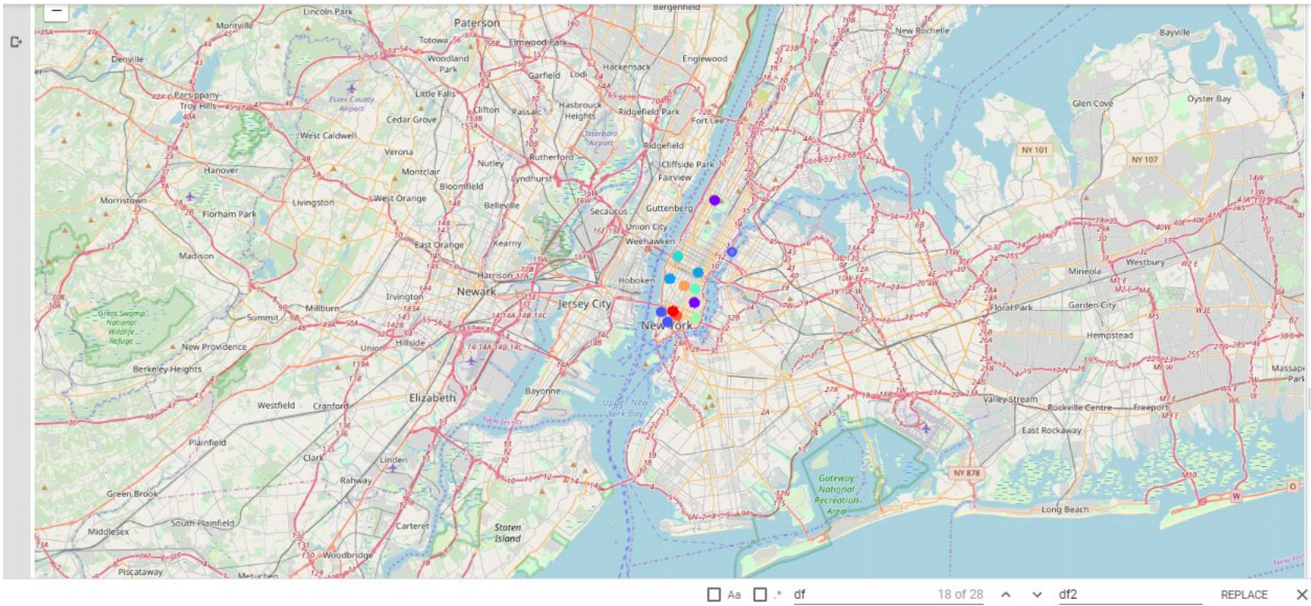
	Neighborhood	HousingID	Neighborhood	BUILDING_CLASS_CATEGORY	SALE_PRICE	YEAR_BUILT	Borough	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Co V
49	CIVIC CENTER	1295	CIVIC CENTER	13 CONDOS - ELEVATOR APARTMENTS	2800000	2010	Manhattan	40.715229	-74.005415	2	Gym / Fitness Center	Coffee Shop	Falafel Restaurant	Spa	Yoga Studio	Nail Salon	Monument / Landmark	Mole Gastron Restau
50	CIVIC CENTER	1296	CIVIC CENTER	13 CONDOS - ELEVATOR APARTMENTS	2498000	2010	Manhattan	40.715229	-74.005415	2	Gym / Fitness Center	Coffee Shop	Falafel Restaurant	Spa	Yoga Studio	Nail Salon	Monument / Landmark	Mole Gastron Restau
51	CIVIC CENTER	1297	CIVIC CENTER	13 CONDOS - ELEVATOR APARTMENTS	2525000	2010	Manhattan	40.715229	-74.005415	2	Gym / Fitness Center	Coffee Shop	Falafel Restaurant	Spa	Yoga Studio	Nail Salon	Monument / Landmark	Mole Gastron Restau
52	CIVIC CENTER	1298	CIVIC CENTER	13 CONDOS - ELEVATOR APARTMENTS	1225000	2010	Manhattan	40.715229	-74.005415	2	Gym / Fitness Center	Coffee Shop	Falafel Restaurant	Spa	Yoga Studio	Nail Salon	Monument / Landmark	Mole Gastron Restau

Cluster 9

Cluster 9

```
manhattan_merged.loc[manhattan_merged['Cluster Labels'] == 9, manhattan_merged.columns[[1] + list(range(manhattan_merged.shape[1]))]]
```

	Neighborhood	HousingID	Neighborhood	BUILDING_CLASS_CATEGORY	SALE_PRICE	YEAR_BUILT	Borough	Latitude	Longitude	Cluster Labels	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
129	LITTLE ITALY	6709	LITTLE ITALY	13 CONDOS - ELEVATOR APARTMENTS	3350000	2012	Manhattan	40.719324	-73.997305	9	Café	Wine Bar	Sandwich Place	Clothing Store	Karaoke Bar	Gourmet Shop	Pizza Place	Optical Shop	Noodle House
130	LITTLE ITALY	6710	LITTLE ITALY	13 CONDOS - ELEVATOR APARTMENTS	6500000	2012	Manhattan	40.719324	-73.997305	9	Café	Wine Bar	Sandwich Place	Clothing Store	Karaoke Bar	Gourmet Shop	Pizza Place	Optical Shop	Noodle House
131	LITTLE ITALY	6724	LITTLE ITALY	13 CONDOS - ELEVATOR APARTMENTS	1200000	2010	Manhattan	40.719324	-73.997305	9	Café	Wine Bar	Sandwich Place	Clothing Store	Karaoke Bar	Gourmet Shop	Pizza Place	Optical Shop	Noodle House



Clustered Neighborhoods of Manhattan Borough

6.OBSERVATION

We have observed using EDA that new houses for sale in 2016 which has been built after 2010 are highest in Brooklyn. Queens has the highest number of houses for sale in 2016 followed by Brooklyn. If we see the average sale price, Manhattan has the highest mean sale price for houses and the Bronx has the least mean sale price for the house.

7.CONCLUSION

This project helps a person get a better understanding of the neighborhoods with respect to the most common venues in that neighborhood. It is always helpful to make use of technology to stay one step ahead i.e. finding out more about places before moving into a neighborhood. We have just taken safety as a primary concern to shortlist the borough of London. The future of this project includes taking other factors such as the cost of living in the areas into consideration to shortlist the borough based on the house sale price, amenities, and venue around the location of the house.