House Sales Prices & Venues Data Analysis of New York City

Xia Zhang

June 2020

1. Introduction

New York City (NYC) is the most populous city in the United States. The estimated population in NYC in 2019 was 8,336,817 distributed over about 302.6 square miles. NYC is also the most densely populated major city in the United States. With almost 20 million people in its metropolitan statistical area and approximately 23 million in its combined statistical area, it is one of the world's most populous megacities. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports [1].

Situated on one of the world's largest natural harbors, New York City is composed of five boroughs, each of which is a county of the State of New York. The five boroughs are: Brooklyn, Queens, Manhattan, the Bronx, and Staten Island. At its core is Manhattan, a densely populated borough that's among the world's major commercial, financial and cultural centers. Its iconic sites include skyscrapers such as the Empire State Building and sprawling Central Park. Broadway theater is staged in neon-lit Times Square.

With the above statistics and information, one kind easily imagine that living in New York City can be very exciting especially if you are a city person. However, it can also be very expensive, especially the housing. While on the other hand, NYC has always been very attractive for investors all over the world, whether to invest in the housing market, opening restaurants, bars, entertainment center, etc. Whatever the perspective objectives are, a comprehensive understanding of different areas in NYC with the corresponding statistics including housing prices, & venues information will be very critical for either a personal buyer or an investor. Although one can easily search for information on the internet for an area of interest, combined information of average housing price and venues information is not readily available. Thus, the primary objective for this study is to perform a comprehensive investigation and analysis of the mean unit housing prices in different neighborhoods of New York City as well as to explore the surrounding venues for each neighborhood and combine this two information to show on the same map.

In this study, eventually, two maps will be created:

1.1 A map of housing located at different neighborhood of New York City for the five boroughs, each borough shown with a different color. A popup label was created for each building labeled with building class category name, borough, neighborhood, and the mean unit housing price.

1.2 A map of clustered neighborhoods of NYC with combined information of mean unit housing price and venues information: popup labels showing the borough, neighborhood, top surrounding venues and the mean unit housing prices. The neighborhoods will also be clustered into 4 clusters by DBSCAN method and a name will be given to each cluster based on the appearance frequency of different venues for each cluster.

2. Data Description

The following data were used for the present study:

- The Department of Finance's Rolling Sales files of New York City, which lists properties that sold in the last twelve-month period in New York City for all tax classes [2]. These files include information on:
    - The neighborhood
    - Building type
    - Square footage
    - Sale price
    - Other data

  Data source : link1 and link2
- Geocoder.arcgis method was used to get the latitudes and longitudes pairs of each neighborhood in the housing sales file.
- Foursquare API was used to explore nearby venues for each neighborhood of NYC [3].

3. Methodology

3.1 Data acquisition and cleaning

In order to be able to combine the housing sales information of New York City with nearby venues data, a file containing both housing sales prices as well as the housing location including borough, neighborhood as well as the zip code will be needed. The nyc-rolling-sales.csv file, which is available both at the Kaggle website as well as the nyc.gov site shown in the links above, contains all this information.

After data was downloaded, data cleaning was performed. There are several issues with the dataset. First, some of the housing/building has 0 sale prices. Based on the data explanation sheet, a $0 sale indicates that there was a transfer of ownership without a cash consideration. There can be a number of reasons for a $0 sale including transfers of ownership from parents to children. Thus, these building sales will not serve my purpose of study and was dropped from the dataset.

Second, for some buildings sold, it has a 0 zip code in the file. As the zip code will be used to retrieve the latitudes and longitudes data, I decided to drop those rows with 0 zip code data as well.

Third, the dataset also has a column indicating the building class code and the building type, e.g, rentals-walk up apartments, condominium, dwelling house, etc. I thought this is very useful information to show along with the sales price on one of the map I was intending to create. But the building code, which is a number code, might be not very useful for the current study. Thus, I decide to split the column name to only show the building class category name.

Fourth, the sales price contained in this file is the sale price for the building, which can be 1 or multiple units. Thus, I divided the sales price with the total units to obtain the unit sale price, whose mean (average across several similar type of sales) will be used to be shown on the maps created.

Furthermore, although there are also square footage information, including the land square feet and the gross square feet, these information was not used for price/square feet calculation, due to the fact that it was not clear whether the square footage really correspond to the perspective units sold.

Finally, the dataset was grouped by borough, neighborhood, zip code and building class category name and the mean/average unit sales price was obtained, which will be shown on the maps to be created.

3.2 Explore the neighborhoods in New York City

First, geocoder.arcgis was used to obtain the latitudes and longitudes coordinates of NYC neighborhoods through the zip codes. To accomplish this, all the unique zip code was extracted and converted to a list. The corresponding latitudes and longitudes coordinates was obtained with arcgis. Then, a dictionary was created with the zip codes as the keys and the latitudes and longitudes pairs as the corresponding values. Finally, the latitudes and longitudes values were filled into the nyc_grouped datasets containing mean house sales prices explained above.

3.3 Explore the nearby venues of NYC to combine with house sales information

Foursquare API was used to explore the neighborhoods and segment them. To further simplify the dataset, I grouped the dataset again by only the borough, neighborhood and zip code. Doing this mean that we are not going to be able to show

the building class category names, which reduce the rows of data from over 4000 to about 600. Then the mean unit sales price was obtained. The drawback of this mean unit sale price to be shown on the second map to be created is that it does not show what type of building was sold. Rather, it is just an average of building sales price based on neighborhood. Thus, the mean unit price information shown on the 2nd map will be much more general compared to the mean unit price shown on the 1st map created above.

To explore nearby venues, I set the radius to be **180 meters** and limit to **30 venues** for each item from their latitude and longitude information to be searched. As there are 613 row in the new grouped dataset, the code will run 613 searches, although some row have duplicate latitude and longitude pairs.

```
nyc_grouped2.head(10)
```

```
(613, 7)
```

	BOROUGH	NEIGHBORHOOD	ZIP CODE	SALE PRICE MEAN	SALE PRICE MAX	SALE PRICE MIN	PRICE/UNIT MEAN
0	Bronx	BATHGATE	10451	4.461000e+06	4461000.0	4461000.0	171576.920000
1	Bronx	BATHGATE	10456	4.000000e+05	400000.0	400000.0	400000.000000
2	Bronx	BATHGATE	10457	6.871700e+05	3000000.0	40000.0	465802.968125
3	Bronx	BATHGATE	10458	9.413929e+05	4052000.0	18000.0	229476.034167
4	Bronx	BAYCHESTER	10466	4.748387e+05	3000000.0	10.0	328506.865736
5	Bronx	BAYCHESTER	10469	3.938067e+05	4750000.0	1.0	281702.618589
6	Bronx	BAYCHESTER	10475	8.860018e+05	11000000.0	123000.0	722394.645417
7	Bronx	BEDFORD PARK/NORWOOD	10458	6.830143e+05	7000000.0	10.0	259213.154110
8	Bronx	BEDFORD PARK/NORWOOD	10467	9.544206e+05	7375000.0	10.0	438735.046731
9	Bronx	BEDFORD PARK/NORWOOD	10468	9.497747e+05	11118000.0	72000.0	275719.651379

```
nyc_grouped2.head(3)
```

	BOROUGH	NEIGHBORHOOD	ZIP CODE	LATITUDE	LONGITUDE	SALE PRICE MEAN	SALE PRICE MAX	SALE PRICE MIN	PRICE/UNIT MEAN
0	Bronx	BATHGATE	10451	40.819986	-73.918433	4.461000e+06	4461000.0	4461000.0	171576.920000
1	Bronx	BATHGATE	10456	40.833955	-73.896685	4.000000e+05	400000.0	400000.0	400000.000000
2	Bronx	BATHGATE	10457	40.848111	-73.903813	6.871700e+05	3000000.0	40000.0	465802.968125

The output file of the searched venues was called “nyc_venues”. A snap shot of the file was shown below:

```
nyc_venues.shape
```

```
(4621, 9)
```

```
nyc_venues.head()
```

	Borough	Neighborhood	Zip Code	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bronx	BATHGATE	10451	40.819986	-73.918433	Pepe Joes	40.820838	-73.917523	Pizza Place
1	Bronx	BATHGATE	10456	40.833955	-73.896685	JC Wines & liqours	40.834490	-73.896980	Liquor Store
2	Bronx	BATHGATE	10456	40.833955	-73.896685	Prospect Gourmet Deli	40.833674	-73.896681	Deli / Bodega
3	Bronx	BATHGATE	10456	40.833955	-73.896685	Fine Fare Supermarket	40.834043	-73.894640	Supermarket
4	Bronx	BATHGATE	10456	40.833955	-73.896685	FOOT LOCKER - CONCOURSE PLAZA	40.834428	-73.895602	Shoe Store

In summary, 247 unique venue categories was returned from the search.

3.4 Analyze each neighborhood

By applying pandas.get_dummies method, a file (nyc_onehot) containing the encoded venue category information was obtained. The borogh, neighborhood, zip code and a combined location information was also included into this nyc_onehot file. Then the nyc_onehot file was grouped by borough, neighborhood, zip code and location with the mean unit sale price obtained.

To better understand venues distribution characteristics for each neighborhood, I sorted the venues to return different levels of most common venues for each neighborhood. Below is the table created:

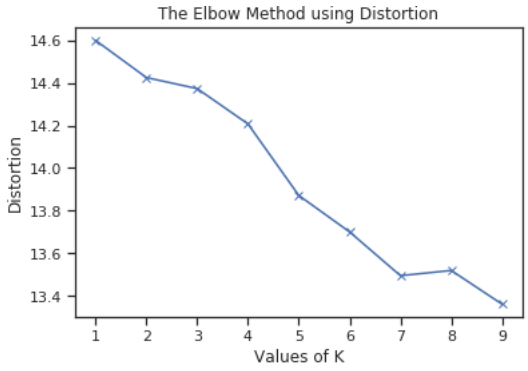
	Borough	Neighborhood	Zip_Code	Location	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Bronx	BATHGATE	10451	Bronx,BATHGATE,10451	Pizza Place	Yoga Studio	Entertainment Service	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Filipino Restaurant	Fast Food Restaurant
1	Bronx	BATHGATE	10456	Bronx,BATHGATE,10456	Deli / Bodega	Liquor Store	Shoe Store	Supermarket	Bus Station	Yoga Studio	Fast Food Restaurant	Exhibit	Falafel Restaurant	Farmers Market
2	Bronx	BATHGATE	10457	Bronx,BATHGATE,10457	Deli / Bodega	Sandwich Place	Grocery Store	Yoga Studio	Entertainment Service	Food & Drink Shop	Food	Flower Shop	Flea Market	Filipino Restaurant
3	Bronx	BATHGATE	10458	Bronx,BATHGATE,10458	Coffee Shop	Sandwich Place	Wine Bar	Yoga Studio	Fast Food Restaurant	Event Space	Exhibit	Falafel Restaurant	Farmers Market	Filipino Restaurant
4	Bronx	BAYCHESTER	10466	Bronx,BAYCHESTER,10466	Diner	Yoga Studio	Ethiopian Restaurant	Food Truck	Food Court	Food & Drink Shop	Food	Flower Shop	Flea Market	Filipino Restaurant

3.5 Cluster the neighborhoods

Clustering, an unsupervised learning process, is unique way for us to understand data by dividing the entire data into groups/clusters based on the patterns in the data.

First, I tried K-means clustering method to cluster the neighborhoods. K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid [4].

To determine the best number of K, the elbow method using distortion was performed. However, no elbow was found from the graph. This does not mean that there are no clusters in the data. No elbow means that the algorithm used cannot separate clusters. Thus, I decided to use the DBSCAN method instead.



DBSCAN [5]:

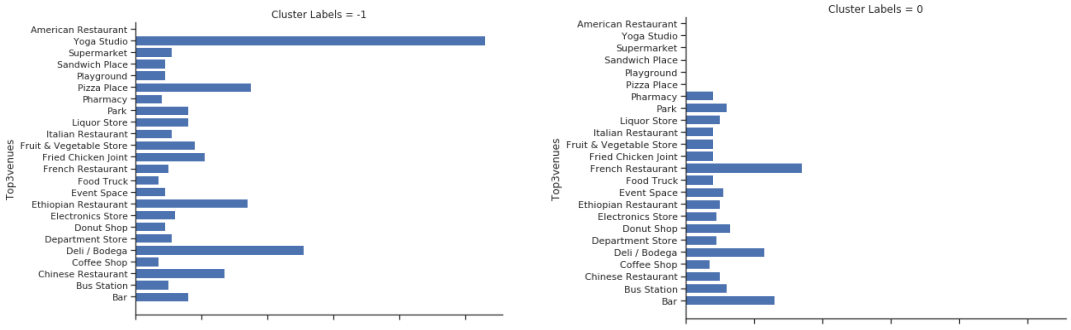
- stands for “density based spatial clustering of applications with noise”
- does not require the user to set the number of clusters a priori
- can capture clusters of complex shapes
- can identify points that are not part of any cluster (very useful as outliers detector)
- is somewhat slower than agglomerative clustering and k-means, but still scales to relatively large datasets.
- works by identifying points that are in crowded regions of the feature space, where many data points are close together (dense regions in feature space)
- Points that are within a dense region are called core samples (or core points)

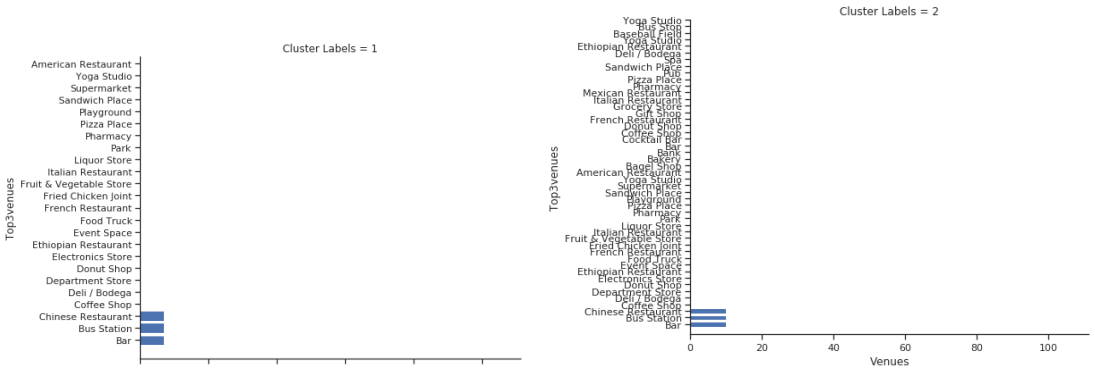
A total of 4 clusters was obtained from DBSCAN:

```
unique_labels = set(labels_db)
unique_labels
```

```
{-1, 0, 1, 2}
```

However, I would like to show the cluster on the popup label of the map as well. It would be more useful to show the cluster with a name indicating its features by analyzing the venues data. Thus, a bar chart of the venues frequency which is the sum of the top 1, 2 & 3 most common venues (only selected the top 20 venues for the top 1, 2 & 3 most common venues by sorting the venues frequency) was created using seaborn facetgrid method. See the chart below:





Based on the bar charts above, I name the clusters as the following:

- Cluster -1: Social, Shopping & Entertainment Complex
- Cluster 0: French Restaurant, Bar & Misc
- Cluster 1: Chinese Restaurant & Bar_1
- Cluster -1: Chinese Restaurant & Bar_2

Note that as cluster 1 and -1 show pretty simiar features and they both don't have a lot of venues nearby, I just name them based on the several venues they had and give a number 1 and 2 to distinguish them.

4. Results

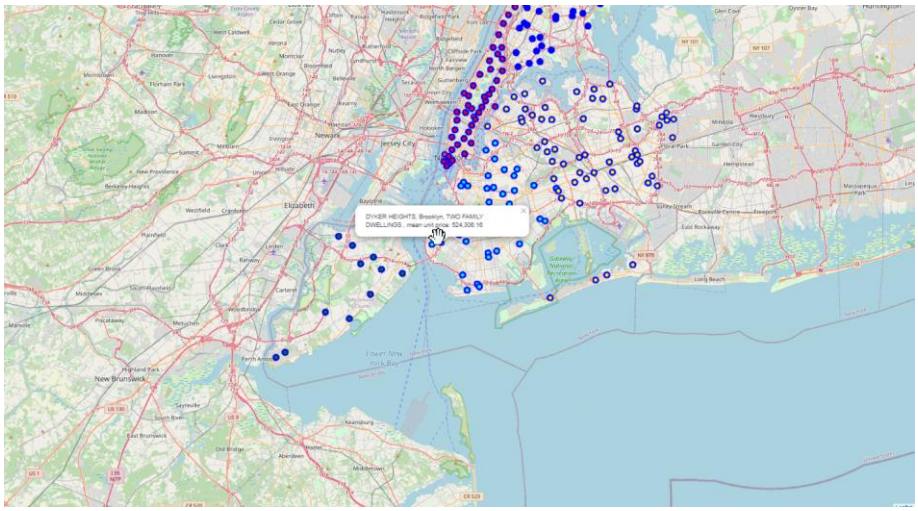
4.1 Map 1 (correspond to section 1.1):

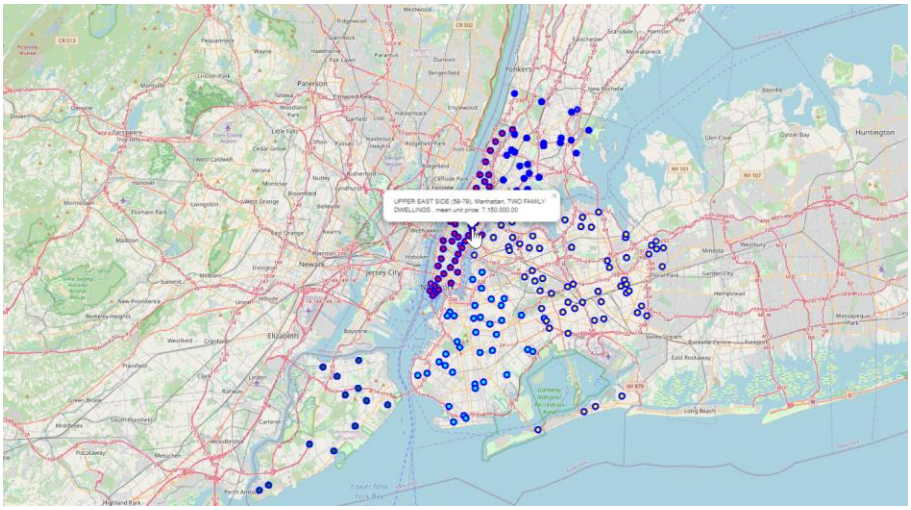
The grouped data used to create the 1st map I explained in section 1.1 is shown below:

```
nyc_grouped.head(3)
```

	BOROUGH	NEIGHBORHOOD	ZIP CODE	BUILDING CLASS CATEGORY NAME	LONGITUDE	LATITUDE	SALE PRICE MEAN	SALE PRICE MAX	SALE PRICE MIN	PRICE/UNIT MEAN	PRICE/GROSS SQFT MEAN	PRICE/LAND SQFT MEAN
0	Bronx	BATHGATE	10451	RENTALS - WALKUP APARTMENTS	-73.918433	40.819986	4461000.0	4461000.0	4461000.0	171576.92	244.0	1323.0
1	Bronx	BATHGATE	10456	RELIGIOUS FACILITIES	-73.896685	40.833955	400000.0	400000.0	400000.0	400000.00	57.0	160.0
2	Bronx	BATHGATE	10457	COMMERCIAL GARAGES	-73.903813	40.848111	2100000.0	2500000.0	1700000.0	1675000.00	inf	573.5

The 1st map was created with python **folium** library: 1.1 A map of housing located at different neighborhood of New York City for the five boroughs, each borough shown with a different color. A couple of snapshot of the map are shown below:





As can be seen, each borough was denoted by a different fill color. And each group of building sold was denoted by its location, especially its zip code on the map. A popup label was created to display some key information including the neighborhood, borough name, building class category name, and the mean unit price sold for this type of building. This map can be very useful for people who are looking for buying a either an apartment, a house, or investing apartments. Building clicking on the map, one can easily find the mean unit sale prices as well as its neighborhood and how the price is compared to similar or different building in the same or different neighborhoods.

4.2 Map 2 (correspond to section 1.2)

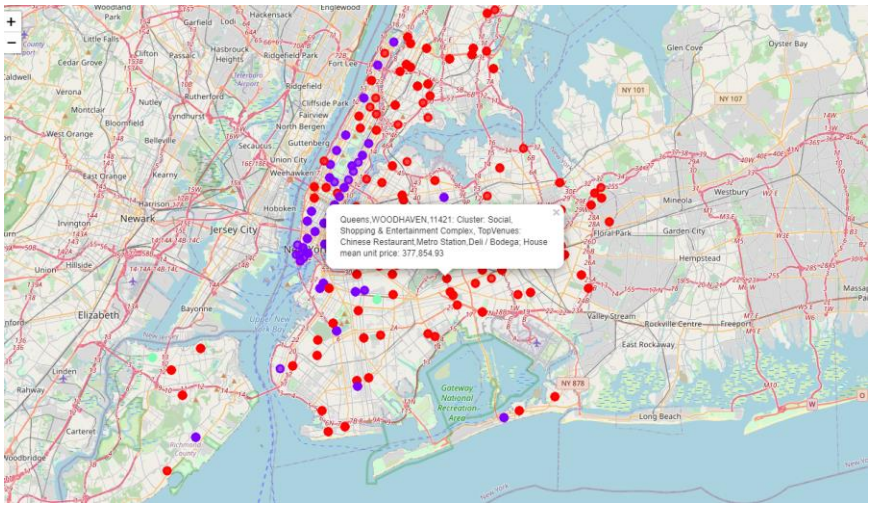
The venues data searched was merged with the nyc_grouped 2 data, which includes the mean unit building sales price info. A snap shot of the merged data was shown below:

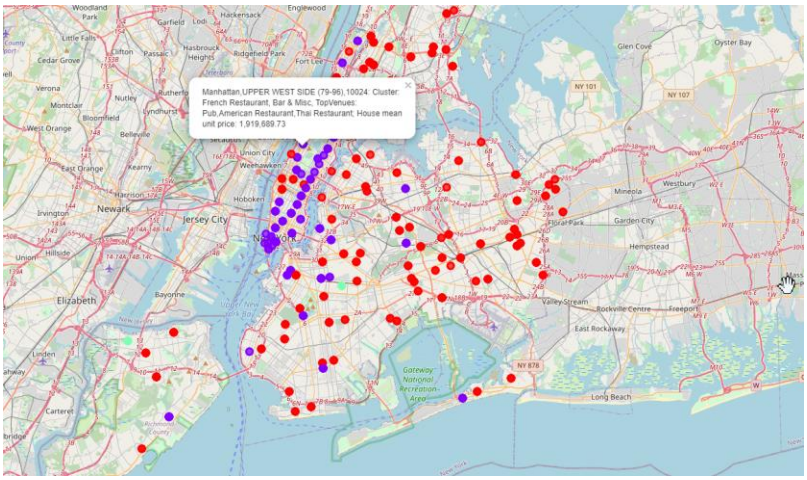
	Borough	Neighborhood	Zip_Code	Location	Latitude	Longitude	Sale Price Mean	Sale Price Max	Sale Price Min	Price/Unit Mean	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6 C
0	Bronx	BATHGATE	10451	Bronx,BATHGATE,10451	40.819986	-73.918433	4.461000e+06	4461000.0	4461000.0	171576.920000	Social, Shopping & Entertainment Complex	Pizza Place	Yoga Studio	Ethiopian Restaurant	French Restaurant	Food Truck	
1	Bronx	BATHGATE	10456	Bronx,BATHGATE,10456	40.833955	-73.896685	4.000000e+05	400000.0	400000.0	400000.000000	Social, Shopping & Entertainment Complex	Liquor Store	Playground	Deli / Bodega	Supermarket	Shoe Store	
2	Bronx	BATHGATE	10457	Bronx,BATHGATE,10457	40.848111	-73.903813	6.871700e+05	3000000.0	40000.0	465802.968125	Social, Shopping & Entertainment Complex	Deli / Bodega	Grocery Store	Sandwich Place	Yoga Studio	Ethiopian Restaurant	
3	Bronx	BATHGATE	10458	Bronx,BATHGATE,10458	40.862059	-73.887575	9.413929e+05	4052000.0	18000.0	229476.034167	Social, Shopping & Entertainment Complex	Coffee Shop	Theater	Sandwich Place	Café	Yoga Studio	
4	Bronx	BAYCHESTER	10466	Bronx,BAYCHESTER,10466	40.887857	-73.827943	4.748387e+05	3000000.0	10.0	328506.865736	Social, Shopping & Entertainment Complex	Diner	Business Service	Yoga Studio	Event Space	French Restaurant	

Another thing I did was that I created another column which shows the joined name of the top 3 most common venues, named “Top_Common_Venues”. This column is included in the merged file shown above.

Most ımon /enue	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	Top_Common_Venues
opian ıurant	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Filipino Restaurant	Pizza Place,Yoga Studio,Ethiopian Restaurant
Deli / ɔdega	Supermarket	Shoe Store	Bus Station	Yoga Studio	Fast Food Restaurant	Exhibit	Falafel Restaurant	Liquor Store,Playground,Deli / Bodega
dwich Place	Yoga Studio	Ethiopian Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Deli / Bodega,Grocery Store,Sandwich Place
dwich Place	Café	Yoga Studio	Farmers Market	Exhibit	Falafel Restaurant	Fast Food Restaurant	Ethiopian Restaurant	Coffee Shop,Theater,Sandwich Place
Yoga ıtudio	Event Space	French Restaurant	Food Truck	Food Court	Food & Drink Shop	Flower Shop	Flea Market	Diner,Business Service,Yoga Studio

Finally, we can visualize the data in a map using **Folium**:





As can be seen, the map shows clustered neighborhoods of NYC with combined information of mean unit housing price and venues information. The popup labels display the borough, neighborhood, top surrounding venues and the mean unit housing prices. This map will be very useful for house buyer or investors to review and compare housing sale prices as well as what is the characteristic venues nearby. The top venues information shown on the popup label will help the buyers or investors decide where might better suite their purpose when making decision on buying or investing a property.

5. Discussion

In the current study, two maps were created: one shows the NYC housing sold at different neighborhoods in the 5 boroughs with different color notations. Each house sold was shown on the map with a popup label with information including: neighborhood, borough name, building class category name, and the mean unit price sold for this type of building. This map will be useful to house buyers or investors to explore the type of buildings/housing they are interested in and to compare the neighborhoods as well as the price difference across different neighborhoods and boroughs.

For the second map created, nearby venues information was combined with the average unit sale price for housing in that neighborhood. The detailed information of building class category name was sacrificed to only obtain an overall average price for all the different building sold in that neighborhood. The reason for this is to reduce the amount of data for nearby venues exploration and to simply the display on the map. Then both K-means clustering and DBSCAN methods was applied to the data for neighborhood clustering and DBSCAN method was chosen because no elbow was found in the elbow method chart. The popup labels display the borough, neighborhood, top surrounding venues and the mean unit housing prices. This map will be very useful for house buyer or investors to review and compare housing sale prices as well as what is the characteristic venues nearby.

Combined use of Map 1 and Map 2 created in this study will be even more powerful for house buyers or investors to get more comprehensive information on

average housing price, building class category names, neighborhood and top venues nearby.

6. Conclusion

In this study, two maps were created to help especially house buyers or investors to explore different types of building units and compare the prices based on building types as well neighborhoods changes. The second map created can be further used to explore nearby venues to get more comprehensive understanding of the buildings/housing in New York City.

References:

[1] https://en.wikipedia.org/wiki/New_York_City

[2] https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

[3] Foursquare API

[4]https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

[5]https://towardsdatascience.com/dbscan-clustering-for-data-shapes-k-means-cant-handle-well-in-python-6be89af4e6ea