Correlation between home price and nearby facilities in New York City

1. Background

The price of a property could be affected by many factors, including property size, age and condition, the market, and its location. Among which, location is one of the most important influencers. For example, both school system quality and crime rates in a neighborhood are correlated with home price. Besides, nearby features, such as shopping centers, restaurants and other pulic entertainments could also affect home price.

New York City is decribed as the cultutal, financial and media capital of the world, and it is a prominent location for the America entertainment industry. There are five boroughs in New York City: Manhattan, Brooklyn, Queens, The Bronx and Staten Island. Each of them has its own distinct characteristics. For example, Manhattan is the most famous borough among the five, it is the symbol of New York City, and most of the city's skyscrapers and prominent landmarks are located in this borough. Brooklyn is known for its cultural, social and ethnic diversity, and it has the largest central core in the outer boroughs. Queens is geographically the largest borough, and it is one of the most busiest borough because two of the three busiest airports are located here. The Bronx is the northernmost borough and it is the location of Yankee Stadium, Bronx Zoo, and the New York Botanical Garden. Staten Island is the most suburban in the five boroughs, but it has the most beautiful natural scenery.

Here, I used New York City as an example, to examine the venue characteristics for each borough, and how nearby facilities, such as entertainments, restaurants and shops affect home price in different boroughs.

According to this project, real estate developer could better understand whether and how nearby facilities affect home prices, and this project can be a reference for them to choose real estate locations.

2. Data

The home price data in New York City used in this project was from NYC OpenData website (https://data.cityofnewyork.us/City-Government/DOF-Summary-of-Neighborhood-Sales-by-Neighborhood-/5ebm-myj7). Home price data was collected and maintained by the

Department of Finance, covering sales of family homes in each borough, and it is a summary of neighborhood sales for 1,2 and 3 family homes.

The location and venue data was from Foursquare (https://foursquare.com). I mianly used venue location and category data in my project.

New York City neighborhood information and location data was from data.beta.nyc website (https://data.beta.nyc/dataset/pediacities-nyc-neighborhoods/resource/35dd04fb-81b3-479b-a074-a27a37888ce7). This dataset includes neighborhood name, borough name, and neighborhood coordinates.

New York City Borough boundaries data was downloaded from Data.gov (https://catalog.data.gov/dataset/borough-boundaries). I used geojson data file which includes borough name and boundaries coordinates features.

3. Methodology

3.1 Data cleaning

Neighborhood information and location dataset is a json data file, and I extracted borough, neighborhood, latitude and longitude features from this json file, and transformed the data into a pandas data frame. Fortunately, there was no missing values in this dataset, and there were 310 entries in total.

Home sales data set includes 9 features: borough, neighborhood, type of home, number of sales, lowest sale price, highest sale price, average sale price, median sale price and year. And there are 5989 entries in total.

I was not interested in type of home and year, so I dropped these columns. Because I mainly focused on neighborhood differences and borough differences, so I grouped the data by neighborhood and borough respectively, then calculated the mean value.

In order to merge this home sales data frame with neighborhood and borough location data frame, I set the first character in every word in neighborhood and borough column to upper case, then renamed the column name with uppercase first character as well.

I did not deal with outliers, because one of my purpose is to find out where these extremely high or low sale price homes are located in, and whether nearby facility is the reason why they have such high or low prices.

I used Foursquare API to get the venue data. Because borough has a quiet large area, so I just explored the venues on the basis of neighborhood in each borough. First I extracted neighborhood information for each borough, and assigned them to new data frames. Secondly, I put neighborhood latitude and longitude as input, to make a GET request in order to get venue information for each neighborhood location in the radius of 500. Finally, I transformed venue data into data frame, with features including neighborhood name, neighborhood coordinate, venue name, venue coordinate and venue category.

Because I focused on venue category and neighborhood, I transform venue data frame into dummy table. By doing this, categorical variable can be analyzed as numeric variable, and this is suitable for further analysis and model building.

3.2 Exploratory data analysis

In order to better understand the difference in home price between neighborhoods, I created a choropleth map with neighborhood and average sale price. By doing this, I can directly find out where high price homes are located in. However, some neighborhoods are lack of home price data. Besides, I also created a choropleth map with borough and median sale price. Median sale price is less likely to be affected by outliers, by doing this, I can easily find out the differences in home price between boroughs.

Then I firstly explored the venue differences between boroughs. Because all the boroughs have quiet large areas, so I got nearby venue data based on neighborhood location in each borough. Then I calculated the frequency of each venue category in every neighborhood. Then I got the sum of neighborhood nearby venue frequency in each borough, and get a table with the 20 highest venue frequency in descending order for each borough.

After doing this analysis for each borough, I merged five borough venue frequency tables into one table in order to better analyze the venue differences between boroughs.

Finally, I tried to explore the differences between neighborhoods. I set all the neighborhoods into three bins according to their median home price: low price, medium price, and high price. Afterwards, I got the nearby venue information for each group, and found the top 20 frequent venues. As I did for the boroughs, I also merged three tables into one, in order to directly compare the differences among these three home price groups.

3.3 Machine learning

Firstly, I used k-means to cluster five boroughs into 3 groups based on their home prices. Then, I used also applied k-means on venue category to cluster five boroughs into 3 clusters. By doing this, I can see whether the clusters for borough based on their venue category is similar to clusters based on their home price.

I also used k-means to cluster neighborhoods into 3 clusters based on their venue categories. Then I can check whether there is overlap between these three clusters and their home price (low, medium, and high) groups.

Besides, I also tried to use K-nearest-neighbors (KNN) model and logistic regression to classify neighborhoods home prices based on venue categories. I divided the whole dataset into training set and testing set, in order to test the accuracy of these models. Then I trained the model using training set, and test the model accuracy using accuracy score.

4. Result

4.1 Borough differences

According to the choropleth map, Manhattan is the borough with the highest home price (Figure 1). Brooklyn is the second one, but its home price is much lower than Manhattan’s. The following three boroughs are Queens, the Bronx and Staten Island. However, the difference between boroughs with low and medium home price is not very significant (Figure 2). Therefore, my goal here is to find out whether there is significant difference in venue category between Manhattan and other boroughs.

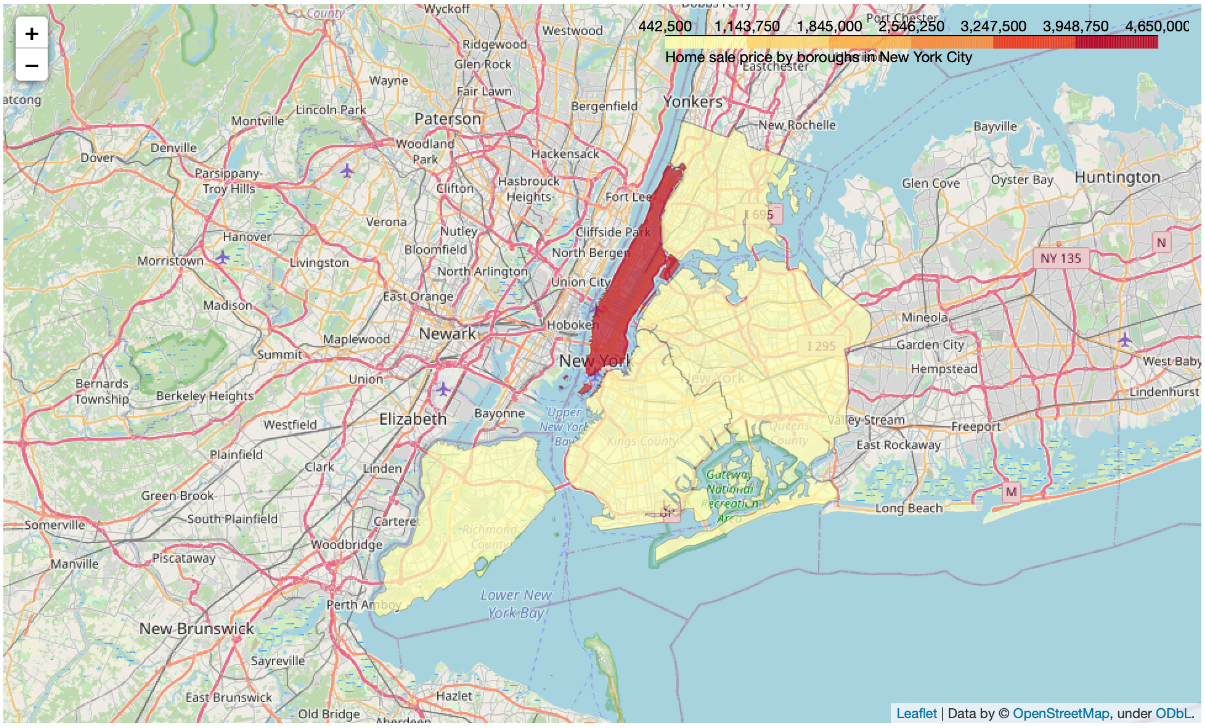

442,500
1,143,750
1,845,000
2,546,250
3,247,500
3,948,750
4,650,000
Home sale price by boroughs in New York City
+
−
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 1. Choropleth map on median home sale price difference between five boroughs in New York City

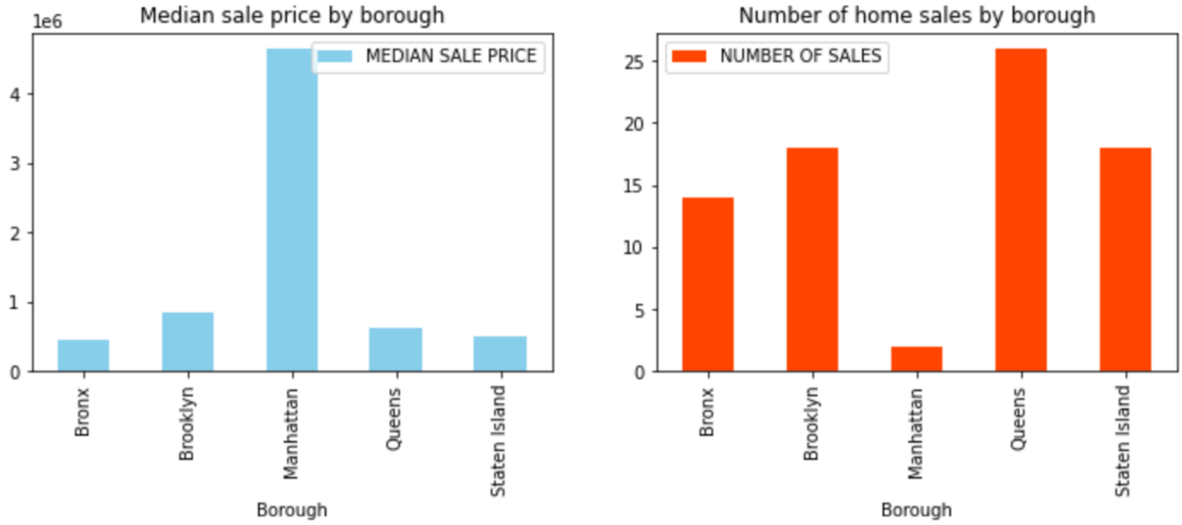
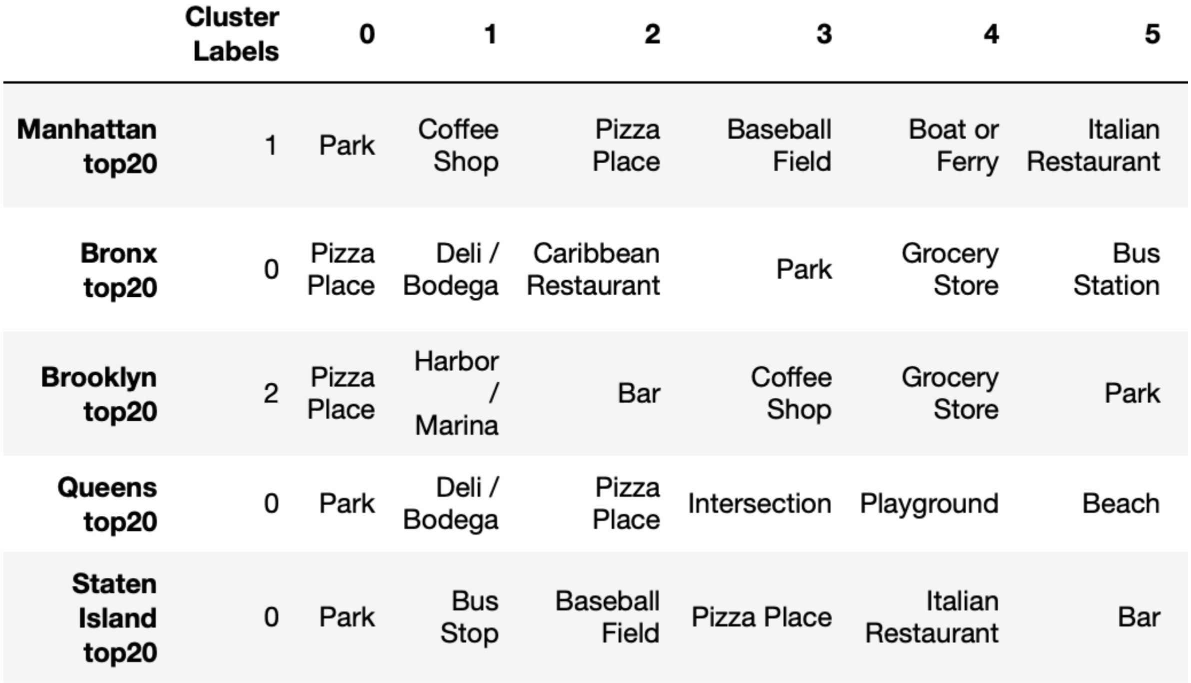


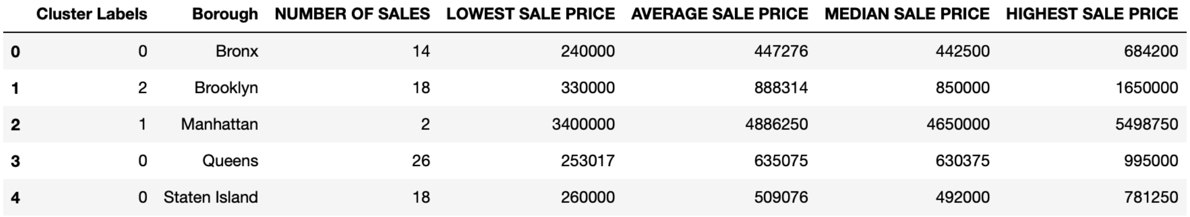
Figure 2. Median home sale price in five boroughs (left); number of home sale in five boroughs (right).

Then I used k-means to cluster boroughs on the basis of their home price. The result shows that Manhattan is a unique cluster, Brooklyn is another cluster, and the Bronx, Queens, and Staten Island formed one cluster (Table 1). The result of k-means based on venue category and their frequency illustrated the same pattern, which is Manhattan itself is one cluster, Brooklyn is one cluster, and another cluster consist of the Bronx, Queens, and Staten Island (Table 2).



	Cluster Labels	0	1	2	3	4	5
Manhattan top20	1	Park	Coffee Shop	Pizza Place	Baseball Field	Boat or Ferry	Italian Restaurant
Bronx top20	0	Pizza Place	Deli / Bodega	Caribbean Restaurant	Park	Grocery Store	Bus Station
Brooklyn top20	2	Pizza Place	Harbor / Marina	Bar	Coffee Shop	Grocery Store	Park
Queens top20	0	Park	Deli / Bodega	Pizza Place	Intersection	Playground	Beach
Staten Island top20	0	Park	Bus Stop	Baseball Field	Pizza Place	Italian Restaurant	Bar

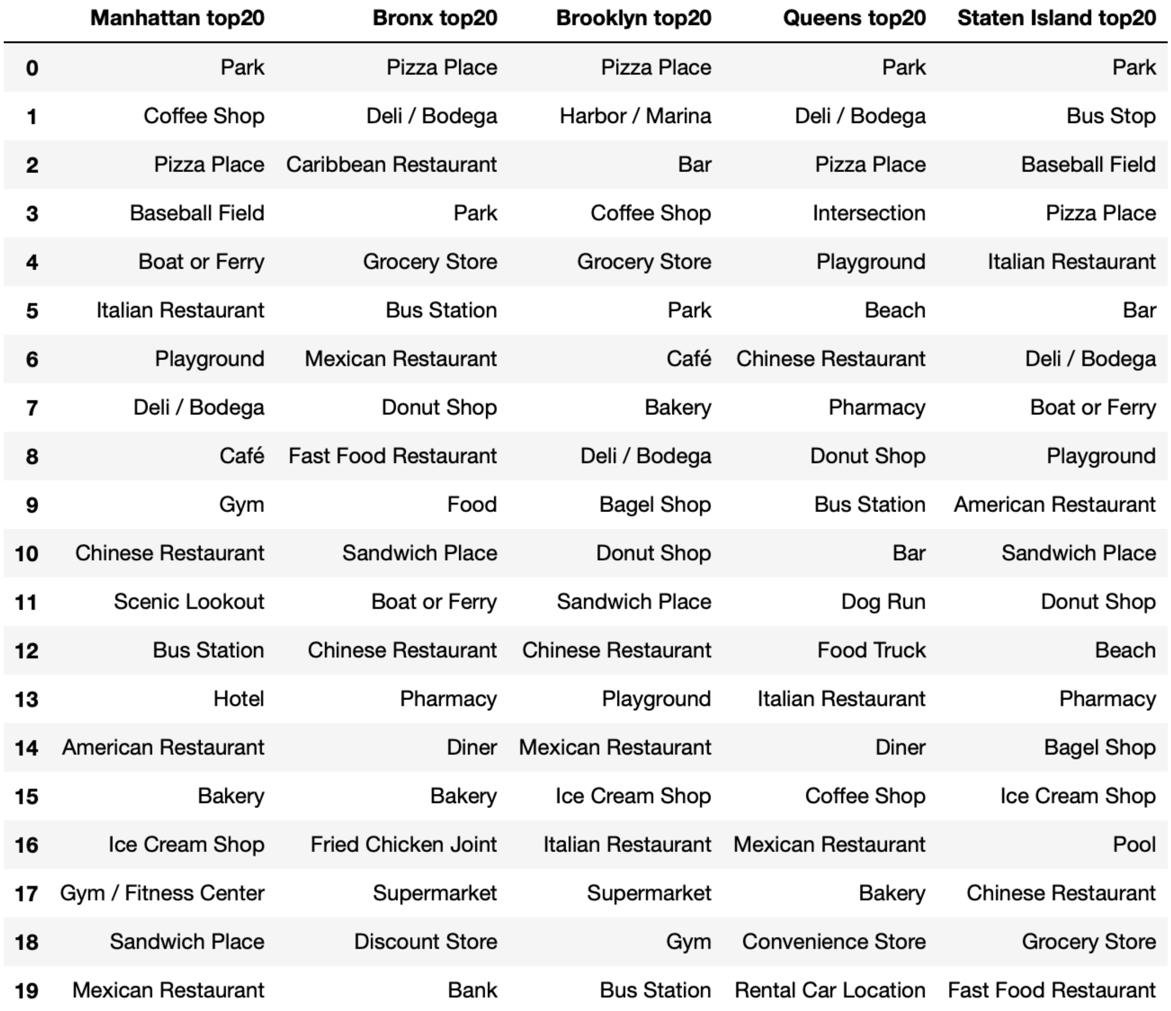
Table 1. K-means clustering result based on nearby venue category and venue category frequency.



	Cluster Labels	Borough	NUMBER OF SALES	LOWEST SALE PRICE	AVERAGE SALE PRICE	MEDIAN SALE PRICE	HIGHEST SALE PRICE
0	0	Bronx	14	240000	447276	442500	684200
1	2	Brooklyn	18	330000	888314	850000	1650000
2	1	Manhattan	2	3400000	4886250	4650000	5498750
3	0	Queens	26	253017	635075	630375	995000
4	0	Staten Island	18	260000	509076	492000	781250

Table 2. K-means clustering result based on home sale price.

According to top 20 frequent venue category in each borough, I found that Manhattan has more parks and fields, as well as fitness facilities than other boroughs (Table 3). However, it has less venues for shopping than other boroughs.



	Manhattan top20	Bronx top20	Brooklyn top20	Queens top20	Staten Island top20
0	Park	Pizza Place	Pizza Place	Park	Park
1	Coffee Shop	Deli / Bodega	Harbor / Marina	Deli / Bodega	Bus Stop
2	Pizza Place	Caribbean Restaurant	Bar	Pizza Place	Baseball Field
3	Baseball Field	Park	Coffee Shop	Intersection	Pizza Place
4	Boat or Ferry	Grocery Store	Grocery Store	Playground	Italian Restaurant
5	Italian Restaurant	Bus Station	Park	Beach	Bar
6	Playground	Mexican Restaurant	Café	Chinese Restaurant	Deli / Bodega
7	Deli / Bodega	Donut Shop	Bakery	Pharmacy	Boat or Ferry
8	Café	Fast Food Restaurant	Deli / Bodega	Donut Shop	Playground
9	Gym	Food	Bagel Shop	Bus Station	American Restaurant
10	Chinese Restaurant	Sandwich Place	Donut Shop	Bar	Sandwich Place
11	Scenic Lookout	Boat or Ferry	Sandwich Place	Dog Run	Donut Shop
12	Bus Station	Chinese Restaurant	Chinese Restaurant	Food Truck	Beach
13	Hotel	Pharmacy	Playground	Italian Restaurant	Pharmacy
14	American Restaurant	Diner	Mexican Restaurant	Diner	Bagel Shop
15	Bakery	Bakery	Ice Cream Shop	Coffee Shop	Ice Cream Shop
16	Ice Cream Shop	Fried Chicken Joint	Italian Restaurant	Mexican Restaurant	Pool
17	Gym / Fitness Center	Supermarket	Supermarket	Bakery	Chinese Restaurant
18	Sandwich Place	Discount Store	Gym	Convenience Store	Grocery Store
19	Mexican Restaurant	Bank	Bus Station	Rental Car Location	Fast Food Restaurant

Table 3. Top 2 frequent venue categories in five boroughs.

4.2 Neighborhood differences

According to the choropleth map (Figure 3) and box plot (Figure 4), I got the neighborhoods with relative high home price: Tribeca and Little Italy. Not surprisingly, they both located in Manhattan. And I also found in the top 10 home price neighborhoods, 8 of them are located in Manhattan.

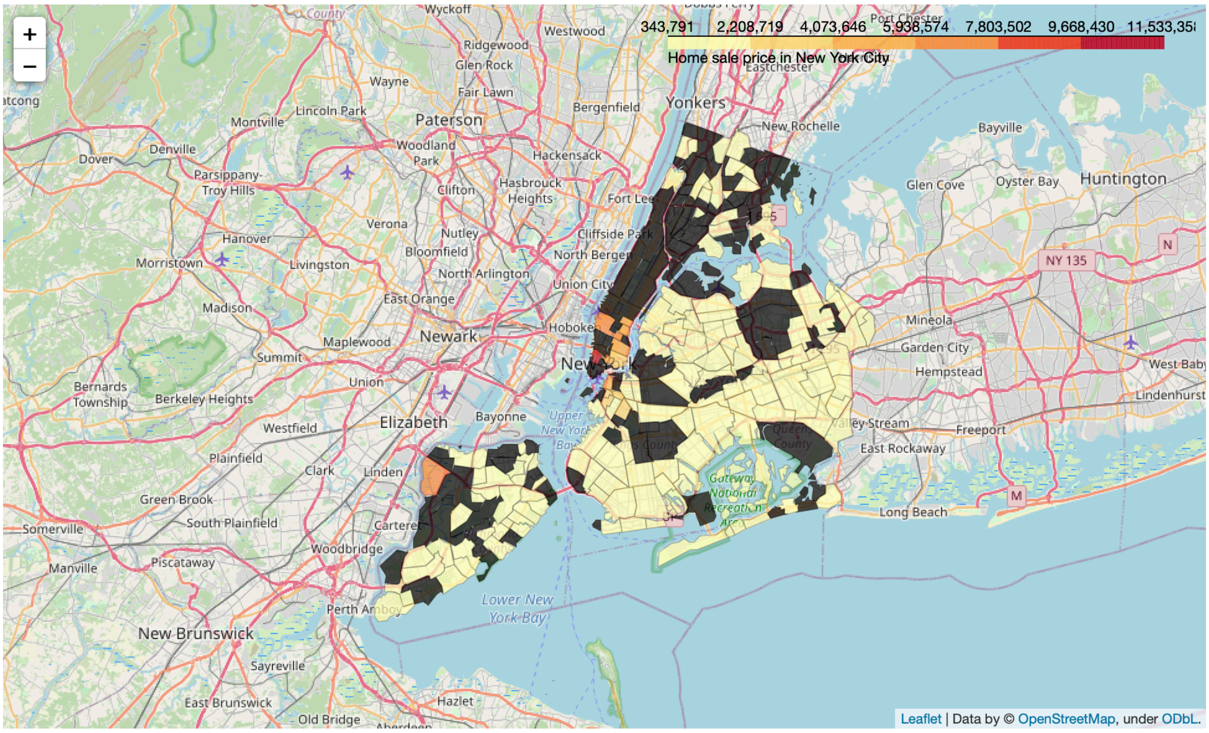


Figure 3. Choropleth map on median home sale price in neighborhoods of New York City.

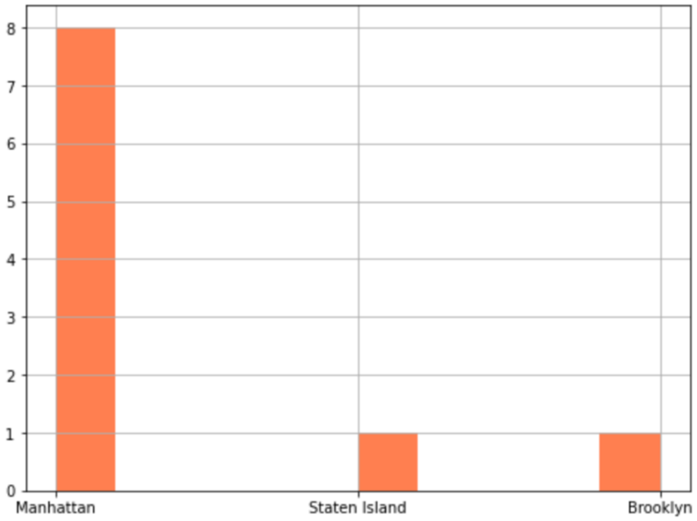
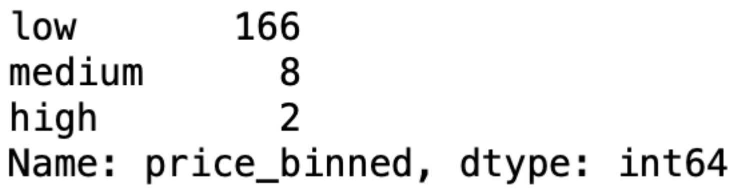


Figure 4. Boroughs of the top 10 highest home sale price neighborhoods.

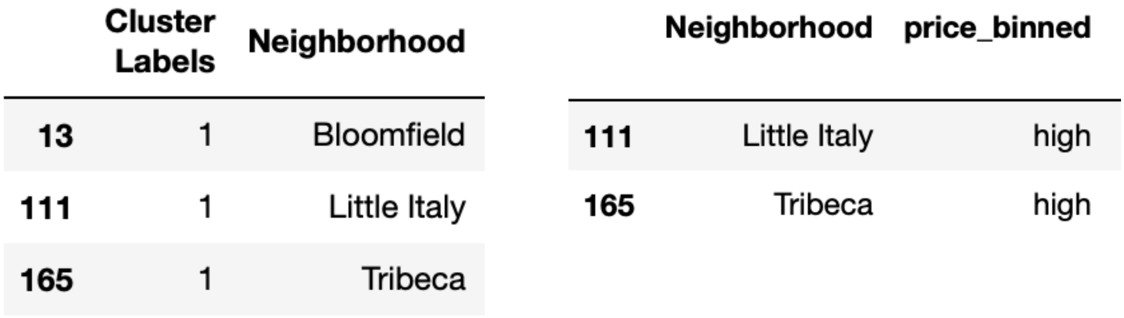
Then I set neighborhoods into three bins on the basis of their median sale price (Table 4). And there are 2 neighborhoods with high home sale prices, 8 neighborhoods with medium home sale prices, and 166 neighborhoods with low home sale prices.



```
low       166
medium      8
high        2
Name: price_binned, dtype: int64
```

Table 4. Bins based on home price. The count refers to how many neighborhoods are set into this bin.

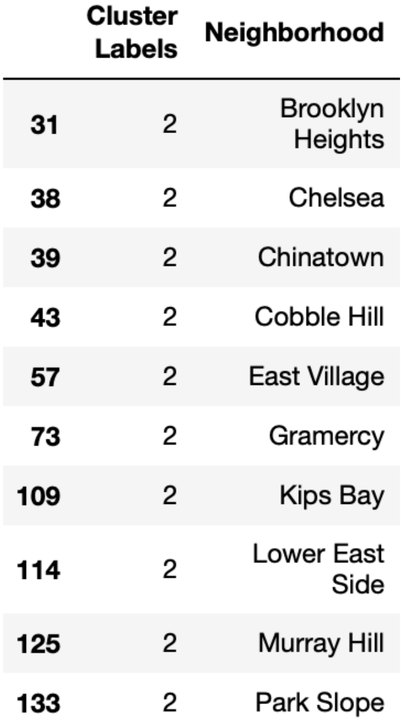
I applied k-means to cluster the neighborhoods into 3 clusters on the basis of their venue category and frequency. According to the result, I found the clusters are very similar to the sale price groups. There are 3 neighborhoods in cluster labelled as 1, and both 2 neighborhoods with high home sale prices are included in this cluster (Table 5). And there are 10 neighborhoods in cluster 2, interestingly, all the 8 neighborhoods with medium home price are included in this cluster (Table 6). And finally, there are 163 neighborhoods in cluster labelled as 0, and all of the neighborhoods in this cluster are also labelled as low home sale price.



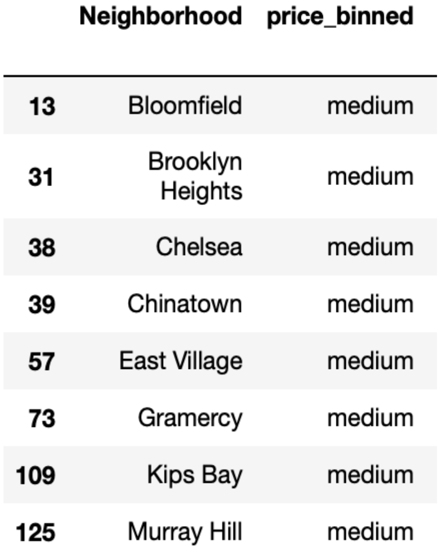
	Cluster Labels	Neighborhood
13	1	Bloomfield
111	1	Little Italy
165	1	Tribeca

	Neighborhood	price_binned
111	Little Italy	high
165	Tribeca	high

Table 5. K-means clustering result based on venue category, showing neighborhoods with cluster label 1 (left); Neighborhoods with high home sale prices (right).



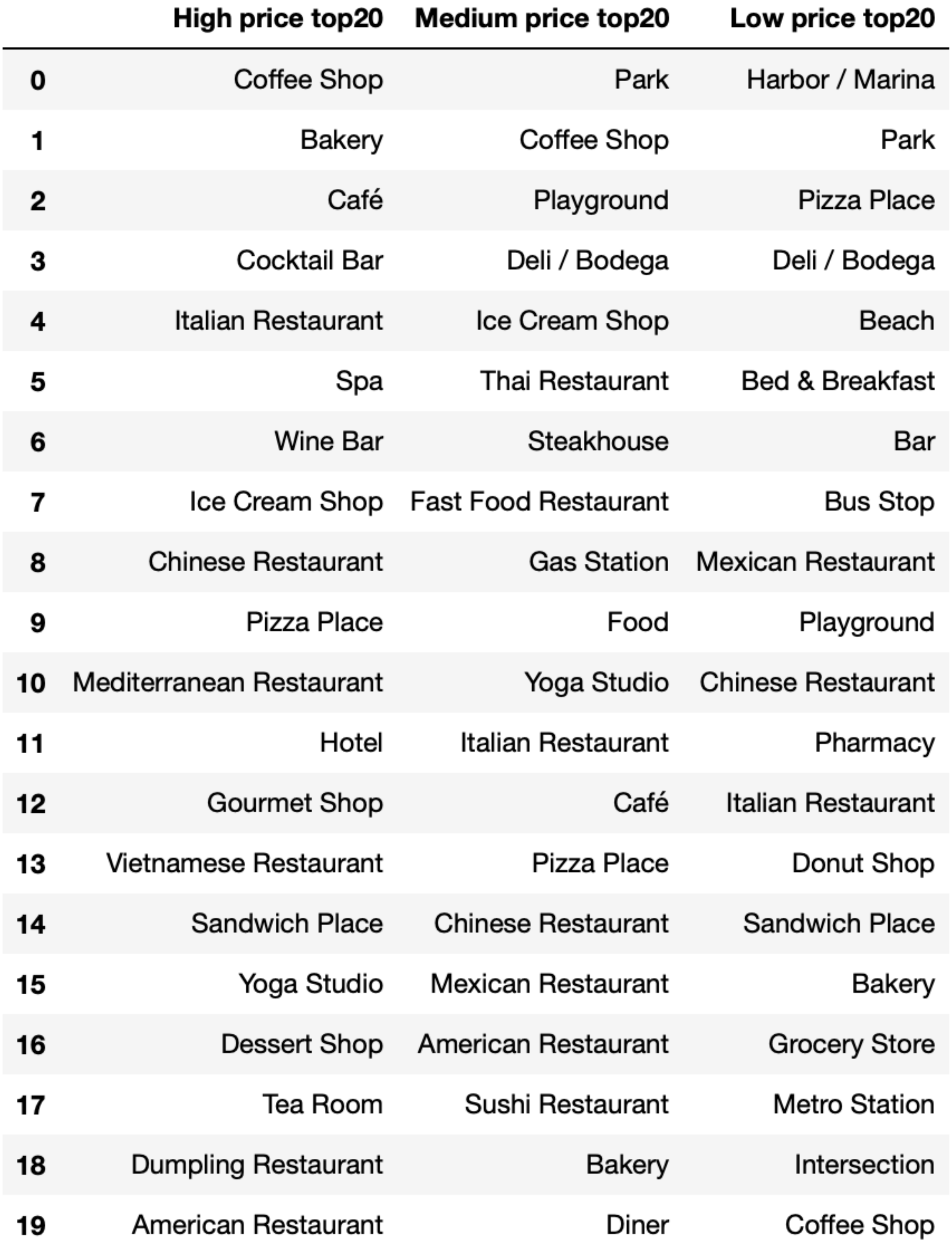
	Cluster Labels	Neighborhood
31	2	Brooklyn Heights
38	2	Chelsea
39	2	Chinatown
43	2	Cobble Hill
57	2	East Village
73	2	Gramercy
109	2	Kips Bay
114	2	Lower East Side
125	2	Murray Hill
133	2	Park Slope



	Neighborhood	price_binned
13	Bloomfield	medium
31	Brooklyn Heights	medium
38	Chelsea	medium
39	Chinatown	medium
57	East Village	medium
73	Gramercy	medium
109	Kips Bay	medium
125	Murray Hill	medium

Table 6. K-means clustering result based on venue category, showing neighborhoods with cluster label 2 (left); Neighborhoods with medium home sale prices (right).

By comparing the top 20 frequent venue categories among high, medium, low home price groups (Table 7), I found high home sale price neighborhoods have more various and larger amount of restaurants, but fewer park or field places than other neighborhoods.



	High price top20	Medium price top20	Low price top20
0	Coffee Shop	Park	Harbor / Marina
1	Bakery	Coffee Shop	Park
2	Café	Playground	Pizza Place
3	Cocktail Bar	Deli / Bodega	Deli / Bodega
4	Italian Restaurant	Ice Cream Shop	Beach
5	Spa	Thai Restaurant	Bed & Breakfast
6	Wine Bar	Steakhouse	Bar
7	Ice Cream Shop	Fast Food Restaurant	Bus Stop
8	Chinese Restaurant	Gas Station	Mexican Restaurant
9	Pizza Place	Food	Playground
10	Mediterranean Restaurant	Yoga Studio	Chinese Restaurant
11	Hotel	Italian Restaurant	Pharmacy
12	Gourmet Shop	Café	Italian Restaurant
13	Vietnamese Restaurant	Pizza Place	Donut Shop
14	Sandwich Place	Chinese Restaurant	Sandwich Place
15	Yoga Studio	Mexican Restaurant	Bakery
16	Dessert Shop	American Restaurant	Grocery Store
17	Tea Room	Sushi Restaurant	Metro Station
18	Dumpling Restaurant	Bakery	Intersection
19	American Restaurant	Diner	Coffee Shop

Table 7. Top 2 frequent venue categories in three home sale price groups.

At the last, I also built classification models, trying to classify home sale price groups (high, medium and low) according to the nearby venue categories. I used KNN and logistic regression to build the model, and test the accuracy of these models using accuracy score. The results for these two models both show very high accuracy (both larger than 0.98).

5. Discussion

The most important finding of my project is that nearby facilities is likely to be correlated with home sale price in neighborhoods in New York City. According to my k-means clustering model for neighborhood venue category, the clusters are very similar to home price groups. In other words, neighborhoods with similar home sale prices tend to have similar venue categories. This pattern is also shown in borough differences analysis: boroughs in the same home price cluster are also fall into the same venue category cluster.

Meanwhile, out classification models also indicated that I can predict neighborhoods' home price groups on the basis of their nearby venues, and the accuracy is pretty high.

Manhattan is the center of New York City, so it is not very surprising that this borough has the highest home sale price. However, the most interesting part is that the venue categories in Manhattan are also different from other boroughs. As shown in Table 3, there are more outdoor recreation places in Manhattan, such as park, boat and ferry, playground, baseball field or even scenic lookout. On the other hand, there are more gyms and fitness centers and less fast food restaurants in Manhattan than other boroughs. Therefore, it is possible that getting in touch with nature and keeping fit are more important to people living in Manhattan. And they are more likely to value mental and physical health.

When I looked into neighborhood differences in nearby venue category among three home price groups (Table 7), I found that the importance of outdoor recreational venues and fitness centers disappears. Instead, there are more various and larger amounts of restaurants near these neighborhoods with high home prices.

One explanation is that, the amount of neighborhoods with high price homes is relatively small (with only 2 neighborhoods), and lack of data might cause underrepresented and bias. Another possible explanation is that Manhattan is a large borough with several neighborhoods, and not all the neighborhoods in Manhattan has high home sale price. The high home sale price in Manhattan might be the result of extremely high price in some neighborhoods, this bias could lead to underrepresented, and studying the venue category and frequency in the whole Manhattan might not be helpful for us to understand the correlation between nearby facilities and home prices.

However, my project has some drawbacks. Firstly, the home price dataset is still lack of data for some neighborhoods, so the result might be not very comprehensive. Secondly, because there are fewer neighborhoods with high home sale price, so the result might be biased. Thirdly, even I draw the conclusion that it is possible that there is a correlation between home price and nearby venues, it is not clear that how nearby venues affect home prices. Finally, in my opinion, the importance of nearby facilities is varied in different cultures, so my conclusion is just suitable for New York City.

6. Conclusion

Overall, I found that there is a correlation between nearby venues and home sale prices in both neighborhoods and boroughs in New York City. And my classification models can also be applied to predict home price groups (high price, medium price and low price) with high accuracy. Therefore, I suggest that real estate developers in New York City can take nearby venues as an important factor when they choose real estate locations or appraise home values.