Cogs 109 Final Project: New York Airbnb Data

Andrew Pesek, Chen Xu, Andrei Sebald, Boyuan Zheng

Background

Dataset:

The dataset contains many samples taken from airbnb's database containing information on host listings in the New York city area. Samples include information on the name of each listing, the host name, neighborhood and neighborhood area, latitude and longitude, space type, and the price of the listing in dollars. Included is also information about reviews, such as last review, total number, and average per month; and also the minimum number of nights, days available throughout the year, and each host's total number of listings.

Link to dataset: https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

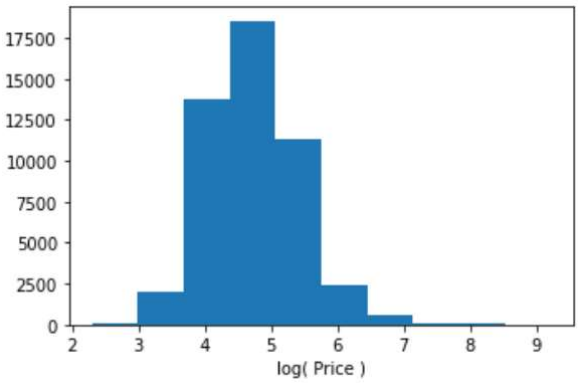
Research Question:

The goal of our analysis is to use primarily linear regression and clustering to understand which variables in the dataset, along with which prediction method, do the best at predicting the listing price, and provide insights into the nature of other associated variables in terms of how they correlate or cluster together.

Methods

Data cleaning and processing:

- Removal of samples that had a price value of 0, this probably shouldn't make sense for an airbnb listing and most likely represents a missing value in the data.
- We are also using log(price) to visualize and train some of our models because the distribution of prices in that dataset appears to be exponential, which makes it difficult to visualize this data against other variables.
    - (for the sake of time we were only able to use this method with the clustering models)

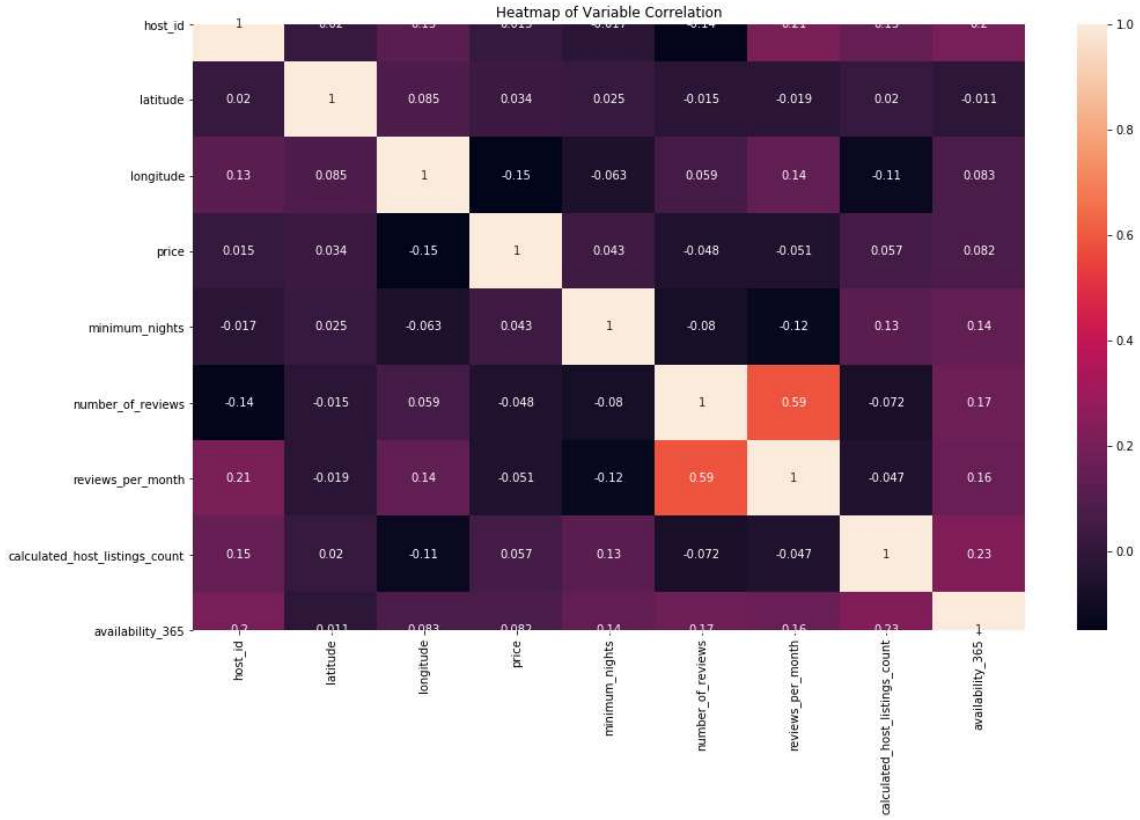


Distribution of prices after applying log function

Basic Analysis:

After data analysis and processing, we conduct some basic data analysis to have an intuition about what the dataset is like and build our model based on the analysis result.

Firstly, we calculate the Pearson Correlation between each variable, so we can have a basic intuition about which variables are more important to price prediction. We draw a heatmap to visualize the correlation.

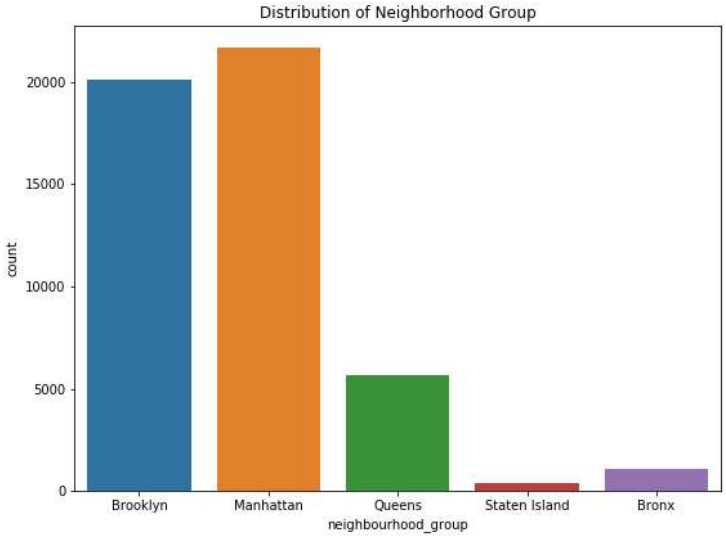


Pearson Correlation between each variable

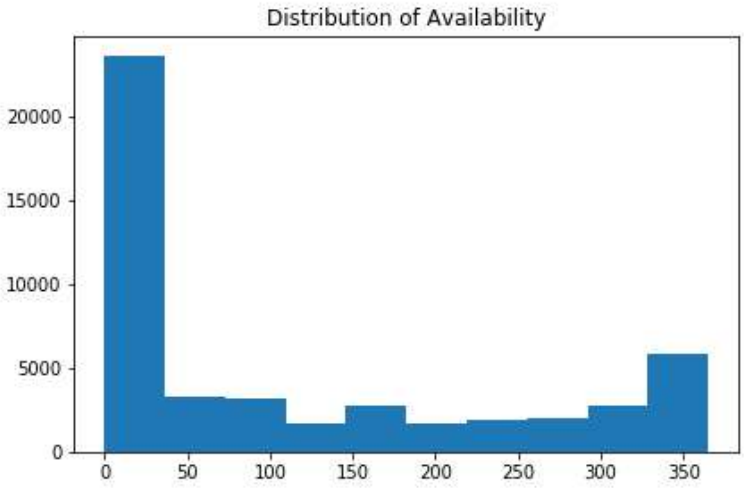
We find most of the variables are weakly correlated. So we need to combine multiple variables to predict the price.

Secondly, we analyze the distribution of each variable and calculate histogram for each of them.

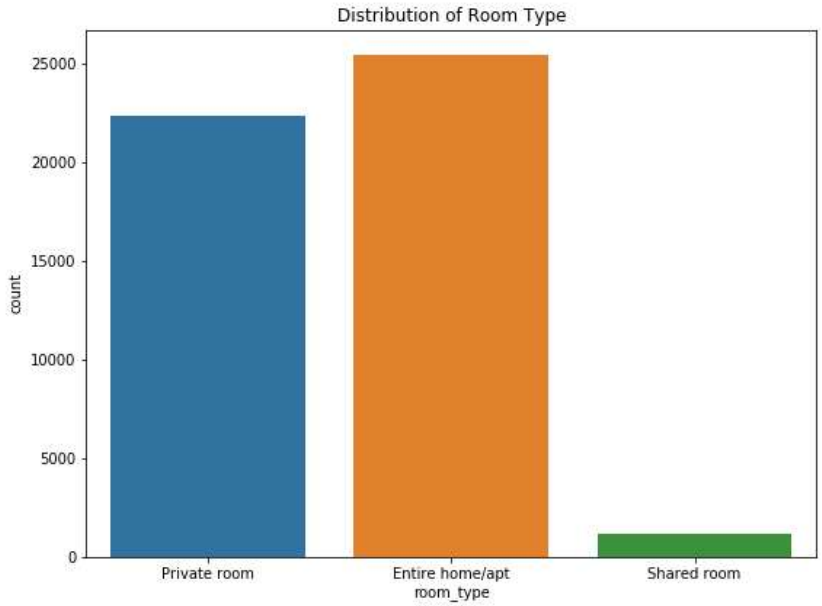
The first graph is about neighborhood group distribution. We can find that the neighborhood group is not evenly distributed.



The second graph is about the distribution of availability days. We can find the available day is also not evenly distributed. A large number of examples is lower than 50.



We also analyze the distribution of room type and draw the histogram.



Analysis:

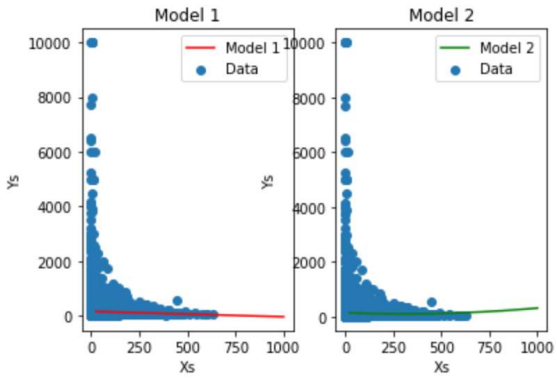
- Linear Regression:
    - We first split the data set into training and test sets, for the training set, we simply chose the first 30000 samples in the data set and used the remaining samples as our test set.
    - Based on the initial look at the data set, we came up with two models to train our training set.
    - Model 1: price = w0 + w1 x number_of_reviews
    - Model 2: price = w0 + w1 x number_of_reviews + w2 x availability_365^2
    - After plotting the two models on our training set, we found that model 2 has lower SSE and therefore fits more data points than model 1.
    - Next we came up with two other models that made more sense on paper, model 3 put more weight on the number of reviews. However, we found out that it actually has higher SSE than the first two models. Model 4 put more emphasis on host listing counts, but this model has much higher SSE than the other three models.
    - Model 3: price = w0 + w1 x availability_365 + w2 x number_of_reviews^2
    - Model 4: price = w0 + w1 x number_of_reviews + w2 x calculated_host_listings_count^2
    - Since model 2 has the lowest SSE, we decided to use its weight vector to run cross validation on our test dataset.

    - After the cross validation and the plots were done, we found out that the MSE for the test set is actually lower than the training set. This means our model 2 fits better on our test set.
- Clustering:
    - We chose a few different methods of clustering the data, and for each method we generated some statistics describing the numerical features of each sample within a cluster.
    - We stuck with doing 5 clusters each time the k-means algorithms was used because there are 5 unique neighborhood group values in the dataset, corresponding to well known areas of New York City (Manhattan, Brooklyn, Queens, Staten Island, and Bronx) and were curious to see if there would be any alignment between these neighborhoods and the clusters identified.
    - Firstly we clustered the data solely based on the coordinate data of each listing (2 features: latitude and longitude)
    - Then we generated a set statistics on the clusters identified here, and another set of statistics using the neighborhood group feature as labels.
    - Finally there is a clustering generated by 5 other numerical features. (price, minimum nights, number of reviews, calculated host listings, and availability) Here we have displayed these clusters onto the map and generated a similar set of statistics.

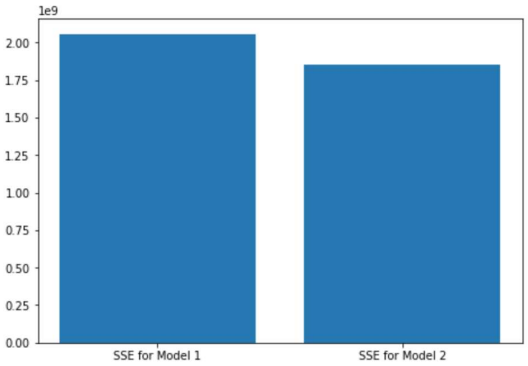
Results

Linear Regression:

Model 1: price = w0 + w1 x number_of_reviews
Model 2: price = w0 + w1 x number_of_reviews + w2 x availability_365^2
Plot comparison of model 1 and model 2:

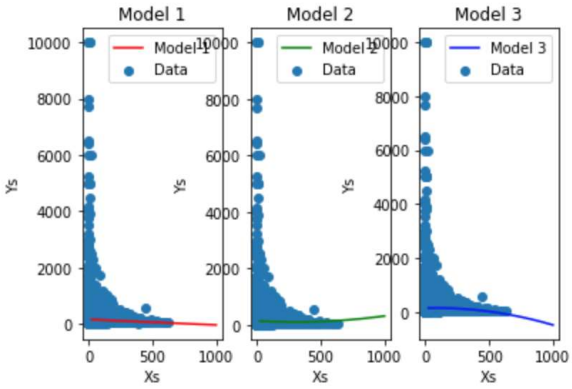


Plot comparison of the SSEs of model 1 and model 2:

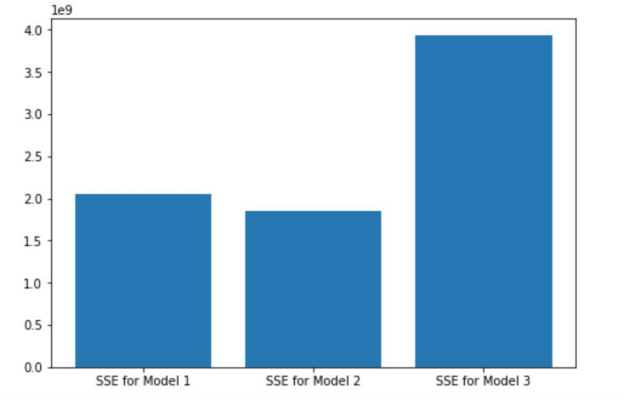


Model 3: price = w0 + w1 x availability_365 + w2 x number_of_reviews^2

Plot comparison of model 1, model 2, and model 3:

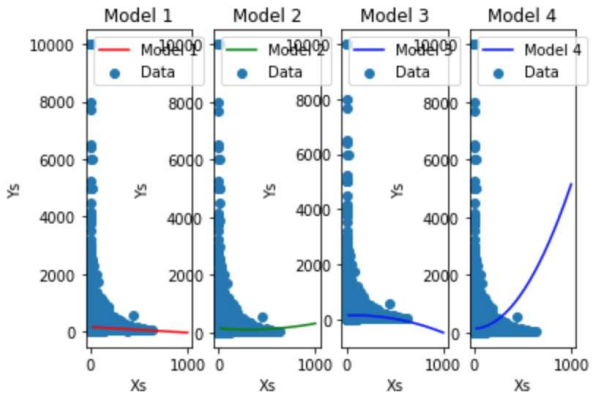


Plot comparison of the SSEs of model 1, 2 and 3:

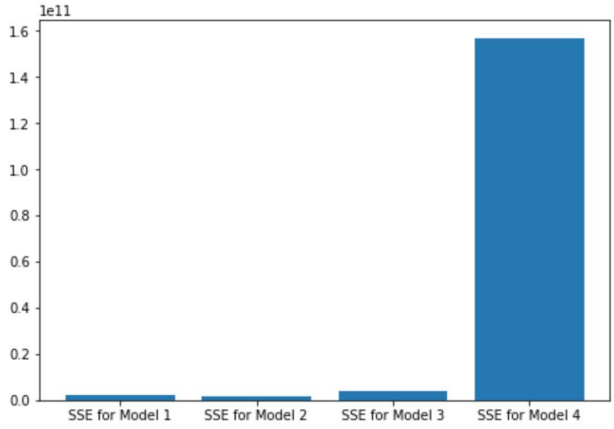


Model 4: price = w0 + w1 x number_of_reviews + w2 x calculated_host_listings_count^2

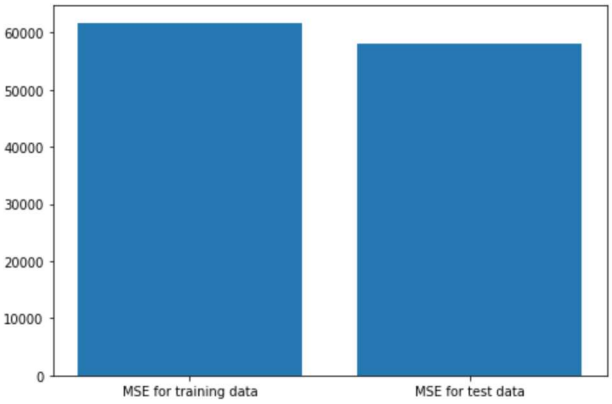
Plot comparison of all 4 models:



Plot comparison of the SSEs of all 4 models:

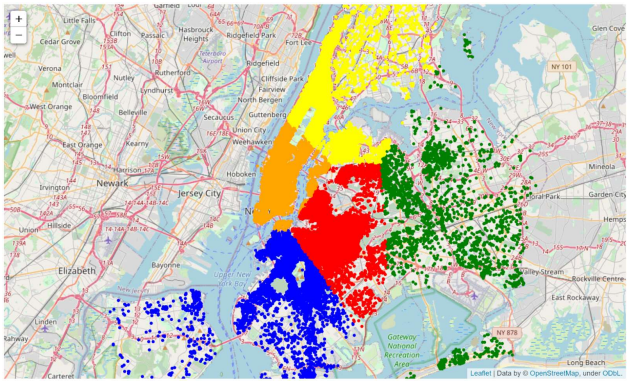


Plot comparison of the MSEs of the training set and test set:



Clustering:

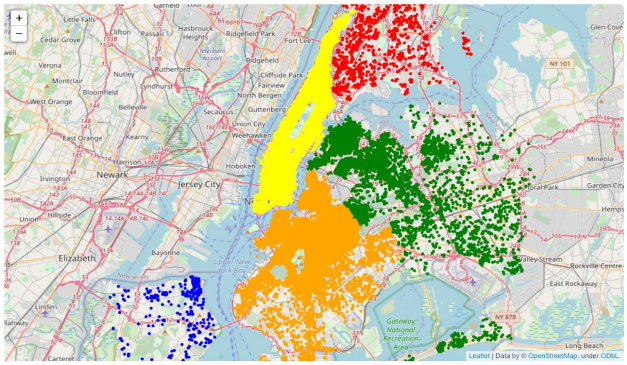
Method 1 - clustering by latitude and longitude



K-means clustering by location, 5 clusters

	avg price	avg minimum nights	avg number of reviews	avg host listings	avg availability
0: red	114.75, (std=169.7)	6.02, (std=16.5)	24.38, (std=44.8)	2.86, (std=8.4)	103.75, (std=127.9)
1: orange	227.92, (std=327.6)	9.28, (std=26.4)	19.56, (std=41.1)	17.05, (std=57.1)	116.13, (std=134.9)
2: yellow	126.53, (std=210.2)	6.56, (std=18.5)	24.26, (std=46.2)	3.29, (std=12.4)	110.17, (std=129.8)
3: green	98.66, (std=117.4)	3.69, (std=11.1)	32.15, (std=56.0)	3.56, (std=8.1)	172.38, (std=133.7)
4: blue	130.81, (std=171.3)	6.21, (std=18.7)	24.19, (std=42.8)	1.85, (std=2.6)	106.70, (std=128.2)

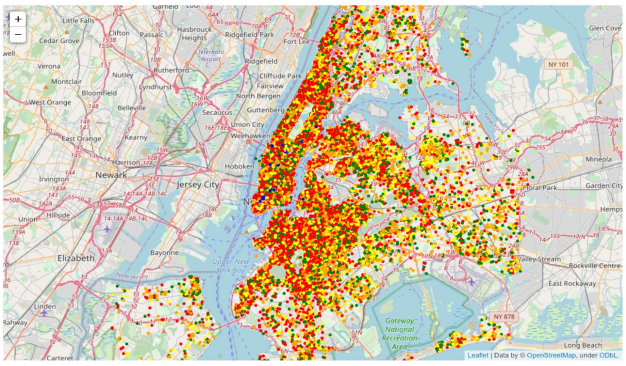
Method 2 - samples labeled by neighborhood group



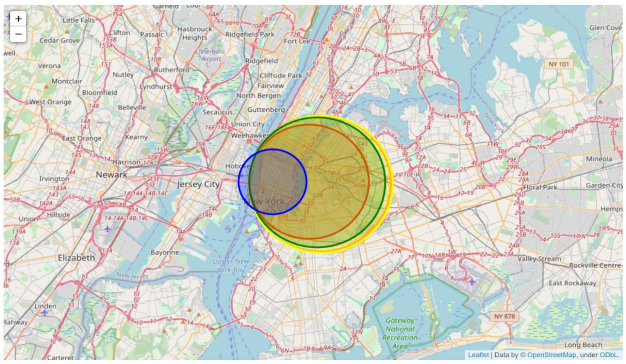
Displaying neighborhood group
(Red: Bronx, Orange: Brooklyn, Yellow: Manhattan, Green: Queens, Blue: Staten Island)

	avg price	avg minimum nights	avg number of reviews	avg host listings	avg availability
0: red	87.58, (std=106.7)	4.56, (std=15.6)	25.98, (std=42.2)	2.23, (std=2.4)	165.79, (std=135.2)
1: orange	124.44, (std=186.9)	6.06, (std=17.6)	24.20, (std=44.3)	2.28, (std=5.3)	100.22, (std=126.3)
2: yellow	196.88, (std=291.4)	8.58, (std=24.1)	20.99, (std=42.6)	12.79, (std=48.2)	111.98, (std=132.7)
3: green	99.52, (std=167.1)	5.18, (std=15.0)	27.70, (std=52.0)	4.06, (std=12.4)	144.45, (std=135.5)
4: blue	114.81, (std=277.2)	4.83, (std=19.7)	30.94, (std=44.8)	2.32, (std=1.9)	199.68, (std=131.7)

Method 3 - clustering by price, minimum nights, number of reviews, host listings and availability



K-means clustering by 5 numerical features, 5 clusters



Clusters centers and relative dispersion of locations

	avg price	avg minimum nights	avg number of reviews	avg host listings	avg availability
0: red	136.73, (std=190.3)	4.83, (std=13.2)	12.28, (std=26.8)	1.66, (std=6.5)	5.20, (std=10.9)
1: orange	178.29, (std=311.3)	11.84, (std=33.4)	28.70, (std=49.9)	9.73, (std=21.3)	330.53, (std=29.8)
2: yellow	146.36, (std=251.1)	6.10, (std=16.8)	34.98, (std=51.2)	2.54, (std=6.3)	87.18, (std=26.9)
3: green	170.97, (std=271.2)	7.95, (std=20.4)	43.86, (std=64.6)	5.50, (std=16.3)	202.18, (std=36.1)
4: blue	273.57, (std=101.7)	20.96, (std=16.0)	2.39, (std=3.8)	288.79, (std=46.6)	287.76, (std=72.7)

Discussion

Linear Regression overall is somewhat inconclusive. While model 2 appears to be the best in comparison to the error metrics of the other models, graphically it appears to be difficult to entirely represent the shape of the data through this method. It does tell us however that the number of reviews and availability features are likely to be more closely related to the listing price. Although, the significance appears to be somewhat vague, and linear regression models don't appear to be the best descriptors of this particular dataset.

From the clustering analysis it can be seen that grouping the locations together using the k-means algorithm based solely on location doesn't appear to point out particularly notable clusters of locations. The 5 clusters generated here appear to somewhat line up with a similar map shown below it that distinguishes each location by its associated neighborhood group in the data, however it would make sense that with how densely populated most of New York city is, the density of airbnb listings don't vary too much throughout most of the metropolitan area. The statistics generated for both of these labeling methods however do point out some areas of interest. The samples corresponding to the orange cluster of the first map have a much higher average price and host listings than the other clusters, slightly higher minimum nights, and slightly lower number of reviews. Interestingly the green area of this first map contains samples with a lower average price, minimum nights, and higher availability. A similar result shows up in the second map, where the samples corresponding to the yellow region have a much higher average price and host listings. The trends pointed out previously in the green region are not quite as pronounced within the green region of the second map, corresponding to the Queens neighborhood region.

In the third map shown, samples are clustered by the 5 features described in the statistics in each table, not by location. When shown on the map each cluster appears to just be evenly dispersed for the most part, telling that these clusters are not particularly notable when compared against each other, however there is one cluster corresponding to the blue locations that is much more separated from the other clusters. The table below this map shows that the blue labeled points have a significantly higher average price, minimum nights, host listings, and much fewer number of reviews than the other clusters. The next map displays the centers and rough size of each of these clusters on the map, and the smaller blue region exists roughly in the same place as the similar orange and yellow regions in the first and second maps respectively. Each of these areas noted by each clustering method are located in the lower Manhattan area of New York.

Link to Poster:
https://drive.google.com/file/d/1GkHPAxKh1WOqVkdiRiSP3uZL1sPEsLtx/view?usp=sharing
Links to Python Code:
https://drive.google.com/file/d/1zvtHsEJ7sbsy4ySX3iI5aEZnSN0kEyiO/view?usp=sharing
https://drive.google.com/file/d/1wwdoBD_EBE1Eu3ZZygKBLKuIvP5BDf23/view?usp=sharing