Statistical Exploration of Airbnb Listings in NewYork City in 2019 using Exploratory Data Analysis and Regression.

Hyunjae Kwon,[1] Luke Yee,[2] and Aly Kapasi[3]

[1] *Department of Statistics, Student ID: 915084195, Email: kevkwon@ucdavis.edu*

[2] *Department of Statistics, Student ID: 914913445, Email: leyee@ucdavis.edu*

[3] *Department of Statistics, Student ID: 915223613, Email: amkapasi@ucdavis.edu*

Abstract:

Airbnb is an online marketplace that offers alternative forms of accommodation to consumers beyond the conventional choices, such as hotels, hostels, and motels. In order to improve user experience, the company allows the public to access data through their API. In the current research paper, New York City Airbnb dataset (N = 48,895) is analyzed using an array of data analysis techniques to find interesting or useful patterns that could potentially assist with improving customer experience. The exploratory analysis section of the paper looks at the number of listings and price at the neighborhood group and room type level. Hierarchical clustering is also employed to find the patterns of similarities and differences among neighborhood groups. Predictive models are also built to further analyze the relationship between the variables in the dataset. Lastly, confounding trends between variables and group types are explored.

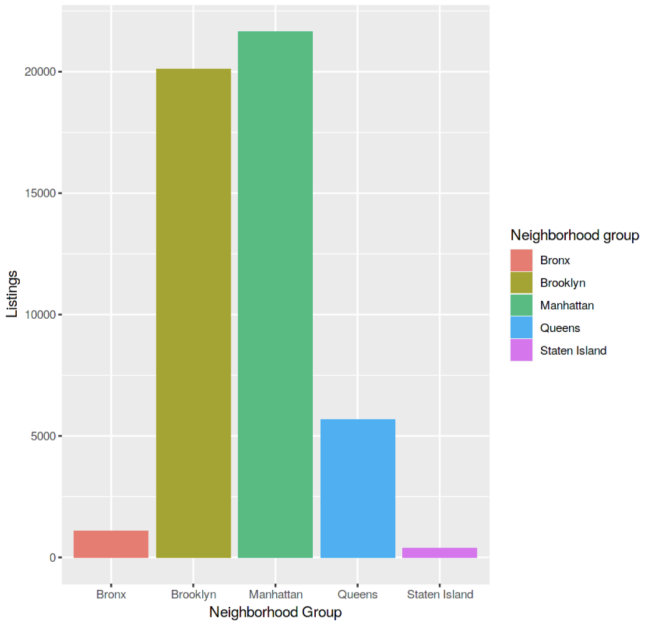
I. Introduction

A. Dataset

The original dataset consists of 48,895 listings and 16 variables. Each listing has *id, name, host id, host name, neighborhood group, neighborhood, latitude, longitude, room type, price, minimum nights, number of reviews, last review date, reviews per month, host listings count, and the number of available days*. *Neighborhood group* represents the boroughs in New York City and has 5 levels: Brooklyn, Manhattan, Queens, Staten Island, and Bronx. *Neighborhoods* represents the neighborhoods in New York City and consists of 211 levels. *Room type* has 3 levels: entire home/apartment, private room, and shared room. *Host listings* count is the number of listings a host of the given listing has.

B. Exploratory Data Analysis

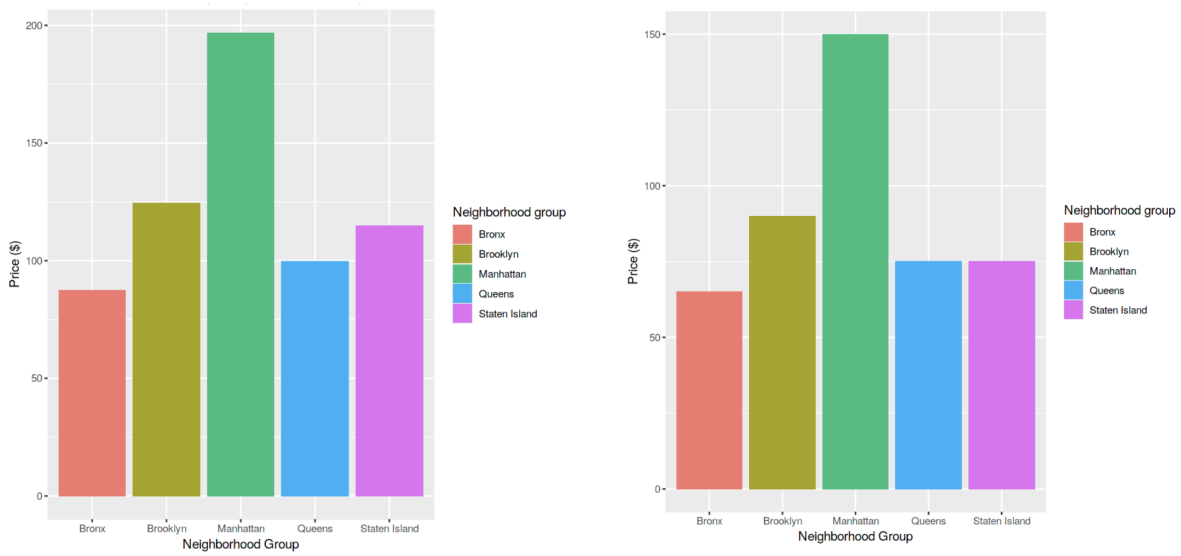
FIG. 1. Number of Listings by Neighborhood Group



One of the measures that can be used to understand the popularity of neighborhood groups is the *number of listings*. As seen in FIG. 1., Manhattan and Brooklyn have significantly higher numbers

of listings than the other neighborhoods at 21,661 and 20,104, respectively (TABLE 1.). Another interesting insight can be obtained from looking at the average price across neighborhood groups. The left bar plot in FIG. 2. shows that listings in Manhattan have the highest mean price of $197.50.

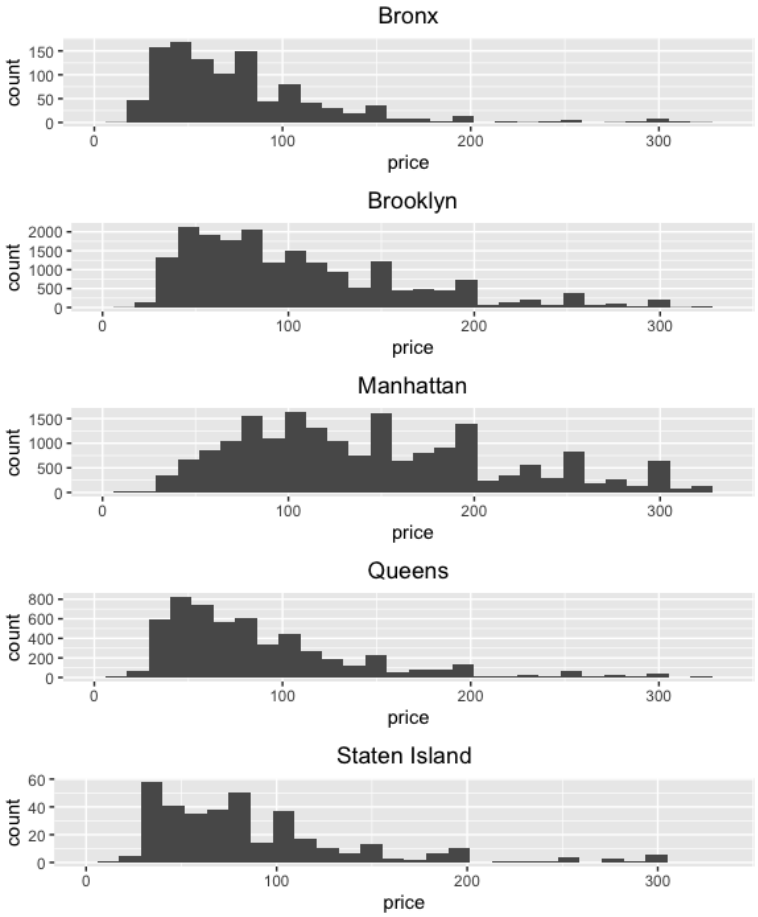
FIG. 2. Mean and Median Price by Neighborhood Group



Rest of the neighborhood groups have similar mean prices: Bronx at $87.50, Brooklyn at $124.00, Queens at $99.50, and Staten Island at $115.00 (TABLE 1.). The median prices are also observed because it is less sensitive to outliers—2972 listings, or 6% of the dataset are outliers when the IQR method is employed. With this information, as expected, the shape of the bar plots of mean prices and median prices on FIG. 2. do not show much difference. Manhattan still has the highest average price, and Bronx has the lowest average price. The bar plots in FIG. 2. provide insight into the general price difference across neighborhood groups; however, it does not provide information about the distribution, spread, minimum, maximum prices of the listings.

To find more about the prices of listings in each neighborhood, we look at the histograms of each neighborhood group. As expected, FIG. 3. shows that the distribution of prices is skewed to the right in all neighborhood groups, albeit the distribution of listing prices is noticeably less skewed to the right compared to others. Colloquially speaking, there are more cheaper listings than expensive listings.

FIG. 3. Distribution of Prices by Neighborhood Group with Removal of Outliers



Outliers—which was calculated using the IQR method—were removed for the interpretability of the histograms. From the histograms, we learn valuable information not accessible in the bar plots. Although Manhattan's average listing price is significantly higher than those of other neighborhood groups, there are still many affordable listings. Users can still find more affordable listings in one of the most popular neighborhood groups in Airbnb. Also, the histogram of Manhattan's price listings shows

that there are more expensive listings compared to other neighborhood groups. This fact can partially explain the high average listing prices in Manhattan. However, there can be other factors that contribute to this pattern. One such factor is the distribution of room types in a neighborhood group.

FIG. 4. Geospatial Representation of Room Types

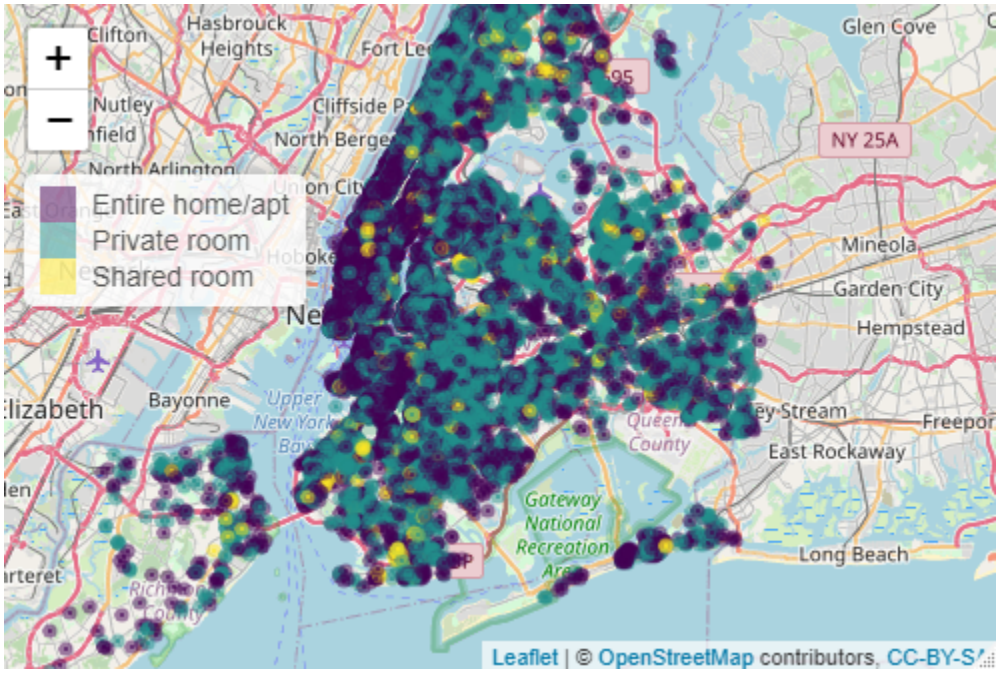


FIG. 5.. Room Type by Neighborhood Group

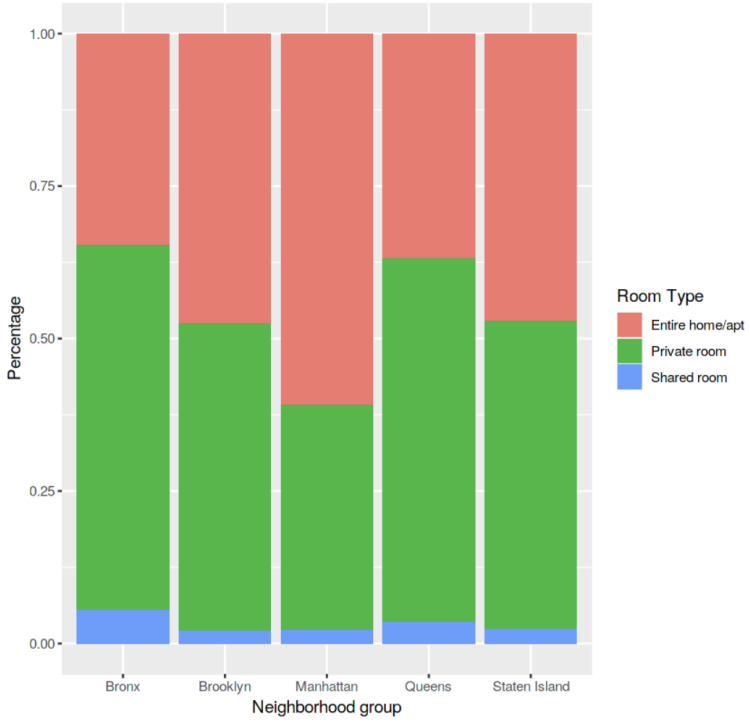


FIG. 5. shows that Manhattan has the highest proportion of entire home or apartment units of 60.9% (TABLE. 2.). The proportions of entire home or apartment units are 34.7% in Bronx, 47.5% in Brooklyn, 37.0% in Queens, and 47.2% in Staten Island (TABLE. 2.). Note that the entire home or apartment category has the highest mean price of $212.00, followed by $89.80 for private room and $70.1 for shared room (TABLE. 1.). Thus, the proportion of room type provides further explanation for Manhattan's high average listing price. Other numerical variables are also compared by neighborhood groups.

Heatmap is especially useful for visualizing the differences between neighborhood groups across all numerical variables. FIG. 7. shows the heatmap of the mean values of all numerical variables by neighborhood groups. Latitude, longitude, minimum nights, and calculated host listings count

FIG. 6. Mean Price by Room Type

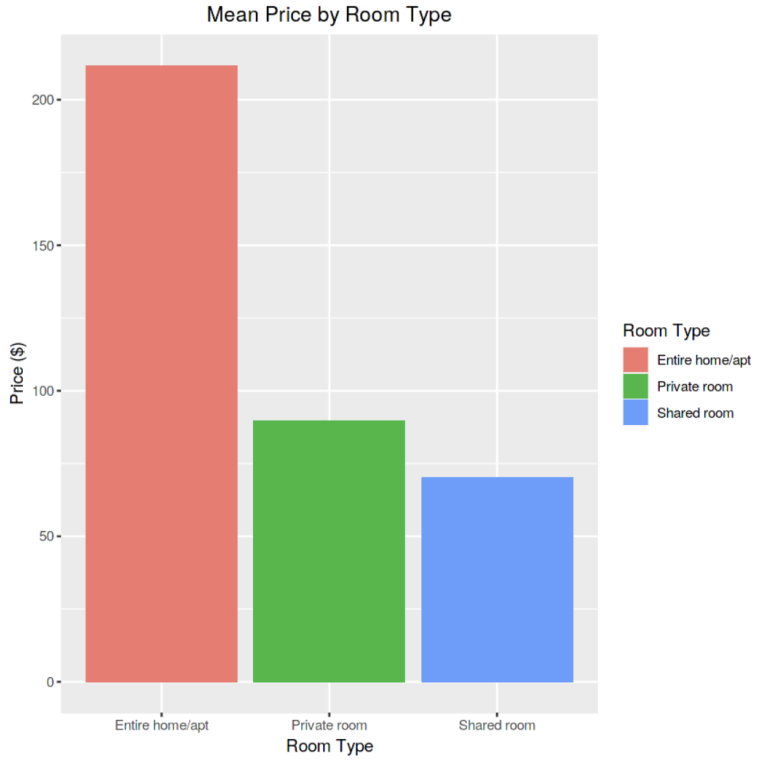
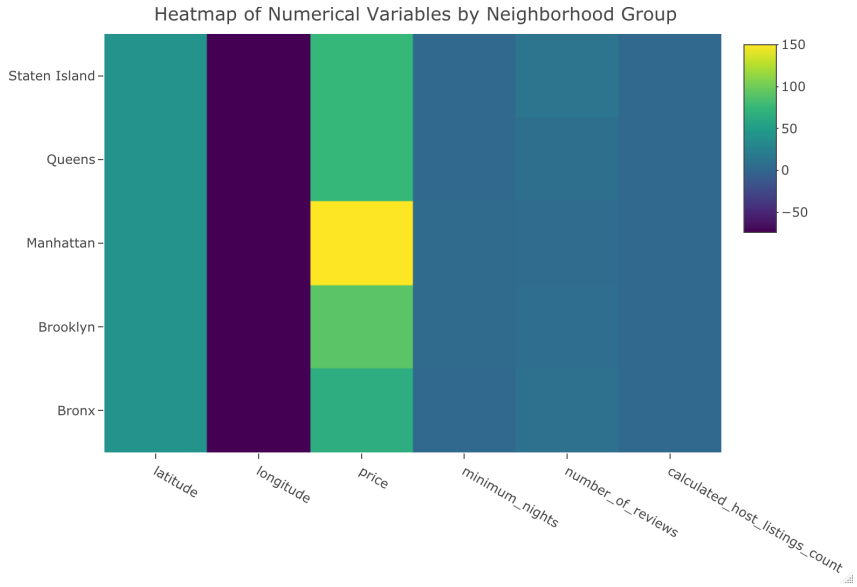


FIG. 7. Heatmap of Numerical Variables by Neighborhood Group



have no noticeable differences across neighborhood groups. The differences are slightly more noticeable for the number of reviews. Based on the color scale, Staten Island seems to have the highest mean number of reviews.

C. Clustering and Modeling

Neighborhoods groups can be further compared against one another through hierarchical clustering. Two dendrograms are created using 1) agglomerative clustering and complete linkage clustering and 2) agglomerative clustering and single linkage clustering. In agglomerative clustering, each observation is treated as a single cluster, and at each step of the algorithm, cluster merges with the closest cluster to form a bigger cluster, and this process is repeated until the root node is reached (Hierarchical cluster analysis). The similarity or the distance between each cluster can be measured in many ways: two of which are complete linkage and single linkage clustering. Complete linkage clustering “computes all pairwise dissimilarities between the elements in cluster 1 and the elements in

cluster 2, and considers the largest value… of these dissimilarities as the distance between the two clusters" (Hierarchical cluster analysis). Single linkage clustering "computes all pairwise dissimilarities between the elements in cluster 1 and the elements in cluster 2, and considers the smallest of these dissimilarities as a linkage criterion" (Hierarchical cluster analysis). FIG. 8. shows the dendrogram of neighborhood groups using complete linkage clustering, and FIG. 9. shows the dendrogram of neighborhood groups using single linkage clustering.

FIG. 8. Dendrogram of Neighborhood Groups – Complete Linkage

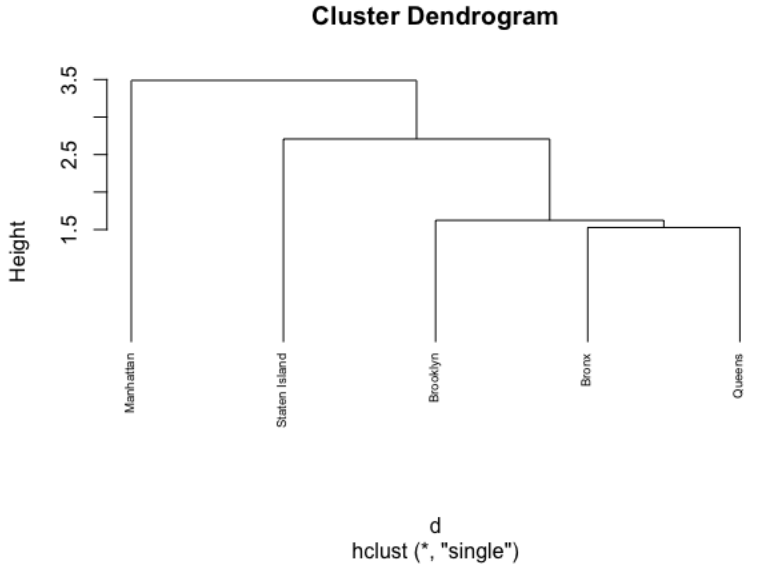
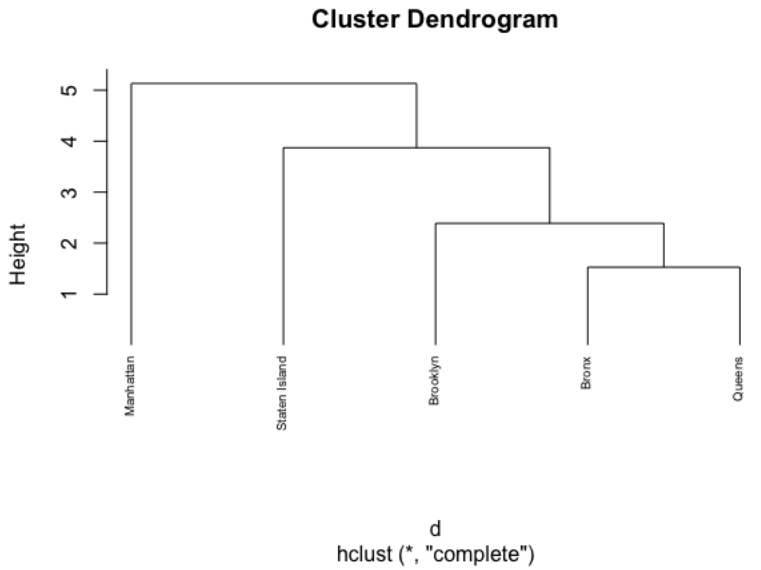


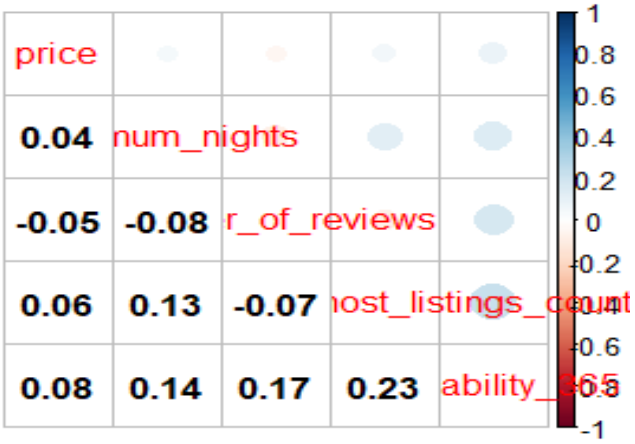
FIG. 9. Dendrogram of Neighborhood Groups – Single Linkage



D. Predictive Models

To determine whether the data was usable for prediction, we applied a linear regression model. The linear model was selected by stepwise regression using the BIC criteria. The models were regressed with the multiple response variables. The linear regression models resulted without any significant findings. The linear model with price as the response variable had a r^2 value of 0.09271, which is a clear indication that the model was an extremely ineffective fit. The other models had even worse fits, except for the model with the number of available days per year, that model was only slightly better with a r^2 value of 0.1341. When running diagnostics to verify if the assumptions for the multiple linear regression were met, we noticed that all the variables seemed to be extremely loosely correlated to each other, if at all; this is a major problem for regression analysis, because if there is no relationship between the variables, if is very difficult to predict.

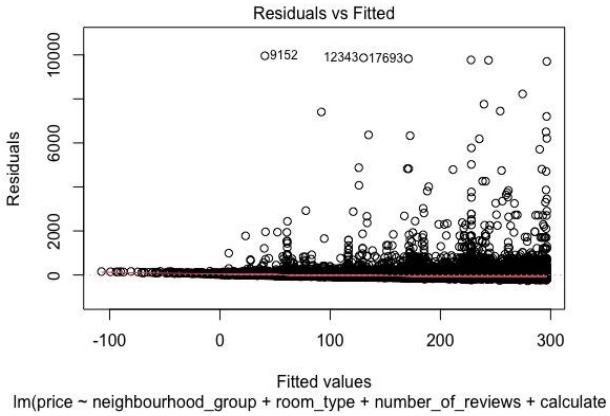
FIG. 10. Correlation Matrix

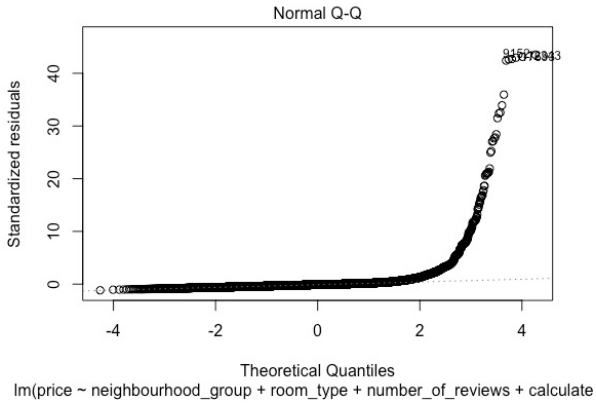


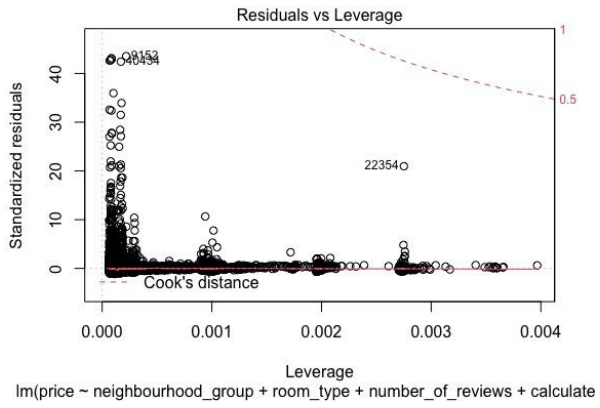
Using stepwise regression on the price linear regression model, we found the best model contained boroughs, room types, number of reviews, number of listings by the host and number of days available in the year. The stepwise regression also determined in both cases that the best model was the full

model, though that could also be because all the models were extremely ineffective, so the full model was the best choice.

FIG. 11. Residual Plot and QQ Plot








```
Call:
lm(formula = price ~ neighbourhood_group + room_type + number_of_reviews +
    calculated_host_listings_count + availability_365, data = airbnb)

Residuals:
   Min     1Q Median     3Q    Max
-268.3  -64.0  -22.7   16.3 9959.2

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1.394e+02  7.194e+00  19.378  < 2e-16 ***
neighbourhood_groupBrooklyn     3.319e+01  7.138e+00   4.649 3.34e-06 ***
neighbourhood_groupManhattan    8.877e+01  7.144e+00  12.425  < 2e-16 ***
neighbourhood_groupQueens       1.357e+01  7.568e+00   1.793 0.072901 .
neighbourhood_groupStaten Island 7.613e+00  1.373e+01   0.555 0.579192
room_typePrivate room          -1.115e+02  2.140e+00 -52.085  < 2e-16 ***
room_typeShared room           -1.448e+02  6.900e+00 -20.994  < 2e-16 ***
number_of_reviews              -3.146e-01  2.376e-02 -13.240  < 2e-16 ***
calculated_host_listings_count -1.112e-01  3.300e-02  -3.369 0.000755 ***
availability_365                1.876e-01  8.318e-03  22.549  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 228.8 on 48885 degrees of freedom
Multiple R-squared:  0.09271,   Adjusted R-squared:  0.09254
F-statistic:   555 on 9 and 48885 DF,  p-value: < 2.2e-16
```

From the diagnostic plots, we can determine that linear regression may not be the best model for this data. This can be determined by the graph of residuals vs fitted points, where there are many outliers. The Q-Q plot also shows the theoretical quantiles straying greatly from the expected range near the upper theoretical quantiles. The scale location graph has a diagonal line, which indicates that the model lacks homoscedasticity. According to the diagnostics, this data is insufficient to run any meaningful linear regression, though it may be possible to use a robust linear regression model instead.

E. Confounding Trends Between Groups

When analyzing the number of listings by neighbourhood groups, a confounding trend was spotted. The Manhattan group had more listings than the Brooklyn group, but a lower number of reviews. To test the significance of this observation, the two smaller datasets were formed, each composed purely of Manhattan or Brooklyn observations. A t-test was run comparing the number of reviews from each group, and returned a p-value of 4.3E-14, indicating that there was a significant difference in the number of reviews. To look further into the matter, the mean number of minimum nights per listing in each group was taken, with Manhattan requiring an average of 8.6 nights while Brooklyn required an average minimum of 6.0 nights. What does this observation reveal? It means that Manhattan listings require tenants to stay longer, which means that each listing place cannot be rented out as frequently. On the other hand, Brooklyn listings have a lower minimum nights requirement, which allows each listing to be rented out more frequently within a time span. This helps explain why Manhattan may have more listings, but less reviews than Brooklyn because the number of tenants and reviews a listing can receive in a given time span is higher in Brooklyn.

Another interesting feature we decided to investigate while on this topic was the calculated number of host listings between the Manhattan and Brooklyn group. Simply taking the mean of this column led us to believe that the average number of listings per host in Manhattan was 12.7, while Brooklyn had an average of 2.2 listings per host. However, we noticed that just taking the mean was over representing hosts who had a large number of listings. For example, an observation that corresponded to a host with 10 listings would add a value of 10 when summing up the total, but this also meant that there were 9 other observations that contributed a value of 10 to the total. After only

selecting rows in the groups with unique host ID, we took the average number of listings per host again. This time, Manhattan had an average of 1.32 listings per host, while Brooklyn had an average of 1.29 listings per host. Taking the mean almost led to a false conclusion that hosts in manhattan listed their properties for rental far more often than hosts in Brooklyn, but the corrected mean reveals that there is very little difference between the two neighborhood groups in this category.

II. Conclusion

From our exploration of the 2019 New York City Airbnb dataset, we were able to see the distributions and relationships of the variables. We discovered that the majority of the listings were either in Manhattan or Brooklyn, we also found that most of the listings were full units, followed by private rooms with shared rooms being the most infrequent. During our data exploration, we discovered a problem that there were many outliers that greatly impacted the averages of the data, for example most hosts were only able to let out properties a few times a year, but due to the presence of outliers such as hosts that would let out properties 200 or more times a year, so the data was very unclean and needed some doctoring. Our hierarchical clustering dendrogram models, we can see the closeness of the borough clusters in terms of overall similarity.

During the modeling and fitting stage, we noticed that the variables were very poor predictors of price. This was confirmed when we ran a correlation plot and saw that there was very little correlation between variables. We also noticed that some of the assumptions for linear regression were not being satisfied, especially due to the high amount of outliers that caused the data to be skewed. Our price model had a r2 value of 0.0927, while our most accurate model was the availability_365 variable

which had an r2 value of 0.1341, although this model was more accurate than the price model, it still was not accurate enough to be considered a useful model.

Our exploratory analysis also revealed some confounding trends in the data, with certain groups appearing to be over represented. By dividing the observations into neighbourhood groups, we were able to compare certain variables such as minimum nights that could help explain the over representation of Brooklyn's number of reviews. Finally, we also were able to find the actual mean listings per host of Brooklyn and Manhattan groups, a difference that initially appeared to be great but was actually exaggerated due to duplicate counting when summing up totals for the mean.

III. Appendix

TABLE I. Descriptive Statistics by Neighborhood Group

	Number of Listings	Mean Price	Median Price
Bronx	1,091	87.50	65.00
Brooklyn	20,104	124.00	90.00
Manhattan	21,661	197.00	150.00
Queens	5,666	99.50	75.00
Staten Island	373	115.00	75.00

TABLE 2. Room Types by Neighborhood Group

	Entire Home/Apt	Private Room	Shared Room
Bronx	379 (34.73%)	652 (59.76%)	60 (5.50%)
Brooklyn	9,559 (47.55%)	10,132 (50.40%)	413 (2.05%)
Manhattan	13,199 (60.93%)	7,982 (36.85%)	480 (2.21%)
Queens	2,096 (36.99%)	3,372 (59.51%)	198 (3.49%)

Staten Island	176 (47.18%)	188 (50.40%)	9 (2.41%)

IV. Supplementary Material

The dataset used in the current research paper is accessible on the 'Airbnb_NYC_2019.csv' file, and the programming is done via R which is available on the "STA160_Midterm_Project.R."

V. References

Hierarchical cluster analysis. (n.d.). Retrieved May 03, 2021, from https://uc-r.github.io/hc_clustering.