MSCI 718 Project

Topic: Pricing Mechanism of Airbnb in New York City

Chengcheng (Eric) Jiang

Wenying (Mandy) Chen

Date of Submission: 2020-04-22

Background Information and Dataset

The short-term rental industry is booming, especially in big cities and popular tourist attractions. Airbnb is a leading company in this industry, operating in more than 81,000 cities and 191 countries of the world. Also, it imposes a great threat to the hotel industry, since many Airbnb properties offer cheaper price and more of a home feeling. The evolving phenomenon attracts our attention: to be competitive in the market, how do Airbnb homeowners decide their prices? Is there something traceable in the pricing mechanism?

For this project, we narrow down the topic to Airbnb's pricing mechanism in New York City (NYC). We obtain the dataset "NYC Airbnb Open Data" from Kaggle (Dgomonov, 2019), which describes the listing activity and metrics in NYC of the year 2019.

Data Description and Data Cleaning

There are 48895 observations and 16 variables in the dataset. Aside from the host labels (*host_id*, *host_name*) and the property labels (*id*, *name*), there are eight continuous variables including *price, minimum_night, number_of_reviews* and so on; three categorical (nominal) variables including *neighbourhood_group (5 levels), neighbourhood (221 levels), and room_type (3 levels)*; one date-format variable *last_review.*

The three categorical variables are transformed into factor variables for our analysis, and the missing values are only found in *last_review* and *reviews_per_month*. Also, the boxplot of *price* (Appendix A) shows a lot of extreme outliers. In particular, the listing price over $500 per night only account for 2.5% of the total Airbnb listing properties in NYC. Through research on Airbnb application, we find that the extreme values are from those super luxury lofts or townhouses. However, in this dataset, we have no other features to distinguish the average and the luxury listing properties other than price, so we focus our analysis on the "normal" price Airbnb listing and consider those over $500 as outliers needing to be removed. The abnormal values of 0 in *price* are removed as well.

Breakdown of the topic

We split the report into three main parts to answer the following questions respectively:

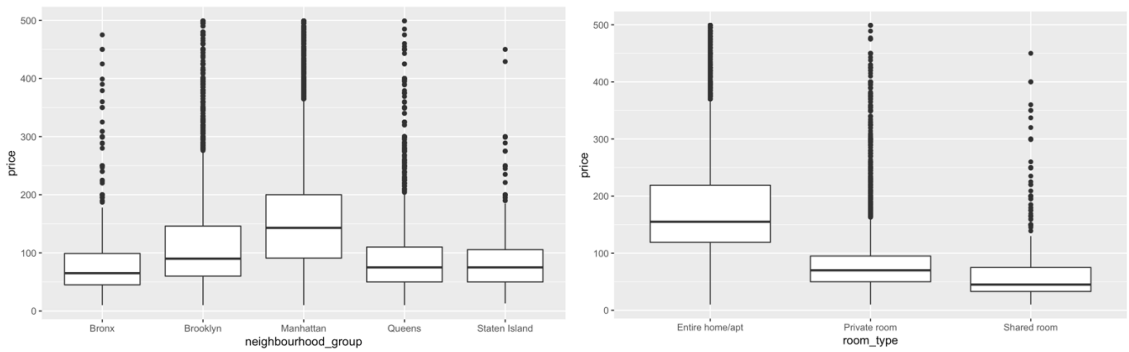
a) Is there a price difference among different neighbourhood groups and/or room types?
b) What are the various categories of Airbnb properties and their main features?
c) How to predict the price for a listing property based on the information given to us?

Part I – Descriptive Analysis - Price difference among different neighbourhood group and/or room type

In this part, we want to study whether different neighbourhood groups and/or room types can affect the listing price. Our initial intention is to use ANOVA for analysis.

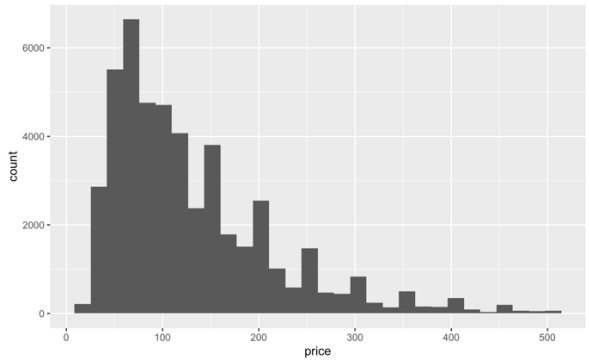
1.1 Data visualization

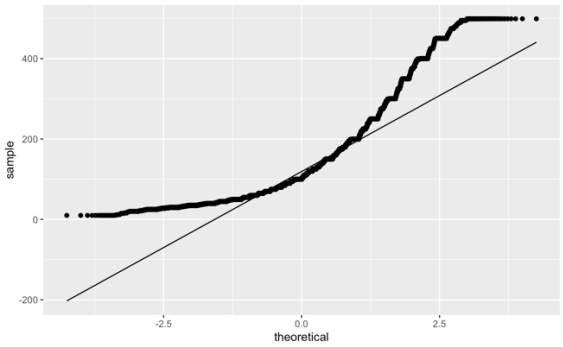
After the data cleaning, the boxplots drawn using the new dataset are more readable as following. The boxplots show that Airbnb properties in Manhattan tend to have the highest average price and the entire home/apartment room type tend to result in the highest listing price.



1.2 Assumption check before using ANOVA

Before using ANOVA to compare means between groups, we need to check its three assumptions: independence, normality, and homoescedascity. First, all observations are independent as we assume that individual property owners decide their prices independently. Second, both the histogram and Q-Q plot show that the normality test fails as the data is very positively skewed.





Third, the Levene's test is significant at the level of 0.01, so we reject the null hypothesis and conclude that the variances are significantly different between groups – that is, the assumption of homogeneity of variances has been violated.

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    4  453.89 < 2.2e-16 ***
      47644
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1.3 Transfer to a non-parametric method

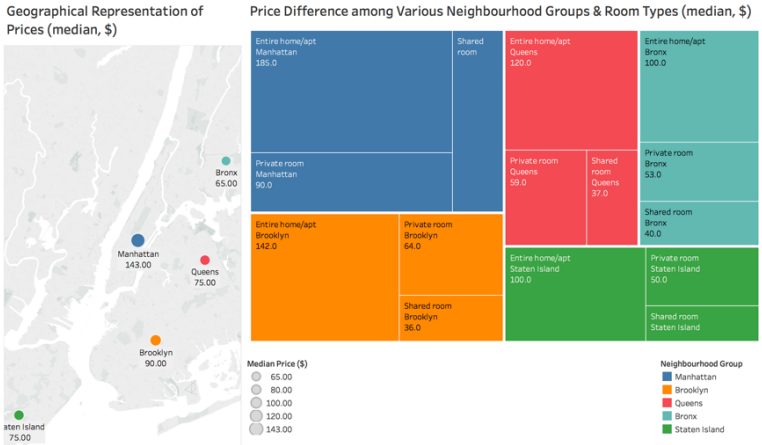
As the assumptions of ANOVA fail, we proceed with non-parametric method: the Kruskal-Wallis test, which is based on ranked data. The test finds whether the median price between different *neighbourhood_group* and *room_type* differs. The results show that the Kruskal-Wallis test is significant for different neighbourhood groups and different room types at the level of 0.01. So, we conclude that the location and room type of Airbnb listing in NYC significantly affect the listing price.

```
        Kruskal-Wallis rank sum test

data:  price by neighbourhood_group
Kruskal-Wallis chi-squared = 6604.5, df = 4, p-value < 2.2e-16
```

```
        Kruskal-Wallis rank sum test

data:  price by room_type
Kruskal-Wallis chi-squared = 22166, df = 2, p-value < 2.2e-16
```

1.4 Use Tableau Dashboard to visualize the overall picture

The following figure is a dashboard produced by Tableau. It shows that Airbnb properties in Manhattan have the highest median price ($) and those in Bronx have the lowest median price ($). In all the five neighbourhood groups, the entire home/apartment has the highest median price, and the private room has a higher medium price than the shared room. The result from this figure also matches the interpretation of our previous findings from the boxplots.



Part II - Exploratory Analysis - Discover different categories of Airbnb listing in NYC and their main features in price, etc.

Following Part I, we are interested in going one step further, using cluster analysis to explore the different categories of Airbnb listings in NYC.

2.1 Selection of clustering methods and variables

Considering the scale of our dataset, k-means clustering is a better choice, as hierarchical clustering has more time complexity and is more suitable for small datasets.

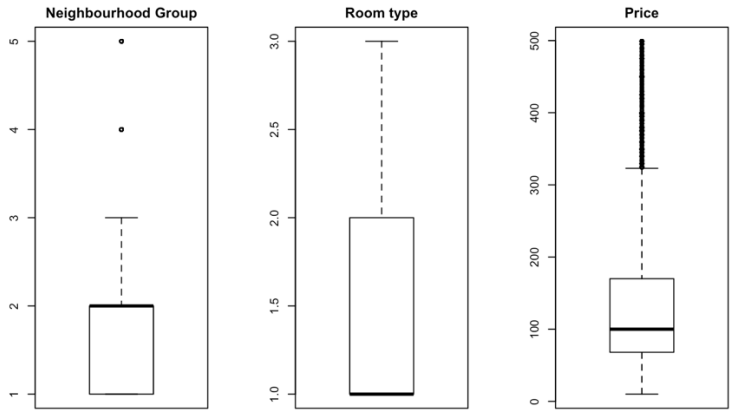
As to variable selection, we would like to focus our analysis on the three variables that have been studied in Part I to make it straightforward when interpreting the results.

2.2 Assumption check and data modification

First, *neighbourhood_group* is encoded in a way that the factor level follows the descending order of the median price in each group. And then we convert the factor variables *neighbourhood_group* and *room_type* to numeric values to meet the requirement of k-means clustering.

Second, missing values need to be checked. Since there is no missing value in these three variables, we continue with our analysis.

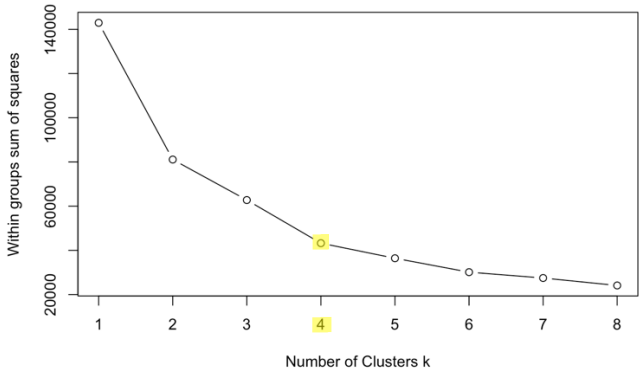
Third, extreme outliers will affect the performance of k-means clustering. The following figure shows that there are no extreme values in these three variables (as we consider the prices below $500 and neighbourhood group number above three not as outliers for practical reasons).



Fourth, we rescale the data to meet the comparability requirement of k-means algorithm.

2.3 Determine the optimal k

K-means clustering requires us to specify the number of clusters k before analysis. After comparing k being different values from 1 to 8, we find that when k = 4, the result is optimal based on a graphical representation (see figure below). The plot shows there is an apparent reduction in variation before k = 4, but after that, the variation does not go down as quickly.



2.4 K-means cluster analysis and plotting

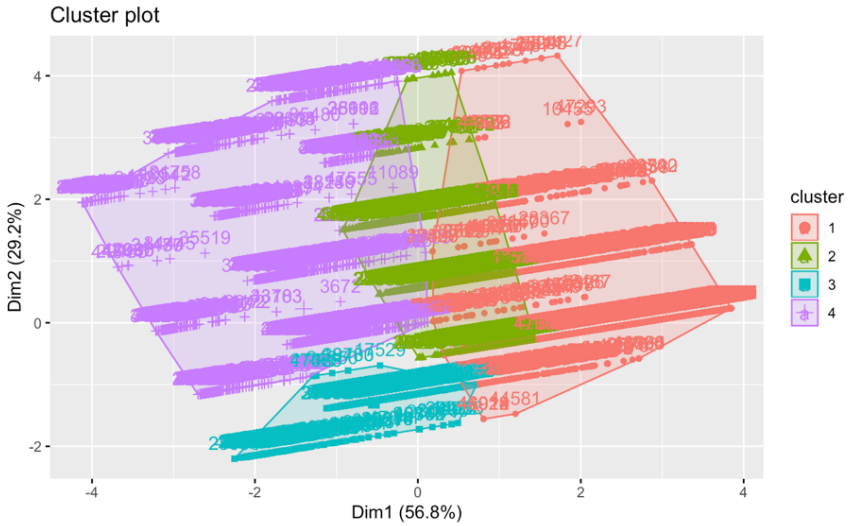
After implementing k-means clustering setting k = 4, we get the values of the four cluster centroids as well as the cluster plot as following.

Figure 1 – K means clustering output

Group.1	neighbourhood_group	room_type	price
1	-0.6073854	-0.8406076	1.77659930
2	-0.0478909	-0.9395793	0.04756088
3	-0.8928241	1.0016186	-0.43986218
4	0.8312960	0.9163861	-0.73136256

Figure 2 – Summary statistics of variables

neighbourhood_group	room_type	price
Min. :-0.8941	Min. :-0.9396	Min. :-1.4115
1st Qu.:-0.8941	1st Qu.:-0.9396	1st Qu.:-0.7298
Median : 0.2731	Median :-0.9396	Median :-0.3536
Mean : 0.0000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 0.2731	3rd Qu.: 0.8923	3rd Qu.: 0.4692
Max. : 3.7747	Max. : 2.7242	Max. : 4.3364



2.5 Explanation of the results

Compared with the summary statistics (Figure 2 above) of the three variables, and after connecting with the factor level value of *neighbourhood_group* (smaller number indicating the neighbourhood group with higher median price) and *room_type* (smaller number indicating the room type with higher median price), we can safely conclude what the four main clusters are and their main features.

Cluster	Main features
1	High price; Located in high price neighbourhood such as Manhattan; Most are entire home/apartment
2	Medium price; Located in medium price neighbourhood such as Brooklyn and Queens; Most are entire home/apartment
3	Moderately low price; Located in high price neighbourhood such as Manhattan; Most are private or shared rooms
4	Low price; Located in low price neighbourhoods such as Staten Island and Bronx; Most are private or shared rooms

Part III - Predictive Analysis - Price prediction model

Last but not least, we would like to combine all of the information in the dataset to establish a price prediction model using multiple regression analysis.

3.1 Variable selection

Initially, we include all reasonably related variables in our multiple regression model except labels (*id, name, host_id, host_name*), *neighbourhood* (too many levels to interpret and do not make much sense in the regression model), *latitude* and *longitude* (all listing properties are quite closely gathered in NYC and these variables do not make much sense in the regression model), *last review* (in date format which is not suitable for regression).

Referring to Appendix B, we find the correlation coefficient between *number_of_reviews* and *reviews_per_month* = 0.552 > 0.5. To solve the problem of multicollinearity between predictor variables, we decide to remove *reviews_per_month* as it has more missing values and we keep *number_of_reviews*. Appendix B also reveals some outliers exist in minimum_nights, etc. So, we will log all of the non-nominal variables before running the regression model.

3.2 Dummy variable creation

We convert the two factor variables—*neighbourhood_group* and *room_type*—to dummy variables based on their factor level so that we could use them in the regression model.

3.3 Pre-model assumption check

The predictor variables are quantitative or categorical, and the outcome (*price*) is quantitative, continuous, and unbounded. The non-zero variance assumption is met since the data clearly varies.

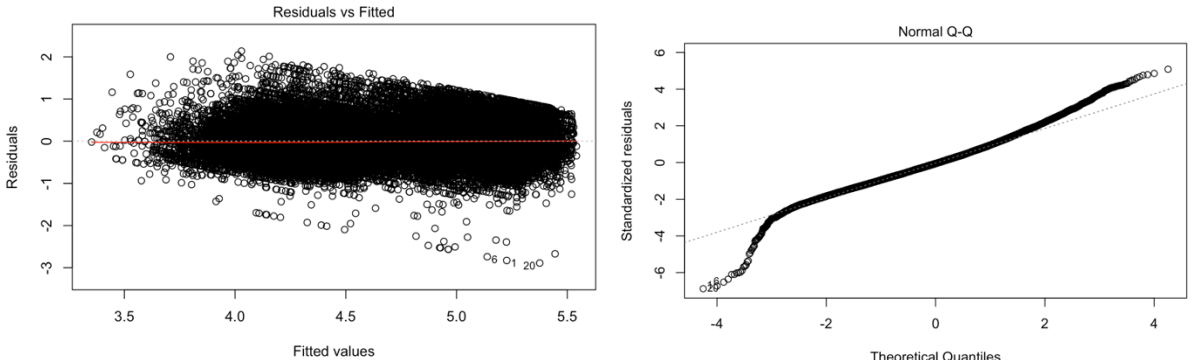
3.4 Model generation and model result

The following R output shows the summary of our multiple regression model.

```
Call:
lm(formula = price ~ neighbourhood_group + room_type + minimum_nights +
    number_of_reviews + calculated_host_listings_count + availability_365,
    data = ab_3_sub)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89065 -0.28024 -0.02312  0.25429  2.13672

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                           3.642586   0.018140 200.806  < 2e-16 ***
neighbourhood_groupManhattan_v_NMA    0.557307   0.013231  42.122  < 2e-16 ***
neighbourhood_groupBrooklyn_v_NMA     0.267618   0.013203  20.270  < 2e-16 ***
neighbourhood_groupQueens_v_NMA       0.126800   0.013970   9.077  < 2e-16 ***
neighbourhood_groupStaten_Island_v_NMA -0.001809   0.025402  -0.071  0.94324
room_typeEntire_home_v_NMA            1.160511   0.012816  90.553  < 2e-16 ***
room_typePrivate_room_v_NMA           0.399060   0.012779  31.227  < 2e-16 ***
minimum_nights                       -0.096164   0.002501 -38.447  < 2e-16 ***
number_of_reviews                    -0.039146   0.001365 -28.668  < 2e-16 ***
calculated_host_listings_count        0.007567   0.002582   2.930  0.00339 **
availability_365                      0.037390   0.000910  41.088  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4203 on 47638 degrees of freedom
Multiple R-squared:  0.5342,    Adjusted R-squared:  0.5341
F-statistic:  5464 on 10 and 47638 DF,  p-value: < 2.2e-16
```

3.5 Assumption check for the model generated (Diagnosing of the model)

We check the multicollinearity of the model and find all the VIFs are lower than 10 and the average VIF (=1.3279) is close to 1, showing no existence of multicollinearity. The “Residuals vs. Fitted” plot shows that a random array of dots evenly dispersed around zero, so the homoscedascity assumption is met. The majority of points lie along with the straight line in the Q-Q plot, so the normality assumption is checked.



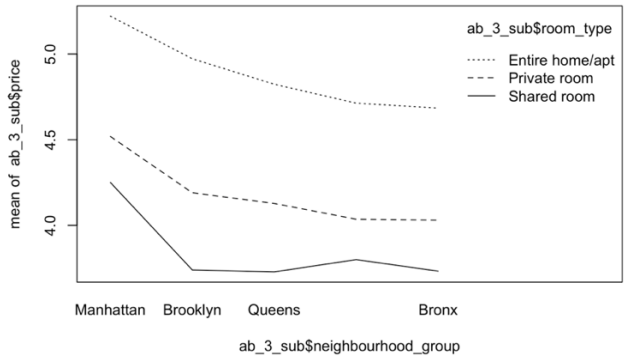
The Durbin Watson test result (d=1.92) is close to 2, indicating independence of residuals (see Appendix C for the result). Appendix D also shows that all points fall within the Cook’s distance line, so there is no concern for influential points.

3.6 Cross validation

We use k-fold cross validation to test the robustness of our model, and a typical value of k=10 is chosen to train the model. As can be seen from Appendix E, the R squared value is almost the same with our original model, indicating that the accuracy of our model is robust.

3.7 Interaction effect

We go one step further to investigate whether there is an interaction effect between *neighbourhood_group* and *room_type.* An interaction plot shows that all the three rooms types are approximately parallel with each other, indicating no interaction effect between them.



3.8 Explanation of the results

We conclude that there is generally a price increase from Bronx to Manhattan. Price of Airbnb properties in Staten Island is not significantly different from that in Bronx ($p=0.94$). However, the price in all other three locations are higher than that in Bronx ($p<0.01$). Manhattan has the highest price, then Brooklyn does, and Queens follows. Referring back to the dashboard produced by Tableau, Manhattan has the highest median price in the centre of NYC. Bronx and Staten Island being the furthest from Manhattan, have relatively the lowest price. Brooklyn and Queens tend to have medium price being closer to Manhattan.

Also, we conclude that when the room type is more exclusively being used, the price tends to be higher. That is, both of the entire home room type and the private room type have a higher price than the shared room at 1% significance level.

The other predictor variables are all significant at 1% level to the model as well. For instance, the price significantly decreases as the requirement for minimum nights increases. Our prediction model can explain 53.41% of the variation in the price, which is desirable.

Conclusions and Implications

From Part I to Part III, we answer the three questions raised at the beginning respectively:

a) There is a price difference among different types of neighbourhood group and room type. In other words, both the location and the room type have an influence on prices.
b) There are mainly four categories of Airbnb listings in NYC, each having different features based on price level, location and room type.
c) A price prediction model is established, and several factors significantly affect the price.

Our results also have important practical insights. The Airbnb hosts can set a higher price if their property is close to Manhattan and if they offer an exclusive right for the property use. If a person only has limited budget, he/she could choose to live a bit further from Manhattan or to share some common areas with others.

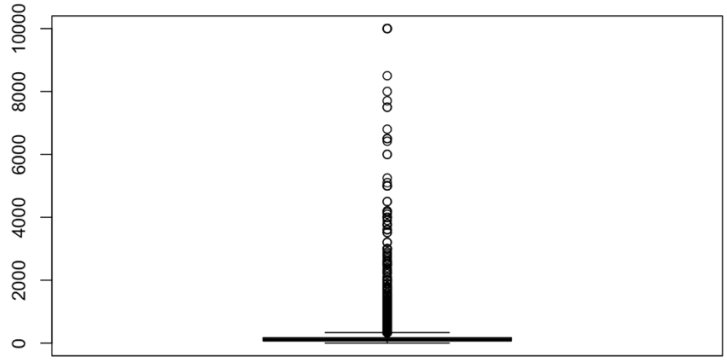
Limitation and Further Studies

We use k-means clustering since it is infeasible to do hierarchical clustering in R with such a big dataset. However, hierarchical clustering can provide a more comprehensive picture of the data. We may use other distributed systems such as Spark to implement it in the future.

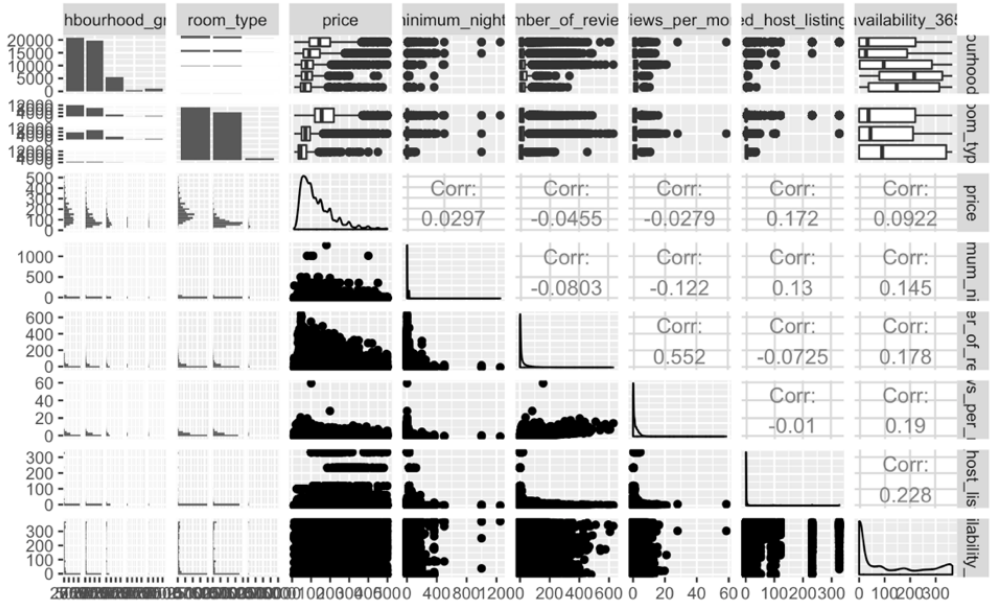
Also, more efforts will be put into explaining the mechanism of the prediction model. For example, we may study why number of reviews negatively affects the price, dividing reviews into good ones and bad ones. Besides, a sub-study on the pricing mechanism of shared rooms may be conducted, as it does not strictly follow the pattern with the other two room types through the interaction effect plot.

Appendix

Appendix A – Boxplot of *price* before data cleaning



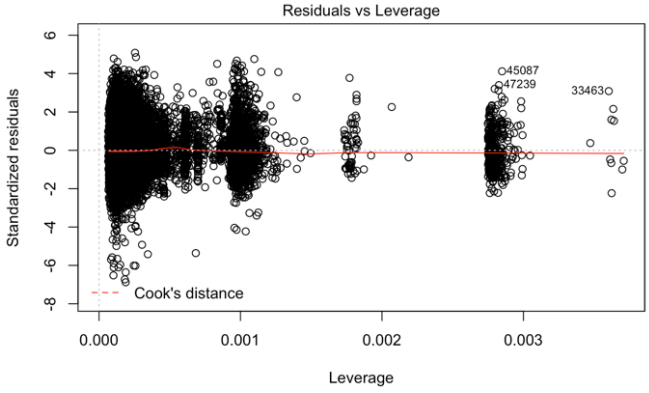
Appendix B – Visual inspection of all the predictor variables



Appendix C – the Durbin Watson test result

```
 lag Autocorrelation D-W Statistic p-value
   1      0.04218697       1.91558       0
Alternative hypothesis: rho != 0
```

Appendix D – Residuals vs. Leverage plot



Appendix E – k-fold cross validation result

```
Linear Regression

47649 samples
    6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 42884, 42884, 42885, 42884, 42884, 42884, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.4203367  0.5340689  0.3265587

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Appendix F – Please find the R code for this report in the file named "code.Rmd" under the "eric-jiang-1997/stats-and-machine-learning/pricing_analytics_nyc_airbnb" folder.

Reference

Dgomonov. (2019, August 12). New York City Airbnb Open Data. Retrieved April 15, 2020, from https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data