IEOR E4650-002 Project Report
Business Analysis on NYC Airbnb

Submitted by: Jiyin Chen (jc5498), Xinyi Lin (xl3024), Gaole Lyu (gl2704), Yifan Xia (yx2610), and Yifei Zhu (yz3919)

I. Introduction

Airbnb, known as a vacation rental online platform, provides both travelers and guests a unique and personalized way of experiencing the world. Data analysis on millions of listings provided by the platform is a crucial factor for the company to construct business decisions, guiding marketing initiatives and so forth. In order to consistently provide high-quality service and expand the business, it requires the company to understand what customers value the most when choosing a host/accommodation. Therefore, our team takes the initiative to analyze the factors that impact the review score ratings in NYC specifically by implementing Airbnb's Open Data resource "Detailed Review Data for listings in New York City" dataset. The main goal for this project is to help the company have a better understanding of what may potentially affect the users' picks by analyzing the key factors that influence the review score rating of a listing. The main programming language used for analyzing the data is R.

II. Data Cleansing

Our data cleansing trims down covariates that we think have little impact on the review score ratings, converts some values to another format, and generates some new values based on raw ones.

Starting with **74** columns, we perform data cleansing as the following steps:
1) We change all the "True" or "False" values from "t"/"f" to 1/0.
2) We handle N/A, either replace N/A with values that are consistent with column content or remove the entire row. For example, for column **host_response_time**, we convert all "N/A" into "a few days or more", for columns **host_response_rate** and **host_acceptance_rate**, we replace "N/A" with 0.
3) We remove superfluous words and symbols such as '%', '$','private',' shared'.
4) For **amenities** and **host_verifications**, the list-like values show us what these amenities/verifications are, but we care more about the number of amenities/verifications the host has to offer, so we replace texts with list length and convert values to numeric.
5) Columns **first_review** and **last_review** indicate the date of first and last review of this listing, so we combine them and generate a new column **review_duration** by computing the difference in days, and then divide the resulting by 365 to get the duration of reviews in years.
6) We remove all the other unwanted columns. **45** columns are removed in total and we make sure that there are no columns containing N/A values.

Now we have 74 + 1 - 2 - 45 = **28** columns left:

```
 [1] "host_response_time"            "host_response_rate"
 [3] "host_acceptance_rate"          "host_is_superhost"
 [5] "host_total_listings_count"     "host_verifications"
 [7] "host_has_profile_pic"          "host_identity_verified"
 [9] "neighbourhood_group_cleansed" "room_type"
[11] "accommodates"                  "bathrooms_text"
[13] "bedrooms"                      "beds"
[15] "amenities"                     "price"
[17] "minimum_nights"                "maximum_nights"
[19] "availability_30"               "availability_60"
[21] "availability_90"               "availability_365"
[23] "number_of_reviews"             "number_of_reviews_ltm"
[25] "number_of_reviews_l30d"        "review_scores_rating"
[27] "instant_bookable"              "review_duration"
```

Figure 1: 28 Column Names

with each containing no N/A values. We have then finished our data cleansing and can proceed to further steps of analysis.

III. Exploratory Data Analysis

To have a better understanding of the relationship between the covariates before modeling the data, we generate a correlation heatmap for all the variables (Figure 2). In the correlation heatmap, every cell represents the correlation between the column factor and the row factor. According to the figure, it is obvious that most of the input variables that we keep in the data are not strongly correlated with one another; in other words, we cannot represent one factor from the other factors. Although some covariates are correlated, they are not collinear. Therefore, we do not need to delete any more repetitive input variables before we build the model.

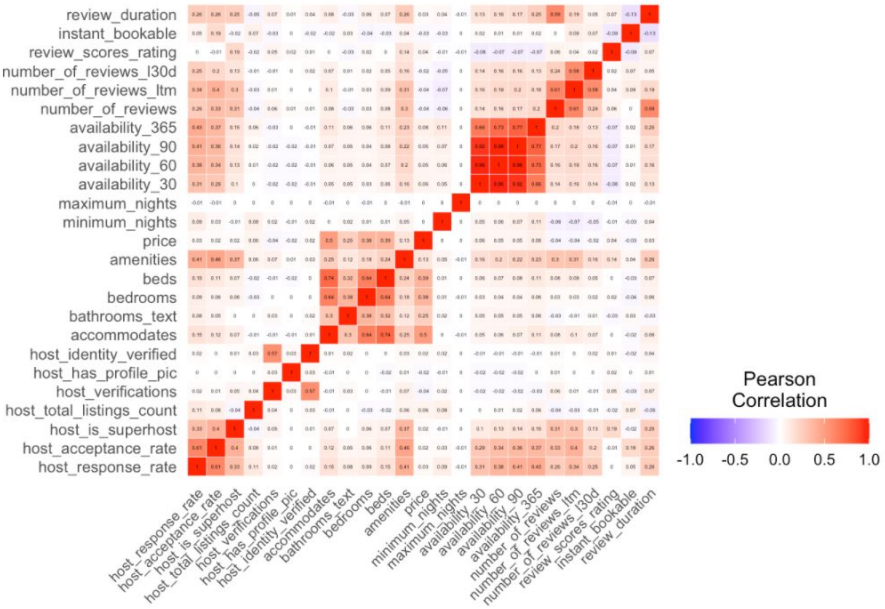

Figure 2: Correlation Heat Map

Weak correlation can result from either two variables being independent or their relationship being non-linear. Acknowledging that the covariates are not strongly correlated to one another, we would like to know the relationship or trend between the covariates and the dependent variable. In order to do so, we plot every input variable against our output variable, **review_scores_rating**. Reading the plots, we can grasp three main patterns:

1) Figure 3 shows the distribution of a continuous covariate that skews toward the upper corner.
2) Figure 4 is a dummy variable which distributes differently at 0 and 1.
3) Figure 5 represents a relatively even spread of a discrete variable.

We can definitely see that there are relationships between input and dependent variables. However, it is not obvious how the covariates can affect the dependent variable. Therefore, we decide to perform further analysis, such as multiple regression analysis.

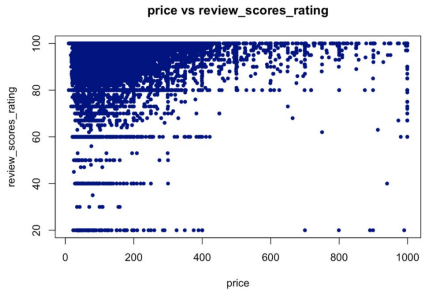

Figure 3: Price vs RSR

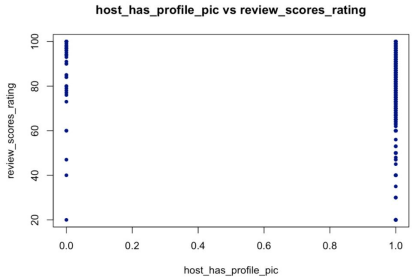

Figure 4: Hosts has profile pic vs RSR

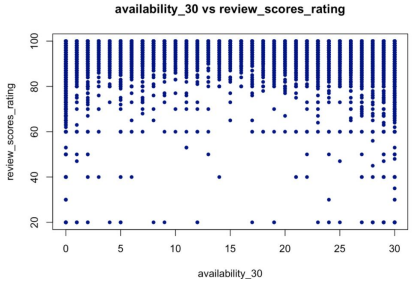

Figure 5: Availability_30 vs. RSR

IV. Multi-regression analysis + Lasso

We follow a regular model selection and assessment step. We randomly generate 50% of our data as the training set, 25% as the validation set, and 25% as the test set. Then, we use the training set and all variables (after data collection and clean process) to create a multiple regression model to analyze the relation between those variables and final rating score. The following figures are parts of the result of our base model.

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.68 on 15022 degrees of freedom
## Multiple R-squared:  0.09233,    Adjusted R-squared:  0.09027
## F-statistic: 44.94 on 34 and 15022 DF,  p-value: < 2.2e-16
```

Figure 6

Firstly, the P-value of the regression model is less than 2.2e^-16, which means that the regression relation between dependent variable and independent variables exists. Adjusted R-squared is 0.09027

and it means only 9% of the total variation in reviews scores rating can be explained by the regression model.

```
## (Intercept)                                  < 2e-16 ***
## host_response_timewithin a day               0.52049
## host_response_timewithin a few hours         0.29079
## host_response_timewithin an hour             0.17731
## host_response_rate                           0.88582
## host_acceptance_rate                         1.26e-10 ***
## host_is_superhost                            < 2e-16 ***
## host_total_listings_count                    0.18181
## host_verifications                           0.00160 **
## host_has_profile_pic                         0.00224 **
## host_identity_verified                       0.01703 *
## neighbourhood_group_cleansedBrooklyn         0.02543 *
## neighbourhood_group_cleansedManhattan        0.47839
## neighbourhood_group_cleansedQueens           0.07728 .
## neighbourhood_group_cleansedStaten Island    0.98331
## room_typeHotel room                          8.20e-09 ***
## room_typePrivate room                        3.05e-08 ***
## room_typeShared room                         0.09393 .
## accommodates                                 1.92e-08 ***
## bathrooms_text                               2.61e-09 ***
## bedrooms                                     0.03746 *
## beds                                         0.06844 .
## amenities                                    < 2e-16 ***
## price                                        7.29e-15 ***
## minimum_nights                               0.05823 .
## maximum_nights                               0.01891 *
## availability_30                              0.00169 **
## availability_60                              0.66300
## availability_90                              0.41282
## availability_365                             9.47e-09 ***
## number_of_reviews                            0.15709
## number_of_reviews_ltm                        0.22033
## number_of_reviews_l30d                       0.20968
## instant_bookable                             3.32e-13 ***
## review_duration                              0.00286 **
```

Figure 7

Then we go through the P value of the independent variables. We found at 1% significance level, there 19 variables are statistically significant. Among those variables, **amenities** (0.1560), **host_is_superhost** (4.083), **accommodates** (-0.4826) have the top 3 highest coefficient values.

After building the base model, we decide that Lasso is the best parameter selection method for our data. Since the cleaned data has 28 columns and over 30000 rows, it is overly computationally intense to use best subset selection. Also, the number of observations is far greater than the number of predictors, so all relevant predictors are possible to be picked by Lasso. Lastly, as we can see from the correlation heatmap, only a few variables are highly correlated (**beds** & **bathrooms**, **availbility_365** & **availbility_90** & **availbility_60** & **availbility_30**) and none of them are highly collinear. Thus Lasso will not need to randomly pick one from many similar predictors.

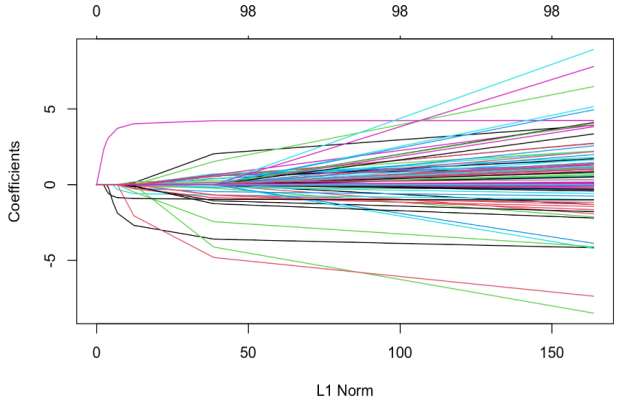

Figure 8

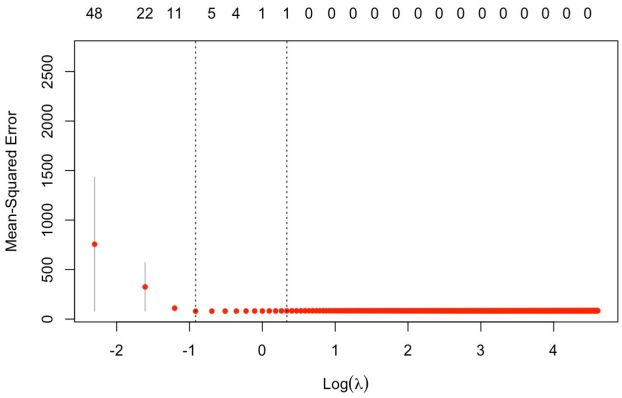

Figure 9

The L1 norm (Figure 8) is the regularization term for Lasso. A better way to look at the x-axis is the maximum permissible value the L1 norm can take. Therefore, an L1 norm of zero gives an empty model, and as one increases the L1 norm, variables will "enter" the model as their coefficients take non-zero values. In our model, we try 1000 lambda values from 0 to 100 and find the first log-lambda with the smallest MSE located around -1 (Figure 9) which makes our best lambda 0.4.

V. Result

Although we expect the Lasso regression model to have better performance, we ultimately find that our base model has slightly smaller validation MSE (Figure 10) which indicates higher prediction accuracy. Therefore we choose the Ordinary Least Squares (OLS) model as our final model and retrain it with both training and validation data to obtain prediction results on test data.

Lasso MSE	Base Model MSE
78.9356572264889	75.7249638003555

Figure 10

Following figures (Figure 11 & Figure 12) are optimal model coefficients and prediction MSE on test data is 82.26686.

```
Residuals:
    Min      1Q  Median      3Q     Max
-76.051  -2.176   1.446   5.113  18.834

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 8.895e+01  1.261e+00  70.548  < 2e-16 ***
host_response_timewithin a day             -1.610e+00  6.485e-01  -2.483 0.013052 *
host_response_timewithin a few hours       -2.393e+00  7.266e-01  -3.294 0.000990 ***
host_response_timewithin an hour           -2.523e+00  7.407e-01  -3.407 0.000659 ***
host_response_rate                          1.598e-02  7.487e-03   2.134 0.032826 *
host_acceptance_rate                       -1.545e-02  1.904e-03  -8.111 5.28e-16 ***
host_is_superhost                           4.083e+00  1.609e-01  25.375  < 2e-16 ***
host_total_listings_count                  -1.325e-03  8.443e-04  -1.569 0.116675
host_verifications                          1.819e-01  3.868e-02   4.702 2.59e-06 ***
host_has_profile_pic                        3.722e+00  1.185e+00   3.142 0.001682 **
host_identity_verified                     -3.903e-01  1.792e-01  -2.177 0.029454 *
neighbourhood_group_cleansedBrooklyn        6.376e-01  3.748e-01   1.701 0.088893 .
neighbourhood_group_cleansedManhattan       3.918e-03  3.781e-01   0.010 0.991734
neighbourhood_group_cleansedQueens          5.568e-01  3.968e-01   1.403 0.160563
neighbourhood_group_cleansedStaten Island   5.916e-01  7.778e-01   0.761 0.446868
room_typeHotel room                        -4.500e+00  6.398e-01  -7.033 2.08e-12 ***
room_typePrivate room                      -1.174e+00  1.510e-01  -7.773 8.02e-15 ***
room_typeShared room                       -2.229e+00  4.366e-01  -5.105 3.34e-07 ***
```

Figure 11

```
accommodates                               -4.826e-01  5.849e-02  -8.250  < 2e-16 ***
bathrooms_text                             -6.750e-01  1.626e-01  -4.150 3.33e-05 ***
bedrooms                                    3.402e-01  1.175e-01   2.896 0.003781 **
beds                                       -1.293e-01  8.726e-02  -1.482 0.138402
amenities                                   1.412e-01  8.163e-03  17.298  < 2e-16 ***
price                                       5.848e-03  7.353e-04   7.953 1.90e-15 ***
minimum_nights                             -4.722e-03  3.172e-03  -1.489 0.136577
maximum_nights                             -9.569e-09  4.078e-09  -2.347 0.018951 *
availability_30                            -6.069e-02  2.079e-02  -2.918 0.003521 **
availability_60                             3.427e-03  2.288e-02   0.150 0.880965
availability_90                             1.059e-02  1.154e-02   0.918 0.358621
availability_365                           -3.928e-03  6.992e-04  -5.618 1.96e-08 ***
number_of_reviews                          -3.084e-03  1.743e-03  -1.769 0.076887 .
number_of_reviews_ltm                       1.139e-02  8.633e-03   1.320 0.186953
number_of_reviews_l30d                      8.716e-02  7.322e-02   1.190 0.233886
instant_bookable                           -9.888e-01  1.329e-01  -7.440 1.04e-13 ***
review_duration                             1.428e-01  4.086e-02   3.495 0.000475 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.754 on 22551 degrees of freedom
Multiple R-squared:  0.08759,   Adjusted R-squared:  0.08621
F-statistic: 63.67 on 34 and 22551 DF,  p-value: < 2.2e-16
```

Figure 12

VI. Conclusion

Based on the analysis from the previous section, it is evident that some factors are more important than other factors on affecting the dependent variable. **Review_scores_rating**, known as a direct way for customers to reflect on their user experience to future customers, is important for Airbnb to maintain for their future market growth.

Through Exploratory Data Analysis, we could clearly notice that some factors, such as **price**, may have a main spotting area: for **price** varied from 400 to 700, the review score rating tends to be on the higher end while the rest may have slightly lower score ratings when **price** is below 400 or above 700. Through this recognition, we would recommend Airbnb to have a suggestion for the hosts who price their listings below 400 or above 700 to be more careful when they are serving the customers.

In addition to **price**, **host_response_rate** is also comparably obvious on its trends, where the host with higher response rate tends to have a bigger range of review scores ratings and the higher scores ratings mainly spot on the higher end of their response rates. Based on this finding, we would recommend Airbnb to encourage the hosts to actively interact with customers while also being careful about messages they send to their customers.

Furthermore, we would suggest Airbnb to actively promote the benefits of obtaining a superhost badge, which can not only increase the score ratings for the host but also potentially elevate the overall reputation of Airbnb.

Reference:

The dataset we use: **reviews.csv.gz**, under the category “New York City, New York, United States”, with 05 October, 2020 as “Date Compiled”, through link: http://insideairbnb.com/get-the-data.html.