Pricing Your Airbnb in NYC

Mengyao Liu

Oct. 1st, 2019

Keywords: business — internet — regression analysis — New York City — hotels and accommodations

1 Introduction

Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world. In a big city like New York City (NYC), which attracts millions of tourists every year[1], there are over 10,000 hosts posting over 45,000 house listings on Airbnb in 2019 alone[2]. Suppose you have a spare room/house in NYC, you would like to take advantage of the prospering tourism in NYC and rent the room/house at a reasonable price, so that the spare room/house can be made use of most of the time to gain profit. However, making an appropriate price is no easy task. On the one hand, the host would prefer to set the price high to gain more profit. On the other hand, too high a price will decrease the willingness of the potential guest booking the house and thus result in fewer sales. Fortunately, thanks to Internet, we can find public dataset from Airbnb and help make a wise decision by learning from the history.

This report aims to predict the best price range of a room/house for a host on Airbnb using regression analysis to make sure the room/house will be rented more than two thirds of the time of a year.

We use the dataset describing the listing activity and metrics in NYC, NY for 2019 from Kaggle[2]. Essentially the data file is a `csv` table. This table includes all needed information to find out more about hosts, geographical availability, necessary metrics to make predictions and draw conclusions. This public dataset is part of Airbnb, and the original source can be found on this website.

From the viewpoint of a host, the predictors we will use are `neighborhood_group`, `room_type` and `minimum_nights`. `neighborhood_group` is a string of the neighborhood location name. `room_type` is a string of space type. The string will be one of the following three: Entire home/apt, Private room, Shared room. `minimum_nights` is a number specifying the amount of nights minimum. The response will be `price`. `price` denotes the price in dollars.

2 Exploratory Data Analysis

2.1 Data Cleaning

There are 48,895 observations in total and there are 16 variables. First we check the missing data. It seems the columns that miss data are `name`, `host_name`, `last_review` and `review_per_month` as shown in Figure 1. `name` and `host_name` are not useful for the problem so we do not need to worry about them. As to the 10,052 observations with missing `last_review` and `review_per_month`, after

scrutinization it turns out that they all have `number_of_reviews = 0`. This means our dataset is quite clean and we do not have the problem of missing data.

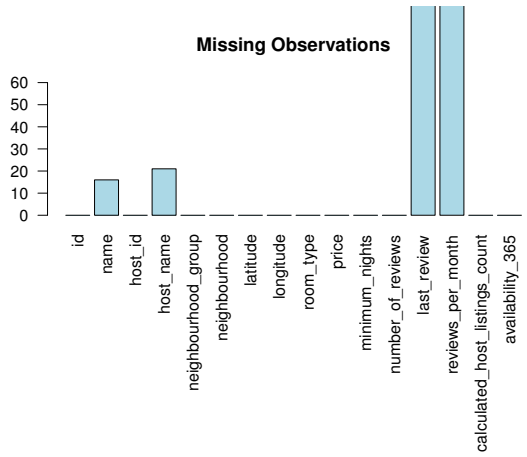

Figure 1

We have also found that there are 11 observations that have `price = 0`. This is unusual and we do not aim to investigate the reasons here. Thus we remove those 11 observations.

2.2 Exploratory Summary

To investigate pricing, we plot the distribution of `price` by `neighbourhood_group` (see Figure 2), `room_type` (see Figure 3), and `minimum_nights` (see Figure 4), respectively. Note that we only use observations that are "Popular". We made these plots to gain insight on the significance of each predictor to help determine whether to reduce any of them later as well as insight on the signs of the coefficients.

Apparently, there is difference in `price` between different `neighbourhood_group` and `room_type`. From Figure 2, it can be shown that the rooms/houses located in Manhattan tend to have the highest price while the rooms/houses located in Bronx tend to be priced low. From Figure 3, there is a clear order of price among the three room types with "Entire home/apt" pricing the highest, followed by "Private room", and "Shared room" with the lowest price. Therefore, we can expect `neighbourhood_group` and `room_type` will be strong predictors for `price`. The variation of `price` with `minimum_nights` based on Figure 4 is not obvious so far.

3 Regression Model Building

3.1 Initial Manipulation

Since we expect the room/house to be rented most of the time, we would treat the listings that are rented most of the time in 2019 as successful examples to learn from. In the dataset, `availability_365` represents the number of days when listing is available for booking. Thus we only use part of the observations which satisfy `availability_365` < 122. There are 30,632 satisfactory observations.

`neighborhood` and `neighborhood_group` are correlated and we should only use one of them. After a test of simple linear regression with only the predictor `neighborhood`, we notice that the p-values for the majority of `neighborhood` are insignificant (> 0.3). There are 221 categories in `neighborhood`, which also significantly increases the parameter number and thus leads to high variance. We notice the number of observations for each category of `neighborhood` varies from 10 to 800 and most of

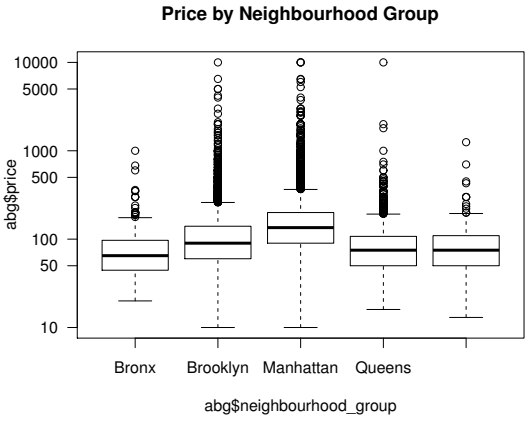

Figure 2

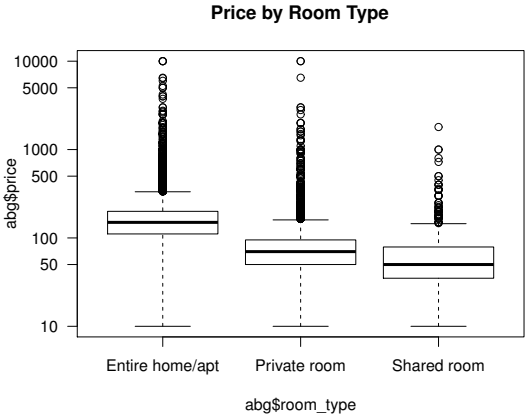

Figure 3

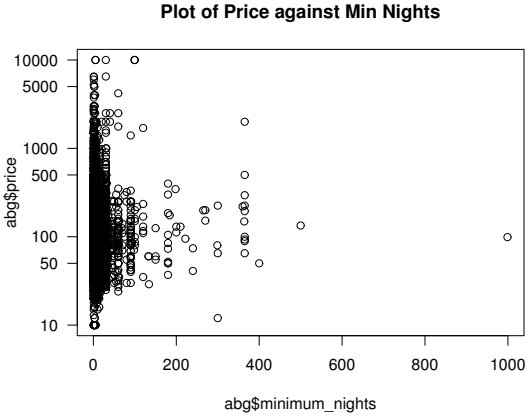

Figure 4

them have fewer than 100 observations. The limited number of observations for each category makes it hard to derive significant coefficient for each category. Thus we decide to use `neighborhood_group` instead of `neighborhood` as one of the predictors.

We perform least squares regression and the results are shown in Figure 5. We see is R^2 very small and RSE is very large. That means the fitting is not working well.

```
Call:
lm(formula = price ~ room_type + minimum_nights + neighbourhood_group)

Residuals:
   Min     1Q Median     3Q    Max
-484.2  -51.8  -18.8   11.7 9907.5

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   142.19500    8.98187  15.831  < 2e-16 ***
room_typePrivate room         -98.07529    2.32904 -42.110  < 2e-16 ***
room_typeShared room         -108.01571    8.28506 -13.037  < 2e-16 ***
minimum_nights                  0.37215    0.08127   4.579 4.69e-06 ***
neighbourhood_groupBrooklyn    24.99183    9.02521   2.769  0.00562 **
neighbourhood_groupManhattan   69.26126    9.03223   7.668 1.80e-14 ***
neighbourhood_groupQueens      11.11895    9.58126   1.160  0.24586
neighbourhood_groupStaten Island  13.53212   19.54262   0.692  0.48867
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199.4 on 30624 degrees of freedom
Multiple R-squared:  0.07845,   Adjusted R-squared:  0.07824
F-statistic: 372.4 on 7 and 30624 DF,  p-value: < 2.2e-16
```

Figure 5

3.2 Diagnostics

We plot the fitting results in Figure 6. From the residual plot, we notice there are definitely a few outliers that result in high residuals. Also the variance is not constant as the residuals are more spread out for larger fitted values. Based on the Box Cox output shown in Figure 7, we attempt raising the response variable to an exponent of -0.28.

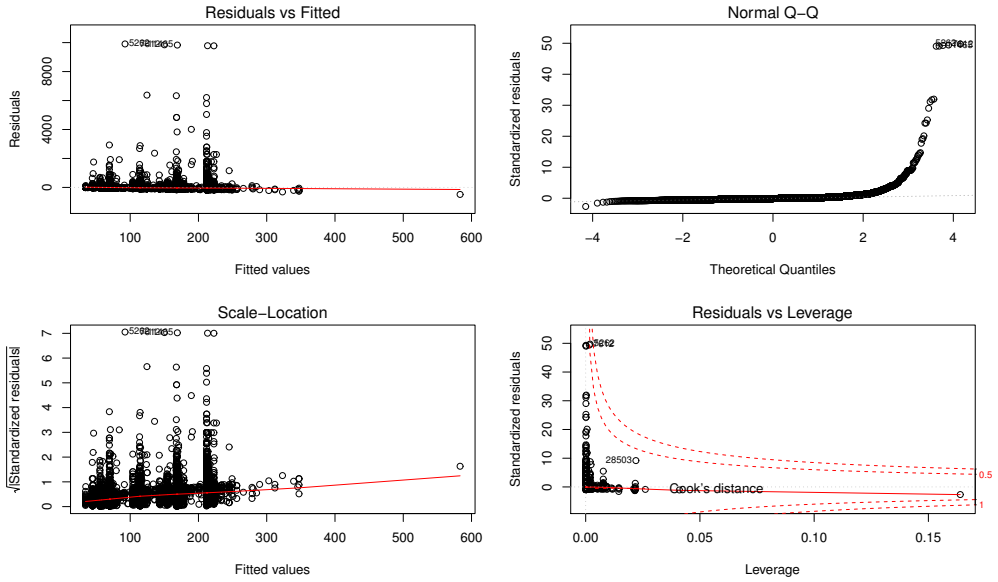

Figure 6

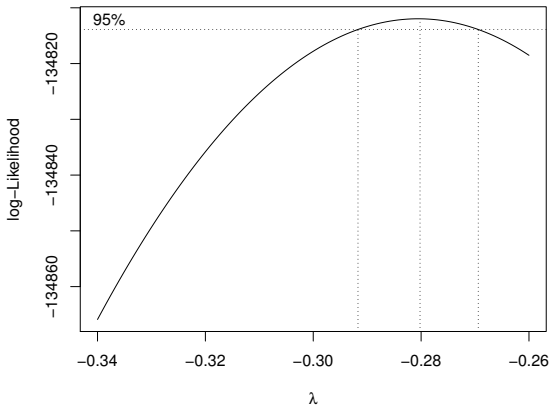

Figure 7

3.3 Improved Modeling Results

We display the new fitting results in Figure 8. According to the residual plots shown in Figure 9, which look a lot better after the transformation, the variance appears constant and the residuals are close to 0. Although there are some high leverage points seen from the leverage plot in Figure 9, they are not influential so that is not a problem. We notice RSE significantly drops and R^2 significantly increases compared with the initial fitting results shown in Figure 5.

```
Call:
lm(formula = price^(-0.28) ~ room_type + minimum_nights + neighbourhood_group)

Residuals:
      Min        1Q    Median        3Q       Max
-0.257405 -0.019473  0.000803  0.022134  0.292496

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    2.740e-01  1.587e-03 172.619  < 2e-16 ***
room_typePrivate room          5.738e-02  4.116e-04 139.412  < 2e-16 ***
room_typeShared room           8.276e-02  1.464e-03  56.529  < 2e-16 ***
minimum_nights                 1.291e-04  1.436e-05   8.992  < 2e-16 ***
neighbourhood_groupBrooklyn   -1.997e-02  1.595e-03 -12.523  < 2e-16 ***
neighbourhood_groupManhattan  -4.205e-02  1.596e-03 -26.346  < 2e-16 ***
neighbourhood_groupQueens     -1.100e-02  1.693e-03  -6.498 8.29e-11 ***
neighbourhood_groupStaten Island -6.320e-03  3.453e-03  -1.830   0.0672 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03524 on 30624 degrees of freedom
Multiple R-squared:  0.4746,    Adjusted R-squared:  0.4744
F-statistic:  3951 on 7 and 30624 DF,  p-value: < 2.2e-16
```

Figure 8

According to the p-values the coefficients of all the predictors are significant at 0.1% level except the coefficient for the last class of `neighborhood_group` is significant at 10% level. Thus there is no need to reduce any predictors. In order to investigate the influence of `room_type` and `neighborhood_group` on `price` more precisely, we compare the means in `price` between each pair of `room_type` in Figure 10 and `neighborhood_group` in Figure 11, respectively. From Figure 10, we see that all the coefficients are significant at 0.1% and positive. That means different `room_type` definitely makes a difference. In particular, the `price` of "Entire home/apt" would be higher than that of "Private room" and the `price` of "Private room" would be higher than that of "Shared room". Note that the response is $\texttt{price}^{-0.28}$ so the interpretation of the sign of the coefficients should flip from normal. From Figure 11, we can see not all pairs are significantly different. The

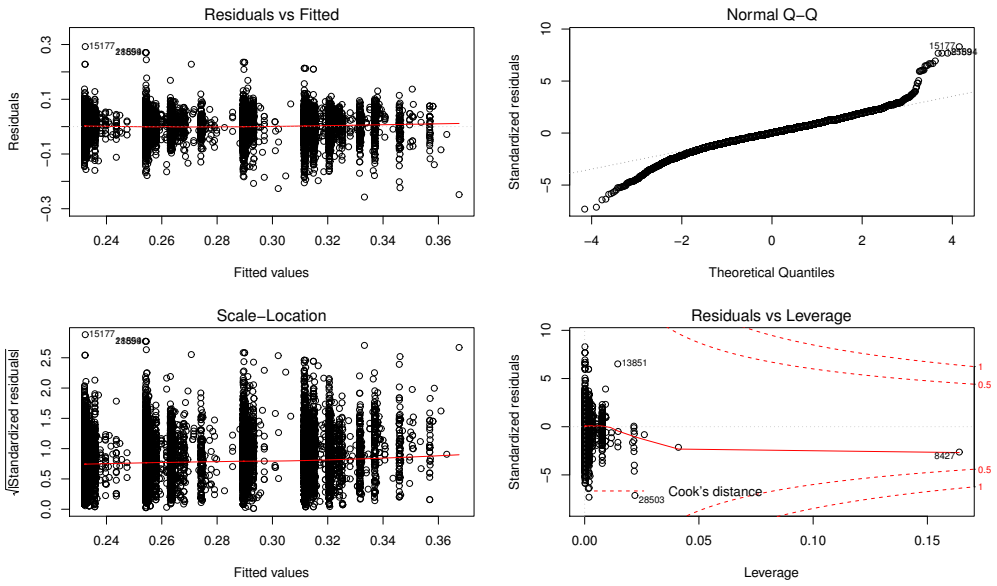

Figure 9

order of `price` based on `neighborhood_group` assuming other predictors hold the same should be Manhattan > Brooklyn > Queens > StatenIsland $\gtrsim$ Bronx.

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = price^(-0.28) ~ room_type + minimum_nights + neighbourhood_group)

Linear Hypotheses:
                                       Estimate Std. Error t value Pr(>|t|)
Private room - Entire home/apt == 0 0.0573760  0.0004116  139.41   <2e-16 ***
Shared room - Entire home/apt == 0  0.0827602  0.0014640   56.53   <2e-16 ***
Shared room - Private room == 0     0.0253841  0.0014662   17.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Figure 10

4 Conclusions

We have built a linear model to estimate `price` based on `neighborhood_group`, `room_type` and `minimum_nights`. According to the fitting results shown in Figure 8, a higher `minimum_nights` tends to lead to a lower `price`. The order of `price` based on `room_type` assuming other predictors hold the same should be Entirehome/apt > Privateroom > Shareroom. The order of `price` based on `neighborhood_group` assuming other predictors hold the same should be Manhattan > Brooklyn > Queens > StatenIsland $\gtrsim$ Bronx.

We also carry out simple linear regressions with the predictor `number_of_reviews` and `reviews_per_month`, respectively and it turns out they are both related to `price`. So maybe after the host puts his listing into the market and starts to get reviews gradually, soon the host would need another model at least adding in `number_of_reviews` and `reviews_per_month` as predictors to adjust the `price`.

As mentioned in §3.2, we notice there are a number of outliers and high leverage data points. We calculate the studentized residuals for the new fitting results and list the observations with

```
        Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = price^(-0.28) ~ room_type + minimum_nights + neighbourhood_group)

Linear Hypotheses:
                                Estimate Std. Error t value Pr(>|t|)
Brooklyn - Bronx == 0          -0.0199713  0.0015948 -12.523   <0.001 ***
Manhattan - Bronx == 0         -0.0420502  0.0015961 -26.346   <0.001 ***
Queens - Bronx == 0            -0.0110008  0.0016931  -6.498   <0.001 ***
Staten Island - Bronx == 0     -0.0063199  0.0034533  -1.830    0.305
Manhattan - Brooklyn == 0      -0.0220789  0.0004314 -51.182   <0.001 ***
Queens - Brooklyn == 0          0.0089705  0.0007160  12.529   <0.001 ***
Staten Island - Brooklyn == 0   0.0136514  0.0030944   4.412   <0.001 ***
Queens - Manhattan == 0         0.0310494  0.0007191  43.179   <0.001 ***
Staten Island - Manhattan == 0  0.0357303  0.0030942  11.547   <0.001 ***
Staten Island - Queens == 0     0.0046809  0.0031468   1.488    0.517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Figure 11

studentized residuals greater than 3. We find the `price` of these observations varies from 1100 to 9999, which is quite unusual for the price of accommodation of one night. We display the five numbers of `minimum_nights` and notice the median is 2 and the 3rd-quantile number is 4 but the maximum is 999 which is obviously a high-leverage data point. Then we calculate the leverage statistic h_i of `minimum_nights` for each observation and consider those observations with $h_i > (p+1)/n$, where $n = 30,632$ and $p = 7$, are high leverage data points. We notice the high leverage `minimum_nights` vary from 43 to 999, which is a quite unusual demand for guests in reality. Thus we tried fitting the model without those outliers and high leverage data points. The coefficients for each predictor are similar with the results shown in Figure 8. The residual plots look similar except the leverage plot gets improved. R^2 increases a little bit. That means even if including these unusual observations, the model built in §3.3 is not significantly skewed.

We did not expect there will be houses with `price` $= 0$. This reminds us to be cautious with the data examination before starting building a model.

The influence of `minimum_nights` on `price` is not obvious intuitively, but from the fitting results this predictor seems significant and relate to the response both from the multiple linear regression and the t-statistic in the simple linear regression we have carried out for a hypothesis test.

A challenge would be an attempt to try the KNN method but there is no function available and the categorical predictors make the process tricky.

Reference

[1] https://nycfuture.org/research/destination-new-york
[2] https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data