Exploration of Factors that Affecting Airbnb Price Using Multiple Linear Regression

Liuyi Pan, 1006211573

December 15, 2021

1 Introduction

Airbnb takes a unique approach to accommodation, offering travelers a home-like B&B rather than a traditional hotel. Since its inception in 2008, it has grown rapidly but has also created a competitive market environment where hosts need to use appropriate pricing to make revenue. Compared to the traditional accommodation industry, the uniqueness of Airbnb makes it more difficult for hosts to set an optimal price. Therefore, the goal of this study is to explore what factors affect Airbnb prices. This can help the hosts to explore the buyers' willingness to pay, and to what extent separately, which benefits the hosts to set the price that both satisfies consumers and maximizes their revenue.

In the published research literature, most of the literature on the analysis of Airbnb price determinants uses ordinary least squares (OLS) and geographically-weighted regression (GWR), with Airbnb's listing characteristics (Chen & Xie, 2017), host features (Chen & Xie, 2017), and customer reviews (Kwok & Xie, 2019) are significant predictors. However, few people have applied the multiple linear regression model to study Airbnb prices with data from the Inside Airbnb website. Therefore, this paper investigates which of the above characteristics "best" explain the variation in the Airbnb price by using the multiple linear regression model.

2 Method

2.1 Data Source

The Toronto Airbnb listings dataset used in this study, which includes 15,155 observations and 74 variables as of November 6, 2021, was downloaded from Inside Airbnb website (Cox, 2018). This study is interested in the variables about:

- listing characteristics (price, type, maximum capacity, number of bathrooms and bedrooms);
- host features (response rate, acceptance rate, whether host is a super host, total number of listings owned, whether host has profile picture and whether host's is verified);
- customer feedbacks (number of reviews and review scores).

2.2 Multiple Linear Regression

The multiple linear regression model not only explains how the mean of the response conditions changes when the values of the predictors change, but also generates predictions. Hence, it was employed to investigate the factors affecting Airbnb prices in Toronto, where the response is price and the initial predictors are the other 12 variables mentioned in 2.1 data source section. The specific mathematical model is as follows:

$$Y = \beta0 + \beta_1X_1 + \beta_2X_2 + \ldots+\beta_iX_i + \epsilon$$

where Y is the Airbnb daily price, X_i is a vector of variables that include listing characteristics, host features and customer feedbacks, ϵ is the random error term.

2.3 Model Violations and Diagnostics

The premise of using linear regression is that these four assumptions need to hold: 1) Linearity relationship should exist, 2) Errors should be uncorrelated, 3) Error variance should be the same, and 4) Errors should be normal. Moreover, each time a new model is fitted, all assumptions need to be examined to proceed to our next step of the analysis. Here, model violations and diagnostics are tested in the following four ways:

Residual Plots

The first method is to use residual plots, which can check whether the first three assumptions are valid or not. We don't want see any discernible pattern in these plots, otherwise, we may consider doing transformation by the use of Box-Cox method, which can also refine the normality.

It is worth mentioning that when using residual plots, we must ensure that the following two conditions hold:

- Condition 1: there is a linear or obvious pattern in the response against fitted value plot.
- Condition 2: there are no non-linear relationships among pairs of predictors.

If they do not hold, we need to consider doing transformations on the problematic variables.

QQ plot

Plotting QQ plot is the second method, which is used to test the normality assumption. Here, we hope to see points is almost fitted on the straight diagonal line with minimal deviations at the ends. As mentioned above, we can use Box-Cox to fix the violation.

Multicollinearity

Multicollinearity will reduce the significance of our parameter estimates, so we want no predictors with variance inflation factor (VIF) is larger than 5.

Problematic Observation

Problematic Observation will also negatively impact the goodness of regression model. We can check as follows:

- leverage points: if leverage is greater than 2(p+1)/n,
- outliers: if standard residual is not between -4 to 4,
- influential points: if cook distance is larger than $F_{0.5}$(p+1, n-p-1).

where n is the observation number, p is the variable number.

If these points are due to contextual reason, we need remove them, otherwise, we should keep it as the limitation.

2.4 Variable Selection

Six techniques will be used for the variable selection.

F, T and Partial F Test

Firstly, F-test, with null hypothesis is none of the predictors are linearly related to the response, is used to see whether a significant linear relationship exists overall. We will perform this test for each model and reject H0 if a p-value is below 0.05, otherwise, this model will not be considered.

Next, we look at individual T-tests to see whether each predictor has a linear relationship in the presence of the other predictors and remove the predictors have p-value that is larger than 0.05. Also, check if we can delete these variables by using the partial F test. If the p-value is larger than 0.05, then we can remove these predictors at once.

AIC and BIC

AIC and BIC are two criterions balancing the goodness of fit of the model. When comparing models, we would prefer the one with lowest AIC and BIC, but always bringing in context as well.

R_{adj}^2

R_{adj}^2 computes the proportion of the variation in the response explained by the predictors. A higher value is better when comparing models, but need to consider over-fit as well.

2.5 Model Validation

For model validation, we perform a 50:50 random split on the whole dataset into training and test sets. The training dataset is used to do model building and diagnostics, while the test dataset is only for evaluating the performance of models. Here are some validations we want to see between two sets:

- Estimated coefficients have minimal differences
- Same predictors appear as significant
- No new model violations are appeared
- Gain similar R_{adj}^2

3 Result

3.1 Data Description

After removing the missing values, a total of 6143 observations and 13 variables were finally obtained and were divided into two sets of data, train and test. As shown in Table 1A and Box Plot 1B in Appendix, we can see that the 9 numerical variables and the 4 categorical variables all have similar distributions between two sets.

3.2 Process of Obtaining Final Model

We first constructed a full linear regression model (model1) with all 12 variables as predictors, but condition 2 of residual plots was not satisfied. As we can see in Figure1 below, most of the variables are clustered on one side, so we applied Box-Cox method to find the appropriate transformation. Hence, we log transformed the price, accommodates, bedrooms, review number and total number of landlord listings, square rooted the bathroom, and applied powers of 2 and 9 to the response rate and review score respectively. This gave us model 2, which satisfied condition 2 (as can be seen in Figure2) and held all other assumptions.

Because the p-value of model 2 F-test was less than 0.05, we further selected the variables by individual T-tests, and then removed 4 unsignificant variables (shaded in Table1A). Also, the p-value of the partial F test was greater than 0.05. Finally, we obtained model 3 and check its assumptions, which are almost valid as shown in the Figure3.

Finally, comparing these three models, we find that they all have small multicollinearity, and model 3 has the largest adjusted R2 and the smallest AIC, BIC (as shown in Table2A). Additionally, compared to model 1, the number of problematic observations for models 2, 3 is greatly reduced (as shown in Table2B), further confirming the decision to choose model 3 as the final model.

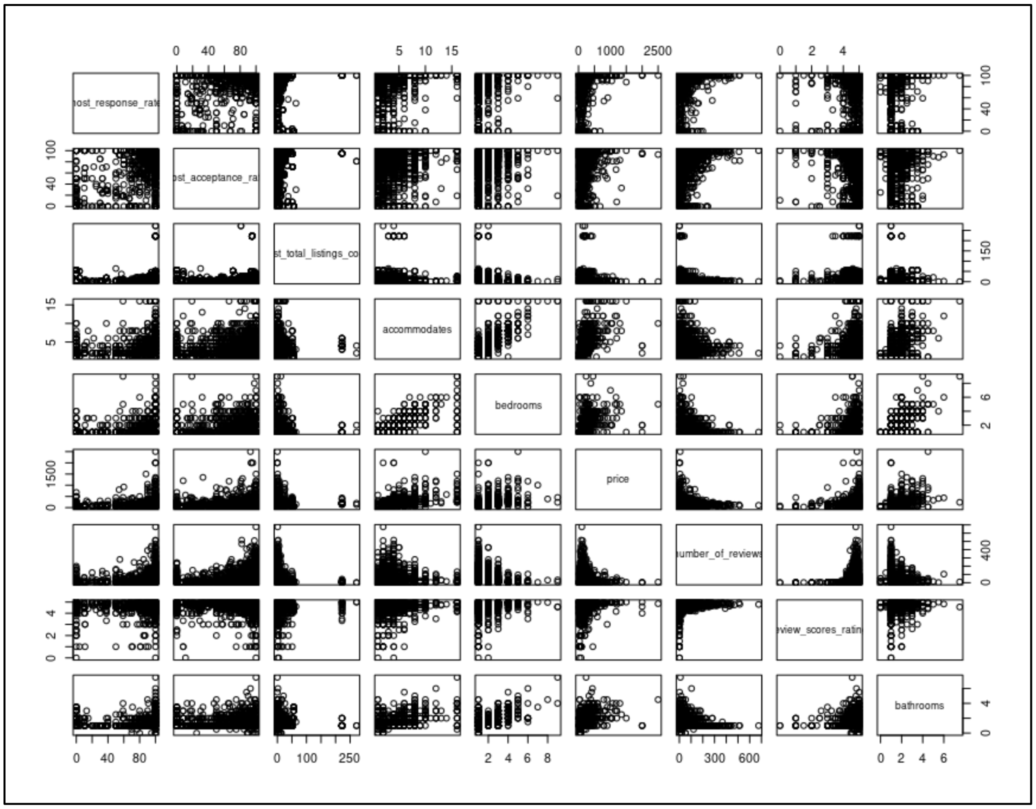


Figure1. Scatterplots of pairs of predictors in Model1

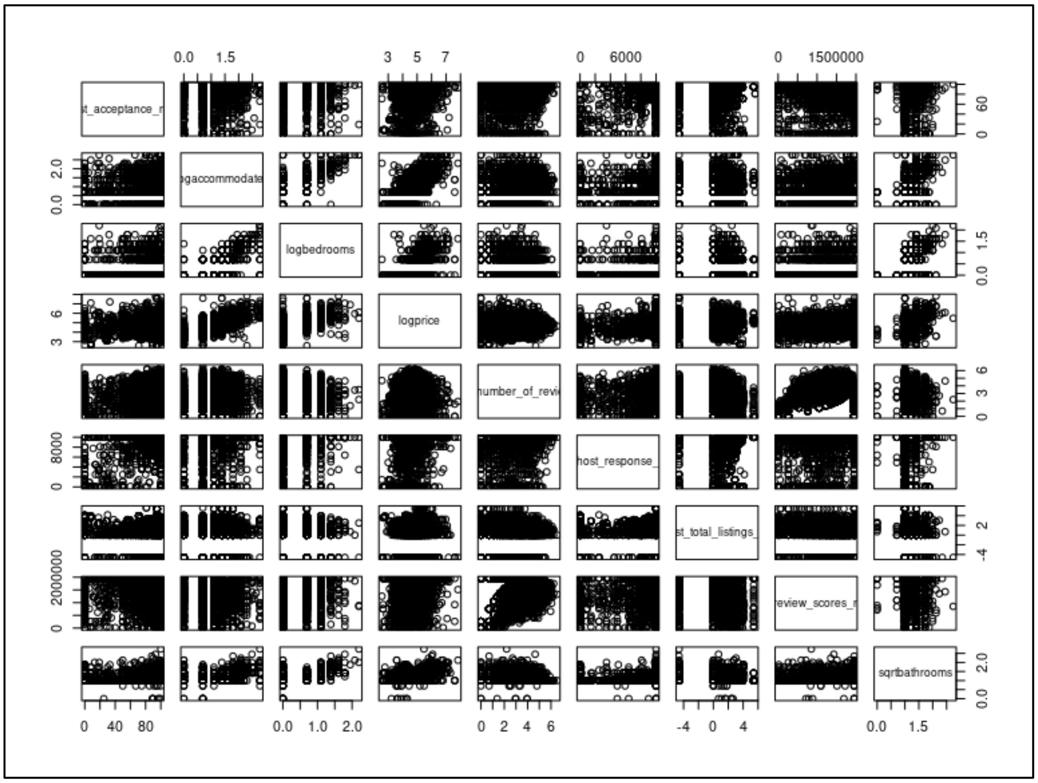


Figure2. Scatterplots of pairs of predictors in Model2

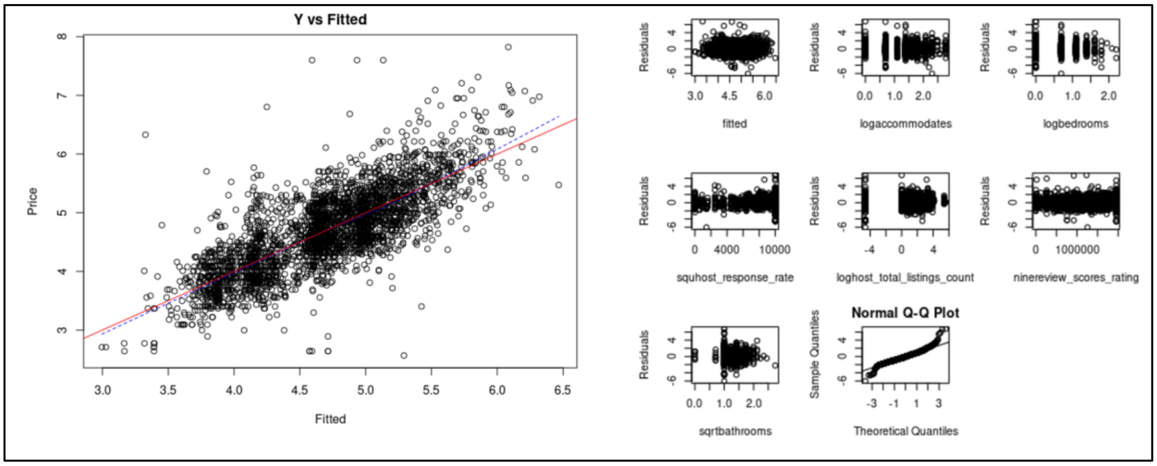


Figure3. Residual Plots and QQ-Plot of Model3

Table1. Model 2&3 Summary

Table A Model 2 Summary

	Estimate	Pvalue
(Intercept)	3.52E+00	< 2e-16
host_acceptance_rate	9.54E-05	7.79E-01
host_is_superhostt	-8.02E-02	2.29E-05
host_has_profile_pict	6.91E-02	6.69E-01
host_identity_verifiedt	1.02E-02	7.48E-01
room_typeHotel room	-1.06E+00	7.00E-05
room_typePrivate room	-4.78E-01	< 2e-16
room_typeShared room	-1.03E+00	< 2e-16
logaccommodates	4.73E-01	< 2e-16
logbedrooms	1.31E-01	2.98E-05
lognumber_of_reviews	-7.54E-03	1.81E-01
squhost_response_rate	2.19E-05	8.09E-09
loghost_total_listings_count	1.02E-02	5.88E-03
ninereview_scores_rating	1.14E-07	8.08E-12
sqrtbathrooms	3.99E-01	< 2e-16

Table B Model 3 Summary

	Estimate	Pvalue
(Intercept)	3.58E+00	< 2e-16
host_is_superhostt	-8.67E-02	1.37E-06
room_typeHotel room	-1.07E+00	6.20E-05
room_typePrivate room	-4.78E-01	< 2e-16
room_typeShared room	-1.02E+00	< 2e-16
logaccommodates	4.70E-01	< 2e-16
logbedrooms	1.35E-01	1.28E-05
squhost_response_rate	2.23E-05	1.86E-11
loghost_total_listings_count	1.02E-02	5.06E-03
ninereview_scores_rating	1.18E-07	6.54E-13
sqrtbathrooms	4.00E-01	< 2e-16

Table2. Evaluations of Three Models

Table A Comparison of Criteria

	R_{adj}2	AIC	BIC
Model 2	0.59591	3862.6966	3959.17798
Model 3	0.59616	3856.7887	3929.14968

Table B Comparison of Problematic Observations

	number of Leverage points	number of ouliers	number of influential points
Model 1	211	24	0
Model 2	126	14	0
Model 3	114	14	0

3.3 Goodness of Final Model

As mentioned in 3.2, the assumptions of model3 all hold, and there is no strong multicollinearity and the number of problematic observations is reduced. Then, we applied model3 to the test dataset and Table1 in Appendix shows that it obtained similar results on training set, indicating that the final model has been validated.

4 Discussion

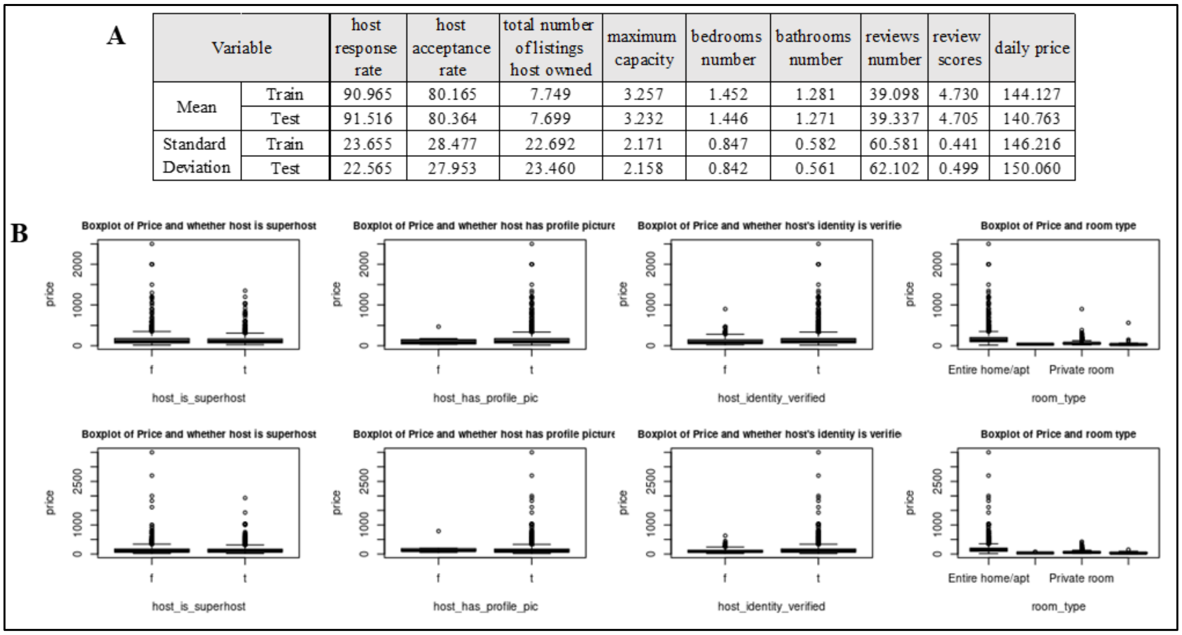
4.1 Interpretation and Importance

We finally get our model containing 8 predictors with some transformations to explain log of Airbnb price, as shown in Table 1B. It indicates that Airbnb prices are mainly related to the intrinsic features, especially the type, accommodates and number of bathrooms. Hosts should carefully check these attributes and then develop appropriate pricing strategies. Furthermore, host effort also plays a role in influencing prices: host attributes, including superhost status, high response rate and more listings, are positively correlated with the price of listings. Lastly, review score is shown to be positively correlated with price as well, suggesting hosts should strive to get better reviews to increase their Airbnb value.

4.2 Limitation

- Although the number of problematic observations in model 3 is reduced, the total number is still large, which impacts the coefficients to have different sign as shown in model 3 summary on test set.
- After the transformation, the constant variance and normality assumptions may still be slightly violated despite the improvement, which could lead to bias.
- This paper does not analyze the categorical variable too much, such as multicollinearity, which may cause bias to the model.

Appendix



Variable		host response rate	host acceptance rate	total number of listings host owned	maximum capacity	bedrooms number	bathrooms number	reviews number	review scores	daily price
Mean	Train	90.965	80.165	7.749	3.257	1.452	1.281	39.098	4.730	144.127
	Test	91.516	80.364	7.699	3.232	1.446	1.271	39.337	4.705	140.763
Standard Deviation	Train	23.655	28.477	22.692	2.171	0.847	0.582	60.581	0.441	146.216
	Test	22.565	27.953	23.460	2.158	0.842	0.561	62.102	0.499	150.060

B
Boxplot of Price and whether host is superhost
price
0
1000
2000
f
t
host_is_superhost
Boxplot of Price and whether host has profile picture
price
0
1000
2000
f
t
host_has_profile_pic
Boxplot of Price and whether host's identity is verifie
price
0
1000
2000
f
t
host_identity_verified
Boxplot of Price and room type
price
0
1000
2000
Entire home/apt
Private room
room_type
Boxplot of Price and whether host is superhost
price
0
1000
2500
f
t
host_is_superhost
Boxplot of Price and whether host has profile picture
price
0
1000
2500
f
t
host_has_profile_pic
Boxplot of Price and whether host's identity is verifie
price
0
1000
2500
f
t
host_identity_verified
Boxplot of Price and room type
price
0
1000
2500
Entire home/apt
Private room
room_type

Figure1. Numerical and Visual Summaries of Variables

Table1. Results of Model3 on Test Dataset

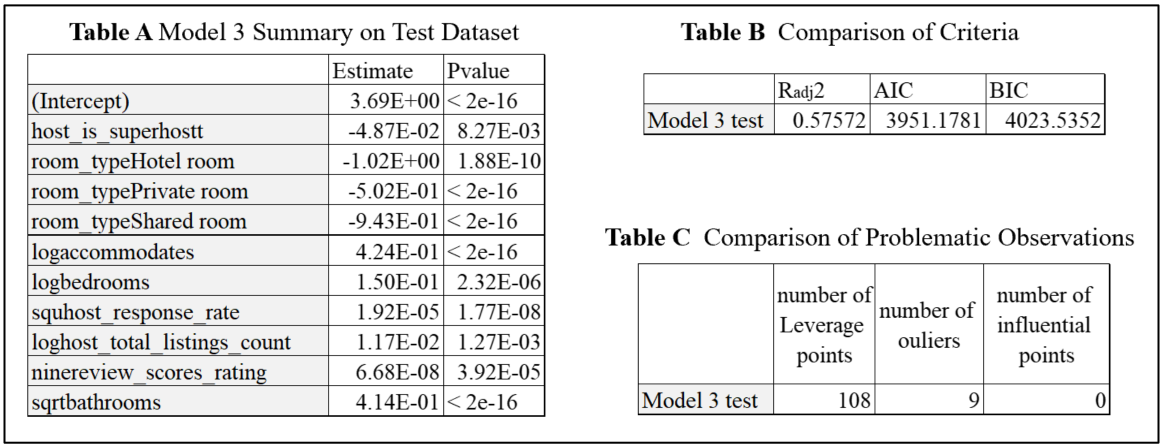


Table A Model 3 Summary on Test Dataset

	Estimate	Pvalue
(Intercept)	3.69E+00	< 2e-16
host_is_superhostt	-4.87E-02	8.27E-03
room_typeHotel room	-1.02E+00	1.88E-10
room_typePrivate room	-5.02E-01	< 2e-16
room_typeShared room	-9.43E-01	< 2e-16
logaccommodates	4.24E-01	< 2e-16
logbedrooms	1.50E-01	2.32E-06
squhost_response_rate	1.92E-05	1.77E-08
loghost_total_listings_count	1.17E-02	1.27E-03
ninereview_scores_rating	6.68E-08	3.92E-05
sqrtbathrooms	4.14E-01	< 2e-16

Table B Comparison of Criteria

	Radj2	AIC	BIC
Model 3 test	0.57572	3951.1781	4023.5352

Table C Comparison of Problematic Observations

	number of Leverage points	number of ouliers	number of influential points
Model 3 test	108	9	0

References

Chen, Y., & Xie, K. (2017). Consumer valuation of Airbnb listings: A hedonic pricing approach. *International journal of contemporary hospitality management*.

Cox, M. (2018). *Inside Airbnb*. Retrieved December 15, 2021, from http://insideairbnb.com/get-the-data.html

Kwok, L., & Xie, K. L. (2019). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros?. *International Journal of Hospitality Management*, *82*, 252-259.