STA302 Course Project Proposal: Prediction of Total Revenue generated by AirBnB Units using Linear Regression

Hanrui Zhang, Ray(Zirui) Liu, Fanshi Lin, Qinjue Jiang, Cruise Chen

May 21, 2025

0 Contribution

All group members made a significant contribution to each part of the project, with the specifics as follows:

1. Data Wrangling and Data Cleaning: Ray Liu
2. Data Analysis and Data Plots: Ray Liu, Hanrui Zhang, Fanshi Lin
3. Residual Analysis: Hanrui Zhang
4. Introduction: Qinjue Jiang, Ray Liu
5. Data Description: Fanshi Lin
6. Preliminary Results: Cruise Chen, Hanrui Zhang

1 Introduction

Airbnb is quickly changing the travel industry with its multi-dimensional characteristics including features like price flexibility, various location choices, and reviews from guests and hosts. With low entry barriers and flexible business structure as a means of extra income, more people have been motivated to list their properties. Under this circumstance, the understanding of determinants of the total revenue has become significantly important for hosts and investors to make business decisions.

There have been some studies on the performance of Airbnb. Wang and Nicolau investigated price strategies and found that listing price are mainly influenced by room type and location [Wang and Nicolau, 2017]. Kwok and Xie also found that multi-unit hosts could have higher revenues compared with single-unit hosts, because they gained more experiences on price setting [Kwok and Xie, 2019]. However, revenue was not only influenced by price but also other explanatory variables. Gunter and Önder studied the elasticity of Airbnb demand and found that the revenue of hosts depended on the host responsiveness and listing capacity [Gunter and Önder, 2017].

Our study integrates these studies and overcomes their limitations in the scope. We aim to study the relationship between total revenue and a set of explanatory variables including host experience, nightly price, customer review scores, listing capacity, host responsiveness, and room type. We expect that the listings with higher price, better scores, more capacity, more private personal space and faster responsiveness would have higher total revenue. To model the relationship between total revenue and explanatory variables, we use multiple linear regression approach. MLR is suitable for our study as we want to study the marginal effect of each explanatory variable on total revenue while holding others constant. We aim to maintain model interpretability by applying only one transformation per variable. However, our objective is to have more precise predictions. With this model, hosts could input fixed listing characteristics and see how the expected total revenue would change if it responded to changes in more controllable variables.

Understanding Airbnb revenue is useful for hosts who would like to make more earnings with their decisions on price, quality of service, and the amenities of the listing. It would also be useful for investors who would like to rent their properties to Airbnb.

2 Data Description

To investigate the total revenue of an Airbnb property, we used the dataset on Airbnb properties in the Netherlands [Cannata, 2017]. Due to its detailed information, we suspected it was sourced from Airbnb logs to predict customer ratings for Airbnb properties. It was sufficient for our purpose of predicting the total revenue of a property to advise optimal business strategies.

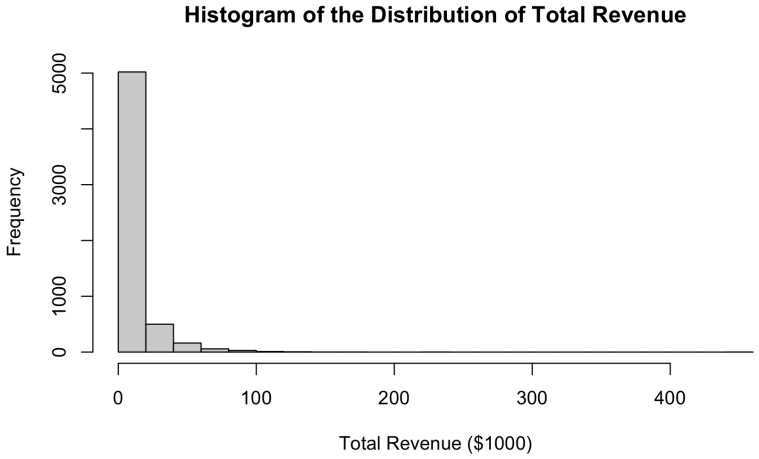

Figure 2.1: Summary of total revenue

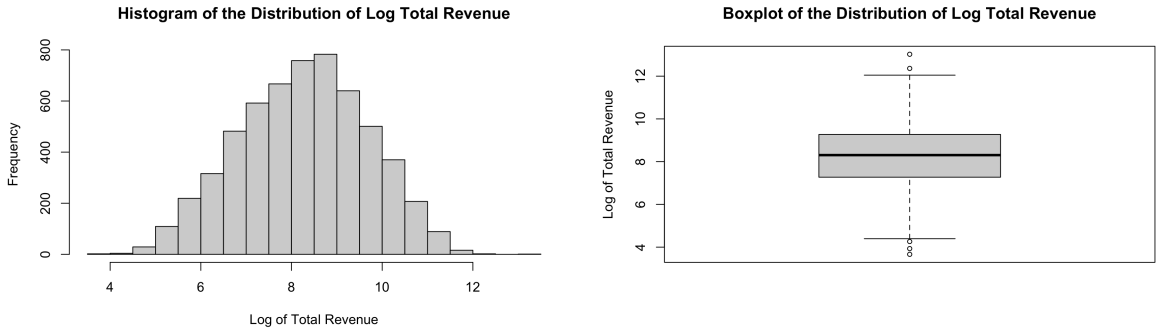

Figure 2.2: Summary of the log of total revenue

The total revenue distribution was right-skewed, with most properties having a total revenue below $100,000 (Figure 2.1). The logarithm of total revenue distribution removed the skewness and became roughly normally distributed (Figure 2.2(a)). Its distribution also exhibits very few outliers (Figure 2.2(b)). As a continuous variable, the log of total revenue showed its potential as a reasonable response for our linear regression model.

Among all variables, we chose five continuous variables that positively interact with total revenue: host age of experience, price, review scores value, number of accommodates, and host response rate (Figure 2.3). After taking the logarithm of total revenue, both review scores value and host response

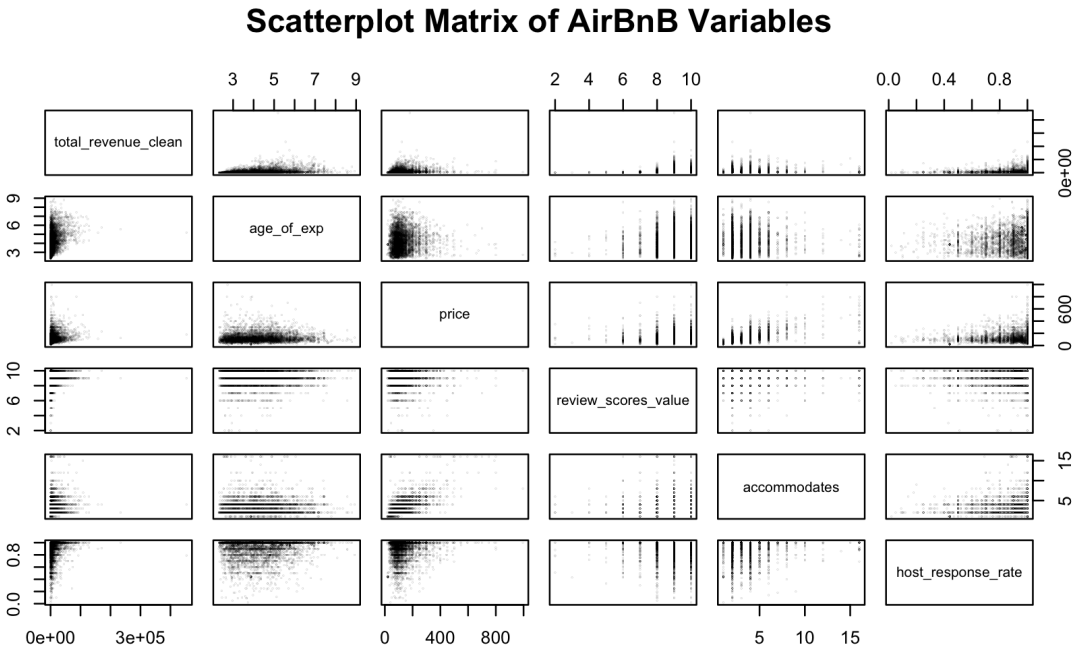

Figure 2.3: Analysis of relations between candidate variables

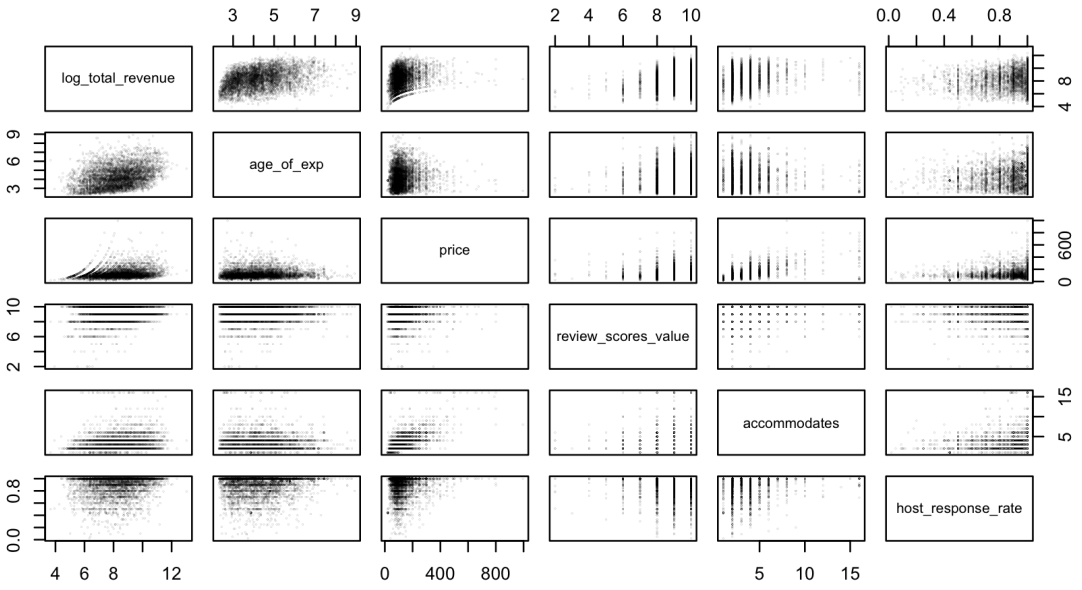

Figure 2.4: Analysis of relations between response and candidate variables

rate were linearized (Figure 2.4). The plots of experience and price against log total revenue were significantly right-skewed and linearized by a logarithmic transformation (Figures 2.4, 2.5). The number of accommodates showed a parabolic shape in its scatterplot versus the log of total revenue, and a square root transformation linearized it (Figures 2.4, 2.5).

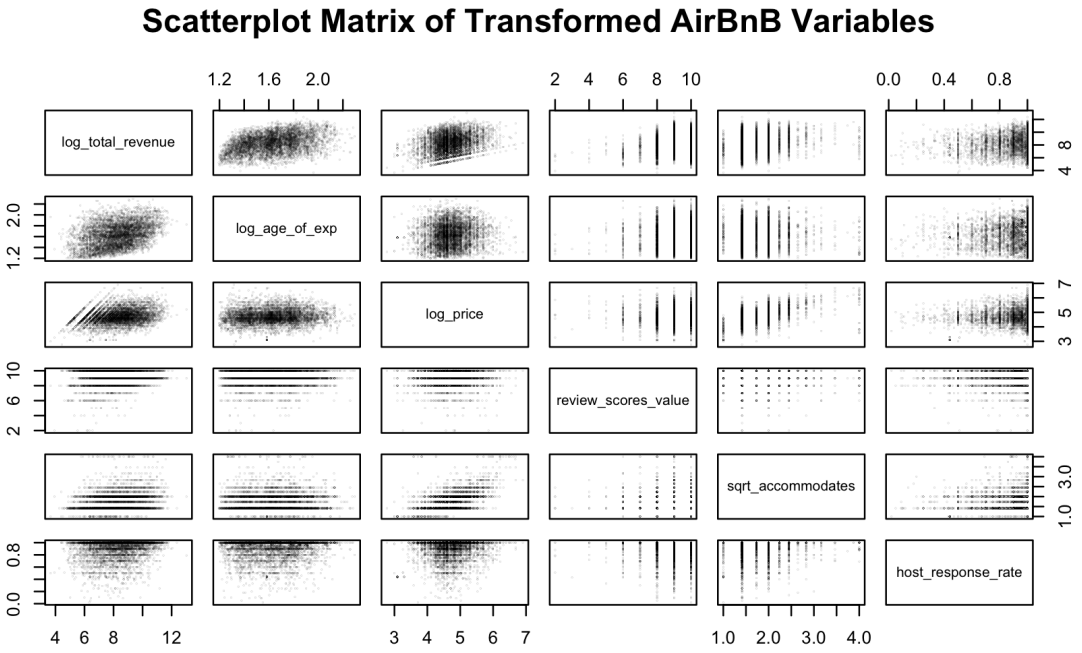

Figure 2.5: Analysis of relations between transformed preliminary predictors and the log of total revenue

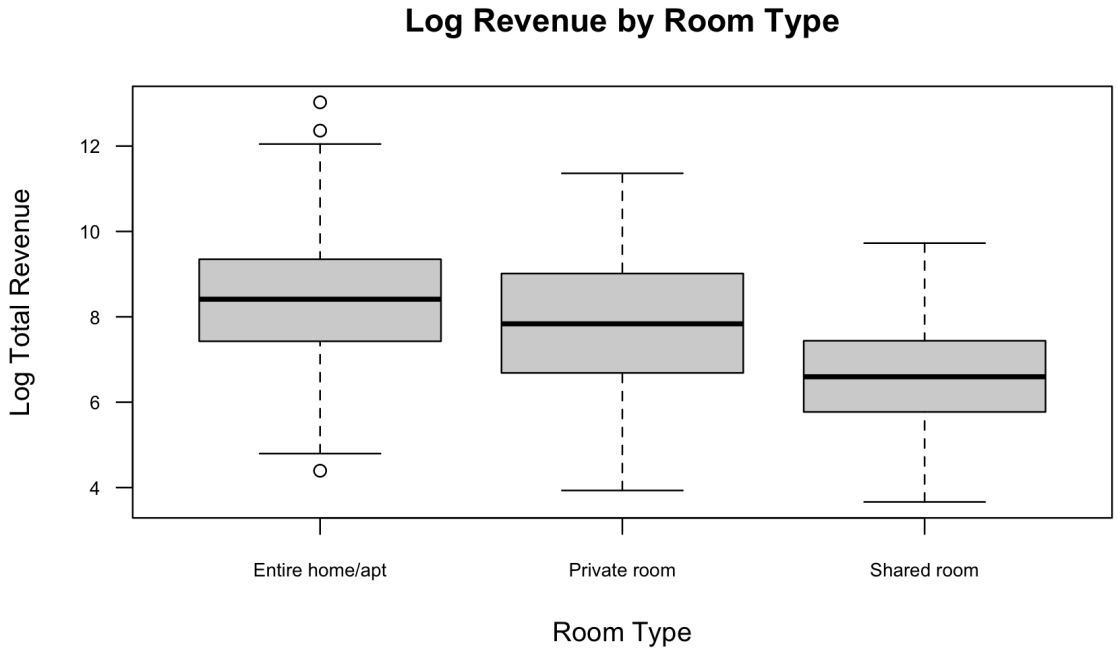

Figure 2.6: Analysis of relations between log revenue and room type.

Analysis of room types showed that the logarithms of total revenues of entire homes or apartments centred at the highest value amongst the three room types, followed by private rooms and shared rooms

(Figure 2.6). Even with slight variations in interquartile range and outliers, the distributions for all categories were reasonably similar with similar spreads(Figure 2.6). This indicated the relationship between a property's room types and the log total revenue. Room type was a reasonable categorical predictor.

3 Preliminary Results

To check the validity of our linearity assumption, we will be using Residual vs Fitted Scatterplot, and Standardized Residual Histogram, QQ plot (Figure 3.1).

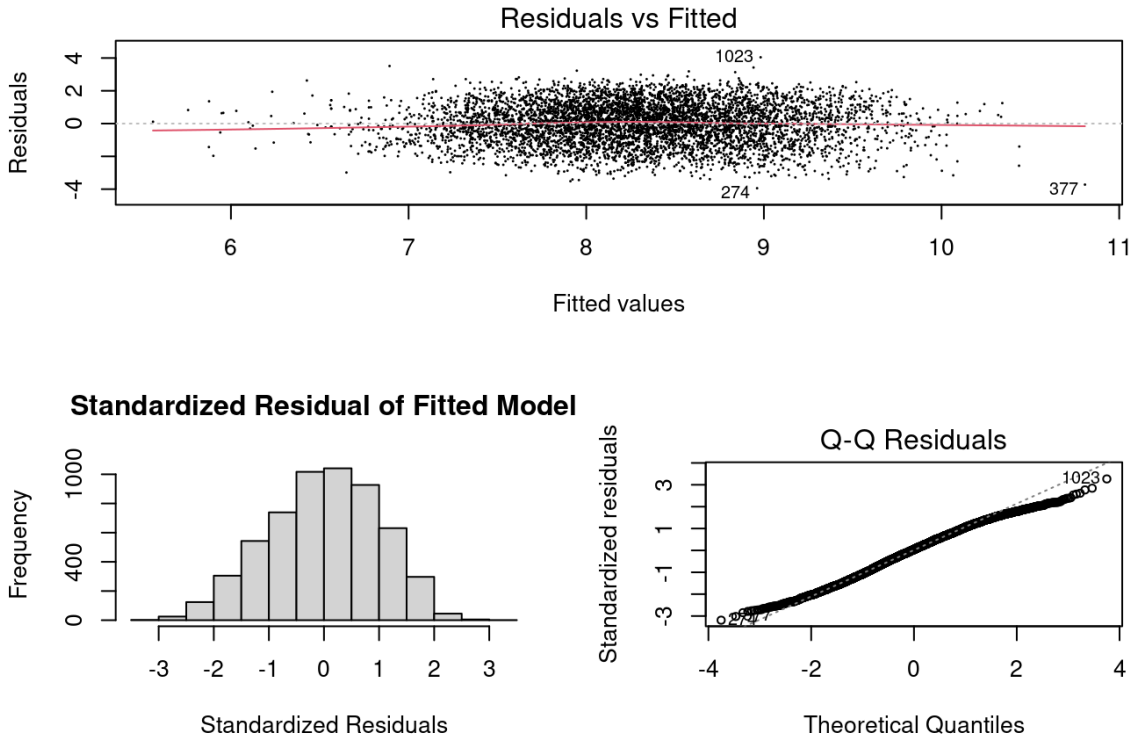

Figure 3.1: Residual vs Fitted Scatterplot, QQ-Plot and Standardized Residual Histogram of our Linear Model

Despite exhibiting a slight U-shape, the Residual vs Fitted model is nearly 0-mean throughout and homoskedastic, meaning that it is close to a null plot; this demonstrates minimal correlation between the residual and the fitted values. Both characteristics suggest that the linear assumption is valid.

Although our QQ-plot suggests that our model might not exactly have a $N(0,1)$ standardized residual with heavy tails, the distribution look approximately like $N(0,1)$, as required by linear assumption.

Therefore, our residual analysis shows that a linear relationship is plausible, with further room for explainability present.

In conclusion, our preliminary fitted model is given by:

$$\begin{aligned}\log(\text{Total Revenue}) = &-1.80055 + 2.31110 \cdot \log(\text{Years of Experience}) + 0.48899 \cdot \log(\text{Price}) \\ &+ 0.10537 \cdot (\text{Review Scores Value}) + 0.32137 \cdot \sqrt{\text{Accommodates}} \\ &+ 0.94504 \cdot (\text{Host Response Rate}) - 0.07089 \cdot (\text{is Private Room}) \\ &- 0.85165 \cdot (\text{is Shared Room}).\end{aligned}$$

This demonstrates that any increase in a continuous predictor variable while holding all other predictors constant, such as Years of Experience as a host, would lead to an increase total revenue. In particular we see that increasing log(Price) by 1 would increase log(Total Revenue) by 0.48899 (all else constant), which demonstrates that AirBnB is demand-inelastic, which agrees with literature [Wang and Nicolau, 2017].

Moreover, by our categorical variable Room Type, we see that leasing out the entire home, private room, and shared room generates revenue in decreasing order. This is expected as entire home and private rooms give more privacy, translating to better homestay experience, thus leading to higher revenue generated.

References

[Cannata, 2017] Cannata, P. E. (2017). Gaairbnb - dataset by cannata.

[Gunter and Önder, 2017] Gunter, U. and Önder, I. (2017). Determinants of airbnb demand in vienna and their implications for the traditional accommodation industry. *Tourism Economics*, 24(3):270–293.

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

[Wang and Nicolau, 2017] Wang, D. and Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on airbnb.com. *International Journal of Hospitality Management*, 62:120–131.