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

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

June 25, 2025

1 Introduction (361 words)

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 (310 words)

To investigate the total revenue of an Airbnb property, we used the dataset on Airbnb properties in the Netherlands [Cannata, 2017]. The details suggested that the data was likely sourced from Airbnb backend 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.

The total revenue distribution was right-skewed, mostly below \$100,000 (Figure 1). Since total revenue was positive, a logarithmic transformation was able to remove its skewness. Log total revenue appeared roughly normal(Figure 2(a)), exhibiting very few outliers (Figure 2(b)). As a continuous variable, the log of total revenue showed its potential as a response for our linear regression model.

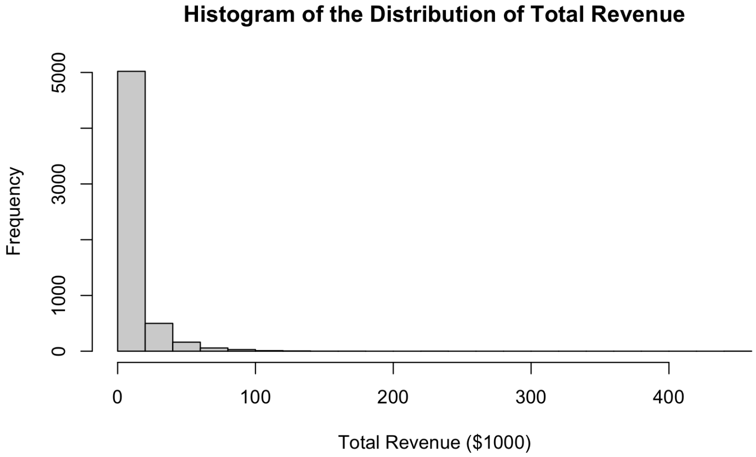

Figure 2.1: Summary of total revenue

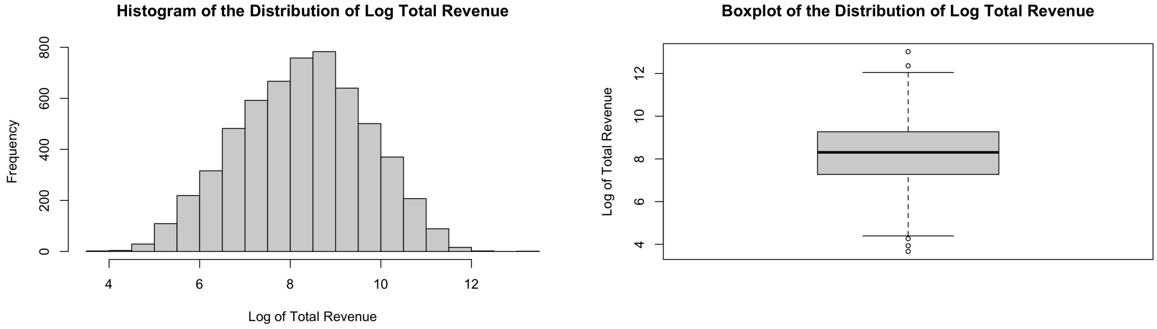

Figure 2.2: Summary of the log of total revenue

We chose five continuous variables that potentially impact total revenue–age of experience of host, review scores value, accommodates and host response rate–and verified their interactions via a scatterplot (Figure 2.3). When checking their relations with log total revenue, both review scores value and host response rate exhibited linearity (Figure 2.4). Age of experience and price against log total revenue plots were right-skewed and were linearized by a logarithmic transformation (Figures 2.4, 2.5). The accommodates versus log total revenue plot showed a parabolic shape, and a square root transformation linearized it (Figures 2.4, 2.5). Since all variables mentioned are positive, the above transformations were legal.

Analysis of another potential factor, room types, showed that the log total revenue of entire homes/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 IQR and outliers, the distributions for all categories were reasonably similar with similar spreads(Figure 2.6). This verified the relationship between room type and log total revenue. Hence, room type was a reasonable categorical predictor.

3 Preliminary Results (301 words)

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).

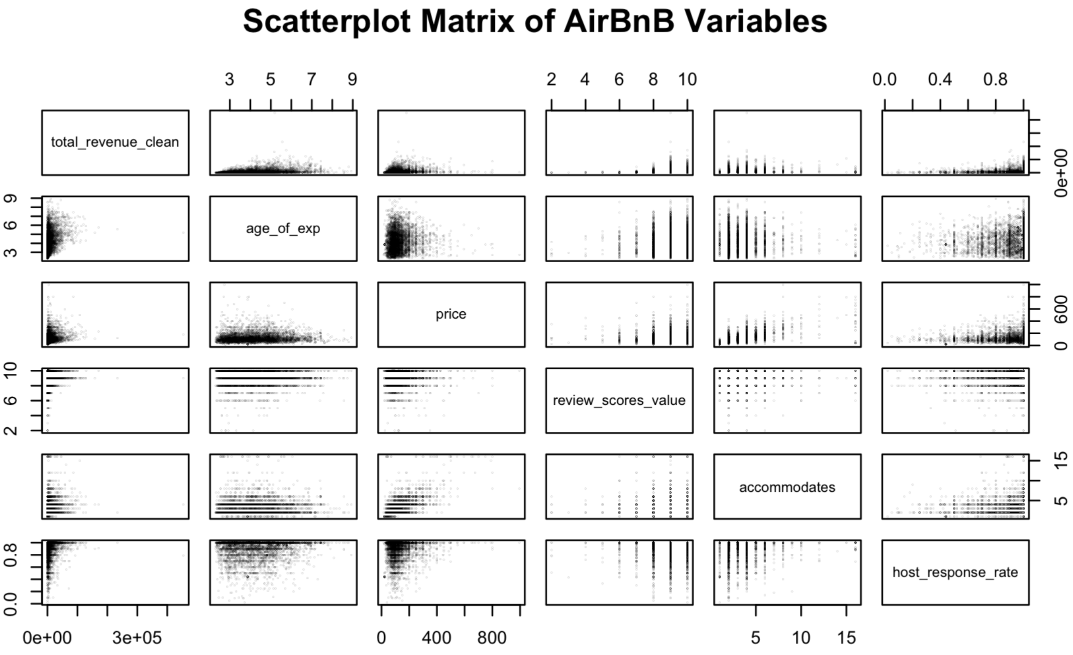

Figure 2.3: Analysis of relations between candidate variables

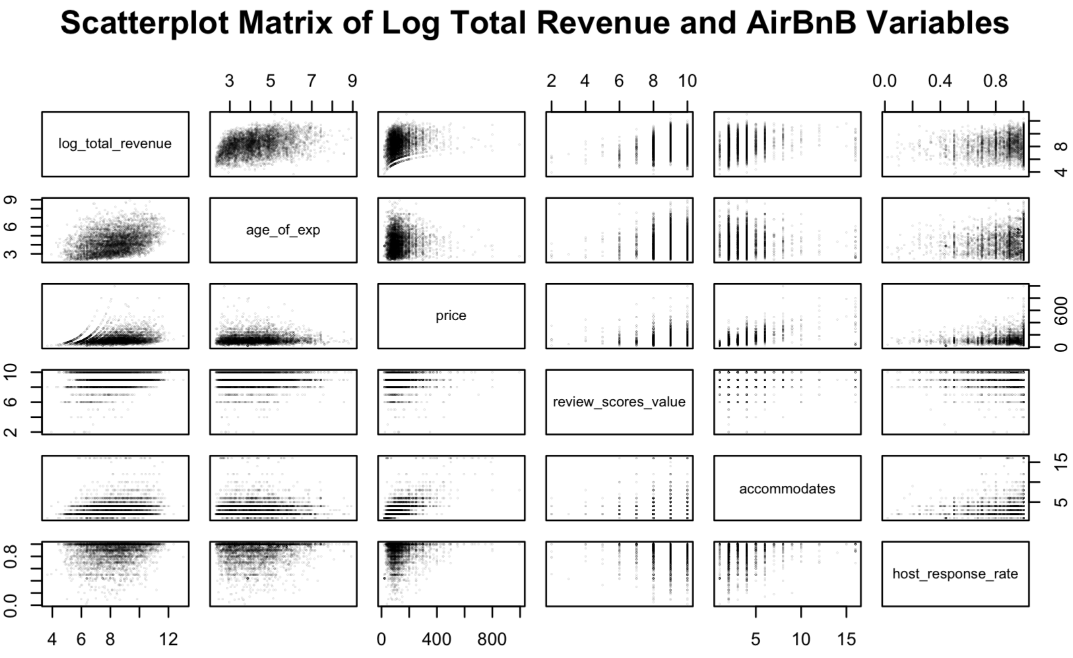

Figure 2.4: Analysis of relations between response and candidate variables

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.

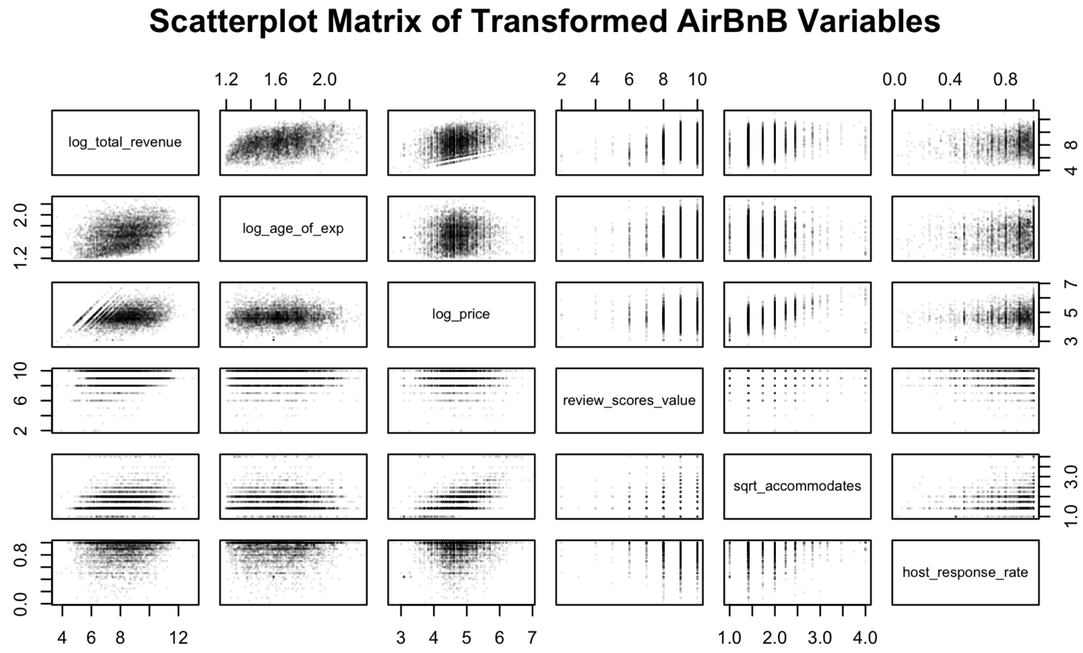

Figure 2.5: Analysis of relations between transformed preliminary predictors and the log of total revenue

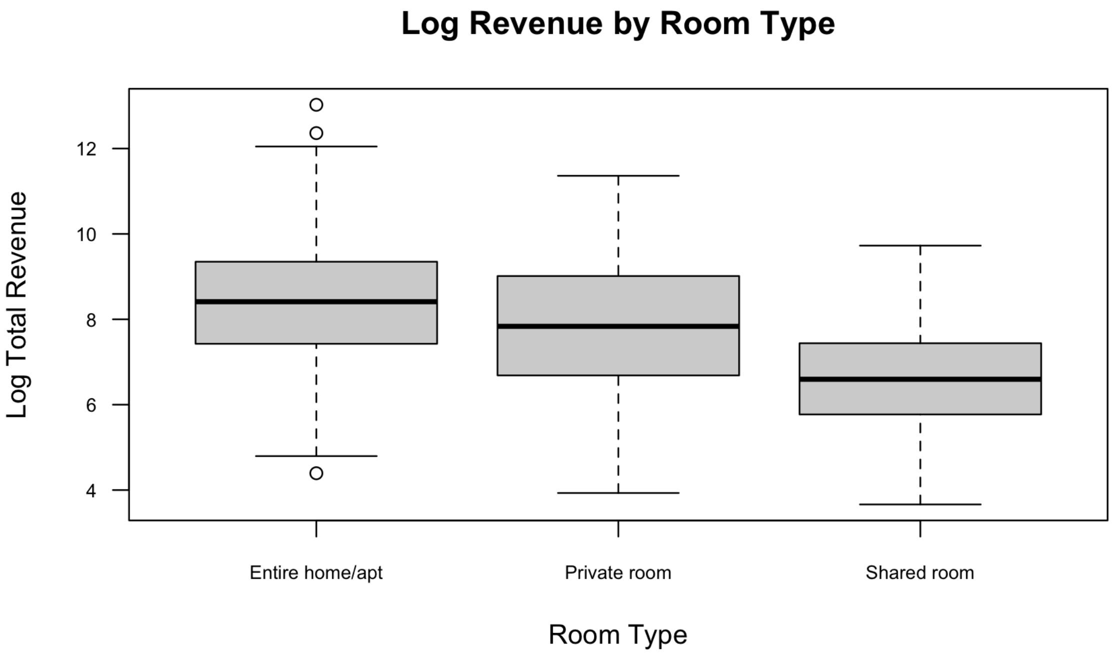

Figure 2.6: Analysis of relations between log revenue and room type.

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.

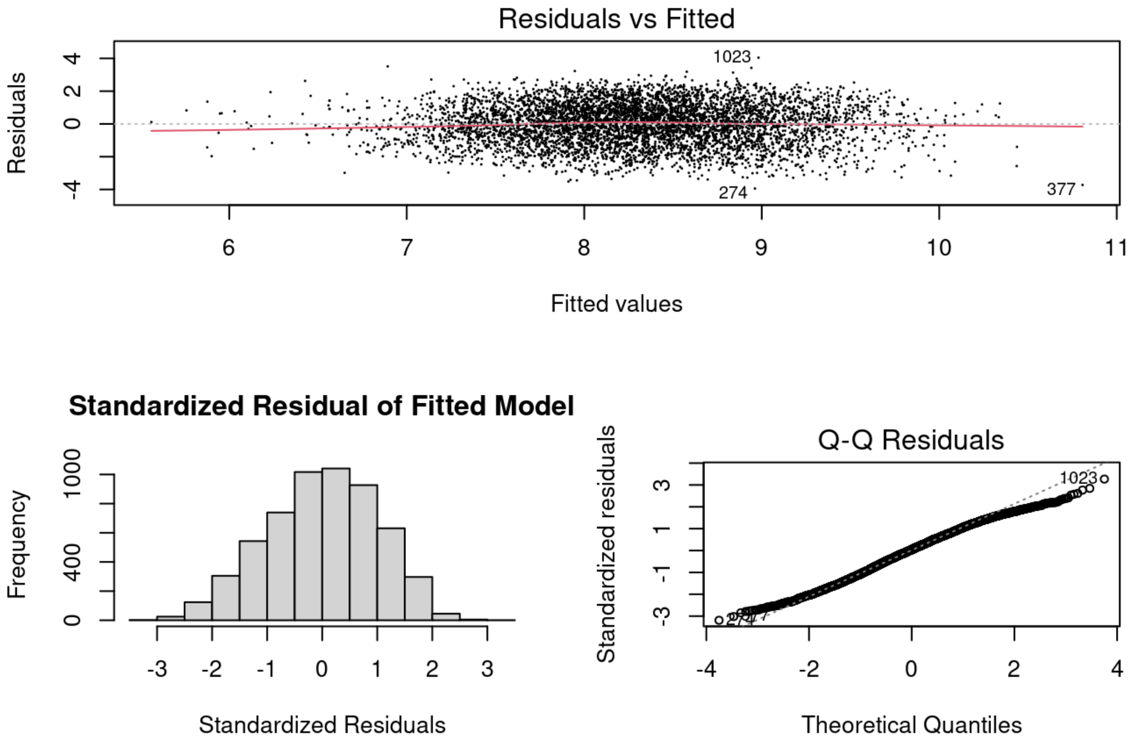

Figure 3.1: Residual vs Fitted Scatterplot, QQ-Plot and Standardized Residual Histogram of our Preliminary Model

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.

4 Model Selection (892 words)

4.1 Transformations

In fitting our multiple linear regression model (MLR), we carefully examined each continuous and categorical predictor to ensure the model meets all assumptions of linear regression and is correctly specified.

Reiterating what we have previously discussed in section 2 in further detail, we introduced several transformations of the predictors, each of which can be justified based on the data and model requirements.

The first transformation we apply is a logarithmic transformation on our response variable total_revenue to reduce its heavy right-skew, as seen in Figure 2.1 and 2.2 before. Although normality of response variable is not assumed in MLR, we still find advantages of doing this. For instance, if the response variable is highly skewed, the residual is likely to be skewed as well, violating an important regression assumption. In such cases, a log transformation can normalize the distribution of residuals and even improve model fit and inference. Furthermore, after comparing the two scatterplot matrices in Figure 2.3 and 2.4 respectively, we observe an increase in linearity and homoskedasticity between predictor and response variables as a result of our transformation, especially in age_of_exp and price.

Moreover, we apply a logarithmic transformation on age_of_exp and price and a square root transformation on accommodates to further improve linearity between these predictors and the response variable, as shown in Figure 2.5.

To further explore the potential of transformations beyond what we've already done in section 2, we tested the Box-Cox transformation on top of our log-transformed response variable. Our expectation is that it can stabilize variance and improve model fit for regression. However, this method yields minimal significant improvements while making interpretation much more difficult. Taking these factors into account, we decided that the transformations in our preliminary model are sufficient.

4.2 Further Inclusions and Exclusions of Predictors

To ensure that our model fully utilizes the information in our given dataset while maintaining simplicity, we performed an analysis on the performance of our model after potential inclusions and exclusions of predictor variables using partial F-tests.

One potential candidate for predictor exclusion was the categorical variable room_type (3 categories), as it was the only one that contributed a predictor variable with a large coefficient p-value of 0.112. However, our partial F-test between our original model ($p + 1 = 8$) and reduced model ($p + 1 = 6$) gives a p-value of 0.000189, which is significantly smaller than 0.05, providing evidence against removing room_type. This leads us to deciding to keep the predictor in our final model.

We also have a categorical variable potential candidate to include as a predictor in our final model: The response time of a host (host_response_time). This potential predictor contains 4 categories: {within an hour, within a few hours, within a day, a few days or more}. As shown in Figure 4.1, our box-plot indicates a potential pattern with our response variable, but seemed to be minute (this was the primary reason for its exclusion in our preliminary model). However, our partial F-test indicates that adding the new category to our model ($p + 1 = 11$) compared to our original model ($p + 1 = 8$) yields a p-value of $9.605 \cdot 10^{-12}$, which is also significantly smaller than 0.05 providing evidence for adding host_response_time.

Furthermore, a 10-fold Cross Validation tells us our new model sees an improvement in bias-corrected estimate by 0.01255771; our new model has a VIF of less than 2 for each predictor variable, telling us that multicollinearity is not a problem. This leads us to deciding to add the categorical variable host_response_time as 3 indicator predictors to our model.

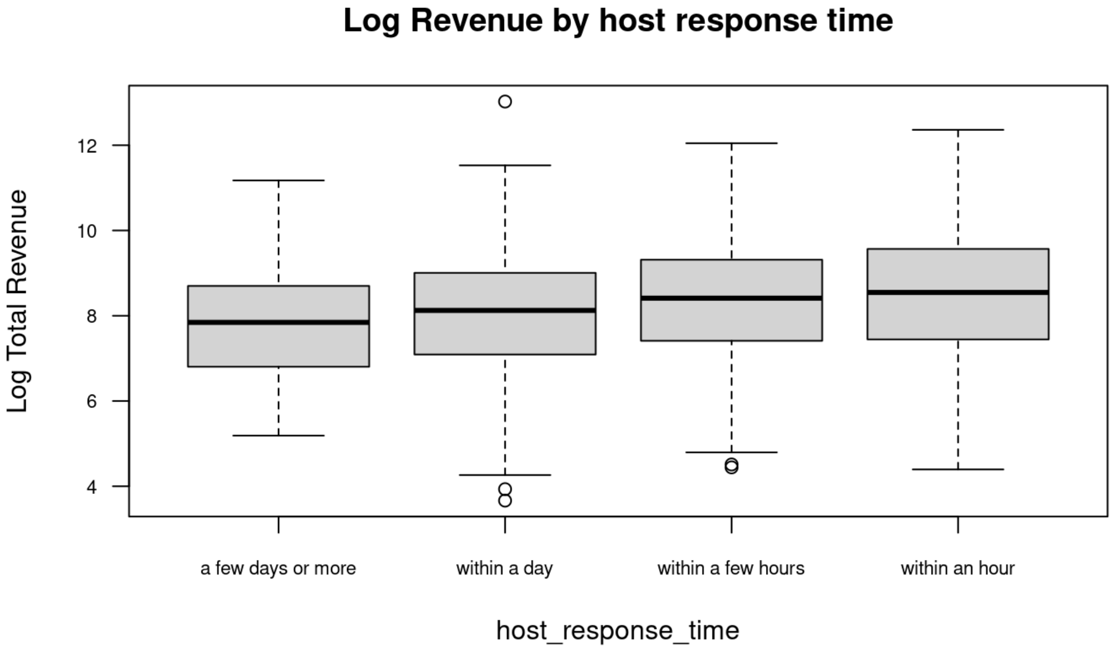

Figure 4.1: Box-plot of Log Total Revenue by Host Response Time

A quick check of the residual plots of our new model shows that the conclusions reached about our preliminary model regarding assumptions of linear regression holds for our new model as well, as shown in Figure 4.2. Thus, we will finalize our model's predictor variables as the same as our preliminary model and including host_response_time.

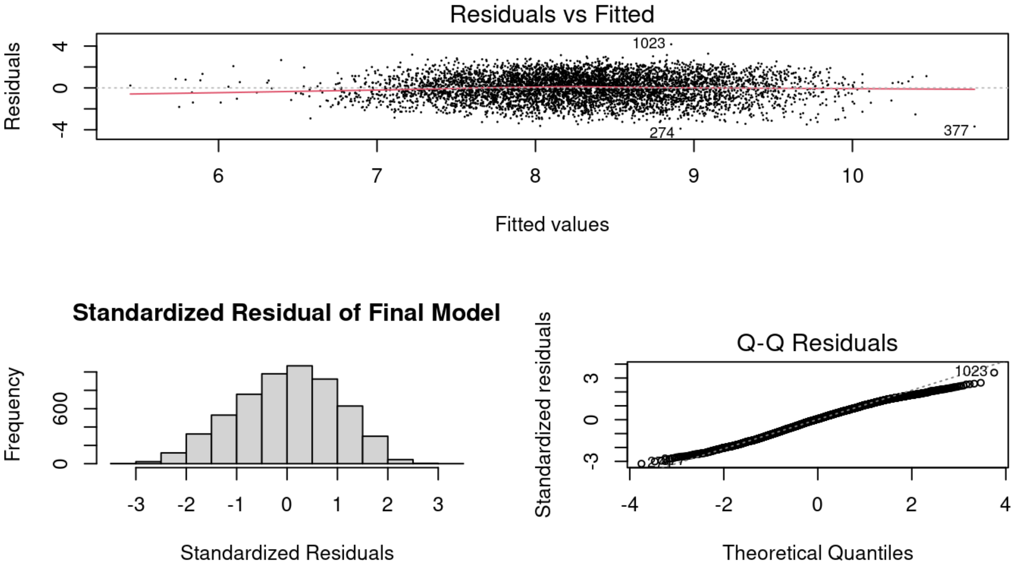

Figure 4.2: Residual vs Fitted Scatterplot, QQ-Plot and Standardized Residual Histogram of our Final Linear Model

4.3 Problematic Observations

To examine the potential existence of problematic observations in our data, we ran an analysis on our finalized model to find outliers and influential points, leading us to decide to keep all observations in our data.

For outliers, we use the criterion $|r_i| > 4$, where r_i is the standardized residual of the i-th observation, to determine an outlier. This is the standard convention for datasets with more than 50 observations (our dataset contains 5702 observations). As shown in Figure 4.3, all points have a standard residual between 4 and -4, indicating that no outliers are present. This leads to no motivation for finding leverage points as any such points are considered beneficial for the model since they are not outliers.

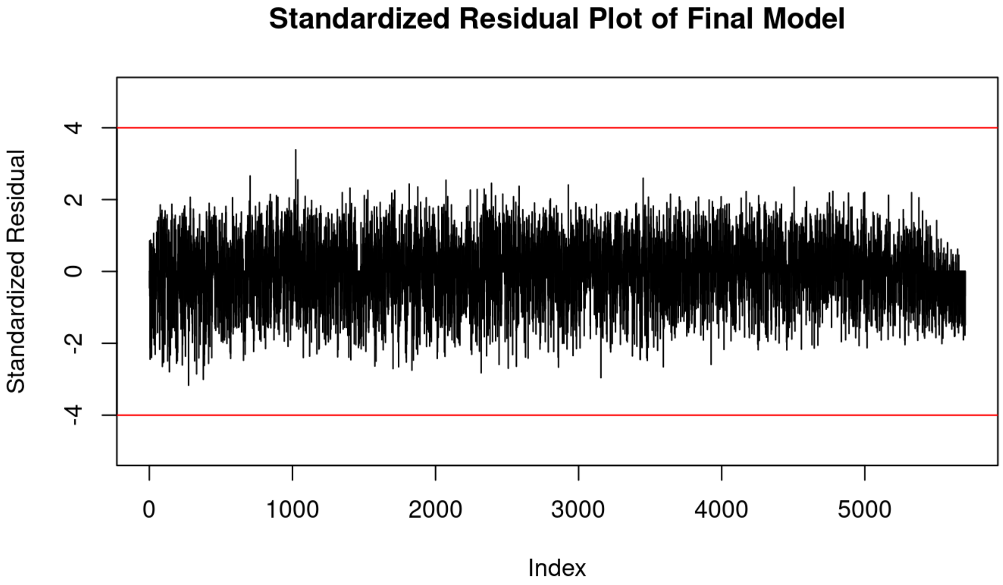

Figure 4.3: Standardized Residual Plot of all observations with respect to our Final Model.

For influential points, we use the Cook's distance D_i as an indicator, with the criterion

$$D_i > F_{0.5}(11, 5702 - 11) = 0.9402015$$

, where $F_q(m, n)$ is the cumulative distribution function for the F-distribution at quantile q with degree of freedom m and n respectively. As observed in Figure 4.4, the Cook's distance of each observation is clearly below our defined threshold, indicating that no highly influential points are present.

Combining our results above, we have no outliers nor influential points, indicating no problematic observations in our data. Thus any removal of data will be unjustified.

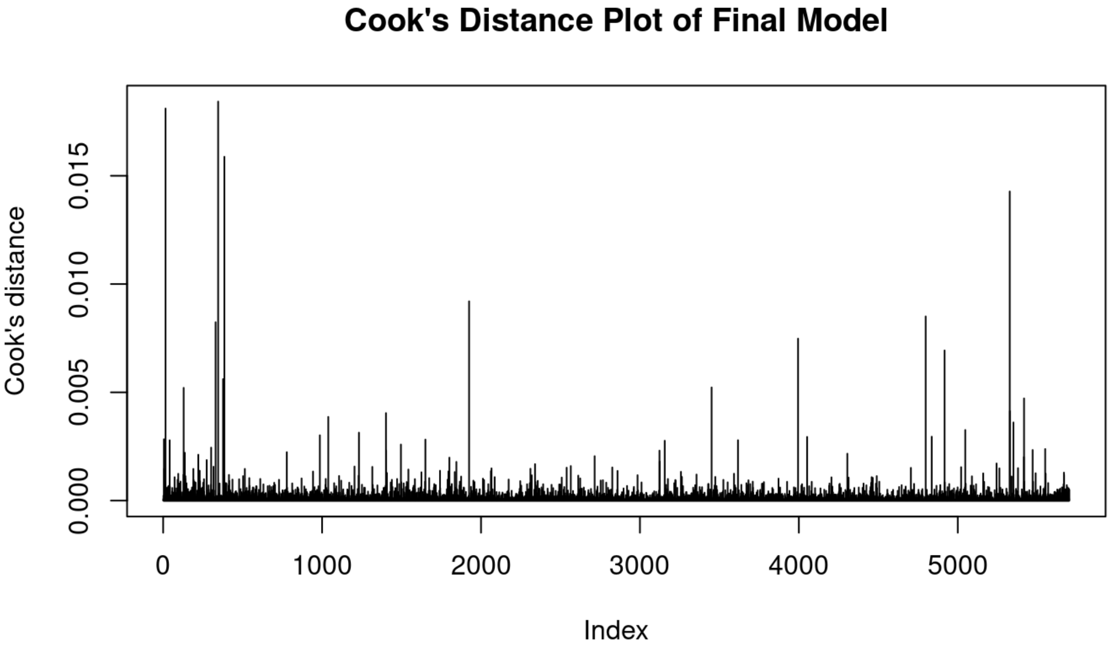

Figure 4.4: Cook's Distance Plot of all observations with respect to our Final Model.

5 Final Model Inference and Results (938 words)

5.1 Final Model

Our final model provides an estimate of the relations between log of total revenue and 7 predictors. Among the 7 predictors, 5-log of experience age, log of price, value of review scores, square root of accommodations, host response rate-are continuous, and 2-room type and host response time-are categorical (Table 1).

5.2 Coefficients Interpretation

5.2.1 Intercept

The intercept of this model is 0.24367, suggesting that an entire home or apartment with host response time of a few days or more with all continuous variables being 0 has an estimate of expected log of total revenue of 0.24367. However, this intercept lacks practical meaning, as all continuous variables are positive in real life.

5.2.2 Continuous Predictors

The value of all 5 continuous predictors are positively associated with log of total revenue according to their coefficients, meaning that the increase in each continuous predictor results in a release in log of total revenue. An increase of 1 unit in log age of experience, log price, review scores value, square root of accommodates and host response rate will result in an increase of 2.31945, 0.49851, 0.11014, 0.30480 and 0.74058 in our estimation of average log of total revenue, respectively. All coefficients of the continuous variables have a p-value less than 0.05, suggesting that we reject null hypothesis for each coefficient at 95% significance level. Alternatively, the 95% confidence interval also supports this observation. We are 95% confident that the coefficients fall in such intervals greater than 0, confirming

Table 1: Coefficients, standard errors, confidence intervals, p-values for final model

	Estimate	Std. Error	t value	$\Pr(>\|t\|)$	95% CI
(Intercept)	0.24367	0.30	0.818	0.41327	$[0.34, 0.83]$
log(age of experience)	2.31945	0.08	29.438	$<$ 2e$-$16	$[2.16, 2.47]$
log(price)	0.49851	0.04	10.995	$<$ 2e$-$16	$[0.41, 0.59]$
review scores value	0.11014	0.02	5.760	8.83e$-$9	$[0.07, 0.15]$
$\sqrt{\text{accommodates}}$	0.30480	0.05	6.459	1.14e$-$10	$[0.21, 0.40]$
host response rate	0.74058	0.16	4.469	8.02e$-$6	$[0.42, 1.06]$
(room type) private room	-0.10541	0.048	-2.191	0.02847	$[-0.20, -0.01]$
(room type) shared room	-0.86998	0.22	-3.979	7.00e$-$05	$[-1.30, -0.44]$
(host response time) within a day	-0.39293	0.14	-2.891	0.00386	$[-0.66, -0.13]$
(host response time) within a few hours	-0.25007	0.15	-1.680	0.09298	$[-0.54, 0.04]$
(host response time) within an hour	-0.08290	0.15	-0.541	0.58820	$[-0.38, 0.22]$

the positive relations between these continuous predictors and log of total revenue.

5.2.3 Categorical Predictors

Room type also affects log of total revenue. The model shows that holding all other predictors constant, changing from renting out an entire home or apartment to a private room causes a decrease of 0.10541 in log of total revenue, and changing to renting out a shared room causes a decrease of 0.86998. Both coefficients have a p-value lof of less than 0.05, suggesting that we reject null hypothesis for both coefficients at 95% significant level. We are 95% confidence that the true coefficient for private room falls in [-0.20, -0.01], and the true coefficient for shared room falls in [-1.30, -0.44], confirming the negative impact on log of total revenue when switching from entire home or apartment to these two room types.

The model also reveals relations between host response time and log of total revenue. Holding all other predictors constant, host response time changing from a few days or more to within a day causes a decrease of -0.39293 in estimated average of log of total revenue. Its p-value of 0.00386 suggests that the coefficient is at 95% significance level, and we are 95% confident that the true coefficient lies between -0.65942 and -0.12645. Changing from a few days or more to a few hours and to within an hour have less significant negative impact. The estimated average of decrease in log of total revenue is 0.25007 when changing to a few hours, and 0.08290 when changing to within an hour. The coefficient of a few hours has a significance level of 90%, while that of within an hour has a significance level of less than 50%. The impact of host response time on log of total revenue is less reliable statistically, especially the response time of within an hour. This is also reflected by the 95% confidence interval–we are 95% confident that the true coefficient of a few hours falls in [-0.54, 0.04] and that of less than an hour falls in [-0.38, 0.22]. Since 0 in both interval, it is likely that these two response time don't have a significant impact on log of total revenue. However, since we observed am association between host response time and log of total revenue from the boxplots (Figure x.x), and the estimate of coefficients align with the observation, where host response time causes a relatively higher estimate in average of log of total revenue, and response time within the day causes the largest decrease, we keep this categorical predictor to explain more differences in our estimation.

5.3 Model Performance assessment

5.3.1 R^2 Analysis

According to the R^2 score, the model explains 22.4% of the differences in log of total revenue. The remaining 77.6% might stem from remaining variables or random noise. Previous VIF diagnostics suggest that there is no multicollinearity between our current predictors, and hence one way to improve this model could be screening through more predictors to explain the differences in our response.

Table 2: Comparison between final model and preliminary model

Model	Adjusted R^2	AIC	BIC
Preliminary Model	21.56%	18621	18680
Final Model	22.26%	18572	18652
Differences	0.70%	-49	-28

However, due to the nature of social science research, an R^2 of 22.4% with most predictors being significant is an acceptable range for this model [Ozili, 2022].

5.3.2 Model Improvement (Adjusted R^2, AIC, BIC)

Comparing to our preliminary model, our final model slightly improves in its performance (Table 2).

It improves adjusted R^2 by 0.007, suggesting that the final model explains slightly more differences in our response, log of total revenue. With a sample size of 5702, we further assess model performance by Akaike Information Criterion (AIC) and Beysian Information Criterion (BIC). A decrease of 48.41689 in AIC suggests that our final model is a better fit than our preliminary model with current complexity level, and a decrease of 28.47117 in BIC suggests that favouring simpler model, our final model with one more predictor still presents a better fit. All these three criteria combined suggest that though very slightly, our final model presents a better fit than the preliminary model. Including the host response time in our final model improves our model performance in general.

6 Discussion and Conclusion (719 words)

6.1 Final Model and Coefficients:

$$\begin{aligned}\log(\text{Total Revenue}) = {} & 0.24367 + 2.31945 \cdot \log(\text{Age of Experience}) + 0.29852 \cdot \log(\text{Price}) \\ & + 0.11014 \cdot (\text{Review Score}) + 0.30480 \cdot \sqrt{\text{Accommodate}} \\ & + 0.74058 \cdot (\text{Host Response Rate}) - 0.10541 \cdot I(\text{Room Type: Private Room}) \\ & - 0.86998 \cdot I(\text{Room Type: Is Shared Room}) - 0.39293 \cdot I(\text{Response Time: within a day}) \\ & - 0.25007 \cdot I(\text{Response Time: within a few hours}) \\ & - 0.08290 \cdot I(\text{Response Time: within an hour})\end{aligned}$$

6.2 Conclusions and Key Results

This study aims to identify dominant factors that influence total revenue for Airbnb using multiple linear regression models. The final model incorporated five continuous variables: host business experience, price, customer review scores, accommodation capacity, and host responding speed to be the key predictors. Among all these predictors, year of experience and host response rate seem to be the most influential variables. The coefficient of price reflects the price elasticity of revenue. With an estimated positive coefficient, it suggests that Airbnb revenue is price inelastic and implies that modest price increases may improve earnings [Gunter and Önder, 2017].

Besides, the model involves two categorical variables: room types and host response time. We find that private rooms and the entire home structure do not show a significant difference in the impact on total revenue. In contrast, shared rooms are associated with a substantial reduction in earnings, suggesting guests place a premium on privacy. Also longer the responding time reflects the more negative impact on total revenue.

The final linear regression model presents multiple R^2 equal to 0.224 and adjusted R^2 equal to 0.2226, meaning the model approximately explains 22% of the variance in log-transformed total revenue. The statistical outcome suggests our final model comparably captures meaningful and strong relationships while maintaining the simplicity of the model. Although the figure may not appear to be significant, it is still reasonable in a real-world context, where a large proportion of variation is due to external factors, such as seasonal pattern, regional policy, holiday date, etc. [Ozili, 2022].

6.3 Suggestions and Recommendations to Current and Prospecting AirBnB Hosts

Our finding offers several practical insights for Airbnb hosts and prospective property investors. Firstly, our model identifies both response time and rate as strong predictors of total revenue. The host should respond to customers as quickly and consistently as possible to maintain substantial income. For example, an automatic reply system is encouraged. The hosting experience is another key determinant, meaning the longer a host has been active in doing Airbnb business, the more revenue they tend to generate, as they tend to be more reliable and trustworthy. Therefore, we suggest that new hosts should learn from more experienced business owners and try their best to maintain consistent quality to construct a long-term performance.

Moreover, the model suggests that higher review scores tend to increase revenue, despite the influence not being as significant as other factors. The host should prioritize guest experience, offer positive communication, maintain cleanliness, and make sure that all the facilities operate normally. Price strategy is important when making business decisions. The corresponding coefficient is modest at around 0.3, reflecting that revenue will increase with price, but at the same time, the quality of service and room facilities that match the price are of vital importance. Lastly, the host should consider the room type carefully. A shared room tends to generate significantly less income. If feasible, offering private rooms or entire homes may substantially boost earnings.

6.4 Limitations of Analysis and Future Improvements

Although the model is statistically robust and interpretable, several limitations still exist. Our analysis is based on Airbnb listings in the Netherlands, which may limit generalizability to other regions with different tourist dynamics or regulations. Therefore, further improvement could expand the dataset to include multiple countries.

Another limitation is that we did not apply automated variable selection techniques during initial model development, such as best subset selection, forward selection, or backward elimination. Instead, we manually selected predictors, and only later used best subset selection to evaluate the final set of predictors. Despite our final model performing well and being supported by statistical metrics, a more rigorous approach would involve automated selection methods at the beginning, which might reduce subjective bias through selecting variables based on data-driven criteria. Automated selection does not necessarily lead to a global optimum, but might be beneficial.

Moreover, we did not split the original dataset into training and testing subsets before performing model selection. Ideally, a portion of the data should be reserved for testing, so that the model can be trained on one subset and then validated on unseen data; this would allow for checking for overfitting and underfitting of models. In future studies, implementing a formal train-test would help assess how well the model generalizes beyond the training sample and guarantee the model's validity.

Despite these limitations, our analysis provides a valuable starting point for future Airbnb revenue prediction.

7 Contributions and Remarks

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

1. Hanrui Zhang: R Markdown Contribution: Data Analysis and Data Plots; Report Contribution: Preliminary Results, Transformations; LaTeX Compilation.

2. Ray Liu: R Markdown Contribution: Data Wrangling and Data Processing, Data Analysis and Data Plots; Report Contribution: Introduction.

3. Fanshi Lin: R Markdown Contribution: Data Plots; Report Contribution: Data Description, Final Model Inference and Results.

4. Donna Jiang: Report Contribution: Introduction, Discussion and Conclusion.

5. Cruise Chen: Report Contribution: Preliminary Results, Transformation.

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.

[Ozili, 2022] Ozili, P. K. (2022). The acceptable r-square in empirical modelling for social science research. *SSRN Electron. J.*

[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.