Final Group Project:
How do different listing characteristics (room type, number of reviews, and availability) influence Airbnb prices in New York City?

STAT 306 201
G15: Vaibhav Hari, Palak Mahendru, Yirui Wang, Perry Zhu
Kenny Chiu
April 6, 2025

1. Introduction

1.1 The data

The dataset we chose to analyze is the New York City Airbnb Open Data, obtained from Kaggle's public dataset repository. This dataset provides a comprehensive snapshot of Airbnb listings across the five boroughs of New York City: Manhattan, Brooklyn, Queens, the Bronx, and Staten Island, as of 2019. Dgomonov published it and includes various features related to rental listings, hosts, and property characteristics.

The data includes 48,895 listings and 16 variables, such as listing ID, name, host ID, etc. The variables we will be exploring include:

- Response Variable (Y):
- Price (US $): The listing price in US dollars.
- Explanatory Variables (X):
- Room Type (Categorical: Entire home/apt, private room, shared room): The type of Airbnb listing.
- Number of Reviews (Continuous): The total number of reviews the listing has received.
- Availability (days per year) (Continuous): The number of days the listing is available for booking in a year.

1.2 Motivation & Research Question

All data points are publicly available and reflect information from Airbnb's listings in 2019. The dataset is suitable for exploring pricing determinants and market behaviors in short-term rental platforms. The study draws its motivation from Yong and Karen (2018) who used a hedonic pricing model on Airbnb listings to determine that price is affected by location, host reputation and listing characteristics. The authors stress that these factors are crucial for understanding how customers value products.

New York, one of the largest cities in the world, has exceptionally high housing prices. Many tourists visit NYC and look for places to stay. Airbnb listings have gained popularity in recent years because they are usually more affordable than hotels. The New York Airbnb dataset can help us understand how room type, number of reviews, and availability influence pricing. This analysis can assist hosts in optimizing their pricing strategies and provide insights into market trends and demand patterns across different neighborhoods. It can also assist tourists in making better decisions by balancing cost and quality when selecting accommodations.

2. Analysis

2.1 Data Cleaning

The original dataset contained 16 variables. We refined that by only focusing on the variables most relevant to our response (price): room_type, number_of_reviews, and availability_365. The variables dropped were either unique identifiers or text-based and were not directly useful for modeling price variation. Since room_type was a categorical variable with three levels (Entire home/apt, private room, shared room), we converted it into a factor. Additionally, we treated the rest of the variables along with the response as numeric (integer) values. No missing values were observed for the observations of our chosen variables.

2.2 Exploratory Data Analysis

After cleaning the data, we began our explanatory analysis by exploring the distribution of the response variable, price. The summary statistics revealed a wide range in prices, from $0 to $10,00 per night, with a median of $106 and a mean of approximately $153. Extreme values were also present, either due to outliers or high-end luxury listings likely skewing the distribution.

In order to better understand the distributions of all key variables, we created histograms for price, number_of_reviews, and availability_365, and a bar plot for the categorical variable room_type. The histogram for price showed a strong right skew, with many listings priced under $500, leaving a long tail because of higher values. Similarly, number_of_reviews also displayed a right-skewed distribution, with most listings having relatively low reviews. On the contrary, availability_365 showed clustering at 0 and 365 days, likely indicating that several listings

are either rarely available or available year-round. The bar plot for room_type suggested that the majority of listings are “Entire home/apt” or “Private room,” with relatively fewer observations for “Shared room.”

2.3 Statistical Analysis

2.3.1 Correlation and Multicollinearity Checks

We calculated pairwise Pearson correlations among all numeric variables and the correlation matrix (fig 1) highlighted weak linear relationships between price and the covariates of interest.

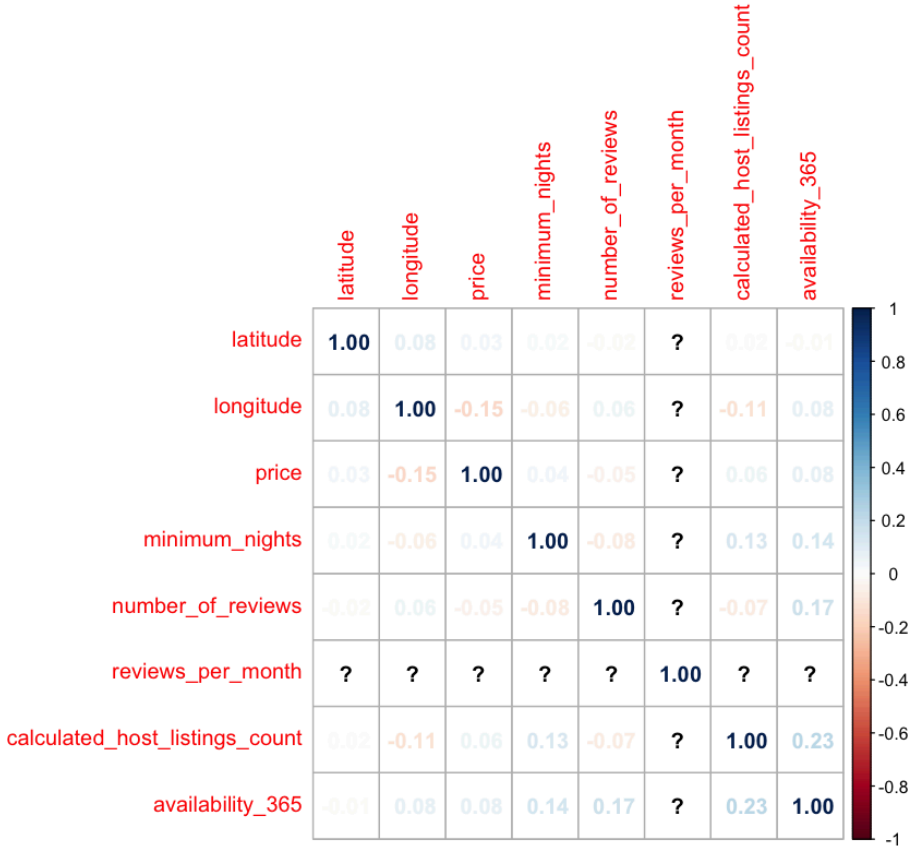

Fig 1. Correlation matrix of selected Airbnb listing features. The values represent Pearson correlation coefficients between numerical variables such as price, number of reviews, availability, and location. Stronger positive correlations are shown in darker blue, while negative correlations are shown in red. Missing values are indicated with “?”.

None of the correlations suggest a strong linear association with price, though availability_365 had the strongest positive correlation out of all.

To visualize potential relationships and check for multicollinearity, we created a scatterplot matrix and also computed partial correlations. Analyzing the partial correlations allowed us to isolate the linear relationship between each pair of variables, while controlling for the others. Once again, we found weak partial correlations between price and all covariates. Furthermore, we assessed all Variance Inflation Factor (VIF) values to be below the threshold of 10, which indicates that multicollinearity is not a concern in our set of covariates.

longitude	price	minimum_nights	number_of_reviews	reviews_per_month	calculated_host_listings_count	availability_365
0.1018489	0.04837526	0.03049811	0.006272811	-0.01363	0.01679294	-0.03930321

Table 1. Partial correlation coefficients between each covariate and the response variable (price). These values quantify the linear relationship between each variable and price while controlling for the influence of other variables.

2.3.2 Model Selection

In order to examine how our covariates influence Airbnb prices, we fit several linear models:

- Model 1: price ~ room_type
- Model 2: price ~ number_of_reviews
- Model 3: price ~ availability_365
- Model 4: price ~ room_type + number_of_reviews + availability_365 (Additive model)
- Model 5: price ~ room_type * number_of_reviews * availability_365 (Full interaction model)

We used R-squared and adjusted R-squared values to evaluate the performance of each model, the results are shown in table 2 below.

A data.frame: 5 × 3

model_name	r_squared	adjusted_r_squared
<chr>	<dbl>	<dbl>
lm_type	0.065613666	0.065575444
lm_num_reviews	0.002299608	0.002279202
lm_availability	0.006695957	0.006675641
lm_all	0.076499902	0.076424344
lm_interaction	0.081243624	0.081036879

Table 2. Comparison of multiple linear models using R-squared value and adjusted R-squared value metrics. The interaction model (Model 5) achieved the highest performance, which suggests that interaction effects may play a meaningful role in explaining variation in Airbnb prices.

The interaction model (**lm_interaction**) had the highest R-squared value, though the overall explanatory power remained low as it could only explain 8.1% of the variation in the response. This suggested that the possibility of unobserved variables that likely play an important role in determining Airbnb prices. Nonetheless, we retained the interaction model for further diagnostic analysis due to the relatively better fit.

2.3.3 Model Diagnostics

We examined residual behaviour for model diagnostics. The residuals exhibited quite a wide range, with a minimum of -284 and a maximum of 9917, which confirmed the impact of high-price outliers. After generating diagnostic plots to evaluate model assumptions, we found:

- *Fitted vs Response*: large deviations from the line, indicating a weak fit as shown in Fig 2.

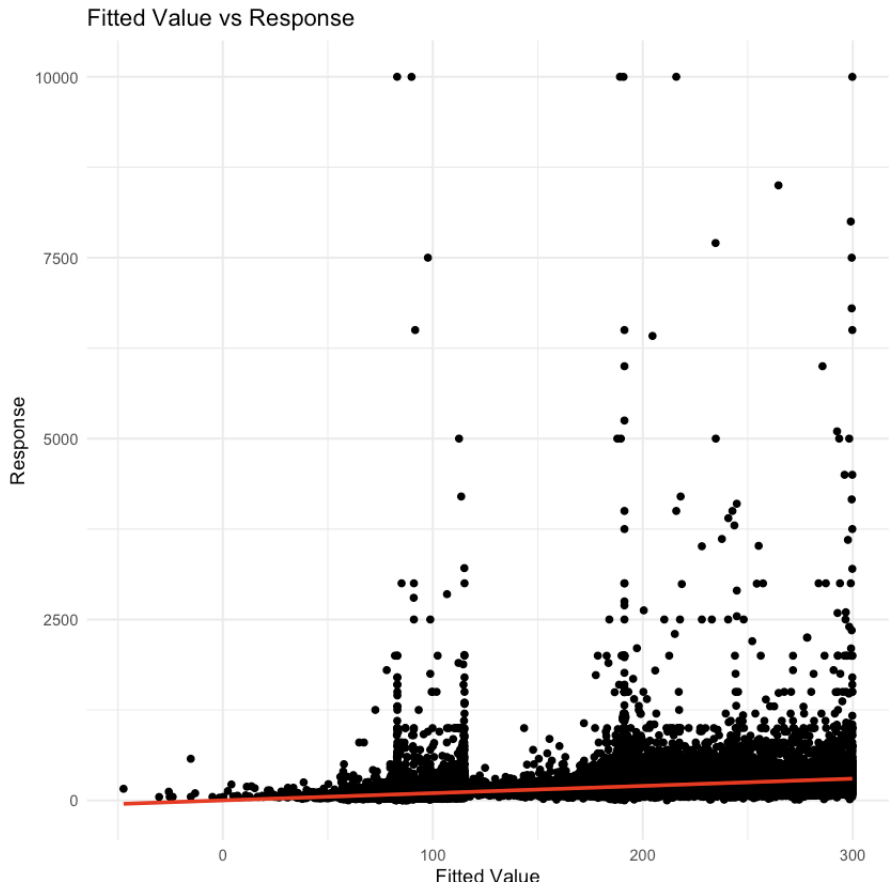

Fig 2. Scatter Plot of fitted values versus observed response values.

- *Residuals vs Fitted*: signs of heteroscedasticity, with increasing variance at higher fitted values.
- *Residuals vs Covariates*: plots against all three covariates showed no strong patterns but confirmed non-constant variance.
- *Normal Q-Q plot*: noticeable departures from normality, particularly in the tails, suggesting the presence of skewness and a potential heavy-tailed distribution as seen in Fig 3.

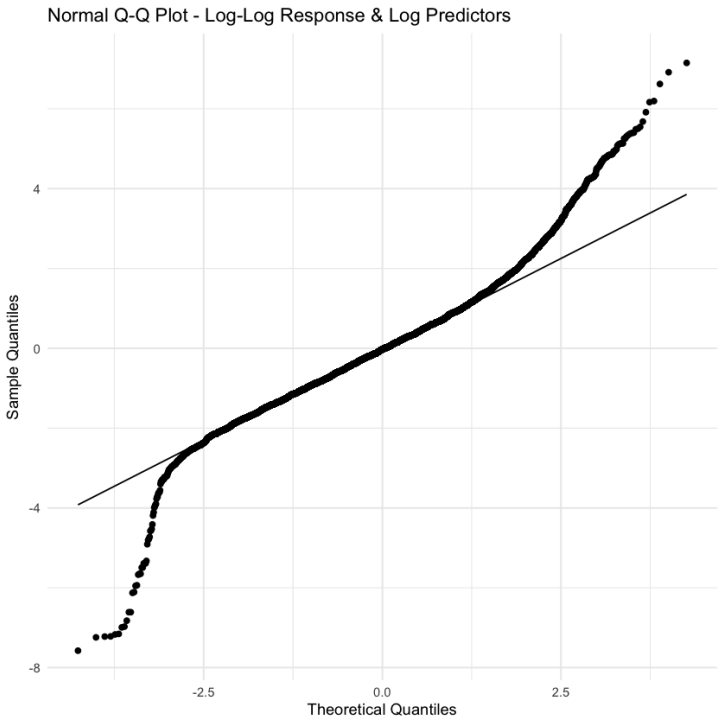

Fig 3. Normal Q-Q plot of residuals from the log-log linear regression model. Deviations from the reference line, especially in the tails, indicate that the residuals are not normally distributed and suggest the presence of outliers or heavy tails. This violates the normality assumption of linear regression.

These diagnostics suggest that the assumptions of homoscedasticity and normality under the linear model are not fully met. Consequently, we explored data transformations.

2.3.4 Data Transformation

Our goal was to improve the model assumptions and address poor performance in the initial models. We applied a series of transformations starting with the response variable. We observed that the price variable had extreme values and violated assumptions of normality and constant variance in fig 4, so we removed listings with a price of zero and log-transformed the remaining values. This transformation helped reduce the influence of outliers and stabilized variance across observations. After applying the log transformation to price, the adjusted R-squared value improved substantially.

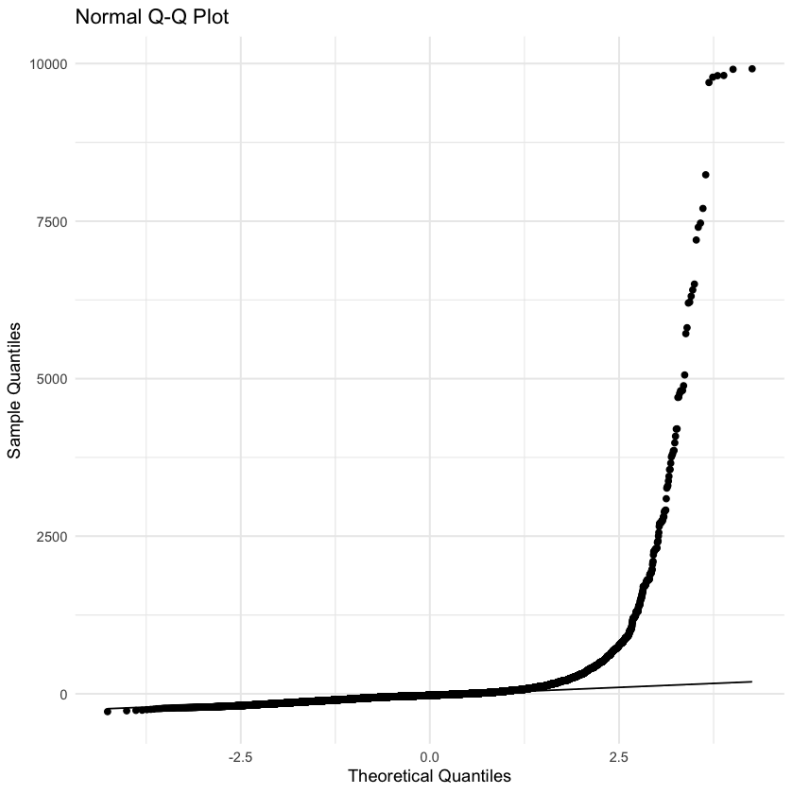

Fig 4. Normal Q-Q plot of residuals from the initial linear regression model with the three covariates and their interactions. Offsetting significantly from the theoretical quantile line, suggesting that the residuals are not normally distributed. This significantly violates the normality assumption of linear regression.

Incentivized by this improvement, we further explored whether transforming the covariates variables could enhance the model. We log-transformed the **number_of_reviews** and **availability_365**, while ensuring they had no zero values by replacing them with a small constant (0.0001). These transformations aimed to reduce skewness in the covariates.

The model with both log-transformed covariates and response variable (**lm_log_y_log_x**) achieved an adjusted R-squared value of approximately 40.5%, and when interaction terms were added, it improved slightly to 40.8%. To experiment further, we tested a double log transformation on the response variable (**log(log(price))**) along with the already log-transformed covariates. This model performed even better, achieving the highest adjusted R-squared value of 42.8%. These results confirmed that the combination of transformations and interaction terms significantly enhanced the model's ability to capture price variation, even though some nonlinear effects were still there.

Furthermore, to support this, we plotted Q-Q plots and residual plots which showed improved normality and reduced heteroscedasticity in the transformed models.The standardized residuals appeared more evenly spread, and the number of extreme outliers dropped in the log-transformed models. Overall, the transformations helped meet key linear regression assumptions and revealed more meaningful relationships between covariates and Airbnb prices.

2.3.5 Final Model

After exploratory data analysis and model comparison, we selected the following final model:
log(log(Price)) ~ room_type +log(number_of_reviews) + log(availability_365)

Model Performance:

- As in Table 3 below, the R-squared value of 0.428 indicates that 42.8% of the variability in log-log(prices) is explained by the model.
- The p-value confirms the model is statistically significant as a whole. Only one interaction (log(reviews) × log(availability) × room_typePrivate room) has a p-value that is bigger than 0.05, which confirms most interaction effects are statistically significant.
- The adjusted R-squared value is very close to R-squared value, showing the model generalizes well despite the number of covariates.

```
=== Coefficient Table (Estimate & p-value) ===
                                                                Estimate
(Intercept)                                                      1.63254
log(number_of_reviews)                                          -0.00402
log(availability_365)                                            0.00265
room_typePrivate room                                           -0.18037
room_typeShared room                                            -0.27205
log(number_of_reviews):log(availability_365)                    -0.00009
log(number_of_reviews):room_typePrivate room                     0.00178
log(number_of_reviews):room_typeShared room                      0.00077
log(availability_365):room_typePrivate room                     -0.00101
log(availability_365):room_typeShared room                      -0.00469
log(number_of_reviews):log(availability_365):room_typePrivate room  0.00002
log(number_of_reviews):log(availability_365):room_typeShared room   0.00035
                                                                Pr(>|t|)
(Intercept)                                                      0.00000
log(number_of_reviews)                                           0.00000
log(availability_365)                                            0.00000
room_typePrivate room                                            0.00000
room_typeShared room                                             0.00000
log(number_of_reviews):log(availability_365)                     0.00002
log(number_of_reviews):room_typePrivate room                     0.00000
log(number_of_reviews):room_typeShared room                      0.24187
log(availability_365):room_typePrivate room                      0.00000
log(availability_365):room_typeShared room                       0.00000
log(number_of_reviews):log(availability_365):room_typePrivate room  0.42914
log(number_of_reviews):log(availability_365):room_typeShared room   0.00048

=== Model Performance ===
      R_squared Adjusted_R_squared F_statistic df1   df2 p_value
value   0.42806            0.42793    3325.183  11 48872       0
```

Table 3. Estimated coefficients and p-values from the full interaction log-log linear regression model estimating Airbnb price. The model includes interaction terms between room type, number of reviews, and availability. Significant coefficients ($p < 0.05$) suggest notable influence on the response. The model achieves an adjusted R-squared value of 0.4279, indicating moderate explanatory power.

This final model captured key structural relationships and interaction effects relevant to Airbnb pricing. While it does not explain all variation in prices (which is expected in real-world datasets), it offers an interpretable and statistically sound summary of the data. The significance of interaction terms suggested that the relationship between reviews, availability, and price depends on room type.

3. Discussion & Conclusion

Initial models showed weak explanatory power. The best-performing untransformed model, the full interaction model, explained only 8.1% of the variance in Airbnb prices, while the additive model with all three covariates (room type, number of reviews, and availability) explained just 7.6%. These low R-squared values indicate that the selected covariates have limited ability to explain variation in price in their original form. This

suggests that the relationship between these features and price is likely nonlinear or influenced by additional unmeasured variables such as location, seasonality, or host ratings. It also highlights the importance of data transformation and model refinement when dealing with skewed or complex real-world data.

We began the analysis by fitting several linear models to explore how room type, number of reviews, and availability affected Airbnb prices. First we used raw, untransformed variables, but these models offered limited explanatory power. The additive model (**lm_all**) explained only 7.6% of the variance in price (adjusted R^2), and the full interaction model (**lm_interaction**) slightly improved that to 8.1%. These results showed that the chosen covariates alone could not explain much of the variation, indicating a nonlinear relationship, heteroscedasticity, or missing variables.

To address these problems, we log-transformed the response variable (price) to stabilize the variance and reduce the impact of outliers. This step significantly boosted model performance. The basic log-transformed model (**lm_log_y**) raised the adjusted R^2 to 40.1%, and with the addition of interaction terms (**lm_log_y_interaction**), it increased slightly to 40.7%. These improvements showed that the log transformation helped the model capture more of the variation in price.

Next, we log-transformed the covariates as well, creating double-log models. The model that included both logged covariates and response (**lm_log_log_y_log_x**) achieved an adjusted R-squared value of 42.5%. When they added interaction terms (**lm_log_log_y_log_x_interaction**), the adjusted R-squared value rose to 42.8%, the highest among all models. These results confirmed that combining transformations with model complexity improved the ability to explain Airbnb price variation. Re-examining Q-Q and residual plots showed better normality and reduced heteroscedasticity in the transformed models. Despite these improvements, we acknowledged that unmeasured factors, such as location, amenities, and host reputation, likely played a key role in setting prices but were not included in our analysis.

3.2 Limitations

Temporal Constraints: The data is from 2019, so it does not show how the market changed after COVID-19. Since the pandemic had a big impact on travel and housing, the prices and patterns today might look very different from what the data shows.

Nonlinear Relationships: Even though we used transformations (like log) to try and make things more linear, real-world data can have patterns that are too complex for a straight-line model to fully capture. So, the model might have missed some of those complicated relationships.

Skewed Data & Outliers: Some listings had really high prices that were way above the average. Even after trying to adjust the data, those extreme values still could have pulled the results in a way that could have made the predictions less accurate

3.3 Potential Future Research

Future research could benefit from using more detailed location-based variables, such as specific neighborhoods, proximity to subway stations, tourist spots, or local crime rates. These spatial features likely play a big role in pricing and could boost the model accuracy. It might also be helpful to apply natural language processing (NLP) to listing descriptions and reviews, which could reveal quality indicators not captured in the standard data. Lastly, access to time-series or panel data would allow for analysis of price trends over time which would account for seasonality, market changes, or new policies like rental regulation.

Appendix

Unused variables in the original dataset:

- Listing ID: A unique identifier for each listing.
- Name: Name of the listing.
- Host ID: A unique identifier for each host.
- Host Name: Name of the host.
- Neighborhood Group: The borough in which the listing is located.
- Neighborhood: The specific neighborhood within the borough.
- Latitude: One geographical coordinate of the listing.
- Longitude: Another geographical coordinate of the listing.
- Minimum Nights: The minimum number of nights required per stay.
- Last Review: Date of last review by guest.
- Reviews per month: The number of reviews per month.
- Calculated host listings count: The number of listings managed by the same host.

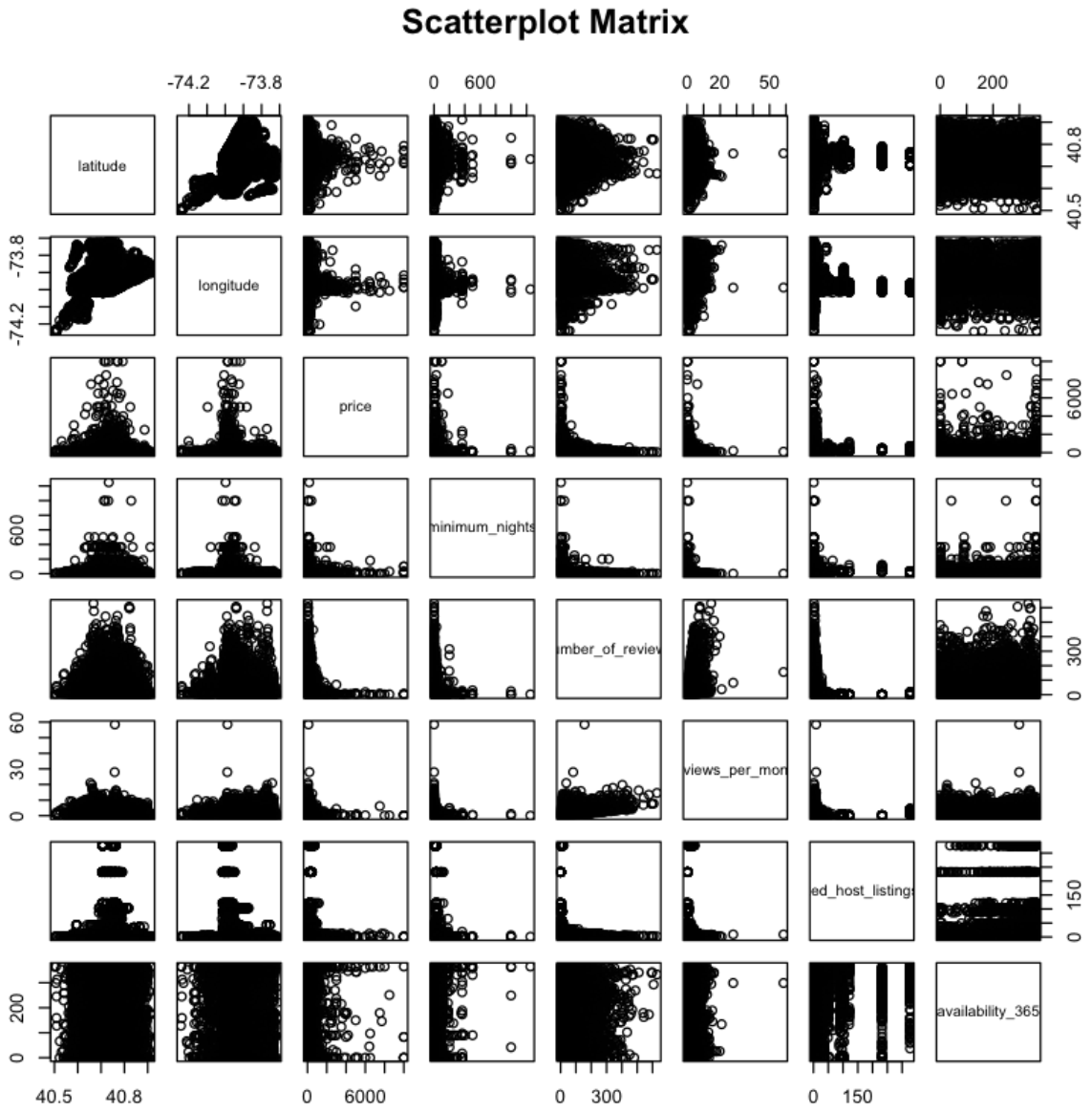

Fig 5. Scatterplot matrix displaying pairwise relationships among all numeric variables in the dataset. This visualization helps assess potential linear and nonlinear associations, detect outliers, and identify collinearity patterns across features such as location, price, review counts, and availability.

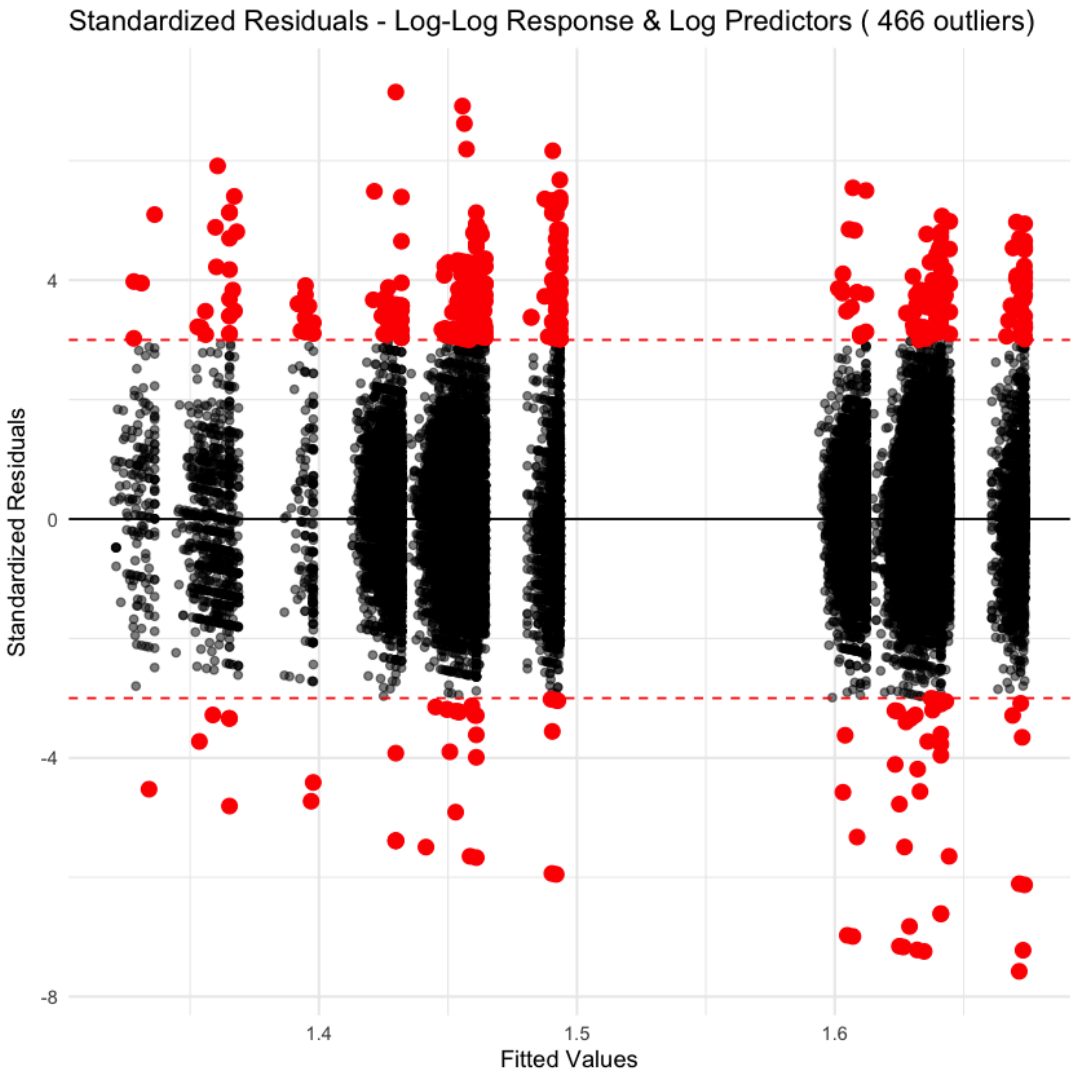

Fig 6. Standardized residuals plotted against fitted values for the log-log linear regression model. Red points represent 466 outliers with standardized residuals exceeding ±3, suggesting violations of model assumptions and the presence of extreme observations. The horizontal dashed lines mark the common threshold for outlier detection.

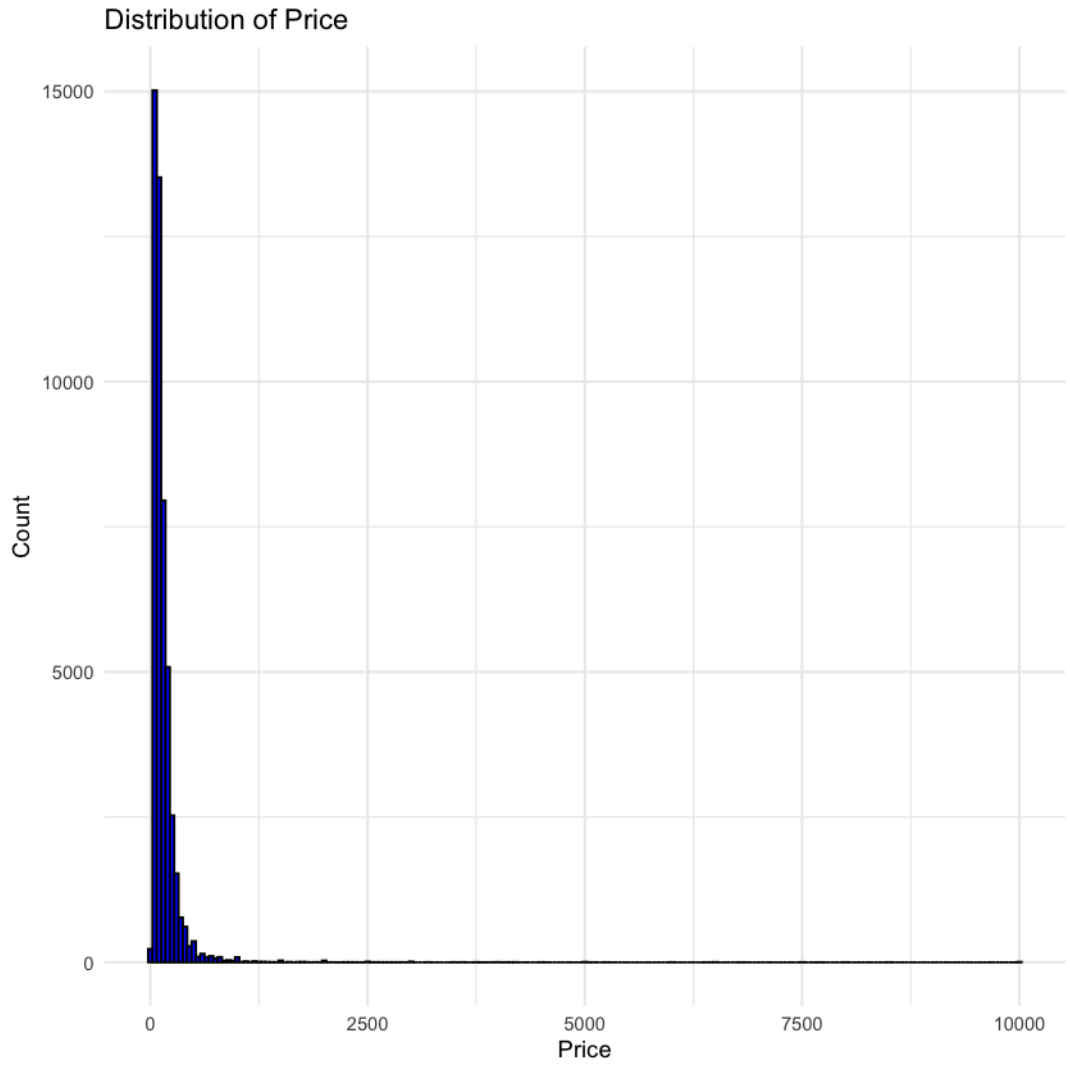

Fig 7. Histogram showing the distribution of Airbnb listing prices. The distribution is heavily right-skewed, with most listings priced below $500 and a long tail of high-priced outliers. This skewness suggests the need for a log transformation to stabilize variance and improve model fit.

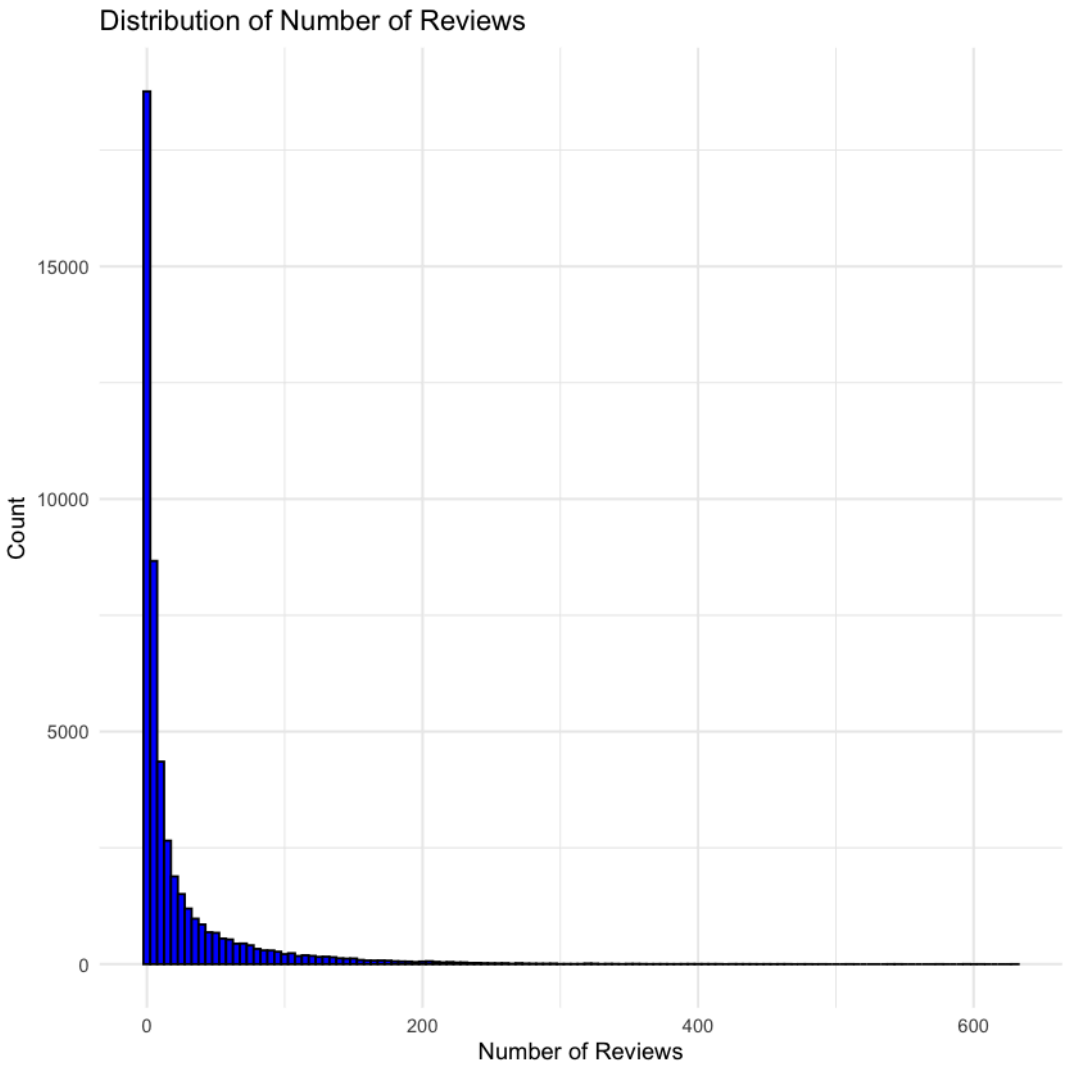

Fig 8. Histogram showing the distribution of number of reviews. The distribution is heavily right-skewed, suggesting the need for a log transformation to stabilize variance and improve model fit.

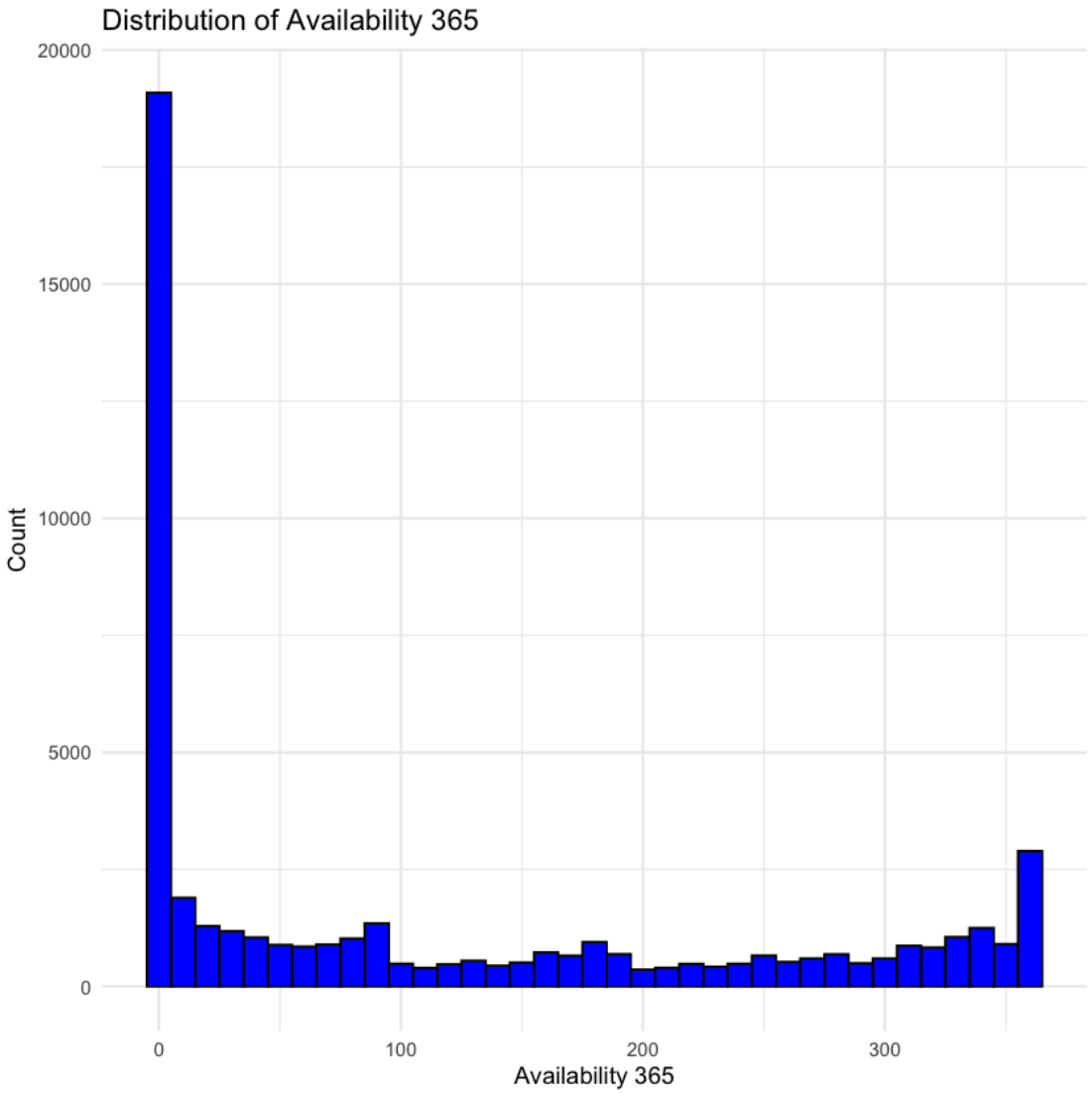

Fig 9. Histogram showing the distribution of availability_365, representing how many days per year a listing is available. The spikes at 0 and 365 indicate a large number of listings that are either never available or available year-round, suggesting strong heterogeneity in host activity or listing strategy.

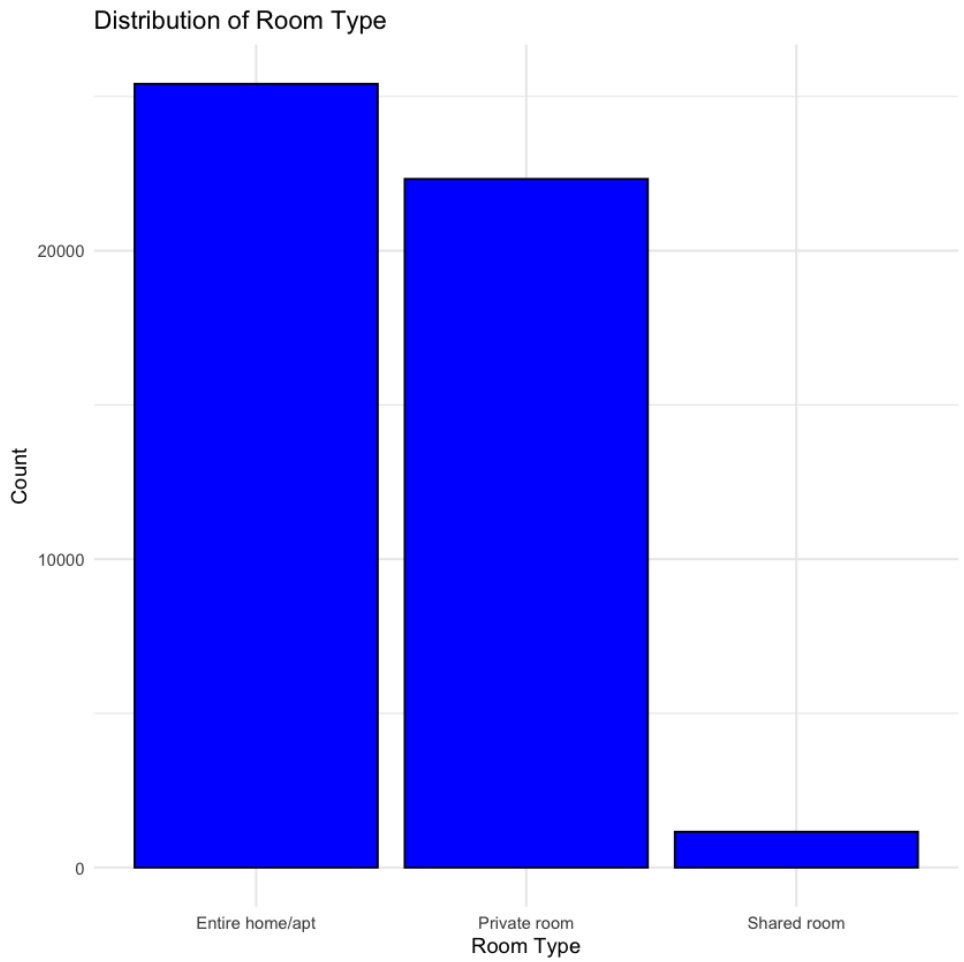

Fig 10. Bar chart showing the distribution of room types in the dataset. Entire home/apartment listings are the most common, followed by private rooms. Shared rooms make up a small fraction of all listings, indicating that most Airbnb accommodations offer a higher degree of privacy.

References

Chen, Y., & Xie, K. (2017, September 11). Consumer Valuation of Airbnb listings: A hedonic pricing approach. International Journal of Contemporary Hospitality Management. https://www.emerald.com/insight/content/doi/10.1108/ijchm-10-2016-0606/full/html

Dgomonov. (2019, August 12). New York City airbnb open data. Kaggle. https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data?resource=download

Acknowledgement

Yirui Wang submitted the code and data files to Canvas.