Predicting Airbnb Booking Rates through EDA and Regression Analysis

Antoinette Nguyen

Stat 418 Tools in Data Science

1 Introduction

A friend of a friend named Richie owns a house in San Diego that he operates as an Airbnb rental. It is located in Ocean Beach, a coastal neighborhood known for its beach-town vibe that attracts beach-goers and tourists. The area has a historic downtown filled with restaurants and boutiques, making it a popular destination for both short-term visitors and locals. Richie has no problem keeping his Airbnb rental booked during the spring and summer months. However, like many hosts, he experiences a noticeable decline in bookings during the off-season.

Richie is interested in understanding what factors are associated with higher booking rates. To help him achieve this goal, I explored patterns in Airbnb listings within his neighborhood that could explain variations in guest popularity. By identifying significant predictors of listings across Ocean Beach, Richie, and hosts like him, can make informed decisions to increase the visibility and appeal of their rentals year-round.

For my effort, I performed exploratory data analysis (EDA) and built regression models aimed at predicting the monthly rate of reviews, a proxy for booking frequency, for listings in Ocean Beach. I also provide tailored suggestions for Richie and fellow hosts looking to boost their rental performance.

1.1 Dataset

The dataset was sourced from the Inside Airbnb website [1], a public platform that provides detailed data on Airbnb listings for cities around the world. For this analysis, I used the San Diego dataset and focused specifically on listings in the Ocean Beach neighborhood, resulting in a total of 548 listings.

The dataset contains a comprehensive set of characteristics related to each listing and its respective host. These include host reputation indicators, listing ratings, and various property characteristics such as pricing, availability, and amenities.

1.2 Methods Used

- Text Manipulation with `dplyr` in R
- Exploratory Data Analysis in R
- Left Joins
- Ordinary Least Squares Regression Model
- Lasso Regression Model
- Ridge Regression Model

For this analysis, I built and compared three linear regression models. The first is an Ordinary Least Squares (OLS) model, developed using a backward selection process to identify significant predictors of reviews per month. In addition to this, I used two regularized regression models, Lasso and Ridge regression model, to assess whether automated variable selection methods can lead to comparable or improved performance in explaining the variation in number of reviews per month

2 Exploratory Data Analysis Findings

2.1 Target Variable

The target variable for this analysis is the monthly review rate, labeled as `reviews_per_month`. This variable serves as a proxy for the frequency of bookings a rental receives over time. Since guests typically leave reviews after their stay, a higher number of reviews per month suggests more frequent bookings.

2.2 Predictor Variables

To begin the modeling process, I performed a preliminary variable selection based on some first-stage EDA and a brief interview with Richie, the Airbnb host.

After listening to the strategies Richie used to make his rental more popular, I developed several hypotheses regarding potential predictors of the reviews per month variable. I expected

that the number of amenities, having a Superhost label, high host response and acceptance rates, enabling instant booking, and strong review scores would be positive and significant predictors of the number of reviews per month.

On the other hand, I hypothesized that higher prices and longer minimum night requirements could act as negative predictors.

While exploring the dataset, I observed that several variables seem autocorrelated like the relationship between the number of beds, bedrooms, and guest capacity. To avoid redundancy and multicollinearity, I chose to exclude related measurements during the initial stage of model development.

In the data preprocessing step, I used text manipulation techniques to convert formatted values. For example, extracting numeric values from columns that used symbols like `$` or `%`. I also created a new variable by parsing the long string in the `amenities` column to compute the total number of amenities listed for each Airbnb listing.

Below is a list of predictors that were used during the model-development stage:

- **bedrooms**: Does the number of bedrooms affect booking frequency?
- **amenities_count**: Is the number of amenities associated with the booking frequency?
- **minimum_nights**: Are guests more likely to book when fewer nights are required?
- **price**: Will reducing the nightly price rate encourage more bookings?
- **host_is_superhost**: Does being a Superhost lead to higher booking frequency?
- **instant_bookable**: Does offering instant booking increase booking frequency?
- **host_acceptance_rate**: Does a higher acceptance rate correlate with more bookings?
- **host_response_rate**: Is a responsive host more likely to attract bookings?
- **review_scores**: Do optics of having highly rated reviews affect booking frequency?
  - **value**
  - **rating**
  - **cleanliness**
  - **accuracy**
  - **location**

Table 1: Summary Statistics of Airbnb Dataset Variables

Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
host_response_rate	0.00	100.00	100.00	98.84	100.00	100.00
host_acceptance_rate	0.0	96.0	99.0	95.6	100.0	100.0
bathrooms	0.000	1.000	1.000	1.405	2.000	7.000
beds	0.00	1.00	2.00	2.62	3.00	19.00
price	47.0	153.0	216.0	281.2	316.2	2200.0
minimum_nights	1.000	2.000	2.000	4.383	3.000	31.000
maximum_nights	3.0	28.0	90.0	304.4	365.0	1125.0
availability_30	0.00	4.00	8.00	9.87	14.00	30.00
review_scores_rating	2.000	4.760	4.880	4.831	4.960	5.000
review_scores_accuracy	2.000	4.810	4.900	4.856	4.960	5.000
review_scores_cleanliness	2.000	4.760	4.860	4.825	4.950	5.000
review_scores_checkin	1.00	4.87	4.93	4.90	4.98	5.00
review_scores_communication	4.000	4.870	4.940	4.908	4.990	5.000
review_scores_location	1.000	4.860	4.920	4.887	4.960	5.000
review_scores_value	2.000	4.670	4.780	4.735	4.862	5.000
calculated_host_listings_count	1.00	2.00	5.00	21.48	41.00	168.00
reviews_per_month	0.030	1.067	2.005	2.366	3.525	14.620
amenities_count	7.00	37.00	47.00	46.35	57.00	83.00

3 Model

A multiple linear regression model was used as a baseline model for predicting the number of reviews per month, which I'll call the OLS Model. I initially began with a full model, including all relevant predictors that were not highly correlated with one another to avoid multicollinearity.

Table 2: OLS Regression Results: Predicting `log1p(reviews_per_month)`

Variable	**Estimate**	**Std. Error**	**t value**	**Pr($>\|t\|$)**
(Intercept)	1.946	0.406	4.797	2.11×10^{-6}
I(amenities_count^2)	8.83e-05	1.56e-05	5.666	2.43×10^{-8}
log1p(minimum_nights)	-0.267	0.029	-9.350	$< 2 \times 10^{-16}$
room_typePrivate room	-0.272	0.094	-2.909	0.0038
room_typeShared room	0.214	0.218	0.979	0.328
log1p(price)	-0.282	0.036	-7.895	1.75×10^{-14}
host_is_superhostTRUE	0.152	0.044	3.483	0.00054
instant_bookableTRUE	-0.085	0.039	-2.188	0.029
review_scores_value	0.161	0.074	2.165	0.031
Residual Std. Error: 0.419 (df = 516)				
Multiple R-squared: 0.322 Adjusted R-squared: 0.311				
F-statistic: 30.63 on 8 and 516 DF, $p < 2.2 \times 10^{-16}$				

To improve the model's performance and address potential multicollinearity, I used two regularization techniques: ridge regression and lasso regression. Ridge regression penalizes large coefficients and is designed to handle multicollinearity by shrinking coefficients toward zero. In addition, Lasso regression not only shrinks coefficients but has the potential to set some to exactly zero, thus performing an automated variable selection.

However, as shown on Table 3, regularization led to only minimal gains in adjusted R^2 and RMSE compared to the ordinary least squares model. As a result, I chose to proceed with the OLS model for interpretation and evaluation.

The OLS model provides a reasonable-moderate fit for predicting the number of reviews per month. I evaluated the model using residual diagnostics in R. The residual plots indicate that the model assumptions are largely satisfied. In particular, the Normal Q-Q plot shows that the residuals are approximately normally distributed, with only minor deviations from the diagonal line.

Additionally, I identified three influential observations using the residual-leverage plot. Upon closer inspection, these points corresponded to listings for reservations at the same rental unit at a hostel. To reduce dependence between observations, I removed the observations from the dataset.

3.1 Results with Cross Validation Applied

The data was split into an 80/20 train-test split to evaluate model performance on test data. Additionally, 10-fold cross-validation was applied to the training set to improve the stability and reliability of the OLS model's estimates.

Model	R-squared	RMSE
OLS (CV)	0.302	0.418
Ridge	0.340	0.399
Lasso	0.339	0.399

Table 3: Performance Comparison of Models

4 Conclusion

In this analysis, I built a multiple linear regression model to predict the log-transformed monthly review rate of Airbnb listings using a range of listing and host-related characteristics. Significant predictors included squared amenities count, log-transformed minimum nights and price, room type, Superhost status, instant bookability, and review scores.

The final model explained approximately 30.2% of the variance in the monthly review rate (Adjusted $R^2 = 0.302$), indicating a moderate level of predictive power.

Listings with more amenities and higher review scores tend to receive more reviews.

Longer minimum night requirements and higher prices are associated with fewer reviews.

Superhosts and entire home listings tend to perform better than others.

Instant bookable listings showed a slight negative association with review frequency.

Most of my hypothesis on the predictors appear to have strong evidence, except for instant bookable. I initially hypothesized that having a listing set with "instant bookable" would be related to higher booking frequency. However, it appears that it is the opposite. This might be useful for a Airbnb host to know in case they're reluctant to turn on the instant bookable feature.

4.1 Future Considerations

To enhance predictive performance of the model to understand predictors Airbnb popularity, I could incorporate external factors such as:

- Neighborhood crime rates that may influence guest comfort or safety.
- Proximity to major tourist attractions, like beaches or landmarks
- Weather fluctuations, as it may affect travel behavior and booking frequency.

Additionally, I want perform sentiment analysis on guest review text to understand what drives satisfaction. This could provide actionable insights to help Airbnb hosts improve the guest experience.

5 Plots

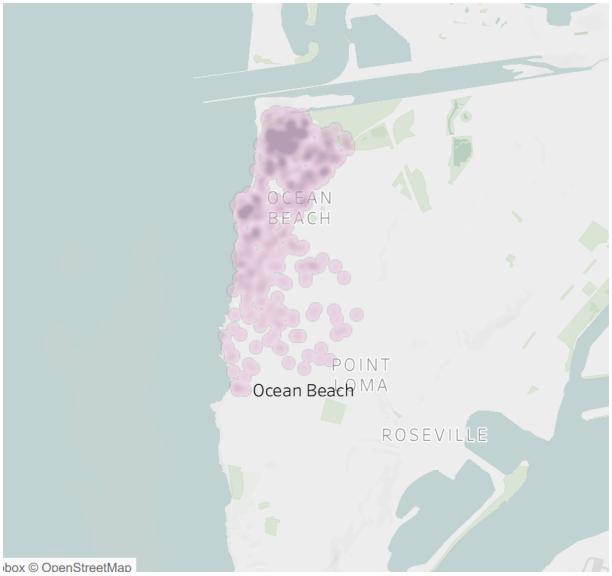

Figure 1: Map of Airbnb Listings in Ocean Beach

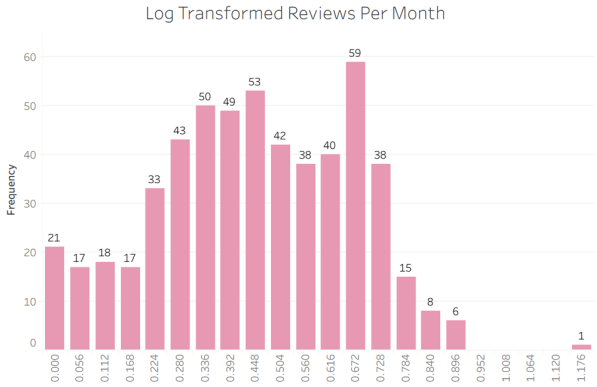

Figure 2: Distribution of Log Transformed Reviews Per Month

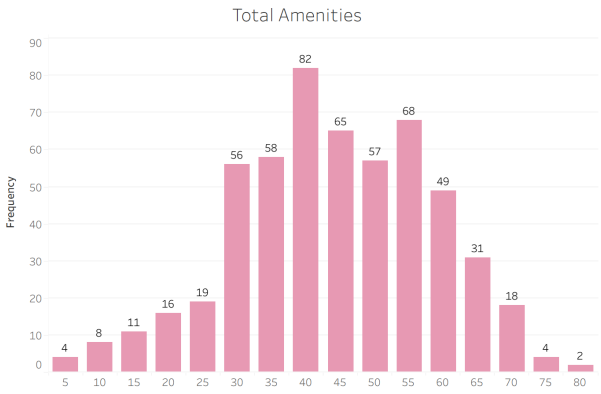

Figure 3: Distribution of Number of Amenities

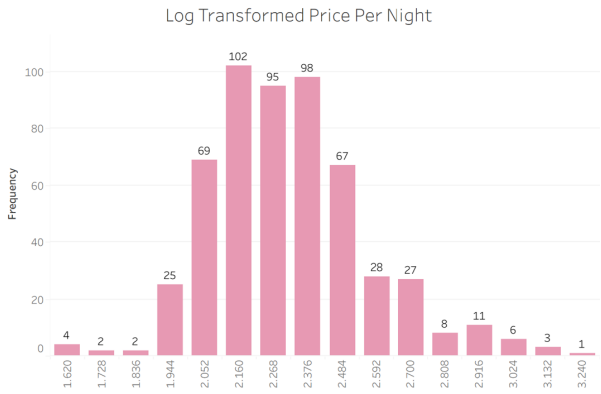

Figure 4: Distribution of Log Transformed Price Per Night

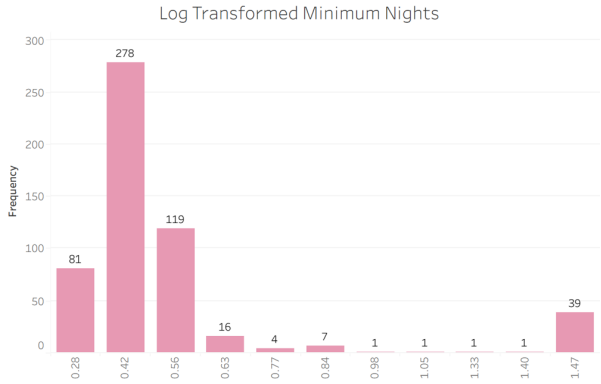

Figure 5: Distribution of Log Transformed Minimum Nights

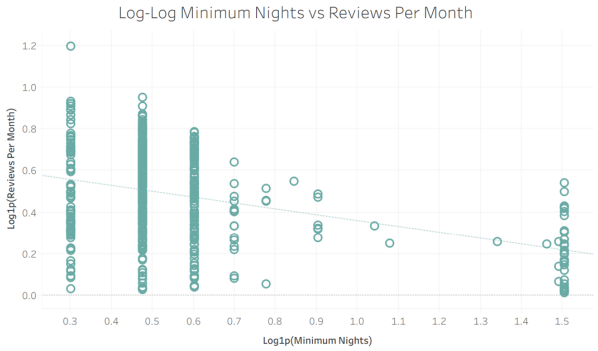

Figure 6: Log-Log Minimum Nights vs Review Per Month

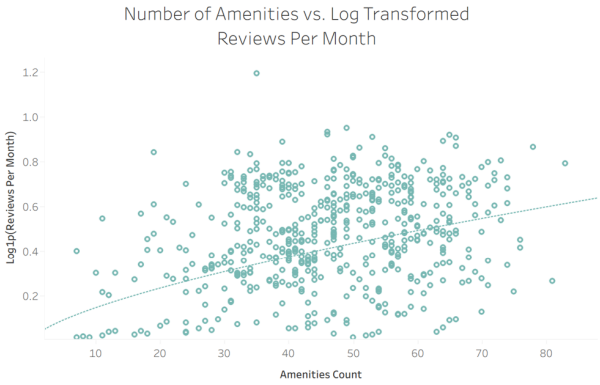

Figure 7: Amenities vs Log Transformed Reviews Per Month

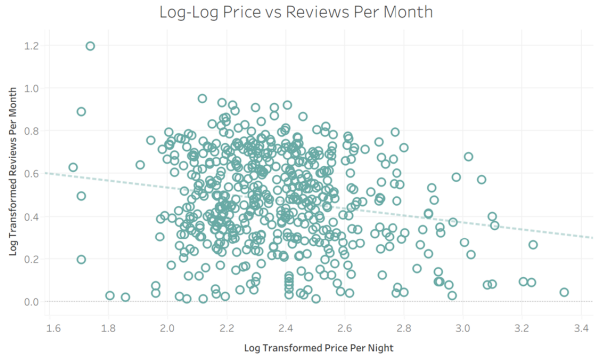

Figure 8: Log-Log Price vs Review Per Month

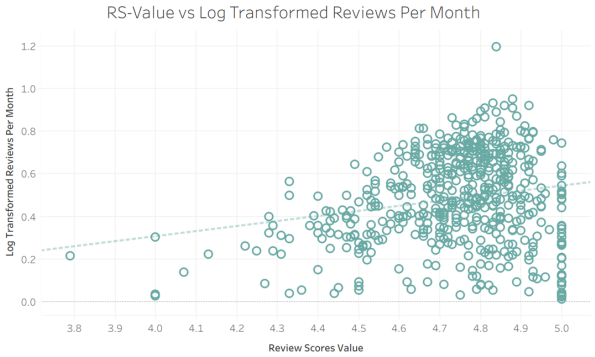

Figure 9: Review Score Value vs Log Transformed Reviews Per Month

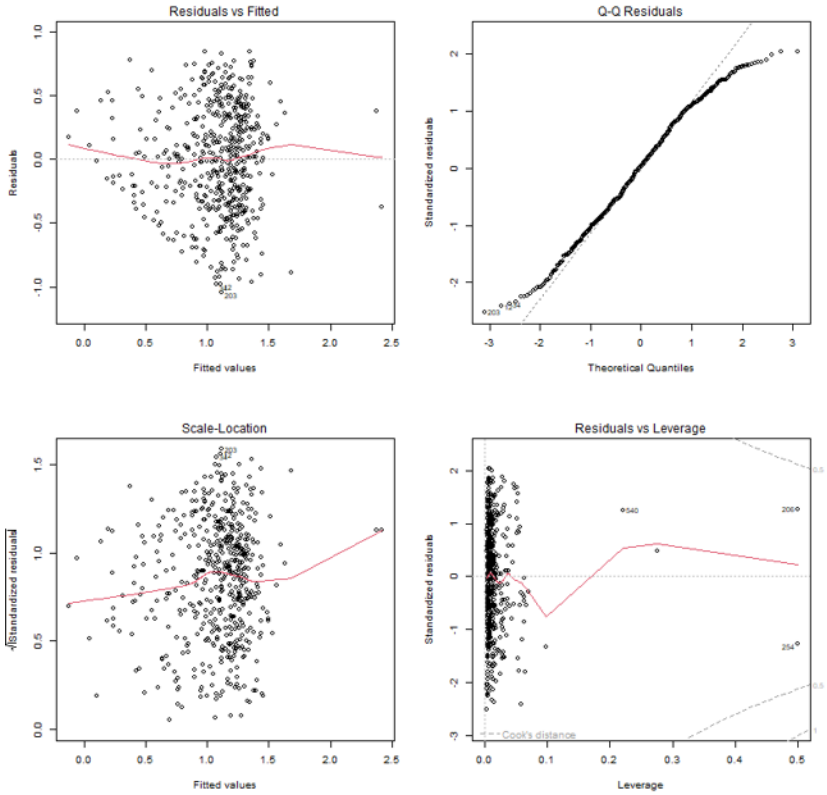

Figure 10: OLS Model Diagnostics Plots

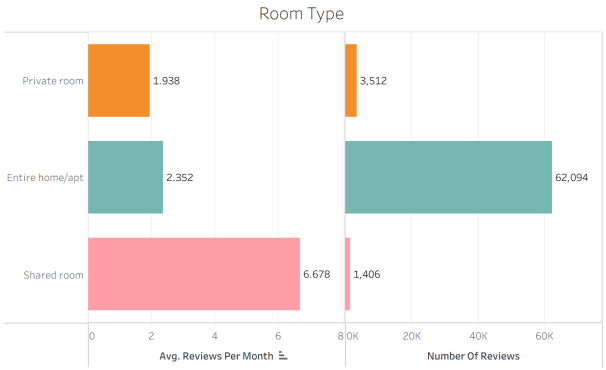

Figure 11: Avg Reviews Per Month by Room Type

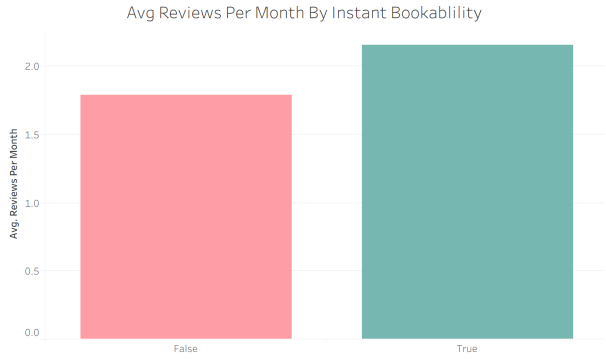

Figure 12: Average Reviews Per Month by Instant Bookable

Figure 13: Correlation Matrix

	Reviews Per Month
Reviews Per Month	1.000
amenities_count	0.212
rs_value	0.177
host_acceptance_rate	0.146
rs_rating	0.122
rs_accuracy	0.120
rs_communication	0.113
rs_cleanliness	0.108
host_response_rate	0.102
longitude	0.091
rs_checkin	0.057
latitude	0.038
rs_location	0.025
availability_30	-0.076
maximum_nights	-0.086
calculated_host_listings_count	-0.115
accommodates	-0.130
beds	-0.137
bedrooms	-0.154
bathrooms	-0.185
price	-0.206
minimum_nights	-0.307

Figure 14: Correlation Coefficients w.r.t Reviews Per Month

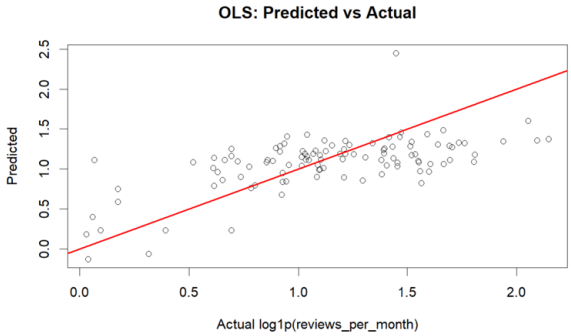

Figure 15: Correlation Coefficients w.r.t Reviews Per Month

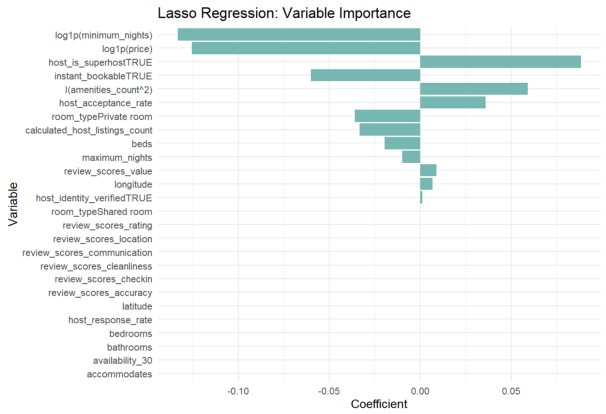

Figure 16: Correlation Coefficients w.r.t Reviews Per Month

References

[1] Inside Airbnb. San diego, ca dataset. https://insideairbnb.com/get-the-data/, 2025.