Predicting Nightly Listing Prices on Chicago Airbnb

DSC 423 Final Project

Sylvia Ifeoma Ozoede

Professor Thiru Ramaraj

Executive Summary

This project looks at factors that affect the nightly price of Airbnb listings in Chicago. Using Chicago Airbnb data, five regression models were built, starting with simple models and then building up to a full model and a log-transformed model. The analysis showed four main findings:

- Room type has the biggest effect on price. Entire homes are usually the most expensive, while shared rooms are usually the cheapest.All room-type coefficients are highly statistically significant.
- Neighbourhood and geographic location contribute meaningfully. Adding neighbourhood as a variable raised R-squared from 16.87% to 25.41%, an 8.5 percentage-point improvement.
- The log(price) model provides the best fit. Log-transforming the dependent variable improved R-squared to 39.37% and made the model results more reliable.
- The total number of reviews does not strongly predict price. Its p-value was 0.636, which means it was not statistically significant. However, reviews per month had a small effect in some models.

1. Introduction

Short-term rental platforms such as Airbnb have fundamentally altered urban accommodation markets. Understanding what drives listing prices has practical implications for hosts seeking to optimize revenue, for guests budgeting travel expenses, and for researchers studying housing market dynamics. Chicago, as one of the largest metropolitan rental markets in the United States, provides a rich context for this analysis.

The goal of this project is to identify and quantify the factors that best predict nightly Airbnb listing prices in Chicago. The analysis starts with simple regression models using one factor at a time, then moves to larger models with multiple factors, including a log-transformed model. This step-by-step approach helps show how much each factor adds to explaining Airbnb prices.

2. Data & Methodology

2.1 Dataset

The dataset (listings.csv from https://insideairbnb.com/get-the-data/) contains Chicago Airbnb listings sourced from the Inside Airbnb project. After loading the data, the following variables were selected: nightly price (response/dependent), room type, neighbourhood, latitude, longitude, minimum nights, number of reviews, reviews per month, calculated host listings count, availability in the past 365 days, and number of reviews in the last twelve months. Price was also cleaned.

2.2 Data Cleaning

Categorical variables, room type, neighbourhood, were converted to factors. Rows with missing values were removed to limit the influence of extreme outliers, the extremely expensive listings were excluded so they would not mess up the results. Because of that, the highest price analyzed became lower than the original highest price of $1,331. After cleaning, the analysis dataset contained 6,285 complete observations.

2.3 Analytical Approach

Five ordinary least squares (OLS) regression models were estimated in R:

- Model 1 (model1): price ~ availability_365 (simple regression baseline)
- Model 2 (model2): price ~ number_of_reviews (tests review activity as a predictor)
- Model 3 (model3): price ~ room_type (tests room type as a predictor)

- Model 4 (model4): price ~ room_type + latitude + longitude + minimum_nights + number_of_reviews + reviews_per_month + calculated_host_listings_count + availability_365 + number_of_reviews_ltm (full model without neighbourhood)
- Model 5 (model5): Same as Model 4 plus neighbourhood

- Log Model (model_log): log(price) ~ same predictors as Model 4, testing a log-transformed response to address skewness

Diagnostic plots and the Durbin-Watson test were used to assess residual behavior. Variance Inflation Factors (VIF) were computed for model 4 to check for multicollinearity.

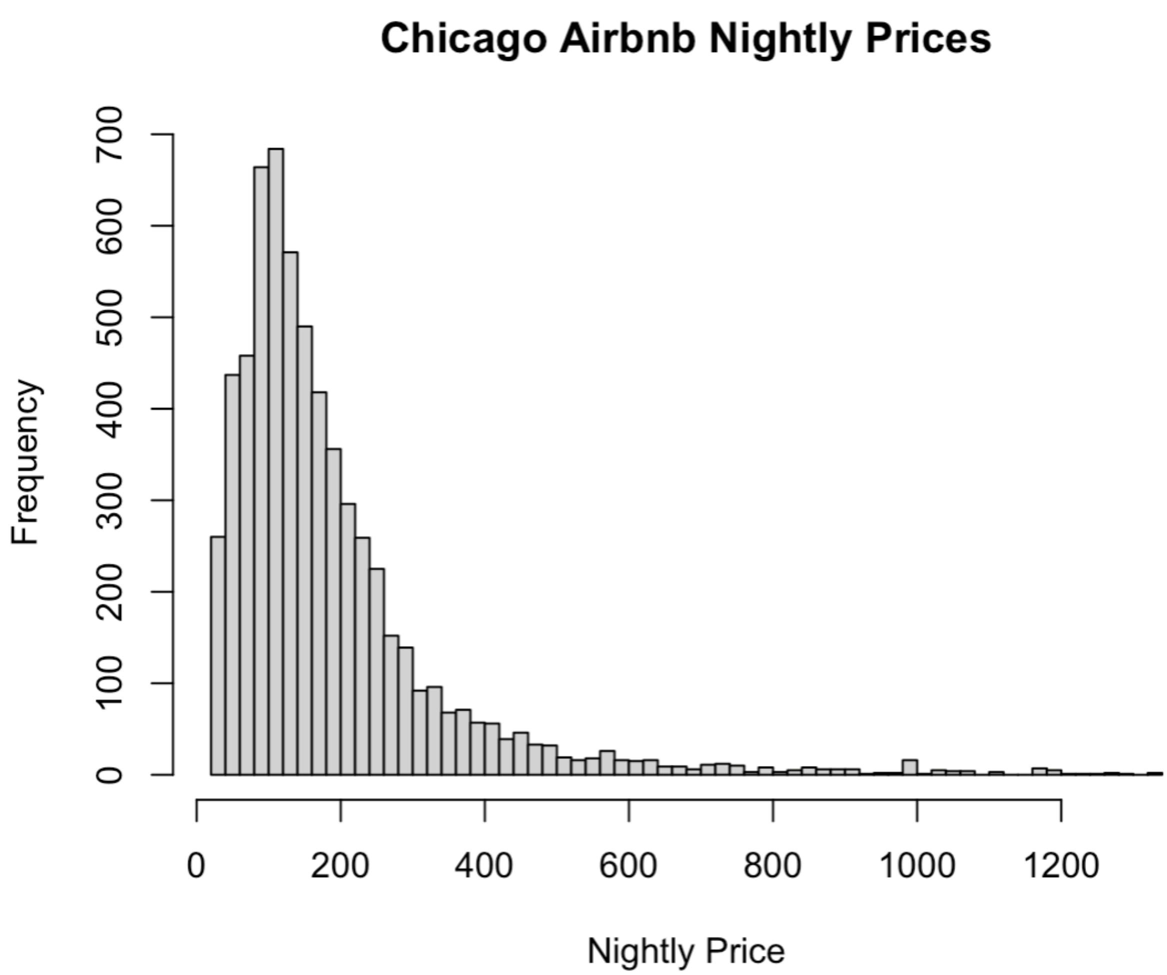
3. Exploratory Data Analysis

3.1 Price Distribution

The distribution of nightly prices is strongly right-skewed. The summary statistics are presented below:

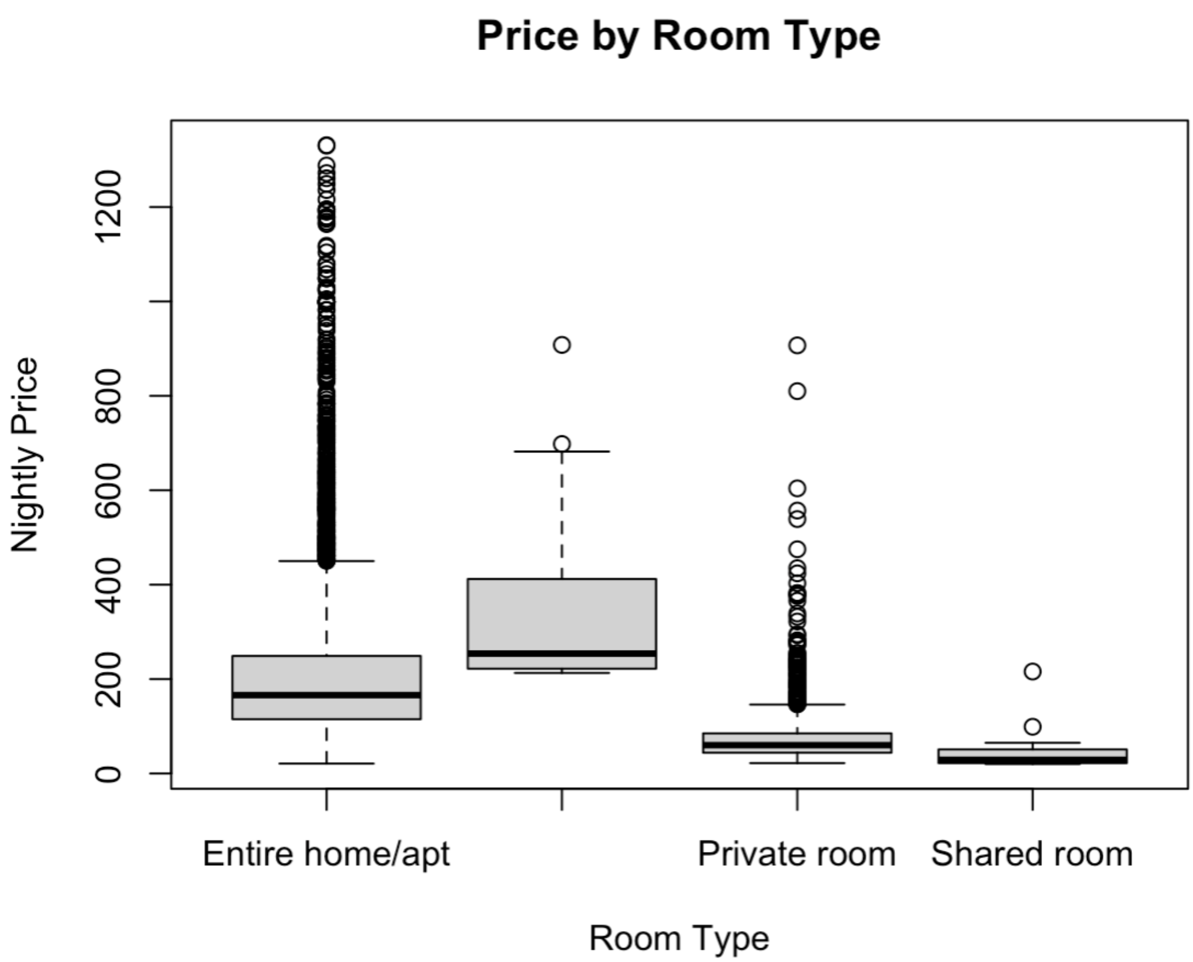
Min	Q1	Median	Mean	Max
$20.00	$95.00	$143.00	$188.50	$1,331.00

The median nightly price is $143 while the mean is $188.50. This $45.50 difference is mainly caused by a small number of expensive luxury listings that raise the average. The histogram confirms that the majority of listings are priced below $300, with a long right tail extending past $1,000. This pattern in the price data does not fit the normality assumption for OLS regression which motivated the log-price model tested in Section 4.5.



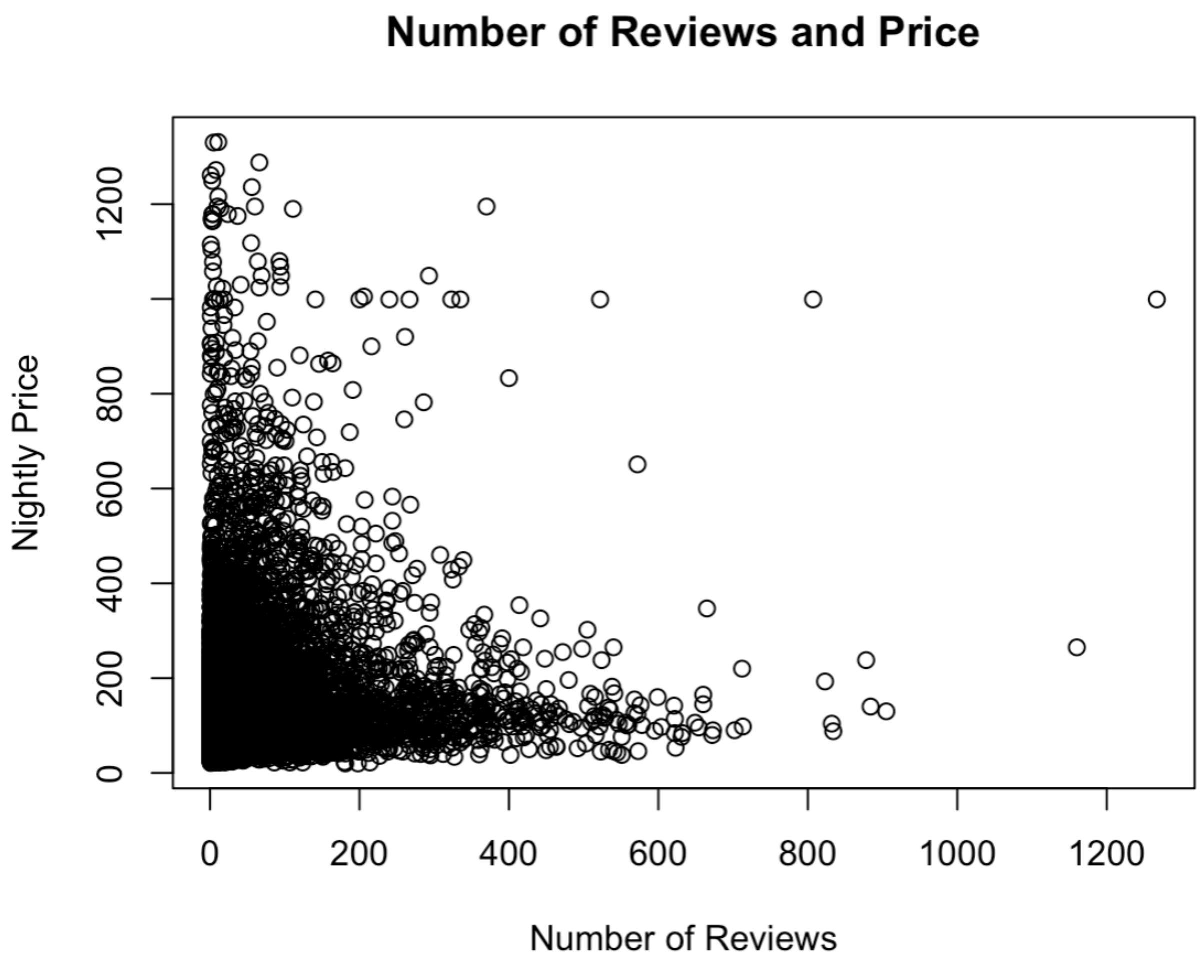
3.2 Price by Room Type

A boxplot of price by room type reveals that prices are different depending on room type. Entire homes and apartments have the highest median prices, followed by hotel rooms, private rooms, and shared rooms. Entire homes and apartments also have the widest range of prices and the most expensive outliers. This makes sense because this group includes many different types of properties, from small studio apartments to large homes with multiple bedrooms.



3.3 Number of Reviews and Price

A scatterplot of number of reviews against nightly price reveals no discernible linear relationship. Listings with many reviews are usually around the middle price range, not the most expensive or the cheapest. The regression results in Section 4.2 also support this.



4. Regression Models

4.1 Simple Regression: Room Type

A simple regression of price on room type (Model 3) yields an R-squared of 11.9%, which means room type explains about 11.9% of the differences in nightly prices. Out of the single factors tested, room type explains price the best.The intercept of $214.83 represents the The models estimates that entire homes and apartments cost about $214.83 per night on average. Compared to entire homes and apartments:

- Hotel rooms are about $125.31 more expensive (Predicted mean: $340.14).
- Private rooms are $136.50 cheaper (predicted mean: $78.33).
- Shared rooms are $170.11 cheaper (predicted mean: $44.72).

All coefficients are statistically significant providing strong evidence that room type systematically affects listing price.

4.2 Simple Regression: Number of Reviews

Model 2 looks at whether the number of reviews affects price. The coefficient on number of reviews is −0.009 with a p-value of 0.636, which is not statistically significant. The R-squared is essentially zero (3.56×10^{-5}). Overall, total review count does not meaningfully predict Airbnb nightly price.

4.3 Multiple Regression without Neighbourhood

Model 4 includes room type, latitude, longitude, minimum nights, total reviews, reviews per month, host listing count, year-round availability, and reviews in the last twelve months. This model explains 16.87% of price variation (Adjusted R^2 = 16.72%). Key findings:

- Room type remains highly significant. The results are similar to Model 3, meaning room type continues to have a strong effect on price.
- Latitude and longitude are both statistically significant, confirming that geographic position matters for price.
- Minimum nights carries a negative coefficient (−1.05, $p < 0.001$): listings with higher minimum stays tend to be priced lower per night, possibly to attract guests who stay longer.
- Reviews per month is significant and negative (−10.26, $p < 0.001$), suggesting that listings with more frequent reviews are not always more expensive.

- Total number of reviews is still not statistically significant, suggesting that total review count does not strongly predict nightly price, even when other factors are included. (p = 0.35).

4.4 Multiple Regression with Neighbourhood

Adding neighbourhood as a factor variable (Model 5) increases R-squared to 25.41% (Adjusted $R^2 = 24.37\%$). This is an improvement of approximately 8.5% over Model 4. Chicago has many distinct neighbourhoods with varying price profiles, and their categorical dummies add information that latitude and longitude alone do not fully capture. The F-statistic for the full model is 24.55 ($p < 2.2 \times 10^{-16}$), confirming joint the model is statistically significant overall, meaning the variables in the model work together to help explain price.

4.5 Log(Price) Regression Model

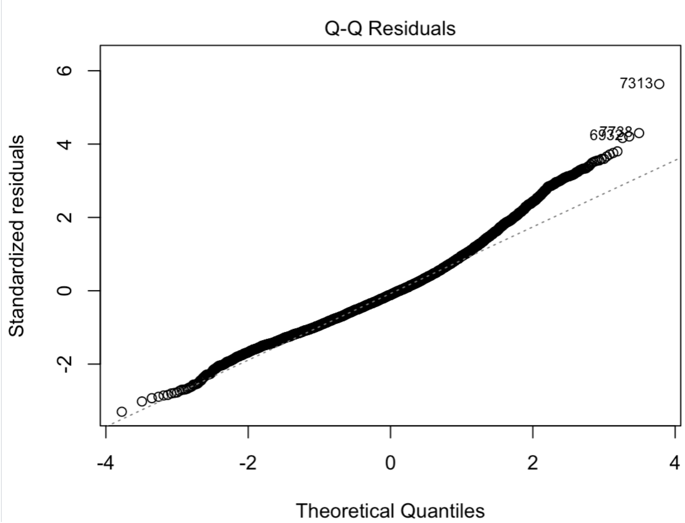
Given the right-skewed distribution of nightly prices, a log-transformed model was estimated with log(price) as the response variable (model_log). This model uses the same predictors as Model 4 (without neighbourhood). The results showed an improvement:

- R-squared increases to 39.37% (Adjusted $R^2 = 39.26\%$), the highest of all models tested.

- Residuals vs. Fitted plot shows more even spread compared to the price-level models.

- The Q-Q plot shows improvement in the middle quantiles, although the extreme values were still not perfectly normal.

In this log model, the coefficients are understood more like percentage changes instead of dollar changes. Room type still had the same general effect as in the earlier models.Specifically, private rooms show a coefficient of −0.965 ($\exp(-0.965) \approx 0.38$), implying private rooms are predicted to cost about 38% of the price of an entire home or apartment, after controlling for the other factors in the model.



5. Model Comparison

The table below summarizes R-squared values across all models estimated in this project, illustrating the contribution of each modeling choice.

Model	R-squared	Key Addition
Model 3 — Room Type Only	11.90%	Baseline predictor
Model 4 — Multiple (no neighbourhood.)	16.87%	Geographic coordinates, reviews, availability
Model 5 — Multiple + Neighbourhood	25.41%	+8.5 pp from neighbourhood variable
Log(Price) Model	39.37%	Log-transform on response, best fit

The log(price) model is the best-fitting overall. However, even though it explains 39.37% of the price differences, most of the variation in nightly prices is still not explained by the variables in this dataset. This suggests that other important factors may also affect price, such as amenities, property quality, cleanliness ratings, listing photos, and how responsive the host is. These factors were not included in the current model.

6. Final Discussion

6.1 Room Type

Room type was one of the most important predictors of price. This makes sense because guests usually pay more for an entire home or apartment than for a private or shared room. Hotel rooms being more expensive than entire homes may seem surprising at first, but it makes sense because hotels often provide extra services, such as professional hospitality, concierge support, and more consistent service.

6.2 Geographic Location

Adding neighbourhood to the model improved the results, which shows that location matters a lot. Airbnb prices in Chicago can vary depending on the neighbourhood. Areas like River North, Gold Coast, and Lincoln Park are usually more expensive, which likely explains part of this effect. Latitude and longitude were still important too. This means that smaller location details may also affect price, such as being close to the lakefront, public transportation, or popular attractions.

6.3 Review Activity

Total number of reviews did not strongly predict price. This suggests that hosts may set prices based more on the property, location, market demand, and season instead of how many reviews the listing has. A high number of reviews may simply mean the listing has been active for a long time. It does not necessarily mean the listing can charge a higher price. Reviews per month had a small effect in some models, which suggests that recent demand may matter a little more than total review count.

6.4 Log Transformation

The log(price) model improved the results and fit the data better than the regular price models. This makes sense because price data often has a few very expensive listings that can affect the model. However, the Q-Q plot still showed some issues at the extreme ends. This means that a small number of very expensive listings still had a strong effect, even after using log(price).

7. Conclusion, Limitations & Recommendations

7.1 Conclusion

Room type and location are the main factors that affect nightly Airbnb prices in Chicago. Entire homes usually have the highest prices, followed by hotel rooms. When neighbourhood was added to the model, the R-squared improved by about 8.5%. The log(price) model worked best overall, with an R-squared of 39.37%. The number of reviews did not have a strong or statistically significant effect on price.

7.2 Limitations

- The dataset does not include detailed amenity information like WiFi, parking, or kitchen access. These features may affect price.
- The dataset also does not include some quality-related factors, such as cleanliness ratings or host response rate.
- Another limitation is that the data does not capture seasonal changes or price increases during major events.
- Even after using the log(price) model, some very expensive listings still affected the results.
- Finally, some review-related variables may overlap with each other.

7.3 Recommendations for Future Research

- Future research should include more details about listing amenities and quality, such as WiFi, parking, cleanliness, and host response rate.
- It would also be useful to study Airbnb prices over several months to understand seasonal changes.