Predicting Airbnb Listing Prices in New York City

Lei Duli
Guo Jingyi

1. Introduction

Airbnb has changed the short-term rental market by letting individual hosts set prices for many different types of properties. In New York City, this market is especially useful for price analysis because listings are spread across five boroughs with clear differences in demand, tourism, housing type, and local amenities. A nightly price is therefore not only a number chosen by a host. It also reflects location, room type, review activity, host scale, and yearly availability.

This project asks: what factors most strongly influence the nightly price of Airbnb listings in New York City? We study this question with the 2019 New York City Airbnb dataset. The main response is the natural log of price, because the raw price variable has a long right tail. The main hypothesis is that room type and borough location are the strongest predictors of price. We expect entire-home listings to have higher prices than private rooms, and we expect Manhattan listings to have a price premium over Brooklyn and the other boroughs. This setup follows the basic idea of hedonic price modeling, where a price is viewed as a function of the characteristics attached to the good or service.

2. Methods

2.1 Data Source and Study Design

The data come from the New York City Airbnb Open Data set. The data were originally collected by Inside Airbnb through public web scraping from Airbnb and were later made available through Kaggle. The original data contain 48,895 listings from 2019 and cover Brooklyn, Manhattan, Queens, the Bronx, and Staten Island. This is a cross-sectional observational study. The unit of analysis is one Airbnb listing.

2.2 Variables and Cleaning

The response variable is price, measured as the listed nightly rate in U.S. dollars. Since price is strongly right-skewed, the main response used in the models is log(price). The categorical predictors are room_type and neighbourhood_group. Private room is used as the reference group for room type, and Brooklyn is used as the reference group for borough. The quantitative predictors are minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, and availability_365.

The cleaning process keeps only variables used in the pricing analysis. Listings with price equal to zero are removed because a zero nightly rate is not a valid market price. Rows with missing values in the modeling variables are removed by complete-case analysis. After cleaning, the analytic sample contains 38,833 listings. This sample is large enough for stable model estimation, but it does exclude listings without review-frequency information.

2.3 Model Plan and Bias

The main analysis uses ordinary least squares regression. Model 1 includes only room type and borough. Model 2 adds the quantitative controls. Model 3 adds an interaction between room type and borough and applies log transformations to skewed numeric predictors. Standard regression diagnostics are then used to assess linearity, homoscedasticity, normality of residuals, multicollinearity, and influential observations.

Several sources of bias remain. The data include listed prices, not final transaction prices. Discounts, seasonal pricing, cleaning fees, and actual bookings are not observed. Important property features such as unit size, bedrooms, amenities, distance to transit, and exact neighborhood demand are also missing. The results should therefore be read as associations, not causal effects. Spatial correlation is also likely because listings near each other may share unobserved local conditions.

3. Data Analysis

The descriptive analysis first summarizes the cleaned sample and then checks the shape of the response variable. The cleaned sample contains 38,833 listings after removing zero prices and missing modeling variables. The raw price variable has a large gap between the mean and median, which shows that a small number of expensive listings pull the distribution upward. The log transformation makes the response more regular and more appropriate for OLS.

Table 1. *Summary of cleaned sample and variables.*

Metric	Value
Raw observations	48,895
Cleaned observations	38,833
Rows removed	10,062
Mean price (USD)	142.35
Median price (USD)	101.00
Mean log(price)	4.697
SD of log(price)	0.664
Entire home/apt share	52.35%
Private room share	45.47%
Manhattan share	42.83%
Brooklyn share	42.33%

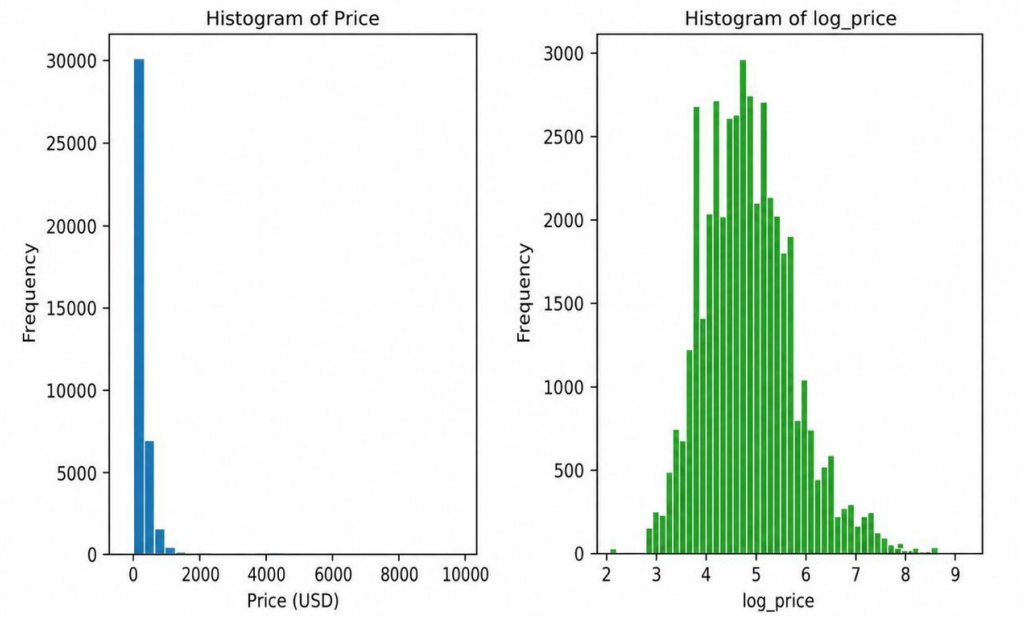

Figure 1. Histogram of price and histogram of log(price)

The first figure supports the use of log(price). Raw price is concentrated at lower values and has a long right tail. Log(price) is more balanced. This does not make the data perfectly normal, but it removes most of the extreme skew that would otherwise dominate the OLS fit.

The next step compares log(price) across listing groups. Entire-home listings have a higher central price than private rooms. Shared rooms have the lowest central price. Borough differences are also clear, with Manhattan listings centered higher than Brooklyn, the Bronx, Queens, and Staten Island.

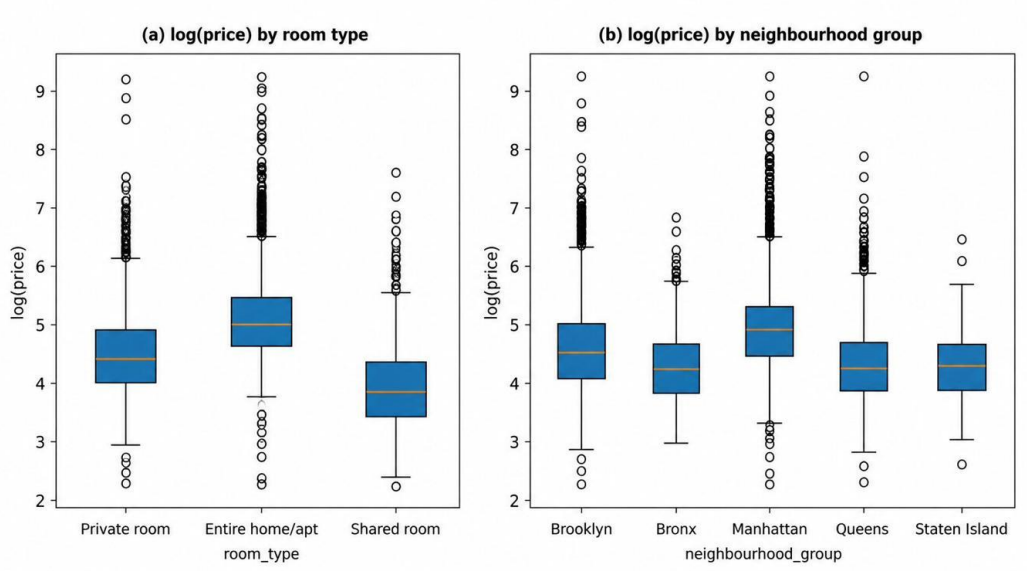

Figure 2. Boxplots of log(price) by room type and borough group

Notes: The boxes show the interquartile range (IQR) with the median as the orange line. Whiskers extend to 1.5 × IQR; points beyond are outliers.

The quantitative predictors show weaker patterns. The scatterplots do not show strong linear relationships between log(price) and the activity variables. The correlation matrix also shows that pairwise correlations among numeric predictors are generally modest, suggesting limited severe multicollinearity.

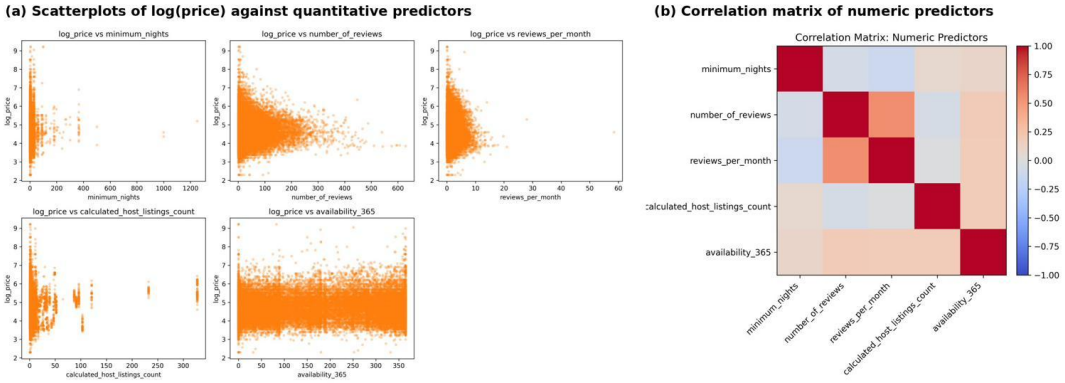

Figure 3. Scatterplots of log(price) against quantitative predictors and correlation matrix

Notes: The left panel shows log(price) against five quantitative predictors. The right panel shows pairwise correlations among the numeric predictors.

4. Linear Models

Model 1 uses room_type and neighbourhood_group only. This model tests whether listing type and borough alone explain a meaningful part of log-price. Model 2 adds minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, and availability_365 as quantitative controls.

Table 2. *Key OLS results for Model 1 and Model 2.*

Variable	Model 1 Coef.	Model 2 Coef.	Model 2 % Effect
Entire home/apt	0.774***	0.779***	118.0%
Shared room	-0.386***	-0.423***	-34.5%
Manhattan	0.305***	0.305***	35.6%
Bronx	-0.246***	-0.290***	-25.2%
Queens	-0.118***	-0.147***	-13.7%
Staten Island	-0.242***	-0.311***	-26.7%
R-squared	0.480	0.498	
Adjusted R-squared	0.480	0.498	

Model 1 explains about 48% of the variation in log(price). This is strong for a simple model using only room type and borough. The signs also match market expectations. Entire-home listings are higher than private rooms, shared rooms are lower, and Manhattan is higher than Brooklyn. The Bronx, Queens, and Staten Island are lower than Brooklyn.

Model 2 improves the fit, but the improvement is limited. R-squared rises from about 0.480 to about 0.498. This means that the quantitative controls add useful information, but they do not change the main result. In Model 2, entire-home listings are about 118% higher than private rooms on the expected price scale. Manhattan is about 35.6% higher than Brooklyn. These large effects remain after the controls are added.

The quantitative controls are smaller in size. Minimum nights and number of reviews are negative and significant. Availability is positive and significant. Reviews per month and host listing count are also significant but with small coefficients. None of these effects are large enough to change the main story driven by room type and borough.

Model 3 is added to improve predictive flexibility. It includes a room_type by neighbourhood_group interaction, so the room-type effect can vary by borough. It also replaces skewed numeric predictors with their log transformations. The model comparison table summarizes adjusted R-squared, AIC, BIC, and RMSE for the three specifications, and the selected Model 3 coefficients are reported afterward.

Table 3. *Model comparison for OLS specifications.*

Model	Adjusted R-squared	AIC	BIC	RMSE
Model 1	0.48	53000.55	53060.52	0.4787
Model 2	0.498	51591.63	51694.43	0.47
Model 3	0.507	50910.83	51082.17	0.4658

Table 4. *Selected Model 3 coefficient estimates.*

Variable	Coefficient	p-value
Entire home/apt	0.841***	<0.001
Shared room	-0.546***	<0.001
Manhattan	0.341***	<0.001
Queens	-0.114***	<0.001
Bronx	-0.259***	<0.001
Staten Island	-0.305***	<0.001
Entire home/apt × Manhattan	-0.067***	<0.001
Entire home/apt × Queens	-0.104***	<0.001
log(minimum nights + 1)	-0.092***	<0.001
log(number of reviews + 1)	-0.024***	<0.001
availability_365	0.001***	<0.001

Because Model 3 includes interactions, the main coefficients require careful reading. The entire-home coefficient is the effect within Brooklyn, because Brooklyn and private room are the reference groups. The Manhattan coefficient is the Manhattan difference for private rooms. The interaction terms show how the room-type premium changes in each borough relative to Brooklyn. Several interaction terms are significant, so the room-type gap is not identical across boroughs. This supports Model 3 as a better predictive model, while Model 2 remains easier to interpret.

The final diagnostics are based on Model 3. The residuals are centered around zero, so the model captures the main mean structure. The Q-Q plot shows tail departures, which means the residuals are not fully normal. The scale-location plot suggests mild heteroscedasticity. The leverage plot shows a small number of influential observations, but most observations have low leverage. The VIF results stay below common concern thresholds, so severe multicollinearity is not evident. These diagnostics support using the model for broad pricing patterns, but not for very precise individual predictions.

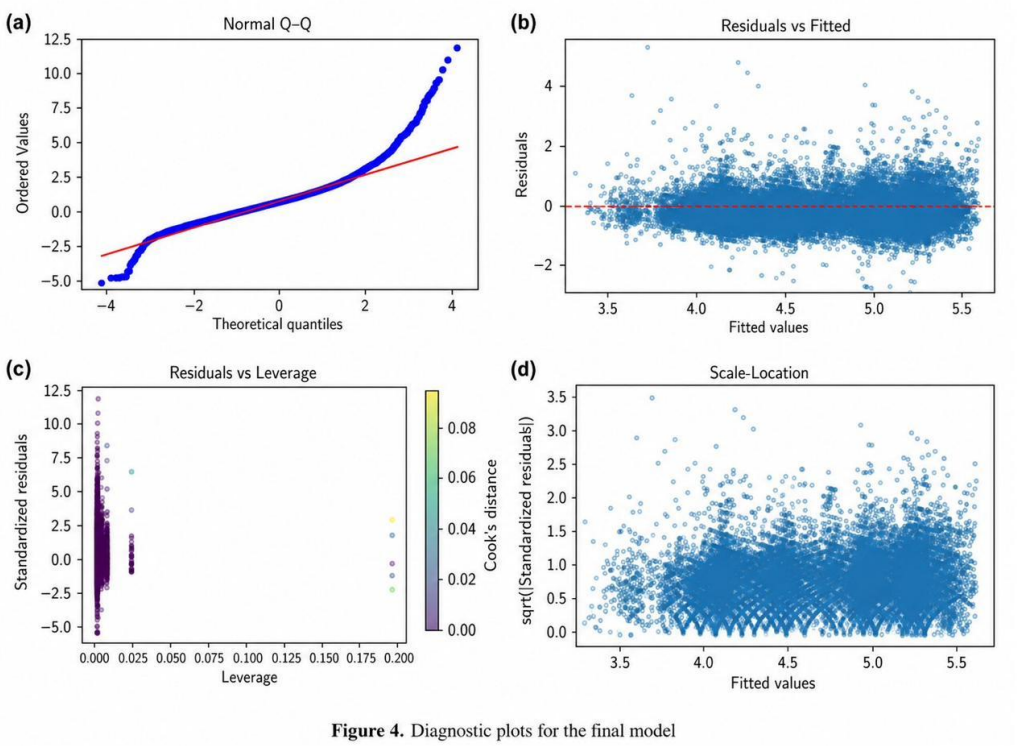

Figure 4. Diagnostic plots for the final model

For prediction, Model 3 is applied to a representative Manhattan entire-home listing with a three-night minimum, 20 reviews, 1.5 reviews per month, two host listings, and 180 available days per year. The predicted price is about $204.05. The 95% prediction interval is wide, about $81.86 to $508.60. This wide interval shows that the model can summarize the expected price level, but individual listing prices still contain large unexplained variation.

5. General Linear Model

As a second analysis, we use a logistic regression model. The response is a binary high-price indicator. A listing is coded as 1 if its price is above the median price of the cleaned sample and 0 otherwise. This model changes the question from predicting the level of price to predicting whether a listing belongs to the high-price group.

The logistic model uses the same main predictors as the OLS analysis: room type, borough, minimum nights, number of reviews, reviews per month, host listing count, and availability. The coefficients are reported as odds ratios. An odds ratio greater than 1 means that the predictor increases the odds that a listing is high-priced. An odds ratio less than 1 means that the predictor lowers those odds, holding the other variables fixed.

The logistic model serves as a check on the main OLS results. If the OLS findings are stable, entire-home listings and Manhattan listings should produce odds ratios above 1, while shared rooms and lower-price boroughs should produce odds ratios below 1. This gives a different but related view of the same pricing structure and avoids relying only on the conditional mean of log(price). The GLM result therefore reads as a classification-based confirmation of the main conclusion: room type and borough location are the main factors separating high-price listings from lower-price listings.

6. Conclusion

The results support the main hypothesis. Airbnb prices in New York City are primarily structured by room type and borough location. Entire-home listings have a large price premium over private rooms, and Manhattan has a clear location premium over Brooklyn. These patterns appear in the descriptive figures, remain strong in the OLS models, and are expected to appear again in the logistic regression through higher odds of being a high-price listing.

The quantitative listing variables provide some additional predictive power, but they do not change the main story. Model 3 improves the fit by allowing room-type effects to differ by borough and by reducing skewness in the numeric predictors. But the improvement is moderate. The model is better for explaining broad price patterns than for predicting the exact price of a single listing.

This study is limited by the available data. Listed price may not equal the paid price, and important features such as amenities, bedrooms, transit distance, seasonality, and local demand are missing. Future work should add finer location measures, amenity variables, and spatial methods. These additions would likely improve prediction and reduce the remaining unexplained variation.

References

Bernardi, M., & Guidolin, M. (2023). The determinants of Airbnb prices in New York City: A spatial quantile regression approach. Journal of the Royal Statistical Society Series C: Applied Statistics, 72(1), 104–138. https://doi.org/10.1093/jrsssc/qlad001

Inside Airbnb. (2019). New York City Airbnb listing data. Retrieved from http://insideairbnb.com/get-the-data

Kaggle. (2019). New York City Airbnb open data. Retrieved from https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55. https://doi.org/10.1086/260169