Airbnbs in New York City

Team lol: Tamsin Connerly, Hannah Lee, Jasmine Jiang

2025-03-17

Introduction

The rise of short-term rental platforms, particularly Airbnb, has significantly disrupted the traditional hospitality industry and transformed urban housing markets worldwide. In New York City, one of the world's most popular tourist destinations, the impact of Airbnb has been particularly pronounced, raising questions about its effects on local communities, housing affordability, and the broader urban economy.

Previous research has identified several factors that impact Airbnb pricing. One study found that host attributes, site and property attributes, amenities and services, rental rules, and online review ratings all play significant roles in determining listing prices (Wang and Nicolau 2017). Furthermore, recent studies have provided evidence of Airbnb's influence on housing markets. Another study found that a 1% increase in Airbnb listings leads to a 0.018% increase in rents and a 0.026% increase in house prices (Barron, Kung, and Proserpio 2018). This effect is more pronounced in areas with a lower share of owner-occupiers, suggesting that non-owner-occupiers are more likely to reallocate their properties from long-term to short-term rentals.

Our research question is: "How do various factors, such as bedroom number, room type, review scores, and neighborhood, influence the price of Airbnb listings in New York City?"

Price is the total price per night including fees (quantitative). Bedroom number is the total number of bedrooms in the rental (quantitative). Room type is whether the rental is a hotel room, entire home/apartment, private room, or shared room (categorical). Review score is the average review score of the rental from 1-5 stars (quantitative). Neighborhood is the borough of New York City that the rental is located in (categorical).

The Airbnb dataset that we are utilizing can be found on Inside Airbnb (https://insideairbnb.com/). Inside Airbnb has randomly collected data on dozens of countries and cities, but we decided to focus on New York City. The data was sourced from publicly available data on the Airbnb website on March 1, 2025.

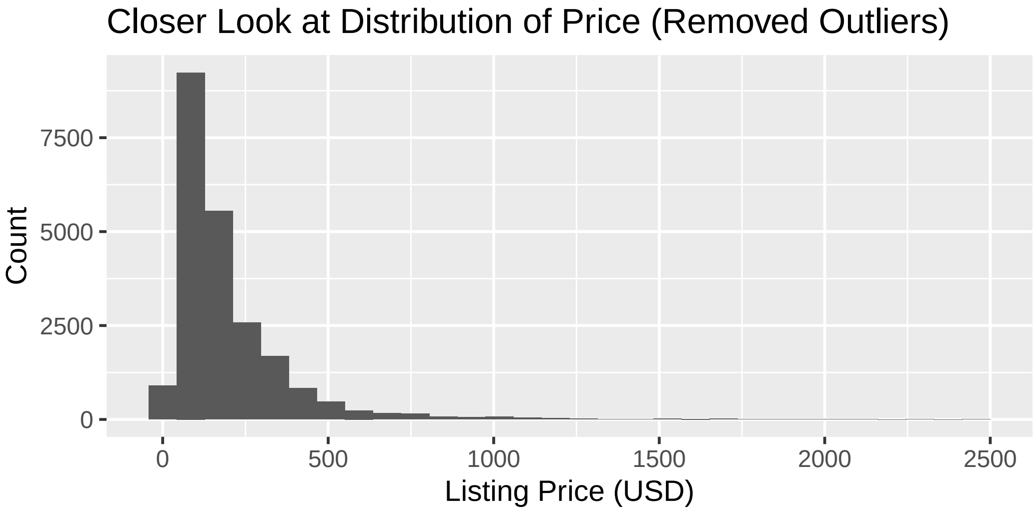
Understanding the determinants of Airbnb pricing in New York City is crucial for several reasons. Firstly, it can provide valuable insights for policymakers grappling with the challenges posed by the growth of short-term rentals, including potential impacts on housing affordability and neighborhood character (Toader et al. 2021). Secondly, it can help hosts make more informed pricing decisions, potentially leading to more efficient market outcomes.

Univariate Exploratory Data Analysis

Response Variable - Price

minimum	q1	median	mean	q3	maximum
7	85	140	213.835	240	20000

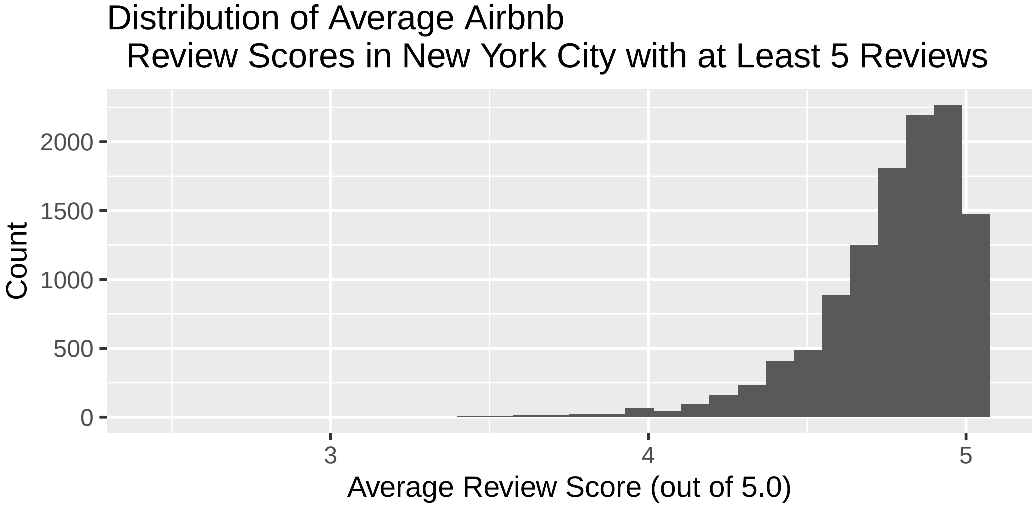
The distribution is pretty heavily right skewed. There is an outlier at $20,000 that impacts the mean, since the median of $140 is quite a bit less than the mean of around $213.84, and the mean is roughly equal to the 3rd quartile which is also around $240. The second highest price value is less than $1,000. We have removed this outlier for our analysis.



We can see that the distribution is still right skewed, and the vast majority of the listings seem to cost between $50-$200. Because of this skewedness, we also plan to apply log transformation to this variable to address the skew of the response variable.

Predictor Variable - Review Scores

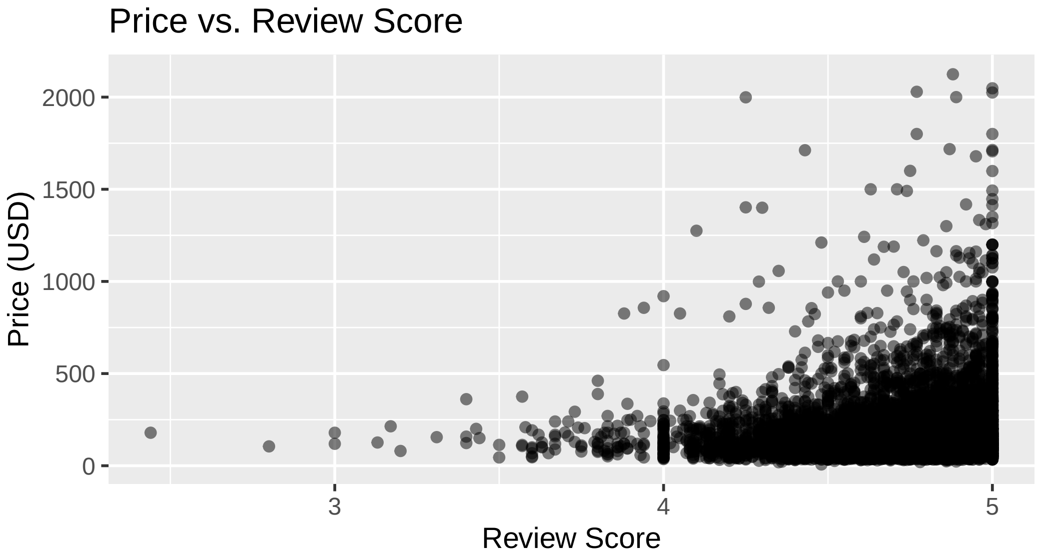
To account for the 6733 NA values for review scores, we will filter the dataset to include only listings with 5 or more reviews, since the median number of reviews for a listing is 5.



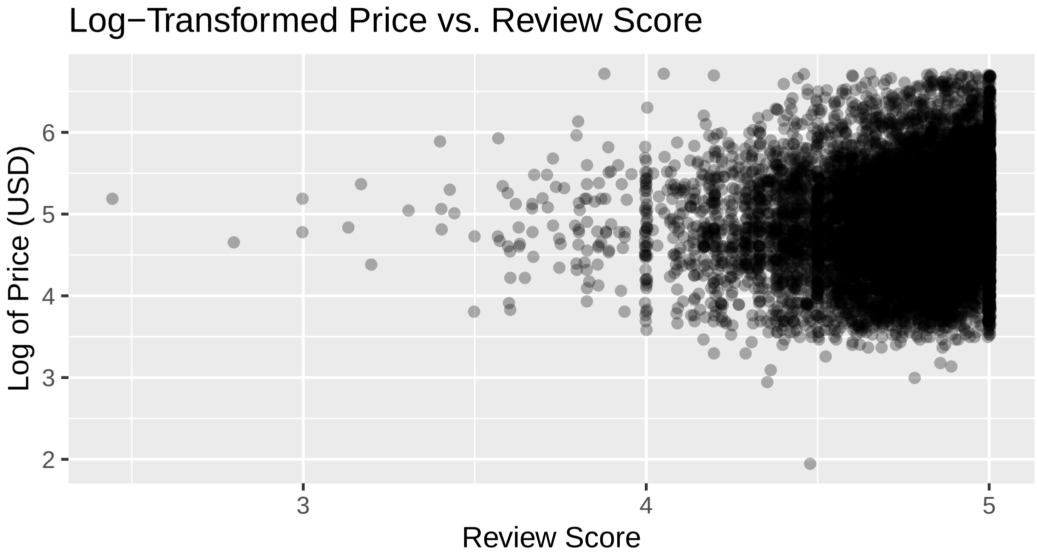
The distribution of review scores is skewed left after the transformation, with a median of 4.82 and a mean of 4.765. The minimum review score has increased from 1 to 2.44, and the third quartile review score has decreased from 5 to 4.93.

Bivariate Exploratory Data Analysis

Response (Price) vs Predictor Variable (review scores)

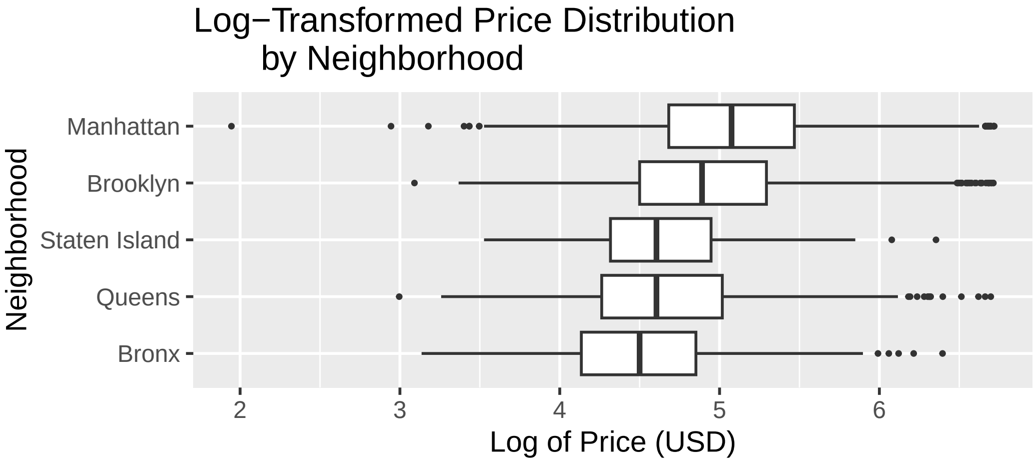


This scatter plot shows the relationship between price and review scores. However, it looks a little problematic and hard to interpret because high data density at certain score levels, especially between 4 and 5. Since the price variable is highly skewed, we applied a log transformation to try to help spread out values and make trends more visible.



The majority of listings have review scores between 4 and 5, this shows that most listings have scores within this range. Higher-rated listings tend to have slightly higher prices, but the effect is weak.

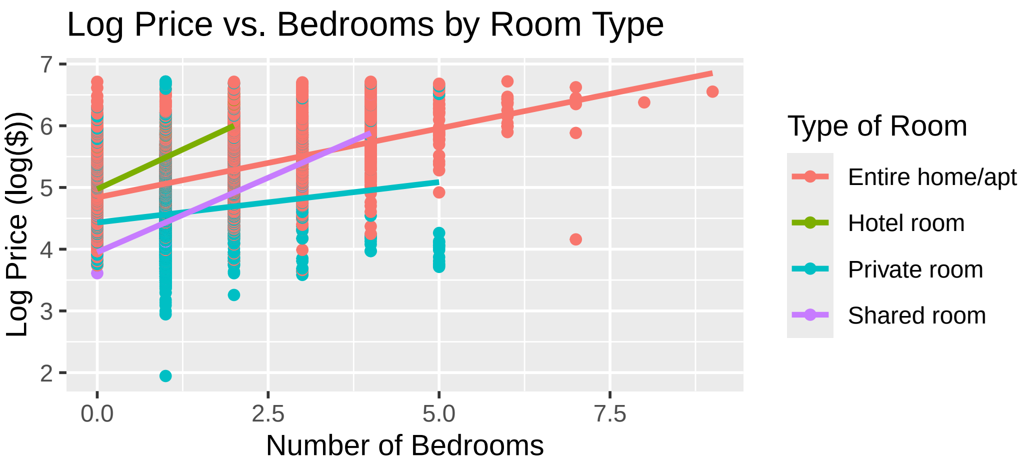
Response (Price) vs Predictor Variable (Neighborhood)



According to this plot, we can see that Manhattan has the highest median price, showing that it is the most expensive borough for Airbnb listings. It alsoexhibits the widest IQR, suggesting a high variation in listing prices. The median of Brooklyn follows Manhattan, with a slightly lower median price but still a wider spread. There are still some outliers shown in the plot, but the interpretability is much better.

Interaction Effects

Bedrooms and Room Type



Based on the graph, it appears that the rate at which price increases per number of bedrooms varies across room types. The slope of the shared room especially seems to differ from the others. Thus, there may be an interaction effect here.

Methodology

We chose to build a multiple linear regression model, as we wanted to use multiple explanatory variables to predict our continuous response variable: airbnb price.

To build the model, we kept used the log transformed price as the response variable due to our findings from the EDA above. The predictor variables we included in the model were the number of bedrooms, room type, review score ratings, and neighborhood. We decided to wanted to build a model that did not include the interaction between room type and number of bedrooms (this was the most obvious potential interaction effect from our EDA), and then build a second model that included the interaction and see which one performed better before selecting our final model.

Multiple Linear Regression With No Interaction

term	estimate	std.error	statistic	p.value
(Intercept)	3.204	0.098	32.533	0.000
bedrooms	0.241	0.005	44.060	0.000
room_typeHotel room	0.413	0.072	5.765	0.000

term	estimate	std.error	statistic	p.value
room_typePrivate room	-0.438	0.009	-46.361	0.000
room_typeShared room	-0.551	0.068	-8.130	0.000
review_scores_rating	0.271	0.020	13.382	0.000
neighbourhood_group_cleansedQueens	0.090	0.023	3.908	0.000
neighbourhood_group_cleansedStaten Island	0.014	0.040	0.342	0.732
neighbourhood_group_cleansedBrooklyn	0.253	0.022	11.546	0.000
neighbourhood_group_cleansedManhattan	0.495	0.022	22.457	0.000

Multiple Linear Regression with Interaction Effects

term	estimate	std.error	statistic	p.value
(Intercept)	3.171	0.098	32.237	0.000
bedrooms	0.255	0.006	43.337	0.000
room_typeHotel room	0.205	0.191	1.073	0.284
room_typePrivate room	-0.316	0.020	-15.949	0.000
room_typeShared room	-0.828	0.130	-6.377	0.000
review_scores_rating	0.273	0.020	13.539	0.000
neighbourhood_group_cleansedQueens	0.089	0.023	3.858	0.000
neighbourhood_group_cleansedStaten Island	0.020	0.040	0.505	0.614
neighbourhood_group_cleansedBrooklyn	0.253	0.022	11.606	0.000
neighbourhood_group_cleansedManhattan	0.497	0.022	22.572	0.000
bedrooms:room_typeHotel room	0.205	0.169	1.210	0.226
bedrooms:room_typePrivate room	-0.108	0.015	-7.022	0.000
bedrooms:room_typeShared room	0.251	0.099	2.540	0.011

Evaluating the Models

To evaluate both models' performance, we decided to take a look at their r^2 and adjusted r^2 values to see which model had higher values.

r.squared	adj.r.squared
0.387	0.387

r.squared	adj.r.squared
0.391	0.39

The model that includes the interaction effect between bedrooms and room type seems to perform slightly better, with higher r^2 and adjusted r^2 values of 0.391 and 0.39, respectively. To further validate this, we wanted to conducted a drop in deviance test to see if there was a difference between the two models' performance.

Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)
11313	2574.040	NA	NA	NA
11310	2560.848	3	13.192	0

The results from the drop in deviance test also support this, as the p-value is less than the threshold, indicating that including the interaction effect is able to significantly improve the model's ability to explain variation in log price. As a result, we decided to select the model with the interaction effect as our final model.

Predictor	VIF
bedrooms	1.077
room_typeHotel room	1.010
room_typePrivate room	1.091
room_typeShared room	1.006
review_scores_rating	1.040
neighbourhood_group_cleansedQueens	4.048
neighbourhood_group_cleansedStaten Island	1.347
neighbourhood_group_cleansedBrooklyn	5.659
neighbourhood_group_cleansedManhattan	5.573

Predictor	VIF
bedrooms	1.256
room_typeHotel room	7.202
room_typePrivate room	4.821
room_typeShared room	3.703
review_scores_rating	1.040
neighbourhood_group_cleansedQueens	4.048
neighbourhood_group_cleansedStaten Island	1.347
neighbourhood_group_cleansedBrooklyn	5.659
neighbourhood_group_cleansedManhattan	5.574

Predictor	VIF
bedrooms:room_typeHotel room	7.191
bedrooms:room_typePrivate room	4.574
bedrooms:room_typeShared room	3.695

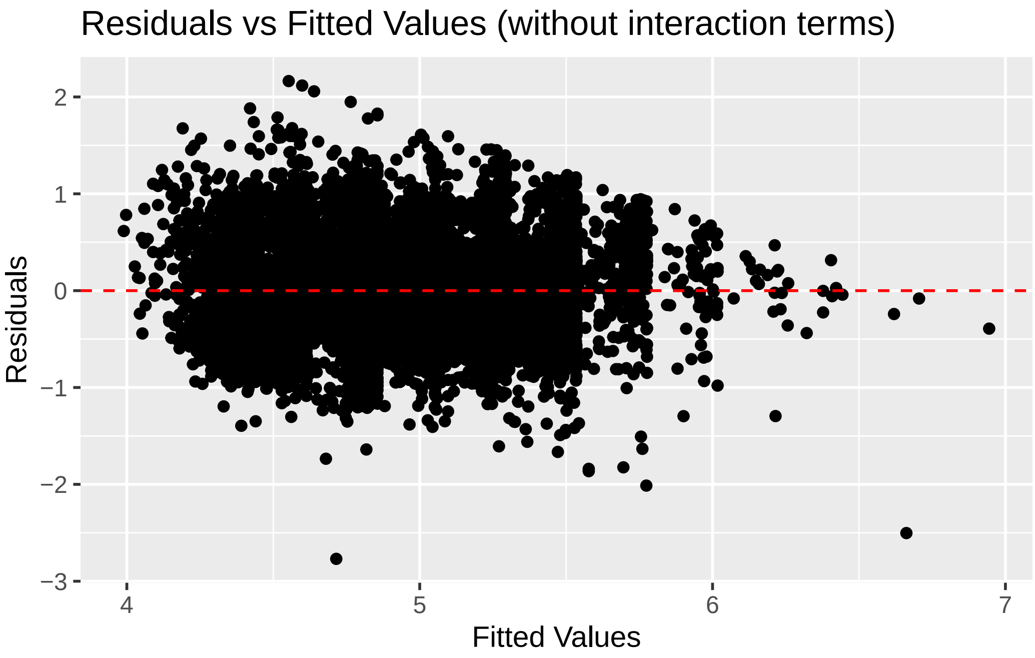
To assess multicollinearity, we calculated Variance Inflation Factors (VIFs) for both the two models. In the model without interaction terms, all predictors had VIF values below 6. The highest values were observed for the neighborhood dummy variables neighbourhood_group_cleansedBrooklyn (VIF = 5.659) and neighbourhood_group_cleansedManhattan (VIF = 5.573), indicating moderate multicollinearity, but still within acceptable limits. All other predictors had VIFs close to 1, suggesting low collinearity.

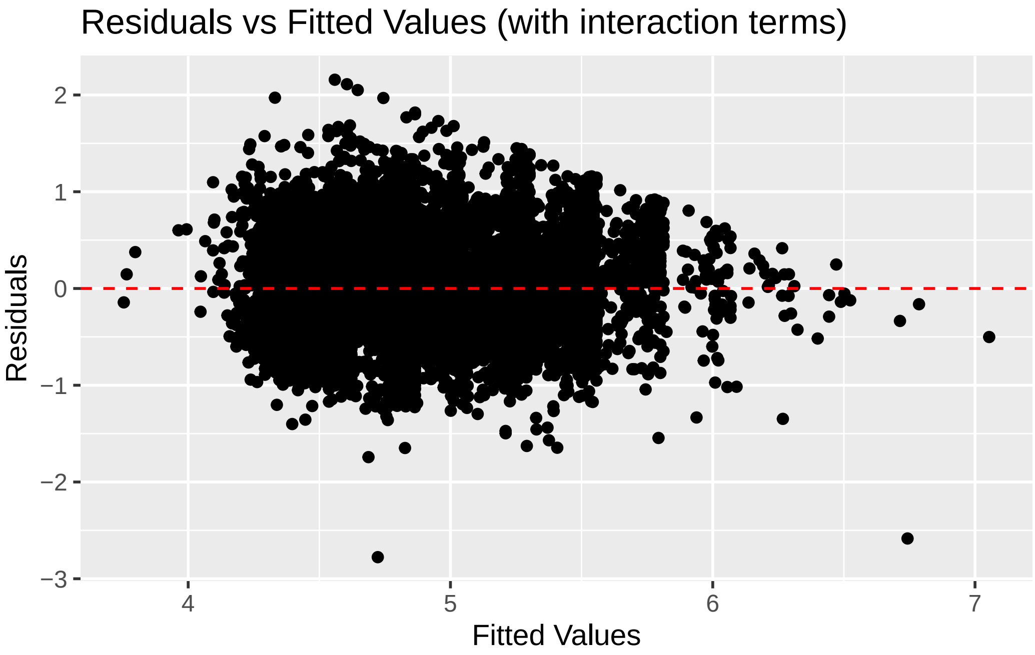
In the interaction-effects model, VIF values slightly increased, particularly for `room_typeHotel room` (VIF = 7.202) and its interaction with `bedrooms` (VIF = 7.191). This is expected due to the inclusion of interaction terms, which can introduce redundancy and inflate variance when the interacting variables are correlated or when one category has relatively fewer observations. `room_typePrivate room` and its interaction term also showed moderately elevated VIFs (4.821 and 4.574 respectively). However, **n**one of the predictors exceeded the common VIF threshold of 10, indicating that severe multicollinearity is not present in either model.

Results

After selecting the interaction effect model to be our final model, we checked model assumptions, diagnostics, and model fit statistics.

Assumption Check



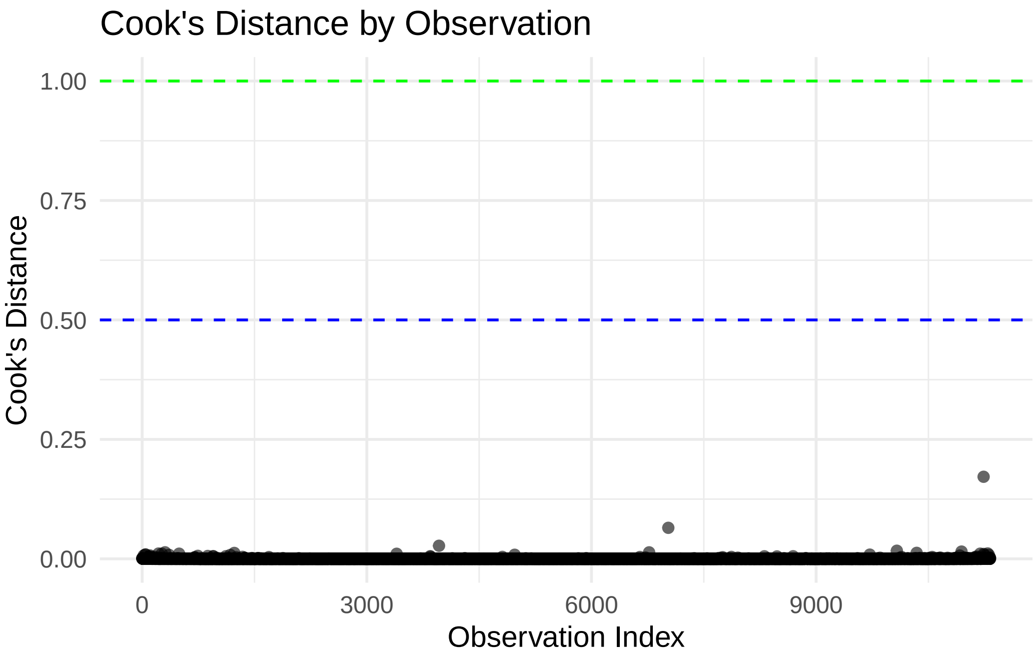


According to the residual plots, for both models without and with interaction terms, the residuals are centered around 0, which suggests that linearity is satisfied. However, both plots show a mild equal variance because there's a slight funnel shape – residuals seem more spread out at lower fitted values and slightly tighter at higher fitted values. There are also few vertical lines of the residuals, indicating discrete fitted values, likely due to categorical variables like room type or neighborhood.

The normality is fine because our dataset has more than 10,000 data points, which is large enough (n > 30) to satisfy the normality assumption.

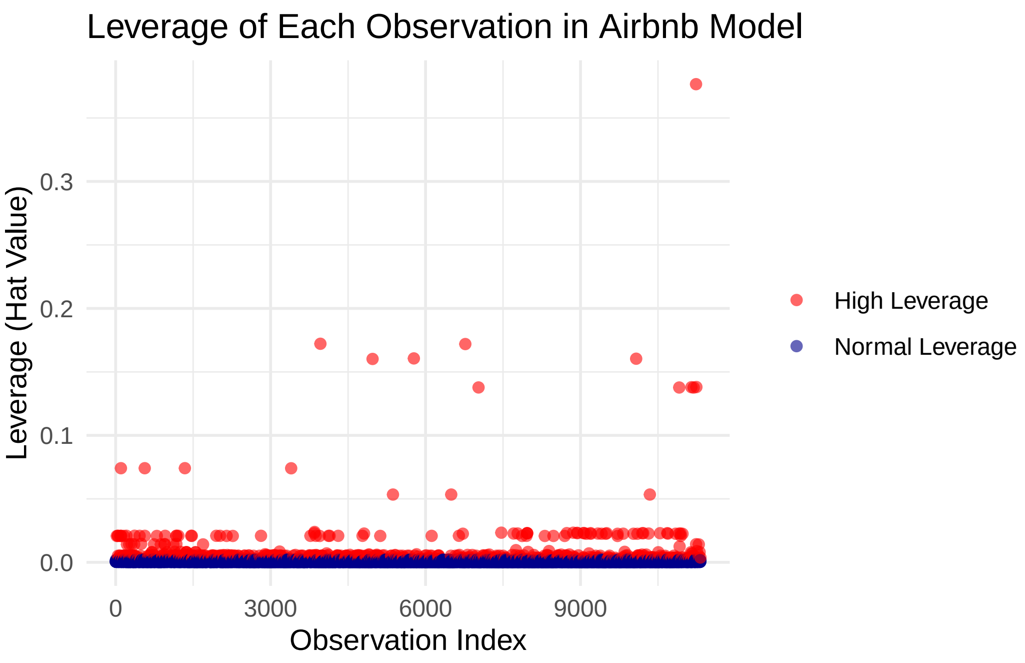
The independence is reasonably satisfied because each row in the dataset represents a distinct Airbnb listing, and the data is randomly collected. There is no indication of temporal or spatial autocorrelation, and there are no repeated measurements from the same listing.

Model Diagnostics



.rownames	log_price	bedrooms	room_type	review_scores_rating	neighborhood_group_cleansed	.fitted	.resid	.hat	.sigma	.cooksd	.std.resid

We used Cook's Distance to assess whether any individual observations had an undue influence on the overall model fit. As shown in the "Cook's Distance by Observation" plot, none of the observations exceed the commonly used thresholds of 0.5 (blue line) or 1 (green line). Additionally, the table listing points with Cook's Distance greater than 0.5 returned no results. It conforms that there are no overly influential data points in our dataset. Therefore, our fitted model is stable and not disproportionately affected by any single observation.



We also examined leverage values (the hat values) to identify observations with unusual combinations of predictor values. We calculated the leverage using the standard threshold:

$$\text{Leverage Threshold} = \frac{2(p+1)}{n}$$

We used the plot to visualize the leverage values for each observation, distinguishing between "High Leverage" (red) and "Normal Leverage" (blue) points. While a number of high leverage points were identified, they do not correspond to high Cook's Distance values, meaning they are not both unusual and influential. These points do not exert strong pull on the model's regression coefficients. Therefore, our model is robust, with no extreme outliers distorting the regression estimates.

Model Fit Statistics

To evaluate how well our model explains variation in Airbnb listing prices, we compared two multiple linear regression models: one without and one with interaction terms between room type and number of bedrooms. The final model we chose is the model includes the interaction effects, which shows slightly better performance based on both the r-squared:

$$R^2 = 0.391$$

$$\text{Adj } R^2 = 0.390$$

These values indicate that approximately 39% of the variability in log-transformed Airbnb prices is explained by our model, which includes predictors such as room type, number of bedrooms, neighborhood, review scores, and their interactions.

Additionally, we conducted a drop-in-deviance test to validate the improvement from including the interaction terms. According to the results we got, the full model provides a statistically significant improvement in explaining the variability in price because it has a p-value smaller than 0.001.

While the r-squared is moderate, this is under our expectation given the complexity and variability in Airbnb pricing. Other unobserved factors (such as amenities, host reputation, seasonal demand, or listing descriptions) likely contribute to price fluctuations and are not captured in this dataset.

Barron, Kyle, Edward Kung, and Davide Proserpio. 2018. "The Sharing Economy and Housing Affordability: Evidence from Airbnb." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3006832.

Toader, Valentin, Adina Letiţia Negruşa, Oana Ruxandra Bode, and Rozalia Veronica Rus. 2021. "Analysis of Price Determinants in the Case of Airbnb Listings." *Economic Research-Ekonomska Istraživanja* 35 (1): 2493–2509. https://doi.org/10.1080/1331677x.2021.1962380.

Wang, Dan, and Juan L. Nicolau. 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 (April): 120–31. https://doi.org/10.1016/j.ijhm.2016.12.007.