Airbnbs in New York City

Team lol: Tamsin Connerly, Hannah Lee, Jasmine Jiang

2025-03-17

title: "Airbnbs in New York City" author: "Team lol: Tamsin Connerly, Hannah Lee, Jasmine Jiang" date: "3/17/25" format: pdf execute: warning: false message: false echo: false editor: visual bibliography: references.bib

---

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 (**barron2018a?**). 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?"

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 (**toader2021a?**). Secondly, it can help hosts make more informed pricing decisions, potentially leading to more efficient market outcomes.

Based on existing literature and our understanding of the New York City housing market, we hypothesize that:

- Listings with more bedrooms will command higher prices, reflecting the premium placed on space in urban environments.
- The type of room (entire home/apartment vs. private room) will significantly impact pricing, with entire homes/apartments having a higher price.
- Higher review scores will be associated with higher prices, as positive feedback may justify premium pricing.
- Properties in more affluent neighborhoods like Manhattan will have higher prices compared to less affluent ones like the Bronx because of real estate price differences in each borough.

Exploratory Data Analysis

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.

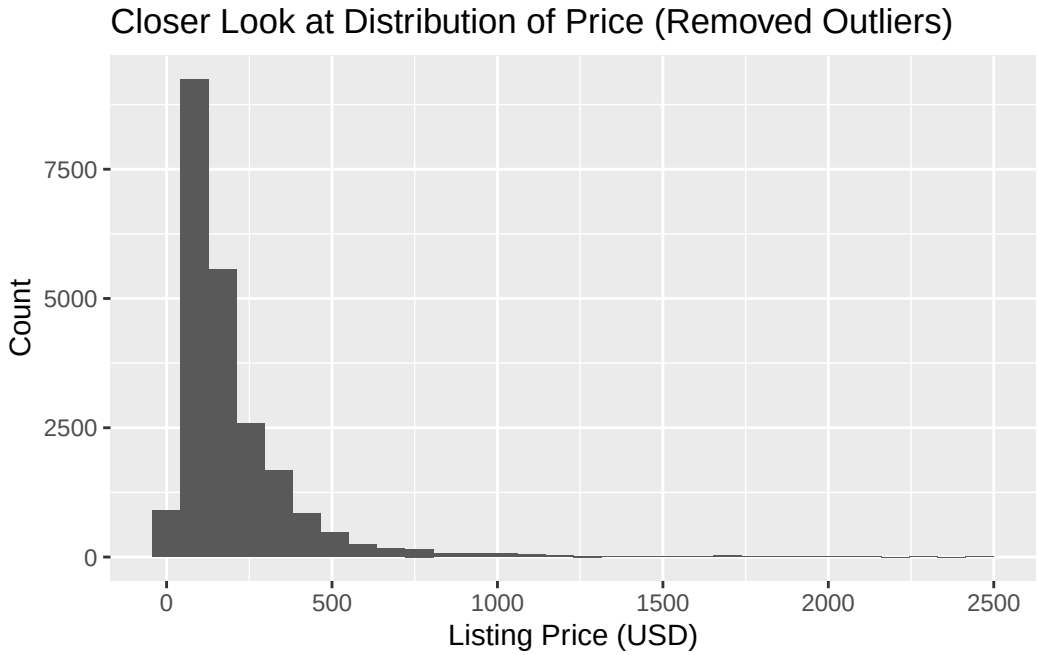
Each row in the dataset represents a unique Airbnb listing in New York City. Each of these correspond to individual properties available for rental on the platform and have many (58) variables such as name of the listing, latitude and longitude, room type, price, minimum number of nights required for booking, total number of reviews the listing has reviewed, and more.

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, as can be seen from both of the histograms. It is difficult to analyze the distribution in the first histogram because there is an outlier at $20,000 and makes the bins and binwidth very narrow and zoomed out (since the range of the data is too large). It is also clear that this outlier 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. 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 - Bedrooms

minimum	q1	median	mean	q3	maximum	na
0	1	1	1.313	2	15	49

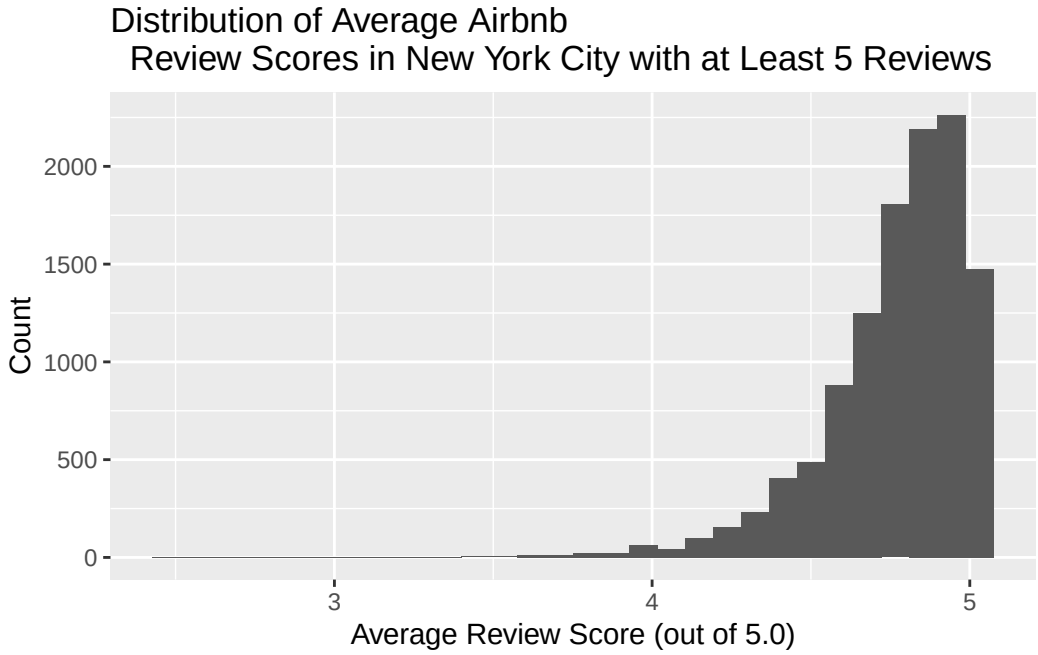
The distribution of the number of bedrooms for Airbnb listings in New York City is skewed to the right, as the mean is greater than the median. Since there are 17 NA values, we will drop them for our analysis.

Predictor Variable - Review Scores

minimum	q1	median	mean	q3	maximum	na
1	4.66	4.85	4.724	5	5	6733

The distribution of the predictor variable review score is skewed left, with the median of 4.85 being slightly higher than the mean of 4.724. The majority of the reviews are around 5 (over 6000), and the vast majority of the observations are between 4 and 5, with very few of them being below 3. Additionally, there are 6733 NA values.

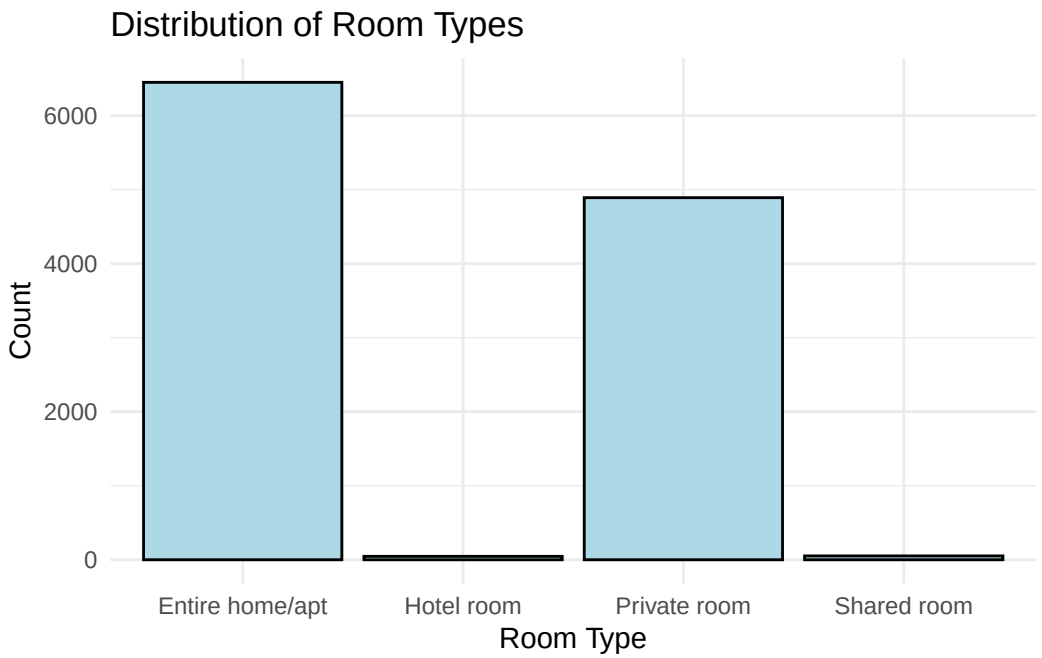
To account for the NAs, we will filter the dataset to include only listings with 5 or more reviews, since the median number of reviews for a listing is 5.



minimum	q1	median	mean	q3	maximum
2.44	4.67	4.82	4.765	4.93	5

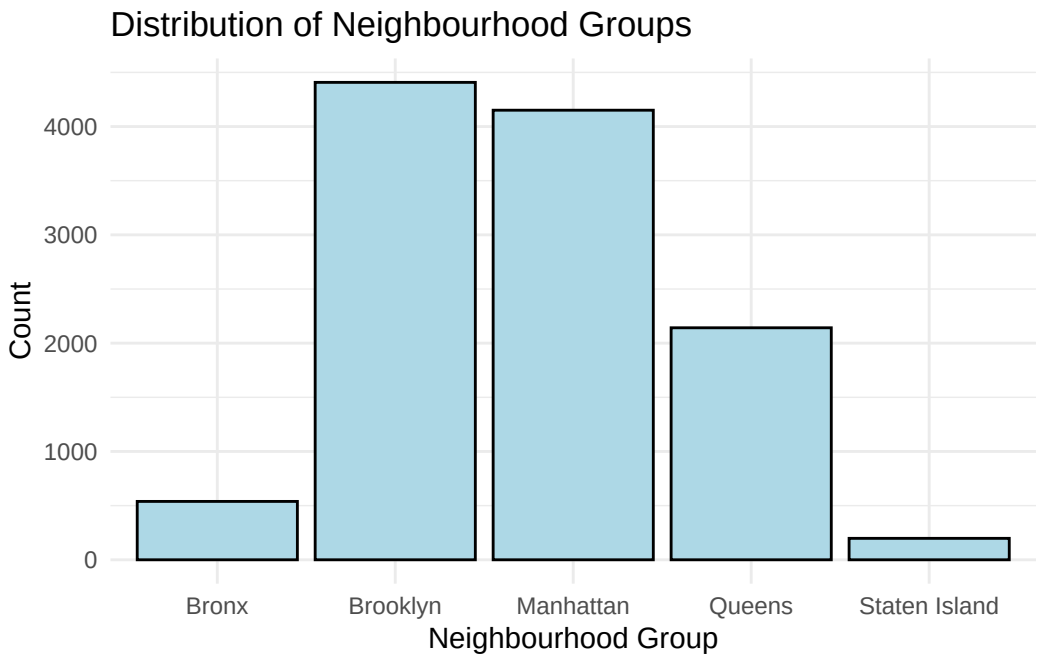
The distribution of review scores is still skewed left, with a median of 4.82 and a mean of 4.765. The minimum review score has increased from 1 to 2.44, and the third quartile review score has decreased from 5 to 4.93.

Predictor Variable - Room Type



The most frequent room type in this dataset is Entire home/apt, followed by private room; these may be more popular and sought out. There are very few hotel rooms and even fewer shared rooms.

Predictor Variable - Neighborhood

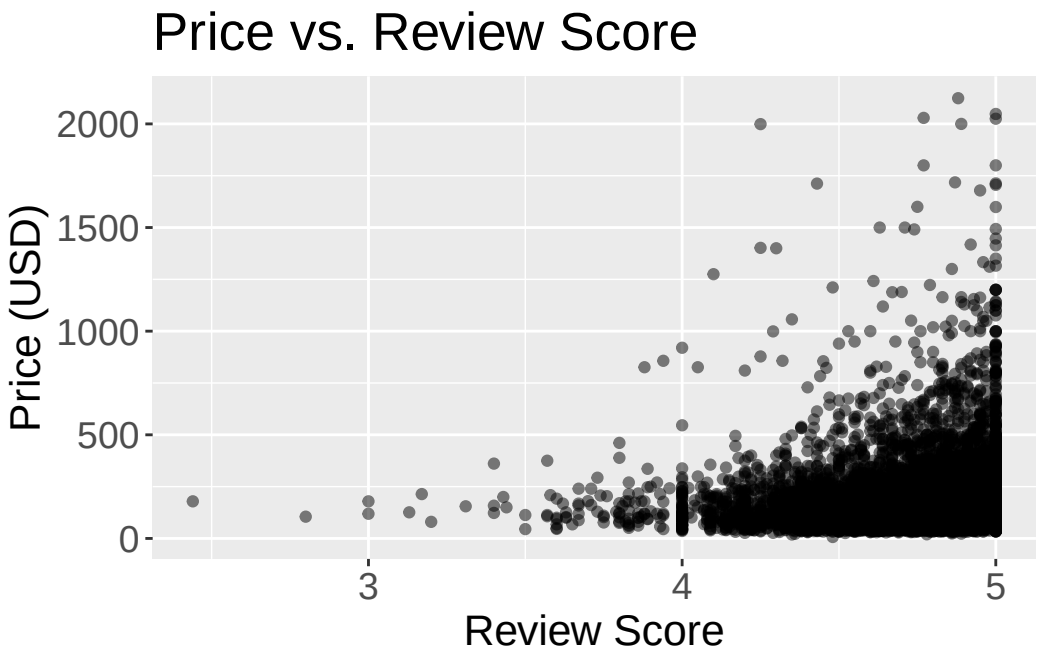


Var1	Freq
Bronx	539
Brooklyn	4408
Manhattan	4151
Queens	2142
Staten Island	198

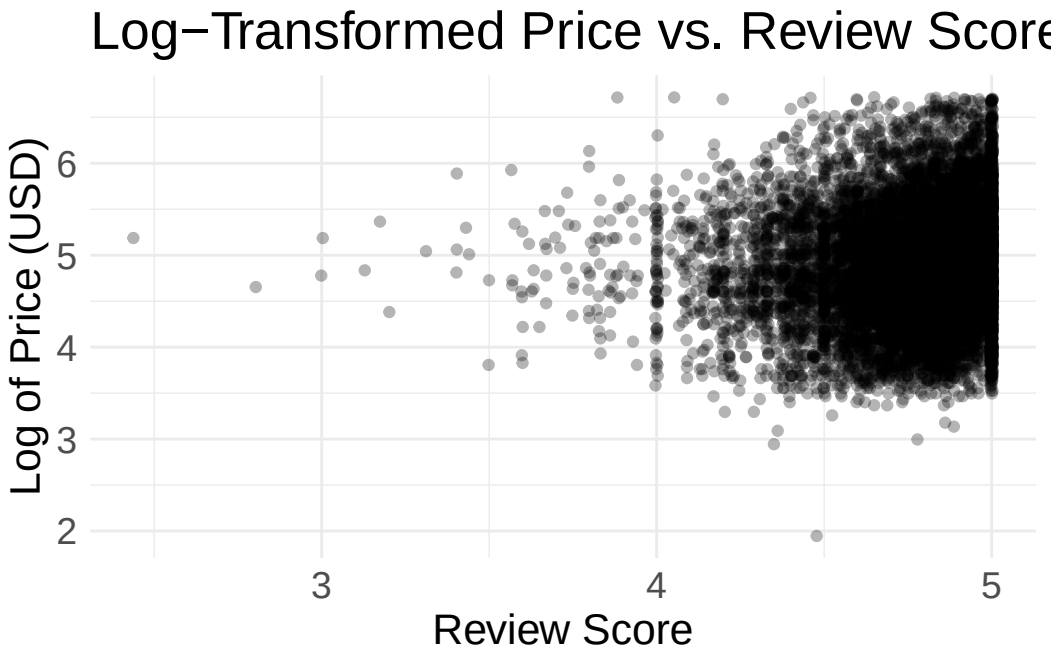
The greatest number listings are in Brooklyn (4408), followed by Brooklyn (4151), and Queens (2142). A few of them are in Bronx and even fewer in Staten Island. This is to be expected, as Manhattan and Brooklyn are prime areas for tourism and business, while other areas might be less popular for short-term rentals.

Bivariate Exploratory Data Analysis

Response (Price) vs Predictor Variable (review scores)

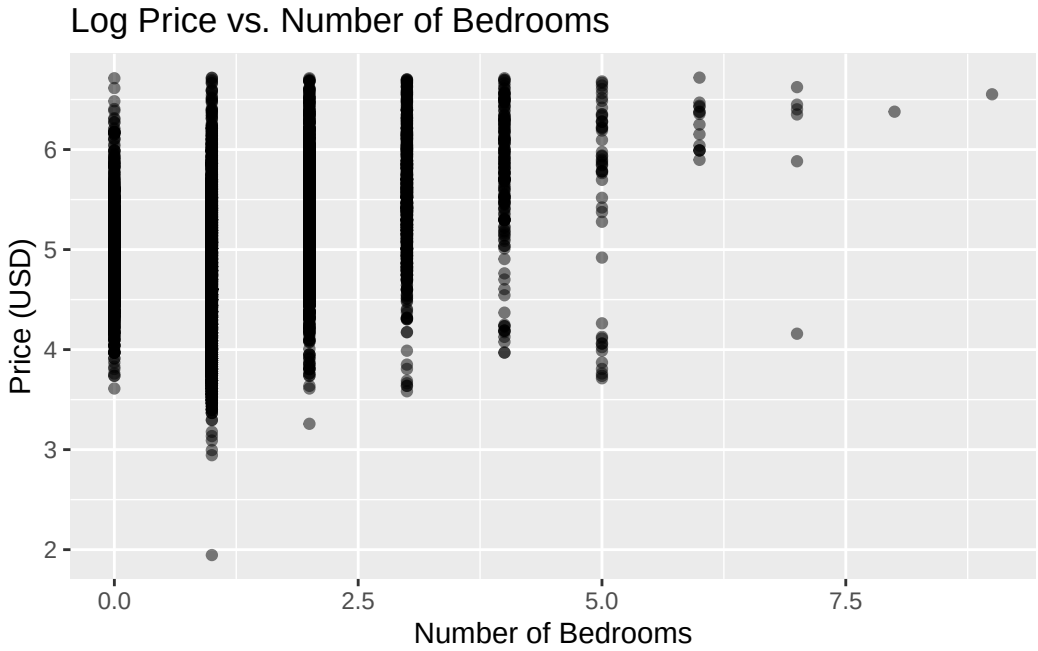


This scatter plot shows the relationship between price and review scores. However, it looks a little problematic and hard to interpret because high data density at certain score levels, especially between 4 and 5. Since the price variable is highly skewed, we applied a log transformation to try to help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



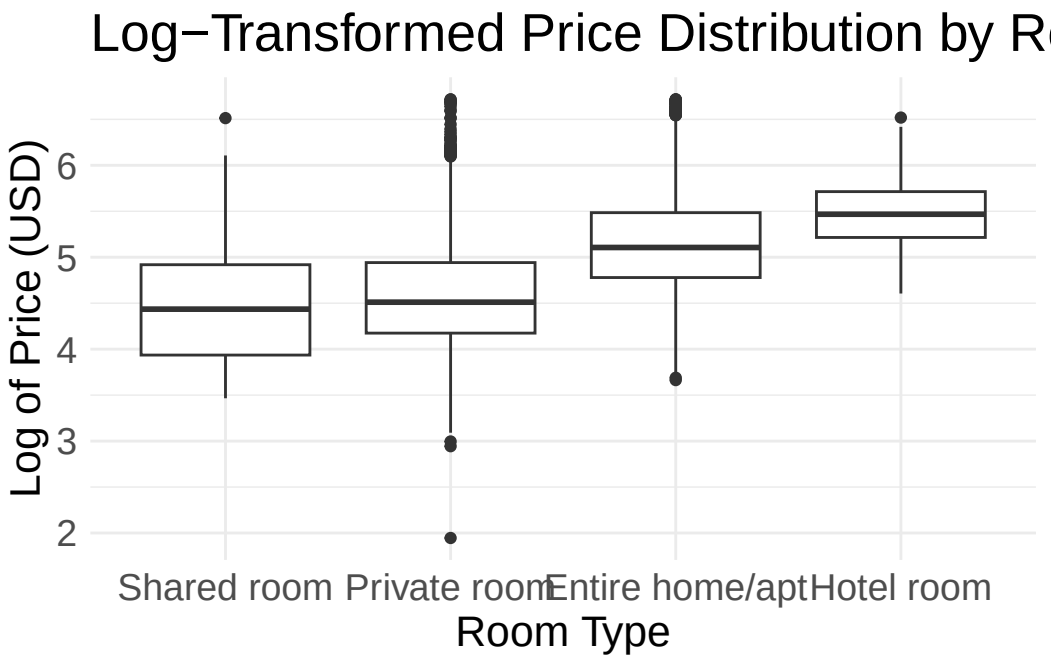
The majority of listings have review scores between 4 and 5, this shows that most listings have scores within this range. Higher-rated listings tend to have slightly higher prices, but the effect is weak.

Response (Price) vs Predictor Variable (number of prices)



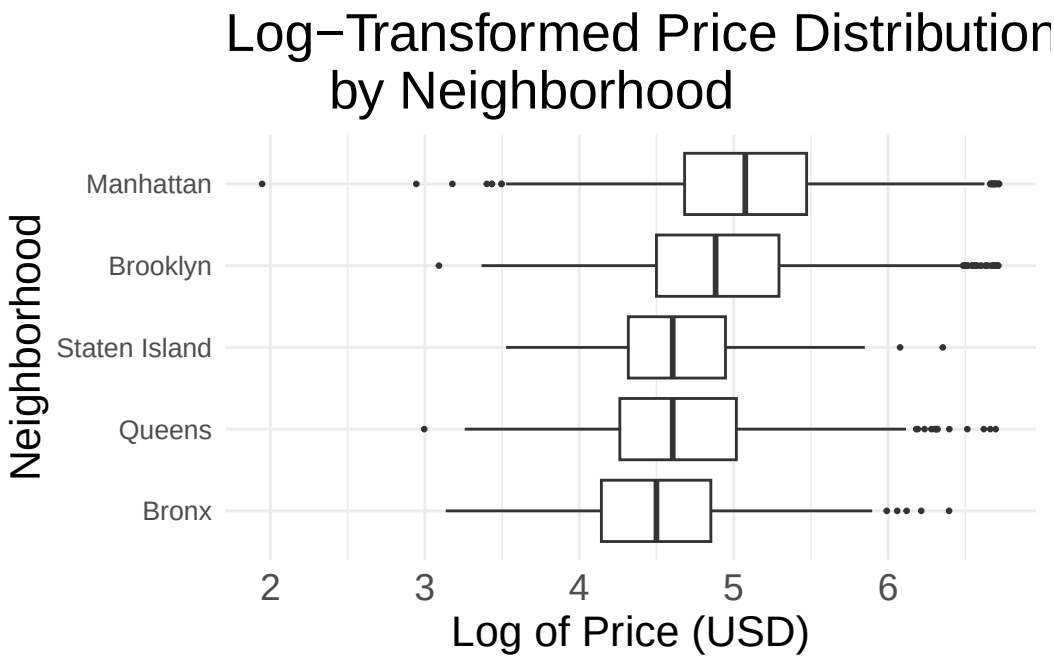
This scatter plot visualizes the relationship between the number of bedrooms and the price of Airbnb listings in New York City. There is not a clear trend. However, listings with 0 to 5 bedrooms exhibit a wide range of prices, with some listings priced significantly higher than the majority. Some outliers even exceed $10,000 per night, likely representing luxury or highly unique accommodations. This indicates that other factors such as location, room type, and review scores may have a stronger influence on pricing.

Response (Price) vs Predictor Variable (Room Type)



According to this plot, hotel rooms have the highest median price and general highest price among all room types, suggesting that they are generally priced higher than other Airbnb listings such as shared room and private room, etc. Also, the IQR for hotel rooms and entire homes/apartments is larger compared to private and shared rooms, indicating greater variation in pricing.

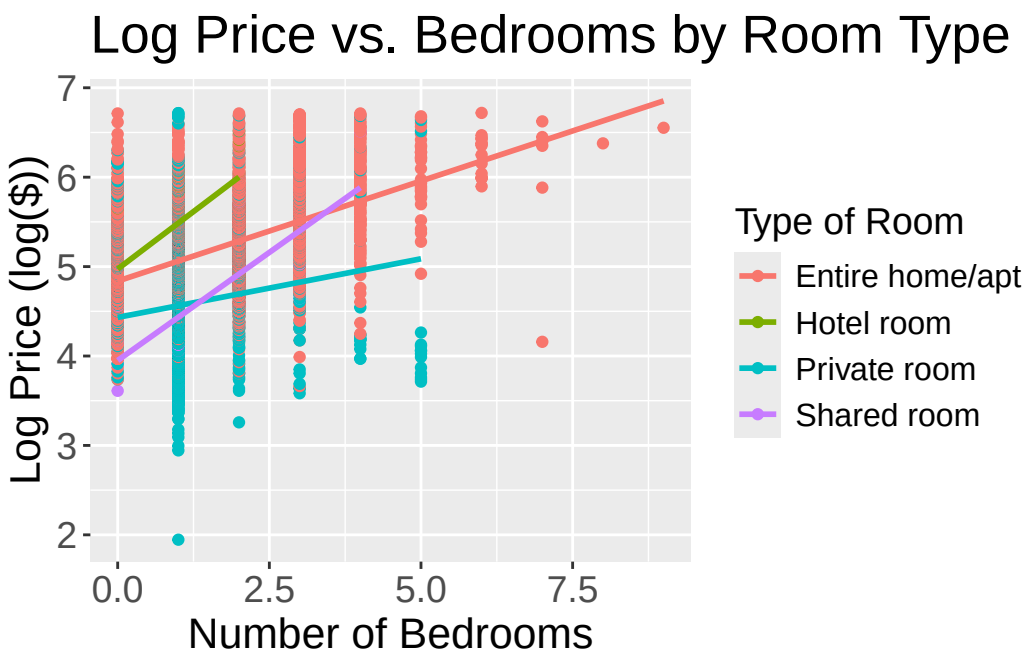
Response (Price) vs Predictor Variable (Neighborhood)



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.

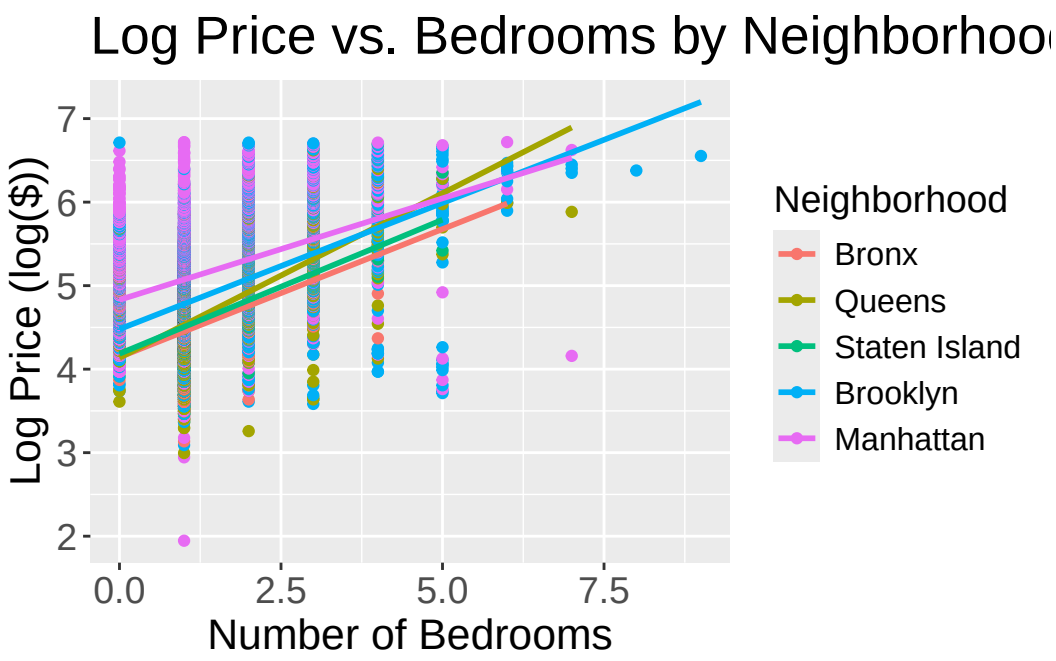
Interaction Effects

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.

Bedrooms and Neighborhood



Based on the graph visually, it appears that the rate at which price increases per bedroom does not greatly vary based on neighborhood, especially compared to the previous graph by room type. Staten Island does have a noticeably different slope, though. Thus, there is a potential for an interaction effect between bedroom number and neighborhood, but not as much as the previous graph with bedroom and room type.

Building the 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
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

term	estimate	std.error	statistic	p.value
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

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.

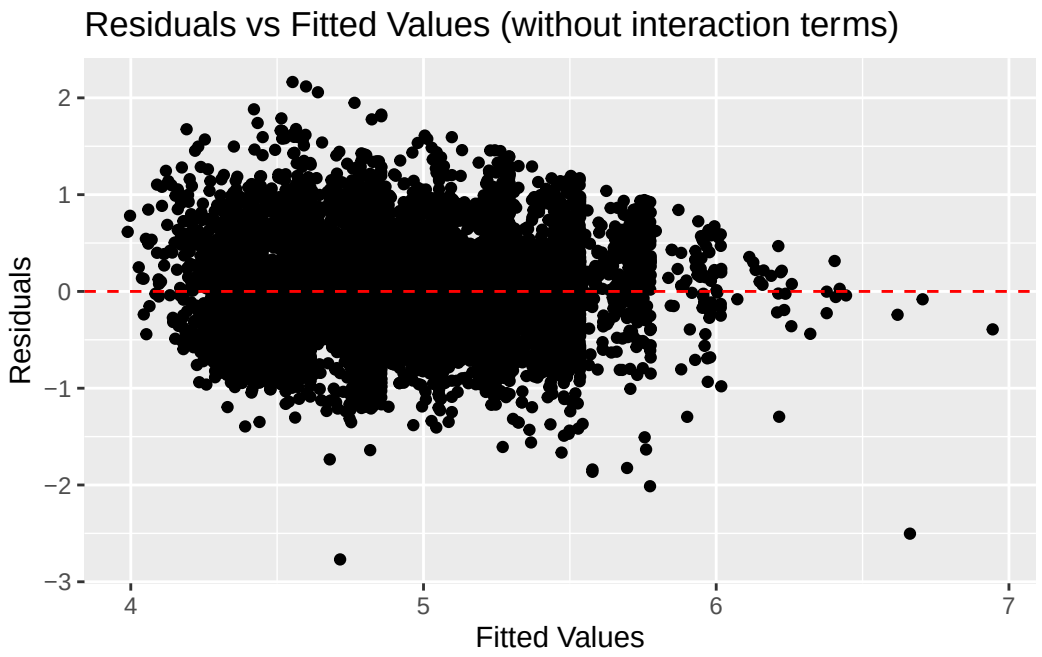
Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)
11313	2574.040	NA	NA	NA
11310	2560.848	3	13.192	0

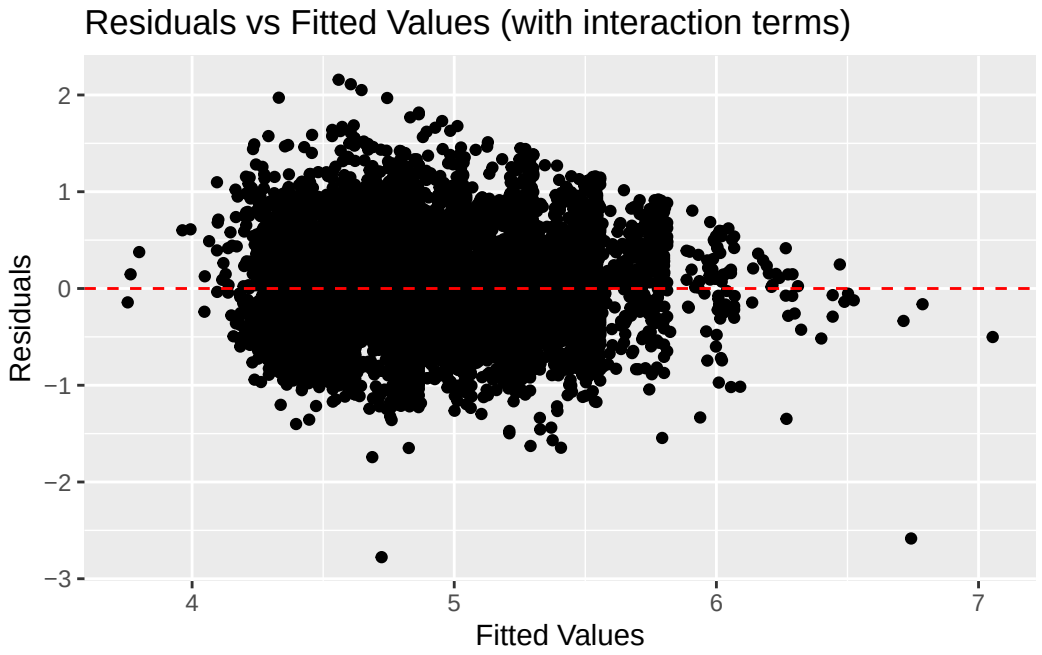
The results from the drop in deviance test also support this, as the p-value is less than the threshold, indicating that including the interaction terms is able to significantly improve the model's ability to explain variation in log price.

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

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.

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.