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 (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?"

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.

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. We are particularly interested in the following explanatory variables that we believe could impact the price of an Airbnb listing:

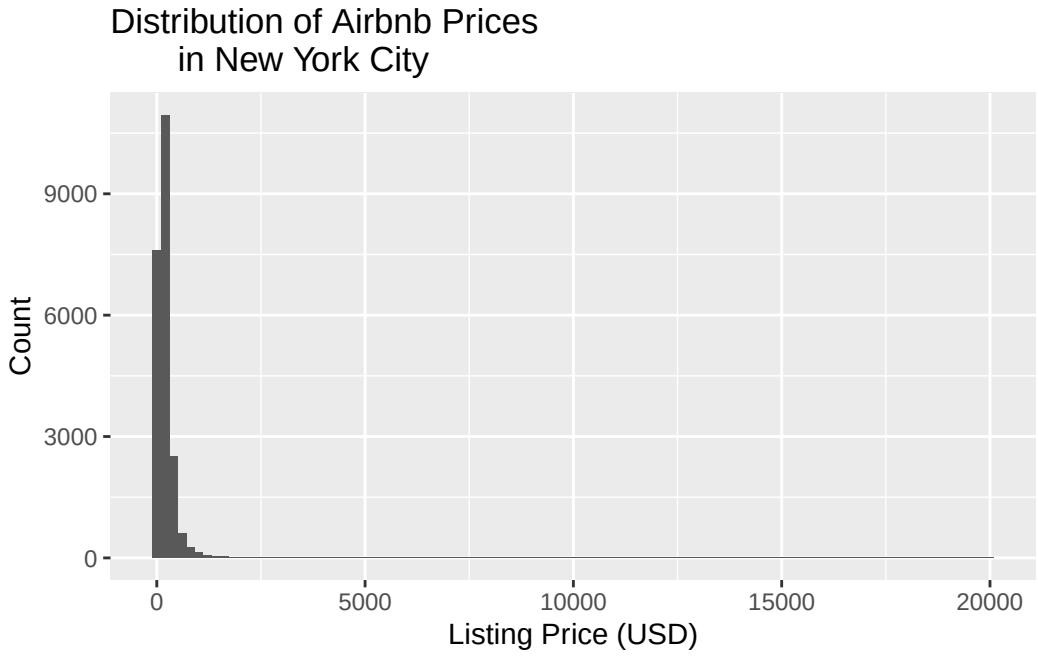
- Number of bedrooms
  - These could give insights into the size and comfort level of the Airbnb, likely affecting the price.
- Room type
  - Type of room (whether the listing is an entire home/apartment or a private room) can impact pricing
- Review scores (overall rating)

  - Listings with better or higher number of reviews could be priced higher because of higher perceived value and trustworthiness

- Neighborhood
  - Which New York City Neighborhood the Airbnb is located in. More affluent neighborhoods like Manhattan may have higher prices than neighborhoods like the Bronx.

Univariate Exploratory Data Analysis

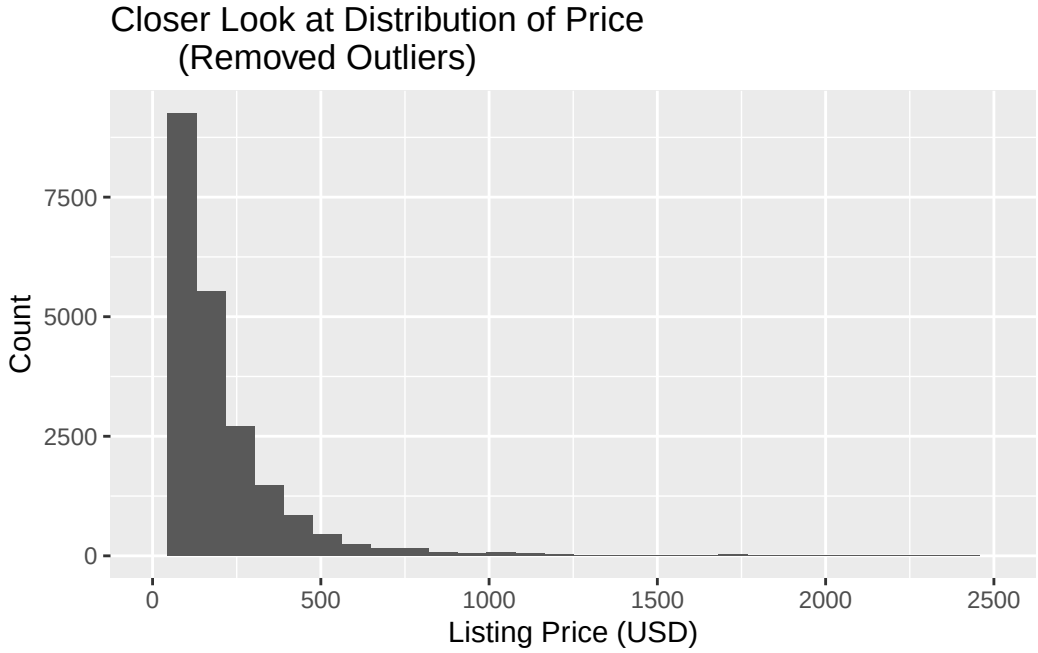
Response Variable - Price



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

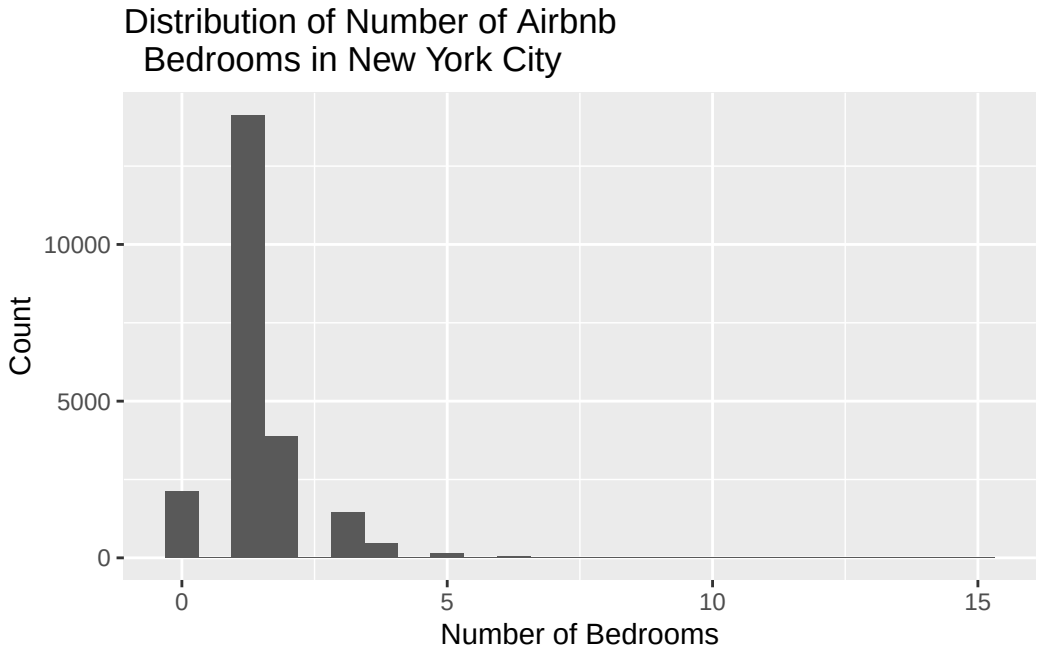
The distribution is pretty heavily right skewed, as can be seen from both of the histograms. It is difficult to analyze the distribution in the first histogram because there is an outlier at $20,000 and makes the bins and binwidth very narrow and zoomed out (since the range of the data is too large). 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 plan to to remove this outlier as a result when doing our analysis, and we will go into more depth later and check for additional outliers.



From the initial histogram, since the majority of the listings seem to fall between 0 and 2500 dollars, we visualized the distribution of the listings between this price range to get a better view of it. We can see that the distribution is still right skewed, and the vast majority of the listings seem to cost between $50-$200. There are a few more points visible that were not in the initial histogram that appear to be quite far from where the majority of the listings are clustered, so we will go back and check to see if they are greater than 1.5 x IQR and can be classified as outliers. We also plan to apply log transformation to this variable to address the skew of the response variable.

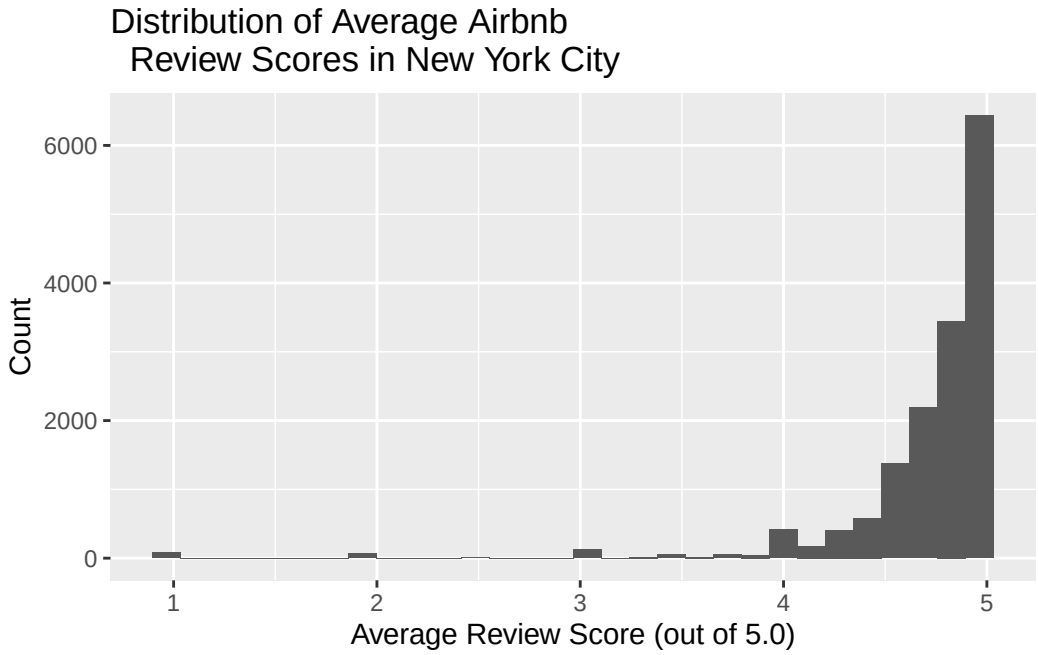
Predictor Variable - Bedrooms



The distribution of the number of bedrooms for Airbnb listings in New York City is skewed to the right. Most properties are one- or two-bedroom units. This aligns with the typical demand for smaller accommodations, which are more suited for individuals or small groups that might be more common. Listings with more bedrooms, such as 5 or more, are far less common. Larger apartments or homes, typically with 3 or more bedrooms, are less frequent, but they most likely consist of listings that may be priced much higher.

minimum	q1	median	mean	q3	maximum	na
0	1	1	1.315	2	15	49

Predictor Variable - Review Scores



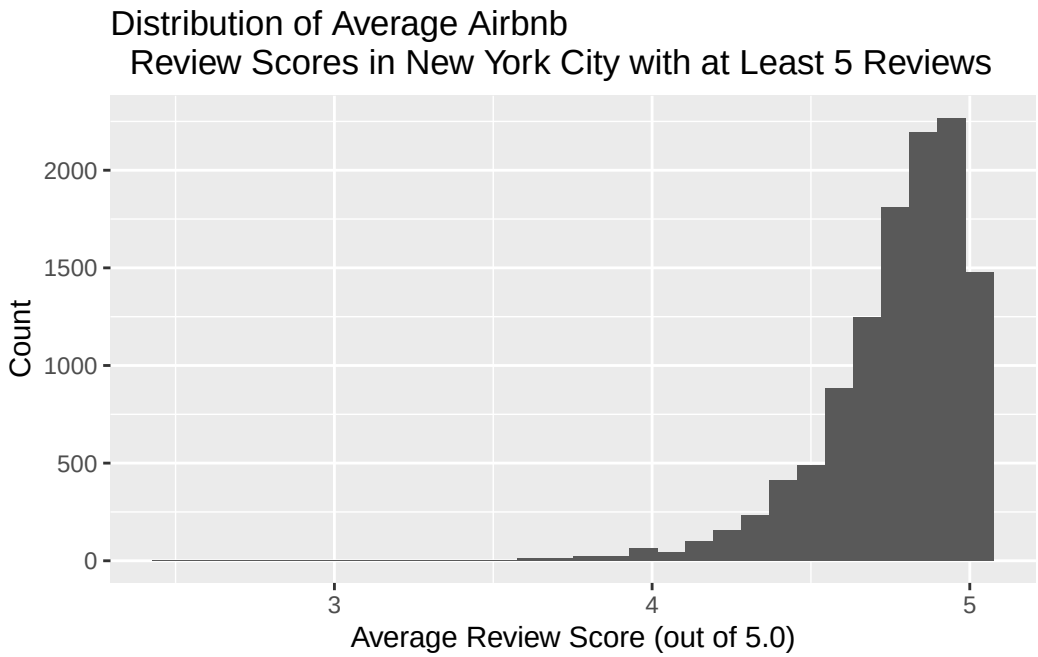
minimum	q1	median	mean	q3	maximum	na
1	4.662	4.85	4.724	5	5	6798

The distribution of the predictor variable "review score" is skewed left, with the median of 4.85 being slightly higher than the mean of 4.72. The majority of the reviews are around 5 (over 6000 of them), and the vast majority of the observations are between 4 and 5, with very few of them being below 3. Additionally, there are quite a few NA values for this variable (around a third of the rows have NAs).

Table of the distribution of the number of reviews

minimum	q1	median	mean	q3	maximum
0	0	5	34.353	35	2749

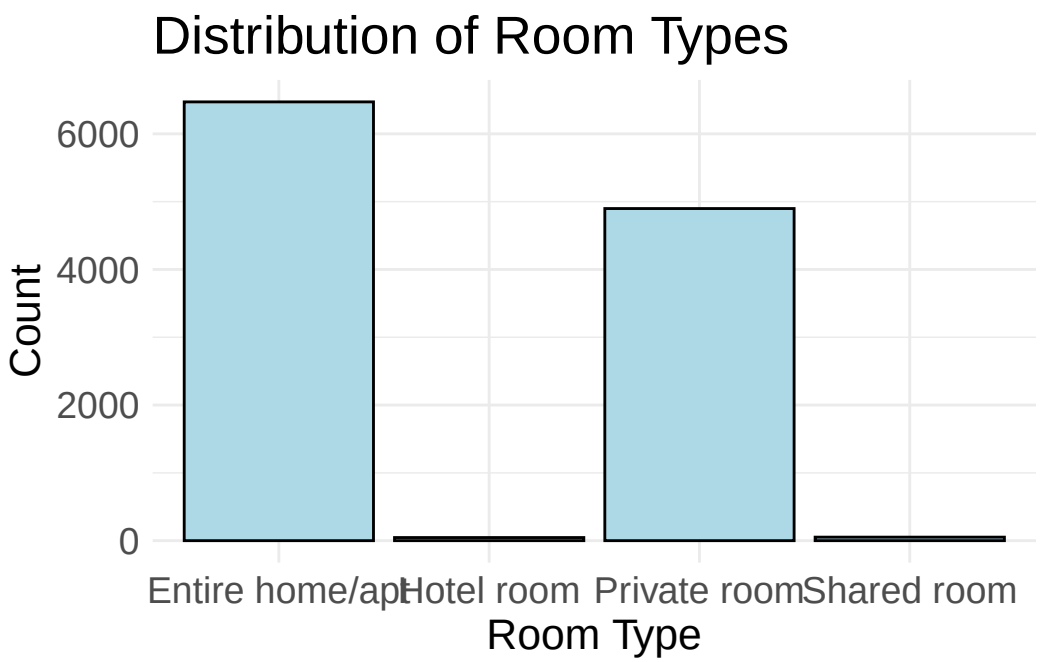
To account for the NAs, we will filter the dataset to include only listings with 5 or more reviews. The median number of reviews is 5, and the average is 34.35.



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

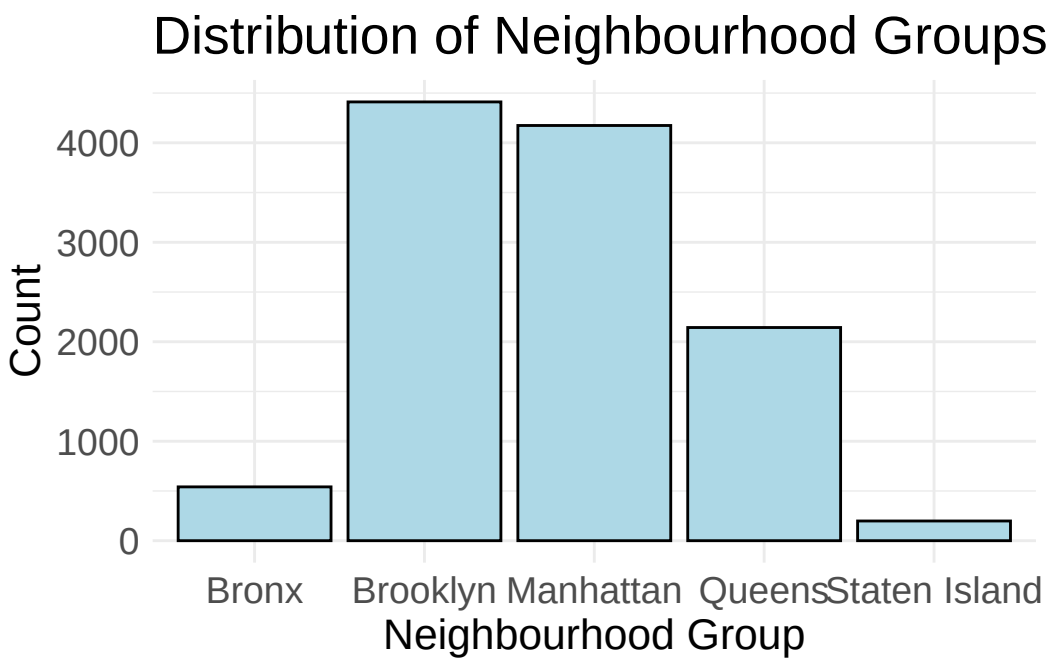
Predictor Variable - Room Type



Var1	Freq
Entire home/apt	6471
Hotel room	46
Private room	4898
Shared room	52

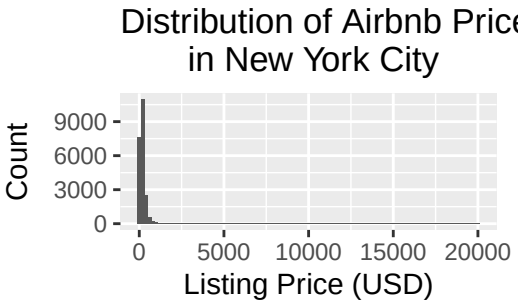
The mode, or most frequent room type in this dataset is Entire home/apt, followed by private room; these may be more popular and sought out, and might tend to be more expensive than the hotel or shared rooms. There are very few hotel rooms and even fewer shared rooms.

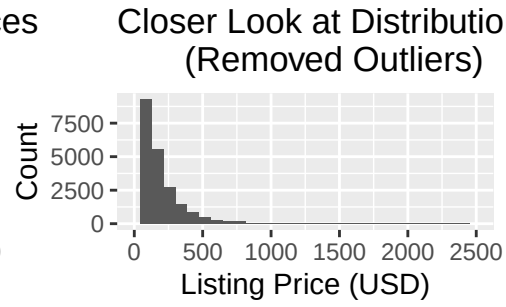
Predictor Variable - Neighborhood

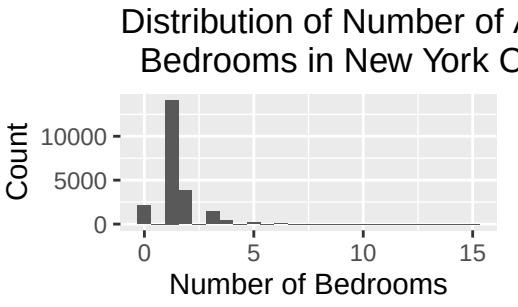


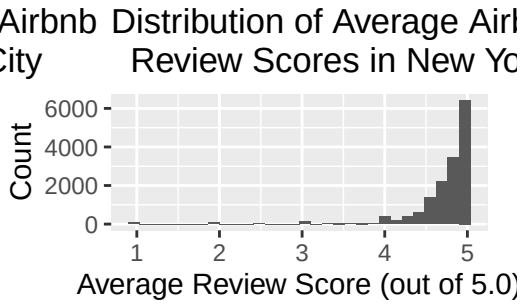
Var1	Freq
Bronx	541
Brooklyn	4411
Manhattan	4174
Queens	2143
Staten Island	198

The majority of the airbnb listings are in Manhattan (just over 10,000), followed by Brooklyn (around 7500), and Queens (a bit over 2500). 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.


Distribution of Airbnb Prices
in New York City
Count
9000
6000
3000
0
0
5000
10000
15000
20000
Listing Price (USD)

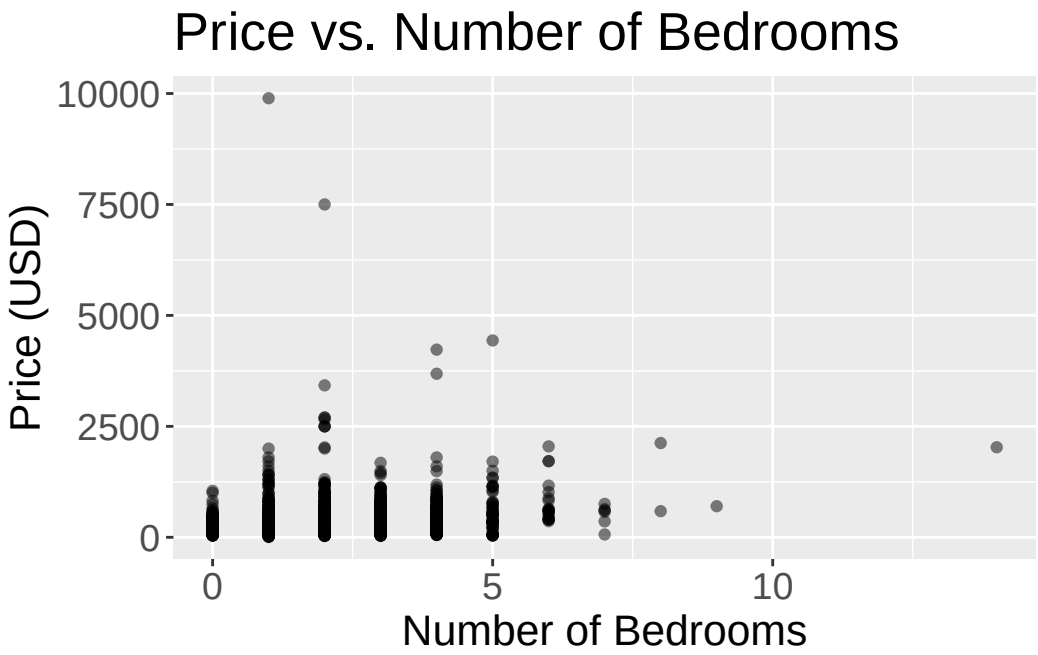

Closer Look at Distribution
(Removed Outliers)
Count
7500
5000
2500
0
0
500
1000
1500
2000
2500
Listing Price (USD)


Distribution of Number of Airbnb
Bedrooms in New York City
Count
10000
5000
0
0
5
10
15
Number of Bedrooms


Distribution of Average Airl
Review Scores in New Yo
Count
6000
4000
2000
0
1
2
3
4
5
Average Review Score (out of 5.0)

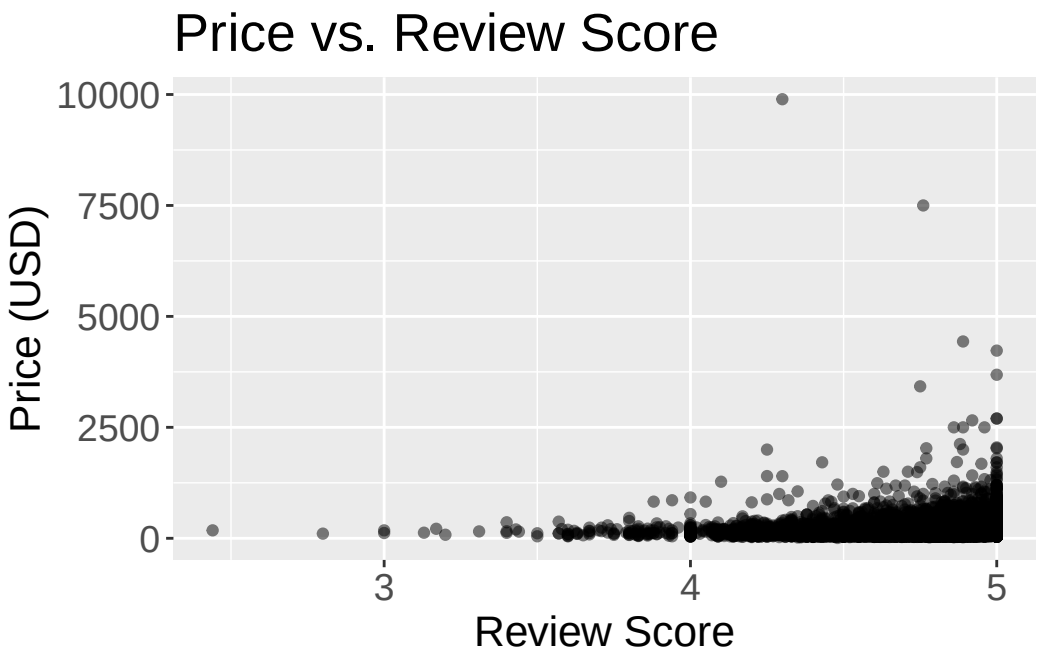
Bivariate Exploratory Data Analysis

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

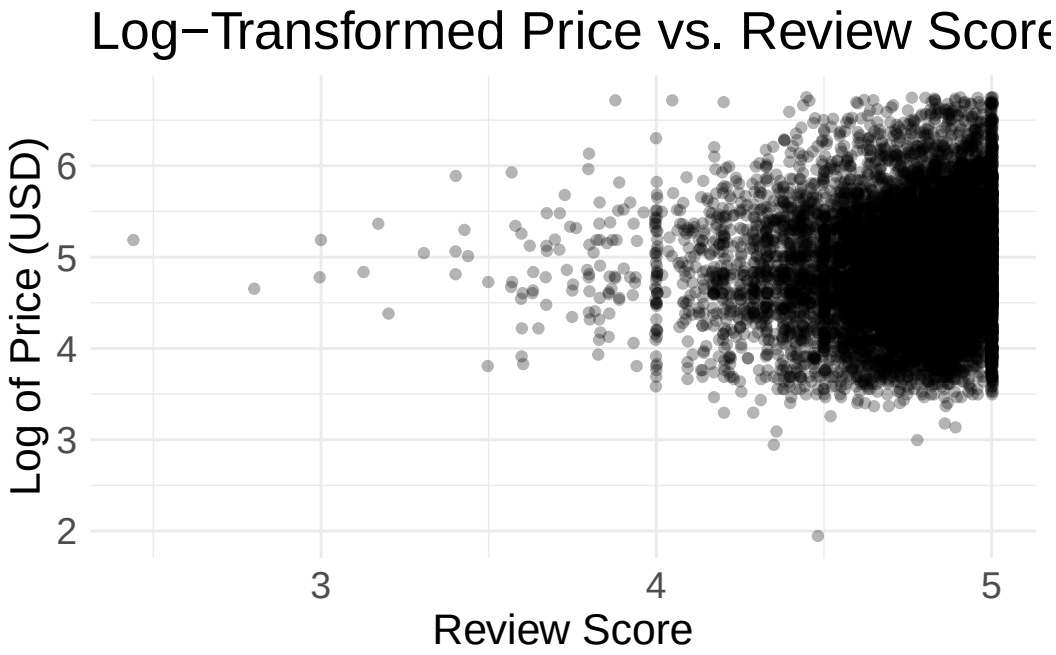


This scatter plot visualizes the relationship between the number of bedrooms and the price of Airbnb listings in New York City. There is not a clear trend. However, listings with 0 to 5 bedrooms exhibit a wide range of prices, with some listings priced significantly higher than the majority. Some outliers 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 (review scores)

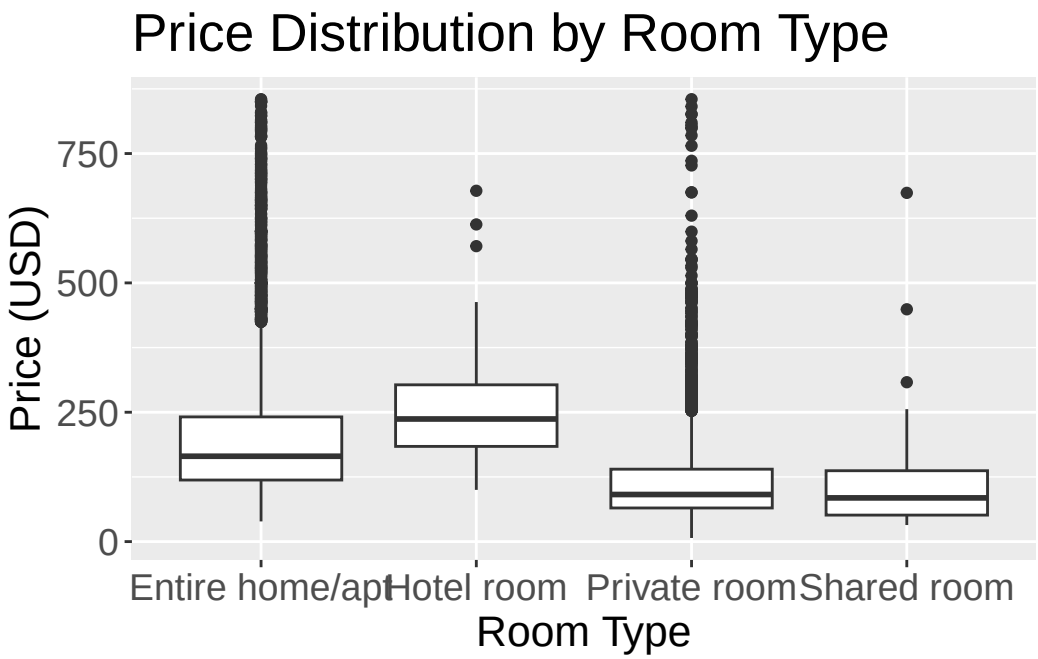


This scatter plot shows the relationship between price and review scores. However, it looks a little problematic and hard to interpret because high data density at certain score levels, especially between 4 and 5. Since the price variable is highly skewed, we applied a log transformation to try to help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

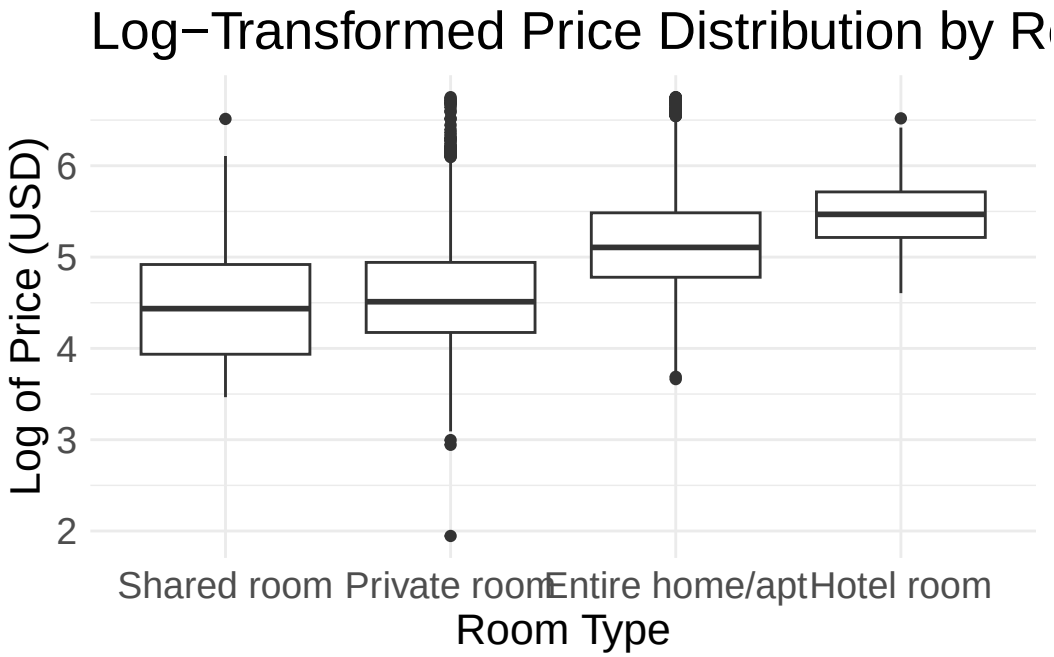


The majority of listings have review scores between 4 and 5, this shows that most listings have scores within this range. Higher-rated listings tend to have slightly higher prices, but the effect is weak.

Response (Price) vs Predictor Variable (Room Type)

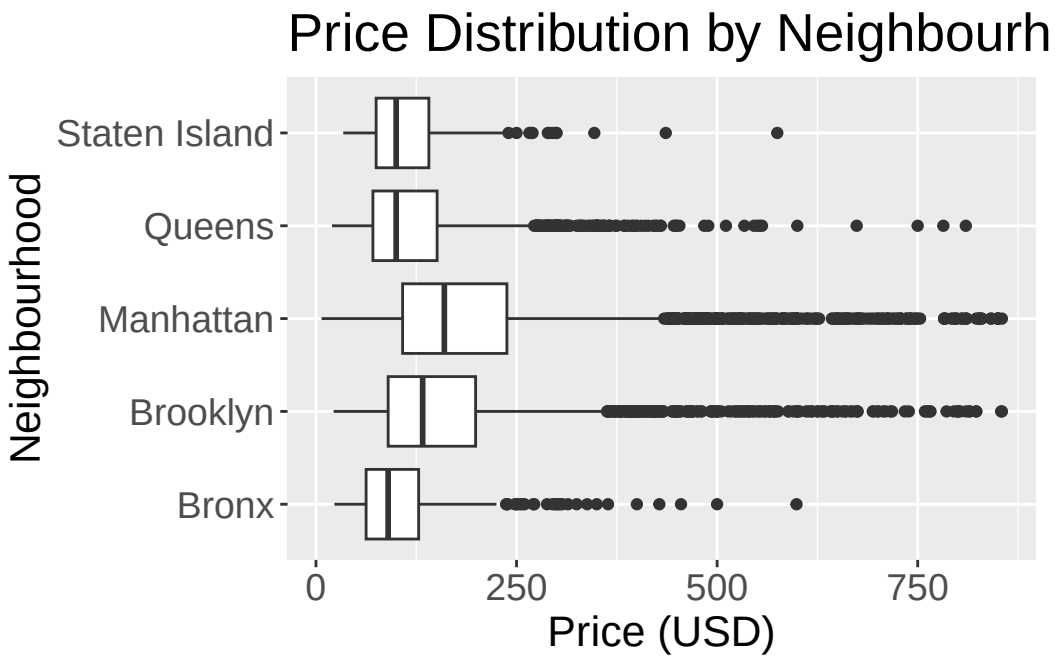


This box plot tries to visualize the distribution of prices across different room types, but it appears compressed due to the presence of extreme outliers. This makes it difficult to interpret the main trends effectively. To improve the visualization, We applied a log transformation, which may help spread out the values and make the differences between room types clearer, because prices are highly skewed.

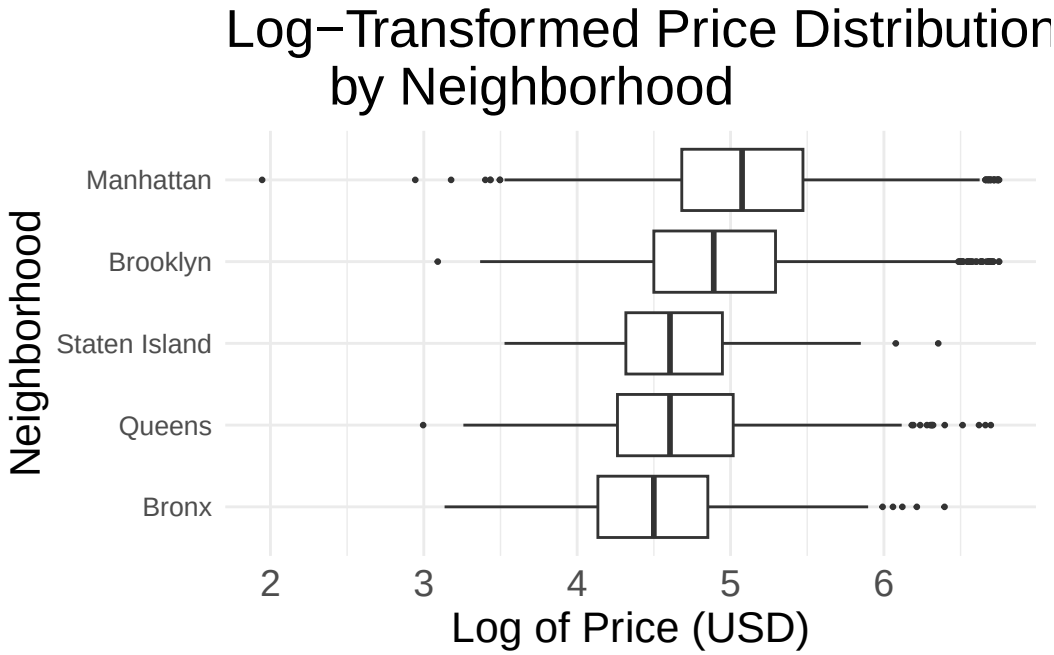


This box plot provides a clearer comparison of price distributions across different room types. According to this plot, hotel rooms have the highest median price and general highest price among all room types, suggesting that they are generally priced higher than other Airbnb listings such as shared 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. Even after log transformation, private rooms and entire homes/apartments still exhibit a notable number of higher-end outliers. However, the log transformation has greatly reduced the impact of extreme values, making the distribution more interpretable.

Response (Price) vs Predictor Variable (Neighborhood)



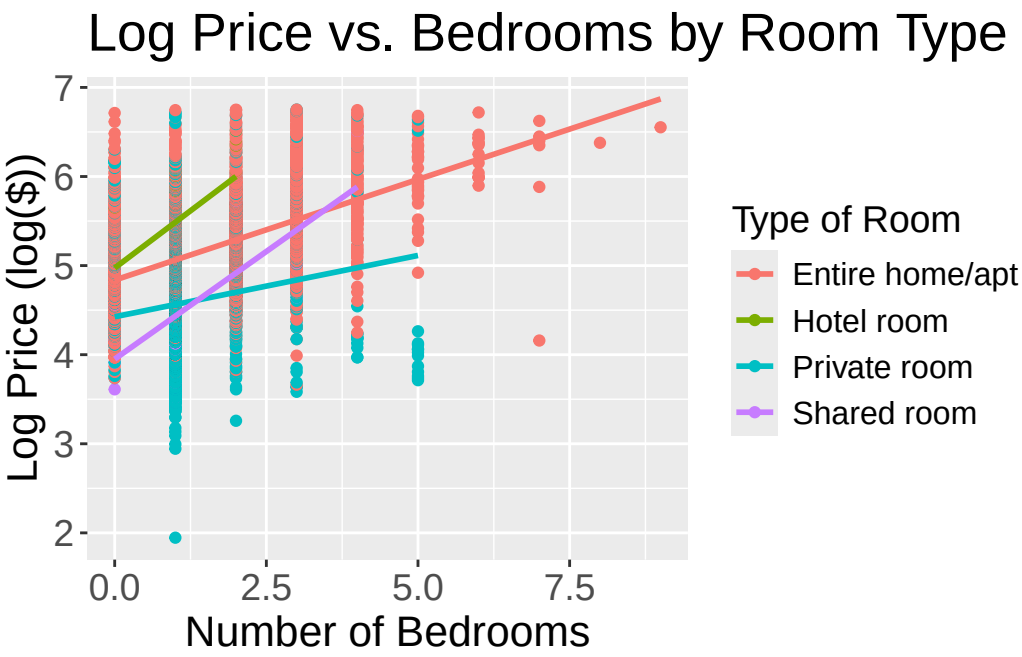
The current plot has similar issues as plots that are shown above, which is difficult to interpret due to issues like extreme outliers and others. We removed the extreme values and also did log transformation.



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.

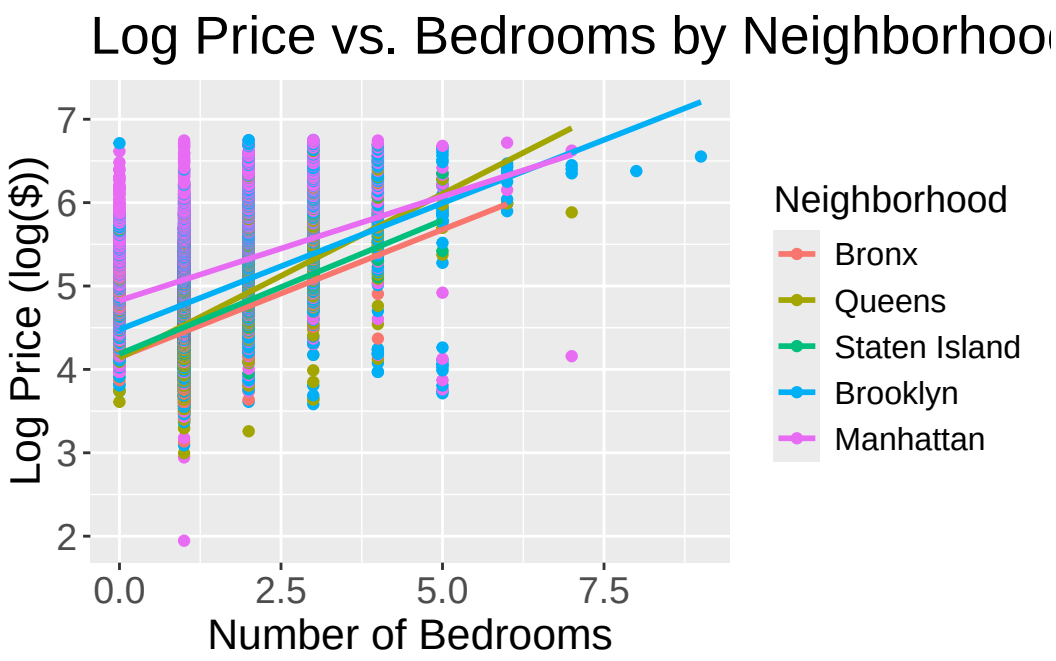
Interaction Effects

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

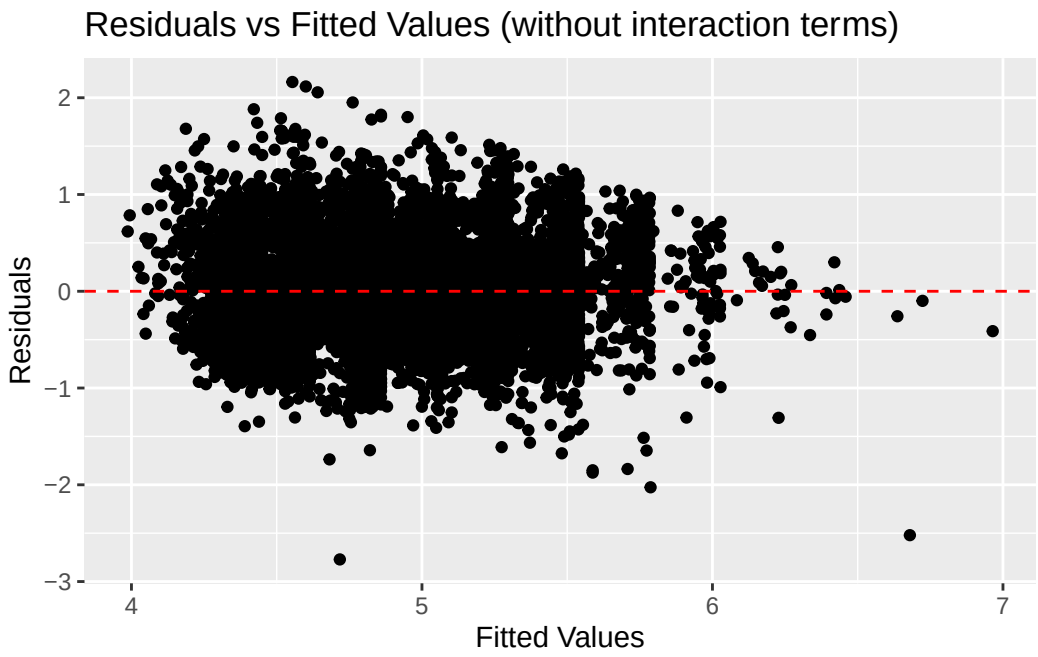
term	estimate	std.error	statistic	p.value
(Intercept)	3.193	0.099	32.333	0.000
bedrooms	0.244	0.005	44.488	0.000
room_typeHotel room	0.411	0.072	5.728	0.000
room_typePrivate room	-0.438	0.009	-46.210	0.000
room_typeShared room	-0.552	0.068	-8.113	0.000
review_scores_rating	0.272	0.020	13.433	0.000
neighbourhood_group_cleansedQueens	0.090	0.023	3.881	0.000
neighbourhood_group_cleansedStaten Island	0.013	0.040	0.328	0.743

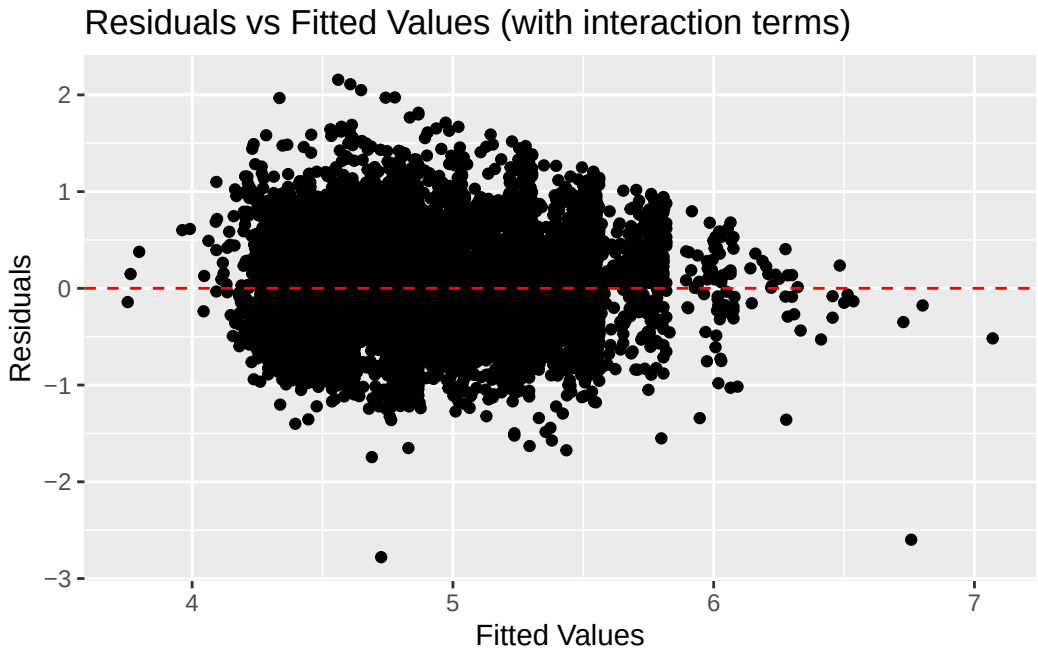
term	estimate	std.error	statistic	p.value
neighbourhood_group_cleansedBrooklyn	0.253	0.022	11.524	0.000
neighbourhood_group_cleansedManhattan	0.499	0.022	22.538	0.000

Multiple Linear Regression with Interaction Effects

term	estimate	std.error	statistic	p.value
(Intercept)	3.162	0.099	32.051	0.000
bedrooms	0.257	0.006	43.615	0.000
room_typeHotel room	0.206	0.191	1.074	0.283
room_typePrivate room	-0.321	0.020	-16.156	0.000
room_typeShared room	-0.826	0.130	-6.347	0.000
review_scores_rating	0.275	0.020	13.577	0.000
neighbourhood_group_cleansedQueens	0.088	0.023	3.833	0.000
neighbourhood_group_cleansedStaten Island	0.019	0.040	0.483	0.629
neighbourhood_group_cleansedBrooklyn	0.254	0.022	11.582	0.000
neighbourhood_group_cleansedManhattan	0.500	0.022	22.648	0.000
bedrooms:room_typeHotel room	0.202	0.170	1.191	0.234
bedrooms:room_typePrivate room	-0.104	0.015	-6.737	0.000
bedrooms:room_typeShared room	0.249	0.099	2.514	0.012

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.

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.