Data-Driven Analysis and Forecasting of the Los Angeles Airbnb Market

MA678 Midterm Project

Fengyuan Shen (Vincent)

2023-12-09

Abstract

This report conducts a data-driven analysis to forecast and understand the determinants of Airbnb listing prices in Los Angeles. With the rise of Airbnb as a significant player in the short-term rental market, this study provides a detailed examination of the factors influencing pricing strategies. It employs a robust statistical approach, starting from meticulous data cleaning to exploratory data analysis, and progresses to the development and comparison of multiple predictive models. The models range from simple linear regressions to more sophisticated partial pooling models, each analyzed for its predictive accuracy using the Root Mean Square Error (RMSE) metric. The findings are synthesized to reveal insights about the Los Angeles Airbnb market, contributing to the body of knowledge with practical analytical methodologies for pricing predictions in the sharing economy. The report's structured approach—from literature review to model evaluation—offers a clear narrative, making it a valuable resource for academics and practitioners interested in the economics of peer-to-peer accommodation platforms.

Introduction

Airbnb has revolutionized the landscape of short-term accommodation, presenting a flexible alternative to traditional hotel stays. Its emergence as a major player in the hospitality industry has prompted extensive research into its economic impact, regulatory challenges, and influence on local real estate markets.

This project aims to dissect and predict the pricing mechanisms of Airbnb listings in Los Angeles, a city with a vibrant and diverse short-term rental market. By leveraging statistical models, I endeavor to understand the factors that drive listing prices and to forecast them accurately. The methodology employed encompasses data cleaning, exploratory data analysis, and the construction of various predictive models, ranging from simple averages to complex hierarchical models.

The report is structured to walk the reader through the research process systematically. It begins with a literature review that sets the stage for understanding the current academic and practical perspectives on Airbnb. This is followed by a detailed description of the data preprocessing steps, ensuring the robustness of subsequent analyses. The main body of the report delves into exploratory data analysis, model building and validation, where each model's assumptions, strengths, and limitations are carefully examined. Finally, the report culminates in a discussion that synthesizes the findings, offering insights to the Airbnb market in Los Angeles.

Literature Review

In recent years, the rise and development of Airbnb has had a profound impact on urban housing markets and community dynamics. Studies have shown that the distribution of Airbnb listings is influenced by a number of factors. Deboosere et al.[1] uses hedonic regression models to analyze Airbnb transactions in New York City. This study finds that locational factors, particularly transit accessibility, greatly influence listing prices and revenues. It also indicates a trend towards the professionalization of the short-term rental market, leading to increased revenue for a smaller segment of hosts and potentially displacing long-term residents in accessible neighborhoods.

According to a comprehensive review by Guttentag [2][3] of 132 peer-reviewed journal articles, research on Airbnb has been categorized into six primary domains: the preferences and motivations of Airbnb guests, the incentives and objectives of hosts, the impact of Airbnb on local destinations, regulatory aspects, its influence on the tourism sector, and the operational dynamics of the company itself. The study highlights that Airbnb rentals tend to be located in areas with better transit services, proximity to city centers, and higher median house values and incomes, suggesting a risk of social inequality in the sharing economy.

In summary, these studies highlight a variety of factors that can affect Airbnb's listing prices. These insights have informed the modeling and analysis in my project, which focuses on identifying and analyzing the factors that can impact the pricing of Airbnb listings. This involves examining how elements like location, socioeconomic status, and broader market trends.

Method

The initial dataset was segmented into ten distinct sections. To facilitate a comprehensive analysis, I consolidated these segments into a single dataset, subsequently named **`listings.csv`**.

Data Cleaning

In the data cleaning phase, the dataset comprised 75 variables, presenting a complexity that necessitated reduction for effective model fitting. Drawing upon insights gathered from the literature review, I meticulously chose over 30 variables deemed pertinent for predicting Airbnb's listing prices. The following table delineates a subset of these selected variables.

host_name	host_since	bathrooms_text	bedrooms	beds	price	review_scores_rating
Shiela	2019-07-21	1 bath	NA	1	$90.00	NA
Yahide	2022-10-07	3 baths	NA	1	$35.00	NA
Lin	2023-01-13	1 bath	1	2	$69.00	NA
Paola	2014-09-30	1.5 baths	1	1	$120.00	5.00
Mojdeh	2018-12-22	1.5 baths	NA	2	$450.00	NA
Mark	2020-01-15	4.5 baths	5	7	$1,400.00	4.75

This is a summary of the dataset.

```
##        id                host_id           host_name          host_since
##  Min.   :1.090e+02   Min.   :      521   Length:44594       Length:44594
##  1st Qu.:2.902e+07   1st Qu.: 25138314   Class :character   Class :character
##  Median :5.219e+07   Median :107434423   Mode  :character   Mode  :character
##  Mean   :3.609e+17   Mean   :177057751
##  3rd Qu.:7.851e+17   3rd Qu.:321708775
##  Max.   :9.724e+17   Max.   :534987552
```

```
##
##  host_response_time host_response_rate host_acceptance_rate host_is_superhost
##  Length:44594       Length:44594       Length:44594         Length:44594
##  Class :character   Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character   Mode  :character     Mode  :character
##
##
##
##
##  host_listings_count host_identity_verified    latitude       longitude
##  Min.   :   0.0      Length:44594           Min.   :33.34   Min.   :-118.9
##  1st Qu.:   1.0      Class :character       1st Qu.:34.00   1st Qu.:-118.4
##  Median :   3.0      Mode  :character       Median :34.06   Median :-118.3
##  Mean   : 104.6                             Mean   :34.05   Mean   :-118.3
##  3rd Qu.:  12.0                             3rd Qu.:34.11   3rd Qu.:-118.2
##  Max.   :4576.0                             Max.   :34.81   Max.   :-117.7
##  NA's   :2
##  property_type       room_type          accommodates   bathrooms_text
##  Length:44594       Length:44594       Min.   : 1.00   Length:44594
##  Class :character   Class :character   1st Qu.: 2.00   Class :character
##  Mode  :character   Mode  :character   Median : 3.00   Mode  :character
##                                        Mean   : 3.94
##                                        3rd Qu.: 5.00
##                                        Max.   :16.00
##
##     bedrooms           beds           price           minimum_nights
##  Min.   : 1.000   Min.   : 1.000   Length:44594       Min.   :   1.00
##  1st Qu.: 1.000   1st Qu.: 1.000   Class :character   1st Qu.:   2.00
##  Median : 2.000   Median : 2.000   Mode  :character   Median :   7.00
##  Mean   : 2.062   Mean   : 2.182                      Mean   :  17.73
##  3rd Qu.: 3.000   3rd Qu.: 3.000                      3rd Qu.:  30.00
##  Max.   :32.000   Max.   :50.000                      Max.   :1124.00
##  NA's   :13343    NA's   :503
##  maximum_nights   number_of_reviews review_scores_rating review_scores_accuracy
##  Min.   :   1.0   Min.   :   0.00   Min.   :0.000        Min.   :0.000
##  1st Qu.:  90.0   1st Qu.:   0.00   1st Qu.:4.690        1st Qu.:4.750
##  Median : 365.0   Median :   5.00   Median :4.890        Median :4.910
##  Mean   : 518.3   Mean   :  34.38   Mean   :4.711        Mean   :4.771
##  3rd Qu.:1125.0   3rd Qu.:  31.00   3rd Qu.:5.000        3rd Qu.:5.000
##  Max.   :3650.0   Max.   :2472.00   Max.   :5.000        Max.   :5.000
##                                     NA's   :11632        NA's   :11863
##  review_scores_cleanliness review_scores_checkin review_scores_communication
##  Min.   :0.000             Min.   :0.000         Min.   :0.000
##  1st Qu.:4.670             1st Qu.:4.850         1st Qu.:4.850
##  Median :4.870             Median :4.970         Median :4.970
##  Mean   :4.712             Mean   :4.839         Mean   :4.831
##  3rd Qu.:5.000             3rd Qu.:5.000         3rd Qu.:5.000
##  Max.   :5.000             Max.   :5.000         Max.   :5.000
##  NA's   :11862             NA's   :11870         NA's   :11863
##  review_scores_location review_scores_value instant_bookable
##  Min.   :0.000          Min.   :0.000       Length:44594
##  1st Qu.:4.740          1st Qu.:4.600       Class :character
##  Median :4.900          Median :4.800       Mode  :character
##  Mean   :4.778          Mean   :4.669
```

```
##  3rd Qu.:5.000          3rd Qu.:4.950
##  Max.   :5.000          Max.   :5.000
##  NA's   :11871          NA's   :11874
##  host_neighbourhood
##  Length:44594
##  Class :character
##  Mode  :character
##
##
##
##
```

Upon preliminary examination of the dataset, it becomes evident that the removal of missing values and outliers is imperative. Their presence could potentially skew the predictive accuracy of the model.

For the **price** column, I removed all non-numeric characters and calculated the first quartile (Q1) and third quartile (Q3) of the column, as well as the interquartile (IQR). They are then used to define a range of outliers. Finally, the dataset is filtered to retain only the price values in this range.

In the **host_response_time**, **host_is_superhost**, **host_identity_verified**, **instant_bookable**, **bathrooms_text**, and **room_type** columns, I converted categorical variables into numeric types, simultaneously filtering out rows containing NA values. This transformation not only streamlined the dataset but also enhanced its suitability for quantitative analysis and modeling.

Regarding the **host_response_rate** column, I transformed percentage strings into numeric values, thereby standardizing the data format for more effective subsequent analyses.

EDA

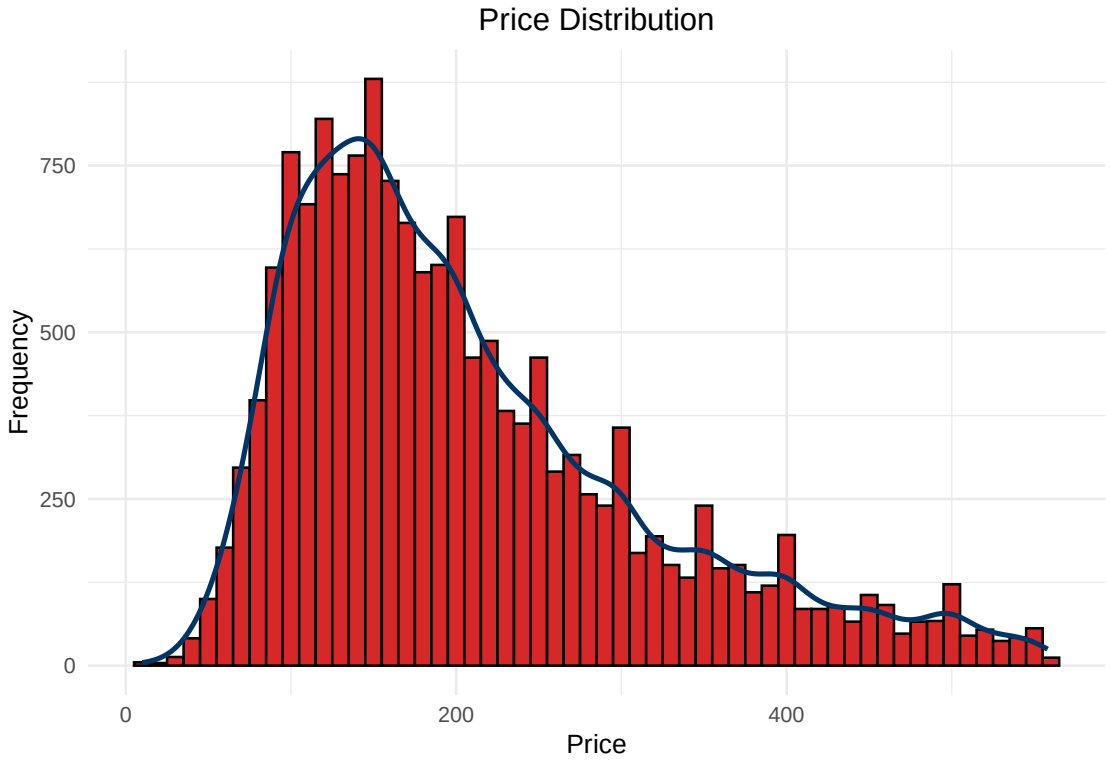
Following the completion of the data cleaning process, the subsequent phase involves conducting exploratory data analysis (EDA). This critical step will facilitate a deeper and more nuanced understanding of the dataset.

Initially, to ascertain the linear relationships between various variables, I constructed a heatmap. This graphical representation elucidates the correlation coefficients among different variables, providing a visual overview of their interrelationships.

The heatmap analysis reveals that the **price** variable exhibits significant correlations with several factors, including **host_is_superhost**, **room_type**, **accommodates**, **bathrooms**, **bedrooms**, **beds**, **minimum_nights**, **number_of_reviews**, **review_scores_ratings**, and **review_scores_location**.

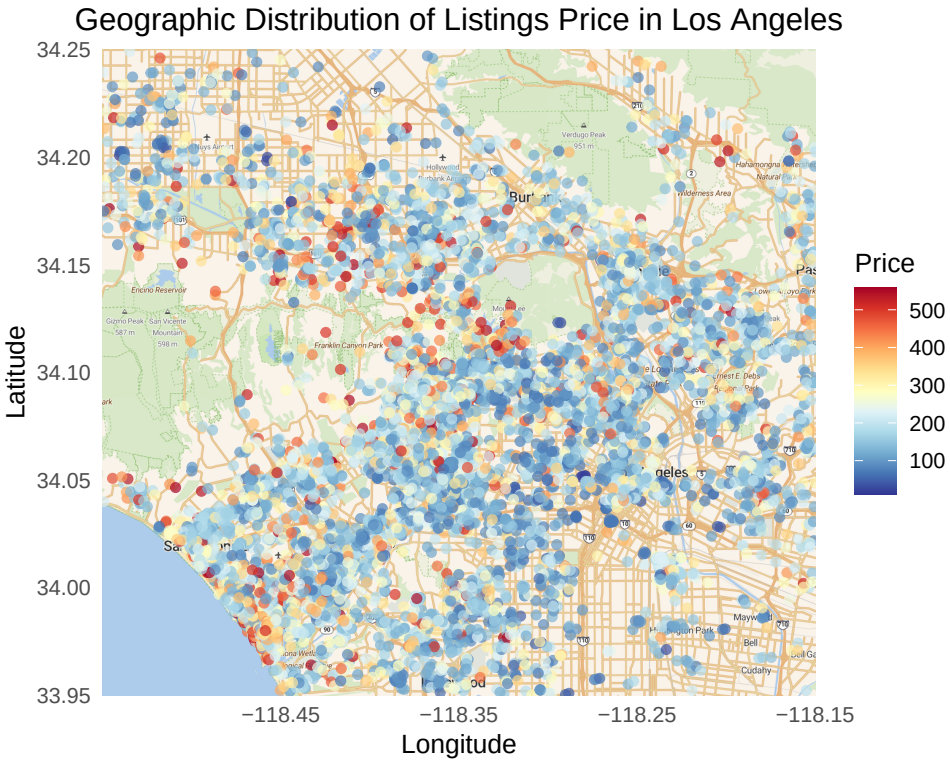
However, it is crucial to note that the correlation matrix primarily reflects linear relationships between variables. In some instances, the associations might be non-linear, as exemplified by the relationship between **price** and geographical coordinates (**latitude** and **longitude**). Consistent with insights from existing literature, Airbnb's listing price is influenced by geographic location. Accordingly, incorporating **latitude** and **longitude** into the model building process is a critical step to capture these complex spatial dynamics.

Before delving into the exploration of other variables, an initial analysis was conducted to examine the distribution of Airbnb's listing price.

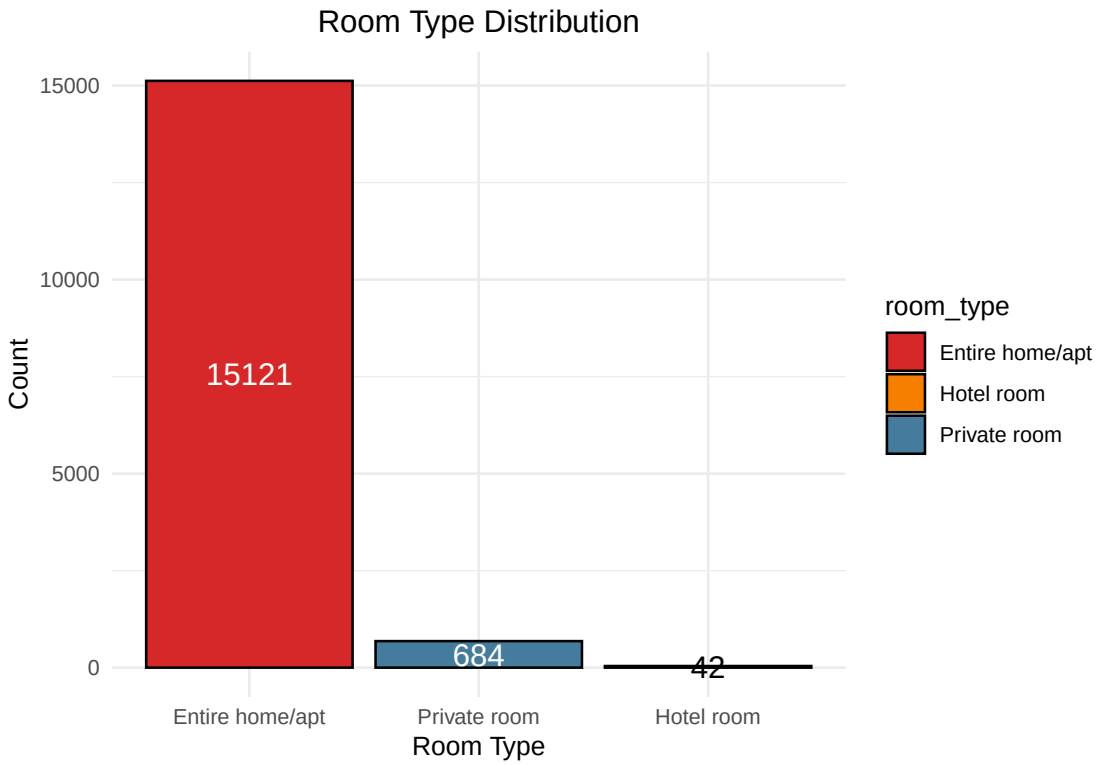


The histogram indicates a right-skewed distribution, with a high frequency of listings priced at the lower end of the spectrum, suggesting that budget-friendly options are more common in the market. Most listings in Los Angeles are priced under $200, although there are also listings priced over $400. Notably, there is a long tail extending towards the higher price range, highlighting the presence of premium-priced listings.

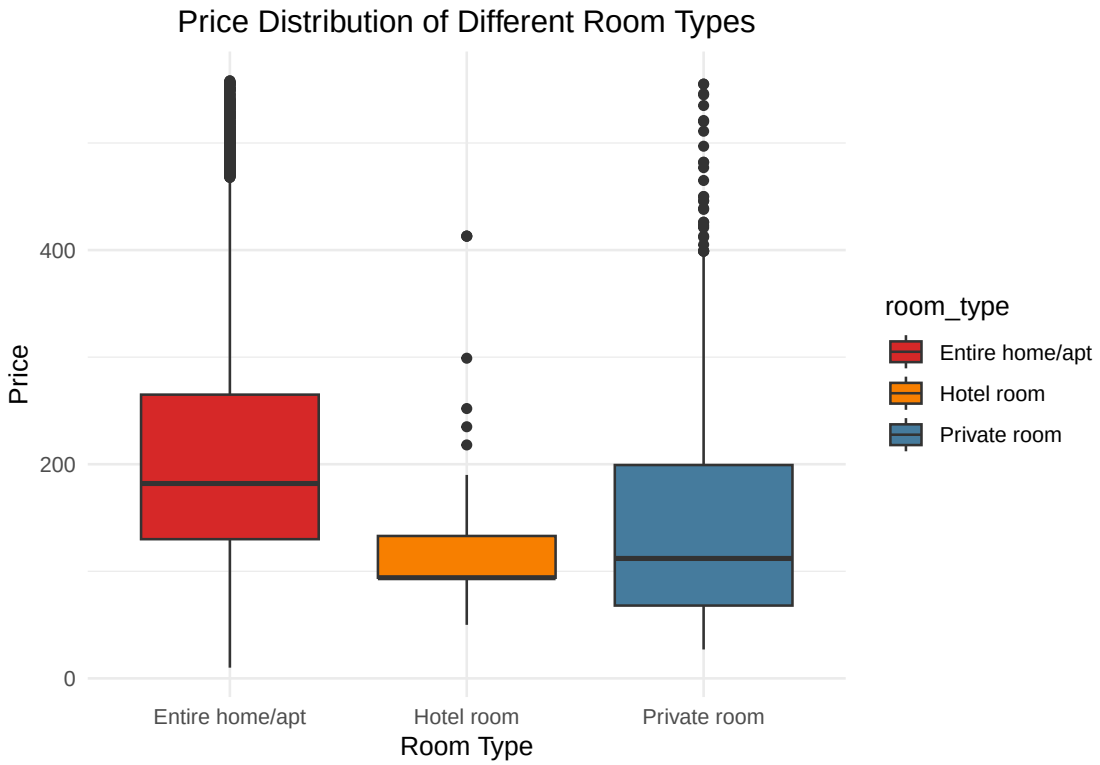
Given the right-skewed distribution here, it is prudent to consider a logarithmic transformation of the price. This transformation aims to normalize the distribution, thereby satisfying one of the key assumptions of linear regression models which is the normality of the error terms.



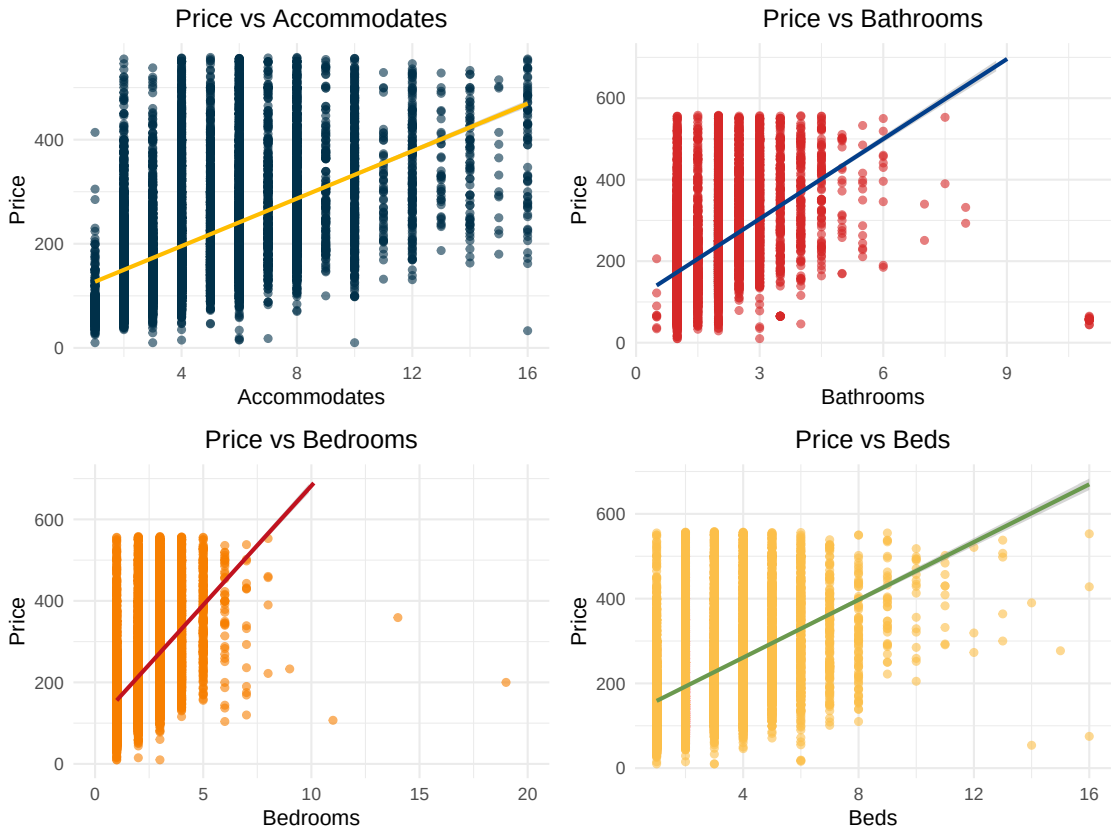
The geographic distribution map of Airbnb's listing prices in Los Angeles visualizes the variation in pricing across different locations, with a higher density of listings in central and coastal areas. A gradient in pricing is observed, with more expensive listings often clustered in specific neighborhoods, potentially reflecting the desirability or accessibility of these areas. Conversely, more affordable listings appear to be more dispersed throughout the city.



The bar chart depicts the room type distribution of Airbnb listings in Los Angeles, highlighting a predominant preference for **`Entire home/apartments`** which far outnumbers the other room types. This distribution suggests a significant market demand for entire homes or apartments, with limited availability or demand for hotel-style accommodations on the platform.



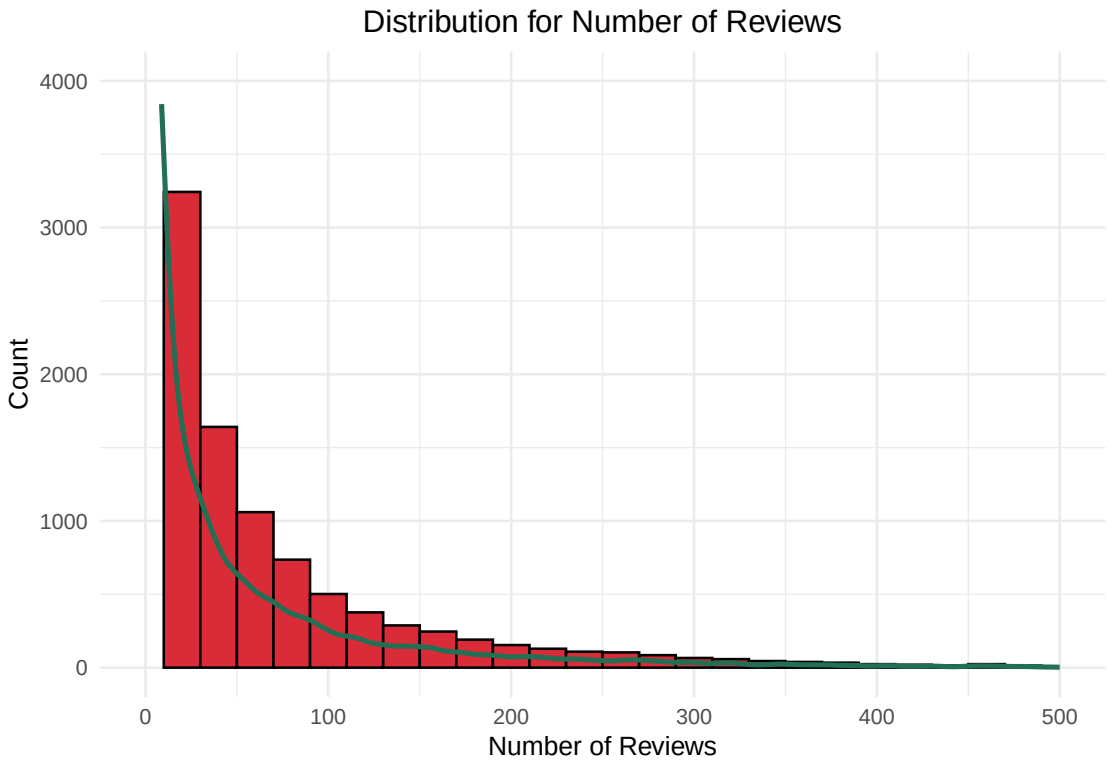
I used boxplot to provide a compelling visualization of the price variance among different room types offered on Airbnb in Los Angeles. The distinct median prices and interquartile ranges across **`Entire home/apartments`**, **`Private room`**, and **`Hotel room`** indicate that room type is a significant factor influencing listing prices.



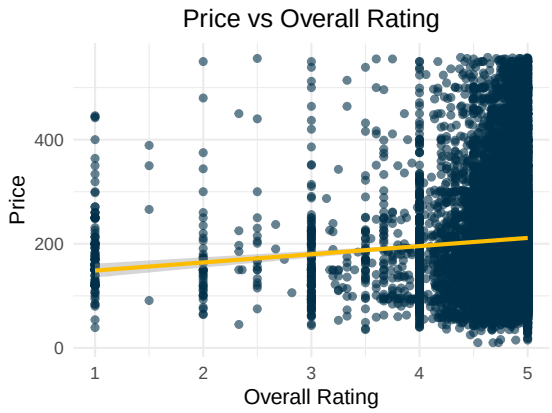
The scatter plots above illustrate the relationship between Airbnb listing prices and property features such as the number of people the property can accommodate, and the number of bathrooms, bedrooms, and beds available.

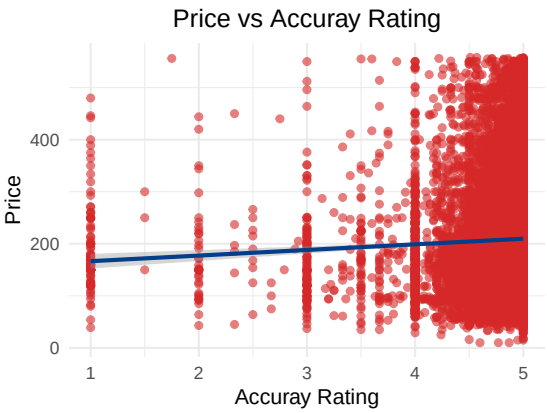
A positive trend is observed across all variables, with listing prices increasing in tandem with the number of accommodations, bathrooms, bedrooms, and beds. This suggests that properties with greater capacity and amenities are priced higher, reflecting their increased value to renters.

The clear trends demonstrated in these plots justify the inclusion of these variables in the model, as they are likely to be significant predictors of price.

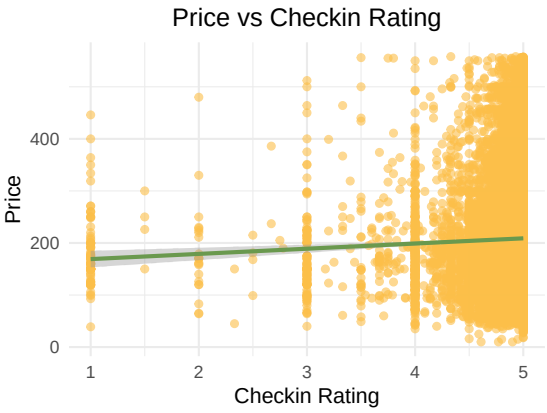


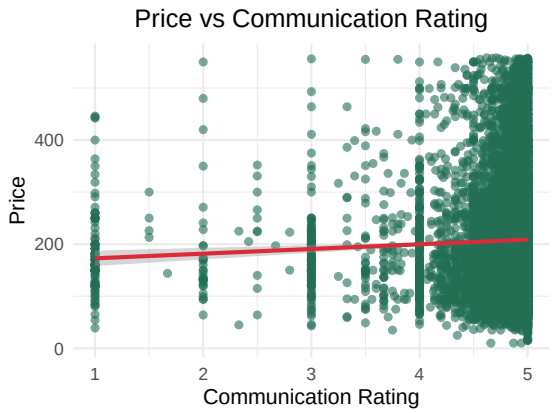
From the histogram which illustrates a right-skewed distribution for the number of reviews per Airbnb listing, the presence of listings with an exceptionally high number of reviews suggests that there are a few highly popular or long-established listings. Given the significance of customer reviews in influencing rental desirability and trustworthiness, the number of reviews will be included as a variable in the model to capture its potential impact on pricing.

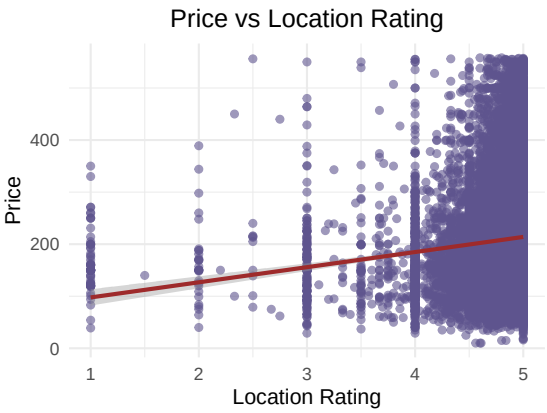

Price vs Overall Rating
Price
Overall Rating
0
200
400
1
2
3
4
5

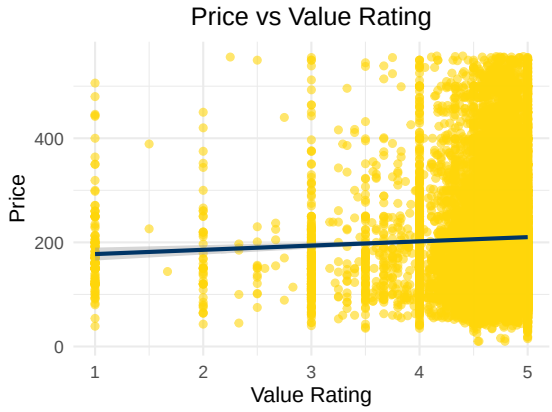

Price vs Accuray Rating
Price
Accuray Rating
0
200
400
1
2
3
4
5


Price vs Cleanliness Rating
Price
Cleanliness Rating
0
200
400
1
2
3
4
5


Price vs Checkin Rating
Price
Checkin Rating
0
200
400
1
2
3
4
5


Price vs Communication Rating
Price
Communication Rating
0
200
400
1
2
3
4
5


Price vs Location Rating
Price
Location Rating
0
200
400
1
2
3
4
5


Price vs Value Rating
Price
Value Rating
0
200
400
1
2
3
4
5

The scatter plots above analyze the relationship between Airbnb listing prices in Los Angeles and various customer ratings. The upward trend in the price versus overall rating plot suggests a positive correlation. Meanwhile, the location rating also shows a pronounced positive correlation with price, highlighting the potential influence of location on pricing decisions. In contrast, the plots for accuracy, cleanliness, check-in, communication, and value ratings show a less distinct relationship with price, suggesting that these factors might not be as influential in the pricing model.

As a result, given the visual trend of the overall and location rating, it is prudent to consider incorporating overall ratings and location-related ratings into the predictive model.

Fitting Model

Based on the insights garnered from the literature and EDA process, I decided to include the following variables as predictors in the model: **`host_is_superhost`**, **`room_type`**, **`accommodates`**, **`bathrooms`**, **`bedrooms`**, **`beds`**, **`minimum_nights`**, **`number_of_reviews`**, **`review_scores_ratings`**, **`review_scores_location`**, along with geographical coordinates **`latitude`** and **`longitude`**.

Moving forward, to facilitate the validation of the model's predictive efficacy post-construction, I have partitioned the data into train set and test set. The model will be built on the train set, thereafter its performance will be assessed on the test set.

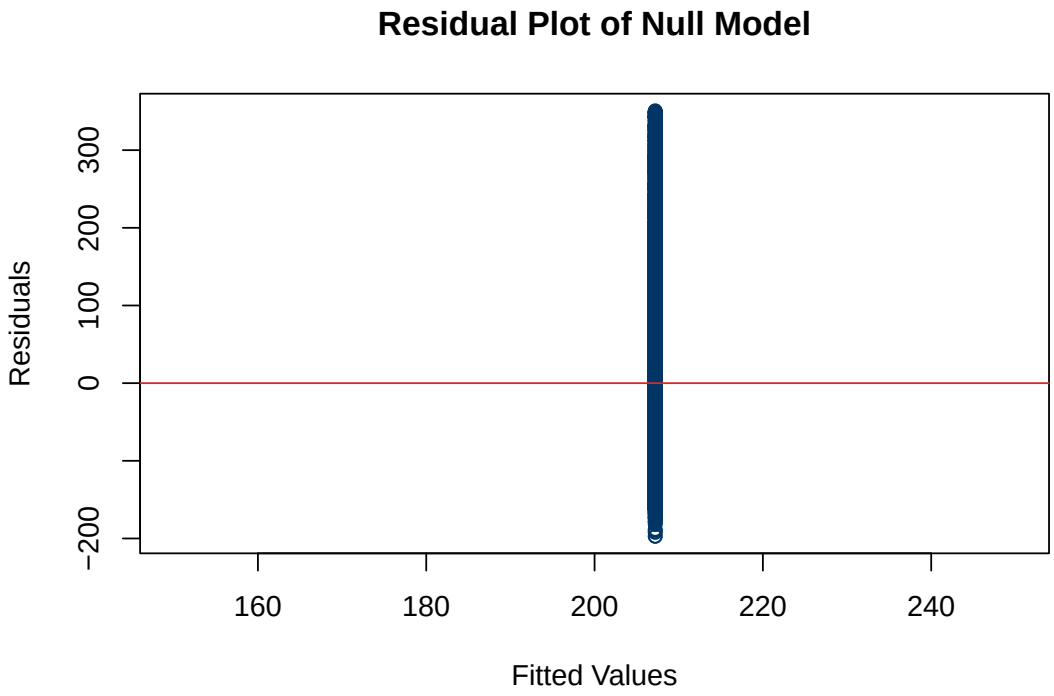
Split the Dataset into train_data and test_data

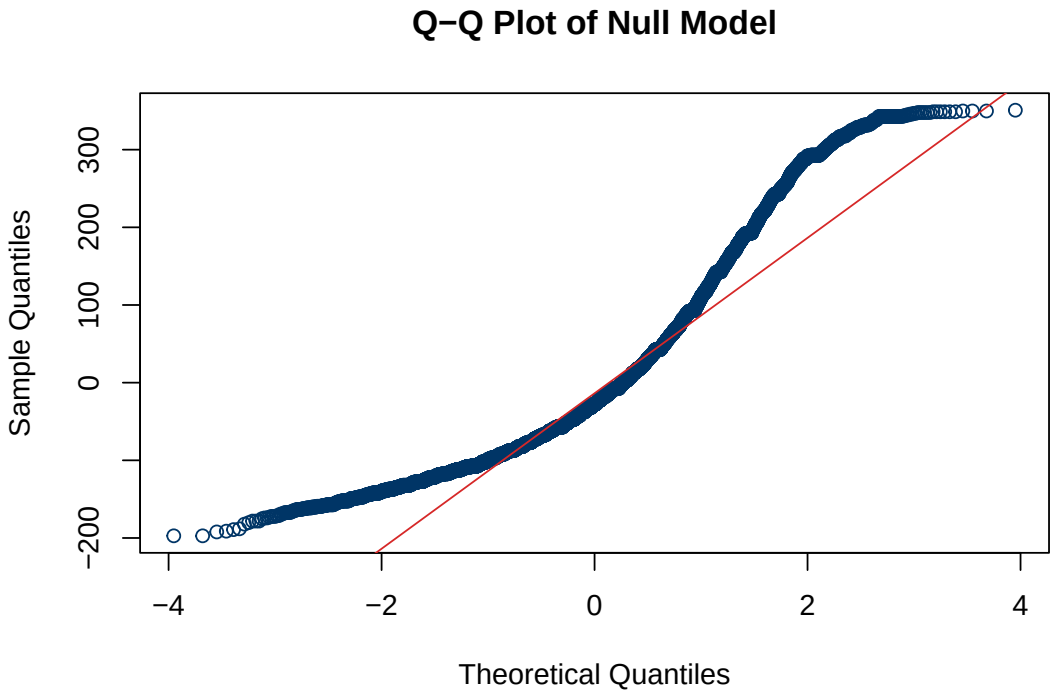
To ensure class balance within the dataset, I utilized the **`createDataPartition`** function from the **`caret`** package, which performs stratified sampling to maintain representative proportions of classes in both the train and test sets. I allocated 80% of the data for training purposes and reserved 20% for testing.

Model1: Null Model

$$lm(Price = \alpha)$$

```
##
## Call:
## lm(formula = price ~ 1, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.21  -81.21  -27.21   53.79  350.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 207.2075     0.9585   216.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 108.8 on 12874 degrees of freedom
```





The Null Model serves as a baseline for my regression analysis, assuming no predictors other than the intercept. The summary output indicates that the intercept, representing the average listing price when no predictors are included, is estimated at 207.2075 with a standard deviation of 0.9585. This model is

statistically significant, as evidenced by the p-value of less than 2e-16. However, the large residual standard error of 108.8 suggests considerable unexplained variability in listing prices.

For the residual plot, I observe a very narrow vertical distribution of residuals, showing that the model does not capture the variance in the data. Also, the Q-Q Plot of the Null Model indicates that the residuals are not normally distributed.

Based on the above analysis, I would say that the Null Model, without any predictors, is insufficient for explaining the variance in the Airbnb listing prices.

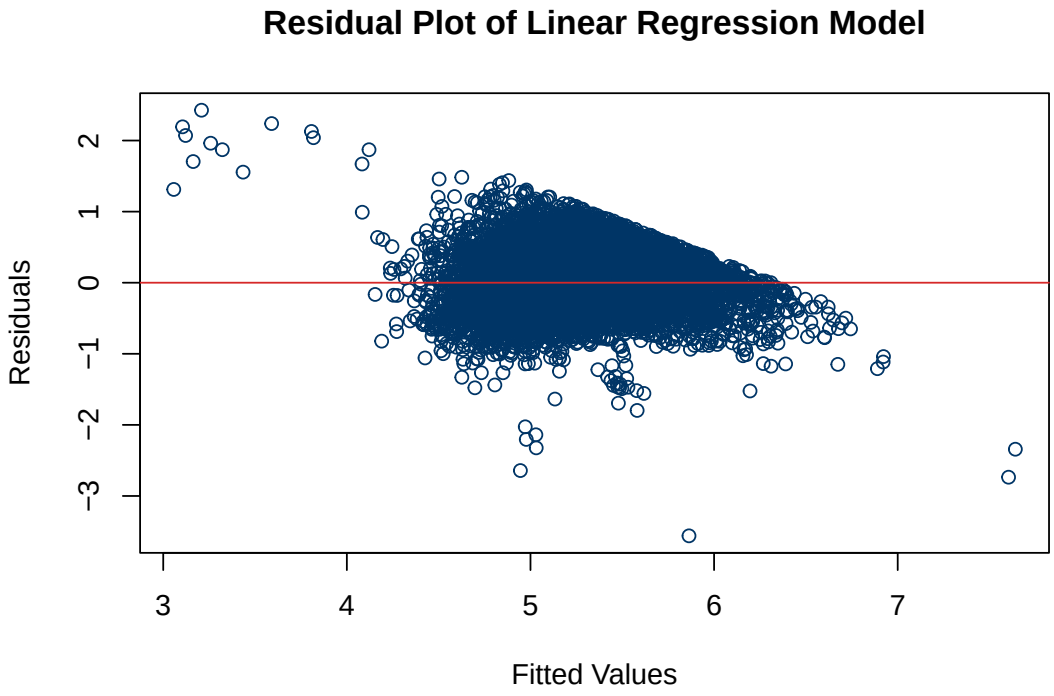
Model2: Complete Pooling Model (Linear Regression)

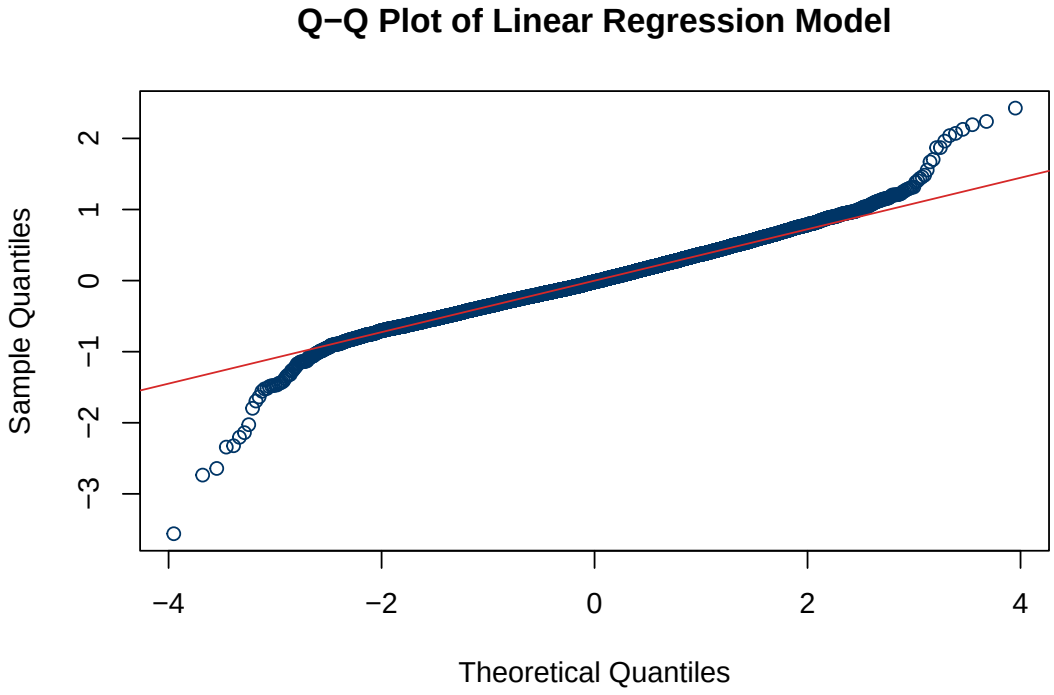
During the EDA process, it became evident that the distribution of listing prices was substantially right-skewed, with the majority of listings being lower-priced, while a minority were notably higher-priced. Given this skewed distribution, a linear model (LM) may not be the best choice because LM assumes that the response variables are approximately normally distributed. So I used a logarithmic transformation to process the response variable to make it closer to a normal distribution and thus fit a linear model.

$$lm(log(price) = \alpha + \beta_1 \cdot host\_is\_superhost + \beta_2 \cdot latitude + \beta_3 \cdot longitude + \beta_4 \cdot room\_type + \beta_5 \cdot accommodates$$
$$+ \beta_6 \cdot bathrooms + \beta_7 \cdot bedrooms + \beta_8 \cdot beds + \beta_9 \cdot minimum\_nights + \beta_{10} \cdot number\_of\_reviews$$
$$+ \beta_{11} \cdot review\_scores\_rating + \beta_{12} \cdot review\_scores\_location)$$

```
##
## Call:
## lm(formula = log(price) ~ host_is_superhost + latitude + longitude +
##     room_type + accommodates + bathrooms + bedrooms + beds +
##     minimum_nights + number_of_reviews + review_scores_rating +
##     review_scores_location, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5608 -0.2454 -0.0149  0.2430  2.4264
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -7.682e+01  2.493e+00 -30.813  < 2e-16 ***
## host_is_superhost       2.985e-02  7.182e-03   4.157 3.24e-05 ***
## latitude               -4.190e-01  2.449e-02 -17.107  < 2e-16 ***
## longitude              -8.050e-01  2.042e-02 -39.428  < 2e-16 ***
## room_typeHotel room    -2.463e-01  6.449e-02  -3.820 0.000134 ***
## room_typePrivate room  -2.053e-01  1.714e-02 -11.978  < 2e-16 ***
## accommodates            5.653e-02  2.949e-03  19.172  < 2e-16 ***
## bathrooms               8.103e-02  5.842e-03  13.870  < 2e-16 ***
## bedrooms                1.328e-01  6.216e-03  21.360  < 2e-16 ***
## beds                   -1.812e-02  4.556e-03  -3.977 7.03e-05 ***
## minimum_nights         -5.552e-03  1.964e-04 -28.262  < 2e-16 ***
## number_of_reviews      -2.512e-04  3.870e-05  -6.493 8.75e-11 ***
## review_scores_rating   -3.745e-02  1.016e-02  -3.686 0.000229 ***
## review_scores_location  1.499e-01  1.145e-02  13.090  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3828 on 12861 degrees of freedom
## Multiple R-squared:  0.4531, Adjusted R-squared:  0.4525
## F-statistic: 819.5 on 13 and 12861 DF,  p-value: < 2.2e-16
```





The Complete Pooling model was estimated using a Linear Regression approach. The summary of the regression model indicates all the variables are statistically significant predictors of the logarithmically transformed price, as demonstrated by the p-values less than 0.05.

The coefficients from the model provide insights into the relative impact of each predictor. For instance, being a superhost and having additional bathrooms and bedrooms are associated with higher prices, while the type of room being either a hotel room or a private room typically corresponds with a lower price relative to entire homes/apartments.

The Adjusted R-squared value of the model is 0.4525, indicating that the model explains about 45.25% of the variability of the price variable.

The Residual Plot of the Linear Regression Model shows residuals scattered around the zero line, which is indicative of a good fit, but the pattern suggests potential heteroscedasticity as the variance of residuals appears to increase with the fitted values.

The Q-Q Plot of the Linear Regression Model reveals some deviation from normality, particularly in the tails of the distribution, suggesting that the normality assumption may not fully hold. This could be due to outliers.

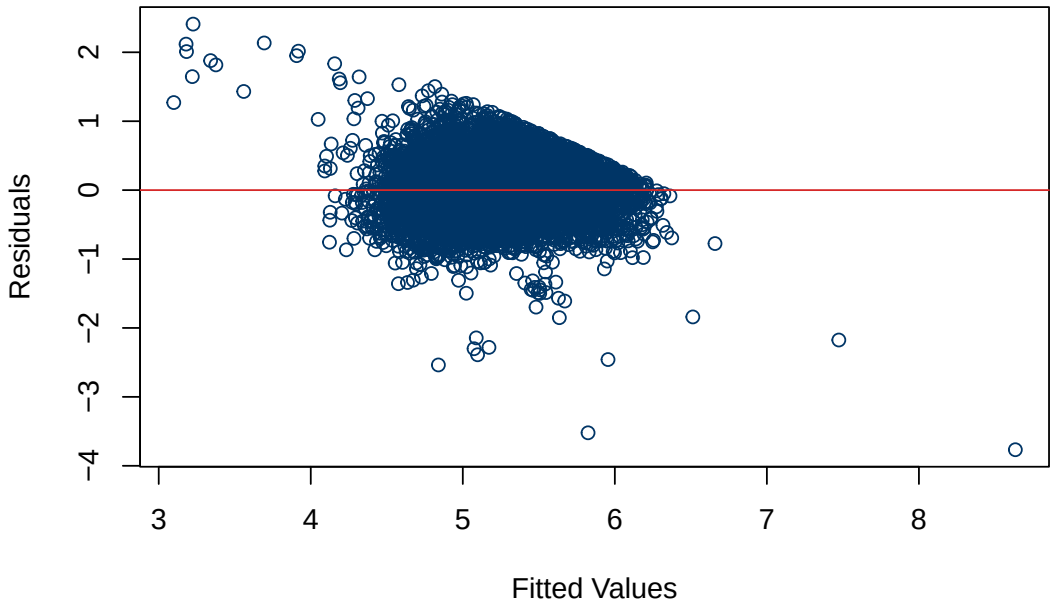
Model3: Linear Regression Model with Interaction Term

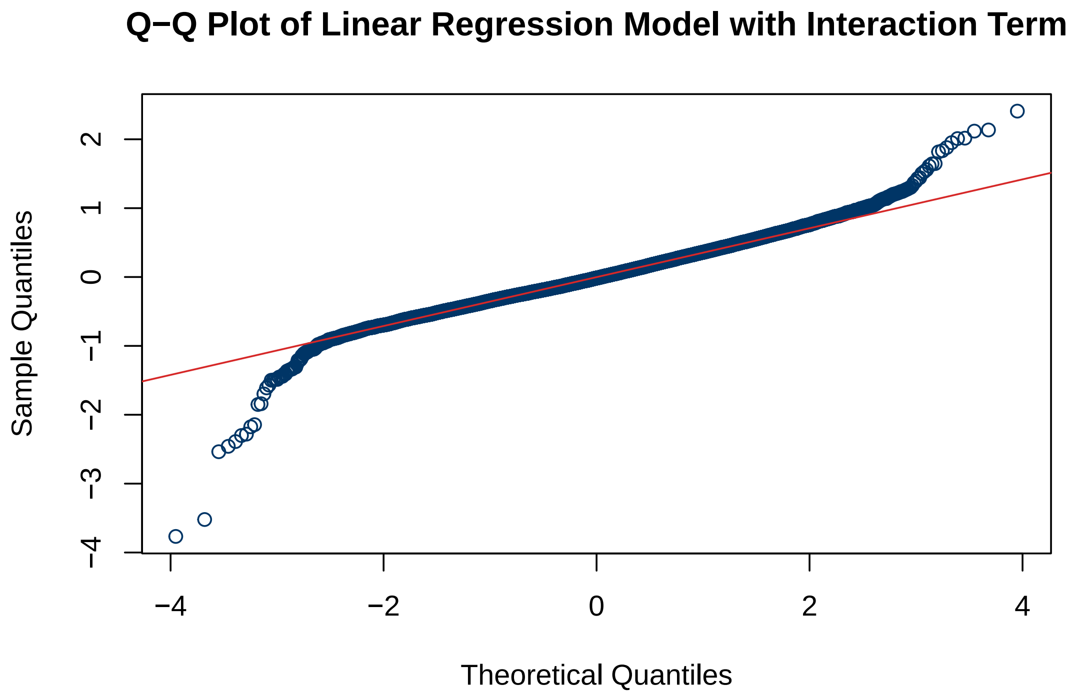
$$lm(log(price) = \alpha + \beta_1 \cdot host\_is\_superhost + \beta_2 \cdot latitude + \beta_3 \cdot longitude + \beta_4 \cdot room\_type + \beta_5 \cdot accommodates + \beta_6 \cdot bathrooms + \beta_7 \cdot bedrooms + \beta_8 \cdot beds + \beta_9 \cdot minimum\_nights + \beta_{10} \cdot number\_of\_reviews + \beta_{11} \cdot review\_scores\_rating + \beta_{12} \cdot review\_scores\_location + \beta_{13} \cdot latitude * longitude + \beta_{14} \cdot accommodates * bedrooms + \beta_{15} \cdot review\_scores\_rating * review\_scores\_location)$$

```
##
## Call:
## lm(formula = log(price) ~ host_is_superhost + latitude + longitude +
##     room_type + accommodates + bathrooms + bedrooms + beds +
##     minimum_nights + number_of_reviews + review_scores_rating +
##     review_scores_location + latitude:longitude + accommodates:bedrooms +
##     review_scores_rating:review_scores_location, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7667 -0.2411 -0.0149  0.2377  2.4087
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                -7.046e+03  6.495e+02 -10.849
## host_is_superhost                           1.978e-02  7.097e-03   2.787
## latitude                                    2.042e+02  1.906e+01  10.714
## longitude                                  -5.973e+01  5.491e+00 -10.877
## room_typeHotel room                        -1.223e-01  6.329e-02  -1.933
## room_typePrivate room                      -1.576e-01  1.689e-02  -9.330
## accommodates                                9.812e-02  3.733e-03  26.284
## bathrooms                                   8.712e-02  5.742e-03  15.173
## bedrooms                                    2.147e-01  7.732e-03  27.768
## beds                                       -1.215e-02  4.471e-03  -2.717
## minimum_nights                             -5.444e-03  1.922e-04 -28.323
## number_of_reviews                          -2.171e-04  3.787e-05  -5.732
## review_scores_rating                       -3.240e-01  2.341e-02 -13.839
```

```
## review_scores_location                       -1.111e-01  2.214e-02  -5.019
## latitude:longitude                            1.730e+00  1.611e-01  10.735
## accommodates:bedrooms                        -1.585e-02  9.530e-04 -16.632
## review_scores_rating:review_scores_location   7.337e-02  5.425e-03  13.524
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## host_is_superhost                             0.00533 **
## latitude                                      < 2e-16 ***
## longitude                                     < 2e-16 ***
## room_typeHotel room                           0.05331 .
## room_typePrivate room                         < 2e-16 ***
## accommodates                                  < 2e-16 ***
## bathrooms                                     < 2e-16 ***
## bedrooms                                      < 2e-16 ***
## beds                                          0.00659 **
## minimum_nights                                < 2e-16 ***
## number_of_reviews                            1.01e-08 ***
## review_scores_rating                          < 2e-16 ***
## review_scores_location                       5.26e-07 ***
## latitude:longitude                            < 2e-16 ***
## accommodates:bedrooms                         < 2e-16 ***
## review_scores_rating:review_scores_location    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3743 on 12858 degrees of freedom
## Multiple R-squared:  0.4772, Adjusted R-squared:  0.4766
## F-statistic: 733.5 on 16 and 12858 DF,  p-value: < 2.2e-16
```

Residual Plot of Linear Regression Model with Interaction Term





This model is an extension from the basic model to include interactions between certain variables. Specifically, interaction terms between latitude and longitude, accommodates and bedrooms, and review_scores_rating and review_scores_location have been included, suggesting that I believe the influence of location on price is not uniform across the geographic space, that the relationship between the number of people accommodated and the price varies with the number of bedrooms, and that the impact of review scores on price is moderated by their respective locations.

This model retains many of the predictors from the previous model, with all variables showing statistical significance in predicting the logarithmically transformed price.

The Residual Plot and Q-Q Plot show the similar pattern to the previous model. However, the adjusted R-squared value of 0.4766 represents a slight improvement over the previous model, explaining approximately 47.66% of the variability in the logarithmically transformed price. Therefore, the inclusion of interaction terms in the model did slghtly improve the model.

Model4: No Pooling Model

$$lm(log(price) = \alpha + \beta_1 \cdot host\_is\_superhost + \beta_2 \cdot latitude + \beta_3 \cdot longitude + \beta_4 \cdot room\_type + \beta_5 \cdot accommodates$$
$$+\beta_6 \cdot bathrooms + \beta_7 \cdot bedrooms + \beta_8 \cdot beds + \beta_9 \cdot minimum\_nights + \beta_{10} \cdot number\_of\_reviews$$
$$+\beta_{11} \cdot review\_scores\_rating + \beta_{12} \cdot review\_scores\_location + \beta_{13} \cdot latitude * longitude$$
$$+\beta_{14} \cdot accommodates * bedrooms + \beta_{15} \cdot review\_scores\_rating * review\_scores\_location$$
$$+factor(host\_neighbourhood))$$

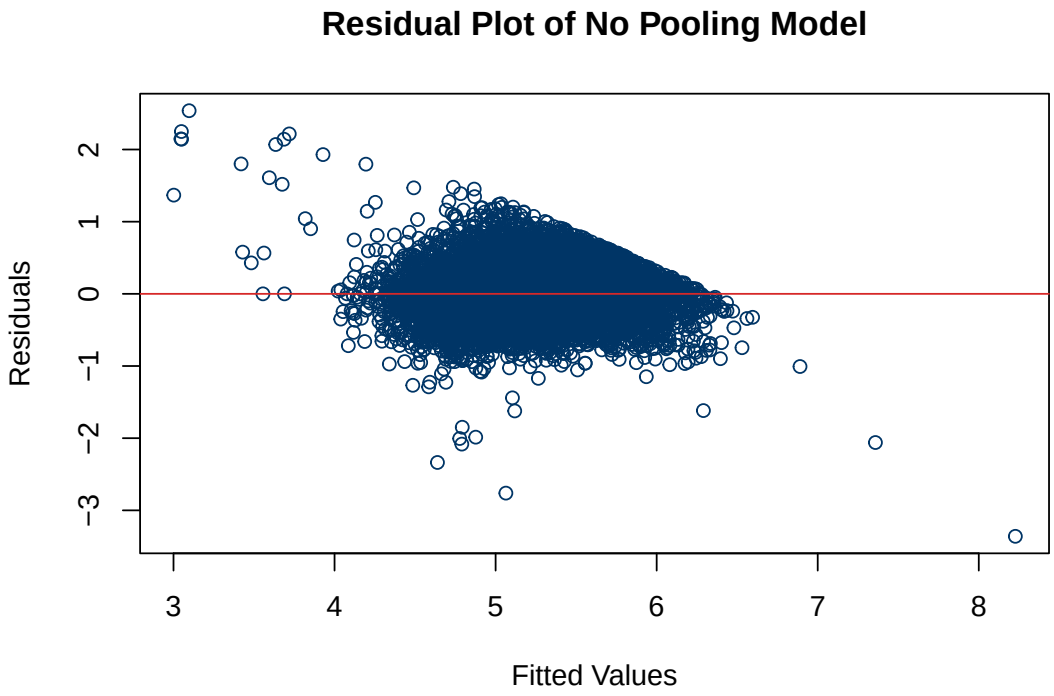
```
##                                                       Estimate   Std. Error
## (Intercept)                                      -5.262200e+03 8.124601e+02
## host_is_superhost                                  4.578235e-02 7.022021e-03
```

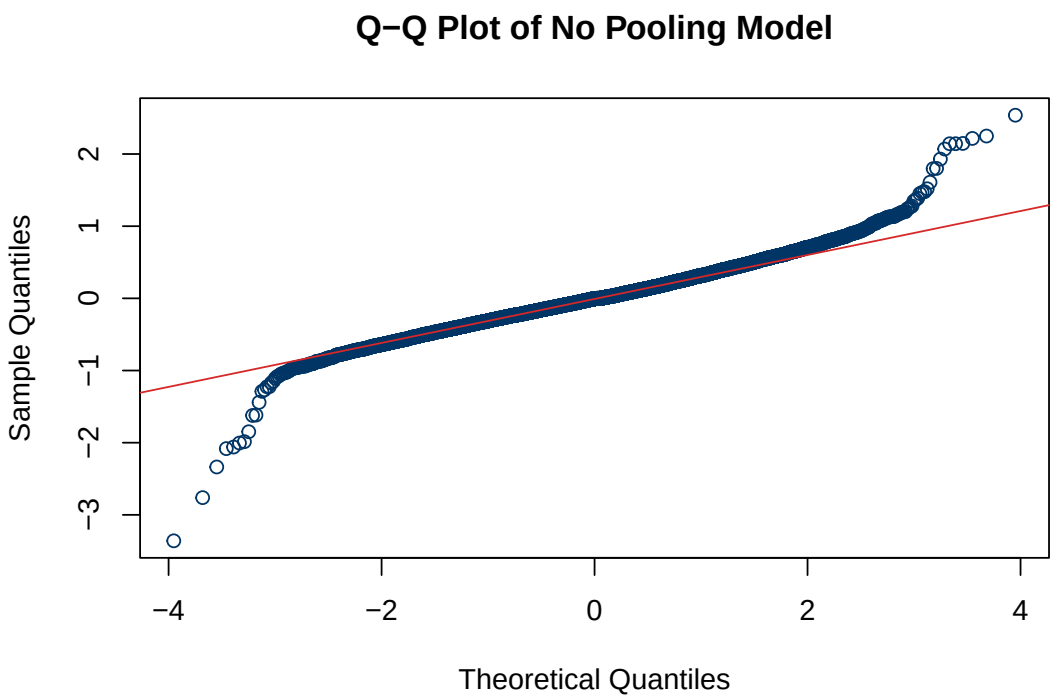
```
## latitude                                          1.518170e+02 2.384195e+01
## longitude                                        -4.465590e+01 6.870231e+00
## room_typeHotel room                              -4.144566e-02 6.135072e-02
## room_typePrivate room                            -1.383752e-01 1.640986e-02
## accommodates                                      9.869390e-02 3.632973e-03
## bathrooms                                         1.058100e-01 6.217445e-03
## bedrooms                                          2.174707e-01 7.470817e-03
## beds                                             -7.070951e-03 4.291129e-03
## minimum_nights                                   -5.853548e-03 2.047675e-04
## number_of_reviews                                -2.607702e-04 3.757819e-05
## review_scores_rating                             -2.340232e-01 2.229265e-02
## review_scores_location                           -9.471262e-02 2.078959e-02
## factor(host_neighbourhood)4th Street Corridor -3.846858e-01 1.730627e-01
## factor(host_neighbourhood)Adams Hill          -9.315552e-02 1.222926e-01
## factor(host_neighbourhood)Alamitos Beach       9.706419e-02 7.825405e-02
## factor(host_neighbourhood)Alamitos Heights     9.259089e-02 3.454199e-01
## factor(host_neighbourhood)Algiers             -2.737055e-01 3.453917e-01
## factor(host_neighbourhood)Alhambra            -2.130752e-01 3.267388e-02
## factor(host_neighbourhood)Alondra Park        -2.152120e-02 2.443397e-01
## factor(host_neighbourhood)Alphabet City        7.495600e-01 3.452795e-01
## factor(host_neighbourhood)Altadena             1.340861e-01 5.527891e-02
##                                                     t value      Pr(>|t|)
## (Intercept)                                      -6.47687216  9.720906e-11
## host_is_superhost                                 6.51982495  7.314112e-11
## latitude                                          6.36764198  1.987874e-10
## longitude                                        -6.49991318  8.346999e-11
## room_typeHotel room                              -0.67555294  4.993372e-01
## room_typePrivate room                            -8.43244335  3.760849e-17
## accommodates                                     27.16615516 7.055674e-158
## bathrooms                                        17.01824129  3.260039e-64
## bedrooms                                         29.10936394 3.248674e-180
## beds                                             -1.64780680  9.941786e-02
## minimum_nights                                  -28.58631113 4.537725e-174
## number_of_reviews                                -6.93940180  4.135149e-12
## review_scores_rating                            -10.49777356  1.135385e-25
## review_scores_location                           -4.55577247  5.269548e-06
## factor(host_neighbourhood)4th Street Corridor  -2.22281200  2.624659e-02
## factor(host_neighbourhood)Adams Hill           -0.76174261  4.462282e-01
## factor(host_neighbourhood)Alamitos Beach        1.24037278  2.148612e-01
## factor(host_neighbourhood)Alamitos Heights      0.26805313  7.886629e-01
## factor(host_neighbourhood)Algiers              -0.79244961  4.281138e-01
## factor(host_neighbourhood)Alhambra             -6.52126822  7.244307e-11
## factor(host_neighbourhood)Alondra Park         -0.08807903  9.298153e-01
## factor(host_neighbourhood)Alphabet City         2.17087871  2.995940e-02
## factor(host_neighbourhood)Altadena              2.42562752  1.529624e-02

##
## Residual standard error:  0.345055  on  12301  degrees of freedom

## Multiple R-squared:  0.5749739   Adjusted R-squared:  0.5551755

## F-statistic:  29.04145  on  573  and  12301  DF,  p-value:  NA
```





For the No Pooling Model, I included a neighborhood-specific effect, thereby capturing local variations that the Complete Pooling Model might miss. The model output suggests that all predictors are statistically significant, as indicated by their p-values.

The Residual Plot shows a random scatter of residuals around the zero line, indicating a well-fitted model with no apparent patterns that would suggest model misspecification, Which looks similar to the previous one.

The Q-Q Plot shows that the residuals are mostly aligned with the theoretical quantiles, but deviations are observed in the tails, which indicates that there may still be outliers or influential points affecting the model's assumptions.

The model's adjusted R-squared of approximately 0.5552 suggests that over 55.52% of the variability in the logarithmically transformed price is explained by the model, marking an improvement in fit compared to the Complete Pooling Model.

Model5: Partial Pooling Model

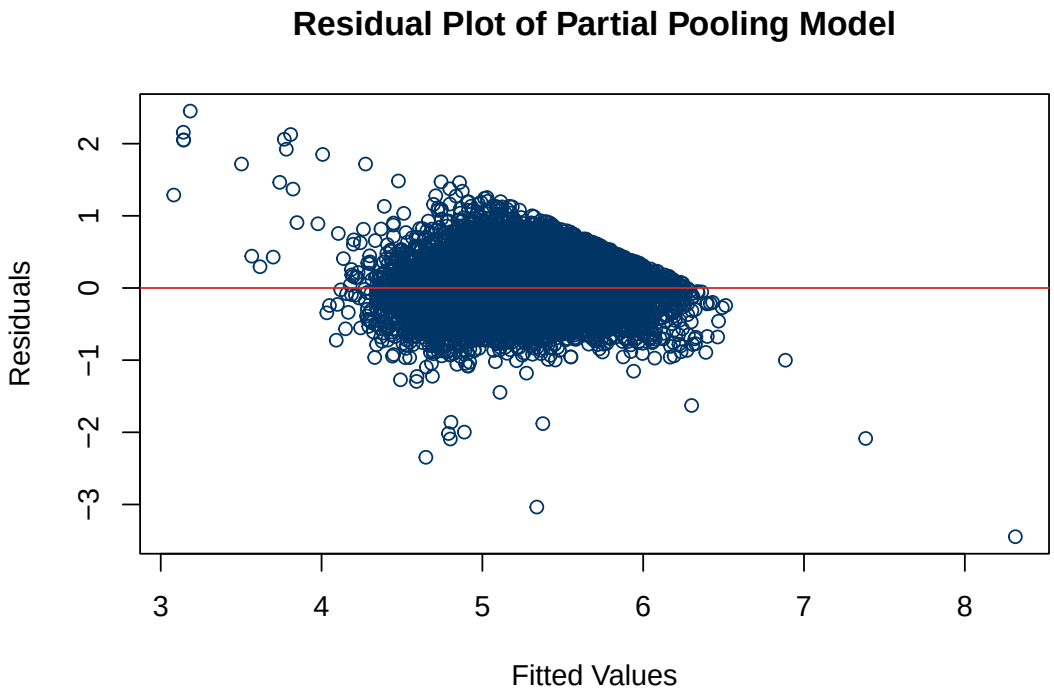
$$lm(log(price) = \alpha + \beta_1 \cdot host\_is\_superhost + \beta_2 \cdot latitude + \beta_3 \cdot longitude + \beta_4 \cdot room\_type + \beta_5 \cdot accommodates + \beta_6 \cdot bathrooms + \beta_7 \cdot bedrooms + \beta_8 \cdot beds + \beta_9 \cdot minimum\_nights + \beta_{10} \cdot number\_of\_reviews + \beta_{11} \cdot review\_scores\_rating + \beta_{12} \cdot review\_scores\_location + \beta_{13} \cdot latitude * longitude + \beta_{14} \cdot accommodates * bedrooms + \beta_{15} \cdot review\_scores\_rating * review\_scores\_location + (1|host\_neighbourhood))$$

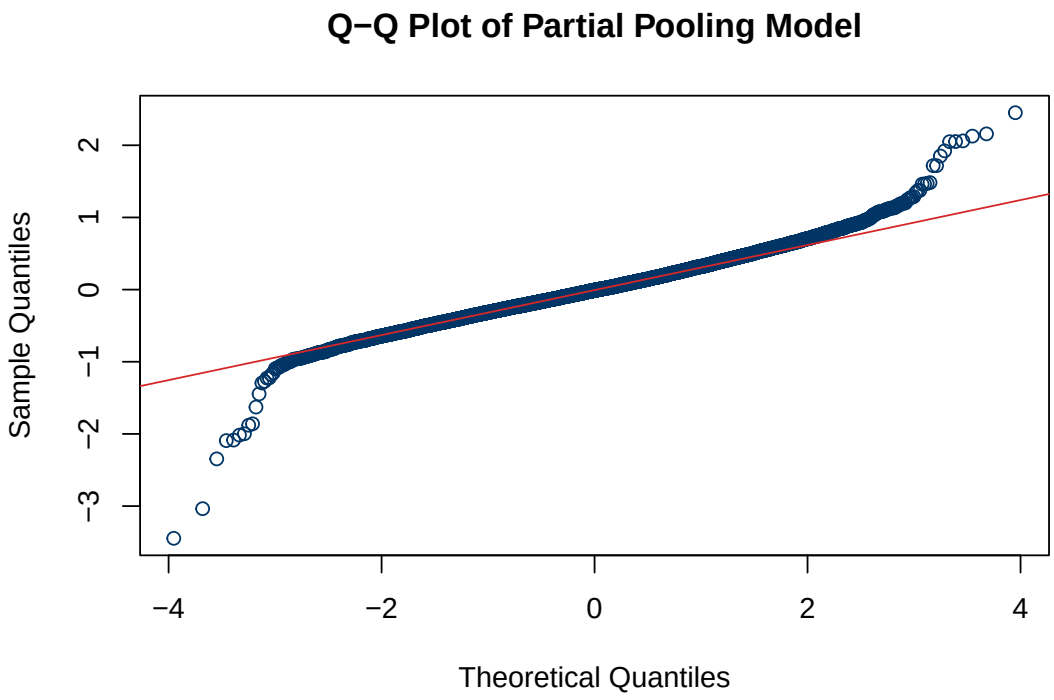
```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(price) ~ host_is_superhost + latitude + longitude + room_type +
##     accommodates + bathrooms + bedrooms + beds + minimum_nights +
##     number_of_reviews + review_scores_rating + review_scores_location +
##     latitude:longitude + accommodates:bedrooms + review_scores_rating:review_scores_location +
##     (1 | host_neighbourhood)
##    Data: train_data
##
## REML criterion at convergence: 9941.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -9.9870 -0.6292 -0.0274  0.5900  7.1031
##
## Random effects:
##  Groups             Name        Variance Std.Dev.
##  host_neighbourhood (Intercept) 0.0431   0.2076
##  Residual                       0.1191   0.3451
## Number of obs: 12875, groups:  host_neighbourhood, 558
##
## Fixed effects:
##                                          Estimate Std. Error t value
## (Intercept)                            -5.506e+03  7.634e+02  -7.213
## host_is_superhost                       4.567e-02  6.908e-03   6.611
## latitude                                1.590e+02  2.240e+01   7.096
## longitude                              -4.672e+01  6.455e+00  -7.237
## room_typeHotel room                    -5.114e-02  6.123e-02  -0.835
## room_typePrivate room                  -1.393e-01  1.617e-02  -8.615
## accommodates                            9.796e-02  3.571e-03  27.429
## bathrooms                               1.041e-01  6.041e-03  17.238
## bedrooms                                2.174e-01  7.383e-03  29.447
## beds                                   -7.074e-03  4.233e-03  -1.671
```

```
## minimum_nights                                -5.602e-03  1.962e-04 -28.557
## number_of_reviews                             -2.630e-04  3.716e-05  -7.077
## review_scores_rating                          -2.433e-01  2.214e-02 -10.988
## review_scores_location                        -9.777e-02  2.066e-02  -4.731
## latitude:longitude                             1.348e+00  1.894e-01   7.115
## accommodates:bedrooms                         -1.630e-02  9.052e-04 -18.003
## review_scores_rating:review_scores_location    5.803e-02  5.111e-03  11.355

##
## Correlation matrix not shown by default, as p = 17 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0020504 (tol = 0.002, component 1)
```





The Partial Pooling Model introduces random effects to account for variations within Airbnb's host neighborhoods, on top of the fixed effects considered in previous models. This approach allows me to capture inherent group-level variability in the price-setting behavior across different neighborhoods.

The Residual Plot and Q-Q Plot show the similar pattern to the previous one.

For the fixed effects, it can be seen that the predictors like **`host_is_superhost`**, **`latitude`**, **`longitude`**, **`room_type`**, and property characteristics such as **`accommodates`**, **`bathrooms`**, **`bedrooms`**, and **`beds`** remain significant.

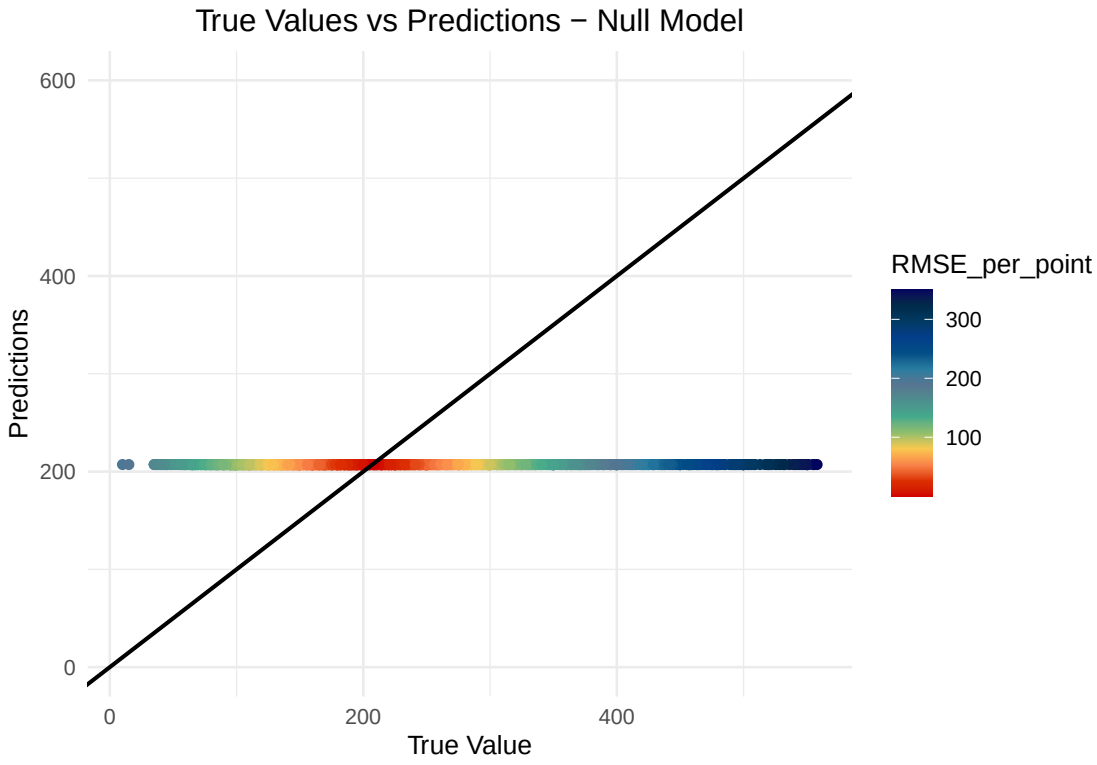
The random effects for **`host_neighborhood`** suggest that there is a significant variance in price that can be attributed to the neighborhood level, justifying the inclusion of random effects in the model.

Results Analysis

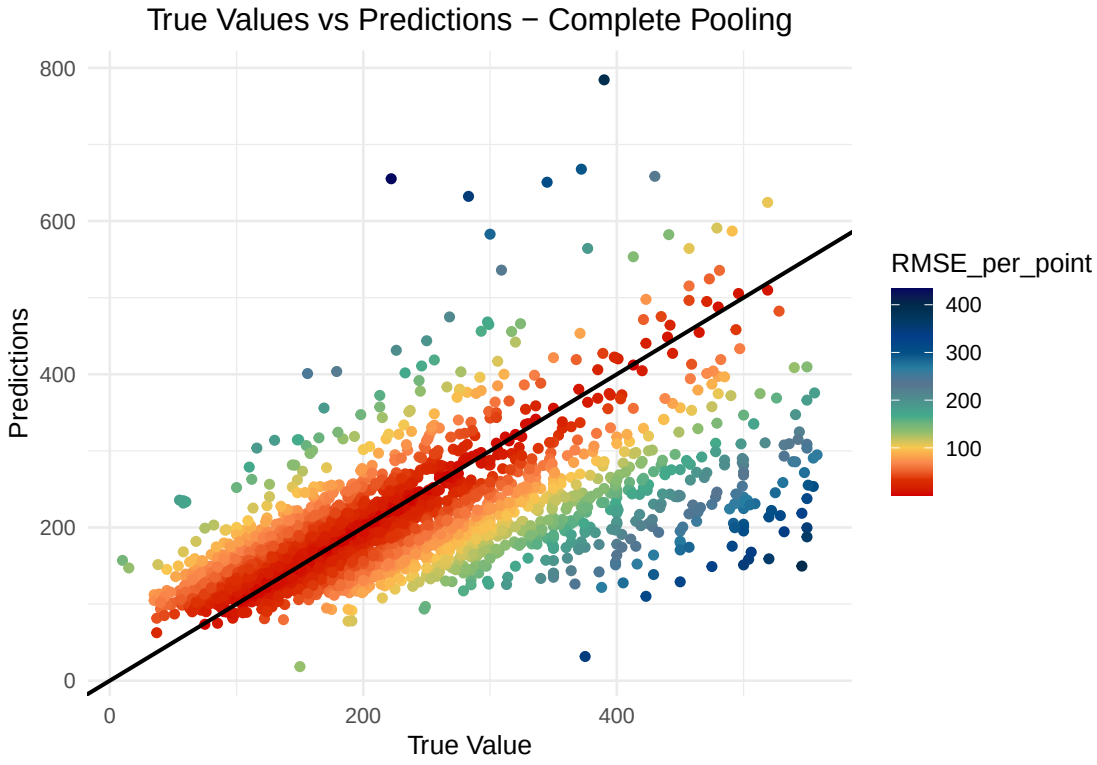
In the preceding analysis, I constructed five distinct models, each possessing unique strengths and weaknesses. To facilitate a more direct comparison of their predictive capabilities, I will deploy these models to forecast outcomes on the test set and compare the predictions against the actual values.

The Root Mean Square Error (RMSE) for each model has been computed to quantify their prediction accuracy. Additionally, I have created scatter plots to visually represent the predictive performance of each model. This comparative approach will enable an assessment of which model most effectively captures the nuances of the data, thereby guiding the selection of the optimal model for further application.

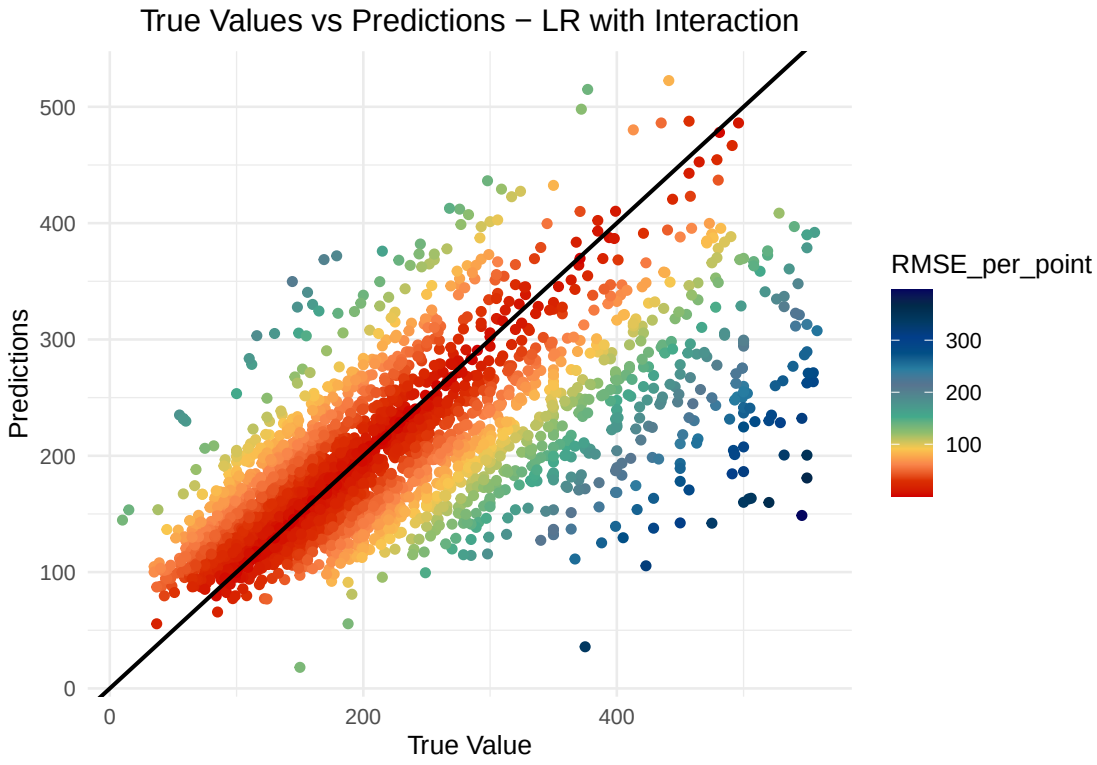
Null Model


True Values vs Predictions − Null Model
Predictions
True Value
0
200
400
600
RMSE_per_point
300
200
100

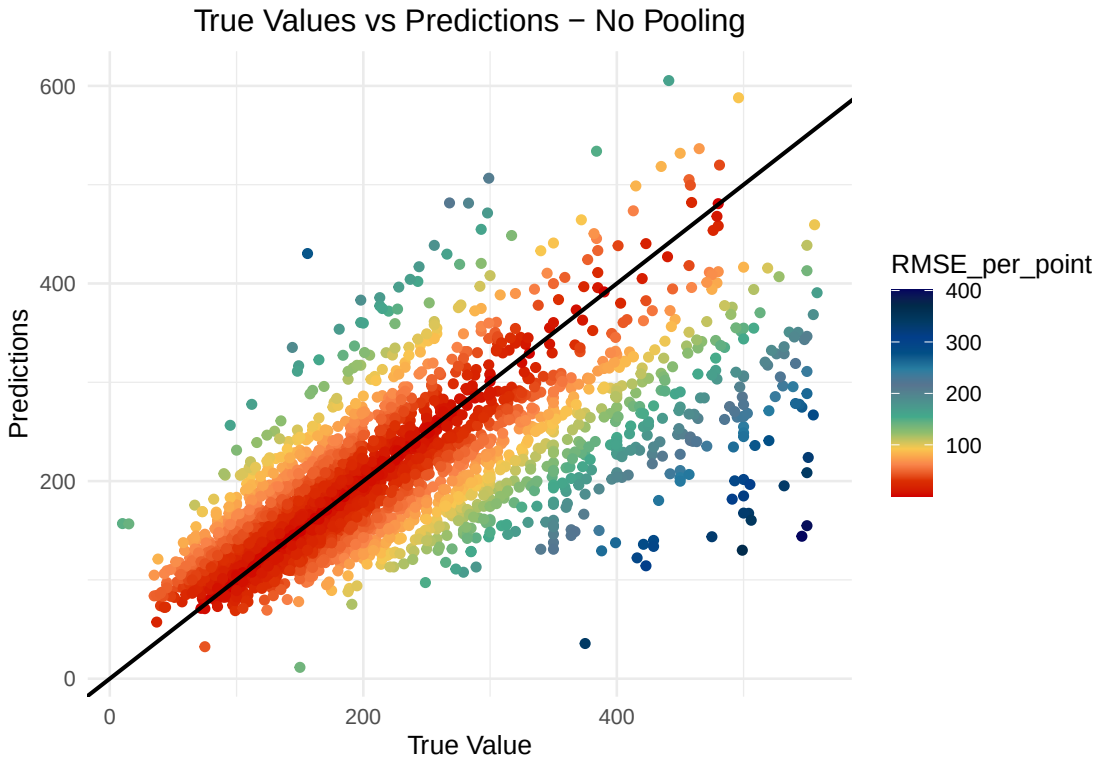
Complete Pooling Model (Linear Regression)



Linear Regression Model with Interaction Term



No Pooling Model



Partial Pooling Model

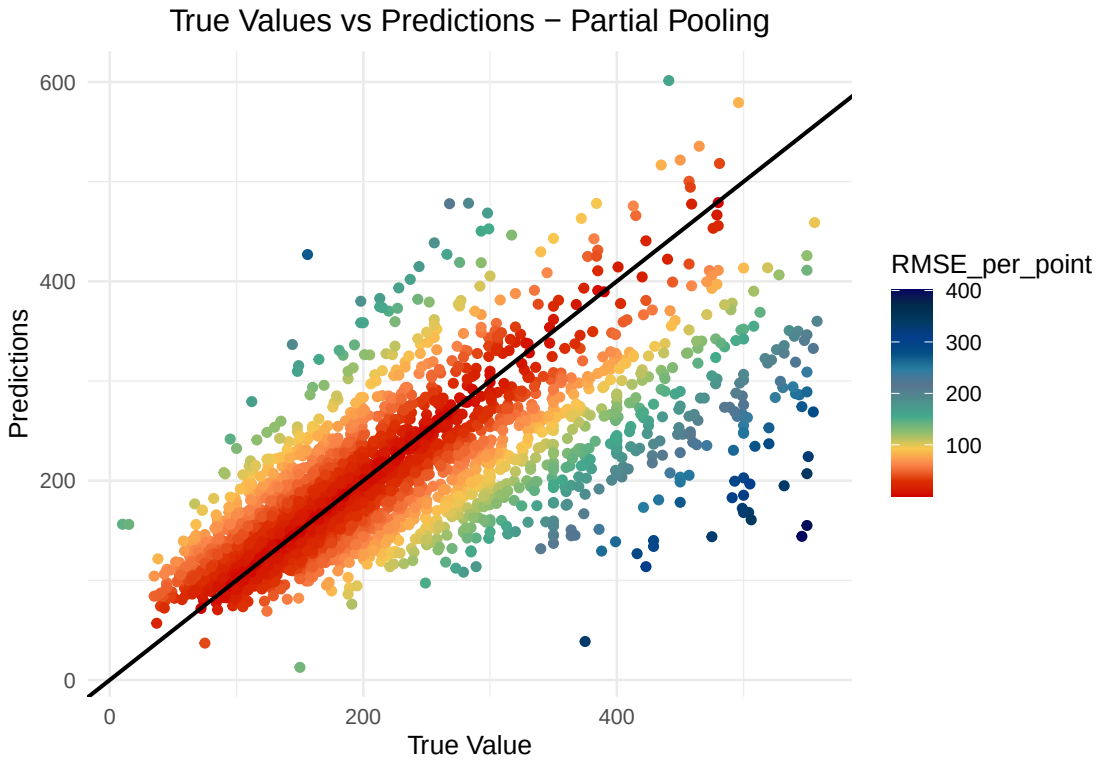


Table 2: Comparison of Different Models

Model	RMSE
Null Model	107.05792
Complete Pooling Model (Linear Regression)	84.35703
Linear Regression Model with Interaction Term	80.32188
No Pooling Model	77.19066
Partial Pooling Model	77.04541

The scatter plots illustrate the predictive accuracy of the five models compared to the actual Airbnb listing prices, with the line of perfect prediction as a reference. The color intensity represents the RMSE for each prediction point, with darker colors indicating higher errors.

The Null Model shows a horizontal line of predictions, reflecting its simplistic assumption that all listings will have the average price. This model has the highest RMSE, suggesting it's the least accurate.

The Complete Pooling Model (Linear Regression) shows a greater spread of predictions, with many points aligning closer to the line of perfect prediction. However, there are still many points with high RMSE values, indicating substantial error for those predictions.

The Linear Regression Model with Interaction Term has a slightly improved spread of predictions, with more points close to the line of perfect prediction and fewer points with very high RMSE values, indicating that considering interaction effects has enhanced the model's predictive power.

The No Pooling Model shows a further improved alignment with the line of perfect prediction, suggesting that treating each listing individually, rather than assuming they are all the same, provides a better fit.

The Partial Pooling Model appears to be the most accurate, as indicated by the denser cluster of points around the line of perfect prediction and the lower RMSE values. This suggests that accounting for both fixed effects and random effects due to neighborhoods provides the best balance between capturing the common trend and accommodating the variability in prices across different neighborhoods. The RMSE comparison chart solidifies these observations, confirming that the Partial Pooling Model has the lowest RMSE.

In conclusion, the detailed analysis points to the Partial Pooling Model as the most effective for predicting Airbnb's listing prices in Los Angeles, offering the best balance between complexity and accuracy.

Discussion and Limitation

While the Partial Pooling Model is superior in terms of predictive accuracy, the choice of model may also depend on the specific application and the interpretability requirements. If the goal is to understand specific neighborhood effects, the No Pooling or Partial Pooling models would be more appropriate. If interpretability and simplicity are more critical, the Complete Pooling or Interaction models might be preferred, despite the slight sacrifice in prediction accuracy.

These models also have some limitations, almost all the models are unduly influenced by outliers or points with high leverage that do not represent the general trend. Besides, some complex models, especially those with many predictors and interactions, are at risk of overfitting to the train data and may not generalize well to unseen data. What's more, as models become more complex, they often become harder to interpret. The Partial Pooling Model, for instance, can be difficult to explain to stakeholders without a statistical background.

Reference

[1] Deboosere, R., Kerrigan, D. J., Wachsmuth, D., & El-Geneidy, A. (2019). *Location, location and professionalization: a multilevel hedonic analysis of Airbnb listing prices and revenue.* Regional Studies, Regional Science, 6(1), 143-156.

[2] Jiao, J., & Bai, S. (2020). *An empirical analysis of Airbnb listings in forty American cities.* Cities, 99, 102618.

[3] Guttentag, D. (2019). *Progress on Airbnb: A literature review.* Journal of Hospitality and Tourism Technology, 10(4)

[4] Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories.* Cambridge University Press.

[5] Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.