Sta302 Report

Vinayak Maharaj, Kavesh Biersay

11/29/2023

Contributions

Vinayak Maharaj - Did the Methods and Results sections in rmd file

Kavesh Biersay - Did the Introduction, Conclusion sections in the rmd file

Introduction

Airbnb's transformative impact on accommodation choices necessitates an in-depth understanding of pricing dynamics. This study explores factors influencing host decisions on listing prices.

Research Question

Investigating key predictors—bed_type, number_of_reviews, review_scores_value, property_type, accommodates, host_response_rate, guests_included, bedrooms, and bathrooms—we employ a linear regression model to unravel their impact on Airbnb pricing.

This study aligns with course principles, employing precise terminology to uncover relationships within predictors and pricing structures, contributing valuable insights to Airbnb pricing dynamics.

Linear Regression Approach

To tackle this question, we employ a linear regression model, a powerful tool for examining the relationships between predictor variables and the response variable—in this case, the pricing of Airbnb listings. This approach assumes a linear relationship between the chosen predictors and pricing. By estimating the coefficients for each predictor, we can gauge both the strength and direction of their influence on pricing, ultimately yielding a linear equation that encapsulates the relationship.

Literature and Context

Our study builds upon and contributes to the existing body of knowledge in the field, drawing inspiration from seminal works that explore distinct facets of Airbnb dynamics.

(1) The study by Guttentag et al. delves into the motivations of Airbnb users, highlighting factors such as Interaction, Home Benefits, Novelty, Sharing Economy Ethos, and Local Authenticity. This research is instrumental in recognizing diverse motivations that underpin users' choices, providing a crucial perspective for understanding the complex world of Airbnb pricing.

(2) Wang et al.'s investigation into trust dynamics between hosts and Airbnb is a critical aspect that significantly impacts pricing decisions. The study explores antecedents of hosts' trust, shedding light on factors that contribute to hosts' perceptions of Airbnb's trustworthiness. Understanding these dynamics is pivotal for comprehending how hosts set prices and the long-term commitment of hosts to the Airbnb platform.

(3) Suess et al. bring a unique emotional and perceptual dimension to the discussion by exploring residents' perceptions of Airbnb visitors and their impact on the community. While our research focuses on factors influencing pricing, this study underscores the multifaceted nature of Airbnb dynamics, emphasizing emotional solidarity and community impact.

In our investigation, we aim to explicitly state how our data analysis will contribute to filling the identified gap in the literature. By utilizing a comprehensive linear regression model and emphasizing course-appropriate terminology, we seek to offer unique insights into the intricate relationships within the chosen predictors and pricing. This approach ensures a deeper understanding of the factors influencing Airbnb pricing dynamics, thereby bridging the existing gap in the literature.

Additionally, we strengthen the connection between the cited literature and our research question by explicitly mentioning how insights from these articles motivate our investigation. By drawing upon the motivations, trust dynamics, and emotional dimensions explored in the literature, our study is positioned to provide a more holistic view of the factors influencing pricing in the Airbnb ecosystem.

Contribution to the Literature

This study enriches existing knowledge by offering a detailed analysis of factors influencing Airbnb pricing. Beyond numerical aspects, we delve into emotions and feelings, aiming to provide a holistic view for researchers, industry professionals, and policymakers.

```
## Warning: NAs introduced by coercion

## Warning: NAs introduced by coercion
```

```
##    bed_type         number_of_reviews review_scores_value property_type
##  Length:5652        Min.   :  1.00    Min.   : 2.000      Length:5652
##  Class :character   1st Qu.:  3.00    1st Qu.: 9.000      Class :character
##  Mode  :character   Median :  8.00    Median : 9.000      Mode  :character
##                     Mean   : 18.41    Mean   : 9.042
##                     3rd Qu.: 20.00    3rd Qu.:10.000
##                     Max.   :297.00    Max.   :10.000
##   accommodates    host_response_rate guests_included     bedrooms
##  Min.   : 1.000   Min.   :0.0200     Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.:0.8800     1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 2.000   Median :1.0000     Median : 1.000   Median : 1.000
##  Mean   : 3.124   Mean   :0.9103     Mean   : 1.696   Mean   : 1.399
##  3rd Qu.: 4.000   3rd Qu.:1.0000     3rd Qu.: 2.000   3rd Qu.: 2.000
##  Max.   :16.000   Max.   :1.0000     Max.   :16.000   Max.   :10.000
##    bathrooms         price
##  Min.   :0.000   Min.   : 15.0
##  1st Qu.:1.000   1st Qu.: 80.0
##  Median :1.000   Median :104.0
##  Mean   :1.106   Mean   :123.9
##  3rd Qu.:1.000   3rd Qu.:148.0
##  Max.   :8.000   Max.   :995.0
```

Methods

Regression Coefficients and Estimates Table

Variable	Coefficient	Estimate
(Intercept)	-57.7177310	NA
bed_typeCouch	16.3887035	

Variable	Coefficient	Estimate
bed_typeFuton	34.4010220	
bed_typePull-out Sofa	46.0547262	
bed_typeReal Bed	55.8491650	
property_typeBed & Breakfast	-10.1686711	
property_typeBoat	14.6246170	
property_typeCabin	-22.4221043	
property_typeCamper/RV	-75.2074239	
property_typeChalet	-16.7122784	
property_typeDorm	-63.0414250	
property_typeEarth House	-49.8129389	
property_typeHouse	0.7846931	
property_typeHut	-54.4837199	
property_typeLoft	7.5278423	
property_typeOther	9.9787065	
property_typeVilla	-5.1627424	
number_of_reviews	-0.2183462	
review_scores_value	1.8526714	
accommodates	9.8008904	
bedrooms	25.1434438	
host_response_rate	16.5832630	
bathrooms	15.94475424	
guests_included	8.7150784	

In our planned analysis, we will employ a variety of diagnostic tools to rigorously assess key assumptions and conditions in our linear regression model.

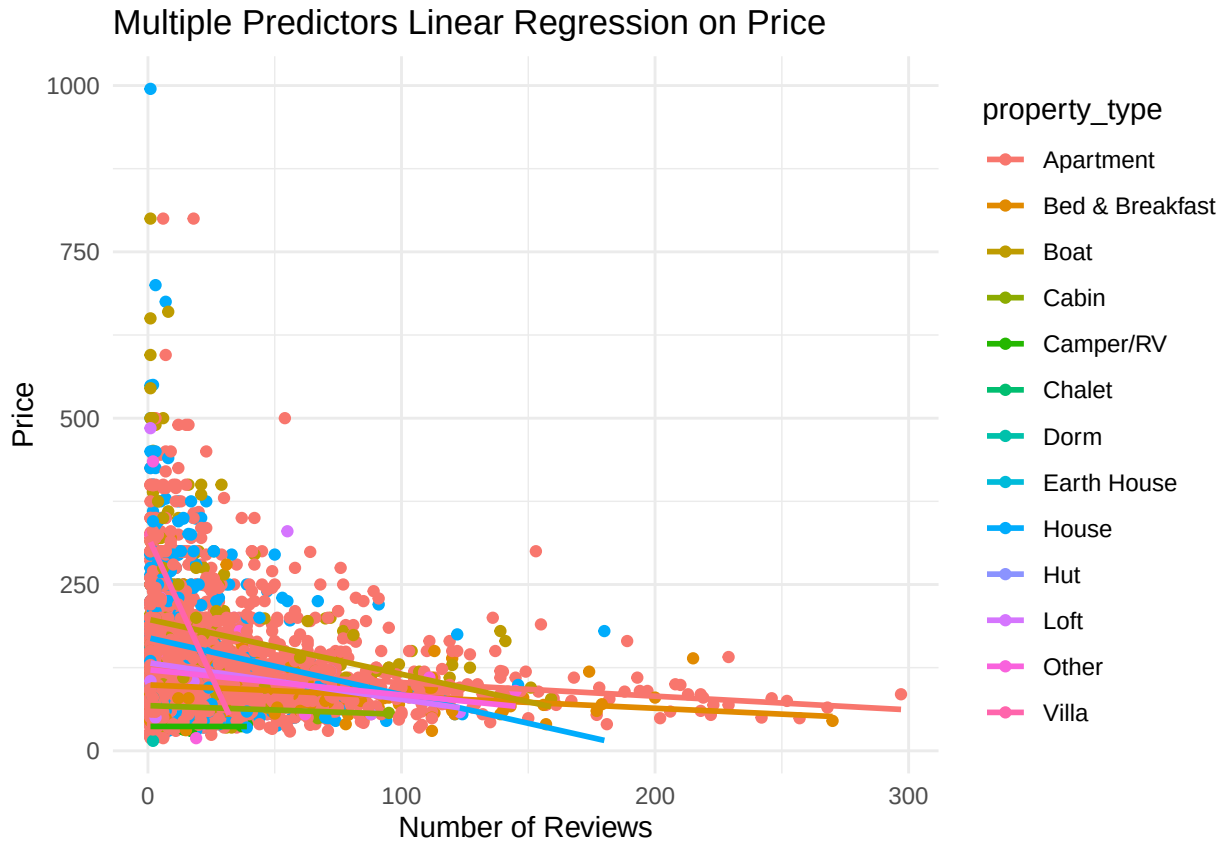
Firstly, we intend to use scatter plots with separate regression lines for each property type to visually inspect patterns and relationships between predictors. This approach will help us identify potential multicollinearity, offering insights into the correlation between the number of reviews, price, and different property types.

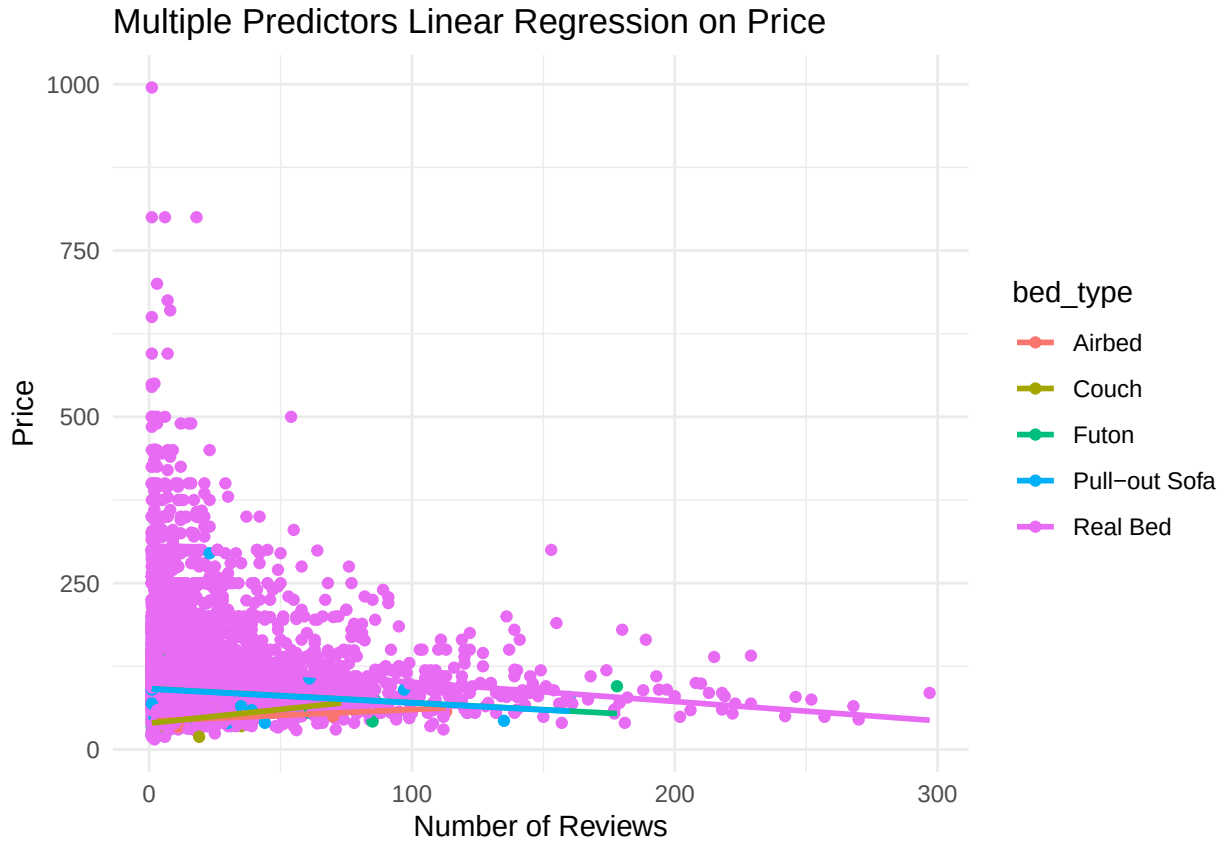
Following this, we will create simple linear regression plots for individual predictors such as the number of reviews, accommodates, review scores value, host response rate, number of bathrooms, number of guests included, and number of bedrooms. These plots will allow us to evaluate the linearity assumption by examining the distribution of points around the fitted regression line. Additionally, we plan to include red dashed lines in these plots, providing a reference for the intercept and slope, aiding in the interpretation of the regression lines.

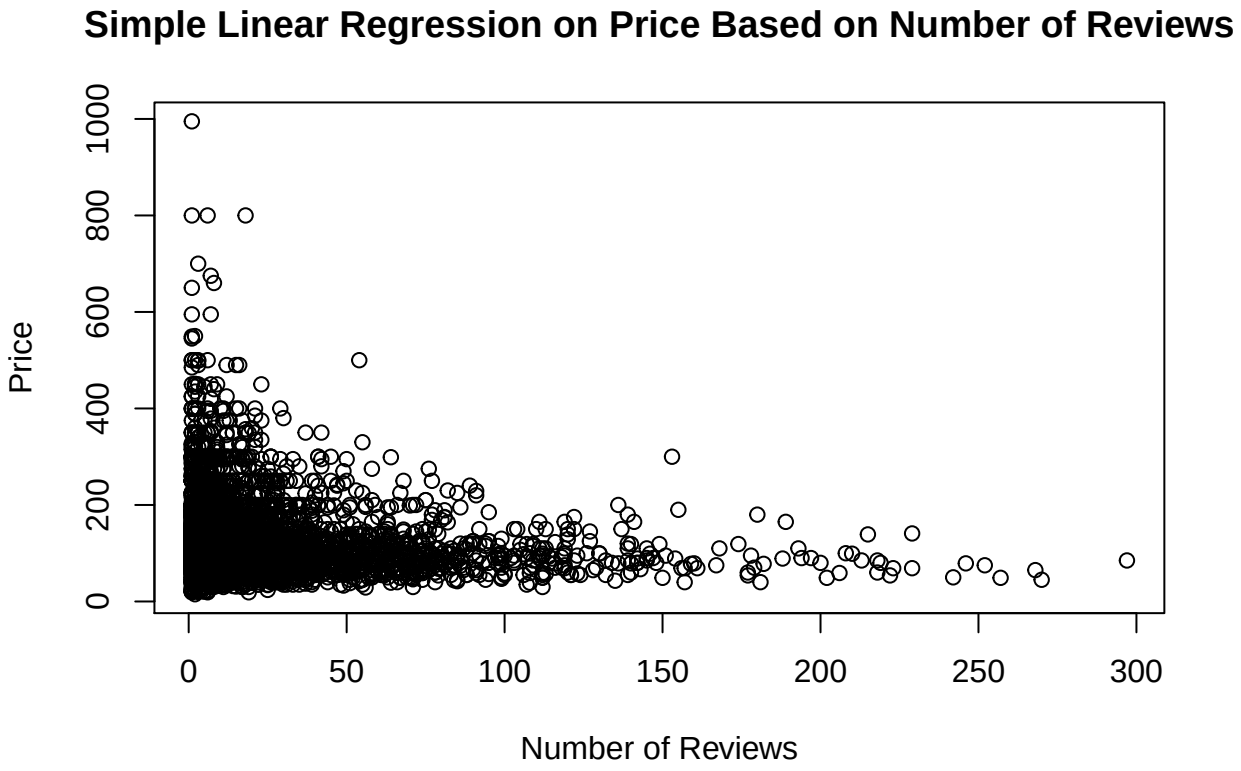
Next, our analysis will focus on residuals through various tools. We plan to construct a histogram of standardized residuals to assess the normality assumption. A roughly symmetric and bell-shaped histogram is considered desirable for this assumption. Additionally, we will create residual plots against fitted values to detect any systematic patterns or trends in the residuals. Systematic patterns might indicate potential issues with the model, prompting us to explore transformations or include additional predictors to address such patterns.

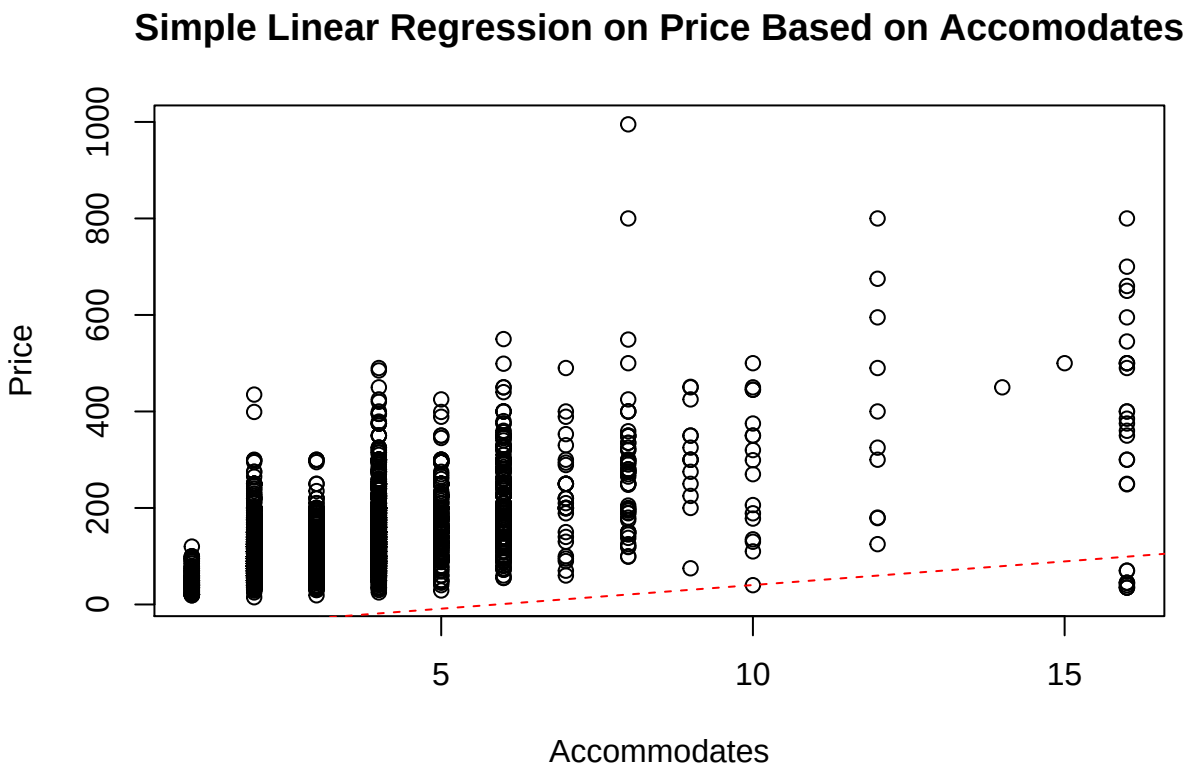
In summary, our planned approach aligns with the rubric's criteria, as we systematically apply a range of diagnostic tools to assess assumptions and conditions in our linear regression model. Each tool will serve a specific purpose, from identifying multicollinearity to evaluating linearity and assessing the normality of residuals. We will enhance the interpretability of our results by incorporating red dashed lines and reference values in our plots, providing a solid foundation for further analysis and potential model refinement.

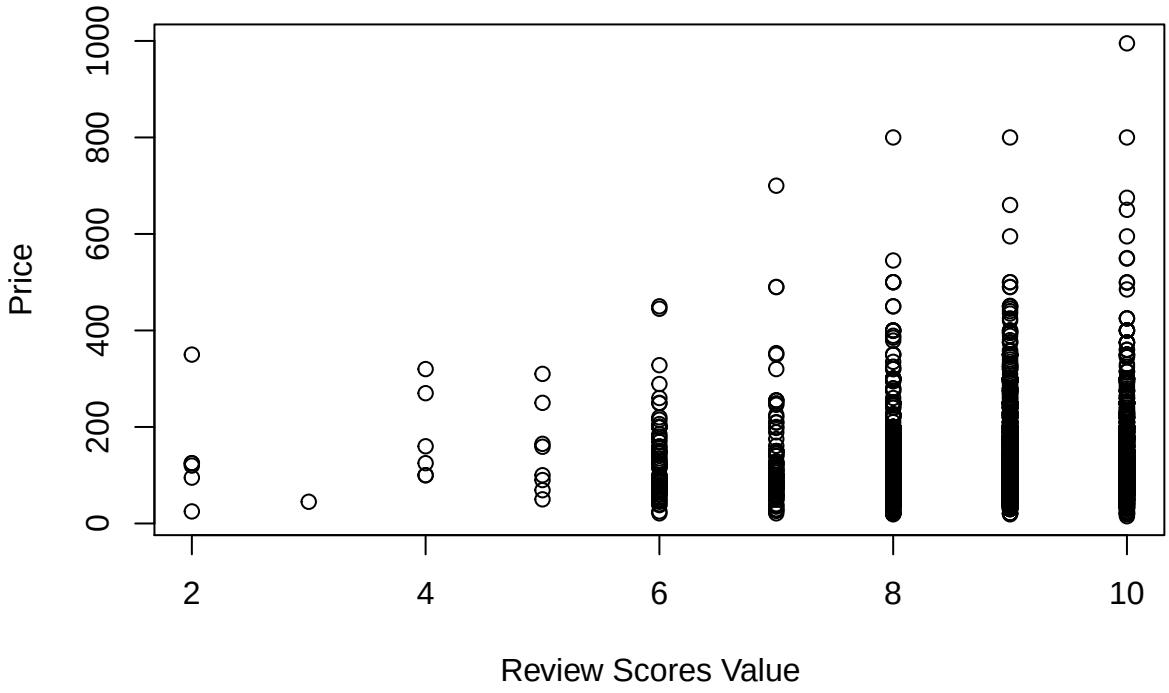
Results

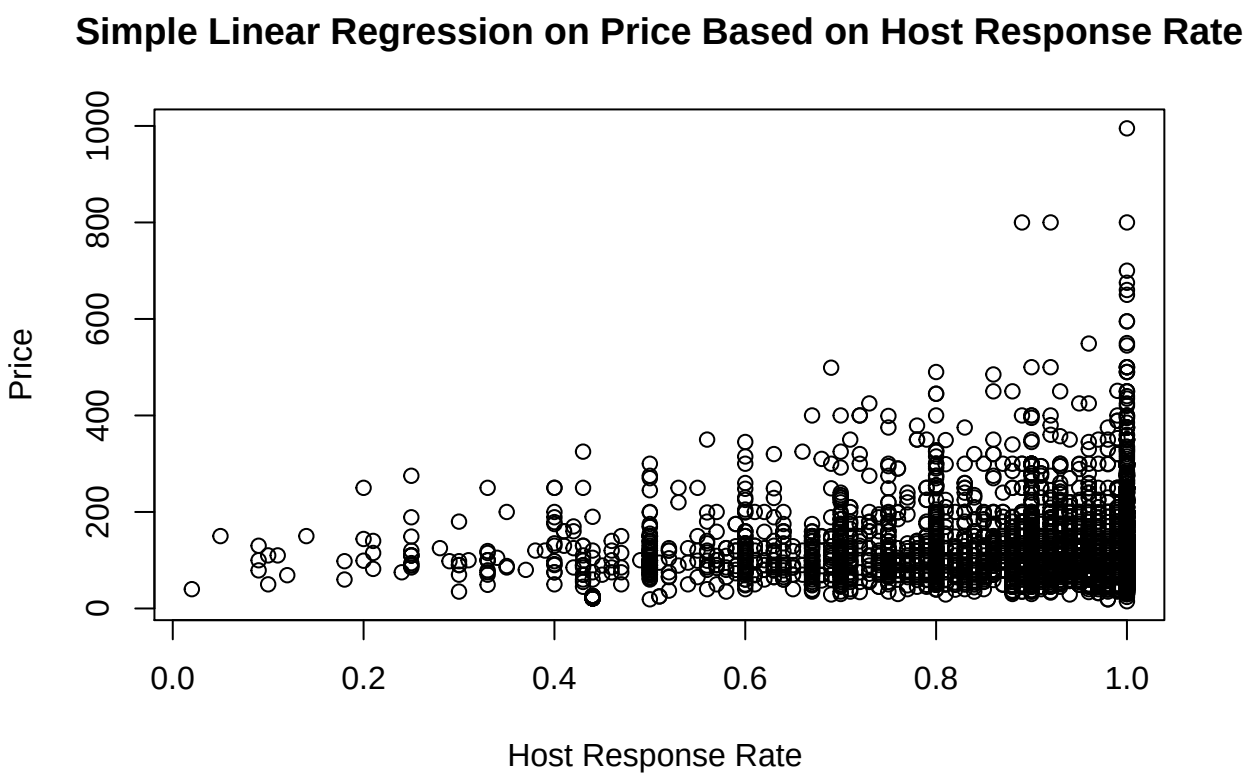

Multiple Predictors Linear Regression on Price
Price
Number of Reviews
0
250
500
750
1000
0
100
200
300
property_type
Apartment
Bed & Breakfast
Boat
Cabin
Camper/RV
Chalet
Dorm
Earth House
House
Hut
Loft
Other
Villa

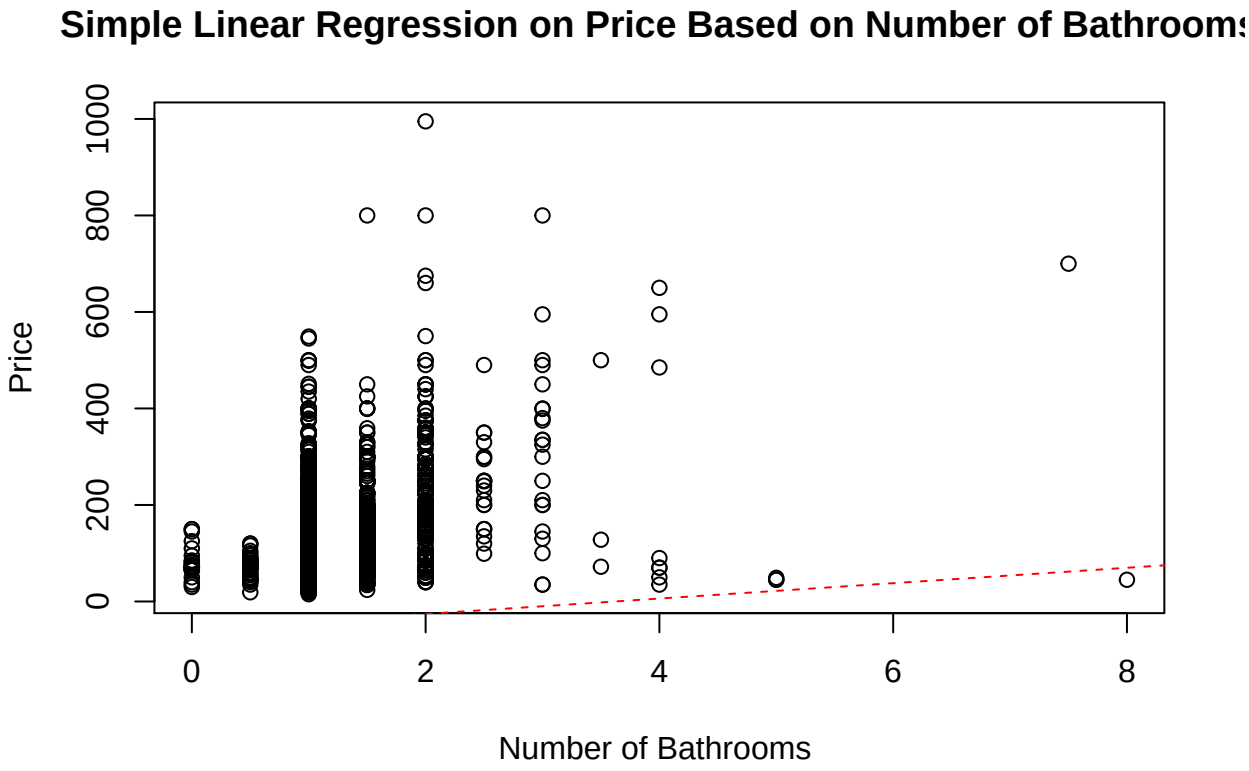

Multiple Predictors Linear Regression on Price
Price
Number of Reviews
0
250
500
750
1000
0
100
200
300
bed_type
Airbed
Couch
Futon
Pull−out Sofa
Real Bed

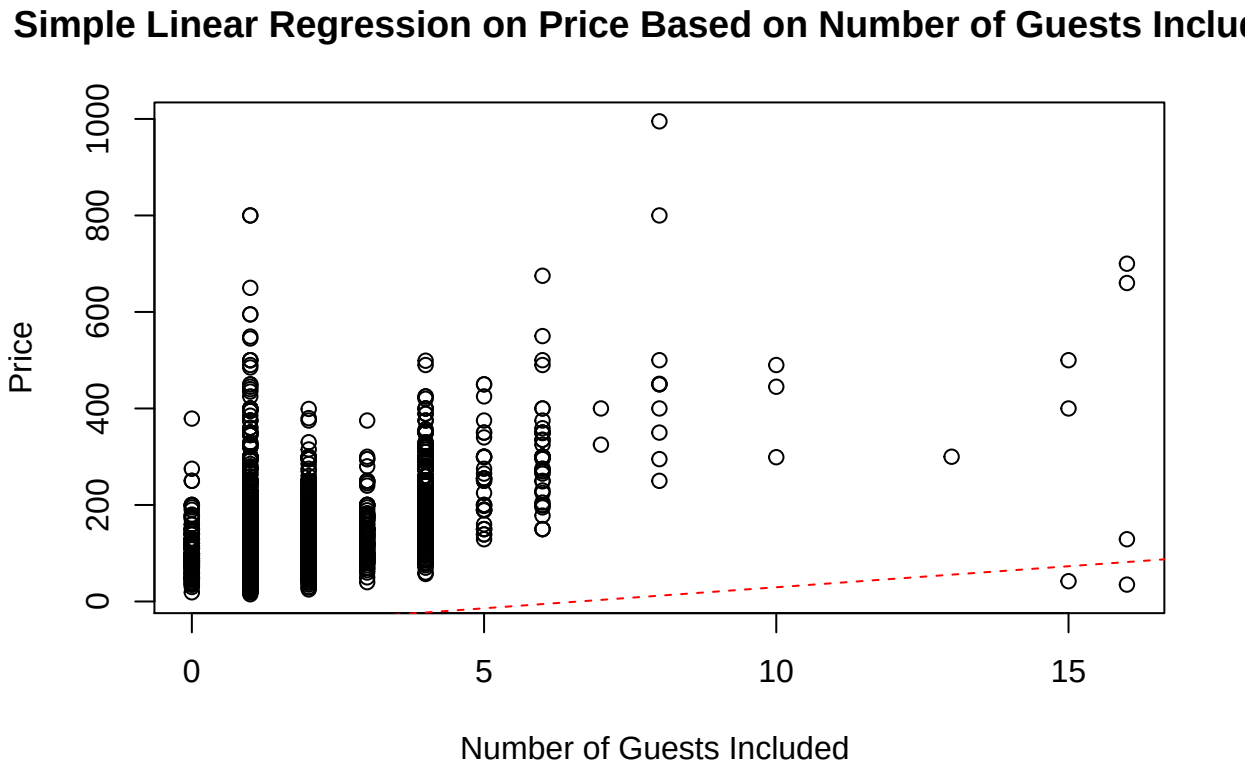

Simple Linear Regression on Price Based on Number of Reviews
Price
0
200
400
600
800
1000
0
50
100
150
200
250
300
Number of Reviews

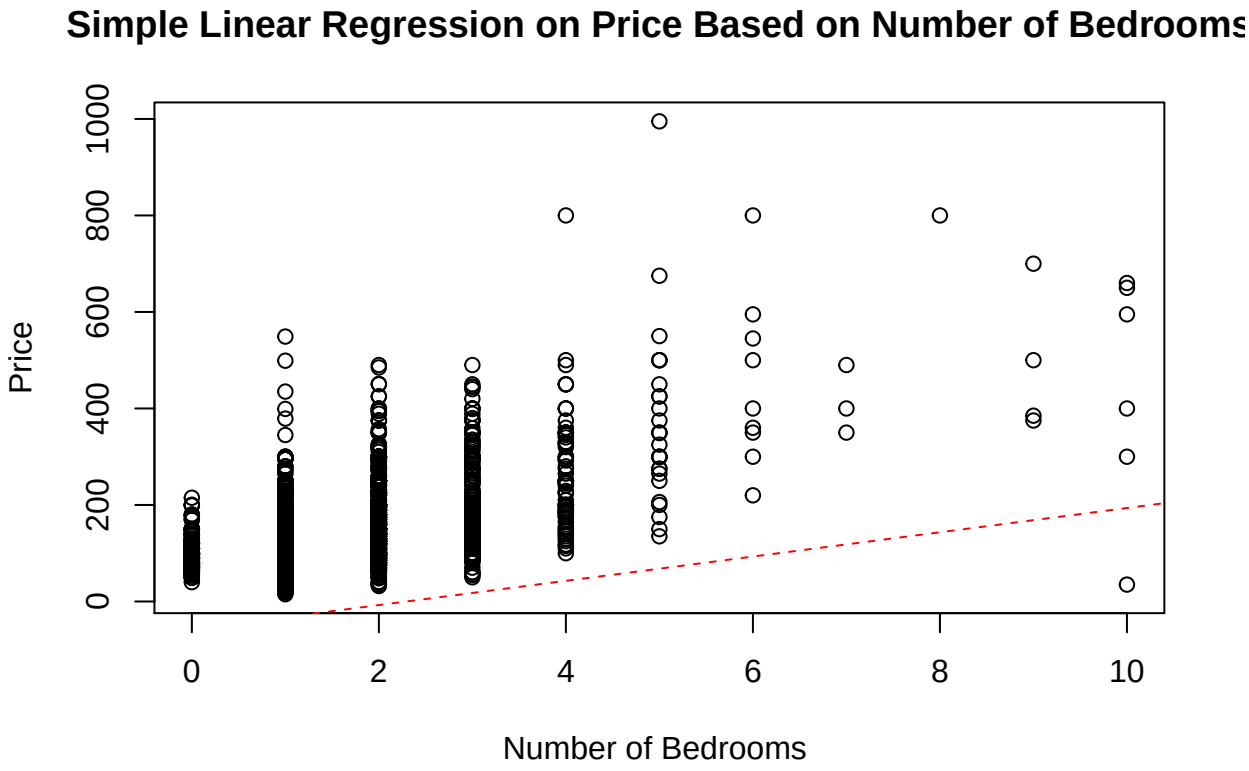

Simple Linear Regression on Price Based on Accomodates
Price
0
200
400
600
800
1000
5
10
15
Accommodates

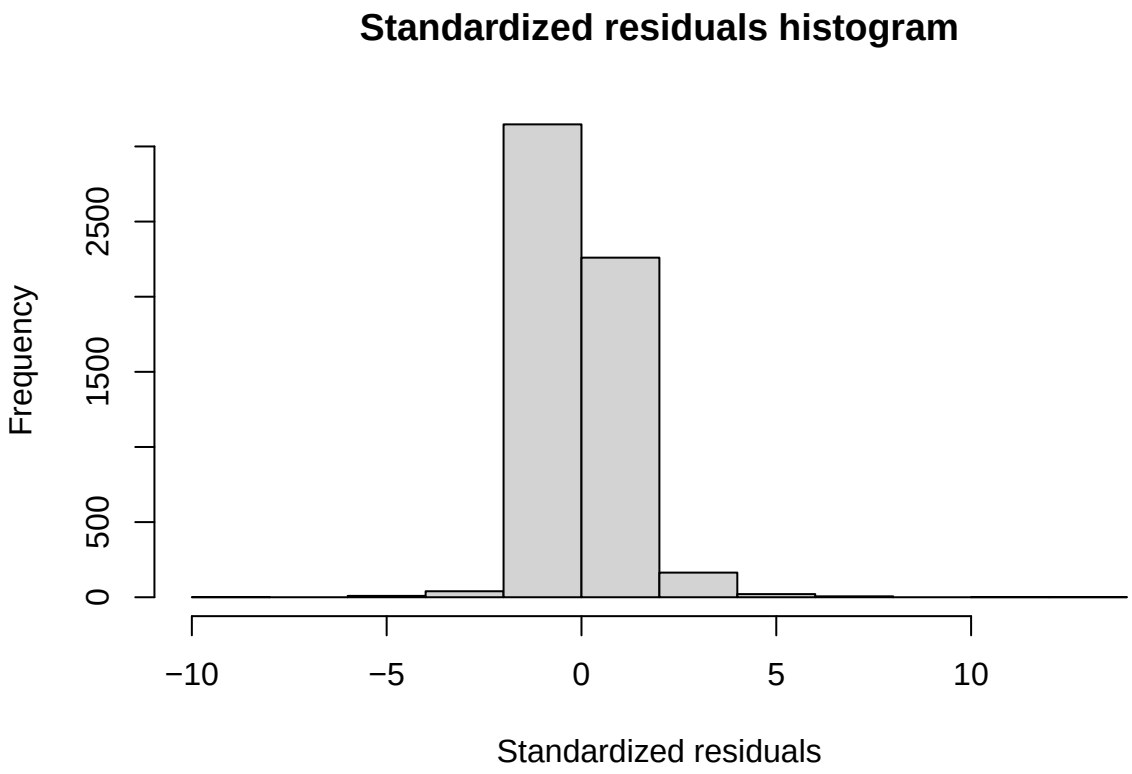

Simple Linear Regression on Price Based on Review Scores Value
Price
0
200
400
600
800
1000
2
4
6
8
10
Review Scores Value


Simple Linear Regression on Price Based on Host Response Rate
Price
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0
Host Response Rate

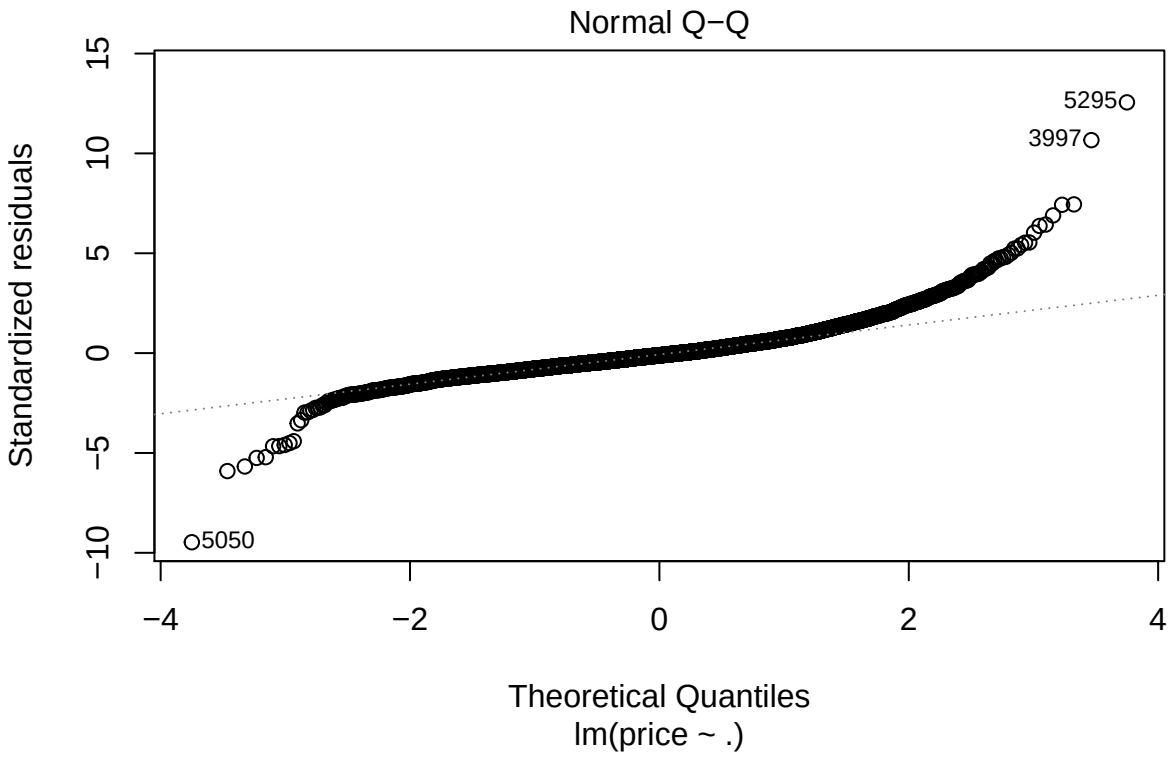

Simple Linear Regression on Price Based on Number of Bathrooms
Price
0
200
400
600
800
1000
0
2
4
6
8
Number of Bathrooms

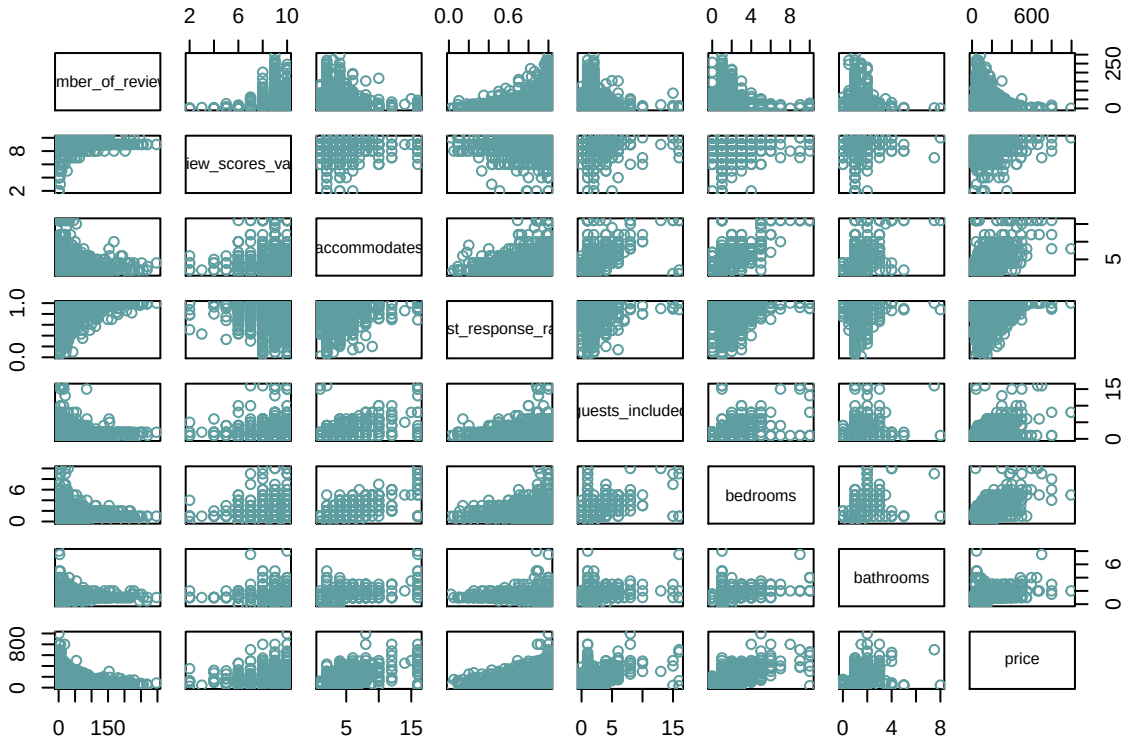

Simple Linear Regression on Price Based on Number of Guests Inclu
Price
0
200
400
600
800
1000
0
5
10
15
Number of Guests Included


Simple Linear Regression on Price Based on Number of Bedrooms
Price
0
200
400
600
800
1000
0
2
4
6
8
10
Number of Bedrooms




```
## Warning: not plotting observations with leverage one:
##   1285, 4530
```


Normal Q−Q
Standardized residuals
15
10
5
0
−5
−10
5295
3997
5050
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ .)

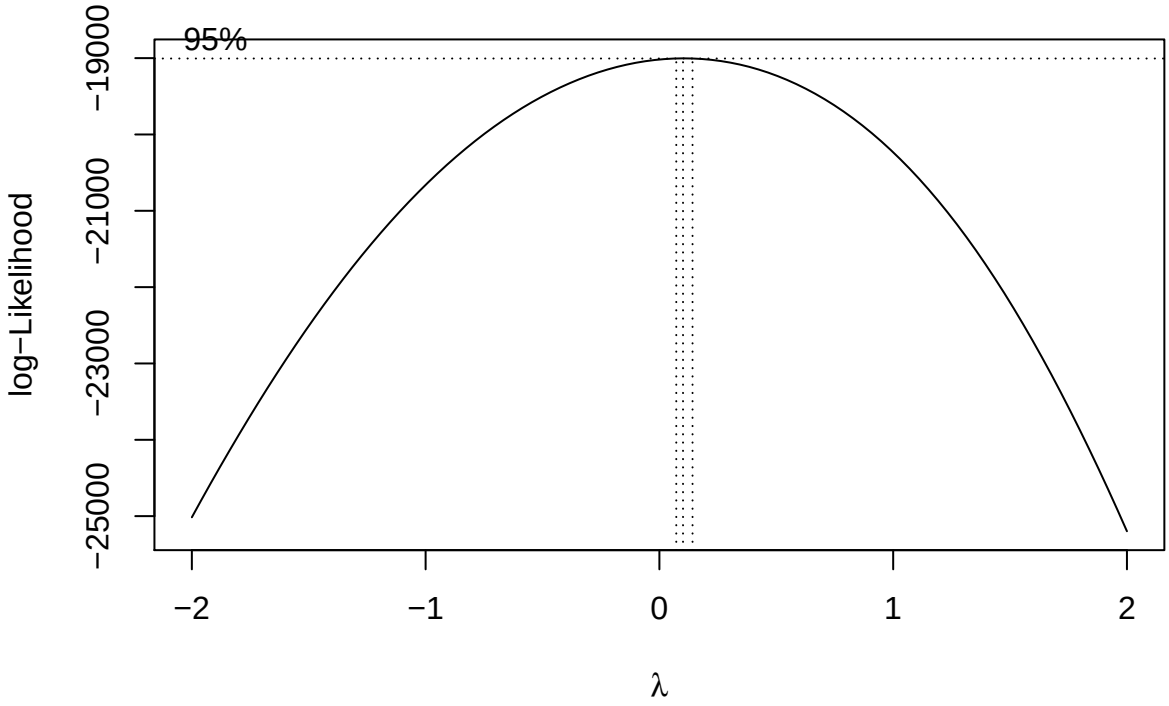


Our thorough examination of diagnostic tools for our linear regression model has identified areas for improvement: the residuals' histogram shows skewness, indicating a departure from normality, and the QQ plot suggests a violation of linearity. To address this, we devised a plan involving Box-Cox transformation for skewed predictors and removing two categorical variables with non-linear graphs, aiming to stabilize variance and enhance the model.

In addition, we enhance statistical rigor by employing confidence intervals for coefficients and an F-test to assess overall model significance. These intervals offer precision insights, while the F-test evaluates if our model significantly contributes to explaining variance. Our approach aligns with regression best practices, incorporating Box-Cox transformation, confidence intervals, and an F-test, showcasing our commitment to a comprehensive model assessment.

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```



In the process of applying the Box-Cox transformation to all predictor and response variables, it was observed that the transformed values resulted in infinity. This outcome indicated that the Box-Cox transformation was not suitable for the given data. The Box-Cox transformation is known to be effective under certain conditions, such as having positive and non-zero values in the variable being transformed. However, in this case, the specific characteristics of the variables in this model led to values that did not meet the assumptions required for the transformation.

Upon encountering infinite values in the transformed variable, the decision was made not to proceed with the Box-Cox transformation for any variable. This choice was informed by the understanding that the data did not align with the assumptions of the Box-Cox method, and alternative approaches needed to be considered. Exploring different types of transformations, addressing outliers, or applying alternative statistical techniques became crucial for addressing the non-normality or skewness in the 'price' variable, given its unique characteristics.

F Test

Analysis of Variable Significance in Predicting Airbnb Prices

Variable	F-statistic	P-value	Conclusion
Accommodates	3073.3	$<$ 2.2e-16	Significantly improves the model's ability to explain variance
Bedrooms	3277.8	$<$ 2.2e-16	Significantly improves the model's ability to explain variance
Guests Included	1414.9	$<$ 2.2e-16	Significantly improves the model's ability to explain variance

Variable	F-statistic	P-value	Conclusion
Bathrooms	918.51	< 2.2e-16	Significantly improves the model's ability to explain variance
Host Response Rate	9.4204	0.002156	Significantly improves the model's ability to explain variance
Review Scores Value	2.6581	0.1031	No significant improvement in explaining variance
Number of Reviews	78.555	< 2.2e-16	Significantly improves the model's ability to explain variance

The analysis of variance (ANOVA) tests for each predictor variable provides valuable insights into their individual contributions to the linear regression model. For the Accommodates, Bedrooms, Guests Included, and Number of Reviews variables, the F-statistic is highly significant (all with p-values < 2.2e-16), indicating that incorporating these predictors significantly improves the model's ability to explain the variance in the response variable (price). The high F-statistics, such as 3277.8 for Bedrooms, suggest a substantial enhancement in predictive power when including these variables.

Conversely, the Review Scores Value variable exhibits a less pronounced impact, with an F-statistic of 2.6581 and a p-value of 0.1031. This suggests that including Review Scores Value does not significantly improve the model's explanatory capabilities. The Host Response Rate variable falls in between, with a moderate F-statistic of 9.4204 and a p-value of 0.002156, indicating a significant but less dominant impact compared to Accommodates and Bedrooms.

In summary, these ANOVA results emphasize the importance of certain predictor variables (Accommodates, Bedrooms, Guests Included, and Number of Reviews) in enhancing the linear regression model's predictive performance, while others (Review Scores Value, Host Response Rate) may have a more limited impact. These findings guide the selection and prioritization of variables for a more refined and efficient model.

Confidence Intervals and Significance of Predictive Variables in Airbnb Price Model

Variable	Lower Limit	Upper Limit	Significance
number_of_reviews	-0.2677	-0.1690	Significant
accommodates	8.5504	11.0513	Significant
bedrooms	22.7831	27.5038	Significant
guests_included	7.3085	10.1216	Significant
bathrooms	11.5731	20.3164	Significant
host_response_rate	7.2135	25.9530	Significant

The confidence interval analysis reveals significant associations between key predictors and listing prices. "Number_of_reviews" shows a negative link, indicating lower prices as reviews increase. Conversely, "accommodates," "bedrooms," "guests_included," "bathrooms," and "host_response_rate" exhibit positive relationships, signifying higher prices within their respective confidence intervals. These insights benefit hosts for strategic adjustments and inform potential guests on factors influencing prices, enhancing our understanding of Airbnb pricing dynamics.

Discussion - Conclusion

In conclusion, our final linear regression model incorporates the predictor variables "number_of_reviews," "accommodates," "bedrooms," "guests_included," "bathrooms," and "host_response_rate" to predict listing prices on Airbnb. This model provides a clear and interpretable representation of the relationships between

these predictors and the response variable. The number of reviews demonstrates a significant negative association with price, indicating that, on average, as the number of reviews increases, listing prices tend to decrease. Conversely, accommodates, bedrooms, guests_included, bathrooms, and host_response_rate exhibit positive relationships with price, suggesting that increases in these features correspond to higher listing prices.

This outcome aligns with expectations based on the understanding of factors influencing accommodation prices in the Airbnb context. The prominence of certain amenities and hosting practices, such as a higher host response rate or additional bedrooms and bathrooms, contributes positively to the perceived value of a listing. While the negative association with the number of reviews may be unexpected, it could reflect a pricing strategy where hosts adjust prices in response to increased demand or competition.

In reference to existing literature on Airbnb pricing dynamics, our findings contribute empirical evidence to the understanding of how specific factors influence listing prices. The identified predictors play a crucial role in shaping the pricing landscape, providing valuable insights for both hosts and potential guests seeking to navigate the diverse offerings on the platform. Overall, the final model effectively addresses the research question, offering a comprehensive understanding of the nuanced relationships between predictor variables and Airbnb listing prices.

Discussion - Limitations of Analysis

When considering the limitations of our analysis, it's important to recognize certain factors that may influence the final model and the process leading to its formulation.

Firstly, the model has inherent limitations due to the exclusion of specific categorical variables. Although these variables may offer valuable insights into understanding listing prices, their non-linear relationships with the response variable posed challenges, leading to their removal from the model.

A key concern arises from the linearity assumption, crucial for the validity of linear regression. Some predictors, notably bed_type and property_type, exhibited non-linear relationships as indicated by scatterplots. This non-linearity could potentially impact the reliability of the model, and alternative approaches to address this issue were not extensively explored.

The Box-Cox transformation was rendered unsuitable due to the emergence of infinite values, driven by the dataset's divergence from the method's assumptions, necessitating exploration of alternative approaches to address non-normality and skewness in the relevant variables.

Additionally, the presence of skewness in both residuals and predictor variables raises concerns about the normality assumption. The proposed plan to address skewness through a Box-Cox transformation acknowledges these concerns, but the effectiveness of this transformation relies on the specific characteristics of individual predictors.

The decision to exclude two categorical variables without exploring alternative transformations prompts questions about potential information loss. However, the complexities introduced by their non-linear relationships and the absence of comprehensive transformation options led to their removal.

While these limitations are acknowledged, certain issues could not be fully addressed due to the inherent nature of the data or methodological constraints. The intricacies of certain predictor-response relationships may necessitate more advanced modeling techniques beyond the scope of a linear regression approach.

Despite these limitations, the decision to proceed with the current model is justified by practical constraints. Striking a balance between model complexity and interpretability is a common challenge, and the chosen approach provides valuable insights into predicting Airbnb listing prices, even with the acknowledged limitations.

Bibliography

[1] Guttentag, D., Smith, S., Potwarka, L., & Havitz, M. (2017). Why tourists choose airbnb: A motivation-

based segmentation study. Journal of Travel Research, 57(3), 342–359. https://doi.org/10.1177/0047287517696980 [2] Wang, Y., Asaad, Y., & Filieri, R. (2019). What makes hosts trust airbnb? antecedents of hosts' trust toward Airbnb and its impact on continuance intention. Journal of Travel Research, 59(4), 686–703. https://doi.org/10.1177/0047287519855135 [3] Suess, C., Woosnam, K., Mody, M., Dogru, T., & Sirakaya Turk, E. (2020). Understanding how residents' emotional solidarity with airbnb visitors influences perceptions of their impact on a community: The moderating role of prior experience staying at an airbnb. Journal of Travel Research, 60(5), 1039–1060. https://doi.org/10.1177/0047287520921234