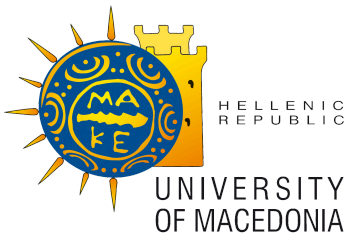


Understanding Airbnb Pricing: A Comprehensive Statistical Analysis

GEORGIOS MANOLIS eco21338

University of Macedonia

Table of Contents

Abstract

This study aims to investigate how various factors influence the pricing of Airbnb listings in the United States. The analysis utilizes a comprehensive dataset containing over 50,000 entries, encompassing diverse variables such as the number of bedrooms, number of beds, city, bed type, house type, room type, and accommodation capacity. To discern the relationships between these variables and the listing price, we employed a suite of statistical techniques including descriptive statistics, correlation analysis, regression models, training and testing, prediction, ANOVA, and Tukey's Honest Significant Difference test. The findings from this analysis provide valuable insights into the determinants of Airbnb pricing, offering implications for hosts, guests, and platform strategists in optimizing listing attributes and pricing strategies.

Introduction

In recent years, platforms like Airbnb have transformed the hospitality industry, offering property owners the opportunity to capitalize on their assets by providing accommodation to travelers. However, despite the widespread adoption of Airbnb, there remains a lack of comprehensive statistical analyses examining the factors influencing listing prices. This study aims to address this gap by conducting a thorough investigation into Airbnb listing prices across the United States. With a dataset from Kaggle that comprises over 50,000 Airbnb listings, this analysis focuses on elucidating the relationship between listing prices and various property attributes. We explore factors such as the number of bedrooms, beds, city location, bed type, house type, room type, and accommodation capacity to discern their impact on pricing dynamics. To achieve this, we employ a diverse array of statistical techniques, including descriptive statistics, correlation analysis, regression modeling, training and testing, prediction, ANOVA, and Tukey's Honest Significant Difference test.

Data Description

The dataset contains 29 variables but for the purpose of this analysis we do not need all of them. After removing the columns for `amenities`, `id`, `description`, `first_review`, `last_review`, `host_since`, `name`, `thumbnail_url`, `neighborhood` and `zipcode` we left with 19 variables. As we can observe from the table below, we have 10 quantitative variables while the remaining 9 are qualitative. It is important to note that we transformed the `host_response_rate` from a percentage representation into numerical values, using basic arithmetic operations, to help us with the statistical analysis and the data visualization.

QUANTITATIVE	QUALITATIVE
`log_price`: The logarithm of the price of the Airbnb listing	`property_type`: The type of the property being listed
`accommodates`: The number of quests that the Airbnb can accommodate	`room_type`: The type of the room being offered within a property
`bathrooms`: The number of bathrooms in the Airbnb	`bed_type`: The type of bed available in each listing
`host_response_rate`: The percentage of messages that the host responds to	`cancellation_policy`: Whether the listing has a cancellation option
`latitude`: The geographic latitude coordinates of the location of each Airbnb listing	`cleaning_fee`: Whether the listing has a cleaning fee
`longitude`: The geographic longitude coordinates of the location of each Airbnb listing	`city`: The city where each listing is located
`number_of_reviews`: The total count of reviews that an Airbnb listing has received	`host_has_profile_pic`: Whether the host of each listing has a profile picture
`review_scores_rating`: The overall rating score to an Airbnb listing	`host_identity_verified`: Whether the identity of the host of each listing has been verified
`bedrooms`: The number of bedrooms in the Airbnb	`instant_bookable`: Whether a listing is instant bookable
`beds`: The number of beds in the Airbnb	

Descriptive Analysis

Summary of the Dataset

The summary statistics for the quantitative variables are presented below, displaying the minimum and maximum values, the median and mean for each variable, as well as the first and third quartiles.

```
   log_price      accommodates      bathrooms    host_response_rate    latitude
 Min.   :0.000   Min.   : 1.000   Min.   :0.00   Min.   :0.0000     Min.   :33.34
 1st Qu.:4.304   1st Qu.: 2.000   1st Qu.:1.00   1st Qu.:1.0000     1st Qu.:34.11
 Median :4.700   Median : 2.000   Median :1.00   Median :1.0000     Median :40.65
 Mean   :4.751   Mean   : 3.325   Mean   :1.24   Mean   :0.9556     Mean   :38.36
 3rd Qu.:5.170   3rd Qu.: 4.000   3rd Qu.:1.00   3rd Qu.:1.0000     3rd Qu.:40.75
 Max.   :7.600   Max.   :16.000   Max.   :8.00   Max.   :1.0000     Max.   :42.39
   longitude       number_of_reviews review_scores_rating    bedrooms           beds
 Min.   :-122.51   Min.   :  1.00    Min.   : 20.00       Min.   : 0.000   Min.   : 0.000
 1st Qu.:-118.35   1st Qu.:  5.00    1st Qu.: 92.00       1st Qu.: 1.000   1st Qu.: 1.000
 Median : -77.03   Median : 14.00    Median : 96.00       Median : 1.000   Median : 1.000
 Mean   : -93.21   Mean   : 30.78    Mean   : 94.23       Mean   : 1.282   Mean   : 1.795
 3rd Qu.: -73.95   3rd Qu.: 39.00    3rd Qu.:100.00       3rd Qu.: 1.000   3rd Qu.: 2.000
 Max.   : -71.00   Max.   :605.00    Max.   :100.00       Max.   :10.000   Max.   :18.000
```

The summary statistics for the qualitative variables indicate the number of elements present at each level of the factors.

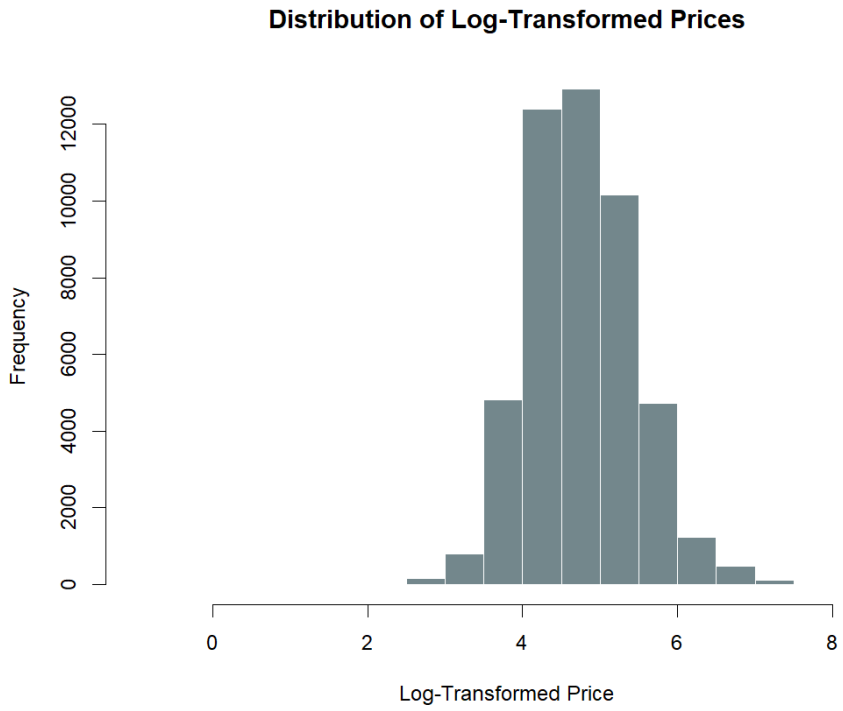
```
    property_type            room_type             bed_type         cancellation_policy cleaning_fee       city
Apartment  :30018   Entire home/apt:27445   Airbed       :  254   flexible       : 9163    False: 8502   Boston : 2469
House      :11777   Private room   :19090   Couch        :  131   moderate       :13869    True :39285   Chicago: 3069
Condominium: 1764   Shared room    : 1252   Futon        :  495   strict         :24667                  DC     : 3515
Townhouse  : 1199                           Pull-out Sofa:  410   super_strict_30:   78                  LA     :15340
Loft       :  869                           Real Bed     :46497   super_strict_60:   10                  NYC    :19402
Guesthouse :  397                                                                                        SF     : 3992
(Other)    : 1763
host_has_profile_pic host_identity_verified instant_bookable
TRUE :47714          TRUE :34934            TRUE :14014
FALSE:   73          FALSE:12853            FALSE:33773
```

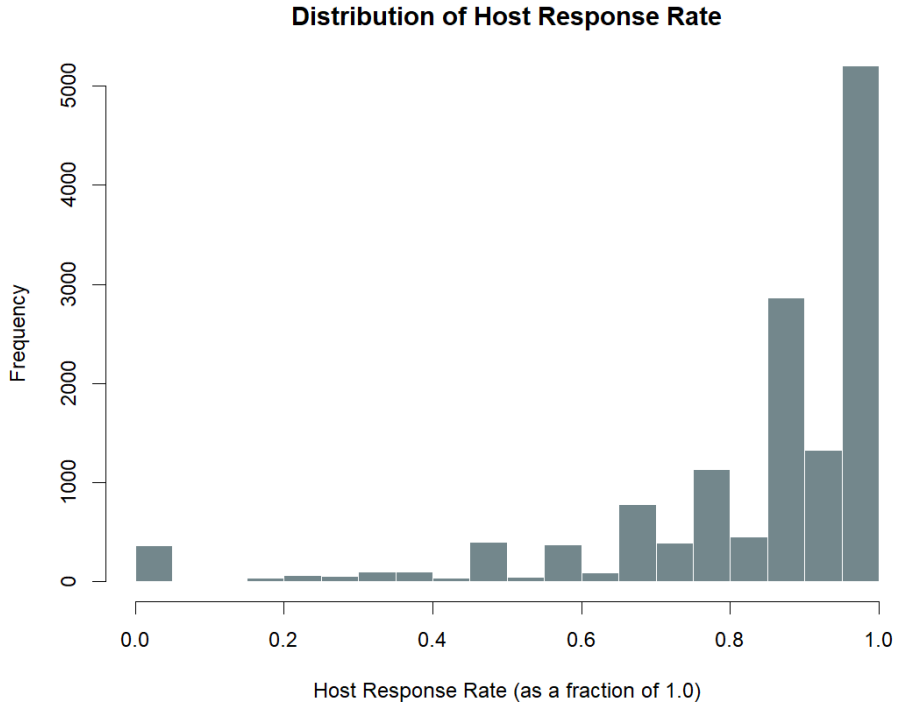
Visualization of the Dataset

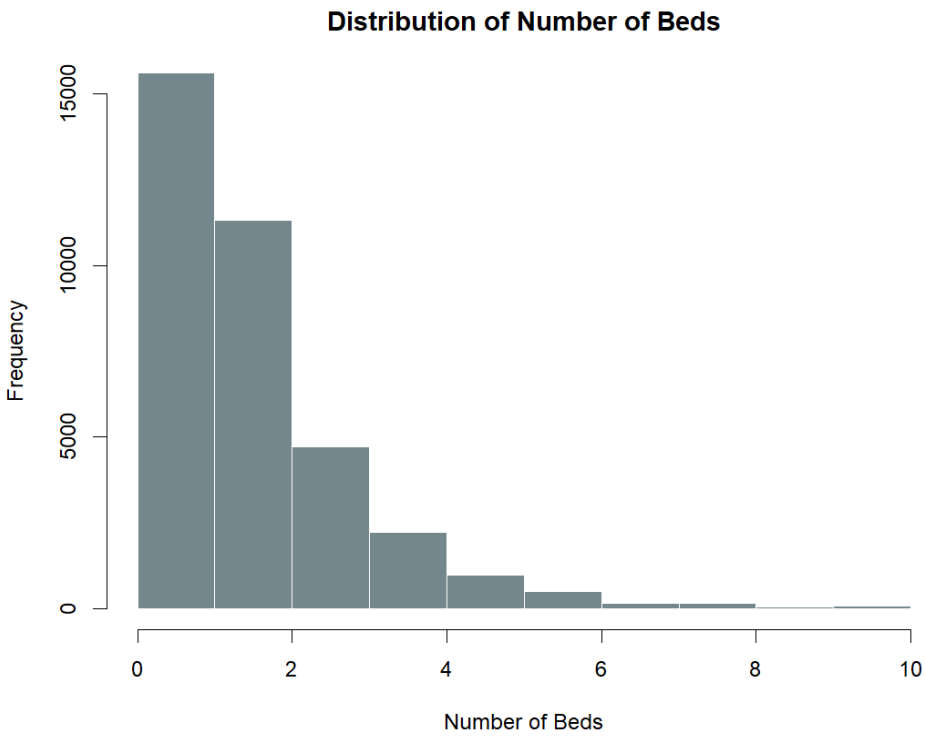
Histograms of the Quantitative Variables

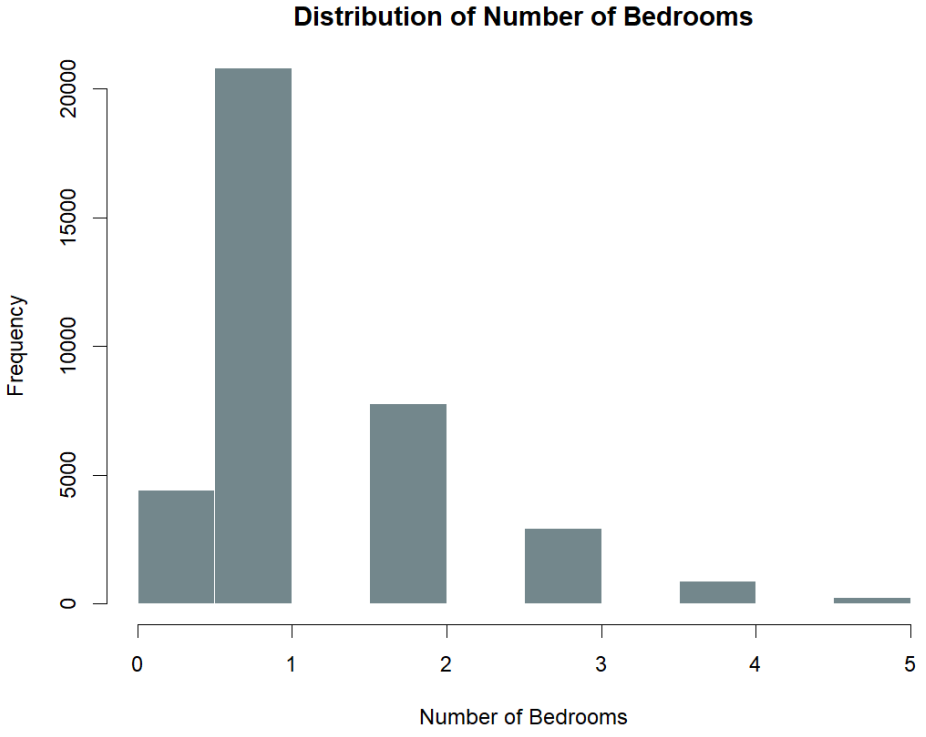
Histograms are essential graphical tools for providing visual representations of the distribution of the quantitative variables by condensing data frequency or density into intervals. By examining the shape and characteristics of the histogram, we can identify patterns such as symmetry, skewness, modality, and outliers in the data distribution.

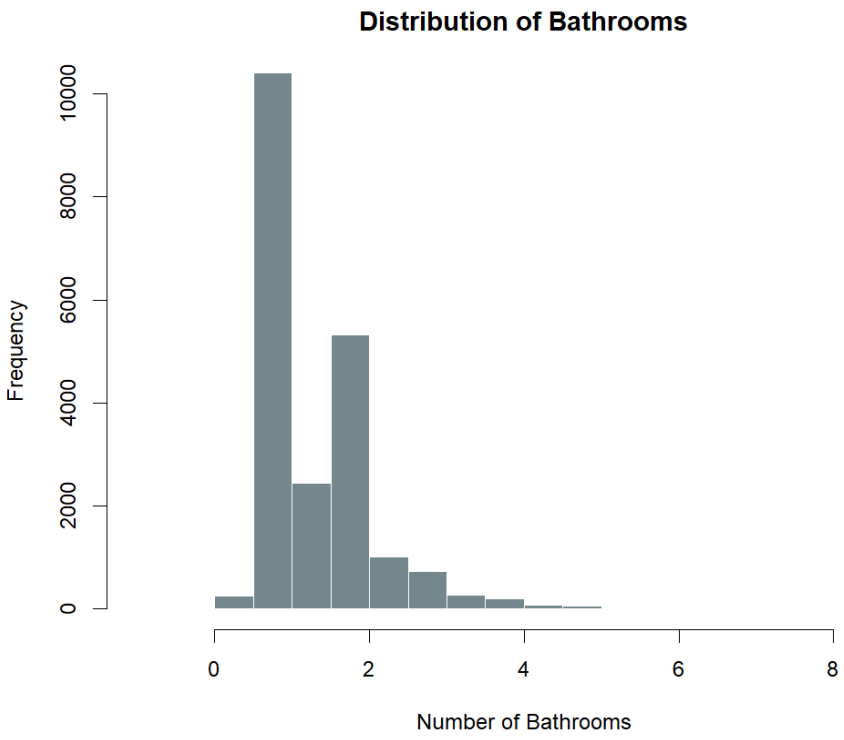
As we observe only the Log-Transformed Prices have a normal distribution while the other variables skew either to the left or the right of the graph.

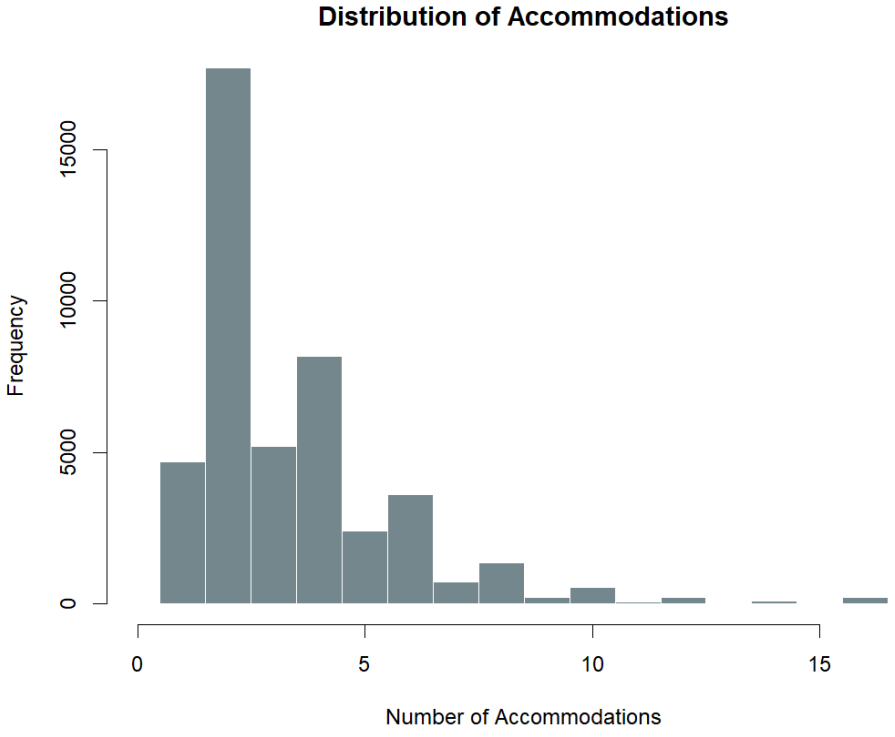


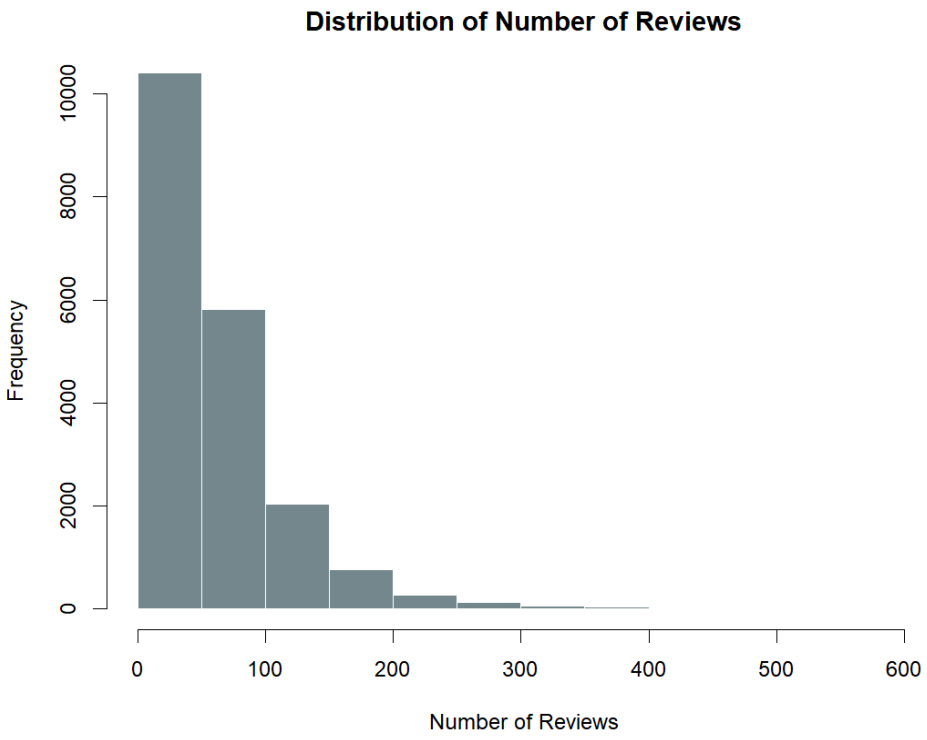


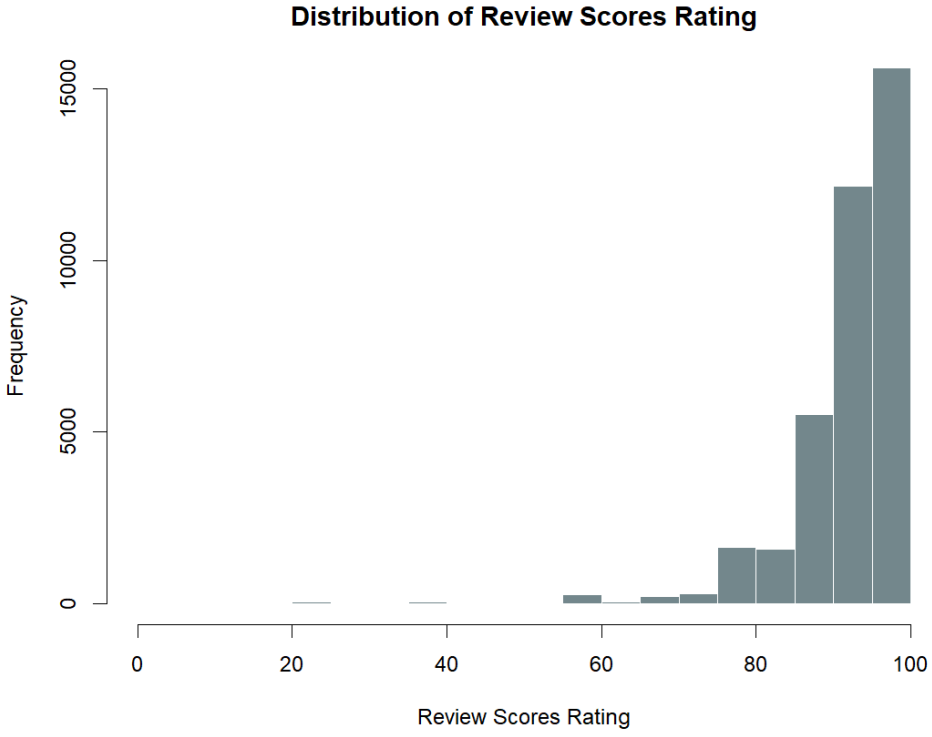

Distribution of Number of Beds
Frequency
0
5000
10000
15000
0
2
4
6
8
10
Number of Beds


Distribution of Number of Bedrooms
Frequency
0
5000
10000
15000
20000
0
1
2
3
4
5
Number of Bedrooms


Distribution of Bathrooms
Frequency
0
2000
4000
6000
8000
10000
0
2
4
6
8
Number of Bathrooms


Distribution of Accommodations
Frequency
0
5000
10000
15000
0
5
10
15
Number of Accommodations

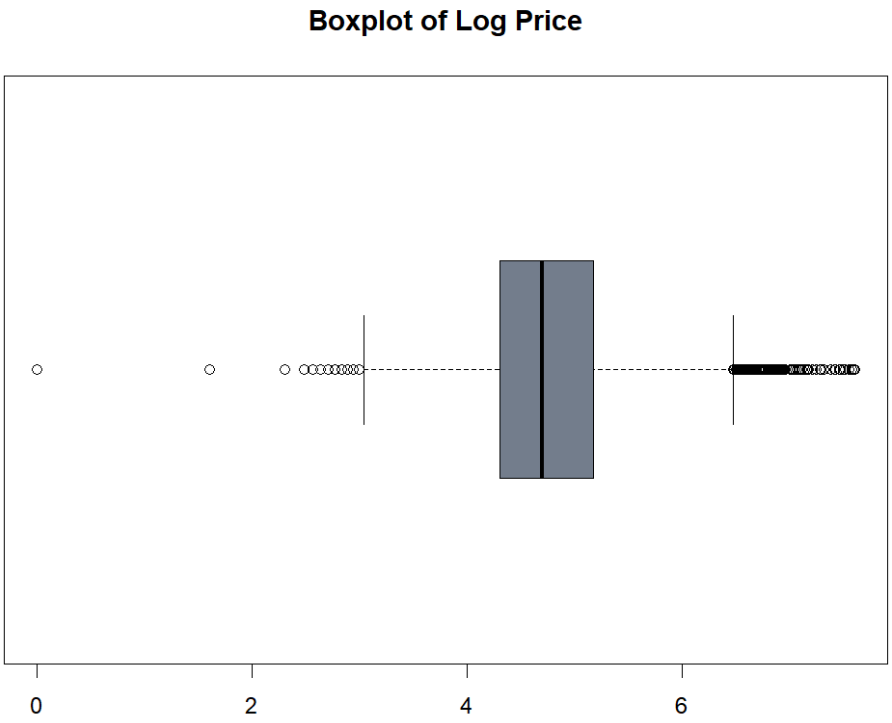

Distribution of Number of Reviews
Frequency
0
2000
4000
6000
8000
10000
0
100
200
300
400
500
600
Number of Reviews


Distribution of Review Scores Rating
Frequency
0
5000
10000
15000
0
20
40
60
80
100
Review Scores Rating

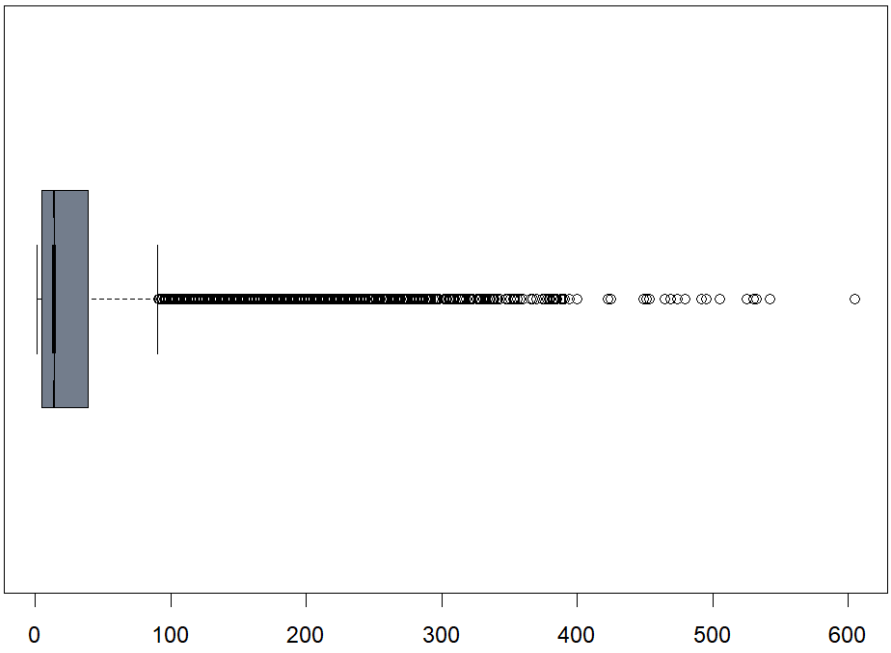
Boxplots of the Quantitative Variables

Boxplots, also referred to as box-and-whisker plots, are valuable graphical tools to visualize the distribution and variability of quantitative variables. They provide a concise visual representation of key statistical measures, including central tendency, spread, variability and outlier presence. Boxplots offer a clear visual depiction of median, quartiles, and data range, making them valuable for exploratory data analysis. The boxplots provided indicate the presence of numerous outliers across all quantitative variables. Notably, the variables `bathrooms`, `bedrooms` and `host_response_rate` consist exclusively of outlier values.

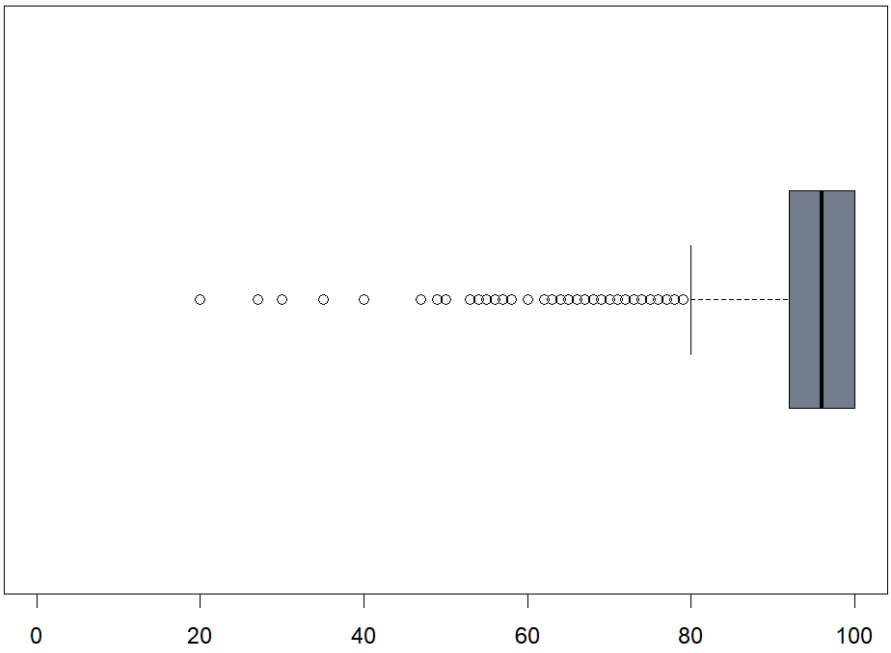
```
   log_price
Min.   :0.000
1st Qu.:4.304
Median :4.700
Mean   :4.751
3rd Qu.:5.170
Max.   :7.600
```



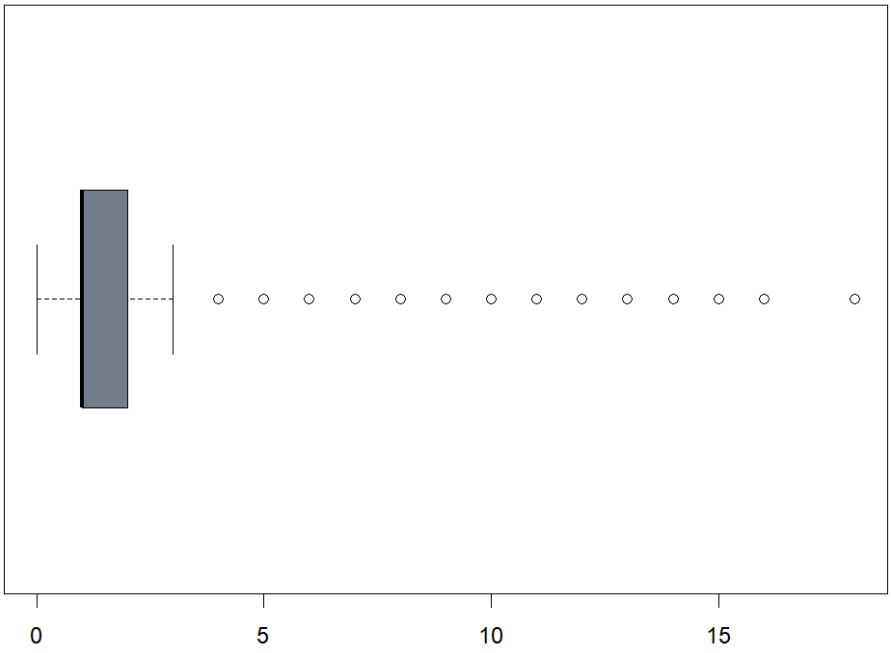
Boxplot of Number of Reviews



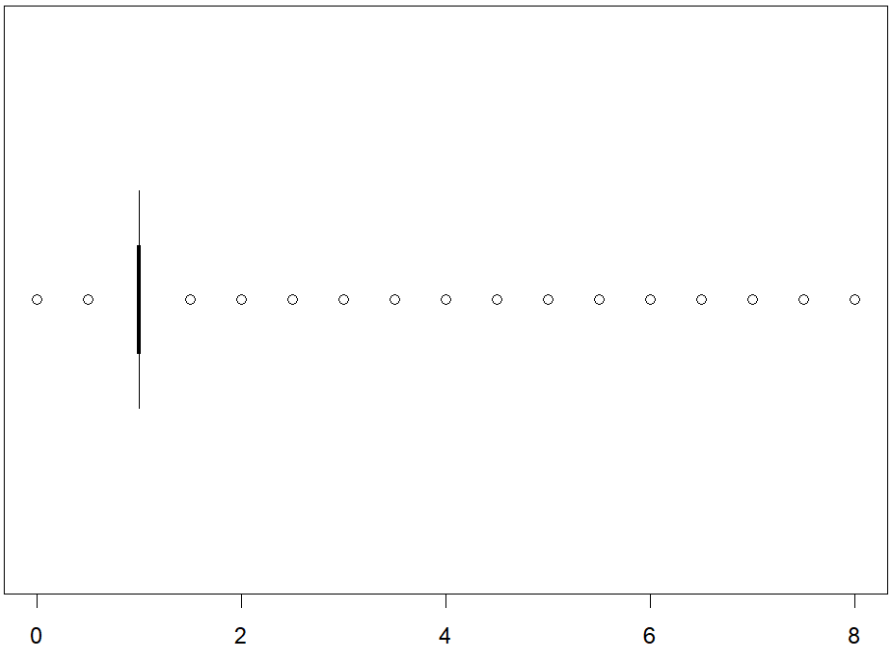
Boxplot of Review Scores Rating



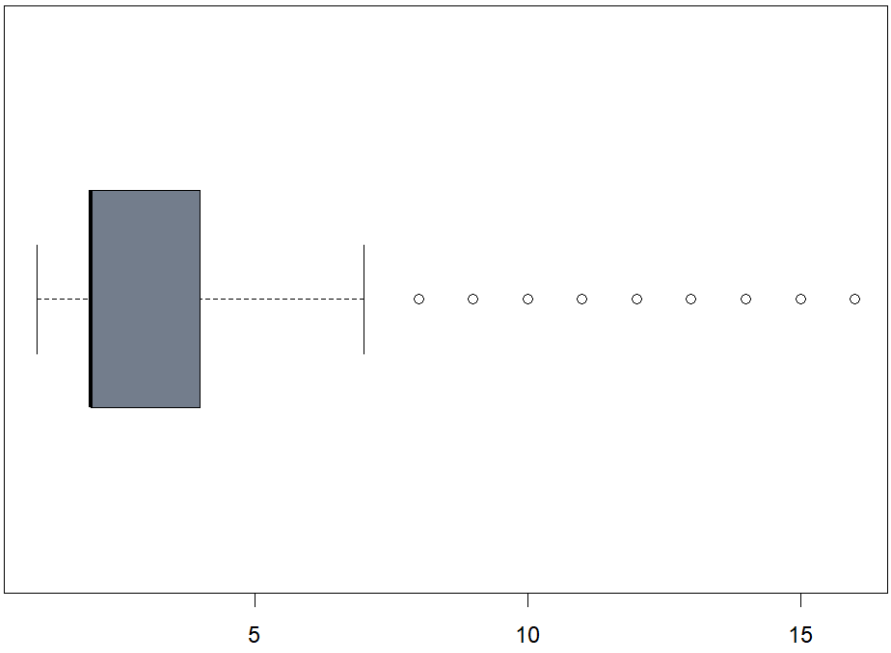
Boxplot of Beds



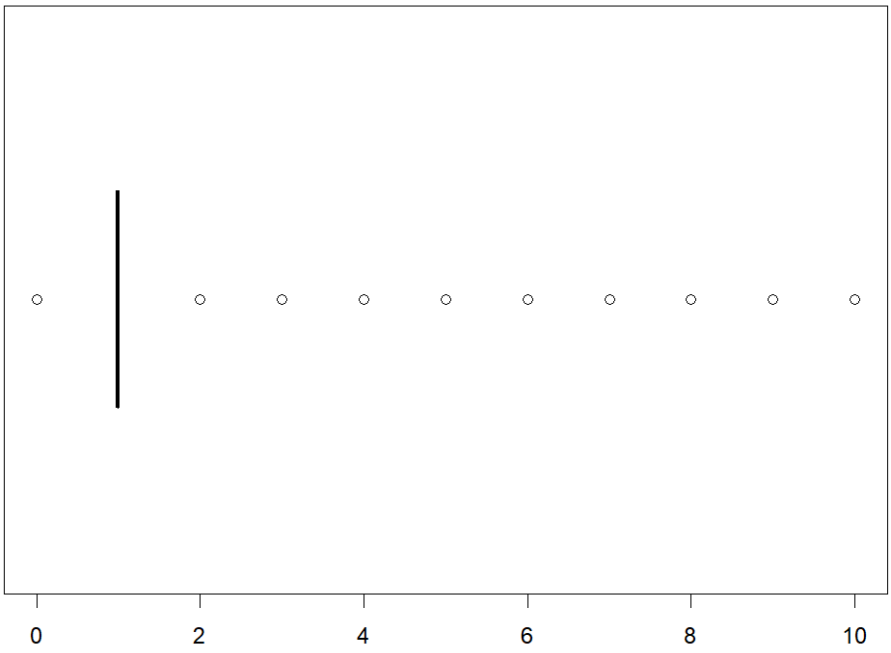
Boxplot of Bathrooms

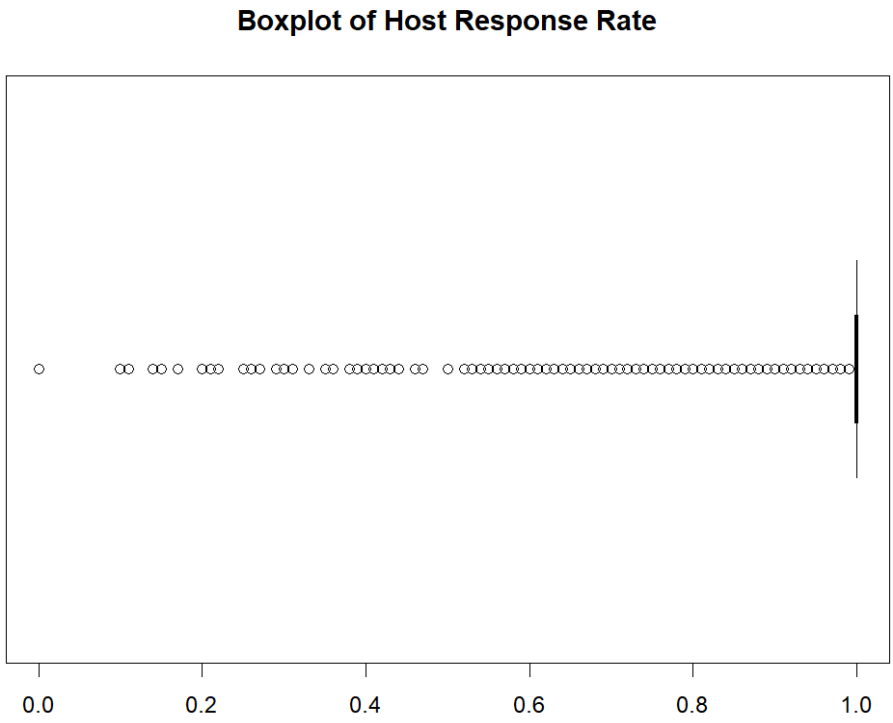


Boxplot of Accommodates



Boxplot of Bedrooms





Statistical Analysis of the Quantitative Variables

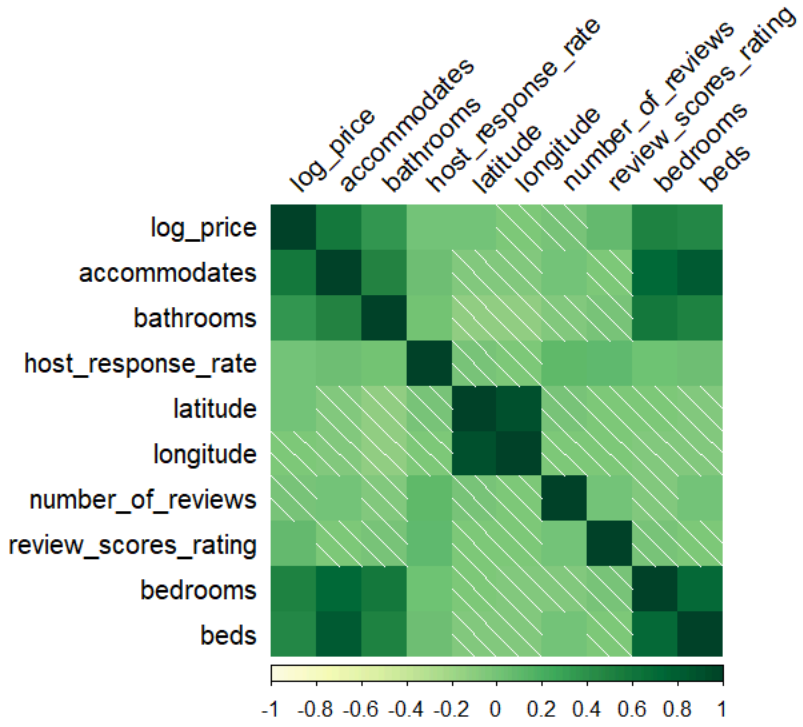
Correlation Analysis

Correlation analysis is a fundamental aspect of statistical data analysis, used to examine the strength and direction of relationships between quantitative variables. By measuring how closely two variables are related, this analysis reveals patterns, trends, and potential causal links within the dataset. This section of the report explores the correlation analysis, employing statistical tools, using visual aids such as correlation plots, the correlation analysis aims to uncover significant associations between variables, which can help with data-driven decisions.

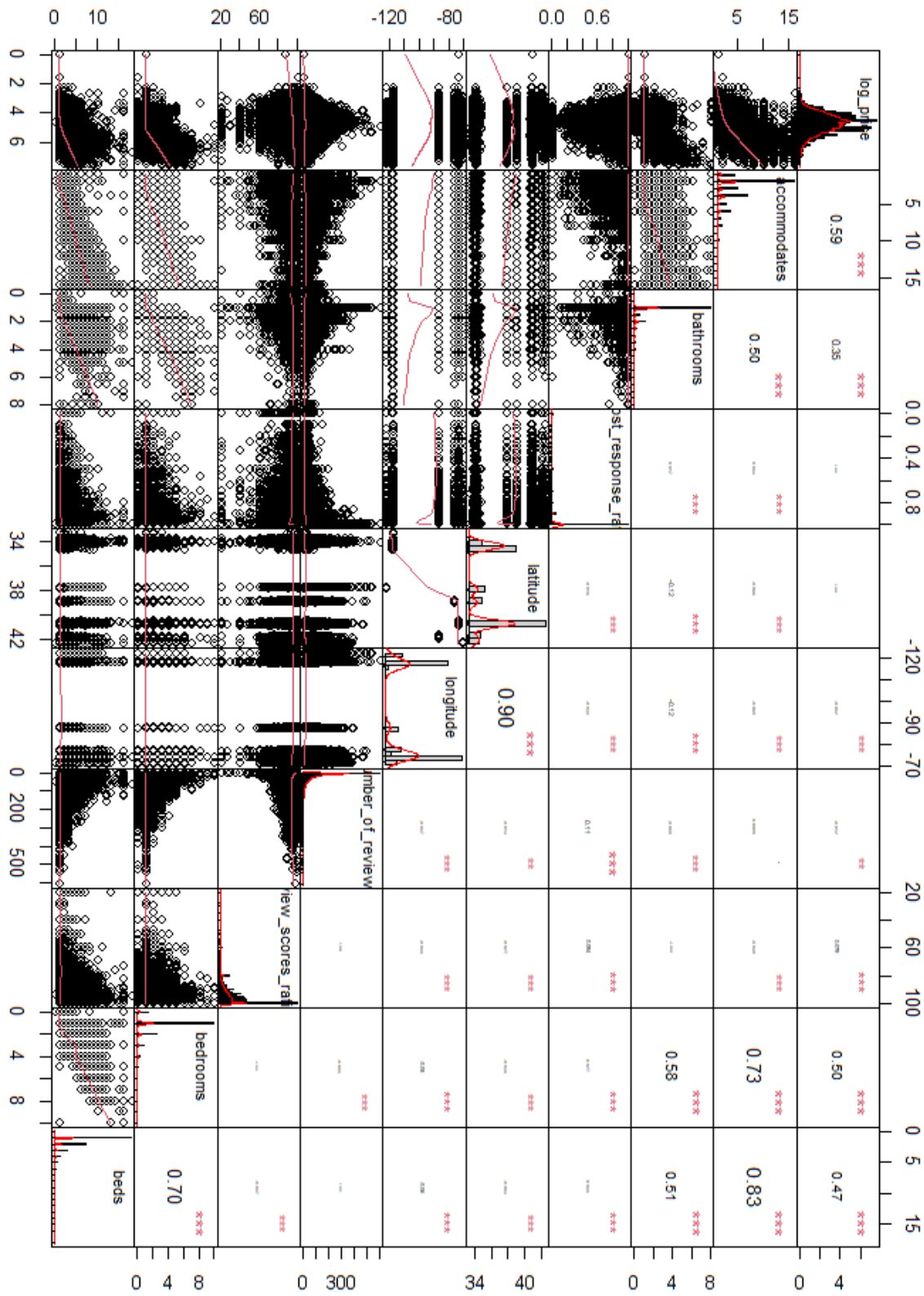
	log_price	accommodates	bathrooms	host_response_rate	latitude	longitude	number_of_reviews	review_scores_rating	bedrooms	beds
log_price	1.000000000	0.593480770	0.3522794319	0.002655097	0.00349665	-0.04224858	-0.0123776612	0.0792684626	0.503405536	0.465855590
accommodates	0.593480770	1.000000000	0.4982958789	0.042917023	-0.06560903	-0.06827546	0.0084591974	-0.0377213506	0.725349380	0.825103391
bathrooms	0.352279432	0.498295879	1.0000000000	0.017156625	-0.12208499	-0.12068317	-0.0545581066	-0.0005195305	0.575605942	0.511197577
host_response_rate	0.002655097	0.042917023	0.0171566252	1.000000000	-0.01636337	-0.03421797	0.1052561486	0.0944818283	0.027313269	0.044102294
latitude	0.003496650	-0.065609034	-0.1220849947	-0.016363374	1.00000000	0.89516190	-0.0127400215	-0.0270260327	-0.034149378	-0.063371660
longitude	-0.042248585	-0.068275458	-0.1206831706	-0.034217968	0.89516190	1.00000000	-0.0471828223	-0.0438045603	-0.050437601	-0.060428088
number_of_reviews	-0.012377661	0.008459197	-0.0545581066	0.105256149	-0.01274002	-0.04718282	1.0000000000	0.0008991657	-0.053576423	0.002530212
review_scores_rating	0.079268463	-0.037721351	-0.0005195305	0.094481828	-0.02702603	-0.04380456	0.0008991657	1.0000000000	-0.001119229	-0.046684725
bedrooms	0.503405536	0.725349380	0.5756059416	0.027313269	-0.03414938	-0.05043760	-0.0535764227	-0.0011192294	1.000000000	0.704525207
beds	0.465855590	0.825103391	0.5111975770	0.044102294	-0.06337166	-0.06042809	0.0025302118	-0.0466847248	0.704525207	1.000000000

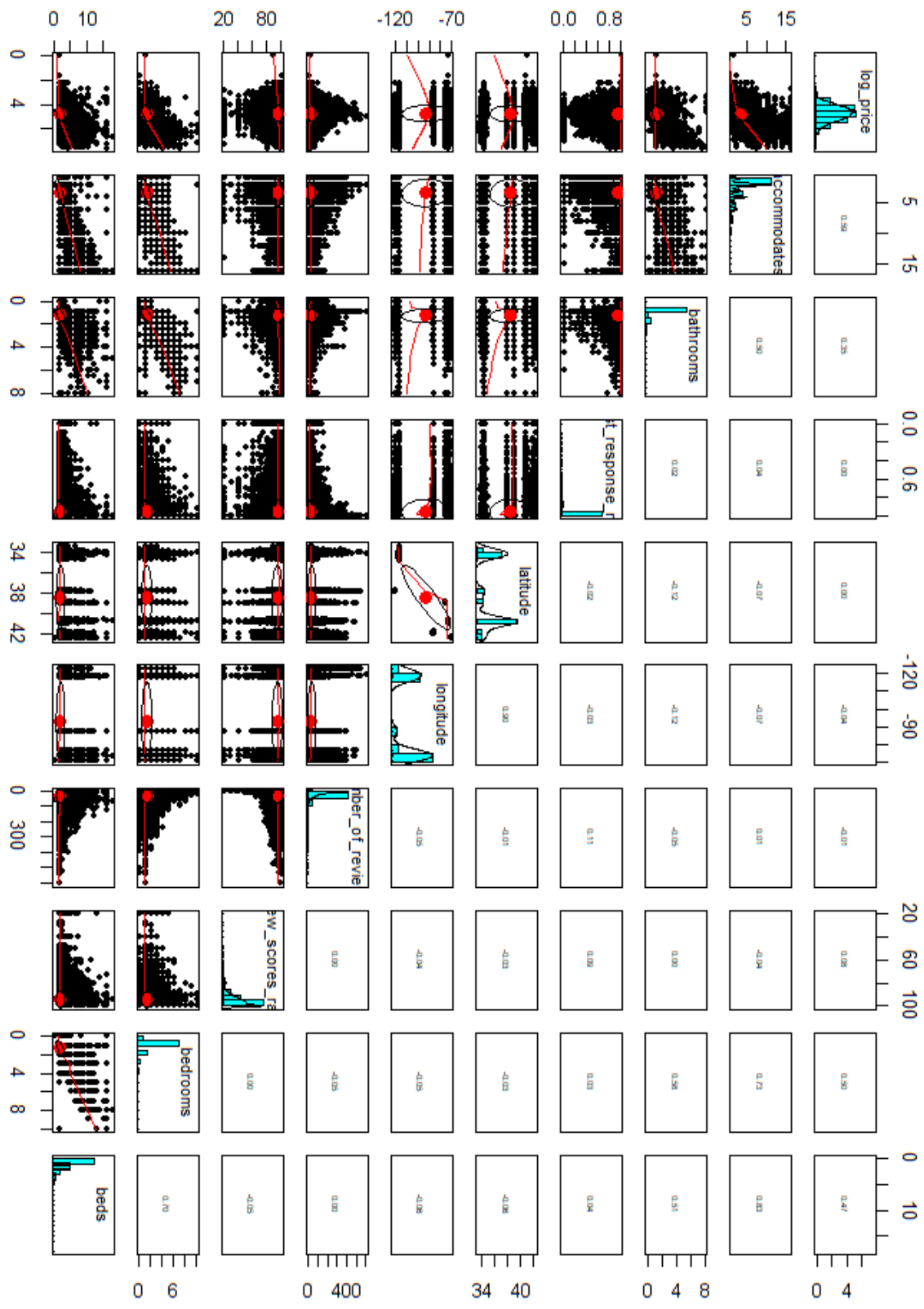
The table presented above illustrates the correlation between all the variables with each other. The correlation coefficient is expressed with “r” and ranges between -1 and +1. However, due to the difficulty in observing each correlation individually, we will visualize the correlations in the heatmap below.

Correlation Heatmap (Numerical Columns Only)



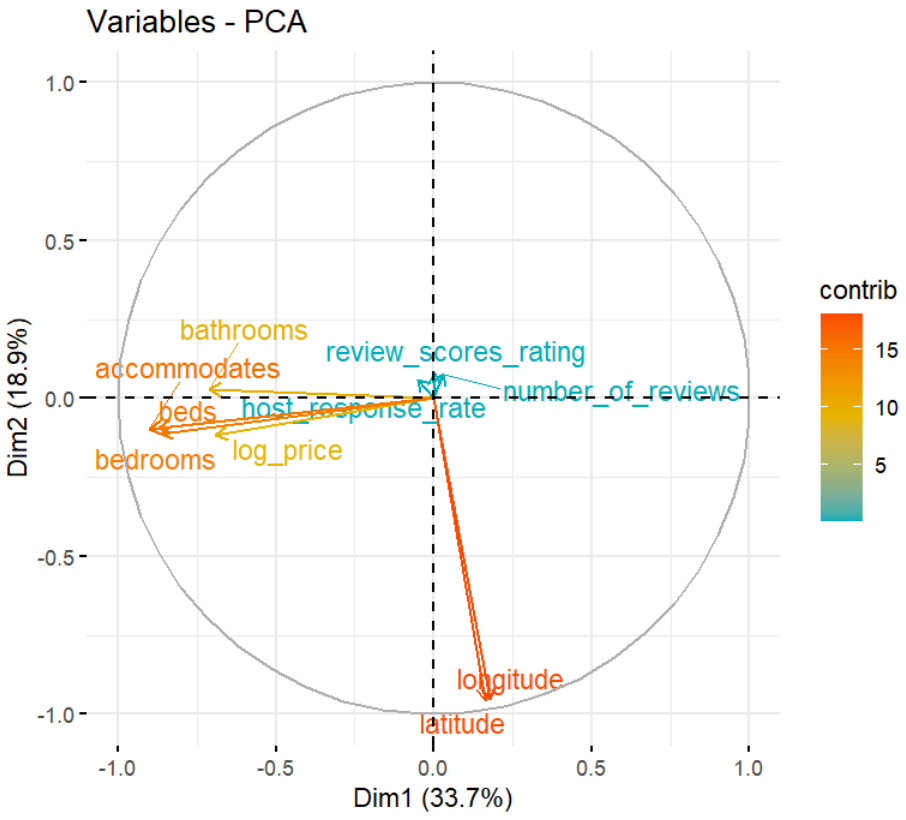
Now, it is evident that the variable `log_price` has a strong correlation with `accommodates`, `bathrooms`, `bedrooms` and `beds` by the color of the square. It is important to note that `bedrooms` and `beds` also correlate with `accommodates`, `bathrooms` and each other. Be aware that the shade squares are when the correlation variable is negative. The strong correlation between `latitude` and `longitude` is expected, as they are location coordinates. The next two correlation charts verify everything that we said through the histograms and scatter plots.


log_price
accommodates
bathrooms
bedrooms
beds
latitude
longitude
number_of_reviews
review_scores_rating
host_response_rate
0.59
0.35
0.50
0.90
0.58
0.73
0.50
0.51
0.83
0.47
0.70


log_price
accommodates
bathrooms
t_response_r
latitude
longitude
nber_of_revie
w_scores_re
bedrooms
beds

Principal Components Analysis

Principal Component Analysis (PCA) is a robust statistical method used in data analysis and machine learning to reduce the complexity of high-dimensional datasets while retaining their key patterns and relationships. It achieves this by transforming the original variables into a new set of uncorrelated variables known as principal components. These components capture the maximum variance in the data, allowing for a more manageable and insightful analysis. PCA simplifies data visualization and interpretation, making it a valuable tool in various fields such as finance, genetics, and image processing. It is commonly used for tasks such as noise reduction, feature extraction, and data compression.

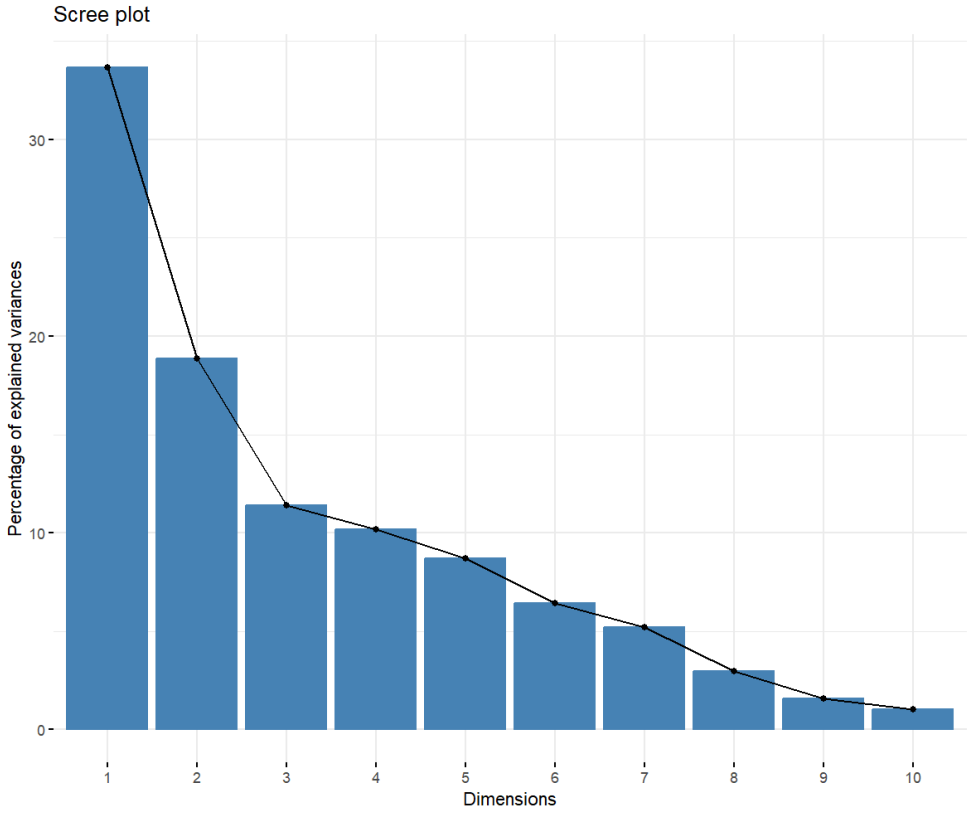


Overall, the graph illustrates the correlations between variables. Variables that are irrelevant to each other are positioned far apart, whereas those that are strongly and highly correlated are aligned in the same direction as the dependent variable, `log_price`

This table outlines the significance of each principal component. Firstly, we observe that there are ten components. The first component accounts for 33.65% of the total variance. The combined variance of the first two components is approximately 52.51%. When all ten principal components are considered, they collectively explain 100% of the variance.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     1.8344 1.3732 1.0680 1.0085 0.9338 0.80197 0.72019 0.54628 0.39707 0.31875
Proportion of Variance 0.3365 0.1886 0.1141 0.1017 0.0872 0.06432 0.05187 0.02984 0.01577 0.01016
Cumulative Proportion  0.3365 0.5251 0.6391 0.7409 0.8280 0.89236 0.94423 0.97407 0.98984 1.00000
```

Additionally, the Principal Components Analysis plot indicates that the largest percentage of variance is attributed to the first component. This is evident from the table below, which displays the standard deviations of each component.




```
Call:
princomp(x = df, cor = T, scores = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7    Comp.8    Comp.9   Comp.10
1.8344103 1.3732188 1.0680272 1.0084859 0.9337937 0.8019724 0.7201856 0.5462794 0.3970661 0.3187518

 10  variables and  47787 observations.
```

Selecting variables with the stepwise regression methods

The stepwise method, including forward selection, backward elimination, and stepwise selection, is a systematic approach used in regression analysis to identify the most significant variables from a larger set. These techniques help in creating more parsimonious models by including only the most relevant predictors, enhancing interpretability and predictive performance. Forward Selection starts with no variables in the model, adding them one by one based on specific criteria such as the lowest p-value or highest F-statistic, until no further significant improvement is observed.

```
Call:
lm(formula = y ~ accommodates + bedrooms + review_scores_rating +
    beds + latitude + longitude + bathrooms + host_response_rate +
    number_of_reviews, data = df[, -1])

Coefficients:
         (Intercept)          accommodates              bedrooms  review_scores_rating                 beds
           1.1491545             0.1738290             0.1147574             0.0089623           -0.0666105
            latitude             longitude             bathrooms    host_response_rate     number_of_reviews
           0.0446636            -0.0056298             0.0596631            -0.1630380           -0.0001375
```

Backward Elimination begins with all potential variables included, removing the least significant variables step-by-step, until only those that contribute meaningfully to the model remain.

```
Call:
lm(formula = y ~ accommodates + bathrooms + host_response_rate +
    latitude + longitude + number_of_reviews + review_scores_rating +
    bedrooms + beds, data = df[, -1])

Coefficients:
         (Intercept)          accommodates             bathrooms    host_response_rate              latitude
           1.1491545             0.1738290             0.0596631            -0.1630380             0.0446636
           longitude     number_of_reviews  review_scores_rating              bedrooms                  beds
          -0.0056298            -0.0001375             0.0089623             0.1147574           -0.0666105
```

Stepwise Selection combines elements of both forward and backward methods, adding and removing variables iteratively to find the optimal model.

```
Call:
lm(formula = y ~ accommodates + bedrooms + review_scores_rating +
    beds + latitude + longitude + bathrooms + host_response_rate +
    number_of_reviews, data = df[, -1])

Coefficients:
         (Intercept)          accommodates              bedrooms  review_scores_rating                 beds
           1.1491545             0.1738290             0.1147574             0.0089623           -0.0666105
            latitude             longitude             bathrooms    host_response_rate     number_of_reviews
           0.0446636            -0.0056298             0.0596631            -0.1630380           -0.0001375
```

The stepwise method concludes that all variable coefficients are statistically significant. As evidenced by the results below, all p-values are below the chosen alpha level of significance (0.05 or 5%). Therefore, no variables should be excluded from the final model.

```
Call:
lm(formula = y ~ accommodates + bathrooms + host_response_rate +
    latitude + longitude + number_of_reviews + review_scores_rating +
    bedrooms + beds, data = df[, -1])

Residuals:
    Min      1Q  Median      3Q     Max
-4.2868 -0.3385  0.0021  0.3449  3.2046

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.149e+00  9.357e-02  12.281   <2e-16 ***
accommodates          1.738e-01  2.054e-03  84.621   <2e-16 ***
bathrooms             5.966e-02  5.204e-03  11.465   <2e-16 ***
host_response_rate   -1.630e-01  1.839e-02  -8.863   <2e-16 ***
latitude              4.466e-02  1.732e-03  25.781   <2e-16 ***
longitude            -5.630e-03  2.515e-04 -22.388   <2e-16 ***
number_of_reviews    -1.375e-04  5.649e-05  -2.434   0.0149 *
review_scores_rating  8.962e-03  3.374e-04  26.566   <2e-16 ***
bedrooms              1.148e-01  4.484e-03  25.590   <2e-16 ***
beds                 -6.661e-02  3.390e-03 -19.649   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5283 on 47777 degrees of freedom
Multiple R-squared:  0.3887,    Adjusted R-squared:  0.3886
F-statistic:  3376 on 9 and 47777 DF,  p-value: < 2.2e-16
```

Training and Testing model

In machine learning and statistical modeling, the process of training and testing a model is essential to ensure accuracy and generalizability. During the training phase, a portion of the data, known as the training set, is used to develop the model by identifying patterns and relationships. This step is essential for optimizing the model's parameters and enhancing its predictive performance. After training, the model is evaluated using a different portion of the data called the testing set. This testing phase assesses the model's performance on new, unseen data, providing an unbiased evaluation of its effectiveness and robustness. By dividing the data into training and testing sets, we can accurately determine the model's ability to generalize to real-world data, ensuring its reliability and accuracy in practical applications. The training set will consist of 70% of the quantitative variables, while the remaining 30% will be allocated to the testing set.

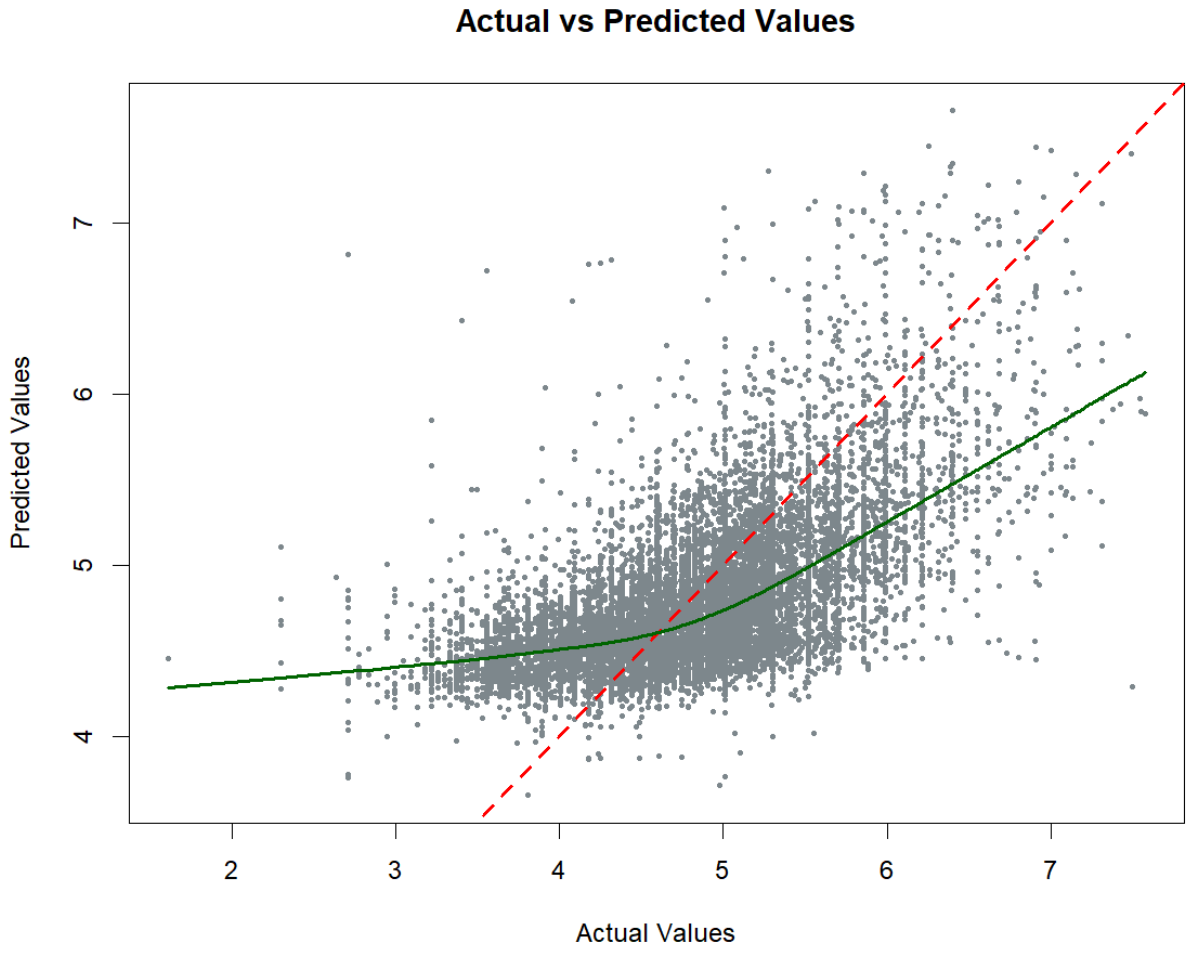
Prediction Metrics

Prediction metrics are essential tools in machine learning and statistical modeling for evaluating the performance of models. They help in comparing models, monitoring performance over time, and identifying areas for improvement. In regression tasks, key metrics include Mean Absolute Error (MAE), which measures the average magnitude of errors, Mean Squared Error (MSE), which gives higher weight to larger errors and highlights outliers, Root Mean Squared Error (RMSE), which is the square root of MSE and makes error measurements in the same units as the target variable, R-squared (R^2), which indicates the proportion of variance explained by the model, and Mean Absolute Percentage Error (MAPE), which expresses prediction accuracy as a percentage and is useful for understanding the relative error magnitude. These metrics provide objective measures of model performance, guiding the selection and refinement of predictive models for better accuracy and reliability.

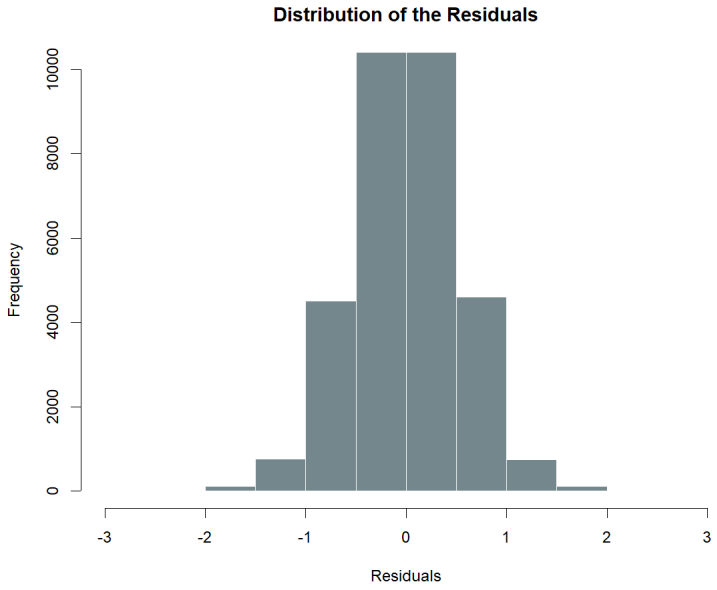
Summary

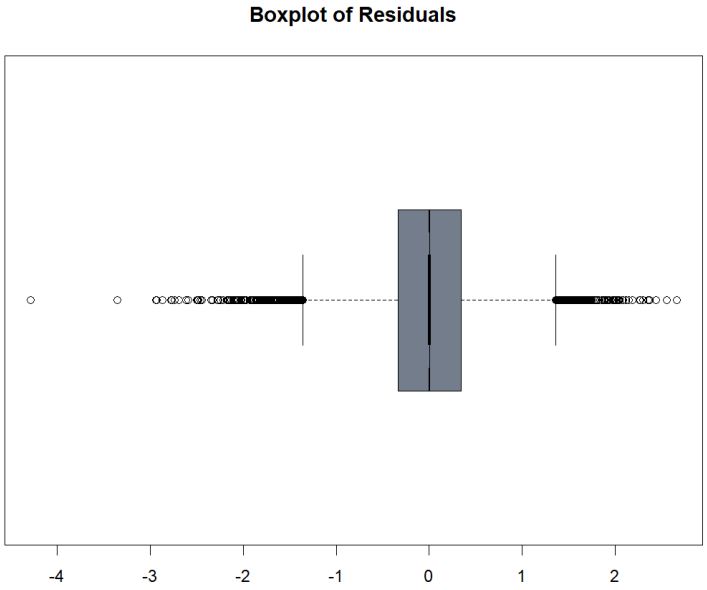
Given that the `log_price` values range from 0 to 7, the prediction metrics offer insights into the model's performance. The Mean Absolute Percentage Error (MAPE) of 9.10% indicates that, on average, predictions deviate by approximately 9.10% from the actual values, which is considered excellent. Similarly, the Mean Absolute Error (MAE) of 0.42 and Mean Squared Error (MSE) of 0.29, along with the Root Mean Squared Error (RMSE) of 0.54, suggest relatively small errors relative to the target value range. The high correlation coefficient of 0.62 indicates a moderate to strong positive linear relationship between predicted and actual values. Additionally, the Symmetric Mean Absolute Percentage Error (sMAPE) of 8.88% further underscores the model's accuracy. The Min-Max Accuracy and low Percentage Bias confirm the model's ability to make predictions close to the actual values with minimal systematic error. Overall, these metrics suggest that the model performs well on the test set, providing accurate predictions within the range of the target variable.

MAPE	MAE	MSE	RMSE	Manual MAPE
0.09101797 (9.10%)	0.4177904	0.2870882	0.5358061	0.08796185 (8.80%)
Min - Max	Correlation Co	sMAPE	Percentage Bias	
0.9172145 (91.72%)	0.6193422	0.08876155 (8.88%)	0.001334413 (0.13%)	



As evidenced by the histogram and boxplot, the residuals of the training set follow a normal distribution.





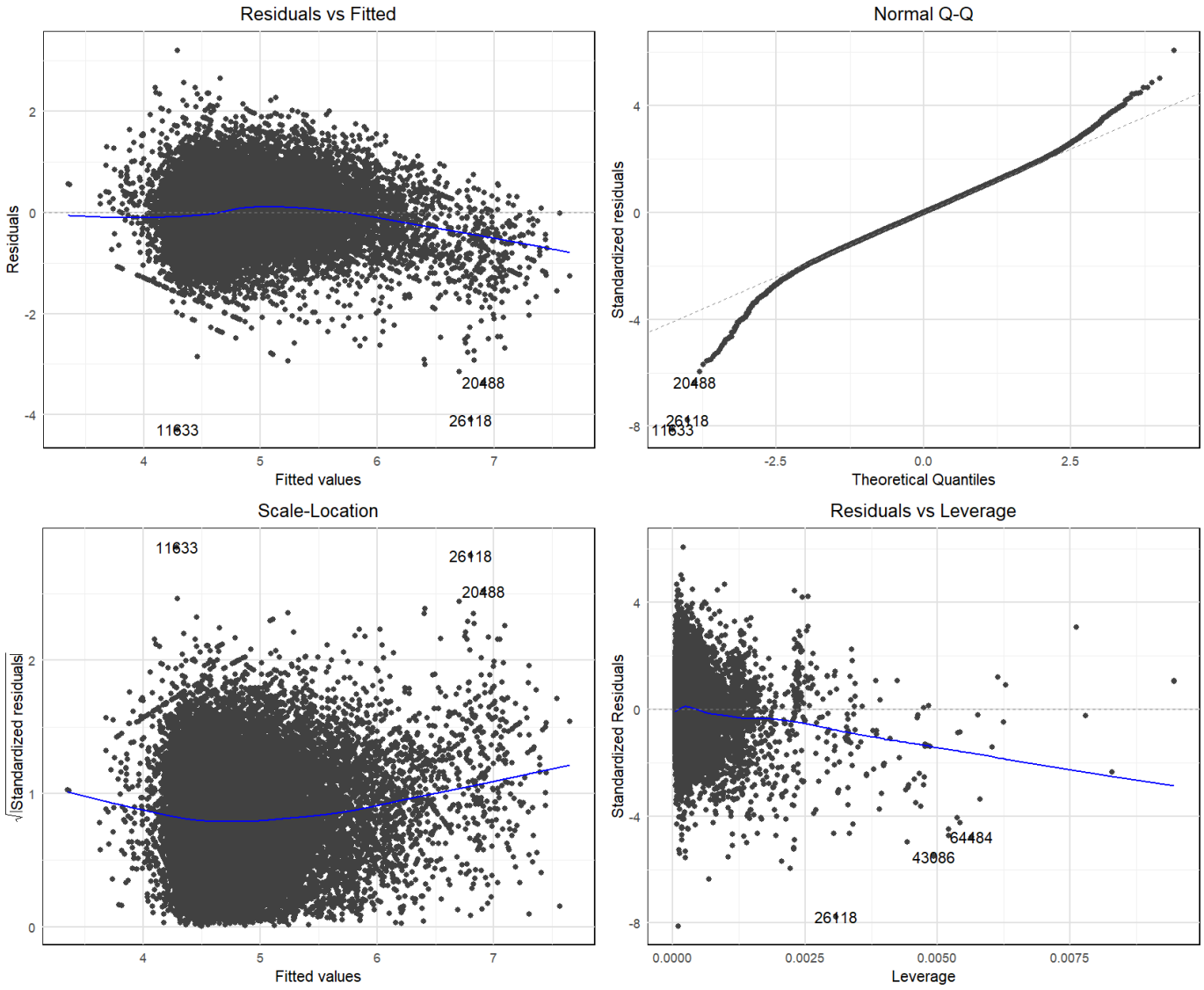
Multiple Linear Regression Model

Multiple linear regression is a statistical method designed to model the relationship between a dependent variable and multiple independent variables. By fitting a linear equation to observed data, this approach helps to understand how changes in the independent variables affect the dependent variable. Unlike simple linear regression, which involves just one predictor, multiple linear regression can handle more complex datasets and provide insights into the individual impact of each predictor. Based on the results from the stepwise method, all variables have demonstrated statistical significance, warranting their inclusion in the model. Therefore, we will incorporate all these variables. The dependent variable is `log_price`, while the independent variables consist of the remaining quantitative variables.

```
Call:
lm(formula = df$log_price ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2868 -0.3385  0.0021  0.3449  3.2046

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.149e+00  9.357e-02  12.281   <2e-16 ***
accommodates         1.738e-01  2.054e-03  84.621   <2e-16 ***
bathrooms            5.966e-02  5.204e-03  11.465   <2e-16 ***
host_response_rate  -1.630e-01  1.839e-02  -8.863   <2e-16 ***
latitude             4.466e-02  1.732e-03  25.781   <2e-16 ***
longitude           -5.630e-03  2.515e-04 -22.388   <2e-16 ***
number_of_reviews   -1.375e-04  5.649e-05  -2.434   0.0149 *
review_scores_rating 8.962e-03  3.374e-04  26.566   <2e-16 ***
bedrooms             1.148e-01  4.484e-03  25.590   <2e-16 ***
beds                -6.661e-02  3.390e-03 -19.649   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5283 on 47777 degrees of freedom
Multiple R-squared:  0.3887,    Adjusted R-squared:  0.3886
F-statistic:  3376 on 9 and 47777 DF,  p-value: < 2.2e-16
```

The residuals range from -4.2868 to 3.2046, with a median close to zero, indicating a relatively symmetrical distribution around the fitted values. The coefficients for all variables are statistically significant, with p-values less than 0.05, suggesting that each variable is a meaningful predictor of `log_price`.

The model's residual standard error is 0.5283, and it achieves an R-squared value of 0.3887, indicating that about 38.87% of the variance in `log_price` is explained by the model. The overall F-statistic is highly significant (p-value < 2.2e-16), demonstrating that the model is robust and the included variables collectively contribute to predicting `log_price`.



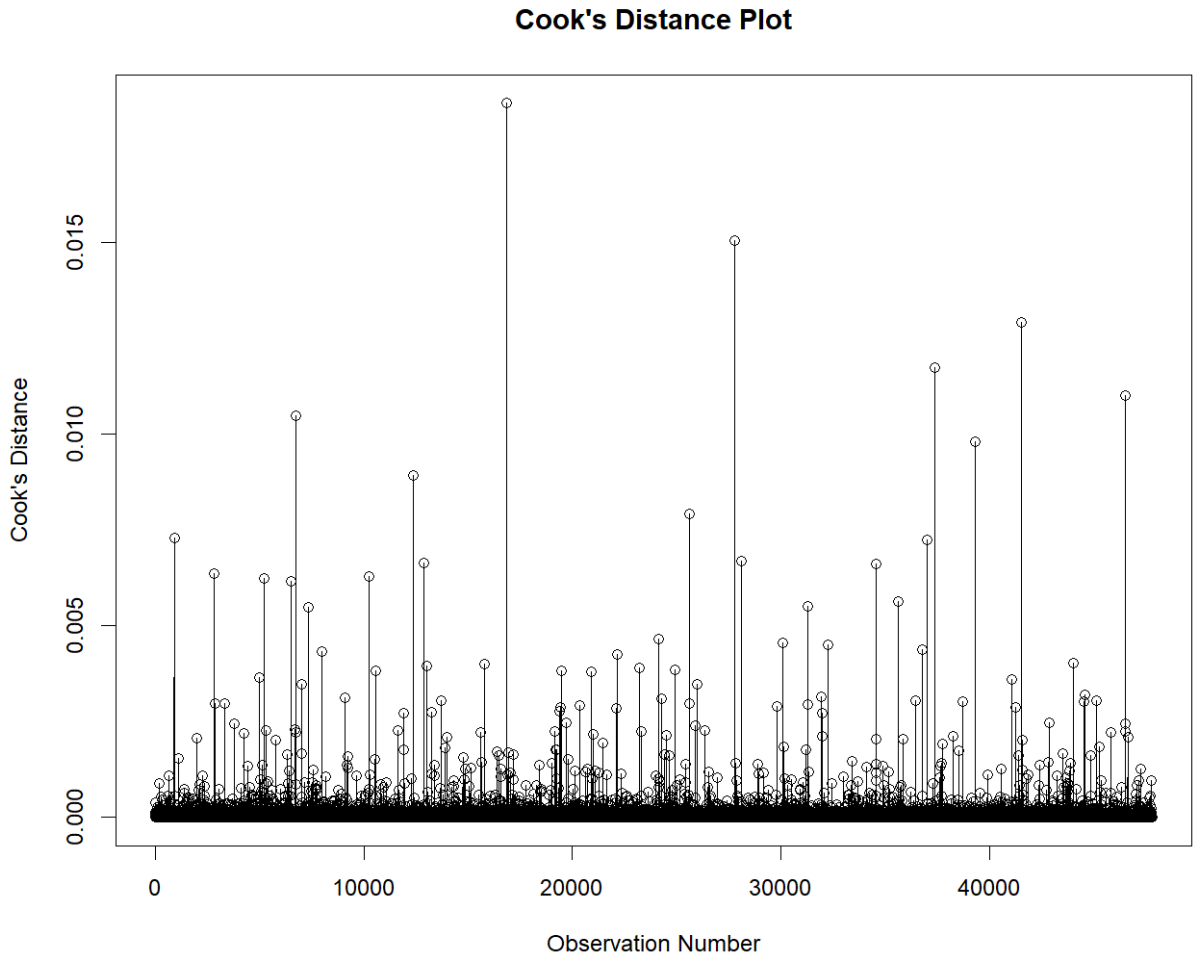
Residuals vs Fitted:
This plot is used to check for non-linearity, heteroscedasticity and outliers. On the x-axis, it plots the fitted values from the regression model, and on the y-axis, it plots the residuals. The residuals are randomly scattered around the horizontal line at y = 0, without any discernible pattern. This indicates that the model's assumptions are satisfied, and the residuals have constant variance. Although a curvature is present, its magnitude does not indicate non-linearity.

Normal Q-Q Plot (Quantile-Quantile Plot):
This plot assesses whether the residuals are normally distributed. It compares the quantiles of the residuals to the quantiles of a normal distribution. The x-axis represents the theoretical quantiles, and the y-axis represents the standardized residuals. The points lie roughly along the reference line. This suggests that the residuals are normally distributed. A minor deviation is present at the beginning of the line but given the scale of the observations it does not violate the assumptions of the regression model.

Scale-Location Plot (or Spread-Location Plot):
This plot checks for homoscedasticity. It plots the square root of the standardized residuals on the y-axis against the fitted values on the x-axis. The points are randomly spread around a horizontal line. This indicates that the variance of the residuals is constant across all levels of the fitted values.

Residuals vs Leverage:
This plot identifies influential data points that have a significant impact on the regression model. It plots leverage, which is a measure of how far an independent variable deviates from its mean, on the x-axis against the standardized residuals on the y-axis. Most points lie within the middle.

Cook's Distance
A Cook's distance plot is a tool used in regression analysis to identify influential data points. It measures the impact of removing each observation on the regression coefficients. It is noticeable that most of the points are clustered near zero.



Assumptions of the Regression

For multivariate linear regression to provide trustworthy results, we must adhere to certain assumptions.

Multicollinearity

The Variance Inflation Factor (VIF) test helps detect multicollinearity in regression analysis by measuring how much the variance of a regression coefficient is inflated due to correlation with other independent variables. High VIF values indicate multicollinearity issues, typically above 5. Identifying and addressing multicollinearity is crucial for ensuring the reliability of regression results.

```
     accommodates             bathrooms   host_response_rate              latitude             longitude
         3.612459              1.573175              1.024080              5.092619              5.099596
number_of_reviews review_scores_rating              bedrooms                  beds
         1.027831              1.016621              2.575882              3.444738
```

As observed, all VIF values are below 5, except `latitude` and `longitude`, as anticipated. To address this, we propose employing Principal Component Analysis (PCA) to transform `latitude` and `longitude` into uncorrelated components. Initially, we standardize `longitude` and `latitude` to ensure that PCA treats both dimensions equally, then we perform PCA on the standardized `latitude` and `longitude` and the resulting principal components are incorporated into the dataset, replacing the original latitude and longitude variables. Last, we check the VIF values to ensure multicollinearity is reduced

```
accommodates             bathrooms   host_response_rate    number_of_reviews review_scores_rating
    3.612459              1.573175              1.024080              1.027831              1.016621
    bedrooms                  beds                  pca1                  pca2
    2.575882              3.444738              1.021470              1.012026
```

Normality of the Residuals

The assumption of normality of residuals is a fundamental aspect of regression analysis, including multivariate linear regression. Residuals, which are the differences between observed and predicted values, should ideally follow a normal distribution. This assumption is crucial because it underpins the validity of various statistical tests and confidence intervals derived from the regression model. Ensuring that residuals are normally distributed allows for accurate estimation of parameters, reliable hypothesis testing, and robust inference. Let's perform some commonly used tests.

Kolmogorov-Smirnov Test

This test compares the sample distribution with a reference probability distribution, in this case, a normal distribution and is suitable for larger sample sizes.

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  resid.reg
D = 0.011197, p-value = 1.251e-05
alternative hypothesis: two-sided
```

The results of the Kolmogorov-Smirnov test indicate that the residuals do not follow a normal distribution since the p-value is much less than the significance level (1.251e-05 < 0.05), so we reject the null hypothesis.

Anderson-Darling Test

This test gives more weight to the tails of the distribution than the Kolmogorov-Smirnov test, making it more sensitive to deviations in the tails.

```
        Anderson-Darling normality test

data:  resid.reg
A = 14.969, p-value < 2.2e-16
```

The results of the Anderson-Darling test indicate that the residuals do not follow a normal distribution. The extremely small p-value ,less than 2.2e-16, provides strong evidence against the null hypothesis, suggesting a significant deviation from normality.

Jarque-Bera Test

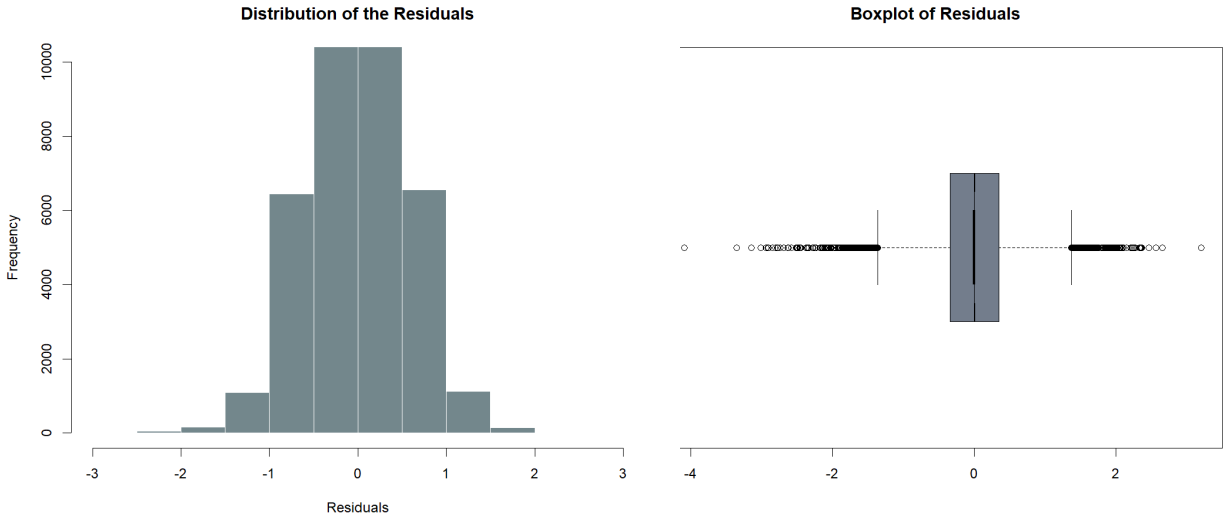
This test is based on the skewness and kurtosis of the sample data. It tests whether the data has the skewness and kurtosis matching a normal distribution.

```
        Jarque Bera Test

data:  resid.reg
X-squared = 2042.5, df = 2, p-value < 2.2e-16
```

The Jarque-Bera test results show that the residuals are not normally distributed. The very small p-value ,less than 2.2e-16, strongly rejects the null hypothesis, indicating a significant deviation from normality in the residuals.

Although the following histogram and boxplot show that the residuals are normally distributed is probably because of the large scale of the data. Below we can see the difference between the Median and the Mean.

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-4.286848 -0.338456  0.002063  0.000000  0.344939  3.204590
```



Homoscedasticity

The variance of the residuals should be constant across all levels of the independent variables. This implies that the spread of the residuals should remain consistent throughout the range of the predicted values. The **Breusch-Pagan** test is a statistical test used to assess the presence of heteroscedasticity in the residuals of a regression model.

```
        studentized Breusch-Pagan test

data:  regressor
BP = 1674.7, df = 9, p-value < 2.2e-16
```

The results of the Breusch-Pagan test indicate that there is heteroscedasticity in the residuals. The extremely small p-value ,less than 2.2e-16, provides strong evidence against the null hypothesis, suggesting that the variance of the residuals is not constant across all levels of the independent variables, thus homoscedasticity exists.

Autocorrelation

Autocorrelation refers to the correlation between observations, it indicates whether there is a relationship between each data point and its preceding or succeeding data points. Positive autocorrelation suggests that adjacent data points are similar, while negative

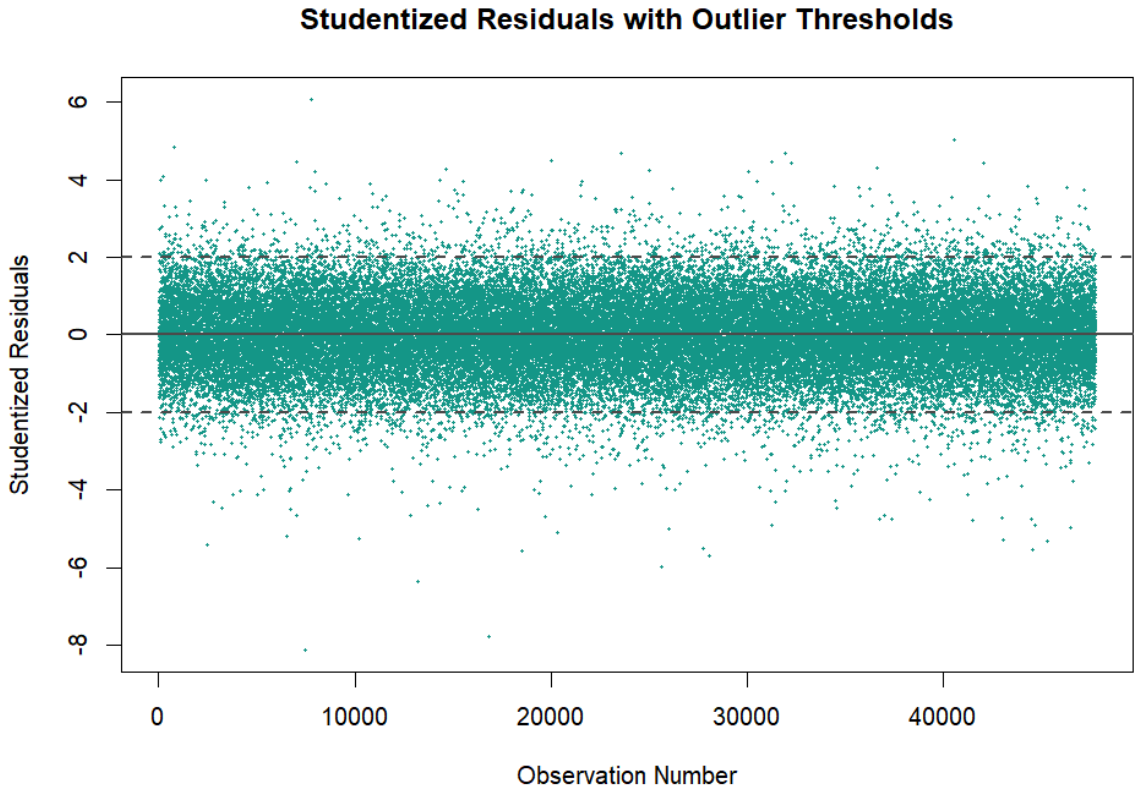
autocorrelation implies dissimilarity. The **Durbin-Watson** test is a statistical test used to detect autocorrelation in the residuals of a regression model.

```
 lag Autocorrelation D-W Statistic p-value
   1     0.007810014      1.984377    0.14
Alternative hypothesis: rho != 0
```

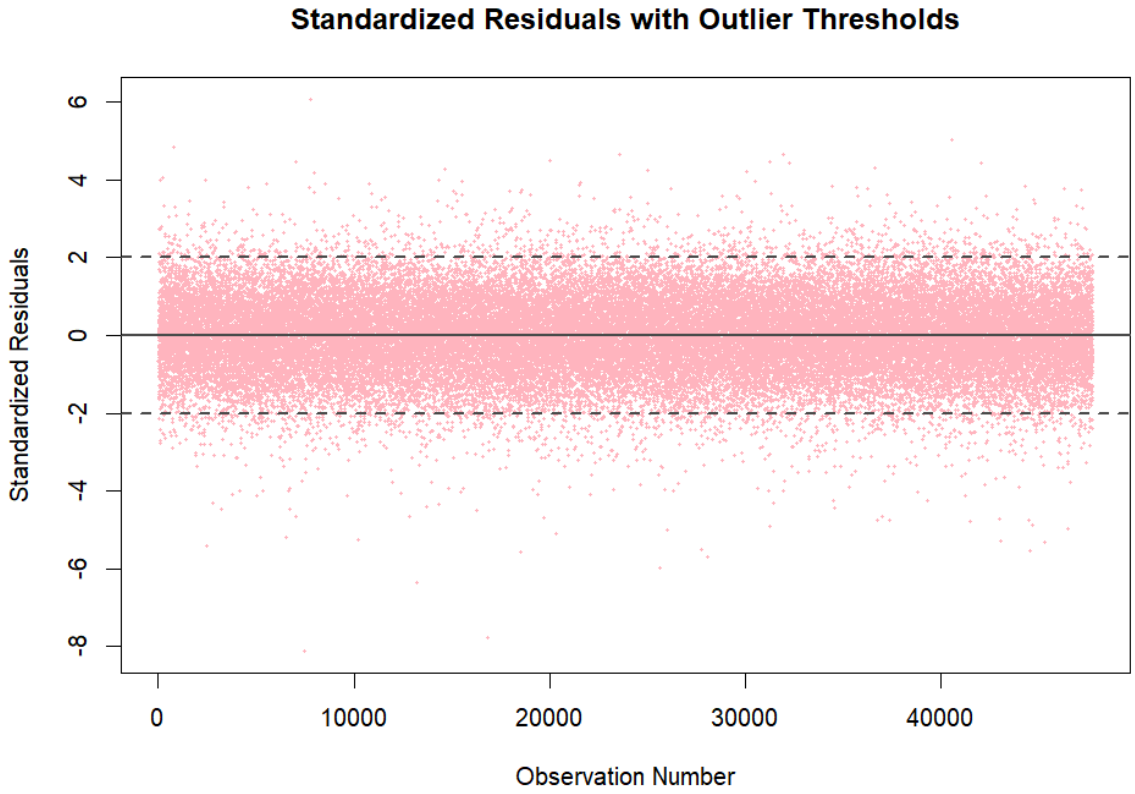
The results of the Durbin-Watson test suggest that there is no significant autocorrelation in the residuals of the regression model. The p-value exceeding 0.05 indicates that we do not have enough evidence to reject the null hypothesis of no autocorrelation. Thus, there is no autocorrelation in the residuals.

Analysis of the Residuals

In this section of the report, we will utilize two plots to analyze the studentized and standardized residuals. We start this process by concentrating on the studentized residuals from our regression analysis.



From the scatter plot above, we understand that there are some outliers in the studentized residuals, which makes sense given the number of observations. Similarly, the standardized residuals scatter plot reveals a comparable pattern.



Conclusions of Regression

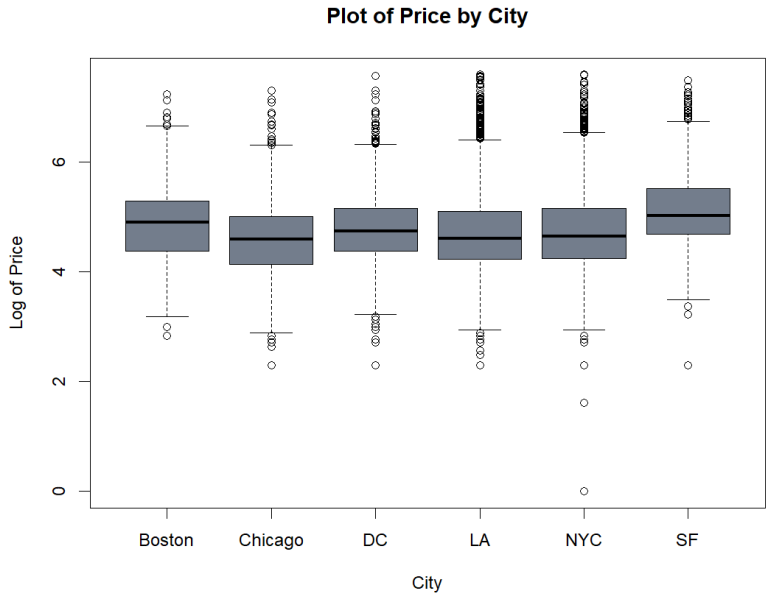
To initiate our multivariate linear regression analysis, we employed the stepwise method to identify which independent variables should be retained in our final regression model. This method indicated that all variables should be included. After dividing the data into training and testing sets, we developed the final multiple linear regression model to perform predictions. The predictions on the testing set yielded several metrics, with MAPE (Mean Absolute Percentage Error) and Min-Max Accuracy being the most significant. Notably, the mean accuracy between actual and predicted values was nearly 0%, indicating a highly accurate prediction rate of approximately 99.9% close to the actual values.

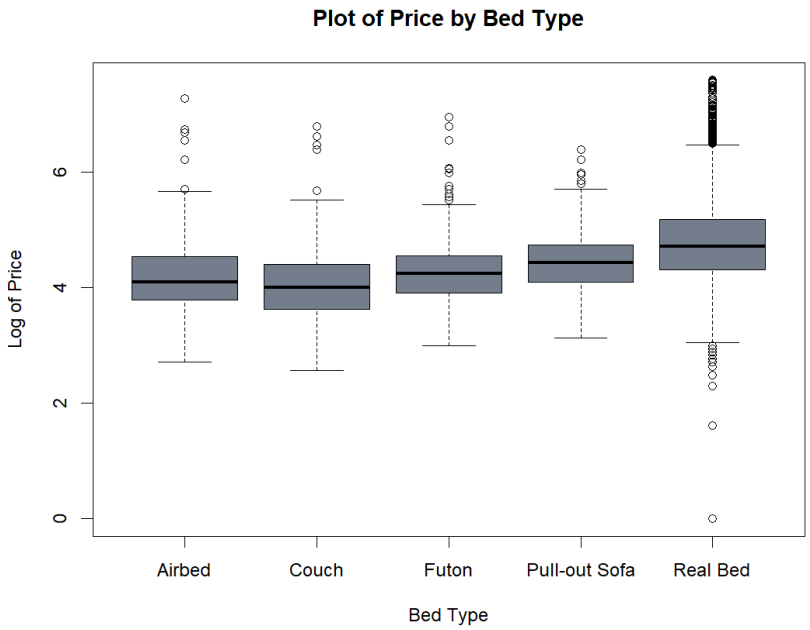
Our analysis determined that approximately 39% of the variability in the dependent variable is explained by the independent variables. Furthermore, the model demonstrated statistical significance with a p-value below 5%. Upon examining the regression assumptions, we found that the residuals were not normally distributed and that there was no autocorrelation. Although initial multicollinearity was detected, we resolved it by scaling and applying PCA to the `longitude` and `latitude` variables. However, heteroscedasticity was present in the model. Finally, plots and tables of studentized and standardized residuals indicated the presence of some outliers. Given the scale of the observations, these outliers are deemed acceptable.

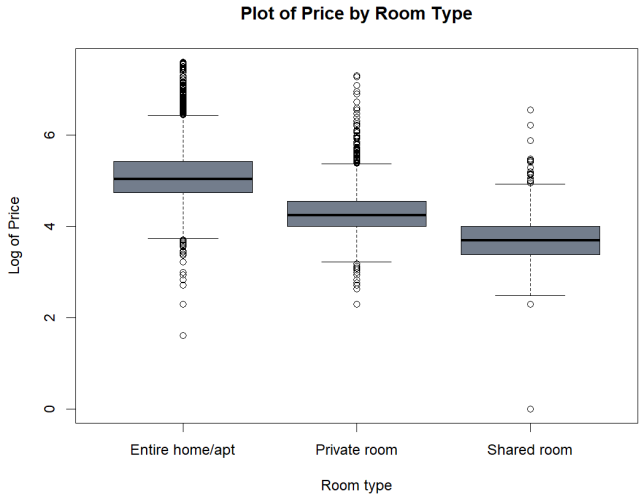
Empirical Analysis of the Qualitative Variables

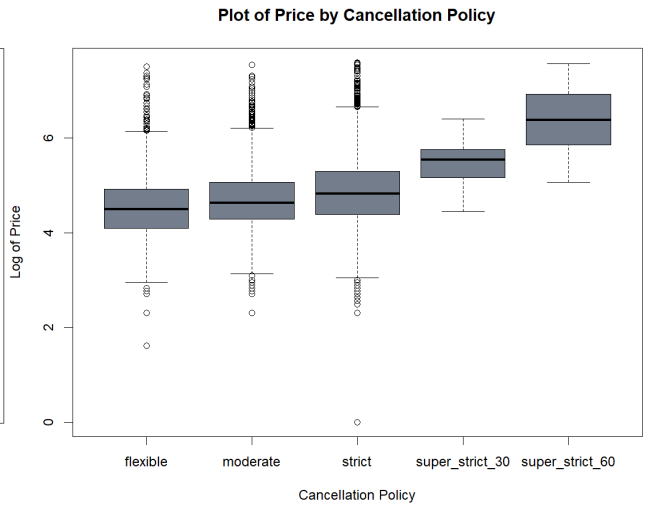
Descriptive Statistics of the Qualitative Variables

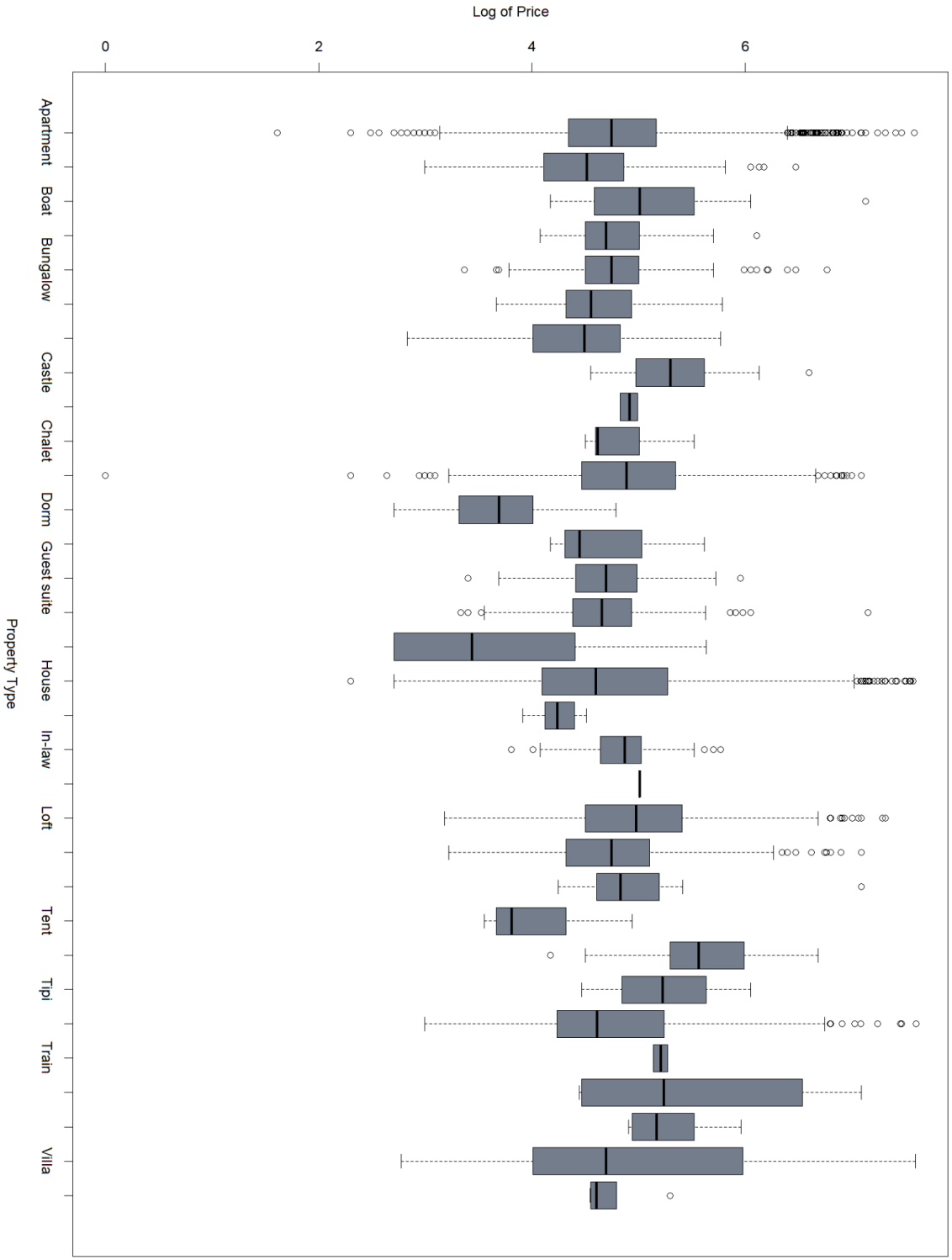
Descriptive statistics are essential for summarizing and understanding qualitative variables, which represent data divided into distinct categories based on attributes. By analyzing qualitative data, we aim to identify key insights and establish a foundation for more complex analyses, using detailed tables and visual representations to present a clear overview of the categorical data. Below we can see that the qualitative variables affect the log_price.




Plot of Price by Bed Type
Log of Price
0
2
4
6
Airbed
Couch
Futon
Pull-out Sofa
Real Bed
Bed Type


Plot of Price by Room Type
Log of Price
0
2
4
6
Entire home/apt
Private room
Shared room
Room type


Plot of Price by Cancellation Policy
Log of Price
0
2
4
6
flexible
moderate
strict
super_strict_30
super_strict_60
Cancellation Policy



Plot of Price by Property Type

Statistical Significance with ANOVA

Statistical significance with ANOVA (Analysis of Variance) determines if there are significant differences between the means of three or more groups. By analyzing the variance within and between groups, ANOVA tests the hypothesis that all group means are equal. A low p-value, typically below 0.05, indicates significant differences, suggesting at least one group mean is different. This method is widely used to compare effects across various fields.

Room Type and Price

We have performed an ANOVA test to compare the means of log-transformed prices across different room types.

```
                Df Sum Sq Mean Sq F value Pr(>F)
room_type        2   9233    4616   17528 <2e-16 ***
Residuals    47784  12585       0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA summary results indicate a significant difference in the mean log-transformed prices across different room types. The analysis shows that the sum of squares for room type is 9233 with 2 degrees of freedom, and the sum of squares for residuals is 12585 with 47784 degrees of freedom. The mean square for room type is 4616, while the mean square for residuals is effectively zero. The F value of 17528, with a p-value of less than 2e-16, strongly suggests that the observed differences in means are not due to random chance. This extremely low p-value, far below the conventional threshold of 0.05, leads us to reject the null hypothesis that the mean log-transformed prices are equal across room types. Therefore, we conclude that at least one room type has a significantly different mean log-transformed price compared to the others.

```
        Anderson-Darling normality test

data:  myaov_room_type$residuals
A = 209.27, p-value < 2.2e-16
```

The Anderson-Darling test suggests that the data significantly deviate from a normal distribution.

```
        studentized Breusch-Pagan test

data:  myaov_room_type
BP = 761.24, df = 2, p-value < 2.2e-16
```

The Breusch-Pagan test suggests evidence against the null hypothesis of homoscedasticity, indicating that heteroscedasticity may be present in the model.

```
 lag Autocorrelation D-W Statistic p-value
   1    -0.003324989      2.006644   0.538
 Alternative hypothesis: rho != 0
```

The high p-value in the Durbin-Watson test and the Statistic value close to 2 indicates that there is no autocorrelation.

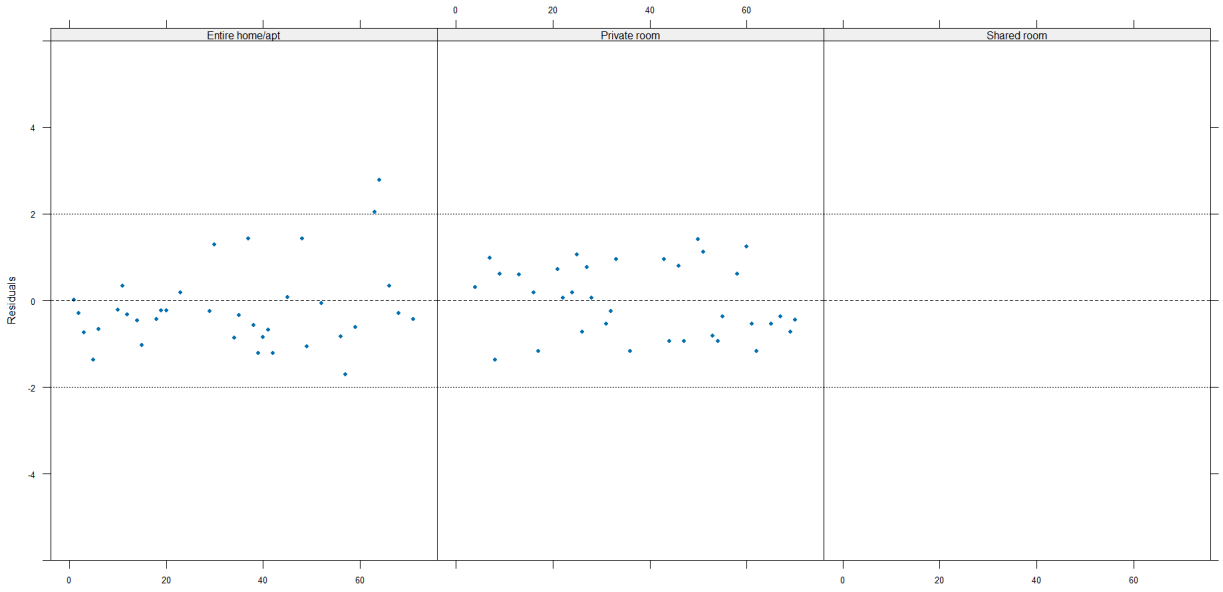
```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     2  459.48 < 2.2e-16 ***
      47784
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

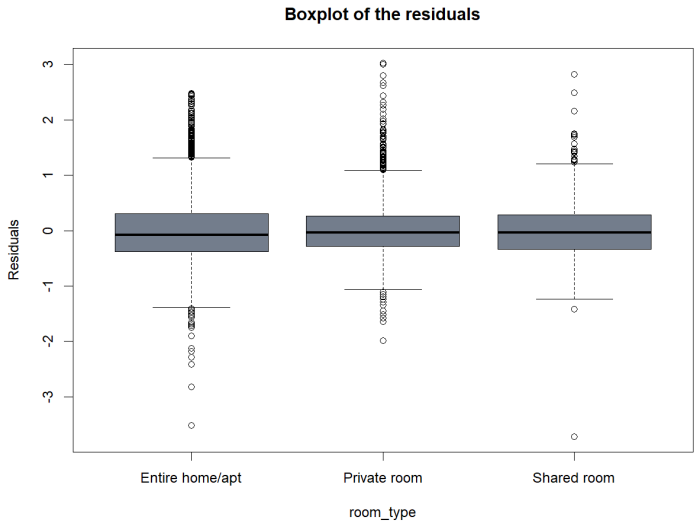
Levene’s test suggests that there is evidence against the null hypothesis of homogeneity of variances.

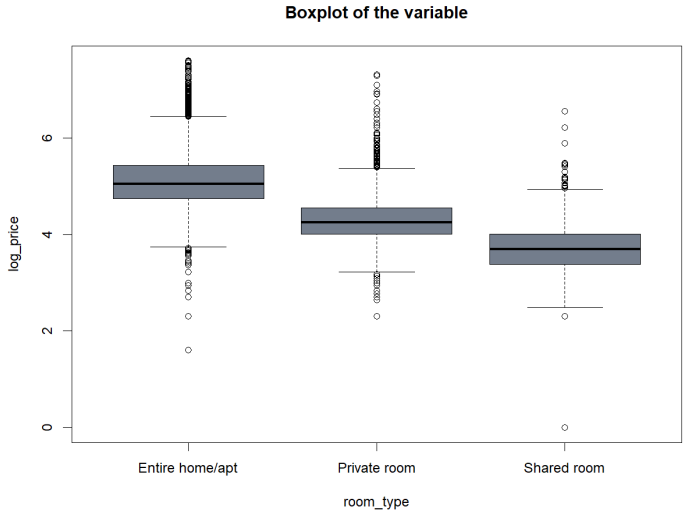
```
    One-way analysis of means (not assuming equal variances)

data:  log_price and room_type
F = 18230, num df = 2.0, denom df = 3380.7, p-value < 2.2e-16
```

The One-way test shows that there are significant differences among the group means.

Below we can observe the distribution of the residuals through different types of plots.


Entire home/apt
Private room
Shared room
Residuals
0
20
40
60
4
2
0
-2
-4


Boxplot of the residuals
Residuals
3
2
1
0
-1
-2
-3
Entire home/apt
Private room
Shared room
room_type


Boxplot of the variable
log_price
6
4
2
0
Entire home/apt
Private room
Shared room
room_type

Bed Type and Price

```
               Df Sum Sq Mean Sq F value Pr(>F)
bed_type        4    289   72.21   160.3 <2e-16 ***
Residuals   47782  21529    0.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA summary results for bed type indicate a significant difference in the mean log-transformed prices among the different bed types. The analysis reveals that the sum of squares for bed type is 289 with 4 degrees of freedom, and the sum of squares for residuals is 21529 with 47782 degrees of freedom. The mean square for bed type is 72.21, whereas the mean square for residuals is 0.45. The F value is 160.3, with a p-value less than 2e-16, demonstrating that the differences in means are statistically significant. This exceptionally low p-value, much smaller than the standard threshold of 0.05, leads us to reject the null hypothesis that the mean log-transformed prices are the same across bed types. Consequently, we conclude that at least one bed type has a significantly different mean log-transformed price compared to the others.

```
        Anderson-Darling normality test

data:  myaov_bed_type$residuals
A = 92.037, p-value < 2.2e-16
```

The Anderson-Darling test indicates that the data significantly diverge from a normal distribution.

```
        studentized Breusch-Pagan test

data:  myaov_bed_type
BP = 51.768, df = 4, p-value = 1.543e-10
```

The Breusch-Pagan test provides evidence against the null hypothesis of homoscedasticity, suggesting that heteroscedasticity might be present in the model.

```
 lag Autocorrelation D-W Statistic p-value
   1     0.002711649       1.99457   0.614
 Alternative hypothesis: rho != 0
```

The high p-value in the Durbin-Watson test, along with a statistic value close to 2, indicates the absence of autocorrelation.

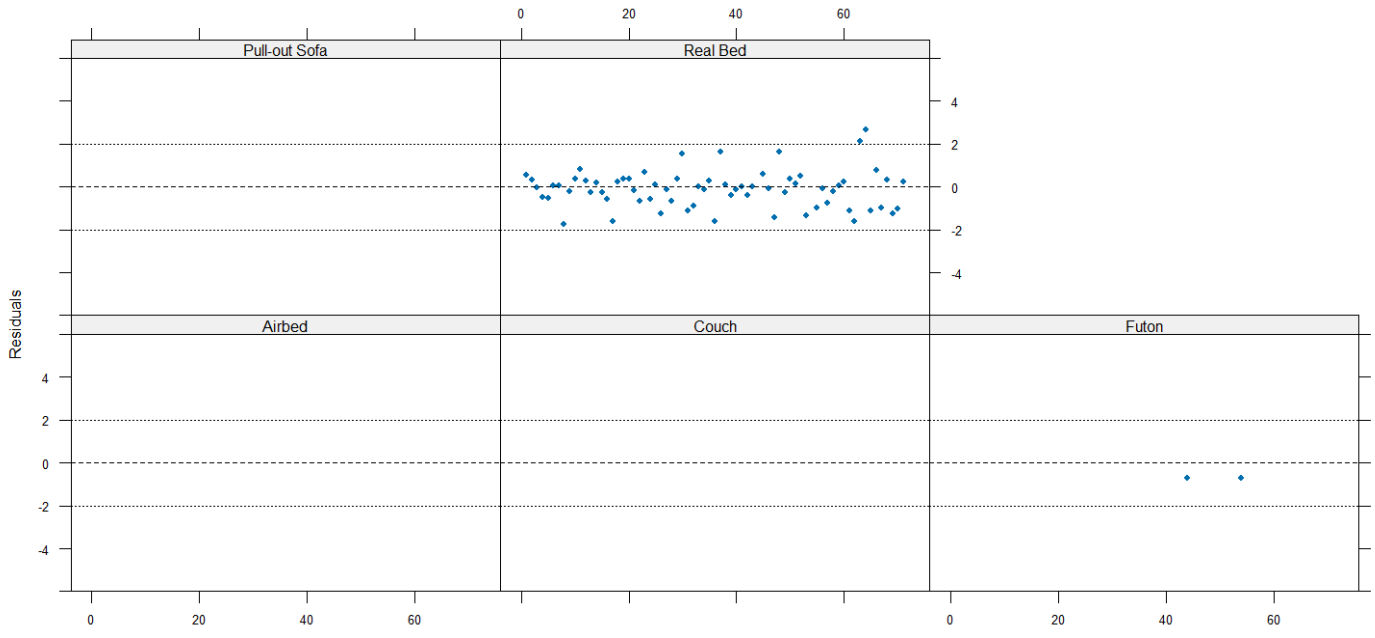
```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     4  19.981 < 2.2e-16 ***
      47782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

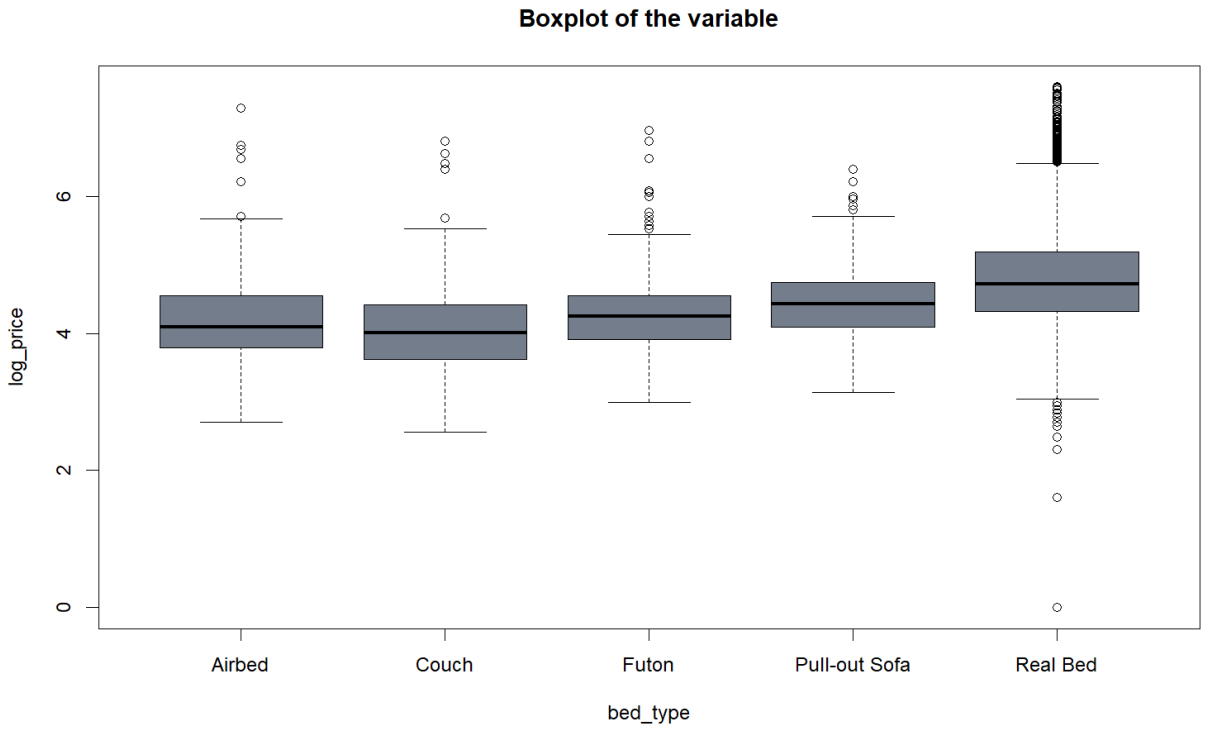
Levene's test indicates evidence against the null hypothesis of homogeneity of variances.

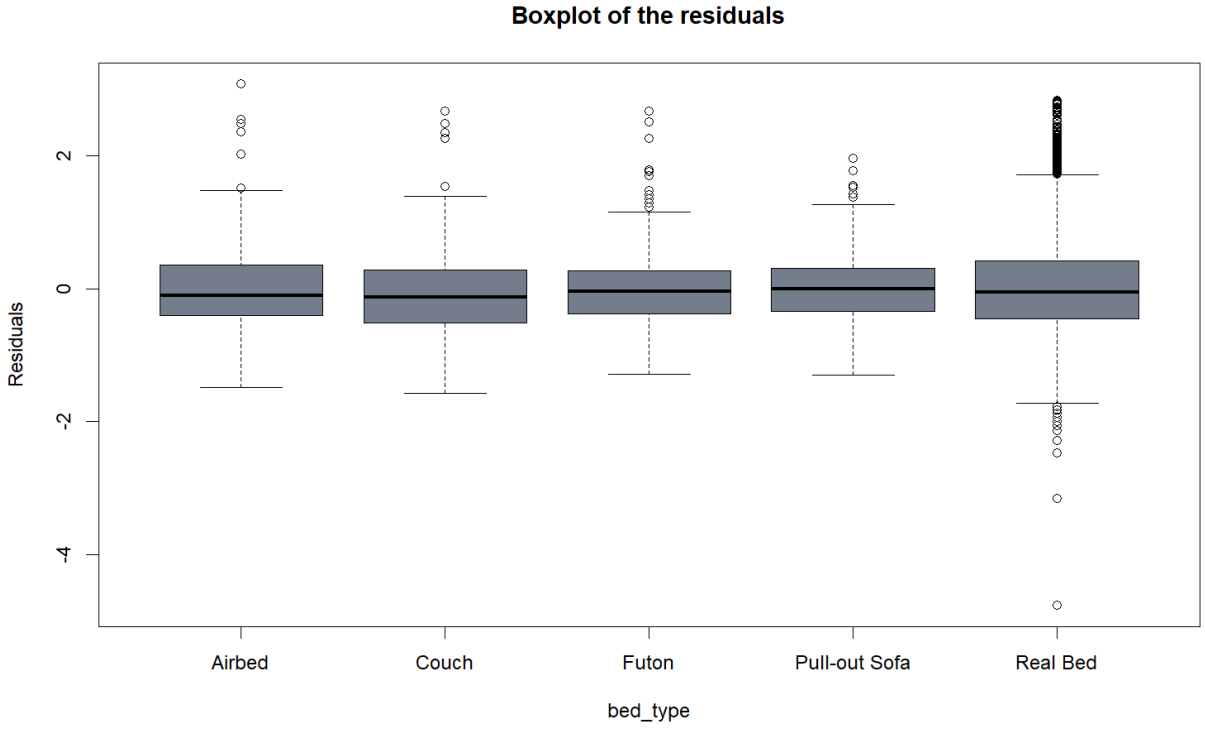
```
        One-way analysis of means (not assuming equal variances)

data:  log_price and bed_type
F = 203.02, num df = 4.00, denom df = 502.93, p-value < 2.2e-16
```

The One-way test indicates significant differences among the group means.

Below, we can observe the distribution of the residuals using various types of plots.




Boxplot of the variable
log_price
0
2
4
6
Airbed
Couch
Futon
Pull-out Sofa
Real Bed
bed_type


Boxplot of the residuals
Residuals
-4
-2
0
2
Airbed
Couch
Futon
Pull-out Sofa
Real Bed
bed_type

Cancellation Policy and Price

The ANOVA results for cancellation policies show a significant difference in mean log-transformed prices among different policies. With a very low p-value <2e-16 and an F value of 474.6, the analysis indicates that at least one cancellation policy has a significantly different mean log-transformed price compared to the others.

```
                      Df Sum Sq Mean Sq F value Pr(>F)
cancellation_policy     4    834  208.43   474.6 <2e-16 ***
Residuals           47782  20984    0.44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The Anderson-Darling test suggests that the data significantly deviate from a normal distribution.

```
        Anderson-Darling normality test

data:  myaov_cancellation_policy$residuals
A = 65.408, p-value < 2.2e-16
```

The Breusch-Pagan test suggests evidence against the null hypothesis of homoscedasticity, indicating that heteroscedasticity may be present in the model.

```
        studentized Breusch-Pagan test

data:  myaov_cancellation_policy
BP = 452.56, df = 4, p-value < 2.2e-16
```

The high p-value in the Durbin-Watson test and the Statistic value close to 2 indicates that there is no autocorrelation.

```
 lag Autocorrelation D-W Statistic p-value
   1     0.005418206      1.989159   0.216
 Alternative hypothesis: rho != 0
```

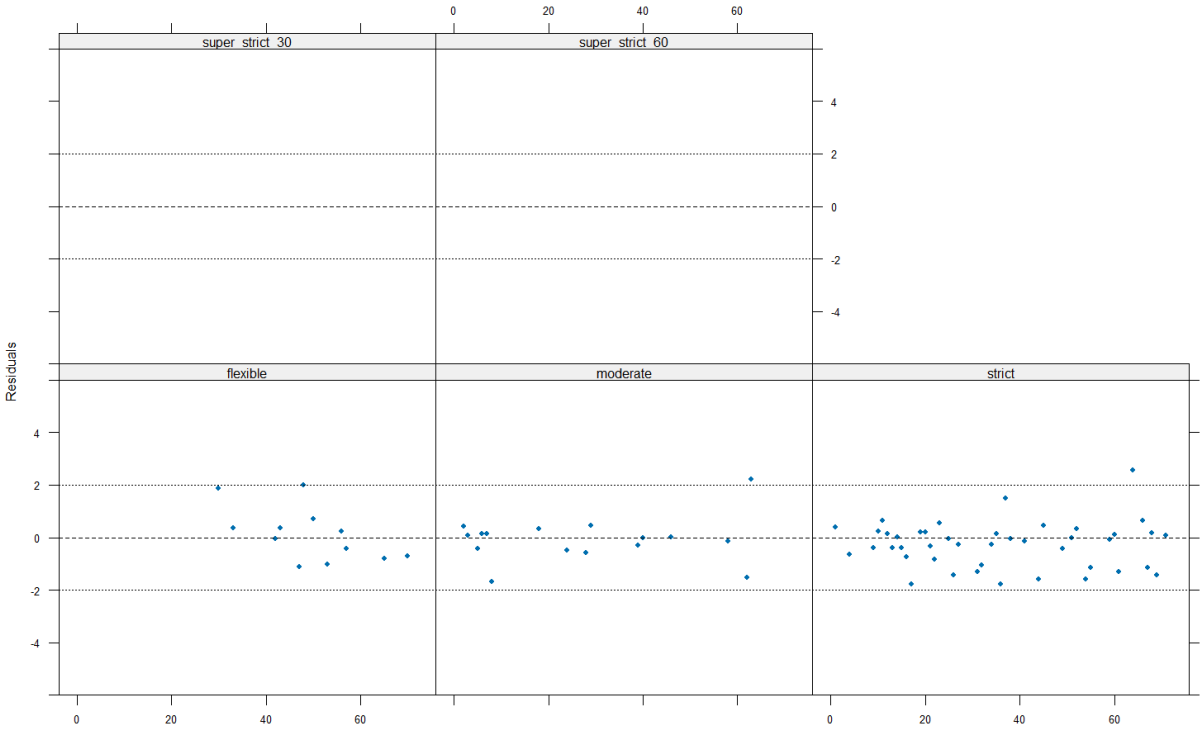
Levene’s test suggests that there is evidence against the null hypothesis of homogeneity of variances.

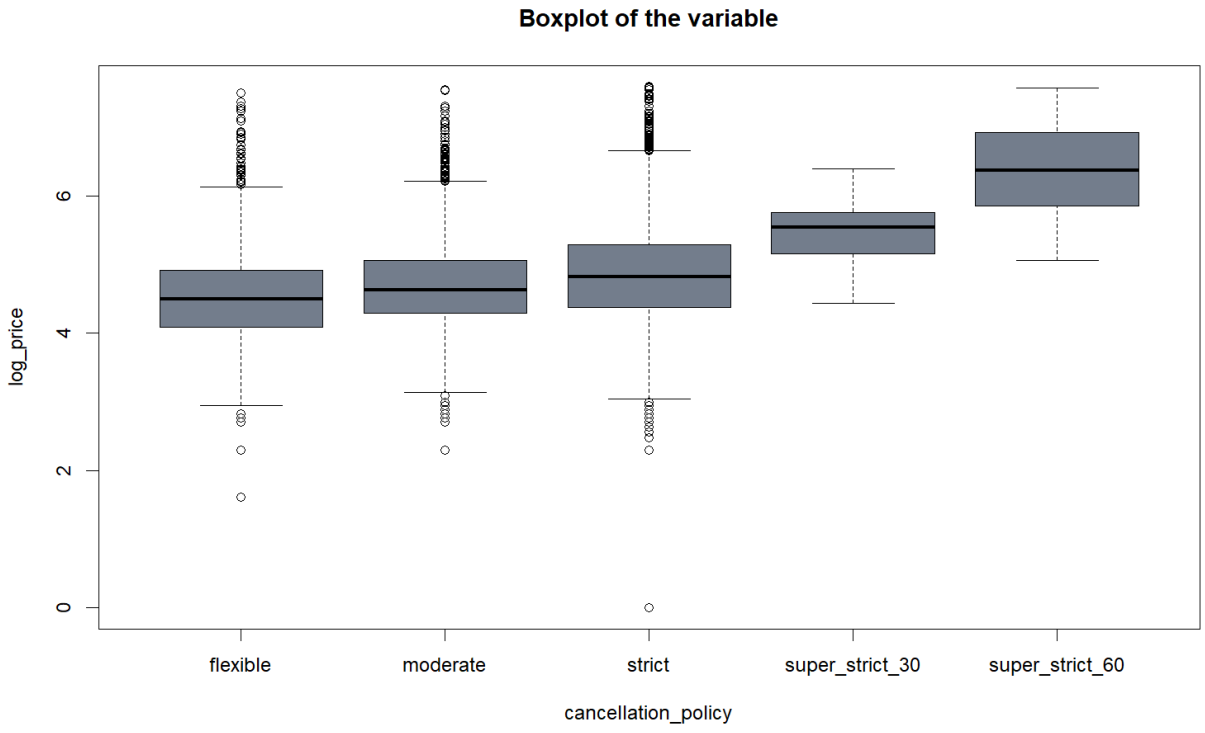
```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     4  129.94 < 2.2e-16 ***
      47782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

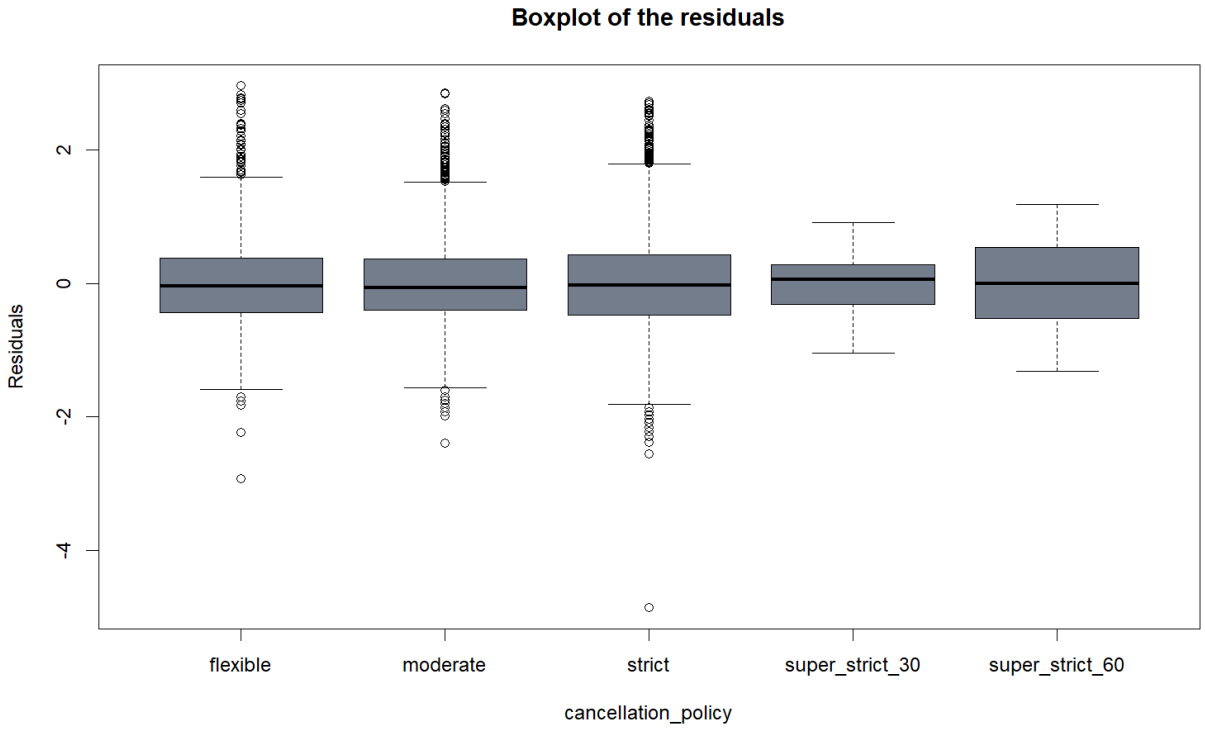
The One-way test shows that there are significant differences among the group means.

```
        One-way analysis of means (not assuming equal variances)

data:  log_price and cancellation_policy
F = 504.97, num df = 4.000, denom df = 64.473, p-value < 2.2e-16
```

Below we can observe the distribution of the residuals through different types of plots.




Boxplot of the variable
log_price
0
2
4
6
flexible
moderate
strict
super_strict_30
super_strict_60
cancellation_policy


Boxplot of the residuals
Residuals
-4
-2
0
2
flexible
moderate
strict
super_strict_30
super_strict_60
cancellation_policy

City and Price

The ANOVA analysis for cities indicates a significant variation in mean log-transformed prices among the different cities. The extremely low p-value <2e-16 and high F value of 319.7 suggest that at least one city has a significantly different mean log-transformed price compared to the others.

```
                Df Sum Sq Mean Sq F value Pr(>F)
city             5    706  141.26   319.7 <2e-16 ***
Residuals    47781  21112    0.44
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
            Anderson-Darling normality test

data:  myaov_city$residuals
A = 97.943, p-value < 2.2e-16
```

```
            studentized Breusch-Pagan test

data:  myaov_city
BP = 190.5, df = 5, p-value < 2.2e-16
```

```
 lag Autocorrelation D-W Statistic p-value
   1     0.002479425      1.995032   0.612
 Alternative hypothesis: rho != 0
```

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     5  24.602 < 2.2e-16 ***
      47781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

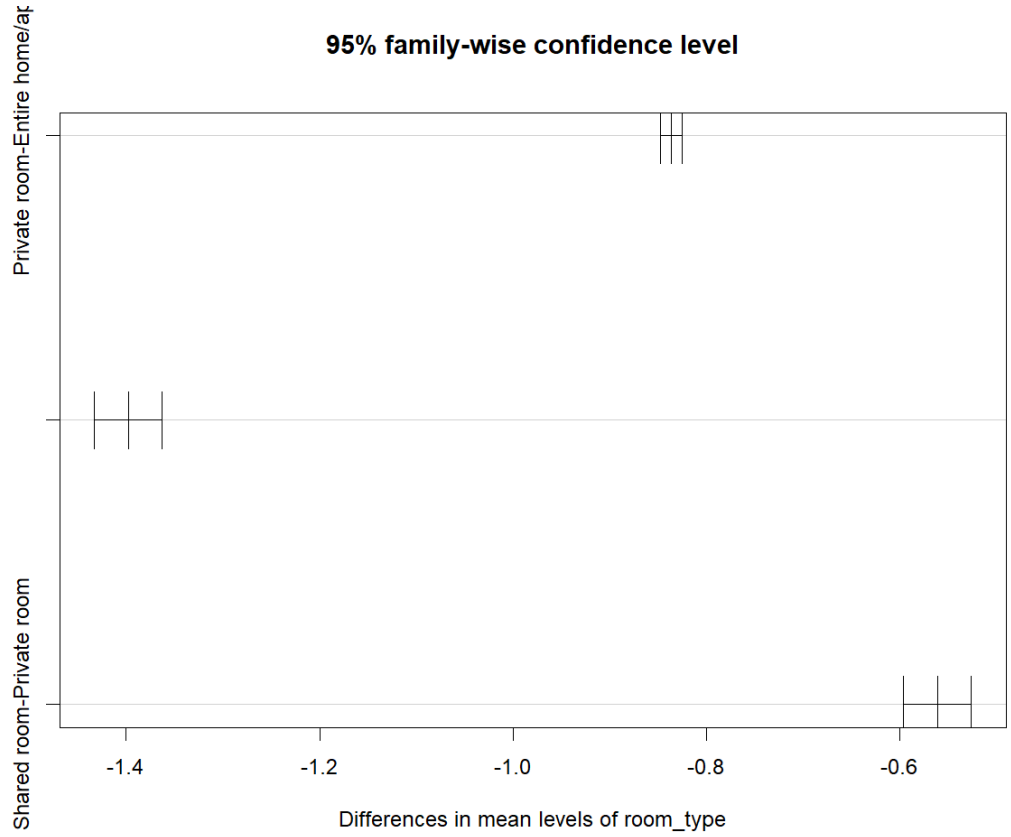
```
        One-way analysis of means (not assuming equal variances)

data:  log_price and city
F = 367.22, num df = 5, denom df = 10216, p-value < 2.2e-16
```

The results of the tests indicate that the data significantly deviate from a normal distribution, providing evidence against the null hypothesis of homoscedasticity and suggesting the presence of heteroscedasticity in the model. Furthermore, the tests reveal no autocorrelation, demonstrate evidence against the null hypothesis of homogeneity of variances, and identify significant differences among the group means.

Tukey HSD

Tukey's Honestly Significant Difference (HSD) test is a statistical method used after ANOVA to compare multiple group means and pinpoint significant differences. It manages the family-wise error rate, minimizing the chance of false positives. By computing a critical value, it identifies which group means are significantly different from one another, serving as an important tool in data analysis and scientific studies.

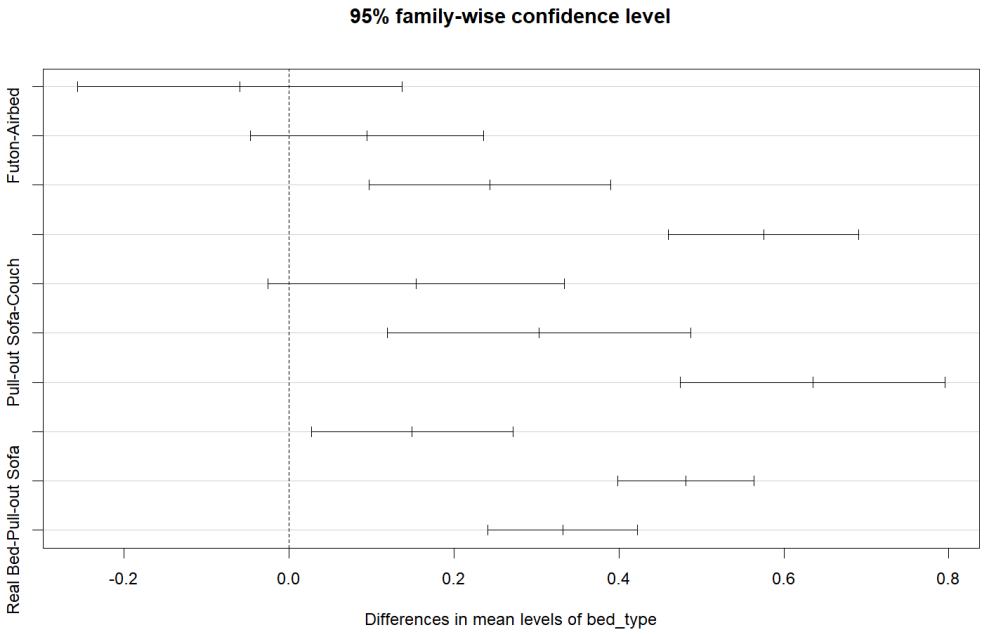
Room Type and Price




```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_price ~ room_type, data = data)

$room_type
                                    diff        lwr        upr p adj
Private room-Entire home/apt -0.8363957 -0.8477314 -0.8250600     0
Shared room-Entire home/apt  -1.3979786 -1.4327384 -1.3632189     0
Shared room-Private room     -0.5615829 -0.5966730 -0.5264929     0
```

The negative differences indicate that `Entire home/apt` has higher log-transformed prices compared to both `Private room` and `Shared room`, and `Private room` has higher log-transformed prices compared to `Shared room`. The strongest level of the factor variable `room_type`, in terms of log-transformed price, is `Entire home/apt`, followed by `Private room`, and then `Shared room`.

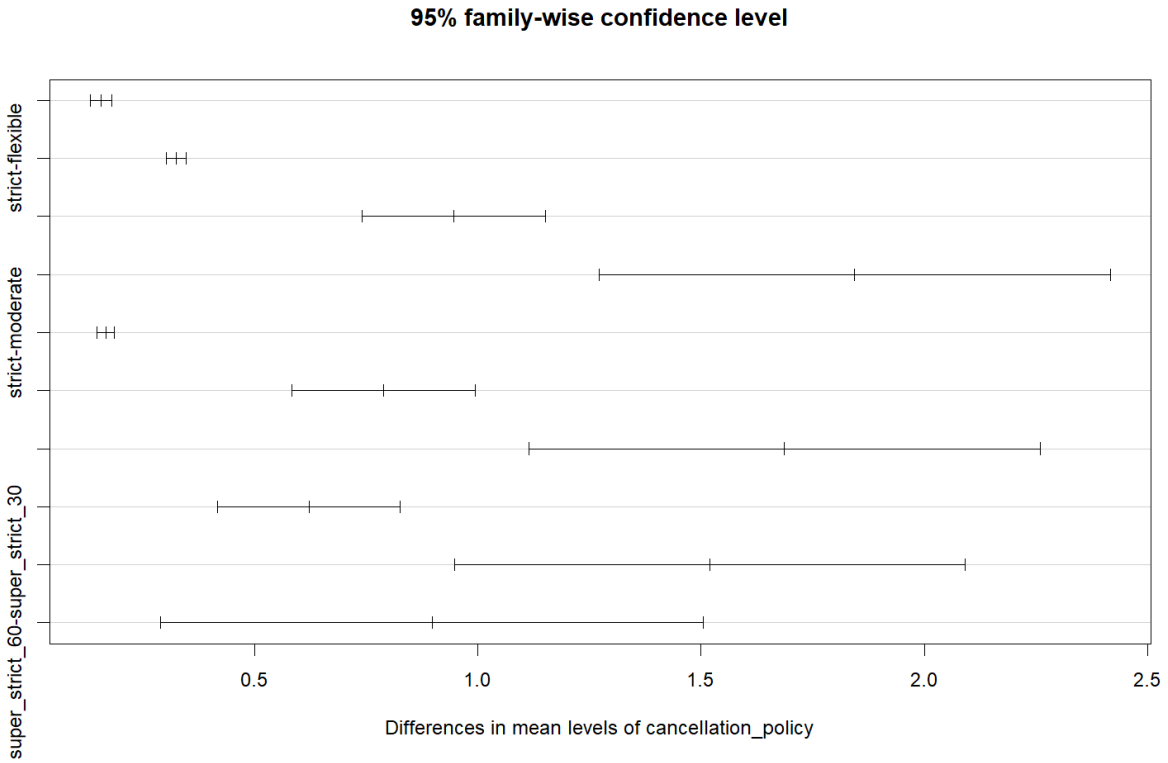
Bed Type and Price




```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_price ~ bed_type, data = data)

$bed_type
                           diff         lwr       upr     p adj
Couch-Airbed          -0.05962376 -0.25657954 0.1373320 0.9227538
Futon-Airbed           0.09442254 -0.04690028 0.2357454 0.3604186
Pull-out Sofa-Airbed   0.24356410  0.09735779 0.3897704 0.0000541
Real Bed-Airbed        0.57528077  0.46007957 0.6904820 0.0000000
Futon-Couch            0.15404630 -0.02585727 0.3339499 0.1336353
Pull-out Sofa-Couch    0.30318786  0.11942323 0.4869525 0.0000664
Real Bed-Couch         0.63490453  0.47470319 0.7951059 0.0000000
Pull-out Sofa-Futon    0.14914155  0.02687135 0.2714118 0.0078156
Real Bed-Futon         0.48085822  0.39812351 0.5635929 0.0000000
Real Bed-Pull-out Sofa 0.33171667  0.24089169 0.4225416 0.0000000
```

The Tukey test results indicate that `Real Bed` and `Pull-out Sofa` have significantly higher log-transformed prices compared to other bed types. `Real Bed` shows the highest prices, followed by `Pull-out Sofa`. Differences among `Futon`, `Couch`, and `Airbed` are not statistically significant. Thus, `Real Bed` is the strongest bed type in terms of price, with `Pull-out Sofa` also performing well.

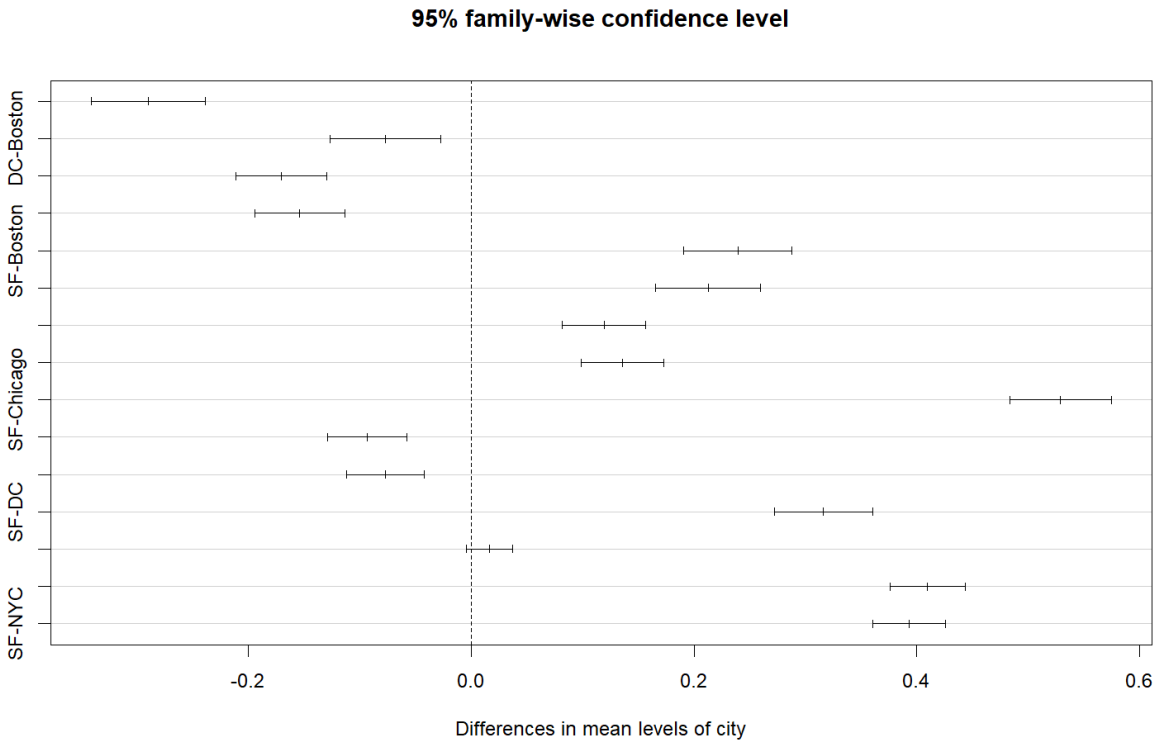
Cancellation Policy and Price




```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_price ~ cancellation_policy, data = data)

$cancellation_policy
                                      diff       lwr       upr     p adj
moderate-flexible                0.1570830 0.1327470 0.1814190 0.0000000
strict-flexible                  0.3244256 0.3023100 0.3465411 0.0000000
super_strict_30-flexible         0.9464858 0.7409357 1.1520358 0.0000000
super_strict_60-flexible         1.8440390 1.2720850 2.4159930 0.0000000
strict-moderate                  0.1673426 0.1481569 0.1865283 0.0000000
super_strict_30-moderate         0.7894028 0.5841473 0.9946583 0.0000000
super_strict_60-moderate         1.6869560 1.1151078 2.2588042 0.0000000
super_strict_30-strict           0.6220602 0.4170561 0.8270643 0.0000000
super_strict_60-strict           1.5196134 0.9478554 2.0913714 0.0000000
super_strict_60-super_strict_30 0.8975532 0.2903721 1.5047343 0.0005294
```

The strongest level of the factor variable `cancellation_policy`, in terms of log-transformed price, is `super_strict_60`, followed by `super_strict_30`, `strict`, `moderate`, and `flexible` policies. All differences between policies are statistically significant, demonstrating clear price variations based on the stringency of the cancellation policy.

City and Price




```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = log_price ~ city, data = data)

$city
                       diff          lwr         upr     p adj
Chicago-Boston -0.28952861 -0.340738327 -0.23831889 0.0000000
DC-Boston      -0.07692047 -0.126660660 -0.02718027 0.0001525
LA-Boston      -0.17027834 -0.211353702 -0.12920297 0.0000000
NYC-Boston     -0.15369609 -0.194170955 -0.11322122 0.0000000
SF-Boston       0.23927785  0.190779274  0.28777642 0.0000000
DC-Chicago      0.21260814  0.165811071  0.25940522 0.0000000
LA-Chicago      0.11925027  0.081792767  0.15670778 0.0000000
NYC-Chicago     0.13583252  0.099034505  0.17263054 0.0000000
SF-Chicago      0.52880646  0.483331291  0.57428162 0.0000000
LA-DC          -0.09335787 -0.128779849 -0.05793589 0.0000000
NYC-DC         -0.07677562 -0.111499472 -0.04205177 0.0000000
SF-DC           0.31619831  0.272384588  0.36001204 0.0000000
NYC-LA          0.01658225 -0.003883442  0.03704794 0.1902718
SF-LA           0.40955618  0.375899963  0.44321240 0.0000000
SF-NYC          0.39297394  0.360053269  0.42589460 0.0000000
```

The strongest level of the factor variable `city`, in terms of log-transformed price, is `SF`, followed by `Boston`, `DC`, `NYC` and `LA`. `Chicago` has the lowest log-transformed prices. All differences between cities are statistically significant except for the comparison between `NYC` and `LA`.

Final Conclusions

This study provides a comprehensive analysis of the factors influencing the pricing of Airbnb listings in the United States, leveraging a substantial dataset of over 50,000 entries. Our investigation reveals several key determinants of listing prices, including the number of bedrooms and beds, the city location, bed type, house type, room type, and accommodation capacity. Through the application of descriptive statistics, correlation analysis, regression models, training and testing predictions, ANOVA, and Tukey's Honest Significant Difference test, we have identified significant relationships between these variables and listing prices.

The results indicate that certain features, such as larger accommodation capacity and premium bed types, generally command higher prices. Additionally, the type of house and room, as well as the specific city of the listing, play crucial roles in price determination. These findings offer valuable insights for hosts aiming to optimize their listing attributes to achieve competitive pricing. For guests, understanding these price determinants can aid in making informed decisions when booking accommodations. Moreover, the implications for platform strategists are significant, as these insights can guide the development of more effective pricing algorithms and strategies to enhance market efficiency and user satisfaction.

Overall, this study contributes to the growing body of literature on the economics of short-term rentals and provides actionable recommendations for various stakeholders in the Airbnb ecosystem.

Bibliography

Dataset: https://www.kaggle.com/datasets/paramvir705/airbnb-data

Ιωαννίδης Δ. και Αθανασιάδης Ι.(2017) : "Στατιστική και Μηχανική Μάθηση με την R : Θεωρία και εφαρμογές", Εκδόσεις Τζιόλα. Θεσσαλονίκη.

ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ : Θεωρία και Εφαρμογές με Χρήση Excel & R (2018), Εκδόσεις Τζιόλα, Θεσσαλονίκη.

https://ggplot2.tidyverse.org/reference/ggtheme.html

https://r-graph-gallery.com/index.html

https://ftp.cc.uoc.gr/mirrors/CRAN/