
# Final Project Report

**Rutwiz Gangadhar Gullipalli**

**Kalyan Kumar Bhogi**

Intermediate Analytics- *ALY 6015*

Amin Karimpour

May 18th, 2023

# Introduction

The bulk of Airbnb datasets give information regarding the houses that are offered on the market as well as any related attributes. These features could be specifications like the kind of property (such an apartment or a house), the number of beds and bathrooms, the amenities provided, the location information, the price information, the host information, and reviews. In the Barcelona dataset, specific information about Airbnb listings in Barcelona is included, such as how much it costs to list a room during the week. Having access to the real data set is required in order to provide more in-depth information or perform any data analysis.

# Analysis

## Preliminary Analysis:

**Summary of the dataset:**

```
      ...1          realSum        room_type         room_shared       room_private      person_capacity host_is_superhost     multi
 Min.   :   0.0   Min.   :  69.59   Length:1555        Mode :logical   Mode :logical   Min.   :2.000   Mode :logical     Min.   :0.0000
 1st Qu.: 388.5   1st Qu.: 161.99   Class :character   FALSE:1547      FALSE:370       1st Qu.:2.000   FALSE:1274        1st Qu.:0.0000
 Median : 777.0   Median : 208.53   Mode  :character   TRUE :8         TRUE :1185      Median :2.000   TRUE :281         Median :0.0000
 Mean   : 777.0   Mean   : 288.39                                                      Mean   :2.756                     Mean   :0.3768
 3rd Qu.:1165.5   3rd Qu.: 335.37                                                      3rd Qu.:3.000                     3rd Qu.:1.0000
 Max.   :1554.0   Max.   :6943.70                                                      Max.   :6.000                     Max.   :1.0000
      biz          cleanliness_rating guest_satisfaction_overall    bedrooms          dist          metro_dist       attr_index
 Min.   :0.0000   Min.   : 2.000     Min.   : 20.00              Min.   :0.000   Min.   :0.1199   Min.   :0.0130   Min.   :  93.82
 1st Qu.:0.0000   1st Qu.: 9.000     1st Qu.: 88.00              1st Qu.:1.000   1st Qu.:1.0906   1st Qu.:0.2521   1st Qu.: 282.77
 Median :0.0000   Median :10.000     Median : 93.00              Median :1.000   Median :1.7518   Median :0.3705   Median : 389.20
 Mean   :0.3505   Mean   : 9.286     Mean   : 90.93              Mean   :1.217   Mean   :2.1173   Mean   :0.4349   Mean   : 464.37
 3rd Qu.:1.0000   3rd Qu.:10.000     3rd Qu.: 97.00              3rd Qu.:1.000   3rd Qu.:2.9492   3rd Qu.:0.5542   3rd Qu.: 591.59
 Max.   :1.0000   Max.   :10.000     Max.   :100.00              Max.   :6.000   Max.   :8.4440   Max.   :2.4028   Max.   :2934.13
 attr_index_norm    rest_index     rest_index_norm       lng             lat
 Min.   :  3.198   Min.   : 159.8   Min.   :  3.518   Min.   :2.105   Min.   :41.35
 1st Qu.:  9.637   1st Qu.: 494.4   1st Qu.: 10.883   1st Qu.:2.156   1st Qu.:41.38
 Median : 13.265   Median : 801.8   Median : 17.650   Median :2.171   Median :41.39
 Mean   : 15.827   Mean   : 877.7   Mean   : 19.320   Mean   :2.169   Mean   :41.39
 3rd Qu.: 20.162   3rd Qu.:1211.3   3rd Qu.: 26.663   3rd Qu.:2.179   3rd Qu.:41.40
 Max.   :100.000   Max.   :4542.8   Max.   :100.000   Max.   :2.226   Max.   :41.46
```

- The dataset contains 1,555 observations (listings), each of which provides information on a specific amenity of the lodgings. Significant information includes the following:
- Each listing has a unique numeric identity called Real Sum, which has a range of 0 to 1,554. There is one for each observation since it serves as an index or identifier.
- Room_type is a category variable that describes the type of lodging, such as a shared room or a whole apartment, a private room, or both. The mode of this variable represents the most prevalent room type.
- person_capacity variable specifies the maximum number of people the listing is capable of accommodating. The capacity for the average individual ranges from 2.76 to 6. This demonstrates that the bulk of ads, with person capacities ranging from 2.76 to 6, cater to small groups.

- The binary variable host_is_superhost can be used to detect if a host is designated as a superhost. Superhosts are seasoned hosts with excellent ratings and reviews.
- The variables cleanliness_rating and guest_satisfaction_overall, respectively, describe the evaluations of cleanliness and overall guest satisfaction. Both variables range from 2 to 10, with higher values indicating better ratings.
- bedrooms: The number of bedrooms in the accommodation is indicated by this variable, which has a range of 0 to 6. The median number reveals that one-bedroom listings are most common.
- Dist and metro_dist: These variables display how distant a lodging is from a certain location or metro stop. Their measuring methods and parameters vary.
- lng and lat, rest_index, rest_index_norm, attr_index, and attr_index_norm: These variables likely relate to latitude and longitude as well as the distances to nearby sites of interest or convenience. To analyse them precisely, further data from the dataset would be required.

| | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> | <dbl> |
| ...1 | 1 | 1555 | 777.00 | 449.03 | 777.00 | 777.00 | 576.73 | 0.00 | 1554.00 | 1554.00 | 0.00 | -1.20 | 11.39 |
| realSum | 2 | 1555 | 288.39 | 321.18 | 208.53 | 243.05 | 102.48 | 69.59 | 6943.70 | 6874.11 | 13.28 | 253.78 | 8.14 |
| room_type* | 3 | 1555 | 1.77 | 0.43 | 2.00 | 1.83 | 0.00 | 1.00 | 3.00 | 2.00 | -1.10 | -0.22 | 0.01 |
| room_shared | 4 | 1555 | NaN | NA | NA | NaN | NA | Inf | -Inf | -Inf | NA | NA | NA |
| room_private | 5 | 1555 | NaN | NA | NA | NaN | NA | Inf | -Inf | -Inf | NA | NA | NA |
| person_capacity | 6 | 1555 | 2.76 | 1.28 | 2.00 | 2.47 | 0.00 | 2.00 | 6.00 | 4.00 | 1.51 | 0.96 | 0.03 |
| host_is_superhost | 7 | 1555 | NaN | NA | NA | NaN | NA | Inf | -Inf | -Inf | NA | NA | NA |
| multi | 8 | 1555 | 0.38 | 0.48 | 0.00 | 0.35 | 0.00 | 0.00 | 1.00 | 1.00 | 0.51 | -1.74 | 0.01 |
| biz | 9 | 1555 | 0.35 | 0.48 | 0.00 | 0.31 | 0.00 | 0.00 | 1.00 | 1.00 | 0.63 | -1.61 | 0.01 |
| cleanliness_rating | 10 | 1555 | 9.29 | 1.01 | 10.00 | 9.46 | 0.00 | 2.00 | 10.00 | 8.00 | -2.54 | 11.12 | 0.03 |
| guest_satisfaction_overall | 11 | 1555 | 90.93 | 8.70 | 93.00 | 92.11 | 7.41 | 20.00 | 100.00 | 80.00 | -2.36 | 10.66 | 0.22 |
| bedrooms | 12 | 1555 | 1.22 | 0.57 | 1.00 | 1.11 | 0.00 | 0.00 | 6.00 | 6.00 | 2.07 | 6.18 | 0.01 |
| dist | 13 | 1555 | 2.12 | 1.35 | 1.75 | 1.97 | 1.19 | 0.12 | 8.44 | 8.32 | 1.02 | 0.80 | 0.03 |
| metro_dist | 14 | 1555 | 0.43 | 0.28 | 0.37 | 0.40 | 0.21 | 0.01 | 2.40 | 2.39 | 1.63 | 4.27 | 0.01 |
| attr_index | 15 | 1555 | 464.37 | 268.32 | 389.20 | 431.76 | 214.34 | 93.82 | 2934.13 | 2840.31 | 2.21 | 9.98 | 6.80 |
| attr_index_norm | 16 | 1555 | 15.83 | 9.14 | 13.26 | 14.71 | 7.31 | 3.20 | 100.00 | 96.80 | 2.21 | 9.98 | 0.23 |
| rest_index | 17 | 1555 | 877.66 | 461.29 | 801.81 | 839.64 | 494.72 | 159.84 | 4542.75 | 4382.92 | 0.94 | 2.17 | 11.70 |
| rest_index_norm | 18 | 1555 | 19.32 | 10.15 | 17.65 | 18.48 | 10.89 | 3.52 | 100.00 | 96.48 | 0.94 | 2.17 | 0.26 |
| lng | 19 | 1555 | 2.17 | 0.02 | 2.17 | 2.17 | 0.02 | 2.11 | 2.23 | 0.12 | 0.06 | 0.14 | 0.00 |
| lat | 20 | 1555 | 41.39 | 0.02 | 41.39 | 41.39 | 0.02 | 41.35 | 41.46 | 0.11 | 0.71 | 0.58 | 0.00 |

The dataset's summary statistics lead to the following significant conclusions:

- Each listing corresponds to the index or identifier range of the variable "realSum", which is from 0 to 1,554. The standard deviation is around 321.18, while the mean value is almost 288.39.
- The variable "room_type" indicates the kind of lodging, with a mean value of 1.77. Regarding the meaning of each category, further information would need to be given.

- The variable "person_capacity" indicates how many people the listings can accommodate in total. The average person capacity is around 2.76, with a standard variation of 1.28.
- "Cleanliness_rating" and "guest_satisfaction_overall" are ratings for cleanliness and overall guest satisfaction, respectively. Average ratings for cleanliness are 9.29 and 90.93 respectively for guest satisfaction.
- The variable "Bedrooms" indicates the number of bedrooms that are available. The average number of bedrooms is around 1.22, with a standard deviation of 0.57.
- The variables "dist" and "metro_dist" describe the distances from the accommodations to particular locations or metro stations. The average distances are 2.12 and 0.43 miles, respectively.

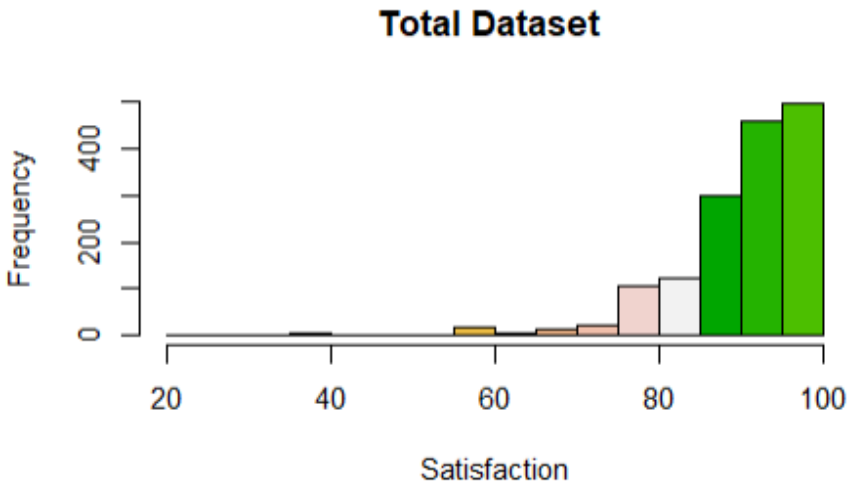
Other terms like "attr_index", "attr_index_norm", "rest_index", and "rest_index_norm" most likely refer to indices or normalized indices related to attractions or amenities. More understanding of these factors is needed to provide more in-depth insights. These summary statistics give a quick overview of the dataset, but a comprehensive study would need further research, data cleansing, and visualization to uncover more noteworthy patterns and connections.

# Analysis Methods

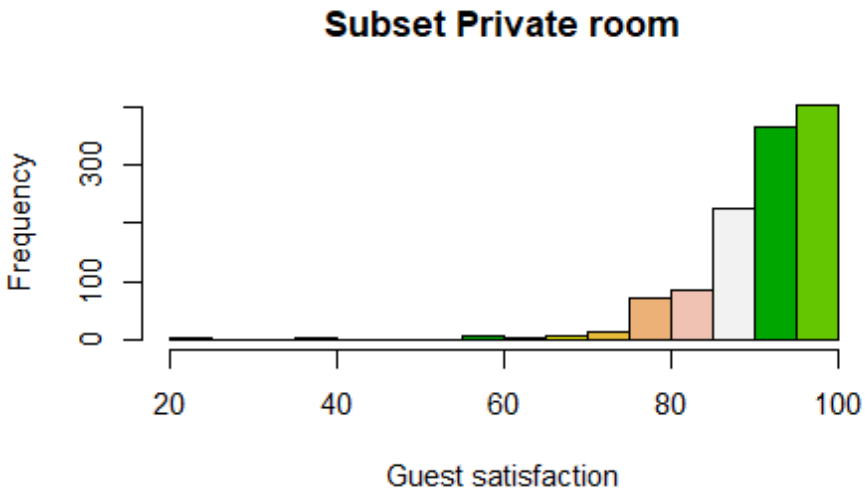
### Subset Analysis:

```
#subset analysis ----
subset_analysis<-subset(barcelona_weekdays,room_private ==
"TRUE")
subset_analysis_1 <-subset(barcelona_weekdays,room_private ==
"TRUE" & host_is_superhost =="TRUE")
subset_analysis_2 <-subset(barcelona_weekdays,room_private ==
"TRUE" & host_is_superhost =="FALSE" )
```

The first subset subset_analysis is created by subsetting the original dataset when the value of the room_private variable is "TRUE". Only the rows in this subgroup have private rooms listed as their room type. The second subset, subset_analysis_1, is created when both the host_is_superhost and the room_private variables are set to "TRUE". The only rows with a private room type and a superhost host are included in this subset. By subsetting the original dataset with the values "TRUE" for the room_private variable and "FALSE" for the host_is_superhost variable, the third subset, "subset_analysis_2," is created. Only the rows with a private room type and a non-superhost host are contained in this subset.

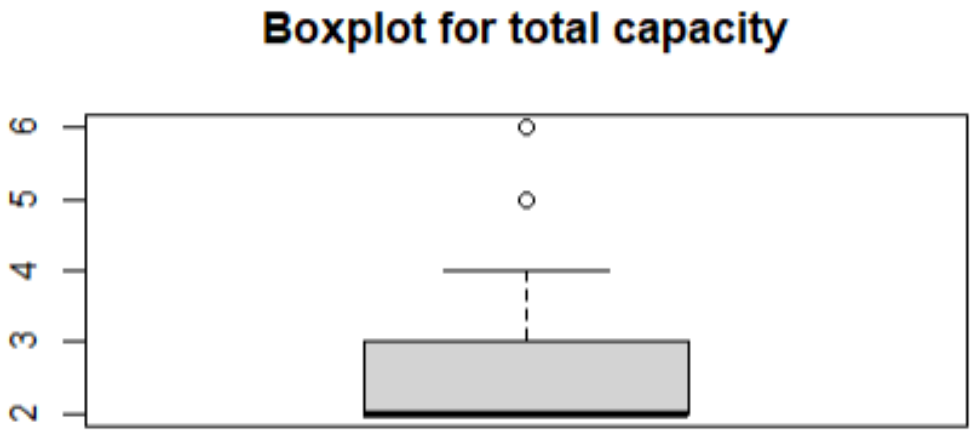


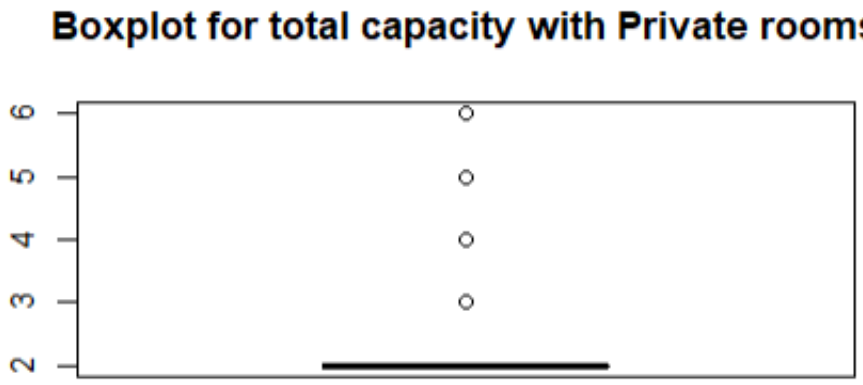

The first histogram displays the distribution of customer satisfaction scores throughout the full dataset. The x-axis displays the satisfaction rating, while the y-axis depicts the frequency or number of listings with that rating. The color scheme for the bars is terrain.colors(13), which provides a variety of distinctive tones. The "Satisfaction" x-axis and the "Frequency" y-axis are both displayed. "Total Dataset" is the main heading of the histogram.




The distribution of reviews from guests for a sample of private room listings is shown in the second histogram. Both the y-axis and the x-axis display the frequency of listings with that rating and the level of visitor satisfaction, respectively. In contrast to the first histogram, this histogram's bars are colored using the more diverse terrain.colors(7) color scheme. The y-axis is labeled "Frequency," the x-axis is "Guest satisfaction," and the major title is "Subset Private room." Looking at the histograms will reveal further information about the distribution of the guest satisfaction ratings throughout the overall dataset as well as the subset of private room listings. The distributions of

the full dataset and the private room subset may be compared, along with any significant peaks or trends, and the concentration of ratings in different satisfaction bands can be displayed.

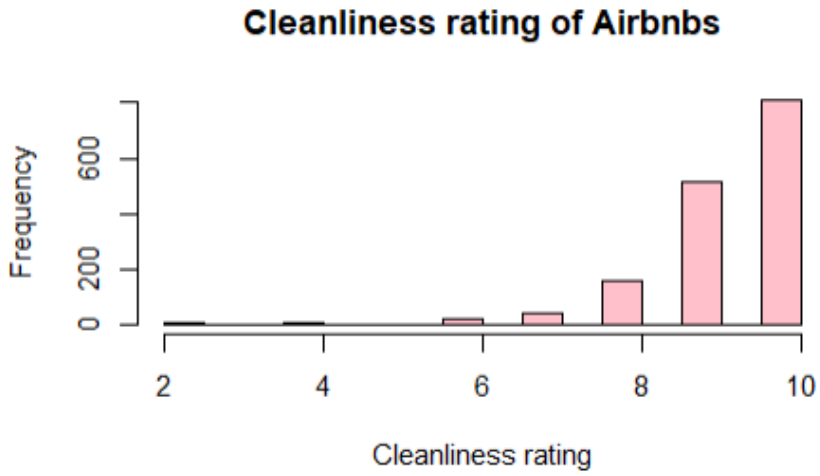


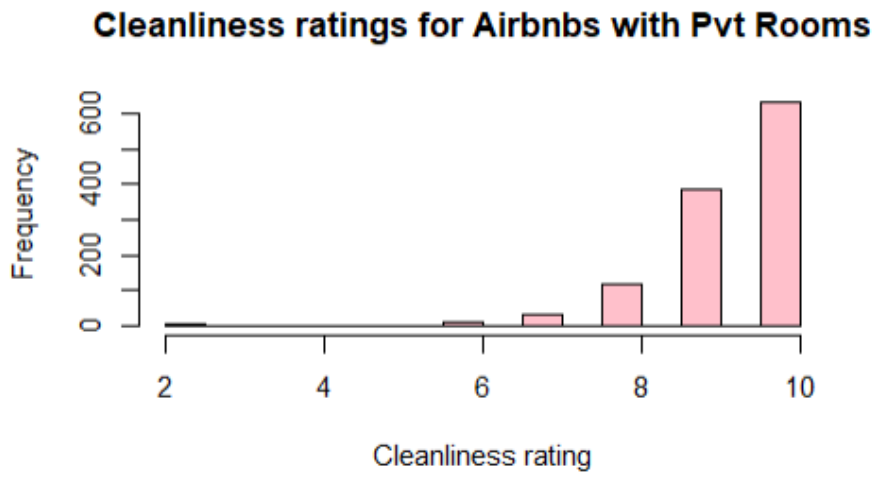




The initial boxplot depicts the distribution of individuals over the whole dataset. The median value is displayed as a horizontal line inside the box that contains the interquartile range (IQR), which is represented as a box. The whiskers extend outward from the box to the values that are 1.5 times the IQR apart on the lowest and highest axis. Any data points outside of this range are flagged as outliers and shown individually. The main title of the boxplot is "Boxplot for total capacity."

The second boxplot shows the distribution of person capacity, particularly for the subgroup of private room listings. This box shows the interquartile range, the horizontal line inside it the median, and the whiskers the range within 1.5 times the interquartile range. Any outliers are shown as individual points. The headline for this boxplot reads, "Boxplot for total capacity with Private rooms."

These boxplots help explain how person capacity is distributed across the overall dataset and, more specifically, across the subset of private room listings. You may view the distribution of the data, the median value, and any potential outliers. Comparing the two boxplots will allow you to see if there are any differences in the distribution of person capacity between the total dataset and the subset of private rooms.



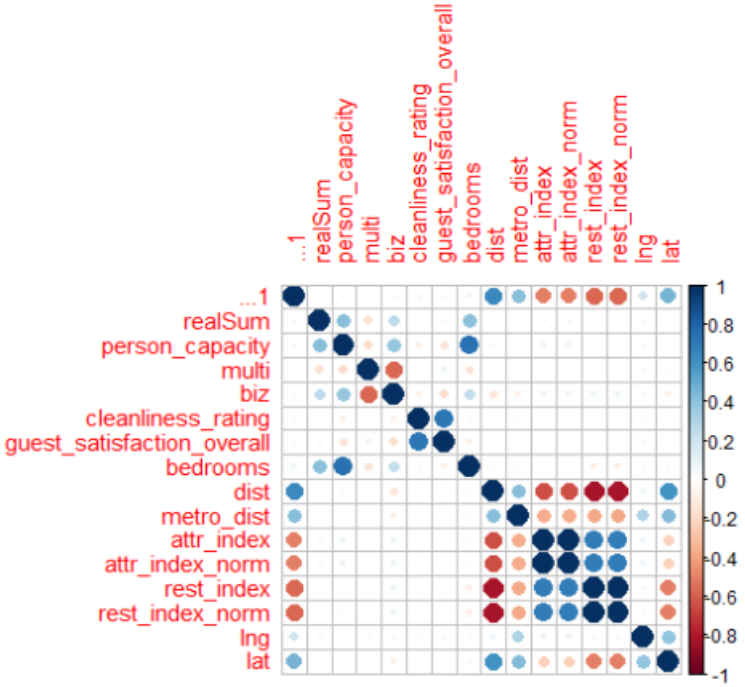




The distribution of cleanliness ratings for all Airbnbs in the dataset barcelona_weekdays is depicted in the first histogram. The number of Airbnbs with a particular cleanliness rating is indicated on the y-axis, and the frequency of those accommodations is shown on the x-axis. There are pink bars in the histogram. The histogram's subtitle is "Cleanliness rating of Airbnbs."

The distribution of cleanliness ratings for the subset of Airbnbs with private rooms is clearly shown in the second histogram (subset_analysis). It focuses on the cleanliness ratings for Airbnbs that have been designated as private rooms. The frequency or number of Airbnbs with a particular cleanliness rating is displayed on the y-axis, and the cleanliness rating is represented on the x-axis. Like the last histogram, this one also has pink bars. This histogram's heading is "Cleanliness ratings for Airbnbs with Pvt Rooms."

These histograms show the distribution and frequency of cleanliness ratings for the entire dataset as well as specifically for the subset of Airbnbs with private rooms. One may view the range of cleanliness ratings, the most popular ratings, and the overall rating distribution. Comparing the two histograms will show you if there are any differences in the cleanliness ratings between the total dataset and the subset of private room listings.

**Correlation table:**




A correlation plot is the name for this graphic representation of the correlation matrix. The correlation between two variables is depicted in each cell of a correlation plot. The intensity and color of the cell reflect the strength and direction of the relationship. Positive correlations are frequently represented by blue hues, while negative correlations are frequently illustrated by red hues. The size of the circles in the cells reveals how extensive the link is.

Looking at the correlation plot can help you understand the relationships between the different numerical variables in the dataset. Dark blue tones, which denote significant positive correlations, demonstrate a favourable association between the variables. Strong negative associations are indicated by the use of dark red colours. The correlation map, which identifies variables that are heavily related or have significant associations, can help with understanding the dataset and exploring potential patterns or links between variables.

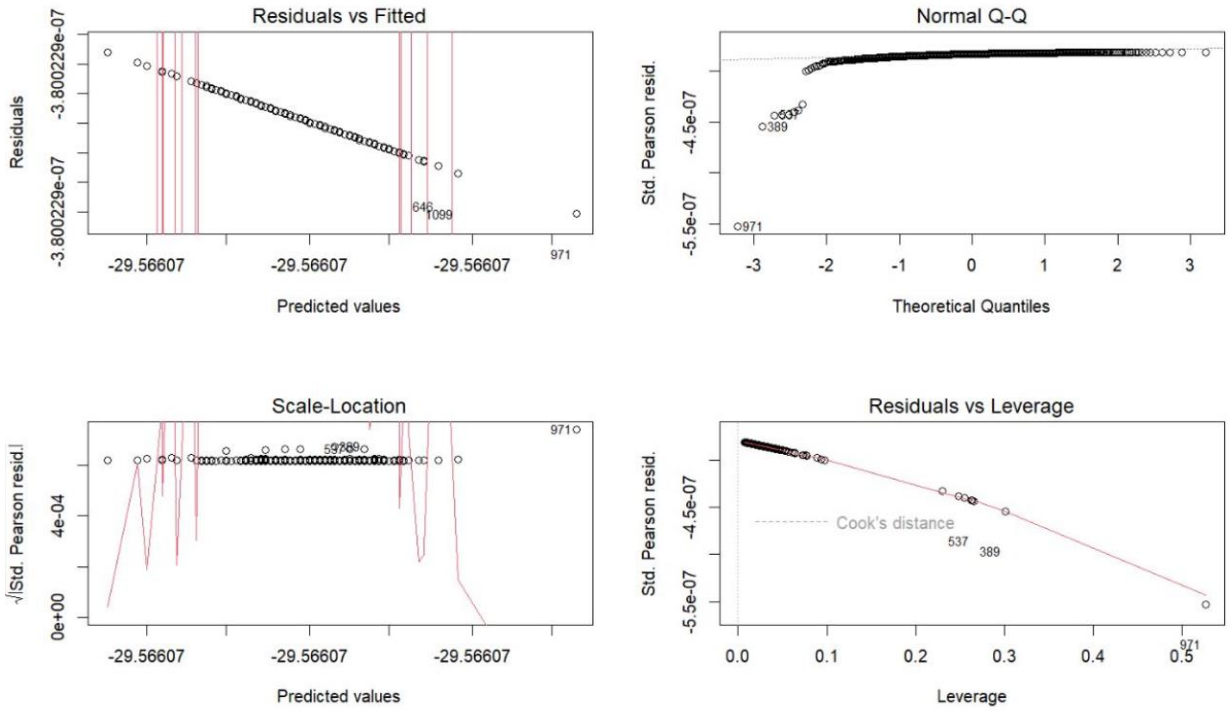
## Generalised Linear Model:

Inorder to simulate the link between a group of predictor variables and a response variable, statisticians use generalized linear models (GLMs). The linear regression model provides a flexible framework for figuring out complex relationships and making predictions by extending its capabilities to encompass multiple response variables, such as binary, count, or categorical data.

| | Estimate | Std. Error | t value |
|---|---|---|---|
| (Intercept) | 411.79362 | 40.007717 | 10.292855 |
| room_typePrivate room | -289.03513 | 25.107486 | -11.511910 |
| room_typeShared room | -395.66637 | 100.739904 | -3.927603 |
| person_capacity | 35.88005 | 8.307799 | 4.318840 |

| | Pr(>\|t\|) |
|---|---|
| (Intercept) | 4.427294e-24 |
| room_typePrivate room | 1.744494e-29 |
| room_typeShared room | 8.956163e-05 |
| person_capacity | 1.668004e-05 |

The result shows the coefficients, standard errors, t-values, and p-values for a multiple linear regression model with three predictor variables: room type (with Private room as the reference category), room capacity, and shared room. The intercept term denotes the predicted value of the response variable (price) when all predictor variables are equal to zero. The coefficients for each predictor variable show the projected change in the response variable (price) associated with a one-unit increase in each predictor variable, holding all other predictors constant. The p-value of each predictor variable indicates how relevant it is, with lower p-values denoting more significance. Each predictor variable's significant influence on the response variable is shown by the model. Private rooms, when compared to the reference group, have a negative impact on the price, whereas shared rooms have a large negative impact. However, the number of guests has a favourable effect on the price.




The resulting plot serves as a diagnostic plot for the logistic regression model (GLM). It aids in measuring the model's effectiveness and identifying any potential problems or model assumption breaches. A number of panels make up the plot:

- **Residuals vs. Fitted:** This panel displays the correlation between the model's predicted probability and the related residuals. It aids in finding patterns or nonlinearity in the data.

- **Normal Q-Q:** This panel evaluates the residuals' supposed normalcy. A relatively straight line denotes a reasonable adherence to the assumption.

- **Scale-Location:** This panel looks at the spread (variance) of the residuals in comparison to the values that were fitted. The ideal distribution of points along a horizontal line should show constant variance.

- **Residuals vs. Leverage:** This panel aids in locating noteworthy observations or outliers that could have a major effect on the model's coefficients.

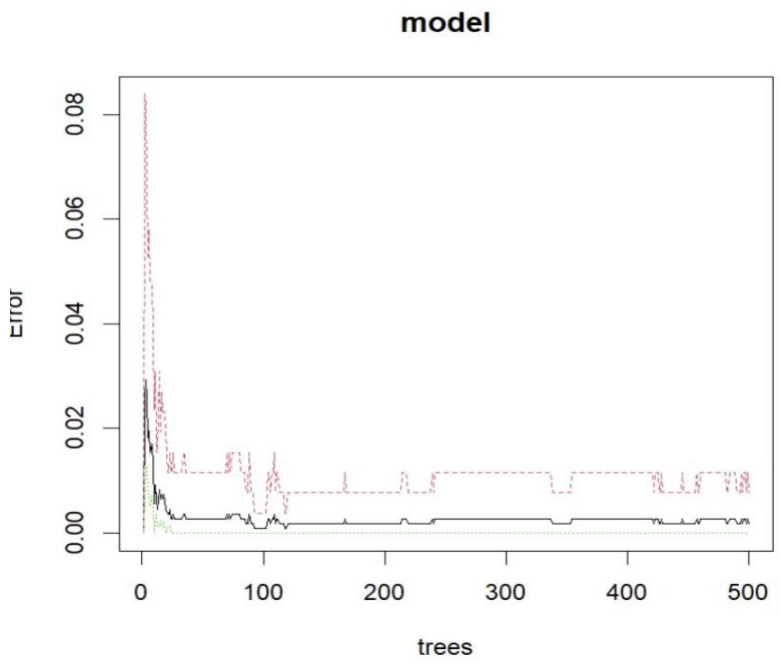
  We can assess the effectiveness of the logistic regression model and spot any potential problems, such as heteroscedasticity, nonlinearity, or significant observations, by examining these panels.

## Random Forest Regression:

A strong prediction model is created by integrating different decision trees with the use of the machine learning technique known as random forest regression. Support is provided for both continuous and categorical answer variables. The response variable in this code, room_private, is predicted using additional dataset variables. The randomForest() function builds an ensemble of decision trees by training each tree on a unique random subset of the training data. This ensemble strategy reduces overfitting and improves the model's ability to generalize. The ntree parameter displays the number of trees in the forest. After the model has been trained, predictions on the testing dataset are generated using the predict() method. These forecasts can then be evaluated or given further scrutiny.

```
Call:
 randomForest(formula = as.factor(room_private) ~ ., data = training,
ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 0.09%
Confusion matrix:
      FALSE TRUE class.error
FALSE   258    1 0.003861004
TRUE      0  830 0.000000000
> |
```




The estimated error rate according to the out-of-bag (OOB) method is 0.09%, which is quite accurate. A class error rate of 0.0039% was achieved with 258 of the 259 occurrences labelled as FALSE being correctly identified, according to the confusion matrix. In a similar vein, all 830 occurrences marked as TRUE were correctly categorised, yielding a class error rate of 0.

# Conclusion:

Random forest regression, subset analysis, and generalized linear models (GLM) are statistical modelling techniques used to predict outcomes or understand correlations in data.

Random forest regression, an ensemble learning technique, blends many decision trees to produce predictions. It is a strong, flexible algorithm that can handle complex links and interactions in data. The relative significance of each feature can be determined using random forest regression on high-dimensional datasets.

Subset analysis, also known as variable selection or feature selection, involves selecting relevant variables from a larger pool of potential predictors. This approach aims to improve model performance by reducing the number of dimensions in the data and focusing on the most informative characteristics. Subset analysis can be performed using a number of methods, such as stepwise regression, backward elimination, and forward selection.

The generalized linear model (GLM) is a framework that extends the traditional linear regression model to allow different response variables, including binary, count, or categorical outcomes. When modelling the relationship between predictors and the projected response value using a link function, it makes an explicit assumption about the probability distribution of the response variable. Non-linear relationships can be incorporated into GLM as well as other data distributions.

Finally, whereas random forest regression is a flexible ensemble method that aids in the selection of relevant predictors, GLM extends linear regression for a variety of response types. Any technique may be utilized since it has benefits depending on the analysis's requirements and the characteristics of the dataset.

# References


1. Airbnb Prices in European Cities. (2023, February 20). Kaggle. https://www.kaggle.com/datasets/thedevastator/airbnb-prices-in-european-cities?resource=download&select=barcelona_weekdays.csv
2. Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.
3. McCullagh, P., & Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.
4. Bluman, A. (2017). Elementary Statistics: A Step By Step Approach (10th ed.). McGraw Hill.
5. Kabacoff, R. (2022). R in Action, Third Edition: Data Analysis and Graphics with R and Tidyverse. Manning.