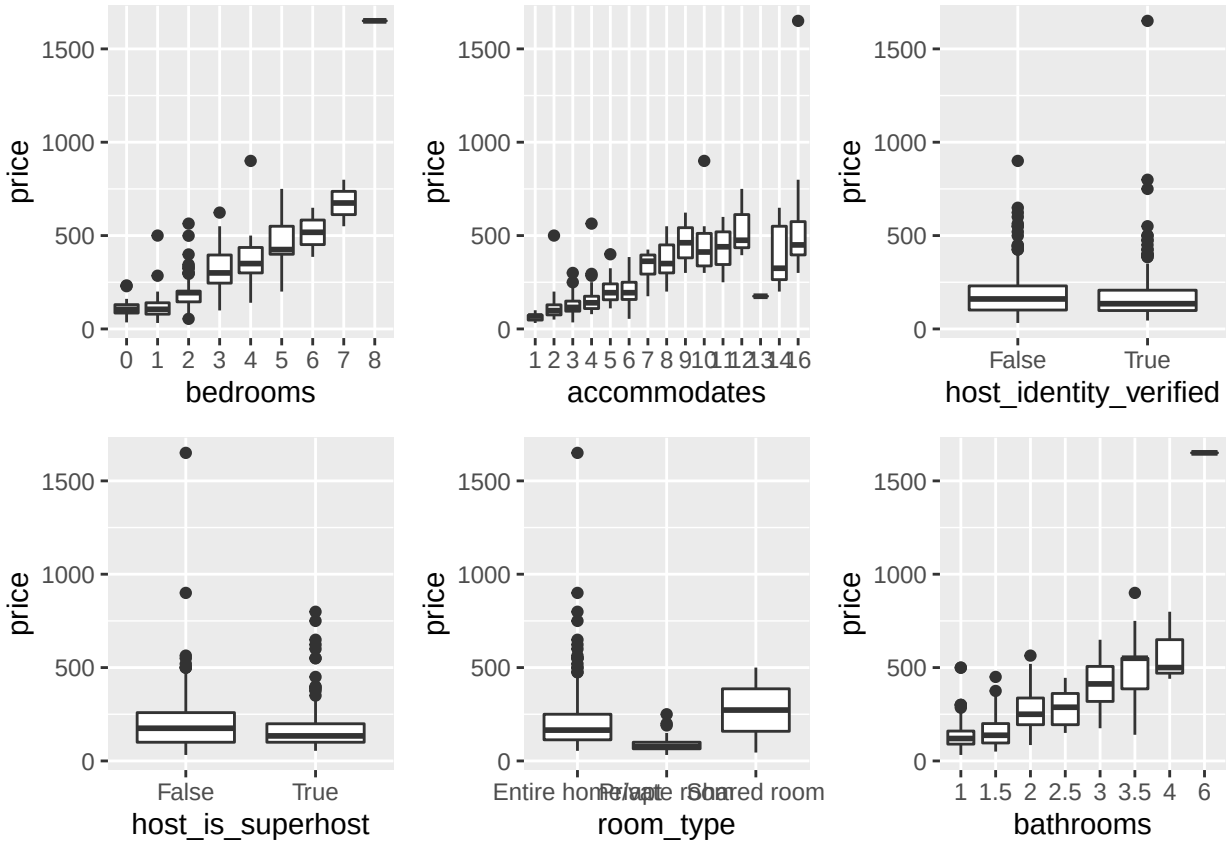
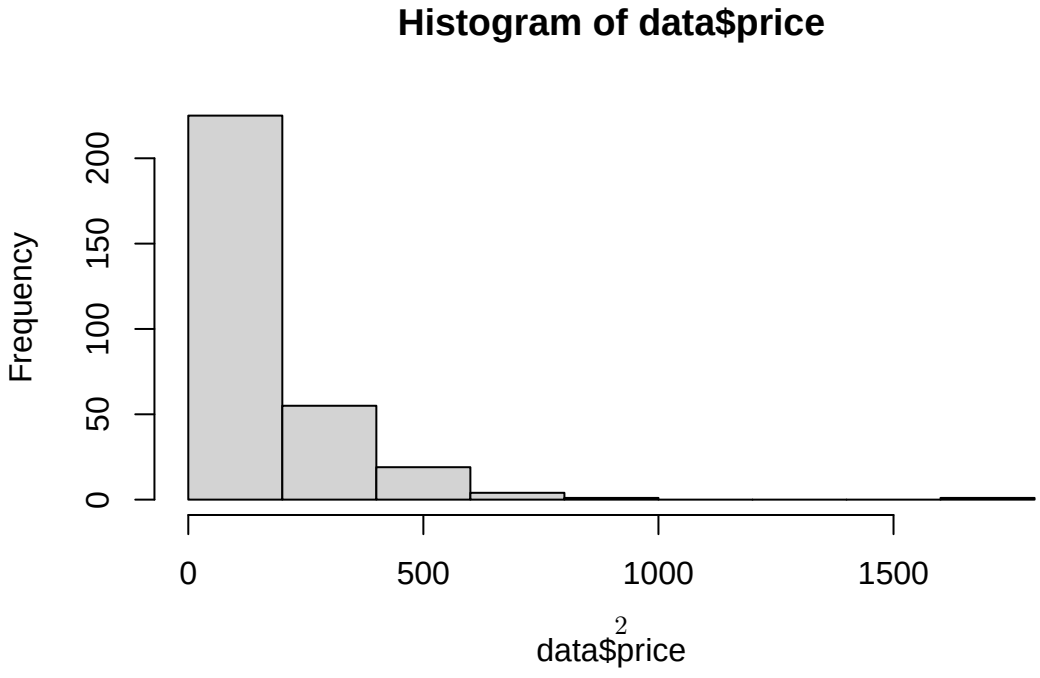
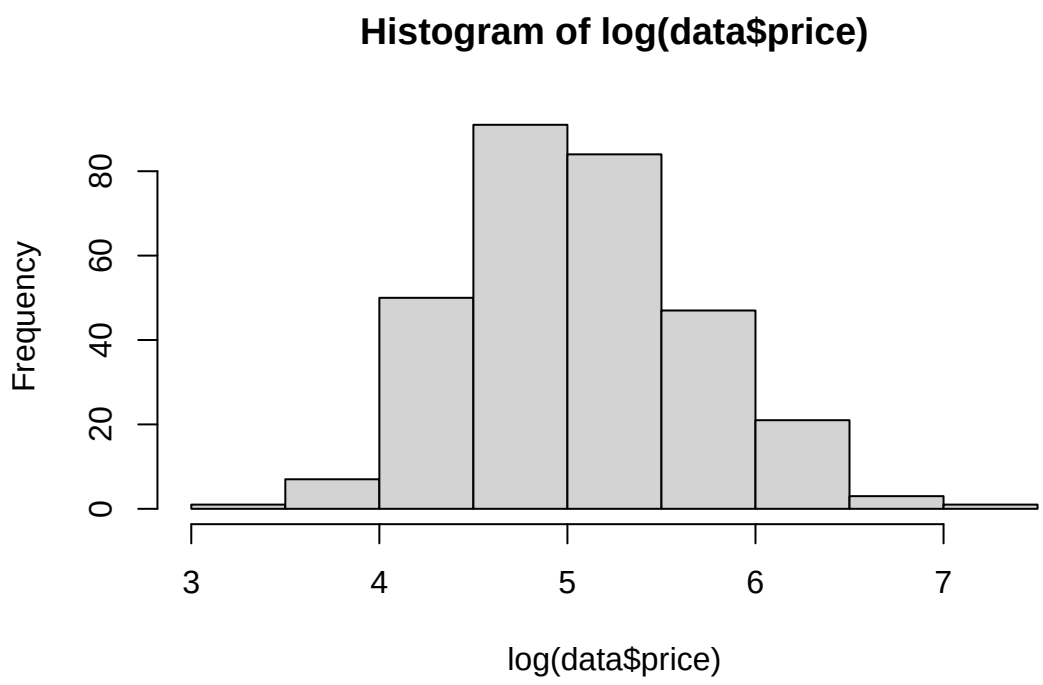
# Question 3

## (A)

**Analyze the data using host_is_superhost,host_identity_verified, room_type, accommodates, bathrooms and bedrooms as predictors. You should start by doing EDA, then model fitting, and model assessment. You should consider transformations if needed.**






log(data$Price) follows the normal distribution and will use this as the response variable for the regression model.

# (B)

**Include the output from the final regression model that you used, as well as evidence that the model fits the assumptions reasonably well. Your regression output should includes a table with coefficients and SEs, and p-values or confidence intervals.**

**Regression Model Output**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **(Intercept)** | 4.265 | 0.1425 | 29.93 | 4.871e-88 |
| **bedrooms1** | 0.07419 | 0.07686 | 0.9653 | 0.3352 |
| **bedrooms2** | 0.2421 | 0.09789 | 2.473 | 0.014 |
| **bedrooms3** | 0.3619 | 0.1384 | 2.615 | 0.009419 |
| **bedrooms4** | 0.2 | 0.1766 | 1.133 | 0.2582 |
| **bedrooms5** | 0.4047 | 0.2456 | 1.647 | 0.1006 |
| **bedrooms6** | 0.6794 | 0.3578 | 1.899 | 0.05865 |
| **bedrooms7** | 0.7685 | 0.3642 | 2.11 | 0.03574 |
| **bedrooms8** | 2.257 | 0.4338 | 5.203 | 3.872e-07 |
| **accommodates2** | 0.4071 | 0.1194 | 3.408 | 0.0007538 |
| **accommodates3** | 0.4467 | 0.1413 | 3.162 | 0.001743 |
| **accommodates4** | 0.5679 | 0.1311 | 4.332 | 2.077e-05 |
| **accommodates5** | 0.7461 | 0.1579 | 4.724 | 3.706e-06 |
| **accommodates6** | 0.571 | 0.1512 | 3.776 | 0.0001959 |

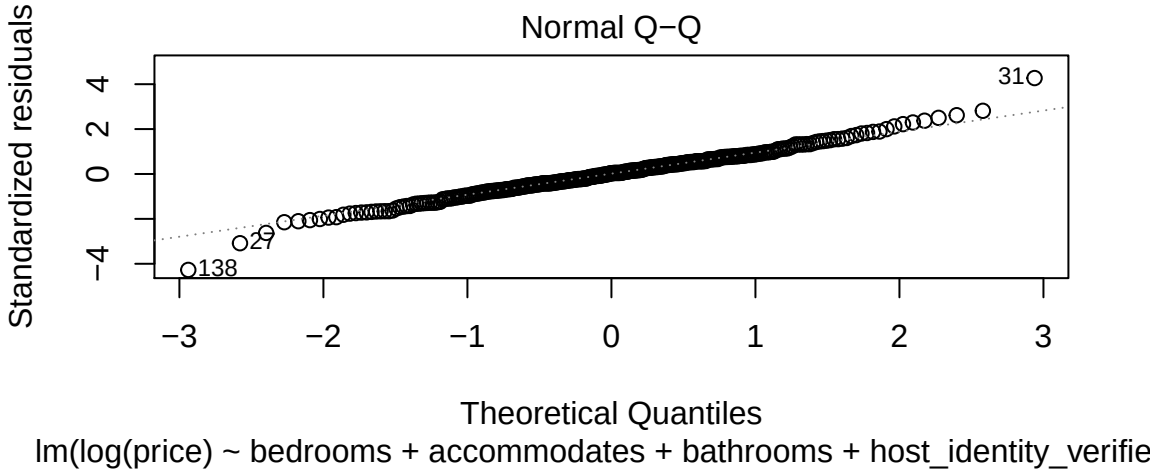
| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| **accommodates7** | 1.021 | 0.1938 | 5.271 | 2.768e-07 |
| **accommodates8** | 1.024 | 0.1913 | 5.353 | 1.843e-07 |
| **accommodates9** | 1.361 | 0.3015 | 4.515 | 9.444e-06 |
| **accommodates10** | 1.274 | 0.2138 | 5.957 | 7.911e-09 |
| **accommodates11** | 0.9601 | 0.3051 | 3.147 | 0.001833 |
| **accommodates12** | 1.418 | 0.2813 | 5.042 | 8.441e-07 |
| **accommodates13** | 0.4317 | 0.3954 | 1.092 | 0.2758 |
| **accommodates14** | 0.8269 | 0.259 | 3.193 | 0.001573 |
| **accommodates16** | 0.9477 | 0.2877 | 3.294 | 0.001119 |
| **bathrooms1.5** | 0.1412 | 0.07508 | 1.881 | 0.06101 |
| **bathrooms2** | 0.3293 | 0.07619 | 4.322 | 2.173e-05 |
| **bathrooms2.5** | 0.2012 | 0.109 | 1.846 | 0.06599 |
| **bathrooms3** | 0.3778 | 0.1441 | 2.621 | 0.009257 |
| **bathrooms3.5** | 0.4439 | 0.1637 | 2.712 | 0.007108 |
| **bathrooms4** | 0.7616 | 0.3066 | 2.484 | 0.0136 |
| **host_identity_verifiedTrue** | -0.06077 | 0.04338 | -1.401 | 0.1624 |
| **host_is_superhostTrue** | 0.00398 | 0.04288 | 0.09283 | 0.9261 |
| **room_typePrivate room** | -0.3169 | 0.0688 | -4.606 | 6.326e-06 |
| **room_typeShared room** | 0.5292 | 0.2548 | 2.077 | 0.03877 |

Table 2: Fitting linear model: log(price) ~ bedrooms + accommodates + bathrooms + host_identity_verified + host_is_superhost + room_type

| Observations | Residual Std. Error | $R^2$ | Adjusted $R^2$ |
|---|---|---|---|
| 305 | 0.3417 | 0.7346 | 0.7034 |

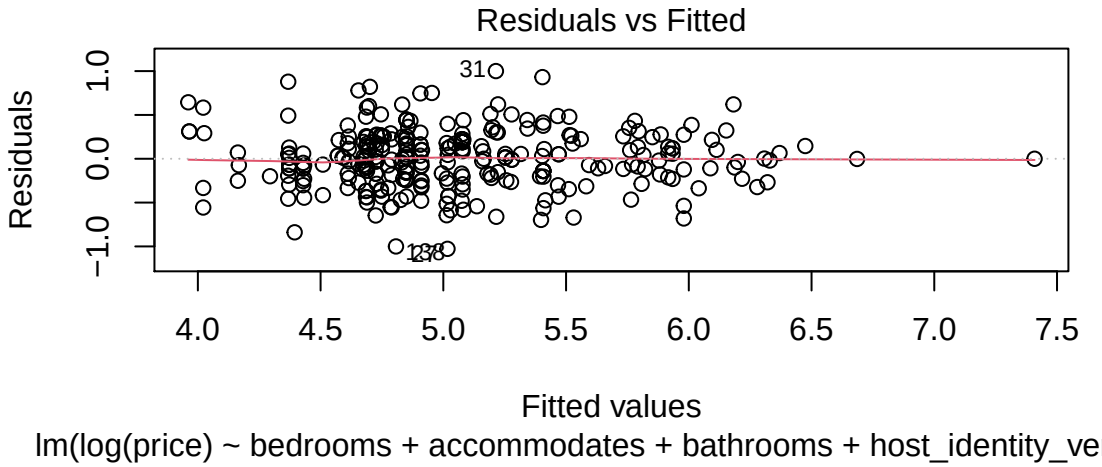
**Normality Assumption**

```
## Warning: not plotting observations with leverage one:
##   148, 217
```



As majority of the data points are on the 45 degree line, the model does not violate the normality assumption.

**Independence and Equal Variance**




As the points are random and spread throughout the X axis, the model does not violate the independence and equal variance assumption.

# (C)

### Interpret the results of your fitted model in the context of the data.

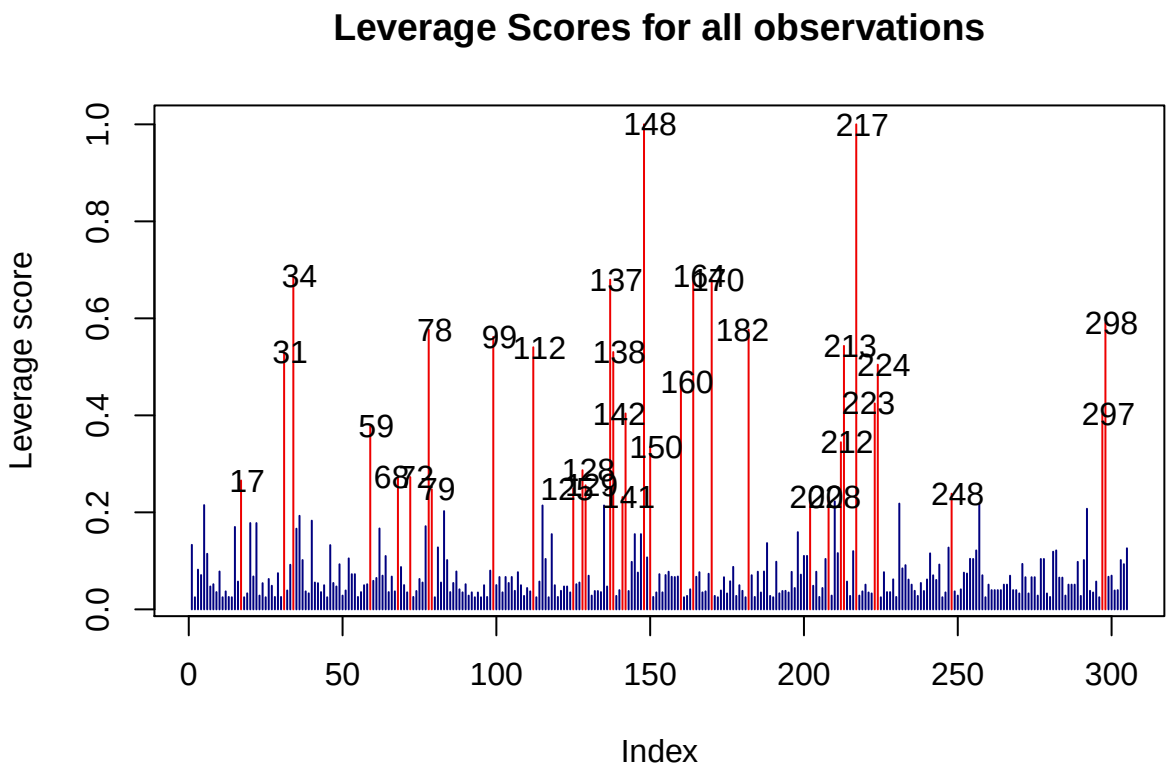
Intercept is 4.265 which means that if a house has no bedrooms, 0 accomodates, 0 bathrooms, host_identify_verified and host_is_superhost is False, the price of the house is $4.265. Among the various number of bedrooms(0-8), a 8th bedrooms house is the strongest predictor of the outcome (Abs(t value) = 5.203) Among the various accommodates (0-16), accommodates10 is the strongest predictors of the outcome (Abs(t value) = 5.957). Among the various number of bathrooms (1-6), a 2 bathroom house is the strongest predictor (Abs(t value) = 4.322). The percent of the variability in price explained by the regression model (R-squared) is 73.46%.

# (D)

### Are there any (potential) outliers, leverage points or influential points? Provide evidence to support your response. Also, if there are influential points and/or outliers, exclude the points, fit your model without them, and report the changes in your overall conclusions.

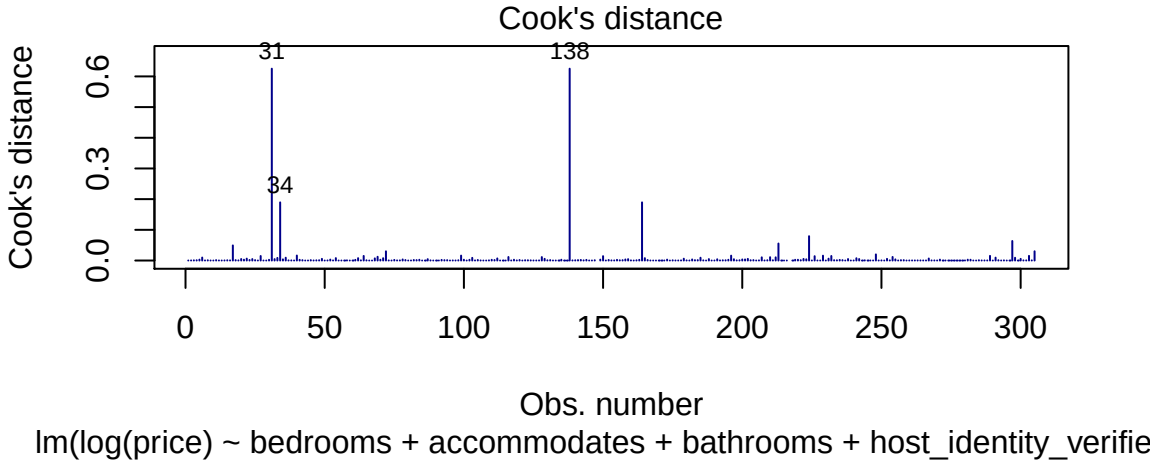
**Identifying Leverage Points with leverage score = 2(p+1)/n**

```
## Warning in c(1:n)[lev_scores > (2 * p/n)] + c(rep(2, 4), -2, 2): longer object
## length is not a multiple of shorter object length
```




There are many leverage points in the data.

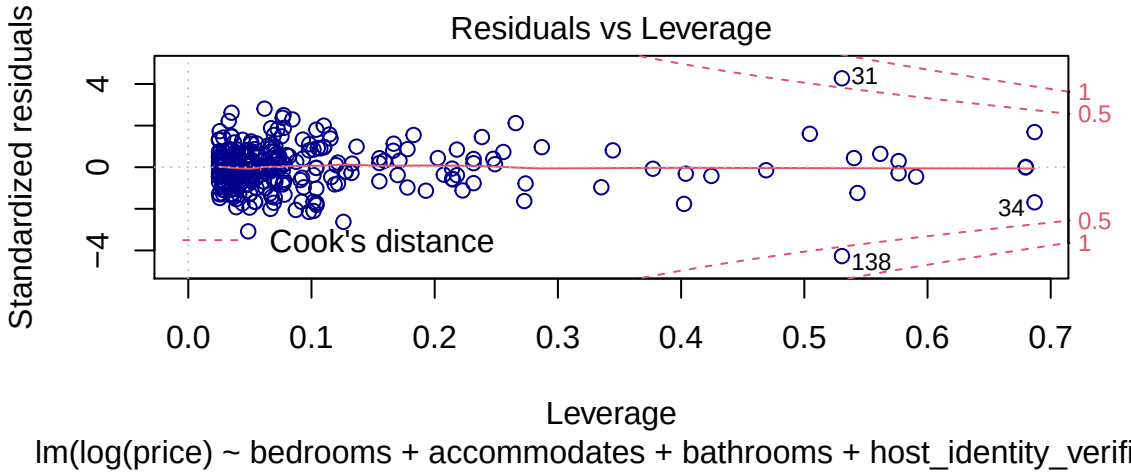
**Identifying Influence Points having Cook's distance>0.5**




Data Points 138 and 31 are high influence points.

**Identifying Outliers**

```
## Warning: not plotting observations with leverage one:
##   148, 217
```




Data Points 138 and 31 seem to also be outliers as their standardized residuals are at 4 and -4. Hence, points 138 and 31 are removed from the data and the updated data is fitted again.

Even after removing the high influence points and outliers from the data, the linear regression model does not change significantly. Hence, there is not much change between removing and not removing the high influence points.

- This
- is
- a test