Superhost Worth It?

Results of the study

This file provides users of the data set with the main findings of the research that has been done. The first part is about the structure of the data set, then some assumptions are tested. After that comes the analysis, which consists of two parts: the first part concludes that there is a significant difference in prices of super hosts and non-super hosts. The second part of the analysis results in the significant difference in prices when considering price classes. Hereby, the results show that the classes > $ 100 to $ 150 and > $ 250 see significant differences in prices between super hosts and non-super hosts.

Composition

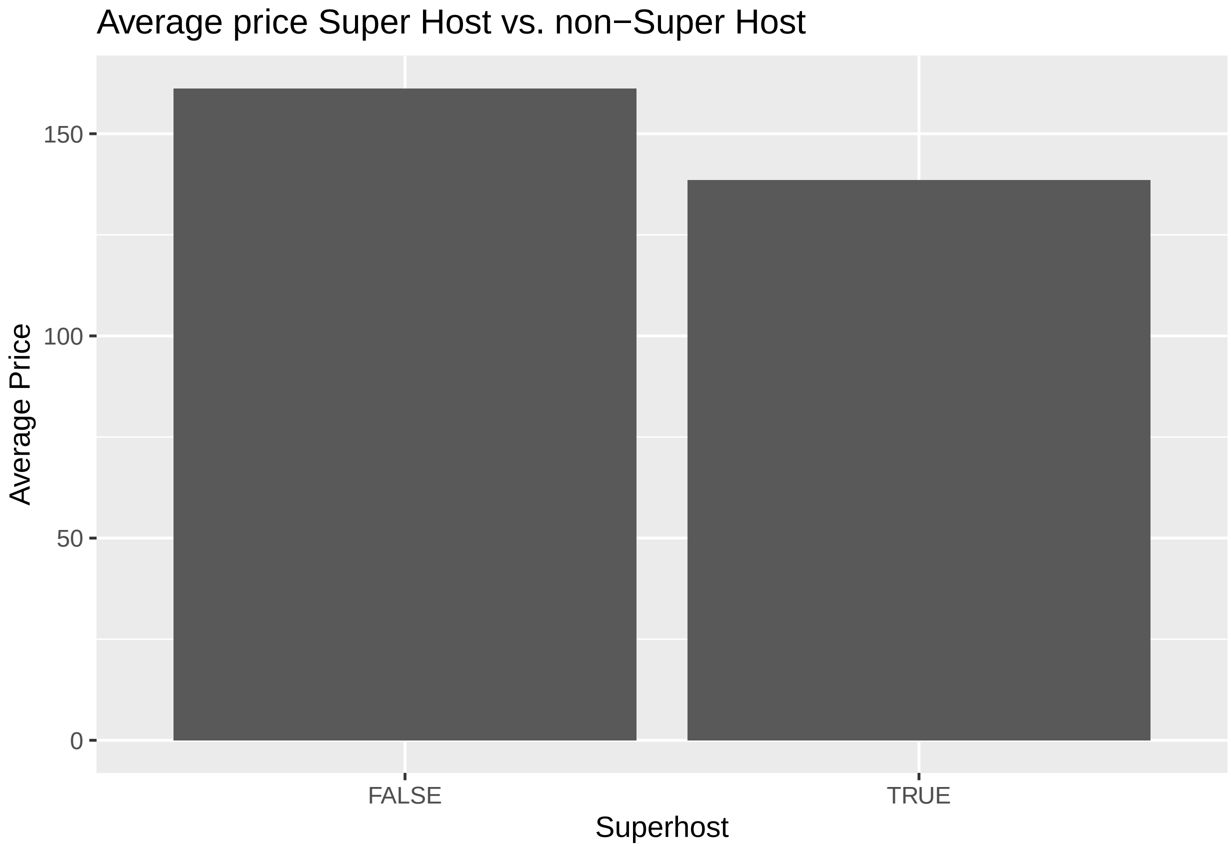
First of all, there will be provided some insights in the data set that is used. This data set has been cleaned from errors, NA's and abundant data. Underneath are summary statistics of different variables in the data set. A first notable insight is the difference between the minimum and maximum prices (min. is $9, max. is $9999). This is the reason for the difference in mean ($153.3156428) and the median ($108), where some strikingly high prices increase the average considerably.

Table 1: Summary data

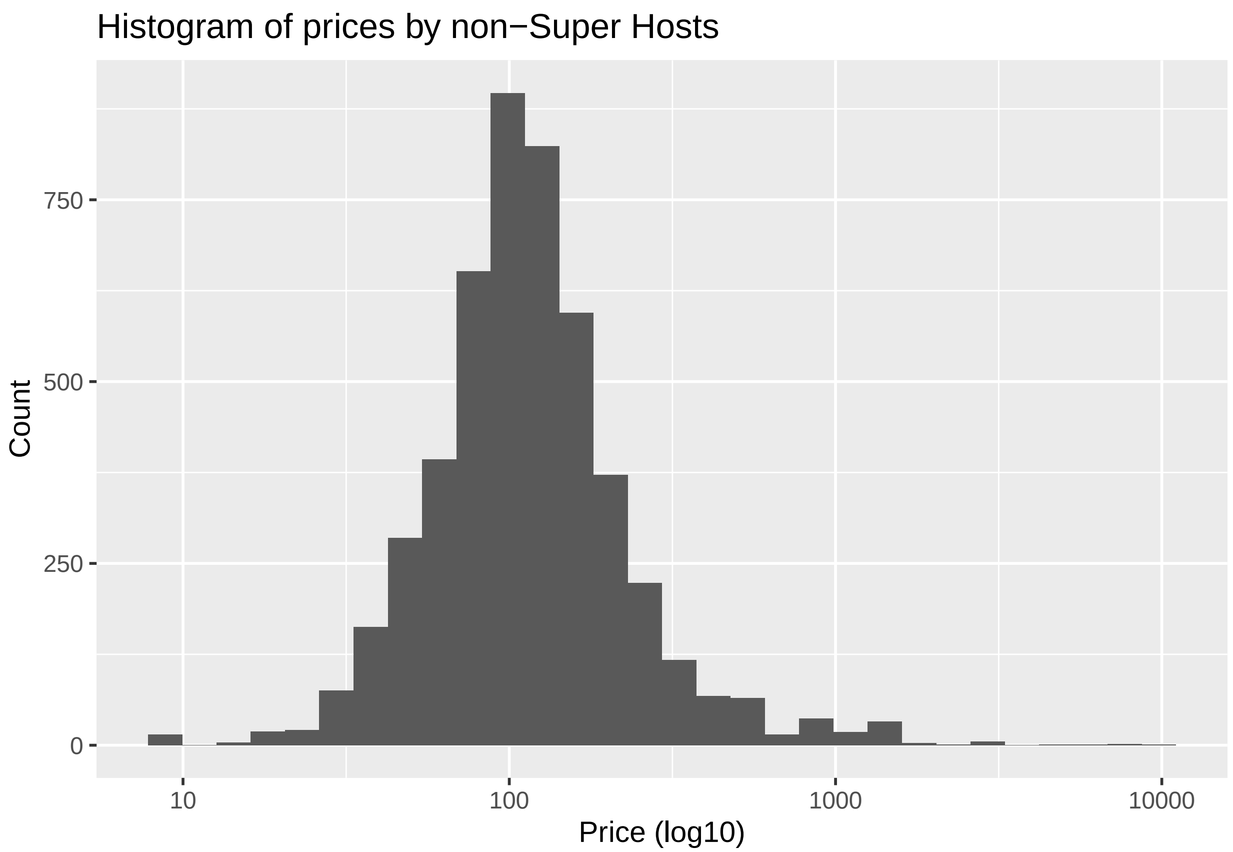
Superhost	Bedrooms	Beds	Number of reviews	Review scores	Reviews per month	Price
Mode :logical	Min. : 1.0	Length:7537	Min. :0.000	Min. : 0.010	Min. : 1.000	Min. : 1.0
FALSE:4905	1st Qu.: 1.0	Class :character	1st Qu.:4.580	1st Qu.: 0.360	1st Qu.: 1.000	1st Qu.: 1.0
TRUE :2632	Median : 2.0	Mode :character	Median :4.800	Median : 1.030	Median : 1.000	Median : 2.0
NA	Mean : 2.5	NA	Mean :4.666	Mean : 1.514	Mean : 1.615	Mean : 2.5
NA	3rd Qu.: 3.0	NA	3rd Qu.:4.930	3rd Qu.: 2.160	3rd Qu.: 2.000	3rd Qu.: 3.0
NA	Max. :19.0	NA	Max. :5.000	Max. :17.720	Max. :14.000	Max. :19.0
NA	NA	NA	NA's :1393	NA's :1393	NA	NA

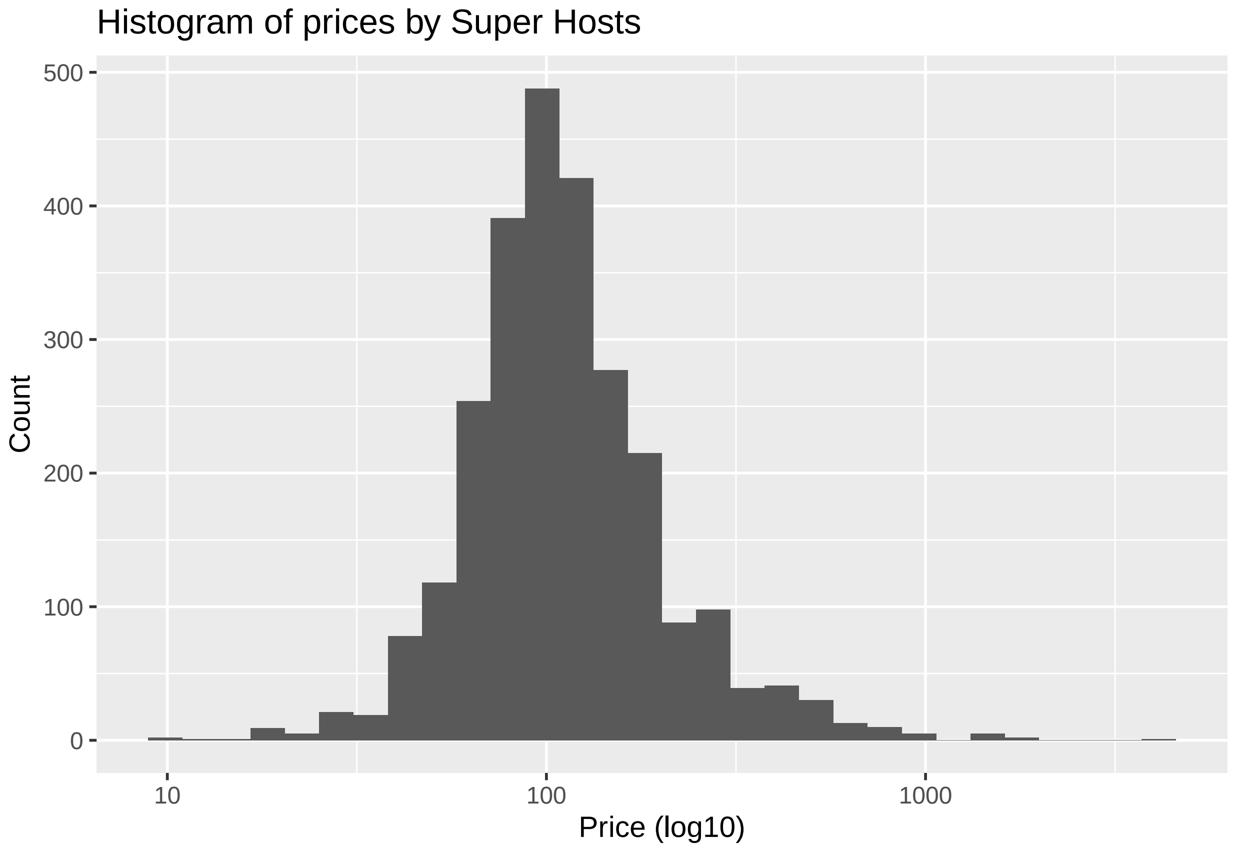
Average prices Super Host vs. non-Super Host

To get insight into the difference between prices by Super Hosts and non-Super Hosts, there's a box plot and two histograms provided underneath. Note that the x-axis at the histograms are in logarithmic scale, due to the range in price.



The findings are that the average price by a super host is $138.57 and the average price asked by a non-super host is $161.23. Meaning that on average hosts with the label non-super host ask a higher price for their listings. But as could be seen in the summary statistics, there is a considerable range in the prices. To get insight in this, there following histograms are provided:


Histogram of prices by non−Super Hosts
Count
750
500
250
0
10
100
1000
10000
Price (log10)



Analysis

First study

Levene’s test

To formally test the variance in price between superhosts and non-superhosts, a t-test will be performed. Hereby, we first need to check an assumption: whether the variances of the dependent variable (“price_numeric”) are equal across the two groups (super host vs non-super host) . This will be done by a LeveneTest.

```
## Levene’s Test for Homogeneity of Variance (center = mean)
##         Df F value    Pr(>F)    
## group    1  28.017 1.237e-07 ***
##       7535                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The null hypothesis is that the variances are equal and the findings are that this test has a p-value of 1.237e-07 , meaning that we can reject the null hypothesis. So we can assume that the variances are not equal across these two groups.

T-test

Now the actual t-test will be performed. Where the var.equal will be set to FALSE, as we assumed in the LeveneTest that the variances were not equal across the two groups.

```
##
##  Welch Two Sample t-test
##
## data:  price_numeric by superhost
## t = 4.2836, df = 7411.6, p-value = 1.862e-05
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to
## 95 percent confidence interval:
##  12.28790 33.02389
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            161.2273            138.5714
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 1.862e-05, meaning that the null hypothesis can be rejected. Indicating that we can assume that the means in these two groups are not similar. So we can say that the data shows some significant differences between being a super host and not being a super host.

Linear regression

Next the actual linear regression model will be estimated. This will give an insight in the relationship between the dependent variable, "price_numeric", and the independent variable, "superhost". Immediately after the model has been estimated, a summary will be asked to check how well the model fits the data.

```
##
## Call:
## lm(formula = price_numeric ~ superhost, data = df_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -152.2  -76.2  -42.6    3.8 9837.8
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    161.227      3.805  42.375  < 2e-16 ***
## superhostTRUE  -22.656      6.439  -3.519 0.000436 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 266.5 on 7535 degrees of freedom
## Multiple R-squared:  0.001641,   Adjusted R-squared:  0.001508
## F-statistic: 12.38 on 1 and 7535 DF,  p-value: 0.0004361
```

the summary shows the F-statistic of the model, which indicates that the model is significant (p-value < .05, 0.0004361).

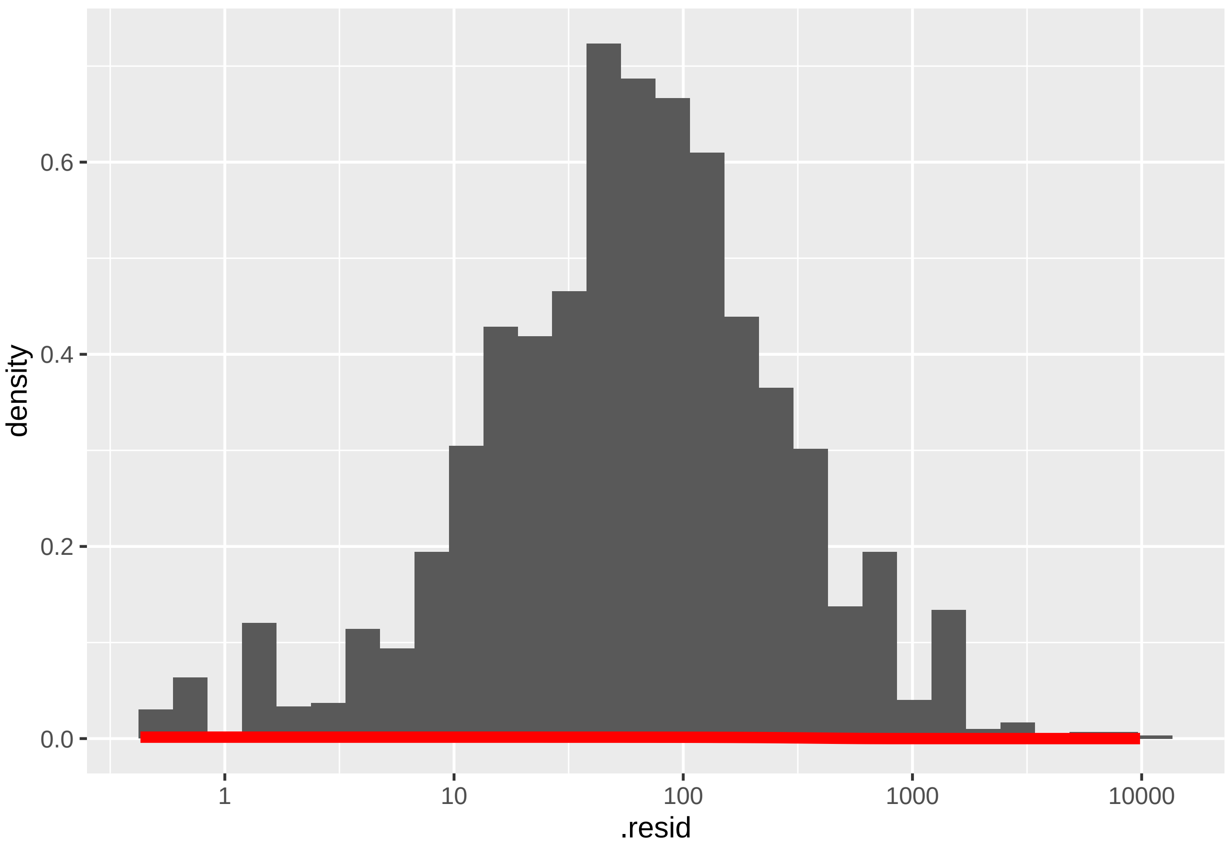
The coefficients brings us to the conclusion that whenever hosts are super hosts the price is on average $22.66 lower than hosts that are labelled as non-super hosts. This effect is clearly significant, as the p-value of the coefficient is clearly below 0.05 (p-value = 0.000436).

We also tried to extend the model to look for some interaction between the “superhost” variable and other variables: “number_of_reviews”, “review_scores_rating”, “reviews_per_month. Whenever estimating a model with the interaction between “superhost” and one of the other three named variables, the coefficient of the “superhost” turned out to be no significant anymore. Therefore, no valid conclusion can be drawn.

```
##
## Call:
## lm(formula = price_numeric ~ superhost * reviews_per_month, data = df_cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -138.7  -56.8  -27.0   13.9 8118.5
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   148.03924    3.47135  42.646  < 2e-16 ***
## superhostTRUE                   9.63077    6.03816   1.595    0.111
## reviews_per_month             -12.94908    1.92327  -6.733 1.81e-11 ***
## superhostTRUE:reviews_per_month  -0.04572    2.70246  -0.017    0.987
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 158.4 on 6140 degrees of freedom
##   (1393 observations deleted due to missingness)
## Multiple R-squared:  0.0148, Adjusted R-squared:  0.01432
## F-statistic: 30.75 on 3 and 6140 DF,  p-value: < 2.2e-16
```

Normality test

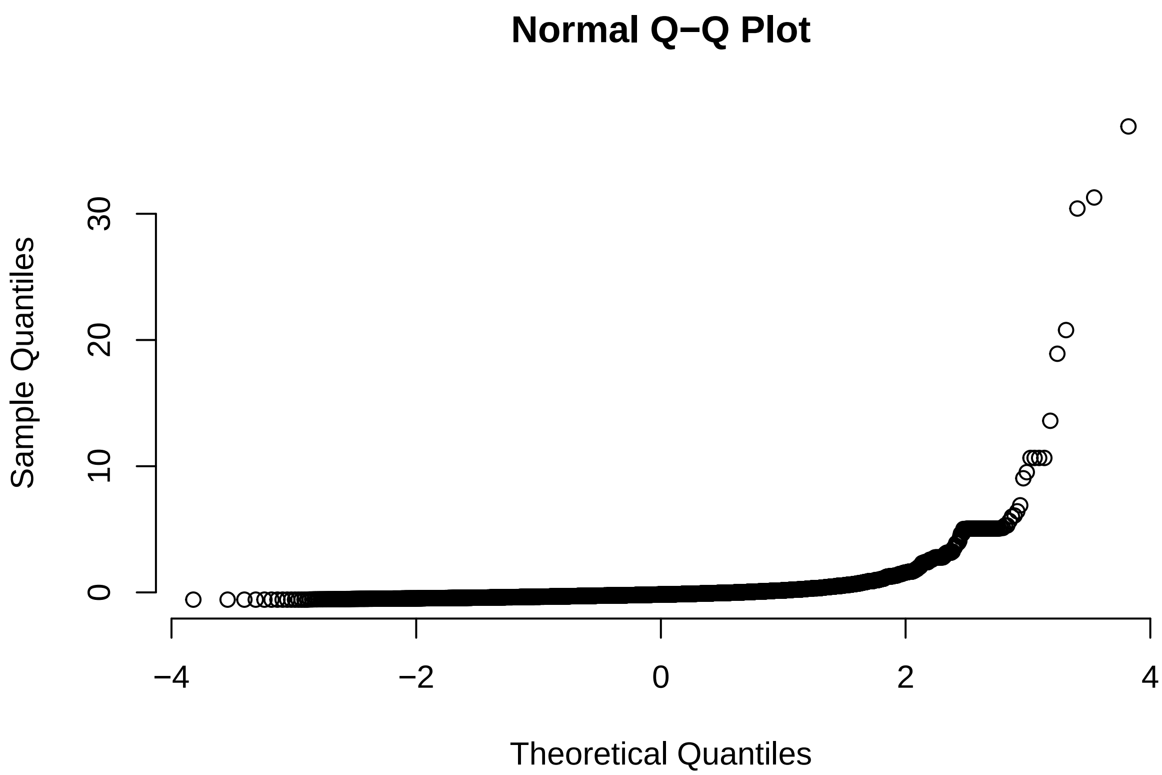
So we just continue with the first model, where we want to check on the assumption whether the data is normally distributed. To check for the normality assumption, the residuals should be distributed normally. Therefore, the residuals of model 1 have to be accessed.



The check for the normality, a histogram of the residuals will be created. The idea behind this is to look for a bell shaped curve. The findings were that there is a bell shaped curve, although it has a very long tail on the right hand side. This indicates that the data is right skewed.

The findings were that there is a bell shaped curve, although it has a very long tail on the right hand side. This indicates that the data is right skewed.

Another way to check for normality, is the normal QQ plot. The idea behind this plot is to search for a straight line that moves diagonal. Indicating that the data is normally distributed



The findings are consistent with the previous (histogram) plot, the data is right skewed. The plotted line is curved because of some extreme outliers.

So we found out that the distribution of the data is positively skewed, meaning that the most frequent values are low, with a tail towards the high values. Therefore, it could be considered to transform the data. However, transforming data can make the interpretation of the analysis much more difficult , when comparing the mean of the two groups after transforming the data it is not to say that there is a difference in the two groups means (Data Novia, 2021). Also, according to Data Novia, analyses like the F or t family of tests (i.e., independent and dependent sample t-tests, ANOVAs, MANOVAs, and regressions), violations of normality are not usually a death sentence for validity. With large enough sample sizes (> 30 or 40), there's a pretty good chance that the data will be normally distributed. In line of these findings we keep the data how it is.

Interim conclusion

There are significant differences in the asking price between the two groups (super host vs non-super host). Meaning that on average the price asked by a host with the label super host is \$22.66 lower.

Second study

Based the conclusion in the first stage of the analysis, we like to have a closer look on how this conclusion relates to more specific price classed. Therefore, the following price classes will be considered: > \$ 0 to \$ 50, > \$ 50 to \$ 100, > \$ 100 to \$ 150, > \$ 150 to \$ 200, > \$ 200 to \$ 250 and > \$ 250. Six price classes in total. We perform for every price class the same analysis as for the whole data set in the first stage. But now the intention is to see for which price classes there is a significance difference between the average asking price of a super host vs a non-super host.

Price class: : > $ 0 to $ 50

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1  0.0076 0.9304
##       709
```

```
##
##  Two Sample t-test
##
## data:  price_numeric by as.factor(superhost)
## t = -1.07, df = 709, p-value = 0.285
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -2.5960661  0.7645336
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            38.96402            39.87978
```

```
##
## Call:
## lm(formula = price_numeric ~ superhost, data = df_cleaned1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.880  -5.880   2.036   7.120  11.036
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    38.9640     0.4342   89.74   <2e-16 ***
## superhostTRUE   0.9158     0.8558    1.07    0.285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.977 on 709 degrees of freedom
## Multiple R-squared:  0.001612,   Adjusted R-squared:  0.0002041
## F-statistic: 1.145 on 1 and 709 DF,  p-value: 0.285
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 0.285, meaning that the null hypothesis cannot be rejected. Indicating that we can assume that the means in these two groups are similar. So we can say that the data does not show some significant differences between being a super host and not being a super host.

Price class: : > $ 50 to $ 100

```
## Levene's Test for Homogeneity of Variance (center = mean)
##         Df F value    Pr(>F)
## group    1  11.719 0.0006275 ***
##       2791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  Welch Two Sample t-test
##
## data:  price_numeric by as.factor(superhost_binary)
## t = -1.3369, df = 2376.5, p-value = 0.1814
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.7946993  0.3396159
## sample estimates:
## mean in group 0 mean in group 1
##        79.09452        79.82206
```

```
##
## Call:
## lm(formula = price_numeric ~ superhost_binary, data = df_cleaned2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.8221 -10.0945   0.9055  10.9055  20.9055
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       79.0945     0.3421 231.179   <2e-16 ***
## superhost_binary   0.7275     0.5505   1.322    0.186
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.16 on 2791 degrees of freedom
## Multiple R-squared:  0.0006255,  Adjusted R-squared:  0.0002674
## F-statistic: 1.747 on 1 and 2791 DF,  p-value: 0.1864
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 0.1814, meaning that the null hypothesis cannot be rejected. Indicating that we can assume that the means in these two groups are similar. So we can say that the data does not show some significant differences between being a super host and not being a super host.

Price class: : > $ 100 to $ 150

```
## Levene’s Test for Homogeneity of Variance (center = mean)
##         Df F value Pr(>F)
## group    1   0.013 0.9093
##       2011
```

```
##
##  Two Sample t-test
##
## data:  price_numeric by as.factor(superhost_binary)
## t = 2.586, df = 2011, p-value = 0.009779
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  0.4087169 2.9741871
## sample estimates:
## mean in group 0 mean in group 1
##        124.8595        123.1680
```

```
##
## Call:
## lm(formula = price_numeric ~ superhost_binary, data = df_cleaned3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23.86 -12.17  -2.86  10.83  26.83
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      124.8595     0.3944 316.563  < 2e-16 ***
## superhost_binary  -1.6915     0.6541  -2.586  0.00978 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.12 on 2011 degrees of freedom
## Multiple R-squared:  0.003314,   Adjusted R-squared:  0.002819
## F-statistic: 6.688 on 1 and 2011 DF,  p-value: 0.009779
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 0.009779, meaning that the null hypothesis can be rejected. Indicating that we can assume that the means in these two groups are not similar. So we can say that the data does show some significant differences between being a super host and not being a super host.

Price class: : > $ 150 to $ 200

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1   0.592 0.4418
##       924
```

```
##
##  Two Sample t-test
##
## data:  price_numeric by as.factor(superhost_binary)
## t = -0.54209, df = 924, p-value = 0.5879
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.695736  1.528836
## sample estimates:
## mean in group 0 mean in group 1
##        174.3702        174.9536
```

```
##
## Call:
## lm(formula = price_numeric ~ superhost_binary, data = df_cleaned4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.9536 -14.3702  -0.3702  12.0464  25.6298
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      174.3702     0.6147 283.687   <2e-16 ***
## superhost_binary   0.5835     1.0763   0.542    0.588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.35 on 924 degrees of freedom
## Multiple R-squared:  0.0003179,  Adjusted R-squared:  -0.000764
## F-statistic: 0.2939 on 1 and 924 DF,  p-value: 0.5879
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 0.5879, meaning that the null hypothesis cannot be rejected. Indicating that we can assume that the means in these two groups are similar. So we can say that the data does not show some significant differences between being a super host and not being a super host.

Price class: : > $ 200 to $ 250

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   1  0.0048 0.9447
##       385
```

```
##
##  Two Sample t-test
##
## data:  price_numeric by as.factor(superhost_binary)
## t = 1.3176, df = 385, p-value = 0.1884
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -1.132763  5.735799
## sample estimates:
## mean in group 0 mean in group 1
##        228.1268        225.8252
```

```
##
## Call:
## lm(formula = price_numeric ~ superhost_binary, data = df_cleaned5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.127 -12.476  -1.127  13.873  24.175
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      228.1268     0.9011 253.159   <2e-16 ***
## superhost_binary  -2.3015     1.7467  -1.318    0.188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.19 on 385 degrees of freedom
## Multiple R-squared:  0.004489,   Adjusted R-squared:  0.001904
## F-statistic: 1.736 on 1 and 385 DF,  p-value: 0.1884
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 0.1884, meaning that the null hypothesis cannot be rejected. Indicating that we can assume that the means in these two groups are similar. So we can say that the data does not show some significant differences between being a super host and not being a super host.

Price class: : > $ 250

```
## Levene’s Test for Homogeneity of Variance (center = mean)
##        Df F value    Pr(>F)
## group   1  19.775 1.012e-05 ***
##       705
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
##  Welch Two Sample t-test
##
## data:  price_numeric by as.factor(superhost_binary)
## t = 4.0915, df = 677.08, p-value = 4.802e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##   95.9625 273.0439
## sample estimates:
## mean in group 0 mean in group 1
##        635.3143        450.8112
```

```
##
## Call:
## lm(formula = price_numeric ~ superhost_binary, data = df_cleaned6)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -384.3 -302.3 -154.8   14.7 9363.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        635.31      33.50  18.964  < 2e-16 ***
## superhost_binary  -184.50      58.36  -3.162  0.00164 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 729.4 on 705 degrees of freedom
## Multiple R-squared:  0.01398,    Adjusted R-squared:  0.01258
## F-statistic: 9.996 on 1 and 705 DF,  p-value: 0.001636
```

The null hypothesis here is that the means in these two groups are similar to each other. The findings are that this test has a p-value = 4.802e-05, meaning that the null hypothesis can be rejected. Indicating that we can assume that the means in these two groups are not similar. So we can say that the data does show some significant differences between being a super host and not being a super host.

Conclusion

In 2 price classes we see a significant difference in the average asking price between the two groups, in the price classes: $>$ \$ 100 to \$ 150 and $>$ \$ 250. But the outcome for the other price classes, it seems the pricing does not matter whether you are a super host or not a super host.