Data Science and Analytics Assignment

Luigui Gallardo-Becerra

4/14/2022

Statistical analysis

Is there a statistically significant relationship between the cancellation policy and unit price?

First, we need to import our data (listings.csv) into R.

Looking at the data, the "price" contains the "$" and "," symbols. It is easier to manage the data if we remove these character.

Before evaluate the relationship between these variables, we must know the distribution of our data, and with it select a parametric or non-parametric method. To achieve this we can use the Shapiro-Wilk's test.

```
##
##  Shapiro-Wilk normality test
##
## data:  listings$price
## W = 0.74283, p-value < 2.2e-16
```

From this result we can assume that the data is significantly different to the normal distribution (p-value $<$ 0.05), so we must use non-parametric tests.

Now we can answer the question. We can use a Kruskal-Wallis to compare the means between groups.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  price by cancellation_policy
## Kruskal-Wallis chi-squared = 256.6, df = 2, p-value < 2.2e-16
```

With this result we can conclude that there are differences in the price between the cancellation policy groups.

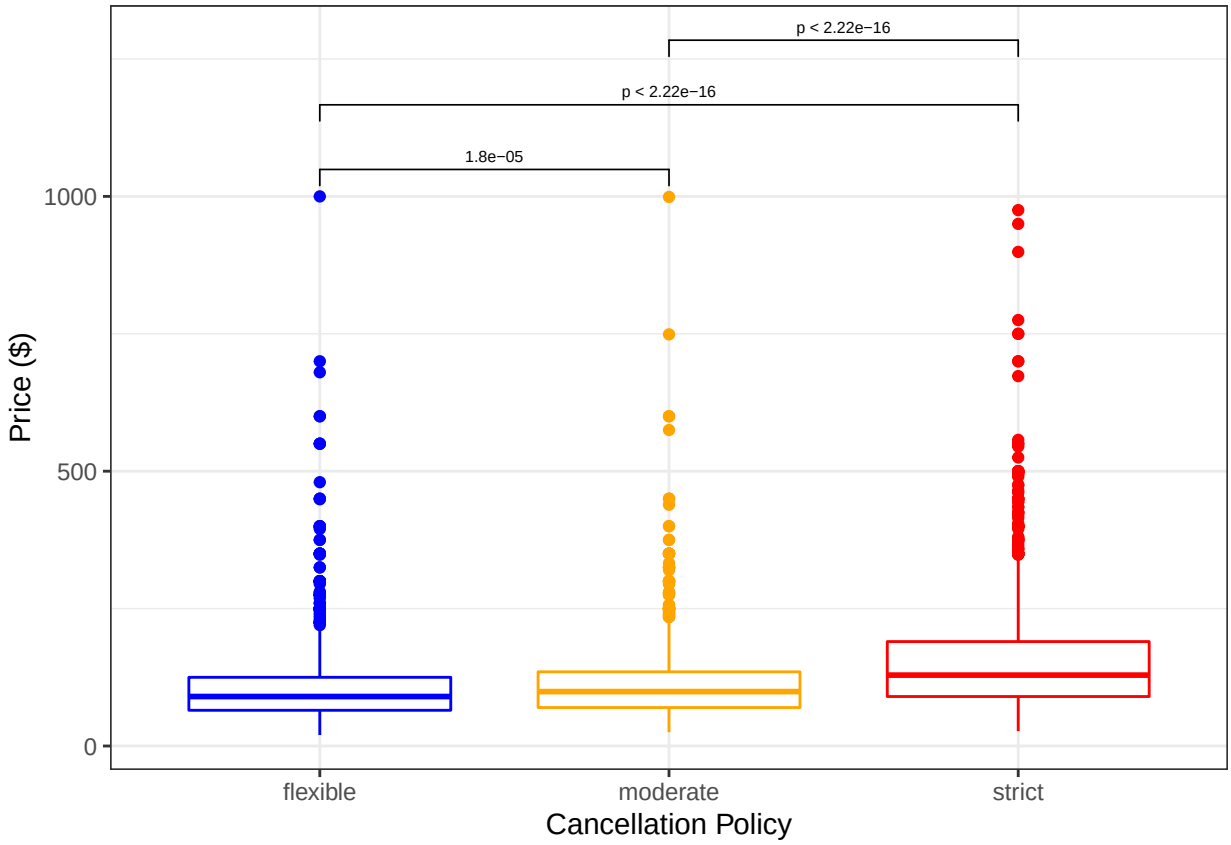
To know which groups differ from each other, we can use a Wilconxon's test. Additionally we can obtain a p-value adjustment with FDR (False Discovery Rate).

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  listings$price and listings$cancellation_policy
##
##          flexible moderate
## moderate 1.8e-05  -
```

```
## strict   < 2e-16  < 2e-16
##
## P value adjustment method: fdr
```

The p-values obtained were much lower than 0.05, so whit analysis we can conclude that the three groups means differ from each other.

Additionally, we can obtain a boxplot to visualize these results.



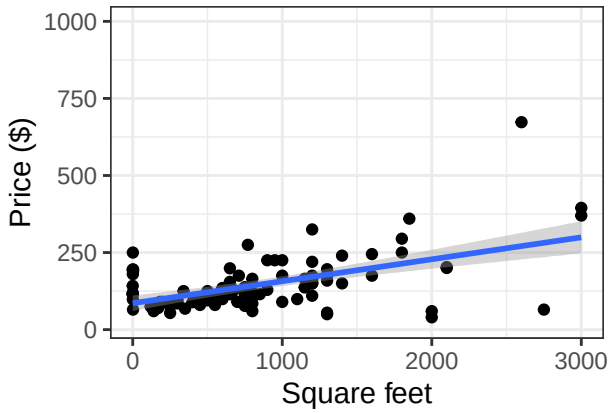
Construct a linear model to evaluate how the price of a rental is influenced by unit size, the number of bedrooms, the number of bathrooms, and the maximum occupancy. What can you conclude about the influence of these attributes on price?

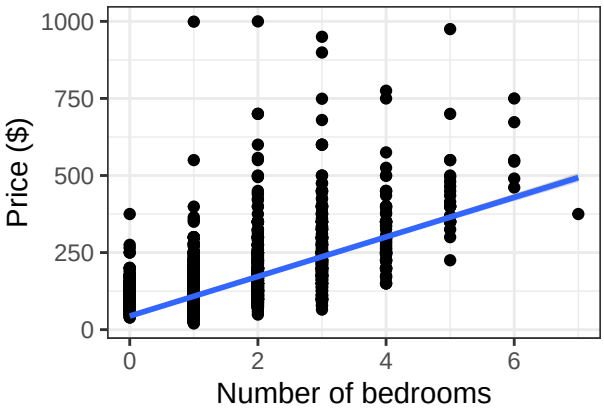
We can use the same data that the previous analysis. Now we can make a Linear Model considering these factors.

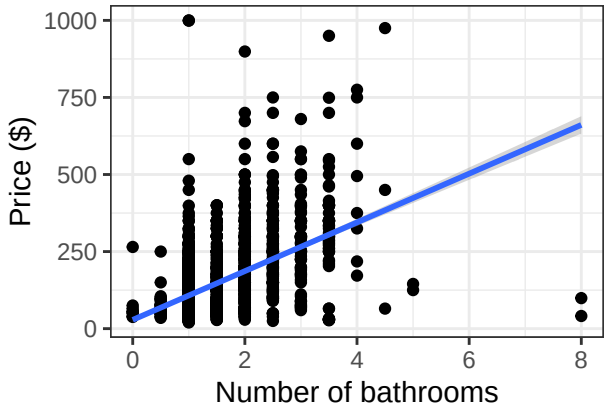
```
##
## Call:
## lm(formula = price ~ square_feet + bedrooms + bathrooms + guests_included,
##     data = listings)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.458  -29.722    1.251   30.987  156.079
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     25.681765  12.871953   1.995 0.049083 *
## square_feet      0.021184   0.009658   2.193 0.030893 *
## bedrooms        46.878254   6.830920   6.863  8.7e-10 ***
## bathrooms       -1.005877   8.540683  -0.118 0.906511
## guests_included 13.036911   3.608481   3.613 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.91 on 89 degrees of freedom
##   (3724 observations deleted due to missingness)
## Multiple R-squared:  0.7003, Adjusted R-squared:  0.6869
## F-statistic:    52 on 4 and 89 DF,  p-value: < 2.2e-16
```

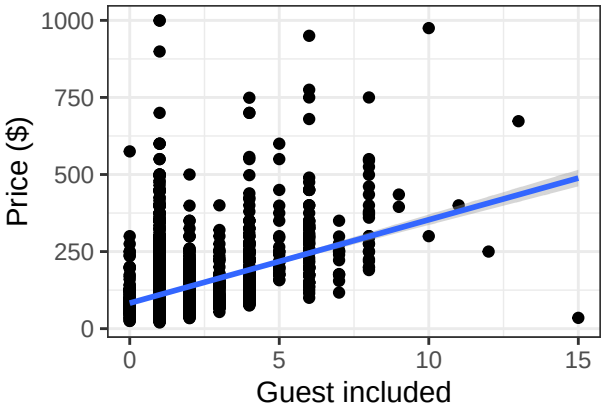
With these results we can conclude than there is a significant association between the price and the unit size (0.03), maximum occupancy (0.0005), number of bedrooms (8.7e-10). The number of bathrooms was not significantly associated with the price.

Additionally, we can obtain plots to visualize these results.


Price ($)
1000
750
500
250
0
0
1000
2000
3000
Square feet


Price ($)
1000
750
500
250
0
0
2
4
6
Number of bedrooms


Price ($)
1000
750
500
250
0
0
2
4
6
8
Number of bathrooms


Price ($)
1000
750
500
250
0
0
5
10
15
Guest included