STATISTICS 301 - Course Project

Is Price of Airbnbs in Chicago Associated to Airbnb's Accommodates, Property Type, or Bed Type?

Group: II

Students' Name: Jaeyeon Won, Charles Krebs, Christopher Wesseln

Professor: Dr. Laura Ziegler

Due Date: November 09, 2018

Word Count: 2097

Table of Contents

1. Introduction

According to Airbnb, it has been increasing its accommodations to over 5 million places in more than 81,000 cities and 191 countries since 2008 when it was founded. Each Airbnb that has different property types, accommodates, amenities, and etc. also has various price range. In this report, we will analyze the data and conduct some tests with the ultimate purpose of determining the association between the price and accommodates, property type, or bed type.

2. Hypothesis

There are many factors which might be associated to the low price or to the high price of Airbnbs. Out of the potential factors, we hypothesize that Real Bed, Apartment, and larger accommodates are associated with higher Price of Airbnbs.

3. Description of Data

The data set was collected by the people who host a place to stay on Airbnb in Chicago, from August, 2008 and May, 2017. The sample was taken using SRS (Simple Random Sample) with the sample size of 500. Also, data on 59 different variables were gathered including 36 categorical and 23 numerical variables. Particularly, Price was selected as the response variable. It is price per night and is collected in US dollars. Moreover, 2 categorical variables (Property Type, Bed Type) and 1 quantitative variable (Accommodates) were chosen as explanatory variables. Specifically, Property Type variable, which originally had 12 categories, was modified by collapsing the levels to 4 categories including Apartments, House, Condominium, and Other, because the number of other 8 levels were small. Among the levels chosen, Apartments was selected as the baseline group. Similarly, Bed Type variable originally had 5 levels including Airbed, Couch, Futon, Pull-out Sofa, and Real Bed. However, the categories were collapsed to 2 levels, which are Not Real Bed and Real Bed, since the beds that are not real beds have very few units. In addition, Real Bed Type was used as the reference group.

4. Data Analysis and Interpretation

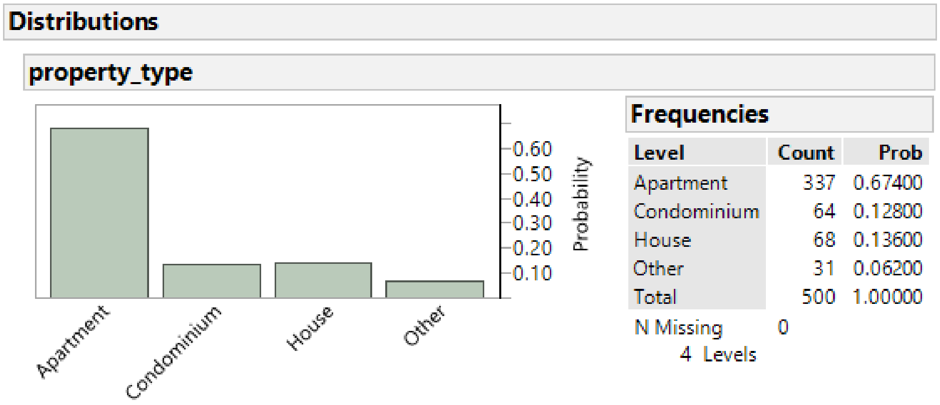
4.1 Bar Graphs and Histograms

4.1.1 Bar Graphs

4.1.1.1 Property Type

The following bar graph shows the number of Airbnbs that has different type of property. The property types have 4 categories in total, including Apartment, Condominium, House, and Other. In particular, the most common property type Chicago Airbnbs provide is Apartment that accounts for approximately 67% of 500 Airbnbs. Moreover, both Condominium and House type of property have a proportion of around 13%, individually. Also, Other types of property shows the least proportion of about 6%.

Bar Graph of Property Type



Distributions

property_type

Frequencies

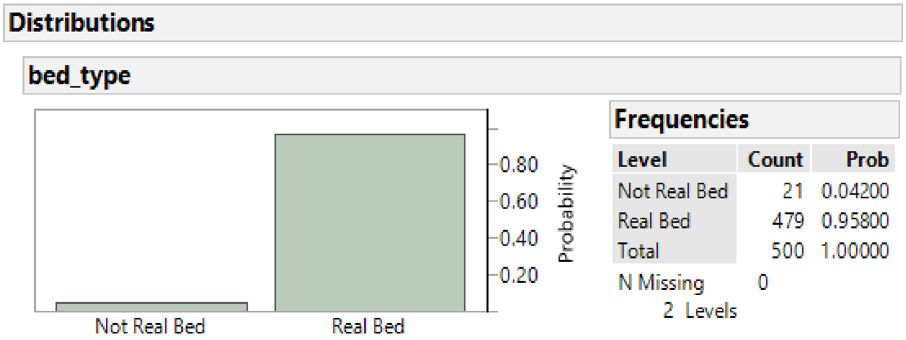
Level	Count	Prob
Apartment	337	0.67400
Condominium	64	0.12800
House	68	0.13600
Other	31	0.06200
Total	500	1.00000

N Missing 0
4 Levels

4.1.1.2 Bed Type

The bar graph below indicates the number of Airbnbs that has different type of bed. There are 2 levels of bed types, which are Not Real Bed and Real Bed individually. It is obvious that most of Airbnbs offer Real Bed type (95.8%) while Not Real Bed type is barely provided (4.2%). Particularly, they supply 479 Real Beds and 21 Not Real Beds.

Bar Graph of Bed Type



Distributions

bed_type

Frequencies

Level	Count	Prob
Not Real Bed	21	0.04200
Real Bed	479	0.95800
Total	500	1.00000

N Missing 0
2 Levels

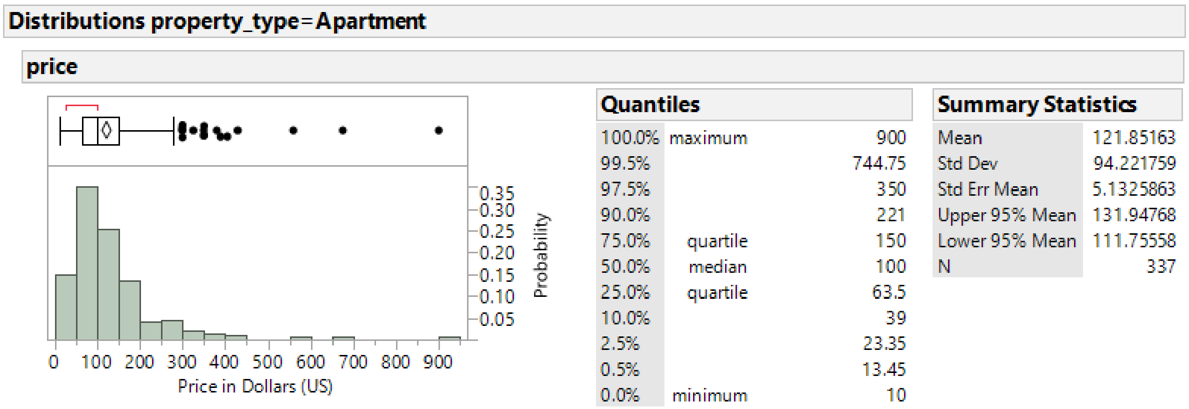
4.1.2 Histograms

4.1.2.1 Property Type

(a) Apartment

The distribution of Price by Apartment is unimodal and highly skewed to the right with a few outliers in high prices. Additionally, the box plot evidently shows the presence of more outliers. Most of the Chicago Apartment type Airbnbs cost a lower price, starting from $10. Considering the skewness and a few distinct outliers, the middle 50% price of apartment type property Airbnbs is $86.5, and the price of the 50th percentile is $100.

Histogram Price by Apartment



Distributions property_type=Apartment

price

Quantiles

100.0%	maximum	900
99.5%		744.75
97.5%		350
90.0%		221
75.0%	quartile	150
50.0%	median	100
25.0%	quartile	63.5
10.0%		39
2.5%		23.35
0.5%		13.45
0.0%	minimum	10

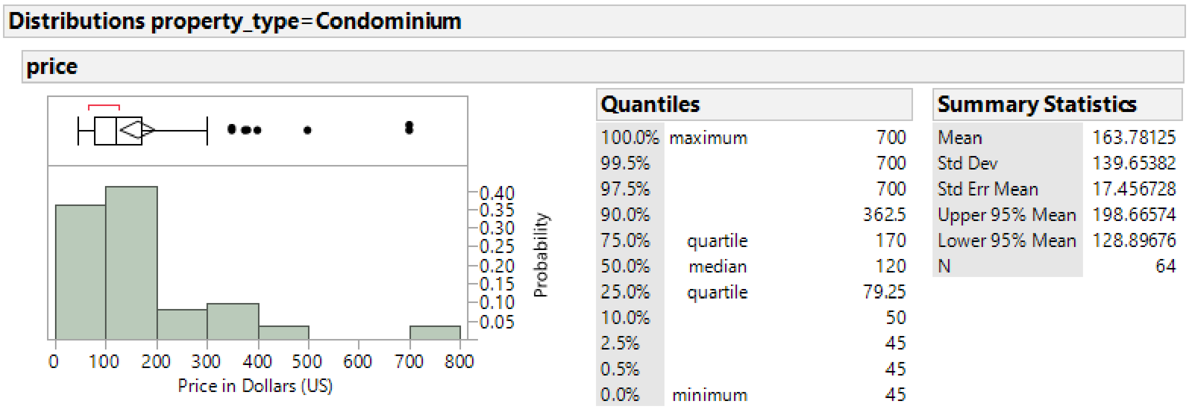
Summary Statistics

Mean	121.85163
Std Dev	94.221759
Std Err Mean	5.1325863
Upper 95% Mean	131.94768
Lower 95% Mean	111.75558
N	337

(b) Condominium

The distribution of Price by Condominium shows a unimodal and skewed to the right shape with 2 obvious outliers around the highest prices. Also, there are several potential outliers around the price range of $200 which are showed both in Histogram and box plot. Moreover, it has an IQR of $90.75 and a median of $120.

Histogram Price by Condominium



Distributions property_type=Condominium

price

Quantiles

100.0%	maximum	700
99.5%		700
97.5%		700
90.0%		362.5
75.0%	quartile	170
50.0%	median	120
25.0%	quartile	79.25
10.0%		50
2.5%		45
0.5%		45
0.0%	minimum	45

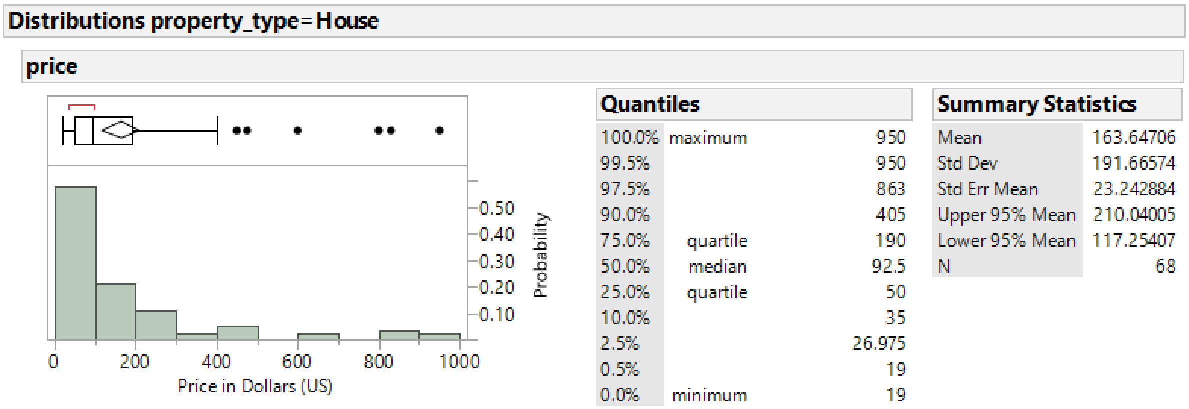
Summary Statistics

Mean	163.78125
Std Dev	139.65382
Std Err Mean	17.456728
Upper 95% Mean	198.66574
Lower 95% Mean	128.89676
N	64

(c) House

The Price by House type Airbnb has an extremely right-skewed and unimodal distribution. Starting from $19, a large proportion of Airbnbs with House property type costs low prices while a small proportion of them have expensive prices. With several outliers in the high values, it has an IQR of $140 and a median of $92.5.

Histogram Price by House



Distributions property_type=House

price

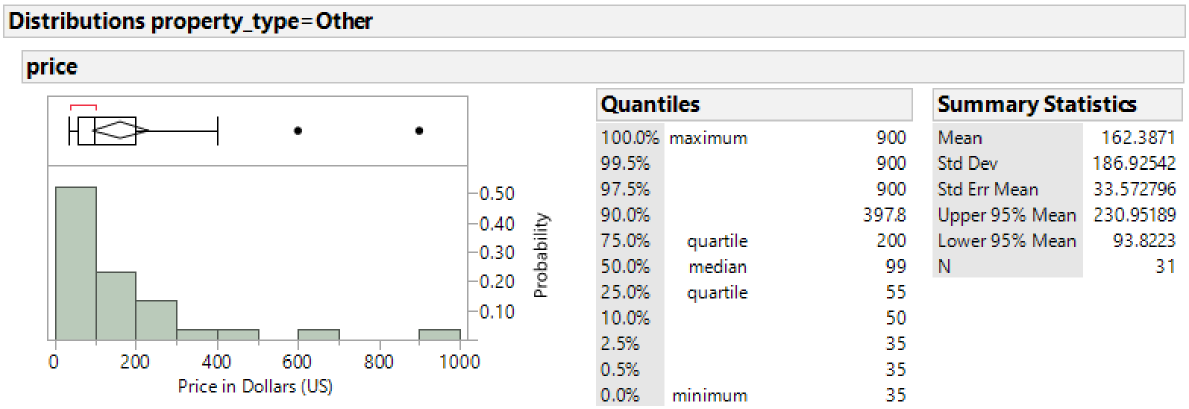
Quantiles		
100.0%	maximum	950
99.5%		950
97.5%		863
90.0%		405
75.0%	quartile	190
50.0%	median	92.5
25.0%	quartile	50
10.0%		35
2.5%		26.975
0.5%		19
0.0%	minimum	19

Summary Statistics	
Mean	163.64706
Std Dev	191.66574
Std Err Mean	23.242884
Upper 95% Mean	210.04005
Lower 95% Mean	117.25407
N	68

(d) Other

The distribution of Price by Other property types is skewed to the right and unimodal with 2 outliers above the price range of $600. Due to the fact that the distribution has a highly right - skewed shape with a few outliers, the median of $99 shows a great difference from the mean of $162.39. In addition, an IQR of $145 has a distinct difference from the range of $865.

Histogram Price by Other



Distributions property_type=Other

price

Quantiles		
100.0%	maximum	900
99.5%		900
97.5%		900
90.0%		397.8
75.0%	quartile	200
50.0%	median	99
25.0%	quartile	55
10.0%		50
2.5%		35
0.5%		35
0.0%	minimum	35

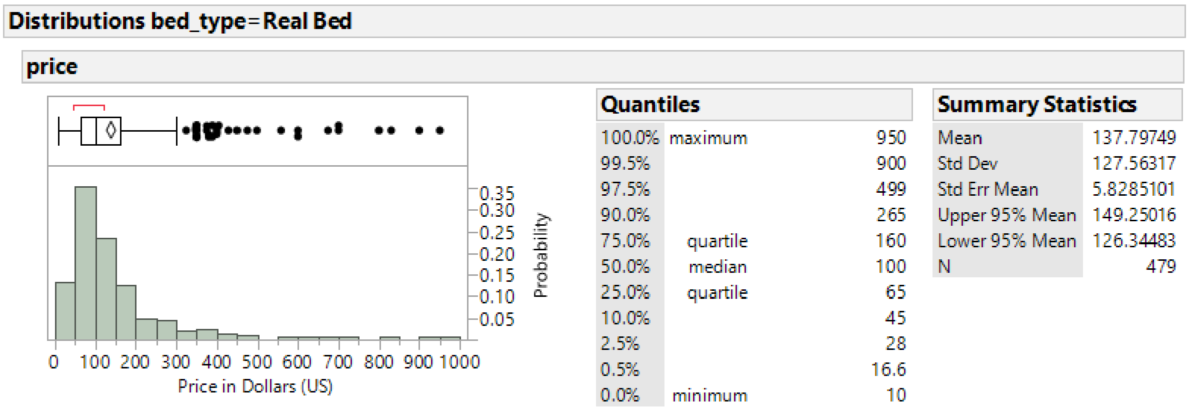
Summary Statistics	
Mean	162.3871
Std Dev	186.92542
Std Err Mean	33.572796
Upper 95% Mean	230.95189
Lower 95% Mean	93.8223
N	31

4.1.2.2 Bed Type

(a) Real Bed

The distribution of Price by Real Bed is skewed to the right and unimodal with an IQR of $95 and a median of 100. Moreover, there are lots of apparent and potential outliers above the price range of $600 represented both in the histogram and the box plot.

Histogram Price by Real Bed



Distributions bed_type=Real Bed

price

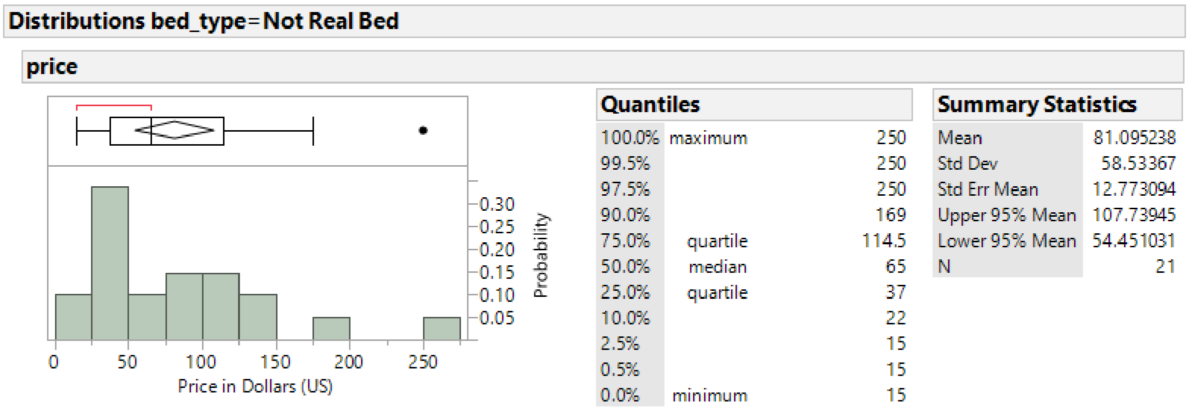
Quantiles		
100.0%	maximum	950
99.5%		900
97.5%		499
90.0%		265
75.0%	quartile	160
50.0%	median	100
25.0%	quartile	65
10.0%		45
2.5%		28
0.5%		16.6
0.0%	minimum	10

Summary Statistics	
Mean	137.79749
Std Dev	127.56317
Std Err Mean	5.8285101
Upper 95% Mean	149.25016
Lower 95% Mean	126.34483
N	479

(b) Not Real Bed

The distribution of Price by Not Real Bed is skewed to the right and unimodal with an IQR of $77.5 and a median of $65. Also, there are 2 outliers around the price range of $175.

Histogram Price by Not Real Bed



Distributions bed_type=Not Real Bed

price

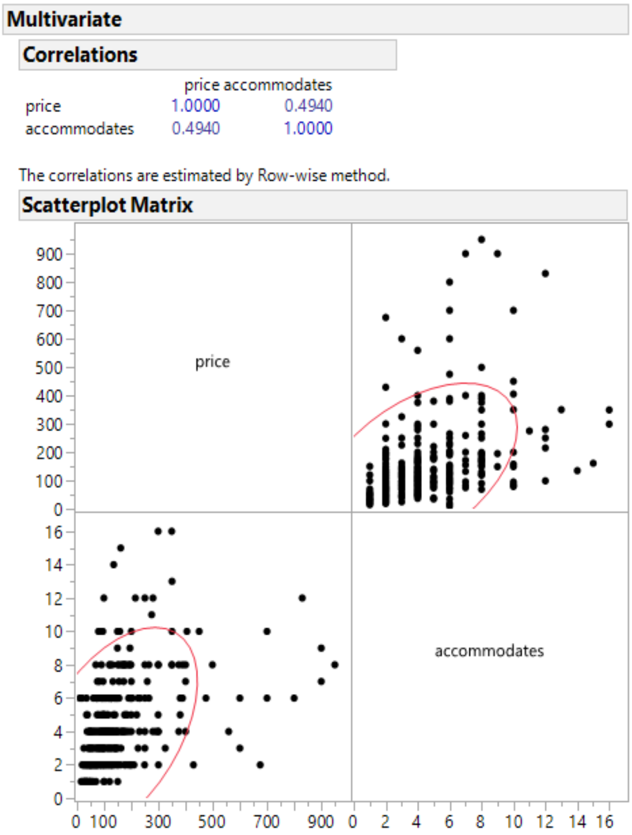
Quantiles		
100.0%	maximum	250
99.5%		250
97.5%		250
90.0%		169
75.0%	quartile	114.5
50.0%	median	65
25.0%	quartile	37
10.0%		22
2.5%		15
0.5%		15
0.0%	minimum	15

Summary Statistics	
Mean	81.095238
Std Dev	58.53367
Std Err Mean	12.773094
Upper 95% Mean	107.73945
Lower 95% Mean	54.451031
N	21

4.2 Correlation and Scatterplot Matrix

Based on the scatterplot matrix, firstly, Price and Accommodates are non-linearly, moderately, and positively associated. Furthermore, there are a few regression outliers, and multiple outliers both in high prices and in large accommodates (X-direction and in Y-direction). Since Price and Accommodates have non-linear relationship, it is not appropriate to look at the correlation between the two quantitative variables.

ScatterPlot Matrix between Price and Accommodates



Multivariate

Correlations

	price	accommodates
price	1.0000	0.4940
accommodates	0.4940	1.0000

The correlations are estimated by Row-wise method.

Scatterplot Matrix

4.3 Multiple Regression Equation

From our data with 2 categorical variables including 6 levels and 1 quantitative variable with the sample size of 500, the Multiple Regression Equation is

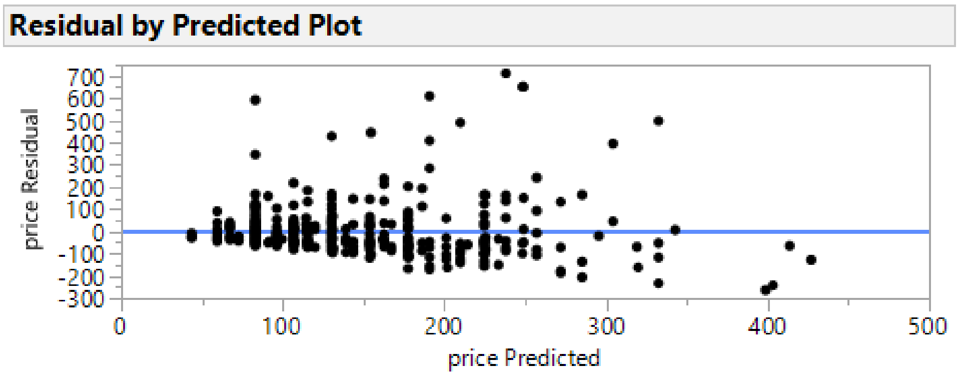
$$\widehat{Price} = 35.6703 + 32.3789 * (Condominium) + 13.3236 * (House) + 47.9660 * (Other) - 15.7772 \ * (Not\ Real\ Bed) + 23.6255 * (Accommodates)$$

Multiple regression equation above demonstrates that the predicted Price of an Airbnb per night with the property type of Apartment and with the Real Bed type is $35.7 when the Accommodates is 0 people. However, it is not appropriate to interpret this value not only because the Accommodates cannot be 0 people but also because the minimum Accommodates in this data was 1 person. Regarding the Property Types, it indicates that the mean Price for Condominiums is $32.4 more expensive than the Apartments while the Houses costs $13.3 more than Apartment types, on average, after controlling for Bed Type and Accommodates. Additionally, the predicted Price of other property types is $48.0 more than Apartments, when Bed Type and Accommodates are held constant. Moreover, regarding Bed Types, the predicted Price for Not Real Bed is $15.8 less expensive than the Real Bed, assuming Property Type and Accommodates are held constant. Lastly, when the Accommodates increases by 1 person, the predicted Price increases by $23.6, after controlling for Property Type and Bed Type.

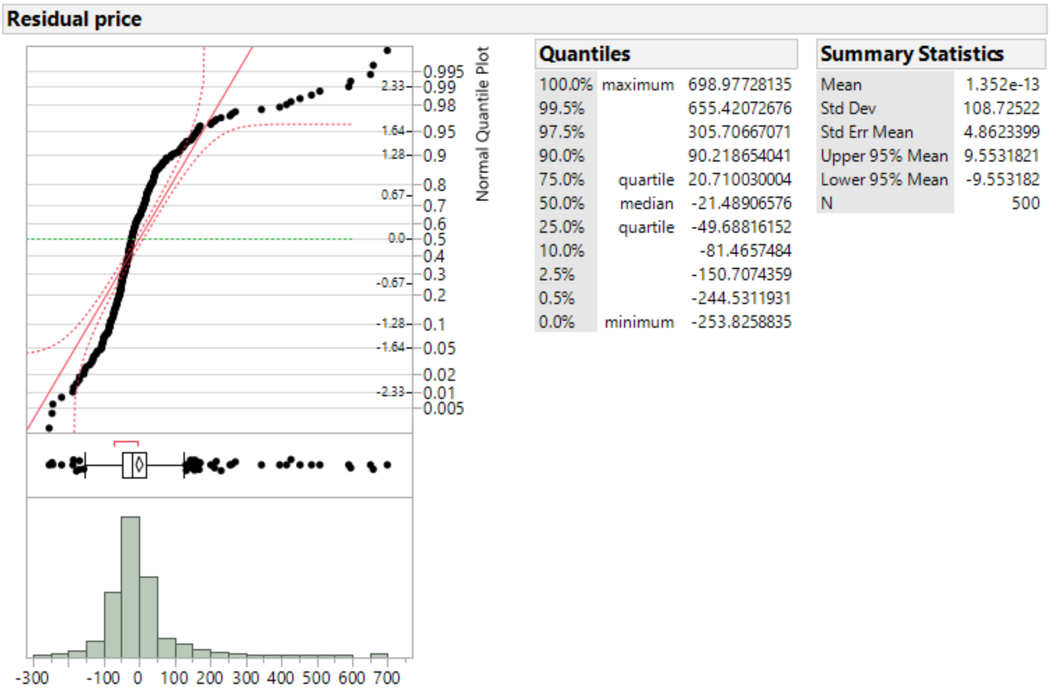
4.4 Conditions

In order to use the Multiple Regression Model in section 4.3 for inference, estimation, or prediction, four conditions must be met. Firstly, the use of SRS (Simple Random Sample) enables the values of Prices to be independent from each other. Moreover, we assume the data from an Airbnb would not affect the data from another Airbnb, which also satisfies the independence condition. Secondly, looking at the residual plot, the linearity condition is reasonably met since there is no curved pattern. In addition, even though there are more residual points below the value of 0, multiple points with high values probably even out the points below, enabling the mean residual to be approximately equal to 0. Thirdly, homoscedasticity condition is violated since a distinct fan shape in the residual plot demonstrates that the variance of price is not constant for all predicted values of price. Last but not least, normality condition is violated since the normal quantile plot shows some obvious curvatures, and most of the data are outside of the recommended bounds. Furthermore, the residual histogram does not follow a normal model, showing a right-skewed distribution. Even though the QQPlot does not show a normally distributed shape and the residual histogram has a right-skewed shape, it is not much of a concern because of the large sample size of 500.

Residual Plot by Predicted Plot



Normal Quantile Plot with Residual Histogram



Quantiles

100.0%	maximum	698.97728135
99.5%		655.42072676
97.5%		305.70667071
90.0%		90.218654041
75.0%	quartile	20.710030004
50.0%	median	-21.48906576
25.0%	quartile	-49.68816152
10.0%		-81.4657484
2.5%		-150.7074359
0.5%		-244.5311931
0.0%	minimum	-253.8258835

Summary Statistics

Mean	1.352e-13
Std Dev	108.72522
Std Err Mean	4.8623399
Upper 95% Mean	9.5531821
Lower 95% Mean	-9.553182
N	500

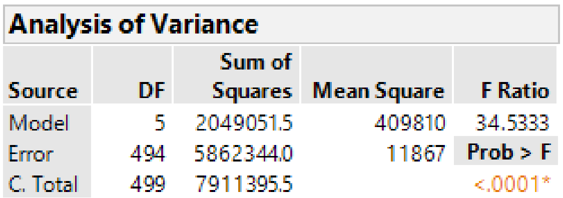
4.5 Significance

We have independence and linearity conditions met in section 4.4. Also, the normality condition is not concerned. Therefore, we can examine the significance of overall model as well as individual slopes.

4.5.1 Overall Model Significance

Looking at the ANOVA, overall, the multiple regression model (section 4.3) is significant in predicting the Price of Airbnb based on the F Ratio of 34.5333, which produces a p-value less than 0.0001, however, the condition for constant variance was not met.

ANOVA



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	2049051.5	409810	34.5333
Error	494	5862344.0	11867	Prob > F
C. Total	499	7911395.5		<.0001*

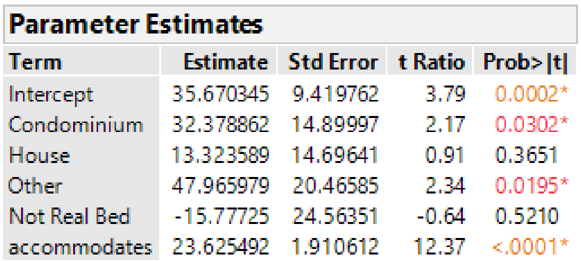
4.5.2 Individual Slopes

In this report, we surmised that Apartments will require higher price than Condominium, House, and other types of Airbnbs. Firstly, the null hypothesis, that the population slope β_1 was 0, was failed to be rejected in favor of the alternative, the population slope β_1 was less than 0 since the Test Statistic of t is 2.17 and the P-value is $\left(1-\frac{0.0302}{2}\right) = 0.9849$, which is greater than any reasonable α. Therefore, we do not have evidence that the population mean Price of Apartments costs more than Condominiums, assuming Bed Type and Accommodates are held constant. Secondly, we have $H_0\colon \beta_2 = 0, H_a\colon \beta_2 < 0$ with the t-ratio of 0.91 and the P-value of $\left(1-\frac{0.3651}{2}\right) = 0.81745$. Because the P-value is greater than any reasonable α, we decide the fail to reject the null hypothesis. Hence, there is no evidence that the population mean Price of Apartments is more expensive than Houses, when Bed Type and Accommodates are held constant. Lastly, $H_0\colon \beta_3 = 0$ was failed to be rejected in favor of $H_a\colon \beta_3 < 0$ due to the t-ratio of 2.34 and the P-value of $\left(1-\frac{0.0195}{2}\right) = 0.99025$, which is significantly greater than any reasonable α. Therefore, when Bed Type and Accommodates are held constant, we do not have evidence that the population mean Price of Apartments is more expensive than other types of Airbnbs.

Next, in terms of Bed Type, Real Beds were assumed to cost more than Not Real Beds. So, we have $H_0: \beta_4 = 0, H_a: \beta_4 < 0$. Then the Test Statistic of $\mathrm{t\,is} - 0.64$, and the $\mathrm{P} -$ value is 0.2605. Thus, we fail to reject $H_0: \beta_4 = 0$ because p-value is greater than α of 0.05. Thus, we do not have evidence that the population mean Price of Real Beds is more expensive than Not Real Bed, when Property Type and Accommodates are held constant.

Lastly, we presumed that the larger Accommodates are related to higher Price of Airbnbs. We have hypothesis statements of $H_0: \beta_5 = 0, H_a: \beta_5 > 0$. Also, Test Statistic of t is 12.37, and the $\mathrm{P} -$ value is less than 0.00005. Since the P-value is less than any reasonable α, we reject $H_0: \beta_5 = 0$. Therefore, as hypothesized, we have evidence that larger Accommodates are significantly associated with higher Airbnb Price on average, for the population, when Property Type and Bed Type are held constant.

Parameter Estimates



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	35.670345	9.419762	3.79	0.0002*
Condominium	32.378862	14.89997	2.17	0.0302*
House	13.323589	14.69641	0.91	0.3651
Other	47.965979	20.46585	2.34	0.0195*
Not Real Bed	-15.77725	24.56351	-0.64	0.5210
accommodates	23.625492	1.910612	12.37	<.0001*

4.6 Leverage Plot

With the purpose of examining the influential points in changing the placement of simple linear regression line, Leverage Plot was created.

4.6.1 Leverage

In order to find the leverage points that have the potentials to influence the regression line, we used a leverage cutoff value of

$$2\left(\frac{3+1}{500}\right) = 0.016$$

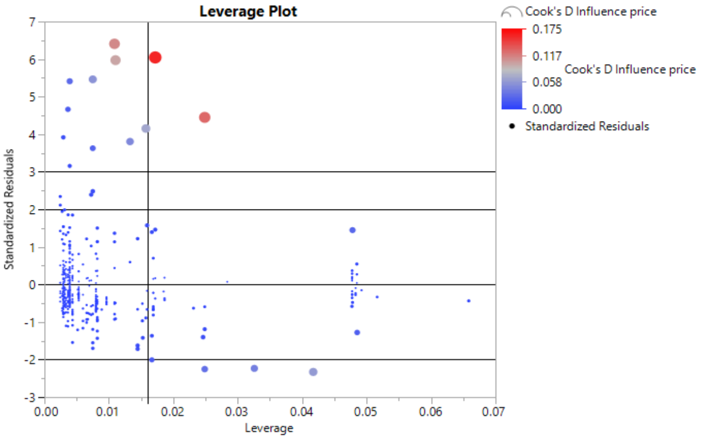
As can be found in the leverage plot, there are numerous high leverage points which have a bigger h (hat) values than the leverage cutoff value. Specifically, the point at the very right side of the plot shows the strong evidence that the point is extremely unusual and is a high leverage point with a hat value of 0.0659, which is a significantly greater than 0.016.

4.6.2 Cook's D

In the leverage plot, there is not a concern about the points that are considered to exert high influence based on Cook's D cutoff value of 1 since the highest D value is 0.163, which is significantly smaller than 1.

4.6.3 Standardized Residuals

There are 12 extremely unusual points that have a normal standardized residual greater than 3. Specifically, the highest standardized residual value is 6.409. Moreover, there are about 8 points with the standardized residuals of between 2 and 3 or between -2 and -3 which are considered somewhat unusual.



Leverage Plot

5. Conclusion

In conclusion, the histograms of Price by each property types, individual slopes with positive coefficients in multiple regression model, and hypothesis tests for slopes disproved our assumption that Apartment type of Airbnb will cost more than Condominiums, Houses, and other property types of Airbnbs, on average. On the other hand, the histograms of Price by each bed types as well as the negative slope of Not Real Bed (non-reference group) supported our hypothesis that the Price of Airbnbs with real beds will be more expensive than the not real beds while the hypothesis test for the slope was not supportive. Lastly, even though the Price and Accommodates were not linearly related, the positive relation between Price and Accommodates, the positive slope of Accommodates in the regression line and the hypothesis for the slope of Accommodates validated that the larger accommodates are significantly associated with the higher Airbnb prices, as we initially hypothesized.

Works Cited

“About Us.” Airbnb Press Room, https://press.airbnb.com/about-us/. Accessed 23 Oct. 2018.