airbnb

Amit Gupta - 673965500
Parichay Jain - 668555331
Abhinov Singh- 670092950
Indresh Triphati- 652815251
Shivanshu - 671083164

Executive Summary:

This report summarizes the statistical modeling and analysis results of price modeling for various listings on Airbnb in Chicago. The purpose of this report is to document the design implementation of the data model with linear regression technique used for this statistical analysis.

The initial dataset for designing our model has been taken from the Airbnb official site. This dataset contained 5147 observations from 69 neighborhoods with 16 variables. After data cleaning and adding two extra calculated attributes(log(price) and log(number_of_reviews)), the final data set contains 5147 observations across 18 variables. 4 independent variables and one control variable (neighborhood) have been chosen from this final dataset to model our dependent variable ("price") as per our research question.

The univariate statistics that summarize the distribution across the data and affect the price of a listing in Chicago have been discussed later in the report. The density plot of the price variable is very right skewed. Therefore, log(price) was used in the regression model which is almost normal and has been added as the calculated variable in the dataset. Other independent variables have also been modified and adjusted to model the price.

Various bivariate analyses on dependent and independent variables have been conducted. Welch two samples T-test and ANOVA are used to determine the effect of independent variables on the dependent variable i.e. price. The relationship between the dependent variable and independent variables have been displayed by the plots. The natural log transformation technique has been implemented to model price as a dependent variable as price is heavily right skewed.

Finally, a linear regression model is proposed for modeling the listing price while simultaneously adjusting for independent variables that were hypothesized to also (possibly) influence the price. The results agree with the earlier tests (ANOVA, t-test) and they confirm that all the three hypotheses are true.

Introduction

About Airbnb -

Airbnb is a peer-to-peer online marketplace and homestay network enabling people to list or rent short-term lodging in residential properties, with the cost of such accommodation set by the property owner. The company receives percentage service fees from both guests and hosts in conjunction with every booking. It has over 2,000,000 listings in 34,000 cities and 191 countries. Airbnb was founded in August 2008, is headquartered in San Francisco, California, and is privately owned and operated.

About our dataset -
Characteristics

Airbnb released this data from their company database. The data on the insideairbnb.com site is sourced from publicly available information from the Airbnb site.

The data has 16 variables with 5147 observations.
10 of the variables are numeric or can be converted to numeric.
Some variables like neighborhood_group contain several NA/missing values.

Variable	**Type**	**Description**
ID	Numeric	ID of the listing
Name	Character	Name of the host
Host_id	Integer	ID of the host
Host_name	Factor	Name of the host
Neighborhood_group	Logical	Name of the neighborhood group for a particular listing
Neighborhood	Factor	Name of the neighborhood for a particular listing
Latitude	Numeric	Latitude of a listing
Longitude	Numeric	Longitude of a listing
Room_type	Factor	Type of the room
Price	Integer	Price of the listing
Minimum_nights	Integer	Minimum nights for which the listing must be booked
Number_of_reviews	Integer	Number of total reviews for a particular listing
Last_review	Factor	Date on which last review was posted

Reviews_per_month	Numeric	Reviews per month for a particular listing
Calculated_host_listings_count	Integer	Number of distinct listings for a single host
Availability	Integer	Availability of the listing in a year (365 days)

Steps taken to clean the data -

1) Dropped the neighborhood_group column as it was entirely NA
2) Converted last_review from a factor to date format
3) Converted the room_type to a factor
4) The listings for which the number_of_reviews were 0, for those the reviews_per month were set to 0
5) Host_names which were blank or had a '-' were replaced with an NA

Research Question
What are the factors that affect the price of a listing in a Chicago neighbourhood?

Hypothesis

- The price of an Airbnb listing in a Chicago neighbourhood is higher when the listing has more reviews.
- The price of a listing is higher when there are more listings in a neighbourhood.
- The price is higher when the entire home/apartment is being rented.

Limitations:

There is no information regarding the square footage of any of the listings in the Airbnb dataset which could be one of the factors affecting the price of a listing in Chicago. Neither is there any information regarding the amenities provided in each of the listings. It would help to model the price more accurately if the dataset had a variable like 'amenities ratings'.

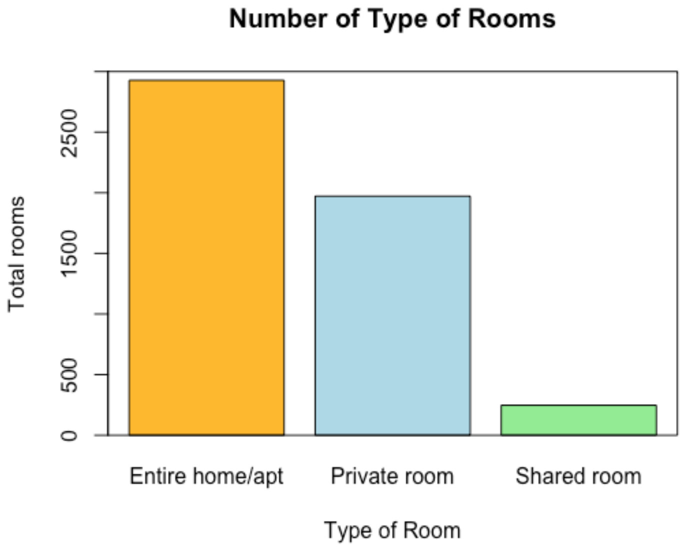
The dataset only had the number of reviews for each of the listings. It would help to model the price more accurately if the positivity / negativity of the review could be measured via text or ratings of the review.

Also, time of year (seasonality) affects the price of a particular listing not provided in the dataset. Since listing price is seasonal in nature, it could be predicted more accurately if the time of the year was also present.

Analysis:

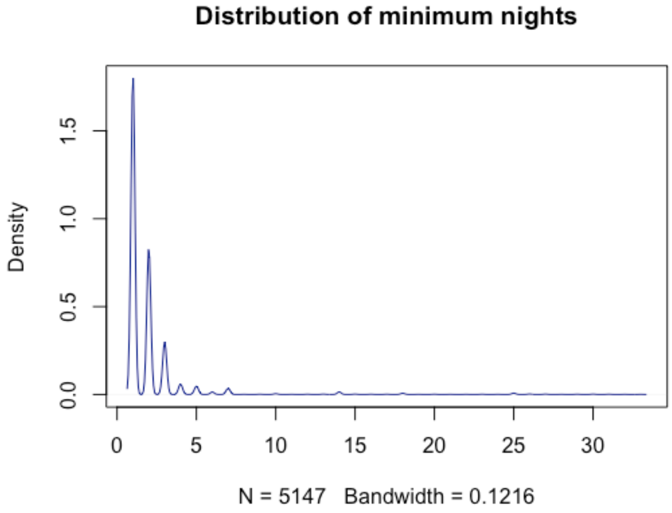
Univariate Analysis

Roomtype

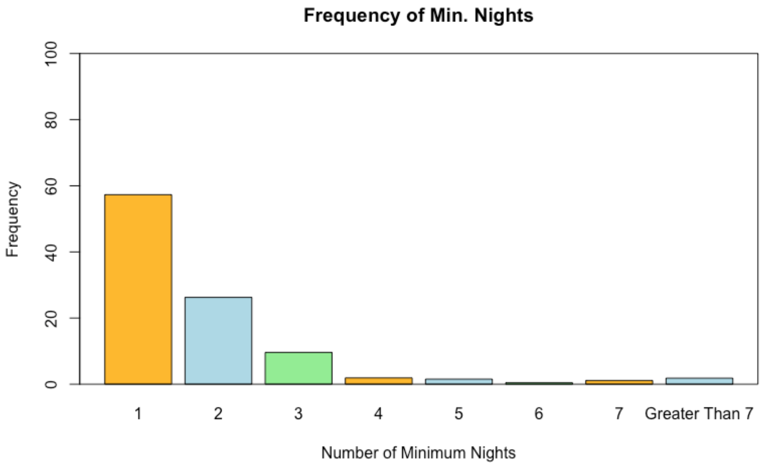


56.8% of all the listings available were for the entire home/apartment. Private rooms were 38.3% and only 4.8% were shared rooms.

Minimum_nights



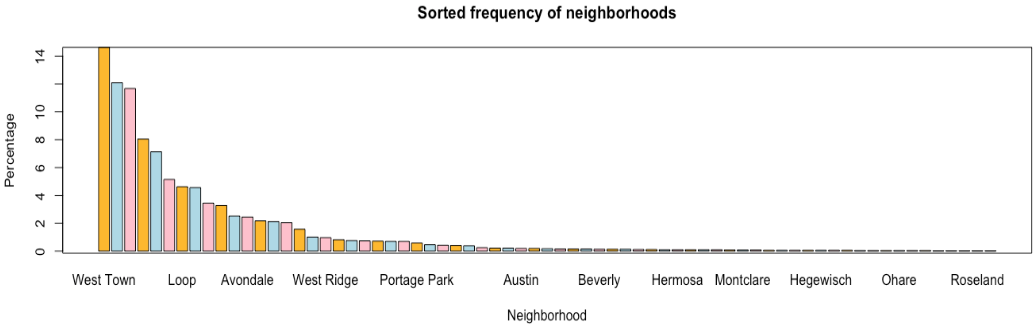
The variable is skewed to the right with a skewness of 6.77. The majority of the listings were up to 7 minimum nights of stay. After 7 nights, there were a lot of outliers up to 300 minimum nights. To simplify the analysis, we added a new column - Minimumnights_cat, a factor, and made minimum nights greater than 7 a factor.



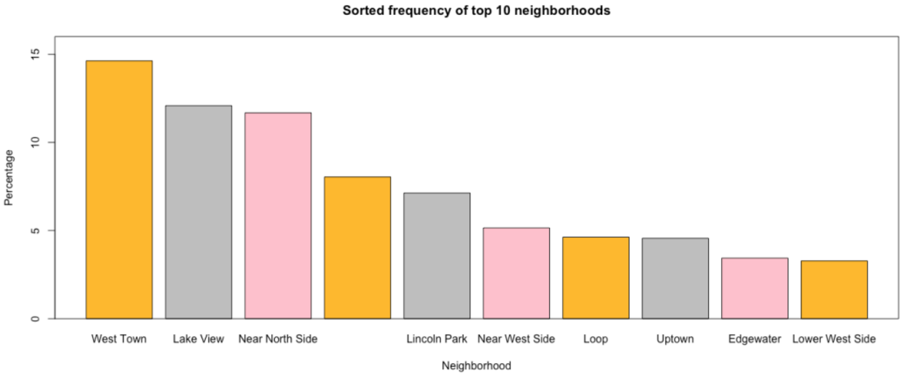
Most of listings require minimum 1 night's stay which is 2950 or approximately 57.3% of the total, followed by minimum 2 nights stay (26.24%). We can see that minimum nights greater than 7 are only 1.82%.

Neighborhood

Before cleaning the data

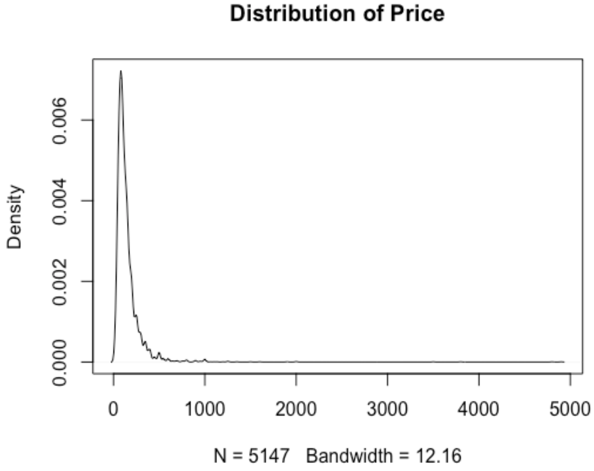


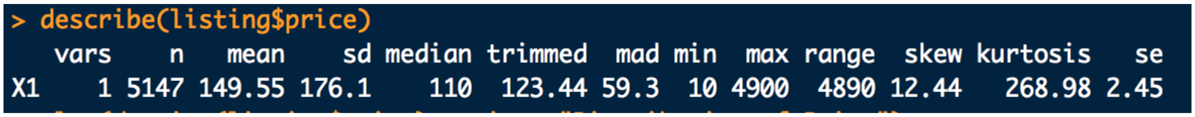
After cleaning the data, sorting and taking the top 10 neighborhoods in terms of the frequency of listings



The maximum number of listings are in West Town with 14.6% of all listings followed by Lake View with 12%.

Price

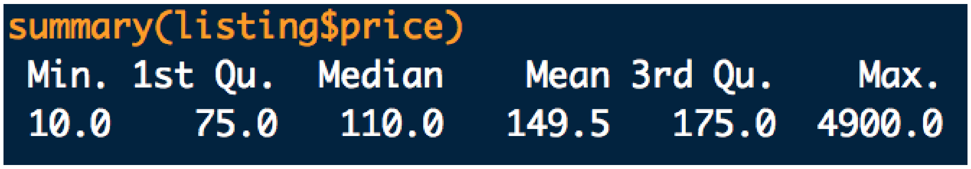




```
> describe(listing$price)
   vars    n   mean    sd median trimmed  mad min  max range  skew kurtosis   se
X1    1 5147 149.55 176.1    110  123.44 59.3  10 4900  4890 12.44   268.98 2.45
```

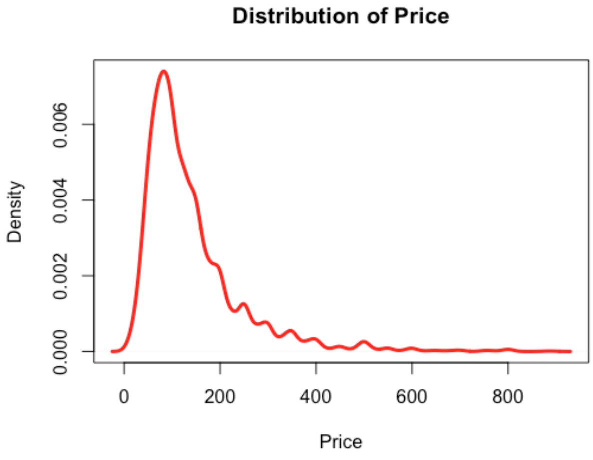
The median is smaller than the mean, so it is right skewed. The skewness is 12.44 which shows that it is heavily skewed. The range is very large between 10 and 4900. Kurtosis is 268.98 which is a lot higher than a normally distributed plot.

A summary of the price variable shows that 75% of the values are below 175.

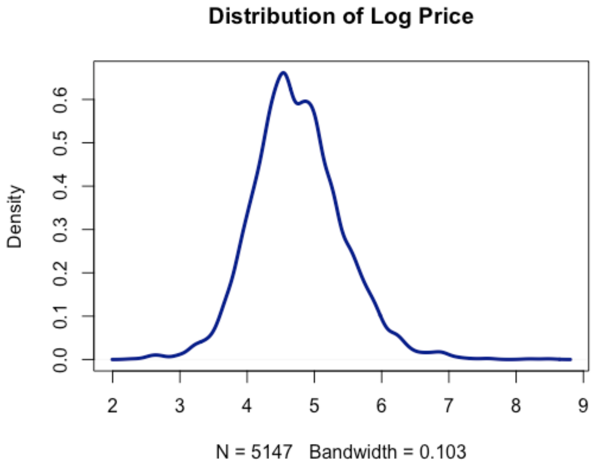


```
summary(listing$price)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 10.0    75.0   110.0   149.5   175.0  4900.0 
```

As it can be seen from the distribution, majority of the values are under 900. So only listings with a price under 900 will be considered.

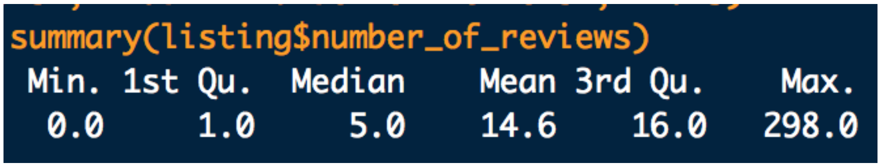


To convert the price to a more normal distribution we use the log function.

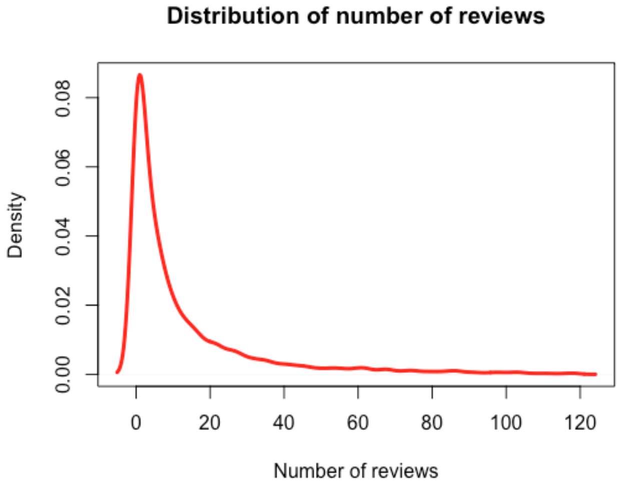


Number of Reviews

A summary of number_ of_ reviews shows that the mean number of reviews are 14.6 and the median is 5. The mean is greater than the median so the plot is right skewed.



```
summary(listing$number_of_reviews)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.0     1.0     5.0    14.6    16.0   298.0
```



Bivariate analysis:

1) **Room Type (Dependent variable) vs Price (Independent Variable)**

By using the aggregate function, the mean and standard deviation of room type vs price can be found.

	Room_Type	**Mean price of listing ($)**
1	Entire home/apt	199.38081
2	Private room	86.10345
3	Shared room	65.37247

The above table shows that there are 3 types of rooms available and among these entire home/apt has the highest mean price followed by the private room and the shared room has the least mean price.

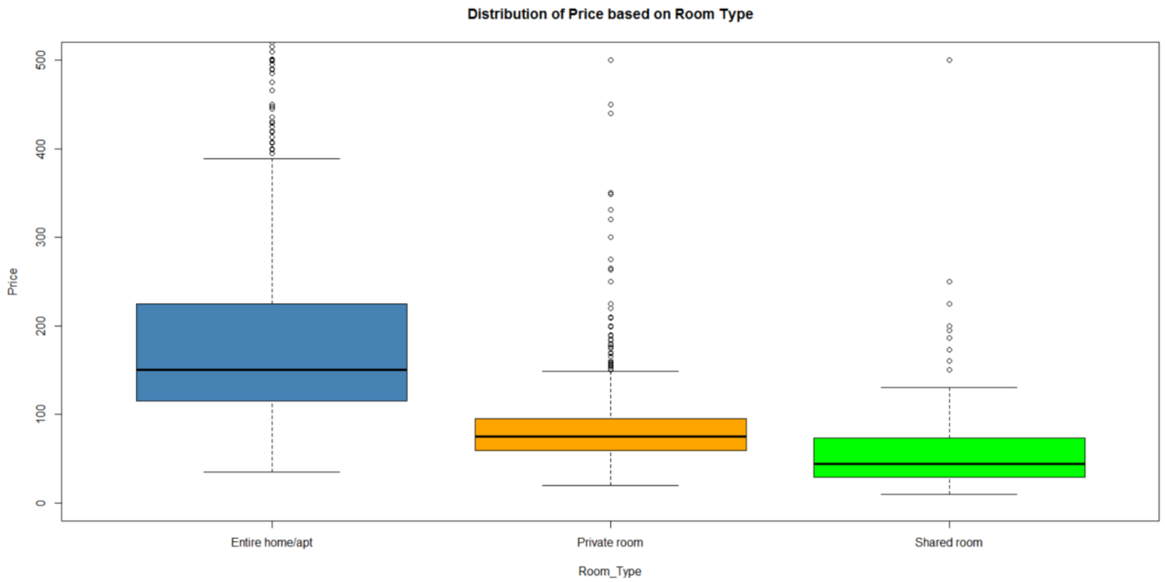
	Room_Type	**Std. dev of listing price ($)**
1	Entire home/apt	195.65780
2	Private room	119.64064
3	Shared room	98.10836

Similarly, the standard deviation of Price for room type follows the same order as the mean.

Below is the distribution of the price for the 3 room_types -

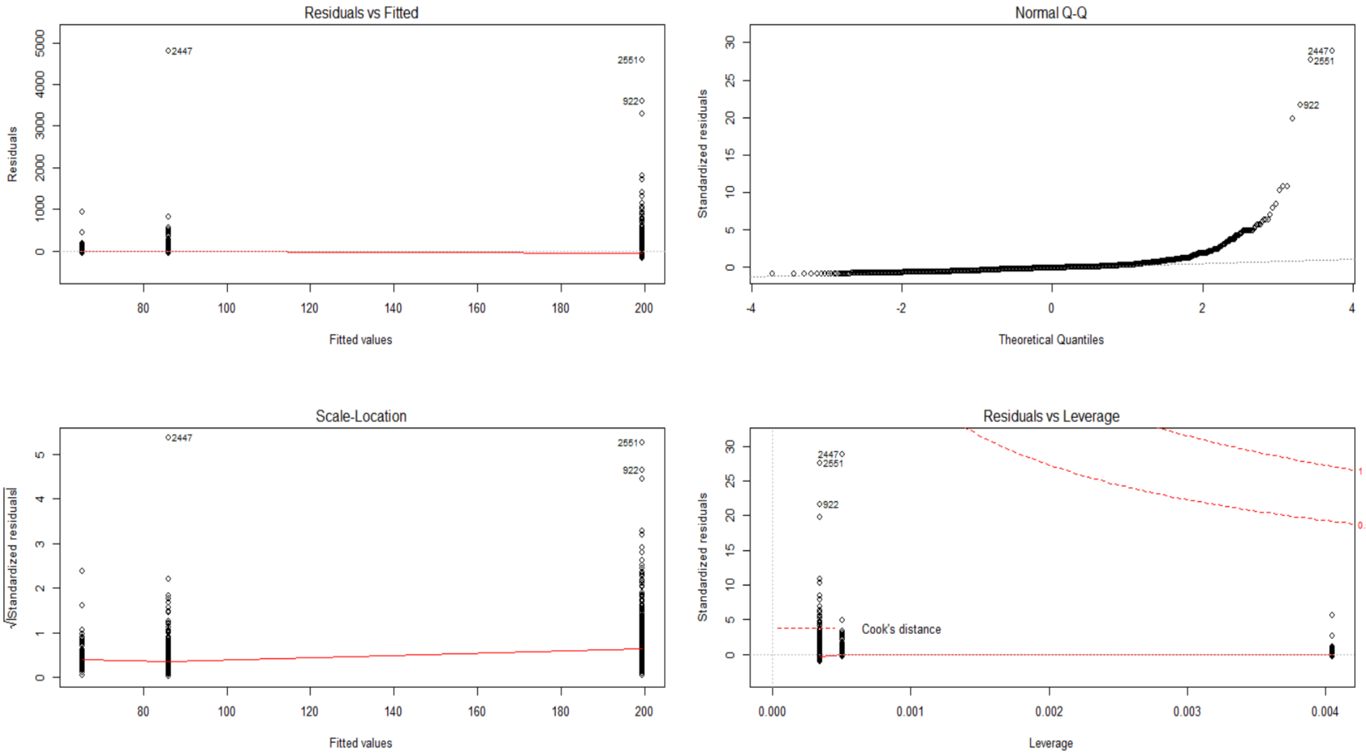
```
group: Entire home/apt
   vars    n   mean     sd median trimmed   mad min  max range  skew kurtosis   se
X1    1 2928 199.38 195.66    150  168.98 74.13  35 4800  4765 10.24   180.87 3.62
--------------------------------------------------------------------
group: Private room
   vars    n mean     sd median trimmed   mad min  max range  skew kurtosis   se
X1    1 1972 86.1 119.64     75   76.55 26.69  20 4900  4880 33.47  1328.19 2.69
--------------------------------------------------------------------
group: Shared room
   vars   n  mean    sd median trimmed   mad min  max range skew kurtosis   se
X1    1 247 65.37 98.11     44   49.98 28.17  10 1000   990 7.44    65.43 6.24
```

It can be seen that a shared room has the least price range of 990. Another interesting find is the minimum price for an entire home/apartment is $35.

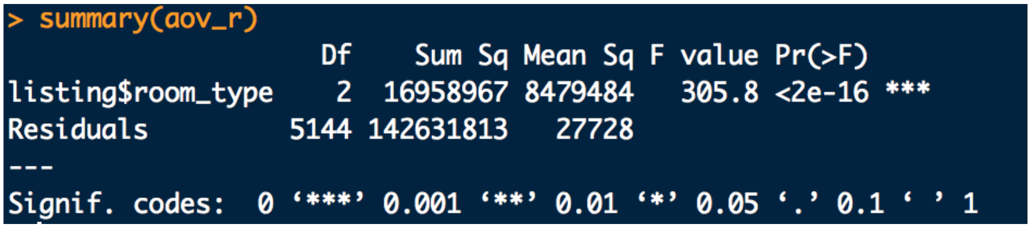


The above is the distribution of the all the different room types. From the box plot it can be seen that the median of entire home/apt is around $150 with many outliers above $400. The median price of private room is 90 with many outliers above $120. The shared room has median at $50 which is the lowest among all the room type and has with many outliers above $110.

All the plots show that rows 2447, 2552, 922 are outliers.



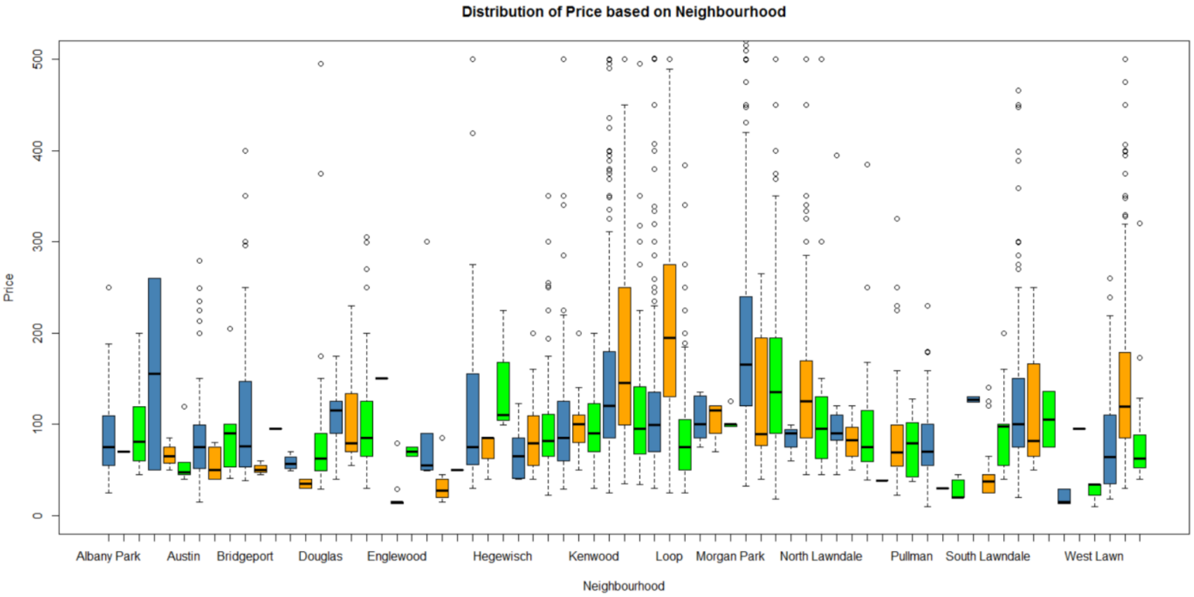
Result of ANOVA



```
> summary(aov_r)
                    Df    Sum Sq Mean Sq F value Pr(>F)    
listing$room_type    2  16958967 8479484   305.8 <2e-16 ***
Residuals         5144 142631813   27728                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

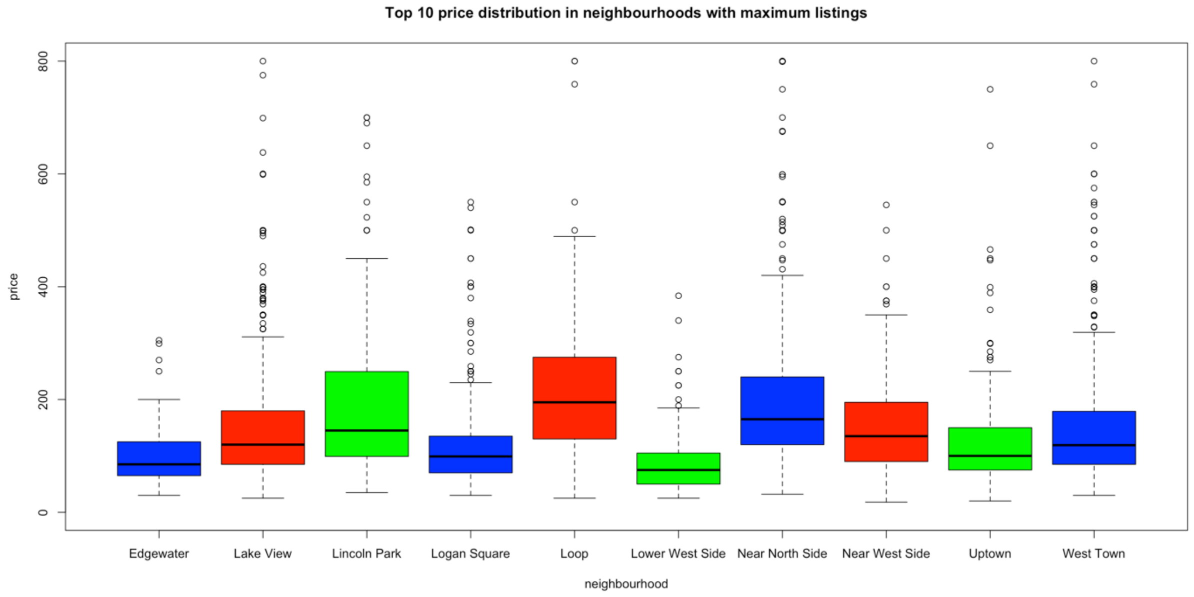
The p value is lesser than 2e-16. The null hypothesis that the mean of the price of the 3 room types is same and can be rejected. The alternate hypothesis that there is a difference in the mean price of the room types can be considered.

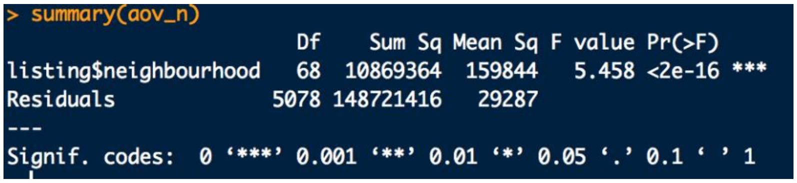
2) **Neighborhood vs. Price**



The boxplot depicts the distribution of prices in different neighbourhoods. It can be seen that there are a lot of outliers in each neighborhood.

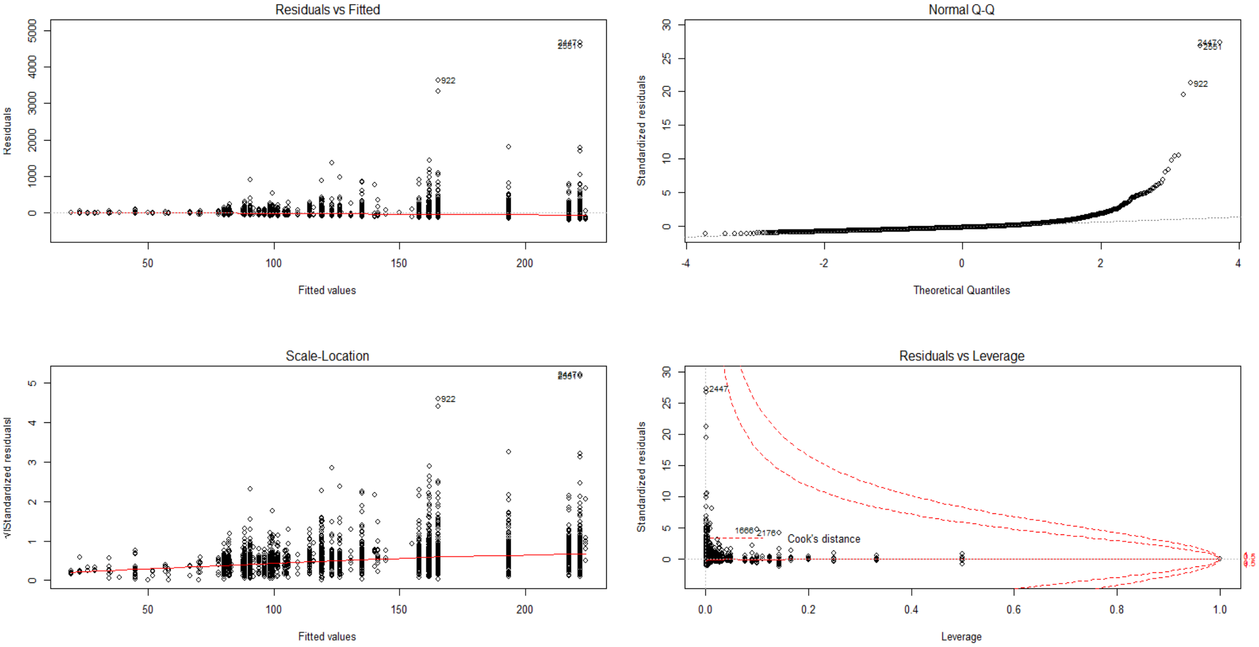
The price of top 10 neighbourhoods with maximum listing has been plotted for better clarity. The the median of Loop is the highest and Lower West side has the lowest median.





```
> summary(aov_n)
                        Df    Sum Sq Mean Sq F value Pr(>F)    
listing$neighbourhood   68  10869364  159844   5.458 <2e-16 ***
Residuals             5078 148721416   29287                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

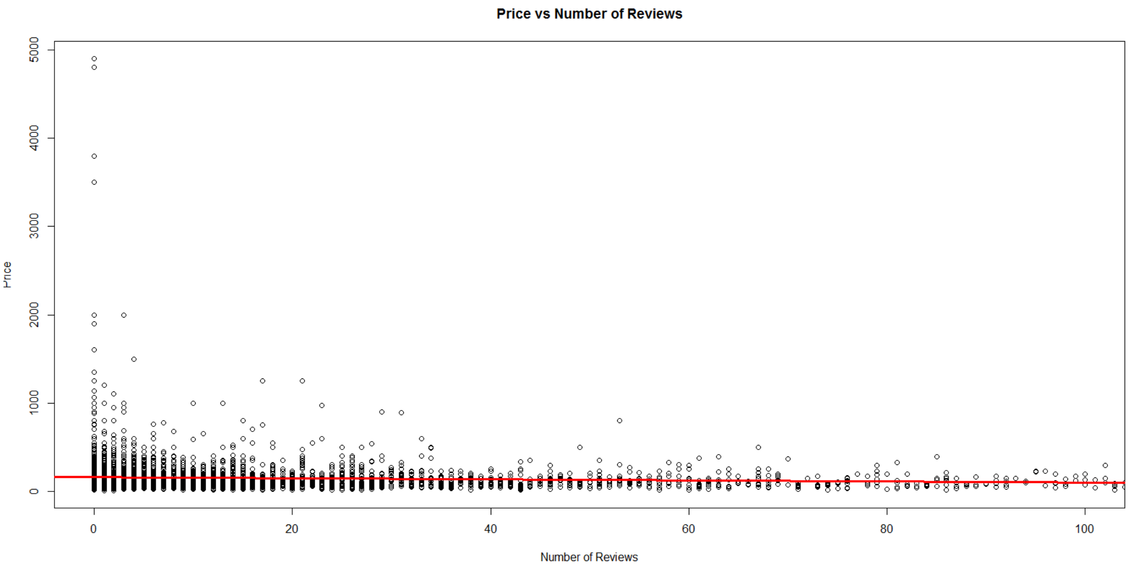
This above result shows that the P value is very low (<2e-16) and that the null hypothesis can be rejected over the alternate. (null hypothesis being; the average price for every neighbourhood is the same).

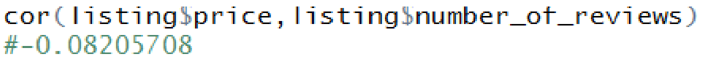


Again, there are same outliers; rows 2447, 1666, 2176, 922, 2551.

The listings on rows 2447 and 2551, have 0 reviews but a very high price of 4900 and 4800 respectively.

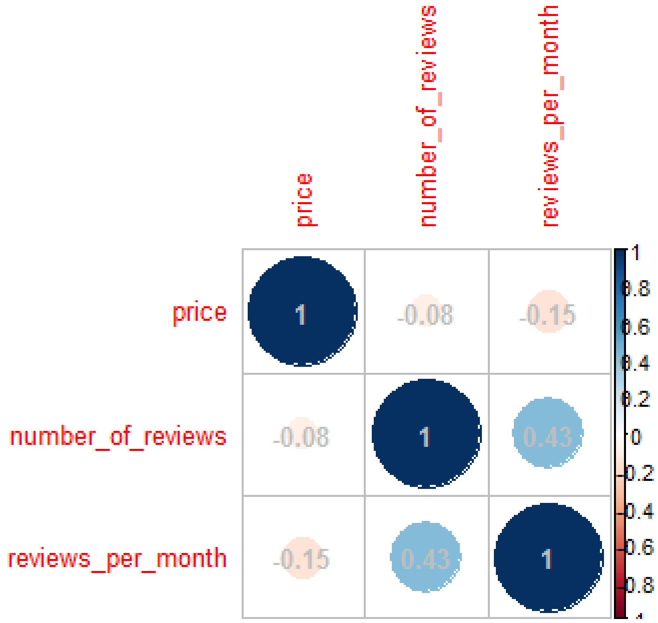
3) **Number of Reviews vs. Price**



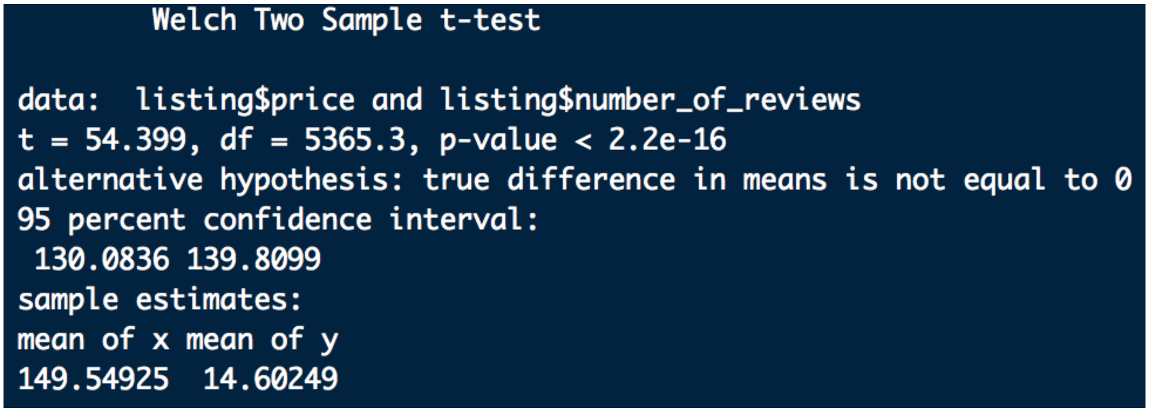


```
cor(listing$price,listing$number_of_reviews)
#-0.08205708
```

The correlation between the dependent variable, price and number of reviews is -0.082. This is not a good correlation (negative). The correlation between number_of_reviews and reviews_per_month is 0.43 which is fairly good.



To analyse further, a t.test has been performed on price and number of reviews to see whether number of reviews should be included in further analysis. Based on the t.test result, a decision will be made if variable number of reviews should be considered in the analysis.



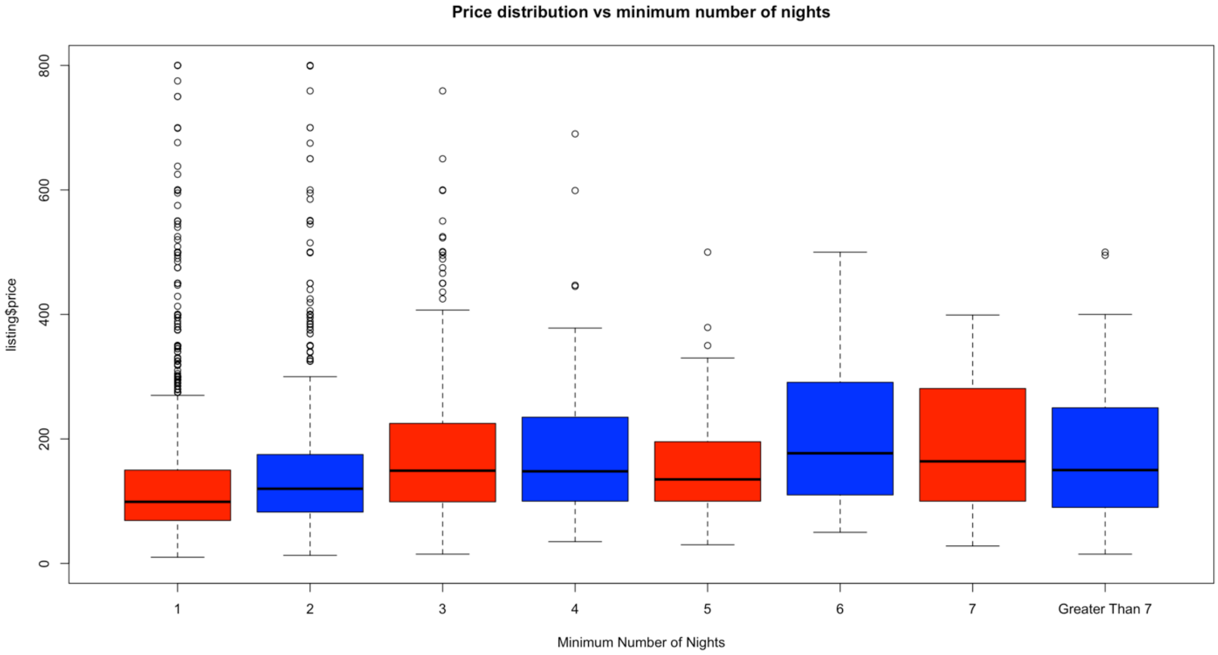
```
        Welch Two Sample t-test

data:  listing$price and listing$number_of_reviews
t = 54.399, df = 5365.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 130.0836 139.8099
sample estimates:
mean of x mean of y
149.54925  14.60249
```

Since the p value is very less(<2e-16) so the null hypothesis can be rejected over the alternate hypothesis.

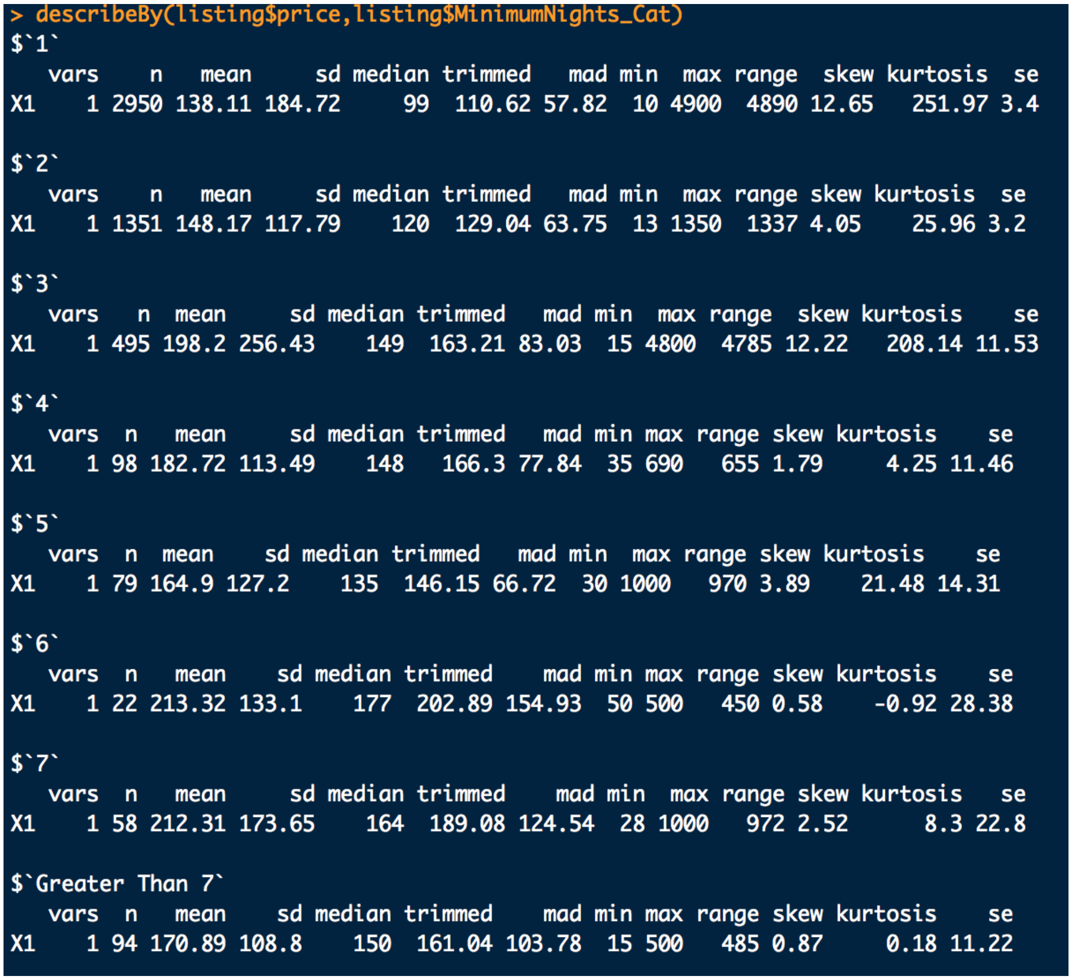
This means there is a significant different in price when there is a change in number of reviews. This can be used as an independent variable while modelling the price.

Another interesting find, a listing with a price of $4900 has 0 reviews, and a listing with 100 reviews has a very low price of $80.

4) **Minimum Nights Category vs. Price**

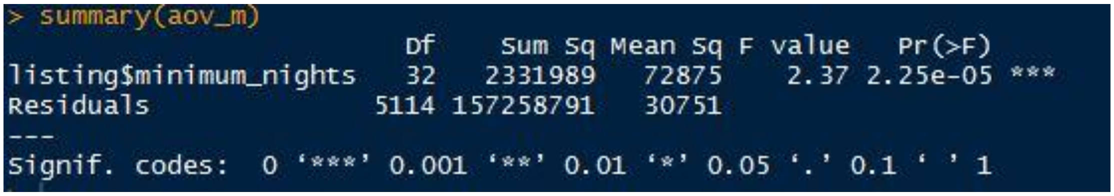


From the plot, it can be seen the highest mean is for a listing having 6 minimum number of nights.



```
> describeBy(listing$price,listing$MinimumNights_Cat)
$`1`
   vars    n   mean     sd median trimmed   mad min  max range  skew kurtosis  se
X1    1 2950 138.11 184.72     99  110.62 57.82  10 4900  4890 12.65   251.97 3.4

$`2`
   vars    n   mean     sd median trimmed   mad min  max range skew kurtosis  se
X1    1 1351 148.17 117.79    120  129.04 63.75  13 1350  1337 4.05    25.96 3.2

$`3`
   vars   n  mean     sd median trimmed   mad min  max range  skew kurtosis    se
X1    1 495 198.2 256.43    149  163.21 83.03  15 4800  4785 12.22   208.14 11.53

$`4`
   vars  n   mean     sd median trimmed   mad min max range skew kurtosis    se
X1    1 98 182.72 113.49    148   166.3 77.84  35 690   655 1.79     4.25 11.46

$`5`
   vars  n  mean    sd median trimmed   mad min  max range skew kurtosis    se
X1    1 79 164.9 127.2    135  146.15 66.72  30 1000   970 3.89    21.48 14.31

$`6`
   vars  n   mean    sd median trimmed    mad min max range skew kurtosis    se
X1    1 22 213.32 133.1    177  202.89 154.93  50 500   450 0.58    -0.92 28.38

$`7`
   vars  n   mean     sd median trimmed    mad min  max range skew kurtosis   se
X1    1 58 212.31 173.65    164  189.08 124.54  28 1000   972 2.52      8.3 22.8

$`Greater Than 7`
   vars  n   mean    sd median trimmed    mad min max range skew kurtosis    se
X1    1 94 170.89 108.8    150  161.04 103.78  15 500   485 0.87     0.18 11.22
```

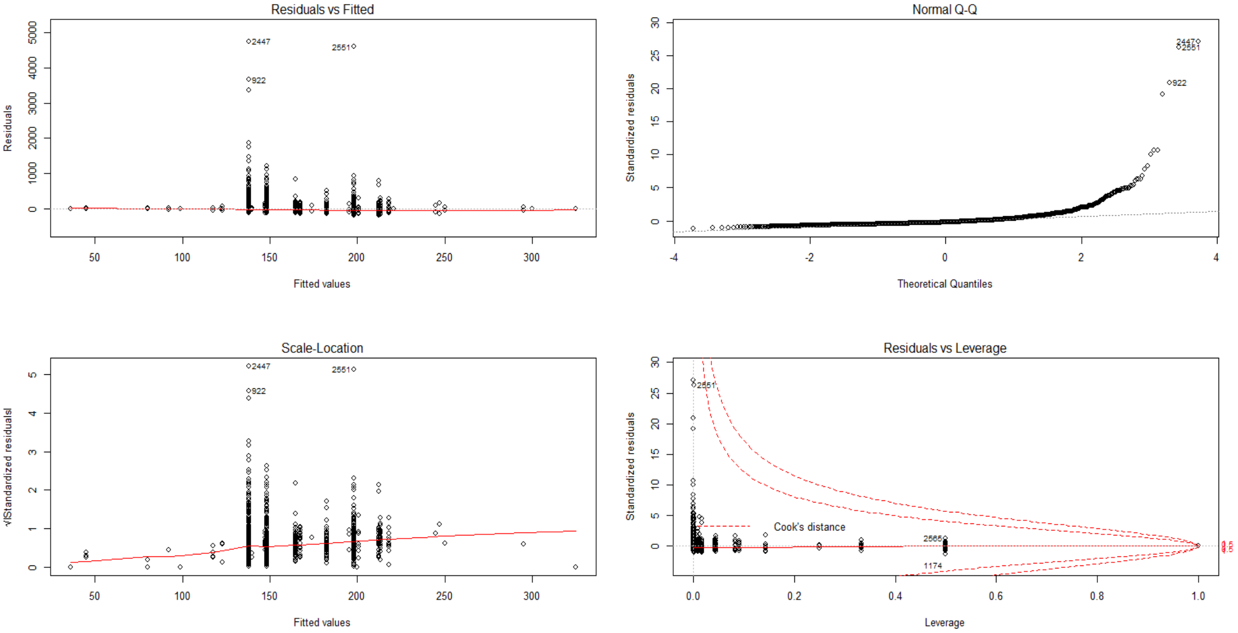
As the data is highly skewed to the right, medians can be used to compare each of the factors. A minimum of one night has a median price of $99 and the highest median price is for a minimum of 6 nights is $177.



```
> summary(aov_m)
                         Df    Sum Sq Mean Sq F value   Pr(>F)    
listing$minimum_nights   32   2331989   72875    2.37 2.25e-05 ***
Residuals              5114 157258791   30751                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p value after performing the ANOVA test is <2.25e-05. Since the p value is lesser than 0.05 we can reject the null hypothesis that there is no difference between the average price and minimum_nights of a listing.

From the below plots we can look there are outliers, from the plot (2551, 2565, 1174,922)



Analysis

Linear regression models:

Model 1 - Price vs. Independent variables

First, we tried to model the price - our dependent variable, of a listing with all the following variables: MinimumNights_Cat, number_of_reviews, neighbourhood, reviews_per_month and room_type.

The result of this model was that we got an adjusted r-squared value of 14.21%, which is very low. This could be attributed to the fact that our dependent variable is heavily right skewed so we have to improve our model further by taking into consideration log(price) rather than just price because it will be more normally distributed as seen from the univariate analysis.

Model 2 - log(price) vs independent variables

For model 2, we ran the same model with a new calculated variable, i.e., log_price against the same independent variables: MinimumNights_Cat, number_of_reviews, neighborhood, reviews_per_month and room_type.

This led to an improvement in the adjusted r-squared value from 14.21% to 52.06%. The adjusted r-squared value is not as good as expected, so we further need to fine tune our model.

As suggested by Professor Zack, we further converted the independent categorical variable, MinimumNights_Cat into two categories as follows:

1. 1 - which indicates the minimum number of nights is 1, as we can see from the univariate analysis, over 50% of the dataset is in this category
2. Greater_than_1 - which indicates the minimum number of nights greater than 1

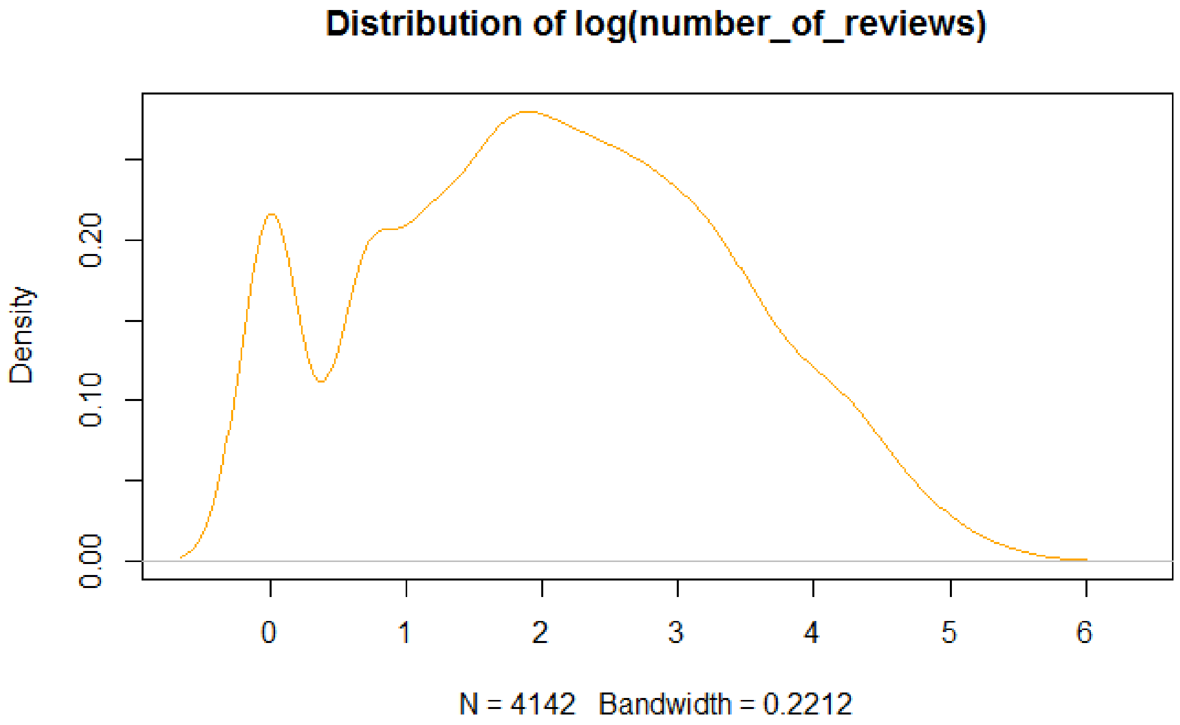
Model 3 - Converting MinimumNights_Cat to two categories

For model 3, we ran the same model with a new calculated independent variable, i.e., MinimumNights_Cat considering the above two categories as before with the same independent variables: MinimumNights_Cat, number_of_reviews, neighbourhood, reviews_per_month and room_type.

After considering this new independent variable in our model, the adjusted r-squared value has slightly decreased to 51.88%. But this is good as we now only have two categories and the adjusted R squared has not been affected much.

Model 4 - Taking log(number_of_reviews) in price model

For model 4, as we know from the univariate analysis of the independent variable number_of_reviews is heavily right skewed. We replaced number_of_reviews with log(number_of_reviews) in our linear model. Many of the observations have zero as the number of reviews, which were replaced with NA. This allows us to take log(number_of_reviews) successfully. Since log(number_of_reviews) is close to a normal distribution (as seen from the plot) there is a significant improvement in the adjusted r-squared value from 51.8% to 57.13%. We can also see that the correlation between log(number_of_reviews) and log(price) is greater than correlation between number_of_reviews and log(price).



Model 5 - Eliminating the outliers from bivariate analysis

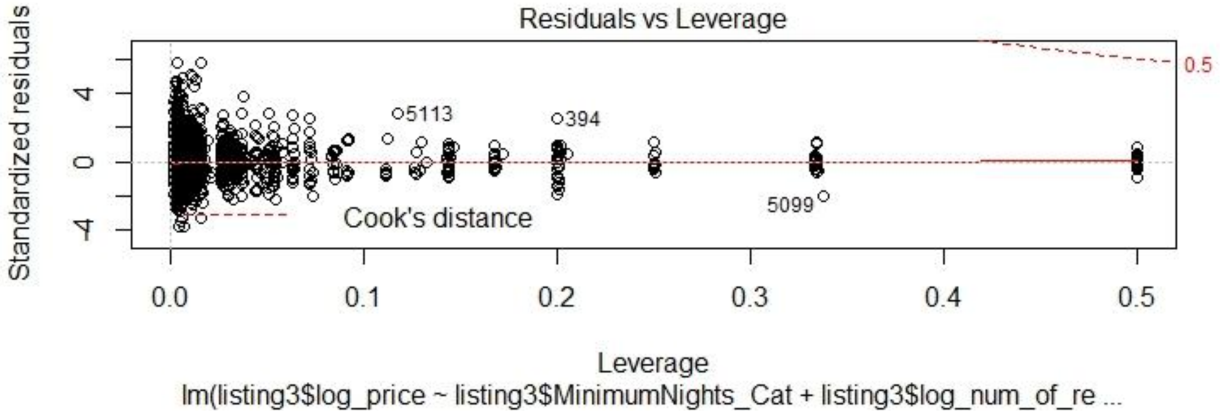
As seen from the bivariate analysis of the independent versus the dependent variable price, we can see that there are several outliers. To further fine tune our model, we have eliminated the outliers which were selected in bivariate analysis and remodeled the dependent variable.

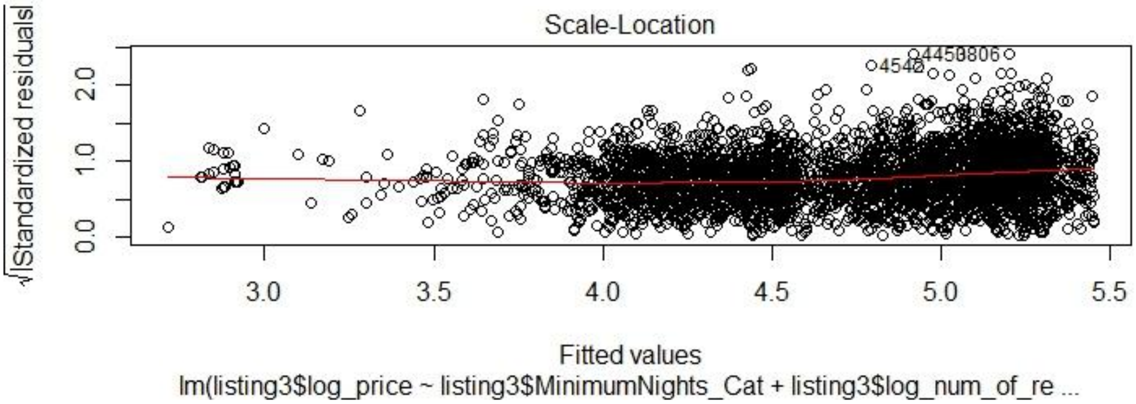
This model results in an adjusted r-squared that is almost the same, i.e. the adjusted r-squared value (57.12%).

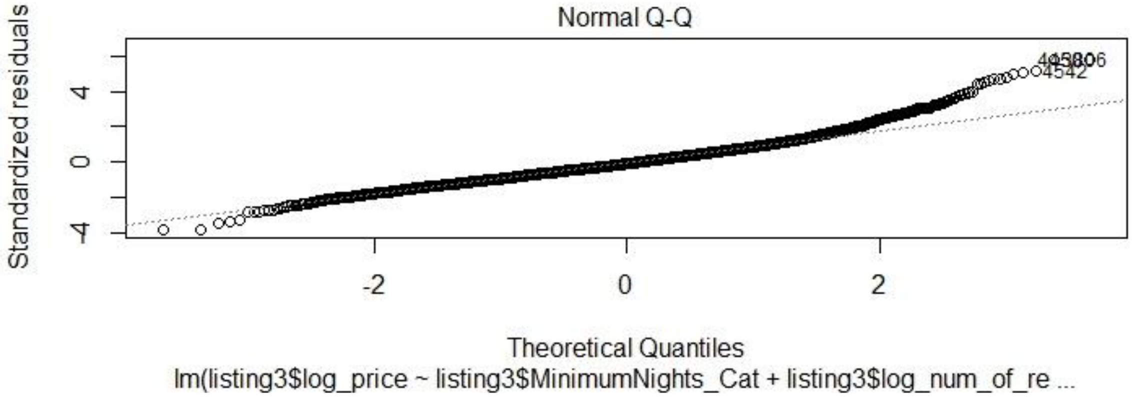
The following row numbers contain the outliers which we have eliminated:
922, 2551, 2565, 1174 ,2447, 1666, 2176, 2552

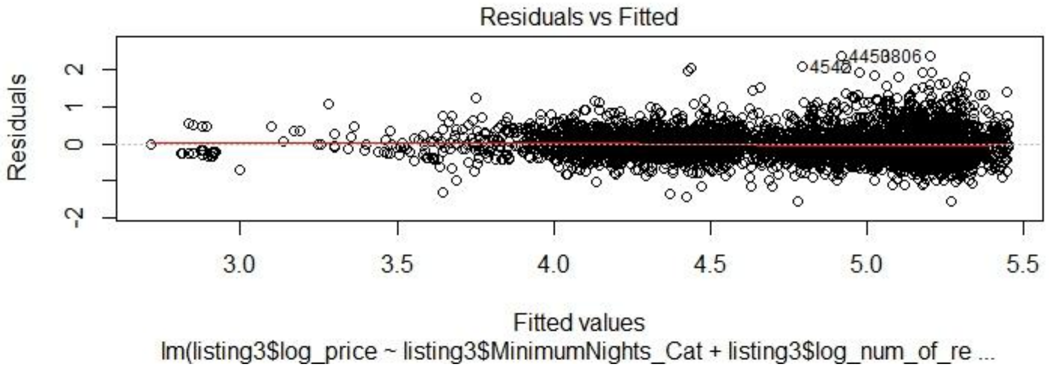
Since there isn't much difference in adjusted r-squared values after removing the outliers, we stop at this linear regression model.

Final model plots are as follows:









Conclusion:

The adjusted R-Squared score is 57.12% as seen in the final model. This indicates that the final model is able to explain the 57.12% variance of the price of a listing in Chicago. As seen from the linear model the factors that affect the listing price are MinimumNights_Cat, number_of_reviews, neighborhood, reviews_per_month and room_type.

Hypothesis 1:

The bivariate analysis explains that there is a significant difference in the price of a listing when there is a change in the number of reviews (p-value in t-test is very low). As indicated by the coefficient of the independent variable log_num_of_reviews (i.e. log(number_of_reviews))= **0.0043812** in our final model, it can be said, keeping all the other independent variables constant if the log(number_of_reviews) increases by 1, the log(price) will go up by 0.0043812 .

Hypothesis 2:

The bivariate analysis explains that the p-value of the ANOVA test is very low which indicates that the price across different neighbourhoods is not the same. As indicated from the last model (model5), the coefficient of neighbourhood = ”Near North Side” is **0.5470091**. This means the average difference in log(price) for a listing in Near North Side as compared to a listing which is not in the near north side neighborhood is 0.5470091, while keeping all the other independent variables constant. It can be inferred that a price of a listing is higher when it is in a neighborhood with a higher number of listings.

Hypothesis 3:

The bivariate analysis explained that the p-value of the ANOVA test is very low which indicates that price across different room_type is not the same. As indicated from our last model the coefficient of room_type = “Shared room” is **-1.0729652.** This means that average difference in log(price) for a listing for shared room is lesser by -1.07 as compared to not a shared room, while keeping all the other independent variable constant. It can be inferred that the price of a listing is lower for a shared room.