## 1. **Synopsis**

The Data was collected from [**Kaggle**](https://www.kaggle.com/). In this project we present to you exploratory data analysis, visualizations of New York Airbnb data. We focus on New York City’s data as we wish to perform an in-depth analysis on one of the most densely populated cities in the world.

In this project, we also try to predict the factors that affect the pricing of the airbnbs around New York. This includes creating different kind of models, model specification, transformation, variable selection and many more.

We carried out the project in the following steps:

* Data Cleaning and Preparation
* Data Visualization
* Modelling and Model Checking
* Finalising the Model
* Prediction using the Final Model.

The original Data Set can be found here —> [**New York AirBnB Data 2019**](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data)



## 2. **Packages Required**

#install DAAG from archived source  
if(!is.element("DAAG", installed.packages()[,1])){  
 packageurl <- "https://cran.r-project.org/src/contrib/Archive/DAAG/DAAG\_1.22.tar.gz"  
 install.packages("latticeExtra")  
 install.packages(packageurl, repos=NULL, type="source")  
}  
  
  
library(tidyr)  
library(DT)  
library(ggplot2)  
library(dplyr)  
library(tidyverse)  
library(kableExtra)  
library(lubridate)  
library(readxl)  
library(highcharter)  
library(lubridate)  
library(scales)  
library(RColorBrewer)  
library(wesanderson)  
library(plotly)  
library(shiny)  
library(readxl)  
library(readr)  
library(choroplethr)  
library(choroplethrMaps)  
library(GGally)  
library(zoo)  
library(scales)  
library(ggmap)  
library(stringr)  
library(gridExtra)  
library(caret)  
library(treemap)  
library(psych)  
library(DAAG)  
library(leaps)  
library(corrplot)  
library(glmnet)

Package

Description

library(tidyr)

For changing the layout of your data sets, to convert data into the tidy format

library(DT)

For HTML display of data

library(ggplot2)

For customizable graphical representation

library(dplyr)

For data manipulation

library(tidyverse)

Collection of R packages designed for data science that works harmoniously with other packages

library(kableExtra)

To display table in a fancy way

library(lubridate)

Lubridate makes it easier to do the things R does with date-times and possible to do the things R does not

library(readxl)

The readxl package makes it easy to get data out of Excel and into R

library(highcharter)

Highcharter is a R wrapper for Highcharts javascript libray and its modules

library(scales)

The idea of the scales package is to implement scales in a way that is graphics system agnostic

library(RColorBrewer)

RColorBrewer is an R package that allows users to create colourful graphs with pre-made color palettes that visualize data in a clear and distinguishable manner

library(wesanderson)

A Wes Anderson is color palette for R

library(plotly)

Plotly’s R graphing library makes interactive, publication-quality graphs

library(shiny)

Shiny is an R package that makes it easy to build interactive web apps straight from R

## 3. **Data Preparation**

### 3.1 Loading and Reading the Data

#### Summary and Glimpse of the Data

summary(airbnb\_data)

## id name   
## Min. : 2539 Hillside Hotel : 18   
## 1st Qu.: 9471945 Home away from home : 17   
## Median :19677284 : 16   
## Mean :19017143 New york Multi-unit building : 16   
## 3rd Qu.:29152178 Brooklyn Apartment : 12   
## Max. :36487245 Loft Suite @ The Box House Hotel: 11   
## (Other) :48805   
## host\_id host\_name neighbourhood\_group  
## Min. : 2438 Michael : 417 Bronx : 1091   
## 1st Qu.: 7822033 David : 403 Brooklyn :20104   
## Median : 30793816 Sonder (NYC): 327 Manhattan :21661   
## Mean : 67620011 John : 294 Queens : 5666   
## 3rd Qu.:107434423 Alex : 279 Staten Island: 373   
## Max. :274321313 Blueground : 232   
## (Other) :46943   
## neighbourhood latitude longitude   
## Williamsburg : 3920 Min. :40.50 Min. :-74.24   
## Bedford-Stuyvesant: 3714 1st Qu.:40.69 1st Qu.:-73.98   
## Harlem : 2658 Median :40.72 Median :-73.96   
## Bushwick : 2465 Mean :40.73 Mean :-73.95   
## Upper West Side : 1971 3rd Qu.:40.76 3rd Qu.:-73.94   
## Hell's Kitchen : 1958 Max. :40.91 Max. :-73.71   
## (Other) :32209   
## room\_type price minimum\_nights   
## Entire home/apt:25409 Min. : 0.0 Min. : 1.00   
## Private room :22326 1st Qu.: 69.0 1st Qu.: 1.00   
## Shared room : 1160 Median : 106.0 Median : 3.00   
## Mean : 152.7 Mean : 7.03   
## 3rd Qu.: 175.0 3rd Qu.: 5.00   
## Max. :10000.0 Max. :1250.00   
##   
## number\_of\_reviews last\_review reviews\_per\_month  
## Min. : 0.00 :10052 Min. : 0.010   
## 1st Qu.: 1.00 2019-06-23: 1413 1st Qu.: 0.190   
## Median : 5.00 2019-07-01: 1359 Median : 0.720   
## Mean : 23.27 2019-06-30: 1341 Mean : 1.373   
## 3rd Qu.: 24.00 2019-06-24: 875 3rd Qu.: 2.020   
## Max. :629.00 2019-07-07: 718 Max. :58.500   
## (Other) :33137 NA's :10052   
## calculated\_host\_listings\_count availability\_365  
## Min. : 1.000 Min. : 0.0   
## 1st Qu.: 1.000 1st Qu.: 0.0   
## Median : 1.000 Median : 45.0   
## Mean : 7.144 Mean :112.8   
## 3rd Qu.: 2.000 3rd Qu.:227.0   
## Max. :327.000 Max. :365.0   
##

glimpse(airbnb\_data)

## Observations: 48,895  
## Variables: 16  
## $ id <int> 2539, 2595, 3647, 3831, 5022, 5...  
## $ name <fct> "Clean & quiet apt home by the ...  
## $ host\_id <int> 2787, 2845, 4632, 4869, 7192, 7...  
## $ host\_name <fct> John, Jennifer, Elisabeth, Lisa...  
## $ neighbourhood\_group <fct> Brooklyn, Manhattan, Manhattan,...  
## $ neighbourhood <fct> Kensington, Midtown, Harlem, Cl...  
## $ latitude <dbl> 40.64749, 40.75362, 40.80902, 4...  
## $ longitude <dbl> -73.97237, -73.98377, -73.94190...  
## $ room\_type <fct> Private room, Entire home/apt, ...  
## $ price <int> 149, 225, 150, 89, 80, 200, 60,...  
## $ minimum\_nights <int> 1, 1, 3, 1, 10, 3, 45, 2, 2, 1,...  
## $ number\_of\_reviews <int> 9, 45, 0, 270, 9, 74, 49, 430, ...  
## $ last\_review <fct> 2018-10-19, 2019-05-21, , 2019-...  
## $ reviews\_per\_month <dbl> 0.21, 0.38, NA, 4.64, 0.10, 0.5...  
## $ calculated\_host\_listings\_count <int> 6, 2, 1, 1, 1, 1, 1, 1, 1, 4, 1...  
## $ availability\_365 <int> 365, 355, 365, 194, 0, 129, 0, ...

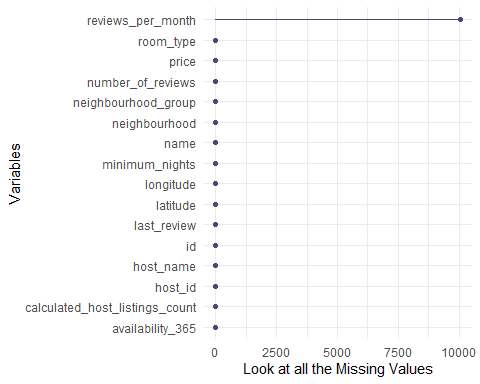
#### Checking for NA

summary(is.na(airbnb\_data))

## id name host\_id host\_name   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:48895 FALSE:48895 FALSE:48895 FALSE:48895   
##   
## neighbourhood\_group neighbourhood latitude longitude   
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:48895 FALSE:48895 FALSE:48895 FALSE:48895   
##   
## room\_type price minimum\_nights number\_of\_reviews  
## Mode :logical Mode :logical Mode :logical Mode :logical   
## FALSE:48895 FALSE:48895 FALSE:48895 FALSE:48895   
##   
## last\_review reviews\_per\_month calculated\_host\_listings\_count  
## Mode :logical Mode :logical Mode :logical   
## FALSE:48895 FALSE:38843 FALSE:48895   
## TRUE :10052   
## availability\_365  
## Mode :logical   
## FALSE:48895   
##

We visualise the number of missings in each variable using naniar gg\_miss\_var

naniar::gg\_miss\_var(airbnb\_data) +  
 theme\_minimal()+  
 labs(y = "Look at all the Missing Values")



### 3.2 Data Cleaning

### 3.3 Cleaned Dataset

The final cleaned dataset can be found below in an interactive table.

datatable(head(airbnb\_data, 20), class = 'cell-border stripe')

### 3.4 Summary of Variables

#Reading the variable summary excel File  
var\_sum <- read\_excel("variable\_summary.xlsx")  
  
kable(var\_sum) %>%  
 kable\_styling(bootstrap\_options = c("striped", "hover", "condensed", "responsive"), full\_width = F, fixed\_thead = T, )

Variable

Class

Description

anime\_id

int

anime ID (as in <https://myanimelist.net/anime/animeID>)

anime\_name

character

anime title - extracted from the site.

anime\_type

factor

anime type (e.g. TV, Movie, OVA)

source

character

source of anime (i.e original, manga, game, music, visual novel etc.)

producers

character

producers

genre

factor

genre

studio

character

studio

no\_of\_episodes

int

number of episodes

airing\_status

character

True/False- is still airing?

start\_date

date

start date

episode\_duration

character

per episode duration or entire duration, text string

MPAA\_rating

factor

age rating

viwers\_rating

num

score (higher = better)

rated\_by\_no\_of\_viewers

int

number of users that scored

rankings

int

rank - weight according to MyAnimeList formula

popularity\_index

int

based on how many members/users have the respective anime in their list

wishlisted\_members

int

number members that added this anime in their list

favorites

int

number members that favorites these in their list

premiered\_season

factor

the season the show first premiered

premiered\_year

num

the year the show first premiered

broadcast\_day

factor

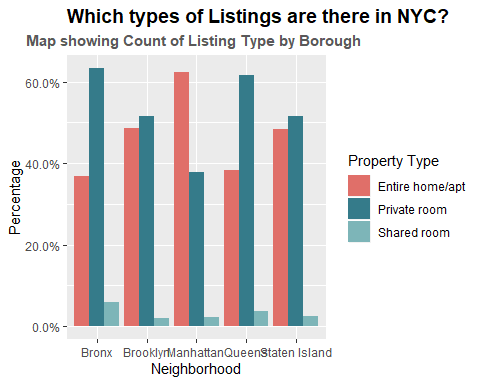
the day of the week it gets broadcasted

broadcast\_time

character

the time of the day it gets broadcasted

## 4. **Exploratory Data Analysis**



## 5. **Modelling**

### 5.1 Data Splitting

Training set will be 70% percent of the original data. Objects with price equal to 0 will be ommited since price can’t be 0 (faulty records). They would make predictive models significantly weaker.

airbnb\_data <- airbnb\_data %>% mutate(id = row\_number())  
  
airbnb\_train <- airbnb\_data %>% sample\_frac(.7) %>% filter(price > 0)  
  
airbnb\_test <- anti\_join(airbnb\_data, airbnb\_train, by = 'id') %>% filter(price > 0)

# sanity check  
nrow(airbnb\_train) + nrow(airbnb\_test) == nrow(airbnb\_data %>% filter(price > 0))

## [1] TRUE

### 5.2 Model Building Process

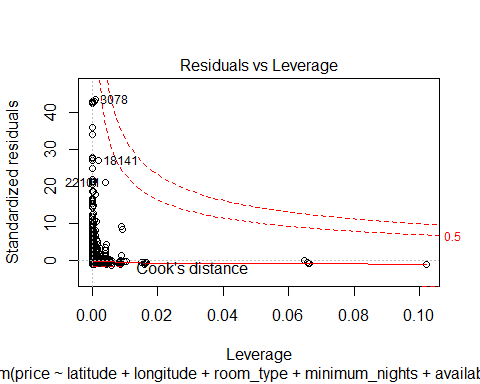
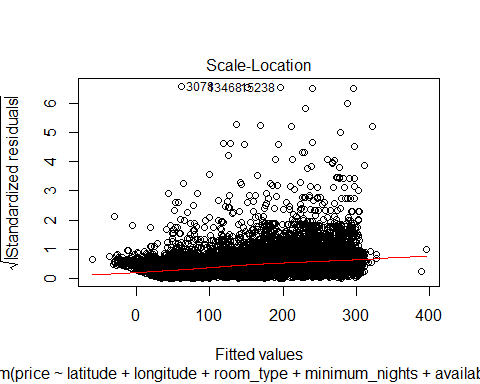
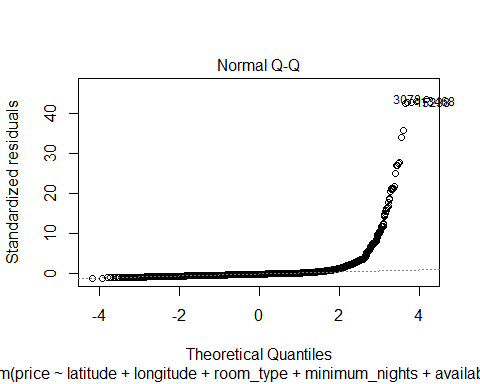
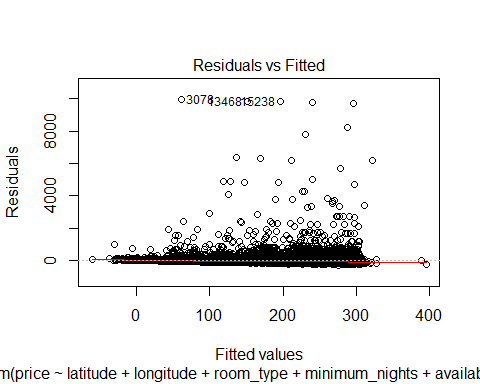
#### 5.2.1 1st Linear Regression Model

airbnb\_model\_1 <- lm (price ~ latitude + longitude + room\_type + minimum\_nights + availability\_365 + neighbourhood\_group, data = airbnb\_train)  
  
summary(airbnb\_model\_1)

##   
## Call:  
## lm(formula = price ~ latitude + longitude + room\_type + minimum\_nights +   
## availability\_365 + neighbourhood\_group, data = airbnb\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -263.8 -62.3 -24.7 14.6 9938.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2.949e+04 3.893e+03 -7.574 3.71e-14  
## latitude -1.998e+02 3.769e+01 -5.301 1.16e-07  
## longitude -5.114e+02 4.360e+01 -11.728 < 2e-16  
## room\_typePrivate room -1.048e+02 2.592e+00 -40.445 < 2e-16  
## room\_typeShared room -1.373e+02 8.387e+00 -16.371 < 2e-16  
## minimum\_nights 7.824e-02 5.933e-02 1.319 0.18725  
## availability\_365 1.634e-01 9.699e-03 16.842 < 2e-16  
## neighbourhood\_groupBrooklyn -3.487e+01 1.062e+01 -3.284 0.00103  
## neighbourhood\_groupManhattan 2.498e+01 9.651e+00 2.589 0.00964  
## neighbourhood\_groupQueens -5.065e+00 1.025e+01 -0.494 0.62111  
## neighbourhood\_groupStaten Island -1.448e+02 2.007e+01 -7.214 5.53e-13  
##   
## (Intercept) \*\*\*  
## latitude \*\*\*  
## longitude \*\*\*  
## room\_typePrivate room \*\*\*  
## room\_typeShared room \*\*\*  
## minimum\_nights   
## availability\_365 \*\*\*  
## neighbourhood\_groupBrooklyn \*\*   
## neighbourhood\_groupManhattan \*\*   
## neighbourhood\_groupQueens   
## neighbourhood\_groupStaten Island \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 229.5 on 34209 degrees of freedom  
## Multiple R-squared: 0.09362, Adjusted R-squared: 0.09336   
## F-statistic: 353.3 on 10 and 34209 DF, p-value: < 2.2e-16

#### Plot of the 1st Linear Regression Model

plot(airbnb\_model\_1)



#### 5.2.2 **Transformation** - 2nd Linear Regression Model -

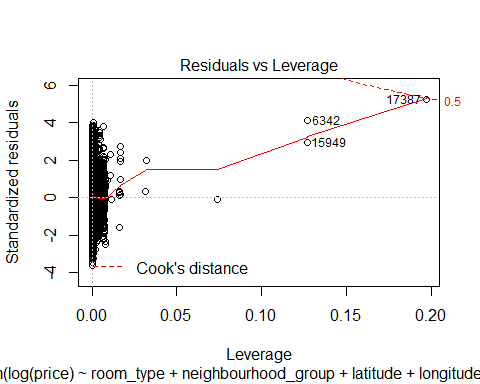
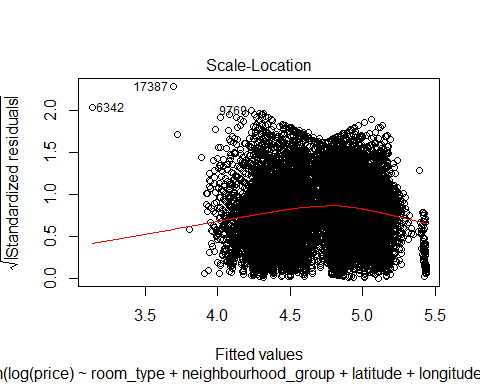
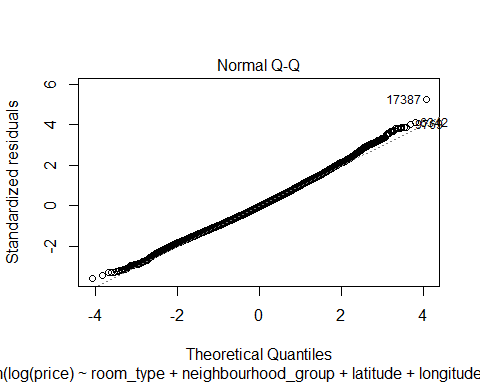
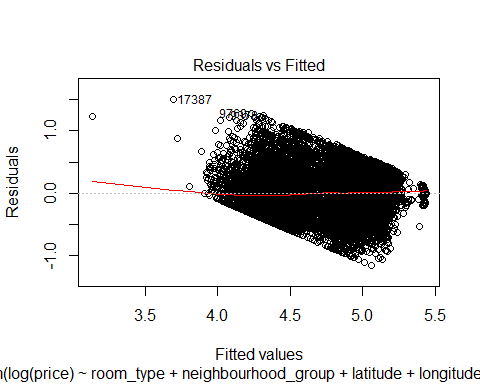
airbnb\_data\_2 <- airbnb\_train %>%   
 filter(price < quantile(airbnb\_train$price, 0.9) & price > quantile(airbnb\_train$price, 0.1)) %>%   
 drop\_na()

airbnb\_model\_2 <- lm(log(price) ~ room\_type + neighbourhood\_group + latitude + longitude + number\_of\_reviews + availability\_365   
 + reviews\_per\_month + calculated\_host\_listings\_count + minimum\_nights, data = airbnb\_data\_2)  
  
# Summarize the results  
summary(airbnb\_model\_2)

##   
## Call:  
## lm(formula = log(price) ~ room\_type + neighbourhood\_group + latitude +   
## longitude + number\_of\_reviews + availability\_365 + reviews\_per\_month +   
## calculated\_host\_listings\_count + minimum\_nights, data = airbnb\_data\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.14759 -0.22319 -0.01558 0.20590 1.50107   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.272e+02 6.777e+00 -18.765 < 2e-16  
## room\_typePrivate room -5.388e-01 4.418e-03 -121.961 < 2e-16  
## room\_typeShared room -6.695e-01 2.089e-02 -32.055 < 2e-16  
## neighbourhood\_groupBrooklyn -5.243e-02 1.974e-02 -2.656 0.00790  
## neighbourhood\_groupManhattan 1.483e-01 1.817e-02 8.164 3.43e-16  
## neighbourhood\_groupQueens 3.568e-02 1.914e-02 1.864 0.06238  
## neighbourhood\_groupStaten Island -6.195e-01 3.630e-02 -17.069 < 2e-16  
## latitude -5.859e-01 6.524e-02 -8.980 < 2e-16  
## longitude -2.108e+00 7.604e-02 -27.721 < 2e-16  
## number\_of\_reviews -1.505e-04 5.320e-05 -2.829 0.00467  
## availability\_365 3.242e-04 1.808e-05 17.935 < 2e-16  
## reviews\_per\_month 4.477e-04 1.581e-03 0.283 0.77708  
## calculated\_host\_listings\_count 5.133e-04 8.757e-05 5.862 4.64e-09  
## minimum\_nights -1.230e-03 1.146e-04 -10.738 < 2e-16  
##   
## (Intercept) \*\*\*  
## room\_typePrivate room \*\*\*  
## room\_typeShared room \*\*\*  
## neighbourhood\_groupBrooklyn \*\*   
## neighbourhood\_groupManhattan \*\*\*  
## neighbourhood\_groupQueens .   
## neighbourhood\_groupStaten Island \*\*\*  
## latitude \*\*\*  
## longitude \*\*\*  
## number\_of\_reviews \*\*   
## availability\_365 \*\*\*  
## reviews\_per\_month   
## calculated\_host\_listings\_count \*\*\*  
## minimum\_nights \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3192 on 22107 degrees of freedom  
## Multiple R-squared: 0.4941, Adjusted R-squared: 0.4938   
## F-statistic: 1661 on 13 and 22107 DF, p-value: < 2.2e-16

#### Plot of the Transformed Linear Regression Model

plot(airbnb\_model\_2)

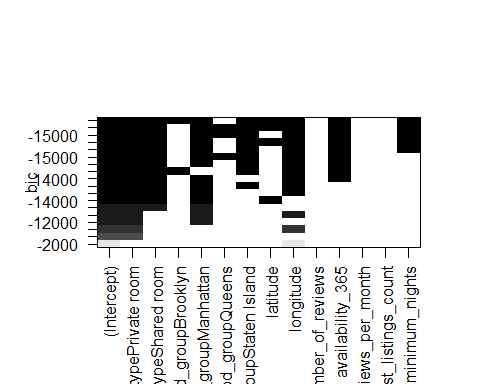


#### 5.2.3 Model Building by **Variable Selection** Method

#### **Best Subset Regression Method**

best\_fit\_model <- regsubsets (log(price) ~ room\_type + neighbourhood\_group + latitude + longitude + number\_of\_reviews + availability\_365   
 + reviews\_per\_month + calculated\_host\_listings\_count + minimum\_nights, data = airbnb\_data\_2, nbest = 2, nvmax = 9)  
  
summary(best\_fit\_model)

plot(best\_fit\_model, scale="bic")



#### Model Building by **Stepwise Regression with AIC/BIC** (direction = forward/backward/both)

#### **stepwise selection using AIC** (Direction = “both”)

null <- lm(log(price)~1, data = airbnb\_train)  
full <- lm(log(price) ~ room\_type + neighbourhood\_group + latitude + longitude + number\_of\_reviews + availability\_365   
 + reviews\_per\_month + calculated\_host\_listings\_count + minimum\_nights, data = airbnb\_train)  
  
step(null, scope =list(lower=null, upper= full), direction = "both")

## Start: AIC=-24586.22  
## log(price) ~ 1

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, :  
## using the 27193/34220 rows from a combined fit

## Df Sum of Sq RSS AIC  
## + room\_type 2 4885.8 7087.2 -36560  
## + neighbourhood\_group 4 1516.0 10457.0 -25978  
## + longitude 1 1279.8 10693.2 -25377  
## + calculated\_host\_listings\_count 1 116.2 11856.7 -22568  
## + availability\_365 1 78.2 11894.7 -22481  
## + latitude 1 52.3 11920.6 -22422  
## + reviews\_per\_month 1 17.8 11955.1 -22343  
## + number\_of\_reviews 1 8.3 11964.6 -22322  
## + minimum\_nights 1 2.2 11970.7 -22308  
## <none> 11972.9 -22305  
##   
## Step: AIC=-41219.7  
## log(price) ~ room\_type

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, :  
## using the 27193/34220 rows from a combined fit

## Df Sum of Sq RSS AIC  
## <none> 10258.3 -41220  
## + neighbourhood\_group 4 855.8 6231.4 -40051  
## + longitude 1 559.3 6527.9 -38793  
## + availability\_365 1 118.1 6969.1 -37015  
## + latitude 1 80.4 7006.8 -36868  
## + calculated\_host\_listings\_count 1 34.0 7053.2 -36688  
## + minimum\_nights 1 11.2 7075.9 -36601  
## + number\_of\_reviews 1 3.1 7084.1 -36570  
## + reviews\_per\_month 1 1.1 7086.1 -36562  
## - room\_type 2 6422.8 16681.1 -24586

##   
## Call:  
## lm(formula = log(price) ~ room\_type, data = airbnb\_train)  
##   
## Coefficients:  
## (Intercept) room\_typePrivate room room\_typeShared room   
## 5.1436 -0.8448 -1.1871

#### **stepwise selection using BIC** (Direction = “both”)

null <- lm(log(price)~1, data = airbnb\_train)  
full <- lm(log(price) ~ room\_type + neighbourhood\_group + latitude + longitude + number\_of\_reviews + availability\_365   
 + reviews\_per\_month + calculated\_host\_listings\_count + minimum\_nights, data = airbnb\_train)  
  
n=dim(airbnb\_train[1])  
step(null, scope =list(lower=null, upper= full), k=log(n), direction = "both")

## Start: AIC=-24577.78  
## log(price) ~ 1

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, :  
## using the 27193/34220 rows from a combined fit

## Df Sum of Sq RSS AIC  
## + room\_type 2 4885.8 7087.2 -36566  
## + neighbourhood\_group 4 1516.0 10457.0 -25936  
## + longitude 1 1279.8 10693.2 -25360  
## + calculated\_host\_listings\_count 1 116.2 11856.7 -22551  
## + availability\_365 1 78.2 11894.7 -22464  
## + latitude 1 52.3 11920.6 -22426  
## + reviews\_per\_month 1 17.8 11955.1 -22347  
## + number\_of\_reviews 1 8.3 11964.6 -22326  
## + minimum\_nights 1 2.2 11970.7 -22312  
## <none> 11972.9 -22296  
##   
## Step: AIC=-41194.38  
## log(price) ~ room\_type

## Warning in add1.lm(fit, scope$add, scale = scale, trace = trace, k = k, :  
## using the 27193/34220 rows from a combined fit

## Warning in k \* dfs: longer object length is not a multiple of shorter  
## object length

## Df Sum of Sq RSS AIC  
## <none> 10258.3 -41194  
## + neighbourhood\_group 4 855.8 6231.4 -40065  
## + longitude 1 559.3 6527.9 -38801  
## + availability\_365 1 118.1 6969.1 -37023  
## + latitude 1 80.4 7006.8 -36834  
## + calculated\_host\_listings\_count 1 34.0 7053.2 -36696  
## + minimum\_nights 1 11.2 7075.9 -36567  
## + number\_of\_reviews 1 3.1 7084.1 -36536  
## + reviews\_per\_month 1 1.1 7086.1 -36528  
## - room\_type 2 6422.8 16681.1 -24588

##   
## Call:  
## lm(formula = log(price) ~ room\_type, data = airbnb\_train)  
##   
## Coefficients:  
## (Intercept) room\_typePrivate room room\_typeShared room   
## 5.1436 -0.8448 -1.1871

## 6. **Conclusion**