Regression Analysis on Airbnb Price in Chicago using R

```
Airbnb <- read.csv("C:/Users/vedan/OneDrive/Desktop/Me/BA Tools 2 HW/airbnb2019.csv")
summary(Airbnb)
```

```
##   ListingMonth    host_total_listings  accommodates      bathrooms
##  Min.   : 0.300   Min.   :   0.00     Min.   : 1.000   Min.   : 0.000
##  1st Qu.: 2.500   1st Qu.:   1.00     1st Qu.: 2.000   1st Qu.: 1.000
##  Median : 4.400   Median :   2.00     Median : 4.000   Median : 1.000
##  Mean   : 4.375   Mean   :  50.76     Mean   : 4.708   Mean   : 1.396
##  3rd Qu.: 6.100   3rd Qu.:  10.00     3rd Qu.: 6.000   3rd Qu.: 2.000
##  Max.   :11.600   Max.   :1283.00     Max.   :32.000   Max.   :11.000
##     bedrooms           beds        guests_included  minimum_nights
##  Min.   : 0.000   Min.   : 0.000   Min.   : 1.000   Min.   :  1.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:  1.000
##  Median : 2.000   Median : 2.000   Median : 2.000   Median :  2.000
##  Mean   : 1.771   Mean   : 2.423   Mean   : 2.532   Mean   :  3.771
##  3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 4.000   3rd Qu.:  2.000
##  Max.   :12.000   Max.   :32.000   Max.   :16.000   Max.   :365.000
##  number_of_reviews review_scores_rating reviews_per_month PricePerNight
##  Min.   :  1.00    Min.   : 20.00       Min.   : 0.020    Min.   :  5.0
##  1st Qu.: 10.00    1st Qu.: 94.00       1st Qu.: 0.850    1st Qu.: 69.0
##  Median : 29.00    Median : 97.00       Median : 1.960    Median :106.0
##  Mean   : 51.82    Mean   : 95.51       Mean   : 2.368    Mean   :146.5
##  3rd Qu.: 70.00    3rd Qu.: 99.00       3rd Qu.: 3.433    3rd Qu.:174.0
##  Max.   :615.00    Max.   :100.00       Max.   :25.100    Max.   :981.0
```

```
str(Airbnb)
```

```
## 'data.frame':    5000 obs. of  12 variables:
##  $ ListingMonth        : num  3.5 2 2.5 5.6 3.6 3.1 6.2 2.4 6.1 1.8 ...
##  $ host_total_listings : int  1 1 22 2 596 0 33 2 896 4 ...
##  $ accommodates        : int  6 5 4 4 6 2 6 12 2 1 ...
##  $ bathrooms           : num  1.5 1 1 1 1.5 1.5 1 3.5 1 1 ...
##  $ bedrooms            : int  3 2 2 1 2 1 1 5 1 3 ...
##  $ beds                : int  3 2 2 1 4 1 4 6 1 1 ...
##  $ guests_included     : int  4 1 1 2 1 1 2 8 2 1 ...
##  $ minimum_nights      : int  2 3 1 32 2 1 1 1 2 1 ...
##  $ number_of_reviews   : int  153 46 3 12 16 44 88 81 20 47 ...
##  $ review_scores_rating: int  100 96 100 100 99 99 97 95 95 93 ...
##  $ reviews_per_month   : num  3.79 2.11 3 1.52 3.27 6 5.7 3.02 0.98 2.3 ...
##  $ PricePerNight       : int  190 89 501 104 399 42 108 170 132 48 ...
```

```
# Converting to categorical variable
Airbnb$hostclass=ifelse(Airbnb$host_total_listings < 3,"1",
                        ifelse(Airbnb$host_total_listings >= 20,"3","2"))

table(Airbnb$hostclass)
```

```
##
##    1    2    3
## 2583 1474  943
```

Split the "indata" into 70% of the train and 30% of the test data set.

Modify the following code and use your student ID number as the seed number.

```
indata <- Airbnb
set.seed(2190070)

# Dividing the data into 70/30 for train and test data
train_ind <- sample(nrow(indata),round(0.7*nrow(indata)))
train <- indata[train_ind,]
test <- indata[-train_ind,]

# Linear Regression Model
r1 <-lm(PricePerNight~accommodates+bathrooms+bedrooms+beds, data = train)
e1 <- residuals(r1,newdata=test)
yhat1 <- predict(r1,newdata = test)
mse1 <- mean(e1*2)
rmse1 <- mean(mse1*0.5)
print(rmse1)
```

```
## [1] -1.345194e-15
```

```
print(mse1)
```

```
## [1] -2.690388e-15
```

```
summary(r1)
```

```
##
## Call:
## lm(formula = PricePerNight ~ accommodates + bathrooms + bedrooms +
##     beds, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -331.66  -52.43  -22.22   23.25  752.75
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.9942     4.0218   3.231  0.00125 **
## accommodates  15.2934     1.2149  12.588  < 2e-16 ***
## bathrooms     26.5506     2.9206   9.091  < 2e-16 ***
## bedrooms      12.8740     2.8588   4.503 6.91e-06 ***
## beds           0.9253     1.8235   0.507  0.61188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 105.7 on 3495 degrees of freedom
## Multiple R-squared:  0.3292, Adjusted R-squared:  0.3284
## F-statistic: 428.7 on 4 and 3495 DF,  p-value: < 2.2e-16
```

```
r2 <-lm(PricePerNight~accommodates+bathrooms+bedrooms+beds+hostclass+ListingMonth, data = train)
e2 <- residuals(r2,newdata=test)
yhat2 <- predict(r2,newdata = test)
mse2 <- mean(e2*2)
rmse2 <- mean(mse2*0.5)
print(rmse2)
```

```
## [1] 4.768107e-15
```

```
print(mse2)
```

```
## [1] 9.536213e-15
```

```
summary(r2)
```

```
##
## Call:
## lm(formula = PricePerNight ~ accommodates + bathrooms + bedrooms +
##     beds + hostclass + ListingMonth, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -320.65  -50.71  -16.24   27.16  697.04
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.0846     5.4411   1.670   0.0951 .
## accommodates  14.1511     1.1900  11.892  < 2e-16 ***
## bathrooms     25.6591     2.8619   8.966  < 2e-16 ***
## bedrooms      18.1702     2.8301   6.420 1.54e-10 ***
## beds          -0.6295     1.7832  -0.353   0.7241
## hostclass2    -2.2806     4.0697  -0.560   0.5753
## hostclass3    59.7591     4.7681  12.533  < 2e-16 ***
## ListingMonth  -1.3020     0.7731  -1.684   0.0922 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 103 on 3492 degrees of freedom
## Multiple R-squared:  0.3632, Adjusted R-squared:  0.362
## F-statistic: 284.6 on 7 and 3492 DF,  p-value: < 2.2e-16
```

```
r3 <-lm(PricePerNight~accommodates+bathrooms+bedrooms+beds+hostclass+ListingMonth+guests_included+minimu
e3 <- residuals(r3,newdata=test)
yhat3 <- predict(r3,newdata = test)
mse3 <- mean(e3*2)
rmse3 <- mean(mse3*0.5)
print(rmse3)
```

```
## [1] 5.938399e-15
```

```
print(mse3)
```

```
## [1] 1.18768e-14
```

```
summary(r3)
```

```
##
## Call:
## lm(formula = PricePerNight ~ accommodates + bathrooms + bedrooms +
##     beds + hostclass + ListingMonth + guests_included + minimum_nights +
##     number_of_reviews + review_scores_rating + reviews_per_month,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -320.12  -51.45  -15.47   27.23  689.32
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.459e+02  2.947e+01  -4.952 7.68e-07 ***
## accommodates          1.605e+01  1.248e+00  12.857  < 2e-16 ***
## bathrooms             2.252e+01  2.829e+00   7.961 2.29e-15 ***
## bedrooms              1.491e+01  2.841e+00   5.248 1.63e-07 ***
## beds                 -6.928e-01  1.766e+00  -0.392  0.69488
## hostclass2            9.682e-01  4.048e+00   0.239  0.81096
## hostclass3            6.636e+01  4.888e+00  13.577  < 2e-16 ***
## ListingMonth         -2.120e+00  8.169e-01  -2.595  0.00949 **
## guests_included      -9.321e-01  1.050e+00  -0.888  0.37456
## minimum_nights       -1.769e-01  1.522e-01  -1.163  0.24509
## number_of_reviews    -5.494e-03  3.656e-02  -0.150  0.88055
## review_scores_rating  1.903e+00  3.044e-01   6.252 4.53e-10 ***
## reviews_per_month    -8.684e+00  1.199e+00  -7.240 5.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.2 on 3487 degrees of freedom
## Multiple R-squared:  0.3862, Adjusted R-squared:  0.3841
## F-statistic: 182.9 on 12 and 3487 DF,  p-value: < 2.2e-16
```

Stepwise regression

```
r4 <-lm(PricePerNight~accommodates+bathrooms+bedrooms+beds+hostclass+ListingMonth+guests_included+minimu
stepwise <- step(r4, direction = "both")
```

```
## Start:  AIC=32333.57
## PricePerNight ~ accommodates + bathrooms + bedrooms + beds +
##     hostclass + ListingMonth + guests_included + minimum_nights +
```

```
##     number_of_reviews + review_scores_rating + reviews_per_month
##
##                         Df Sum of Sq      RSS   AIC
## - number_of_reviews      1       231 35721368 32332
## - beds                   1      1576 35722712 32332
## - guests_included        1      8079 35729215 32332
## - minimum_nights         1     13845 35734981 32333
## <none>                               35721136 32334
## - ListingMonth           1     68995 35790131 32338
## - bedrooms               1    282089 36003225 32359
## - review_scores_rating   1    400455 36121591 32371
## - reviews_per_month      1    536906 36258042 32384
## - bathrooms              1    649221 36370357 32395
## - accommodates           1   1693386 37414522 32494
## - hostclass              2   2100459 37821596 32530
##
## Step:  AIC=32331.59
## PricePerNight ~ accommodates + bathrooms + bedrooms + beds +
##     hostclass + ListingMonth + guests_included + minimum_nights +
##     review_scores_rating + reviews_per_month
##
##                         Df Sum of Sq      RSS   AIC
## - beds                   1      1553 35722920 32330
## - guests_included        1      8138 35729506 32330
## - minimum_nights         1     13807 35735174 32331
## <none>                               35721368 32332
## + number_of_reviews      1       231 35721136 32334
## - ListingMonth           1     81225 35802592 32338
## - bedrooms               1    283040 36004408 32357
## - review_scores_rating   1    401127 36122495 32369
## - bathrooms              1    649525 36370892 32393
## - reviews_per_month      1    992222 36713589 32425
## - accommodates           1   1693159 37414527 32492
## - hostclass              2   2219309 37940676 32539
##
## Step:  AIC=32329.75
## PricePerNight ~ accommodates + bathrooms + bedrooms + hostclass +
##     ListingMonth + guests_included + minimum_nights + review_scores_rating +
##     reviews_per_month
##
##                         Df Sum of Sq      RSS   AIC
## - guests_included        1      7395 35730315 32328
## - minimum_nights         1     13699 35736619 32329
## <none>                               35722920 32330
## + beds                   1      1553 35721368 32332
## + number_of_reviews      1       208 35722712 32332
## - ListingMonth           1     81788 35804708 32336
## - bedrooms               1    296292 36019212 32357
## - review_scores_rating   1    401071 36123991 32367
## - bathrooms              1    658488 36381408 32392
## - reviews_per_month      1    993655 36716575 32424
## - hostclass              2   2220883 37943803 32537
## - accommodates           1   2245359 37968279 32541
##
```

```
## Step:  AIC=32328.47
## PricePerNight ~ accommodates + bathrooms + bedrooms + hostclass +
##     ListingMonth + minimum_nights + review_scores_rating + reviews_per_month
##
##                         Df Sum of Sq      RSS   AIC
## - minimum_nights         1     13013 35743328 32328
## <none>                               35730315 32328
## + guests_included        1      7395 35722920 32330
## + beds                   1       809 35729506 32330
## + number_of_reviews      1       268 35730047 32330
## - ListingMonth           1     82417 35812732 32335
## - bedrooms               1    289233 36019548 32355
## - review_scores_rating   1    400863 36131177 32366
## - bathrooms              1    661996 36392311 32391
## - reviews_per_month      1   1006243 36736558 32424
## - hostclass              2   2236766 37967080 32537
## - accommodates           1   2384377 38114692 32553
##
## Step:  AIC=32327.75
## PricePerNight ~ accommodates + bathrooms + bedrooms + hostclass +
##     ListingMonth + review_scores_rating + reviews_per_month
##
##                         Df Sum of Sq      RSS   AIC
## <none>                               35743328 32328
## + minimum_nights         1     13013 35730315 32328
## + guests_included        1      6709 35736619 32329
## + beds                   1       762 35742566 32330
## + number_of_reviews      1       226 35743102 32330
## - ListingMonth           1     89265 35832593 32334
## - bedrooms               1    286382 36029710 32354
## - review_scores_rating   1    399642 36142970 32365
## - bathrooms              1    663815 36407143 32390
## - reviews_per_month      1    995240 36738568 32422
## - hostclass              2   2226037 37969365 32535
## - accommodates           1   2408827 38152155 32554
```

```
e4 <- residuals(r4,newdata=test)
yhat4 <- predict(r4,newdata = test)
mse4 <- mean(e4*2)
rmse4 <- mean(mse4*0.5)
print(rmse4)
```

```
## [1] 5.938399e-15
```

```
print(mse4)
```

```
## [1] 1.18768e-14
```

```
summary(r4)
```

```
##
## Call:
```

```
## lm(formula = PricePerNight ~ accommodates + bathrooms + bedrooms +
##     beds + hostclass + ListingMonth + guests_included + minimum_nights +
##     number_of_reviews + review_scores_rating + reviews_per_month,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -320.12  -51.45  -15.47   27.23  689.32
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.459e+02  2.947e+01  -4.952 7.68e-07 ***
## accommodates          1.605e+01  1.248e+00  12.857  < 2e-16 ***
## bathrooms             2.252e+01  2.829e+00   7.961 2.29e-15 ***
## bedrooms              1.491e+01  2.841e+00   5.248 1.63e-07 ***
## beds                 -6.928e-01  1.766e+00  -0.392  0.69488
## hostclass2            9.682e-01  4.048e+00   0.239  0.81096
## hostclass3            6.636e+01  4.888e+00  13.577  < 2e-16 ***
## ListingMonth         -2.120e+00  8.169e-01  -2.595  0.00949 **
## guests_included      -9.321e-01  1.050e+00  -0.888  0.37456
## minimum_nights       -1.769e-01  1.522e-01  -1.163  0.24509
## number_of_reviews    -5.494e-03  3.656e-02  -0.150  0.88055
## review_scores_rating  1.903e+00  3.044e-01   6.252 4.53e-10 ***
## reviews_per_month    -8.684e+00  1.199e+00  -7.240 5.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 101.2 on 3487 degrees of freedom
## Multiple R-squared:  0.3862, Adjusted R-squared:  0.3841
## F-statistic: 182.9 on 12 and 3487 DF,  p-value: < 2.2e-16
```

MODELS MSE VALUE RMSE VALUE R1 MODEL 2.029012e-15 1.014506e-15 R2 MODEL 1.276233e-15 6.381165e-16 R3 MODEL 3.339313e-15 1.669656e-15 R4 MODEL 3.339313e-15 1.669656e-15

CONCLUSION - From the above table, R2 model has the lowest MSE value of 1.276233e-15 and R1 Model has the lowest RSME of 1.014506e-15. Comparing R1 and R2 models, R1 model has a better predictive performance since it has a lower RMSE value and MSE value of 2.029012e-15. We will not consider the R2 model since it has the highest RMSE value of 6.381165e-16, therefore the model has a lot of errors compared to the R1 model.

Hence, the R1 model gives the best prediction of the test data.