Business Analytics Assignment 2

Rakhee Moolchandani

11/15/2020

Run the following code in R-studio to create two variables X and Y

```
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
X
```

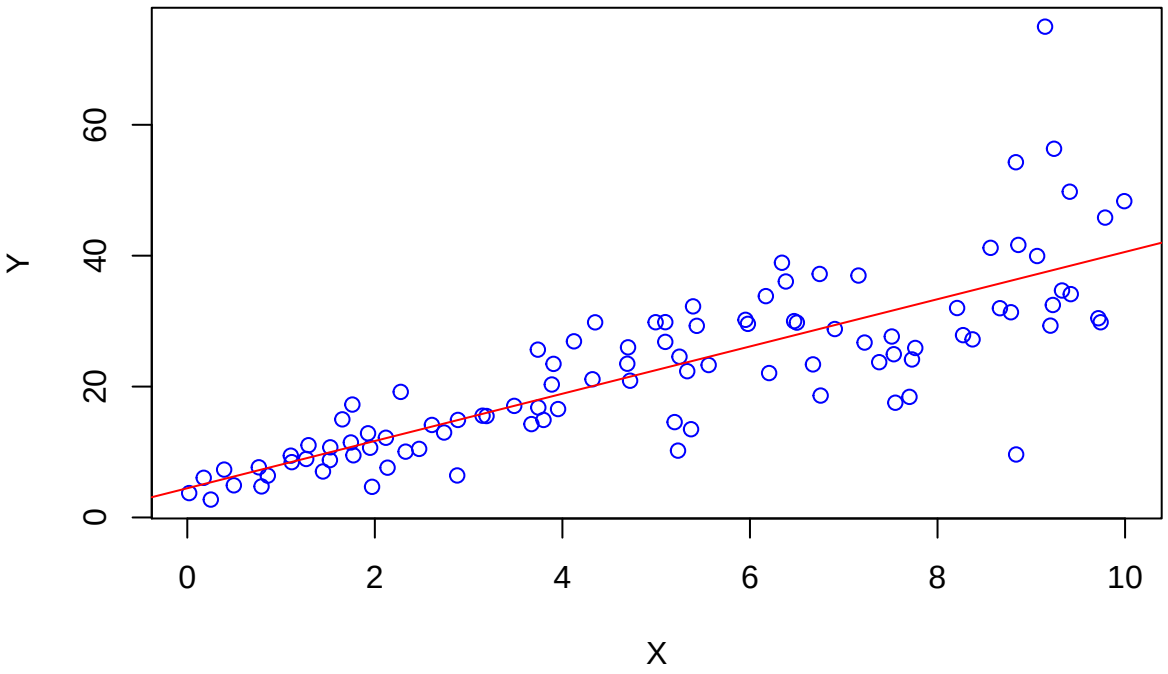
```
##   [1] 9.24242609 5.37176413 4.69195646 2.88626176 7.70088162 7.72768713
##   [7] 0.39322336 4.34905600 4.72166386 2.73833123 6.74331481 0.02020766
##  [13] 0.25093514 4.32077786 4.99391912 3.88681932 3.95375316 7.15707325
##  [19] 9.40999879 8.27229161 6.34113521 3.79867441 9.42074033 7.54993688
##  [25] 2.27611845 9.14666027 6.20445041 3.19104576 0.76288815 2.60839318
##  [31] 8.83844421 0.17528790 3.14878806 2.11763026 9.06225134 6.50003779
##  [37] 1.29273602 1.10404957 1.77131691 5.33112156 5.55950241 5.97766741
##  [43] 1.97042522 0.79169898 9.71505629 5.09695036 0.49657258 6.75366414
##  [49] 7.76260542 7.53371964 5.24834395 8.83536608 3.48626534 6.38246757
##  [55] 7.37828671 2.87824398 1.92755898 9.78701948 5.39325110 9.22954194
##  [61] 2.47248332 3.74299080 5.95257096 7.51236601 7.22214210 4.69996306
##  [67] 1.76023807 1.11417533 6.16943032 6.67186700 9.99147375 6.90462846
##  [73] 1.52084375 8.78294065 9.73919514 3.73830438 3.90523002 5.23268295
##  [79] 4.12317001 9.32749002 1.44675789 8.56598966 5.09706556 9.20411842
##  [85] 2.32821523 1.94989021 1.52485729 1.65339545 3.67003424 0.85871050
##  [91] 8.66551321 6.47087535 5.19656023 5.43282776 8.37357287 1.26816635
##  [97] 8.20879745 2.13562298 8.86173036 1.74446203
```

```
Y
```

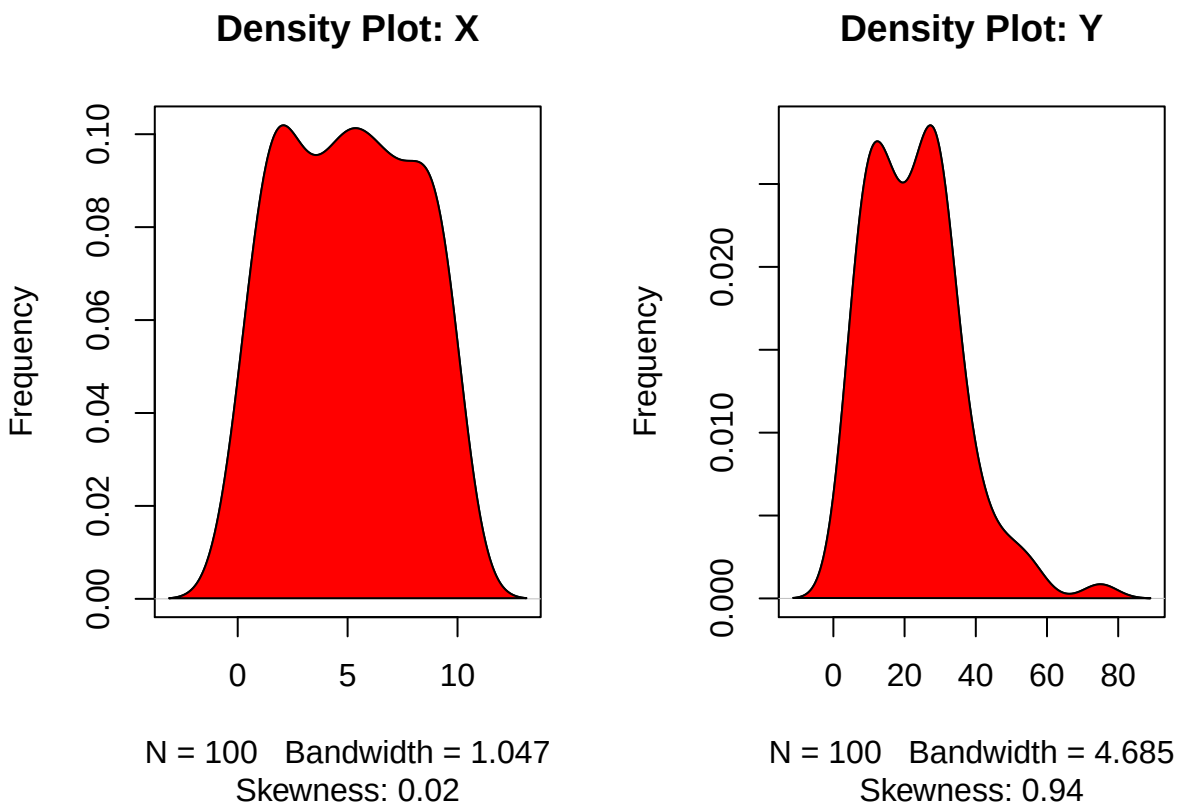
```
##   [1] 56.339338 13.481743 23.491286 14.900097 18.424419 24.142820  7.312874
##   [8] 29.813024 20.894231 12.979686 37.204723  3.722200  2.735146 21.116100
##  [15] 29.831496 20.317106 16.565079 36.964220 49.778322 27.856923 38.922793
##  [22] 14.901612 34.126732 17.533639 19.186412 74.994732 22.066461 15.501982
##  [29]  7.664396 14.133430  9.624374  6.059640 15.547511 12.178370 39.947159
##  [36] 29.774544 11.036792  9.441514  9.497847 22.348027 23.292688 29.580988
##  [43]  4.680640  4.741384 30.440391 26.829325  4.911111 18.624106 25.900178
##  [50] 24.940094 24.567106 54.280785 17.048523 36.055410 23.726551  6.415862
##  [57] 12.858711 45.821756 32.253871 32.469412 10.475566 16.800574 30.197694
##  [64] 27.650414 26.727964 25.996029 17.252984  8.450433 33.822233 23.388623
##  [71] 48.330995 28.791944  8.764633 31.355354 29.826326 25.642803 23.467743
##  [78] 10.210921 26.915454 34.680875  7.025347 41.200639 29.843081 29.309346
##  [85] 10.056939 10.650237 10.716962 14.990493 14.266479  6.390429 31.973424
##  [92] 30.002584 14.569665 29.282199 27.212813  8.937213 32.003817  7.608283
##  [99] 41.645110 11.459822
```

a) Plot Y against X. Include a screen shot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X?

```
plot(X, Y, xlab = 'X', ylab = 'Y', col = "blue")  #plot Y against X
abline(lsfit(X, Y),col = "red")
```




```
library(e1071)
par(mfrow=c(1, 2))  # divide graph area in 2 columns
plot(density(X), main="Density Plot: X", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness
polygon(density(X), col="red")
plot(density(Y), main="Density Plot: Y", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness
polygon(density(Y), col="red")
```




```
cor(X, Y) #calculate correlation between X and Y
```

```
## [1] 0.807291
```

The plot along with the line above suggests a linearly increasing relationship between the 'X' and 'Y' variables. This is a good thing, because, one of the underlying assumptions in linear regression is that the relationship between the response and predictor variables is linear and additive. Based on the plot we can fit a linear model to explain Y based on X.

b) Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model?

```
linearMod <- lm(Y ~ X)  # build linear regression model on full data
print(linearMod) # print the linear model
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611
```

```
summary(linearMod) # show the summary of the model
```

```
##
## Call:
## lm(formula = Y ~ X)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

Now that we have built the linear model, we also have established the relationship between the predictor and response in the form of a mathematical formula for Y as a function for X. For the above output, you can notice the 'Coefficients' part having two components: Intercept: 4.4655, X: 3.6108 These are also called the beta coefficients. In other words,

Equation of the Model Y = 4.4655 + 3.6108*X and Accuracy of the Model = 65.17%. Therefore 65% Variability in Y can be explained in X.

c) How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y?

```
SSYY=sum((Y-mean(Y))^2)
SSXY=sum((X-mean(X))*(Y-mean(Y)))
SSX=sum((X-mean(X))^2)
b1= SSXY/SSX
b0=mean(Y)-b1*mean(X)
Y_Estimated=X*b1+b0
Residuals= Y-Y_Estimated
SSE=sum((Residuals -mean(Residuals))^2)
r2=1-SSE/SSYY
r2
```

```
## [1] 0.6517187
```

```
cor(X,Y)^2
```

```
## [1] 0.6517187
```

The coefficient of determination, R2, is used to analyze how differences in one variable can be explained by a difference in a second variable. More specifically, R-squared gives the percentage variation in Y explained by X-variables. The range is 0 to 1 (i.e. 0% to 100% of the variation in Y can be explained by the X-variables).

The coefficient of determination, R2, is similar to the correlation coefficient, R. The correlation coefficient formula will tell how strong of a linear relationship there is between two variables. R Squared is the square of the correlation coefficient, r (hence the term r squared).Therefore,

Coefficient of Determination= (Correlation Coefficient)^2

Here, the values of the Square of the correlation coefficient is 0.6517187 which is equivalent to 65.17%

2) We will use the 'mtcars' dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset.

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a) James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question.

```
JamesCarModel = lm(mtcars$hp~mtcars$wt)
summary(JamesCarModel)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## mtcars$wt     46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

```
ChrisCarModel = lm(mtcars$hp~mtcars$mpg)
summary(ChrisCarModel)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mtcars$mpg     -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Accuracy of James Car Model which is based on the weight of the car is 43.39%

Accuracy of Chris Car Model which is based on the fuel consumption expressed in Mile per Gallon (mpg) is 60.24%

Therefore Chris is right.

b) Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22?

```
HPModel = lm(mtcars$hp~mtcars$cyl+mtcars$mpg)
summary(HPModel)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$cyl + mtcars$mpg)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## mtcars$cyl    23.979      7.346   3.264  0.00281 **
## mtcars$mpg    -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

```
anova(HPModel)
```

```
## Analysis of Variance Table
##
## Response: mtcars$hp
##            Df Sum Sq Mean Sq F value    Pr(>F)
## mtcars$cyl  1 100984  100984 69.1203 3.644e-09 ***
## mtcars$mpg  1   2374    2374  1.6249    0.2125
## Residuals  29  42369    1461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
HPModel
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$cyl + mtcars$mpg)
##
## Coefficients:
## (Intercept)   mtcars$cyl   mtcars$mpg
##      54.067       23.979       -2.775
```

```
cyl = 4
mpg = 22
hp = 54.067 + 23.979*cyl - 2.775*mpg
hp
```

```
## [1] 88.933
```

Based on the above model, the estimated Horse Power of a car with 4 cylindar and mpg of 22 is 88.933.

3) For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to instal the package, call the library and the load the dataset using the following commands

```
#install.packages('mlbench')
library(mlbench)
data(BostonHousing)
```

a) Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model?

```
BHModel = lm(medv~crim+zn+ptratio+chas, data = BostonHousing)
summary(BHModel)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

Model's Accuracy is only 35.99% which is very low and can be improved.

b) Use the estimated coefficient to answer these questions?

I. Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much?

Estimated Coefficient value of chas1 is 4.58393. Since the price is expressed in 1000 units of dollars. The increase in price will be 1000*4.58393 which will be 4583. Therefore one bound of chase river will be $4583 more expensive.

II. Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much?

Estimated Coefficient value of ptratio is -1.49367 Therefore for every unit increase in ptratio there will be 1.4937 unit decrease. One of the ptratio = 15 and other = 18. Therefore, (18-15)*1.4937 = 4.4811. If it is expressed in units of $1000 it will be $4481. Hence ptratio with 15 is $4481 more expensive than ptratio with 18.

c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer

```
head(BostonHousing)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
```

```
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

```
MyData <- BostonHousing[,1:14]
HousingModel = lm(medv~., data = MyData)
summary(HousingModel)
```

```
##
## Call:
## lm(formula = medv ~ ., data = MyData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas1        2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## b            9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

The regression output shows that the predictor variables are statistically significant when their p-values equal 0.000. On the other hand, it is not statistically significant because its p-value is greater than the usual significance level of 0.05.

Hence, In the above model, P-values of the coefficients of most of the variables is low and hence all the variables are statistically important. But the variables indus and age have p-values 0.738288 and 0.958229 respectively and therefore they are statistically less importatnt.

Also, to answer this question, based on the above model (created in Q3 part(a)) with four variables crim, zn, ptratio & chas, all the variables have low p-value and so all are statistically important.

d) Use the anova analysis and determine the order of importance of these four variables.

```
anova(BHModel)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
T=anova(BHModel)
T$Variable_Importance_Perentage=T[,2]/sum(T[,2])
T
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value     Pr(>F) Variable_Importance_Perentage
## crim        1  6440.8  6440.8 118.007 0.00000000                        0.15078
## zn          1  3554.3  3554.3  65.122 0.00000000                        0.08321
## ptratio     1  4709.5  4709.5  86.287 0.00000000                        0.11025
## chas        1   667.2   667.2  12.224 0.00051369                        0.01562
## Residuals 501 27344.5    54.6                                          0.64014
```

Order of Importance of these four variables : crim, ptratio, zn, chas.