Assignment 3

Kiran Kour

2022-11-10

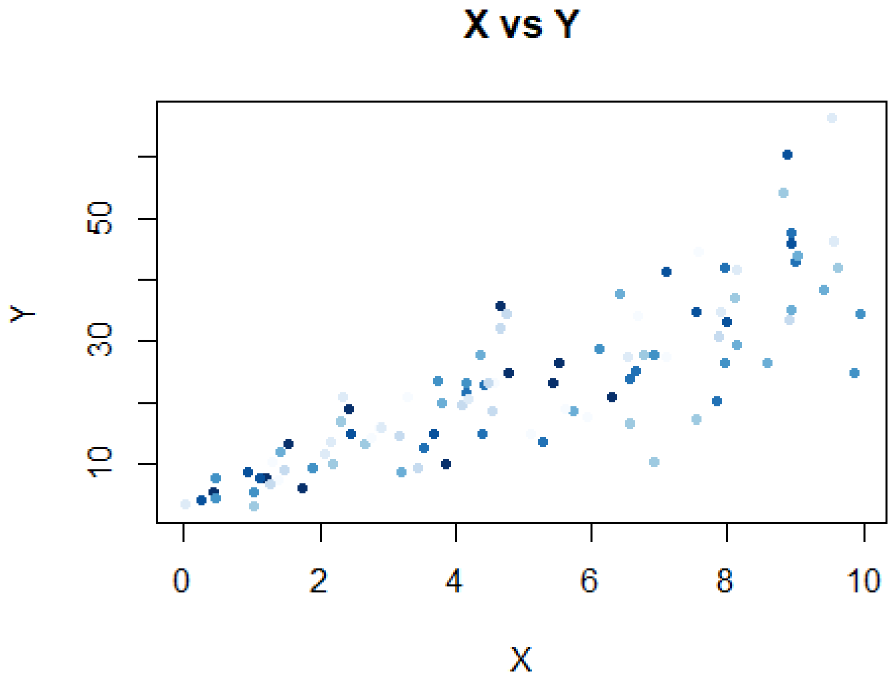
Questions

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

```
set.seed(123)

X <- runif(100)*10
Y <- X*4+3.45
Y <- rnorm(100)*0.29*Y+Y
```

a) Plot Y against X. Include a screenshot 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? (8% of total points)

```
#Plotting X vs Y
plot(X , Y, main="X vs Y",xlab="X",ylab="Y",col = blues9,pch = 20)
```



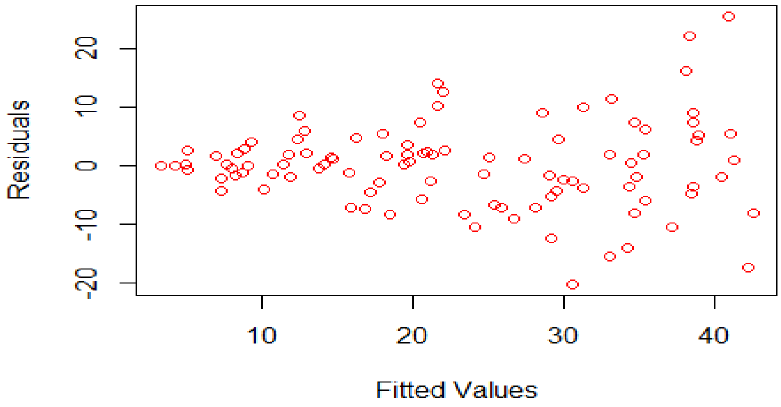
Based on the Scatter plot above, we can fit a linear model to explain Y based on X, as we can see a positive relationship between X and Y. As X increases, Y seems to increase as well.

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? (8% of total points)

```
set.seed(123)

# To run the linear model
linear_model <- lm(Y ~ X)

# To get descriptive statistics of the model
summary(linear_model)

##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3132  -4.0022   0.1144   3.0670  25.4482
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2746     1.4828   2.208   0.0296 *
## X             3.9452     0.2585  15.260   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.331 on 98 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.7008
## F-statistic: 232.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

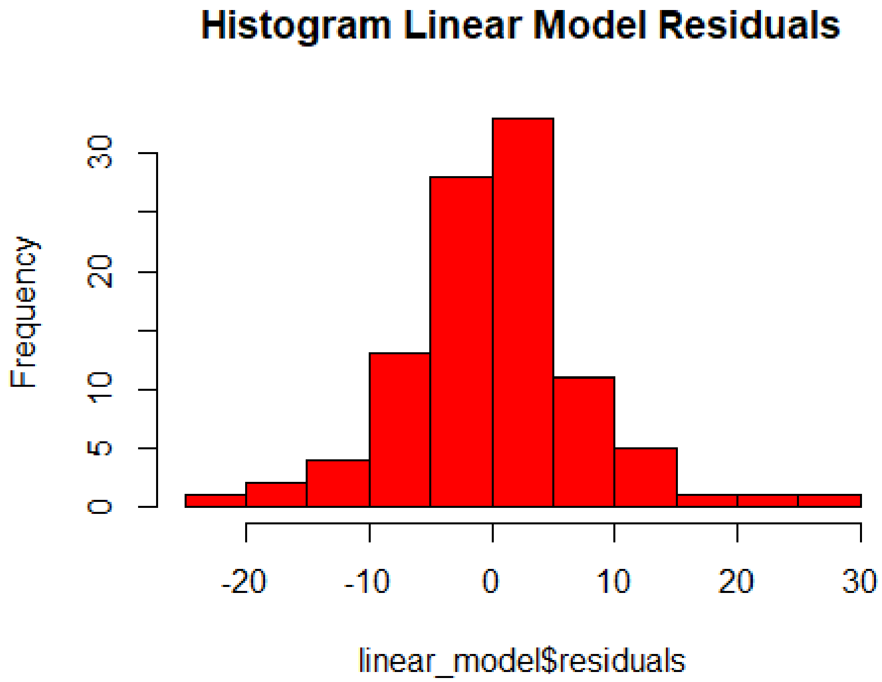
As we can see above, We have the model's summary statistics. But before analyzing this variables, we will see the pattern of the residuals using a scatter plot, Histogram, and Q-Q plot.

```
# Residuals plot
plot(linear_model$fitted.values, linear_model$residuals,
xlab = "Fitted Values",
ylab= "Residuals",
col= "red")
```



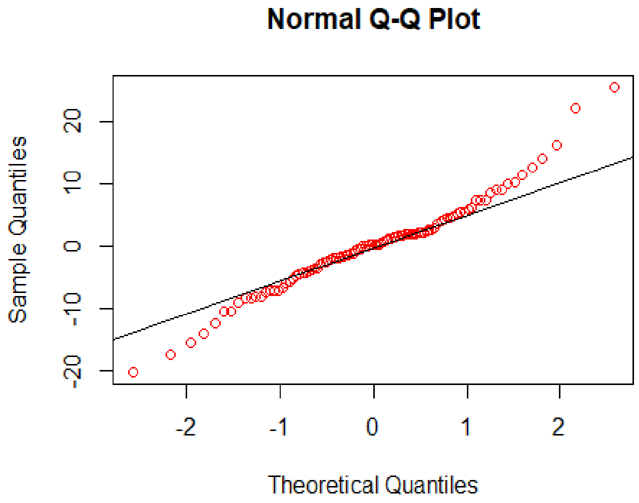
As we can see, the residual plot might show some shape.

```
# Histogram plot
hist(linear_model$residuals,
main="Histogram Linear Model Residuals",
col="red")
```



While analyzing the histogram, we can see the residuals seem to have a symmetric distribution and look like a bell-shaped normal distribution which is what we want.

```
# QQ plot
qqnorm(linear_model$residuals, col = "red")
qqline(linear_model$residuals)
```



As we can see, we have the theoretical quantiles on the X-axis, and on the Y-axis, we have the sample quantiles. From the result, we can see a good match between the theoretical and sample distribution, suggesting the residuals follow a normal distribution as we expected.

Now that we can see that the assumptions of the regression model are satisfied, we see the summary statistics of the model. So ideally, if you want to see any statistically significant relationship between the dependent variable and any of the independent variables, you would like to see the p-value as small as possible. So, the p-value is the probability that the hypothesis that a particular coefficient is equal to 0 is true. So, if this probability is small, we can reject that hypothesis and conclude that there are statistically significant relationships between the two variables. In this case, we can see that the intercept and the coefficient for x have a very small p-value. And therefore, we would consider them statistically significant.

Regarding the accuracy of the model, we can see that R2 is 70.38%, which we can affirm it is highly accurate.

c) How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y? (8% of total points)

Here is the formula for the calculation of R2,

$$R^2 = 1 - \frac{RSS}{TSS}$$

In other words,

$$R^2 = Coefficient\ of\ Determination = (Correlation\ Coefficient)^2$$

```
# Finding the value of Correlation Coefficient
cor(X,Y)

## [1] 0.8389348

# Finding the value of R^2
cor(X,Y)^2

## [1] 0.7038116
```

The coefficient of Determination states that it is the proportion of the variability of the dependent variable (y) accounted for or explained by the independent variable (x). The coefficient of determination ranges from 0 to 1. On the hand, the Correlation Coefficient measures the strength and the direction of a linear relationship between two variables. The value of r varies between -1 and 1. In our model value of R is 0.8389, and R2 is 0.7038, which means the correlation coefficient of X and Y is related.

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.

```
# Showing 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. (17% of total points)

James's point of view

```
set.seed(123)

# To run the linear model
james <- lm(mtcars$hp ~ mtcars$wt)

# To get descriptive statistics of the model
summary(james)

##
## 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
```

Chris's point of view

```
set.seed(123)

#Running the linear model
chris <- lm(mtcars$hp ~ mtcars$mpg)

#Getting the summary statistics of linear model
summary(chris)

##
## 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
```

By comparing both the models, we can see that the P value for Jame's model (weight of the car) and Chris's model (miles per gallon) is very small, which means that both are statistically significant. But the value of R2 for Jame's model is very low, which is 0.4339. On the other hand, Chris's statement that miles per gallon are a better predictor of horse car's power is right because this linear regression model shows higher accuracy than James' opinion, giving an accuracy of 0.6024 compared to 0.4339.

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? (17% of total points)

```
# Building a linear model

set.seed(123)

# To run the linear model
cyl_mpg <- lm(hp ~ cyl + mpg, data = mtcars)

# To get descriptive statistics of the model
summary(cyl_mpg)
```

```
## 
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
## 
## 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   
## cyl           23.979      7.346   3.264  0.00281 **
## 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
```

Building the Prediction.

```
set.seed(123)

# Predicted model
predicted_hp <- predict(cyl_mpg, newdata = data.frame(cyl = 4, mpg = 22))

#To see the result
predicted_hp

##        1 
## 88.93618
```

The estimated Horse Power of a car with 4 cylinders and an mpg of 22 is 88.93618.

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

```
# Installing required packages
#install.packages('mlbench')
library(mlbench)

## Warning: package 'mlbench' was built under R version 4.2.2

#Loading the dataset
data(BostonHousing)

#Showing the dataset
head(BostonHousing)
```

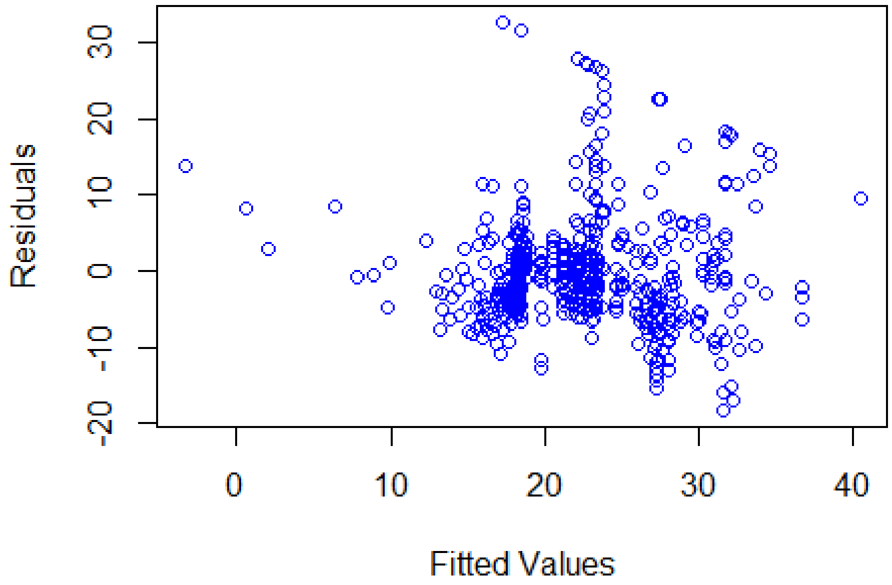
```
##       crim zn indus chas   nox    rm  age    dis rad tax ptratio      b
lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    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
```

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? (Hint check R2) (8% of total points)

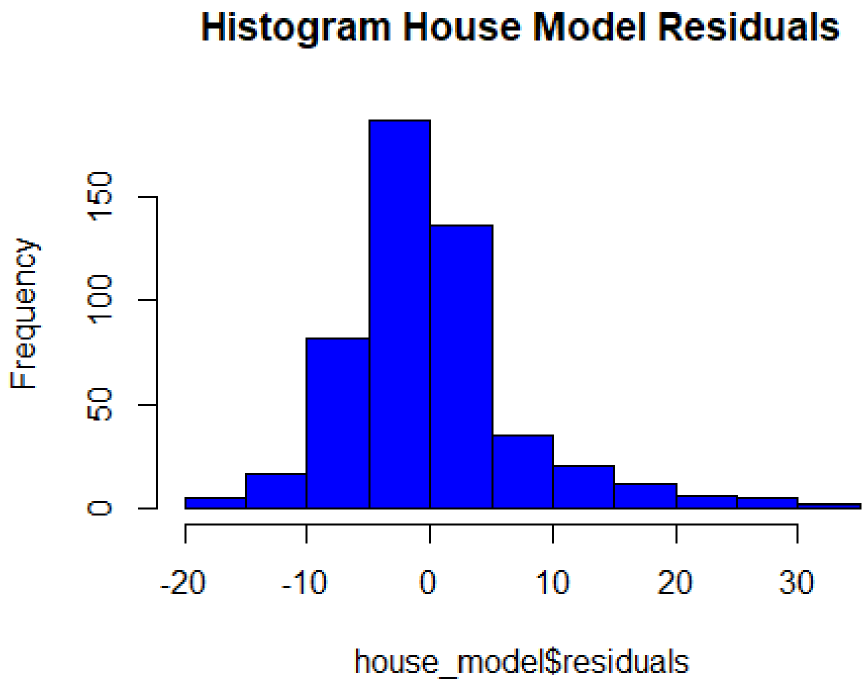
```
set.seed(123)

# To run the linear model
house_model <- lm(BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn
+ BostonHousing$ptratio + BostonHousing$chas)

# To get descriptive statistics of the model
summary(house_model)

##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##     BostonHousing$ptratio + BostonHousing$chas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             49.91868    3.23497  15.431  < 2e-16 ***
## BostonHousing$crim      -0.26018    0.04015  -6.480 2.20e-10 ***
## BostonHousing$zn       0.07073    0.01548   4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144  -8.712  < 2e-16 ***
## BostonHousing$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
```

Before analyzing the summary statistics of the model, let's analyze the assumptions of the regression model and plot scatter plot, histogram, and Q-Q plot.

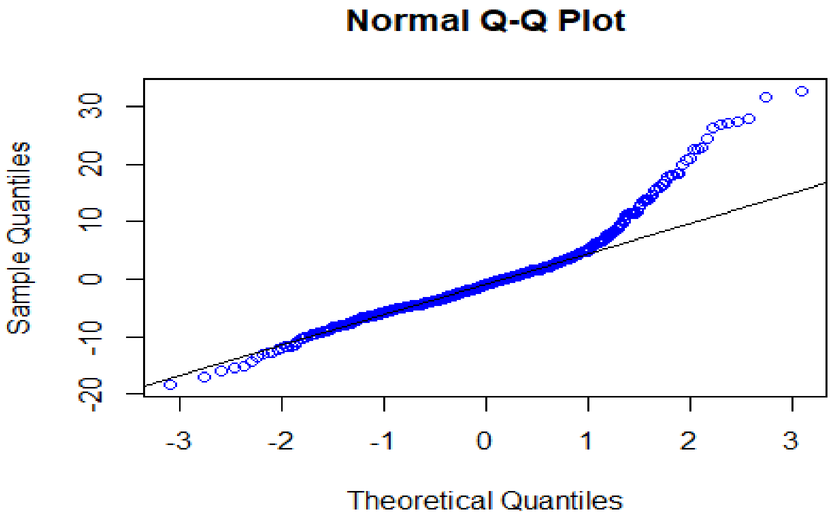
```
# Residuals plot
plot(house_model$fitted.values, house_model$residuals,xlab = "Fitted
Values",ylab= "Residuals",col= "blue")
```



```
# Histogram plot
hist(house_model$residuals,
main="Histogram House Model Residuals",
col="blue")
```



```
# QQ plot
qqnorm(house_model$residuals, col = "blue")
qqline(house_model$residuals)
```



By looking at the summary statistics, scatterplot, histogram, and Q-Q plot, it seems to be a partial model. The reason is the value of R2 is too low and gives an accuracy of just 35.99%. If we look at the residual plot, the values are concentrated in just one place, the histogram is also right skewed tailed, and the values in the Q-Q plot are not in a line.

b) Use the estimated coefficient to answer these questions.

```
# Getting the coefficient values

house_model$coefficients

##           (Intercept)    BostonHousing$crim      BostonHousing$zn
##           49.91868439           -0.26017612            0.07072809
## BostonHousing$ptratio   BostonHousing$chas1
##           -1.49367255            4.58392591
```

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? (8% of total points)

: *As we can see, the value of chas is positive. So the house which bounds the chas river will be more expensive, and its price will increase by 4.5839 compared to the one which does not bound chas river.*

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? (Golden Question: 4% extra)

: *The house with a pupil-teacher ratio of 15, the price for it is going to change by 15(-1.49367255)= -22.40509. In a house with a pupil-teacher ratio of 18, its price will change by 18(-1.49367255)= -26.88611. Therefore, a house with a ratio of 15 will be more expensive than a house with a ratio of 18. The difference between the price will be 4.48102 (-22.40509 -(-26.88611)= 4.48102)*

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. (8% of total points)

: *The P-values help us to determine whether the variables are statistically significant. As we know that the p-values should be small to be statistically significant. In this model, we can confirm that the p-values for all the variables are small and, therefore, statistically significant.*

d.) Use the anova analysis and determine the order of importance of these four variables. (18% of total points)

```
set.seed(123)

# Run ANOVA analysis
anova(house_model)

## Analysis of Variance Table
##
## Response: BostonHousing$medv
##                        Df Sum Sq Mean Sq F value    Pr(>F)
## BostonHousing$crim      1 6440.8  6440.8 118.007 < 2.2e-16 ***
## BostonHousing$zn        1 3554.3  3554.3  65.122 5.253e-15 ***
## BostonHousing$ptratio   1 4709.5  4709.5  86.287 < 2.2e-16 ***
## BostonHousing$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
```

The order of importance based on ANOVA analysis is determined using the Sum Sq. So the order of importance is as follows:

1.) Crime Rate (crim) - 6440.8

2.) Pupil- Teacher Ratio (ptratio) - 4709.5

3.) Land Zoned (zn) - 3554.3

4.) Chas River (chas) - 667.2