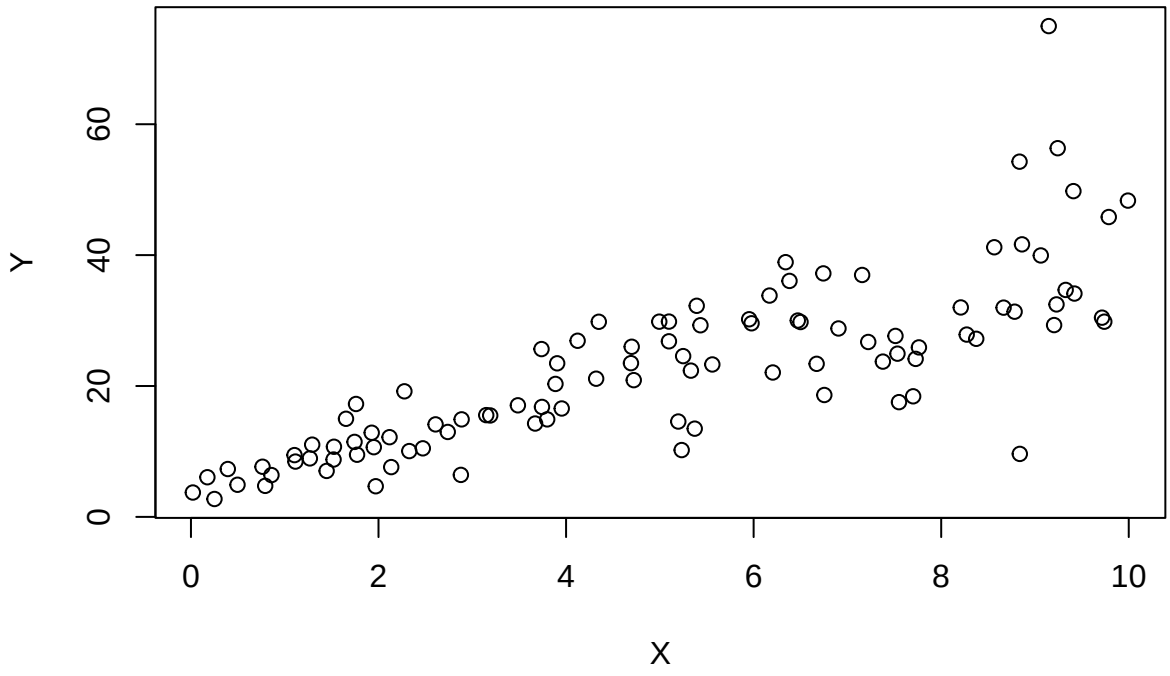
Assignment-Regression Analytics

1.

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

(a)Plot Y against X:

```
plot(Y~X)
```



Based on the plot, it is clear that the value of Y keeps increasing with the value of X indicating that there is a positive correlation between X and Y. Hence, we can fit a linear model to explain Y based on X.

(b) Simple linear model of Y based on X:

```
 linear_model<- lm(Y~X)
linear_model$coefficients
```

```
## (Intercept)           X
##    4.465490    3.610759
```

Equation explaining Y based on X:

$$Y = 3.61075 * X + 4.465490$$

Accuracy of the model:

```
summary(linear_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
```

Considering R square value to predict the accuracy of the model as R square is the measure of performance of a linear Regression model. In this case, R square is 0.6517 indicates that the model is 65.17 percent accurate.

(c) How the Coefficient of determination R^2,of the model is related to the correlation coefficient of X and Y?

```
cor(Y,X)
```

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

The correlation coefficient calculated above indicates that there is a positive correlation of 0.807291 between X and Y. Coefficient of determination is the square of this value.

2

(a)

```
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
```

```
#constructing linear regression model to determine hp based on weight of the car:
linear_model1<- lm(hp~wt,data=mtcars)
summary(linear_model1)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## 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
## 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
```

```
#constructing linear regression model to determine hp based on Mile per Gallon(mpg) of the car:
linear_model2<-lm(hp~mpg,data=mtcars)
summary(linear_model2)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = mtcars)
##
## 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 ***
## 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
```

To best determine which variable can be used to estimate the horse power of a car, we are considering R square value as it implies the proportion of variability of the dependent variable accounted for the independent variable.

R square value to estimate horse power based on weight is 43.39 percent whereas the R square value to estimate horse power based on miles per Gallon is 60.24 percent.

Therefore, it is clear to say that the horse power can be best estimated with the value of mpg and not based on the weight of the car.

Hence, Chris is right about estimating the horse power of the car

(b)Constructing a model to predict the car horse power based on number of cylinders and miles per Gallon:

```
linear_model3<- lm(hp~cyl+mpg,data = mtcars)
summary(linear_model3)
```

```
##
## 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
```

Linear equation:

$$hp = 54.067 + 23.979 * X_1 - 2.775 * X_2$$
$$where X_1 = cyl, X_2 = mpg$$

Estimated horsepower of a car with 4 cylinders and mpg of 22:

```
predicted_hp_value<-predict(linear_model3,data.frame(cyl=c(4),mpg=c(22)))

predicted_hp_value
```

```
##        1
## 88.93618
```

The estimated horse power of a car with 4 cylinders and 22 mpg is 88.93618

3 (a) building a model to estimate the median value of owner occupied homes based on crime rate, proportion of residential land zoned for lots over 25,000 sq.ft,the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River:

```
library(mlbench)
```

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

```
data(BostonHousing)

linear_model4<-lm(medv~crim+zn+ptratio+chas,data=BostonHousing)

summary(linear_model4)
```

```
##
## 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
```

The above model is 35.99 percent accurate which is a relatively low percent. Hence, we don't consider this as a good model

(b) Identifying which of the two identical houses is more expensive:

To identify which home is expensive comparing the one that bounds the Chas river and the one's do not, we consider the coefficient of the chas value in the above linear model. The coefficient is 4.58393, indicates that the houses that bounds by the Chas river are 4.58393 times more expensive than the houses which do not bounds by the river.

Moreover, in the dataset, the values of chas river are 1 or 0 which means the houses which bounds by the river are assigned a value of 1, otherwise 0. So for the houses which do not bounds by the river are going to have 0 times change in their value

(c) Finding which of the variables are statistically important:

All the variables including crime rate,proportion of residential land zoned for lots over 25,000 sq.ft,the local pupil-teacher ratio,the tract bounds Chas River are statistically important as all of them has very low P value

(d)Determining the order of importance of the 4 variables using ANOVA analysis:

```
anova_lm<-anova(linear_model4)
anova_lm
```

```
## 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
```

The importance of variables can be determined by their Sum of Squares value. Higher the Sum of squares, the more important is the variable in estimating the value of a dependent variable

Order of importance of variables:

crim-per capita crime rate by town

ptratio-pupil-teacher ratio by town.

zn-proportion of residential land zoned for lots over 25,000 sq.ft.

Chas-Charles River dummy variable