EXPERIMENT NO:- 3

AIM:- Using multiple linear regression perform the following tasks on the Auto data set.

(a) Produce a scatter plot matrix which includes all of the variable sin the data set.
(b) Compute the matrix of correlations between the variables using the function cor() . You will need to exclude the name variable, cor() which is qualitative.
(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output, That is

i. Is there a relationship between the predictors and the response?
ii. Which predictors appear to have a statistically significant relationship to the response?
iii. What does the coefficient for the year variable suggest?

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?
(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?
(f) Try a few different transformations of the variables, such aslog(X),√ X, X 2 . Comment on your findings.

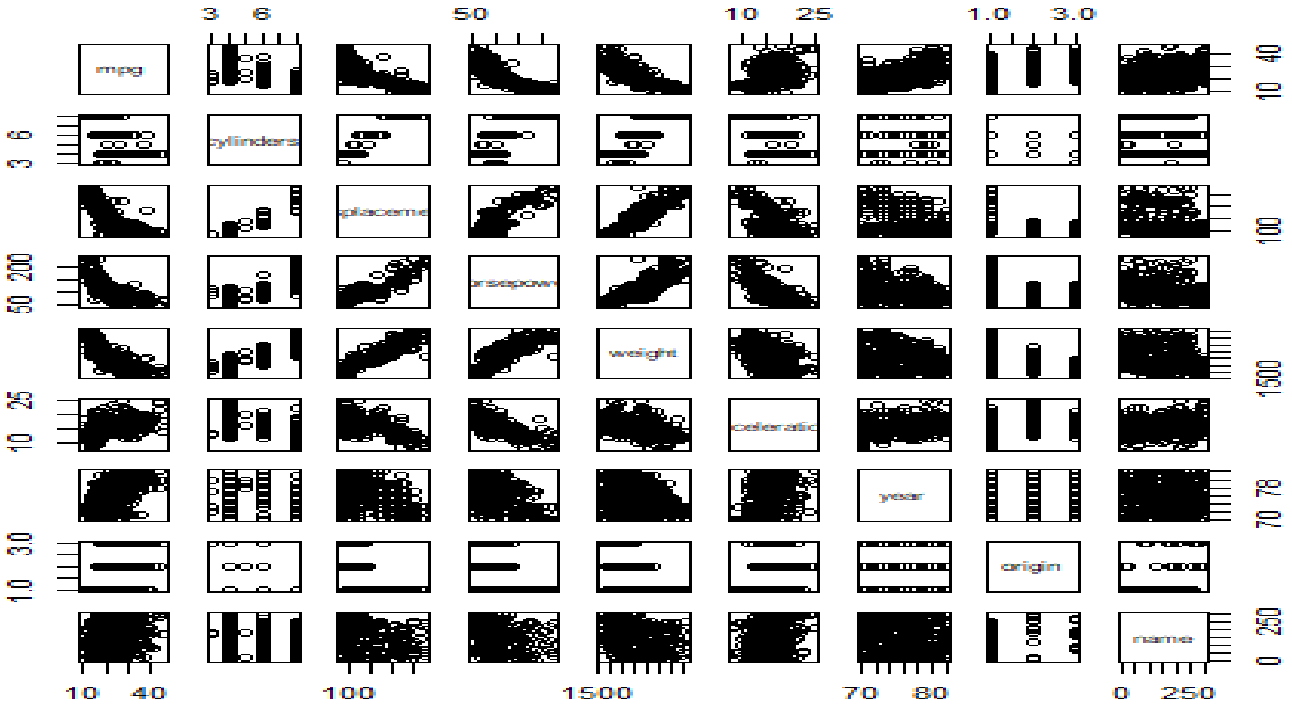
(3a) AIM:- Produce a scatter plot matrix which includes all of the variable sin the data set.

DESCRIPTION: A scatterplot is a set of dotted points representing individual data pieces on the horizontal and vertical axis. In a graph in which the values of two variables are plotted along the axis X and Y, the pattern of the resulting points reveals a correlation between them. We can create a scatter plot in R programming Language using the plot() function.

CODE:

```
data("Auto", package = "ISLR")
pairs(Auto)
```

OUTPUT:



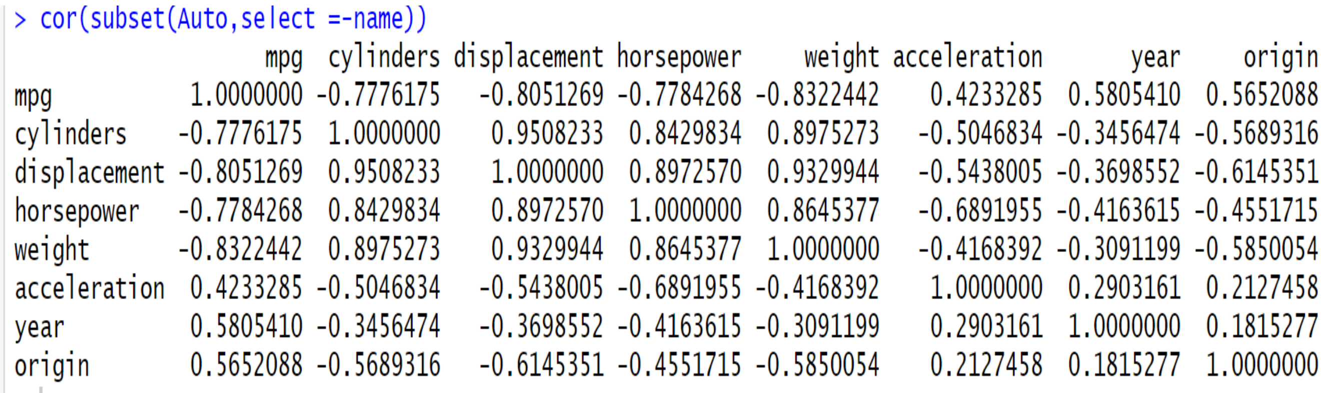
(3b) AIM:- Compute the matrix of correlations between the variables using the function cor() . You will need to exclude the name variable, cor() which is qualitative.

DESCRIPTION: cor() function in R Language is used to measure the correlation coefficient value between two vectors. This returns a simple correlation matrix showing the correlations between pairs of variables (devices). You can choose the correlation coefficient to be computed using the method parameter. The default method is Pearson, but you can also compute Spearman or Kendall coefficients.

CODE:

```
cor(subset(Auto, select = -name))
```

OUTPUT:



```
> cor(subset(Auto,select =-name))
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

(3c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output, That is

i. Is there a relationship between the predictors and the response?

ii. Which predictors appear to have a statistically significant relationship to the response?

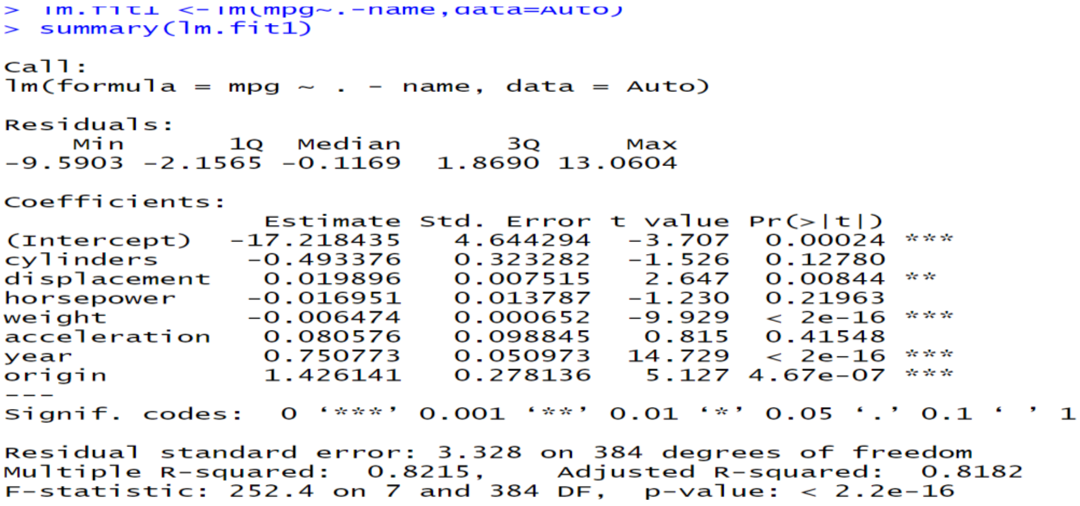
iii. What does the coefficient for the year variable suggest?

DESCRIPTION: The lm() function to fit a simple linear regression lm() model, with medv as the response and lstat as the predictor. The basic syntax is lm(y~x,data), where y is the response, x is the predictor, and data is the data set in which these two variables are kept.

CODE:

```
lm.fit1 <- lm(mpg ~ . - name, data = Auto)
summary(lm.fit1)
```

OUTPUT:



```
> lm.fit1 <-lm(mpg~.-name,data=Auto)
> summary(lm.fit1)

Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5903 -2.1565 -0.1169  1.8690 13.0604 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780    
displacement   0.019896   0.007515   2.647  0.00844 ** 
horsepower    -0.016951   0.013787  -1.230  0.21963    
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548    
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,	Adjusted R-squared:  0.8182 
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

(3.c)
(i). Is there a relationship between the predictors and the response?

Yes, there is a relationship between the predictors and the response by testing the null hypothesis of whether all the regression coefficients are zero. The F-statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis.

(3.c)
(ii). Which predictors appear to have a statistically significant relationship to the response?

Looking at the p-values associated with each predictor’s t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship , while cylinders, horsepower, and acceleration do not.

(3.c)
(iii). What does the coefficient for the `year` variable suggest?

The regression coefficient for year, 0.7507727, suggests that for every one year, mpg increases by the coefficient. In other words, cars become more fuel efficient every year by almost 1 mpg / year.

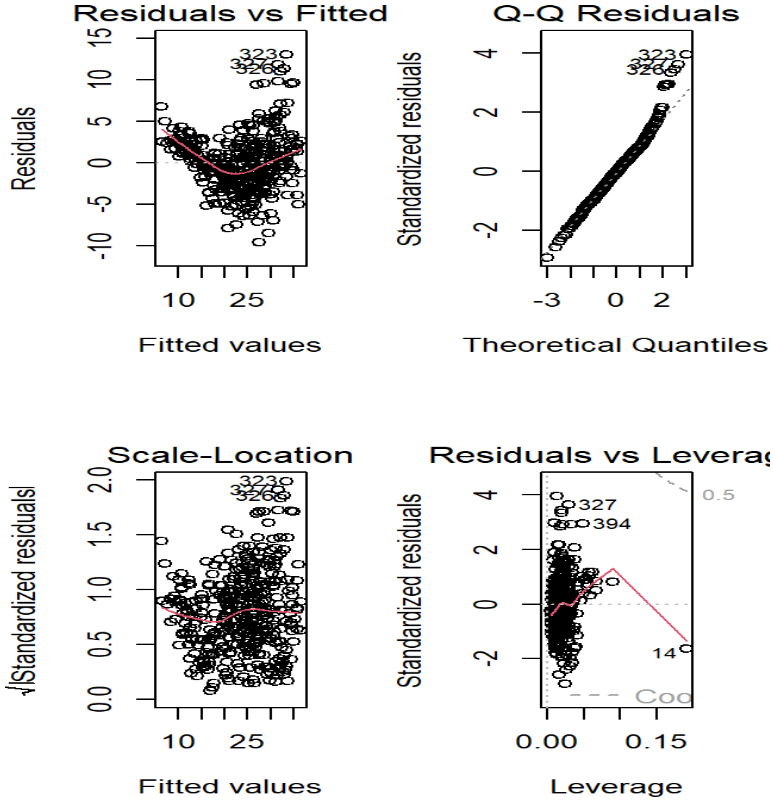
(3d) AIM:- Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

DESCRIPTION: Based on the plot of Residuals versus fitted values there appears to be non linearity in the data and based on the Residuals vs Leverage plot below, there does appear to be outliers due to the fact that there are data points outside of the -2 and 2 standardized residuals. And there does appear to be high leverage observations greater then 0.05.

CODE:

```
par(mfrow = c(2, 2))
plot(lm.fit1)
```

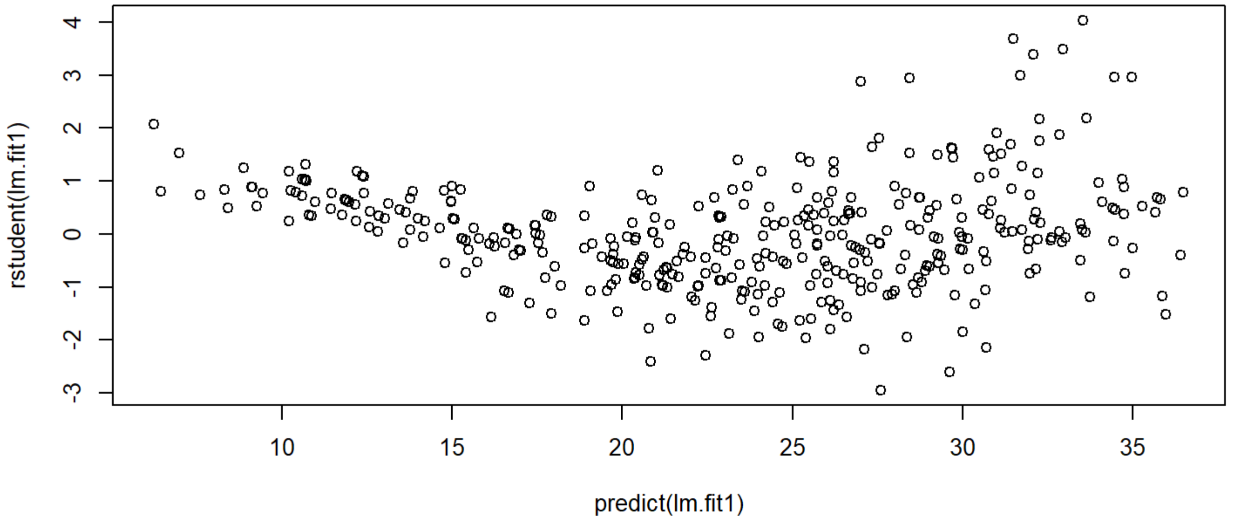
OUTPUT:



CODE:

plot(predict(lm.fit1),rstudent(lm.fit1))

OUTPUT:



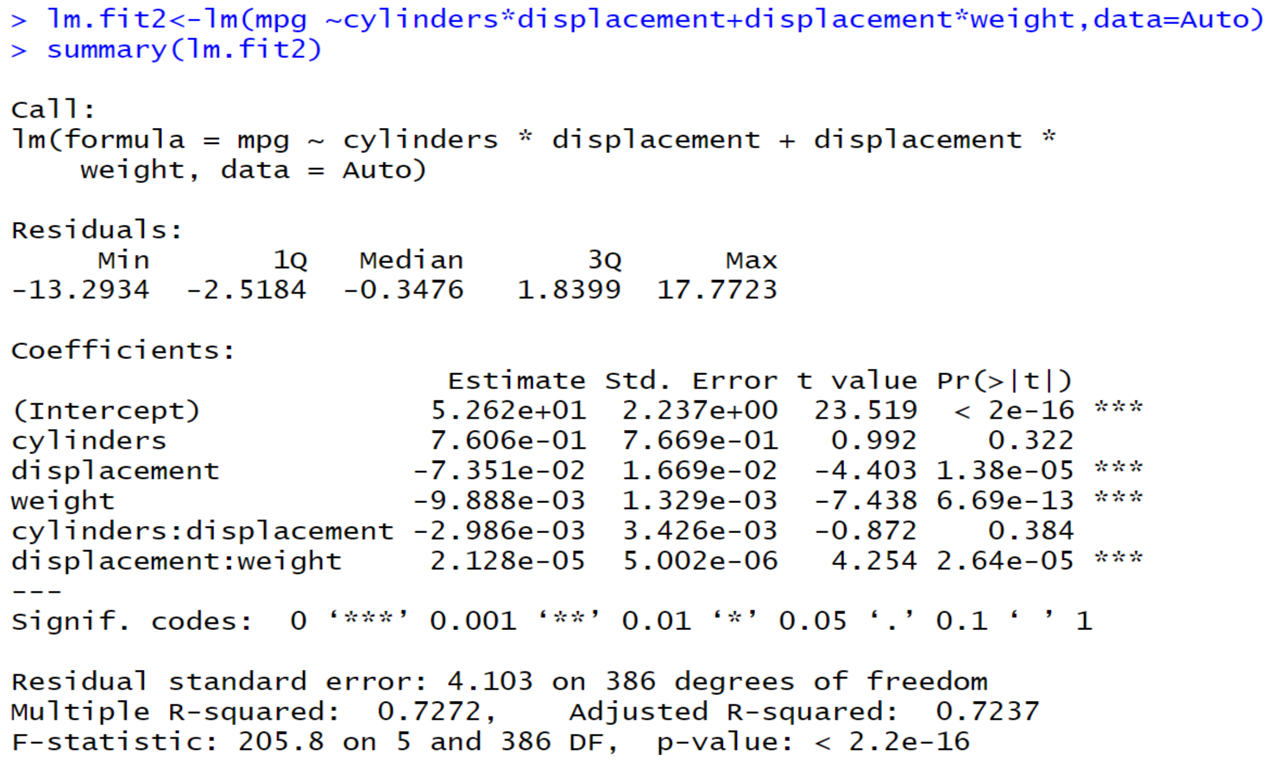
(3e) AIM:- Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

DESCRIPTION: Based on here it appears that the interactions between ‘cylinders’ and ‘displacement are not statistically significant but the interactions between ’horsepower’ and ‘weight’ is significant and the interaction between ‘acceleration’ and ‘year’ is significant.

CODE:

```
lm.fit2 <- lm(mpg ~ cylinders * displacement + displacement * weight, data = Auto)
summary(lm.fit2)
```

OUTPUT:



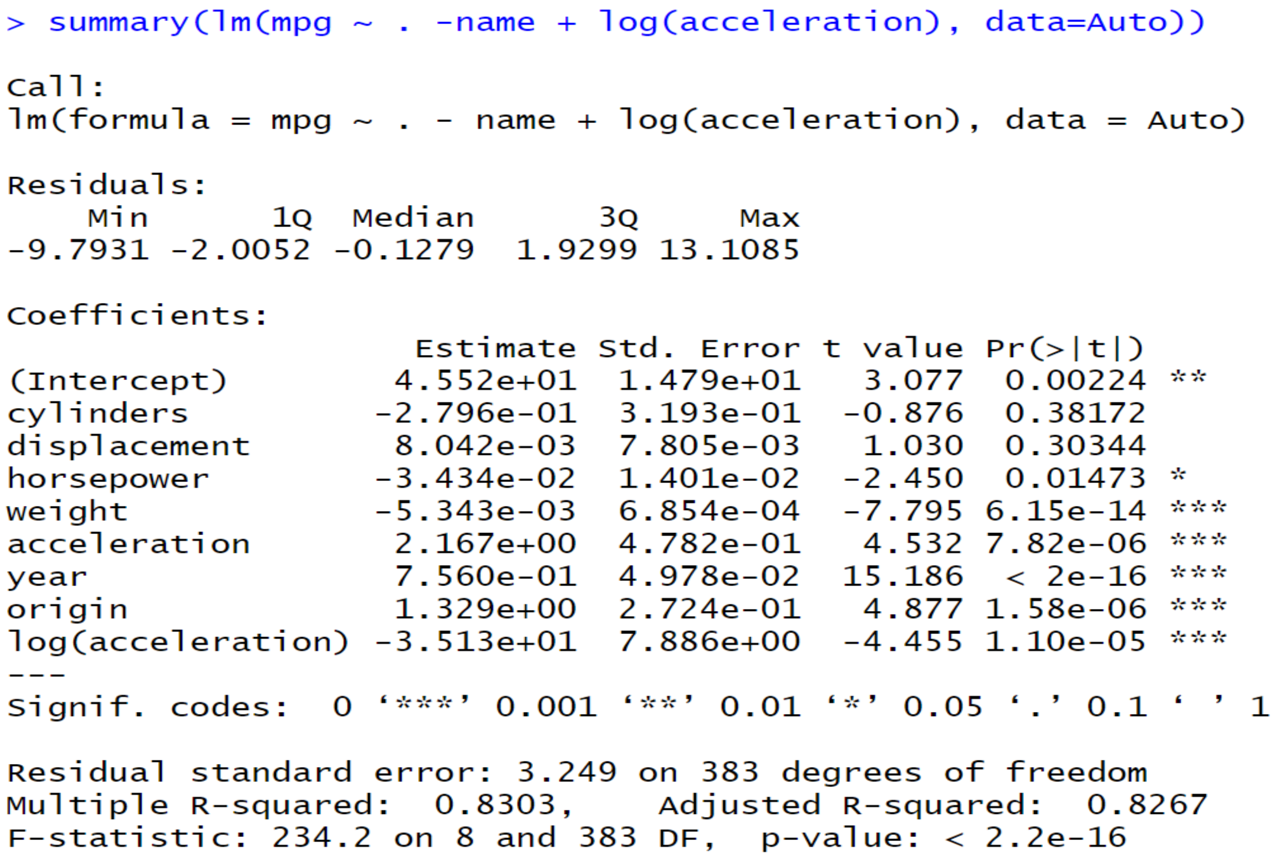
```
> lm.fit2<-lm(mpg ~cylinders*displacement+displacement*weight,data=Auto)
> summary(lm.fit2)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement * 
    weight, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.2934  -2.5184  -0.3476   1.8399  17.7723 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
cylinders               7.606e-01  7.669e-01   0.992    0.322    
displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384    
displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.103 on 386 degrees of freedom
Multiple R-squared:  0.7272,    Adjusted R-squared:  0.7237 
F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

(3f) AIM:- Try a few different transformations of the variables, such aslog(X),√ X, X 2 . Comment on your findings.

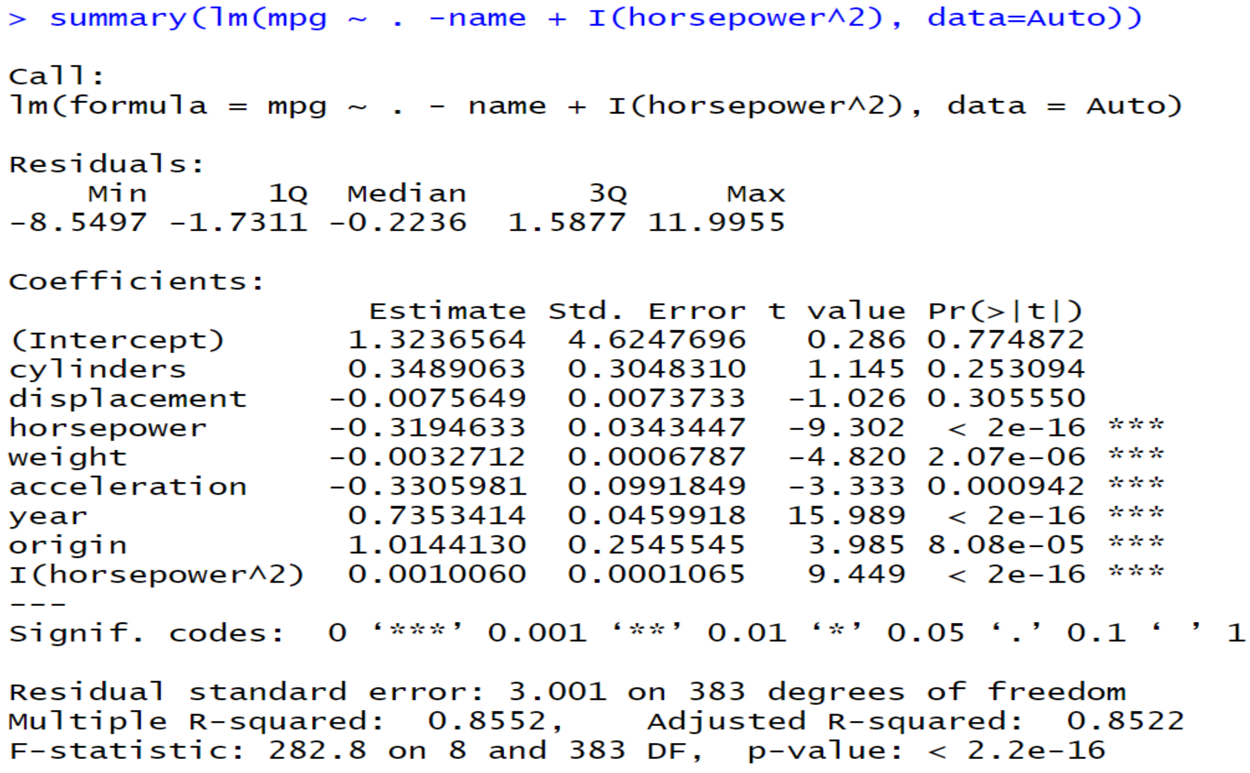
DESCRIPTION: : Applying the log() function on a vector, data frame, or other data set in R results in a log transformation. To avoid applying a logarithm to a 0 number, 1 is added to the base value prior to applying the logarithm.

CODE: summary(lm(mpg ~ . -name + log(acceleration), data=Auto))
summary(lm(mpg ~ . -name + I(horsepower^2), data=Auto))

OUTPUT:



```
> summary(lm(mpg ~ . -name + log(acceleration), data=Auto))

Call:
lm(formula = mpg ~ . - name + log(acceleration), data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.7931 -2.0052 -0.1279  1.9299 13.1085 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)        4.552e+01  1.479e+01   3.077  0.00224 ** 
cylinders         -2.796e-01  3.193e-01  -0.876  0.38172    
displacement       8.042e-03  7.805e-03   1.030  0.30344    
horsepower        -3.434e-02  1.401e-02  -2.450  0.01473 *  
weight            -5.343e-03  6.854e-04  -7.795 6.15e-14 ***
acceleration       2.167e+00  4.782e-01   4.532 7.82e-06 ***
year               7.560e-01  4.978e-02  15.186  < 2e-16 ***
origin             1.329e+00  2.724e-01   4.877 1.58e-06 ***
log(acceleration) -3.513e+01  7.886e+00  -4.455 1.10e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.249 on 383 degrees of freedom
Multiple R-squared:  0.8303,	Adjusted R-squared:  0.8267 
F-statistic: 234.2 on 8 and 383 DF,  p-value: < 2.2e-16
```



```
> summary(lm(mpg ~ . -name + I(horsepower^2), data=Auto))

Call:
lm(formula = mpg ~ . - name + I(horsepower^2), data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.5497 -1.7311 -0.2236  1.5877 11.9955 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.3236564  4.6247696   0.286 0.774872    
cylinders        0.3489063  0.3048310   1.145 0.253094    
displacement    -0.0075649  0.0073733  -1.026 0.305550    
horsepower      -0.3194633  0.0343447  -9.302  < 2e-16 ***
weight          -0.0032712  0.0006787  -4.820 2.07e-06 ***
acceleration    -0.3305981  0.0991849  -3.333 0.000942 ***
year             0.7353414  0.0459918  15.989  < 2e-16 ***
origin           1.0144130  0.2545545   3.985 8.08e-05 ***
I(horsepower^2)  0.0010060  0.0001065   9.449  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.001 on 383 degrees of freedom
Multiple R-squared:  0.8552,	Adjusted R-squared:  0.8522 
F-statistic: 282.8 on 8 and 383 DF,  p-value: < 2.2e-16
```