Data representation analysis

Step 1 Create a directory on your computer where the home work will be implemented. That means all input files should be placed there and R will create the output files in this directory.

Step 2 Open R- studio. In the command prompt select "Session" then "Set working directory" and "Choose directory". Using the command prompt with file manager select the directory you created in Step 1.

```
> setwd("D:/ABI Assignments")
```

Step 3(1 mark) Read Auto.csv file. Use command dim to check how many observations and variables are in this file. Report the output below:

COMMAND:

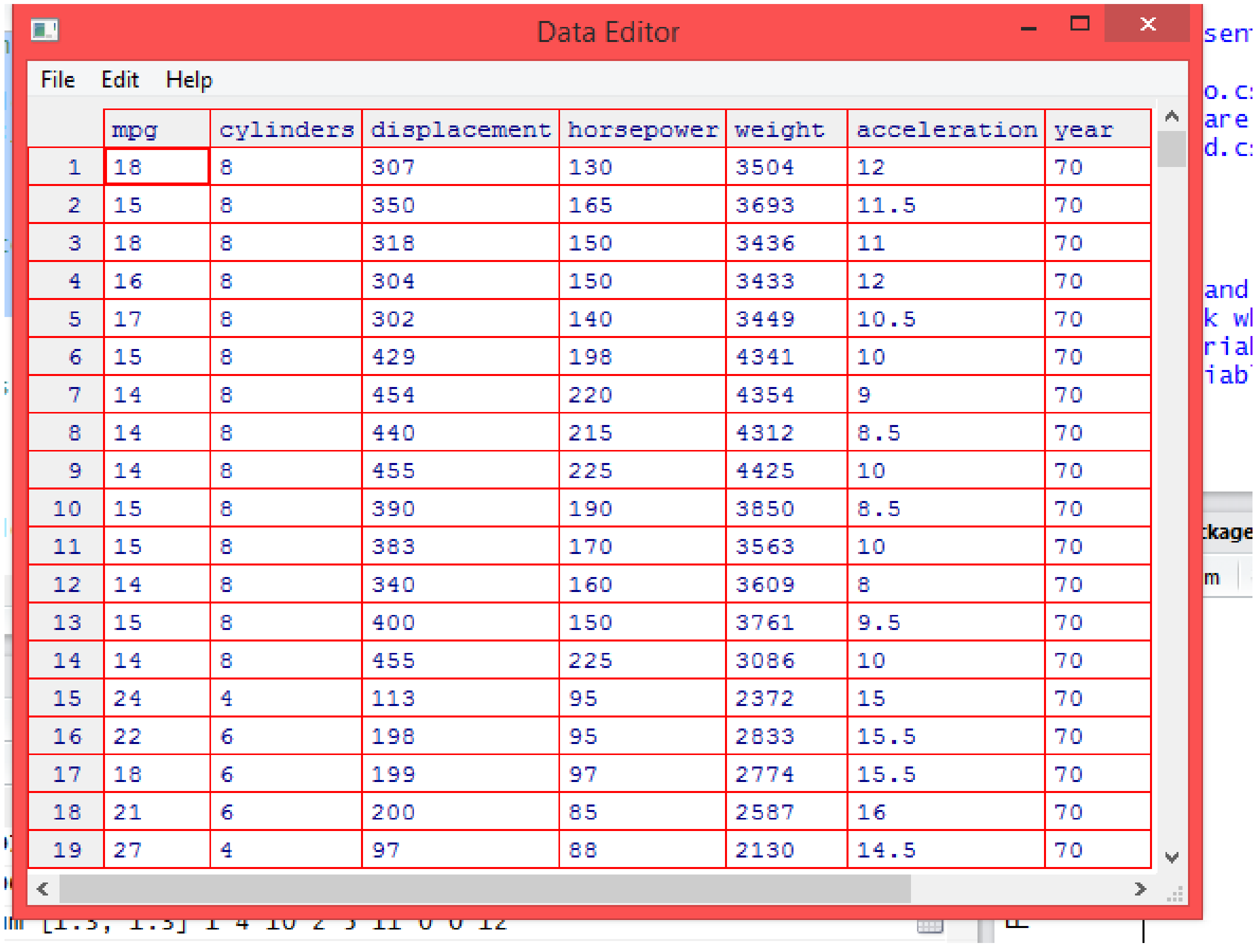
```
> Auto <- read.csv("D:/ABI Assignments/Auto.csv")
> dim(Auto)
[1] 397   9
```

OUTPUT: 397 Observations and 9 Variables.

Step 4(1 mark) Use command fixto open the file editor. The program will open you this file, check what variables you have there and close the file. Change the name of variable "cylinders" to "cyl".Use command names to print the name of the variables. Report the output below:

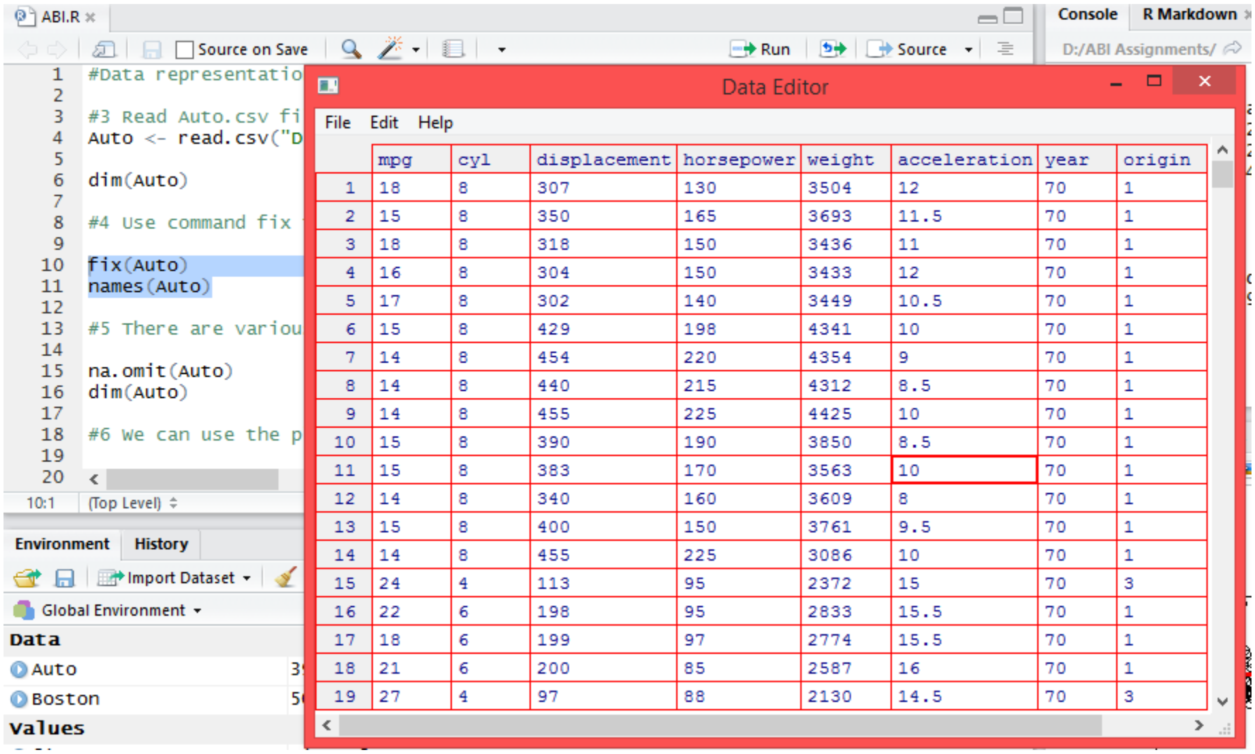
```
> fix(Auto)

> names(Auto)
[1] "mpg"          "cyl"          "displacement" "horsepower"
[5] "weight"       "acceleration" "year"         "origin"
[9] "name"
```



	mpg	cylinders	displacement	horsepower	weight	acceleration	year
1	18	8	307	130	3504	12	70
2	15	8	350	165	3693	11.5	70
3	18	8	318	150	3436	11	70
4	16	8	304	150	3433	12	70
5	17	8	302	140	3449	10.5	70
6	15	8	429	198	4341	10	70
7	14	8	454	220	4354	9	70
8	14	8	440	215	4312	8.5	70
9	14	8	455	225	4425	10	70
10	15	8	390	190	3850	8.5	70
11	15	8	383	170	3563	10	70
12	14	8	340	160	3609	8	70
13	15	8	400	150	3761	9.5	70
14	14	8	455	225	3086	10	70
15	24	4	113	95	2372	15	70
16	22	6	198	95	2833	15.5	70
17	18	6	199	97	2774	15.5	70
18	21	6	200	85	2587	16	70
19	27	4	97	88	2130	14.5	70

Name changed from cylinders to cyl.



	mpg	cyl	displacement	horsepower	weight	acceleration	year	origin
1	18	8	307	130	3504	12	70	1
2	15	8	350	165	3693	11.5	70	1
3	18	8	318	150	3436	11	70	1
4	16	8	304	150	3433	12	70	1
5	17	8	302	140	3449	10.5	70	1
6	15	8	429	198	4341	10	70	1
7	14	8	454	220	4354	9	70	1
8	14	8	440	215	4312	8.5	70	1
9	14	8	455	225	4425	10	70	1
10	15	8	390	190	3850	8.5	70	1
11	15	8	383	170	3563	10	70	1
12	14	8	340	160	3609	8	70	1
13	15	8	400	150	3761	9.5	70	1
14	14	8	455	225	3086	10	70	1
15	24	4	113	95	2372	15	70	3
16	22	6	198	95	2833	15.5	70	1
17	18	6	199	97	2774	15.5	70	1
18	21	6	200	85	2587	16	70	1
19	27	4	97	88	2130	14.5	70	3

Step 5 (1 mark) There are various ways to deal with the missing data. In this case, only five of the rows contain missing observations, and so we choose to use the na.omit() function to simply remove these rows. Check the file with dim() command again. How many missing row do you have in this file?

OUTPUT:

```
> na.omit(Auto)
> dim(Auto)
[1] 397   9
```

CONCLUSION: **There are no missing values in the file.**

Step 6 (1 mark) We can use the plot() function to produce *scatterplots* of the quantitative variables. To indicate the file from which variables should be plotted use command attach() before. Report the results including the scatterplot of cylinders vs mpg. Save the output to the *.pdf file then download it to this report.

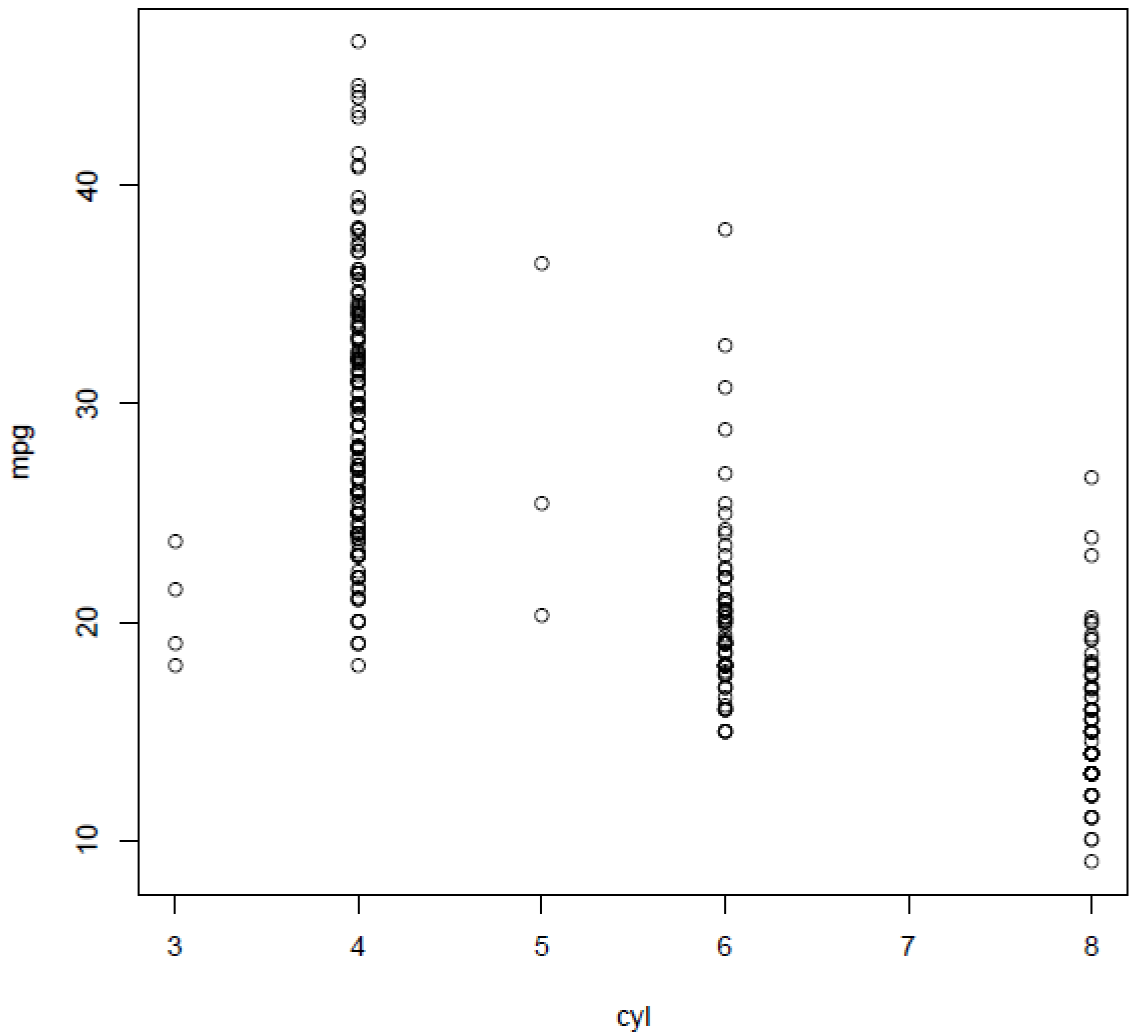
COMMAND

```
attach(Auto)
plot(cyl, mpg, main="Scatterplot",
    xlab="cyl ", ylab="mpg ")

pdf (" scatterplot.pdf ")
plot(cyl,mpg)
dev.off ()
```

OUTPUT:

```
> attach(Auto)
> plot(cyl, mpg, main="Scatterplot", xlab="cyl ", ylab="mpg ")
>
> pdf (" scatterplot.pdf ")
> plot(cyl,mpg)
> dev.off ()
RStudioGD
        2
```



Step 7 (1 mark) Produce the histogram for horsepower variable using hist() command. Use blue color (col) and 6 bars (breaks).

COMMAND

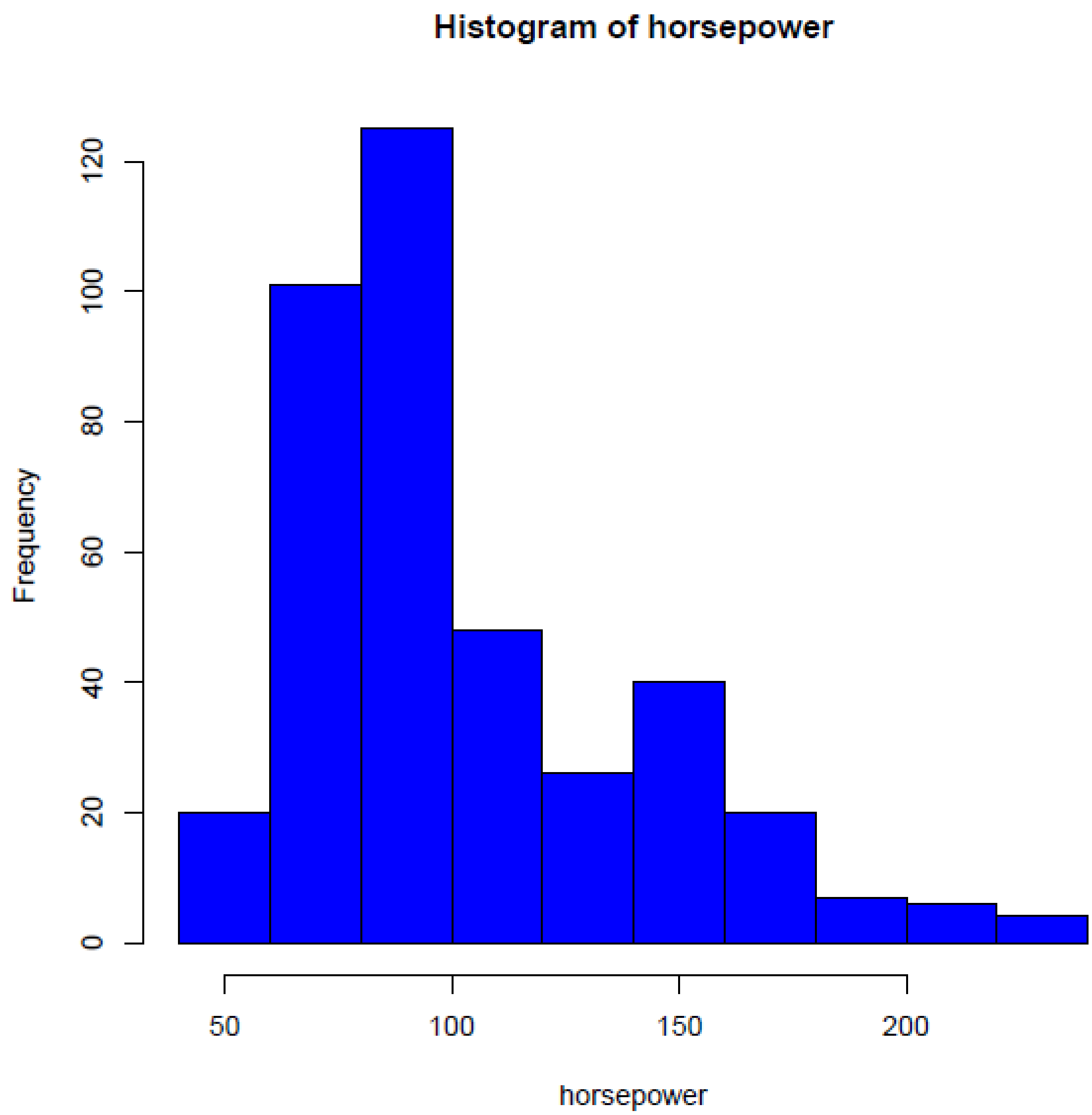
```
plot(cyl, mpg, main="Scatterplot",
      xlab="cyl ", ylab="mpg ")

hist(horsepower, breaks=6, col="blue")

pdf (" histogram .pdf ")
hist(horsepower ,col =" blue ")
dev.off ()
```

OUTPUT

```
> plot(cyl, mpg, main="Scatterplot",
+      xlab="cyl ", ylab="mpg ")
> hist(horsepower, breaks=6, col="blue")
>
> pdf (" histogram .pdf ")
> hist(horsepower ,col =" blue ")
> dev.off ()
RStudioGD
        2
```



Step8 (1 mark) Produce the statistical summary form mpg and acceleration variables using summary() command. Are these data samples symmetric or skewed.

COMMAND:

summary(mpg)

summary(acceleration)

OUTPUT

```
>
> summary (mpg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   17.50   23.00   23.52   29.00   46.60
> summary (acceleration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   13.80   15.50   15.56   17.10   24.80
>
> library(e1071)
>
> skewness(Auto$mpg)
[1] 0.4525649
>
> skewness(Auto$acceleration)
[1] 0.278699
>
```

Observation:

- The data samples are skewed.
- If mean and median does not coincide, it is skewed and not symmetric.
- For mpg, Media =23.00 < Mean =23.52- It is Right Skewed
- For acceleration, Media=15.00 < Mean=15.56- It is Right Skewed
- The data samples are Positive skewed which indicates that the mean of the data values is larger than the median, and the data distribution is right-skewed.

Simple (one variable) linear regression

You need to connect MASS library using the command library(MASS). The Boston data is part of the MASSlibrary. The MASSlibrary contains the Bostondata set, which records medv(median house value) for 506 neighborhoods around Boston. We will seek to predict medvusing 13 predictors such as rm(average number of rooms per house), age(average age of houses), and lstat(percent of households with low socioeconomic status).

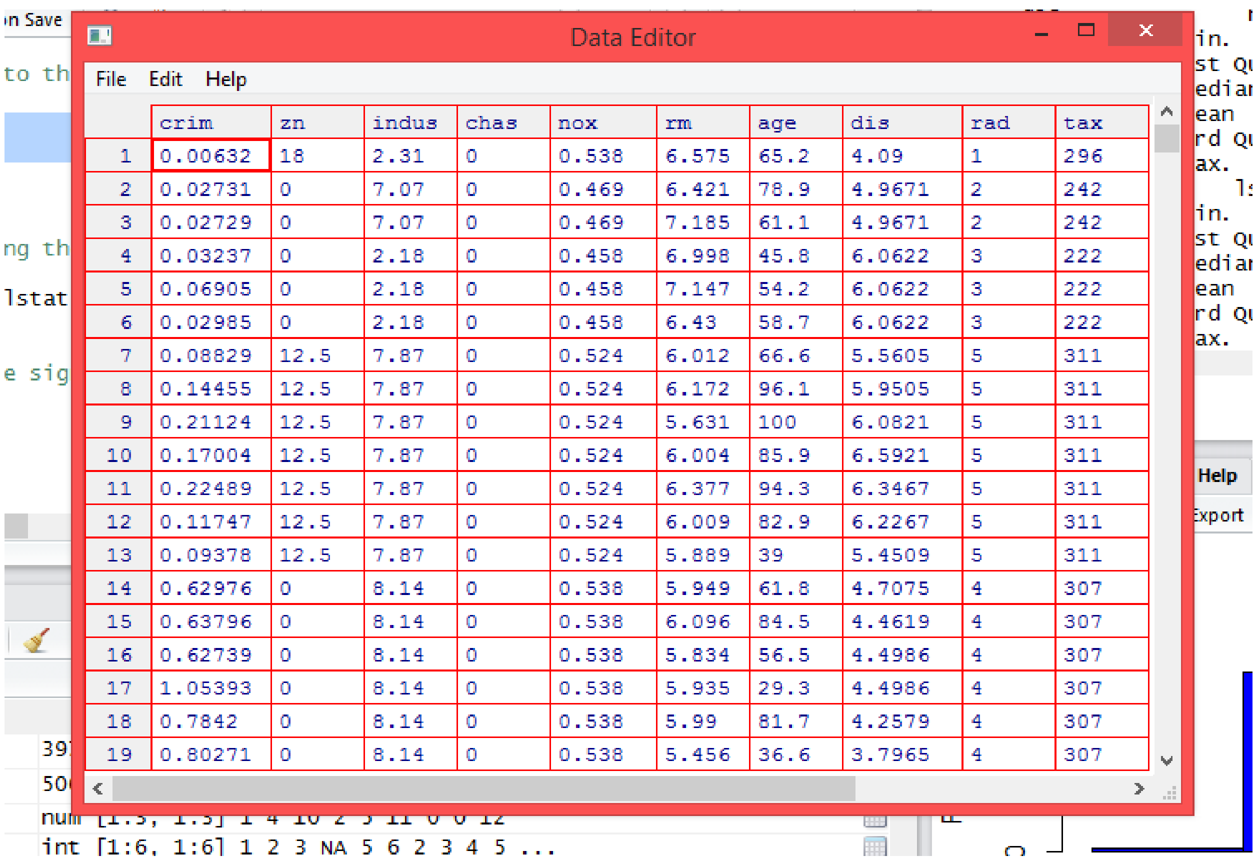
Step 1(1 mark) Have a look to this data using fix command and print summary statistics for all variables.

COMMAND

library(MASS)
fix(Boston)
summary(Boston)

OUTPUT

```
> library(MASS)
> fix(Boston)
```



Data Editor

File Edit Help

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311
13	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311
14	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307
15	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307
16	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307
17	1.05393	0	8.14	0	0.538	5.935	29.3	4.4986	4	307
18	0.7842	0	8.14	0	0.538	5.99	81.7	4.2579	4	307
19	0.80271	0	8.14	0	0.538	5.456	36.6	3.7965	4	307

```
>
> summary(Boston)
      crim                zn             indus            chas              nox
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710
       rm             age              dis              rad              tax
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0
 1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0
 Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0
 Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2
 3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0
    ptratio          black            lstat            medv
 Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
 1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
 Median :19.05   Median :391.44   Median :11.36   Median :21.20
 Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53
 3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00
>
```

Step 2(1 mark) Start by using thelm() function to fit a simple linear regression model, with medvas the response and lstatas the predictor. The basicsyntax is lm(y~x,data), where y is the response, x is the predictor, anddatais the data set in which these two variables are kept. Do not forget to attach Boston data either by separate command or by option "data" in lm command.

COMMAND:

fit <- lm(medv~lstat , data = Boston)
summary(fit)

OUTPUT

```
> lm.fit=lm(medv~lstat,data=Boston)
> attach (Boston )
> |
```

Step 3(1 mark) Interpret the significance of regression model.

```
> fit <- lm(medv~lstat , data = Boston)
> summary(fit)

Call:
lm(formula = medv ~ lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.168  -3.990  -1.318   2.034  24.500 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,	Adjusted R-squared:  0.5432 
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

OBSERVATION

- **p-value** is **2.2e-16** which is very less than 0.05, so it is highly significant, P-value is less than 0.05, this will reject the null hypothesis.
- The T-statistic value must be greater than 2(or less than -2) which indicates the coefficient is significant with >95% coincidence.
- **t** value is significant for medv = 61.41
- **t** value is significant = -24.41

Step 4(1 mark) The model output list can be checked by using command names.

OUTPUT

```
> names(Boston)
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"
 [7] "age"     "dis"     "rad"     "tax"     "ptratio" "black"
[13] "lstat"   "medv"
```

Step 5(2 mark) Plot medvandlstatscatter plot using the plot() command. Plot the least squares regression in red line usingabline() functions, use width parameter lwdof regression line as 3.Report these two graphs.

Command:

attach(Boston)

plot(lstat, medv , main="Scatterplot",

xlab="lstat", ylab="medv ")

pdf (" scatterplot1.pdf ")

plot(lstat, medv)

dev.off ()

OUTPUT

```
> plot(lstat, medv , main="Scatterplot",
+      xlab="lstat", ylab="medv ")
>
> pdf (" scatterplot1.pdf ")
> plot(lstat, medv)
> dev.off ()
RStudioGD
        2
```

COMMAND

attach(Boston)

plot(lstat, medv , main="Scatterplot",

xlab="lstat ", ylab="medv ")

abline(lm(medv~lstat),lwd =3, col="red")

pdf (" scatterplot2.pdf ")

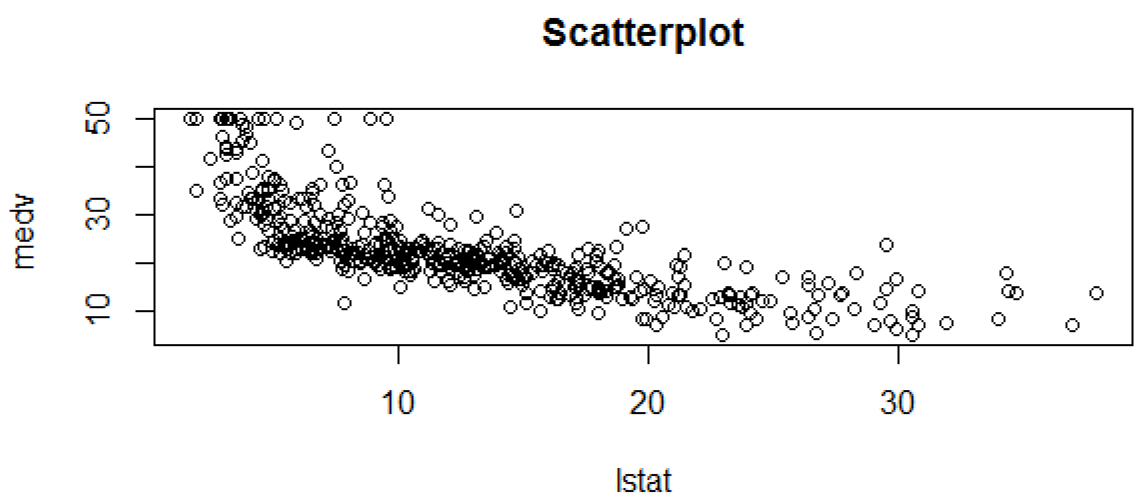
plot(lstat, medv ,col =" red ")

dev.off ()

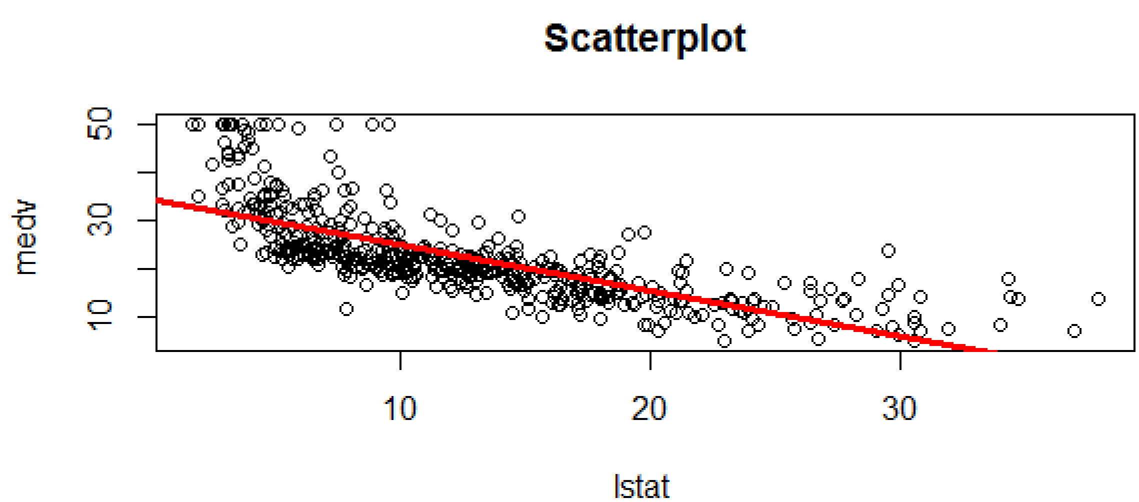
Output:-

```
> plot(lstat, medv , main="Scatterplot",
+      xlab="lstat ", ylab="medv ")
> abline(lm(medv~lstat),lwd =3, col="red")
>
> pdf (" scatterplot2.pdf ")
> plot(lstat, medv ,col =" red ")
> dev.off ()
RStudioGD
        2
```

OUTPUT



OUTPUT



Multiple linear regression

In order to fit a multiple linear regression model using least squares, we again use the lm() function. The syntax lm(y∼x1+x2+x3) is used to fit a model with three predictors, x1, x2, and x3. The summary() function now outputs the regression coefficients for all the predictors.

Step 1(1 mark) Built the regression line of medv variable against lstatand age. Print the summaryoutput of this regression model.

COMMAND:

fit <- lm(medv~lstat+age, data = Boston)
summary(fit)

OUTPUT

```
> fit <- lm(medv~lstat+age, data = Boston)
> summary(fit)

Call:
lm(formula = medv ~ lstat + age, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.981  -3.978  -1.283   1.968  23.158 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
age          0.03454    0.01223   2.826  0.00491 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.173 on 503 degrees of freedom
Multiple R-squared:  0.5513,	Adjusted R-squared:  0.5495 
F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

Step 2(1 mark) Analyze the overall significance of the model.

OBSERVATION:

R square Value:

- R squares measures the variability of the data is captured by the Model.
- R-squares value = 0.5513(55.13%).

F-value

- F- statistic value is 309, which is larger than 1 and we can reject null hypothesis. Also it indicates that there would be relation between the response and predictors.

Step 3(1 mark) Analyze the significance of each individual coefficient.

OBSERVATION:

Y(medv)= 33.22276-1.03207*lstat+0.03454*age

***β*0** = 33.22276 =intercept. It is independent of any predictors.

***β*1** = -1.03207; We can inferred that the response and the predictors are negatively correlated. If any of these will increases the other decreases.

***β*2**= 0.03454; We can inferred that response and the predictors are positively correlated. If any of these will increases the other increases as well.