**Analytics for Business Intelligence**

Last name: **Bhatnagar** First name: **Meru** Section number : 40

**Home Work 1** Total marks are 15 which provide 7.5% to the total assessment. Students must implement the homework using R language, cut-and-paste of outputs from text book are not permitted (and easily detectable).

The Steps of this homework should be implemented using RStudio and printouts inserted in this document after each step.

Data representation and 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.

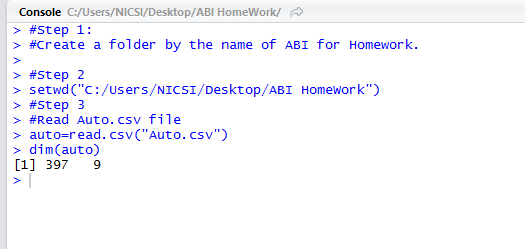
**Step 2** Open R- studio. In the command prompt select “Session” then “Set working directory” and “Choose directory”. Using the command prompt with file manager select the directory you created in Step 1.



**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:



Output:



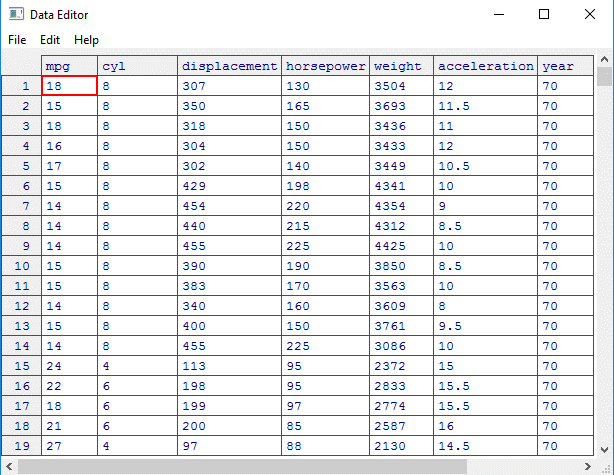
397 observations & 9 variables.

**Step 4** (1 mark) Use command fix to 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

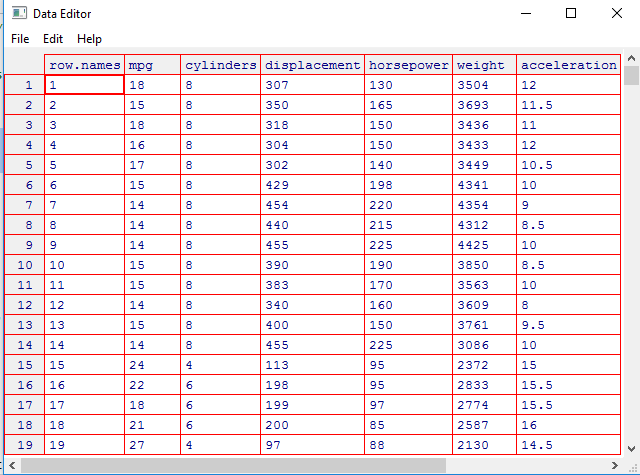
variables. Report the output below:



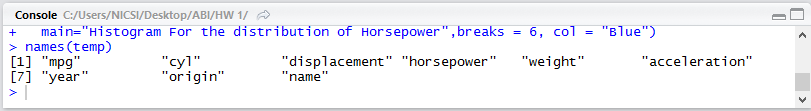
Output1:



Output2:



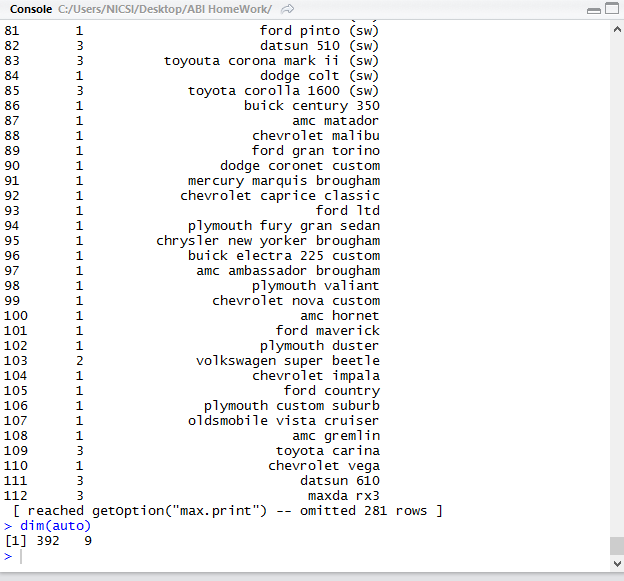
Output3:



**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?



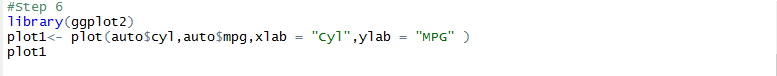
Output:

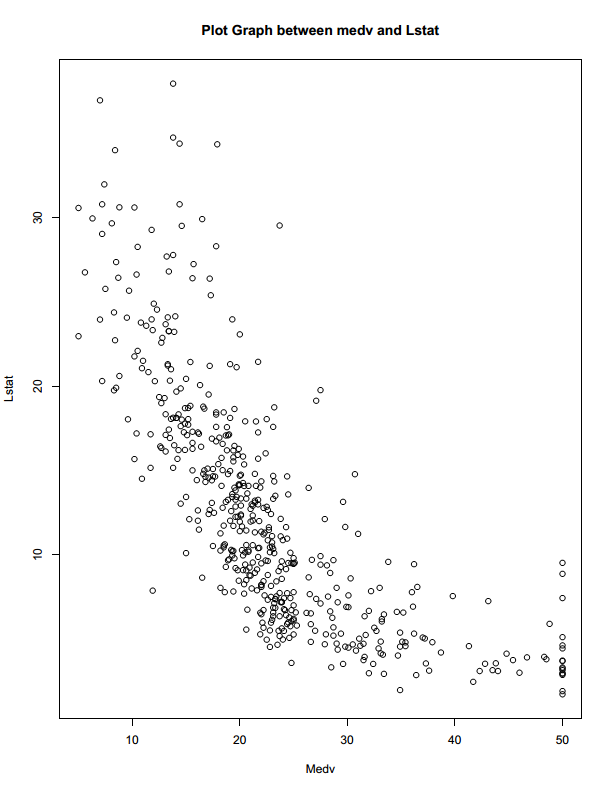


Inference: We have 5 missing rows in the auto.csv file which we have deleted using na.omit() function.

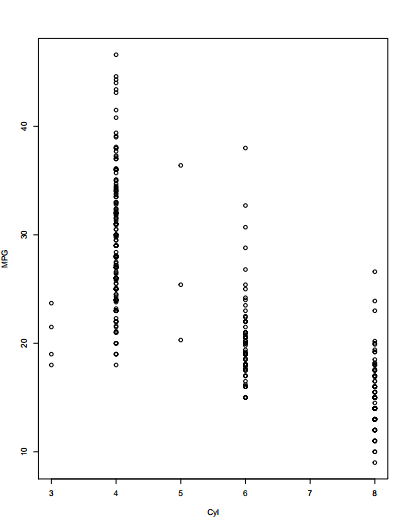
Now we are left with 392 observations and 9 variables.

**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.

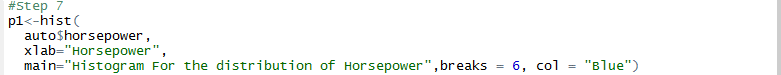




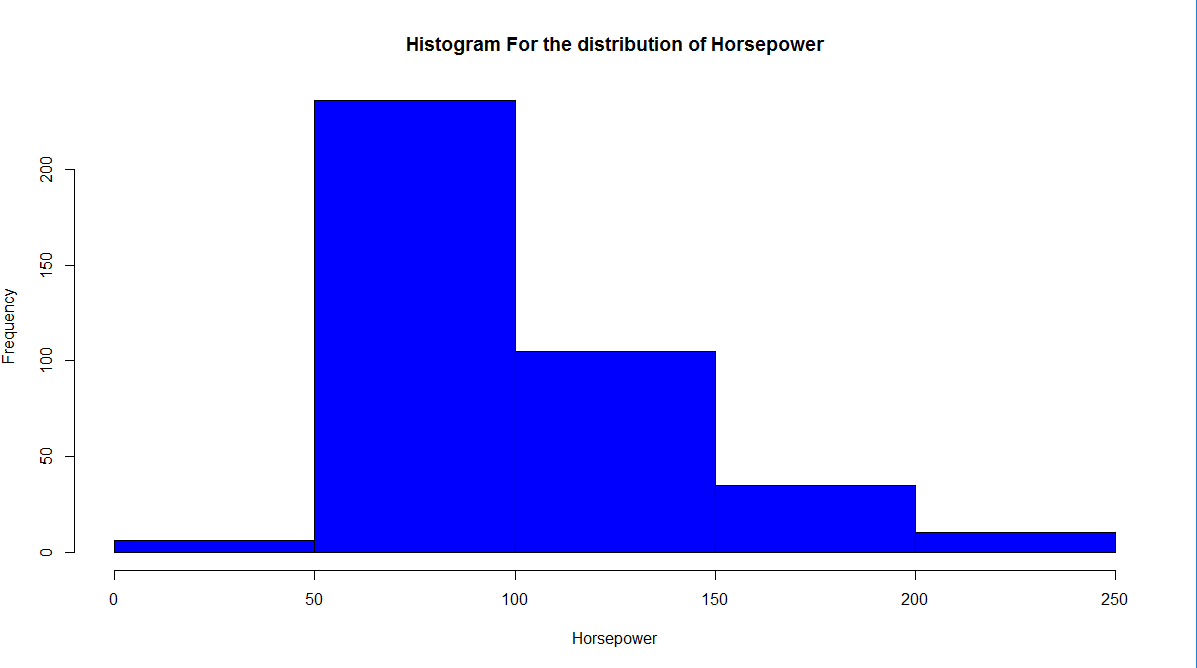
Output:



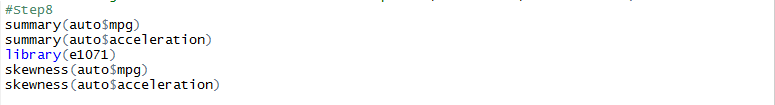
**Step 7** (1 mark) Produce the histogram for horsepower variable using hist() command. Use blue color (col) and 6 bars (breaks).



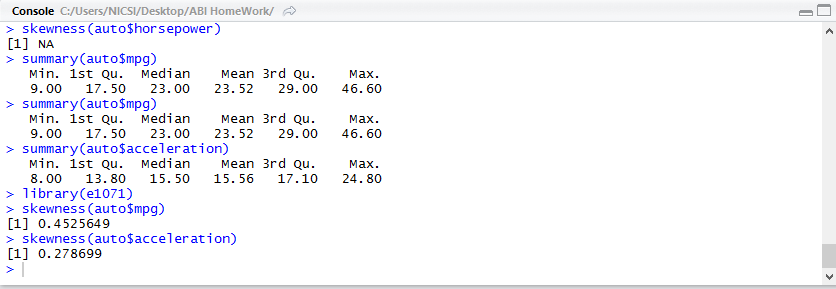
Output:



**Step 8** (1 mark) Produce the statistical summary form mpg and acceleration variables using summary() command. Are these data samples symmetric or skewed.



Output:

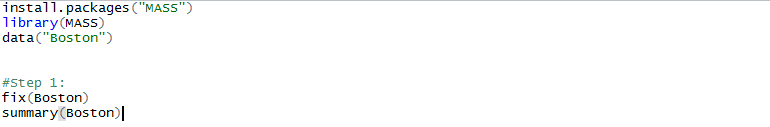


Inference: Both the variables mpg and acceleration are positively skewed as shown in the console using skewness function. We can also check the mean and median value for skewness. Since both the variable are not having common mean and median so it is safe to conclude that the variables are skewed.

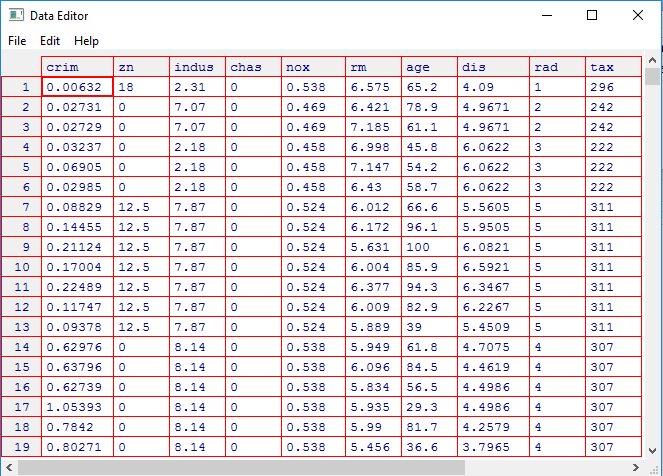
Simple (one variable) linear regression

You need to connect MASS library using the command library(MASS). The Boston data is part of the MASS library. The MASS library contains the Boston data set, which records medv (median house value) for 506 neighborhoods around Boston. We will seek to predict medv using 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).

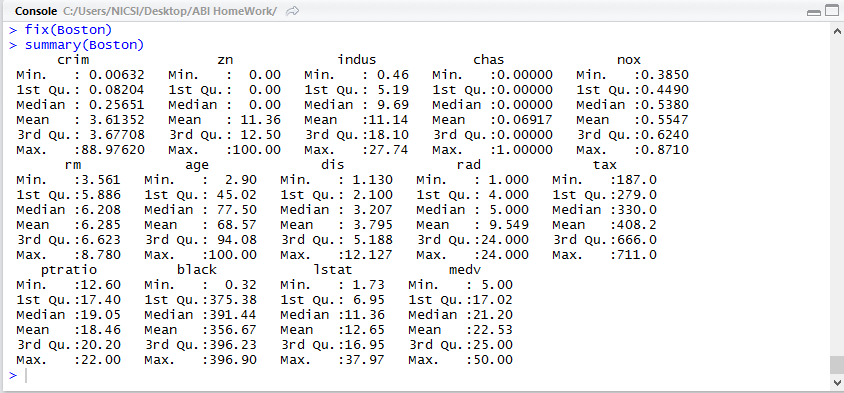
**Step 1** (1 mark) Have a look to this data using fix command and print summary statistics for all variables.



1. Output:1



1. Output 2:

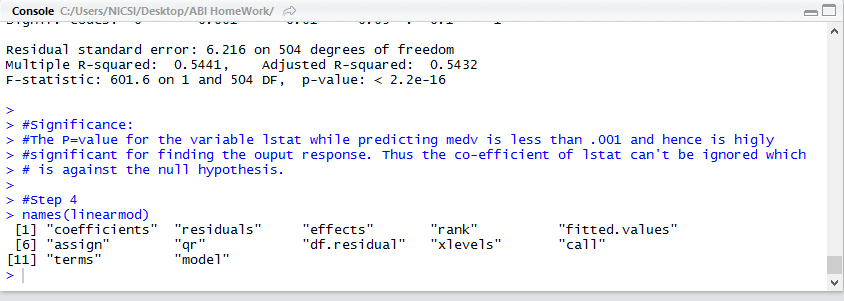


|  |
| --- |
| **Step 2** (1 mark) Start by using the lm() function to fit a simple linear regression model, with medv as the response and lstat as the predictor. The basic syntax is lm(y*∼*x,data), where y is the response, x is the predictor, and data is the data set in which these two variables are kept. Do not forget to attach Boston data either by separate command or by option “data” in lm command. |
| Output: |
| **Step 3** (1 mark) Interpret the significance of regression model.  The P-value for the variable lstat while predicting medv is less than .001 and hence is highly significant for finding the output response. Thus the co-efficient of lstat can't be ignored and the equation comes out to be:  Y(Medv) = 34.55384 - .95005\*lstat  This means that for every 1% increase in lstat, the y(medv) value will get decrease by .95%. |

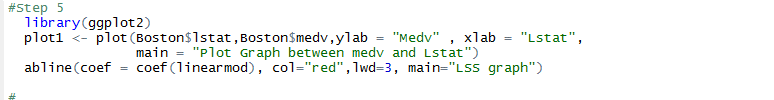
**Step 4** (1 mark) The model output list can be checked by using command names.



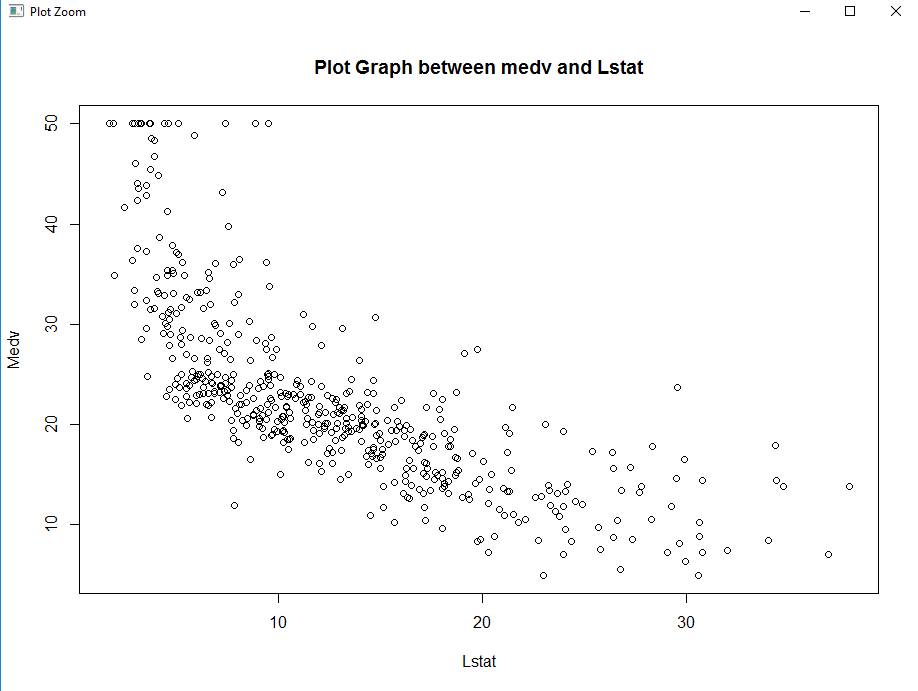
Output:



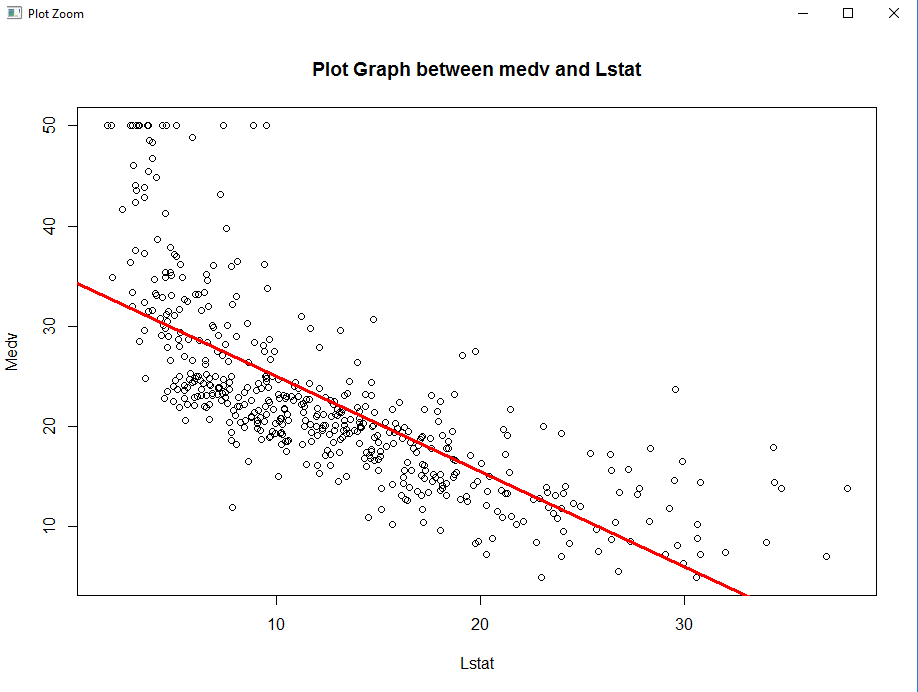
**Step 5** (2 mark) Plot medv and lstat scatter plot using the plot() command. Plot the least squares regression in red line using abline() functions, use width parameter lwd of regression line as 3. Report these two graphs.



Output:



**1a. Medv Vs Lstat Scatter Plot**



**1b. Medv Vs Lstat Scatter Plot with Lease Square Regression Line**

Multiple linear regression

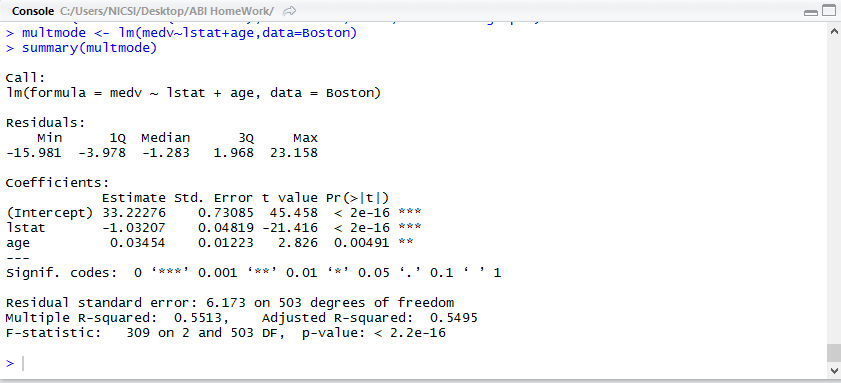
In order to fit a multiple linear regression model using least squares, we again use the lm() function. The syntax lm(y*∼*x1+x2+x3) is used to fit a model with three predictors, x1, x2, and x3. The summary() function now outputs the regression coefficients for all the predictors.

**Step 1** (1 mark) Built the regression line of medv variable against lstat and age. Print the

summary output of this regression model.



Output:



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

The p-values intercept is nearly equal to 0. (2e-16). This clearly defines that the value of response is dependent on atleast a single variable which supports the alternative hypothesis.

Also, overall T-value is relatively large which provides a compelling evidence against null hypothesis. The multi linear regression equation comes out to be:

Y(medv) = 33.22276-1.03207\*lstat + .03454\*age

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

The p-values of lstat variable is nearly equal to 0. (2e-16). Due to negligible p-value the coefficient of lstat is highly significant to predict the medv output. However, the p value of age is a little less than .05 (.004) which does not make it a highly significant factor to predict the response (medy).

Also, T-value of lstat is negative and very less than -2 which suggest that there is strong inverse relationship between lstat and medv with 95% confidence while T-value for age is positive but slightly greater than +2 which indicates a weak direct relationship between age and medv with 95% confidence.