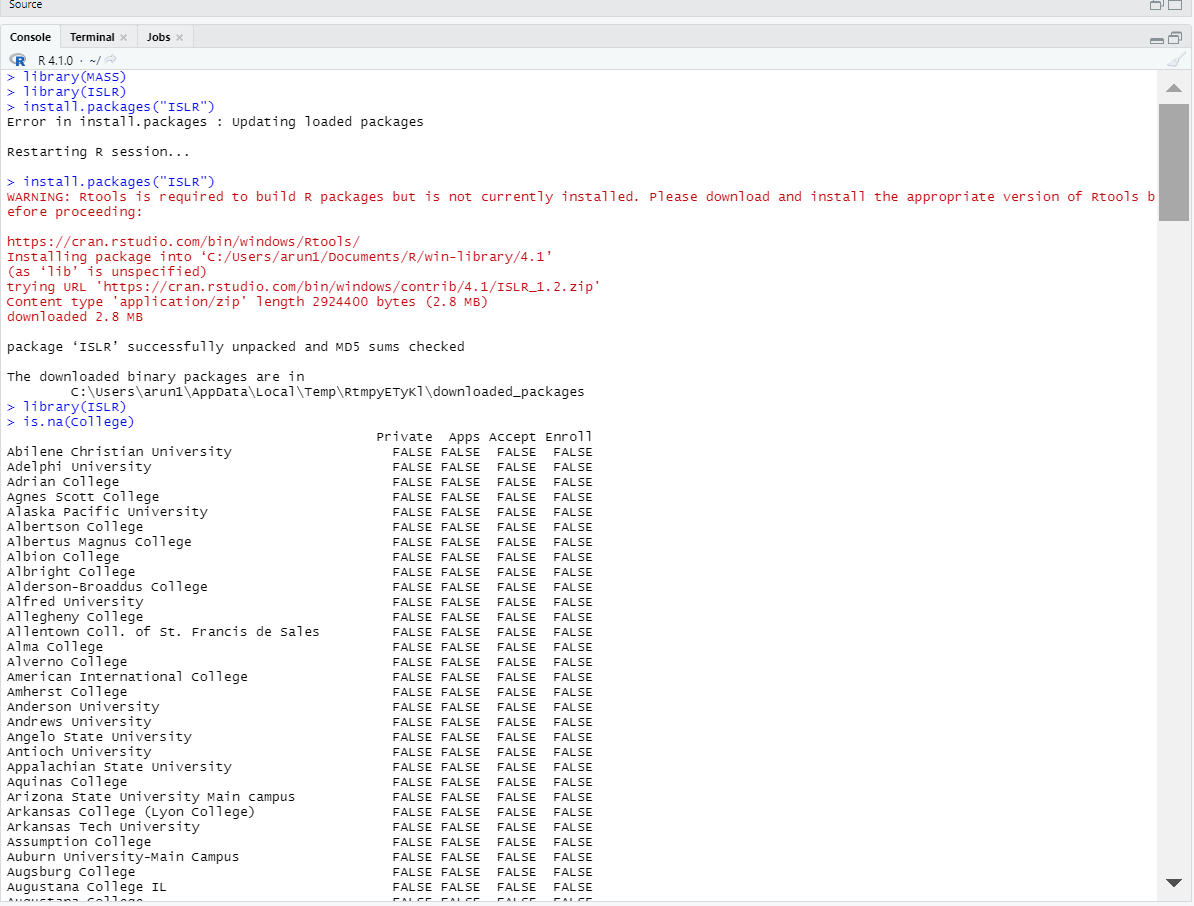
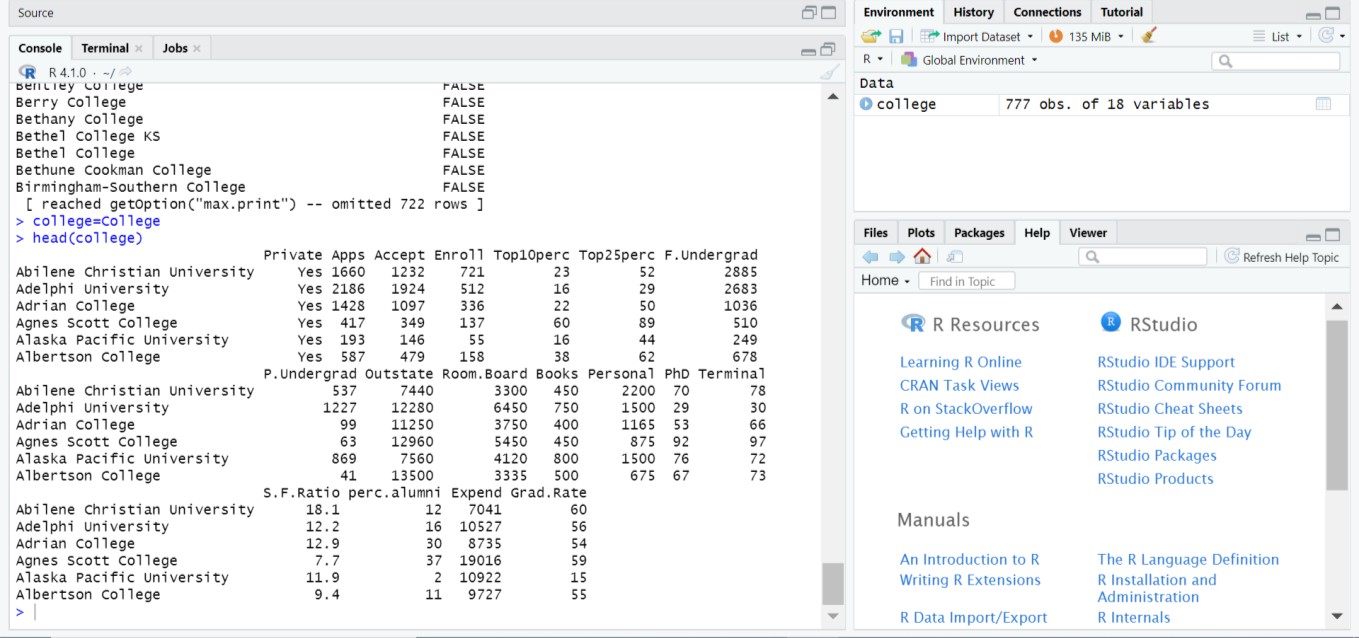
# Introduction to Statistical Learning

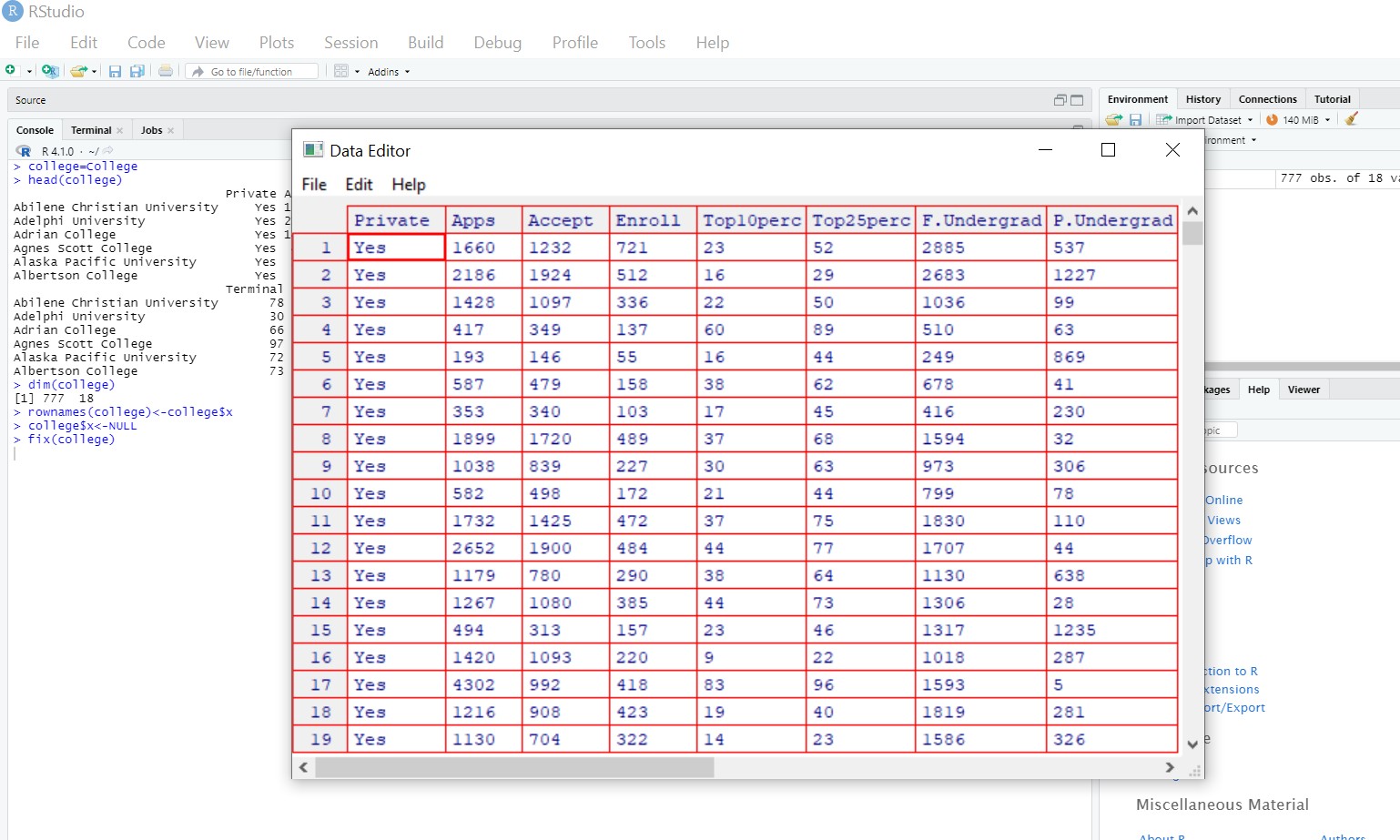
# Viswanadh Gandham---16320368

1. **ISLR 2.4 Applied Problem 8.**
   1. Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

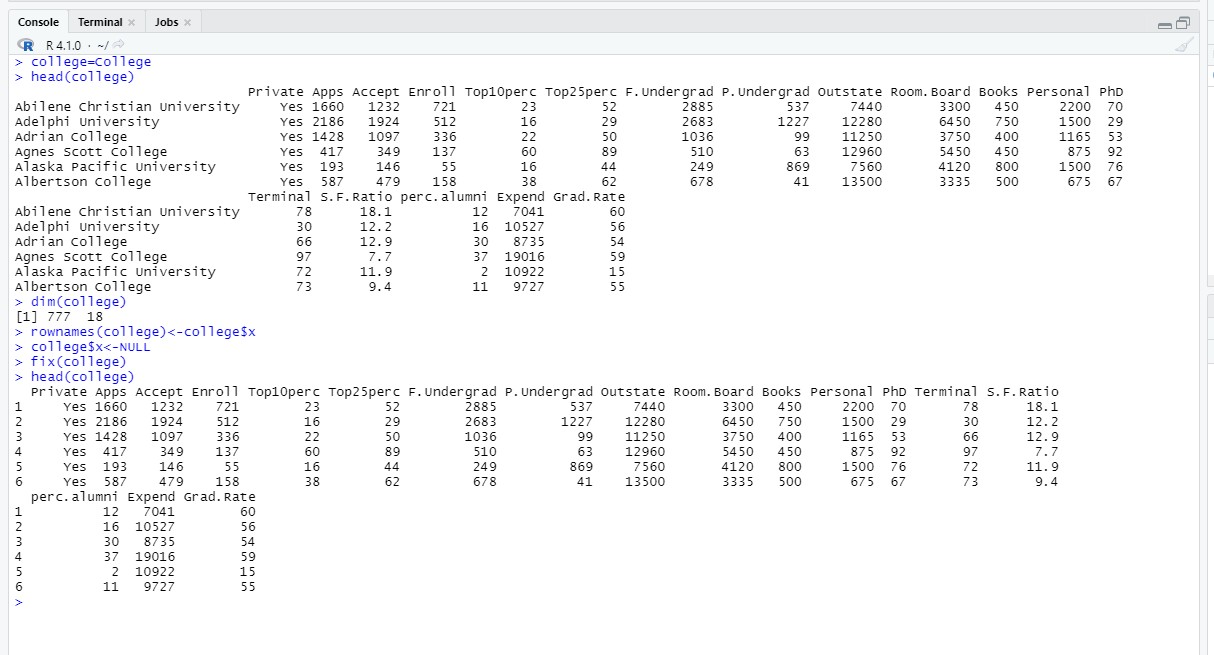


* 1. Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

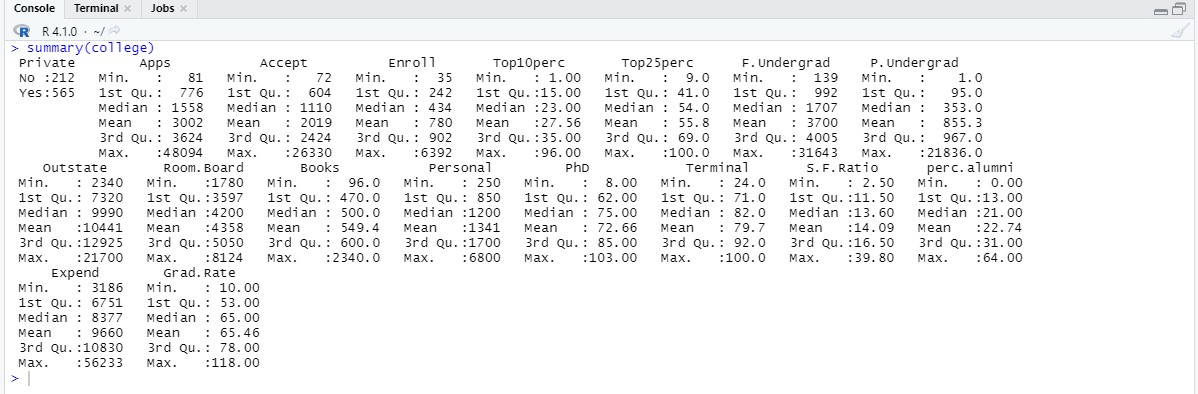




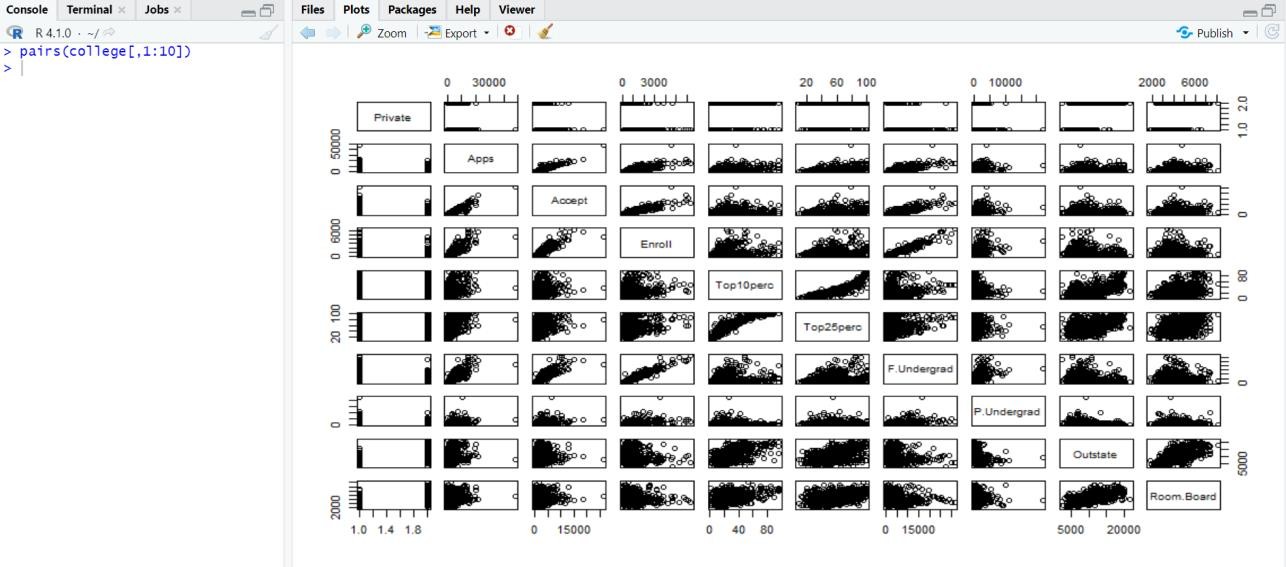
Then we finally got the modified dataset after removal of $X as



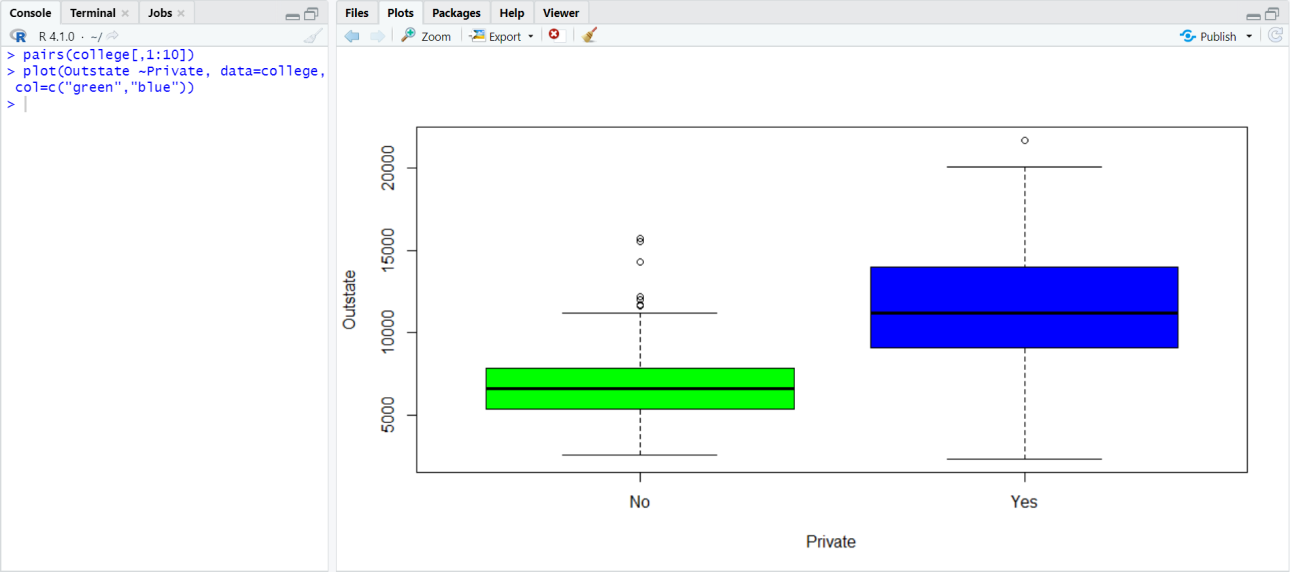
* 1. (c)
     1. Use the summary() function to produce a numerical summary of the variables in the data set.



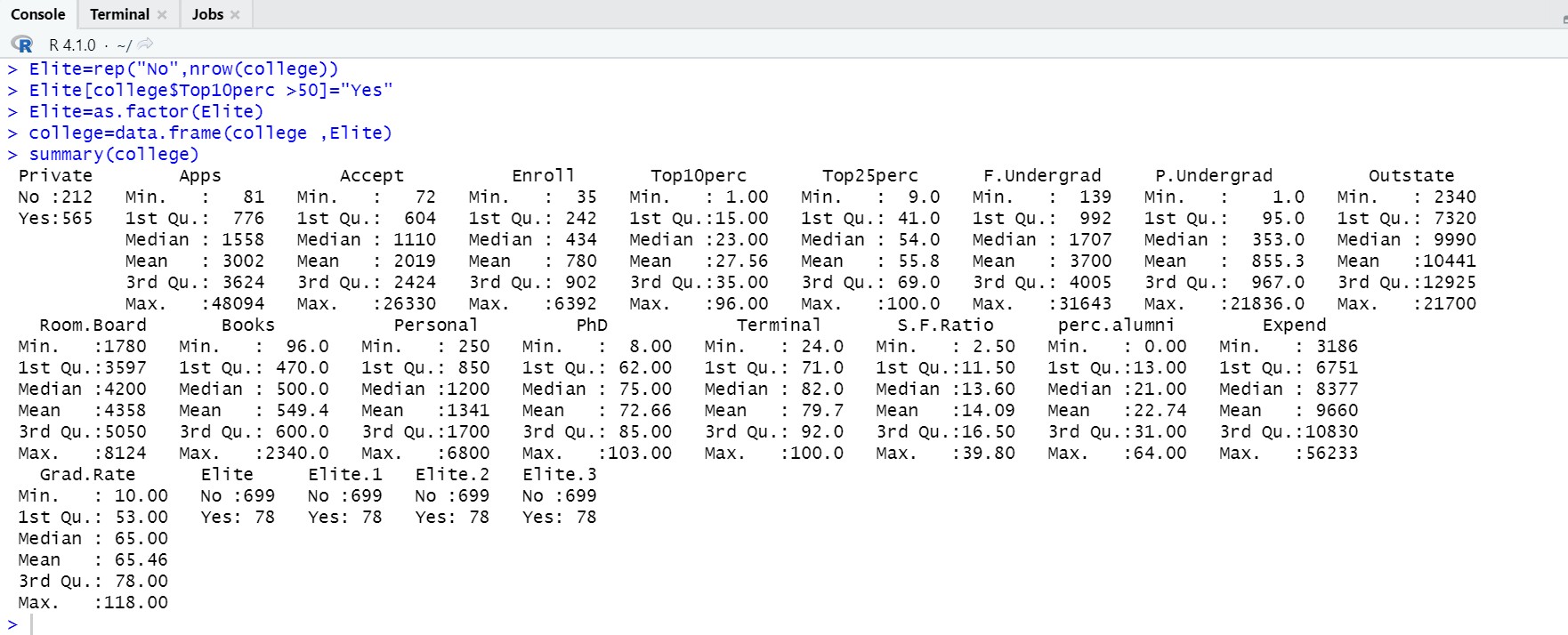
* + 1. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

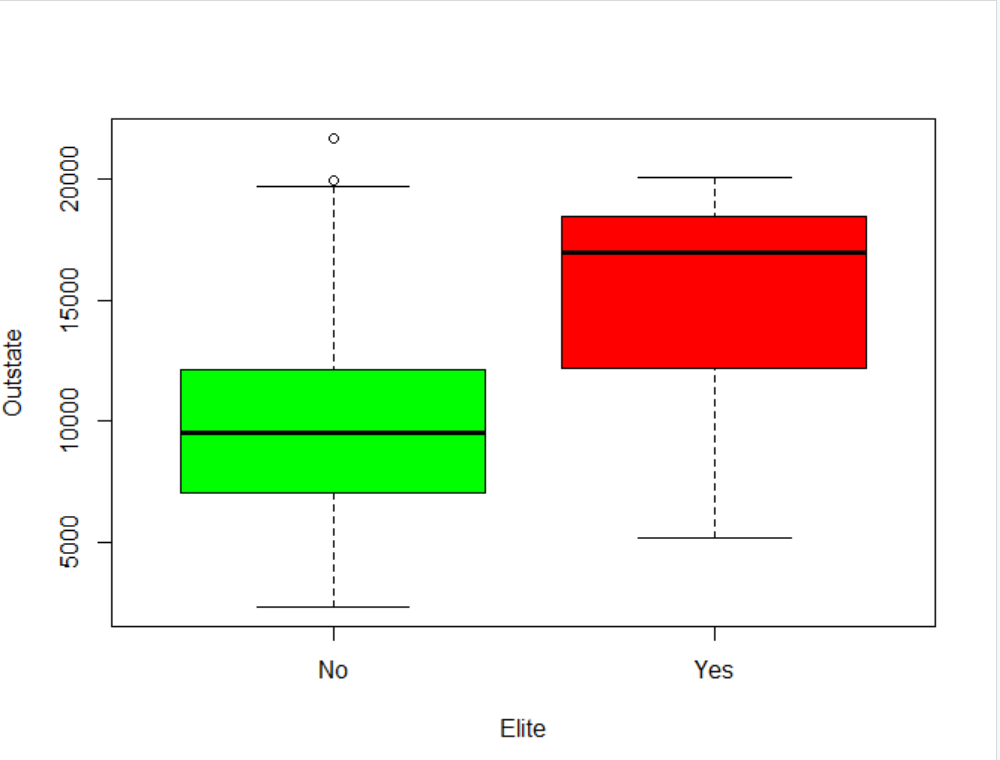
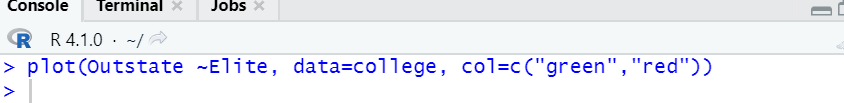


* + 1. Use the plot() function to produce side-by-side boxplots of Outstate versus Private

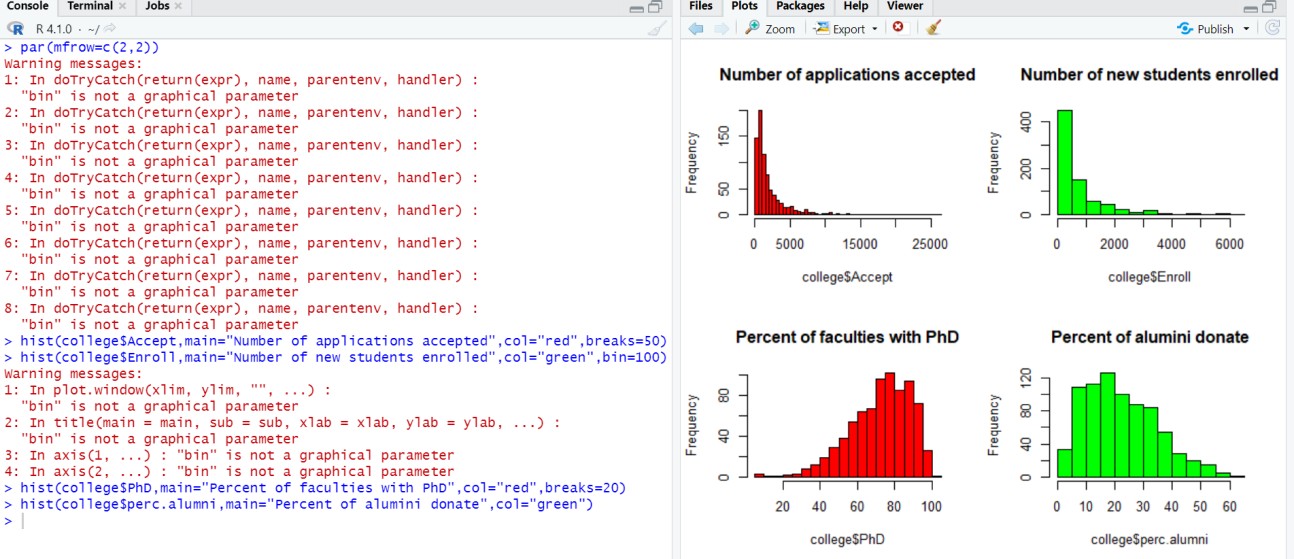


* + 1. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

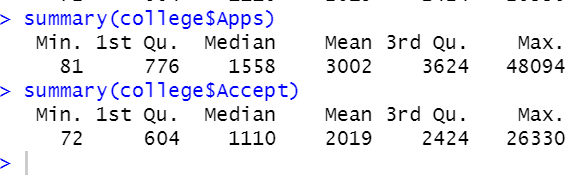




## Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.



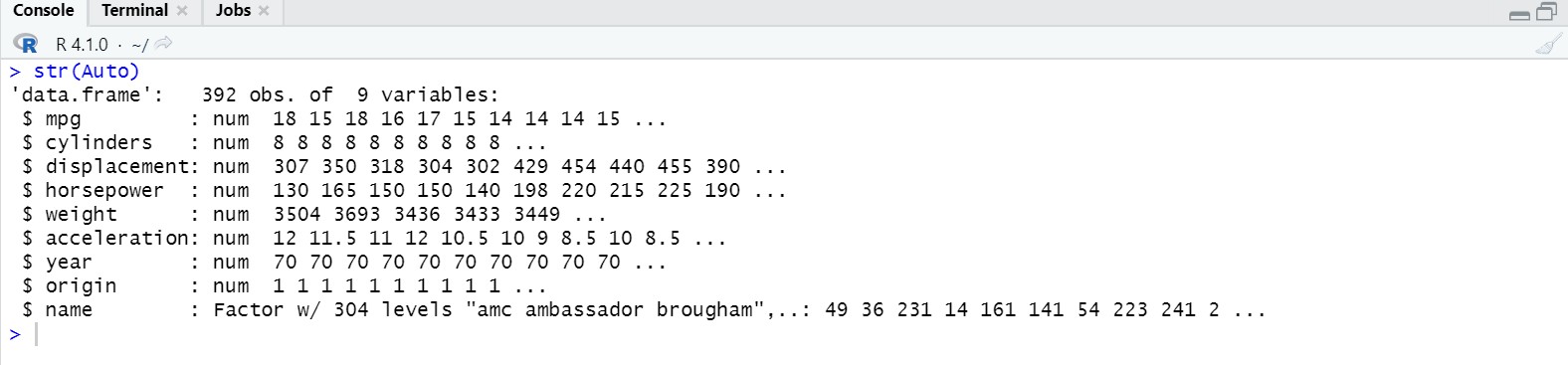
* + 1. Continue exploring the data, and provide a brief summary of what you discover.



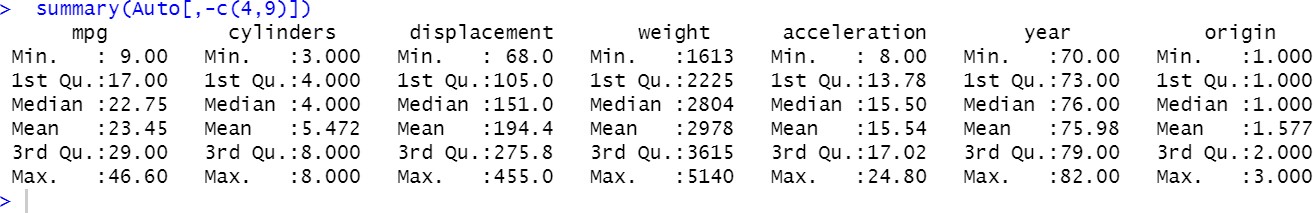
### ISLR 2.4 Applied Problem 9

This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

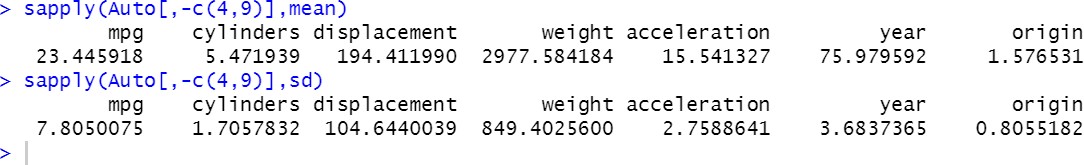
* 1. Which of the predictors are quantitative, and which are qualitative?



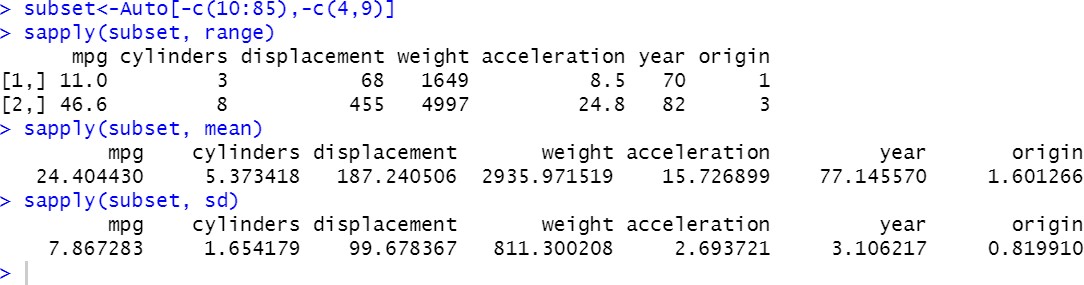
* 1. What is the range of each quantitative predictor? You can answer this using the range() function



* 1. What is the mean and standard deviation of each quantitative predictor?



* 1. Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?



* 1. Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings
* Auto$cy1indersx-as.factor(Auto$cylinders)
* Auto$year<-as. factor (Au to $year)
* Au to$or i gi n<-as. factor (Au to$or i g1n)
* pa1rs{Auto)

Files PIoD Packages Help Viewer

yW DOOR - hXgOFt ”



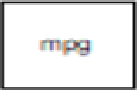
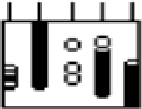
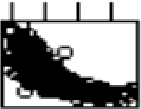
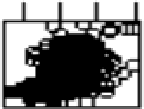
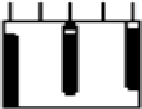
publish

1 3 5

50 *200*

10 ZO

4 . 0 2. 5

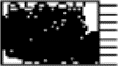
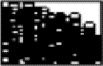
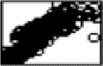
        

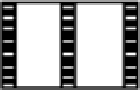
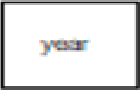
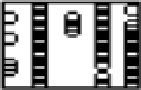
       

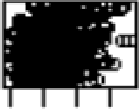
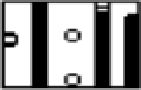
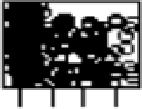
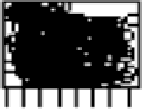
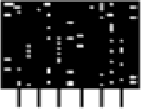
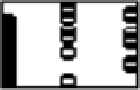






40 40

TOO 400

1500 SOOO

2 8 0 200

* cor{Au to$we 1gh t , Au to$horsepowe r) [1] 0. 864 5 3 77
* cor{Au to$di sp1acenent , Au to$hor sepower )

[1] 0. 89 7257

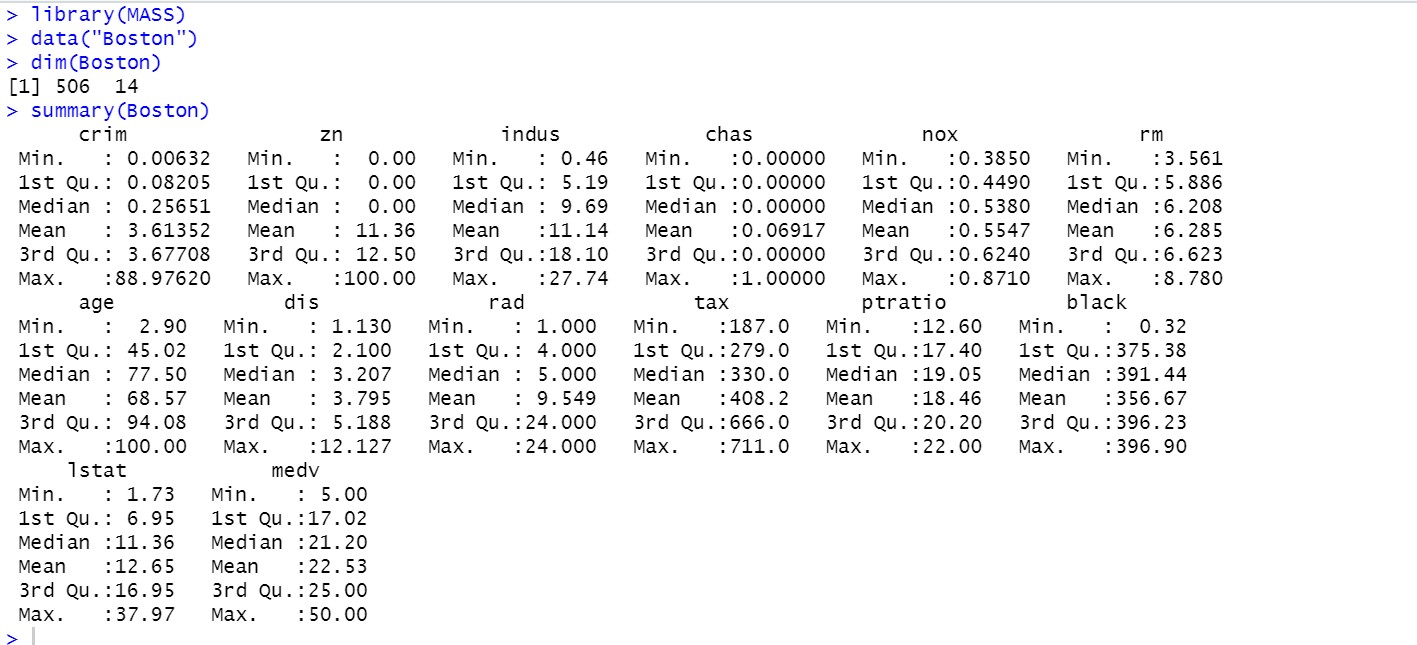
* cor{Au to$we 1gh t , Au to$dl sp1acemen t)

{1] 0. 93 29944

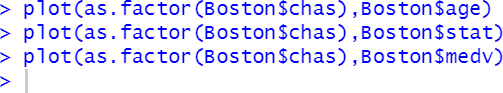
### ISLR 2.4 Applied Problem 10

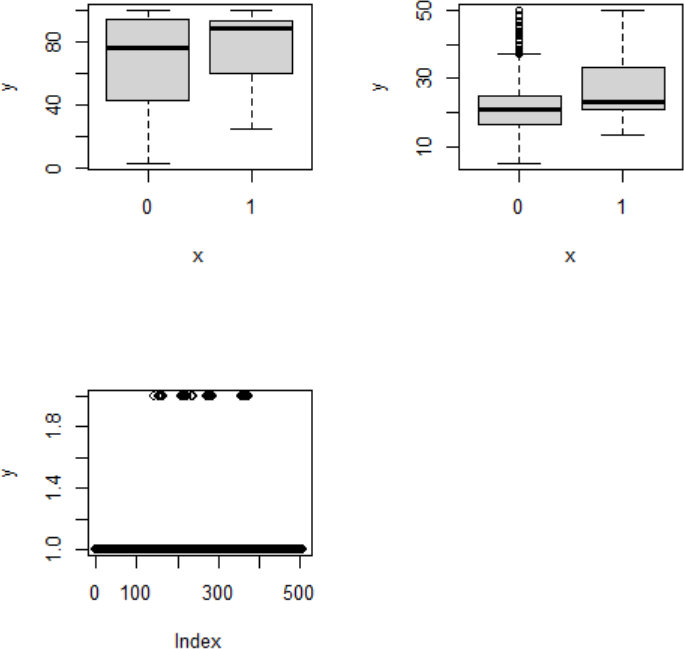
This exercise involves the Boston housing data set.

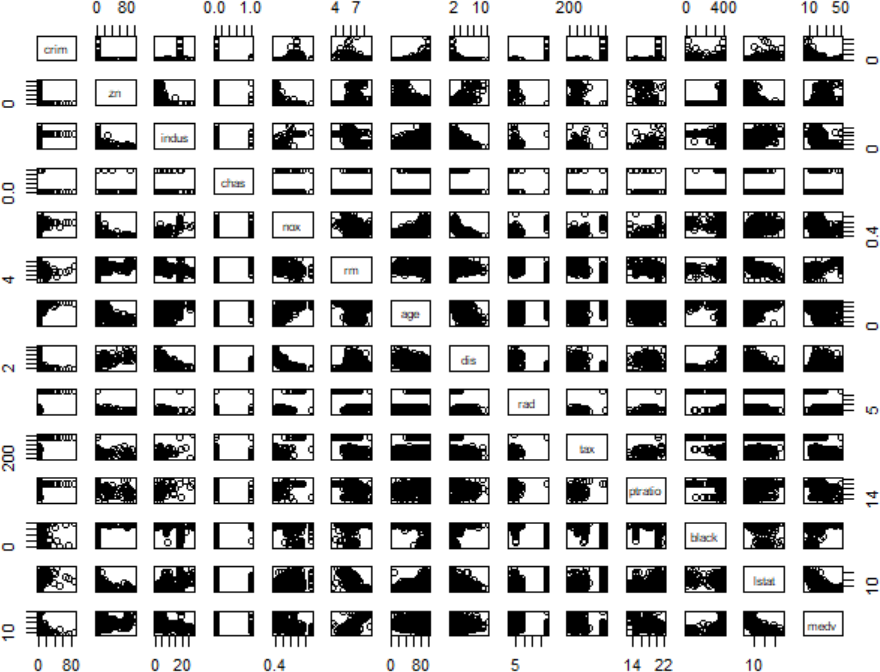
* 1. To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.



* 1. Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

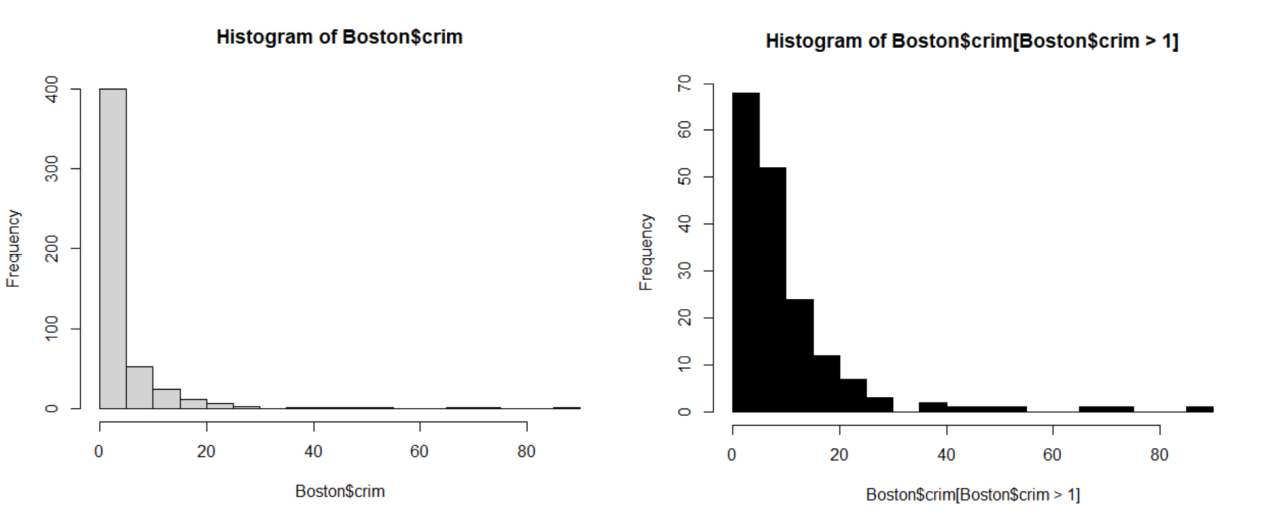




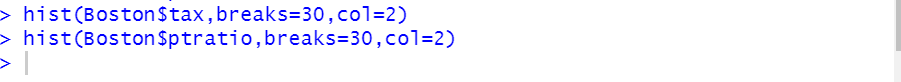


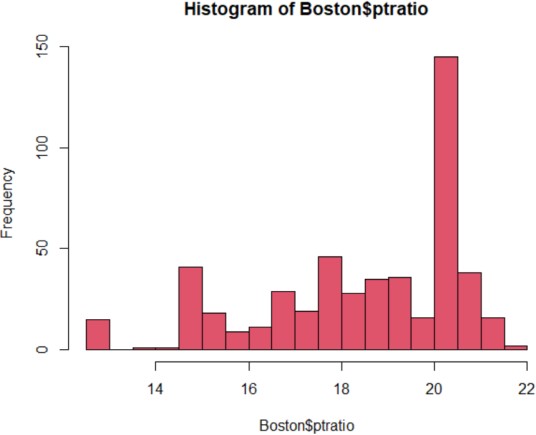
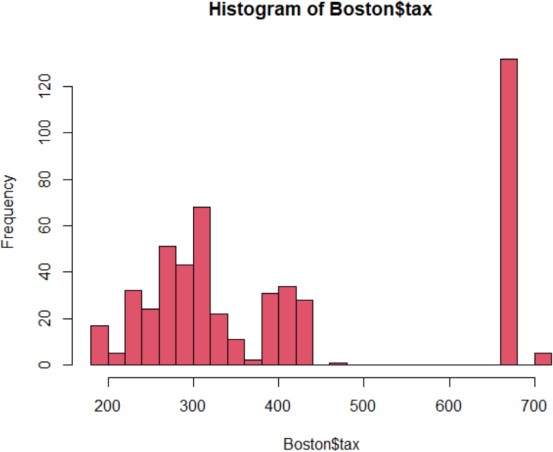
From the graph we can conclude that medv and black are inversely proportional. (c ) Are any of the predictors associated with per capita crime rate?





(D) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.





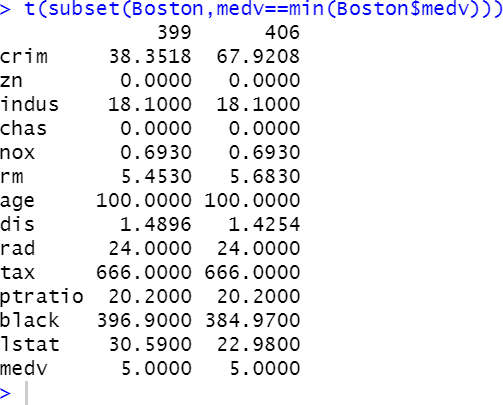
1. How many of the suburbs in this data set bound the Charles river?



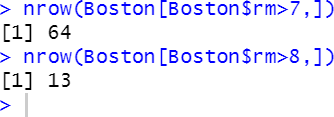
1. What is the median pupil-teacher ratio among the towns in this data set?



1. Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

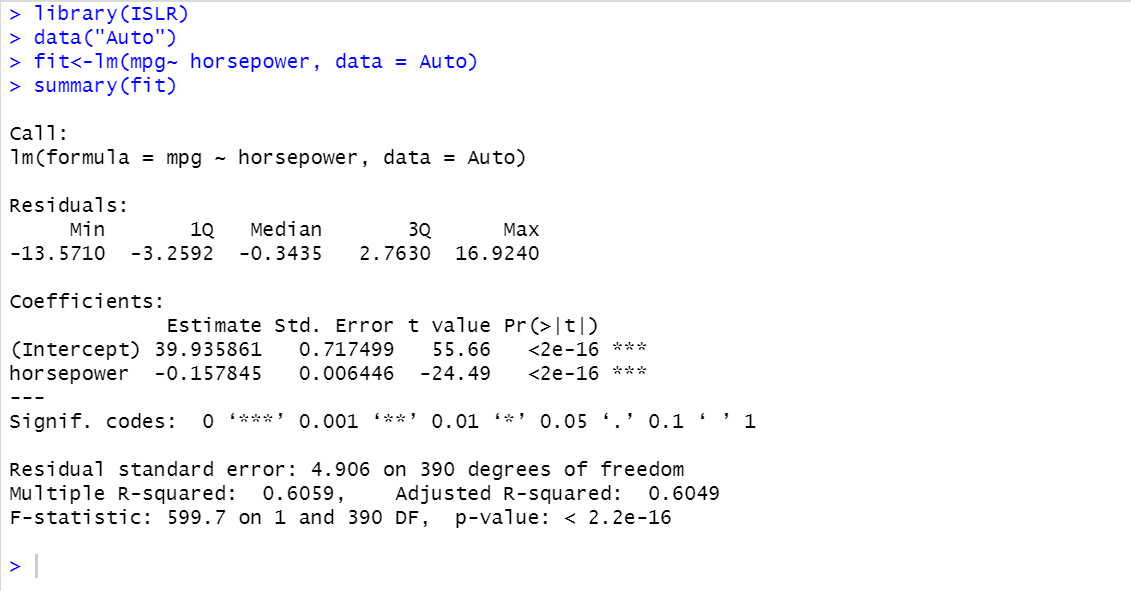


1. In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.



# ISLR 3.7 Applied Problem 8.

1. Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.
2. Is there a relationship between the predictor and the response?



As per the hypothesis testing, The p-value from the above screenshot corresponding to the F-statistic is 7.03198910^{-81}, this indicates a clear evidence of a relationship between “mpg” and “horsepower”.

## How strong is the relationship between the predictor and the response?

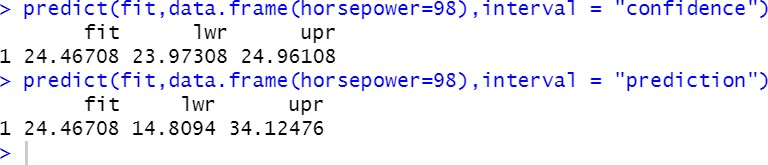
To calculate the leftover blunder relative to the reaction we utilize the cruel of the reaction and the RSE. The cruel of mpg is 23.4459184. The RSE of the lm.fit was 4.9057569 which demonstrates a rate mistake of 20.9237141%. We may moreover note that as the R2 is rise to 0.6059483, nearly 60.5948258% of the inconstancy in “mpg” can be clarified utilizing “horsepower”.

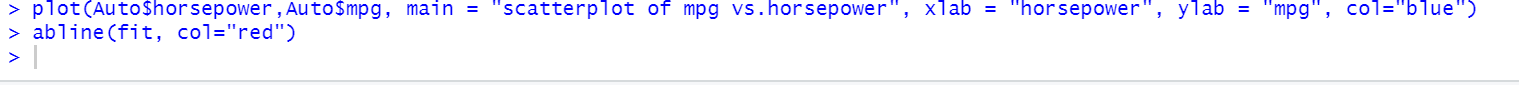
## Is the relationship between the predictor and the response positive or negative?

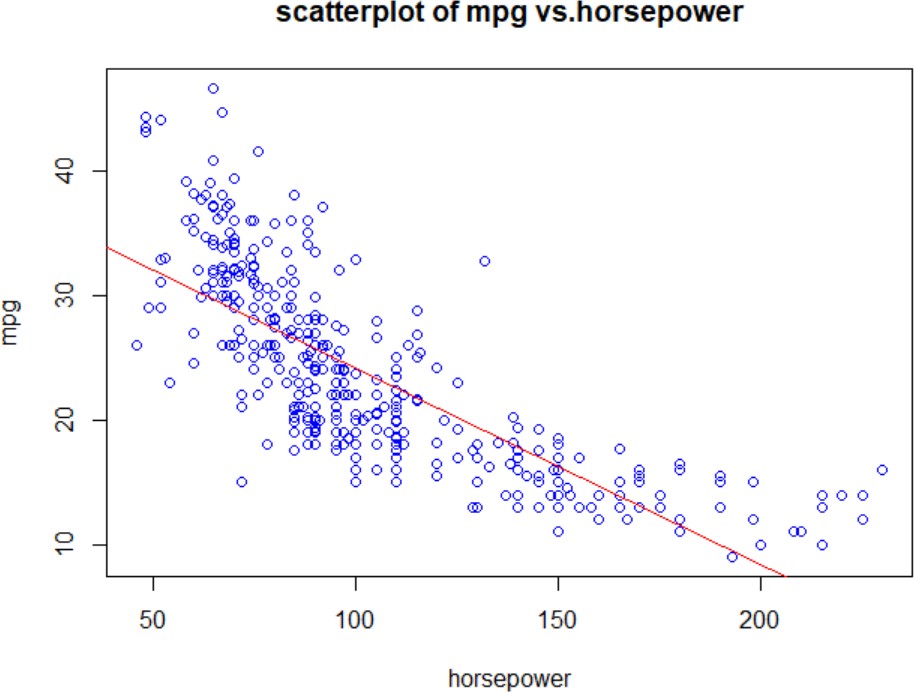
As the coefficient of “horsepower” is negative, the relationship is additionally negative. The more drive a vehicle has the direct relapse shows the less mpg fuel proficiency the vehicle will have.

## What is the predicted mpgmpg associated with a “horsepower” of 98 ? What are the

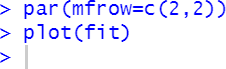
associated 95% confidence and prediction intervals ?

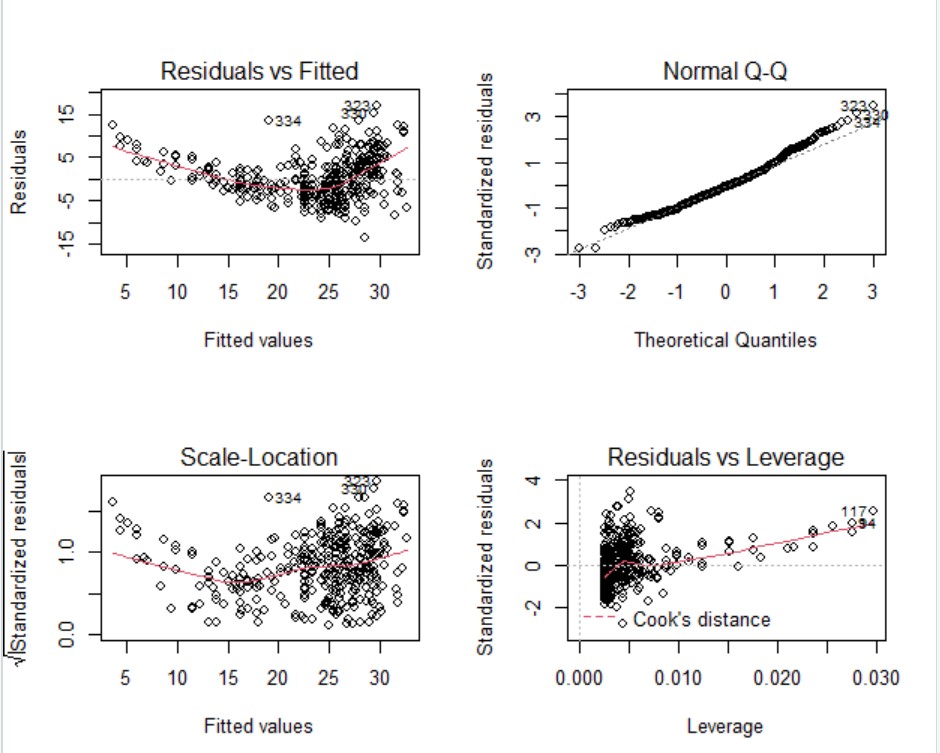


1. Plot the response and the predictor. Use the abline() function to display the least squares regression line.



1. Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.





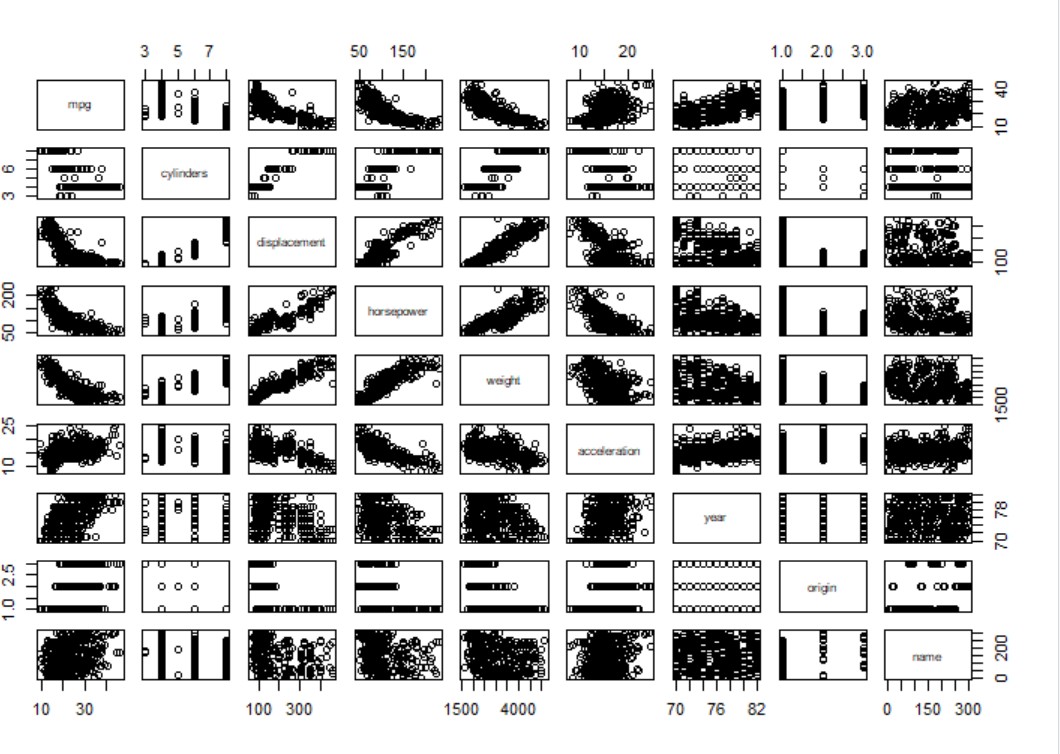
## The plot of residuals versus fitted values indicates the presence of non-linearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and a few high leverage points.

1. **ISLR 3.7 Applied Problem 9.**

This question involves the use of multiple linear regression on the Auto data set.

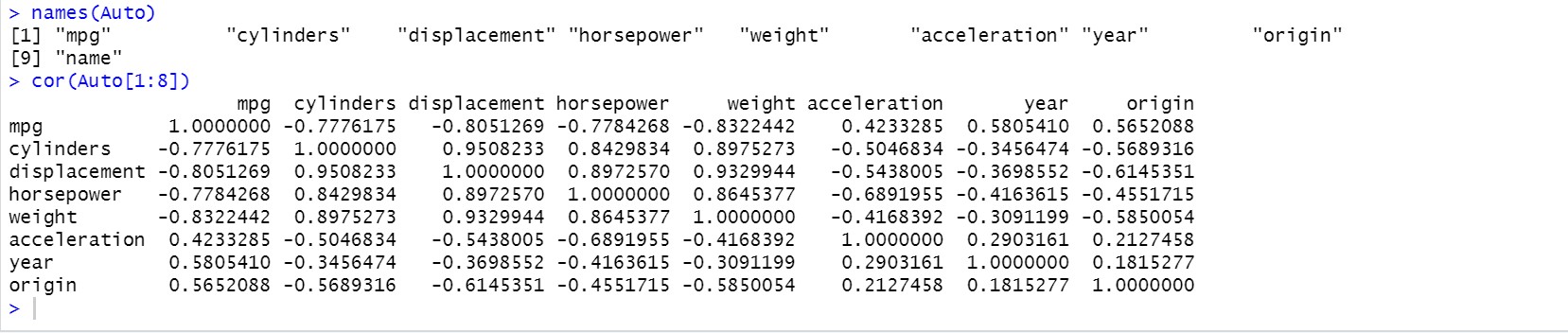
* 1. Produce a scatterplot matrix which includes all of the variables in the data set.





## Compute the matrix of correlations between the variables using the function cor(). You

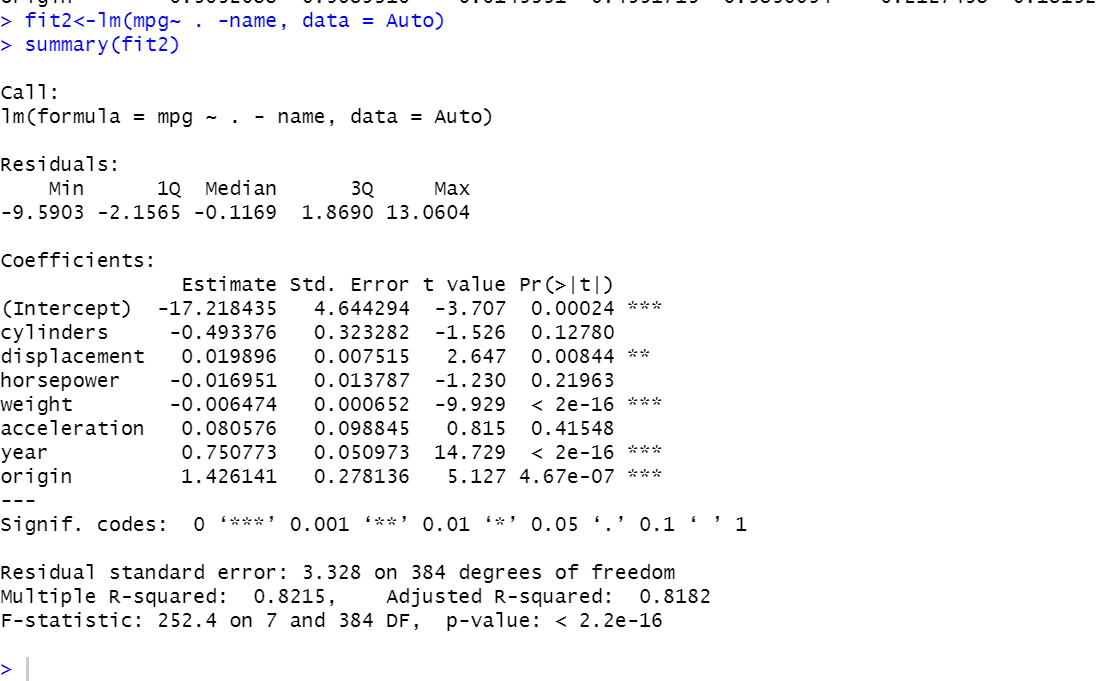
will need to exclude the “name” variable, which is qualitative.



* 1. 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. For instance :

* + 1. Is there a relationship between the predictors and the response ?



By considering the hypothesis testing, the p-value corresponding to the F-statistic is 2.037105910^ {-139}, this indicates a clear evidence of a

relationship between “mpg” and the input predictors.

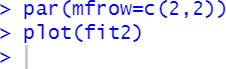
* + 1. Which predictors appear to have a statistically significant relationship to the response?

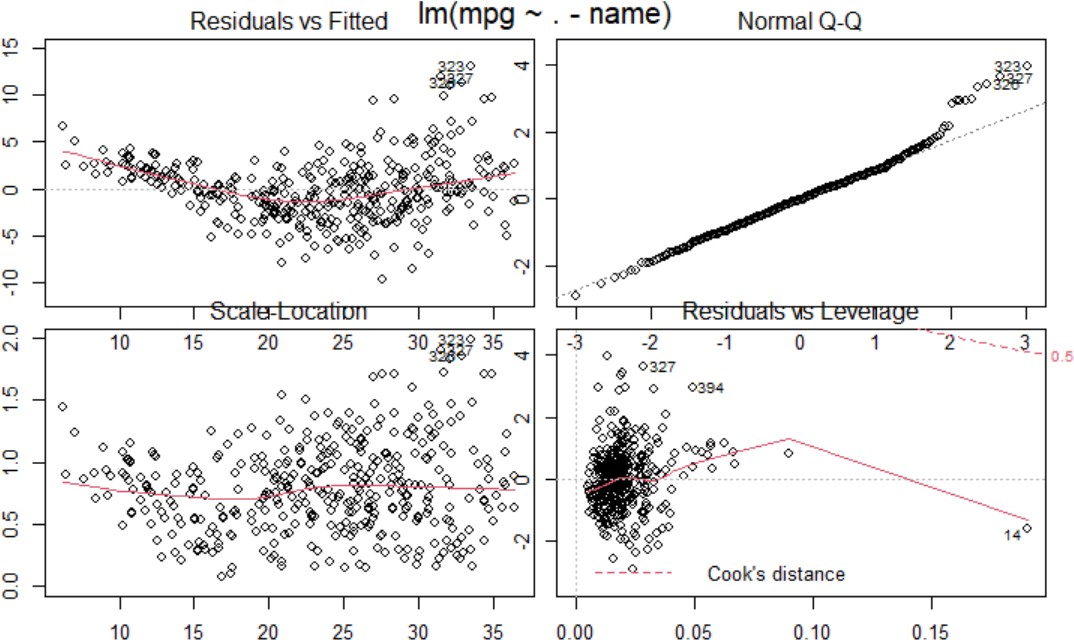
Able to reply this address by checking the p-values related with each predictor’s t-statistic. We may conclude that all indicators are measurably noteworthy but “cylinders”, “horsepower” and “acceleration”.

## What does the coefficient for the “year” variable suggest ?

The coefficient of variable “year” suggesting that a unit increase in year increases 0.75 times the miles per gallon(mpg). With we can say that, cars are becoming fuel efficient year by year.

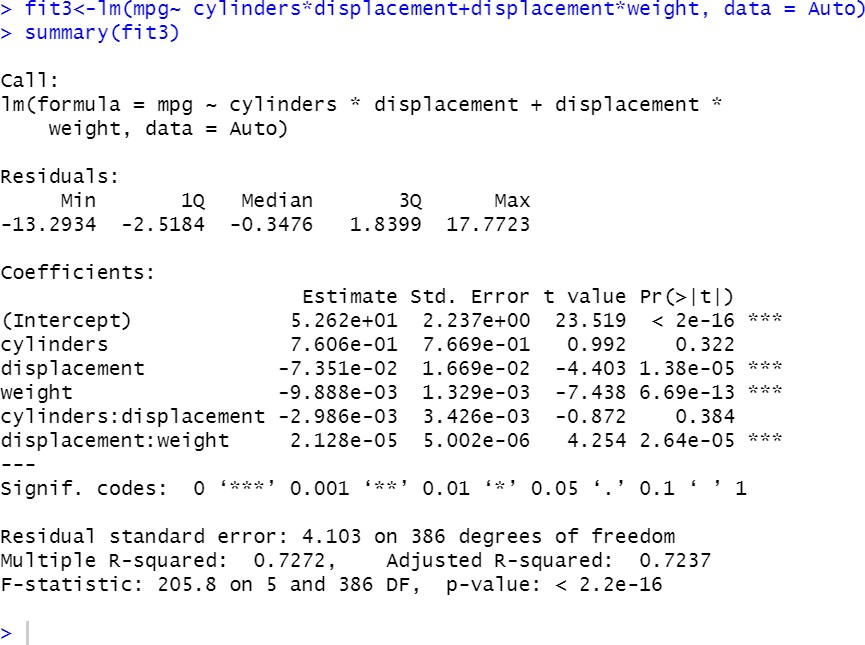
* 1. Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers ? Does the leverage plots identify any observations with unusually high leverages ?





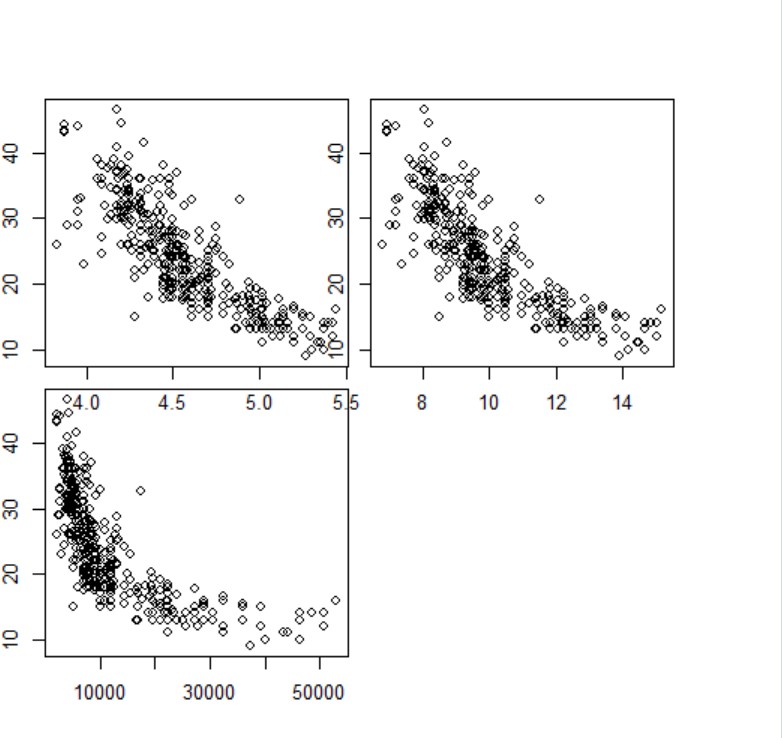
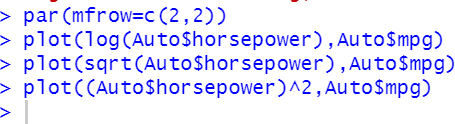
From the above screenshots, we can see that there is some trend in the distribution of residuals which disobeys the assumption of homoscedasticity. Hence, it indicates the mild non-linearity. The standardized residuals versus leverage plot indicates the presence of a few outliers.

* 1. Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant ?



From the above p-values, we could see that the interaction between displacement and weight is statistically significant, while the interaction between cylinders and displacement is not.

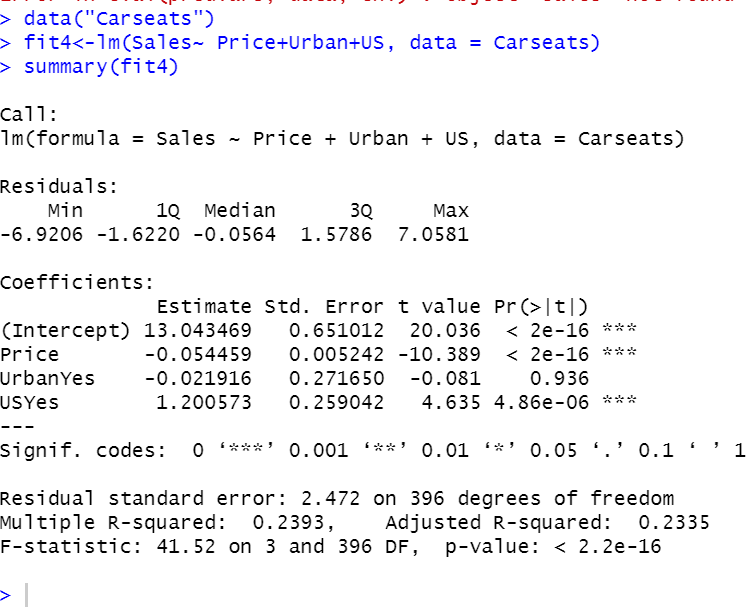
* 1. Try a few different transformations of the variables, such as log(X), √ X, X2. Comment on your findings.



1. **ISLR 3.7 Applied Problem 10.**

This question should be answered using the “Carseats” data set.

* 1. Fit a multiple regression model to predict Sales using Price, Urban, and US.



* 1. Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

The coefficient of the “Price” variable may be translated by saying that the normal impact of a cost increment of 1 dollar may be a diminish of 54.4588492 units in deals all other indicators remaining settled. The coefficient of the “Urban” variable may be translated by saying that on normal the unit deals in urban area are 21.9161508 units less than in rustic area all other indicators remaining settled. The coefficient of the “US” variable may be deciphered by saying that on normal the unit deals in a US store are 1200.5726978 units more than in a non US store all other indicators remaining settled.

* 1. Write out the model in equation form, being careful to handle the qualitative variables properly.

sales= 13.04+(-0.054) \* Price +(-0.021) \*Urban+(1.-20) \*US+E

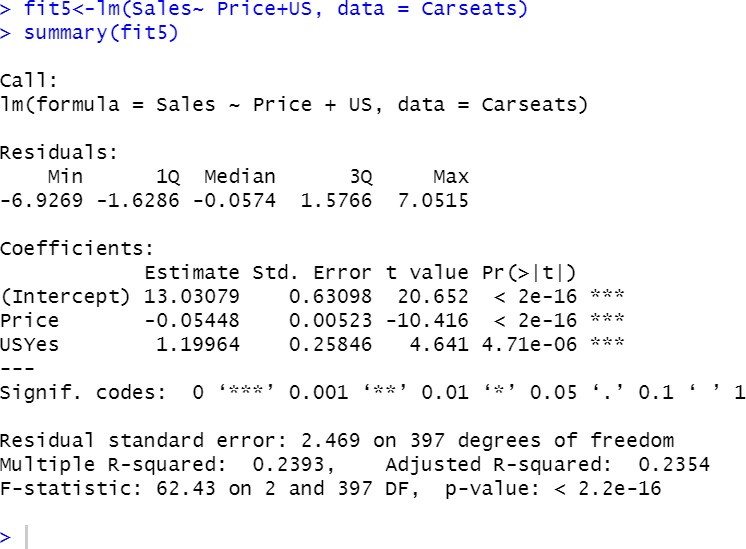
* 1. For which of the predictors can you reject the null hypothesis H0

: βj = 0?

## We can reject the null hypothesis for the ‘price’ and ‘US ‘

variables.

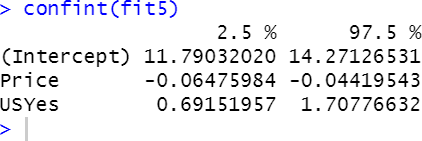
* 1. On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

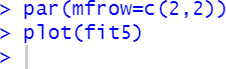


* 1. How well do the models in (a) and (e) fit the data?

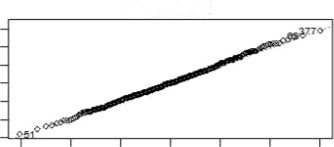
The R^2 for the littler show is imperceptibly way better than for the greater demonstrate. Basically around 23.9262888% of the changeability is clarified by the demonstrate.

* 1. Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).



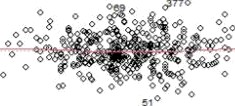
* 1. Is there evidence of outliers or high leverage observations in the model from (e)?

Residuals vs Fitted



Normal Q-Q

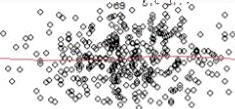




4   10 

Fitted value s

Scak' Location

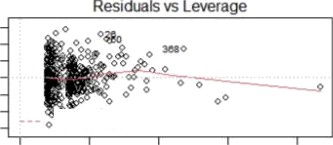


6 8 10 t2

Fitted values

3 -2 -1 0 1 2 3

Theorehc at Ooantiles



•• a

C6ok's d stance

m

000 001 002 003 0 OJ

L ever age