Complete the exercises and answer the questions outlined below. Your submission must be in an R Notebook or Markdown file. Your code must run successfully and you must answer the questions to get credit.

1. This exercise relates to the `College` data set, which can be found in the file `College.csv`. It contains a number of variables for 777 different universities and colleges in the US. The variables are

Private:	Public/private indicator
Apps:	Number of applications received
Accept:	Number of applicants accepted
Enroll:	Number of new students enrolled
Top10perc:	New students from top 10% of high school class
Top25perc:	New students from top 25% of high school class
F.Undergrad:	Number of full-time undergraduates
P.Undergrad:	Number of part-time undergraduates
Outstate:	Out-of-state tuition
Room.Board:	Room and board costs
Books:	Estimated book costs
Personal:	Estimated personal spending
PhD:	Percent of faculty with Ph.D.'s
Terminal:	Percent of faculty with terminal degree
S.F.Ratio:	Student/faculty ratio
perc.alumni:	Percent of alumni who donate
Expend:	Instructional expenditure per student
Grad.Rate:	Graduation rate

   (a) Read the data into R. Make sure that you have the directory set to the correct location for the data.

   (b) Look at the data using the `View()` 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.

   (c) i. Produce a numerical summary of the variables in the data set.

      ii. 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]`.

      iii. Produce side-by-side boxplots of `Outstate` versus `Private`.

      iv. 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%.

       v. Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`.

      vi. 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.

vii. Continue exploring the data, and provide a brief summary of what you discover.

[35 points]

2. This exercise uses the `Auto` data set. Make sure that the missing values have been removed from the data.

   (a) Which of the predictors are quantitative, and which are qualitative?

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

   (c) What is the mean and standard deviation of each quantitative predictor?

   (d) 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?

   (e) Using the full data set, investigate the predictors graphically, using scatterplots or other (R) tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

   (f) Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.

[30 points]

3. This exercise involves the `Boston` housing data set.

   (a) To begin, load in the `Boston` data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

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

   (c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

   (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.

   (e) How many of the suburbs in this data set bound the Charles river?

   (f) What is the median pupil-teacher ratio among the towns in this data set?

   (g) Which suburb of Boston has lowest median value of owner-occupied 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.

   (h) 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.

[35 points]