Assignment 1

Due: Monday, September 7, 2019

Note 1: Provide **outputs** and **interpret** the findings.

Note 2: Turn in your assignment as a **hard copy** in class.

This assignment relates to the ‘College’ data set, which can be found in the ‘ISLR’ library. 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

1. Read the data into R. Call the loaded data “college”.
2. Answer the following sub-questions
3. Use the “summary()” function to produce a numerical summary of the variables in the data set.
4. 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 then columns of a matrix A using A [,1:10].
5. Use the “plot()” function to produce side-by-side boxplots of “Outstate” versus “Private”
6. 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 “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”

1. Use the “hist()” function to produce some histogram 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.
2. Continue exploring the data, and provide a brief summary of what you discover.