CptS 483-04: Introduction to Data Science, Fall 2017

Assignment 2: R basics and Exploratory Data Analysis

Release Date: September 8, 2017 **Due Date**: September 15, 2017 (11:59 pm)

This assignment has **two exercises.** For questions that ask you to produce a specific plot, include that plot along with the code you used to generate it.

1. This exercise relates to the College data set, which can be found in the file College.csv on the course webpage. 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

Before reading the data into R, you can view it in Excel or a text editor. For each of the following questions, include the code you used to complete the task as your response, along with any associated output.

(a) 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.

(b) 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:

```
> rownames (college )=college [,1]
> fix(college)
```

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
> college =college [,-1]
> fix(college)
```

Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

(c)

i. Use the summary() function to produce a numerical summary of the variables in the data set. (Respond to this question with the mean graduation rate included in the summary result).

ii. 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].

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

iv. Create a new qualitative variable, called Top, 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 25% of their high school classes exceeds 50%.

```
> Top=rep("No",nrow(college ))
> Top[college$Top25perc >50]=" Yes"
> Top=as.factor(Top)
> college=data.frame(college, Top)
```

Use the summary() function to see how many top universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Top. Ensure that this figure has an appropriate title and axis labels.

v. 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. Again, ensure that this figure has an appropriate title and axis labels.

vi. Continue exploring the data, and provide a brief summary of what you discover. You may use additional plots or numerical descriptors as needed. Feel free to think outside the box on this one but if you want something to point you in the right direction, look at the summary statistics for various features, and think about what they tell you. Perhaps try plotting various features from the dataset against each other and see if any patterns emerge.

2. This exercise involves the Auto.csv data set found on the course website. Make sure that the missing values have been removed from the data. To do this, consider the na.strings parameter of read.csv(), as well as the na.omit() function.

(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. Hint: consider using R's sapply() function to take the range of multiple features in a single function call.

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

(d) Now remove the 25th through 75th 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 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.