CptS 475/575: Data Science, Fall 2021

Assignment 2: R Basics and Exploratory Data Analysis

Release Date: September 3, 2021 **Due Date**: September 10, 2021 (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. You are strongly encouraged to use R Markdown to prepare your solution. Be sure to clearly number each response in line with the questions, and give each plot appropriate axis labels and title.

1 (**50 points**). This exercise relates to the College data set, which can be found in the file College.csv uploaded on the course's public webpage (https://scads.eecs.wsu.edu/wp-content/uploads/2021/09/College.csv). The dataset 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 : Percentage of new students from top 10% of high school class
- Top25perc : Percentage of 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 or Python, 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 plots or numeric outputs produced. You may omit outputs that are not relevant (such as dataframe contents), but still include all of your code.

(a, **6 points**) Use the read.csv() function to read the data into R, or the csv library to read in the data with python. In R you will load the data into a dataframe. In python you may store it

as a list of lists or use the pandas dataframe to store your data. Call the loaded data college. Ensure that your column headers are not treated as a row of data.

(b, **8 points**) Find the median cost of room and board (Room.Board) for all schools in this dataset. Then find the median cost of room and board (Room.Board) for both public and private (Private) schools.

(c, **8 points**) Produce a scatterplot that shows a relationship between two numeric (not factor or boolean) features of your choice in the dataset. Ensure it has appropriate axis labels and a title.

(d, **8 points**) Produce a histogram showing the overall enrollment numbers (P.Undergrad plus F.Undergrad) for both public and private (Private) schools. You may choose to show both on a single plot (using side by side bars) or produce one plot for public schools and one for private schools. Ensure whatever figures you produce have appropriate axis labels and a title.

(e, **10 points**) Create a new qualitative variable, called Top, by binning the Top10perc variable into two categories (Yes and No). Specifically, divide the schools into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 75%.

Now produce side-by-side boxplots of the schools' acceptance rates (based on Accept and Apps) for each of the two Top categories. There should be two boxes on your figure, one for top schools and one for others. How many top universities are there?

(f, **10 points**) Continue exploring the data, producing two new plots of any type, and provide a brief (one to two sentence) summary of your hypotheses and what you discover. 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 (**50 points**). This exercise involves the forestfires.csv dataset that can be loaded into a dataframe from (https://scads.eecs.wsu.edu/wp-content/uploads/2021/09/forestfires.csv). The features of the dataset are:

- month: month of the year
- day: day of the week
- FFMC: Fine Fuel Moisture Code index
- DMC: Duff Moisture Code index
- DC: Drought code index
- ISI: Initial spread index
- temp: Temperature in degrees Celsius
- RH: Relative Humidity
- wind: Wind speed (km/h)
- rain: Amount of rainfall (mm/m2)
- area: area that got burnt in the forest fire

(a, **6 points**) Specify which of the predictors are quantitative (measuring numeric properties such as size, or quantity), and which are qualitative (measuring non-numeric properties such as color, appearance, type etc.), if any? Keep in mind that a qualitative variable may be represented as a quantitative type in the dataset, or the reverse. You may wish to adjust the types of your variables based on your findings.

(b, **8 points**) What is the range, mean and standard deviation of each quantitative predictor?

(c, **8 points**) Now remove the 20th through 70th (inclusive) observations from the dataset. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(d, **10 points**) Produce a bar plot to show the count of forest fires in each month. During which months are forest fires most common? (Hint: group data by month and calculate count)

(e, **10 points**) Using the full data set, investigate the predictors graphically, using scatterplots, correlation scores or other tools of your choice. Create a correlation matrix for the relevant variables.

(f, **8 points**) Suppose that we wish to predict the area burned by the forest fire (area) on the basis of the other variables. Which, if any, of the other variables might be useful in predicting area? Justify your answer based on the prior correlations.