Assignment: MA717

Data description

The data used in this assignment is taken from ISLR2 library and we modified the data for this assignment. The data contains information of US colleges from the 1995 issue of US News and World Report. The data includes 775 observations on the following 17 variables:

1. Private - A factor with levels Yes and No indicating private or public university.
2. Apps - Number of applications received.
3. Accept - Number of applications accepted.
4. Enroll - Number of new students enrolled.
5. F.UG - Number of full time undergraduates.
6. P.UG - Number of part time undergraduates.
7. Outstate - Out-of-state tuition (dollars).
8. Room.Board - Room and board costs (dollars).
9. Books - Estimated book costs (dollars).
10. Personal - Estimated personal spending (dollars).
11. PhD - Percent of faculty with Ph.D. students.
12. Terminal - Percent of faculty with terminal degree.
13. S.F.Ratio - Student/faculty ratio.
14. alumni.deno - Percent alumni who donate.
15. Expend - Instructional expenditure per student (dollars).
16. Elite - A factor with levels Yes and No indicating Elite university or not.
17. Grad.Rate - Graduation rate in percentage.

Important notes about the assignment:

1. The data is stored in a CSV file with name "College.csv". You should download the data and the Rmarkdown template file "MA717_assignment_template.rmd" into the same folder.

2. You should rename the template file as MA717_yourregistration.rmd. For example, if your registration number is 2000999, then you should change the template file as MA717_2000999.rmd. And you should submit both .rmd and output file (can be .pdf, .html, .doc) to FASER before the deadline. Any late submission will be marked as zero.

3. In your assignment, you will need to use R to work on the data and show the results. You also need to add answers or comments for some questions. Please make sure that you include all required information into rmarkdown file and knit it as an output file. If you cannot knit the file, it may because that you have errors in your codes. If you cannot produce .pdf, then produce .html and .doc file.

4. Please use 'percent' or 'percentage' instead of '%' in your answer to refer to percentages, as '%' will be interpreted as a comment in LaTeX, causing any text following it to be hidden.

5. Please note, in your assignment, you will need to generate a random subset of College data as your own data to work with and details are given in the tasks. Due to randomness, we would not expect seeing same results for any two students. Please be aware that collision or plagiarism is NOT permitted and will be reported as Academic Offence. You should not discuss your work with your classmates or copy the work from your classmates. You should also not to give your work to others to look at.

Task 1: Data reading and simple exploration (25%)

1.1. Read "College.csv" file into R with following command and use dim() and head() to check if you read the data correct. You should report the number of observations and the number of variables. (5%)

1.2. Use your registration number as random seed, generate a random subset of College data with sample size 700, name this new data as mynewdata. Use summary() to output the summarized information about mynewdata. Please report the number of private and public university and the number of Elite university and non-Elite university in this new data. (12%)

1.3. Use mynewdata, plot histogram plots of four variables "Outstate", "Room.Board", "Books" and "Personal". Give each plot a suitable title and label for x axis and y axis. (8%)

Task 2: Linear regression (45%)

2.1. Use mynewdata, do a linear regression fitting when outcome is "Grad.Rate" and predictors are "Private" and "Elite". Show the R output and report what you have learned from this output (you need to discuss significance, adjusted R-squared and p-value of F-statistics). (6%)

2.2. Use linear regression fitting result in 2.1, calculate the confidence intervals for the coefficients. Also give the prediction interval of "Grad.Rate" for a new data with Private="Yes" and Elite="No". (4%)

2.3. Use mynewdata, do a multiple linear regression fitting when outcome is "Grad.Rate", all other variables as predictors. Show the R output and report what you have learned from this output (you need to discuss significance, adjusted R-squared and p-value of F-statistics). Is linear regression model in 2.3 better than linear regression in 2.1? Use ANOVA to justify your conclusion. (14%)

2.4. Use the diagnostic plots to look at the fitting of multiple linear regression in 2.3. Please comment what you have seen from those plots. (7%)

2.5. Use mynewdata, do a variable selection to choose the best model. You should

use plots to justify how do you choose your best model. Use the selected predictors of your best model with outcome "Grad.Rate", do a linear regression fitting and plot the diagnostic plots for this fitting. You can use either exhaustive, or forward, or backward selection method. (14%)

Task 3: Open question (30%)

Use mynewdata, discuss and perform any step(s) that you think that can improve the fitting in Task 2. You need to illustrate your work by using the R codes, output and discussion.