STA 4164-0001

Project Assignment #1

General Guidelines:

1. You need to finish this assignment individually. You may use notes from our class and online resources, such as your preferred programming language documentation and help files for this study.
2. Write your answer to each question separately. Your answer may include the code you have written for that specific part, and your conclusion or interpretation of the results.
3. You will be evaluated according to what you did and not only based on the final result/output. Therefore, it is strongly suggested that you include what you have done, even if you did not get to what is expected. Make sure to explain what you did so that you can at least get the partial credit.
4. The goal here is not just to evaluate your coding or writing skill. Any assignments in this class are part of your learning process. It might be challenging in a few tasks, but it is still doable if you try a little more. You are always welcome to go beyond the materials we discussed in class. Here, creativity matters!
5. You may always reach out during office hours to discuss your issues.
6. You need to submit your final report as a word or pdf file. Please make sure to merge all you want to present and make **only one file**. You need to submit your R markdown file as well. Please write both team members names in the report if you work in a group.

PART 1: EXPLORATORY ANALYSIS

1. Load dataset "Grad_Admission.csv" into R. Variables description are as follows:
    - ID: Unique identification code for each student (DO NOT INCLUDE)
    - GRE: GRE Scores (out of 340)
    - TOEFL: TOEFL Scores (out of 120)
    - Urate: University Rating (out of 5)
    - SOP: Statement of Purpose Strength (out of 5)
    - LOR: Letter of Recommendation Strength (out of 5)
    - CGPA: Undergraduate GPA (out of 4)
    - Chance: Chance of Admission (ranging from 0 to 100)

2. Print the dataset. (Limit your print to include only the first ten observations)

3. Print a table of the min/1st quartile/median/mean/3rd quartile/max of three of the variables (your choice).

4. Make histograms for "Chance", "CGPA", "GRE" and "TOEFL" variables and comment on their distribution. Are these variables normally distributed? What else can we find from histograms?

PART 2: REGRESSION ANALYSIS

In this part, we want to explore the "Grad_Admission" dataset by applying what we learned about regression analysis. The explanation is more critical, so make sure to explain your result for any of the following tasks:

5. Consider "Chance" as the response variable and print scatter plots of each of the other six variables against it.

6. Draw your initial conclusions about the relationship between independent variables and the response variable based on the scatterplot. Is there any relationship? Is it linear enough?

7. Confirm the validity of five major linear regression assumptions and comment on them.

8. Choose the best three independent variables based on your immediate insight into the relationship and list them. Write down the model.

9. Build up a table, including the correlation between all independent variables and the response variable. What can you deduce from these correlation coefficients? Briefly explain the relation between correlation coefficients and the slope of the regression equation.

10. Report the result of hypothesis testing for the correlation coefficient of each independent variable (i.e. $\rho_i = 0$).

11. Build straight-line (univariate) regression models for all six independent variables. Test the null hypothesis that each of the β_i's is equal to zero. (i=1, 2, ...,6) (set $\alpha = 0.05$)

12. Report the ANOVA table for two variables with the highest R-square value? What conclusion is achievable looking at these tables? (hint: compare R-squares between variables and make suitable conclusion)

13. Build up a model, including all six variables available in the dataset. Are all coefficients significant at $\alpha = 0.05$?

14. Remove all non-significant variables from the model and rebuild the model. What has been changed considerably in the ANOVA table compared to the model in task 13?

15. Build up confidence bands and prediction bands for all records. Print the appropriate table into your output.

16. Write the appropriate equation to predict the admission chance with variables included in your final model from task 11. Explain the meaning of intercept and slope in this equation.

17. What conclusion can you arrive at from this exploration in terms of the suitability of descriptive statistics and regression in data exploration? What is the recommendation that you would provide future data explorations to include as a result necessarily?