STA 207 HW-6
Due Date: 11/8 by 10:20 AM

Problem-1 [25 points]: We will work with the data on median housing prices in neighborhoods in the suburbs of Boston (same as hw-5). Regress medv (Y) on lstat (X) and report the fitted model. For this model

- [5 points] Identify high leverage points, if any.
- [5 points] Identify outliers, if any.
- [15 points] Identify influential points, if any. Use Cook's distance and influence plots.

Problem-2 [75 points]: We will need R package ISLR for this problem so install it first. In R package ISLR, we will use dataset called Credit. This is a simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt. The outcome variable of interest is the credit card debt of 400 individuals. Other variables like income, credit limit, credit rating, and age are included as well. Note that the Credit data is not based on real individuals' financial information, but rather is a simulated dataset used for educational purposes. Using outcome variable average credit card debt (called Balance) and three predictor variables cardholder's credit limit (Limit), Income and Credit rating (Rating), do the following:

a.) [7.5 points] Make a scatterplot matrix and explain relationship of outcome variable with all three predictor variables.
b.) [7.5 points] Make a corrplot and see if your explanation in (a) match with the strength and direction of the correlation coefficient for each relationship.
c.) [10 points] Regression outcome variable on the three predictors and report the fitted model. Interpret the regression coefficients.
d.) [10 points] Report goodness of fit for the above model.
e.) [25 points] Report if the LINE conditions are met or not.
f.) [15 points] Check if there is multi-collinearity in the three predictor variables.

Data Analysis Project Problem: Submit one solution for each team under Data Analysis Project

Form a team of two and find a dataset to perform regression analysis. Find the dataset with multiple variables (Y needs to be numeric and some X's numeric and some categorical). Explain the real-world (meaning why one should care about this topic!) motivation for doing analysis in this area. What specific research questions would you like to investigate in this area? Enter this data in R to get it ready for the analysis. NO analysis needed yet.