Homework 2

The purpose of this homework is to have you walk through a complete, applied, example of the process/plot/analyze/plot procedure.

1. Load the `CollegeScorecard.csv` dataset.
2. Process the data with the following steps:
    a) Trim the dataset by selecting only the following variables: INSTNM, STABBR, SAT_AVG_ALL, GRAD_DEBT_MDN_SUPP, md_earn_wne_p10
    b) Ensure that `NULL` and `PrivacySuppressed` values are stored as missing data (e.g. `NA`).
    c) Rename the variables to `Institution`, `State`, `SAT`, `Debt` and `Earnings`.
    d) Create two new vectors `SAT_c` and `Debt_c` by centering each vector. That is, substract the mean of each vector from each individual observation within the vector. Calculate the mean of each of the centered vectors to ensure it was created properly (mean should round to 0).
3. Create the following plots
    a) Scatterplot matrix of `SAT`, `Debt`, and `Earnings`
    b) Histograms of `SAT`, `Debt`, and `Earnings`
    c) Density plots of `SAT`, `Debt`, and `Earnings`
4. Fit the following preliminary models, and inspect the results from each
    a) SAT predicting Earnings
    b) Debt predicting Earnings
    c) Debt predicting SAT
5. How would you interpret the results? Plot the relation between the variables and overlay the regression line for each model. Does anything appear odd? Refit the models with the centered variables. How have the results changed?
6. Fit and inspect the results from a multiple regression model with `SAT_c` and `Debt_c` predicting earnings. Provide a brief description of the results.

Extra Credit. Produce a predictor residual plot that shows the relation between `SAT_c` and `Earnings` after accounting for `Debt_c`. Note that I don't expect you to know how to do this, that's why it's extra credit, but we will cover it in class.