ASSIGNMENT 9
ADVANCED METHODS AND ANALYSIS FOR THE LEARNING AND SOCIAL SCIENCES
PROFESSOR RYAN S.J.d. BAKER
IMPUTATION
DUE NOON, MONDAY APRIL 9

The goal of this assignment is to conduct data imputation to fill in missing data, so that a data set can be appropriately analyzed, using the data in Asgn9-imputation-data-v2.xlsx

This data set has eight variables. Column A, student ID, is an anonymized indicator of which student it is. StandardizedExam represents a percentile ranking on a state standardized exam in Mathematics. The remaining variables represent various quantities which could potentially predict StandardizedExam. However, due to challenges in data collection, a lot of values are missing from this data set. Your goal is to find the best linear regression model which predicts StandardizedExam, using either a statistics package, or a data mining package. However, in order to do that, you must find some way to work around the missing data (Hint: simply dropping any student who is missing at least one data point will NOT get you a good grade on this assignment).

Please turn in:

- The linear regression model obtained predicting StandardizedExam, and an appropriate indicator of the model's goodness
- The data set you used to obtain this linear regression model
- A full description of the model(s) or procedure(s) which you used to fill in missing data points

You will be graded on completeness and comprehensibility of your hand-in, whether you correctly and validly apply the method you choose to this data, and whether the methods you chose fit the requirements of this assignment.

BONUS: I have created a test set generated via the same functions as the data set you are using for this assignment. This test set has no missing data. The student whose linear regression model best predicts this test set (according to linear correlation) will receive the bonus.