EDF 6937 Data Analysis – R Puzzles

For each of the following problems, write a series of R expressions that will perform the indicated operation. You may need separate code for a matrix and a data.frame. Write some code to test your expression. Remember you don’t need to solve each puzzle, but you do need to make an attempt to get us started.

1. Calculate the mean and standard deviation of each column in a data frame.
2. Generate a data set where Y = 2.5\*X+e, mean(X)=10 (approximately) and sd(X)=2, sd(Y) = 5 (approx) and cor(X,Y) = .6.
3. Replace the values 7,8 or 9 with a missing value anywhere they occur in a data matrix and data.frame.
   1. Replace the values 7 with 0 and the values 8 and 9 with a missing value.
4. Scale each column of a matrix (data.frame) to have mean 0 and sd 1.
5. Scale each column of a matrix (data.frame) to have values between 0 and 1.
6. Assume that columns 1-10 are answers to survey questions about topic 1 and 11-16 are answers to survey questions about topic 2. Create two new columns that give the total on each topic.
7. Same problem, but now assume that all questions are on a five point Likert scale (so values range from 1 to 5). Add two columns that give the average scores on each topic.
   1. Don’t count missing values in the numerator or the denominator.
   2. Same problem, but count missing values as 0.
8. Randomly select 10% of the cases in the data set for a test case for cross validtation. Separate the data set into a test set (the 10%) and a training set (the remaining 90%)
9. Create 5 new data sets with the same number of rows as the original data set by sampling (with replacement) from the original data. (This is known as the bootstrap and is a useful trick for generating standard errors.)
10. Create a jackknife estimate of the standard error of the mean by (a) calculating the means of each column leaving out row 1, row 2, row 3, &c in turn and then (b) taking the standard deviation.