EDF 6937 Missing Data – 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. Count the number of missing values in each column of a data matrix and data.frame.
2. Replace the values 7,8 or 9 with a missing value anywhere they occur in a data matrix and data.frame.
    a. Replace the values 7 with 0 and the values 8 and 9 with a missing value.
3. Check to see if the missing data pattern is monotone.
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. Randomly sample 25% of one variable and make the values missing.
7. Assume that column 3 in the data matrix is a probability running between 0 and 1. Make the values in column 4 missing with the probability given in column 3. That is if X[1,3]=.25, then there should be a 25% chance that X[1,4] is missing and if X[2,3]=.5 then there should be a 50% chance that X[2,4] is missing.
8. 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.
9. 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.
    a. Don't count missing values in the numerator or the denominator.
    b. But count missing values as 0.
10. 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%)
11. 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.)
12. 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.