Problem set 5

In this problem set, you will be using the famous Iris dataset, which includes measurements of 3 different species of Iris. Unfortunately, the download file was corrupted, viewed and saved in the wrong format, lost some data when transferred from MacOS to Windows, and I accidentally spilled coffee on it. Fortunately, I believe the bulk of the data is still there. Your task will be to clean and transform the data (i.e., feature engineer) into a format that can be analyzed.

Complete each of the numbers using a separate block and show/print the specific things that are asked (e.g., the shape)

1. Load relevant libraries
2. Load the data and make sure there are no extra index columns. Show the shape of the data AND the head of the data
3. Remove columns and rows where there are more than 50% of the data missing. Show the shape of the data after you have removed those columns/rows
4. Remove duplicate data, if there is any. Show the shape of the data
5. Dummy code the categorical data. Show the head of the data
6. Drop the redundant columns from the dataframe so that you are left with just the dummy coded columns. Show the head of the data
7. Remove outliers above 2 standard deviations.
8. For the remaining missing data, replace with the median value. Or Bonus 0.5 points – Instead of replacing the missing data with the overall median value, replace them with the median value with respect to species. In other words, in the first row, if petal length is missing, then replace that petal length value with the median petal length value for the species setosa.
9. Bonus 0.5 points – Bin or bucket the petal width column into low, medium, and high depending on the following bins = 0 to 0.9, 0.9 to 1.8, 1.8 to 2.7. These should not be dummy coded, but instead, you should have one column with pedal\_width as 1 for 0 to 0.9, 2 for 0.9 to 1.8, and 3 for 1.8 to 2.7.
10. Show (i.e., print) that there are no missing data and also use the describe method to show off the final dataset.