Problem set 9

Imagine working at Yellowstone National Park as a scientist researching Old Faithful Geyser. Your boss wants you to evaluate the predictiveness of the relationship between duration of an eruption in minutes (“duration”) and the duration in minutes until the next eruption (“waiting”). You collected the data (data = sns.load\_dataset('geyser')), but the internet in Yellowstone is not very good, so you cannot access the crossvalidation library in python. You will write your own k-fold cross validation code to look at the average r^2 value between duration and waiting.

Block 1. Import only “seaborn”, “pandas”, “numpy”, “matplotlib.pyplot”, and “from sklearn.linear\_model import LinearRegression”. Import the data here as well using seaborn (data = sns.load\_dataset('geyser'))

Block 2. Create your own k-fold cross validation code with the number of folds being 10. You can implement this however you think best. To help you out though, think about the following implementation:

1. The raw data needs to be split into training and testing.
2. The number of folds is the number of times you should go through a for loop.
3. Inside of the for loop, index into your raw data to pull out training and testing data based on the fold (think about how the folds are organized and how much data is in each fold, which is based on the total number of folds)
4. Now, you can do the normal linear regression model fitting that we did a couple weeks ago
5. Keep track of the R^2 scores of the folds

Block 3. Print the average and standard deviation of the R^2 value for the different folds

Block 4. In a text block, interpret this result

EXTRA CREDIT (1 point). In block 2 above, the number of folds is hardcoded to be 10. Instead, for extra credit, create a function in block 2 that let’s you pass in (1) the X feature matrix, (2) the y predicted variable, and (3) the number of folds. Then, in Block 3, print the averages and standard deviations for k-folds with k being 3, 5, 10, and 20