CMSC 352 Machine Learning
Assignment #3: Regression and Data Processing **Due: 2.Oct.20**

The data in DataProblem.csv consists of a header of seven columns followed by 10,000 data points, one per row. Using SciKitLearn, load the data from the local directory and build a regression model to predict the final column from the other columns. You should use the first 8000 items for training and the final 2000 for testing.

Write your solution in a Jupyter Notebook named HW3.ipynb. You should put your name and the homework identifier in markdown in the first cell. As your process the data and develop your solution, you should provide explanation prior to the command you execute, explaining why you are taking that specific step. Please use markdown to write your explanations. Describe any relationships you observe in the data based on scatterplots.

Look into R^2 scoring of the linear regression model and use it as a summary statistic. Explain how you interpret the R^2 values. You should also provide a plot of your predicted testing values versus the actual testing values in a 2-D plot. Your solution will be judged on your approach to solving the problem as well as how well you correctly process your data, including the ideas of data preparation we discussed, and correct treatment of training and testing data.

Can you eliminate any of the variables and still obtain a good solution? Rank the variables in order of importance as predictors in your final solution.

You will submit HW3.ipynb. It should load the data CSV file and perform all processing with no errors when I restart the kernel and run all cells.