CST 383 - Intro to Data Science

Dr. Glenn Bruns

# Lab: Linear regression, part 3

In this lab we willcontinue to perform linear regression using the CPU data set.

1. Read the CPU data:

df = pd.read\_csv("https://raw.githubusercontent.com/grbruns/cst383/master/machine.csv")

df.index = df['vendor']+' '+df['model']

df.drop(['vendor', 'model'], axis=1, inplace=True)

df['cs'] = np.round(1e3/df['myct'], 2) # clock speed in MHz

1. From the machine data, create training and test sets, using a 70/30 split. See the linear regression 2 slides if needed.
2. This lab is open-ended. I want you to experiment with building linear models using **interaction**s and **nonlinear transformations**, as discussed in class. The goal is still to predict performance (feature ‘prp’). Start simple.
3. Predict prp using your test set, and write your own code to compute the RMSE.
4. For each experiment, I want you to pick some features, transform them in some way (or not at all), and then build a linear model using the training data. Write down the RMSE you get on the test data. Note: if you transform ‘prp’, you will need to account for that when you compute RMSE!
5. See which features and transformations give you the best RMSE scores.