**EE5373: Data Modeling Using R**

**Fall, 2017**

Department of Electrical and Computer Engineering

**University of Minnesota**

Lab 2: Data Cleaning

Due date: See the due date shown on the class moodle page.

Goal: This lab introduces the concepts of data cleaning and sanity checking using the CPU DB database.

What to do:

Download the CPU DB database from z.umn.edu/lrur. Read this data into R using the read-data.R program available at the same web site. Experiment with built-in R functions to analyze the various columns of data looking for unusual patterns, outliers, and anything else that might look amiss. For example, try the following for every column in the data frame:

* Compute the mean, variance, minimum, and maximum.
* Sort the column to look for outliers or unusual patterns.
* Compute the fraction of NA values to see if there are enough values.
* Use the table() function to determine the distribution of values.

How else could you verify that the data looks reasonable?

What to turn in for grading:

Write a 1-2 page lab report describing all of the tests that you did for each of the dataframes. Be sure to include a representative set of your specific results. Based on your analysis, determine which columns of the database are “unusual” and describe how they are unusual. Also, describe any anomalies that you might see in the data. Explain how you would fix the unusual or anomalous data. You must include one interesting type of analysis beyond the list above. Do not forget to upload the pdf file with your report to moodle by the due date.