**Homework One - Due February 3rd,, 2018**

**Description:** This assignment will involve the concepts covered in chapters 2, 3, 4. You will work with two assigned data set. The data sets are derived from Compustat for selected companies. The companies were identified by the Securities Exchange Commission (SEC) for committing fraud of one form or another. The first data set is simply data downloaded for a select set of firms. The data variables are listed. The second data set is composed of firms who were cited for fraud. Additionally, some of these firms went bankrupt after being cited. This information is indicated in a variable.

You will perform several types of analysis in two parts. Part one will use the part one data set. Part two will use the part two data set.

**Part one**

Using AAERDATAHWPart1 file:

1. Identify missing values and take appropriate action – eliminate, average value, etc. Justify your decision.
2. Look for redundancies and errors.
3. Identify mean, count, sum, median, standard deviation for – SALES, Price\_close, Employees
4. Identify outliers for – SALES, Price\_close, Employees
5. Calculate
   1. Skewness for Sales
   2. Skewness for Z-Score standardized SALES
   3. Is there evidence of SKEWNESS
6. Normal probability plots
   1. Construct a normal probability plot for Employees
   2. Derive a new variable – CASH/Total Assets
   3. Construct a normal probability plot for new Variable.

**Part Two**

Using AAERDATAHWPART2 file:

1. Ensure no missing values or errors
2. Standardize the data where relevant – ie you don’t standardize GVKEY or SIC codes
3. Which variables are categorical and which are numeric
4. Examine numeric variables for correlation
5. For the categorical variables construct a bar chart
6. For each pair of categorical variables construct a cross tabulation
7. Construct a histogram of three numeric variables with an overlay of bankruptcy variable (bktype)
8. For three numeric variables construct a scatter plot for each pair. (Should be three)
9. Conduct a correlation analysis of the “Data” variables – exclude the calculated variables
10. Generate Principal components based on the numeric data – both DATA and calculated
11. How many Principal components should you use?
12. Plot the factor scores for the number of principal components you identified in #11
13. Apply PCA with varimax rotation for the number of principal components in #10
14. Examine the difference in PCA results if you use only the Data variables versus using the constructed (calculated variables).