Stat 666
Mini-Project #1
Due Date: TBA

On the course webpage, you will find the file `fraud.zip` which contains the file `fraud.csv`. The fraud dataset contains data for 250,000 financial transactions over a two-day period. Because the original data are proprietary and confidential, we have 12 aggregate variables instead of the original, easily interpretable financial features (variables). This data set comprises both normal and fraudulent transactions, with the fraudulent transactions believed to account for less than 0.2% (i.e., <500) of the records. In this project, you will use the descriptive statistics (both graphical and numerical) discussed in class as well as in the text to assess the assumption of *multivariate normality* and the *presence of outliers* in the dataset. You may treat the observations (rows) as a random sample of records from the time and area of interest.

Two main issues need to be addressed but neither your analyses nor your report need to address them in the order presented below. First, identification of outliers (i.e., potentially fraudulent records). In your analysis, you should identify the 250 most potentially fraudulent transactions. Instead of listing the 250 suspected cases of fraud in your short report, you will email me a vector of length 250,000 with 0's in every location EXCEPT the 250 locations that are suspected to be fraudulent. Denote your suspected cases of fraud with 1's. Part of your grade will be based on the number among your 250 flagged records that have actually been verified to be fraudulent. This is a bit tricky but should make the project more fun....I'll leave it to you to come up with a strategy for identifying potential outliers within a mixture. With no training data, this is unquestionably difficult. You are welcome to discuss ideas with me. In your report, make sure to carefully justify your chosen methodology for identifying abnormal values.

Second, are the data distributed normally? (If not, are there transformations for making the data look more normal?) Transform the data and then evaluate whether or not the multivariate normality assumption is plausible for the transformed data. Note: univariate and bivariate evaluations of normality are useful, but make sure that you also include some diagnostic for multivariate normality.

You should turn in a write-up of no longer than 4 pages (11 pt font, 1 inch margins, single spaced), including all tables and graphs. Use the standard formatting conventions for the M.S. project. You may assume that your client has an MBA and a basic understanding of statistical methods (anova, regression, hypothesis testing, confidence intervals, etc.), but the reader is NOT a professional statistician. However, your client employs an M.S. level statistician that will also be reading the report to check for accuracy. Hence, a careful explanation of the problem, techniques, and results are crucial. Keep in mind that I (Prof. C.) am still another reader of the report—you need to *convince* me that you understand the principles discussed in class and in the text. You are not allowed to discuss the project with other members of the class. You may, however, discuss the project with me.

You should put figures and tables directly into the body of the report as a LaTeX "float"—not in an appendix. All figures and tables should have captions and should be referred to in the text. Match your tone and style to a professional or scholarly report (see *The American Statistician* or Prof. C. for examples).