**Computational Mathematics**

Your final is due by the end of day on 12/20/2017. You should post your solutions to your GitHub account. You are also expected to make a short presentation during our last meeting (3-5 minutes) or post a recording to the board. This project will show off your ability to understand the elements of the class.

You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition. <https://www.kaggle.com/c/house-prices-advanced-regression-techniques> . I want you to do the following.

Pick **one** of the quantitative independent variables from the training data set (train.csv) , and define that variable as X. Pick **SalePrice** as the dependent variable, and define it as Y for the next analysis.

*Probability.* Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the 1st quartile of the X variable, and the small letter "y" is estimated as the 2d quartile of the Y variable. Interpret the meaning of all probabilities.

1. P(X>x | Y>y) b. P(X>x, Y>y) c. P(X<x | Y>y)

Does splitting the training data in this fashion make them independent? In other words, does P(X|Y)=P(X)P(Y))? Check mathematically, and then evaluate by running a Chi Square test for association. You might have to research this.

*Descriptive and Inferential Statistics.* Provide univariate descriptive statistics and appropriate plots for both variables. Provide a scatterplot of X and Y. Transform both variables simultaneously using Box-Cox transformations. You might have to research this.

*Linear Algebra and Correlation.*  Using at least three untransformed variables, build a correlation matrix. Invert your correlation matrix. (This is known as the precision matrix and contains variance inflation factors on the diagonal.) Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix.

*Calculus-Based Probability & Statistics*. Many times, it makes sense to fit a closed form distribution to data. For your *non-transformed* independent variable, location shift (if necessary) it so that the minimum value is above zero. Then load the MASS package and run fitdistr to fit a density function of your choice. (See <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html> ). Find the optimal value of the parameters for this distribution, and then take 1000 samples from this distribution (e.g., rexp(1000, ) for an exponential). Plot a histogram and compare it with a histogram of your non-transformed original variable.

*Modeling*. Build some type of regression model and submit your model to the competition board. Provide your complete model summary and results with analysis. **Report your Kaggle.com user name and score.**