**Computational Mathematics.**

Your final is due by the end of day on 12/23/2016. 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 quanititative independent variables from the training data set (train.csv) , and define that variable as X.  *Make sure this variable is skewed to the right!* Pick the dependent variable and define it as Y.

*Probability.* Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the 3d 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. In addition, make a table of counts as shown below.

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

|  |  |  |  |
| --- | --- | --- | --- |
| x/y | <=2d quartile | >2d quartile | Total |
| <=3d quartile |  |  |  |
| >3d quartile |  |  |  |
| Total |  |  |  |

Does splitting the training data in this fashion make them independent? Let A be the new variable counting those observations above the 3d quartile for X, and let B be the new variable counting those observations above the 2d quartile for Y. Does P(A|B)=P(A)P(B)? Check mathematically, and then evaluate by running a Chi Square test for association.

*Descriptive and Inferential Statistics.* Provide univariate descriptive statistics and appropriate plots for the training data set. Provide a scatterplot of X and Y. Provide a 95% CI for the difference in the mean of the variables. Derive a correlation matrix for two of the quantitative variables you selected. Test the hypothesis that the correlation between these variables is 0 and provide a 99% confidence interval. Discuss the meaning of your analysis.

*Linear Algebra and Correlation.* 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. Conduct principle components analysis (research this!) and interpret. Discuss.

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

Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).

Also generate a 95% confidence interval from the empirical data, assuming normality. Finally, provide the empirical 5th percentile and 95th percentile of the data. Discuss.

*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.