Assignment 3: Data Analytics (Spring 2016) (15% written)
Due: FRIDAY March 4, 2016 (by 5pm ET)

Submission method: written document by email to pfox@cs.rpi.edu **and** Rahul Divekar divekr@rpi.edu

Please use the following file naming for electronic submission:
DataAnalytics2016_A3_YOURFIRSTNAME_YOURLASTNAME.xxx

Late submission policy: first time with valid reason – no penalty, otherwise 20% of score deducted each late day

Note: Your report for this assignment should be the result of your own **individual** work. Take care to avoid plagiarism ("copying"), and include references to all web resources, texts, and class presentations. You may discuss the problems with other students, but do not take written notes during these discussions, and do not share your written solutions.

General assignment: Distribution analysis and comparison of distributions, visual analysis, statistical model fitting and testing of the nyt2, … nyt31 datasets. The weighting score for each question is included below. Please use the question numbering below for your written responses for this assignment.

Please include code (fragments and/or scripts) and the plots you generate for the questions below.

1. For any **5** of the nyt datasets except nyt1, perform the following:
    a. Create boxplots for all 5 datasets for each of two key variables (you choose these; i.e. two sets of plots with 5 boxplots per plot). Describe/summarize the distributions. min. 3-4 sentences (3%)
    b. Create histograms for all 5 datasets each of for two key variables (you choose the histogram bin width). Describe the distributions in terms of known parametric distributions and similarities/ differences among them. min. 3-4 sentences (3%)
    c. Plot the ECDFs for your two key variables. Plot the quantile-quantile distribution using a suitable parametric distribution you chose in 1b. Describe features of these plots. min. 3-4 sentences (ugrad 5%, grad 3%)
    d. Perform a significance test that is suitable for the variables you are investigating. Discuss the test results and indicate whether the null hypothesis is valid. min. 3-4 sentences (ugrad 4%, grad 3%)
    e. Discuss any observations you had about the datasets/ variables, other data in the dataset (0% ;-))
2. Graduate 6600-level question (3%). Filter the distributions you explored in Q1 using one or more of the other variables for only **2** (not 5) of the nyt datasets. Repeat Q1b, Q1c and Q1d and draw any conclusions from this study. min. 3-4 sentences