Assignment 3: Data Analytics (Fall 2023) (15% written)
Due: October 17th , 2023 (by 11:59pm EST) Submission method: written document posted on LMS Assignment.

Please use the following file naming for electronic submission:
DataAnalytics_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 nyt3, … nyt31 datasets (nyt datasets are available class RPI Box repository) . 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. Choose any **7** of the nyt datasets except nyt1 and nyt2 (do not used the nyt1 and nyt2), perform the following:

a). Create boxplots for all 7 datasets for each of two key variables (you choose the variables), i.e. two figures (one for each variable) with 7 boxplots (for the 7 different datasets) in each. Describe and run summary statistics on the two chosen variables and explain them in your words. min. 3-4 sentences (3%)

b). Conduct the applicable normality test (i.e Shapiro Wilk, Anderson Darling, Kolmogorov-Smirnov) for the chosen two variables for all 7 datasets for two key variables – can be the same variables in 1a or different.
Create histograms for those two variables in the 7 datasets (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 (Empirical Cumulative Distribution Function 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 (4000-level 5%, 6000-level 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 (4000-level 4%, 6000-level 3%)

e). Discuss any observations you had about the datasets/ variables, other data in the dataset (0% ;-))

2. 6600-level question (3%). Filter the distributions you explored in Q1 using one or more of the other variables for only **4** (not 7) of the nyt datasets. Repeat Q1b, Q1c and Q1d and draw any conclusions from this study. min. 3-4 sentences