Assignment 3: Data Analytics (Fall 2024) (15% written)

Due: October 22nd, 2024 (by 08:00pm EST) Submission method: written document submitted to (eleisa2@rpi.edu)

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 New York Times Covid-19 dataset and the NY house dataset.

https://github.com/nytimes/covid-19-data

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 2 of the covid-19 us-counties datasets, perform the following:

a). Create boxplots for the "Cases" and "Deaths" variables comparing the variables between the 2 datasets, i.e. two figures (one for each variable) with 2 boxplots (for the 2 different datasets) in each. Describe and run summary statistics on the two chosen variables and explain them in your words. min. 2-3 sentences (2%)

b). Create histograms for those two variables in the 2 datasets (you choose the

histogram bin width). Describe the distributions in terms of known parametric distributions and similarities/ differences among them. Plot the distribution you think matches the histogram (e.g. normal, chis-square, gamma, t-distribution, etc.) overlayed on the histogram. min. 2-3 sentences (3%)

c). Plot the ECDFs (Empirical Cumulative Distribution Function) for the two variables in both datasets. Plot the quantile-quantile distribution using a suitable parametric distribution you chose in 1b. Describe features of these plots. min. 2-3 sentences

(4000-level 5%, 6000-level 2%).

2. 6600-level question (3%). Filter the distributions you explored in Q1 by a number of states or counties. Repeat Q1b, Q1c and Q1d and draw any conclusions from this study. min. 3-4 sentences

3. Using the NY house dataset:

https://rpi.box.com/s/h3tfkjov93kga1b384mvjgz32915loj7

a) Fit a linear model using the formula *PRICE ~ BEDS + BATH + PROPERTYSQFT* and identify the variable most significantly influencing house price. Produce a scatterplot of that variable with another and overlay the best fit line. Plot the residuals of the linear model. min. 2-3 sentences (2%)

b) Derive a subset of the dataset according to any criteria (e.g. PRICE > VALUE or BEDS < NUMBER) and repeat the linear model with its plots. Explain how the significance of the input variables changes and your interpretation of the change. min. 2-3 sentences (3%)