CIT 660 - Statistical Analysis and Visualization
Spring 2022
Assignment 02
Deadline: Monday - May 30th, 2022, 11:59 pm.

Part 1:
Assume you have a gene expression dataset as in the attached file titled “Assignment_02_Data.csv”. The values in the file are separated by commas “,”. Write an R script called “GE_Data_Modification.R” and an R function called “GE_Data_Normalization.R” as described below:

1. “GE_Data_Modification.R” reads the data from the “Assignment_01_Data.csv” file and stores them in a data frame.
2. Then, “GE_Data_Modification.R” imputes the missing data (NA) in the file with the median value of their corresponding columns; across samples. You can use any package that can impute data such as the “randomForest” package.
3. Send the imputed data to the “GE_Data_Normalization.R” function. The input to “GE_Data_Normalization.R” is a data frame.
4. “GE_Data_Normalization.R” replaces each value in the data frame by a new value computed as:

$$x_{new} = \frac{x_{old} - x_{avg}}{x_{sd}},$$

   where x_{old} is the original gene expression (GE) value (or the imputed value in case of missing data), x_{avg} is the average value across the samples of the column containing x_{old}, x_{sd} is the standard deviation of the same column, and x_{new} is the new GE value to be kept in the data frame.
5. The output of “GE_Data_Normalization.R” is a data frame containing the normalized GE values.
6. “GE_Data_Modification.R” receives the output data frame from “GE_Data_Normalization.R” and then stores it in a tab “\” delimited text file “.txt” that to be saved on the C partition of your hard disk.

Notes:

- The format of the input file is “.csv”, where values are separated by “,”. The format of the output file is “.txt”, where values are separated by “\t”.

Part 2:
Using the expression data in **Part 1**:

- Compute the mean and standard deviation of GE across samples for each gene. This step should lead to 2 vectors (one is for the mean and one is for the standard deviation) of length 5 each.
- Plot three sub-figures, one figure occupies the left half of the figure, and the remaining two sub-figures are positioned on the right half one above the other.
  1. On the left sub-figure:
     - Plot the mean and standard deviation values for each gene as connected points. The mean values are connected with solid lines and the standard deviation values are connected by dashed lines.
     - The figure must have one x-axis and one y-axis,

        - Let your code determine the y-axis limits automatically (do not write the limits value by hand in the code),
        - Add legend, labels to the axes, and a title to the sub-figure.
    2. On the right sub-figures:
        - On the upper right sub-figure, plot the mean and standard deviations values as two box plots.
            - Let your code determine the y-axis limits automatically (do not write the limits value by hand in the code),
            - Add labels to the axes and a title to the figure.
        - On the lower right sub-figure, plot the cumulative distribution function for the mean values only.
- Save the figure to a PNG file and submit it with the code script.

Good luck!