CS544 Module 1 Assignment

General Rules for Homework Assignments

- You are strongly encouraged to add comments for the code portions. Doing so will help your facilitator to understand your programming logic and grade you more accurately.
- You must work on your assignments individually. You are not allowed to copy the answers from the others.
- Each assignment has a strict deadline. However, you are still allowed to submit your assignment within 2 days after the deadline with a penalty.
  15% of the credit will be deducted unless you made previous arrangements with your facilitator and professor. Assignments submitted 2 days after the deadline will not be graded.
- When the term *lastName* is referenced in an assignment, please replace it with your last name.

Part1) 50 points

The data set *rivers* contains the lengths (in miles) of "major" rivers in North America, as compiled by the US Geological Survey. Use the data set to answer the following questions using R:

a) How many data points are there in the data set?
b) Compute the mean, median, and mode.
c) Compute the variance and the standard deviation.
d) Compute the five number summary, the interquartile range, and outliers, if any.
e) Compute the standardized version (z-scores) of the above data.
f) Create a matrix of size 2 x 30 using the first 60 data points in *rivers*. The first 30 values belong to the first row of the matrix. Assign the result to the variable, *rivers.60*, and display the result.
g) Without hardcoding, displaying the first and last columns of the matrix.
h) Assign row names for the *rivers.60* as Row_1 and Row_2 and column names as Length_1, Length_2, ….Length_30. The code should not hard code the values of the numbers in the row and column names.

Part 2) 50 points

The data file Johnson.csv contains quarterly earnings (dollars) per Johnson & Johnson share 1960–80.

a) Read the data from johnson.csv into a data frame. In the data frame, the data in "Year" column should be used as row names and "Qtr1", 'Qtr2", "Qtr3", and "Qtr4" should be column names.
b) Show the summary for earnings for each quarter.
c) Add a new column, Yearly, showing the earnings for the whole year (the sum of earnings for the 4 quarters). Display the new resulting data frame.

d) Which was the best performing year (in terms of highest earning) and worst performing year?
e) Show all rows of the data frame whose “Yearly” is greater than 20.

Submission:

Create a folder, CS544_HW1_lastName and place the following files in this folder.

Write the solution in a Word document, HW1_lastName.doc.

For the code portions (Part1 and Part2), provide all R code in a single file, HW1_lastName.r. For this homework only, you can earn an extra credit of 10 points by providing the solutions in a Jupyter Notebook, HW1_lastName.ipynb, instead. (You can include the R file and the Jupyter notebook if you wish).

Archive the folder (CS544_HW1_lastName.zip). Upload the zip file to the Assignments section of Blackboard.