DSCC 201/401 Homework Assignment #5
Due: **October 25, 2021 at 9 a.m. EDT**

Answers to these questions should be submitted via Blackboard. Only one submission for this assignment will be allowed. Revised submissions will not be allowed. So please make sure you only submit your final answers. All answers must be shown with the corresponding code using R. It is recommended that version 3.6.1 be used to complete the assignment on BlueHive. Please provide a text file or PDF of your R code showing BOTH input and output.

1. Given two matrices as follows:

$$X = \begin{pmatrix} 8 & 1 & 3 \\ 5 & -3 & 11 \\ -6 & 7 & 1 \end{pmatrix} \quad Y = \begin{pmatrix} -1 & 3 & 8 \\ 7 & -6 & 4 \\ -8 & 3 & 1 \end{pmatrix}$$

a) Compute the matrix algebra product (i.e. not the element-wise product) of X and the transpose of Y. Store the output in a matrix Z.
b) Calculate the trace of the multiplicative inverse of matrix Z.

2. A group of 25 doctors was sampled for their systolic blood pressure and the following values were recorded: 117, 119, 114, 123, 132, 109, 113, 129, 124, 122, 124, 115, 138, 130, 135, 154, 118, 119, 134, 128, 116, 119, 132, 123, 120.

a) Given that the mean systolic blood pressure is 120 ($\mu = 120$), is there a statistical significance for the mean of systolic blood pressures for this particular group of doctors? Check at the 95% confidence levels using R's t-test functionality.
b) Now check the statistical significance at the 99% confidence level. Based on the data, should we be concerned about the blood pressure readings of this group of doctors?

3. Copy the files `/public/bmort/R/summer2014.csv` and `/public/bmort/R/summer2021.csv` to your own directory on BlueHive. and answer the following questions:

a) Load the data from the files into two separate data frames. Run the summary command to observe the range of high temperatures for each of the data frames. What is the maximum high temperature in 2014? What is the maximum high temperature in 2021?

b) Combine the two data frames into one large data frame. First, remove the unnecessary SNW column by storing the value `NULL` in this column. This will essentially delete it.
c) Notice that the header names are different for the two data frames. You can verify this with the `names()` function on the data frame. The output of this function is a vector of strings. Decide on a consistent naming convention for the column labels of your data frames and convert one or the other data frames (or both) to follow this scheme.
d) Create a column in each of the data frames that represents the year for the data. The data from `summer2014.csv` will have an entry "2014" in a new column called `YEAR` for all rows of data. The data from `summer2021.csv` will have an entry "2021" in a new column for `YEAR` for all rows of the data from 2021. Use the `cbind()` function to add the vector of the corresponding years to each of the data frames.
e) Combine the two separate data frames into one data frame using the `rbind()` function.
f) Is there any missing data in the data frame? If so, where is it located? If there is missing data, compute the median value for the column containing the data for the missing column and replace the missing value with the median value.
g) For the combined data frame, which month showed lesser variation in high temperatures: June or July? What is the difference in standard deviation for the high temperatures for the two months?
h) What is the average daily rainfall and standard deviation for the period July 1, 2021 – July 31, 2021? How did you consider days when a trace (T) of rain was recorded?