

Problem

1. **(19 points)** For the data set **NationalParks.xlsx**, *Table 1.11 National Park Data: Johnson and Wichern, Applied Multivariate Statistical Analysis, Pearson, 6th edition, 2019.*
 - (a) Present a matrix scatter plot.
 - (b) Compute and display the sample correlation matrix.
 - (c) Identify the park that is unusual. Give a reason.
 - (d) Produce a new data set by removing the data for the unusual park. Present a matrix scatter plot from the new data.

In the reference for data files on the cover page, please add a note referring to the original data file and add a note that you also used the data obtained by removing an outlier
 - (e) Compute the sample correlation matrix for the new data set. How much impact did removing the unusual data point have?
 - (f) Include your code.

2. (27 points) Read in data set **survival_data.xlsx**, *A. Rencher and W. Christensen, “Methods of Multivariate Analysis”, Table 6.22, Wiley, 2012*. The data for variables B , C , D , E , and F are in columns 2 – 6. There are 5 groups of data indicated in the first column.
- (a) Display a boxplot of the data. Is there potential for scale and outlier issues? Explain.
 - (b) Display a matrix scatter plot, indicating data from different groups with different color symbols.
 - (c) Compute the sample mean vector and standard deviations across all groups, display.
 - (d) Compute the matrix $\overline{\mathbf{X}}^*$ of sample means within groups using the standardized values, display.
 - (e) Compute the distance matrix $\overline{\mathbf{D}}^*$ between the groups using the standardized values, display.
 - (f) Are there any groups that is distinguished by the results?
 - (g) Include your code.

3. (17 points) Read in data set **project3_3.txt**, *Data set for project 3 problem 3 made by Don, 2022*. The data points are provided in the first two columns and the group of the data is provided in the third column.
- (a) Display a matrix scatter plot, indicating data from different groups with different color symbols.
 - (b) Compute the matrix $\overline{\mathbf{X}}$ of sample means within groups, display.
 - (c) Compute the distance matrix $\overline{\mathbf{D}}$ between the groups, display.
 - (d) Should you conclude that the groups of data are fairly close? Explain your answer.
 - (e) Include your code.