Data Analysis & Statistics for Earth Scientists
Nanjing University, Spring 2026

Lab 1 Assignment

1. Open the Kentucky.dat datafile in the PAST program or the Kentucky.csv file in another mathematical application (e.g., MS-Excel, *Mathematica*™, MatLab™).

2. Create 10-bin histograms for all seven variables. (60 points)
    a. Identify the variable whose distribution most closely resembles a normal distribution. Be sure to justify your answer. (10 points)

3. Create scatterplots for the following variable pairs. (40 points)
    - Basin Magnitude vs. Basin Relief
    - Basin Magnitude vs. Basin Area
    - Basin Relief vs. Stream Length
    - Basin Area vs. Drainage Density

    a. Characterize each scatterplot using the following categories: wide scatter, narrow scatter, boundary-constrained scatter, categorical scatter, homoscedastic scatter, heteroscedastic scatter. Remember each scatterplot involves two variables. (40 points)

4. Create a 3D scatterplot of the Outlet Elevation, Stream Length and Drainage Density variables. (10 points)
    a. Describe the geometric relations among the basins. Are there any clusters? If so identify them. (30 points)

5. For each variable, list or calculate the following parameters and present these summaries in the form of a table: maximum value, arithmetic mean, geometric mean, median, minimum value, range of observations, variance, standard deviation, coefficient of variation. (20 points)

6. Using the table you created above, answer the following questions. (35 points)
    a. Variable with the greatest range?
    b. Variable with the smallest range?
    c. Variable with the greatest difference between its arithmetic and geometric mean?
    d. Variable with the greatest variance?
    e. Variable with the least variance?
    f. Variable with the greatest variability?
    g. Variable with the least variability?

7. Create ternary density plots of the following variable triplets. (30 points)
    - Outlet Elevation, Basin Relief, Stream Length
    - Basin Magnitude, Basin Area, Basin Shape
    - Basin Magnitude, Basin Area, Stream Length

    a. Using the plots you created above, identify any evidence for data clustering and provide an interpretation for the patterns seen in the plots (in terms of the influence of different variables. (20 points)

8. Make a network plot of the first ten basins using all variables. (10 points)
    Based on the plot you created above, along with the data table, identify and explain your reasoning for the following identifications. List all basins with equivalent characteristics. (30 points)
    a. Most distinct basin.
    b. Least distinct basin.
    c. Set of most similar basins.

Total = 335 points