SC 4024 Tutorial 6: Exploratory Data Analysis

AY2024-2025 Semester 1

1. Fundamental Concepts of Correlation Analysis

For the following Pearson correlation coefficients (r):

0.32, -0.76, 0.13, -0.25, 0.04, -0.03, 0.01

a) Which is the strongest correlation?

b) Which is the weakest correlation?

2. Exploratory Data Analysis

During the lecture of Chapter 9.1, we take the course scores of students as an example to illustrate how we can conduct exploratory data analysis via data visualization and statistical analysis. Now you are provided with a dataset file called "students_performance.csv", and you are asked to conduct exploratory data analysis on it. Suppose you are interested in **checking if there is a difference between the math scores and reading scores of students from group A**, please answer the following questions:

2.1 By referring to what we have learned about exploratory data analysis and correlation analysis, briefly describe **what kind of steps** you will take to answer such a question, and what kind of **visualizations** and **statistical tests** you will use.

2.2 What are your **null hypothesis (H_0)** and **alternate hypothesis (H_1)** for such an analysis?

2.3 By following the steps in 2.1, you are asked to write Python code to create appropriate visualizations and conduct the corresponding relevant statistic tests. Note: You are requested to use the Python packages like **scipy.stats** and **statsmodels.stats**.

2.4 Suppose you are asked to check there is a difference between the math scores and reading scores of **students from group B or group C**, what kind of changes do you need to do in your procedures of exploratory data analysis in 2.3?