Basic Inferential Analysis
Real-World Application Using Kaggle Dataset

1 Dataset: Students Performance in Exams

Source:
`https://www.kaggle.com/datasets/spscientist/students-performance-in-exams`

This dataset contains records for 1,000 students, including demographic characteristics and scores in mathematics, reading, and writing. Additional attributes include gender, parental level of education, lunch type, and completion of a test preparation course.

You are required to use this dataset to perform the following statistical analyses using either **R** or **Python**. All answers must include: relevant code, statistical output (e.g., tables or plots), and concise, context-driven interpretations.

2 Question (a): Hypothesis Testing

Objective: Test whether completing a test preparation course affects students' math scores.

1. Formulate the null and alternative hypotheses to compare the mean math scores of students who completed the test preparation course and those who did not.
2. Choose and justify the appropriate statistical test.
3. Report the test statistic, and p-value.
4. State the assumptions of the test and check whether they are met.
5. Interpret the result and conclude at the 5% significance level.
6. **Further analysis:** Repeat the hypothesis test for reading score and writing score. Comment on whether the pattern of results is consistent across all subjects.

3 Question (b): Correlation Analysis

Objective: Explore the relationships among students' scores.

1. Compute the Pearson correlation coefficients between math score, reading score, and writing score.

2. Display the results using a correlation matrix or heatmap.

3. Identify the strongest positive correlation and interpret its academic relevance.

4. Briefly discuss whether multicollinearity may be a concern in this dataset.

4 Question (c): Linear Regression Modeling

Objective: Build a multiple linear regression model to predict math scores.

1. Use math score as the dependent variable. Select at least two independent variables, including one continuous (e.g., reading score) and one categorical (e.g., test preparation course). Fit a multiple linear regression model using either R or Python.

2. Report the model formula, coefficients, R-squared value, and p-values.

3. Interpret the meaning of at least one predictor coefficient in context.

4. Briefly comment on the overall model fit and check residuals for basic assumptions.

General Instructions:

- Include all code and statistical outputs with appropriate formatting.

- Use correct statistical terminology and notation throughout.

- Ensure that all interpretations are directly related to the dataset context.

- Upload your completed analysis and code to a public GitHub repository and share the link along with your assignment.