Statistical Analysis and Visualization with Python

Objective:

This project aims to perform statistical analysis and data visualization using Python. You will generate a synthetic dataset, perform descriptive and inferential statistics, and visualize the results using Matplotlib and Seaborn.

Instructions for Students:

Project Overview

You will generate a synthetic dataset and use it to perform various statistical analyses. Your final submission should include a presentation highlighting your findings, visualizations, and interpretations.

Deliverables

- A presentation summarizing your findings, visualizations, and interpretations.
- Ensure your presentation addresses the technical questions provided.

Tools and Libraries

You will use the following Python libraries:

- Pandas
- Numpy
- Matplotlib
- Seaborn

Tasks and Guidelines

Task 1: Data Generation

1. **Generate a synthetic dataset**:
    - Use Numpy to create a dataset with 1000 samples.
    - The dataset should include the following columns: Age, Height, Weight, Gender, and Income.
        - Age: Normally distributed with a mean of 35 and a standard deviation of 10.
        - Height: Normally distributed with a mean of 170 cm and a standard deviation of 15 cm.
        - Weight: Normally distributed with a mean of 70 kg and a standard deviation of 10 kg.
        - Gender: Randomly assigned with 50% probability for 'Male' and 'Female'.
        - Income: Normally distributed with a mean of 50,000 and a standard deviation of 15,000.

Task 2: Descriptive Statistics

2. **Calculate basic descriptive statistics**:
    - Calculate mean, median, standard deviation, and variance for Age, Height, Weight, and Income.
3. **Calculate the mode for Gender**.

Task 3: Data Visualization

4. **Visualize the distributions**:
    - Plot histograms for Age, Height, Weight, and Income.
    - Use Seaborn to create KDE plots for Age, Height, Weight, and Income.
5. **Boxplots to identify outliers**:
    - Create boxplots for Age, Height, Weight, and Income.

Task 4: Correlation Analysis

6. **Calculate correlation**:
    - Calculate the Pearson correlation coefficient between Age, Height, Weight, and Income.
7. **Visualize the correlation matrix**:
    - Use a heatmap to visualize the correlation matrix.

Task 5: Inferential Statistics

8. **Hypothesis Testing**:
    - Perform a t-test to see if there is a significant difference in Income between Male and Female.

Technical Questions

Ensure your presentation addresses the following questions:

1. How did you generate the synthetic dataset, and why did you choose normal distributions for the variables?
2. What insights can you draw from the descriptive statistics calculated for Age, Height, Weight, and Income?
3. What do the KDE plots and histograms tell you about the distribution of the data?
4. How can you interpret the boxplots, and what do they reveal about potential outliers in the dataset?
5. Discuss the significance of the correlation matrix and what relationships you can infer from it.
6. Explain the results of the t-test. What does the p-value indicate about the difference in Income between Male and Female?

Submission Guidelines

- Prepare a presentation in PowerPoint, Google Slides, or any other preferred format.
- Include slides that explain your methodology, present your visualizations, and answer the technical questions.
- Ensure your presentation is clear, concise, and well-organized.
- Submit your presentation by the specified deadline.