Python Homework Assignment: Introduction to Data Analysis

Sybil Prince Nelson

January 27, 2025

Objective

This assignment introduces you to the basics of data analysis using Python. You will perform simple data analysis tasks, calculate descriptive statistics, and perform hypothesis tests such as t-tests, z-tests, and ANOVA.

Instructions

Use Python in your environment (preferably RStudio Workbench) to complete this assignment. Make sure to import the necessary libraries and provide explanations for each step. All answers should be clearly presented and include code snippets where applicable.

Questions

1. **Setting up your Python Environment:**
    - Open RStudio Workbench and set up a new Python project.
    - Import the following libraries:
        - pandas
        - numpy
        - scipy
        - matplotlib
        - seaborn
    - Verify that you can import these libraries without errors.
2. **Data Loading:**
    - Download the provided CSV file: `data.csv`.
    - Load the data into a pandas DataFrame and display the first 5 rows of the dataset.

3. **Descriptive Statistics:**

   - Calculate the mean, median, and standard deviation of the numerical columns in your dataset.
   - Generate a summary of the dataset that includes the count, mean, standard deviation, minimum, and maximum values for each numerical column.

4. **Data Visualization:**

   - Create a histogram for one of the numerical columns in your dataset.
   - Create a boxplot to visualize the distribution of this column and check for any outliers.

5. **t-test:**

   - The dataset includes two groups of values in one of the numerical columns. Perform an independent two-sample t-test to check if the means of these two groups are significantly different.
   - State the null and alternative hypotheses.
   - Perform the t-test and report the p-value and conclusion at a significance level of 0.05.

6. **z-test:**

   - In the same dataset, assume the population mean for a numerical column is 50. Perform a one-sample z-test to determine if the sample mean is significantly different from the population mean.
   - State the null and alternative hypotheses.
   - Perform the z-test and report the p-value and conclusion at a significance level of 0.05.

7. **ANOVA:**

   - The dataset includes a categorical variable called Region, which contains four levels: North, South, East, and West.
   - Perform a one-way ANOVA to determine if there is a significant difference in Income across these four regions.
   - State the null and alternative hypotheses.
   - Report the F-statistic and the p-value, and provide a conclusion at a significance level of 0.05.

8. **Confidence Intervals:**

   - Calculate a 95% confidence interval for the mean of one numerical column in your dataset.

- Interpret the confidence interval and explain what it means in the context of your data.

9. **Correlation Analysis:**
    - Calculate the Pearson correlation coefficient between two numerical columns in your dataset.
    - Visualize the relationship between these two variables with a scatter plot.
    - Based on the correlation, provide a brief explanation of the relationship between the two variables.
10. **Bonus Regression Analysis:**
    - Perform a simple linear regression analysis where you predict one numerical column using another numerical column.
    - Interpret the coefficients of the regression model.
    - Calculate the R-squared value and interpret its meaning in the context of your data.