Assignment 2, data analysis

All assignments must be done in anaconda environment by the use of Jupyter notebook. All operations, datasets should be MEANINFGUL AND USEFUL

Exploratory Data Analysis (EDA) with Pandas: [20 points]

- Load and Inspect any Data that you find: Load a dataset (e.g., CSV) using Pandas. Display basic information about the dataset, such as the number of rows and columns, data types, and the first few rows of data.
- Data Cleaning: Clean the dataset by handling missing values, duplicates, or outliers.
- Summary Statistics: Compute summary statistics (mean, median, standard deviation) for numerical columns in the dataset. Identify any trends or insights.

Data Visualization with Matplotlib and Seaborn: [40 points]

- Histogram: Create a histogram to visualize the distribution of a numerical variable in your dataset. Choose appropriate bins and labels.
- Box Plot: Generate a box plot to visualize the distribution of a numerical variable, grouped by a categorical variable. For example, you can compare the distribution of salaries for different job positions.
- Scatter Plot: Create a scatter plot to explore the relationship between two numerical variables. Add a regression line to visualize any trends.
- Bar Plot: Create a bar plot to show the frequency of unique values in a categorical variable. Label the bars and axes.

Hypothesis Testing with t-test: [40 points]

- Data Splitting: Split your dataset into two groups based on a categorical variable (e.g., treatment group and control group).
- t-test: Perform an independent two-sample t-test to determine if there is a statistically significant difference between the means of two groups for a numerical variable. Interpret the results and state your hypothesis.
- Visualization: Visualize the distributions of the two groups using histograms and box plots. Overlay the means and confidence intervals to visually assess the t-test results.