Statistics for Data Analysis /

Assignment set by @Zainalysis for Rimsha.

Instructions:

- This assignment is designed to guide you from basic to advanced levels of statistical analysis used in data projects.
- Each section builds on the previous one, and practical exercises/projects are included to reinforce learning.
- Refer to the provided links for guidance and solutions.

Section 1: Basic Descriptive Statistics

1. **Question:** Calculate the mean, median, mode, range, variance, and standard deviation for the following dataset: `[12, 15, 14, 10, 8, 11, 18, 14, 16, 19, 21]`
   Help: Khan Academy - Measures of Central Tendency
2. **Question:** Create a frequency distribution table and histogram for the dataset above.
   Help: Statology - Frequency Distribution
3. **Question:** Calculate the 25th, 50th, and 75th percentiles for the dataset.
   Help: Khan Academy - Quartiles and Percentiles

Section 2: Inferential Statistics

4. **Question:** Conduct a t-test to compare two datasets: `[23, 21, 25, 22, 27, 29, 30]` and `[18, 20, 17, 19, 15, 22, 24]`. Determine if there is a significant difference between them.
   Help: SciPy - T-test
5. **Question:** Perform a chi-square test on the following contingency table:

Category	A	B
X	20	30
Y	25	35

Help: Statology - Chi-square Test

6. **Question:** Conduct a one-way ANOVA for the following groups: `Group 1: [5, 7, 9, 6, 8]`, `Group 2: [10, 12, 11, 14, 13]`, `Group 3: [15, 17, 16, 18, 19]`.
   Help: Khan Academy - ANOVA

Section 3: Regression Analysis

7. **Question:** Perform a simple linear regression using the following dataset:
   `X = [1, 2, 3, 4, 5, 6, 7]`
   `Y = [2, 4, 5, 4, 6, 8, 9]`
   Help: Statology - Linear Regression
8. **Question:** Evaluate the goodness of fit for the regression model by calculating R-squared.
   Help: Real Python - R-squared Calculation
9. **Question:** Implement multiple linear regression with the following dataset:
   `X1 = [1, 2, 3, 4, 5]`,
   `X2 = [5, 3, 6, 7, 8]`,
   `Y = [10, 12, 14, 15, 17]`
   Help: Statology - Multiple Linear Regression

Section 4: Advanced Topics

10. **Question:** Perform a logistic regression analysis for the following binary classification dataset:

```
X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
Y = [0, 0, 0, 1, 1, 1, 1, 1, 1, 1]
```

Help: Towards Data Science - Logistic Regression

11. **Question:** Conduct a principal component analysis (PCA) for the following matrix:

```
Matrix = [[1, 2, 3], [4, 5, 6], [7, 8, 9], [10, 11, 12]]
```

Help: Medium - PCA Tutorial

12. **Question:** Implement k-means clustering for the following dataset:

```
X = [[1, 2], [2, 3], [3, 4], [8, 9], [9, 10], [10, 11]]
```

Help: Scikit Learn - k-means Clustering

Project: End-to-End Data Analysis

Task:

- Use any open dataset (e.g., Titanic, Iris, or your own) to perform exploratory data analysis (EDA), apply inferential statistics, build regression models, and present findings.
- Write a full report including visualizations, statistical summaries, and insights.

Help:

- Kaggle - Free Datasets
- Towards Data Science - EDA Guide