Title:  Data Analysis I

Name:  Vanisha Prasad

Course Information:  PST107 Probability and Statistics

## Introduction

This project demonstrates my ability to perform data analysis using R programming. The aim is to clean, analyze, and extract meaningful insights from a small dataset containing information about individuals' names, ages, and scores.

The dataset was loaded, cleaned, and processed to ensure no missing values, and summary statistics were calculated to provide an overview of the data.

## Source Code Explanation

### *1. Loading the Data*

The dataset was loaded from a CSV file named `example_data.csv` using the `read.csv()` function. The structure of the dataset was examined to ensure proper formatting.

**Code:**

```
# Load the CSV file into a data frame
data <- read.csv("example_data.csv")

# Check the structure of the data
str(data)

# View the first few rows to inspect the data
head(data)
```

**Explanation:**

- `read.csv()` reads the file into R.
- `str()` checks the structure of the dataset.
- head() provides a preview of the first few rows.

## 2. Cleaning the Data

The dataset was checked for missing values and ensured that numeric columns were properly formatted. Missing values were handled by replacing them with the mean for numeric variables.

**Code:**

```
# Check for missing values
sum(is.na(data))

# Replace missing Age values with the mean
data$Age[is.na(data$Age)] <- mean(data$Age, na.rm =
TRUE)

# Convert Age and Score to numeric
data$Age <- as.numeric(data$Age)
data$Score <- as.numeric(data$Score)
```

**Explanation:**

- `is.na()` checks for missing values.
- The mean of the Age column was calculated and used to replace missing values.
- Numeric columns were explicitly converted to numeric types to avoid errors.

### Generating Summary Statistics

Summary statistics were calculated to provide an overview of the dataset, including minimum, maximum, and mean values for numeric columns.

**Code:**

```
# Summary statistics
summary(data)
```

**Explanation:** The `summary()` function displays a statistical summary for each column in the dataset.

### *4. Saving the Cleaned Data*

The cleaned data was saved to a new CSV file for record-keeping and further analysis.

**Code:**

```
# Save the cleaned data to a new file
write.csv(data, "cleaned_data.csv", row.names = FALSE)
```

**Explanation:**

- `write.csv ()` writes the modified dataset to a new file, excluding row names for simplicity

## Results

- **Dataset Overview:** The dataset contains 3 observations and 3 variables: Name (character), Age (numeric), and `Score` (numeric)

Summary Statistics

| Variable | Min | 1st Qu. | Median | Mean | 3rdQu. | **Max.** | Class | Mode |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Name | - | - | - | - | - | - | character | character |
| Age | 22.0 | 22.5 | 23.0 | 24.0 | 25.0 | 27.0 | numeric | - |
| Score | 78.0 | 82.5 | 87.0 | 86.67 | 91.0 | 95.0 | numeric | - |

### **Conclusion**

This project successfully demonstrated key data analysis steps, including data loading, cleaning, and summarizing using R. The cleaned dataset was saved for future use, ensuring data integrity and usability.

### **Appendix: Cleaned Dataset**

The cleaned dataset is included in the file `cleaned_data.csv` with the following content:

Cleaned Dataset

| NAME | AGE | SCORE |
|---|---|---|
| ALICE | 23 | 95 |
| BOB | 27 | 87 |
| CHARLIE | 22 | 78 |