Part 1: Calculating Descriptive Statistics

1. **Create a Sample Dataset**:
   - Start by creating a simple dataset in R to calculate basic descriptive statistics:

```
# Sample dataset
data <- c(5.2, 6.8, 5.9, 7.1, 6.0, 6.5, 7.3, 5.7, 6.2, 6.9)
```

2. **Calculate Descriptive Statistics**:
   - Calculate the mean, median, and standard deviation of the dataset:

```
# # Mean
mean_value <- mean(data)

# Median
median_value <- median(data)

# Standard Deviation
sd_value <- sd(data)

# Summary statistics
summary(data)
```

   - **Task**: Run the script to generate the scatter plot. What does the plot reveal about the relationship between Variable1 and Variable2?

3. **Interpret the results:**
   - Reflect on the calculated statistics. How do the mean and median compare? What does the standard deviation tell you about the data spread?

Part 2: Handling Missing Data

1. **Introduce Missing Data**:

   - Modify your dataset to include some missing values:

```
# Dataset with missing values
data_with_na <- c(5.2, 6.8, NA, 7.1, 6.0, 6.5, NA, 5.7, 6.2,
6.9)
```

2. **Handle Missing Data by Imputation**:
   - Replace the missing values with the mean of the non-missing data:

```
# Handling missing data by imputation (mean)
data_imputed <- ifelse(is.na(data_with_na), mean(data_with_na,
na.rm = TRUE), data_with_na)
```

   - **Task**: Run the script to impute the missing values. How does imputing missing values affect the dataset? Why might this be important for analysis?

3. **Compare the Original and Imputed Datasets:**
    - Compare the summary statistics of the original dataset with missing values and the imputed dataset:

```
# Summary of the original dataset with missing values
summary(data_with_na)

# Summary of the imputed dataset
summary(data_imputed)
```

    - **Task**: Compare the two summaries. How does the imputation process change the descriptive statistics?

Part 3: Detecting and Handling Outliers

1. **Detect Outliers Using Z-Scores:**
    - Calculate Z-scores for the imputed dataset and identify any outliers:

```
# Calculate Z-scores
z_scores <- (data_imputed - mean(data_imputed)) /
sd(data_imputed)

# Identify outliers
outliers <- abs(z_scores) > 2

# Display outliers
data_imputed[outliers]
```

    - **Task** Task: Run the script to generate the bar chart. What does the bar chart show about the distribution of values across categories?

2. **Handle outliers**:
    - One way to handle outliers is to remove them from the dataset:

```
# Remove outliers
data_cleaned <- data_imputed[!outliers]
```

    - **Task:** Run the script to remove the outliers. Compare the descriptive statistics of the cleaned dataset to the original dataset. How does removing outliers affect your data?

Part 4: Reflect and Document

1. **Reflect**:
    - Reflect on the process of cleaning and analyzing your data. What challenges did you encounter? What insights did you gain from this exercise?
2. **Document Your Work**:
    - Add comments to your script explaining each step.

    - Save your script as week3_exercise.R and be prepared to discuss your experience in the next class.

Submission Instructions

- **Save your script**: Ensure your script is well-commented and saved as `week3_exercise_name.R`.
- **Upload**: Submit your script via TEAMS.
- **Discussion**: Be ready to discuss your experiences and any questions you have during the next class.

Expected Outcome

By the end of this exercise, you should be comfortable with calculating descriptive statistics, handling missing data, detecting and dealing with outliers, and preparing your dataset for further analysis. These skills are essential for any data analysis task, especially in the context of ecological data, where data quality can significantly impact results.
