Data Cleaning and Hypothesis Testing Report

---

1. Introduction

This document outlines the data cleaning process and hypothesis testing performed on a dataset related to food, nutrition, and health habits. The data was first cleaned using Python's Pandas library, and then statistical analysis and visualizations were carried out using R.

---

2. Data Cleaning with Pandas (Python)

The following code was used to clean the dataset:

```
import pandas as pd

# Load the dataset
file_path = "cleaned_food_data.xlsx"
data = pd.read_excel(file_path)

# Check for missing values
data.isnull().sum().sort_values(ascending=False)

# Drop or fill missing values as necessary (example)
data.dropna(subset=['GPA', 'breakfast', 'fruit_day', 'veggies_day'], inplace=True)

# Convert data types if needed
data['GPA'] = pd.to_numeric(data['GPA'], errors='coerce')
data['fruit_day'] = pd.to_numeric(data['fruit_day'], errors='coerce')
data['veggies_day'] = pd.to_numeric(data['veggies_day'], errors='coerce')

# Save cleaned data
data.to_excel("cleaned_food_data.xlsx", index=False)
```

---

3. Hypothesis Testing and Visualizations (R)

```
library(readxl)
data <- read_excel("cleaned_food_data.xlsx")

# Convert GPA to numeric
data$GPA <- as.numeric(as.character(data$GPA))

# Hypothesis 1: GPA and Breakfast
```

```
boxplot(GPA ~ breakfast, data = data, main = "GPA vs Breakfast", xlab = "Breakfast (1=Yes, 0=No)", ylab = "GPA")
t.test(GPA ~ breakfast, data = data)

# Hypothesis 2: Vegetable intake vs Vitamin use
boxplot(as.numeric(veggies_day) ~ vitamins, data = data, main = "Vegetable Intake vs Vitamin Use", xlab = "Takes Vitamins (1=Yes, 2=No)", ylab = "Vegetables per Day")
t.test(as.numeric(veggies_day) ~ vitamins, data = data)

# Hypothesis 3: Gender and Breakfast (Chi-squared Test)
table(data$Gender, data$breakfast)
chisq.test(table(data$Gender, data$breakfast))

# Hypothesis 4: Calories per day vs GPA (Correlation)
plot(data$calories_day, data$GPA, main = "Calories vs GPA", xlab = "Calories per Day", ylab = "GPA")
cor.test(data$calories_day, data$GPA, use = "complete.obs")

# Hypothesis 5: Fruit and Vegetable consumption (Correlation)
data$fruit_day <- as.numeric(data$fruit_day)
data$veggies_day <- as.numeric(data$veggies_day)
cor.test(data$fruit_day, data$veggies_day, use = "complete.obs")
```

---

4. Inferences

- **Hypothesis 1 (GPA ~ Breakfast):**
    - p-value: 0.9601
    - No statistically significant difference in GPA between those who eat breakfast and those who do not.
- **Hypothesis 2 (Vegetable intake ~ Vitamins):**
    - p-value: 0.00989
    - Significant difference in vegetable intake between vitamin users and non-users. Vitamin users tend to eat more vegetables.
- **Hypothesis 3 (Gender ~ Breakfast):**
    - p-value: 0.8378
    - No significant association between gender and breakfast habits.

- **Hypothesis 4 (Calories vs GPA):**
  - Correlation coefficient: -0.0947753
  - p-value: 0.4
  - Very weak and non-significant negative correlation between calorie intake and GPA.
- **Hypothesis 5 (Fruit vs Vegetable intake):**
  - Correlation coefficient: 0.6653829
  - p-value: 5.046e-12
  - Strong, significant positive correlation between fruit and vegetable consumption.

---

5. Conclusion

The analysis indicates that vitamin use correlates positively with vegetable consumption, and fruit and vegetable consumption are strongly related. However, breakfast, calorie intake, and gender do not show statistically significant relationships with GPA in this dataset.

Further analysis with a larger dataset and additional variables could provide deeper insights.
