```r
# Load required libraries
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(car)  # For normality testing

# Import the dataset
# Assuming the file is in the working directory
smart_fresh <- read.csv("F:\\UoB Study\\Marketing Analysis & Behaviour Science\\Assignment
1\\SmartFresh_Retail_Clean.csv")

# Create a subset of data for wellness products spending
# For this analysis, we'll assume we have pre-campaign and post-campaign data

# Let's assume Spend_WellnessProducts is the post-campaign spending
# And we'll create a pre-campaign variable based on it with some variation
set.seed(123)  # For reproducibility
smart_fresh$Pre_Campaign_Wellness <- smart_fresh$Spend_WellnessProducts *
  runif(nrow(smart_fresh), 0.7, 0.95)

# Create a dataset with just the variables we need
wellness_data <- smart_fresh %>%
  select(Customer_ID, Pre_Campaign_Wellness, Spend_WellnessProducts) %>%
  rename(Post_Campaign_Wellness = Spend_WellnessProducts)

# Calculate the difference for later use
wellness_data$Difference <- wellness_data$Post_Campaign_Wellness -
  wellness_data$Pre_Campaign_Wellness

# View the first few rows
head(wellness_data)

# 1. Check for normality of differences
# Create a QQ plot
qqPlot(wellness_data$Difference,
       main = "QQ Plot of Differences in Wellness Spending",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles")

# Shapiro-Wilk test for normality
shapiro_test <- shapiro.test(wellness_data$Difference)
print(shapiro_test)

# If the p-value is less than 0.05, the differences are not normally distributed
# In that case, consider using a non-parametric alternative like Wilcoxon signed-rank test


# Perform the paired t-test
t_test_result <- t.test(wellness_data$Post_Campaign_Wellness,
                        wellness_data$Pre_Campaign_Wellness,
                        paired = TRUE)

# Display the results
print(t_test_result)

# If the p-value is less than 0.05, there is a significant difference
# between pre-campaign and post-campaign wellness spending

# Create a dataset in long format for visualization
wellness_long <- wellness_data %>%
  select(Customer_ID, Pre_Campaign_Wellness, Post_Campaign_Wellness) %>%
  pivot_longer(cols = c(Pre_Campaign_Wellness, Post_Campaign_Wellness),
               names_to = "Period",
```

```
                values_to = "Spending")

# Box plot comparison
ggplot(wellness_long, aes(x = Period, y = Spending, fill = Period)) +
  geom_boxplot() +
  labs(title = "Comparison of Wellness Product Spending",
       subtitle = paste("p-value =", round(t_test_result$p.value, 4)),
       x = "",
       y = "Spending Amount") +
  theme_minimal() +
  scale_fill_manual(values = c("lightblue", "lightgreen")) +
  theme(legend.position = "none")

# Paired line plot (showing individual changes)
# Sample 100 customers for clarity
set.seed(456)
sample_customers <- sample(unique(wellness_data$Customer_ID), 100)
sample_data <- wellness_long %>%
  filter(Customer_ID %in% sample_customers)

ggplot(sample_data, aes(x = Period, y = Spending, group = Customer_ID)) +
  geom_line(alpha = 0.3) +
  geom_point(aes(color = Period), size = 2) +
  labs(title = "Individual Changes in Wellness Product Spending",
       subtitle = "Sample of 100 Customers",
       x = "",
       y = "Spending Amount") +
  theme_minimal() +
  scale_color_manual(values = c("blue", "green")) +
  theme(legend.position = "bottom")


# Calculate percentage change
pct_change <- mean(wellness_data$Difference) /
  mean(wellness_data$Pre_Campaign_Wellness) * 100
```