```r
# Hypothesis Testing Assigment
# Daniel Blumberg

# Install packages
install.packages("rstatix")
install.packages("ggplot2")
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(lubridate)
library(car)      # Required for Levene's Test
library(rstatix)  # Required for eta_squared and other statistical functions
alpha <- 0.05     # Set the significance level

# --- SECTION 1: DATA LOADING AND PREPARATION ---

# 1. Load the data from the CSV file
library(readxl)
fda_data <- read_excel("C:/Users/deblu/Downloads/FDAcleaned (1).xlsx")
View(fda_data)

# 2. Corrected: Convert 'recall_date' to a proper Date object and filter out missing dates
# The column name is corrected to 'recall_date'
fda_data <- fda_data %>%
  mutate(RecallDate_Clean = as.Date(recall_date, format = "%m/%d/%Y")) %>%
  filter(!is.na(RecallDate_Clean))

# 3. Create 'Quarter' and aggregate the total count of recalls per quarter-year
# This creates the dataset needed for ANOVA
quarterly_recalls <- fda_data %>%
  mutate(Year = year(RecallDate_Clean),
         Quarter = quarter(RecallDate_Clean, with_year = FALSE)) %>%
  group_by(Year, Quarter) %>%
  summarise(RecallCount = n(), .groups = 'drop') %>%
  mutate(Quarter_Group = factor(paste0("Q", Quarter)))

print("Head of Prepared Quarterly Recall Data:")
print(head(quarterly_recalls))


# --- SECTION 2: ASSUMPTION CHECKS (ANOVA) ---
print("\n--- SECTION 2: ASSUMPTION CHECKS (ANOVA) ---")

# 1. Normality Check (Shapiro-Wilk Test per Quarter - H0: Data is Normal)
print("1. Normality Check (Shapiro-Wilk p-values):")
normality_check <- quarterly_recalls %>%
  group_by(Quarter_Group) %>%
  filter(n() >= 3) %>%
  summarise(Shapiro_p = shapiro.test(RecallCount)$p.value)
print(normality_check)

# 2. Homogeneity of Variance (Levene's Test - H0: Variances are Equal)
print("\n2. Homogeneity of Variance Check (Levene's Test):")
levene_result <- leveneTest(RecallCount ~ Quarter_Group, data = quarterly_recalls)
print(levene_result)


# --- SECTION 3: CONDUCT THE STATISTICAL TEST (ANOVA) ---
print("\n--- SECTION 3: CONDUCT THE STATISTICAL TEST (ANOVA) ---")
print(paste("Significance Level (alpha):", alpha))

# 1. Perform the One-Way ANOVA
anova_model <- aov(RecallCount ~ Quarter_Group, data = quarterly_recalls)
print("\n1. ANOVA Summary Table:")
```

```
anova_summary <- summary(anova_model)
print(anova_summary)

# 2. Calculate the Effect Size (Eta-squared)
print("\n2. Effect Size (Eta-squared):")
effect_size <- anova_model %>%
  eta_squared()
print(effect_size)

# 3. Post-Hoc Test (Tukey HSD) - for pairwise comparisons
print("\n3. Post-Hoc Test (Tukey HSD - Pairwise Comparisons):")
print(as.data.frame(TukeyHSD(anova_model)$Quarter_Group))


# --- SECTION 4: INTERPRETATION AND RESULTS SUMMARY ---
print("\n--- SECTION 4: INTERPRETATION AND RESULTS SUMMARY ---")

# Extracting key results
p_value_anova <- anova_summary[[1]]$'Pr(>F)'[1]
f_stat <- anova_summary[[1]]$'F value'[1]
df1 <- anova_summary[[1]]$'Df'[1]
df2 <- anova_summary[[1]]$'Df'[2]

if (p_value_anova <= alpha) {
  interpretation <- paste0(
    "1. DECISION: REJECT the Null Hypothesis (H0).",
    "\n2. CONCLUSION: There IS a statistically significant association between the time of
year (quarter) and the number of recalls."
  )
} else {
  interpretation <- paste0(
    "1. DECISION: FAIL TO REJECT the Null Hypothesis (H0).",
    "\n2. CONCLUSION: There is NO statistically significant association between the time
of year (quarter) and the number of recalls.",
    "\n   The observed differences are likely due to random chance."
  )
}

print(paste0(
  "ANOVA Test Result: F(", df1, ", ", df2, ") = ", round(f_stat, 2), ", p = ",
round(p_value_anova, 4)
))
print(interpretation)


# Visuals

# --- VISUALIZATION 1: BOX PLOT ---

# Purpose: Shows the distribution (median, quartiles, outliers) of recalls for each
quarter.
# This visually confirms the Normality and Homogeneity of Variance assumptions.

plot_box <- ggplot(quarterly_recalls, aes(x = Quarter_Group, y = RecallCount, fill =
Quarter_Group)) +
  geom_boxplot(alpha = 0.7) +
  geom_point(position = position_jitter(width = 0.1), size = 2, alpha = 0.6) + # Adds
individual data points
  labs(
    title = "Distribution of Quarterly FDA Recalls",
    x = "Quarter of the Year",
    y = "Number of Recalls (per Quarter-Year)",
    caption = "Individual data points are shown."
  ) +
  theme_minimal() +
```

```
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5))

print(plot_box)

# --- VISUALIZATION 2: BAR CHART WITH ADJUSTED LABEL ---

# --- VISUALIZATION 2: BAR CHART WITH MULTIPLE COLORS & LABELED GRAND MEAN ---

# Calculate the Grand Mean (from the 'quarterly_recalls' data frame)
grand_mean <- mean(quarterly_recalls$RecallCount)

# Start the plot by mapping 'fill' to Quarter_Group in the main aes() call
plot_bar_multi_color <- ggplot(quarterly_recalls,
                               aes(x = Quarter_Group, y = RecallCount, fill =
Quarter_Group)) +

  # Use stat_summary to calculate the mean and plot the bar (NO fixed 'fill' here)
  stat_summary(fun = mean, geom = "bar", alpha = 0.8) +

  # Plot the standard error (SE) bars
  stat_summary(fun.data = mean_se, geom = "errorbar", width = 0.2, color = "black") +

  # Add the Grand Mean reference line
  geom_hline(yintercept = grand_mean, linetype = "dashed", color = "red", linewidth = 0.8)
+

  # Add the non-overlapping, blue label for the Grand Mean
  geom_text(x = 0.05, y = grand_mean * 1.1,
            label = paste0("Grand Mean: ", round(grand_mean, 0)),
            color = "red", size = 4, hjust = -0.5) +

  labs(
    title = "Mean Quarterly FDA Recalls with Standard Error",
    x = "Quarter of the Year",
    y = "Mean Number of Recalls (per Quarter)"
  ) +
  # Use a color palette and suppress the redundant legend
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position = "none") +
  # Use the 'Set2' color palette for distinct colors
  scale_fill_brewer(palette = "Set2")

# Print the final plot
print(plot_bar_multi_color)
```