```r
# R course for beginners
#Final assignment by Anastasia Keidar, id 322044082

# in MAC Shift + Command + 0 to restart the R session (clear workspace and packages)
# ^ + L to clear the console.

# Loading required packages

library(dplyr)
library(gridExtra)
library(ggplot2)
library(skimr)
library(broom)
library(pROC)

# laod data

data <- read.csv("./OSA-childhood trauma - Raw Data.csv")

# Step 1 - Defining the research question ----

# 1. Dataset - Write a few sentences describing the data set and why you chose it.

# The dataset used in this analysis examines various health-related factors, including
daytime sleepiness (ESS), presence of obstructive sleep apnea (OSA) and history of
childhood trauma.
# It includes measures of trauma history (ETISR-SF subscales), demographic variables
(Age, BMI, Menopause status), and clinical outcomes.
# I chose this dataset because my research focuses on childhood abuse, and I wanted to
work with existing data rather than my own.
# Additionally, since my doctoral research does not involve experimental or control
groups, I wanted to gain experience analyzing data that includes distinct comparison
groups (OSA vs. Control).
# This allowed me to explore methodologies and statistical approaches that I might not
typically use in my work.

# 2. Preview data using the ggdist or ggplot package.

# Descriptive statistics

summary_stats <- data |>
  group_by(Group) |>
  summarise(
    across(c(Age, BMI, ESS, REIeventshour, nadir, baseline, GAD7, ETISRSF_General,
ETISRSF_Physical, ETISRSF_Emotional, ETISRSF_Sexual),
           list(mean = ~mean(.x, na.rm = TRUE),
                sd = ~sd(.x, na.rm = TRUE),
                range = ~quantile(.x, probs = c(0,1), na.rm = TRUE) |> paste(collapse =
" - ")),
           .names = "{.col}_{.fn}"),
    Menopause_mode = names(sort(table(Menopause), decreasing = TRUE))[1],
    .groups = "drop"
    )
print(summary_stats)

# Graph display for the distribution of variables
# Click "zoom" for optimal viewing.

variables <- c("Age", "BMI", "ESS", "REIeventshour", "nadir", "baseline", "GAD7",
"ETISRSF_General", "ETISRSF_Physical", "ETISRSF_Emotional", "ETISRSF_Sexual")

group_sizes <- data |>
  group_by(Group) |>
  summarise(n = n()) |>
  mutate(label = paste0(Group, "\n(n=", n, ")"))

group_labels <- setNames(group_sizes$label, group_sizes$Group)
```

```r
plot_list <- lapply(variables, function(var) {
  ggplot(data, aes(x = as.factor(Group), y = .data[[var]], fill = as.factor(Group))) +
    geom_boxplot(alpha = 0.7) +
    labs(x = "Group", y = var) +
    scale_x_discrete(labels = group_labels) +
    theme_minimal()
  })

grid.arrange(grobs = plot_list, ncol = 2)

# Displaying a variable "Menopause" separately due to it being categorical

menopause_labels <- c("1" = "Pre", "2" = "Peri", "3" = "Post")

group_sizes <- data |>
  count(Menopause, Group) |>
  mutate(Menopause = factor(Menopause, levels = c("1", "2", "3"), labels = c("Pre",
"Peri", "Post")))

ggplot(group_sizes, aes(x = Menopause, y = n, fill = as.factor(Group))) +
  geom_bar(stat = "identity", position = "dodge", alpha = 0.6, color = "black") +
  geom_text(aes(label = paste0("n = ", n)),
            position = position_dodge(width = 0.9), vjust = -0.5, size = 3, fontface =
"bold") +
  scale_fill_manual(values = c("CONTROL" = "#F8766D", "OSA" = "#00BFC4")) +
  labs(x = "Menopause Status", y = "Count", fill = "Group") +
  theme_minimal(base_size = 14)

# 3. The research question

# Does OSA mediate the relationship between childhood trauma history and daytime
sleepiness?

# Step 2 - Pre-processing the data ----

# 1. Predictor/predicted variables for the analysis.
# Dependent Variable: daytime sleepiness as measured by ESS (Epworth Sleepiness Scale).
# Predictors: Childhood trauma as measured by ETISR-SF (General,Physical, Emotional and
Sexual).
# Mediating variable: OSA as measured by "Group".
# Variables to monitor: Age, BMI.

# 2. Process the variables

data <- data |>
  mutate(
    Trauma_total = ETISRSF_General + ETISRSF_Physical + ETISRSF_Emotional +
ETISRSF_Sexual,
    Group_binary = as.integer(Group == "OSA")
    )

# 3. Creating a function

calculate_summary <- function(df, vars) {
  df |>
    group_by(Group_binary) |>
    summarise(
      across(all_of(vars),
             list(mean = ~mean(.x, na.rm = TRUE),
                  sd = ~sd(.x, na.rm = TRUE)),
             .names = "{.col}_{.fn}"),
      .groups = "drop"
    )
  }

summary_stats <- calculate_summary(data, c("Age", "BMI", "Trauma_total", "ESS"))

print(summary_stats)
```

```
# 4. Using a function from a package we haven't learned about

skim(data[data$Group_binary == 1, c("Age", "BMI", "ESS", "Trauma_total")])

# Step 3 - Data analysis ----
# 1. a. Regression Analysis - Multiple Linear Regression

linear_model <- lm(ESS ~ Trauma_total + Age + BMI, data = data)
summary(linear_model)

# 2. a. Interpretation of the results

tidy(linear_model, conf.int = TRUE)

# Trauma_total is not significant (p = 0.171), but the direction is positive (B =
0.184).
# The model explains only 16.4% of the variance in ESS (R² = 0.164).
# Conclusion: The relationship between trauma and ESS is very weak and not significant.
# Here we already see that there is not a strong enough relationship to justify
mediation.

# 3. a. A graph

ggplot(data, aes(x = Trauma_total, y = ESS)) +
  geom_point(alpha = 0.6, color = "#00BFC4") +
  geom_smooth(method = "lm", se = TRUE, color = "#F8766D") +
  labs(title = "Effect of Trauma on ESS",
       x = "Trauma_total",
       y = "ESS") +
  theme_minimal()

# 1. b. Regression Analysis - Logistic Regression.

logistic_model <- glm(Group_binary ~ Trauma_total + Age + BMI, data = data, family =
binomial)
summary(logistic_model)

# 2. b. Interpretation of the results

logistic_results <- tidy(logistic_model, conf.int = TRUE, exponentiate = TRUE)
model_diagnostics <- glance(logistic_model)
list(results = logistic_results, diagnostics = model_diagnostics)

# Trauma_total is not significant (p = 0.781, B = 0.023).
# Age is almost significant (p = 0.056, B = 0.074).
# BMI is significant (p = 0.005, B = 0.258).
# Conclusion: Trauma does not predict OSA at all, age almost influences and BMI
strongly and significantly influences.
# Since trauma does not predict OSA, there cannot be mediation of OSA on the
relationship between trauma and ESS.

# 3. b. A graph

ggplot(data, aes(x = Trauma_total, y = Group_binary, color = as.factor(Group_binary)))
+
  geom_jitter(height = 0.05, alpha = 0.6) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = TRUE) +
  scale_color_manual(values = c("0" = "#F8766D", "1" = "#00BFC4"), labels =
c("Control", "OSA")) +
  labs(title = "Effect of Trauma on Probability of OSA",
       x = "Trauma_total",
       y = "Probability of OSA",
       color = "Group") +
  theme_minimal()

# 4. ROC graph for logistic regression analysis
```

```
data$predicted_prob <- predict(logistic_model, type = "response")

roc_curve <- roc(data$Group_binary, data$predicted_prob)

ggroc(roc_curve) +
  ggtitle("ROC Curve for Logistic Regression") +
  theme_minimal()
```