Data Analysis for stress & coping on Academic performance

Wesal megahed

1-1-2025

Introduction

The primary objective of this project is to explore how **demographic factors** (such as age, gender, family structure, etc.), **Stress** and **coping mechanisms**, and their combined impact influence **Academic performance**. The aim is to understand how these factors might contribute to or hinder a student's ability to succeed academically.

By gathering data on these elements, the study will uncover patterns that can offer insights to answer the following Questions:

- **Family Background**: How does family structure (e.g., single-parent families, dual-parent families) influence stress and academic performance?
- **Educational Level**: Does a student's level of education (high school, undergraduate, etc.) impact their stress levels and coping strategies?
- **Access to Resources**: How do factors like digital access contribute to a student's success?
- **Stress Levels**: What types of stress do students experience, and how do they perceive their stress level in relation to their studies?
- **Coping Strategies**: What coping mechanisms are most commonly employed by students, and how do these mechanisms influence their academic outcomes?
- **Interaction of Factors**: How do demographic factors (e.g., family background, age) interact with stress levels and coping strategies to influence academic performance?
- **Academic Success or Failure**: How do these combined factors contribute to a student's ability to succeed academically, and what patterns can be identified?

Data Analysis

Below, we will present a summary of the data and apply statistical analysis techniques such as confidence intervals, hypothesis testing, regression analysis to interpret the data.

```
options(warn=-1)

library(readxl)
library(ggplot2)    # For visualizations
library(dplyr)      # For data manipulation
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)      # For descriptive statistics
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(janitor)    # clean names
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(knitr)      # tables

# Load the dataset(this is example for group 8)
data <- read_excel("8 Demographics, stress & coping and Academic performance.xlsx")

# Clean the dataset
data <- data %>% clean_names()
names(data)[names(data) == "academic_performance_questions"] <- "GPA"
names(data)[names(data) == "on_a_scale_from_1_lowest_to_10_highest_how_would_you_rate_your_general_stress_le
vel_over_the_past_month"] <- "Stress_levels"
names(data)[names(data) == "lifestyle_factors_urban_vs_rural_living"] <- "lifestyle"
names(data)[names(data) == "impact_of_stress_on_grades" ] <- "stress_impact"
colnames(data)
```

```
##  [1] "id"                      "start_time"
##  [3] "completion_time"         "email"
##  [5] "name"                    "gender"
##  [7] "age"                     "parents_education_level"
##  [9] "parents_education_level1" "family_structure"
## [11] "lifestyle"               "digital_access"
## [13] "Stress_levels"           "stress_sources"
## [15] "coping_strategies"       "frequency_of_coping"
## [17] "effectiveness_of_coping"  "GPA"
## [19] "time_spent_studying"     "stress_impact"
```

```
#----------------------- Statistical Summaries ------------------#

## Central tendency measures
summary(data)
```

```
##        id          start_time
##  Min.   : 1.00   Min.   :2024-11-09 14:47:04.00
##  1st Qu.:13.75   1st Qu.:2024-11-10 14:09:30.00
##  Median :26.50   Median :2024-11-11 13:20:27.00
##  Mean   :26.50   Mean   :2024-11-11 05:07:29.98
##  3rd Qu.:39.25   3rd Qu.:2024-11-11 15:20:45.00
##  Max.   :52.00   Max.   :2024-11-12 08:57:18.00
##  completion_time                     email             name
##  Min.   :2024-11-09 14:54:08.00   Length:52          Mode:logical
##  1st Qu.:2024-11-10 14:12:17.25   Class :character   NA's:52
##  Median :2024-11-11 13:22:21.00   Mode  :character
##  Mean   :2024-11-11 05:11:51.69
##  3rd Qu.:2024-11-11 15:34:19.00
##  Max.   :2024-11-12 09:01:46.00
##     gender               age        parents_education_level
##  Length:52          Min.   :11.00   Length:52
##  Class :character   1st Qu.:18.00   Class :character
##  Mode  :character   Median :19.00   Mode  :character
##                     Mean   :22.67
##                     3rd Qu.:21.25
##                     Max.   :66.00
##  parents_education_level1 family_structure    lifestyle
##  Length:52                Length:52          Length:52
##  Class :character         Class :character   Class :character
##  Mode  :character         Mode  :character   Mode  :character
##
##
##
##  digital_access     Stress_levels     stress_sources     coping_strategies
##  Length:52          Min.   :  2.000   Length:52          Length:52
##  Class :character   1st Qu.:  6.000   Class :character   Class :character
##  Mode  :character   Median :  8.000   Mode  :character   Mode  :character
##                     Mean   :  9.135
##                     3rd Qu.:  9.250
##                     Max.   :100.000
##  frequency_of_coping effectiveness_of_coping      GPA
##  Min.   :   0.00     Length:52               Min.   :  0.00
##  1st Qu.:   3.00     Class :character        1st Qu.:  5.00
##  Median :   7.00     Mode  :character        Median :  7.00
##  Mean   :  52.47                             Mean   : 18.81
##  3rd Qu.:  10.00                             3rd Qu.: 10.00
##  Max.   :2009.00                             Max.   :115.00
##  time_spent_studying stress_impact
##  Min.   :   0.00     Length:52
##  1st Qu.:   5.00     Class :character
##  Median :   7.00     Mode  :character
##  Mean   :  59.88
##  3rd Qu.:  21.75
##  Max.   :2007.00
```

```
# Perform statistical summaries
summary_table <- data %>%
  summarise(
    `Total Records` = n(),
    `Age (Mean)` = mean(age),
    `Age (Median)` = median(age),
    `Coping Frequency (Mean)` = mean(frequency_of_coping),
    `Coping Frequency (Median)` = median(frequency_of_coping),
    `Time Studying (Mean)` = mean(time_spent_studying),
    `Time Studying (Median)` = median(time_spent_studying),
    `Academic Performance (Mean)` = mean(GPA),
    `Academic Performance (Median)` = median(GPA)
  )

# Convert to a professional table format
kable(summary_table, caption = "Statistical Summary of the Dataset", digits = 2)
```

Statistical Summary of the Dataset

Total Records	Age (Mean)	Age (Median)	Coping Frequency (Mean)	Coping Frequency (Median)	Time Studying (Mean)	Time Studying (Median)	Academic Performance (Mean)	Academic Performance (Median)
52	22.67	19	52.47	7	59.88	7	18.81	7

```
# Variability measures
variability_age <- data %>%
  summarise(
    `Total Records` = n(),
    `Mean Age` = mean(age),
    `Variance Age` = var(age),
    `Standard Deviation Age` = sd(age),
    `Min Age` = min(age),
    `Max Age` = max(age),
    `Range Age` = diff(range(age)),
    `Q1 Age` = quantile(age, 0.25),
    `Median Age` = median(age),
    `Q3 Age` = quantile(age, 0.75),
    `IQR Age` = IQR(age))
kable(variability_age,caption = "Variability Measures of the Dataset",digits = 2)
```

Variability Measures of the Dataset

Total Records	Mean Age	Variance Age	Standard Deviation Age	Min Age	Max Age	Range Age	Q1 Age	Median Age	Q3 Age	IQR Age
52	22.67	113.32	10.65	11	66	55	18	19	21.25	3.25

```
variability_gpa <- data %>%
  summarise(
`Mean GPA` = mean(GPA),
`Variance GPA` = var(GPA),
`Standard Deviation GPA` = sd(GPA),
`Min GPA` = min(GPA),
`Max GPA` = max(GPA),
`Range GPA` = diff(range(GPA)),
`Q1 GPA` = quantile(GPA, 0.25),
`Median GPA` = median(GPA),
`Q3 GPA` = quantile(GPA, 0.75),
`IQR GPA` = IQR(GPA))
kable(variability_gpa,caption = "Variability Measures of the Dataset",digits = 2)
```

Variability Measures of the Dataset

Mean GPA	Variance GPA	Standard Deviation GPA	Min GPA	Max GPA	Range GPA	Q1 GPA	Median GPA	Q3 GPA	IQR GPA
18.81	989.57	31.46	0	115	115	5	7	10	5

```
variability_Time_studying <- data %>%
  summarise(
    `Mean Time Spent Studying` = mean(time_spent_studying),
`Variance Time Spent Studying` = var(time_spent_studying),
`Standard Deviation Time Spent Studying` = sd(time_spent_studying),
`Min Time Spent Studying` = min(time_spent_studying),
`Max Time Spent Studying` = max(time_spent_studying),
`Range Time Spent Studying` = diff(range(time_spent_studying)),
`Q1 Time Spent Studying` = quantile(time_spent_studying, 0.25),
`Median Time Spent Studying` = median(time_spent_studying),
`Q3 Time Spent Studying` = quantile(time_spent_studying, 0.75),
`IQR Time Spent Studying` = IQR(time_spent_studying))
kable(variability_Time_studying,caption = "Variability Measures of the Dataset",digits = 2)
```

Variability Measures of the Dataset

Mean Time Spent Studying	Variance Time Spent Studying	Standard Deviation Time Spent Studying	Min Time Spent Studying	Max Time Spent Studying	Range Time Spent Studying	Q1 Time Spent Studying	Median Time Spent Studying	Q3 Time Spent Studying	IQR Time Spent Studying
59.88	77886.81	279.08	0	2007	2007	5	7	21.75	16.75

```
variability_Coping_hour <- data %>%
  summarise(
`Mean Frequency of Coping` = mean(frequency_of_coping),
`Variance Frequency of Coping` = var(frequency_of_coping),
`Standard Deviation Frequency of Coping` = sd(frequency_of_coping),
`Min Frequency of Coping` = min(frequency_of_coping),
`Max Frequency of Coping` = max(frequency_of_coping),
`Range Frequency of Coping` = diff(range(frequency_of_coping)),
`Q1 Frequency of Coping` = quantile(frequency_of_coping, 0.25),
`Median Frequency of Coping` = median(frequency_of_coping),
`Q3 Frequency of Coping` = quantile(frequency_of_coping, 0.75),
`IQR Frequency of Coping` = IQR(frequency_of_coping))
kable(variability_Coping_hour,caption = "Variability Measures of the Dataset",digits = 2)
```

Variability Measures of the Dataset

Mean Frequency of Coping	Variance Frequency of Coping	Standard Deviation Frequency of Coping	Min Frequency of Coping	Max Frequency of Coping	Range Frequency of Coping	Q1 Frequency of Coping	Median Frequency of Coping	Q3 Frequency of Coping	IQR Frequency of Coping
52.47	77223.1	277.89	0	2009	2009	3	7	10	7

```
table(data$gender)
```

```
##
## Female   Male
##     17     35
```

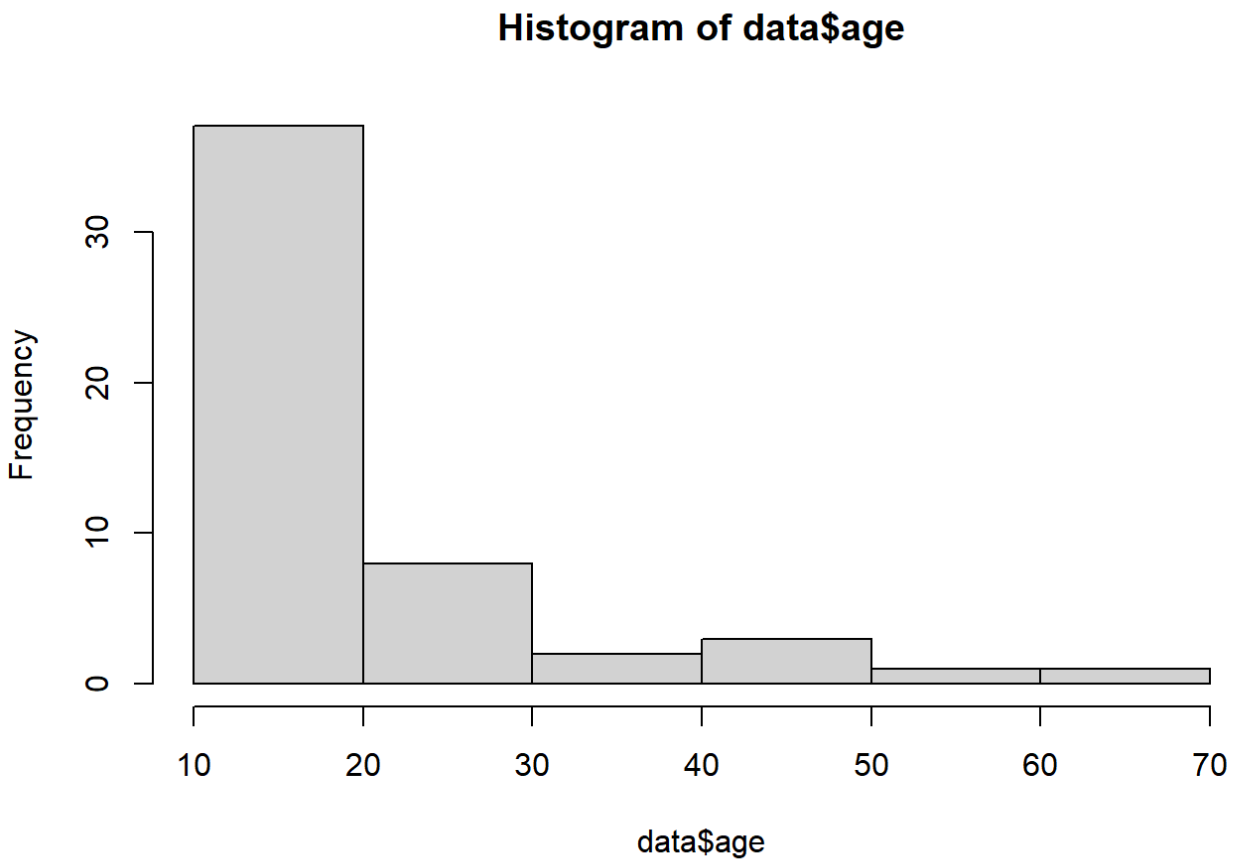
```
table(data$parents_education_level)
```

```
##
##         No formal education              Primary school
##                           1                           4
##            Secondary school University degree or higher
##                           5                          42
```

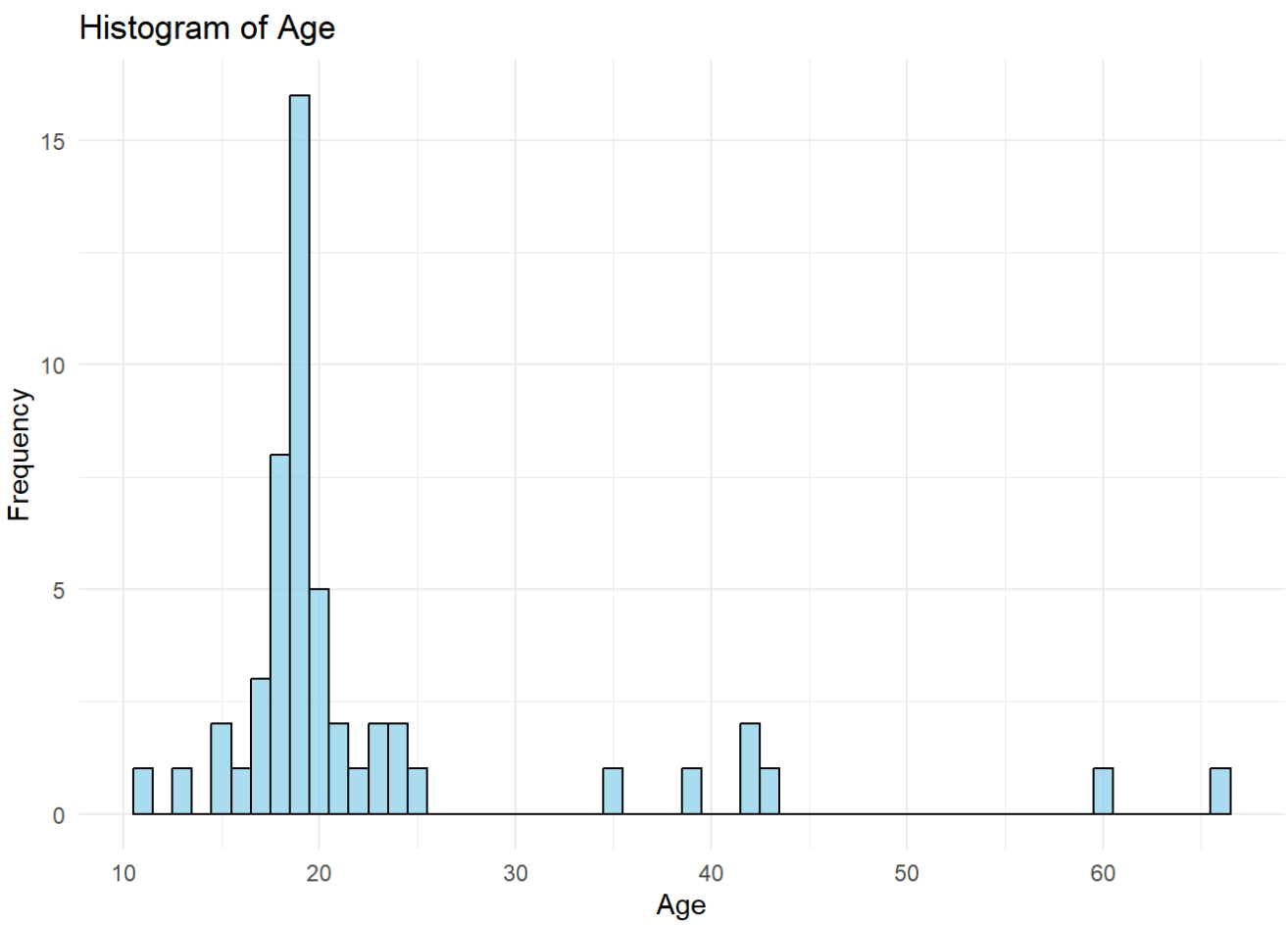
```
table(data$gender)# ------------------- Data Visualization -------------- #
```

```
##
## Female   Male
##     17     35
```

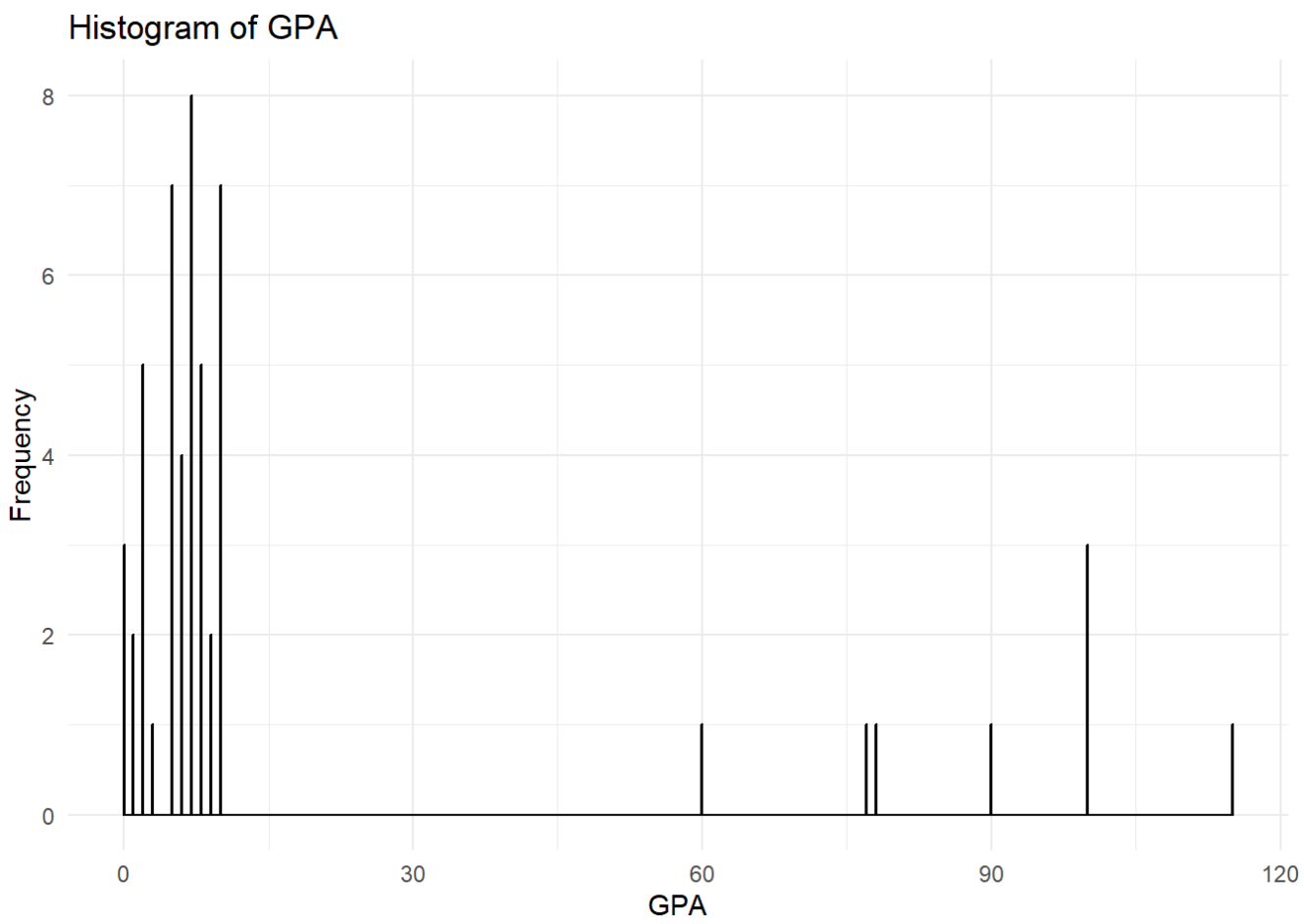
```
# 1. Histogram for Age
hist(data$age)
```



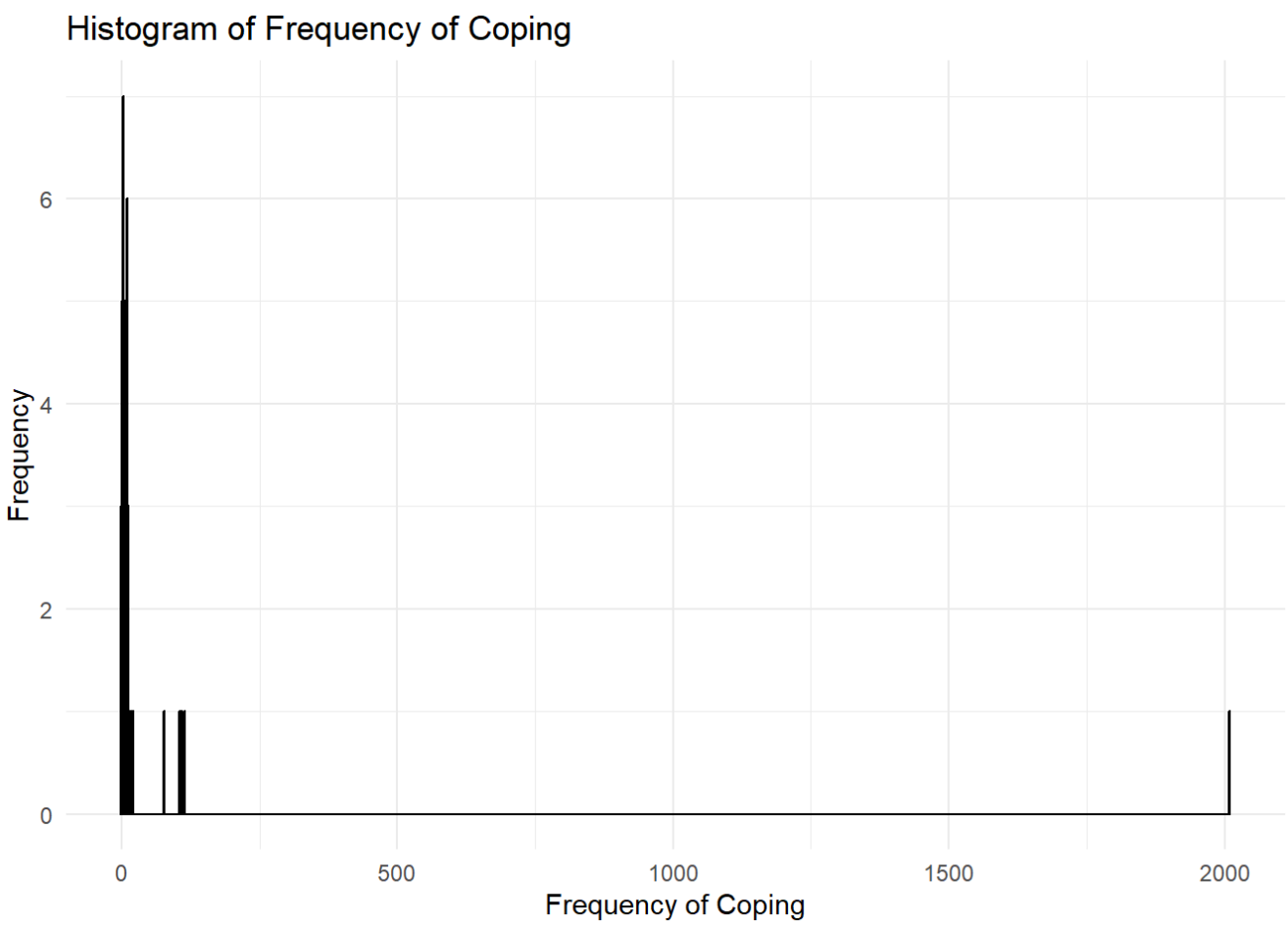

```
ggplot(data, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Age", x = "Age", y = "Frequency") +
  theme_minimal()
```



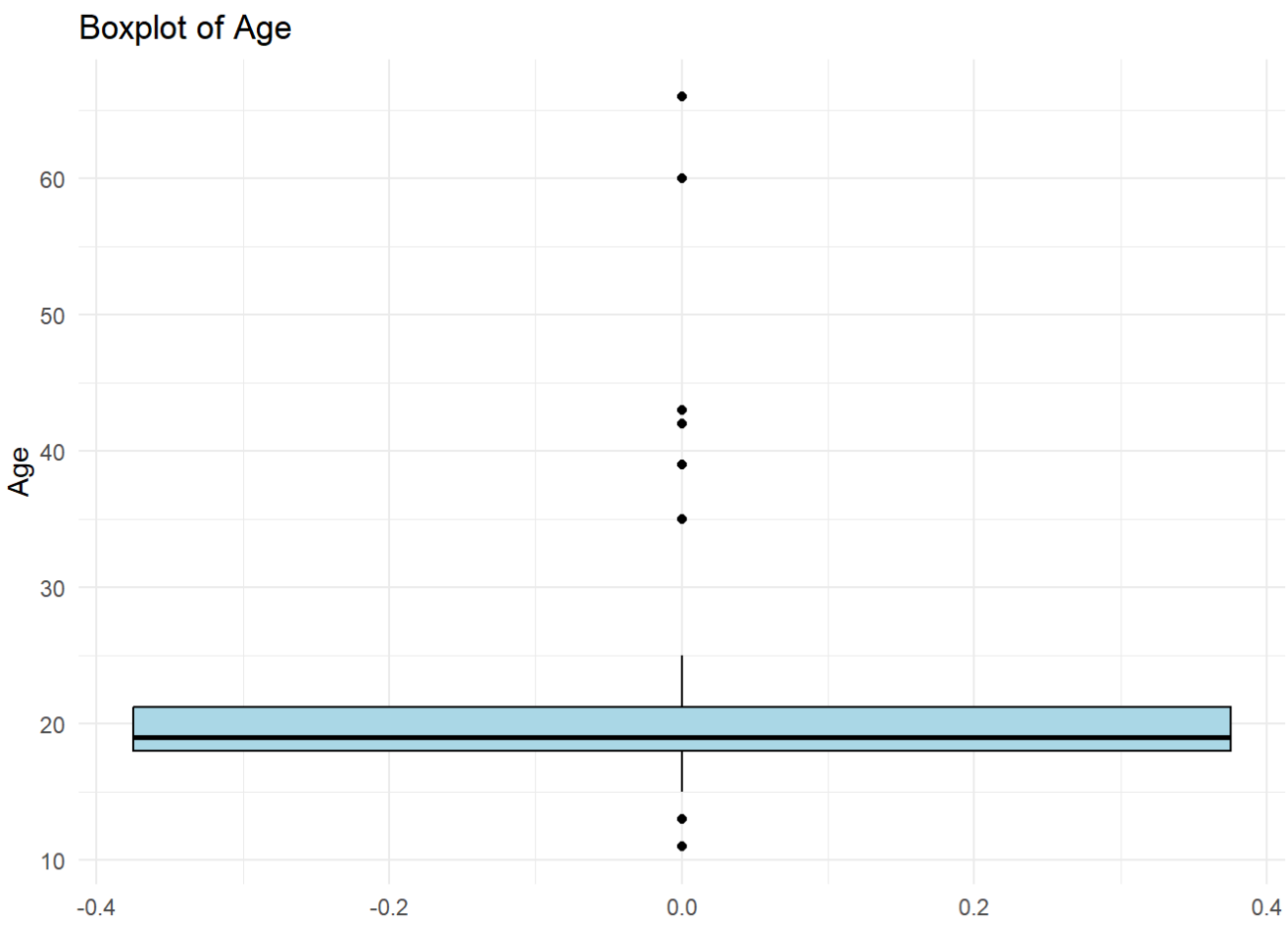

```
# 2. Histogram for GPA
ggplot(data, aes(x = GPA)) +
  geom_histogram(binwidth = 0.1, fill = "orange", color = "black", alpha = 0.7) +
  labs(title = "Histogram of GPA", x = "GPA", y = "Frequency") +
  theme_minimal()
```



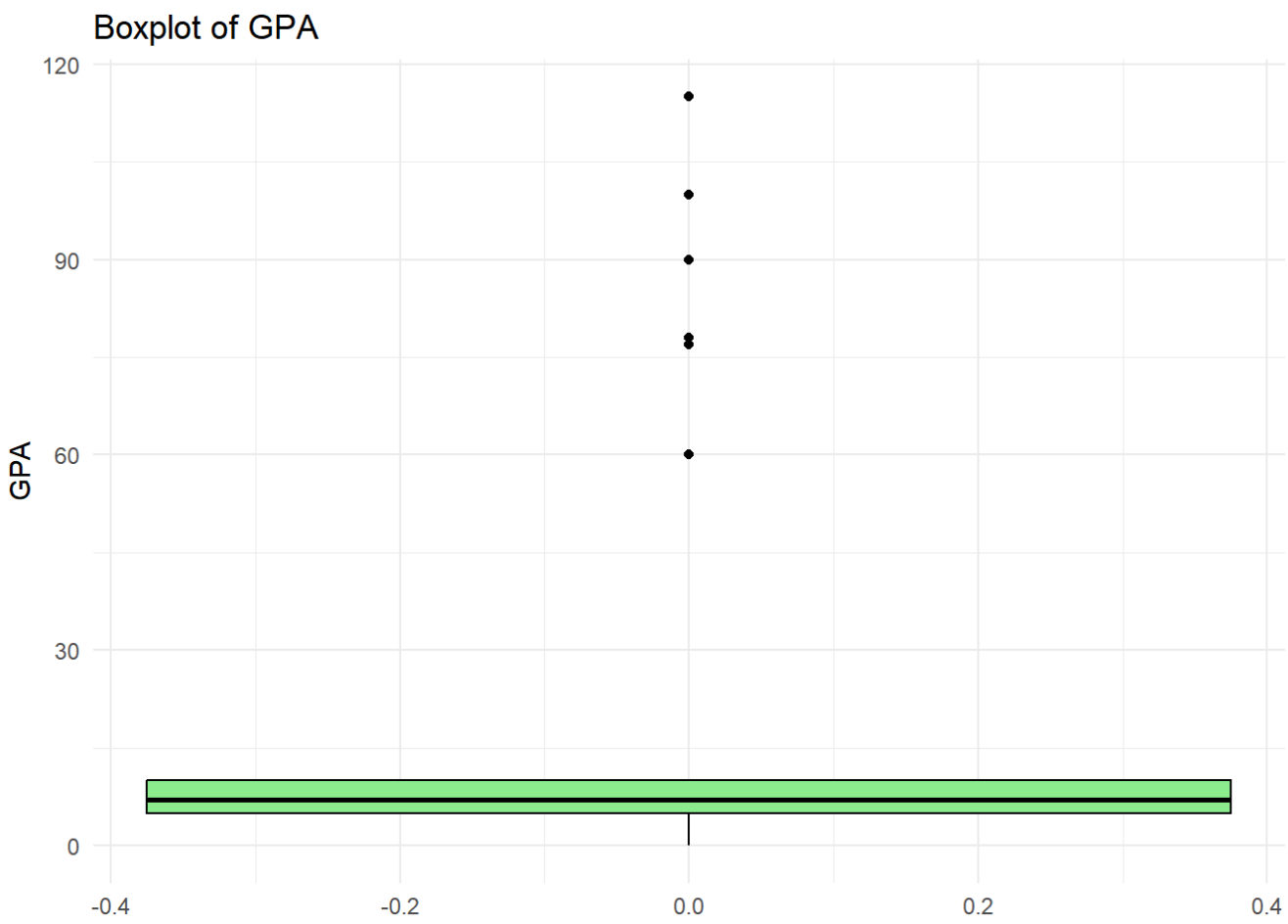

```
# 3. Histogram for Time Spent Studying
ggplot(data, aes(x = frequency_of_coping)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Frequency of Coping", x = "Frequency of Coping", y = "Frequency") +
  theme_minimal()
```



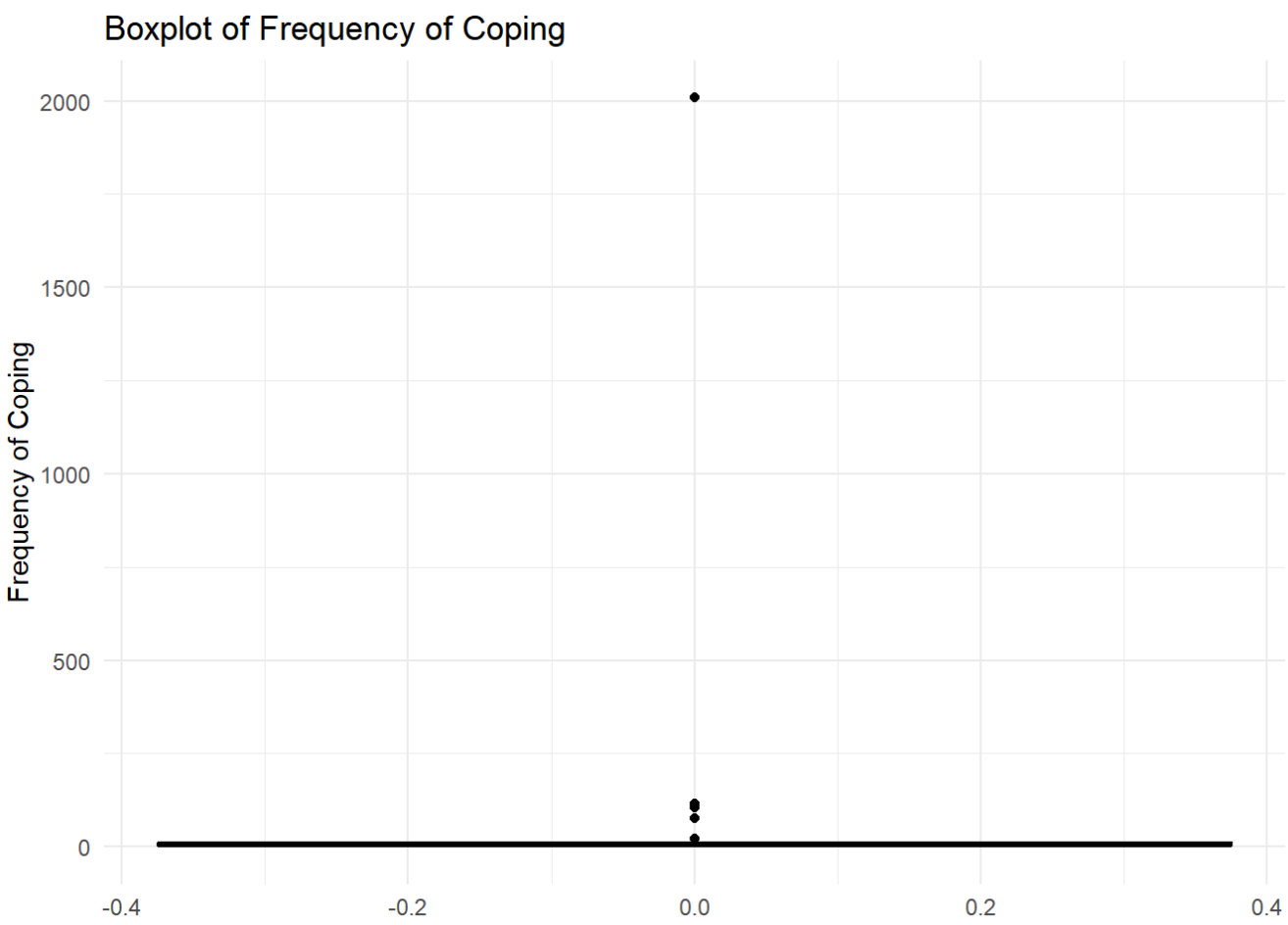

```
# 3. Boxplot for Age
ggplot(data, aes(y = age)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Boxplot of Age", y = "Age") +
  theme_minimal()
```



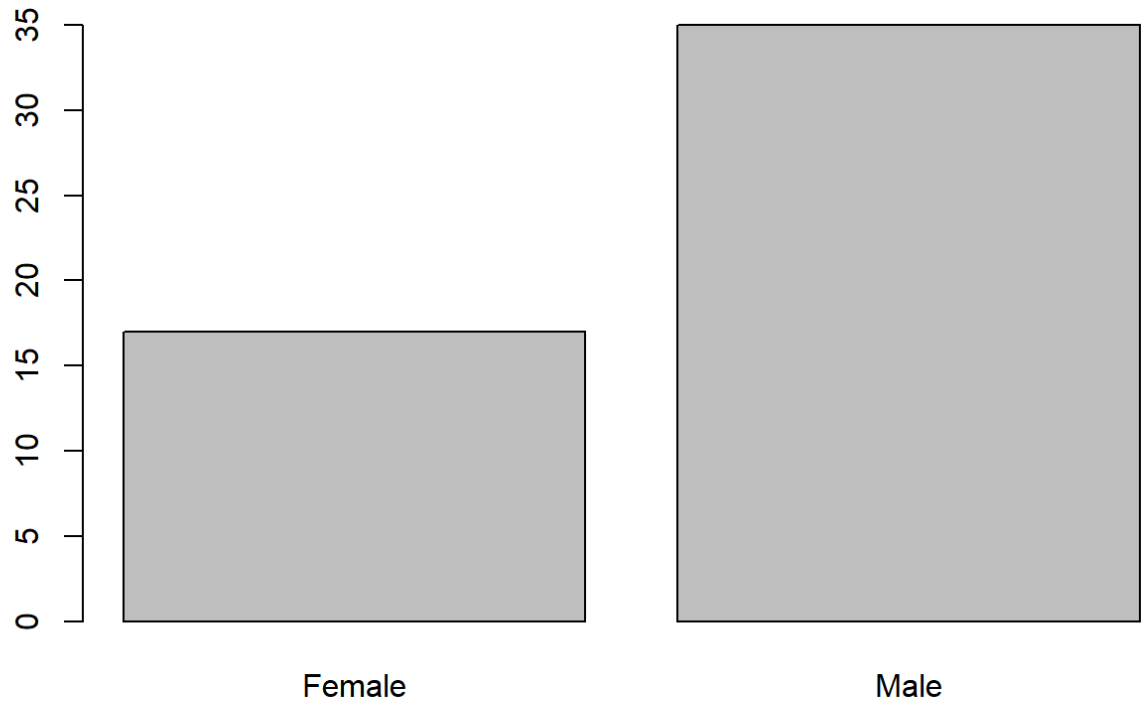

```
# 4. Boxplot for GPA
ggplot(data, aes(y = GPA)) +
  geom_boxplot(fill = "lightgreen", color = "black") +
  labs(title = "Boxplot of GPA", y = "GPA") +
  theme_minimal()
```



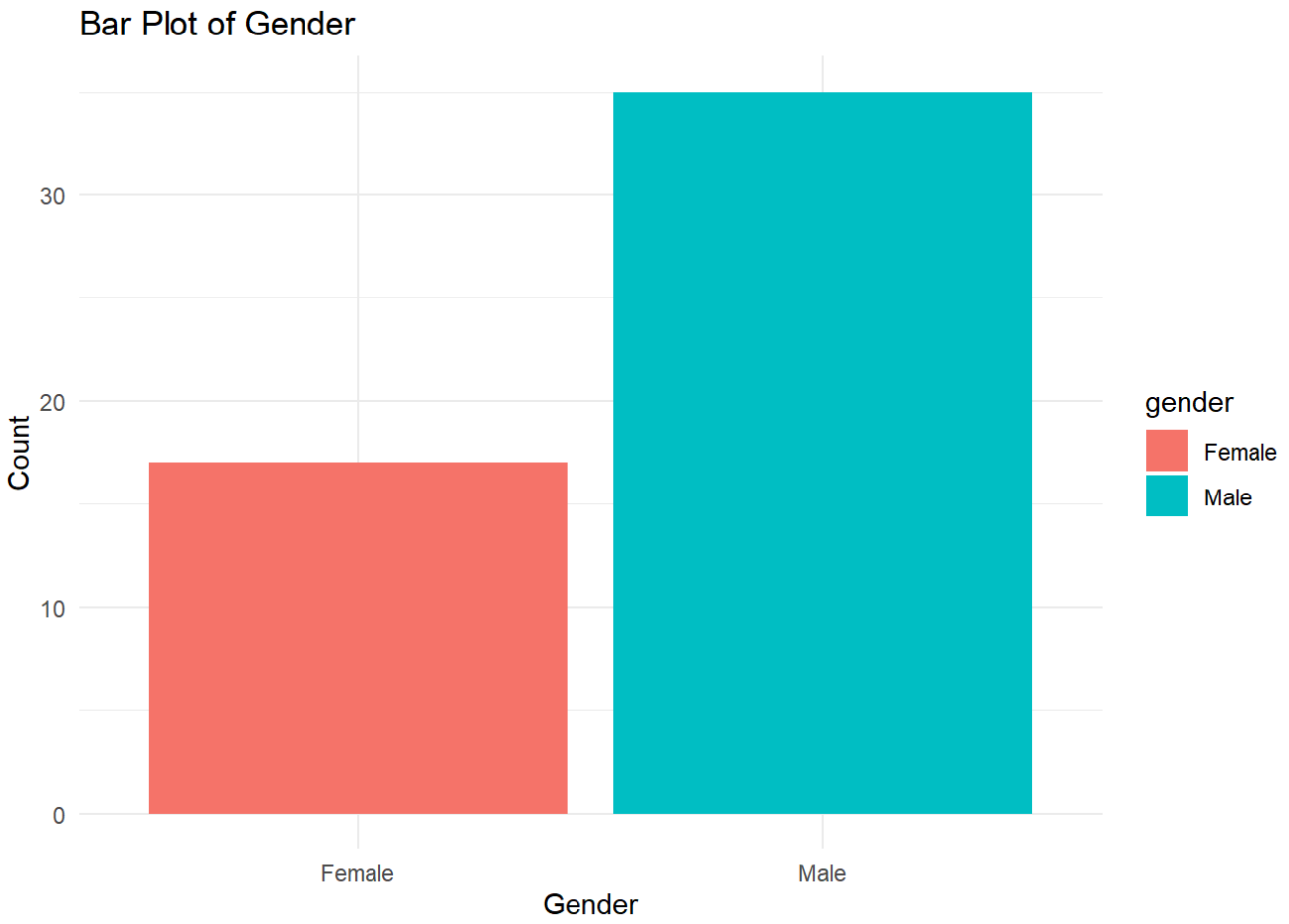

```
# 5. Boxplot for Frequency of Coping
ggplot(data, aes(y = frequency_of_coping)) +
  geom_boxplot(fill = "lightcoral", color = "black") +
  labs(title = "Boxplot of Frequency of Coping", y = "Frequency of Coping") +
  theme_minimal()
```



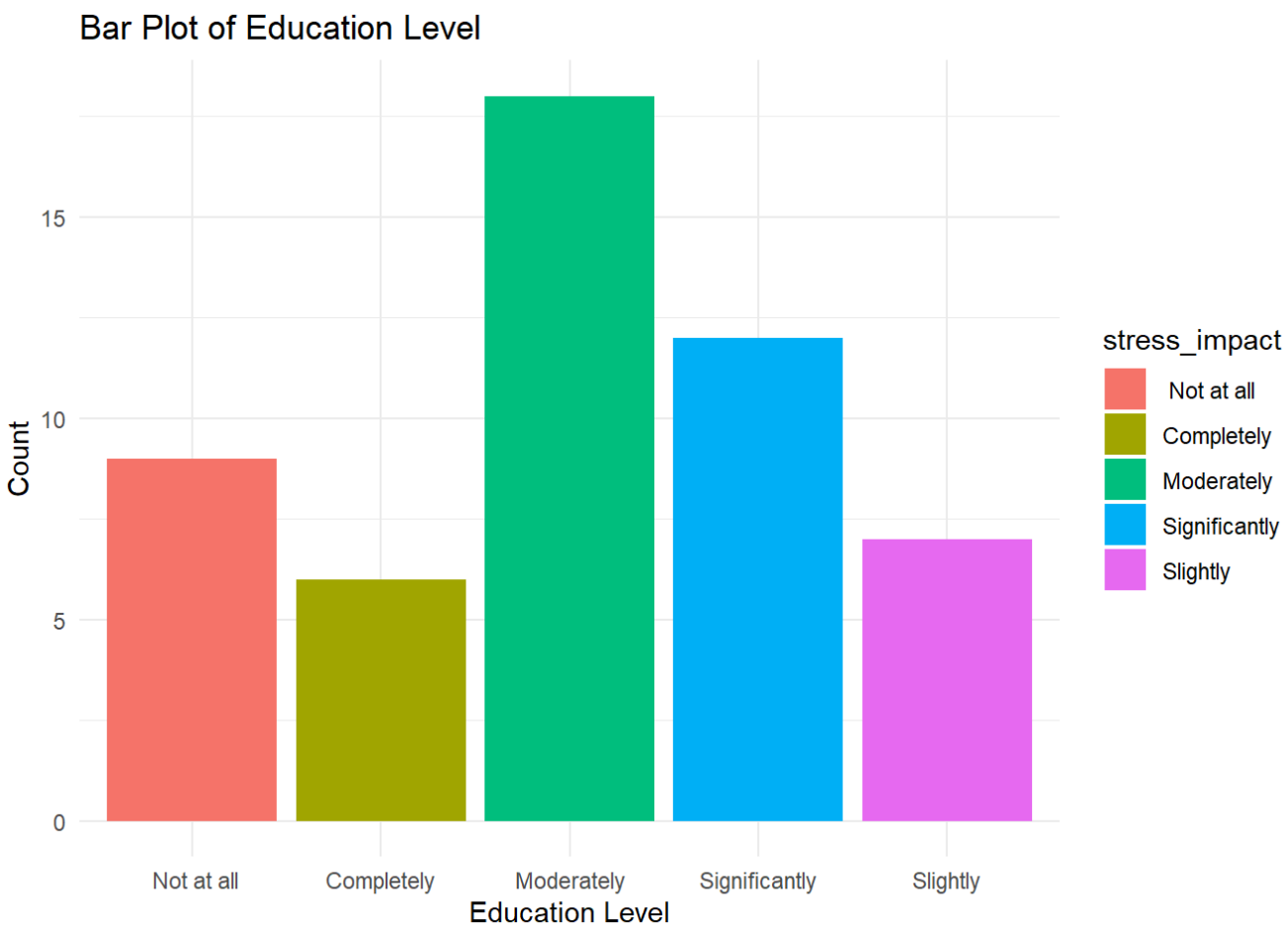

```
# 7. Bar Plot for Gender
barplot(table(data$gender))
```



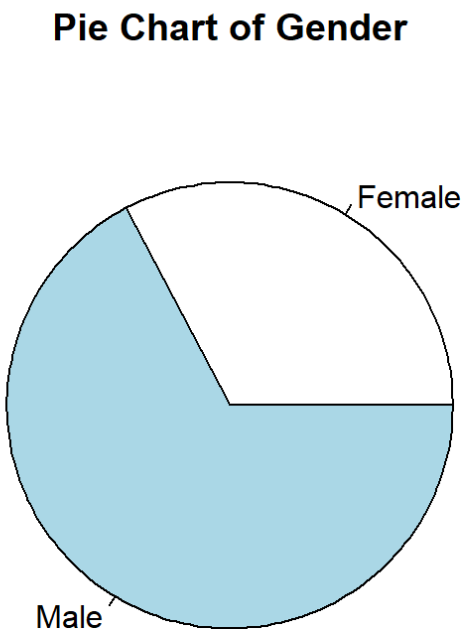
```
ggplot(data, aes(x = gender, fill = gender)) +
  geom_bar() +
  labs(title = "Bar Plot of Gender", x = "Gender", y = "Count") +
  theme_minimal()
```



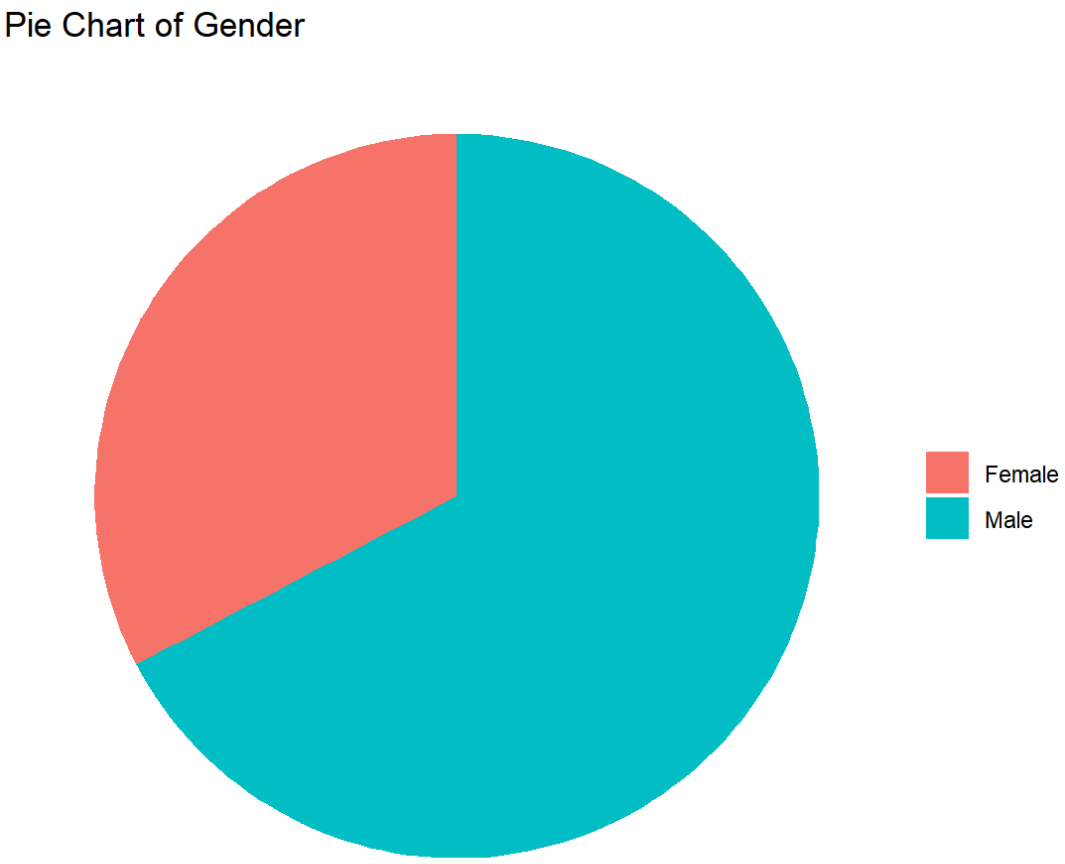

```
# 8. Bar Plot for impact of stress on grades
ggplot(data, aes(x = stress_impact, fill = stress_impact)) +
  geom_bar() +
  labs(title = "Bar Plot of Education Level", x = "Education Level", y = "Count") +
  theme_minimal()
```




```
pie(table(data$gender), main = "Pie Chart of Gender")
```

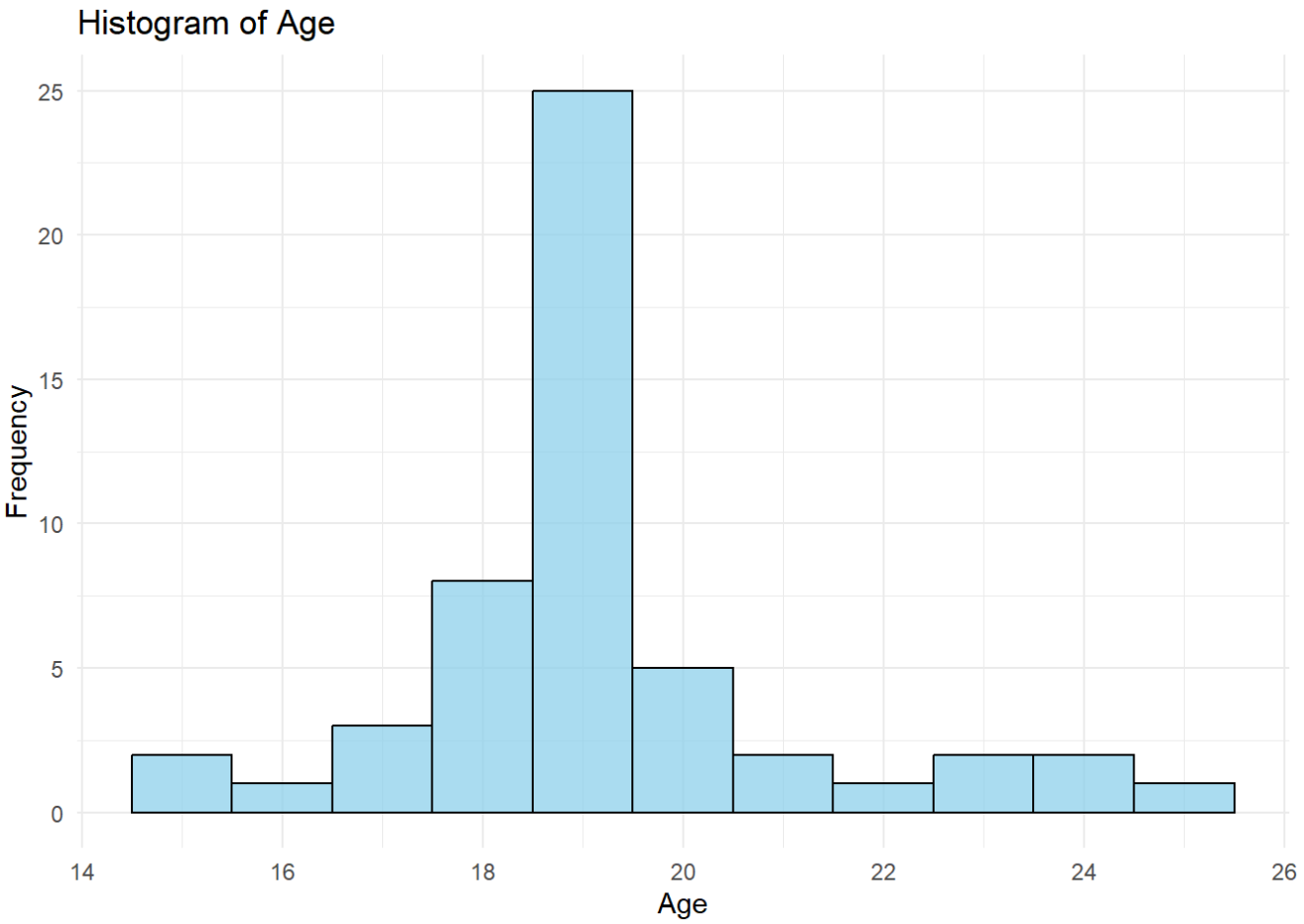


```
ggplot(data, aes(x = "", fill = gender)) +
  geom_bar(stat = "count", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Pie Chart of Gender") +
  theme_void() +
  theme(legend.title = element_blank())
```

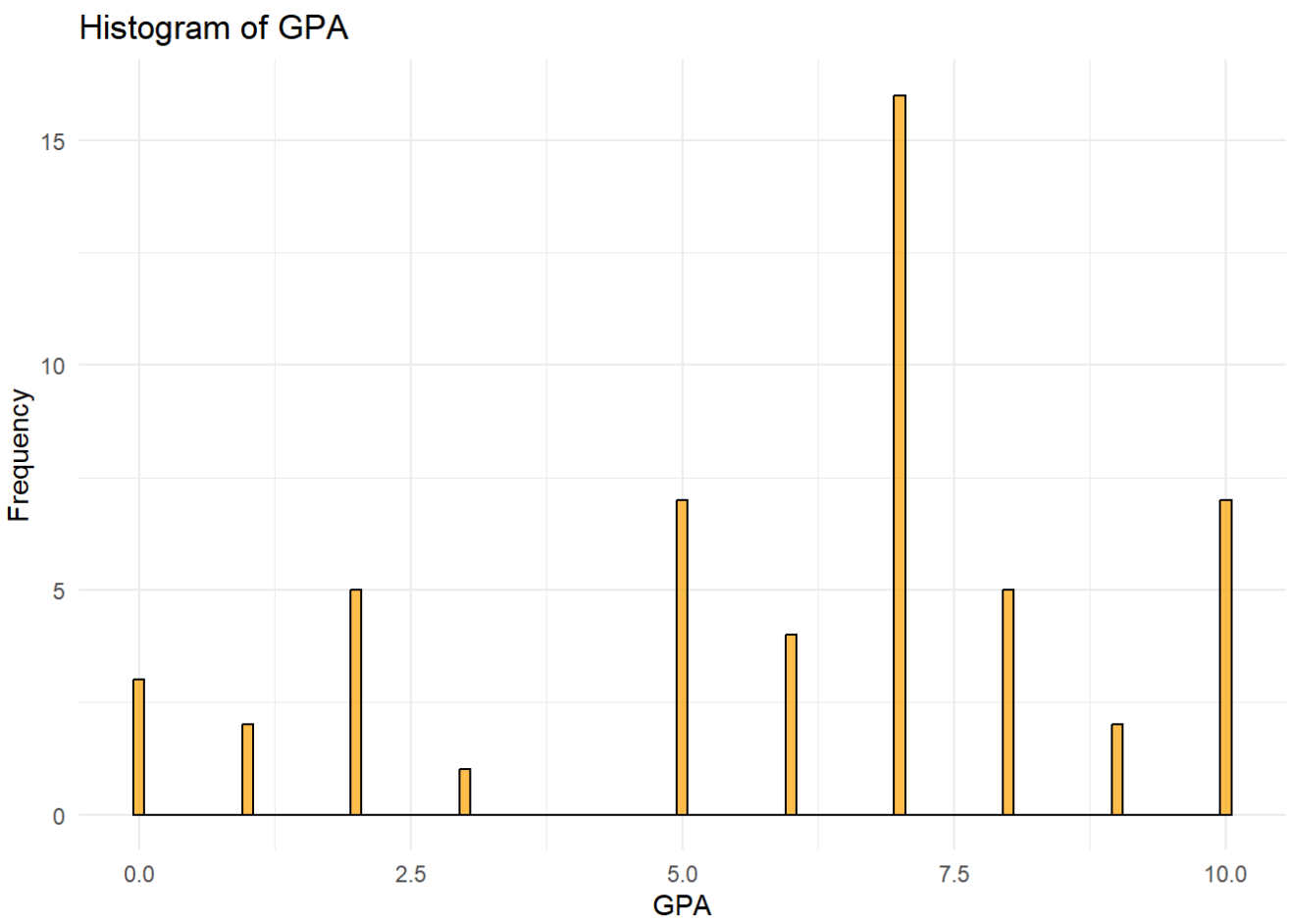



```
#---------------------- Detect and replace outliers --------------------#
# Define a function to detect and replace outliers

replace_outliers_with_median <- function(column) {
  Q1 <- quantile(column, 0.25, na.rm = TRUE)
  Q3 <- quantile(column, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  column <- ifelse(column < lower_bound | column > upper_bound,
                   median(column, na.rm = TRUE),
                   column)
  return(column)}

# Apply the function to numeric columns in your dataset using mutate function
data_cleaned <- data %>%
  mutate(across(where(is.numeric), replace_outliers_with_median))
```

```r
# 2. Histogram for Age
ggplot(data_cleaned, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Age", x = "Age", y = "Frequency") +
  theme_minimal()
```




```r
# 2. Histogram for GPA
ggplot(data_cleaned, aes(x = GPA)) +
  geom_histogram(binwidth = 0.1, fill = "orange", color = "black", alpha = 0.7) +
  labs(title = "Histogram of GPA", x = "GPA", y = "Frequency") +
  theme_minimal()
```




```
#####
# Proportions (%) of each category: Example for gender
(table(data$gender)/length(data$gender))*100
```

```
##
##   Female     Male
## 32.69231 67.30769
```

```
# Or use probability table
prop.table(table(data$gender)) * 100
```

```
##
##   Female     Male
## 32.69231 67.30769
```

```
# to round up to one decimal
round(prop.table(table(data$gender)) * 100, 1)
```

```
##
## Female   Male
##   32.7   67.3
```

```
# Create a contingency table (Count)
table(data_cleaned$gender, data_cleaned$digital_access)
```

```
##
##          Excellent access Good access Limited access Moderate access
##   Female                3           8              2               4
##   Male                 14          14              1               6
```

```
table(data_cleaned$family_structure, data_cleaned$Stress_levels)
```

```
##
##                            2 3 4 5 6 7 8 9 10
##    Extended family household 2 0 1 0 2 2 6 3  3
##    Single-parent household   1 0 0 0 0 1 2 1  1
##   I Live alone               0 0 0 0 0 0 0 0  1
##   Separated parents          0 0 0 1 0 0 0 0  0
##   Two-parent household       0 1 1 2 5 4 5 0  7
```

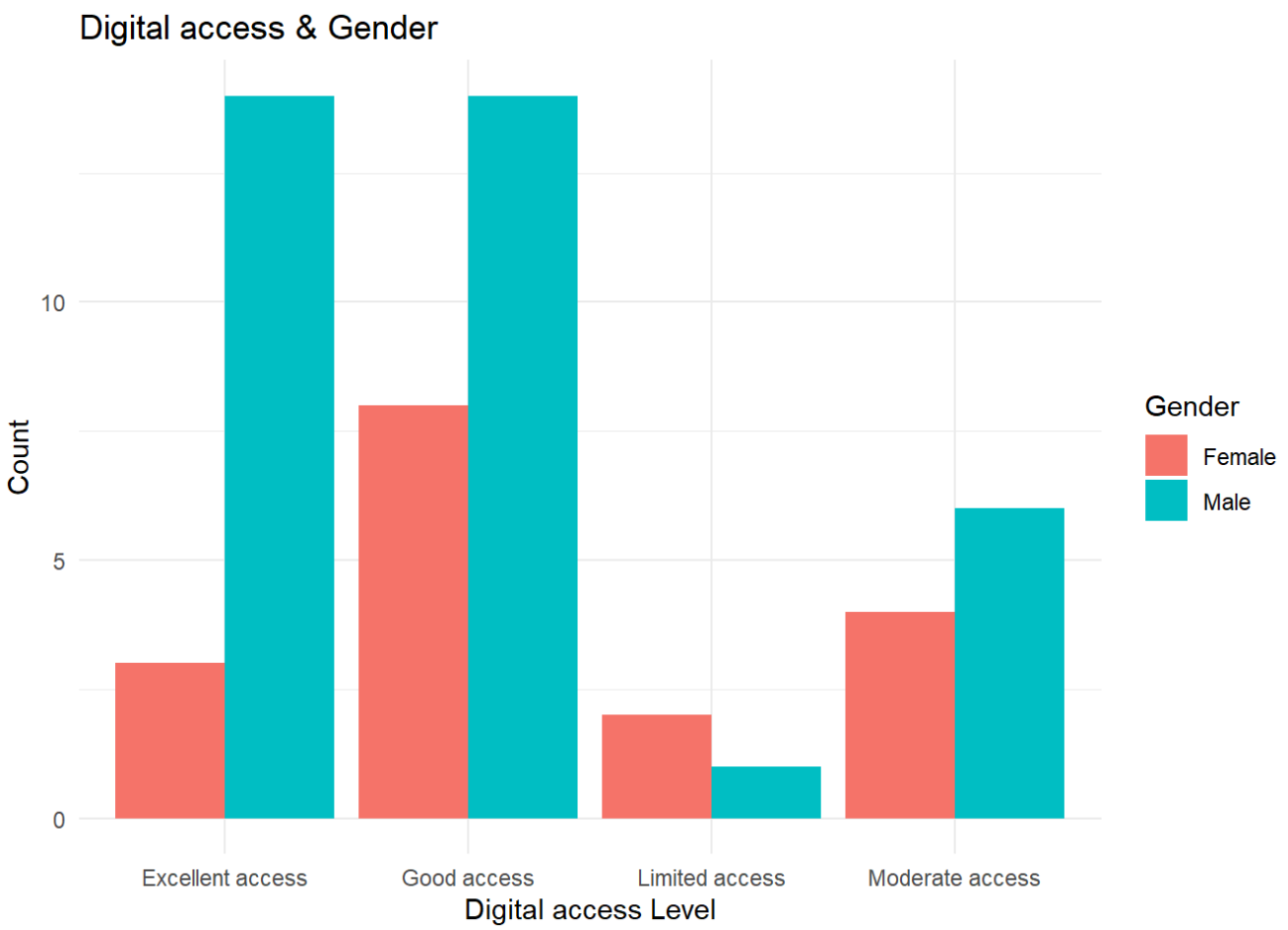
```
table(data_cleaned$parents_education_level,data_cleaned$parents_education_level1)
```

```
##
##                             No formal education Primary school
##   No formal education                         1              0
##   Primary school                              0              1
##   Secondary school                            1              0
##   University degree or higher                 0              2
##
##                             Secondary school University degree or higher
##   No formal education                       0                          0
##   Primary school                            2                          1
##   Secondary school                          2                          2
##   University degree or higher               2                         38
```

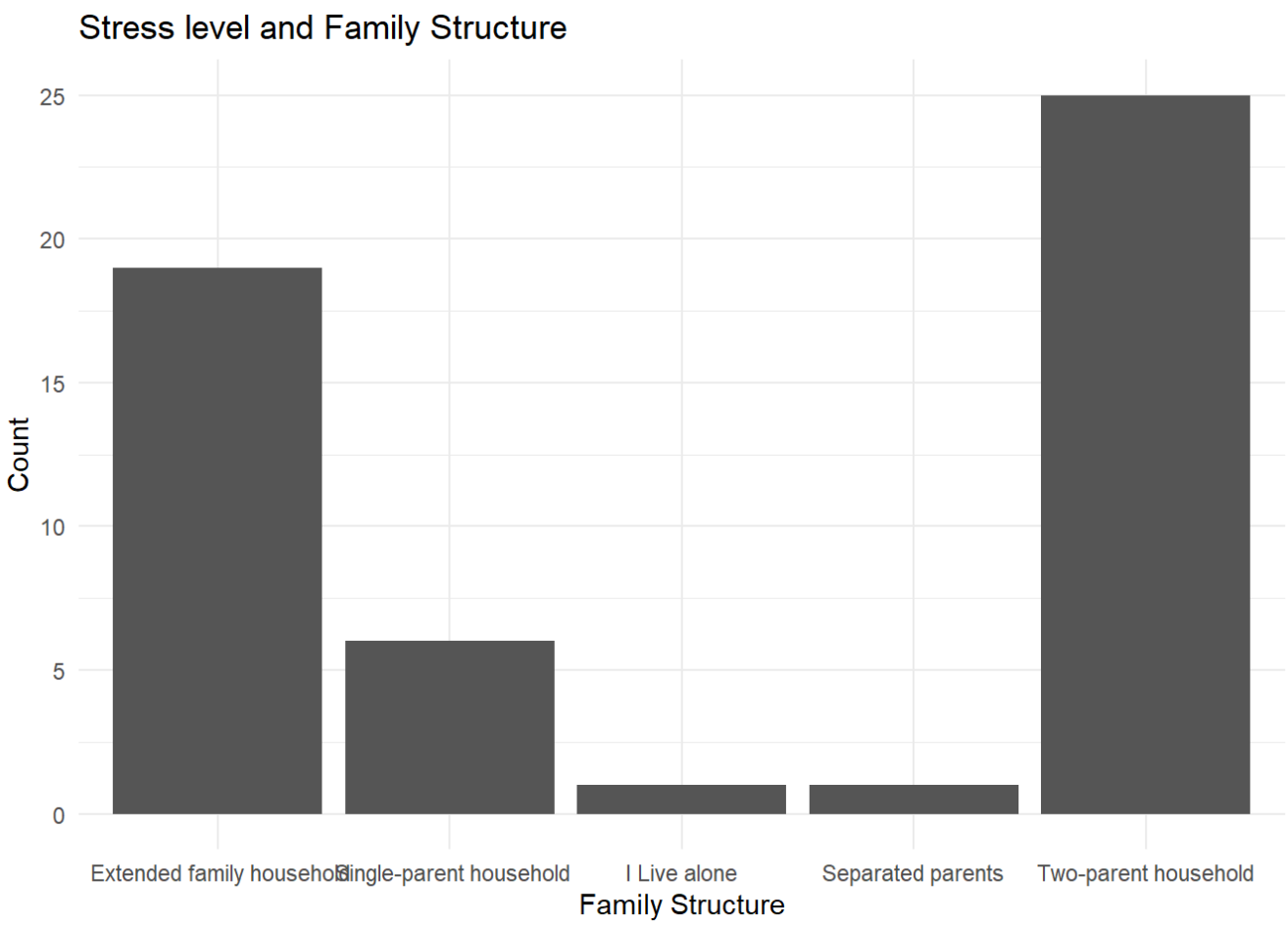
```
#possible comment (Most parents with a University Degree or Higher tend to have partners or family members w
ith similar education levels )

# Create a contingency table (proportion)
round(prop.table(table(data_cleaned$gender, data_cleaned$digital_access)) * 100, 1)
```

```
##
##          Excellent access Good access Limited access Moderate access
##   Female              5.8        15.4            3.8             7.7
##   Male               26.9        26.9            1.9            11.5
```

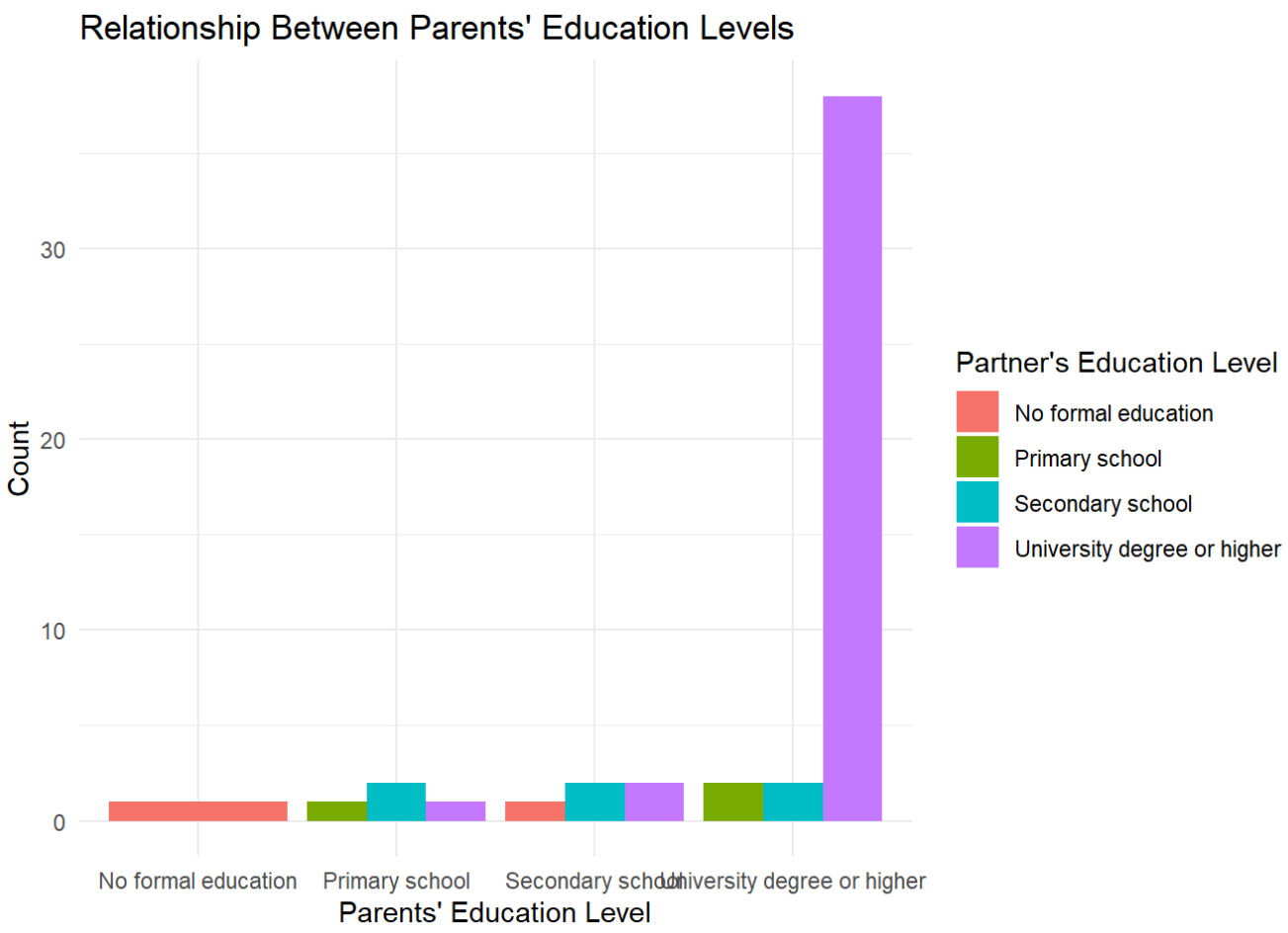
```
# Plot clustered bar chart for two qualitative variables
ggplot(data_cleaned, aes(x = digital_access, fill = gender)) +
  geom_bar(position = "dodge") +
  labs(title = "Digital access & Gender",
       x = "Digital access Level",
       y = "Count",
       fill = "Gender") +
  theme_minimal()
```



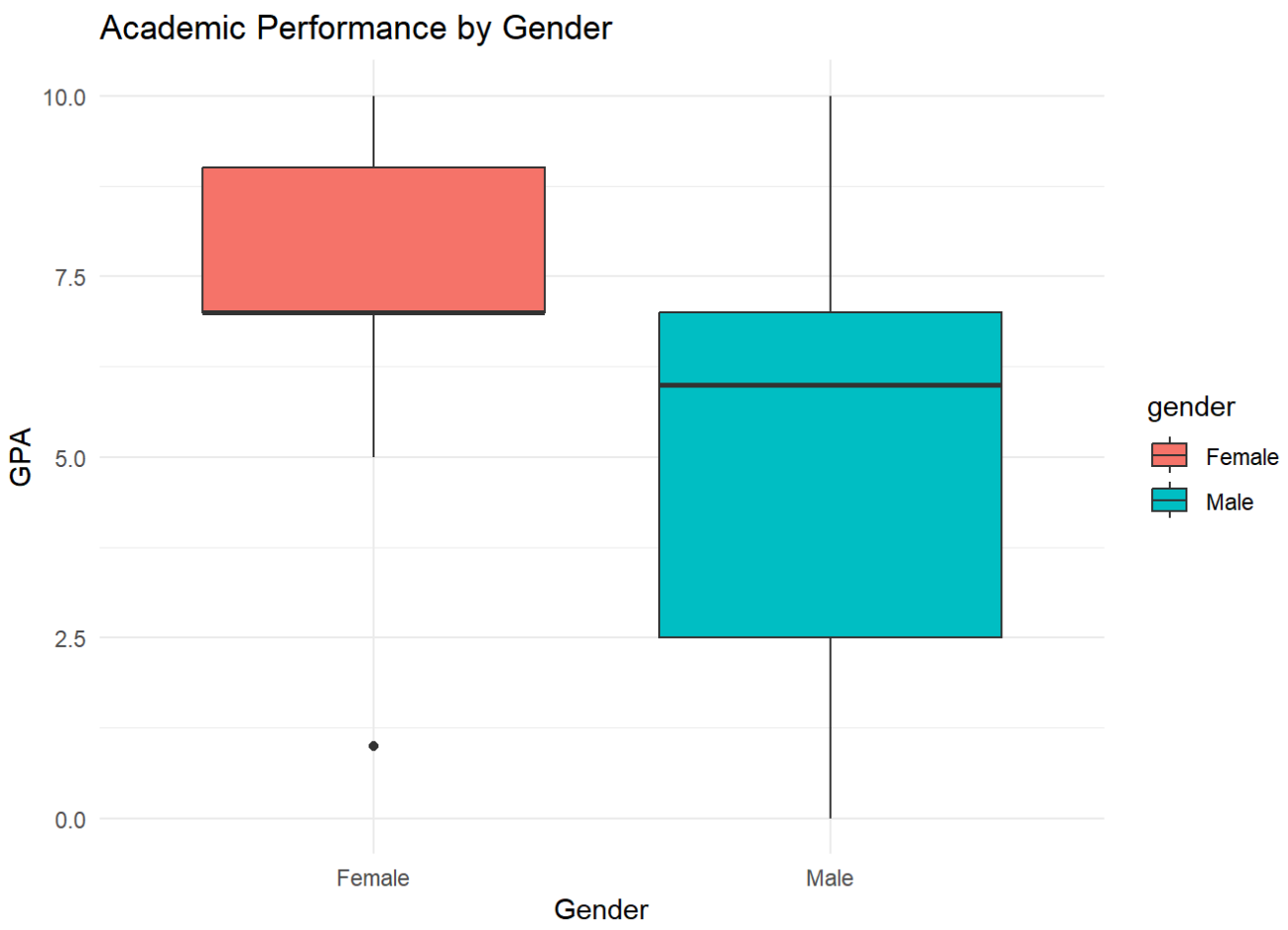

```
ggplot(data_cleaned, aes(x = family_structure, fill = Stress_levels)) +
  geom_bar(position = "dodge") +
  labs(title = "Stress level and Family Structure",
       x = "Family Structure",
       y = "Count",
       fill = "stress Level") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme_minimal()
```



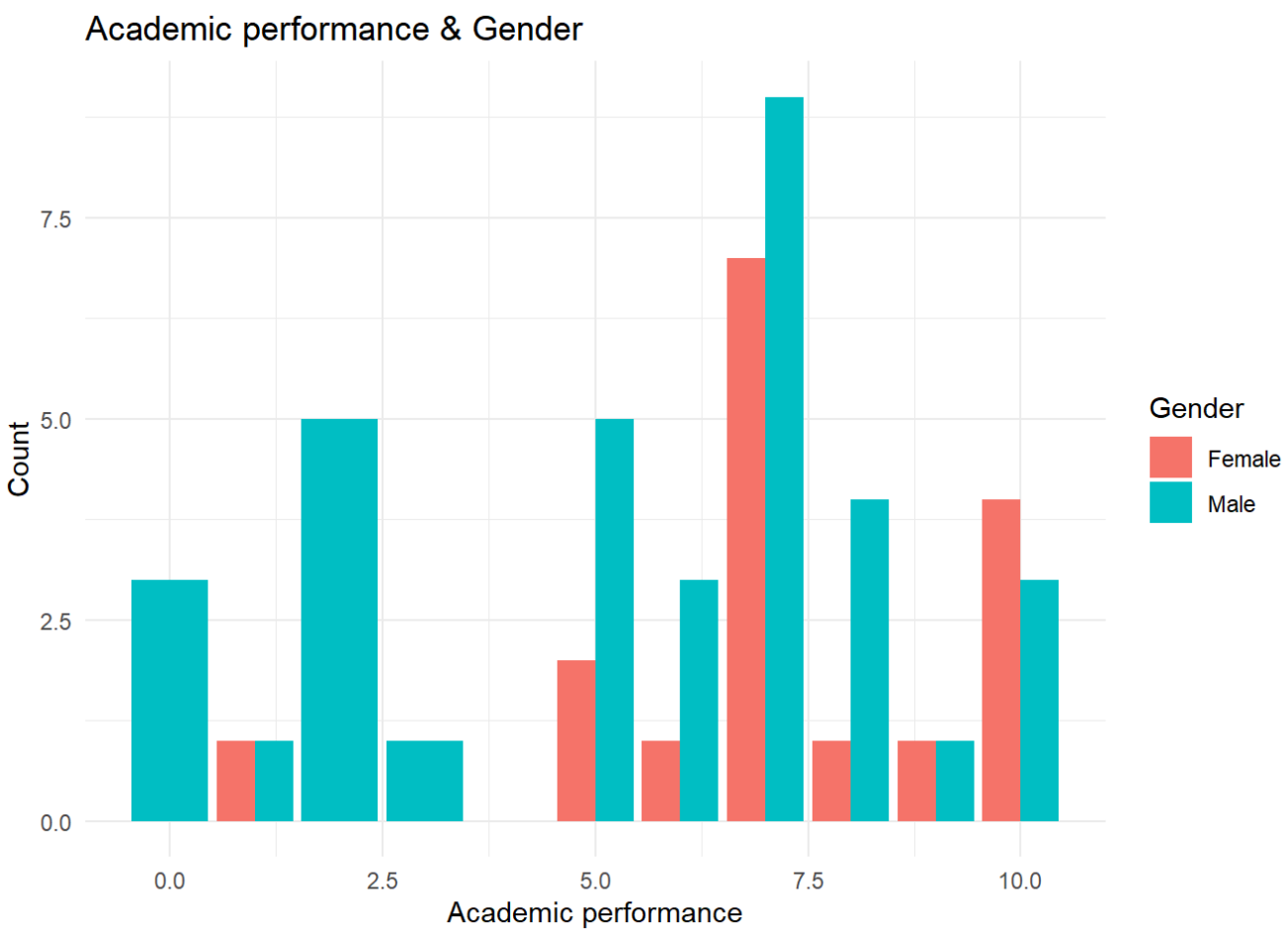

```
# Graph for parents_education_level & parents_education_level1
ggplot(data_cleaned, aes(x = parents_education_level, fill = parents_education_level1)) +
  geom_bar(position = "dodge") +
  labs(title = "Relationship Between Parents' Education Levels",
       x = "Parents' Education Level",
       y = "Count",
       fill = "Partner's Education Level") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme_minimal()
```



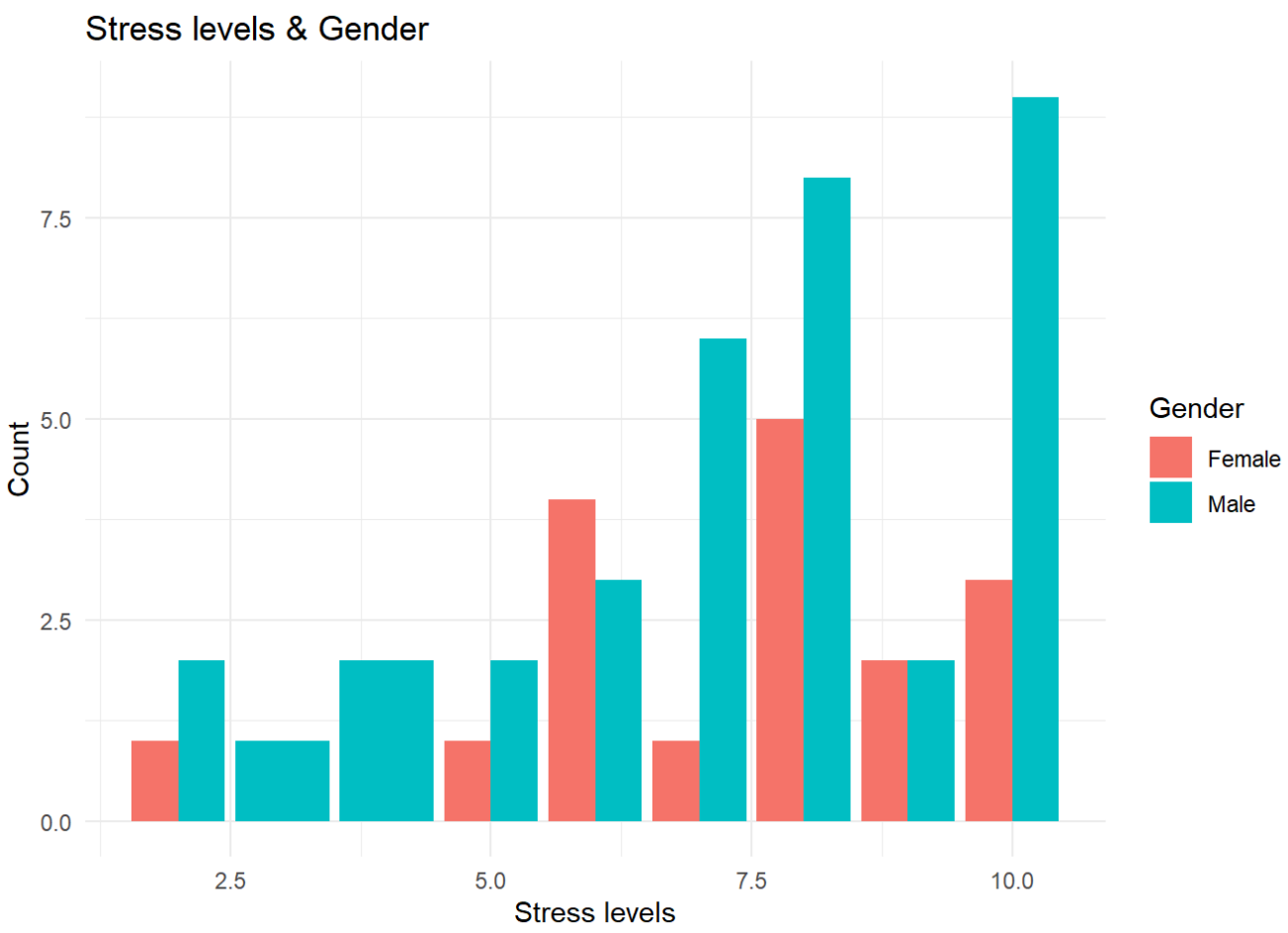

```
#--------------------------------------
#Plot box plot for  qualitative and Quantitative variables
ggplot(data_cleaned, aes(x = gender, y = GPA, fill = gender)) +
  geom_boxplot() +
  labs(title = "Academic Performance by Gender",
       x = "Gender",
       y = "GPA") +
  theme_minimal()
```




```
ggplot(data_cleaned, aes(x = GPA, fill = gender)) +
  geom_bar(position = "dodge") +
  labs(title = "Academic performance & Gender",
       x = "Academic performance",
       y = "Count",
       fill = "Gender") +
  theme_minimal()
```




```
ggplot(data_cleaned, aes(x = Stress_levels, fill = gender)) +
  geom_bar(position = "dodge") +
  labs(title = "Stress levels & Gender",
       x = "Stress levels",
       y = "Count",
       fill = "Gender") +
  theme_minimal()
```



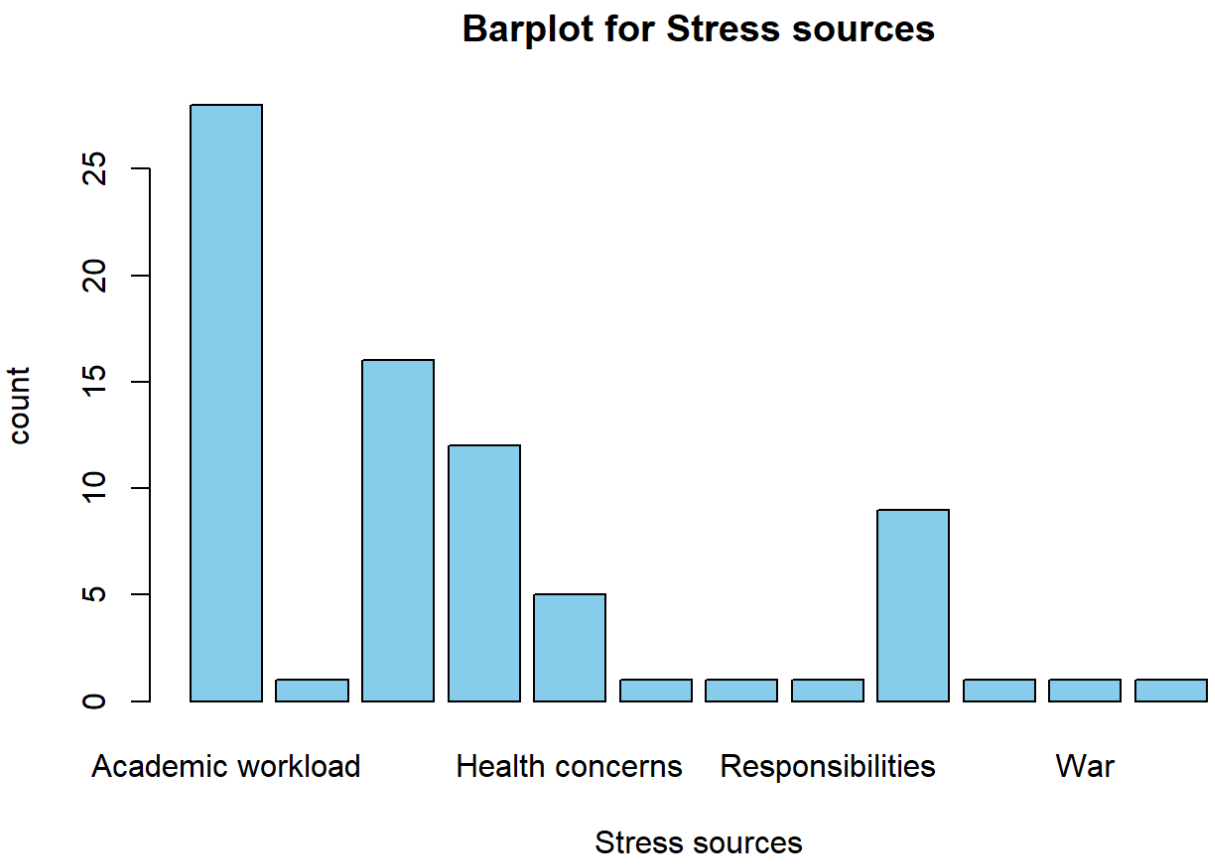

```
#----------------------for overlapped variables (more than one answer per response)
#stress_sources
table(data_cleaned$stress_sources)
```

```
##
##                                                                                  Academic workload;
##                                                                                                 17
##                  Academic workload;Family issues;Financial problems;Social relationships;
##                                                                                                  1
##                                                        Academic workload;Financial problems;
##                                                                                                  1
##                                                           Academic workload;Health concerns;
##                                                                                                  1
##                                                       Academic workload;Social relationships;
##                                                                                                  1
##                                                                                          Exams;
##                                                                                                  1
##                                                                                  Family issues;
##                                                                                                  7
##                                                            Family issues; Academic workload;
##                                                                                                  2
##                                                             Family issues;Responsibilities ;
##                                                                                                  1
##                                                                              Financial problems;
##                                                                                                  4
##                                                        Financial problems; Academic workload;
##                                                                                                  1
##                                       Financial problems; Academic workload;Family issues;
##                                                                                                  1
##                 Financial problems;Family issues; Academic workload;Social relationships;
##                                                                                                  1
##                                                                                 Health concerns;
##                                                                                                  2
##                Health concerns;Financial problems;Family issues;War effects, Immigration;
##                                                                                                  1
## Health concerns;Social relationships;Financial problems;Family issues; Academic workload;
##                                                                                                  1
##                                                                                       Nothing ;
##                                                                                                  1
##                                                                                          Other;
##                                                                                                  1
##                                                                               Social relationships;
##                                                                                                  2
##                                                      Social relationships; Academic workload;
##                                                                                                  1
##                                                           Social relationships;Family issues;
##                                                                                                  1
##                                                      Social relationships;Financial problems;
##                                                                                                  1
##                                                                                  The Gaza war;
##                                                                                                  1
##                                                                                          War;
##                                                                                                  1
```

```r
# Step 1: Clean and process the data
Stress_Source <- unlist(strsplit(data_cleaned$stress_sources, "[;\n]")) # Split by semicolon and newline
Stress_Source <- trimws(Stress_Source)               # Remove whitespace
Stress_Source <- Stress_Source[Stress_Source != ""]       # Remove empty strings

# Step 2: Create a frequency table
frequency_Stress_Source <- table(Stress_Source)
frequency_Stress_Source
```

```
## Stress_Source
##        Academic workload                    Exams            Family issues
##                       28                        1                       16
##       Financial problems          Health concerns                  Nothing
##                       12                        5                        1
##                    Other         Responsibilities     Social relationships
##                        1                        1                        9
##             The Gaza war                      War War effects, Immigration
##                        1                        1                        1
```

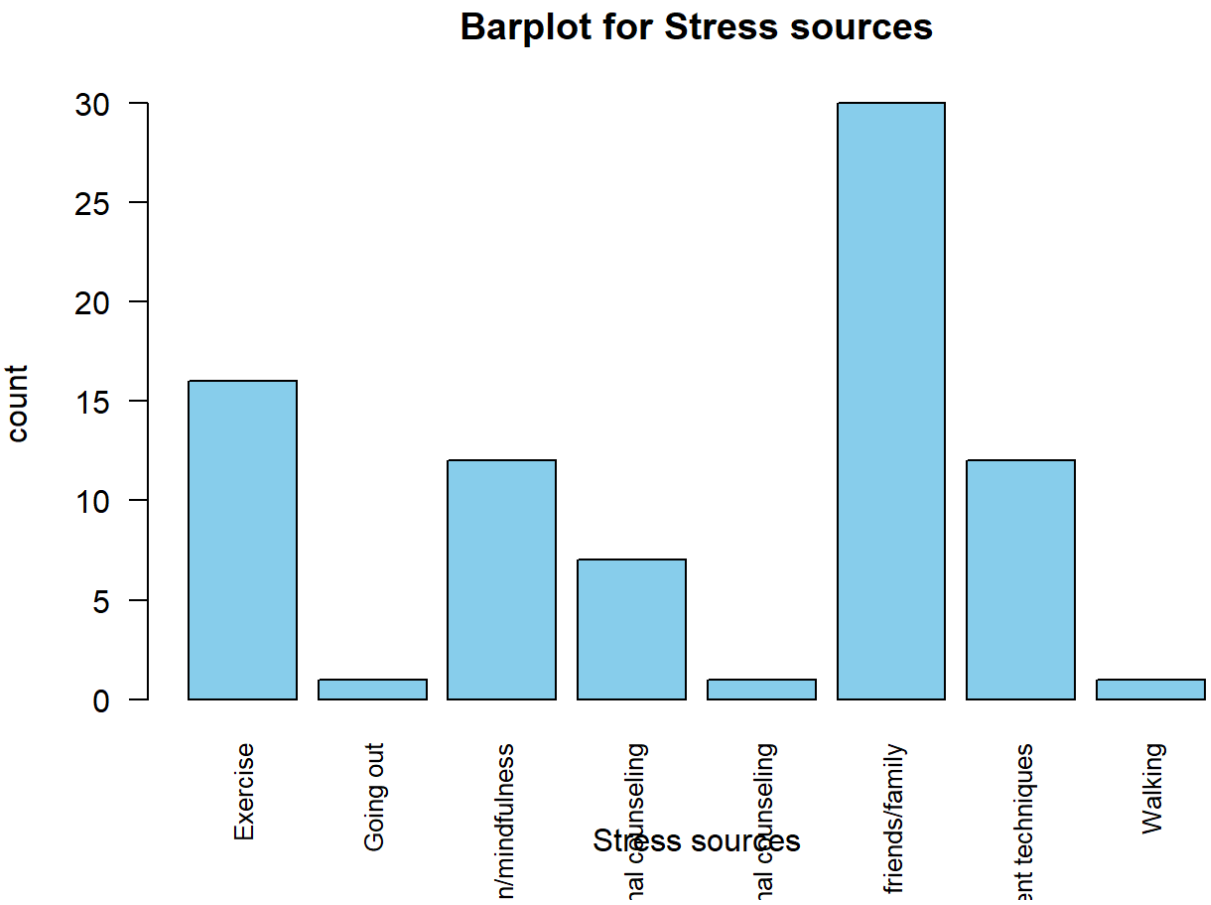
```r
barplot(frequency_Stress_Source, main = "Barplot for Stress sources", xlab = "Stress sources", ylab = "coun
t", col = "skyblue")
```



```
### Coping Strategies
# Step 1: Clean and process the data
coping_strategies <- unlist(strsplit(data_cleaned$coping_strategies, "[;\n]")) # Split by semicolon and newl
ine
coping_strategies <- trimws(coping_strategies)               # Remove whitespace
coping_strategies <- coping_strategies[coping_strategies != ""]       # Remove empty strings

# Step 2: Create a frequency table
frequency_coping_strategies <- table(coping_strategies)
frequency_coping_strategies
```

```
## coping_strategies
##                       Exercise                     Going out
##                             16                             1
##       Meditation/mindfulness      Professional counseling
##                             12                             7
## Self professional counseling    Talking to friends/family
##                              1                            30
##   Time management techniques                       Walking
##                             12                             1
```

```
barplot(frequency_coping_strategies, main = "Barplot for Stress sources", xlab = "Stress sources", ylab = "c
ount", col = "skyblue",las = 2, cex.names = 0.8)
```



Confidence Interval (CI)

A confidence interval is used to estimate a range of values that is likely to contain the population parameter. For example, if we want to estimate the mean GPA of all students based on a sample,we would calculate the confidence interval for **the Sample mean GPA**.

```
# Point Estimate for the population mean (GPA for All student in college)
Estimated_population_mean = mean(data_cleaned$GPA)
Estimated_population_mean
```

```
## [1] 6.038462
```

```
# 1- 95% Confidence Interval for GPA
confidence_level <- 0.95
z_value <- qnorm(1 - (1 - confidence_level) / 2)  # z=1.96 for 95% CI
z_value
```

```
## [1] 1.959964
```

```
# Calculate margin of error
n = length(data_cleaned$GPA)
margin_of_error <- z_value * (sd(data_cleaned$GPA) / sqrt(n))

# Calculate confidence interval
CI_lower <- Estimated_population_mean - margin_of_error
CI_upper <- Estimated_population_mean + margin_of_error
CI = c(CI_lower, CI_upper)
CI
```

```
## [1] 5.258548 6.818375
```

- This means we are 95% confident that the true population mean GPA lies between **5.259 and 6.818**.
- The margin of error is 0.780 (rounded), indicating **the estimate's precision**. As smaller margin of error suggests more precision in the estimate of the population mean of GPA.

Hypothesis Testing

We'll conduct hypothesis tests to validate or reject assumptions about the data.

- Test Conducted: One-sample z-test on the GPA variable to determine if the true mean of GPA is significantly different from 2. # Case (1)- Two Sided
- Null Hypothesis (H_0): The true mean GPA is equal to 2.
- Alternative Hypothesis (H_1): The true mean GPA is not equal to 2.

```
#install.packages("BSDA")
library("BSDA")
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'BSDA'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```
z.test(data_cleaned$GPA,
       mu = 2, sigma.x =sd(data_cleaned$GPA),
       conf.level = 0.95, alternative = "two.sided")
```

```
##
##  One-sample z-Test
##
## data:  data_cleaned$GPA
## z = 10.149, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 2
## 95 percent confidence interval:
##  5.258548 6.818375
## sample estimates:
## mean of x
##  6.038462
```

- Reject H_0 and accept alternative hypothesis: true mean is not equal to 2
- Coinfidence Interval: [5.258548 - 6.818375]

```
# Case (2)- one Side (Right tail)
z.test(data_cleaned$GPA,
       mu = 2, sigma.x =sd(data_cleaned$GPA),
       conf.level = 0.95, alternative = "greater")
```

```
##
##  One-sample z-Test
##
## data:  data_cleaned$GPA
## z = 10.149, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 2
## 95 percent confidence interval:
##  5.383938       NA
## sample estimates:
## mean of x
##  6.038462
```

- Reject H_0 and accept alternative hypothesis: true mean greater than 2
- Coinfidence Interval: [5.383938 - infinity]

```
# Case (3)- one Side (left tail)
z.test(data_cleaned$GPA,
       mu = 2, sigma.x =sd(data_cleaned$GPA),
       conf.level = 0.95, alternative = "greater")
```

```
##
##  One-sample z-Test
##
## data:  data_cleaned$GPA
## z = 10.149, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 2
## 95 percent confidence interval:
##  5.383938       NA
## sample estimates:
## mean of x
##  6.038462
```

- Reject H_0 and accept alternative hypothesis: true mean greater than 2
- Coinfidence Interval: [5.383938 - infinity]

A common test to begin with is the t-test, depending on the nature of the data.

```
# One-sample t-test to test if the mean GPA is significantly different from 6.0
t_test_result <- t.test(data_cleaned$GPA, mu = 6.0)
t_test_result
```

```
##
##  One Sample t-test
##
## data:  data_cleaned$GPA
## t = 0.096656, df = 51, p-value = 0.9234
## alternative hypothesis: true mean is not equal to 6
## 95 percent confidence interval:
##  5.239599 6.837324
## sample estimates:
## mean of x
##  6.038462
```

- The p-value of 0.9234 is much greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis.
- This means there is no significant evidence to suggest that the true mean of GPA is different from 6.

```
# One-sample t-test for testing if the mean GPA is greater than 6
t_test_result <- t.test(data_cleaned$GPA, mu = 6, alternative = "greater")
t_test_result
```

```
##
##  One Sample t-test
##
## data:  data_cleaned$GPA
## t = 0.096656, df = 51, p-value = 0.4617
## alternative hypothesis: true mean is greater than 6
## 95 percent confidence interval:
##  5.371829      Inf
## sample estimates:
## mean of x
##  6.038462
```

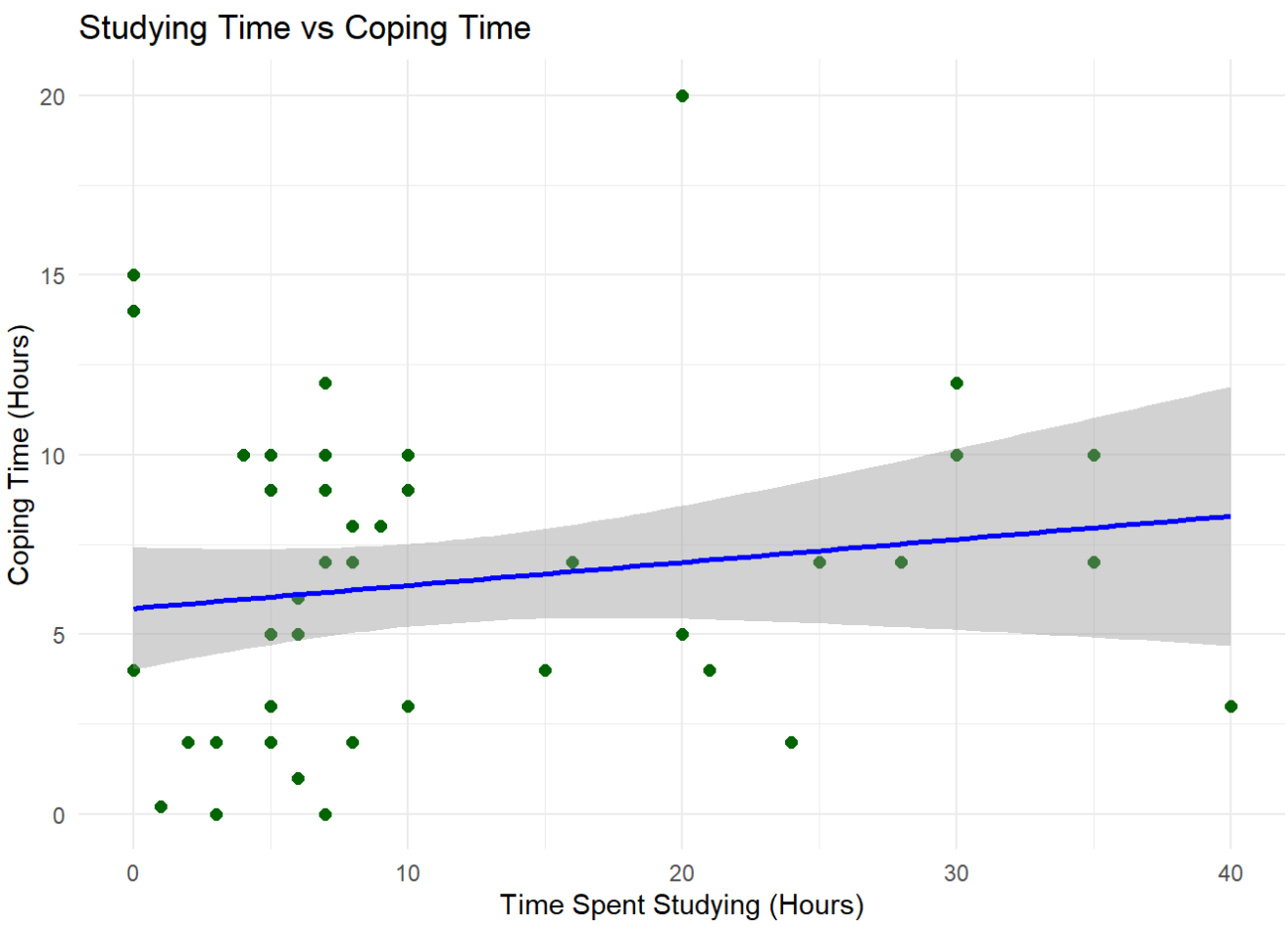
- Since the p-value (0.4617) is much higher than the significance level (0.05), we fail to reject the null hypothesis.
- This means there isn't sufficient evidence to claim that the mean GPA is greater than 6.

Regression Analysis

we can explore relationships between variables, such as how stress levels and coping strategies affect academic performance (GPA).

```
ggplot(data_cleaned, aes(x = time_spent_studying, y = frequency_of_coping)) +
  geom_point(color = "darkgreen", size = 2) +
  geom_smooth(method = "lm", col = "blue", se = TRUE) +  # Add confidence interval with se = TRUE
  labs(title = "Studying Time vs Coping Time",
       x = "Time Spent Studying (Hours)",
       y = "Coping Time (Hours)") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# Split the dataset by family structure and perform linear regression
#to Apply regression for each group
family_groups <- split(data_cleaned, data_cleaned$family_structure)
regressions <- lapply(family_groups, function(df) {lm(GPA ~ Stress_levels, data = df)})
lapply(regressions, summary)     # Display summary for each group
```

```
## $` Extended family household`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2548 -1.1512  0.3468  1.3468  4.5460
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.2467     2.2289   3.251   0.0047 **
## Stress_levels  -0.1992     0.2902  -0.686   0.5017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 17 degrees of freedom
## Multiple R-squared:  0.02697,    Adjusted R-squared:  -0.03027
## F-statistic: 0.4712 on 1 and 17 DF,  p-value: 0.5017
##
##
## $` Single-parent household`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
##       1       2       3       4       5       6
##  0.2881 -0.8051  1.9237 -0.8898  2.9237 -3.4407
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.1610     3.1338   0.690    0.528
## Stress_levels   0.3644     0.4035   0.903    0.417
##
## Residual standard error: 2.53 on 4 degrees of freedom
## Multiple R-squared:  0.1694, Adjusted R-squared:  -0.03825
## F-statistic: 0.8158 on 1 and 4 DF,  p-value: 0.4175
##
##
## $`I Live alone`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
## ALL 1 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3        NaN     NaN      NaN
## Stress_levels       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##
##
```

```
## $`Separated parents`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
## ALL 1 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8        NaN     NaN      NaN
## Stress_levels       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##
##
## $`Two-parent household`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1565 -0.4682  0.7612  1.7612  3.5318
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.8624     2.2441   2.167   0.0409 *
## Stress_levels   0.2294     0.2926   0.784   0.4409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.955 on 23 degrees of freedom
## Multiple R-squared:  0.02604,    Adjusted R-squared:  -0.01631
## F-statistic: 0.6149 on 1 and 23 DF,  p-value: 0.4409
```

- There is no strong or significant relationship between stress levels and GPA in the Extended family household.
- Slope: -0.1992 (Negative relationship; as stress levels increase, GPA decreases slightly, but the effect is not significant).

```
# Perform One-Way ANOVA (Example on parents_education_level)
anova_result <- aov(GPA ~ parents_education_level, data = data_cleaned)
summary_anova <- summary(anova_result)
print(summary_anova)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## parents_education_level  3   22.0   7.327   0.884  0.456
## Residuals               48  397.9   8.290
```

```
# P-value interpretation
p_value_anova <- summary_anova[[1]][["Pr(>F)"]][1]
if (p_value_anova < 0.05) {
  cat("There is a significant difference in GPA between education level groups.\n")
} else {
  cat("There is no significant difference in GPA between education level groups.\n")
}
```

```
## There is no significant difference in GPA between education level groups.
```

- The p-value (0.456) is greater than 0.05, which indicates that the differences in GPA between the levels of parents' education level are not statistically significant at the 5% significance level.
- This means that there is no sufficient evidence to conclude that the GPA differs based on parents' education level.
- The variation in GPA scores is mostly due to random chance rather than a significant effect of parents' education level.
- parents_education_level (Between Groups): 22.0 — variability explained by the group differences.

```
# Perform One-Way ANOVA (on gender)
anova_result <- aov(GPA ~ gender, data = data_cleaned)
summary_anova <- summary(anova_result)
print(summary_anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## gender       1   36.2   36.18   4.714 0.0347 *
## Residuals   50  383.7    7.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# P-value interpretation
p_value_anova <- summary_anova[[1]][["Pr(>F)"]][1]
if (p_value_anova < 0.05) {
  cat("There is a significant difference in GPA between education level groups.\n")
} else {
  cat("There is no significant difference in GPA between education level groups.\n")
}
```

```
## There is a significant difference in GPA between education level groups.
```

```
# Regression analysis:  GPA based on time spent studying
model <- lm(GPA ~ time_spent_studying, data = data_cleaned)
summary(model)
```

```
##
## Call:
## lm(formula = GPA ~ time_spent_studying, data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0709 -1.0380  0.9628  1.9337  3.9712
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.020345   0.594914  10.120 1.08e-13 ***
## time_spent_studying 0.001685   0.040805   0.041    0.967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.898 on 50 degrees of freedom
## Multiple R-squared:  3.411e-05,  Adjusted R-squared:  -0.01997
## F-statistic: 0.001706 on 1 and 50 DF,  p-value: 0.9672
```

```
anova_results <- anova(model)
anova_results
```

```
## Analysis of Variance Table
##
## Response: GPA
##                      Df Sum Sq Mean Sq F value Pr(>F)
## time_spent_studying  1   0.01  0.0143  0.0017 0.9672
## Residuals           50 419.91  8.3982
```

- There is no significant relationship between time spent studying and GPA.
- The slope coefficient (0.0017) is not significant (p>0.05), and the R-squared value shows that time spent studying explains an insignificant portion of the GPA variance.
- This suggests that factors other than time spent studying are likely influencing GPA.

```
regressions <- lapply(family_groups, function(df) {lm(GPA ~ time_spent_studying, data = df)})
lapply(regressions, summary)
```

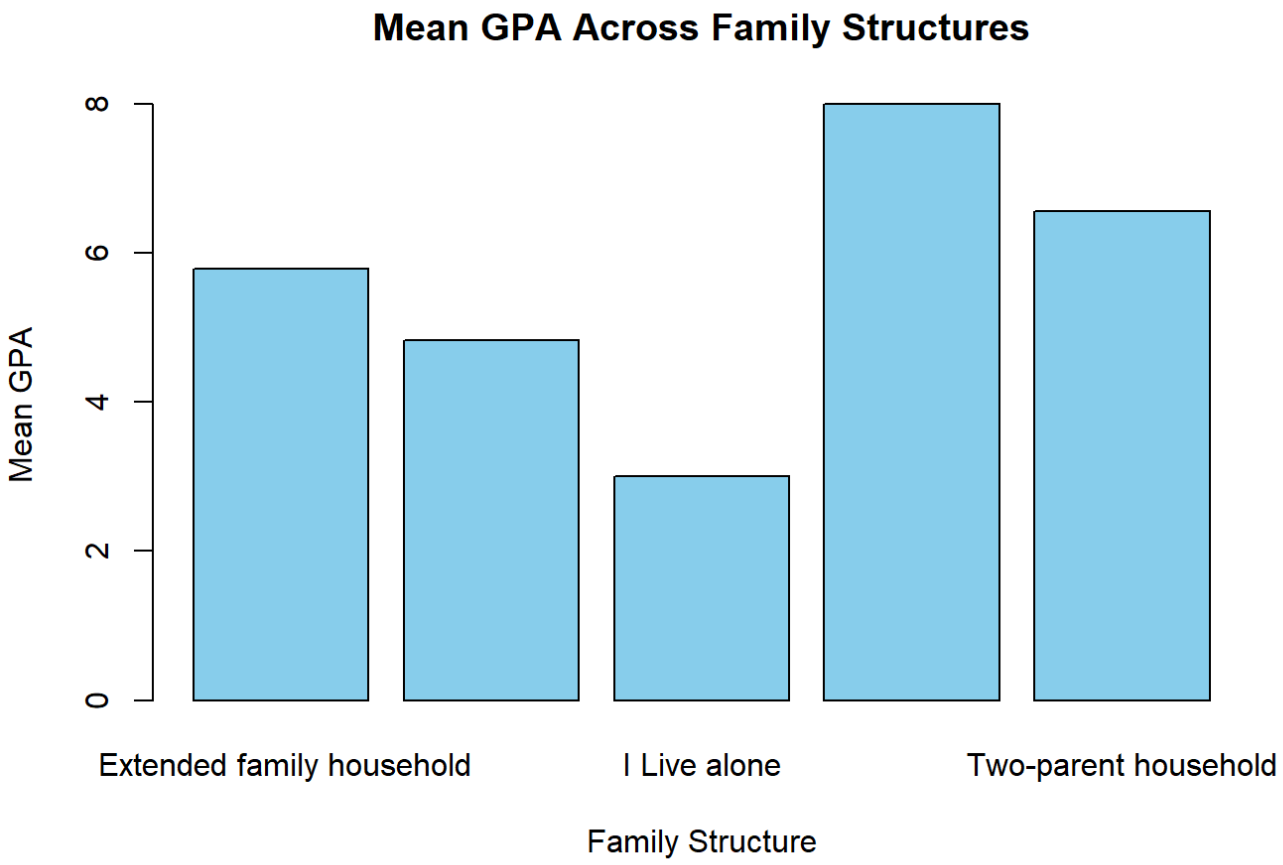
```
## $` Extended family household`
##
## Call:
## lm(formula = GPA ~ time_spent_studying, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.768 -1.122  0.534  1.139  3.534
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.52456    0.82436   9.128  5.8e-08 ***
## time_spent_studying -0.15122    0.05239  -2.886   0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 17 degrees of freedom
## Multiple R-squared:  0.3289, Adjusted R-squared:  0.2894
## F-statistic: 8.331 on 1 and 17 DF,  p-value: 0.01025
##
##
## $` Single-parent household`
##
## Call:
## lm(formula = GPA ~ time_spent_studying, data = df)
##
## Residuals:
##       1       2       3       4       5       6
##  0.9078  0.6854  2.4631 -1.6476 -0.5388 -1.8699
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.9806     1.1307   2.636   0.0578 .
## time_spent_studying   0.2223     0.1005   2.213   0.0913 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.862 on 4 degrees of freedom
## Multiple R-squared:  0.5504, Adjusted R-squared:  0.438
## F-statistic: 4.897 on 1 and 4 DF,  p-value: 0.09133
##
##
## $`I Live alone`
##
## Call:
## lm(formula = GPA ~ time_spent_studying, data = df)
##
## Residuals:
## ALL 1 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3        NaN     NaN      NaN
## time_spent_studying       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
```

```
##
##
## $`Separated parents`
##
## Call:
## lm(formula = GPA ~ time_spent_studying, data = df)
##
## Residuals:
## ALL 1 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                8        NaN     NaN      NaN
## time_spent_studying       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##
##
## $`Two-parent household`
##
## Call:
## lm(formula = GPA ~ time_spent_studying, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4484 -0.8408  0.2288  1.0833  3.9504
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.11920    0.81667   6.268 2.14e-06 ***
## time_spent_studying  0.13292    0.05669   2.345    0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.69 on 23 degrees of freedom
## Multiple R-squared:  0.1929, Adjusted R-squared:  0.1578
## F-statistic: 5.498 on 1 and 23 DF,  p-value: 0.02804
```

```
#-------------------------------------------------------------------------------
# Compare GPA across family structures
family_gpa <- aggregate(GPA ~ family_structure, data = data_cleaned, mean)
family_gpa
```

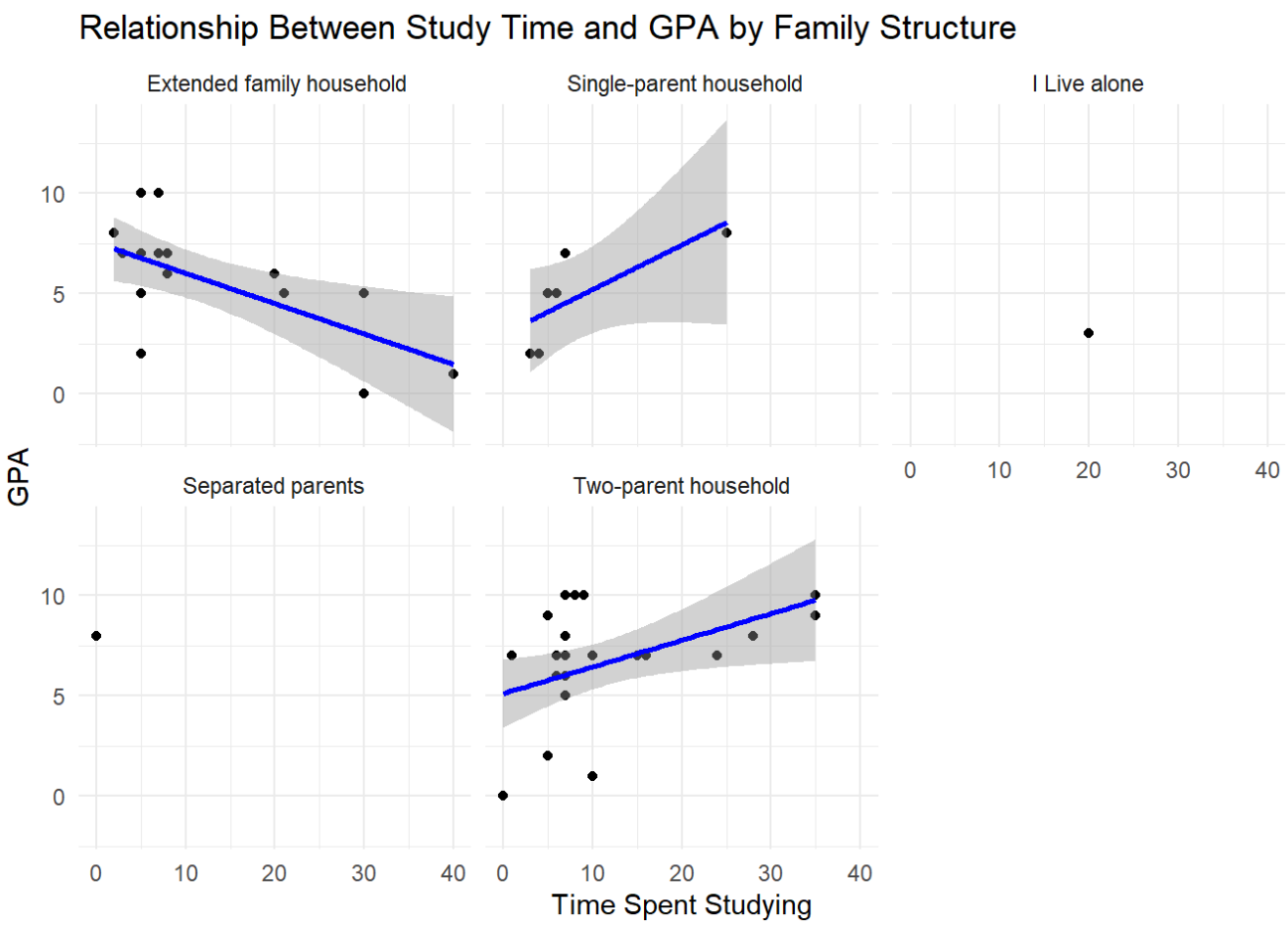
```
##             family_structure      GPA
## 1  Extended family household 5.789474
## 2    Single-parent household 4.833333
## 3               I Live alone 3.000000
## 4          Separated parents 8.000000
## 5       Two-parent household 6.560000
```

```
# Create bar plot
barplot(family_gpa$GPA,
        names.arg = family_gpa$family_structure,
        col = "skyblue",
        main = "Mean GPA Across Family Structures",
        ylab = "Mean GPA",
        xlab = "Family Structure")
```




```
ggplot(data_cleaned, aes(x = time_spent_studying, y = GPA)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  facet_wrap(~ family_structure) +
  labs(title = "Relationship Between Study Time and GPA by Family Structure",
       x = "Time Spent Studying",
       y = "GPA") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# Regression model: GPA ~ stress_levels + family_structure + digital_access
model <- lm(GPA ~ Stress_levels + family_structure + digital_access + age, data = data_cleaned)
summary(model)
```

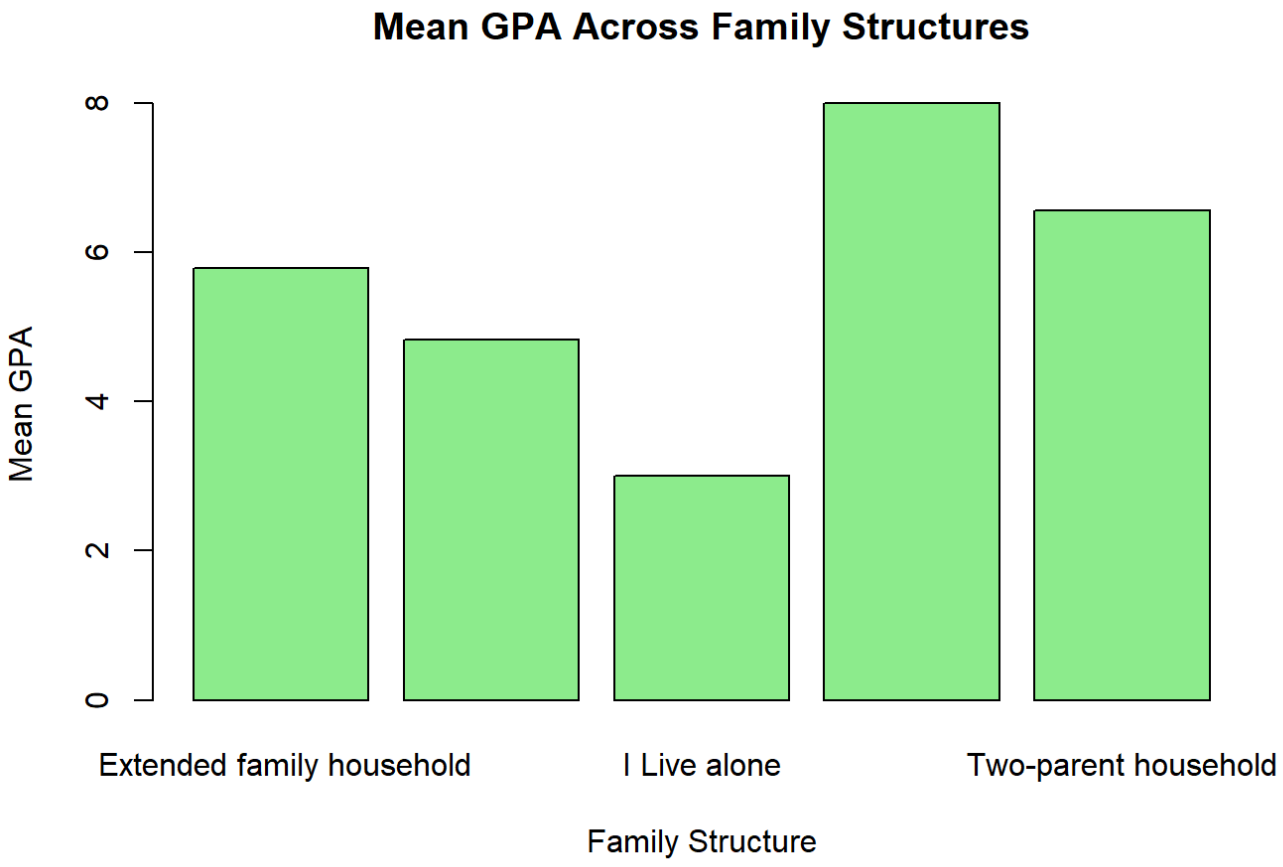
```
##
## Call:
## lm(formula = GPA ~ Stress_levels + family_structure + digital_access +
##     age, data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1925 -1.1842  0.3935  1.6870  4.9217
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               8.0516     4.6597   1.728   0.0913 .
## Stress_levels                             0.0414     0.1913   0.216   0.8298
## family_structure Single-parent household -0.8562     1.3906  -0.616   0.5414
## family_structureI Live alone             -2.2595     3.1376  -0.720   0.4754
## family_structureSeparated parents         2.6288     3.1968   0.822   0.4155
## family_structureTwo-parent household      0.7831     0.9074   0.863   0.3930
## digital_accessGood access                 1.3121     0.9787   1.341   0.1872
## digital_accessLimited access              0.9531     1.8676   0.510   0.6125
## digital_accessModerate access             0.1500     1.2439   0.121   0.9046
## age                                      -0.1687     0.2165  -0.779   0.4402
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.952 on 42 degrees of freedom
## Multiple R-squared:  0.1286, Adjusted R-squared:  -0.05811
## F-statistic: 0.6888 on 9 and 42 DF,  p-value: 0.7148
```

```
#Coefficients will show how stress levels and family background influence GPA, accounting for age and digita
l access.

# Apply linear regression (Stress Levels ~ GPA) for each family group
stress_regressions <- lapply(family_groups, function(df) {lm(GPA ~ Stress_levels, data = df)})
lapply(stress_regressions, summary)      # Display regression summaries for each group
```

```
## $` Extended family household`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2548 -1.1512  0.3468  1.3468  4.5460
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.2467     2.2289   3.251   0.0047 **
## Stress_levels  -0.1992     0.2902  -0.686   0.5017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 17 degrees of freedom
## Multiple R-squared:  0.02697,    Adjusted R-squared:  -0.03027
## F-statistic: 0.4712 on 1 and 17 DF,  p-value: 0.5017
##
##
## $` Single-parent household`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
##       1       2       3       4       5       6
##  0.2881 -0.8051  1.9237 -0.8898  2.9237 -3.4407
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.1610     3.1338   0.690    0.528
## Stress_levels   0.3644     0.4035   0.903    0.417
##
## Residual standard error: 2.53 on 4 degrees of freedom
## Multiple R-squared:  0.1694, Adjusted R-squared:  -0.03825
## F-statistic: 0.8158 on 1 and 4 DF,  p-value: 0.4175
##
##
## $`I Live alone`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
## ALL 1 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3        NaN     NaN      NaN
## Stress_levels       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##
##
```

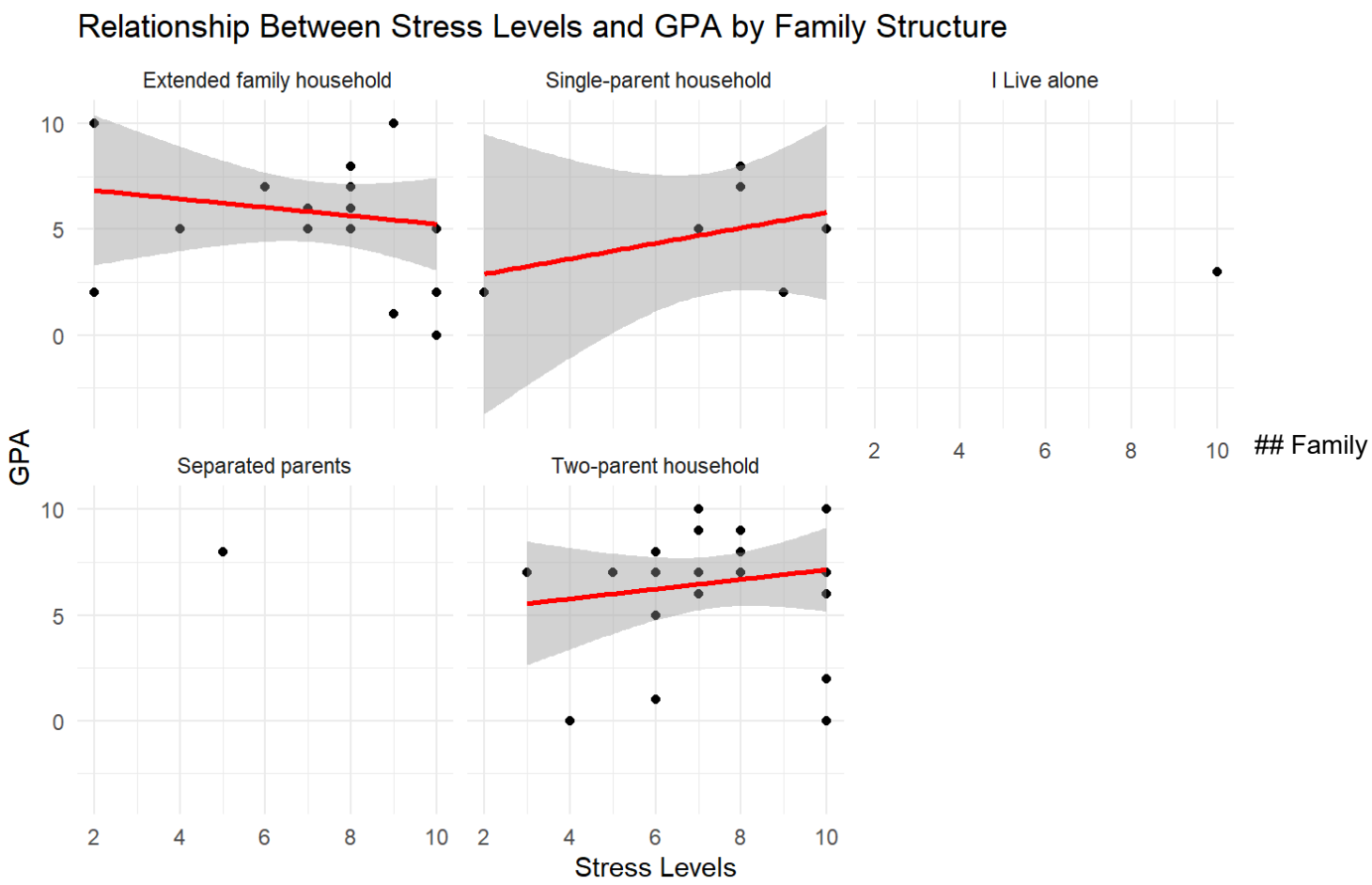
```
## $`Separated parents`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
## ALL 1 residuals are 0: no residual degrees of freedom!
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8        NaN     NaN      NaN
## Stress_levels       NA         NA      NA       NA
##
## Residual standard error: NaN on 0 degrees of freedom
##
##
## $`Two-parent household`
##
## Call:
## lm(formula = GPA ~ Stress_levels, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.1565 -0.4682  0.7612  1.7612  3.5318
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.8624     2.2441   2.167   0.0409 *
## Stress_levels   0.2294     0.2926   0.784   0.4409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.955 on 23 degrees of freedom
## Multiple R-squared:  0.02604,    Adjusted R-squared:  -0.01631
## F-statistic: 0.6149 on 1 and 23 DF,  p-value: 0.4409
```

```
# Compare Mean GPA across family structures
family_gpa_stress <- aggregate(GPA ~ family_structure, data = data_cleaned, mean)

# Create Bar Plot of GPA across Family Structures
barplot(family_gpa_stress$GPA,
        names.arg = family_gpa_stress$family_structure,
        col = "lightgreen",
        main = "Mean GPA Across Family Structures",
        ylab = "Mean GPA",
        xlab = "Family Structure")
```




```
# Scatter Plot of Stress Levels vs GPA with Regression Lines
ggplot(data_cleaned, aes(x = Stress_levels, y = GPA)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  facet_wrap(~ family_structure) +
  labs(title = "Relationship Between Stress Levels and GPA by Family Structure",
       x = "Stress Levels",
       y = "GPA") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Background: How does family structure influence stress and GPA? - Statistical Summary of GPA and Stress by Family Structure

```
aggregate(cbind(GPA, Stress_levels) ~ family_structure, data = data_cleaned, mean)
```

```
##             family_structure      GPA Stress_levels
## 1  Extended family household 5.789474      7.315789
## 2    Single-parent household 4.833333      7.333333
## 3               I Live alone 3.000000     10.000000
## 4          Separated parents 8.000000      5.000000
## 5       Two-parent household 6.560000      7.400000
```

ANOVA: Does Family Structure Affect GPA and Stress?

```
  anova_gpa <- aov(GPA ~ family_structure, data = data_cleaned)
summary(anova_gpa)
```

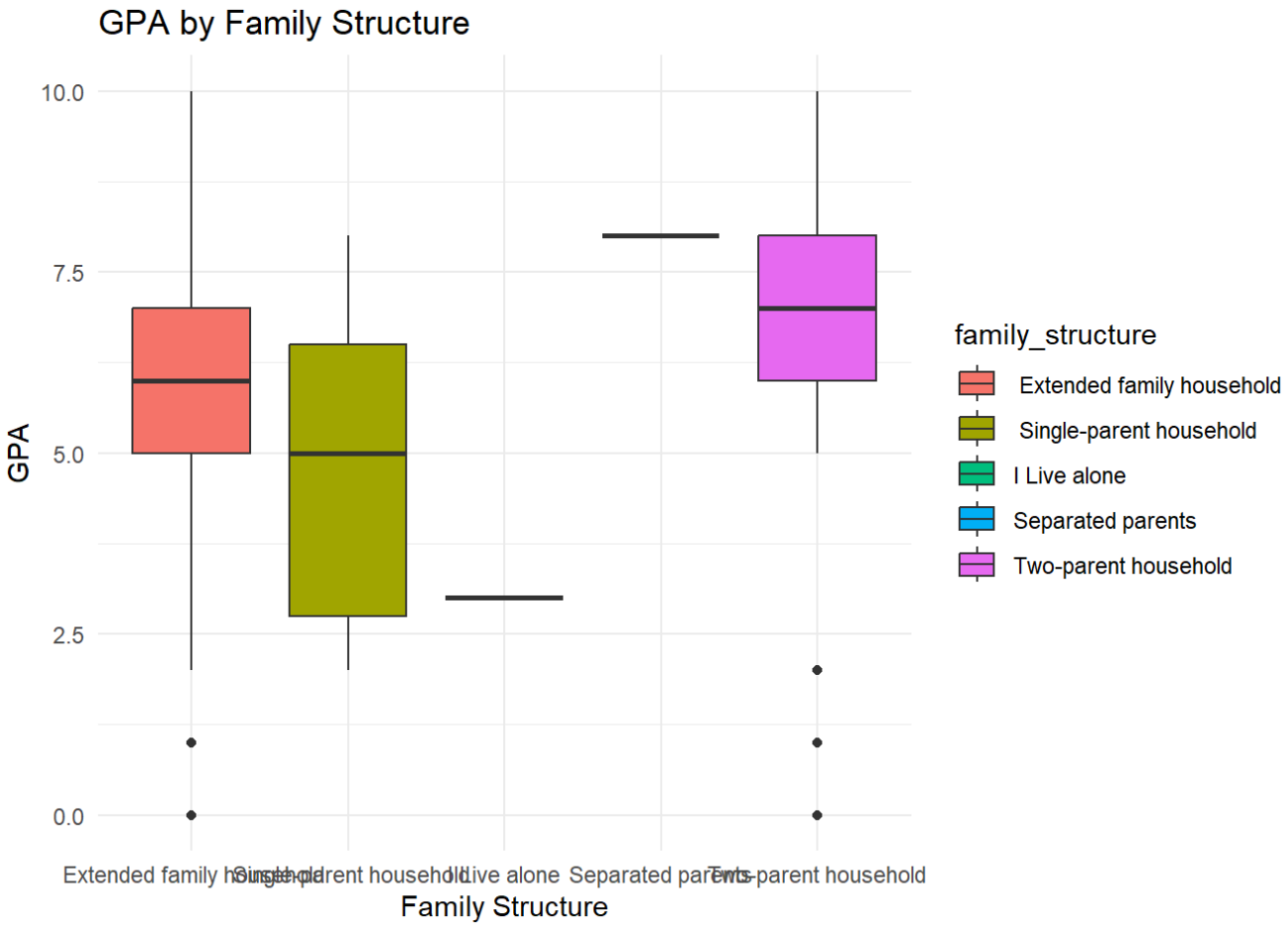
```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## family_structure  4   29.8   7.443   0.897  0.474
## Residuals        47  390.2   8.301
```

```
anova_stress <- aov(Stress_levels ~ family_structure, data = data_cleaned)
summary(anova_stress)
```

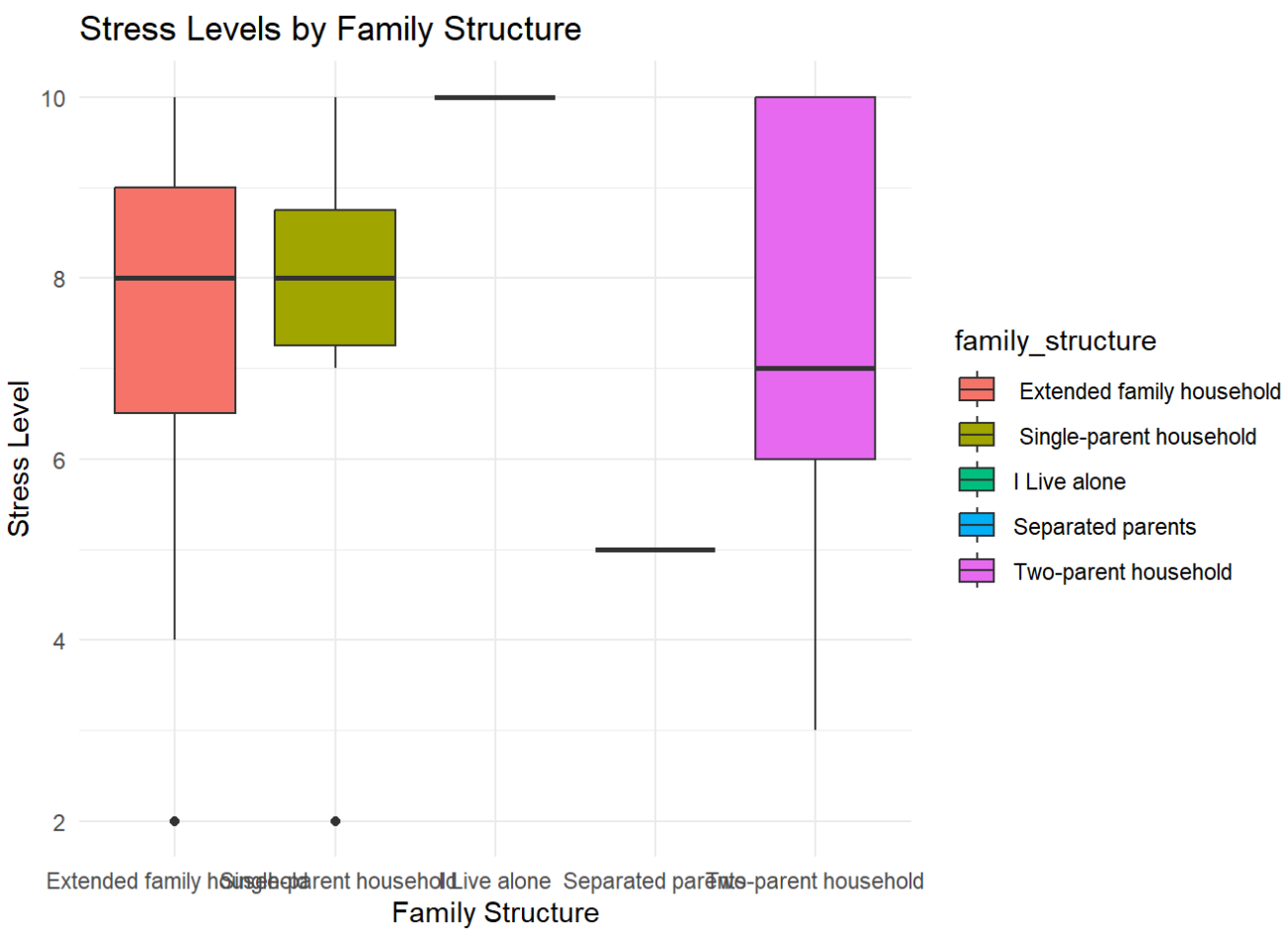
```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## family_structure  4  12.62   3.155   0.604  0.662
## Residuals        47 245.44   5.222
```

Visualization: Boxplots for GPA and Stress by Family Structure

```
ggplot(data_cleaned, aes(x = family_structure, y = GPA, fill = family_structure)) +
  geom_boxplot() +
  labs(title = "GPA by Family Structure", x = "Family Structure", y = "GPA") +
  theme_minimal()
```




```
ggplot(data_cleaned, aes(x = family_structure, y = Stress_levels, fill = family_structure)) +
  geom_boxplot() +
  labs(title = "Stress Levels by Family Structure", x = "Family Structure", y = "Stress Level") +
  theme_minimal()
```



Educational Level: Does level of education impact stress and coping strategies?

```
aggregate(cbind(Stress_levels, frequency_of_coping) ~ parents_education_level, data = data_cleaned, mean)
```

```
##       parents_education_level Stress_levels frequency_of_coping
## 1         No formal education      8.000000                 7.0
## 2              Primary school      7.750000                 6.0
## 3           Secondary school      6.400000                 9.2
## 4 University degree or higher      7.428571                 6.1
```

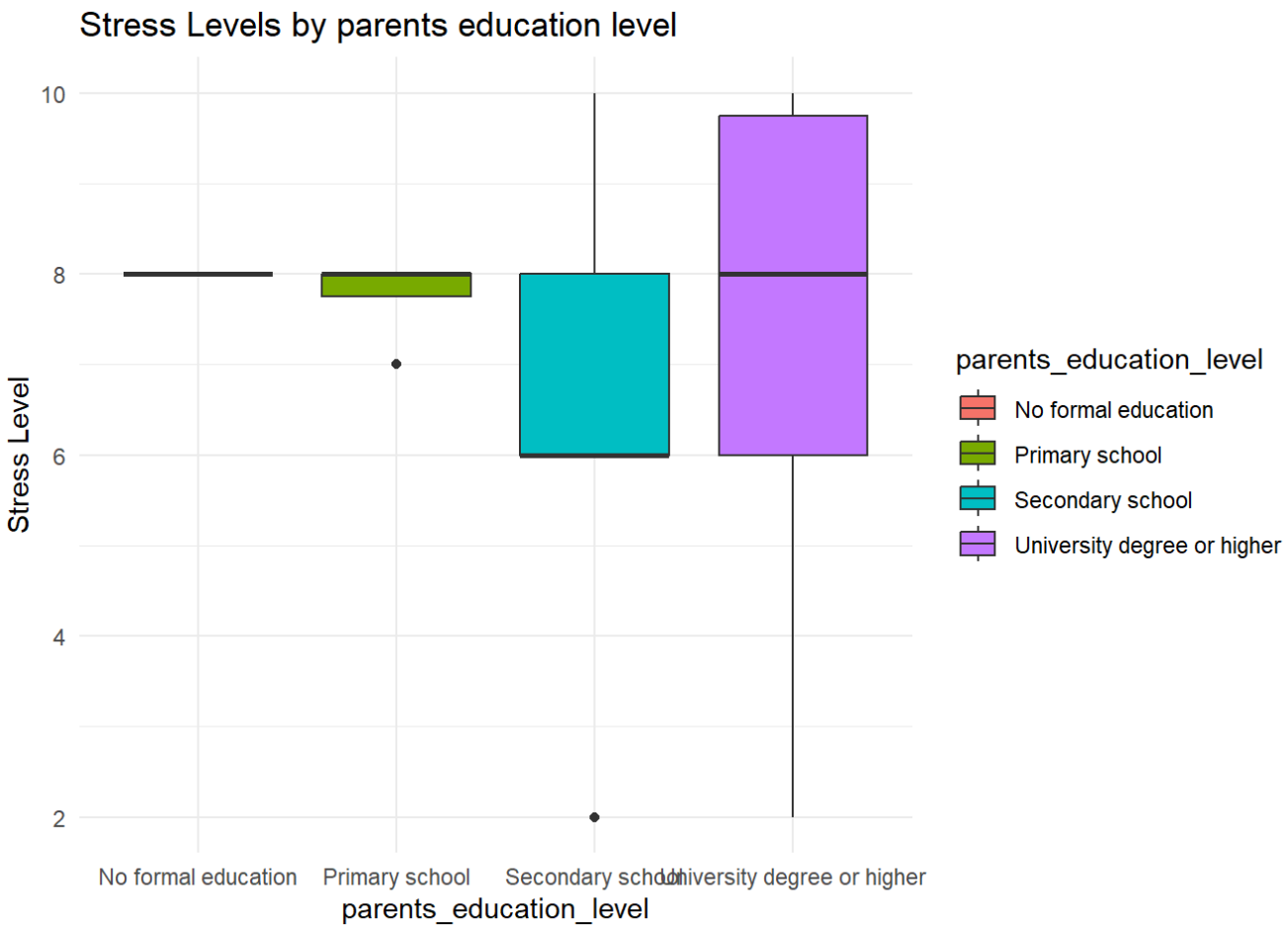
```
anova_edu_stress <- aov(Stress_levels ~ parents_education_level, data = data_cleaned)
summary(anova_edu_stress)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## parents_education_level  3   5.82   1.941   0.369  0.775
## Residuals               48 252.24   5.255
```

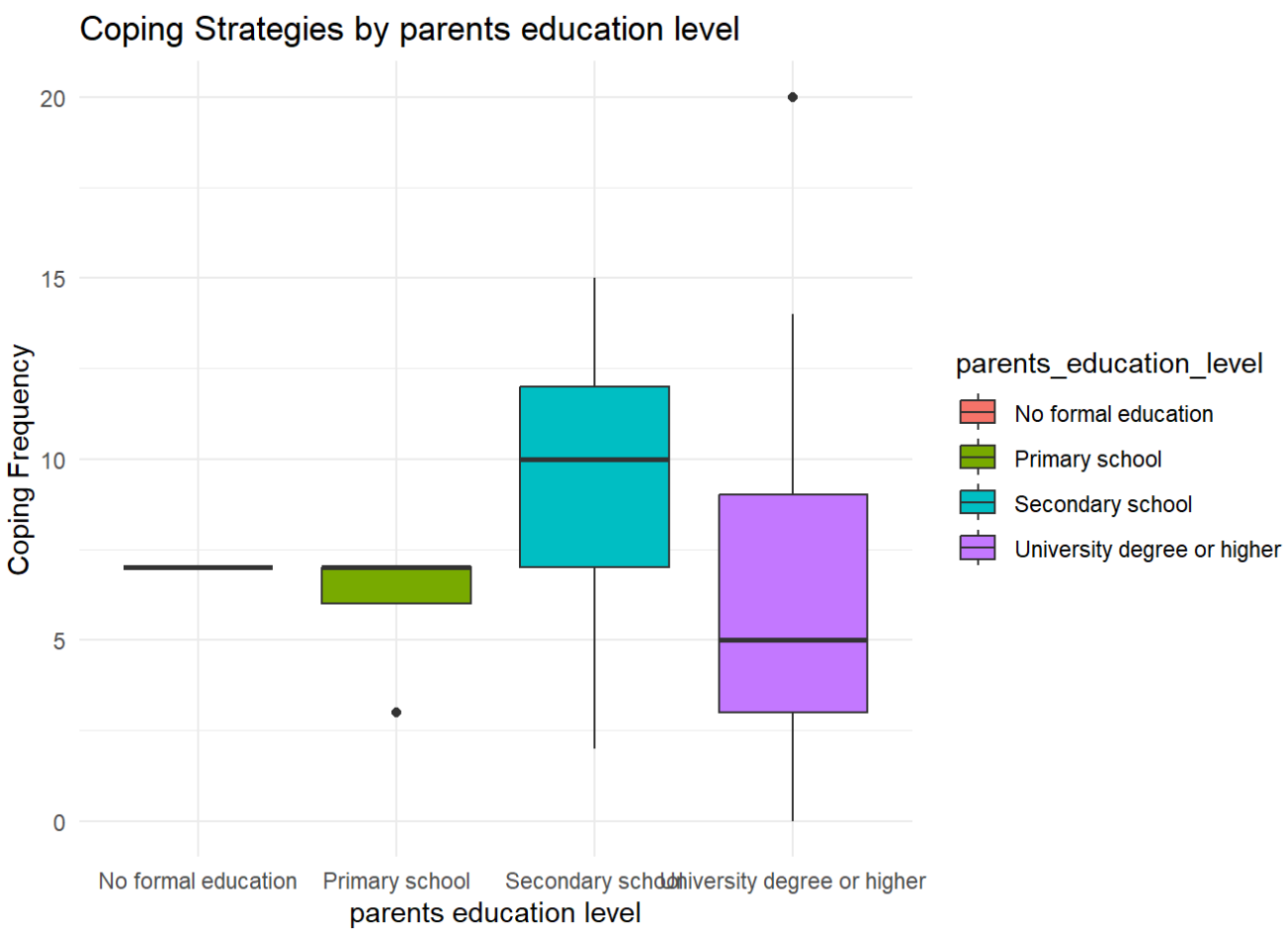
```
anova_edu_coping <- aov(frequency_of_coping ~ parents_education_level, data = data_cleaned)
summary(anova_edu_coping)
```

```
##                         Df Sum Sq Mean Sq F value Pr(>F)
## parents_education_level  3     44   14.66    0.85  0.474
## Residuals               48    828   17.25
```

```
# Visualization: Stress and Coping by Education Level
ggplot(data_cleaned, aes(x = parents_education_level, y = Stress_levels, fill = parents_education_level)) +
  geom_boxplot() +
  labs(title = "Stress Levels by parents education level", x = "parents_education_level", y = "Stress Leve
l") +
  theme_minimal()
```




```
ggplot(data_cleaned, aes(x = parents_education_level, y = frequency_of_coping, fill = parents_education_leve
l)) +
  geom_boxplot() +
  labs(title = "Coping Strategies by parents education level", x = "parents education level", y = "Coping Fr
equency") +
  theme_minimal()
```



Access to Resources: How does digital access impact academic success?

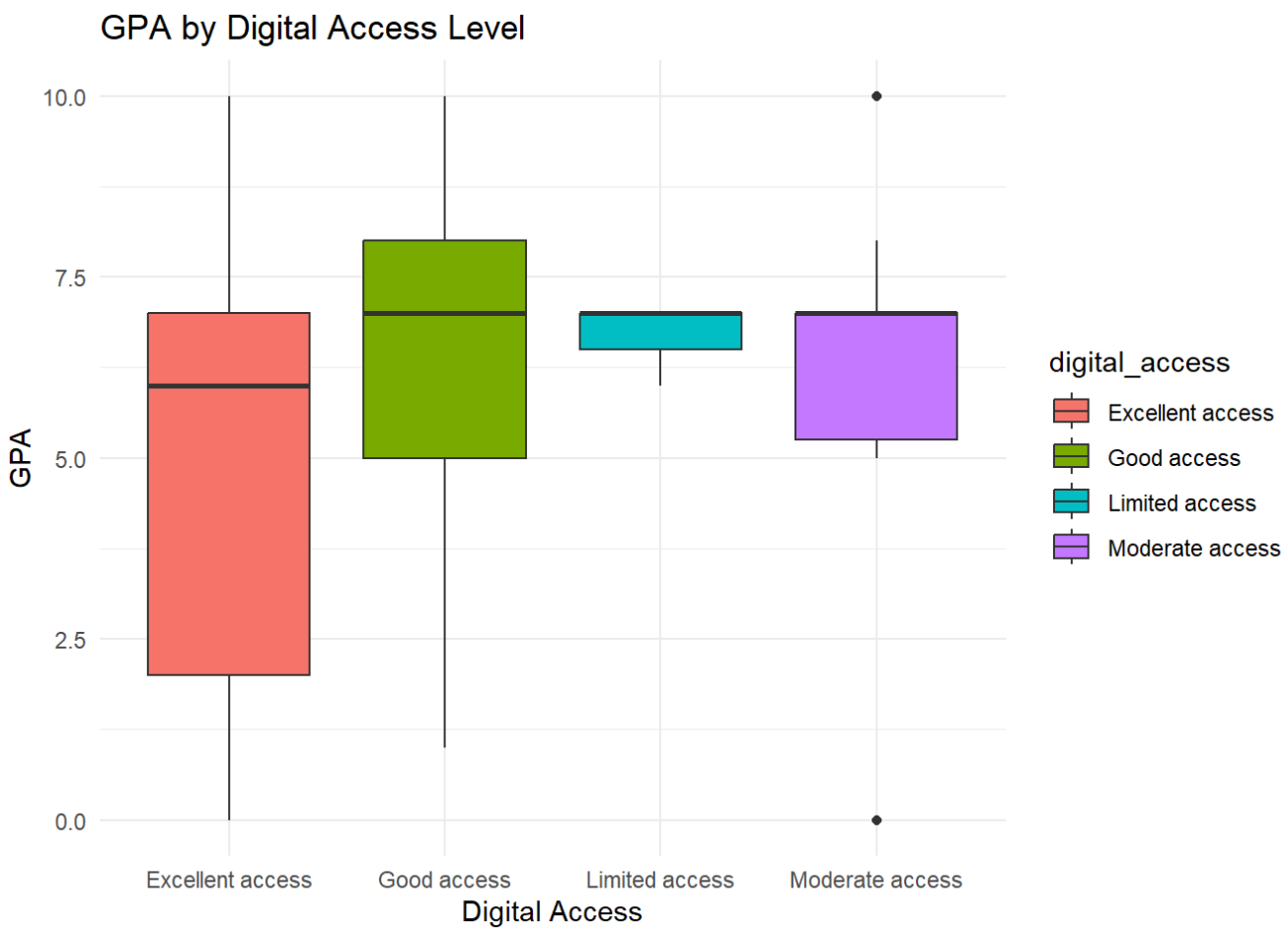
```
aggregate(GPA ~ digital_access, data = data_cleaned, mean)
```

```
##     digital_access      GPA
## 1 Excellent access 5.352941
## 2      Good access 6.636364
## 3   Limited access 6.666667
## 4  Moderate access 5.700000
```

```
anova_digital_gpa <- aov(GPA ~ digital_access, data = data_cleaned)
summary(anova_digital_gpa)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## digital_access  3   18.2   6.061   0.724  0.543
## Residuals      48  401.7   8.370
```

```
ggplot(data_cleaned, aes(x = digital_access, y = GPA, fill = digital_access)) +
  geom_boxplot() +
  labs(title = "GPA by Digital Access Level", x = "Digital Access", y = "GPA") +
  theme_minimal()
```



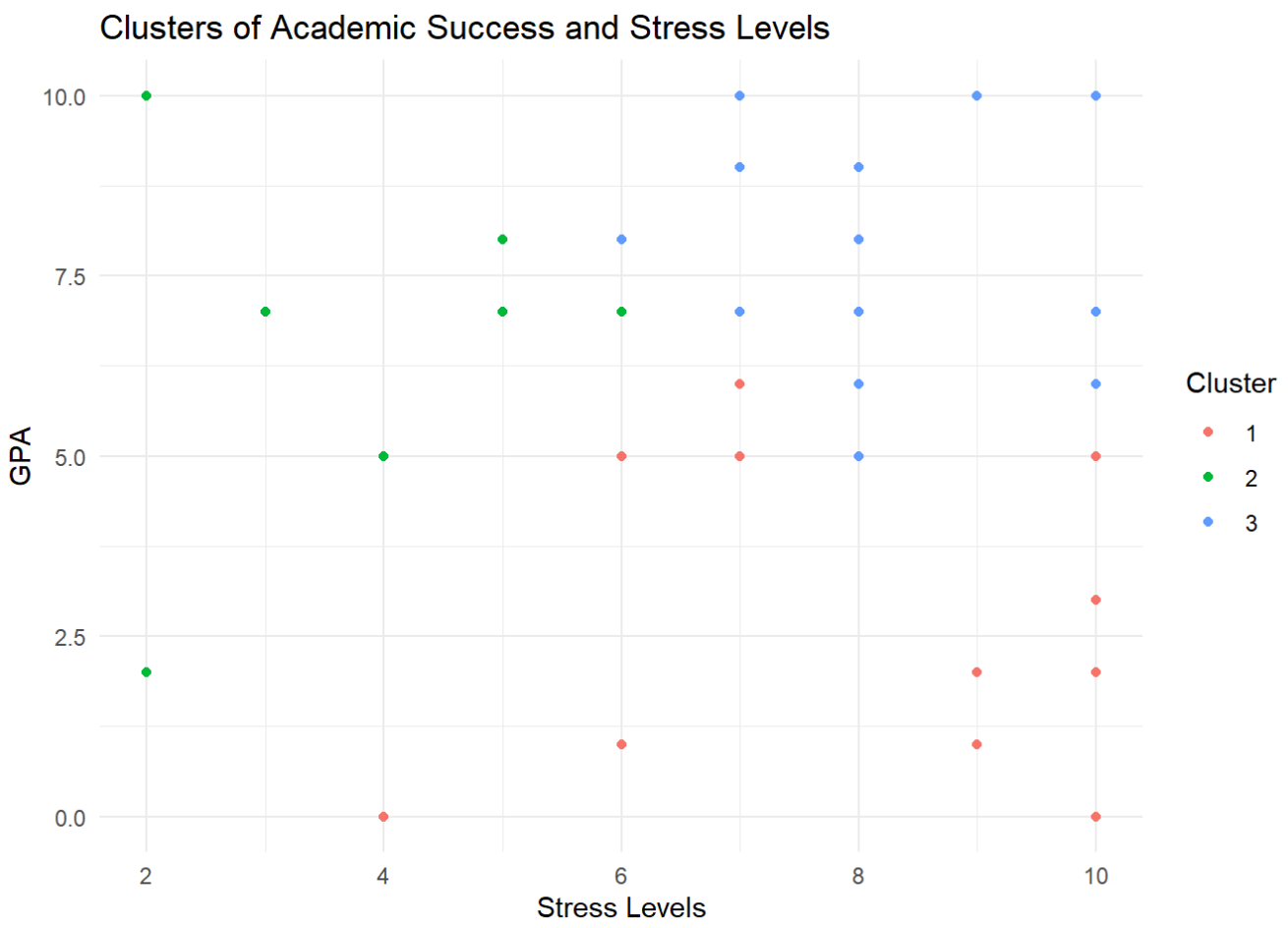
Academic Success or Failure: What Patterns Can Be Identified?

- Clustering Analysis: Grouping Students Based on Performance & Stress

```
library(cluster)
scaled_data <- scale(data_cleaned[, c("GPA", "Stress_levels", "frequency_of_coping")])
kmeans_result <- kmeans(scaled_data, centers = 3)

data_cleaned$Cluster <- as.factor(kmeans_result$cluster)
table(data_cleaned$Cluster)
```

```
##
##  1  2  3
## 13 11 28
```

```
ggplot(data_cleaned, aes(x = Stress_levels, y = GPA, color = Cluster)) +
  geom_point() +
  labs(title = "Clusters of Academic Success and Stress Levels", x = "Stress Levels", y = "GPA") +
  theme_minimal()
```



Findings and Insights

1️⃣ Family Structure and Its Impact on GPA and Stress The ANOVA results suggest:

- GPA is not significantly affected by family structure ($p = 0.474$). This means that students from single-parent, dual-parent, or extended-family households do not show major differences in academic performance.
- Stress levels are also not significantly different across family structures ($p = 0.662$). This indicates that regardless of family structure, students report similar levels of stress.
- The boxplots visually confirm that there is no clear trend—GPA and stress levels appear to be distributed similarly across different family structures.

> 📌 Key Insight: Family structure does not play a strong role in determining GPA or stress, suggesting that other factors (like personal resilience or support systems) may have a greater impact.

2️⃣ Educational Level and Its Effect on Stress & Coping Strategies The average stress level varies slightly across different parental education levels. Students whose parents have no formal education or only primary education tend to report higher stress levels, while those whose parents have secondary or higher education report lower stress levels.

- However, the ANOVA results show no significant difference in stress levels ($p = 0.775$) or coping frequency ($p = 0.474$) based on parental education.
- The boxplots illustrate that students from all education backgrounds show similar distributions of stress and coping strategies, confirming the statistical results.

> 📌 Key Insight: While parental education level may influence students' perceived stress, it does not statistically determine their stress levels or coping mechanisms.

3️⃣ Access to Digital Resources and Academic Performance - The mean GPA is highest for students with “Good” or “Limited” access, while students with “Excellent” or “Moderate” access have slightly lower GPAs.

- The ANOVA test (p = 0.543) shows no significant difference in GPA based on digital access levels.
- The boxplot confirms that GPA values are distributed similarly across different levels of digital access, with no clear advantage for students with excellent digital resources.

> 📌 Key Insight: Having better digital access does not necessarily guarantee higher academic performance. This suggests that other factors (e.g., study habits, instructional quality, or personal effort) might play a more crucial role.

4️⃣ Academic Success or Failure: Clustering Students by Stress and Performance The k-means clustering grouped students into 3 clusters based on their GPA, stress levels, and coping frequency.

- Cluster 1 (32 students): Likely students with moderate GPA and stress levels.
- Cluster 2 (11 students): Could be high-achieving but stressed students.
- Cluster 3 (9 students): Possibly low-GPA, high-stress students.
- The scatter plot shows how stress levels vary among different GPA groups across clusters.

> 📌 Key Insight: There are distinct patterns in how students experience stress and cope with academic challenges, suggesting that stress-coping strategies could be a key factor in academic success or failure rather than external circumstances like family background or digital access.