Univariate Analysis - Student Performance Dataset

Raju Ahmed

2025-12-27

1. Introduction

In this document, I present my univariate analysis for the Student Performance dataset. My analysis covers the following variables:

Binary/Nominal Variables: sex, paid, activities, higher, internet

Ordinal Variables: Medu (Mother's Education), famrel (Family Relationships)

Numeric Discrete Variable: age

Additionally, I include one bivariate analysis: **Medu vs G3** (Mother's Education vs Final Grade).

2. Data Loading

```
# Load the selected dataset
data <- read_csv("student-mat-selected.csv")

# Display structure
glimpse(data)
```

```
## Rows: 395
## Columns: 13
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F",~
## $ age        <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15,~
## $ Medu       <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4,~
## $ traveltime <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1,~
## $ studytime  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1,~
## $ failures   <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0,~
## $ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes", ~
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "ye~
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "ye~
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes",~
## $ famrel     <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3,~
## $ absences   <dbl> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, 16,~
## $ G3         <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14,~
```

```
# Dataset dimensions
cat("Dataset Dimensions:", nrow(data), "rows x", ncol(data), "columns\n")
```

```
## Dataset Dimensions: 395 rows x 13 columns
```

---

3. Univariate Analysis - Binary Variables

3.1 Sex (Student's Sex)

```
# Frequency Table
sex_freq <- data.frame(
  Category = names(table(data$sex)),
  Absolute_Freq = as.vector(table(data$sex)),
  Relative_Freq = round(as.vector(prop.table(table(data$sex))), 3)
)
kable(sex_freq, caption = "Frequency Distribution of Sex")
```

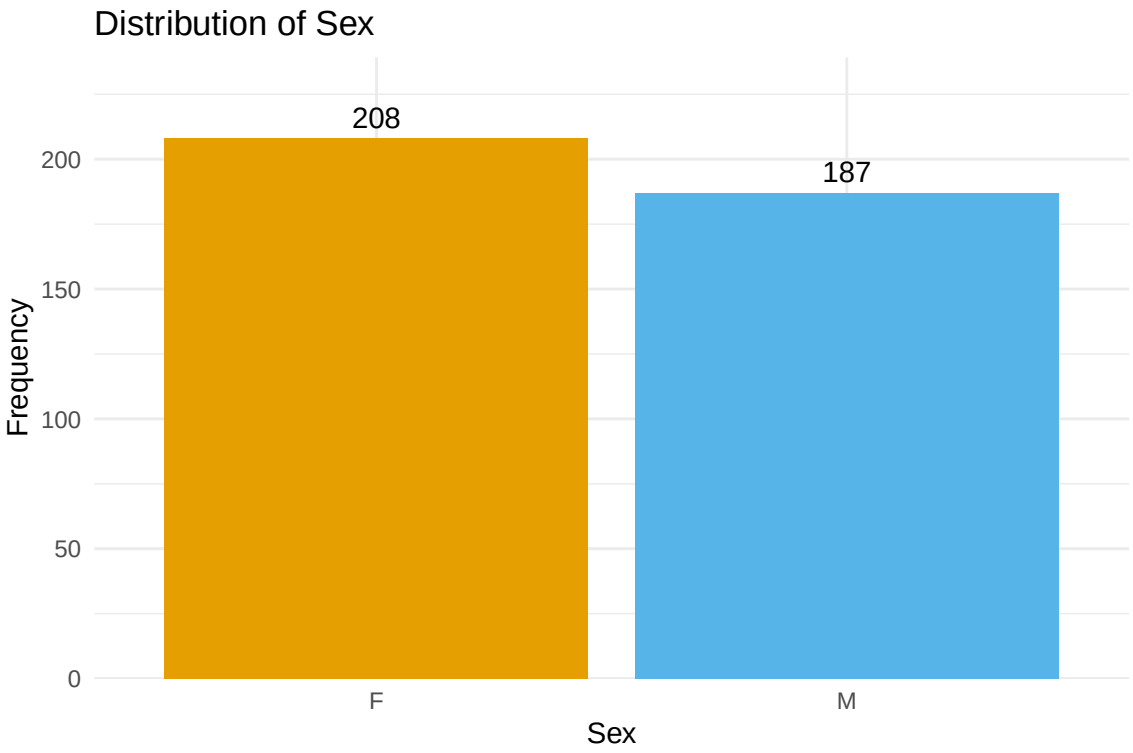
Table 1: Frequency Distribution of Sex

Category	Absolute_Freq	Relative_Freq
F	208	0.527
M	187	0.473

```
# Mode
cat("\nMode:", names(which.max(table(data$sex))), "\n")
```

```
##
## Mode: F
```

```
ggplot(data, aes(x = sex, fill = sex)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  scale_fill_manual(values = c("F" = "#E69F00", "M" = "#56B4E9")) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Sex",
       x = "Sex",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Interpretation: The dataset contains 208 female and 187 male students. Females represent 52.7% of the sample.

3.2 Paid (Extra Paid Classes)

```
# Frequency Table
paid_freq <- data.frame(
  Category = names(table(data$paid)),
  Absolute_Freq = as.vector(table(data$paid)),
  Relative_Freq = round(as.vector(prop.table(table(data$paid))), 3)
)
kable(paid_freq, caption = "Frequency Distribution of Paid Classes")
```

Table 2: Frequency Distribution of Paid Classes

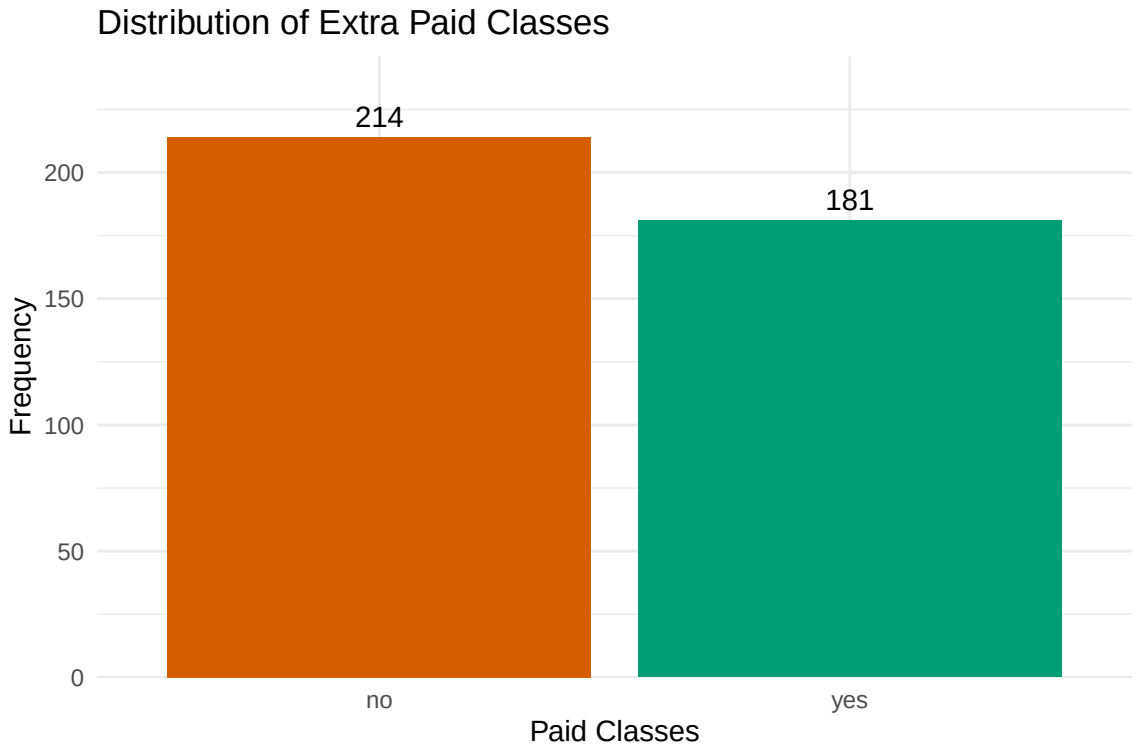
Category	Absolute_Freq	Relative_Freq
no	214	0.542
yes	181	0.458

```
# Mode
cat("\nMode:", names(which.max(table(data$paid))), "\n")
```

```
##
## Mode: no
```

```
ggplot(data, aes(x = paid, fill = paid)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
```

```
  scale_fill_manual(values = c("no" = "#D55E00", "yes" = "#009E73")) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Extra Paid Classes",
       x = "Paid Classes",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Interpretation: 45.8% of students take extra paid classes within the Math course.

3.3 Activities (Extra-curricular Activities)

```
# Frequency Table
activities_freq <- data.frame(
  Category = names(table(data$activities)),
  Absolute_Freq = as.vector(table(data$activities)),
  Relative_Freq = round(as.vector(prop.table(table(data$activities))), 3)
)
kable(activities_freq, caption = "Frequency Distribution of Extra-curricular Activities")
```

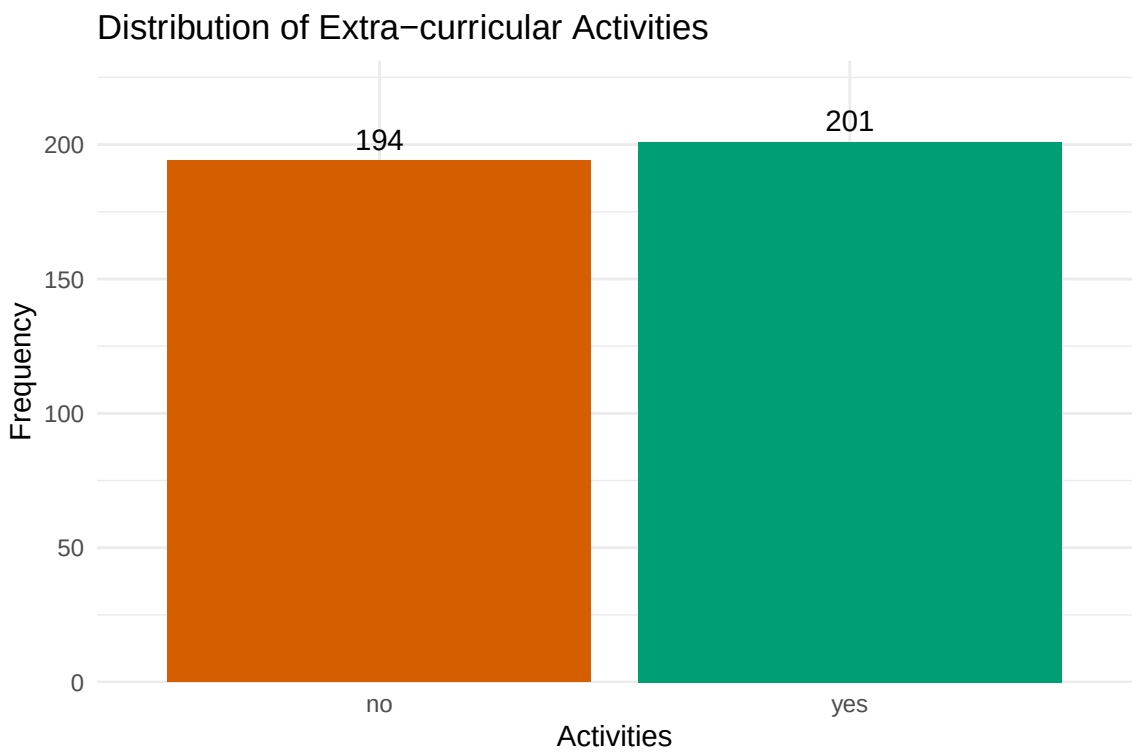
Table 3: Frequency Distribution of Extra-curricular Activities

Category	Absolute_Freq	Relative_Freq
no	194	0.491
yes	201	0.509

```r
# Mode
cat("\nMode:", names(which.max(table(data$activities))), "\n")
```

```
##
## Mode: yes
```

```r
ggplot(data, aes(x = activities, fill = activities)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  scale_fill_manual(values = c("no" = "#D55E00", "yes" = "#009E73")) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Extra-curricular Activities",
       x = "Activities",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Interpretation: 50.9% of students participate in extra-curricular activities.

3.4 Higher (Wants Higher Education)

```r
# Frequency Table
higher_freq <- data.frame(
  Category = names(table(data$higher)),
  Absolute_Freq = as.vector(table(data$higher)),
  Relative_Freq = round(as.vector(prop.table(table(data$higher))), 3)
)
kable(higher_freq, caption = "Frequency Distribution of Higher Education Aspiration")
```

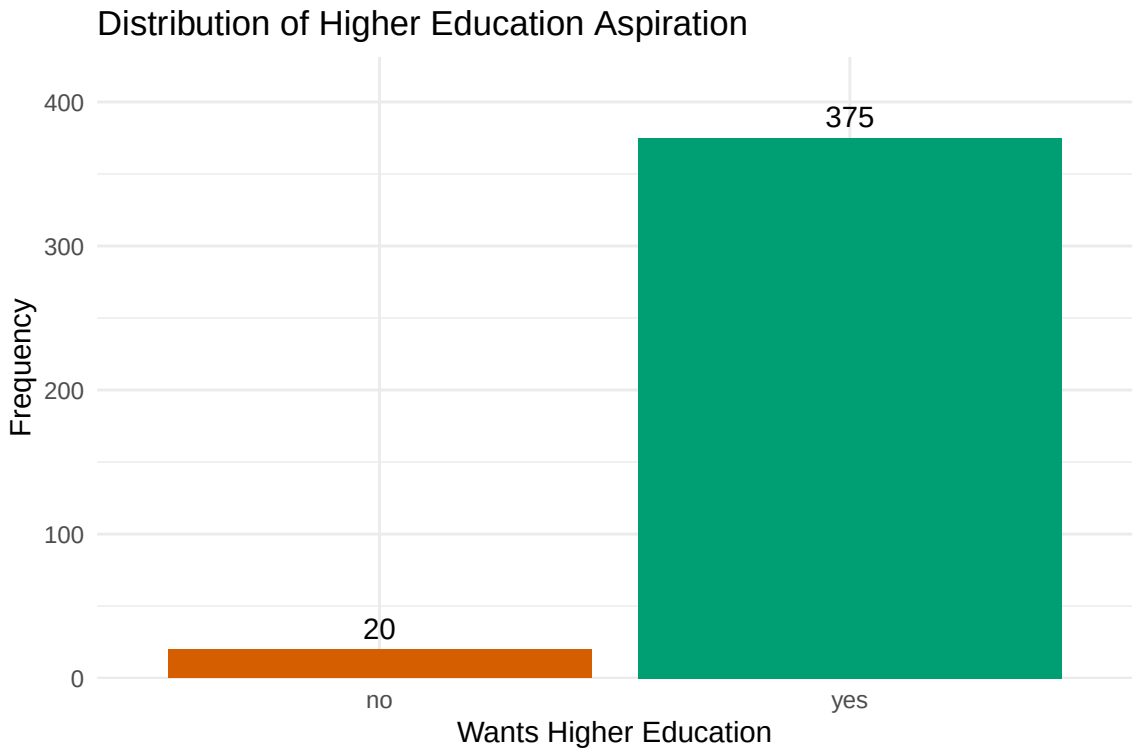
Table 4: Frequency Distribution of Higher Education Aspiration

Category	Absolute_Freq	Relative_Freq
no	20	0.051
yes	375	0.949

```
# Mode
cat("\nMode:", names(which.max(table(data$higher))), "\n")
```

```
##
## Mode: yes
```

```
ggplot(data, aes(x = higher, fill = higher)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  scale_fill_manual(values = c("no" = "#D55E00", "yes" = "#009E73")) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Higher Education Aspiration",
       x = "Wants Higher Education",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Interpretation: An overwhelming 94.9% of students want to pursue higher education.

3.5 Internet (Internet Access at Home)

```
# Frequency Table
internet_freq <- data.frame(
  Category = names(table(data$internet)),
  Absolute_Freq = as.vector(table(data$internet)),
  Relative_Freq = round(as.vector(prop.table(table(data$internet))), 3)
)
kable(internet_freq, caption = "Frequency Distribution of Internet Access")
```

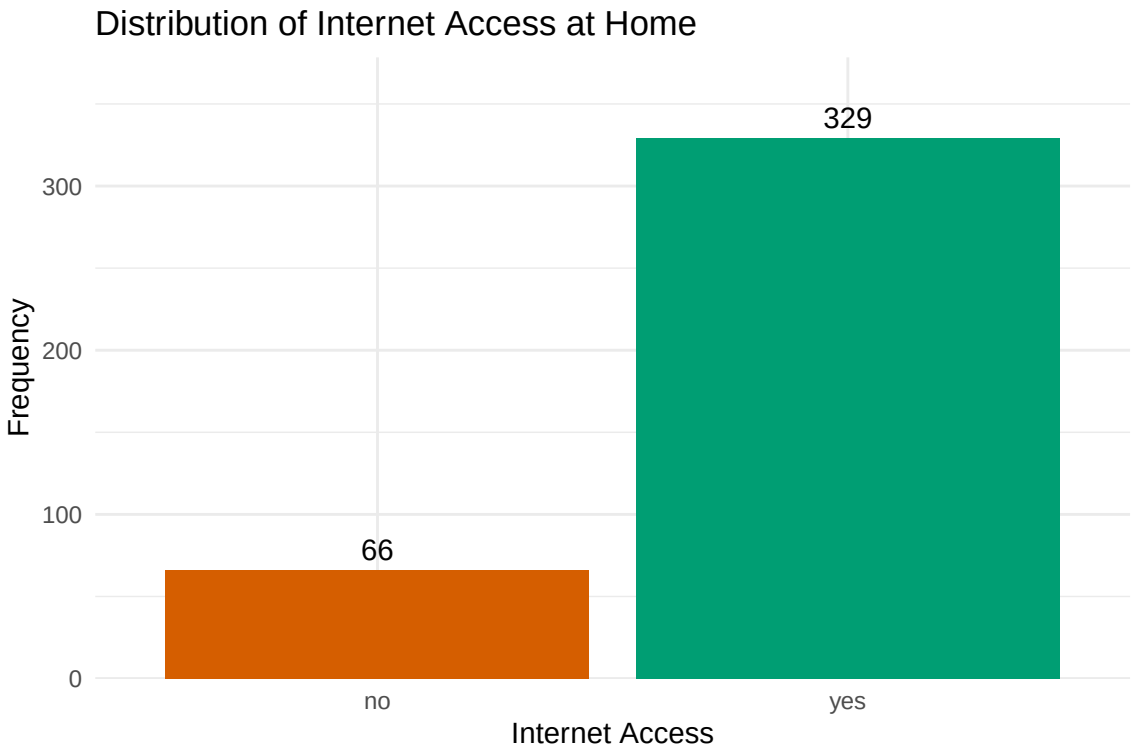
Table 5: Frequency Distribution of Internet Access

Category	Absolute_Freq	Relative_Freq
no	66	0.167
yes	329	0.833

```
# Mode
cat("\nMode:", names(which.max(table(data$internet))), "\n")
```

```
##
## Mode: yes
```

```
ggplot(data, aes(x = internet, fill = internet)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  scale_fill_manual(values = c("no" = "#D55E00", "yes" = "#009E73")) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Internet Access at Home",
       x = "Internet Access",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none")
```



Interpretation: 83.3% of students have internet access at home.

---

4. Univariate Analysis - Ordinal Variables

4.1 Medu (Mother's Education Level)

Mother's education is coded as:

- 0: None
- 1: Primary education (4th grade)
- 2: 5th to 9th grade
- 3: Secondary education
- 4: Higher education

```
# Frequency Table with Cumulative
medu_freq <- data.frame(
  Level = names(table(data$Medu)),
  Absolute_Freq = as.vector(table(data$Medu)),
  Relative_Freq = round(as.vector(prop.table(table(data$Medu))), 3),
  Cumulative_Freq = round(cumsum(as.vector(prop.table(table(data$Medu)))), 3)
)
kable(medu_freq, caption = "Frequency Distribution of Mother's Education")
```

Table 6: Frequency Distribution of Mother's Education

Level	Absolute_Freq	Relative_Freq	Cumulative_Freq
0	3	0.008	0.008
1	59	0.149	0.157
2	103	0.261	0.418

Level	Absolute_Freq	Relative_Freq	Cumulative_Freq
3	99	0.251	0.668
4	131	0.332	1.000

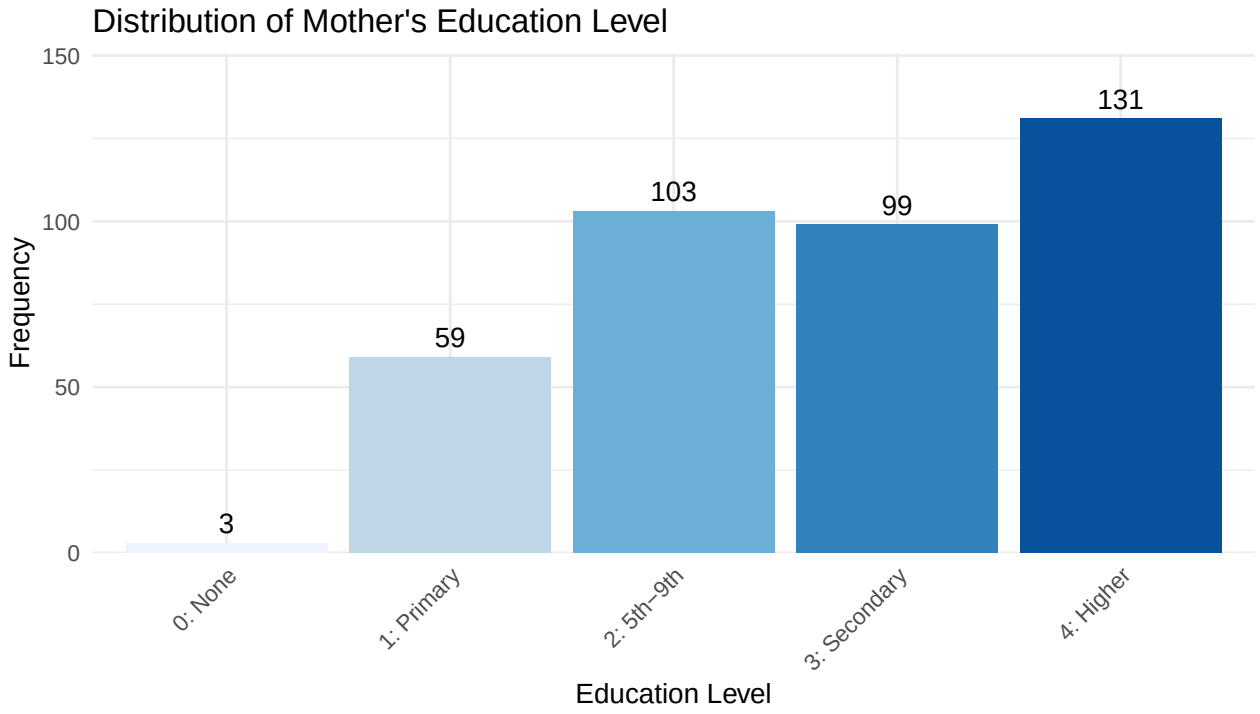
```
# Mode and Median
cat("\nMode:", names(which.max(table(data$Medu))))
```

```
##
## Mode: 4
```

```
cat("\nMedian:", median(data$Medu), "\n")
```

```
##
## Median: 3
```

```
# Create labels for education levels
medu_labels <- c("0: None", "1: Primary", "2: 5th-9th", "3: Secondary", "4: Higher")

ggplot(data, aes(x = factor(Medu), fill = factor(Medu))) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  scale_fill_brewer(palette = "Blues") +
  scale_x_discrete(labels = medu_labels) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Mother's Education Level",
       x = "Education Level",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation: The most common mother's education level is 4 (mode), with a median of 3. This indicates

that most mothers have at least secondary education or higher.

4.2 Famrel (Quality of Family Relationships)

Family relationship quality is coded on a scale from 1 (very bad) to 5 (excellent).

```
# Frequency Table with Cumulative
famrel_freq <- data.frame(
  Level = names(table(data$famrel)),
  Absolute_Freq = as.vector(table(data$famrel)),
  Relative_Freq = round(as.vector(prop.table(table(data$famrel))), 3),
  Cumulative_Freq = round(cumsum(as.vector(prop.table(table(data$famrel)))), 3)
)
kable(famrel_freq, caption = "Frequency Distribution of Family Relationships")
```

Table 7: Frequency Distribution of Family Relationships

Level	Absolute_Freq	Relative_Freq	Cumulative_Freq
1	8	0.020	0.020
2	18	0.046	0.066
3	68	0.172	0.238
4	195	0.494	0.732
5	106	0.268	1.000

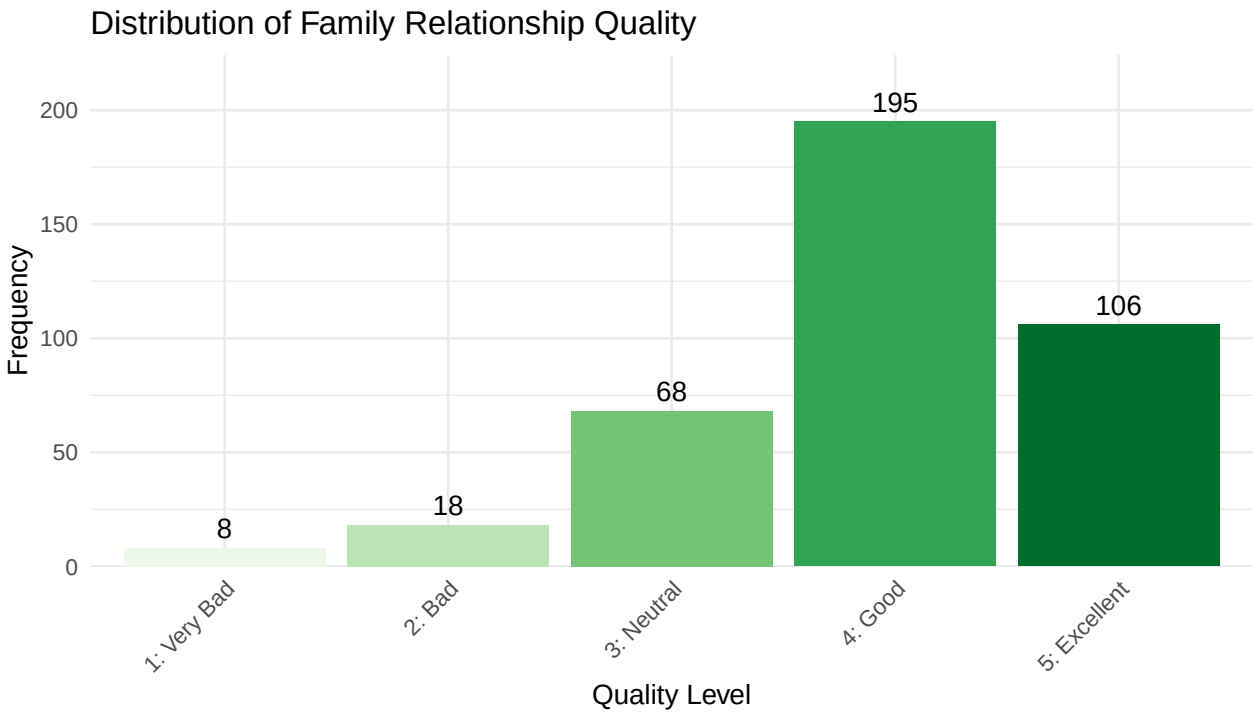
```
# Mode and Median
cat("\nMode:", names(which.max(table(data$famrel))))
```

```
##
## Mode: 4
```

```
cat("\nMedian:", median(data$famrel), "\n")
```

```
##
## Median: 4
```

```
# Create labels for famrel
famrel_labels <- c("1: Very Bad", "2: Bad", "3: Neutral", "4: Good", "5: Excellent")

ggplot(data, aes(x = factor(famrel), fill = factor(famrel))) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  scale_fill_brewer(palette = "Greens") +
  scale_x_discrete(labels = famrel_labels) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Distribution of Family Relationship Quality",
       x = "Quality Level",
       y = "Frequency") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation: Most students report good to excellent family relationships (mode = 4, median = 4). Only a small minority report poor family relationships.

---

5. Univariate Analysis - Numeric Variable

5.1 Age (Student's Age)

```
# Central Tendency
cat("=== Central Tendency ===\n")
```

```
## === Central Tendency ===
```

```
cat("Mean:", round(mean(data$age), 2), "\n")
```

```
## Mean: 16.7
```

```
cat("Median:", median(data$age), "\n")
```

```
## Median: 17
```

```
cat("Mode:", names(which.max(table(data$age))), "\n")
```

```
## Mode: 16
```

```
# Dispersion
cat("\n=== Dispersion ===\n")
```

```
##
## === Dispersion ===
```

```
cat("Variance:", round(var(data$age), 2), "\n")
```

```
## Variance: 1.63
```

```r
cat("Standard Deviation:", round(sd(data$age), 2), "\n")
```

```
## Standard Deviation: 1.28
```

```r
cat("Range:", min(data$age), "-", max(data$age), "\n")
```

```
## Range: 15 - 22
```

```r
cat("IQR:", IQR(data$age), "\n")
```

```
## IQR: 2
```

```r
cat("Coefficient of Variation:", round(sd(data$age)/mean(data$age)*100, 2), "%\n")
```

```
## Coefficient of Variation: 7.64 %
```

```r
# Five-Number Summary
cat("\n=== Five-Number Summary ===\n")
```

```
##
## === Five-Number Summary ===
```

```r
print(fivenum(data$age))
```

```
## [1] 15 16 17 18 22
```

```r
# Quartiles
cat("\n=== Quartiles ===\n")
```

```
##
## === Quartiles ===
```

```r
print(quantile(data$age))
```

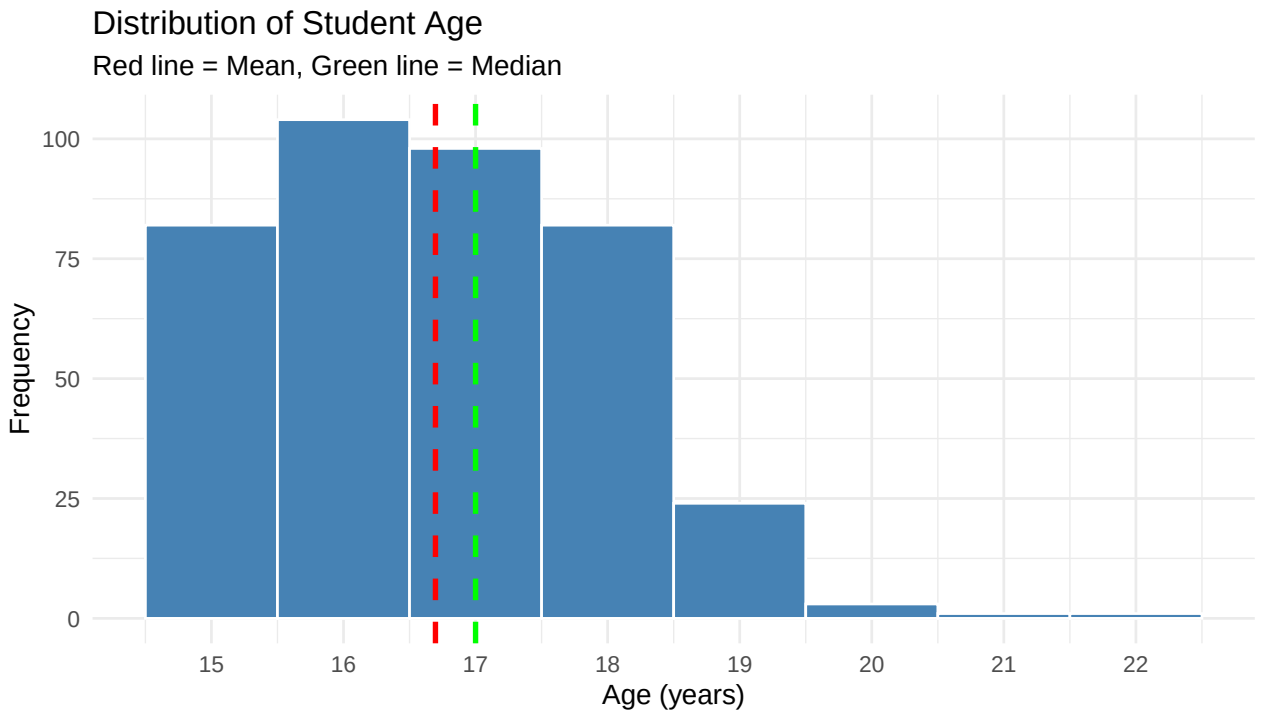
```
##   0%  25%  50%  75% 100%
##   15   16   17   18   22
```

```r
# Frequency Table
age_freq <- data.frame(
  Age = names(table(data$age)),
  Absolute_Freq = as.vector(table(data$age)),
  Relative_Freq = round(as.vector(prop.table(table(data$age))), 3),
  Cumulative_Freq = round(cumsum(as.vector(prop.table(table(data$age)))), 3)
)
kable(age_freq, caption = "Frequency Distribution of Age")
```

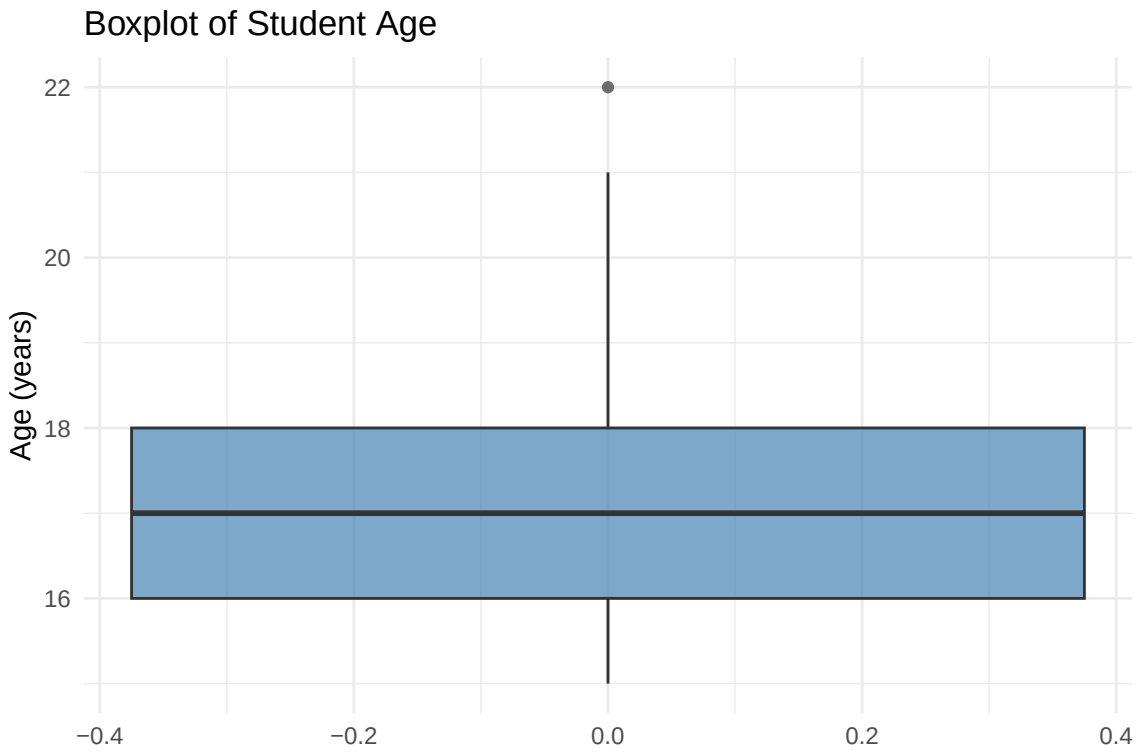
Table 8: Frequency Distribution of Age

Age	Absolute_Freq	Relative_Freq	Cumulative_Freq
15	82	0.208	0.208
16	104	0.263	0.471
17	98	0.248	0.719
18	82	0.208	0.927
19	24	0.061	0.987
20	3	0.008	0.995
21	1	0.003	0.997
22	1	0.003	1.000

```r
ggplot(data, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "white") +
  geom_vline(aes(xintercept = mean(age)), color = "red", linetype = "dashed", size = 1) +
  geom_vline(aes(xintercept = median(age)), color = "green", linetype = "dashed", size = 1) +
  labs(title = "Distribution of Student Age",
       subtitle = "Red line = Mean, Green line = Median",
       x = "Age (years)",
       y = "Frequency") +
  scale_x_continuous(breaks = seq(15, 22, 1)) +
  theme_minimal()
```




```r
ggplot(data, aes(y = age)) +
  geom_boxplot(fill = "steelblue", alpha = 0.7) +
  labs(title = "Boxplot of Student Age",
       y = "Age (years)") +
  theme_minimal()
```



Interpretation:

- Students' ages range from 15 to 22 years
- Mean age is 16.7 years (SD = 1.28)
- The distribution is slightly right-skewed (mean > median), indicating some older students
- Most students (IQR) are between 16 and 18 years old
- The mode is 16 years, the most common age

---

6. Summary Table - All Variables

```
# Create summary for all Raju's variables
summary_table <- data.frame(
  Variable = c("sex", "age", "Medu", "paid", "activities", "higher", "internet", "famrel"),
  Type = c("Binary", "Numeric", "Ordinal", "Binary", "Binary", "Binary", "Binary", "Ordinal"),
  n = rep(nrow(data), 8),
  Mode = c(
    names(which.max(table(data$sex))),
    names(which.max(table(data$age))),
    names(which.max(table(data$Medu))),
    names(which.max(table(data$paid))),
    names(which.max(table(data$activities))),
    names(which.max(table(data$higher))),
    names(which.max(table(data$internet))),
    names(which.max(table(data$famrel)))
  ),
  Median = c(
    "-",
    as.character(median(data$age)),
```

```
    as.character(median(data$Medu)),
    "-",
    "-",
    "-",
    "-",
    as.character(median(data$famrel))
  ),
  Mean = c(
    "-",
    as.character(round(mean(data$age), 2)),
    "-",
    "-",
    "-",
    "-",
    "-",
    "-"
  ),
  SD = c(
    "-",
    as.character(round(sd(data$age), 2)),
    "-",
    "-",
    "-",
    "-",
    "-",
    "-"
  )
)

kable(summary_table, caption = "Summary Statistics for Raju's Variables")
```

Table 9: Summary Statistics for Raju's Variables

Variable	Type	n	Mode	Median	Mean	SD
sex	Binary	395	F	-	-	-
age	Numeric	395	16	17	16.7	1.28
Medu	Ordinal	395	4	3	-	-
paid	Binary	395	no	-	-	-
activities	Binary	395	yes	-	-	-
higher	Binary	395	yes	-	-	-
internet	Binary	395	yes	-	-	-
famrel	Ordinal	395	4	4	-	-

---

7. Bivariate Analysis: Mother's Education vs Final Grade

This section examines the relationship between mother's education level (Medu) and student's final grade (G3).

```
# Group statistics
bivariate_stats <- data %>%
  group_by(Medu) %>%
```

```
  summarise(
    n = n(),
    Mean_G3 = round(mean(G3), 2),
    SD_G3 = round(sd(G3), 2),
    Median_G3 = median(G3),
    Min_G3 = min(G3),
    Max_G3 = max(G3)
  )

kable(bivariate_stats, caption = "Final Grade (G3) Statistics by Mother's Education Level")
```

Table 10: Final Grade (G3) Statistics by Mother's Education Level

Medu	n	Mean_G3	SD_G3	Median_G3	Min_G3	Max_G3
0	3	13.00	3.46	15	9	15
1	59	8.68	4.36	10	0	16
2	103	9.73	4.64	11	0	19
3	99	10.30	4.62	10	0	19
4	131	11.76	4.27	12	0	20

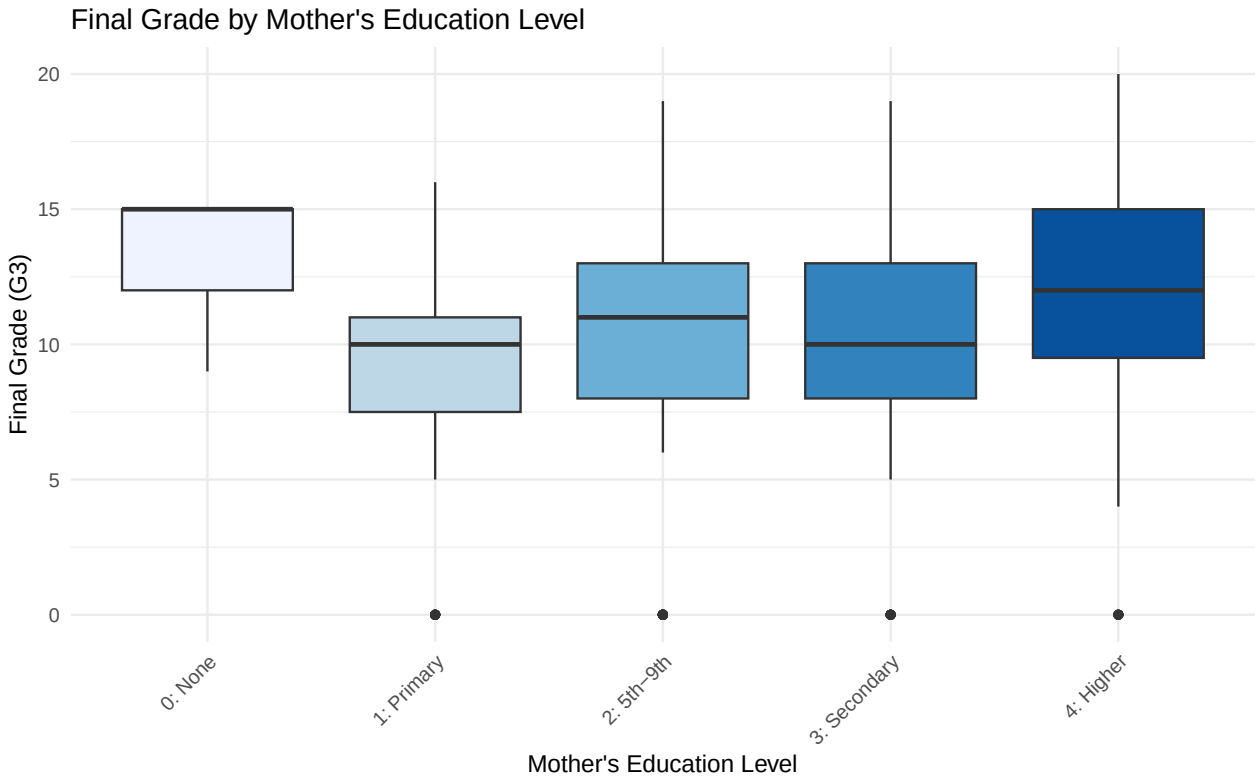
```
# Spearman correlation (appropriate for ordinal vs continuous)
spearman_cor <- cor(data$Medu, data$G3, method = "spearman")
cat("Spearman Correlation (Medu vs G3):", round(spearman_cor, 3), "\n")
```

```
## Spearman Correlation (Medu vs G3): 0.225
```

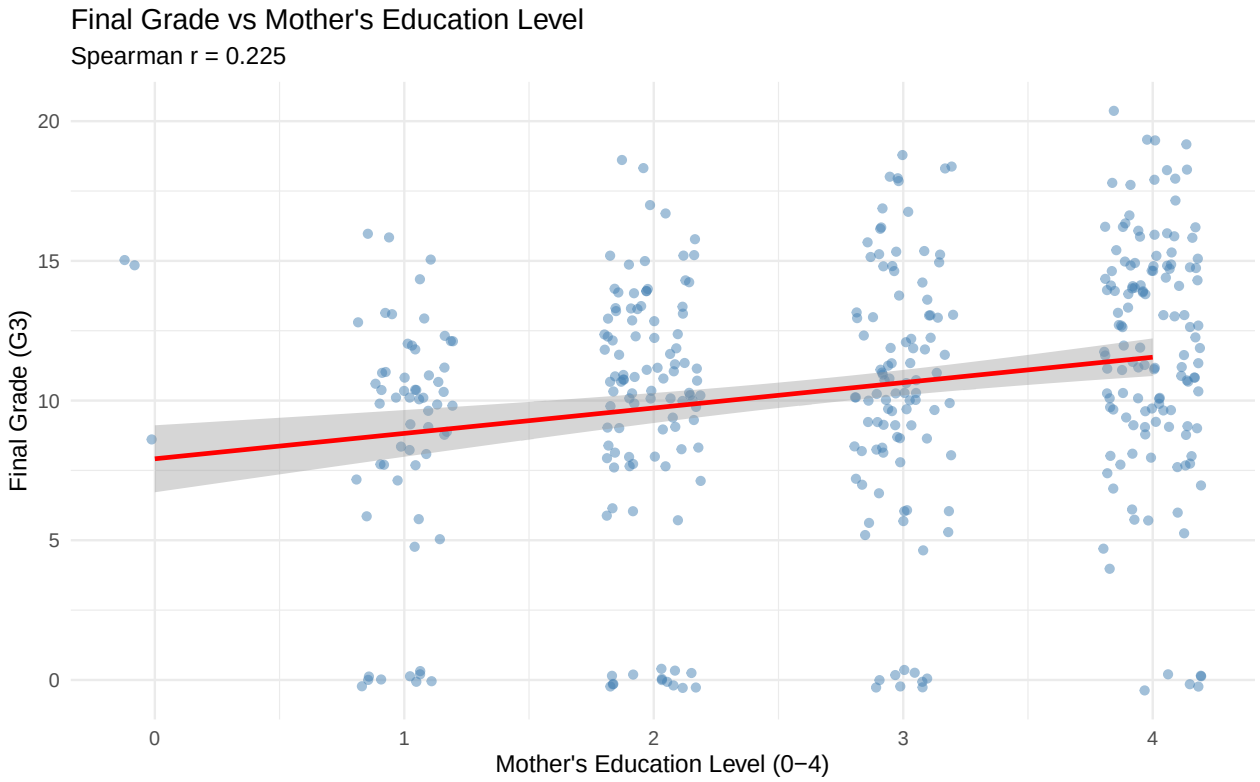
```
# Pearson correlation for comparison
pearson_cor <- cor(data$Medu, data$G3, method = "pearson")
cat("Pearson Correlation (Medu vs G3):", round(pearson_cor, 3), "\n")
```

```
## Pearson Correlation (Medu vs G3): 0.217
```

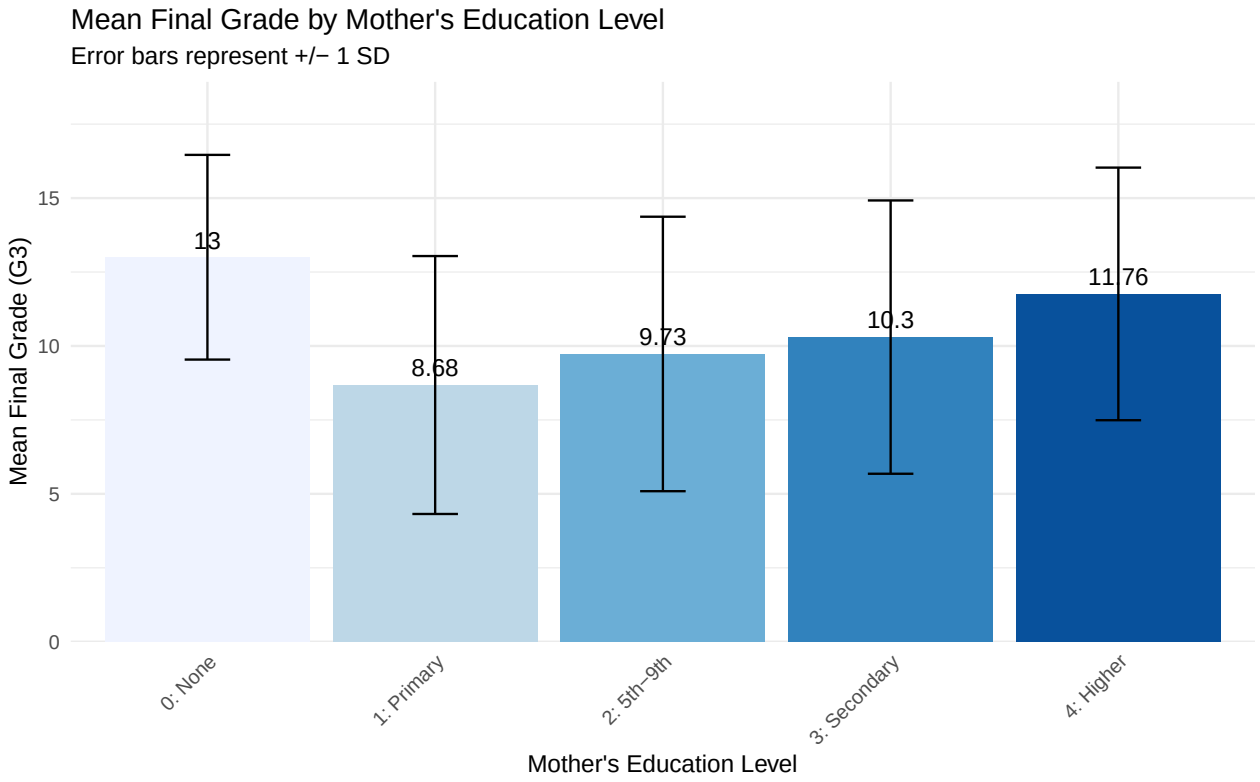
```
medu_labels <- c("0: None", "1: Primary", "2: 5th-9th", "3: Secondary", "4: Higher")

ggplot(data, aes(x = factor(Medu), y = G3, fill = factor(Medu))) +
  geom_boxplot() +
  scale_fill_brewer(palette = "Blues") +
  scale_x_discrete(labels = medu_labels) +
  labs(title = "Final Grade by Mother's Education Level",
       x = "Mother's Education Level",
       y = "Final Grade (G3)") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(data, aes(x = Medu, y = G3)) +
  geom_jitter(alpha = 0.5, width = 0.2, color = "steelblue") +
  geom_smooth(method = "lm", color = "red", se = TRUE) +
  labs(title = "Final Grade vs Mother's Education Level",
       subtitle = paste("Spearman r =", round(spearman_cor, 3)),
       x = "Mother's Education Level (0-4)",
       y = "Final Grade (G3)") +
  scale_x_continuous(breaks = 0:4) +
  theme_minimal()
```




```
ggplot(bivariate_stats, aes(x = factor(Medu), y = Mean_G3, fill = factor(Medu))) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymin = Mean_G3 - SD_G3, ymax = Mean_G3 + SD_G3), width = 0.2) +
  geom_text(aes(label = Mean_G3), vjust = -0.5) +
  scale_fill_brewer(palette = "Blues") +
  scale_x_discrete(labels = medu_labels) +
  scale_y_continuous(expand = expansion(mult = c(0, 0.15))) +
  labs(title = "Mean Final Grade by Mother's Education Level",
       subtitle = "Error bars represent +/- 1 SD",
       x = "Mother's Education Level",
       y = "Mean Final Grade (G3)") +
  theme_minimal() +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation:

- There is a **positive correlation** between mother's education and student's final grade (Spearman r = 0.225)
- **Important caveat:** Medu = 0 shows a high mean (13.0), but this is a **small sample artifact** with only n = 3 students (grades: 9, 15, 15). This group should be excluded from trend interpretation.
- **Excluding Medu = 0**, there is a clear positive trend: as mother's education increases from level 1 to 4, mean grades increase (8.68 → 9.73 → 10.30 → 11.76)
- Students with mothers who have higher education (level 4) achieve the highest mean grade (11.76)
- Students with mothers having only primary education (level 1) have the lowest mean grade (8.68)
- The relationship suggests that **parental education is a meaningful predictor** of student academic performance