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 covering: **Binary Variables** (sex, paid, activities, higher, internet), **Ordinal Variables** (Medu, famrel), **Numeric Variable** (age), and one **Bivariate Analysis** (Medu vs G3).

2. Data Loading

```
data <- read_csv("student-mat-selected.csv")
cat("Dataset:", nrow(data), "rows x", ncol(data), "columns\n")
```

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

3. Binary Variables

Table 1: Binary Variables Summary

Variable	Yes/F (n)	Yes/F (%)	No/M (n)	Mode
sex (F/M)	208	52.7	187	F
paid	181	45.8	214	no
activities	201	50.9	194	yes
higher	375	94.9	20	yes
internet	329	83.3	66	yes

Interpretation: Gender is balanced (52.7% female, 47.3% male). Less than half (45.8%) take paid Math classes. Extracurricular participation is evenly split (50.9%). A striking 94.9% aspire to higher education. Internet access is available to 83.3% of students.

4. Ordinal Variables

Table 2: Ordinal Variables Summary

Variable	Mode	Median	Distribution
Medu (0-4)	4	3	3>59, 103>99>131 (skewed to higher)
famrel (1-5)	4	4	4>195, 5>86 (mostly good)

Interpretation: Mother's education (Medu) is skewed toward higher levels with mode=4 (higher education, n=131) and median=3 (secondary). Only 3 mothers have no formal education. Family relationship quality
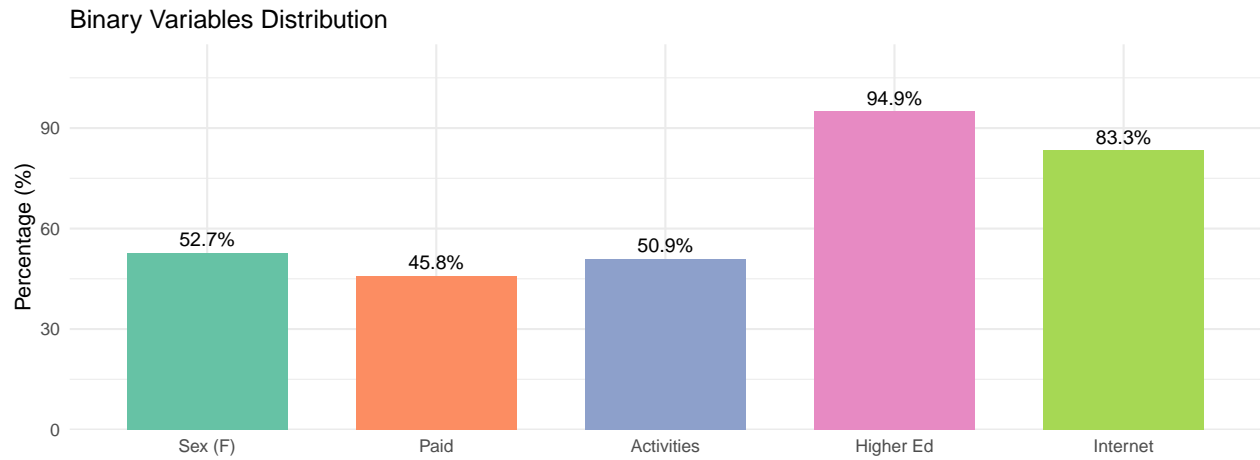
1

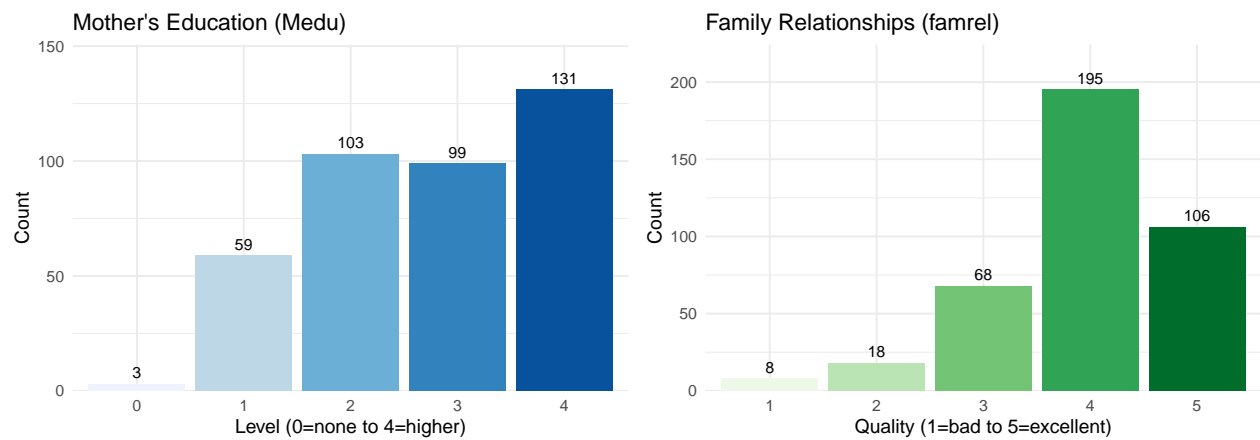Figure 1: Distribution of Binary Variables - Percentage of Yes/Female responses

Figure 2: Distribution of Ordinal Variables - Mother's Education (left) and Family Relationships (right)

(famrel) is predominantly positive with mode=4 (good, n=195) and median=4. Over 71% report good to excellent relationships, suggesting supportive home environments.

5. Numeric Variable: Age

Mean: 16.7 | Median: 17 | Mode: 16 | SD: 1.28 | Range: 15 - 22 | IQR: 2
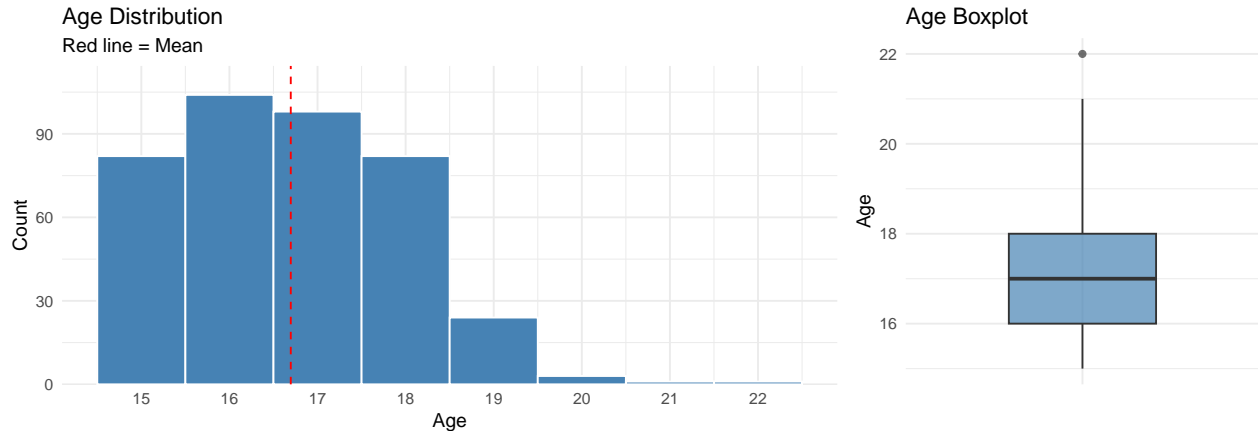

Figure 3: Age Distribution - Histogram with mean line (left) and Boxplot (right)

```
## TableGrob (1 x 2) "arrange": 2 grobs
##   z     cells      name              grob
## 1 1 (1-1,1-1) arrange gtable[layout]
## 2 2 (1-1,2-2) arrange gtable[layout]
```

Interpretation: Ages range from 15-22 years with mean=16.70, median=17, mode=16, and SD=1.28. The distribution is slightly right-skewed with most students in the typical 15-18 age range. Older students (19-22) may have repeated grades. The IQR of 2 years confirms low variability, with potential outliers at the upper end.

6. Bivariate: Mother's Education vs Final Grade

Table 3: Final Grade (G3) Statistics by Mother's Education Level

Medu	n	Mean_G3	SD
0	3	13.00	3.46
1	59	8.68	4.36
2	103	9.73	4.64
3	99	10.30	4.62
4	131	11.76	4.27

Spearman r = 0.225

Interpretation: Spearman correlation r=0.225 indicates a weak positive relationship between mother's education and final grades. The high mean for Medu=0 (13.0) is a small sample artifact (n=3). Excluding this group, grades increase consistently from 8.68 (primary) to 11.76 (higher education), suggesting mother's education is a meaningful predictor of student performance.
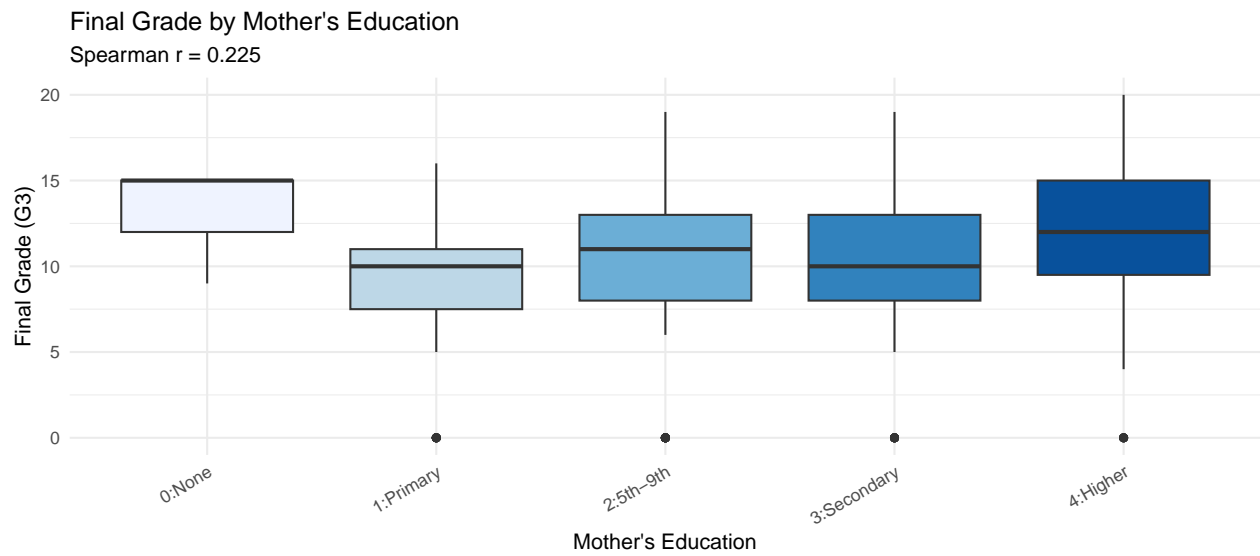
Figure 4: Boxplot of Final Grades by Mother's Education Level with Spearman Correlation