Univariate Analysis - Student Performance Dataset

Kezia Fernandes, Raju Ahmed

2025-12-30

Contents

1 Introduction

This dataset originates from the secondary education domain and focuses on analyzing factors associated with student academic performance in Mathematics at two Portuguese secondary schools (Cortez 2008; Cortez and Silva 2008). The data capture multiple dimensions of a student's profile, combining academic outcomes, demographic characteristics, and socio-educational factors. Information was collected through a combination of school records (such as grades and absences) and student questionnaires, providing both objective and self-reported measures relevant to educational performance.

For this analysis, a subset of 13 variables was selected to reflect key aspects influencing student achievement while maintaining analytical clarity. These variables include demographic attributes (sex, age), family and background indicators (mother's education level, quality of family relationships), school-related factors (study time, travel time, past failures, absences), support and engagement variables (paid classes, extracurricular activities, internet access), educational aspirations (desire for higher education), and the final Mathematics grade (G3) as the outcome variable.

The dataset contains a mix of binary nominal variables (e.g., sex, internet access), ordinal categorical variables (e.g., study time, travel time, family relationship quality), and numeric discrete variables (e.g., age, failures, absences). The final grade (G3), measured on a scale from 0 to 20, is treated as a continuous numeric variable. This structure makes the dataset well suited for univariate statistical analysis, allowing for an initial exploration of distributions, central tendencies, and variability across different types of educational and socio-demographic factors.

2 Continuous Variable: G3 (Final Grade)

2.1 Descriptive Statistics

Total = 395 | Mean = 10.42 | Median = 11 | Mode = 10 | SD = 4.58 | Variance = 20.99 | CV = 0.44

Five-Number Summary:
Min = 0 | Q1 = 8 | Median = 11 | Q3 = 14 | Max = 20 | IQR = 6

Shape:
Skewness = -0.73

2.2 Visualizations

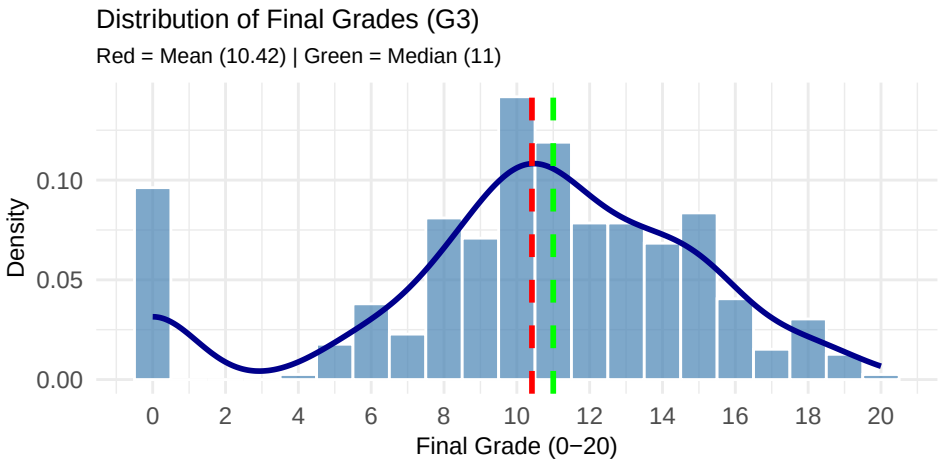

Figure 1: Distribution of Final Grades showing left skewness

2.3 Interpretation

The mean final grade is 10.42 with median of 11. The SD of 4.58 shows considerable variability (CV = 0.44). The skewness of -0.73 indicates a left-skewed distribution with more high-performing students. Grades span 0 to 20, with 50% scoring between 8 and 14 (IQR = 6).

3 Numeric Discrete Variable: Absences

3.1 Descriptive Statistics

N = 395 | Mean = 5.71 | Median = 4 | Mode = 0 | SD = 8 | Variance = 64.05 | CV = 1.402

Five-Number Summary:
Min = 0 | Q1 = 0 | Median = 4 | Q3 = 8 | Max = 75 | IQR = 8

Shape:
Skewness = 3.658 | Zero absences: 29.1 %

3.2 Visualizations

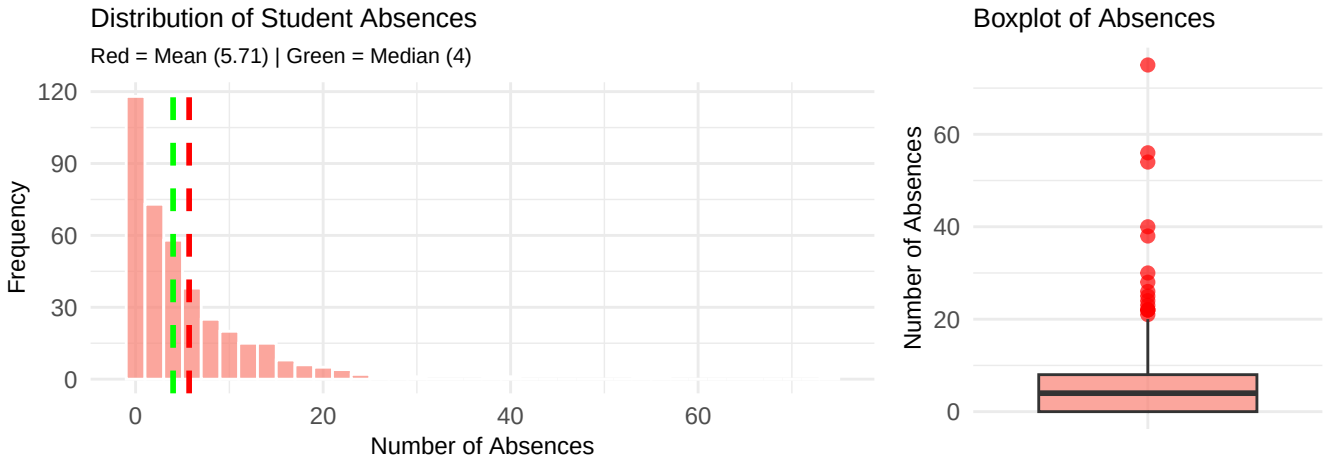

Figure 2: Distribution and outliers in student absences

3.3 Interpretation

Students average 5.71 absences with median of 4, but mode is 0 (29.1% had perfect attendance). The SD of 8.0 is notably large with CV = 1.402, indicating extremely high variability. The skewness of 3.658 shows an extremely right-skewed distribution. Absences range from 0 to 75, with 50% having 0-8 absences (IQR = 8).

4 Numeric Discrete Variable: Failures

4.1 Descriptive Statistics

N = 395 | Mean = 0.33 | Median = 0 | Mode = 0 | SD = 0.74 | Variance = 0.55 | Range = 0 - 3

4.2 Frequency Distribution

Table 1: Frequency Distribution of Past Class Failures

Failures	Count	Percentage	Cumulative
0	312	78.99	78.99
1	50	12.66	91.65
2	17	4.30	95.95
3	16	4.05	100.00

4.3 Visualizations

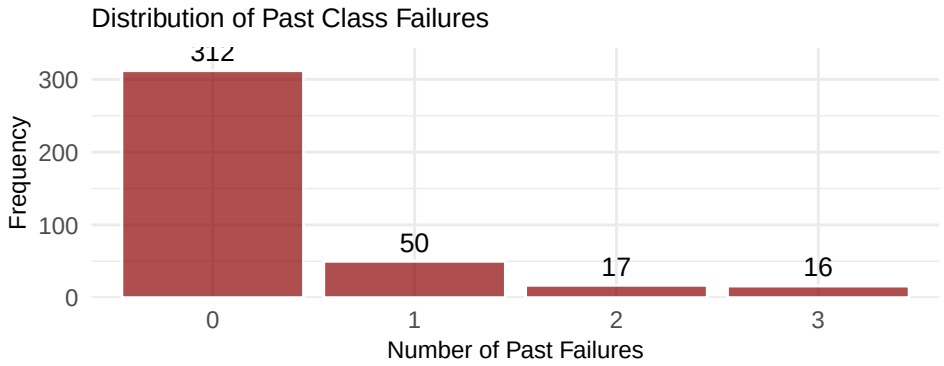

Figure 3: Distribution showing most students never failed

4.4 Interpretation

Mean is 0.33 failures with median and mode of 0. An impressive 78.99% have never failed a class. Only 8.35% have failed 2+ classes, representing a small at-risk group. The SD of 0.74 indicates limited variability.

5 Ordinal Variables: Study Time and Travel Time

Study time categories: 1 = <2 hours/week, 2 = 2-5 hours, 3 = 5-10 hours, 4 = >10 hours.

Travel time categories: 1 = <15 min, 2 = 15-30 min, 3 = 30-60 min, 4 = >60 min.

5.1 Frequency Distributions

Table 1: Study Time Distribution

Category	Count	Percentage	Cumulative
1	105	26.58	26.58
2	198	50.13	76.71

Category	Count	Percentage	Cumulative
3	65	16.46	93.16
4	27	6.84	100.00

Table 2: Travel Time Distribution

Category	Count	Percentage	Cumulative
1	257	65.06	65.06
2	107	27.09	92.15
3	23	5.82	97.97
4	8	2.03	100.00

Study Time - Central Tendency:
Mode = 2 | Median = 2

Travel Time - Central Tendency:
Mode = 1 | Median = 1

5.2 Visualizations

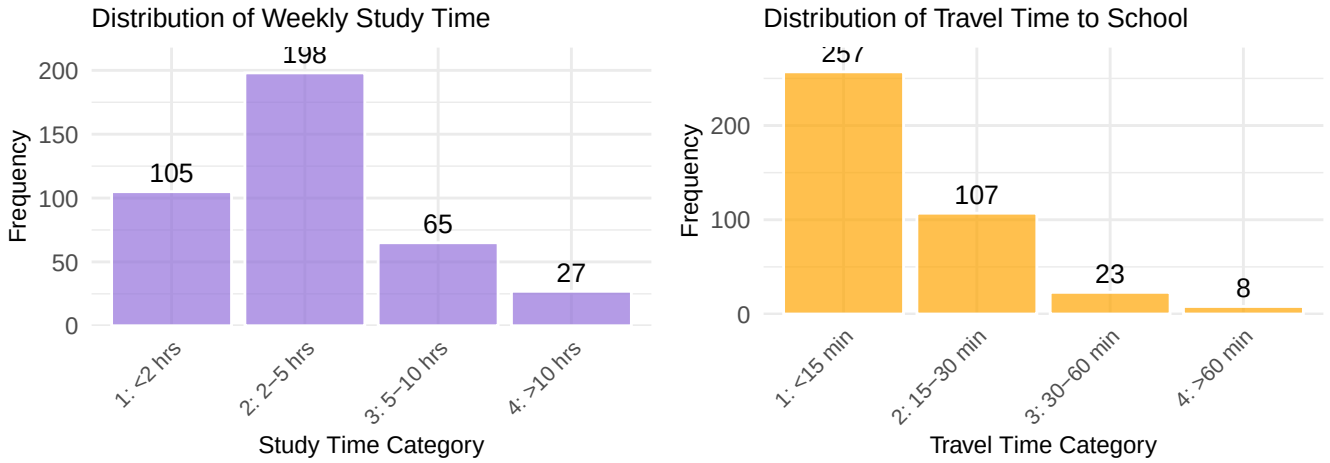

Figure 4: Study time and travel time distributions

5.3 Interpretation

Study Time: Most common is category 2 (2-5 hours/week) with 50.13% of students. Over 26% study <2 hours weekly (potentially insufficient). Only 6.84% study >10 hours. Median of 2 confirms typical student studies 2-5 hours weekly.

Travel Time: Majority (65.06%) live very close to school (<15 min commute). Mode and median are both 1. Additional 27.09% have 15-30 min commutes (92.15% total within 30 min). Only 7.85% face commutes >30 min. Short travel times likely minimize fatigue.

6 Binary Variables (Raju): Sex, Paid, Activities, Higher, Internet

6.1 Frequency Distribution

Table 4: Binary Variables Summary

Variable	Yes/F (n)	Yes/F (%)	No/M (n)	Mode
sex (F/M)	208	52.7	187	F

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

6.2 Visualization

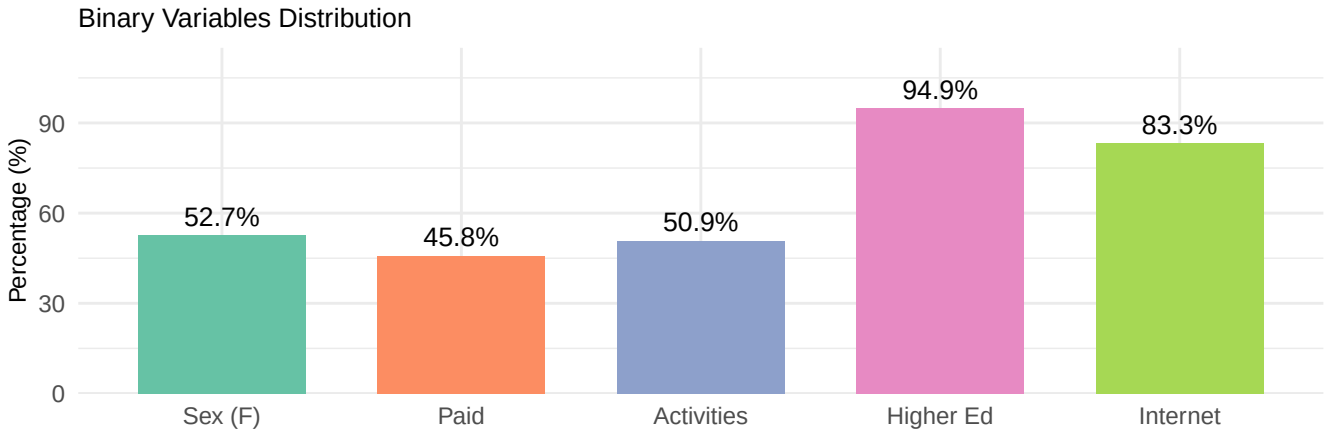

Figure 5: Distribution of Binary Variables - Percentage of Yes/Female responses

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.

7 Ordinal Variables (Raju): Medu and Famrel

Medu categories: 0 = none, 1 = primary (4th grade), 2 = 5th-9th grade, 3 = secondary, 4 = higher education.

Famrel categories: 1 = very bad, 2 = bad, 3 = neutral, 4 = good, 5 = excellent.

7.1 Frequency Distribution

Table 5: Ordinal Variables Summary

Variable	Mode	Median
Medu (0-4)	4	3
famrel (1-5)	4	4

7.2 Visualizations

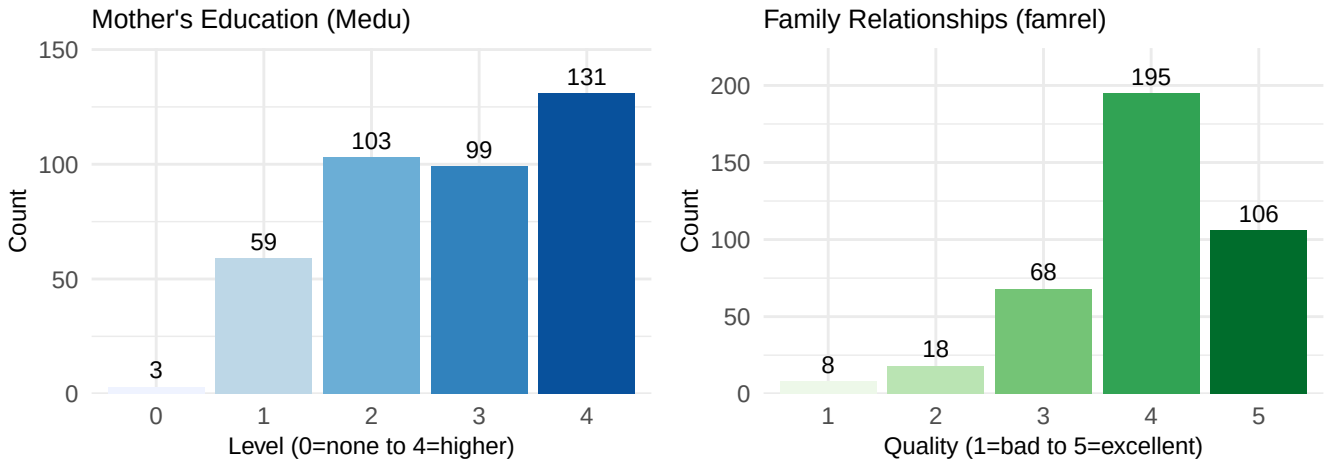

Figure 6: Distribution of Ordinal Variables - Mother's Education (left) and Family Relationships (right)

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

8 Numeric Variable (Raju): Age

8.1 Descriptive Statistics

N = 395 | Mean = 16.7 | Median = 17 | Mode = 16 | SD = 1.28 | Range = 15 - 22 | IQR = 2

8.2 Visualizations

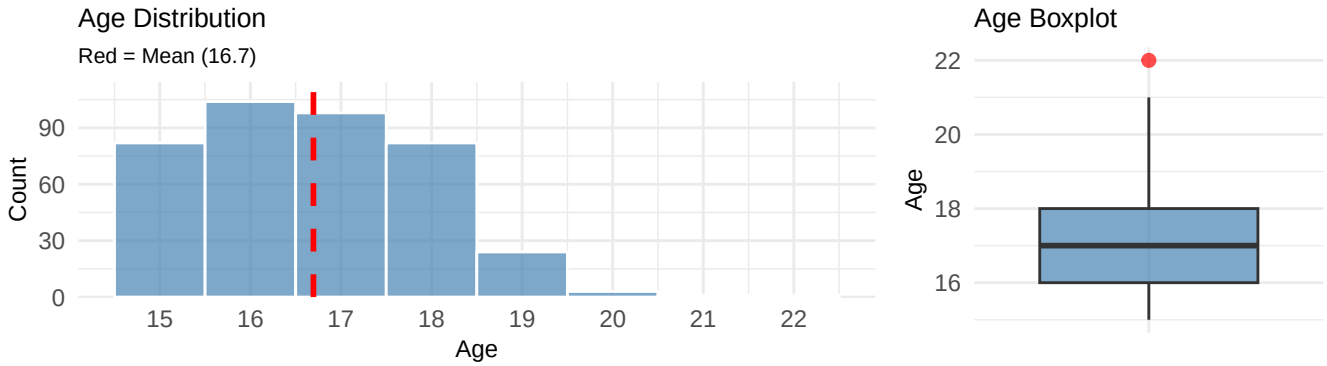

Figure 7: Age Distribution - Histogram with mean line (left) and Boxplot (right)

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.

9 Bivariate Analysis (Raju): Medu vs G3

9.1 Descriptive Statistics

Table 6: 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 correlation: r = 0.225

9.2 Visualization

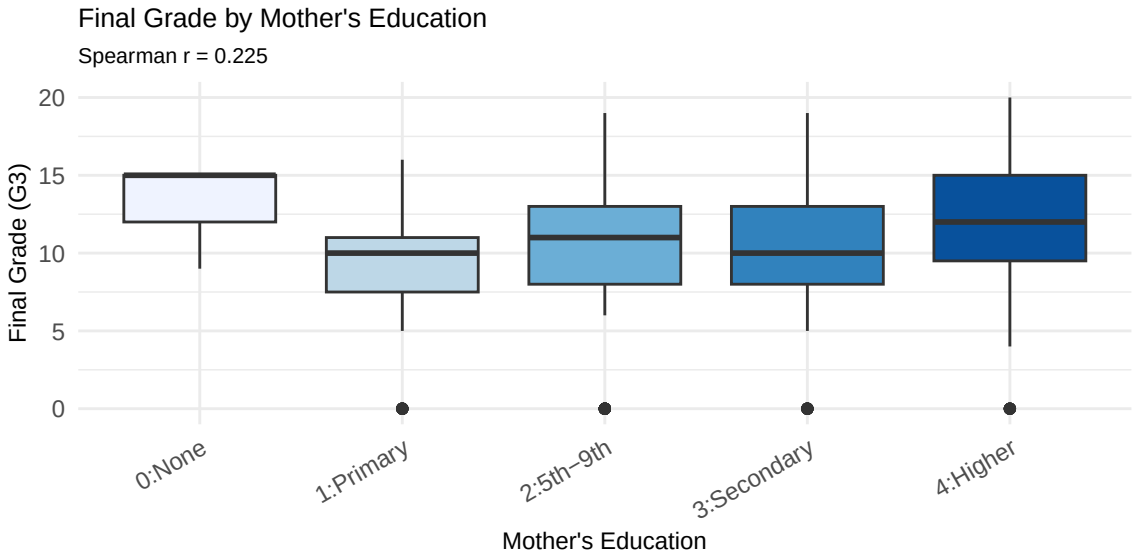

Figure 8: Boxplot of Final Grades by Mother's Education Level

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.

10 Literature

Cortez, Paulo. 2008. "Student Performance." UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T.

Cortez, Paulo, and Alice Silva. 2008. "Using Data Mining to Predict Secondary School Student Performance." In *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, 5–12. Porto, Portugal: EUROSIS.

11 Appendix

11.1 System Information

```
## R version 4.4.1 (2024-06-14)
## Platform: aarch64-apple-darwin20
## Running under: macOS 26.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK version
##
```

```
## locale:
## [1] C.UTF-8/C.UTF-8/C.UTF-8/C/C.UTF-8/C.UTF-8
##
## time zone: Europe/Berlin
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gridExtra_2.3  knitr_1.48     moments_0.14.1 ggplot2_4.0.0  dplyr_1.1.4
##
## loaded via a namespace (and not attached):
##  [1] vctrs_0.6.5        cli_3.6.5          rlang_1.1.6        xfun_0.48
##  [5] highr_0.11         generics_0.1.3     S7_0.2.0           labeling_0.4.3
##  [9] glue_1.8.0         htmltools_0.5.8.1  tinytex_0.53       scales_1.4.0
## [13] fansi_1.0.6        rmarkdown_2.28     grid_4.4.1         evaluate_1.0.1
## [17] tibble_3.2.1       fastmap_1.2.0      yaml_2.3.10        lifecycle_1.0.4
## [21] compiler_4.4.1     RColorBrewer_1.1-3 pkgconfig_2.0.3    farver_2.1.2
## [25] digest_0.6.37      R6_2.5.1           tidyselect_1.2.1   utf8_1.2.4
## [29] pillar_1.9.0       magrittr_2.0.3     withr_3.0.2        tools_4.4.1
## [33] gtable_0.3.6
```