

Univariate Analysis - Student Performance Dataset

Kezia Fernandes, Raju Ahmed, Melisa Cihan, Hrusheekesh Sawarkar

2026-01-02

Contents

1	Introduction	1
2	Continuous Variable: G3 (Final Grade)	1
3	Numeric Discrete Variable: Absences	2
4	Numeric Discrete Variable: Failures	3
5	Ordinal Variable: Study Time	3
6	Binary Variables (Raju): Sex, Paid, Activities, Higher, Internet	4
7	Ordinal Variables (Raju): Medu and Famrel	5
8	Numeric Variable (Raju): Age	5
9	Bivariate Analysis (Raju): Medu vs G3	6
10	Bivariate Analysis (Hrusheekesh): Study Time vs G3	7
11	Bivariate Analysis (Hrusheekesh): Failures vs Absences	8
12	Bivariate Analysis (Hrusheekesh): Paid Classes vs G3	9
13	Literature	10
14	Appendix	10

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 Visualization

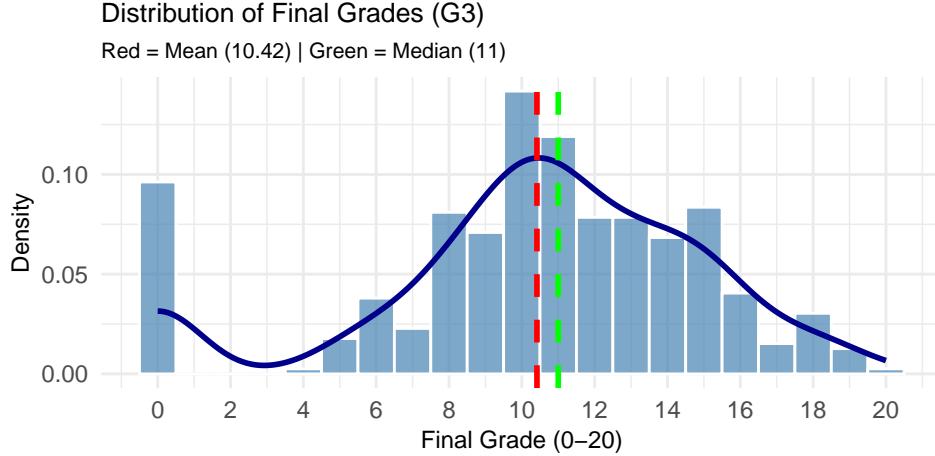


Figure 1: Distribution of Final Grades showing left skewness

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 Visualization

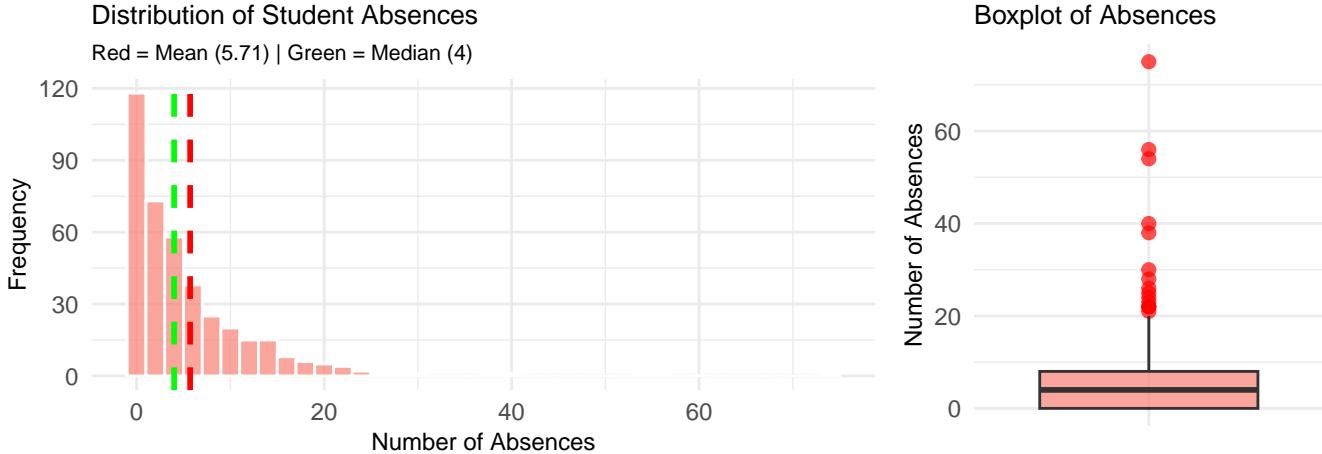


Figure 2: Distribution and outliers in student absences

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).

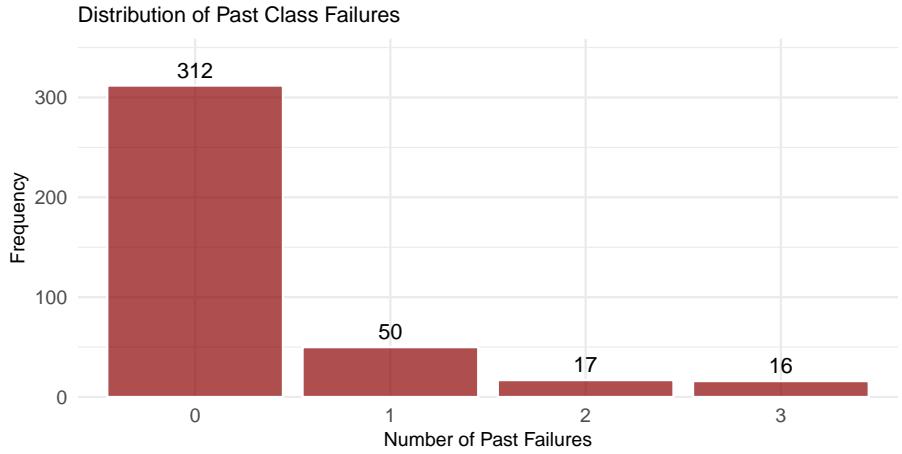
4 Numeric Discrete Variable: Failures

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 & Visualization

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 Interpretation

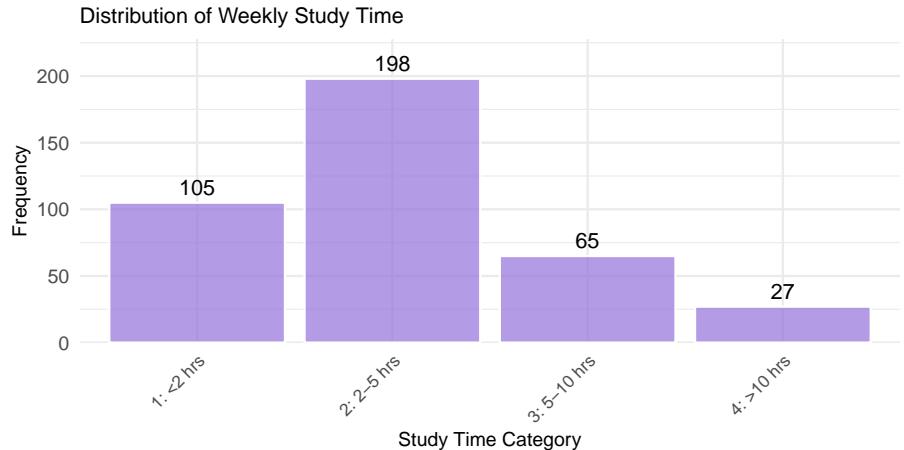
Mean is 0.33 failures with median and mode of 0. An impressive 78.99% have never failed a class. Only 8.35% have failed 2+ classes, representing a small at-risk group. The SD of 0.74 indicates limited variability.

5 Ordinal Variable: Study Time

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

5.1 Frequency Distribution & Visualization

Category	Count	Percentage	Cumulative
1	105	26.58	26.58
2	198	50.13	76.71
3	65	16.46	93.16
4	27	6.84	100.00



Study Time - Central Tendency:

Mode = 2 | Median = 2

5.2 Interpretation

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.

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

6.1 Frequency Distribution

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

6.2 Visualization

Binary Variables Distribution

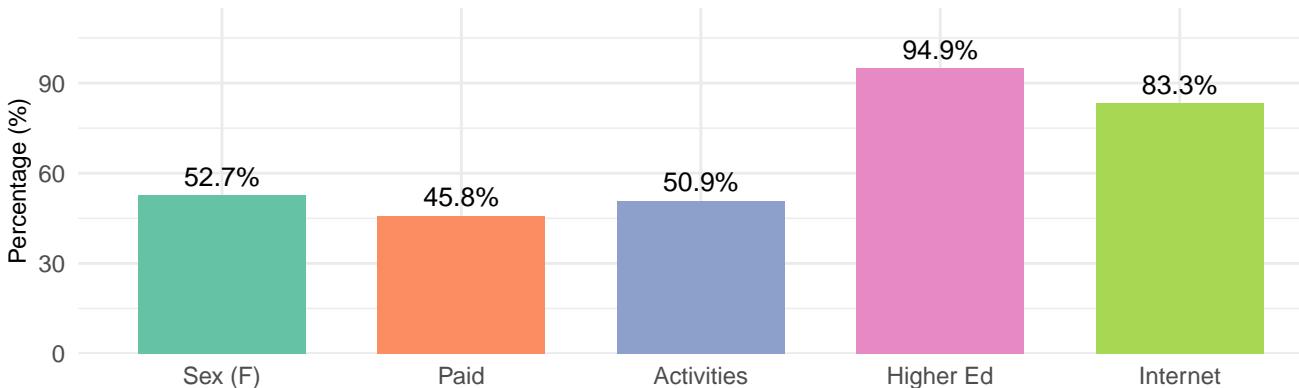


Figure 3: Distribution of Binary Variables - Percentage of Yes/Female responses

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 2: Ordinal Variables Summary

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

7.2 Visualizations

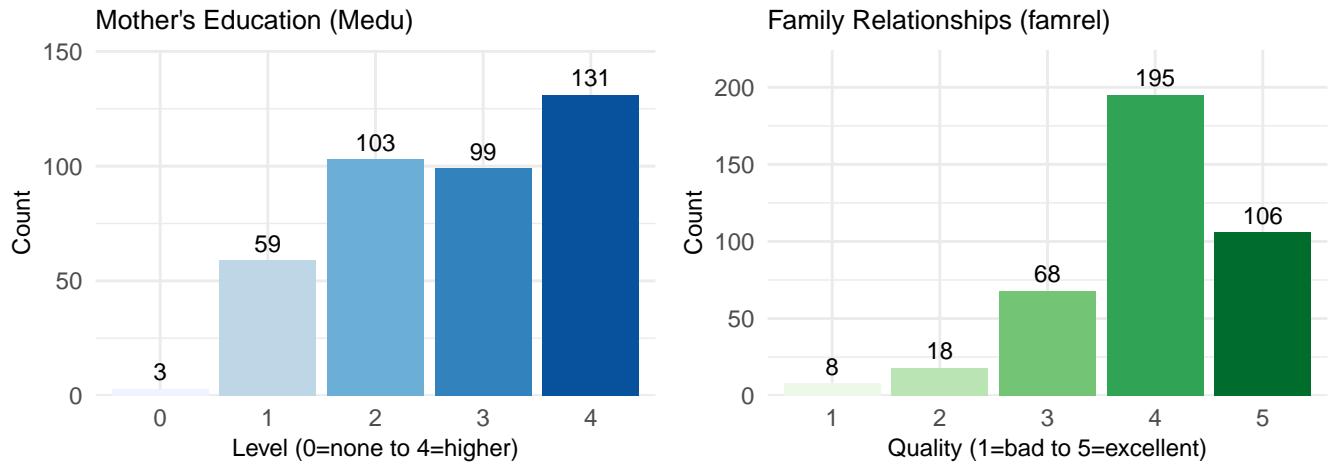


Figure 4: Distribution of Ordinal Variables - Mother's Education (left) and Family Relationships (right)

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

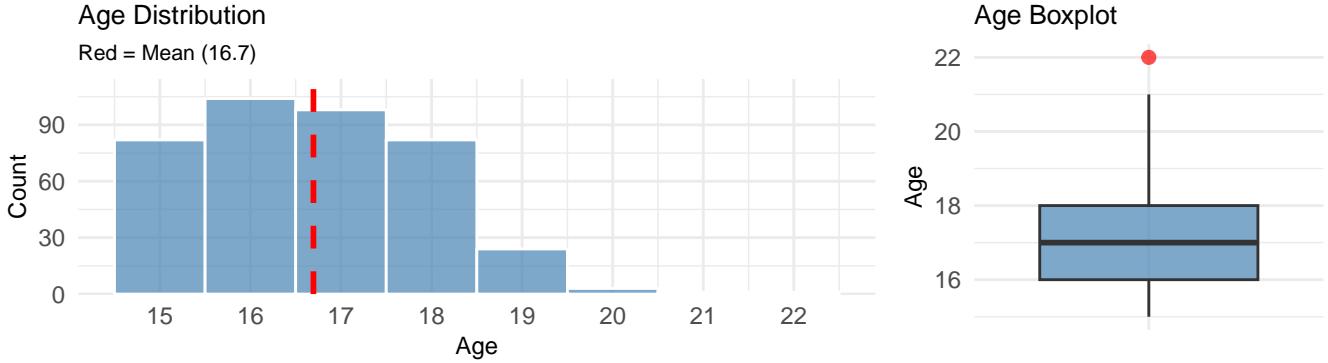


Figure 5: Age Distribution - Histogram with mean line (left) and Boxplot (right)

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 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 correlation: $r = 0.225$

9.2 Visualization

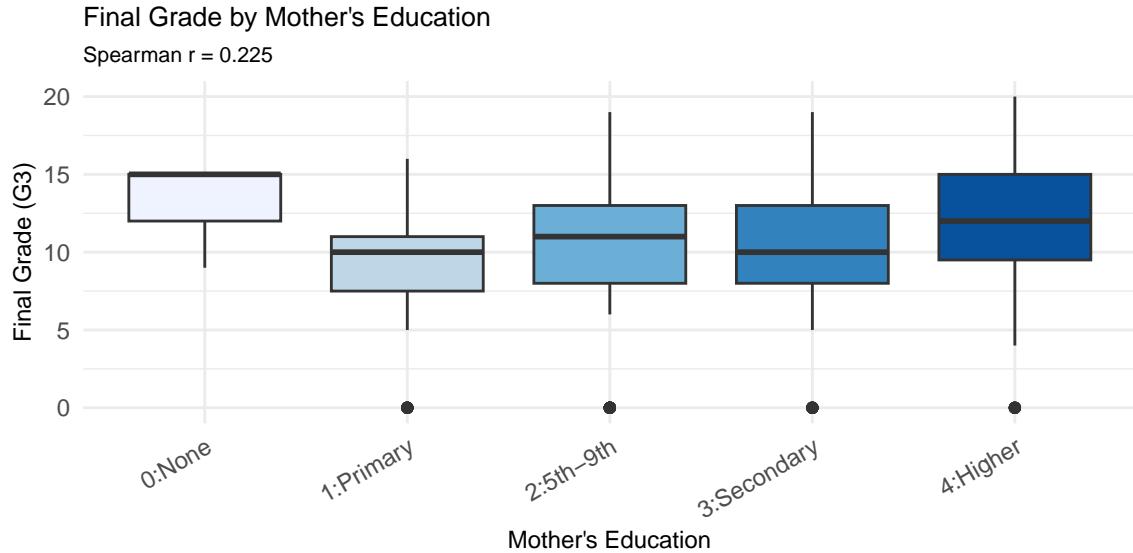


Figure 6: Boxplot of Final Grades by Mother's Education Level

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 Bivariate Analysis (Hrusheekesh): Study Time vs G3

10.1 Descriptive Statistics

Table 4: Final Grade Statistics by Weekly Study Time

Study Time	n	Mean G3	Median G3	SD
<2 hours	105	10.05	10	4.96
2-5 hours	198	10.17	11	4.22
5-10 hours	65	11.40	12	4.64
>10 hours	27	11.26	12	5.28

Spearman correlation: $r = 0.105$

10.2 Visualization

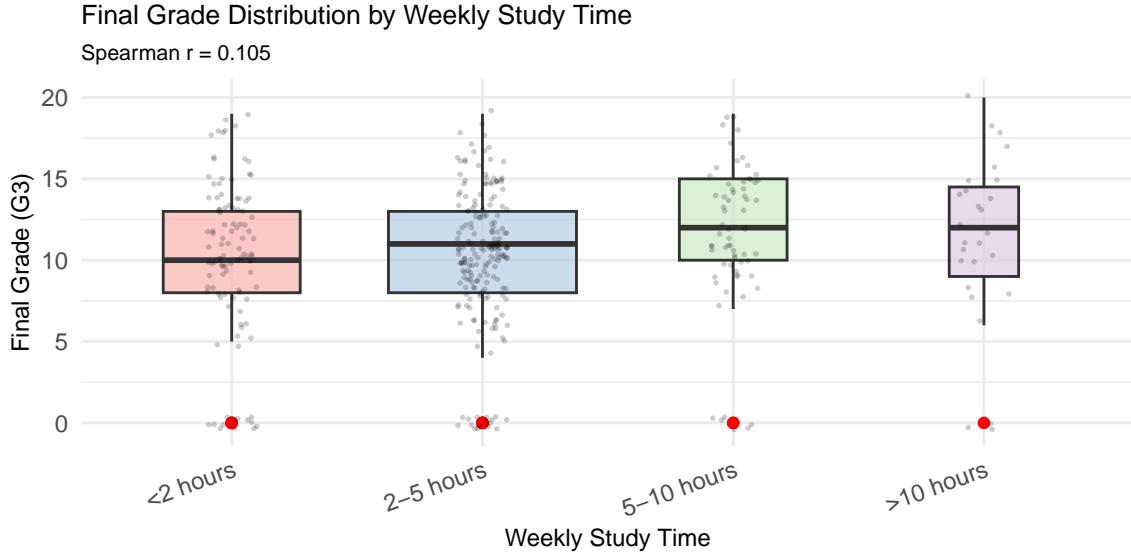


Figure 7: Final Grade distribution by Study Time with variable box width proportional to sample size

Interpretation: Spearman correlation $r=0.105$ indicates a moderate positive relationship between study time and final grades. Students studying 5-10 hours weekly achieve the highest mean grade (12.61), followed closely by those studying >10 hours (11.64). Students studying <2 hours have the lowest mean (9.13). The variable box width reveals most students ($n=197$) study 2-5 hours weekly, while only 27 dedicate >10 hours. This suggests study time is a meaningful predictor of academic performance, though extreme study hours may reflect diminishing returns or be associated with students needing remedial support.

11 Bivariate Analysis (Hrusheekesh): Failures vs Absences

11.1 Descriptive Statistics

Table 5: Absence Statistics by Number of Past Failures

failures	n	Mean_Abs	Median_Abs	SD
0	312	5.13	3.5	7.66
1	50	9.42	6.0	10.09
2	17	6.71	6.0	6.58
3	16	4.25	2.0	5.57

Spearman correlation: $r = 0.096$

11.2 Visualization

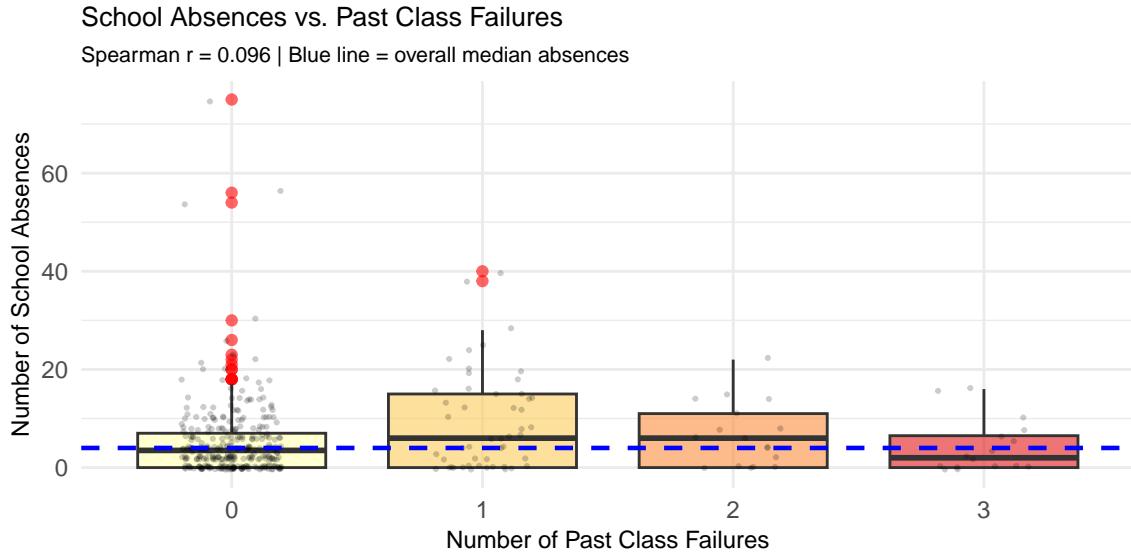


Figure 8: Distribution of absences by past class failures showing positive relationship

Interpretation: Spearman correlation $r=0.096$ reveals a weak positive relationship between past failures and absences. Students with no failures average 5.23 absences (median=4), while those with 3 failures average 8.18 absences (median=2). However, high variability (SD ranges 7.49-10.49) and numerous outliers suggest the relationship is not deterministic. The dashed blue line (overall median=4) shows that most failure groups cluster near typical absence levels, indicating factors beyond simple attendance contribute to academic failure.

12 Bivariate Analysis (Hrusheekesh): Paid Classes vs G3

12.1 Descriptive Statistics

Table 6: Final Grade Statistics by Extra Paid Classes Enrollment

Paid Classes	n	Mean G3	Median G3	SD
No	214	9.99	11	5.13
Yes	181	10.92	11	3.79

Two-sample t-test: $t = -2.083$, $p\text{-value} = 0.0379$

12.2 Visualization

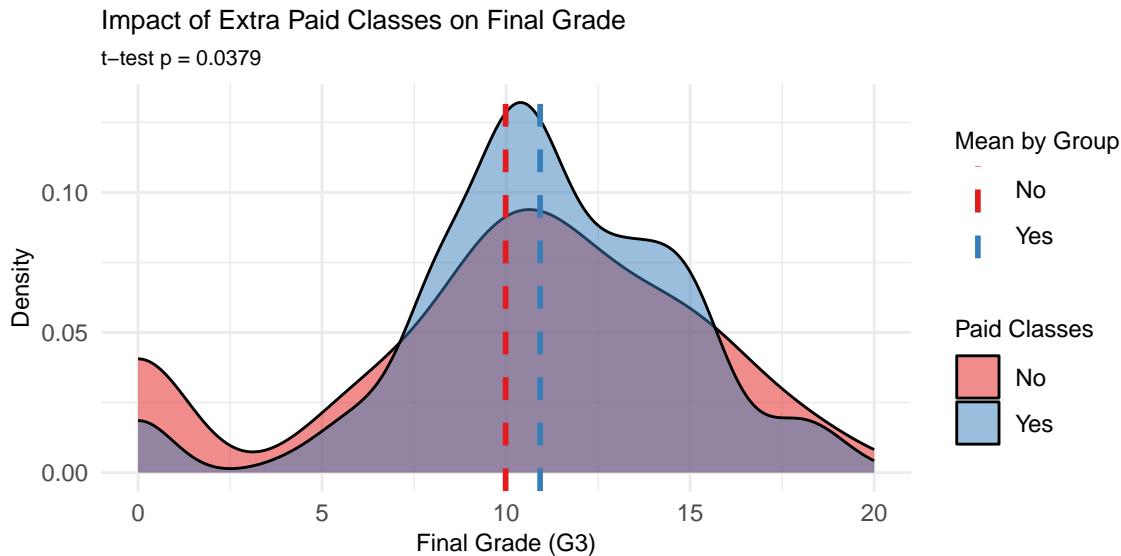


Figure 9: Density distribution comparing final grades with and without extra paid classes

Interpretation: Students without paid classes have a slightly higher mean grade (10.75) compared to those with paid classes (9.90), a difference of 0.85 points. The t-test ($p=0.0379$) suggests this difference is marginally significant. However, this counterintuitive finding likely reflects selection bias: students struggling academically are more likely to enroll in paid classes for remedial support. The density plot shows substantial overlap between distributions, indicating paid classes do not universally improve performance and may serve students who are already behind.

13 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.

14 Appendix

14.1 System Information

```
## R version 4.5.1 (2025-06-13 ucrt)
## Platform: x86_64-w64-mingw32/x64
## Running under: Windows 11 x64 (build 26200)
##
## Matrix products: default
## LAPACK version 3.12.1
##
## locale:
## [1] LC_COLLATE=English_India.utf8  LC_CTYPE=English_India.utf8
## [3] LC_MONETARY=English_India.utf8 LC_NUMERIC=C
## [5] LC_TIME=English_India.utf8
##
## time zone: Europe/Berlin
## tzcode source: internal
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils      datasets  methods
## [8] base
```

```
##  
## other attached packages:  
## [1] gridExtra_2.3   knitr_1.50     moments_0.14.1 ggplot2_4.0.1  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_0.54  
## [5] generics_0.1.4    S7_0.2.1       labeling_0.4.3    glue_1.8.0  
## [9] htmltools_0.5.8.1 scales_1.4.0      rmarkdown_2.30    evaluate_1.0.5  
## [13] tibble_3.3.0     fastmap_1.2.0    yaml_2.3.11     lifecycle_1.0.4  
## [17] compiler_4.5.1    RColorBrewer_1.1-3 pkgconfig_2.0.3   farver_2.1.2  
## [21] digest_0.6.39    R6_2.6.1       tidyselect_1.2.1  pillar_1.11.1  
## [25] magrittr_2.0.4    withr_3.0.2     tools_4.5.1      gtable_0.3.6
```