Statistical Analysis - Student Performance Dataset

Kezia Fernandes, Raju Ahmed, Melisa Cihan, Hrusheekesh Sawarkar

2026-01-04

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 Univariate Statistics

2.1 Continuous Variable: G3 (Final Grade)

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

2.1.1 Visualization

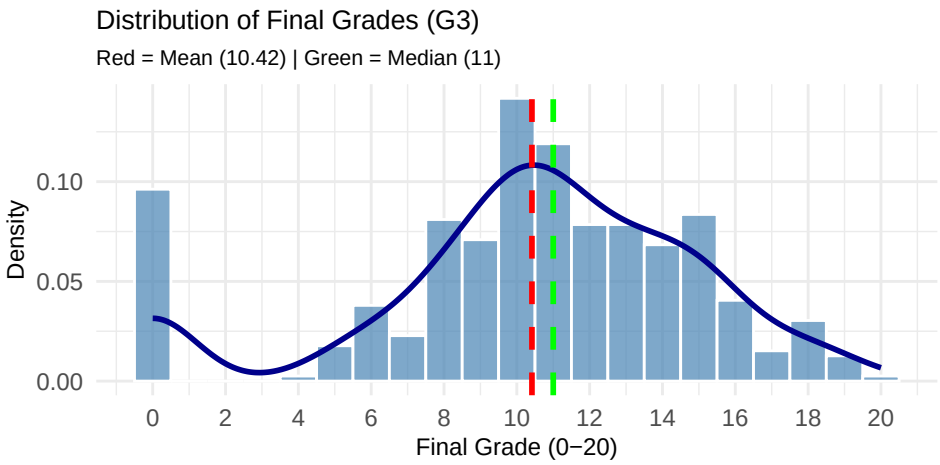


Figure 1: Distribution of Final Grades showing left skewness

2.1.2 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).

2.2 Numeric Discrete Variable: Absences

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 %

2.2.1 Visualization

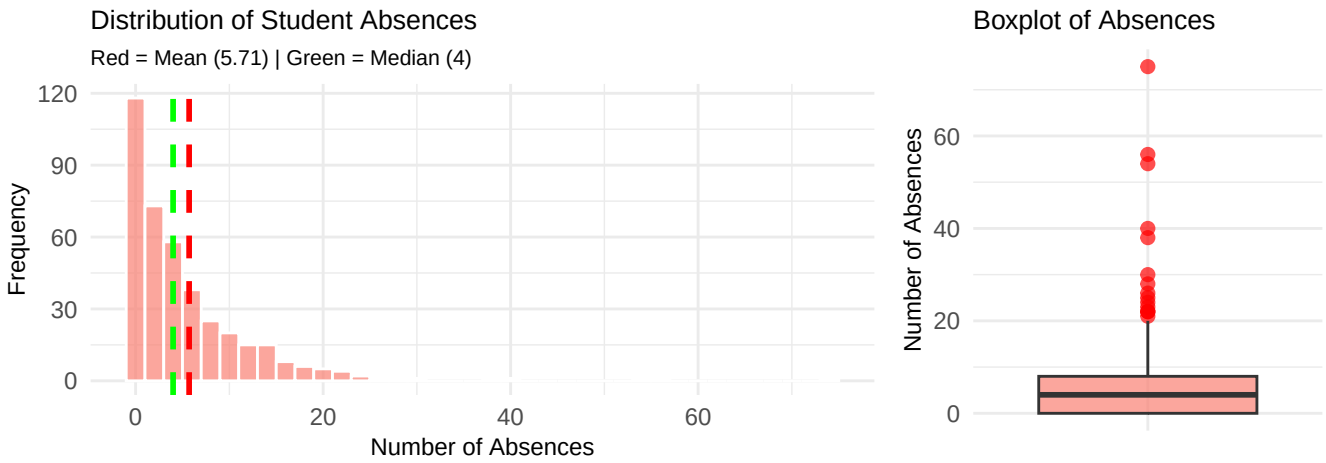


Figure 2: Distribution and outliers in student absences

2.2.2 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).

2.3 Numeric Discrete Variable: Failures

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

2.3.1 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

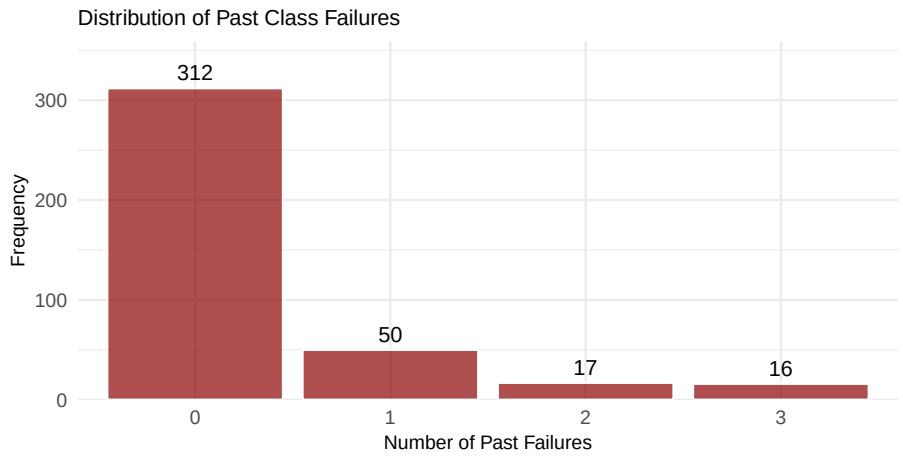


Figure 3: Frequency distribution and visualization of past class failures

2.3.2 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.

2.4 Ordinal Variable: Study Time

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

2.4.1 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

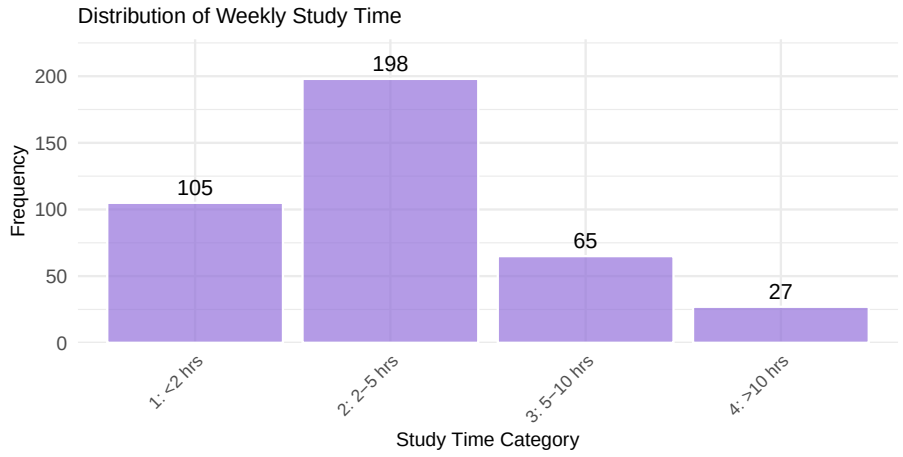


Figure 4: Frequency distribution and visualization of weekly study time

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

2.4.2 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.

2.5 Binary Variables: Paid Extra Classes, Higher Education Aspiration, Internet Access

2.5.1 Frequency Distribution & Visualization

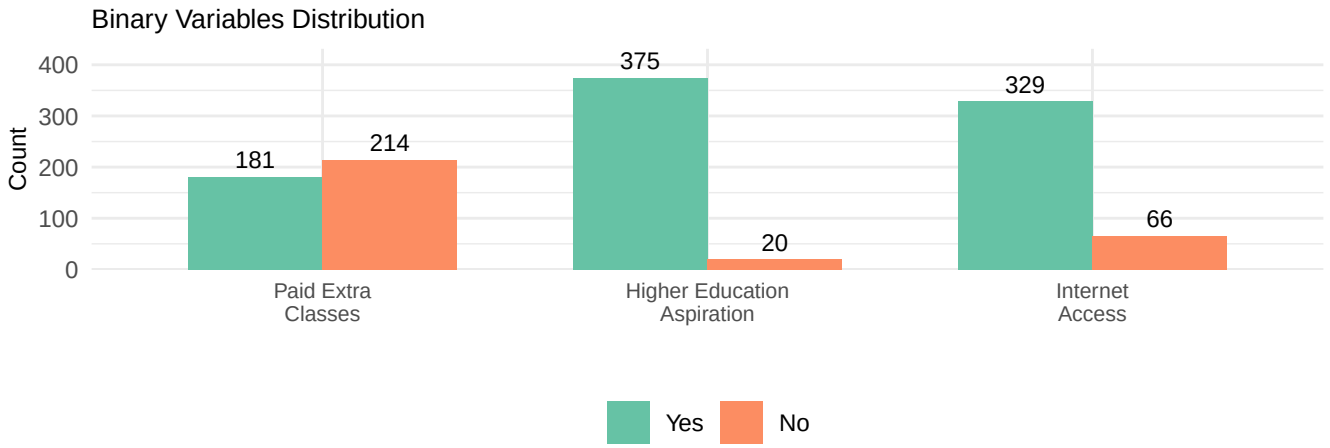


Figure 5: Binary Variables - Grouped Bar Chart

Interpretation: Less than half (45.8%) take paid extra Math classes. A striking 94.9% aspire to higher education. Internet access is available to 83.3% of students.

2.6 Ordinal Variables: Mother's Education (Medu) and Family Relationship (Famrel)

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

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

2.6.1 Visualization

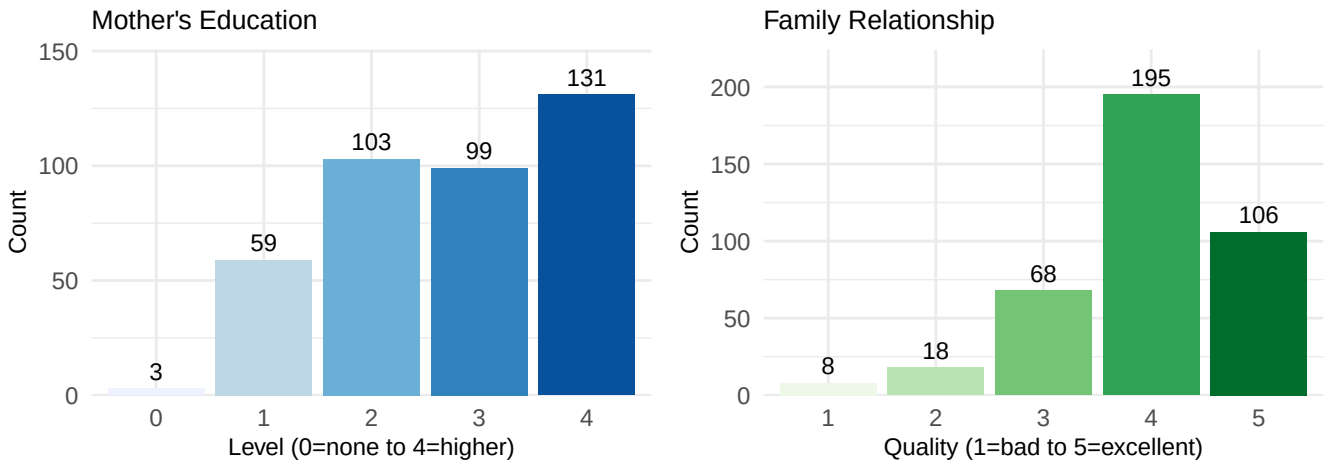


Figure 6: Distribution of Ordinal Variables - Mother's Education (left) and Family Relationship (right)

Interpretation: Mother's education is skewed toward higher levels with mode=4 (higher education, n=131) and median=3 (secondary education). Only 3 mothers have no formal education. Family relationship quality is predominantly positive with mode=4 (good, n=195) and median=4 (good). Over 71% report good to excellent relationships, suggesting supportive home environments.

2.7 Numeric Variable: Age

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

2.7.1 Visualization

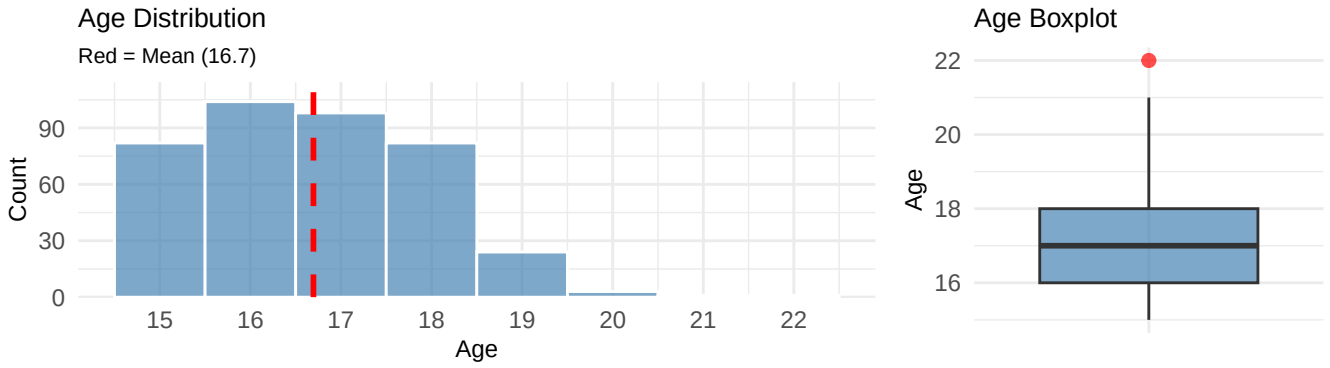


Figure 7: Age Distribution - Histogram with mean line (left) and Boxplot (right)

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.

3 Bivariate Statistics

3.1 Mother's Education vs Final Grade (G3)

Medu	n	Mean G3	SD
0	3	13	3.46
1	59	8.68	4.36
2	103	9.73	4.64
3	99	10.3	4.62
4	131	11.76	4.27

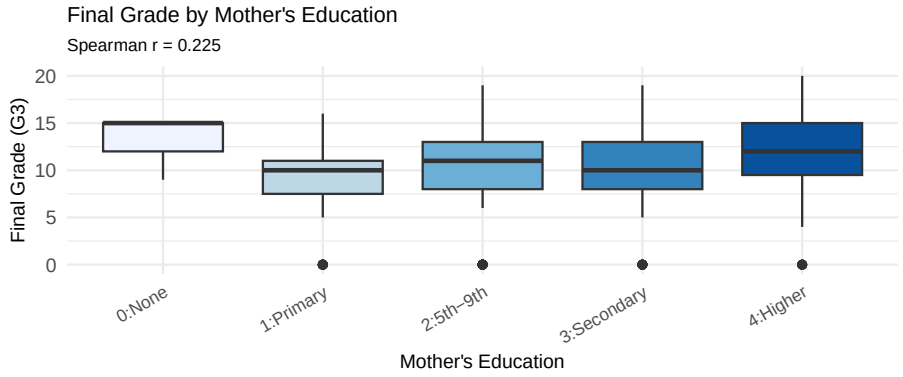


Figure 8: Final Grade by Mother's Education - Summary Table (left) and Boxplot (right)

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.

3.2 Study Time vs G3

Study Time	n	Mean	Median	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

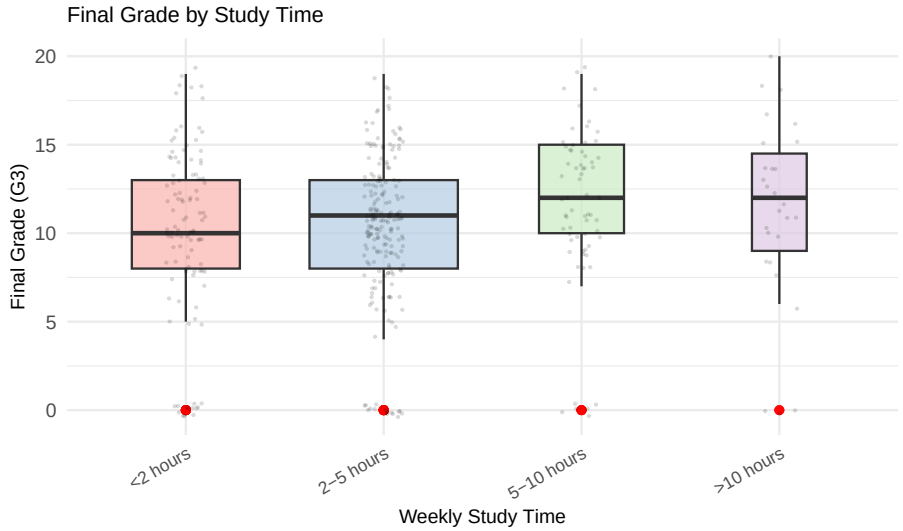


Figure 9: Study Time vs G3: Statistics (left) and distribution (right)

Interpretation: Students studying 5-10 hours achieve highest mean (11.4), while <2 hours group has lowest (10.05). Most students (n=198) study 2-5 hours weekly. Study time is a meaningful predictor, though extreme hours may not reflect significant positive returns, or may reflect students needing remedial support.

3.3 Failures vs Absences

Failures	n	Mean	Median	SD
0	312	5.13	3.5	7.66
1	50	9.42	6	10.09
2	17	6.71	6	6.58
3	16	4.25	2	5.57

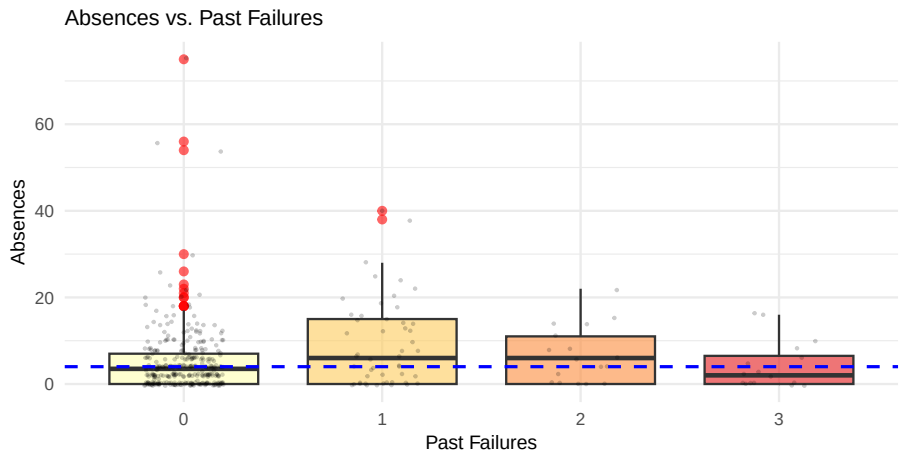


Figure 10: Failures vs Absences: Statistics (left) and distribution (right)

Interpretation: Students with no failures average 5.13 absences, while those with 3 failures average 4.25. High variability and outliers suggest the relationship is not deterministic. Most failure groups cluster near typical absence levels, indicating factors beyond attendance contribute to academic failure.

3.4 Paid Classes vs G3

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

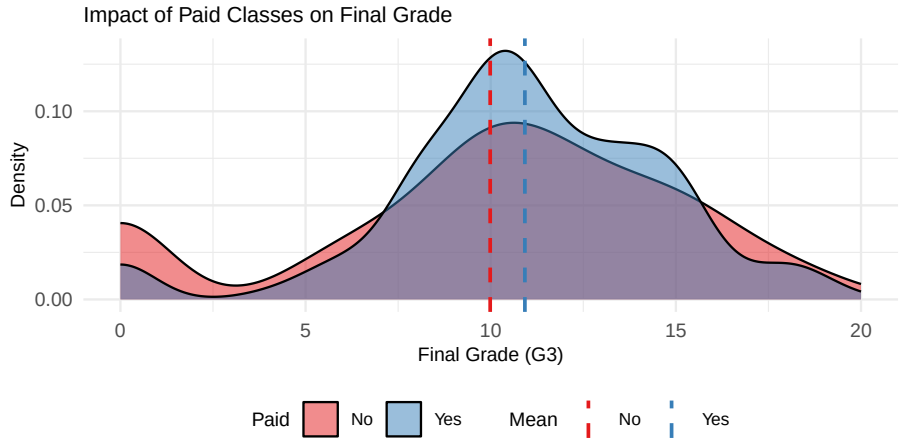


Figure 11: Paid Classes vs G3: Statistics (left) and density distribution (right)

Interpretation: Students without paid classes have higher mean grade (9.99) vs. those with paid classes (10.92), difference of -0.93 points. The visualizations suggest marginally significant difference. Counterintuitive finding likely reflects selection bias: struggling students more likely enroll for remedial support. Substantial distribution overlap indicates paid classes do not universally improve performance.

3.5 G3 vs Internet Access

Internet	n	Mean	Median	SD
No	66	9.41	10	4.49
Yes	329	10.62	11	4.58

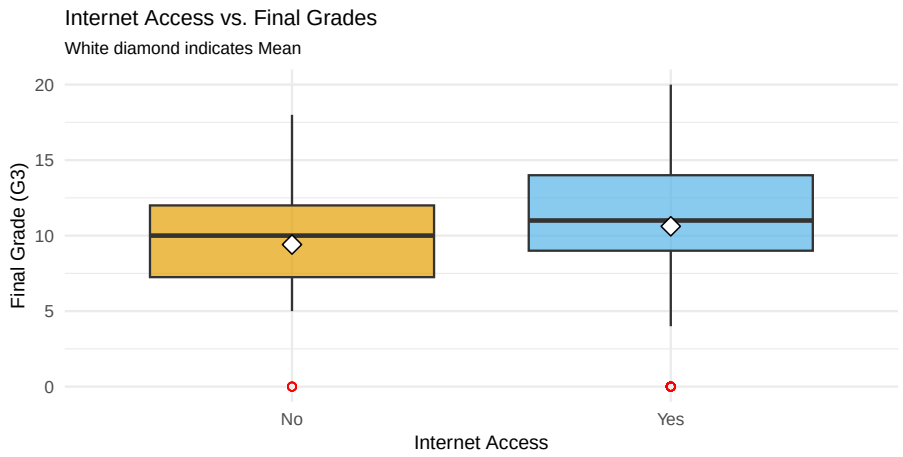


Figure 12: Internet Access vs G3: Statistics (left) and Boxplot (right)

Interpretation: Students with internet access at home have a higher mean grade (10.62) compared to those without (9.41), a difference of roughly 1.2 points. The boxplot visually confirms this shift, with the "Yes" group showing a higher median and mean (white diamond). This suggests that access to digital resources may positively influence academic performance, potentially by facilitating research and study materials.

3.6 G3 vs Wish For Higher Education

Higher Education?	n	Mean	Median	SD
No	20	6.80	8	4.83
Yes	375	10.61	11	4.49

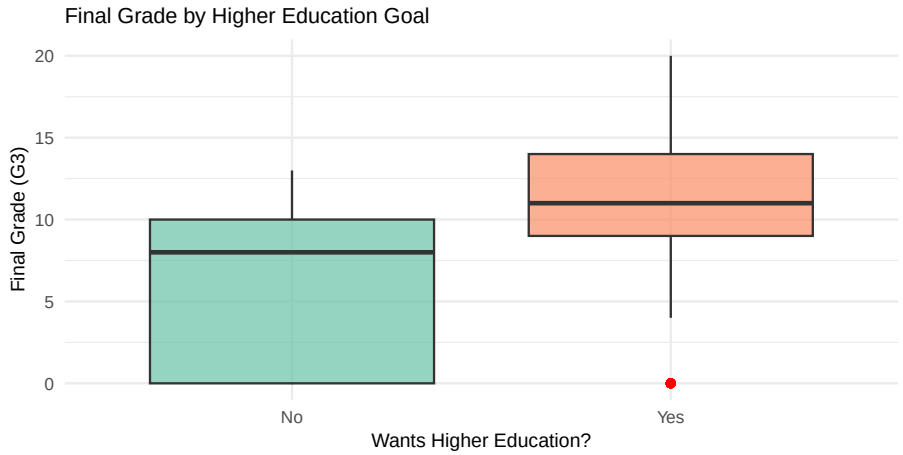


Figure 13: Higher Education Goal vs G3: Statistics (left) and Boxplot (right)

Interpretation: The analysis of the descriptive statistics and the boxplot reveals a clear distinction between the two groups. Students who intend to pursue higher education (`higher = yes`) demonstrate a noticeably higher level of academic performance compared to those who do not (`higher = no`).

Observing the descriptive table, the mean and median grades for the "Yes" group are elevated, suggesting that motivation for future studies is a strong factor in current performance. The boxplot further confirms this trend:

- **Median Differences:** The median line for the "Yes" group is positioned higher on the G3 scale than that of the "No" group.
- **Distribution Shift:** The entire interquartile range (the box) for students aiming for higher education is shifted upwards, indicating that the bulk of these students perform better than the majority of the "No" group.

- **Variability:** While both groups show some spread, the lower quartile for the "Yes" group is often higher than the median of the "No" group, highlighting a significant gap in achievement.

Overall, the desire to attend higher education appears to be positively associated with higher final grades.

3.7 Family Relationship vs Age

Age	n	Mean	Median	SD
15	82	4	4	0.89
16	104	3.84	4	0.98
17	98	3.91	4	0.86
18	82	4.02	4	0.89
19	24	3.88	4	0.68
20	3	5	5	0
21	1	5	5	NA
22	1	5	5	NA

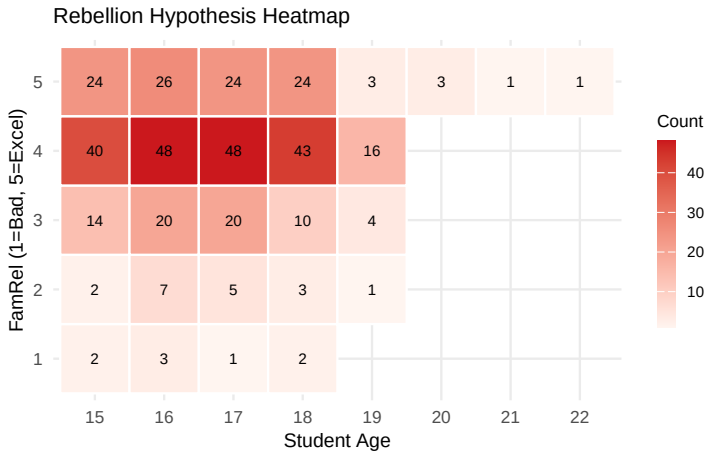
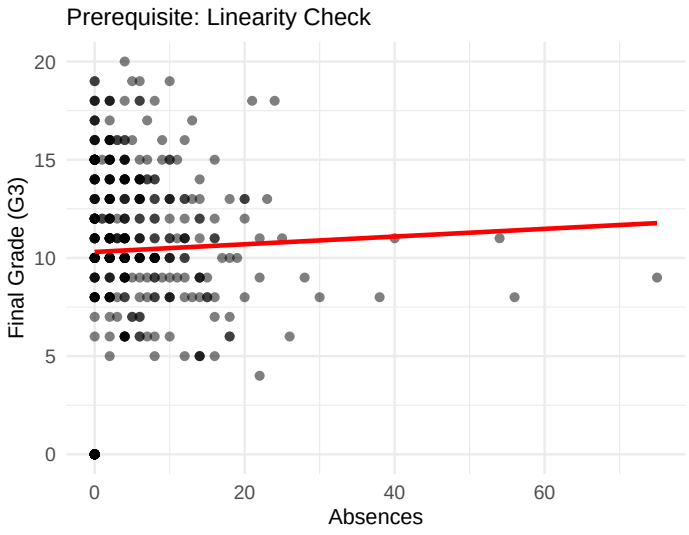


Figure 14: Family Relationship vs Age: Statistics (left) and Heatmap (right)

Interpretation The heatmap suggests that family relationships do not significantly deteriorate with age, contradicting the "Rebellion Hypothesis." The highest concentration of students consistently reports good to excellent relationships (levels 4 and 5) across all age groups, with no clear downward trend in relationship quality as students get older.

3.8 Simple Linear Regression G3~Absences



Regression Coefficients: Impact of Absences on Final Grade

term	estimate	std.error	statistic	p.value
(Intercept)	10.303	0.283	36.347	0.000
absences	0.020	0.029	0.679	0.497

Figure 15: Prerequisite Linearity Check (left) and Regression Coefficients (right)

Interpretation: The simple linear regression analysis was conducted to assess the effect of student absences on final mathematics grades (G3).

- **Significance:** The p-value for the absences coefficient is **0.497**. Since this value is significantly higher than the standard alpha level of 0.05, the relationship is **not statistically significant**.

- **Coefficient (Slope):** The estimated coefficient for absences is **0.020**. While this technically suggests a negligible positive increase in grades per absence, the lack of statistical significance indicates that this result is likely due to chance or random variation rather than a meaningful effect.
- **Conclusion:** Based on this model, we cannot conclude that the number of absences is a reliable predictor of the final mathematics grade. The data suggests that absences alone do not have a significant linear impact on academic performance in this context.

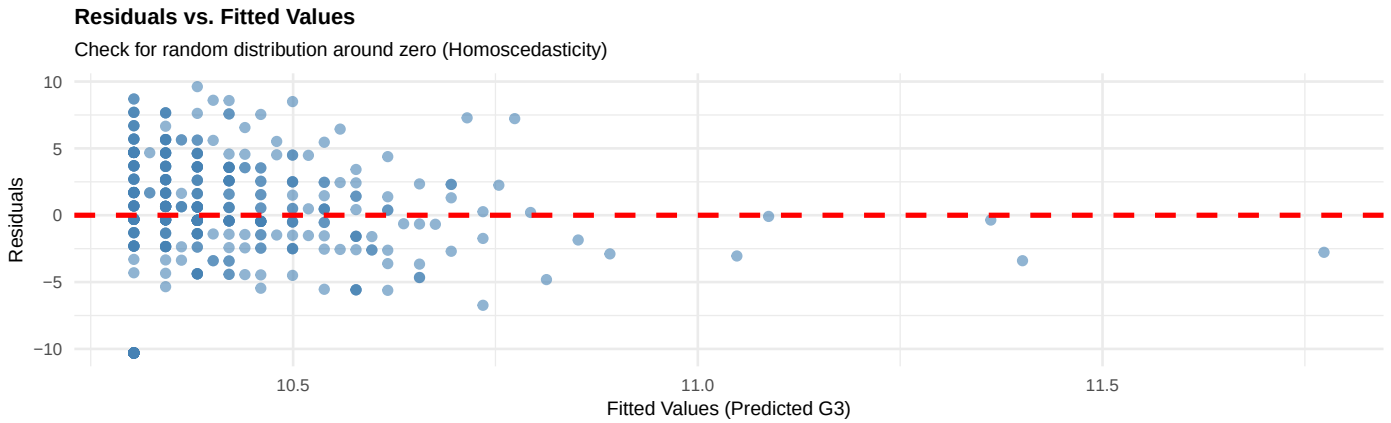


Figure 16: Model Diagnostics: Residuals vs Fitted Values

4 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.