What Impacts Student Performance?

...

By: Kevin Wehrle
DSC 101

Data Set Overview

- UC Irvine Machine Learning Repository —> “Student Performance”
- Student achievement in secondary education of two Portuguese schools
- Collected using school reports and questionnaires
- In regard to two subjects: Mathematics and Portuguese language
- Shape:
  - 649 instances
  - 30 features
- Target Variable —-> **Final Grade (G3)**
- Data types include...
  - Numerical (integer)
  - Categorical
  - Binary

Goal

Determine which variable(s) have the greatest impact on students' final grade in an attempt to find ways to help students who are academically struggling find ways to potentially improve their grades.

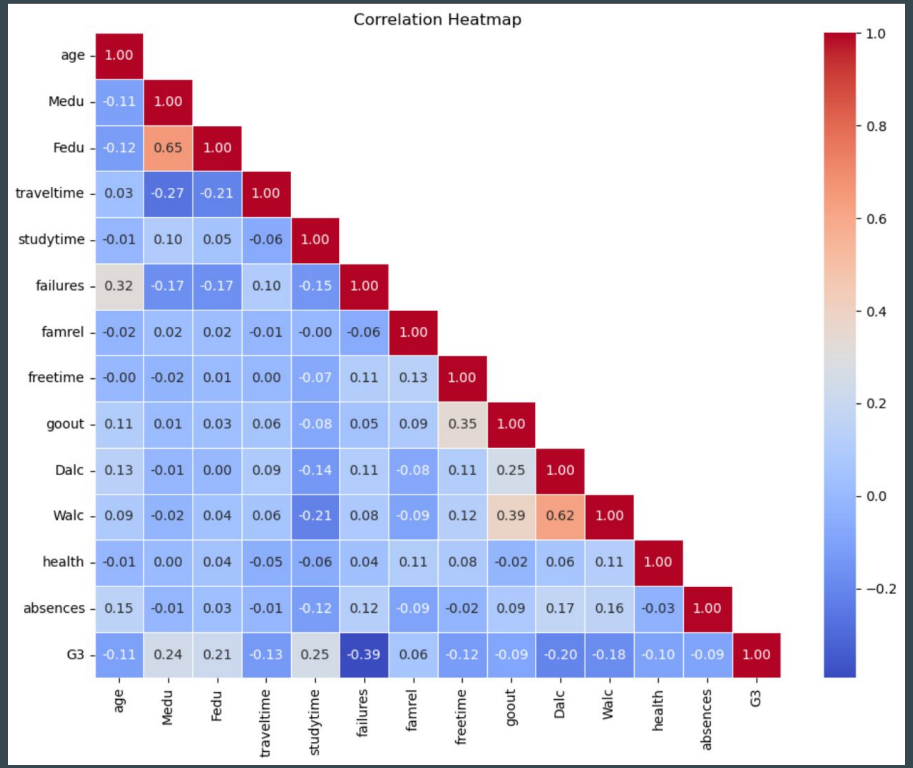
Cleaning and Preprocessing

- Removed G1 (first period grade) and G2 (second period grade) from target variables
  - Focusing on ONLY G3 (final grade)
- Check for and removed potential missing values
  - None found
- Converted categorical variables to numerical

Heatmap Overview

- All numerical values compared

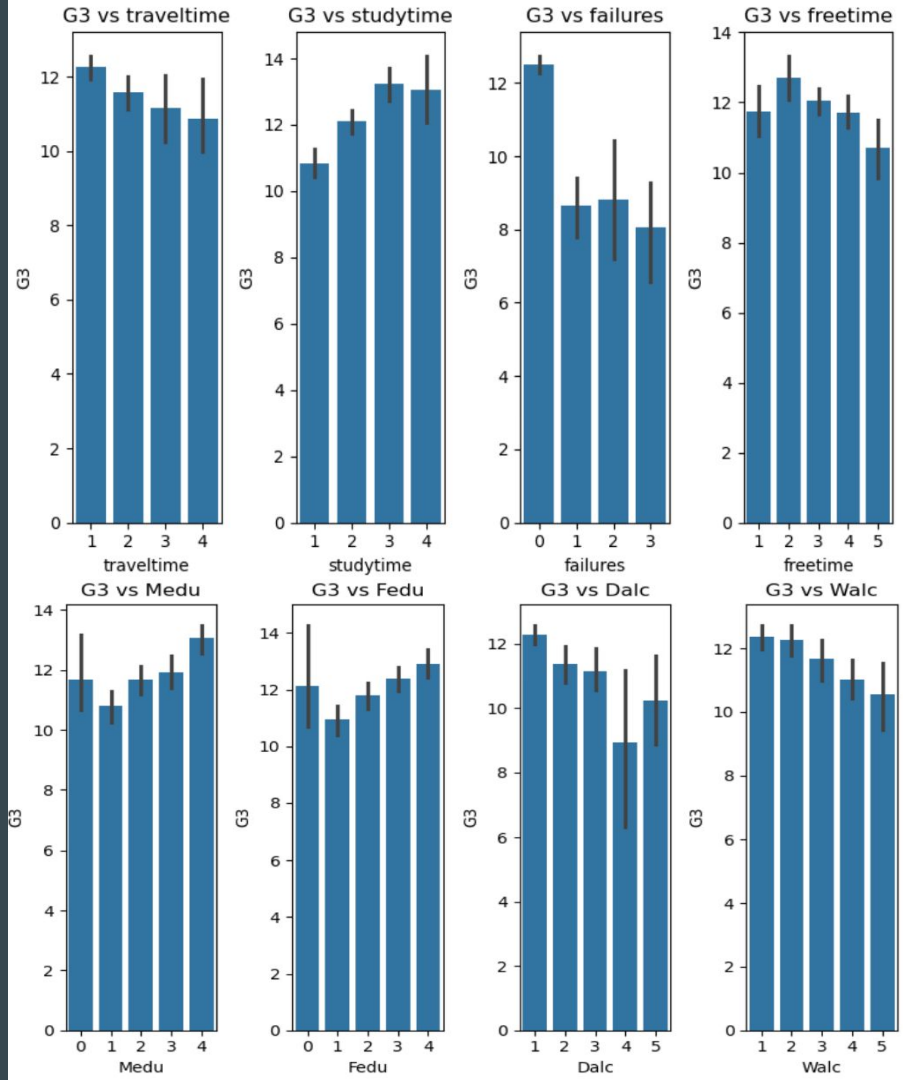
- Key Observations w/ G3
  - “Failures” and final grade have highest negative correlation
  - “Freetime” and final grade have a negative correlation
  - Quality of Family Relationship (“famrel”) and final grade have the least correlation



Key Numerical Values

- Traveltime
- Studytime
- Failures
- Freetime
- Mother/Father education levels
- Workday/Weekend alcohol consumption levels

Most seem expected, but “freetime” stands out...

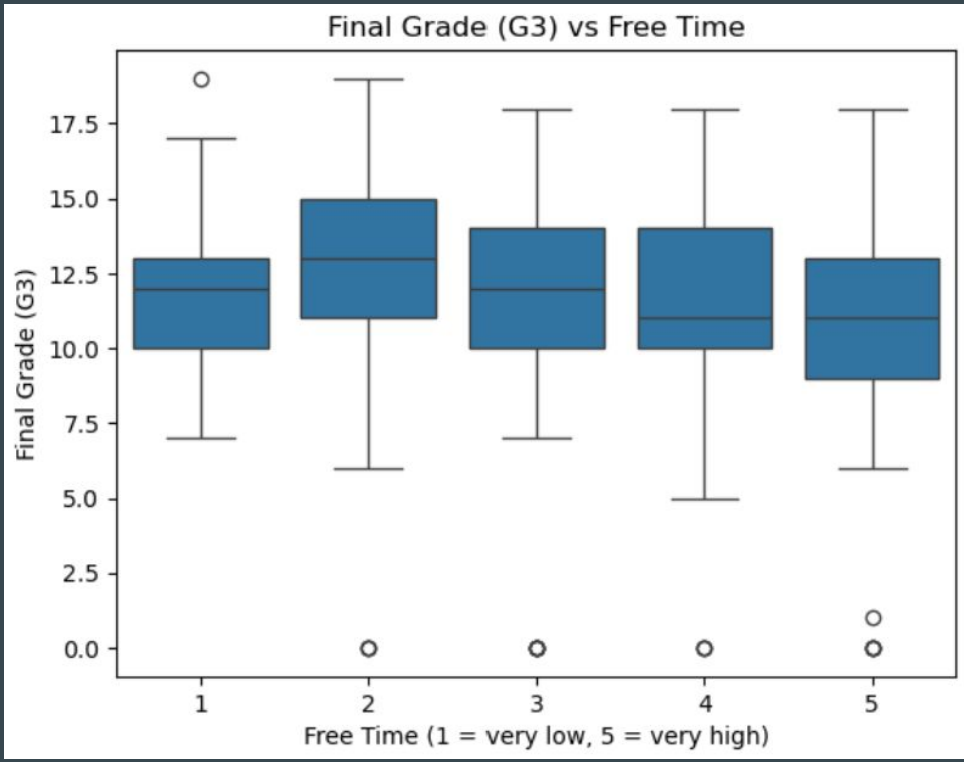


Closer look at Freetime vs. Final Grade

Average grade value per freetime level (rounded):

- 1 → 11.73
- 2 → 12.71
- 3 → 12.06
- 4 → 11.71
- 5 → 10.69

The most freetime actually leads to the lowest average grade!



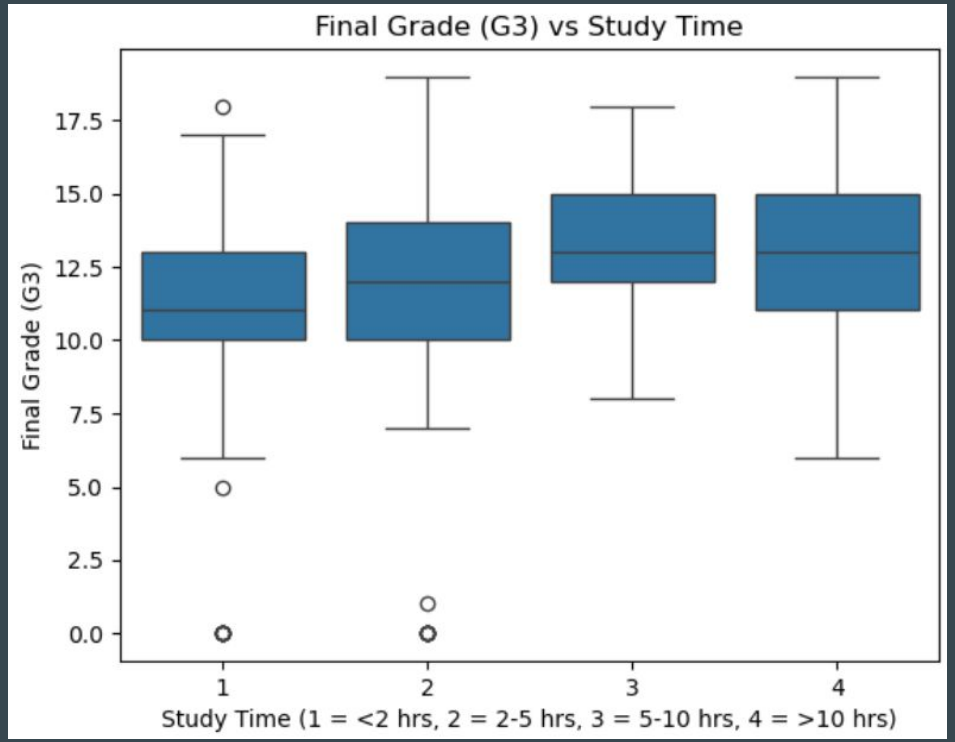
Closer look at Study Time vs. Final Grade

Average grade value per study time (rounded):

- < 2 hours per week → 10.84
- 2-5 hours per week → 12.09
- 5-10 hours per week → 13.23
- > 10 hours per week → 13.06

Grade level seems to plateau after so much studying.

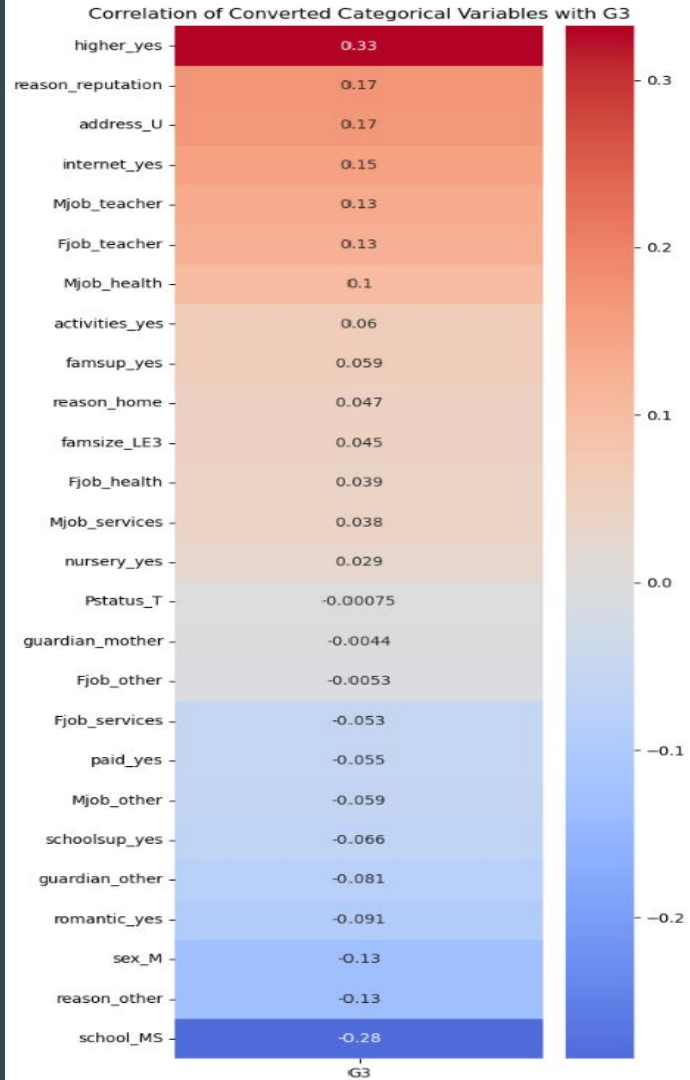
- Interesting how more freetime and more time spent studying both don't necessarily lead to a higher final grade!



Categorical Values Heatmap

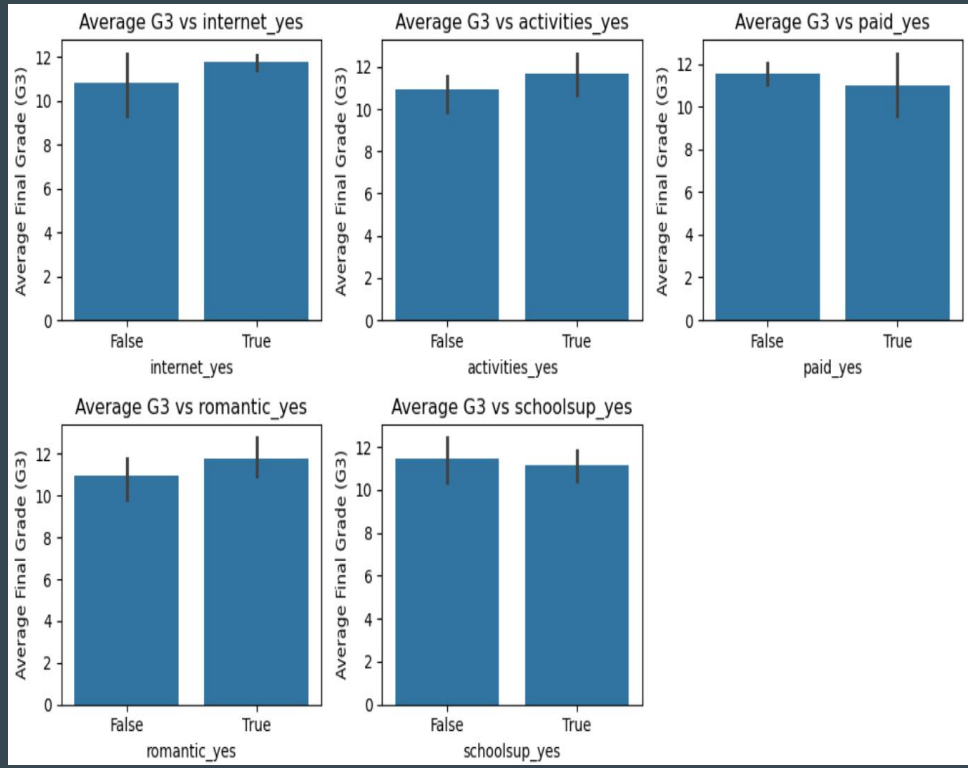
Key Observations w/ G3:

- Most negative correlation
  - Students going to Mousinho da Silveira
- Most positive correlation
  - Those who do want to go into higher education



Key Categorical Values

- Having internet (“internet_yes”)
- Being involved in after school activities (“activities_yes”)
- Extra paid classes within the course subject (“paid_yes”)
- With a romantic relationship (“romantic_yes”)
- Extra educational support (“schoolsup_yes”)

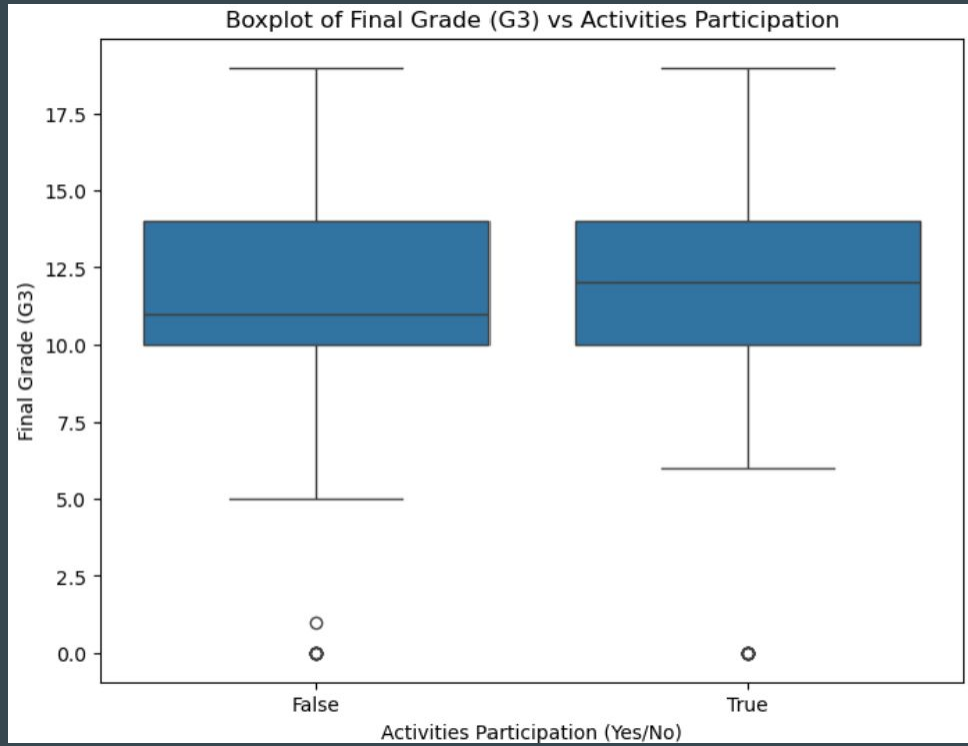


Closer look at Activities vs. Final Grade

Average grade value for after school activity (rounded):

- Not involved in after school activities → 11.72
- Involved in after school activities → 12.10

Surprisingly, those with after school activities on average have a higher final grade



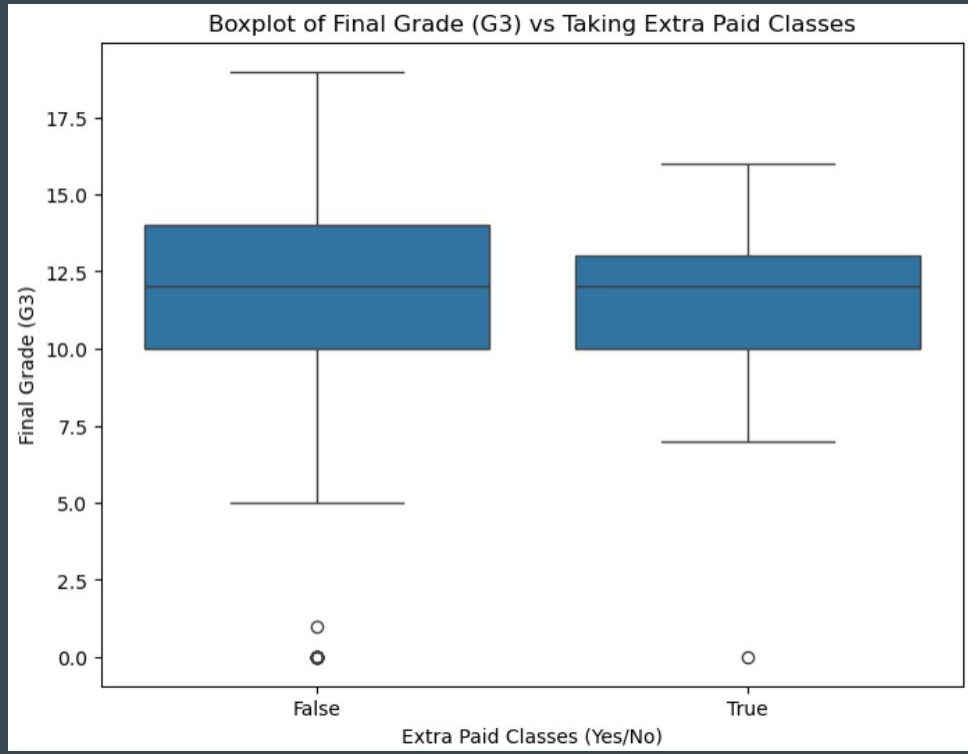
Closer look at Extra Paid Classes vs. Final Grade

Average grade value for extra paid classes (rounded):

- Not taking extra paid classes → 11.95
- Taking extra paid classes → 11.2

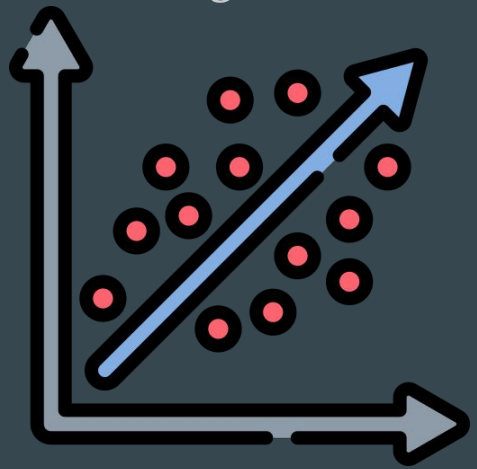
Slightly lower final grade for those taking extra paid classes within a course

Could be because students taking extra classes actually do need more help in the given subject compared others?



Linear Regression Model

- Target Variable
  - **Final Grade** → **G3** (integer/numerical)
  - Will use Linear Regression, not logistic
- Reintroduced G1 and G2 into data set as features, not targets
  - This helped improve my adjusted R^2 value for they are large predictors in the data set
- Calculated Root Mean Squared Value (RMSE), Mean Absolute Percentage Error (MAPE), and adjusted R^2 value



Linear Regression Model (cont.)

Selected Features:

- First period grade, second period grade, travel time, study time, free time, past class failures, weekday alcohol consumption, weekend alcohol consumption, internet, after school activities, paid extra classes, relationship status, and extra school support

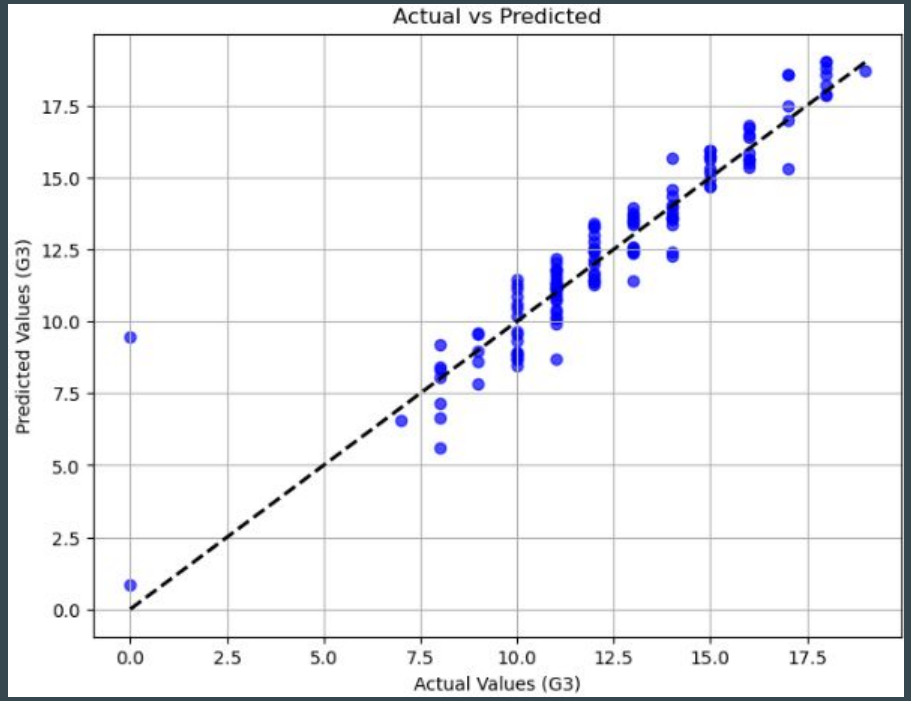
Key Results:

- Root Mean Squared Value (RMSE) → 1.172
- Mean Absolute Percentage Error (MAPE) → .05785 or 5.785%
- Adjusted R^2 (adjusted is optimized for only relevant features) → 0.846

Linear Regression Model Visualization

- Adjusted R^2 → 0.846

Model seems to be reasonably accurate and realistic, but can of course still be improved.



Model Summary

Adjusted R^2 of 0.846

- My model explains about 84.6% of the variability in final grades (G3), adjusting for the number of predictors
- This is a strong result

Root Mean Squared Error of 1.172

- On average, model predicts are about 1.17 grade points off from the actual values, which is good considering G3 ranges from 0 to 20

Mean Absolute Percentage Error of 5.785%

- Model's predictions are on average within 5.785% of the true values, which is pretty accurate

Overall Conclusion

Notable Positive Correlations on Students Final Grade

- Study Time, Wanting to Take Higher Education, Mother and Father Education Levels

Notable Negative Correlations on Students Final Grade

- Past Failures, Alcohol Consumption (weekend or weekday)

Various variables have reasonable effects on a student's potential final grade, and actionable steps should be taken using the data to attempt to improve these scores for the students.