Jason Antal, Grace Lin, Sarah Dominguez,
Quinlan O'Connell, Sam Chen, Cole Brown

Student Success Data Analysis

Dataset Description

We are investigating a dataset that summarizes student performance from two schools in Portugal. The dataset looks at statistics related to a student's study behavior, personal life, family background, interests, and gives their final grade for the school year. Our goal is to identify the best predictors for a student's individual success in school.

Importance of the Problem

The community that would benefit most directly from our study is the Portuguese population; however, our analysis can also provide valuable insights for any sort of education system. By identifying the pattern of what increases student performance, educational institutions can implement target strategies that are more efficient at enhancing learning outcomes.

Additionally, we could consider the results from a different perspective. When two students have similar grades, but one has overcome significant personal challenges, this individual might offer greater personal values. Further discussions about whether these outliers tend to have more entrepreneurial success can be an interesting extension of this study. By exploring this data, we can understand further what predictors outside of school are impactful on a student's performance.

Exploratory Analysis

Our first initiative was to explore the distribution of the target variable: final_grade. With this, we found that this predictor is normally distributed, with most students falling between a final grade of 10-15 (10 being passing). However, it should also be noted that there is a chunk of students who fail, with final grades of 0.0. The dataset also includes predictors G1 and G2, with explicit values of how the student performed in the first ⅓ and ⅔ of the course. However, the data guide mentioned that dropping these predictors made predicting final_grade more difficult, so we decided to drop both predictors.

After this initial analysis of the target variable and what predictors to remove, we further explored the characteristics of a wide range of data. For example, we found that most students' ages fall between 16 and 18; however, there is a max age of 22. While most students have never failed a class, a small number of students have 1 to 3 failures. Most students are generally happy with their family situation, and tend to not consume alcohol

during the weekdays. 79% of students have internet access, 36% are in a romantic relationship, and 11% receive additional school support. Parents with similar education levels are likely to be together because there is a strong positive correlation at .64. Older students are a little more likely to fail with a correlation of .28. Study time and final grade are only correlated at .16. Internet access and final grade are also weakly correlated at .11.

Through our EDA, we also engineered a new feature called "growth", obtained by calculating how much a student's scores improved from one semester to the next. The goal in creating this predictor was to take into account the impact of a student's grade improvement across a short period of time and apply that growth across the entire school year.

Interestingly, students that displayed negative growth usually failed the course - showing that there is a high likelihood that either parents or teachers should step in if they see a student performing worse in consecutive semesters.

Our solution to the problem of improving educational performance was to apply a variety of machine learning methods to successfully predict which factors influenced a student's final grade.

Solution:

The problem in question is a regression problem, where we were looking to predict the column “final_grade” from our predictors. Our data consisted of 40 columns, but we decided to leave all columns except G1 and G2, as many of these columns consisted of dummy variables. For our error metrics, we decided to use RMSE to understand how accurate our predictions may be, and R-squared to measure the amount of variance we can accurately predict.

Summary of Methods Applied:

- **Linear, Lasso, and Ridge Regression:** Using an alpha of 0.01 for Lasso and Ridge, these three models all reached a RMSE of approximately 3.422, with an R squared of 0.318.
- **KNN:** This model showed moderate performance with an R-squared of 0.409, meaning it explained 40.9% of the variance in final grades. The RMSE of 3.427 indicates that the predictions are on average 3.4 units off from the actual grades. While the model provides some predictive accuracy, the MSE of 11.744 suggests there's room for improvement.
- **Regression Tree:** With an initial tree depth of 5, we obtain a RMSE of 3.429. Conducting a grid search over just the size of the tree from 1 to 10, we find that an optimal tree size of 3 yields an RMSE of 3.200. Trimming the tree works, but overfits onto the 'failures' predictor,

which measures if a student has failed before.

- **Bagging:** We employed hyperparameter tuning to enhance this model's performance. The best combination of hyperparameters identified by GridSearchCV were: bootstrap=False, bootstrap_features=False, max_features=0.7, max_samples=0.5, and n_estimators=200. The model achieved a RMSE of approximately 3.098 on the test set. The model's R-squared value was 0.441, indicating that approximately 44.1% of the variance in the final grades can be explained by the model.
- **Gradient Boost:** Using an initial pass of 100 trees with a max depth of 5, we obtained an RMSE of 3.194. After conducting cross-validation over a grid search with 10 folds, the optimal RMSE over the grid was 3.095. Optimal parameters were learning rate of 0.1, max depth of 7, min samples split of 9, and number of estimators at 200 trees.
- **Random Forest:** The most accurate model was a Random Forest Regressor for prediction, alongside hyperparameter tuning. The optimal hyperparameters here were: bootstrap=True, max_depth=None, min_samples_leaf=1, min_samples_split=5, and n_estimators=200. This model achieved a RMSE of 3.084, which was our best performing model overall. The associated R-squared value was 0.446. This model would be the one we select for future predictions.

Further Investigation:

The best model that we found after converting the G1 and G2 grades into a growth metric was the random forest model. A random forest model takes only a few variables to fit weak learning trees. This produces a sizable improvement from bagging since selecting fewer variables than the bagging case decreases co-correlation and influence from noisy variables that don't really matter. Interestingly, in such noisy data, gradient boost performs marginally worse than random forest, since a gradient boost model can fit sequentially on the errors of previous models.

As we add back in G1 and G2, we find that all the models fit directly onto those two variables and they become by far the most important variables. This makes sense because G1 and G2 are themselves target variables of the rest of the predictors at the time of their recording. Implementing G1 and G2 as predictors only serves to bake in the information that we already knew. It can be argued that fitting a regression model with G1 and G2 as predictors tends to have the models overfit on a limited set of variables.

Again, random forest performs the best. An explanation is that there simply

isn’t that much error on each iteration of a gradient boost for additional fitting. Provided that a random forest tree captures either G1 or G2, the result would be a more accurate model.

Insights of the Study

There are a few key insights important to the overall study apart from a student's previous failures:

- Negative growth is a strong signal for student course failures
- Absences are highly correlated with student failures
- Higher education aspirations influence the decision of students to invest in their studies
- Parental education can have an influence on educational success

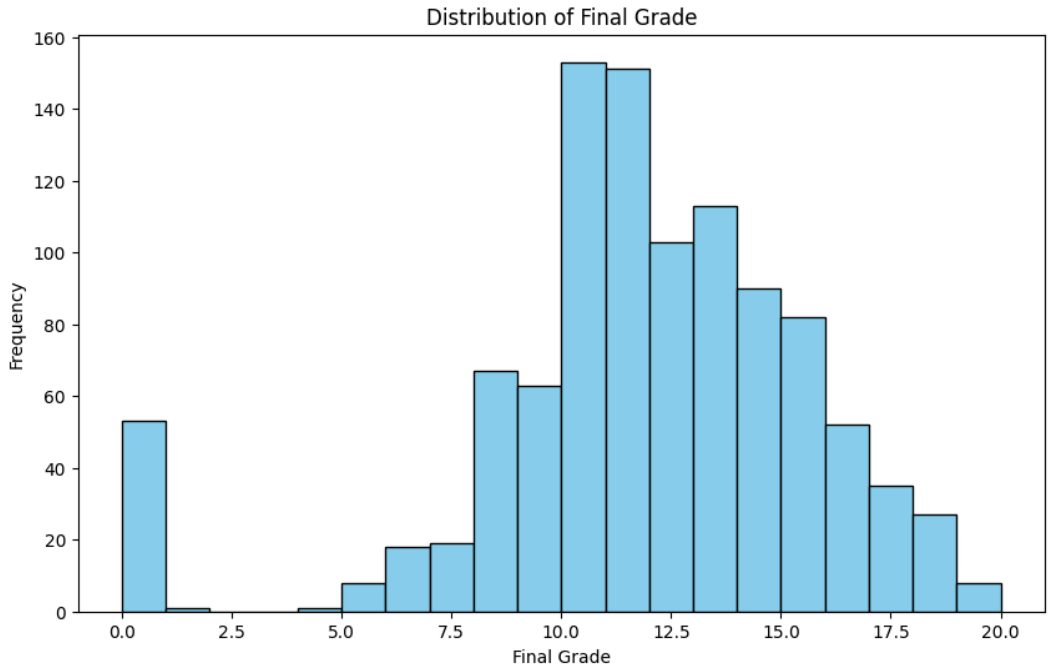
Applications, Future Potential, & Problem Conclusion

Through this analysis, it became clear that while a student’s performance in the classroom is highly indicative of their final grade, other external factors also contribute to setting them up for successful performance in the classroom.

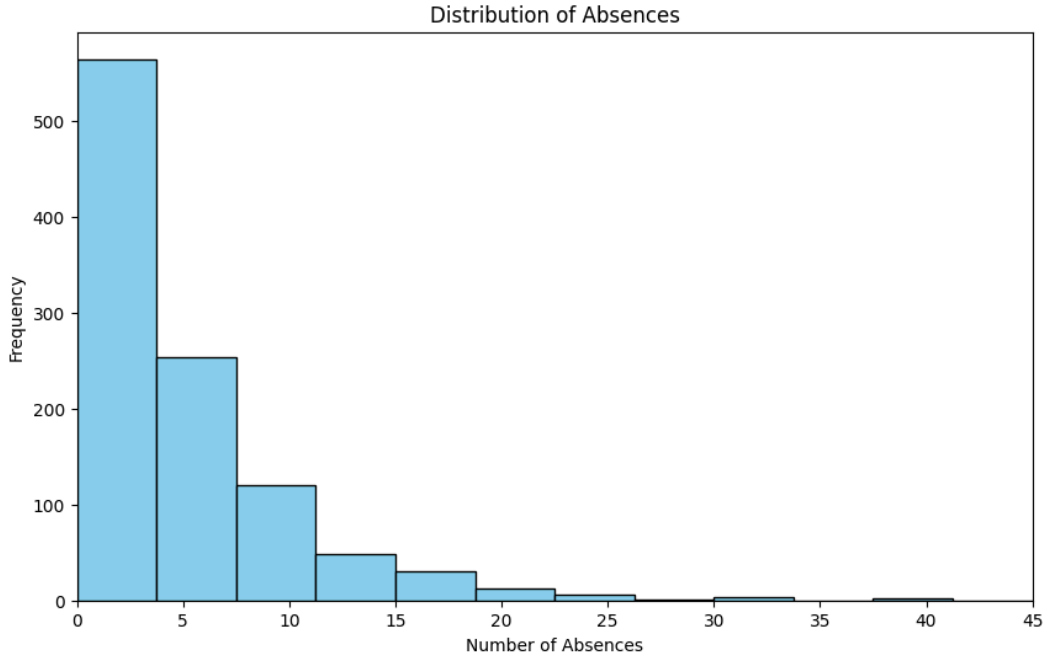
Our analysis confirmed logical explanations of performance. Students who were focused on their studies and a college education tended to do better. Students who had difficult circumstances at home or outside of school tended to perform worse. Things like a lack of Wi-Fi, a high number of absences, or a previously failed class indicated a worse predicted grade, which may indicate a student who is not very “serious” about their studies. Moving forward, these models could be used to identify students at risk of failing, and provide extra support and resources towards helping these students succeed. Models like these could also identify those likely to be high performers, which could allow schools to allocate more challenging coursework to driven students. Schools consistently gather data on students' current grades in classes, and by incorporating this data into predictive models, the models' accuracy and performance could be further enhanced.

Overall, this problem was an interesting exploration into the school system. Many students throughout the world lack proper education, or the tools to succeed in their education. Underdeveloped countries struggle with creating systems that produce successful students. Through research like this, we can make meaningful strides in identifying the most effective ways to support struggling students, creating an immediate and lasting impact on building the future leaders of tomorrow.

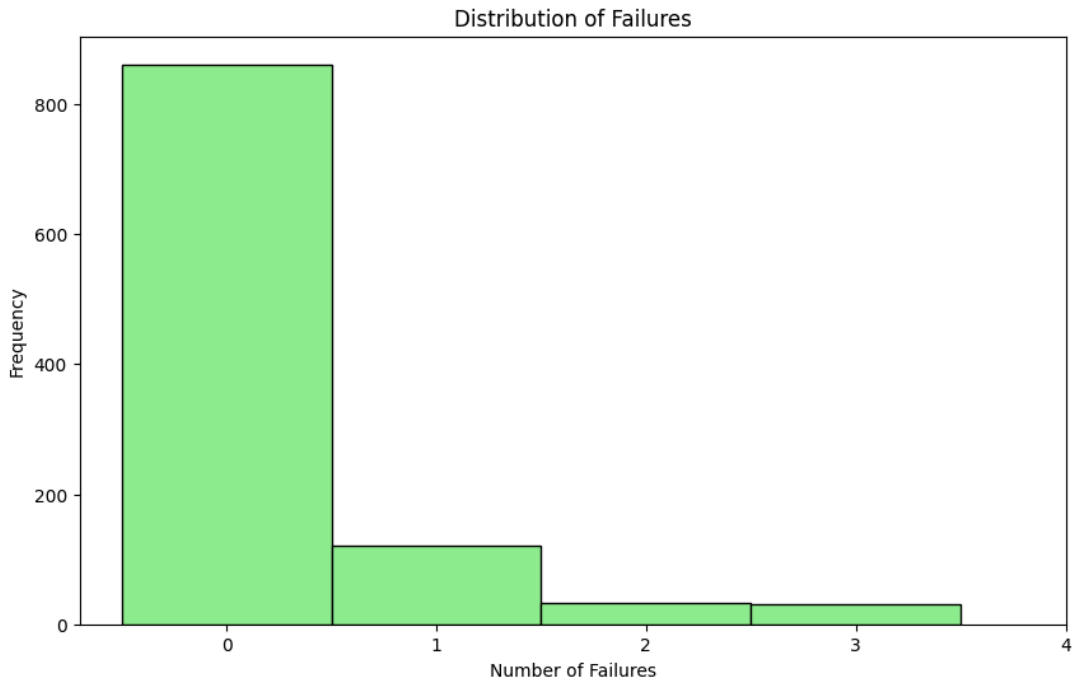
APPENDIX



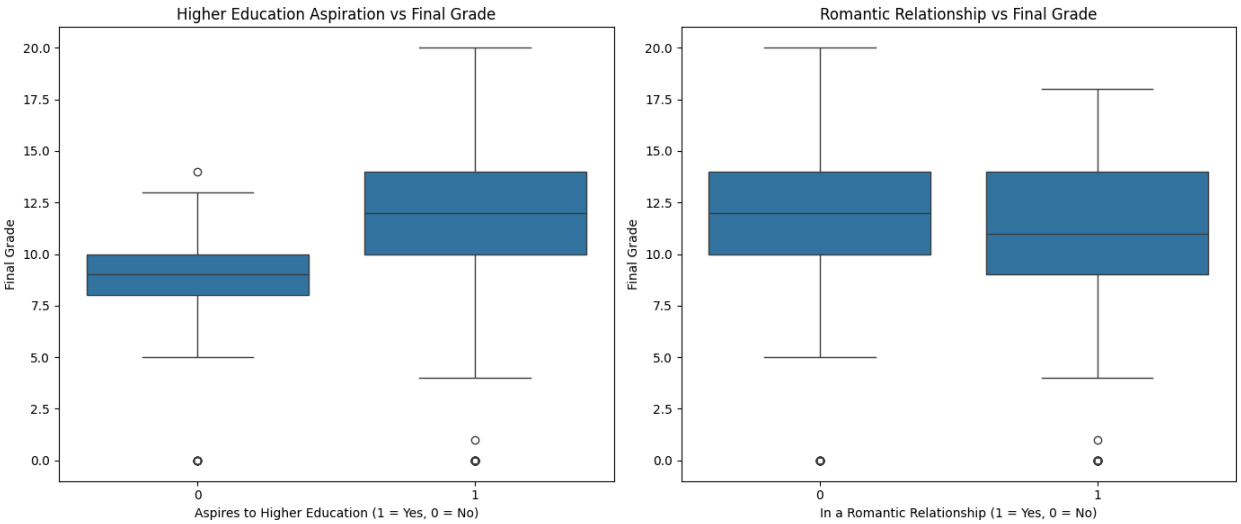
Distribution of Target Variable



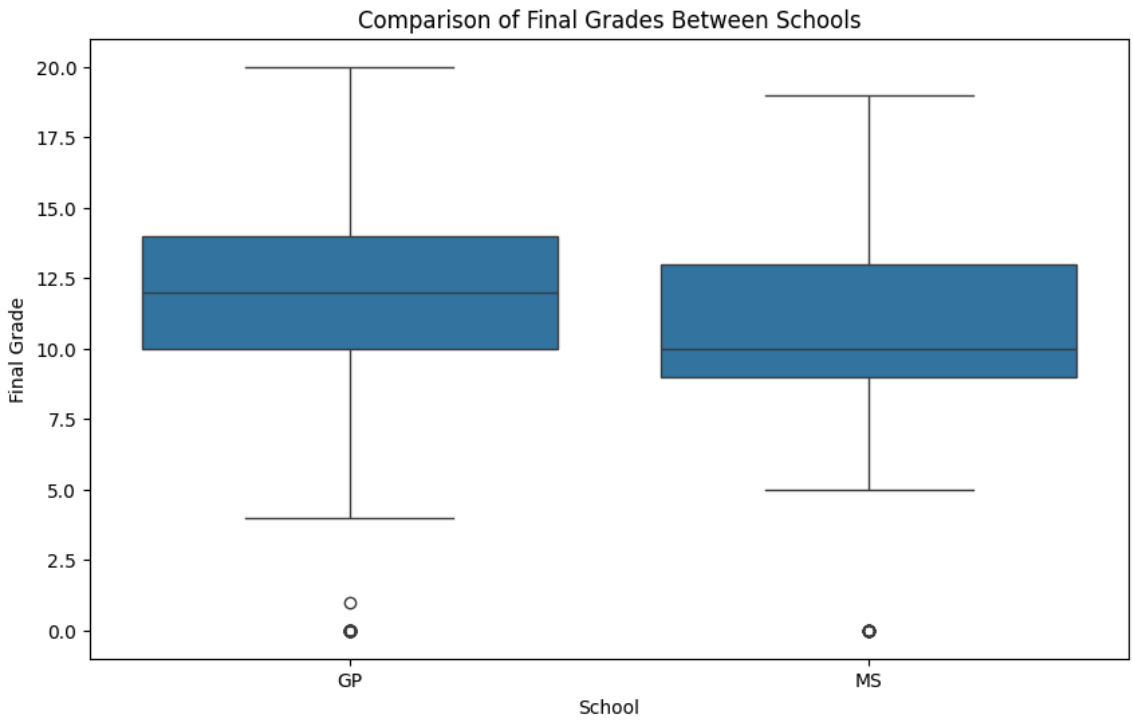
Distribution of Student Absences



Distribution of Student Failures



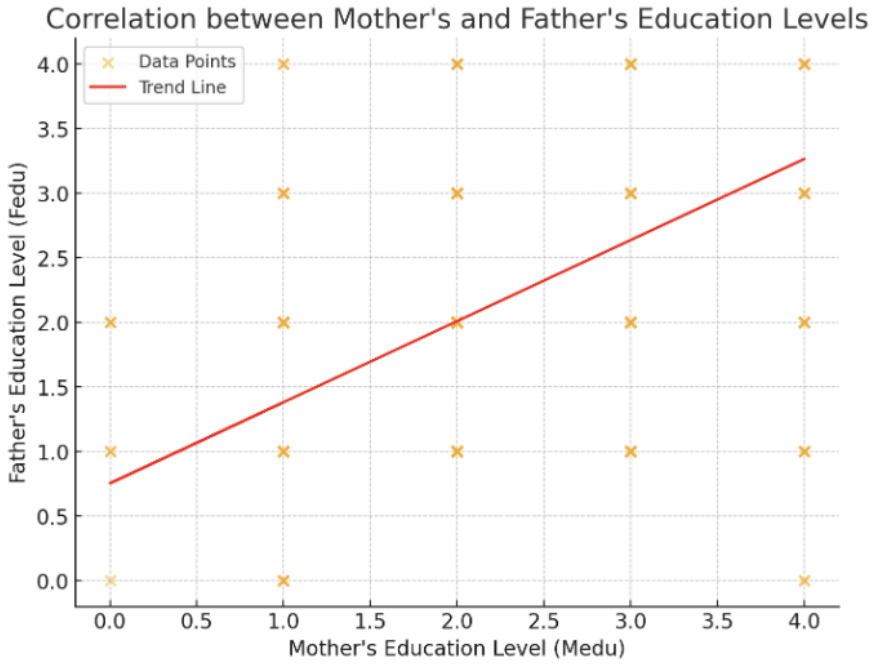
Higher Education vs. Final Grade & Relationship vs. Final Grade



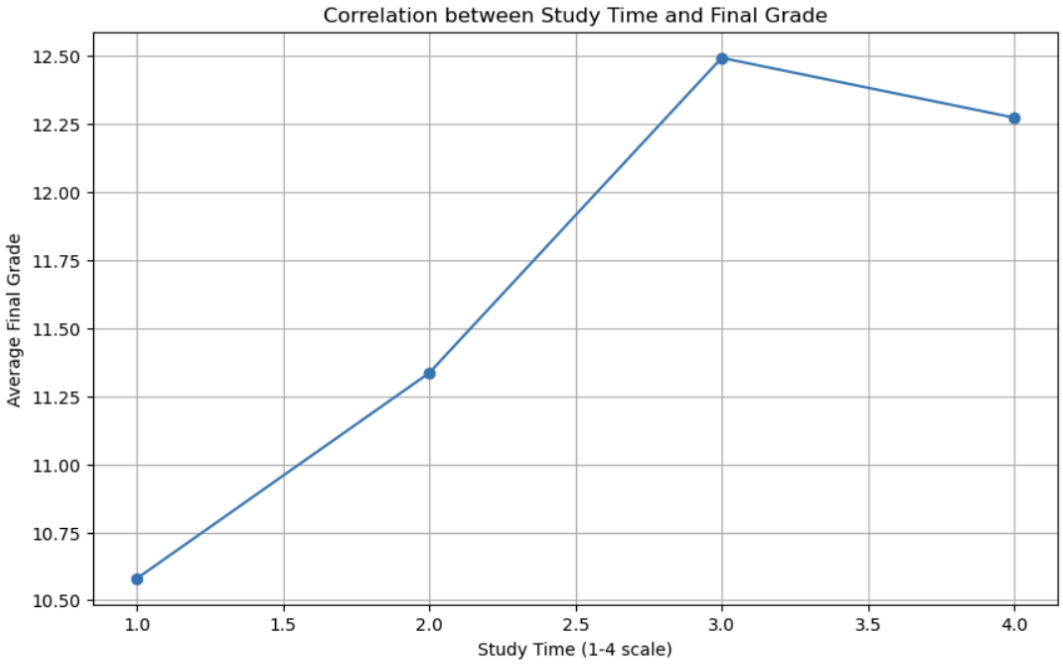
School Grade Comparison

	final_grade
age	0.125282
Medu	0.201472
Fedu	0.159796
traveltime	0.102627
studytime	0.161629
failures	0.383145
Dalc	0.129642
Walc	0.115740
school_MS	0.127114
address_U	0.117696
Mjob_health	0.101349
Fjob_teacher	0.101361
reason_reput…	0.121303
higher_yes	0.236578
internet_yes	0.107064
growth	0.341705

Highly Correlated Predictors with Final Grade

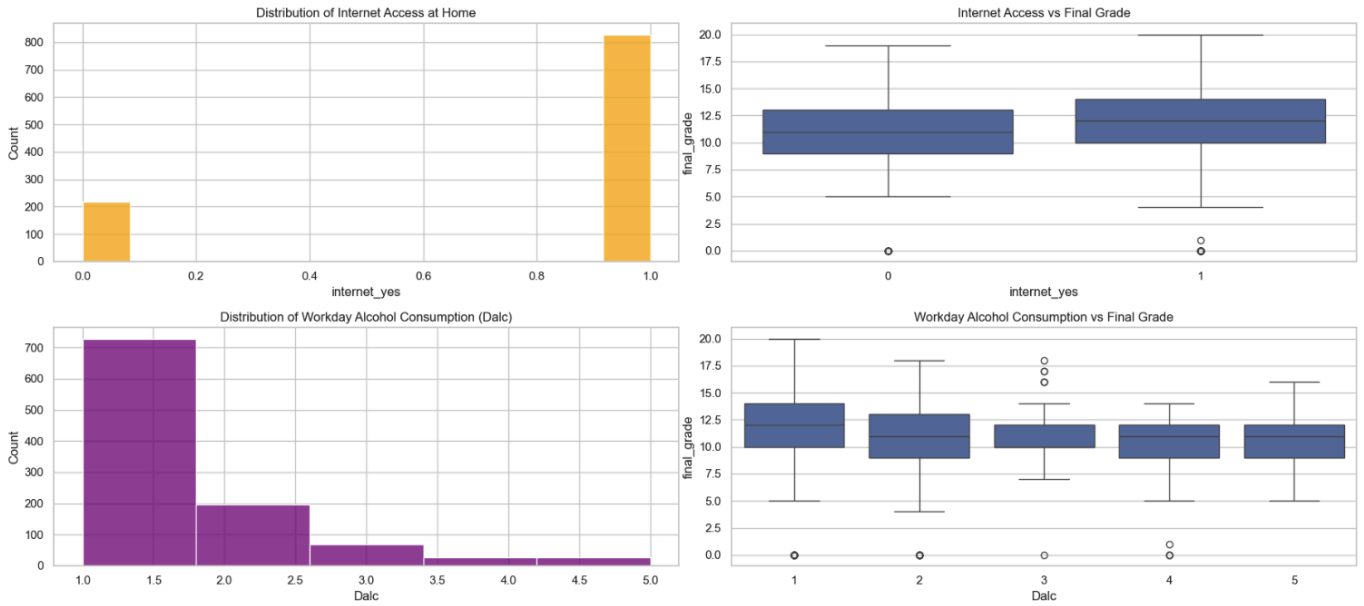


Education Correlation

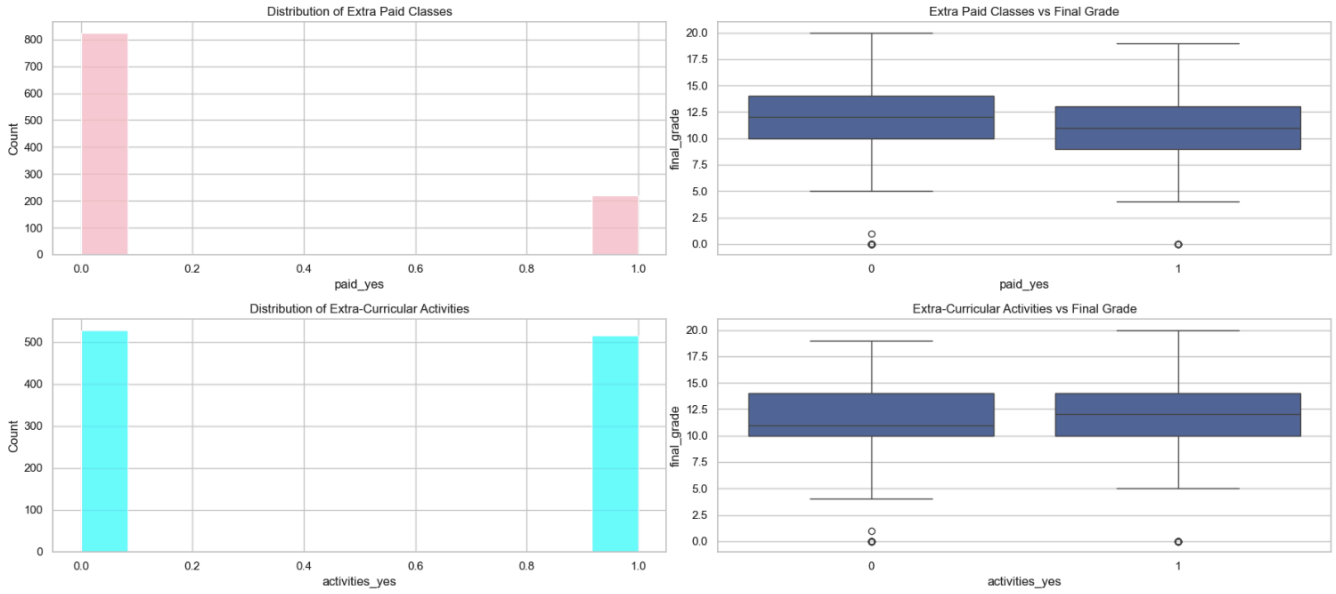



```
Correlation: 0.1616289352029381
```

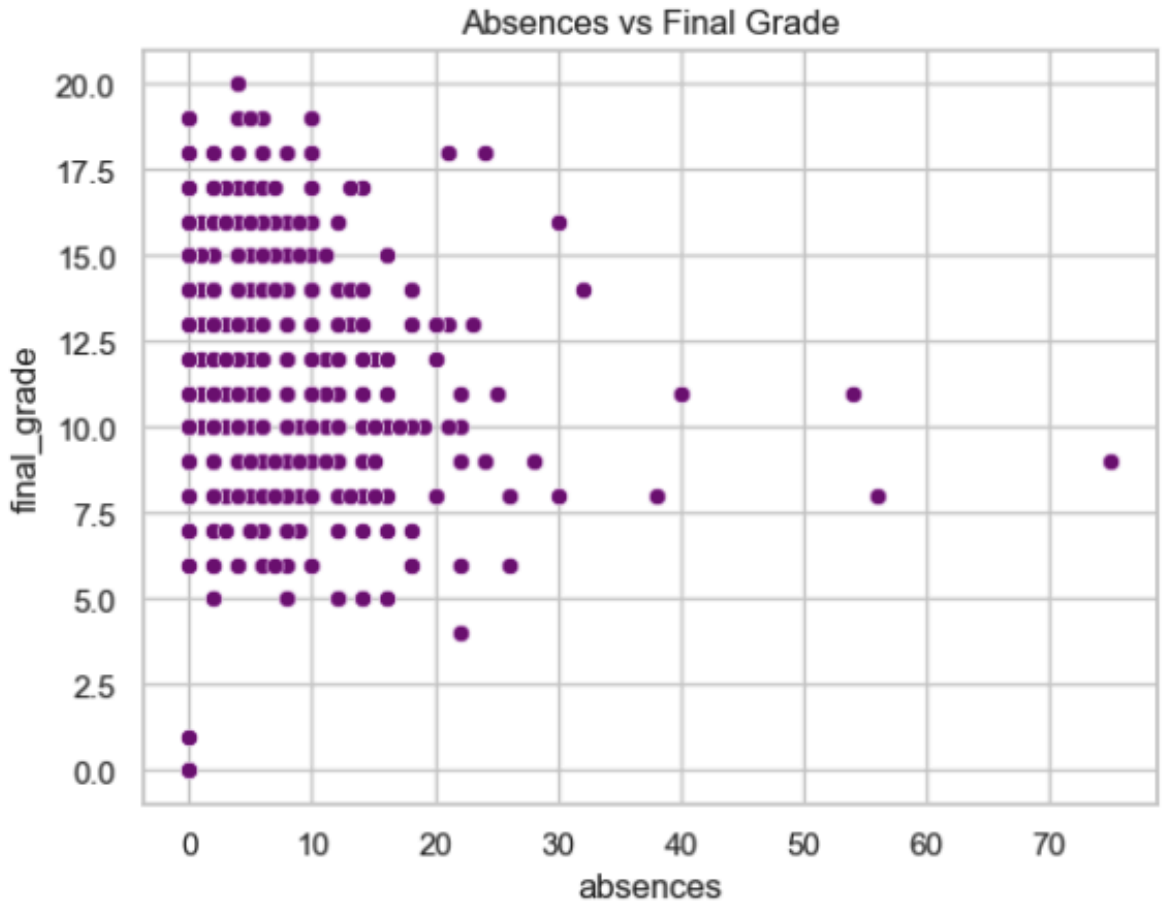
Amount of Studying vs. Grade Received



Internet Access at Home (top), Workday Alcohol Consumption (bottom)



Additional Tutored Classes (top), Extra-Curricular Activities (bottom)



Effect of Absences on Performance