1. Introduction

1.1 The Data

The dataset we chose to conduct our analysis is the ‘Student Performance’ dataset obtained from the UC Irvine Machine Learning Repository. It measures student achievement in secondary education across two Portuguese schools, Gabriel Pereira and Mousinho da Silveira, during the 2005-2006 school year (Cortez, 2014). The dataset consists of 33 variables, including student grades, demographic information, and social and school-related factors, all collected from school reports and questionnaires, with no missing values. The variables we will be exploring are as follows:

Response variable:

- G3: Final grade, ranging from 0 to 20 (output target).

Explanatory variables:

- sex: Student's sex: 'F' (female) or 'M' (male).
- age: Student's age, ranging from 15 to 22.
- address: Student's home address type: 'U' (urban) or 'R' (rural).
- famsize: Family size: 'LE3' (less or equal to 3) or 'GT3' (greater than 3).
- Pstatus: Parent's cohabitation status: T' (together) or 'A' (apart).
- Medu: Mother's education level, coded from 0 (none) to 4 (higher education).
- Fedu: Father's education level, coded from 0 (none) to 4 (higher education).
- Mjob: Mother's job, coded as 'teacher', 'health', 'services', 'at_home', or 'other'.
- Fjob: Father's job, coded as 'teacher', 'health', 'services', 'at_home', or 'other'.
- reason: Reasons to choose the school, coded as 'home', 'reputation', 'course', or 'other'.
- guardian: Student's guardian, coded as 'mother', 'father', or 'other'.
- traveltime: Home to school travel time, coded from 1 (<15 min) to 4 (>1 hour).
- studytime: Weekly study time, coded from 1 (<2 hours) to 4 (>10 hours).
- failures: Number of past class failures, coded from 0 (none) to 4 (4 or more).
- school_support: Extra educational support, coded as 'yes' or 'no'.
- family_support: Family educational support, coded as 'yes' or 'no'.
- paid: Extra paid classes within the course subject, coded as 'yes' or 'no'.
- activities: Participation in extracurricular activities, coded as 'yes' or 'no'.
- nursery: Attended nursery school, coded as 'yes' or 'no'.
- higher: Plans for higher education, coded as 'yes' or 'no'.
- internet: Home internet access, coded as 'yes' or 'no'.
- romantic: In a romantic relationship, coded as 'yes' or 'no'.
- famrel: Quality of family relationships, rated from 1 (very bad) to 5 (excellent).
- freetime: Free time after school, rated from 1 (very low) to 5 (very high).
- goout: Going out with friends, rated from 1 (very low) to 5 (very high).
- Dalc: Workday alcohol consumption, rated from 1 (very low) to 5 (very high).
- Walc: Weekend alcohol consumption, rated from 1 (very low) to 5 (very high).
- health: Current health status, rated from 1 (very bad) to 5 (very good).
- absences: Number of school absences, ranging from 0 to 93.

Note: We have excluded the variables 'school' (name of school; included as part of background of data) and 'G1' and 'G2' (first and second period grade; G3 is a function of G1 and G2 so they are strongly correlated).

1.2 Motivation & Research Question

As students ourselves, naturally, we are interested in understanding the factors that contribute to academic success, and how our backgrounds and choices influence our performance. As such, this dataset caught our eye, as it can

provide insight into our own lives and our academic careers. Two data files are included in the dataset, one for the subject of Math and one for Portuguese. We chose to focus on the data file for 'Math', as our group consists of Math and Statistics students, and this is more applicable to our fields of study. This leads us to our research question:

What external factors are associated with a student's academic performance?

2. Analysis

2.1.1 Exploratory Data Analysis: Covariates

To get a preliminary idea of which variables are interesting to explore, we begin by visualizing each covariate to see if there are potential associations with our response variable. As there are more than 20 covariates in the data, we can use this step to do an initial manual selection of covariates, and then investigate the significance of the ones we have chosen through inferential statistical methods.

After plotting the covariates, we discovered the ones we would like to explore are:

1. Sex
2. Address
3. Mjob
4. Fjob
5. Guardian
6. Schoolsup
7. Higher
8. Internet
9. Age
10. Medu
11. Fedu
12. Traveltime
13. Studytime
14. Failures
15. Famrel
16. Free_time
17. Goout
18. Dalc
19. Walc
20. health

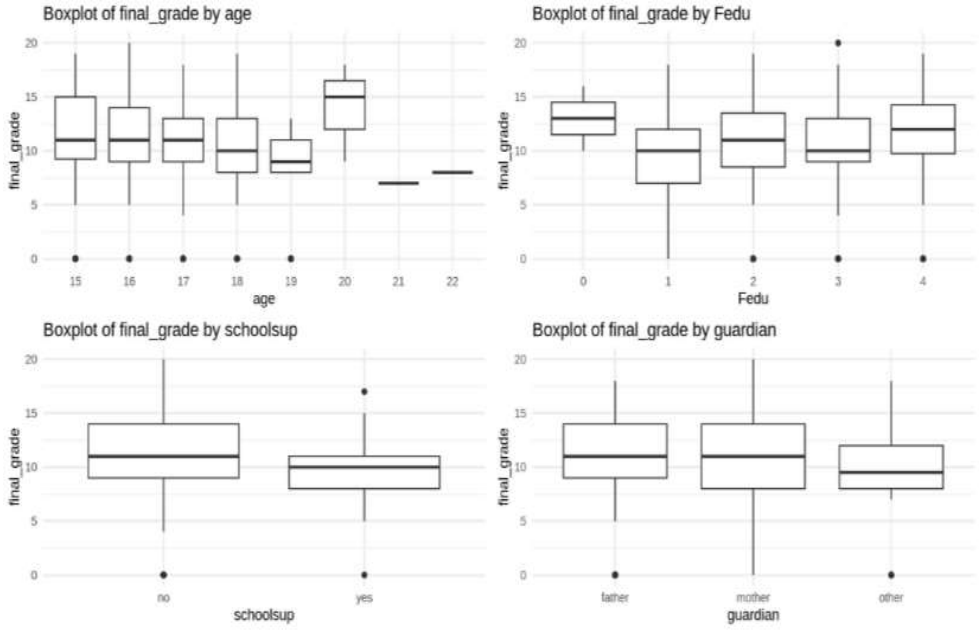

Figure 1. *Boxplots of Some Covariates We Chose to Use*

As we can see in Figure 1, there are differences in medians between the different levels, suggesting there are differences that could be significant and lead to further insights. Upon observing the boxplots, we observe the following for each covariate:

- Sex: Males seem to achieve higher academic performance than females.

- Address: Students who live in urban areas achieve higher performance than those who live in rural areas.
- Mjob: Students whose mothers are in health related occupations perform better.
- Fjob: Students whose fathers are in teaching related occupations perform better.
- Guardian: Students whose guardians are not their parents perform worse.
- Schoolsup: Students with no school support perform better.
- Higher: Students who plan to pursue higher education perform better than those who do not.
- Internet: Having access to the internet appears to lead to higher performance
- Medu/Fedu: Students whose mothers and fathers who did not receive education perform better
- Traveltime: Students who travel less than 15 min to school perform better.
- Studytime: Students who study less than 2 hours perform worse.
- Failures: Students' performance decreases as the number of past class failures increases.
- Famrel: Students who have very bad quality of family relationships seem to perform better.
- Free_time: Students with low free time after school perform better.
- Goout: Those who do not go out with their friends often perform better.
- Dalc: Those who consume very low amounts of alcohol on weekdays perform better.
- Walc: Those who consume very low to low amounts of alcohol on weekends perform better.
- Health: Students with very bad health perform better.

The ones we found do not seem to be associated with academic performance are:

1. Famsize
2. Pstatus
3. Reason
4. Famsup
5. Paid
6. Activities
7. Nursery
8. Romantic
9. Absences

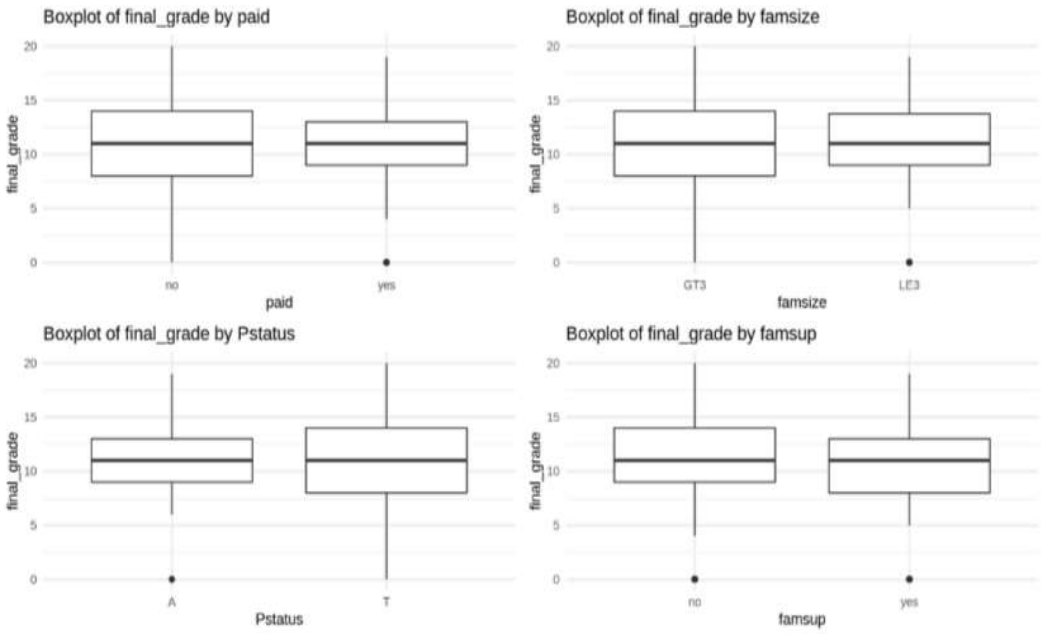

Figure 2. *Boxplots of Some Covariates We Chose Not to Use*

As we can see in Figure 2, the medians between the different levels are approximately the same, and the quartiles are similar as well.

2.1.2 Exploratory Data Analysis: Data as a Whole

Before fitting the model, we prepared the data by examining the distribution of the response variable. The initial distribution is shown in Figure 3:

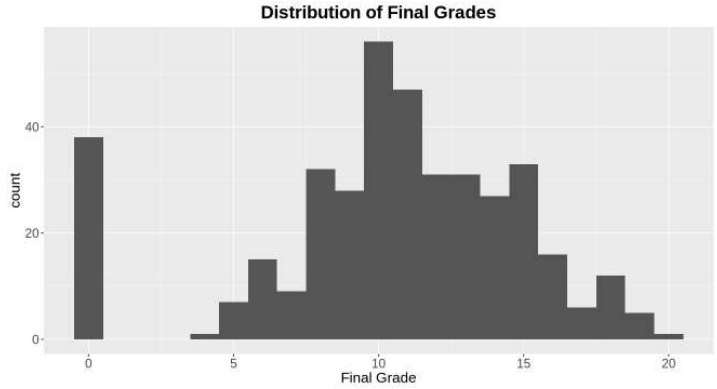

Figure 3. *Distribution of Final Grades Before Removing Outliers*

As we can see in Figure 3, the distribution of the data exhibits an extreme on the left side of the distribution, corresponding to a grade of 0. This anomaly likely reflects cases where students did not attempt the exam, faced extenuating circumstances, or dropped out. To focus on data that provides meaningful insights for our research question, this extreme value was removed, resulting in the distribution shown in Figure 4:

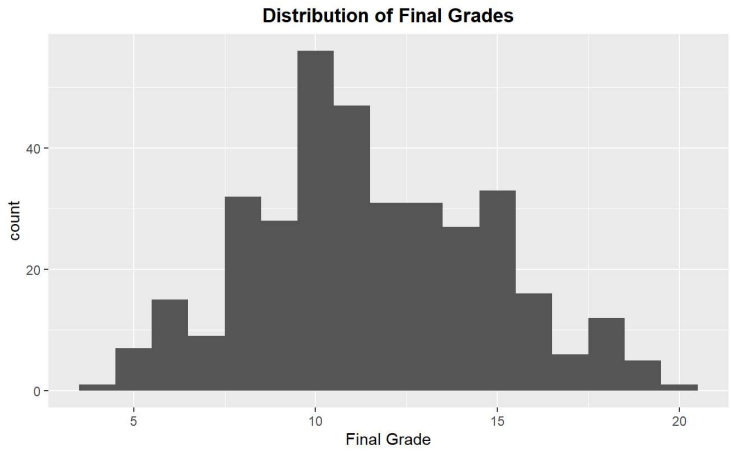

Figure 4. *Distribution of Final Grades After Removing Outliers*

From Figure 4, we can see that the distribution is roughly symmetric and bell-shaped. This suggests a linear regression *may* be appropriate for our data. However, in Section 2.2.2, we will check if the linear assumptions have been met to confirm.

2.2.1 Linear Regression

To explore how the chosen covariates are associated with students' academic performance, we fit a linear regression model. We will treat our response variable 'final_grade' as a continuous response variable, despite it being inherently discrete with a range of 0 to 20. This approach was chosen for simplicity and interpretability, as it allows us to model the relationship between students' grades and predictors using regression methods. Since each grade represents a consistent increase of one point for academic performance, it is reasonable to treat the variable as

continuous for modeling purposes as it aligns with how continuous variables generally behave. This approach allows us to model the relationship between students' grades and the predictors in the dataset, providing insights into how various factors influence academic performance.

2.2.2 Model Diagnostics

To apply linear regression, several assumptions need to be met:

1. **Linearity**: The relationship between predictors and the response variable must be linear.
2. **Independence**: Observations should be independent, though this assumption should ideally be validated.
3. **Homoscedasticity**: Residuals should exhibit constant variance across levels of the independent variables.
4. **Normality**: Residuals should be approximately normally distributed.

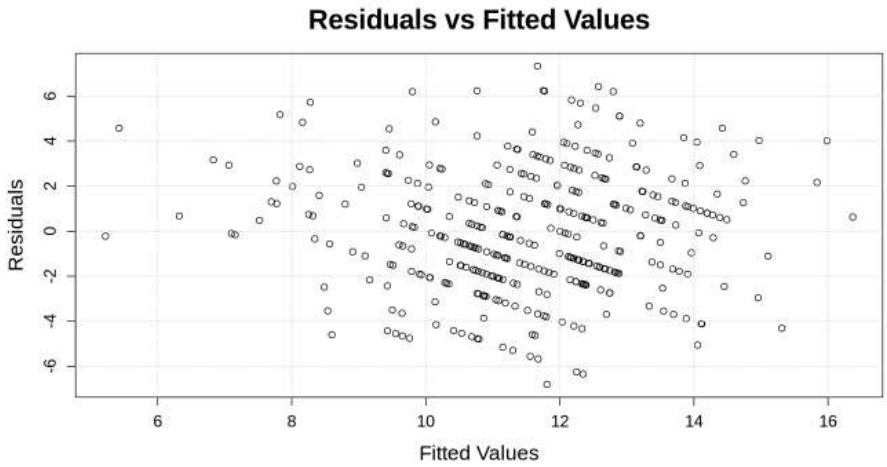

Figure 5. *Residuals vs. Fitted Values: Diagnostic Plot for Homoscedasticity*

As we can see in Figure 5, the residuals of the linear model are randomly spread, indicating that the homoscedasticity assumption was met. While there is an appearance of distinct lines, it is likely the result of the discrete nature of the response variable. The randomness remains so it does not seem to indicate a violation of the model assumptions.

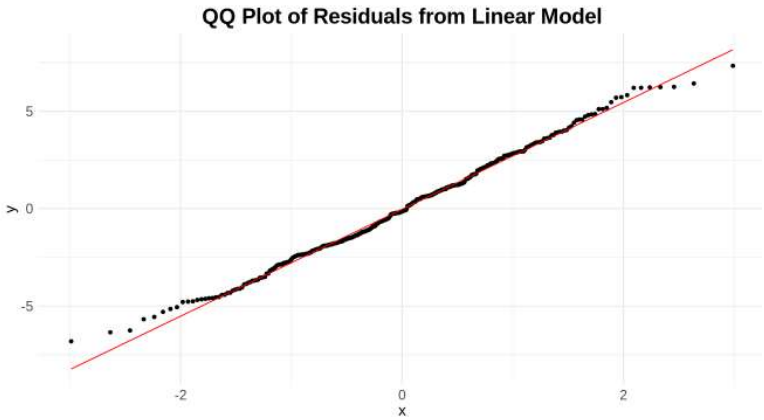

Figure 6. *QQ-Plot For Assessing Normality of Residuals*

To assess the normality assumption, a QQ-plot of residuals was used. From Figure 6 we can see that the residuals closely follow the reference line, with minor deviations in the tails, which suggests some light-tailed behaviour. However, these deviations are small and unlikely to significantly affect the validity of the model's coefficient

estimates or hypothesis tests. Overall, the assumption of normality of residuals appears reasonably satisfied for the purposes of this analysis.

These diagnostic checks indicate that, while there are minor deviations due to the discrete nature of the response variable, the assumptions of linear regression were sufficiently met. As such, we can proceed with the analysis and draw reliable inferences.

2.2.3 Model Selection

To pick the best model to infer the relationship between the response variable of academic performance and the covariates we have chosen to investigate, we applied the best subset selection algorithm. This method evaluates all possible combinations of predictors to identify the subset that best balances explanatory power and model simplicity. The selected model is then compared with the full model to ensure that the reduced model retains the ability to provide meaningful insights into the factors influencing student performance.

After performing best subset selection we identified the 'best' model by evaluating R^2, adjusted R^2 and Mallows' C_p across all the different combinations of these variables. The selection process considers all levels of the categorical variables which resulted in 26 models. Mallows' C_p helped us evaluate the fit of a regression model relative to the full model; as we can see in Figure 7, the C_p (10.004246) of Model 8 is the closest to $p + 1 = 9$ (where p is the number of covariates in each specific model), indicating a good balance between interpretability and accuracy. In contrast, the other models had C_p values that deviated significantly from p +1.

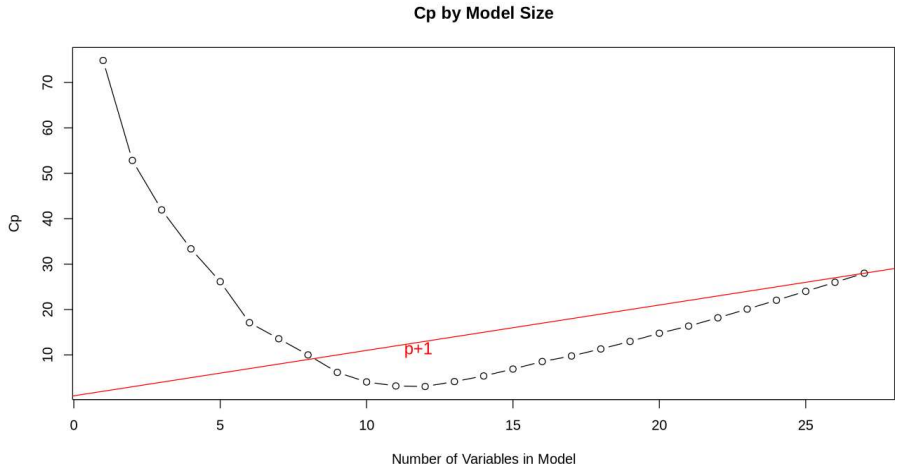

Figure 7. *C_p by Model Size*

For the R^2 value, it increased as the number of covariates increased, which was expected, therefore we observed the adjusted R^2 for more insight. While Model 12 (with 12 covariates) had the highest adjusted R^2 (0.2617665), its C_p (3) was significantly lower than p +1 (13), suggesting there may be a bad model fit. In comparison, Model 8 had an adjusted R^2 of 0.2375837, which is only slightly lower than the highest adjusted R^2 value.

Therefore, by applying the principle of parsimony, we ultimately chose the simpler model, Model 8, as it offers similar explanatory power compared to more complex models while having fewer covariates (and a better C_p) and therefore being easier to interpret.

Finally, when comparing Model 8 with the full model, we can see that our chosen model offers a comparable adjusted R^2 while sacrificing very little explanatory power for the benefit of simplicity. This makes our model simpler to work with and easier to interpret.

Model	R-squared	Adjusted R-Squared	Mallow's Cp
Full Model	0.2974348	0.2397775	28.000000
Best Selection Model	0.2547166	0.2375837	10.004246

2.2.4 Interpretation: Reduced Model Selected

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   12.3813     0.9605  12.890  < 2e-16 ***
sexM           0.7890     0.3250   2.428 0.015705 *
Mjobhealth     1.7888     0.6653   2.689 0.007523 **
Mjobother     -0.1324     0.4834  -0.274 0.784308
Mjobservices   1.2501     0.5082   2.460 0.014398 *
Mjobteacher   -0.1414     0.5913  -0.239 0.811084
Fjobhealth    -0.8359     0.9866  -0.847 0.397474
Fjobother     -0.8160     0.7262  -1.124 0.261997
Fjobservices  -0.8402     0.7546  -1.113 0.266313
Fjobteacher    1.3336     0.9059   1.472 0.141886
studytime      0.5579     0.1925   2.898 0.003991 **
failures      -1.1916     0.2330  -5.114 5.25e-07 ***
schoolsupyes  -2.1219     0.4480  -4.736 3.19e-06 ***
goout         -0.4998     0.1398  -3.576 0.000399 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.833 on 343 degrees of freedom
Multiple R-squared:  0.258,     Adjusted R-squared:  0.2299
F-statistic: 9.173 on 13 and 343 DF,  p-value: 2.848e-16
```

Figure 8. *Summary Statistics of the Reduced Model*

$Intercept$: The expected intercept is 12.3813, which indicates that the expected final grade is 12.3813 points when all the other predictors are zero.

Assuming all other variables are held constant, the following coefficients are significant at a 5% significance level:

- **sexM***:* Being male contributes to an increase in the final grade by 0.7890 points, suggesting that gender has a significant impact on final grades.
- **Mjobhealth**: Students whose mothers work in a health-related field tend to receive a higher final grade by 1.7888 points than those whose mothers stay at home.
- **Mjobservices**: Students whose mothers have a service-related occupation tend to receive a higher final grade by 1.2501 points than those whose mothers stay at home.
- **studytime**: More study time is associated with an increase in the student's final grade by 0.5579 points.
- **failures**: Each additional failure is associated with a lower final grade by 1.1916 points.
- **schoolsupyes**: Receiving school support tends to decrease the student's final grade by 2.1219 points.
- **goout**: Each additional outing is associated with a decrease in the student's final grade by 0.4998 points.

Assuming all other variables are held constant, **Fjobteacher** is *not* significant at a 5% significance level. As such, having a father with a teaching job does not have a statistically significant impact on the students' final grades.

3. Conclusion

3.1.1 Our Findings

Our analysis aimed to uncover external factors associated with students' academic performance, using data from the 'Student Performance' dataset. From our exploration, we discovered, through inferential statistical methods, that there are indeed external factors that may be associated with a student's academic performance. To answer our research question, there are 7 variables in particular that impact students' grades. The external factors that impacted a student's academic performance positively are:

1. Being male.
2. Having a mother in a health-related field.
3. Having a mother in a service-related occupation.
4. Spending more time studying.

While factors that impacted a student's performance negatively are:

1. Failing past classes.
2. Receiving more school support.
3. Going out more often.

While some factors are more obvious, such as the positive impact of more study time, we discovered many factors we found interesting, such as having a mother in specific occupations. Additionally, we found the result of one variable particularly surprising, which is that having school support negatively impacts student performance. This could be due to several reasons such as:

1. Too much reliance on school support: This may hinder a student's ability to build good study habits and independence outside of school support.
2. Inadequate support: The type of support may not suit the students' needs, or may not be of high quality.
3. Confounding factors: There may be underlying factors, such as school support being offered more frequently to individuals of low socioeconomic status.

Most of our findings align with the literature findings. For instance, consistent findings show that students who dedicate more time to studying generally achieve higher grades, whereas academic setbacks, like failing classes, lead to lower performance due to reduced confidence and cumulative knowledge gaps (Aljaffer et al., 2024). Additionally, parents' professional backgrounds can influence their children's attitude toward academics (Masud, 2019).

3.2 Limitations

- **Sample:** The results of our linear regression could be skewed if the sample does not accurately represent the target population. For instance, overrepresentation or underrepresentation of certain demographic groups in the sample can impact how well the results generalize to a broader population.
- **Assumption of independence:** We presume that observations are independent of one another. However, in reality, data points may be correlated due to clustering or repeated measurements. Violating this assumption can lead to inaccurate confidence intervals and p-values.
- **Linearity of relationship:** If the true relationship between the predictors and the response variable is nonlinear, the linear model may not properly capture the true underlying patterns.
- **Model fit:** While selected for its balance between simplicity and interpretability, Model 8 has relatively low R^2 (0.2547166) and adjusted R^2 (0.2375837) values, indicating that the model explains only a small proportion of the variance in academic performance. Additionally, a predictor, Fjobteacher was not statistically significant at the 5% significance level, which suggests that certain covariates may not meaningfully contribute to the model's explanatory power.

3.3 Potential Future Research Questions

Future research could explore how external factors influence academic performance over time. For instance, a longitudinal study could investigate whether the effects of extracurricular activities, parental involvement, or socioeconomic status vary as students progress through different educational stages. Additionally, examining datasets from diverse cultural or geographical contexts could also provide insights into how external factors like family expectations, access to technology, or education policies influence academic outcomes in different regions.

*Member who submitted the code file: **Suyeon Choi**

References

Aljaffer, M.A., Almadani, A.H., AlDughaither, A.S. et al. (2024). The impact of study habits and personal factors on the academic achievement performances of medical students. *BMC Med Educ 24,* 888 (2024). https://doi.org/10.1186/s12909-024-05889-y

Cortez, Paulo. (2014). *Student Performance* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T

Masud S, Mufarrih SH, Qureshi NQ, Khan F, Khan S and Khan MN. (2019). Academic performance in adolescent students: The role of parenting styles and socio-demographic factors – A cross sectional study from Peshawar, Pakistan. *Front. Psychol*. 10:2497. doi: 10.3389/fpsyg.2019.02497