Key Factors to a Student's Success

Sayalee Joshi
Instructor: Jim Herold

Abstract

This project aims to analyze Student Performance Data. The goal is to understand the patterns that can help in recognizing students at risk of under performing, thereby facilitating the development of targeted interventions and support strategies to improve their academic achievements. This analysis contributes to a deeper understanding of how different aspects of student life impact academic performance and aims to support educational stakeholders in enhancing student outcomes.

Contents

1 Introduction

The Student Performance Data from Kaggle has information on 2,392 high school students. This dataset encompasses a wide array of variables including demographics, study habits, parental involvement, and extracurricular activities, all of which play an important role in the academic performance of students. The primary objective of this project is to conduct an analysis to identify the key factors that contribute to academic success. By identifying these factors, the aim is to pinpoint students who may be struggling academically and provide them with targeted interventions and support as needed.

2 Data Summary

- **Dataset:**
  - **StudentID:** A unique identifier for each student, serving as the key of the table.
  - **Demographic Details:**
    i. **Age:** Reflects the age range of high school students, which is between 15 and 18 years.
    ii. **Gender:** A binary variable with 0 representing males and 1 representing females.
    iii. **Ethnicity:** A categorical variable coded as 0 for Caucasian, 1 for African-American, 2 for Asian, and 3 for Other.
  - **Study Habits:**
    i. **StudyTimeWeekly:** Weekly study time in hours, ranging from 0 to 20.
    ii. **Absences:** The number of absences during the school year, ranging from 0 to 30.
    iii. **Tutoring:** Indicator of tutoring status, where 0 denotes no tutoring and 1 indicates participation in tutoring.
  - **Parental Involvement:**
    i. **ParentalSupport:** An indicator of parental support, ranging from 0 (None) to 4 (Very high).
    ii. **ParentalEducation:** A categorical variable indicating the highest level of education attained by the parents:
      · 0: None
      · 1: High School
      · 2: Some College
      · 3: Bachelor's
      · 4: Higher
  - **Extracurricular Activities:** These are binary categorical variables (1 for Yes, 0 for No) indicating participation in:

i. **Extracurricular**
ii. **Sports**
iii. **Music**
iv. **Volunteering**

– **Academic Performance:** This section includes the response variables:

i. **GPA:** The grade point average, ranging from 2.0 to 4.0.
ii. **GradeClass:** Classification of students' grades based on GPA:
· 0: 'A' (GPA $\geq$ 3.5)
· 1: 'B' (3.0 $\leq$ GPA $<$ 3.5)
· 2: 'C' (2.5 $\leq$ GPA $<$ 3.0)
· 3: 'D' (2.0 $\leq$ GPA $<$ 2.5)
· 4: 'F' (GPA $<$ 2.0)

- **Primary Findings:**

– The analysis reveals that over 50% of students are failing their classes, highlighting a significant issue within the academic environment. The objective of this study is to conduct a comprehensive analysis to understand the factors contributing to this high failure rate.

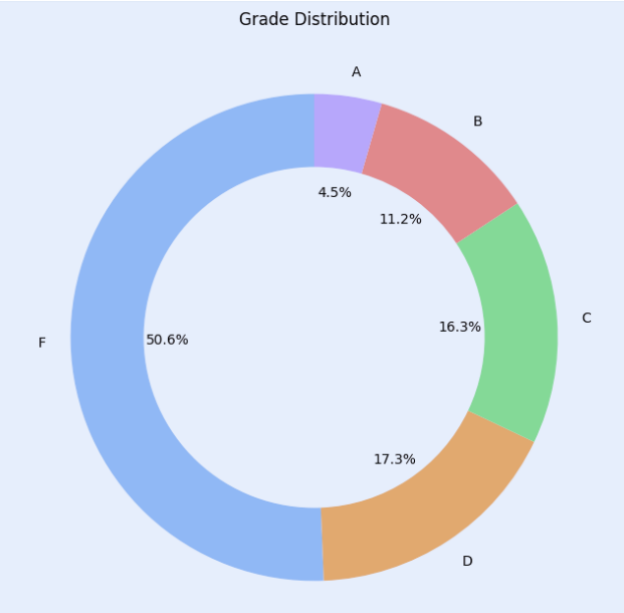


Figure 1: Distribution of grades

– From the distribution, we can see that the data is well balanced between male and female students.About a third of the students receive tutoring.

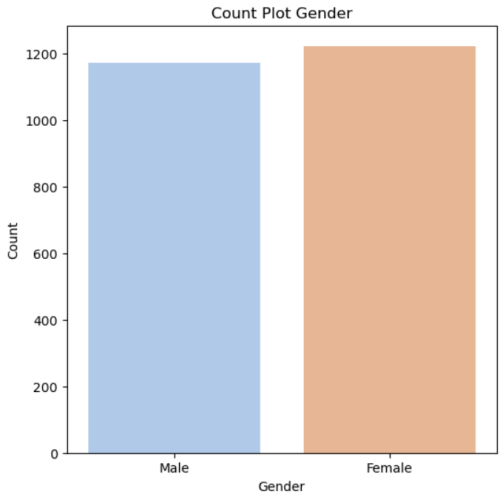


Figure 2: Gender Count

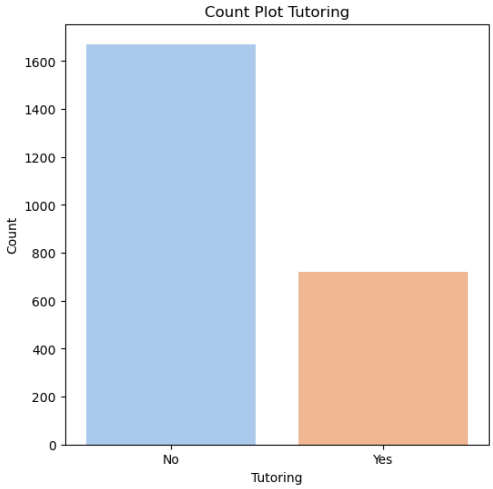


Figure 3: Tutoring

– Parental education and Parental Support somewhat follow a normal distribution.

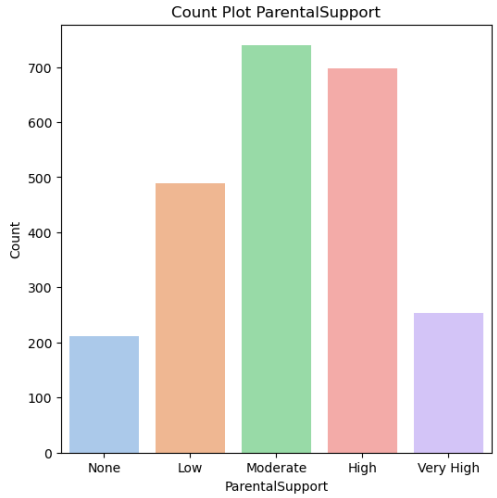


Figure 4: Parental Education Distribution

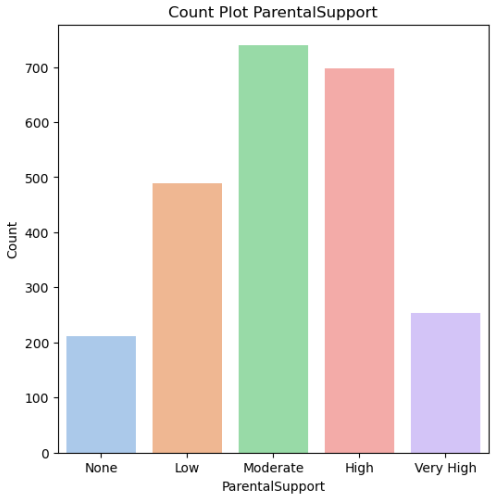


Figure 5: Parental Support Distribution

3 Methodology

- **Data Pre-processing:** The dataset is complete with no missing values, eliminating the need for data imputation or removal of incomplete rows. Key numerical variables include age, with an average of 16.47 years and a standard deviation of 1.12 years;

StudyTimeWeekly, averaging 9.77 hours; and Absences, with an average of 15 days. The StudentID column is an identifier and provides no analytical value. To enhance analysis, a new binary feature, Any_Extracurricular, will be added to indicate whether a student participates in any extracurricular activities, combining information from Extracurricular, Sports, Music, and Volunteering.

- **Model:**

  The analysis began with a multiple linear regression model that included all available covariates to predict GPA. Initial results indicated that covariates such as Gender, Ethnicity, and Parental Education had high p-values, suggesting they were not statistically significant predictors of GPA.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.6656	0.069	38.621	0.000	2.530	2.801
Age	-0.0095	0.004	-2.360	0.018	-0.017	-0.002
Gender	0.0113	0.009	1.247	0.213	-0.006	0.029
Ethnicity	0.0034	0.004	0.759	0.448	-0.005	0.012
ParentalEducation	0.0040	0.005	0.892	0.373	-0.005	0.013
StudyTimeWeekly	0.0288	0.001	35.962	0.000	0.027	0.030
Absences	-0.0994	0.001	-185.541	0.000	-0.100	-0.098
Tutoring	0.2531	0.010	25.615	0.000	0.234	0.272
ParentalSupport	0.1507	0.004	37.284	0.000	0.143	0.159
Any_Extracurricular	0.2083	0.010	20.723	0.000	0.189	0.228

Figure 6: Regression Model with all covariates

To further validate these findings, Lasso (L1) regularization was applied, which penalizes less significant covariates by shrinking their coefficients towards zero. This process resulted in the elimination of the coefficients for Age, Gender, Ethnicity, and Parental Education.

	Coefficient
Age	-0.000000
Gender	-0.000000
Ethnicity	0.000000
StudyTimeWeekly	0.144601
Absences	-0.814068
Tutoring	0.096583
ParentalSupport	0.141302
ParentalEducation	-0.000000
Any_Extracurricular	0.067927

Figure 7: Lasso Regularization

The final reduced model retained covariates including Weekly Study Time, Number of Absences, Tutoring, Parental Support, and Participation in Extracurricular Activities. This model demonstrated an R-squared value of 0.941, indicating that 94% of the

variation in GPA is explained by these covariates. The Adjusted R-squared value of 0.941 confirms the model's effectiveness even after accounting for the number of features.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.5239	0.016	153.826	0.000	2.492	2.556
StudyTimeWeekly	0.0289	0.001	35.965	0.000	0.027	0.030
Absences	-0.0993	0.001	-185.560	0.000	-0.100	-0.098
Tutoring	0.2527	0.010	25.574	0.000	0.233	0.272
ParentalSupport	0.1504	0.004	37.218	0.000	0.142	0.158
Any_Extracurricular	0.2088	0.010	20.774	0.000	0.189	0.229

Figure 8: Reduced Model

Although the AIC values for both the full and reduced models are similar, the BIC is lower for the reduced model, suggesting a more efficient balance between model fit and complexity.

- **Diagnostics:** The model is robust and does not violet any regression assumptions.
  - **R-squared:** 0.941

    94.1% of the variability in GPA is explained by the model, indicating a very good fit.
  - **Adjusted R-squared:** 0.941

    With 94.1%, the model explains the variability in GPA well, adjusted for the number of predictors.
  - **F-statistic:** 7676.0

    A high F-statistic indicates that the model is significant and at least one predictor is significantly related to GPA.The very low p-value (< 0.05) indicates that the model is statistically significant.
  - **Skewness:** -0.008

    A skewness close to 0 indicates that the residuals are symmetrically distributed.
  - **Kurtosis:** 3.157

    A kurtosis close to 3 indicates that the residuals have a normal distribution; the value here is close to 3, suggesting normality.
  - **VIF:**

    All the features have VIF values close to 1, indicating that there is no significant multicollinearity among these predictors. This means that the predictors are not highly correlated with each other and can be reliably used in the regression model.

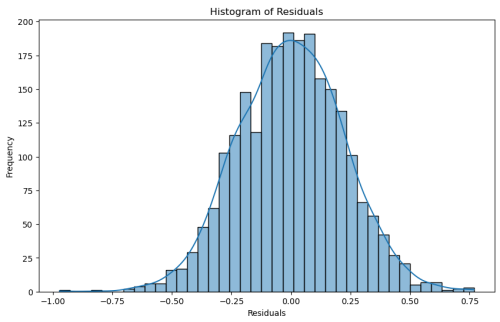


Figure 9: A roughly bell-shaped histogram indicates that residuals are normally distributed.

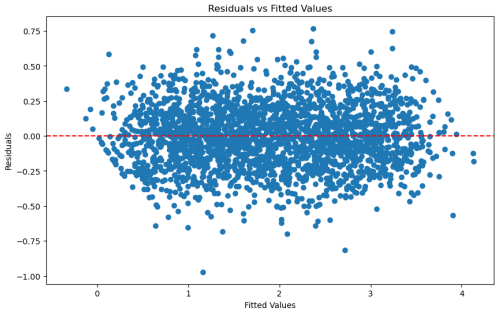


Figure 10: This plot does not show any pattern. A random scatter indicates that the model is appropriately specified.

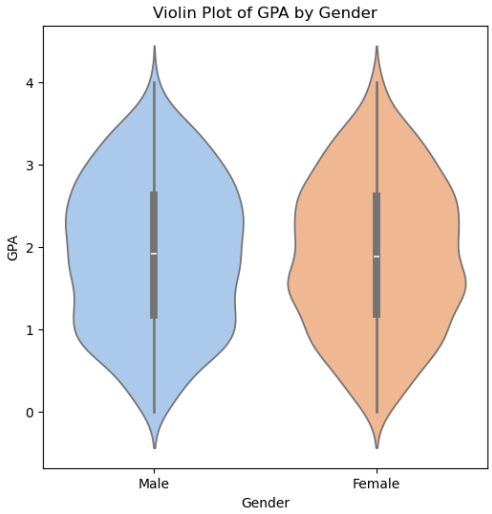
	Feature	VIF
0	const	13.109138
1	StudyTimeWeekly	1.002279
2	Absences	1.000345
3	Tutoring	1.001164
4	ParentalSupport	1.001658
5	Any_Extracurricular	1.000499

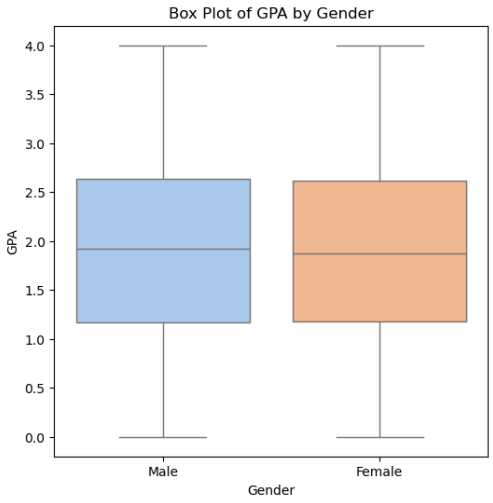
Figure 11: Variation Inflation Factor

4 Analysis and Findings

- **Claim:** Gender does not play an important role in predicting the GPA.

  Even though the mean GPA for male students is slightly higher than that of female students, it is not a very helpful covariate to predict the GPA. The p-value for the t-statistic for the covariate is greater than 0.05. Also when Lasso regularization is applied, the coefficient for Gender shrinks to zero.





- **Claim:** Parents' education is not important but parental support is pivotal.

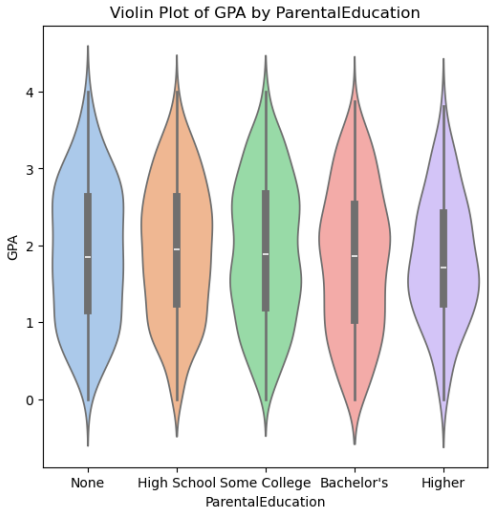


Figure 12: Parental Education Distribution

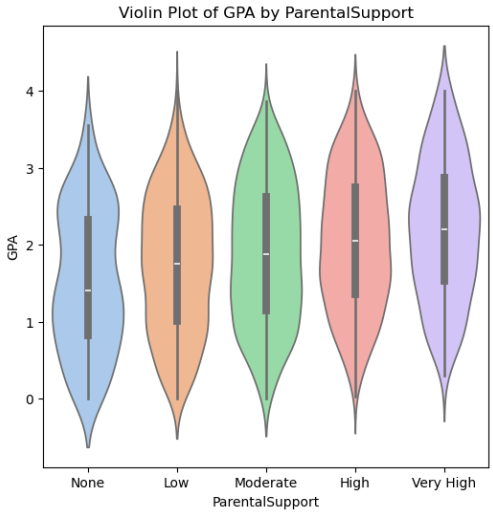


Figure 13: Parental Support Distribution

As we can see the mean GPA across parental support shows an increasing trend. Whereas Parental education doesn't show any trend. Also while performing Lasso regularization, ParentalEducation coefficient shrinks to zero but ParentalSupport is not penalized.

- **Claim:** Absence from classes hinder a students performance.

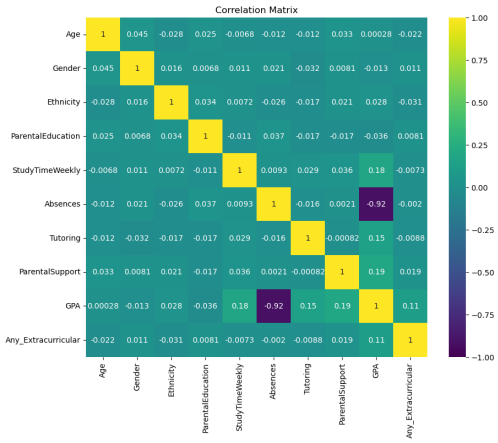


Figure 14: Absence and GPA show a high negative correlation.

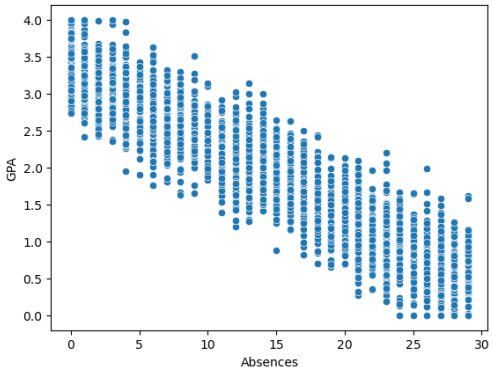
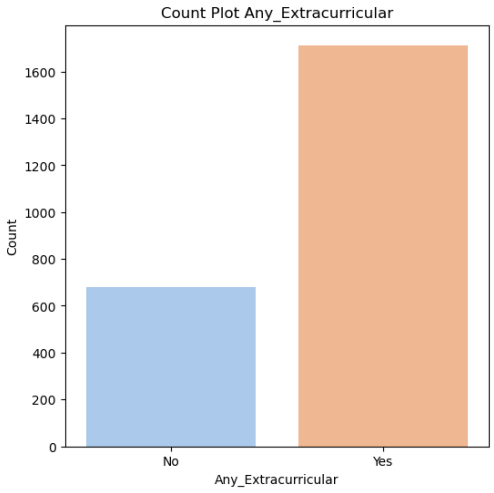
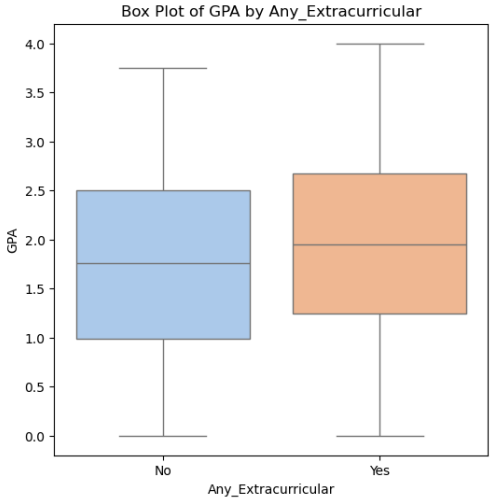


Figure 15: As the number of absences increase, GPA decreases.

- **Claim:** Participation in extracurricular activities have a positive impact on the GPA.

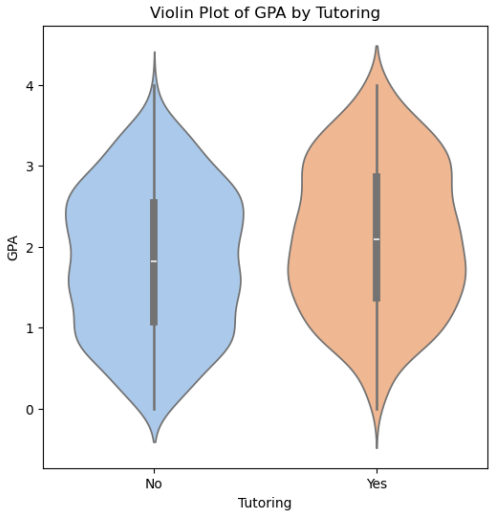
  The t-test results reveal a significant difference in GPA between students who participated in extracurricular activities and those who did not, with a p-value of 2.17e-7. This implies that participation in extracurricular activities is positively associated with higher academic performance.

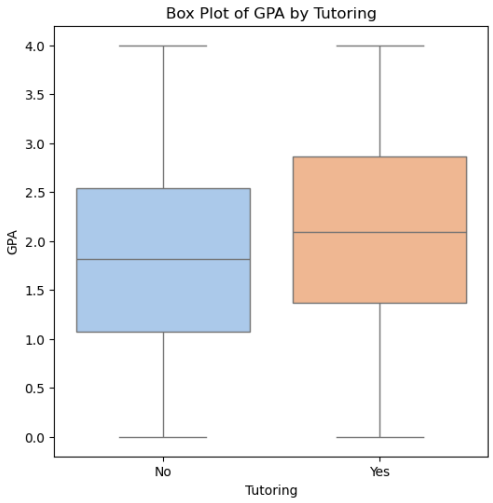




- **Claim:** Tutoring helps a student improve academically.

  The t-test results indicate a highly significant difference in GPA between students who received tutoring and those who did not, with a p-value of 1.2e-12 . This suggests that tutoring has a substantial positive impact on academic performance.





5 Future Scope

The data source is confidential, and hence getting further information on this dataset is not currently possible. But this type of analysis can be performed on similar datasets with more information about the socioeconomic status of students, some additional qualitative information on parental support etc. With more information on the data source, we can decide if this analysis can be generalised or is school specific.

6 Recommendations and Next Steps

- Guide parents on being supportive thus fostering open communication between children, parents, and teachers.
- Provide counseling for students with frequent absences to understand and solve their issues.
- Create incentives for students to improve school attendance. Offer wide range of classes, workshops, competitions etc.
- Encourage students to join extracurricular activities.
- Offer tutoring to struggling students from peers, volunteer teachers or parents.
- Repeating this analysis after six months might help to understand if the changes implemented are effective or not.

7 Conclusion

In conclusion, this analysis highlights the critical role of tutoring and extracurricular activities in enhancing students' academic performance, as evidenced by the statistically sig-

nificant differences in GPA between those who engage in these activities and those who do not. To further support student success, it is essential to guide parents in fostering open communication with their children and teachers, provide counseling for students with frequent absences, and create incentives to improve school attendance. By offering a variety of classes, workshops, and competitions, schools can create a more engaging learning environment. Implementing these changes and repeating the analysis after six months will help evaluate their effectiveness and ensure continuous improvement in student outcomes.

8 Appendices

- Link to the dataset:
  - `https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data`
- Link to the data analysis code:
  - `https://github.com/sayaleej/Student-Performance-Data-DATA-294P-.git`