Kayley Reith
Patrick Campbell
Jake Konrad
MATH 4780

Final Project Proposal: High School Student Performance & Demographics

The goal of our project is aimed to determine if there are specific factors that contribute more towards the overall success of a high school student's performance. Our dataset carries large variability with 395 students ranging from ages of 15 to 21 years old. This is highly relevant and significant as in the United States alone from 2021 there will be 16.9 million high school students (High School of America 2022). The percentage of students who don't graduate high school is 6% accounting for 414,000 students never completing high school (Craft 2022). Thus it is critical to explore what factors play an influential role in determining high school student success and performances.

Our methodology will explore the dataset through data visualization to discern patterns and relationships between variables like internet access, family's education, and even level of socialness, aiding in the informed selection of regression models.This will also allow us to check if there are any outstanding outliers in students performance or highlight any minority demographics. We can start with a simple linear regression by predicting final grades based on family education and study time. Then we can create a more complex multi-linear regression model testing for individual and joint significance to see student final grades and next steps in higher education dependent on the predictors, social time, age, and alcohol consumption. After the regression diagnostics phase we will encompass model adequacy checks, thorough residual diagnostics for randomness and homoscedasticity, and assess if there are any leverage points or factors of influence. To address potential violations of assumptions and collinearity issues, we will employ suitable transformations and variable selection strategies dependent on the skewness and non-normality of the data. The final stage will involve model evaluation, demonstrating adherence to assumptions through inference methods, and providing transparent discussions on any residual limitations and suggestions for further analysis improvement. We are hoping to employ hypothesis testing, ANOVA tests, and the multiple linear regression tests. This structured approach ensures a comprehensive understanding of the factors influencing high school student performances and contributes valuable insights to educational policies and practices.

Citations:

Craft, S. (2022, February 11). High School Statistics. Think Impact. https://www.thinkimpact.com/high-school-statistics/#:~:text=The%20gender%20parity%20for%20US,15%25%20college%20drop%20out%20rate

High School of America. (2022, October 7). High School Statistics in the United States. High School of America.https://www.highschoolofamerica.com/united-states-high-school-statistics/#:~:text=In%202021%2C%20approximately%2016.9%20million,enrolled%20in%20public%20high%20schools

Data: High School Student Performance & Demographics. Kaggle. 2023. https://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics