Study Behaviors and Academic Achievement: An Exploratory Analysis Using Multi-Subject Student Assessment Data

Yumo Bai, Suhan Liu, Xu Huang, Beichen Wan

December 2025

Abstract

Study habits, absenteeism, and extracurricular participation are often seen as key factors influencing academic success, yet their combined effects across different subjects remain unclear. In this report, we examine how weekly self-study hours, absence days, and extracurricular involvement relate to the academic performance of 2,000 high school students using a publicly available dataset from Kaggle. After cleaning the data and creating an overall performance measure based on seven subject scores, we conduct exploratory analyses with quartile-based boxplots and scatterplots. Next, we estimate an extended multiple regression model that includes demographic controls, a cutoff-based measure of excessive absences, and an interaction term to assess whether study effort mitigates the impact of absenteeism. The findings reveal a clear and statistically significant positive relationship between self-study hours and average scores, while neither absence days nor extracurricular participation show significant partial effects, even when using the cutoff transformation or interaction term. However, a joint F-test indicates that the behavioral variables collectively help explain achievement. We conclude by discussing the dataset's limitations, such as measurement error, missing covariates, and the absence of subject-specific study-hour data. We also suggest future research directions that incorporate multiple hypothesis testing procedures and nonlinear model components.

1 Introduction

Academic performance is influenced by a multidimensional set of behavioral, cognitive, and motivational factors. Among these, *study habits*, *attendance patterns*, and *extracurricular participation* are key behaviors that students can directly control. Understanding how these behaviors collectively shape academic outcomes is critically important for developing evidence-based study strategies and institutional interventions.

The dataset considered in this project consists of 2,000 students, each with detailed demographic information, behavioral metrics (weekly self-study hours, number of absence days, extracurricular participation), and exam scores across seven subjects: mathematics, history, physics, chemistry, biology, English, and geography. This dataset enables a broad, multi-subject investigation into how different domains of academic behavior may influence performance outcomes.

A growing body of quantitative work has examined the link between study time and academic achievement. In a meta-analysis of 49 studies comprising 77 independent samples and 19,219 students, Huang (2015) estimated the average correlation between self-reported study time and academic achievement to be approximately $r \approx 0.12$, with somewhat larger effects in mathematics and language-intensive subjects. The association is positive and statistically reliable but small, implying that variation in hours studied explains only a modest share of the variation in grades. This finding motivates a more comprehensive modeling strategy that considers additional behavioral predictors beyond study time alone. In particular, our study focuses on self-study hours, absence days, and extracurricular participation in a unified framework and will examine their joint contribution to academic performance across multiple subjects in future work.

The goal of this report is therefore not to present final results, but to clarify the research question, describe the data and its limitations, explore basic patterns, and propose an analysis plan.

2 Research Question and Hypotheses

2.1 Research Question

We focus on the following overarching question:

To what extent do self-study hours, absence days, and extracurricular participation predict academic performance across subjects?

2.2 Substantive Hypotheses

Based on prior empirical findings, we consider three substantive hypotheses. Hypothesis H1 states that weekly self-study hours positively impact academic performance. Hypothesis H2 states that absence days negatively impact academic performance. Hypothesis H3 posits that extracurricular participation has a modest positive impact on academic performance. These hypotheses are qualitative in nature and are intended to guide the subsequent choice of statistical models rather than to make immediate formal claims.

2.3 Null Hypotheses

For the planned regression analysis, these informal claims translate into formal null hypotheses about the regression coefficients. Under $H_{0,1}$, and holding other variables fixed, self-study hours are not associated with the average score; equivalently, the regression coefficient on self-study hours equals zero. Under $H_{0,2}$, and again holding other variables fixed, absence days are not associated with the average score. Under $H_{0,3}$, extracurricular participation is not associated with the average score once other predictors are controlled for. These null hypotheses will be formally tested in a subsequent stage of the project.

3 Data Source, Wrangling, and Limitations

3.1 Data Source

The data come from the *Student scores* dataset on Kaggle, compiled by Medhat (2023). The dataset is publicly available and contains information on 2,000 high school students. For each student, the dataset includes demographic variables such as age, gender, and career aspiration; behavioral variables including weekly self-study hours, number of absence days, and an indicator of participation in extracurricular activities; and exam scores in seven subjects: mathematics, history, physics, chemistry, biology, English, and geography.

For the purposes of this project we construct an overall performance measure, `average_score`, defined as the mean of the seven subject scores. This aggregate measure serves as our primary outcome and allows us to summarize multi-subject achievement in a single continuous variable.

3.2 Data Cleaning

Several data wrangling steps were carried out to prepare the dataset for exploration and future modeling. First, all variable names were converted to lower case with underscores to simplify coding and ensure consistent reference across scripts. Second, observations with missing values for any of the key variables (study hours, absences, extracurricular indicator, or subject scores) were removed, and basic range checks were applied to ensure that exam scores and counts of absences fell within plausible bounds. Third, the extracurricular participation variable was converted to numeric format, with 0 indicating no reported participation and 1 indicating at least one extracurricular activity. Finally, for continuous variables such as `weekly_self_study_hours` and `average_score`, values above four standard deviations from the mean were winsorized to the four–standard-deviation cutoff. This approach reduces the influence of extreme points that may reflect data entry errors or rare cases not representative of typical students, while retaining those observations in the analysis.

3.3 Concerns About the Data

Although the dataset is rich and convenient, several concerns may affect its ability to fully answer our research question and should be kept in mind when interpreting any patterns. First, as a Kaggle dataset, the sample may not be nationally representative. The students could come from a specific region or school system, which would limit generalizability of the findings. Second, weekly self-study hours and possibly absence counts are self-reported, so measurement error is likely. Such error could attenuate estimated relationships, particularly for study time. Third, important predictors such as intrinsic motivation, prior achievement, parental education, and school-level characteristics are not included in the data. These omitted variables may bias simple regression estimates if they are correlated with both the behaviors of interest and academic performance. Fourth, the data are cross-sectional, with all variables measured at a single time point. This design prevents strong causal conclusions about whether behavior changes lead to achievement changes. Finally, averaging across seven subjects implicitly weights each subject equally and may mask subject-specific patterns. For

example, behaviors might matter more for mathematics than for history, but this distinction is not captured by the `average_score` measure.

4 Exploratory Data Analysis

Before fitting regression models, we conducted exploratory data analysis to inspect how average achievement varies across levels of absence days and self-study hours. For both variables we divided students into quartiles so that each group contains roughly 25% of the sample, and then plotted boxplots of `average_score` within each quartile. In all boxplots, the central line denotes the median, the box spans the interquartile range (IQR), and the whiskers extend to $1.5 \times$ IQR with points beyond this range shown as outliers.

4.1 Average Score by Absence Days Quartile

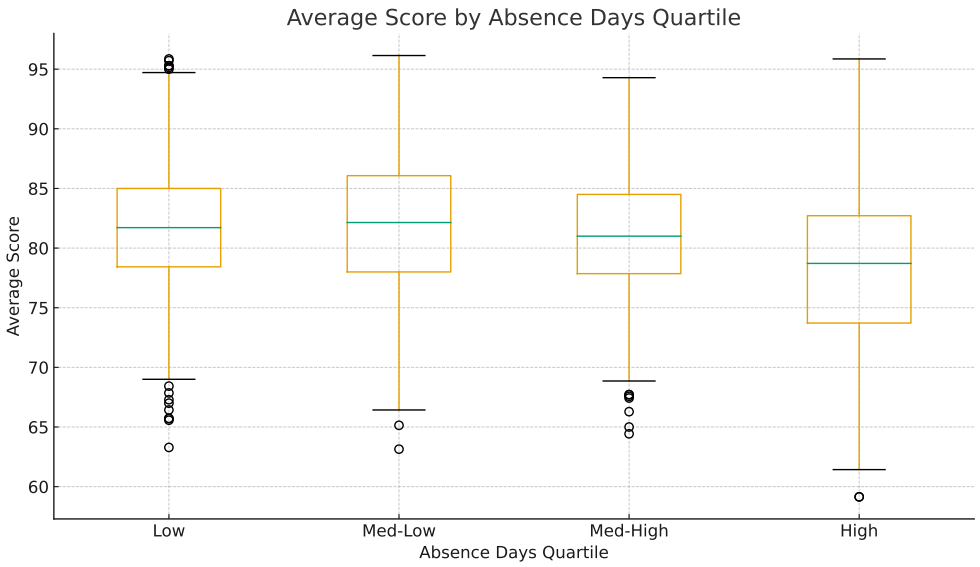

Figure 1: Average score by absence days quartile.

To construct Figure 1, we computed quartiles of `absence_days` and assigned students to four ordered categories: Low, Med–Low, Med–High, and High absence. The resulting boxplots show a gradual downward shift in median average score as absence increases. Students in the High absence group have both a lower median and a thicker lower tail, indicating more low-performing students in that category. At the same time, the boxes and whiskers overlap across quartiles, and some

high-achieving students appear even in the higher-absence groups. This pattern suggests a clear but not deterministic negative relationship between absenteeism and performance.

4.2 Average Score by Study Hours Quartile

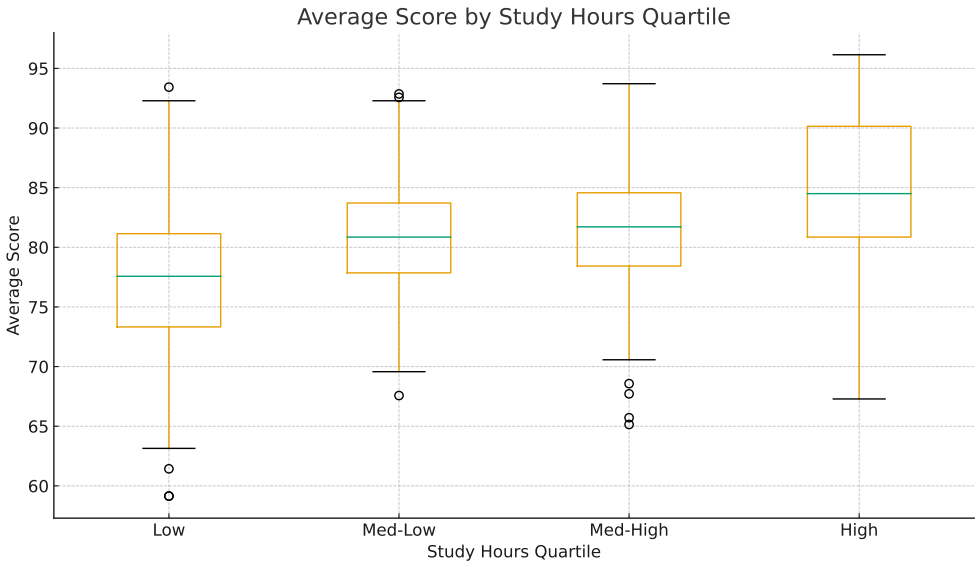

Figure 2: Average score by self-study hours quartile.

Figure 2 is based on quartiles of `weekly_self_study_hours`, again labeled Low, Med–Low, Med–High, and High. Here the medians and IQRs shift upward as study hours increase: students in the High study group tend to have higher average scores and fewer very low outliers. Nevertheless, there is substantial overlap across quartiles, with some low performers in the High study group and some high performers in the Low study group. This visual evidence is consistent with meta-analytic results that study time has a positive but modest association with achievement (Huang, 2015). These patterns strengthen the case for fitting multivariable models that can quantify the relationships while controlling for other behaviors and demographics.

4.3 More General Explanation on the Discretizations of EDA

In the previous EDA section, we temporarily converted the continuous variables (weekly self-study hours and absence days) into quartile-based categories in order to make the patterns in the data easier to see and interpret. Working with four ordered groups such as “Low,” “Med–Low,” “Med–

High," and "High" allows us to summarize how average scores change across the distribution of each predictor, instead of relying on a cloud of individual points that may be noisy and affected by outliers. Boxplots by quartile provide a clear visual comparison of median performance, the spread of scores, and the presence of extreme values in each group, which is useful for descriptive purposes and for communication to non-technical readers. This grouping also helps reveal potential non-linear or threshold effects—for example, whether performance only drops markedly once absences move into the highest range—without forcing us to assume any particular functional form at this stage. This discretization is only an exploratory visualization choice; in the planned regression models, these variables will be kept in their original continuous form so that we retain all available information for formal inference.

4.4 Scatterplots with Continuous Predictors

In addition to the quartile-based boxplots, we also examined scatterplots that treat weekly self-study hours and absence days as continuous predictors. This provides a complementary view of the data by showing the full cloud of individual observations together with fitted linear trends, rather than only summaries within four groups.

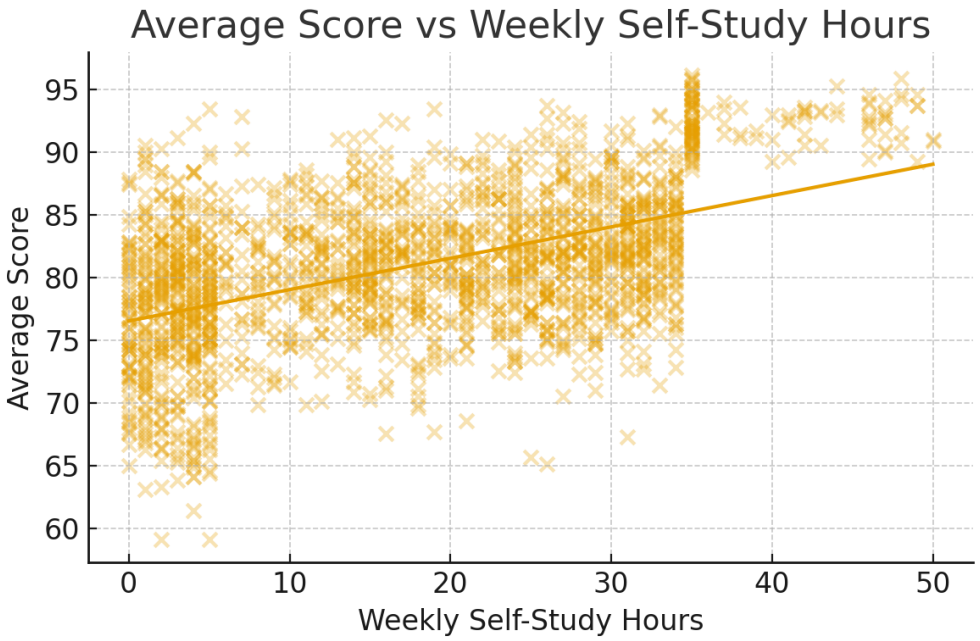

Figure 3: Average score vs. weekly self-study hours (continuous).

Figure 3 plots each student's average score against their weekly self-study hours, along with a fitted regression line. The individual points are fairly dispersed, indicating substantial heterogeneity in achievement at any given level of study time. Nonetheless, the fitted line slopes upward, echoing the boxplot pattern in Figure 2: on average, students who report more self-study hours tend to achieve higher scores. The scatterplot also makes it clear that increases in study time are associated with only gradual gains in performance, which is consistent with the interpretation of a positive but modest effect.

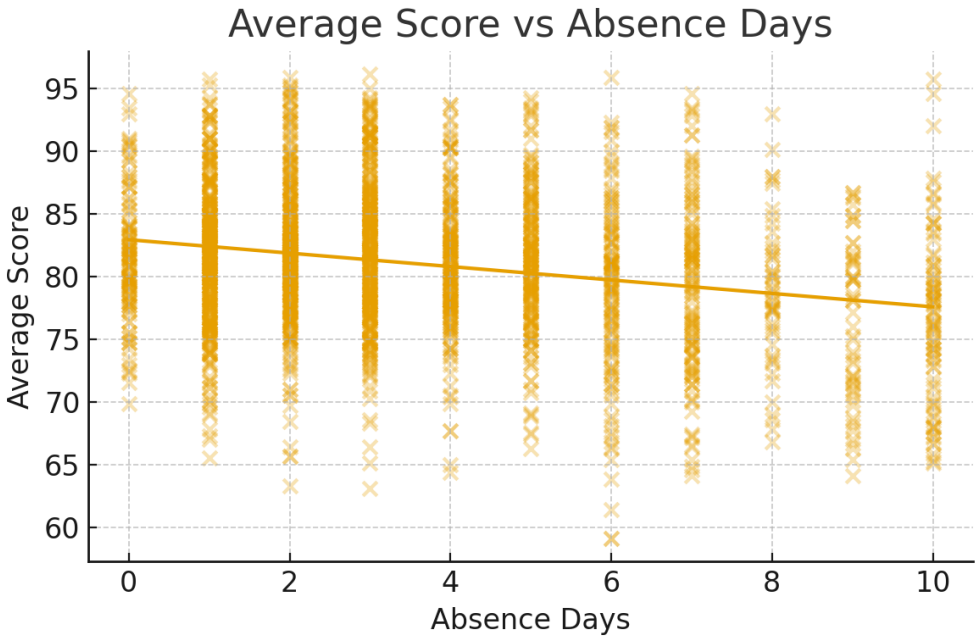

Figure 4: Average score vs. absence days (continuous).

Figure 4 displays average score against absence days, again with an overlaid linear trend. Here the fitted line slopes downward, reinforcing the message from Figure 1 that greater absenteeism is associated with lower performance on average. As with the study hours plot, the cloud of points shows considerable variability, including some students with high scores despite relatively many absences and some low scorers even among students with very few absences. Together with the quartile-based boxplots, these scatterplots suggest that absence and self-study are meaningfully related to achievement, but that they do not fully determine outcomes, which motivates the use of multivariable regression models that jointly account for multiple behaviors and student characteristics.

Table 1: Summary Statistics of Student Performance and Habits

	weekly self study hours	average score	absence days	extracurricular activities
mean	17.76	80.98	3.67	0.20
std	12.13	6.04	2.63	0.40
min	0.00	59.14	0.00	0.00
25%	5.00	77.29	2.00	0.00
50%	18.00	81.00	3.00	0.00
75%	28.00	84.71	5.00	0.00
max	50.00	96.14	10.00	1.00

4.5 Summary Statistics

Table 1 presents the descriptive statistics for student performance and behavioral habits, specifically focusing on weekly self-study hours, average academic scores, absence days, and extracurricular participation. The data reveals substantial variability in study habits, with students reporting an average of 17.76 hours of self-study per week ($\sigma \approx 12.13$) and a range extending from 0 to 50 hours.

5 Regression Model, Derived Variables, and Hypothesis Tests

5.1 Model Specification

To examine the determinants of academic performance, we estimate the following ordinary least squares (OLS) regression model:

$$\begin{aligned} \text{AverageScore}_i = \beta_0 &+ \beta_1 \,\text{StudyHours}_i + \beta_2 \,\text{AbsenceDays}_i + \beta_3 \,\text{AbsenceOverCutoff}_i \\ &+ \beta_4 \,(\text{AbsenceOverCutoff}_i \times \text{StudyHours}_i) \\ &+ \beta_5 \,\text{Extracurricular}_i + \beta_6 \,\text{PartTimeJob}_i \\ &+ \gamma_{\text{gender}(i)} + \delta_{\text{aspiration}(i)} + \varepsilon_i. \end{aligned} \tag{1}$$

Each term corresponds directly to a column in the dataset or a constructed variable described below.

5.2 Derived Variables

Absence Over Cutoff. Exploratory data analysis suggested that absenteeism does not significantly affect academic performance until it becomes excessive. To model this nonlinear behavior, we define:

$$\text{AbsenceOverCutoff}_i = \max(0,\ \text{AbsenceDays}_i - c),$$

where $c = 3$ in this analysis. This variable equals zero for students with three or fewer absences, and increases linearly only for absences beyond the cutoff.

Interaction Term. To test whether self-study effort compensates for excessive absenteeism, we include an interaction term:

$$\text{Interaction}_i = \text{AbsenceOverCutoff}_i \times \text{StudyHours}_i.$$

A significant coefficient β_4 would indicate that the marginal effect of absenteeism depends on a student's study behavior.

Extracurricular Participation and Job Status. Both variables originate as binary indicators in the dataset and are coded as:

$$1 = \text{Yes}, \qquad 0 = \text{No}.$$

This allows them to enter linearly into the regression as dummy variables.

5.3 Individual Hypothesis Tests

Table 2 displays the t-tests for the four coefficients of primary interest: study hours, absence days, extracurricular participation, and the interaction term.

Table 2: Individual t-Tests for Key Predictors

Hypothesis	**Coef.**	**Std. Err.**	**t**	**p-value**	**95% CI**
$H_{0,1}$: Study hours= 0	0.1387	0.017	8.126	0.000	[0.105, 0.172]
$H_{0,2}$: Absence days= 0	0.0983	0.128	0.767	0.443	[–0.153, 0.350]
$H_{0,3}$: Extracurricular= 0	–0.0955	0.265	–0.360	0.719	[–0.616, 0.425]
$H_{0,4}$: AbsenceOverCutoff ×Study= 0	–0.0051	0.005	–0.967	0.334	[–0.015, 0.005]

Study hours is the only predictor with a statistically significant individual effect.

5.4 Joint Hypothesis Test

To evaluate whether the behavioral predictors collectively contribute to the model, we conduct the following joint test:

$$H_0 : \beta_{\text{study}} = \beta_{\text{absence}} = \beta_{\text{extracurricular}} = 0.$$

Table 3 reports the results.

Table 3: Joint F-Test of Three Behavioral Predictors

Statistic	**Value**
F-statistic	18.59
p-value	6.80×10^{-12}
Numerator df	3
Denominator df	1980

Although absence days and extracurricular participation do not show significant individual effects, the extremely small joint p-value indicates that the three predictors together contribute significantly to the explanation of academic performance.

5.5 Interpretation

Overall, the regression results indicate that weekly self-study hours exert a clear and statistically significant positive influence on students' academic performance. In contrast, absenteeism—whether modeled as total absence days or as absences beyond a specified cutoff—does not demonstrate a strong or consistent effect within this sample. Participation in extracurricular activities similarly shows no measurable association with academic outcomes once other factors are controlled for. Finally, the interaction term between excessive absenteeism and self-study hours provides no evidence that increased study effort reduces or offsets the potential negative effects of missing classes.

These findings imply that students' independent study habits are substantially more important for academic outcomes than moderate differences in absenteeism or participation in activities outside the classroom.

6 Discussion

In this section, we will try to answer the following questions:

1. Were our answers conclusive? If not, what could have been done to make them conclusive?
2. Were our assumptions reasonable?
3. What limitations did our data and/or analysis have?

Finally, we will discuss some future extensions of our work.

6.1 Conclusiveness

Generally, the results are not fully conclusive. Our exploratory data analysis and regression results show a clear and statistically significant positive association between weekly self-study hours and the average score. However, neither absence days nor extracurricular participation demonstrated statistically significant partial effects after including demographic controls. We also tried to adjust the "absence days" variable by applying a cutoff as the EDA suggested, but this adjustment did not produce a strong or statistically meaningful relationship either. These findings limit the strength of any conclusions about the independent effects of absences and extracurricular participation on academic performance.

Nevertheless, these limitations are partially offset by the joint F-test of hypothesis $H_{0,\text{all}}$, which tests whether all three behavioral coefficients are simultaneously equal to zero. This joint null hypothesis is strongly rejected, indicating that although absences and extracurricular participation are not individually significant, the behavioral variables collectively provide meaningful explanatory power to the model.

To make the results more conclusive, we could incorporate richer data, particularly key unobserved variables such as prior academic achievement, motivation, parental background, and school-level traits. Additionally, obtaining more accurate measurements of behavioral variables—rather than depending on self-reported study hours and absence counts—would help minimize measurement errors and produce more dependable estimates.

6.2 Assumption Reasonability

The assumptions were **reasonable** for a preliminary analysis. The linearity assumption is roughly supported by the scatterplots in EDA, which is also reflected in the final results. The additivity of the effects is also supported by the insignificant interaction between extreme absences and study hours.

Therefore, the assumptions are acceptable for the scope of an exploratory project, but additional diagnostics and richer data would be necessary for more rigorous inference.

6.3 Limitations of the Data and Analysis

One of the main limitations of the data is that study hours for each subject could not be distinguished. Because of this limitation, we couldn't perform a more detailed analysis on each subject; instead, we used the average score. Also, the Kaggle dataset may reflect a specific school system rather than a national population, limiting generalizability.

For the analysis, the weak individual effects of key predictors remain a major issue. Neither absence days nor extracurricular participation shows statistically significant partial effects, making it hard to draw firm conclusions about their independent contributions to academic performance. Even after adding an "absence over cutoff" variable and interaction terms to account for possible nonlinearities, absenteeism still did not demonstrate a strong or consistent link with the outcome.

6.4 Future Work

A potential direction for future work is to implement multiple hypothesis testing methods to more thoroughly evaluate the combined significance of behavioral predictors and lower the chance of false positives. Additionally, including higher-order terms—such as quadratic functions of study hours or absence days—may better capture nonlinear relationships that the current linear model overlooks, potentially enhancing model fit and uncovering more subtle behavioral effects.

Acknowledgments

We thank all group members for their contributions to this project. Bryce Bai identified the Kaggle dataset that best matched our research interests and took the lead in organizing the report. Xu Huang prepared the README files and worked with Bryce on writing and organizing the report documentation. Suhan Liu created the presentation slides and produced the exploratory boxplots, with assistance from ChatGPT in clarifying the x-axis definitions and labeling. Beichen Wan coordinated the project workflow, organized the shared files, and helped guide the group toward refining and focusing the research question.

References

Huang, C. (2015). Meta-analysis of the relation between study time and academic achievement. In *Proceedings of the European Conference on Education*. The International Academic Forum.

Medhat, M. (2023). Student scores. `https://www.kaggle.com/datasets/markmedhat/student-scores`. Accessed 17 November 2025.