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

Yumo Bai, Suhan Liu, Xu Huang, Beichen Wan

November 2025

Abstract

Study habits, absenteeism, and extracurricular involvement are widely discussed as influential factors in academic achievement, yet their combined effects across multiple subjects are not well understood. In this preliminary report we (i) formalize a research question about how self-study hours, absence days, and extracurricular participation relate to academic performance; (ii) describe a publicly available Kaggle dataset of 5,000 students that will be used to address this question; (iii) summarize existing literature, including meta-analytic evidence that the correlation between study time and achievement is positive but small; (iv) conduct exploratory data analysis via boxplots of average score across quartiles of absence days and study hours; and (v) propose a multiple regression analysis plan for future work. No formal inferential results are reported here; instead, the focus is on clarifying the question, documenting data wrangling decisions, and outlining the next steps in the analysis.

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 5,000 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

> **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 propose:

- **H1 :** Weekly self-study hours positively impact academic performance.
- **H2 :** Absence days negatively impact academic performance.
- **H3 :** Extracurricular participation has a modest positive impact on academic performance.

2.3 Null Hypotheses

For the planned regression analysis, these informal claims translate into the following null hypotheses:

- $H_{0,1}$**:** Holding other variables fixed, self-study hours are not associated with average score (the regression coefficient on self-study hours equals zero).
- $H_{0,2}$**:** Holding other variables fixed, absence days are not associated with average score.
- $H_{0,3}$**:** Holding other variables fixed, extracurricular participation is not associated with average score.

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 5,000 high school students, including:

- Demographic variables (e.g., age, gender, career aspiration).
- Behavioral variables: weekly self-study hours, number of absence days, and an indicator of participation in extracurricular activities.
- 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.

3.2 Data Cleaning and Wrangling

Several data wrangling steps were carried out to prepare the dataset for exploration and future modeling:

1. **Standardization of column names:** All variable names were converted to lower case with underscores to simplify coding.
2. **Handling missing and invalid values:** Observations with missing values for any of the key variables (study hours, absences, extracurricular indicator, or subject scores) were removed. Basic range checks were applied to ensure that exam scores and counts of absences were within plausible bounds.

3. **Binary coding:** The extracurricular participation variable was converted to numeric format (0 = no participation, 1 = at least one activity).

4. **Outlier treatment:** For continuous variables such as `weekly_self_study_hours` and `average_score`, values above 4 standard deviations from the mean were winsorized to the 4–SD cutoff. This reduces the influence of extreme points that may reflect data entry errors or rare cases not representative of typical students.

3.3 Concerns About the Data

While the dataset is rich and convenient, several concerns may affect its ability to fully answer our research question:

- **Representativeness:** As a Kaggle dataset, the sample may not be nationally representative. The students could come from a specific region or school system, limiting generalizability.

- **Self-reporting error:** Weekly self-study hours and possibly absence counts are self-reported, so measurement error is likely. This could attenuate estimated relationships, particularly for study time.

- **Unobserved confounding:** Important predictors such as intrinsic motivation, prior achievement, parental education, and school-level characteristics are not included. These omitted variables may bias simple regression estimates.

- **Cross-sectional structure:** The data are cross-sectional, with all variables measured at one time point. This prevents strong causal conclusions about whether behavior changes lead to achievement changes.

- **Construct validity of `average_score`:** Averaging across seven subjects implicitly weights each subject equally and may mask subject-specific patterns (e.g., behaviors might matter more for math than for history).

These limitations will be kept in mind when interpreting any patterns and when designing the planned analysis.

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 $1.5 \times \mathrm{IQR}$ with points beyond this range shown as outliers.

4.1 Average Score by Absence Days Quartile

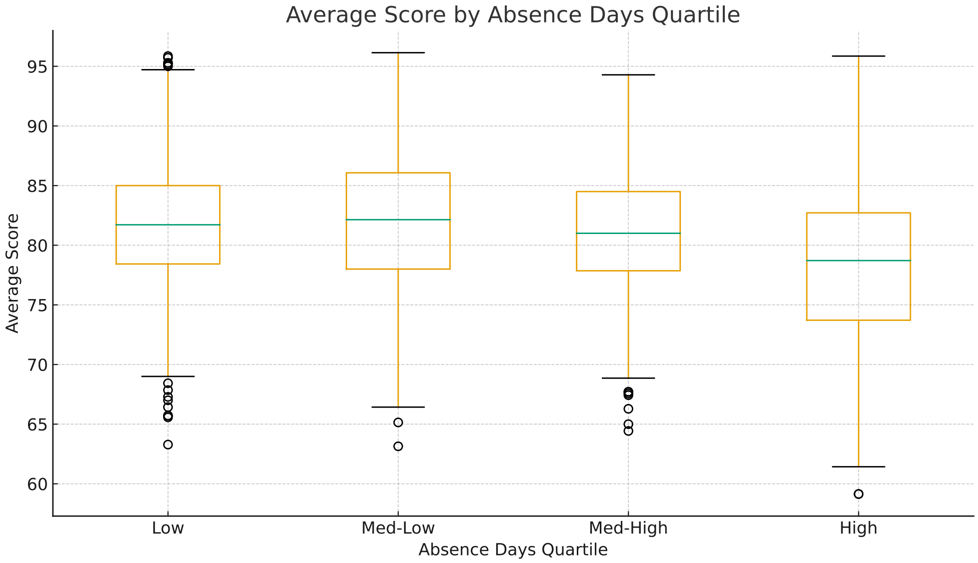

Figure 1: Average score by absence days quartile.

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 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. However, 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

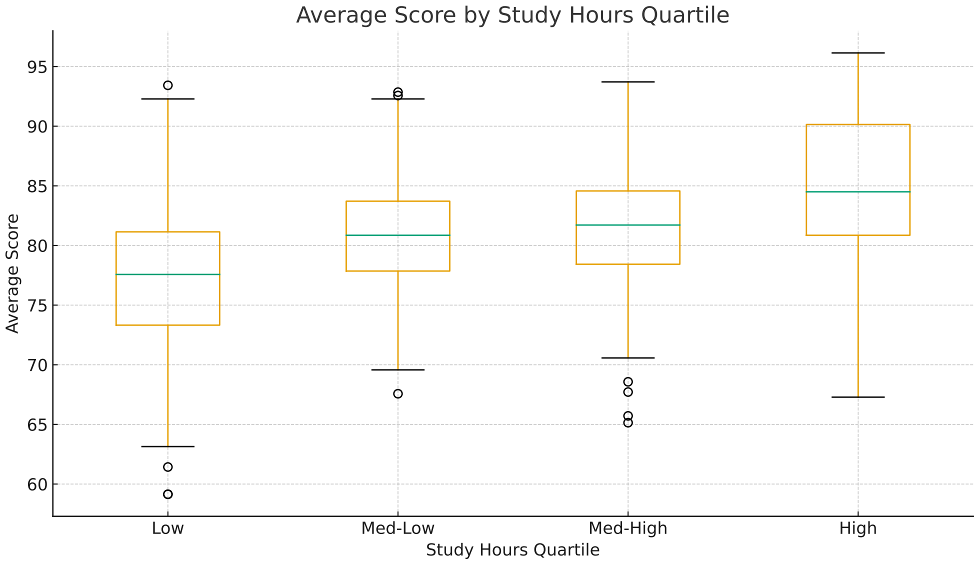

Figure 2: Average score by self-study hours quartile.

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. At the same time, 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.

5 Planned Statistical Analysis

In the next stage of the project, we plan to use multiple linear regression to quantify the associations suggested by the exploratory plots.

5.1 Baseline Model

Let average_score_i denote the mean exam score of student i across the seven subjects. Our baseline model will be:

$$\text{average\_score}_i = \beta_0 + \beta_1(\text{self\_study\_hours}_i) + \beta_2(\text{absence\_days}_i) + \beta_3(\text{extracurricular}_i) + \epsilon_i,$$

where ϵ_i is an error term with mean zero and constant variance. The coefficients β_1, β_2, and β_3 correspond directly to the null hypotheses specified in Section 3.

5.2 Model with Demographic Controls

To partially address confounding by demographics, an expanded model will include a vector of control variables X_i:

$$\text{average\_score}_i = \beta_0+\beta_1(\text{self\_study\_hours}_i)+\beta_2(\text{absence\_days}_i)+\beta_3(\text{extracurricular}_i)+\gamma' X_i+\epsilon_i,$$

where X_i includes gender, age, and career aspiration indicators. Comparing the baseline and expanded models will help assess how sensitive the estimated behavioral effects are to these controls.

5.3 Diagnostics and Extensions

After fitting the models, we plan to:

- Inspect residual plots to assess linearity, constant variance, and potential outliers.
- Check for multicollinearity among predictors using variance inflation factors (VIFs).
- Explore possible interactions, for example between study hours and absences, to see whether the association of study time with performance differs by attendance level.
- Consider subject-specific models (e.g., math score as the outcome) if time permits, motivated by findings that study time effects can vary by domain.

These steps will form the core of the analysis plan for the next phase of the project.

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.