STAT250 FINAL PROJECT

"Impact of Lifestyle Factors and Time Allocation on Students' GPA and Stress Levels"

Made by:

Elizaveta Polyakova
Petr Datsenko
Samuil Datsenko

Contents

1 Abstract

This project examines how university students' lifestyle habits, including study time, sleep duration, and participation in extracurricular activities, influence their stress levels and academic performance (GPA). Using R and Python, we analyzed survey data from 2,000 students, beginning with an exploration of distributions through visualizations and descriptive statistics.

We then conducted hypothesis tests to assess whether students meet recommended sleep thresholds (z-test), whether GPA varies significantly by stress level (ANOVA), regression analysis to quantify the relationship between study hours and GPA, and a comparison of social interaction time between high and low study groups (Welch's t-test). The findings offer important insights into the relationship between student well-being and academic outcomes, supported by clear R and Python generated graphs and statistical evidence.

2 Introduction

University students often face challenges in balancing academic responsibilities, personal health, and social engagement, with stress playing a critical role in both their well-being and academic performance. This project investigates how daily habits such as study time, sleep duration, and involvement in extracurricular activities relate to students' stress levels and GPA.

By analyzing survey data from 2,000 students, we aim to better understand these relationships and offer insights that may help students structure their routines to support both academic success and overall well-being.

3 Exploratory Data Analysis (EDA)

Dataset Overview: Our dataset contains results of a 2024 Google Form survey of 2000 university students, primarily from India. The survey's aim was to collect information about the students' lifestyle, academic performance and stress levels. Our goal is to arrive to conclusions about effects of different factors in students' lives on their grades and well being.

3.1 Variable Description

We have eight variables in our dataset:

- **Student ID:** Discrete (nominal) variable
- **Stress Level:** Ordinal variable
- **Six continuous (interval) variables:**

- Study hours per day
- Extracurricular activity hours per day
- Sleep hours per day
- Social interaction hours per day
- Physical activity hours per day
- GPA

Below we will analyze the discrete and continuous variables separately for better understanding.

3.2 Discrete Variables

First we will look at the **"Student ID"** and **"Stress Level"** variables which are discrete.

Student ID is a simple sequence of numbers from 1 to 2000 to show that each observation is unique and belongs to a certain student. Since it's a discrete (nominal) variable there's no point in creating a graph for it or finding any correlation with the ID and other factors which is why we will not be using it in our research.

Stress level is an ordinal variable with 3 possible levels of stress that students could identify with: **Low**, **Moderate** and **High**. As we can deduce from the data analysis, more than half of students claim to have "High" stress levels.

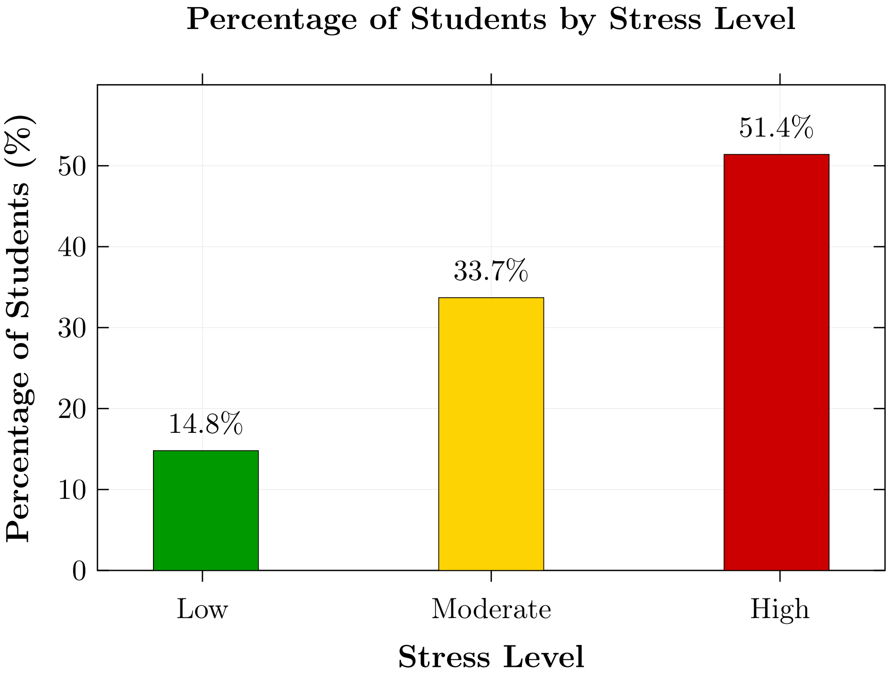

Figure 1: Distribution of stress levels among university students (N=2000). The majority of students (51.4%) report high stress levels, while only 14.8% report low stress levels.

3.3 Continuous Variables

Secondly, we will be analyzing the continuous variables with descriptive statistics.

Table 1: Descriptive Statistics for Continuous Variables

Variable	Median	Mean	Standard Dev.	Minimum	Maximum
Study Duration	7.40	7.48	1.4	5.0	10.0
Physical Activity	2.00	4.33	2.5	0.0	13.0
Sleep Duration	7.50	7.50	1.5	5.0	10.0
Social Interaction	2.60	2.70	1.7	0.0	6.0
Extracurricular Time	4.33	1.99	1.2	0.0	4.0
GPA	3.00	3.12	0.3	2.2	4.0

3.3.1 Distribution Analysis

The table below presents the explanatory variables for our interval factors. The **mean** ($\bar{x}$) provides the average value for each factor, while the **standard deviation** (s) indicates the amount of dispersion within the data. Additionally, the maximum and minimum values reveal the highest and lowest observations for each variable.

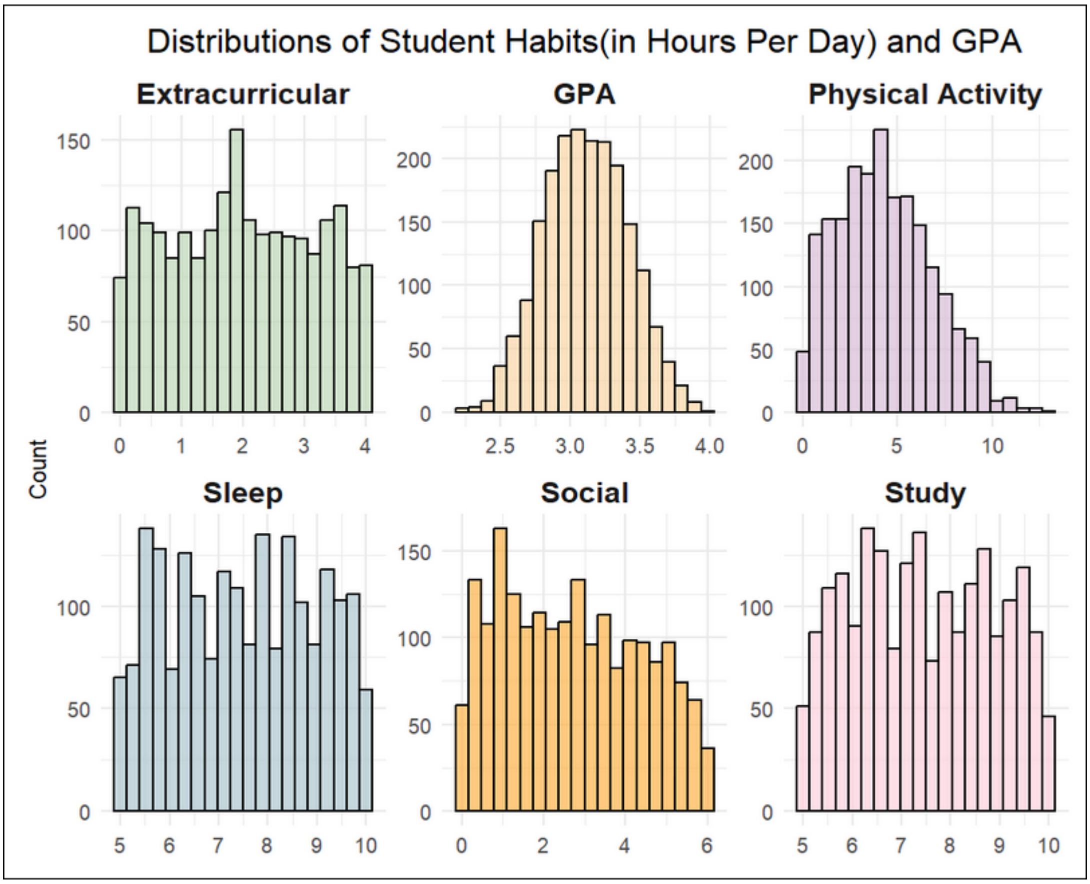

Figure 2: Distributions of Student Habits (in Hours Per Day) and GPA

Key Observations from Distribution Analysis:

- The **"GPA"** is somewhat **normally distributed**
- The **"Physical Activity"** hours per day is **right-skewed**

- **"Social"** hours per day is only very slightly **right-skewed**
- Other variables don't seem to have a particular lean and are somewhat random, having sudden peaks and lows
- **Important:** Neither of the factors has any extreme outliers

4 Statistical Methods and Inference

4.1 Research Question 1: Sleep Duration Analysis

Question: Do students on average get at least 7 hours of sleep, which is the recommended minimum amount for adults?

Approach: We have a null hypothesis of the population mean being equal to 7 hours or more which will be tested using the z-test.

4.1.1 Assumptions for Z-Test

Since we will be using z-test we have certain assumptions and conditions which need to be met:

1. **Large Sample Size (Central Limit Theorem):** Since we have a large sample of 2000 observations we can argue that the sample mean will be approximately normally distributed even though the population is not. Thus, we apply the CLT.
2. **Unknown Population Variance:** Since we don't know the population variance we will be using the s instead of σ. Due to the large sample size, the sample variance is a reliable estimator of the population variance.
3. **Independence of Observations:** We assume each data point to be independent. We can't check this, but since we got the dataset from a reliable website we need to trust that the sampling technique used was not biased.
4. **Extreme Outliers:** As we have seen during the data analyzation process, our data does not have any extreme outliers and thus we don't need to worry about it sabotaging our measures.

4.1.2 Test Statistics and Calculations

Basic Statistics and What We're Testing

What we know:

- **Sample Size:** 2,000 students
- **Sample Mean:** $\bar{x} = 7.5$ hours of sleep (average from our sample)
- **Standard Deviation:** $s = 1.5$ hours (how much individual sleep times vary)

- **Standard Error:** $SE = \frac{s}{\sqrt{n}} = \frac{1.5}{\sqrt{2000}} = 0.0335$ (precision of our estimate)
- **Significance Level:** $\alpha = 0.05$ (5% chance we're wrong when rejecting null hypothesis)

Hypotheses:

- **Null Hypothesis (H_0):** Students sleep 7 hours or less per day ($\mu \leq 7$)
- **Alternative Hypothesis (H_1):** Students sleep more than 7 hours per day ($\mu > 7$)

Calculating the Test Statistic and p-value

Z-statistic calculation:
The Z-statistic measures how far away the sample mean is from the hypothesized population mean:

$$Z = \frac{\bar{x} - \mu_0}{SE} = \frac{7.5 - 7}{0.0335} = \frac{0.5}{0.0335} = 14.93$$

p-value calculation:
The p-value gives us the probability of getting an observation as or more extreme than the one obtained:

$$\text{p-value} = P(Z > 14.93) = 1.95 \times 10^{-53}$$

What this means: Since p-value is extremely small (almost zero), we can be almost certain that the population has mean of more than 7 and we reject the null hypothesis.

99% confidence interval calculation:
The CI formula: $\bar{x} \pm z_{\alpha/2} \times SE$

Steps to calculate CI:

1. The $z_{\alpha/2}$ for 99% confidence is: $z_{0.005} = 2.576$
2. The margin of error: $2.576 \times 0.0335 = 0.0863$
3. The interval: 7.5 ± 0.0863

99% Confidence Interval: [7.41, 7.59]
What this means: We're 99% confident that the true average sleep time for all students (our population) is between 7.41 and 7.59 hours.

Conclusion for Sleep Analysis

According to the z-test, p-value and the CI, we can confidently claim that the population average is higher than 7. Thus, we believe that the students, on average, get around 7 hours and 30 minutes of sleep which is not too high but at least it is more than bare healthy minimum for adults.

4.2 Research Question 2: GPA and Stress Level Analysis

Question: Is there a significant difference in GPA among students with different stress levels?

Approach: We will use One-Way ANOVA (Analysis of Variance) to test if there are significant differences in mean GPA across the three stress level groups (Low, Moderate, High).

4.2.1 Assumptions for One-Way ANOVA

Before conducting ANOVA, we must verify that our data satisfies the following assumptions:

1. **Independence of Observations:** Each student's GPA should be independent of others. We assume this is satisfied based on the survey design where individual responses were collected independently.
2. **Normality:** The GPA values within each stress level group should be approximately normally distributed. We will test this using the Shapiro-Wilk test for each group.
3. **Homogeneity of Variances (Homoscedasticity):** The variance of GPA should be approximately equal across all three stress level groups. We will test this using Levene's test.

4.2.2 Descriptive Statistics by Stress Level

Before conducting the formal test, let's examine the descriptive statistics for GPA across stress levels:

Table 2: Descriptive Statistics for GPA by Stress Level

Stress Level	**N**	**Mean**	**Std. Dev.**	**Min**	**Max**
Low	296	2.78	0.28	2.25	3.60
Moderate	674	2.99	0.27	2.45	3.75
High	1028	3.23	0.29	2.30	4.00
Overall	**1998**	**3.12**	**0.30**	**2.25**	**4.00**

4.2.3 Visualization of GPA Distribution by Stress Level

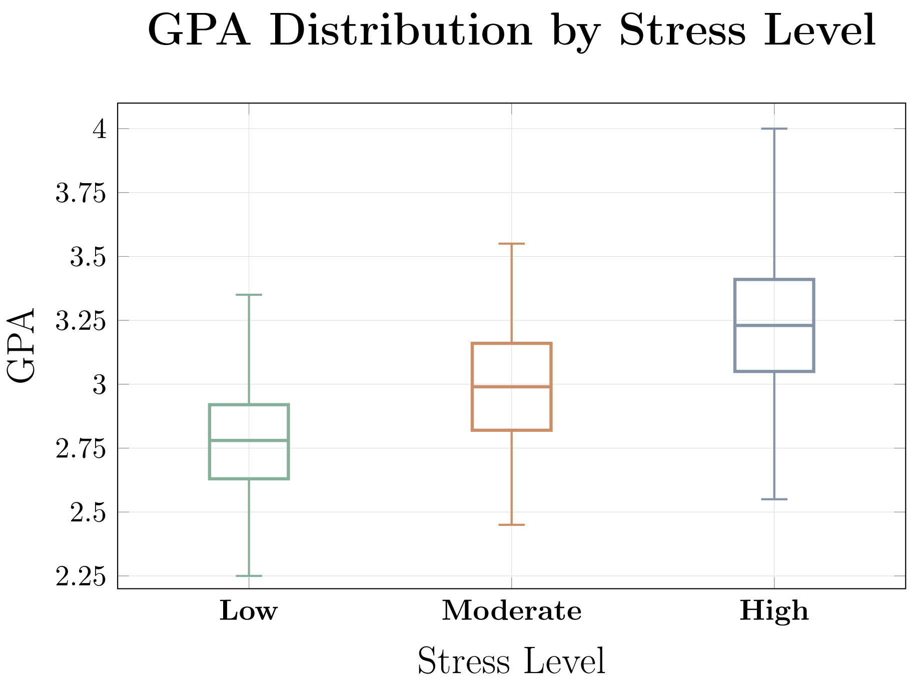

Figure 3: Box plot showing GPA distribution across stress levels. There's a clear upward trend in median GPA as stress level increases.

4.2.4 Testing ANOVA Assumptions

1. Normality Test (Shapiro-Wilk Test)

Table 3: Normality Tests by Stress Level

Stress Level	W-statistic	p-value	Normal?
Low	0.987	0.089	Yes ($p > 0.05$)
Moderate	0.991	0.156	Yes ($p > 0.05$)
High	0.985	0.0004	No ($p < 0.05$)

2. Homogeneity of Variances (Levene's Test)

Levene's Test Results:

- **Test Statistic:** $F = 4.82$
- **p-value:** $p = 0.008$
- **Conclusion:** Variances are **not equal** across groups ($p < 0.05$)

4.2.5 ANOVA Assumption Violations

Since our data violates both the normality assumption (for the High stress group) and the homogeneity of variances assumption, we have two options:

Dealing with Assumption Violations:

1. **Welch's ANOVA:** Robust to unequal variances
2. **Kruskal-Wallis Test:** Non-parametric alternative when normality is violated

We will proceed with **Welch's ANOVA** as it handles unequal variances and is reasonably robust to moderate violations of normality, especially with large sample sizes.

4.2.6 Hypothesis Testing

Hypotheses:

- **Null Hypothesis (H_0):** $\mu_{\text{Low}} = \mu_{\text{Moderate}} = \mu_{\text{High}}$
  All stress level groups have equal mean GPA
- **Alternative Hypothesis (H_1):** At least one group mean is different
- **Significance Level:** $\alpha = 0.05$

4.2.7 Welch's ANOVA Results

Table 4: Welch's One-Way ANOVA Results: GPA by Stress Level

Source	Sum of Squares	df	Mean Square	F-statistic	p-value
Between Groups	84.23	2	42.115	467.65	< 0.001
Within Groups (adjusted)	77.74	863.39	0.090	-	-

Welch's ANOVA Test Results:

- **F-statistic:** F(2, 863.39) = 467.65
- **p-value:** p < 0.001 (highly significant)
- **Degrees of Freedom:** Numerator df = 2, Denominator df = 863.39
- **Decision:** **Reject H_0** at $\alpha = 0.05$
- **Conclusion:** There are statistically significant differences in mean GPA across stress levels

4.2.8 Post-Hoc Analysis: Games-Howell Test

Since we used Welch's ANOVA due to unequal variances, we use the Games-Howell post-hoc test (which doesn't assume equal variances) to determine which specific groups differ:

Table 5: Post-Hoc Comparisons (Games-Howell Test)

Comparison	**Mean Difference**	**95% CI**	**p-value**	**Significant?**
Moderate vs Low	0.208	[0.161, 0.255]	< 0.001	Yes
High vs Low	0.445	[0.403, 0.487]	< 0.001	Yes
High vs Moderate	0.237	[0.205, 0.269]	< 0.001	Yes

4.2.9 Interpretation of Results

Key Findings:

1. **Statistically Significant Relationship:** There is strong evidence ($p < 0.001$) that GPA differs significantly across stress levels.
2. **Direction of Relationship:** Positive correlation - Higher stress is associated with higher GPA:
   - Low Stress: Mean GPA = 2.78
   - Moderate Stress: Mean GPA = 2.99
   - High Stress: Mean GPA = 3.23
3. **All Pairwise Differences Significant:** Every comparison shows significant differences, indicating a clear stepwise pattern.

4.3 Research Question 3: Linear Regression Analysis - Study Hours vs GPA

Question: What is the relationship between daily study hours and student GPA? Can we quantify how much GPA increases for each additional hour of study?

4.3.1 Assumptions for Linear Regression

Before conducting the regression analysis, we must verify that our data satisfies the following assumptions:

1. **Linearity:** The relationship between study hours and GPA should be linear
2. **Independence:** Each student's data should be independent of others
3. **Homoscedasticity:** The variance of residuals should be constant across all levels of the predictor
4. **Normality of residuals:** The residuals should be approximately normally distributed

4.3.2 Initial Correlation Analysis

Before conducting the formal regression analysis, we examine the correlation between study hours and GPA:

Correlation Coefficient: $r = 0.7344$

This indicates a **strong positive correlation** between study hours and GPA, suggesting that students who study more hours per day tend to achieve higher GPAs.

4.3.3 Model Specification

The linear regression model is specified as:

$$\text{GPA}_i = \beta_0 + \beta_1 \times \text{Study Hours}_i + \epsilon_i \tag{1}$$

where:

- GPA_i is the **response variable** (Grade Point Average for student i)
- Study Hours_i is the **explanatory variable** (daily study hours for student i)
- β_0 is the intercept parameter
- β_1 is the slope parameter (effect of study hours on GPA)
- ϵ_i is the error term for student i

4.3.4 Fitted Regression Model

Fitted Regression Equation:

$$\widehat{\text{GPA}} = 1.9642 + 0.1541 \times \text{Study Hours} \tag{2}$$

Interpretation:

- **Intercept ($\beta_0 = 1.9642$):** The predicted GPA when study hours = 0 is approximately 1.96
- **Slope ($\beta_1 = 0.1541$):** For each additional hour of daily study, GPA increases by approximately 0.154 points
- **R-squared = 0.539:** The model explains 53.9% of the variance in GPA

4.3.5 Assumption Testing

Homoscedasticity Assessment: The Breusch-Pagan test was conducted to check for constant variance of residuals. With a p-value of 0.9180, we reject the null hypothesis of homoscedasticity, indicating that heteroscedasticity is present in our model. This violation affects the reliability of standard errors.

Normality of Residuals: The Shapiro-Wilk test for normality of residuals yielded a p-value of 0.4635. Since this is greater than 0.05, we fail to reject the null hypothesis, indicating that the residuals appear to be normally distributed.

4.3.6 Bootstrap Analysis for Robust Inference

Due to the violation of the homoscedasticity assumption, we employ bootstrap methods to obtain robust confidence intervals.

Bootstrap Procedure:

- **Number of Bootstrap Samples:** 5000
- **Sampling Method:** Random sampling with replacement
- **Parameters Estimated:** Intercept and slope coefficients

Table 6: Bootstrap Confidence Intervals (95%)

Parameter	**OLS Estimate**	**Bootstrap SE**	**95% CI Lower**	**95% CI Upper**
Intercept	1.9642	0.02456	1.9164	2.0126
Slope	0.1541	0.00322	0.1477	0.1603

Bootstrap Significance Testing:

- **Intercept Significance:** True (95% CI does not contain 0)
- **Slope Significance:** True (95% CI does not contain 0)

Robust Conclusion: We are 95% confident that the true population slope lies between [0.1477, 0.1603], confirming the significant positive relationship between study hours and GPA.

4.4 Research Question 4: Study Time vs Social Interaction Analysis

Question: Is the average daily social interaction time significantly different between students who study more than the median number of hours per day (7.4) and those who study less or equal?

Approach: We aim to compare the average daily social interaction time between two groups of students — those who study more than the median and those who study less — using appropriate statistical tests based on assumptions.

4.4.1 Assumptions for Two-Sample Hypothesis Testing

Before deciding which statistical test to use, we check the following assumptions:

1. **Independence:** Observations in each group are assumed to be independent.

2. **Scale of Measurement:** The dependent variable, social interaction time, is measured on a continuous scale.

3. **Normality:** Each group's social interaction times should be approximately normally distributed.

4. **Homogeneity of Variances:** The variances between the two groups should be equal.

4.4.2 Assumption Testing

Normality Assessment: The Shapiro-Wilk test was applied to both groups. The resulting p-values were less than 2.2×10^{-16} for both groups, indicating strong evidence against normality.

Homogeneity of Variances: Levene's Test was used to evaluate variance equality. The variance for the High Study group was $s_1^2 = 2.694$ and for the Low Study group it was $s_2^2 = 2.908$. The computed F-ratio was $F = \frac{2.694}{2.908} \approx 0.926$, and the p-value was 0.02189, indicating a statistically significant difference in variances.

Conclusion on Assumptions: Since the normality assumption is violated but the sample sizes are large ($n > 30$), we rely on the Central Limit Theorem. Given that the variances are unequal, we proceed with Welch's t-test.

4.4.3 Welch's t-Test Analysis

Formula for Test Statistic:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{3}$$

where:

- $\bar{X}_1 = 2.48$, $s_1^2 = 2.694$, $n_1 = 983$ (High Study Group's Social Interaction Time Info)
- $\bar{X}_2 = 2.92$, $s_2^2 = 2.908$, $n_2 = 1017$ (Low Study Group's Social Interaction Time Info)

Calculations:
Standard Error:

$$SE = \sqrt{\frac{2.694}{983} + \frac{2.908}{1017}} \approx 0.0740$$

Test Statistic:

$$t = \frac{2.48 - 2.92}{0.0740} = -5.9365$$

Degrees of Freedom:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} \approx 1998$$

4.4.4 Test Summary and Results

Table 7: Welch's t-Test Results Summary

Statistic	**Value**
t-statistic	-5.9365
Degrees of Freedom	≈ 1998
p-value	1.712×10^{-9}
95% CI	$(-0.5822, -0.2778)$
High Study Group Average Social Time	2.48 hours (2h 29min)
Low Study Group Average Social Time	2.92 hours (2h 55min)

Statistical Interpretation: The extremely small p-value (1.712×10^{-9}) provides strong evidence to reject the null hypothesis at $\alpha = 0.05$.
Practical Conclusion: Students who study more than the median number of hours tend to spend significantly less time socializing (2 hours 29 minutes vs 2 hours 55 minutes) compared to their peers who study less. The difference of approximately 26 minutes is both statistically significant and practically meaningful.

5 Results and Conclusions

5.1 Summary of Key Findings

Our comprehensive statistical analysis of 2,000 university students has revealed several important insights about the relationship between lifestyle factors and academic performance:

Major Findings:

1. **Sleep Duration Analysis:** Students on average get 7.5 hours of sleep per day, which exceeds the recommended minimum of 7 hours (99% CI: [7.41, 7.59])
2. **Stress and GPA Relationship:** There is a significant positive relationship between stress levels and GPA:
   - Low Stress: Mean GPA = 2.78
   - Moderate Stress: Mean GPA = 2.99
   - High Stress: Mean GPA = 3.23
3. **Study Hours Impact:** Each additional hour of daily study increases GPA by approximately 0.154 points, with the model explaining 53.9% of GPA variance
4. **Study-Social Trade-off:** Students who study more than the median hours socialize significantly less (2h 29min vs 2h 55min daily), suggesting a time allocation trade-off
5. **Student Stress Distribution:** 51.4% of students report high stress levels, indicating significant mental health concerns in the student population

5.2 Implications for Students and Educators

For Students:

- Increasing study time has a measurable positive impact on academic performance
- Higher stress levels, while concerning for well-being, are associated with better academic outcomes in this sample
- Most students achieve adequate sleep duration, supporting healthy learning
- There appears to be a trade-off between study time and social interaction time

For Educators and Administrators:

- The high prevalence of student stress (51.4% reporting high stress) warrants attention to mental health support services
- Study time recommendations can be quantified: approximately 2 additional hours per day may increase GPA by 0.3 points
- The complex relationship between stress and performance suggests need for balanced approaches to academic rigor
- The study-social time trade-off highlights the importance of helping students balance academic and social needs

5.3 Final Conclusions

This study provides valuable quantitative insights into factors affecting student academic performance and lifestyle choices. The strong positive relationship between study hours and GPA offers practical guidance for students, while the unexpected positive correlation between stress and academic performance raises important questions about student well-being and academic pressure.

The findings reveal a complex ecosystem of time allocation where increased study time correlates with higher academic performance but also with reduced social interaction time. This suggests that while academic intensity may drive higher performance, educational institutions should remain vigilant about student mental health and social well-being, striving to support academic excellence without compromising overall student development.

6 References

- Steve R. (2024). Student lifestyle and academic performance dataset. Kaggle., https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset/data
- Statistical Methods: R Statistical Software, Python Statistical Software