Student Lifestyle Factors and their Effects

Jonas Priolo, Will Bolger, Alejandro Herraiz Sen

2025-04-25

Introduction

Research Topic: Student Lifestyle Factors and their Effects

This research focuses on Student lifestyle trends and their effects on things such as their grades and stress levels. With depression and anxiety rates being at all time highs and finals week just around the corner this subject is not only relevant to society but to everyone in our group/class. Our research revolves around gaining a better understanding on what effects student outcomes such as stress and GPA pertaining to a students daily study hours, extracurricular hours, sleep hours, social hours, and physical activity. Via researching our topic in depth, mapping our data in data sets, and by utilizing data visualizations we can explore the relations between our key attributes and student well being. Our goal is to increase our understanding of the effects of different lifestyle factors and how much they affect our lives as well as presenting our research in a way our reader could deepen their own knowledge for this topic.

Research Questions

The first research question we will explore is how different lifestyle factors, such as hours spent studying, sleeping, socializing, exercising, or time spent doing extracurriculars affect both student stress levels and academic performance. We will create different visualizations to present our research findings and explain the relationships between each of our key attribute lifestyle factors and their effects. In addition to this we will also examine gender and its effects on the different items we are analyzing. We are also interested in examining a student's major and its impact on their mental health. For example, is a student who is studying an engineering related discipline more stressed or depressed then a student who is studying business? Another question we want to answer is weheter an individual who is on a scholorship has higher levels of stress compared to someone without one. For our research we will also need to be aware of potential biases we may have. All three of us are males and are pursuing STEM fields and we cant let our own experiences alter our conclusion.

Provenance Of Our Data

The data sets we are using came from the Website Kaggle, a data science website that offers open source resources and data sets with the goal to help others learn more about data. The author of our first and primary data set is Charlotte Bennett. The author describes there data as a “detailed view of student lifestyle patterns and their correlation with academic performance, represented by GPA.” The data contains detailed student survey data across a variety of student lifestyle factors, student demographics, and academic outcomes. This data was last updated 21 days ago and was sourced from a Google Form survey focusing on students across different educational institutions, primarily focusing on those in India and other South Asian countries. All in all there were 2000 voluntary participants for the survey used in the data set and the respondents were informed that the data would be used for educational purposes only. No personally identifiable information was collected for the data set. In this data set, a case is an individual student. The data set includes the attributes of Student ID, Study Hours per Day, Extracurricular Hours per Day, Sleep Hours Hours per Day, Social Hours per Day, Physical Activity Hours per Day, Stress Level, Gender, And Grades (CGPA). For this data set we will touch upon all of these attributes. For this data set we intend on converting the ten point GPA scale provided in the data set, which is the standard grading scale in India and South Asia, to the four point GPA scale we are more familiar with.

Our second and supplementary data set also originates from Figshare. The author of this data set is Mahbubul and he describes the data set as “a statistical research on the effects of mental health.” The data for this set was collected via survey form from students studying at the top 15 ranked universities in Bangladesh. The data set includes the attributes of a age, gender, University, major, academic year, GPA, scholarship status, answers to survey questions, student stress levels, student anxiety levels, and student depression levels. For this data set we are primarily interested in the students major, there scholarship status, and there stress level.

CARE Principles

The data we are utilizing meets the CARE principles. The data can be used for the Collective benefit because it allows for us to identify what are the main factors that contribute to student well being. When properly analyzed as a society we can improve student resources and improve student mental health. This data meets the Authority to control because the data was collected via an optional survey. This means that participants had the autonomy to choose whether or not to share their experiences, ensuring respect for an individual’s data and consent. The data aligns with Responsibility because it emphasizes the ethical use for this information such as prioritizing student welfare and protecting student privacy. Lastly this data is ethical with it promoting equity and positive change with its ability to inform policies to support student well-being.

Main Data Set

1. Data Tidying

For our main data set, it was already tidy, and didn't require any changes to it. Each row was considered a single case, with each case being a student. Each column represented a single variable, with no cell containing multiple data points. The variables in this case was student id, study hours per day, extracurricular activity per day, sleep hours per day, social hours per day, physical activity hours per day, reported stress level, their gender, and their GPA.

2. Data Wrangling

For data wrangling, we did reshape the data to make it more understandable to US audiences. In the US, the primary GPA scale is a 4 point scale, while our data was in a 10 point scale. We converted it into a 4 point scale by multiplying it by 0.4.

3. Data Cleaning

There was no data cleaning required as there were no missing values, no incorrect values, and no duplicates among the data set. BY summing the duplicate function on our data set, we can see that there are no duplicates.

```
[1] 0
```

Exploratory Data Analysis

Before creating any data visualizations, we created a frequency that alalyise a series of statistics to better understand the relation between student stress levels and a students grades. Table 1 shows a summary table that includes *count*, *minimum*, *Q1*, *median*, *Q3*, *maximum*, *median absolute deviation*, *mean*, and *standard deviation* for every stress level and student GPA. We are interested in using this data to gain a better understanding of student stress and its impacts and we thought that a students GPA was a good metric to start with.

Scatter Plot Matrices

We created two different scatter plot matrices to help us get a baseline overview of the data to help us better understand what we were analyzing. We took out three columns from the original data set, as they would not give valuable information in these visualizations. These three were the student id, gender, and GPA. We took out student id as it is a number assigned to each student and has no impact on any of the other variables. Next, we used gender and stress level to create 2 different scatter plot matrices as they were both categorical data, and would be unhelpful to directly graph against continuous data. After eliminating those three variables, we plotted the other variables of study hours per day, extracurricular hours per day, sleep hours per day, social hours per day, physical activity hours per day, and the students GPA on a 4.0 scale.

For this first scatter plot matrix, the different colored points represents the different stress levels reported by the students. The green points represent those with a reported high stress, red with a reported moderate stress level, and blue representing a reported low stress level.

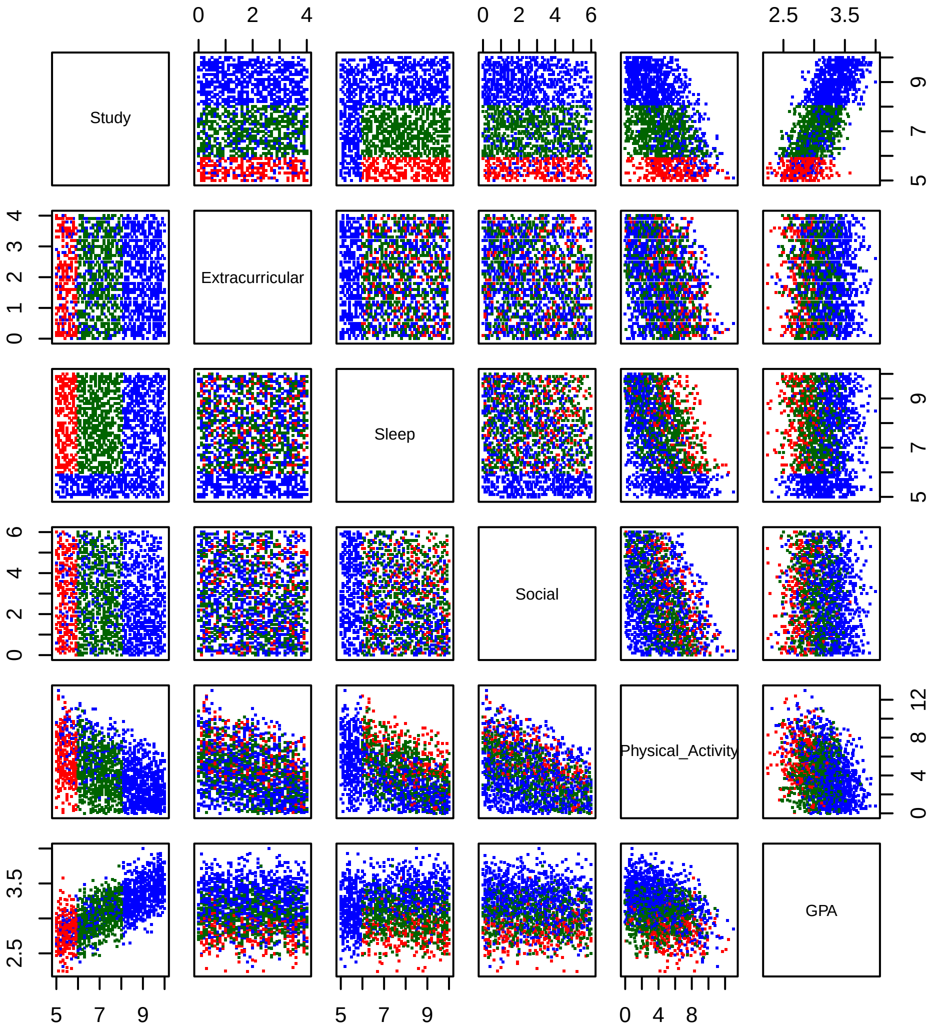


Figure 1: Scatter plot matrix with stress level

For this first scatter plot matrix, the different colored points represents the different stress levels reported by the students. The green points represent those with a reported high stress, red with a reported moderate stress level, and blue representing a reported low stress level.

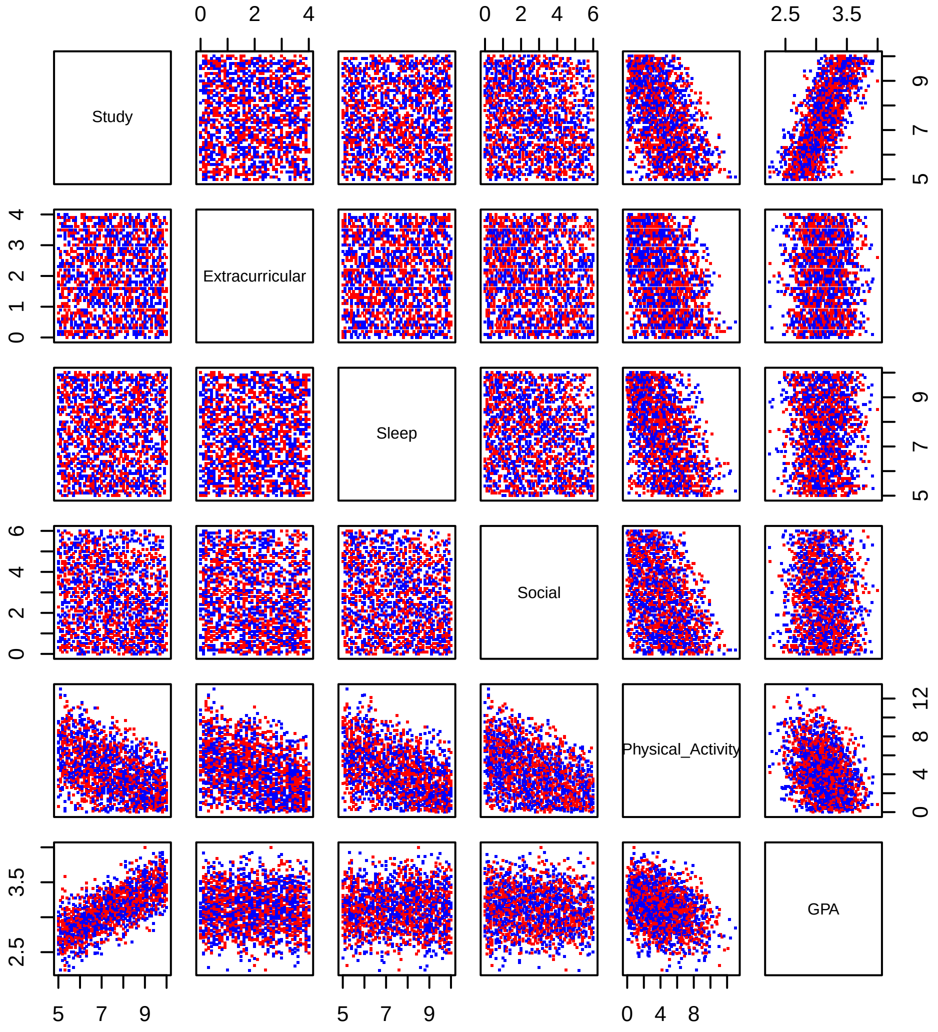


Figure 2: Scatter plot matrix with Gender

For the second scatter plot matrix, we mapped the color to the gender of each student in the data set. Each blue point represents a blue student and each red point represents a female student. This was done to help meet CARE principles. We wanted to make sure that there was not only an even representation of male and females, but that it also wasn't skewed in any particular way that would lead to results that could be interpreted as sexist against either gender.

In these scatter plots, we noticed a noticeable possible relation between the variables with the GPA and stress. Study hours and physical activity have the strongest visible relation with grades. An increase in stress seemingly occurs with an increase of the hours spent studying, doing physical activity, being social, and physical activity, while the decrease of sleep hours also increases stress.

Frequency table

Table 1: Summary Statistics on Grades by Stress Level

Stress_Level	count	min	Q1	median	max	mad	mean	Q3	sd
High	1029	2.312	3.088	3.272	4.000	0.2787288	3.261936	3.460	0.2750148
Low	297	2.240	2.680	2.820	3.580	0.2075640	2.816835	2.952	0.2154800
Moderate	674	2.440	2.872	3.020	3.752	0.2194248	3.024819	3.180	0.2206716

From the summery table we can get a better understanding of the data we are working with. If we look at the *count* column we can see that the highest stress level among students is High, followed by Moderate, then low. It was also interesting to see how stress level impacted the various statistics in relation to a students GPA. From the table we were able to find out that there is a relation between a students stress level and GPA with higher levels of stress being related to higher GPAs. We also found it interesting how students with low stress typically had lower GPAs, signaling that caring about your grades makes you more stressed and vice versa.

Frequency graphs

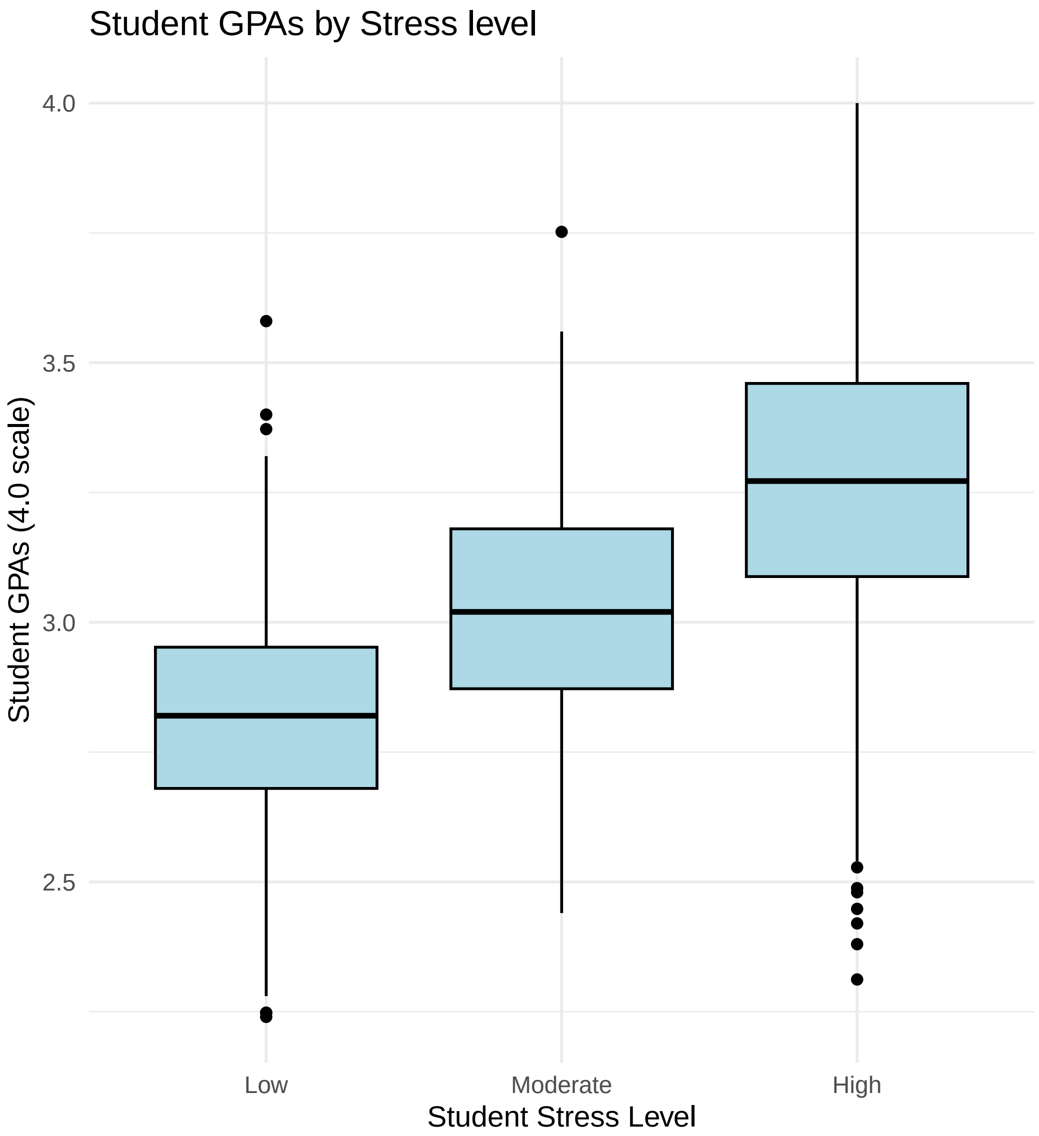


Figure 3: Student GPAs by stress level

The Box and Whisker plot (**gpa-stress_plot?**) allows us to confirm what we learned about the relationship between a students stress level and there GPA from our summery table. The plot visually displays the distributions for each stress level by showing the median, upper and lower quartiles, and any outlier points. From this plot we are able to able to verify that students with higher levels of stress tend to have higher GPAs, while those with lower stress levels tend to have lower GPAs. This reinforces the positive relation observed in the data.

Main Analysis

Hypothesis

This study examines how different aspects of students' daily routines, specifically the amount of time dedicated to studying, sleep duration, physical activity, social interaction, and participation in extracurricular activities, are associated with both self-reported stress levels and academic performance as measured by GPA. To drive the analysis, the following detailed hypotheses are proposed:

1. Stress Level and GPA
   - H0: The mean GPA is equal across all stress categories (Low, Moderate, High).
   - H1: The mean GPA differs across stress categories.
2. Study Hours
   - H0 (gpa): There is no linear relationship between daily study hours and GPA .
   - H1 (gpa): Daily study hours are correlated with GPA .
   - H0 (stress): Mean daily study hours are equal across stress categories.
   - H1 (stress): Mean daily study hours differ across stress categories.
3. Sleep Duration
   - H0 (gpa): There is no linear relationship between sleep hours and GPA .
   - H1 (gpa): Sleep hours are correlated with GPA .
   - H0 (stress): Mean sleep hours are equal across stress categories
   - H1 (stress): Mean sleep hours differ across stress categories.
4. Physical Activity
   - H0 (gpa): No correlation between physical activity hours and GPA .
   - H1 (gpa): Physical activity hours are correlated with GPA .
   - H0 (stress): Mean physical activity hours are equal across stress categories.
   - H1 (stress): Mean physical activity hours differ across stress categories.
5. Social Engagement

- H0 (gpa): No correlation between social hours and GPA.
- H1 (gpa): Social hours are correlated with GPA.
- H0 (stress): Mean social hours are equal across stress categories.
- H1 (stress): Mean social hours differ across stress categories.

6. Extracurricular Involvement

   - H0 (gpa): No linear association between extracurricular hours and GPA.
   - H1 (gpa): Extracurricular hours are correlated with GPA.
   - H0 (stress): Mean extracurricular hours are equal across stress categories.
   - H1 (stress): Mean extracurricular hours differ across stress categories.

7. Anxiety

   - H0: Mean GPA is equal for students with and without anxiety.
   - H1: Mean GPA differs between students with and without anxiety.

8. Depression

   - H0: Mean GPA is equal for students with and without depression.
   - H1: Mean GPA differs between students with and without depression.

9. Panic Attacks

   - H0: Mean GPA is equal for students with and without panic attacks.
   - H1: Mean GPA differs between students with and without panic attacks.

Data analysis

In our preliminary analyses, we employed one-way analysis of variance (ANOVA) to compare mean outcomes between categorical groups and simple linear regression to quantify relationships between continuous variables. ANOVA is appropriate when the independent variable (for example, stress level) is categorical(has more than 2 groups) and the outcome (such as GPA) is quantitative ; regression is used when both predictor and outcome are quantitative. We set our significance threshold at $p < 0.05$, meaning that any test yielding a p-value below this cutoff leads us to reject the null hypothesis of no association or no difference. Only those relationships for which the null hypothesis is rejected will be subjected to further investigation.

Table 2: Summary of preliminary statistical tests

Analysis	Types	Test	Pvalue	H0	Investigation
GPA vs Stress	Categorical vs Quantitative	ANOVA	0.0000	Reject	Conduct Tukey HSD
Study Hours vs Stress	Categorical vs Quantitative	ANOVA	0.0000	Reject	Conduct Tukey HSD
Sleep Hours vs Stress	Categorical vs Quantitative	ANOVA	0.0000	Reject	Conduct Tukey HSD
Physical Activity vs Stress	Categorical vs Quantitative	ANOVA	0.0000	Reject	Conduct Tukey HSD
Social Hours vs Stress	Categorical vs Quantitative	ANOVA	0.0489	Reject	Conduct Tukey HSD
Extracurricular vs Stress	Categorical vs Quantitative	ANOVA	0.8977	Do not Reject	None
Study Hours vs GPA	Quantitative vs Quantitative	Linear regression	0.0000	Reject	Examine regression coefficients
Sleep Hours vs GPA	Quantitative vs Quantitative	Linear regression	0.8491	Do not Reject	None
Physical Activity vs GPA	Quantitative vs Quantitative	Linear regression	0.0000	Reject	Examine regression coefficients
Social Hours vs GPA	Quantitative vs Quantitative	Linear regression	0.0001	Reject	Examine regression coefficients
Extracurricular vs GPA	Quantitative vs Quantitative	Linear regression	0.1516	Do not Reject	None

Further Investigation

For categorical group comparisons, we will conduct Tukey's HSD (honest significant difference) post-hoc tests to identify which specific pairs of group means differ and to estimate effect sizes.For regression models, we will examine estimated coefficients, R^2, and residual diagnostics to assess the strength, direction, and robustness of the continuous associations. These follow-up tests complement the initial omnibus analyses by clarifying where and how variables are related.

Table 3: Combined Tukey HSD Pairwise Comparisons for Stress Level vs. Quantitative Outcomes

Variable	Comparison	Mean_Diff	Lower_CI	Upper_CI	P	Conclusive
GPA	Moderate-Low	0.5200	0.4181	0.6218	0.0000	Yes
GPA	High-Low	1.1128	1.0164	1.2091	0.0000	Yes

GPA	High-Moderate	0.5928	0.5203	0.6653	0.0000	Yes
Study Hours	Moderate-Low	1.4952	1.3384	1.6520	0.0000	Yes
Study Hours	High-Low	2.9106	2.7623	3.0589	0.0000	Yes
Study Hours	High-Moderate	1.4154	1.3039	1.5270	0.0000	Yes
Sleep Hours	Moderate-Low	-0.1163	-0.3424	0.1097	0.4491	No
Sleep Hours	High-Low	-1.0175	-1.2313	-0.8037	0.0000	Yes
Sleep Hours	High-Moderate	-0.9012	-1.0620	-0.7403	0.0000	Yes
Physical Activity	Moderate-Low	-1.2450	-1.6459	-0.8441	0.0000	Yes
Physical Activity	High-Low	-1.6209	-2.0001	-1.2417	0.0000	Yes
Physical Activity	High-Moderate	-0.3759	-0.6611	-0.0906	0.0057	Yes
Social Hours	Moderate-Low	-0.1513	-0.4268	0.1243	0.4022	No
Social Hours	High-Low	-0.2631	-0.5237	-0.0025	0.0472	Yes
Social Hours	High-Moderate	-0.1118	-0.3079	0.0842	0.3744	No
Extracurricular	Moderate-Low	0.0175	-0.1714	0.2064	0.9743	No
Extracurricular	High-Low	-0.0091	-0.1878	0.1696	0.9922	No
Extracurricular	High-Moderate	-0.0266	-0.1610	0.1078	0.8881	No

In Table 3, none of the three pairwise contrasts for Extracurricular Hours per Day by stress level were significant (Adjusted p = 0.9743, 0.9922, 0.8881), so we did not pursue any further comparisons for that variable.

Although the overall ANOVA for Sleep Hours per Day versus stress was highly significant (p < 0.001), only two of its three contrasts reached significance: High vs Low (mean difference = –1.0175, Adjusted p < 0.001) and High vs Moderate (mean difference = –0.9012, Adjusted p < 0.001)

The Moderate vs Low contrast was not significant (–0.1163, Adjusted p = 0.4491). This pattern—overall variability without every category contrast achieving significance—suggests that the largest drop in sleep hours is between the highest-stress group and the others, rather than a linear trend across all levels.

For Social Hours per Day, only the High vs Low contrast proved conclusive (mean difference = –0.2631, Adjusted p = 0.0472), while Moderate vs Low (–0.1513, p = 0.4022) and High vs Moderate (–0.1118, p = 0.3744) were not. Again, this points to a threshold effect at the upper social bracket.

By contrast, all pairwise contrasts for GPA, Study Hours per Day, and Physical Activity per Day yielded Adjusted p < 0.001 and will be examined in detail. Non-significant findings (Extracurricular Hours per Day, the Moderate–Low Sleep contrast, and the two non-significant Social contrasts) will be reported but not subjected to additional subgroup analysis. To display these results succinctly, we will use a forest-style plot of each mean difference with its 95 % confidence interval, faceted by outcome variable—this will clearly show which intervals exclude zero and allow direct comparison of effect sizes.

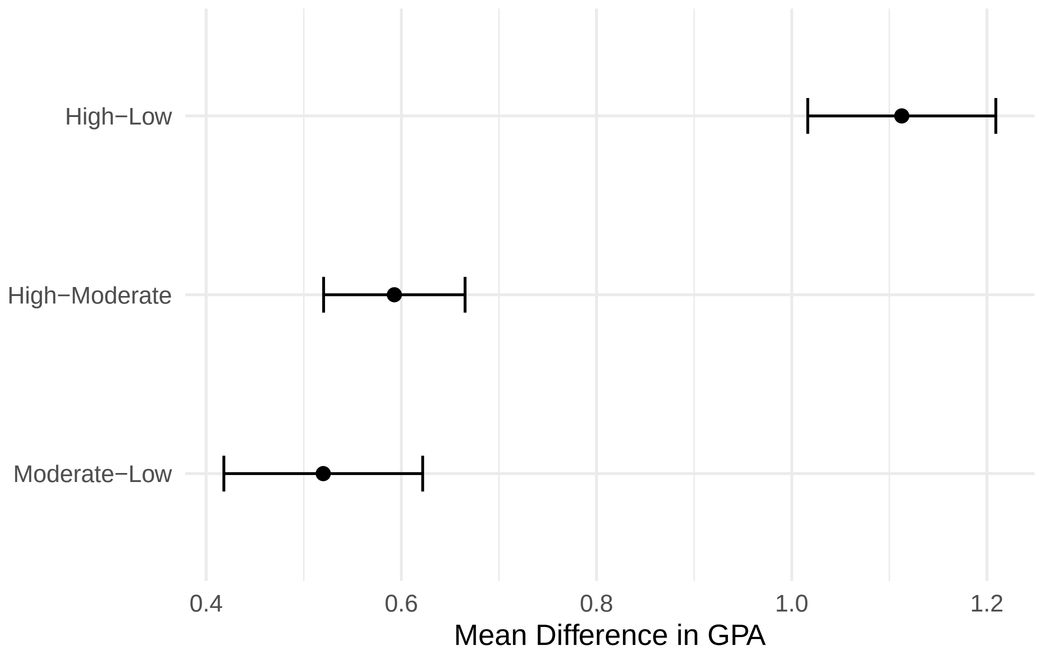


Figure 4: Tukey HSD Mean Differences for GPA by Stress Level

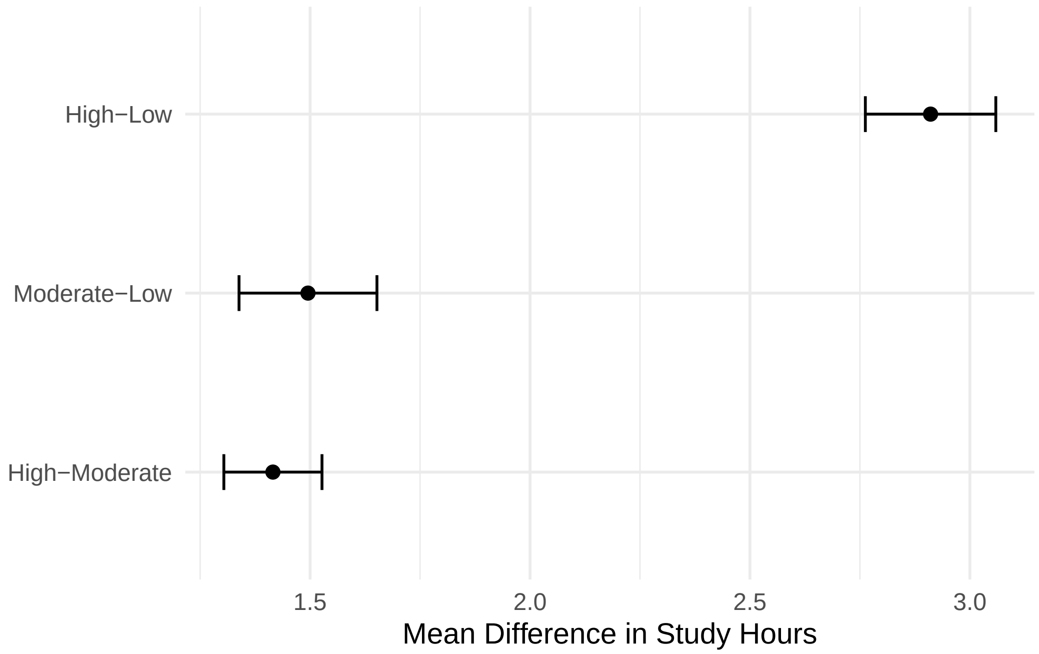


Figure 5: Tukey HSD Mean Differences for Study Hours per Day by Stress Level

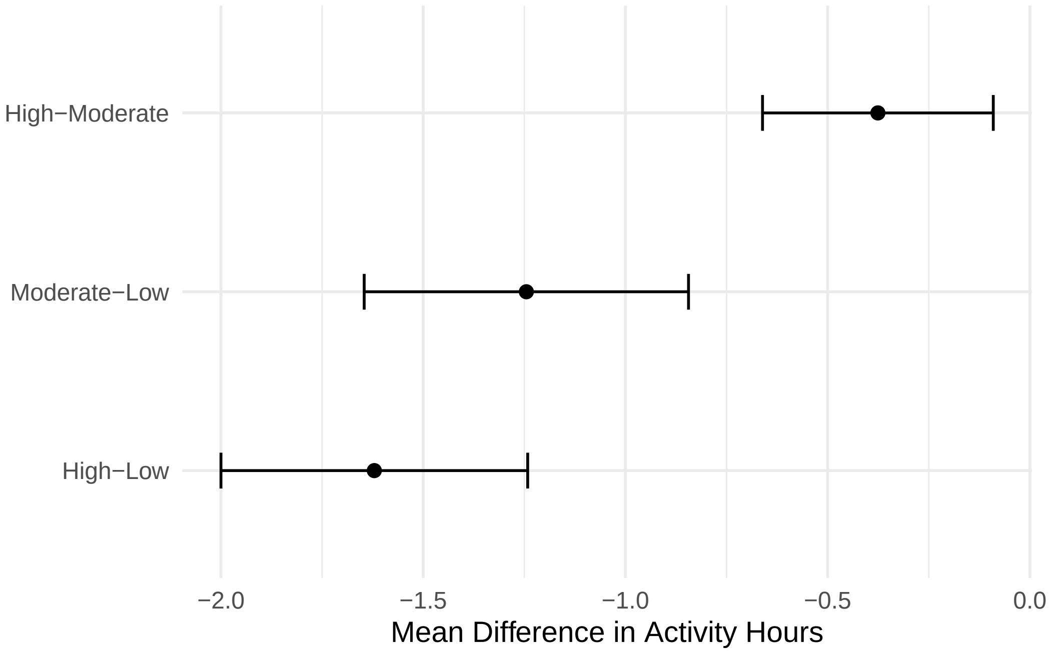


Figure 6: Tukey HSD Mean Differences for Physical Activity Hours per Day by Stress Level

In our regression analyses, we fit separate linear models predicting GPA from each quantitative predictor. For each model, we report the estimated slope coefficient, the coefficient of determination (R^2), and the overall model p-value, accompanied by a conclusive flag indicating whether the association is statistically significant at $= 0.05$. These metrics allow us to assess the direction, strength, and explanatory power of each predictor and will be supplemented by residual diagnostics to confirm model validity. This detailed reporting extends our initial tests by precisely quantifying continuous relationships.

Table 4: Detailed Linear Regression Results for Quantitative Predictors

Predictor	Estimate	P value	R squared	Conclusive
Study Hours per Day	0.3852	0.0000	0.5393	Yes
Sleep Hours per Day	-0.0022	0.8491	0.0000	No
Physical Activity per Day	-0.1013	0.0000	0.1164	Yes
Social Hours per Day	-0.0379	0.0001	0.0073	Yes
Extracurricular Hours per Day	-0.0207	0.1516	0.0010	No

In examining our detailed regression results (Table X), we see that Study Hours per Day, Physical Activity per Day, and Social Hours per Day each yield statistically significant associations with GPA (all $p < 0.001$) and explain non-trivial portions of variance ($R^2 = 0.5393$, 0.1164, and 0.0073, respectively). Specifically, the slope estimate for Study Hours is Beta = 0.3852, indicating that each additional hour of daily study predicts a 0.39-point increase in GPA on our four-point scale; with over 53% of GPA variation accounted for, this is our strongest continuous predictor and will be the centerpiece of further analysis.

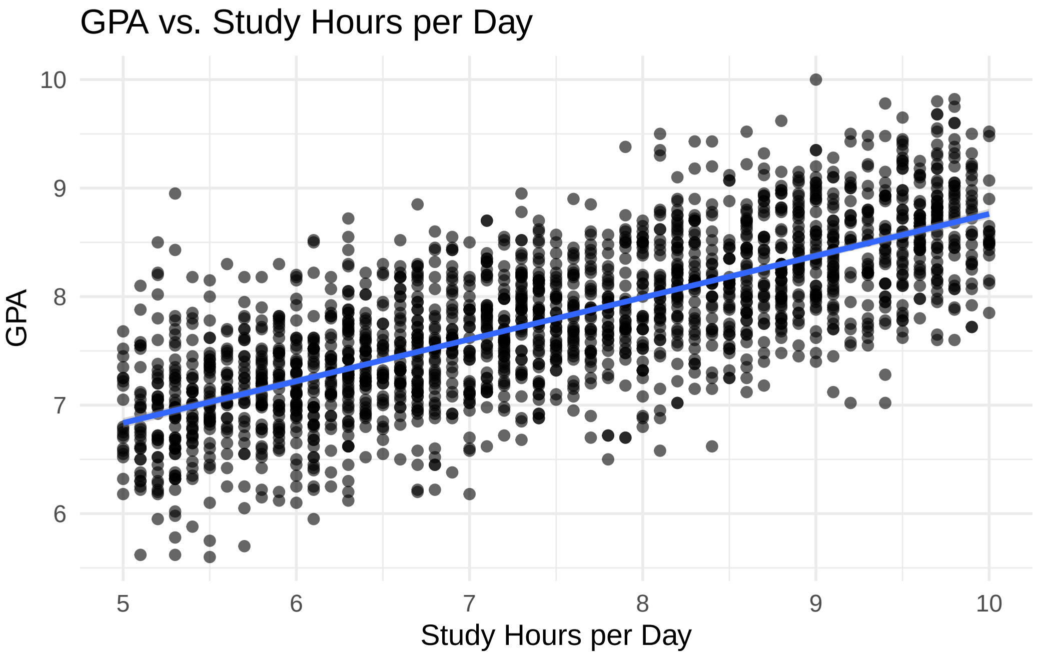
By contrast, Sleep Hours per Day shows a near-zero slope (Beta = –0.0022), a non-significant p-value (p = 0.8491), and an effectively zero R^2 (0.000), demonstrating no discernible linear relationship between sleep duration and GPA in this sample. Because the confidence interval around the slope includes zero and the model explains none of the outcome variance, we will report this null finding but will not go any further in-depth plotting or diagnostics.

Similarly, Extracurricular Hours per Day yields a small negative slope (Beta = –0.0207) that is not statistically significant (p = 0.1516) and explains only 0.1% of GPA variance (R^2 = 0.0010). This suggests that time spent in extracurricular activities has no reliable linear impact on academic performance here; accordingly, we will also limit our reporting to the tabulated summary for this predictor without additional regression visuals.

Physical Activity per Day, in contrast, has = –0.1013 ($p < 0.001$, R^2 = 0.1164), indicating a modest but reliable inverse relationship in which each extra hour of exercise predicts about a tenth-point decrease in GPA. Though smaller in magnitude than the study-hours effect, this result warrants further graphical exploration and examination of potential nonlinearities .

Finally, Social Hours per Day produces = –0.0379 (p = 0.0001, R^2 = 0.0073), a small but statistically robust negative association; each additional hour of socializing corresponds to roughly a 0.04-point drop in GPA. Given its significance, we will include a fitted-line plot to assess whether the negative social-time effect holds uniformly across the stress spectrum or whether it reflects a subgroup pattern.

In summary, only the three predictors with $p < 0.05$ (Study Hours, Physical Activity, and Social Hours) will be subjected to further model inspection and visualization.



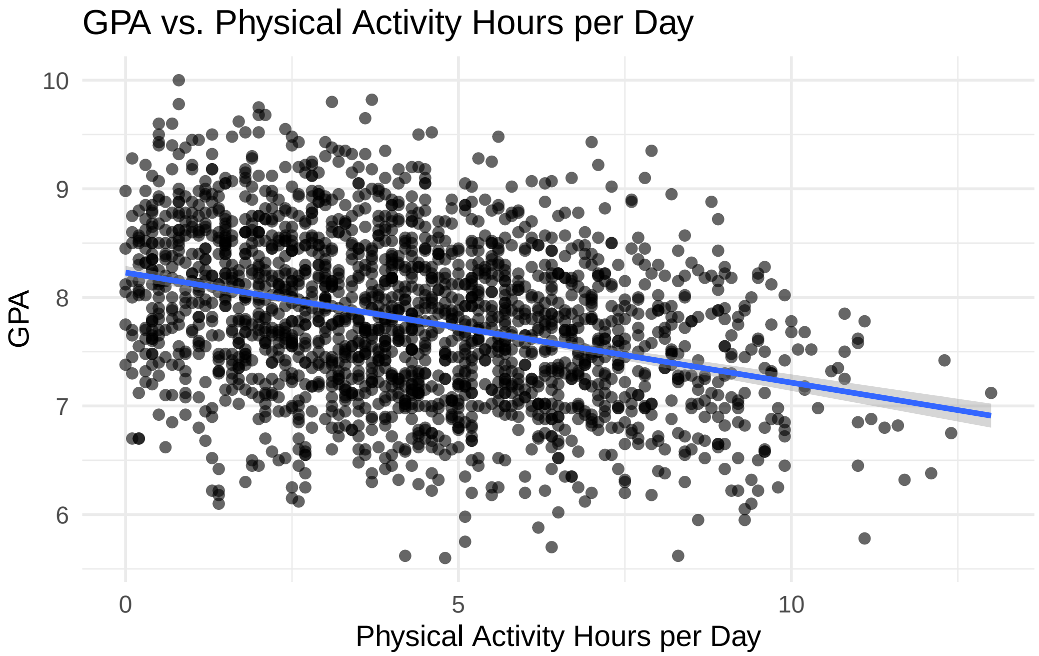


Figure 7: Linear Regression of GPA on Physical Activity Hours per Day

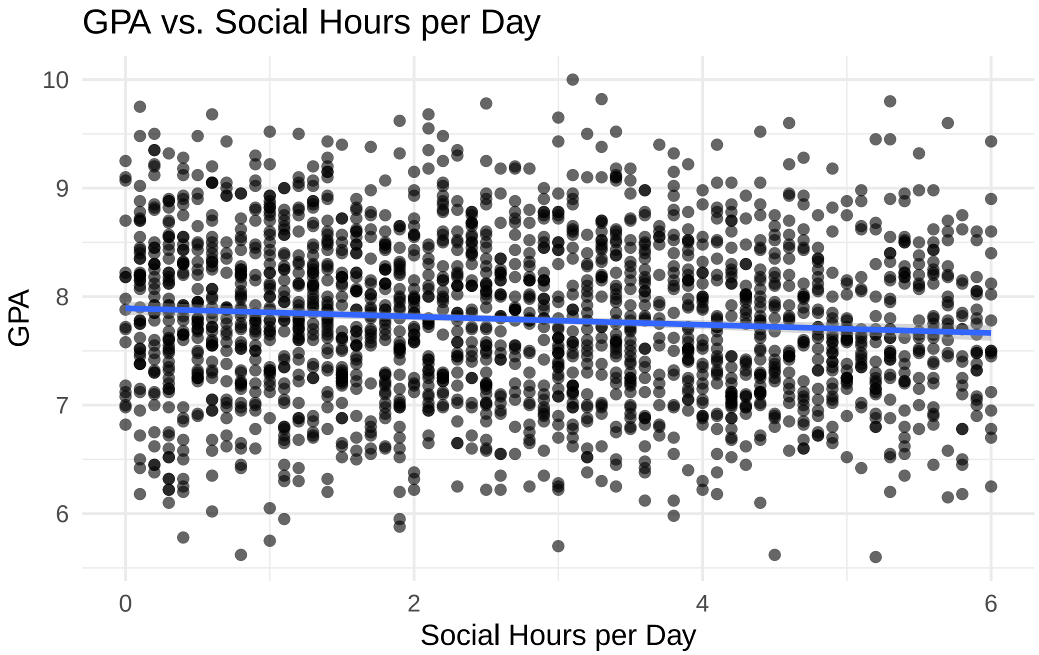


Figure 8: Linear Regression of GPA on Social Hours per Day

Supporting research

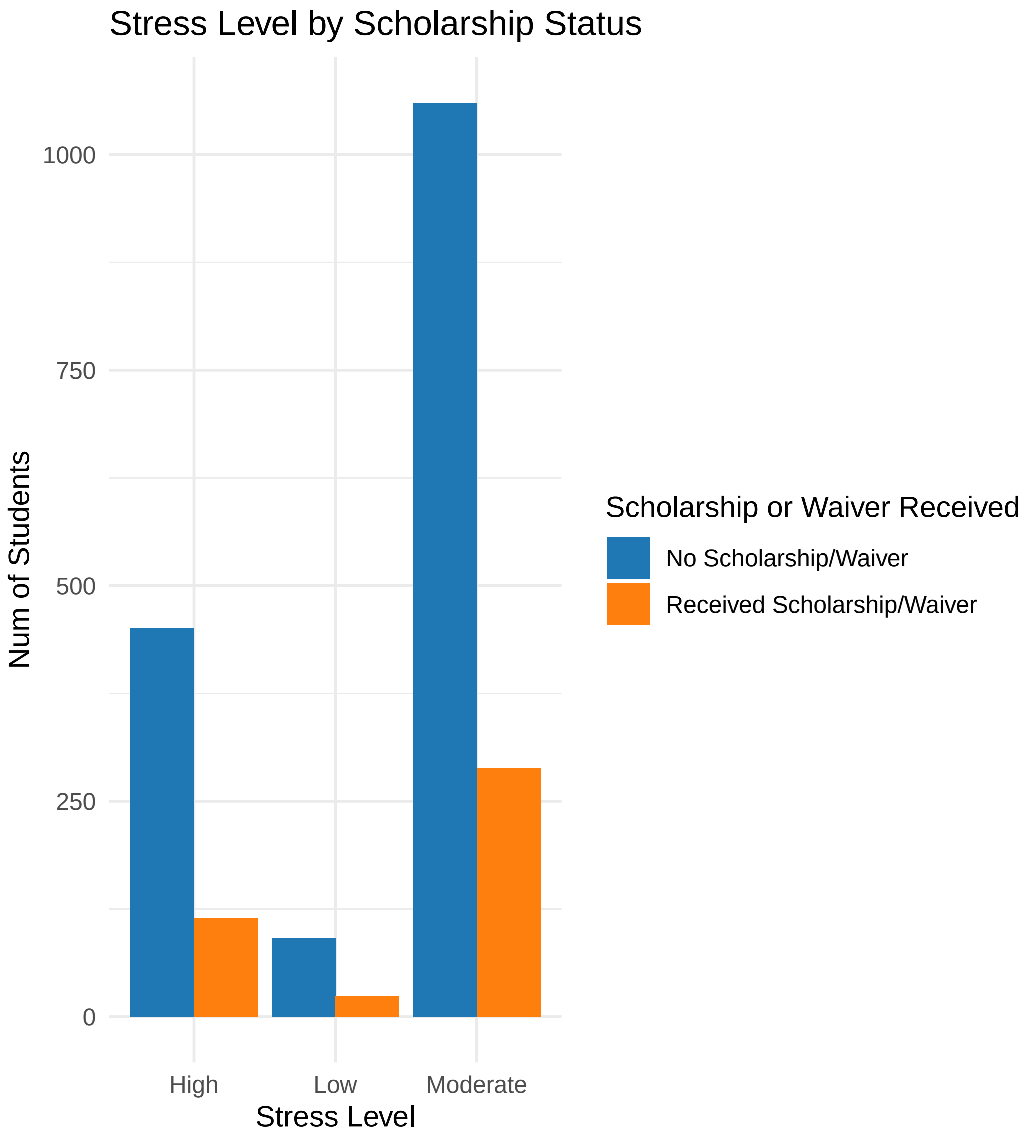


Figure 9: stress level by student scholarship status

(**scholarship-stress_plot?**) shows for all stress levels that stress levels do not have a direct relationship with weather a student has a scholarship or not. We were surprised to see that there was not as large of a positive relations as we thought. This plot leads us to believe that scholarship status/waiver status is not a primary factor in determining student stress level.

The following visualizations we will create two plots for each mental health disorder we analyse.

One visualization will include the data for students enrolled in the computer science/ computer engineering major and one will not include the data. The reason for this is due to the data set we used had a majority of students enrolled as that major and generating two visualizations allows us to better analyse and comment on the data.

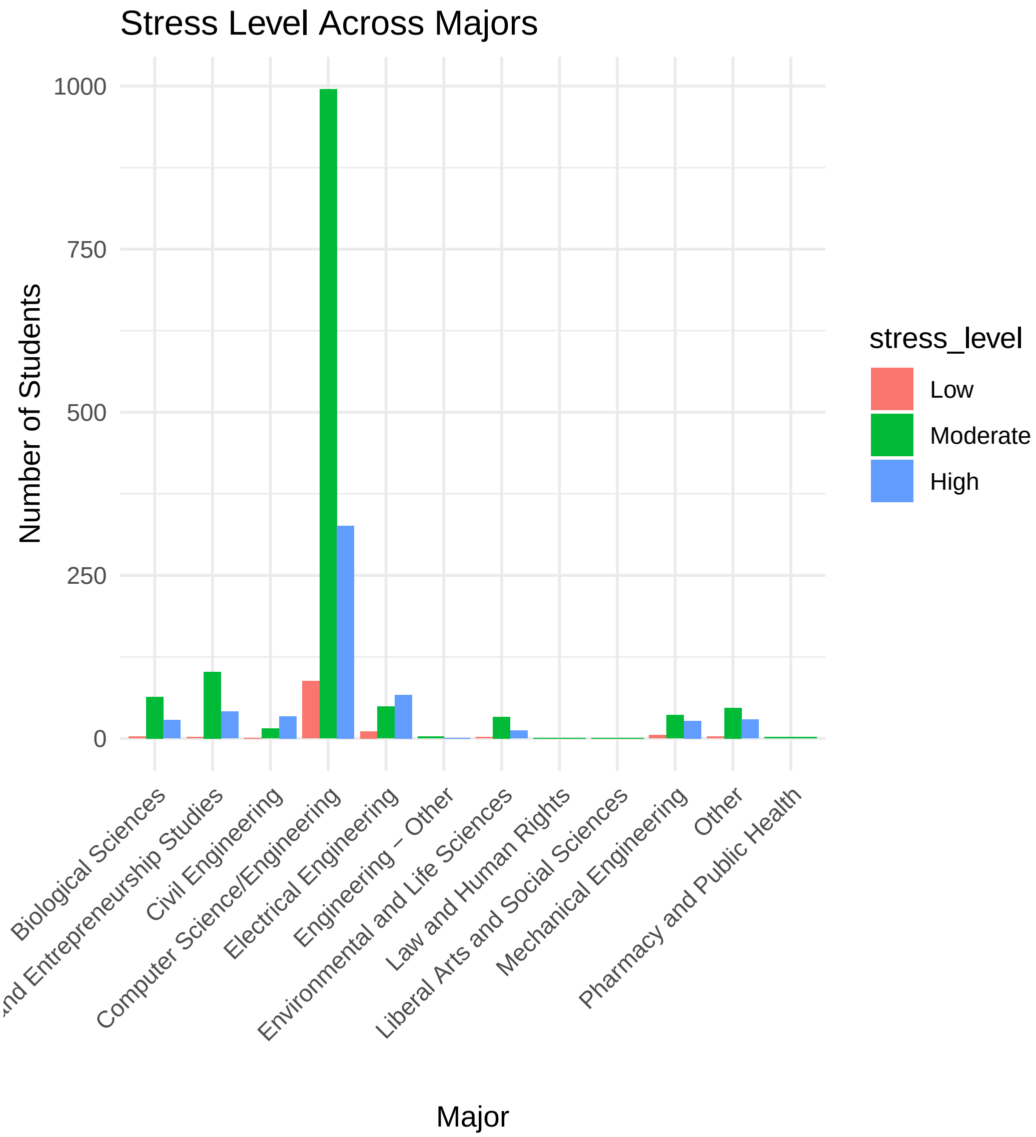


Figure 10: stress level by major

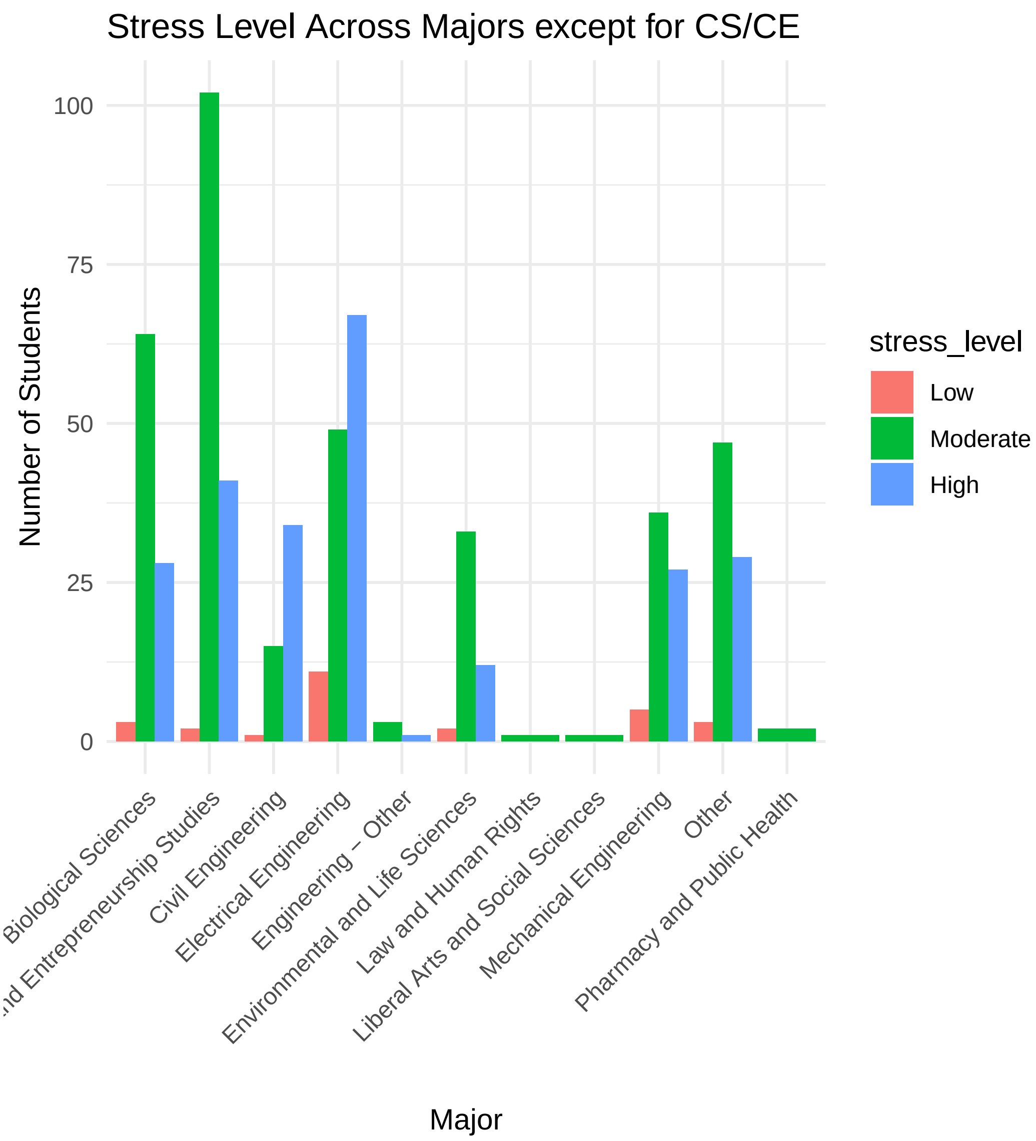


Figure 11: stress level by majors except for CS/CE

From (**major-stress_plot?**) and (**no-cs-stress_plot?**) we are able to learn how a students major effects there stress level. From the visualizations we are able to learn that students majoring in Civil or electrical engineering typically have higher levels of stress compared to all other majors. We are also able to see that most students perceive them selves as having moderate amounts of stress along most majors. This visualization made sense to us due to majors such as Electrical/Civil engineering are typacly seen as harder STEM major, thus causing more stress. It was interesting to see however other STEM majors such as mechanical engineering and Computer science/engineering deviate from the other engineering majors.

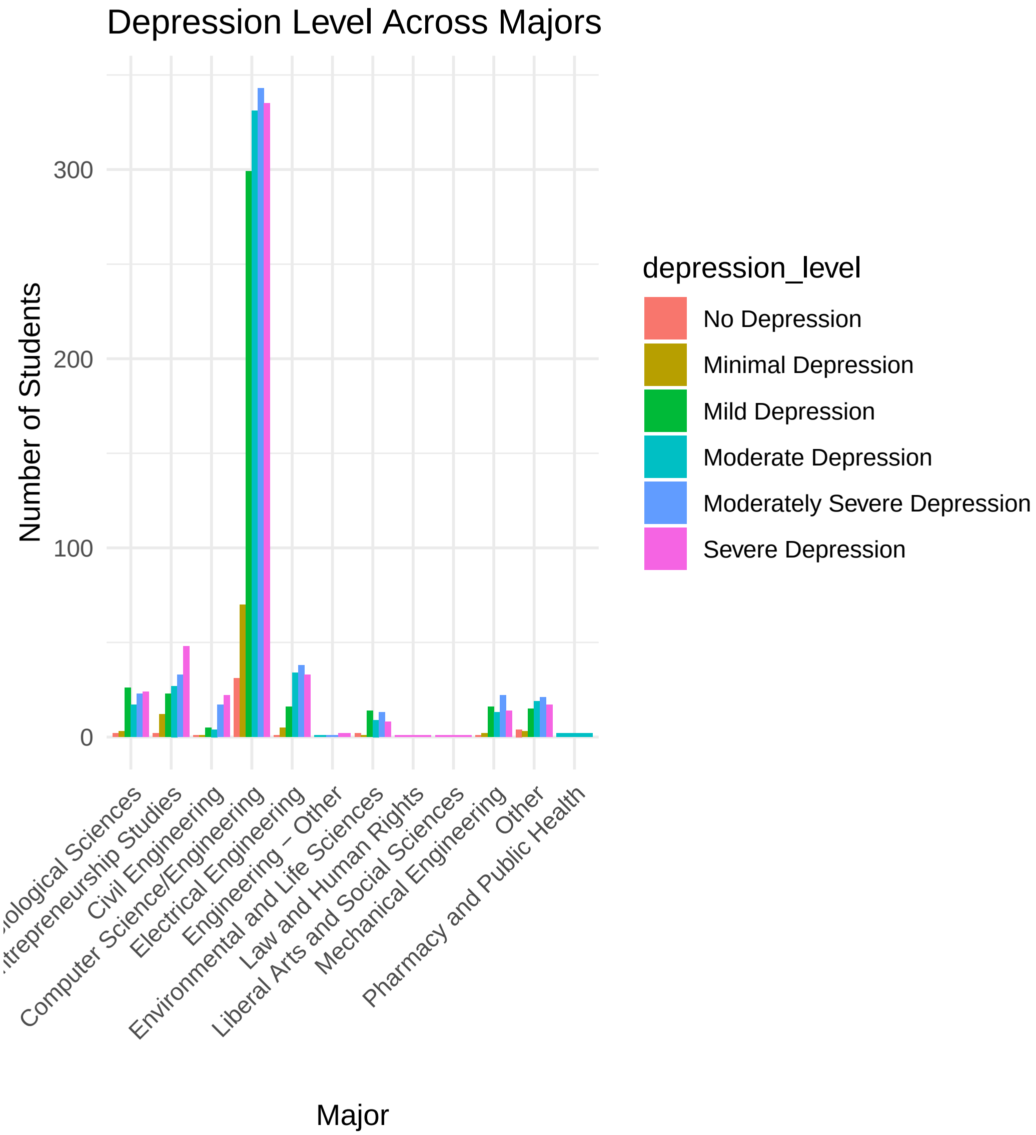


Figure 12: depression level by major

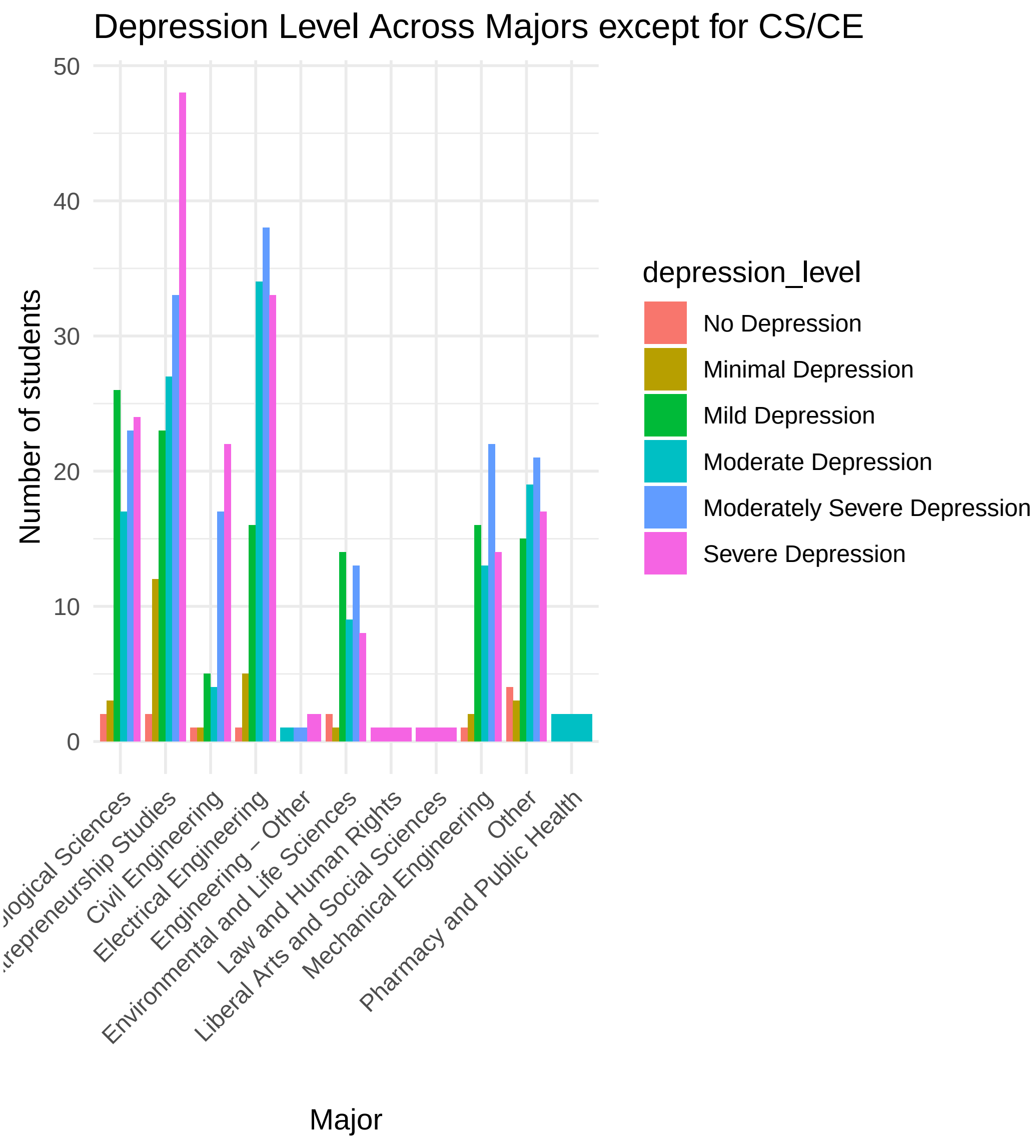


Figure 13: depression level by majors except for CS/CE

From (**major-depression_plot?**) and (**no-cs-depression_plot?**) we are able to learn how a students major effects there depression level. From these visualizations we were able to find out that students studding business and entrepreneurship studies had the highest levels of depression out of all the various majors. This surprised us due to us believing that all the various mental health would all have the similar distributions among the various majors.

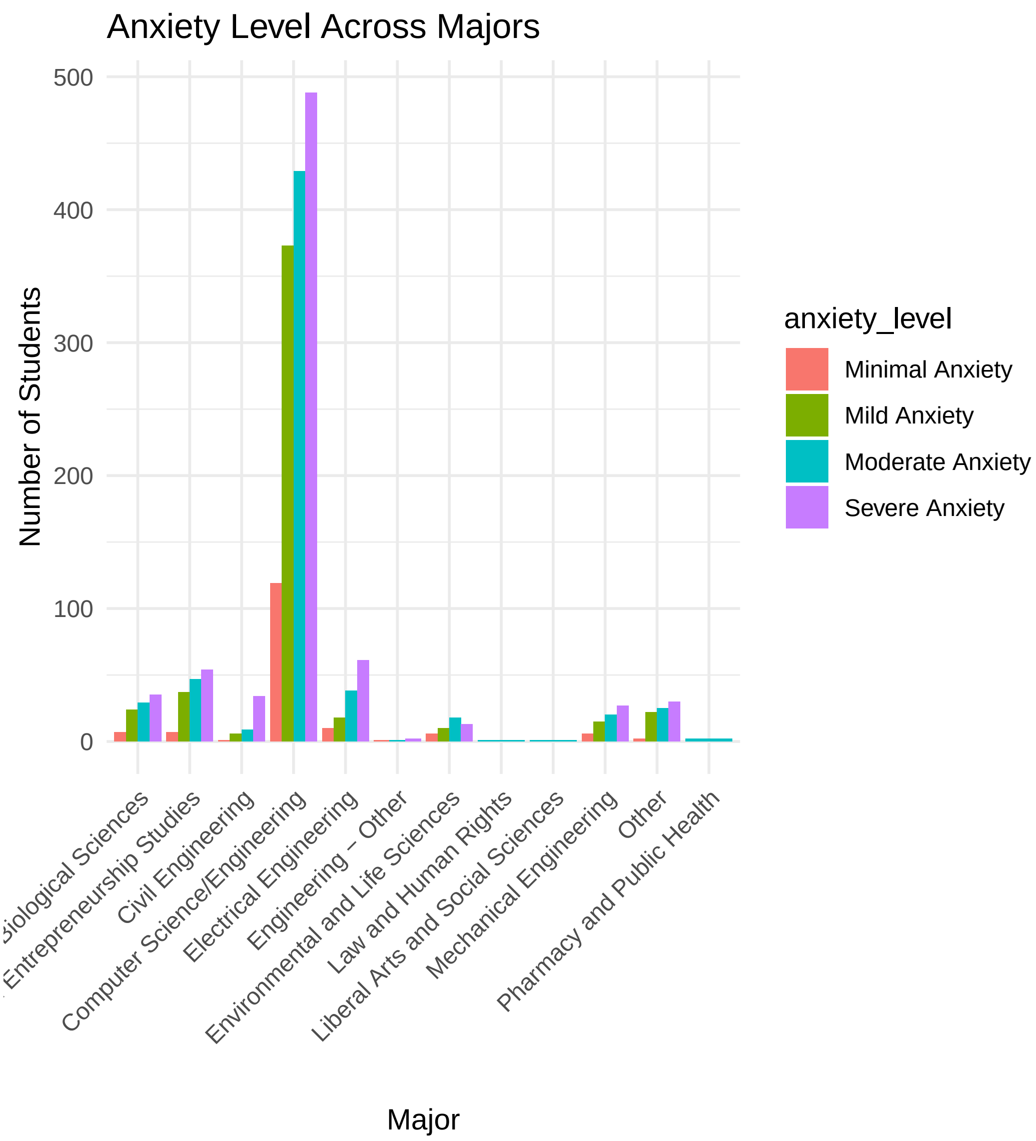


Figure 14: anxiety level by major

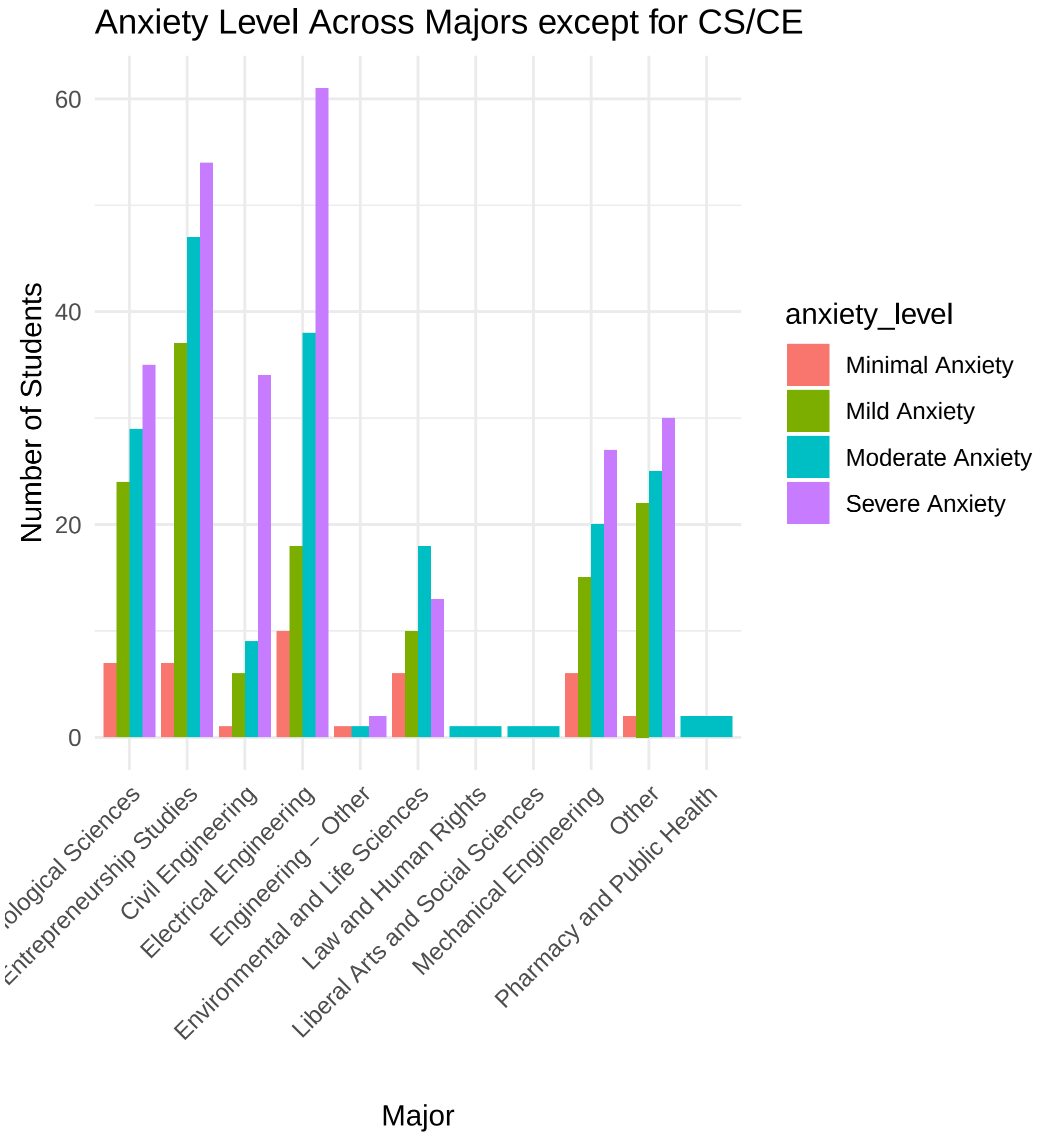


Figure 15: depression level by majors except for CS/CE

From (**major-anxiety_plot?**) and (**no-cs-anxiety_plot?**) we are able to learn how a students major effects there anxiety level. From this graphic we are able to see that for the majority of majors a plurality of students have severe anxiety. The two majors with the highest levels of severe anxiety were civil engineering and electrical engineering. This relation show there being a greater relation between a students stress and there anxiety compared to a students stress and depression or there anxiety and depression. Another interesting observation is that students studding Environmental and Life Sciences are the only group who has multiple levels of measured anxiety to have higher levels of moderate anxiety compared to severe anxiety. From this knowledge we potentially apply what students do for that major towards majors with higher levels of anxiety.

Conclusion

This research set out to examine how various student lifestyle factors—including study time, sleep duration, physical activity, social interaction, and extracurricular involvement—relate to both academic performance, as measured by GPA, and self-reported stress levels. The investigation was guided by specific research questions focused on identifying whether and how these behavioral variables predict significant differences in student outcomes. Additionally, the role of gender, academic major, and scholarship status was considered to provide a more comprehensive view of the student experience across demographic lines.

The main statistical analyses included a combination of one-way ANOVA tests for categorical group comparisons and simple linear regression for associations between continuous variables. These methods enabled rigorous testing of the hypotheses for each factor. Notably, the ANOVA test for GPA by stress level yielded a statistically significant result, leading to the rejection of the null hypothesis. Post-hoc Tukey HSD tests further demonstrated that students reporting high stress levels had significantly higher GPAs than their peers with moderate or low stress. These findings suggest that academic performance and psychological stress are not independent but may instead be intertwined in complex ways, potentially reflecting heightened academic effort accompanied by increased psychological burden.

Regression analysis provided additional clarity by quantifying the relationships between each lifestyle factor and GPA. Study hours per day stood out as the strongest predictor, with a statistically significant positive association and an R^2 of 0.5393, indicating that over 53% of GPA variance could be explained by this single variable. Physical activity and social hours showed weaker but still statistically significant negative associations with GPA, suggesting that increased time in these domains may modestly detract from academic outcomes. Conversely, sleep hours and extracurricular involvement were not found to be significantly associated with GPA, and their explanatory power was negligible. These results contribute to a more precise understanding of which aspects of student behavior are most relevant for academic achievement.

Further investigation of group-level differences in stress level across behavioral factors revealed additional patterns. Tukey HSD tests showed that both study hours and physical activity hours significantly differed across stress levels, particularly between the high-stress group and the other two. These differences were not uniformly linear but instead reflected threshold effects, wherein high stress levels were associated with disproportionately higher study time and reduced sleep and leisure. These results underscore the importance of contextualizing stress within behavioral trade-offs rather than assuming uniform trends across all categories.

Supporting data drawn from the secondary dataset provided additional insights into the broader academic environment. For instance, no significant relationship was found between scholarship status and stress level, challenging common assumptions about financial pressure. Furthermore, comparisons across academic majors revealed consistently elevated stress and anxiety among students in civil and electrical engineering, suggesting that discipline-specific demands may mediate the effects of general lifestyle factors. While these findings fall slightly outside the main scope of the paper, they align with and reinforce the broader patterns identified in the primary analysis.

In summary, this research demonstrates that while certain lifestyle factors—especially study time—have a direct and measurable impact on GPA, they also correlate with elevated stress levels. The coexistence of strong academic performance and psychological strain indicates that academic success may come at a psychological cost for some students. By employing both inferential statistics and targeted post-hoc analyses, this study provides evidence-based insights into the behavioral determinants of student outcomes. These results highlight the need for academic institutions to consider how behavioral expectations intersect with mental health, and to design policies and interventions that support both performance and well-being. ## Code Appendix

Citations

```
# Load necessary packages ----
library(ggplot2)
library(dplyr)
library(knitr)
library(tinytex)
library(tidyverse)
library(tibble)
library(broom)
library(purrr)
library(kableExtra)

# Google R style guide used
#Primary Dataset
student_raw <-  read.csv(
  "https://raw.githubusercontent.com/Stat184-Spring2025/Sec4_FP_JonasP_WillB_AlexH/refs/heads
  sep = ","
)
#Supplementary Dataset
mental_health_raw <- read.csv(
  "https://raw.githubusercontent.com/Stat184-Spring2025/Sec4_FP_JonasP_WillB_AlexH/refs/heads
  sep = ","
)

#Converts the provided 10 point GPA scale to a 4 point scale for the primary dataset
student_4 <- student_raw %>%
  mutate(Grades = Grades * (4/10))
#Converts the provided 10 point GPA scale to a 4 point scale for the primary data set
student_4 <- student_raw %>%
  mutate(Grades = Grades * (4/10))
sum(duplicated(student_4))
# Remove Student_ID column
students_scatter_matrix <- student_4 %>%
  subset(, select = -c(Student_ID))
# Define colors
my_cols <- c("Blue", "Red", "dark Green")

#Rename labels
students_scatter_matrix <- students_scatter_matrix %>%
  rename(
    Study = Study_Hours_Per_Day,
    Extracurricular = Extracurricular_Hours_Per_Day,
    Sleep = Sleep_Hours_Per_Day,
    Social = Social_Hours_Per_Day,
    Physical_Activity = Physical_Activity_Hours_Per_Day,
    GPA = Grades
  )

# Make Stress_Level a factor
students_scatter_matrix$Stress_Level <- as.factor(students_scatter_matrix$Stress_Level)

# Create a color map
color_map <- setNames(
  my_cols[
    1:length(
      levels(
  students_scatter_matrix$Stress_Level
  ))],
  levels(
    students_scatter_matrix$Stress_Level
    ))

# Plot Stress_Level as colors and without Gender
```

```r
plot(
  subset(
    students_scatter_matrix,
    select = -c(Gender, Stress_Level)
    ),
  pch = ".",
  col = color_map[students_scatter_matrix$Stress_Level],
)
# Remove Student_ID column
students_scatter_matrix <- student_4 %>%
  subset(, select = -c(Student_ID))
# Define colors
my_cols <- c("Red", "Blue")

#Rename labels
students_scatter_matrix <- students_scatter_matrix %>%
  rename(
    Study = Study_Hours_Per_Day,
    Extracurricular = Extracurricular_Hours_Per_Day,
    Sleep = Sleep_Hours_Per_Day,
    Social = Social_Hours_Per_Day,
    Physical_Activity = Physical_Activity_Hours_Per_Day,
    GPA = Grades
  )

# Make Stress_Level a factor
students_scatter_matrix$Gender <- as.factor(students_scatter_matrix$Gender)

# Create a color map
color_map <- setNames(
  my_cols[
    1:length(
      levels(
  students_scatter_matrix$Gender
  ))],
  levels(
    students_scatter_matrix$Gender
    )
  )

# Plot Gender as colors and without Stress_Level
plot(
  subset(
    students_scatter_matrix,
    select = -c(Gender, Stress_Level)
    ),
  pch = ".",
  col = color_map[students_scatter_matrix$Gender],
)
# Make summary table for Grades by Stress Level ----
student_stress_summary <- student_4 %>%
  select(Stress_Level, Grades) %>%
  group_by(Stress_Level) %>%
  summarize(
    count = n(),
    min = min(Grades),
    Q1 = quantile(Grades, 0.25),
    median = median(Grades),
    Q1 = quantile(Grades, 0.75),
    max = max(Grades),
    mad = mad(Grades),
    mean = mean(Grades),
    count = n(),
    min = min(Grades),
    Q1 = quantile(Grades, 0.25),
    median = median(Grades),
    Q3 = quantile(Grades, 0.75),
    max = max(Grades),
    mad = mad(Grades),
    mean = mean(Grades),
    sd = sd(Grades)
  )
student_stress_summary %>%
  knitr::kable()
#relevels stress level by low, moderate, high
student_4$Stress_Level <- fct_relevel(student_4$Stress_Level, "Low", "Moderate", "High")

#generates box and whisker plot
ggplot(
  student_4,
  aes(
    x = Stress_Level,
    y = Grades,
    fill = Stress_Level
```

```r
  )
) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Student GPAs by Stress level", x = "Student Stress Level", y = "Student GPAs
  theme_minimal()
df <- student_raw %>%
  rename(
    Grades = Grades,
    Stress_Level = Stress_Level,
    Study_Hours_Per_Day = Study_Hours_Per_Day,
    Sleep_Hours_Per_Day = Sleep_Hours_Per_Day,
    Physical_Activity_Hours_Per_Day = Physical_Activity_Hours_Per_Day,
    Social_Hours_Per_Day = Social_Hours_Per_Day,
    Extracurricular_Hours_Per_Day = Extracurricular_Hours_Per_Day
  ) %>%
  mutate(
    Stress_Level = factor(Stress_Level, c("Low", "Moderate", "High"))
  )

# Define models
anova_stress <- aov(Grades ~ Stress_Level, data = df)
anova_study  <- aov(Study_Hours_Per_Day ~ Stress_Level, data = df)
anova_sleep  <- aov(Sleep_Hours_Per_Day ~ Stress_Level, data = df)
anova_phys   <- aov(Physical_Activity_Hours_Per_Day ~ Stress_Level, data = df)
anova_soc    <- aov(Social_Hours_Per_Day ~ Stress_Level, data = df)
anova_extra  <- aov(Extracurricular_Hours_Per_Day ~ Stress_Level, data = df)

lm_study     <- lm(Grades ~ Study_Hours_Per_Day, data = df)
lm_sleep     <- lm(Grades ~ Sleep_Hours_Per_Day, data = df)
lm_phys      <- lm(Grades ~ Physical_Activity_Hours_Per_Day, data = df)
lm_soc       <- lm(Grades ~ Social_Hours_Per_Day, data = df)
lm_extra     <- lm(Grades ~ Extracurricular_Hours_Per_Day, data = df)

alpha <- 0.05

# Build summary table
summary_tbl <- tibble(
  Analysis = c(
    "GPA vs Stress",
    "Study Hours vs Stress",
    "Sleep Hours vs Stress",
    "Physical Activity vs Stress",
    "Social Hours vs Stress",
    "Extracurricular vs Stress",
    "Study Hours vs GPA",
    "Sleep Hours vs GPA",
    "Physical Activity vs GPA",
    "Social Hours vs GPA",
    "Extracurricular vs GPA"
  ),

  Types = c(
    rep("Categorical vs Quantitative", 6),
    rep("Quantitative vs Quantitative", 5)
  ),
  Test = c(
    rep("ANOVA", 6),
    rep("Linear regression", 5)
  ),
  Pvalue = c(
    broom::tidy(anova_stress)$p.value[1],
    broom::tidy(anova_study)$p.value[1],
    broom::tidy(anova_sleep)$p.value[1],
    broom::tidy(anova_phys)$p.value[1],
    broom::tidy(anova_soc)$p.value[1],
    broom::tidy(anova_extra)$p.value[1],
    broom::glance(lm_study)$p.value,
    broom::glance(lm_sleep)$p.value,
    broom::glance(lm_phys)$p.value,
    broom::glance(lm_soc)$p.value,
    broom::glance(lm_extra)$p.value
  )
) %>%
  mutate(
    H0 = if_else(Pvalue < alpha, "Reject", "Do not Reject"),
    Investigation = if_else(
      H0 == "Reject",
        if_else(grepl("ANOVA", Test),
          "Conduct Tukey HSD",
          "Examine regression coefficients"),
      "None"
    )
  )
```

```
# Output table
knitr::kable(
  summary_tbl,
  caption = "Summary of preliminary statistical tests",
  digits = 4
)
# Mapping of display labels to quantitative outcomes
mapping <- list(
  "GPA"                         = "Grades",
  "Study Hours"         = "Study_Hours_Per_Day",
  "Sleep Hours"         = "Sleep_Hours_Per_Day",
  "Physical Activity"   = "Physical_Activity_Hours_Per_Day",
  "Social Hours"        = "Social_Hours_Per_Day",
  "Extracurricular" = "Extracurricular_Hours_Per_Day"
)

# Compute Tukey HSD for Stress_Level → each quantitative variable
tukey_all <- purrr::imap_dfr(
  mapping,
  function(col_name, display_label) {
    # 1) Fit ANOVA: outcome ~ Stress_Level
    formula <- as.formula(paste(col_name, "~ Stress_Level"))
    aov_mod <- aov(formula, data = df)

    # 2) Extract Tukey HSD for Stress_Level
    tukey_res <- TukeyHSD(aov_mod, "Stress_Level")$Stress_Level

    # 3) Tidy and flag conclusiveness
    as.data.frame(tukey_res) %>%
      rownames_to_column("Comparison") %>%
      rename(
        Mean_Diff = diff,
        Lower_CI   = lwr,
        Upper_CI   = upr,
        P      = `p adj`
      ) %>%
      mutate(
        Variable   = display_label,
        Conclusive = if_else(P < 0.05, "Yes", "No")
      ) %>%
      select(Variable, Comparison, Mean_Diff, Lower_CI, Upper_CI, P, Conclusive)
  }
```

```
)

# 4) Output a single formatted table
tukey_all %>%
  kable(
    caption = "Combined Tukey HSD Pairwise Comparisons for Stress Level vs. Quantitative Out
    digits  = 4
  ) %>%
  kable_styling(position = "left")
# Fit ANOVA and extract Tukey HSD for GPA
df <- student_raw %>%
  rename(
    Grades                         = Grades,
    Stress_Level                   = Stress_Level
  ) %>%
  mutate(
    Stress_Level = factor(Stress_Level, c("Low", "Moderate", "High"))
  )

# 1) ANOVA and Tukey
aov_gpa      <- aov(Grades ~ Stress_Level, data = df)
tukey_gpa_df <- as.data.frame(TukeyHSD(aov_gpa, "Stress_Level")$Stress_Level) %>%
  rownames_to_column("Comparison") %>%
  rename(
    mean_diff = diff,
    lower_ci  = lwr,
    upper_ci  = upr,
    adj_p     = `p adj`
  )

# 2) Plot
ggplot(tukey_gpa_df, aes(x = mean_diff, y = reorder(Comparison, mean_diff))) +
  geom_point(size = 2) +
  geom_errorbarh(aes(xmin = lower_ci, xmax = upper_ci), height = 0.2) +
  labs(
    x = "Mean Difference in GPA",
    y = NULL
  ) +
  theme_minimal()
# 0) Prep libraries (df already defined above)
library(dplyr); library(ggplot2); library(tibble)
```

```r
# 1) ANOVA and Tukey
aov_study    <- aov(Study_Hours_Per_Day ~ Stress_Level, data = df)
tukey_study_df <- as.data.frame(TukeyHSD(aov_study, "Stress_Level")$Stress_Level) %>%
  rownames_to_column("Comparison") %>%
  rename(
    mean_diff = diff,
    lower_ci  = lwr,
    upper_ci  = upr,
    adj_p     = `p adj`
  )

# 2) Plot
ggplot(tukey_study_df, aes(x = mean_diff, y = reorder(Comparison, mean_diff))) +
  geom_point(size = 2) +
  geom_errorbarh(aes(xmin = lower_ci, xmax = upper_ci), height = 0.2) +
  labs(
    x = "Mean Difference in Study Hours",
    y = NULL
  ) +
  theme_minimal()
# 0) Prep libraries (df already defined above)
library(dplyr); library(ggplot2); library(tibble)

# 1) ANOVA and Tukey
aov_phys     <- aov(Physical_Activity_Hours_Per_Day ~ Stress_Level, data = df)
tukey_phys_df <- as.data.frame(TukeyHSD(aov_phys, "Stress_Level")$Stress_Level) %>%
  rownames_to_column("Comparison") %>%
  rename(
    mean_diff = diff,
    lower_ci  = lwr,
    upper_ci  = upr,
    adj_p     = `p adj`
  )

# 2) Plot
ggplot(tukey_phys_df, aes(x = mean_diff, y = reorder(Comparison, mean_diff))) +
  geom_point(size = 2) +
  geom_errorbarh(aes(xmin = lower_ci, xmax = upper_ci), height = 0.2) +
  labs(
    x = "Mean Difference in Activity Hours",
    y = NULL
  ) +
  theme_minimal()
# Mapping of display labels to quantitative predictors
tmp_mapping <- list(
  "Study Hours per Day"       = "Study_Hours_Per_Day",
  "Sleep Hours per Day"       = "Sleep_Hours_Per_Day",
  "Physical Activity per Day" = "Physical_Activity_Hours_Per_Day",
  "Social Hours per Day"      = "Social_Hours_Per_Day",
  "Extracurricular Hours per Day" = "Extracurricular_Hours_Per_Day"
)

linreg_details <- purrr::imap_dfr(
  tmp_mapping,
  function(var_name, display_label) {
    model <- lm(as.formula(paste("Grades ~", var_name)), data = df)
    gl    <- broom::glance(model)
    td    <- broom::tidy(model) %>%
      filter(term == var_name) %>%
      select(estimate, p.value)

    tibble(
      Predictor       = display_label,
      Estimate        = td$estimate,
      `P value`       = gl$p.value,
      `R squared`     = gl$r.squared,
      Conclusive      = if_else(gl$p.value < 0.05, "Yes", "No")
    )
  }
)

linreg_details %>%
  kable(
    digits    = 4,
    caption   = "Detailed Linear Regression Results for Quantitative Predictors"
  ) %>%
  kable_styling(
    full_width = FALSE,
    position   = "left"
  )

library(ggplot2)

ggplot(df, aes(x = Study_Hours_Per_Day, y = Grades)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    x     = "Study Hours per Day",
    y     = "GPA",
    title = "GPA vs. Study Hours per Day"
  ) +
  theme_minimal()
library(ggplot2)

ggplot(df, aes(x = Physical_Activity_Hours_Per_Day, y = Grades)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    x     = "Physical Activity Hours per Day",
    y     = "GPA",
    title = "GPA vs. Physical Activity Hours per Day"
  ) +
  theme_minimal()
library(ggplot2)

ggplot(df, aes(x = Social_Hours_Per_Day, y = Grades)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(
    x     = "Social Hours per Day",
    y     = "GPA",
    title = "GPA vs. Social Hours per Day"
  ) +
  theme_minimal()
#Cleans/wrangles supplementary dataset by selecting useful columns and renaiming/mutating cas
mental_health_clean <- mental_health_raw %>%
  select(1, 2, 3, 4, 7, 19, 28, 39) %>%
  rename(
    major = Department,
    stress_level = Stress.Label,
    anxiety_level = Anxiety.Label,
    depression_level = Depression.Label
  ) %>%
  mutate(
    stress_level = case_when(
      stress_level == "High Perceived Stress" ~ "High",
      stress_level == "Moderate Stress" ~ "Moderate",
      stress_level == "Low Stress" ~ "Low",
      TRUE ~ stress_level
    )
  ) %>%
  mutate(
    major = case_when(
      str_detect(major, "Similar to CS") ~ "Computer Science/Engineering",
      str_detect(major, "Similar to EEE") ~ "Electrical Engineering",
      str_detect(major, "Similar to ME") ~ "Mechanical Engineering",
      str_detect(major, "Similar to CE") ~ "Civil Engineering",
      TRUE ~ major
    )
  )

#generates a bar chart
ggplot(mental_health_clean, aes(x = stress_level, fill = as.factor(waiver_or_scholarship))) +
  geom_bar(position = "dodge") +
  labs(
    title = "Stress Level by Scholarship Status",
    x = "Stress Level",
    y = "Num of Students",
    fill = "Scholarship or Waiver Received"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("#1f77b4", "#ff7f0e"),
                    labels = c("No Scholarship/Waiver", "Received Scholarship/Waiver"))
student_no_CS <- mental_health_clean %>%
  filter(major != "Computer Science/Engineering")
#re levels stress levels so that they are in order
mental_health_clean$stress_level <- fct_relevel(mental_health_clean$stress_level, "Low", "Mod

#generates  bar plot
ggplot(mental_health_clean, aes(x = major, fill = stress_level)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Stress Level Across Majors",
    x = "Major",
    y = "Number of Students"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
#re levels stress levels so that they are in order
student_no_CS$stress_level <- fct_relevel(student_no_CS$stress_level, "Low", "Moderate", "Hig

#generates bar plot
ggplot(student_no_CS, aes(x = major, fill = stress_level)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Stress Level Across Majors except for CS/CE",
    x = "Major",
    y = "Number of Students"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#re levels depression levels so that they are in order
mental_health_clean$depression_level <- fct_relevel(mental_health_clean$depression_level, "N

#generates bar plot
ggplot(mental_health_clean, aes(x = major, fill = depression_level)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Depression Level Across Majors",
    x = "Major",
    y = "Number of Students"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

#re levels depression levels so that they are in order
student_no_CS$depression_level <- fct_relevel(student_no_CS$depression_level, "No Depression

#generates bar plot
ggplot(student_no_CS, aes(x = major, fill = depression_level)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Depression Level Across Majors except for CS/CE",
    x = "Major",
    y = "Number of students"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#re levels anxiety levels so that they are in order
mental_health_clean$anxiety_level <- fct_relevel(mental_health_clean$anxiety_level, "Minimal
#generates bar plot
ggplot(mental_health_clean, aes(x = major, fill = anxiety_level)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Anxiety Level Across Majors",
    x = "Major",
    y = "Number of Students"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
#re levels anxiety levels so that they are in order
student_no_CS$anxiety_level <- fct_relevel(student_no_CS$anxiety_level, "Minimal Anxiety", "

#generates bar plot
ggplot(student_no_CS, aes(x = major, fill = anxiety_level)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Anxiety Level Across Majors except for CS/CE",
    x = "Major",
    y = "Number of Students"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Bennett, Charlotte. “Lifestyle Factors and Their Impact on Students.” *Kaggle*, Apr. 2025, https://www.kaggle.com/datasets/charlottebennett1234/lifestyle-factors-and-their-impact-on-students/data.

Syeed, Mahbubul. *MHP (Anxiety, Stress, Depression) Dataset of University Students*. 2024, https://figshare.com/articles/dataset/MHP_Anxiety_Stress_Depression_Dataset_of_University_Students/25771164?file=46172346.