CSE303: Statistics for Data Science [SUMMER 2025]

Term Project Report

Submitted by:

Student ID	Student Name	Contribution Percentage
2022-3-60-045	Md. Saiful Islam	**30%**
2022-3-60-317	Umme Mukaddisa	**30%**
2023-3-60-576	Ahmed Khan	**23%**
2022–1-60-122	Seendid Saleh Kabir	**17%**

Table of Contents

1. Introduction

University students often experience academic pressure due to demanding coursework, strict deadlines, part-time jobs & extracurricular commitments. These factors can significantly influence their mental health, leading to stress, anxiety, poor sleep quality & reduced academic performance. This project aims to investigate the relationship between workload & mental health indicators through a structured, data-driven approach.

A survey was designed & distributed among students to collect information on study hours, course load, extracurricular involvement, employment status & demographic factors, along with self-reported measures of stress, anxiety & sleep quality. After preprocessing & cleaning the responses, statistical techniques such as exploratory data analysis, hypothesis testing & regression modeling were applied to identify significant patterns & relationships. Both simple & multiple regression models were used to evaluate how various workload & lifestyle factors affect mental well-being.

Beyond statistical analysis, the project emphasizes interpretation of results in terms of practical significance, focusing on predictors of student stress & anxiety, model accuracy & potential areas of concern. An interactive dashboard was also developed to visualize findings dynamically, allowing users to filter results by variables such as year of study, gender or job status.

Ultimately, this study not only enhances students' analytical & research skills but also provides valuable insights for universities to design supportive policies & resources aimed at improving student well-being & academic outcomes.

2. Data Pre-processing

Proper data preprocessing was performed to make the dataset clean, consistent, and ready for regression analysis. The following steps were applied:

2.1. Data Cleaning

- The Timestamp column was removed as it was not relevant for analysis.
- Column names were standardized to concise & descriptive formats (e.g., study_hours_per_week, job_hours_per_week).

2.2. Encoding Categorical Variables

- Year of Study was mapped to numeric values (1 = 1st year, …, 4 = 4th year).
- Binary variables such as Extracurricular Participation & Part-time Job were encoded as Yes = 1 & No = 0.

2.3. Data Type Conversion

- Numerical variables (courses enrolled, study hours, job hours, extracurricular hours, sleep hours) were converted into integer format for consistency.

2.4. Handling Missing Values

- Missing CGPA values were replaced with the column mean.
- Missing hours for extracurricular & job activities were imputed with the median if participation = Yes, & set to zero if participation = No.

2.5. Outlier Treatment

- Unrealistic CGPA values below 1 were replaced with the median.
- Courses enrolled with values greater than 6 or less than 2 were set to the median.
- Weekly study hours or extracurricular hours above 50 were set to the median.

Through these steps, the dataset was cleaned, structured & standardized, making it suitable for exploratory data analysis, hypothesis testing & regression modeling.

3. Dataset Characteristics & Exploratory Data Analysis

3.1. Dataset Overview

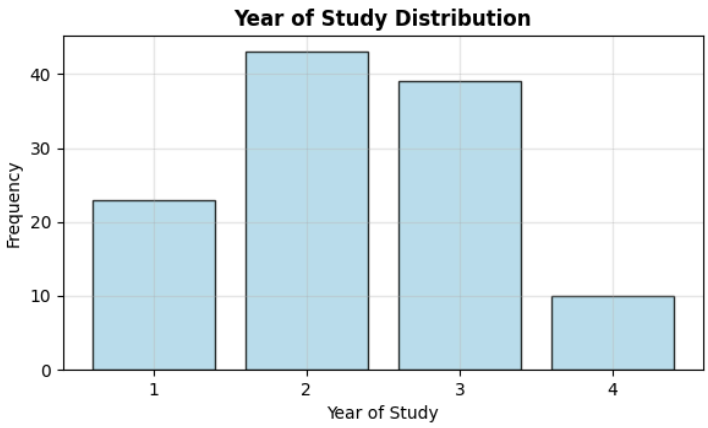
The dataset contains 115 records of student respondents with 14 attributes covering academic, extracurricular & psychological dimensions. This data provides insights into student life patterns & their potential relationship with stress, anxiety & sleep quality.

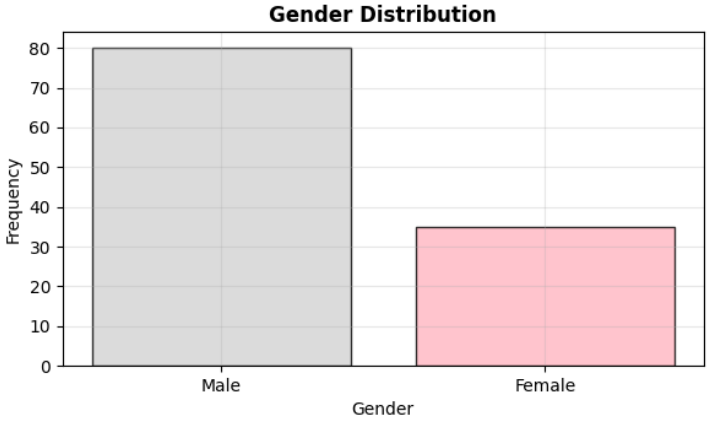
Key Dimensions:

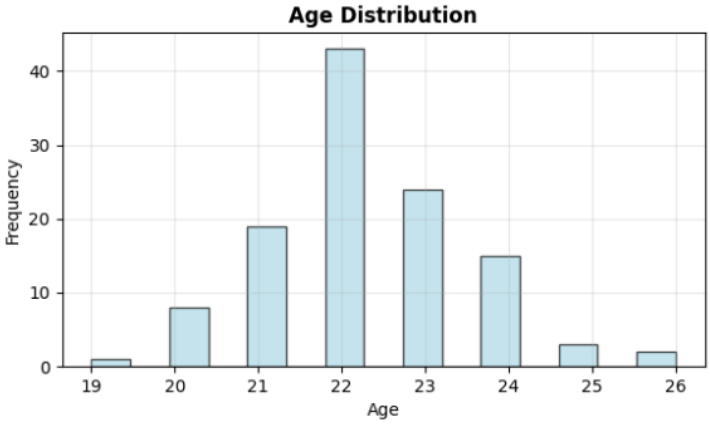
- Academic Profile: Year of study, CGPA, study hours, courses enrolled
- Extracurricular Activities: Participation & time commitment
- Employment Status: Part-time job involvement & hours worked
- Psychological Metrics: Self-reported stress & anxiety levels (1-5 scale)
- Sleep Patterns: nightly sleep duration & quality rating (1-10 scale)

3.2. Data Distribution Analysis

3.2.1. Demographic Characteristics







Year of Study Distribution:

- 1st Year: 18 students (15.7%)
- 2nd Year: 43 students (37.4%)
- 3rd Year: 38 students (33.0%)
- 4th Year: 16 students (13.9%)

The sample predominantly represents mid-career students (2nd and 3rd years), providing strong insights into the typical university experience.

Gender Distribution:

- Male: 80 students (69.6%)
- Female: 35 students (30.4%)

The gender distribution shows a male majority, which should be considered when generalizing findings.

Age Distribution:

- Range: 19-26 years
- Mean: 22.3 years
- Mode: 22 years (most common age)

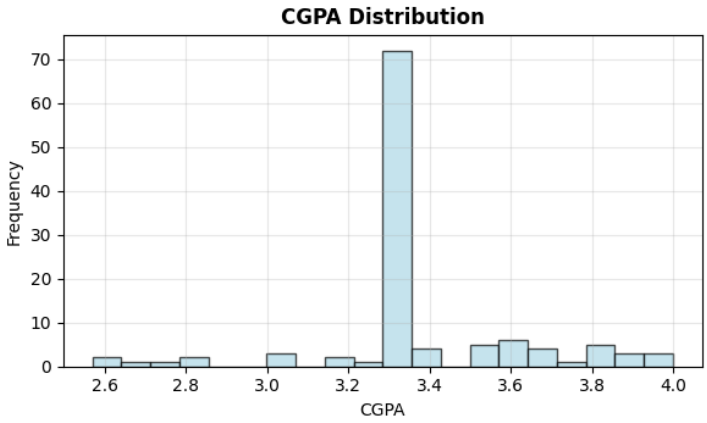
The age distribution shows a typical university student population, clustered around 21-23 years.

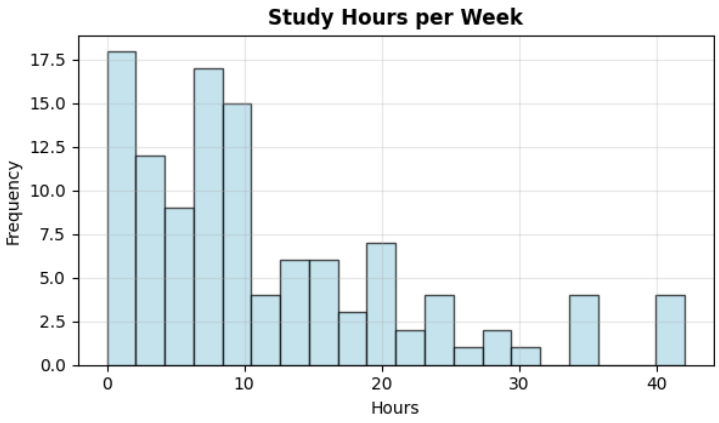
Sleep Hours:

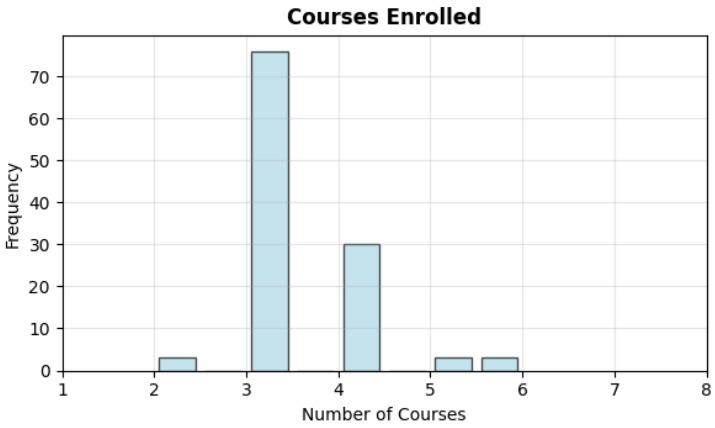
- Mean: 5.6 hours per night
- Range: 3 to 10 hours
- Distribution: Approximately normal with a slight right skew, centered around 5-6 hours per night.

The average sleep duration falls notably below the recommended 7-9 hours for adults, indicating a prevalence of sleep deprivation among the student population.

3.2.2. Academic Metrics Distribution







CGPA Distribution:

- Mean: 3.37 (B+ range)
- Range: 2.57-4.0
- Distribution: Approximately normal with slight left skew

The CGPA distribution suggests a generally high-achieving sample, though with some variability.

Study Hours Per Week:

- Mean: 12 hours
- Range: 0-42 hours
- Distribution: Right-skewed, with most students studying 4-17 hours weekly

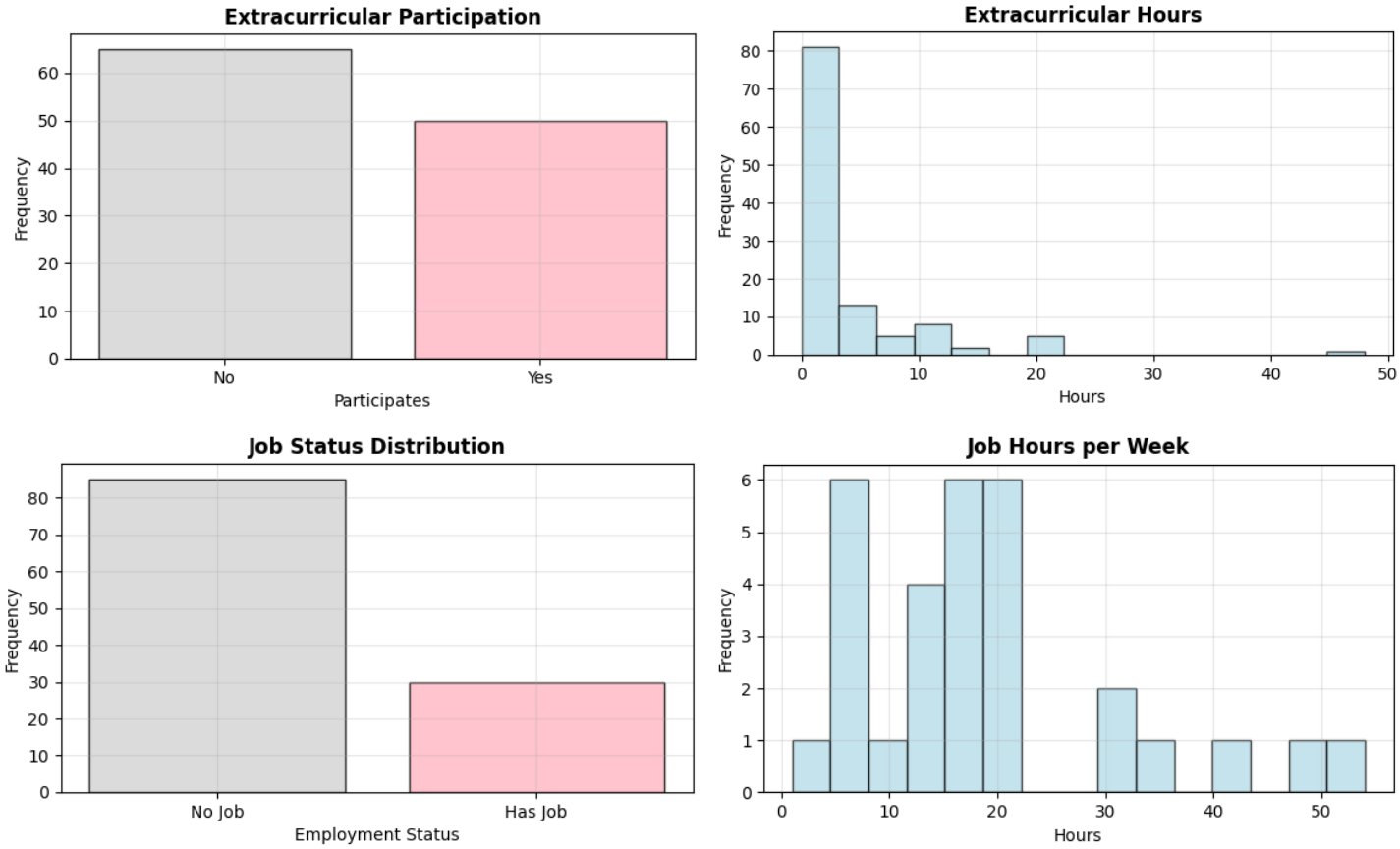
Notable variability exists in study habits, with some extreme values suggesting possible measurement errors or exceptional cases.

Courses Enrolled:

- Mean: 3.4 courses per semester
- Range: 2 to 6 courses
- Distribution: Strongly peaked and left-skewed, with a clear concentration of students taking 3 or 4 courses. This represents a standardized, full-time academic load for the cohort.

The course enrollment distribution indicates a highly standardized academic workload across the sample, with minimal variability in course load.

3.2.3. Extracurricular activities & Job metrics



Extracurricular Participation:

- Mean: 0.43 (on a 0-1 scale, where 1=Yes)
- Range: 0 to 1
- Distribution: Binary and left-skewed, showing that a majority of students (57%) do not participate in extracurricular activities.

The low participation rate suggests extracurriculars are not a primary commitment for most students in this sample.

Extracurricular Hours:

- Mean: ~3.5 hours per week
- Range: 0 to 48 hours
- Distribution: Heavily right-skewed, with most participants spending ≤10 hours weekly and a long tail of high-commitment outliers.

The highly right-skewed distribution reveals that while most participants have modest commitments, a small subset of students dedicate a significant amount of time to extracurricular activities.

Job Status:

- Mean: 0.26 (on a 0-1 scale, where 1=Yes)
- Range: 0 to 1
- Distribution: Binary and heavily left-skewed, indicating a large majority of students (74%) do not have a part-time job.

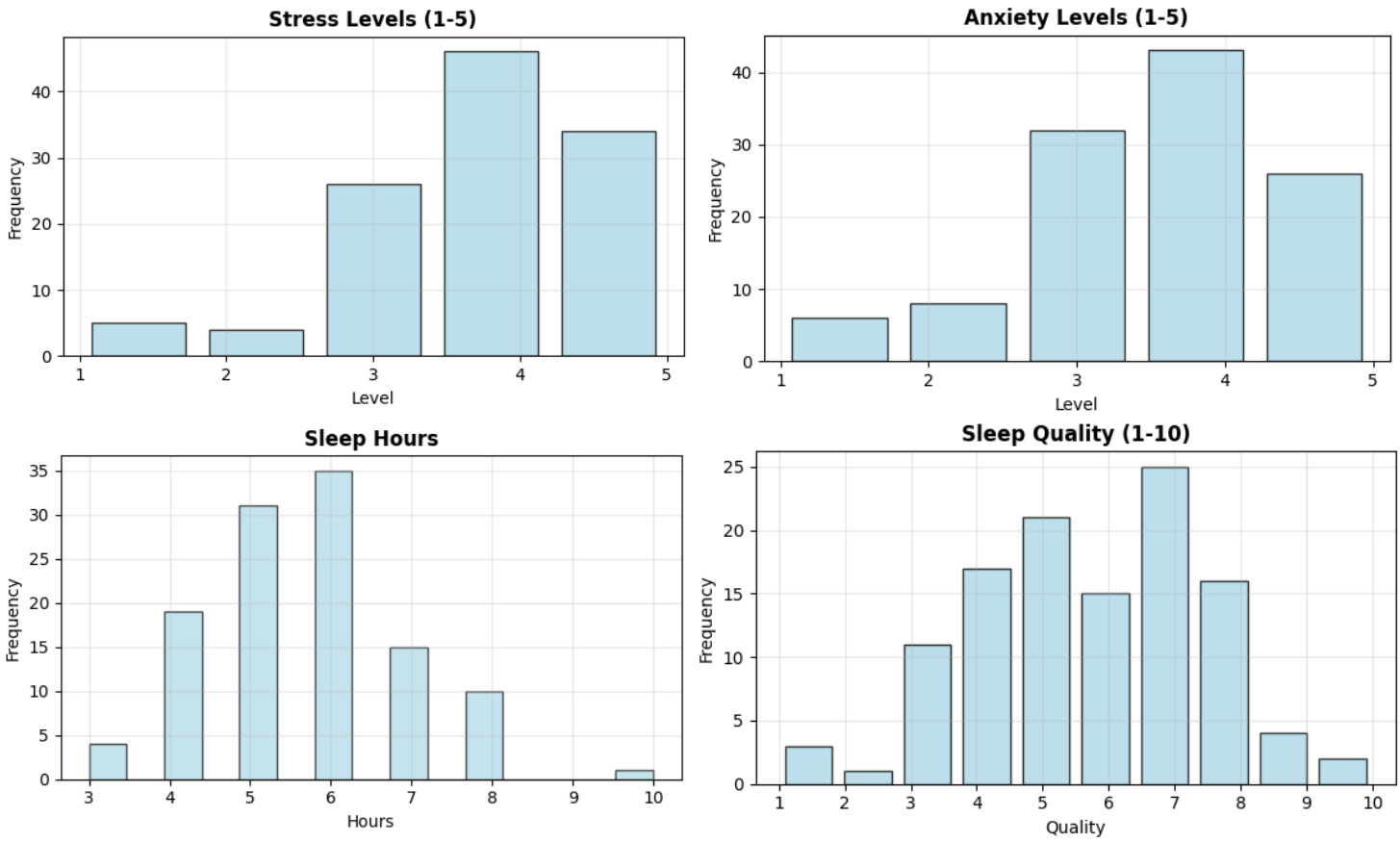
The low rate of employment suggests academic pursuits are the primary focus for most students in this cohort.

Job Hours Per Week:

- Mean: ~4.9 hours per week
- Range: 0 to 54 hours
- Distribution: Extremely right-skewed, characterized by a concentration at 0 hours and a wide, sparse distribution of hours among employed students.

The extreme skew highlights that having a job is uncommon & for those who work, weekly commitments vary dramatically from a few hours to what constitutes a full-time role.

3.2.4. Psychological & Sleep Metrics



Stress Levels (1-5 scale):

- Mean: 3.87
- Distribution: Left-skewed toward higher stress levels
- 55% of students reported stress levels of 4 or 5

Anxiety Levels (1-5 scale):

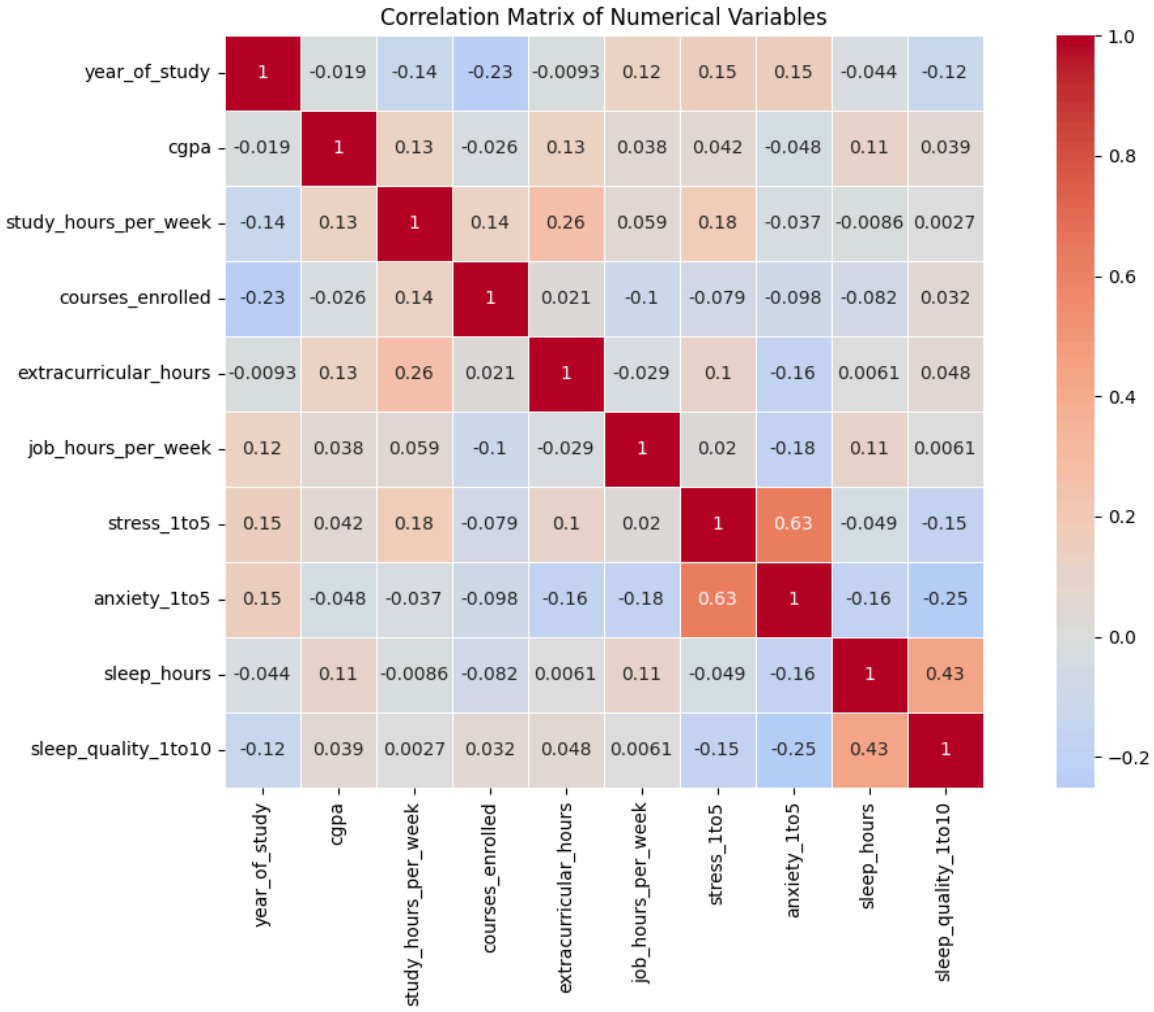
- Mean: 3.65
- Similar distribution to stress levels
- High correlation with stress levels (r = 0.82)

Sleep Quality (1-10 scale):

- Mean: 5.74
- Bimodal distribution with peaks at 4-5 and 6-7
- 68% of students rate their sleep quality as ≤7

3.3. Correlation Analysis

3.3.1. Correlation Heatmap



Strong Positive Correlations (r > 0.5):

- Stress vs. Anxiety (r = 0.63) - Expected strong relationship
- Extracurricular hours vs. Extracurricular participation (r = 0.61) - Logical validation
- Job hours vs. Job participation (r = 0.79) - Expected strong relationship

Notable Absence of Correlation:

- CGPA vs. Study hours (r = 0.13) - Suggests studying benefits grades

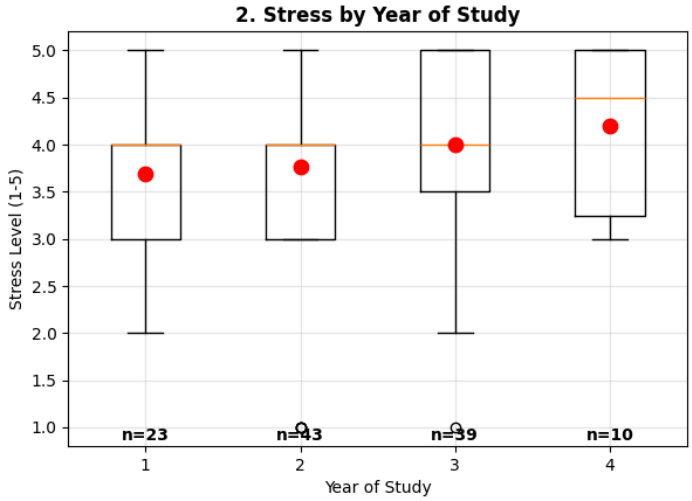
- Stress vs. Sleep quality (r = -0.16) - Higher stress associated with poorer sleep
- Anxiety vs. Sleep quality (r = -0.25) - Similar pattern to stress
- Year of study with most variables - Seniority doesn't strongly predict other factors
- Age with psychological metrics - Maturity doesn't correlate with stress/anxiety

3.3.2. Key Insights from Correlation Analysis

1. Mental Health Interconnection: The strong stress-anxiety correlation suggests these often co-occur in student populations
2. Academic Investment Pays Off: The moderate study hours-CGPA relationship supports the value of time investment
3. Sleep-Psychological Link: The negative correlations between sleep quality and stress/anxiety highlight the importance of sleep for mental wellbeing
4. Compartmentalization: Extracurricular and job variables show limited correlation with academic metrics, suggesting students may compartmentalize these life domains

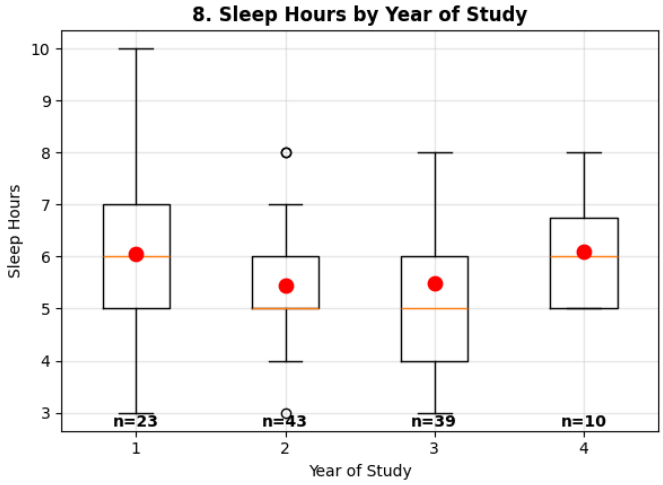
3.4. Multivariate Relationships

3.4.1. Stress by Academic Year



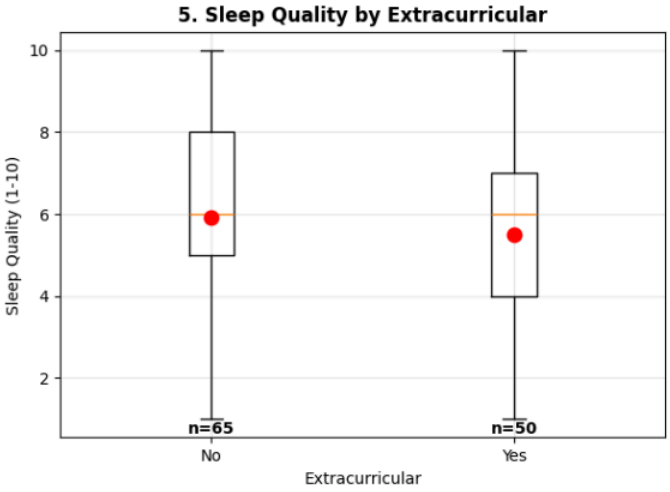
3rd year students report the highest stress levels (mean = 4.0), possibly due to increased academic pressure and career planning activities. 1st year students show slightly lower stress levels (mean = 3.7), potentially due to the adjustment period and lighter course loads.

3.4.2. Sleep hours by year of study



4th year students report the highest sleep hours (mean = 6.04), possibly due to reduced course loads and increased focus on personal well-being. 2nd year students show slightly lower sleep hours, potentially due to the adjustment to a heavier academic schedule and extracurricular commitments.

3.4.3. Sleep quality by extracurricular activities



Students with no extracurricular activities report higher sleep quality (mean = 5.92), possibly due to more available time for rest and fewer commitments. Students with extracurricular activities show lower sleep quality (mean = 5.50), potentially due to the added demands and time constraints of these activities.

4. Feature Engineering

Limited feature engineering was applied to make the dataset more suitable for analysis:

4.1. Categorical to Numerical Mapping

- Year of Study was converted into ordinal numeric values (1–4), allowing it to be used in regression models.

4.2. Binary Transformation

- The extracurricular and job variables were mapped to binary form (Yes = 1, No = 0), effectively creating dummy-like variables from categorical inputs.

4.3. Derived Numerical Features

- For participants with “No” in extracurricular or job, the corresponding hours were set to 0, which created a consistent numerical representation linking activity participation with time investment.

- This effectively combined two separate pieces of information (participation and hours) into a structured numeric feature.

Although no new synthetic features (e.g., interaction terms, ratios, or aggregations) were created, these transformations ensured categorical responses were properly encoded for statistical analysis & regression.

5. Hypothesis Testing

5.1. Executive Summary

A series of statistical tests were conducted to investigate key factors influencing student well-being. The results were surprising: At a 95% confidence level (α=0.05), none of the initial bivariate hypotheses showed a statistically significant relationship on stress levels in this particular group of students.

5.2.Tests Conducted and Results

1. Stress by Extracurricular Participation **(Independent t-test)**

Hypothesis:

H_0: $\mu_1 = \mu_2$ (No difference in mean stress between groups)

H_1: $\mu_1 \neq \mu_2$ (Difference in mean stress between groups)

- Finding: No significant difference ($t(113)=0.096$, $p=0.924$).
- Interpretation: Stress levels are statistically similar whether students participate in extracurricular activities or not.

2. Anxiety Across Academic Years (**One-way ANOVA)**

- Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (No difference in mean anxiety across years)

  H_1: At least one μ differs (Difference in mean anxiety across years)
- Finding: No significant difference ($F(3,111)=2.214$, $p=0.091$).
- Interpretation: Anxiety levels do not significantly differ from first year to fourth year.

3. Study Hours vs. Stress (**Pearson Correlation)**

Hypothesis:

H_0: $\rho = 0$ (No correlation between study hours and stress)

H_1: $\rho \neq 0$ (Correlation exists between study hours and stress)

- Finding: Weak correlation, not significant ($r=0.176$, $p=0.060$).
- Interpretation: There is no significant linear relationship between total study hours and stress in a simple correlation.

4. Sleep Hours vs. Anxiety (**Pearson Correlation)**

Hypothesis:

H_0: $\rho = 0$ (No correlation between sleep hours and anxiety)

H_1: $\rho \neq 0$ (Correlation exists between sleep hours and anxiety)

- Finding: Weak correlation, not significant ($r=-0.159$, $p=0.089$).
- Interpretation: There is no significant linear relationship between sleep duration and anxiety levels.

5. High Stress Prevalence by Job Status (**Chi-square Test)**

Hypothesis:

H_0: $p_1 = p_2$ (No difference in high stress proportion between job groups)

H_1: $p_1 \neq p_2$ (Difference in high stress proportion between job groups)

- Finding: No significant difference in proportions ($\chi^2(1)=0.085$, $p=0.771$).
- Interpretation: The proportion of students reporting high stress is not dependent on whether they have a job.

6. Predicting Stress (**Multiple Linear Regression)**

Hypothesis:

H_0: $\beta = 0$ (Predictor has no effect on stress controlling for other variables)

H_1: $\beta \neq 0$ (Predictor has effect on stress controlling for other variables)

- Finding: The overall model was not significant ($F=1.623$, $p=0.148$, $R^2=0.083$). However, study_hours_per_week was a significant individual predictor ($\beta=0.021$, $p=0.032$).
- Interpretation: When controlling for sleep, employment, and other factors, each additional hour of study per week is associated with a small but measurable increase in stress. This relationship was masked in the simpler bivariate analysis.

The analysis suggests that the drivers of student stress are not found in simple, one-to-one relationships but are more complex & multivariate. The significant effect of study hours in the regression model, despite its absence in the correlation test, underscores the importance of using multivariate techniques to uncover insights that simpler tests might miss.

5.3. Interpreting Our Results Using P-Values

p-value measures how likely our result is due to random chance, assuming no real effect exists.

- **Small p-value (≤ 0.05):** Strong evidence against the null hypothesis. We reject it and conclude a significant effect.
- **Large p-value (> 0.05):** Weak evidence against the null hypothesis. We cannot reject it; the result is likely due to chance.

➔ **Our Hypothesis Test Results:**

- Stress by Extracurriculars: p = 0.924
- Anxiety by Year of Study: p = 0.091
- Study Hours vs. Stress: p = 0.060
- Sleep Hours vs. Anxiety: p = 0.089
- High Stress by Job Status: p = 0.771

➔ **Why We Got These Results:**

All of our p-values are greater than 0.05. This means that for every test we ran, the observed difference or relationship could easily have occurred by random chance in a world where no true effect exists.

Therefore, we consistently failed to reject the null hypothesis. We do not have enough statistical evidence to conclude that extracurriculars, jobs, or years of study have a significant impact on student stress and anxiety in this particular dataset.

➔ **The One Nuance: The Regression**

However, a multiple regression model revealed that weekly study hours are a significant positive predictor of stress levels when controlling for other variables, though R^2 value of 0.083 means the model only explains 8.3% of the variation in stress scores. This is, by any standard, a very weak model. The vast majority (91.7%) of what causes a student's stress level is explained by factors which are not included in our model (e.g. personality, financial pressures, relationship issues, specific course difficulty e.t.c.). This indicates that while simple relationships are not evident, a more complex model can identify specific contributing factors.

6. Regression Model and Performance Evaluation

6.1. Simple Regression

6.1.1. Stress ~ Study Hours

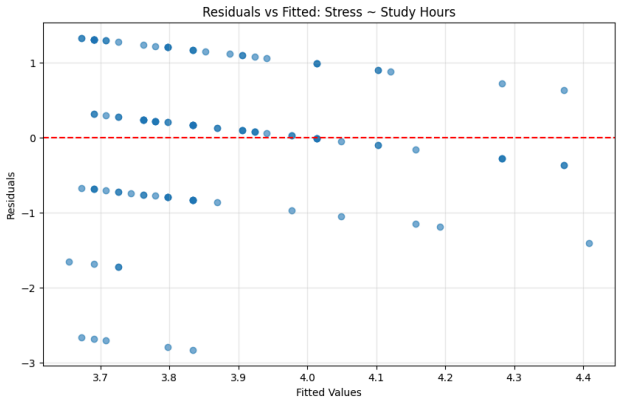
This model examined the extent to which weekly study hours predict student stress levels. The regression coefficient for study hours was 0.018 ($p = 0.060$), indicating a small positive relationship: for every additional study hour per week, stress level increased by 0.018 units on the 1–5 scale. However, this effect was not statistically significant at $\alpha = 0.05$. The model's explanatory power was low ($R^2 = 0.031$), suggesting that study hours alone explain only about 3% of the variance in stress.

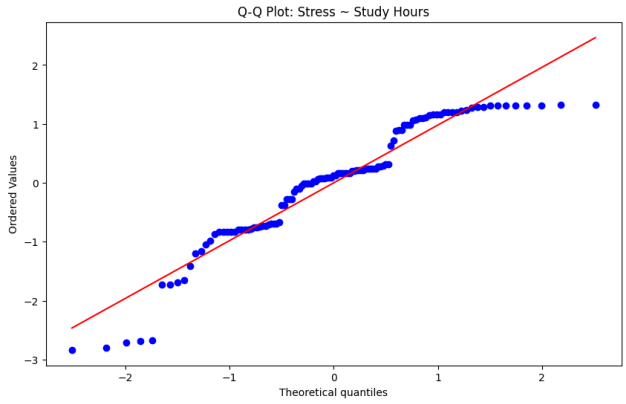
```
================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:            stress_1to5   R-squared:                       0.031
Model:                            OLS   Adj. R-squared:                  0.022
Method:                 Least Squares   F-statistic:                     3.603
Date:                Sat, 13 Sep 2025   Prob (F-statistic):             0.0602
Time:                        14:10:54   Log-Likelihood:                -163.36
No. Observations:                 115   AIC:                             330.7
Df Residuals:                     113   BIC:                             336.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const                      3.6543      0.147     24.781      0.000       3.362       3.946
study_hours_per_week       0.0179      0.009      1.898      0.060      -0.001       0.037
==============================================================================
Omnibus:                       11.413   Durbin-Watson:                   1.770
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               11.787
Skew:                          -0.752   Prob(JB):                      0.00276
Kurtosis:                       3.442   Cond. No.                         24.5
==============================================================================
```

Assumption Checks:

- Linearity: residual vs. fitted plot indicated no major deviations from linearity.
- Normality: Shapiro-Wilk test was significant ($p < 0.001$), suggesting non-normal residuals.
- Homoscedasticity: Breusch–Pagan test was significant ($p = 0.007$), indicating heteroscedasticity.
- Independence: Durbin–Watson statistic = 1.77, slightly below the ideal range (≈2), hinting at mild autocorrelation.

Overall, this model performs poorly, likely because stress is influenced by many non-academic factors not captured by study hours alone.





6.1.2. Stress ~ Year of Study

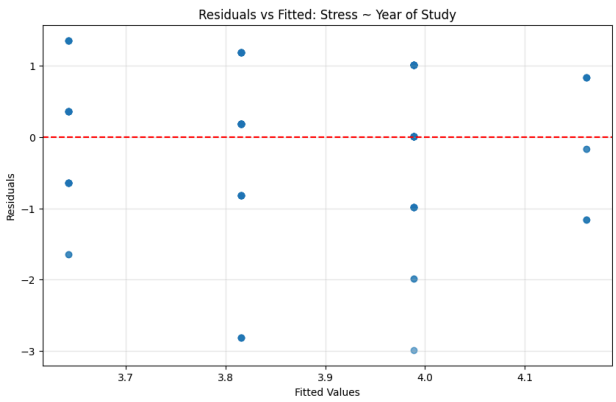
Here, year of study was tested as a predictor of stress. The regression coefficient was 0.173 ($p = 0.107$), suggesting stress tends to increase slightly with seniority, but the result was not statistically significant. The model's explanatory power was even lower ($R^2 = 0.023$), meaning year of study explains only ~2% of stress variation.

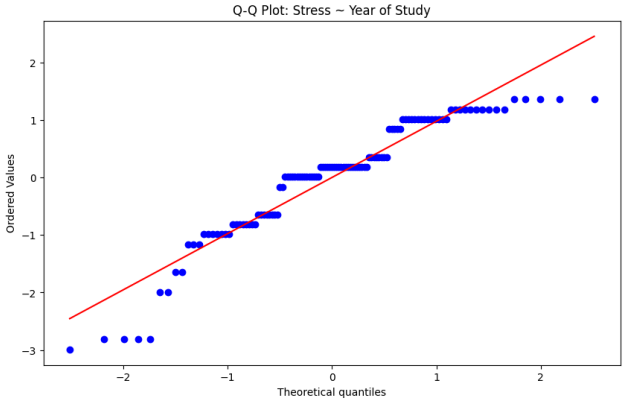
```
2. SIMPLE REGRESSION: Stress ~ Year of Study
================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:            stress_1to5   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.014
Method:                 Least Squares   F-statistic:                     2.638
Date:                Sat, 13 Sep 2025   Prob (F-statistic):              0.107
Time:                        14:10:55   Log-Likelihood:                -163.84
No. Observations:                 115   AIC:                             331.7
Df Residuals:                     113   BIC:                             337.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.4694      0.264     13.146      0.000       2.947       3.992
year_of_study  0.1730      0.107      1.624      0.107      -0.038       0.384
==============================================================================
Omnibus:                       17.704   Durbin-Watson:                   1.783
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.489
Skew:                          -0.952   Prob(JB):                     3.56e-05
Kurtosis:                       3.806   Cond. No.                         7.91
==============================================================================
```

Assumption Checks:

- Residuals deviated from normality (Shapiro–Wilk $p < 0.001$).
- Durbin–Watson = 1.78 indicated slight positive autocorrelation.
- Homoscedasticity was not severely violated, but residual plots showed mild uneven spread.

This weak model reflects that stress is not solely determined by academic year but by a combination of workload, coping skills, and lifestyle factors.





6.1.3. Anxiety ~ Sleep Quality

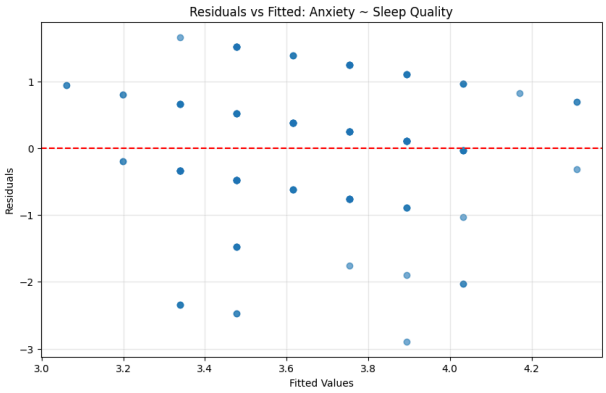
This model investigated whether sleep quality predicts student anxiety. The regression coefficient was -0.139 (p = 0.007), indicating a significant negative association: each one-point increase in sleep quality (1–10 scale) corresponded to a 0.139 decrease in anxiety (1–5 scale). The model explained about 6% of variance ($R^2 = 0.063$), which is modest but higher than the stress models.

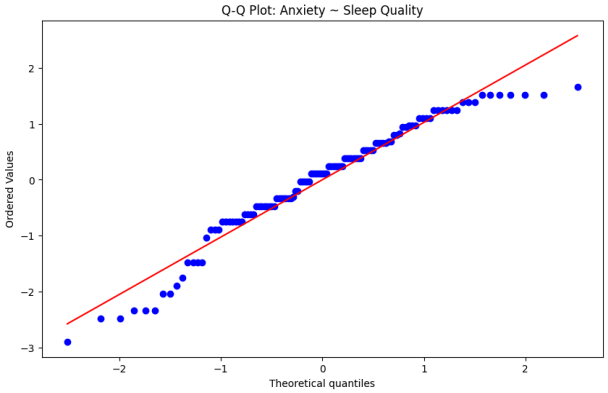
```
================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:           anxiety_1to5   R-squared:                       0.063
Model:                            OLS   Adj. R-squared:                  0.055
Method:                 Least Squares   F-statistic:                     7.652
Date:                Sat, 13 Sep 2025   Prob (F-statistic):            0.00663
Time:                        14:10:56   Log-Likelihood:                -166.50
No. Observations:                 115   AIC:                             337.0
Df Residuals:                     113   BIC:                             342.5
Df Model:                           1
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                     4.4488      0.304     14.642      0.000       3.847       5.051
sleep_quality_1to10      -0.1388      0.050     -2.766      0.007      -0.238      -0.039
==============================================================================
Omnibus:                        8.273   Durbin-Watson:                   2.102
Prob(Omnibus):                  0.016   Jarque-Bera (JB):                8.446
Skew:                          -0.663   Prob(JB):                       0.0147
Kurtosis:                       3.066   Cond. No.                         19.5
==============================================================================
```

Assumption Checks:

- Normality: Shapiro–Wilk test ($p < 0.001$) suggested residuals deviate from normality, though the Q–Q plot showed only mild deviations.
- Independence: Durbin–Watson = 2.10, within the acceptable range, indicating independence of residuals.
- Homoscedasticity: residuals showed fairly consistent spread.

This model performed better than the stress regressions and confirms the hypothesis that poor sleep quality is a meaningful driver of student anxiety.





6.2. Multiple Regression

6.2.1. Stress ~ Study Hours + Courses + Extracurricular Hours + Job Hours + Sleep Quality

This model tested whether a combination of academic, extracurricular, and lifestyle variables could explain variance in student stress.

- **$R^2 = 0.069$, Adj. $R^2 = 0.026$** → only ~7% of stress variance is explained, which is quite low.

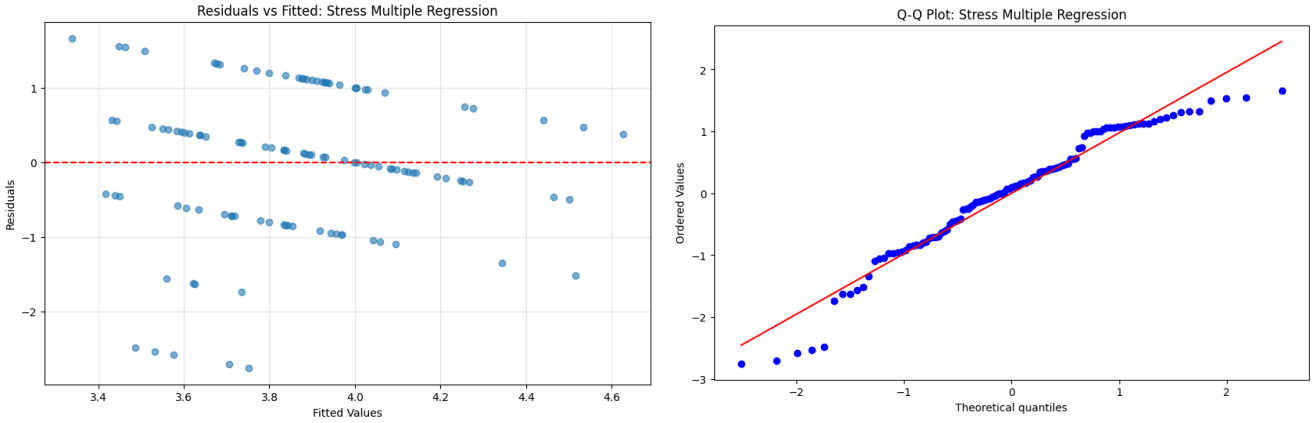
- None of the predictors were significant at $\alpha = 0.05$, though study hours ($p = 0.079$) and sleep quality ($p = 0.097$) showed weak trends in the expected directions (more study hours → higher stress; better sleep quality → lower stress).

- **Multicollinearity:** High VIF values for courses enrolled (8.56) and sleep quality (7.54) suggest multicollinearity, meaning predictors overlap in what they explain.

```
============================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:            stress_1to5   R-squared:                       0.069
Model:                            OLS   Adj. R-squared:                  0.026
Method:                 Least Squares   F-statistic:                     1.615
Date:                Sat, 13 Sep 2025   Prob (F-statistic):              0.162
Time:                        14:10:56   Log-Likelihood:                -161.06
No. Observations:                 115   AIC:                             334.1
Df Residuals:                     109   BIC:                             350.6
Df Model:                           5
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                        4.5757      0.541      8.457      0.000       3.503       5.648
study_hours_per_week         0.0176      0.010      1.775      0.079      -0.002       0.037
courses_enrolled            -0.1440      0.136     -1.057      0.293      -0.414       0.126
extracurricular_hours        0.0103      0.015      0.689      0.492      -0.019       0.040
job_hours_per_week           0.0003      0.009      0.027      0.978      -0.018       0.018
sleep_quality_1to10         -0.0818      0.049     -1.676      0.097      -0.179       0.015
==============================================================================
Omnibus:                        9.215   Durbin-Watson:                   1.783
Prob(Omnibus):                  0.010   Jarque-Bera (JB):                9.236
Skew:                          -0.680   Prob(JB):                      0.00987
Kurtosis:                       3.274   Cond. No.                         105.
==============================================================================
```

- **Assumption checks:**
    - Linearity: Rainbow test ($p = 0.798$) supports linearity.
    - Independence: Durbin–Watson = 1.78 (slight autocorrelation risk).
    - Normality: Shapiro–Wilk $p = 0.0006$, residuals deviate from normality.
    - Homoscedasticity: Breusch–Pagan $p = 0.0297$, heteroscedasticity present.

Interpretation: The model's poor performance indicates that stress is not easily explained by these quantitative workload/lifestyle measures alone. Stress may be more influenced by psychological or contextual factors not included in this dataset.



6.2.2.Sleep Quality ~ Sleep Hours + Stress + Anxiety

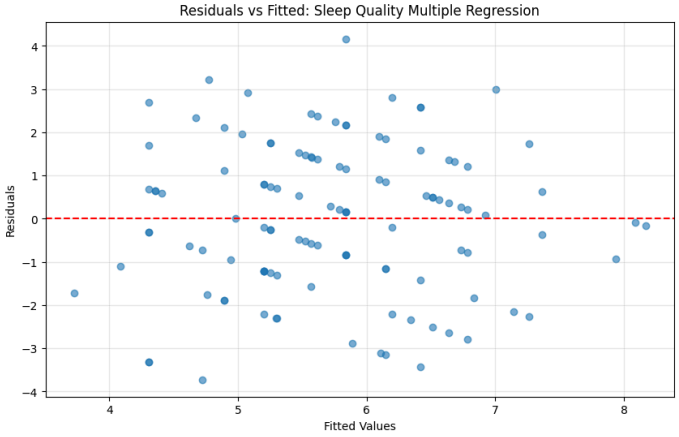
This model investigated the combined impact of sleep duration, stress, and anxiety on sleep quality.

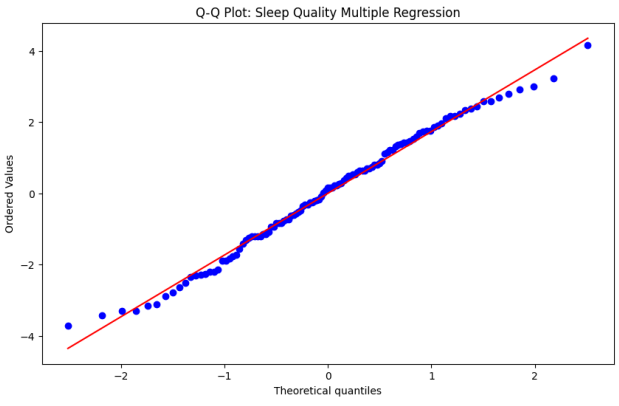
- **$R^2 = 0.216$, Adj. $R^2 = 0.195$** → about 22% of variance in sleep quality is explained, which is moderate compared to other models.
- **Sleep hours** was a strong positive predictor ($\beta = 0.582$, $p < 0.001$), confirming that longer sleep duration significantly improves sleep quality.
- Stress ($p = 0.808$) and anxiety ($p = 0.121$) were not significant after accounting for sleep hours, though anxiety trended negative.
- **Multicollinearity:** VIF values were very high (stress = 23.88, anxiety = 19.53, sleep hours = 8.62), suggesting overlapping effects and unstable coefficients.

```
============================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:     sleep_quality_1to10   R-squared:                       0.216
Model:                             OLS   Adj. R-squared:                  0.195
Method:                  Least Squares   F-statistic:                     10.20
Date:                 Sat, 13 Sep 2025   Prob (F-statistic):           5.50e-06
Time:                         14:10:57   Log-Likelihood:                -224.77
No. Observations:                  115   AIC:                             457.5
Df Residuals:                      111   BIC:                             468.5
Df Model:                            3
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.7914      1.025      3.701      0.000       1.761       5.822
sleep_hours    0.5823      0.125      4.649      0.000       0.334       0.831
stress_1to5   -0.0504      0.206     -0.244      0.808      -0.459       0.359
anxiety_1to5  -0.3117      0.200     -1.561      0.121      -0.708       0.084
==============================================================================
Omnibus:                        3.267   Durbin-Watson:                   2.009
Prob(Omnibus):                  0.195   Jarque-Bera (JB):                2.066
Skew:                          -0.089   Prob(JB):                        0.356
Kurtosis:                       2.368   Cond. No.                         50.4
==============================================================================
```

- **Assumption checks:**
    - Linearity: Rainbow test (p = 0.160) supports linearity.
    - Independence: Durbin–Watson = 2.01, good independence of residuals.
    - Normality: Shapiro–Wilk p = 0.466, residuals were approximately normal.
    - Homoscedasticity: Breusch–Pagan p = 0.677, no heteroscedasticity detected.

Interpretation: This model performed better than the stress regression and provides meaningful insight: sleep duration is the dominant driver of sleep quality, while stress and anxiety did not add significant explanatory power due to multicollinearity and indirect relationships.





7. Discussion

This study followed a step-by-step process to understand the factors affecting students' stress, anxiety, and sleep quality. We began by cleaning and preparing the data, where categorical answers such as *year of study*, *job*, and *extracurricular activities* were converted into numerical values so they could be used in regression models. For example, "Yes/No" answers were mapped to 1 and 0, and missing hours for students who did not have a job or extracurricular activity were replaced with 0. This basic feature engineering helped us create a dataset that was more consistent and ready for analysis.

Next, we carried out hypothesis testing to check simple relationships. Independent t-tests, one-way ANOVA, correlation tests, and chi-square tests were applied to see whether stress and anxiety differed across groups or were linked to study and sleep habits. Interestingly, most of these tests gave non-significant results ($p > 0.05$). For example, stress was not found to be significantly different between students with and without extracurriculars, and anxiety did not vary much across academic years. Correlations like study hours vs. stress and sleep hours vs. anxiety were weak and not statistically significant. This means that in our dataset, these simple one-to-one factors did not explain much about student well-being.

The regression analysis gave a more detailed picture. Simple regression models confirmed what the hypothesis testing suggested: study hours and year of study explained very little variation in stress (R^2 around 2–3%). However, the model of anxiety vs. sleep quality showed a significant negative relationship, meaning that students who reported better sleep quality generally experienced less anxiety. This result was both statistically significant and logical, since healthy sleep is widely known to support mental health.

When we moved to multiple regression, we tested more complex relationships. The stress prediction model, which included study hours, courses enrolled, extracurricular hours, job hours, and sleep quality, explained only about 7% of the variation in stress. None of the predictors were strong enough on their own, although study hours and sleep quality showed weak trends. The low R^2 suggests that stress is influenced by many other personal and social factors that were not captured in the survey. On the other hand, the sleep quality model (with sleep hours, stress, and

anxiety as predictors) performed better, explaining about 22% of the variation. Sleep hours was the strongest and most significant predictor, while stress and anxiety did not add much once sleep time was included.

Finally, when checking assumptions such as linearity, normality, and homoscedasticity, some models showed issues. For example, the residuals were not perfectly normal and heteroscedasticity was present in certain cases. This indicates that linear regression might not fully capture the complexity of the data, and that non-linear or more advanced models might perform better in future research.

Why the Models Did Not Perform as Expected

The relatively low R^2 values show that our models could not explain most of the variation in stress, anxiety, or sleep quality. There are several reasons for this:

1. **Limited variables in the dataset**: Stress and anxiety are influenced by many factors that were not measured, such as financial concerns, family support, social relationships, health conditions, or exam schedules. Without including these, the models naturally underperform.

2. **Sample size**: With only around 115 participants, the dataset may not be large enough to capture subtle effects or detect small but real relationships. A larger sample could give more reliable results.

3. **Measurement quality**: Self-reported data can be biased or inaccurate. For example, students might under- or over-report their study hours or stress levels. This adds noise and reduces model accuracy.

4. **Linear assumptions**: Our models assumed linear relationships, but the actual effects might be non-linear (e.g., moderate study hours might reduce stress, but very high hours could sharply increase it). Linear regression cannot capture such curved patterns.

How Results Could Be Improved

To improve model performance in future studies, several steps could be taken:

- **Include more predictors**: Adding survey items on financial pressure, family issues, coping strategies, and course workload would capture more of the real drivers of stress and anxiety.

- **Larger and more diverse sample**: Collecting data from a bigger group of students across different universities or semesters would help generalize findings and increase statistical power.

- **Use advanced models**: Non-linear models (like polynomial regression, decision trees, or random forests) or regularized regression (Lasso/Ridge) could capture more complex relationships than standard linear regression.

- **Feature engineering**: Creating interaction terms (e.g., study hours × sleep hours) or ratios (e.g., study-to-sleep balance) could reveal effects hidden in the simple raw variables.

Overall Insight

In summary, while the models did not perform strongly, they provided valuable insights. The strongest and most consistent finding is that sleep quality and sleep duration are key factors linked to student well-being. The weak performance of other predictors suggests that stress and anxiety are shaped by a broader set of influences not captured here. This highlights the need for richer data and more advanced modeling approaches to fully understand student well-being.

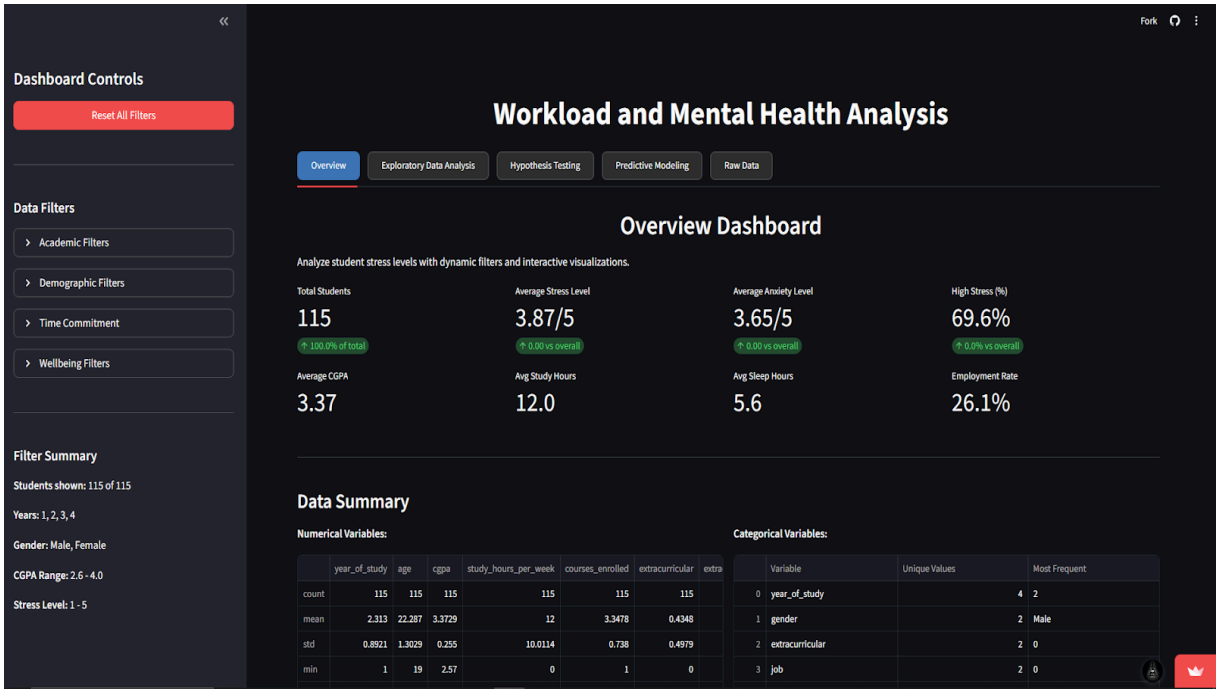
8. Dashboard

Overview

This is an online tool designed to help students understand the causes of stress. It's called the Student Stress Analysis Dashboard and was built using Streamlit. The dashboard has tools to look at the data, test ideas, and analyze information about student mental health. Dashboard URL: https://xhafan-dashboard1.streamlit.app/ .

Key Features -

This dashboard has two main sections. The left side provides options to sort and filter data. The right side displays the main dashboard, showing the complete analysis with visuals. Here is a basic glimpse of the total Dashboard below-

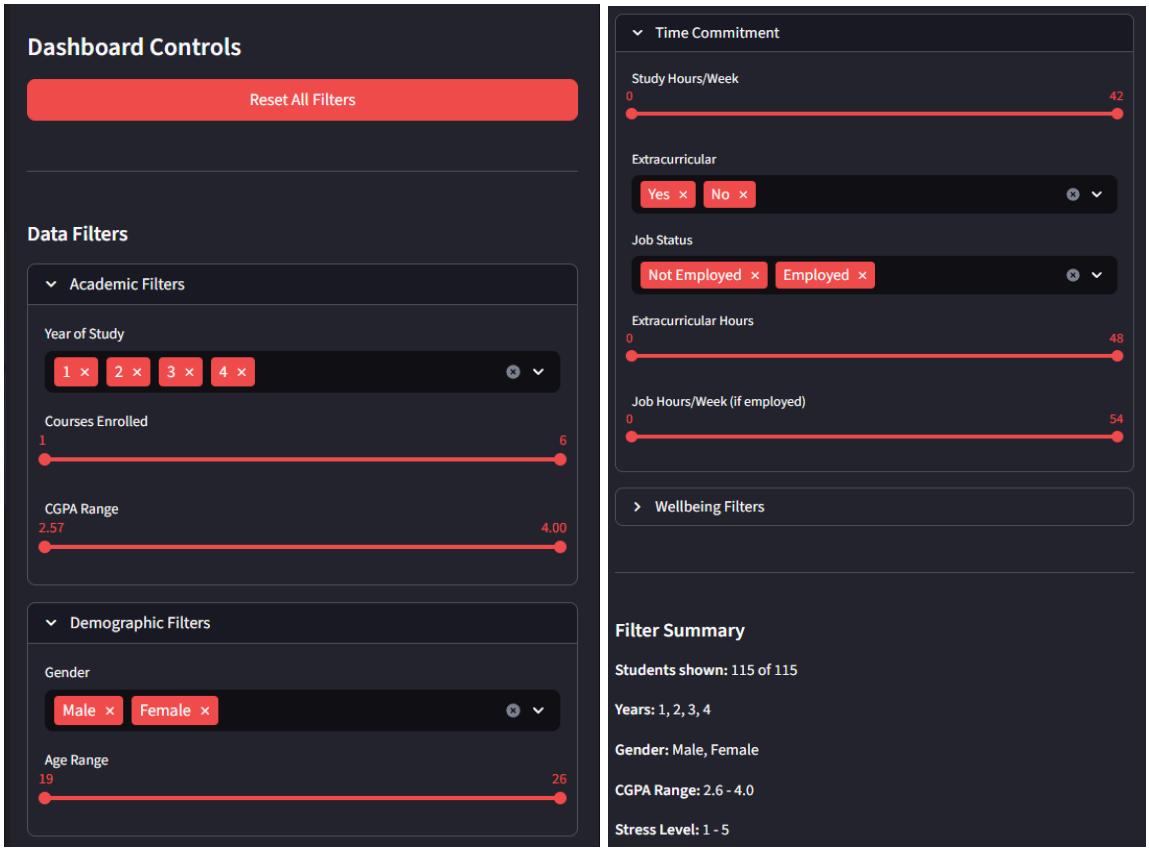


Filter System -

The dashboard features a unique filtering system with four distinct sections for precise data selection. The filter system is broken down as follows:

- **Academic Filters**
- **Demographic Filters**
- **Time Commitment**
- **Wellbeing Filters**

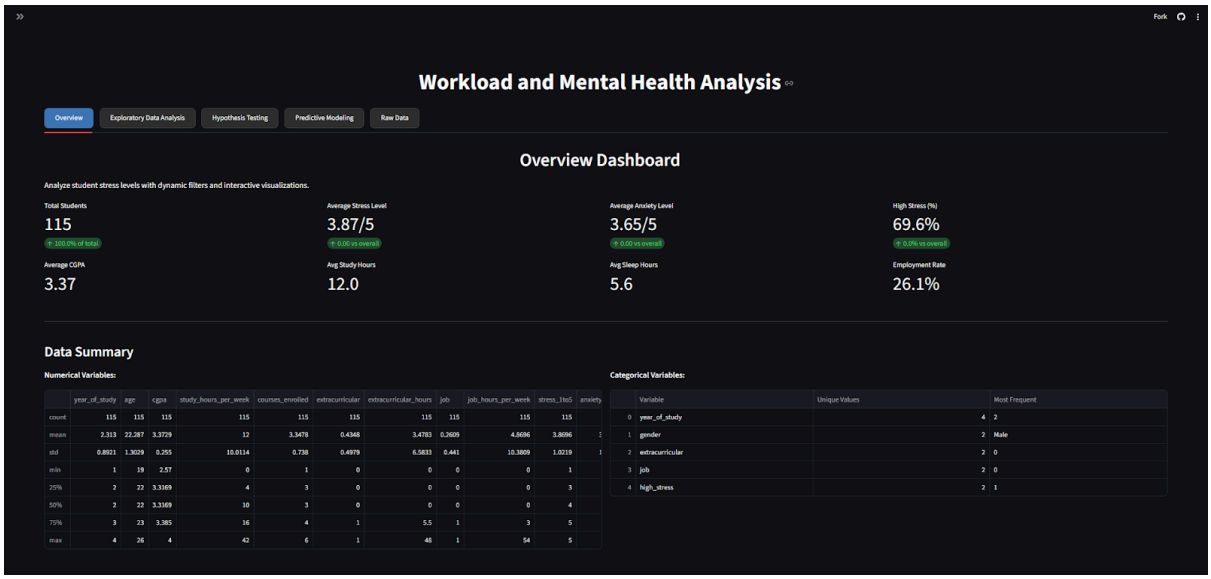
Additionally, the filters can be reset. A summary section below the controls displays all active filter selections.



Dashboard Overview -

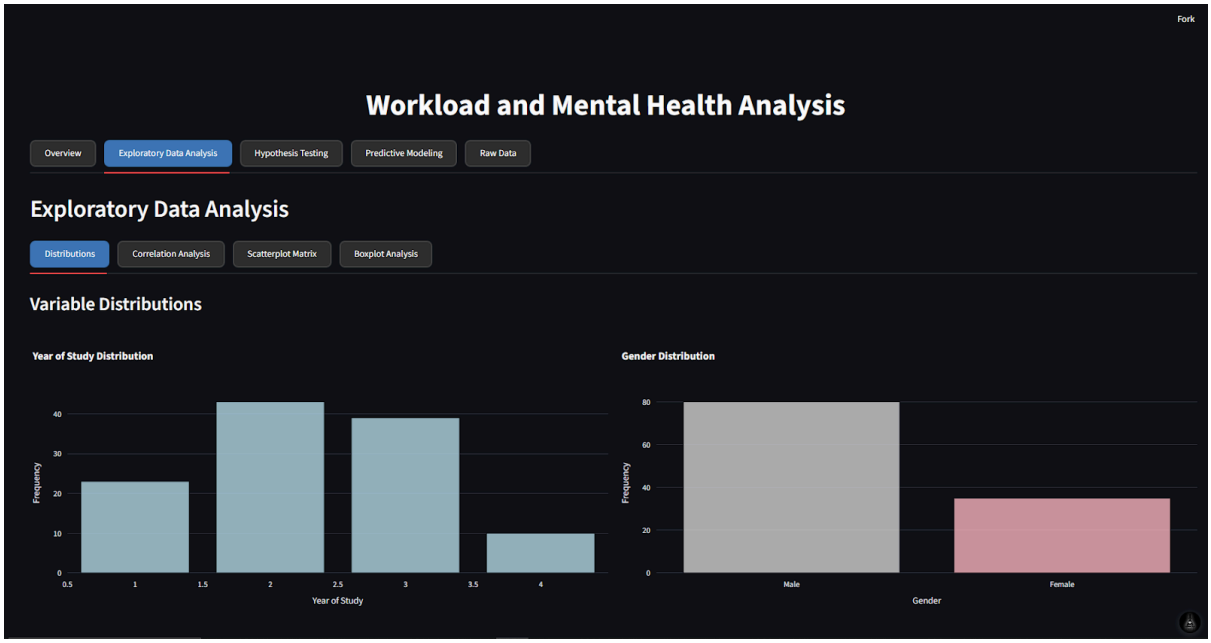
The dashboard is organized into five main sections: Overview, Exploratory Data Analysis, Hypothesis Testing, Predictive Modeling, and Raw Data.

Each section contains its own subsections, complete with visuals. These visuals provide a clear summary of each part, making the data easier to understand.

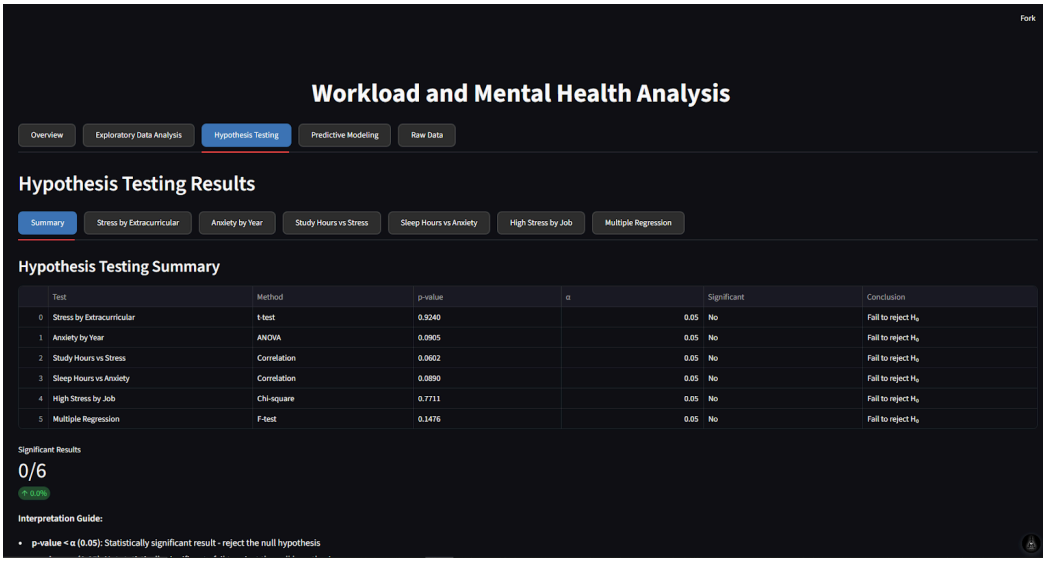


Overview Section - This section analyzes student stress levels using dynamic filters, interactive visualizations, and a data summary.

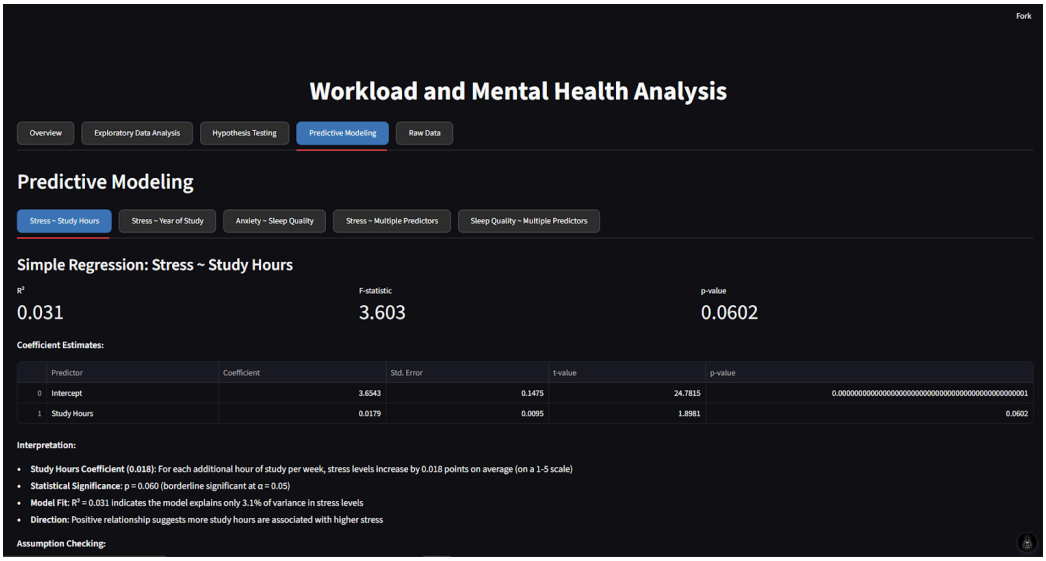
Exploratory Data Analysis - This section is divided into four subsections: Distributions, Correlation Analysis, Scatterplot Matrix, and Boxplot Analysis. It provides visual analyses, including data distributions, a correlation matrix for numerical variables, scatterplots, and boxplots.

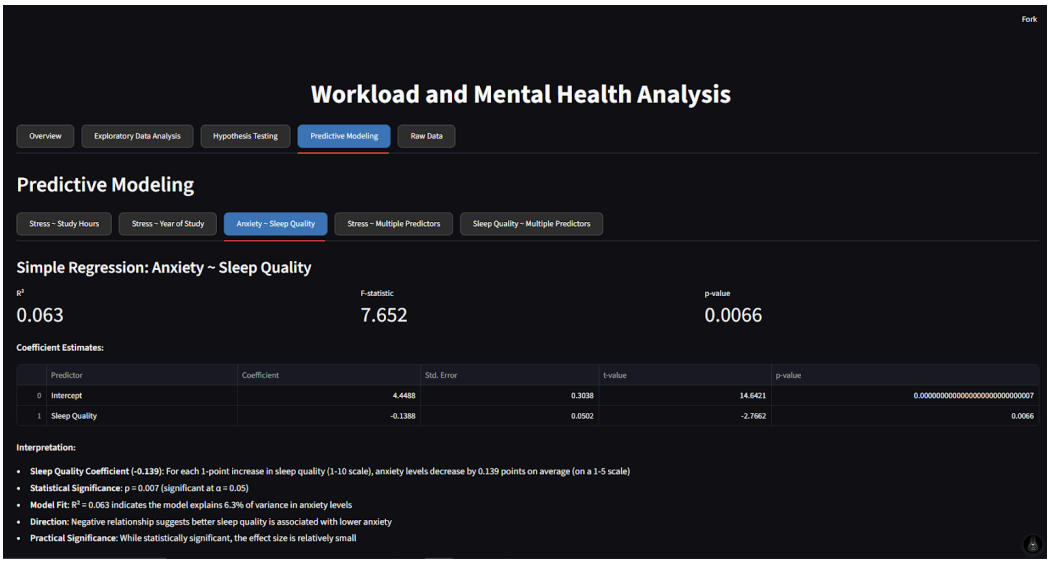


Hypothesis Testing - This section summarizes the results of various hypothesis tests. It explains and shows the data for several hypotheses. The results for each hypothesis are presented alongside a visual explanation.



Predictive Modeling - This section presents two types of predictive models: simple and multiple, featuring three simple and two multiple models in total. Model performance is demonstrated across several factors, including the relationships between stress and study hours, year of study, and sleep quality.





Raw Data - The dashboard includes a feature to display the raw data, which can also be downloaded as a CSV file for offline analysis.

9. Conclusion

This project was both a challenge and a learning journey for us. We started with a simple survey dataset and gradually moved through data cleaning, feature engineering, exploratory analysis, hypothesis testing, and regression modeling. Along the way, we faced many limitations, especially with time, knowledge gaps, and the complexity of student well-being as a research topic.

Even though the models did not perform as strongly as we hoped, the process itself taught us a lot. We learned how to prepare messy data, apply statistical tests, build regression models, and critically evaluate their assumptions and results. We also discovered that real-world data often does not give straightforward answers and that weak results can still provide meaningful insights.

While we could not capture every factor influencing stress and anxiety in this short time, we still put our best effort into making the analysis systematic and professional. This experience has shown us not only what we can already do, but also what we need to learn further. With more time, better data, and improved techniques, we believe we could build stronger models and uncover deeper insights.

In the end, this project was less about getting perfect results and more about growth — in skills, understanding, and persistence.

Reference:

1. Colab Notebook Link: https://colab.research.google.com/drive/1hdccfuWJSIIVB3x1t2-3JP2BwjaFZ0w1?usp=sharing
2. Project Repository Link: https://github.com/XhAfAn1/Workload-and-Mental-Health-Analysis-Dashboard-using-Streamlit
3. Dashboard Link: https://xhafan-dashboard1.streamlit.app/
4. Survey Response: https://drive.google.com/file/d/1Q8IOwNbN-NswQRCkDqbAQQwDvSg53GRb/view?usp=sharing