

**CSE303: Statistics for Data Science**

**[SUMMER 2025]**

**Term Project Report**

**Modelling the Relationship Between Workload and Mental**

**Health in University Students**

**Submitted by:**

| **Student ID** | **Student Name** | **Contribution Percentage** |
| --- | --- | --- |
| **2023-1-60-097** | **Hassan Ahmed Shawon** | **45%** |
| **2023-1-60-212** | **Albin Israt Siddique** | **35%** |
| **2023-1-60-236** | **Yeamin Sultana** | **20%** |
| **2023-1-60-215** | **Abdullah Saleh Mahmud** | **0%** |

# Introduction

In this project, We collected data from over 200+ students and prepared many simple linear and multiple linear regression model. Our goal was to find some relations between the features of our servey and visualizing them. Simple and Multiple linear regression models were created based on the correlation between the features which we found through heatmap. 4-types of Hypothesis Testing were also added here(eg: Independent t-test, One-way ANOVA, Correlation test,Proportion test (Chi-square).

We collected our dataset from conducting a severy among university students from different institutions of BD using Google-Forms. Integrity of data was our main priority, we tried to ensure all data’s are filled by different students to get dynamic and real data. Overall, this project is capable of showing relationship between workload and mental health in university students by analysis different aspects.

# 2. Data Preprocessing

Filling up missing values was a quite challenging part; we first analyzed the particular feature distribution type and what does the missing value indicates. based on that result, we fulfilled CGPA column empty values with the mean of CGPA column as the distribution is symmetric. PartTime\_Job\_Hours, Extracurricular\_Hours column empty values were filled with 0 as the empty value indicates that they donot work or participate in extracurricular activities so their invested hours is empty. Categorical values were encoded with numerical values(ex: Years\_of\_Study→{"1st":1,"2nd":2,"3rd":3,"4th":4},Gender→{"Male":1,”Female":0},Extracurricular\_pariticipation→{"Yes":1,"No":0},PartTime\_Job\_Status→{"Yes":1, "No":0}. We also dropped the rows where Years\_of\_Study is null. Feature Scaling Strandardization process is used here for scaling the data.

# 3.Dataset Characteristics and Exploratory Data Analysis

Our Student\_Servery dataset has (205,14) rows and columns and each rows are different.

In this section, introduce your dataset. Mention the number of rows, columns, and other characteristics. Provide the histograms of data distribution and correlations among the variable with a suitable discussion. Try to stand out and be creative in your presentation!

# Feature Engineering

In this project , Standardizatin of Feature Scaling is used to scale the all numeric data . It makes the data range between -1 to +1. It helped to obtain the correct r^2 value.

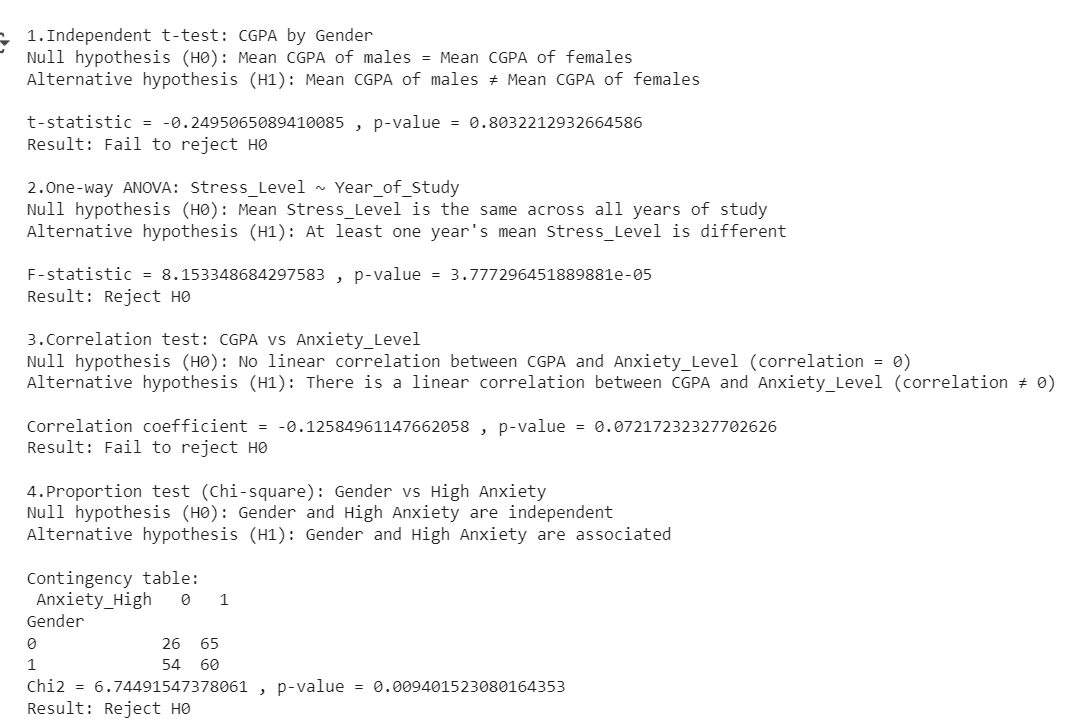
# Hypothesis Testing

**Indepndent t-test:** CGPA by Gender

H0: Mean CGPA of males = Mean CGPA of females

Ha : Mean CGPA of males != Mean CGPA of females  
t-statistic = -.25 , p-value= 0.803

p-value> 0.05 → **Fail to reject H0**  
Interpretation: There is no difference or mean CGPA of males is equal to females.



**One-way ANOVA:** Stress\_Level by Year\_of\_Study

H0: Mean Stress\_Level is same across all years of study  
Ha : At least one year’s mena Stress\_Level is different  
F-statistic = 8.15 , p-value=3.78e^-05

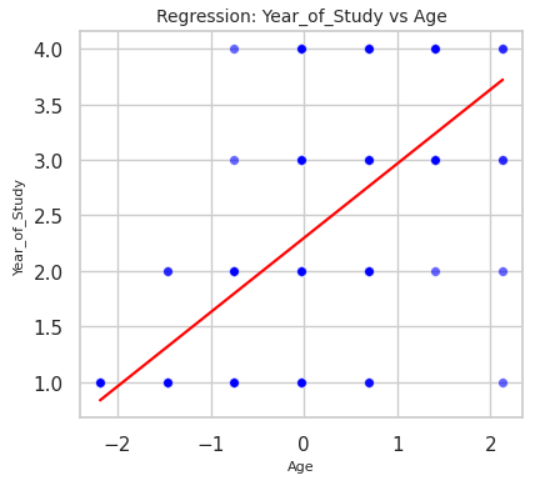
p-value< 0.05 → **reject H0**  
Interpretation: At least one year’s mena Stress\_Level is different

**Correlation Test:** CGPA by Anxiety\_Level

H0: No linear correlation between CGPA and Anxiety\_Level  
Ha : There is linear correlation between CGPA and Anxiety\_Level  
Correlation coefficient= -0.126, p-value=0.072

p-value>0.05 → **fail to reject H0**  
Interpretation: There is no linear correlation between CGPA and Anxiety\_Level

**Proportion Test(Chi-square):** Gender vs High Anxiety  
H0: Gender and High Anxiety are independent  
Ha : Gender and High Anxiety are associated  
Chi2 = 6.74, p-value = 0.009

p-value < 0.05 → **Reject H₀**  
Interpretation:

* Gender is significantly associated with anxiety level and Males are moore like to have high anxiety than female.
* Male has **65%** high anxiety vs **26%** low anxiety
* Female has **60%** high anxiety vs **54%** low anxiety

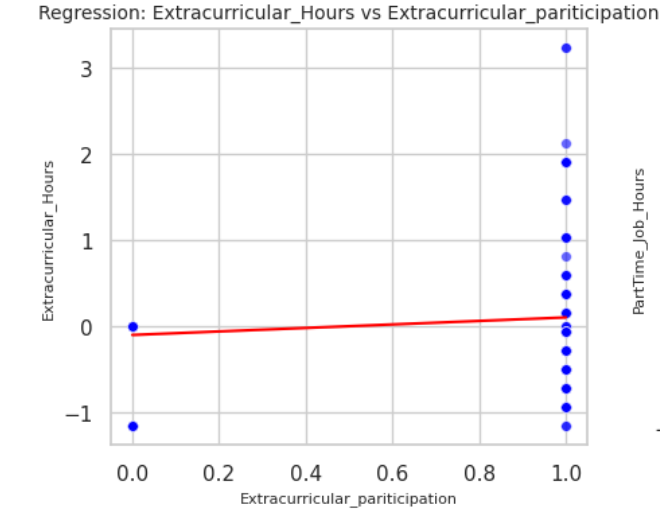
# Regression Model and Performance Evaluation

Here our top 5 best simple and multiple linear regression model is shown

## Simple Linear Regression Models **1. Years\_of\_Study ~ Age**

* **Intercept ≈ 0** ; when Age is zero, Year\_of\_Study is also zero means study is not started yet.
* **Age Coefficient ≈ 0.67** ; means for every 1   
  unit increase in age will increase the Year\_of\_Study by 0.67 on average
* **R2- Value ≈ 0.457**; about 45.7% f the variation in Year\_of\_Study can be explained by Age.

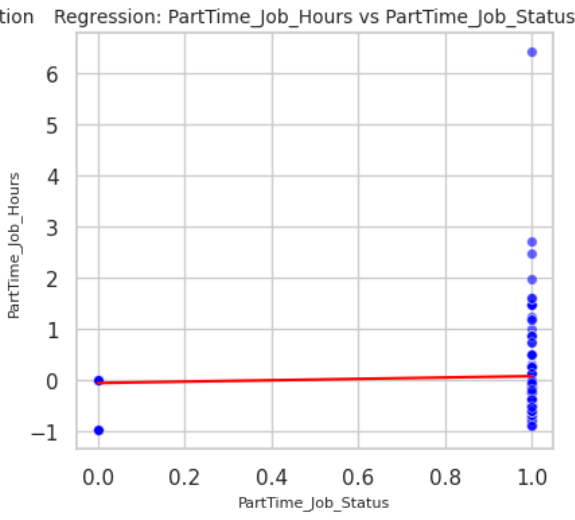
**2.Extracurricular\_Hours~Extracurricular\_pariticipation**



* **Intercept ≈ -0.64**; non-paritcipants do not have any hours dedicated for extracurricular activities.
* **Extracurricular\_pariticipation ≈ 1.34**;

means for every 1 unit increase in Extracurricular\_pariticipation will increase the Extracurricular\_Hours by 1.34 on average.

* **R2- Value ≈ 0.447**; about 44.7% of the variation in Extracurricular\_Hours can be explained by Extracurricular\_pariticipation.



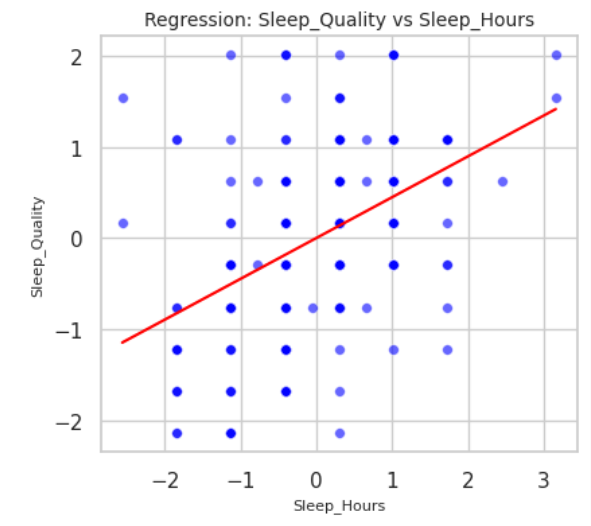
**3. PartTime\_Job\_Hours ~ PartTime\_Job\_Status**

* **Intercept ≈ -0.60** ; This means who doesn’t have partime job **≈** donot spent any job hours.
* **PartTime\_Job\_Status Coefficient ≈ 1.19** ; means for every 1 unit increase in partTime\_Job\_Status will increase the PartTime\_Job\_Hours by 1.19 on average .
* **R² ≈ 0.355**; about 45.7% f the variation in Year\_of\_Study can be explained by Age.



**4. Anxiety\_Level~ Stress\_Level**

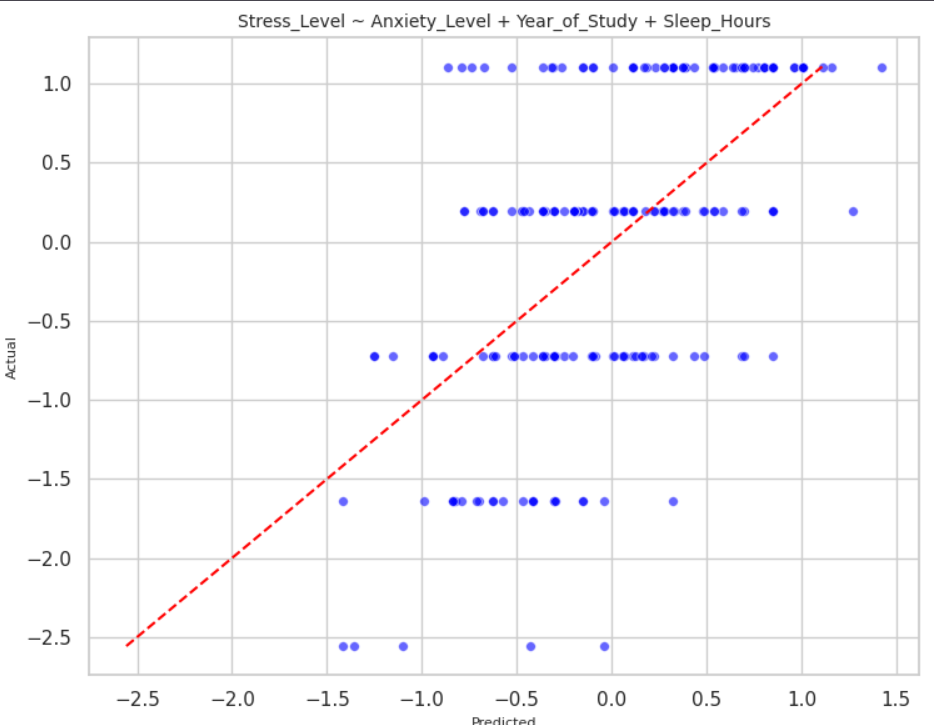
* **Intercept ≈ 0** ; When Stress\_Level = 0, Anxiety\_Level **≈**  0.
* **Stress Coefficient ≈ 0.48** ; means for every 1   
  unit increase in Stress\_Level will increase the Anxiety\_Level by 0.48 on average
* **R² ≈ 0.233** ;About 23.3% of the variation in Anxiety\_Level is explained by Stress\_Level.

**5.Sleep\_Quality ~ Sleep\_Hours**

* **Intercept ≈ 0** ; when Age is zero, Year\_of\_Study is also zero means study is not started yet.
* **Sleep\_Hours Coefficient ≈ 0.45**; means for every 1 unit increase in Sleep\_Hours will increase the sleep\_Quality by 0.45 on average
* **R² ≈ 0.200** ; About 20% of the variation in sleep\_Quality is explained by Sleep\_Hours .

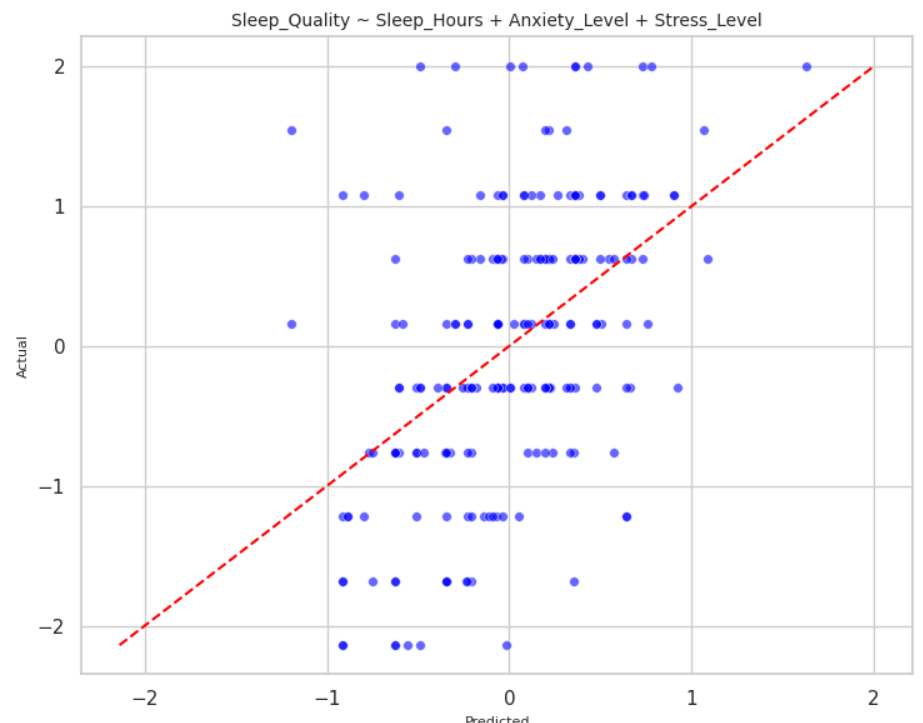
## Multiple Linear Regression Models

1. **Stress\_Level ~ Anxiety\_Level + Year\_of\_Study + Sleep\_Hours**



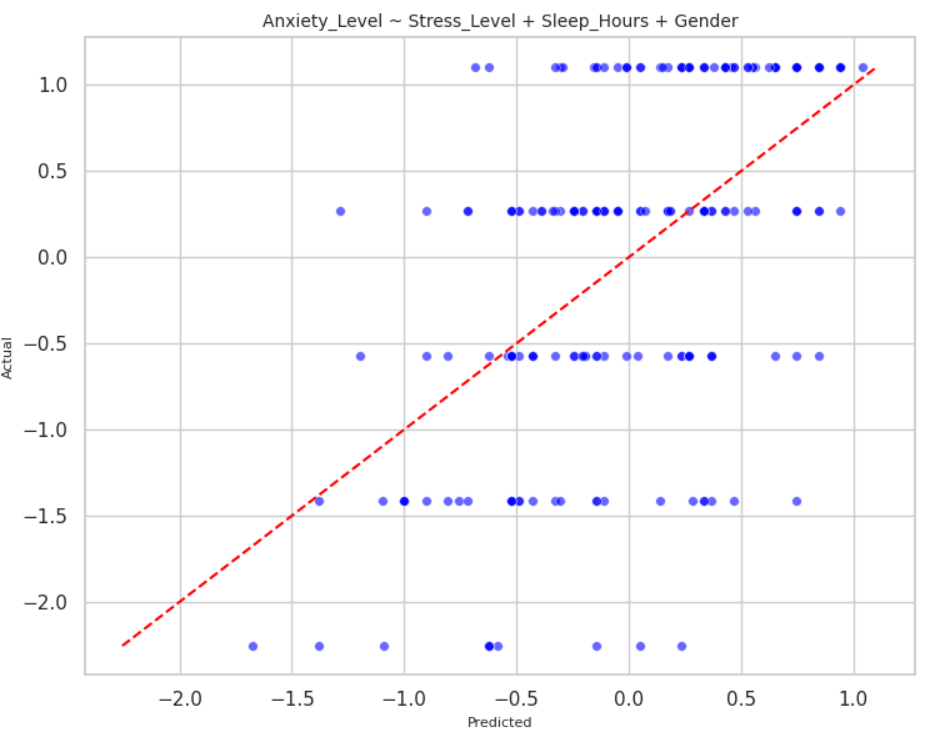
* **Intercept ≈ 0** ; When predictors are 0, stress is ~0 (not meaningful baseline).
* **Anxiety\_Level Coefficient ≈ 0.38** ;Each 1-unit increase in anxiety increases stress by ~0.38.
* **Year\_of\_Study Coefficient ≈ 0.26** ; Each 1-year increase in study year increases stress by ~0.26.
* **Sleep\_Hours Coefficient ≈ -0.22** ; Each extra hour of sleep decreases stress by ~0.22.
* **R² ≈ 0.345** ; About 34.5% of variation in stress is explained by anxiety, year of study, and sleep hours.

### Sleep\_Quality ~ Sleep\_Hours + Anxiety\_Level + Stress\_Level

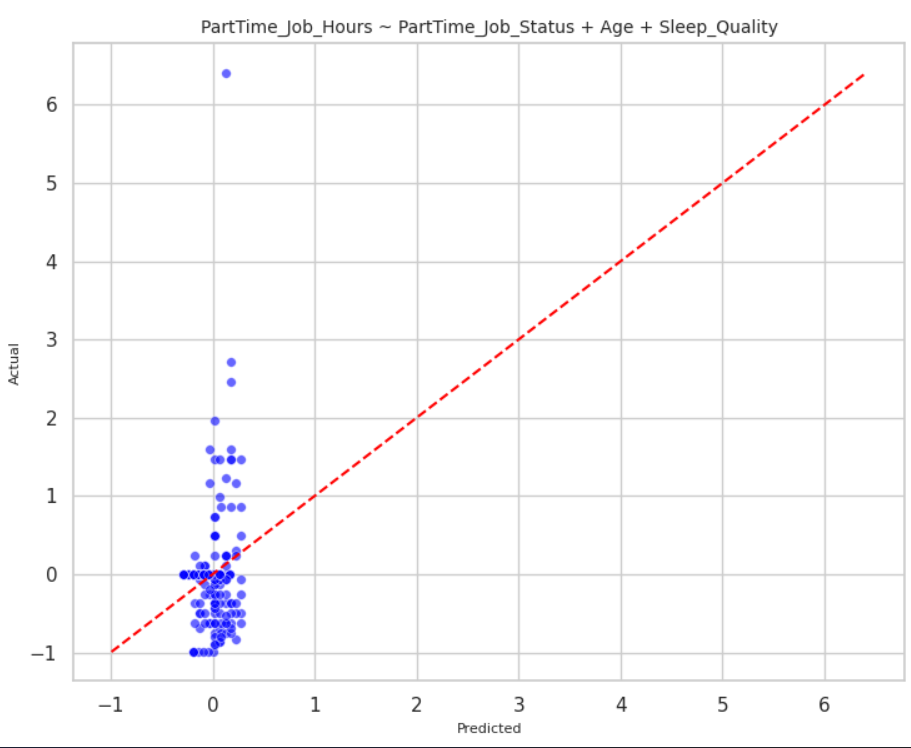


* **Intercept ≈ 0** ;When predictors are 0, sleep quality is ~0 (baseline).
* **Sleep\_Hours Coefficient ≈ 0.40** ; Each extra hour of sleep improves sleep quality by ~0.40.
* **Anxiety\_Level Coefficient ≈ -0.14** Higher anxiety tends to lower sleep quality (weak evidence).
* **Stress\_Level Coefficient ≈ -0.02** ;Stress has almost no effect on sleep quality.
* **R² ≈ 0.221** ;About 22.1% of variation in sleep quality is explained by these predictors.

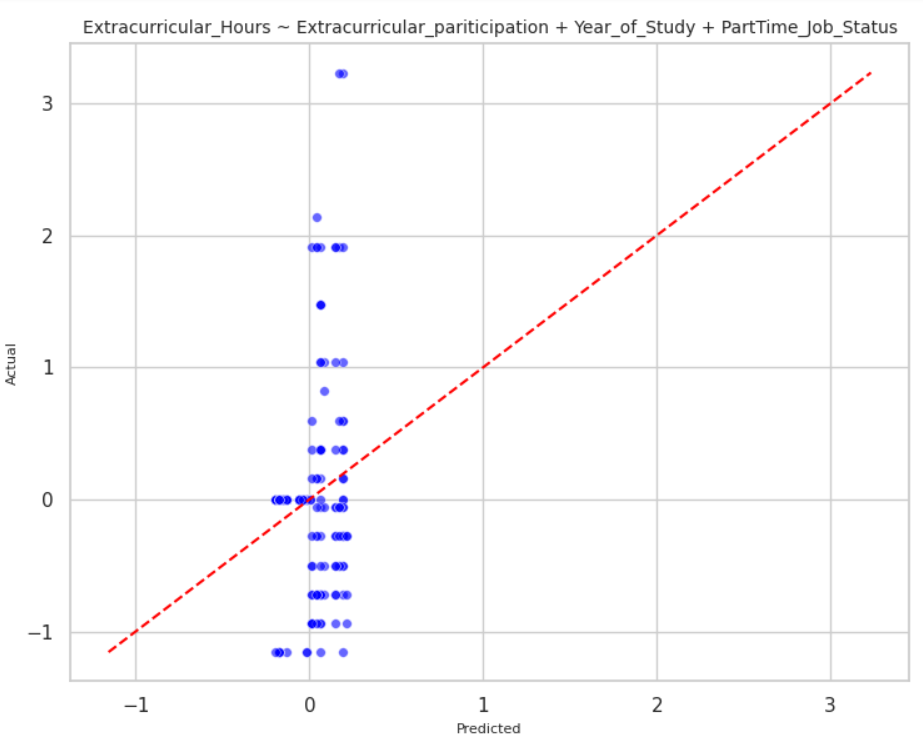
### Anxiety\_Level ~ Stress\_Level + Sleep\_Hours + Gender

* **Intercept ≈ 0.23** ;Baseline anxiety level when predictors = 0.
* **Stress\_Level Coefficient ≈ 0.41** ; Each 1-unit increase in stress increases anxiety by ~0.41.
* **Sleep\_Hours Coefficient ≈ -0.14** ;Each extra hour of sleep decreases anxiety by ~0.14.
* **Gender Coefficient ≈ -0.41** ;Gender difference: coded group (likely female vs male) has ~0.41 lower anxiety on average.
* **R² ≈ 0.295** ;About 29.5% of variation in anxiety is explained by stress, sleep, and gender.

### PartTime\_Job\_Hours ~ PartTime\_Job\_Status + Age + Sleep\_Quality

* **Intercept ≈ -0.58** ;Baseline job hours when predictors = 0 (not realistic, just model baseline).
* **PartTime\_Job\_Status Coefficient ≈ 1.16** ;Students with a job work ~1.16 more hours than those without.
* **Age Coefficient ≈ 0.02 (ns)** ;Age has no meaningful effect on job hours.
* **Sleep\_Quality Coefficient ≈ 0.10 ;**Higher sleep quality slightly increases job hours (weak evidence).
* **R² ≈ 0.366** ; About 36.6% of variation in job hours is explained by these predictors.

### Extracurricular\_Hours ~ Extracurricular\_participation + Year\_of\_Study + PartTime\_Job\_Status

* **Intercept ≈ -0.61**;Baseline extracurricular hours when predictors = 0.
* **Participation Coefficient ≈ 1.34**; Participants spend ~1.34 more hours in extracurriculars than non-participants.
* **Year\_of\_Study Coefficient ≈ 0.01;**Year of study has no effect on extracurricular hours.
* **PartTime\_Job\_Status Coefficient ≈ -0.06 ;** Job status has no effect on extracurricular hours.
* **R² ≈ 0.447** ;About 44.7% of variation in extracurricular hours is explained by these predictors (but mostly participation).

# Discussion

Most important section!

Analyse the performance of your regression systems and provide the hypothesis behind their performance, e.g., why the model is performing in this way or why it is not performing better? Provide appropriate reasoning for your hypothesis.

# Dashboard

Provide a description of your dashboard using screenshots as necessary. Also, include the link to your code and dashboard.

# Conclusion

For collecting data, we had some debate whether to collect data offline or online as on online survey people don’t give actual data and offline data collection much more time consuming and hard working also costly. Finally, we decided to collect data using Google-Forms. Cleaning data and adjusting the R2  value was completely unknown to us , we had study a lot for this. EDA part was also challenging because dataset is not so organized. Lastly, building and running the application was quite frustrating and brainstorming for us.

In our project, regression models are underfitted and has low-r^2 value because of the dataset. For this reason , many potential important relationships were not found in our project. This project has some potential to predict and describe student’s mental health with respect to workload features only if the data are collected sincerely and accurately.   
  
This project might not be perfect but by building this project we got to understand statistics and its related topic, deeply which will help us to strengthen our knowledge in the Data Science filed in Future. Thanks to our Rezwan sir for his effort.