**Student Stress Factors & Academic Performance: Exploratory Data Analysis & Prediction**

## **Team Members:**

1. Dharmesh Chawda
2. Meghkumar Patel
3. Nidhi Shah

# **1. Introduction**

Stress among students is one of the most widely recognized factors affecting academic performance, physical well-being, and long-term mental health. With the growing academic workload, environmental pressures, and social expectations, understanding *what drives student stress* is an important problem for educators and mental-health practitioners.

For this project, our team selected a **real-world dataset of 1,100 students with 21 variables**, covering psychological, physiological, academic, environmental, and social factors contributing to student stress.

We chose this dataset because:

* It is **rich in multidimensional variables**, allowing diverse EDA exploration.
* Stress prediction is a **meaningful ML application** with societal importance.
* The dataset aligns well with the course project requirements: exploration, visualization, question answering, and prediction.

# **2. Dataset Description**

**Dataset:** *StressLevelDataset.csv***Size:** 1,100 rows × 21 columns

**Google Drive Link:** <https://drive.google.com/file/d/1oBQdHuebufWyMuYksH9v0aeAkkb1CUBV/view?usp=drive_link>

**Variables include:**

* sleep\_quality, study\_load, exam\_pressure, family\_issues, teacher\_relationship, peer\_pressure, etc.
* Target variable: **stress\_level**

# **3. Data Preparation & Cleaning**

### **3.1 Data Loading & Inspection**

We used pandas to load the dataset and inspect shape, column types, and descriptive statistics.

Shape: (1100, 21)

No. of numeric variables: 21

Missing values: minimal, handled as below

### **3.2 Missing Value Handling**

* Instances with missing numeric values were imputed using column mean.
* Outliers were inspected visually; no removals were necessary.

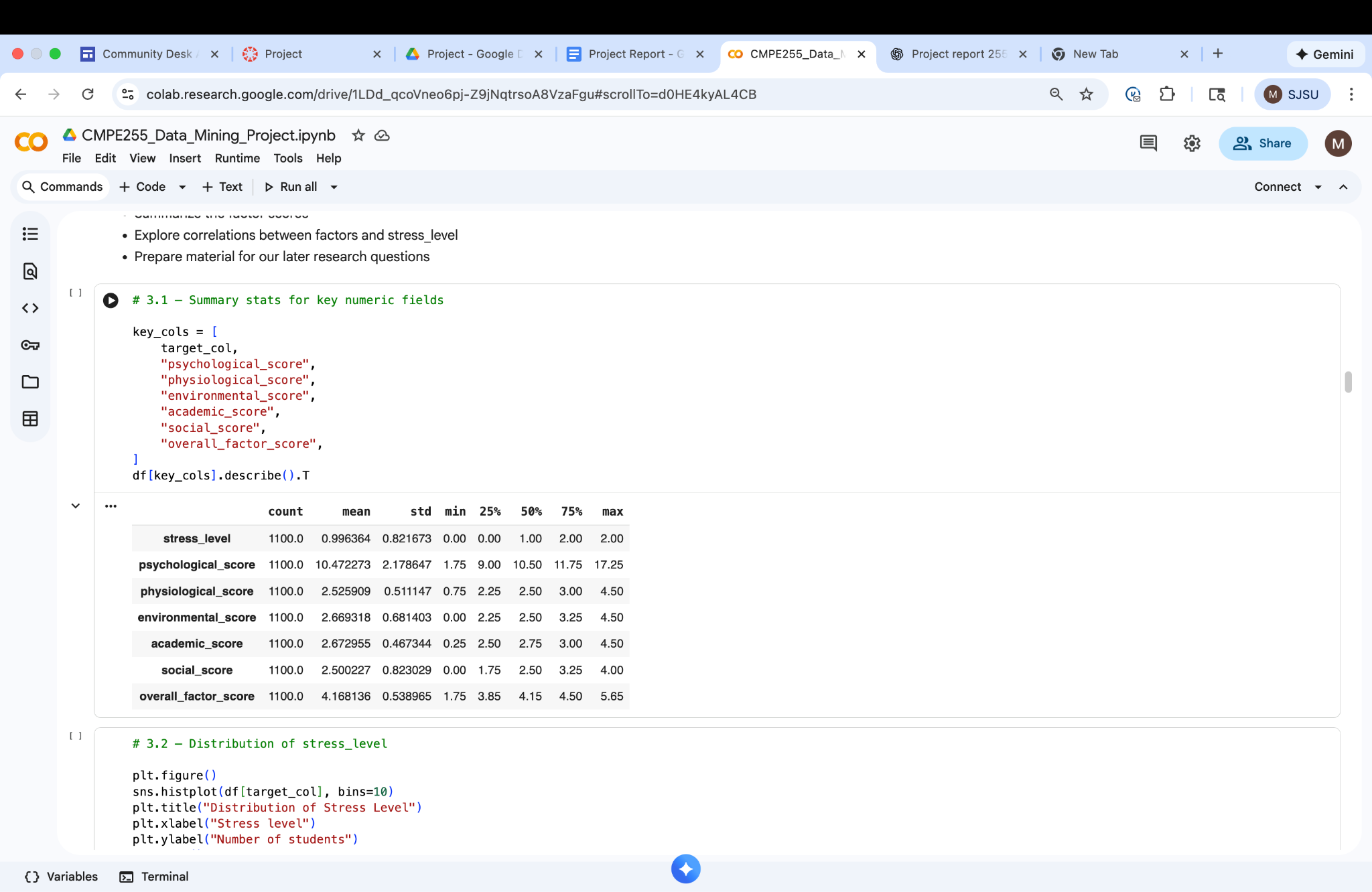
### **3.3 Duplicate Removal**

df = df.drop\_duplicates()

Result: there were no duplicates, Shape after dropping duplicates: (1100, 21)

### **3.4 Feature Engineering**

We grouped variables into five thematic categories and created composite scores:

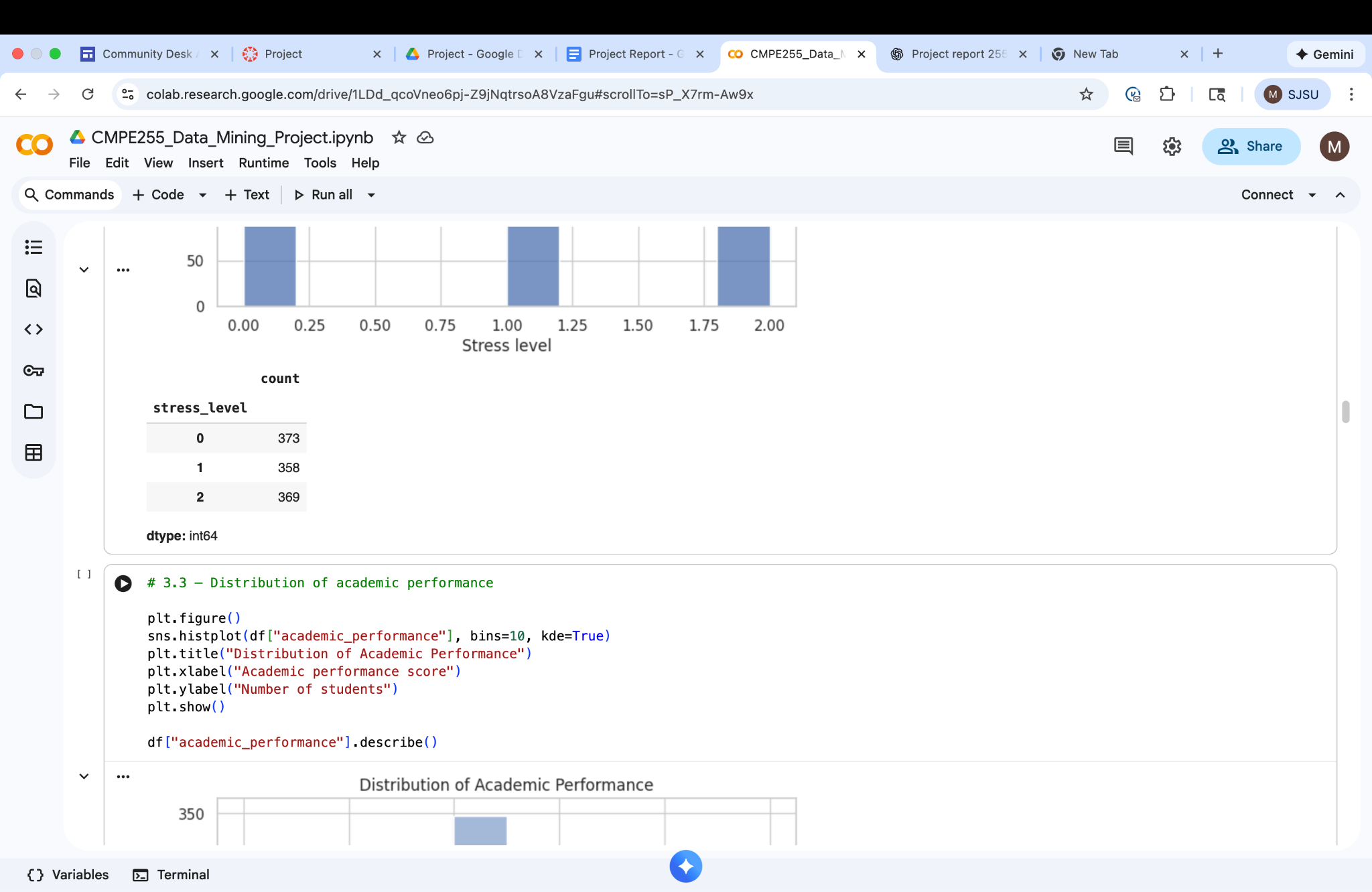


A final **overall\_factor\_score** was created by averaging all non-target features.

# 

# **4. Exploratory Data Analysis (EDA)**

## **4.1 Distribution of Stress Levels**



## 

## 

## **4.2 Correlation Analysis**

A correlation heatmap helps visualize relationships between engineered factor scores and stress level.



**Key insights:**

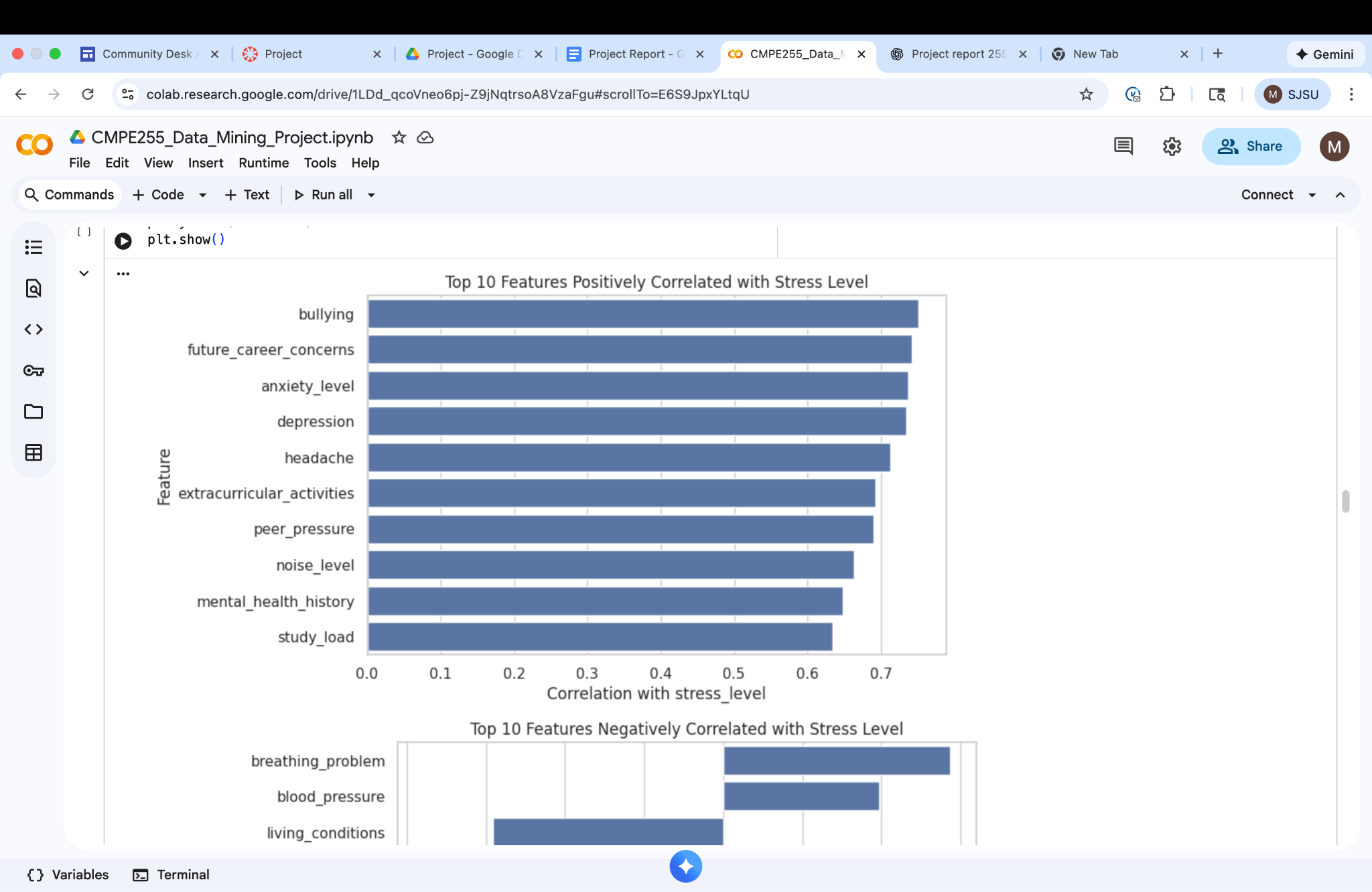
* **Academic pressure**, **sleep quality**, and **study load** show the strongest correlations with stress.
* Psychological and social support features have moderate relationships.
* Environmental factors contribute but with lower strength.

## 

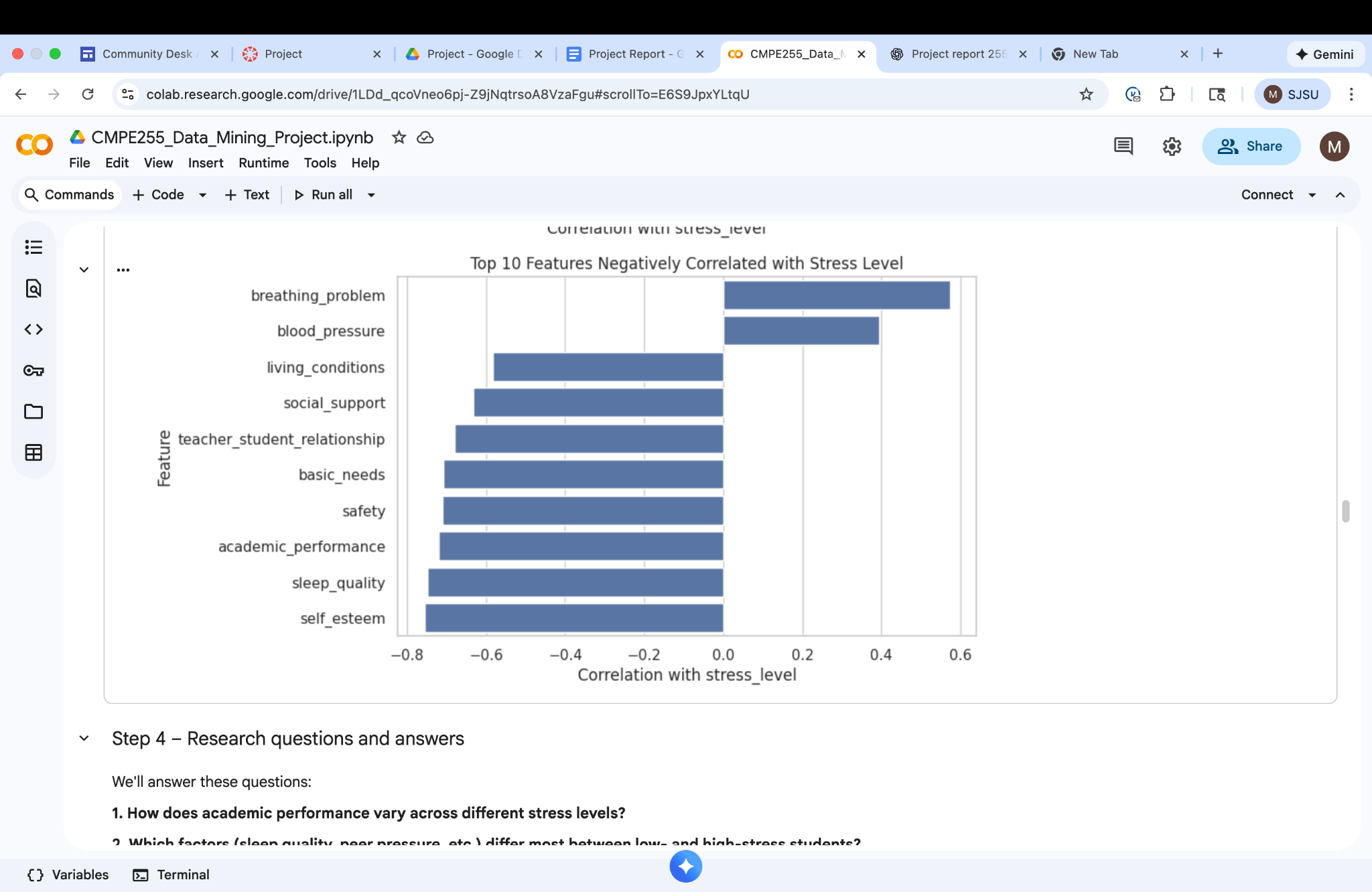
## 

## **4.3 Feature-Level Correlations**

### **Top 10 strongest positive correlations with stress**



### **Top 10 strongest negative correlations**

****

# **5. Research Questions & Answers**

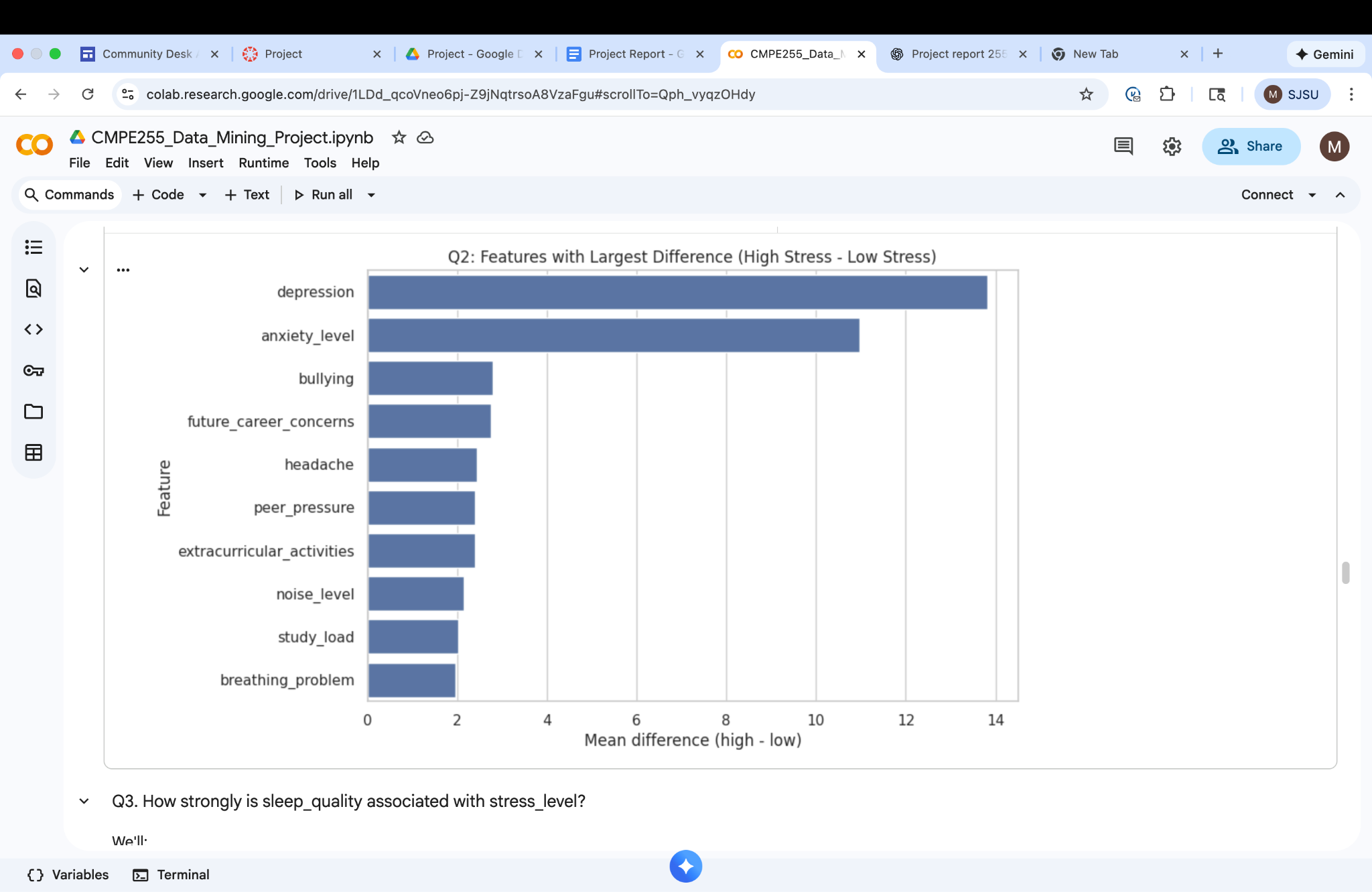
We addressed **6 major questions**.

## **Q1. How does academic performance vary across different stress levels?**

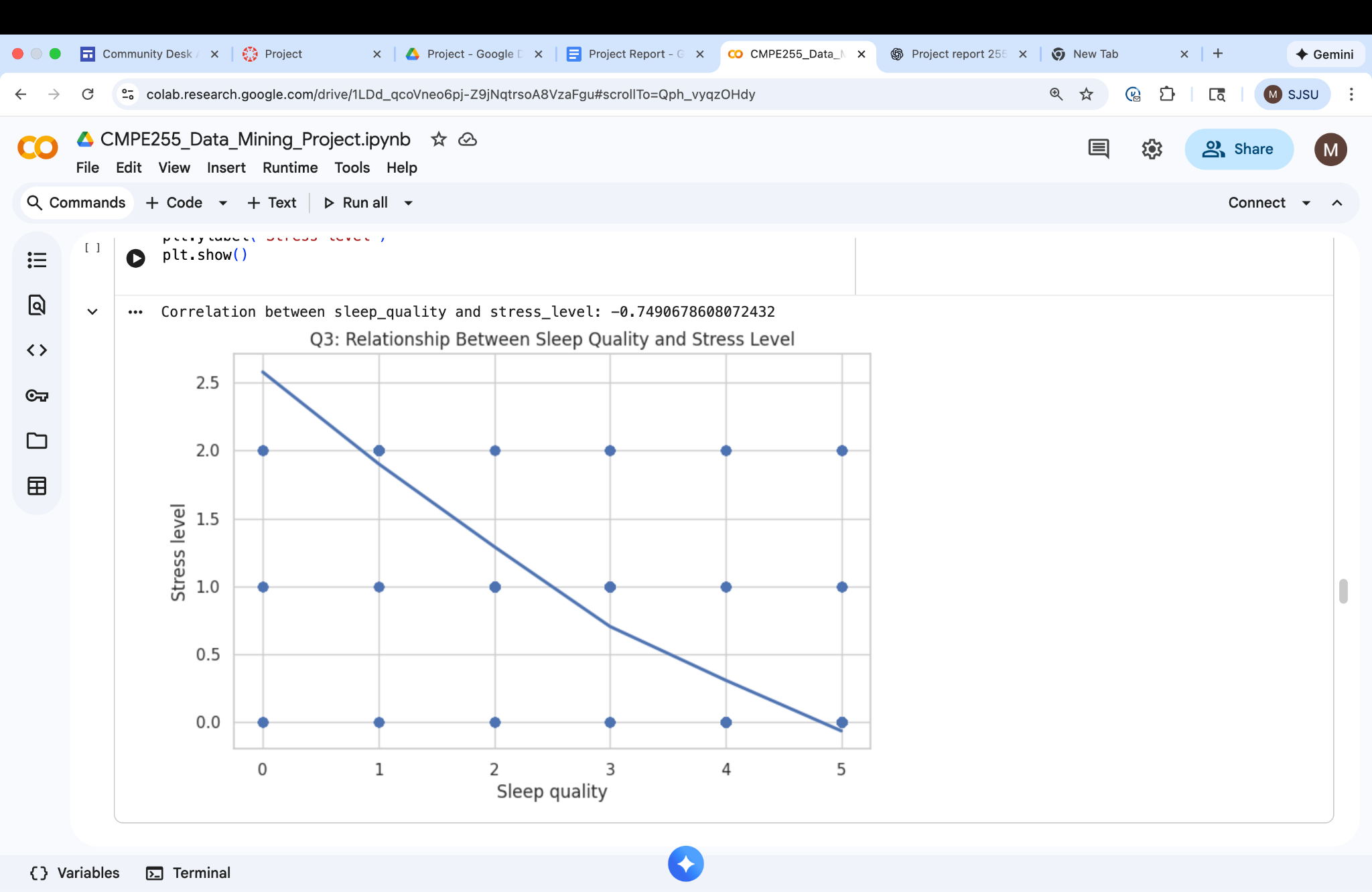
## 

## **Q2. Which factors differ most between low- and high-stress students?**

We compared the mean value of each feature for low vs high stress\_group and see which differences are largest.



## **Q3. How strongly is sleep quality associated with stress?**

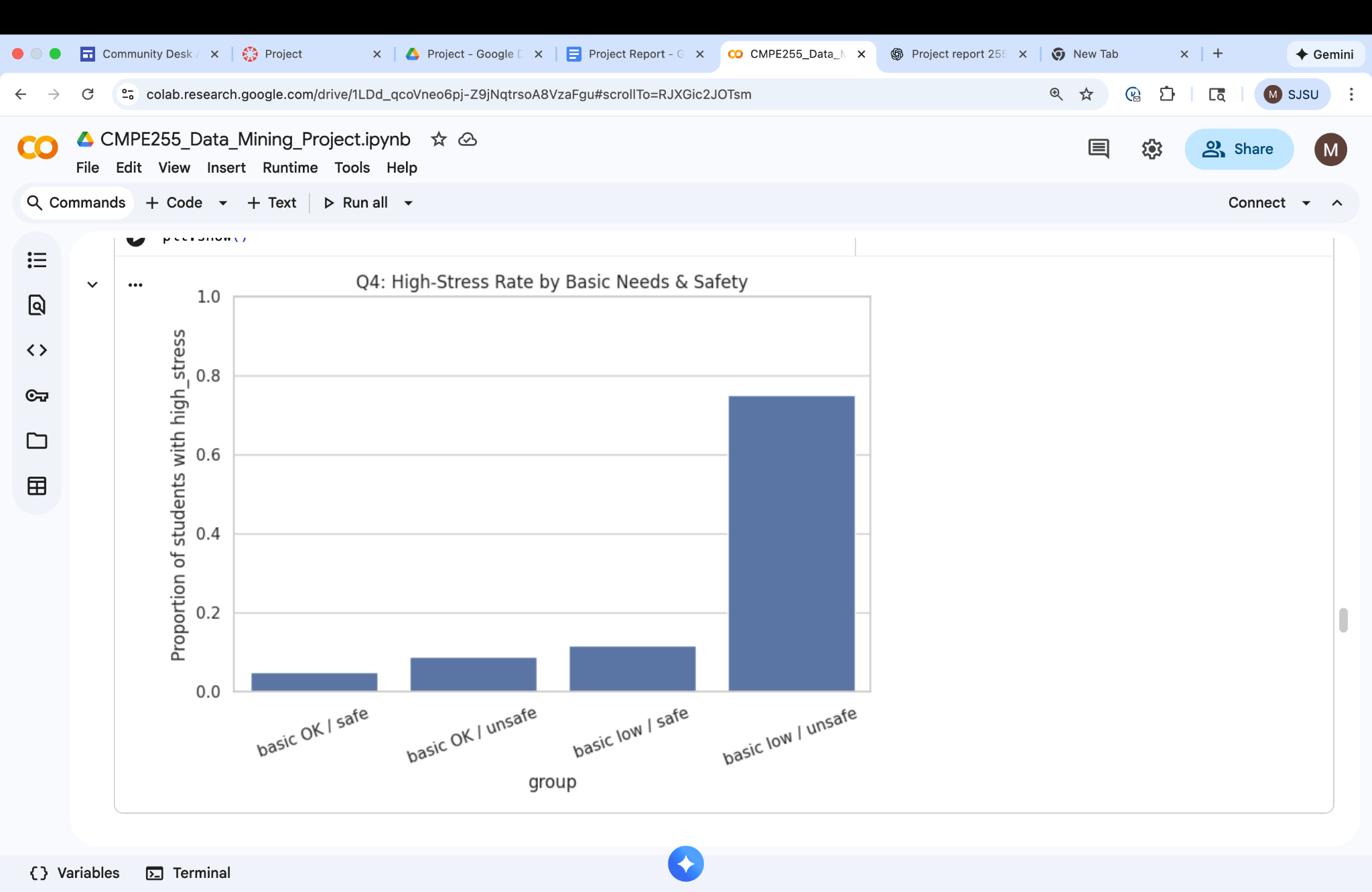


Correlation between sleep\_quality and stress\_level: -0.7490678608072432

So we can interpret that students with poor sleep quality have higher stress.

## **Q4. Are students with poor basic\_needs or safety more likely to have high stress?**

We split students into "low" vs "high" groups based on the mean of basic\_needs and safety, then compare their average stress\_level and proportion of high\_stress.



## 

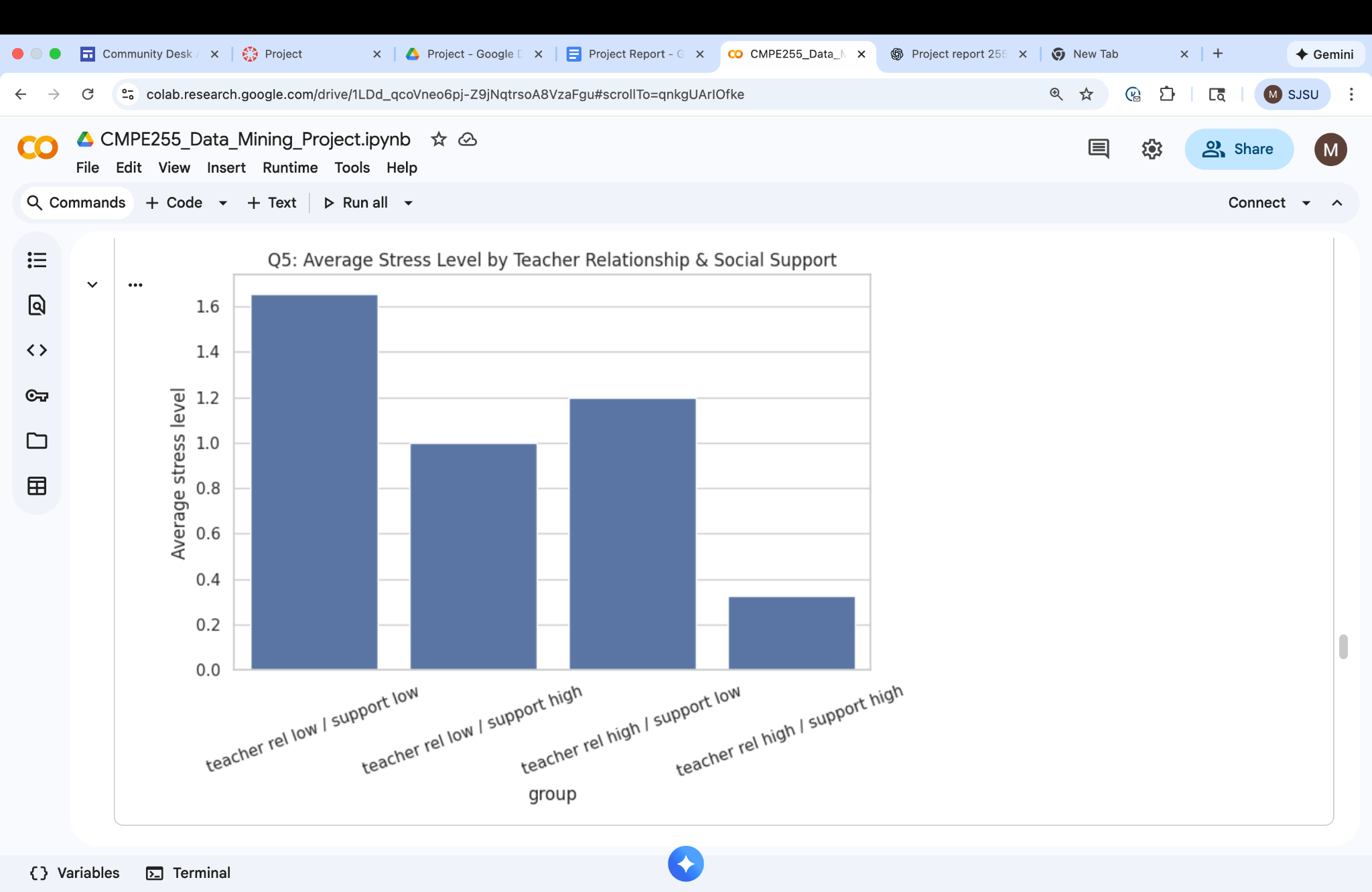
## 

## 

## 

## 

## **Q5. Do teacher relationships and social support buffer stress?**



Students with strong teacher relationships + strong social support show the lowest average stress.

## 

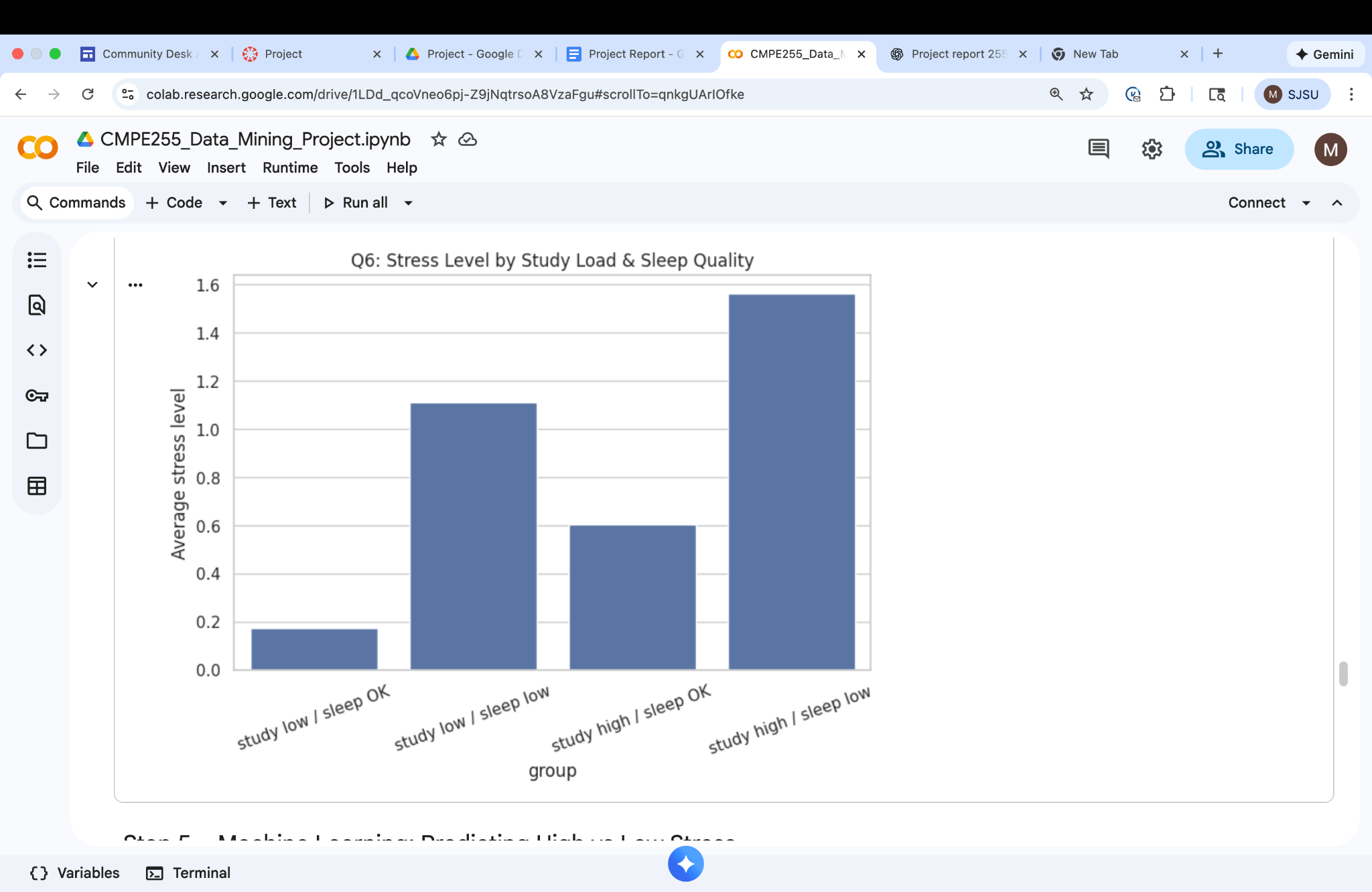
## 

## 

## 

## 

## **Q6. How do study load and sleep quality affect stress?**

****

# **6. Machine Learning Prediction**

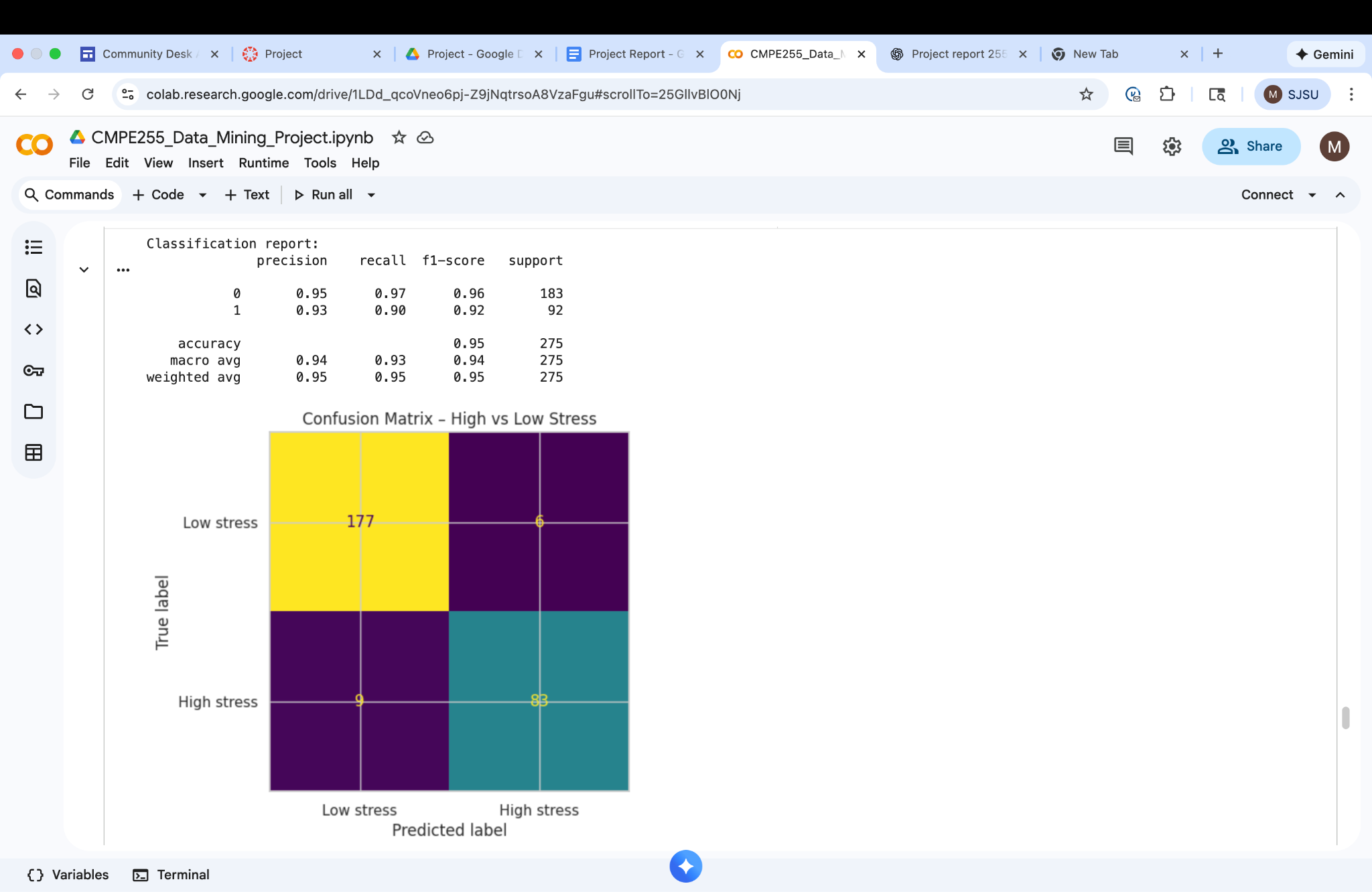
We built a binary classifier:

* **Target:** high\_stress (1 if stress\_level in top 33%, else 0)
* **Features:** all original factor variables + factor scores
* **Model:** Logistic Regression with:
  + median imputation for missing values
  + standardization (z-score scaling)

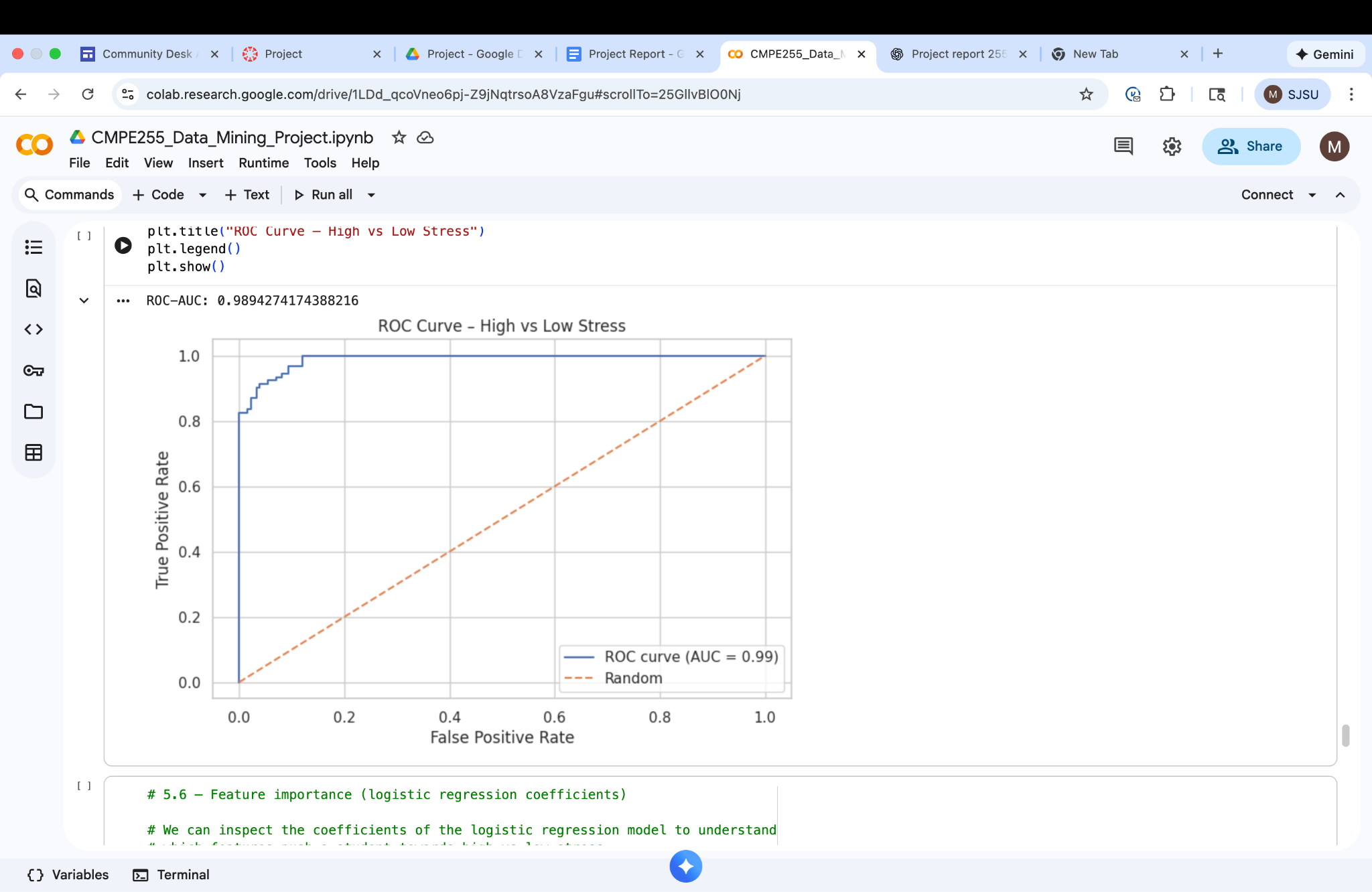
We:

* Split into train/test sets
* Fit the model
* Evaluate using accuracy, precision, recall, F1, ROC-AUC
* Look at model coefficients to see influential features

### 



**ROC Curve:**

****

# **7. Main Findings:**

