Muhsin Efe Tarım – Phase II

**Section 1 -- Introduction:**

Depression is a pervasive mental health concern that affects students’ academic performance, well‑being, and overall quality of life. Despite growing awareness, much of our understanding remains anecdotal, highlighting the need for rigorous, data‑driven insights. In this project, we leverage two complementary student survey datasets to quantify how socio‑economic and behavioral factors—such as academic achievement, financial support, and sleep habits—relate to depression severity.

Our first dataset (uni\_depression.csv) captures PHQ‑9 scores alongside demographic and academic variables like CGPA ranges and scholarship status, enabling us to investigate how educational performance and financial aid correlate with mental health outcomes. The second dataset (second\_Student\_Depression\_Dataset.csv) offers a broader perspective, including numeric CGPA values, sleep duration categories, and additional stress indicators (e.g., work pressure, financial stress) to validate and extend findings across different student populations.

Through exploratory visualizations—violin plots, boxplots, scatterplots, and bar charts—we will illuminate distributional patterns and associations. Building on these insights, we will apply hypothesis tests (two‑sample t‑tests, chi‑square tests, and correlation analyses) to assess the statistical significance of observed relationships and determine which factors most strongly predict depression.

This report is structured as follows: Section 2 details the datasets; Section 3 presents exploratory data analysis and visualizations; Section 4 is for hypotheses.

**Section 2 -- datasets:**

We leverage two complementary student survey datasets to examine predictors of depression:

*uni\_depression.csv*

• Description: University‑specific survey capturing PHQ‑9 depression scores alongside demographic and academic variables.

**Key fields:**

• Depression Value – total PHQ‑9 score

• Current CGPA – binned CGPA ranges (e.g., “2.5–3.0”)

• Did you receive a waiver or scholarship at your university? – indicator of financial aid

*second\_Student\_Depression\_Dataset.csv*

• Description: Broader survey of student well‑being, including academic, lifestyle, and mental health measures.

**Key fields:**

• CGPA – numeric cumulative GPA

• Sleep Duration – categorical sleep ranges (e.g., “5–6 hrs”)

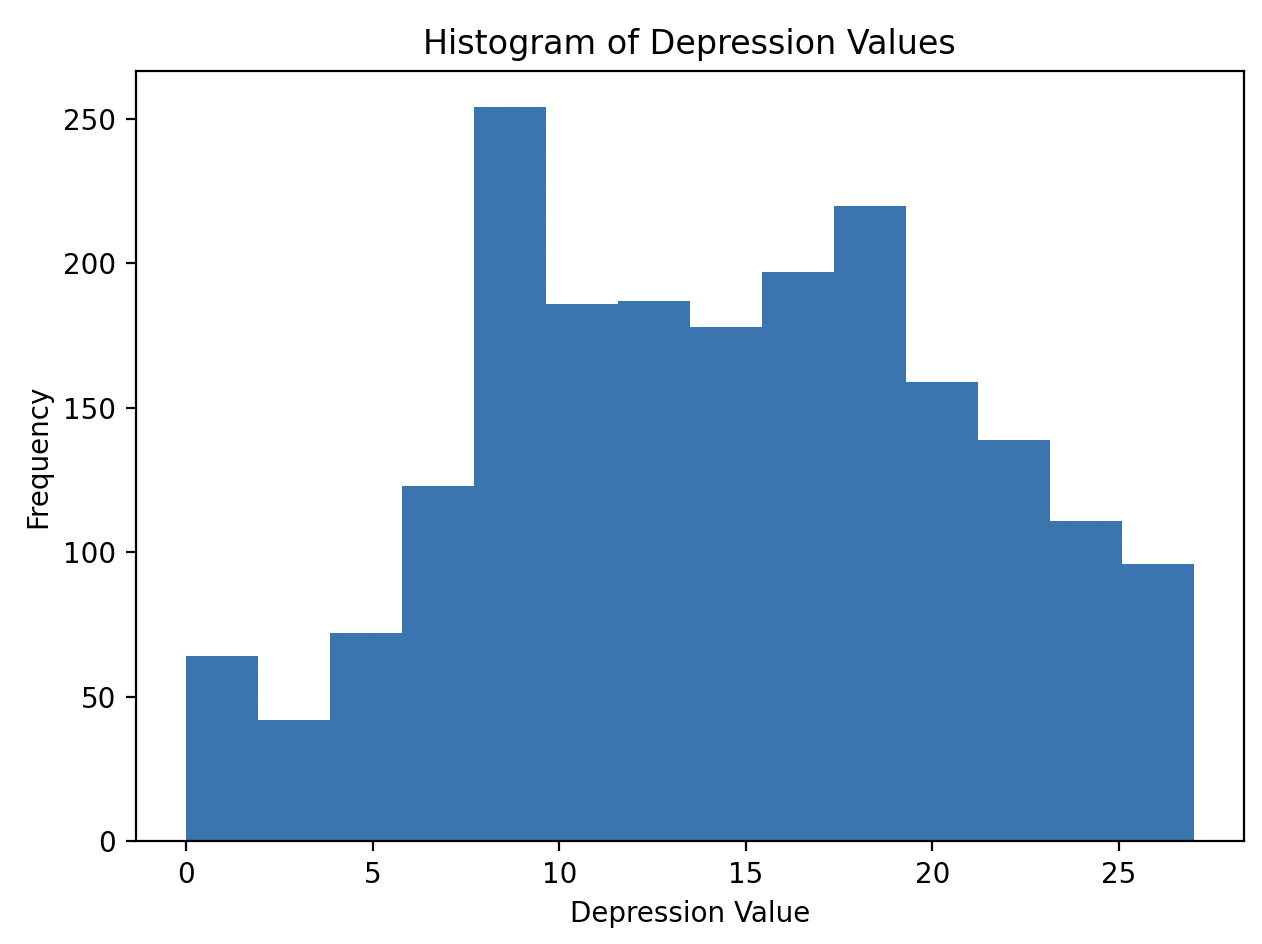
• Depression – depression score analogous to PHQ‑9

• Additional socio‑economic and stress indicators such as Work Pressure and Financial Stress

**Section 3 – Exploratory Data Analysis and Visualizations:**

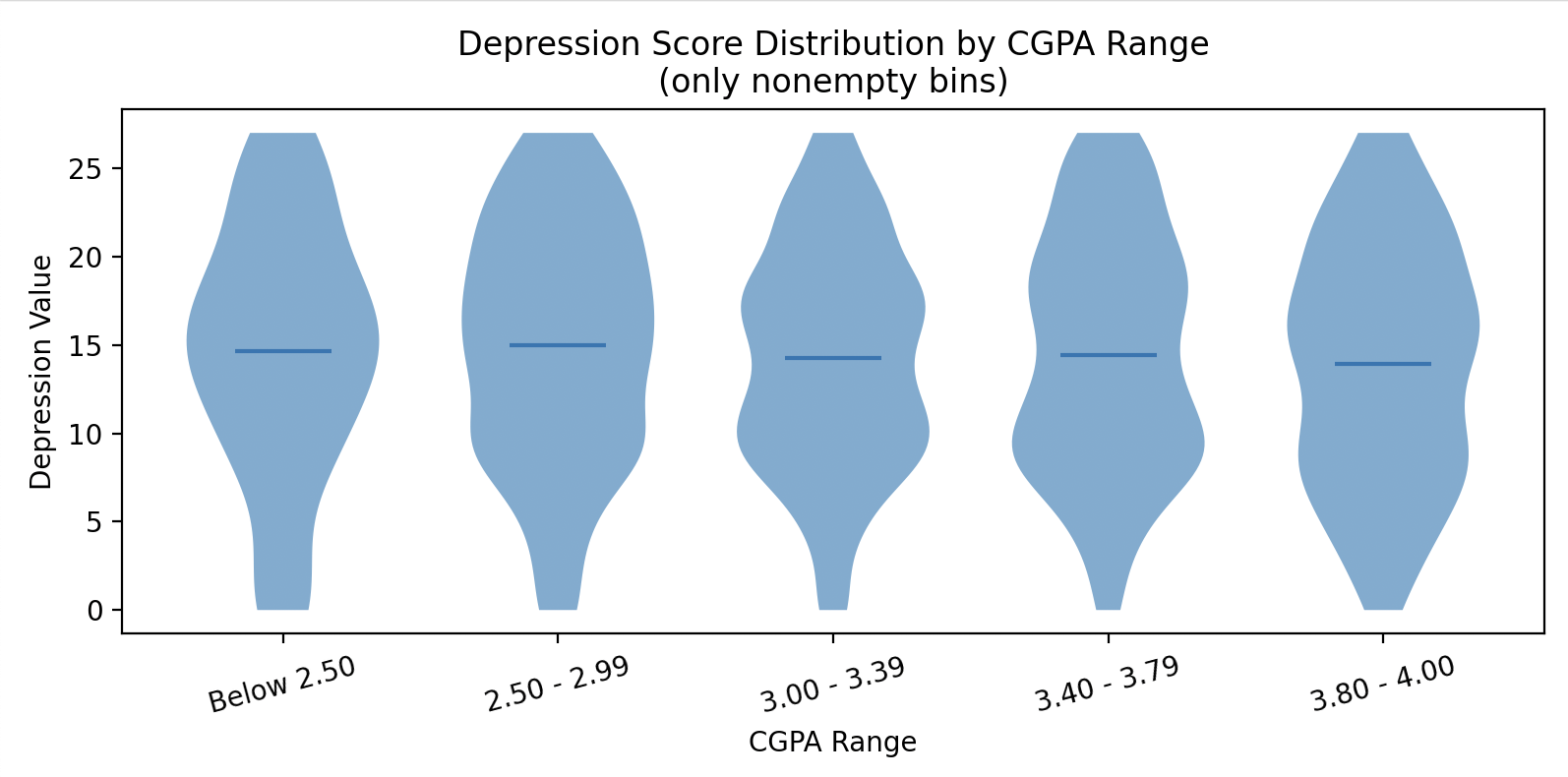
*for uni\_depression.csv*

**1) Histogram of Depression Values:**



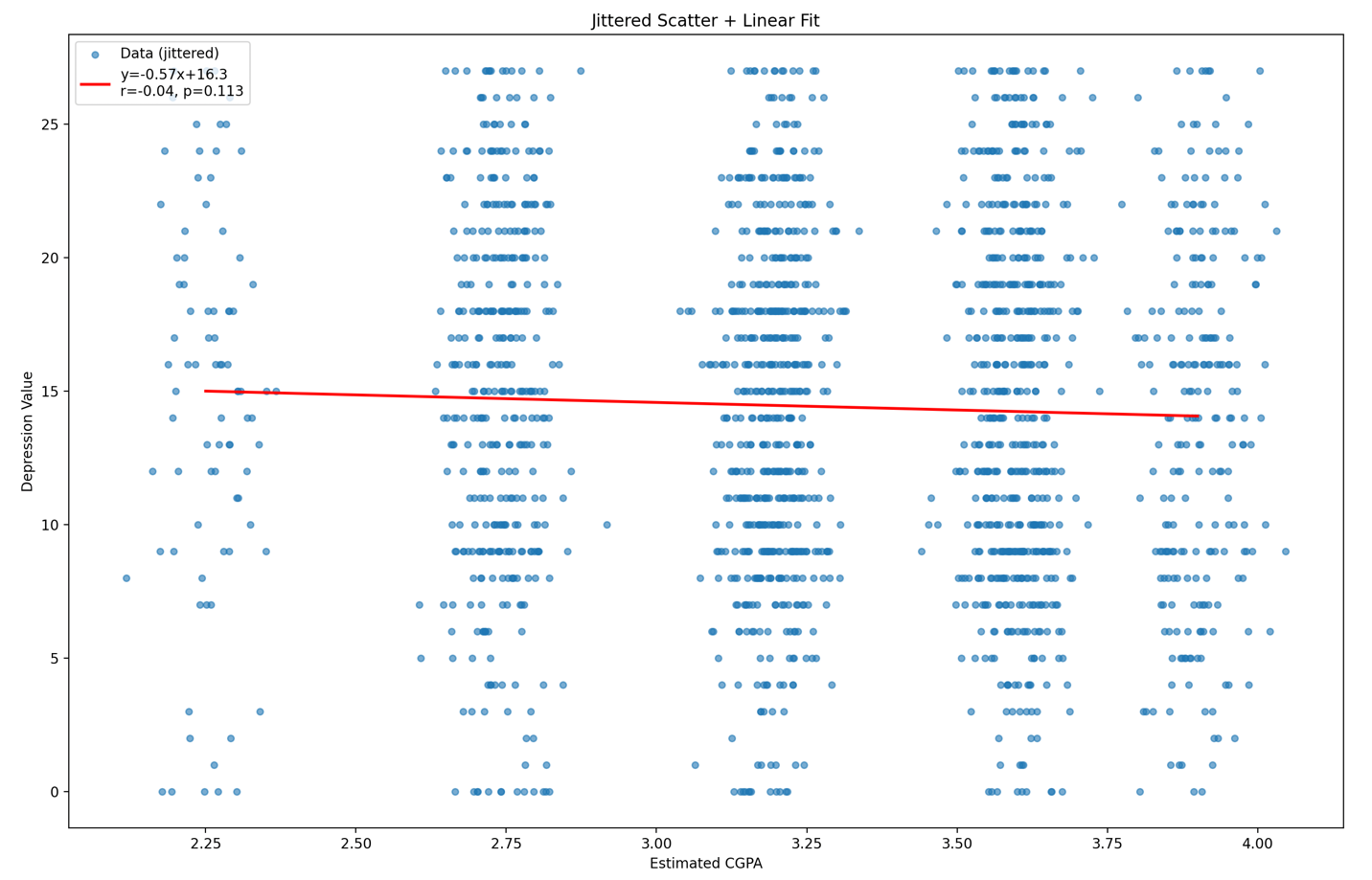
**Understanding:** The histogram reveals that most students’ PHQ‑9 scores fall between 8 and 20, indicating a concentration in the mild‑to‑moderate range. A noticeable right skew (long tail toward higher scores) shows that while severe depression is less common, it still affects a significant minority. The relative frequency of very low scores (0–4) is small, suggesting that few students report minimal symptoms. This overall profile establishes the baseline distribution against which subgroup comparisons can be made.

**2) Violin Plot of Depression by CGPA Range:**

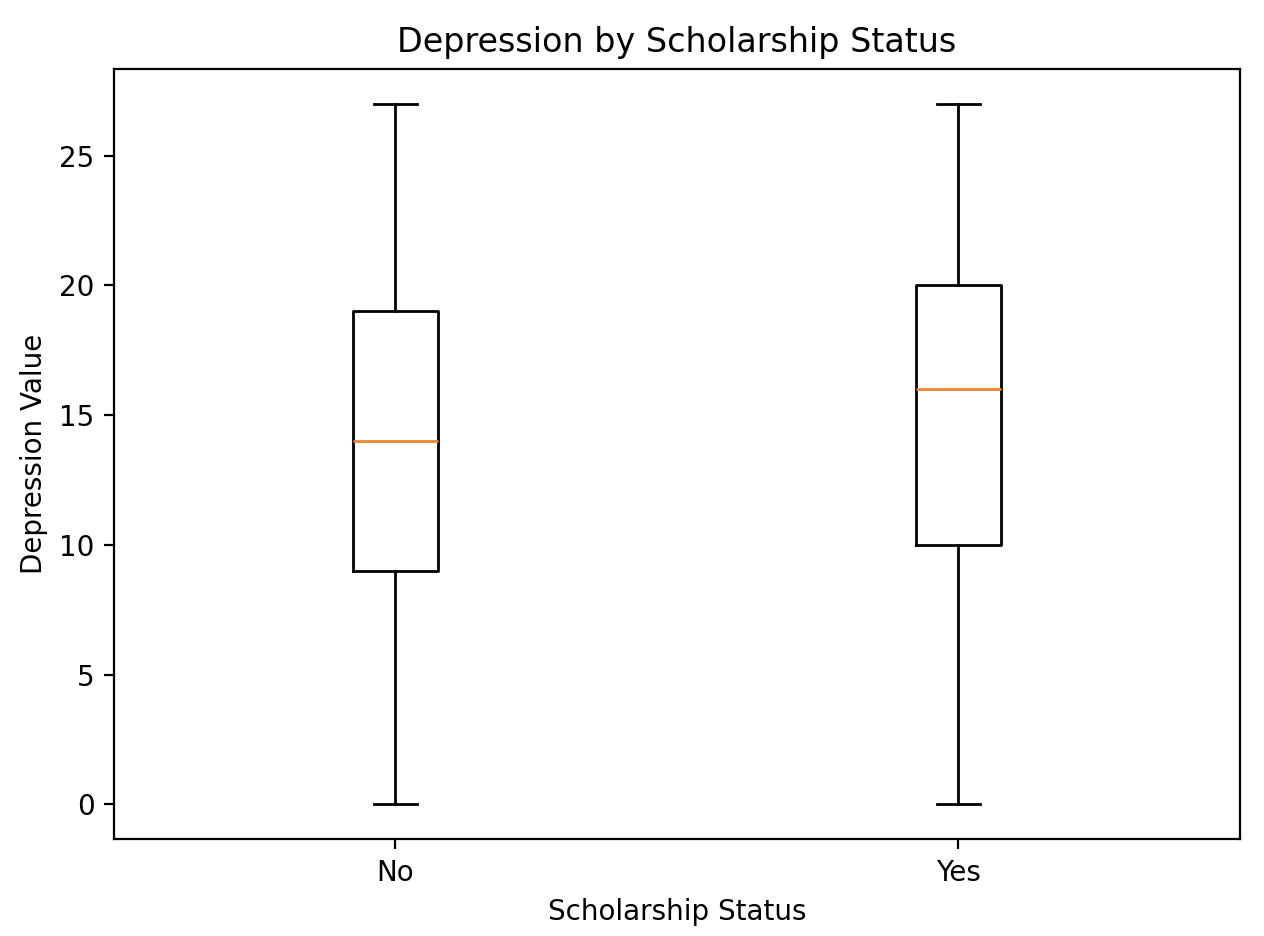


**Understanding:** The violin plot overlays density estimates for each GPA bin, highlighting both central tendency and distributional spread. Students in the “Below 2.50” and “2.50 – 2.99” categories exhibit slightly wider tails toward higher scores, suggesting more variability and occasional severe symptoms in lower‑GPA groups. In contrast, the “3.80 – 4.00” bin appears somewhat narrower, with its median shifted lower, hinting at a modest protective effect of higher academic performance. However, overlap among all bins indicates that CGPA alone is not a strong separator of depression levels.

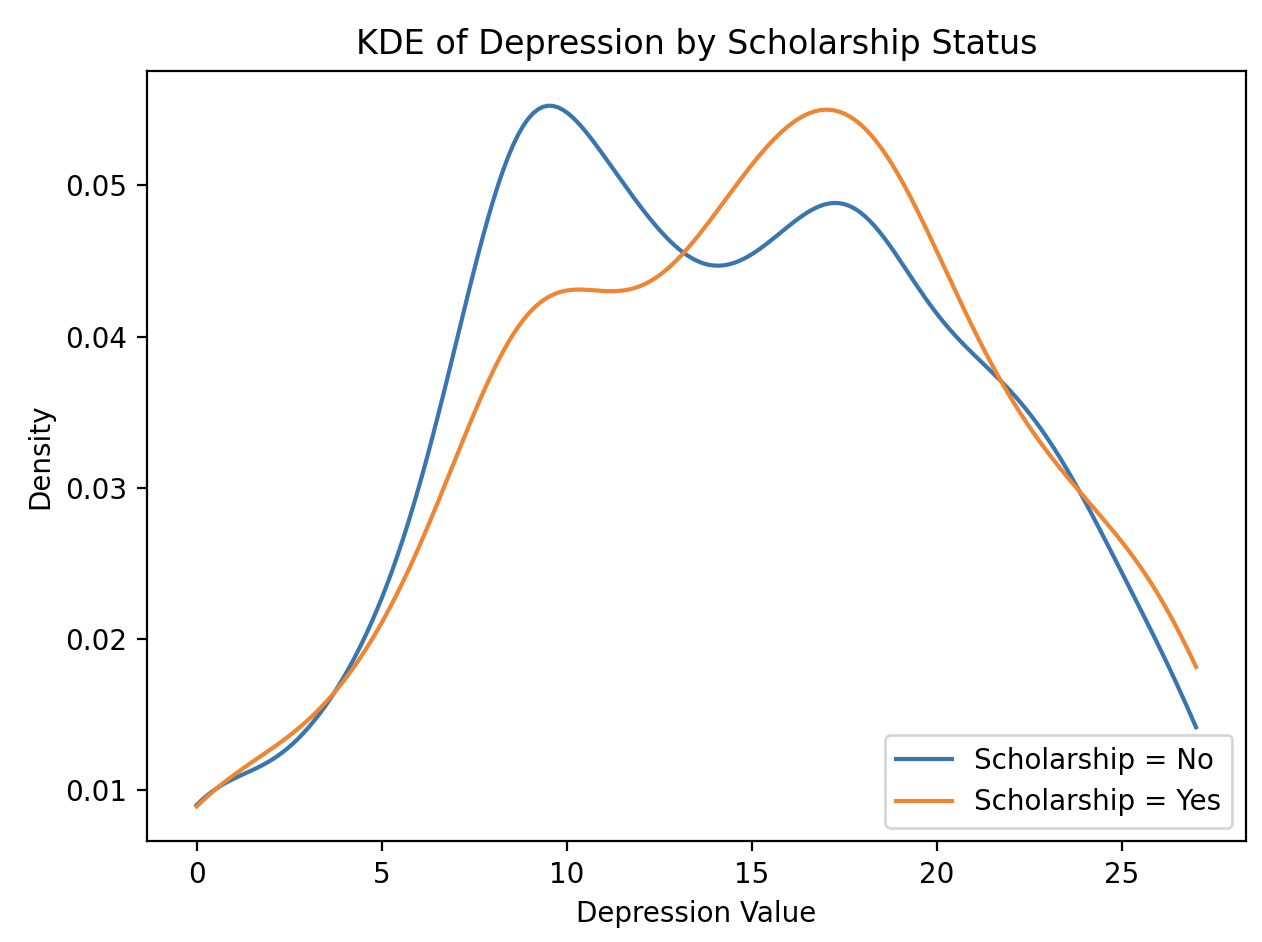
**3) Jittered Scatterplot with Linear Fit (Estimated CGPA vs. Depression):**



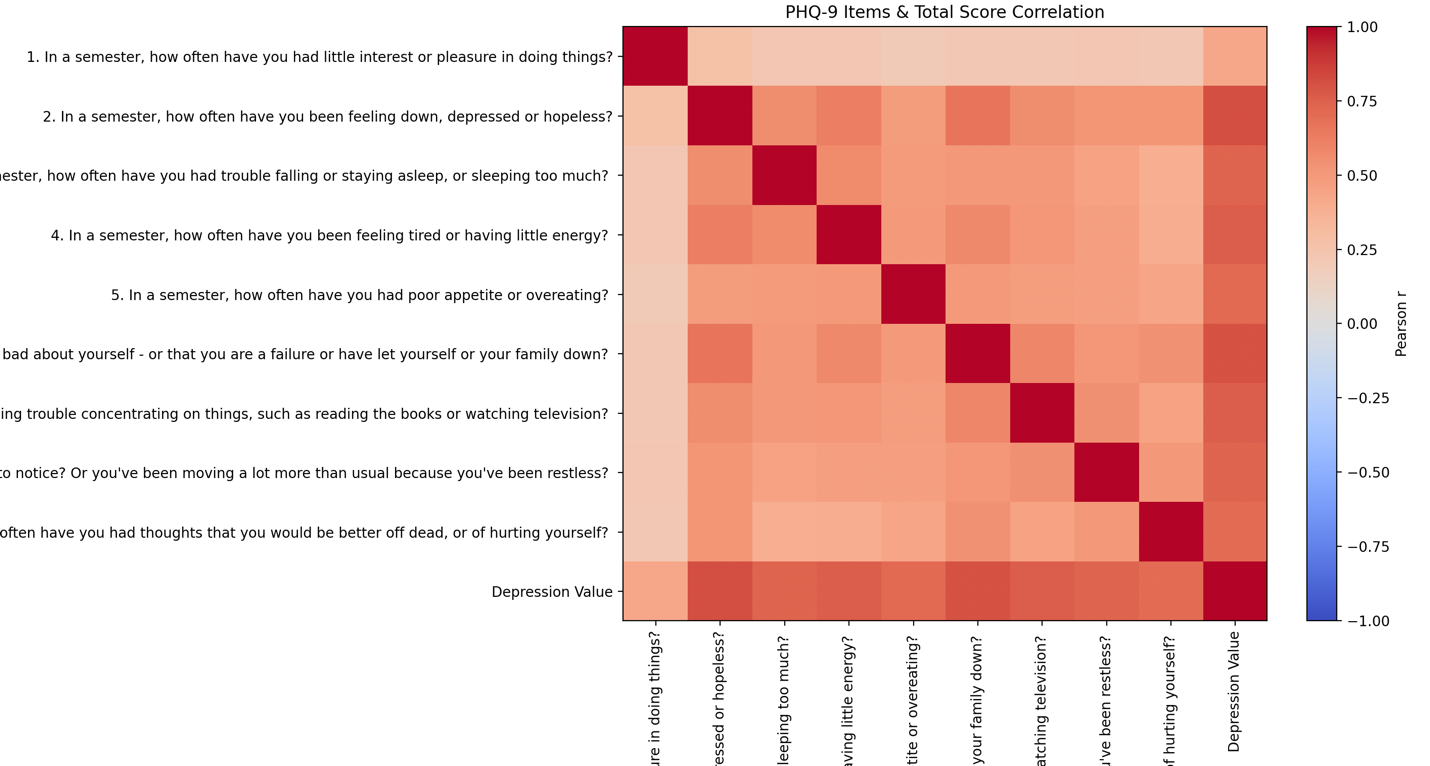
**Understanding:** Plotting individual points (with slight horizontal jitter for visibility) alongside a best‑fit regression line quantifies the CGPA–depression relationship. The fitted line has a small negative slope (≈ –0.6 depression points per 1 CGPA) and a Pearson correlation of r ≈ –0.04 (p = 0.11), confirming a very weak, non‑significant negative association. In plain terms, higher GPAs are marginally—but not reliably—associated with lower depression scores. The cloud of points underscores the high degree of scatter and suggests many other factors influence students’ well‑being.

**4) Boxplot of Depression by Scholarship Status:**

**Understanding:** Comparing students with versus without a scholarship reveals that scholarship recipients actually show a higher median PHQ‑9 score (around 16 vs. 14) and a slightly larger interquartile range. This counterintuitive result may reflect additional pressures faced by scholarship holders—such as maintaining performance standards—or underlying financial stress not alleviated by aid. Whisker lengths and outliers indicate that extreme scores appear in both groups, so scholarship status alone does not capture the full story.

**5) Kernel Density Estimate (KDE) of Depression by Scholarship Status:**

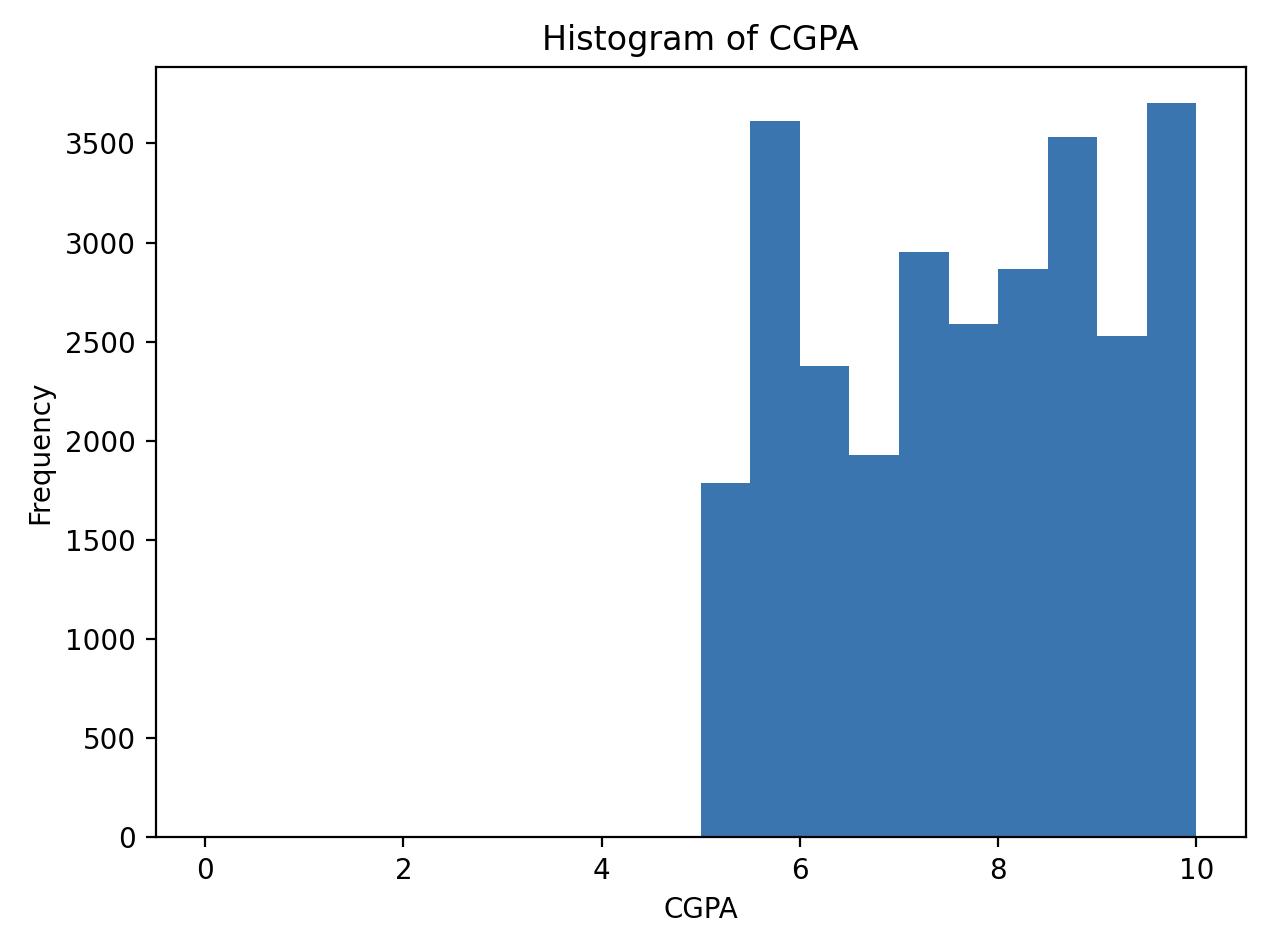
**Understanding:** The KDE plot smooths each group’s score distribution, making subtle differences more visible. The density curve for scholarship recipients is shifted rightward by 1–2 points relative to non‑recipients and shows a slightly sharper peak around moderate symptoms (15–18). In contrast, non‑recipients have a broader shoulder toward lower scores (8–12). These nuanced shifts corroborate the boxplot findings and suggest that financial aid intersects with mental health in complex ways—potentially mediated by performance expectations or conditionality.

**6) Heatmap of PHQ‑9 Item‑to‑Total Correlations**

**Understanding:** The heatmap displays Pearson correlations among each PHQ‑9 item and between items and the total depression score. Items assessing “feeling down, depressed, or hopeless” and “little interest or pleasure in doing things” exhibit the strongest correlations with the total score (r > 0.70), indicating they are core drivers of overall depression severity. By contrast, somatic items like “poor appetite or overeating” correlate more weakly (r ≈ 0.45), suggesting they tap into a distinct symptom cluster. This analysis confirms the internal consistency of the PHQ‑9 while highlighting which symptoms most heavily influence composite scores.

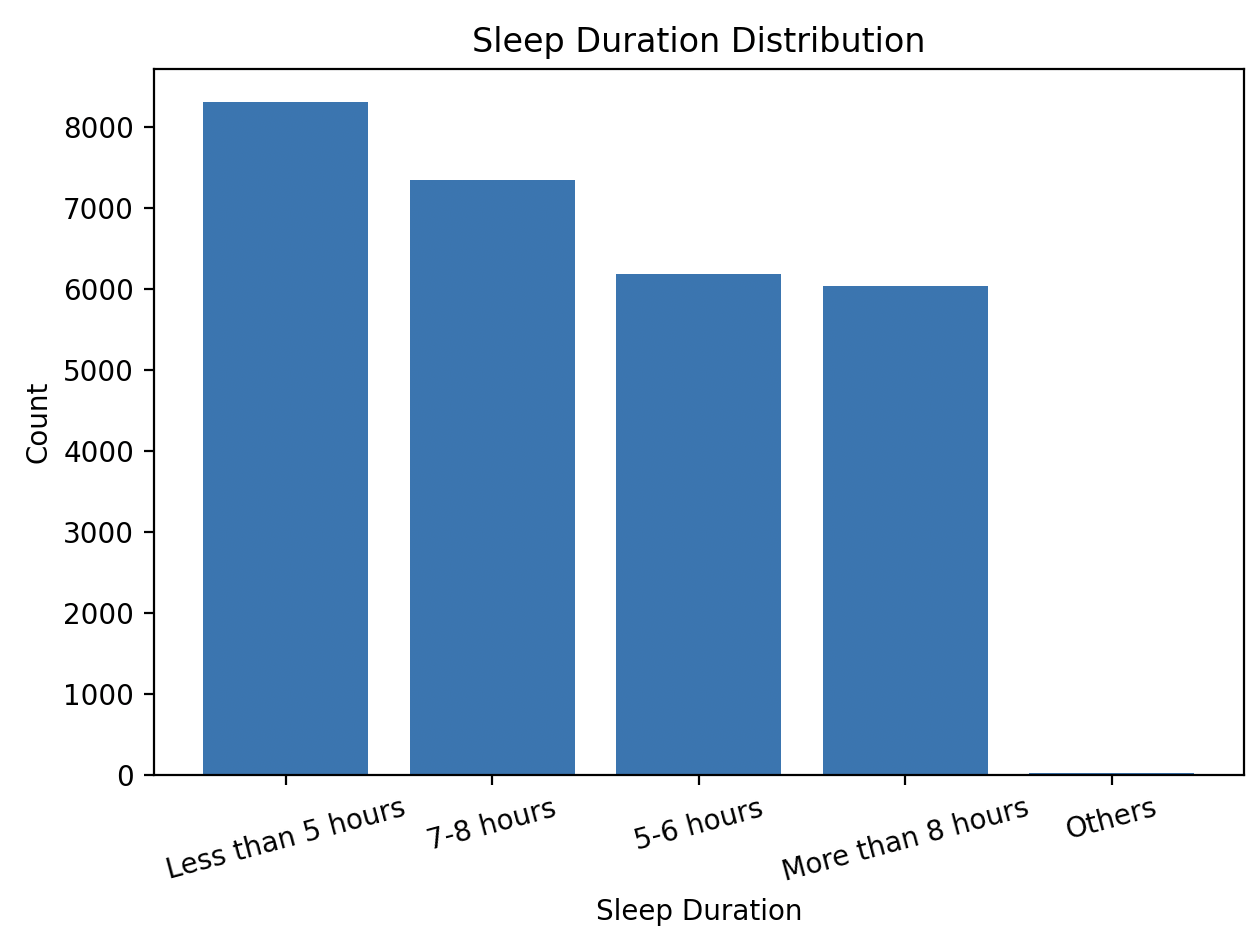
*for second student depression dataset*

**1) Histogram of CGPA:**



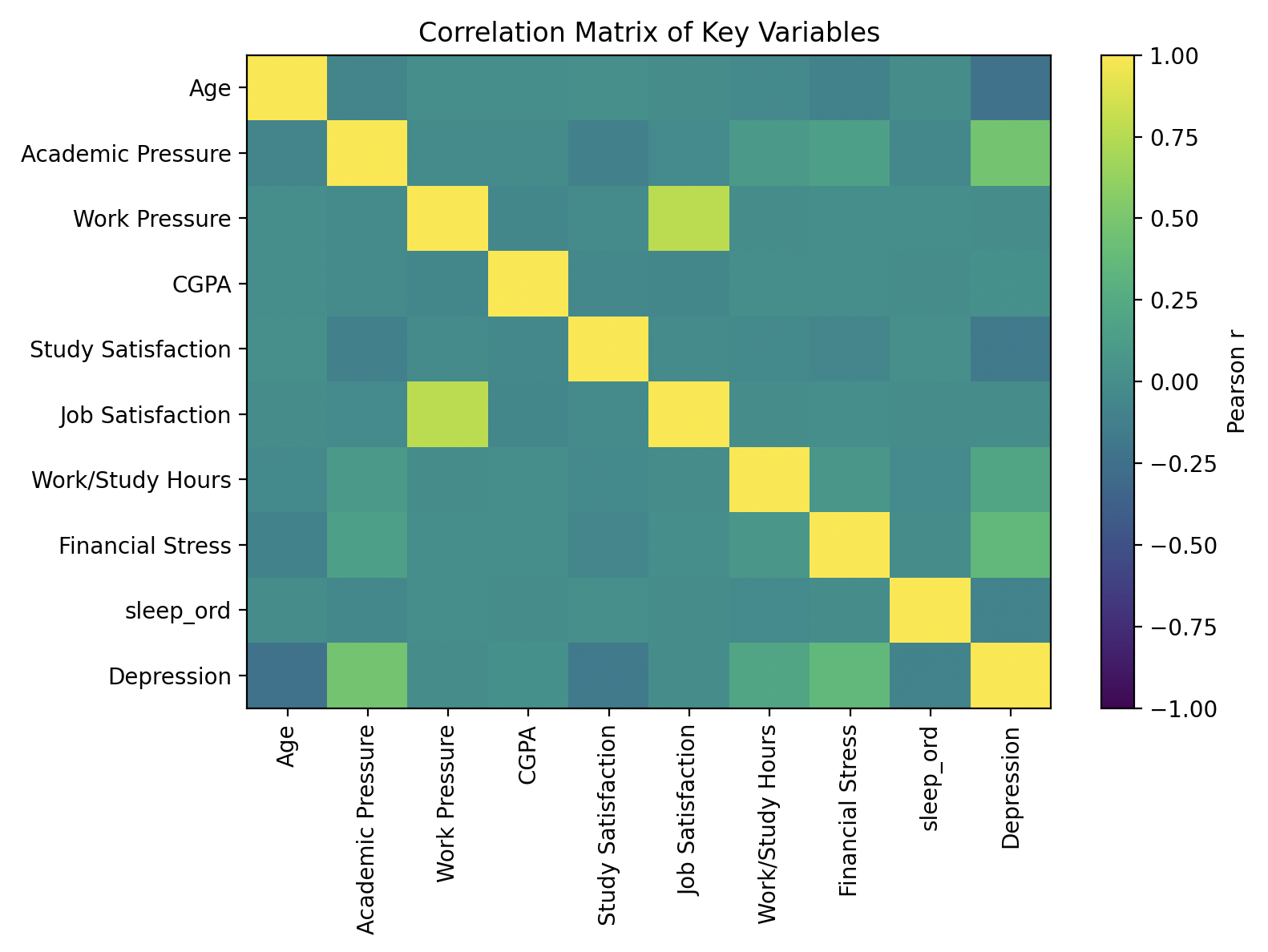
**Understanding:** The numeric CGPA distribution is heavily right‑skewed, with the majority of students clustering between 8.0 and 10.0. A smaller tail extends downward to about 5.0, indicating relatively few low‑GPA respondents. This concentration at higher GPAs suggests grade inflation or a high‑achieving sample, and sets expectations for interpreting CGPA’s role in subsequent depression analyses.

**2) Sleep Duration Distribution:**



**Understanding:** Sleep patterns are unevenly distributed: the largest group (~8,300) reports “Less than 5 hours”, followed by “7–8 hours” (~7,350), “5–6 hours” (~6,200), and “More than 8 hours” (~6,000). This skew toward very short sleep highlights widespread sleep deprivation among students and suggests sleep duration may be a key behavioral predictor to explore against depression scores.

**3) Correlation Matrix of Key Variables:**



**Understanding:** This heatmap of Pearson r values reveals a complex web of associations:

• Financial Stress (r ≈ +0.50)) shows the strongest positive correlation with depression, indicating that economic burdens closely track symptom severity.

• Academic Pressure and Work/Study Hours also correlate moderately (r ≈ +0.30–0.40), suggesting that high demands are linked to poorer mental health.

• Sleep Ordinal (short → long) correlates negatively (r ≈ –0.20), reinforcing that longer sleep tends to accompany lower depression scores.

**4) Depression Rate by CGPA Bin:** A graph with a line and a blue line

AI-generated content may be incorrect.

**Understanding:** Grouping students into CGPA bins and plotting the proportion meeting a clinical depression threshold uncovers a non‑monotonic pattern:

• Rates rise from ~56% at a bin midpoint of 5.5 to ~60% at 6.5, dip slightly at 7.5, then peak around 64% at 8.5 before settling near 58% at 9.5.

• Overlapping error bars (95% CI) indicate these fluctuations are not statistically distinct, so we conclude that CGPA alone does not reliably predict depression prevalence.

**5) CGPA Distribution by Depression Status:**

A graph of a graph with lines and a line in the middle

AI-generated content may be incorrect.

**Understanding:** Finally, comparing CGPA among students above versus below the depression cutoff shows nearly identical boxplots:

• Depressed students have a median CGPA of ~7.8 vs. ~7.6 for non‑depressed, with similar interquartile ranges and whisker lengths.

• The heavy overlap underscores that academic performance—at least as measured by CGPA—is not a strong discriminator of depression status in this larger sample.

**Section 3 – Hypotheses and Statistical Tests:**

Building on insights from exploratory visualizations, we conduct formal hypothesis tests to assess the statistical significance of observed patterns. These tests help confirm or refute visual trends and identify factors that may meaningfully predict depression scores.

**4.1 Academic Performance and Depression**

***a)*** *Depressed vs. Not Depressed: Mean CGPA Comparison (t-test)*

H₀: Mean CGPA of depressed students equals that of non-depressed students

H₁: Mean CGPA differs between the two groups

• Result: t = 3.69, p = 0.0002

• Conclusion: Since p < 0.05, we reject H₀. Depressed students exhibit significantly lower average CGPA, although the effect size is small.

• Note: While the t-test shows significance, visualizations (e.g., jittered scatterplot, boxplots) suggest high overlap and weak practical association (r ≈ –0.04). Academic performance statistically correlates, but not strongly.

**4.2 Sleep Duration and Depression**

***a)*** *Correlation (Spearman Rank): Sleep Ordinal vs. Depression Score*

H₀: No monotonic association between sleep and depression

H₁: Sleep and depression are monotonically associated

• Result: ρ = –0.087, p ≪ 0.001

• Conclusion: Reject H₀. There is a weak negative association: longer sleep correlates slightly with lower depression.

***b)*** *Oneway ANOVA: Depression Score Across Sleep Groups*

H₀: All sleep groups have the same mean depression score

H₁: At least one group differs

• Result: F = 93.01, p ≪ 0.001

• Conclusion: Reject H₀. Depression scores differ across sleep categories. Combined with the skewed sleep distribution, this supports sleep as a meaningful factor.

**4.3 Family History of Mental Illness**

***a)*** *Chi-Square Test: Family History vs. Depression Status*

H₀: Depression is independent of family history of mental illness

H₁: Depression is associated with family history

• Result: χ² = 79.43, p ≪ 0.001

• Conclusion: Reject H₀. Students with a family history of mental illness are significantly more likely to be depressed.

**4.4 CGPA and Depression (Linear Correlation)**

a) Pearson Correlation: Estimated CGPA vs. Depression

H₀: No linear correlation between CGPA and depression score

H₁: CGPA is linearly correlated with depression

• Result: r = –0.037, p = 0.113

• Conclusion: Fail to reject H₀. No statistically significant linear relationship was found in the broader dataset, reinforcing that CGPA is a poor standalone predictor.

**4.5 Financial Support and Depression**

a) Two-Sample t-Test: Scholarship vs. No Scholarship

H₀: Depression scores are equal for students with and without scholarships

H₁: Depression scores differ by scholarship status

• Result: t = 1.82, p = 0.068

• Conclusion: Fail to reject H₀. No statistically significant difference found, despite visual patterns showing slightly higher depression among scholarship recipients.

**Section 5 – Conclusion**

In this report, we combined exploratory visualizations with formal hypothesis tests to investigate factors associated with student depression. These statistical tests—such as t-tests, ANOVA, chi-square, and correlation analyses—allowed us to move beyond surface-level patterns and assess the reliability of our observations.

We found that students who sleep less tend to report significantly higher depression scores. This was supported by both the Spearman rank correlation (ρ = –0.087, p ≪ 0.001) and a highly significant ANOVA test (F = 93.01, p ≪ 0.001), confirming that sleep duration is a meaningful predictor.

The chi-square test (χ² = 79.43, p ≪ 0.001) also revealed a strong association between depression and family history of mental illness, emphasizing the role of background risk factors.

In contrast, while the t-test comparing CGPA between depressed and non-depressed students was statistically significant (p = 0.0002), the Pearson correlation across the full sample (r = –0.037, p = 0.113) showed no meaningful linear relationship—highlighting the limits of academic performance as a predictor.

Similarly, although visualizations suggested differences in depression by scholarship status, the two-sample t-test (p = 0.068) showed no significant effect.

Together, these hypothesis tests helped clarify which relationships are statistically supported and which may be misleading without rigorous validation.