Camille Lewis

INF 6490

Final Project

Analysis of Student Depression Data

Dataset and Research Questions

I chose a dataset titled *Student Depression Dataset* that is available on Kaggle here: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset

The *Student Depression Dataset* compiles information from 27,901 students, including data on their demographics, academics, lifestyle, and whether they experience depression. A full table of the variables and their definitions can be found in Appendix 1 of this document.

My primary objective of this project is to examine diverse variables of students and to identify which variables most strongly predict depression among students. My key research questions to accomplish this investigation are:

- How prevalent is depression among students, and how does it vary by demographics?
- How are academic pressure and academic performance related to depression?
- What lifestyle and health-related factors are correlated with depression among students?
- Which factors best predict depression among students?

Data Cleaning and Preprocessing

I checked the dataset for N/A values, and there were none. I decided to check to make sure that all numerical variables had numerical values. Since there is a large number of records, I decided to remove any records that did not pass this check. Three records were removed due to non-numerical values in the financial stress field.

Next, I encoded the categorical variables relevant to my research questions as numeric values that were not already encoded, using one-hot encoding for binary variables and label encoding for all others. These variables included sleep duration, dietary habits, suicidal thoughts, and family history of mental illness.

Descriptive Statistics and Data Visualization

Descriptive analyses were conducted to explore the distribution of depression and key demographic and academic variables among students. The sample included 27,880 complete cases after removing observations with missing data. Depression prevalence in the sample was approximately 41%, indicating that a substantial proportion of students reported experiencing depressive symptoms. Age differed noticeably by depression status, with students who were depressed being younger on average (≈ 24.9 years) than those who were not depressed (≈ 27.1 years). Academic pressure also showed meaningful differences between groups, with depressed students reporting significantly higher levels of academic pressure compared to their non-depressed peers. These descriptive patterns suggested that age and academic stress may play an important role in students' mental health and warranted further statistical testing. See Appendices 3 and 4 for data and visualizations.

Statistical Analysis and Hypothesis Testing

I conducted several tests to examine relationships between depression and demographic or academic variables. A chi-square test of independence found no statistically significant association between gender and depression ($\chi^2 \approx 0.085$, $p = 0.77$), indicating that depression prevalence did not differ meaningfully by gender.

In contrast, t-tests revealed statistically significant differences for both age and academic pressure. Students with depression were significantly younger than those without depression ($t \approx$

38.45, $p < .001$), with a mean age difference of just over two years. Academic pressure was also significantly higher among depressed students ($t \approx -89.30$, $p < .001$), indicating a strong association between academic stress and depression status. These results suggest that while gender may not be a key factor, age and academic pressure are strongly related to student depression.

Model Building

I used a logistic regression model to examine the combined effects of academic, demographic, and lifestyle factors on the likelihood of depression. The model included academic pressure, CGPA, study satisfaction, sleep duration, financial stress, family history of mental illness, work/study hours, gender, and age as predictors. During the logistic regression analysis, 18 observations were unexpectedly dropped by the model, even after initial data cleaning. To avoid mismatches between the model output and predicted values, these rows were removed from the dataset and excluded from both the regression model and prediction steps. As a result, the final analysis was conducted using the remaining complete cases.

Results indicated that higher academic pressure, greater financial stress, longer work/study hours, and a family history of mental illness significantly increased the odds of depression. Conversely, older age and higher study satisfaction were associated with lower odds of depression. Gender showed a very small effect and was only marginally significant.

Model performance was evaluated using a confusion matrix. The model achieved an overall classification accuracy of approximately 79%, with a balanced accuracy of about 78%. Sensitivity (≈ 0.72) indicated that the model correctly identified most depressed students, while specificity (≈ 0.84) showed strong performance in identifying non-depressed students. These results suggest the model has good predictive capability and captures meaningful relationships between student characteristics and depression outcomes.

Statistical Conclusions

This analysis demonstrates that depression among students is both prevalent and strongly associated with academic and lifestyle factors. While gender was not significantly related to depression, younger age, higher academic pressure, financial stress, and workload intensity emerged as key risk factors. The regression model confirmed that these variables jointly predict depression with substantial accuracy. Overall, the findings highlight the importance of addressing academic stressors and financial pressures in efforts to support student mental health and suggest that targeted interventions focusing on these areas may help reduce depression risk among students.

Appendix 1: Dataset Variables

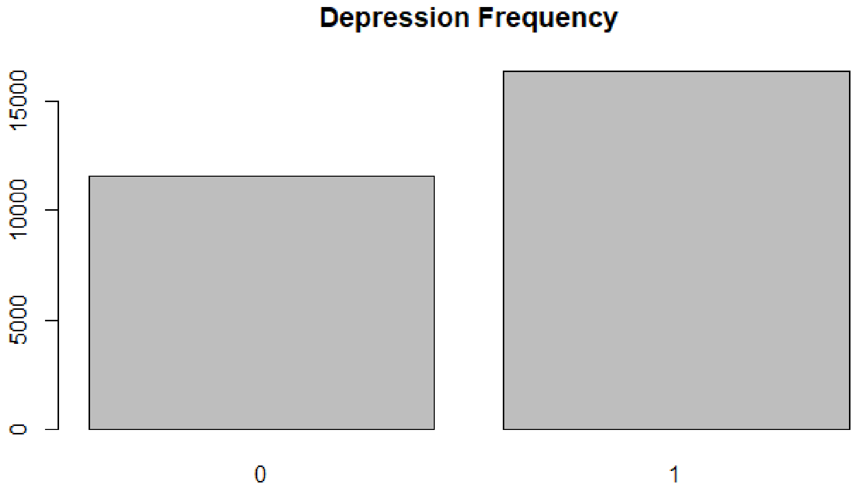
ID	Unique marker for each student.
Gender	Male or female.
Age	Age in years.
City	City or region where the student resides.
Profession	Primary job position. The vast majority of records show "Student."
Academic Pressure	A number between 0 and 5 that indicates the student's level of pressure they face in academic settings.
Work Pressure	A number between 0 and 5 that indicates the student's level of pressure related to work responsibilities.
CGPA	Cumulative grade point average. Its scale is from 0 to 10.
Study Satisfaction	A number between 0 and 5 that indicates how satisfied the student is with their studies.
Job Satisfaction	A number between 0 and 4 that indicates how satisfied the student is with their job environment, if applicable.
Sleep Duration	The average number of hours the student sleeps per day. Categories include: "Less than 5 hours," "5-6 hours," "7-8 hours," "More than 8 hours," and "Others."
Dietary Habits	A categorical variable assessing the student's eating patterns and nutritional habits. Possible values are "Healthy," "Moderate," "Unhealthy," and "Others."

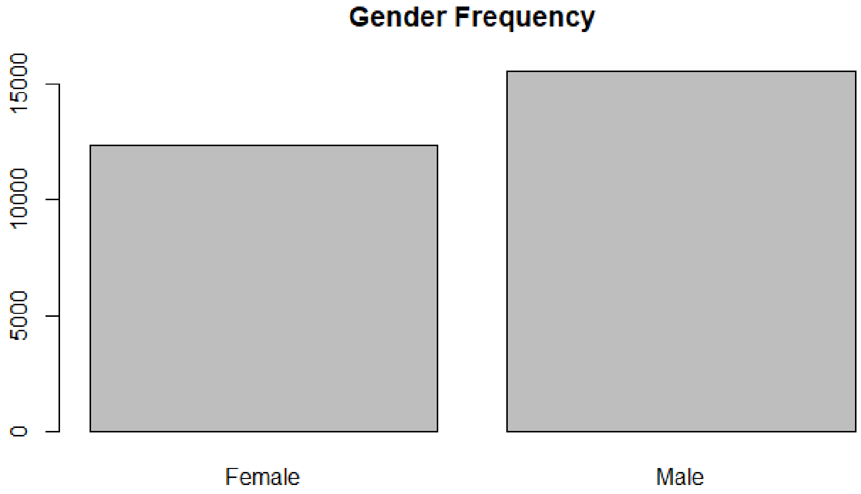
Degree	The academic degree or program that the student is pursuing.
Suicidal Thoughts	A binary indicator (Yes/No) that reflects whether the student has ever experienced suicidal ideation.
Work/Study Hours	The average number of hours per day the student dedicates to work or study.
Financial Stress	A number between 0 and 5 that indicates the level of stress experienced due to financial concerns.
Family History of Mental Illness	Indicates whether there is a family history of mental illness (Yes/No).
Depression	The target variable that indicates whether the student is experiencing depression (Yes/No).

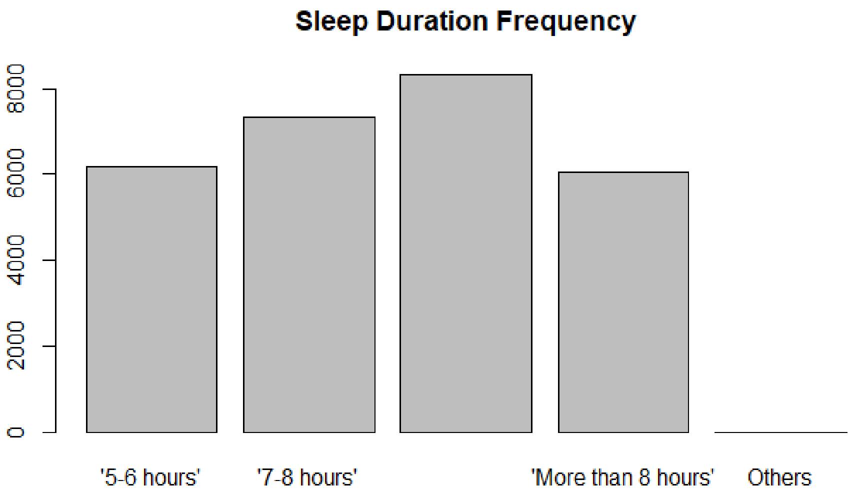
Appendix 2: GitHub Project Link

https://github.com/CamilleLewis-hr2270/INF-6490-Final-Project

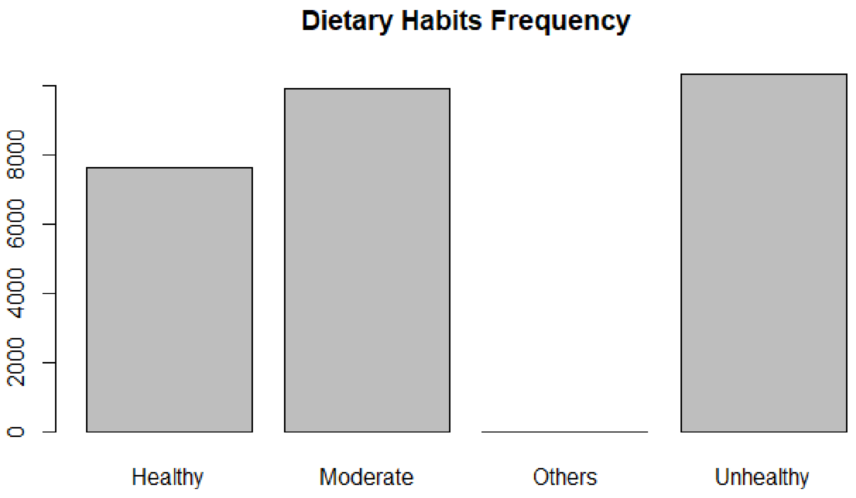
Appendix 3: Descriptive Statistics



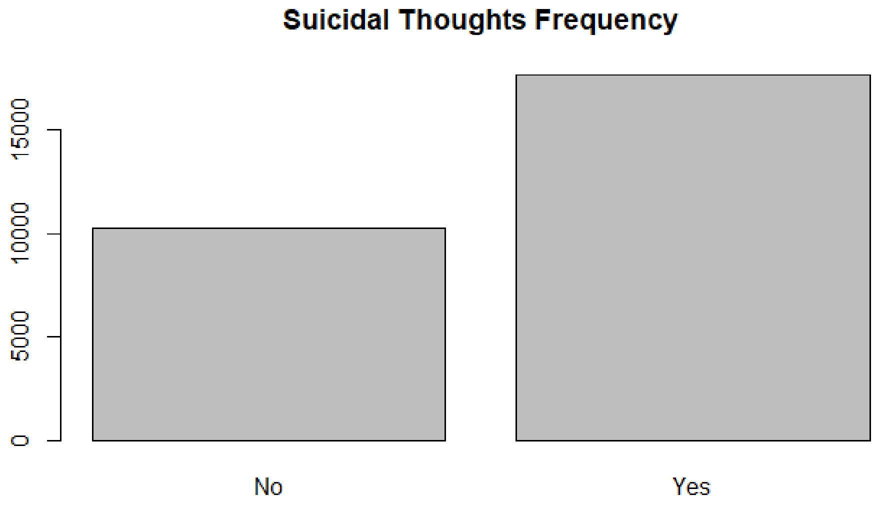


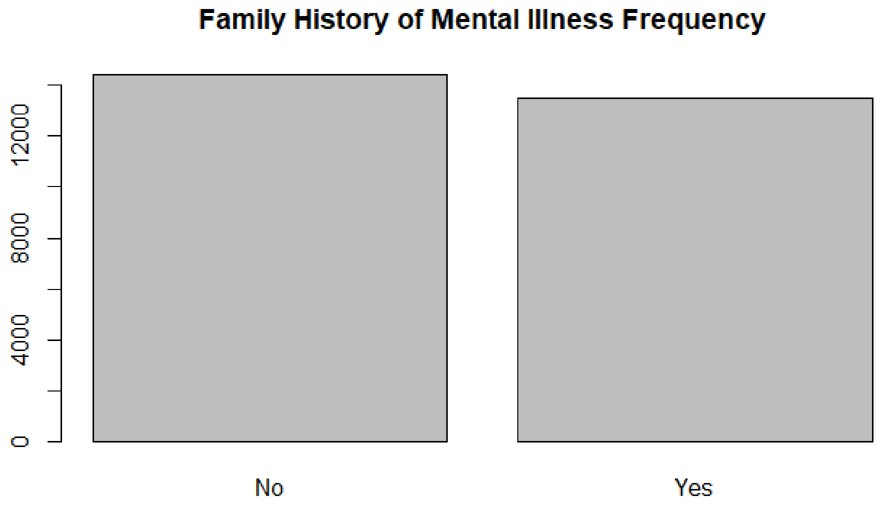


Sleep Duration Frequency



Dietary Habits Frequency

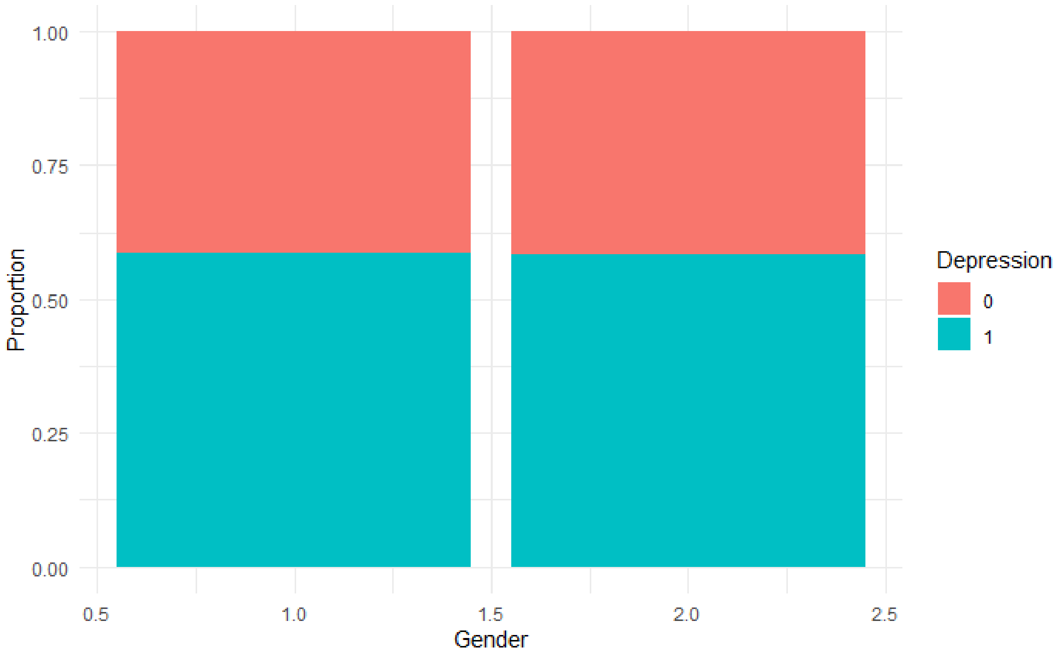

Suicidal Thoughts Frequency

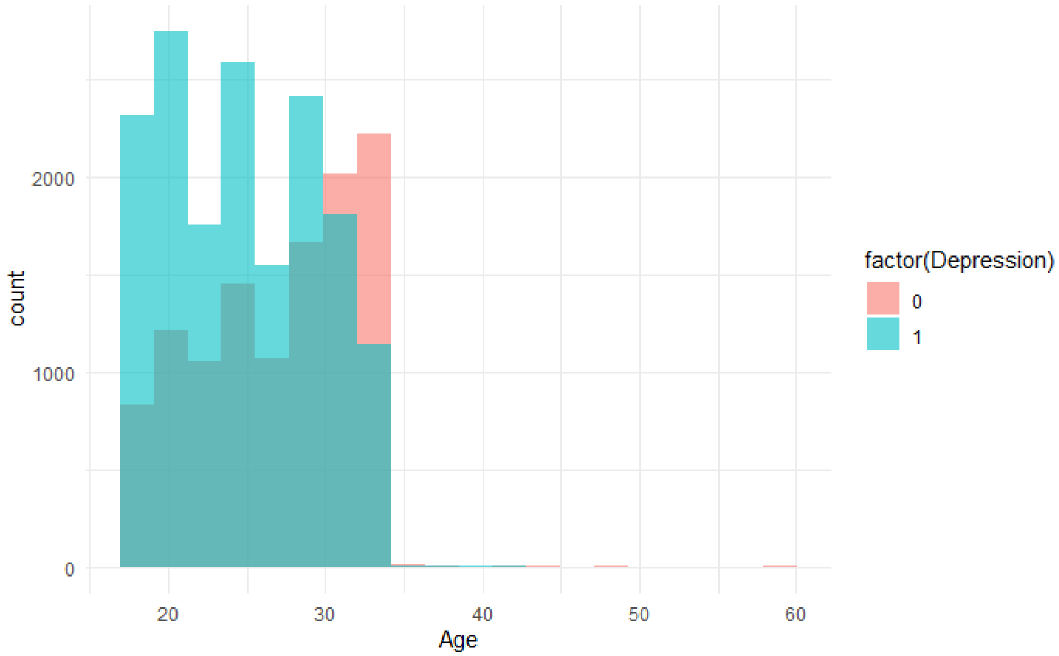

Family History of Mental Illness Frequency

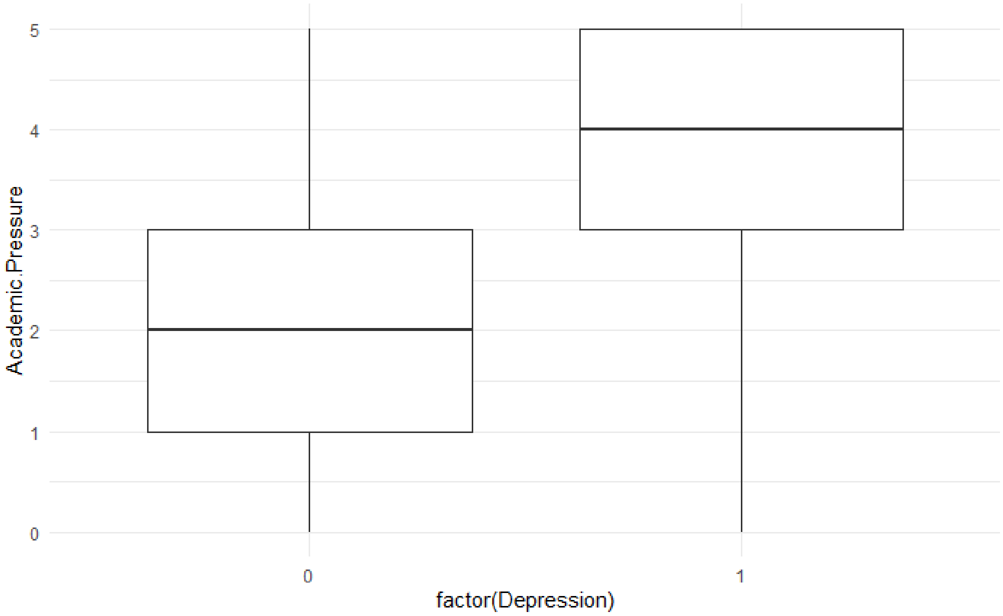
Dispersion of Continuous Variables

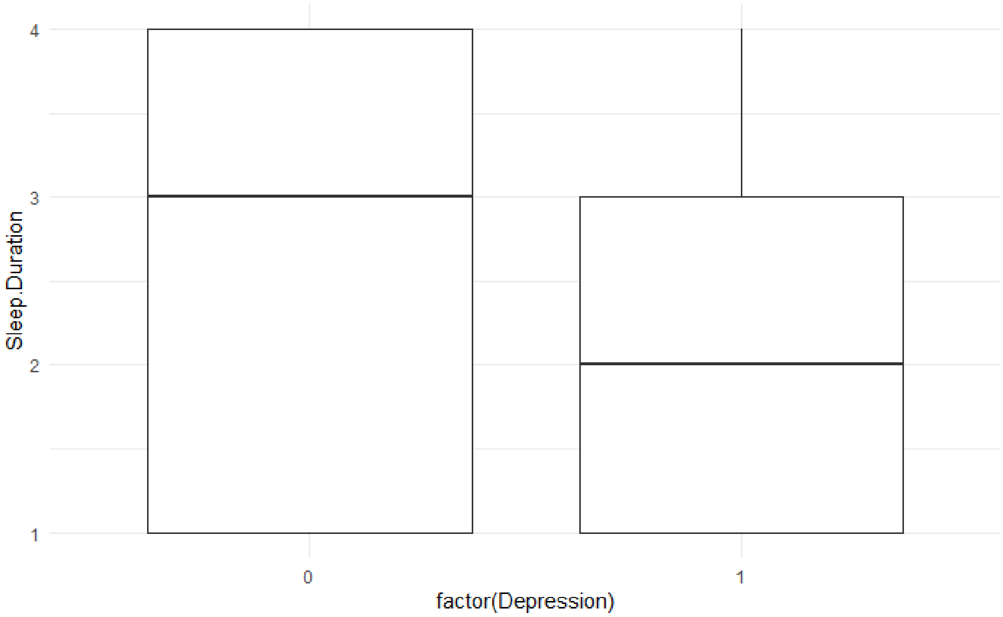
	Variance	Standard Deviation
Age	24.0654121730266	4.90565104476731
Academic Pressure	1.90843857701065	1.3814624775978
Work Pressure	0.00193550698425157	0.0439943971915921
CGPA	2.16298344409072	1.47070848372161
Study Satisfaction	1.85265197368278	1.36112158666402
Job Satisfaction	0.00197107430921968	0.0443967826449133
Work/Study Hours	Variance: 13.7462845701106	3.70759822123577
Financial Stress	2.06596552690126	1.43734669683457

Appendix 4: Data Visualization

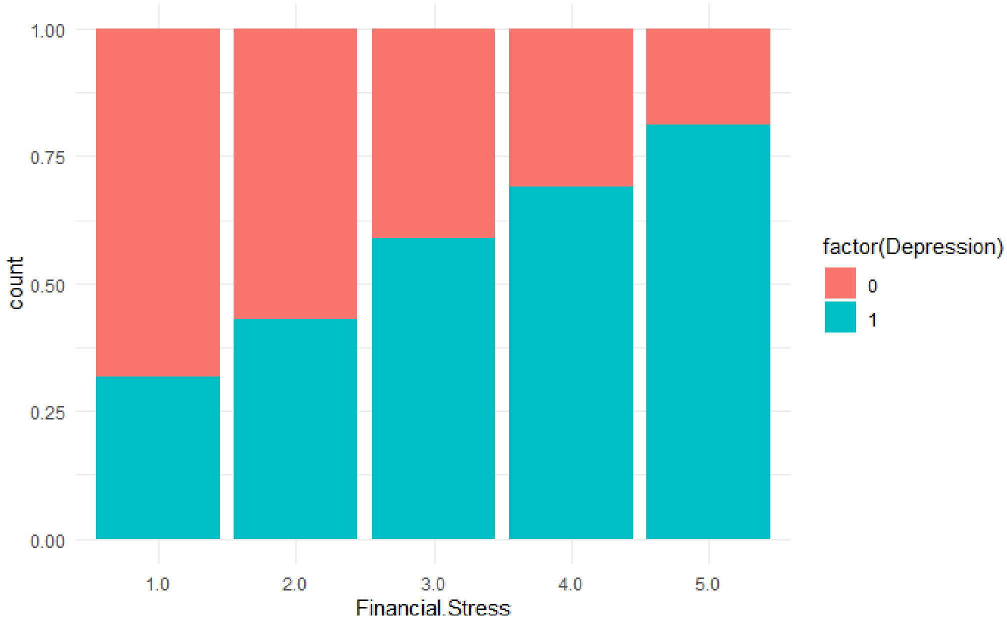



count
2000
1000
0
20
30
40
50
60
Age
factor(Depression)
0
1


Academic.Pressure
5
4
3
2
1
0
0
1
factor(Depression)



Sleep.Duration
4
3
2
1
0
1
factor(Depression)



count
1.00
0.75
0.50
0.25
0.00
1.0
2.0
3.0
4.0
5.0
Financial.Stress
factor(Depression)
🟥 0
🟦 1