Final Project - Step 2 (15 Points)

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

- Valerie De La Fuente (valeriedelafuente)
- Matthew Arteaga (matthewarteaga)
- Phuc Lu (pdlu)
- William Nelson (williamnelson)
- Hayden Galletta (haydengalletta)

Due Date

The deadline for this step is **Friday, May 9, 2025**.

Instructions

In this step, you will develop clear research questions and hypotheses based on your selected dataset, and conduct a thorough Exploratory Data Analysis (EDA). This foundational work is crucial for guiding your analysis in the following steps.

1 Step 2: Research Questions, Hypotheses, and Exploratory Data Analysis (EDA)

1.1 Research Questions

Question 1

Do certain dietary habits coincide with an increased rate of depression among students?

Question 2

Is there a correlation between the amount of sleep a student gets and the proportion of them that are depressed?

Question 3

Does the presence (and magnitude) of certain stressors have an impact on the rate at which students are depressed?

1.2 Hypotheses

Hypothesis 1

Students with moderate to healthy dietary habits will have lower rates of depression compared to students with unhealthy dietary habits.

Hypothesis 2

Students who average more sleep per night will have lower rates of depression compared to students who average less.

Hypothesis 3

Students with the highest collective reported stressors (`Academic Pressure + Work Pressure + Financial Stress`) will have higher rates of depression compared to students with lower collective reported stressors.

1.3 Exploratory Data Analysis (EDA)

1.3.1 Data Cleaning

1.3.1.1 Viewing the Data

```r
# Load necessary packages
library(readr)
library(tidyverse)
library(naniar)
library(janitor)

# Load in the data
depression_data <- read.csv("data/student_depression_dataset.csv")

# View the dataset
head(depression_data)
```

```
  id Gender Age          City Profession Academic.Pressure Work.Pressure CGPA
1  2   Male  33 Visakhapatnam    Student                 5             0 8.97
2  8 Female  24     Bangalore    Student                 2             0 5.90
3 26   Male  31      Srinagar    Student                 3             0 7.03
4 30 Female  28      Varanasi    Student                 3             0 5.59
5 32 Female  25        Jaipur    Student                 4             0 8.13
6 33   Male  29          Pune    Student                 2             0 5.70
  Study.Satisfaction Job.Satisfaction      Sleep.Duration Dietary.Habits
1                  2                0         '5-6 hours'        Healthy
2                  5                0         '5-6 hours'       Moderate
3                  5                0 'Less than 5 hours'        Healthy
4                  2                0         '7-8 hours'       Moderate
5                  3                0         '5-6 hours'       Moderate
6                  3                0 'Less than 5 hours'        Healthy
   Degree Have.you.ever.had.suicidal.thoughts.. Work.Study.Hours
1 B.Pharm                                   Yes                3
2     BSc                                    No                3
3      BA                                    No                9
4     BCA                                   Yes                4
5  M.Tech                                   Yes                1
6     PhD                                    No                4
  Financial.Stress Family.History.of.Mental.Illness Depression
1              1.0                               No          1
2              2.0                              Yes          0
3              1.0                              Yes          0
4              5.0                              Yes          1
5              1.0                               No          0
6              1.0                               No          0
```

```r
# Examine the dimensions
dim(depression_data)
```

```
[1] 27901    18
```

There are 27901 observations and 18 variables in this dataset. The list of variables is as follows:

- `id`: A unique identifier assigned to each student record in the dataset.
- `Gender`: The gender of the student (e.g., Male, Female, Other). This helps in analyzing gender-specific trends in mental health.
- `Age`: The age of the student in years.
- `City`: The city or region where the student resides, providing geographical context for the analysis.
- `Profession`: The field of work or study of the student, which may offer insights into occupational or academic stress factors.
- `Academic Pressure`: A measure indicating the level of pressure the student faces in academic settings. This could include stress from exams, assignments, and overall academic expectations.
- `Work Pressure`: A measure of the pressure related to work or job responsibilities, relevant for students who are employed alongside their studies.
- `CGPA`: The cumulative grade point average of the student, reflecting overall academic performance.
- `Study Satisfaction`: An indicator of how satisfied the student is with their studies, which can correlate with mental well-being.
- `Job Satisfaction`: A measure of the student's satisfaction with their job or work environment, if applicable.
- `Sleep Duration`: The average number of hours the student sleeps per day, which is an important factor in mental health.
- `Dietary Habits`: An assessment of the student's eating patterns and nutritional habits, potentially impacting overall health and mood.
- `Degree`: The academic degree or program that the student is pursuing.
- `Have you ever had 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, which can influence stress levels.
- `Financial Stress`: A measure of the stress experienced due to financial concerns, which may affect mental health.
- `Family History of Mental Illness`: A measure of the stress experienced due to financial concerns, which may affect mental health.
- `Depression`: The target variable that indicates whether the student is experiencing depression (Yes/No). This is the primary focus of the analysis.

1.3.1.2 Fixing Column Names

```r
# Fix column names
depression_data <- depression_data %>%
  clean_names() %>%
  rename(
    cum_gpa = cgpa,
    suicidal_thoughts = have_you_ever_had_suicidal_thoughts,
    fam_mental_illness = family_history_of_mental_illness
  )

```

```r
# Check if names were fixed
names(depression_data)
```

```
 [1] "id"                 "gender"             "age"               
 [4] "city"               "profession"         "academic_pressure" 
 [7] "work_pressure"      "cum_gpa"            "study_satisfaction"
[10] "job_satisfaction"   "sleep_duration"     "dietary_habits"    
[13] "degree"             "suicidal_thoughts"  "work_study_hours"  
[16] "financial_stress"   "fam_mental_illness" "depression"        
```

1.3.1.3 Missing Data

```r
# View missing data
sum(is.na(depression_data))
```

```
[1] 0
```

On the surface, there is no missing data. However, when looking at the categories and their unique values, there are some signs of missingness.

For example, some categories have the `Other` category. Since there is not way of figuring out what `Other` mean precisely, it can be considered as an unknown category. To deal with this, it won't remove but will still be concluded to not harm the data integrity.

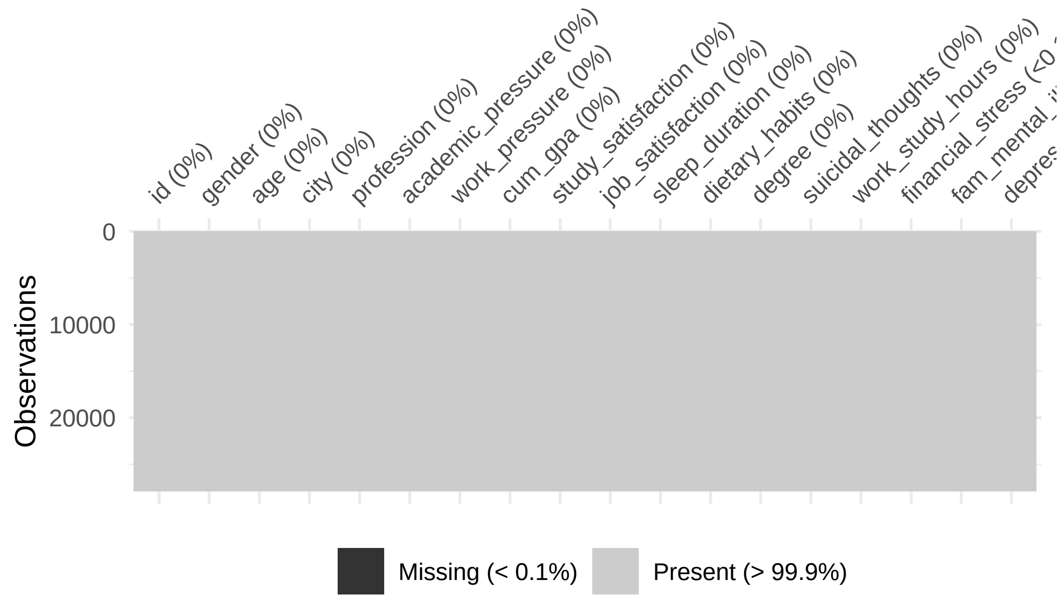
Besides the other category, there is one variable that encodes its missing value as `?`. This is a placeholder for missing value under the `financial stress` variable. We'll deal with this by encoding it as `NA`.

```r
# Fixing the `financial_stress` variable
depression_data <- depression_data %>%
  mutate(
    financial_stress = as.numeric(financial_stress),
    # convert string numbers to integers
    financial_stress = case_when(
      financial_stress == "?" ~ NA,
      # convert "?" to NA values
      .default = financial_stress))

sum(is.na(depression_data))
```

```
[1] 3
```

Now, the total number of missing observation is 3, which comes from the `financial stress` variable.

```r
library(naniar)
depression_data %>% vis_miss()
```



This missingness only makes up $\frac{3}{27,901}$ values or much less than 0.1% of the data set. We can simply remove these values without a problem.

```r
depression_data <- depression_data %>% na.omit()
depression_data %>% dim()
```

```
[1] 27898    18
```

1.3.1.4 Checking Data Types

```r
# Check data types of the variables
str(depression_data)
```

According to the output, we must mutate some variables. This includes factorization and fixing some values that the variables take in.

1.3.1.5 Mutating Variables

```r
# Factorizing the `gender` variable
depression_data$gender <- factor(depression_data$gender)

# Fixing the `city` variable to change invalid entries
depression_data <- depression_data %>%
  mutate(city = case_when(
    city == "Khaziabad" ~ "Ghaziabad",
    city == "Nalyan" ~ "Kalyan",
    city == "'Less Delhi'" ~ "Delhi",
    city == "'Less than 5 Kalyan'" ~ "Kalyan",
    city == "3.0" ~ "Other",
    city == "Saanvi" ~ "Other",
    city == "M.Tech" ~ "Other",
    city == "Bhavna" ~ "Other",
    city == "City" ~ "Other",
    city == "Mira" ~ "Other",
    city == "Harsha" ~ "Other",
    city == "Vaanya" ~ "Other",
    city == "Gaurav" ~ "Other",
    city == "Harsh" ~ "Other",
    city == "Reyansh" ~ "Other",
    city == "Kibara" ~ "Other",
    city == "Rashi" ~ "Other",
    city == "ME" ~ "Other",
    city == "M.Com" ~ "Other",
    city == "Mihir" ~ "Other",
    city == "Nalini" ~ "Other",
    city == "Nandini" ~ "Other",
    TRUE ~ city  # Leave valid entries as they are
  ))

# Fixing the `profession` variable to change invalid entries
depression_data <- depression_data %>%
  mutate(profession = case_when(
    profession == "'Civil Engineer'" ~ "Civil Engineer",
    profession == "'UX/UI Designer'" ~ "UX/UI Designer",
    profession == "'Digital Marketer'" ~ "Digital Marketer",
    profession == "'Content Writer'" ~ "Content Writer",
    profession == "'Educational Consultant'" ~ "Educational Consultant",
    TRUE ~ profession # Leave valid entries as they are
  ))

# Fixing the `sleep_duration` variable to change invalid entries
depression_data <- depression_data %>%
  mutate(sleep_duration = case_when(
    sleep_duration == "'5-6 hours'" ~ "5-6 hours",
    sleep_duration == "'Less than 5 hours'" ~ "Less than 5 hours",
    sleep_duration == "'7-8 hours'" ~ "7-8 hours",
    sleep_duration == "'More than 8 hours'" ~ "More than 8 hours",
    sleep_duration == "Others" ~ "Other"
  ))

# Factorizing the `sleep_duration` variable
depression_data <- depression_data %>%
  mutate(sleep_duration = factor(sleep_duration,
                                 levels = c("Less than 5 hours",
                                            "5-6 hours",
                                            "7-8 hours",
                                            "More than 8 hours",
                                            "Other"),
                                 ordered = TRUE))

# Fixing the `dietary_habits` variable to change misspelling
depression_data <- depression_data %>%
  mutate(dietary_habits = case_when(
    dietary_habits == "Others" ~ "Other",
    TRUE ~ dietary_habits
  ))

```

```r
# Factorizing the `dietary_habits` variable
depression_data <- depression_data %>%
  mutate(dietary_habits = factor(dietary_habits,
                                 levels = c("Healthy", "Moderate", "Unhealthy",
                                            "Other"),
                                 ordered = TRUE))

# Fixing the `degree` variable to change invalid entries
depression_data <- depression_data %>%
  mutate(degree = case_when(
    degree == "'Class 12'" ~ "Class 12",
    degree == "Others" ~ "Other",
    # Others could less than HS education or totally unknown.
    .default = degree
  ))

# Factorizing the `suicidal_thoughts` variable
depression_data$suicidal_thoughts <- factor(depression_data$suicidal_thoughts)

# Factorizing the `fam_mental_illness` variable
depression_data$fam_mental_illness <- factor(depression_data$fam_mental_illness)

# Turning the `depression` variable back to "yes" and "no" for visualization purposes
depression_data <- depression_data %>%
  mutate(depression = case_when(
    depression == 0 ~ "No",
    depression == 1 ~ "Yes"
  ))

# Factorizing the `depression` variable
depression_data$depression <- factor(depression_data$depression)
```

```r
# Check data types of the variables again to ensure everything was properly done
str(depression_data)
```

According to the output, the data was successfully cleaned and the variables are ready for visualization.

1.4 Descriptive Statistics

1.4.1 Opinion Rating Variables

These are variables whose values were collected by asking the subjects to rate their experience from a scale of 1-5. These include `academic pressure`, `work pressure`, `study satisfaction`, `financial stress`. We chose to isolate these to study their summary statistics because they're in a different class compare to the other kind of numeric statistics. For example, their values are bounded between 0 to 5 because that's the range of the rating scale, where as age and cumulative GPA are not bound to the same constraints.

```r
rating_var <- depression_data %>%
  select(academic_pressure, work_pressure, job_satisfaction, study_satisfaction, financial_stress)
```

```r
rating_var %>% summary()
```

Statistic	Academic Pressure	Work Pressure	Job Satisfaction	Study Satisfaction	Financial Stress
Min.	0.000	0.0000000	0.000000	0.000	1.00
1st Qu.	2.000	0.0000000	0.000000	2.000	2.00
Median	3.000	0.0000000	0.000000	3.000	3.00
Mean	3.141	0.0004300	0.000681	2.944	3.14
3rd Qu.	4.000	0.0000000	0.000000	4.000	4.00
Max.	5.000	5.0000000	4.000000	5.000	5.00

A category that immediately jumps out is the extremely low overall rating for `work pressure`. This means that the majority of students in the data set reported that they have no pressure with their work or job related responsibilities.

```r
# table of work pressure and rating
depression_data$work_pressure %>% table()
```

Work Pressure	0	1	2	3	4	5
Frequency	27710	0	1	0	0	2

There are 2 students that reported 5 for work pressure. These are male students from Rajkot and Chennai. They also share the same rating of 3 for financial stress, cumulative GPA of 0.0, they seem to be very satisfied with their jobs with a rating of 4, no family history of mental illness, and both have had high school education.

The one student who rated a 2 for work pressure is a male student from from Lucknow, experiences academic pressure, 0.0 cumulative GPA, 1 for job satisfaction, does have suicidal thoughts, 3 hours of work/study a day, and does have family history of mental illness.

`Job satisfaction` is another category with the majority of students express a rating of 0. For those with high job satisfaction, we want to know what these type of people are and what they do.

```r
# people with job satisfaction > 0
good_job <- depression_data %>% filter(job_satisfaction > 0)
```

There are 8 students in this data who rated themselves as having job satisfaction greater than 0.

We're most interested in those who have the highest job satisfaction.

```r
good_job %>% filter(job_satisfaction == max(job_satisfaction))
```

There are two male students in this data who rated their job satisfaction a 4 which is higher than the majority of people in this data set. Besides high job satisfaction, the other attributes that they have in common are rated 3 for academic pressure, rated 3 for work pressure, 0 study satisfaction, education up to grade 12, and no family history of mental illness.

The first person is a 38 years old healthy male from Chennai, India. This person sleeps for 5-6 hours, no suicidal thoughts, 2 hours of work-study, rated a 3 for financial stress, and no depression.

The second person is an 18 years old moderately healthy male from Rajkot. This person sleeps for 7-8 hours, does have suicidal thoughts, 9 hours of work-study, rated a 4 for financial stress, and does have depression.

```r
depression_data %>%
  filter(work_pressure == max(work_pressure),
         job_satisfaction == max(job_satisfaction))
```

Between `work pressure` and `job satisfaction`, the two students with the highest rating 5 for work pressure is also the person with the highest job satisfaction with a rating of 4. They both have at least high school education, 0 cumulative GPA, 0 study satisfaction, and no family history of having a mental illness.

1.4.2 Other Numeric Variables

These variables include id, age, cumulative GPA and work/study hours.

```r
other_numeric_var <- depression_data %>% select_if(is.numeric) %>% select(-(rating_var %>% colnames))
```

```r
other_numeric_var %>% summary()
```

Statistic	ID	Age	Cumulative GPA	Work/Study Hours
Min.	2	18.0	0.000	0.000
1st Qu.	35053	21.0	6.290	4.000
Median	70694	25.0	7.770	8.000
Mean	70449	25.8	7.659	7.158
3rd Qu.	105828	30.0	8.920	10.000
Max.	140699	59.0	10.000	12.000

`ID` can also be omitted from summary statistics because its numeric value doesn't carry the same meaning as all of the other variables. It's purpose is simply to provide a unique identity to each observation.

The student with the oldest age in the data set is 59 years old and the youngest student is 18 years old. On average, the students in the data set tend to be 26 years old. Since the mean is slightly higher than the median, this reveals that the shape of the distribution is slightly right-skewed with tail. The age with the highest number of students is 24 and is followed at students who are 20 years old.

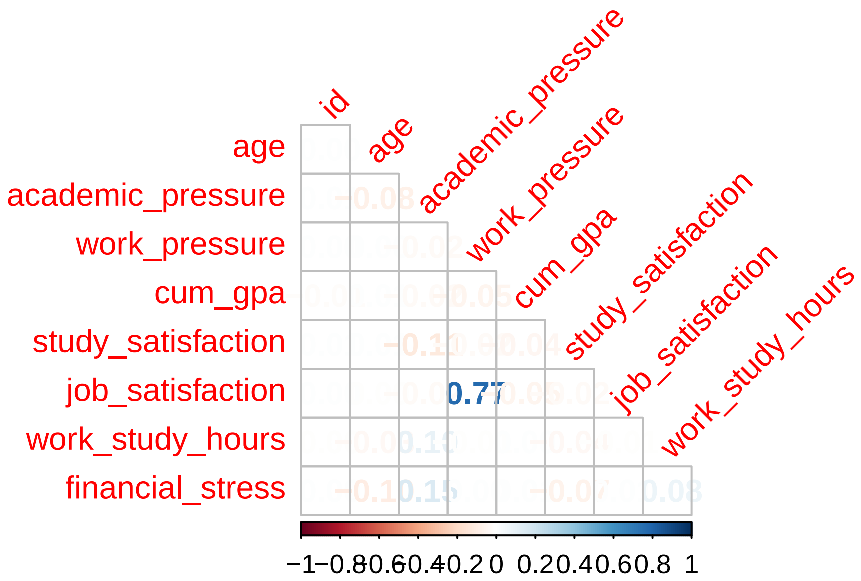
Since this data set was based in India, their GPA scale system is not the same as the U.S. Instead of using the 4.0 scale, India's GPA scale uses the 10.0 scale where it's out of 10. Under this system, the highest GPA is 10.0 decreases from there. In this data set, the highest cumulative GPA is 10.0 while the lowest is 0. The mean cumulative GPA is 7.656 or terms of the letter grade scale, it's approximately a B letter grade.

There are 9 students who reported having a cumulative GPA of 0.0, which is almost impossible to receive if they went to school. However, a cumulative GPA of 0 is possible if the students had just started school. Among these 9 students, the most shocking trait that they share is that 2/3 of them did express having suicidal thoughts and the majority of them does not have education beyond high school.

For work or study hours, the max number is 12 hours can reflect intensive work or academic schedules. On the other hand, the lowest number of hour for work or study is 0, which means that these students do not engage with work or study in a day. On average, the students tend to work 7.158 hours. Since the middle of 0 to 12 is 6, but the mean and median are above that, it can be inferred that students tend to a high academic or work-related workload.

1.4.3 Correlation Plot of Numeric Variables

```r
library(corrplot)
cor(depression_data %>%
      select_if(is.numeric)) %>%
  corrplot(type="lower", diag = FALSE, method = "number", tl.srt = 45)
```



There is a large correlation of 0.77 between the responses of `Job Satisfaction` and `Work Pressure`. We should explore this correlation:

```
table(depression_data$work_pressure)
table(depression_data$job_satisfaction)
```

Work Pressure Level	Frequency
0	27,895
1	0
2	1
3	0
4	0
5	2

Job Satisfaction Level	Frequency
0	27,890
1	2
2	3
3	1
4	2

When looking at the frequency table for `work pressure` and `job satisfaction`, it is the case that nearly all of the students expressed that they have the lowest level (0) in both variables.The majority of students reported 0 in both variable could also explain the high correlation. That is, if a student reported having 0 work pressure, it is also extremely likely that they will also report having 0 job satisfaction. A reason for this probably because the students do not have jobs and are full-time students.

1.4.4 Binary Variables

Since these variables are continuous, it is better to analyze them by comparing proportions of the two categories.

1.4.4.1 Depression

```r
round(prop.table(table(depression_data$depression)), digits = 3)
```

Depression	No	Yes
Proportion	0.414	0.586

We see just below 60% of students in the data set responded they experience Depression. We proportion is not necessarily imbalanced, but it's not totally balanced either. It's not necessary to be concern about misrepresentation.

1.4.4.2 Suicidal Thoughts

```r
round(prop.table(table(depression_data$suicidal_thoughts)), digits = 3)
```

Suicidal Thoughts	No	Yes
Proportion	0.367	0.633

We see 63.3% of students respond they have had suicidal thoughts. This means that there are more students who reported having suicidal thoughts than those who did not. This proportion is imbalanced but it's doesn't seem to be arbitrary. It would be interesting to explore the common attributes among those who reported having suicidal thoughts. On the other hand, the same thing could be done for those who reported not having suicidal thoughts. Learning about the main indicator(s) for suicidal thoughts or beneficial habits of those without suicidal thoughts can potentially help students with suicidal thoughts later on by incorporating the good habits into their lives.

1.4.4.3 Family Mental Illness

```r
round(prop.table(table(depression_data$fam_mental_illness)), digits = 3)
```

Family History Mental Illness	No	Yes
Proportion	0.516	0.484

It is nearly an even split between responses for the presence of mental illness in the student's family, with a slightly higher frequency of "No" responses. Since both groups are fairly even, it could be worthwhile to explore if having a family history of a mental illness could potentially contribute to a student developing depression in their lives.

1.4.5 Categorical Variables

These are variables with categories and numerical values to show how many students fall under each category.

1.4.5.1 Dietary Habits

```r
round(prop.table(table(depression_data$dietary_habits)), digits = 4)
```

Dietary Habit	Healthy	Moderate	Unhealthy	Other
Proportion	0.2742	0.3556	0.3698	0.0004

More students have moderate and unhealthy dietary habits than healthy dietary habits. There is a very small amount of students reported having a more nuanced dietary habit. It is difficult to tell what they eat, but we decided to just leave this category be instead of removing it.

1.4.5.2 Sleep Duration

```r
round(prop.table(table(depression_data$sleep_duration)), digits = 3)
```

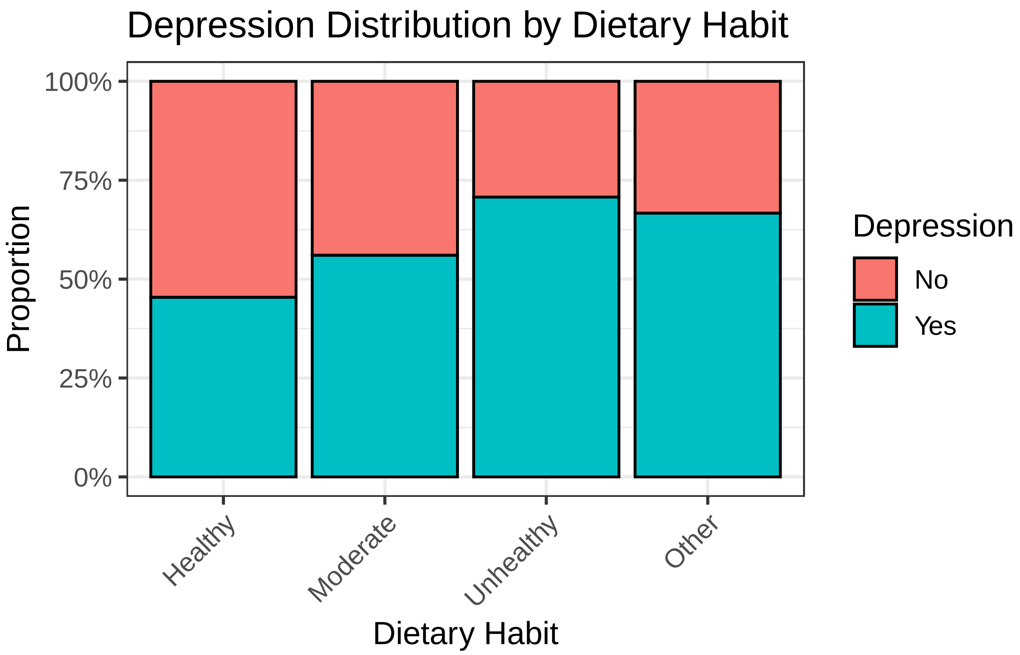
Sleep Duration	Less than 5 hours	5-6 hours	7-8 hours	More than 8 hours	Other
Proportion	0.298	0.222	0.263	0.217	0.001

We see that almost a third of students said they get less than 5 hours of sleep on average. There is again, a small group of students who report having a sleep duration that's different from the other categories. They could have a very unstable sleeping habit that fluctuates a lot every night. This category could be valuable for looking into if having a regular sleep schedule as opposed to one that fluctuates can contribute to having depression.

1.5 Data Visualization

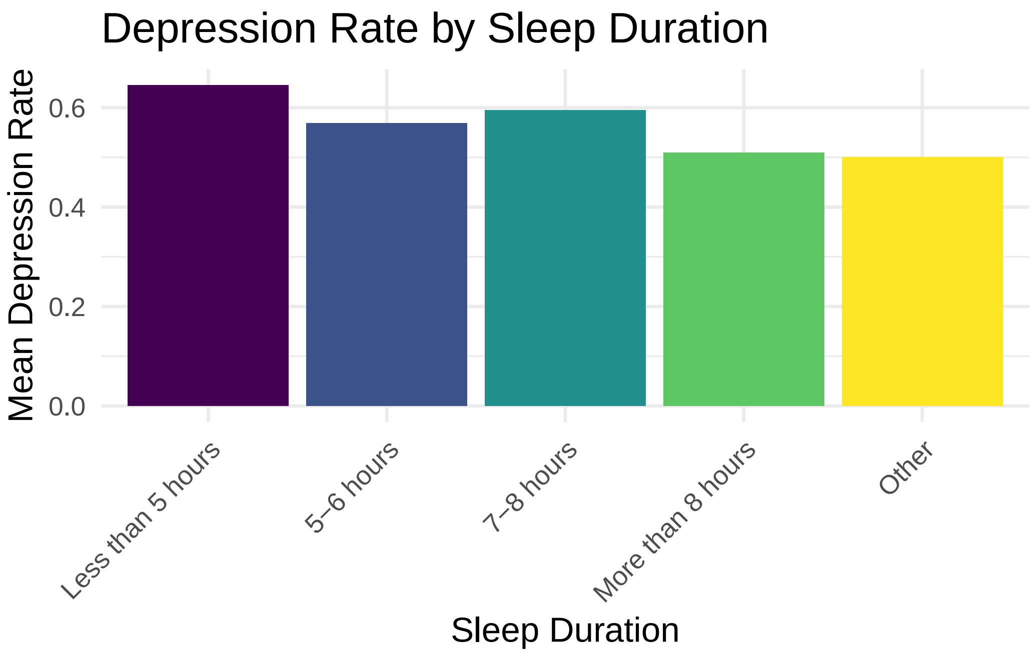
This first graph is a bar plot that helps to answer hypothesis 1 by visualizing the correlation between a healthy diet and depression. The results of this bar plot clearly indicate a strong correlation with a unhealthy diet and rates of depression.

```r
# For dietary habits
ggplot(depression_data, aes(x = dietary_habits, fill = factor(depression))) +
  geom_bar(position = "fill", color = "black") +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Depression Distribution by Dietary Habit",
       x = "Dietary Habit", y = "Proportion",
       fill = "Depression") +
  theme_bw(base_size = 12) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



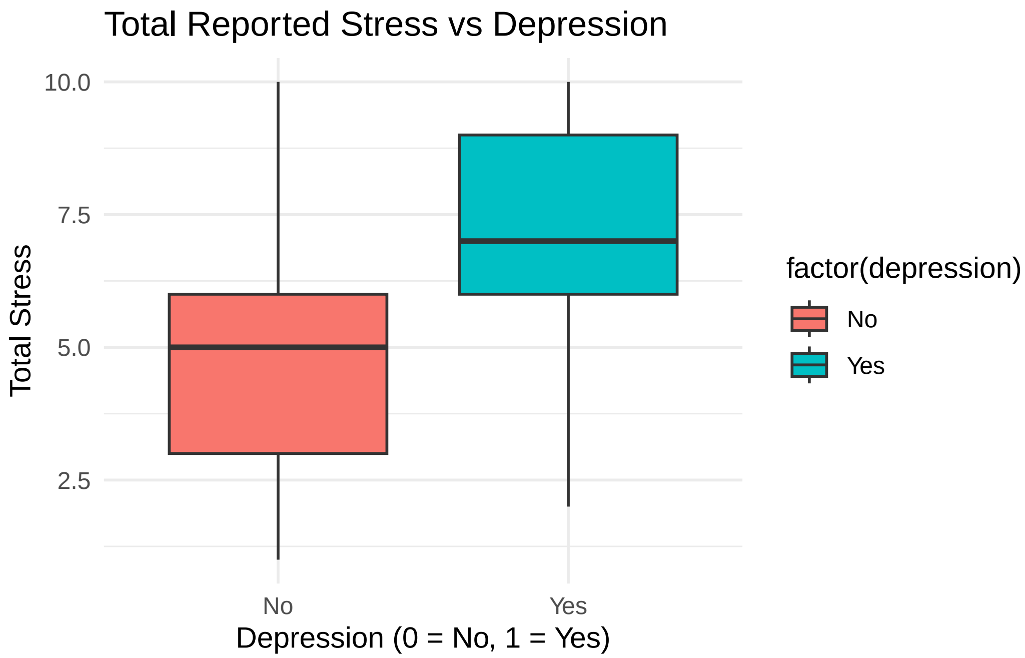
The second graph is a bar plot that helps to answer hypothesis 2 by visualizing the correlation between sleep patterns and depression. The results of this bar plot seem to indicate that people getting less than 5 hours of sleep have a significantly higher rate of depression and people who get more than 8 hours of sleep have a significantly lower rate of depression.

```r
# Convert 'depression' factor to numeric: "No" = 0, "Yes" = 1
depression_data <- depression_data %>%
  mutate(depression_numeric = as.numeric(depression) - 1)

# Create summarized depression rates and standard errors by sleep duration
sleep_summary <- depression_data %>%
  group_by(sleep_duration) %>%
  summarise(
    mean_dep = mean(depression_numeric, na.rm = TRUE),
    se = sd(depression_numeric, na.rm = TRUE) / sqrt(n())
  )

# Bar plot with error bars
ggplot(sleep_summary, aes(x = sleep_duration, y = mean_dep, fill = sleep_duration)) +
  geom_col(show.legend = FALSE) +

  labs(
    title = "Depression Rate by Sleep Duration",
    x = "Sleep Duration",
    y = "Mean Depression Rate"
  ) +
  theme_minimal(base_size = 13) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This final graph is a boxplot that helps visualize hypothesis 3 and shows the correlation between multiple stressor factors and depression. The results seem to indicate a correlation with the total stressors and depression.

```r
depression_data$financial_stress <- as.numeric(depression_data$financial_stress)
depression_data$total_stress <- depression_data$academic_pressure +
                                 depression_data$work_pressure +
                                 depression_data$financial_stress

ggplot(depression_data, aes(x = factor(depression), y = total_stress, fill = factor(depression))) +
  geom_boxplot() +
  labs(title = "Total Reported Stress vs Depression", x = "Depression (0 = No, 1 = Yes)", y = "Total St
  theme_minimal()
```


Total Reported Stress vs Depression
Total Stress
No
Yes
Depression (0 = No, 1 = Yes)
factor(depression)
No
Yes