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)

1.4 Data Cleaning

1.4.1 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.

1.4.2 Fixing Column Names

```
# 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
  )

# 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"
```

1.4.3 Missing Data

```
# 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.

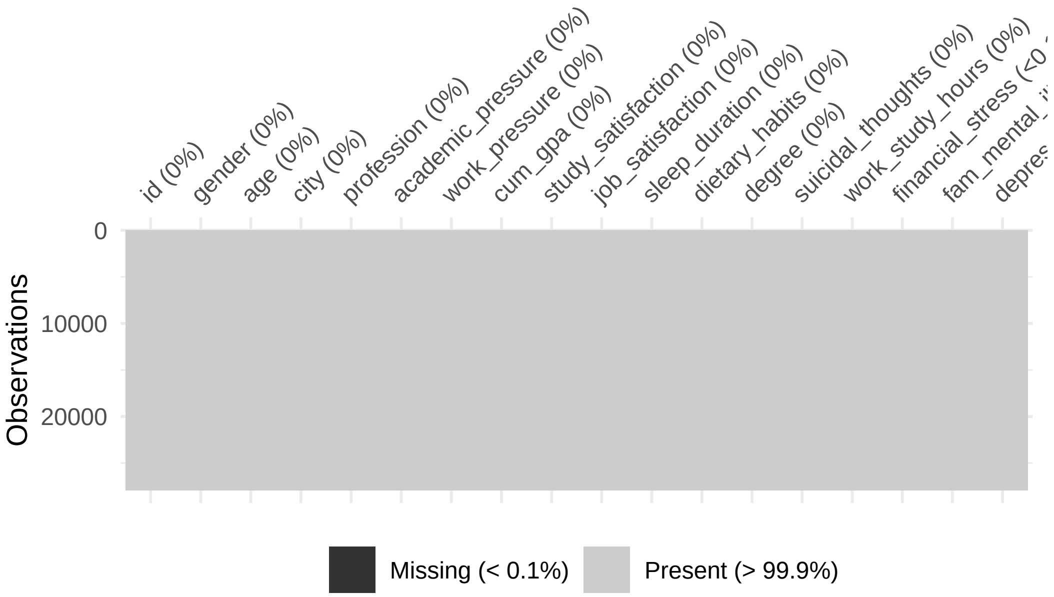
Besides the other category, there is one variable that encodes its missing value as `?`. This is a placeholder for missing value under the `financial stress` variable. We'll deal with this by encoding it as `NA`.

```
# 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
```

1.4.4 Checking Data Types

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

```
'data.frame':   27898 obs. of  18 variables:
 $ id                : int  2 8 26 30 32 33 52 56 59 62 ...
 $ gender            : chr  "Male" "Female" "Male" "Female" ...
 $ age               : num  33 24 31 28 25 29 30 30 28 31 ...
 $ city              : chr  "Visakhapatnam" "Bangalore" "Srinagar" "Varanasi" ...
 $ profession        : chr  "Student" "Student" "Student" "Student" ...
 $ academic_pressure : num  5 2 3 3 4 2 3 2 3 2 ...
 $ work_pressure     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ cum_gpa           : num  8.97 5.9 7.03 5.59 8.13 5.7 9.54 8.04 9.79 8.38 ...
 $ study_satisfaction: num  2 5 5 2 3 3 4 4 1 3 ...
 $ job_satisfaction  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sleep_duration    : chr  "'5-6 hours'" "'5-6 hours'" "'Less than 5 hours'" "'7-8 hours'" ...
 $ dietary_habits    : chr  "Healthy" "Moderate" "Healthy" "Moderate" ...
 $ degree            : chr  "B.Pharm" "BSc" "BA" "BCA" ...
 $ suicidal_thoughts : chr  "Yes" "No" "No" "Yes" ...
 $ work_study_hours  : num  3 3 9 4 1 4 1 0 12 2 ...
 $ financial_stress  : num  1 2 1 5 1 1 2 1 3 5 ...
 $ fam_mental_illness: chr  "No" "Yes" "Yes" "Yes" ...
 $ depression        : int  1 0 0 1 0 0 0 0 1 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:3] 4459 13597 19267
  ..- attr(*, "names")= chr [1:3] "4459" "13597" "19267"
```

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

1.4.5 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
  ))

# Since we're interested in Student Depression,
# we'll removing observations that are not Student.
depression_data <- depression_data %>% filter(profession == "Student")

# Fixing the `work_pressure` variable for proper scaling
depression_data <- depression_data %>%
  mutate(work_pressure = case_when(
    work_pressure == 0 ~ 0,
    work_pressure == 2 ~ 1,
    work_pressure == 5 ~ 3
  ))

```

```r
# 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
  ))

# 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'" ~ "High School",
    degree == "Others" ~ "Other",
    # Others could less than HS education or totally unknown.
    .default = degree
  ))

# Factorizing the `degree variable`
degree_levels <- c(
  "High School",
  "BA", "BSc", "B.Com", "BCA", "B.Pharm", "B.Ed", "B.Tech", "BE", "BHM", "B.Arch", "BBA",
  "MA", "MSc", "MBA", "M.Com", "MCA", "M.Tech", "M.Ed", "M.Pharm", "MHM",
  "LLB", "LLM", "MD", "MBBS",
  "PhD",
  "Other"
)
depression_data <- depression_data %>%
  mutate(degree = factor(degree, levels = degree_levels, ordered = TRUE))
```

```r

# 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)

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

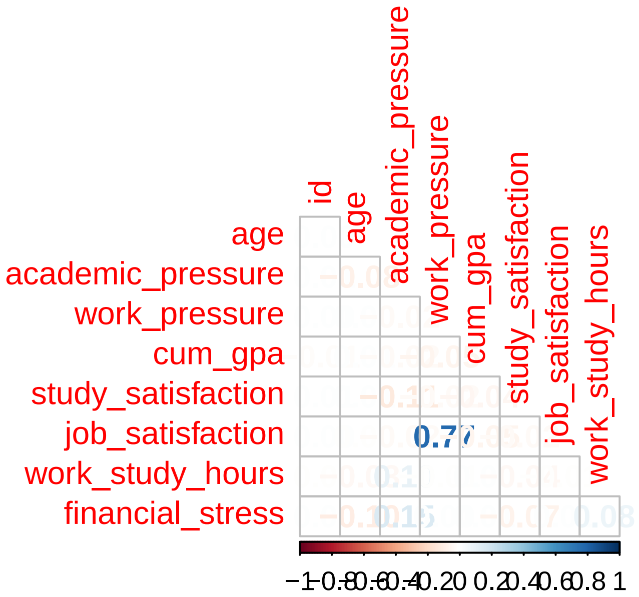
```
'data.frame':   27867 obs. of  18 variables:
 $ id                : int  2 8 26 30 32 33 52 56 59 62 ...
 $ gender            : Factor w/ 2 levels "Female","Male": 2 1 2 1 1 2 2 1 2 2 ...
 $ age               : num  33 24 31 28 25 29 30 30 28 31 ...
 $ city              : chr  "Visakhapatnam" "Bangalore" "Srinagar" "Varanasi" ...
 $ profession        : chr  "Student" "Student" "Student" "Student" ...
 $ academic_pressure : num  5 2 3 3 4 2 3 2 3 2 ...
 $ work_pressure     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ cum_gpa           : num  8.97 5.9 7.03 5.59 8.13 5.7 9.54 8.04 9.79 8.38 ...
 $ study_satisfaction: num  2 5 5 2 3 3 4 4 1 3 ...
 $ job_satisfaction  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sleep_duration    : Ord.factor w/ 5 levels "Less than 5 hours"<..: 2 2 1 3 2 1 3 1 3 1 ...
 $ dietary_habits    : Ord.factor w/ 4 levels "Healthy"<"Moderate"<..: 1 2 1 2 2 1 1 3 2 2 ...
 $ degree            : Ord.factor w/ 27 levels "High School"<..: 6 3 2 5 18 26 3 1 7 22 ...
 $ suicidal_thoughts : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 1 1 2 2 ...
 $ work_study_hours  : num  3 3 9 4 1 4 1 0 12 2 ...
 $ financial_stress  : num  1 2 1 5 1 1 2 1 3 5 ...
 $ fam_mental_illness: Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 1 ...
 $ depression        : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 1 2 2 ...
 - attr(*, "na.action")= 'omit' Named int [1:3] 4459 13597 19267
  ..- attr(*, "names")= chr [1:3] "4459" "13597" "19267"
```

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

1.5 Descriptive Statistics

1.5.0.1 Correlation Plot of Numeric Variables

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



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

```r
table(depression_data$work_pressure)
```

```

    0     1     3 
27864     1     2 
```

```r
table(depression_data$job_satisfaction)
```

```

    0     1     2     3     4 
27859     2     3     1     2 
```

Nearly all the students responded they have the lowest level (0) of both Work Pressure and Job Satisfaction. This likely means the students do not have jobs and are full-time students.

1.5.0.2 Age

```r
summary(depression_data$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   21.00   25.00   25.82   30.00   59.00 
```

Students have a mean age of about 26 years old, with the large majority of students being between 18 and 30 years old.

1.5.0.3 GPA

```r
summary(depression_data$cum_gpa)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   6.290   7.770   7.656   8.920  10.000 
```

```r
# There are a handful of students with 0 GPA
# This slightly affects mean but minimally

depression_data %>% filter(cum_gpa != 0) %>%
  select(cum_gpa) %>%
  summary()
```

```
    cum_gpa
 Min.   : 5.030
 1st Qu.: 6.290
 Median : 7.770
 Mean   : 7.659
 3rd Qu.: 8.920
 Max.   :10.000
```

On this GPA scale where the minimum seems to be 5 or lower and the maximum is 10, students have an average GPA of about 7.7.

1.5.0.4 Depression

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

```

   No   Yes
0.415 0.585
```

We see just below 60% of students responded they experience Depression.

1.5.0.5 Suicidal Thoughts

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

```

   No   Yes
0.367 0.633
```

We see 63.3% of students respond they have had suicidal thoughts.

1.5.0.6 Family Mental Illness

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

```

   No   Yes
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.

1.5.0.7 Work Study Hours

```
summary(depression_data$work_study_hours)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   4.000   8.000   7.158  10.000  12.000 
```

1.5.0.8 Dietary Habits

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

```
  Healthy  Moderate Unhealthy     Other 
   0.2741    0.3556    0.3699    0.0004 
```

More students have moderate and unhealthy dietary habits than healthy dietary habits.

1.5.0.9 Sleep Duration

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

```
Less than 5 hours         5-6 hours         7-8 hours More than 8 hours 
            0.298             0.222             0.263             0.217 
            Other 
            0.001 
```

We see that almost a third of students said they get less than 5 hours of sleep on average.

1.5.0.10 Academic Pressure

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

```
     0      1      2      3      4      5 
0.0003 0.1722 0.1498 0.2673 0.1848 0.2256 
```

There seems to be a pretty even distribution of responses for Academic Pressure, with the highest frequency of responses in 3, 5, and 4 respectively.

1.5.0.11 Financial Stress

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

```
    1     2     3     4     5 
0.184 0.182 0.187 0.207 0.241 
```

The distribution of responses for Financial Stress appears very uniform, with each level receiving almost one fifth of the responses, but the highest frequency of responses was for the highest level of Financial Stress.

1.5.0.12 Study Satisfaction

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

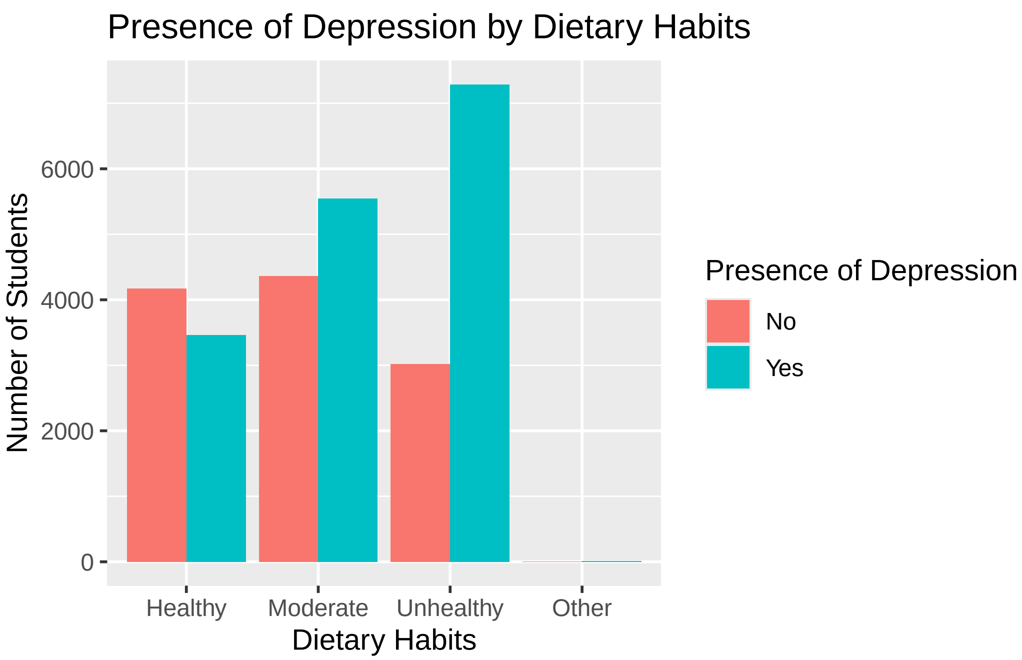
```
     0      1      2      3      4      5 
0.0004 0.1954 0.2094 0.2085 0.2279 0.1585 
```

The distribution of responses for Study Satisfaction are similarly quite uniform, with a small decrease in responses for the highest level of Study Satisfaction.

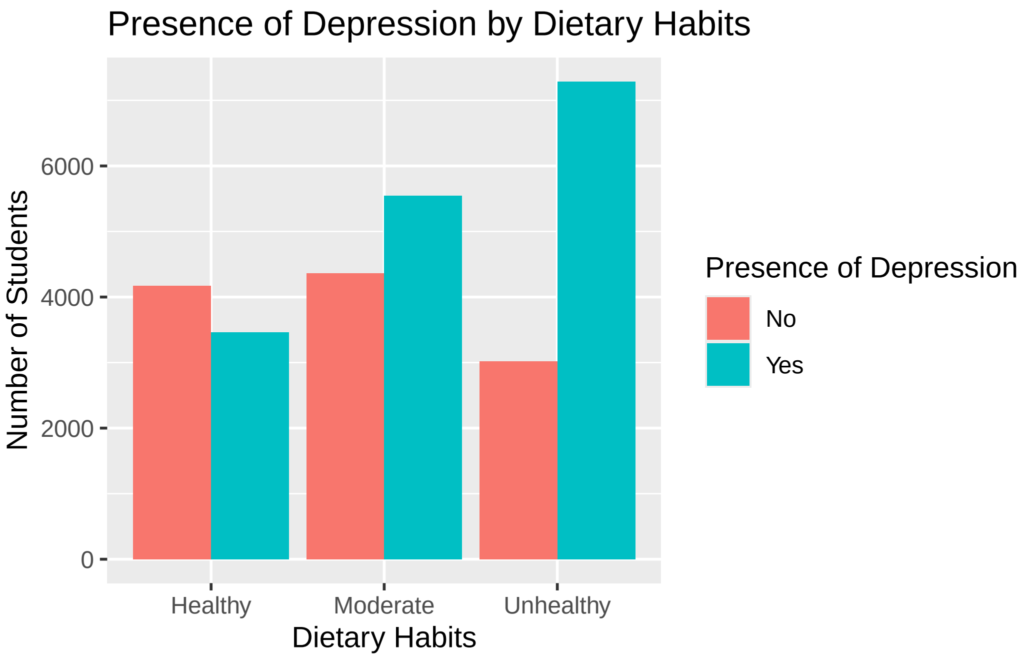
1.6 Data Visualization

- Depression by Dietary Habits

```
depression_data %>%
  ggplot(aes(x=dietary_habits, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Dietary Habits", y = "Number of Students",
       title = "Presence of Depression by Dietary Habits",
       fill = "Presence of Depression")
```

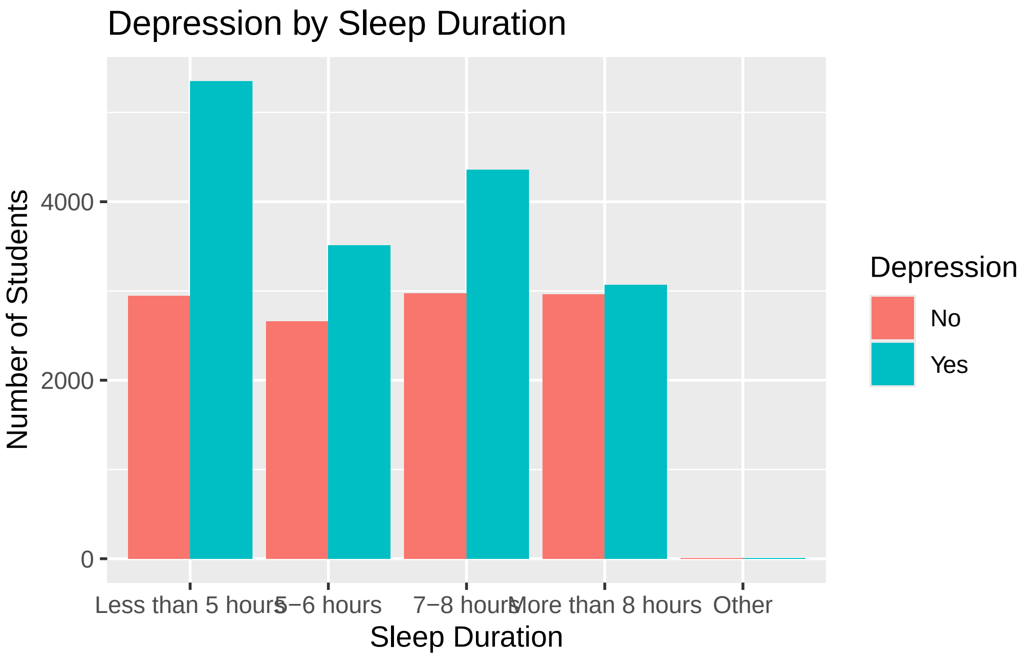


```
depression_data %>%
  filter(dietary_habits != "Other") %>%
  ggplot(aes(x=dietary_habits, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Dietary Habits", y = "Number of Students",
       title = "Presence of Depression by Dietary Habits",
       fill = "Presence of Depression")
```

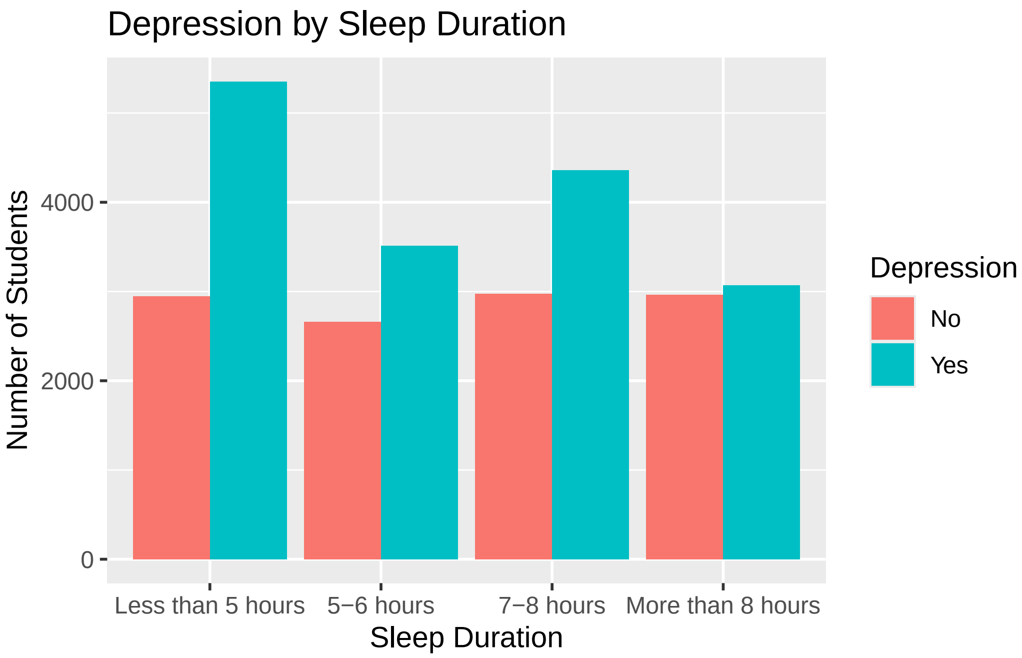


- Depression by Sleep

```r
depression_data %>%
  ggplot(aes(x=sleep_duration, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Sleep Duration", y = "Number of Students",
       title = "Depression by Sleep Duration",
       fill = "Depression")
```

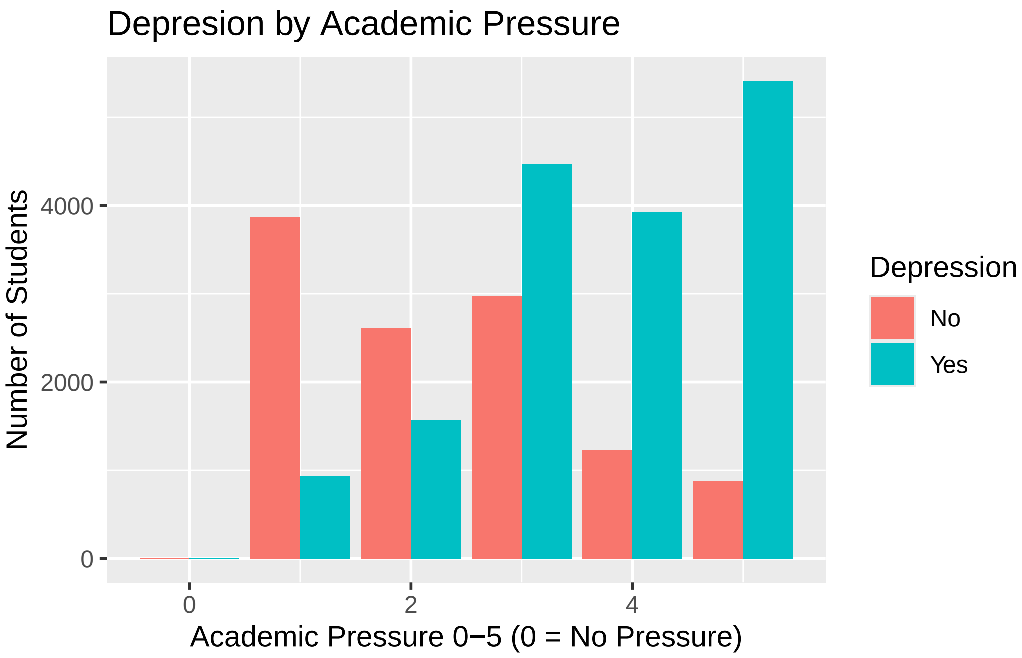


```r
depression_data %>%
  filter(sleep_duration != "Other") %>%
  ggplot(aes(x=sleep_duration, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Sleep Duration", y = "Number of Students",
       title = "Depression by Sleep Duration",
       fill = "Depression")
```

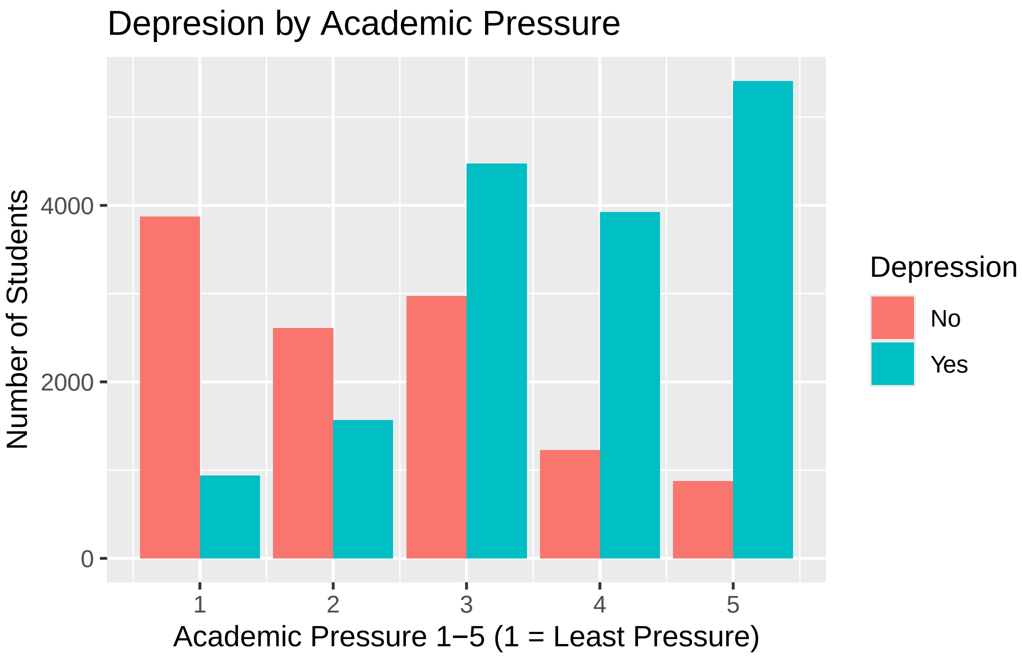


- Depression by Academic Pressure

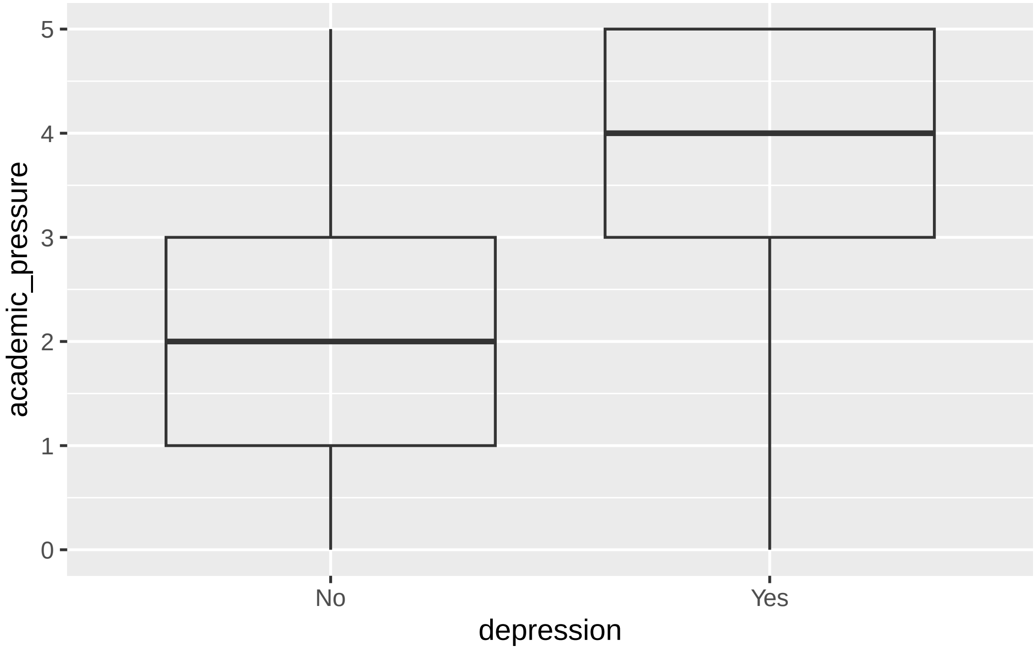
```r
depression_data %>%
  ggplot(aes(x=academic_pressure, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Academic Pressure 0-5 (0 = No Pressure)", y = "Number of Students",
       title = "Depresion by Academic Pressure",
       fill = "Depression")
```



```r
depression_data %>%
  mutate(academic_pressure = case_when(
    academic_pressure == 0 ~ 1,
    TRUE ~ academic_pressure)) %>%
  ggplot(aes(x=academic_pressure, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Academic Pressure 1-5 (1 = Least Pressure)", y = "Number of Students",
       title = "Depression by Academic Pressure",
       fill = "Depression")
```

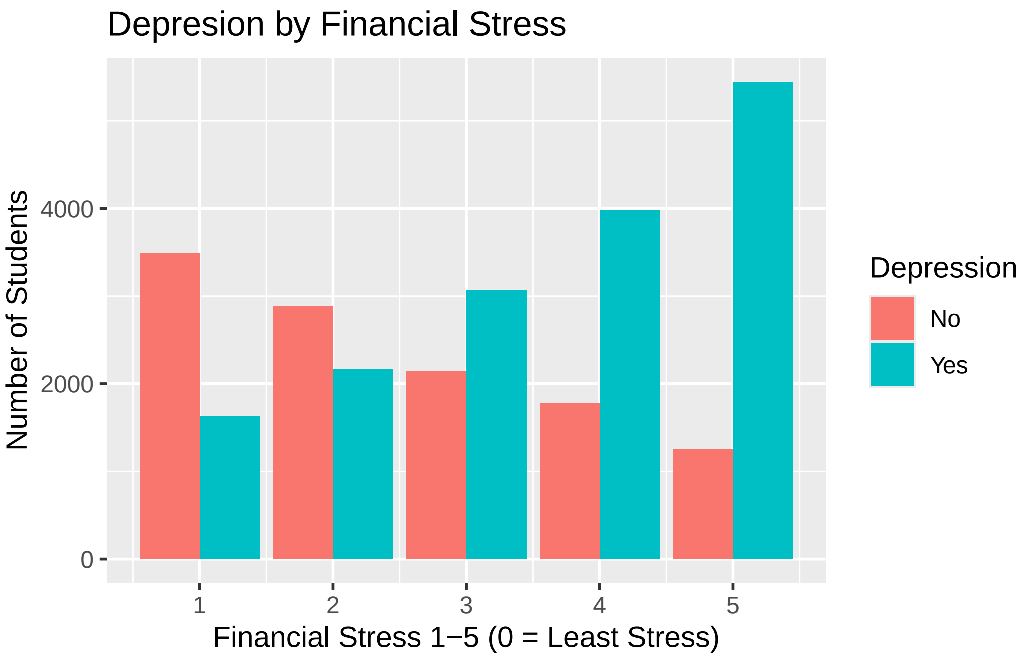


```
depression_data %>%
  ggplot(aes(x = depression, y = academic_pressure)) +
    geom_boxplot(alpha = 0)
```

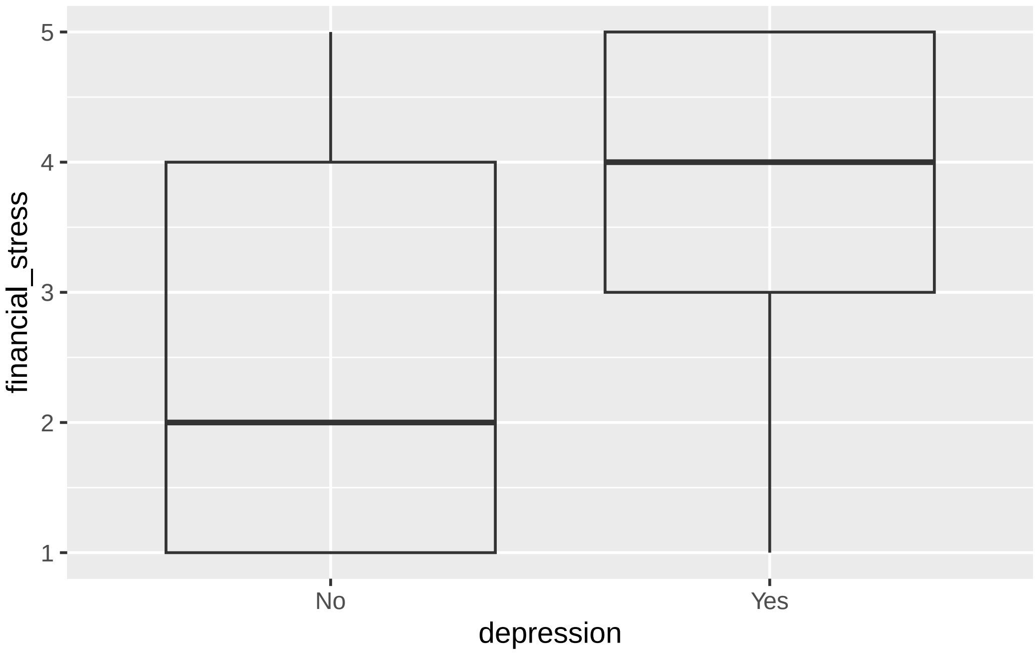


- Depression by Financial Stress

```
depression_data %>%
  ggplot(aes(x=financial_stress, fill = depression)) +
  geom_bar(position = "dodge") +
  labs(x = "Financial Stress 1-5 (0 = Least Stress)", y = "Number of Students",
       title = "Depression by Financial Stress",
       fill = "Depression")
```



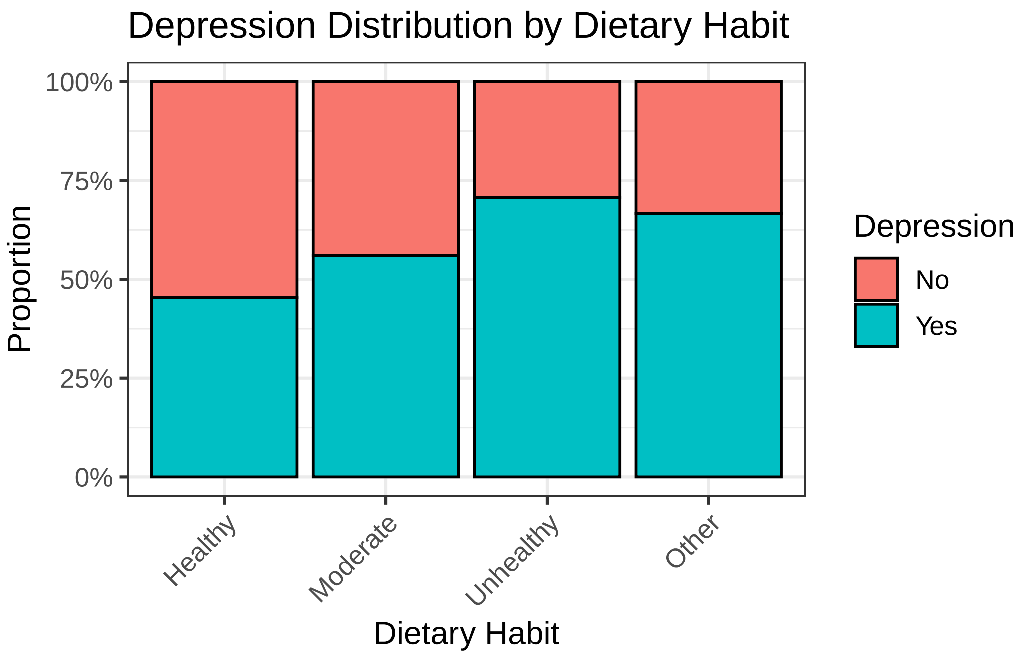
```r
depression_data %>%
  ggplot(aes(x = depression, y = financial_stress)) +
    geom_boxplot(alpha = 0)
```



1.7 Data Visualization (Hayden)

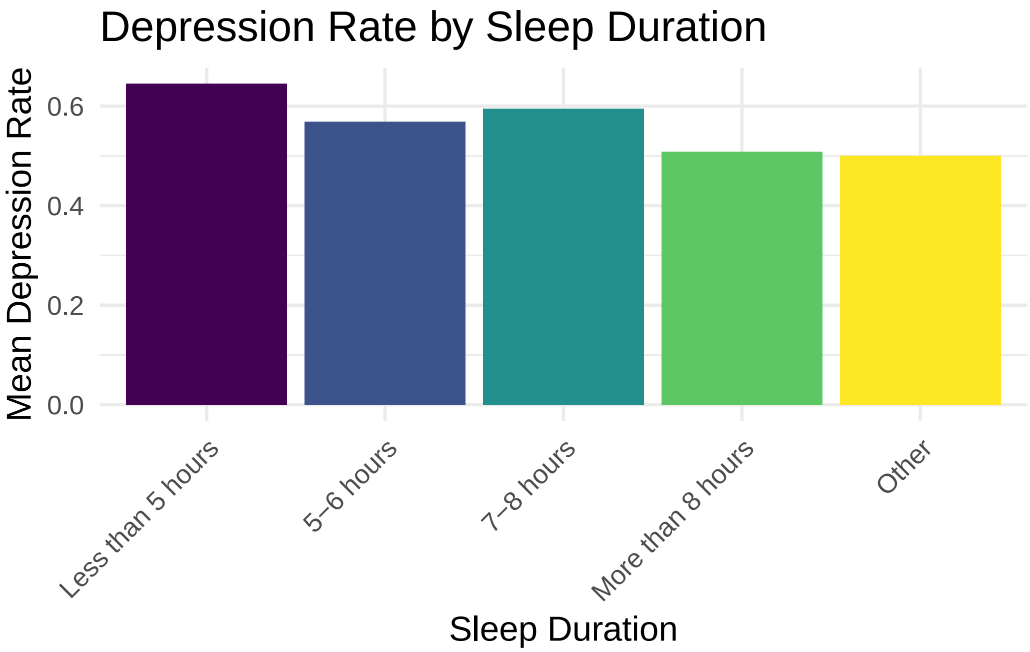
This first graph is a bar plot that helps to answer hypothesis 1 by visualizing the correlation between a healthy diet and depression. The results of this bar plot clearly indicate a strong correlation with a unhealthy diet and rates of depression.

```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))
```



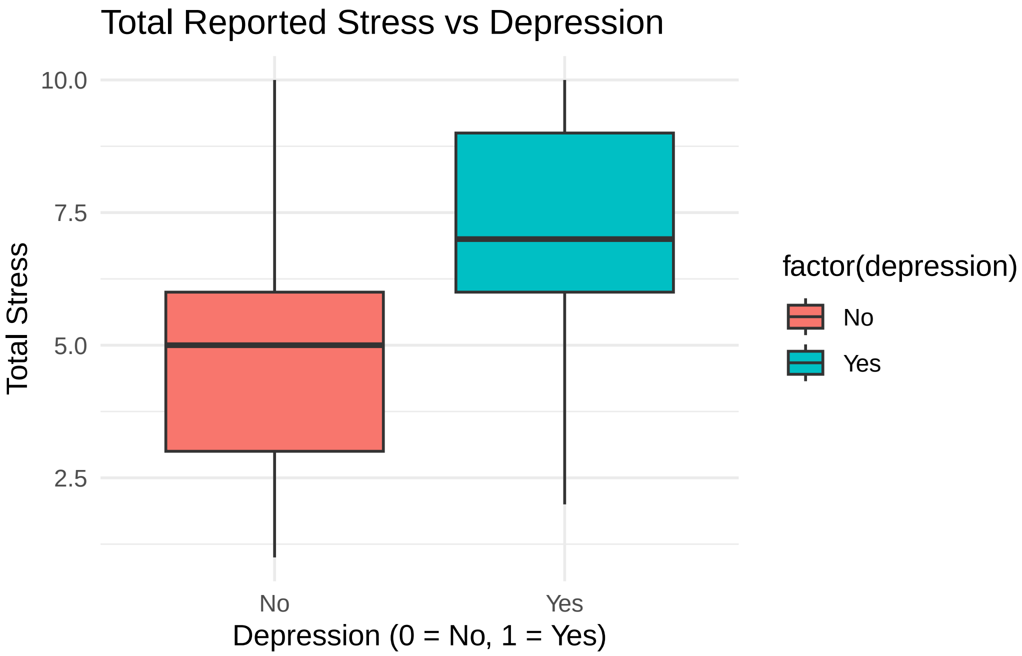
The second graph is a bar plot that helps to answer hypothesis 2 by visualizing the correlation between sleep patterns and depression. The results of this bar plot seem to indicate that people getting less than 5 hours of sleep have a significantly higher rate of depression and people who get more than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

```
# 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.

```
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