Final Project - Step 2 (15 Points)

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

- Valerie De La Fuente (valeriedelafuente)
- Matthew Arteaga (matthewarteaga)
- STUDENT 3 (NetID 3)
- STUDENT 4 (NetID 4)
- STUDENT 5 (NetID 5)

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
```

```
# 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

```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
  )

# 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

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

```
[1] 0
```

There is no missing data present.

1.4.4 Checking Data Types

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

```
'data.frame':	27901 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  : chr  "1.0" "2.0" "1.0" "5.0" ...
 $ fam_mental_illness: chr  "No" "Yes" "Yes" "Yes" ...
 $ depression        : int  1 0 0 1 0 0 0 0 1 1 ...
```

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
  ))

# 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
  ))

```

```r
# 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
  ))

# 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'" ~ "Diploma",
    degree == "ME" ~ "M.Tech",
    degree == "BSc" ~ "B.Sc.",
    degree == "BCA" ~ "B.C.A.",
    degree == "High School" ~ "Other",
    TRUE ~ degree
  ))

# Factorizing the `degree variable`
```

```r
degree_levels <- c(
  "High School",
  "BA", "B.Sc.", "B.Com", "B.C.A.", "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",
  "Others"
)
depression_data <- depression_data %>%
  mutate(degree = factor(degree, levels = degree_levels, ordered = TRUE))

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

# Fixing the `financial_stress` variable
depression_data$financial_stress <- as.numeric(depression_data$financial_stress)

# 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':   27901 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 NA 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 ...
```

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

1.5 Descriptive Statistics

1.6 Data Visualization