National Research University Higher School of Economics

Master's Programme 'Data Analytics and Social Statistics (DASS)'

Exploratory Data Analysis (prof. Batagelj)

Project 3: Large Dataset Analysis

Contents

The project was prepared by ***Timofei Korovin****, DASS student*

Short formulation of the task:

We should choose and analyze the large dataset that contains at least 10000 observations. We should conduct EDA, in particular study our variables and relationships between them. Finally, we need to formulate some suggestions for deeper analysis.

Creation date: 20/04/2025

The last change date: 27/04/2025

Introduction

In this project, we will conduct an exploratory data analysis on a dataset about depression among students, using different demographic, lifestyle, academic indicators. We will examine the variables as well as check the relationship between the variables. Finally, fter the exploratory analysis, we will make suggestions for further analysis

Dataset description

This dataset compiles a wide range of information aimed at understanding, analyzing, and predicting depression levels among students. It is designed for research in psychology, data science, and education, providing insights into factors that contribute to student mental health challenges and aiding in the design of early intervention strategies.

P.S. the description is taken from Kaggle. It can be found with the following link: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset/data

Variables description

“id” - students ids

“age” - students age, numerical (continuous).

“gender” - students gender, nominal.

“profession” - student professional status, nominal.

“academic_pressure” - a measure indicating the level of pressure the student faces in academic settings (1-5 scale), ordinal.

“cgpa” - the cumulative grade point average of the student, reflecting overall academic performance, numerical (continuous).

“study_satisfaction” - an indicator of how satisfied the student is with their studies (1-5 scale), ordinal.

“sleep_duration” - the average number of hours the student sleeps per day, ordinal.

“dietary_habits” - an assessment of the student’s eating patterns and nutritional habits, ordinal.

“work_study_hours” - the average number of hours per day the student dedicates to work or study, numerical (continuous).

“family_history_of_mental_ilness” - Indicates whether there is a family history of mental illness, binary.

“have_you_ever_had_suicidal_thoughts” - an indicator that reflects whether the student has ever experienced suicidal ideation, binary.

“depression” - the target variable that indicates whether the student is experiencing depression, binary.

Data preparation

```
library(tidyverse)
library(ggplot2)
library(corrplot)
```

```r
df <- read.csv("student_depression_dataset.csv")
summary(df)
```

```
##        id            Gender               Age            City
##  Min.   :     2   Length:27901       Min.   :18.00   Length:27901
##  1st Qu.: 35039   Class :character   1st Qu.:21.00   Class :character
##  Median : 70684   Mode  :character   Median :25.00   Mode  :character
##  Mean   : 70442                      Mean   :25.82
##  3rd Qu.:105818                      3rd Qu.:30.00
##  Max.   :140699                      Max.   :59.00
##   Profession        Academic.Pressure Work.Pressure          CGPA
##  Length:27901       Min.   :0.000     Min.   :0.00000   Min.   : 0.000
##  Class :character   1st Qu.:2.000     1st Qu.:0.00000   1st Qu.: 6.290
##  Mode  :character   Median :3.000     Median :0.00000   Median : 7.770
##                     Mean   :3.141     Mean   :0.00043   Mean   : 7.656
##                     3rd Qu.:4.000     3rd Qu.:0.00000   3rd Qu.: 8.920
##                     Max.   :5.000     Max.   :5.00000   Max.   :10.000
##  Study.Satisfaction Job.Satisfaction   Sleep.Duration     Dietary.Habits
##  Min.   :0.000      Min.   :0.000000   Length:27901       Length:27901
##  1st Qu.:2.000      1st Qu.:0.000000   Class :character   Class :character
##  Median :3.000      Median :0.000000   Mode  :character   Mode  :character
##  Mean   :2.944      Mean   :0.000681
##  3rd Qu.:4.000      3rd Qu.:0.000000
##  Max.   :5.000      Max.   :4.000000
##     Degree          Have.you.ever.had.suicidal.thoughts.. Work.Study.Hours
##  Length:27901       Length:27901                          Min.   : 0.000
##  Class :character   Class :character                      1st Qu.: 4.000
##  Mode  :character   Mode  :character                      Median : 8.000
##                                                           Mean   : 7.157
##                                                           3rd Qu.:10.000
##                                                           Max.   :12.000
##  Financial.Stress   Family.History.of.Mental.Illness   Depression
##  Length:27901       Length:27901                      Min.   :0.0000
##  Class :character   Class :character                  1st Qu.:0.0000
##  Mode  :character   Mode  :character                  Median :1.0000
##                                                       Mean   :0.5855
##                                                       3rd Qu.:1.0000
##                                                       Max.   :1.0000
```

```r
table(df$Work.Pressure)
```

```
##
##     0     2     5
## 27898     1     2
```

```r
table(df$Job.Satisfaction)
```

```
##
##     0     1     2     3     4
## 27893     2     3     1     2
```

We will drop variables Work.Pressure and Job.Satisfaction from the final dataset, since they contain a lot of missing values. Next, we format the variable names into a basic format and preprocces our data for further analysis

```
df <- df %>%
  select(id, Gender, Age, Profession, Academic.Pressure, CGPA, Study.Satisfaction,
         Sleep.Duration, Dietary.Habits, Work.Study.Hours,
         Family.History.of.Mental.Illness,
         Have.you.ever.had.suicidal.thoughts.., Depression) %>%
  rename_with(~ .x %>%
                tolower() %>%
                str_replace_all("\\.", "_") %>%
                str_replace_all("__+", "_") %>%
                str_replace("_$", "")
              )
```

```
df_clean <- df  %>%
  mutate(
    sleep_duration = case_when(
      sleep_duration == "'Less than 5 hours'" ~ "low",
      sleep_duration == "'5-6 hours'" ~ "moderate",
      sleep_duration == "'7-8 hours'" ~ "normal",
      sleep_duration == "'More than 8 hours'" ~ "above_normal",
      TRUE ~ NA_character_
    ),
    family_history_of_mental_illness =
      if_else(family_history_of_mental_illness == "Yes", 1, 0),
    have_you_ever_had_suicidal_thoughts =
      if_else(have_you_ever_had_suicidal_thoughts == "Yes", 1, 0)
  )

df_clean <- df_clean %>%
  filter(profession == "Student") %>%
  mutate(cgpa = ifelse(cgpa == 0, NA, cgpa)) %>%
  mutate(work_study_hours = ifelse(work_study_hours == 0, NA, work_study_hours)) %>%
  mutate(academic_pressure = ifelse(academic_pressure == 0, NA, academic_pressure)) %>%
  mutate(study_satisfaction = ifelse(study_satisfaction == 0, NA, study_satisfaction)) %>%
  mutate(dietary_habits = ifelse(dietary_habits == "Others", NA, dietary_habits))
```

Dealing with NA values

After data preprocessing we can detect NA values for our variables

```
colSums(is.na(df_clean))
```

```
##                                 id                             gender
##                                  0                                  0
##                                age                         profession
##                                  0                                  0
##                  academic_pressure                               cgpa
##                                  9                                  9
##                 study_satisfaction                     sleep_duration
```

```
##                                   10                                   18
##                       dietary_habits                    work_study_hours
##                                   12                                 1699
##    family_history_of_mental_illness have_you_ever_had_suicidal_thoughts
##                                    0                                    0
##                           depression
##                                    0
```

For academic_pressure, dietary_habits, cgpa, study_satisfaction and sleep_duration variables we will simply delete NA values, since it is a very small fraction of the values from the total number of observations for these variables. In case of work_study_hours, from our point of view, the best choice is to replace the missing values with the median values. In this case, we will not create outliers and will not lose a significant proportion of observations.

```
df_clean <- df_clean %>%
  filter(
    !is.na(academic_pressure),
    !is.na(dietary_habits),
    !is.na(cgpa),
    !is.na(study_satisfaction),
    !is.na(sleep_duration)
  ) %>%
  mutate(
    work_study_hours = ifelse(
      is.na(work_study_hours),
      median(work_study_hours, na.rm = TRUE),
      work_study_hours
    )
  )

colSums(is.na(df_clean))
```

```
##                                   id                               gender
##                                    0                                    0
##                                  age                           profession
##                                    0                                    0
##                   academic_pressure                                 cgpa
##                                    0                                    0
##                  study_satisfaction                       sleep_duration
##                                    0                                    0
##                       dietary_habits                    work_study_hours
##                                    0                                    0
##    family_history_of_mental_illness have_you_ever_had_suicidal_thoughts
##                                    0                                    0
##                           depression
##                                    0
```

Now our dataset is ready for analysis.

Central tendency measures

For our continuous variables lets calculate some central tendency measures (mean, median).

```r
num_vars <- c("age", "cgpa", "academic_pressure")
df_clean %>%
  select(all_of(num_vars)) %>%
  summary()
```

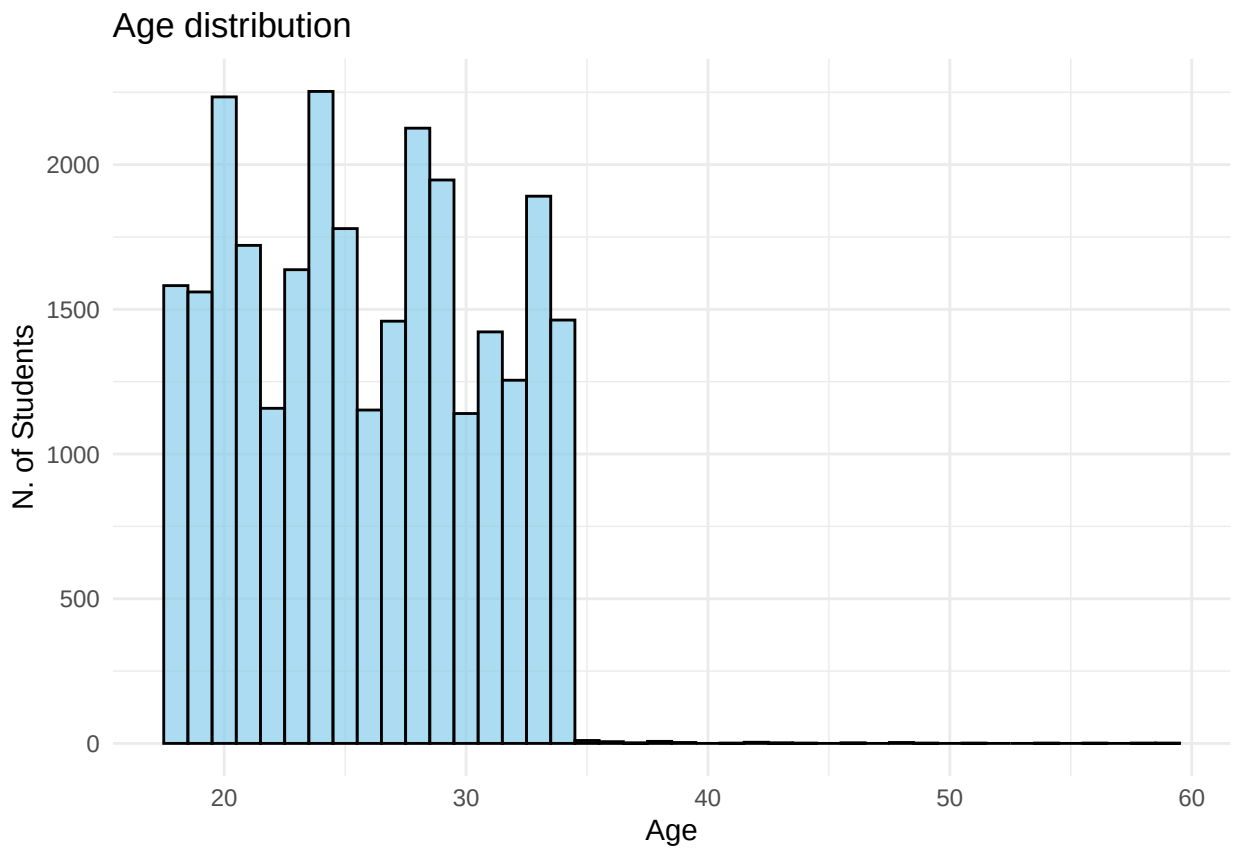
```
##       age             cgpa        academic_pressure
##  Min.   :18.00   Min.   : 5.030   Min.   :1.000
##  1st Qu.:21.00   1st Qu.: 6.290   1st Qu.:2.000
##  Median :25.00   Median : 7.770   Median :3.000
##  Mean   :25.82   Mean   : 7.659   Mean   :3.142
##  3rd Qu.:30.00   3rd Qu.: 8.920   3rd Qu.:4.000
##  Max.   :59.00   Max.   :10.000   Max.   :5.000
```

For our nominal and ordinal variables we will use visualizations to demonstrate distributions.

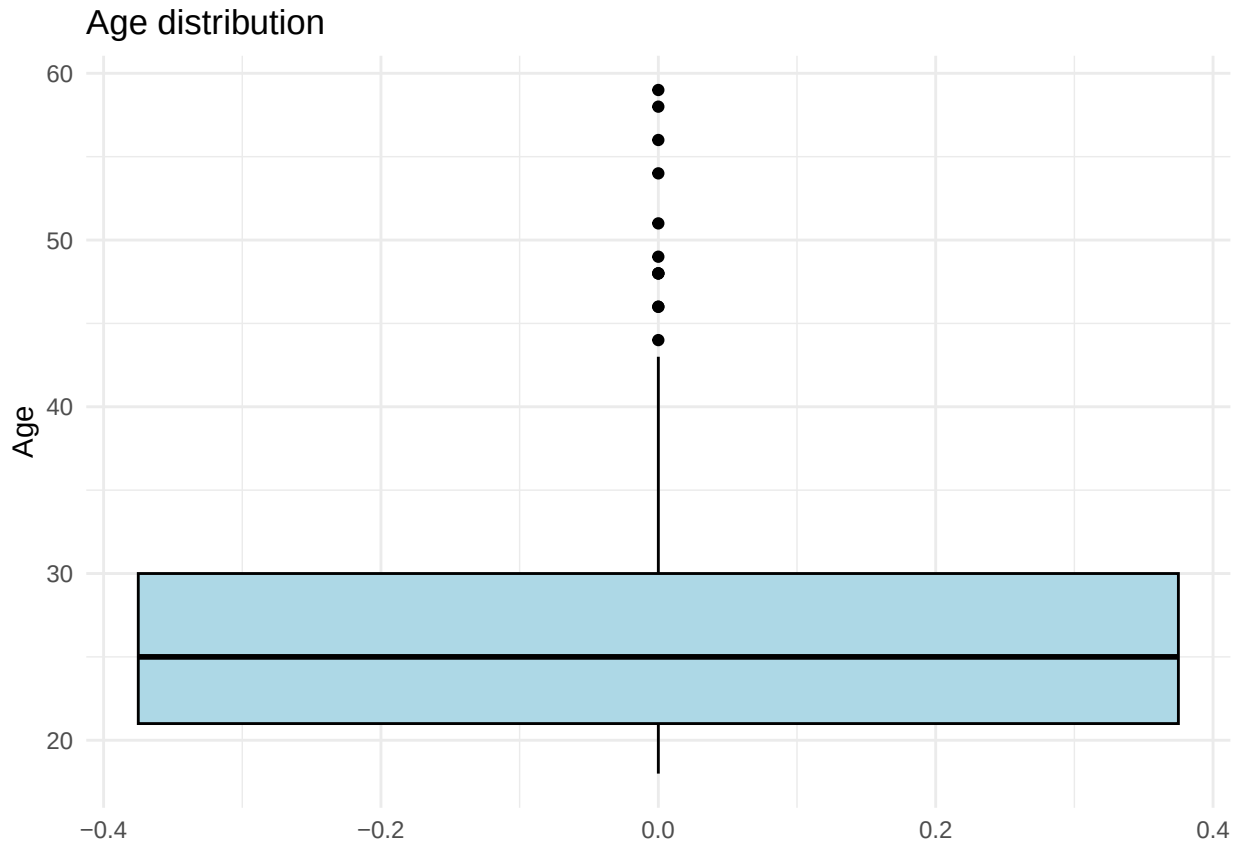
Variables distribuitons

In this section we will examine the distributions of our variables. In addition, we will plot our ordinal and nominal variables against target variables (depression).

Age distribution



The age distribution in the sample is heavily skewed towards younger respondents: the vast majority of participants are between 18 and 35 years of age. Which is logical, since dataset mainly describes school/college/university students. In addition, in the range the values are distributed relatively uniformly. However, to better visualize outliers in our data we will build boxplot.



As it can be seen on the graph, the median age is 25. Indeed values slightly higher than 40 are clear outliers. We suggest that these values can be simply deleted from our dataset, since we are more interested in studying the tendency of depression in young respondents.

However, let's also use the table() function to check the distribution of unique values in the age variable.

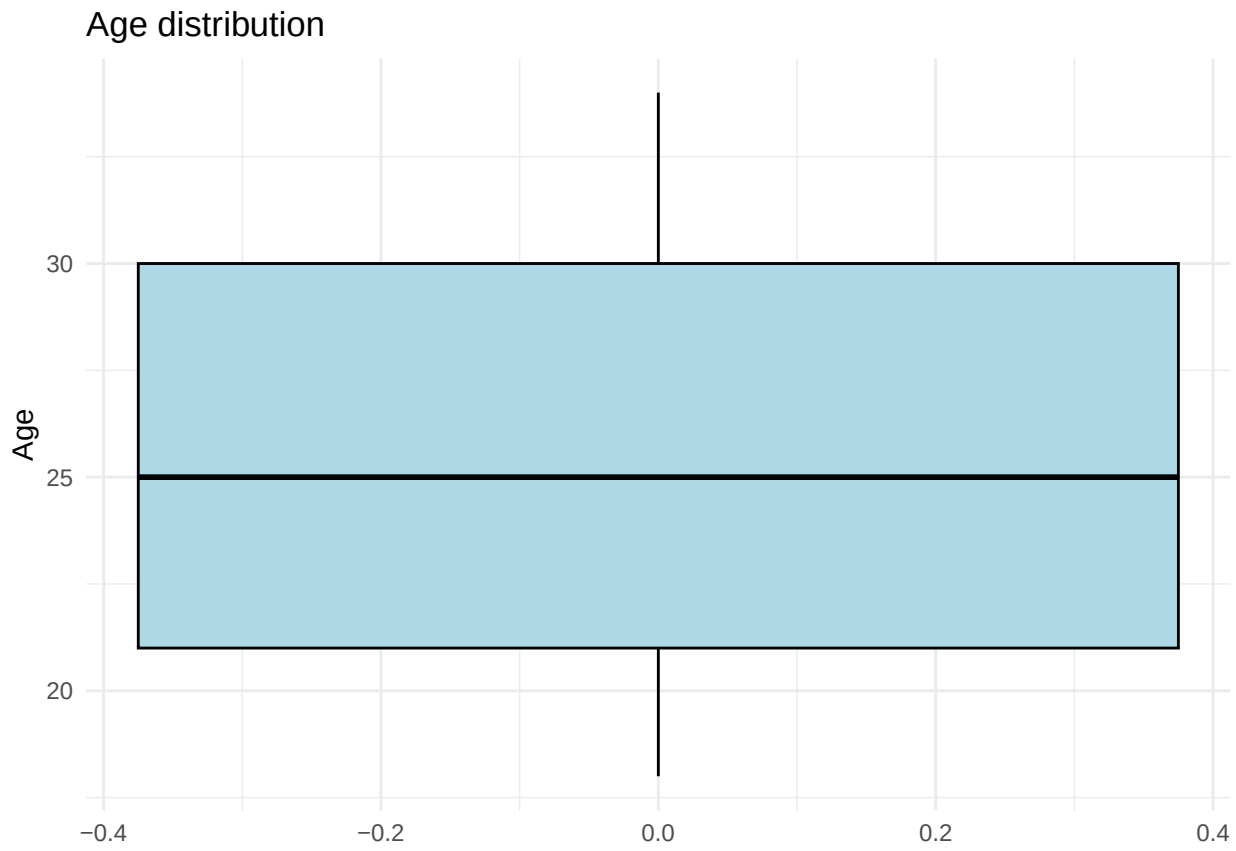
```
table(df_clean$age)
```

```
##
##   18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33
## 1582 1560 2234 1721 1158 1637 2253 1779 1152 1459 2126 1947 1140 1422 1255 1891
##   34   35   36   37   38   39   41   42   43   44   46   48   49   51   54   56
## 1463   10    6    2    7    3    1    4    2    1    2    3    1    1    1    1
##   58   59
##    1    1
```

Our suggestion is that all values above 34 should be excluded from the data set, since they are clear outliers.

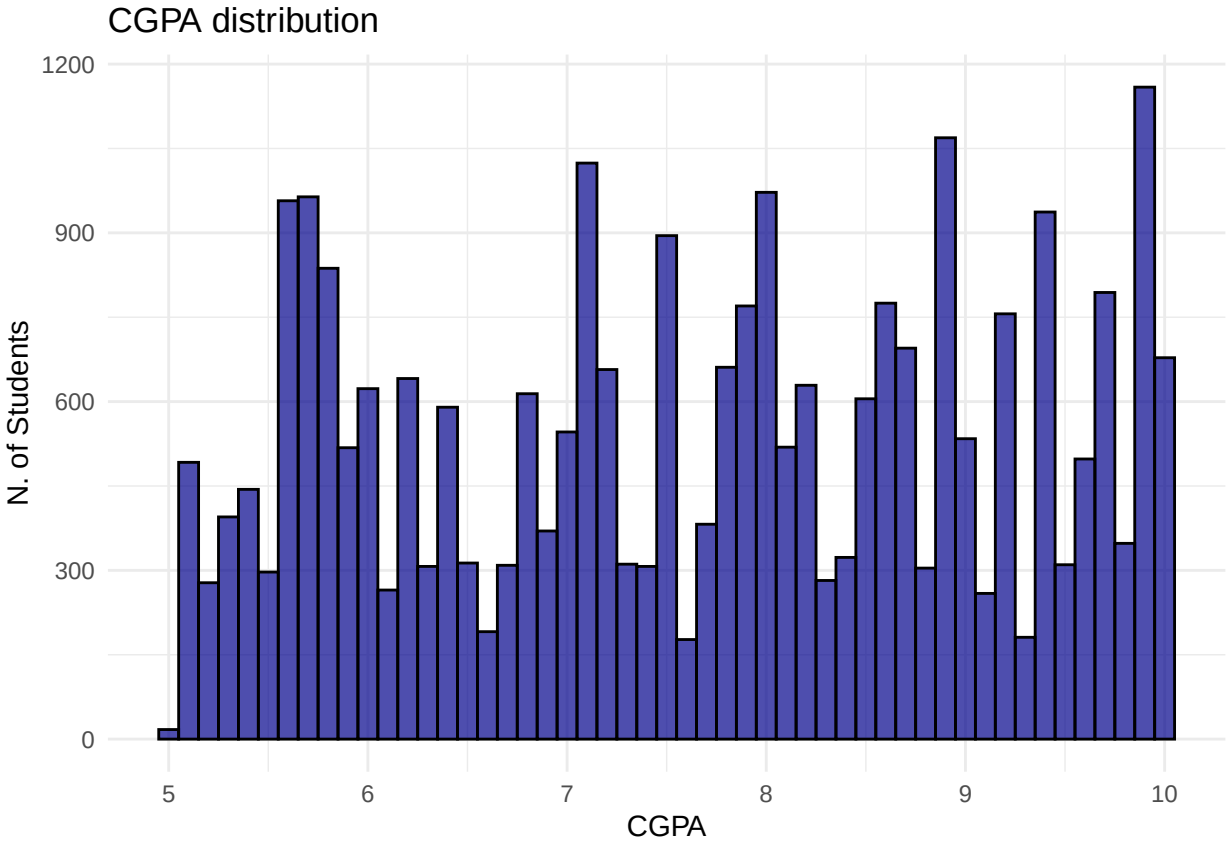
```
df_clean <- df_clean %>%
  filter(age <= 34)
```

```r
ggplot(df_clean, aes(y = age)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Age distribution",
       y = "Age") +
  theme_minimal()
```

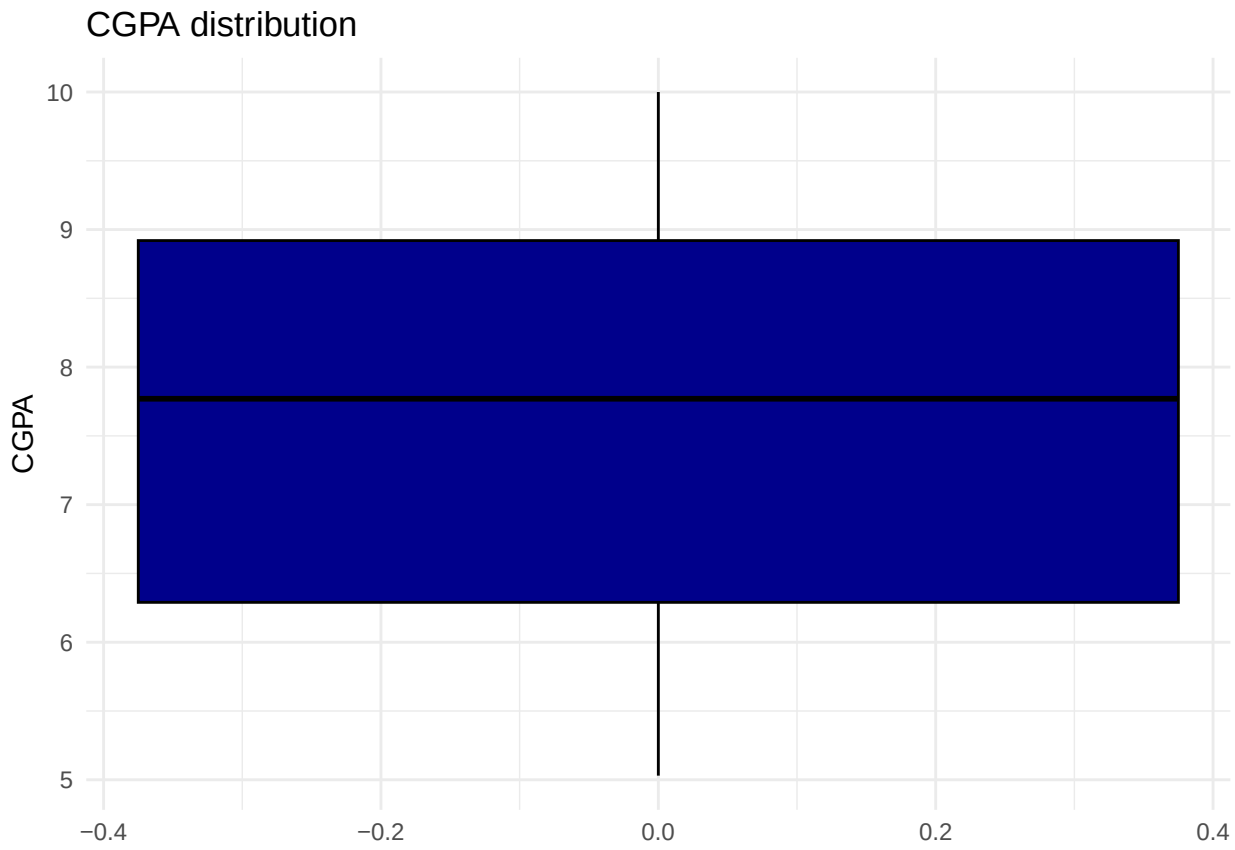


It is evidenced from the box plot above that we have no outliers for “age” variable anymore.

CGPA distribution

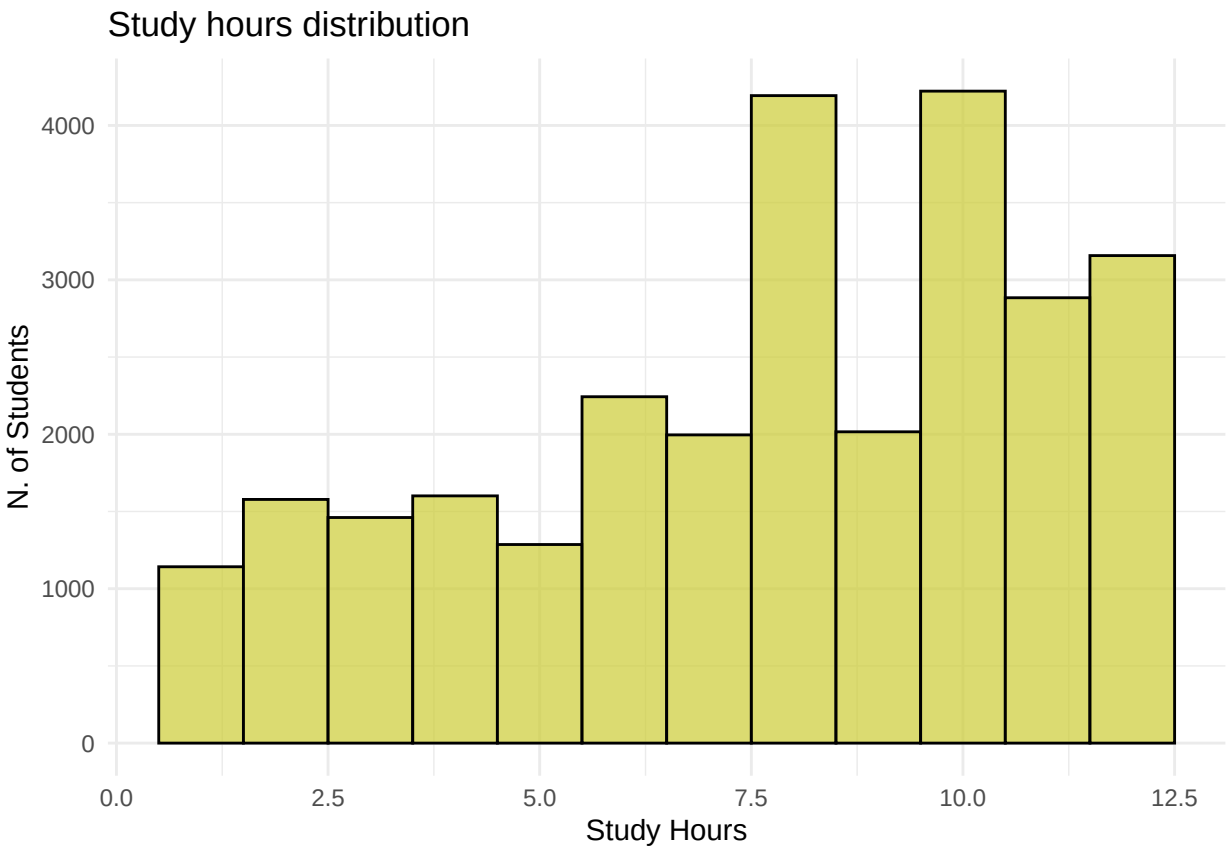


Overall, CGPA is fairly uniformly distributed across the whole sample, there is no clear bias toward high or low scores. Futhermore, no extreme values below 5 or above 10 have been observed. However, to be sure lets also build box plot again.

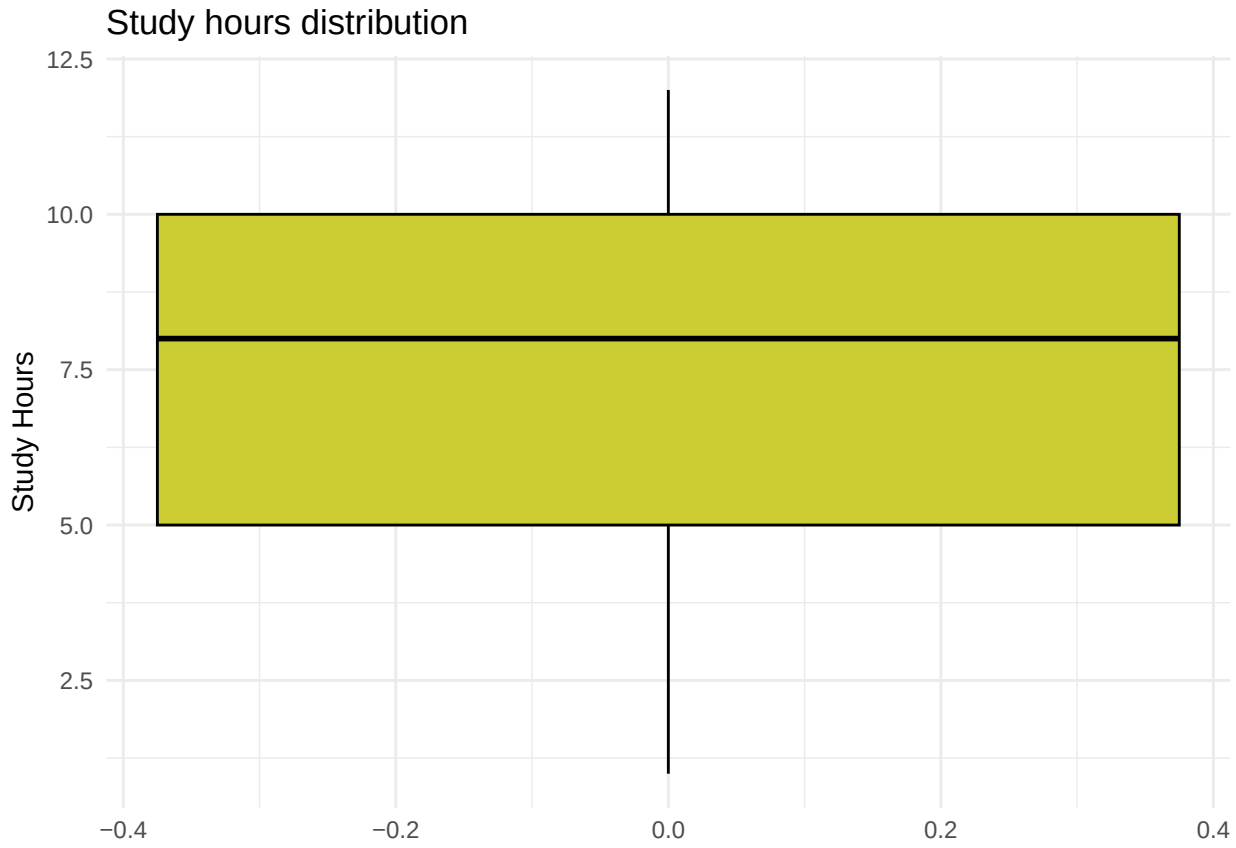


The median CGPA is about 7.8. No outliers have been detected on the graph. IQR is approximately between 6.5 and 9, indicating that the biggest share of students have CGPA in that particular range.

Study Hours distribution

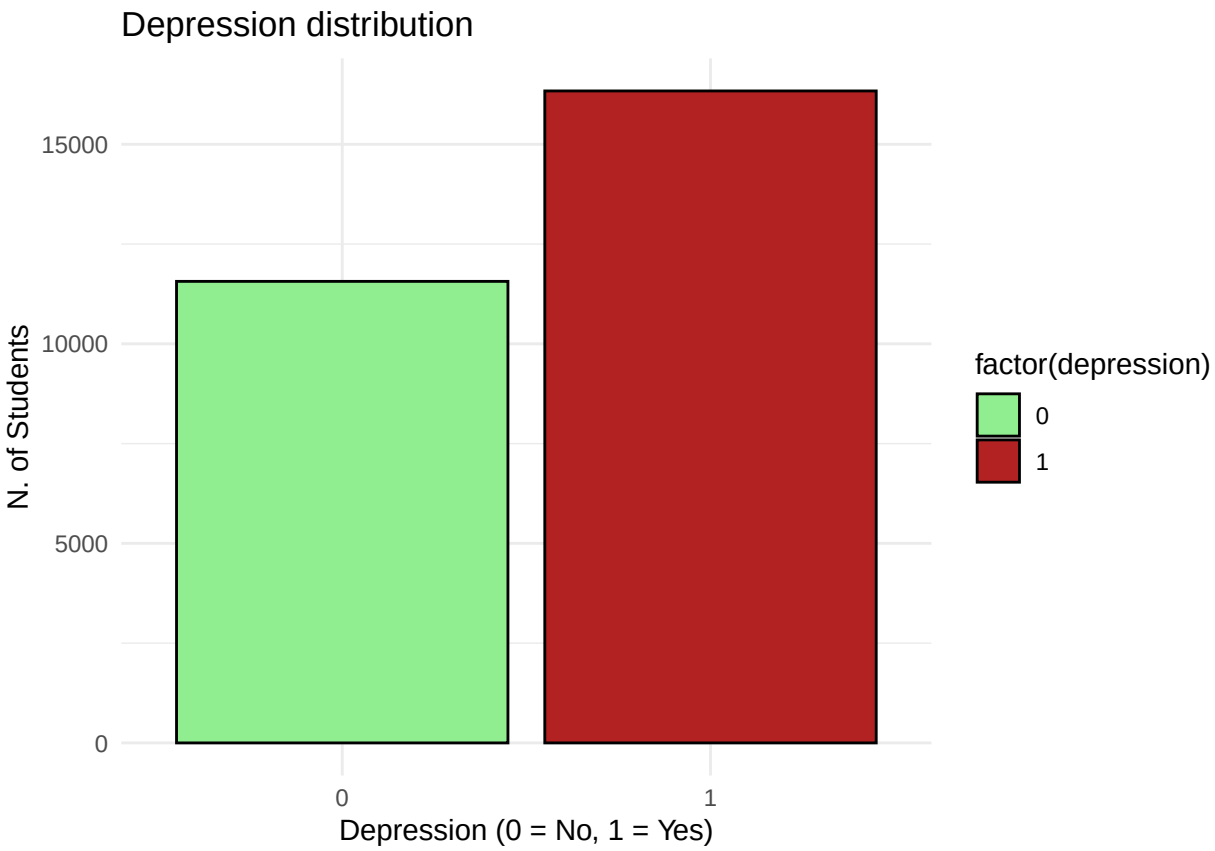


According to the histogram above, the distribution for Study Hours is uneven. In general, there is an increase in the number of students as the number of hours increases up to 10 (clear pick value in our data), after which the number students decreases. Thus, the graph shows that more students tend to spend a significant amount of time studying (more than 5 hours). None extreme values have been observed.



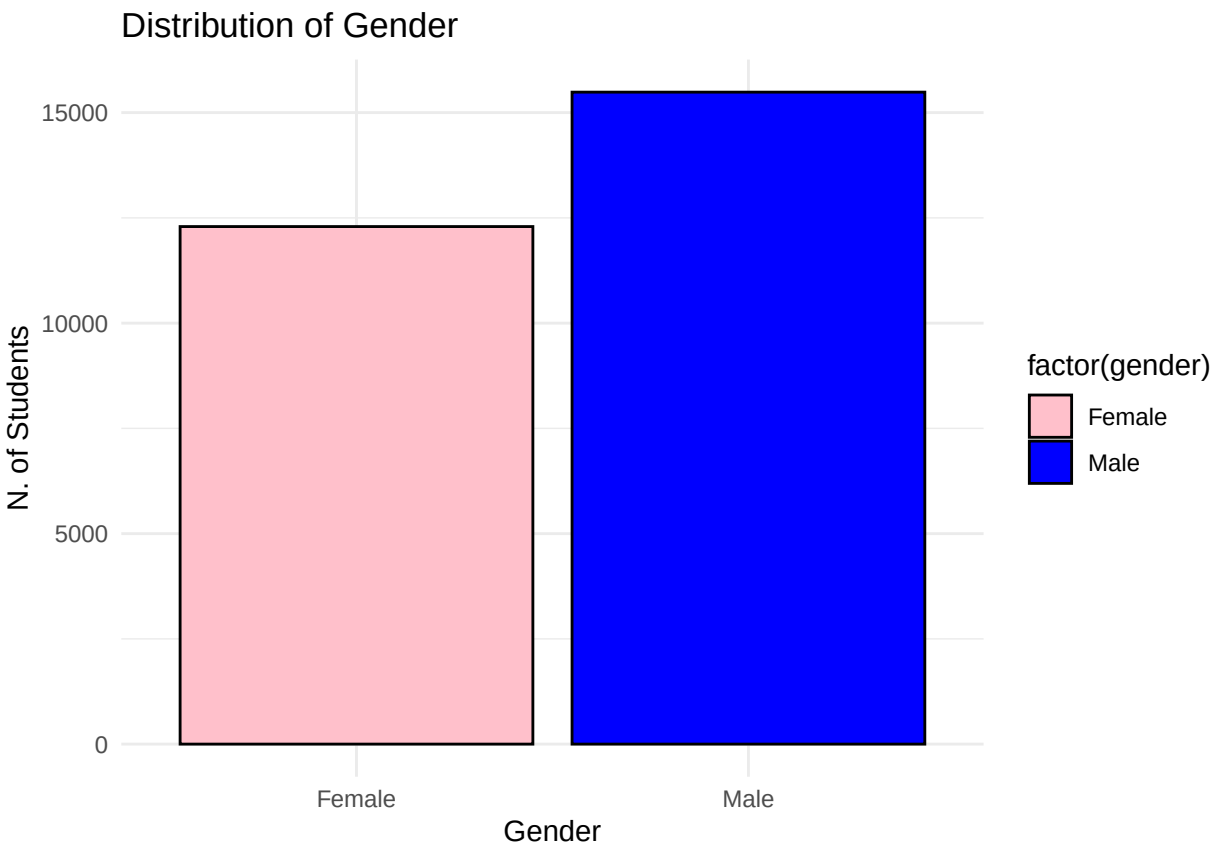
The median value of Study hours is slightly above 7,5. The IQR is between 5 to 10 hours, which confirms that most students study for more than 5 hours per day.

Depression distribution



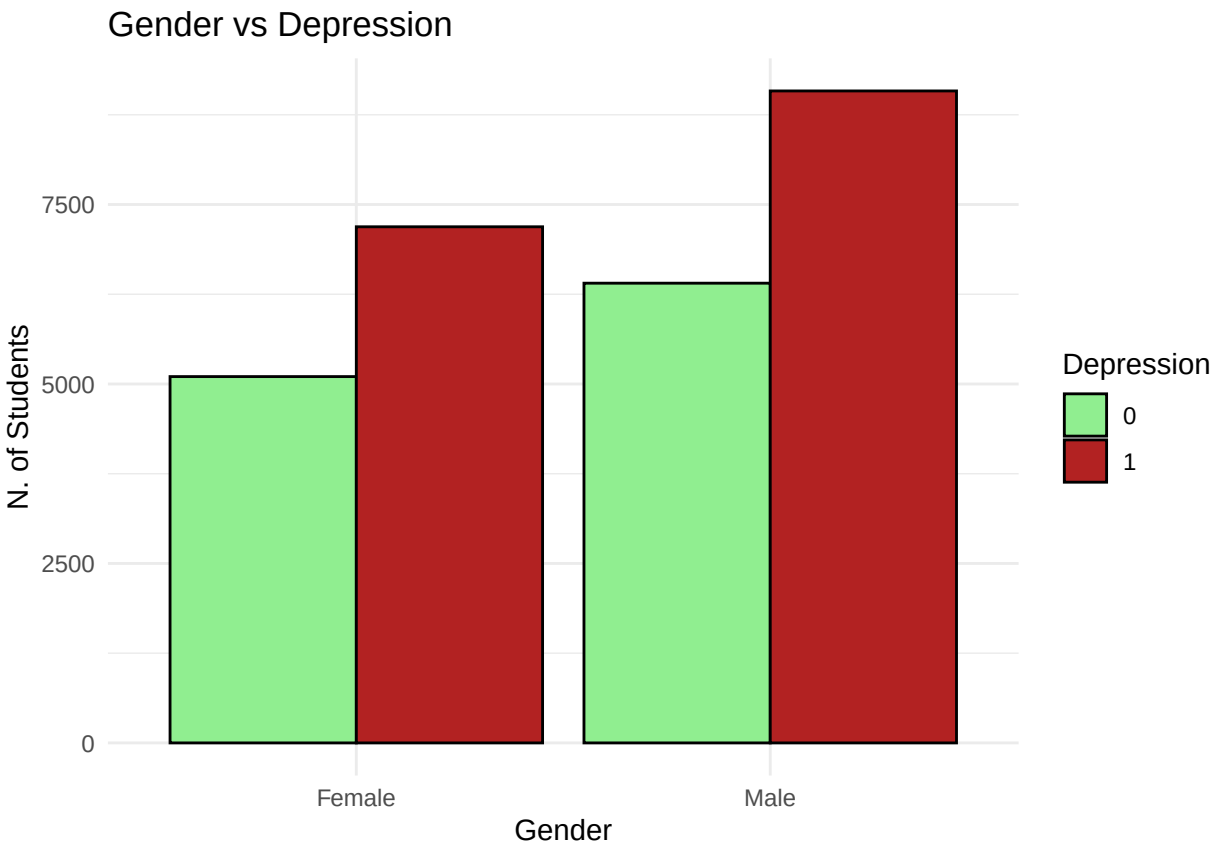
The distribution of depression in the dataset is imbalanced. The number of students with depression (the red bar, around 16.000 students) is higher than the number of students without depression (the light green bar, around 11.000 students).

Gender distribution



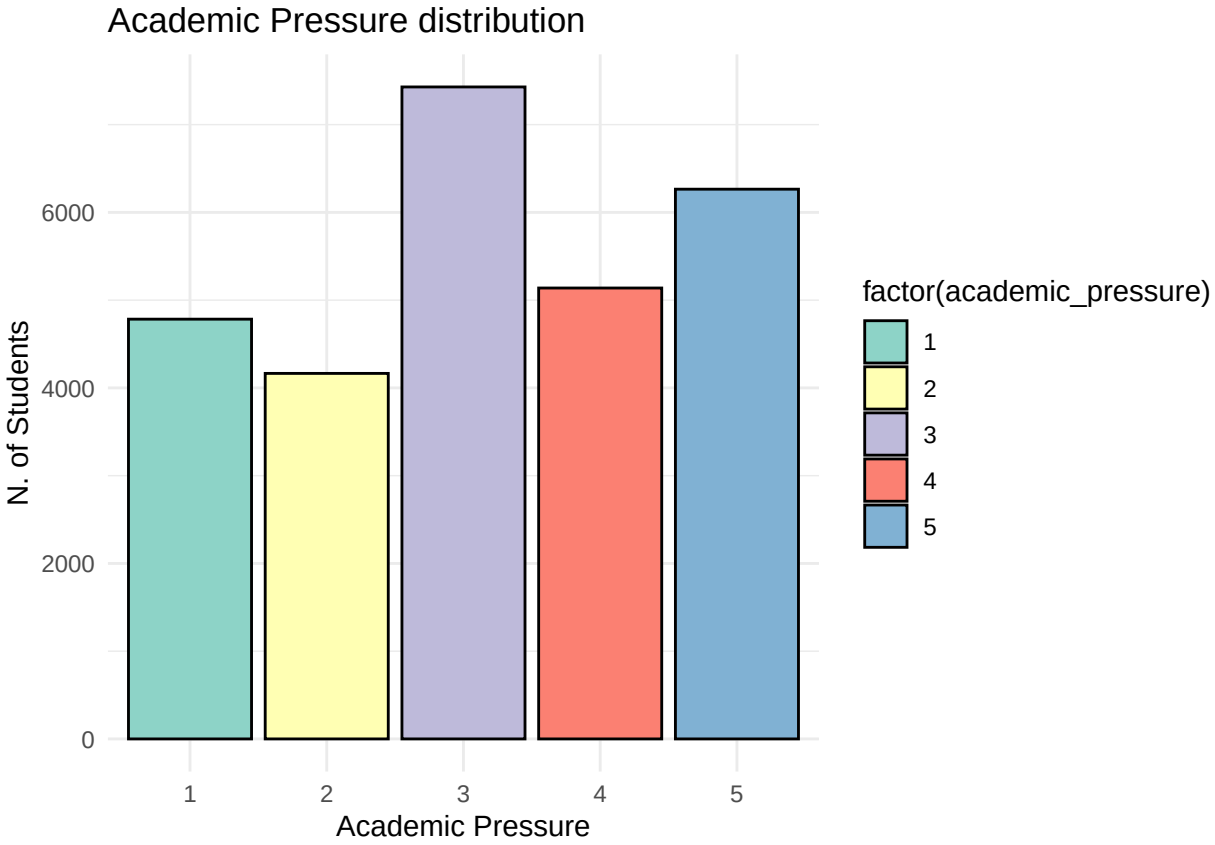
The gender distribution is imbalanced as well. The number of males is slightly below 15.000, whereas there are about 12.500 females students.

Gender vs Depression



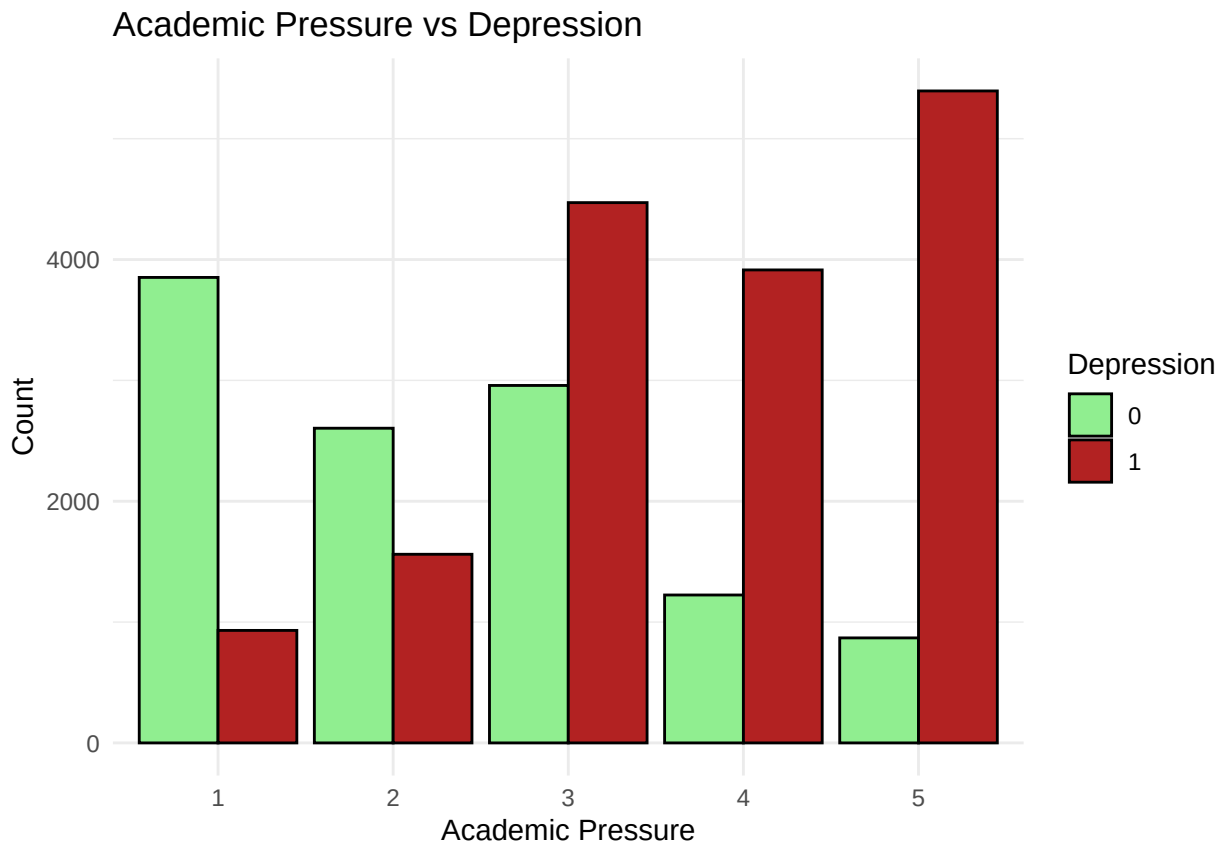
In context of the depression distribution by gender factor, both genders shows higher proportion of students with depression compared to those without it. In addition, males have slightly higher rate of depression compared to females.

Academic Pressure distribution



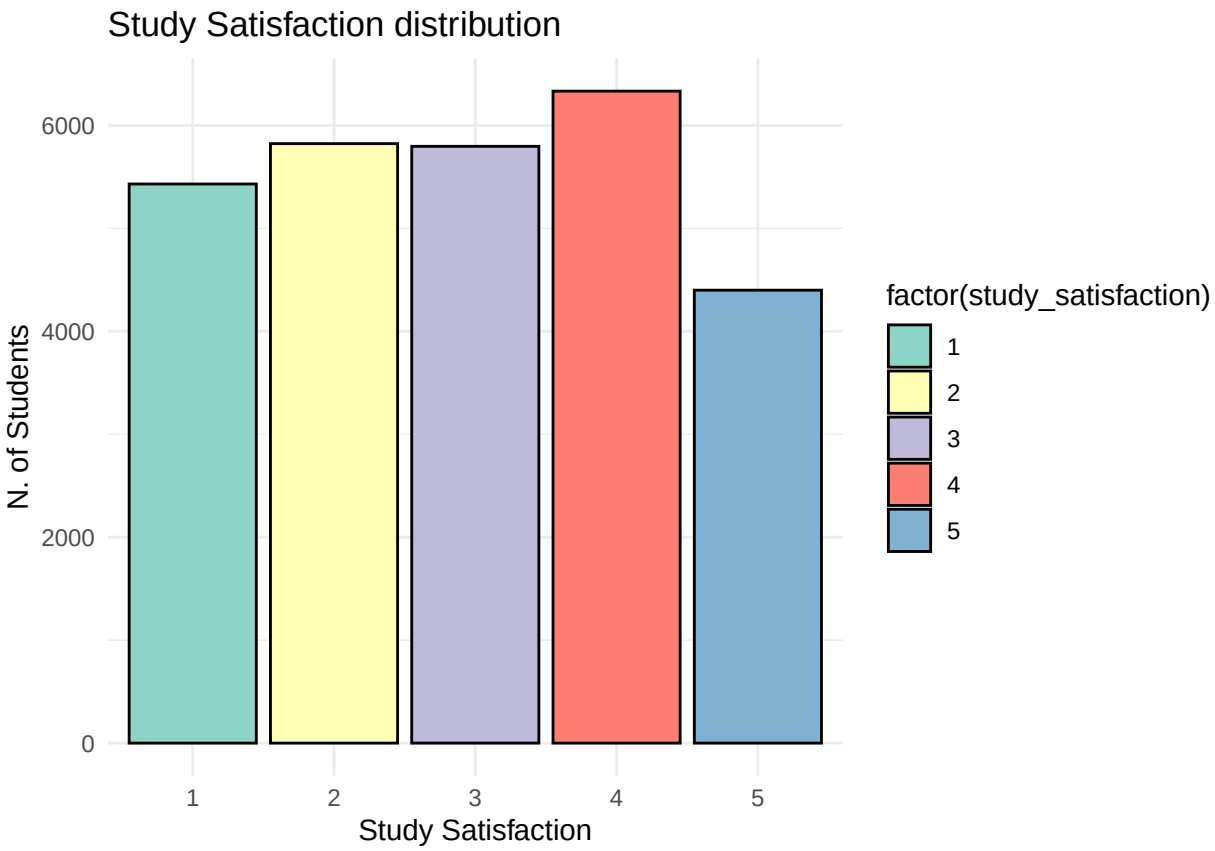
The bar chart above demonstrates the distribution of Academic Pressure. It is noticeable that 3 category demonstrate the highest count, with number exceeding 7000 students. It is followed by the 5th category with about 6000 students for this level of pressure. Level 1 and 4 demonstrate almost similar shares, about 5000 students. 2nd level demonstrate the smallest number of students.

Academic Pressure vs Depression



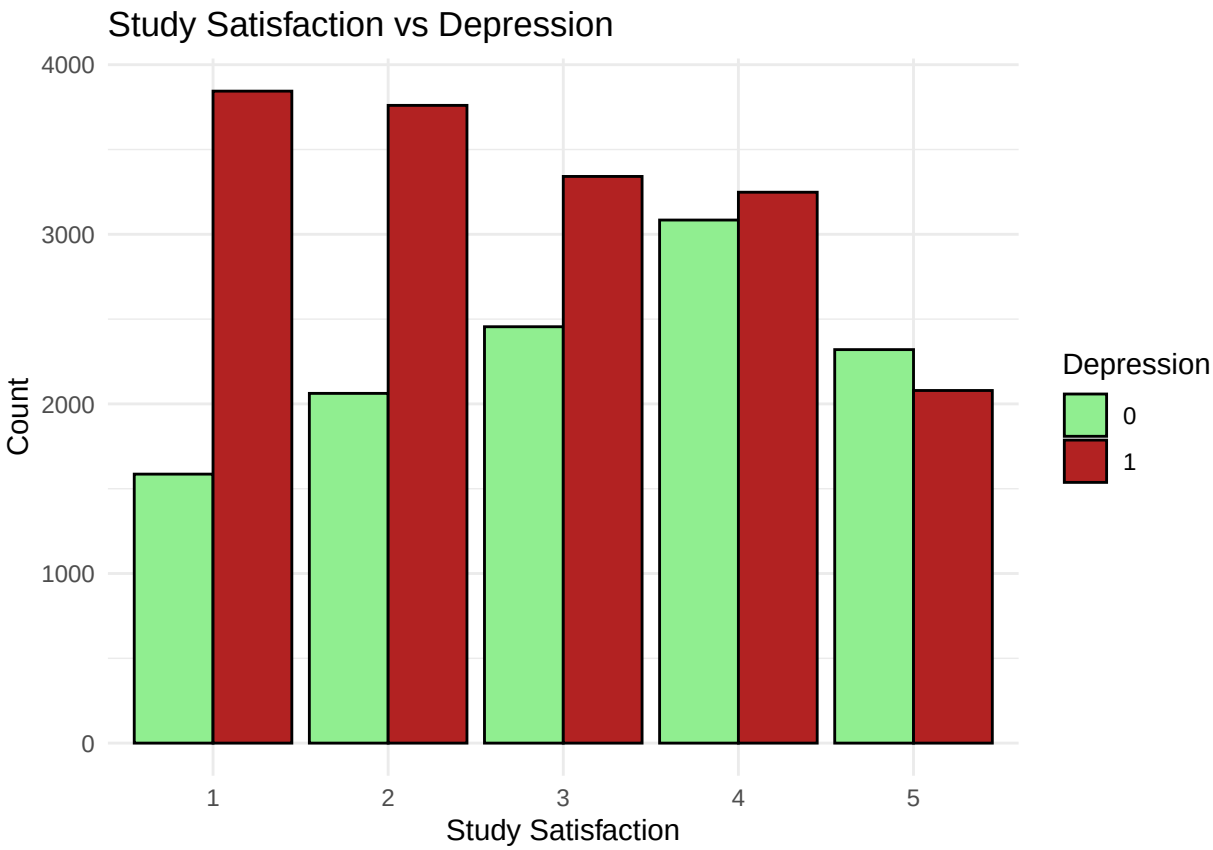
The bar chart demonstrates the Depression levels for each category of Academic Pressure. Overall, it can be seen that while academic pressure increases, the share of students with depression also increases. However, it is important to note, that this visualization does not prove casual relationship between these two variables.

Study Satisfaction



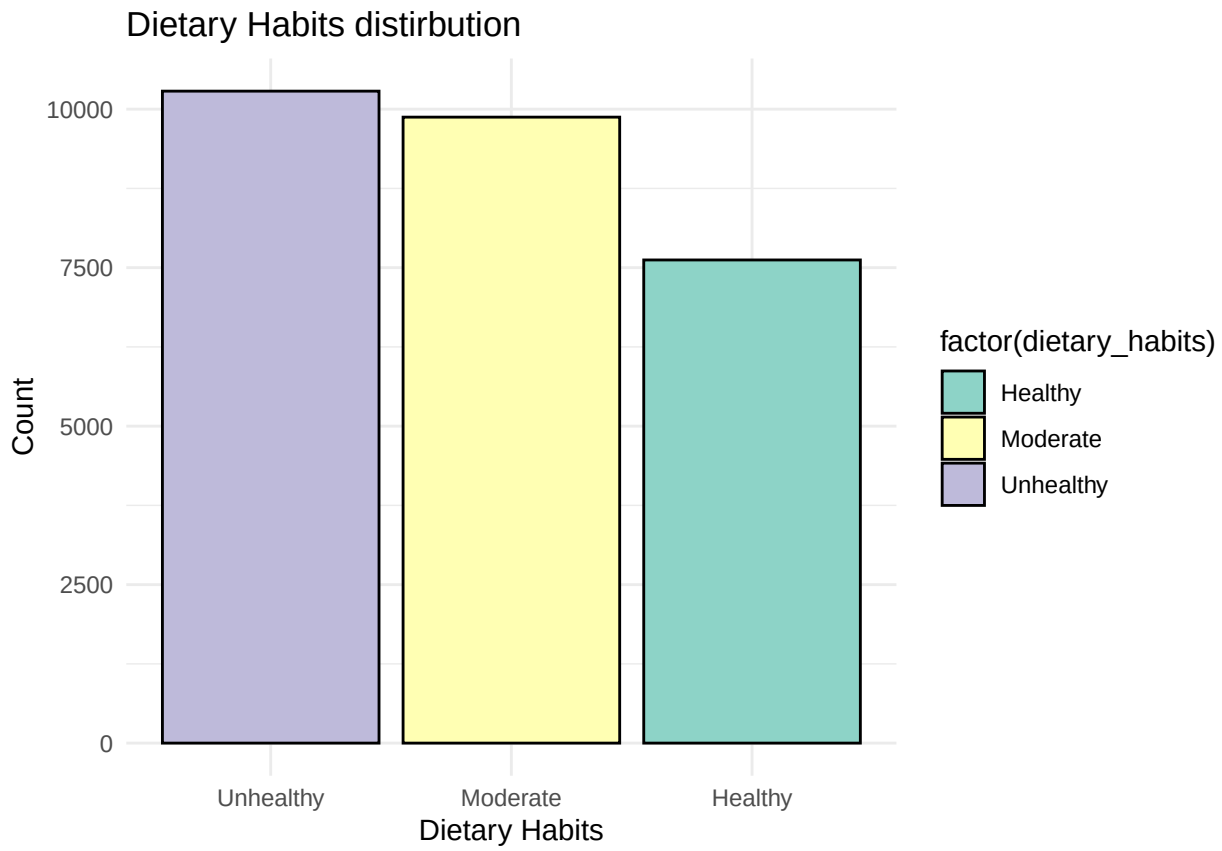
The provided bat chart demonstrates distribution of study satisfaction levels. The 4th level demonstrate the highest share, exceeding 6000 students. It is followed by the satisfaction levels 2 and 3, that have the same count of students (approximately 5800 students). 1st level demonstrate slightly lower number of students, approximately 5400 students. The smallest share of students (approximately 4400 students) belongs to the 5th level.

Study Satisfaction vs Depression



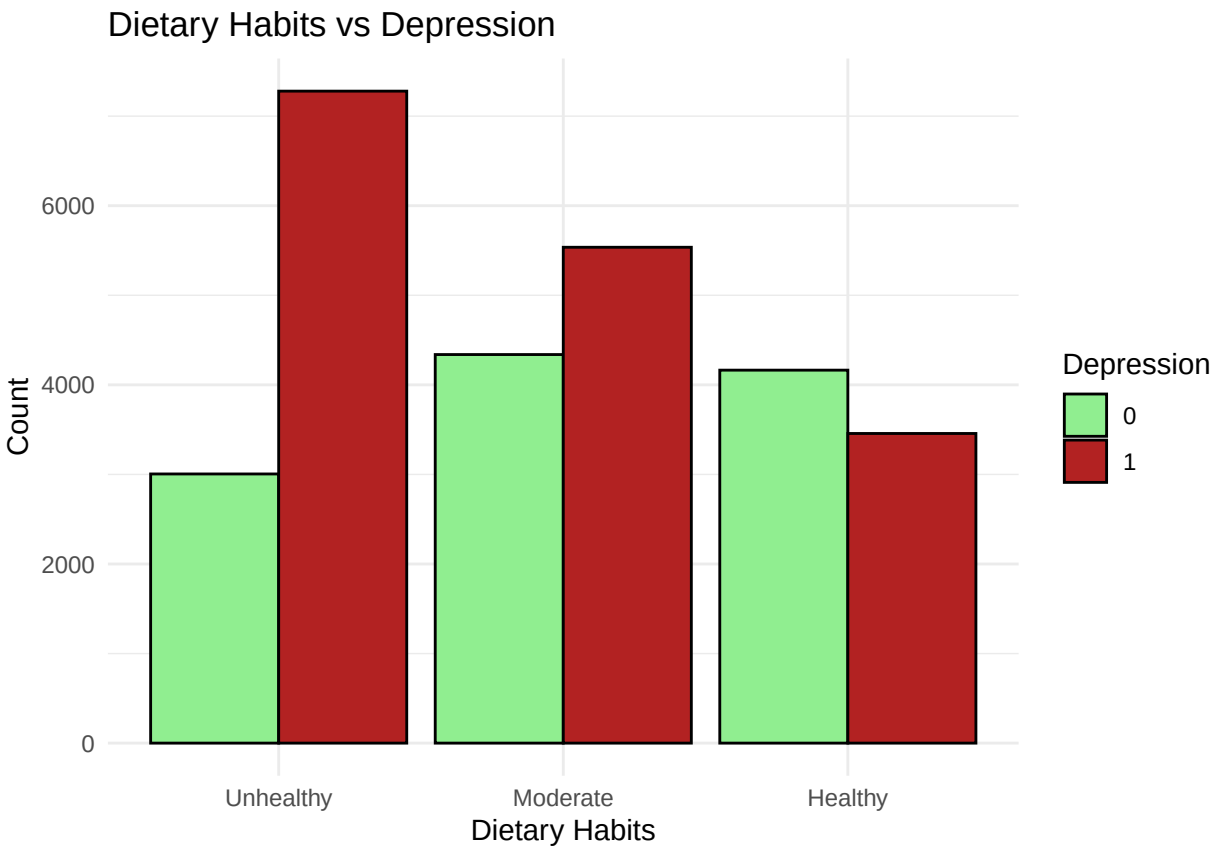
The bar chart above demonstrates the shares of depression for each category of Study Satisfaction level. Overall, it can be seen that the share of students with depression decreases with higher satisfaction level.

Dietary Habits distirbution



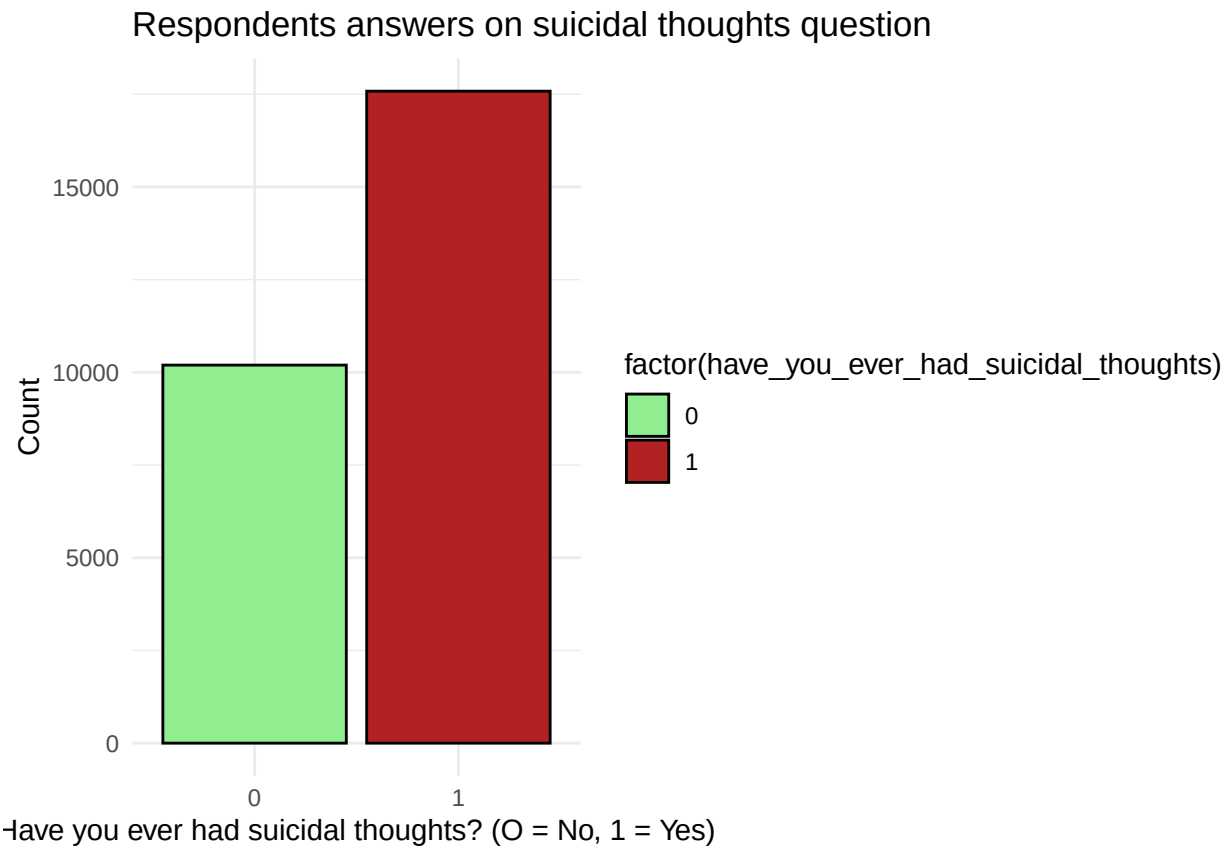
As it can be seen on the graph, the biggest part of students have unhealthy dietary habits, the number exceeds 10000 students. It is followed by the moderate level of dietary habits, with number of students slightly below 10000. The lowest number of students, around 7500, have healthy dietary habits.

Dietary Habits vs Depression



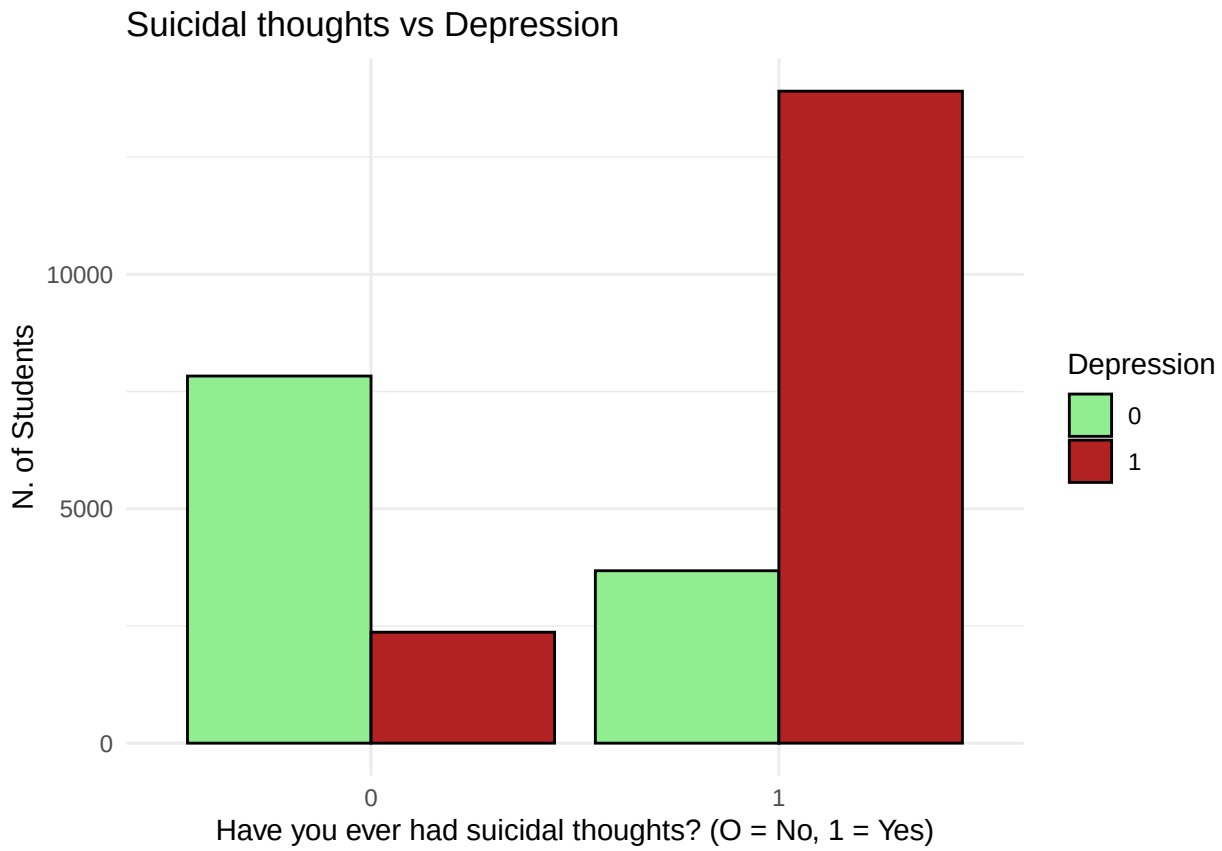
Overall, it evidenced from the graph above that the positive depression cases decreases with healthier dietary habits.

Suicidal Thougths distribution



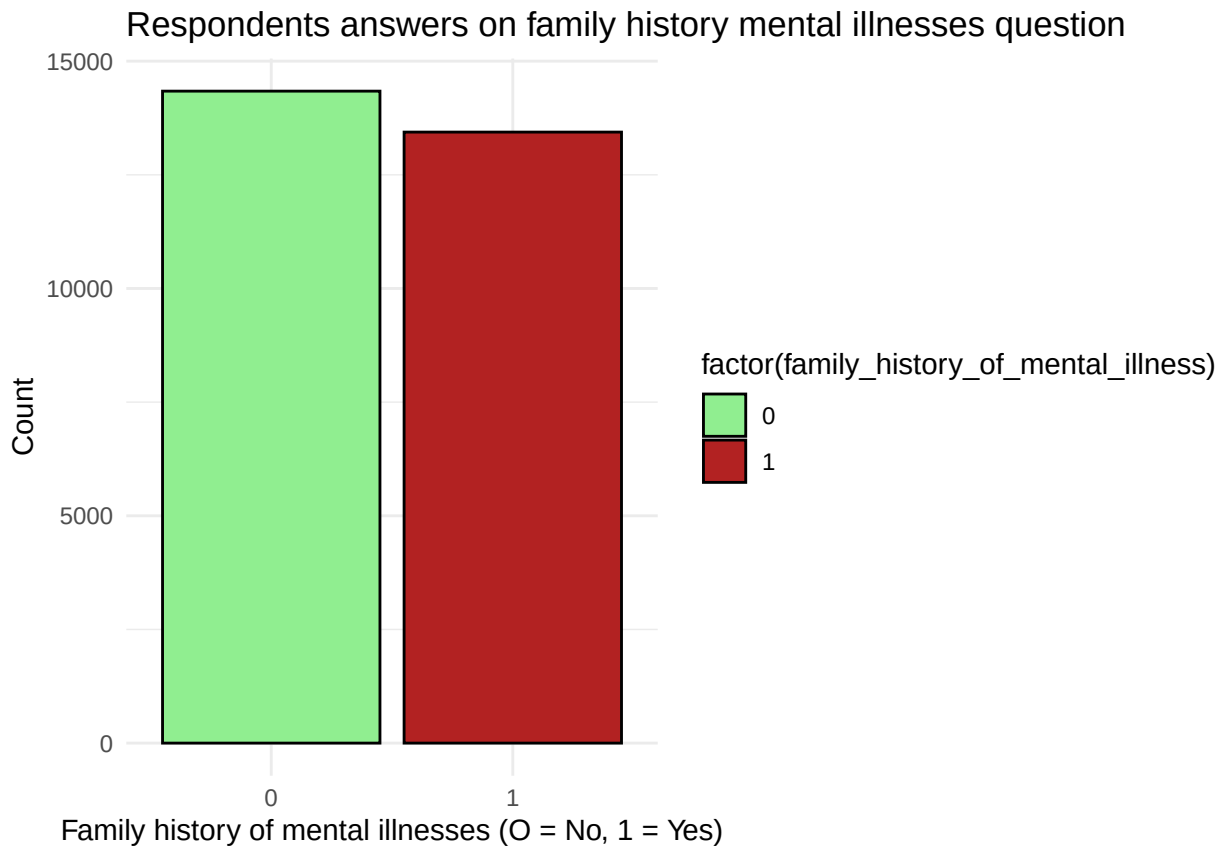
The bar chart shows the distribution of students who have ever had suicidal thoughts. Most students, about 20000, encountered suicidal thoughts. Half as many students, about 10000, have never had such thoughts.

Suicidal Thougths vs Depression



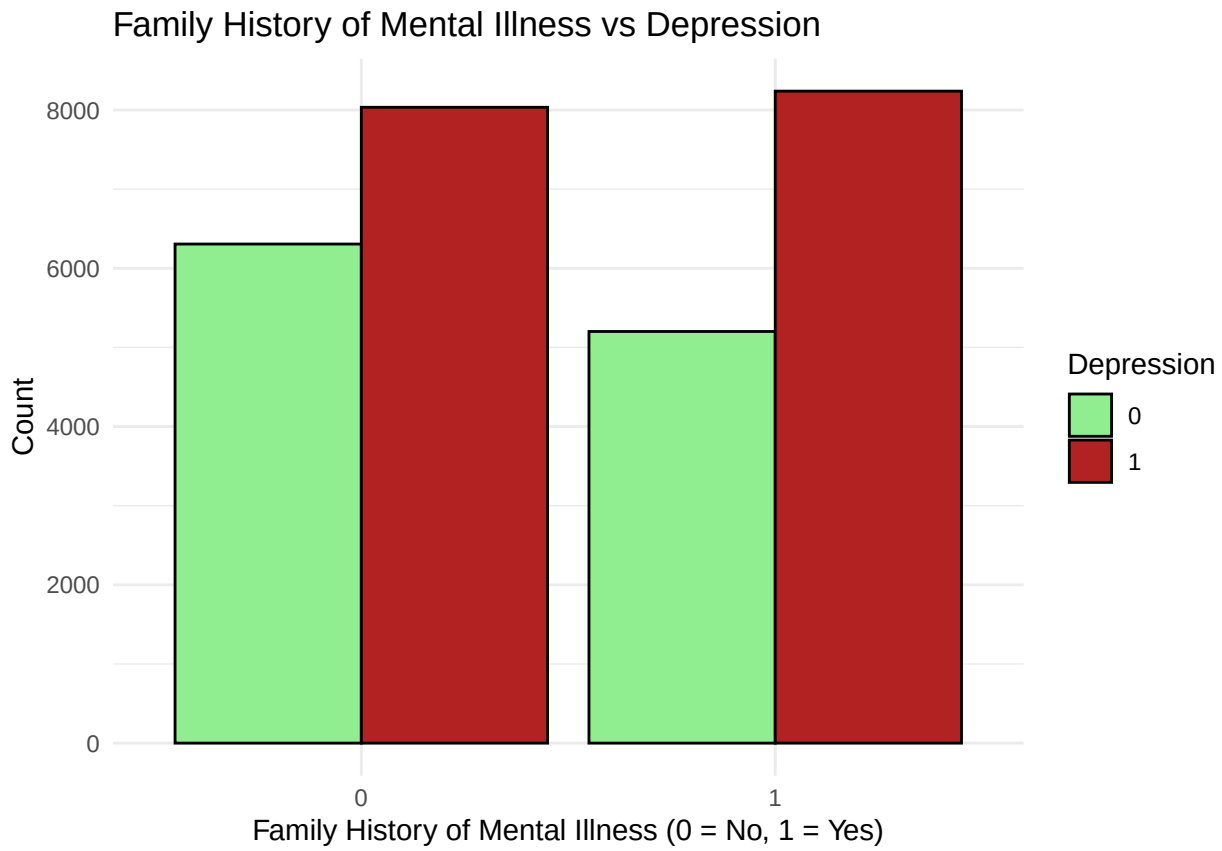
The bar chart demonstrates depression rates for two groups of students (that have encountered suicidal thoughts and not). Overall, it clearly seen on the graph, that the share of positive depression cases is extremely higher for students who have encountered such thoughts compared to those who have not.

Family history of mental illnessess



The distribution of students with and without family history of mental illnesses is almost equal, with slightly higher number of students for the 0 category.

Family history of mental illnessess vs Depression



The Family History of Mental Illness vs Depression plot shows, that the proportion of positive cases of depression is slightly higher for those students whose family members have experienced mental health problems.

Correaltion matrix and pairs diagram

Correlation matrix for continous and ordinal variables

Firstly, we should re-code some of our ordinal variables, in case to include them in our correlation matrix.

```
df_clean2 <- df_clean %>%
  mutate(
    diet_ordinal = case_when(
      dietary_habits == "Unhealthy" ~ 1,
      dietary_habits == "Moderate" ~ 2,
      dietary_habits == "Healthy" ~ 3
    ),
    sleep_ordinal = case_when(
      sleep_duration == "low" ~ 1,
      sleep_duration == "moderate" ~ 2,
      sleep_duration == "normal" ~ 3,
      sleep_duration == "above_normal" ~ 4
```

```
    )
  )
```

Now we can create a correlation matrix using Spearman method. The reason behind choosing these method is pretty simple. Firstly, we have ordinal varaibles and it is better to use mentioned method when we want to calculate correlation for such type of variables. Secondly, since none of our continuous variables are distributed normally, using Pearson method (for example separate correlation matrix for continuous variables) would be a violation of the normality assumption of this method. In turn, Spearman method is robust for non-normality.

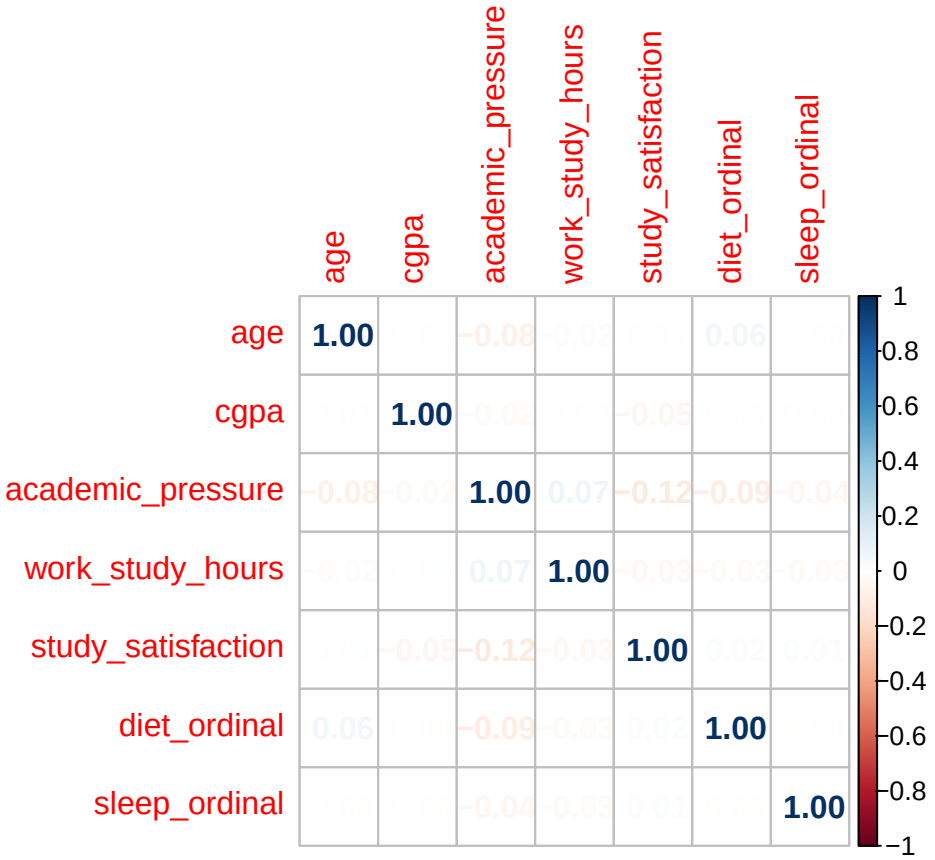
```
df_numeric <- df_clean2 %>%
  select(age, cgpa, academic_pressure, work_study_hours, study_satisfaction, diet_ordinal, sleep_ordinal

cor_matrix <- cor(df_numeric, method = "spearman")

cor_matrix
```

```
##                           age         cgpa academic_pressure work_study_hours
## age                1.000000000  0.007603389       -0.07791441     -0.018227977
## cgpa               0.007603389  1.000000000       -0.02457446      0.001660629
## academic_pressure -0.077914405 -0.024574464        1.00000000      0.067372694
## work_study_hours  -0.018227977  0.001660629        0.06737269      1.000000000
## study_satisfaction 0.007869731 -0.046975117       -0.11678257     -0.030865738
## diet_ordinal       0.057994842 -0.001821500       -0.09091517     -0.029688871
## sleep_ordinal     -0.003678560 -0.004809684       -0.04330413     -0.025239913
##                   study_satisfaction diet_ordinal sleep_ordinal
## age                      0.007869731   0.05799484  -0.003678560
## cgpa                    -0.046975117  -0.00182150  -0.004809684
## academic_pressure       -0.116782568  -0.09091517  -0.043304135
## work_study_hours        -0.030865738  -0.02968887  -0.025239913
## study_satisfaction       1.000000000   0.01951491   0.012386504
## diet_ordinal             0.019514911   1.00000000  -0.002069790
## sleep_ordinal            0.012386504  -0.00206979   1.000000000
```

```
corrplot(cor_matrix, method = "number")
```



From the correlation matrix above we can make the following conclusions. The Spearman correlation matrix demonstrate weak relationships between the variables. No strong correlations, positive or negative, can be observed. However, let's consider even minor correlations:

1) Age demonstrate a very weak positive correlation with dietary habits (0.06), indicating a minor tendency for older individuals to have slightly better diets. It is also negatively correlated (also weakly, -0.08) with academic pressure, suggesting that older students tend to experience less academic pressure.

2) Cgpa correlates with study satisfaction (-0.05), suggesting that higher academic performance is weakly associated with lower levels of study satisfaction. Also it correlates positively with academic pressure (0.02), suggesting that the higher a student's grade point average, the greater the academic load.

3) Academic pressure correlates weakly with study satisfaction (-0,12 - however, it is the highest value within matrix). Also it is positively associated with study hours (0,07), suggesting that students who spend more time studying tend to experience slightly higher levels of academic pressure. Additionally, academic pressure is negatively correlated with dietary habits (-0.09), indicating that higher academic pressure may be linked to poorer dietary habits. Finally, it is negatively correlated to sleep hours (-0,04), indicating that higher academic pressure associated with less our slept among students.

4) Study hours is negatively correlated with study satisfaction (-0.03), suggesting that more time spent on studying associated with lower rate of study satisfaction. Additionally, it is negatively associated with dietary habits, which shows the minimal effect of study time on dietary habits. Finally, study hours is negatively associated with sleep duration (-0.025), which may indicate a small negative effect of study time on sleep duration.

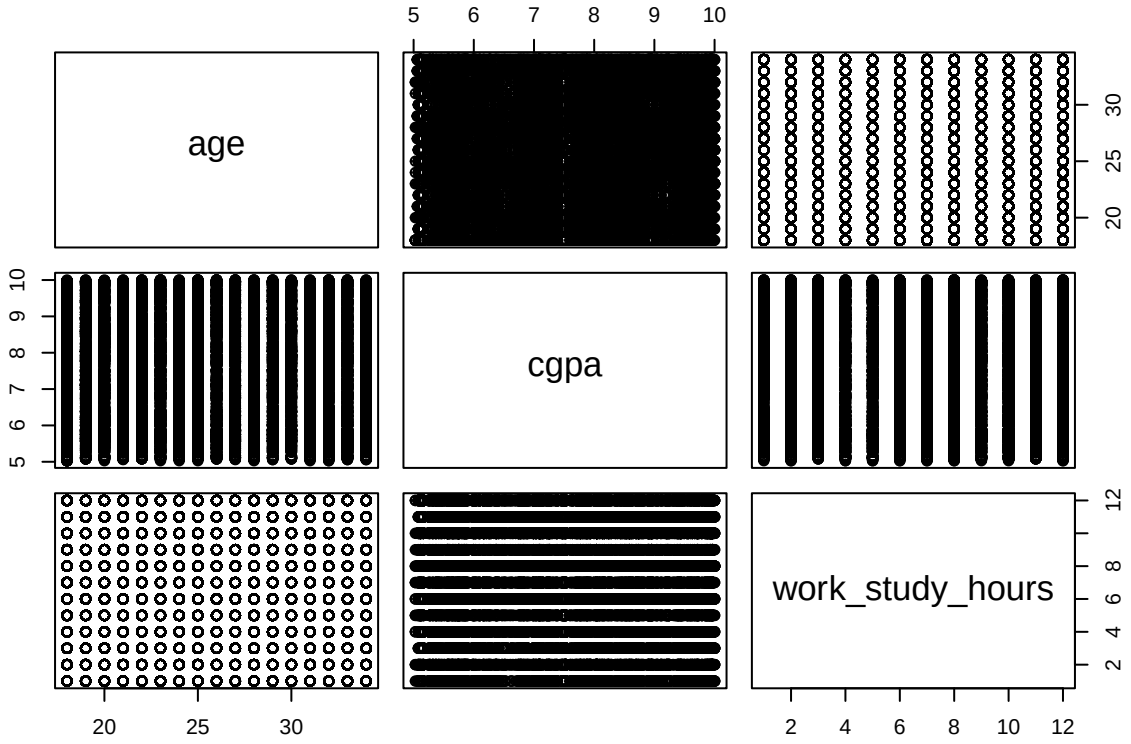
5) Study Satisfaction is negatively correlated with dietary habits (-0,2), indicating that those who are more satisfied with their studies tend to have slightly poorer dietary habits.

Nevertheless, the correlations between the variables are too weak to be considered statistically significant.

Pairs diagram for continuous variables

```
df_cont <- df_clean %>%
  select(age, cgpa, work_study_hours)

pairs(df_cont)
```



According to pairs diagram, there are no direct and linear relationships between age, CGPA and hours studied.

Suggestions for further analysis

This dataset was originally created to predict the target variable depression. Therefore, in a sense, it is logical that we did not find multicollinearity or direct linear relationships. In real life data, especially in social and medical research, predictors often have complex, non-linear and interdependent effects on the target variable (which is usually binary). A suitable method to deeply analyze this dataset would be machine learning models like logistic regression or random forest.