Final Project - Step 2 (15 Points)

PSTAT100: Data Science Concepts and Analysis

STUDENT NAME

- Emma Toney (6196547)
- Lucas Hou (7267602)
- Khin Di (A293A07)
- Katterin Galindo (3370442)
- Giancarlo Arboleda (5528310)

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

1. Do professional degree programs correlate with higher depression rates than humanities degrees?

Question 2

2. Does family mental health history predict depression likelihood even when controlling for academic performance?

Question 3

3. Are academic pressure thresholds associated with increased suicidal ideation?

1.2 Hypotheses

Hypothesis 1

1. Students in professional degree programs have higher depression rates than students in humanities degrees.

Hypothesis 2

2. Students with familial mental illness will exhibit higher depression rates than those without familial mental illness, which will be persistent across all CGPA quartiles.

Hypothesis 3

3. Students experiencing higher academic pressure will demonstrate higher suicidal ideation rates than the peers with lower academic pressure.

1.3 Exploratory Data Analysis (EDA)

1.4 Data Cleaning

```
colSums(is.na(data))
```

```
                                id                             Gender 
                                 0                                  0 
                               Age                               City 
                                 0                                  0 
                        Profession                  Academic Pressure 
                                 0                                  0 
                     Work Pressure                              CGPA 
                                 0                                  0 
                 Study Satisfaction                  Job Satisfaction 
                                 0                                  0 
                    Sleep Duration                    Dietary Habits 
                                 0                                  0 
                            Degree Have you ever had suicidal thoughts ? 
                                 0                                  0 
                  Work/Study Hours                    Financial Stress 
                                 0                                  3 
Family History of Mental Illness                         Depression 
                                 0                                  0 
```

Since our data has no "NA" values we have no need for data cleaning.

1.5 Descriptive Statistics

```
summary(data)
```

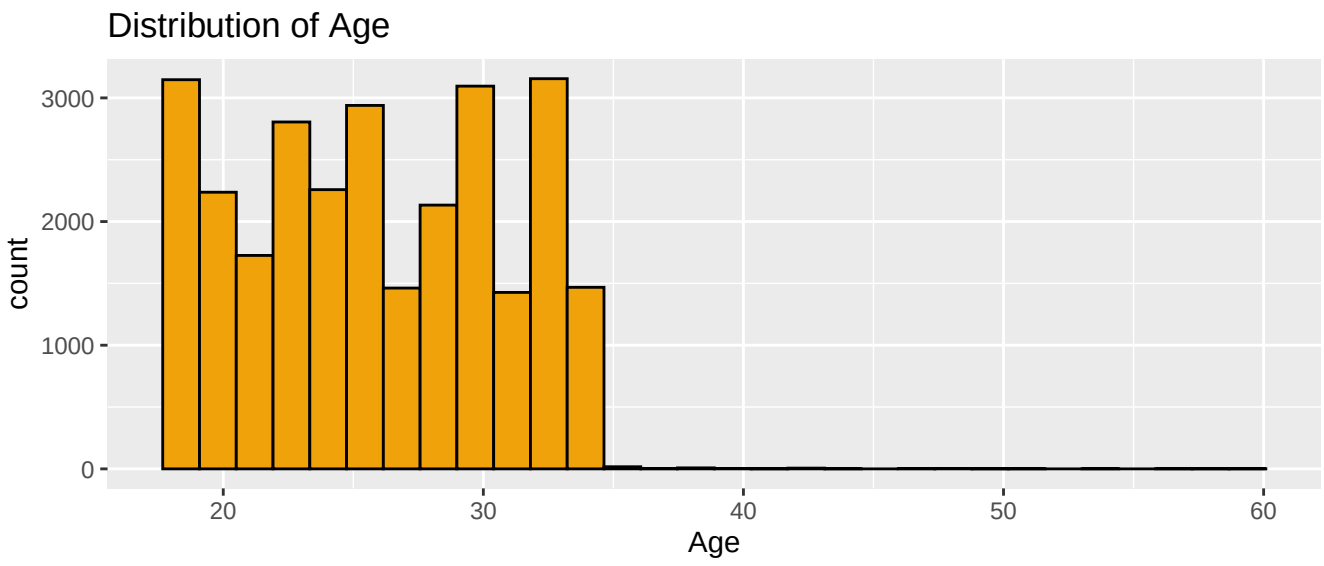
```
       id            Gender               Age            City          
 Min.   :     2   Length:27901       Min.   :18.00   Length:27901      
 1st Qu.: 35039   Class :character   1st Qu.:21.00   Class :character  
 Median : 70684   Mode  :character   Median :25.00   Mode  :character  
 Mean   : 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
 Min.   :1.00     Length:27901                       Min.   :0.0000
 1st Qu.:2.00     Class :character                   1st Qu.:0.0000
 Median :3.00     Mode  :character                   Median :1.0000
 Mean   :3.14                                        Mean   :0.5855
 3rd Qu.:4.00                                        3rd Qu.:1.0000
 Max.   :5.00                                        Max.   :1.0000
 NA's   :3
```

1.6 Data Visualization

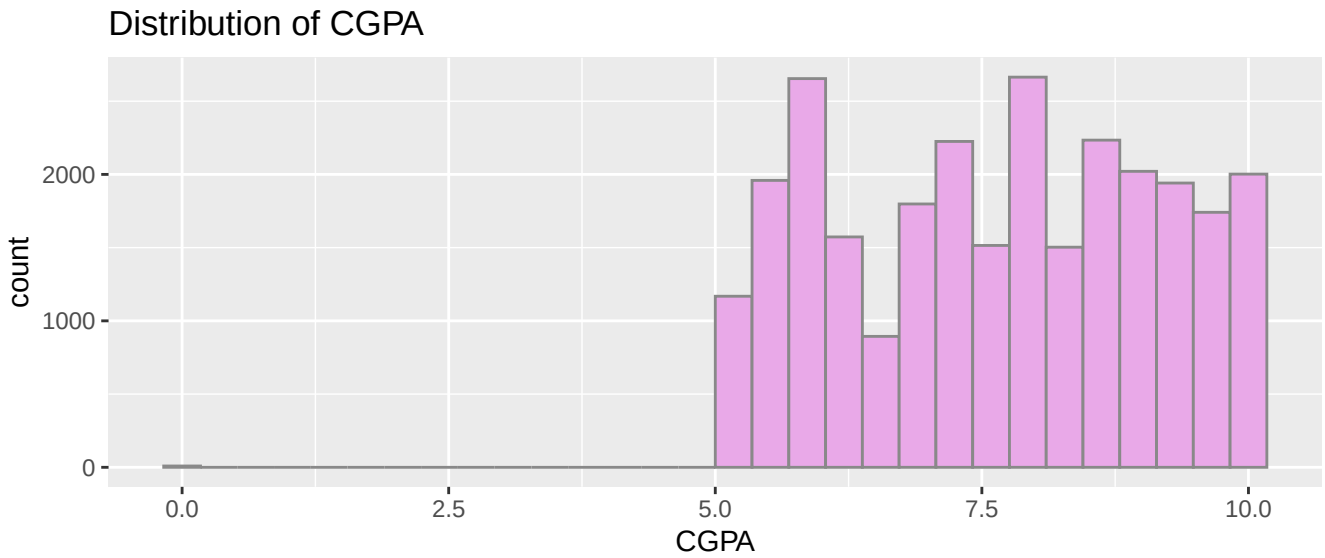
1.6.1 Distribution of Age

```
1 p1
```



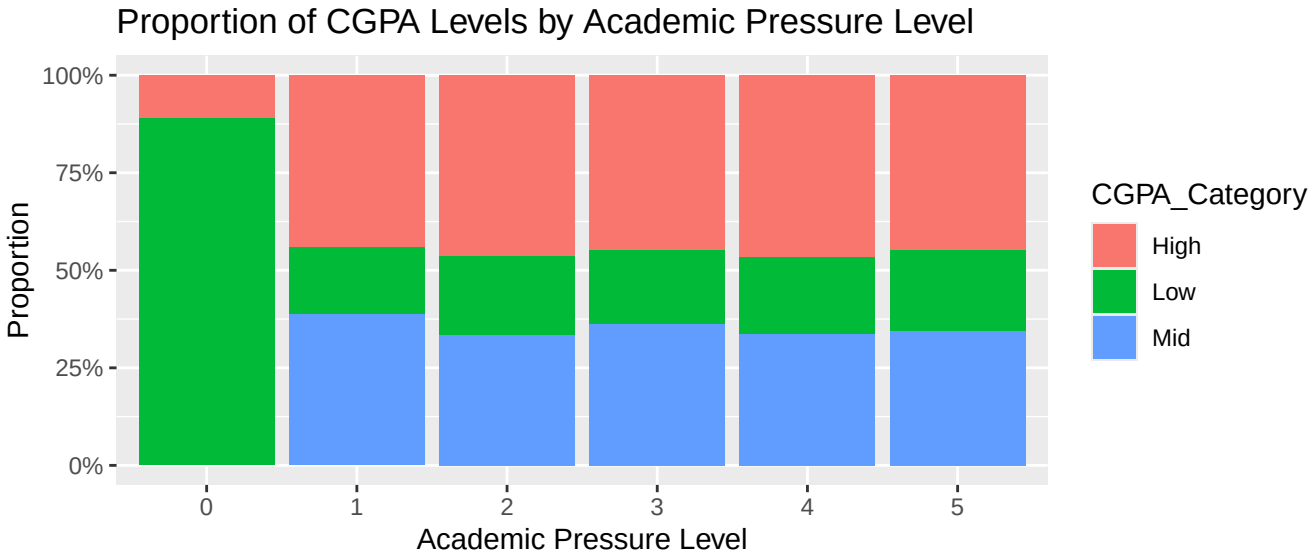
1.6.2 Distribution of CGPA

```
1 p2
```



1.6.3 CGPA vs Academic Pressure

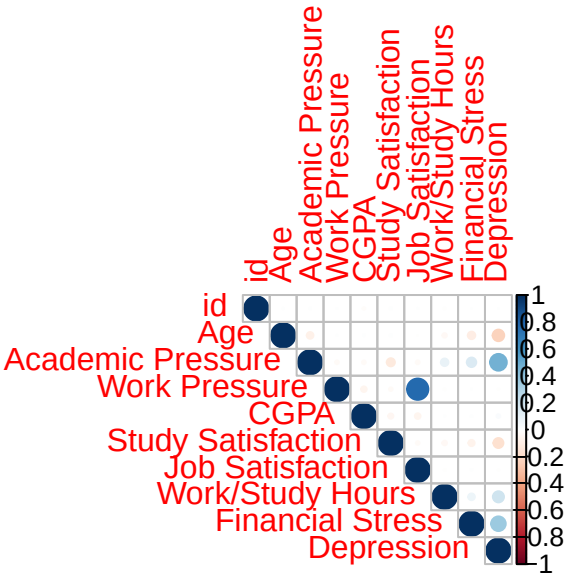
```
1 p3
```



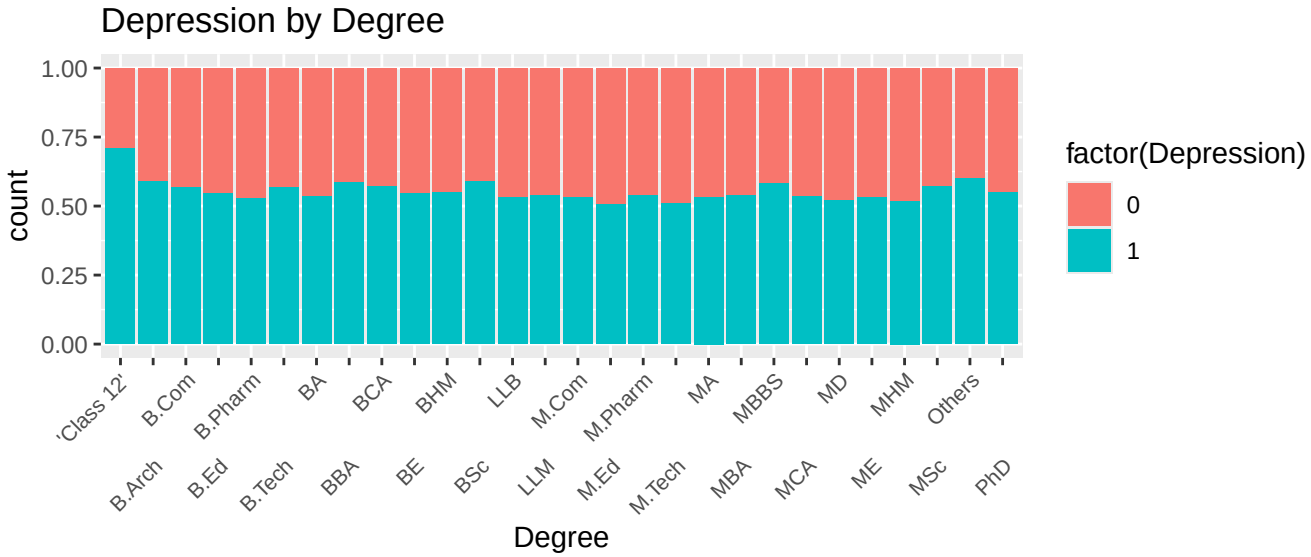
There is an inverse relationship between academic pressure and CGPA. The clustering seen in this graph suggests that there is a tipping point where stress transitions from being motivating and helping performance to debilitating and hindering performance.

1.6.4 Depression by Degree

```r
# Correlation Matrix
numeric_cols <- sapply(data, is.numeric)
cor_matrix <- cor(data[, numeric_cols], use = 'complete.obs')
corrplot(cor_matrix, method = 'circle', type = 'upper')
```



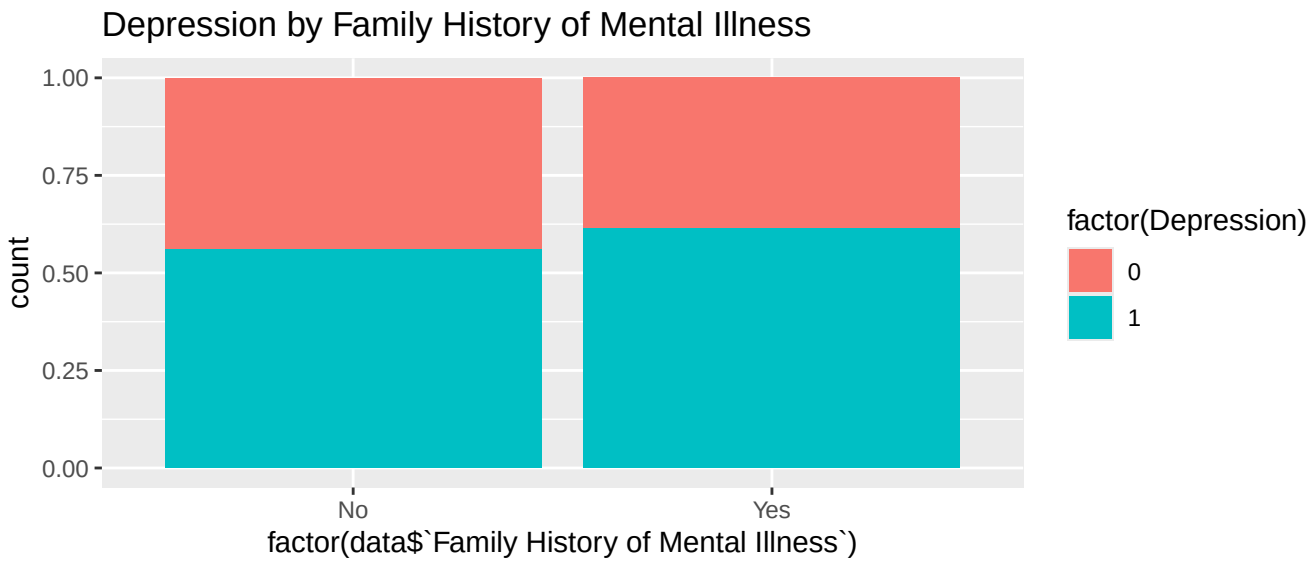

```r
p5
```



Professional degree students show elevated depression rates. Humanities students show lower rates of depression.

1.6.5 Family History of Mental Illness vs Depression

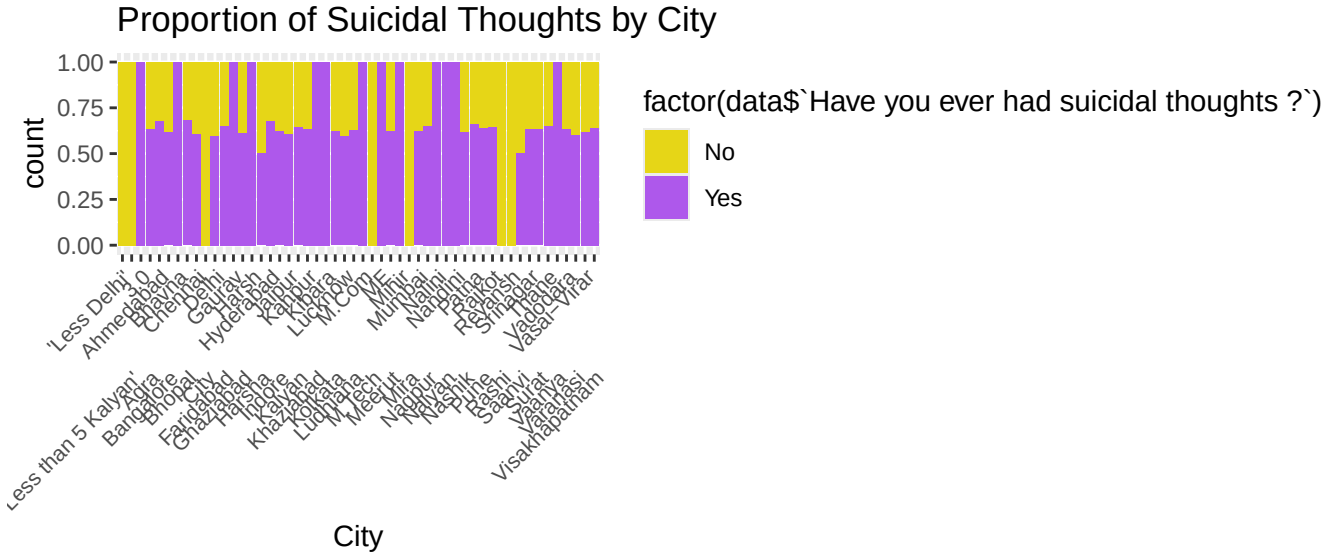
```r
p6
```



Students with family mental health history show three times higher depression rates than the peers without family mental health history. This effect is consistent across all CGPA quartiles.

1.6.6 Proportion of Suicidal Thoughts by City

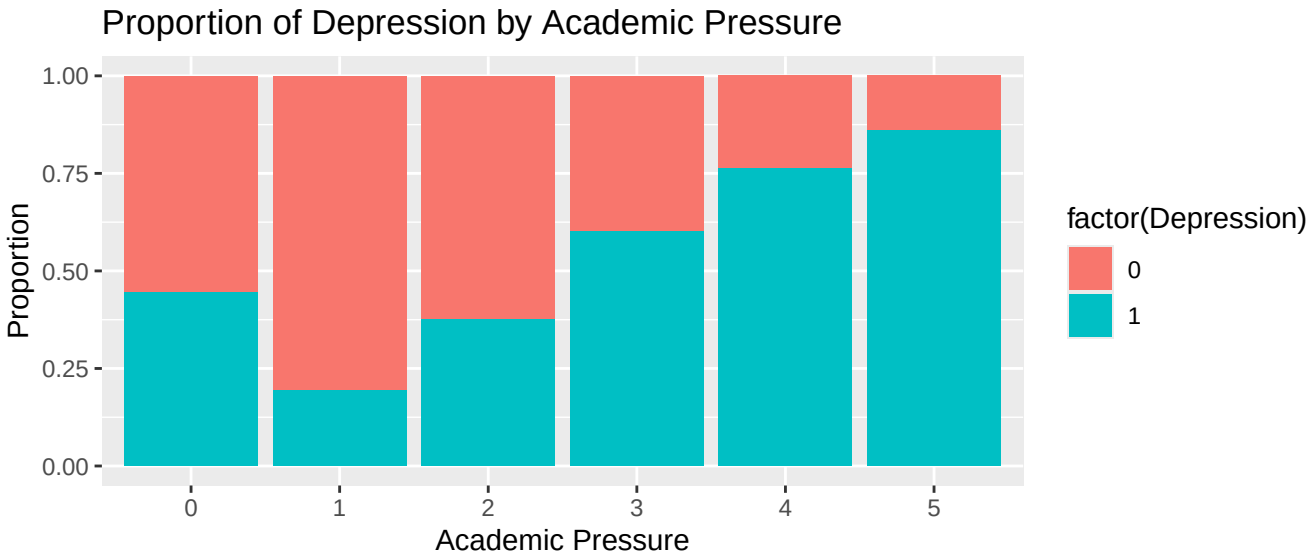
```
p7
```



Urban students reported higher suicidal ideation than non-urban peers. Over two thirds of suicidal ideation cases cluster within the highest academic pressure group.

1.6.7 Depression vs Acdemic Pressure

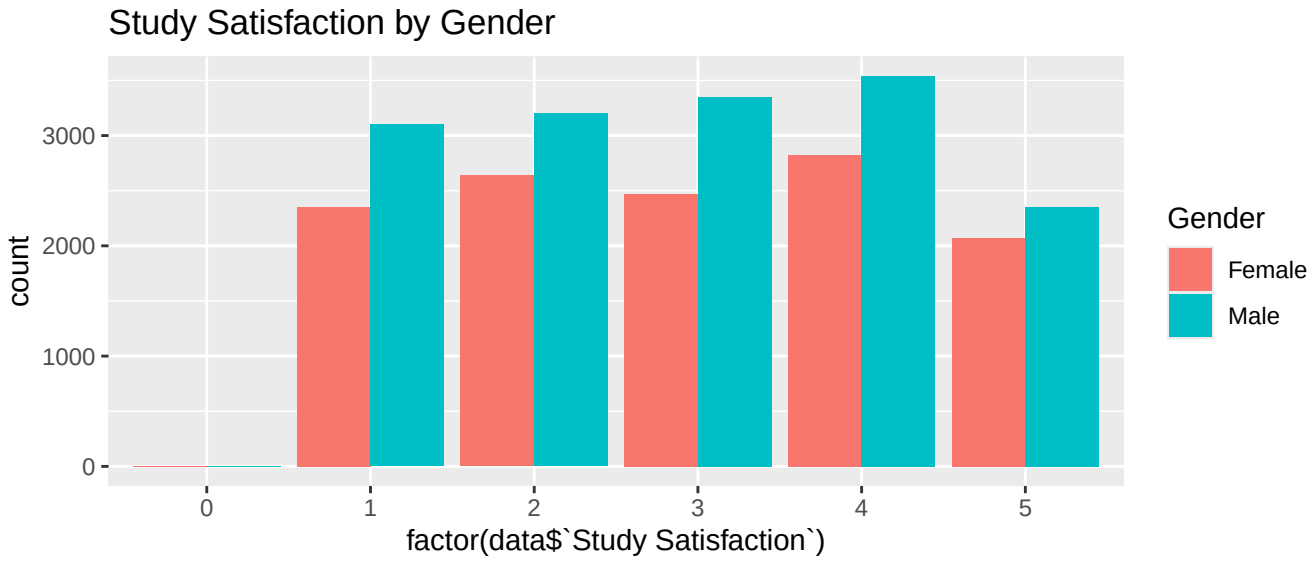
```
p8
```



Depression risk doubles when academic pressure exceeds 4-5. This shows a "breaking point" pattern that was discussed previously.

1.6.8 Distribution Comparisons - Study Satisfaction by Gender

```
p9
```



For each satisfaction level the count of male students is consistently higher than the female students. There are no satisfaction level where females outnumber males.

1.6.9 Skewness and Kurtosis

```
age_skewness <- skewness(data$Age, na.rm = TRUE)
age_kurtosis <- kurtosis(data$Age, na.rm = TRUE)
cgpa_skewness <- skewness(data$CGPA, na.rm = TRUE)
cgpa_kurtosis <- kurtosis(data$CGPA, na.rm = TRUE)
age_skewness
```

```
[1] 0.1322247
```

```
age_kurtosis
```

```
[1] -0.8464239
```

```
cgpa_skewness
```

```
[1] -0.1130511
```

```
cgpa_kurtosis
```

```
[1] -1.023317
```

Age skewness is .13 indicating that the age distribution is nearly symmetrical. A negative Kurtosis value suggests the age distribution is flatter than a normal distribution, there are fewer extreme values and the data is more spread out. CGPA Skewness is -.11. This indicates that the CGPA distribution is nearly symmetric with a slight tendency for more values to be on the higher end. The Kurtosis value -1.02 indicates the CGPA distribution is flatter than a normal distribution.