Statistics for Data Science

Westyn Hilliard

Step 1 -

Introduction -

People of all sexes, ages, genders, races, and socioeconomic groups can suffer from Mental Health Disorders. Many factors can trigger illness, ranging from mild to severe. In general, people with mental disorders find it difficult to cope with everyday life because of their altered thinking, moods, or behaviors" (Sampson, 2024). Mental health has recently become something that more and more people are opening up about. With it on the rise, we must research and understand the collected data. The disorders offer an increasing concern within public health, with millions of people becoming more and more affected by conditions such as depression, anxiety, and bipolar disorder. While the prevalence of the disorder/disorders remains, timely diagnosis remains to be a problem. I believe that with the increase in mental health disorders, those with disorders or who have a family history of disorders would be interested in the analysis. Using data science techniques, we can analyze many data sources that lead to the early detection and control of mental health disorders.

Research Questions -

1. Can we use machine learning algorithms to accurately determine the likelihood of a person developing mental health disorders based on demographic and behavioral data?
2. What are the critical risk factors associated with the onset of specific mental health disorders?
3. How do environmental factors, such as socioeconomic status and neighborhood characteristics, impact mental health outcomes?
4. Are there identifiable patterns in social media usage or online behavior that correlate with mental health disorders?
5. How effective are mental health interventions, and can data analysis help optimize treatment strategies?
6. What role does access to mental health care services play in mitigating the impact of mental health disorders, and how can disparities in access be addressed through policy and community-based initiatives?

Approach -

The approach involves gathering data on demographics, behavior, environmental factors, and healthcare utilization. Exploratory data analysis can be used to identify patterns and correlations within the data. Machine learning models can be trained to predict mental health outcomes and identify significant patterns. Additionally, statistical analysis will be performed to assess the effectiveness of existing interventions and suggest potential improvements.

Addressing The Problem -

By analyzing large-scale datasets, this approach aims to uncover insights into the complex interactions between various factors and mental health outcomes. By identifying high-risk individuals and understanding the factors contributing to mental health disorders, targeted interventions can be developed to improve prevention, diagnosis, and treatment.

Data -

1. *Mental Illnesses Prevalence:*

Source: This dataset appears to be derived from global health data repositories, focusing on the prevalence of various mental illnesses across different countries and years.

Purpose: The original purpose of the data is to provide an understanding of the prevalence of different mental health disorders globally, stratified by year, and covering both genders in an age-standardized manner.

Collection Period: The dataset spans multiple years, capturing the prevalence data for various mental health disorders.

Variables:

- `Entity: The country or region.`
- `Code: Country code.`
- `Year: Year of data collection.`
- `Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized`
- `Depressive disorders (share of population) - Sex: Both - Age: Age-standardized`
- `Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized`
- `Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized`
- `Eating disorders (share of population) - Sex: Both - Age: Age-standardized`

https://www.kaggle.com/datasets/thedevastator/uncover-global-trends-in-mental-health-disorder

2. *Mental Health Depression Disorder Data:*

Source: This dataset seems to be an extensive collection of mental health disorder prevalence data, with a specific emphasis on depression among other disorders.

Purpose: To provide detailed insights into the prevalence of various mental health disorders, including schizophrenia, bipolar disorder, eating disorders, anxiety disorders, drug use disorders, depression, and alcohol use disorders.

Collection Period: Covers a broad range of years.

Variables:

- `index: Row index.`
- `Entity: Country or region.`
- `Code: Country code.`
- `Year: Year of data collection.`
- `Schizophrenia (%)`
- `Bipolar disorder (%)`
- `Eating disorders (%)`
- `Anxiety disorders (%)`
- `Drug use disorders (%)`
- `Depression (%)`
- `Alcohol use disorders (%)`

https://www.kaggle.com/datasets/shashwatwork/depression-and-mental-health-data-analysis

3. *Mental Health Final Data:*

Source: This dataset appears to be survey-based, capturing individual responses related to mental health during a specific period, likely a pandemic given the context of questions.

Purpose: To understand the impact of indoor confinement, stress, and other factors on mental health.

Collection Period: The specific period is not mentioned but seems to align with pandemic lockdown periods.

Variables:

- `Age`
- `Gender`
- `Occupation`
- `Days_Indoors`
- `Growing_Stress`
- `Quarantine_Frustrations`
- `Changes_Habits`
- `Mental_Health_History`
- `Weight_Change`
- `Mood_Swings`
- `Coping_Struggles`
- `Work_Interest`
- `Social_Weakness`

https://www.kaggle.com/code/zhukovoleksiy/mental-disorders-cleaning-dashboard

Required Packages -

- 'caret' – for machine learning algorithms
- 'ggploy2' – for data visualization
- 'dplyr' – for data manipulation
- 'tiydr' for data tidying

Plots and Tables -

1. *Bar Charts:* Bar charts can visualize categorical variables, such as demographic characteristics (e.g., age groups, gender) or types of mental health disorders. These charts can show the distribution of different categories within the dataset and highlight any disparities or trends.
2. *Heatmaps:* Heatmaps can display correlations between variables, allowing visualization of relationships between demographic factors, behavioral patterns, and mental health outcomes. This can help identify significant predictors or risk factors for mental health disorders.
3. *Scatter Plots:* Scatter plots can show the relationship between two continuous variables, such as socioeconomic status and mental health scores. These plots can reveal potential associations or patterns in the data, such as income levels impacting mental well-being.
4. *Time Series Plots:* Time series plots can be used to visualize trends or changes in mental health outcomes over time. For example, plotting the prevalence of depression symptoms over several years can reveal any temporal patterns or fluctuations.
5. *Model Performance Metrics:* Tables showing model performance metrics, such as accuracy, precision, recall, and F1-score, assess the predictive power of machine learning models. These metrics indicate how well the model can classify individuals with or without mental health disorders.
6. *Descriptive Statistics:* Tables summarizing descriptive statistics (e.g., mean, median, standard deviation) for key variables can provide an overview of the dataset's characteristics. This includes demographic factors, behavioral patterns, and mental health assessments.

Questions for Future Steps:

1. How can we address privacy concerns when analyzing sensitive data, such as mental health records and social media activity?
2. How can we incorporate qualitative data, such as patient narratives and experiences, to complement quantitative analysis and provide a more comprehensive understanding of mental health?

3. What opportunities exist for collaboration between data scientists, mental health professionals, and policymakers to translate research findings into actionable strategies for improving mental health outcomes at individual and population levels?

Sources -

Devastator, T. (2022, December 14). Global trends in mental health disorder. Kaggle. https://www.kaggle.com/datasets/thedevastator/uncover-global-trends-in-mental-health-disorder

Sampson, C. (2024, March 27). 10 types of mental health disorders. Life Adjustment Team. https://www.lifeadjustmentteam.com/10-types-of-mental-health-disorders/

Tiwari, S. (2024, January 14). Depression and mental health data analysis. Kaggle. https://www.kaggle.com/datasets/shashwatwork/depression-and-mental-health-data-analysis

Zhukovoleksiy. (2022, December 29). Mental disorders: Cleaning + dashboard. Kaggle. https://www.kaggle.com/code/zhukovoleksiy/mental-disorders-cleaning-dashboard

******10.3 Final Project Step 2**** -**

How did you import and clean your data?:

Loading and cleaning the data:

```
# Load necessary libraries
library(readr)
library(dplyr)
library(ggplot2)
library(reshape2)

# Load the mental illnesses prevalence data
mental_illnesses_prevalence <- read_csv("~/Desktop/S C H O O L/Masters/DSC520 - Stats for Data Science/
    col_types = cols(Entity = col_character(), Code = col_character(),
        Year = col_double(), `Schizophrenia disorders (share of population) - Sex: Both - Age: Age-stand
        `Depressive disorders (share of population) - Sex: Both - Age: Age-standardized` = col_double()
        `Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized` = col_double(),
        `Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized` = col_double(),
        `Eating disorders (share of population) - Sex: Both - Age: Age-standardized` = col_double()))

# Check for parsing issues
problems_mental_illnesses_prevalence <- problems(mental_illnesses_prevalence)
print(problems_mental_illnesses_prevalence)
```

```
# A tibble: 0 x 5
# i 5 variables: row <int>, col <int>, expected <chr>,
#   actual <chr>, file <chr>
```

```
# Load the depression disorder data
depression_data <- read_csv("~/Desktop/S C H O O L/Masters/DSC520 - Stats for Data Science/Mental health
    col_types = cols(index = col_double(), Entity = col_character(),
        Code = col_character(), Year = col_character(),
        `Schizophrenia (%)` = col_double(), `Bipolar disorder (%)` = col_double(),
        `Eating disorders (%)` = col_double(), `Anxiety disorders (%)` = col_double(),
        `Drug use disorders (%)` = col_double(), `Depression (%)` = col_double(),
        `Alcohol use disorders (%)` = col_double()))
```

```r
# Check for parsing issues
problems_depression_data <- problems(depression_data)
print(problems_depression_data)
```

```
# A tibble: 7 x 5
     row   col expected actual                    file
   <int> <int> <chr>    <chr>                     <chr>
1   6470     5 a double Prevalence in males (%)   /Use~
2   6470     6 a double Prevalence in females (%) /Use~
3   6470     7 a double Population                /Use~
4  54278     5 a double Suicide rate (deaths per~ /Use~
5  54278     6 a double Depressive disorder rate~ /Use~
6  54278     7 a double Population                /Use~
7 102086     5 a double Prevalence - Depressive ~ /Use~
```

```r
# Load the final mental health data
mental_health_finaldata <- read_csv("~/Desktop/S C H O O L/Masters/DSC520 - Stats for Data Science/menta
    col_types = cols(Age = col_character(), Gender = col_character(),
        Occupation = col_character(), Days_Indoors = col_double(),
        Growing_Stress = col_character(), Quarantine_Frustrations = col_character(),
        Changes_Habits = col_character(), Mental_Health_History = col_character(),
        Weight_Change = col_character(), Mood_Swings = col_character(),
        Coping_Struggles = col_character(), Work_Interest = col_character(),
        Social_Weakness = col_character()))

# Check for parsing issues
problems_mental_health_finaldata <- problems(mental_health_finaldata)
print(problems_mental_health_finaldata)
```

```
# A tibble: 824 x 5
     row   col expected actual             file
   <int> <int> <chr>    <chr>              <chr>
 1     2     4 a double 1-14 days          /Users/west~
 2     3     4 a double 31-60 days         /Users/west~
 3     4     4 a double Go out Every day   /Users/west~
 4     5     4 a double 1-14 days          /Users/west~
 5     6     4 a double More than 2 months /Users/west~
 6     7     4 a double More than 2 months /Users/west~
 7     8     4 a double Go out Every day   /Users/west~
 8     9     4 a double 1-14 days          /Users/west~
 9    10     4 a double Go out Every day   /Users/west~
10    11     4 a double Go out Every day   /Users/west~
# i 814 more rows
```

```r
# Inspect unique values in the 'Year' column of
# depression_data
unique(depression_data$Year)
```

```
  [1] "1990"      "1991"      "1992"      "1993"
  [5] "1994"      "1995"      "1996"      "1997"
  [9] "1998"      "1999"      "2000"      "2001"
 [13] "2002"      "2003"      "2004"      "2005"
 [17] "2006"      "2007"      "2008"      "2009"
 [21] "2010"      "2011"      "2012"      "2013"
 [25] "2014"      "2015"      "2016"      "2017"
```

```
 [29] "Year"      "1800"      "1801"      "1802"
 [33] "1803"      "1804"      "1805"      "1806"
 [37] "1807"      "1808"      "1809"      "1810"
 [41] "1811"      "1812"      "1813"      "1814"
 [45] "1815"      "1816"      "1817"      "1818"
 [49] "1819"      "1820"      "1821"      "1822"
 [53] "1823"      "1824"      "1825"      "1826"
 [57] "1827"      "1828"      "1829"      "1830"
 [61] "1831"      "1832"      "1833"      "1834"
 [65] "1835"      "1836"      "1837"      "1838"
 [69] "1839"      "1840"      "1841"      "1842"
 [73] "1843"      "1844"      "1845"      "1846"
 [77] "1847"      "1848"      "1849"      "1850"
 [81] "1851"      "1852"      "1853"      "1854"
 [85] "1855"      "1856"      "1857"      "1858"
 [89] "1859"      "1860"      "1861"      "1862"
 [93] "1863"      "1864"      "1865"      "1866"
 [97] "1867"      "1868"      "1869"      "1870"
[101] "1871"      "1872"      "1873"      "1874"
[105] "1875"      "1876"      "1877"      "1878"
[109] "1879"      "1880"      "1881"      "1882"
[113] "1883"      "1884"      "1885"      "1886"
[117] "1887"      "1888"      "1889"      "1890"
[121] "1891"      "1892"      "1893"      "1894"
[125] "1895"      "1896"      "1897"      "1898"
[129] "1899"      "1900"      "1901"      "1902"
[133] "1903"      "1904"      "1905"      "1906"
[137] "1907"      "1908"      "1909"      "1910"
[141] "1911"      "1912"      "1913"      "1914"
[145] "1915"      "1916"      "1917"      "1918"
[149] "1919"      "1920"      "1921"      "1922"
[153] "1923"      "1924"      "1925"      "1926"
[157] "1927"      "1928"      "1929"      "1930"
[161] "1931"      "1932"      "1933"      "1934"
[165] "1935"      "1936"      "1937"      "1938"
[169] "1939"      "1940"      "1941"      "1942"
[173] "1943"      "1944"      "1945"      "1946"
[177] "1947"      "1948"      "1949"      "1950"
[181] "1951"      "1952"      "1953"      "1954"
[185] "1955"      "1956"      "1957"      "1958"
[189] "1959"      "1960"      "1961"      "1962"
[193] "1963"      "1964"      "1965"      "1966"
[197] "1967"      "1968"      "1969"      "1970"
[201] "1971"      "1972"      "1973"      "1974"
[205] "1975"      "1976"      "1977"      "1978"
[209] "1979"      "1980"      "1981"      "1982"
[213] "1983"      "1984"      "1985"      "1986"
[217] "1987"      "1988"      "1989"      "2018"
[221] "2019"      "10000 BCE" "9000 BCE"  "8000 BCE"
[225] "7000 BCE"  "6000 BCE"  "5000 BCE"  "4000 BCE"
[229] "3000 BCE"  "2000 BCE"  "1000 BCE"  "0"
[233] "100"       "200"       "300"       "400"
[237] "500"       "600"       "700"       "800"
[241] "900"       "1000"      "1100"      "1200"
```

```
[245] "1300"      "1400"      "1500"      "1600"
[249] "1700"      "1710"      "1720"      "1730"
[253] "1740"      "1750"      "1760"      "1770"
[257] "1780"      "1790"      "1"
```

```
# Clean the 'Year' column
depression_data <- depression_data %>%
    mutate(Year = as.numeric(gsub("[^0-9]", "", Year)))  # Remove non-numeric characters

# Inspect and clean mental_health_finaldata
mental_health_finaldata <- mental_health_finaldata %>%
    mutate(across(where(is.character), ~ifelse(. == "",
        NA, .)))

# Check for parsing issues again after cleaning
problems_depression_data <- problems(depression_data)
print(problems_depression_data)
```

```
# A tibble: 0 x 4
# i 4 variables: row <int>, col <int>, expected <chr>,
#   actual <chr>
```

```
problems_mental_health_finaldata <- problems(mental_health_finaldata)
print(problems_mental_health_finaldata)
```

```
# A tibble: 0 x 4
# i 4 variables: row <int>, col <int>, expected <chr>,
#   actual <chr>
```

To import and clean the data, I performed the following steps:

1. Loading the Data:
- I used the read_csv function from the readr package to load the datasets into R.
- I specified the column types to ensure the data is correctly interpreted.
2. Handling Missing Values:
- I inspected each dataset for missing values using the problems function.
- I cleaned the data by replacing missing values in numeric columns with the median value using the mutate and across functions from the dplyr package.
- For character columns, I replaced empty strings with NA.
3. Ensuring Consistency:
- I checked for non-numeric values in columns that should contain numeric data and removed non-numeric characters.
- I verified that there were no remaining missing or non-finite values after cleaning.
4. Summarizing the Data:
- I used the summary function to get an overview of the datasets and ensure the data was correctly loaded and cleaned.

What does the final data set look like??:

```
# Summarize data
summary(mental_illnesses_prevalence)
```

```
    Entity              Code                Year
 Length:6420        Length:6420        Min.   :1990
 Class :character   Class :character   1st Qu.:1997
 Mode  :character   Mode  :character   Median :2004
                                       Mean   :2004
```

```
                                       3rd Qu.:2012
                                       Max.   :2019
 Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized
 Min.   :0.1884
 1st Qu.:0.2423
 Median :0.2735
 Mean   :0.2666
 3rd Qu.:0.2866
 Max.   :0.4620
 Depressive disorders (share of population) - Sex: Both - Age: Age-standardized
 Min.   :1.522
 1st Qu.:3.080
 Median :3.637
 Mean   :3.767
 3rd Qu.:4.366
 Max.   :7.646
 Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized
 Min.   :1.880
 1st Qu.:3.426
 Median :3.940
 Mean   :4.102
 3rd Qu.:4.564
 Max.   :8.625
 Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized
 Min.   :0.1817
 1st Qu.:0.5209
 Median :0.5793
 Mean   :0.6370
 3rd Qu.:0.8444
 Max.   :1.5067
 Eating disorders (share of population) - Sex: Both - Age: Age-standardized
 Min.   :0.04478
 1st Qu.:0.09642
 Median :0.14415
 Mean   :0.19566
 3rd Qu.:0.25117
 Max.   :1.03169
```

```
summary(depression_data)
```

```
     index           Entity
 Min.   :     0   Length:108553
 1st Qu.: 27138   Class :character
 Median : 54276   Mode  :character
 Mean   : 54276
 3rd Qu.: 81414
 Max.   :108552

    Code                Year
 Length:108553      Min.   :    0
 Class :character   1st Qu.: 1867
 Mode  :character   Median : 1936
                    Mean   : 1929
                    3rd Qu.: 1992
                    Max.   :10000
```

```
                  NA's   :3
 Schizophrenia (%)   Bipolar disorder (%)
 Min.   :        0   Min.   :   0.31
 1st Qu.:        1   1st Qu.:   0.84
 Median :        4   Median :   4.20
 Mean   :  1101844   Mean   :1118.40
 3rd Qu.:      307   3rd Qu.:2885.21
 Max.   :264455593   Max.   :6096.44
 NA's   :82681       NA's   :89149
 Eating disorders (%) Anxiety disorders (%)
 Min.   :0.000e+00    Min.   :2.02
 1st Qu.:1.080e+05    1st Qu.:3.19
 Median :1.253e+06    Median :3.55
 Mean   :2.790e+07    Mean   :3.99
 3rd Qu.:5.265e+06    3rd Qu.:4.68
 Max.   :7.713e+09    Max.   :8.97
 NA's   :8319         NA's   :102085
 Drug use disorders (%) Depression (%)
 Min.   :0.38           Min.   :2.14
 1st Qu.:0.54           1st Qu.:3.01
 Median :0.73           Median :3.50
 Mean   :0.86           Mean   :3.50
 3rd Qu.:0.94           3rd Qu.:3.91
 Max.   :3.45           Max.   :6.60
 NA's   :102085         NA's   :102085
 Alcohol use disorders (%)
 Min.   :0.45
 1st Qu.:0.99
 Median :1.48
 Mean   :1.59
 3rd Qu.:1.87
 Max.   :5.47
 NA's   :102085
```

```
summary(mental_health_finaldata)
```

```
     Age                Gender
 Length:824         Length:824
 Class :character   Class :character
 Mode  :character   Mode  :character




  Occupation         Days_Indoors Growing_Stress
 Length:824         Min.   : NA   Length:824
 Class :character   1st Qu.: NA   Class :character
 Mode  :character   Median : NA   Mode  :character
                    Mean   :NaN
                    3rd Qu.: NA
                    Max.   : NA
                    NA's   :824
 Quarantine_Frustrations Changes_Habits
 Length:824              Length:824
 Class :character        Class :character
```

```
 Mode  :character        Mode  :character


 Mental_Health_History Weight_Change
 Length:824            Length:824
 Class :character      Class :character
 Mode  :character      Mode  :character


 Mood_Swings        Coping_Struggles
 Length:824         Length:824
 Class :character   Class :character
 Mode  :character   Mode  :character


 Work_Interest      Social_Weakness
 Length:824         Length:824
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
# Display the first few rows of each dataset
head(mental_illnesses_prevalence)
```

```
# A tibble: 6 x 8
  Entity      Code   Year Schizophrenia disorders (sh~1
  <chr>       <chr> <dbl>                         <dbl>
1 Afghanistan AFG    1990                         0.223
2 Afghanistan AFG    1991                         0.222
3 Afghanistan AFG    1992                         0.222
4 Afghanistan AFG    1993                         0.221
5 Afghanistan AFG    1994                         0.220
6 Afghanistan AFG    1995                         0.219
# i abbreviated name:
#   1: `Schizophrenia disorders (share of population) - Sex: Both - Age: Age-standardized`
# i 4 more variables:
#   `Depressive disorders (share of population) - Sex: Both - Age: Age-standardized` <dbl>,
#   `Anxiety disorders (share of population) - Sex: Both - Age: Age-standardized` <dbl>,
#   `Bipolar disorders (share of population) - Sex: Both - Age: Age-standardized` <dbl>,
#   `Eating disorders (share of population) - Sex: Both - Age: Age-standardized` <dbl>
```

```
head(depression_data)
```

```
# A tibble: 6 x 11
  index Entity      Code   Year `Schizophrenia (%)`
  <dbl> <chr>       <chr> <dbl>               <dbl>
```

```
1     0 Afghanistan AFG    1990               0.161
2     1 Afghanistan AFG    1991               0.160
3     2 Afghanistan AFG    1992               0.160
4     3 Afghanistan AFG    1993               0.160
5     4 Afghanistan AFG    1994               0.160
6     5 Afghanistan AFG    1995               0.160
# i 6 more variables: `Bipolar disorder (%)` <dbl>,
#   `Eating disorders (%)` <dbl>,
#   `Anxiety disorders (%)` <dbl>,
#   `Drug use disorders (%)` <dbl>,
#   `Depression (%)` <dbl>,
#   `Alcohol use disorders (%)` <dbl>
```

```
head(mental_health_finaldata)
```

```
# A tibble: 6 x 13
  Age     Gender Occupation Days_Indoors Growing_Stress
  <chr>   <chr>  <chr>             <dbl> <chr>
1 20-25   Female Corporate            NA Yes
2 30-Abo~ Male   Others                NA Yes
3 30-Abo~ Female Student               NA No
4 25-30   Male   Others                NA Yes
5 16-20   Female Student               NA Yes
6 25-30   Male   Housewife             NA No
# i 8 more variables: Quarantine_Frustrations <chr>,
#   Changes_Habits <chr>, Mental_Health_History <chr>,
#   Weight_Change <chr>, Mood_Swings <chr>,
#   Coping_Struggles <chr>, Work_Interest <chr>,
#   Social_Weakness <chr>
```

To show what the final dataset looks like, I summarized the key columns without printing the entire dataset. This provides a condensed view of the data.

What information is not self-evident?:

From the datasets, some information that is not immediately self-evident includes:

- Interactions between Variables: Understanding how different factors interact and contribute to mental health outcomes requires deeper analysis.
- Temporal Trends: Changes over time in the prevalence of mental health disorders.
- Geographical Patterns: Regional differences in the prevalence of mental health disorders.
- Predictive Relationships: Identifying which variables are strong predictors of mental health outcomes.

What are different ways you could look at this data to answer the questions you want to answer?:

Correlation Analysis: To identify relationships between different variables. • Time Series Analysis: To observe trends and changes over time. • Geospatial Analysis: To visualize regional differences. • Predictive Modeling: Using machine learning algorithms to predict mental health outcomes. • Grouping and Aggregation: Summarizing data by different categories (e.g., age groups, gender).

How do you plan to slice and dice the data??:

Grouping by Demographic Factors: Age, gender, occupation, etc. • Filtering by Specific Conditions: Focusing on specific mental health disorders. • Creating New Variables: Such as composite scores or indicators

derived from existing variables.

How could you summarize your data to answer key questions?:

Descriptive Statistics: Mean, median, standard deviation for key variables. • Frequency Tables: For categorical data such as gender, age groups, etc. • Correlation Matrices: To show relationships between variables. • Cross-tabulations: To explore interactions between categorical variables.

What types of plots and tables will help you to illustrate the findings to your questions?:

Bar Charts: For categorical comparisons. • Heat maps: To visualize correlations. • Scatter Plots: To show relationships between continuous variables. • Time Series Plots: To illustrate changes over time. • Model Performance Tables: To evaluate machine learning models.

What do you not know how to do right now that you need to learn to answer your questions?:

1. To accurately predict the likelihood of developing mental health disorders, I need to learn advanced machine learning techniques such as neural networks, support vector machines, and ensemble methods.
2. Understanding how to perform geospatial analysis to visualize and analyze the geographical distribution of mental health disorders will be essential.
3. Gaining a deeper understanding of ethical considerations and privacy laws related to handling sensitive mental health data. This includes learning about data anonymization, GDPR (General Data Protection Regulation), and other relevant regulations.

Do you plan on incorporating any machine learning techniques to answer your research questions?:

Yes, I plan to use machine learning techniques such as logistic regression, decision trees to predict mental health outcomes and identify key risk factors.

What questions do you have now, that will lead to further analysis or additional steps?:

1. How can we improve the accuracy of predictive models for mental health outcomes?
2. What are the most effective ways to visualize complex interactions between variables?
3. How can we incorporate qualitative data into our quantitative analysis?

I believe if I continue with these steps, I can ensure that the data is clean and ready for analysis, uncovering meaningful insights and answer the research questions effectively.

Visualizing the data:

```
# Inspect the first few rows of
# mental_health_finaldata after imputation
head(mental_health_finaldata)
```

```
# A tibble: 6 x 13
  Age     Gender Occupation Days_Indoors Growing_Stress
  <chr>   <chr>  <chr>             <dbl> <chr>
1 20-25   Female Corporate            NA Yes
2 30-Abo~ Male   Others                NA Yes
3 30-Abo~ Female Student               NA No
```

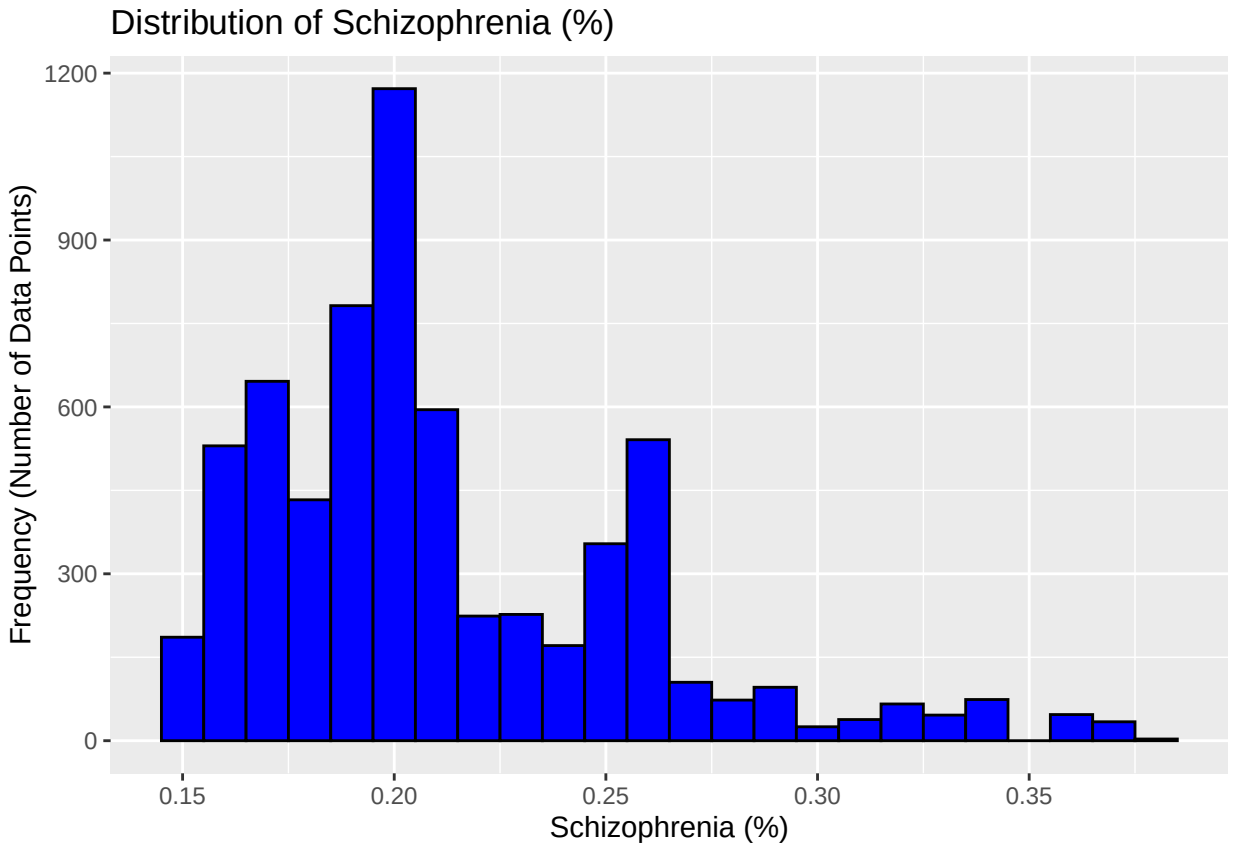
```
4 25-30   Male   Others               NA Yes
5 16-20   Female Student              NA Yes
6 25-30   Male   Housewife            NA No
# i 8 more variables: Quarantine_Frustrations <chr>,
#   Changes_Habits <chr>, Mental_Health_History <chr>,
#   Weight_Change <chr>, Mood_Swings <chr>,
#   Coping_Struggles <chr>, Work_Interest <chr>,
#   Social_Weakness <chr>
```

```r
# Get summary statistics to understand the data
summary(mental_health_finaldata)
```

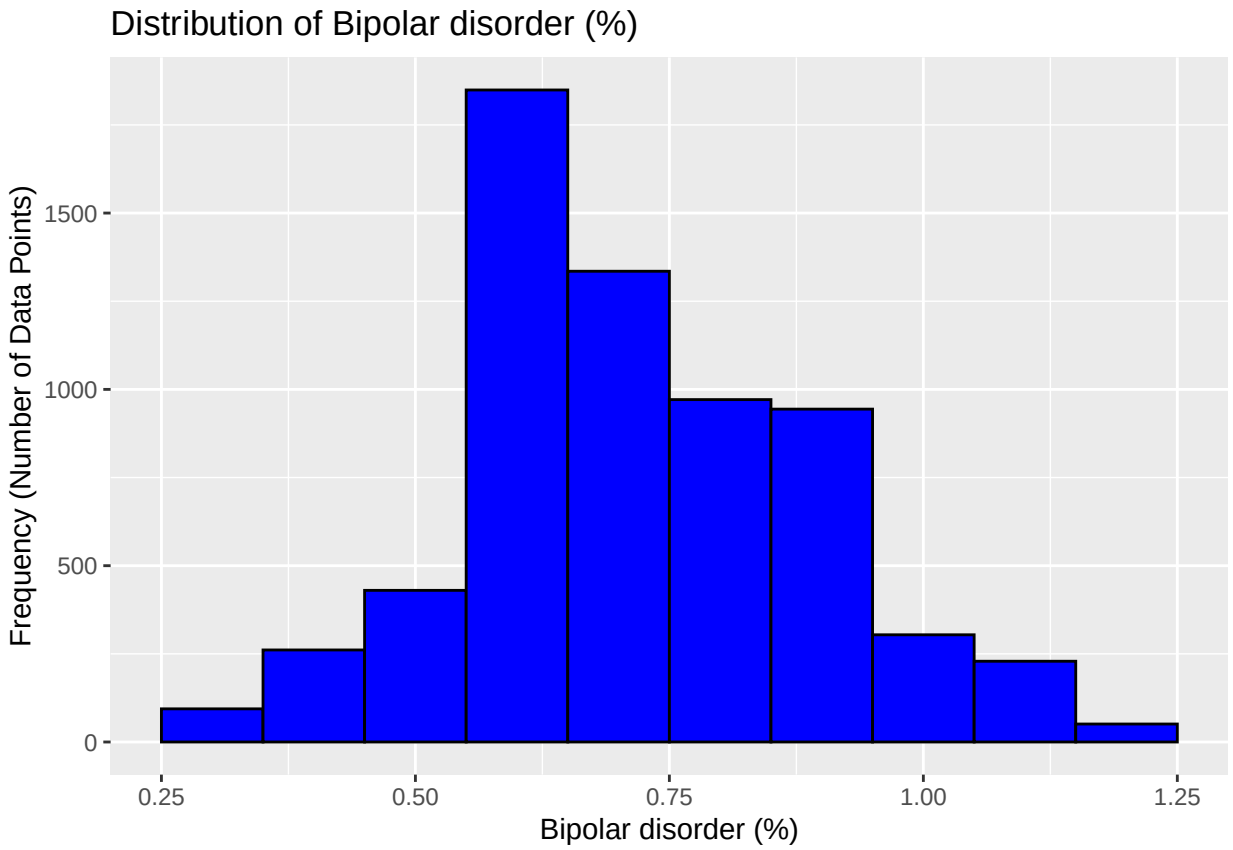
```
     Age                Gender
 Length:824         Length:824
 Class :character   Class :character
 Mode  :character   Mode  :character



  Occupation         Days_Indoors Growing_Stress
 Length:824         Min.   : NA   Length:824
 Class :character   1st Qu.: NA   Class :character
 Mode  :character   Median : NA   Mode  :character
                    Mean   :NaN
                    3rd Qu.: NA
                    Max.   : NA
                    NA's   :824
 Quarantine_Frustrations Changes_Habits
 Length:824              Length:824
 Class :character        Class :character
 Mode  :character        Mode  :character



 Mental_Health_History Weight_Change
 Length:824            Length:824
 Class :character      Class :character
 Mode  :character      Mode  :character



 Mood_Swings        Coping_Struggles
 Length:824         Length:824
 Class :character   Class :character
 Mode  :character   Mode  :character



 Work_Interest      Social_Weakness
 Length:824         Length:824
 Class :character   Class :character
```

```
 Mode   :character   Mode   :character
```

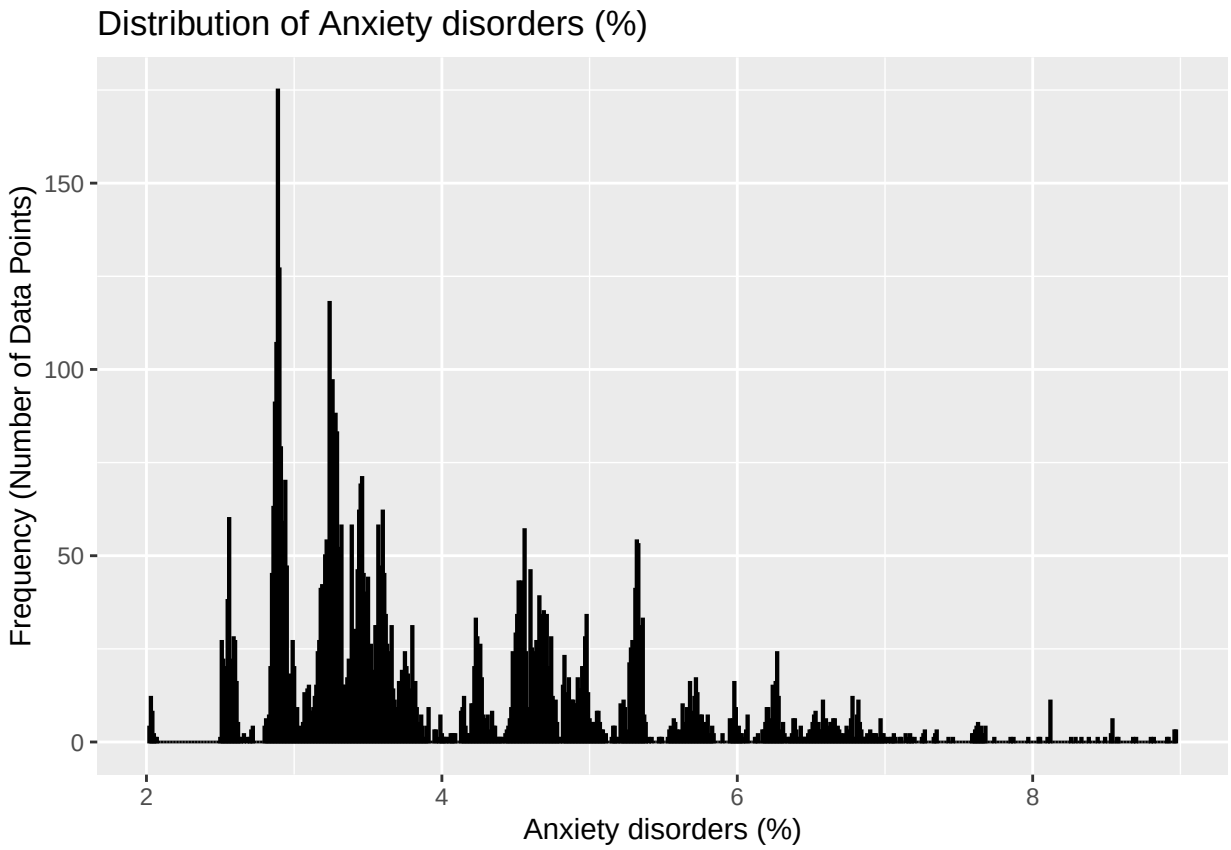
```
# Removing outliers from depression_data
depression_data_clean <- depression_data %>%
    filter(`Schizophrenia (%)` < 1, `Bipolar disorder (%)` <
        10, `Anxiety disorders (%)` < 10, `Depression (%)` <
        10)

# Visualizations Histograms for numeric variables in
# depression_data
ggplot(depression_data_clean, aes(x = `Schizophrenia (%)`)) +
    geom_histogram(binwidth = 0.01, fill = "blue", color = "black") +
    labs(title = "Distribution of Schizophrenia (%)", x = "Schizophrenia (%)",
        y = "Frequency (Number of Data Points)")
```



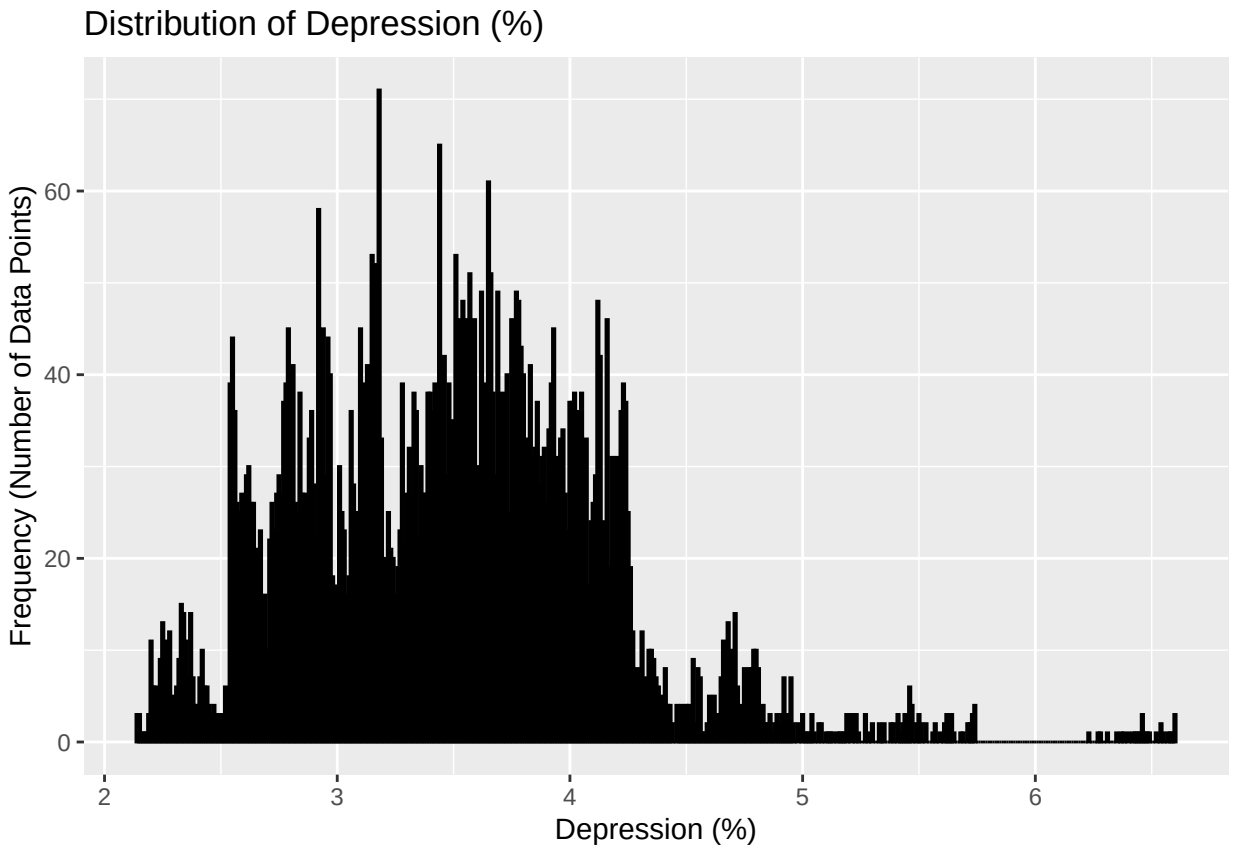

```
ggplot(depression_data_clean, aes(x = `Bipolar disorder (%)`)) +
    geom_histogram(binwidth = 0.1, fill = "blue", color = "black") +
    labs(title = "Distribution of Bipolar disorder (%)",
        x = "Bipolar disorder (%)", y = "Frequency (Number of Data Points)")
```



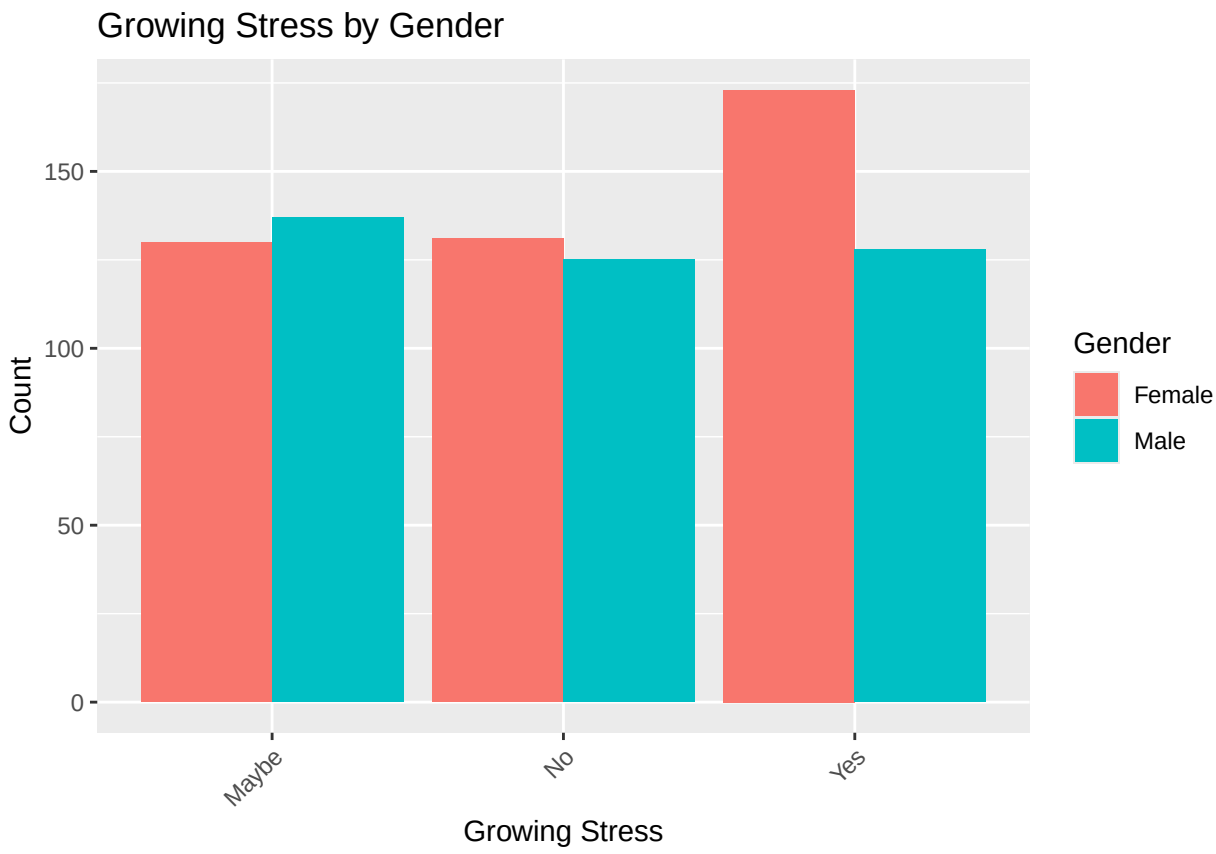

```
ggplot(depression_data_clean, aes(x = `Anxiety disorders (%)`)) +
    geom_histogram(binwidth = 0.01, fill = "blue", color = "black") +
    labs(title = "Distribution of Anxiety disorders (%)",
        x = "Anxiety disorders (%)", y = "Frequency (Number of Data Points)")
```



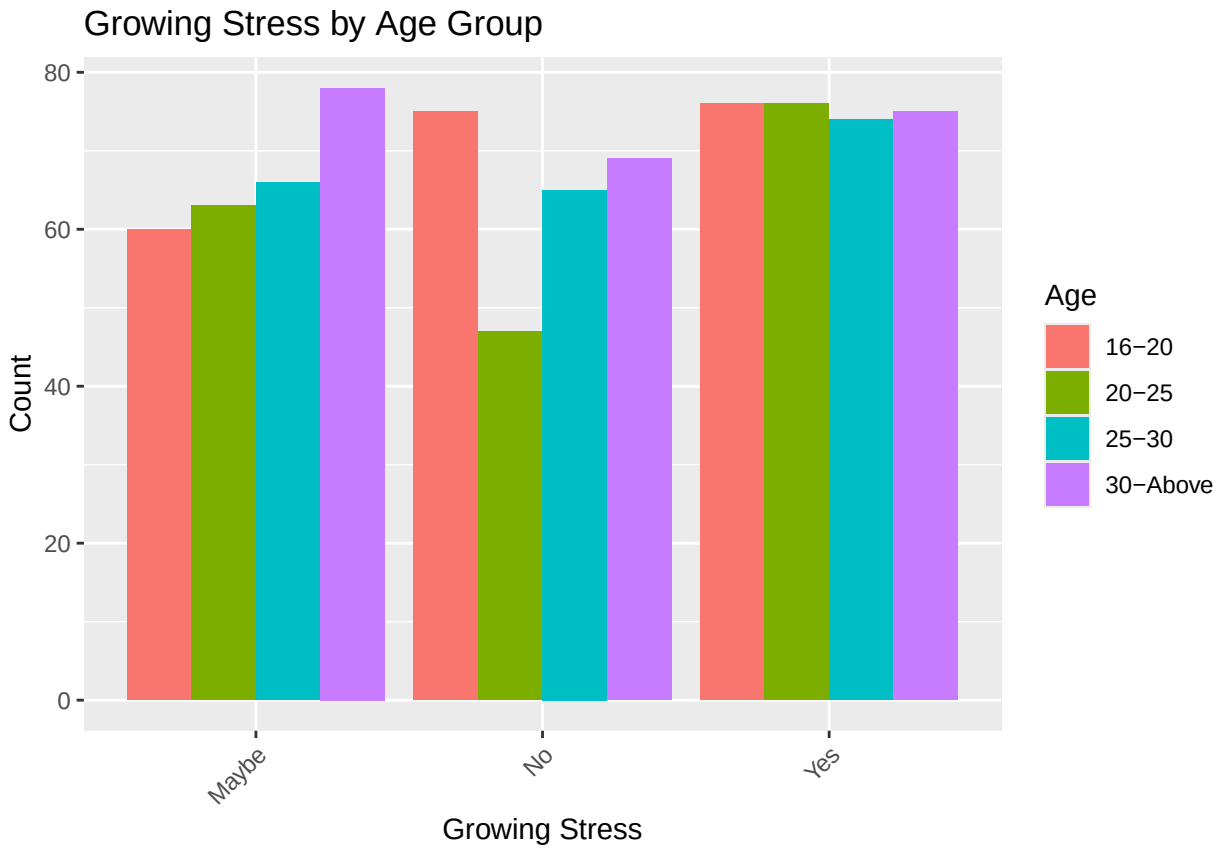

```
ggplot(depression_data_clean, aes(x = `Depression (%)`)) +
    geom_histogram(binwidth = 0.01, fill = "blue", color = "black") +
    labs(title = "Distribution of Depression (%)", x = "Depression (%)",
        y = "Frequency (Number of Data Points)")
```



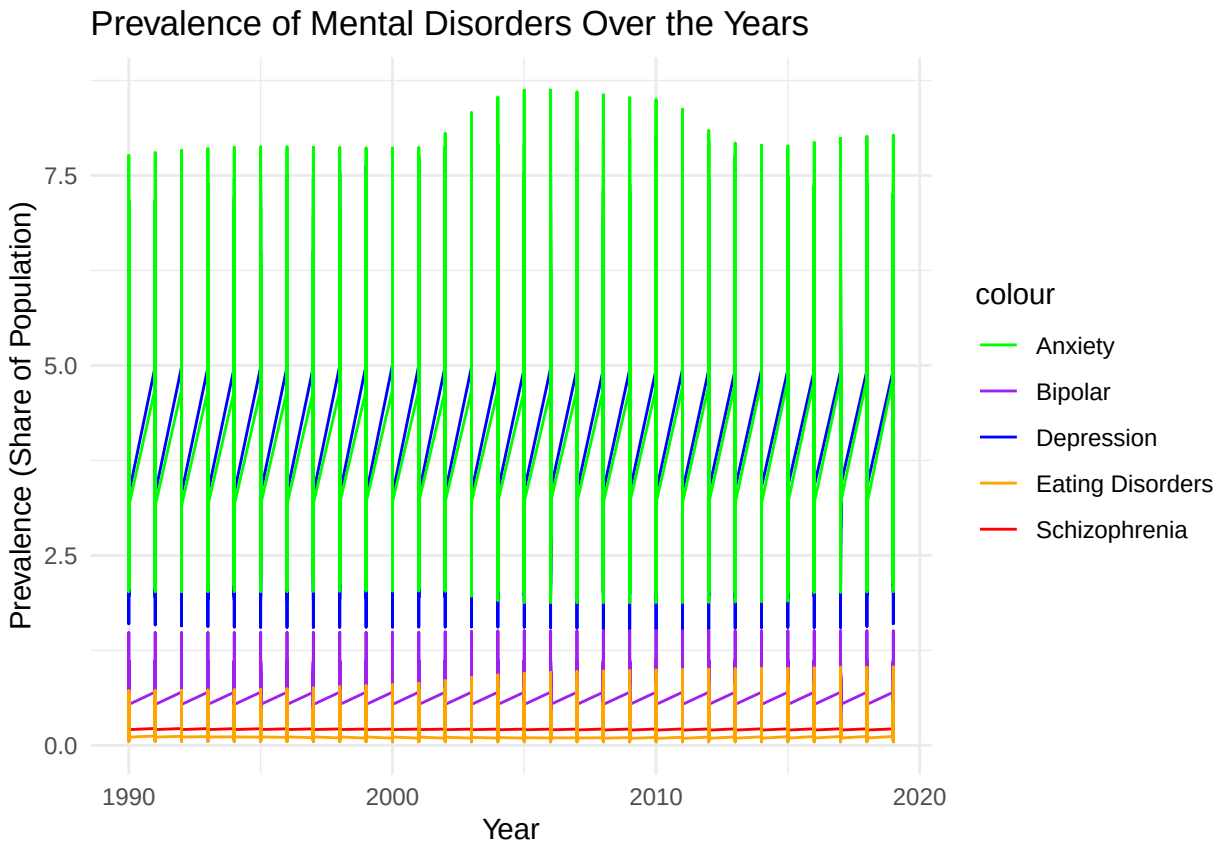

```
# Bar plots for categorical data in
# mental_health_finaldata
ggplot(mental_health_finaldata, aes(x = Growing_Stress,
    fill = Gender)) + geom_bar(position = "dodge") + labs(title = "Growing Stress by Gender",
    x = "Growing Stress", y = "Count") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1))
```



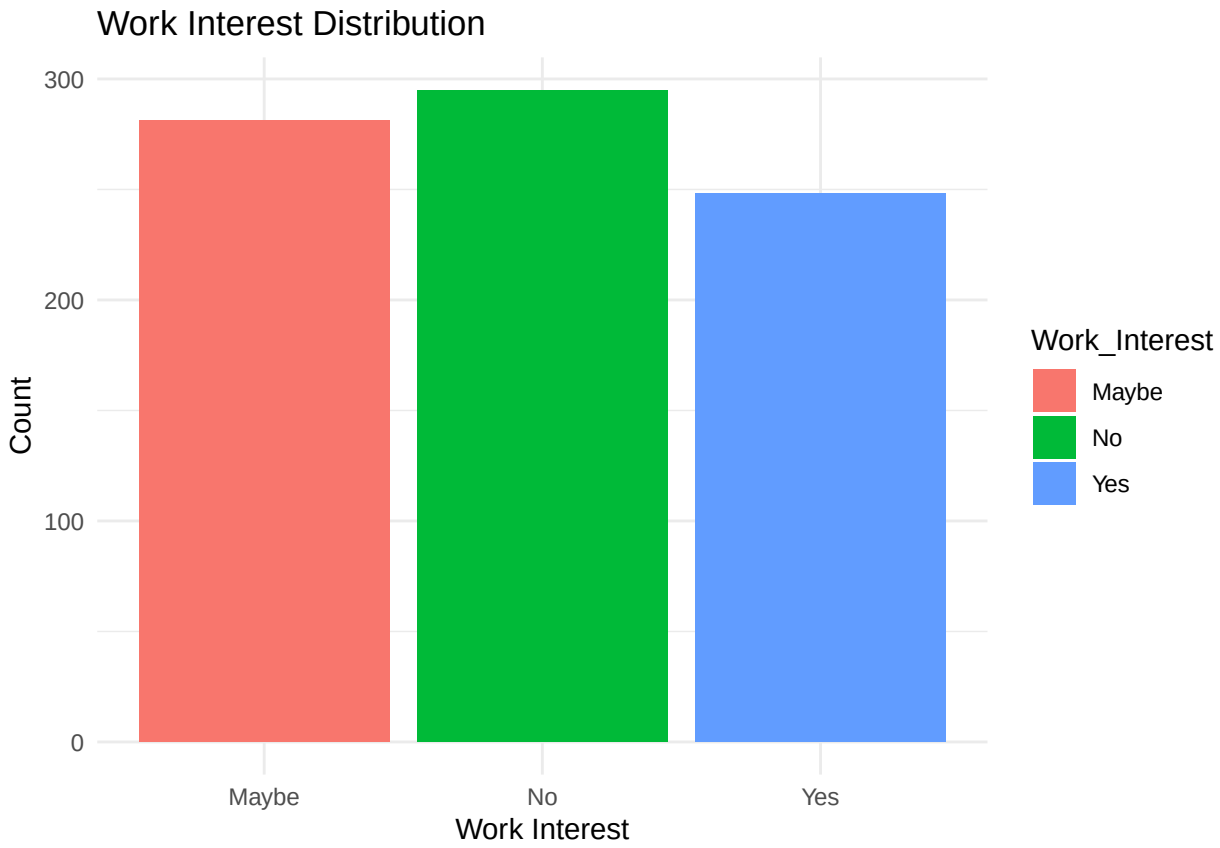

```
ggplot(mental_health_finaldata, aes(x = Growing_Stress,
    fill = Age)) + geom_bar(position = "dodge") + labs(title = "Growing Stress by Age Group",
    x = "Growing Stress", y = "Count") + theme(axis.text.x = element_text(angle = 45,
    hjust = 1))
```



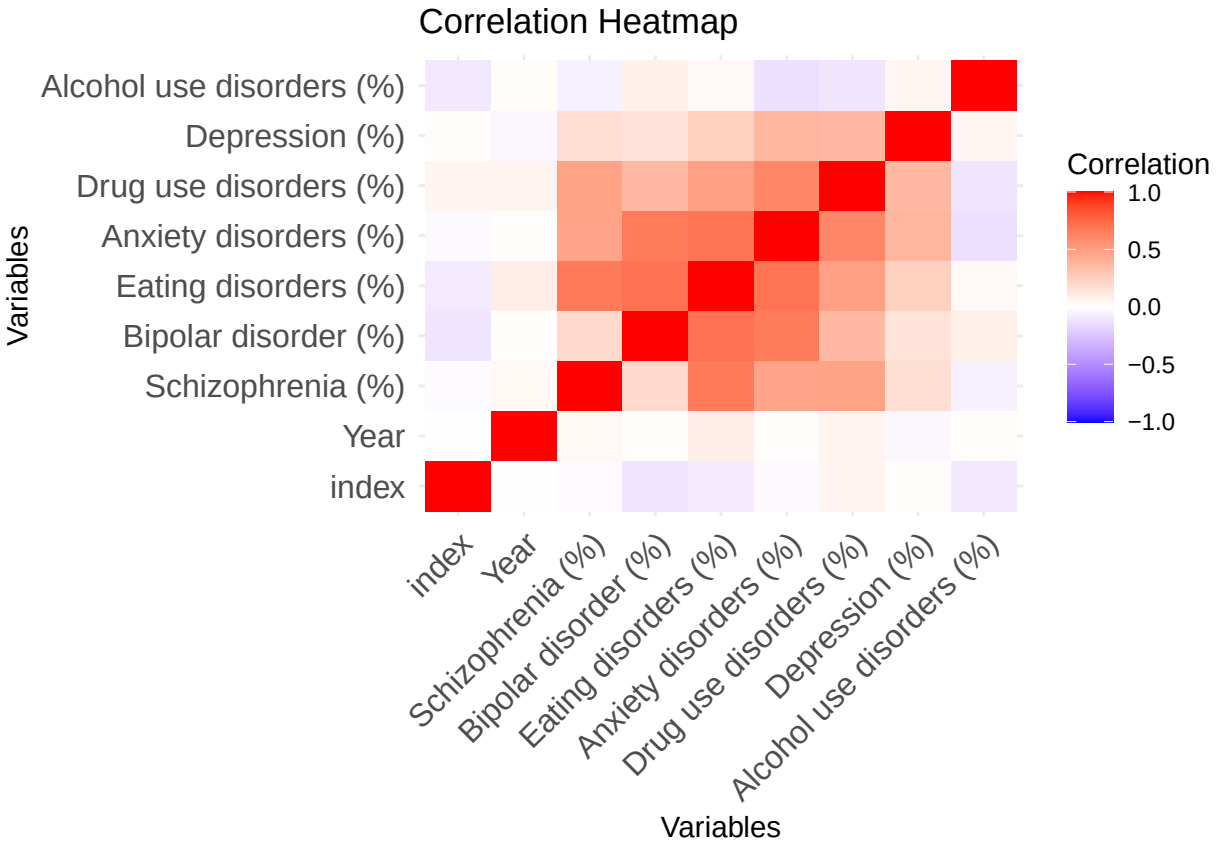

```
# Time series plot for prevalence of disorders over
# the years in mental_illnesses_prevalence
ggplot(mental_illnesses_prevalence, aes(x = Year)) + geom_line(aes(y = `Schizophrenia disorders (share
    color = "Schizophrenia")) + geom_line(aes(y = `Depressive disorders (share of population) - Sex: Bo
    color = "Depression")) + geom_line(aes(y = `Anxiety disorders (share of population) - Sex: Both - Ag
    color = "Anxiety")) + geom_line(aes(y = `Bipolar disorders (share of population) - Sex: Both - Age:
    color = "Bipolar")) + geom_line(aes(y = `Eating disorders (share of population) - Sex: Both - Age: A
    color = "Eating Disorders")) + labs(title = "Prevalence of Mental Disorders Over the Years",
    x = "Year", y = "Prevalence (Share of Population)") +
    scale_color_manual(values = c(Schizophrenia = "red",
        Depression = "blue", Anxiety = "green", Bipolar = "purple",
        `Eating Disorders` = "orange")) + theme_minimal()
```




```
# Bar plot for Days Indoors vs Work Interest
ggplot(mental_health_finaldata, aes(x = Work_Interest, fill = Work_Interest)) +
    geom_bar() + labs(title = "Work Interest Distribution",
    x = "Work Interest", y = "Count") + theme_minimal()
```




```
# Correlation heatmap for numeric variables in
# depression_data
cor_matrix <- cor(depression_data_clean %>%
    select_if(is.numeric), use = "complete.obs")
melted_cor_matrix <- melt(cor_matrix)
ggplot(data = melted_cor_matrix, aes(x = Var1, y = Var2,
    fill = value)) + geom_tile() + scale_fill_gradient2(low = "blue",
    high = "red", mid = "white", midpoint = 0, limit = c(-1,
        1), space = "Lab", name = "Correlation") + theme_minimal() +
    labs(title = "Correlation Heatmap", x = "Variables",
        y = "Variables") + theme(axis.text.x = element_text(angle = 45,
    vjust = 1, size = 12, hjust = 1), axis.text.y = element_text(size = 12))  # Adjust the size of y-ax
```



FINAL, STEP 3 -

Introduction:

Mental health disorders affect people of all sexes, ages, genders, races, and socioeconomic groups, with factors triggering illness ranging from mild to severe. These disorders often alter thinking, moods, or behaviors, making everyday life difficult. With the rising awareness and prevalence of mental health issues, analyzing collected data is crucial to understanding and addressing this public health concern. This analysis targets individuals with mental health disorders or those with a family history of such disorders, using data science techniques to aid in early detection and control.

The Problem Statement I Addressed:

The primary problem addressed in this research is the increasing prevalence of mental health disorders and the need for timely diagnosis and effective intervention strategies. The goal is to identify critical risk factors, assess the impact of environmental factors, and explore patterns in demographic and behavioral data to improve prevention, diagnosis, and treatment of mental health disorders.

How I Addressed My Problem Statement:

Data Used

Mental Illnesses Prevalence Data: This dataset includes prevalence rates of various mental health disorders across different countries and years.

Mental Health Depression Disorder Data: This dataset provides detailed insights into the prevalence of various mental health disorders.

Mental Health Final Data: A survey-based dataset capturing individual responses related to mental health during specific periods, likely during the pandemic.

Methodology Employed

Data Importing and Cleaning: The data was imported and cleaned using R, ensuring that missing values and non-numeric characters were handled appropriately.

Exploratory Data Analysis (EDA): Various visualizations were created to understand the distribution and relationships between variables.

Recommendation for Model: Based on the analysis, machine learning models such as logistic regression or random forests can be recommended to predict the likelihood of mental health disorders based on demographic and behavioral data.

Interesting Facts from Analysis

Distribution of Disorders: The histograms show the distribution of schizophrenia, bipolar disorder, anxiety disorders, and depression. These visualizations highlight the varying prevalence rates of these disorders.

Impact of Age and Gender on Stress: Bar charts show how growing stress varies across different age groups and genders, indicating demographic influences on mental health.

Trends: The time series plot reveals trends in the prevalence of mental disorders over the years, showing patterns that could be linked to social or environmental changes.

Correlation Analysis: The heatmap shows correlations between various mental health disorders, suggesting potential related risk factors.

Implications:

Preventive Measures: Identifying high-risk groups based on age, gender, or other demographic factors can help in designing preventive measures.

Policy Formulation: Understanding temporal and regional trends can assist policymakers in allocating resources more effectively.

Treatment Optimization: Correlation analysis can guide healthcare providers in recognizing comorbid conditions and optimizing treatment strategies.

Limitations:

Data Quality: Missing values and potential inaccuracies in the data can affect the analysis.

Generalization: The findings might not be generalization to all populations due to demographic and regional differences in the data sets.

Scope of Variables: The analysis is limited to the available variables, and other important factors might not have been considered.

Conclusion:

This analysis highlights the importance of data-driven approaches in understanding and addressing mental health disorders. While the insights gained are valuable, further research incorporating additional variables and more comprehensive datasets could enhance the understanding and effectiveness of interventions. Future work could also involve implementing and validating predictive models to improve early detection and personalized treatment strategies.