Analyzing the Impact of Technology Usage on Mental Health Across Different Demographics

Ezra Kipkurui Bii

2025-04-10

1.0. Executive Summary

This project investigates the relationship between technology usage and mental health across different demographic groups, based on a dataset of 10,000 individuals. A range of methods, including feature engineering, exploratory data analysis, statistical testing, and predictive modeling, were applied. While predictive models demonstrated limited accuracy, the analysis uncovered meaningful patterns, notably the significant role of support systems in influencing mental health outcomes. These findings highlight the complexity of mental health factors and suggest the need for broader approaches in future research.

2.0. Data Importation & Exploration

```r
#2.1. Loading the necessary libraries
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(plotly)
library(ggpubr)
library(car)
library(cluster)
library(forcats)
library(randomForest)
library(nnet)
library(xgboost)
library(MASS)
msg <- ("All Libraries Loaded Successfully")
print(msg)
```

```
## [1] "All Libraries Loaded Successfully"
```

the above code loads the necessary libraries required in the entire analysis. Below is the summary of the work of each package loaded;

readr: For reading CSV files.

dplyr: For data manipulation tasks such as filtering, summarizing, and arranging data.

tidyr: For tidying data, including functions for handling missing values and reshaping data.

ggplot2: For creating high-quality static visualizations.

plotly: For creating interactive plots.

ggpubr: Provides easy-to-use functions for creating publication-ready plots.

stats: For various statistical tests and modeling (this is a base package in R).

car: Companion to Applied Regression, useful for advanced regression diagnostics.

cluster: For cluster analysis.

forcats: For working with categorical variables (factors).

randomForest: For building ensembles of decision trees to perform classification and regression accuracy

nnet: For modeling patterns in data

MASS: For advanced regression analysis

```
#2.2.Dataset loading
tryCatch({
  Dataset <- read.csv("C:/Users/ADMIN/Documents/EZ/PROJECT/mental_health_and_technology_usage_2024.csv"
  print("Data Loaded successfully and Ready For Analysis")
}, error = function(e) {
  print("Failed to load data. Review your file path.")
})
```

```
## [1] "Data Loaded successfully and Ready For Analysis"
```

```
#2.3. Preview of Dataset content
head(Dataset)
```

```
##      User_ID Age Gender Technology_Usage_Hours Social_Media_Usage_Hours
## 1 USER-00001  23 Female                   6.57                     6.00
## 2 USER-00002  21   Male                   3.01                     2.57
## 3 USER-00003  51   Male                   3.04                     6.14
## 4 USER-00004  25 Female                   3.84                     4.48
## 5 USER-00005  53   Male                   1.20                     0.56
## 6 USER-00006  58   Male                   5.59                     5.74
##   Gaming_Hours Screen_Time_Hours Mental_Health_Status Stress_Level Sleep_Hours
## 1         0.68             12.36                 Good          Low        8.01
## 2         3.74              7.61                 Poor         High        7.28
## 3         1.26              3.16                 Fair         High        8.04
## 4         2.59             13.08            Excellent       Medium        5.62
## 5         0.29             12.63                 Good          Low        5.55
## 6         0.11              1.34                 Poor          Low        8.61
##   Physical_Activity_Hours Support_Systems_Access Work_Environment_Impact
## 1                    6.71                     No                Negative
## 2                    5.88                    Yes                Positive
## 3                    9.81                     No                Negative
## 4                    5.28                    Yes                Negative
## 5                    4.00                     No                Positive
## 6                    6.54                    Yes                 Neutral
##   Online_Support_Usage
## 1                  Yes
## 2                   No
## 3                   No
## 4                  Yes
## 5                  Yes
## 6                  Yes
```

```
#2.4. checking for missing values
cat(paste("The total missing values is:", sum(is.na(Dataset))))
```

```
## The total missing values is: 0
```

3.0. Data Cleaning & Feature Engineering

3.1. Data Cleaning

```
#3.1.1 Converting categorical variables to factors
Dataset <- Dataset %>%
  mutate(
    Gender = as.factor(Gender),
    Mental_Health_Status = as.factor(Mental_Health_Status),
    Stress_Level = as.factor(Stress_Level),
    Support_Systems_Access = as.factor(Support_Systems_Access),
    Work_Environment_Impact = as.factor(Work_Environment_Impact),
    Online_Support_Usage = as.factor(Online_Support_Usage)
  )

#3.1.2. Ensuring the numeric columns are properly formatted
Dataset <- Dataset %>%
  mutate(across(c(Age, Technology_Usage_Hours, Social_Media_Usage_Hours,
                  Gaming_Hours, Screen_Time_Hours, Sleep_Hours, Physical_Activity_Hours),
                as.numeric))

#3.1.3. Checking the structure of the cleaned Dataset
str(Dataset)
```

```
## ’data.frame':    10000 obs. of  14 variables:
##  $ User_ID                 : chr  "USER-00001" "USER-00002" "USER-00003" "USER-00004" ...
##  $ Age                     : num  23 21 51 25 53 58 63 51 57 31 ...
##  $ Gender                  : Factor w/ 3 levels "Female","Male",..: 1 2 2 1 2 2 1 1 3 3 ...
##  $ Technology_Usage_Hours  : num  6.57 3.01 3.04 3.84 1.2 ...
##  $ Social_Media_Usage_Hours: num  6 2.57 6.14 4.48 0.56 5.74 2.55 4.1 4.11 7.23 ...
##  $ Gaming_Hours            : num  0.68 3.74 1.26 2.59 0.29 0.11 3.79 4.74 0.08 0.81 ...
##  $ Screen_Time_Hours       : num  12.36 7.61 3.16 13.08 12.63 ...
##  $ Mental_Health_Status    : Factor w/ 4 levels "Excellent","Fair",..: 3 4 2 1 3 4 1 1 2 1 ...
##  $ Stress_Level            : Factor w/ 3 levels "High","Low","Medium": 2 1 1 3 2 2 3 3 3 1 ...
##  $ Sleep_Hours             : num  8.01 7.28 8.04 5.62 5.55 8.61 8.61 7.11 7.19 5.09 ...
##  $ Physical_Activity_Hours : num  6.71 5.88 9.81 5.28 4 6.54 1.34 5.27 5.22 0.47 ...
##  $ Support_Systems_Access  : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 2 2 1 1 ...
##  $ Work_Environment_Impact : Factor w/ 3 levels "Negative","Neutral",..: 1 3 1 1 3 2 2 2 3 3 ...
##  $ Online_Support_Usage    : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 1 1 1 1 ...
```

```
#3.1.4. Summary of the cleaned Dataset
summary(Dataset)
```

```
##    User_ID               Age           Gender     Technology_Usage_Hours
##  Length:10000       Min.   :18.00   Female:3286   Min.   : 1.000        
##  Class :character   1st Qu.:29.00   Male  :3350   1st Qu.: 3.760        
##  Mode  :character   Median :42.00   Other :3364   Median : 6.425        
##                     Mean   :41.52                 Mean   : 6.474        
##                     3rd Qu.:54.00                 3rd Qu.: 9.213        
##                     Max.   :65.00                 Max.   :12.000        
##  Social_Media_Usage_Hours  Gaming_Hours   Screen_Time_Hours
##  Min.   :0.000            Min.   :0.000   Min.   : 1.000   
##  1st Qu.:1.980            1st Qu.:1.260   1st Qu.: 4.520   
##  Median :3.950            Median :2.520   Median : 7.900   
##  Mean   :3.972            Mean   :2.516   Mean   : 7.976   
##  3rd Qu.:5.990            3rd Qu.:3.790   3rd Qu.:11.500   
##  Max.   :8.000            Max.   :5.000   Max.   :15.000   
##  Mental_Health_Status Stress_Level   Sleep_Hours    Physical_Activity_Hours
##  Excellent:2518       High  :3330   Min.   :4.000   Min.   : 0.000         
##  Fair     :2490       Low   :3332   1st Qu.:5.260   1st Qu.: 2.490         
##  Good     :2508       Medium:3338   Median :6.500   Median : 4.990         
##  Poor     :2484                     Mean   :6.501   Mean   : 5.004         
##                                     3rd Qu.:7.760   3rd Qu.: 7.540         
##                                     Max.   :9.000   Max.   :10.000         
##  Support_Systems_Access Work_Environment_Impact Online_Support_Usage
##  No :5006               Negative:3378           No :5013            
##  Yes:4994               Neutral :3312           Yes:4987            
##                         Positive:3310                               
##                                                                     
##                                                                     
##                                                                     
```

This data set contains 10,000 observations across 14 variables related to digital behavior and mental health.

Variable Breakdown:

1. *User_ID:* Unique identifier for each participant.
2. *Age:* Ranges from 18 to 65 years, with a mean of 41.52. The age distribution shows a relatively even spread among adults.
3. *Gender:* Balanced distribution across three categories – Female (3286), Male (3350), and Other (3364).
4. *Technology_Usage_Hours:* Mean usage is 6.47 hours/day, with most users ranging between 3.76 and 9.21 hours.
5. *Social_Media_Usage_Hours:* Average is 3.97 hours/day, indicating moderate engagement with social platforms.
6. *Gaming_Hours:* Mean gaming time is 2.52 hours/day, with a maximum of 5 hours.
7. *Screen_Time_Hours:* High average of 7.98 hours/day, showing significant screen exposure.
8. *Mental_Health_Status:* Fairly even distribution across statuses – Excellent (2518), Fair (2490), Good (2508), Poor (2484).

9. *Stress_Level:* Almost equally distributed – High (3330), Low (3332), Medium (3338).
10. *Sleep_Hours:* Average sleep duration is 6.5 hours, slightly below the recommended 7–9 hours.
11. *Physical_Activity_Hours:* Mean of 5 hours, suggesting moderate physical activity.
12. *Support_Systems_Access:* Nearly equal access – Yes (4994), No (5006).
13. *Work_Environment_Impact:* Even split among Negative (3378), Neutral (3312), and Positive (3310).
14. *Online_Support_Usage:* Slightly more respondents use online support systems – Yes (4987), No (5013).

General Insight:

The data set is well-balanced across demographics and categorical variables. It provides a rich ground for analyzing correlations between technology usage, mental health, and lifestyle behaviors.

3.2. Feature Engineering

```
Dataset$Technology_Usage_Hours <- Dataset$Technology_Usage_Hours +
                        Dataset$Social_Media_Usage_Hours +
                        Dataset$Gaming_Hours

Dataset$Screen_to_Sleep_Ratio <- Dataset$Screen_Time_Hours / Dataset$Sleep_Hours

Dataset$Active_Lifestyle <- ifelse(Dataset$Physical_Activity_Hours >= 5, 1, 0)

Dataset$Youth_Category <- ifelse(Dataset$Age <= 25, 1, 0)

Dataset$Tech_Addiction_Risk <- ifelse(Dataset$Technology_Usage_Hours >= 10, 1, 0)

Dataset$Support_Index <- ifelse(Dataset$Support_Systems_Access == "Yes", 1, 0) +
                    ifelse(Dataset$Online_Support_Usage == "Yes", 1, 0)

Dataset$Work_Impact_Score <- ifelse(Dataset$Work_Environment_Impact == "Positive", 1,
                          ifelse(Dataset$Work_Environment_Impact == "Neutral", 0, -1))
# Creating Age_Group based on Age
Dataset$Age_Group <- cut(Dataset$Age,
                         breaks = c(0, 18, 25, 35, 50, Inf),
                         labels = c("Teenager", "Youth", "Young Adult", "Adult", "Senior Adult"),
                         right = TRUE)

# preview of the engineered data
head(Dataset[c("Screen_to_Sleep_Ratio", "Active_Lifestyle",
          "Youth_Category", "Tech_Addiction_Risk", "Support_Index", "Work_Impact_Score", "Age_Group")])
```

```
##   Screen_to_Sleep_Ratio Active_Lifestyle Youth_Category Tech_Addiction_Risk
## 1             1.5430712                1              1                   1
## 2             1.0453297                1              1                   0
## 3             0.3930348                1              0                   1
## 4             2.3274021                1              1                   1
## 5             2.2756757                0              0                   0
## 6             0.1556330                1              0                   1
##   Support_Index Work_Impact_Score    Age_Group
## 1             1                -1        Youth
## 2             1                 1        Youth
## 3             0                -1 Senior Adult
## 4             2                -1        Youth
## 5             1                 1 Senior Adult
## 6             2                 0 Senior Adult
```

Summary of New Features:

Technology_Usage_Hours: Total time spent on general technology, social media, and gaming.

Screen_to_Sleep_Ratio: Measures balance between screen exposure and sleep.

Active_Lifestyle: Binary indicator (1 = 5 or more hours of physical activity per day).

Youth_Category: Binary indicator for respondents aged 25 or younger.

Tech_Addiction_Risk: Flags users with10 hours of technology usage daily.

Support_Index: Score (0–2) indicating access to both offline and online support.

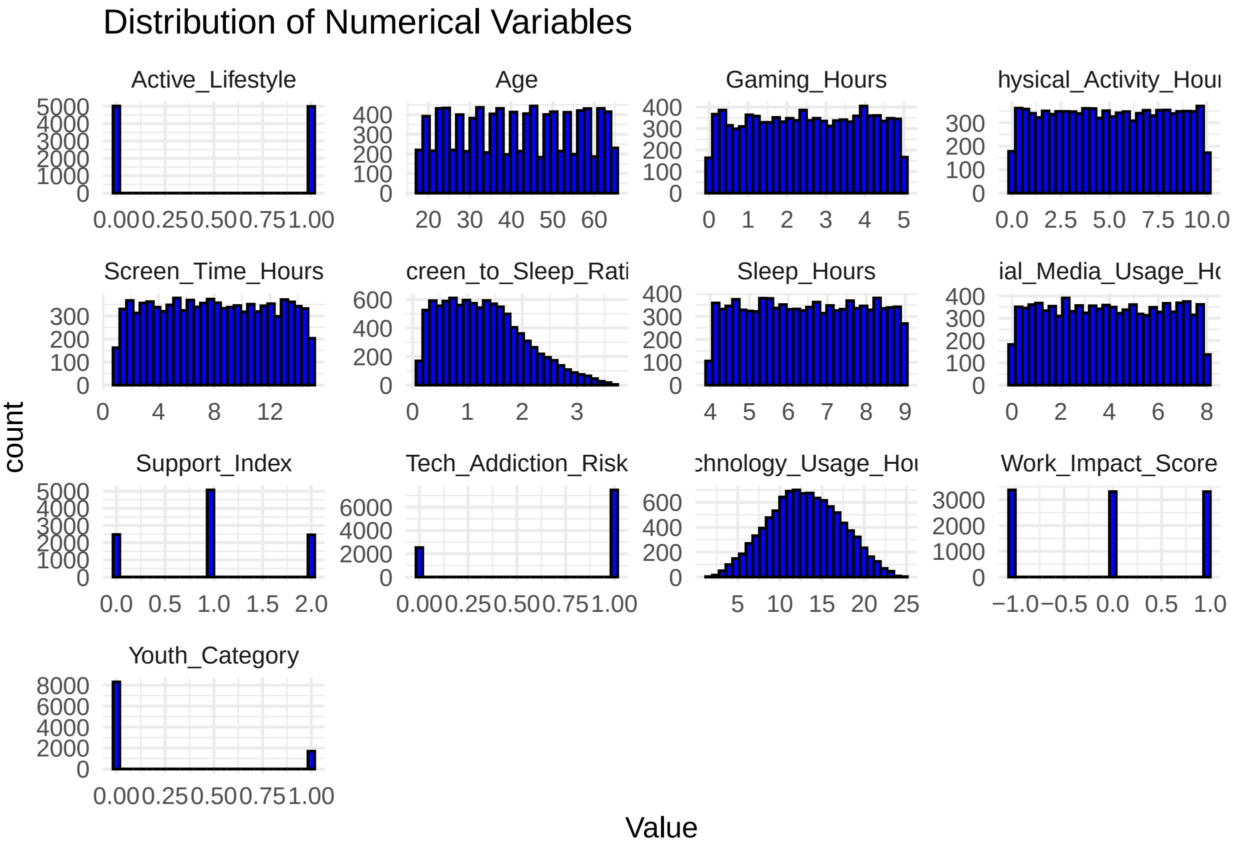
Work_Impact_Score: Encodes the perceived impact of the work environment (-1 = Negative, 0 = Neutral, 1 = Positive).

4.0. Exploratory Data Analysis

4.1. General Distribution

4.1.1. Histograms for numerical variables

```r
# i. Selecting numeric columns
numeric_data <- Dataset[, sapply(Dataset, is.numeric)]

#ii. Plot
numeric_data %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  facet_wrap(~ Variable, scales = 'free') +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Numerical Variables")
```



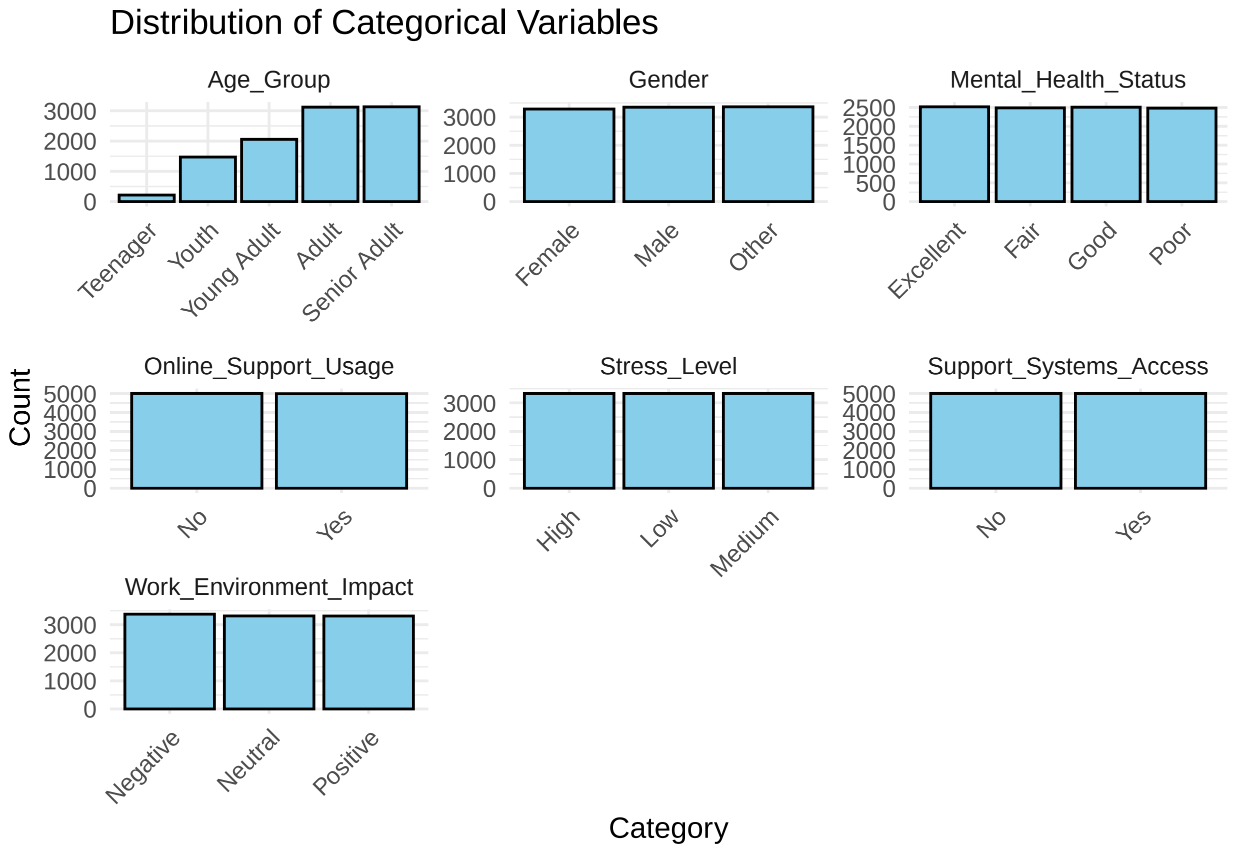
Interpretation:

- Age shows fair and uniform distribution between 15 and 60 years, suggesting a balanced sample across age groups.
- ‘Gaming_Hours, Physical_Activity_Hours, Screen_Time_Hours, Sleep_Hours, and Social_Media_Usage_Hours: These are almost evenly spread, suggesting good variability in habits across individuals.
- Screen_to_Sleep_Ratio: Right-skewed — most people have a lower screen-to-sleep ratio, but a few have very high ratios high screen time compared to sleep.
- Technology_Usage_Hours: Bell-shaped (normal distribution), indicating most people have moderate technology use, with fewer individuals using very little or a lot.
- Work_Impact_Score: Highly categorical (values clustered at -1, 0, and 1), probably representing negative, neutral, and positive work impacts.
- Active_Lifestyle, Tech_Addiction_Risk, Youth_Category, and Support_Index: are binary or categorical.

4.1.2. Bar plots for categorical variables

```
#i.  Selecting only categorical columns
categorical_data <- Dataset[, sapply(Dataset, is.factor)]

# ii. Plot
```

```r
categorical_data %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Category") %>%
  ggplot(aes(x = Category)) +
  facet_wrap(~ Variable, scales = 'free') +
  geom_bar(fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Categorical Variables", x = "Category", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

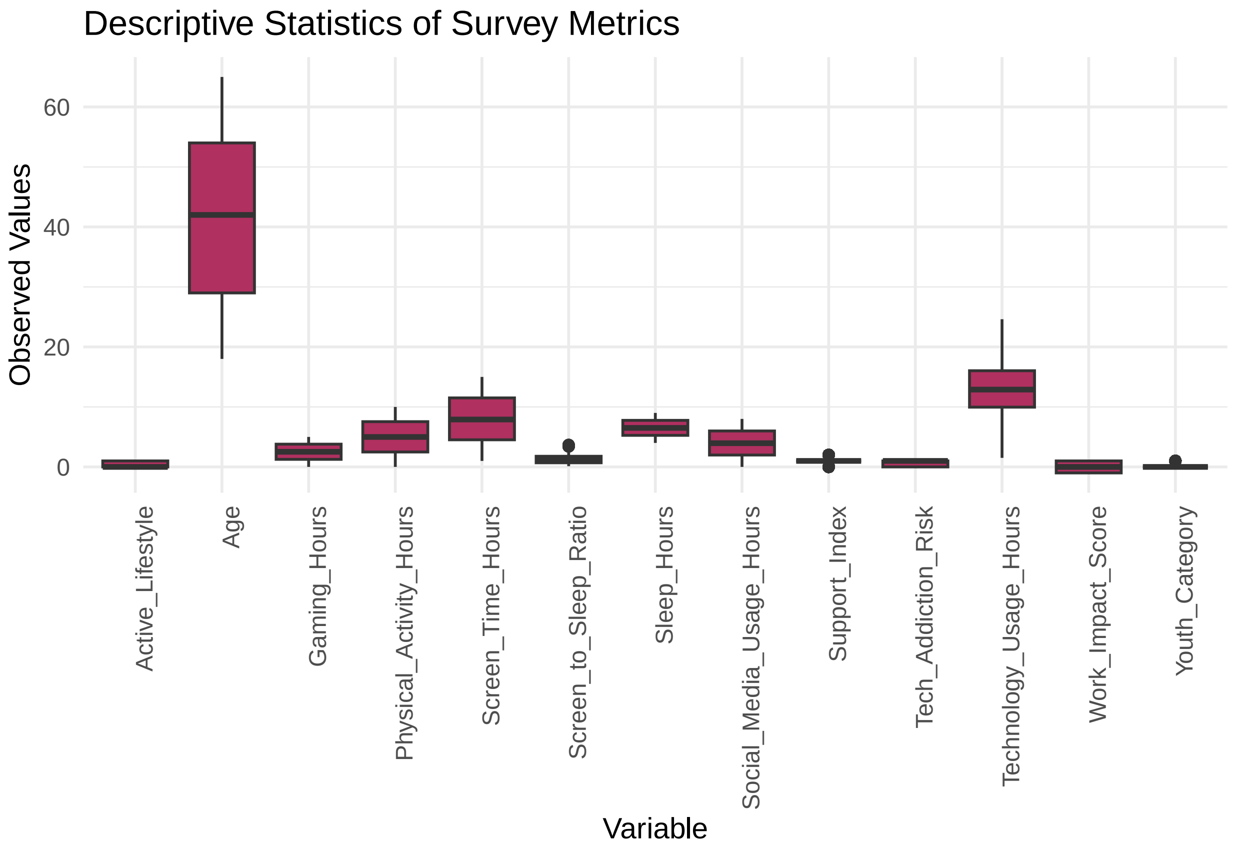


Interpretation:

- *Age_Group*: There is a low number of teenagers. More balanced distribution among Youth, Young Adults, Adults, and Senior Adults.
- *Gender:* Male and Female are almost equa there is also a small representation of "Other" gender category.
- *Mental_Health_Status:* There is fair spread across Excellent, Good, Fair, and Poor. No single mental health status dominates.
- *Online_Support_Usage:* Participants who do not use online support systems are almost equal with the ones who does.
- *Stress_Level:* High, Medium, and Low stress levels are almost evenly distributed.
- *Support_Systems_Access:* Nearly equal access and non-access to support systems.
- *Work_Environment_Impact:* Negative, Neutral, and Positive impacts are evenly distributed.

4.1.3. Boxplots to detect outliers

```r
# i. Selecting numeric data
numeric_data <- Dataset[, sapply(Dataset, is.numeric)]

# ii. Creating boxplots
numeric_data %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value") %>%
  ggplot(aes(x = Variable, y = Value)) +
  geom_boxplot(fill = "maroon") +
  theme_minimal() +
  labs(title = "Descriptive Statistics of Survey Metrics", x = "Variable", y = "Observed Values") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Insights:

- Screen_Time_Hours and Technology_Usage_Hours have the widest spread, indicating high variability among participants.
- Physical_Activity_Hours, Sleep_Hours, and Social_Media_Usage_Hours show moderate variability.
- Support_Index, Stress_Level_Num, Tech_Addiction_Risk, Youth_Category, and Active_Lifestyle have very small ranges, suggesting these measures are quite consistent across participants.
- There are several outliers in variables like Support_Index, Screen_to_Sleep_Ratio, and Technology_Usage_Hours
- Age appears to consist of tight age group with no outliers

4.2. Correlation analysis

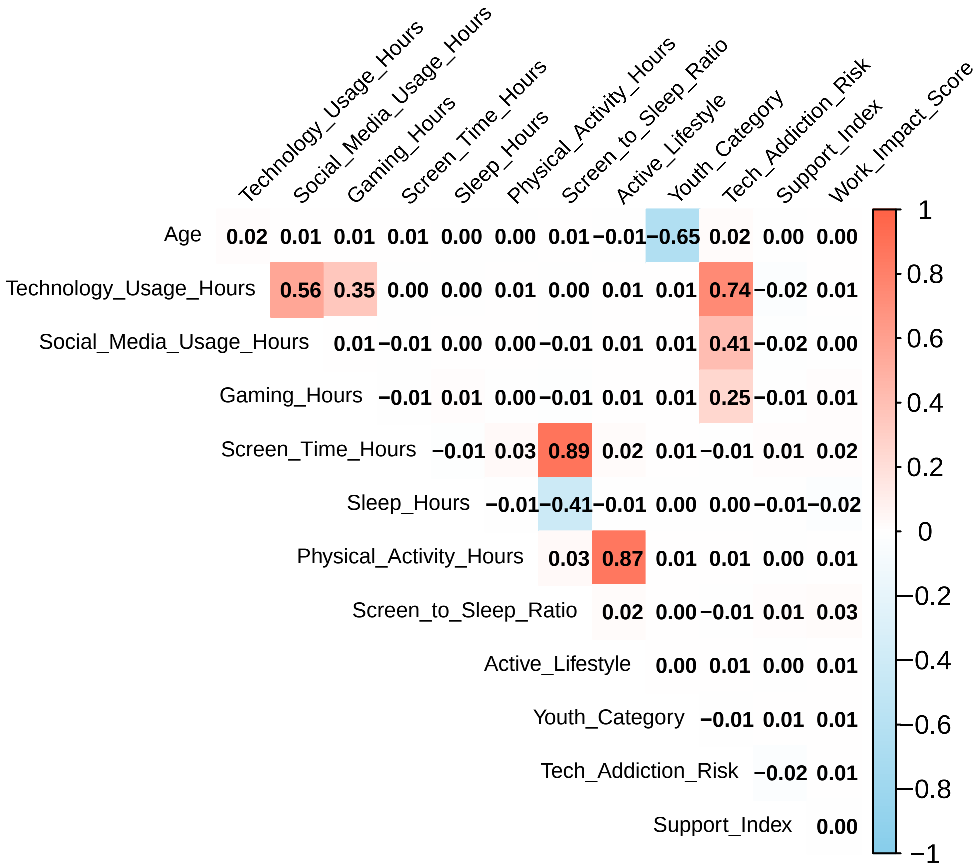
```r
# Selecting numeric variables
numeric_data <- Dataset[, sapply(Dataset, is.numeric)]

# Calculating correlation matrix
cor_matrix <- cor(numeric_data, use = "complete.obs")

# View of matrix
print(cor_matrix)
```

```
##                                    Age Technology_Usage_Hours
## Age                       1.0000000000            0.019803403
## Technology_Usage_Hours    0.0198034029            1.000000000
## Social_Media_Usage_Hours  0.0091514913            0.564658703
## Gaming_Hours              0.0052042845            0.354902447
## Screen_Time_Hours         0.0071687953           -0.001311492
## Sleep_Hours              -0.0015125049           -0.001355965
## Physical_Activity_Hours  -0.0045262945            0.008766976
## Screen_to_Sleep_Ratio     0.0076579099           -0.003291245
## Active_Lifestyle         -0.0070447338            0.009931399
## Youth_Category           -0.6476670744            0.005155473
## Tech_Addiction_Risk       0.0240515299            0.741120026
## Support_Index            -0.0036853371           -0.020231891
## Work_Impact_Score         0.0006875576            0.006383468
##                          Social_Media_Usage_Hours  Gaming_Hours
## Age                                  0.0091514913  0.0052042845
## Technology_Usage_Hours               0.5646587026  0.3549024471
## Social_Media_Usage_Hours             1.0000000000  0.0058122166
## Gaming_Hours                         0.0058122166  1.0000000000
## Screen_Time_Hours                   -0.0084008348 -0.0078173474
## Sleep_Hours                          0.0044433041  0.0103926737
## Physical_Activity_Hours              0.0023237031 -0.0004205164
## Screen_to_Sleep_Ratio               -0.0087423257 -0.0133384263
## Active_Lifestyle                     0.0072839786  0.0061439715
## Youth_Category                       0.0070080066  0.0081703789
## Tech_Addiction_Risk                  0.4144542959  0.2518132783
## Support_Index                       -0.0171167324 -0.0091136275
## Work_Impact_Score                   -0.0001222082  0.0101238573
##                          Screen_Time_Hours   Sleep_Hours
## Age                            0.007168795 -0.0015125049
## Technology_Usage_Hours        -0.001311492 -0.0013559655
## Social_Media_Usage_Hours      -0.008400835  0.0044433041
## Gaming_Hours                  -0.007817347  0.0103926737
## Screen_Time_Hours              1.000000000 -0.0111805581
## Sleep_Hours                   -0.011180558  1.0000000000
## Physical_Activity_Hours        0.030501789 -0.0099957161
## Screen_to_Sleep_Ratio          0.886348728 -0.4136224373
## Active_Lifestyle               0.022239703 -0.0070333649
## Youth_Category                 0.007576477 -0.0004035935
## Tech_Addiction_Risk           -0.008274265 -0.0024815757
## Support_Index                  0.010406833 -0.0065321294
## Work_Impact_Score              0.015334448 -0.0233431496
##                          Physical_Activity_Hours Screen_to_Sleep_Ratio
## Age                                -0.0045262945           0.007657910
## Technology_Usage_Hours              0.0087669761          -0.003291245
## Social_Media_Usage_Hours            0.0023237031          -0.008742326
## Gaming_Hours                       -0.0004205164          -0.013338426
## Screen_Time_Hours                   0.0305017892           0.886348728
## Sleep_Hours                        -0.0099957161          -0.413622437
## Physical_Activity_Hours             1.0000000000           0.031473691
## Screen_to_Sleep_Ratio               0.0314736908           1.000000000
## Active_Lifestyle                    0.8674900191           0.022496767
## Youth_Category                      0.0066876268           0.004611810
## Tech_Addiction_Risk                 0.0063450903          -0.007238776
## Support_Index                      -0.0007363165           0.012750609
## Work_Impact_Score                   0.0076901143           0.025128805
##                          Active_Lifestyle Youth_Category Tech_Addiction_Risk
## Age                          -0.007044734  -0.6476670744         0.024051530
## Technology_Usage_Hours        0.009931399   0.0051554731         0.741120026
## Social_Media_Usage_Hours      0.007283979   0.0070080066         0.414454296
## Gaming_Hours                  0.006143972   0.0081703789         0.251813278
## Screen_Time_Hours             0.022239703   0.0075764775        -0.008274265
## Sleep_Hours                  -0.007033365  -0.0004035935        -0.002481576
## Physical_Activity_Hours       0.867490019   0.0066876268         0.006345090
## Screen_to_Sleep_Ratio         0.022496767   0.0046118105        -0.007238776
## Active_Lifestyle              1.000000000   0.0047258369         0.008619034
## Youth_Category                0.004725837   1.0000000000        -0.009367158
## Tech_Addiction_Risk           0.008619034  -0.0093671581         1.000000000
## Support_Index                -0.004991254   0.0057788872        -0.024850823
## Work_Impact_Score             0.005362181   0.0086451918         0.005018593
##                          Support_Index Work_Impact_Score
## Age                      -0.0036853371      0.0006875576
## Technology_Usage_Hours   -0.0202318909      0.0063834679
## Social_Media_Usage_Hours -0.0171167324     -0.0001222082
## Gaming_Hours             -0.0091136275      0.0101238573
## Screen_Time_Hours         0.0104068330      0.0153344481
## Sleep_Hours              -0.0065321294     -0.0233431496
## Physical_Activity_Hours  -0.0007363165      0.0076901143
## Screen_to_Sleep_Ratio     0.0127506087      0.0251288046
## Active_Lifestyle         -0.0049912536      0.0053621813
## Youth_Category            0.0057788872      0.0086451918
## Tech_Addiction_Risk      -0.0248508228      0.0050185925
## Support_Index             1.0000000000     -0.0042027348
## Work_Impact_Score        -0.0042027348      1.0000000000
```

```r
# Visualization of the correlation matrix
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
corrplot(cor_matrix,
         method = "color",
         type = "upper",
         tl.col = "black",
         tl.srt = 45,
         addCoef.col = "black",
         number.cex = 0.7,
         tl.cex = 0.7,
         col = colorRampPalette(c("skyblue", "white", "tomato"))(200),
         diag = FALSE)
```



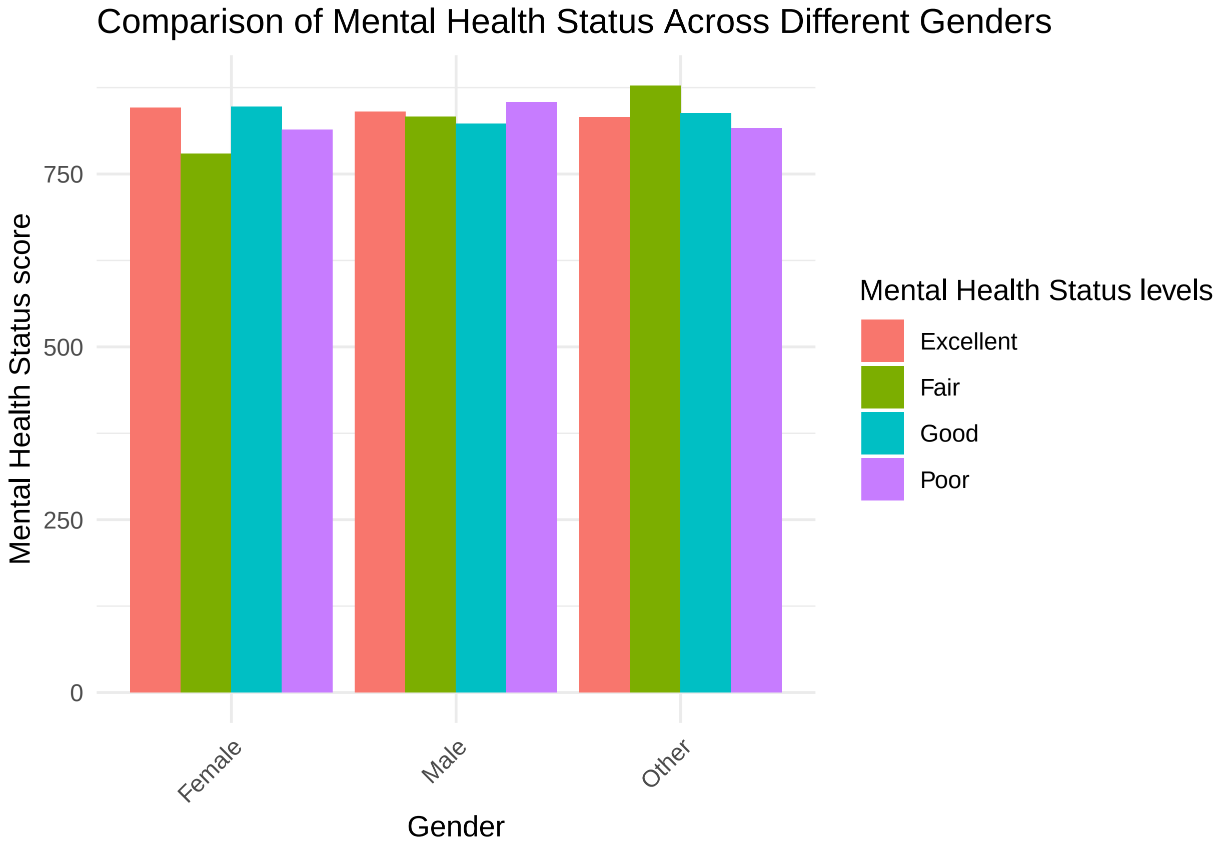
Insights:

- Technology usage have strong correlation with Tech Addiction Risk and moderately with Social Media Usage.
- Age shows a strong negative correlation in Youth Category.
- Physical Activity is strongly linked to an Active Lifestyle. Most other correlations are weak or negligible.

4.3. Mental health by demographics

4.3.1. Mental Health Score by Gender

```
Dataset %>%
  ggplot(aes(x = Gender, fill = Mental_Health_Status)) +
  geom_bar(position = "dodge") +
  theme_minimal() +
  labs(title = "Comparison of Mental Health Status Across Different Genders",
       x = "Gender",
       y = "Mental Health Status score",
       fill = "Mental Health Status levels") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Insights:

- Across all genders, the Mental Health Status scores are relatively similar. `Other` gender seem to have a slightly higher `Fair` mental health score compared to `Female` and `Male`.
- `Excellent` mental health scores are highest among Females, but the difference is small.
- `Poor` mental health status appears fairly consistent across all genders.

5.0. Segmentation Analysis

5.1. Segment by Gender

```
gender_groups <- Dataset %>%
  group_by(Gender) %>%
  summarise(
    Avg_Tech_Usage = mean(Technology_Usage_Hours, na.rm = TRUE),
    Avg_Stress = mean(as.numeric(factor(Stress_Level,
                        levels = c("Low", "Medium", "High"))), na.rm = TRUE),
    Count = n()
  )
print(gender_groups)
```

```
## # A tibble: 3 x 4
##   Gender Avg_Tech_Usage Avg_Stress Count
##   <fct>           <dbl>      <dbl> <int>
## 1 Female           13.0       2.00  3286
## 2 Male             12.9       2.00  3350
## 3 Other            13.0       2     3364
```

5.2. Segment by youth_group

```
youth_groups <- Dataset %>%
  group_by(Youth_Category) %>%
  summarise(
    Avg_Screen_Sleep_Ratio = mean(Screen_to_Sleep_Ratio, na.rm = TRUE),
    Avg_Stress = mean(as.numeric(factor(Stress_Level,
                        levels = c("Low", "Medium", "High"))), na.rm = TRUE),
    Count = n()
  )
print(youth_groups)
```

```
## # A tibble: 2 x 4
##   Youth_Category Avg_Screen_Sleep_Ratio Avg_Stress Count
##            <dbl>                  <dbl>      <dbl> <int>
## 1              0                   1.29       1.99  8306
## 2              1                   1.30       2.02  1694
```

5.3. Segment by Age Group

```
age_groups <- Dataset %>%
  group_by(Age_Group) %>%
  summarise(
    Avg_Tech_Usage = mean(Technology_Usage_Hours, na.rm = TRUE),
    Avg_Sleep = mean(Sleep_Hours, na.rm = TRUE),
    Count = n()
  )
print(age_groups)
```

```
## # A tibble: 5 x 4
##   Age_Group    Avg_Tech_Usage Avg_Sleep Count
##   <fct>                 <dbl>     <dbl> <int>
## 1 Teenager               13.2      6.53   220
## 2 Youth                  13.0      6.49  1474
## 3 Young Adult            12.8      6.53  2056
## 4 Adult                  12.9      6.49  3120
## 5 Senior Adult           13.1      6.49  3130
```

6.0. Statistical Tests

6.1. Chi-Squared Tests: Group Comparison for Categorical Variables

```
# i. Gender vs Mental Health Status
chisq.test(table(Dataset$Gender, Dataset$Mental_Health_Status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Dataset$Gender, Dataset$Mental_Health_Status)
## X-squared = 6.5912, df = 6, p-value = 0.3603
```

```
# ii. Support Systems Access vs Mental Health Status
chisq.test(table(Dataset$Support_Systems_Access, Dataset$Mental_Health_Status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Dataset$Support_Systems_Access, Dataset$Mental_Health_Status)
## X-squared = 8.3275, df = 3, p-value = 0.03971
```

```
# iii. Work Environment Impact vs Mental Health Status
chisq.test(table(Dataset$Work_Environment_Impact, Dataset$Mental_Health_Status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Dataset$Work_Environment_Impact, Dataset$Mental_Health_Status)
## X-squared = 7.3396, df = 6, p-value = 0.2906
```

```
# iv. Online Support Usage vs Mental Health Status
chisq.test(table(Dataset$Online_Support_Usage, Dataset$Mental_Health_Status))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(Dataset$Online_Support_Usage, Dataset$Mental_Health_Status)
## X-squared = 0.59972, df = 3, p-value = 0.8965
```

Chi-square Results interpretation

i. Gender vs. Mental Health Status

X-squared = 6.5912, df = 6, p-value = 0.3603

P > 0.05 suggesting there is no statistically significant association between Gender and Mental Health Status.

ii. Support Systems Access vs. Mental Health Status

X-squared = 8.3275, df = 3, p-value = 0.03971

$p < 0.05$ suggesting there is a statistically significant association between access to support systems and mental health status . Individuals with or without support systems report mental health differently.

iii. Work Environment Impact vs. Mental Health Status

X-squared = 7.3396, df = 6, p-value = 0.2906

No significant association was found between the perceived impact of work environment and mental health status since $p > 0.05$.

iv. Online Support Usage vs. Mental Health Status

X-squared = 0.59972, df = 3, p-value = 0.8965

There is no significant relationship between the use of online support platforms and mental health status $p > 0.05$.

6.2. Normality Testing and Group Comparison Analysis

```
# Picking a Randomly sample 5000 rows
set.seed(123)  # for reproducibility
sample_data <- sample(Dataset$Technology_Usage_Hours, 5000)

# testing for normality
shapiro.test(sample_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample_data
## W = 0.99376, p-value = 5.819e-14
```

```
# If normal:
anova_result <- aov(Technology_Usage_Hours ~ Mental_Health_Status, data = Dataset)
summary(anova_result)
```

```
##                        Df Sum Sq Mean Sq F value Pr(>F)
## Mental_Health_Status    3     53   17.83    0.99  0.396
## Residuals            9996 179927   18.00
```

```
# If NOT normal:
kruskal.test(Technology_Usage_Hours ~ Mental_Health_Status, data = Dataset)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Technology_Usage_Hours by Mental_Health_Status
## Kruskal-Wallis chi-squared = 4.6419, df = 3, p-value = 0.2
```

Results interpretation

W = 0.98787 p-value = 5.819e-14

- Since the p-value is less than 0.05, the null hypothesis of normality is rejected. This suggests the sample data is not normally distributed.
- *One-way ANOVA:* Shows no significant difference in the mean of technology usage hours across mental health status groups. F(3, 9996) = 0.99, p = 0.396. This indicates that the mental health is not influenced by technology usage in this sample.
- *Kruskal-Wallis Rank Sum Test*: p > 0.05 which shows no significance different in the technology usage hours across mental health status groups.

7.0. Predictive Modelling & Analysis

7.1. Linear Regression Model for predicting stress level

```
# Converting Mental_Health_Status and Stress_Level to numeric
Dataset <- Dataset %>%
  mutate(
    Mental_Health_Status_Num = as.numeric(as.factor(Mental_Health_Status)),
    Stress_Level_Num = as.numeric(as.factor(Stress_Level))
  )

# Selecting relevant numeric variables for regression
regression_data <- Dataset %>%
  dplyr::select(Technology_Usage_Hours, Social_Media_Usage_Hours, Gaming_Hours,
                Screen_Time_Hours, Sleep_Hours, Physical_Activity_Hours,
                Stress_Level_Num, Mental_Health_Status_Num)
# model
linear_model <- lm(Stress_Level_Num ~ Technology_Usage_Hours + Social_Media_Usage_Hours +
                   Gaming_Hours + Screen_Time_Hours + Sleep_Hours + Physical_Activity_Hours,
                   data = regression_data)

# summary of the model
summary(linear_model)
```

```
##
## Call:
## lm(formula = Stress_Level_Num ~ Technology_Usage_Hours + Social_Media_Usage_Hours +
##     Gaming_Hours + Screen_Time_Hours + Sleep_Hours + Physical_Activity_Hours,
##     data = regression_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04000 -0.99117 -0.00118  0.98998  1.03963
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.0658584  0.0502516  41.110   <2e-16 ***
## Technology_Usage_Hours   -0.0010068  0.0025788  -0.390    0.696
## Social_Media_Usage_Hours -0.0021480  0.0044208  -0.486    0.627
## Gaming_Hours             -0.0004883  0.0062416  -0.078    0.938
## Screen_Time_Hours         0.0002771  0.0020217   0.137    0.891
## Sleep_Hours               -0.0051735  0.0056307  -0.919    0.358
## Physical_Activity_Hours   -0.0021634  0.0028133  -0.769    0.442
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.8168 on 9993 degrees of freedom
## Multiple R-squared:  0.0002498,  Adjusted R-squared:  -0.0003504
## F-statistic: 0.4162 on 6 and 9993 DF,  p-value: 0.8688
```

Interpretation

- P-values are all greater than 0.05 suggesting that none of the predictors are statistically significant.
- Changes in Technology, Social Media, Gaming, Screen Time, Sleep, or Physical Activity do not significantly predict Stress Level in the dataset.

7.2. Ordinal Logistic Regression Model for Predicting Mental Health Status

```r
# making Mental_Health_Status_Num as an ordered factor
regression_data <- regression_data %>%
  mutate(Mental_Health_Status_Num = factor(Mental_Health_Status_Num, ordered = TRUE))

# Ordinal logistic regression
ordinal_model <- polr(Mental_Health_Status_Num ~ Technology_Usage_Hours + Social_Media_Usage_Hours +
                      Gaming_Hours + Screen_Time_Hours + Sleep_Hours + Physical_Activity_Hours,
                      data = regression_data, method = "logistic")

# summary of the model
summary(ordinal_model)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = Mental_Health_Status_Num ~ Technology_Usage_Hours +
##     Social_Media_Usage_Hours + Gaming_Hours + Screen_Time_Hours +
##     Sleep_Hours + Physical_Activity_Hours, data = regression_data,
##     method = "logistic")
##
## Coefficients:
##                              Value Std. Error t value
## Technology_Usage_Hours    0.011512   0.005641  2.0408
## Social_Media_Usage_Hours -0.010904   0.009661 -1.1287
## Gaming_Hours             -0.017197   0.013675 -1.2576
## Screen_Time_Hours         0.001969   0.004426  0.4450
## Sleep_Hours               0.009541   0.012361  0.7718
## Physical_Activity_Hours  -0.002981   0.006143 -0.4852
##
## Intercepts:
##     Value   Std. Error t value
## 1|2 -0.9641  0.1106    -8.7147
## 2|3  0.1286  0.1102     1.1676
## 3|4  1.2331  0.1109    11.1175
##
## Residual Deviance: 27720.25
## AIC: 27738.25
```

Interpretation

- Only Technology_Usage_Hours is statistically significant (t = 2.04), meaning it has a small but positive effect on Mental Health Status.
- All other variables — Social_Media_Usage_Hours, Gaming_Hours, Screen_Time_Hours, Sleep_Hours, and Physical_Activity_Hours — are not statistically significant predictors.

7.3. Random Forest

```
#splitting the dataset
library(caret)
set.seed(123)
train_index <- createDataPartition(Dataset$Mental_Health_Status, p = 0.7, list = FALSE)
train_data <- Dataset[train_index, ]
test_data <- Dataset[-train_index, ]

#random Forest model

rf_model <- randomForest(Mental_Health_Status ~ Technology_Usage_Hours + Stress_Level + Gender +
                         Age + Support_Systems_Access + Online_Support_Usage + Work_Environment_Impact,
                         data = train_data, importance = TRUE, ntree = 100)

# Prediction
rf_preds <- predict(rf_model, test_data)

# Confusion matrix
confusionMatrix(rf_preds, test_data$Mental_Health_Status)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Excellent Fair Good Poor
##   Excellent       187  218  230  219
##   Fair            226  201  176  181
##   Good            147  119  143  143
##   Poor            195  209  203  202
##
## Overall Statistics
##
##                Accuracy : 0.2444
##                  95% CI : (0.2291, 0.2602)
##     No Information Rate : 0.2518
##     P-Value [Acc > NIR] : 0.828
##
##                   Kappa : -0.0074
```

```
##
##  Mcnemar's Test P-Value : 1.047e-07
##
## Statistics by Class:
##
##                      Class: Excellent Class: Fair Class: Good Class: Poor
## Sensitivity                   0.24768     0.26908     0.19016     0.27114
## Specificity                   0.70276     0.74112     0.81798     0.73070
## Pos Pred Value                0.21897     0.25638     0.25906     0.24969
## Neg Pred Value                0.73520     0.75350     0.75112     0.75205
## Prevalence                    0.25175     0.24908     0.25075     0.24842
## Detection Rate                0.06235     0.06702     0.04768     0.06736
## Detection Prevalence          0.28476     0.26142     0.18406     0.26976
## Balanced Accuracy             0.47522     0.50510     0.50407     0.50092
```

Interpretation:

- The Random Forest model performed poorly, with an accuracy of only 24%, lower than random guessing (which would be around 25% given four classes).
- The negative Kappa indicates that the model is not better than random at classifying mental health categories based on the predictors.
- Sensitivity and specificity values vary between 19% to 27%, showing weak ability to correctly identify any particular mental health class.

7.4. Multinomial Logistic Regression

```
library(nnet)
# Fitting the model
multi_log_model <- multinom(Mental_Health_Status ~ Technology_Usage_Hours + Stress_Level + Gender +
                            Age + Support_Systems_Access + Online_Support_Usage + Work_Environment_Impac
                            data = train_data)
```

```
## # weights:  48 (33 variable)
## initial  value 9705.446822
## iter  10 value 9694.932977
## iter  20 value 9692.714972
## iter  30 value 9691.794444
## final  value 9691.707460
## converged
```

```
# Predict
log_preds <- predict(multi_log_model, test_data)

# Evaluating the model
confusionMatrix(log_preds, test_data$Mental_Health_Status)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Excellent Fair Good Poor
##   Excellent       193  208  208  196
##   Fair            217  232  210  197
##   Good            154  129  156  168
##   Poor            191  178  178  184
##
## Overall Statistics
##
##                Accuracy : 0.2551
##                  95% CI : (0.2396, 0.2711)
##     No Information Rate : 0.2518
##     P-Value [Acc > NIR] : 0.3436
##
##                   Kappa : 0.0068
##
##  Mcnemar's Test P-Value : 6.31e-05
##
## Statistics by Class:
##
##                      Class: Excellent Class: Fair Class: Good Class: Poor
## Sensitivity                   0.25563     0.31058     0.20745     0.24698
## Specificity                   0.72727     0.72291     0.79929     0.75732
## Pos Pred Value                0.23975     0.27103     0.25700     0.25171
## Neg Pred Value                0.74385     0.75968     0.75084     0.75265
## Prevalence                    0.25175     0.24908     0.25075     0.24842
## Detection Rate                0.06435     0.07736     0.05202     0.06135
## Detection Prevalence          0.26842     0.28543     0.20240     0.24375
## Balanced Accuracy             0.49145     0.51674     0.50337     0.50215
```

Interpretation:

- Like Random Forest, the Multinomial Regression model performed poorly with accuracy very close to random chance.
- The Kappa statistic is very close to zero, meaning the model's ability to predict correctly is essentially random.
- Sensitivity and specificity for each class are low, indicating weak discriminatory power.

These results indicate that the model performs no better than random guessing.

Possible Reasons for Poor Performance

- **Feature Limitation**: Mental health is a complex outcome influenced by a wide range of behavioral, social, and environmental factors. The dataset primarily includes basic demographic data and technology usage hours, which may not be sufficient to capture these complexities.
- **Overlapping Class Labels**: The mental health categories (*Excellent*, *Good*, *Fair*, *Poor*) may not be distinct enough in terms of the predictors, leading to overlapping patterns that hinder accurate classification.
- **Noisy or Sparse Data**: Some features may have missing values, inconsistencies, or low variance, which could affect model training and generalization.

Decision and Justification Since the results of the predictive modelling seems meaningless, I will skip the model evaluation part. In this project I will shift from predictive modeling to **insight-based analysis**.

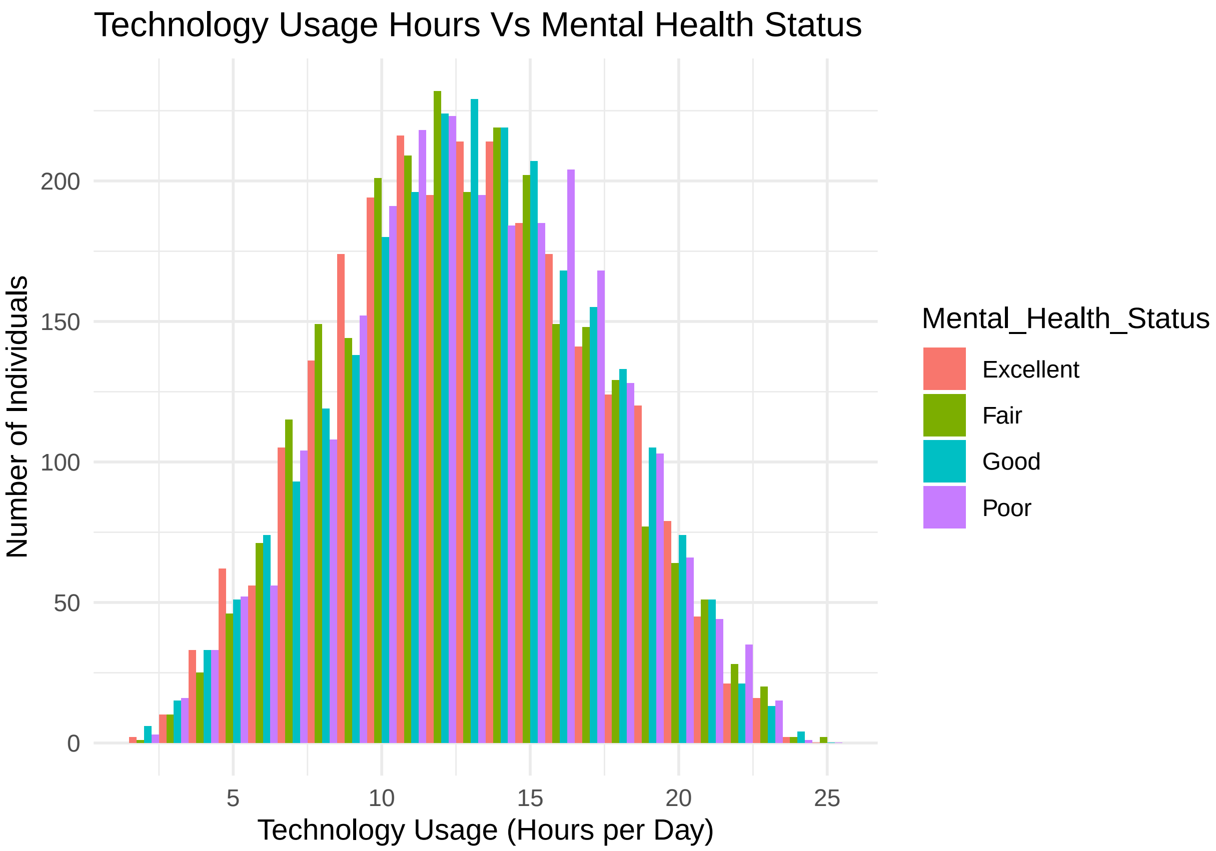
The new focus will be on: -

- Exploring **relationships** between technology usage and mental health status.
- Using **visualization** to highlight meaningful trends.
- Discussing **limitations** and **recommendations** for future studies, including the importance of incorporating additional behavioral and psychological factors.

8.0 Visualization

8.1. Technology Usage vs Mental Health

```
ggplot(Dataset, aes(x = Technology_Usage_Hours, fill = Mental_Health_Status)) +
  geom_histogram(binwidth = 1, position = "dodge") +
  labs(title = "Technology Usage Hours Vs Mental Health Status",
       x = "Technology Usage (Hours per Day)",
       y = "Number of Individuals") +
  theme_minimal()
```

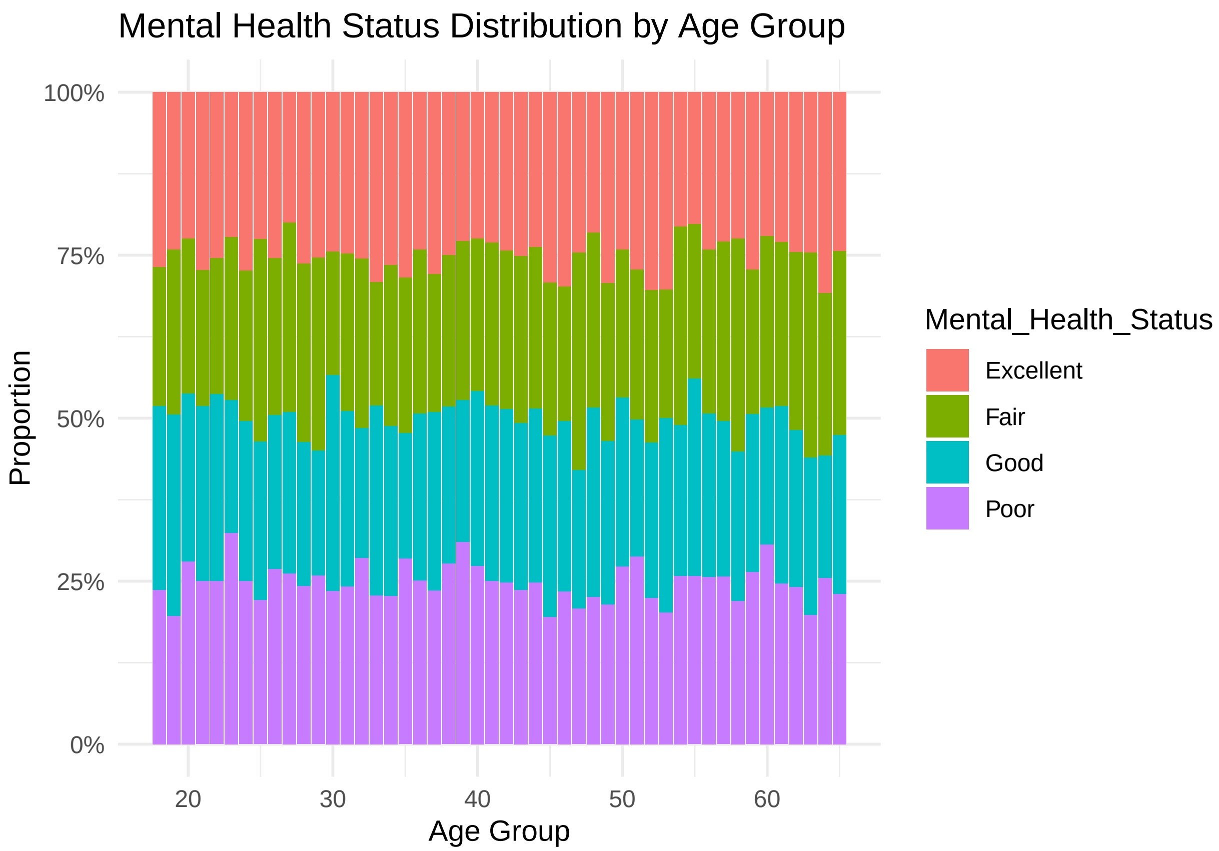


Observations:

- Across most technology usage levels (from 1 to 10+ hours), the mental health statuses are fairly evenly distributed.
- There's a slight decline in "Excellent" mental health as tech usage increases, especially around the higher usage end.
- "Poor" mental health increases slightly in higher usage, suggesting possible negative correlation between excessive tech use and mental well-being.
- Usage around 5 to 9 hours appears to have the highest number of individuals, across all mental health categories.

8.2. Mental Health by Age Groups

```r
ggplot(Dataset, aes(x = Age, fill = Mental_Health_Status)) +
  geom_bar(position = "fill") +
  labs(title = "Mental Health Status Distribution by Age Group",
       x = "Age Group",
       y = "Proportion") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()
```



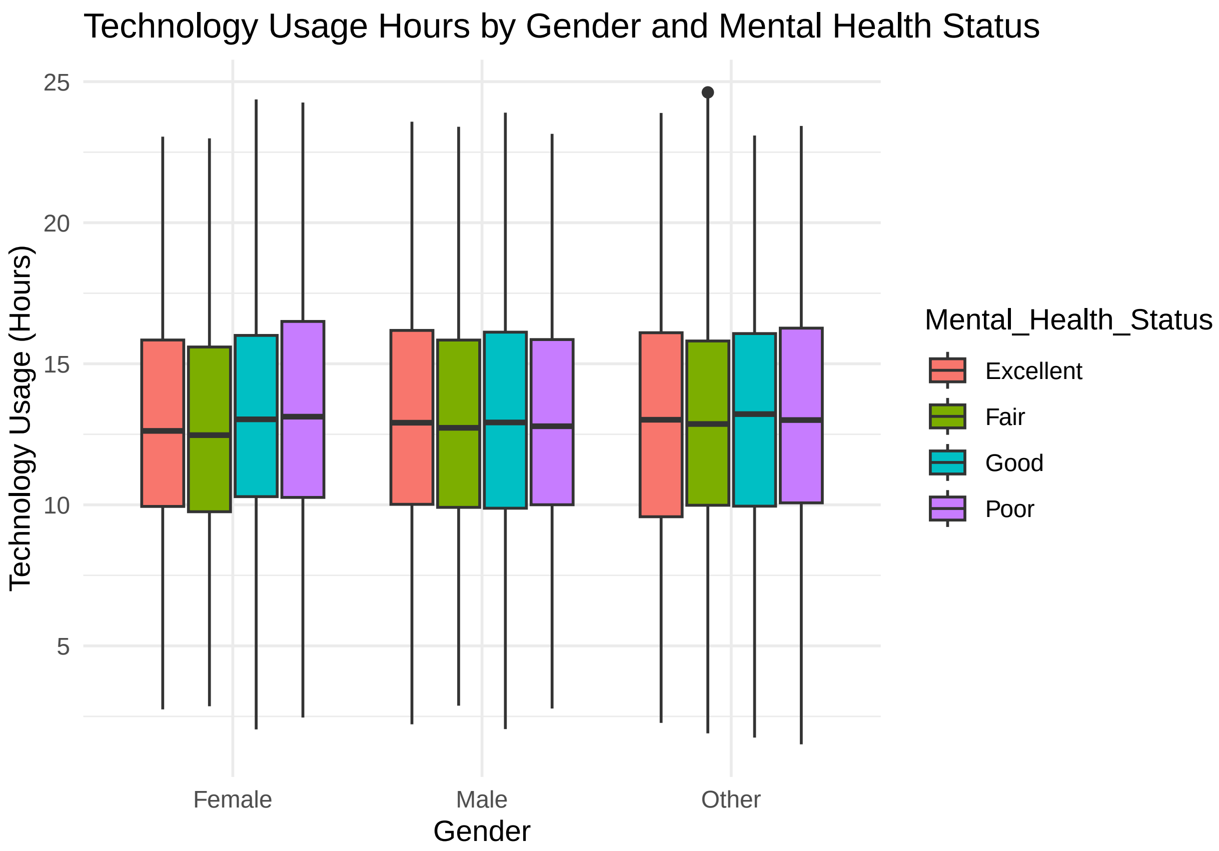
From the bar plot;

- Each age group shows roughly equal proportions of each mental health status. This suggests that mental health status is fairly consistent across age groups.

- There's no significant variation across the age groups, this is a clear indication that mental health distribution is not heavily influenced by age in this dataset.

8.3. relationship between technology usage and Mental Health

```r
ggplot(Dataset, aes(x = Gender, y = Technology_Usage_Hours, fill = Mental_Health_Status)) +
  geom_boxplot() +
  labs(title = "Technology Usage Hours by Gender and Mental Health Status",
       x = "Gender", y = "Technology Usage (Hours)") +
  theme_minimal()
```

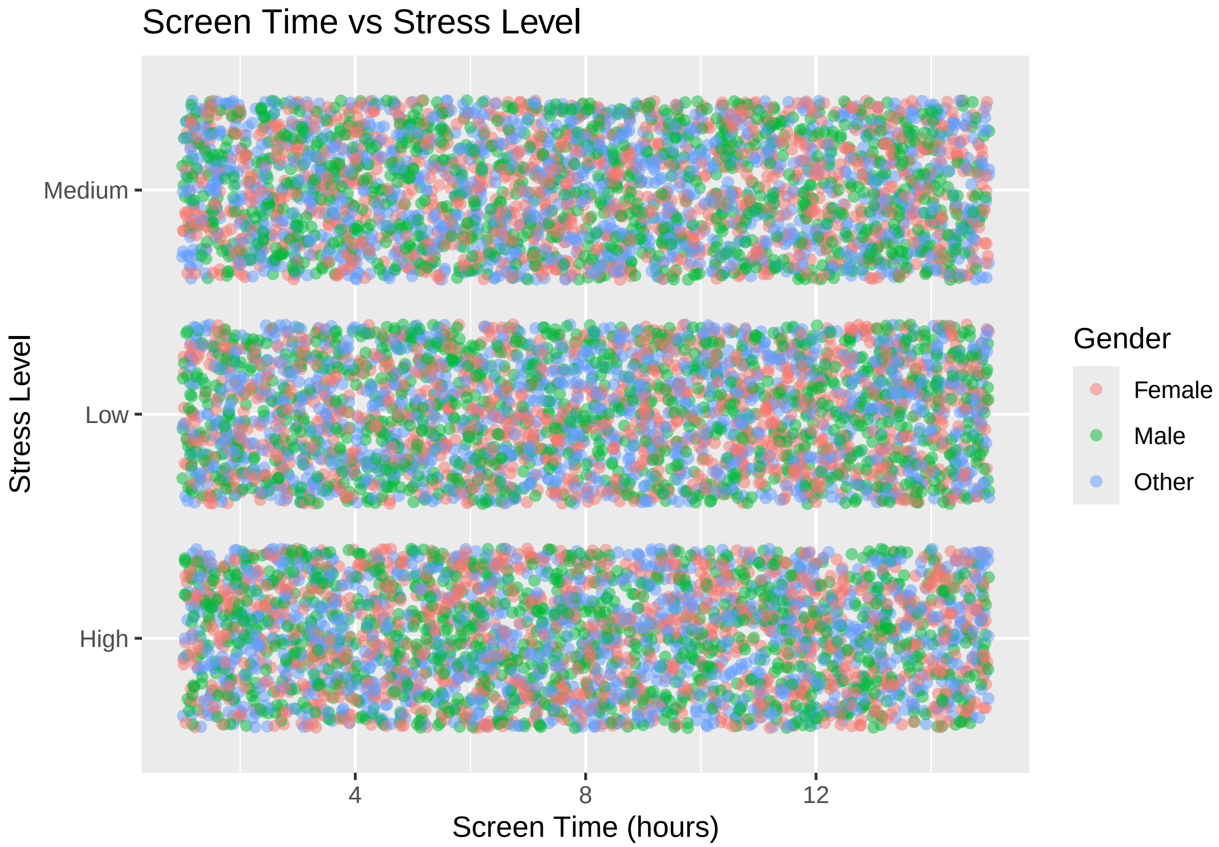


From the boxplot;

- Technology usege is fairly consistent across all genders, with most people using technology between 13 and 15 hours on average.
- Those who reported "Poor" mental health tend to use technology a bit more than others, especially when compared to people with "Excellent" or "Good" mental health.
- There are a few outliers, particularly in the "Other" gender group, where some individuals reported very high usage—over 24 hours—which could indicate unusually heavy or excessive use.
- People with "Excellent" mental health tend to have more consistent and slightly lower tech usage overall, with less variation in their reported hours.

8.4. Relationship between Screen Time and Stress Level

```r
ggplot(Dataset, aes(x = Screen_Time_Hours, y = as.factor(Stress_Level), color = Gender)) +
  geom_jitter(alpha = 0.5) +
  labs(title = "Screen Time vs Stress Level", x = "Screen Time (hours)", y = "Stress Level")
```



From the scatter plot; -

- The distribution of screen time appears fairly uniform across all stress levels, suggesting no clear linear relationship between screen time and stress level.
- The gender distribution is relatively balanced across all stress levels and screen time durations.
- The plot appears to show a dense clustering, indicating a large dataset with a wide spread of screen times within each stress level.

9.0. Discussion

This project investigated the relationship between technology usage and mental health across different demographics. While access to support systems showed a significant association with mental health status, most technology usage metrics (e.g., total technology hours, screen time) had limited predictive power. No major differences were found across gender or age groups.

Predictive modeling attempts, including random forest and multinomial logistic regression, performed poorly, highlighting that mental health outcomes are influenced by broader social and behavioral factors not fully captured in the dataset. Additionally, the cross-sectional nature of the data, lack of control for confounding variables, and absence of longitudinal tracking limit the strength of conclusions.

Ethical considerations, particularly regarding data privacy and responsible use of mental health information, must be emphasized. Future research should explore interaction effects, include more diverse behavioral factors, and apply advanced modeling and validation techniques.

10.0. Conclusion

This study examined the relationship between technology usage and mental health across various demographic groups. While descriptive analysis highlighted some patterns, such as the role of support systems, predictive modeling showed that technology usage alone is a weak predictor of mental health outcomes. The findings emphasize that mental health is influenced by multiple complex factors beyond digital behavior. Future research should incorporate broader social and behavioral variables, adopt longitudinal approaches, and prioritize ethical considerations when handling sensitive data.

11.0. Recommendation

Based on the findings and limitations of this study, the following recommendations are proposed:

- Future studies should incorporate broader behavioral, psychological, and environmental variables to better explain mental health outcomes.
- Longitudinal data collection is recommended to capture changes in technology usage and mental health over time.
- Advanced machine learning techniques with proper feature selection and model tuning should be applied to improve predictive performance.
- Ethical guidelines regarding data privacy, consent, and responsible use of mental health information must be strictly followed.
- Public health interventions should focus not only on reducing technology usage but also on strengthening support systems to promote mental well-being.

12.0. Reference

Kaggle. (2024). Mental Health and Technology Usage 2024 Dataset. Retrieved from https://www.kaggle.com/datasets/waqi786/mental-health-and-technology-usage-dataset