

QDS-Data exploration and cleaning

2025-04-30

Uploading a Dataset in R.

```
#Loading the libraries which we will be using for the data processing.  
library(tidyr)  
library(BayesFactor)
```

```
## Loading required package: coda
```

```
## Warning: package 'coda' was built under R version 4.4.3
```

```
## Loading required package: Matrix
```

```
##
```

```
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
```

```
##
```

```
##      expand, pack, unpack
```

```
## *****
```

```
## Welcome to BayesFactor 0.9.12-4.7. If you have questions, please contact Richard Morey (richarddmorey@ucsd.edu)
```

```
##
```

```
## Type BFManual() to open the manual.
```

```
## *****
```

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
## v dplyr      1.1.4      v purrr      1.0.2
```

```
## v forcats    1.0.0      v readr      2.1.5
```

```
## v ggplot2    3.5.2      v stringr    1.5.1
```

```
## v lubridate  1.9.4      v tibble     3.2.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
```

```
## x Matrix::expand() masks tidyr::expand()
```

```
## x dplyr::filter()  masks stats::filter()
```

```
## x dplyr::lag()     masks stats::lag()
```

```
## x Matrix::pack()   masks tidyr::pack()
```

```
## x Matrix::unpack() masks tidyr::unpack()
```

```
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#Loading the dataset in R.
```

```
mht = read_csv("C:/Users/athar/OneDrive/Desktop/coursework/quantitative reasoning/mht.csv")
```

```
## Rows: 10000 Columns: 14
```

```
## -- Column specification -----
```

```
## Delimiter: ","
```

```
## chr (7): User_ID, Gender, Mental_Health_Status, Stress_Level, Support_System...
```

```
## dbl (7): Age, Technology_Usage_Hours, Social_Media_Usage_Hours, Gaming_Hours...
```

```
##
```

```
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Displaying the first 5 rows to show the data.
```

```
head(mht)
```

```
## # A tibble: 6 x 14
```

```
##   User_ID      Age Gender Technology_Usage_Hours Social_Media_Usage_Hours
```

```
##   <chr>      <dbl> <chr>          <dbl>          <dbl>
```

```
## 1 USER-00001    23 Female          6.57            6
```

```
## 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           1.2            0.56
```

```
## 6 USER-00006    58 Male           5.59           5.74
```

```
## # i 9 more variables: Gaming_Hours <dbl>, Screen_Time_Hours <dbl>,
```

```
## #   Mental_Health_Status <chr>, Stress_Level <chr>, Sleep_Hours <dbl>,
```

```
## #   Physical_Activity_Hours <dbl>, Support_Systems_Access <chr>,
```

```
## #   Work_Environment_Impact <chr>, Online_Support_Usage <chr>
```

```
#exploring the data summary and structure.
```

```
summary(mht)
```

```
##   User_ID      Age      Gender      Technology_Usage_Hours
```

```
## Length:10000   Min.   :18.00   Length:10000   Min.    : 1.000
```

```
## Class :character 1st Qu.:29.00   Class :character 1st Qu.: 3.760
```

```
## Mode :character  Median :42.00   Mode :character  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
```

```
## Length:10000      Length:10000   Min.    :4.000
```

```
## Class :character   Class :character 1st Qu.:5.260
```

```
## Mode :character    Mode :character  Median :6.500
```

```
##                    Mean   :6.501
```

```
##                    3rd Qu.:7.760
```

```
##                               Max.      :9.000
## Physical_Activity_Hours Support_Systems_Access Work_Environment_Impact
## Min.      : 0.000           Length:10000           Length:10000
## 1st Qu.: 2.490             Class :character       Class :character
## Median : 4.990             Mode  :character       Mode  :character
## Mean    : 5.004
## 3rd Qu.: 7.540
## Max.    :10.000
## Online_Support_Usage
## Length:10000
## Class :character
## Mode  :character
##
##
##
```

```
str(mht)
```

```
## spc_tbl_ [10,000 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ User_ID      : chr [1:10000] "USER-00001" "USER-00002" "USER-00003" "USER-00004" ...
## $ Age          : num [1:10000] 23 21 51 25 53 58 63 51 57 31 ...
## $ Gender       : chr [1:10000] "Female" "Male" "Male" "Female" ...
## $ Technology_Usage_Hours : num [1:10000] 6.57 3.01 3.04 3.84 1.2 ...
## $ Social_Media_Usage_Hours: num [1:10000] 6 2.57 6.14 4.48 0.56 5.74 2.55 4.1 4.11 7.23 ...
## $ Gaming_Hours : num [1:10000] 0.68 3.74 1.26 2.59 0.29 0.11 3.79 4.74 0.08 0.81 ...
## $ Screen_Time_Hours : num [1:10000] 12.36 7.61 3.16 13.08 12.63 ...
## $ Mental_Health_Status : chr [1:10000] "Good" "Poor" "Fair" "Excellent" ...
## $ Stress_Level : chr [1:10000] "Low" "High" "High" "Medium" ...
## $ Sleep_Hours : num [1:10000] 8.01 7.28 8.04 5.62 5.55 8.61 8.61 7.11 7.19 5.09 ...
## $ Physical_Activity_Hours : num [1:10000] 6.71 5.88 9.81 5.28 4 6.54 1.34 5.27 5.22 0.47 ...
## $ Support_Systems_Access : chr [1:10000] "No" "Yes" "No" "Yes" ...
## $ Work_Environment_Impact : chr [1:10000] "Negative" "Positive" "Negative" "Negative" ...
## $ Online_Support_Usage : chr [1:10000] "Yes" "No" "No" "Yes" ...
## - attr(*, "spec")=
## .. cols(
## ..   User_ID = col_character(),
## ..   Age = col_double(),
## ..   Gender = col_character(),
## ..   Technology_Usage_Hours = col_double(),
## ..   Social_Media_Usage_Hours = col_double(),
## ..   Gaming_Hours = col_double(),
## ..   Screen_Time_Hours = col_double(),
## ..   Mental_Health_Status = col_character(),
## ..   Stress_Level = col_character(),
## ..   Sleep_Hours = col_double(),
## ..   Physical_Activity_Hours = col_double(),
## ..   Support_Systems_Access = col_character(),
## ..   Work_Environment_Impact = col_character(),
## ..   Online_Support_Usage = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
# Converting character columns to factors so it will be easy for further analysis .
```

```
mht1 <- mht %>%  
  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)  
  )
```

```
#Checking for any missing values and resolving them.
```

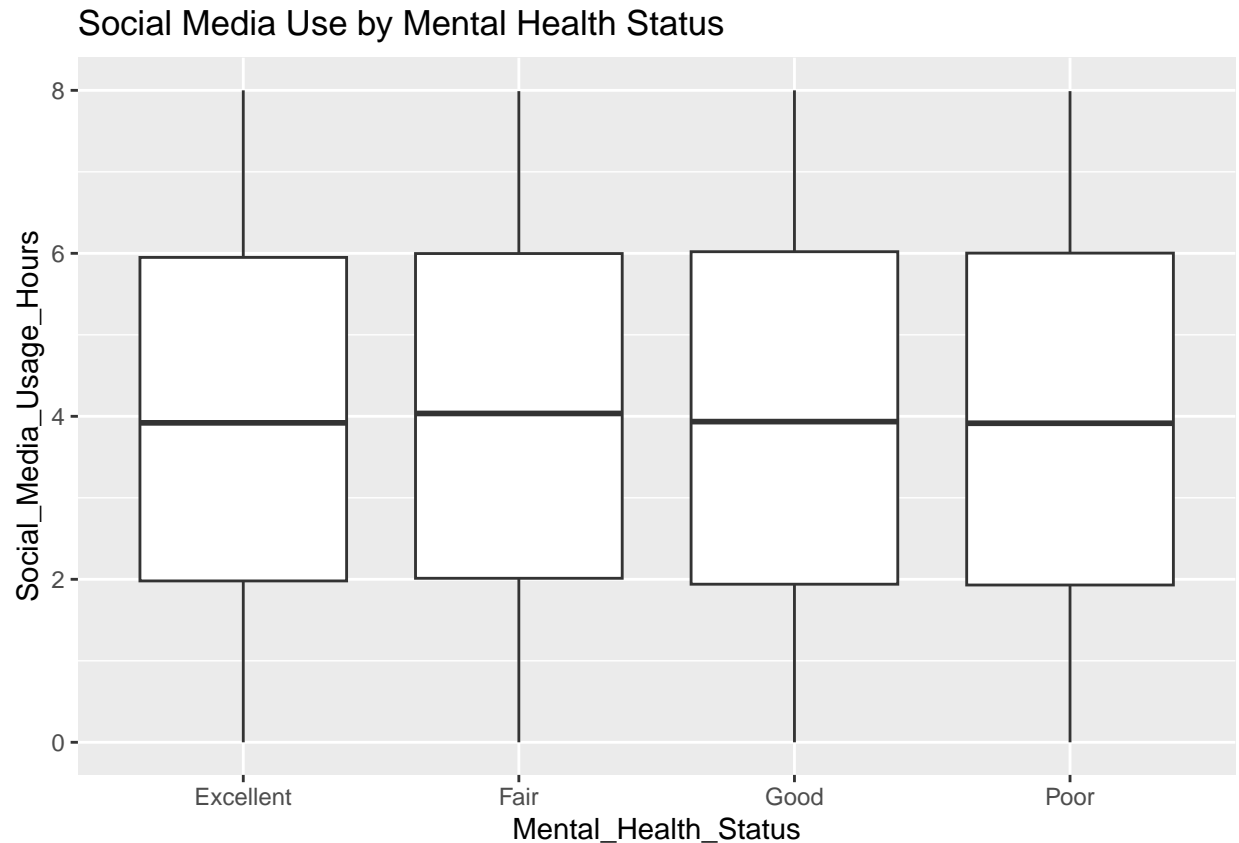
```
colSums(is.na(mht1))
```

```
##           User_ID           Age           Gender  
##           0           0           0  
## Technology_Usage_Hours Social_Media_Usage_Hours           Gaming_Hours  
##           0           0           0  
##           Screen_Time_Hours           Mental_Health_Status           Stress_Level  
##           0           0           0  
##           Sleep_Hours           Physical_Activity_Hours           Support_Systems_Access  
##           0           0           0  
## Work_Environment_Impact           Online_Support_Usage  
##           0           0
```

The Research Question according to the dataset and the Frequentist and Bayesian inferential statistics performed and models used to answer those questions

```
#Does more social media use relate to poorer mental health?
```

```
ggplot(mht1, aes(x = Mental_Health_Status, y = Social_Media_Usage_Hours)) +  
  geom_boxplot() +  
  labs(title = "Social Media Use by Mental Health Status")
```



#The following diagram Shows us social media usage in hours across diff self reported menatal health statuses uch as excellent, good, fair,poor. We can see the plot reflects roughly same use of social media across all the groups. The spread is similar across categories, suggesting no strong difference in usage on mental health.This implies theaet social media alone is not a clear predictor of reported mental health.

#Does more social media use relate to poorer mental health?

#frequentist anova:

```
anova_result <- aov(Social_Media_Usage_Hours ~ Mental_Health_Status, data = mht1)
summary(anova_result)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Mental_Health_Status  3      13    4.330   0.809  0.489
## Residuals          9996  53514    5.354
```

#Bayesian anova

```
anovaBF(Social_Media_Usage_Hours ~ Mental_Health_Status, data = mht1)
```

```
## Warning: data coerced from tibble to data frame
```

```
## Bayes factor analysis
```

```
## -----
```

```
## [1] Mental_Health_Status : 0.0003378527 ±0.07%
```

```
##
```

```
## Against denominator:
```

```
## Intercept only
## ---
## Bayes factor type: BFlinearModel, JZS
```

#Frequentist anova : The p-value > 0.05, this means the difference in average social media usage between mental health groups is not statistically significant, So we can say we cannot conclude that mental health status impacts social media usage.

#Bayesian anova: The bayesfactor = 0.00034 #This is extremely low, meaning the data strongly support the null model, It shows an overwhelming evidence against an effect of mental health status on social media use.

#We can conclude that mental health is not associated with the amount of time someone spends on social media.

#using multinomial logistic regression model for the same question.

```
library(nnet)
multi_logit <- multinom(Mental_Health_Status ~ Social_Media_Usage_Hours, data = mht1)
```

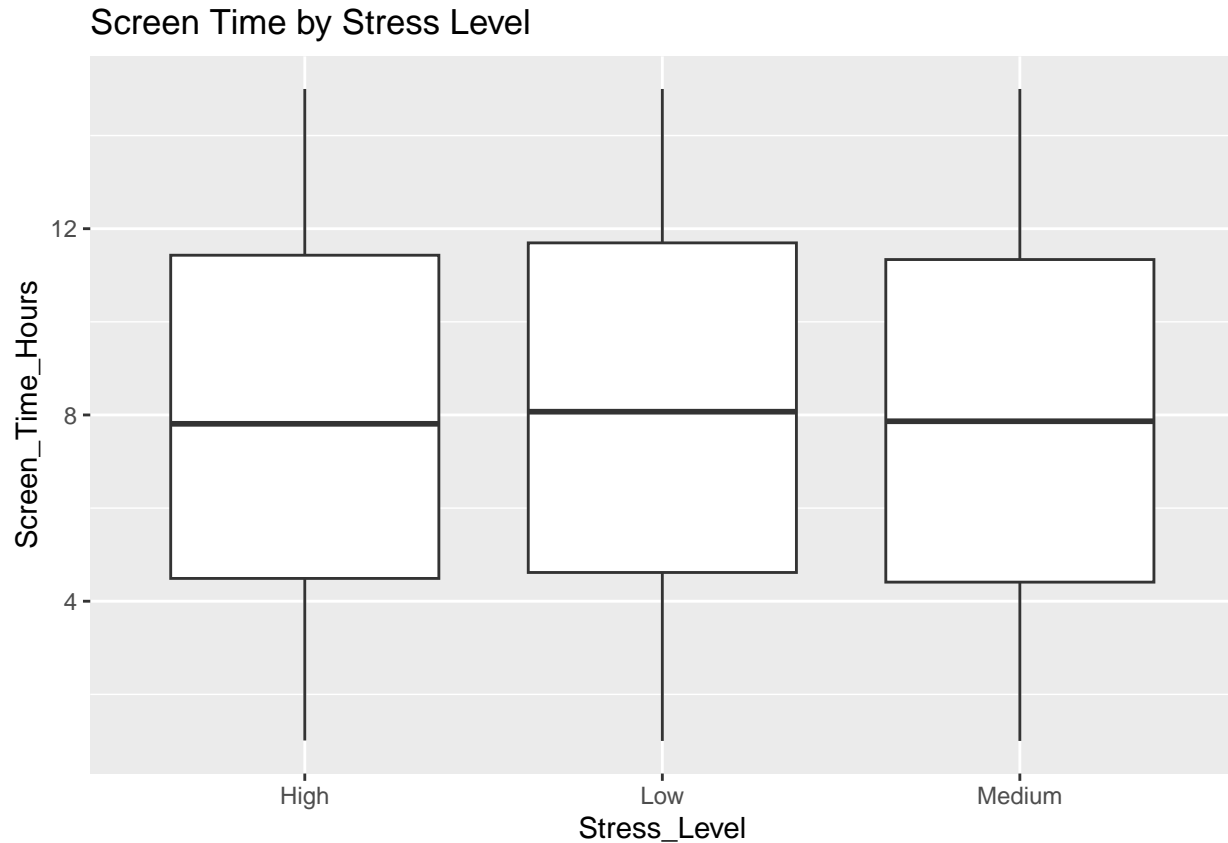
```
## # weights: 12 (6 variable)
## initial value 13862.943611
## final value 13861.581265
## converged
```

```
summary(multi_logit)
```

```
## Call:
## multinom(formula = Mental_Health_Status ~ Social_Media_Usage_Hours,
## data = mht1)
##
## Coefficients:
## (Intercept) Social_Media_Usage_Hours
## Fair -0.08133800 0.017642276
## Good -0.04625317 0.010680765
## Poor -0.02874385 0.003845286
##
## Std. Errors:
## (Intercept) Social_Media_Usage_Hours
## Fair 0.05621596 0.01221871
## Good 0.05591088 0.01219569
## Poor 0.05585334 0.01222547
##
## Residual Deviance: 27723.16
## AIC: 27735.16
```

#The coefficients for Social_Media_Usage_Hours are positive across all mental health categories (Fair, Good, Poor), suggesting that higher social media usage is very slightly associated with increased odds of being in a lower mental health category (vs. Excellent). All the effects are very small and no coefficients are statistically significant. While the direction of the relationship suggests that more usage may slightly increase the odds of reporting poorer mental health (compared to “Excellent”), the effect sizes are small, and likely not statistically meaningful.

```
#Is screen time related to stress level?
ggplot(mht1, aes(x = Stress_Level, y = Screen_Time_Hours)) +
  geom_boxplot() +
  labs(title = "Screen Time by Stress Level")
```



#The following plot compares total daily screen time across three levels of reported stress such as low, medium, high. The median screen time is mostly same across all the stress levels, However higher stress level shows a slightly wider range, indicating more variability in screen time among stressed individuals. This plot suggest that even though screen time may not dramatically differ with stress level, some high-stress user may be outliers with extreme screen time.

```
#Is screen Time related to stress level ?

#frequentist anova:

anova_screen <- aov(Screen_Time_Hours ~ Stress_Level, data = mht1)
summary(anova_screen)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## Stress_Level    2     53   26.30   1.609    0.2
## Residuals  9997 163358   16.34
```

```
#Bayesian anova :
anovaBF(Screen_Time_Hours ~ Stress_Level, data = mht1)
```

```
## Warning: data coerced from tibble to data frame
```

```
## Bayes factor analysis
```

```
## -----
```

```
## [1] Stress_Level : 0.005943686 ±0.06%
```

```
##
```

```
## Against denominator:
```

```
## Intercept only
```

```
## ---
```

```
## Bayes factor type: BFlinearModel, JZS
```

#Frequentist anova : A one-way ANOVA was conducted to examine whether screen time significantly differed across stress levels (Low, Medium, High). The results showed no statistically significant difference. Since the p-value is well above the conventional 0.05 threshold, we fail to reject the null hypothesis. This suggests that, in this dataset, self-reported stress level is not significantly associated with differences in average screen time.

#Bayesian anova : A Bayesian ANOVA was performed to compare models predicting screen time based on stress level versus an intercept-only (null) model. The resulting Bayes Factor (BF) was 0.0059, which provides strong evidence in favor of the null model. This means the data are nearly 170 times more likely under the assumption that screen time is not influenced by stress level, reinforcing the conclusion that no meaningful relationship exists between these variables in the current dataset.

```
#Using ordinal logistic model:
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.4.3
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
## select
```

```
ord_model <- polr(Stress_Level ~ Screen_Time_Hours, data = mht1, Hess = TRUE)
summary(ord_model)
```

```
## Call:
```

```
## polr(formula = Stress_Level ~ Screen_Time_Hours, data = mht1,
```

```
## Hess = TRUE)
```

```
##
```

```
## Coefficients:
```

```
## Value Std. Error t value
```

```
## Screen_Time_Hours 0.0005873 0.004534 0.1295
```

```
##
```

```
## Intercepts:
```

```
## Value Std. Error t value
```

```
## High|Low -0.6899 0.0420 -16.4201
```

```
## Low|Medium 0.6957 0.0420 16.5540
```

```
##
```

```
## Residual Deviance: 21972.22
```

```
## AIC: 21978.22
```

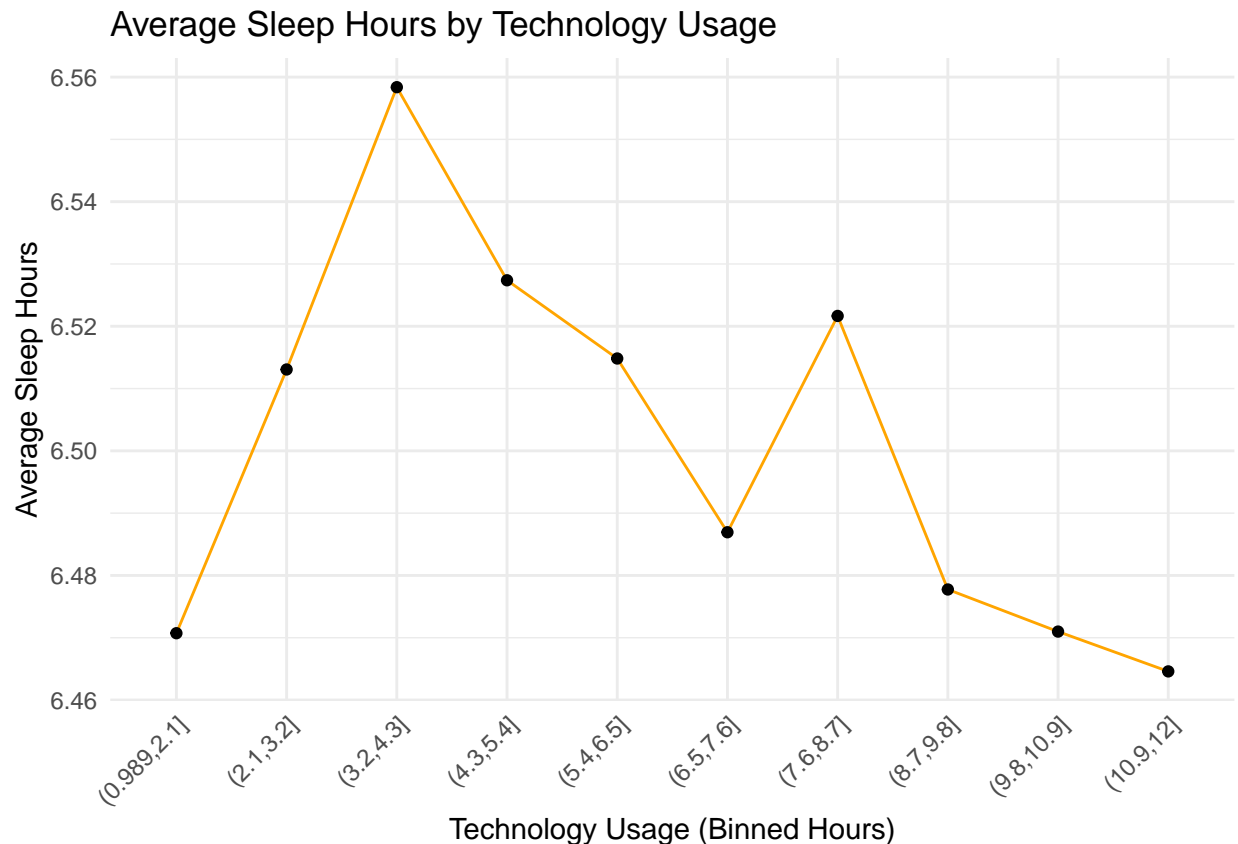
#We get an estimate of 0.0005873, std.error of 0.004534 and the t-value of 0.1295, The coefficient is very close to 0 and not statistically significant, This means that the screen time does not predict changes in stress level in this model, a one hour increase in screen time results in only a 0.06% increase in the odds of being in a higher stress category. This estimates that screen time is not a good predictor of whether someone experiences low, medium or high stress in the dataset.

#Is there a relationship between technology usage hours and sleep duration??

```
mht1 <- mht1 %>%
  mutate(Tech_Usage_Bin = cut(Technology_Usage_Hours, breaks = 10))

avg_sleep_by_tech <- mht1 %>%
  group_by(Tech_Usage_Bin) %>%
  summarise(Avg_Sleep = mean(Sleep_Hours, na.rm = TRUE))

ggplot(avg_sleep_by_tech, aes(x = Tech_Usage_Bin, y = Avg_Sleep, group = 1)) +
  geom_line(color = "orange") +
  geom_point() +
  labs(title = "Average Sleep Hours by Technology Usage",
       x = "Technology Usage (Binned Hours)",
       y = "Average Sleep Hours") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



#The line chart shows the relationship between binned technology usage hours and avg sleep duration. The plot shows that individuals who use fewer hours of technology tend to sleep more than the one who uses

high hours of technology. Although the differences are small, this suggests that higher technology use may be associated with reduced sleep, potentially due to time displacement or mental overstimulation before bedtime.

#Is there a relationship between technology usage hours and sleep duration?

#Frequentist anova:

```
anova_model <- aov(Sleep_Hours ~ Tech_Usage_Bin, data = mht1)
summary(anova_model)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Tech_Usage_Bin    9      8  0.9423   0.447   0.91
## Residuals      9990  21041  2.1063
```

#Bayesian anova:

```
anovaBF(Sleep_Hours ~ Tech_Usage_Bin, data = mht1)
```

```
## Warning: data coerced from tibble to data frame
```

```
## Bayes factor analysis
```

```
## -----
```

```
## [1] Tech_Usage_Bin : 2.908845e-08 ±0%
```

```
##
```

```
## Against denominator:
```

```
##   Intercept only
```

```
## ---
```

```
## Bayes factor type: BFlinearModel, JZS
```

#Frequentist ANOVA: A one-way ANOVA was conducted to examine whether average sleep duration differs across levels of binned technology usage. The analysis showed no statistically significant difference. This high p-value suggests that individuals in different technology usage categories (from lowest to highest usage) do not show meaningful differences in their average sleep hours. Thus, we fail to reject the null hypothesis and conclude that technology usage is not significantly associated with sleep duration based on this grouped comparison.

#Bayesian ANOVA Interpretation: The Bayesian ANOVA yielded a Bayes Factor (BF) of approximately 2.9e-08, indicating overwhelming support for the null model over the alternative. In Bayesian terms, the data provide extremely strong evidence against any differences in sleep duration across technology usage levels. This reinforces the frequentist conclusion that technology usage does not appear to impact sleep hours in this sample when grouped into categories.

#Linear model for Is there a relationship between technology usage hours and sleep duration?

```
lm_model <- lm(Sleep_Hours ~ Technology_Usage_Hours, data = mht1)
summary(lm_model)
```

```
##
```

```
## Call:
```

```
## lm(formula = Sleep_Hours ~ Technology_Usage_Hours, data = mht1)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -2.5151 -1.2403 -0.0023  1.2603  2.5195
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.529786   0.033005  197.84  <2e-16 ***
## Technology_Usage_Hours -0.004489   0.004579   -0.98    0.327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.451 on 9998 degrees of freedom
## Multiple R-squared:  9.612e-05, Adjusted R-squared:  -3.893e-06
## F-statistic: 0.9611 on 1 and 9998 DF, p-value: 0.3269
```

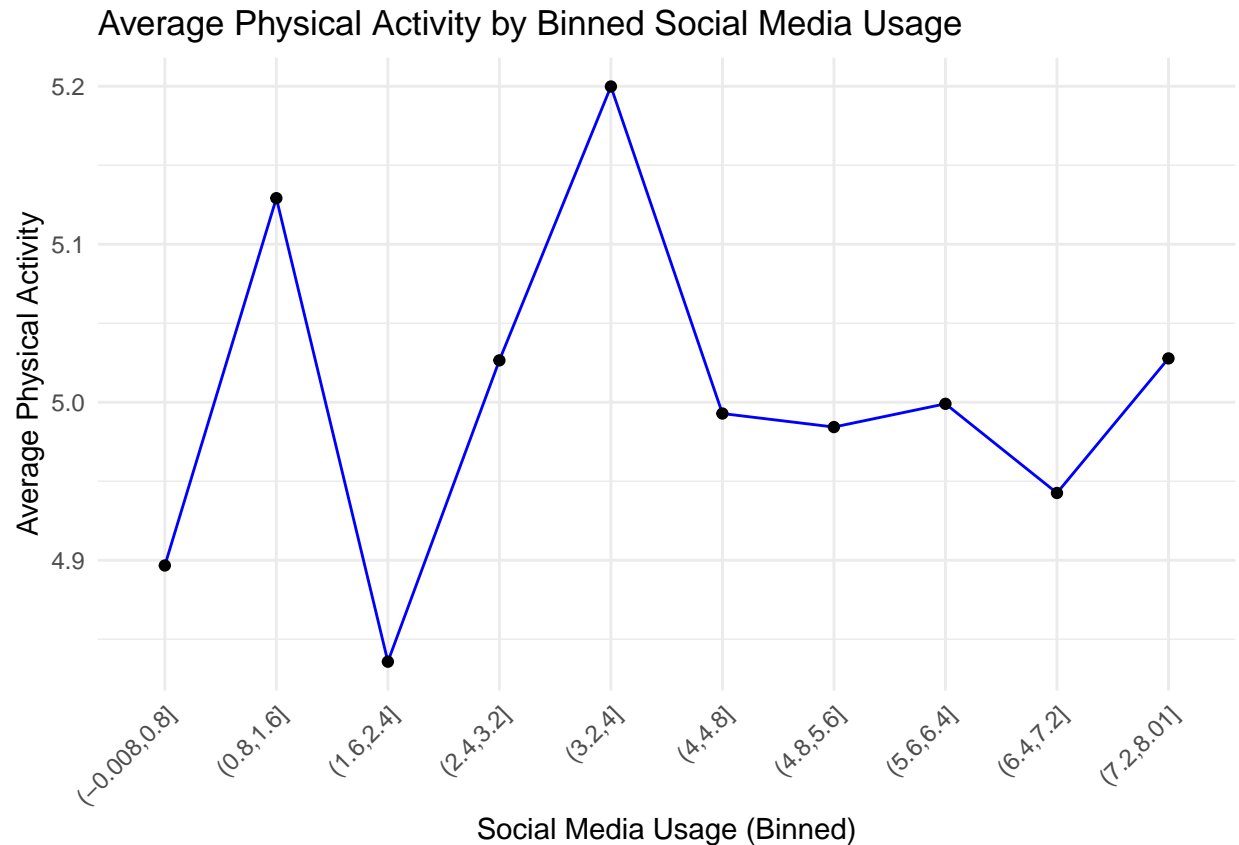
#The slope coefficients for technology_Usage_Hours is -0.0045, indicating that for each additional hour technology use, sleep hours decrease by just 0.0045 hours. this effect is not statistically significant. The R-squared value is extremely low which is 0.000096 indicating which has no predictive power at all. The effect size is negligible, not statistically significant, and the model explains almost none of the variation in sleep duration.

Plot for Is heavy social media usage associated with reduced physical activity?

```
library(dplyr)

# Bin social media usage
data_binned <- mht1 %>%
  mutate(SM_Bin = cut(Social_Media_Usage_Hours, breaks = 10)) %>%
  group_by(SM_Bin) %>%
  summarise(Avg_Physical = mean(Physical_Activity_Hours, na.rm = TRUE))

# Plot
ggplot(data_binned, aes(x = SM_Bin, y = Avg_Physical, group = 1)) +
  geom_line(color = "blue") +
  geom_point() +
  labs(title = "Average Physical Activity by Binned Social Media Usage",
       x = "Social Media Usage (Binned)", y = "Average Physical Activity") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
library(BayesFactor)

bf_result <- correlationBF(mht1$Social_Media_Usage_Hours, mht1$Physical_Activity_Hours)
bf_result
```

```
## Bayes factor analysis
## -----
## [1] Alt., r=0.333 : 0.02413678 ±0%
##
## Against denominator:
##   Null, rho = 0
## ---
## Bayes factor type: BFcorrelation, Jeffreys-beta*
```

#The Bayesian correlation analysis produced a Bayes Factor of $BF = 0.024$, indicating strong evidence in favor of the null hypothesis. In other words, the observed data are approximately 42 times more likely under the assumption that there is no correlation between social media usage and physical activity. This result reinforces the conclusion from the frequentist model, suggesting that social media usage is not meaningfully associated with physical activity levels in this dataset.

#Linear regression for Is heavy social media usage associated with reduced physical activity?

```
lm_model <- lm(Physical_Activity_Hours ~ Social_Media_Usage_Hours, data = mht1)
summary(lm_model)
```

```
##
## Call:
## lm(formula = Physical_Activity_Hours ~ Social_Media_Usage_Hours,
##     data = mht1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.0142 -2.5137 -0.0196  2.5376  5.0060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.992270   0.057724  86.485   <2e-16 ***
## Social_Media_Usage_Hours 0.002918   0.012557   0.232   0.816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.905 on 9998 degrees of freedom
## Multiple R-squared:  5.4e-06,    Adjusted R-squared:  -9.462e-05
## F-statistic: 0.05399 on 1 and 9998 DF,  p-value: 0.8163
```

#A simple linear regression was conducted to examine whether social media usage is associated with physical activity hours per day. The results showed no statistically significant relationship between the two variables, $t(9998) = 0.23$, $p = 0.816$. The model's R^2 value was virtually zero ($R^2 = 0.0000054$), indicating that social media use explains almost none of the variation in physical activity. In short, the data do not support the idea that heavier social media usage predicts lower physical activity levels.

```
#Saving the cleaned data
write_csv(mht1 , "cleaned_mental_health_data.csv")
```