Examining Life Amidst the COVID Era

Grace Esther

```r
source("explorationFunction.R")
library(ggplot2) # data viz
library(dplyr) # data manipulation
library(corrplot) # visualizing correlation matrices
library(ltm) # Item Reponse Theory -> biserial.cor
library(vcd) # Categorical Data Visualization -> assocstats
library(DescTools) # miscellaneous functions for descriptive statistics -> PercTable
library(ggridges) # ridge plots -> geom_density_ridges
#library(plotly) # for plotly interactive visualization
library(tidyr) # for pivot_longer
```

1. Introduction

COVID-19 has affected the whole world, causing big lockdowns and major disruptions in many areas. People had to stay home and avoid social contact to slow the spread of the virus, which changed how we live and interact. Students, especially in the Delhi-National Capital Region (NCR), have been particularly affected. Schools had to switch quickly to online learning, creating serious challenges for traditional education and highlighting the need to understand how this change impacts students' lives.

This study looks at how the COVID-19 pandemic has affected students in the Delhi-NCR, focusing on their mental health, social interactions, and education. The shift to online learning has been tough for regular educational systems, making it important to understand its effects on students. This information is key to creating better educational strategies and mental health services. The research also aligns with Sustainable Development Goal 3 (SDG 3), which aims to ensure healthy lives and promote well-being for all ages.

The report is for researchers, educators, policymakers, and others interested in how COVID-19 has affected students' psychology and education. The main goals are to evaluate how well online learning worked during the pandemic, examine the link between students' daily activities and their health, identify ways students cope with stress and anxiety, and offer suggestions for improving education and mental health services.

The study uses data on students' daily activities, health issues, and online class experiences, applying data analysis methods and statistical tools. There are limitations, such as potential biases in self-reported data and the focus on a specific region. The study assumes that the data accurately represents the experiences of students in the Delhi-NCR.

Despite these limitations, this study is important for understanding the wider impacts of the COVID-19 pandemic on students' lives. It provides valuable insights that could shape future education strategies and mental health services. The report offers a detailed look at current issues and potential solutions, making it a useful resource for anyone concerned about students' well-being during and after the pandemic.

---

2. Data Description

The data for this research was collected from an online survey given to students between July 13 and July 17, 2020, using Google online platforms. The aim was to gather information on how the COVID-19 pandemic affected students' daily lives, health, online class experiences, and other areas. This dataset is referenced in the research article "COVID-19 and its Impact on Education, Social Life, and Mental Health of Students: A Survey" by Chaturvedi, Vishwakarma, and Singh (2020), and is available to the public on Kaggle.

The dataset contains 19 variables and 1,182 responses, each reflecting different aspects of students' experiences during the pandemic. Below is a brief overview of each variable:

- ID: Unique identifier assigned to every respondent.
- Region.of.residence: The region in which the respondent resides (either inside Delhi-NCR or outside of it).
- Age.of.Subject: Age of the respondent in years.
- Time.spent.on.Online.Class: The amount of time (in hours) spent each day on online courses.
- Rating.of.Online.Class.experience: The rating (Very poor, Poor, Average, Good, Excellent) provided by the respondent for their online class experience.
- Medium.for.online.class: The device (such as a laptop, desktop computer, tablet, smartphone, etc.) used to attend online classes.
- Time.spent.on.self.study: The amount of time (in hours) dedicated to self-study each day.
- Time.spent.on.fitness: The number of hours a day dedicated to physical fitness activities.
- Time.spent.on.sleep: Total amount of time (in hours) spent each day sleeping.
- Time.spent.on.social.media: The amount of time (in hours) spent every day on social media.
- Prefered.social.media.platform: The platform of choice for social media (such as Instagram, Youtube, Linkedin, etc.).
- Time.spent.on.TV: The amount of time (in hours) per day spent watching TV.
- Number.of.meals.per.day: Number of meals consumed per day.
- Change.in.your.weight: Variation in weight (increased, decreased, or remained constant) during the lockdown.
- Health.issue.during.lockdown: Whether the respondent had any health issues during the lockdown (Yes/No).
- Stress.busters: Things respondents do to relieve stress (e.g., cook, browse social media, listen to music, etc.).
- Time.utilized: Whether the respondent felt they used their time effectively during the lockdown (Yes/No).
- Do.you.find.yourself.more.connected.with.your.family..close.friends...relatives: Whether the respondent felt more connected with family, close friends, and relatives during the lockdown (Yes/No).
- What.you.miss.the.most: The aspect of life respondents missed the most during the lockdown (e.g., school/college, roaming around freely, travelling, etc).

```r
# import data from directory in data frame format
df.raw <- read.csv("COVID-19 Survey Student Responses.csv")
head(df.raw, n=10)
```

```
##      ID Region.of.residence Age.of.Subject Time.spent.on.Online.Class
## 1   R1           Delhi-NCR             21                          2
## 2   R2           Delhi-NCR             21                          0
## 3   R3           Delhi-NCR             20                          7
## 4   R4           Delhi-NCR             20                          3
## 5   R5           Delhi-NCR             21                          3
## 6   R6           Delhi-NCR             21                          0
## 7   R7           Delhi-NCR             19                          2
## 8   R8   Outside Delhi-NCR             19                          2
## 9   R9           Delhi-NCR             21                          3
## 10 R10   Outside Delhi-NCR             20                          0
##    Rating.of.Online.Class.experience Medium.for.online.class
## 1                               Good          Laptop/Desktop
## 2                          Excellent              Smartphone
## 3                          Very poor          Laptop/Desktop
## 4                          Very poor              Smartphone
## 5                               Good          Laptop/Desktop
## 6                          Very poor              Smartphone
## 7                          Very poor              Smartphone
## 8                          Very poor                  Tablet
## 9                          Very poor          Laptop/Desktop
## 10                         Very poor          Laptop/Desktop
##    Time.spent.on.self.study Time.spent.on.fitness Time.spent.on.sleep
## 1                         4                   0.0                   7
## 2                         0                   2.0                  10
## 3                         3                   0.0                   6
## 4                         2                   1.0                   6
## 5                         3                   1.0                   8
## 6                         6                   0.0                   5
## 7                         2                   1.0                   5
## 8                         1                   1.0                  10
## 9                         4                   1.0                   8
## 10                        1                   0.5                   8
##    Time.spent.on.social.media Prefered.social.media.platform Time.spent.on.TV
## 1                           3                       Linkedin                1
## 2                           3                        Youtube                0
## 3                           2                       Linkedin                0
## 4                           5                      Instagram                0
## 5                           3                      Instagram                1
## 6                           1                        Youtube                0
## 7                           4                      Instagram                0
## 8                           5                      Instagram                0
## 9                           2                       Whatsapp                1
## 10                          5                      Instagram                3
##    Number.of.meals.per.day Change.in.your.weight Health.issue.during.lockdown
## 1                        4             Increased                           NO
## 2                        3             Decreased                           NO
## 3                        3       Remain Constant                           NO
## 4                        3             Decreased                           NO
## 5                        4       Remain Constant                           NO
## 6                        1             Decreased                          YES
## 7                        3             Increased                           NO
## 8                        3             Increased                          YES
## 9                        3             Increased                           NO
## 10                       3             Decreased                          YES
##                    Stress.busters Time.utilized
## 1                         Cooking           YES
## 2  Scrolling through social media           YES
## 3              Listening to music            NO
## 4             Watching web series            NO
## 5                    Social Media            NO
## 6   Coding and studying for exams            NO
## 7             Watching web series            NO
## 8  Scrolling through social media            NO
## 9                  Online surfing            NO
## 10           live stream watching            NO
##    Do.you.find.yourself.more.connected.with.your.family..close.friends...relatives...
## 1                                                                                  YES
## 2                                                                                   NO
## 3                                                                                  YES
## 4                                                                                   NO
## 5                                                                                   NO
## 6                                                                                  YES
## 7                                                                                  YES
## 8                                                                                  YES
## 9                                                                                   NO
## 10                                                                                  NO
##    What.you.miss.the.most
## 1          School/college
## 2   Roaming around freely
## 3              Travelling
## 4     Friends , relatives
## 5              Travelling
## 6          School/college
## 7     Friends , relatives
## 8          Eating outside
## 9     Friends , relatives
## 10         School/college
```

```
# showing the structure of our dataframe
str(df.raw)
```

```
## ’data.frame':    1182 obs. of  19 variables:
##  $ ID                                                                                : chr  "R1" "R2
##  $ Region.of.residence                                                               : chr  "Delhi-N
##  $ Age.of.Subject                                                                    : int  21 21 20
##  $ Time.spent.on.Online.Class                                                        : num  2 0 7 3
##  $ Rating.of.Online.Class.experience                                                 : chr  "Good" "
##  $ Medium.for.online.class                                                           : chr  "Laptop/
##  $ Time.spent.on.self.study                                                          : num  4 0 3 2
##  $ Time.spent.on.fitness                                                             : num  0 2 0 1
##  $ Time.spent.on.sleep                                                               : num  7 10 6 6
##  $ Time.spent.on.social.media                                                        : num  3 3 2 5
##  $ Prefered.social.media.platform                                                    : chr  "Linkedi
##  $ Time.spent.on.TV                                                                  : chr  "1" "0"
##  $ Number.of.meals.per.day                                                           : int  4 3 3 3
##  $ Change.in.your.weight                                                             : chr  "Increas
##  $ Health.issue.during.lockdown                                                      : chr  "NO" "NO
##  $ Stress.busters                                                                    : chr  "Cooking
##  $ Time.utilized                                                                     : chr  "YES" "Y
##  $ Do.you.find.yourself.more.connected.with.your.family..close.friends...relatives...: chr  "YES" "N
##  $ What.you.miss.the.most                                                            : chr  "School/
```

- df.raw is a data frame with 1182 observations/ rows with 19 features/ variables

- it tells which variables has what type of variables (chr: character/ string, int: whole number, num: numeric)
- it also includes some values for each variable

3. Data Preprocessing

Preprocessing was carried out in several steps to ensure the data was ready for analysis and consistent. The steps included:

1. Renaming the Columns: The column names were changed using the dplyr package to improve readability and understanding.

```
# using dplyr library to rename the columns name's
# the left side is the new column name and the right side is the old column name
df.raw = df.raw %>% rename(
  Region = Region.of.residence,
  Age = Age.of.Subject,
  OnlineClass_Time = Time.spent.on.Online.Class,
  OnlineClass_Rating = Rating.of.Online.Class.experience,
  OnlineClass_Medium = Medium.for.online.class,
  SelfStudy_Duration = Time.spent.on.self.study,
  Fitness_Duration = Time.spent.on.fitness,
  Sleeping_Duration = Time.spent.on.sleep,
  SosMed_Duration = Time.spent.on.social.media,
  SosMed_Medium = Prefered.social.media.platform,
  WatchingTV_Duration = Time.spent.on.TV,
  NumberOfMeals = Number.of.meals.per.day,
  WeightChange = Change.in.your.weight,
  Having_HealthIssue = Health.issue.during.lockdown,
  Stress_Busters = Stress.busters,
  Time_Utilized = Time.utilized,
  Connected_with_Family_Friends_Relatives = Do.you.find.yourself.more.connected.with.your.family..close
  Most_Missed_Things = What.you.miss.the.most
)
```

2. Data Subsetting: The “ID” column was removed from the dataset as it was deemed irrelevant for the analysis, resulting in the creation of a new data frame without it.

```
# create a new data frame from subsetting our old data frame and remove the ID column because we think
df <- subset(df.raw, select = -c(ID))

# checking our new data frame structure's
str(df)
```

```
## ’data.frame’:    1182 obs. of  18 variables:
##  $ Region                                 : chr  "Delhi-NCR" "Delhi-NCR" "Delhi-NCR" "Delhi-NCR" ...
##  $ Age                                    : int  21 21 20 20 21 21 19 19 21 20 ...
##  $ OnlineClass_Time                       : num  2 0 7 3 3 0 2 2 3 0 ...
##  $ OnlineClass_Rating                     : chr  "Good" "Excellent" "Very poor" "Very poor" ...
##  $ OnlineClass_Medium                     : chr  "Laptop/Desktop" "Smartphone" "Laptop/Desktop" "Sma
##  $ SelfStudy_Duration                     : num  4 0 3 2 3 6 2 1 4 1 ...
##  $ Fitness_Duration                       : num  0 2 0 1 1 0 1 1 1 0.5 ...
##  $ Sleeping_Duration                         : num  7 10 6 6 8 5 5 10 8 8 ...
##  $ SosMed_Duration                           : num  3 3 2 5 3 1 4 5 2 5 ...
##  $ SosMed_Medium                             : chr  "Linkedin" "Youtube" "Linkedin" "Instagram" ...
##  $ WatchingTV_Duration                       : chr  "1" "0" "0" "0" ...
##  $ NumberOfMeals                             : int  4 3 3 3 4 1 3 3 3 3 ...
##  $ WeightChange                              : chr  "Increased" "Decreased" "Remain Constant" "Decrease
##  $ Having_HealthIssue                        : chr  "NO" "NO" "NO" "NO" ...
##  $ Stress_Busters                            : chr  "Cooking" "Scrolling through social media" "Listenin
##  $ Time_Utilized                             : chr  "YES" "YES" "NO" "NO" ...
##  $ Connected_with_Family_Friends_Relatives: chr  "YES" "NO" "YES" "NO" ...
##  $ Most_Missed_Things                        : chr  "School/college" "Roaming around freely" "Travellin
```

3. Handling Inconsistencies:

- Missing values in the columns "SosMed_Medium" and "WatchingTV_Duration" were set to "NA."
- Inconsistent values were corrected, such as changing "Whatsapp" to "WhatsApp" and setting missing TV watching durations to 0.
- Similar responses in "Most_Missed_Things" were grouped and standardized to ensure consistency.

```r
# function to print the unique value and the total of unique value count in specific column (the column
CekUnique <- function(i)
{
  # cat is a function which stands for concatenate and print
    cat(i,": [",length(unique(df[[i]])),"]\n") # the total of unique value count
    cat(unique(df[[i]]), sep = "; ") # unique value
    cat("\n\n")
}

# check unique for some specific columns in our data frame
CekUnique("SosMed_Medium")
```

```
## SosMed_Medium : [ 16 ]
## Linkedin; Youtube; Instagram; Whatsapp; None; Reddit; Snapchat; Omegle; Twitter; Telegram; Facebook;
```

```r
CekUnique("WatchingTV_Duration")
```

```
## WatchingTV_Duration : [ 25 ]
## 1; 0; 3; 0.5; n; 2; 4.5; 1.5; N; 4; 0.3; 5; No tv; 0.1; 0.25; 6; 0.6; 7; 8; 15;  ; 0.75; 2.5; 3.5; 0
```

```r
CekUnique("Most_Missed_Things")
```

```
## Most_Missed_Things : [ 51 ]
## School/college; Roaming around freely; Travelling; Friends , relatives; Eating outside; Colleagues;
```

This is useful for data preprocessing so that we can identify inconsistencies or missing information.

The value "none" was replaced with "NA" to handle missing data consistently.

```r
df$SosMed_Medium[df$SosMed_Medium %in% c("None", "None ")] <- NA
df$SosMed_Medium[df$SosMed_Medium == "Whatsapp"] <- "WhatsApp"
df$WatchingTV_Duration[df$WatchingTV_Duration %in% c("N", "n", " ", "")] <- NA
df$WatchingTV_Duration[df$WatchingTV_Duration=="No tv"] <- 0
df$Most_Missed_Things[df$Most_Missed_Things %in% c("All ", "all", "All of the above ", "All the above",
df$Most_Missed_Things[df$Most_Missed_Things %in% c("NOTHING", "I have missed nothing ", "nothing", "Not
```

4. Correcting Data Types:

- The data type of “WatchingTV_Duration” was changed from character to numeric.
- To maintain consistency in analysis, categorical variables were transformed into factors with predetermined levels. For example, “OnlineClass_Rating” included levels such as “Very poor,” “Poor,” “Average,” “Good,” and “Excellent.”
- Additional categorical variables converted into factors included “OnlineClass_Medium,” “SosMed_Medium,” “WeightChange,” “Having_HealthIssue,” “Stress_Busters,” “Time_Utilized,” “Connected_with_Family_Friends_Rel and “Most_Missed_Things.”

```
# Change the datatype from char to numeric
df$WatchingTV_Duration = as.numeric(df$WatchingTV_Duration)

df$Region <- factor(df$Region, levels = unique(df$Region))
CekUnique("OnlineClass_Rating")
```

```
## OnlineClass_Rating : [ 6 ]
## Good; Excellent; Very poor; Average; NA; Poor
```

```
df$OnlineClass_Rating <- factor(df$OnlineClass_Rating, levels = c("Very poor","Poor","Average","Good","
df$OnlineClass_Medium <- factor(df$OnlineClass_Medium, levels = unique(df$OnlineClass_Medium))
df$SosMed_Medium <- factor(df$SosMed_Medium, levels = unique(df$SosMed_Medium))
CekUnique("WeightChange")
```

```
## WeightChange : [ 3 ]
## Increased; Decreased; Remain Constant
```

```
df$WeightChange <- factor (df$WeightChange, levels = c("Decreased","Remain Constant","Increased"))
df$Having_HealthIssue <- factor(df$Having_HealthIssue)
df$Stress_Busters <- factor(df$Stress_Busters, levels = unique(df$Stress_Busters))
df$Time_Utilized <- factor(df$Time_Utilized)
df$Connected_with_Family_Friends_Relatives <- factor(df$Connected_with_Family_Friends_Relatives)
df$Most_Missed_Things <- factor(df$Most_Missed_Things, levels = unique(df$Most_Missed_Things))
```

```
# choose the name of columns which datatype is numeric/ integer or categorical/ char
num_cols = c()
cat_cols = c()
for (i in names(df))
{
  if (is.numeric(df[[i]]) || is.integer(df[[i]]))
  {
    num_cols <- c(num_cols,i)
  }
  else
  {
    cat_cols <- c(cat_cols,i)
  }
}

cat("Numeric: ",paste(num_cols,collapse=", "),"\n\n")
```

```
## Numeric:  Age, OnlineClass_Time, SelfStudy_Duration, Fitness_Duration, Sleeping_Duration, SosMed_Dura
```

```
cat("Categorical: ",paste(cat_cols,collapse=", "),"\n")
```

```
## Categorical:  Region, OnlineClass_Rating, OnlineClass_Medium, SosMed_Medium, WeightChange, Having_Hea
```

5. Dealing with Missing Data:

- The mean of the column was used to fill in the missing data for numerical columns which had missing values.
- The mode, or most frequent value, of a categorical column, was used to fill in the missing value.

```
# Count how many NA value in each column
for (i in names(df))
{
  cat(i,": ",sum(is.na(df[[i]])),"\n")
}
```

```
## Region :  0
## Age :  0
## OnlineClass_Time :  0
## OnlineClass_Rating :  24
## OnlineClass_Medium :  51
## SelfStudy_Duration :  0
## Fitness_Duration :  0
## Sleeping_Duration :  0
## SosMed_Duration :  0
## SosMed_Medium :  18
## WatchingTV_Duration :  10
## NumberOfMeals :  0
## WeightChange :  0
## Having_HealthIssue :  0
## Stress_Busters :  0
## Time_Utilized :  0
## Connected_with_Family_Friends_Relatives :  0
## Most_Missed_Things :  0
```

```
# Fill all the NA value in non numeric variable with mode and using mean to fill all the NA value in nu
for (i in cat_cols)
{
  # names is used to extract the name and not the frequency value
  mode <- names(sort(table(df[[i]]), decreasing=TRUE)[1])
  df[[i]][is.na(df[[i]])] <- mode
}

for (i in num_cols)
{
  mean <- mean(df[[i]], na.rm = TRUE)
  df[[i]][is.na(df[[i]])] <- mean
}
```

Verify the structure of our dataframe after completing the data preprocessing steps.

```r
str(df)
```

```
## ’data.frame':    1182 obs. of  18 variables:
##  $ Region                                 : Factor w/ 2 levels "Delhi-NCR","Outside Delhi-NCR": 1 1
##  $ Age                                    : num  21 21 20 20 21 21 19 19 21 20 ...
##  $ OnlineClass_Time                       : num  2 0 7 3 3 0 2 2 3 0 ...
##  $ OnlineClass_Rating                     : Factor w/ 5 levels "Very poor","Poor",..: 4 5 1 1 4 1 1
##  $ OnlineClass_Medium                     : Factor w/ 5 levels "Laptop/Desktop",..: 1 2 1 2 1 2 2 3
##  $ SelfStudy_Duration                     : num  4 0 3 2 3 6 2 1 4 1 ...
##  $ Fitness_Duration                       : num  0 2 0 1 1 0 1 1 1 0.5 ...
##  $ Sleeping_Duration                      : num  7 10 6 6 8 5 5 10 8 8 ...
##  $ SosMed_Duration                        : num  3 3 2 5 3 1 4 5 2 5 ...
##  $ SosMed_Medium                          : Factor w/ 13 levels "Linkedin","Youtube",..: 1 2 1 3 3 2
##  $ WatchingTV_Duration                    : num  1 0 0 0 1 0 0 0 1 3 ...
##  $ NumberOfMeals                          : num  4 3 3 3 4 1 3 3 3 3 ...
##  $ WeightChange                           : Factor w/ 3 levels "Decreased","Remain Constant",..: 3 1
##  $ Having_HealthIssue                     : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 2 1 2 ...
##  $ Stress_Busters                         : Factor w/ 86 levels "Cooking","Scrolling through social
##  $ Time_Utilized                          : Factor w/ 2 levels "NO","YES": 2 2 1 1 1 1 1 1 1 1 ...
##  $ Connected_with_Family_Friends_Relatives: Factor w/ 2 levels "NO","YES": 2 1 2 1 1 2 2 2 1 1 ...
##  $ Most_Missed_Things                     : Factor w/ 36 levels "School/college",..: 1 2 3 4 3 1 4 5
```

---

4. Data Exploration

```r
summary(df)
```

```
##                Region         Age        OnlineClass_Time OnlineClass_Rating
##  Delhi-NCR        :721   Min.   : 7.00   Min.   : 0.000   Very poor:437
##  Outside Delhi-NCR:461   1st Qu.:17.00   1st Qu.: 2.000   Poor     : 30
##                          Median :20.00   Median : 3.000   Average  :387
##                          Mean   :20.17   Mean   : 3.209   Good     :230
##                          3rd Qu.:21.00   3rd Qu.: 5.000   Excellent: 98
##                          Max.   :59.00   Max.   :10.000
##
##                      OnlineClass_Medium SelfStudy_Duration Fitness_Duration
##  Laptop/Desktop              :596      Min.   : 0.000     Min.   :0.0000
##  Smartphone                  :539      1st Qu.: 2.000     1st Qu.:0.0000
##  Tablet                      : 37      Median : 2.000     Median :1.0000
##  Any Gadget                  :  5      Mean   : 2.912     Mean   :0.7658
##  Smartphone or Laptop/Desktop:  5      3rd Qu.: 4.000     3rd Qu.:1.0000
##                                        Max.   :18.000     Max.   :5.0000
##
##  Sleeping_Duration SosMed_Duration    SosMed_Medium WatchingTV_Duration
##  Min.   : 4.000    Min.   : 0.000   Instagram:370   Min.   : 0.000
##  1st Qu.: 7.000    1st Qu.: 1.000   WhatsApp :337   1st Qu.: 0.000
##  Median : 8.000    Median : 2.000   Youtube  :314   Median : 1.000
##  Mean   : 7.871    Mean   : 2.366   Linkedin : 61   Mean   : 1.023
##  3rd Qu.: 9.000    3rd Qu.: 3.000   Facebook : 52   3rd Qu.: 2.000
##  Max.   :15.000    Max.   :10.000   Twitter  : 28   Max.   :15.000
##                                     (Other)  : 20
##  NumberOfMeals            WeightChange Having_HealthIssue
##  Min.   :1.000   Decreased      :209   NO :1021
##  1st Qu.:2.000   Remain Constant:535   YES: 161
##  Median :3.000   Increased      :438
##  Mean   :2.918
##  3rd Qu.:3.000
##  Max.   :8.000
##
##                         Stress_Busters Time_Utilized
##  Listening to music            :276    NO :608
##  Online gaming                 :175    YES:574
##  Watching web series           :102
##  Reading books                 : 77
##  Scrolling through social media: 74
##  Sleeping                      : 71
##  (Other)                       :407
##  Connected_with_Family_Friends_Relatives             Most_Missed_Things
##  NO :351                                  School/college       :379
##  YES:831                                  Friends , relatives  :223
##                                           Travelling           :183
##                                           Roaming around freely:149
##                                           Eating outside       :104
##                                           Colleagues           : 67
##                                           (Other)              : 77
```

Categorical Variables: The summary shows how often each category was chosen for variables like Region, OnlineClass_Rating, OnlineClass_Medium, and others.

- Region: The dataset includes 461 participants from places outside Delhi-NCR and 721 participants from Delhi-NCR.
- OnlineClass_Rating: Ratings for online classes varied among participants, with responses ranging from very poor (437), poor (30), average (387), good (230), to excellent (98), reflecting diverse opinions on online learning experiences.
- OnlineClass_Medium: Participants used different devices for online classes, such as laptop/desktop (596), smartphone (539), tablet (37), any gadget (5), and both smartphone and laptop/desktop (5), indicating preferences for flexible learning methods.
- SosMed_Medium: Various social media platforms were used by participants, including Instagram (370), WhatsApp (337), YouTube (314), LinkedIn (61), Facebook (52), Twitter (28), and other platforms (20), highlighting diverse social media usage.
- WeightChange: Participants reported weight changes during the pandemic, with decreases (209), no change (535), and increases (438), showing varying impacts on physical health.
- Having_HealthIssue: Participants reported health issues, with responses split into no (1021) and yes (161), underscoring the prevalence of health problems during the pandemic.
- Stress_Busters: Coping strategies included listening to music (276), playing online video games (175), watching web series (102), reading books (77), browsing social media (74), sleeping (71), and other methods (407), indicating diverse ways of managing stress.
- Time_Utilized: Participants reported whether they utilized their time effectively during the pandemic, with responses categorized as no (608) and yes (574), showing different levels of productivity.
- Connected_with_Family_Friends_Relatives: Participants felt connected with family, friends, or relatives, with responses categorized as no (351) and yes (831), emphasizing the importance of social connections during periods of isolation.

- Most_Missed_Things: Activities missed the most during the pandemic included school/college (379), time with friends/relatives (223), traveling (183), freely roaming (149), dining out (104), colleagues (67), and other activities (77), highlighting the impact of restrictions on daily activities and social interactions.

Numeric Variables: The summary gives us information about variables such as Age, OnlineClass_Time, SelfStudy_Duration, and more.

- Age: The participants' ages ranged from 7 to 59 years, with most around 20 years old.
- OnlineClass_Time: During the pandemic, participants spent between 0 to 10 hours daily on online classes, with an average of 3.21 hours.
- SelfStudy_Duration: Study habits varied widely, with participants spending from 0 to 18 hours daily on self-study, averaging 2.91 hours.
- Fitness_Duration: Daily physical activity varied, with participants engaging from 0 to 5 hours, averaging 0.77 hours.
- Sleeping_Duration: Participants slept an average of 7.87 hours daily, with most sleeping around 8 hours.
- SosMed_Duration: Daily social media use ranged from 0 to 10 hours, with an average of 2.37 hours.
- WatchingTV_Duration: Daily TV viewing ranged from 0 to 15 hours, averaging 1.02 hours.
- NumberOfMeals: Participants typically ate 2.92 meals daily, with most having around 3 meals.

These summaries provide insights into how participants spent their time and managed daily activities during the pandemic.

Note: (Other) combine all remaining levels of a factor variable with more than six levels into a single "other" category Numeric variables provide mean value and Tukey's five-number summary which includes the minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum

```r
BasicSummary(df)
```

```
##                                    variable    type levels           topLevel
## 1                                    Region  factor      2          Delhi-NCR
## 2                                       Age numeric     42                 20
## 3                          OnlineClass_Time numeric     21                  4
## 4                        OnlineClass_Rating  factor      5          Very poor
## 5                        OnlineClass_Medium  factor      5     Laptop/Desktop
## 6                        SelfStudy_Duration numeric     23                  2
## 7                          Fitness_Duration numeric     11                  1
## 8                         Sleeping_Duration numeric     18                  8
## 9                           SosMed_Duration numeric     22                  1
## 10                            SosMed_Medium  factor     13          Instagram
## 11                      WatchingTV_Duration numeric     22                  0
## 12                            NumberOfMeals numeric      8                  3
## 13                             WeightChange  factor      3    Remain Constant
## 14                       Having_HealthIssue  factor      2                 NO
## 15                           Stress_Busters  factor     86 Listening to music
## 16                            Time_Utilized  factor      2                 NO
## 17 Connected_with_Family_Friends_Relatives  factor      2                YES
## 18                       Most_Missed_Things  factor     36     School/college
##    topCount topFrac missFreq missFrac
## 1       721   0.610        0        0
## 2       211   0.179        0        0
## 3       222   0.188        0        0
## 4       437   0.370        0        0
## 5       596   0.504        0        0
## 6       346   0.293        0        0
## 7       550   0.465        0        0
## 8       390   0.330        0        0
## 9       343   0.290        0        0
## 10      370   0.313        0        0
## 11      444   0.376        0        0
## 12      610   0.516        0        0
## 13      535   0.453        0        0
## 14     1021   0.864        0        0
## 15      276   0.234        0        0
## 16      608   0.514        0        0
## 17      831   0.703        0        0
## 18      379   0.321        0        0
```

- The majority of the variables in this dataset have clear and meaningful names, which is crucial for understanding the data easily
- Among the 18 variables in the df dataset, 10 are categorical factors, and the rest of the columns contain numeric data
  - variable: name or label of the variable
  - type: data type of the variable (ex: character, numeric)
  - levels: number of unique categories/ values
  - topLevel: most frequent category/value
  - topCount: frequency of the top-level category/value
  - topFrac: proportion of the top-level category/value
  - missFreq: count of missing values
  - missFrac: proportion of missing values

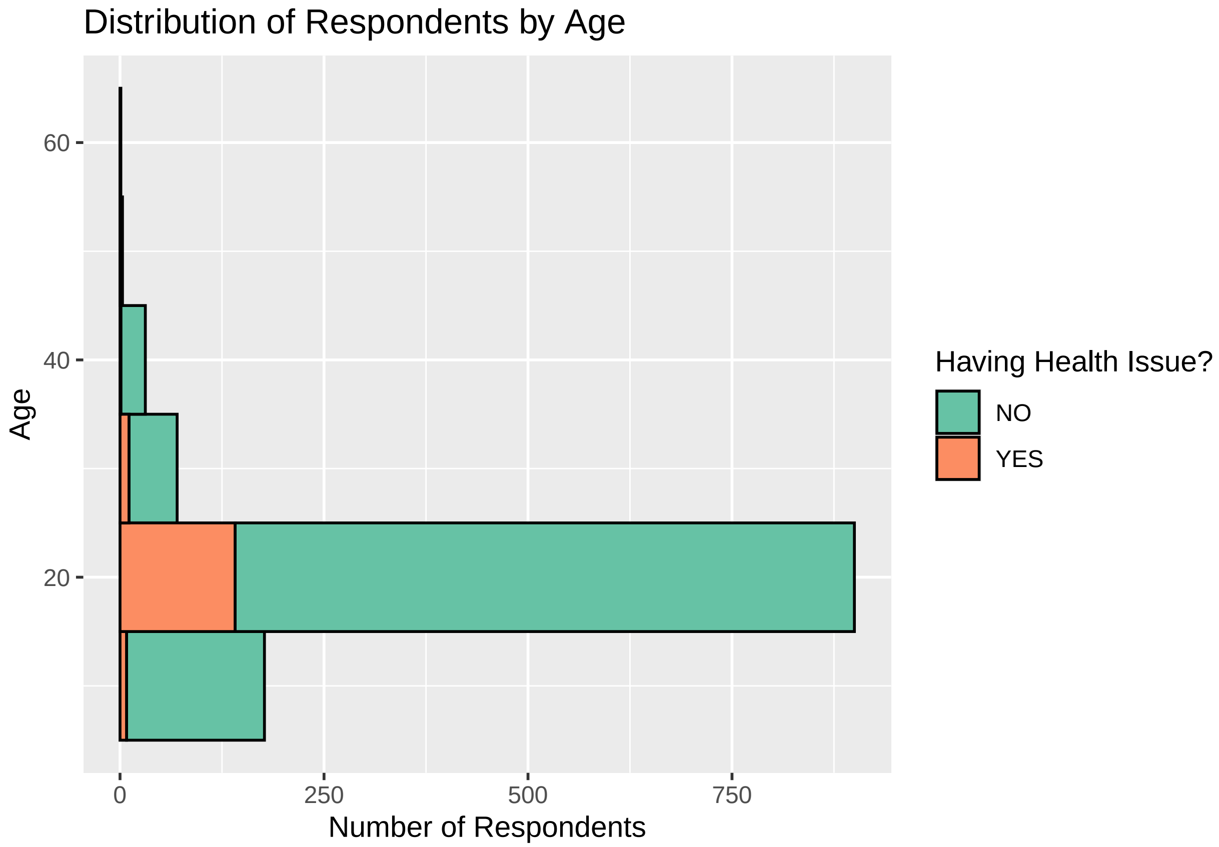
```
num_df <- subset(df,select = num_cols) # make new data frame with consisting only numeric/ integer data
convert_matrix(num_df)
```

```
##                        Age OnlineClass_Time SelfStudy_Duration Fitness_Duration
## Mean                 20.17             3.21               2.91             0.77
## Median               20.00             3.00               2.00             1.00
## Variance             30.43             4.42               4.58             0.52
## Standard Deviation    5.52             2.10               2.14             0.72
## Range                52.00            10.00              18.00             5.00
## Interquartile Range   4.00             3.00               2.00             1.00
##                     Sleeping_Duration SosMed_Duration WatchingTV_Duration
## Mean                             7.87            2.37                1.02
## Median                           8.00            2.00                1.00
## Variance                         2.61            3.12                1.60
## Standard Deviation               1.62            1.77                1.26
## Range                           11.00           10.00               15.00
## Interquartile Range              2.00            2.00                2.00
##                     NumberOfMeals
## Mean                         2.92
## Median                       3.00
## Variance                     0.69
## Standard Deviation           0.83
## Range                        7.00
## Interquartile Range          1.00
```

These statistical measures offer insights into how data is distributed within a dataset:

- Central Tendency: *mean* and *median* help to understand the central value of the data.
- Variability: *variance* and *standard deviation* measure the spread of the data points around the mean
- Range: quick insight into the spread between the minimum and maximum values
- Distribution and Spread: *interquartile range (IQR)* provides insights about the spread of the middle 50% of the data, giving a solid measure of the data's distribution that is less influenced by extreme values or outliers

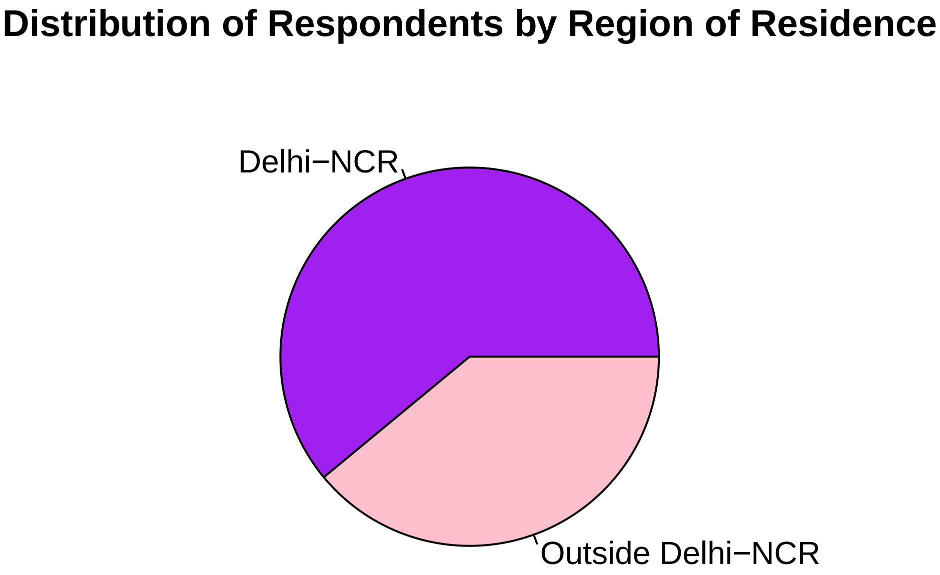
```r
# Age of Respondent visualization (group to some categories with intervals of 10)
age_plot <- ggplot(df, aes(x = Age, fill=Having_HealthIssue)) + # x axis represents age, and the color
  geom_histogram(color = "black", binwidth = 10) + # our histogram will have black outline and the bars
  # add label to our graph
  labs(title = "Distribution of Respondents by Age",
       x = "Age",
       y = "Number of Respondents",
       fill = "Having Health Issue?") +
  scale_fill_brewer(palette = "Set2") + # filling colors using color palettes from the RColorBrewer pac
  coord_flip() # swapping the x and y axis, it is useful when the data categorical name is too long and
print(age_plot)
```



```r
#ggplotly(age_plot)
```

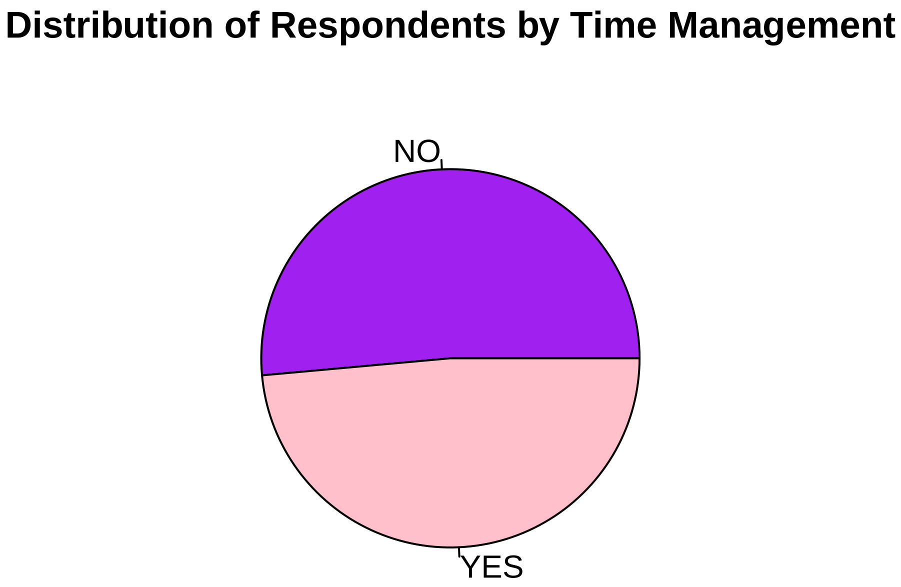
The survey responses indicate that most people who answered are teenagers or in their twenties. This tells us important things about the different issues people of different ages might have. Younger students might find it hard to stay focused and disciplined when they're learning online. Older students might worry about what will happen with their careers in the future.

```
# Respondent's Region of Residence visualization
residence_plot <- pie(table(df$Region), # table is used to create a contingency table of the counts of
    main = "Distribution of Respondents by Region of Residence",
    col = c("purple", "pink"))
```



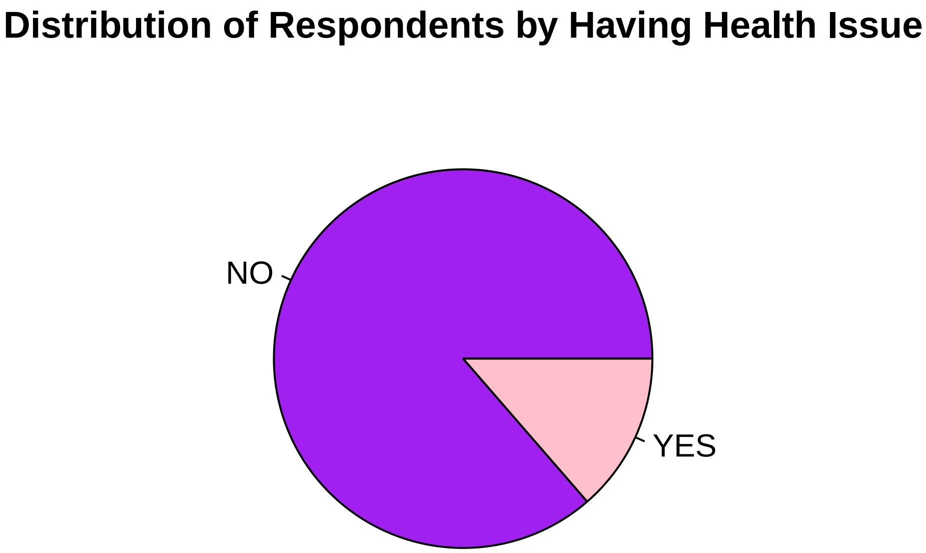
The analysis shows that a large majority, more than half of the respondents, live in the Delhi-National Capital Region (Delhi-NCR).

```
time_plot <- pie(table(df$Time_Utilized),
    main = "Distribution of Respondents by Time Management",
    col = c("purple", "pink"))
```



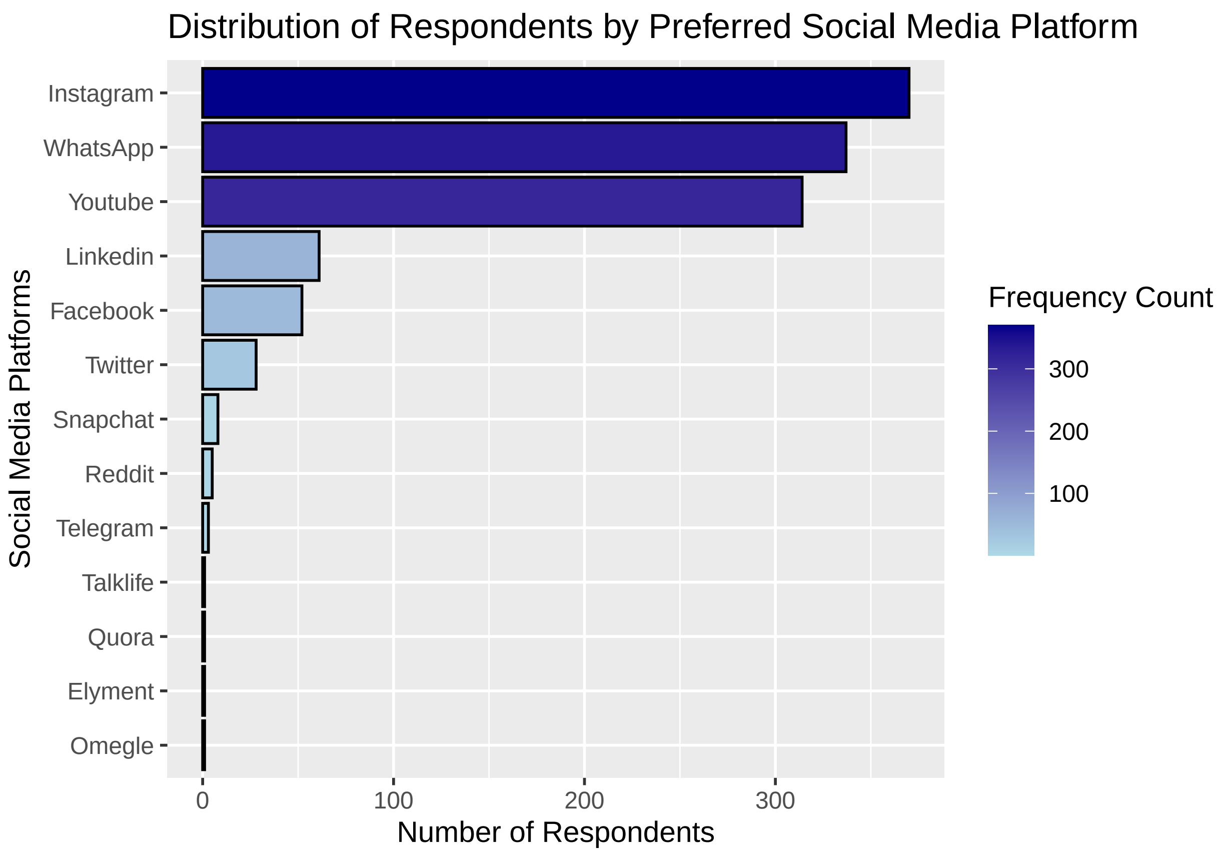
The study reveals that time management has been a significant challenge for students during the pandemic. According to the research, 51.44% of students felt they were not effectively using their time. This finding illustrates the struggle students face in adapting to new routines and managing their time well in an online learning setup.

```r
health_plot <- pie(table(df$Having_HealthIssue),
    main = "Distribution of Respondents by Having Health Issue",
    col = c("purple", "pink"))
```



Based on the survey, about 13.62% of students mentioned having health issues, showing a notable impact on their physical and mental well-being. This finding mirrors a broader worldwide pattern of increased health concerns due to extended lockdowns and lifestyle adjustments.

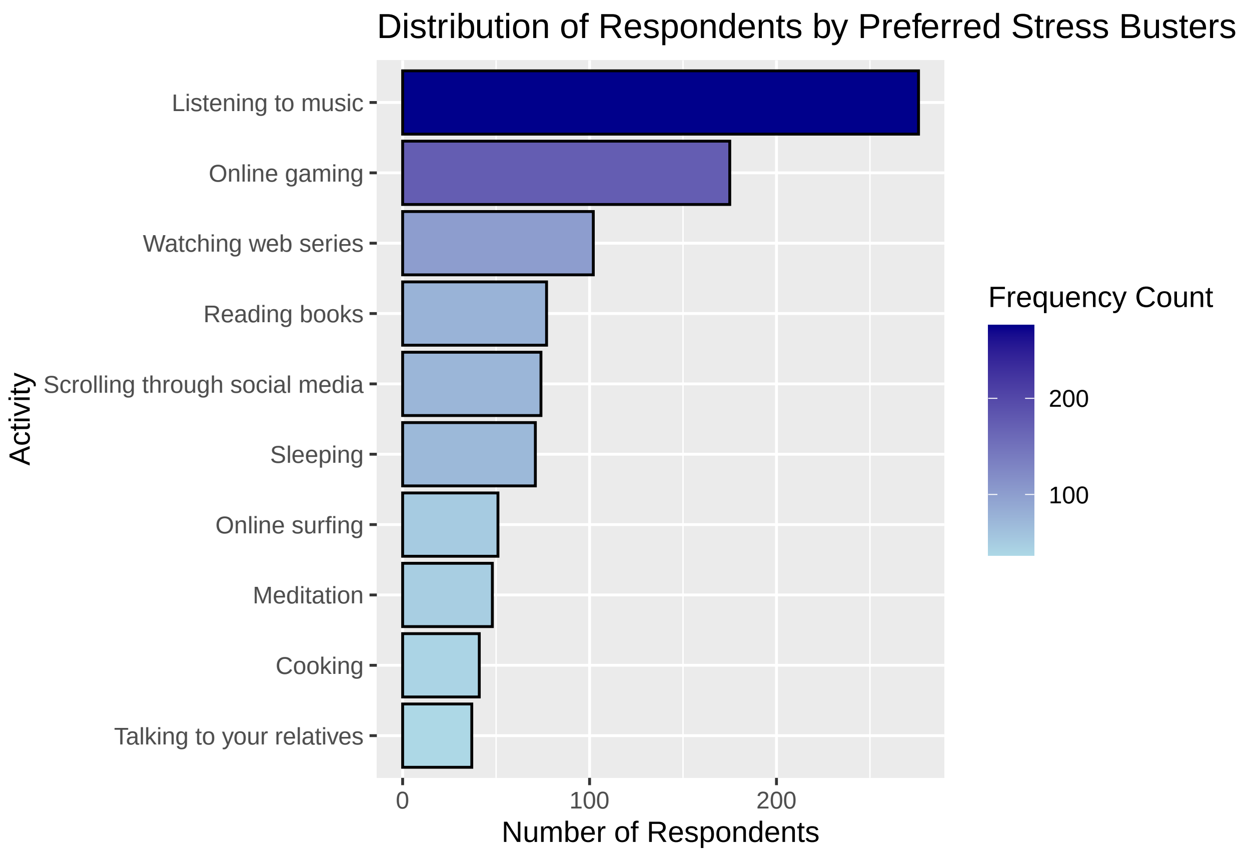
```
# Respondent's Preferred Social Media Platform visualization
sosmed_plot <- df %>% # %>% is a pipe operator to chain multiple operation
  group_by(SosMed_Medium) %>% # group data by SosMed_Medium
  summarise(freqCount = n()) %>% # returns one row for each combination of grouping variables; n() -> c
  ggplot(aes(x=reorder(SosMed_Medium, freqCount),  y=freqCount, fill=freqCount)) + # reorder the sosmed
  geom_bar(color="black", stat="identity") + # stat = "identity" -> height of the bars is determined by
  labs(title = "Distribution of Respondents by Preferred Social Media Platform",
       x = "Social Media Platforms",
       y = "Number of Respondents",
       fill = "Frequency Count") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") + # filling colors using gradient with star
  coord_flip()
print(sosmed_plot)
```



```
#ggplotly(sosmed_plot)
```

The report also examines which social media platforms students preferred most, offering insights into their online habits and how social media helped them stay connected during the pandemic. Instagram and WhatsApp emerged as the top choices, highlighting their significant role in students' social interactions.

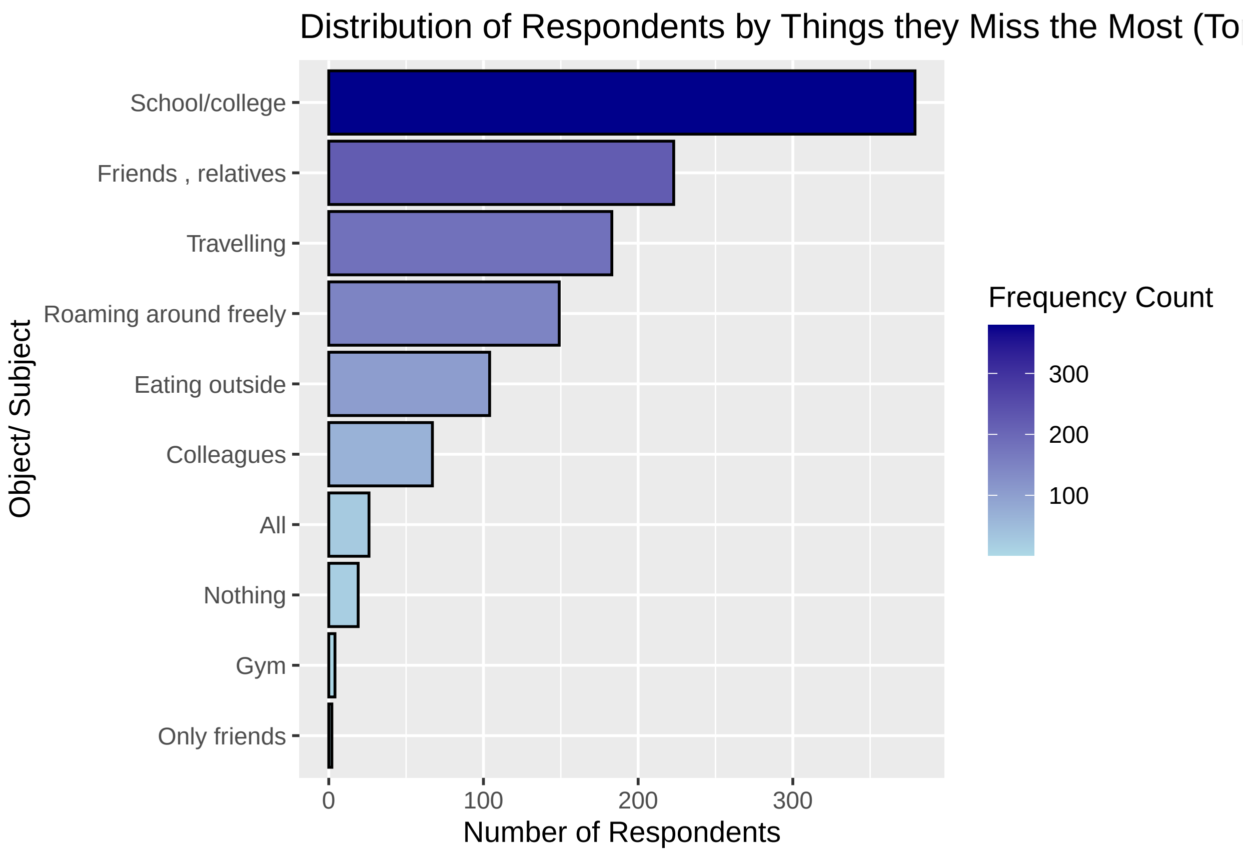
```
# Top 20 Respondent's Preferred Stress Busters
stress_plot <- df %>%
  group_by(Stress_Busters) %>%
  summarise(freqCount = n()) %>%
  top_n(10, freqCount) %>%  # filter the top 10 stress busters
  ggplot(aes(x=reorder(Stress_Busters, freqCount),  y=freqCount, fill=freqCount)) +
  geom_bar(color="black", stat="identity") +
  labs(title = "Distribution of Respondents by Preferred Stress Busters (Top 10)",
       x = "Activity",
       y = "Number of Respondents",
       fill = "Frequency Count") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  coord_flip()
print(stress_plot)
```



```
#ggplotly(stress_plot)
```

Managing stress is another important concern. Throughout the pandemic, students found solace in activities like listening to music and playing online video games, which ranked among their top five ways to cope with stress. These activities played a crucial role in helping students navigate these challenging times by providing relaxation and stress relief.

```
# Top 10 Things the Respondent Miss the Most
miss_plot <- df %>%
  group_by(Most_Missed_Things) %>%
  summarise(freqCount = n()) %>%
  top_n(10, freqCount) %>%
  ggplot(aes(x=reorder(Most_Missed_Things, freqCount),  y=freqCount, fill=freqCount)) +
  geom_bar(color="black", stat="identity") +
  labs(title = "Distribution of Respondents by Things they Miss the Most (Top 10)",
       x = "Object/ Subject",
       y = "Number of Respondents",
       fill = "Frequency Count") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  coord_flip()
print(miss_plot)
```

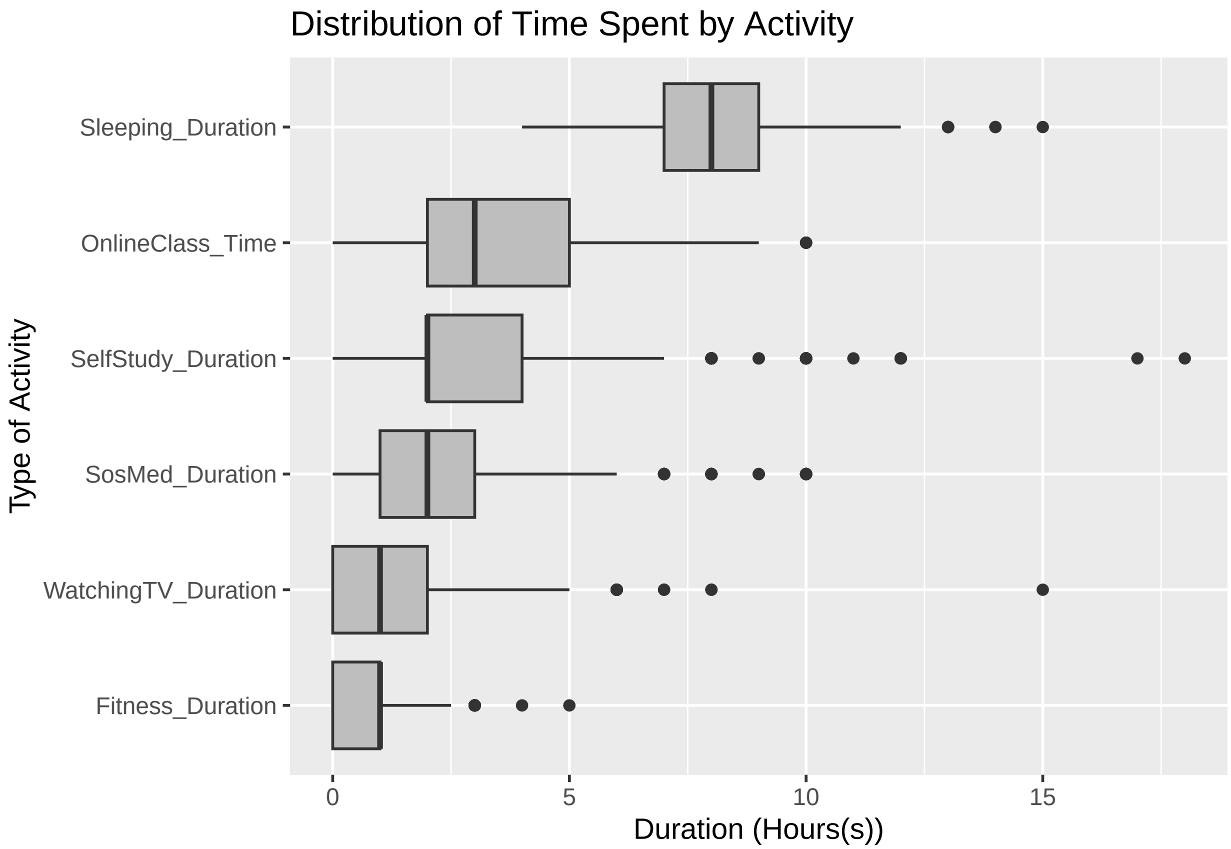


```
#ggplotly(miss_plot)
```

Students expressed longing for aspects of their lives before the pandemic, such as attending school and college and spending time with friends and family. These factors highlight the routines and connections students valued most and illustrate how the pandemic disrupted their social and academic environments. The lack of face-to-face interactions and structured schedules had a significant impact on their daily lives.

```
# store the column name of variables related to duration
duration_cols <- c('OnlineClass_Time', 'SelfStudy_Duration', 'Fitness_Duration',
                   'Sleeping_Duration', 'SosMed_Duration', 'WatchingTV_Duration')

# calculate mean duration for each activity
mean_durations <- sapply(duration_cols, function(col) mean(df[[col]], na.rm = TRUE)) # a function which
names(mean_durations) <- duration_cols # labeling the vector

# create a long-format data frame
df_long <- stack(df[, duration_cols]) # stack the duration_cols -> resulting a data frame with 2 column
names(df_long) <- c("Duration", "Activity")

# reorder activities by mean of duration
ordered_activities <- names(sort(mean_durations))
df_long$Activity <- factor(df_long$Activity, levels = ordered_activities)

# write.table(df_long, file="df_long.csv", col.names = TRUE, row.names = FALSE, sep=",")
```

```r
# create the box plot
duration_plot <- ggplot(df_long, aes(y = Activity, x = Duration)) +
  geom_boxplot(fill="gray") +
  labs(title = "Distribution of Time Spent by Activity",
       x = "Duration (Hours(s))",
       y = "Type of Activity")

print(duration_plot)
```



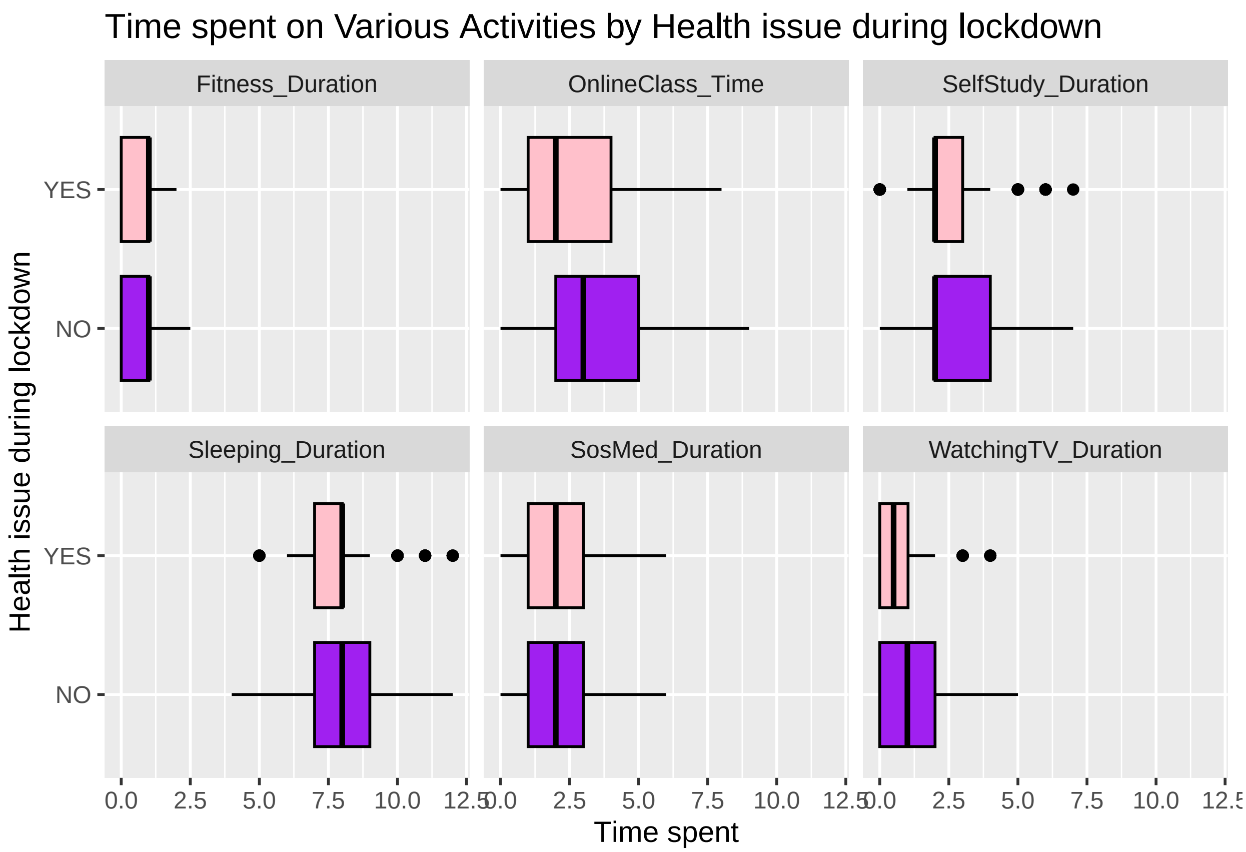
Looking into how students used their time uncovered some concerning patterns. Exercise was the least prioritized activity, suggesting a lack of physical activity. This lack can lead to lasting negative effects on both physical and mental well-being. It’s crucial to encourage children to incorporate more physical activity into their daily schedules.

```r
# remove outliers in duration column based on the outlier we found in the boxplot
for (i in duration_cols)
{
  outliers <- boxplot.stats(df[[i]])$out # identify outliers
  df <- df[!df[[i]] %in% outliers, ] # remove the whole row if it contains outliers
}

str(df)
```

```
## ’data.frame':    1048 obs. of  18 variables:
##  $ Region                                 : Factor w/ 2 levels "Delhi-NCR","Outside Delhi-NCR": 1 1
##  $ Age                                    : num  21 21 20 20 21 21 19 19 21 20 ...
##  $ OnlineClass_Time                       : num  2 0 7 3 3 0 2 2 3 0 ...
##  $ OnlineClass_Rating                     : Factor w/ 5 levels "Very poor","Poor",..: 4 5 1 1 4 1 1
##  $ OnlineClass_Medium                     : Factor w/ 5 levels "Laptop/Desktop",..: 1 2 1 2 1 2 2 3
##  $ SelfStudy_Duration                     : num  4 0 3 2 3 6 2 1 4 1 ...
##  $ Fitness_Duration                       : num  0 2 0 1 1 0 1 1 1 0.5 ...
##  $ Sleeping_Duration                      : num  7 10 6 6 8 5 5 10 8 8 ...
##  $ SosMed_Duration                        : num  3 3 2 5 3 1 4 5 2 5 ...
##  $ SosMed_Medium                          : Factor w/ 13 levels "Linkedin","Youtube",..: 1 2 1 3 3 2
##  $ WatchingTV_Duration                    : num  1 0 0 0 1 0 0 0 1 3 ...
##  $ NumberOfMeals                          : num  4 3 3 3 4 1 3 3 3 3 ...
##  $ WeightChange                           : Factor w/ 3 levels "Decreased","Remain Constant",..: 3 1
##  $ Having_HealthIssue                     : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 2 1 2 ...
##  $ Stress_Busters                         : Factor w/ 86 levels "Cooking","Scrolling through social
##  $ Time_Utilized                          : Factor w/ 2 levels "NO","YES": 2 2 1 1 1 1 1 1 1 1 ...
##  $ Connected_with_Family_Friends_Relatives: Factor w/ 2 levels "NO","YES": 2 1 2 1 1 2 2 2 1 1 ...
##  $ Most_Missed_Things                     : Factor w/ 36 levels "School/college",..: 1 2 3 4 3 1 4 5
```

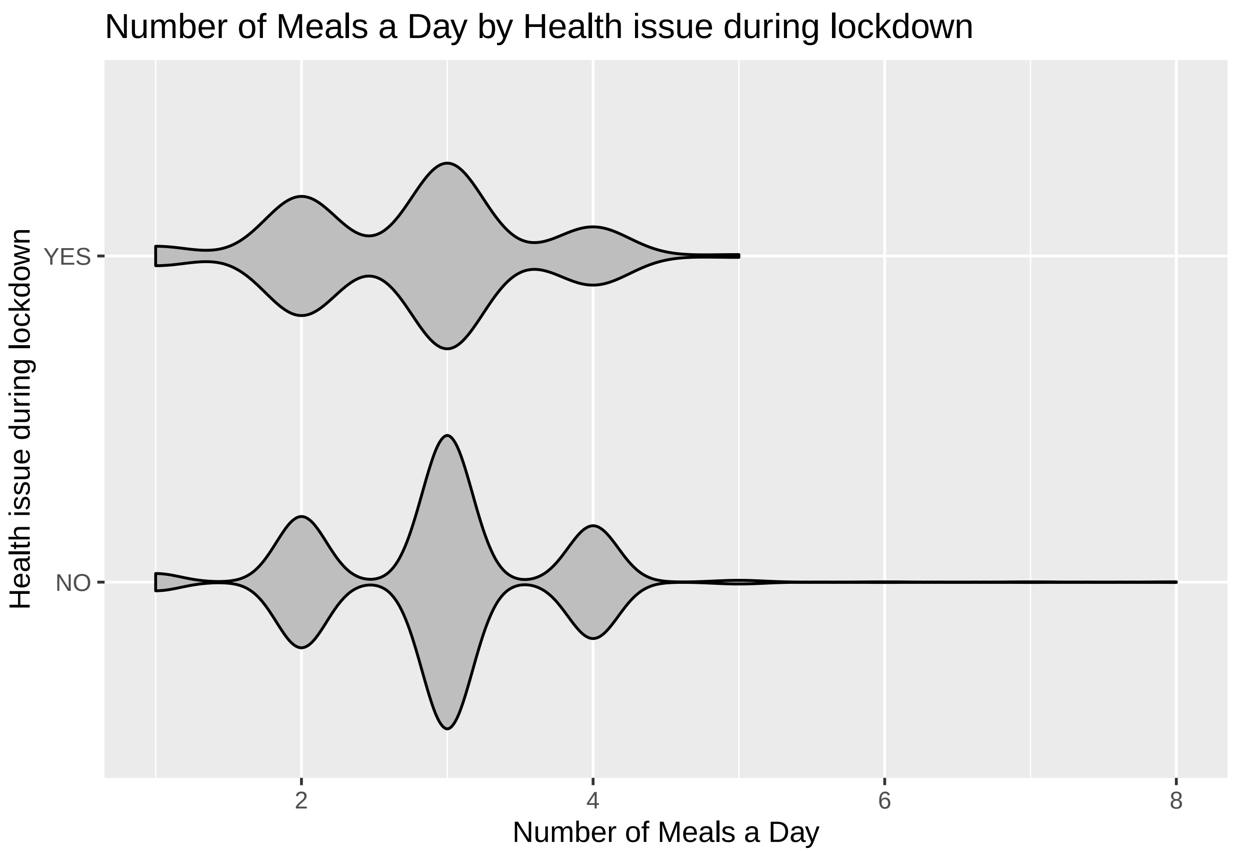
Now our data is reduced to 1048 observations

```
# reshape the data into a long format
df_long <- df %>%
  pivot_longer(
    cols = c(OnlineClass_Time, SelfStudy_Duration, Fitness_Duration, Sleeping_Duration, SosMed_Duration
    names_to = "Activity",
    values_to = "Duration"
  )

p_combined <- ggplot(df_long, aes(x = Having_HealthIssue, y = Duration, fill = Having_HealthIssue)) +
  geom_boxplot(color="black") +
  coord_flip() +
  labs(title="Time spent on Various Activities by Health issue during lockdown",
       x="Health issue during lockdown",
       y="Time spent") +
  facet_wrap(~ Activity) +
  scale_fill_manual(values = c("YES" = "pink", "NO" = "purple")) +
  theme(legend.position = "none")

print(p_combined)
```



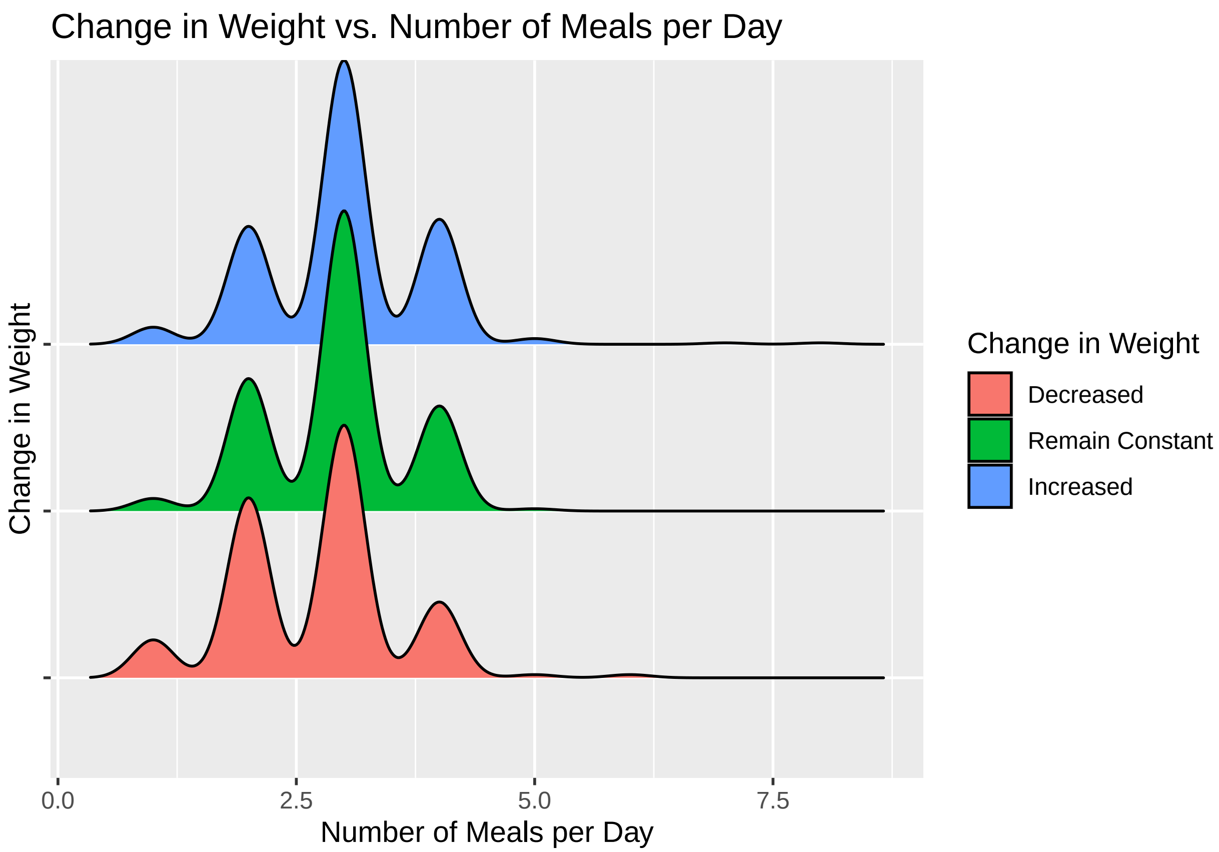
The study did not establish a straightforward connection between activity levels and health issues, suggesting that multiple factors influence students' health, such as mental well-being, diet, and pre-existing medical conditions. To comprehensively grasp these complex dynamics, a thorough approach is necessary.

```
meal_health_plot <- ggplot(df, aes(x = Having_HealthIssue, y = NumberOfMeals)) +
  geom_violin(color="black", fill="gray") +
  coord_flip() +
  labs(title="Number of Meals a Day by Health issue during lockdown",
       x="Health issue during lockdown",
       y="Number of Meals a Day") +
  scale_fill_brewer(palette="Set1") + # set color palette to fill the graph
  theme(legend.position = "none")
print(meal_health_plot)
```



- Majority of the respondents eat 3 times per day, followed by 2 times and 4 times per day
- Respondents with no health issue tend to eat more, than the one who's health issue

```
weight_change_plot <- ggplot(df, aes(x=NumberOfMeals, y=WeightChange, fill=WeightChange)) +
  geom_density_ridges() +
  labs(title = "Change in Weight vs. Number of Meals per Day",
       x = "Number of Meals per Day",
       y = "Change in Weight",
       fill = "Change in Weight") +
  theme(axis.text.y = element_blank()) # hide the y-axis because it isn't used in this graph

print(weight_change_plot)
```



- Respondents who eat less than 3 times a day tend to lose weight Respondents who eat more than 3 times a day tend to gain weight
- Most respondents who eat 3 times a day maintain a stable weight.

Next, we'll look for anomalies

```
FindOutliers(df$NumberOfMeals)$summary
```

```
##        method    n nMiss nOut    lowLim    upLim minNom maxNom
## 1  ThreeSigma 1048     0    3 0.5099625 5.306831      1      5
## 2      Hampel 1048     0  505 3.0000000 3.000000      3      3
## 3 BoxplotRule 1048     0   45 1.5000000 4.500000      2      4
```

The first column lists the name of the outlier detection method used. The second column shows the total number of observations, and the third column indicates how many observations are missing (recorded as NA). The fourth column displays the number of outliers detected by each method. The next two columns present the lower and upper limits used to identify outliers. The final two columns show the lower and upper boundaries of the data values that are not considered outliers.

For example, the three-sigma rule identified only 3 outliers, while the Hampel method found 505 outliers, and the boxplot rule identified 45 outliers.

```
FindOutliers(df$OnlineClass_Time)$summary
```

```
##          method    n nMiss nOut    lowLim    upLim minNom maxNom
## 1  ThreeSigma 1048     0    0 -2.879921 9.336696    0.0      9
## 2      Hampel 1048     0   31 -1.447800 7.447800    0.0      7
## 3 BoxplotRule 1048     0  107  0.500000 9.500000    0.5      9
```

It's evident that the three-sigma rule detected no outliers, whereas the Hampel method identified 31 outliers, and the boxplot rule flagged 107 outliers.

```r
FindOutliers(df$SelfStudy_Duration)$summary
```

```
##          method    n nMiss nOut    lowLim    upLim minNom maxNom
## 1  ThreeSigma 1048     0    0 -2.079668 7.337301      0      7
## 2      Hampel 1048     0   13 -2.447800 6.447800      0      6
## 3 BoxplotRule 1048     0   72  1.000000 7.000000      1      7
```

It's clear that according to the analysis: * The three-sigma rule found 0 outliers * The Hampel method identified 13 outliers * The boxplot rule detected 72 outliers

```r
FindOutliers(df$Fitness_Duration)$summary
```

```
##          method    n nMiss nOut    lowLim    upLim minNom maxNom
## 1  ThreeSigma 1048     0    0 -1.179393 2.620614      0    2.5
## 2      Hampel 1048     0    0 -1.223900 3.223900      0    2.5
## 3 BoxplotRule 1048     0    0 -0.500000 2.500000      0    2.5
```

It's apparent that all the rules identified 0 outliers.

```r
FindOutliers(df$Sleeping_Duration)$summary
```

```
##          method    n nMiss nOut   lowLim    upLim minNom maxNom
## 1  ThreeSigma 1048     0    0 3.233145 12.46304      4     12
## 2      Hampel 1048     0    0 3.552200 12.44780      4     12
## 3 BoxplotRule 1048     0   41 6.000000 12.00000      6     12
```

We can observe that both the three-sigma and Hampel rules identified 0 outliers, while the boxplot rule flagged 41 outliers.

```r
FindOutliers(df$SosMed_Duration)$summary
```

```
##          method    n nMiss nOut    lowLim    upLim minNom maxNom
## 1  ThreeSigma 1048     0    0 -1.913046 6.248638      0      6
## 2      Hampel 1048     0    0 -2.447800 6.447800      0      6
## 3 BoxplotRule 1048     0    0  0.000000 6.000000      0      6
```

It's apparent that all the rules identified 0 outliers.

```r
FindOutliers(df$WatchingTV_Duration)$summary
```

```
##          method    n nMiss nOut    lowLim   upLim minNom maxNom
## 1  ThreeSigma 1048     0   11 -2.204534 4.13603      0      4
## 2      Hampel 1048     0    0 -3.447800 5.44780      0      5
## 3 BoxplotRule 1048     0    0 -1.000000 5.00000      0      5
```

According to the three-sigma rule, there are 11 outliers detected, while neither the Hampel nor the boxplot rules found any outliers.

Now, we'll proceed to remove the outliers identified by the three-sigma rule.

```
# function to remove outliers using three-sigma rule: data outside the range of mean +-3 × standard dev
remove_outliers_3sigma <- function(df, col_name) {
  mean_val <- mean(df[[col_name]], na.rm = TRUE) # calculate mean with NA value removed
  sd_val <- sd(df[[col_name]], na.rm = TRUE) # calculate std with NA value removed

  lower_limit <- mean_val - 3 * sd_val
  upper_limit <- mean_val + 3 * sd_val

  # keep the rows where the values in the specified column (col_name) are within the lower and upper li
  df_filtered <- df[df[[col_name]] >= lower_limit & df[[col_name]] <= upper_limit, ] #buat ngasi syarat

  return(df_filtered)
}

# remove outliers using our function
df <- remove_outliers_3sigma(df, "NumberOfMeals")
df <- remove_outliers_3sigma(df, "NumberOfMeals")
df <- remove_outliers_3sigma(df, "OnlineClass_Time")
df <- remove_outliers_3sigma(df, "SelfStudy_Duration")
df <- remove_outliers_3sigma(df, "Fitness_Duration")
df <- remove_outliers_3sigma(df, "Sleeping_Duration")
df <- remove_outliers_3sigma(df, "SosMed_Duration")
df <- remove_outliers_3sigma(df, "WatchingTV_Duration")
```

5. Statistical Analysis

Numerical Variables

1. Shapiro-Wilk Test for Normality

Before conducting the Pearson correlation, we performed the Shapiro-Wilk test to assess the normality of the numerical variables. None of the variables passed the normality test (all p-values < 0.05), indicating that the data does not follow a normal distribution. Therefore, Pearson correlation, which requires normality, cannot be applied in this case.

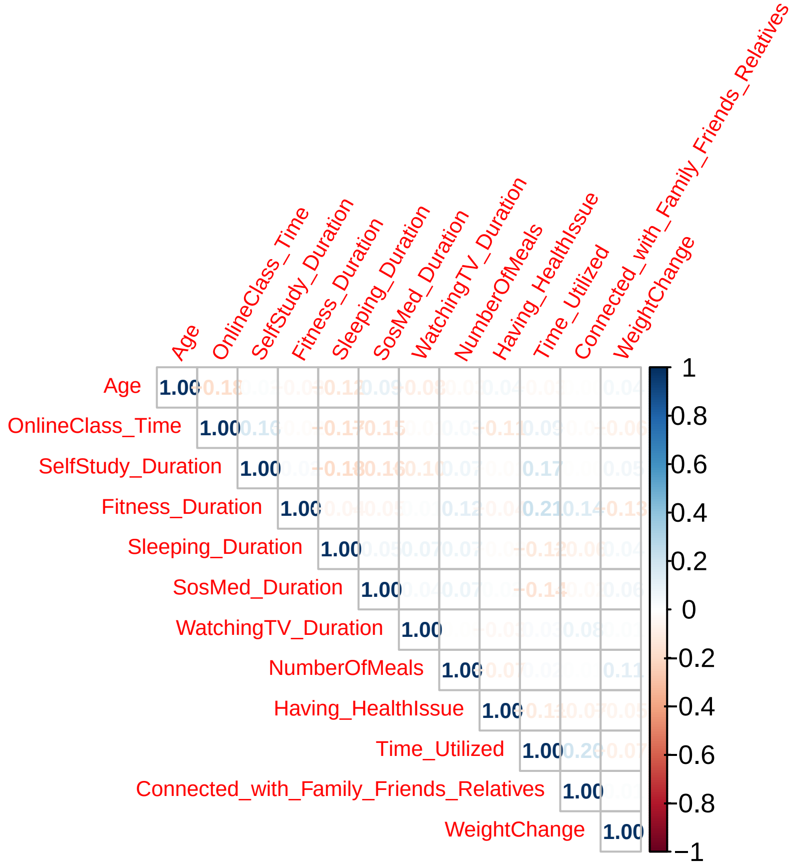
```
col_names = c()
sph_test = c()
for (i in num_cols)
{
  col_names <- c(col_names, i)
  ans <- shapiro.test(df[[i]])$p.value
  sph_test <- c(sph_test, ans)
}
sph_df = data.frame(Name = col_names, PVal = sph_test)
sph_df[order(sph_df$PVal),]
```

```
##                    Name         PVal
## 4    Fitness_Duration 6.107004e-34
## 1                Age 3.445997e-32
## 7 WatchingTV_Duration 1.407972e-31
## 8       NumberOfMeals 9.956070e-30
## 6     SosMed_Duration 3.087665e-25
## 5   Sleeping_Duration 6.143239e-21
## 3  SelfStudy_Duration 1.334916e-20
## 2    OnlineClass_Time 3.777746e-16
```

None of the variables meet the criteria for normal distribution, as all p-values are less than 0.05. Therefore, we can conclude that our numerical data does not follow a normal distribution, which means we cannot use Pearson correlation.

2. Correlation Analysis

```r
num_df <- subset(df,select = num_cols) #update the num_df after removing some of the observations

# convert YES/NO factor columns to numeric (1 for YES, 0 for NO)
num_df$Having_HealthIssue <- as.numeric(df$Having_HealthIssue == "YES")
num_df$Time_Utilized <- as.numeric(df$Time_Utilized == "YES")
num_df$Connected_with_Family_Friends_Relatives <- as.numeric(df$Connected_with_Family_Friends_Relatives

# convert WeightChange to numeric based on mappings
num_df$WeightChange <- ifelse(df$WeightChange == "Increased", 2,
                          ifelse(df$WeightChange == "Decreased", 0, 1))

c <- cor(num_df)
# View(c)
corrplot(c, type="upper", method="number",
         number.cex=0.7,  # font size of correlation numbers
         tl.cex=0.7,      # font size of column/row names
         tl.srt=60)
```



- Calculated the correlation matrix (cor) for numerical variables
- Visualized correlations using corrplot to analyze relationships between variables.

```
# check correlation of having health issue category with quantitative data which is duration of differe
for (i in duration_cols)
{
  cat(i,": ",cor(num_df[[i]], num_df$Having_HealthIssue),"\n")
}
```

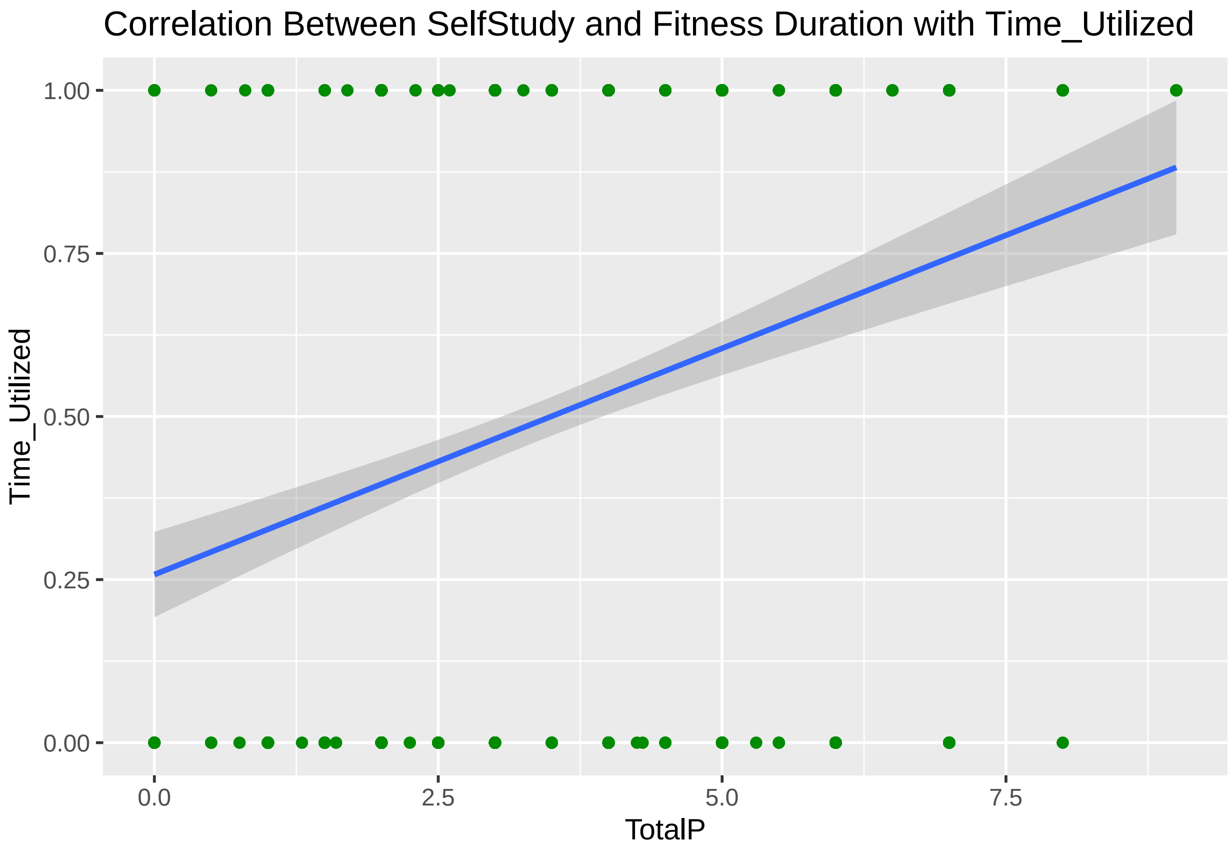
```
## OnlineClass_Time :  -0.113825
## SelfStudy_Duration :  -0.01142817
## Fitness_Duration :  -0.04178206
## Sleeping_Duration :  -0.0176028
## SosMed_Duration :  0.01514649
## WatchingTV_Duration :  -0.03181608
```

There tends to be a mild correlation between the time spent on different activities and the likelihood of experiencing health issues. Longer durations spent on online classes, fitness activities, and watching TV show slight negative associations with health problems. Conversely, spending more time on social media shows a negligible positive correlation with health issues. These correlations, however, are generally weak, suggesting that individual activities may not strongly predict the likelihood of having health issues. It's likely that other factors not measured in this study also influence these relationships.

```
# compute the total time for productive and non productive activity
num_df$TotalP = num_df$SelfStudy_Duration + num_df$Fitness_Duration
num_df$TotalN = num_df$SosMed_Duration + num_df$WatchingTV_Duration
# View(num_df)

# compute correlation
correlations <- cor(num_df[, c("TotalP", "TotalN", "Time_Utilized")])
print(correlations)
```

```
##                   TotalP      TotalN Time_Utilized
## TotalP         1.0000000 -0.18156524    0.23649549
## TotalN        -0.1815652  1.00000000   -0.09479337
## Time_Utilized  0.2364955 -0.09479337    1.00000000
```

```
ggplot(num_df, aes(x = TotalP,
                    y = Time_Utilized)) +
  geom_point(color="green4") +
  labs(title = "Correlation Between SelfStudy and Fitness Duration with Time_Utilized")+
  geom_smooth(method="lm")
```

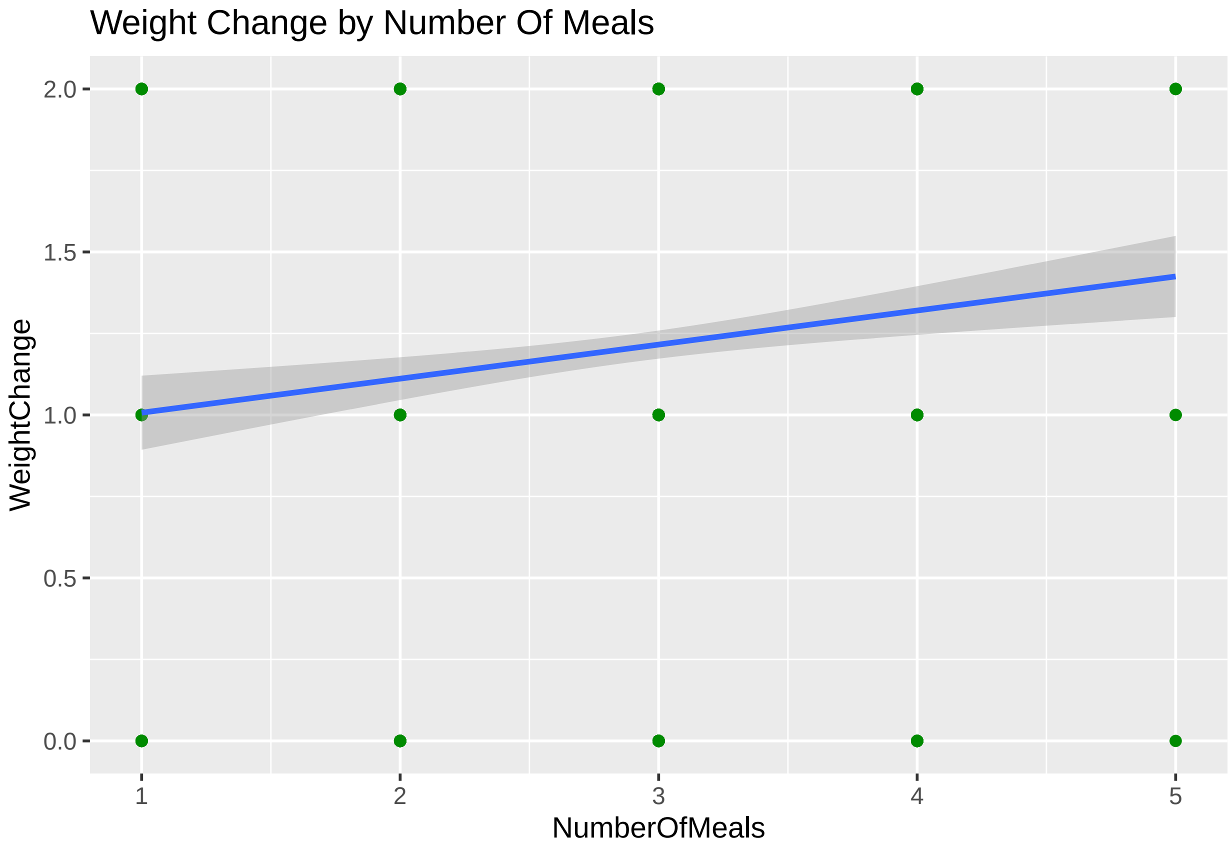


As the combined time spent on productive activities like self-study and fitness increases, there is a slight tendency for the overall time utilized to also increase.

```r
correlations <- cor(num_df[, c("NumberOfMeals", "WeightChange")])
print(correlations)
```

```
##               NumberOfMeals WeightChange
## NumberOfMeals     1.0000000    0.1141691
## WeightChange      0.1141691    1.0000000
```

```r
ggplot(num_df, aes(x = NumberOfMeals,
                     y = WeightChange)) +
  geom_point(color="green4") +
  labs(title = "Weight Change by Number Of Meals")+
  geom_smooth(method="lm")
```



This suggests a weak positive relationship: as the number of meals consumed increases, there is a slight tendency for weight change to also increase.

Categorical Variables

1. Chi-Square Test

- Chi-square tests were performed between all pairs of categorical variables to examine their association.
- The p-value for each pair was calculated using the chisq.test function to assess the significance level of 0.05. A p-value less than 0.05 suggests rejection of the null hypothesis (indicating variables are independent).

```
col_pair = c()
chisq_pv = c()
# harus dikasi ( ) klo ga nanti error
for (i in 2:(length(cat_cols)-1))
{
  for (j in (i+1):length(cat_cols))
  {
    col_pair <- c(col_pair,paste(cat_cols[i],",", cat_cols[j]))
    ans <- chisq.test(df[[cat_cols[i]]], df[[cat_cols[j]]])$p.value
    chisq_pv <- c(chisq_pv, ans)
    # print(chisq.test(df[[cat_cols[i]]], df[[cat_cols[j]]])$p.value)
  }
}

chi_df = data.frame(Name = col_pair, PVal = chisq_pv)
head(chi_df[order(chi_df$PVal),], n=10)
```

```
##                                                             Name          PVal
## 33                         Stress_Busters , Most_Missed_Things 3.969857e-274
## 18                              SosMed_Medium , Stress_Busters 6.805568e-233
## 8                      OnlineClass_Rating , Most_Missed_Things  1.042444e-63
## 5                          OnlineClass_Rating , Stress_Busters  1.740276e-33
## 21                          SosMed_Medium , Most_Missed_Things  1.490637e-22
## 15                     OnlineClass_Medium , Most_Missed_Things  5.783002e-21
## 6                           OnlineClass_Rating , Time_Utilized  3.927739e-17
## 34     Time_Utilized , Connected_with_Family_Friends_Relatives  4.941566e-10
## 12                         OnlineClass_Medium , Stress_Busters  2.114716e-07
## 7  OnlineClass_Rating , Connected_with_Family_Friends_Relatives  6.117431e-06
```

Some variables show a strong connection because we reject the null hypothesis (which states the variables are independent). In statistical terms, we reject the null hypothesis when the p-value is less than 0.05. This indicates that these two variables are not independent in the population from which this sample was taken.

2. Association Statistics

- We calculated association statistics such as the Chi-square test statistic, Cramér's V, and contingency coefficients for specific pairs of variables.

```
tbl1 <- table(df$OnlineClass_Rating, df$Time_Utilized) # creates a contingency table: a frequency table
print(PercTable(tbl1))
```

```
##
##                      NO   YES
##
## Very poor   freq    239   134
##             perc  23.1% 13.0%
##
## Poor        freq     14    11
##             perc   1.4%  1.1%
##
## Average     freq    174   162
##               perc 16.8% 15.7%
##
## Good         freq    89   126
##               perc  8.6% 12.2%
##
## Excellent    freq    11    74
##               perc  1.1%  7.2%
##
```

```
print(PercTable(tbl1, margin=1)) # sum of percentage per row
```

```
##
##                     NO   YES   Sum
##
## Very poor    freq   239   134   373
##               perc 23.1% 13.0% 36.1%
##
## Poor         freq    14    11    25
##               perc  1.4%  1.1%  2.4%
##
## Average      freq   174   162   336
##               perc 16.8% 15.7% 32.5%
##
## Good         freq    89   126   215
##               perc  8.6% 12.2% 20.8%
##
## Excellent    freq    11    74    85
##               perc  1.1%  7.2%  8.2%
##
```

```
print(PercTable(tbl1, margin=2)) # sum of percentage per column
```

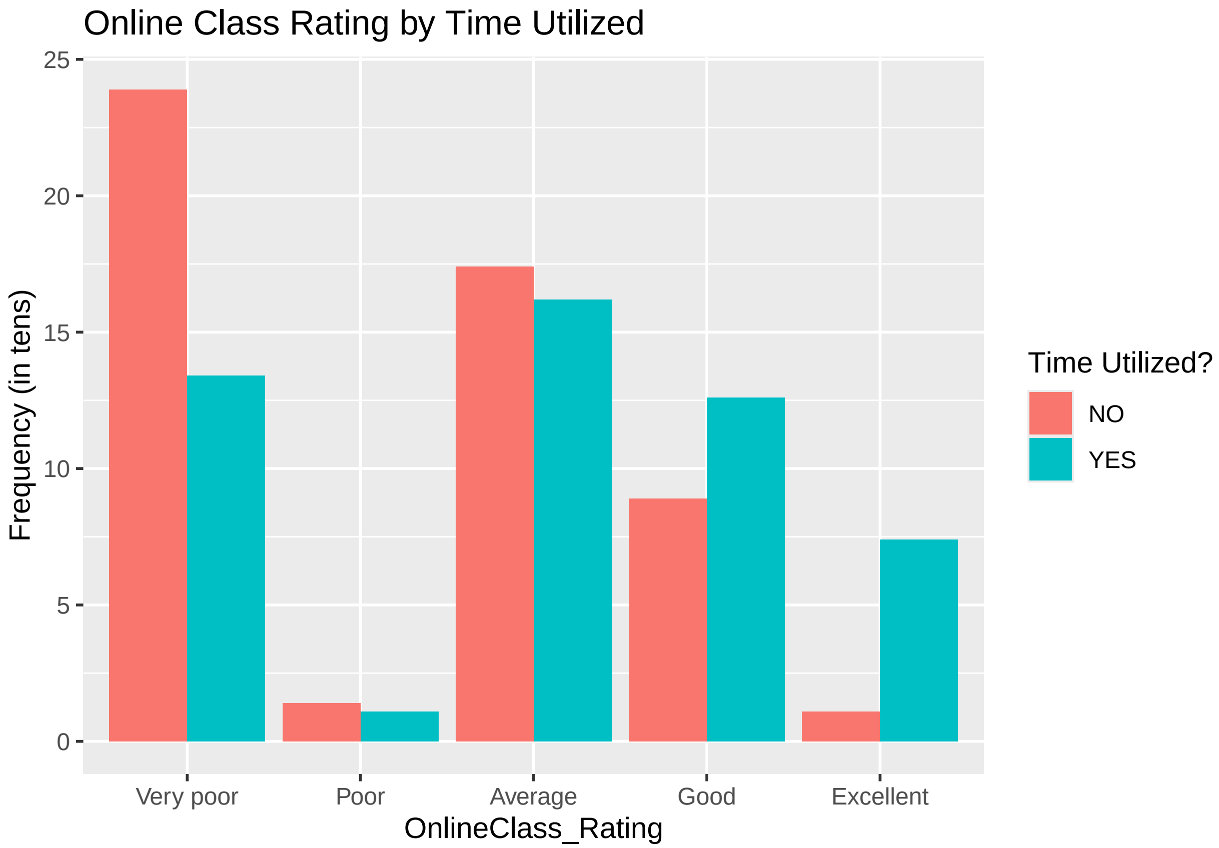
```
##
##                     NO   YES
##
## Very poor    freq   239   134
##               perc 23.1% 13.0%
##
## Poor         freq    14    11
##               perc  1.4%  1.1%
##
## Average      freq   174   162
##               perc 16.8% 15.7%
##
## Good         freq    89   126
##               perc  8.6% 12.2%
##
## Excellent    freq    11    74
##               perc  1.1%  7.2%
##
## Sum          freq   527   507
##               perc 51.0% 49.0%
##
```

Respondents who spent more time utilizing their time generally gave more positive ratings compared to those who spent less time.

```
# computes a variety of association statistics for the contingency table, such as Chi-square test stati
assocstats(tbl1)
```

```
##                       X^2 df P(> X^2)
## Likelihood Ratio 89.102  4        0
## Pearson          83.052  4        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.273
## Cramer's V        : 0.283
```

```
tbl1_plt = ggplot(df, aes(x = OnlineClass_Rating, fill = Time_Utilized)) +
  geom_bar(position = position_dodge(preserve = "single")) + # add a bar plot where bars of different c
  labs(y = "Frequency (in tens)",
       title = "Online Class Rating by Time Utilized",
       fill = "Time Utilized?") +
  scale_y_continuous(labels = function(x) x / 10)
print(tbl1_plt)
```



There are no strict rules for interpreting Cramer's V values, but as a guide: weak (0 to 0.2), moderate (0.2 to 0.3), strong (0.3 to 0.5), and redundant (0.5 to 0.99, indicating the variables likely measure the same concept).

In this case, there is a moderate association between the online class rating and time utilized.

```
tbl2 <- table(df$Connected_with_Family_Friends_Relatives, df$Time_Utilized)
print(PercTable(tbl2))
```

```
##
##                NO   YES
##
## NO    freq   201   103
##       perc 19.4% 10.0%
##
## YES   freq   326   404
##       perc 31.5% 39.1%
##
```

```
print(PercTable(tbl2, margin=1))
```

```
##
##                NO   YES   Sum
##
## NO    freq   201   103   304
##       perc 19.4% 10.0% 29.4%
##
## YES   freq   326   404   730
##       perc 31.5% 39.1% 70.6%
##
```

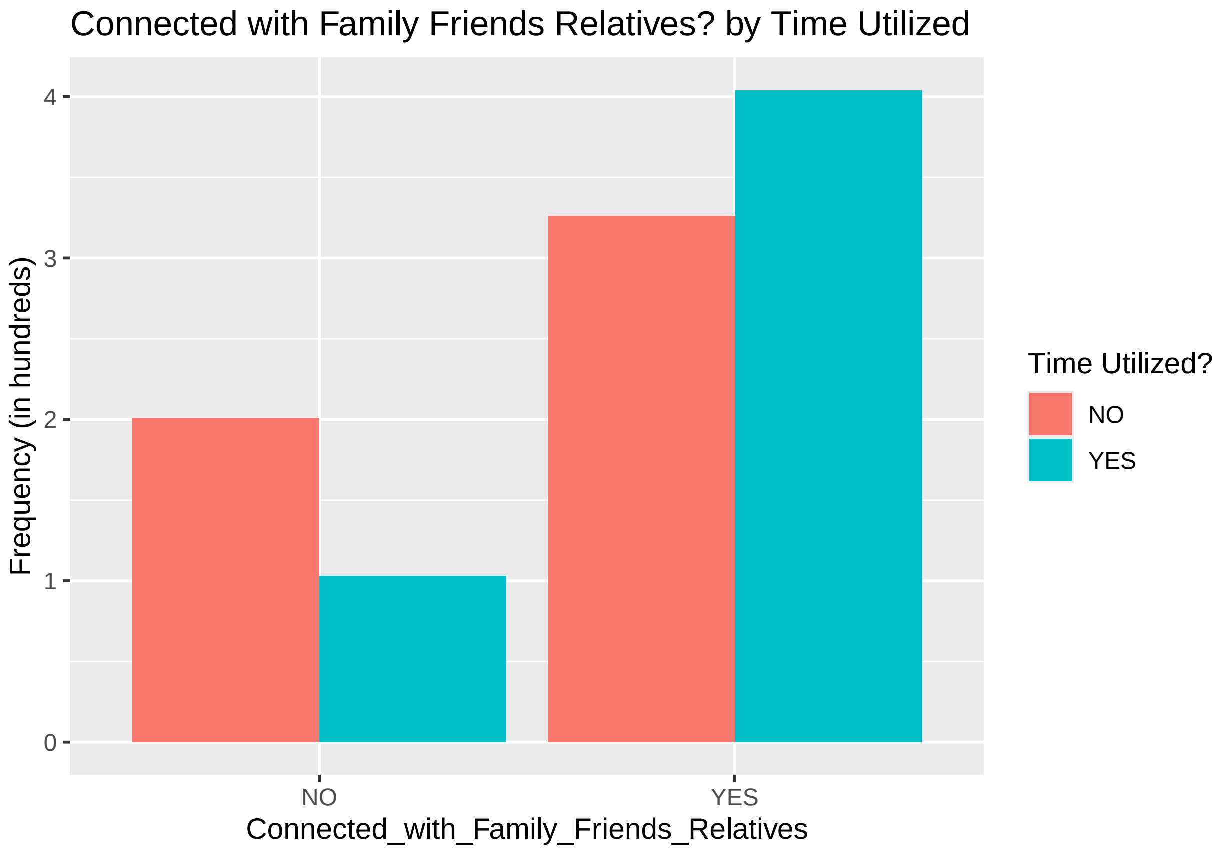
```
print(PercTable(tbl2, margin=2))
```

```
##
##                NO   YES
##
## NO    freq   201   103
##       perc 19.4% 10.0%
##
## YES   freq   326   404
##       perc 31.5% 39.1%
##
## Sum   freq   527   507
##       perc 51.0% 49.0%
##
```

```
assocstats(tbl2)
```

```
##                       X^2 df   P(> X^2)
## Likelihood Ratio 40.127  1 2.3801e-10
## Pearson          39.554  1 3.1906e-10
##
## Phi-Coefficient   : 0.196
## Contingency Coeff.: 0.192
## Cramer’s V        : 0.196
```

```r
tbl2_plt = ggplot(df, aes(x = Connected_with_Family_Friends_Relatives, fill = Time_Utilized)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(y = "Frequency (in hundreds)",
       title = "Connected with Family Friends Relatives? by Time Utilized",
       fill = "Time Utilized?") +
  scale_y_continuous(labels = function(x) x / 100)
print(tbl2_plt)
```



The findings show a connection between effective time management and stronger family bonds. Students who managed their time well reported closer relationships with their families. This is important as it highlights how good time management can enhance personal connections during challenging circumstances.

6. Discussion

A few notable patterns and trends emerged from the investigation, providing valuable insights into the broader impacts of the pandemic.

1. **Time Management Challenges**: During the pandemic, a significant number of respondents (51.44%) reported struggling with time management. This highlights the difficulty many face in effectively organizing their schedules for online learning.
2. **Health Concerns**: Approximately 13.62% of students mentioned experiencing health issues, indicating the pandemic's considerable impact on their mental and physical well-being. This aligns with a global trend of prolonged lockdowns and lifestyle changes contributing to increased health worries.

3. **Activity Preferences**: Students ranked exercise as their least favored activity when dividing their time among other activities. This preference for exercise could have adverse long-term effects on their overall health.
4. **Family Connections**: Effective time management was positively associated with stronger family ties. This suggests that improving time management skills can help students maintain closer relationships with their families, providing crucial emotional support during difficult times.
5. **Stress Management**: Listening to music and playing online video games emerged as popular methods for stress relief. These activities served as essential coping strategies to manage the stress and uncertainties brought by the pandemic.
6. **No Direct Link Between Activity and Health**: The absence of a clear correlation between activity levels and health issues indicates that various factors like nutrition, mental health, and existing health conditions influence students' well-being.

7. Conclusion

The study reveals no direct link between students' health and the time they spend on activities such as self-study and online courses. It's concerning that more than half of the participants did not effectively manage their time during lockdowns, potentially impacting their health negatively. Addressing this issue involves supporting students in learning online and improving their time management skills. Additionally, the study highlights a lack of physical exercise among participants, which is critical for their overall well-being during these challenging times. Encouraging physical activity should be a priority to enhance their overall health.

8. References

Kunal Chaturvedi, Dinesh Vishwakarma, Nidhi Singh. (2020). COVID-19 and its impact on education, social life and mental health of students: A Survey. Children and Youth Services Review, 121. https://doi.org/10.1016/j.childyouth.2020.105866

Kunal Chaturvedi. (2020). COVID-19 and its impact on students. Kaggle. Available at: https://www.kaggle.com/datasets/kunal28chaturvedi/covid19-and-its-impact-on-students/data