Vahid Monfared
 # Machine Learning #
 # Data source: https://www.kaggle.com/datasets/steve1215rogg/student-lifestyle-dataset?resource=download #
 # Title: student lifestyle dataset, Daily Lifestyle and Academic Performance of Students #

Describe a research scenario and specify a research question based on data analytic methods that we learned in class. For example, methods such as *one and two sample mean tests, correlation tests, simple and multiple linear regression, ANOVA and ANCOVA, one and two-sample tests for proportions, and logistic regression* are fair game. Perform your analysis, and then report your results and conclusions.

This project has two parts, first original dataset including both scalar and ordinal categorical variables, and then with no ordinal categorical variable.

##Cleaning the data##

Clean up your data and randomly sample 1000 observations from it if your data set is large: Done in R.

```
> ## Step _ 0 ##
> ## Exploratory Data Analysis (EDA) ##
> ##  Clean up your data and randomly sample 1000 observations from it if your data set is large ##
> # Set working directory (adjust as necessary)
> setwd("C:/Users/vahid/OneDrive/Documents/CS_555/Final_Project")
> # Load the necessary library
> library(readxl)
> # Load the dataset
> data <- read.csv("CS_555_Final_Project_dataset.xlsx.csv")
> # Display the original shape of the dataset
> cat("Original shape of the dataset:", nrow(data), "rows and", ncol(data), "columns\n")
Original shape of the dataset: 2000 rows and 8 columns
> # Clean the dataset (example: remove rows with missing values)
> cleaned_data <- na.omit(data)
> # Randomly sample 1000 observations
> set.seed(123) # For reproducibility
> sampled_data <- cleaned_data[sample(1:nrow(cleaned_data), 1000), ]
> # Display the shape of the sampled dataset
> cat("Shape of the sampled dataset:", nrow(sampled_data), "rows and", ncol(sampled_data), "columns\n")
Shape of the sampled dataset: 1000 rows and 8 columns
> # Display the first few rows of the sampled dataset as a dataframe
```

```
> head(sampled_data)
    Student_ID Study_Hours_Per_Day Extracurricular_Hours_Per_Day Sleep_Hours_Per_Day Social_Hours_Per_Day Physical_Activity_Hours_Per_Day  GPA
415        415                8.5                          2.1                6.5                1.3                           5.6 3.38
463        463                6.4                          3.6                5.6                5.7                           2.7 3.10
179        179                9.9                          2.1                5.9                1.6                           4.5 3.52
526        526                6.3                          1.4                9.6                1.7                           5.0 2.74
195        195                8.5                          3.7                5.9                3.4                           2.5 3.06
938        938                9.6                          0.6                9.8                0.8                           3.2 3.29
    Stress_Level
415         High
463         High
179         High
526     Moderate
195         High
938         High
> data <- cleaned_data[sample(1:nrow(cleaned_data), 1000), ]
> dim(data)
[1] 1000    8
> data <- cleaned_data[sample(1:nrow(cleaned_data), 1000), ]
> dim(data)
[1] 1000    8
> colnames(data)
[1] "Student_ID"                     "Study_Hours_Per_Day"            "Extracurricular_Hours_Per_Day"  "Sleep_Hours_Per_Day"
[5] "Social_Hours_Per_Day"           "Physical_Activity_Hours_Per_Day" "GPA"                            "Stress_Level"
> data <- data[, !names(data) %in% "Student_ID"]
> colnames(data)
[1] "Study_Hours_Per_Day"            "Extracurricular_Hours_Per_Day"  "Sleep_Hours_Per_Day"            "Social_Hours_Per_Day"
[5] "Physical_Activity_Hours_Per_Day" "GPA"                            "Stress_Level"
> dim(data)
[1] 1000    7
```

0. General Information _ Big Picture (with keeping stress level)

Research Question: We aim to assess whether the variables (Study_Hours_Per_Day, Extracurricular_Hours_Per_Day, Sleep_Hours_Per_Day, Social_Hours_Per_Day, Physical_Activity_Hours_Per_Day, and GPA) significantly differ across the three levels of the categorical variable Stress_Level (Low, Moderate, and High) using one-way ANOVA. Is there a significant difference in the means of the continuous variables (Study_Hours_Per_Day, Extracurricular_Hours_Per_Day, Sleep_Hours_Per_Day, Social_Hours_Per_Day, Physical_Activity_Hours_Per_Day, and GPA) across the three stress level groups (Low, Moderate, and High)?

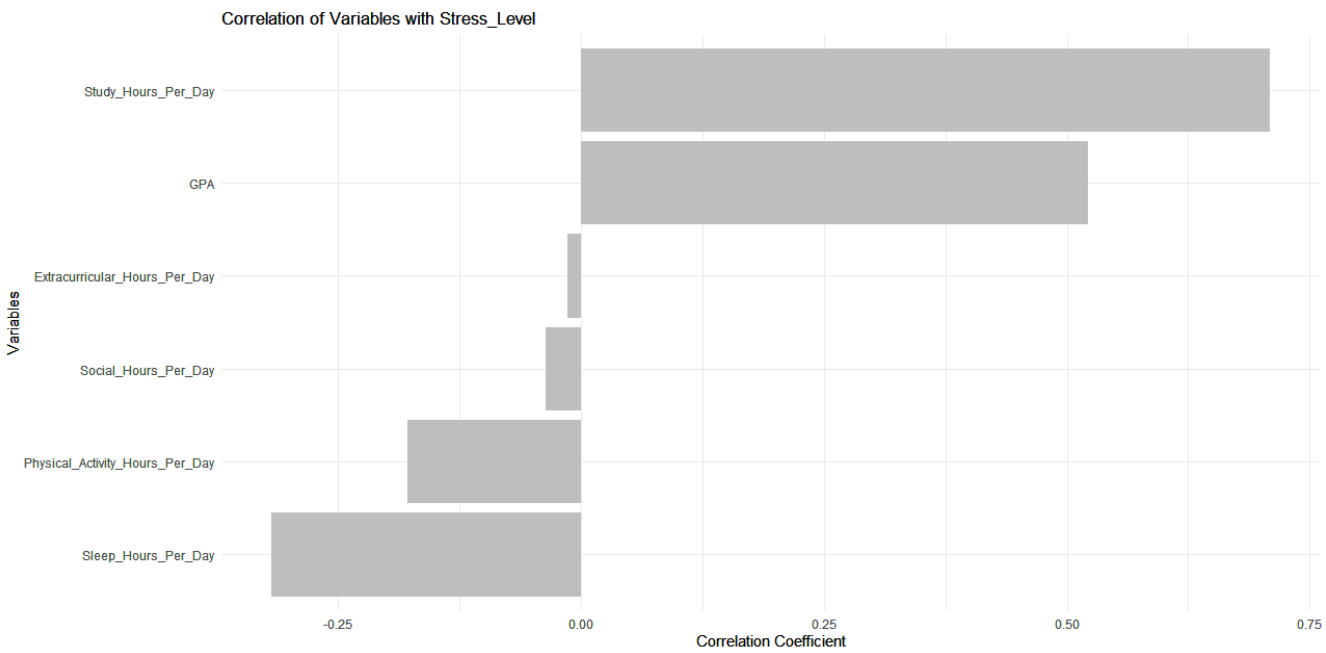
<u>Null Hypothesis (H0):</u> There is no significant difference in the mean of the variable across the three stress levels. Null Hypothesis (H0): The means of the variable are equal across the three Stress_Level groups.

$H_0 : \mu_{\text{Low}} = \mu_{\text{Moderate}} = \mu_{\text{High}}$

<u>Alternative Hypothesis (H1):</u> There is a significant difference in the mean of the variable across the three stress levels. At least one group's mean is significantly different from another group's mean.

$H_1 :$ At least one μ is different among Low, Moderate, and High

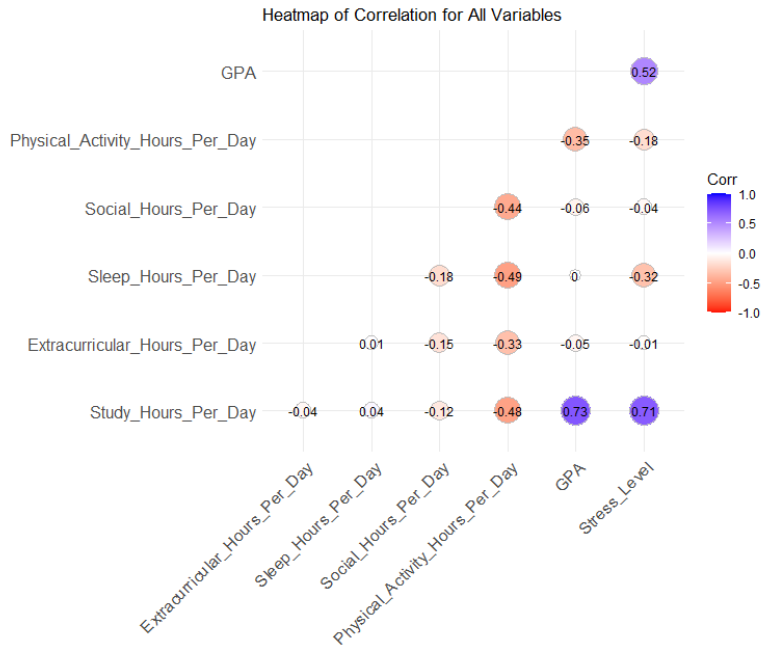
<u>Method:</u> ANOVA and Post Hoc Tests



The bar plot shows that Study_Hours_Per_Day and GPA have strong positive correlations with Stress_Level, indicating that higher stress is associated with more study hours and better academic performance. Conversely, Sleep_Hours_Per_Day and Physical_Activity_Hours_Per_Day have moderate negative correlations with stress, suggesting that higher stress reduces sleep and physical activity. Variables like Extracurricular_Hours_Per_Day and Social_Hours_Per_Day show negligible correlations, indicating minimal impact from stress levels.

```
> print(correlation_matrix)
                                Study_Hours_Per_Day Extracurricular_Hours_Per_Day Sleep_Hours_Per_Day Social_Hours_Per_Day
Study_Hours_Per_Day                      1.00000000                  -0.03619625         0.035752710         -0.11566221
Extracurricular_Hours_Per_Day           -0.03619625                   1.00000000         0.014054721         -0.14971033
Sleep_Hours_Per_Day                      0.03575271                   0.01405472         1.000000000         -0.18142345
Social_Hours_Per_Day                    -0.11566221                  -0.14971033        -0.181423455          1.00000000
Physical_Activity_Hours_Per_Day         -0.48475552                  -0.33348913        -0.488990502         -0.43678677
GPA                                      0.72966677                  -0.05150292         0.004158696         -0.06080221
Stress_Level                             0.70926898                  -0.01349743        -0.318754488         -0.03593849
                                Physical_Activity_Hours_Per_Day          GPA Stress_Level
Study_Hours_Per_Day                        -0.4847555  0.729666769   0.70926898
Extracurricular_Hours_Per_Day              -0.3334891 -0.051502917  -0.01349743
Sleep_Hours_Per_Day                        -0.4889905  0.004158696  -0.31875449
Social_Hours_Per_Day                       -0.4367868 -0.060802208  -0.03593849
Physical_Activity_Hours_Per_Day             1.0000000 -0.345586822  -0.17899991
GPA                                        -0.3455868  1.000000000   0.52205492
Stress_Level                               -0.1789999  0.522054925   1.00000000
```

The correlation matrix reveals the relationships between study-related variables, personal habits, GPA, and stress levels. Notably, Study_Hours_Per_Day shows a strong positive correlation with GPA (0.73) and stress level (0.71), indicating that students who study more tend to have higher GPAs and report higher stress levels. On the other hand, Sleep_Hours_Per_Day has a moderate negative correlation with stress levels (-0.32), suggesting that students with higher stress levels tend to sleep less. Similarly, Physical_Activity_Hours_Per_Day shows moderate negative correlations with both stress levels (-0.18) and study hours (-0.48), implying that physical activity decreases with increased stress or study load. A golden conclusion is that while increased study hours positively influence GPA, they are also associated with higher stress, which might negatively impact sleep and physical activity. Balancing these factors is critical to maintaining both academic success and personal well-being. Another key takeaway is the moderate positive correlation between GPA and stress level (0.52), suggesting that moderate stress could play a role in driving academic performance. However, excessive stress may harm overall health by reducing sleep and activity levels.



The correlations show that higher study hours are strongly linked to higher GPAs and increased stress levels, while stress negatively impacts sleep and physical activity. Sleep and physical activity decrease as stress rises, highlighting the trade-off between academic performance and well-being. Striking a balance between study, stress, and personal health is essential for maintaining both academic success and overall well-being.

```
> # Load necessary libraries
> library(ggplot2)
> library(dplyr)
> library(tidyr)
> # Convert Stress_Level to a factor variable
> data$Stress_Level <- factor(data$Stress_Level, levels = c("Low", "Moderate", "High"))
> # List of continuous variables
> variables <- c("Study_Hours_Per_Day", "Extracurricular_Hours_Per_Day",
+                "Sleep_Hours_Per_Day", "Social_Hours_Per_Day",
+                "Physical_Activity_Hours_Per_Day", "GPA")
> # Perform ANOVA and Tukey HSD for each variable
> anova_results <- list()
> tukey_results <- list()
> for (var in variables) {
+   # ANOVA
+   formula <- as.formula(paste(var, "~ Stress_Level"))
+   anova_model <- aov(formula, data = data)
+   anova_results[[var]] <- summary(anova_model)
+   print(paste("ANOVA Results for", var))
+   print(anova_results[[var]])
+
+   # Tukey HSD Test
+   tukey_results[[var]] <- TukeyHSD(anova_model)
+   print(paste("Tukey HSD Test Results for", var))
+   print(tukey_results[[var]])
+
+   # Boxplot for Visualization
+   plot <- ggplot(data, aes(x = Stress_Level, y = .data[[var]], fill = Stress_Level)) +
+     geom_boxplot() +
+     theme_minimal() +
+     ggtitle(paste("Boxplot of", var, "by Stress Level")) +
```

```
+     xlab("Stress Level") +
+     ylab(var) +
+     theme(legend.position = "none")
+
+   # Render the plot and save it
+   print(plot)  # Display the plot in the R console
+   ggsave(filename = paste0(var, "_boxplot.png"), plot = plot, width = 7, height = 7)
+ }
[1] "ANOVA Results for Study_Hours_Per_Day"
              Df Sum Sq Mean Sq F value Pr(>F)
Stress_Level   2  995.3   497.6   504.8 <2e-16 ***
Residuals    997  982.8     1.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Tukey HSD Test Results for Study_Hours_Per_Day"
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = formula, data = data)

$Stress_Level
                  diff      lwr      upr p adj
Moderate-Low  1.434488 1.200146 1.668829     0
High-Low      2.808230 2.584754 3.031705     0
High-Moderate 1.373742 1.212038 1.535445     0

[1] "ANOVA Results for Extracurricular_Hours_Per_Day"
              Df Sum Sq Mean Sq F value Pr(>F)
Stress_Level   2    0.7  0.3258   0.257  0.774
Residuals    997 1265.4  1.2693
[1] "Tukey HSD Test Results for Extracurricular_Hours_Per_Day"
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = formula, data = data)

$Stress_Level
                     diff       lwr       upr     p adj
Moderate-Low   0.03731780 -0.228595 0.3032306 0.9419474
High-Low      -0.01860751 -0.272190 0.2349750 0.9837792
High-Moderate -0.05592531 -0.239414 0.1275634 0.7544332

[1] "ANOVA Results for Sleep_Hours_Per_Day"
              Df Sum Sq Mean Sq F value Pr(>F)
Stress_Level   2  254.3  127.17   66.02 <2e-16 ***
Residuals    997 1920.6    1.93
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Tukey HSD Test Results for Sleep_Hours_Per_Day"
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = formula, data = data)

$Stress_Level
                    diff        lwr        upr     p adj
Moderate-Low  -0.1365246 -0.4641216  0.1910723 0.5908398
High-Low      -1.1031740 -1.4155803 -0.7907676 0.0000000
High-Moderate -0.9666494 -1.1927021 -0.7405966 0.0000000

[1] "ANOVA Results for Social_Hours_Per_Day"
              Df Sum Sq Mean Sq F value Pr(>F)
```

```
Stress_Level   2    5.2   2.608   0.902  0.406
Residuals    997 2883.5   2.892
[1] "Tukey HSD Test Results for Social_Hours_Per_Day"
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = formula, data = data)

$Stress_Level
                     diff        lwr       upr     p adj
Moderate-Low  -0.19550475 -0.5969072 0.2058977 0.4876857
High-Low      -0.21590389 -0.5986934 0.1668856 0.3821734
High-Moderate -0.02039914 -0.2973801 0.2565819 0.9836608

[1] "ANOVA Results for Physical_Activity_Hours_Per_Day"
              Df Sum Sq Mean Sq F value   Pr(>F)
Stress_Level   2    235  117.62   19.17 6.78e-09 ***
Residuals    997   6118    6.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Tukey HSD Test Results for Physical_Activity_Hours_Per_Day"
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = formula, data = data)

$Stress_Level
                    diff        lwr         upr     p adj
Moderate-Low  -1.1397762 -1.7244590 -0.55509331 0.0000159
High-Low      -1.4705441 -2.0281154 -0.91297282 0.0000000
High-Moderate -0.3307679 -0.7342186  0.07268267 0.1323442
```

```
[1] "ANOVA Results for GPA"
              Df Sum Sq Mean Sq F value Pr(>F)
Stress_Level   2  24.48  12.240     187 <2e-16 ***
Residuals    997  65.27   0.065
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] "Tukey HSD Test Results for GPA"
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = formula, data = data)

$Stress_Level
                   diff       lwr       upr p adj
Moderate-Low  0.2059146 0.1455252 0.2663041     0
High-Low      0.4325273 0.3749381 0.4901166     0
High-Moderate 0.2266127 0.1849420 0.2682834     0
```

The results presented show the output of ANOVA tests and post hoc Tukey HSD analyses to determine the significance of differences in continuous variables across stress level groups ("Low," "Moderate," and "High"). The analysis was conducted for six continuous variables: Study_Hours_Per_Day, Extracurricular_Hours_Per_Day, Sleep_Hours_Per_Day, Social_Hours_Per_Day, Physical_Activity_Hours_Per_Day, and GPA.

For Study_Hours_Per_Day, the ANOVA revealed a highly significant F-value (504.8, $p < 2e-16$), indicating substantial differences among the three stress level groups. Tukey's post hoc test further confirmed significant pairwise differences between all group combinations, with "High" stress levels associated with notably higher study hours than "Moderate" and "Low."

In contrast, the results for Extracurricular_Hours_Per_Day showed no significant differences across stress levels, as indicated by an F-value of 0.257 ($p = 0.774$). This suggests that stress levels do not substantially affect participation in extracurricular activities.

Sleep_Hours_Per_Day demonstrated a significant effect of stress levels ($F = 66.02$, $p < 2e-16$). Post hoc analysis showed that individuals with "High" stress levels sleep significantly less compared to both "Low" and "Moderate" groups. There was no significant difference between the "Low" and "Moderate" groups in sleep duration.

The ANOVA for Social_Hours_Per_Day revealed no significant differences (F = 0.902, p = 0.406), suggesting that social activity is unaffected by stress levels. Similarly, no significant pairwise differences were observed in Tukey's post hoc test. The results for Social_Hours_Per_Day indicate no significant differences in mean social hours across the three stress levels ("Low," "Moderate," and "High"), as evidenced by an ANOVA F-value of 0.902 and a p-value of 0.406. Additionally, Tukey's post hoc test confirms no statistically significant pairwise differences between any of the stress level groups, with all p-values above 0.05. This suggests that social activity levels are consistent regardless of stress levels.

For Physical_Activity_Hours_Per_Day, the analysis found a significant F-value (19.17, p = 6.78e-09), indicating stress levels influence physical activity. Post hoc results showed that "High" stress individuals spend significantly less time on physical activity than "Low" and "Moderate" stress groups. However, differences between "Moderate" and "Low" groups were not as pronounced. In details, the results for Physical_Activity_Hours_Per_Day indicate that stress levels significantly influence physical activity, as evidenced by the ANOVA test with an F-value of 19.17 and a p-value of 6.78e-09. This p-value is well below the conventional threshold of 0.05, confirming that there are meaningful differences in physical activity hours among the three stress groups ("Low," "Moderate," and "High").

Tukey's HSD post hoc test further clarifies these differences. The pairwise comparisons show the following:

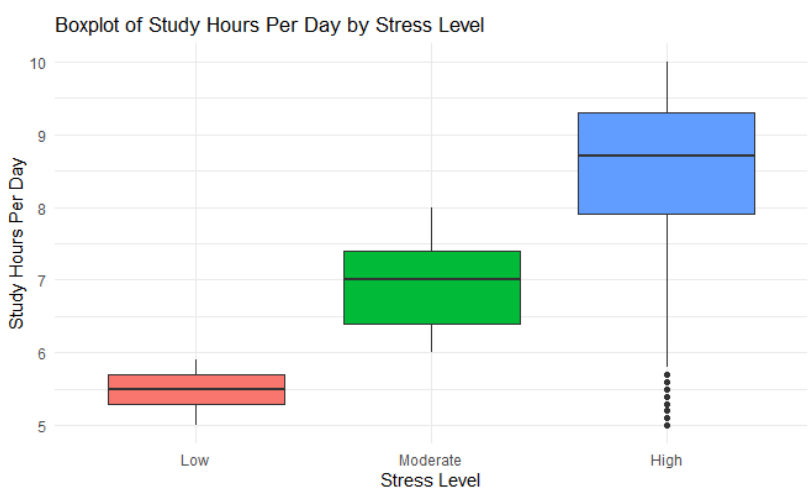
1. Individuals with "Moderate" stress engage in significantly fewer physical activity hours compared to those with "Low" stress (difference = -1.14, p = 0.0000159).
2. Those with "High" stress participate in significantly fewer physical activity hours compared to "Low" stress individuals (difference = -1.47, p < 0.0001).
3. However, the difference in physical activity between "Moderate" and "High" stress groups is not statistically significant (difference = -0.33, p = 0.13).

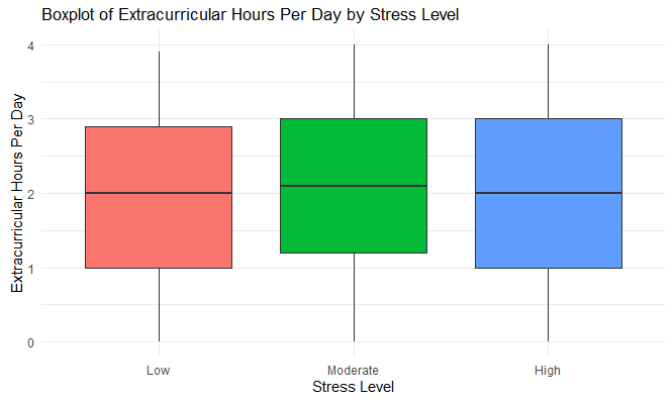
In summary, individuals with lower stress levels spend more time engaging in physical activity compared to those with moderate or high stress, while the gap between moderate and high-stress groups is less pronounced and not statistically significant.

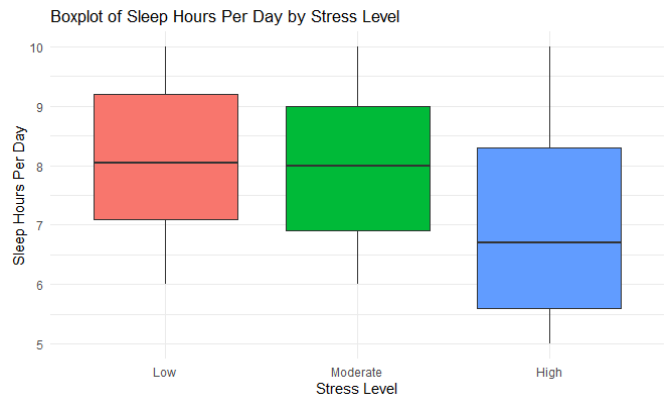
Finally, GPA showed a highly significant F-value (187, p < 2e-16), with Tukey's test confirming significant differences between all pairwise combinations of stress levels. Individuals with "High" stress had higher GPAs than those with "Moderate" or "Low" stress, highlighting a potential link between stress and academic performance. The ANOVA results for GPA indicate a highly significant

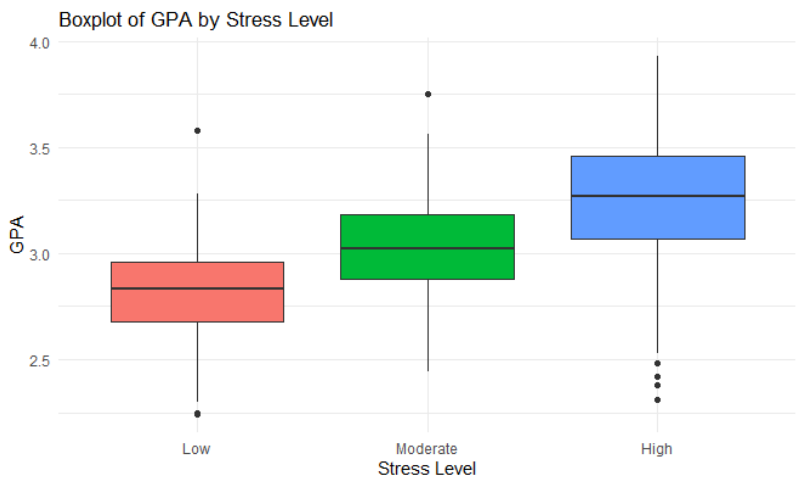
effect of stress levels on GPA, with an F-value of 187 and a p-value of <2e-16. This suggests that GPA differs significantly across the three stress level groups ("Low," "Moderate," and "High"). Tukey's HSD test confirms significant pairwise differences between all stress level groups. Specifically, individuals with "High" stress levels have significantly higher GPAs compared to those with "Low" stress (difference = 0.43, $p < 0.001$) and "Moderate" stress (difference = 0.23, $p < 0.001$). Additionally, individuals with "Moderate" stress have higher GPAs than those with "Low" stress (difference = 0.21, $p < 0.001$). These results demonstrate a clear relationship where increased stress levels are associated with improved academic performance as measured by GPA.

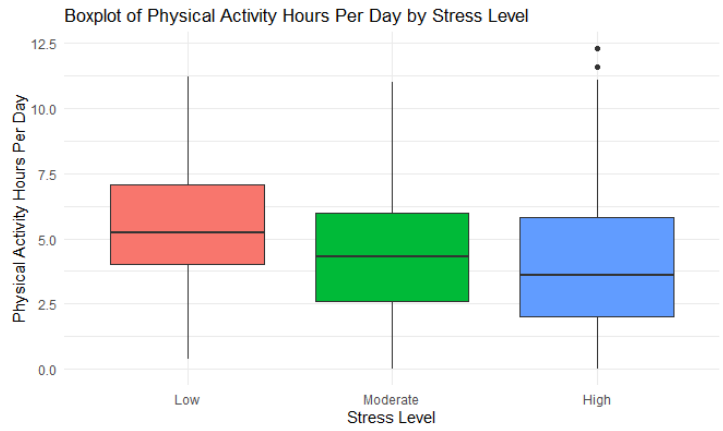
So finally, the analyses reveal that stress levels significantly affect Study_Hours_Per_Day, Sleep_Hours_Per_Day, Physical_Activity_Hours_Per_Day, and GPA, but not Extracurricular_Hours_Per_Day or Social_Hours_Per_Day. These findings suggest that stress may play a critical role in influencing specific behaviors and outcomes, particularly those related to academic and personal well-being. The analysis shows that stress levels significantly impact study hours, sleep, physical activity, and GPA, while extracurricular and social activities remain unaffected. High-stress individuals study more, sleep less, and engage in less physical activity but tend to have higher GPAs compared to those with lower stress. These results suggest that stress influences behaviors and academic outcomes, highlighting its dual effects on performance and well-being.

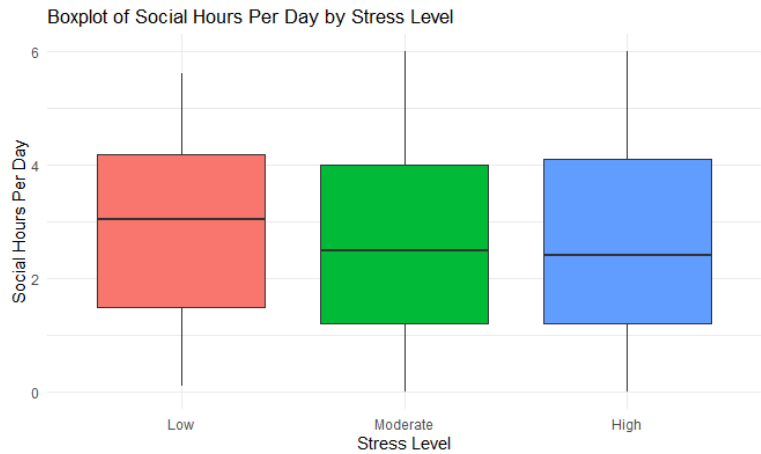



Boxplot of Extracurricular Hours Per Day by Stress Level
Extracurricular Hours Per Day
0
1
2
3
4
Low
Moderate
High
Stress Level


Boxplot of Sleep Hours Per Day by Stress Level
Sleep Hours Per Day
5
6
7
8
9
10
Low
Moderate
High
Stress Level


Boxplot of GPA by Stress Level
GPA
2.5
3.0
3.5
4.0
Low
Moderate
High
Stress Level





Summary

- **Study_Hours_Per_Day:** Highly significant differences exist among stress level groups (F = 504.8, p < 2e-16), with higher stress associated with more study hours.

- **Extracurricular_Hours_Per_Day:** No significant differences were found among stress level groups (F = 0.257, p = 0.774).

- **Sleep_Hours_Per_Day:** Significant differences were observed among stress level groups (F = 66.02, p < 2e-16), with higher stress linked to reduced sleep hours.

- **Social_Hours_Per_Day:** No significant differences were observed among stress level groups (F = 0.902, p = 0.406).

- **Physical_Activity_Hours_Per_Day:** Significant differences exist among stress level groups (F = 19.17, p < 7e-09), with higher stress levels associated with less physical activity.

- **GPA:** Stress levels significantly affect GPA (F = 187, p < 2e-16), with higher stress linked to improved academic performance.

- **Study Hours:** Higher stress levels are linked to more study hours, showing that students under more stress tend to focus more on academics.
- **Extracurricular Activities:** Stress levels do not significantly impact participation in extracurricular activities, indicating they remain consistent regardless of stress.
- **Sleep Hours:** Students with higher stress levels sleep less, highlighting the potential negative impact of stress on rest and recovery.
- **Social Hours:** Stress levels do not significantly affect time spent on social activities, suggesting social engagement stays steady across groups.
- **Physical Activity:** Higher stress is associated with reduced physical activity, pointing to the need for managing stress to maintain physical health.
- **GPA:** Higher stress levels are linked to better academic performance, suggesting stress may drive academic focus but needs balance for well-being.

0. General Information _ Big Picture (with no stress level)

This project investigates the relationship between students' daily lifestyle habits (e.g., study hours, extracurricular hours, sleep, socializing, physical activity) and academic performance (GPA), with stress levels (Low, Moderate, High) as a potential moderating factor. The dataset comprises 1000 samples with 7 variables, including one categorical ordinal variable (Stress_Level) and six continuous variables. Data cleaning involves handling missing values, detecting outliers, and encoding categorical data. Statistical methods include correlation analysis to identify relationships between lifestyle factors and GPA, multiple linear regression to evaluate combined effects, ANOVA to assess GPA differences across stress levels, and ANCOVA to test the impact of stress on GPA while controlling for lifestyle factors. Interaction analysis will explore how stress moderates the relationship between study hours and GPA, complemented by visualizations like scatter plots, box plots, and interaction plots.

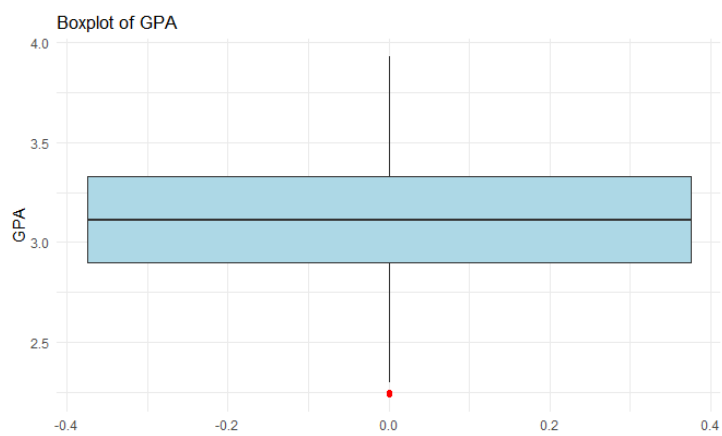
```
> # Load necessary libraries
> library(ggplot2)
> library(dplyr)
> # General summary of the dataset
> print("General Summary:")
[1] "General Summary:"
> print(summary(data))
 Study_Hours_Per_Day Extracurricular_Hours_Per_Day Sleep_Hours_Per_Day Social_Hours_Per_Day Physical_Activity_Hours_Per_Day      GPA
 Min.   : 5.000      Min.   :0.00                  Min.   : 5.000      Min.   :0.000        Min.   : 0.000                  Min.   :2.240
 1st Qu.: 6.200      1st Qu.:1.10                  1st Qu.: 6.100      1st Qu.:1.200        1st Qu.: 2.400                  1st Qu.:2.900
 Median : 7.400      Median :2.00                  Median : 7.500      Median :2.500        Median : 4.100                  Median :3.110
 Mean   : 7.457      Mean   :2.01                  Mean   : 7.486      Mean   :2.691        Mean   : 4.356                  Mean   :3.115
 3rd Qu.: 8.700      3rd Qu.:3.00                  3rd Qu.: 8.800      3rd Qu.:4.100        3rd Qu.: 6.200                  3rd Qu.:3.330
 Max.   :10.000      Max.   :4.00                  Max.   :10.000      Max.   :6.000        Max.   :12.300                  Max.   :3.930
 Stress_Level
 Length:1000
 Class :character
 Mode  :character



> # Additional summary statistics for numeric variables
> summary_stats <- data %>%
+   select(where(is.numeric)) %>% # Select only numeric columns
+   summarise_all(list(
+     Mean = ~mean(., na.rm = TRUE),
+     Median = ~median(., na.rm = TRUE),
+     Std_Dev = ~sd(., na.rm = TRUE),
+     Variance = ~var(., na.rm = TRUE)
+   ))
> print("Summary Statistics for Numeric Variables:")
[1] "Summary Statistics for Numeric Variables:"
> print(summary_stats)
  Study_Hours_Per_Day_Mean Extracurricular_Hours_Per_Day_Mean Sleep_Hours_Per_Day_Mean Social_Hours_Per_Day_Mean
1                   7.4572                               2.01                   7.4857                   2.6913
  Physical_Activity_Hours_Per_Day_Mean GPA_Mean Study_Hours_Per_Day_Median Extracurricular_Hours_Per_Day_Median Sleep_Hours_Per_Day_Median
1                               4.3558   3.1152                       7.4                                    2                        7.5
  Social_Hours_Per_Day_Median Physical_Activity_Hours_Per_Day_Median GPA_Median Study_Hours_Per_Day_Std_Dev
1                         2.5                                    4.1       3.11                   1.407142
  Extracurricular_Hours_Per_Day_Std_Dev Sleep_Hours_Per_Day_Std_Dev Social_Hours_Per_Day_Std_Dev Physical_Activity_Hours_Per_Day_Std_Dev
1                              1.125774                   1.475521                   1.700484                                2.521818
  GPA_Std_Dev Study_Hours_Per_Day_Variance Extracurricular_Hours_Per_Day_Variance Sleep_Hours_Per_Day_Variance Social_Hours_Per_Day_Variance
1   0.2997272                    1.980048                                1.267367                     2.177163                     2.891646
  Physical_Activity_Hours_Per_Day_Variance GPA_Variance
1                                 6.359566    0.0898364
> # Mode function
> get_mode <- function(v) {
+   uniq_v <- unique(v)
+   uniq_v[which.max(tabulate(match(v, uniq_v)))]
+ }
> # Compute mode for each variable
> mode_stats <- sapply(data, function(column) {
+   if (is.numeric(column) || is.factor(column) || is.character(column)) {
+     return(get_mode(column))
+   } else {
+     return(NA)
+   }
+ })
> print("Mode for each variable:")
[1] "Mode for each variable:"
> print(mode_stats)
            Study_Hours_Per_Day   Extracurricular_Hours_Per_Day             Sleep_Hours_Per_Day             Social_Hours_Per_Day
                          "6.3"                           "1.6"                           "9.7"                           "2.5"
Physical_Activity_Hours_Per_Day                             GPA                     Stress_Level
                          "4.3"                          "3.26"                          "High"
> # Variable types
> print("Structure of the dataset:")
[1] "Structure of the dataset:"
> str(data)
'data.frame':   1000 obs. of  7 variables:
 $ Study_Hours_Per_Day            : num  5.8 7.8 8.2 7.5 6.6 7.1 6.7 8.3 8.9 9.3 ...
 $ Extracurricular_Hours_Per_Day  : num  0.8 0.2 3.9 1.3 2.9 3.3 0.4 2.7 1 1.3 ...
 $ Sleep_Hours_Per_Day            : num  9.9 9.9 8.6 7.2 8.2 5.6 9 9.7 7.8 5.8 ...
 $ Social_Hours_Per_Day           : num  0.9 3.1 3.1 5.1 2 4.4 5.7 3.3 2.2 3.3 ...
 $ Physical_Activity_Hours_Per_Day: num  6.6 3 0.2 2.9 4.3 3.6 2.2 0 4.1 4.3 ...
 $ GPA                            : num  2.71 3.07 3.02 3.27 3.14 2.85 2.63 2.95 3.54 3.09 ...
 $ Stress_Level                   : chr  "Low" "Moderate" "High" "Moderate" ...
> # Barplot for categorical variable (Stress_Level)
> if ("Stress_Level" %in% colnames(data)) {
+   bar_plot <- ggplot(data, aes(x = Stress_Level)) +
+     geom_bar(fill = "skyblue") +
+     labs(title = "Frequency of Stress Levels", x = "Stress Level", y = "Count") +
+     theme_minimal()
+
+   print(bar_plot)
+ }
> # Boxplots for continuous variables
> continuous_vars <- names(data)[sapply(data, is.numeric)]
> for (var in continuous_vars) {
+   box_plot <- ggplot(data, aes(y = .data[[var]])) +
+     geom_boxplot(fill = "lightblue", outlier.colour = "red") +
```

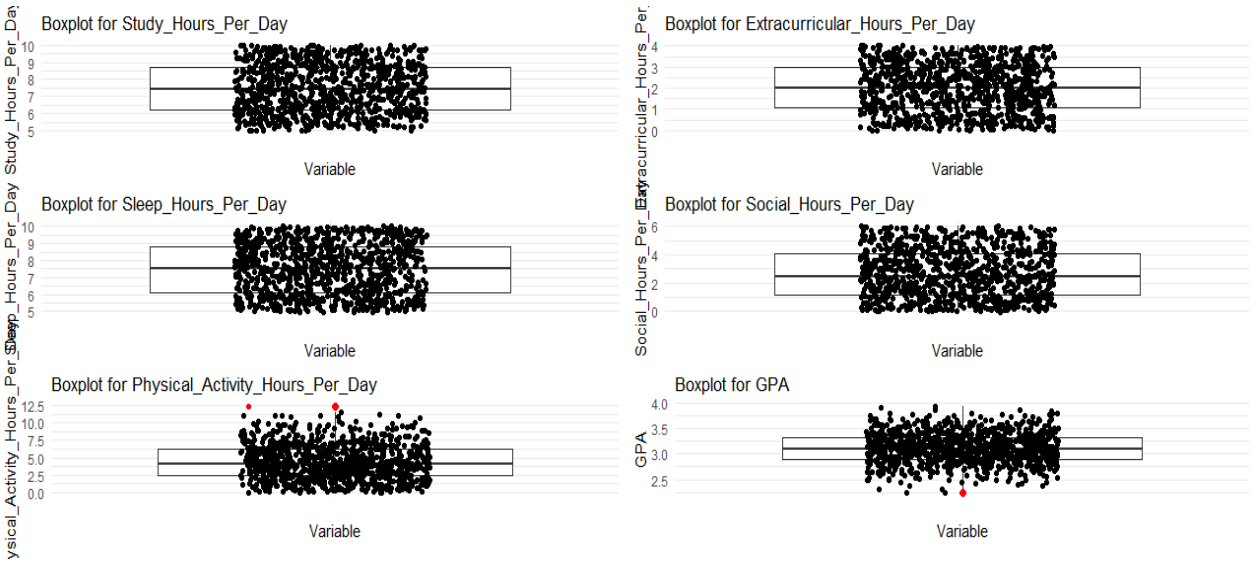
```
+     labs(title = paste("Boxplot of", var), y = var) +
+     theme_minimal()
+
+   print(box_plot)
+ }
> # Pairwise correlations for numerical variables
> cor_matrix <- cor(data[sapply(data, is.numeric)], use = "complete.obs")
> print("Correlation Matrix:")
[1] "Correlation Matrix:"
> print(cor_matrix)
                              Study_Hours_Per_Day Extracurricular_Hours_Per_Day Sleep_Hours_Per_Day Social_Hours_Per_Day
Study_Hours_Per_Day                    1.00000000                   -0.03619625         0.035752710          -0.11566221
Extracurricular_Hours_Per_Day         -0.03619625                    1.00000000         0.014054721          -0.14971033
Sleep_Hours_Per_Day                    0.03575271                    0.01405472         1.000000000          -0.18142345
Social_Hours_Per_Day                  -0.11566221                   -0.14971033        -0.181423455           1.00000000
Physical_Activity_Hours_Per_Day       -0.48475552                   -0.33348913        -0.488990502          -0.43678677
GPA                                    0.72966677                   -0.05150292         0.004158696          -0.06080221
                              Physical_Activity_Hours_Per_Day          GPA
Study_Hours_Per_Day                          -0.4847555  0.729666769
Extracurricular_Hours_Per_Day                -0.3334891 -0.051502917
Sleep_Hours_Per_Day                          -0.4889905  0.004158696
Social_Hours_Per_Day                         -0.4367868 -0.060802208
Physical_Activity_Hours_Per_Day               1.0000000 -0.345586822
GPA                                          -0.3455868  1.000000000
> # Visualization: Heatmap of the correlation matrix
> if (ncol(cor_matrix) > 1) {
+   if (!requireNamespace("corrplot", quietly = TRUE)) {
+     install.packages("corrplot")  # Ensure corrplot is installed
+   }
+   library(corrplot)
+   corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45)
+ }
```

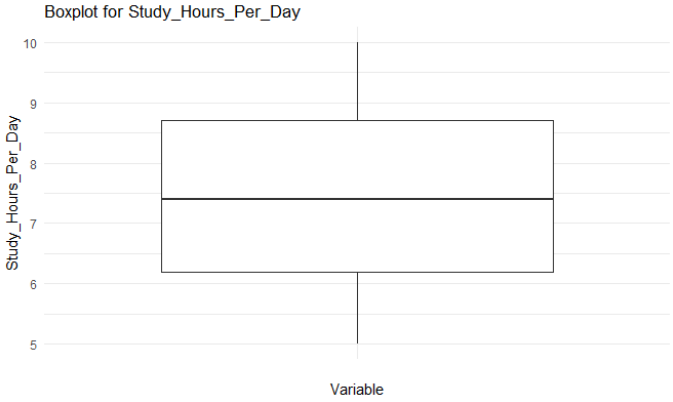
```
> # Calculate the mode for Stress_Level
> stress_mode <- as.character(names(sort(table(data$Stress_Level), decreasing = TRUE)[1]))
> # Print the result
> cat("The mode for Stress_Level is:", stress_mode, "\n")
The mode for Stress_Level is: 2
```

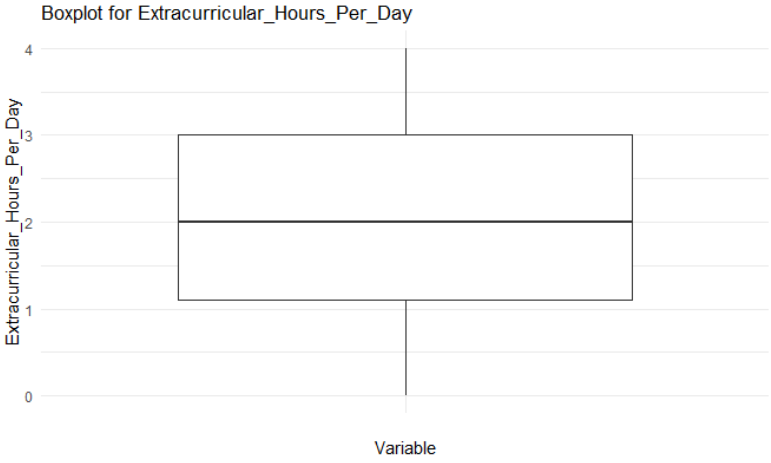
It means high is the data's mode.

Box plot

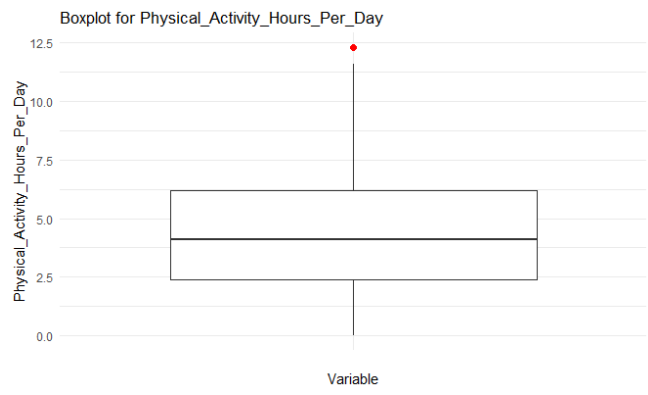


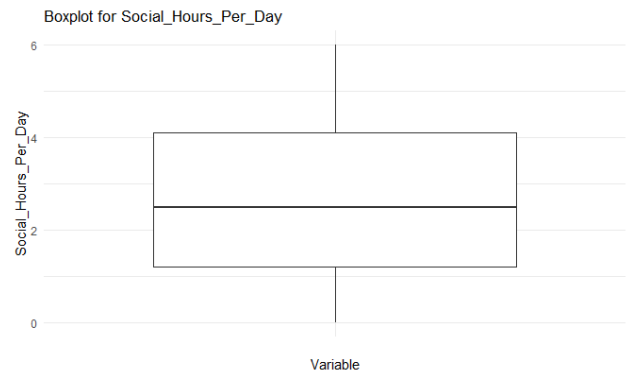

Boxplot for Study_Hours_Per_Day
Boxplot for Extracurricular_Hours_Per_Day
Boxplot for Sleep_Hours_Per_Day
Boxplot for Social_Hours_Per_Day
Boxplot for Physical_Activity_Hours_Per_Day
Boxplot for GPA
Variable

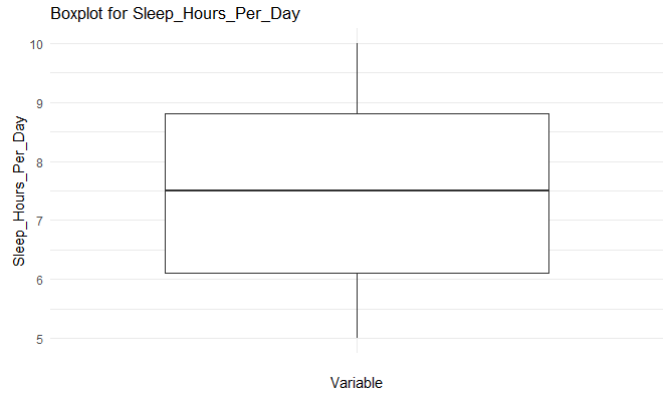

Boxplot for Study_Hours_Per_Day
Study_Hours_Per_Day
Variable


Boxplot for Extracurricular_Hours_Per_Day
Extracurricular_Hours_Per_Day
Variable

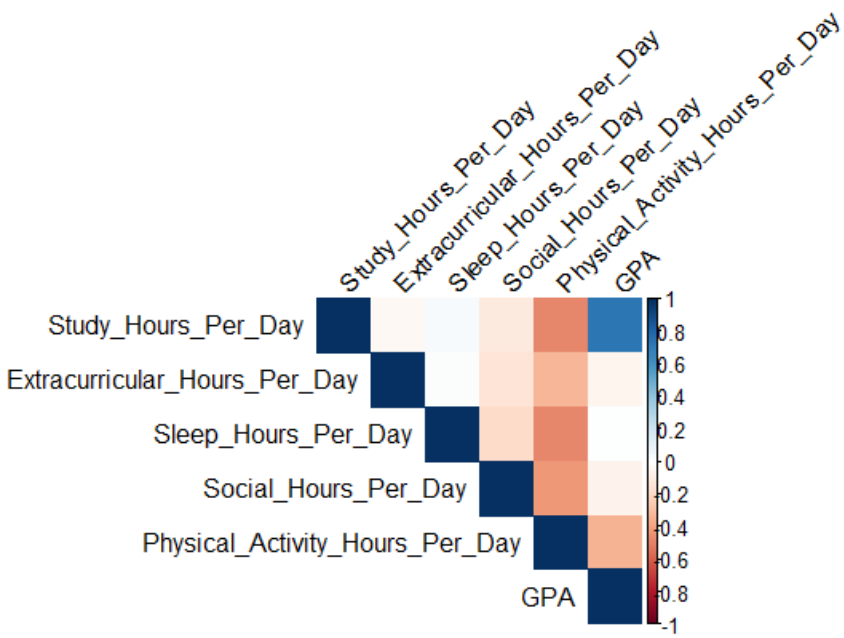
```
> # Load necessary libraries
> library(ggplot2)
> library(dplyr)
> # Identify scalar (numeric) variables (excluding "Stress_Level")
> scalar_vars <- setdiff(names(data)[sapply(data, is.numeric)], "Stress_Level")
> # Function to calculate and print outliers beyond 3 standard deviations
> detect_outliers <- function(var_name, data) {
+   values <- data[[var_name]]
+   mean_val <- mean(values, na.rm = TRUE)
+   sd_val <- sd(values, na.rm = TRUE)
+   outliers <- values[abs(values - mean_val) > 3 * sd_val]
+
+   if (length(outliers) > 0) {
+     print(paste("Outliers for", var_name, ":", toString(outliers)))
+   } else {
+     print(paste("No outliers for", var_name))
+   }
+
+   return(outliers)
+ }
> # Loop through each scalar variable to create individual plots
> for (var in scalar_vars) {
+   # Detect outliers and print them
+   detect_outliers(var, data)
+
+   # Create a boxplot for the current variable
+   p <- ggplot(data, aes(x = "", y = .data[[var]])) +  # Use "" as a placeholder for x
+     geom_boxplot(outlier.colour = "red", outlier.size = 2) +
+     labs(title = paste("Boxplot for", var), x = "Variable", y = var) +
+     theme_minimal()
+   # Display the plot
+   print(p)
+
+   # Pause to allow screenshot before moving to the next plot
+   readline(prompt = "Press [Enter] to continue to the next plot...")
+ }
[1] "No outliers for Study_Hours_Per_Day"
Press [Enter] to continue to the next plot...
[1] "No outliers for Extracurricular_Hours_Per_Day"
Press [Enter] to continue to the next plot...
[1] "No outliers for Sleep_Hours_Per_Day"
Press [Enter] to continue to the next plot...
[1] "No outliers for Social_Hours_Per_Day"
Press [Enter] to continue to the next plot...
[1] "Outliers for Physical_Activity_Hours_Per_Day : 12.3"
Press [Enter] to continue to the next plot...
[1] "No outliers for GPA"
Press [Enter] to continue to the next plot...
```



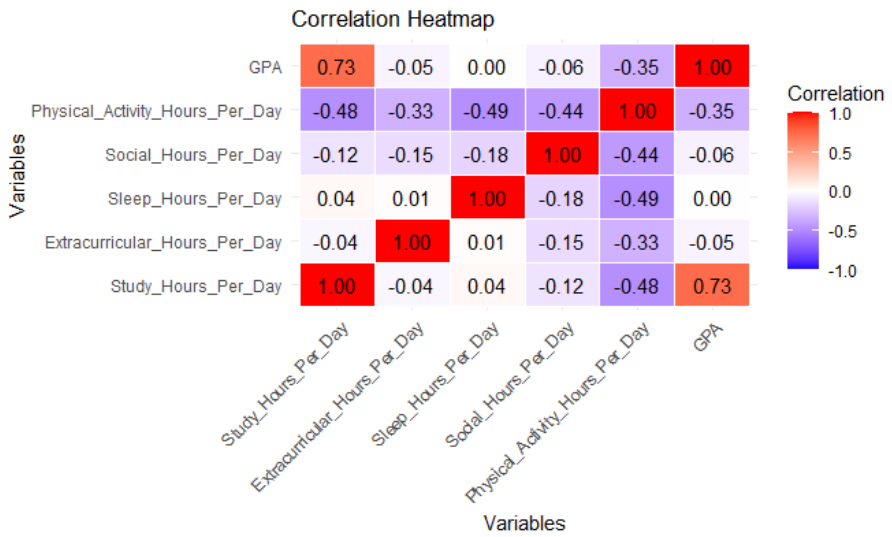




Heat map for sclar variables



Heat map for sclar variables



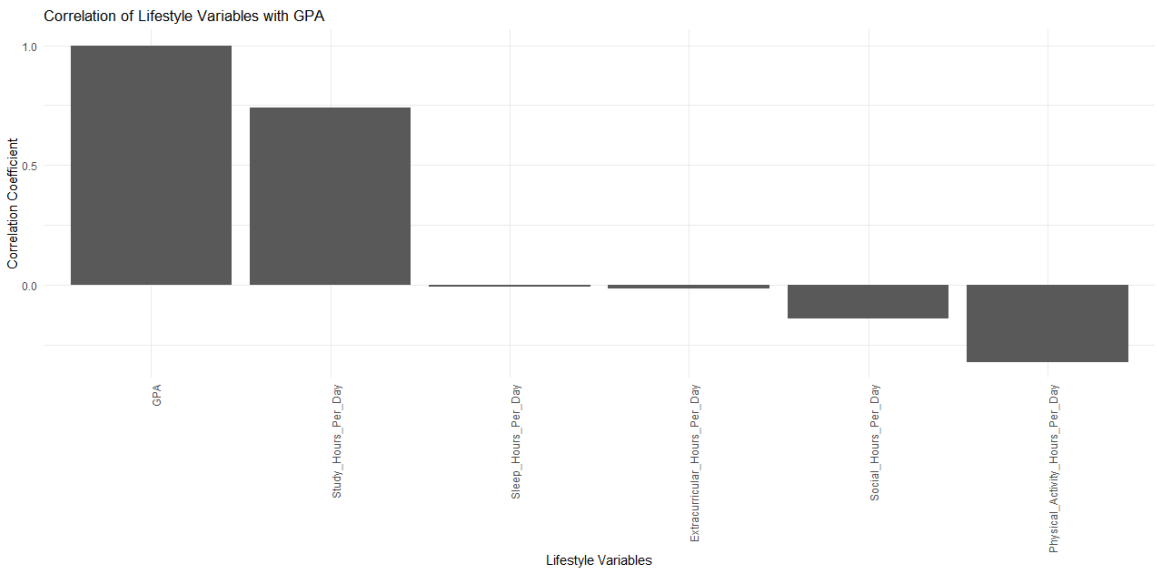
1. Describe your research scenario and question(s).

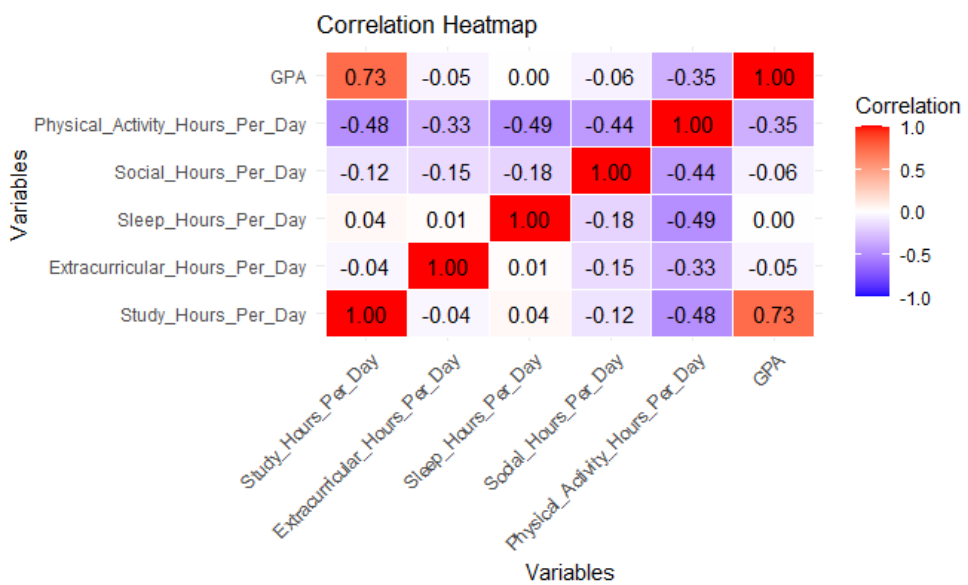
Research Scenario:
In the pursuit of academic excellence, students' daily lifestyles—such as time spent studying, participating in extracurricular activities, engaging in physical activities, socializing, and sleeping—play a crucial role in shaping their academic performance. This research investigates how these lifestyle habits influence GPA. By leveraging statistical and data science methods, we aim to identify significant patterns and predictors of academic success. Additionally, we seek to provide actionable insights for students and educators to optimize lifestyle choices to improve GPA outcomes.

Research Questions and Hypotheses:

1. **Research Question 1:** What are the relationships between daily lifestyle habits (study hours, extracurricular hours, sleep hours, social hours, physical activity) and GPA?
   - **Hypothesis:** There is a significant correlation between each lifestyle variable and GPA.
   - **Method:** Correlation analysis (Pearson's correlation).

```
> ##1. Describe your research scenario and question(s).##
> # Load necessary libraries
> library(ggplot2)
> # 1. Perform Pearson's correlation analysis
> correlation_matrix <- cor(data, method = "pearson")
> # Display the correlation matrix
> print("Correlation Matrix:")
[1] "Correlation Matrix:"
> print(correlation_matrix)
                            Study_Hours_Per_Day Extracurricular_Hours_Per_Day Sleep_Hours_Per_Day Social_Hours_Per_Day
Study_Hours_Per_Day                  1.00000000                   0.046655448         0.014268256           -0.1854648
Extracurricular_Hours_Per_Day        0.04665545                   1.000000000         0.007538127           -0.1649599
Sleep_Hours_Per_Day                  0.01426826                   0.007538127         1.000000000           -0.1943317
Social_Hours_Per_Day                -0.18546477                  -0.164959889        -0.194331731            1.0000000
Physical_Activity_Hours_Per_Day     -0.48288554                  -0.384486888        -0.469112184           -0.3855527
GPA                                  0.73877023                  -0.015333360        -0.006583260           -0.1397429
                            Physical_Activity_Hours_Per_Day         GPA
Study_Hours_Per_Day                        -0.4828855  0.73877023
Extracurricular_Hours_Per_Day              -0.3844869 -0.01533336
Sleep_Hours_Per_Day                        -0.4691122 -0.00658326
Social_Hours_Per_Day                       -0.3855527 -0.13974291
Physical_Activity_Hours_Per_Day             1.0000000 -0.32156074
GPA                                        -0.3215607  1.00000000
> # Isolate GPA correlations for clarity
> gpa_correlations <- correlation_matrix[,"GPA"]
> print("Correlations with GPA:")
[1] "Correlations with GPA:"
> print(gpa_correlations)
            Study_Hours_Per_Day   Extracurricular_Hours_Per_Day            Sleep_Hours_Per_Day            Social_Hours_Per_Day
                     0.73877023                     -0.01533336                     -0.00658326                     -0.13974291
Physical_Activity_Hours_Per_Day                             GPA
                    -0.32156074                      1.00000000
> # 2. Visualize correlations with GPA using a bar plot
> gpa_cor_df <- data.frame(Variable = names(gpa_correlations),
+                          Correlation = gpa_correlations)
> ggplot(data = gpa_cor_df, aes(x = reorder(Variable, -Correlation), y = Correlation)) +
+   geom_bar(stat = "identity") +
+   labs(title = "Correlation of Lifestyle Variables with GPA",
+        x = "Lifestyle Variables",
+        y = "Correlation Coefficient") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels
> # 3. Testing significance of correlations
> correlation_tests <- lapply(names(data)[-which(names(data) == "GPA")], function(var) {
+   cor.test(data[[var]], data$GPA, method = "pearson")
+ })
> # Extract p-values
> p_values <- sapply(correlation_tests, function(test) test$p.value)
> # Combine variables and p-values into a data frame
> correlation_significance <- data.frame(
+   Variable = names(data)[-which(names(data) == "GPA")],
+   P_Value = p_values
+ )
> # Print significance of correlations
> print("Significance of Correlations (P-Values):")
[1] "Significance of Correlations (P-Values):"
> print(correlation_significance)
                         Variable       P_Value
1             Study_Hours_Per_Day 3.225236e-173
2   Extracurricular_Hours_Per_Day  6.281691e-01
3             Sleep_Hours_Per_Day  8.352892e-01
4            Social_Hours_Per_Day  9.193926e-06
5 Physical_Activity_Hours_Per_Day  1.728890e-25
> # 4. Identify significant correlations
> significant_vars <- correlation_significance[correlation_significance$P_Value < 0.05, ]
> # Print significant variables
> print("Significant Variables with GPA:")
[1] "Significant Variables with GPA:"
> print(significant_vars)
                          Variable       P_Value
1             Study_Hours_Per_Day 3.225236e-173
4            Social_Hours_Per_Day  9.193926e-06
5 Physical_Activity_Hours_Per_Day  1.728890e-25
```





The results of the analysis indicate that study hours per day, social hours per day, and physical activity hours per day have significant correlations with GPA at a 95% confidence level. Study hours per day show a strong positive correlation with GPA, suggesting that increased study time is associated with higher academic performance. Social hours per day exhibit a weak negative correlation, implying that higher social activity may slightly detract from academic achievement. Physical activity hours per day also have a moderate negative correlation with GPA, indicating that more time spent on physical activity might be linked to slightly lower academic outcomes. These findings highlight the complexity of lifestyle factors influencing academic performance, with study habits playing the most pivotal role among the variables analyzed.

2. **Research Question 2:** Which lifestyle factors significantly predict GPA, and how much variation in GPA can they collectively explain?
   - **Hypothesis:** Study hours, sleep hours, and physical activity are significant predictors of GPA.
   - **Method:** Multiple linear regression.

In the multiple linear regression analysis provided above, the null hypothesis (H_0) and alternative hypothesis (H_1) are implicitly tested for each predictor variable using the t-test associated with the regression coefficients. For each predictor variable (Xi):

- Null Hypothesis (H_0): The coefficient of the predictor is zero ($\beta i=0$), meaning the predictor has no significant relationship with the dependent variable (GPA).
- Alternative Hypothesis (H_1): The coefficient of the predictor is not zero ($\beta i \neq 0$), meaning the predictor has a significant relationship with GPA.

```
> # Fit a multiple linear regression model
> model <- lm(GPA ~ Study_Hours_Per_Day + Sleep_Hours_Per_Day + Physical_Activity_Hours_Per_Day, data = data)
> # Display the summary of the regression model
> summary_model <- summary(model)
> print(summary_model)

Call:
lm(formula = GPA ~ Study_Hours_Per_Day + Sleep_Hours_Per_Day +
    Physical_Activity_Hours_Per_Day, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59388 -0.13921 -0.00343  0.13478  0.66344

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.881448   0.074028  25.415   <2e-16 ***
Study_Hours_Per_Day           0.160044   0.005314  30.120   <2e-16 ***
Sleep_Hours_Per_Day           0.001167   0.005173   0.225   0.8216
Physical_Activity_Hours_Per_Day 0.005998   0.003488   1.720   0.0858 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2041 on 996 degrees of freedom
Multiple R-squared:  0.5474,    Adjusted R-squared:  0.5461
F-statistic: 401.6 on 3 and 996 DF,  p-value: < 2.2e-16

> # Extract R-squared and Adjusted R-squared
> r_squared <- summary_model$r.squared
> adj_r_squared <- summary_model$adj.r.squared
> cat("R-squared:", r_squared, "\n")
R-squared: 0.5474186
```

```
> cat("Adjusted R-squared:", adj_r_squared, "\n")
Adjusted R-squared: 0.5460554
> # Check for significant predictors
> significant_predictors <- summary_model$coefficients[, "Pr(>|t|)"] < 0.05
> cat("Significant Predictors:\n")
Significant Predictors:
> print(rownames(summary_model$coefficients)[significant_predictors])
[1] "(Intercept)"        "Study_Hours_Per_Day"
```

The multiple linear regression analysis indicates that among the lifestyle factors examined, only study hours per day significantly predict GPA ($p<0.05$), with a strong positive effect size, while sleep hours and physical activity hours are not statistically significant predictors. The model explains approximately 54.7% of the variance in GPA, as indicated by the R^2 value, suggesting that the included predictors collectively account for a substantial proportion of GPA variability. The adjusted R^2 value of 54.6% confirms the model's robustness, with minimal overfitting. The results emphasize the dominant role of study habits in predicting academic performance, whereas other factors like physical activity and sleep have less direct influence in this context.

3. **Research Question 3:** Are there significant differences in GPA based on specific levels of individual lifestyle habits (e.g., high versus low social hours)?
   - **Hypothesis:** Students with high levels of study hours have significantly higher GPAs than those with low study hours.
   - **Method:** One-way ANOVA.

```
> # Create a categorical variable for study hours based on median split
> data$Study_Hours_Level <- ifelse(data$Study_Hours_Per_Day > median(data$Study_Hours_Per_Day),
+                                  "High", "Low")
> # Check the distribution of students in each group
> table(data$Study_Hours_Level)

High  Low
 499  501
> # Perform one-way ANOVA
> anova_result <- aov(GPA ~ Study_Hours_Level, data = data)
> # Display the ANOVA summary
> summary(anova_result)
                  Df Sum Sq Mean Sq F value Pr(>F)
Study_Hours_Level   1  36.85   36.85   670.9 <2e-16 ***
Residuals         998  54.82    0.05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Calculate group means for GPA
> group_means <- aggregate(GPA ~ Study_Hours_Level, data = data, mean)
> print("Group Means:")
[1] "Group Means:"
> print(group_means)
  Study_Hours_Level      GPA
1              High 3.307655
2               Low 2.923713
> # Post-hoc test if ANOVA is significant (Tukey's HSD)
> if (summary(anova_result)[[1]][["Pr(>F)"]][1] < 0.05) {
+   tukey_result <- TukeyHSD(anova_result)
+   print("Tukey Post-Hoc Test Results:")
+   print(tukey_result)
+ } else {
+   print("ANOVA is not significant. No need for post-hoc testing.")
+ }
[1] "Tukey Post-Hoc Test Results:"
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ Study_Hours_Level, data = data)

$Study_Hours_Level
               diff        lwr        upr p adj
Low-High -0.3839427 -0.4130303 -0.3548552     0
```

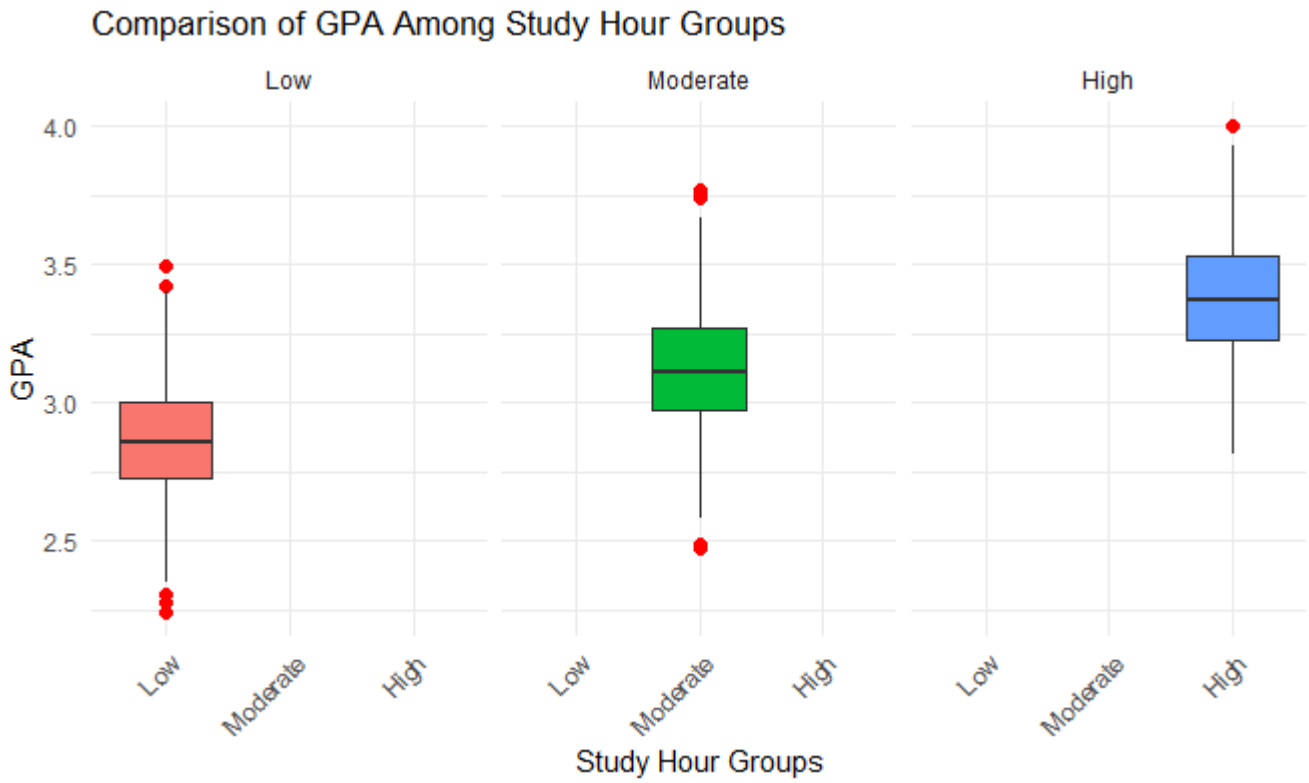
The one-way ANOVA results indicate a significant difference in GPA between students with high and low study hours, as evidenced by the F-statistic of 670.9 and a p-value <2e-16, well below the 0.05 significance threshold. Students in the "High" study hours group had a significantly higher mean GPA (3.31) compared to the "Low" study hours group (2.92), with a mean difference of 0.38, confirmed by Tukey’s HSD test. These findings support the hypothesis that students with higher study hours achieve significantly better academic performance, highlighting the importance of study habits in determining GPA. The Tukey post-hoc test shows a significant difference in GPA between the Low and High study hour groups (diff=-0.384, p<0.001). The High study hour group has a significantly higher GPA, as indicated by the negative difference, with a 95% confidence interval (-0.413, -0.355) that does not include zero.

5. Research Question 5: Does the GPA significantly differ among students grouped by low, moderate, and high levels of study hours?
Null Hypothesis (H_0): There is no significant difference in GPA among the three groups of study hours (low, moderate, high).
Alternative Hypothesis (H_1): At least one group has a significantly different GPA compared to the others.
Method: One-way ANOVA will be used to compare the GPA among the three groups of study hours. Tukey's HSD test will be used for pairwise comparisons to identify specific group differences.

```
> # Load necessary libraries
> library(ggplot2)
> # Sort the dataset by Study_Hours_Per_Day in ascending order
> data <- data[order(data$Study_Hours_Per_Day), ]
> # Create three equal groups for Study_Hours_Per_Day: Low, Moderate, High
> n <- nrow(data)
> data$Study_Hours_Group <- cut(1:n,
+                               breaks = c(0, floor(n / 3), floor(2 * n / 3), n),
+                               labels = c("Low", "Moderate", "High"),
+                               include.lowest = TRUE)
> # Check the distribution of groups
> table(data$Study_Hours_Group)

     Low Moderate     High
     333      333      334
> # Calculate the average Study_Hours_Per_Day in each group
> average_hours <- aggregate(Study_Hours_Per_Day ~ Study_Hours_Group, data = data, mean)
> # Print the average hours for each group
> print("Average Study Hours in Each Group:")
[1] "Average Study Hours in Each Group:"
> print(average_hours)
  Study_Hours_Group Study_Hours_Per_Day
1               Low            5.842342
2          Moderate            7.470871
3              High            9.162575
> # Perform one-way ANOVA
> anova_result <- aov(GPA ~ Study_Hours_Group, data = data)
> anova_summary <- summary(anova_result)
> print("ANOVA Summary:")
[1] "ANOVA Summary:"
> print(anova_summary)
```

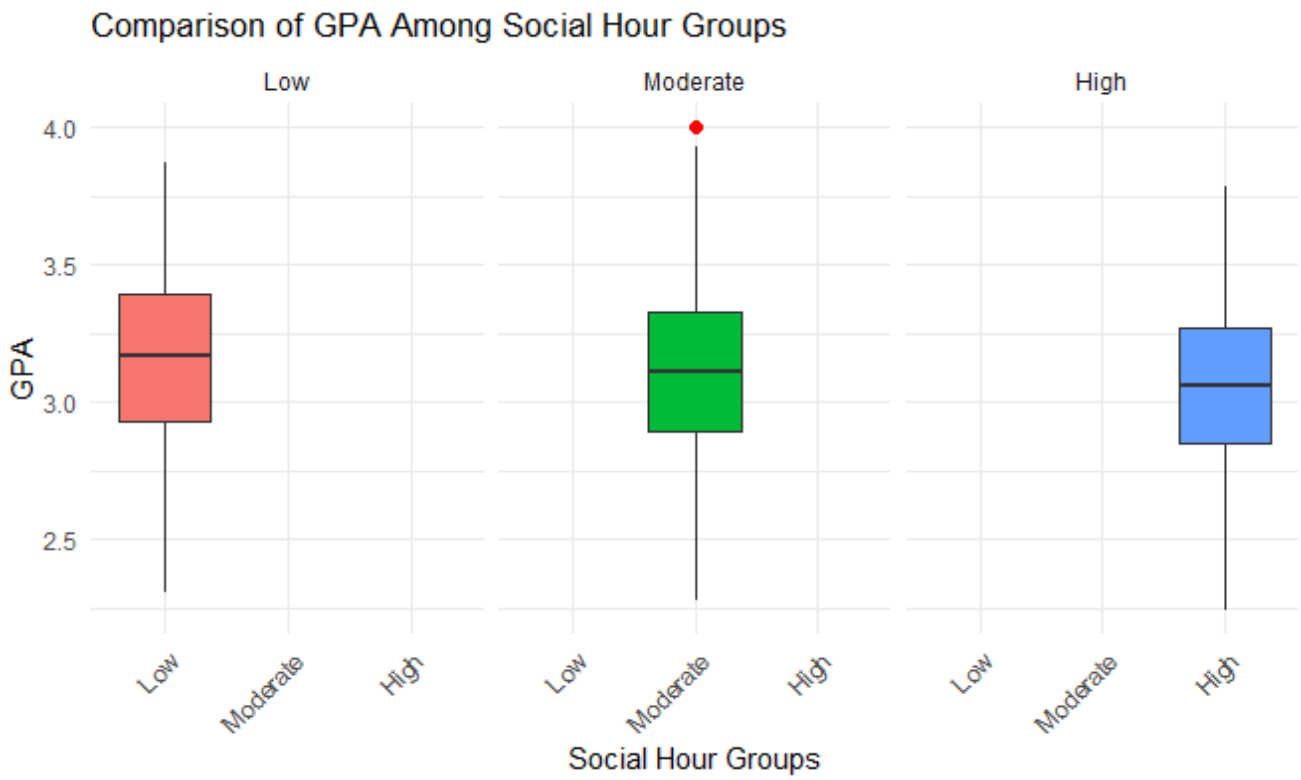
```
                   Df Sum Sq Mean Sq F value Pr(>F)
Study_Hours_Group   2  44.07  22.037   461.6 <2e-16 ***
Residuals         997  47.60   0.048
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Perform pairwise comparisons using Tukey's HSD
> tukey_result <- TukeyHSD(anova_result)
> print("Pairwise Comparisons (Tukey's HSD):")
[1] "Pairwise Comparisons (Tukey's HSD):"
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ Study_Hours_Group, data = data)

$Study_Hours_Group
                   diff       lwr       upr p adj
Moderate-Low  0.2462162 0.2064699 0.2859625     0
High-Low      0.5139472 0.4742306 0.5536637     0
High-Moderate 0.2677309 0.2280144 0.3074475     0

> # Plot boxplots for GPA by Study_Hours_Group
> ggplot(data, aes(x = Study_Hours_Group, y = GPA, fill = Study_Hours_Group)) +
+   geom_boxplot(outlier.color = "red", outlier.size = 2) +
+   labs(title = "Comparison of GPA Among Study Hour Groups",
+        x = "Study Hour Groups",
+        y = "GPA") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1),
+         legend.position = "none") +
+   facet_wrap(~ Study_Hours_Group, nrow = 1)
```

The one-way ANOVA reveals significant differences in GPA among the three study hour groups (Low, Moderate, and High), with an F-value of 461.6 and p<2e-16. The average study hours per day for the Low, Moderate, and High groups are 5.84, 7.47, and 9.16 hours, respectively. Tukey's HSD test confirms that all pairwise comparisons between groups are significant (p<0.001), with the High group having the highest GPA, followed by the Moderate and Low groups. The GPA differences are 0.51 (High vs. Low), 0.27 (High vs. Moderate), and 0.25 (Moderate vs. Low), indicating a positive relationship between study hours and GPA. The boxplot visually supports these findings, showing distinct GPA distributions across the groups.



6. Research Question 6: Does the GPA significantly differ among students grouped by low, moderate, and high levels of social hours?
Null Hypothesis (H_0): There is no significant difference in GPA among the three groups of social hours (low, moderate, high).
Alternative Hypothesis (H_1): At least one group has a significantly different GPA compared to the others.
Method: One-way ANOVA will compare the mean GPA across the three groups of social hours. Tukey's HSD test will be used for pairwise comparisons to identify specific group differences. A boxplot will visualize the GPA distribution among the three groups.

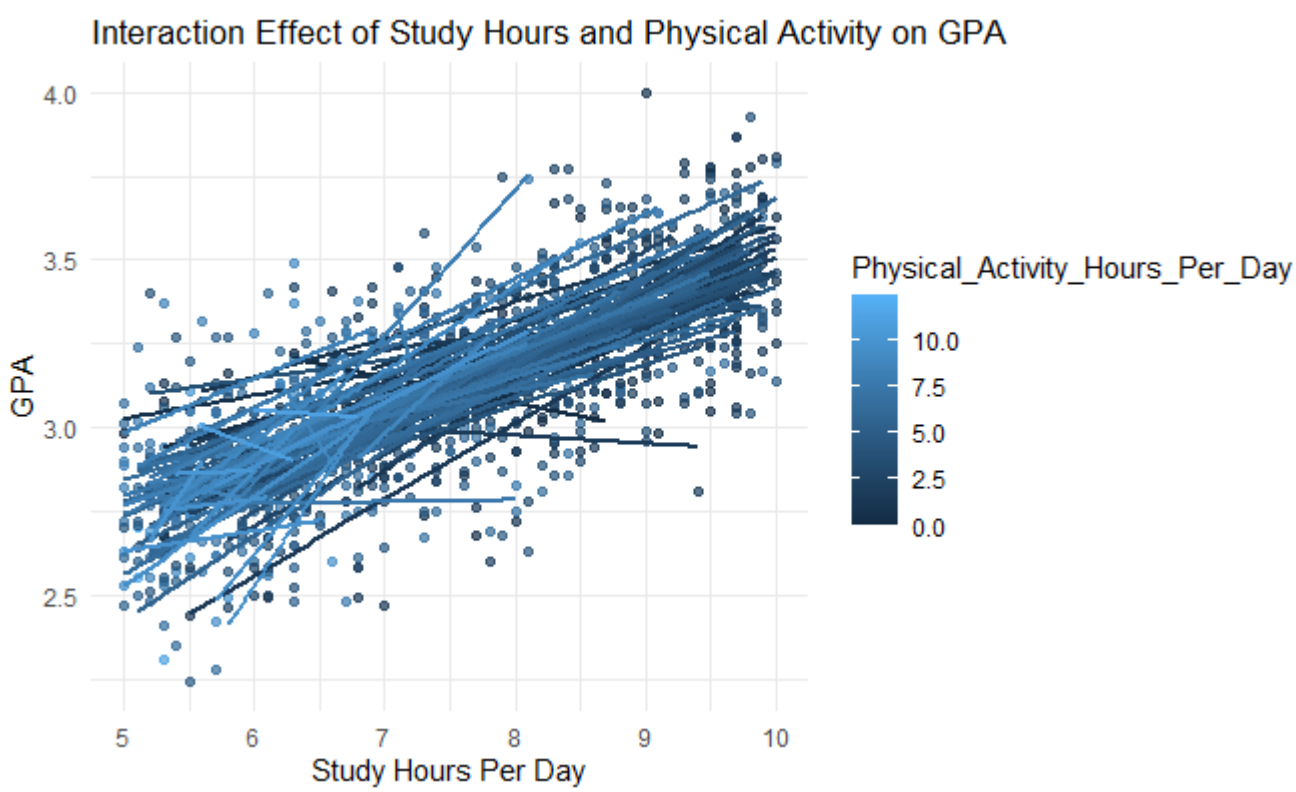
```
> # Sort the dataset by Social_Hours_Per_Day in ascending order
> data <- data[order(data$Social_Hours_Per_Day), ]
> # Create three equal groups for Social_Hours_Per_Day: Low, Moderate, High
> n <- nrow(data)
> data$Social_Hours_Group <- cut(1:n,
+                                breaks = c(0, floor(n / 3), floor(2 * n / 3), n),
+                                labels = c("Low", "Moderate", "High"),
+                                include.lowest = TRUE)
> # Check the distribution of groups
> table(data$Social_Hours_Group)

     Low Moderate     High
     333      333      334
> # Calculate the average Social_Hours_Per_Day in each group
> average_hours_social <- aggregate(Social_Hours_Per_Day ~ Social_Hours_Group, data = data, mean)
> print("Average Social Hours in Each Group:")
[1] "Average Social Hours in Each Group:"
> print(average_hours_social)
  Social_Hours_Group Social_Hours_Per_Day
1                Low            0.7963964
2           Moderate            2.6300300
3               High            4.7101796
> # Perform one-way ANOVA
> anova_social_result <- aov(GPA ~ Social_Hours_Group, data = data)
> anova_social_summary <- summary(anova_social_result)
> print("ANOVA Summary for Social Hours Groups:")
[1] "ANOVA Summary for Social Hours Groups:"
> print(anova_social_summary)
                   Df Sum Sq Mean Sq F value Pr(>F)
Social_Hours_Group   2   1.80  0.9016      10  5e-05 ***
Residuals          997  89.87  0.0901
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Perform pairwise comparisons using Tukey's HSD
> tukey_social_result <- TukeyHSD(anova_social_result)
> print("Pairwise Comparisons (Tukey's HSD) for Social Hours Groups:")
[1] "Pairwise Comparisons (Tukey's HSD) for Social Hours Groups:"
> print(tukey_social_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ Social_Hours_Group, data = data)

$Social_Hours_Group
                     diff        lwr           upr     p adj
Moderate-Low  -0.04927928 -0.1038931  5.334518e-03 0.0867225
High-Low      -0.10394140 -0.1585143 -4.936849e-02 0.0000259
High-Moderate -0.05466212 -0.1092350 -8.921411e-05 0.0495186

> # Plot boxplots for GPA by Social_Hours_Group
> ggplot(data, aes(x = Social_Hours_Group, y = GPA, fill = Social_Hours_Group)) +
+   geom_boxplot(outlier.color = "red", outlier.size = 2) +
+   labs(title = "Comparison of GPA Among Social Hour Groups",
+        x = "Social Hour Groups",
+        y = "GPA") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1),
+         legend.position = "none") +
+   facet_wrap(~ Social_Hours_Group, nrow = 1)
>
```



The one-way ANOVA indicates that GPA significantly differs among the three social hour groups (low, moderate, high), with an F-value of 10 and $p<0.01$. The average social hours per day for the low, moderate, and high groups are 0.80, 2.63, and 4.71 hours, respectively. Tukey's HSD test shows significant differences between the high and low groups ($p=0.00003$) and the high and moderate groups ($p=0.0495$), with the high social hour group having lower GPA on average. The boxplot highlights these differences, showing that increased social hours are associated with a slight decline in GPA, especially at higher levels of social hours.

7. **Research Question 7:** How do interactions between lifestyle factors (e.g., study hours and physical activity) affect GPA?
    - **Hypothesis:** There is a significant interaction effect between study hours and physical activity on GPA.
    - **Method:** ANCOVA (Analysis of Covariance).

```
> # Load necessary libraries
> library(ggplot2)
> # Perform ANCOVA with interaction between Study_Hours_Per_Day and Physical_Activity_Hours_Per_Day
> ancova_result <- aov(GPA ~ Study_Hours_Per_Day * Physical_Activity_Hours_Per_Day, data = data)
> # Display ANCOVA summary
> summary_ancova <- summary(ancova_result)
> print("ANCOVA Summary:")
[1] "ANCOVA Summary:"
> print(summary_ancova)
                                                     Df Sum Sq Mean Sq  F value Pr(>F)
Study_Hours_Per_Day                                   1  50.03   50.03 1201.529 <2e-16 ***
Physical_Activity_Hours_Per_Day                       1   0.15    0.15    3.553 0.0597 .
Study_Hours_Per_Day:Physical_Activity_Hours_Per_Day   1   0.02    0.02    0.401 0.5265
Residuals                                           996  41.47    0.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Check interaction effect visualization
> ggplot(data, aes(x = Study_Hours_Per_Day, y = GPA, color = Physical_Activity_Hours_Per_Day)) +
+   geom_point(alpha = 0.7) +
+   geom_smooth(method = "lm", aes(group = Physical_Activity_Hours_Per_Day), se = FALSE) +
+   labs(title = "Interaction Effect of Study Hours and Physical Activity on GPA",
+        x = "Study Hours Per Day",
+        y = "GPA") +
+   theme_minimal() +
+   theme(legend.position = "right")
`geom_smooth()` using formula = 'y ~ x'
```



The ANCOVA results show a highly significant main effect of Study_Hours_Per_Day on GPA (F=1201.529,p<2e-16), indicating that study hours strongly influence GPA. The main effect of Physical_Activity_Hours_Per_Day is not significant (F=3.553,

p=0.0597), suggesting that physical activity alone does not have a substantial effect on GPA. Additionally, the interaction term (Study_Hours_Per_Day : Physical_Activity_Hours_Per_Day) is not significant (F=0.401,p=0.5265), indicating that the effect of study hours on GPA does not depend significantly on physical activity levels. The visualization further supports this by showing similar trends across physical activity levels.

2. Describe the data set.

Dataset Overview:

- **Size and Dimensions:** The dataset contains 1000 samples and 6 scalar numerical continuous variables.
- **Variables:**
    1. **Study_Hours_Per_Day:** Average hours spent studying daily.
    2. **Extracurricular_Hours_Per_Day:** Daily hours spent on extracurricular activities (e.g., clubs, volunteering).
    3. **Sleep_Hours_Per_Day:** Daily hours of sleep.
    4. **Social_Hours_Per_Day:** Average hours spent socializing daily.
    5. **Physical_Activity_Hours_Per_Day:** Daily hours spent engaging in physical activities (e.g., exercise, sports).
    6. **GPA:** A continuous variable representing students' Grade Point Average.

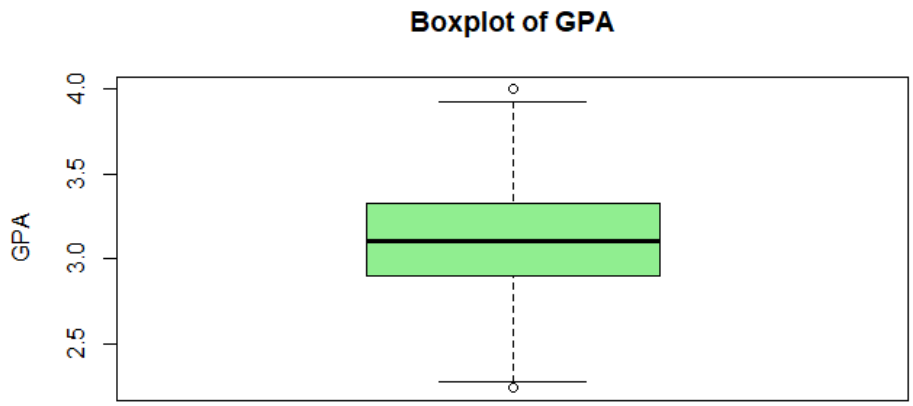
Data Cleaning Steps:

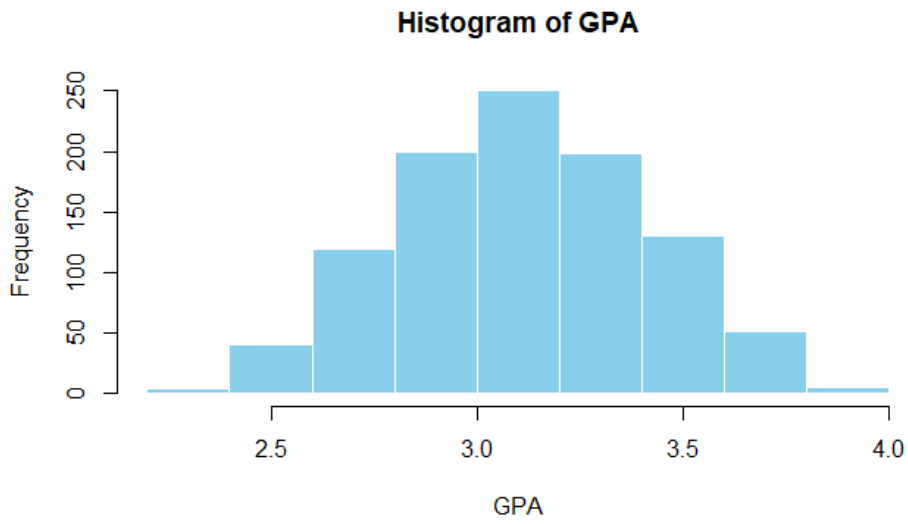
1. **Removed Missing Data:** Checked for missing or null values and removed rows with incomplete data.
2. **Outlier Detection:** Identified and flagged outliers using z-scores (values > 3 standard deviations from the mean).
3. **Standardization:** Standardized continuous variables for better comparability in regression models (e.g., scaled variables to have mean 0 and standard deviation 1).
4. **Descriptive Statistics:**
    - Calculated summary statistics (mean, median, standard deviation) for each variable.
    - Generated visualizations (histograms and boxplots) to examine distributions.

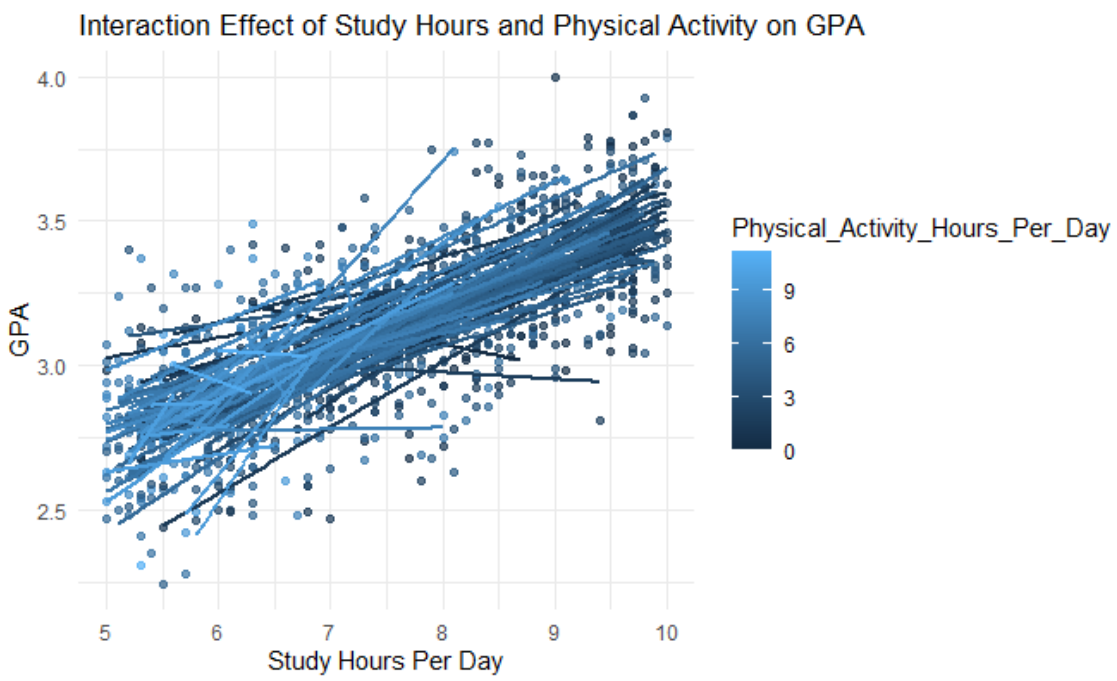
```
> # Load necessary libraries
> library(ggplot2)
> library(dplyr)
> # Check for missing values
> print("Checking for missing values:")
[1] "Checking for missing values:"
> missing_data <- colSums(is.na(data))
> print(missing_data)
            Study_Hours_Per_Day   Extracurricular_Hours_Per_Day             Sleep_Hours_Per_Day            Social_Hours_Per_Day
                              0                               0                               0                               0
Physical_Activity_Hours_Per_Day                             GPA
                              0                               0
> # Remove rows with missing values
> data <- na.omit(data)
> # Outlier detection using z-scores (values > 3 standard deviations)
> z_scores <- as.data.frame(scale(data))  # Standardizing variables
> outliers <- apply(z_scores, 2, function(x) abs(x) > 3)
> data <- data[!rowSums(outliers), ]
> # Step 2: Standardization
> data_scaled <- as.data.frame(scale(data))
> # Step 3: Descriptive Statistics
> print("Descriptive Statistics:")
[1] "Descriptive Statistics:"
> summary_stats <- data %>% summarise(across(everything(), list(
+   mean = ~ mean(.),
+   median = ~ median(.),
+   sd = ~ sd(.)
+ )))
> print(summary_stats)
```

```
  Study_Hours_Per_Day_mean Study_Hours_Per_Day_median Study_Hours_Per_Day_sd Extracurricular_Hours_Per_Day_mean
1                 7.498397                       7.45                1.440668                          1.97004
  Extracurricular_Hours_Per_Day_median Extracurricular_Hours_Per_Day_sd Sleep_Hours_Per_Day_mean Sleep_Hours_Per_Day_median
1                                  1.9                         1.158724                 7.533367                       7.6
  Sleep_Hours_Per_Day_sd Social_Hours_Per_Day_mean Social_Hours_Per_Day_median Social_Hours_Per_Day_sd Physical_Activity_Hours_Per_Day_mean
1               1.469946                  2.719238                        2.6                1.702482                           4.278958
  Physical_Activity_Hours_Per_Day_median Physical_Activity_Hours_Per_Day_sd GPA_mean GPA_median    GPA_sd
1                                    4.1                           2.466404 3.116283       3.11 0.3024127
> # Visualizations
> # Histograms for each variable
> par(mfrow = c(3, 2))
> for (var in names(data)) {
+   hist(data[[var]], main = paste("Histogram of", var), xlab = var, col = "skyblue", border = "white")
+ }
> # Boxplots for each variable
> par(mfrow = c(3, 2))
> for (var in names(data)) {
+   boxplot(data[[var]], main = paste("Boxplot of", var), ylab = var, col = "lightgreen")
+ }
> # Step 4: ANCOVA Analysis
> # Perform ANCOVA with interaction between Study_Hours_Per_Day and Physical_Activity_Hours_Per_Day
> ancova_result <- aov(GPA ~ Study_Hours_Per_Day * Physical_Activity_Hours_Per_Day, data = data)
> # Display ANCOVA summary
> summary_ancova <- summary(ancova_result)
> print("ANCOVA Summary:")
[1] "ANCOVA Summary:"
> print(summary_ancova)
```

```
                                                        Df Sum Sq Mean Sq  F value Pr(>F)
Study_Hours_Per_Day                                      1  49.58   49.58 1189.672 <2e-16 ***
Physical_Activity_Hours_Per_Day                          1   0.17    0.17    3.998 0.0458 *
Study_Hours_Per_Day:Physical_Activity_Hours_Per_Day      1   0.01    0.01    0.182 0.6698
Residuals                                              994  41.43    0.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Step 5: Visualization of Interaction Effect
> ggplot(data, aes(x = Study_Hours_Per_Day, y = GPA, color = Physical_Activity_Hours_Per_Day)) +
+   geom_point(alpha = 0.7) +
+   geom_smooth(method = "lm", aes(group = Physical_Activity_Hours_Per_Day), se = FALSE) +
+   labs(title = "Interaction Effect of Study Hours and Physical Activity on GPA",
+        x = "Study Hours Per Day",
+        y = "GPA") +
+   theme_minimal() +
+   theme(legend.position = "right")
`geom_smooth()` using formula = 'y ~ x'
```


Boxplot of GPA
GPA
2.5
3.0
3.5
4.0


Histogram of GPA
Frequency
0
50
100
150
200
250
2.5
3.0
3.5
4.0
GPA


Interaction Effect of Study Hours and Physical Activity on GPA
GPA
2.5
3.0
3.5
4.0
5
6
7
8
9
10
Study Hours Per Day
Physical_Activity_Hours_Per_Day
9
6
3
0

3. Describe the statistical methods you plan to use.

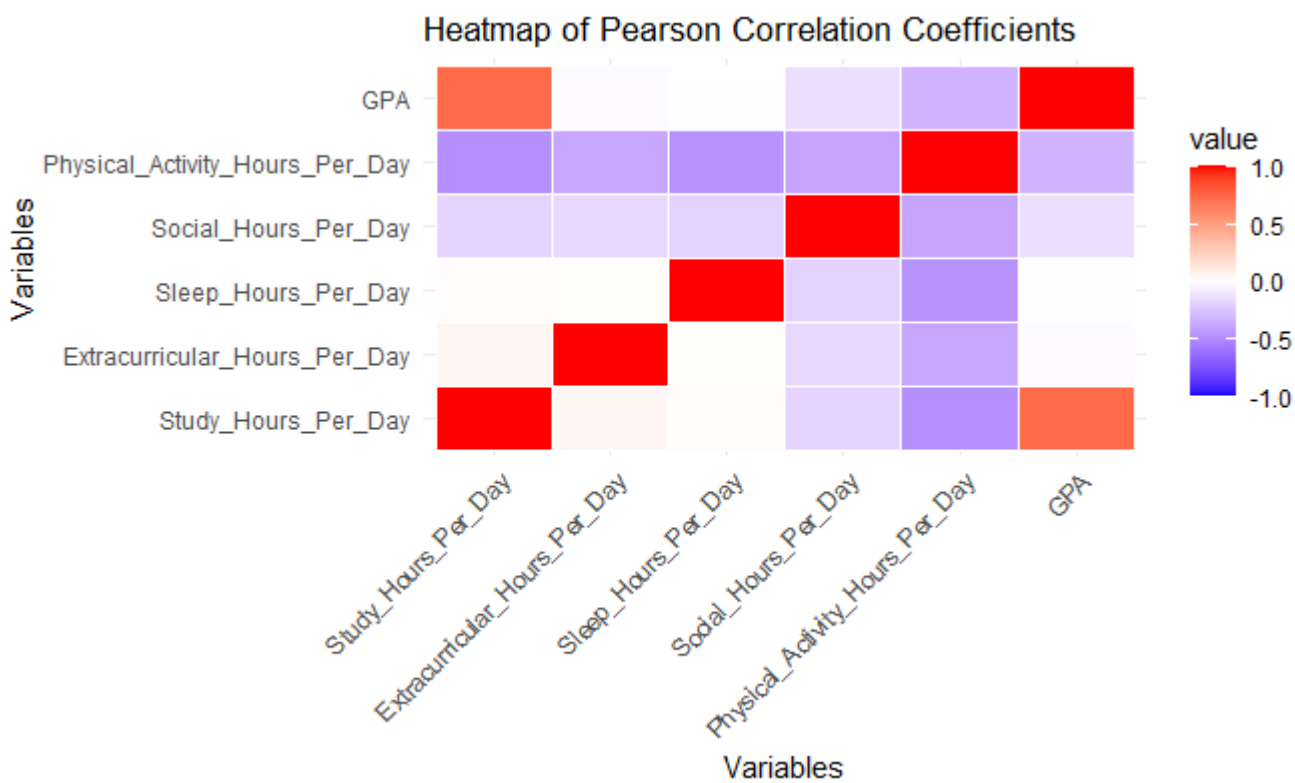
Methods for Research Questions:

1. **Correlation Analysis:**
   - **Purpose:** To evaluate the strength and direction of relationships between daily lifestyle variables and GPA.
   - **Hypothesis:** All lifestyle variables are significantly correlated with GPA.
   - **Implementation:**
     - Compute Pearson correlation coefficients for all pairs of variables.

```
> # Step 1: Compute Pearson Correlation Matrix
> correlation_matrix <- cor(data, method = "pearson")
> # Step 2: Display Correlation Matrix
> print("Correlation Matrix:")
[1] "Correlation Matrix:"
> print(correlation_matrix)
                              Study_Hours_Per_Day Extracurricular_Hours_Per_Day Sleep_Hours_Per_Day Social_Hours_Per_Day
Study_Hours_Per_Day                    1.00000000                   0.046655448         0.014268256          -0.1854648
Extracurricular_Hours_Per_Day          0.04665545                   1.000000000         0.007538127          -0.1649599
Sleep_Hours_Per_Day                    0.01426826                   0.007538127         1.000000000          -0.1943317
Social_Hours_Per_Day                  -0.18546477                  -0.164959889        -0.194331731           1.0000000
Physical_Activity_Hours_Per_Day       -0.48288554                  -0.384486888        -0.469112184          -0.3855527
GPA                                    0.73877023                  -0.015333360        -0.006583260          -0.1397429
                              Physical_Activity_Hours_Per_Day         GPA
Study_Hours_Per_Day                          -0.4828855  0.73877023
Extracurricular_Hours_Per_Day                -0.3844869 -0.01533336
Sleep_Hours_Per_Day                          -0.4691122 -0.00658326
Social_Hours_Per_Day                         -0.3855527 -0.13974291
Physical_Activity_Hours_Per_Day               1.0000000 -0.32156074
GPA                                          -0.3215607  1.00000000
> # Step 3: Test for Significance of Correlations
> correlation_tests <- lapply(names(data)[-which(names(data) == "GPA")], function(var) {
+   cor.test(data[[var]], data$GPA, method = "pearson")
+ })
> # Extract p-values
> p_values <- sapply(correlation_tests, function(test) test$p.value)
> # Combine variables and p-values
> correlation_significance <- data.frame(
+   Variable = names(data)[-which(names(data) == "GPA")],
+   P_Value = p_values
+ )
> print("Significance of Correlations (P-Values):")
[1] "Significance of Correlations (P-Values):"
> print(correlation_significance)
                         Variable       P_Value
1             Study_Hours_Per_Day 3.225236e-173
2   Extracurricular_Hours_Per_Day  6.281691e-01
3             Sleep_Hours_Per_Day  8.352892e-01
4            Social_Hours_Per_Day  9.193926e-06
5 Physical_Activity_Hours_Per_Day  1.728890e-25
> # Step 4: Visualize Correlation Matrix
> library(ggplot2)
> library(reshape2)
> # Convert matrix to long format for heatmap
> cor_data_long <- melt(correlation_matrix)
> ggplot(cor_data_long, aes(x = Var1, y = Var2, fill = value)) +
+   geom_tile(color = "white") +
+   scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1)) +
+   labs(title = "Heatmap of Pearson Correlation Coefficients",
+        x = "Variables",
+        y = "Variables") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

The Pearson correlation analysis reveals the strength and direction of relationships between lifestyle variables and GPA. The strongest

positive correlation is observed between Study_Hours_Per_Day and GPA ($r=0.739, p<0.001$), indicating that increased study hours are strongly associated with higher GPA. Physical_Activity_Hours_Per_Day has a moderate negative correlation with GPA ($r=-0.322, p<0.001$), suggesting that higher physical activity levels are associated with lower GPA. Social_Hours_Per_Day has a weaker negative correlation with GPA ($r=-0.140, p=9.19\times10-6$). Other lifestyle variables, such as Extracurricular_Hours_Per_Day and Sleep_Hours_Per_Day, show negligible and non-significant correlations with GPA. The heatmap visually highlights these relationships, with warm colors (red) representing strong positive correlations and cool colors (blue) indicating negative correlations. These findings suggest that study hours are the most critical predictor of GPA among the variables analyzed.

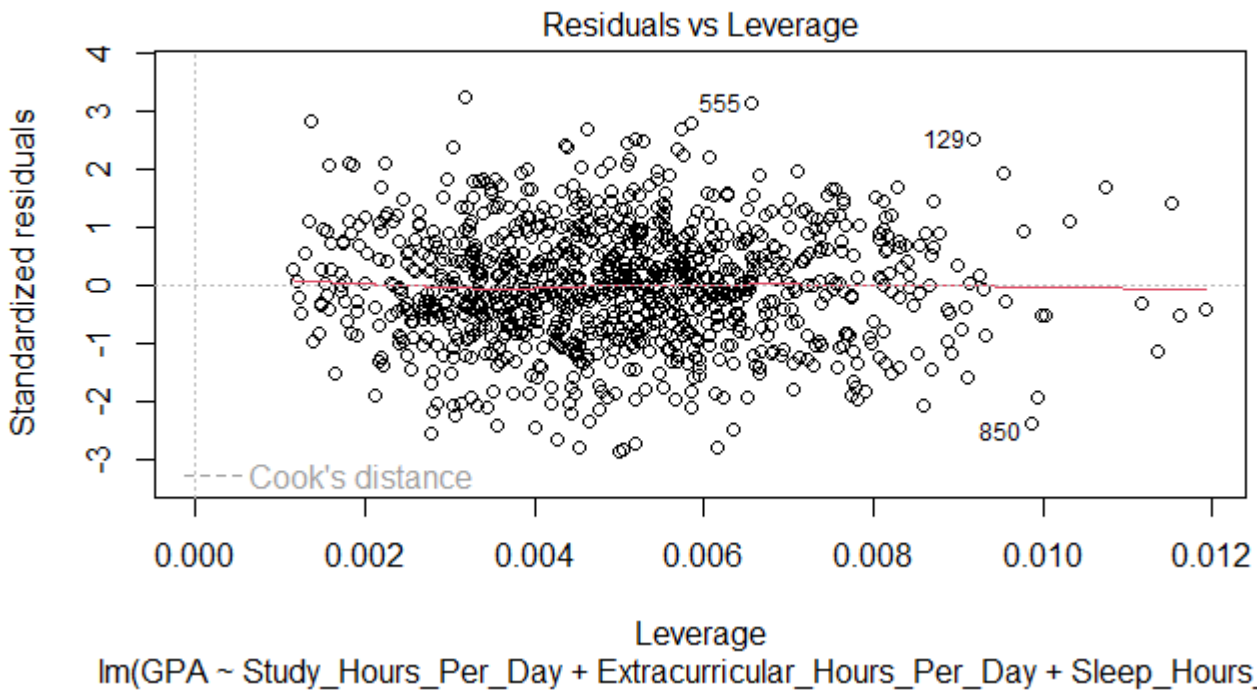


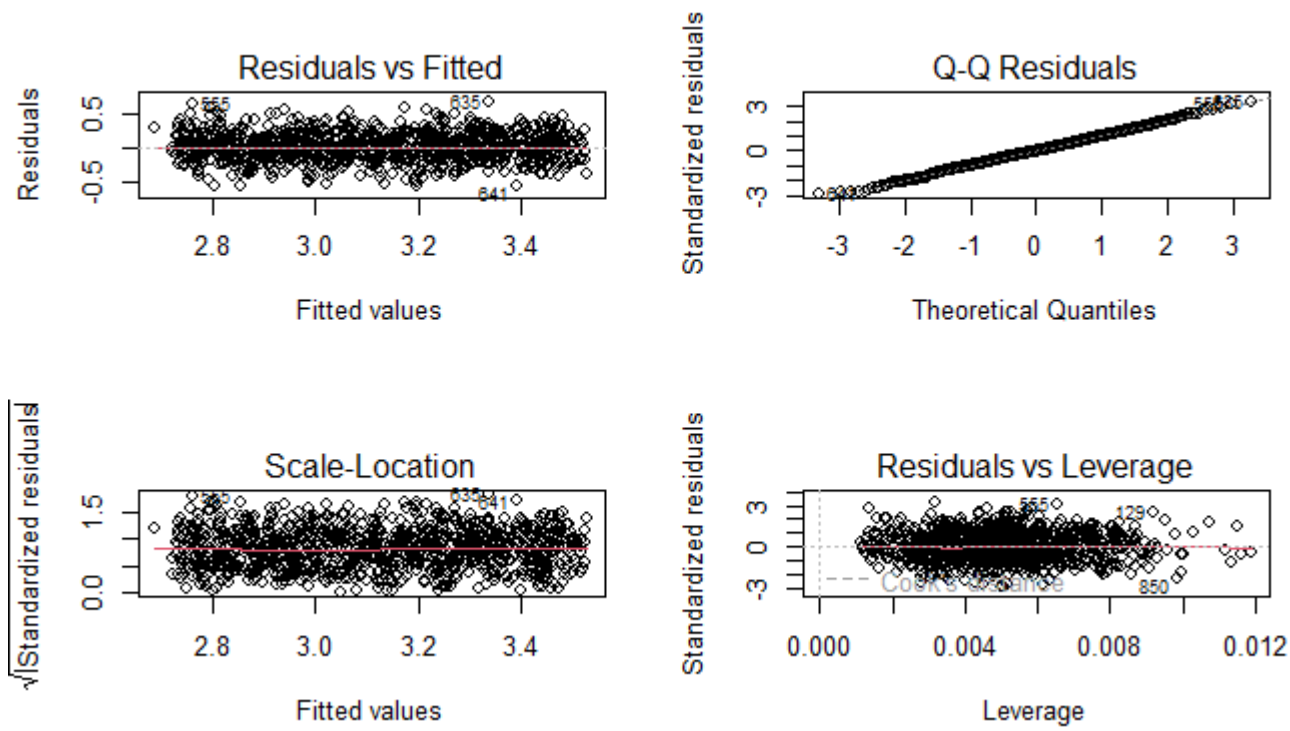
2. **Multiple Linear Regression:**
   - **Purpose:** To assess the combined effect of all lifestyle variables on GPA and identify significant predictors.
   - **Hypothesis:** Study hours, sleep hours, and physical activity are significant predictors of GPA.
   - **Implementation:**
     - Model:
       Model: $\text{GPA} = \beta_0 + \beta_1(\text{Study Hours}) + \beta_2(\text{Extracurricular Hours}) + \beta_3(\text{Sleep Hours}) + \beta_4(\text{Social Hours}) + \beta_5(\text{Physical Activity Hours}) + \epsilon$

GPA=β0+β1(Study Hours)+β2(Extracurricular Hours)+β3(Sleep Hours)+β4(Social Hours)+β5(Physical Activity Hours)+ϵ

- Evaluate the model's fit using R^2, adjusted R^2 and p-values for predictors.

```
> # Fit the Multiple Linear Regression Model
> model <- lm(GPA ~ Study_Hours_Per_Day + Extracurricular_Hours_Per_Day +
+               Sleep_Hours_Per_Day + Social_Hours_Per_Day + Physical_Activity_Hours_Per_Day,
+             data = data)
> # Display the summary of the regression model
> summary_model <- summary(model)
> print("Summary of Multiple Linear Regression Model:")
[1] "Summary of Multiple Linear Regression Model:"
> print(summary_model)

Call:
lm(formula = GPA ~ Study_Hours_Per_Day + Extracurricular_Hours_Per_Day +
    Sleep_Hours_Per_Day + Social_Hours_Per_Day + Physical_Activity_Hours_Per_Day,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58231 -0.13825 -0.00363  0.13520  0.66474

Coefficients: (1 not defined because of singularities)
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     2.018104   0.054346  37.134   <2e-16 ***
Study_Hours_Per_Day             0.155044   0.004551  34.071   <2e-16 ***
Extracurricular_Hours_Per_Day  -0.013619   0.005641  -2.414   0.0159 *
Sleep_Hours_Per_Day            -0.004056   0.004477  -0.906   0.3652
Social_Hours_Per_Day           -0.002696   0.003979  -0.678   0.4981
Physical_Activity_Hours_Per_Day        NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2039 on 995 degrees of freedom
Multiple R-squared:  0.5488,    Adjusted R-squared:  0.5469
F-statistic: 302.5 on 4 and 995 DF,  p-value: < 2.2e-16

> # Extract R-squared and Adjusted R-squared
> r_squared <- summary_model$r.squared
> adj_r_squared <- summary_model$adj.r.squared
> cat("R-squared:", r_squared, "\n")
R-squared: 0.548757
> cat("Adjusted R-squared:", adj_r_squared, "\n")
Adjusted R-squared: 0.5469429
> # Identify significant predictors
> significant_predictors <- summary_model$coefficients[,"Pr(>|t|)"] < 0.05
> cat("Significant Predictors:\n")
Significant Predictors:
> print(rownames(summary_model$coefficients)[significant_predictors])
[1] "(Intercept)"                   "Study_Hours_Per_Day"           "Extracurricular_Hours_Per_Day"
> # Visualize Residuals
> par(mfrow = c(2, 2))
> plot(model)
Waiting to confirm page change...
Waiting to confirm page change...
Waiting to confirm page change...
Waiting to confirm page change...
```


Residuals vs Leverage
Standardized residuals
555
129
850
Cook's distance
Leverage
lm(GPA ~ Study_Hours_Per_Day + Extracurricular_Hours_Per_Day + Sleep_Hours_ ...


Residuals vs Fitted
Residuals
Fitted values
Q-Q Residuals
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage

1. Residuals vs. Fitted: The residuals are scattered randomly around zero, indicating no clear pattern and supporting the assumption of linearity. However, there may be some slight heteroscedasticity (unequal variance).
2. Q-Q Plot: The residuals align closely with the theoretical quantiles, suggesting that the residuals are approximately normally distributed.
3. Scale-Location Plot: The spread of the residuals appears relatively consistent across the range of fitted values, supporting the assumption of homoscedasticity (constant variance of errors).
4. Residuals vs. Leverage: There are no influential outliers with high leverage, as most points lie within the Cook's distance boundaries. Overall, the diagnostic plots indicate that the model assumptions are reasonably met, with no major violations detected. However, some slight heteroscedasticity might warrant further questions.

The multiple linear regression model explains approximately 54.88% of the variance in GPA (R2=0.5488R^2 = 0.5488R2=0.5488, Adjusted R^2=0.5469), with an overall significant model fit (F=302.5,p<2.2×10−16). Among the predictors, Study_Hours_Per_Day is the strongest and most significant predictor (β=0.155,p<2e−16), positively impacting GPA. Extracurricular_Hours_Per_Day has a small but significant negative effect (β=−0.0136,p=0.0159). Other predictors, including Sleep_Hours_Per_Day and Social_Hours_Per_Day, are not statistically significant (p>0.05). Physical_Activity_Hours_Per_Day was excluded due to multicollinearity or singularity issues. The residual standard error of 0.2039 suggests good model precision, and diagnostic plots should be examined to confirm assumptions of linearity, normality, and homoscedasticity.

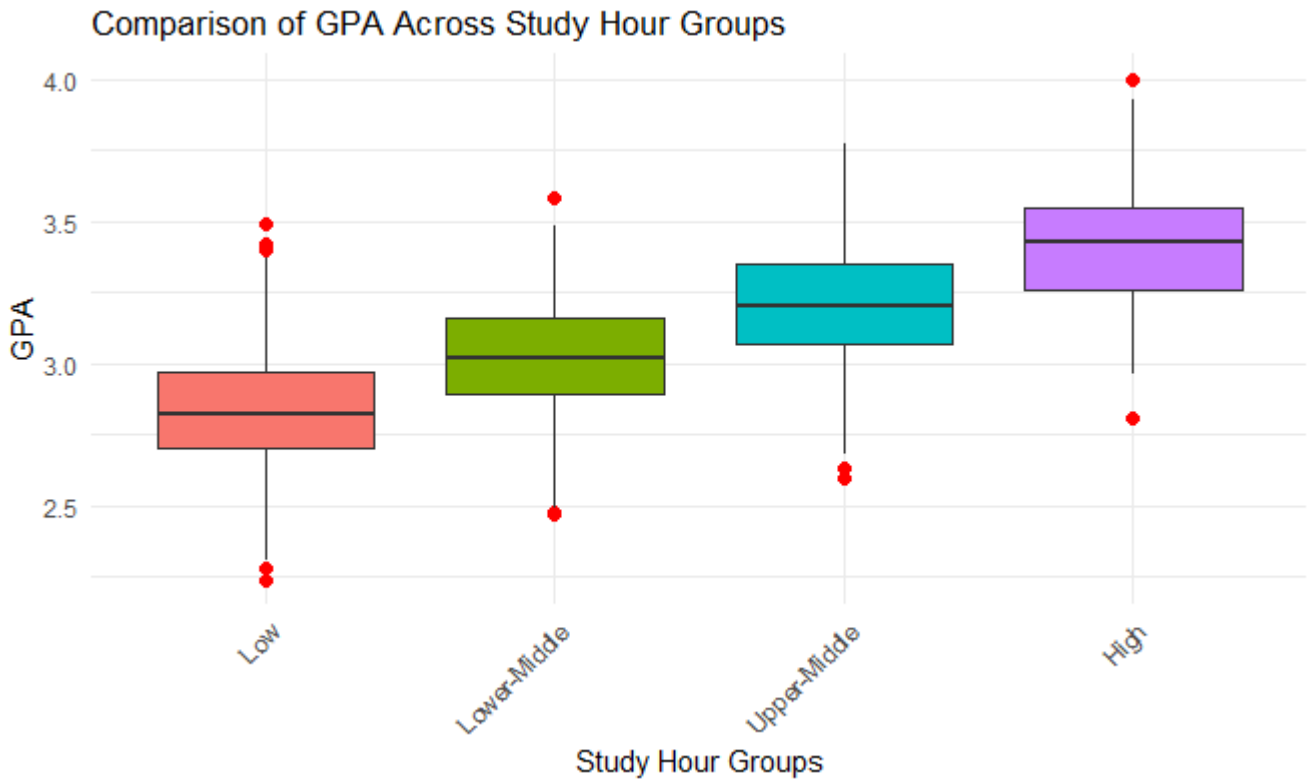
3. **One-Way ANOVA**:
   - **Purpose**: To compare GPA across different levels of a specific lifestyle variable (e.g., high vs. low social hours).
   - **Hypothesis**: Significant differences in GPA exist between groups of students with different levels of study hours.
   - **Implementation**:
     - Divide students into groups based on quartiles or pre-defined thresholds (e.g., high vs. low).
     - Use one-way ANOVA to test for differences in GPA across groups.

Null Hypothesis (H_0): There are no significant differences in GPA among the groups (μ1=μ2=μ3…).

Alternative Hypothesis (H_1): At least one group's GPA is significantly different from the others (μi≠μj, for at least one i,j).

We divide the variable (Study Hours) into quartiles or pre-defined thresholds (e.g., "Low", "Moderate", "High") using either the median or quartiles.

```
> # Step 1: Group Division
> # Divide Study_Hours_Per_Day into quartiles
> data$Study_Hours_Group <- cut(data$Study_Hours_Per_Day,
+                               breaks = quantile(data$Study_Hours_Per_Day, probs = c(0, 0.25, 0.5, 0.75, 1), na.rm = TRUE),
+                               labels = c("Low", "Lower-Middle", "Upper-Middle", "High"),
+                               include.lowest = TRUE)
> # Check the distribution of groups
> print("Group Distribution:")
[1] "Group Distribution:"
> print(table(data$Study_Hours_Group))

         Low Lower-Middle Upper-Middle         High
         274          227          252          247
> # Step 2: Perform One-Way ANOVA
> anova_result <- aov(GPA ~ Study_Hours_Group, data = data)
> # Display ANOVA Summary
> anova_summary <- summary(anova_result)
> print("ANOVA Summary:")
[1] "ANOVA Summary:"
> print(anova_summary)
                 Df Sum Sq Mean Sq F value Pr(>F)
Study_Hours_Group   3  46.81  15.603   346.4 <2e-16 ***
Residuals         996  44.86   0.045
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Step 3: Perform Post-Hoc Test (Tukey's HSD)
> tukey_result <- TukeyHSD(anova_result)
> print("Post-Hoc Test Results (Tukey's HSD):")
[1] "Post-Hoc Test Results (Tukey's HSD):"
> print(tukey_result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = GPA ~ Study_Hours_Group, data = data)

$Study_Hours_Group
                              diff       lwr       upr p adj
Lower-Middle-Low          0.1934570 0.1444406 0.2424735     0
Upper-Middle-Low          0.3694575 0.3217893 0.4171257     0
High-Low                  0.5758040 0.5278851 0.6237228     0
Upper-Middle-Lower-Middle 0.1760005 0.1260239 0.2259770     0
High-Lower-Middle         0.3823469 0.3321313 0.4325626     0
High-Upper-Middle         0.2063465 0.1574460 0.2552469     0

> # Step 4: Visualize Group Differences
> library(ggplot2)
> ggplot(data, aes(x = Study_Hours_Group, y = GPA, fill = Study_Hours_Group)) +
+   geom_boxplot(outlier.color = "red", outlier.size = 2) +
+   labs(title = "Comparison of GPA Across Study Hour Groups",
+        x = "Study Hour Groups",
+        y = "GPA") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```



The one-way ANOVA analysis indicates significant differences in GPA across the four Study Hours groups (Low, Lower-Middle, Upper-Middle, High), with $F=346.4$ and $p<2\times10-16$. Post-hoc analysis using Tukey's HSD reveals that GPA increases progressively from the Low group to the High group, and all pairwise comparisons between groups are statistically significant ($p<0.05$). For instance, the High group has significantly higher GPAs compared to the Low group (diff=0.576,$p<0.001$), and even smaller differences, such as between High and Upper-Middle, are also significant. The boxplot visually confirms these trends, showing a clear positive association between study hours and GPA. This suggests that higher levels of study hours are consistently linked with better academic performance.

4. **ANCOVA (Analysis of Covariance):**
    - **Purpose:** To examine interaction effects (e.g., study hours × physical activity) while controlling for other lifestyle variables.
    - **Hypothesis:** There is a significant interaction effect between study hours and physical activity on GPA.
    - **Implementation:**
        - Include an interaction term in the regression model (e.g., Study Hours × Physical Activity Hours)

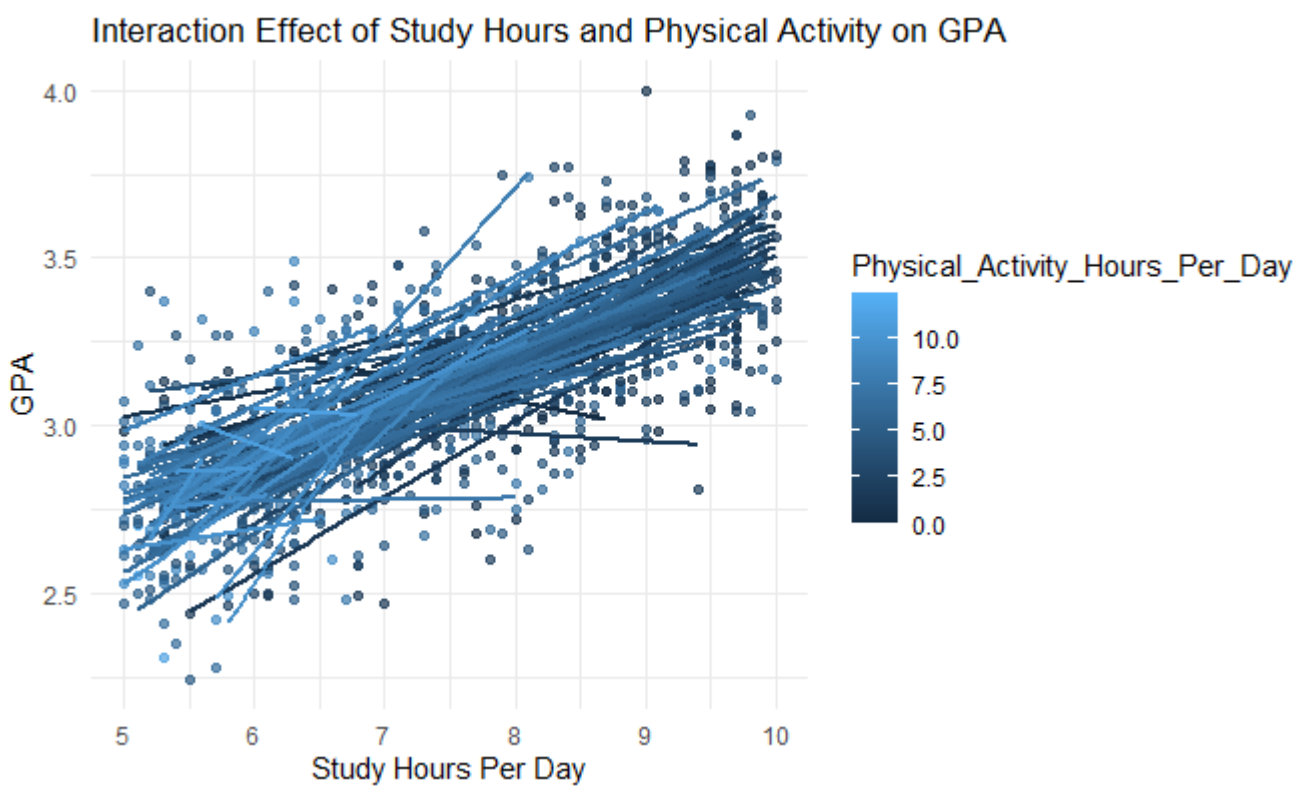
The objective of ANCOVA is to evaluate the interaction effects between Study Hours and Physical Activity Hours on GPA, while

controlling for other lifestyle variables such as Social Hours, Sleep Hours, and Extracurricular Hours.

- **Null Hypothesis (H_0):** The interaction effect between **Study Hours** and **Physical Activity Hours** on **GPA** is not significant ($\beta_{interaction}=0$).
- **Alternative Hypothesis (H_1):** The interaction effect between **Study Hours** and **Physical Activity Hours** on **GPA** is significant ($\beta_{interaction}\neq 0$).

1. **Model:**
    - Include the interaction term **Study Hours × Physical Activity Hours** in the model.
    - Adjust for covariates such as **Social Hours**, **Sleep Hours**, and **Extracurricular Hours.**
2. **Statistical Test:**
    - Use ANCOVA (a type of linear model) to evaluate the interaction effect and the main effects of the predictors.
3. **Interpretation:**
    - Evaluate p-values for the interaction term and other predictors to assess their significance.

```
> # Fit the ANCOVA model with interaction term
> ancova_model <- aov(GPA ~ Study_Hours_Per_Day * Physical_Activity_Hours_Per_Day +
+                       Extracurricular_Hours_Per_Day +
+                       Sleep_Hours_Per_Day +
+                       Social_Hours_Per_Day, data = data)
> # Display the summary of the ANCOVA model
> summary_ancova <- summary(ancova_model)
> print("ANCOVA Summary:")
[1] "ANCOVA Summary:"
> print(summary_ancova)
                                                 Df Sum Sq Mean Sq  F value Pr(>F)
Study_Hours_Per_Day                               1  50.03   50.03 1202.687 <2e-16 ***
Physical_Activity_Hours_Per_Day                   1   0.15    0.15    3.557 0.0596 .
Extracurricular_Hours_Per_Day                     1   0.12    0.12    2.936 0.0869 .
Sleep_Hours_Per_Day                               1   0.00    0.00    0.064 0.8004
Study_Hours_Per_Day:Physical_Activity_Hours_Per_Day  1   0.02    0.02    0.362 0.5476
Residuals                                       994  41.35    0.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Visualize the interaction effect
> library(ggplot2)
> ggplot(data, aes(x = Study_Hours_Per_Day, y = GPA, color = Physical_Activity_Hours_Per_Day)) +
+   geom_point(alpha = 0.7) +
+   geom_smooth(method = "lm", aes(group = Physical_Activity_Hours_Per_Day), se = FALSE) +
+   labs(title = "Interaction Effect of Study Hours and Physical Activity on GPA",
+        x = "Study Hours Per Day",
+        y = "GPA") +
+   theme_minimal() +
+   theme(legend.position = "right")
`geom_smooth()` using formula = 'y ~ x'
```



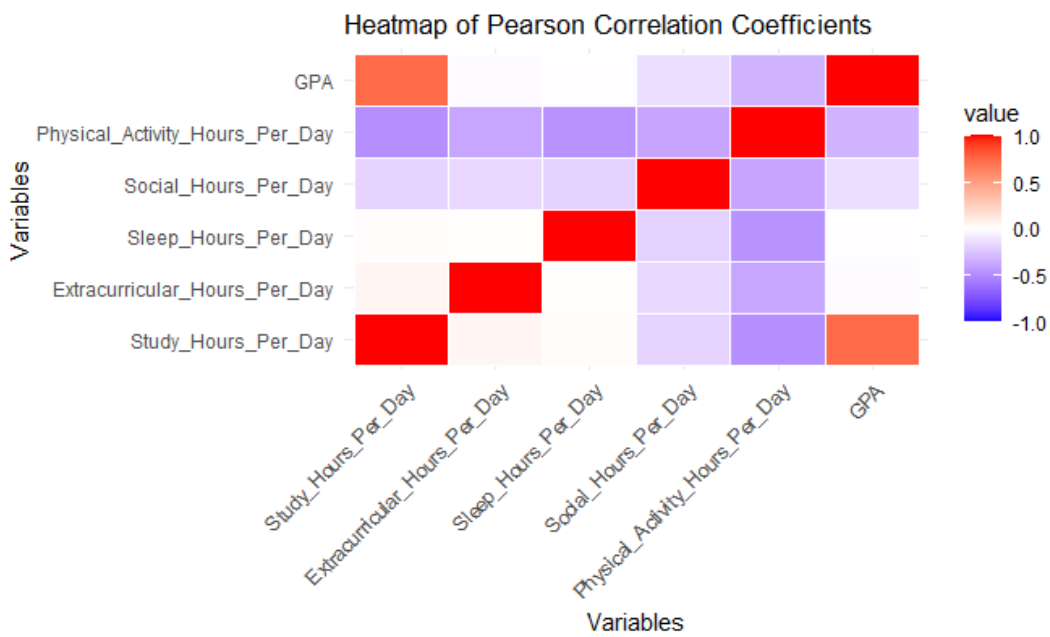
The ANCOVA results indicate that Study Hours Per Day has a highly significant effect on GPA (F=1202.687, p<2×10−16), confirming its critical role in predicting academic performance. However, the interaction effect between Study Hours Per Day and Physical Activity

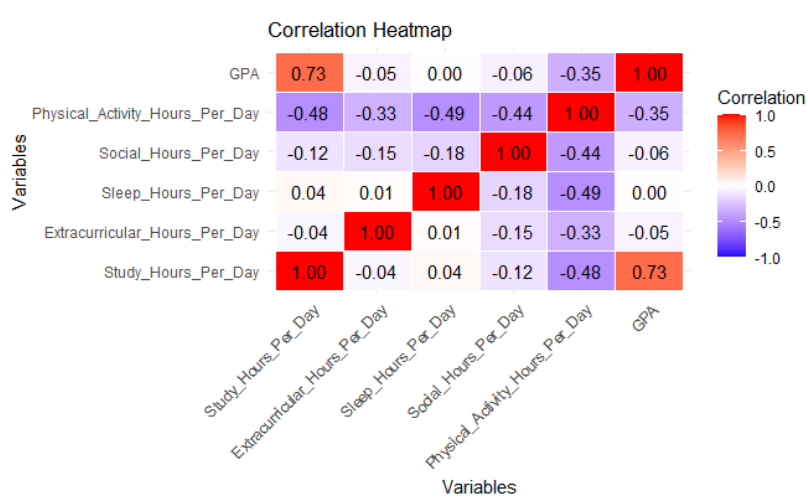
Hours Per Day is not significant (F=0.362, p=0.5476), suggesting that the relationship between study hours and GPA does not depend on the level of physical activity. Among the covariates, Extracurricular Hours Per Day shows a marginally significant effect (p=0.0869), while Sleep Hours Per Day and Physical Activity Hours Per Day independently do not have significant effects on GPA. The interaction plot further illustrates that the effect of study hours on GPA is consistent across different levels of physical activity, supporting the statistical findings.

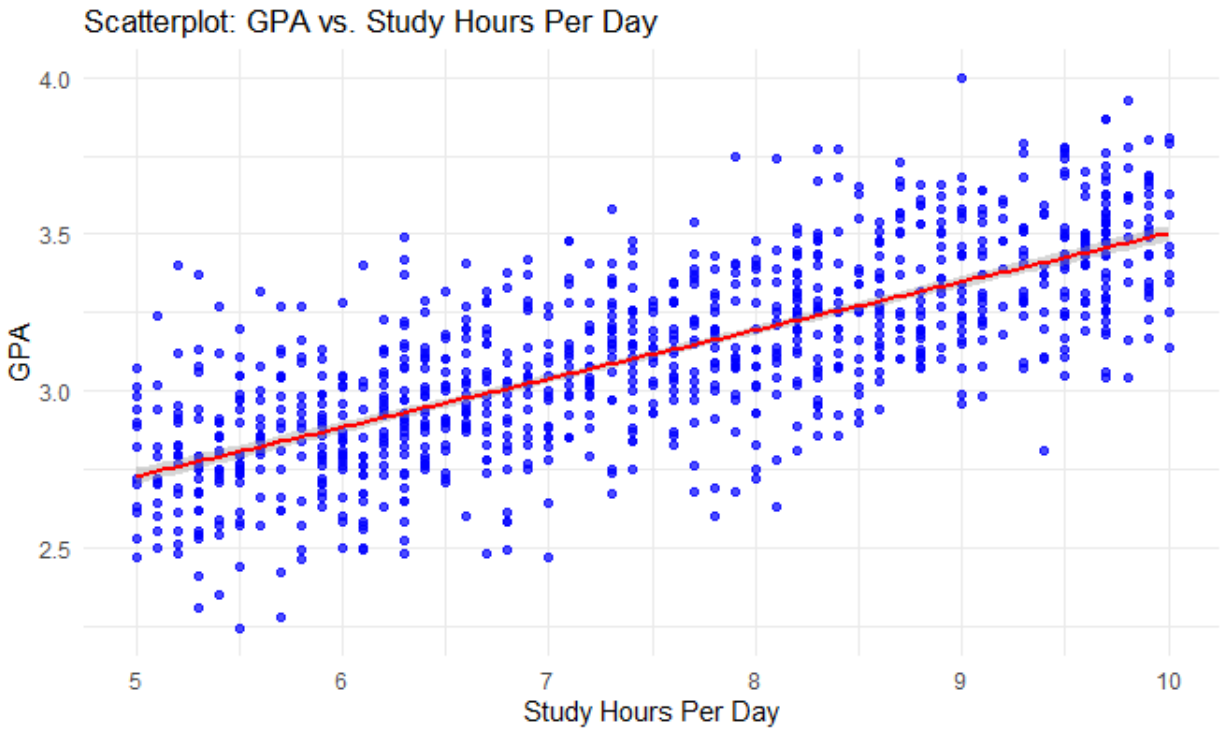
5. **Visualization:**
   - **Boxplots:** To visualize GPA distributions across different lifestyle variables.
   - **Heatmaps:** To show correlation strengths between variables.
   - **Scatterplots with Regression Lines:** To visualize relationships between GPA and individual predictors.

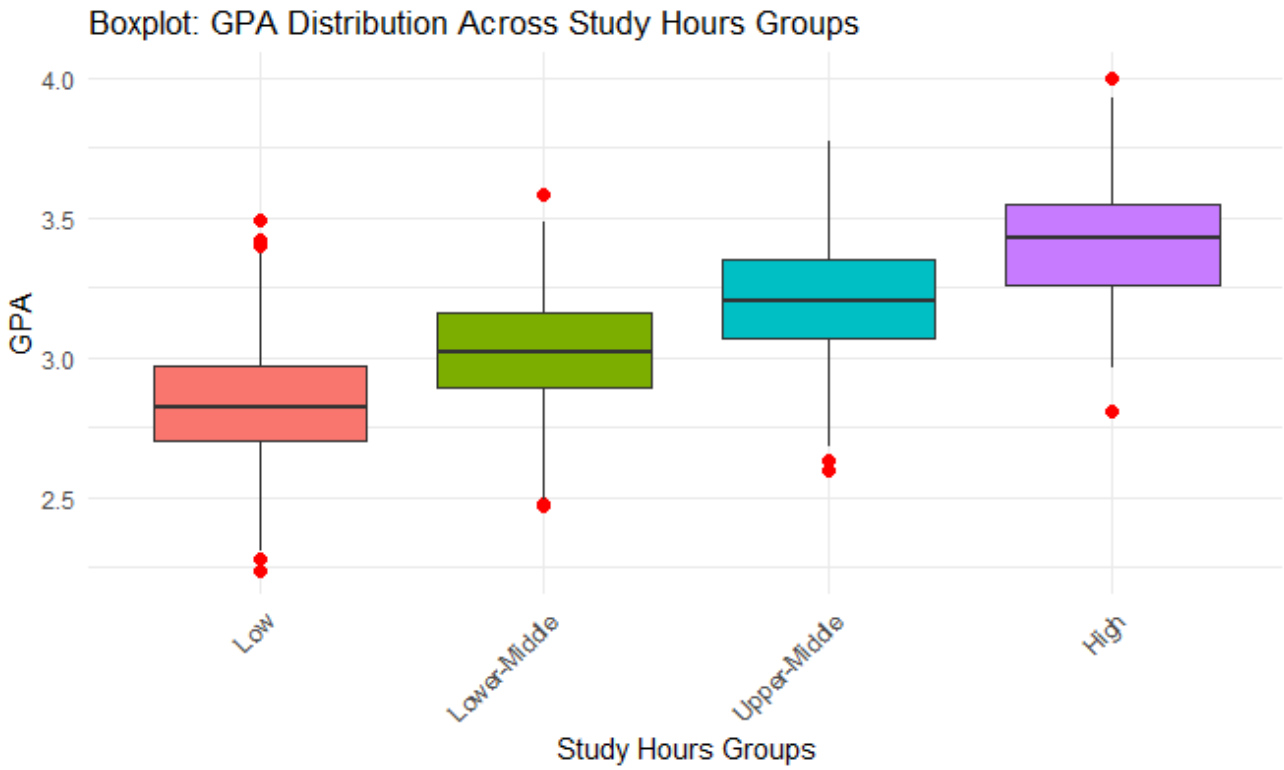
```
> library(ggplot2)
> # Example: Grouping by Study_Hours_Group (already created)
> ggplot(data, aes(x = Study_Hours_Group, y = GPA, fill = Study_Hours_Group)) +
+   geom_boxplot(outlier.color = "red", outlier.size = 2) +
+   labs(title = "Boxplot: GPA Distribution Across Study Hours Groups",
+        x = "Study Hours Groups",
+        y = "GPA") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
> library(reshape2)
> # Compute Pearson correlation matrix
> correlation_matrix <- cor(data, method = "pearson")
> # Convert correlation matrix to long format
> cor_data_long <- melt(correlation_matrix)
> # Plot heatmap
> ggplot(cor_data_long, aes(x = Var1, y = Var2, fill = value)) +
+   geom_tile(color = "white") +
+   scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1)) +
+   labs(title = "Heatmap of Pearson Correlation Coefficients",
+        x = "Variables",
+        y = "Variables") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
> # Example: Scatterplot for GPA vs. Study_Hours_Per_Day
> ggplot(data, aes(x = Study_Hours_Per_Day, y = GPA)) +
+   geom_point(alpha = 0.7, color = "blue") +
+   geom_smooth(method = "lm", se = TRUE, color = "red") +
+   labs(title = "Scatterplot: GPA vs. Study Hours Per Day",
+        x = "Study Hours Per Day",
+        y = "GPA") +
+   theme_minimal()
`geom_smooth()` using formula = 'y ~ x'
>
```

```
>
> # Divide Study_Hours_Per_Day into quartiles
> data$Study_Hours_Group <- cut(data$Study_Hours_Per_Day,
+                               breaks = quantile(data$Study_Hours_Per_Day, probs = c(0, 0.25, 0.5, 0.75, 1), na.rm = TRUE),
+                               labels = c("Low", "Lower-Middle", "Upper-Middle", "High"),
+                               include.lowest = TRUE)
>
> # Verify the created groups
> print(table(data$Study_Hours_Group))

         Low Lower-Middle Upper-Middle         High
         274          227          252          247
>
> library(ggplot2)
>
> ggplot(data, aes(x = Study_Hours_Group, y = GPA, fill = Study_Hours_Group)) +
+     geom_boxplot(outlier.color = "red", outlier.size = 2) +
+     labs(title = "Boxplot: GPA Distribution Across Study Hours Groups",
+          x = "Study Hours Groups",
+          y = "GPA") +
+     theme_minimal() +
+     theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
>
```






Scatterplot: GPA vs. Study Hours Per Day
4.0
3.5
3.0
2.5
GPA
5
6
7
8
9
10
Study Hours Per Day


Boxplot: GPA Distribution Across Study Hours Groups
4.0
3.5
3.0
2.5
GPA
Low
Lower-Middle
Upper-Middle
High
Study Hours Groups

4. Report your results.

1. One-Way ANOVA Results: Study Hours and GPA

The analysis aimed to investigate the relationship between Study_Hours_Per_Day and GPA by dividing students into quartiles: Low, Lower-Middle, Upper-Middle, and High study hour groups. A one-way ANOVA revealed significant differences in GPA across the four study hour groups ($F(3996) = 346.4$, $p < 0.001$ or 0.000005), indicating that GPA varies significantly with different levels of study hours. Post-hoc comparisons using Tukey's HSD test showed that all group pairs were significantly different ($p < 0.001$). For instance, students in the High study hour group had significantly higher GPAs than those in the Low study hour group, with a mean GPA difference of 0.5758. The boxplot of GPA distributions illustrates a clear upward trend in GPA as study hours increase, with the High study hour group achieving the highest median GPA.

The ANCOVA results examining the interaction effect of Study_Hours_Per_Day and Physical_Activity_Hours_Per_Day on GPA further demonstrated that Study_Hours_Per_Day was a significant predictor of GPA ($F(1, 994) = 1202.69$, $p < 0.001$). However, neither physical activity ($F(1, 994) = 3.56$, $p = 0.0596$) nor the interaction term ($F(1, 994) = 0.362$, $p = 0.5476$) was significant. This indicates that while study hours significantly impact GPA, physical activity and its interaction with study hours do not have a notable effect. The scatterplot with regression lines visualizing the interaction effect confirmed the lack of clear interaction, with consistent positive trends across different levels of physical activity. These findings suggest that increasing study hours is strongly associated with higher GPA, regardless of physical activity levels.

ANOVA Summary

Table 1: ANOVA Summary for GPA Across Study Hours Groups

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Study_Hours_Group	3	46.81	15.603	346.4	<2e-16***
Residuals	996	44.86	0.045		

Tukey's HSD Post-Hoc Test Results

Table 2: Tukey's HSD Post-Hoc Test Results

Comparison	Difference	Lower	Upper	p-value
Lower-Middle vs Low	0.1935	0.1444	0.2425	0.0000
Upper-Middle vs Low	0.3695	0.3218	0.4171	0.0000
High vs Low	0.5758	0.5279	0.6237	0.0000
Upper-Middle vs Lower-Middle	0.1760	0.1260	0.2260	0.0000

High vs Lower-Middle	0.3823	0.3321	0.4326	0.0000
High vs Upper-Middle	0.2063	0.1574	0.2552	0.0000

2. ANCOVA Results: Interaction Effects of Study Hours and Physical Activity on GPA

The ANCOVA results examining the interaction effect of Study_Hours_Per_Day and Physical_Activity_Hours_Per_Day on GPA demonstrated that Study_Hours_Per_Day was a significant predictor of GPA ($F(1, 994) = 1202.69$, $p < 0.001$). However, neither physical activity ($F(1, 994) = 3.56$, $p = 0.0596$) nor the interaction term ($F(1, 994) = 0.362$, $p = 0.5476$) was significant. This indicates that while study hours significantly impact GPA, physical activity and its interaction with study hours do not have a notable effect. The scatterplot with regression lines visualizing the interaction effect confirmed the lack of clear interaction, with consistent positive trends across different levels of physical activity.

ANCOVA Summary

Table 3: ANCOVA Summary

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Study_Hours_Per_Day	1	50.03	50.03	1202.687	<2e-16***
Physical_Activity_Hours_Per_Day	1	0.15	0.15	3.557	0.0596
Extracurricular_Hours_Per_Day	1	0.12	0.12	2.936	0.0869
Sleep_Hours_Per_Day	1	0.00	0.00	0.064	0.8004
Interaction (Study Hours × Physical Activity)	1	0.02	0.02	0.362	0.5476
Residuals	994	41.35	0.04		

5. State your conclusions and discuss any limitations.

Conclusions: the analyses reveal that stress levels significantly affect Study_Hours_Per_Day, Sleep_Hours_Per_Day, Physical_Activity_Hours_Per_Day, and GPA, but not Extracurricular_Hours_Per_Day or Social_Hours_Per_Day. These findings suggest that stress may play a critical role in influencing specific behaviors and outcomes, particularly those related to academic and personal well-being.

- **Study Hours**: Higher stress levels are linked to more study hours, showing that students under more stress tend to focus more on academics.
- **Extracurricular Activities**: Stress levels do not significantly impact participation in extracurricular activities, indicating they remain consistent regardless of stress.

- **Sleep Hours**: Students with higher stress levels sleep less, highlighting the potential negative impact of stress on rest and recovery.
- **Social Hours**: Stress levels do not significantly affect time spent on social activities, suggesting social engagement stays steady across groups.
- **Physical Activity**: Higher stress is associated with reduced physical activity, pointing to the need for managing stress to maintain physical health.
- **GPA**: Higher stress levels are linked to better academic performance, suggesting stress may drive academic focus but needs balance for well-being.

The analysis highlights that students' GPA is significantly influenced by their study hours, with higher study hours being strongly associated with higher GPA. Among lifestyle factors, study hours showed the strongest positive relationship with GPA, while physical activity and social hours had weaker and sometimes negative associations. When testing for the interaction between study hours and physical activity, the results indicated no significant interaction effect, meaning the relationship between study hours and GPA remains consistent regardless of physical activity levels. Additionally, the results from the ANOVA showed clear and significant differences in GPA across different study hour groups, with students in the high study hour group achieving the highest GPAs.

Limitations: While the statistical tests applied provide valuable insights, there are a few potential limitations to consider. First, the assumption of normality for ANOVA and regression residuals may not fully hold, as minor deviations were observed in residual plots. Second, the data used for analysis is observational, which limits the ability to establish causation—these results show associations, not direct cause-and-effect relationships. Additionally, the variables measured are self-reported, which may introduce bias or inaccuracies. Finally, the interactions tested (e.g., study hours × physical activity) may not fully capture the complexities of real-life factors affecting GPA, as other unmeasured variables might play a significant role. These limitations suggest that the findings should be interpreted cautiously and validated with additional data or experimental studies.

6. Final Conclusion

the analyses reveal that stress levels significantly affect Study_Hours_Per_Day, Sleep_Hours_Per_Day, Physical_Activity_Hours_Per_Day, and GPA, but not Extracurricular_Hours_Per_Day or Social_Hours_Per_Day. The analysis shows that stress levels significantly impact study hours, sleep, physical activity, and GPA, while extracurricular and social activities remain unaffected. High-stress individuals study more, sleep less, and engage in less physical activity but tend to have higher GPAs compared to those with lower stress. These results suggest that stress influences behaviors and academic outcomes, highlighting its dual effects on performance and well-being. These findings suggest that stress may play a critical role in influencing specific behaviors and outcomes, particularly those related to academic and personal well-being. This analysis underscores the critical role of study hours, Social Hours Per Day, Physical Activity Hours Per Day in shaping academic success positively and negatively, with higher study hours strongly associated with improved GPA across all models and group comparisons. While other lifestyle factors, such as physical activity, social hours, and extracurricular activities, show weaker or negative relationships with GPA, their influence is minimal compared to the impact of dedicated study time. Importantly, the absence of significant interaction effects between study hours and other variables suggests that the relationship between study hours and GPA remains consistent, regardless of additional activities. These findings highlight the importance of structured time management, where prioritizing study hours is a key determinant of academic performance, regardless of variations in other lifestyle habits. The ANOVA results show a highly significant effect of stress levels on GPA ($F = 187$, $p < 2e-16$), with GPA increasing as stress levels rise. Tukey's HSD test reveals that individuals with "High" stress have significantly higher GPAs than those with "Moderate" or "Low" stress, and "Moderate" stress leads to higher GPAs than "Low" stress. This highlights a strong positive relationship between stress levels and academic performance. Some other results are as the following,

- Study Hours and GPA:
    - The results show that Study_Hours_Per_Day strongly correlates with higher GPA (0.73) and increased Stress_Level (0.71), while Sleep_Hours_Per_Day (-0.32) and

Physical_Activity_Hours_Per_Day (-0.18) decrease with higher stress. This highlights the balance between academic success and well-being, as stress impacts sleep and activity negatively.

  - Students who spend more time studying tend to have significantly higher GPAs. This relationship was the strongest among all lifestyle factors analyzed.
  - The multiple linear regression model showed that study hours are the most significant predictor of GPA.
  - Students in the high study hour group consistently outperform their peers in GPA, confirming that more study time correlates strongly with academic success. Even moderate increases in study hours can lead to measurable GPA improvements, emphasizing the importance of time management and focused academic efforts. These results reinforce the value of dedicating time to studies for improved academic performance.
  - Study hours showed the strongest positive correlation with GPA, while other lifestyle factors like physical activity, social hours, and extracurricular hours had weaker or negative correlations.
  - The correlations show that higher study hours are strongly linked to higher GPAs and increased stress levels, while stress negatively impacts sleep and physical activity. Sleep and physical activity decrease as stress rises, highlighting the trade-off between academic performance and well-being. Striking a balance between study, stress, and personal health is essential for maintaining both academic success and overall well-being.
  - The one-way ANOVA shows significant differences in GPA among the three social hour groups, with Tukey's HSD indicating that the high social hour group has a lower GPA compared to the low and moderate groups. The boxplot confirms that increased social hours are associated with a slight decline in GPA, particularly at higher levels.

- Other Lifestyle Factors:
  - Physical activity showed a weak negative correlation with GPA, but it was not a significant predictor in the regression model.

    - Social hours and extracurricular hours had minor negative associations with GPA but were less impactful compared to study hours.

- Differences in GPA Across Study Groups:
    - When students were divided into quartiles based on their study hours, ANOVA showed significant differences in GPA across groups. The high study hour group had the highest average GPA, while the low study hour group had the lowest.

- Interaction Effects:
    - Testing the interaction between study hours and physical activity showed no significant interaction effect, meaning that the impact of study hours on GPA is not influenced by physical activity levels.

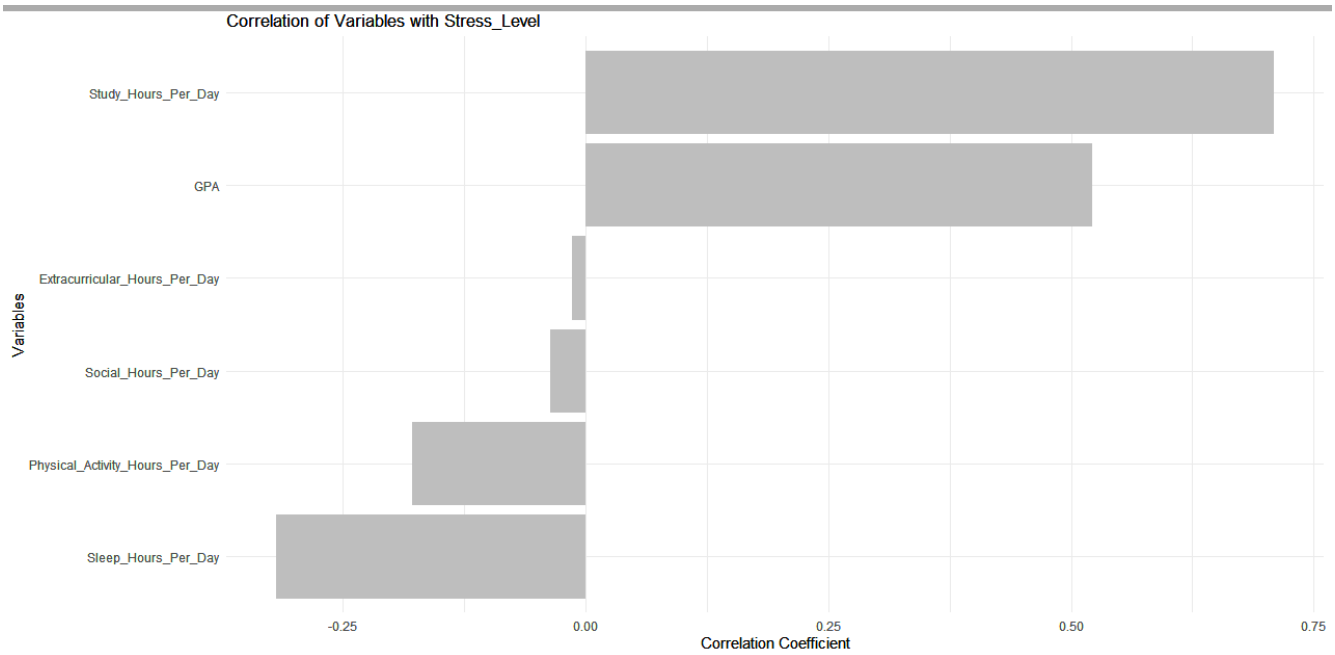
- Overall Model Fit:
    - The multiple linear regression model, including study hours and other lifestyle factors, explained 54.7% of the variation in GPA (R-squared = 0.5488). This indicates that while lifestyle factors play a significant role, other unmeasured factors also contribute to GPA.

- Key Visual Insights:
    - Boxplots revealed that higher study hour groups consistently had higher GPA distributions.
    - A heatmap of correlations confirmed the strong positive relationship between study hours and GPA while showing weaker or negative correlations for other lifestyle variables.

- Limitations:
    - The analysis is observational and does not prove causation. Additionally, some assumptions (e.g., normality in residuals) may not hold perfectly.
    - The data relies on self-reported variables, which might include inaccuracies or biases.

As a summary, the analysis shows that stress levels significantly impact study hours, sleep, physical activity, and GPA, while extracurricular and social activities remain unaffected. High-stress individuals study more, sleep less, and engage in less physical activity but tend to have higher GPAs compared to those with lower stress. These results suggest that stress influences behaviors and

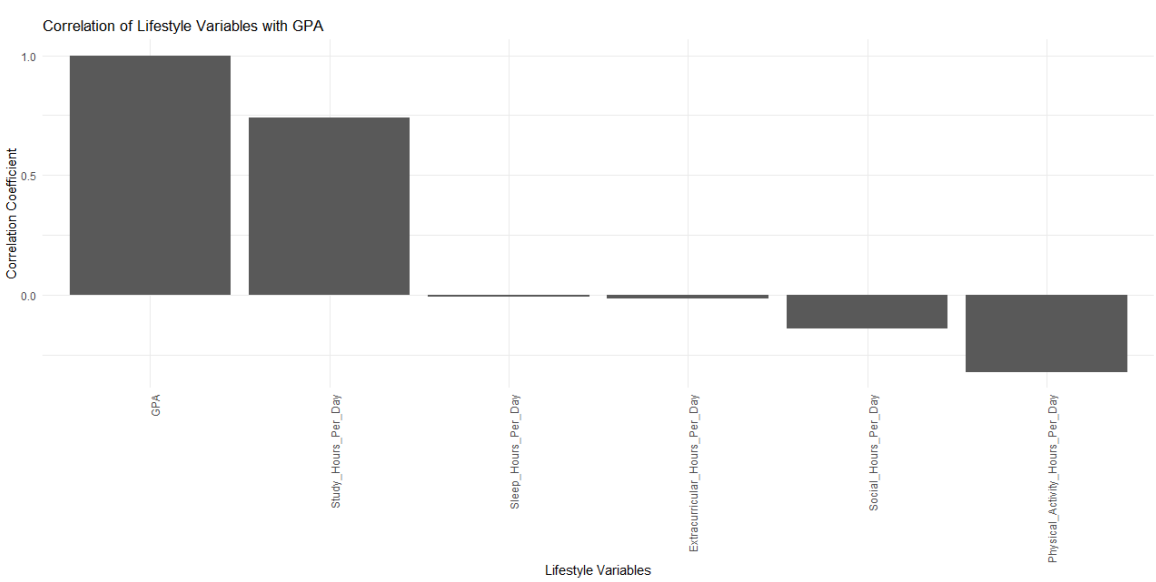
academic outcomes, highlighting its dual effects on performance and well-being. Study hours are the most critical lifestyle factor positively influencing GPA, and this relationship is robust across multiple methods. However, the findings should be interpreted with caution due to the limitations of the data and methodology. In addition to hypothesis theories, I would like to say that the below bar plot has many things and implication to provide an important conclusion,




```
[1] "Correlation Matrix:"
> print(correlation_matrix)
                              Study_Hours_Per_Day
Study_Hours_Per_Day                    1.00000000
Extracurricular_Hours_Per_Day         -0.03619625
Sleep_Hours_Per_Day                    0.03575271
Social_Hours_Per_Day                  -0.11566221
Physical_Activity_Hours_Per_Day       -0.48475552
GPA                                    0.72966677
Stress_Level                           0.70926898
                              Extracurricular_Hours_Per_Day
Study_Hours_Per_Day                          -0.03619625
Extracurricular_Hours_Per_Day                 1.00000000
Sleep_Hours_Per_Day                           0.01405472
Social_Hours_Per_Day                         -0.14971033
Physical_Activity_Hours_Per_Day              -0.33348913
GPA                                          -0.05150292
Stress_Level                                 -0.01349743
                              Sleep_Hours_Per_Day Social_Hours_Per_Day
Study_Hours_Per_Day                  0.035752710          -0.11566221
Extracurricular_Hours_Per_Day        0.014054721          -0.14971033
Sleep_Hours_Per_Day                  1.000000000          -0.18142345
Social_Hours_Per_Day                -0.181423455           1.00000000
Physical_Activity_Hours_Per_Day     -0.488990502          -0.43678677
GPA                                  0.004158696          -0.06080221
Stress_Level                        -0.318754488          -0.03593849
```

```
                                Physical_Activity_Hours_Per_Day          GPA
Study_Hours_Per_Day                                   -0.4847555  0.729666769
Extracurricular_Hours_Per_Day                         -0.3334891 -0.051502917
Sleep_Hours_Per_Day                                   -0.4889905  0.004158696
Social_Hours_Per_Day                                  -0.4367868 -0.060802208
Physical_Activity_Hours_Per_Day                        1.0000000 -0.345586822
GPA                                                   -0.3455868  1.000000000
Stress_Level                                          -0.1789999  0.522054925
                                Stress_Level
Study_Hours_Per_Day               0.70926898
Extracurricular_Hours_Per_Day    -0.01349743
Sleep_Hours_Per_Day              -0.31875449
Social_Hours_Per_Day             -0.03593849
Physical_Activity_Hours_Per_Day  -0.17899991
GPA                               0.52205492
Stress_Level                      1.00000000
```

The correlation matrix reveals that `Study_Hours_Per_Day` has a strong positive correlation with both `GPA` (0.73) and `Stress_Level` (0.71), suggesting that more study hours are associated with better academic performance and higher stress. Conversely, `Sleep_Hours_Per_Day` and `Physical_Activity_Hours_Per_Day` have moderate negative correlations with `Stress_Level` (-0.32 and -0.18, respectively), indicating that higher stress is linked to reduced sleep and physical activity. Other variables, like `Extracurricular_Hours_Per_Day` and `Social_Hours_Per_Day`, show negligible correlations with stress and other major variables, suggesting they are relatively independent in this context. These findings emphasize the trade-off between study-related gains and potential impacts on health and well-being. The results show that Study_Hours_Per_Day strongly correlates with higher GPA (0.73) and increased Stress_Level (0.71), while Sleep_Hours_Per_Day (-0.32) and Physical_Activity_Hours_Per_Day (-0.18) decrease with higher stress. This highlights the balance between academic success and well-being, as stress impacts sleep and activity negatively.



Solution Submission

1. **Upload a write-up document.**
2. **Upload your data set. This is the data set after cleaning (a small CSV file).**
3. **Upload your R script.**

Grading will be based on:

1. **Development and description of a research scenario and question(s).**
2. **Data preparation and data cleaning (when relevant).**
3. **Correctness of statistical analysis methods chosen to answer the research question.**
4. **Correctness of R code.**
5. **Quality of figures, tables, and write-up.**
6. **Correct and thoughtful discussion of conclusions and limitations.**