The Impact of Students' behavior on Exam Performance

STAT 31631 – Statistical Modeling

Department of Statistics & Computer Science

University of Kelaniya

Academic Year 2023/2024

By

Group Number **01**

Sr. No	Student Number	Student Name
01.	PS/2021/225	G.K.S. Munasinghe
02.	PS/2021/032	M.T.S.S. Perera
03.	PS/2021/034	K.D.H.U.N. Sepalika
04.	PS/2021/091	P.A.A. Nishani
05.	PS/2021/122	W.P.C. Navoda
06.	PS/2021/137	U.G. Chiranji
07.	PS/2021/217	K.Thamilpriya
08.	PS/2021/178	B.M.T.M. Jayasinghe
09.	PS/2021/190	P.G.A. Madhushani
10.	PS/2021/194	K.G.H.S. Dissanayake
11.	PS/2021/033	W.A.S.B.W. Arachchi

Table of Contents

Table of Figures

Abstract

This study investigates the relationship between student lifestyle habits and academic performance using a dataset of 1,000 synthetic student records. The research employed comprehensive statistical analysis including exploratory data analysis, multiple linear regression, best subset selection, and stepwise model selection to identify key predictors of exam scores.

The dataset containing 15 variables was analyzed after data preprocessing and outlier assessment. Multiple regression techniques were applied to examine relationships between lifestyle factors (study hours, screen time, sleep, exercise, mental health) and academic outcomes. Model selection procedures including forward selection and stepwise elimination were used to identify the optimal predictive model.

The final model explained approximately 90% of variance in exam scores ($R^2 = 0.9011$) with seven significant predictors. Study hours per day emerged as the strongest positive predictor ($\beta = 9.57$), while social media hours ($\beta = -2.62$) and Netflix hours ($\beta = -2.28$) showed negative associations. Other significant factors included attendance percentage ($\beta = 0.14$), sleep hours ($\beta = 2.00$), exercise frequency ($\beta = 1.45$), and mental health rating ($\beta = 1.95$). All predictors demonstrated statistical significance ($p < 0.001$).

Academic performance is significantly influenced by lifestyle choices, with study habits having the greatest impact. Students who prioritize studying, maintain healthy sleep patterns, exercise regularly, and limit recreational screen time achieve higher academic outcomes. These findings provide evidence-based recommendations for optimizing student lifestyle habits to enhance academic success.

Introduction

Background of the Study

Students today balance academics with various activities such as social media use, streaming content, part-time jobs, and extracurricular involvement. These lifestyle choices ranging from sleep habits to study routines can significantly influence academic performance. For example, excessive screen time or irregular sleep may negatively affect exam results, while disciplined study and healthy routines can enhance them.

Academic achievement is essential for career opportunities and personal growth, yet students face challenges like stress, poor time management, and distractions. Understanding how daily behaviors affect performance can help them make informed choices.

While previous studies often focus on individual factors, few examine their combined effect. This study analyzes data from 1,000 students to explore how habits like study time, sleep, social media use, attendance, diet, and mental health collectively relate to academic success them.

Problem Statement

While individual behavioral factors such as study habits, sleep, and digital consumption have been widely studied in relation to academic performance, there is a lack of comprehensive research exploring how these variables interact and influence outcomes collectively. Most existing studies analyze these behaviors in isolation, failing to account for the complex trade-offs and overlaps students face in real life.

For instance, a student who dedicates significant time to studying may still perform poorly due to insufficient sleep or long work hours. Similarly, high social media usage may not impact grades if offset by strong attendance or healthy lifestyle choices. This fragmented approach limits our understanding of how multiple daily behaviors combine to affect academic success. Therefore, this study addresses the gap by examining the combined influence of seven key behavioral factors, providing a more holistic perspective and generating evidence-based insights to guide students, educators, and policymakers.

Objectives

The primary aim of this research is to evaluate the impact of various daily behavioral habits on students' academic performance. Specifically, the study seeks to:

- Assess the time allocation across key activities such as studying, sleeping, and digital media usage.
- Identify which behaviors are positively or negatively associated with exam outcomes.
- Uncover significant patterns between lifestyle choices and academic success through statistical analysis.
- Provide evidence-based recommendations to help students optimize their habits for improved academic results.

Significance of the Study

This study benefits students, educators, and parents by identifying habits that enhance academic performance. Educators can leverage these insights to improve student support through targeted planning, mentoring, and wellness initiatives. Grounded in empirical data rather than anecdotal evidence, the research offers clear, actionable recommendations. By elucidating the link between behavior and academic outcomes, it empowers students to optimize their routines. Furthermore, in an era of pervasive distractions and stress, the study underscores the critical need for a balanced approach to academics, rest, social engagement, and health to achieve sustained success

Gantt chart

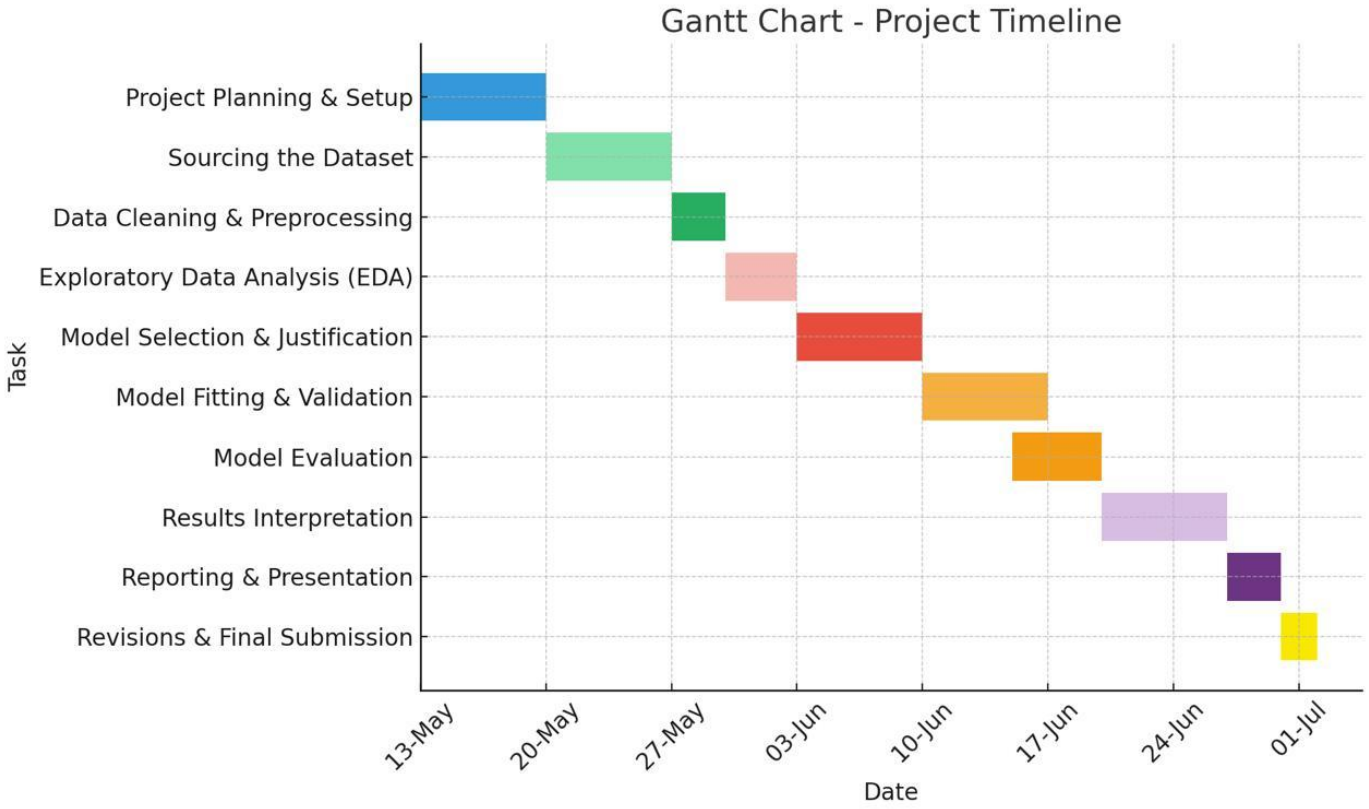


Figure 1: Gantt Chart

Methodology

Data Collection & Preparation

The dataset is publicly available at https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance, contains records of 1,000 students. Each record includes information on student demographics, lifestyle habits, and final exam scores. This data was used to explore how daily student behaviors relate to academic performance. The dataset was provided for the STAT 31631 coursework, and each entry represents an individual student.

This dataset included the following variables:

- Student_ID: Unique identifier for each student.
- Age: Age of the student.
- Gender: Gender of the student.
- Study_Hours_Per_Day: Time spent studying daily.
- Social_Media_Hours: Daily time spent on social media.
- Netflix_Hours: Daily time spent watching Netflix.
- Part_Time_Job: Whether the student has a part-time job (Yes/No).
- Attendance_Percentage: Class attendance rate.
- Sleep_Hours: Average daily hours of sleep.
- Diet_Quality: Quality of diet (Poor, Fair, Good).
- Exercise_Frequency: Frequency of weekly exercise.
- Parental_Education_Level: Parent's highest education level.
- Internet_Quality: Internet connection quality.
- Mental_Health_Rating: Mental well-being score (1-10).
- Extracurricular_Participation: Participation in extracurriculars (Yes/No).
- Exam_Score: Final exam score (response variable).

Methodological Steps

1. Data Acquisition and Preparation

The dataset, student_habits_performance.csv, was imported into R for analysis. Initial data exploration included checking for missing values, examining variable types, and removing irrelevant columns (student ID).

All categorical variables (gender, part-time job, diet quality, parental education, internet quality, and extracurricular participation) were converted to factors, while numerical variables (age, exercise frequency, and mental health rating) were converted to numeric types to ensure appropriate handling in subsequent analyses.

Importing the data set into R.

```
data<-read.csv("student_habits_performance.csv")
head(data)
```

Description: df [6 × 16]

	student_id <chr>	age <int>	gender <chr>	study_hours_per_day <dbl>	social_media_hours <dbl>	netflix_hours <dbl>	part_time_job <chr>	attendance_percentage <dbl>
1	S1000	23	Female	0.0	1.2	1.1	No	85.0
2	S1001	20	Female	6.9	2.8	2.3	No	97.3
3	S1002	21	Male	1.4	3.1	1.3	No	94.8
4	S1003	23	Female	1.0	3.9	1.0	No	71.0
5	S1004	19	Female	5.0	4.4	0.5	No	90.9
6	S1005	24	Male	7.2	1.3	0.0	No	82.9

6 rows | 1-9 of 16 columns

Figure 2: Importing the data into R

Checking for missing values

```
```{r}
missing_value<-sum(is.na(data))
missing_value
```
```

```
[1] 0
```

Figure 3: Checked for missing values

Checking the data type

```
str(data)
```

```
'data.frame':   1000 obs. of  16 variables:
 $ student_id                  : chr  "S1000" "S1001" "S1002" "S1003" ...
 $ age                         : int  23 20 21 23 19 24 21 21 23 18 ...
 $ gender                      : chr  "Female" "Female" "Male" "Female" ...
 $ study_hours_per_day         : num  0 6.9 1.4 1 5 7.2 5.6 4.3 4.4 4.8 ...
 $ social_media_hours          : num  1.2 2.8 3.1 3.9 4.4 1.3 1.5 1 2.2 3.1 ...
 $ netflix_hours               : num  1.1 2.3 1.3 1 0.5 0 1.4 2 1.7 1.3 ...
 $ part_time_job               : chr  "No" "No" "No" "No" ...
 $ attendance_percentage       : num  85 97.3 94.8 71 90.9 82.9 85.8 77.7 100 95.4 ...
 $ sleep_hours                 : num  8 4.6 8 9.2 4.9 7.4 6.5 4.6 7.1 7.5 ...
 $ diet_quality                : chr  "Fair" "Good" "Poor" "Poor" ...
 $ exercise_frequency          : int  6 6 1 4 3 1 2 0 3 5 ...
 $ parental_education_level    : chr  "Master" "High School" "High School" "Master" ...
 $ internet_quality            : chr  "Average" "Average" "Poor" "Good" ...
 $ mental_health_rating        : int  8 8 1 1 1 4 4 8 1 10 ...
 $ extracurricular_participation: chr  "Yes" "No" "No" "Yes" ...
 $ exam_score                  : num  56.2 100 34.3 26.8 66.4 100 89.8 72.6 78.9 100 ...
```

Figure 4: Checking the data type

Assigned data type of the categorical data types to factor type

```
```{r}
data$gender <- as.factor(data$gender)
data$part_time_job <- as.factor(data$part_time_job)
data$diet_quality <- as.factor(data$diet_quality)
data$parental_education_level <- as.factor(data$parental_education_level)
data$internet_quality <- as.factor(data$internet_quality)
data$extracurricular_participation<- as.factor(data$extracurricular_participation)
data$age <- as.numeric(data$age)
data$exercise_frequency <- as.numeric(data$exercise_frequency)
data$mental_health_rating <- as.numeric(data$mental_health_rating)
```
```

Figure 5:Assigned data type of the categorical data types to factor type

Remove the ID variable from the dataset

```
```{r}
data$student_id<- NULL
head(data)
```
```

Description: df [6 × 15]

	age <dbl>	gender <fctr>	study_hours_per_day <dbl>	social_media_hours <dbl>	netflix_hours <dbl>	part_time_job <fctr>	attendance_percentage <dbl>	sleep_hours <dbl>
1	23	Female	0.0	1.2	1.1	No	85.0	8.0
2	20	Female	6.9	2.8	2.3	No	97.3	4.6
3	21	Male	1.4	3.1	1.3	No	94.8	8.0
4	23	Female	1.0	3.9	1.0	No	71.0	9.2
5	19	Female	5.0	4.4	0.5	No	90.9	4.9
6	24	Male	7.2	1.3	0.0	No	82.9	7.4

6 rows | 1-9 of 15 columns

```
```{r}
```

Figure 6:Remove the ID variable from the dataset
```

2. Exploratory Data Analysis (EDA)

Descriptive statistics and summaries were generated to understand the distribution and central tendencies of each variable. Separating categorical (grouped) and numerical (measured) variables ensures you get correct summaries, clear charts, valid tests, and reliable models that accurately reflect your data.

Boxplots and scatterplots were utilized to visualize relationships between exam scores and various predictors, including both categorical and numerical variables. This helped identify patterns, outliers, and potential associations. Outlier analysis was performed using boxplots for all numeric variables, but outliers were retained as they were not deemed to be data entry errors or unrealistic values.

Summary of the data set

```{r}
summary(data)
```

```
  student_id       age           gender    study_hours_per_day social_media_hours netflix_hours   part_time_job attendance_percentage
 S1000  :  1   Min.   :17.00   Female:481   Min.   :0.00        Min.   :0.000      Min.   :0.000   No :785       Min.   : 56.00
 S1001  :  1   1st Qu.:18.75   Male  :477   1st Qu.:2.60        1st Qu.:1.700      1st Qu.:1.000   Yes:215       1st Qu.: 78.00
 S1002  :  1   Median :20.00   Other : 42   Median :3.50        Median :2.500      Median :1.800                 Median : 84.40
 S1003  :  1   Mean   :20.50                Mean   :3.55        Mean   :2.506      Mean   :1.820                 Mean   : 84.13
 S1004  :  1   3rd Qu.:23.00                3rd Qu.:4.50        3rd Qu.:3.300      3rd Qu.:2.525                 3rd Qu.: 91.03
 S1005  :  1   Max.   :24.00                Max.   :8.30        Max.   :7.200      Max.   :5.400                 Max.   :100.00
 (Other):994
  sleep_hours    diet_quality exercise_frequency parental_education_level internet_quality mental_health_rating
 Min.   : 3.20   Fair:437     Min.   :0.000      Bachelor   :350           Average:391      Min.   : 1.000
 1st Qu.: 5.60   Good:378     1st Qu.:1.000      High School:392           Good   :447      1st Qu.: 3.000
 Median : 6.50   Poor:185     Median :3.000      Master     :167           Poor   :162      Median : 5.000
 Mean   : 6.47                Mean   :3.042      None       : 91                            Mean   : 5.438
 3rd Qu.: 7.30                3rd Qu.:5.000                                                 3rd Qu.: 8.000
 Max.   :10.00                Max.   :6.000                                                 Max.   :10.000

 extracurricular_participation   exam_score
 No :682                       Min.   : 18.40
 Yes:318                       1st Qu.: 58.48
                               Median : 70.50
                               Mean   : 69.60
                               3rd Qu.: 81.33
                               Max.   :100.00
```

Figure 7: Summary of the data set

```
#describe the data
library(psych)
describe(data)
boxplot(data)
```

data.frame
15 x 13

	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
age	1	1000	20.50	2.31	20.0	20.50	2.97	17.0	24.0
gender*	2	1000	1.56	0.57	2.0	1.52	1.48	1.0	3.0
study_hours_per_day	3	1000	3.55	1.47	3.5	3.55	1.48	0.0	8.3
social_media_hours	4	1000	2.51	1.17	2.5	2.50	1.19	0.0	7.2
netflix_hours	5	1000	1.82	1.08	1.8	1.79	1.19	0.0	5.4
part_time_job*	6	1000	1.22	0.41	1.0	1.14	0.00	1.0	2.0
attendance_percentage	7	1000	84.13	9.40	84.4	84.37	9.64	56.0	100.0
sleep_hours	8	1000	6.47	1.23	6.5	6.46	1.33	3.2	10.0
diet_quality*	9	1000	1.75	0.75	2.0	1.69	1.48	1.0	3.0
exercise_frequency	10	1000	3.04	2.03	3.0	3.05	2.97	0.0	6.0

1-10 of 15 rows | 1-10 of 13 columns Previous 1 2 Next

Figure 8: Describe the dataset

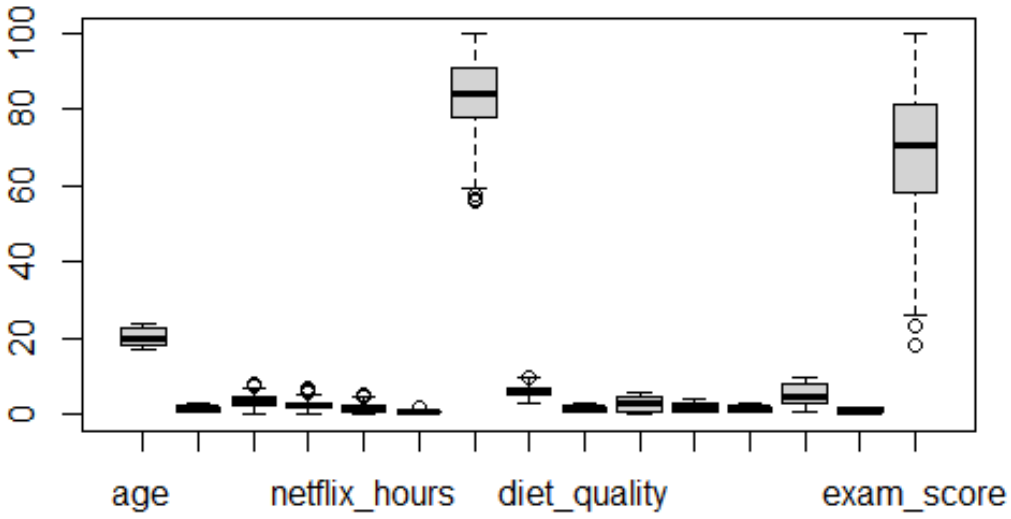


Figure 9 : Boxplot for the all variables

Check the correlation between numeric variables

```r
{r}
library(dplyr)
# Check correlation between numeric variables
numeric_vars <- data %>% select_if(is.numeric)
cor_matrix <- cor(numeric_vars)
print(cor_matrix)

library(ggplot2)
library(GGally)
ggcorr(numeric_vars, label = TRUE, label_size = 3, size = 2.5)
```

R Console

```
                      age study_hours_per_day social_media_hours
netflix_hours attendance_percentage
age                1.000000000         0.003971179       -0.009151199
-0.0011741040      -0.026055201
study_hours_per_day 0.003971179        1.000000000        0.020282314
-0.0311583466       0.026264118
social_media_hours -0.009151199        0.020282314        1.000000000
```

Figure 10 : Check the correlation between numeric variables

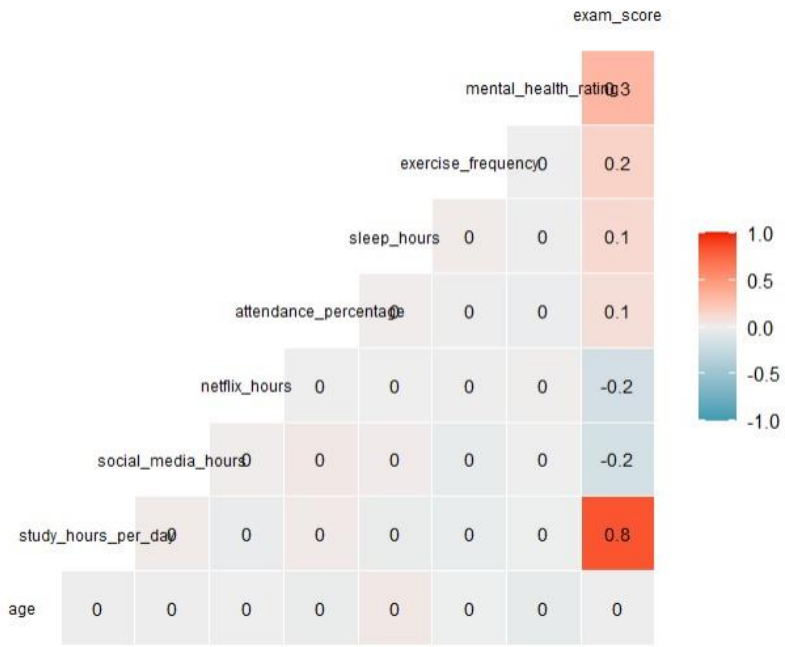


Figure 11 : Visualization of correlation

Separating numerical (continuous/discrete) data

```
#Separating numerical (continuous/discrete) data
v <-c(15,1,3,4,5,7,8,10,13)
pairs(data[,v])
```

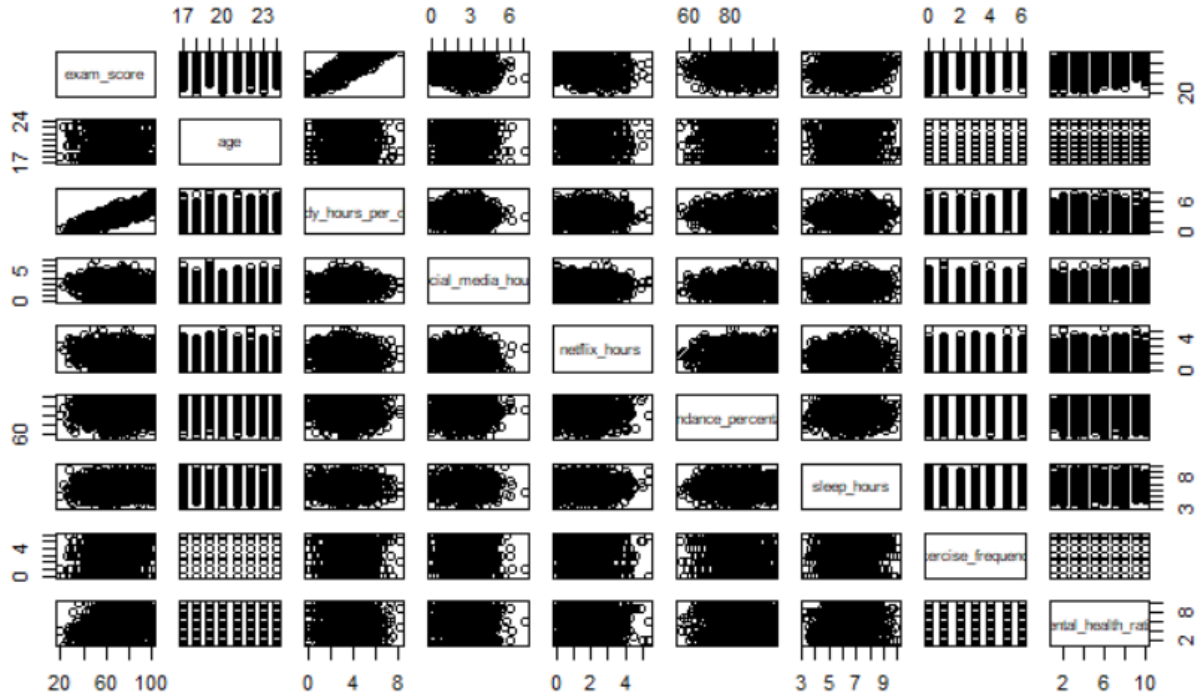


Figure 10 :Separating numerical data

Collinearity of quantitative variables

```
pairs(data[,-v])
```

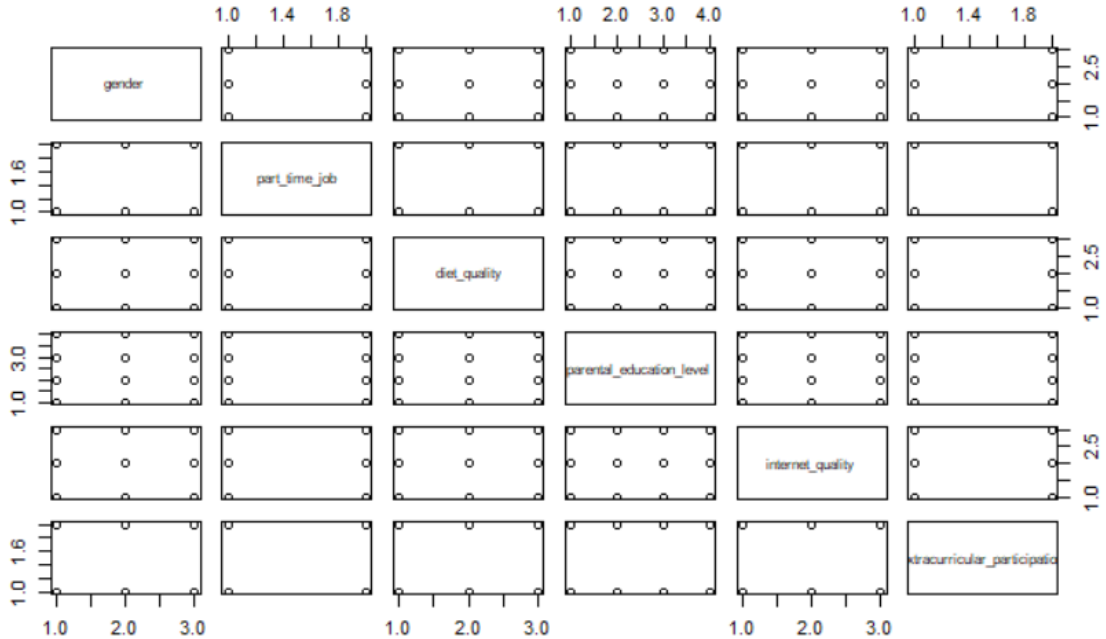


Figure 11 : Collinearity of quantitative variables

Outliers were detected using boxplots (Qualitative variables)

```r
# Replace 'data' with your actual data frame name
numeric_cols <- names(data)[sapply(data, is.numeric)]

# Loop through each numeric variable and create a boxplot with outlier info
for (col in numeric_cols) {
  outliers <- boxplot.stats(data[[col]])$out
  boxplot(data[[col]],
          main = paste("Box plot for", col),
          ylab = col,
          sub = paste("Outliers:", paste(outliers, collapse = ", "))
  )
}
```

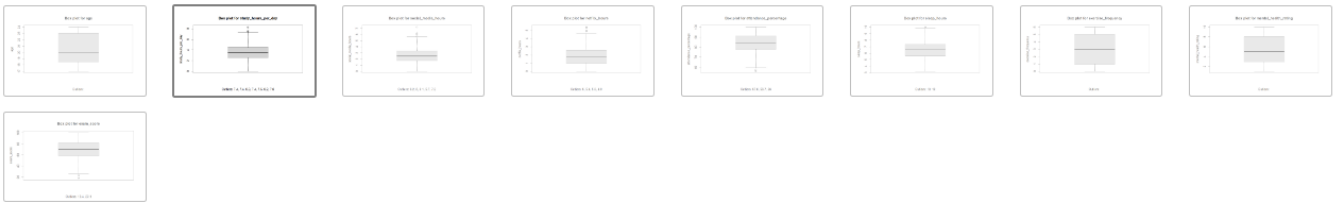


Figure 12:Summery of the data set

We saw some unusually high and low values in our data but didn't remove them. These extremes likely represent real differences in student behavior, not errors. Removing them would throw away valid information, so all data points including these outliers were kept for analysis.

Age **:** Participants are 17 to 24 years old, with an average age of 20.5. This is typical of university students.

Gender**:** There are 481 females, 477 males, and 42 others, showing a balanced gender distribution.

Part time job: 215 students have a part-time job, while 785 do not. Most students are not working alongside their studies.

Study hours per day: Students study between 0 to 8.3 hours per day, with an average of 3.2 hours. There's a wide variation in study habits.

Social media hours**:** Students spend on average 2.8 hours per day on social media, with some spending up to 7.2 hours.

Netflix hours: Average Netflix watching time is 1.8 hours/day, with some students watching up to 5.4 hours/day.

Attendance percentage: Attendance ranges from 56% to 100%, with an average of 84%, suggesting good class participation overall.

Exam score: Exam scores range from 18.4 to 100, with an average of 70.5. Most students are performing fairly well academically.

Parental education level: Most students have parents with high school (392) or bachelor's (350) education. Some parents have master's (167) or no formal education (91).

Sleep hours**:** Students sleep between 3 to 10 hours, with an average of about 6.8 hours slightly below the recommended amount for adults.

Diet quality**:** Most students have a fair or good diet, but a significant number report poor diet quality.
Good: 378 students , Fair: 437 students , Poor: 185 students

Exercise frequency: Students exercise 0 to 6 days per week, averaging 3 days, indicating moderate physical activity.

Mental health rating: Rated from 1 to 10, the median is 5, meaning half of the students rate their mental health as average or lower.

Plots of quantitative variables

Investigate the difference between in Exam score and Gender

```
boxplot(data$exam_score ~ data$gender)
```

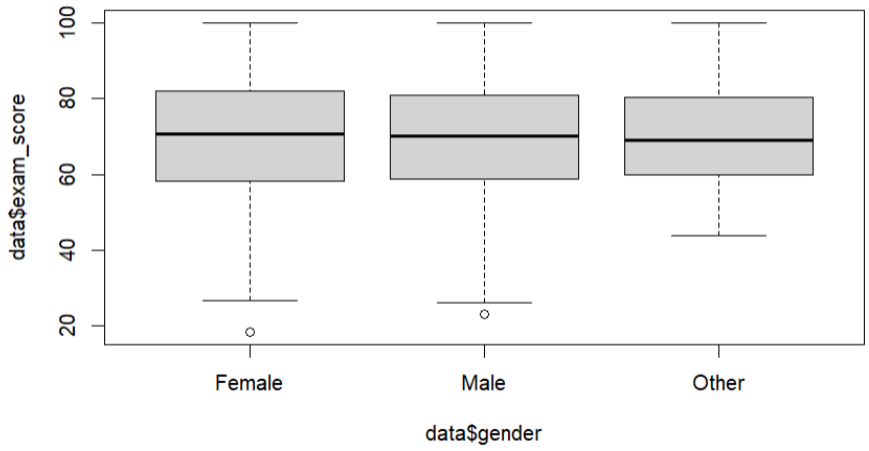


Figure 13:Investigate the difference in exam score and gender

This box plot represents a comparison of exam scores across three different gender categories: Female, Male, and Other. The median exam score for females is around 75, with most scores ranging between 60 and 85. Males show a similar distribution, with a median also near 75 and a slightly wider spread of scores. The “Other” gender group also has a median close to 75, but with a more compact score range and no extreme low outliers. Overall, all three groups show similar performance, suggesting that gender does not have a strong influence on exam scores.

Investigate the difference between Exam score and Part time job

```
boxplot(data$exam_score ~ data$part_time_job)
```

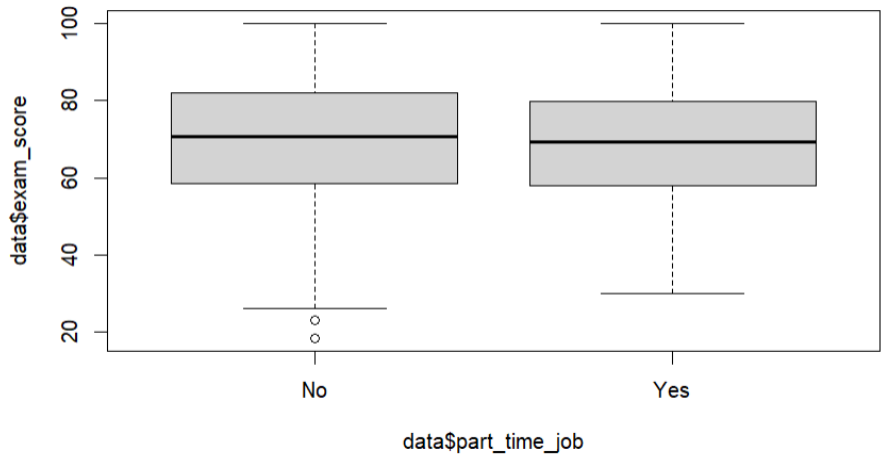


Figure 14:Investigate the difference between in Exam score and Part time job

boxplots show how to affect doing part time job for exam score. The range of doing part time jobs is between approximately 30 and 100. median of not doing part time job is approximately 70

Investigate the difference between Exam score and Diet Quality

```
```{r}
boxplot(data$exam_score ~ data$diet_quality)
```
```

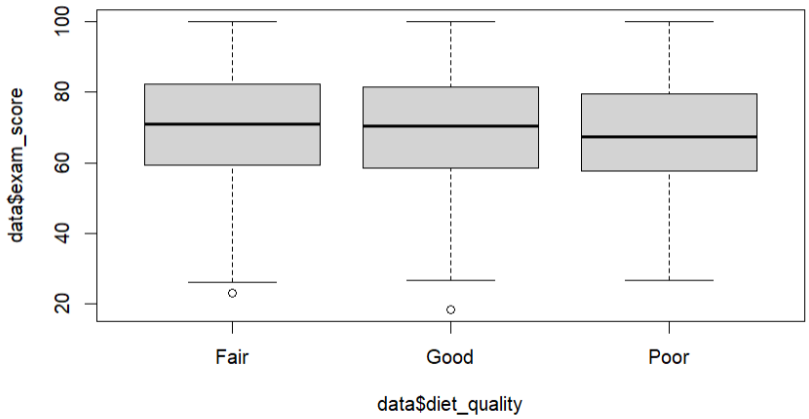


Figure 15 : Investigate the difference between Exam score and Diet Quality

boxplots compare exam scores across three different diet quality categories: Fair, Good, and Poor. The y-axis represents the exam scores ranging from 20 to 100, while the x-axis represents the diet quality categories. It indicates that students with a good diet quality tend to have higher median exam scores compared to those with Fair or Poor diet quality.

Investigate the difference between Exam score and Parenteral Education Level

```
boxplot(data$exam_score ~ data$parental_education_leve)
```

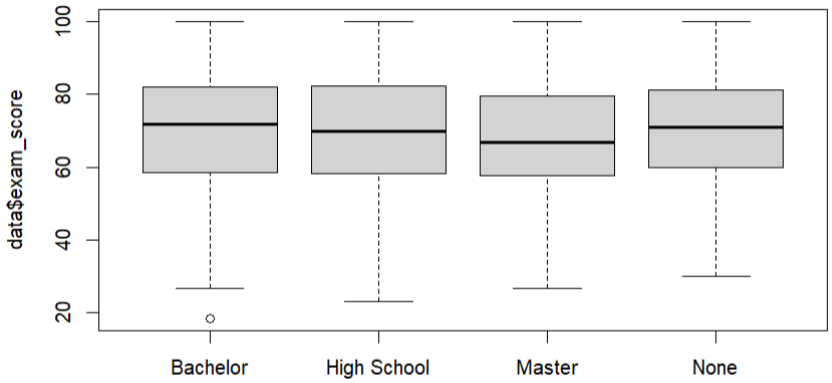


Figure 16 : Investigate the difference between Exam score and Parenteral Education Level

This boxplot displays the distribution of exam scores grouped by parental education level (Bachelor, Master, None). All three groups (Bachelor, Master, None) have similar median scores (around 70).No clear upward or downward trend in medians based on education level. The height of each box (IQR) is quite similar across groups, suggesting comparable score variability. Parental education level does not show a significant visual impact on students' exam performance in this dataset. This suggests that students' scores are relatively independent of whether their parents have Bachelor, Master, or no formal education.

Investigate the difference between Exam score and Internet Quality

```
```{r}
boxplot(data$exam_score ~ data$internet_quality)
```
```

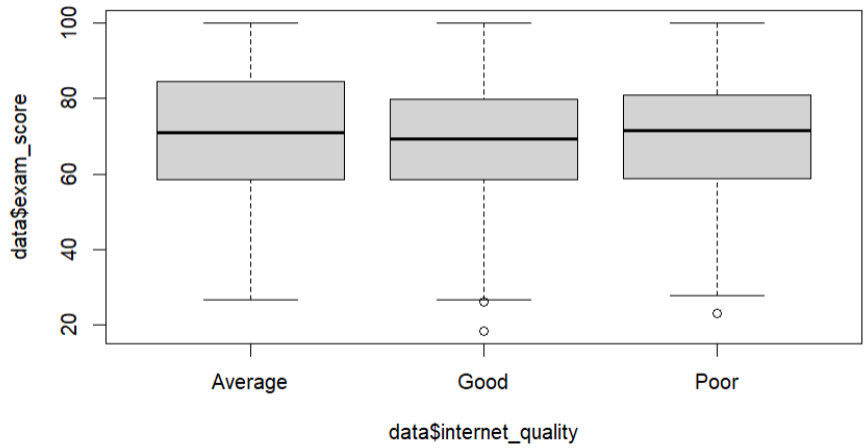


Figure 17 : Investigate the difference between Exam score and Internet Quality

This boxplot shows the distribution of students' exam scores grouped by their reported internet quality: Average, Good, and Poor. All three groups (Average, Good, Poor) have similar medians, around 70.Slightly higher median for students with Average internet. The spread (height of the box) is nearly the same for all groups, suggesting similar variability in scores across internet types. All groups span a wide range (~25 to 100), indicating that both low and high performers exist in all internet quality categories. Internet quality does not appear to have a strong effect on exam performance in this dataset. The presence of high and low scores across all internet types suggests students adapt to their circumstances. While better internet might help with studying, it's not a decisive factor in predicting exam results here.

3. Feature Engineering and Variable Selection

Fit the full model and the dataset was examined for multicollinearity using the Variance Inflation Factor (VIF). All VIF values were below 5, indicating no significant multicollinearity among predictors. Both **best subset selection** and **stepwise model selection** techniques were employed to identify the most significant predictors of exam scores:

Best subset selection was performed using the regsubsets function from the leaps package, evaluating models based on criteria such as Adjusted R^2, Mallows' Cp, and BIC1. Also using the regsubset function for the Forward selection method. Stepwise selection (both forward and backward) was conducted using the step AIC function from the MASS package, optimizing model fit based on AIC1.

The optimal model, as determined by these methods, included the following predictors: study hours per day, social media hours, Netflix hours, attendance percentage, sleep hours, exercise frequency, and mental health rating.

Fit the full model

```
# Fit the model
model<-lm(exam_score~.,data = data)
summary(model)
```

```
Call:
lm(formula = exam_score ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.2097  -3.4768   0.0789   3.4456  15.5955

Coefficients:
                            Estimate Std. Error t value
(Intercept)                 7.17718    2.50263   2.868
age                        -0.01209    0.07365  -0.164
genderMale                  0.14621    0.34805   0.420
genderOther                 0.79264    0.86559   0.916
study_hours_per_day         9.57454    0.11577  82.700
social_media_hours         -2.60220    0.14489 -17.960
netflix_hours              -2.28156    0.15777 -14.462
part_time_jobYes            0.21121    0.41410   0.510
attendance_percentage       0.14339    0.01821   7.876
sleep_hours                 1.99234    0.13849  14.386
```

Figure 18 :Fit the full model

Check the multicollinearity

```
library(car)
vif(model)
```

```
                              GVIF Df GVIF^(1/(2*Df))
age                       1.011421  1        1.005694
gender                    1.031205  2        1.007712
study_hours_per_day       1.012357  1        1.006159
social_media_hours        1.010141  1        1.005058
netflix_hours             1.007106  1        1.003547
part_time_job             1.014125  1        1.007038
attendance_percentage     1.025047  1        1.012446
sleep_hours               1.009766  1        1.004871
diet_quality              1.031849  2        1.007869
exercise_frequency        1.012632  1        1.006296
parental_education_level  1.049969  3        1.008160
internet_quality          1.033362  2        1.008238
```

Figure 19 : Check the multicollinearity

Best subset selection

```
library(leaps)
model_1 <- regsubsets(exam_score~ ., data = data, nvmax = 19)
summary(model_1)
```

```
Subset selection object
Call: regsubsets.formula(exam_score ~ ., data = data, nvmax = 19)
19 Variables  (and intercept)
                                  Forced in Forced out
age                                   FALSE      FALSE
genderMale                            FALSE      FALSE
genderOther                           FALSE      FALSE
study_hours_per_day                   FALSE      FALSE
social_media_hours                    FALSE      FALSE
netflix_hours                         FALSE      FALSE
part_time_jobYes                      FALSE      FALSE
attendance_percentage                 FALSE      FALSE
sleep_hours                           FALSE      FALSE
diet_qualityGood                      FALSE      FALSE
diet_qualityPoor                      FALSE      FALSE
exercise_frequency                    FALSE      FALSE
parental_education_levelHigh School   FALSE      FALSE
parental_education_levelMaster        FALSE      FALSE
parental_education_levelNone          FALSE      FALSE
internet_qualityGood                  FALSE      FALSE
internet_qualityPoor                  FALSE      FALSE
mental_health_rating                  FALSE      FALSE
extracurricular_participationYes      FALSE      FALSE
1 subsets of each size up to 19
```

Figure 20 : model_1 in best subset selection

```
par(mfrow = c(1,2))
plot(summary(model_1)$rss, xlab = "Number of Variables", ylab = "RSS", type =
"l")
RSS_min <- which.min(summary(model_1)$rss)
points(RSS_min,summary(model_1)$rss[RSS_min], col = "blue", cex = 2, pch = 20)
abline(v =RSS_min)
plot(summary(model_1)$adjr2, xlab = "Number of Variables", ylab = "Adjusted
RSq", type = "l")
adjr2_max <- which.max(summary(model_1)$adjr2)
points(adjr2_max, summary(model_1)$adjr2[adjr2_max], col = "red", cex = 2, pch
= 20)
abline(v=adjr2_max)
```

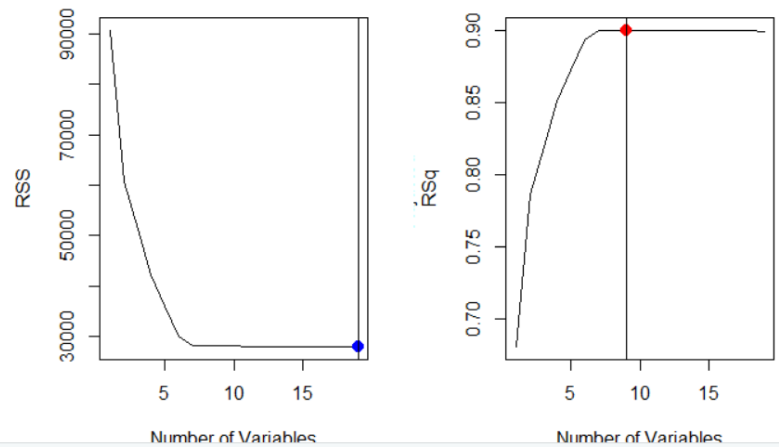


Figure 21 : RSS and adj R2 values

```
par(mfrow = c(1,2))
plot(summary(model_1)$cp, xlab = "Number of Variables", ylab = "Cp", type =
"l")
cp_min <- which.min(summary(model_1)$cp)
points(cp_min, summary(model_1)$cp[cp_min], col = "green", cex = 2, pch = 20)
plot(summary(model_1)$bic, xlab = "Number of Variables", ylab = "BIC", type =
"l")
bic_min <- which.min(summary(model_1)$bic)
points(bic_min, summary(model_1)$bic[bic_min], col = "purple", cex = 2, pch = 20)
```

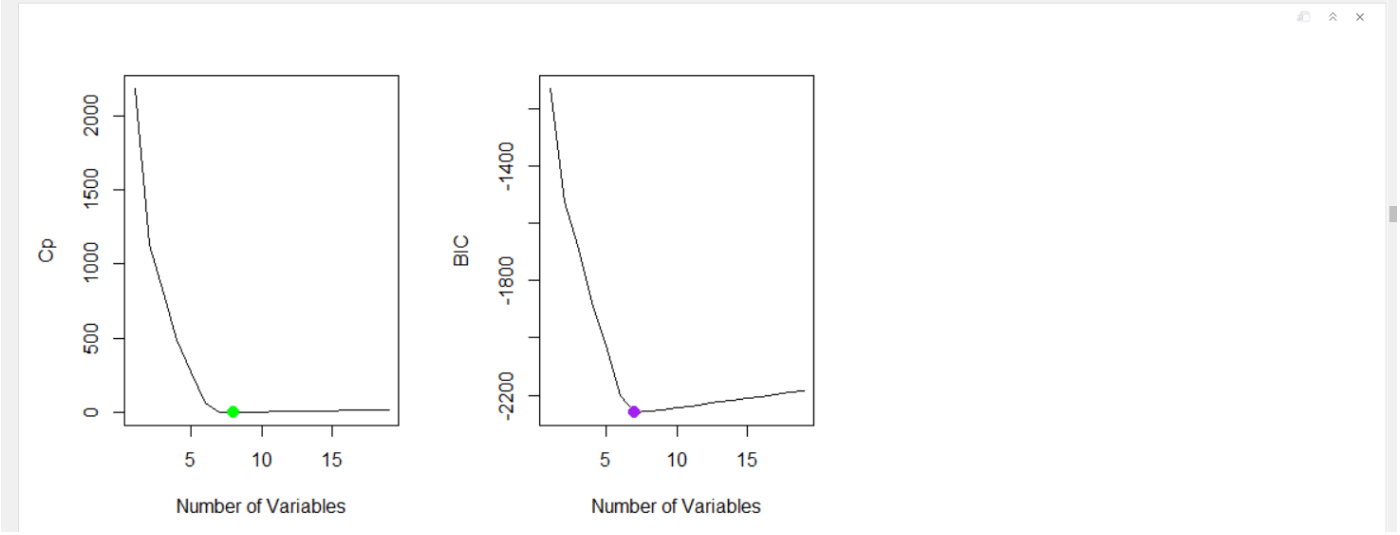


Figure 22 : Cp and BIC values

Create a data frame including all the criterion values for all the models

```
#Create a data frame including all the criterion values for all the models
res.sum <- summary(model_1)
criterion <- data.frame(
  model = 1:19,
  Adj.R2 = (res.sum$adjr2),
  CP = (res.sum$cp),
  BIC = (res.sum$bic),
  RSS = res.sum$rss
)
criterion
```

model <int>	Adj.R2 <dbl>	CP <dbl>	BIC <dbl>	RSS <dbl>
1	0.6809964	2185.826448	-1129.739	90805.38
2	0.7862759	1135.604781	-1524.350	60776.28
3	0.8200278	799.493227	-1690.329	51126.99
4	0.8514238	487.483345	-1876.131	42165.54
5	0.8729408	274.242184	-2026.674	36022.83
6	0.8939647	66.325520	-2201.653	30032.06
7	0.9003729	3.733341	-2258.090	28188.68
8	0.9005771	2.714923	-2254.243	28102.54
9	0.9006441	3.058302	-2249.019	28055.26
10	0.9006326	4.178595	-2243.007	28030.16

1-10 of 19 rows Previous 1 2 Next

Figure 23 : Create a data frame including all the criterion values for all the models

Interpretation and Model Selection:

- Initial improvements: As the number of predictors increases from 1 to 7, there is a rapid improvement in model performance:
- Adj.R2 increases sharply (from 0.68 to 0.90).
- CP drops dramatically (from 2185.83 to 3.73).
- BIC becomes much more negative (from -1129.74 to -2258.09).
- RSS decreases substantially (from 90,805.38 to 28,188.68).
- Optimal range: The best model is typically where:
- CP is close to the number of predictors.
- BIC is at its lowest (most negative).
- Adj.R2 is high and stabilizes.
- RSS does not decrease much with additional predictors.
- In your tables, models 7 and 8 stand out:
- Model 7: Adj.R2 = 0.9004, CP = 3.73, BIC = -2258.09, RSS = 28,188.68
- Model 8: Adj.R2 = 0.9006, CP = 2.71, BIC = -2254.24, RSS = 28,102.54
- Both have high Adj.R2, CP values close to the number of predictors, and among the lowest BIC and RSS values.
- Diminishing returns and overfitting: After model 8, improvements in Adj.R2 and RSS are minimal, while CP starts to exceed the number of predictors, and BIC becomes less negative. This suggests adding more predictors does not enhance model performance and may lead to overfitting.

Conclusion:

Model 7 or Model 8 is optimal, as they balance explanatory power and simplicity, with CP values close to the number of predictors, the lowest BIC, and minimal RSS. Adding more predictors beyond this point does not provide meaningful improvement and increases the risk of overfitting

```r
cp_min <- which.min(summary(model_1)$cp)
cp_min
bic_min <- which.min(summary(model_1)$bic)
bic_min
adjr2_max <- which.max(summary(model_1)$adjr2)
adjr2_max
```

Figure 24 : Cp min, BIC min and adj R2 max

We got 3 models from the best subset selection method.

BIC min

```r
mod_7<-
lm(exam_score~study_hours_per_day+social_media_hours+netflix_hours+attendance_percentage+sleep_hours+exercise_frequency+mental_health_rating
,data=data)
summary(mod_7)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + exercise_frequency +
    mental_health_rating, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9509  -3.3953  -0.0283   3.6680  15.9059

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.15722    1.89252   3.253  0.00118 **
study_hours_per_day   9.57456    0.11503  83.238  < 2e-16 ***
social_media_hours   -2.61978    0.14413 -18.177  < 2e-16 ***
netflix_hours        -2.27708    0.15697 -14.507  < 2e-16 ***
attendance_percentage  0.14473    0.01797   8.054 2.28e-15 ***
sleep_hours           2.00462    0.13764  14.564  < 2e-16 ***
exercise_frequency    1.45187    0.08338  17.413  < 2e-16 ***
mental_health_rating  1.94891    0.05924  32.897  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.331 on 992 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.9004
F-statistic:  1291 on 7 and 992 DF,  p-value: < 2.2e-16
```

Figure 23 : Summary of the BIC min

Cp min

```r
mod_8 <-
lm(exam_score~study_hours_per_day+social_media_hours+netflix_hours+attendance_percentage+sleep_hours+diet_quality+exercise_frequency+mental_he
alth_rating ,data=data)
summary(mod_8)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + diet_quality +
    exercise_frequency + mental_health_rating, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-22.257  -3.454   0.092   3.510  15.848

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.57463    1.90508   3.451 0.000582 ***
study_hours_per_day   9.56948    0.11525  83.035  < 2e-16 ***
social_media_hours   -2.61032    0.14414 -18.110  < 2e-16 ***
netflix_hours        -2.27476    0.15690 -14.498  < 2e-16 ***
attendance_percentage  0.14296    0.01799   7.946 5.23e-15 ***
sleep_hours           2.00072    0.13763  14.537  < 2e-16 ***
diet_qualityGood     -0.68653    0.37625  -1.825 0.068349 .
diet_qualityPoor     -0.26222    0.46963  -0.558 0.576736
exercise_frequency    1.45705    0.08339  17.473  < 2e-16 ***
mental_health_rating  1.95613    0.05935  32.958  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.327 on 990 degrees of freedom
Multiple R-squared:  0.9014,    Adjusted R-squared:  0.9005
F-statistic:  1006 on 9 and 990 DF,  p-value: < 2.2e-16
```

Figure 24 : Summary of the Cp min

Adjusted R2 max

```
mod_9 <-
lm(exam_score~study_hours_per_day+social_media_hours+netflix_hours+attendance_percentage+sleep_hours+diet_quality+exercise_frequency+internet_
quality+mental_health_rating ,data=data)
summary(mod_9)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + diet_quality +
    exercise_frequency + internet_quality + mental_health_rating,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9670  -3.4668   0.1181   3.5208  15.6908

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.84003    1.91742   3.567 0.000378 ***
study_hours_per_day   9.56862    0.11532  82.976  < 2e-16 ***
social_media_hours   -2.60141    0.14431 -18.026  < 2e-16 ***
netflix_hours        -2.26903    0.15705 -14.448  < 2e-16 ***
attendance_percentage 0.14256    0.01802   7.913 6.73e-15 ***
sleep_hours           2.00037    0.13766  14.532  < 2e-16 ***
diet_qualityGood     -0.68036    0.37633  -1.808 0.070929 .
diet_qualityPoor     -0.28866    0.47015  -0.614 0.539380
exercise_frequency    1.45416    0.08345  17.426  < 2e-16 ***
internet_qualityGood -0.47141    0.37050  -1.272 0.203538
internet_qualityPoor -0.08122    0.49889  -0.163 0.870704
mental_health_rating  1.95165    0.05946  32.824  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.328 on 988 degrees of freedom
Multiple R-squared:  0.9016,    Adjusted R-squared:  0.9005
F-statistic: 822.8 on 11 and 988 DF,  p-value: < 2.2e-16
```

Figure 25 : Summary of the adj R2 Max

Comparing the R squared and AIC of the 3 models

```
summary(mod_7)$adj.r.squared
summary(mod_8)$adj.r.squared
summary(mod_9)$adj.r.squared
```

```
[1] 0.9003729
[1] 0.900508
[1] 0.9004833
```

```{r}
AIC(mod_7,mod_8,mod_9)
```

Description: df [3 × 2]

	df <dbl>	AIC <dbl>
mod_7	9	6194.798
mod_8	11	6195.422
mod_9	13	6197.649

3 rows

```{r}
BIC(mod_7,mod_8,mod_9)
```

Description: df [3 × 2]

	df <dbl>	BIC <dbl>
mod_7	9	6238.967
mod_8	11	6249.408
mod_9	13	6261.449

Figure 26 : Comparing the R squared and AIC of the 3 models

Best Model Selection

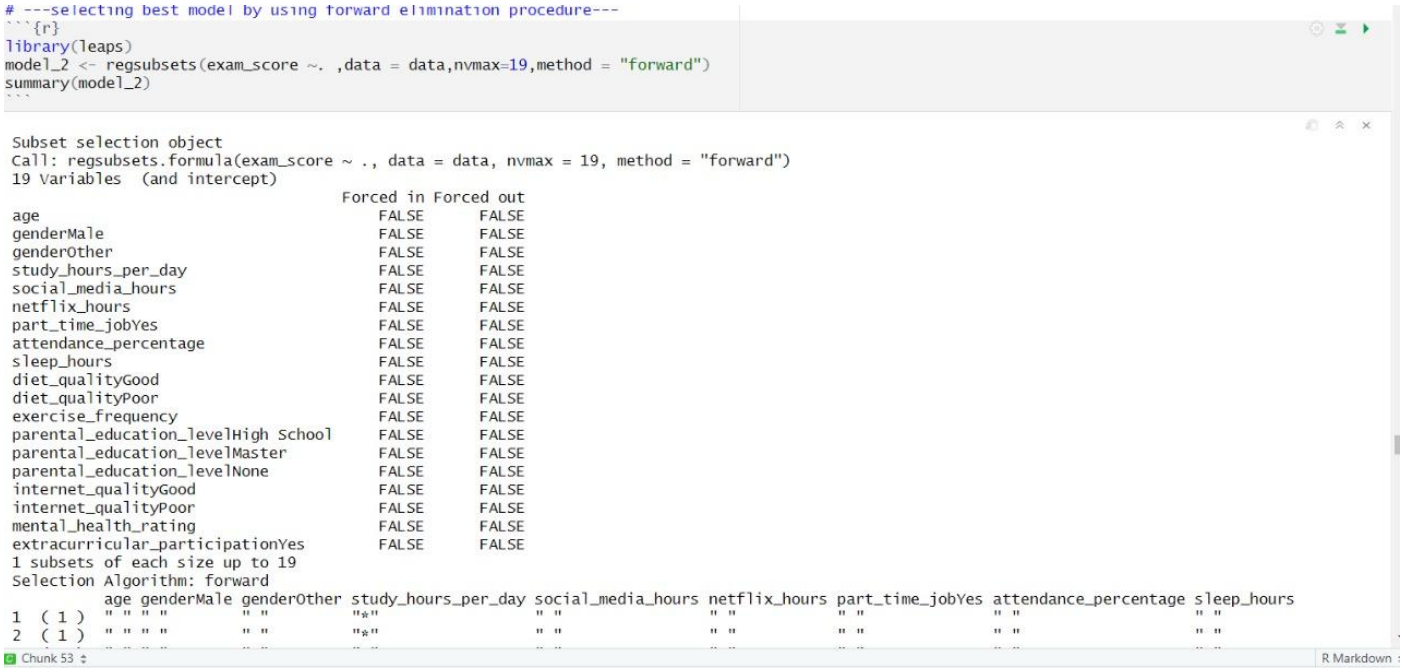
Among the three candidate models (mod_7, mod_8, mod_9), **mod_7** is the best choice. It achieves the lowest values of both AIC and BIC indicating superior parsimony while maintaining an adjusted R^2 comparable to the alternative.

Model	Adjusted R^2	AIC	BIC
mod_7	0.90037	6194.798	6238.967
mod_8	0.90051	6195.422	6249.408
mod_9	0.90048	6197.649	6261.449

Although mod_8 has a marginally higher adjusted R^2 (0.90051 vs. 0.90037), mod_7's notably lower AIC and BIC demonstrate a better balance between goodness-of-fit and model complexity.

Therefore, **mod_7** is the recommended final model.

Forward Selection Method



```
# ---selecting best model by using forward elimination procedure---
```{r}
library(leaps)
model_2 <- regsubsets(exam_score ~. ,data = data,nvmax=19,method = "forward")
summary(model_2)
```

Subset selection object
Call: regsubsets.formula(exam_score ~ ., data = data, nvmax = 19, method = "forward")
19 Variables (and intercept)
                                Forced in Forced out
age   FALSE   FALSE
genderMale   FALSE   FALSE
genderOther   FALSE   FALSE
study_hours_per_day   FALSE   FALSE
social_media_hours   FALSE   FALSE
netflix_hours   FALSE   FALSE
part_time_jobYes   FALSE   FALSE
attendance_percentage   FALSE   FALSE
sleep_hours   FALSE   FALSE
diet_qualityGood   FALSE   FALSE
diet_qualityPoor   FALSE   FALSE
exercise_frequency   FALSE   FALSE
parental_education_levelHigh School FALSE   FALSE
parental_education_levelMaster   FALSE   FALSE
parental_education_levelNone   FALSE   FALSE
internet_qualityGood   FALSE   FALSE
internet_qualityPoor   FALSE   FALSE
mental_health_rating   FALSE   FALSE
extracurricular_participationYes   FALSE   FALSE
1 subsets of each size up to 19
Selection Algorithm: forward
         age genderMale genderOther study_hours_per_day social_media_hours netflix_hours part_time_jobYes attendance_percentage sleep_hours
1 (1) " " " "   " "   "*"   " "   " "   " "   " "   " "
2 (1) " " " "   " "   "*"   " "   " "   " "   " "   " "

Chunk 53 ÷ R Markdown

Figure 27 : Forward selection procedure
```

Create a data frame including all the criterion values for all the models

```
#Create a data frame including all the crieterion values for all the models
res.sum <- summary(model_2)
criterion_2 <- data.frame(
  model = 1:19,
  Adj.R2 = (res.sum$adjr2),
  CP = (res.sum$cp),
  BIC = (res.sum$bic),
  RSS = res.sum$rss
)
criterion_2
```

Description: df [19 × 5]

model <int>	Adj.R2 <dbl>	CP <dbl>	BIC <dbl>	RSS <dbl>
1	0.6809964	2185.826448	-1129.739	90805.38
2	0.7862759	1135.604781	-1524.350	60776.28
3	0.8200278	799.493227	-1690.329	51126.99
4	0.8514238	487.483345	-1876.131	42165.54
5	0.8729408	274.242184	-2026.674	36022.83
6	0.8939647	66.325520	-2201.653	30032.06
7	0.9003729	3.733341	-2258.090	28188.68
8	0.9005771	2.714923	-2254.243	28102.54
9	0.9006441	3.058302	-2249.019	28055.26
10	0.9006326	4.178595	-2243.007	28030.16

1-10 of 19 rows Previous 1 2 Next

Figure 28: : Create a data frame including all the criterion values for all the models

Create the plot for criterion values

```
library(ggplot2)
ggplot(criterion_2, aes(model)) +
geom_line(aes(y = Adj.R2, colour = "Adj.R2")) +
geom_line(aes(y= CP, colour = "Cp")) +
geom_line(aes(y=BIC, colour = "BIC")) +
geom_line(aes(y= RSS, colour = "RSS"))
```

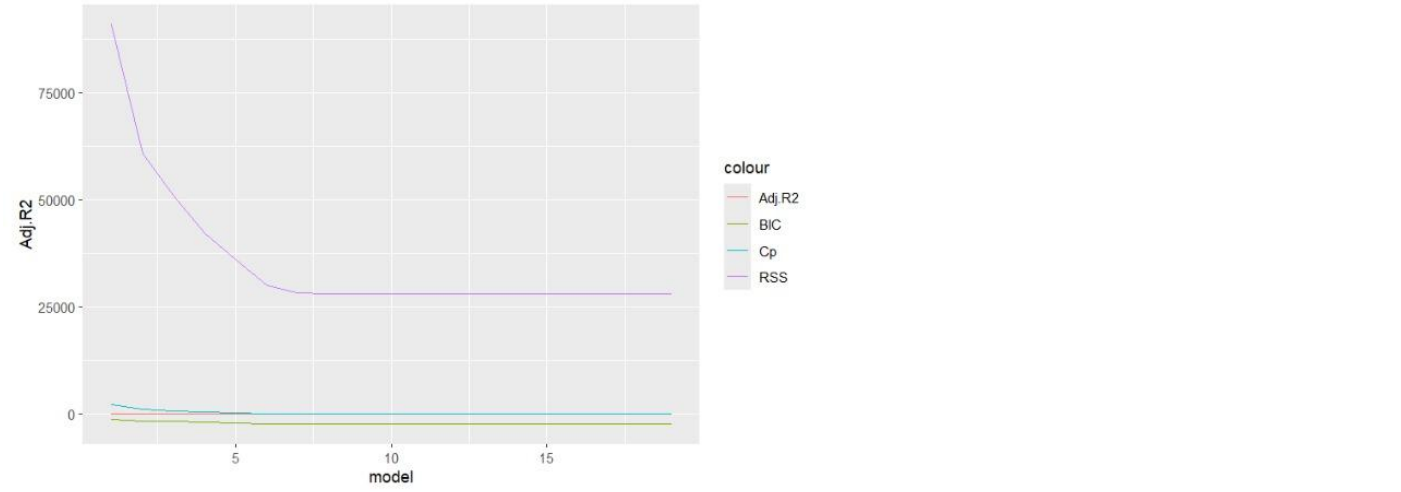


Figure 29 :Create the plot for criterion values

Standardizing criterion values

```r
criterion_std <- cbind(model= criterion$model, scale(criterion[,-1]))
criterion_std <- as.data.frame(criterion_std)
head(criterion_std)
```

Description: df [6 × 5]

	model <dbl>	Adj.R2 <dbl>	CP <dbl>	BIC <dbl>	RSS <dbl>
1	1	-2.8744838	2.8802802	2.565772079	2.8717670
2	2	-1.2297586	1.2264971	1.435870674	1.2313906
3	3	-0.7024711	0.6972225	0.960617815	0.7042866
4	4	-0.2119874	0.2059009	0.428607083	0.2147562
5	5	0.1241614	-0.1298898	-0.002446329	-0.1207971
6	6	0.4526057	-0.4572960	-0.503469769	-0.4480498

Figure 30: : Create a data frame including all the criterion values for all the models after standardizing

```r
library(ggplot2)
ggplot(criterion_std, aes(model)) +
geom_line(aes(y = Adj.R2, colour = "Adj.R2")) +
geom_line(aes(y= CP, colour = "Cp")) +
geom_line(aes(y=BIC, colour = "BIC")) +
geom_line(aes(y= RSS, colour = "RSS"))
```

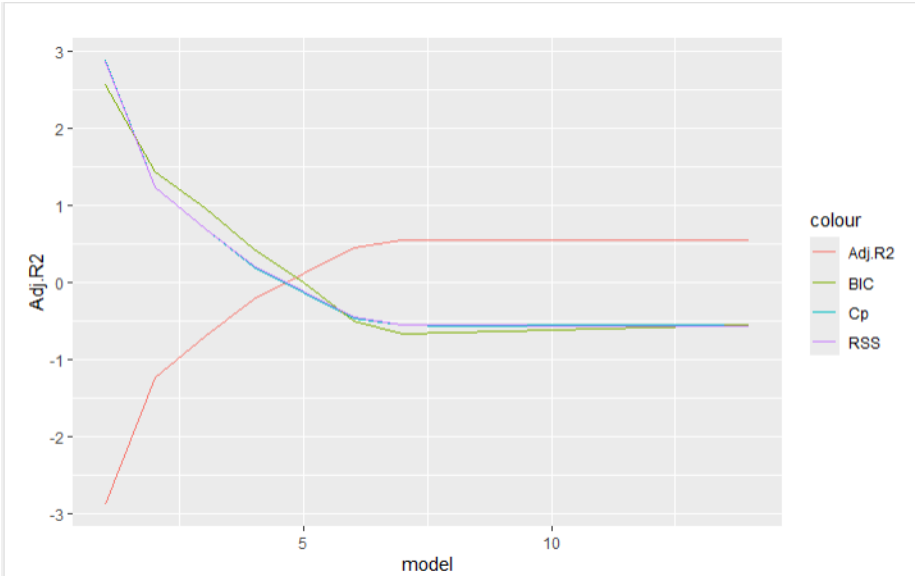


Figure 31 : Plot of criterion values after standardizing

Now it's possible to compare almost all models where model 7 appears as the better model according to the above criterion. So we can say that best model is 7 model

Exam score = study hours per day + social media hours + Netflix hours + attendance percentage + sleep hours + exercise frequency + Mental health rating

Stepwise Model Section

```
install.packages("MASS")
```{r}
# Selecting the best model for the data set using stepwise model selection
library(MASS)

model_step <- lm(exam_score~., data = data)
step_model <- stepAIC(model_step, direction = "both", trace = TRUE)
```
```

```
Start:  AIC=3371.06
exam_score ~ age + gender + study_hours_per_day + social_media_hours +
    netflix_hours + part_time_job + attendance_percentage + sleep_hours +
    diet_quality + exercise_frequency + parental_education_level +
    internet_quality + mental_health_rating + extracurricular_participation

                              Df Sum of Sq    RSS    AIC
- parental_education_level     3        44  28012 3366.6
- gender                       2        26  27994 3368.0
- internet_quality             2        50  28018 3368.8
- extracurricular_participation 1         0  27968 3369.1
- age                          1         1  27969 3369.1
- part_time_job                1         7  27975 3369.3
- diet_quality                 2        94  28062 3370.4
<none>                                     27968 3371.1
- attendance_percentage        1      1770  29738 3430.4
- sleep_hours                  1      5906  33874 3560.7
- netflix_hours                1      5968  33936 3562.5
- exercise_frequency           1      8510  36478 3634.7
- social_media_hours           1      9205  37173 3653.6
- mental_health_rating         1     29934  57902 4096.7
- study_hours_per_day          1    195185 223153 5445.9

Step:  AIC=3366.64
```

Figure 32 : model_2 Stepwise Model Section

Fitted model is -> exam_score =study_hours_per_day + social_media_hours + netflix_hours+attendance_percentage + sleep_hours + exercise_frequency + mental_health_rating

Effect Model-: Y(exam_score) = 6.15722 + 9.57456(study_hours_per_day) - 2.61978(social_media_hours) - 2.27708(netflix_hours) + 0.14473(attendance_percentage) + 2.00462(sleep_hours) + 1.45187(exercise_frequency) + 1.94891(mental_health_rating)

Summary of the Final Model Selected by Stepwise Regression

```
# Summarized the model
summary(step_model)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + exercise_frequency +
    mental_health_rating, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9509  -3.3953  -0.0283   3.6680  15.9059

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.15722    1.89252   3.253  0.00118 **
study_hours_per_day  9.57456    0.11503  83.238  < 2e-16 ***
social_media_hours  -2.61978    0.14413 -18.177  < 2e-16 ***
netflix_hours       -2.27708    0.15697 -14.507  < 2e-16 ***
attendance_percentage  0.14473    0.01797   8.054 2.28e-15 ***
sleep_hours          2.00462    0.13764  14.564  < 2e-16 ***
exercise_frequency   1.45187    0.08338  17.413  < 2e-16 ***
mental_health_rating  1.94891    0.05924  32.897  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.331 on 992 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.9004
F-statistic:  1291 on 7 and 992 DF,  p-value: < 2.2e-16
```

Figure 33 : Summary of the final stepwise regression model

Study hours per day have the strongest positive impact on exam performance. For each additional hour of studying, scores increase by approximately 9.58 points. Social media use and Netflix watching

negatively affect exam scores, indicating that excessive screen time may hinder academic performance. Positive lifestyle habits—such as higher attendance, more sleep, regular exercise, and better mental health—are all significantly associated with higher exam scores. All predictors in the model are statistically significant ($p < 0.001$), suggesting that the relationships observed are highly unlikely to be due to chance. Overall, students who prioritize studying, maintain good health habits, and limit screen time tend to perform better academically. This model can be useful for identifying areas where students might improve their habits to achieve better academic outcomes. Multiple R-squared is 0.9011 which is approximate to 1 so we can say that it is suitable model for data set.

Step Model

```
step_model
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + exercise_frequency +
    mental_health_rating, data = data)

Coefficients:
          (Intercept)    study_hours_per_day     social_media_hours         netflix_hours  attendance_percentage
               6.1572                 9.5746                -2.6198               -2.2771                 0.1447
          sleep_hours     exercise_frequency   mental_health_rating
               2.0046                 1.4519                 1.9489
```

Figure 34 : The final model, selected through stepwise regression, includes seven predictors

Check the multicollinearity in fitted model

```
vif(step_model)
```

```
  study_hours_per_day    social_media_hours         netflix_hours attendance_percentage           sleep_hours     exercise_frequency
             1.003629              1.003859              1.001235              1.002867              1.001751              1.002650
mental_health_rating
             1.000470
```

Figure 35 : Check the multicollinearity

Model Fitting and Evaluation

Multiple linear regression models were fitted to the data, both for individual predictors and for the selected combination of predictors.

Model diagnostics included:

- Checking residual plots for linearity and homoscedasticity.
- Assessing normality of residuals via Q-Q plots.
- Identifying influential observations using leverage and Cook's distance plots

The final model explained approximately 90% of the variance in exam scores (Adjusted $R^2 \approx 0.90$), with all included predictors being statistically significant ($p < 0.001$)

Model diagnostic

```
par(mfrow=c(2,2))
plot(step_model)
```

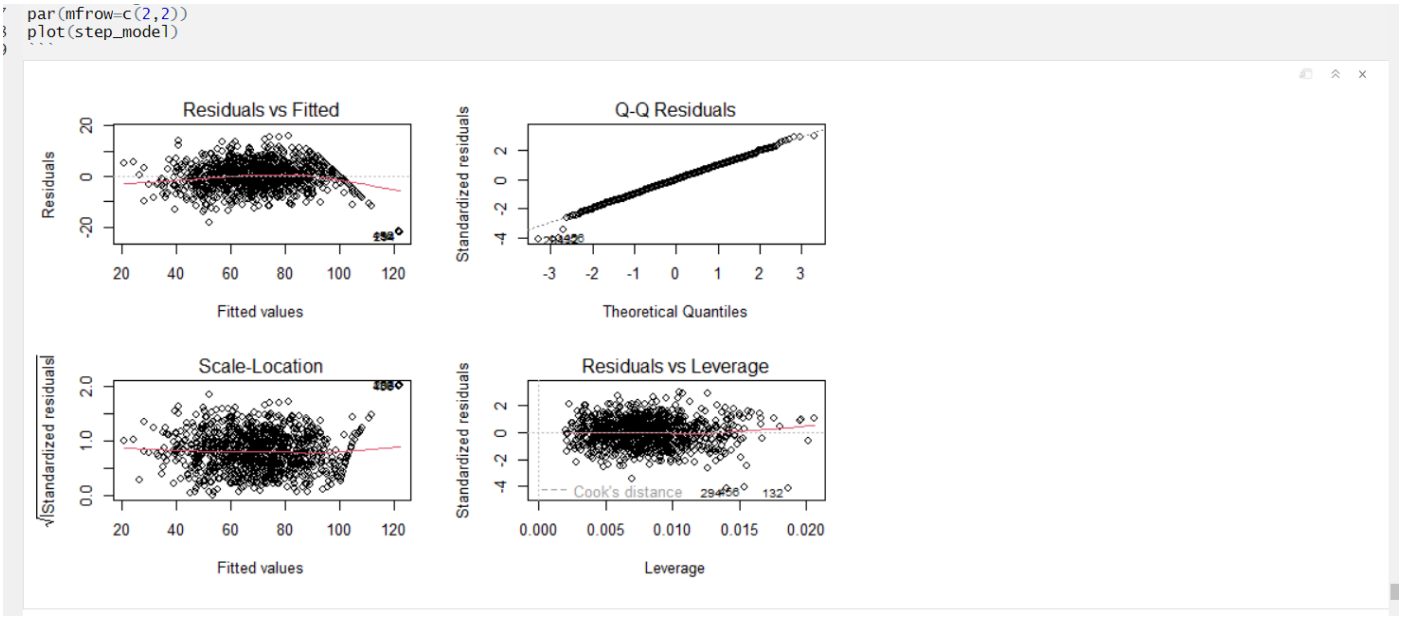


Figure 36 : Model diagnostic

Q-Q Plot

```
# Q-Q Plot
plot(better_model_fwd, which = 2)
```

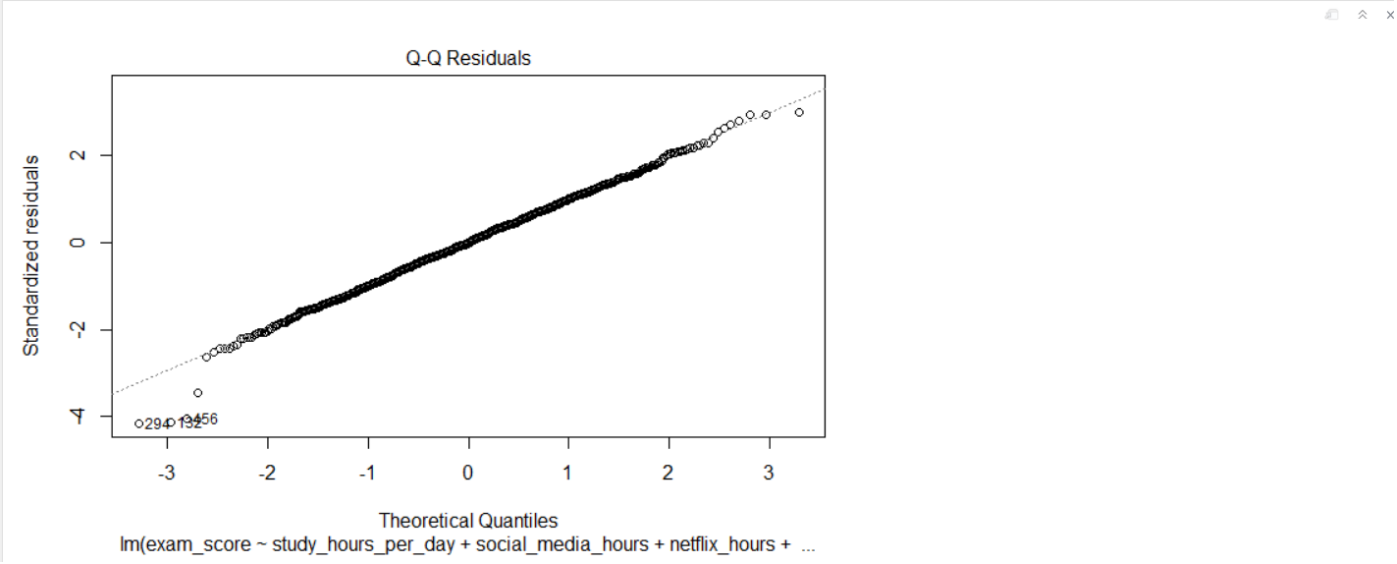


Figure 37 : Q-Q plot

The Q-Q Plot demonstrated approximate normality, with residuals largely following the theoretical line. Minor deviations at the tails were observed but did not substantially violate the normality assumption. In this plot the majority of data align closely on the diagonal reference line we can say that this data set will satisfy the normality assumption.

Check the normality assumption

```
# check Normality Assumption
shapiro.test(residuals(better_model_fwd))
```

```
        Shapiro-Wilk normality test

data:  residuals(better_model_fwd)
W = 0.99523, p-value = 0.003196
```

Figure 38 : Shapiro-Wilk normality test

The Shapiro-Wilk test indicated slight deviation from normality (p = 0.003), likely due to the large sample size. Visual inspection of the Q-Q plot (not shown) confirmed residuals were approximately normal, with minor tail deviations. Residuals vs. Fitted and Run Order plots showed no patterns, supporting homoscedasticity and independence. Collectively, the model assumptions were reasonably met

Residuals vs Fitted

```
plot(better_model_fwd, which = 1)
```

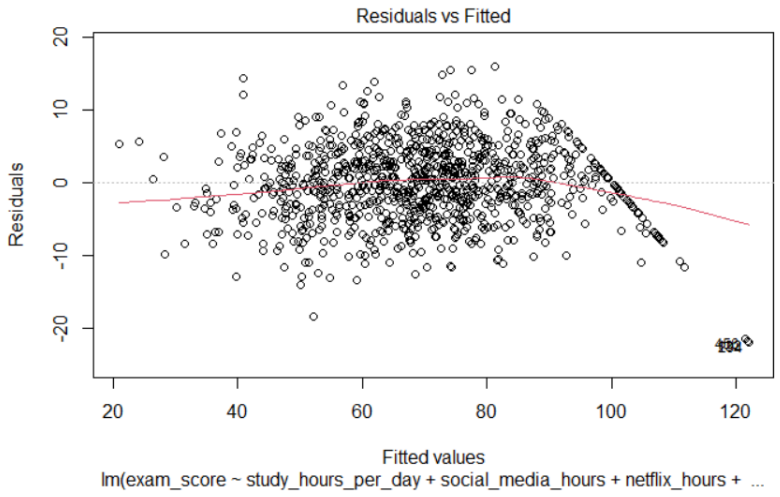


Figure 39 : Residuals vs Fitted

The Residuals vs. Fitted Plot showed no clear patterns, with residuals evenly dispersed around zero. This confirms the assumptions of linearity and homoscedasticity in our model.

Check residuals are uncorrelated

```
#Residual vs Rund order/observation order/Time plot
plot(residuals(better_model_fwd),
xlab = "Observation Order",
ylab = "Residuals",
main = "Residuals vs Rund order/observation order/Time plot")
abline(h = 0, col = "red")
```

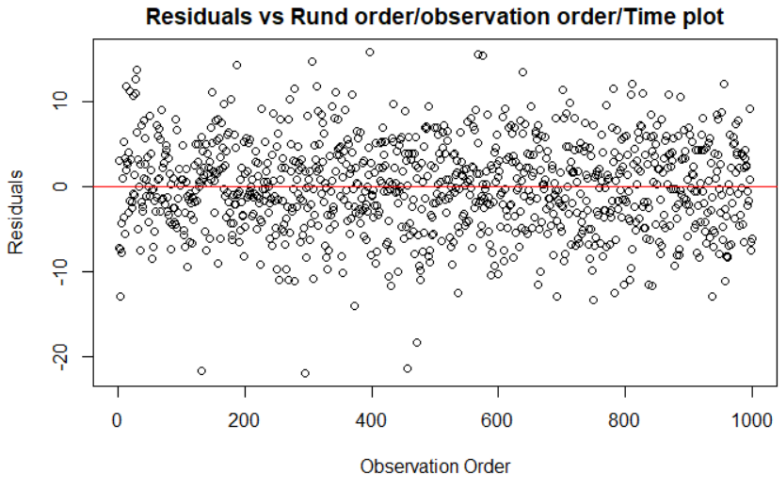


Figure 40 : Check residuals are uncorrelated

Residuals vs. Run Order Plot confirmed the independence assumption, as residuals were randomly distributed without systematic trends, indicating no temporal dependencies in the data.

4. Interpretation

The effects of each predictor were interpreted in terms of their impact on exam scores, holding other variables constant. For example, each additional hour of study per day was associated with a significant increase in exam score, while increased social media or Netflix usage was associated with lower scores. Positive lifestyle factors such as higher attendance, more sleep, regular exercise, and better mental health were all linked to improved academic performance

Results and Discussion

Result

This section presents the key findings from the multiple linear regression analysis conducted to evaluate how various student lifestyle habits affect academic performance, as measured by final exam scores.

Full Model

```r
# Fit the model
model<-lm(exam_score~.,data = data)
summary(model)
```

```
Call:
lm(formula = exam_score ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.2097  -3.4768   0.0789   3.4456  15.5955

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        7.17718    2.50263   2.868  0.00422 **
age                               -0.01209    0.07365  -0.164  0.86959
genderMale                         0.14621    0.34805   0.420  0.67452
genderOther                        0.79264    0.86559   0.916  0.36003
study_hours_per_day                9.57454    0.11577  82.700  < 2e-16 ***
social_media_hours                -2.60220    0.14489 -17.960  < 2e-16 ***
netflix_hours                     -2.28156    0.15777 -14.462  < 2e-16 ***
part_time_jobYes                   0.21121    0.41410   0.510  0.61013
attendance_percentage              0.14339    0.01821   7.876 8.94e-15 ***
sleep_hours                        1.99234    0.13849  14.386  < 2e-16 ***
diet_qualityGood                  -0.68310    0.37777  -1.808  0.07088 .
diet_qualityPoor                  -0.27224    0.47306  -0.575  0.56510
exercise_frequency                 1.45004    0.08397  17.268  < 2e-16 ***
parental_education_levelHigh School -0.15995    0.39605  -0.404  0.68639
parental_education_levelMaster     -0.41093    0.50782  -0.809  0.41860
parental_education_levelNone       -0.70202    0.63313  -1.109  0.26778
internet_qualityGood               -0.47289    0.37292  -1.268  0.20507
internet_qualityPoor               -0.08167    0.50324  -0.162  0.87111
mental_health_rating               1.94417    0.06003  32.386  < 2e-16 ***
extracurricular_participationYes   -0.01412    0.36426  -0.039  0.96909
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.342 on 980 degrees of freedom
Multiple R-squared:  0.9018,    Adjusted R-squared:  0.8999
F-statistic: 473.9 on 19 and 980 DF,  p-value: < 2.2e-16
```

Figure 41 : Full Model

There are some significant and insignificant predictor variables. The adjusted R-squared value is 0.8999.

study_hours_per_day ,social_media_hours ,netflix_hours , attendance_percentage ,sleep_hours, mental_health_rating, exercise_frequency variables are the significant and others are not significant

Best subset Selection Method

```r
final_best<-lm(exam_score~study_hours_per_day+social_media_hours+netflix_hours+attendance_percentage+sleep_hours+exercise_frequency+mental_health_rating ,data=data)
summary(final_best)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + exercise_frequency +
    mental_health_rating, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9509  -3.3953  -0.0283   3.6680  15.9059

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.15722    1.89252   3.253  0.00118 **
study_hours_per_day  9.57456    0.11503  83.238  < 2e-16 ***
social_media_hours  -2.61978    0.14413 -18.177  < 2e-16 ***
netflix_hours       -2.27708    0.15697 -14.507  < 2e-16 ***
attendance_percentage  0.14473    0.01797   8.054 2.28e-15 ***
sleep_hours          2.00462    0.13764  14.564  < 2e-16 ***
exercise_frequency   1.45187    0.08338  17.413  < 2e-16 ***
mental_health_rating  1.94891    0.05924  32.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.331 on 992 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.9004
F-statistic:  1291 on 7 and 992 DF,  p-value: < 2.2e-16
```

Figure 42 : Best subset Selection Method

We had taken three models and compared the cp values, BIC values and the adjusted r squared values of those three models. We then selected the best model as the model with the lowest cp and BIC values and the highest adjusted r-squared value.

Forward Selection Method

```r
better_model_fwd <-
lm(exam_score~study_hours_per_day+social_media_hours+netflix_hours+attendance_percentage+sleep_hours+exercise_frequency+mental_health_rating
,data=data)
summary(better_model_fwd)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + exercise_frequency +
    mental_health_rating, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9509  -3.3953  -0.0283   3.6680  15.9059

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.15722    1.89252   3.253  0.00118 **
study_hours_per_day  9.57456    0.11503  83.238  < 2e-16 ***
social_media_hours  -2.61978    0.14413 -18.177  < 2e-16 ***
netflix_hours       -2.27708    0.15697 -14.507  < 2e-16 ***
attendance_percentage  0.14473    0.01797   8.054 2.28e-15 ***
sleep_hours          2.00462    0.13764  14.564  < 2e-16 ***
exercise_frequency   1.45187    0.08338  17.413  < 2e-16 ***
mental_health_rating  1.94891    0.05924  32.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.331 on 992 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.9004
F-statistic:  1291 on 7 and 992 DF,  p-value: < 2.2e-16
```

Figure 43 : Best subset Selection Method

Fitted Model

Y(exam_score) = 6.1572193 + 9.5745576*(study_hours_per_day) -2.6197830*(social_media_hours) - 2.2770773 *(netflix_hours) + 0.1447283*(attendance_percentage) + 2.0046195*(sleep_hours) + 1.4518742*(exercise_frequency) +1.9489111*(mental_health_rating)

Therefore, we can say that both stepwise and forward selection methods the change in the exam score when a one unit increases or decreases above these predictors.

- study_hours_per_day(9.5745576) - For every additional hour spent per day, the exam score increases by 9.5745576 points, assuming all other variables are held constant. This is the strongest predictor of exam scores in this model.
- social_media_hours(-2.6197830) - For additional hour spent on social media, the exam score decreases by 2.6197830 points, holding other variables constant. Note the negative effect that mean more social media use is associated with lower exam scores.
- netflix_hours(-2.2770773) - For every additional hour spent watching Netflix, the exam score decreases by 2.2770773 points, holding other variables constant. Similar to social media, leisure screen time has a negative impact.
- attendance_percentage(0.1447283) - For every 1% increase in attendance, the exam score increases by0.1447283 points, holding the other variables constant. While positive, the effect is small compared to other predictors.
- sleep_hours(2.0046195) - For every additional unit hour of sleep per night, the exam score increases by 2.0046195points,holding other variables constant. Adequate sleep is important for performance.
- exercise_frequency(1.4518742) - For every additional unit increase in exercise frequency(ex: from "rarely" to "sometimes"), the exam score increases by 1.4518742 points, holding other variables constant. Regular exercise may reduce stress.
- mental_health_rating(1.9489111) - For every one unit improvement in mental health rating , the exam score increases by 1.9489111 points, holding other variables are constant. Better mental health is associated with higher scores.

The strongest positive predictor - study_hours_per_day(9.5745576)

The strongest negative predictors - social_media_hours(-2.6197830) and netflix_hours(-2.2770773) Mental Health Rating - Also positively associated (+1.95), showing that students with better mental well-being tend to perform better in exams.

Model Performance Metrics

Adjusted R^2: Indicates the proportion of variance in exam scores explained by the model

AIC: Used to identify the best-fitting model with the lowest information loss

VIF: Confirmed that multicollinearity is within acceptable limits

Positive habits such as studying regularly, sleeping well, exercising, and maintaining good mental health are associated with higher academic performance.

Negative behaviors like excessive screen time (Netflix, social media) reduce exam scores.

Although attendance has a smaller impact compared to other factors, it still contributes positively.

Stepwise Model Selection

```
# Summarized the model
summary(step_model)
```

```
Call:
lm(formula = exam_score ~ study_hours_per_day + social_media_hours +
    netflix_hours + attendance_percentage + sleep_hours + exercise_frequency +
    mental_health_rating, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-21.9509  -3.3953  -0.0283   3.6680  15.9059

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.15722    1.89252   3.253  0.00118 **
study_hours_per_day   9.57456    0.11503  83.238  < 2e-16 ***
social_media_hours   -2.61978    0.14413 -18.177  < 2e-16 ***
netflix_hours        -2.27708    0.15697 -14.507  < 2e-16 ***
attendance_percentage 0.14473    0.01797   8.054 2.28e-15 ***
sleep_hours           2.00462    0.13764  14.564  < 2e-16 ***
exercise_frequency    1.45187    0.08338  17.413  < 2e-16 ***
mental_health_rating  1.94891    0.05924  32.897  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.331 on 992 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared:  0.9004
F-statistic:  1291 on 7 and 992 DF,  p-value: < 2.2e-16
```

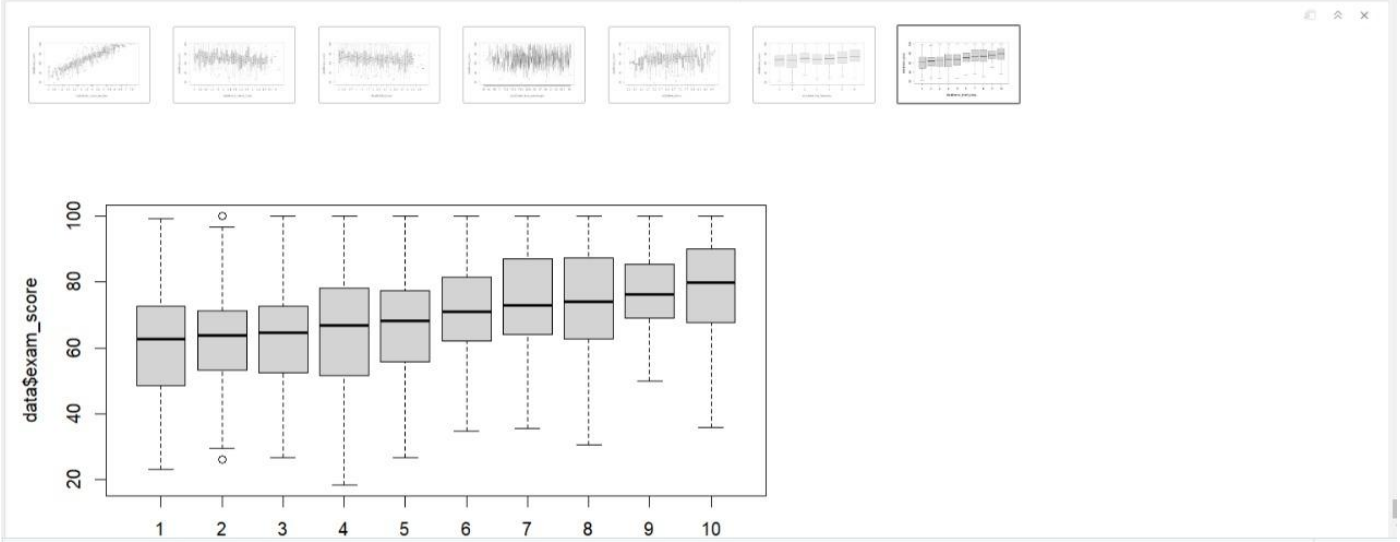
Figure 44 : Stepwise Model Selection

Study hours per day have the strongest positive impact on exam performance. For each additional hour of studying, scores increase by approximately 9.58 points. Social media use and Netflix watching negatively affect exam scores, indicating that excessive screen time may hinder academic performance. Positive lifestyle habits—such as higher attendance, more sleep, regular exercise, and better mental health—are all significantly associated with higher exam scores. All predictors in the model are statistically significant ($p < 0.001$), suggesting that the relationships observed are highly unlikely to be due to chance.

Overall, students who prioritize studying, maintain good health habits, and limit screen time tend to perform better academically. This model can be useful for identifying areas where students might improve their habits to achieve better academic outcomes. Multiple R-squared is 0.9011 which is approximate to 1 so we can say that it is suitable model for data set.

Box-plots of the significant variables

```
boxplot(data$exam_score~data$study_hours_per_day)
boxplot(data$exam_score~data$social_media_hours)
boxplot(data$exam_score~data$netflix_hours)
boxplot(data$exam_score~data$attendance_percentage)
boxplot(data$exam_score~data$sleep_hours)
boxplot(data$exam_score~data$exercise_frequency)
boxplot(data$exam_score~data$mental_health_rating)
```



Discussion

This study shows that exam performance is influenced by both study habits and lifestyle choices.

The most important factor is study time students who spend more time studying tend to get higher exam scores. In contrast, spending too much time on social media or watching Netflix has a negative effect on exam results, likely because it reduces time and focusses for studying. Lifestyle factors such as getting enough sleep, exercising regularly, and having good mental health also help students perform better. These healthy habits improve focus and reduce stress, which supports better learning.

Class attendance was another positive factor. Students who attend classes regularly score better, showing that being present in class helps with understanding and performance. Some factors, like gender, parents' education, internet quality, diet, and part-time jobs, did not show a strong effect on exam scores in this study.

Overall, the results highlight that academic success is not only about studying hard but also about having a balanced and healthy lifestyle. Students, teachers, and parents can use this information to build better routines that support learning and success.

Conclusion

This study successfully investigated the relationship between student lifestyle habits and academic performance using a comprehensive dataset of 1,000 students. Through rigorous statistical analysis including exploratory data analysis, multiple linear regression, and advanced model selection techniques, several key findings emerged that provide valuable insights into factors influencing academic success.

The final model, selected through stepwise regression analysis, demonstrated exceptional predictive power with an adjusted R-squared value of 0.9011, explaining approximately 90% of the variance in exam scores. This high explanatory power indicates that the identified lifestyle factors are strong predictors of academic performance, validating the research hypothesis that daily behavioral habits significantly impact educational outcomes. Study hours per day emerged as the most influential predictor, with each additional hour of studying associated with a 9.57-point increase in exam scores. This finding underscores the fundamental importance of dedicated study time in academic achievement and aligns with educational theory emphasizing the role of deliberate practice in learning.

Conversely, recreational screen time showed significant negative associations with academic performance. Social media usage ($\beta = -2.62$) and Netflix consumption ($\beta = -2.28$) both demonstrated detrimental effects on exam scores. These findings highlight the potential academic costs of excessive digital entertainment and support concerns about screen time's impact on student focus and study effectiveness. The study also revealed the importance of holistic lifestyle factors in academic success. Attendance percentage ($\beta = 0.14$), sleep hours ($\beta = 2.00$), exercise frequency ($\beta = 1.45$), and mental health rating ($\beta = 1.95$) all showed statistically significant positive relationships with exam performance ($p < 0.001$).

Interestingly, several demographic and socioeconomic factors that might be expected to influence academic performance including gender, parental education level, internet quality, and part-time job status did not emerge as significant predictors in the final model. This suggests that personal behavioral choices may be more influential than background characteristics in determining academic outcomes, offering hope that students can improve their performance through lifestyle modifications regardless of their circumstances. The model successfully met key statistical assumptions, with residual analysis confirming appropriate linearity, homoscedasticity, and normality of residuals. Multicollinearity analysis using Variance Inflation Factors (all VIF < 5) confirmed that the predictors were not highly correlated with each other, ensuring the reliability of the regression coefficients.

These findings have important practical implications for students, educators, and educational policymakers. Students can use these evidence-based insights to optimize their daily routines, prioritizing study time while maintaining healthy sleep patterns, regular exercise, and good mental

health. The study's limitations include its reliance on synthetic data and cross-sectional design, which prevents causal inferences. In conclusion, this research demonstrates that academic performance is significantly influenced by modifiable lifestyle behaviors, with study habits, screen time management, and overall wellness playing crucial roles. The strong predictive model developed provides a framework for understanding and improving student academic outcomes through targeted behavioral interventions.

References

1. Sleep Quality, Duration, and Consistency Are Associated with Better Academic Performance in College Students

   Authors: Phillips, A. J., Clerx, W. M., O'Brien, C. S., Sano, A., Barger, L. K., Picard, R. W., ... & Czeisler, C. A. (2017)
   Journal: Nature and Science of Sleep, 9, 205-214
   DOI: https://doi.org/10.2147/NSS.S136915

2. Nightly Sleep Is Key to Student Success

   Authors: Creswell, D., et al. (2023)
   Journal: Proceedings of the National Academy of Sciences
   Institution: Carnegie Mellon University

3. Social Media Usage and Academic Achievement Among Early Adolescents

   Authors: Gordon, M. S., & Ohannessian, C. M. (2021)
   Journal: Youth and Society, 53(6), 883-904
   DOI: https://doi.org/10.1177/0044118X20901533

4. The Influence of Lifestyle on Academic Performance Among Health Profession Students

   Authors: Multiple authors (2024)
   Journal: Available through PMC (PubMed Central)

5. The Impact of Healthy Lifestyles on Academic Achievement Among Italian Adolescents

   Authors: Maniaci, G., La Cascia, C., Giammanco, A., et al. (2021)
   Journal: Current Psychology, 42, 5674-5684
   DOI: https://doi.org/10.1007/s12144-021-01901-z

6. 6The Effects of Physical Activity on Academic Performance in School-Aged Children

   Authors: Multiple authors (2023)
   Journal: Available through PMC (PubMed Central)

7. Associations Between Dietary Intake and Academic Achievement in College Students

   Authors: Multiple authors (2017)
   Journal: Healthcare, 5(4), 60
   DOI: https://doi.org/10.3390/healthcare5040060

Individual Contribution

Task		PS/2021/225	PS/2021/032	PS/2021/034	PS/2021/091	PS/2021/122	PS/2021/137	PS/2021/217	PS/2021/178	PS/2021/190	PS/2021/194	PS/2021/033
Finding a data set												
Activity 01												
Activity 02												
Activity 03												
R project	Data Acquisition and Preparation											
	Exploratory Data Analysis											
	Feature Engineering and Variable Selection											
	Model Fitting and Evaluation											
	Interpretation											
Making presentation												
Presenting												
Final Report	Introduction											
	Methodology											
	Results and discussion											
	Conclusion and References											