Untitled

Sirapu Nandini

2024-11-30

```
options(repos = c(CRAN = "https://cran.r-project.org"))
```

Abstract:

This project investigates how various factors, including motivation level, internet access, study habits, and parental involvement, affect student exam scores. Using a dataset from Kaggle with 6,607 student records, we analyzed key variables such as Hours_Studied, Attendance, Motivation_Level, Internet_Access, and Parental_Involvement. The primary objective was to determine which of these factors significantly contribute to academic performance.

We applied Multiple Linear Regression to explore the relationships between Exam_Score and several independent variables. The analysis revealed that Hours_Studied and Attendance positively impact Exam_Score, with students who study more and attend school regularly achieving higher scores. Additionally, Motivation_Level and Internet_Access also had significant effects, with students having internet access and higher motivation levels performing better.

Further, ANOVA showed that students with different levels of motivation and parental involvement scored differently, confirming that these factors significantly influence academic performance. A T-test for Internet_Access confirmed that students with internet access score higher than those without. Correlation analysis indicated moderate positive correlations between Exam_Score and Hours_Studied and Attendance, but only weak correlations with Sleep_Hours and Previous_Scores.

In conclusion, the findings suggest that factors like study habits, attendance, and access to resources like the internet play a significant role in student success. Educators and policymakers can use these insights to develop strategies that encourage consistent study habits, increase resource access, and foster motivation. The study's limitations include its observational nature, which prevents causal conclusions.

Introduction:

The aim of the project is to find how do various factors like motivational level, Gender, Access-to-resource,Attendance and other explanatory variables effect the student's performance in exams.(Exam_score which is response variable here).

```
#Data loading
url <- "https://raw.githubusercontent.com/srirapunandini/dav---5400/refs/heads/main/StudentPerformanceF
```

```
# Reading the CSV file into R
project_data <- read.csv(url)
```

The Dataset Consists of 6,607 rows and 20 columns.

Printing the first few rows

```r
head(project_data[, 1:5], 6)
```

```
##   Hours_Studied Attendance Parental_Involvement Access_to_Resources
## 1            23         84                  Low                High
## 2            19         64                  Low              Medium
## 3            24         98               Medium              Medium
## 4            29         89                  Low              Medium
## 5            19         92               Medium              Medium
## 6            19         88               Medium              Medium
##   Extracurricular_Activities
## 1                         No
## 2                         No
## 3                        Yes
## 4                        Yes
## 5                        Yes
## 6                        Yes
```

Summary Statistics of the data

```r
# View summary statistics of the dataset
summary(project_data)
```

```
##  Hours_Studied     Attendance     Parental_Involvement Access_to_Resources
##  Min.   : 1.00   Min.   : 60.00   Length:6607          Length:6607
##  1st Qu.:16.00   1st Qu.: 70.00   Class :character     Class :character
##  Median :20.00   Median : 80.00   Mode  :character     Mode  :character
##  Mean   :19.98   Mean   : 79.98
##  3rd Qu.:24.00   3rd Qu.: 90.00
##  Max.   :44.00   Max.   :100.00
##  Extracurricular_Activities  Sleep_Hours     Previous_Scores
##  Length:6607                Min.   : 4.000   Min.   : 50.00
##  Class :character           1st Qu.: 6.000   1st Qu.: 63.00
##  Mode  :character           Median : 7.000   Median : 75.00
##                             Mean   : 7.029   Mean   : 75.07
##                             3rd Qu.: 8.000   3rd Qu.: 88.00
##                             Max.   :10.000   Max.   :100.00
##  Motivation_Level   Internet_Access    Tutoring_Sessions Family_Income
##  Length:6607        Length:6607        Min.   :0.000     Length:6607
##  Class :character   Class :character   1st Qu.:1.000     Class :character
##  Mode  :character   Mode  :character   Median :1.000     Mode  :character
##                                        Mean   :1.494
##                                        3rd Qu.:2.000
##                                        Max.   :8.000
##  Teacher_Quality    School_Type        Peer_Influence     Physical_Activity
##  Length:6607        Length:6607        Length:6607        Min.   :0.000
##  Class :character   Class :character   Class :character   1st Qu.:2.000
##  Mode  :character   Mode  :character   Mode  :character   Median :3.000
##                                                           Mean   :2.968
##                                                           3rd Qu.:4.000
##                                                           Max.   :6.000
##  Learning_Disabilities Parental_Education_Level Distance_from_Home
##  Length:6607           Length:6607              Length:6607
```

```
##  Class :character      Class :character         Class :character
##  Mode  :character      Mode  :character         Mode  :character
##
##
##
##     Gender            Exam_Score
##  Length:6607        Min.   : 55.00
##  Class :character   1st Qu.: 65.00
##  Mode  :character   Median : 67.00
##                     Mean   : 67.24
##                     3rd Qu.: 69.00
##                     Max.   :101.00
```

The dataset contains 6,607 rows and 20 columns, including both numeric and categorical variables. On average, students study for about 20 hours, attend school 80% of the time, and sleep for about 7 hours per day. Exam scores range from 55 to 101, with a mean of 67.24. Other factors include Parental Involvement, Internet Access, and Motivation Level, which are categorical and represent varying levels of support or access.

Data Cleaning :

```
# Check for null values in each column
colSums(is.na(project_data))
```

```
##              Hours_Studied                 Attendance
##                          0                          0
##       Parental_Involvement        Access_to_Resources
##                          0                          0
## Extracurricular_Activities               Sleep_Hours
##                          0                          0
##            Previous_Scores           Motivation_Level
##                          0                          0
##            Internet_Access          Tutoring_Sessions
##                          0                          0
##              Family_Income            Teacher_Quality
##                          0                          0
##                School_Type            Peer_Influence
##                          0                          0
##          Physical_Activity      Learning_Disabilities
##                          0                          0
##   Parental_Education_Level         Distance_from_Home
##                          0                          0
##                     Gender                 Exam_Score
##                          0                          0
```

We found that there are zero null values in the data.

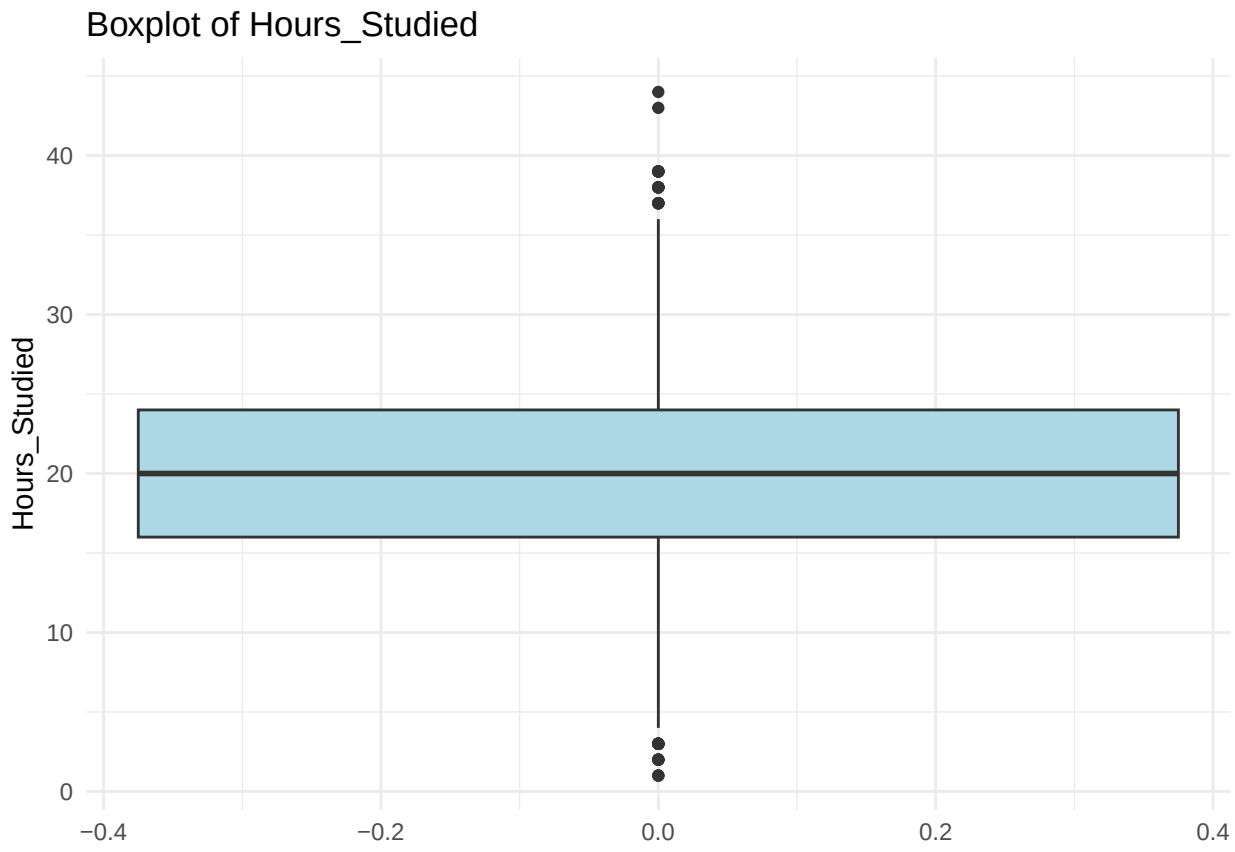
```
#check for outliers
#using Z_scores for checking the outliers
z_scores <- scale(project_data[sapply(project_data, is.numeric)])
outliers <- abs(z_scores) > 3
colSums(outliers)
```

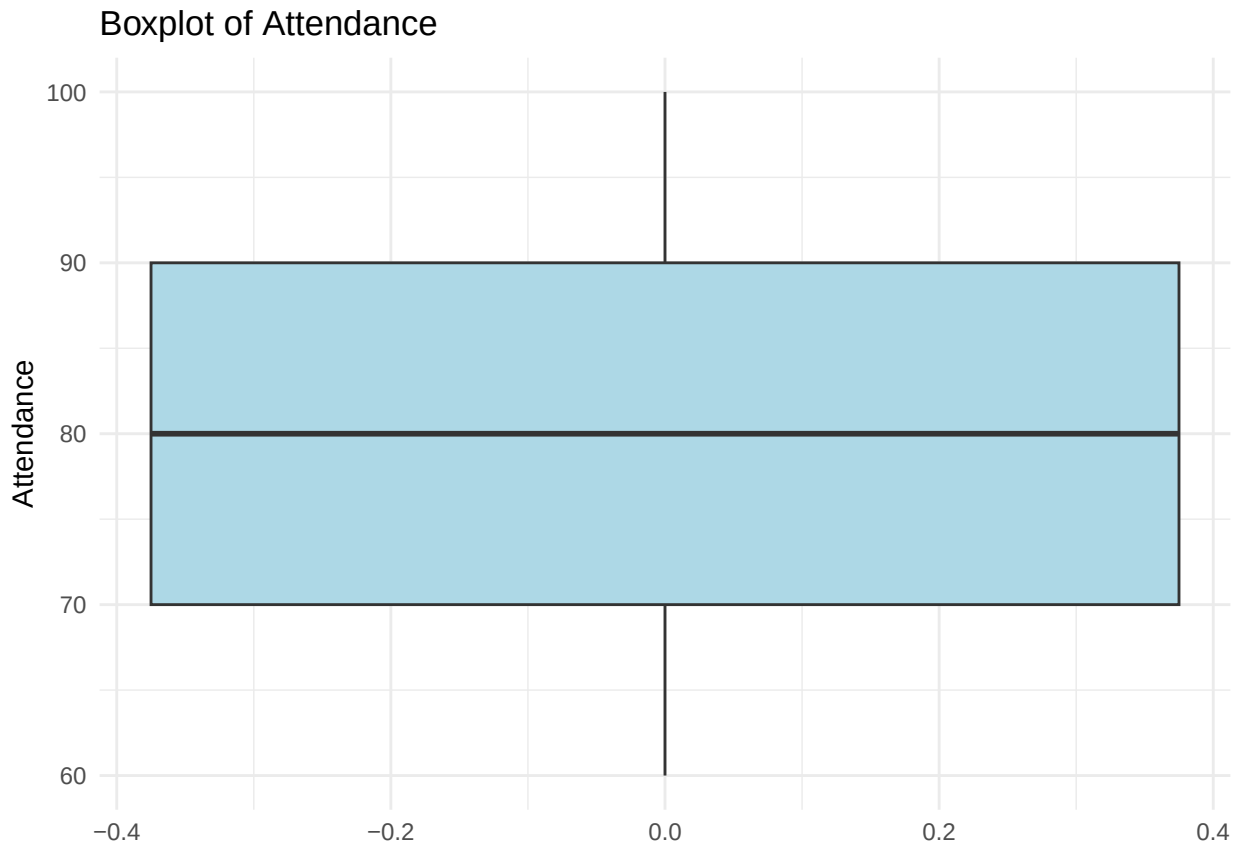
```
##     Hours_Studied        Attendance       Sleep_Hours   Previous_Scores
```

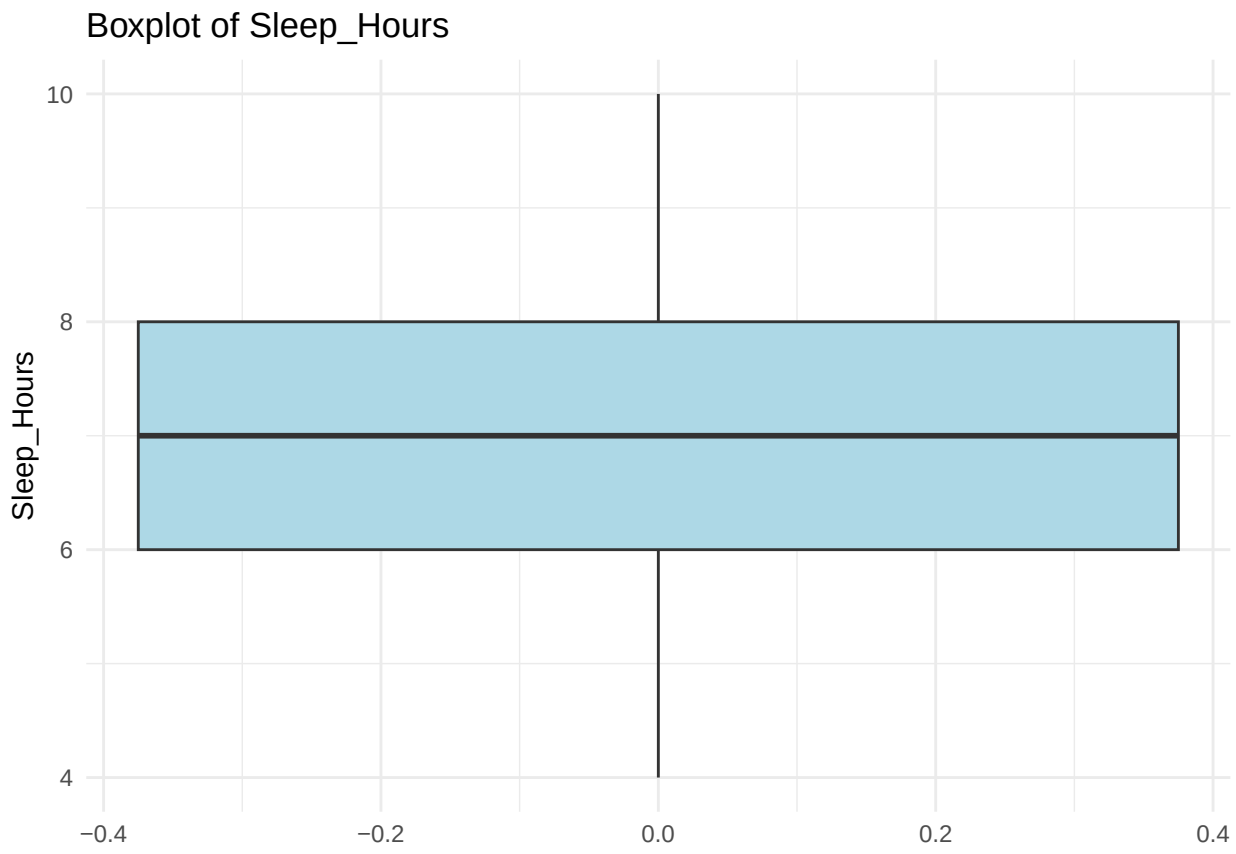
```
##                 25                  0                  0                  0
## Tutoring_Sessions Physical_Activity        Exam_Score
##                 26                  0                 52
```

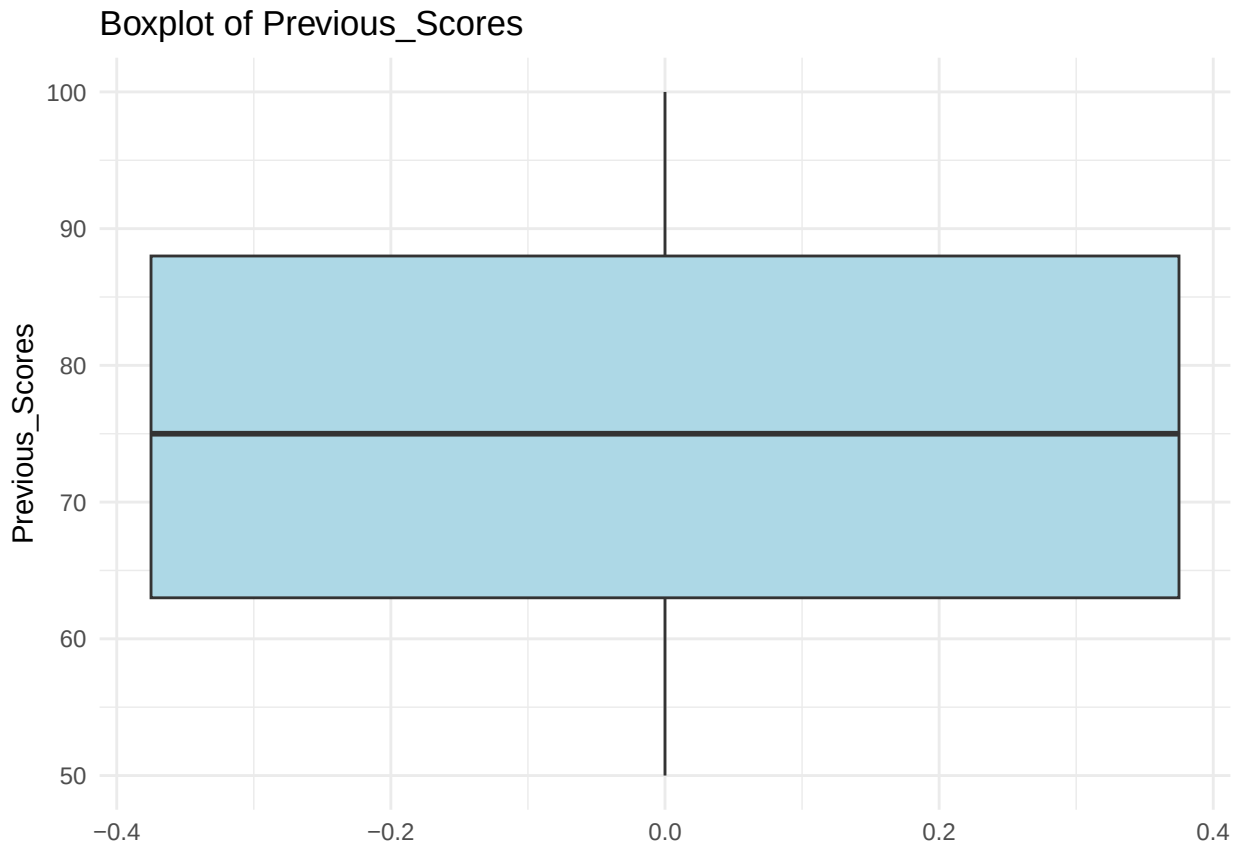
There are some outliers in Hours_Studied, Tutoring_Sessions, Exam_Score columns.

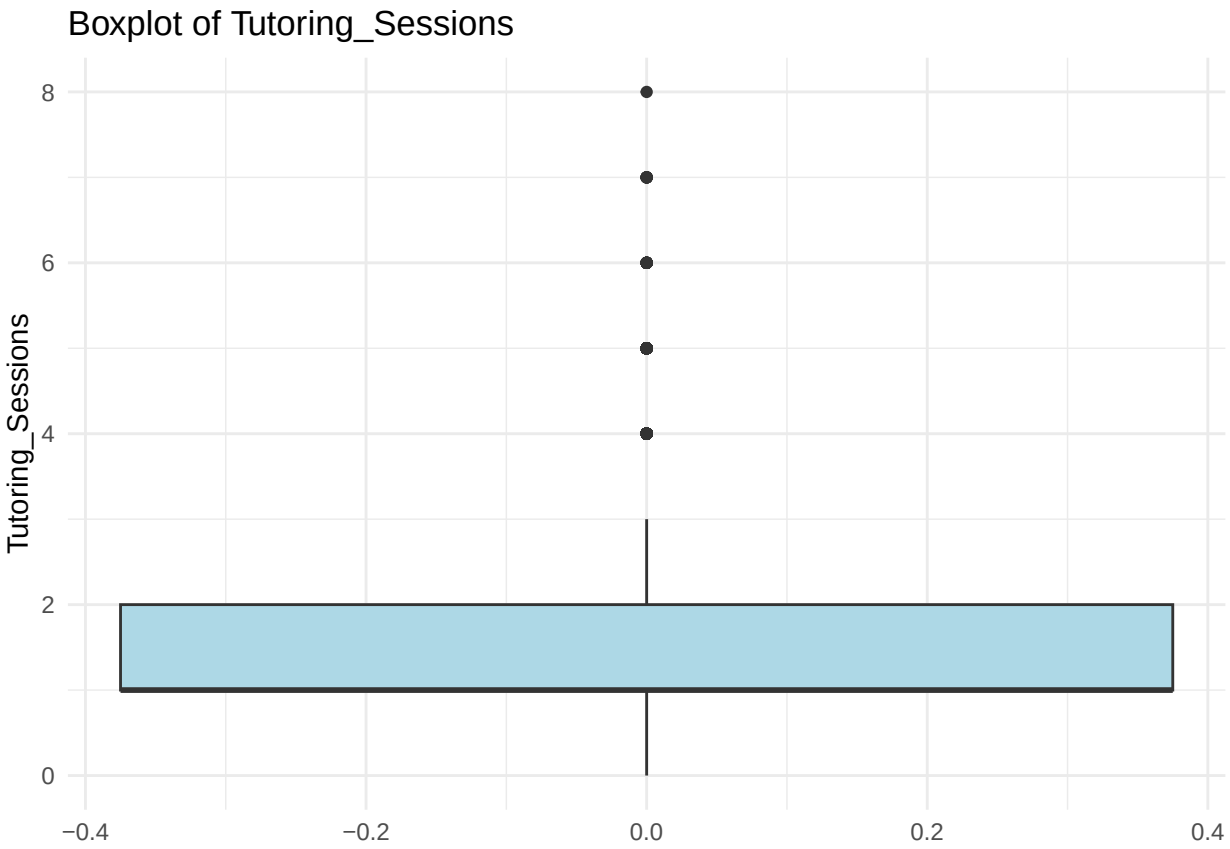
```
#plots showing the outliers
library(ggplot2)

# Variables with outliers
outlier_vars <- c("Hours_Studied", "Attendance", "Sleep_Hours", "Previous_Scores",
                  "Tutoring_Sessions", "Physical_Activity", "Exam_Score")

# Create boxplots for variables with outliers
for (var in outlier_vars) {
  p <- ggplot(project_data, aes(y = .data[[var]])) +
    geom_boxplot(fill = "lightblue") +
    ggtitle(paste("Boxplot of", var)) +
    theme_minimal() +
    ylab(var)
  print(p)
}
```

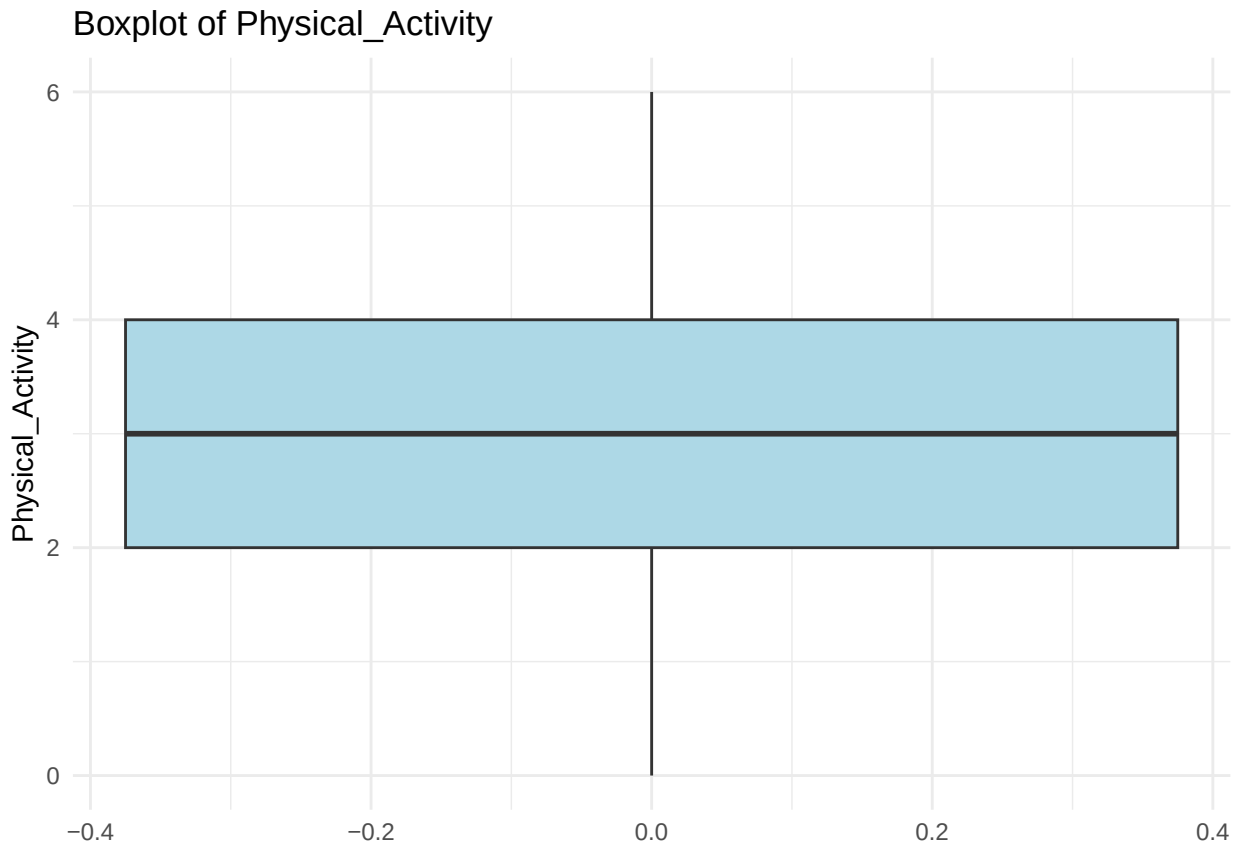


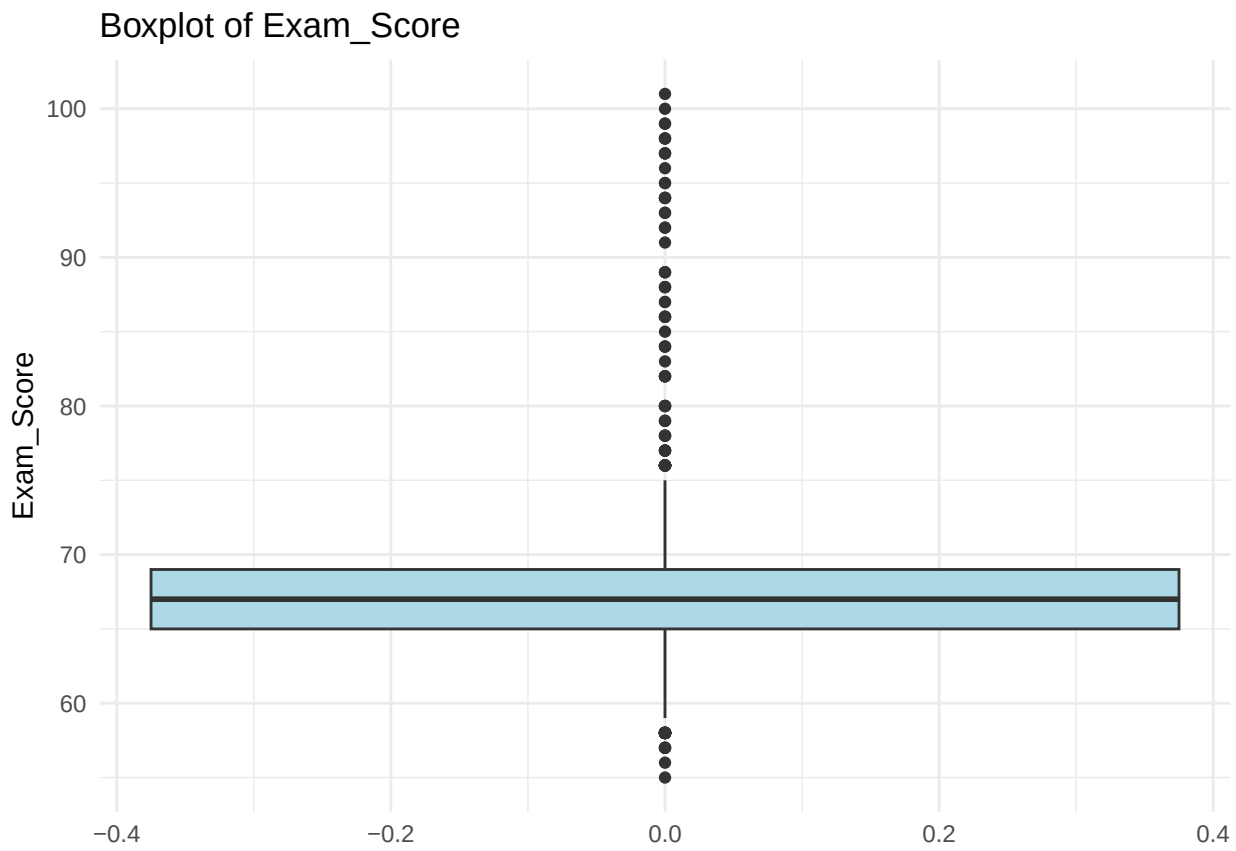

Boxplot of Attendance
Attendance
100
90
80
70
60
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Sleep_Hours
Sleep_Hours
10
8
6
4
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Previous_Scores
Previous_Scores
100
90
80
70
60
50
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Tutoring_Sessions
Tutoring_Sessions
8
6
4
2
0
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Physical_Activity
Physical_Activity
6
4
2
0
−0.4
−0.2
0.0
0.2
0.4

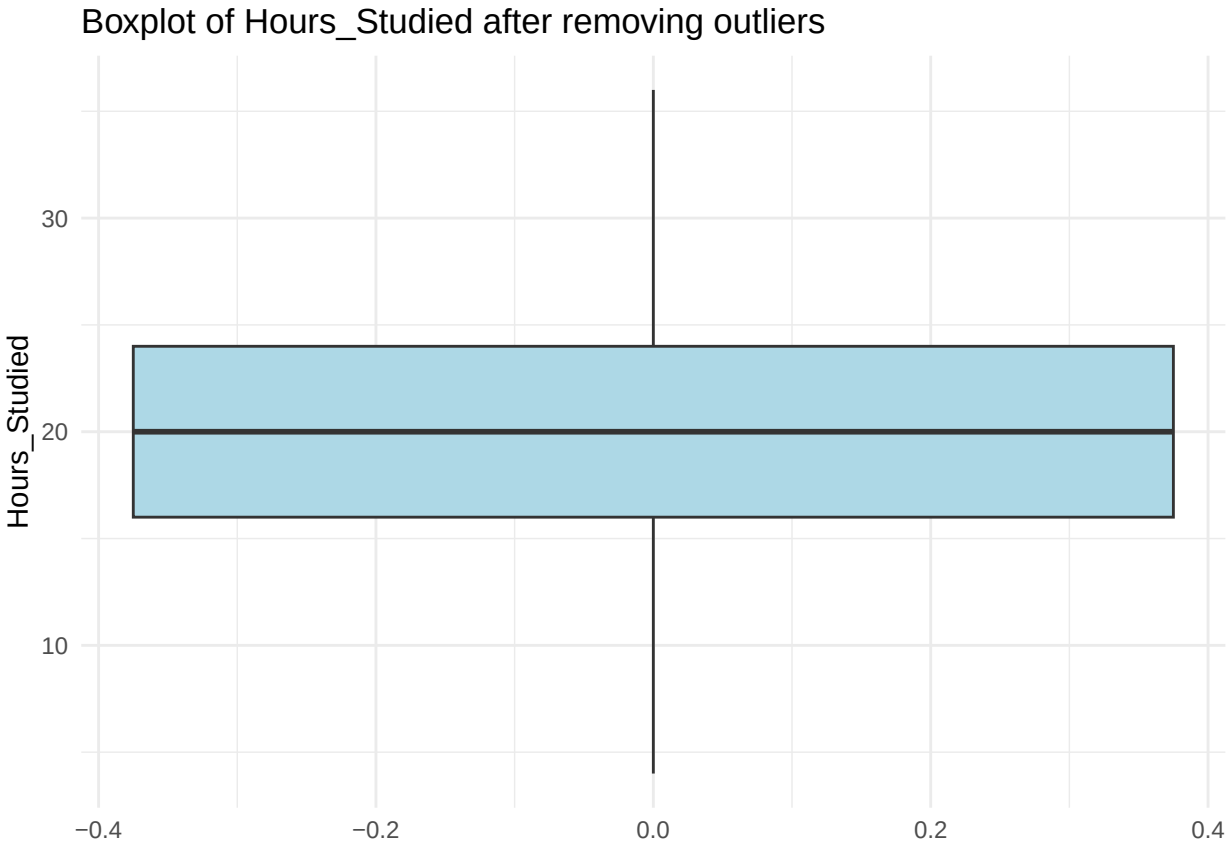


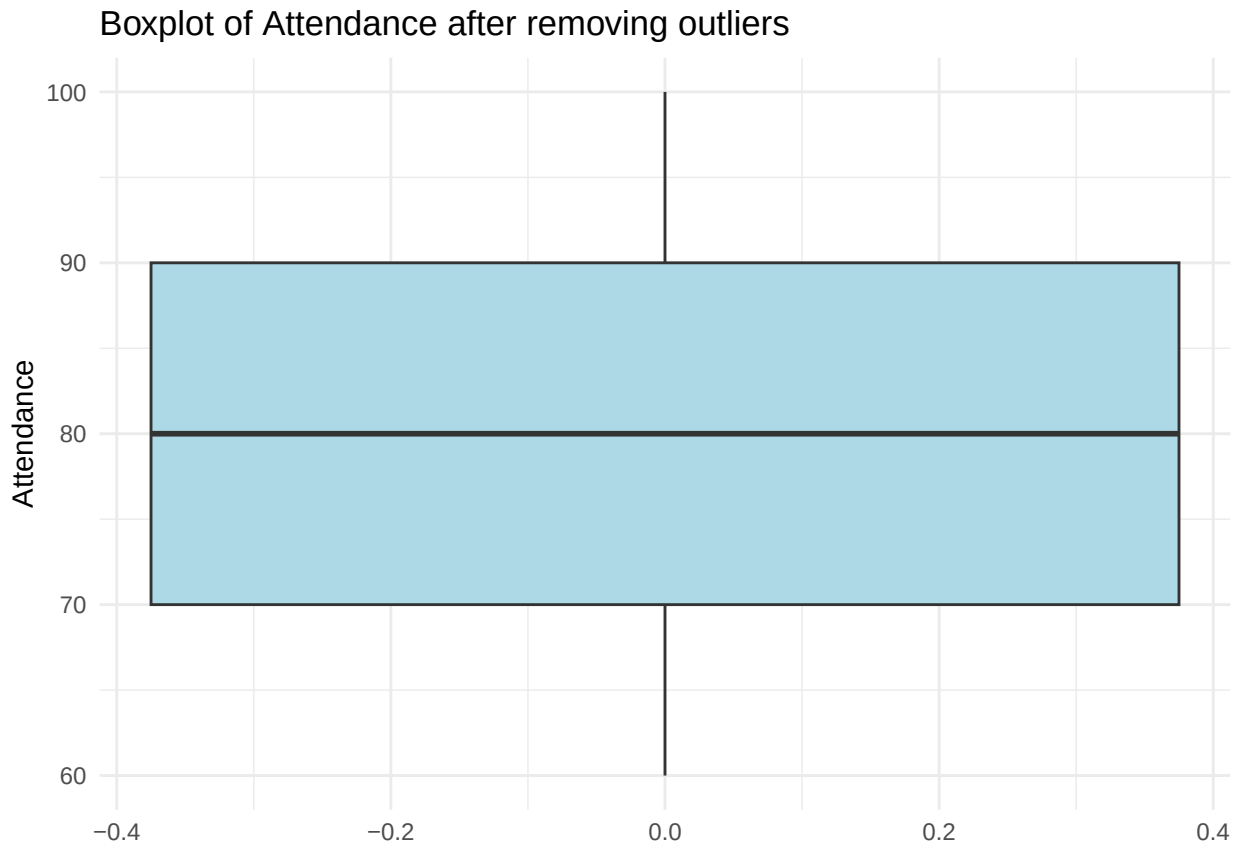
The Plots clearly shows us that the columns like Hours_Studied, Tutoring_Sessions, Exam_Score consists of outliers.These outliers says that a few students are studying far more (or less) than their peers, and similarly, a small number of students might have abnormal exam scores compared to the rest. These outliers could represent exceptional or struggling students.

Removing the outliers using IQR(Inter Quartile Range):

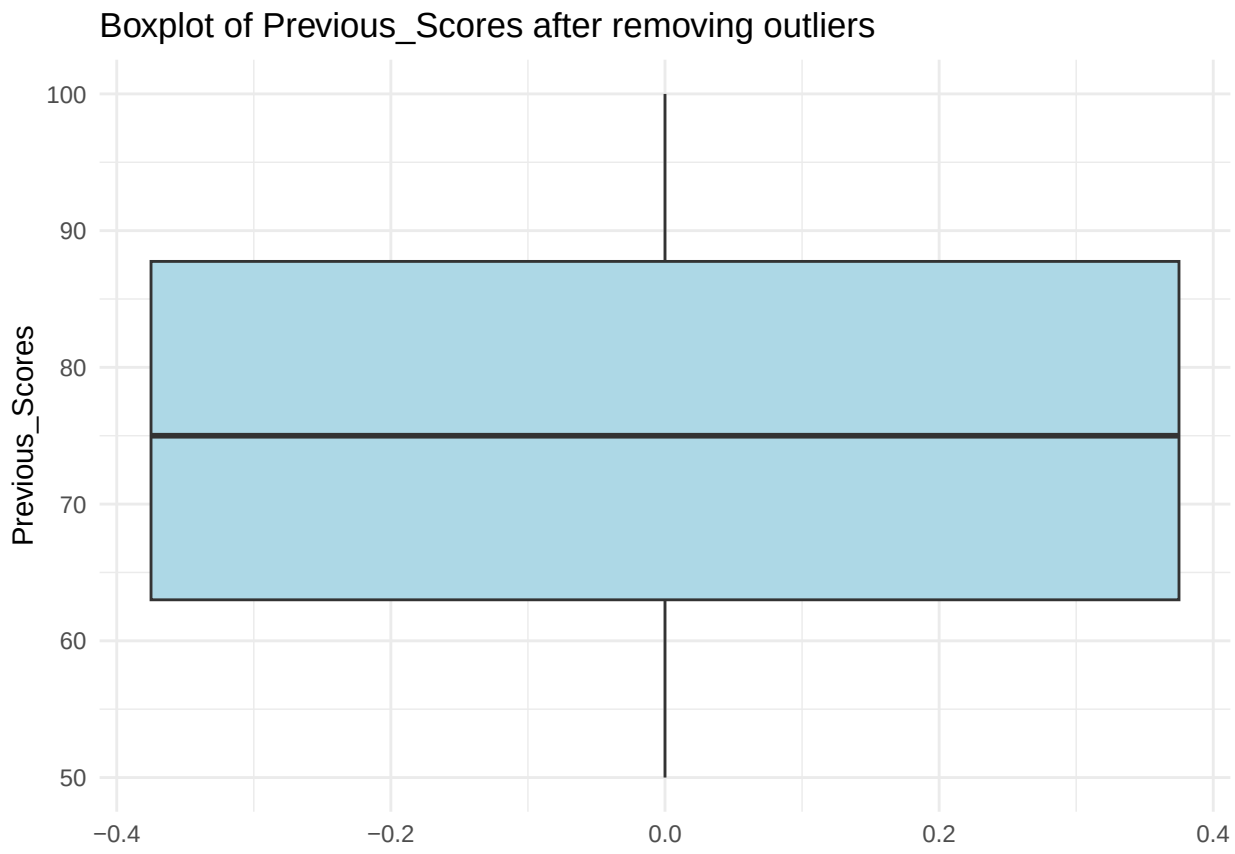
```
#using IQR (INTER QUARTILE RANGE) to remove the outliers
for (var in outlier_vars) {
  Q1 <- quantile(project_data[[var]], 0.25)
  Q3 <- quantile(project_data[[var]], 0.75)
  IQR <- Q3 - Q1
  project_data <- project_data[!(project_data[[var]] < (Q1 - 1.5 * IQR) | project_data[[var]] > (Q3 + 1
}
```

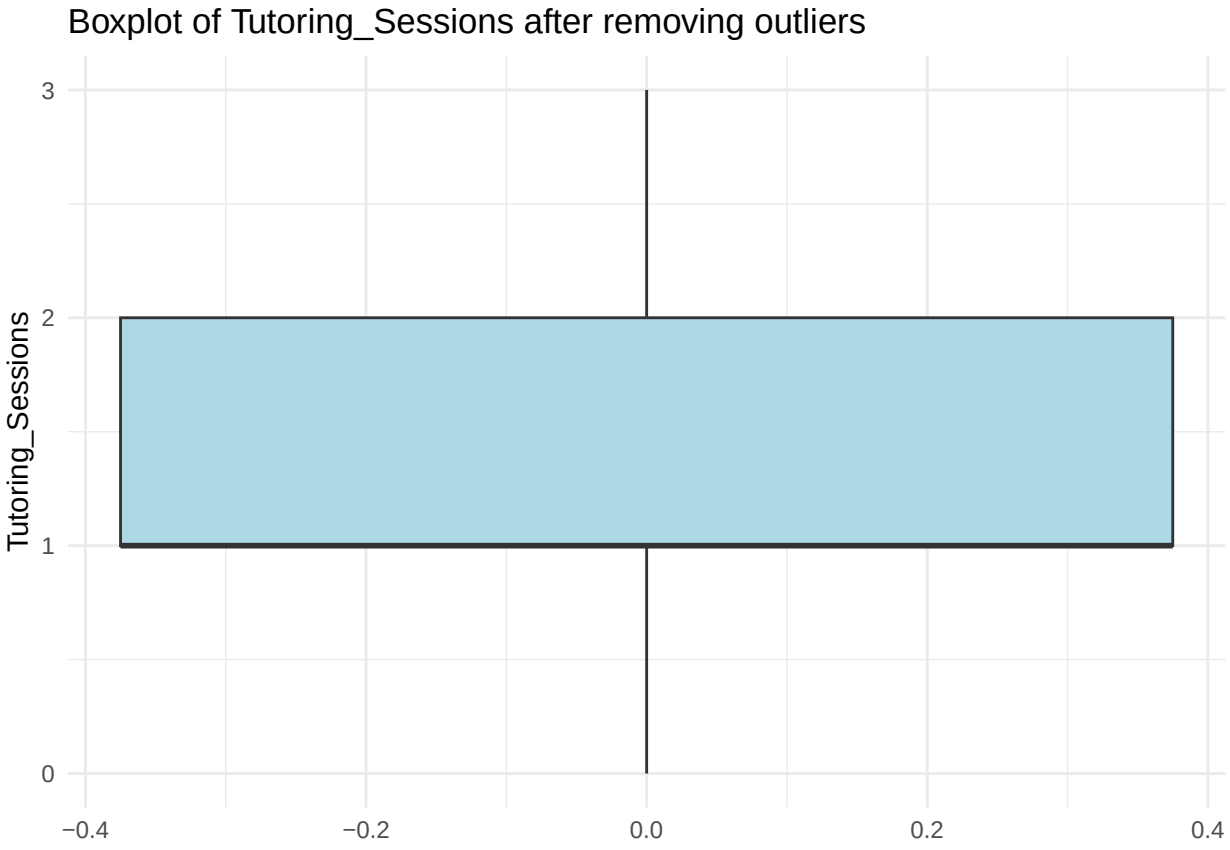
```
#plots to remove outliers
library(ggplot2)

# Create boxplots again for variables after outliers removal
for (var in outlier_vars) {
  p <- ggplot(project_data, aes(y = .data[[var]])) +
    geom_boxplot(fill = "lightblue") +
    ggtitle(paste("Boxplot of", var, "after removing outliers")) +
    theme_minimal() +
    ylab(var)
  print(p)
}
```

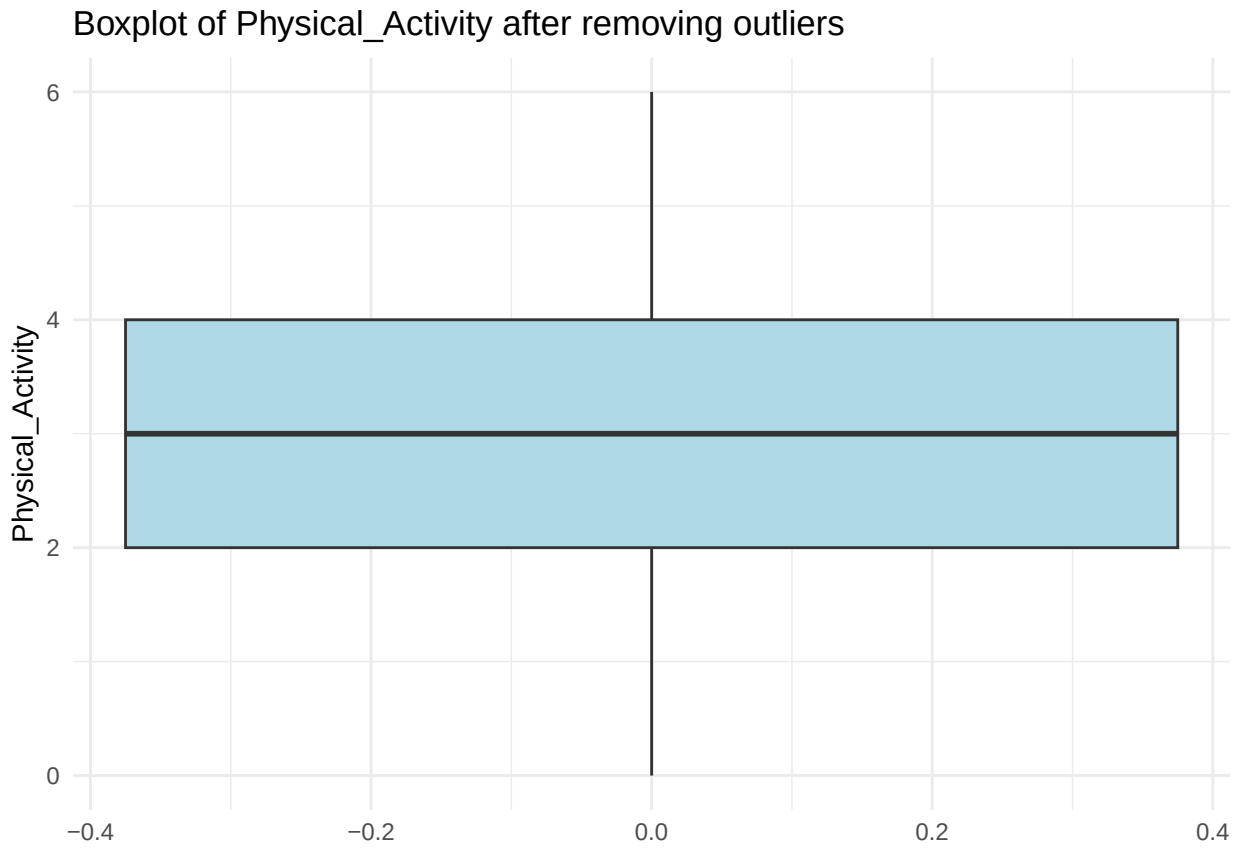

Boxplot of Hours_Studied after removing outliers
Hours_Studied
30
20
10
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Attendance after removing outliers
Attendance
100
90
80
70
60
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Sleep_Hours after removing outliers
Sleep_Hours
10
8
6
4
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Previous_Scores after removing outliers
Previous_Scores
100
90
80
70
60
50
−0.4
−0.2
0.0
0.2
0.4


Boxplot of Tutoring_Sessions after removing outliers
Tutoring_Sessions
3
2
1
0
−0.4
−0.2
0.0
0.2
0.4

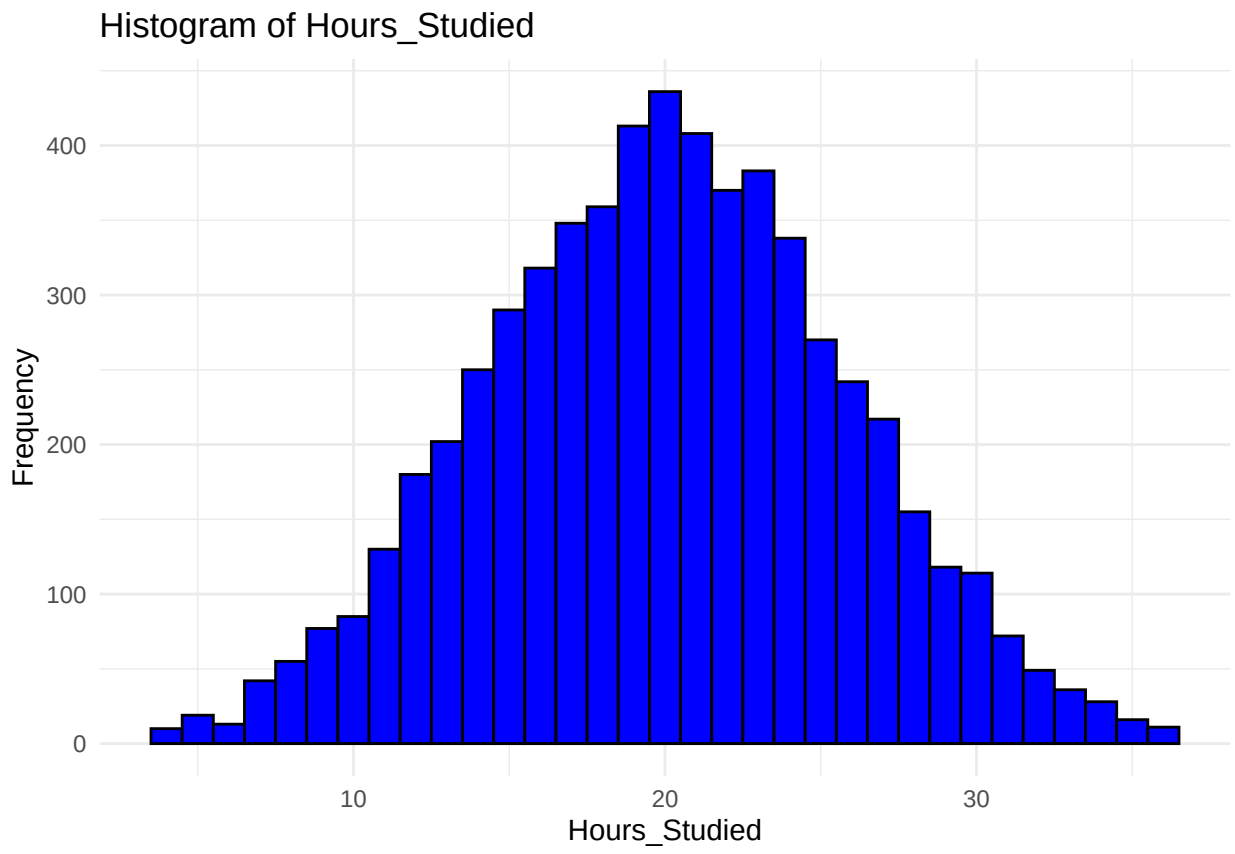

Boxplot of Physical_Activity after removing outliers
Physical_Activity
6
4
2
0
−0.4
−0.2
0.0
0.2
0.4

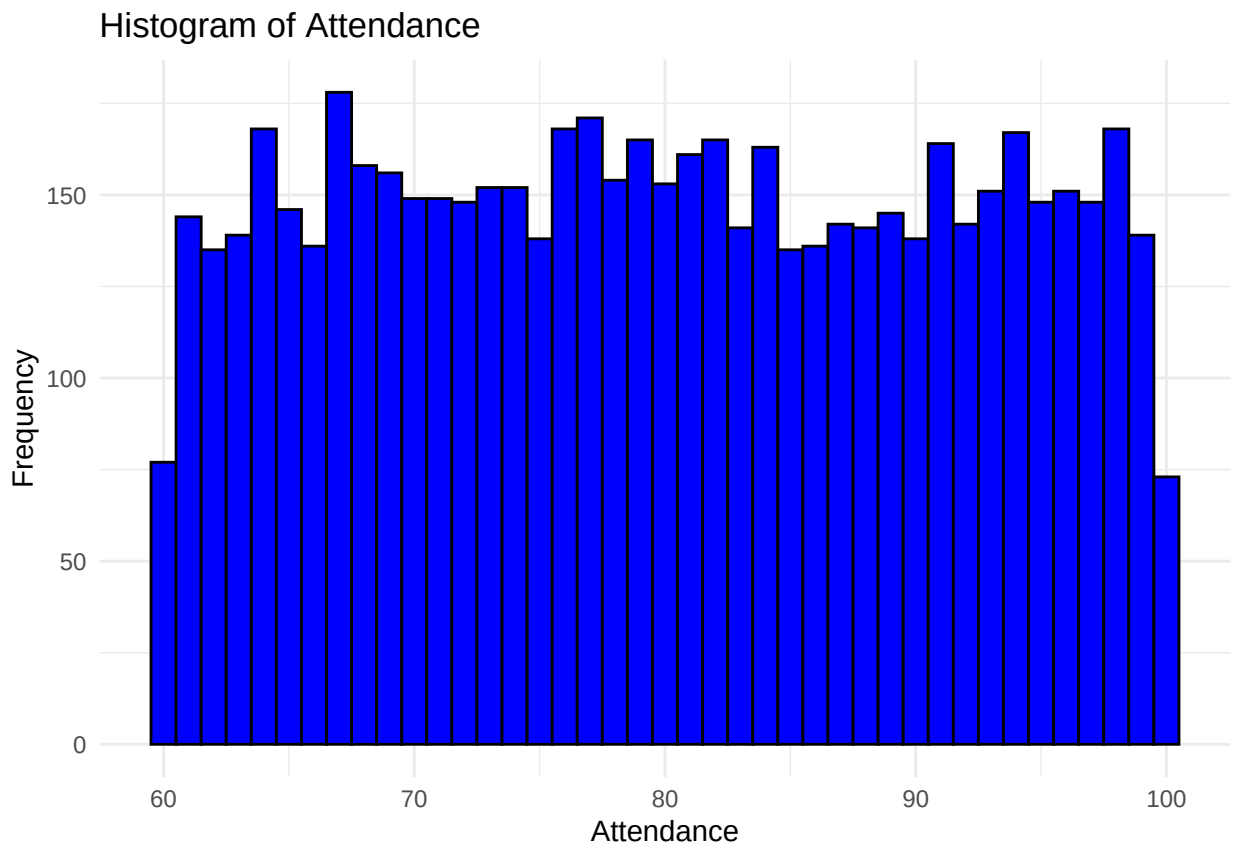
The outliers are removed and the data is cleaned which is quite ready to fit into a model.

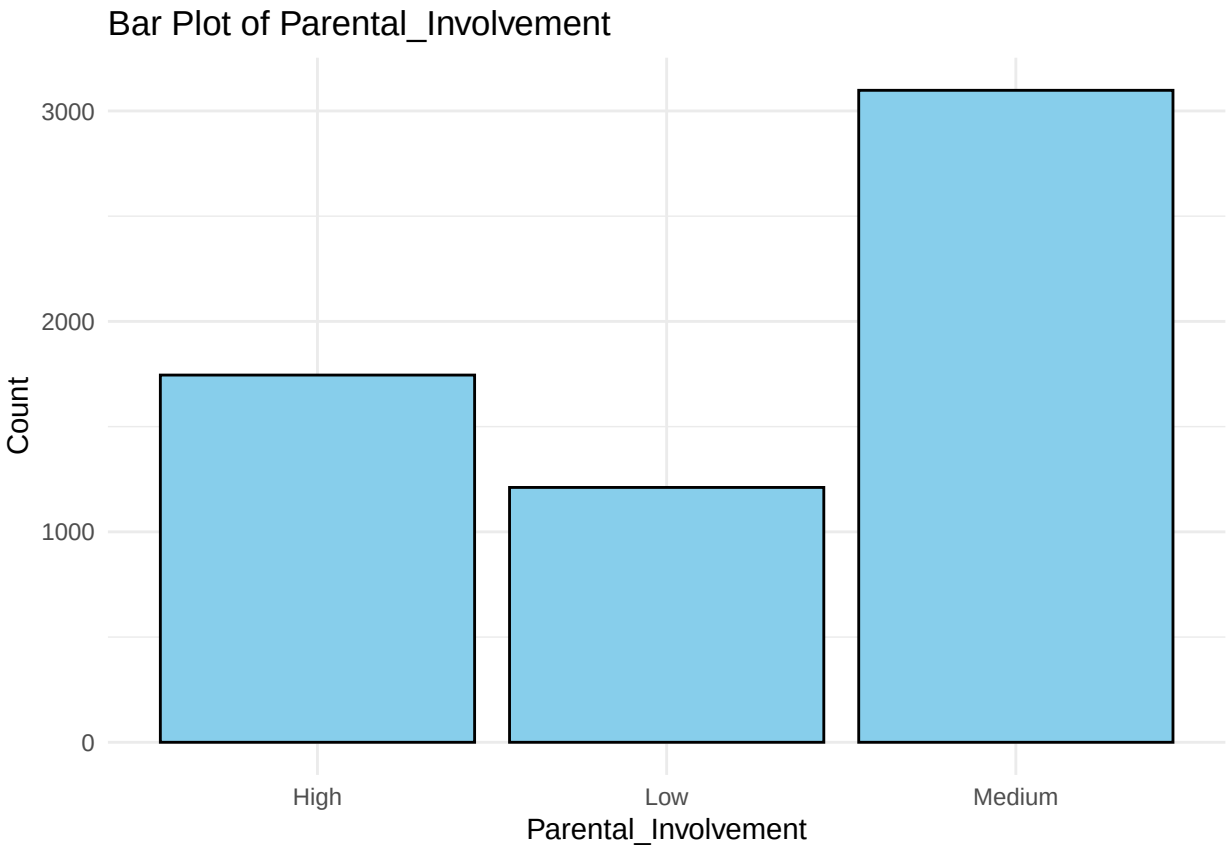
Description of the plots :

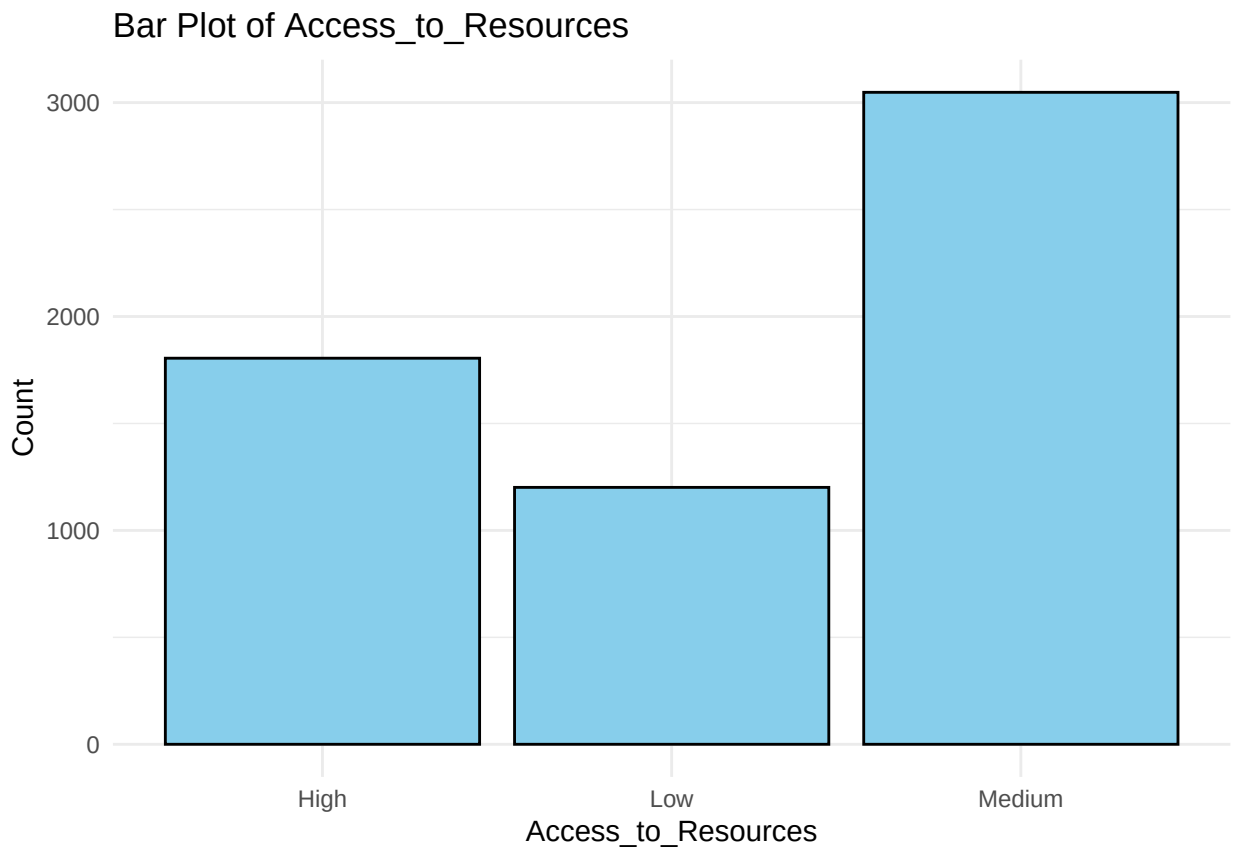
1. Histogram of Hours Studied: The histogram shows that most students studied between 10 and 30 hours, with a few students dedicating significantly more or fewer hours to studying, indicating a skew towards moderate study times.
2. Histogram of Attendance: The attendance histogram reveals that the majority of students have high attendance rates, with most attending between 80% and 100% of the classes, reflecting a generally engaged student body.
3. Histogram of Sleep Hours: The distribution of sleep hours shows that most students sleep between 6 and 8 hours per night, with few extreme values at either end, indicating a typical range of sleep habits.
4. Histogram of Exam Scores: The exam score histogram displays a relatively normal distribution, with most students scoring between 60 and 80, suggesting a concentration around the average exam performance.
5. Bar Plot of Parental Involvement: The bar plot demonstrates that most students have medium to high parental involvement, with fewer students reporting low involvement, highlighting the importance of family support in student performance.
6. Bar Plot of Extracurricular Activities: The bar plot shows that the majority of students engage in extracurricular activities, with a smaller proportion reporting no participation, indicating a positive correlation between involvement outside of class and academic life.
7. Boxplot of Exam Scores by Motivation Level: The boxplot illustrates that students with higher motivation levels tend to score higher on exams, with the lower motivation group showing a wider spread in exam scores.
8. Boxplot of Exam Scores by Internet Access: The boxplot highlights that students with access to the internet generally have higher exam scores, with more consistent results across this group compared to those without internet access.

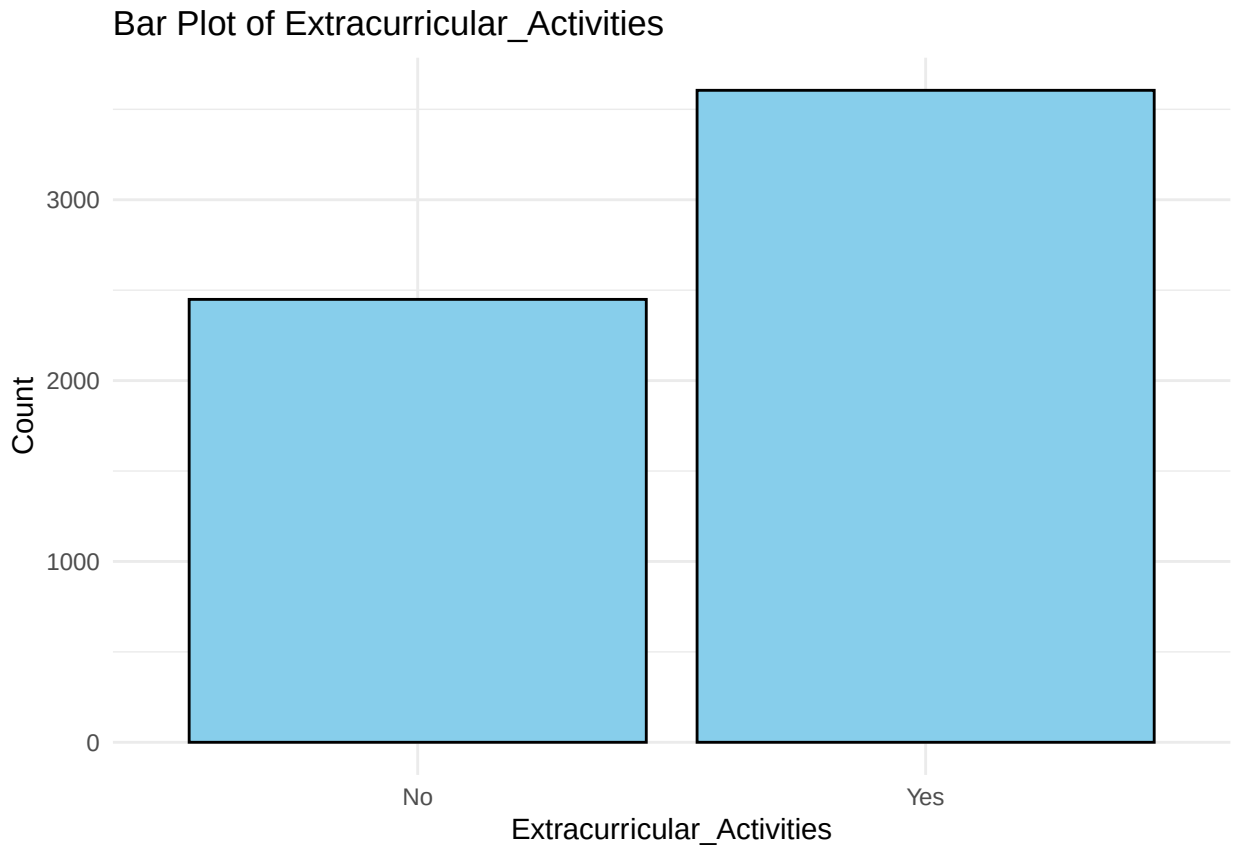
```r
#plots for the variables (numerical and categorical)
library(ggplot2)

# Loop through each variable and create plots
for (var in names(project_data)) {
  if (is.numeric(project_data[[var]])) {
    # Histogram for numeric variables
    p <- ggplot(project_data, aes(x = .data[[var]])) +
      geom_histogram(binwidth = 1, fill = "blue", color = "black") +
      ggtitle(paste("Histogram of", var)) +
      theme_minimal() +
      xlab(var) +
      ylab("Frequency")
  } else {
    # Bar plot for categorical variables
    p <- ggplot(project_data, aes(x = .data[[var]])) +
      geom_bar(fill = "skyblue", color = "black") +
      ggtitle(paste("Bar Plot of", var)) +
      theme_minimal() +
      xlab(var) +
      ylab("Count")
  }
  print(p)
}
```

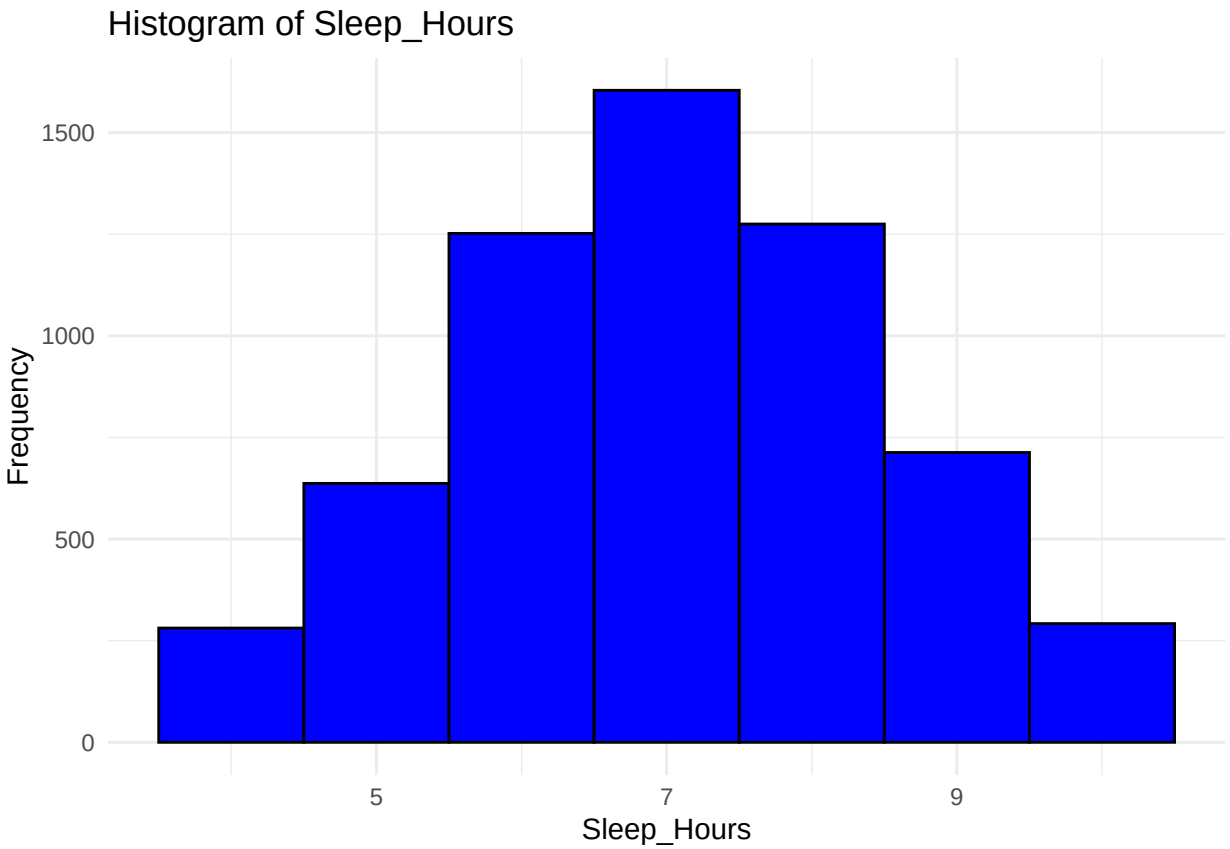


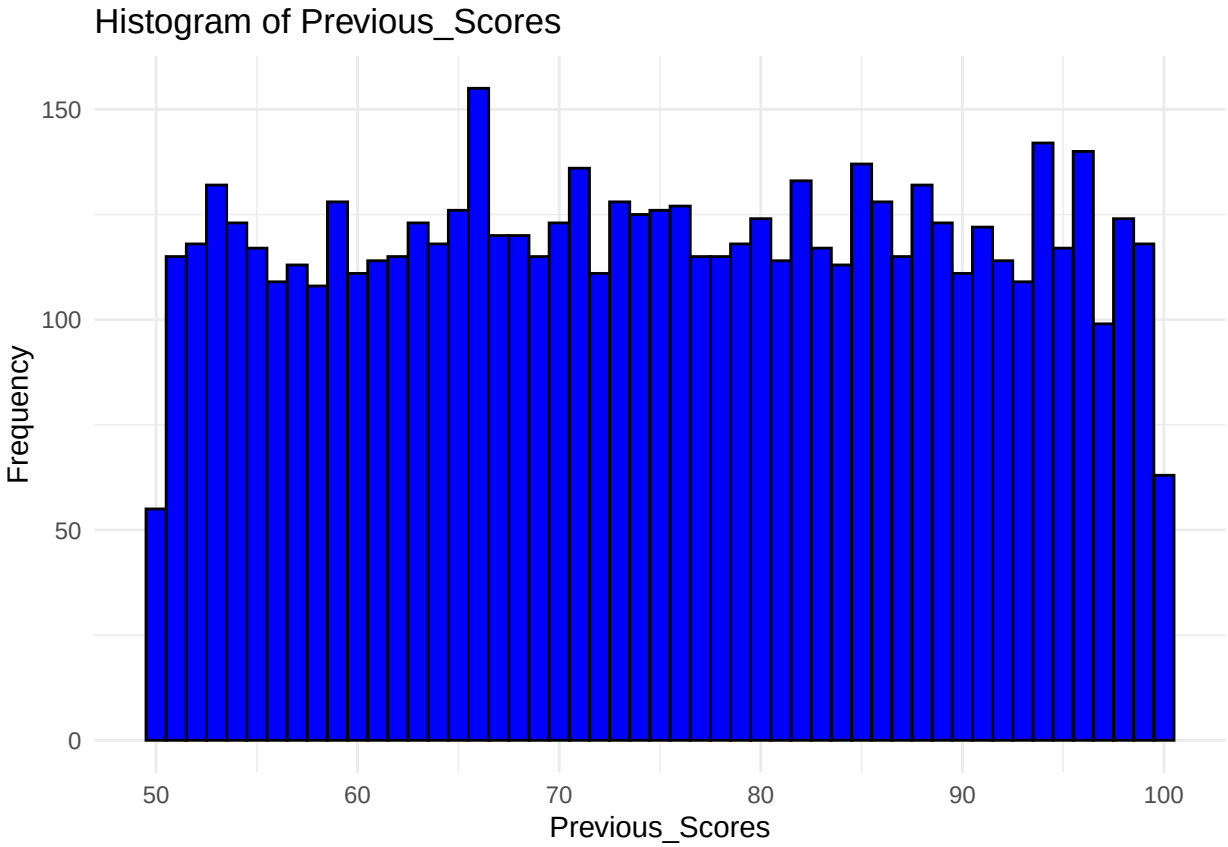

Histogram of Attendance
Frequency
150
100
50
0
60
70
80
90
100
Attendance

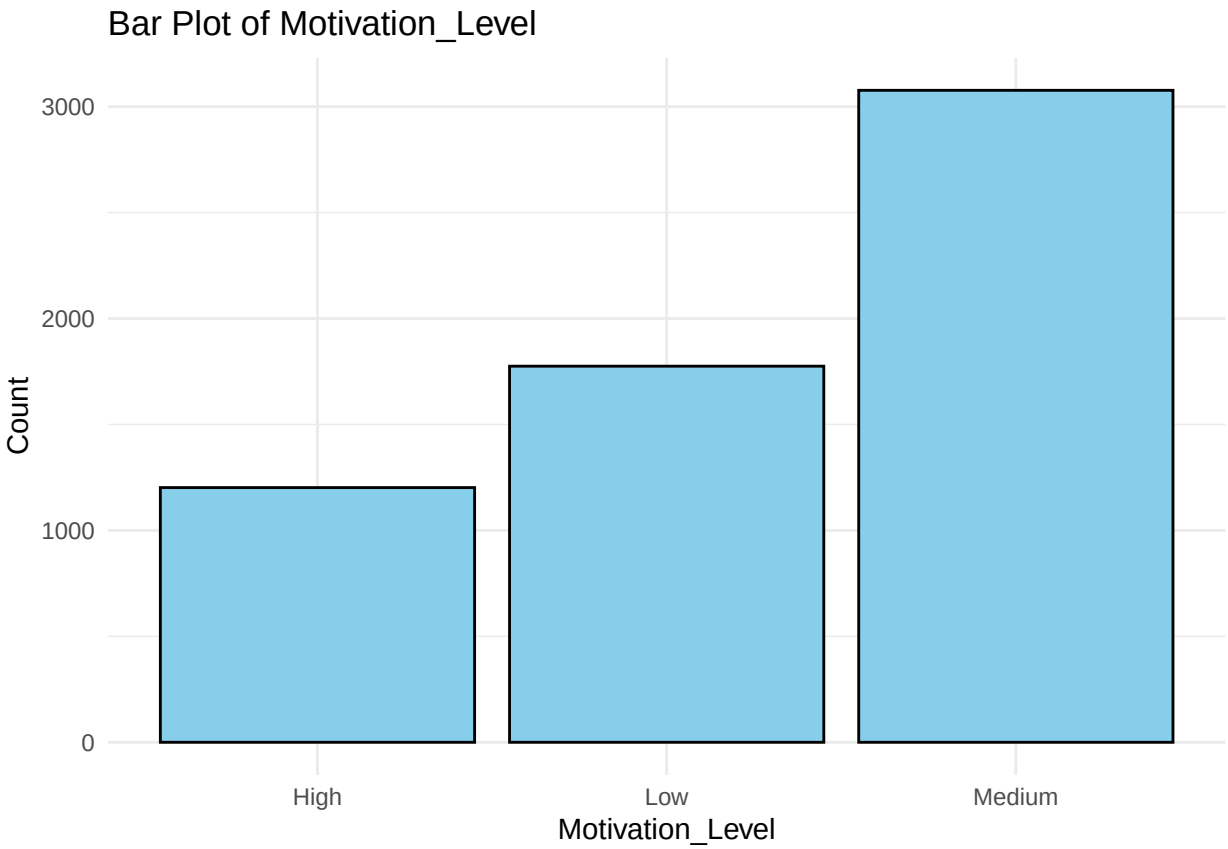

Bar Plot of Parental_Involvement
3000
2000
1000
0
Count
High
Low
Medium
Parental_Involvement

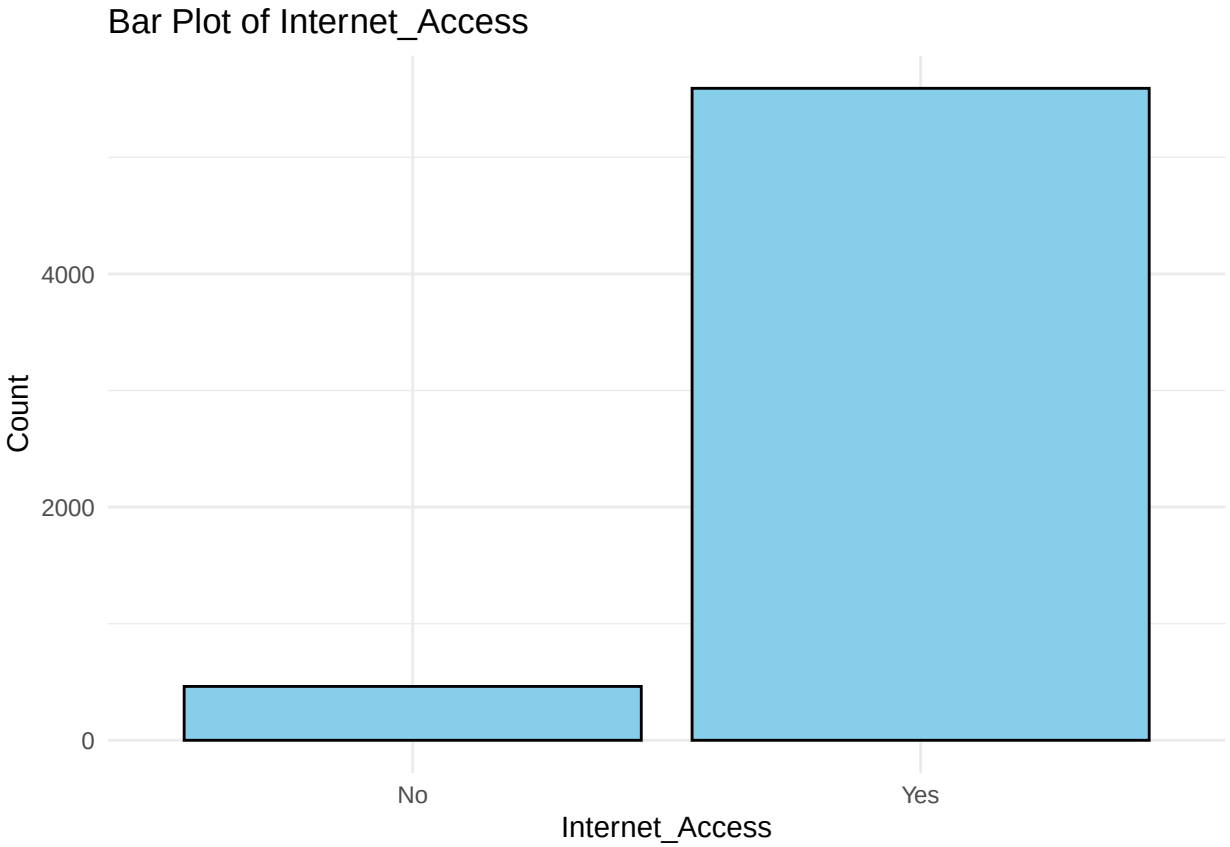

Bar Plot of Access_to_Resources
3000
2000
1000
0
Count
High
Low
Medium
Access_to_Resources

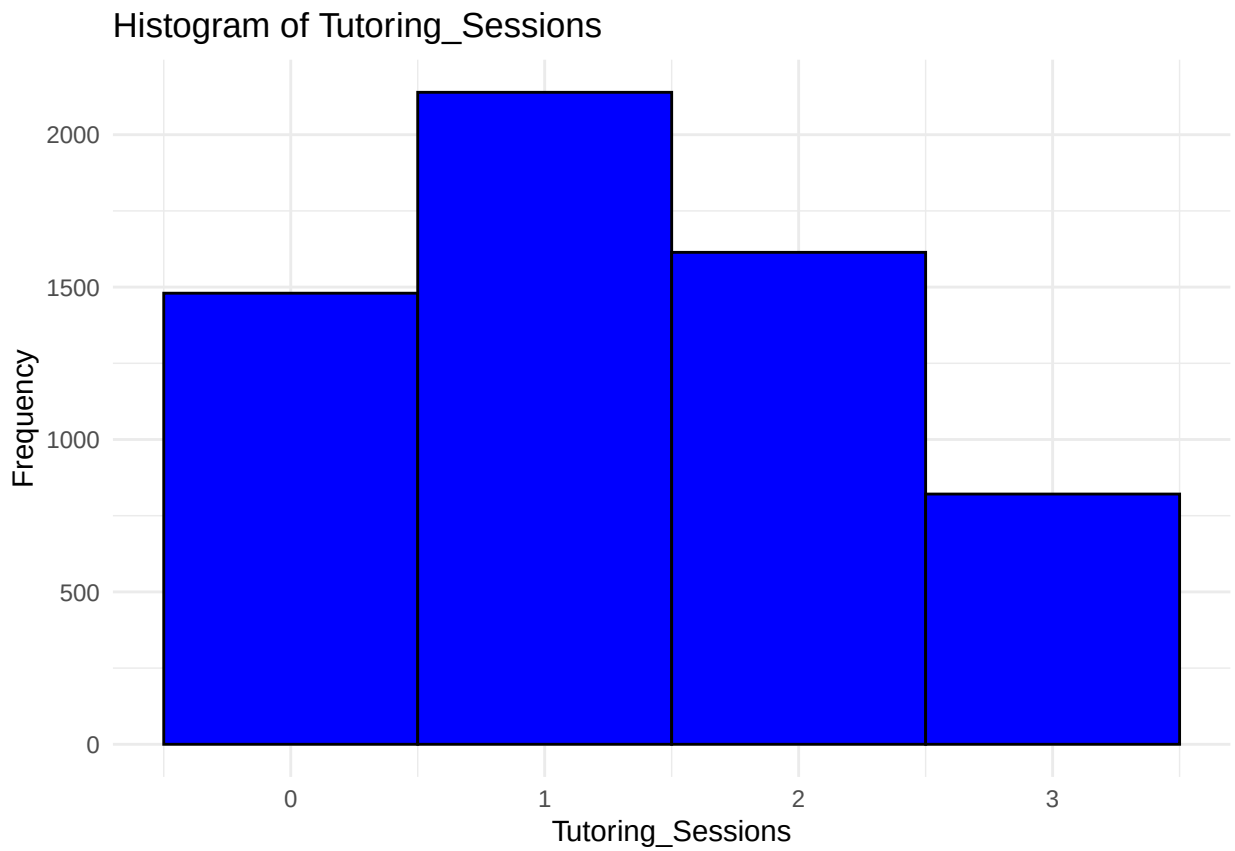

Bar Plot of Extracurricular_Activities
Count
3000
2000
1000
0
No
Yes
Extracurricular_Activities

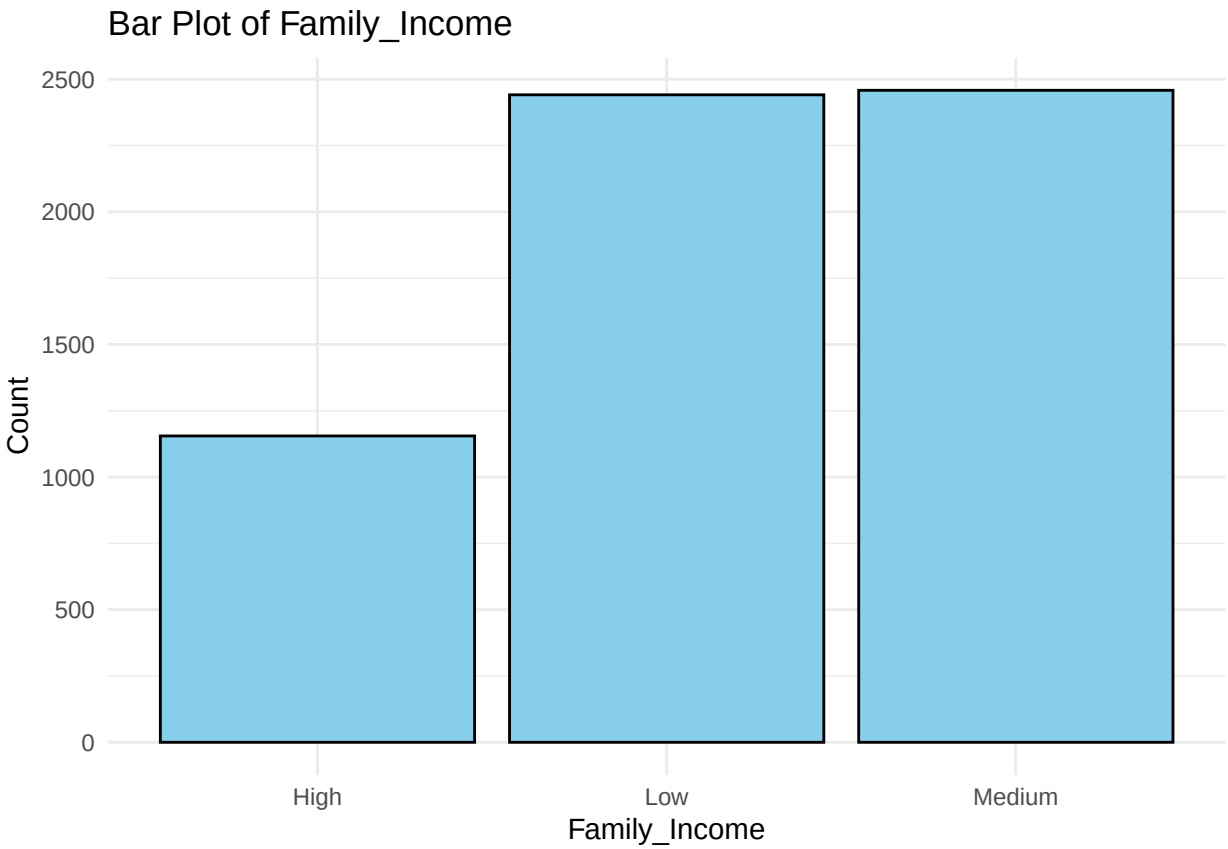

Histogram of Sleep_Hours
Frequency
1500
1000
500
0
5
7
9
Sleep_Hours

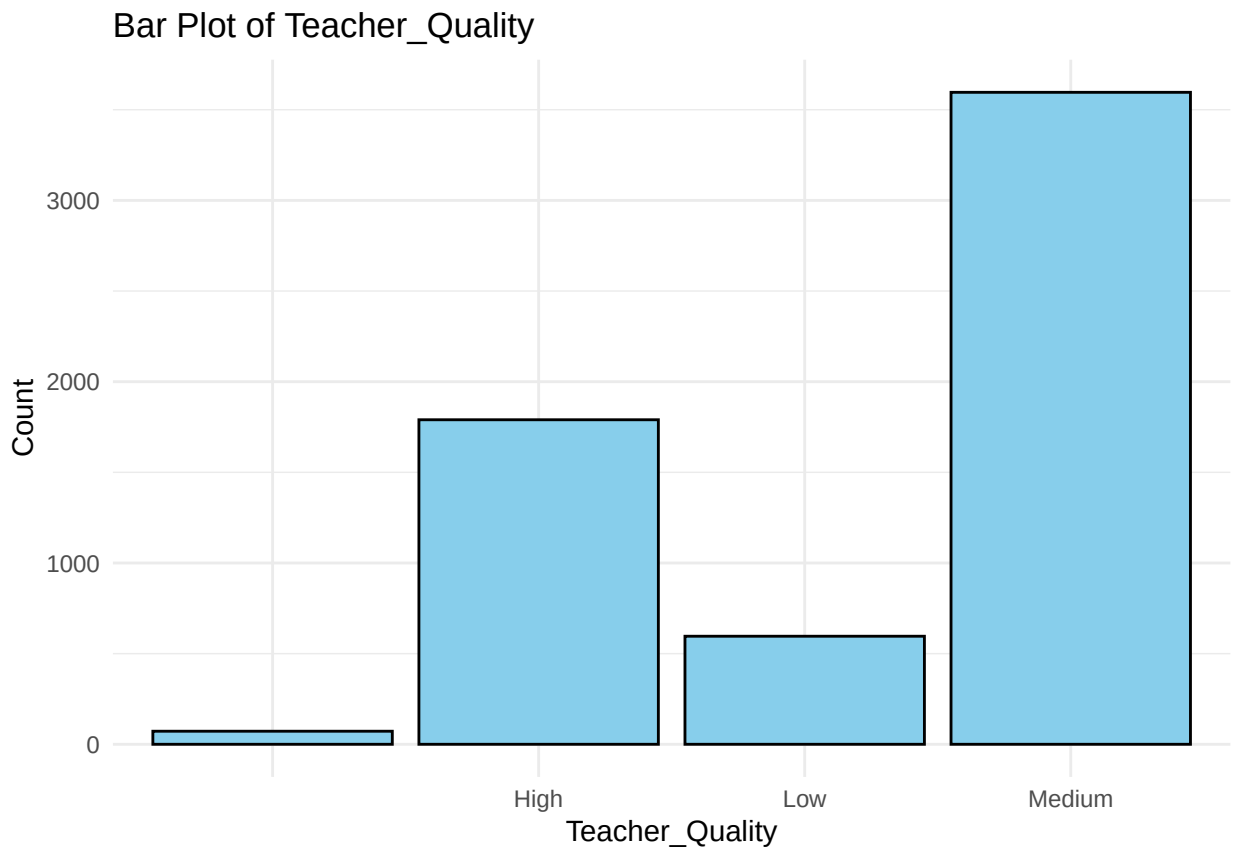

Histogram of Previous_Scores
Frequency
150
100
50
0
50
60
70
80
90
100
Previous_Scores

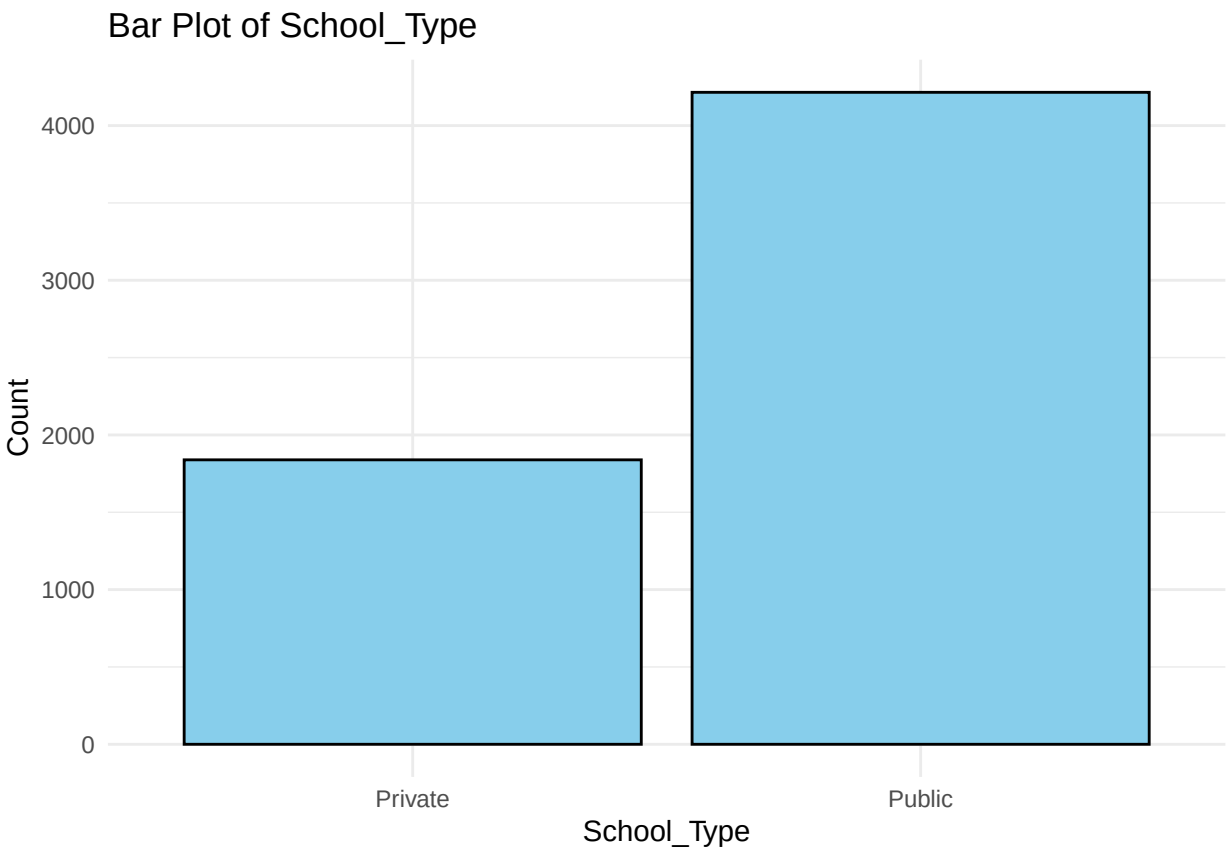

Bar Plot of Motivation_Level
3000
2000
1000
0
Count
High
Low
Medium
Motivation_Level

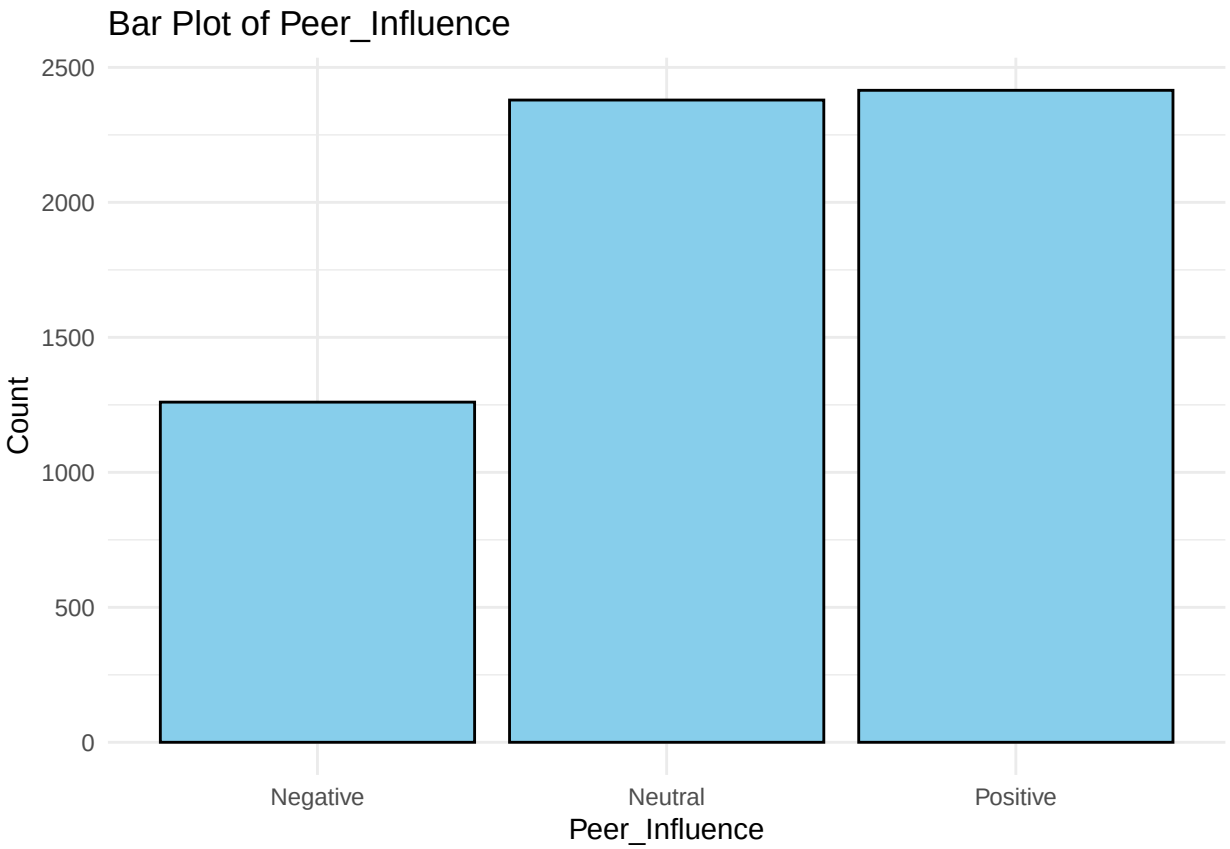

Bar Plot of Internet_Access
Count
4000
2000
0
No
Yes
Internet_Access

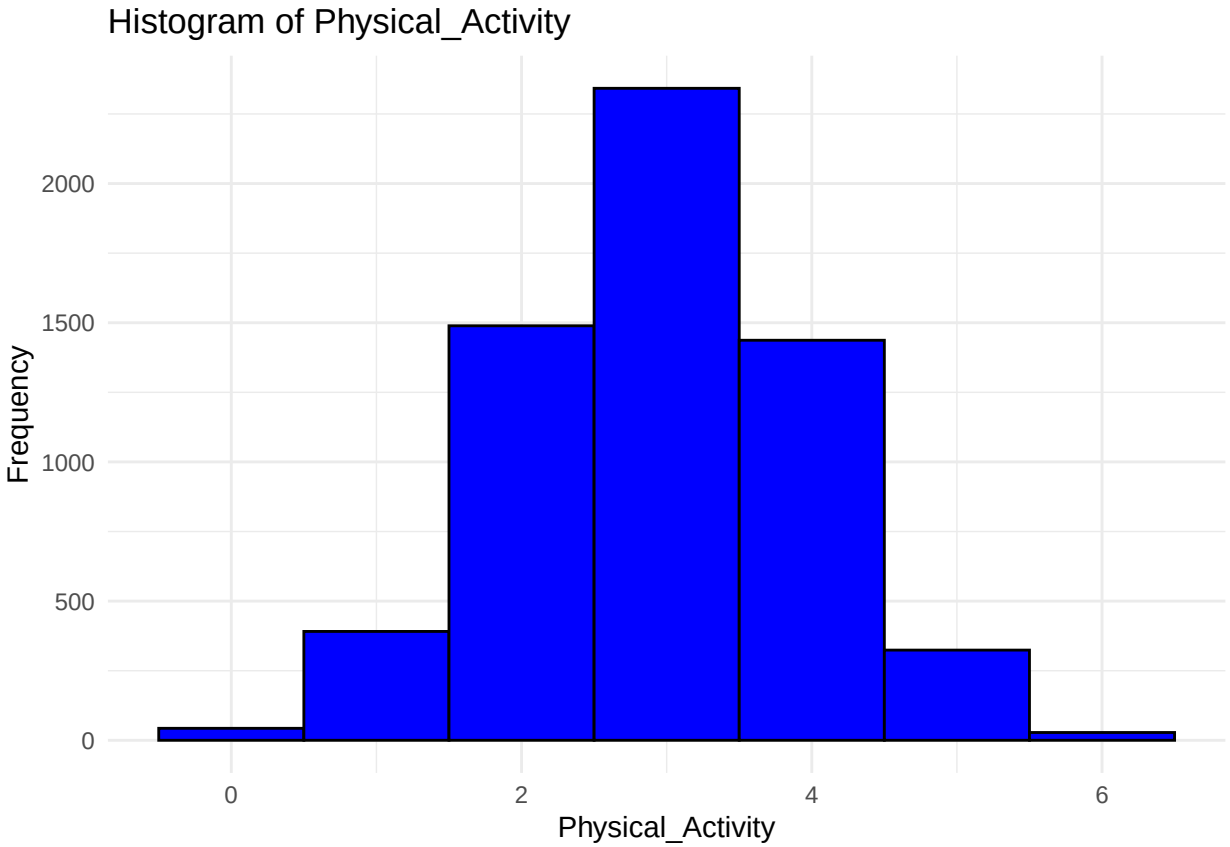

Histogram of Tutoring_Sessions
Frequency
2000
1500
1000
500
0
0
1
2
3
Tutoring_Sessions

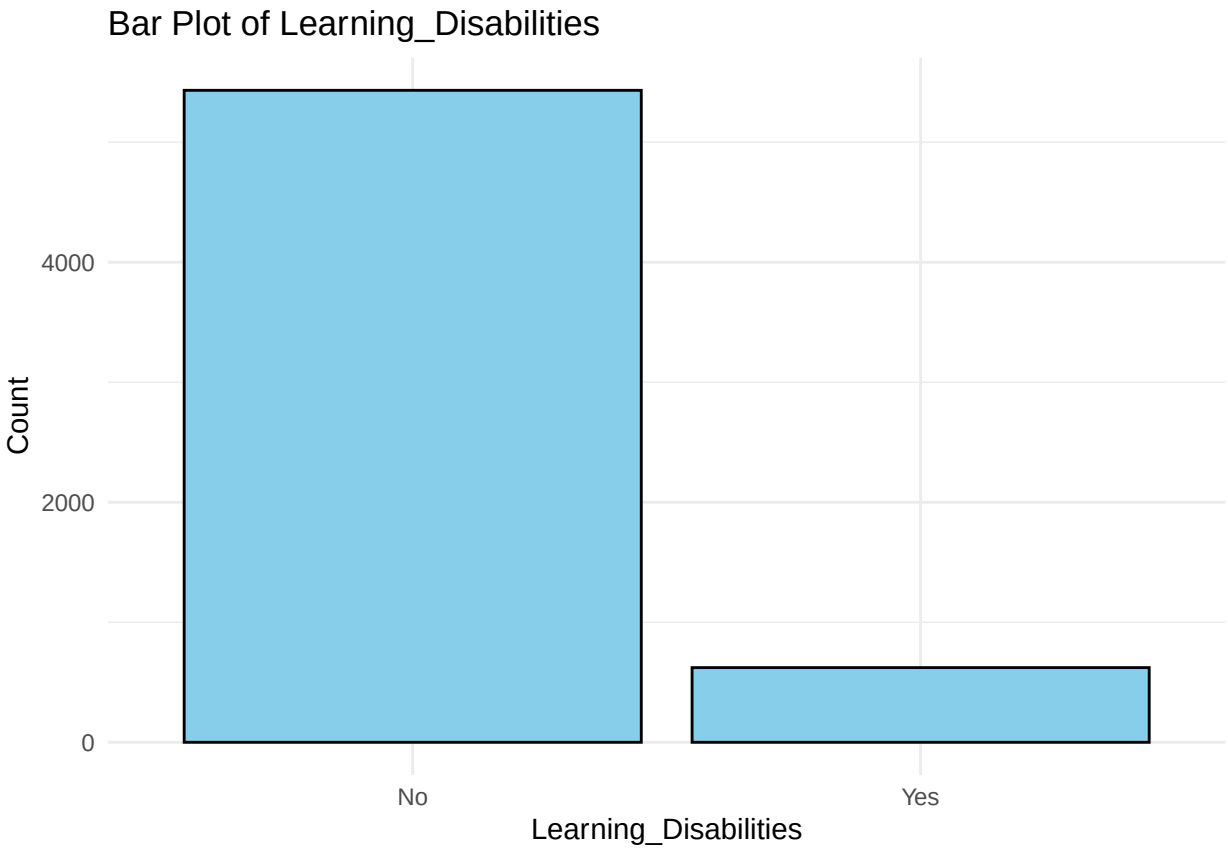

Bar Plot of Family_Income
2500
2000
1500
1000
500
0
Count
High
Low
Medium
Family_Income

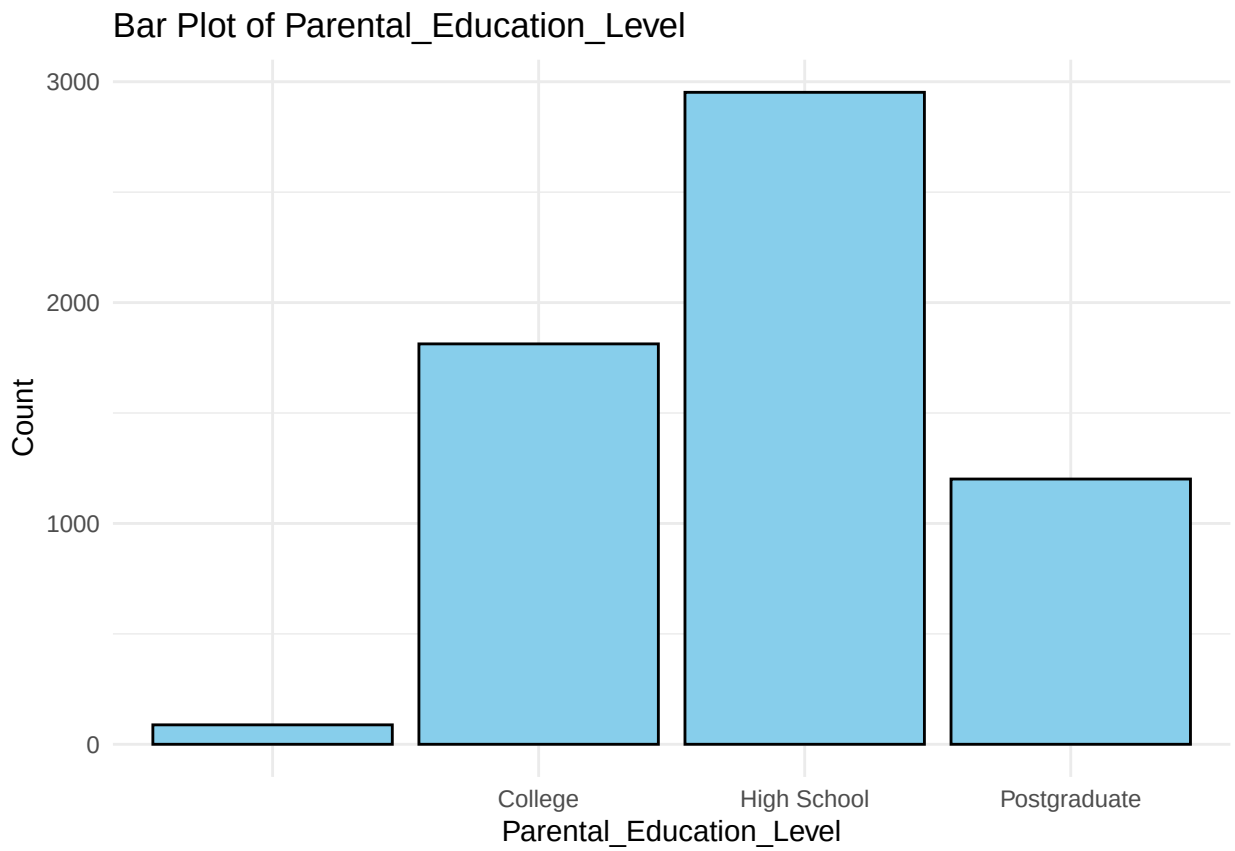

Bar Plot of Teacher_Quality
Count
3000
2000
1000
0
High
Low
Medium
Teacher_Quality

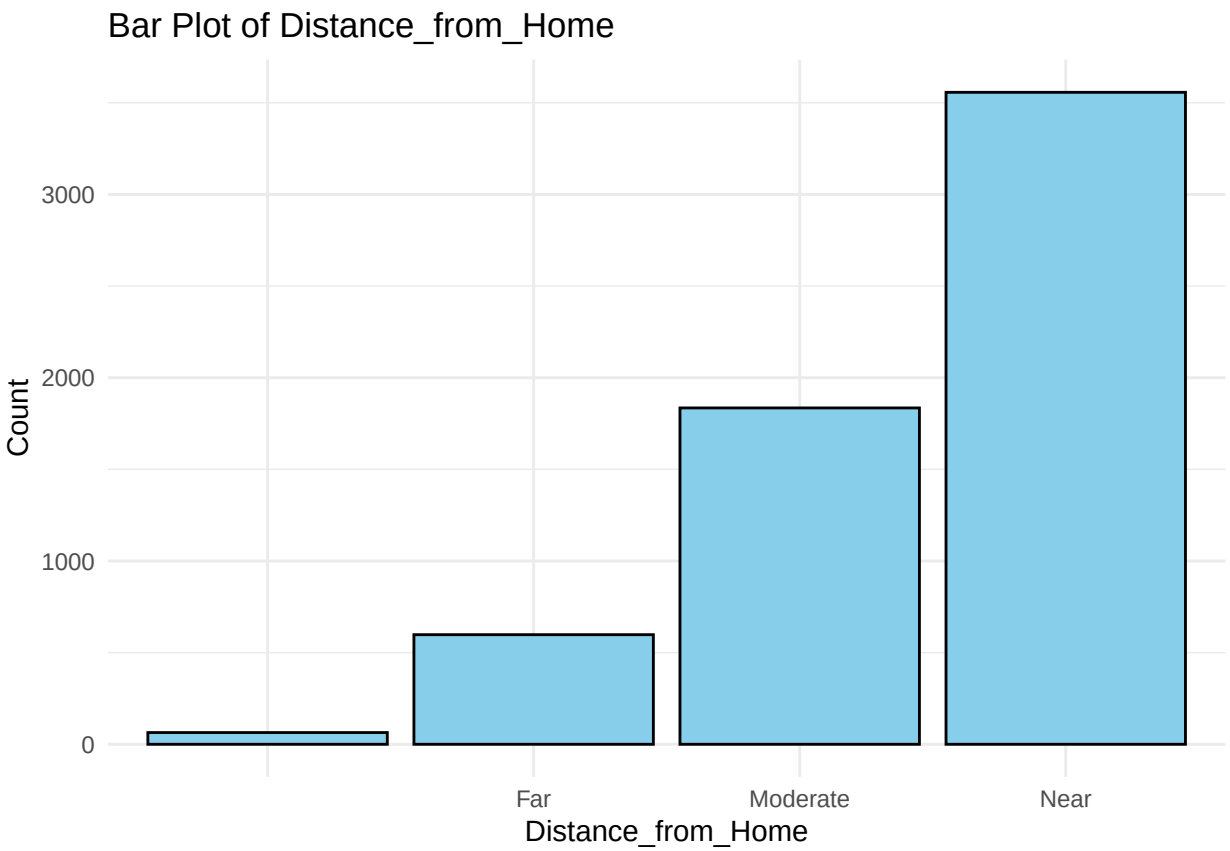

Bar Plot of School_Type
4000
3000
2000
1000
0
Count
Private
Public
School_Type

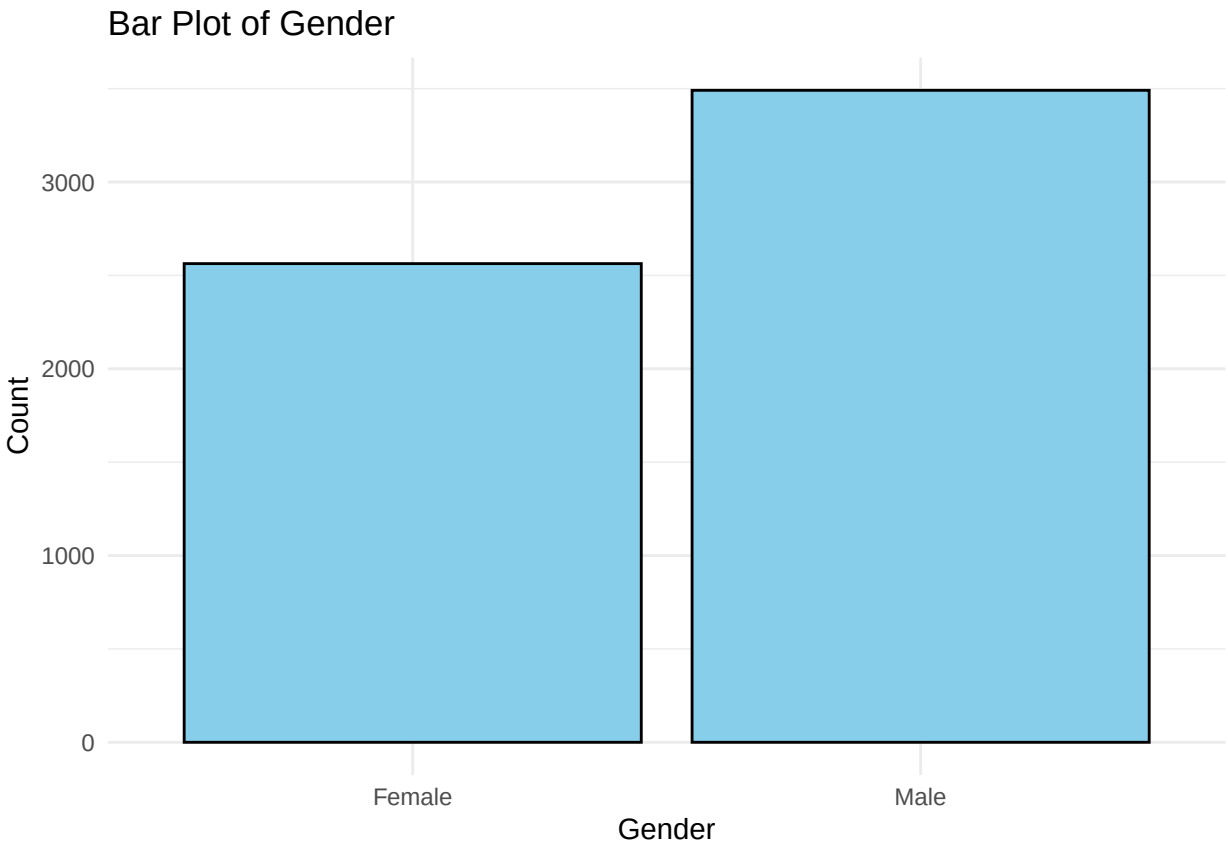

Bar Plot of Peer_Influence
2500
2000
1500
1000
500
0
Count
Negative
Neutral
Positive
Peer_Influence

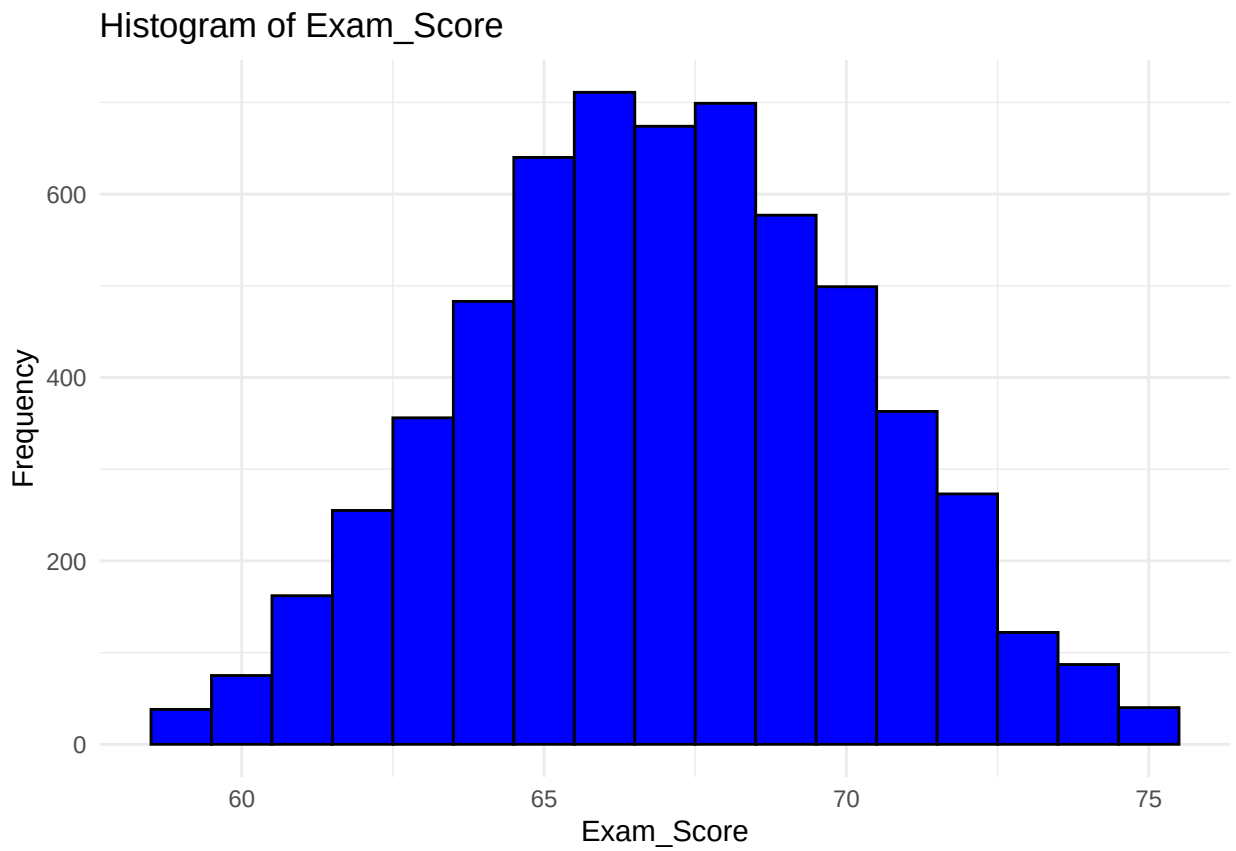

Histogram of Physical_Activity
Frequency
2000
1500
1000
500
0
0
2
4
6
Physical_Activity


Bar Plot of Learning_Disabilities
Count
4000
2000
0
No
Yes
Learning_Disabilities


Bar Plot of Parental_Education_Level
3000
2000
1000
0
Count
College
High School
Postgraduate
Parental_Education_Level


Bar Plot of Distance_from_Home
Count
3000
2000
1000
0
Far
Moderate
Near
Distance_from_Home


Bar Plot of Gender
3000
2000
1000
0
Count
Female
Male
Gender



Histograms for Numeric Variables: Hours Studied, Attendance, Sleep Hours, Previous Scores, Exam Scores. The histograms for these variables show a right-skewed distribution, meaning most students fall into lower-to-mid ranges for hours studied, attendance, and exam scores. We found that this is typical for educational datasets, where most students show average to below-average performance, with fewer students at the extreme high end. This suggests that while many students study for a moderate number of hours, there are a few who might be over-preparing or under-preparing for exams.

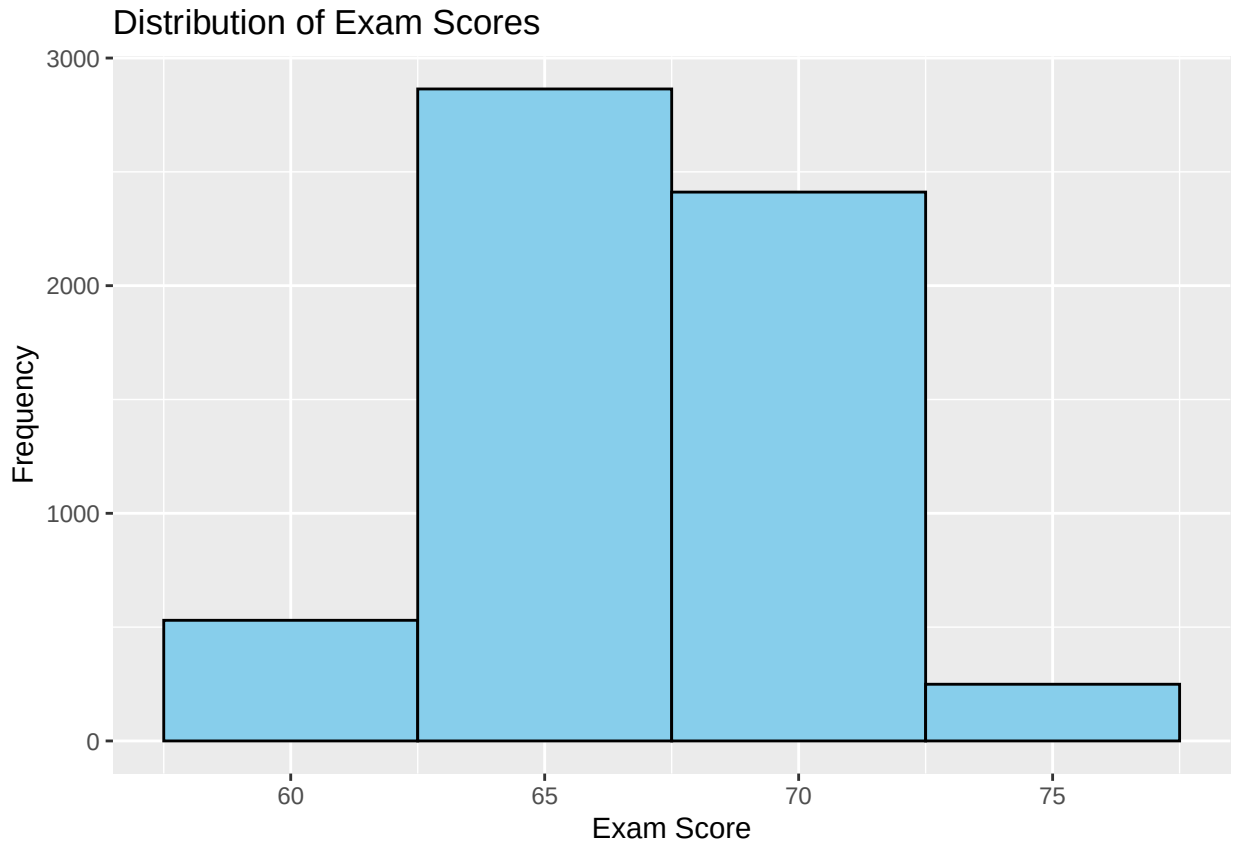
Bar Plots for Categorical Variables: Parental Involvement, Access to Resources, Motivation Level, Family Income, School Type, Teacher Quality: These variables are mostly categorical, showing high counts in medium and high categories for involvement, resources, and motivation. Income is skewed towards medium, while teacher quality has a balanced distribution. Most students have high parental involvement, access to resources, and motivation. The high counts in the "medium" categories for family income and teacher quality suggest that these factors are central to the student population.

```
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Univariate Analysis

```
# Histogram of Exam_Score
ggplot(project_data, aes(x = Exam_Score)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Exam Scores", x = "Exam Score", y = "Frequency")
```



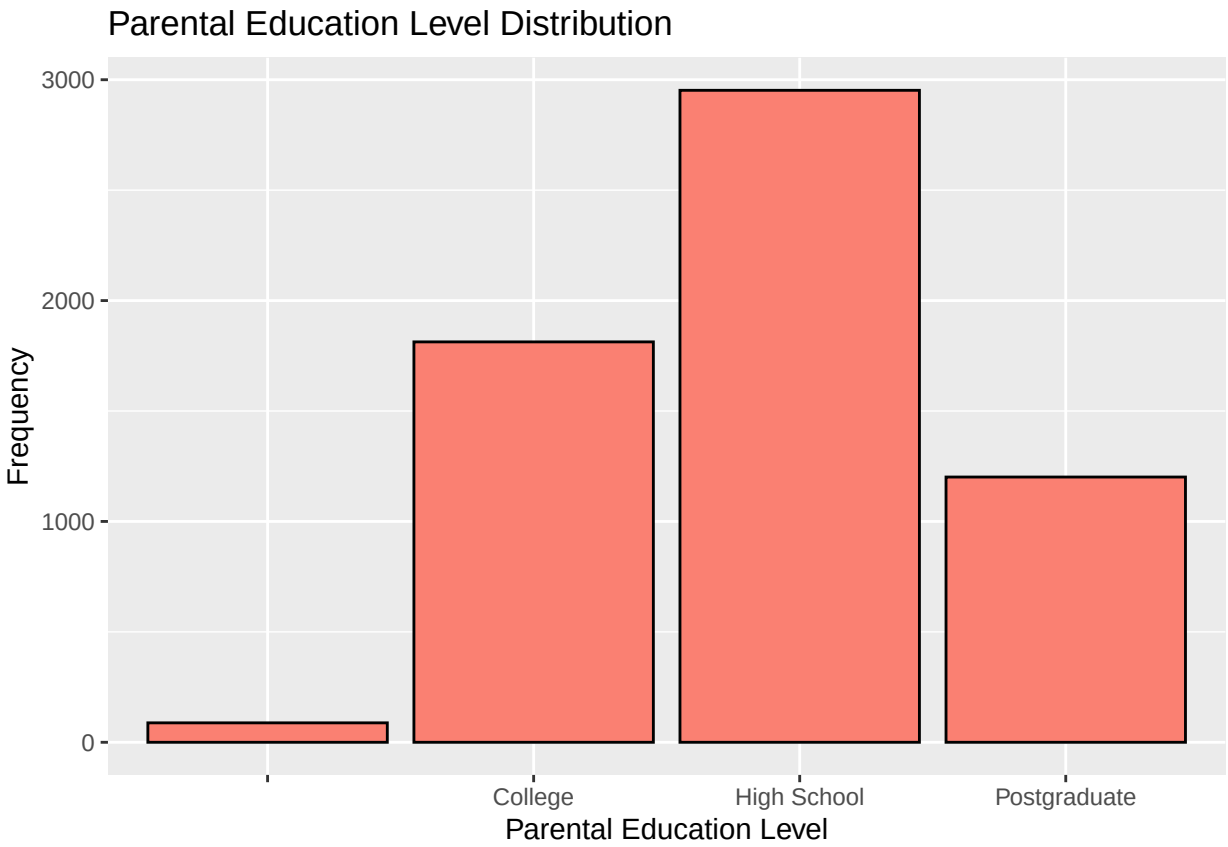
.This histogram shows the distribution of exam scores in the dataset. The majority of scores fall between 65 and 70, with fewer students scoring below 60 or above 75. This indicates that most students' performance is clustered around the average.

```
# Summary statistics for Hours_Studied
summary(project_data$Hours_Studied)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   16.00   20.00   20.02   24.00   36.00
```

The number of hours studied ranges from 4 to 36 hours, with a median of 20 hours and a mean of 20.02 hours, indicating that most students study around 20 hours. The middle 50% of students study between 16 and 24 hours.

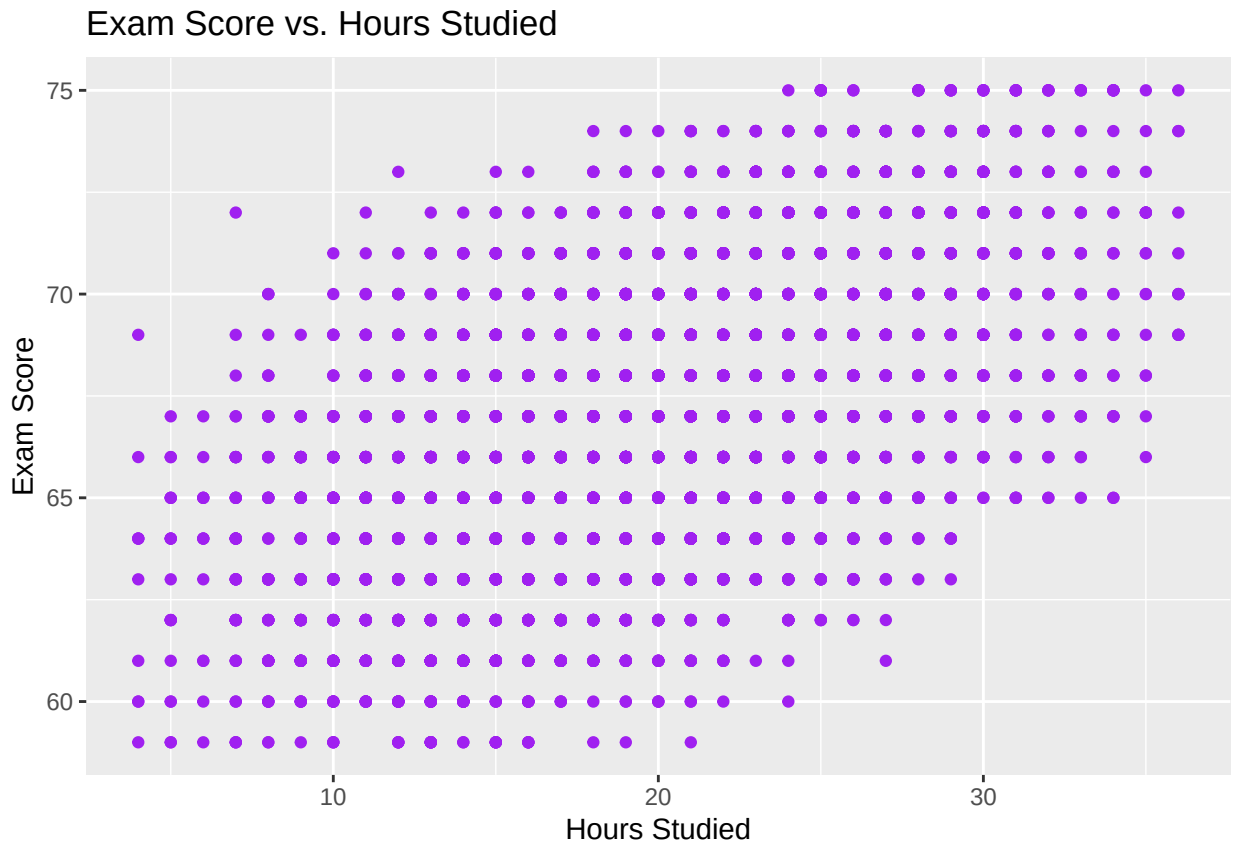
```
# Bar plot for Parental Education Level
ggplot(project_data, aes(x = Parental_Education_Level)) +
  geom_bar(fill = "salmon", color = "black") +
  labs(title = "Parental Education Level Distribution", x = "Parental Education Level", y = "Frequency")
```



The bar plot shows that the parents with high schoool level has a greater frequency than the parents who did postgraduation and who are in college level.

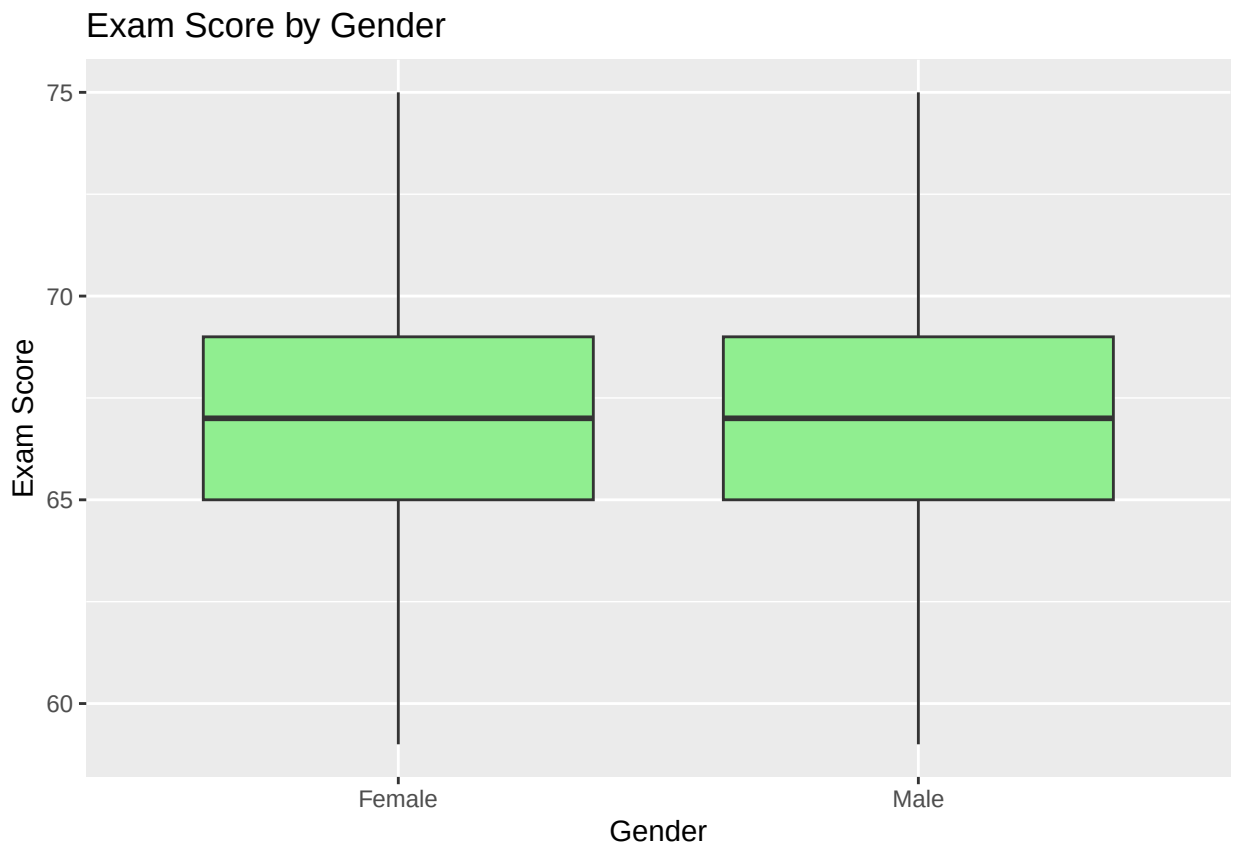
#Bivariate Analysis

```
# Scatter plot of Exam_Score vs. Hours_Studied
ggplot(project_data, aes(x = Hours_Studied, y = Exam_Score)) +
  geom_point(color = "purple") +
  labs(title = "Exam Score vs. Hours Studied", x = "Hours Studied", y = "Exam Score")
```



. Scatter Plot of Hours Studied vs Exam Scores: The scatter plot demonstrates a positive linear relationship between hours studied and exam scores the dots form an upward pattern showing a positive link between studying time and exam performance with students who study more generally achieving higher scores.

```
# Box plot of Exam_Score by Gender
ggplot(project_data, aes(x = Gender, y = Exam_Score)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Exam Score by Gender", x = "Gender", y = "Exam Score")
```



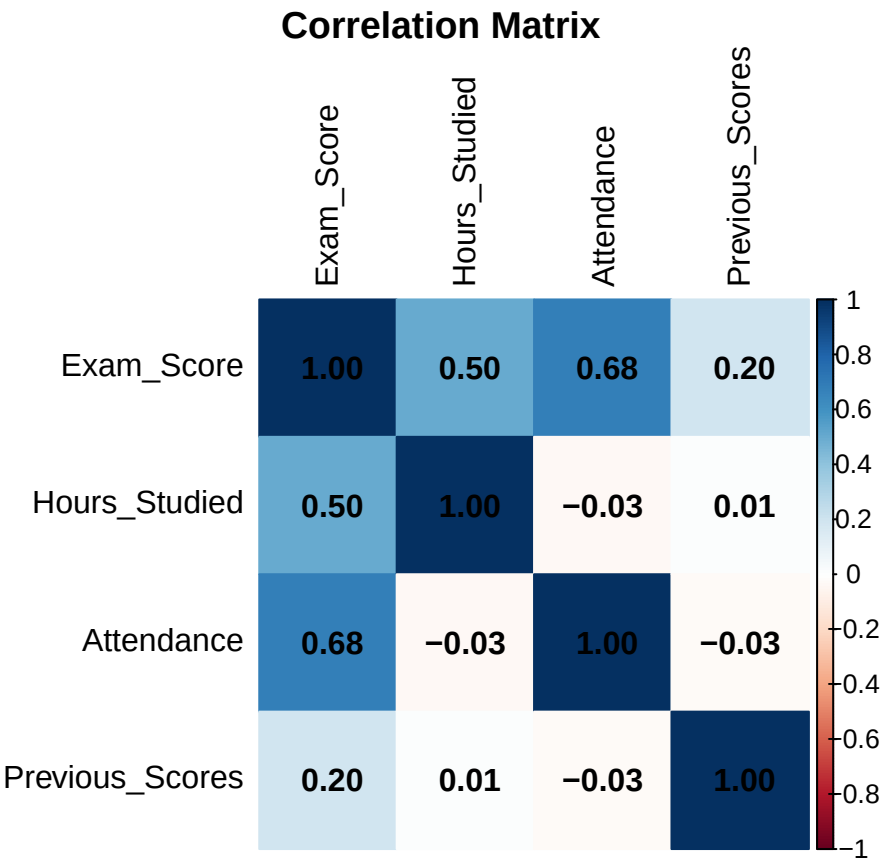
.The boxplot shows exam scores by gender. Both female and male students have similar exam score ranges, with medians around 68-70. There doesn't appear to be a big difference in exam scores between genders.

```
# Correlation matrix for numerical features
# Selecting numeric columns for correlation
num_data <- project_data %>% select(Exam_Score, Hours_Studied, Attendance, Previous_Scores)
cor_matrix <- cor(num_data, use = "complete.obs")

# Display correlation matrix
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.2
```

```
## corrplot 0.95 loaded
```

```
corrplot::corrplot(cor_matrix, method = "color", addCoef.col = "black", tl.col = "black", title = "Corr
```



Correlation Heatmap of Numeric Variables: The correlation heatmap shows a strong positive relationship between hours studied, attendance, and exam scores, suggesting that both study habits and class attendance are key predictors of academic success..The picture shows a correlation matrix, which tells us how different factors relate to exam scores. "Attendance" and "Hours Studied" have a strong positive relationship with "Exam Score" (0.68 and 0.50, respectively), meaning students who attend more and study more tend to score higher. "Previous Scores" has a weaker relationship with "Exam Score" (0.20).

#Splitting the data into training and testing

```
# Set seed for reproducibility
set.seed(123)

# Define the split ratio
split_ratio <- 0.7

# Create an index for training data
train_index <- sample(1:nrow(project_data), size = round(split_ratio * nrow(project_data)))

# Split the data into training and testing
train_data <- project_data[train_index, ]
test_data <- project_data[-train_index, ]

# Confirm the split
dim(train_data)
```

```
## [1] 4238   20
```

```r
dim(test_data)
```

```
## [1] 1816   20
```

```r
# Fit the multiple linear regression model
regression_model <- lm(Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement + Internet_Access

# Display the summary of the regression model
summary(regression_model)
```

```
##
## Call:
## lm(formula = Exam_Score ~ Hours_Studied + Attendance + Parental_Involvement +
##     Internet_Access + Motivation_Level + Previous_Scores, data = project_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0378 -0.8769  0.0216  0.8907  4.2584
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                42.178259   0.172551  244.44   <2e-16 ***
## Hours_Studied               0.293965   0.002847  103.24   <2e-16 ***
## Attendance                  0.199017   0.001430  139.16   <2e-16 ***
## Parental_InvolvementLow    -1.966503   0.047847  -41.10   <2e-16 ***
## Parental_InvolvementMedium -0.975386   0.038263  -25.49   <2e-16 ***
## Internet_AccessYes          0.872409   0.061891   14.10   <2e-16 ***
## Motivation_LevelLow        -1.001464   0.047764  -20.97   <2e-16 ***
## Motivation_LevelMedium     -0.486033   0.043494  -11.18   <2e-16 ***
## Previous_Scores             0.048324   0.001145   42.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.278 on 6045 degrees of freedom
## Multiple R-squared:  0.8433, Adjusted R-squared:  0.8431
## F-statistic:  4066 on 8 and 6045 DF,  p-value: < 2.2e-16
```

The linear regression model shows that Hours_Studied, Attendance, Previous_Scores, Internet_Access, and Motivation_Level significantly impact Exam_Score. Hours_Studied has the largest positive effect, increasing Exam_Score by approximately 0.29 points per hour studied. Attendance and Previous_Scores also positively impact scores, while Parental Involvement (low or medium) and Motivation Level (low or medium) negatively affect scores. The model explains 84.3% of the variance in Exam_Score (R-squared = 0.843).

```r
# Perform ANOVA for Motivation Level
anova_motivation <- aov(Exam_Score ~ Motivation_Level, data = project_data)

#Display the ANOVA Summary
summary(anova_motivation)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)
## Motivation_Level    2    553  276.63   26.82 2.54e-12 ***
```

```
## Residuals        6051  62419   10.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results show that Motivation_Level significantly affects Exam_Score (p-value = 2.54e-12), with students having different motivation levels scoring differently. The F-value of 26.82 indicates that the variation in scores between motivation levels is much greater than the variation within each group, confirming the importance of motivation in exam performance.

```
# Perform ANOVA for Parental Involvement and Exam_Score
anova_parental_involvement <- aov(Exam_Score ~ Parental_Involvement, data = project_data)

# Display the ANOVA summary
summary(anova_parental_involvement)
```

```
##                       Df Sum Sq Mean Sq F value Pr(>F)
## Parental_Involvement    2   2018  1009.1   100.2 <2e-16 ***
## Residuals            6051  60954    10.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results show that Parental Involvement significantly affects Exam_Score (p-value $<$ 2e-16), with students from different levels of parental involvement scoring differently. The F-value of 100.2 indicates that the variation in Exam_Score between parental involvement levels is much greater than the variation within each group, confirming its strong impact on performance.

Both Motivation Level and Parental Involvement significantly affect Exam_Score, which means students with different levels of motivation and parental involvement do score differently.

```
# Perform a t-test for Internet Access
t_test <- t.test(Exam_Score ~ Internet_Access, data = project_data)
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  Exam_Score by Internet_Access
## t = -5.5436, df = 546.73, p-value = 4.616e-08
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1.131260 -0.539313
## sample estimates:
##  mean in group No mean in group Yes
##          66.20996          67.04524
```

The t-test results show that there is a significant difference in Exam_Score between students with and without Internet Access (p-value = 4.616e-08). Students with internet access have a slightly higher mean score (67.05) compared to those without (66.21), with the difference being statistically significant which means students with internet access perform slightly better than those without, as indicated by the significant difference in their exam scores.

```r
# Correlation analysis for numeric variables
correlation_matrix <- cor(project_data[, c("Exam_Score", "Hours_Studied", "Attendance", "Sleep_Hours",
print(correlation_matrix)
```

```
##                   Exam_Score Hours_Studied  Attendance  Sleep_Hours
## Exam_Score        1.000000000    0.50080928  0.68225671 -0.009100561
## Hours_Studied     0.500809277    1.00000000 -0.03067146  0.011480446
## Attendance        0.682256713   -0.03067146  1.00000000 -0.018986041
## Sleep_Hours      -0.009100561    0.01148045 -0.01898604  1.000000000
## Previous_Scores   0.198775257    0.01432053 -0.02715053 -0.018354270
##                  Previous_Scores
## Exam_Score            0.19877526
## Hours_Studied         0.01432053
## Attendance           -0.02715053
## Sleep_Hours          -0.01835427
## Previous_Scores       1.00000000
```

The correlation analysis shows that Exam_Score is moderately positively correlated with Hours_Studied (0.50) and Attendance (0.68), while it has a very weak negative correlation with Sleep_Hours (-0.009). Previous_Scores has a weak positive correlation with Exam_Score (0.20).

Hours_Studied and Attendance are moderately correlated with Exam_Score, indicating that students who study more and attend school regularly tend to have higher exam scores. Sleep_Hours does not have a strong correlation with Exam_Score.

#Conclusion:

The analysis clearly shows that study habits (hours studied), attendance, motivation level, and internet access significantly affect Exam_Score. Students who study more, attend school regularly, and have access to the internet tend to perform better. These findings highlight the importance of improving student engagement and access to resources to enhance academic performance.

IMPORTANCE OF ANALYSIS This analysis is important because it helps identify the key factors that affect student exam scores, such as study habits, attendance, and motivation. Understanding these factors can help educators and policymakers create better strategies to improve student performance. By focusing on what matters most, we can make learning more effective and accessible for all students.

LIMITATIONS OF THE ANALYSIS The analysis cannot prove that the factors directly cause higher exam scores, only that they are linked. It may be biased due to self-reported data, missing information, and may not apply to all student groups.Additionally, it focuses only on certain factors and ignores others like socio-economic status or mental health.