project

Sirapu Nandini

2024-11-04

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

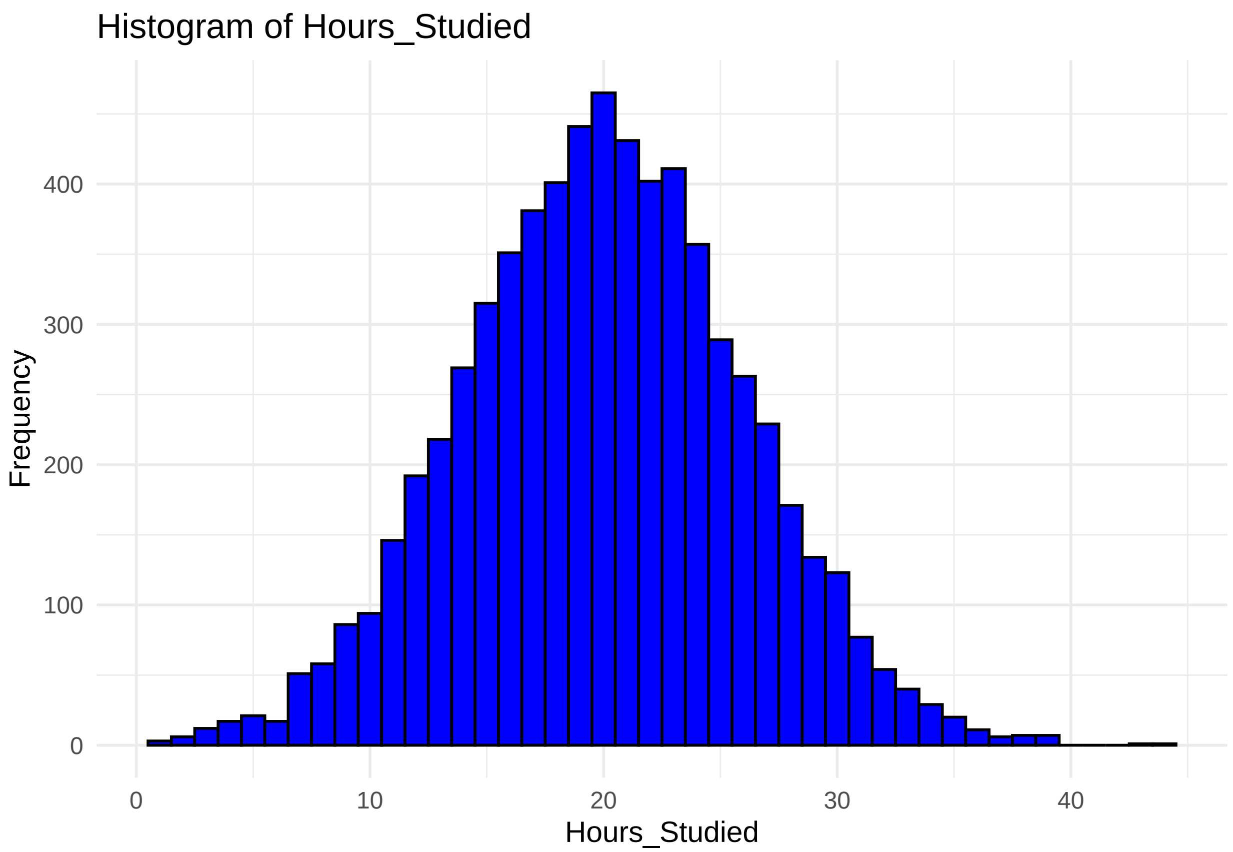
```
#Data loading
url <- "https://raw.githubusercontent.com/srirapunandini/dav---5400/refs/heads/main/StudentPerformanceF

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

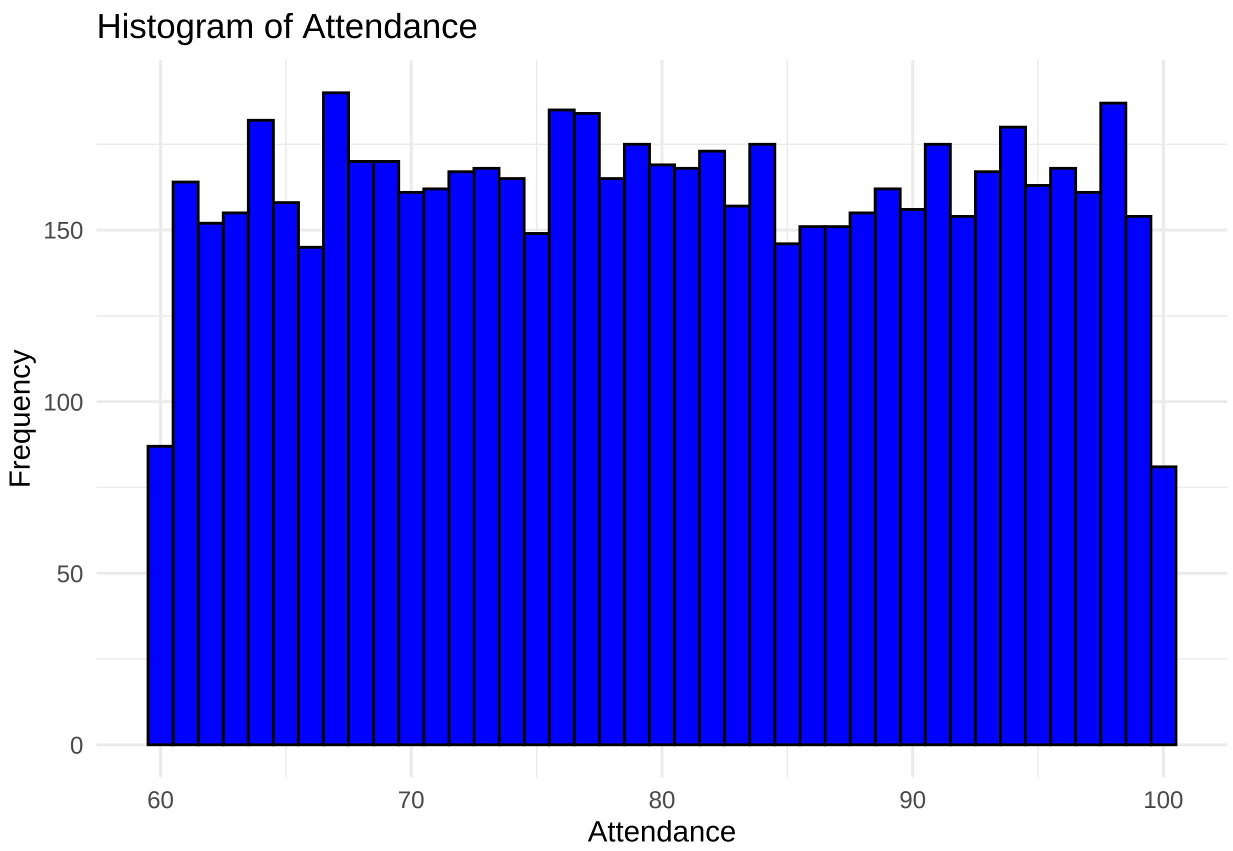
Consists of 6,607 rows and 20 columns.

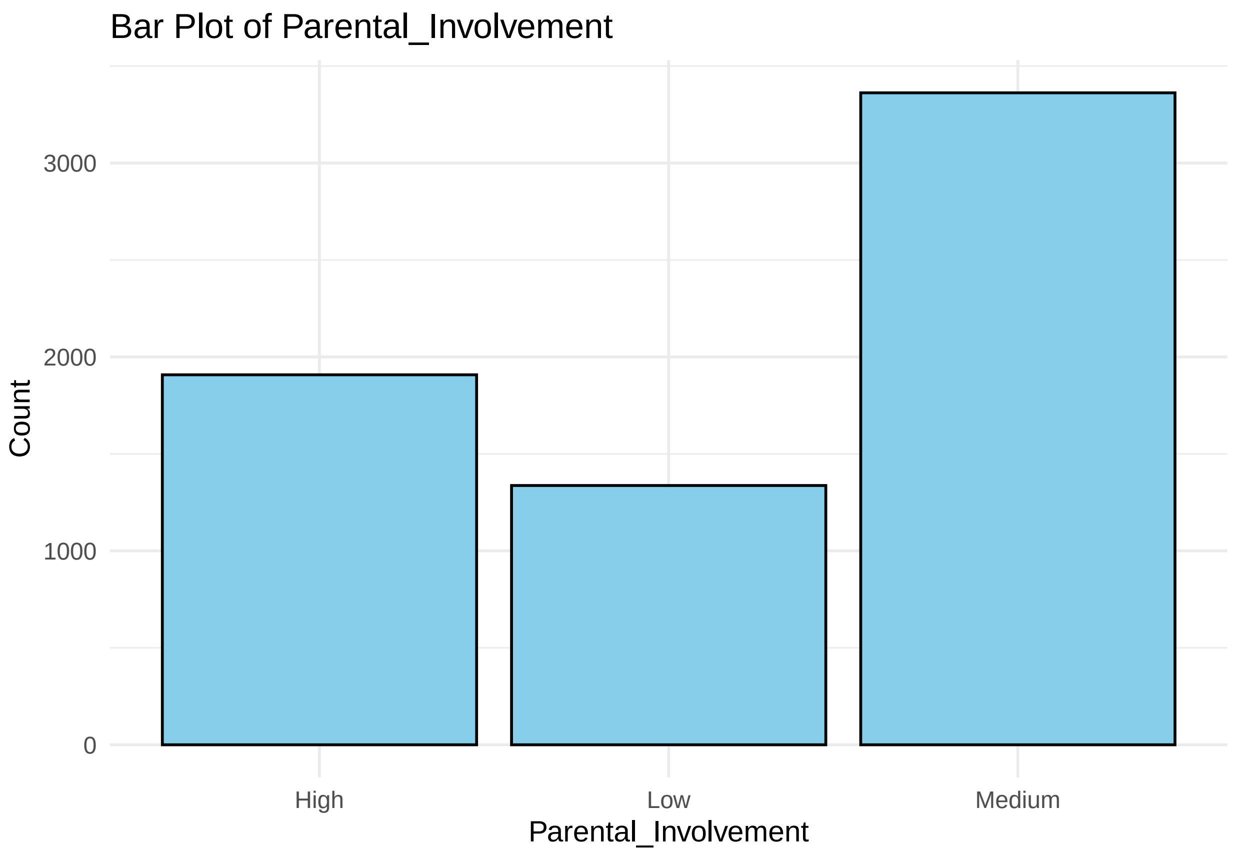
```
# View summary statistics of the dataset
summary(project_data)
```

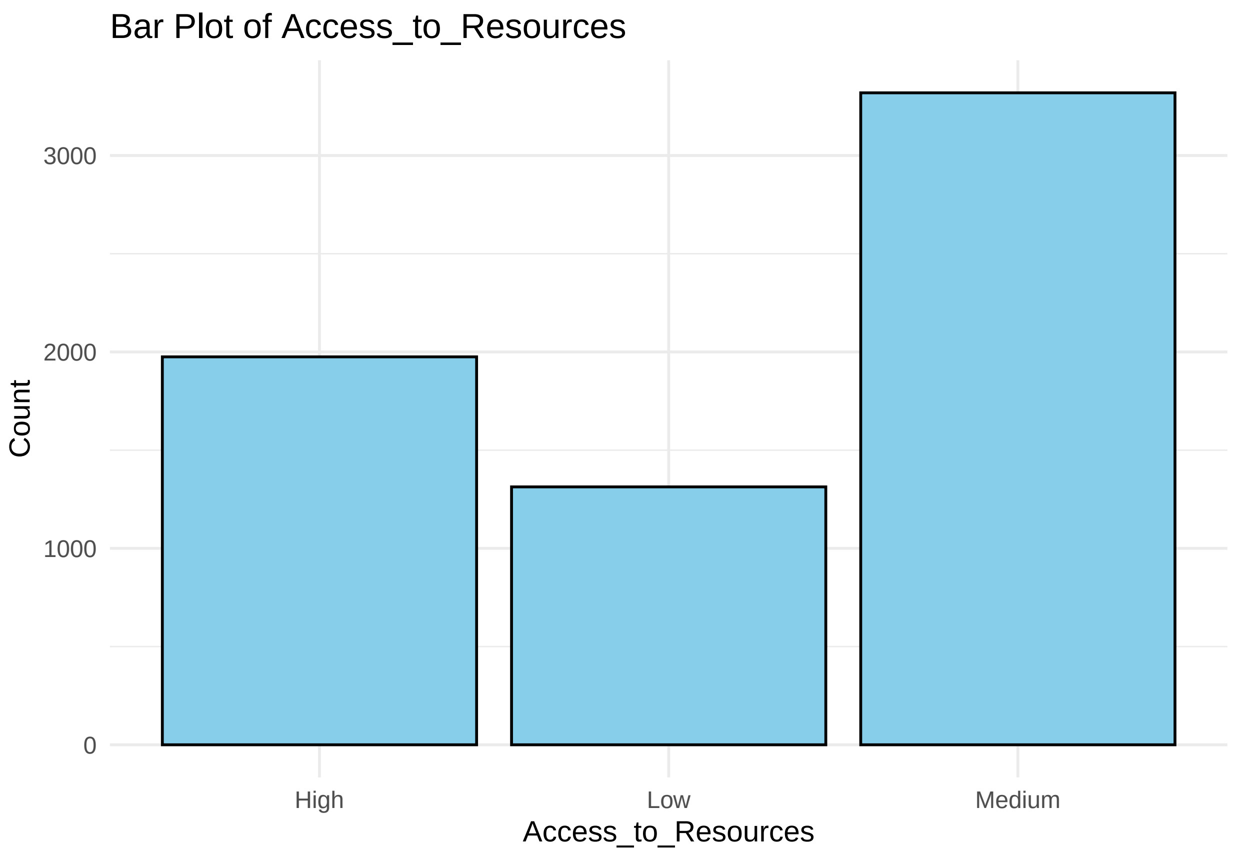
```
##  Hours_Studied    Attendance     Parental_Involvement Access_to_Resources
##  Min.   : 1.00   Min.   : 60.00  Length:6607          Length:6607
##  1st Qu.:16.00   1st Qu.: 70.00  Class :character     Class :character
##  Median :20.00   Median : 80.00  Mode  :character     Mode  :character
##  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
```

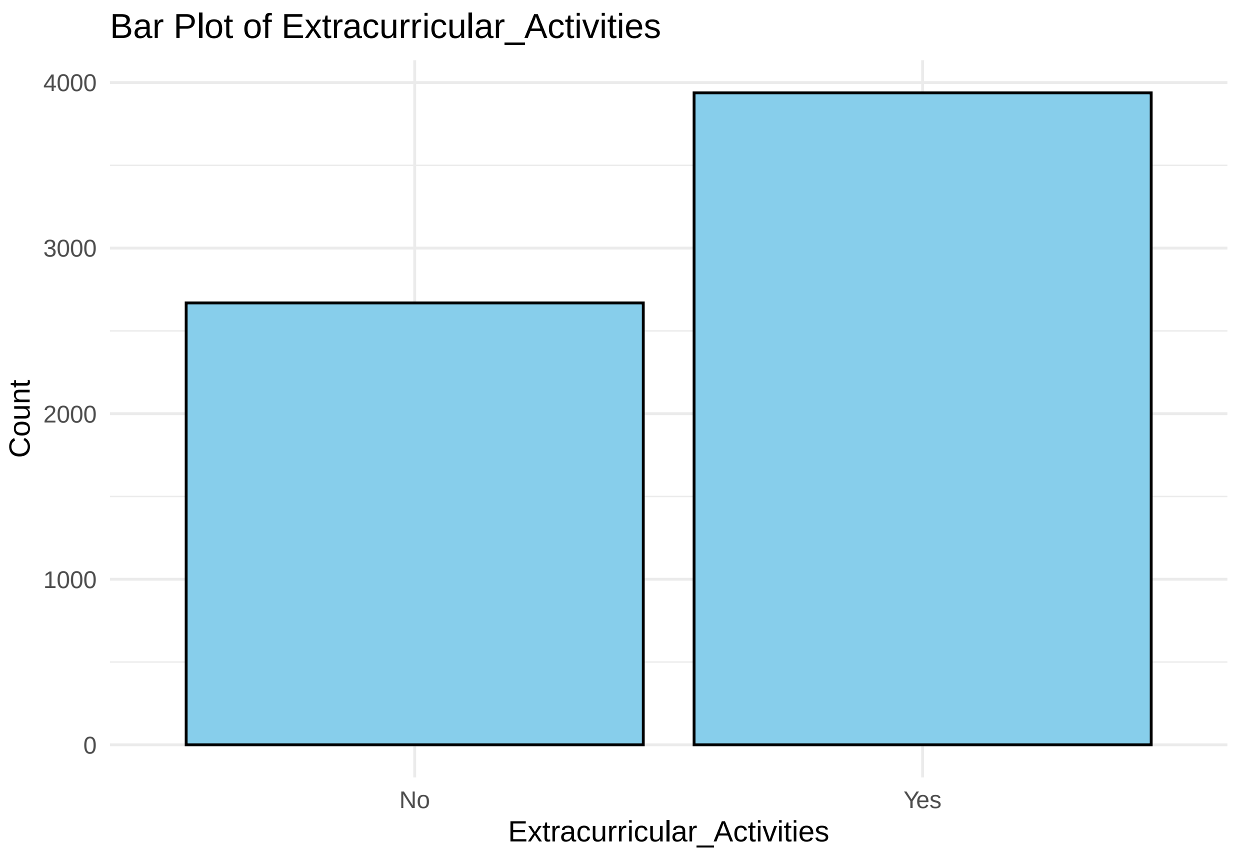
```
#plots for the variables (numerical and categorical)
library(ggplot2)

# 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)
}
```

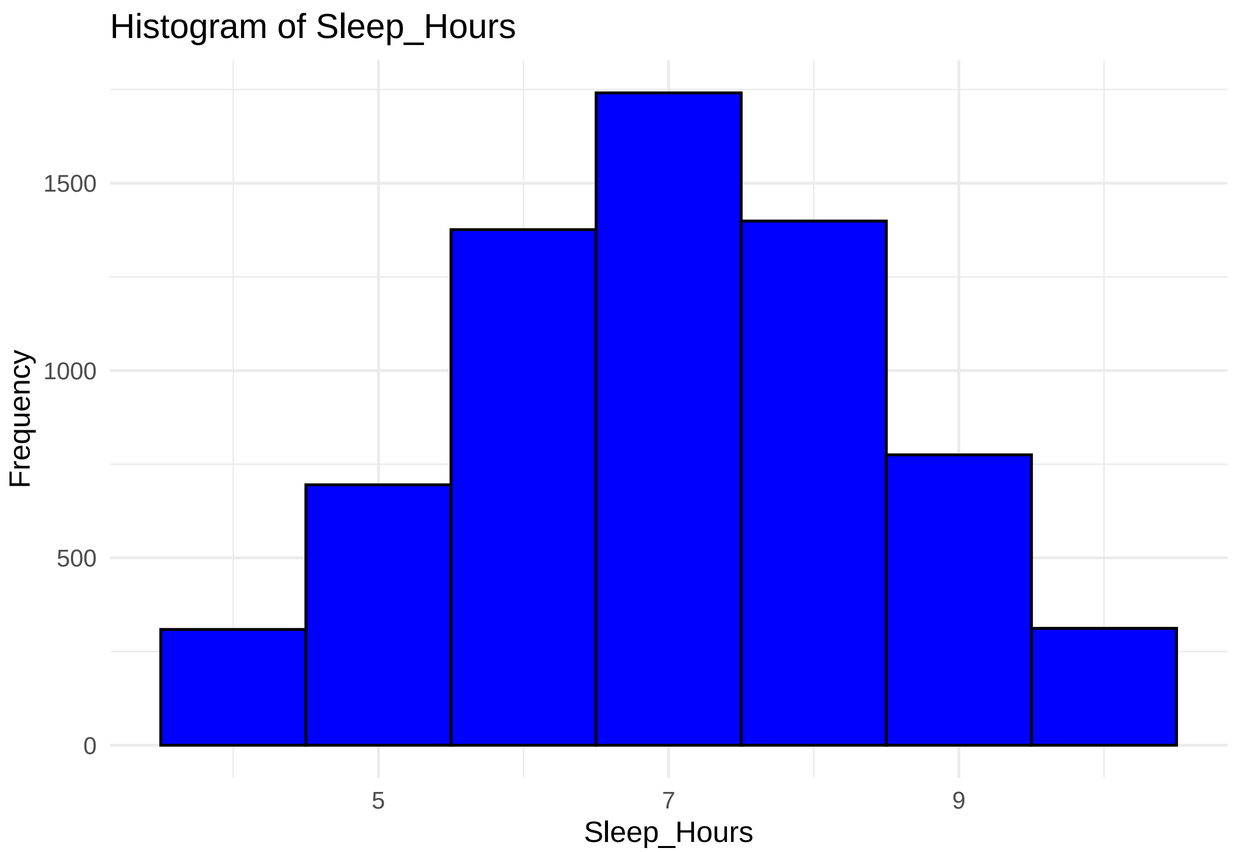

Histogram of Hours_Studied
Frequency
400
300
200
100
0
0
10
20
30
40
Hours_Studied

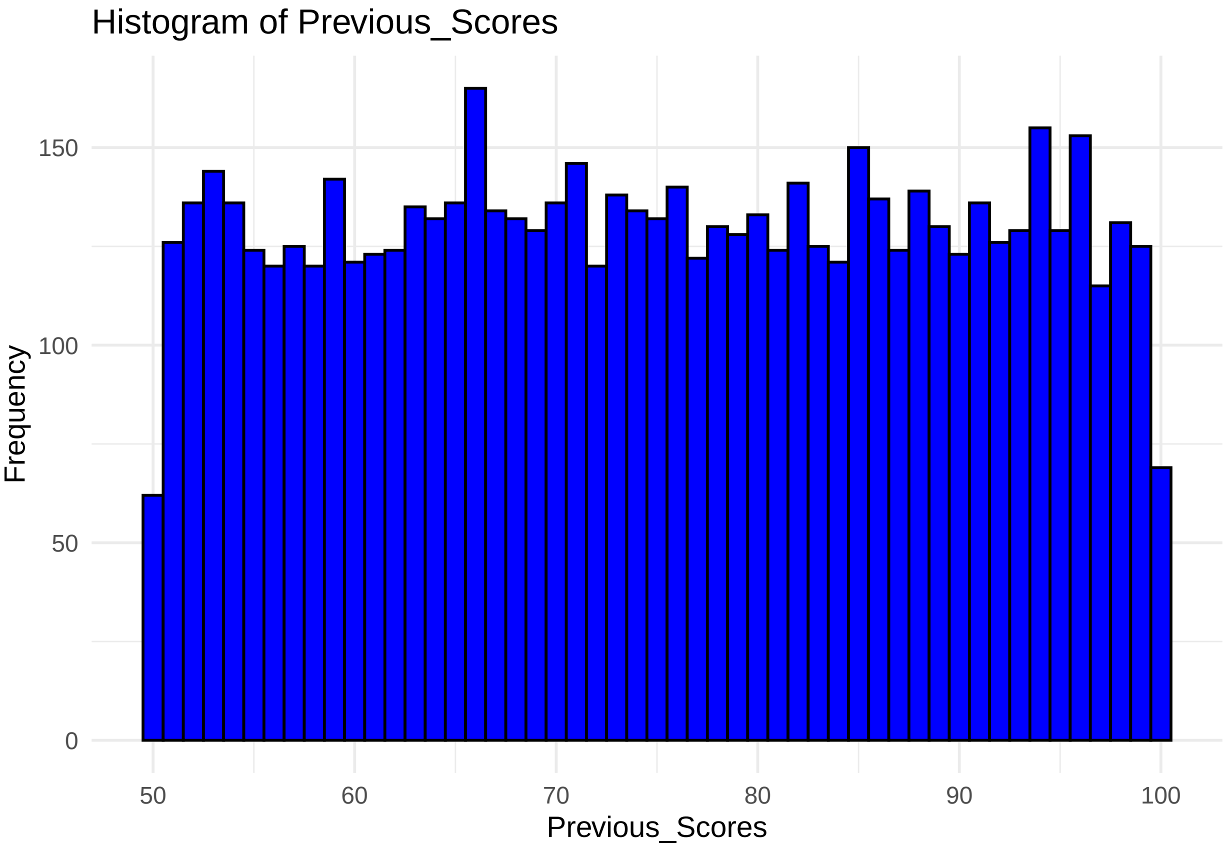

Histogram of Attendance
Frequency
0
50
100
150
60
70
80
90
100
Attendance

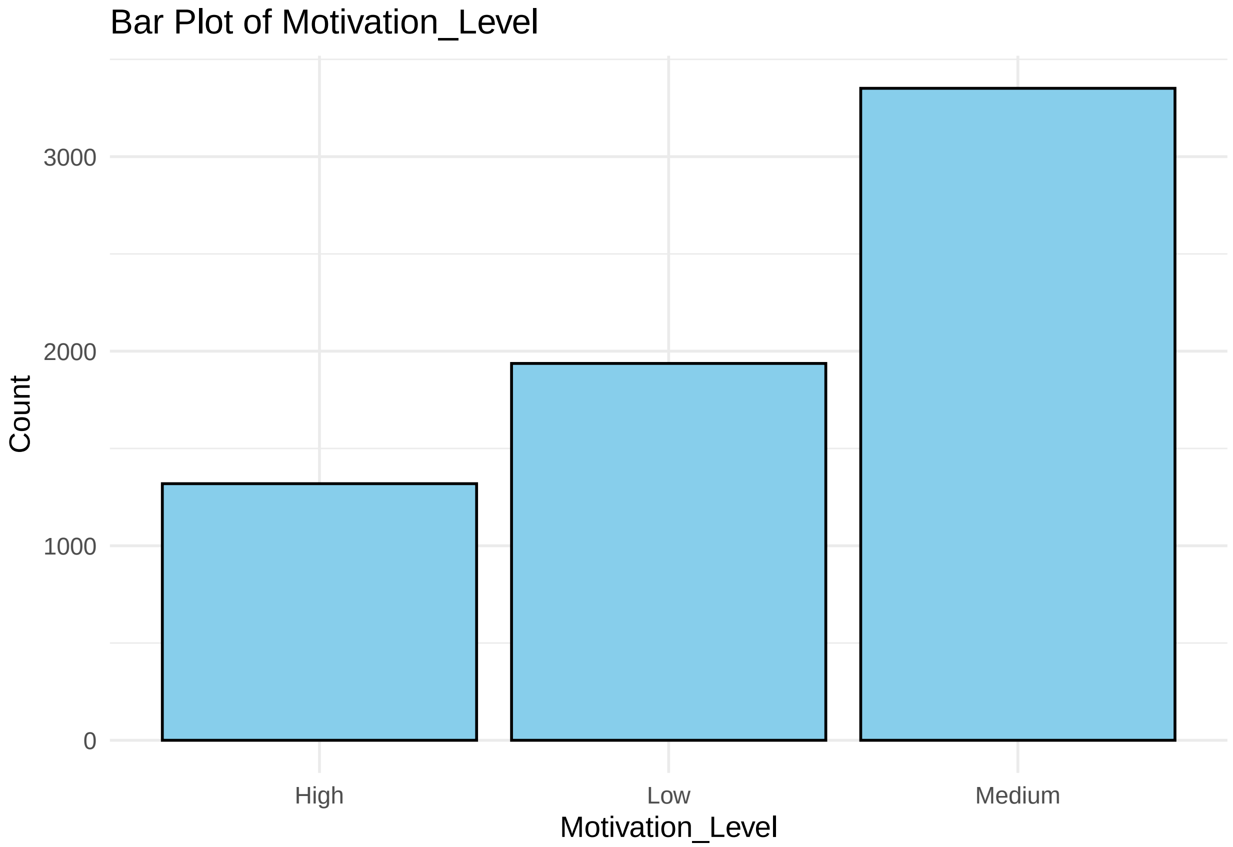

Bar Plot of Parental_Involvement
Count
3000
2000
1000
0
High
Low
Medium
Parental_Involvement

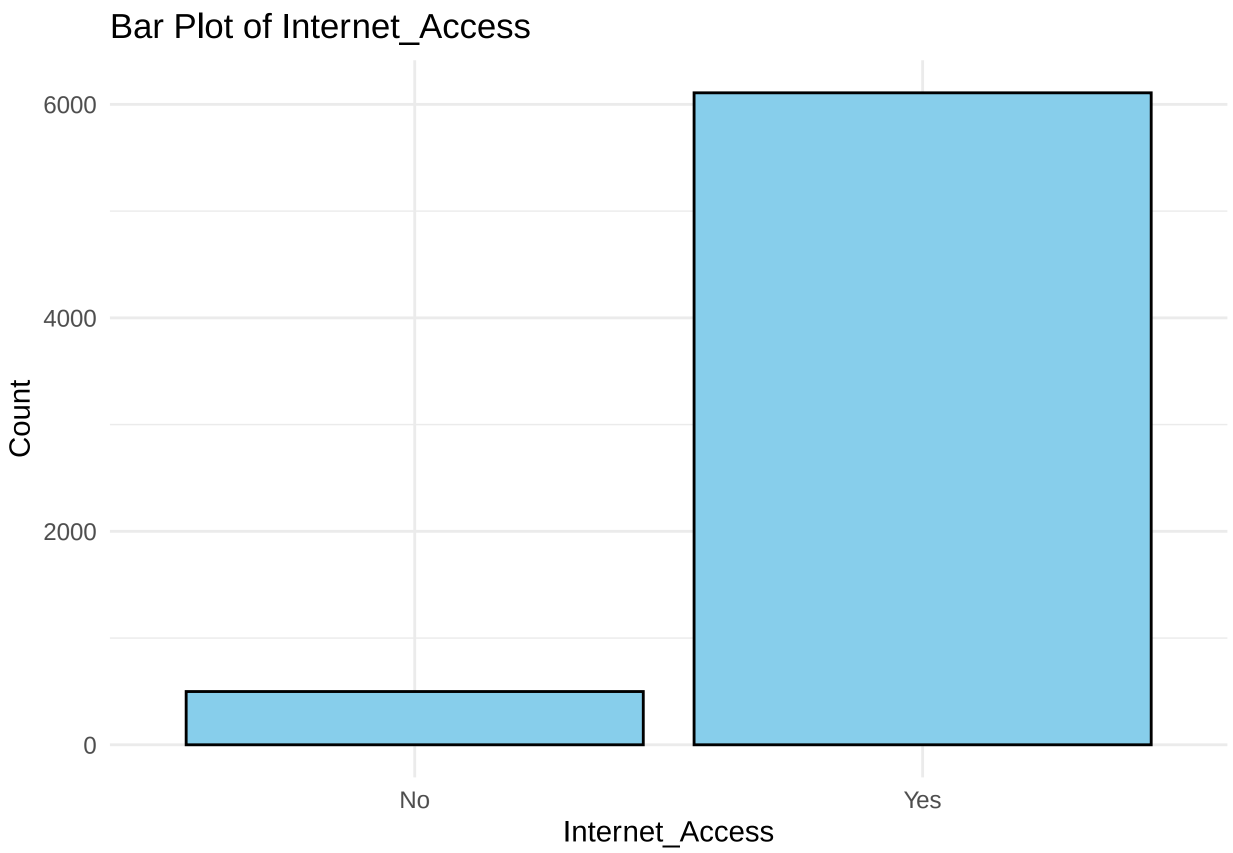

Bar Plot of Access_to_Resources
Count
3000
2000
1000
0
High
Low
Medium
Access_to_Resources

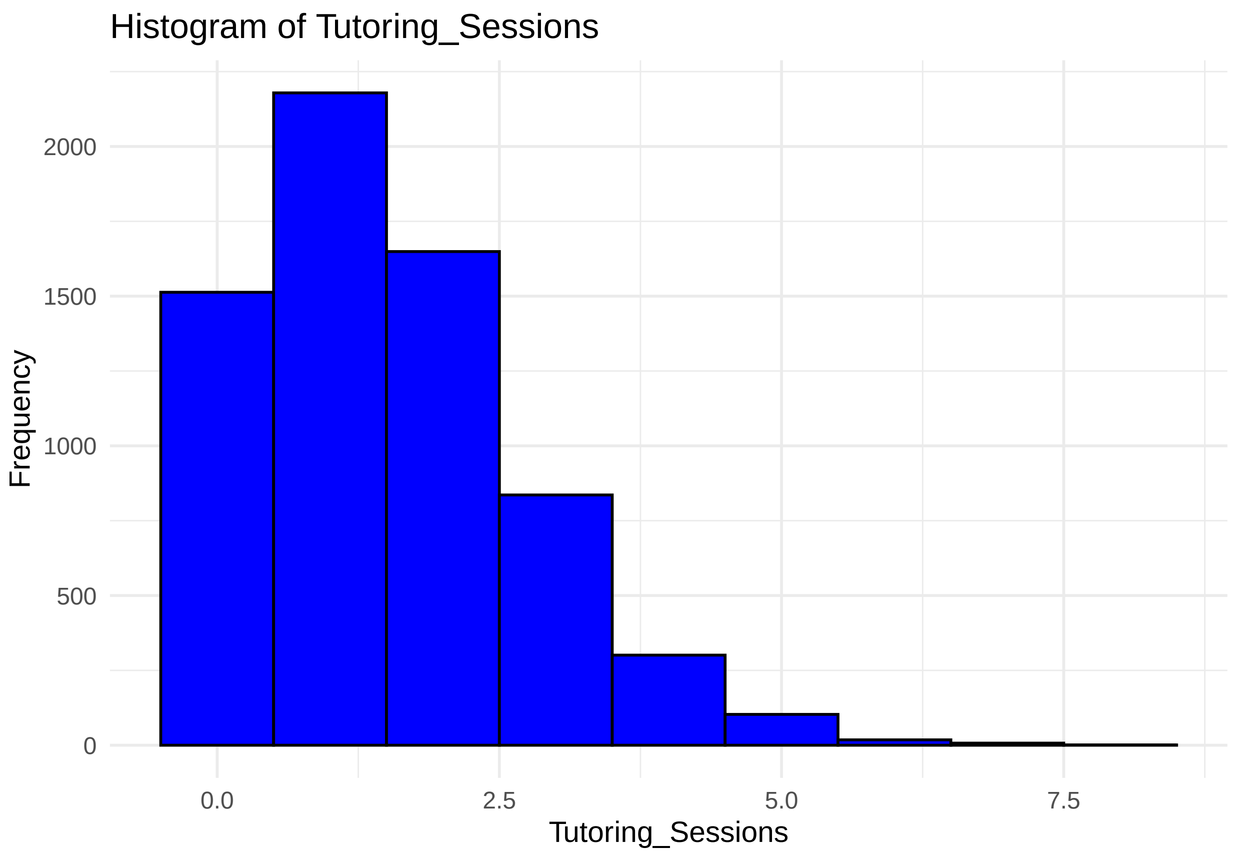

Bar Plot of Extracurricular_Activities
Count
4000
3000
2000
1000
0
No
Yes
Extracurricular_Activities

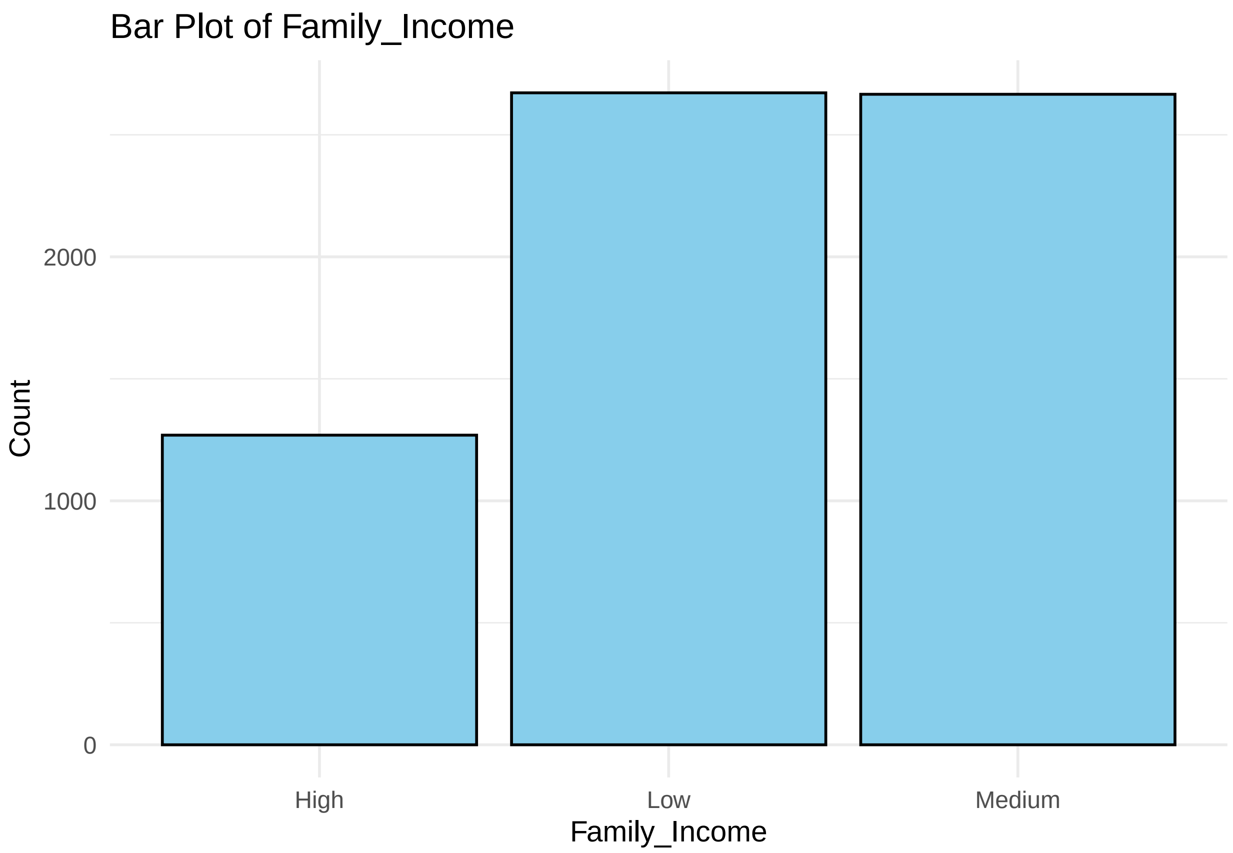

Histogram of Sleep_Hours
Frequency
1500
1000
500
0
5
7
9
Sleep_Hours

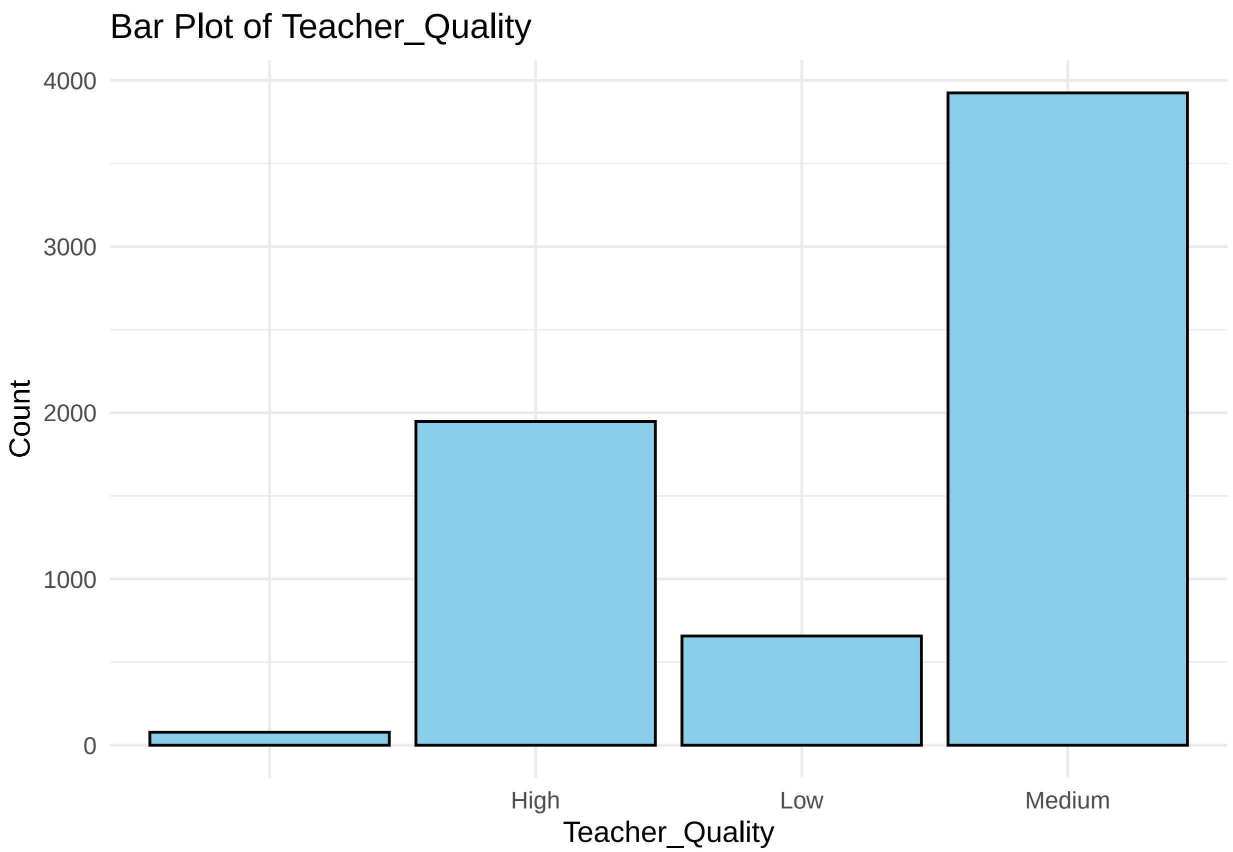

Histogram of Previous_Scores
Frequency
150
100
50
0
50
60
70
80
90
100
Previous_Scores

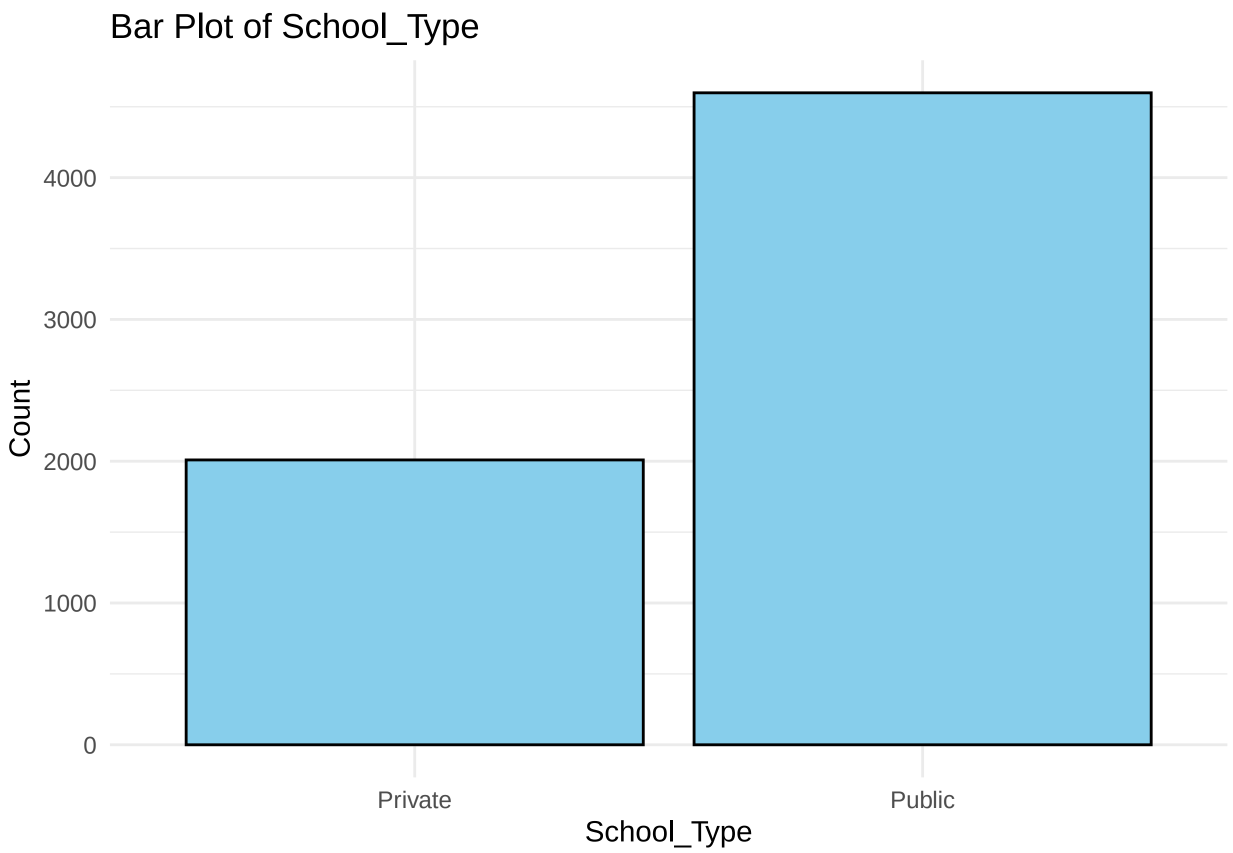

Bar Plot of Motivation_Level
Count
3000
2000
1000
0
High
Low
Medium
Motivation_Level

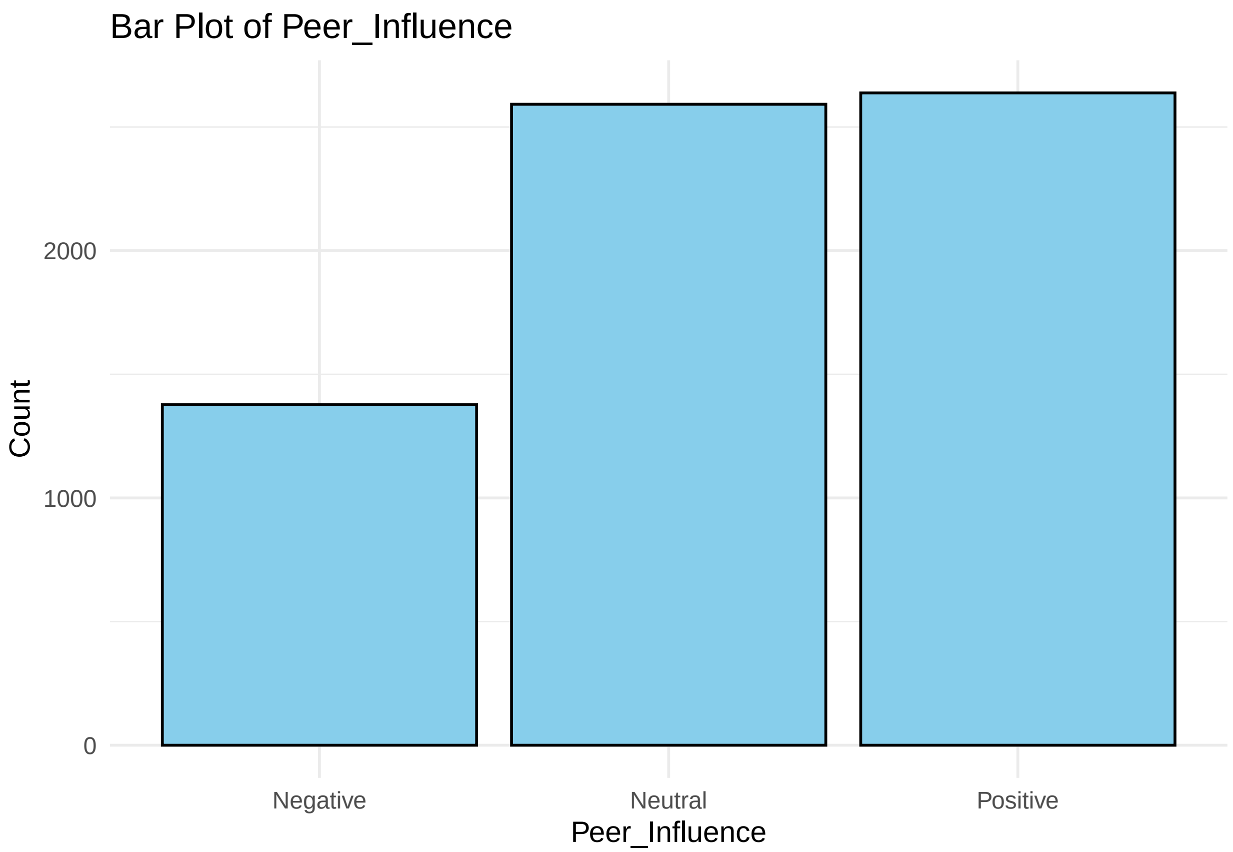

Bar Plot of Internet_Access
Count
6000
4000
2000
0
No
Yes
Internet_Access

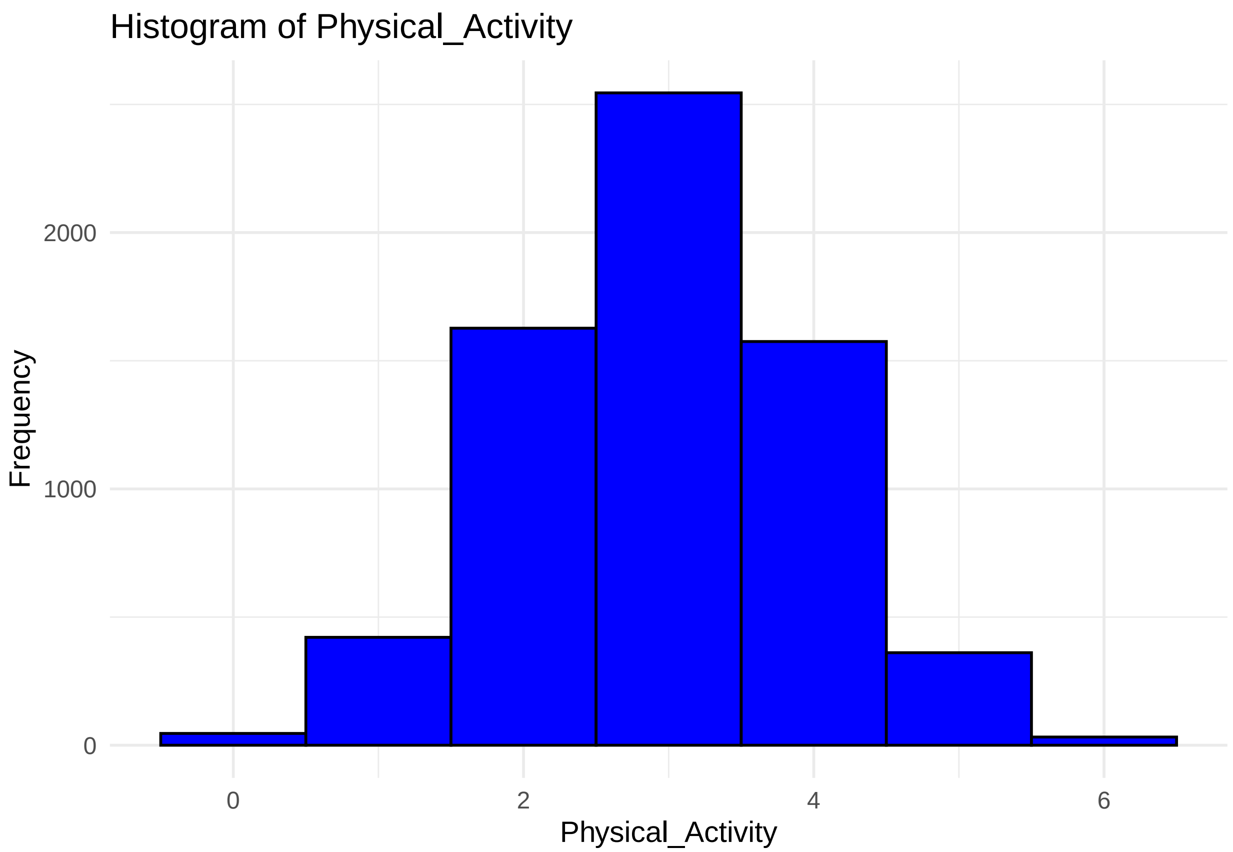

Histogram of Tutoring_Sessions
Frequency
2000
1500
1000
500
0
0.0
2.5
5.0
7.5
Tutoring_Sessions

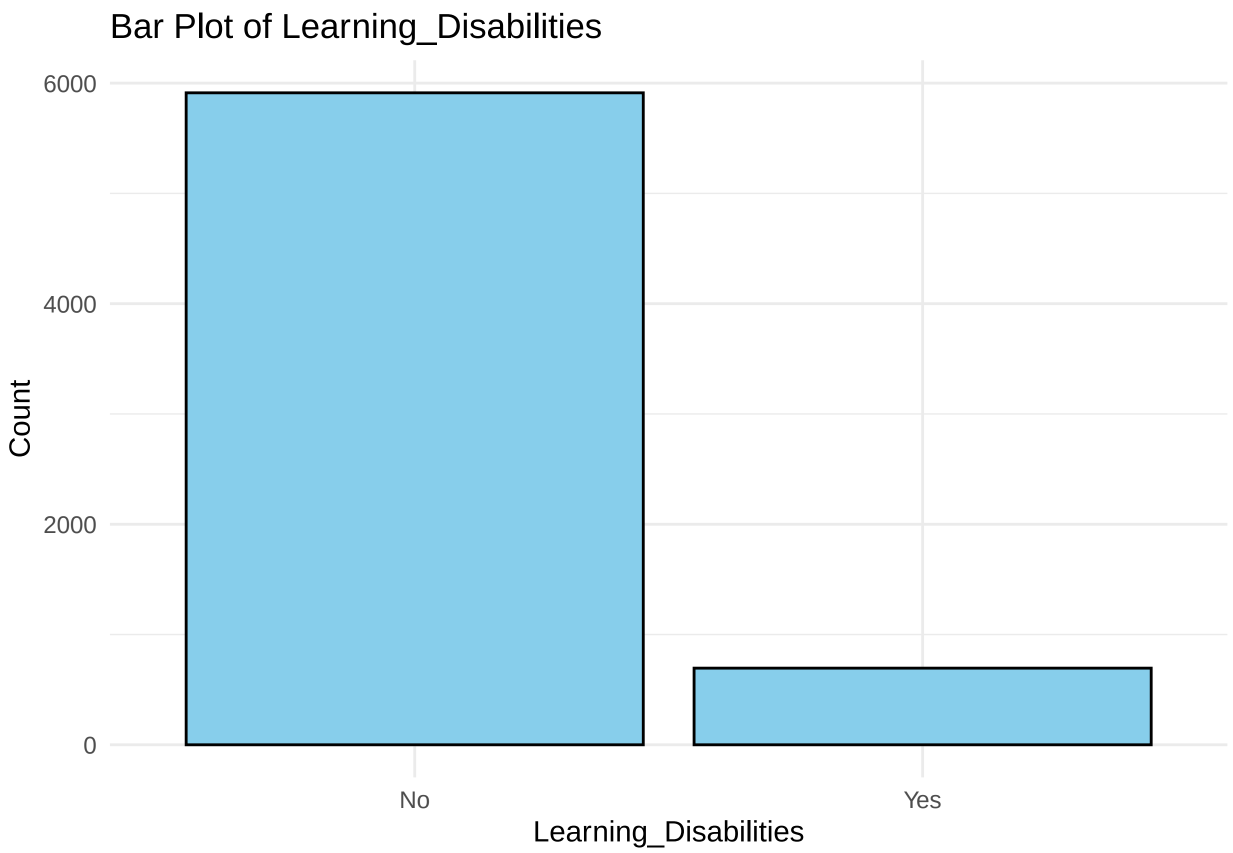

Bar Plot of Family_Income
Count
2000
1000
0
High
Low
Medium
Family_Income

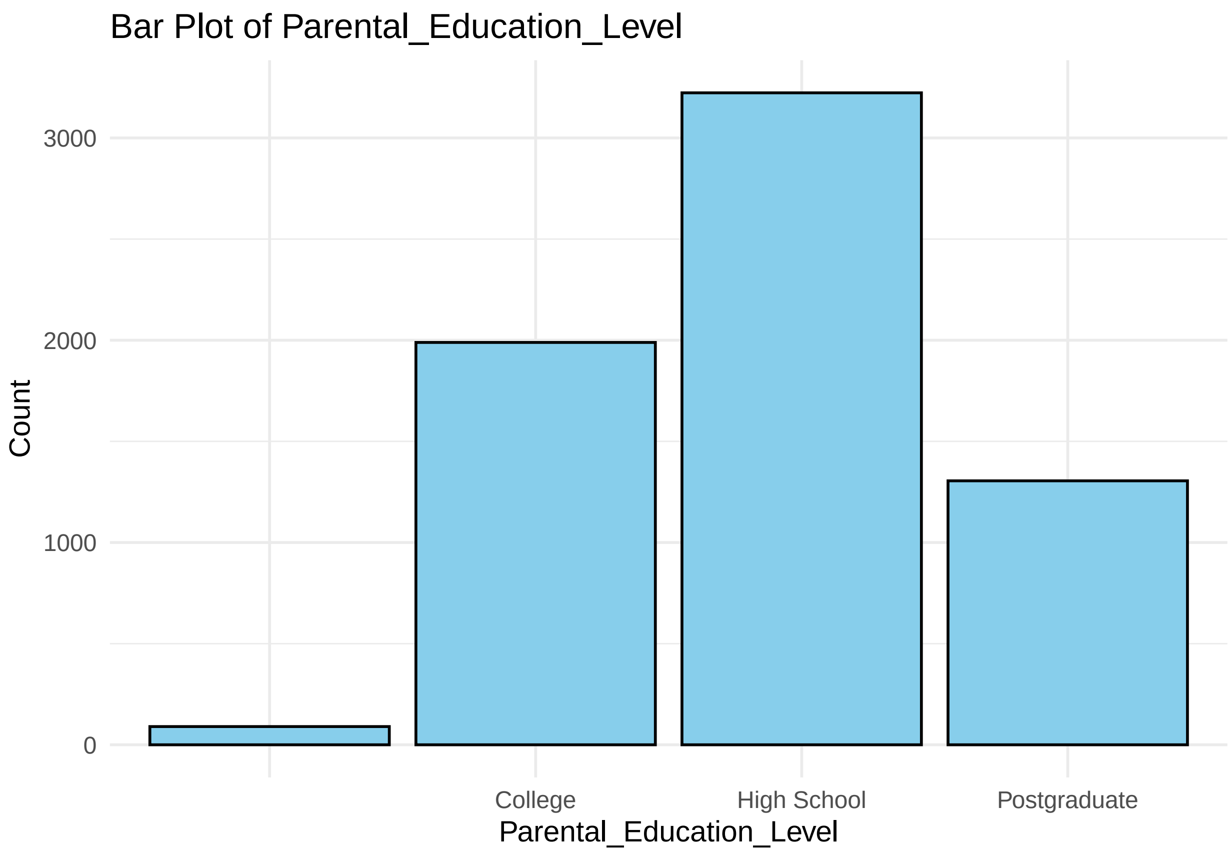

Bar Plot of Teacher_Quality
Count
4000
3000
2000
1000
0
High
Low
Medium
Teacher_Quality

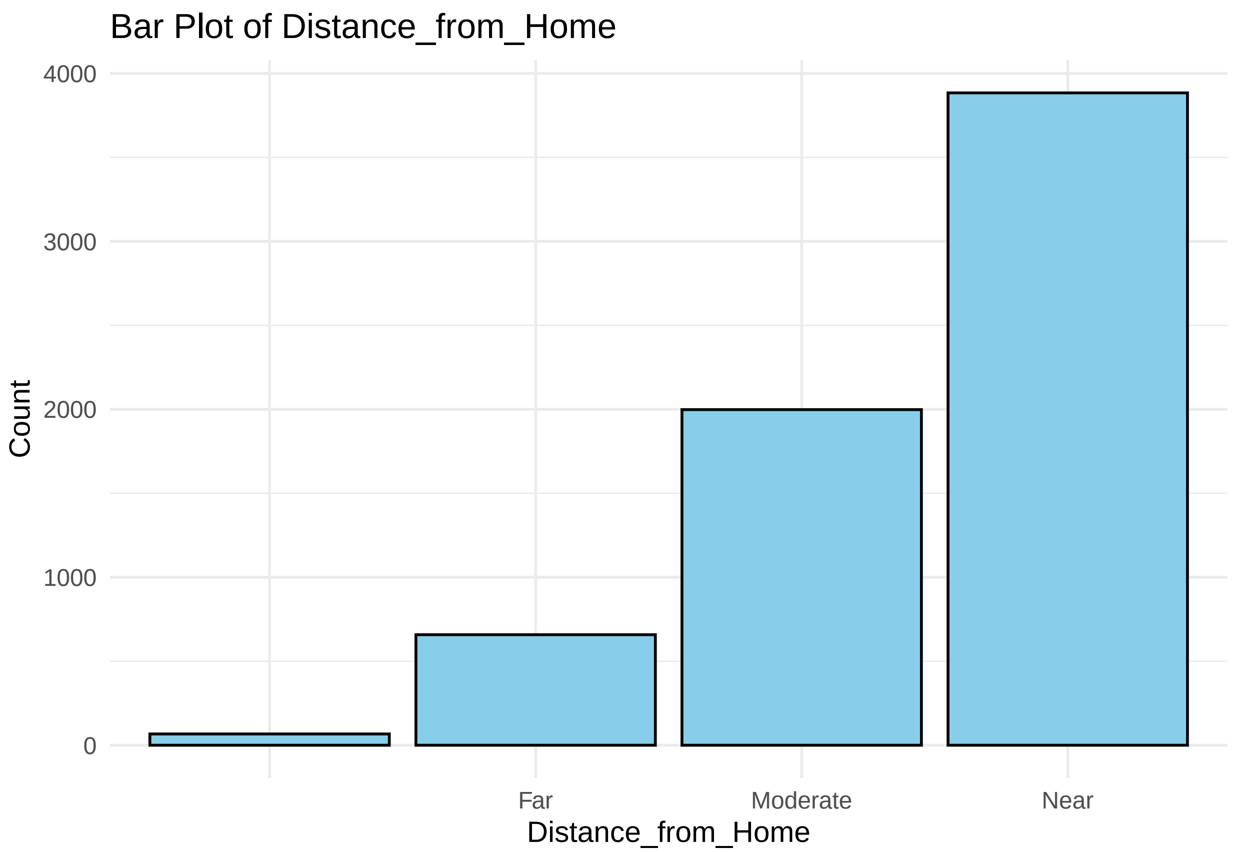

Bar Plot of School_Type
Count
4000
3000
2000
1000
0
Private
Public
School_Type

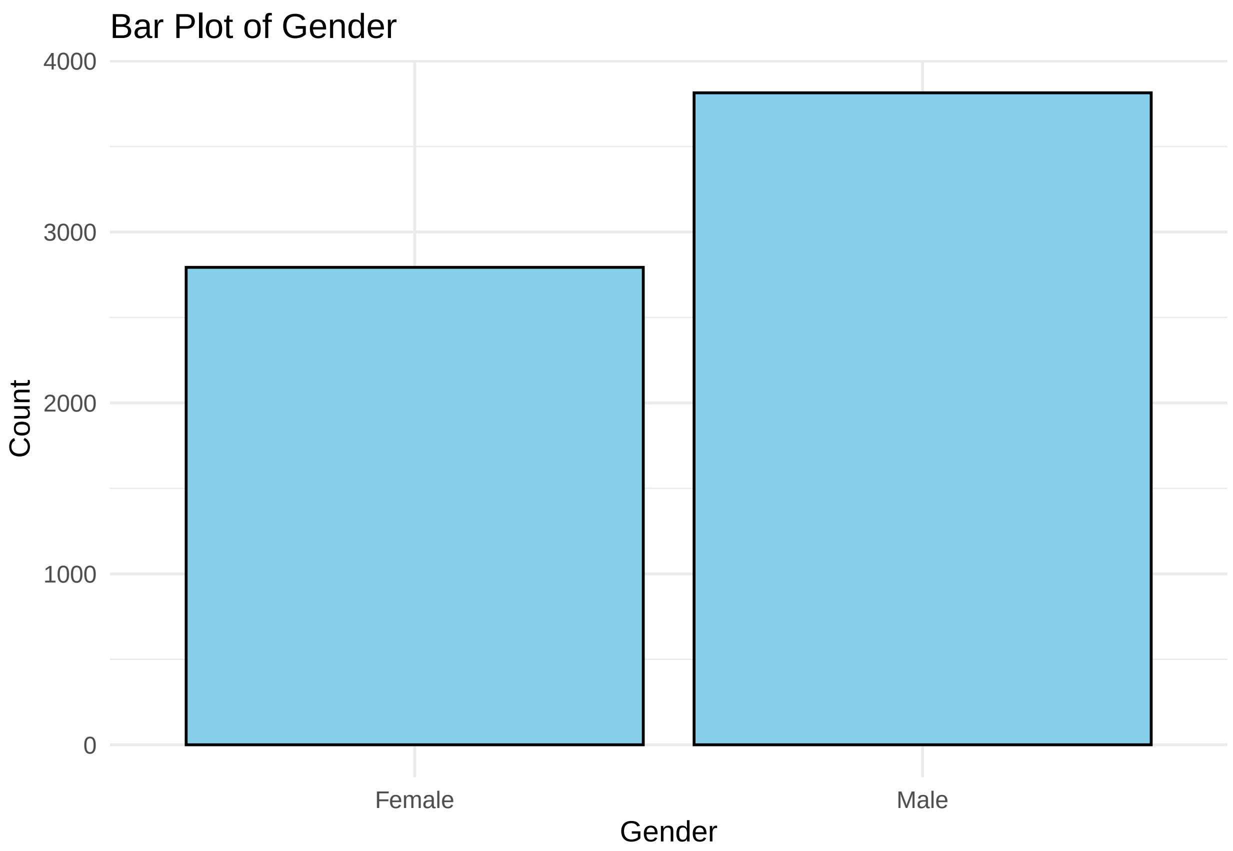

Bar Plot of Peer_Influence
Count
0
1000
2000
Negative
Neutral
Positive
Peer_Influence

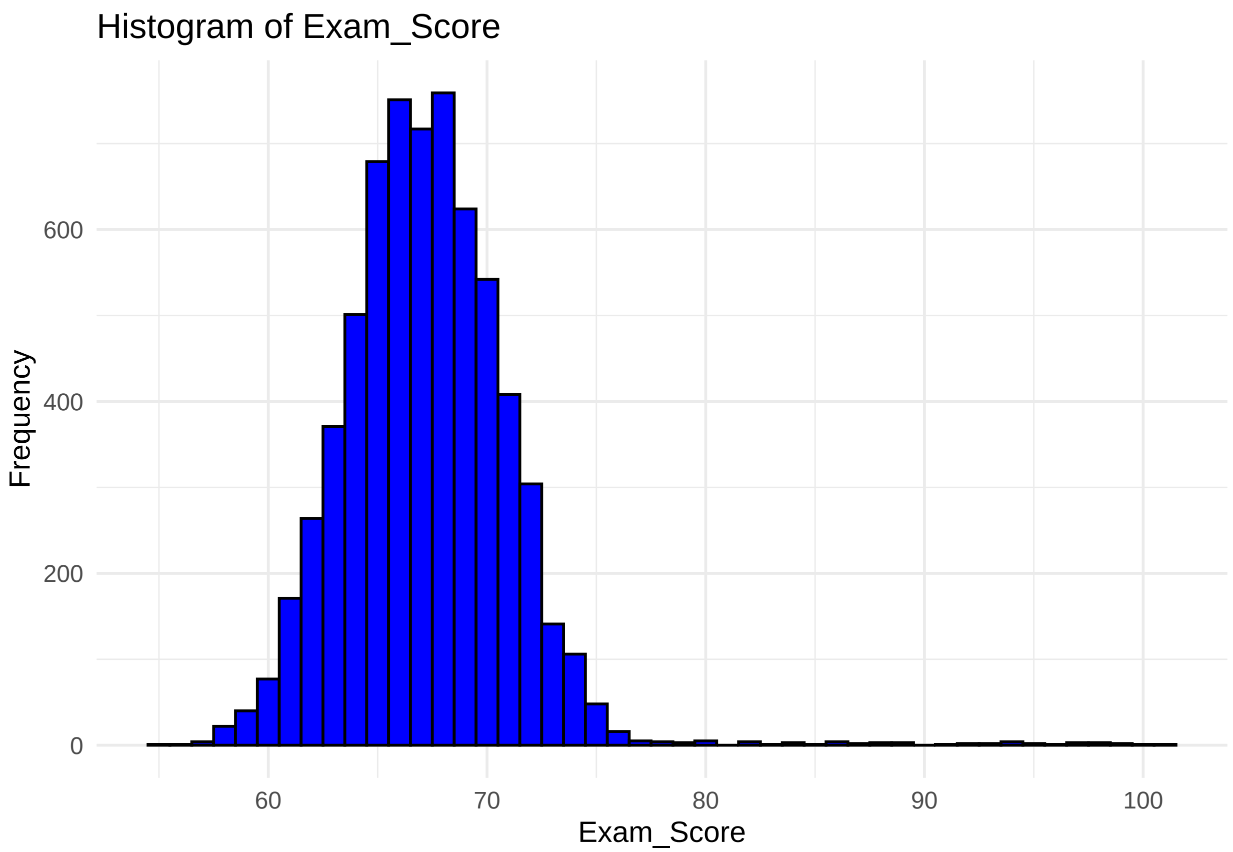

Histogram of Physical_Activity
Frequency
2000
1000
0
0
2
4
6
Physical_Activity


Bar Plot of Learning_Disabilities
6000
4000
2000
0
Count
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
4000
3000
2000
1000
0
Far
Moderate
Near
Distance_from_Home


Bar Plot of Gender
4000
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

```r
# Check for null values in each column
#Data cleaning checking for null values
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
```

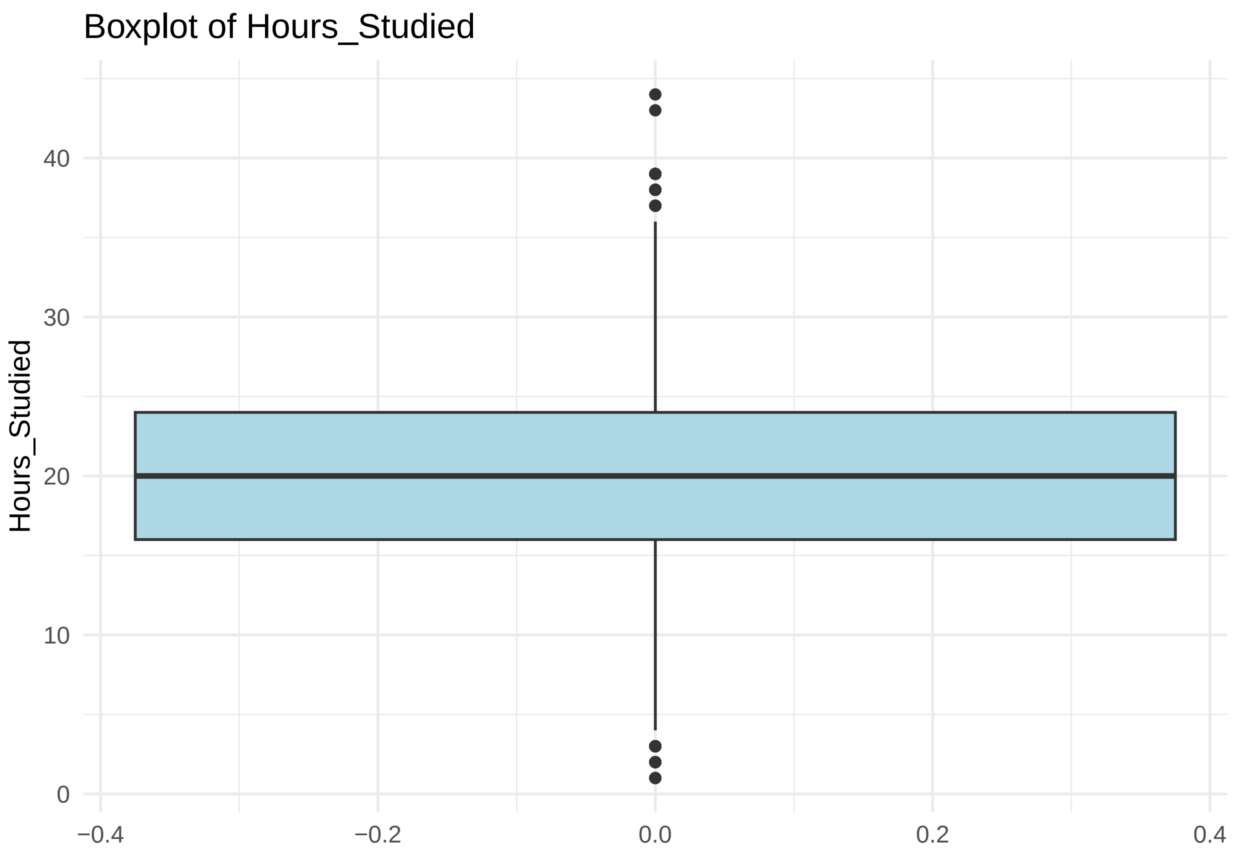
we found that there are zero null values.

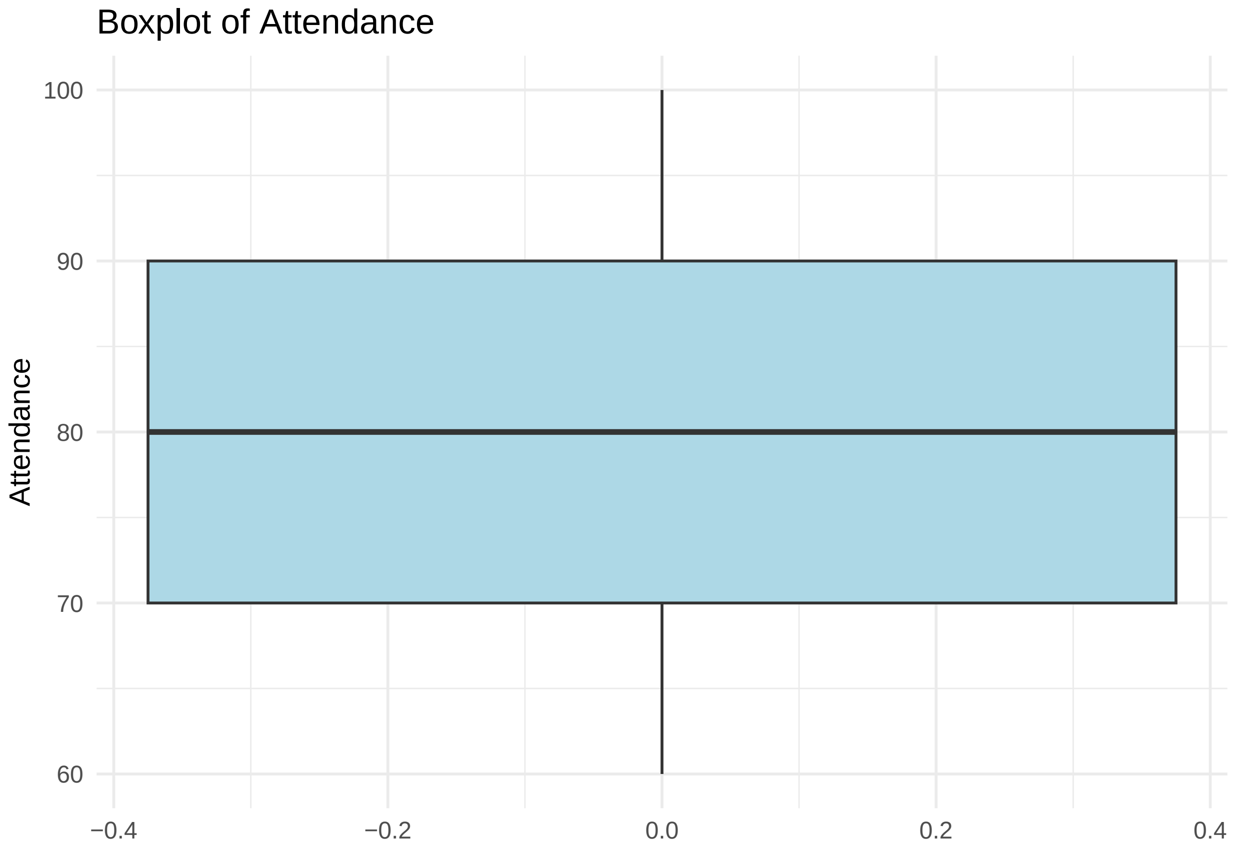
```
#using Z_scores for checking the outliers
z_scores <- scale(project_data[sapply(project_data, is.numeric)])
outliers <- abs(z_scores) > 3
colSums(outliers)
```

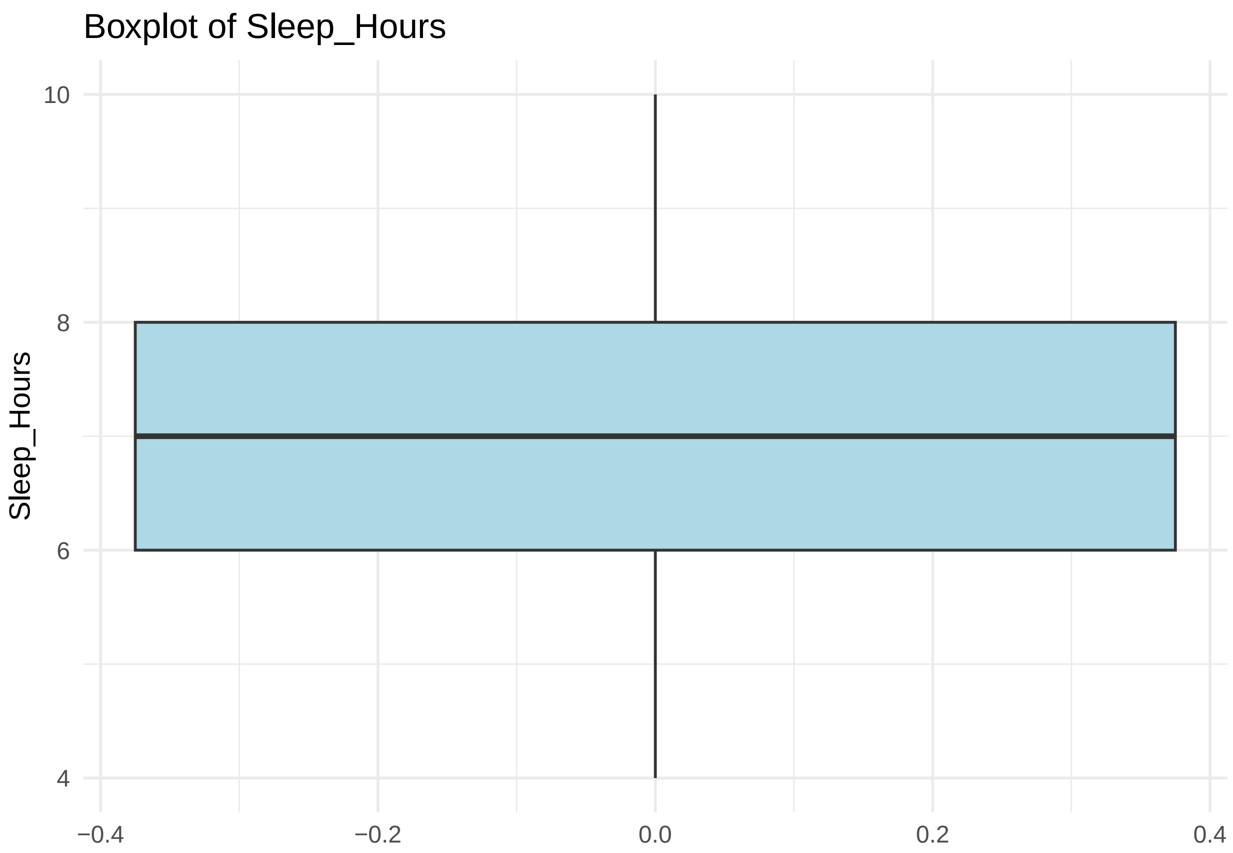
```
##     Hours_Studied        Attendance       Sleep_Hours   Previous_Scores
##                25                 0                 0                 0
## Tutoring_Sessions Physical_Activity        Exam_Score
##                26                 0                52
```

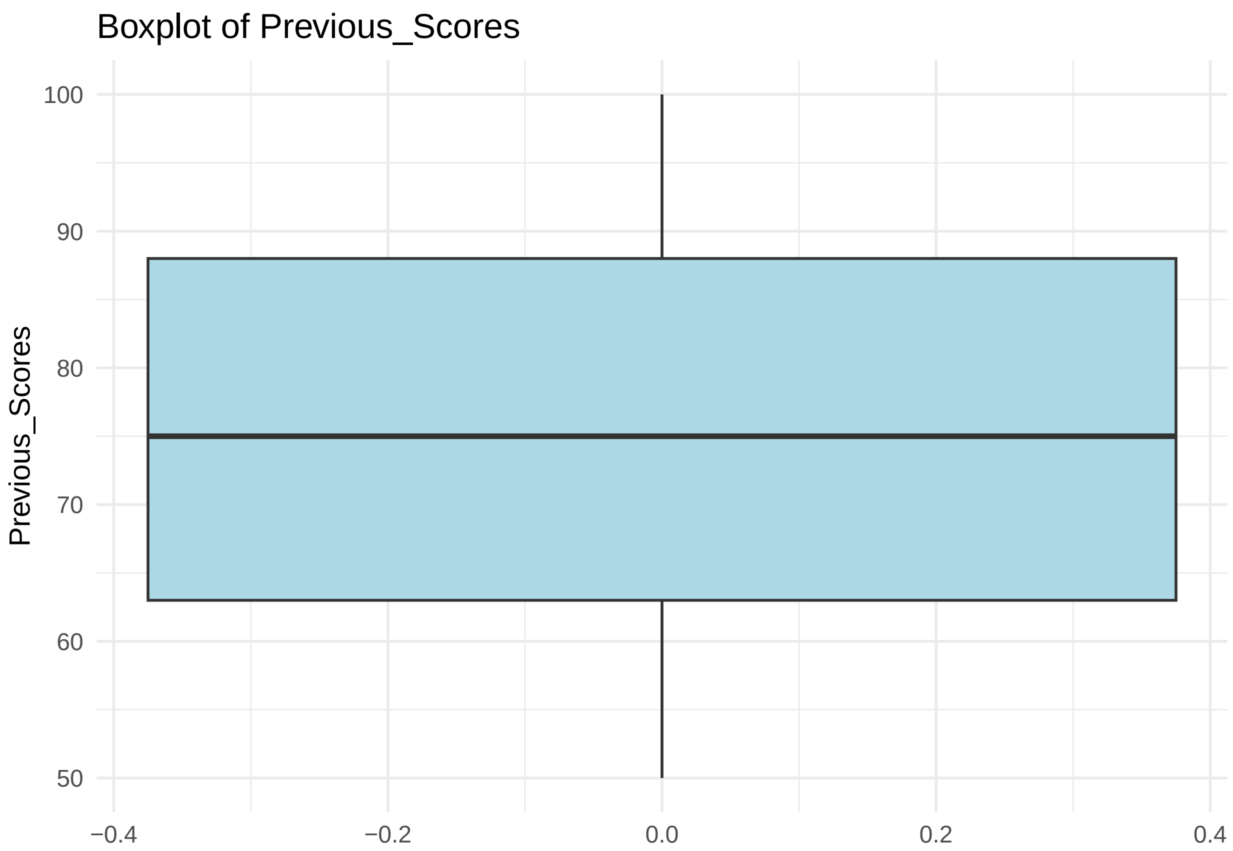
There are outliers in the hours_studied , Tutoring_sessions and the Exam_score columns.

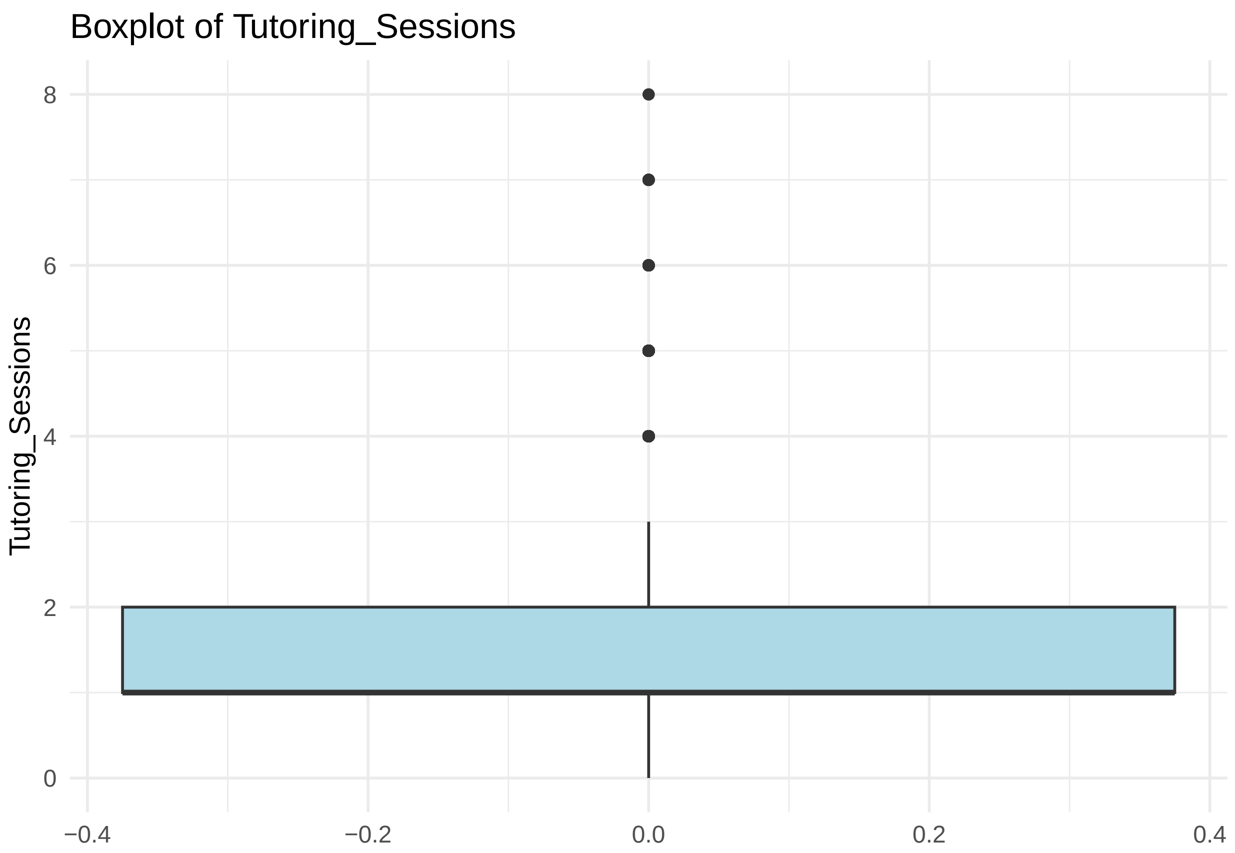
```
#plots showing the outliers
library(ggplot2)

# 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)
}
```

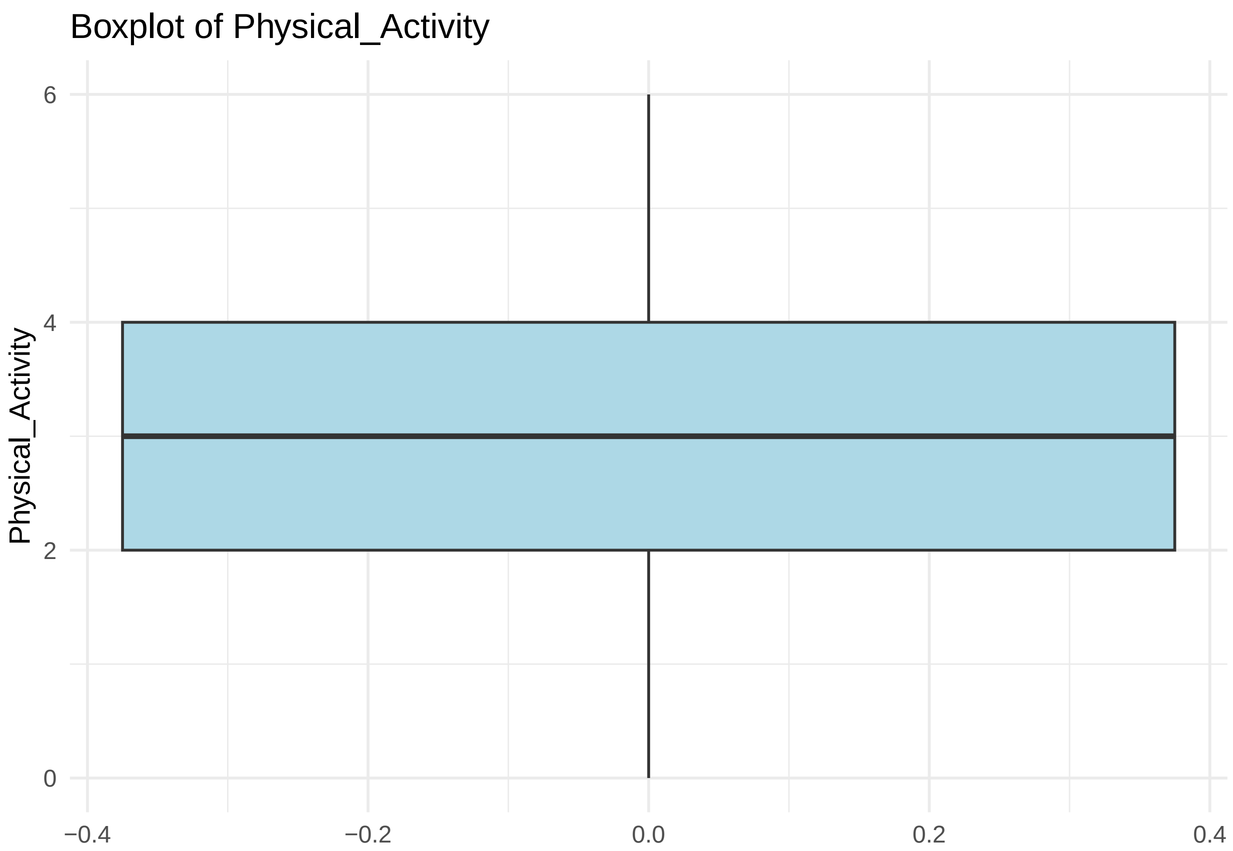

Boxplot of Hours_Studied
Hours_Studied
40
30
20
10
0
−0.4
−0.2
0.0
0.2
0.4

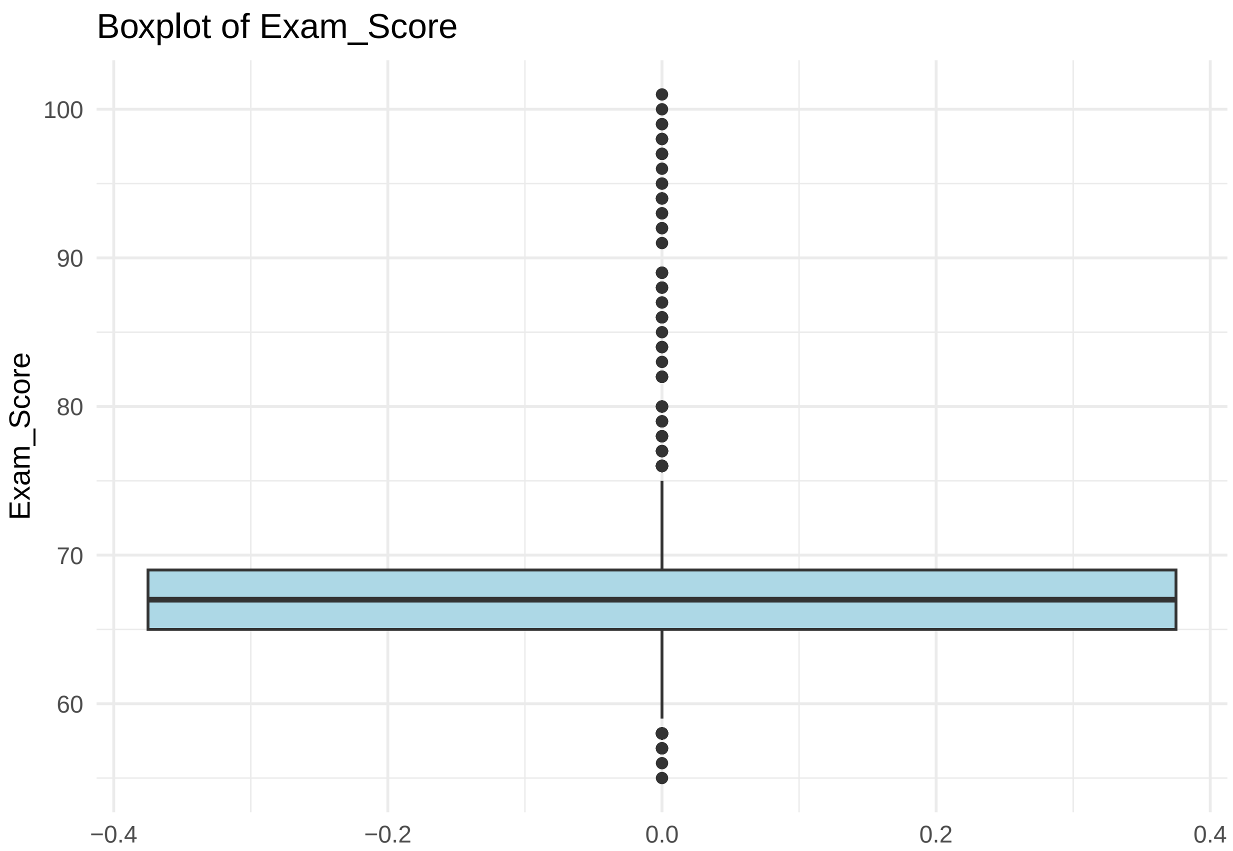

Boxplot of Attendance
Attendance
100
90
80
70
60
−0.4
−0.2
0.0
0.2
0.4


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

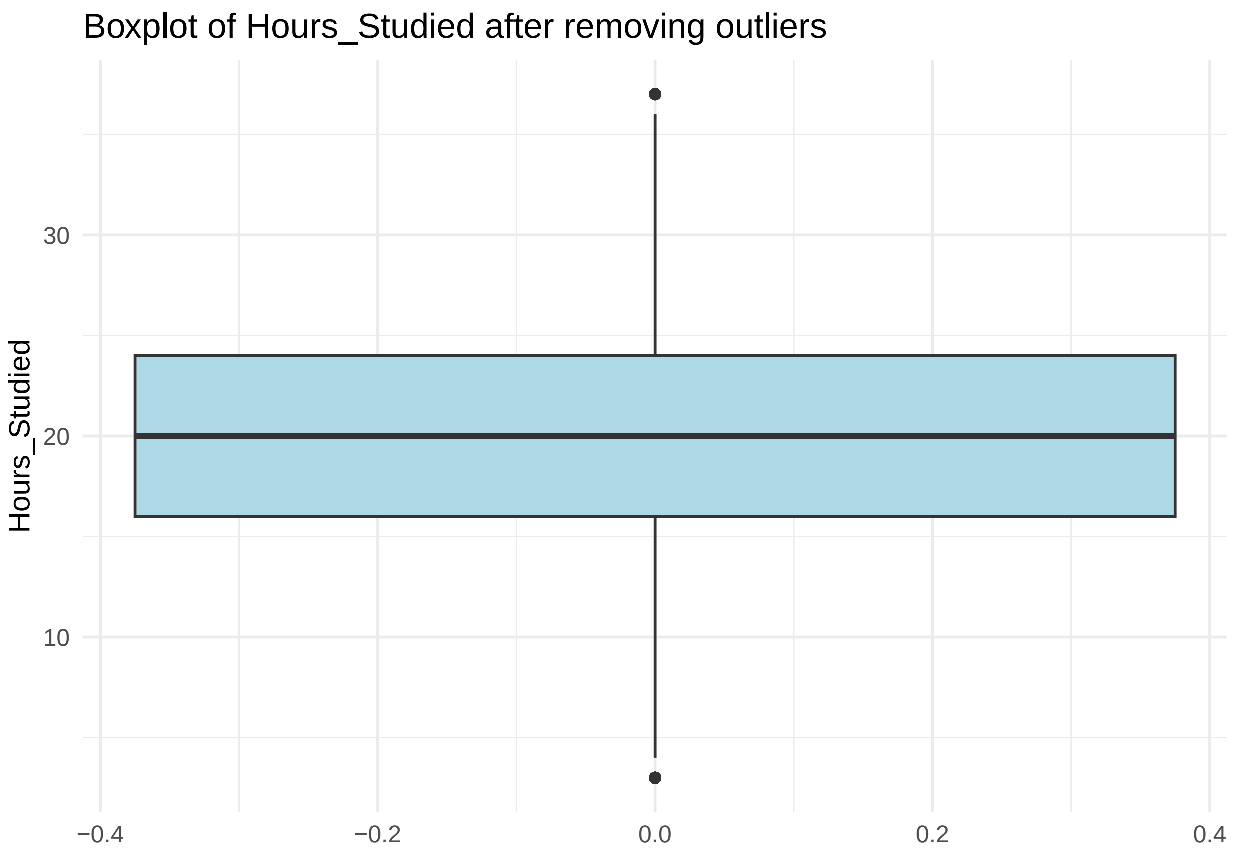

Boxplot of Physical_Activity
Physical_Activity
6
4
2
0
−0.4
−0.2
0.0
0.2
0.4

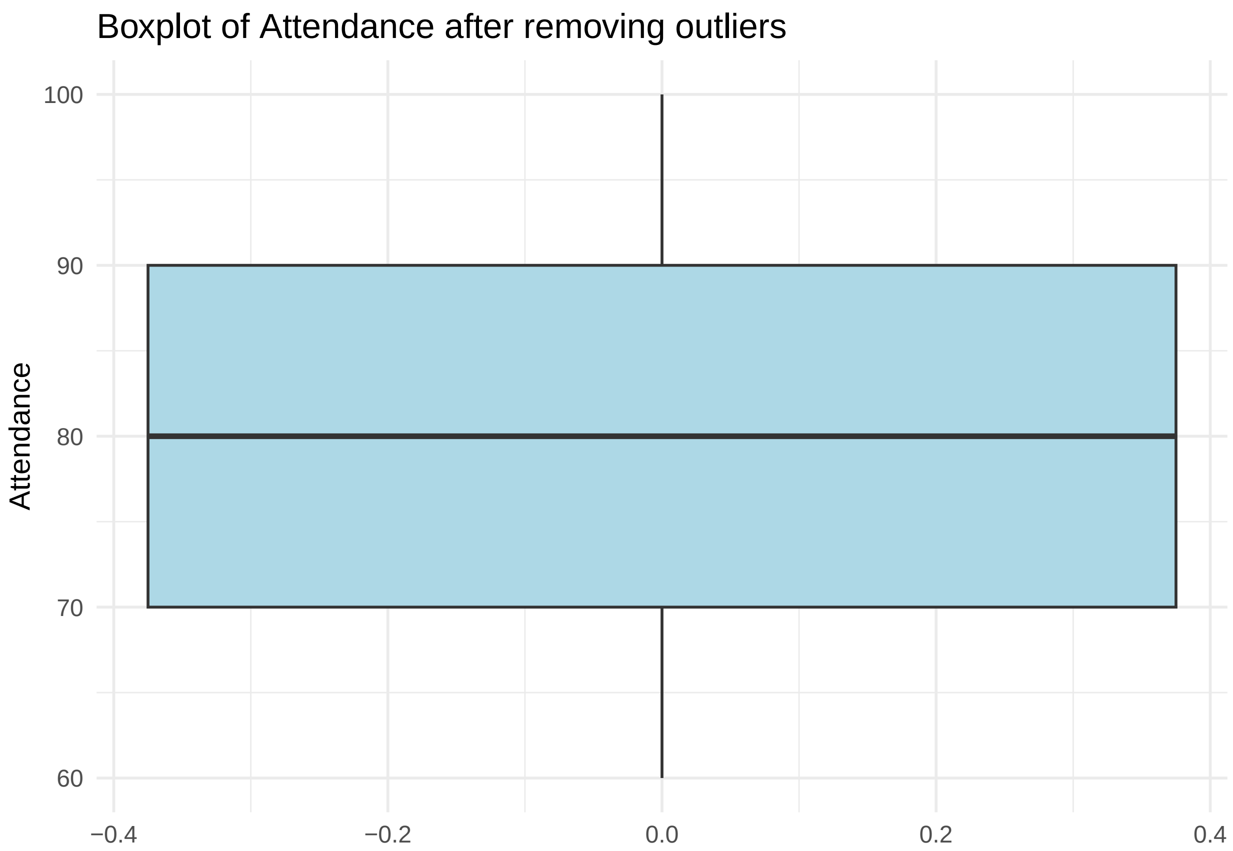


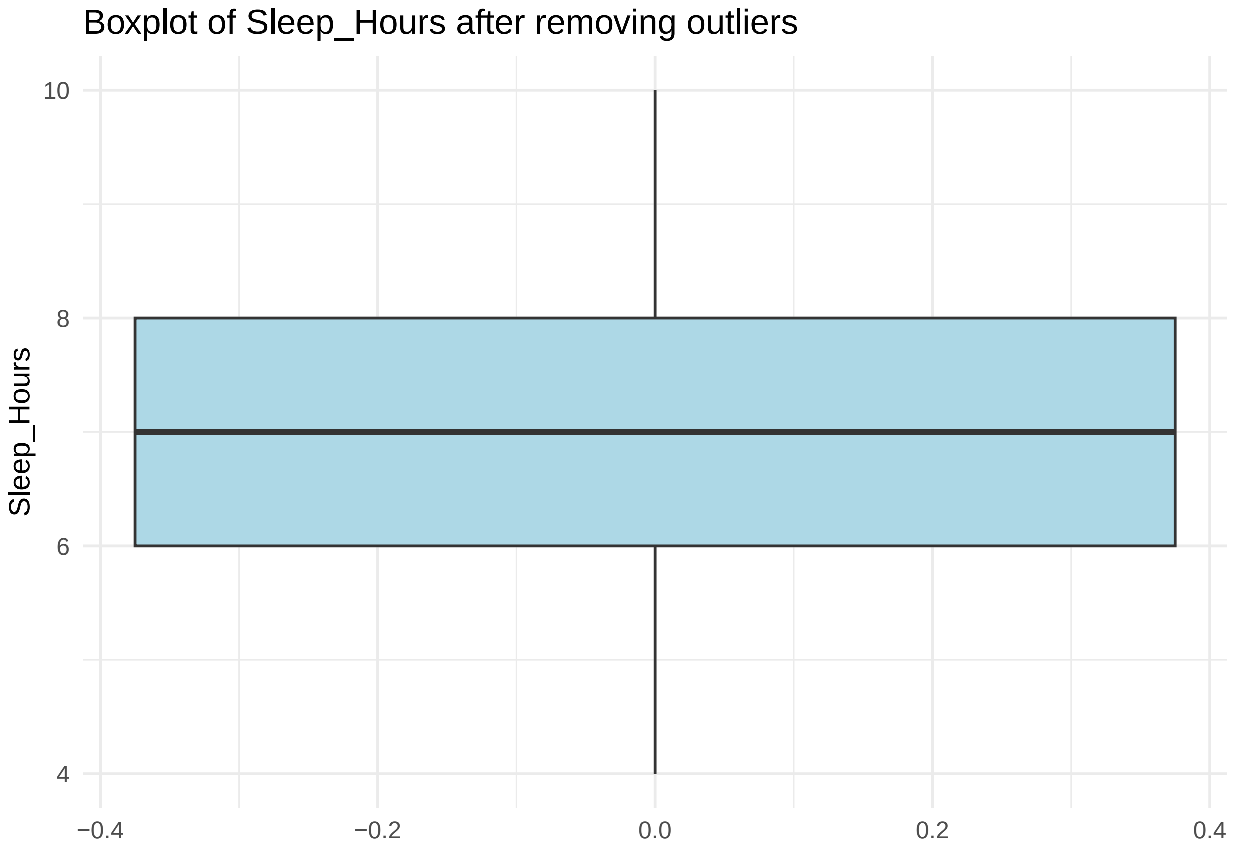
Outliers:Boxplots for Hours Studied, Attendance, Tutoring Sessions, Sleep Hours: Some outliers were identified, especially in hours studied, tutoring sessions, and exam scores.A few students are studying far more (or less) than their peers, and similarly, a small number of students might have abnormal exam scores compared to the rest. These outliers could represent exceptional or struggling students.

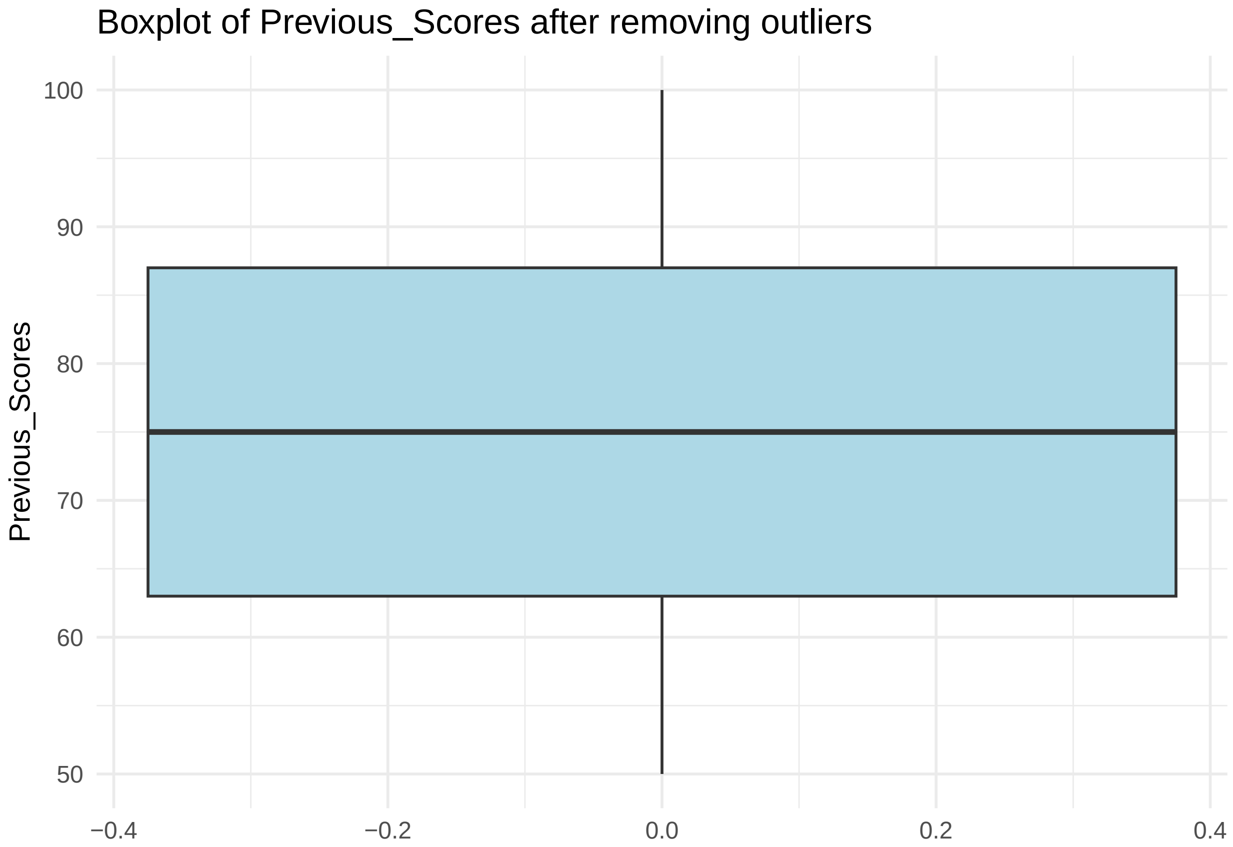
```
#using Z_scores to remove the outliers
z_scores <- scale(project_data[sapply(project_data, is.numeric)])
project_data <- project_data[apply(z_scores, 1, function(x) all(abs(x) <= 3)), ]
```

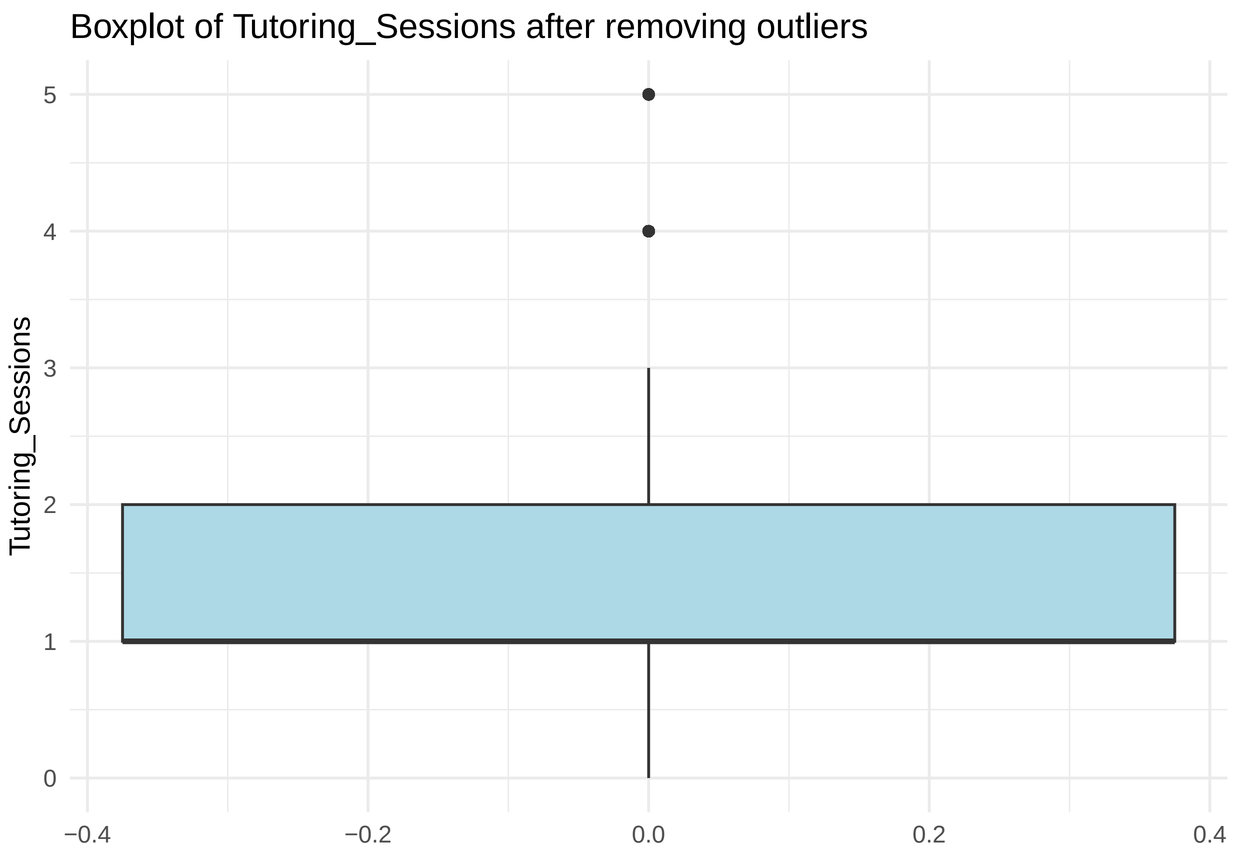
```
#plots to remove the outliers
library(ggplot2)

# 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)
}
```

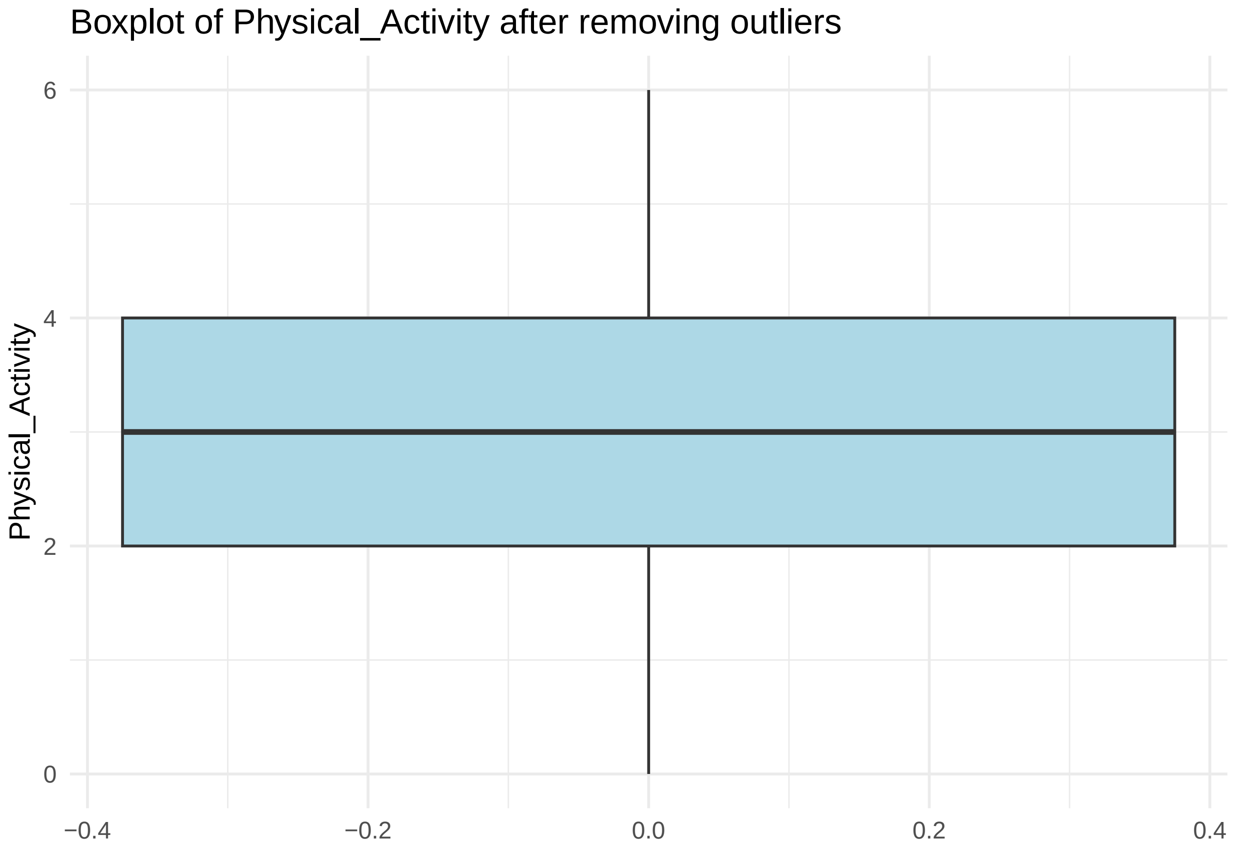

Boxplot of Hours_Studied after removing outliers
Hours_Studied
30
20
10
−0.4
−0.2
0.0
0.2
0.4

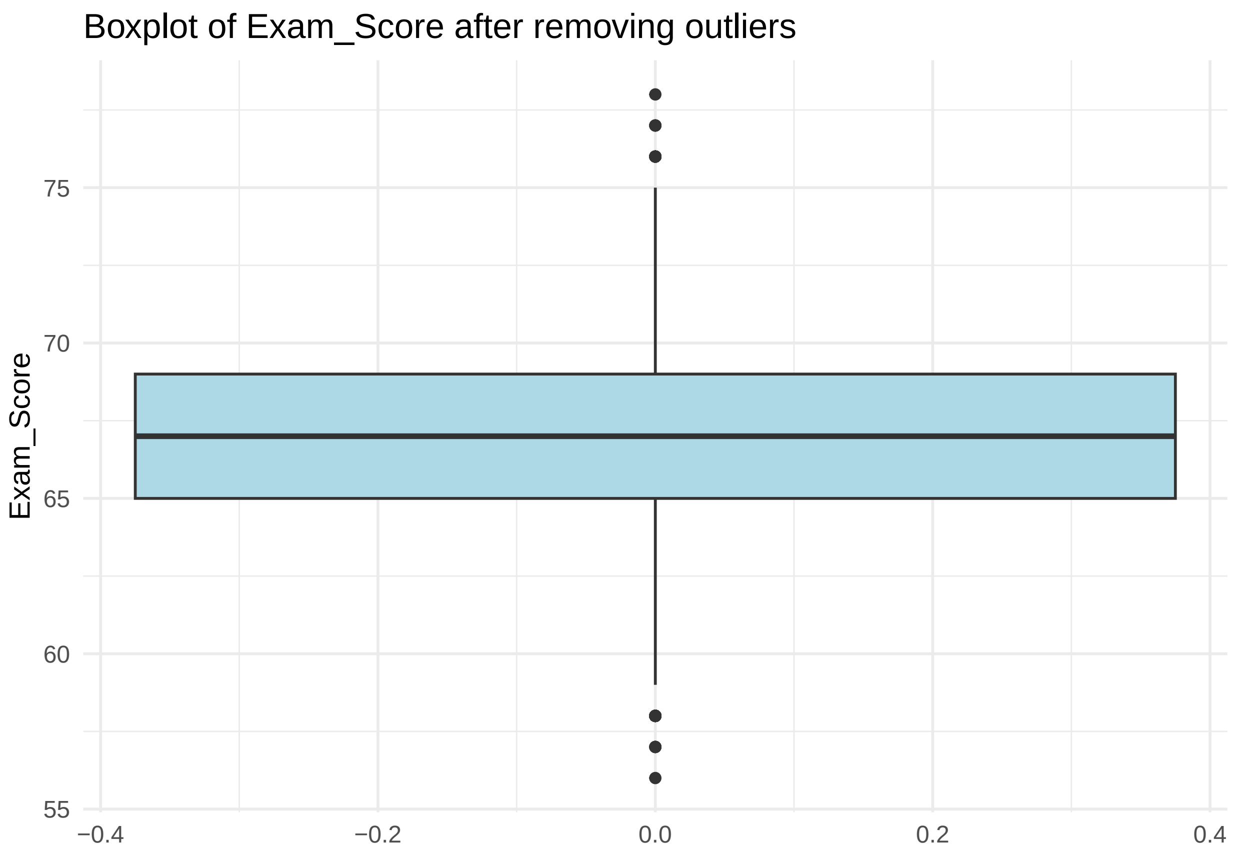

Boxplot of Attendance after removing outliers
Attendance
100
90
80
70
60
−0.4
−0.2
0.0
0.2
0.4


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
5
4
3
2
1
0
−0.4
−0.2
0.0
0.2
0.4

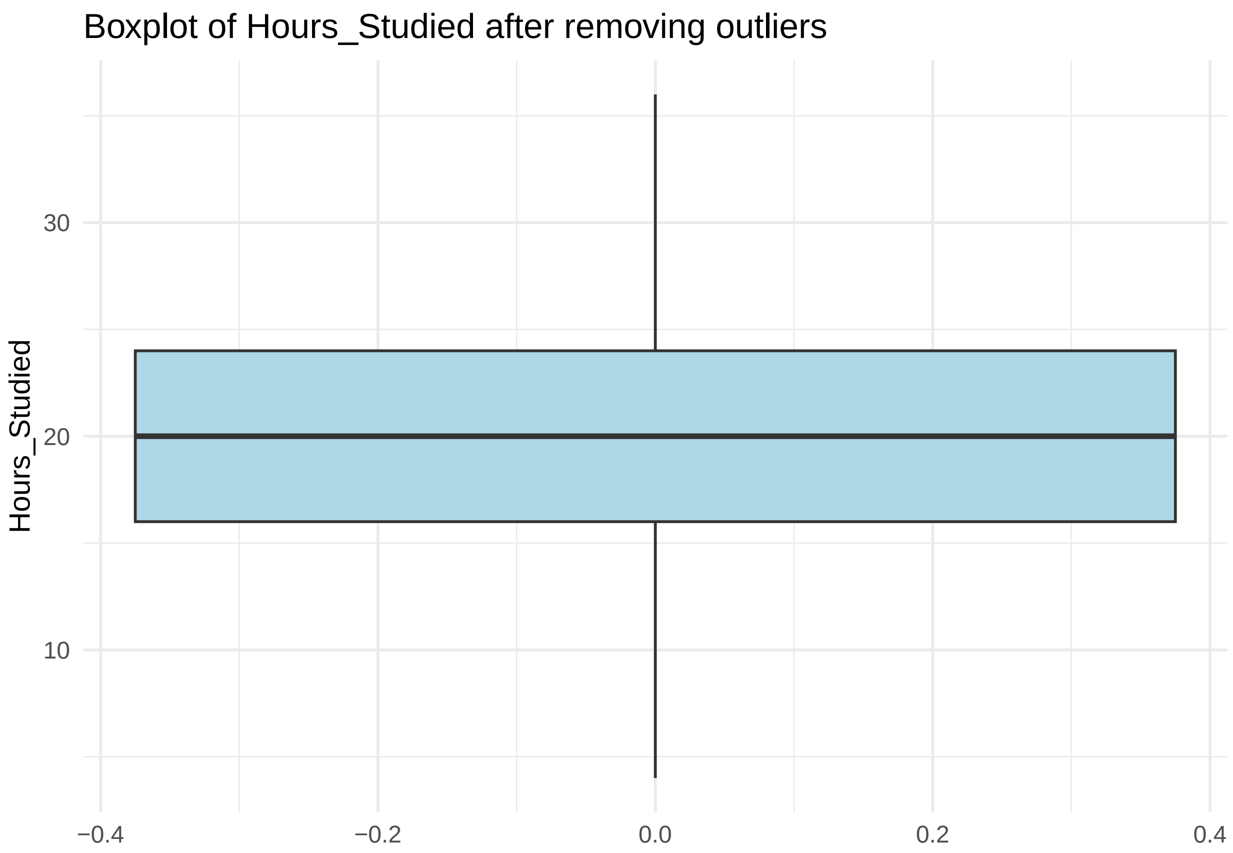

Boxplot of Physical_Activity after removing outliers
Physical_Activity
6
4
2
0
−0.4
−0.2
0.0
0.2
0.4

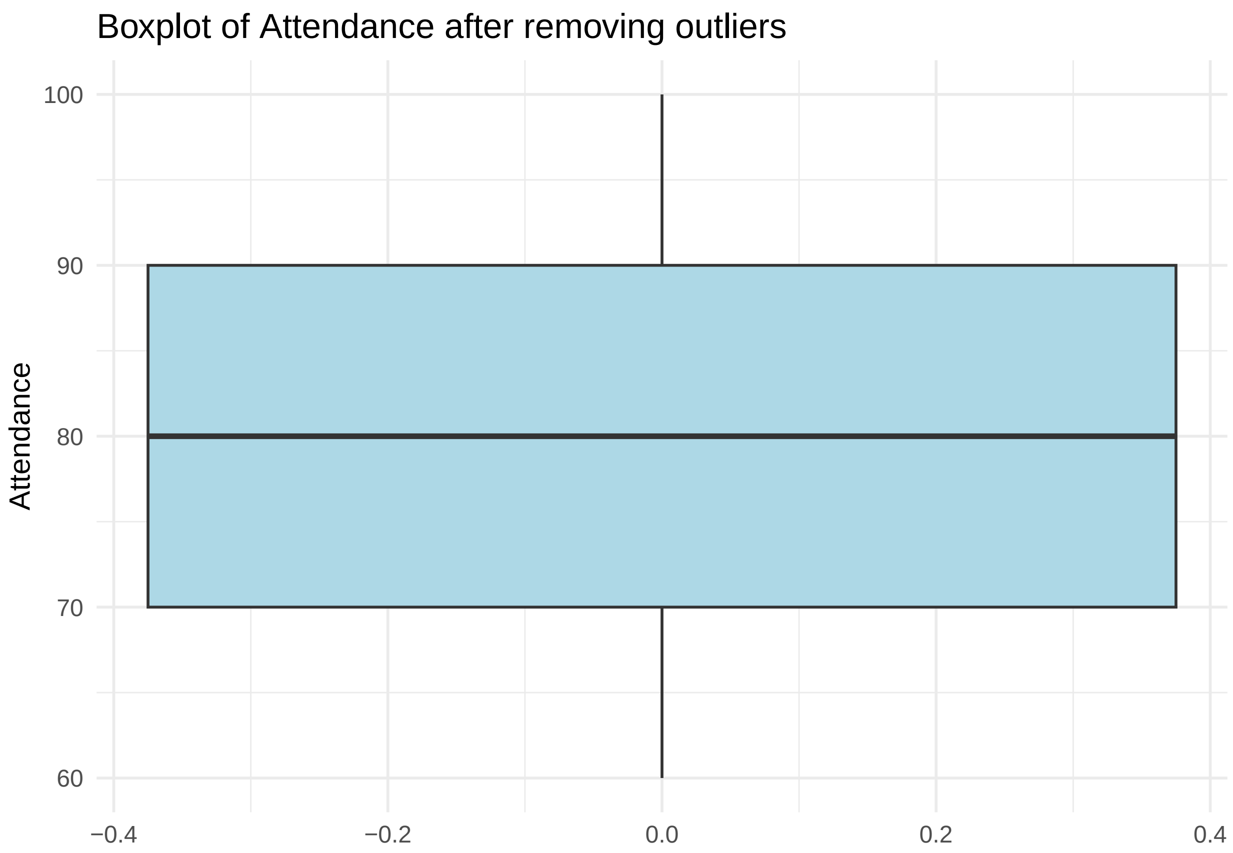


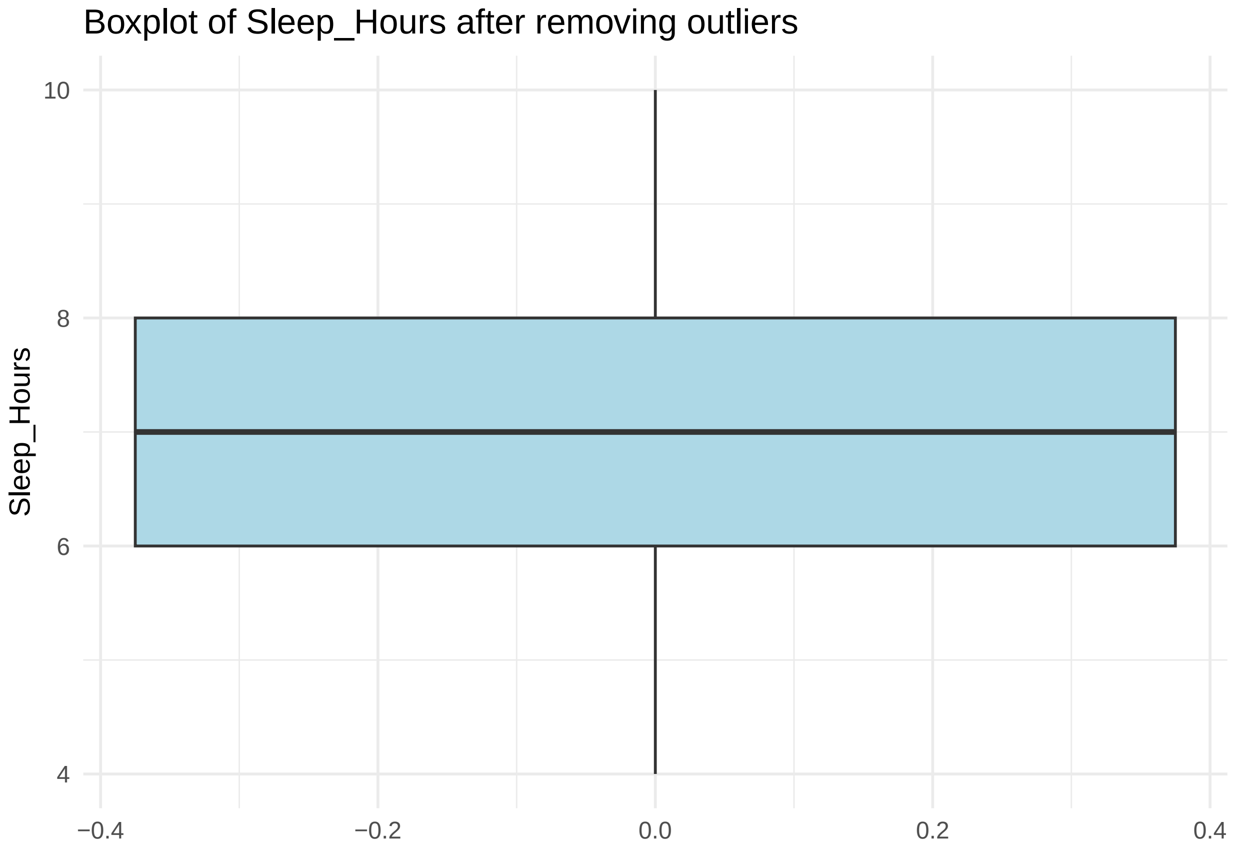
Still we can find the outliers in the particular columns like hours_studied , Tutoring_sessions and the Exam_score.

```r
#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
}
```

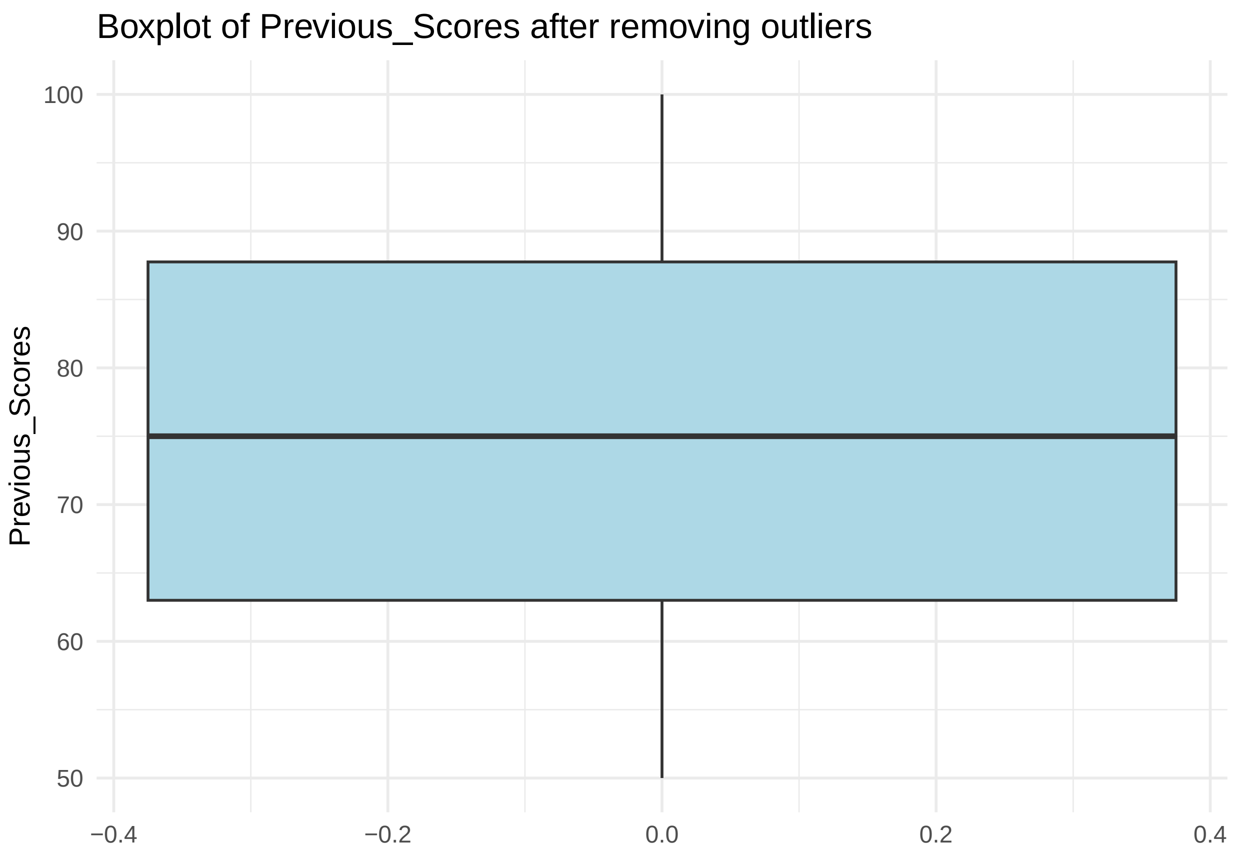
```r
#plots to remove outliers
library(ggplot2)

# 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)
}
```

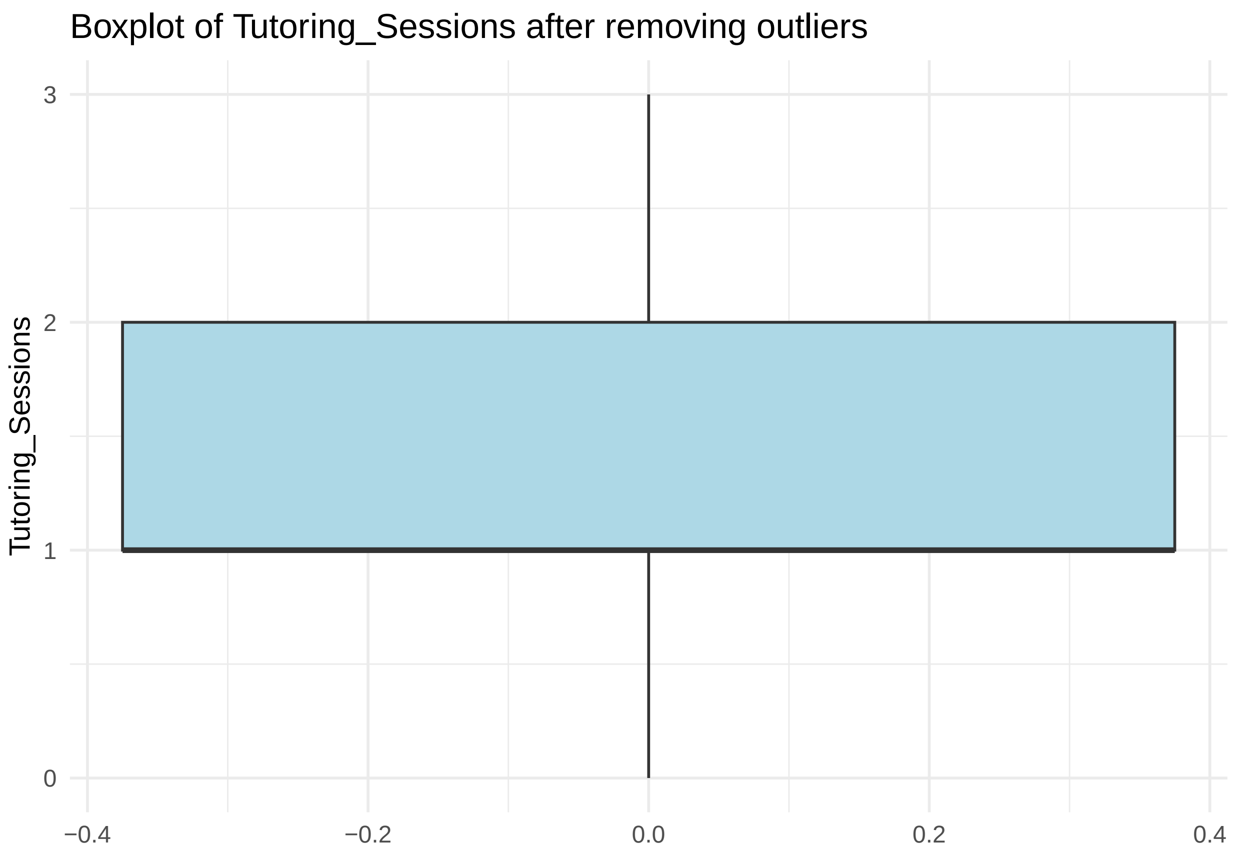

Boxplot of Hours_Studied after removing outliers
Hours_Studied
30
20
10
−0.4
−0.2
0.0
0.2
0.4

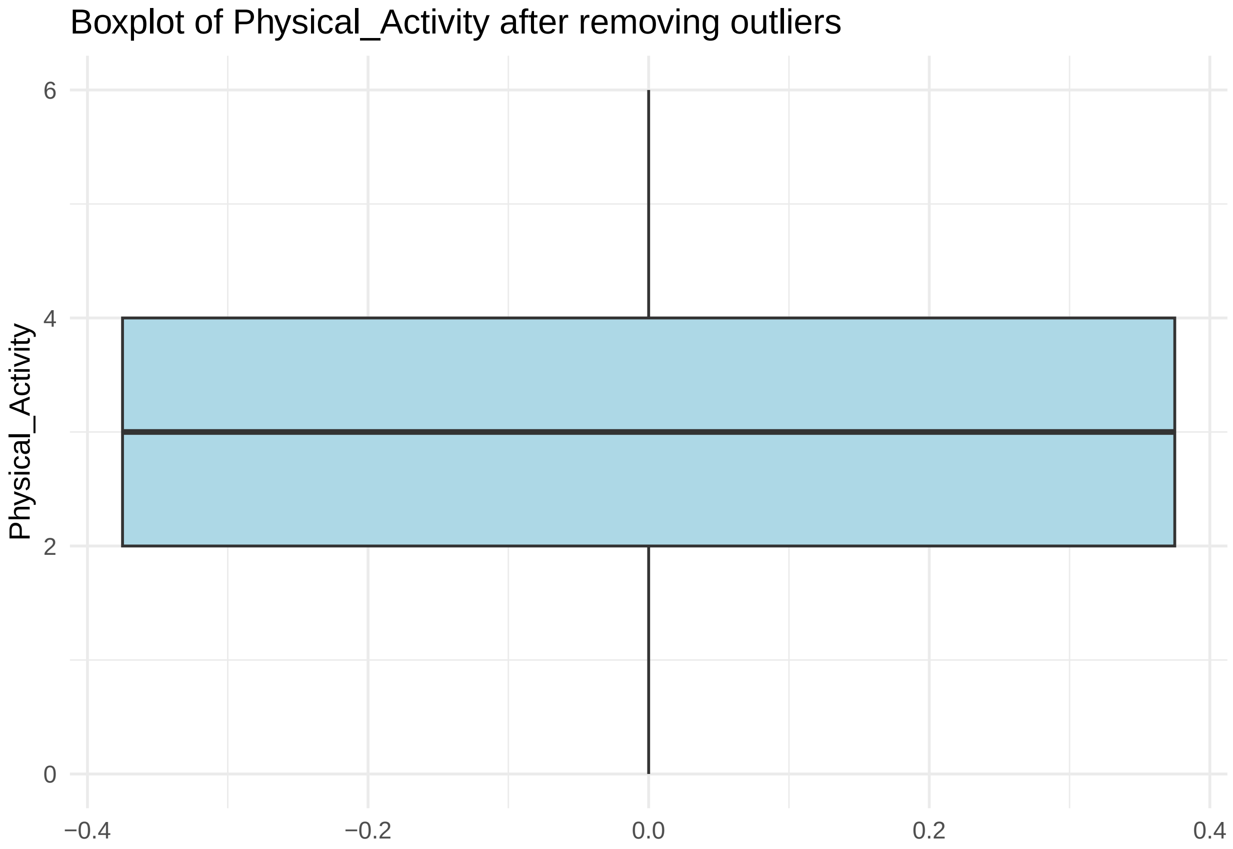

Boxplot of Attendance after removing outliers
Attendance
100
90
80
70
60
−0.4
−0.2
0.0
0.2
0.4

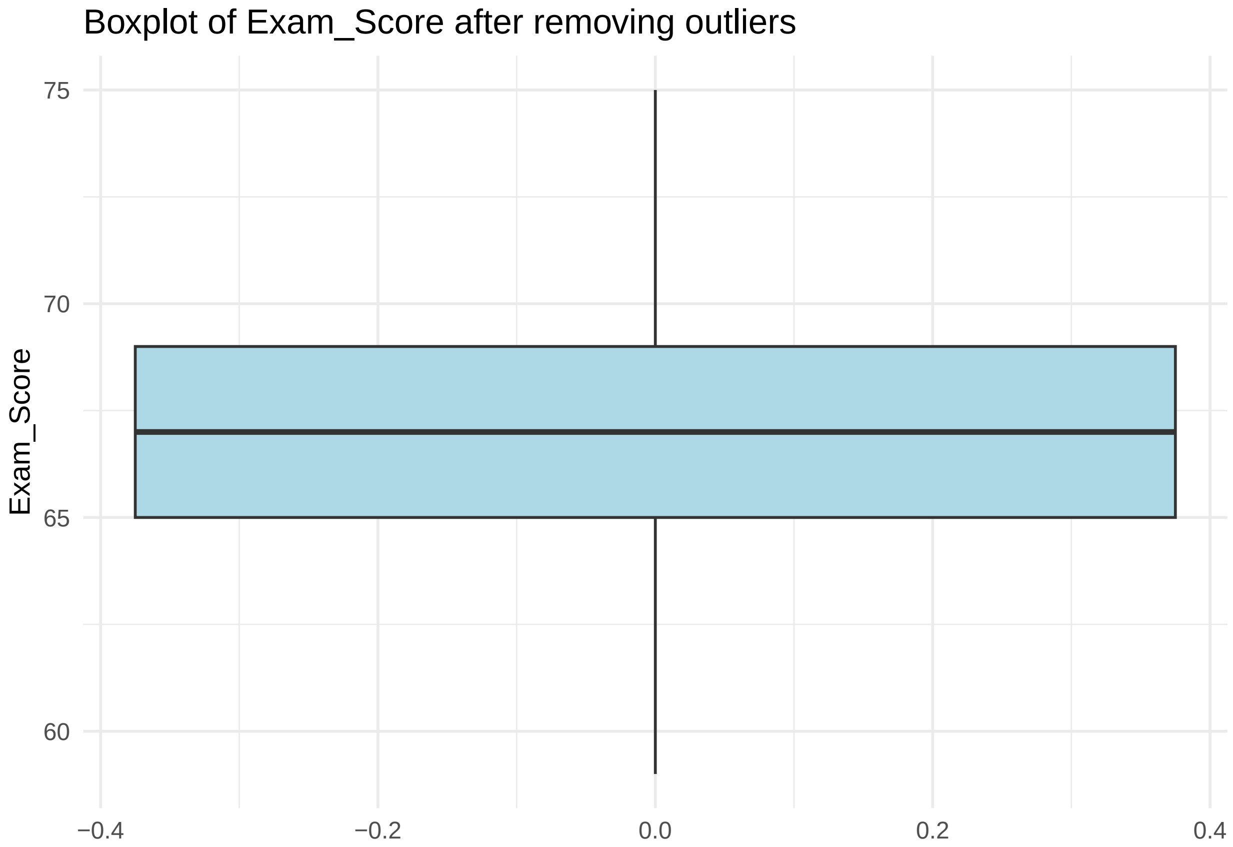


Boxplot of Sleep_Hours after removing outliers


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



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.

```
install.packages("corrplot")
```

```
## Installing package into 'C:/Users/nandi/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'corrplot' successfully unpacked and MD5 sums checked
##
```

```
## The downloaded binary packages are in
##  C:\Users\nandi\AppData\Local\Temp\RtmpaiXbDA\downloaded_packages
```

```
library(corrplot)
```

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

```
## corrplot 0.95 loaded
```

```
install.packages("corrplot", repos = "https://cran.r-project.org")
```

```
## Warning: package 'corrplot' is in use and will not be installed
```

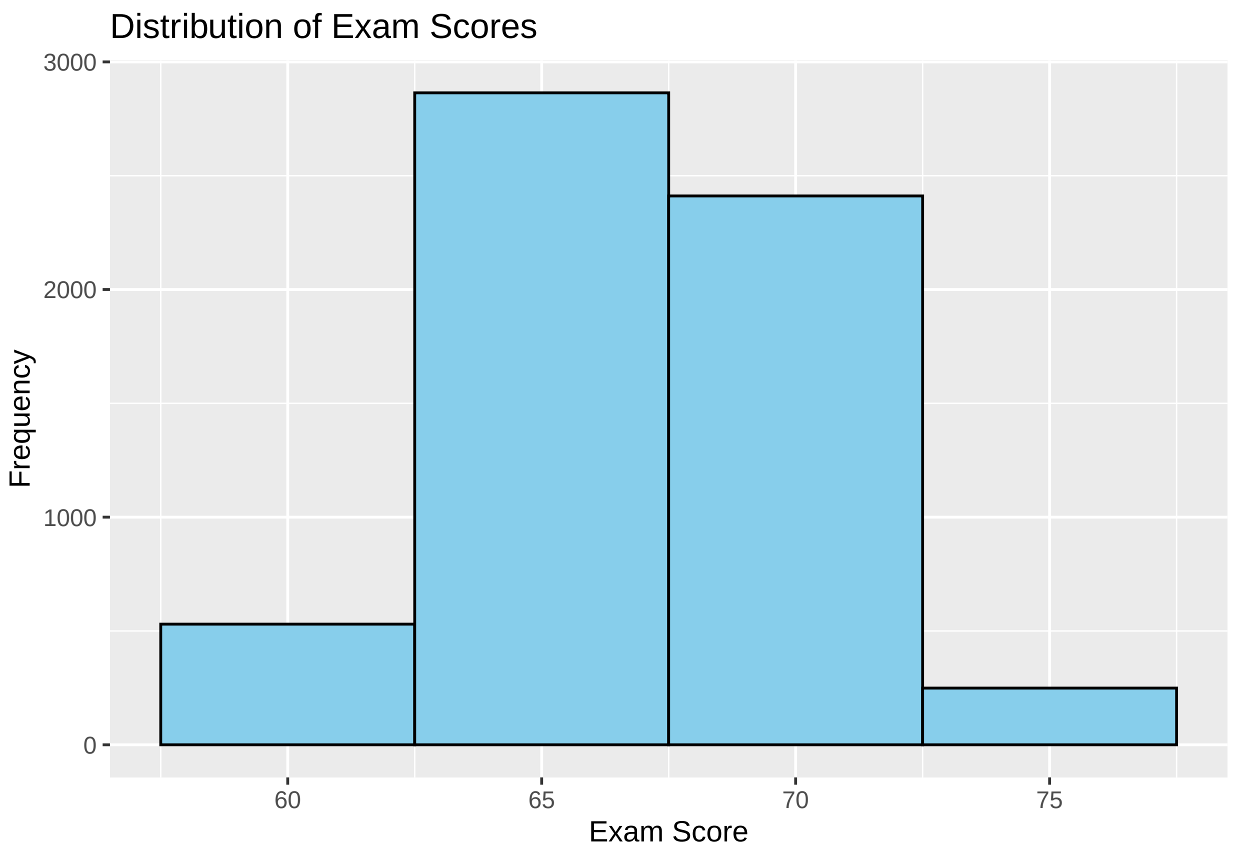
```
# Assuming 'project_data' is your data frame
# Load necessary libraries
library(ggplot2)
library(dplyr)
```

```
##
## 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
```

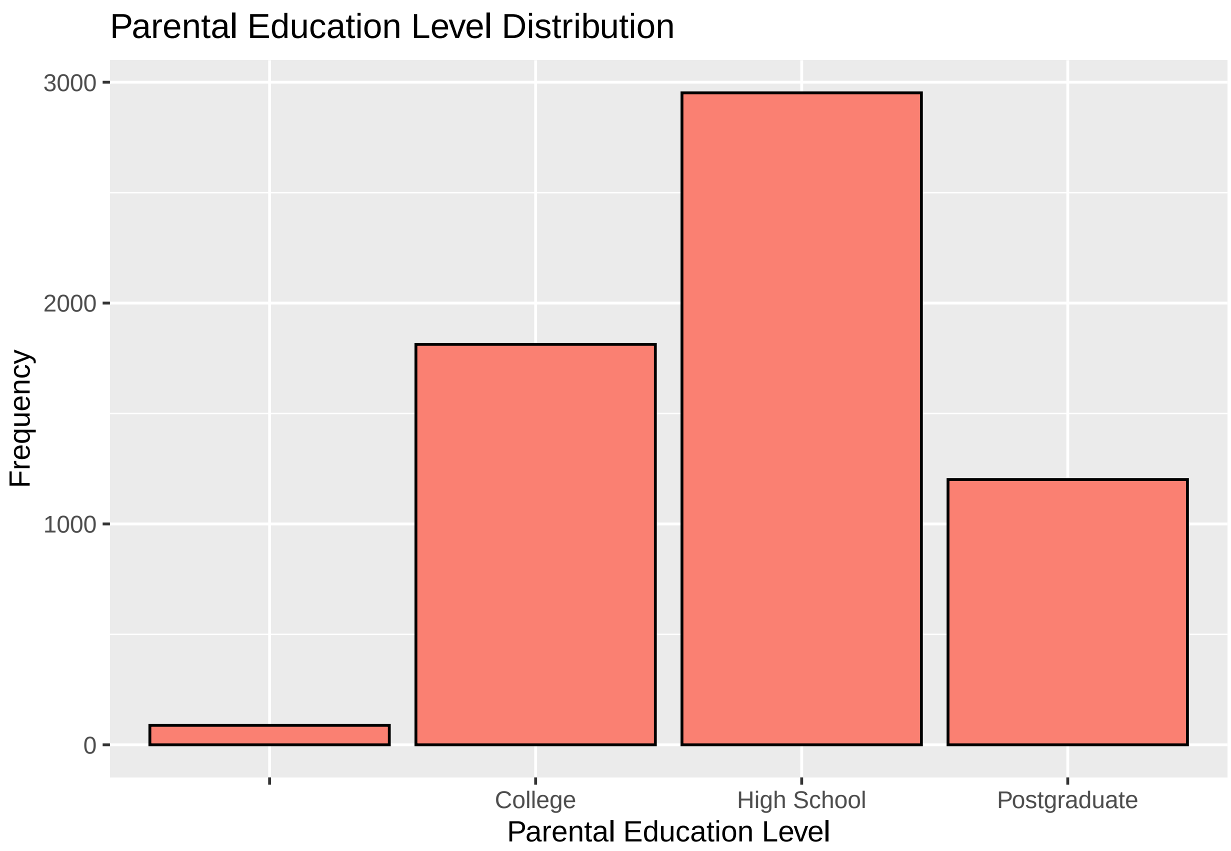
```
# 1. Univariate Analysis
# Histogram of Exam_Score
ggplot(project_data, aes(x = Exam_Score)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Exam Scores", x = "Exam Score", y = "Frequency")
```



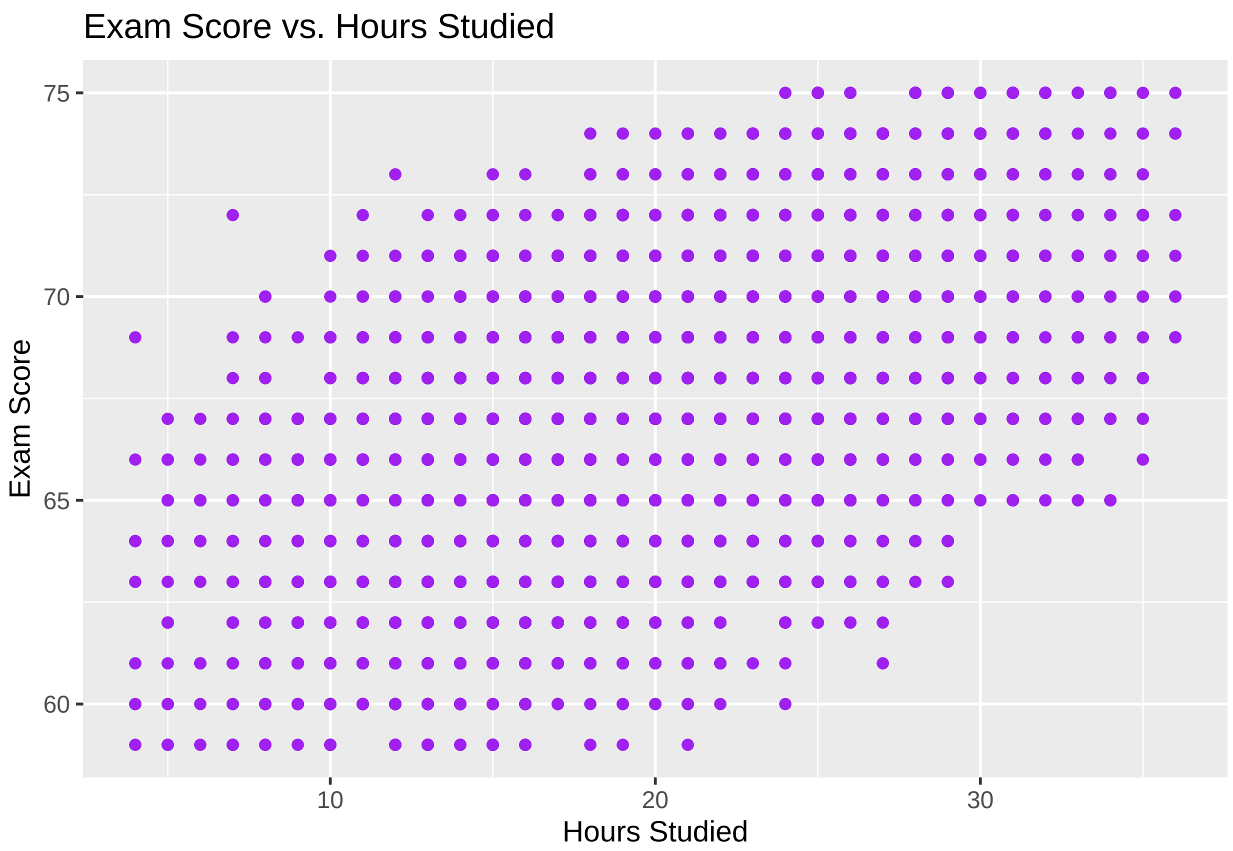
```r
# Summary statistics for Hours_Studied
summary(project_data$Hours_Studied)
```

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

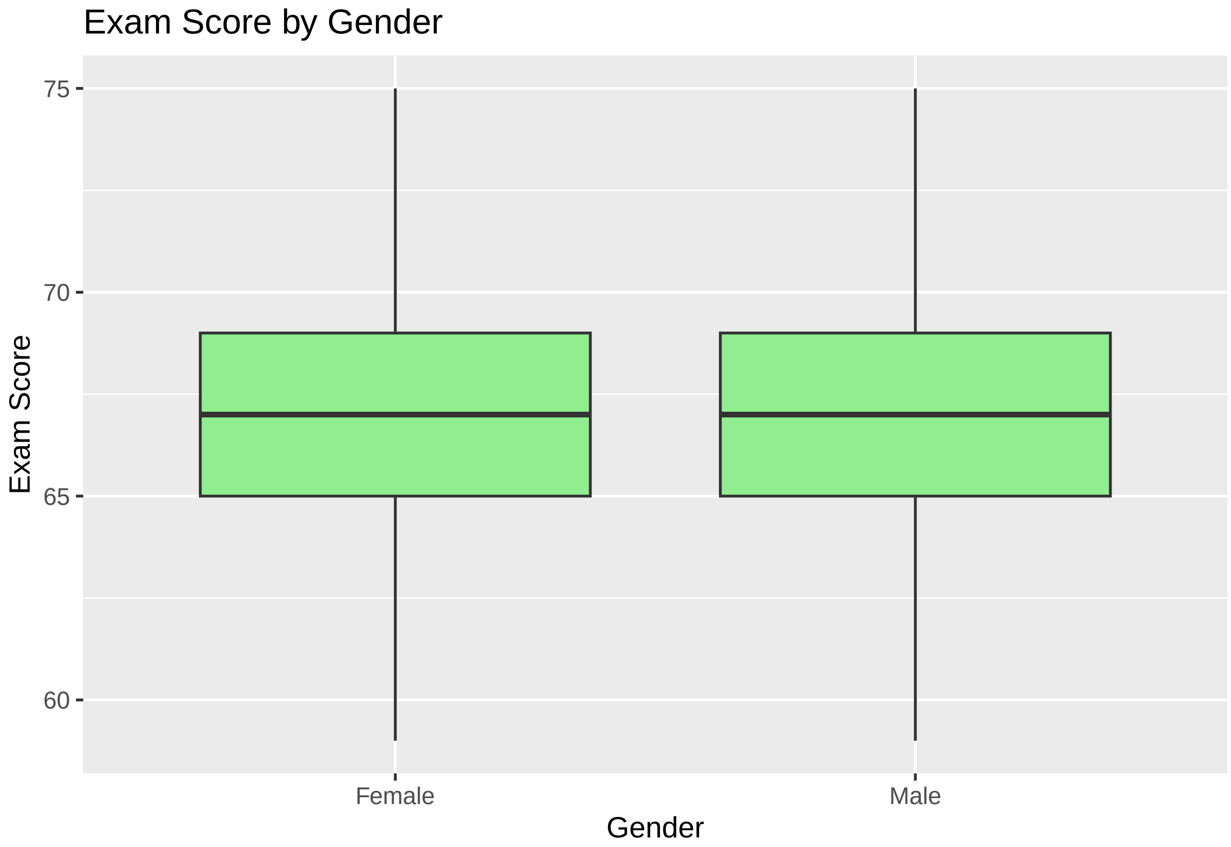
```r
# Bar plot for Parental Education Level
ggplot(project_data, aes(x = Parental_Education_Level)) +
  geom_bar(fill = "salmon", color = "black") +
  labs(title = "Parental Education Level Distribution", x = "Parental Education Level", y = "Frequency")
```



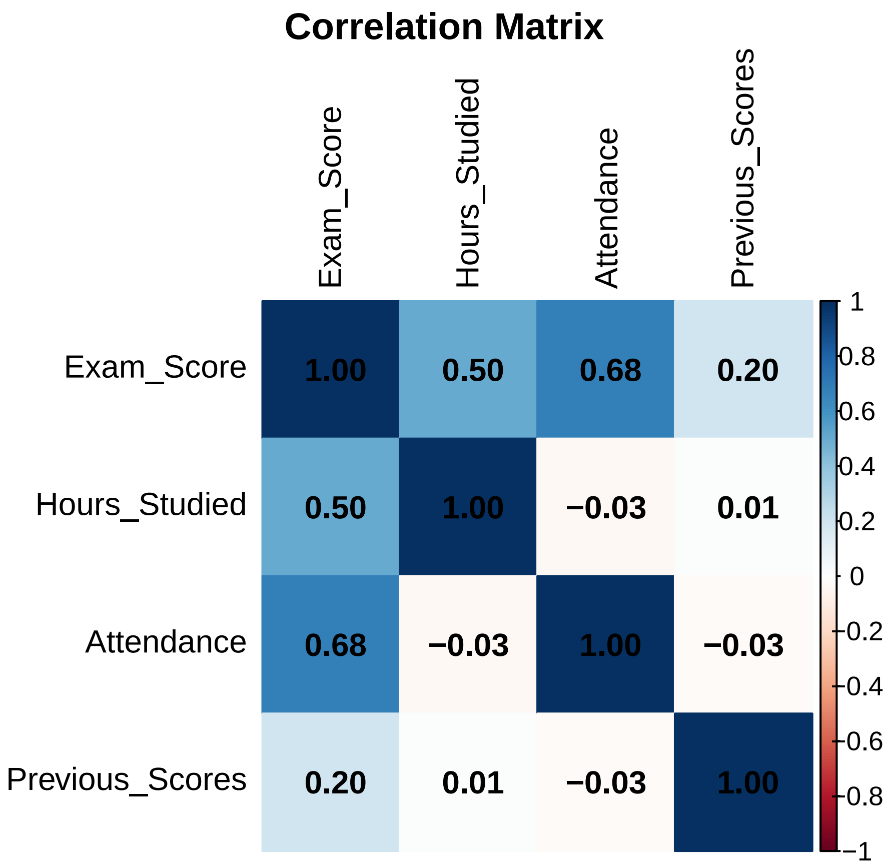
```
# 2. Bivariate Analysis
# Scatter plot of Exam_Score vs. Hours_Studied
ggplot(project_data, aes(x = Hours_Studied, y = Exam_Score)) +
  geom_point(color = "purple") +
  labs(title = "Exam Score vs. Hours Studied", x = "Hours Studied", y = "Exam Score")
```



```
# 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")
```



```r
# 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)
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). . Scatter Plot of Hours Studied vs Exam Scores: The scatter plot demonstrates a positive linear relationship between hours studied and exam scores, with students who study more generally achieving higher scores.

.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, with fewer outliers.

.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. .This scatter plot shows the relationship between hours studied and exam scores. Generally, as students study more hours, their exam scores tend to increase. The dots form an upward pattern, showing a positive link between studying time and exam performance.

```r
# Fit the linear regression model
model <- lm(Exam_Score ~ ., data = project_data)

# Summarize the model to see the significance of each factor
summary(model)
```

```
##
## Call:
```

```
## lm(formula = Exam_Score ~ ., data = project_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.23358 -0.25226 -0.00034  0.25284  1.17216
##
## Coefficients:
##                                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                          41.1614202  0.0818914  502.635  < 2e-16
## Hours_Studied                         0.2980540  0.0007200  413.967  < 2e-16
## Attendance                            0.1999037  0.0003622  551.875  < 2e-16
## Parental_InvolvementLow              -1.9925537  0.0121063 -164.588  < 2e-16
## Parental_InvolvementMedium           -1.0036454  0.0096821 -103.660  < 2e-16
## Access_to_ResourcesLow               -1.9962730  0.0120464 -165.716  < 2e-16
## Access_to_ResourcesMedium            -0.9815247  0.0096106 -102.129  < 2e-16
## Extracurricular_ActivitiesYes         0.5076623  0.0084722   59.921  < 2e-16
## Sleep_Hours                           0.0156972  0.0028269    5.553 2.93e-08
## Previous_Scores                       0.0494658  0.0002899  170.616  < 2e-16
## Motivation_LevelLow                  -1.0152365  0.0120778  -84.058  < 2e-16
## Motivation_LevelMedium               -0.4995216  0.0110032  -45.398  < 2e-16
## Internet_AccessYes                    1.0028872  0.0156554   64.060  < 2e-16
## Tutoring_Sessions                     0.4990209  0.0042243  118.130  < 2e-16
## Family_IncomeLow                     -0.9868852  0.0115502  -85.443  < 2e-16
## Family_IncomeMedium                  -0.4908413  0.0115487  -42.502  < 2e-16
## Teacher_QualityHigh                   0.3198657  0.0388785    8.227 2.33e-16
## Teacher_QualityLow                   -0.6837376  0.0403784  -16.933  < 2e-16
## Teacher_QualityMedium                -0.1789921  0.0384967   -4.650 3.40e-06
## School_TypePublic                    -0.0035462  0.0090419   -0.392    0.695
## Peer_InfluenceNeutral                 0.4936084  0.0112747   43.780  < 2e-16
## Peer_InfluencePositive                0.9976704  0.0112358   88.794  < 2e-16
## Physical_Activity                     0.2365914  0.0040487   58.436  < 2e-16
## Learning_DisabilitiesYes             -0.9990772  0.0136829  -73.017  < 2e-16
## Parental_Education_LevelCollege       0.1572715  0.0353279    4.452 8.67e-06
## Parental_Education_LevelHigh School  -0.3399338  0.0350117   -9.709  < 2e-16
## Parental_Education_LevelPostgraduate  0.6695683  0.0357373   18.736  < 2e-16
## Distance_from_HomeFar                -0.7432570  0.0425300  -17.476  < 2e-16
## Distance_from_HomeModerate           -0.2507359  0.0411476   -6.094 1.17e-09
## Distance_from_HomeNear                0.2549456  0.0408084    6.247 4.46e-10
## GenderMale                           -0.0095182  0.0084071   -1.132    0.258
##
## (Intercept)                          ***
## Hours_Studied                        ***
## Attendance                           ***
## Parental_InvolvementLow              ***
## Parental_InvolvementMedium           ***
## Access_to_ResourcesLow               ***
## Access_to_ResourcesMedium            ***
## Extracurricular_ActivitiesYes        ***
## Sleep_Hours                          ***
## Previous_Scores                      ***
## Motivation_LevelLow                  ***
## Motivation_LevelMedium               ***
## Internet_AccessYes                   ***
## Tutoring_Sessions                    ***
## Family_IncomeLow                     ***
## Family_IncomeMedium                  ***
## Teacher_QualityHigh                  ***
## Teacher_QualityLow                   ***
## Teacher_QualityMedium                ***
## School_TypePublic
## Peer_InfluenceNeutral                ***
## Peer_InfluencePositive               ***
## Physical_Activity                    ***
## Learning_DisabilitiesYes             ***
## Parental_Education_LevelCollege      ***
## Parental_Education_LevelHigh School  ***
## Parental_Education_LevelPostgraduate ***
## Distance_from_HomeFar                ***
## Distance_from_HomeModerate           ***
## Distance_from_HomeNear               ***
## GenderMale
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3227 on 6023 degrees of freedom
## Multiple R-squared:   0.99,  Adjusted R-squared:   0.99
## F-statistic: 1.996e+04 on 30 and 6023 DF,  p-value: < 2.2e-16
```

Coefficients:The estimate values shows how each variable impacts exam_score. For example: Hours_studied has a positive coefficient(0.298) indicating that more study hours are associated with a higher score. Significance Levels: Most predictors are highly significant (indicated by *** in the Pr(>|t|) column), suggesting they meaningfully impact Exam_Score. Model Fit: The high R-squared (0.99) indicates that this model explains about 99% of the variance in Exam_Score, suggesting a strong fit.