Project Regression Models

Mauricio Gómez Macedo

December 6, 2023

Table of content

Introduction

In data science and statistical analysis, the application of regression models is a powerful tool for understanding and predicting relationships between variables. As part of our exploration into the diverse landscape of statistical methods, this project explores the complexities of implementing multiple linear regression. The focal point of our investigation is a comprehensive dataset capturing various facets of student performance.

The foundation of this project is grounded in the knowledge gained from our coursework, where we have assimilated the theoretical foundations and practical applications of multiple linear regression and statistical methods. As we traverse through this endeavour, we aim to apply these learned principles to unravel the underlying patterns within the dataset, shedding light on the factors influencing student performance.

The chosen dataset encompasses six variables, ranging from sleeping and studying to extracurricular activities. By employing multiple linear regression, we seek to construct a model that describes the relationships between these variables and enables us to make informed predictions about future student performance.

Dataset

The Student Performance dataset comprises 10,000 observations, each encapsulating six features. This dataset promises a comprehensive overview of factors influencing academic performance. We aim to extract meaningful insights using statistical models, particularly multiple linear regression, to discern patterns within this extensive dataset.

Variables

- Hours Studied: The total number of hours spent studying by each student.
- Previous Scores: The scores obtained by students in previous tests.
- Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).
- Sleep Hours: The average number of hours of sleep the student had per day.
- Sample Question Papers Practiced: The number of sample question papers the student practiced.

Target Variable

- Performance Index: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Source: https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression (https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression)

```
df <- read.csv("Student_Performance.csv")
head(df)
```

```
##   Hours.Studied Previous.Scores Extracurricular.Activities Sleep.Hours
## 1             7              99                         Yes           9
## 2             4              82                          No           4
## 3             8              51                         Yes           7
## 4             5              52                         Yes           5
## 5             7              75                          No           8
## 6             3              78                          No           9
##   Sample.Question.Papers.Practiced Performance.Index
## 1                                1                91
## 2                                2                65
## 3                                2                45
## 4                                2                36
## 5                                5                66
## 6                                6                61
```

```
nrow(df)
```

```
## [1] 10000
```

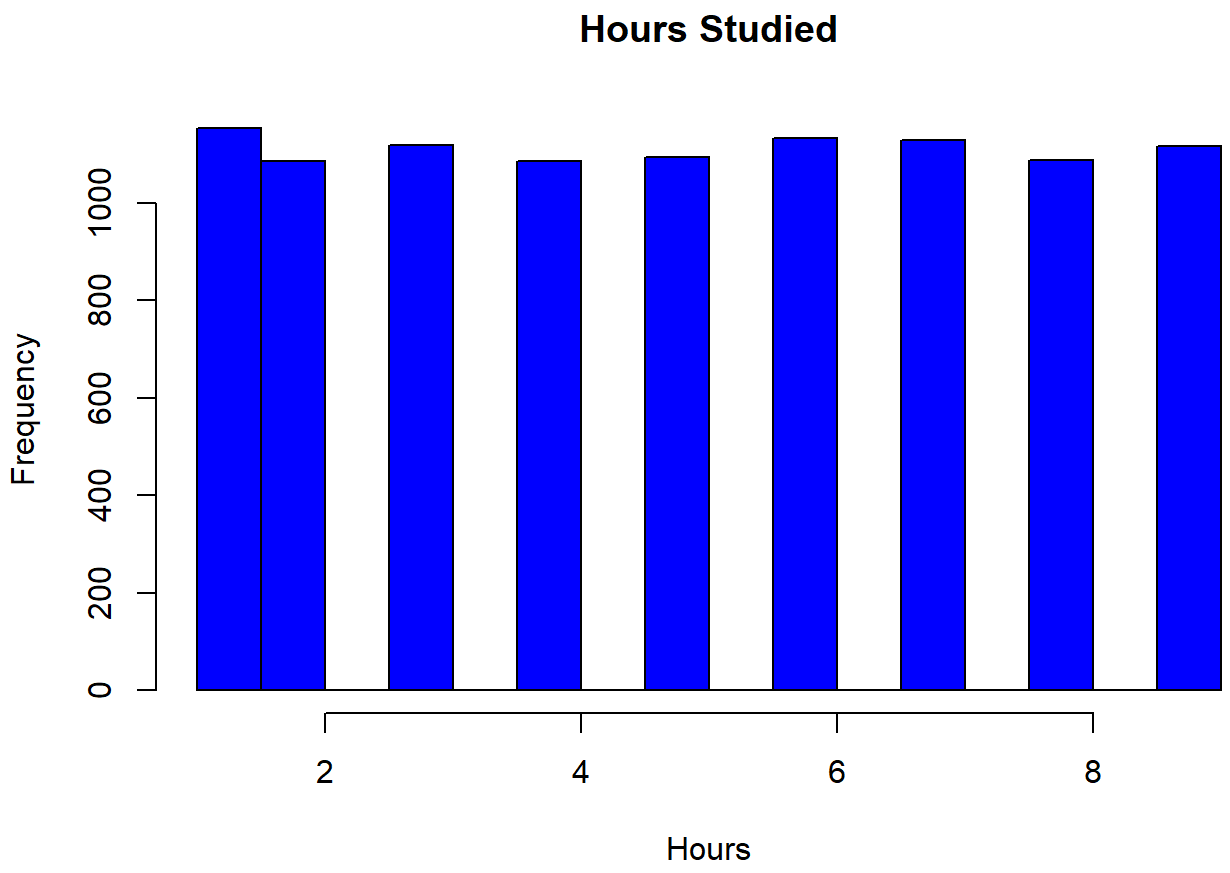
```
colnames(df)
```

```
## [1] "Hours.Studied"                   "Previous.Scores"
## [3] "Extracurricular.Activities"      "Sleep.Hours"
## [5] "Sample.Question.Papers.Practiced" "Performance.Index"
```

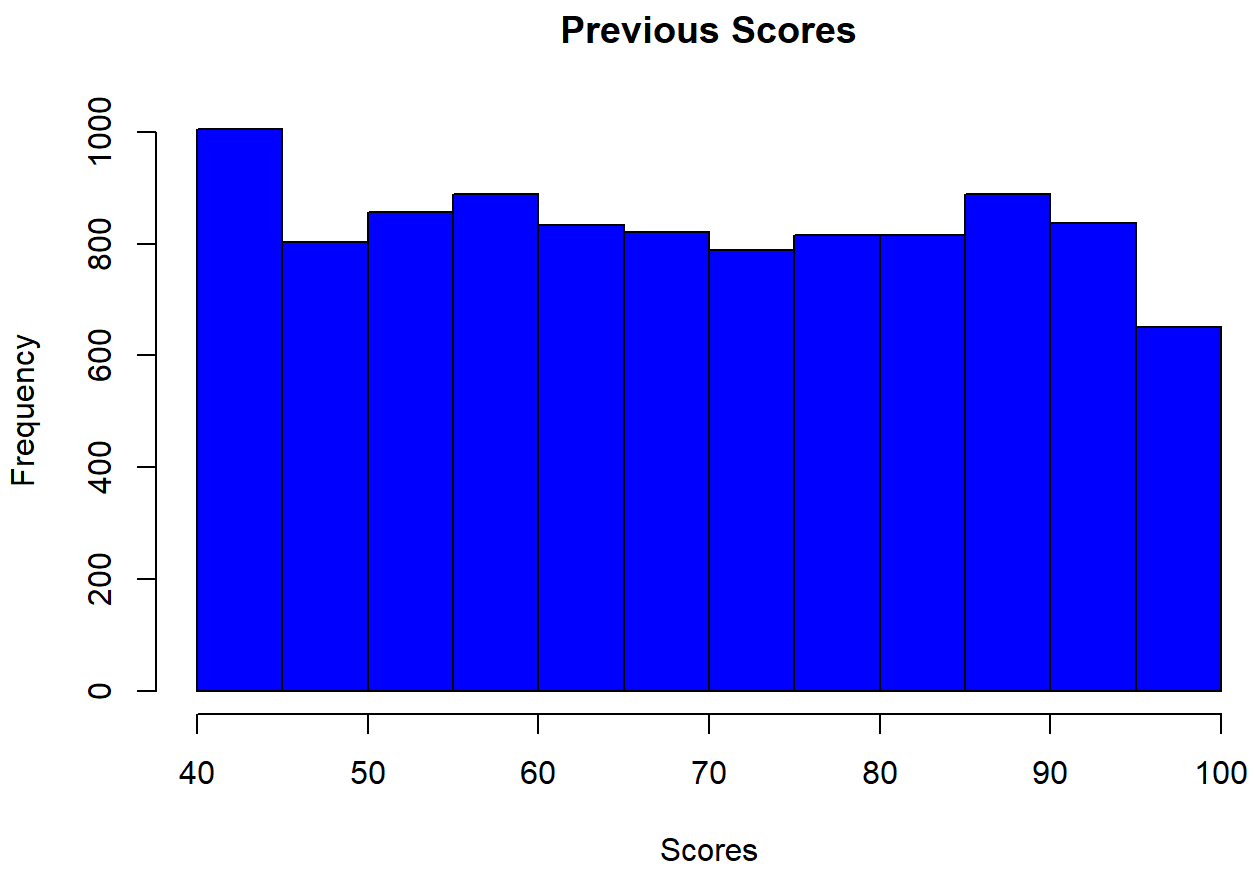
Exploratory Data Analysis

Histograms

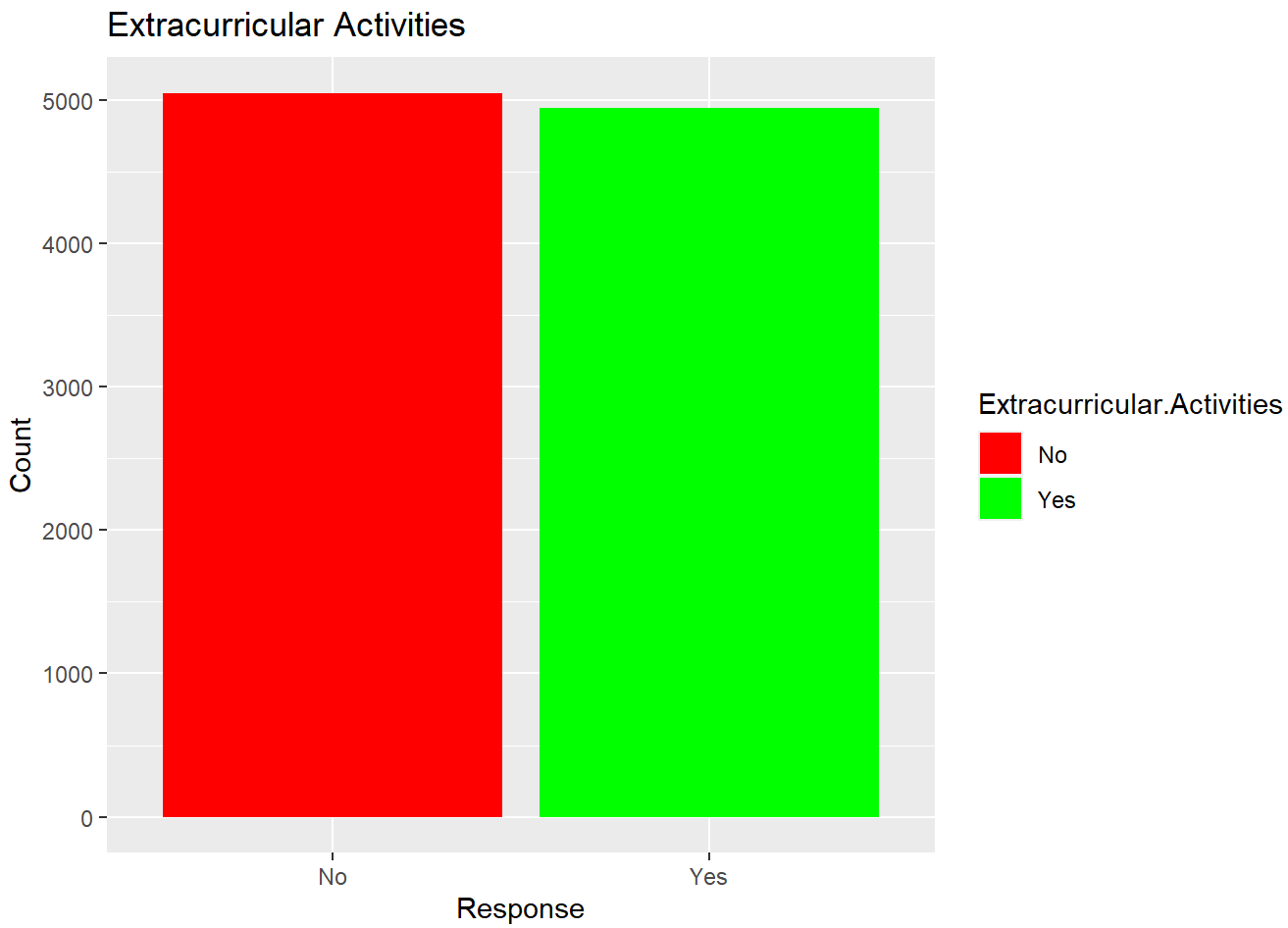
```
hist(df$Hours.Studied,main = "Hours Studied" , xlab = "Hours", ylab = "Frequency", col = "blue",
border = "black")
```



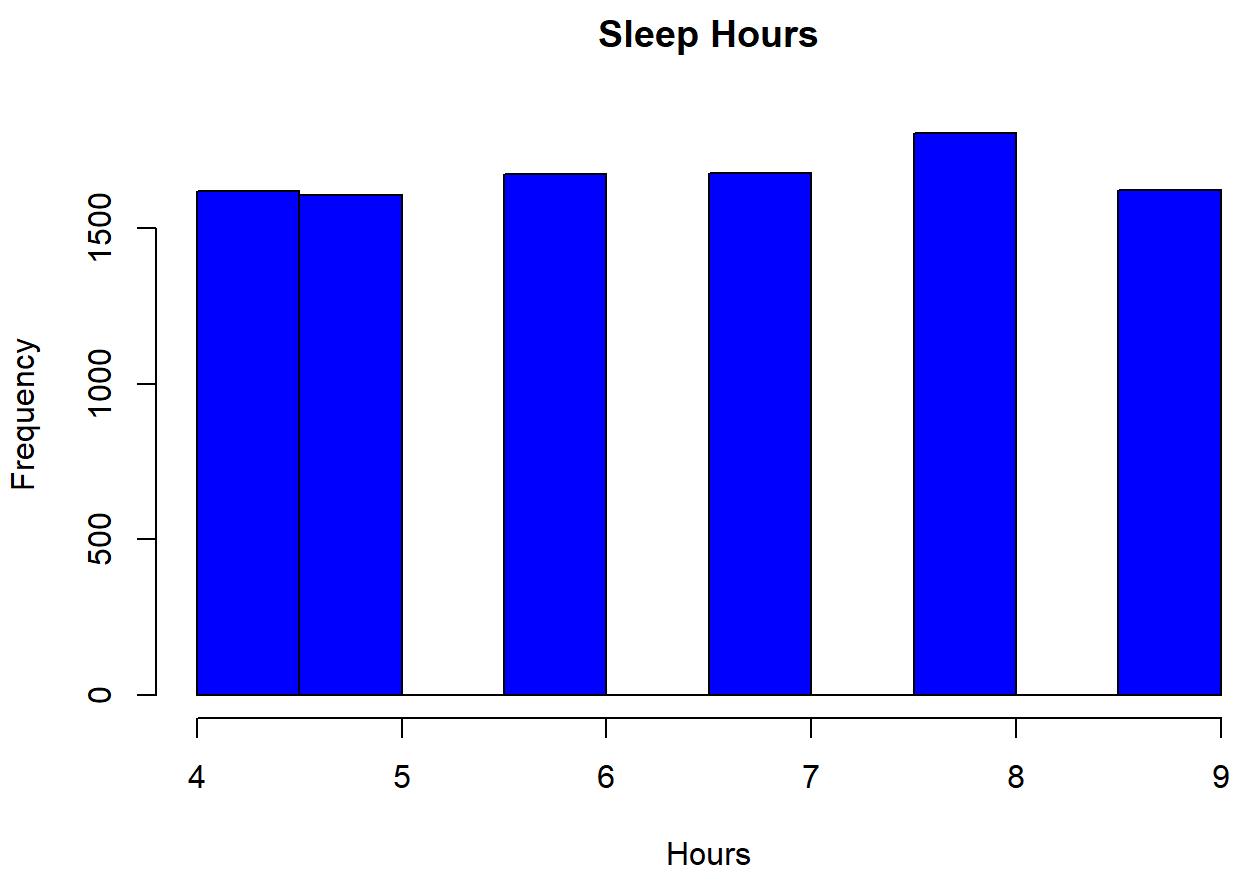

```
hist(df$Previous.Scores,main = "Previous Scores" , xlab = "Scores", ylab = "Frequency", col = "b
lue", border = "black")
```



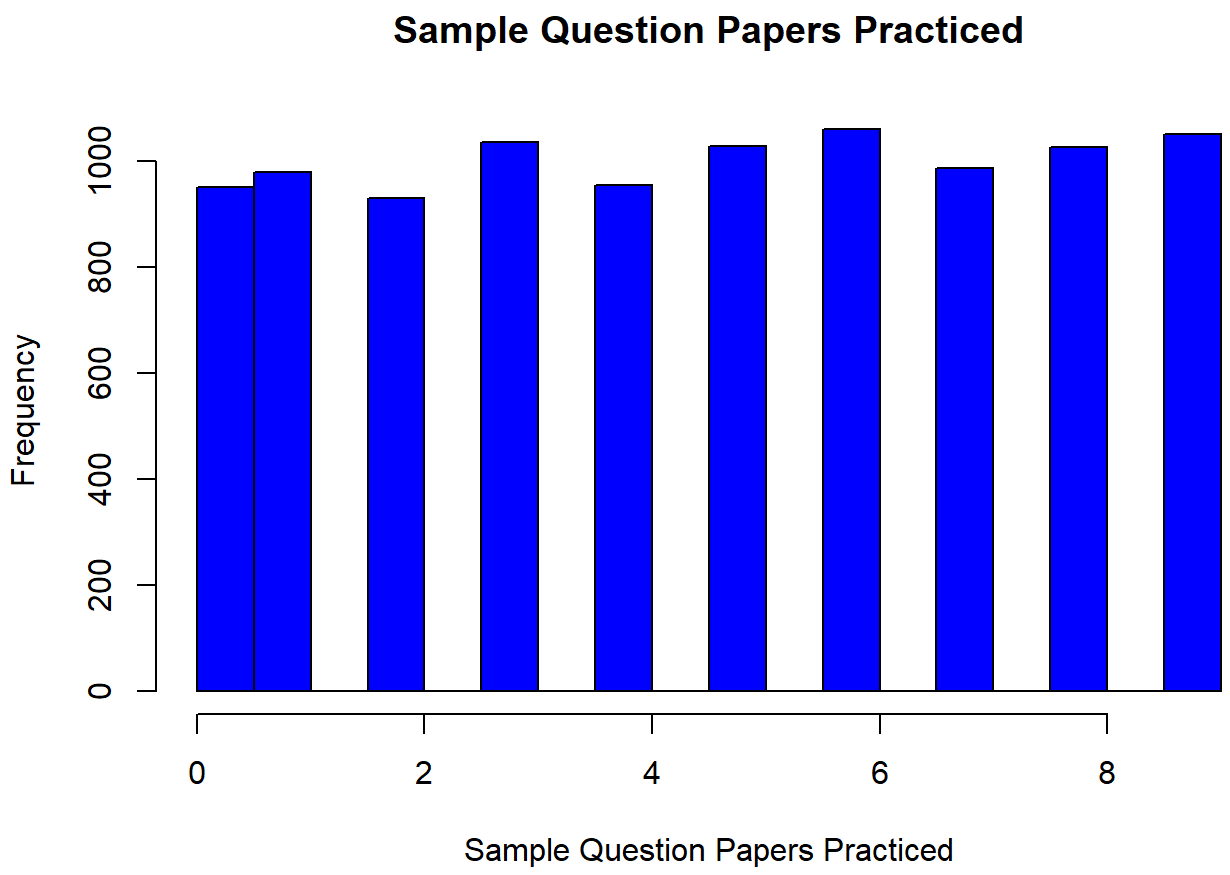

```
library(ggplot2)
ggplot(df, aes(x = Extracurricular.Activities, fill = Extracurricular.Activities)) +
  geom_bar() +
  labs(title = "Extracurricular Activities", x = "Response", y = "Count") +
  scale_fill_manual(values = c("No" = "red", "Yes" = "green"))
```



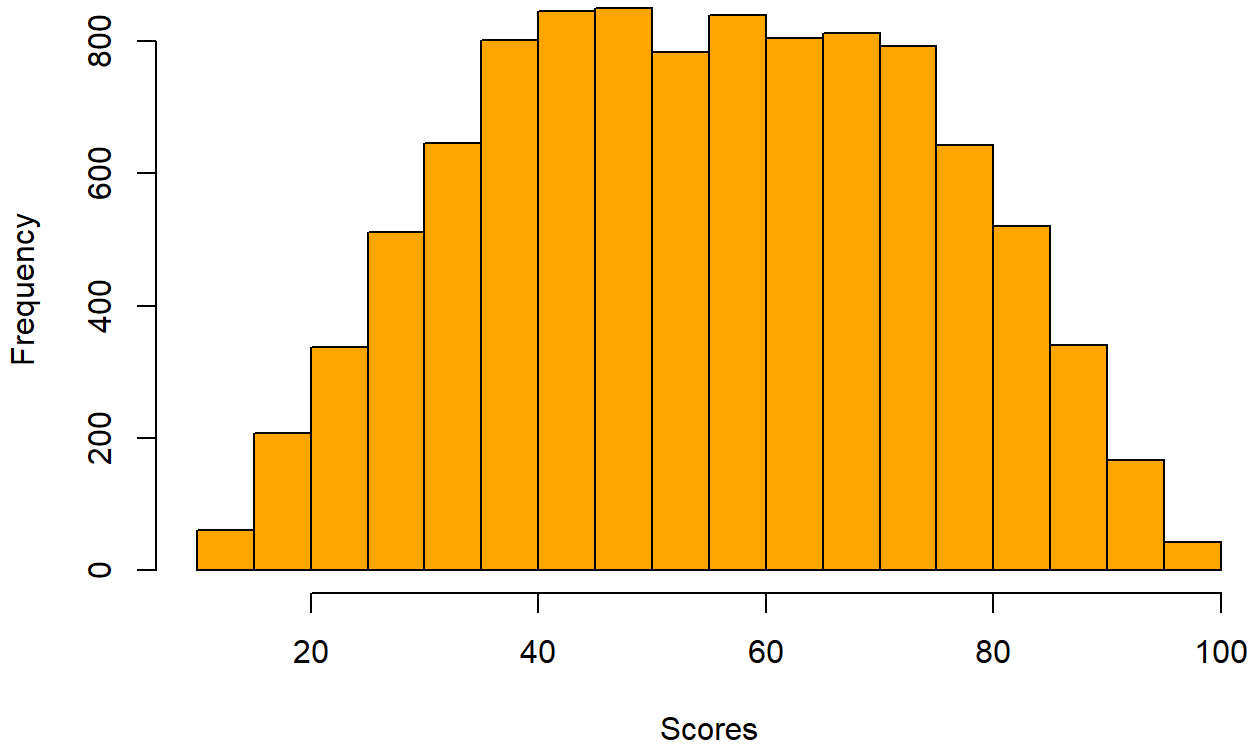

```
hist(df$Sleep.Hours,main = "Sleep Hours" , xlab = "Hours", ylab = "Frequency", col = "blue", bor
der = "black")
```




```
hist(df$Sample.Question.Papers.Practiced,main = "Sample Question Papers Practiced" , xlab = "Sam
ple Question Papers Practiced", ylab = "Frequency", col = "blue", border = "black")
```



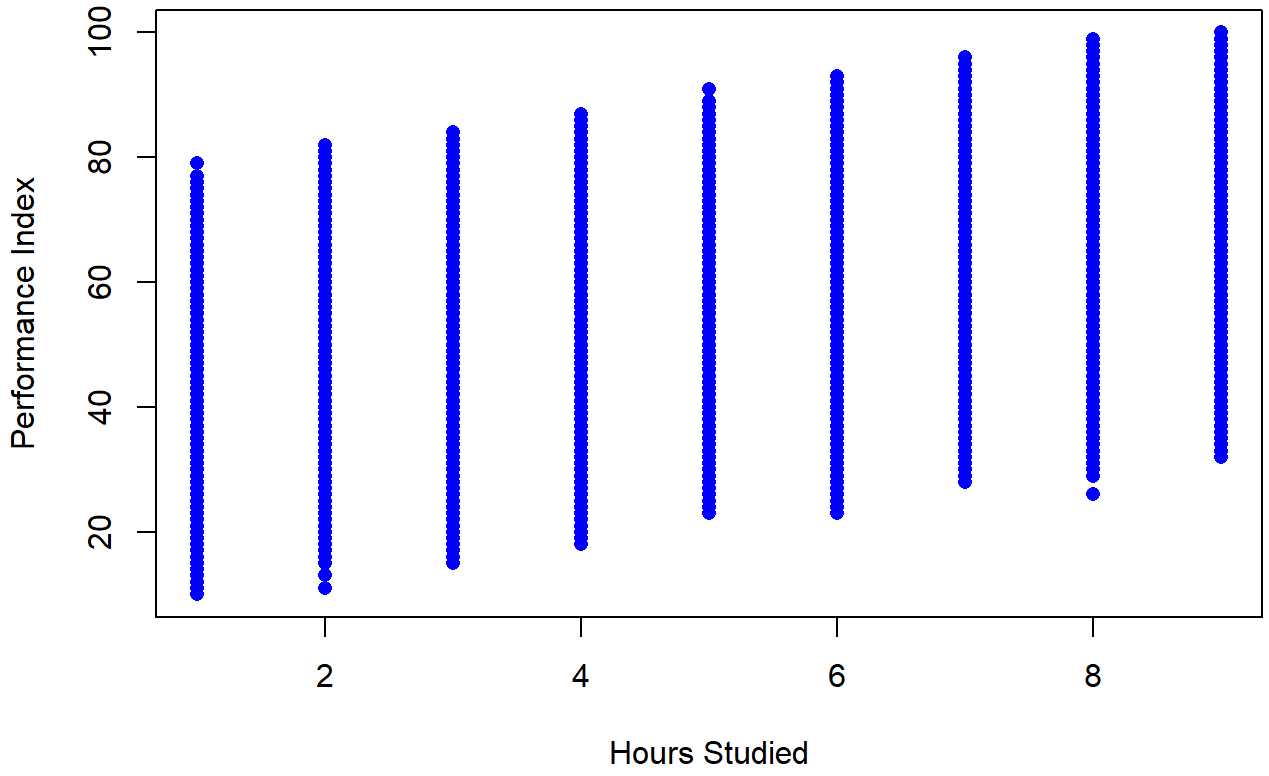

```
hist(df$Performance.Index,main = "Performance" , xlab = "Scores", ylab = "Frequency", col = "ora
nge", border = "black")
```



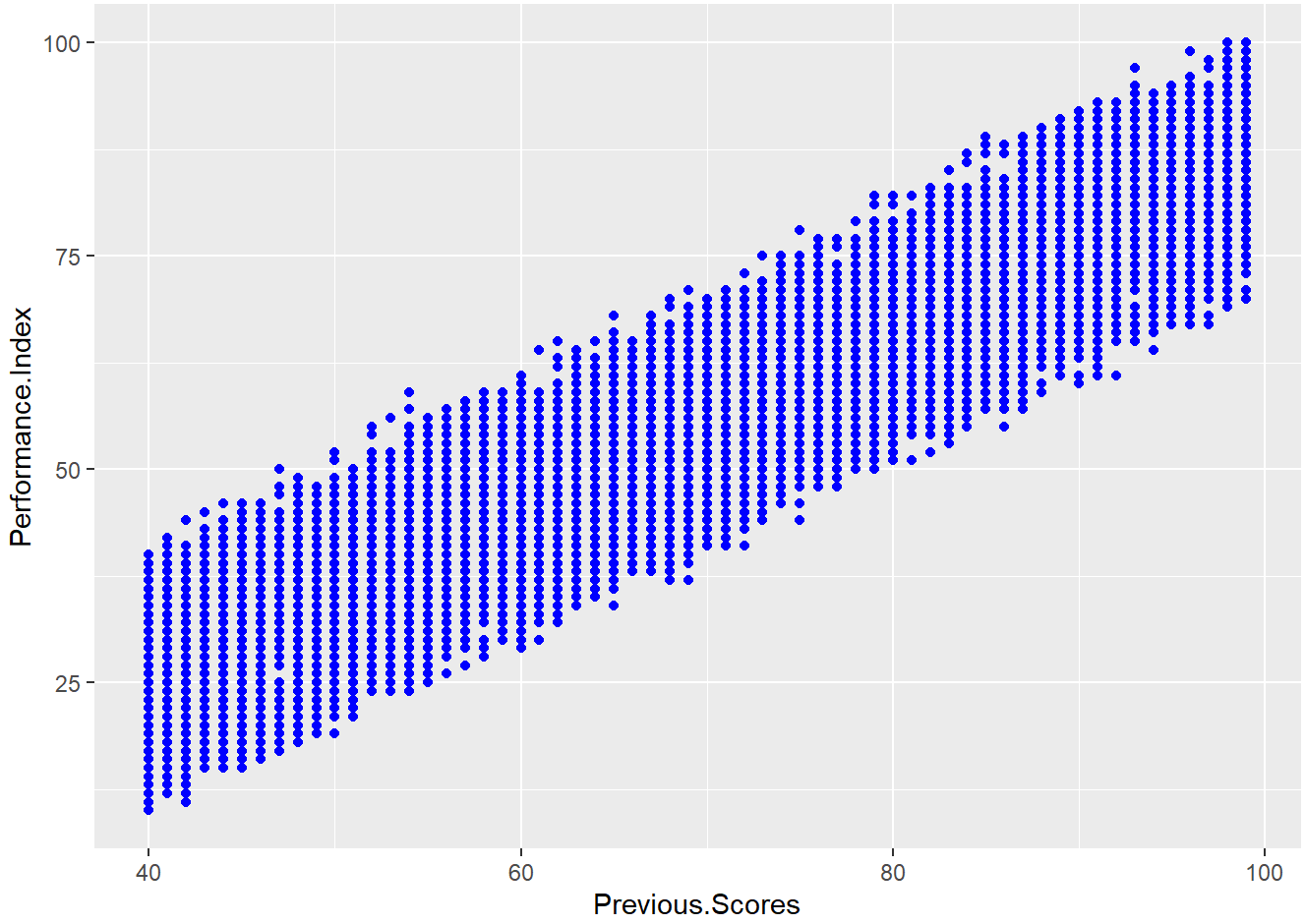
In the histograms, it is evident that Hours.Studied, Previous.Scores, Extracurricular.Activities, Sleep.Hours, and Sample.Question.Papers.Practiced showcase a uniform distribution, implying an even distribution of values.

In contrast, the target variable Performance.Index deviates from this pattern, displaying characteristics consistent with a normal distribution. This deviation suggests a central tendency in Performance.Index, potentially following a bell-shaped curve. These distinct distributional features among predictors and the target variable offer valuable insights, forming a crucial foundation for our subsequent modelling and analytical endeavours.

```
plot(df$Hours.Studied, df$Performance.Index,
     xlab = "Hours Studied",
     ylab = "Performance Index",
     col = "blue",
     pch = 16)
```



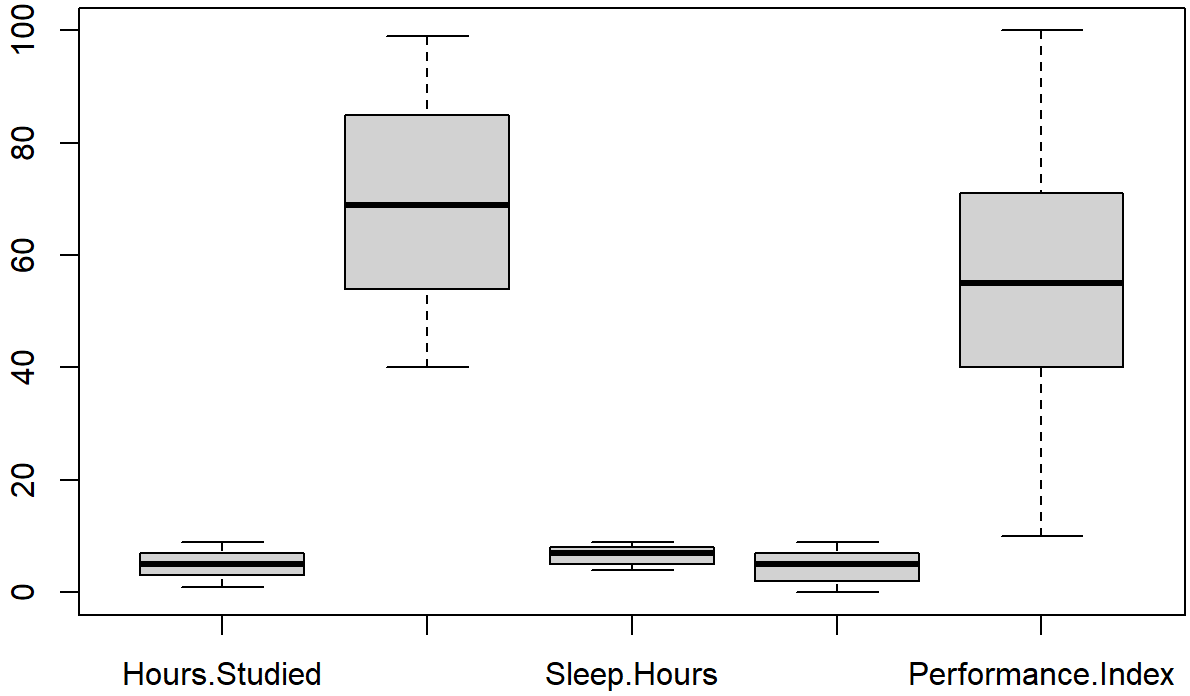

```
ggplot(df, aes(x = Previous.Scores, y = Performance.Index)) +
  geom_point(color = "blue")
```



It becomes apparent that the variables Previous.Scores and Performance.Index exhibit a discernible linear correlation. The plotted data points suggest a trend where an increase in Previous.Scores aligns with a corresponding rise in Performance.Index, indicative of a positive linear relationship between these two variables. This observation underscores the potential predictive power of Previous.Scores in influencing the overall performance index, providing a valuable insight for further analysis and model development.

Boxplots

```
boxplot(df[c(1,2,4,5,6)])
```



The boxplots for the features Hours.Studied, Previous.Scores, Sleep.Hours, Performance.Index, and Sample.Question.Papers.Practiced reveal a centralized distribution of data. This is evident as the medians are positioned near the center of each box, indicating a balanced distribution of values within these features. Notably, the absence of outliers in the boxplots suggests a consistent and concentrated spread of data points, reinforcing the stability of these variables.

Multiple Linear Regression

Dummy variable

As the the feature Extracurricular.Activities is a categorical variable with values of yes or no. It must be changed to a numerical value. That is why a dummy variable has been used. The value yes has been changed for 1 and no for 0.

```
library(fastDummies)
```

```
## Thank you for using fastDummies!
```

```
## To acknowledge our work, please cite the package:
```

```
## Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Ro
ws from Categorical Variables. Version 1.7.1. URL: https://github.com/jacobkap/fastDummies, http
s://jacobkap.github.io/fastDummies/.
```

```
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
```

```
df_dummy <- dummy_columns(df, select_columns = "Extracurricular.Activities")

df_total <- dplyr::select(df_dummy, -Extracurricular.Activities, -Extracurricular.Activities_No)
head(df_total)
```

```
##   Hours.Studied Previous.Scores Sleep.Hours Sample.Question.Papers.Practiced
## 1             7              99           9                                 1
## 2             4              82           4                                 2
## 3             8              51           7                                 2
## 4             5              52           5                                 2
## 5             7              75           8                                 5
## 6             3              78           9                                 6
##   Performance.Index Extracurricular.Activities_Yes
## 1                91                               1
## 2                65                               0
## 3                45                               1
## 4                36                               1
## 5                66                               0
## 6                61                               0
```

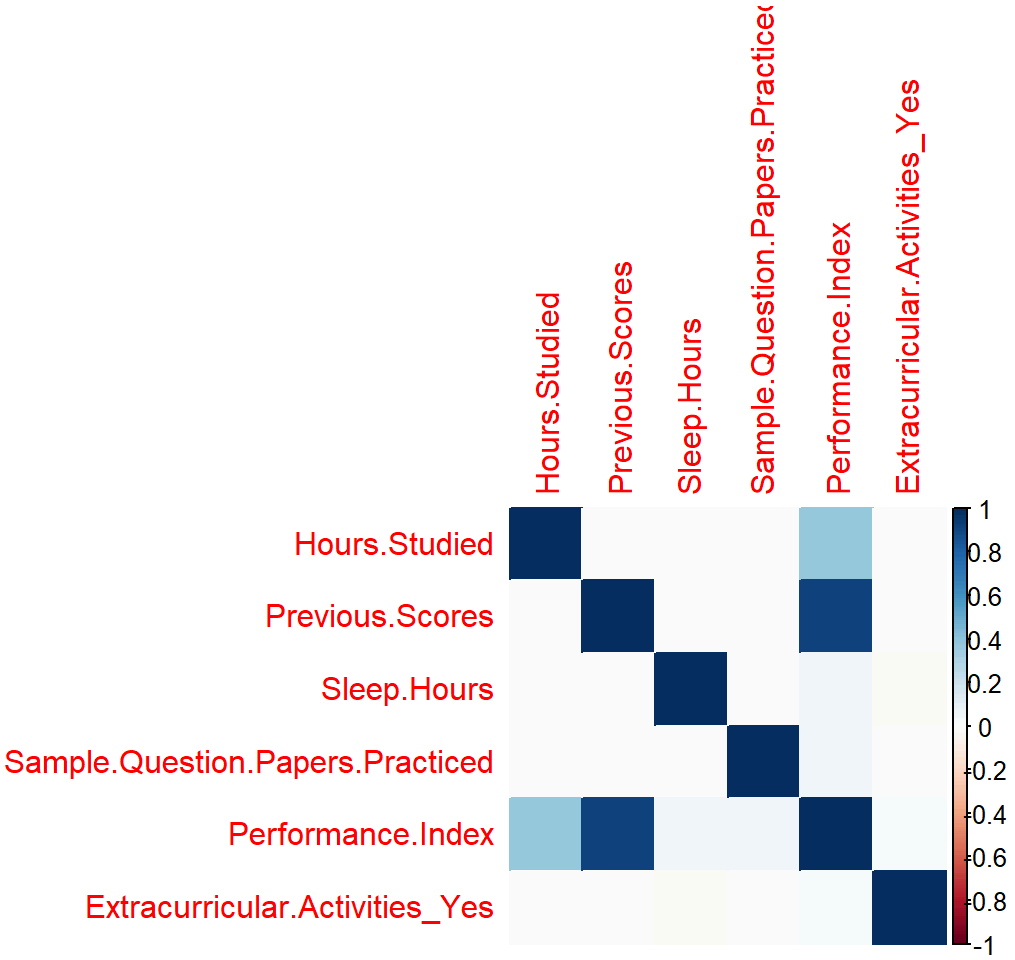
Correlation Map

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
cor_matrix <- cor(df_total)

corrplot(cor_matrix, method = "color")
```



The correlation map highlights a strong association between the target feature and the variables Previous.Scores and Hours.Studied. These predictors exhibit a noteworthy correlation, emphasizing their potential impact on the target variable. This information guides our focus on key contributors as we proceed with the analysis.

Fitting

```
model <- lm(Performance.Index ~ ., data = df_total)
summary(model)
```

```
##
## Call:
## lm(formula = Performance.Index ~ ., data = df_total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6333 -1.3684 -0.0311  1.3556  8.7932
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -34.075588   0.127143 -268.01   <2e-16 ***
## Hours.Studied                     2.852982   0.007873  362.35   <2e-16 ***
## Previous.Scores                   1.018434   0.001175  866.45   <2e-16 ***
## Sleep.Hours                       0.480560   0.012022   39.97   <2e-16 ***
## Sample.Question.Papers.Practiced  0.193802   0.007110   27.26   <2e-16 ***
## Extracurricular.Activities_Yes    0.612898   0.040781   15.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.038 on 9994 degrees of freedom
## Multiple R-squared:  0.9888, Adjusted R-squared:  0.9887
## F-statistic: 1.757e+05 on 5 and 9994 DF,  p-value: < 2.2e-16
```

The P-values for each parameter in our analysis are remarkably low, approaching almost zero. This signifies that we cannot assert the value of any parameter is equal to zero. The low P-values indicate a high level of statistical significance, suggesting that each parameter plays a significant role in influencing the observed outcomes.

The model achieved a remarkable R-squared of 0.9888, signifying that 98.88% of the dependent variable's variance is explained by the independent variables. The Adjusted R-squared, closely trailing at 0.9887, adjusts for predictor count, these high values suggest a strong fit.

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: Performance.Index
##                                   Df  Sum Sq Mean Sq   F value    Pr(>F)
## Hours.Studied                      1  515518  515518 124106.24 < 2.2e-16 ***
## Previous.Scores                    1 3123194 3123194 751880.62 < 2.2e-16 ***
## Sleep.Hours                        1    6560    6560   1579.24 < 2.2e-16 ***
## Sample.Question.Papers.Practiced   1    3131    3131    753.79 < 2.2e-16 ***
## Extracurricular.Activities_Yes     1     938     938    225.88 < 2.2e-16 ***
## Residuals                       9994   41514       4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sse <- sum(summary(model)$residuals^2)
ssr <- sum((predict(model) - mean(df_total$Performance.Index))^2)
sst <- sum((df_total$Performance.Index - mean(df_total$Performance.Index))^2)

# Display the results
print(paste("SSE:", sse))
```

```
## [1] "SSE: 41513.50633946"
```

```
print(paste("SSR:", ssr))
```

```
## [1] "SSR: 3649341.14326057"
```

```
print(paste("SST:", sst))
```

```
## [1] "SST: 3690854.6496"
```

A remarkably low Sum of Squares Error (SSE) suggests that the linear regression model effectively captures and explains the variability in the data, resulting in a highly precise fit.

```
confint(model, level = 0.95)
```

```
##                                       2.5 %      97.5 %
## (Intercept)                     -34.3248136 -33.8263626
## Hours.Studied                     2.8375484   2.8684157
## Previous.Scores                   1.0161302   1.0207382
## Sleep.Hours                       0.4569937   0.5041258
## Sample.Question.Papers.Practiced  0.1798645   0.2077397
## Extracurricular.Activities_Yes    0.5329596   0.6928356
```

These confidence intervals, calculated with a 95% confidence level, delineate the range of plausible values for each parameter in our model. They offer insight into the precision and reliability of our parameter estimates.

Stepwise process

```
step(model, direction = "both")
```

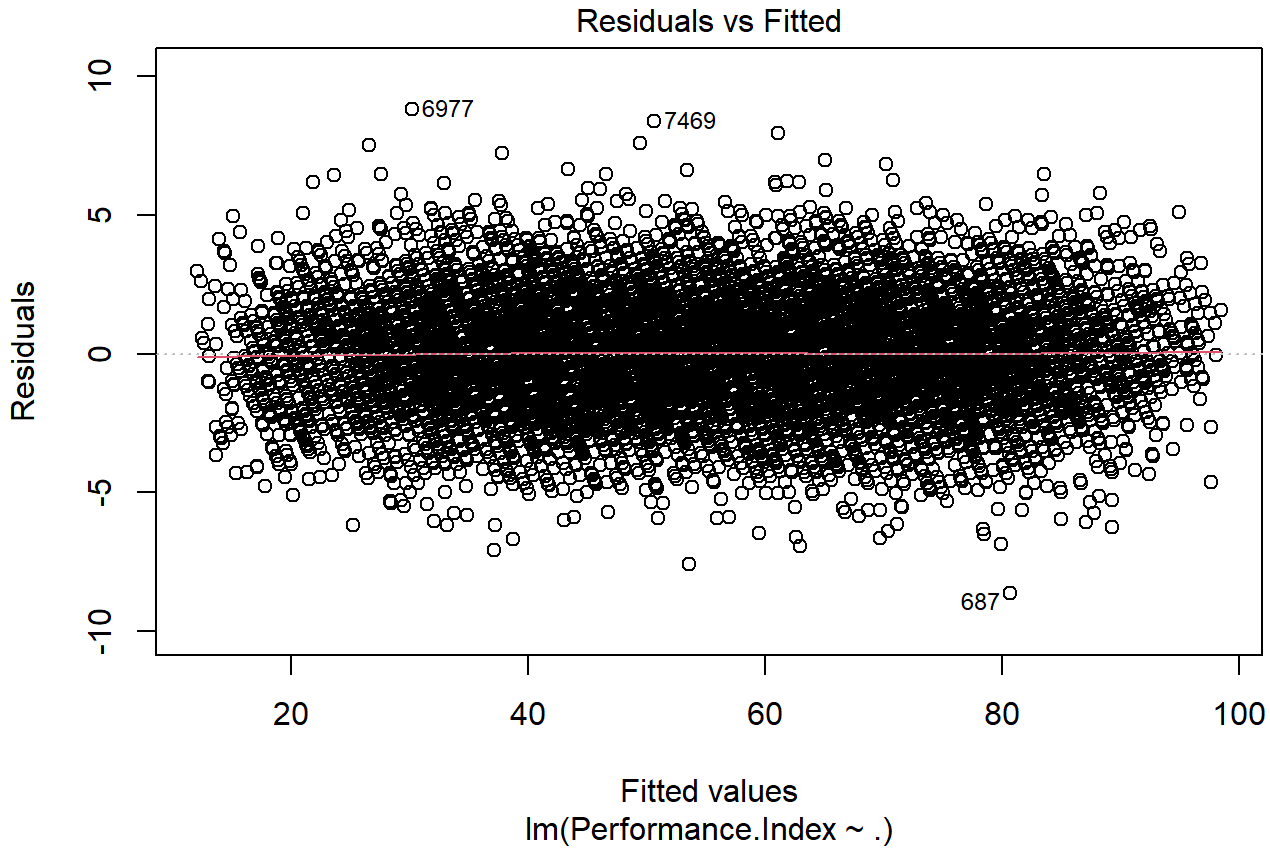
```
## Start:  AIC=14246.34
## Performance.Index ~ Hours.Studied + Previous.Scores + Sleep.Hours +
##     Sample.Question.Papers.Practiced + Extracurricular.Activities_Yes
##
##                                   Df Sum of Sq     RSS   AIC
## <none>                                          41514 14246
## - Extracurricular.Activities_Yes    1       938   42452 14468
## - Sample.Question.Papers.Practiced  1      3086   44599 14961
## - Sleep.Hours                       1      6637   48151 15727
## - Hours.Studied                     1    545400  586913 40733
## - Previous.Scores                   1   3118438 3159951 57567
```

```
##
## Call:
## lm(formula = Performance.Index ~ Hours.Studied + Previous.Scores +
##     Sleep.Hours + Sample.Question.Papers.Practiced + Extracurricular.Activities_Yes,
##     data = df_total)
##
## Coefficients:
##                       (Intercept)                     Hours.Studied
##                          -34.0756                            2.8530
##                  Previous.Scores                       Sleep.Hours
##                            1.0184                            0.4806
## Sample.Question.Papers.Practiced    Extracurricular.Activities_Yes
##                            0.1938                            0.6129
```

The stepwise process, guided by the Akaike Information Criterion (AIC), sequentially optimized the model by adding or removing variables. The decreasing AIC values (14246, 14468, 14961, 15727, 40733, 57567) indicate the iterative refinement, with the final model having the lowest AIC (14246), suggesting a well-balanced when all the predictors are used.

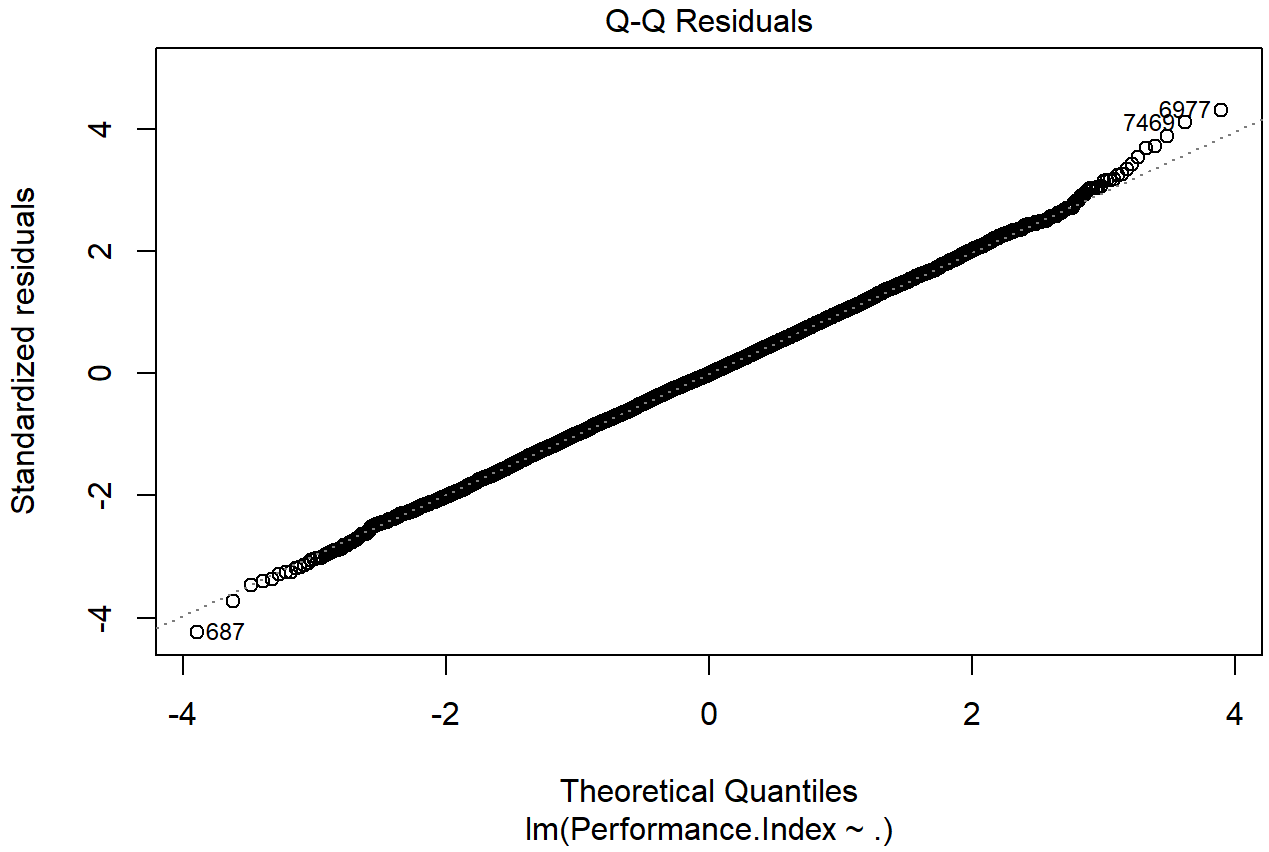
Residuals

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



The residuals vs. fitted values plot shows a random and patternless distribution of dots with a uniform spread, indicating the model captures the relationship well and meets assumptions of linearity and homoscedasticity. This increases confidence in the model's reliability and predictive performance.

```
plot(model, which=2)
```



The points in the QQ plot of residuals are close to the line, which indicates that the residuals follow a normal distribution. Having residuals that closely follow the line in a QQ plot is a good sign for the validity of the linear regression model. It implies that the model assumptions related to the distribution of errors are met, which enhances the reliability of statistical inferences and predictions made by the model.

Conclusion

In this project, we embarked on a comprehensive exploration of a dataset on student performance, applying statistical methods learned in the course, including multiple linear regression. The dataset, comprising 10,000 observations with six features, was a rich source for uncovering patterns influencing academic outcomes. We observed distributions and relationships through histograms, correlation maps, and boxplots, laying the foundation for our regression modelling.

The regression model, with an impressive R-squared of 0.9888 and an adjusted R-squared of 0.9887, showcased its effectiveness in explaining student performance. The low p-values of all parameters underscored their significance, and the confidence intervals at a 95% confidence level provided insights into parameter precision.

Further analysis involved examining residuals through QQ plots and residuals vs. fitted values plots. The QQ plot displayed adherence to normality assumptions, while the residuals vs. fitted values plot revealed a random, patternless scatter with a uniform spread, validating the model's linearity and homoscedasticity assumptions.

In conclusion, the project successfully applied regression techniques to understand and predict student performance with a robust model supported by thorough statistical diagnostics. The adherence to assumptions and the model's reliability encourage confidence in leveraging these insights for informed decision-making in educational contexts.