Student Performance Analysis

Farhana Taiyebah

2025-03-26

Introduction

Why do some students perform better than others? Is study time really the key to success, or do factors like family support and failures in past subjects play a bigger role? This project analyzes student performance data to uncover the strongest predictors of academic success.

Dataset Information

This dataset comes from the UCI Machine Learning Repository and was donated by Paulo Cortez and A. M. G. Silva in 2014. It consists of student achievement data from two Portuguese secondary schools and includes demographic, social, and academic factors. The dataset has been used in multiple studies to predict student performance using data mining techniques.

- **Number of Instances**: 649 students
- **Number of Features**: 30
- **Subject Areas**: Mathematics and Portuguese Language
- **Main Tasks**: Classification, Regression
- **Data Source**: School reports and questionnaires
- **Original Study**: *Using data mining to predict secondary school student performance* by P. Cortez & A. M. G. Silva (2008)
- **License**: Creative Commons Attribution 4.0 International (CC BY 4.0)

The dataset contains attributes such as student grades (`G1`, `G2`, `G3`), demographic details (`age`, `sex`, `famsize`), parental education and job information (`Medu`, `Fedu`, `Mjob`, `Fjob`), and behavioral factors (`studytime`, `failures`, `absences`).

For this analysis, I have chosen to work only with the Mathematics dataset (`student-mat.csv`). This decision was made to maintain a focused analysis and avoid potential inconsistencies when merging two datasets with different subject areas. Additionally, analyzing a single subject allows for deeper insights into factors affecting student performance in Mathematics without introducing subject-specific biases.

Setup

```r
# Install and load required packages
if (!requireNamespace("tidyverse", quietly = TRUE)) install.packages("tidyverse")
if (!requireNamespace("ggplot2", quietly = TRUE)) install.packages("ggplot2")
if (!requireNamespace("dplyr", quietly = TRUE)) install.packages("dplyr")
if (!requireNamespace("readr", quietly = TRUE)) install.packages("readr")
if (!requireNamespace("knitr", quietly = TRUE)) install.packages("knitr")
if (!requireNamespace("rmarkdown", quietly = TRUE)) install.packages("rmarkdown")

library(tidyverse)

# Define file path for cached dataset
cached_file <- "../output/student_mat_clean.rds"

if (file.exists(cached_file)) {
  # Load the preprocessed dataset if it exists
  data_mat_clean <- readRDS(cached_file)
  message("Loaded cached dataset.")
} else {
  # Load raw data
  data_mat <- read.csv("../data/student-mat.csv", sep=";")

  # Data cleaning: Convert categorical variables to factors
  data_mat_clean <- data_mat %>%
    mutate(across(where(is.character), as.factor))

  # Save cleaned dataset
  saveRDS(data_mat_clean, cached_file)
  message("Processed and cached dataset.")
}
```

```
## Loaded cached dataset.
```

1. Exploratory Data Analysis (EDA)

The goal of this section is to explore the dataset, understand its structure, and check for any missing or unusual data points.

```
# Check summary statistics and structure of the data
summary(data_mat_clean)
```

```
##  school   sex          age       address famsize   Pstatus      Medu
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000
##                   Median :17.0                             Median :3.000
##                   Mean   :16.7                             Mean   :2.749
##                   3rd Qu.:18.0                             3rd Qu.:4.000
##                   Max.   :22.0                             Max.   :4.000
##       Fedu             Mjob          Fjob            reason      guardian
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145   father: 90
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109   mother:273
##  Median :2.000   other   :141   other   :217   other     : 36   other : 32
##  Mean   :2.522   services:103   services:111   reputation:105
##  3rd Qu.:3.000   teacher : 58   teacher : 29
##  Max.   :4.000
##    traveltime      studytime        failures      schoolsup famsup     paid
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no :214
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242   yes:181
##  Median :1.000   Median :2.000   Median :0.0000
##  Mean   :1.448   Mean   :2.035   Mean   :0.3342
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000
##  activities nursery   higher    internet  romantic      famrel
##  no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000
##  yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000
##                                                     Median :4.000
##                                                     Mean   :3.944
##                                                     3rd Qu.:5.000
##                                                     Max.   :5.000
##     freetime         goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health        absences          G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

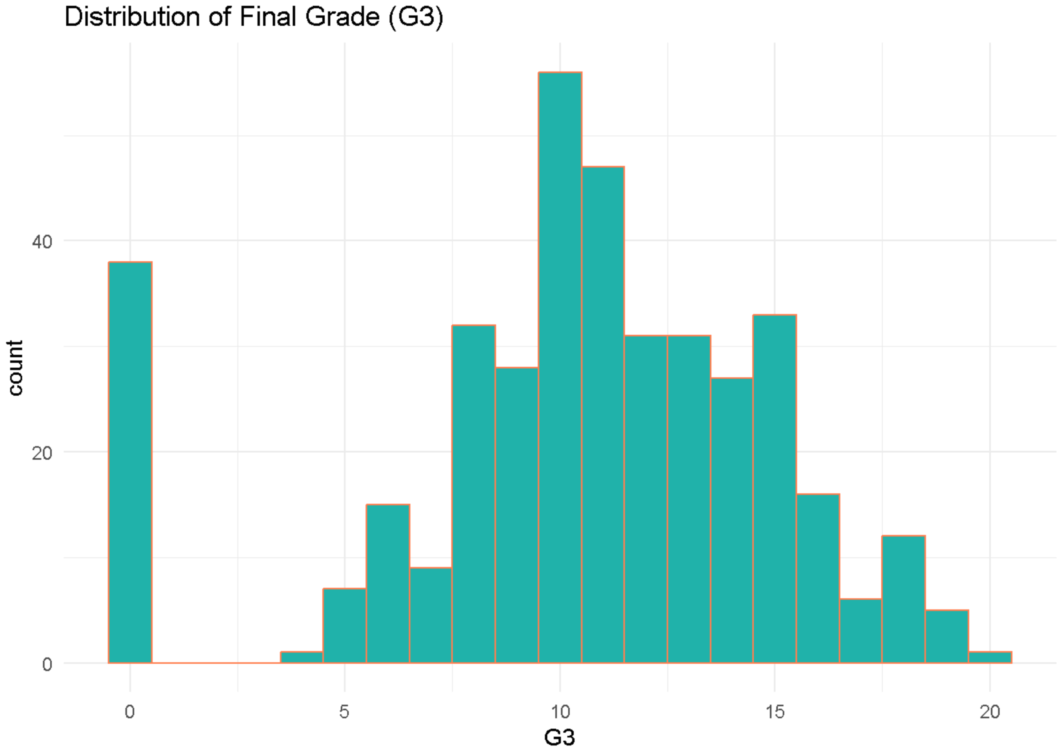
```
str(data_mat_clean)
```

```
## 'data.frame':    395 obs. of  33 variables:
##  $ school    : Factor w/ 2 levels "GP","MS": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ famsize   : Factor w/ 2 levels "GT3","LE3": 1 1 2 1 1 2 2 1 2 1 ...
##  $ Pstatus   : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : Factor w/ 5 levels "at_home","health",..: 1 1 1 2 3 4 3 3 4 3 ...
##  $ Fjob      : Factor w/ 5 levels "at_home","health",..: 5 3 3 4 3 3 3 5 3 3 ...
##  $ reason    : Factor w/ 4 levels "course","home",..: 1 1 3 2 2 4 2 2 2 2 ...
##  $ guardian  : Factor w/ 3 levels "father","mother",..: 2 1 2 2 1 2 2 2 2 2 ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : Factor w/ 2 levels "no","yes": 2 1 2 1 1 1 1 2 1 1 ...
##  $ famsup    : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
##  $ paid      : Factor w/ 2 levels "no","yes": 1 1 2 2 2 2 1 1 2 2 ...
##  $ activities: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
##  $ nursery   : Factor w/ 2 levels "no","yes": 2 1 2 2 2 2 2 2 2 2 ...
##  $ higher    : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ internet  : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
##  $ romantic  : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
# Check for missing values
colSums(is.na(data_mat_clean))
```

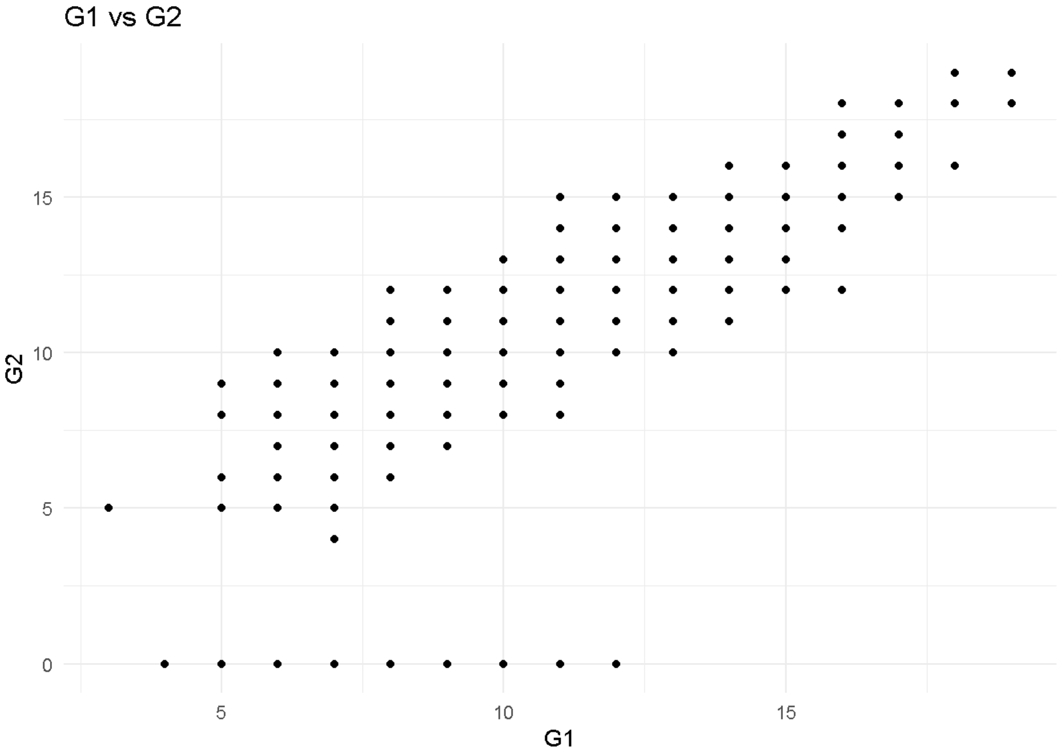
```
##     school        sex        age    address    famsize    Pstatus       Medu 
##          0          0          0          0          0          0          0 
##       Fedu       Mjob       Fjob     reason   guardian traveltime  studytime 
##          0          0          0          0          0          0          0 
##   failures  schoolsup     famsup       paid activities    nursery     higher 
##          0          0          0          0          0          0          0 
##   internet   romantic     famrel   freetime      goout       Dalc       Walc 
##          0          0          0          0          0          0          0 
##     health   absences         G1         G2         G3 
##          0          0          0          0          0
```

```
# Distribution of final grades
ggplot(data_mat_clean, aes(x = G3)) +
  geom_histogram(binwidth = 1, fill = "lightseagreen", color = "coral") +
  theme_minimal() +
  labs(title = "Distribution of Final Grade (G3)")
```



Most students score between 8 and 14, with some outliers on both ends.

```
# Correlation between first and second period grades
ggplot(data_mat_clean, aes(x = G1, y = G2)) +
  geom_point() +
  theme_minimal() +
  labs(title = "G1 vs G2")
```



Strong correlation between G1 and G2 suggests that past performance is a reliable indicator of future performance.

2. Data Cleaning

This section covers cleaning the data by handling missing values and converting categorical variables into factors.

```
# Data already cleaned and loaded from cache, but if not cleaned yet:
data_mat_clean <- data_mat %>%
  mutate(across(where(is.character), as.factor))

# Save cleaned dataset
saveRDS(data_mat_clean, cached_file)
message("Processed and cached dataset.")
```

```
## Processed and cached dataset.
```

3. Correlation Analysis

We analyze correlations between numerical variables to explore relationships that might help predict the final grade (G3).

```
# Calculate correlation matrix for numeric columns
numeric_data <- data_mat_clean %>% select(G1, G2, G3, absences)
cor_matrix <- cor(numeric_data)
print(cor_matrix)
```

```
##                  G1         G2         G3    absences
## G1        1.0000000  0.8521181 0.80146793 -0.03100290
## G2        0.8521181  1.0000000 0.90486799 -0.03177670
## G3        0.8014679  0.9048680 1.00000000  0.03424732
## absences -0.0310029 -0.0317767 0.03424732  1.00000000
```

G2 has the strongest correlation with G3, followed by G1. Absences have almost no correlation with final grades.

4. Statistical Model

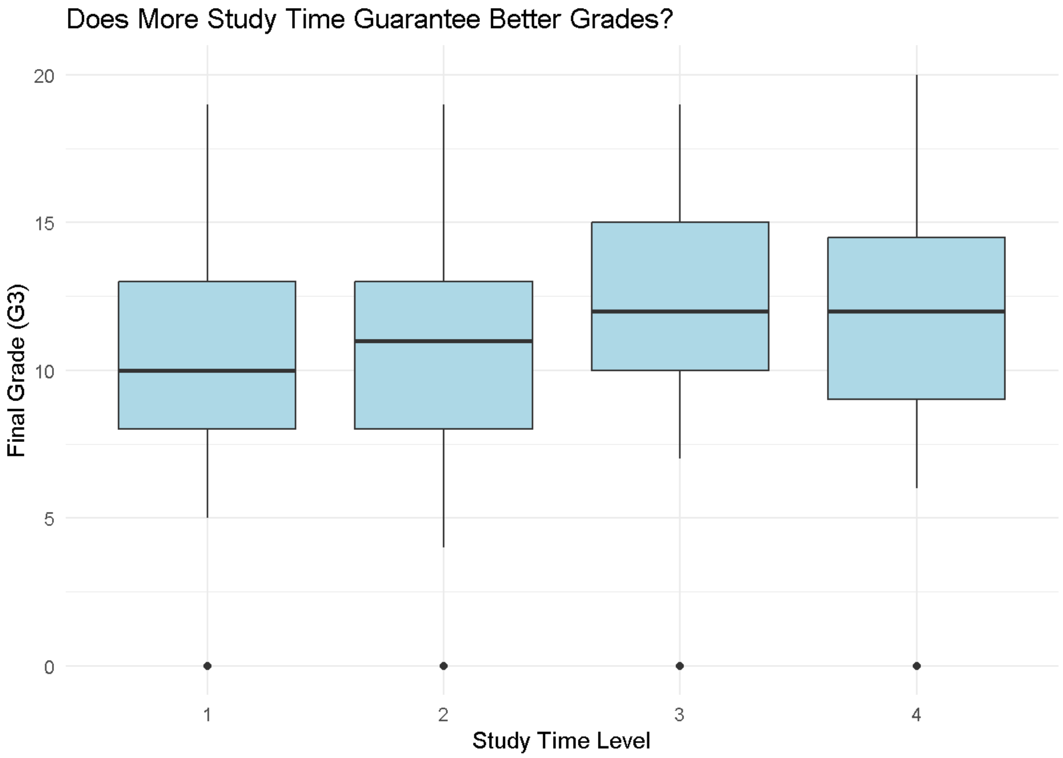
We build a linear regression model to predict the final grade (G3) based on other features like G1, G2, and absences.

```
# Build a linear regression model to predict G3
model <- lm(G3 ~ G1 + G2 + failures + studytime + absences, data = data_mat_clean)
summary(model)
```

```
## 
## Call:
## lm(formula = G3 ~ G1 + G2 + failures + studytime + absences, 
##     data = data_mat_clean)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.1894 -0.3662  0.2649  0.9706  3.6031 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -1.44276    0.43247  -3.336 0.000931 ***
## G1           0.14996    0.05580   2.687 0.007510 ** 
## G2           0.97726    0.04914  19.888  < 2e-16 ***
## failures    -0.28377    0.14041  -2.021 0.043968 *  
## studytime   -0.17817    0.11717  -1.521 0.129177    
## absences     0.03664    0.01205   3.040 0.002530 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.908 on 389 degrees of freedom
## Multiple R-squared:  0.8287, Adjusted R-squared:  0.8265 
## F-statistic: 376.4 on 5 and 389 DF,  p-value: < 2.2e-16
```

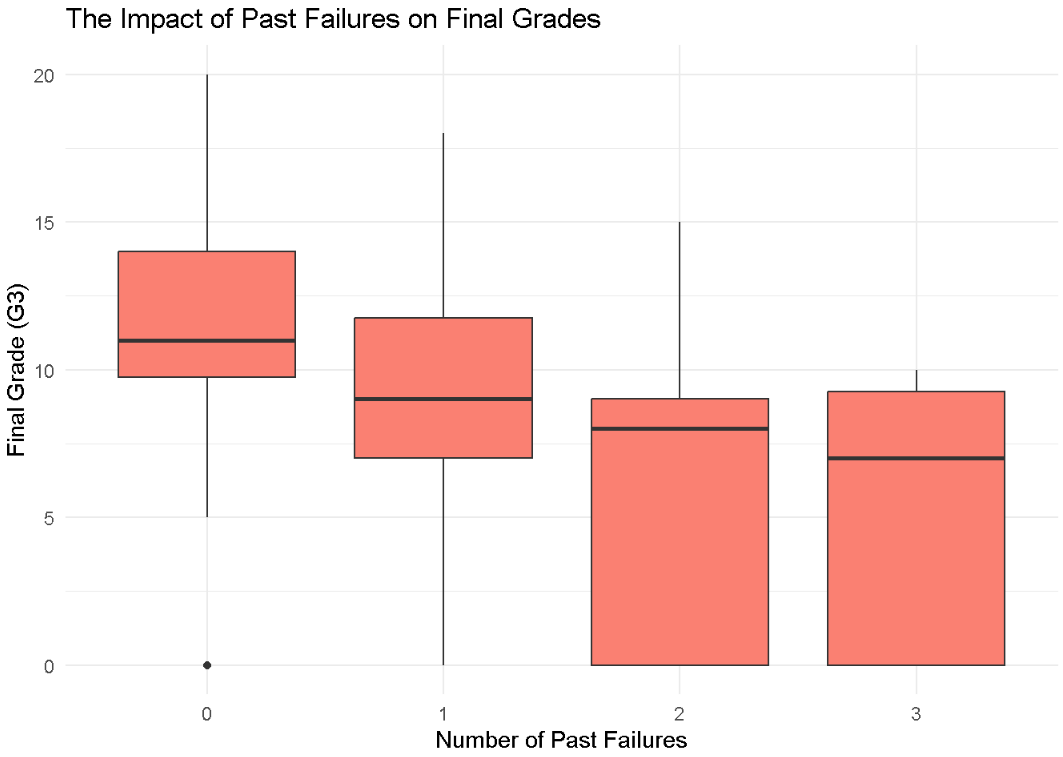
5. Data Visualization

```r
# Study Time vs Final Grade
ggplot(data_mat_clean, aes(x = factor(studytime), y = G3)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Does More Study Time Guarantee Better Grades?",
       x = "Study Time Level", y = "Final Grade (G3)") +
  theme_minimal()
```



Does More Study Time Guarantee Better Grades?

```r
# Failures vs Final Grade
ggplot(data_mat_clean, aes(x = factor(failures), y = G3)) +
  geom_boxplot(fill = "salmon") +
  labs(title = "The Impact of Past Failures on Final Grades",
       x = "Number of Past Failures", y = "Final Grade (G3)") +
  theme_minimal()
```



The Impact of Past Failures on Final Grades

Conclusion

This analysis shows that a student’s past grades are the strongest predictors of their final performance. Surprisingly, study time had little impact on final grades, suggesting that effective learning strategies may be more important than simply spending hours studying. Additionally, students with past failures tend to struggle significantly, highlighting the need for early interventions. Future research could explore whether external support systems (such as tutoring or parental involvement) play a role in improving student outcomes.

Rendering the R Markdown File

```r
# To render the .Rmd file into an HTML report, use the following command:
rmarkdown::render("R/student-performance-analysis.Rmd",
                  output_format = "html_document",
                  output_file = "../output/student-performance-analysis.html")
```