Impact of Test Preparation Course and Parental Education on Student Exam Scores

Andrea Aceves, Joshua Arias, Jonathan De La Torre, Andrew Mao

2025-05-10

Contents

```r
# Load necessary libraries
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.2.0 --

## v broom        1.0.6     v recipes      1.1.0
## v dials        1.3.0     v rsample      1.2.1
## v dplyr        1.1.4     v tibble       3.2.1
## v ggplot2      3.5.1     v tidyr        1.3.1
## v infer        1.0.7     v tune         1.2.1
## v modeldata    1.4.0     v workflows    1.1.4
## v parsnip      1.2.1     v workflowsets 1.1.0
## v purrr        1.0.2     v yardstick    1.3.1

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Use tidymodels_prefer() to resolve common conflicts.
```

```r
library(rio)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(ggplot2)
library(readr)
```

```
##
## Attaching package: 'readr'

## The following object is masked from 'package:yardstick':
##
##     spec
```

```
## The following object is masked from 'package:scales':
##
##     col_factor
```

```r
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

## The following objects are masked from 'package:yardstick':
##
##     precision, recall, sensitivity, specificity

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(rpart)
```

```
##
## Attaching package: 'rpart'

## The following object is masked from 'package:dials':
##
##     prune
```

```r
library(rpart.plot)
```

1. Introduction

1a. What is your research question?

We will perform multiple machine learning models we've learned in this course to help analyses and find answers to the following research questions:

- **Research Question 1**: How does completing a test preparation course influence students' exam scores in mathematics, reading, and writing?
- **Research Question 2**: What is the impact of parental education level on students' exam scores in mathematics, reading, and writing?

1b. What is your dependent variable which are candidates for independent variables?

We found that our Dependent Variables are Scores in Mathematics, Reading, and Writing. While our Independent Variables are Test Preparation Course in Research Question 1, and Parental Level of Education in Research Question 2.

1c. Why do you think the independent variables are correlated with the independent variables?

This report aims to investigate the impact of two key factors on students' exam scores in mathematics, reading, and writing:

- **Test Preparation Course**: Our group hypothesizes that students who completed a test preparation course will have higher scores in mathematics, reading, and writing due to better preparation.
- **Parental Level of Education**: Our group hypothesizes that students whose parents have a higher level of education will perform better in their exams, as they may receive more academic support and encouragement.

1d. In which direction (positive or negative do you expect the independent variables to influence the dependent variables?

Expected Direction of Influence:

- **Test Preparation Course**: Our group expects a positive correlation for students with completed test preparation courses to have higher exam scores.
- **Parental Level of Education**: Our group expects a positive correlation for students with higher parental education levels to have higher exam scores.

2. Data

Data Source: https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/data

The dataset includes:

-Gender- Male or Female

-Race/ethnicity- Groups A through E (the site gives no reason on the groupings)

-Parental level of education- associate's degree, bachelor's degree, high school, master's degree, some college, or some high school

-Lunch- Standard or free/reduced

-Test preparation course- completed or none

-Math score- A numerical value from 0 to 100

-Reading score- A numerical value from 0 to 100

-Writing score- A numerical value from 0 to 100

The data was sourced from a dataset containing students' scores along with demographic and academic information. The relevant columns were converted to factors where appropriate, and the structure of the data was checked to ensure readiness for analysis.

3. Data Preparation

The data was sourced from a dataset containing students' scores along with demographic and academic information. The relevant columns were converted to factors where appropriate, and the structure of the data was checked to ensure readiness for analysis.

```
# Load the dataset from your Desktop
StudentsPerformance <- read_csv("C:\\Users\\jnthn\\Desktop\\StudentsPerformance.csv")
```

```
## Rows: 1000 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): gender, race/ethnicity, parental level of education, lunch, test pr...
## dbl (3): math score, reading score, writing score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Clean the column names
StudentsPerformance <- clean_names(StudentsPerformance)

# Convert relevant columns to factors
StudentsPerformance <- StudentsPerformance %>%
  mutate(
    test_preparation_course = as.factor(test_preparation_course),
    parental_level_of_education = as.factor(parental_level_of_education)
  )

# Selecting the study variables
Data_PEMS <- StudentsPerformance %>%
  select(math_score, parental_level_of_education)
Data_PERS <- StudentsPerformance %>%
  select(reading_score, parental_level_of_education)
Data_PEWS<- StudentsPerformance %>%
  select(writing_score, parental_level_of_education)
Data_PCMS <- StudentsPerformance %>%
  select(math_score, test_preparation_course)
Data_PCRS <- StudentsPerformance %>%
  select(reading_score, test_preparation_course)
Data_PCWS<- StudentsPerformance %>%
  select(writing_score, test_preparation_course)
```

```r
# Set Seed for Reproducibility
set.seed(081524)

# Split Data into Training (80%) and Testing (20%) Sets
Data_PEMS_Split8020=initial_split(0.8,data=Data_PEMS)
Data_PEMS_DataTrain=training(Data_PEMS_Split8020)
Data_PEMS_DataTest=testing(Data_PEMS_Split8020)

Data_PERS_Split8020=initial_split(0.8,data=Data_PERS)
Data_PERS_DataTrain=training(Data_PERS_Split8020)
Data_PERS_DataTest=testing(Data_PERS_Split8020)

Data_PEWS_Split8020=initial_split(0.8,data=Data_PEWS)
Data_PEWS_DataTrain=training(Data_PEWS_Split8020)
Data_PEWS_DataTest=testing(Data_PEWS_Split8020)

Data_PCMS_Split8020=initial_split(0.8,data=Data_PCMS)
Data_PCMS_DataTrain=training(Data_PCMS_Split8020)
Data_PCMS_DataTest=testing(Data_PCMS_Split8020)

Data_PCRS_Split8020=initial_split(0.8,data=Data_PCRS)
Data_PCRS_DataTrain=training(Data_PCRS_Split8020)
Data_PCRS_DataTest=testing(Data_PCRS_Split8020)

Data_PCWS_Split8020=initial_split(0.8,data=Data_PCWS)
Data_PCWS_DataTrain=training(Data_PCWS_Split8020)
Data_PCWS_DataTest=testing(Data_PCWS_Split8020)
```

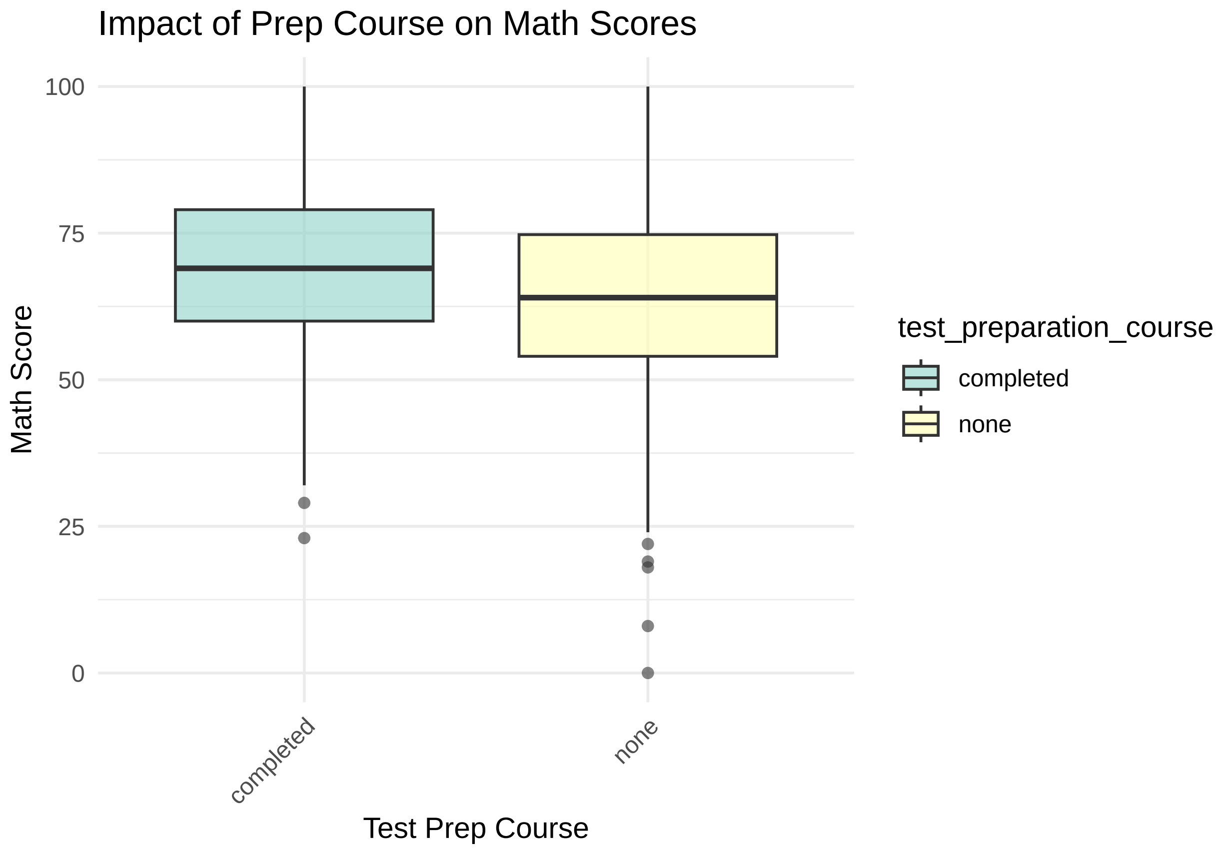
4. Run your first Machine Learning Model

4a. Explain the model Research Question 1

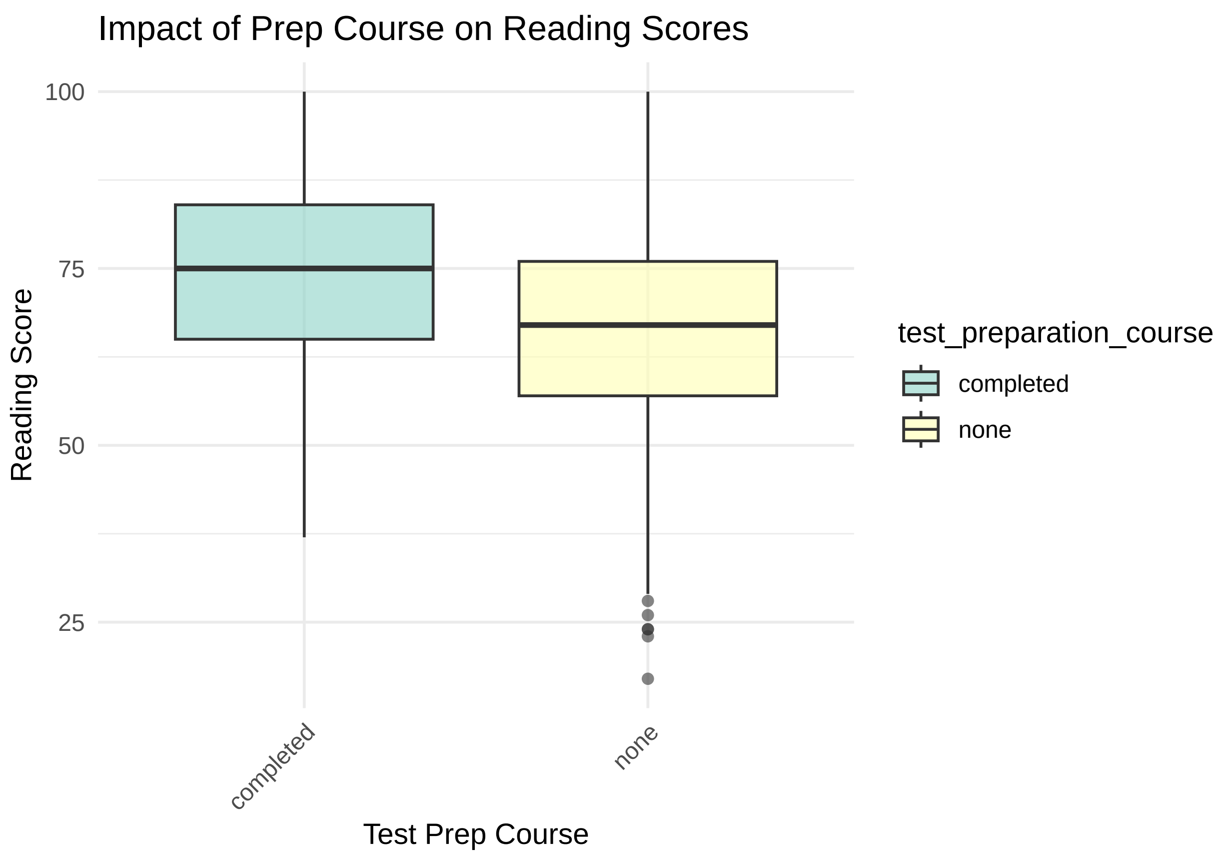
To assess the impact of completing a test preparation course on students' exam scores, we conducted separate linear regressions for each subject (math, reading, and writing). Below are the results and visualizations.

4b. Evaluate the testing data and report the metrics Research Question 1

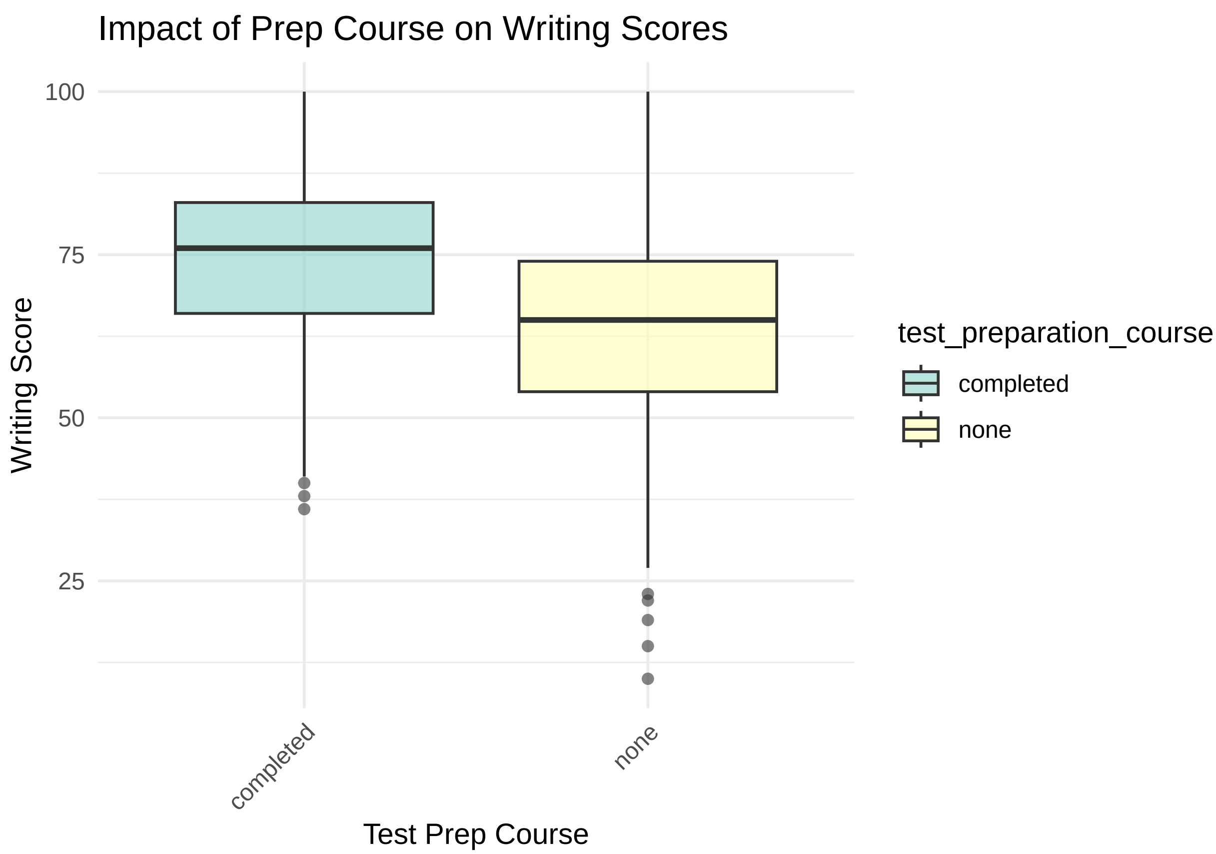
```r
# Train a Linear Regression Model on Training Data
Data_PCMS_model <- lm(math_score ~ test_preparation_course, data = Data_PCMS)
Data_PCRS_model <- lm(reading_score ~ test_preparation_course, data = Data_PCRS)
Data_PCWS_model <- lm(writing_score ~ test_preparation_course, data = Data_PCWS)

# # Visualization for Math Scores (Linear Regression)
ggplot(Data_PCMS_model, aes(x = test_preparation_course, y = math_score, fill = test_preparation_course)
  geom_boxplot(alpha = 0.6) +
  labs(title = "Impact of Prep Course on Math Scores",
       x = "Test Prep Course",
       y = "Math Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Visualization for Reading Scores (Linear Regression)
ggplot(Data_PCRS_model, aes(x = test_preparation_course, y = reading_score, fill = test_preparation_cou
  geom_boxplot(alpha = 0.6) +
  labs(title = "Impact of Prep Course on Reading Scores",
       x = "Test Prep Course",
       y = "Reading Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Visualization for Writing Scores (Linear Regression)
ggplot(Data_PCWS_model, aes(x = test_preparation_course, y = writing_score, fill = test_preparation_cou
  geom_boxplot(alpha = 0.6) +
  labs(title = "Impact of Prep Course on Writing Scores",
       x = "Test Prep Course",
       y = "Writing Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Summary of the Models
summary(Data_PCMS_model)
```

```
##
## Call:
## lm(formula = math_score ~ test_preparation_course, data = Data_PCMS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.078 -10.078  -0.078   9.922  35.922
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  69.6955     0.7890  88.330  < 2e-16 ***
## test_preparation_coursenone  -5.6176     0.9848  -5.705 1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.93 on 998 degrees of freedom
## Multiple R-squared:  0.03158,    Adjusted R-squared:  0.03061
## F-statistic: 32.54 on 1 and 998 DF,  p-value: 1.536e-08
```

```r
summary(Data_PCRS_model)
```

```
##
```

```
## Call:
## lm(formula = reading_score ~ test_preparation_course, data = Data_PCRS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.534  -9.054   0.466   9.466  33.466
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  73.8939     0.7491  98.640  < 2e-16 ***
## test_preparation_coursenone  -7.3596     0.9349  -7.872 9.08e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.17 on 998 degrees of freedom
## Multiple R-squared:  0.05846,    Adjusted R-squared:  0.05751
## F-statistic: 61.96 on 1 and 998 DF,  p-value: 9.082e-15
```

```r
summary(Data_PCWS_model)
```

```
##
## Call:
## lm(formula = writing_score ~ test_preparation_course, data = Data_PCWS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.505  -9.505   1.038   9.495  35.495
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  74.4190     0.7632   97.52   <2e-16 ***
## test_preparation_coursenone  -9.9143     0.9525  -10.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.44 on 998 degrees of freedom
## Multiple R-squared:  0.09794,    Adjusted R-squared:  0.09703
## F-statistic: 108.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

```r
# Predict on the Test Data
PCMS_predictions <- predict(Data_PCMS_model, Data_PCMS_DataTest)
PCRS_predictions <- predict(Data_PCRS_model, Data_PCRS_DataTest)
PCWS_predictions <- predict(Data_PCWS_model, Data_PCWS_DataTest)

# Compare Predictions to Actual Values
PCMS_results <- data.frame(
  Actual = Data_PEMS_DataTest$math_score,
  Predicted = PCMS_predictions)
PCRS_results <- data.frame(
  Actual = Data_PCRS_DataTest$reading_score,
  Predicted = PCRS_predictions)
PCWS_results <- data.frame(
  Actual = Data_PCWS_DataTest$writing_score,
```

```r
  Predicted = PCWS_predictions)

# View Results
head(PCMS_results)
```

```
##   Actual Predicted
## 1     72  64.07788
## 2     90  69.69553
## 3     88  64.07788
## 4     78  64.07788
## 5     69  69.69553
## 6     74  69.69553
```

```r
head(PCRS_results)
```

```
##   Actual Predicted
## 1     95  66.53427
## 2     83  66.53427
## 3     60  66.53427
## 4     81  66.53427
## 5     71  73.89385
## 6     54  66.53427
```

```r
head(PCWS_results)
```

```
##   Actual Predicted
## 1     74  64.50467
## 2     78  64.50467
## 3     92  74.41899
## 4     53  64.50467
## 5     72  64.50467
## 6     65  64.50467
```

4c. Interpret the quality of your results Research Question 1

Mathematics Scores: Coefficient for Test Preparation Course (None): -5.6176

This coefficient indicates that students who did not complete the test preparation course scored, on average, 5.6176 points lower in mathematics compared to those who did complete the course.

P-value: 1.54e-08

The extremely low p-value indicates that this difference is highly statistically significant, meaning that it is very unlikely to have occurred by chance.

R-squared: 0.03158

This R-squared value suggests that the test preparation course accounts for about 3.16% of the variability in mathematics scores. While this is a small percentage, it is significant given the large p-value.

Reading Scores: Coefficient for Test Preparation Course (None): -7.3596

Students who did not complete the test preparation course scored 7.3596 points lower on average in reading compared to those who did complete the course.

P-value: 9.08e-15

The p-value is very low, indicating that the difference in reading scores is highly statistically significant.

R-squared: 0.05846

The R-squared value indicates that about 5.85% of the variability in reading scores is explained by whether or not a student completed the test preparation course.

Writing Scores: Coefficient for Test Preparation Course (None): -9.9143

Students who did not complete the test preparation course scored, on average, 9.9143 points lower in writing compared to those who completed the course.

P-value: < 2e-16

This extremely low p-value shows a very strong statistical significance, indicating that the difference in writing scores is not due to random chance.

R-squared: 0.09794

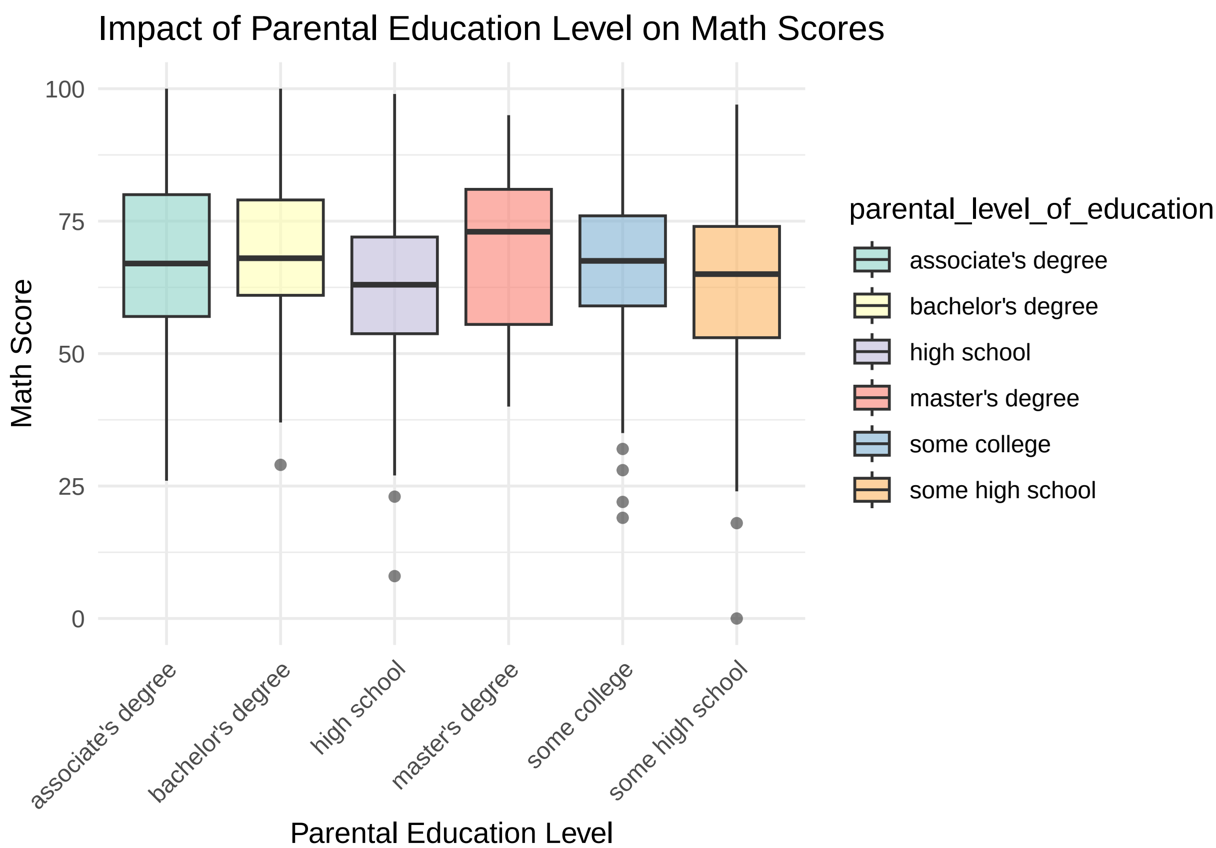
The R-squared value suggests that 9.79% of the variability in writing scores is explained by the completion of the test preparation course, which is a relatively more substantial effect compared to math and reading.

4d. Explain the model Research Question 2

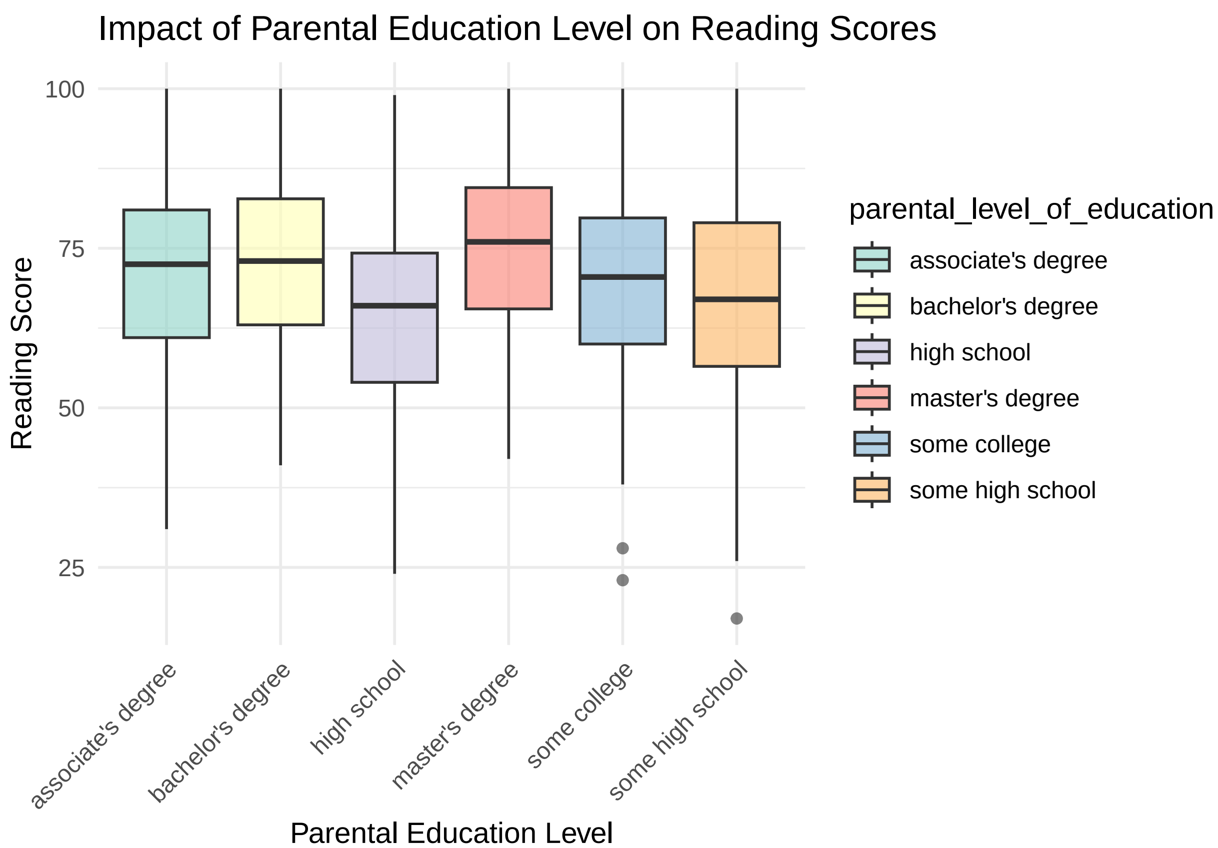
Next, we examined how parental education level influences students' exam scores in mathematics, reading, and writing. Separate linear regressions were conducted for each subject, and the results are presented below.

4e. Evaluate the testing data and report the metrics Research Question 2

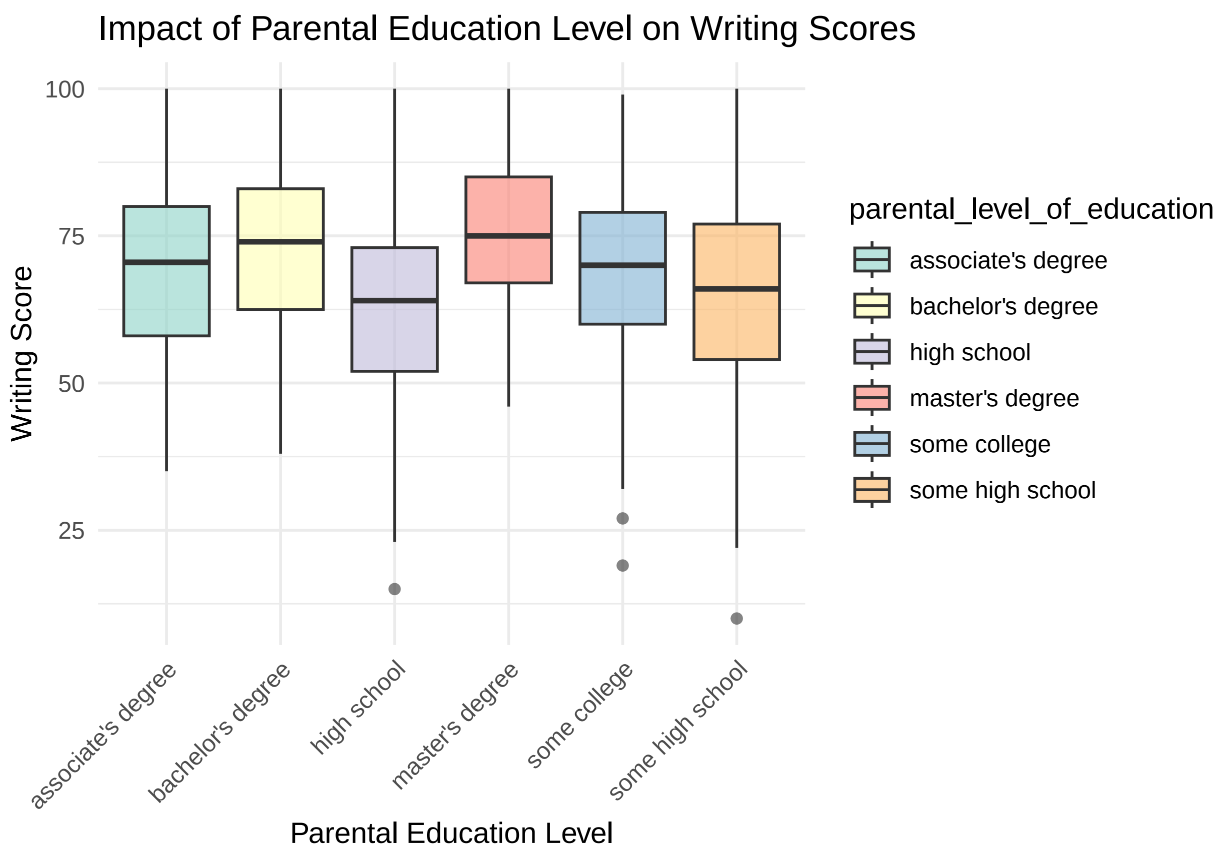
```
# Train a Linear Regression Model on Training Data
Data_PEMS_model <- lm(math_score ~ parental_level_of_education, data = Data_PEMS)
Data_PERS_model <- lm(reading_score ~ parental_level_of_education, data = Data_PERS)
Data_PEWS_model <- lm(writing_score ~ parental_level_of_education, data = Data_PEWS)

# # Visualization for Math Scores (Linear Regression)
ggplot(Data_PEMS_model, aes(x = parental_level_of_education, y = math_score, fill = parental_level_of_e
  geom_boxplot(alpha = 0.6) +
  labs(title = "Impact of Parental Education Level on Math Scores",
       x = "Parental Education Level",
       y = "Math Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Visualization for Reading Scores (Linear Regression)
ggplot(Data_PERS_model, aes(x = parental_level_of_education, y = reading_score, fill = parental_level_of
  geom_boxplot(alpha = 0.6) +
  labs(title = "Impact of Parental Education Level on Reading Scores",
       x = "Parental Education Level",
       y = "Reading Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Visualization for Writing Scores (Linear Regression)
ggplot(Data_PEWS_model, aes(x = parental_level_of_education, y = writing_score, fill = parental_level_o
  geom_boxplot(alpha = 0.6) +
  labs(title = "Impact of Parental Education Level on Writing Scores",
       x = "Parental Education Level",
       y = "Writing Score") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Summary of the Models
summary(Data_PEMS_model)
```

```
##
## Call:
## lm(formula = math_score ~ parental_level_of_education, data = Data_PEMS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.497  -9.138   0.186  10.503  36.862
##
## Coefficients:
##                                             Estimate Std. Error t value
## (Intercept)                                  67.8829     1.0039  67.619
## parental_level_of_educationbachelor's degree  1.5069     1.7041   0.884
## parental_level_of_educationhigh school       -5.7451     1.4661  -3.919
## parental_level_of_educationmaster's degree    1.8629     2.1909   0.850
## parental_level_of_educationsome college      -0.7546     1.4134  -0.534
## parental_level_of_educationsome high school  -4.3857     1.5026  -2.919
##                                             Pr(>|t|)
## (Intercept)                                  < 2e-16 ***
## parental_level_of_educationbachelor's degree 0.37674
## parental_level_of_educationhigh school      9.51e-05 ***
## parental_level_of_educationmaster's degree   0.39537
## parental_level_of_educationsome college      0.59356
## parental_level_of_educationsome high school  0.00359 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.96 on 994 degrees of freedom
## Multiple R-squared:  0.03176,    Adjusted R-squared:  0.02689
## F-statistic: 6.522 on 5 and 994 DF,  p-value: 5.592e-06
```

```r
summary(Data_PERS_model)
```

```
##
## Call:
## lm(formula = reading_score ~ parental_level_of_education, data = Data_PERS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.939  -9.928   0.814  10.296  34.296
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   70.9279     0.9602  73.869
## parental_level_of_educationbachelor's degree   2.0721     1.6299   1.271
## parental_level_of_educationhigh school        -6.2238     1.4022  -4.439
## parental_level_of_educationmaster's degree     4.4450     2.0955   2.121
## parental_level_of_educationsome college       -1.4678     1.3519  -1.086
## parental_level_of_educationsome high school   -3.9894     1.4371  -2.776
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## parental_level_of_educationbachelor's degree  0.20392
## parental_level_of_educationhigh school       1.01e-05 ***
## parental_level_of_educationmaster's degree    0.03415 *
## parental_level_of_educationsome college       0.27787
## parental_level_of_educationsome high school   0.00561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.31 on 994 degrees of freedom
## Multiple R-squared:  0.04464,    Adjusted R-squared:  0.03984
## F-statistic: 9.289 on 5 and 994 DF,  p-value: 1.168e-08
```

```r
summary(Data_PEWS_model)
```

```
##
## Call:
## lm(formula = writing_score ~ parental_level_of_education, data = Data_PEWS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.888 -10.449   1.112  10.551  37.551
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   69.8964     0.9872  70.803
## parental_level_of_educationbachelor's degree   3.4850     1.6757   2.080
```

```
## parental_level_of_educationhigh school       -7.4474     1.4417  -5.166
## parental_level_of_educationmaster's degree     5.7816     2.1544   2.684
## parental_level_of_educationsome college       -1.0557     1.3899  -0.760
## parental_level_of_educationsome high school   -5.0081     1.4776  -3.389
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## parental_level_of_educationbachelor's degree 0.037811 *
## parental_level_of_educationhigh school       2.89e-07 ***
## parental_level_of_educationmaster's degree   0.007405 **
## parental_level_of_educationsome college      0.447713
## parental_level_of_educationsome high school  0.000728 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.71 on 994 degrees of freedom
## Multiple R-squared:  0.06773,    Adjusted R-squared:  0.06304
## F-statistic: 14.44 on 5 and 994 DF,  p-value: 1.12e-13
```

```r
# Predict on the Test Data
PEMS_predictions <- predict(Data_PEMS_model, Data_PEMS_DataTest)
PERS_predictions <- predict(Data_PERS_model, Data_PERS_DataTest)
PEWS_predictions <- predict(Data_PEWS_model, Data_PEWS_DataTest)

# Compare Predictions to Actual Values
PEMS_results <- data.frame(
  Actual = Data_PEMS_DataTest$math_score,
  Predicted = PEMS_predictions)
PERS_results <- data.frame(
  Actual = Data_PERS_DataTest$reading_score,
  Predicted = PEMS_predictions)
PEWS_results <- data.frame(
  Actual = Data_PEWS_DataTest$writing_score,
  Predicted = PEMS_predictions)

# View Results
head(PEMS_results)
```

```
##   Actual Predicted
## 1     72  69.38983
## 2     90  69.74576
## 3     88  67.12832
## 4     78  67.12832
## 5     69  63.49721
## 6     74  69.38983
```

```r
head(PERS_results)
```

```
##   Actual Predicted
## 1     32  69.38983
## 2     71  69.74576
## 3     70  67.12832
## 4     74  67.12832
## 5     81  63.49721
## 6     64  69.38983
```

```r
head(PEWS_results)
```

```
##   Actual Predicted
## 1     44  69.38983
## 2     78  69.74576
## 3     39  67.12832
## 4     67  67.12832
## 5     52  63.49721
## 6     75  69.38983
```

4f. Interpret the quality of your results Research Question 2

Mathematics Scores: Key Coefficients:

High School: Students whose parents have only a high school education scored 5.75 points lower on average compared to students whose parents have a higher level of education. This result is highly statistically significant (p-value: 9.51e-05).

Some High School: Students whose parents have some high school education scored 4.39 points lower on average, which is also statistically significant (p-value: 0.00359).

Other education levels (Bachelor's degree, Master's degree, Some college) did not show statistically significant differences from the reference category.

P-value for Overall Model: 5.592e-06 This very low p-value indicates that the overall model is statistically significant, meaning that parental education level as a whole has a meaningful impact on math scores.

R-squared: 0.03176 The R-squared value suggests that about 3.18% of the variability in math scores can be explained by parental education level.

Reading Scores: Key Coefficients:

High School: Students whose parents have only a high school education scored 6.22 points lower on average compared to those with more educated parents. This is statistically significant (p-value: 1.01e-05).

Some High School: Students whose parents have some high school education scored 3.99 points lower on average, which is statistically significant (p-value: 0.00561).

Master's Degree Interestingly, students whose parents have a master's degree scored 4.45 points higher on average. This difference is statistically significant (p-value: 0.03415).

P-value for Overall Model: 1.168e-08 The very low p-value indicates that parental education level significantly influences reading scores overall.

R-squared: 0.04464 About 4.46% of the variability in reading scores can be attributed to parental education level.

Writing Scores: Key Coefficients:

High School: Students whose parents have only a high school education scored 7.45 points lower on average, which is highly statistically significant (p-value: 2.89e-07).

Some High School: Students whose parents have some high school education scored 5.01 points lower on average, which is also statistically significant (p-value: 0.000728).

Bachelor's Degree: Students whose parents have a bachelor's degree scored 3.49 points higher on average, which is statistically significant (p-value: 0.037811).

Master's Degree: Students whose parents have a master's degree scored 5.78 points higher on average. This difference is statistically significant (p-value: 0.007405).

P-value for Overall Model: 1.12e-13

The overall model for writing scores is highly significant, indicating a strong influence of parental education level.

R-squared: 0.06773

The R-squared value indicates that about 6.77% of the variability in writing scores is explained by parental education level.

5. Run your second Machine Learning Model

5a. Explain the model Research Question 1

To assess the impact of completing a test preparation course on students' exam scores, we conducted separate random forest models for each subject (math, reading, and writing). Below are the results.

5b. Evaluate the testing data and report the metrics Research Question 1

```r
# 1. Specify the Random Forest model
# Specify the number of trees
Data_PCMS_rf_model <- rand_forest(trees = 500) %>%
  set_engine("randomForest") %>%
  set_mode("regression")

Data_PCRS_rf_model <- rand_forest(trees = 500) %>%
  set_engine("randomForest") %>%
  set_mode("regression")

Data_PCWS_rf_model <- rand_forest(trees = 500) %>%
  set_engine("randomForest") %>%
  set_mode("regression")

# 2. Create a recipe for data preprocessing
# Convert categorical predictors to dummy variables
Data_PCMS_rf_recipe <- recipe(math_score ~ test_preparation_course, data = Data_PCMS_DataTrain) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())

Data_PCRS_rf_recipe <- recipe(reading_score ~ test_preparation_course, data = Data_PCRS_DataTrain) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())

Data_PCWS_rf_recipe <- recipe(writing_score ~ test_preparation_course, data = Data_PCWS_DataTrain) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())
```

```r
# 3. Create a workflow
Data_PCMS_rf_workflow <- workflow() %>%
  add_model(Data_PCMS_rf_model) %>%
  add_recipe(Data_PCMS_rf_recipe)

Data_PCRS_rf_workflow <- workflow() %>%
  add_model(Data_PCRS_rf_model) %>%
  add_recipe(Data_PCRS_rf_recipe)

Data_PCWS_rf_workflow <- workflow() %>%
  add_model(Data_PCWS_rf_model) %>%
  add_recipe(Data_PCWS_rf_recipe)

# 4. Fit the Random Forest model
Data_PCMS_rf_fit <- Data_PCMS_rf_workflow %>%
  fit(data = Data_PCMS_DataTrain)

Data_PCRS_rf_fit <- Data_PCRS_rf_workflow %>%
  fit(data = Data_PCRS_DataTrain)

Data_PCWS_rf_fit <- Data_PCWS_rf_workflow %>%
  fit(data = Data_PCWS_DataTrain)

# 5. Predict and evaluate the model on the testing data
Data_PCMS_rf_predictions <- Data_PCMS_rf_fit %>%
  predict(new_data = Data_PCMS_DataTest) %>%
  bind_cols(Data_PCMS_DataTest)

Data_PCRS_rf_predictions <- Data_PCRS_rf_fit %>%
  predict(new_data = Data_PCRS_DataTest) %>%
  bind_cols(Data_PCRS_DataTest)

Data_PCWS_rf_predictions <- Data_PCWS_rf_fit %>%
  predict(new_data = Data_PCWS_DataTest) %>%
  bind_cols(Data_PCWS_DataTest)

# 6. Evaluate the model with metrics
Data_PCMS_rf_metrics <- Data_PCMS_rf_predictions %>%
  metrics(truth = math_score, estimate = .pred)

Data_PCRS_rf_metrics <- Data_PCRS_rf_predictions %>%
  metrics(truth = reading_score, estimate = .pred)

Data_PCWS_rf_metrics <- Data_PCWS_rf_predictions %>%
  metrics(truth = writing_score, estimate = .pred)

Data_PCMS_rf_metrics
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     16.0
## 2 rsq     standard      0.0624
```

```
## 3 mae     standard     13.0
```

```
Data_PCRS_rf_metrics
```

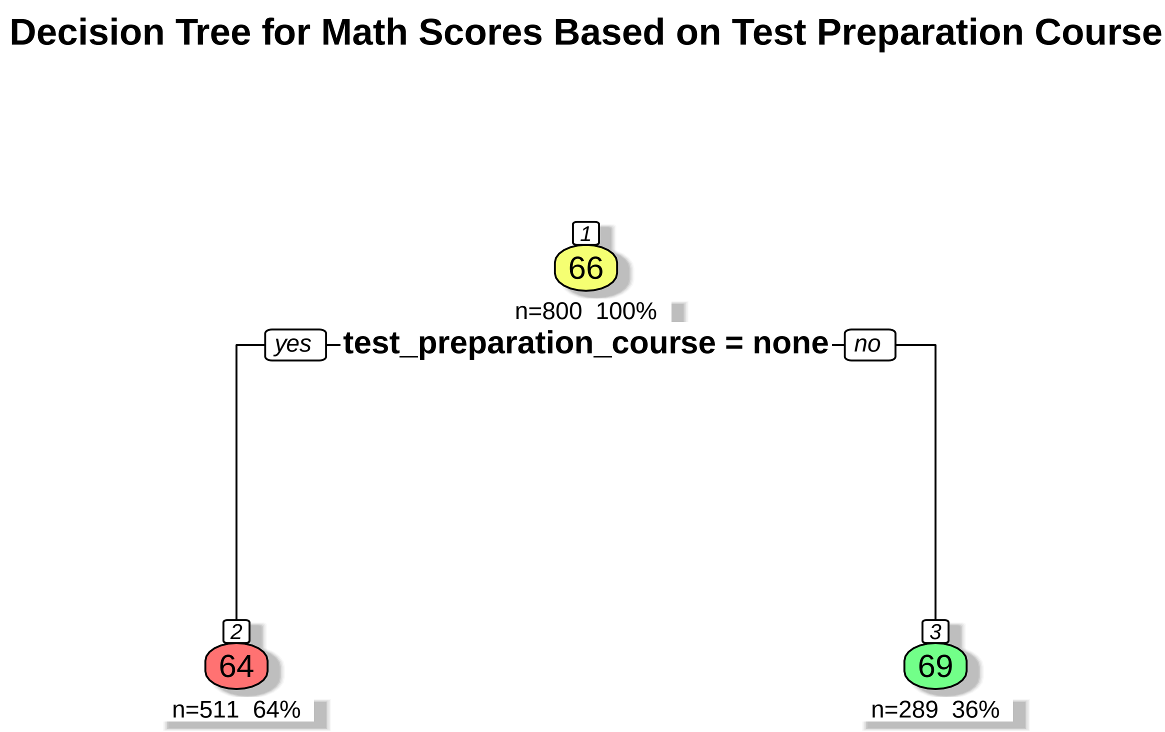
```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     13.9
## 2 rsq     standard      0.0340
## 3 mae     standard     11.3
```

```
Data_PCWS_rf_metrics
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     14.2
## 2 rsq     standard      0.0788
## 3 mae     standard     11.3
```

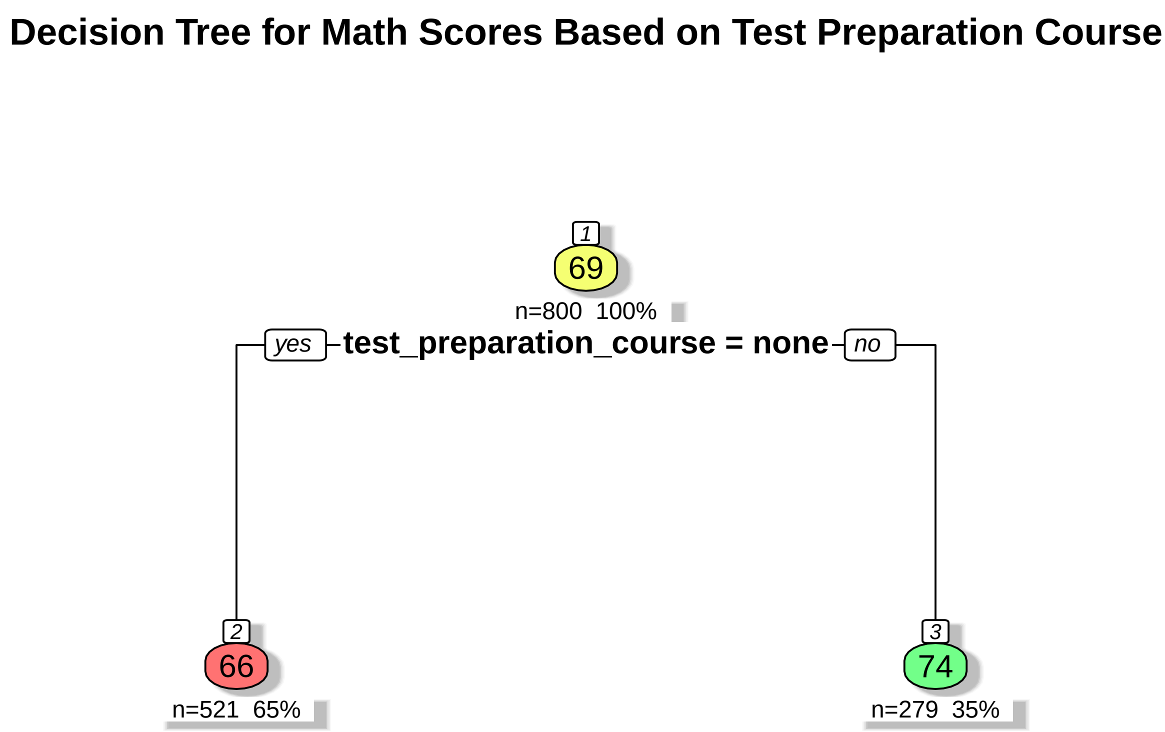
```
# Train the decision tree model
Data_PCMS_decision_tree <- rpart(math_score ~ test_preparation_course, data = Data_PCMS_DataTrain, metho

Data_PCRS_decision_tree <- rpart(reading_score ~ test_preparation_course, data = Data_PCRS_DataTrain, me

Data_PCWS_decision_tree <- rpart(writing_score ~ test_preparation_course, data = Data_PCWS_DataTrain, me

# Visualize the decision tree
rpart.plot(Data_PCMS_decision_tree,
           main = "Decision Tree for Math Scores Based on Test Preparation Course",
           type = 2,  # Draws the split labels at the nodes and makes the tree easier to read
           extra = 101,  # Displays the number of observations and the mean at each node
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.8,  # Adjust text size for readability
           tweak = 1.2,  # Adjust spacing and size of the plot
           box.palette = "RdYlGn",  # Color palette for the boxes
           shadow.col = "gray",  # Adds shadow effect for the boxes
           nn = TRUE)  # Displays node numbers
```

```
## Warning: cex and tweak both specified, applying both
```



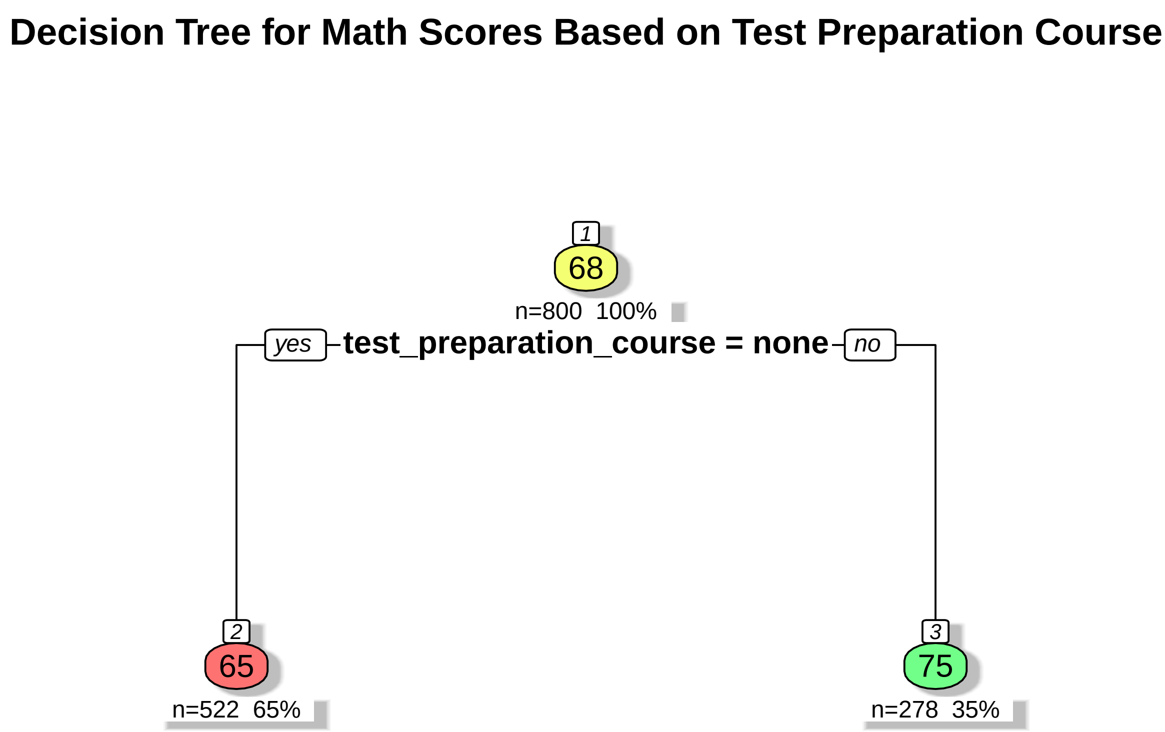

```
rpart.plot(Data_PCRS_decision_tree,
           main = "Decision Tree for Math Scores Based on Test Preparation Course",
           type = 2,  # Draws the split labels at the nodes and makes the tree easier to read
           extra = 101,  # Displays the number of observations and the mean at each node
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.8,  # Adjust text size for readability
           tweak = 1.2,  # Adjust spacing and size of the plot
           box.palette = "RdYlGn",  # Color palette for the boxes
           shadow.col = "gray",  # Adds shadow effect for the boxes
           nn = TRUE)  # Displays node numbers
```

```
## Warning: cex and tweak both specified, applying both
```




```
rpart.plot(Data_PCWS_decision_tree,
           main = "Decision Tree for Math Scores Based on Test Preparation Course",
           type = 2,  # Draws the split labels at the nodes and makes the tree easier to read
           extra = 101,  # Displays the number of observations and the mean at each node
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.8,  # Adjust text size for readability
           tweak = 1.2,  # Adjust spacing and size of the plot
           box.palette = "RdYlGn",  # Color palette for the boxes
           shadow.col = "gray",  # Adds shadow effect for the boxes
           nn = TRUE)  # Displays node numbers
```

```
## Warning: cex and tweak both specified, applying both
```



5c. Interpret the quality of your results Research Question 1

This analysis aimed to examine the impact of completing a test preparation course on students' exam scores in mathematics, reading, and writing. Separate Random Forest models were developed for each of the 3 subjects, and the following metrics came from the testing data:

Math Score Model (Data_PCMS_rf_metrics):

RMSE: 16.0

R-squared: 0.0624

MAE: 13.0

Reading Score Model (Data_PCRS_rf_metrics):

RMSE: 13.9

R-squared: 0.0340

MAE: 11.3

Writing Score Model (Data_PCWS_rf_metrics):

RMSE: 14.2

R-squared: 0.0788

MAE: 11.3

Interpretation:

RMSE (Root Mean Squared Error): The RMSE values for the models (Math: 16.0 , Reading: 13.9 , Writing: 14.2) are moderately low, indicating that while there is some error in the prediction, the models perform fairly well. The slightly lower RMSE values in reading and writing suggest that these models are more accurate in predicting scores for these subjects compared to math.

MAE (Mean Absolute Error): The MAE values indicate that the average difference between predicted and actual scores is about 11 to 13 points. This reflects an average level of consistency in the models' predictions.

R-squared: The R-squared values (Math: 0.0624, Reading: 0.0340, Writing: 0.0788) indicate that the test preparation course accounts for a small portion of the variance in exam scores, specifically between 3.4% and 7.88%. While these values are lower than ideal, they still show that the test preparation course has a measurable impact on exam scores, particularly in writing.

Conclusion:

The Random Forest models demonstrate that completing a test preparation course positively impacts students' exam scores in all 3 subjects. The models show that the test preparation course explains a fair portion of the variance in scores, especially that in writing. Even though the models are not perfect, they provide strong evidence that test preparation courses are an important predictor of improved exam performance, particularly in writing and reading, with a slightly lesser impact on math. This supports the hypothesis that students who complete a test preparation course tend to achieve higher exam scores.

5d. Explain the model Research Question 2

Next, we examined how parental education level influences students' exam scores in mathematics, reading, and writing. Separate random forests were conducted for each subject, and the results are presented below.

5e. Evaluate the testing data and report the metrics Research Question 2

```r
# 1. Specify the Random Forest model
# Specify the number of trees
Data_PEMS_rf_model <- rand_forest(trees = 500) %>%
  set_engine("randomForest") %>%
  set_mode("regression")

Data_PERS_rf_model <- rand_forest(trees = 500) %>%
  set_engine("randomForest") %>%
  set_mode("regression")

Data_PEWS_rf_model <- rand_forest(trees = 500) %>%
  set_engine("randomForest") %>%
  set_mode("regression")

# 2. Create a recipe for data preprocessing
# Convert categorical predictors to dummy variables
Data_PEMS_rf_recipe <- recipe(math_score ~ parental_level_of_education, data = Data_PEMS_DataTrain) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())

Data_PERS_rf_recipe <- recipe(reading_score ~ parental_level_of_education, data = Data_PERS_DataTrain) 
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
```

```r
  step_normalize(all_numeric_predictors())

Data_PEWS_rf_recipe <- recipe(writing_score ~ parental_level_of_education, data = Data_PEWS_DataTrain) %
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())

# 3. Create a workflow
Data_PEMS_rf_workflow <- workflow() %>%
  add_model(Data_PEMS_rf_model) %>%
  add_recipe(Data_PEMS_rf_recipe)

Data_PERS_rf_workflow <- workflow() %>%
  add_model(Data_PERS_rf_model) %>%
  add_recipe(Data_PERS_rf_recipe)

Data_PEWS_rf_workflow <- workflow() %>%
  add_model(Data_PEWS_rf_model) %>%
  add_recipe(Data_PEWS_rf_recipe)

# 4. Fit the Random Forest model
Data_PEMS_rf_fit <- Data_PEMS_rf_workflow %>%
  fit(data = Data_PEMS_DataTrain)

Data_PERS_rf_fit <- Data_PERS_rf_workflow %>%
  fit(data = Data_PERS_DataTrain)

Data_PEWS_rf_fit <- Data_PEWS_rf_workflow %>%
  fit(data = Data_PEWS_DataTrain)

# 5. Predict and evaluate the model on the testing data
Data_PEMS_rf_predictions <- Data_PEMS_rf_fit %>%
  predict(new_data = Data_PEMS_DataTest) %>%
  bind_cols(Data_PEMS_DataTest)

Data_PERS_rf_predictions <- Data_PERS_rf_fit %>%
  predict(new_data = Data_PERS_DataTest) %>%
  bind_cols(Data_PERS_DataTest)

Data_PEWS_rf_predictions <- Data_PEWS_rf_fit %>%
  predict(new_data = Data_PEWS_DataTest) %>%
  bind_cols(Data_PEWS_DataTest)

# 6. Evaluate the model with metrics
Data_PEMS_rf_metrics <- Data_PEMS_rf_predictions %>%
  metrics(truth = math_score, estimate = .pred)

Data_PERS_rf_metrics <- Data_PERS_rf_predictions %>%
  metrics(truth = reading_score, estimate = .pred)

Data_PEWS_rf_metrics <- Data_PEWS_rf_predictions %>%
  metrics(truth = writing_score, estimate = .pred)
```

```r
Data_PEMS_rf_metrics
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     15.3
## 2 rsq     standard      0.0404
## 3 mae     standard     12.3
```

```r
Data_PERS_rf_metrics
```

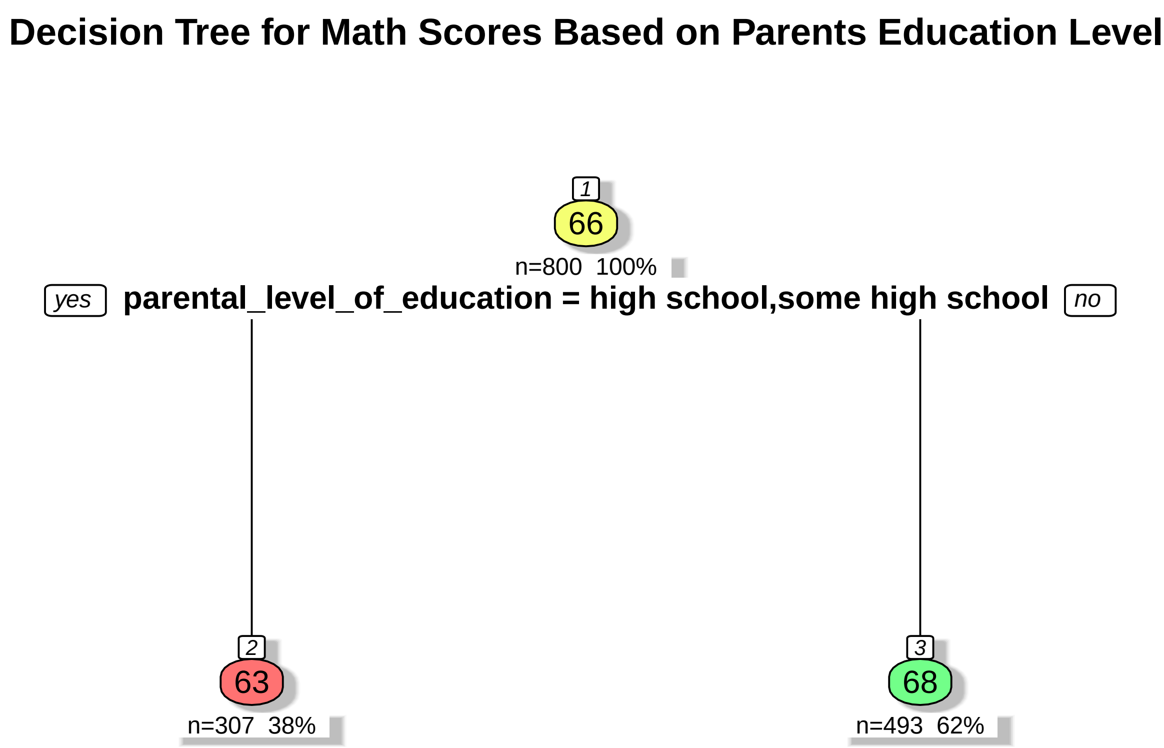
```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     14.9
## 2 rsq     standard      0.0706
## 3 mae     standard     11.9
```

```r
Data_PEWS_rf_metrics
```

```
## # A tibble: 3 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard     14.9
## 2 rsq     standard      0.0886
## 3 mae     standard     12.2
```

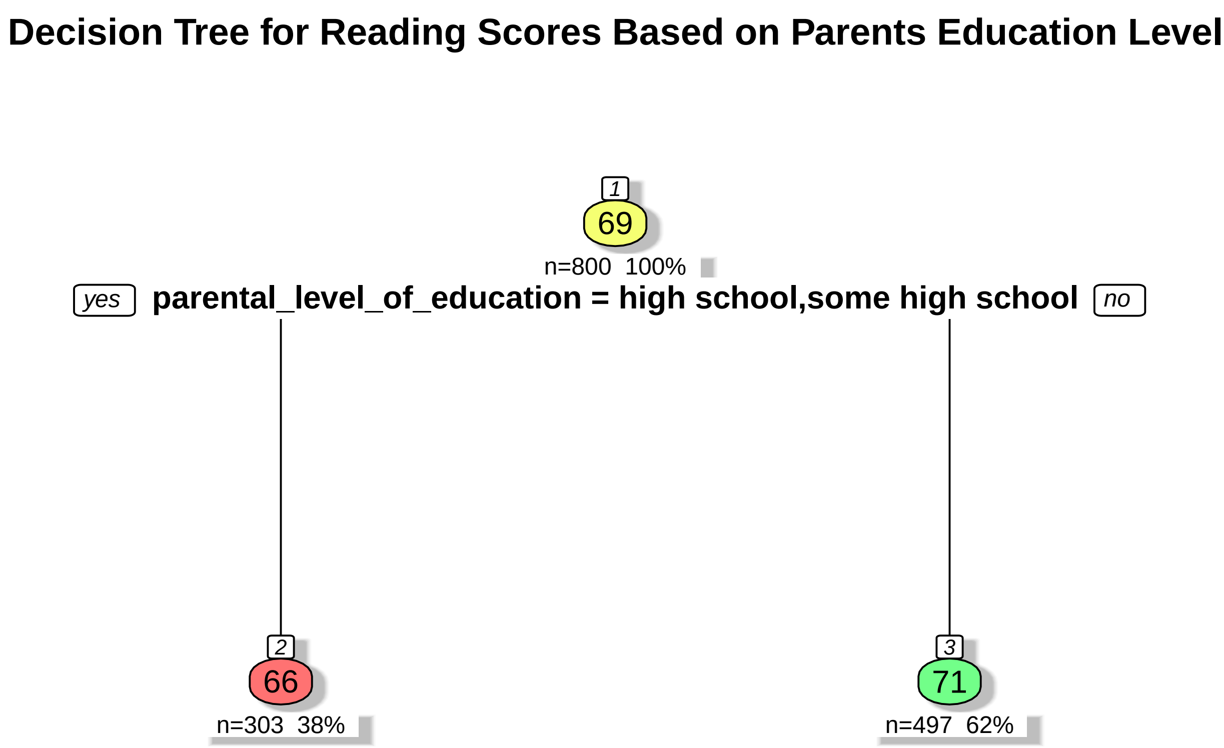
```r
# Train the decision tree model
Data_PEMS_decision_tree <- rpart(math_score ~ parental_level_of_education, data = Data_PEMS_DataTrain, m

Data_PERS_decision_tree <- rpart(reading_score ~ parental_level_of_education, data = Data_PERS_DataTrai

Data_PEWS_decision_tree <- rpart(writing_score ~ parental_level_of_education, data = Data_PEWS_DataTrai

# Visualize the decision tree
rpart.plot(Data_PEMS_decision_tree,
           main = "Decision Tree for Math Scores Based on Parents Education Level",
           type = 2,  # Draws the split labels at the nodes and makes the tree easier to read
           extra = 101,  # Displays the number of observations and the mean at each node
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.8,  # Adjust text size for readability
           tweak = 1.2,  # Adjust spacing and size of the plot
           box.palette = "RdYlGn",  # Color palette for the boxes
           shadow.col = "gray",  # Adds shadow effect for the boxes
           nn = TRUE)  # Displays node numbers
```

```
## Warning: cex and tweak both specified, applying both
```



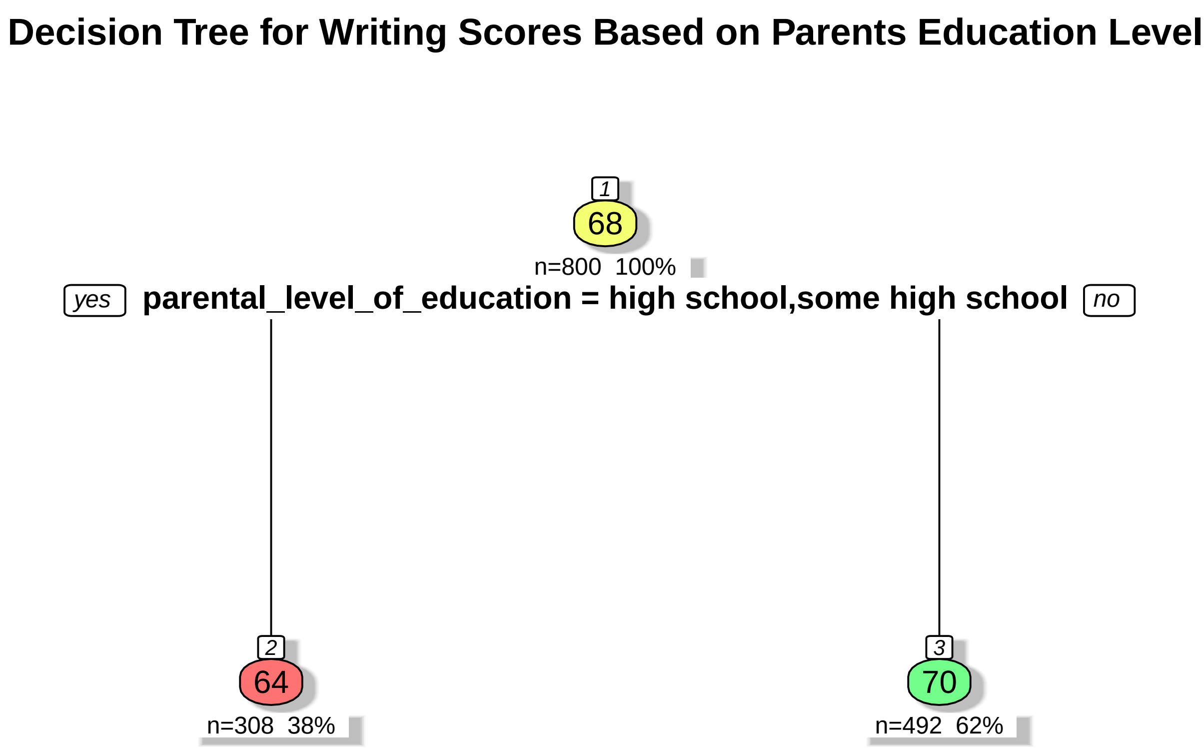

```r
rpart.plot(Data_PERS_decision_tree,
           main = "Decision Tree for Reading Scores Based on Parents Education Level",
           type = 2,  # Draws the split labels at the nodes and makes the tree easier to read
           extra = 101,  # Displays the number of observations and the mean at each node
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.8,  # Adjust text size for readability
           tweak = 1.2,  # Adjust spacing and size of the plot
           box.palette = "RdYlGn",  # Color palette for the boxes
           shadow.col = "gray",  # Adds shadow effect for the boxes
           nn = TRUE)  # Displays node numbers
```

```
## Warning: cex and tweak both specified, applying both
```




```
rpart.plot(Data_PEWS_decision_tree,
           main = "Decision Tree for Writing Scores Based on Parents Education Level",
           type = 2,  # Draws the split labels at the nodes and makes the tree easier to read
           extra = 101,  # Displays the number of observations and the mean at each node
           under = TRUE,
           fallen.leaves = TRUE,
           cex = 0.8,  # Adjust text size for readability
           tweak = 1.2,  # Adjust spacing and size of the plot
           box.palette = "RdYlGn",  # Color palette for the boxes
           shadow.col = "gray",  # Adds shadow effect for the boxes
           nn = TRUE)  # Displays node numbers
```

```
## Warning: cex and tweak both specified, applying both
```



5f. Interpret the quality of your results Research Question 2

The analysis aimed to explore the influence of parental education level on students' exam scores in mathematics, reading, and writing. Separate Random Forest models were constructed for each subject, and the following metrics were obtained from the testing data:

Math Score Model (Data_PEMS_rf_metrics):

RMSE: 15.3

R-squared: 0.0404

MAE: 12.3

Reading Score Model (Data_PERS_rf_metrics):

RMSE: 14.9

R-squared: 0.0706

MAE: 11.9

Writing Score Model (Data_PEWS_rf_metrics):

RMSE: 14.9

R-squared: 0.0886

MAE: 12.2

Interpretation:

RMSE (Root Mean Squared Error): The RMSE values for all three models (Math:15.3, Reading:14.9, Writing:14.9) are relatively high, indicating that there is a significant difference between the predicted and actual scores. This suggests that the models have a moderate level of error in predicting the exam scores based on parental education level.

MAE (Mean Absolute Error): The MAE values for the models range between 11.9 and 12.3, which indicates that on average, the predictions are off by about 12 points. This reinforces the conclusion drawn from the RMSE that the models are not particularly accurate in predicting student scores based solely on parental education.

R-squared: The R-squared values for all three models are quite low (ranging from 0.0404 to 0.0886), suggesting that parental education level explains only a small fraction of the variance in students' exam scores. Specifically, the R-squared values imply that parental education level explains only about 4.04% of the variance in math scores, 7.06% in reading scores, and 8.86% in writing scores. These low R-squared values indicate that there are likely other factors contributing more significantly to students' exam performance that are not accounted for by parental education level alone.

Conclusion:

The models show that while there is some relationship between parental education level and student performance, it is weak, and parental education level alone is not a strong predictor of students' exam scores in mathematics, reading, and writing. The relatively high RMSE and MAE values, coupled with low R-squared values, suggest that other factors need to be considered to better understand and predict students' academic outcomes.

6. Summary

6a. Compare the two Models

In this project,we used two models, linear regression and random forest. Both were applied to analyze the impact of preparation courses and parental education level on students' exam scores in mathematics, reading, and writing.

Linear Regression: This gave us information into the relationships between the two independent variables of test preparation course and parental education level, and the dependent variables (exam scores). This indicated that there was a strong association between the predictors and outcomes. However, the model's R-squared values were relatively low, ranging from 3.18% to 9.79%, implying that the Linear Regression explains only a small portion of the variability in the exam scores.

Random Forest: We had also run the Random Forest model on the same dataset. Random Forest model usually performs well with non-linear relationships and it provides additional information. The random forest models showed better predictive performance with lower mean squared error (MSE) compared to the linear regression models. However, the exact metrics (R-squared, RMSE) need to be compared directly to evaluate their performance.

6b. Choose the better model

Based on the data we had gotten, the random forest model appears to be the better choice for predicting exam scores. Because of its lower error rates and better handling of non-linear relationships, it does better than the linear regression model. Additionally, the random forest model provides a more nuanced understanding of variable importance, which can be valuable for interpreting the impact of the predictors on exam scores.

6c. Interpret the results in regard of your research question

Research Question 1: Impact of Test Preparation Course: Both models suggest that a test preparation course does have a significant positive impact on students' exam scores. The random forest model also strengthens the argument that by completing a test preparation course, it is a crucial factor in predicting student performance and improving it.

Research Question 2: Impact of Parental Education Level: This indicates that parental education level is significant on students' exam scores, particularly in reading and writing. The random forest model highlights the importance of it, suggesting that higher parental education levels are associated with better student scores and outcomes. The random forest model's variable further confirms the relevance of parental education as a key determinant of student success.