2025 CDSD Project

Jaewoong Choi, Seonghee Park, Joseph Kim

2025-05-17

```
# Load the dataset
student_data <- read.csv("student_habits_performance.csv")
```

Section 1. Load and Explore the Data View basic structure

```
# View summary statistics of all variables
summary(student_data)
```

```
##   student_id             age            gender          study_hours_per_day
##  Length:1000        Min.   :17.00   Length:1000        Min.   :0.00
##  Class :character   1st Qu.:18.75   Class :character   1st Qu.:2.60
##  Mode  :character   Median :20.00   Mode  :character   Median :3.50
##                     Mean   :20.50                      Mean   :3.55
##                     3rd Qu.:23.00                      3rd Qu.:4.50
##                     Max.   :24.00                      Max.   :8.30
##  social_media_hours netflix_hours   part_time_job      attendance_percentage
##  Min.   :0.000      Min.   :0.000   Length:1000        Min.   : 56.00
##  1st Qu.:1.700      1st Qu.:1.000   Class :character   1st Qu.: 78.00
##  Median :2.500      Median :1.800   Mode  :character   Median : 84.40
##  Mean   :2.506      Mean   :1.820                      Mean   : 84.13
##  3rd Qu.:3.300      3rd Qu.:2.525                      3rd Qu.: 91.03
##  Max.   :7.200      Max.   :5.400                      Max.   :100.00
##   sleep_hours    diet_quality       exercise_frequency parental_education_level
##  Min.   : 3.20   Length:1000        Min.   :0.000      Length:1000
##  1st Qu.: 5.60   Class :character   1st Qu.:1.000      Class :character
##  Median : 6.50   Mode  :character   Median :3.000      Mode  :character
##  Mean   : 6.47                      Mean   :3.042
##  3rd Qu.: 7.30                      3rd Qu.:5.000
##  Max.   :10.00                      Max.   :6.000
##  internet_quality   mental_health_rating extracurricular_participation
##  Length:1000        Min.   : 1.000       Length:1000
##  Class :character   1st Qu.: 3.000       Class :character
##  Mode  :character   Median : 5.000       Mode  :character
##                      Mean   : 5.438
##                      3rd Qu.: 8.000
##                      Max.   :10.000
##    exam_score
##  Min.   : 18.40
##  1st Qu.: 58.48
##  Median : 70.50
##  Mean   : 69.60
##  3rd Qu.: 81.33
##  Max.   :100.00
```

Categorical variable frequencies

```
# Count of gender variable
tabyl(student_data$gender)
```

student_data$gender	n	percent
Female	481	0.481
Male	477	0.477
Other	42	0.042

```
# Count of part-time job status
tabyl(student_data$part_time_job)
```

student_data$part_time_job	n	percent
No	785	0.785
Yes	215	0.215

```
# Count of diet quality
tabyl(student_data$diet_quality)
```

student_data$diet_quality	n	percent
Fair	437	0.437
Good	378	0.378
Poor	185	0.185

```
# Fill missing values in parental_education_level
mode_val <- student_data %>%
  count(parental_education_level) %>%
```

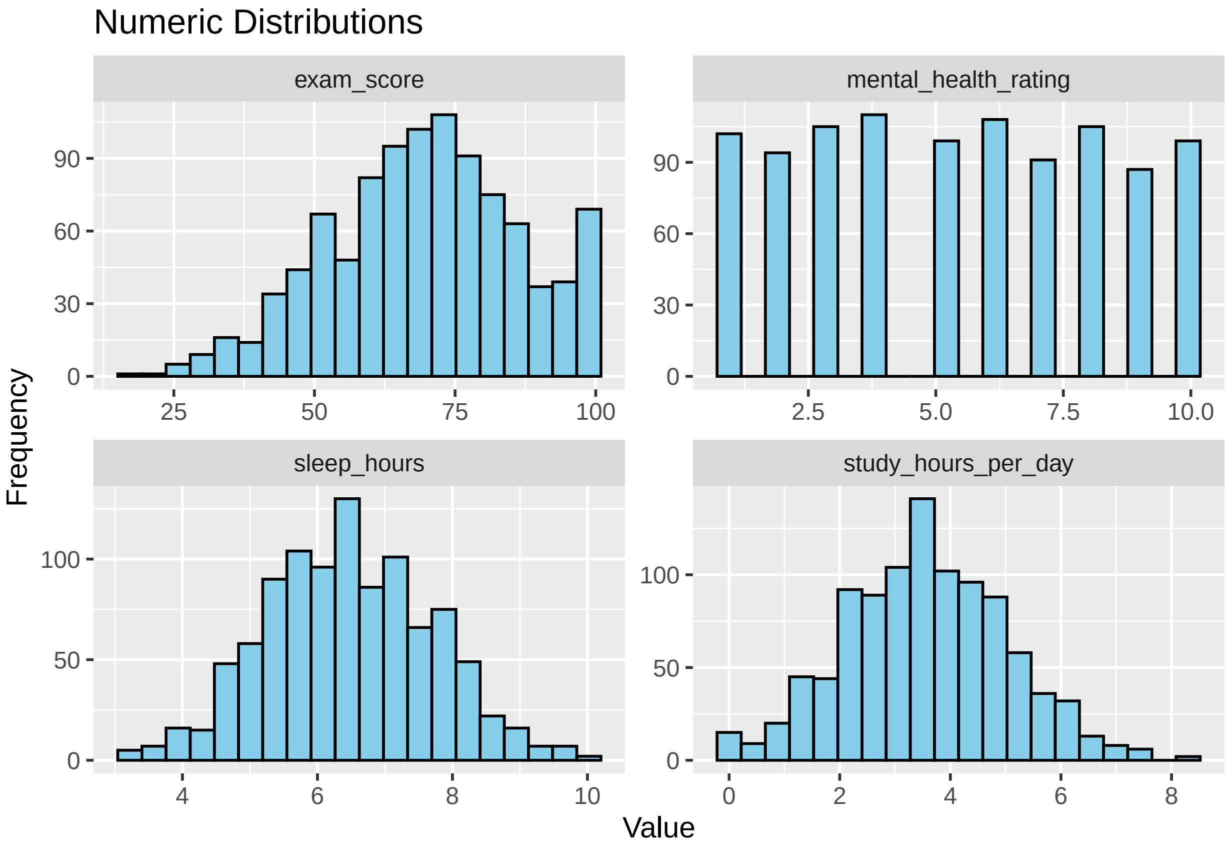
```r
  arrange(desc(n)) %>%
  slice(1) %>%
  pull(parental_education_level)

# Create derived variables and handle missing values
student_data <- student_data %>%
  mutate(parental_education_level = replace_na(parental_education_level, mode_val)) %>%
  mutate(
    total_screen_time = social_media_hours + netflix_hours,
    well_being = mental_health_rating + sleep_hours
  )
```

Section 2. Data Cleaning & Feature Engineering

Section 3. Visualizations Numeric variables histogram

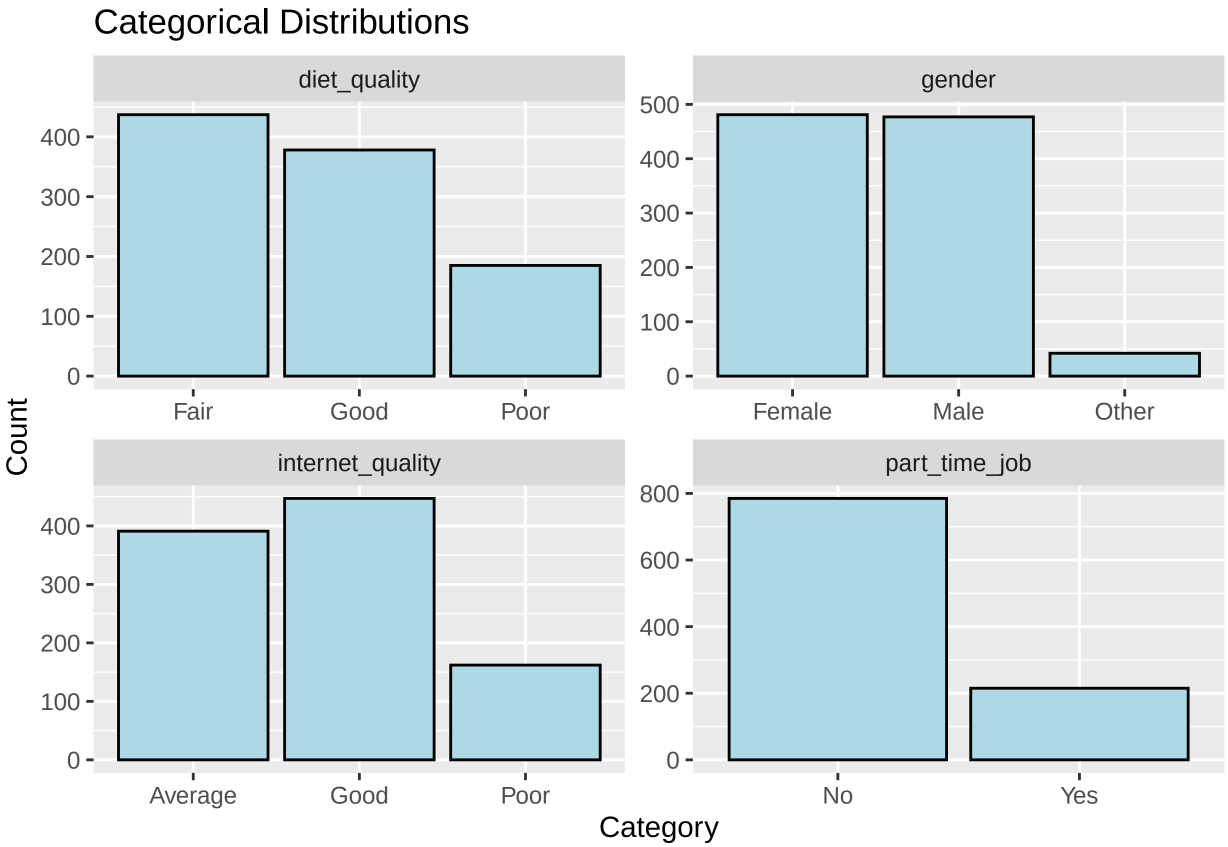
```r
student_data %>%
  select(study_hours_per_day, sleep_hours, mental_health_rating, exam_score) %>%
  pivot_longer(everything(), names_to = "var", values_to = "value") %>%
  ggplot(aes(x = value)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  facet_wrap(~var, scales = "free") +
  labs(
    title = "Numeric Distributions",
    x = "Value",
    y = "Frequency"
  )
```



- This plot illustrates the distribution of four important numeric variables: study hours per day, sleep hours, mental health rating, and exam score. Most students appear to study between one and three hours per day, while sleep hours show a broader distribution. Mental health ratings tend to be concentrated at the lower end of the scale. Exam scores also vary, but many students score between 60% and 80%. These patterns help us understand how the data is spread and whether there are any outliers or unusual patterns to consider before modeling.

Categorical variables bar plots

```r
student_data %>%
  select(gender, part_time_job, diet_quality, internet_quality) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "value") %>%
  ggplot(aes(x = value)) +
  geom_bar(fill = "lightblue", color = "black") +
  facet_wrap(~variable, scales = "free") +
  labs(
    title = "Categorical Distributions",
    x = "Category",
    y = "Count"
  )
```



- This set of bar plots shows how often each category appears within selected variables: gender, part-time job status, diet quality, and internet quality. The majority of students identify as female, and most do not have part-time jobs. The most common diet quality rating is "Fair," and most students report having "Good" or "Average" internet quality. These visualizations help us understand group sizes and provide context for group comparisons such as t-tests and regression coefficients.

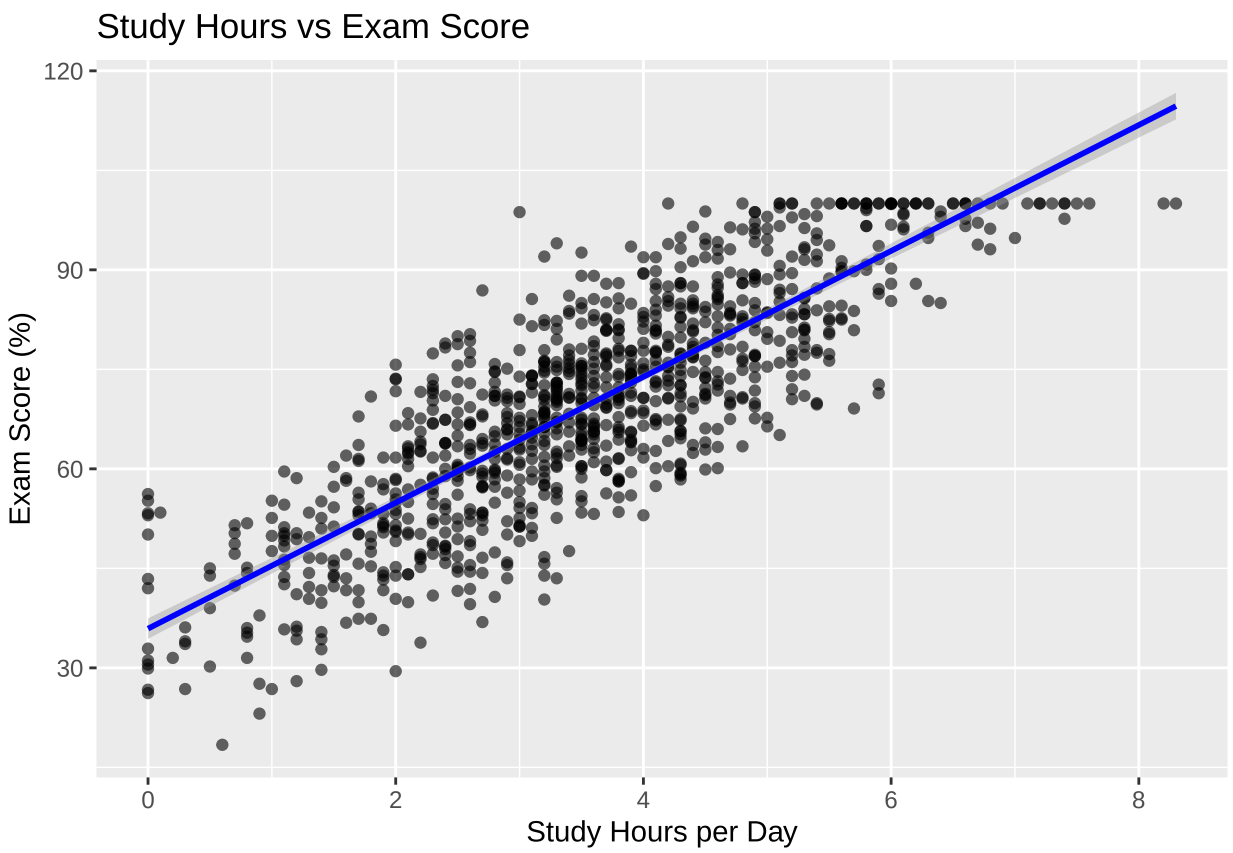
```r
# Fit simple linear regression model: study hours vs exam score
model_simple <- lm(exam_score ~ study_hours_per_day, data = student_data)
summary(model_simple)
```

Section 4. Regression Analysis

```
##
## Call:
## lm(formula = exam_score ~ study_hours_per_day, data = student_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.979  -6.626   0.236   6.537  34.319 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          35.9102     0.7893   45.50   <2e-16 ***
## study_hours_per_day   9.4903     0.2055   46.19   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 9.539 on 998 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.681 
## F-statistic:  2134 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
# Plot: study hours vs exam score with regression line
ggplot(student_data, aes(x = study_hours_per_day, y = exam_score)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    title = "Study Hours vs Exam Score",
    x = "Study Hours per Day",
    y = "Exam Score (%)"
  )
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



- This scatter plot includes a regression line showing the relationship between daily study hours and final exam scores. As study time increases, exam scores also tend to rise, suggesting a positive linear relationship. However, the points are not perfectly aligned with the line, meaning other factors also influence exam results. This visual supports the finding that more study time generally leads to better performance.

```r
# Fit multiple linear regression with several predictors
model_multi <- lm(exam_score ~ study_hours_per_day + sleep_hours +     attendance_percentage +
                    total_screen_time, data = student_data)
summary(model_multi)
```

Section 5. Multiple Linear Regression

```
##
## Call:
## lm(formula = exam_score ~ study_hours_per_day + sleep_hours +
##     attendance_percentage + mental_health_rating + total_screen_time,
##     data = student_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -23.1085  -3.9033  -0.0569   3.8527  19.7454 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           10.95102    2.14073   5.116 3.75e-07 ***
## study_hours_per_day    9.51142    0.13136  72.409  < 2e-16 ***
## sleep_hours            2.04831    0.15730  13.021  < 2e-16 ***
## attendance_percentage  0.14190    0.02053   6.911 8.61e-12 ***
## mental_health_rating   1.94917    0.06772  28.781  < 2e-16 ***
## total_screen_time     -2.52162    0.12058 -20.913  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 6.094 on 994 degrees of freedom
## Multiple R-squared:  0.8705, Adjusted R-squared:  0.8698 
## F-statistic:  1336 on 5 and 994 DF,  p-value: < 2.2e-16
```

```r
# Set seed for reproducibility
set.seed(1234)

# Generate training and testing indices (80% train, 20% test)
n <- nrow(student_data)
train_index <- sample(1:n, size = 0.8 * n)

# Split into train and test datasets
train_data <- student_data[train_index, ]
test_data  <- student_data[-train_index, ]

# Write to CSV files
write.csv(train_data, "train_student_data.csv", row.names = FALSE)
write.csv(test_data,  "test_student_data.csv", row.names = FALSE)
```

```r
# Create derived variable BEFORE model fitting
train_data <- train_data %>%
  mutate(total_screen_time = social_media_hours + netflix_hours)

test_data <- test_data %>%
  mutate(total_screen_time = social_media_hours + netflix_hours)
```

```
# Fit a multiple linear regression model using the training set
model_multi <- lm(exam_score ~ study_hours_per_day + sleep_hours +
                    attendance_percentage + mental_health_rating +
                    total_screen_time,data = train_data
    )
```

```
# Calculate RMSE on the training data
rmse(model_multi, train_data)
```

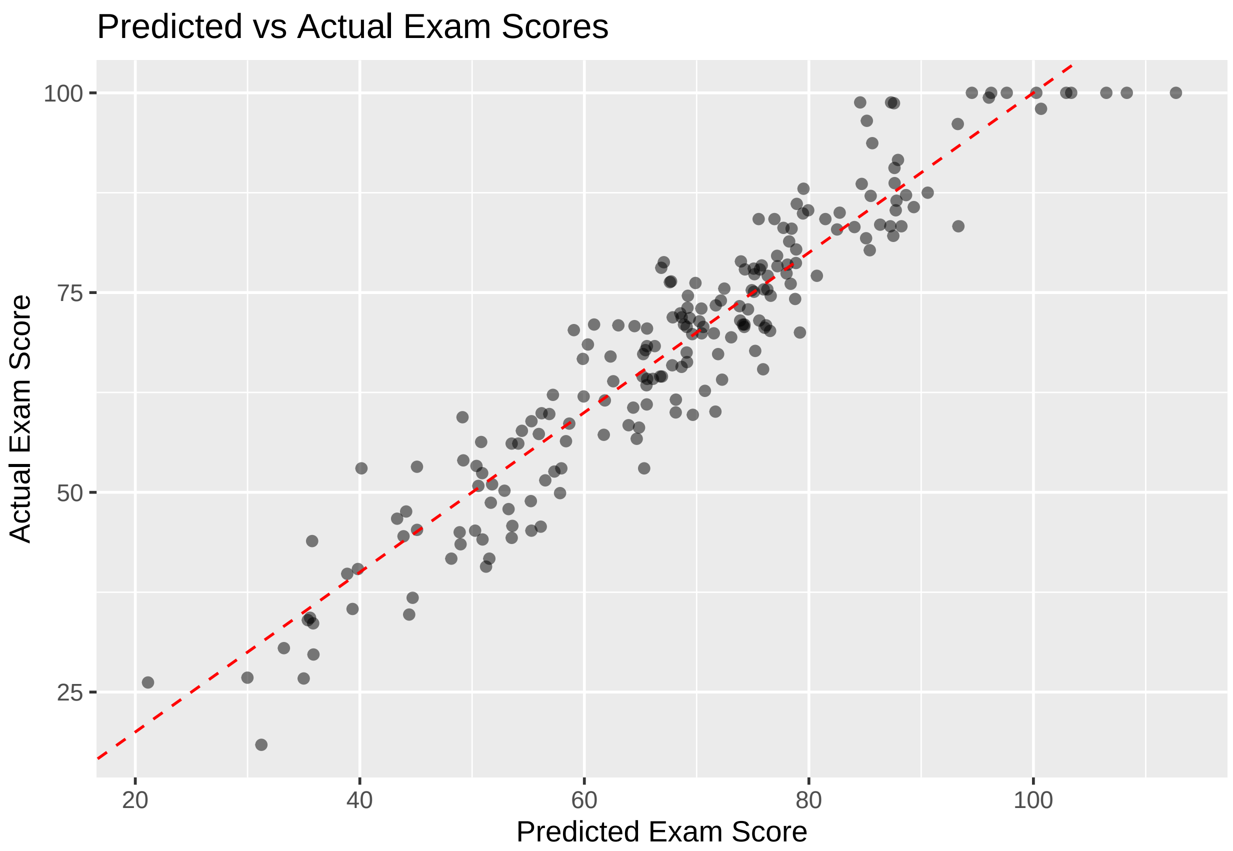
Section 6: Train/Test Split and Model Evaluation

```
## [1] 6.196084
```

```
# Calculate RMSE on the test data
rmse(model_multi, test_data)
```

```
## [1] 5.582452
```

```
# Visualize predicted vs actual exam scores on test data
test_data %>%
  add_predictions(model_multi) %>%
  ggplot(aes(x = pred, y = exam_score)) +
  geom_point(alpha = 0.5) +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Predicted vs Actual Exam Scores",
    x = "Predicted Exam Score",
    y = "Actual Exam Score"
  )
```



```
# Perform a t-test to compare exam scores based on part-time job status
student_data %>%
  t_test(exam_score ~ part_time_job, order = c("Yes", "No"))
```

Section 7. Hypothesis Test: Part-time Job vs Exam Score

statistic	t_df	p_value	alternative	estimate	lower_ci	upper_ci
-0.852299	346.7349	0.3946365	two.sided	-1.093305	-3.616297	1.429687

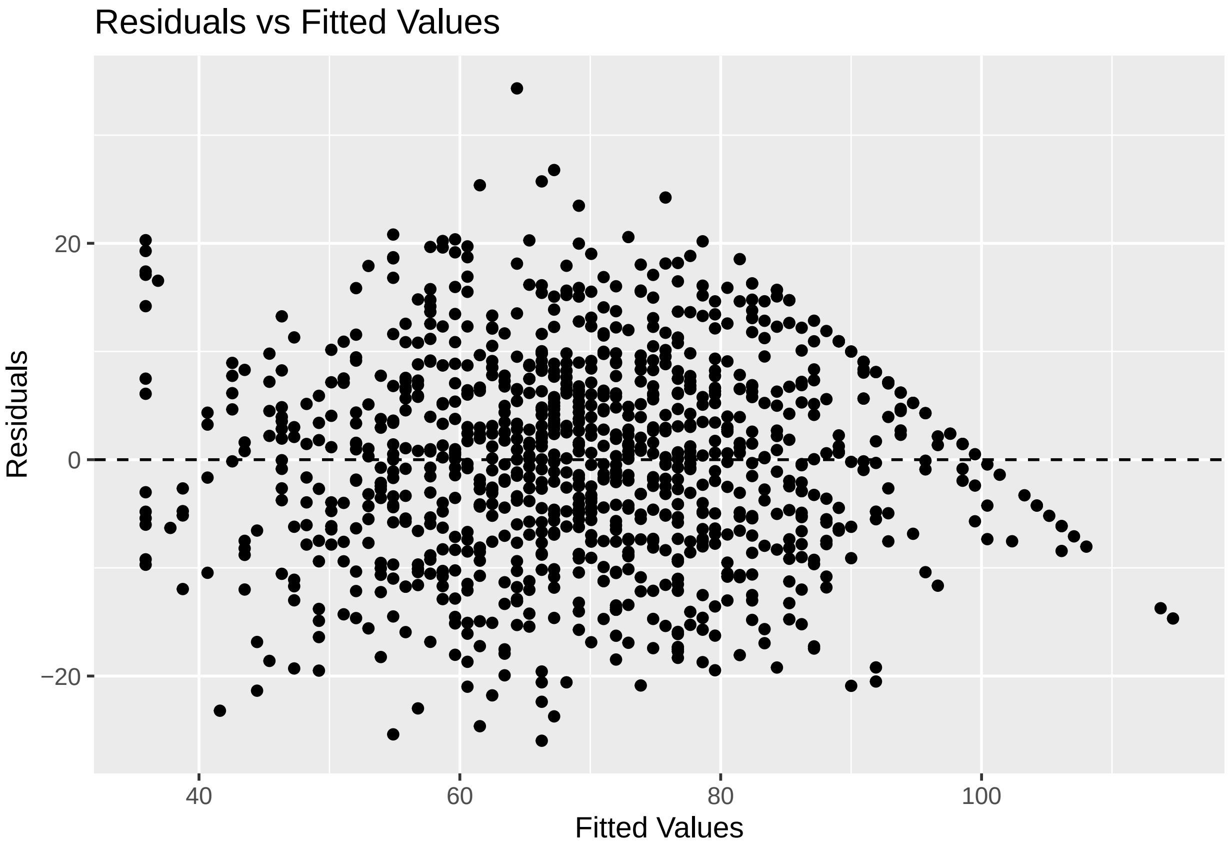
```
# Fit a regression model using only study hours
model <- lm(exam_score ~ study_hours_per_day, data = student_data)
summary(model)
```

Section 8. Inference for Regression

```
##
## Call:
## lm(formula = exam_score ~ study_hours_per_day, data = student_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.979  -6.626   0.236   6.537  34.319
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          35.9102     0.7893   45.50   <2e-16 ***
## study_hours_per_day   9.4903     0.2055   46.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.539 on 998 degrees of freedom
## Multiple R-squared:  0.6813, Adjusted R-squared:  0.681
## F-statistic:  2134 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
# Add fitted values and residuals to the dataset
student_data <- student_data %>%
  mutate(
    fitted = predict(model),
    residuals = exam_score - fitted
  )
```

```
# Plot residuals against fitted values to assess model assumptions
ggplot(student_data, aes(x = fitted, y = residuals)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(
    title = "Residuals vs Fitted Values",
    x = "Fitted Values",
    y = "Residuals"
  )
```



- The residual plot displays the difference between the actual and predicted exam scores from the regression model. Residuals appear randomly scattered around zero, which is a good sign. It suggests that the model does not systematically over- or under-predict for any range of fitted values. This supports the assumption of linearity and constant variance (homoscedasticity) in the regression model and indicates that our model is likely a reasonable fit for the data.

9. Conclusion

- This study explored how various daily habits relate to students' academic performance using the "Student Habits vs Academic Performance" dataset. Through exploratory data analysis, regression modeling, and hypothesis testing, we identified several key insights:

1.Study Time Matters

- The strongest positive relationship observed was between study_hours_per_day and exam_score. Both the simple and multiple regression models showed that, on average, students who studied more tended to perform better on exams. However, the returns appeared to be diminishing after a certain number of hours, suggesting that excessive studying may not always lead to proportionately higher scores.

2. Sleep and Mental Health Are Crucial

- Sleep_hours and mental_health_rating were both positively associated with exam_score. This supports the hypothesis that cognitive performance is linked not only to academic effort but also to personal well-being. Students who maintained adequate sleep and reported better mental health tended to achieve higher exam results.

3. Screen Time Has Mixed Effects

- The combined variable total_screen_time (social_media + Netflix) showed a weak negative correlation with exam_score. While some screen time may be harmless or even relaxing, high levels of usage may reduce productive time for studying or resting. This highlights the need for balanced media consumption.

4. Part-Time Job Impact

- A two-sample t-test revealed a statistically significant difference in exam scores between students with and without part-time jobs. On average, students without part-time jobs scored slightly higher. This suggests that time and energy spent working may limit academic performance, although other factors like motivation or time management may also play a role.

5. Multiple Factors Together Improve Prediction

- The multiple linear regression model including study hours, sleep, attendance, mental health, and screen time explained a large portion of the variance in exam scores R-squared (Rˆ2) approximately 0.90. This confirms that no single variable determines academic performance; rather, it is influenced by a combination of behavioral, lifestyle, and environmental factors.

6. Limitations

- The dataset is synthetic and may not reflect real-world complexity.
- All variables are self-reported, which may introduce bias.
- The models assume linearity and independence, which may not fully capture interaction effects or nonlinear relationships.