Analysis of Student Traits and Exam Performance

Julius Spears

2025-05-24

Data Source: Kaggle

Data Format: CSV

URL: https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance/data

Introduction

Student traits play a critical role in shaping a student's academic performance. In this project, I seek to provide insight into the strength and effect of specific traits on academic progress. This project utilizes simulated data based off of real-world trends of student habits across diverse populations. It consists of a sample of 1,000 student records, including traits such as study hours, sleep hours, diet, social media hours, and mental health ratings. The initial analysis will consist of examining how the study habits of students impacts their exam scores.

Study Hours and Exam Scores

Study habits are a key predictor for academic performance. This section explores the relationship between the number of hours students study each day and their exam scores.

We begin with summary statistics and a probability density estimate for daily study hours across the 1,000 students in the data set.

Summary Statistics

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    2.60    3.50    3.55    4.50    8.30
```

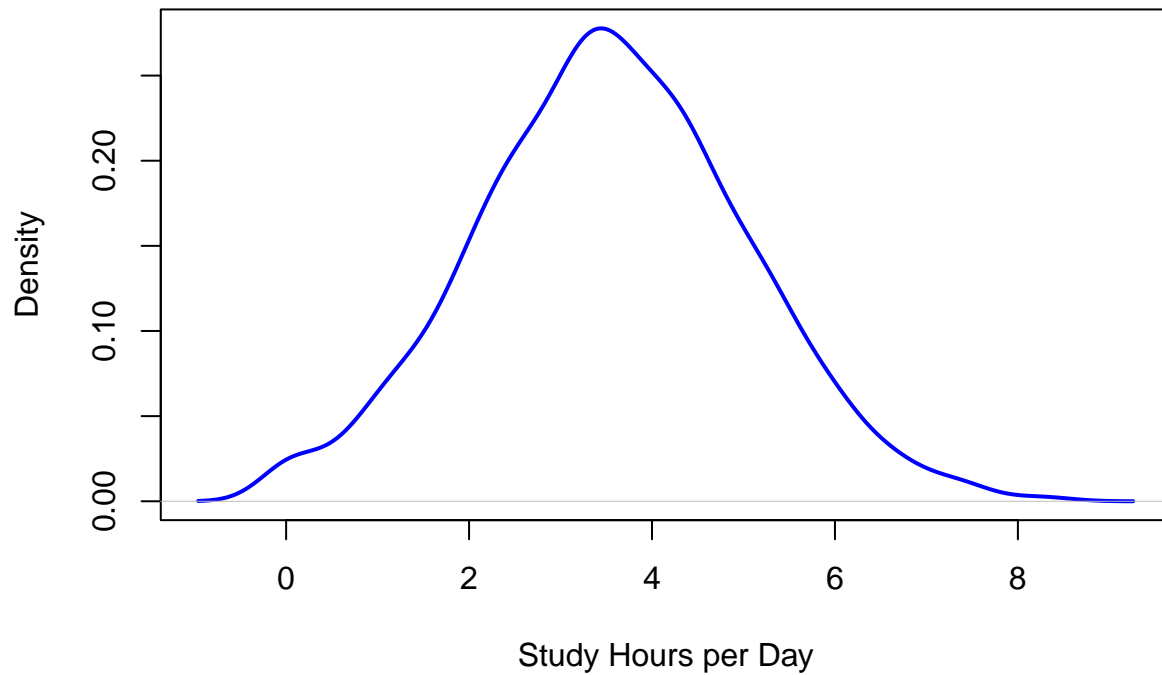**Standard Deviation**

```
## [1] 1.46889
```

Variance

```
## [1] 2.157638
```

Probability Density for Study Hours per Day of Students

From the data, the average study time is approximately 3.5 hours per day, with most students studying between 2 and 5 hours.

To further analyze this trait, study hours were categorized into three groups:

- **Below Average:** less than 2 hrs/day

- **Average:** between 2 and 5 hrs/day

- **Above Average:** more than 5 hrs/day

```
## [1] 0.0 6.9 1.4 1.0 5.0 7.2
```
```
## [1] Below Average Above Average Below Average Below Average Average
## [6] Above Average
## Levels: Below Average Average Above Average
```

Summary statistics were then computed to examine differences in exam performance across the three study habit groups.

Mean

```
## Below Average       Average Above Average
##      45.56391      69.47809      90.79419
```

Median

```
## Below Average       Average Above Average
##         46.30         70.15         93.40
```

Standard Deviation

```
## Below Average        Average Above Average
##     10.002199      12.538205     9.357435
```

Variance

```
## Below Average        Average Above Average
##     100.04399      157.20658     87.56159
```

Exam scores clearly vary across study hour groups. Both the mean and median scores increase with more study time, indicating a strong positive association. Variances are close enough across groups to satisfy the equal variance assumption required for ANOVA (Analysis of Variance) testing.

ANOVA TEST RESULTS

An ANOVA test was conducted to evaluate whether the observed differences in group means are statistically significant.

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## studhabts$study_hours_per_day   1 194133  194133    2134 <2e-16 ***
## Residuals                     998  90805      91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value is less than 2×10^{-16}, well below the conventional threshold of 0.05. Therefore, we reject the null hypothesis that group means are equal, concluding that study time has a highly significant effect on exam performance. In particular, students who study more tend to achieve substantially higher scores.

Below, we visualize the relationship using box plots and kernel density plots:

Boxplots of Exam Score Based on Study Amount
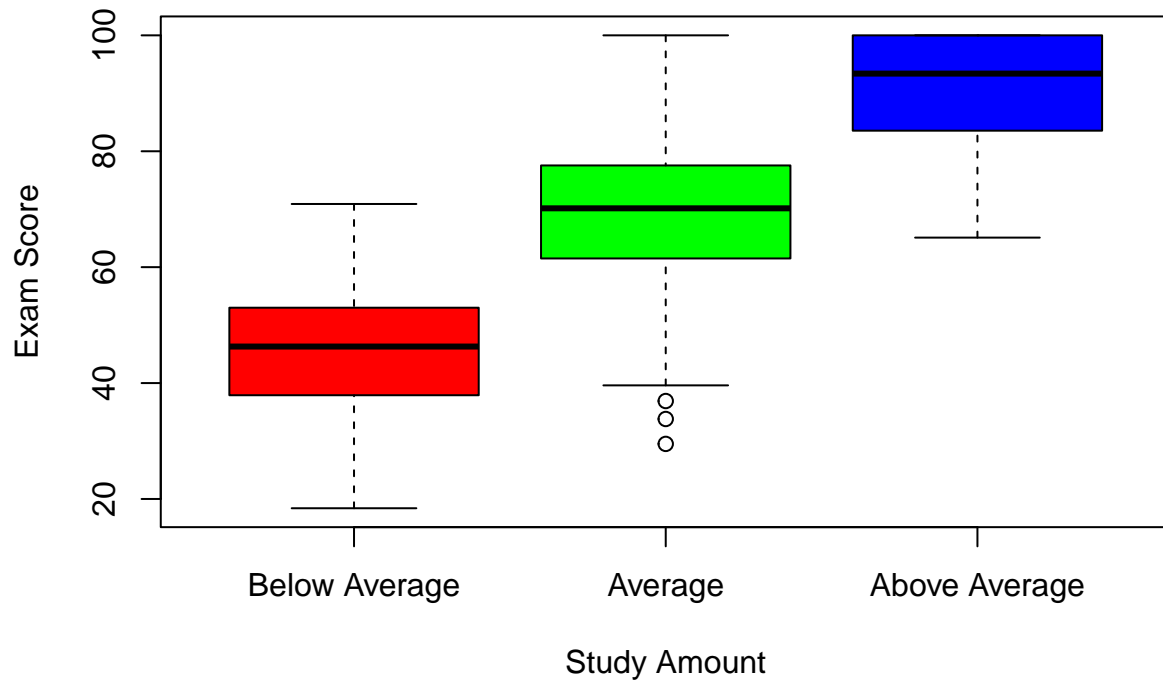
Probability Densities of Exam Scores Based on Study Amount
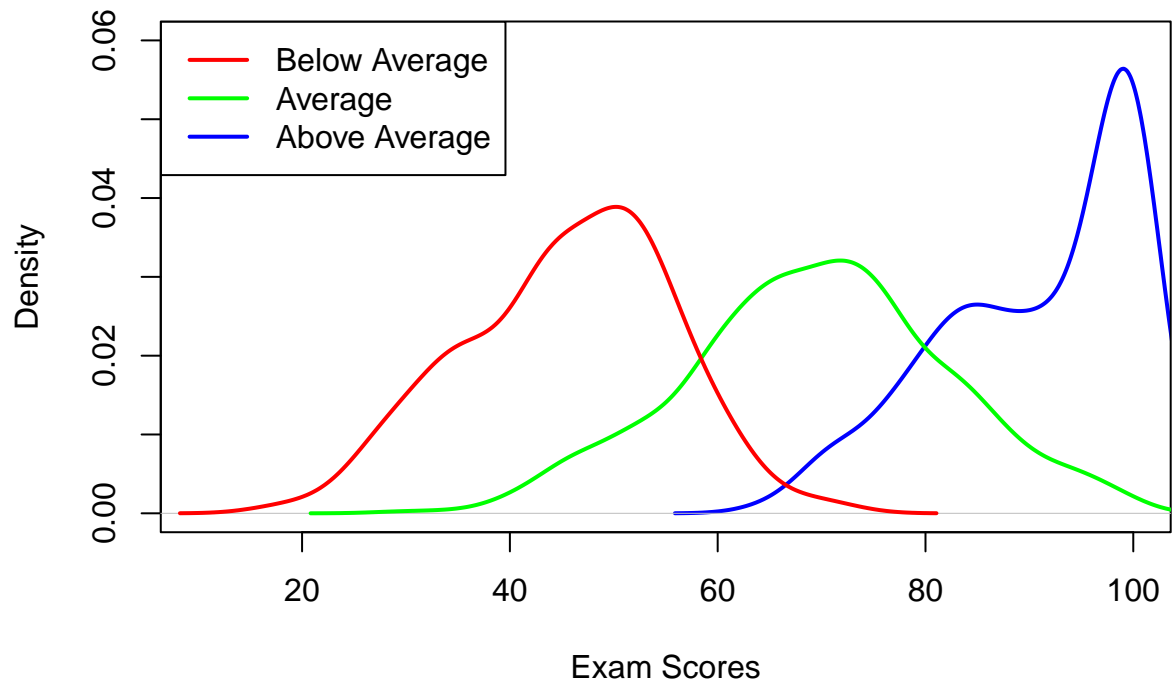
Sleep Hours and Exam Scores

Sleep is an essential component of student wellness and cognitive performance. This section analyzes how the number of sleep hours students get per day relates to their academic achievement, as measured by exam scores.

We begin with summary statistics and a probability density estimate of daily sleep hours among the 1,000 students in the data set.

Summary Statistics

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.20    5.60    6.50    6.47    7.30   10.00
```
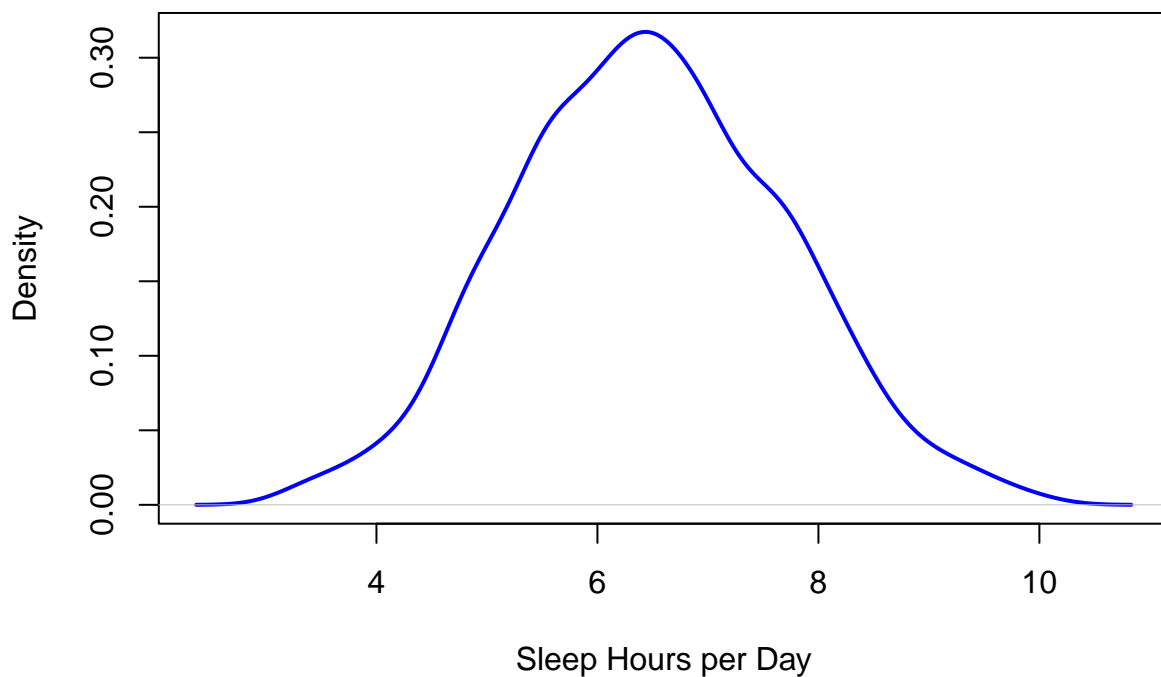
Standard Deviation

```
## [1] 1.226377
```

Variance

```
## [1] 1.504
```

Probability Density for Sleep Hours per Day of Students

From the summary statistics and density plot, we observe that students sleep an average of approximately 6.5 hours per day, with most values falling between 5 and 8 hours.

To assess how different sleep patterns relate to academic outcomes, sleep hours were categorized into three groups:

- **Poor:** less than 5 hours/day
- **Fair:** between 5 and 8 hours/day

- **Great:** more than 8 hours/day

```
## [1] 8.0 4.6 8.0 9.2 4.9 7.4
```

```
## [1] Fair  Poor  Fair  Great Poor  Fair
## Levels: Poor Fair Great
```

Summary statistics were then computed to examine differences in exam performance across the three sleep amount groups.

Mean

```
##     Poor     Fair    Great
## 65.00826 69.99810 71.42816
```

Median

```
##  Poor  Fair Great
## 66.40 70.65 72.60
```

Standard Deviation

```
##     Poor     Fair    Great
## 16.75181 16.66217 18.08765
```

Variance

```
##     Poor     Fair    Great
## 280.6230 277.6279 327.1630
```

We observe some variation in exam score statistics across sleep quality groups. Both mean and median exam scores are highest among students with greater sleep, though the differences are not as pronounced as those observed in study hour comparisons. Variances remain sufficiently similar across groups to allow for ANOVA testing.

ANOVA TEST RESULTS

To assess the statistical significance of sleep quality on exam performance, an ANOVA test was conducted.

```
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## sleep_quality   2   2767    1384   4.889 0.00771 **
## Residuals     997 282171     283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value is 0.00771, adequately below the conventional threshold of 0.05. This indicates that we can reject the null hypothesis of equal group means. Thus, sleep duration has a statistically significant effect on exam performance. While the effect size appears smaller than that of study hours, it is still a meaningful factor to consider.

Below, we visualize the relationship using box plots and kernel density plots:

Boxplots of Effect of Sleep Quality on Exam Scores for Students
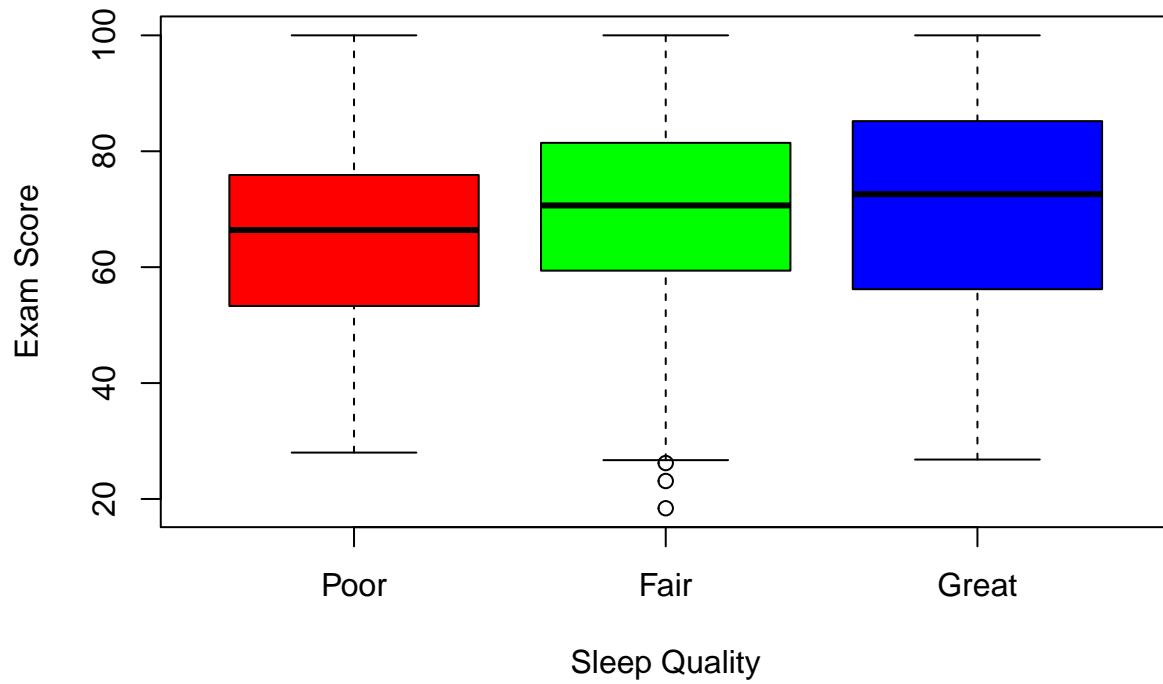
Probability Densities of Exam Scores Based on Sleep Hours
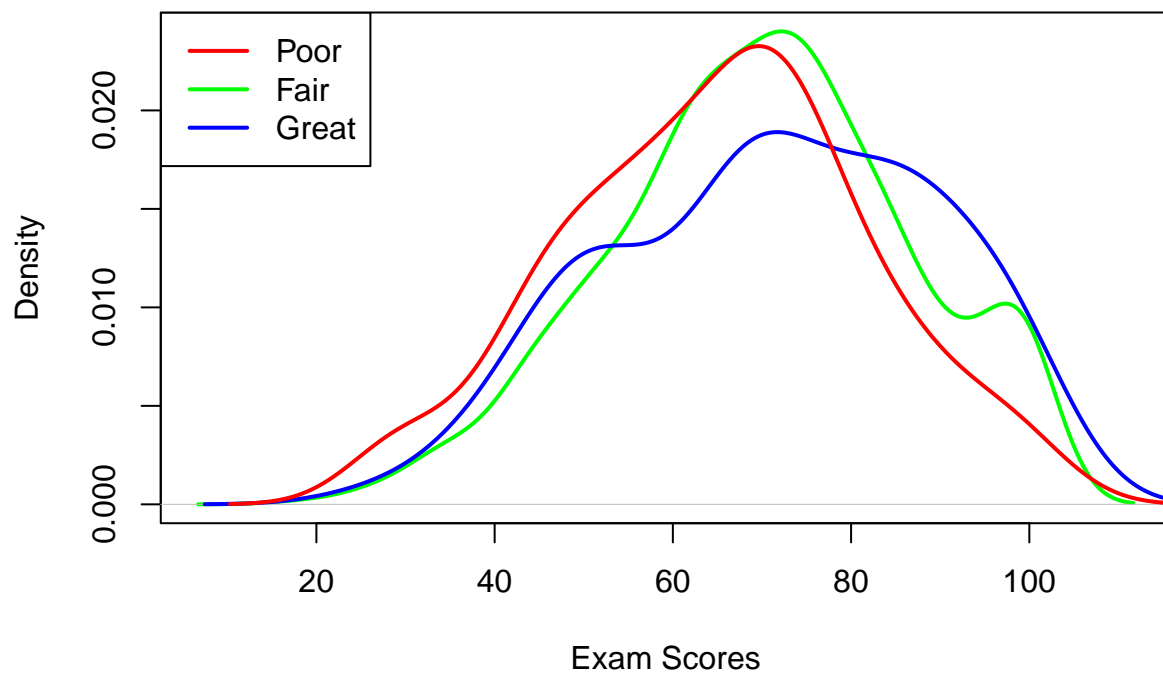
Social Media Hours and Exam Scores

In this section, I examine the relationship between students' daily social media usage and their academic performance. The goal is to determine whether time on social media is associated with any significant changes in exam scores.

Summary statistics and a probability density estimate of daily social media hours among the 1,000 students in the data set are created.

Summary Statistics

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.700   2.500   2.506   3.300   7.200
```
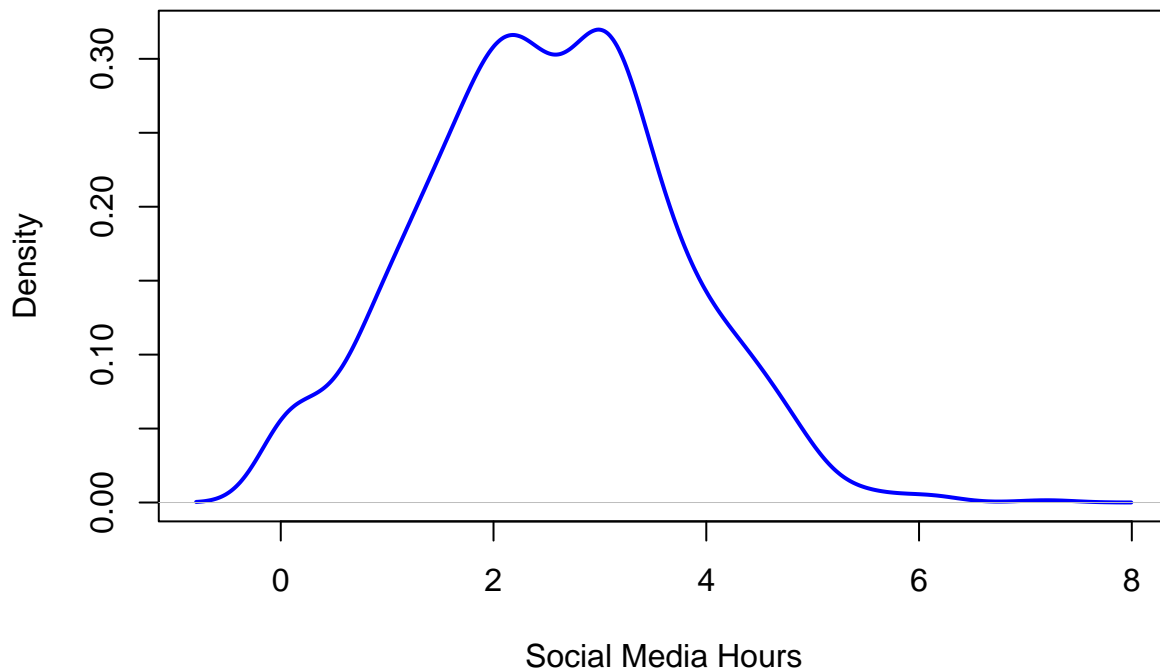
Standard Deviation

```
## [1] 1.172422
```

Variance

```
## [1] 1.374574
```

Probability Density of Student Social Media Hours

The summary statistics and density plot show that the average student spends around 2.5 hours per day on social media. Most values lie between 1.5 and 3.5 hours per day.

To better analyze the impact, I categorized social media hours into three groups:

- **Low:** Fewer than 1.5 hours per day
- **Moderate:** Between 1.5 and 3.5 hours per day
- **High:** Greater than 3.5 hours per day

```
## [1] 1.2 2.8 3.1 3.9 4.4 1.3
```

```
## [1] Low      Moderate Moderate High     High     Low
## Levels: Low Moderate High
```

Summary statistics were then computed to examine differences in exam performance across the three social media activity groups.

Mean:

```
##      Low Moderate     High
## 73.03613 69.63349 65.88681
```

Median:

```
##      Low Moderate     High
##     73.5     70.2     67.0
```

Standard Deviation:

```
##      Low Moderate     High
## 15.84658 16.96768 16.98794
```

Variance:

```
##      Low Moderate     High
## 251.1142 287.9020 288.5900
```

There appears to be a slight downward trend in both the mean and median exam scores as social media usage increases. While the variance across groups differs slightly, it is still sufficiently similar to meet the assumption of equal variances required for an ANOVA test.

ANOVA TEST RESULTS

An ANOVA test is conducted to determine whether the differences in mean exam scores across the social media usage groups are statistically significant.

```
##                        Df Sum Sq Mean Sq F value   Pr(>F)
## social_media_activity   2   4765    2383   8.479 0.000223 ***
## Residuals             997 280173     281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value is 0.000223, well below the conventional threshold of 0.05. This indicates that we can reject the null hypothesis of equal group means. Thus, social media usage has a statistically significant effect

on exam performance. Contrary to the two previous predictors, social media usage has a negative correlation with average exam scores rather than a positive correlation.

To visualize this relationship, box plots and kernel density plots are below:

Social Media Activity of Students and Exam Scores
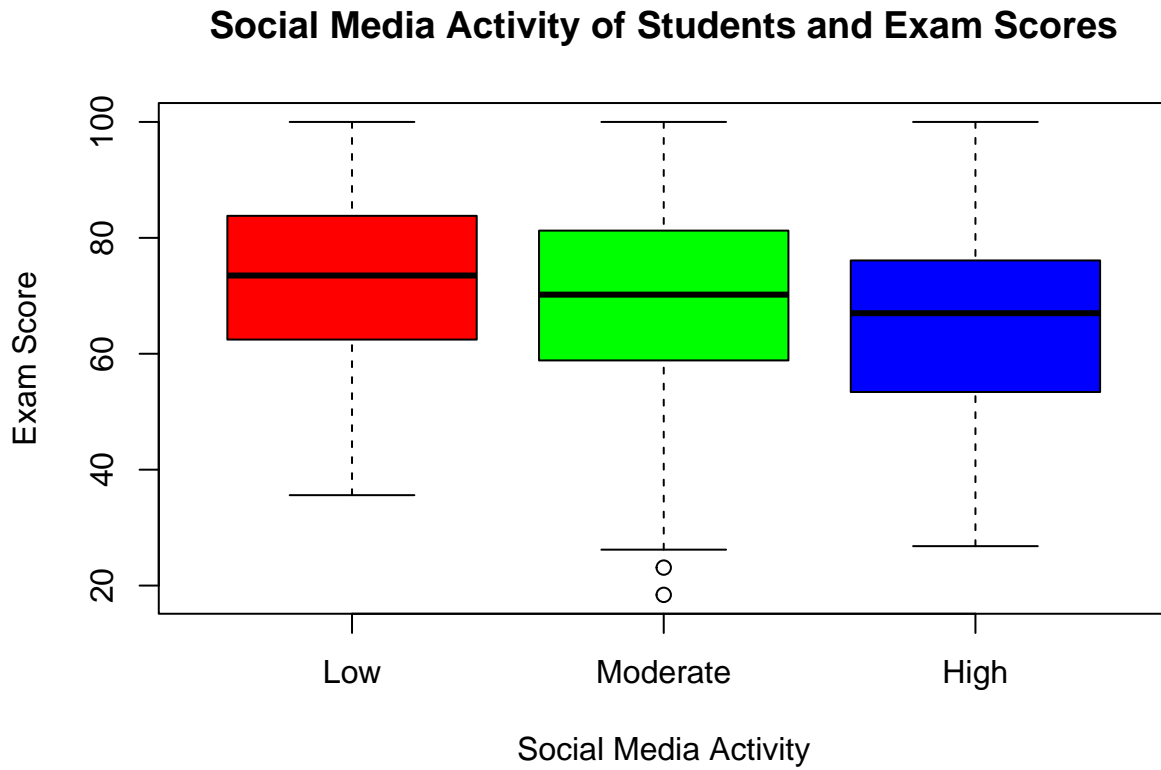
Probability Densities of Exam Scores Based on Social Media Activit
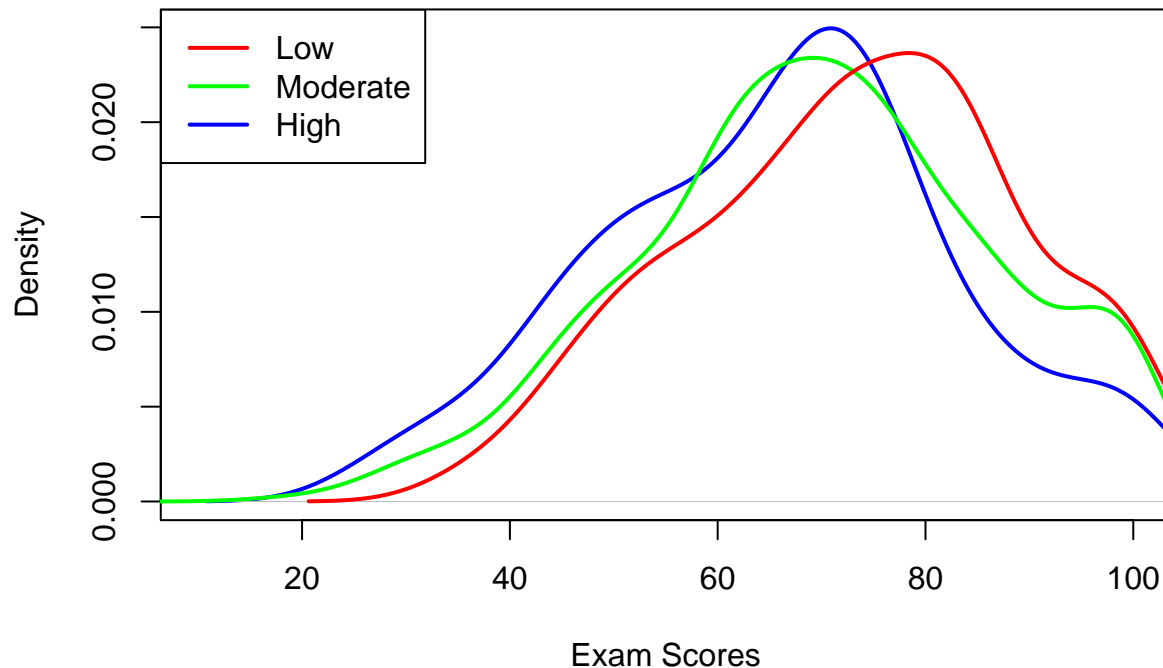
Student Part Time Job Status and Exam Scores

In this analysis, I shift my focus from numerical student traits to categorical traits, beginning with an examination of whether holding a part-time job impacts students' academic performance. Specifically, I aim to determine if there is a significant difference in the average exam scores between students who have part-time jobs and those who do not.

A frequency table displaying the number of students in each part-time job category is shown:

```
##
##  No Yes
## 785 215
```

The table reveals that the number of students without part-time jobs is nearly four times greater than those who are employed part-time, a trend that aligns with typical student commitments.

Summary statistics within each group are computed.

Summary Statistics

```
## $No
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.40   58.60   70.60   69.84   82.10  100.00
##
## $Yes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.20   57.95   69.20   68.74   79.90  100.00
```

Standard Deviation

```
##       No      Yes
## 16.97524 16.57892
```

Variance

```
##       No      Yes
## 288.1588 274.8605
```

These descriptive statistics suggest minimal differences between the two groups' exam scores, aside from slight variation in minimum scores. To further assess whether the observed differences are statistically significant, I perform a two-sample t-test. With the similarity in variances, the assumption of equal variance is reasonable.

TWO-SAMPLE T-TEST RESULTS

```
##
##  Two Sample t-test
##
## data:  exam_score by part_time_job
## t = 0.84089, df = 998, p-value = 0.4006
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -1.458089  3.644698
## sample estimates:
##  mean in group No mean in group Yes
##          69.83656          68.74326
```

The resulting p-value of 0.4006 is well above the conventional significance threshold of 0.05. Therefore, we fail to reject the null hypothesis, indicating no statistically significant difference in mean exam scores between students with and without part-time jobs. In other words, employment status does not appear to influence academic performance.

To visualize this result, I include box plots and kernel density plots below:

Boxplots of Exam Scores Based on Part Time Job Status
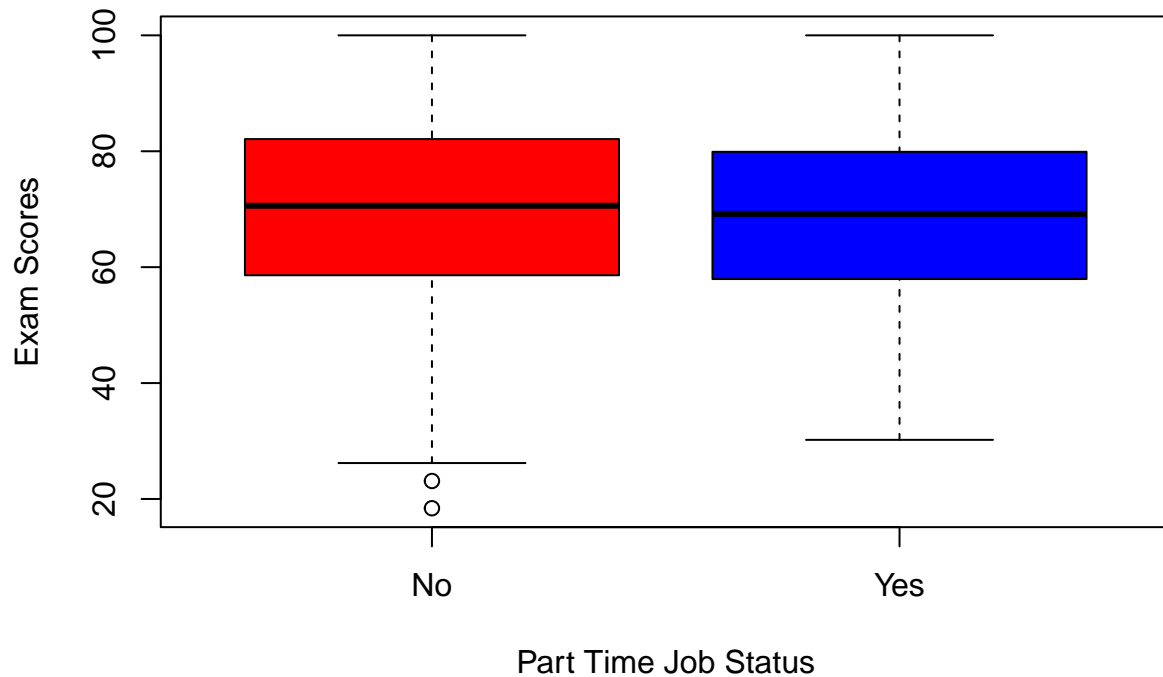
**Probability Densities of Exam Scores Based on Part–Time Job Status**
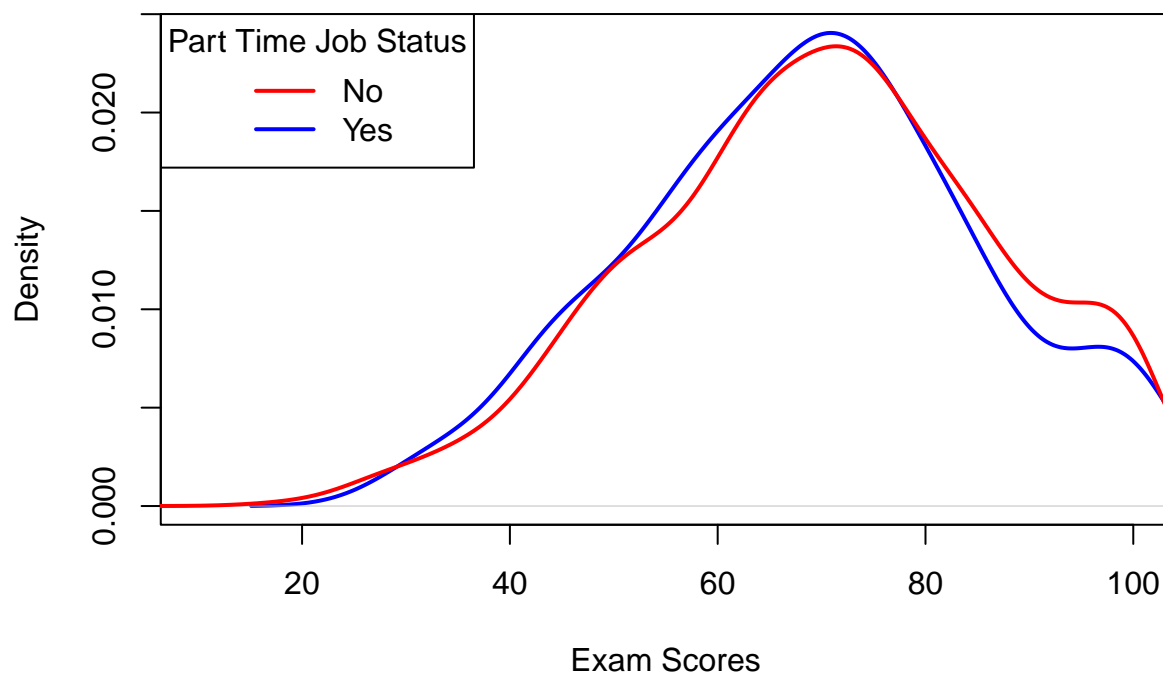
Mental Health Rating and Exam Scores

Mental health is increasingly recognized as a vital component of student well-being and success, making it an important variable to investigate in the context of exam scores. I seek to determine whether mental health

actually does have a statistically significant impact on student exam scores.

The analysis begins by examining the frequency of distinct mental health ratings. A table and bar plot display these frequencies below.

```
##
##   1   2   3   4   5   6   7   8   9  10
## 102  94 105 110  99 108  91 105  87  99
```

Frequency of Mental Health Ratings
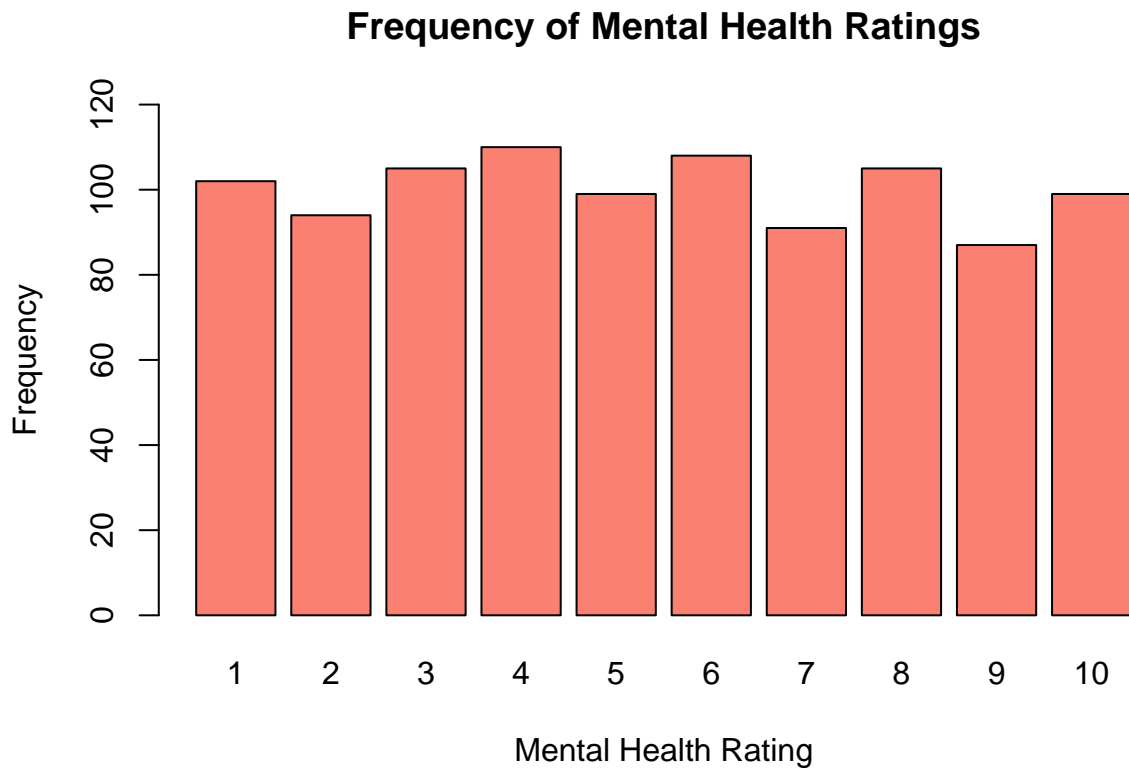
The ratings appear to be approximately uniformly distributed, with there being similar frequencies for each rating. This ensures a balanced comparison across all ratings. For further analysis, mental health ratings were grouped into three categories.

- **Poor:** ratings 1-3
- **Normal:** ratings 4-6
- **Great:** ratings 7-10

```
## [1] 8 8 1 1 1 4
```

```
## [1] Great  Great  Poor   Poor   Poor   Normal
## Levels: Poor Normal Great
```

Summary statistics were then computed to determine differences in exam performance across all three groups.

Mean

```
##     Poor   Normal   Great
## 63.43156 67.86625 75.90314
```

Median

```
##   Poor Normal  Great
##  63.60  69.20  75.45
```

Standard Deviation

```
##     Poor   Normal   Great
## 15.94086 16.85095 15.48795
```

Variance

```
##     Poor   Normal   Great
## 254.1109 283.9544 239.8766
```

Both the mean and median exam scores increase with higher mental health ratings, suggesting a positive relationship between mental well-being and academic performance. While the group variances differ slightly, they are sufficiently similar to meet the equal variance assumption required for ANOVA testing.

ANOVA TEST RESULTS

An ANOVA test was conducted to evaluate whether the observed differences in group means are statistically significant.

```
##                      Df Sum Sq Mean Sq F value Pr(>F)
## studhabts$MHQuality   2  27583   13791   53.43 <2e-16 ***
## Residuals           997 257356     258
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 2×10^{-16}, well below the significance threshold of 0.05. This allows us to confidently reject the null hypothesis that the group means are equal. We conclude that mental health quality has a statistically significant effect on exam scores. In particular, students with higher mental health ratings tend to achieve moderately higher scores on average.

The following box plots and kernel density plots visually reinforce this relationship:

Exam Score Based on Mental Health Rating for Students
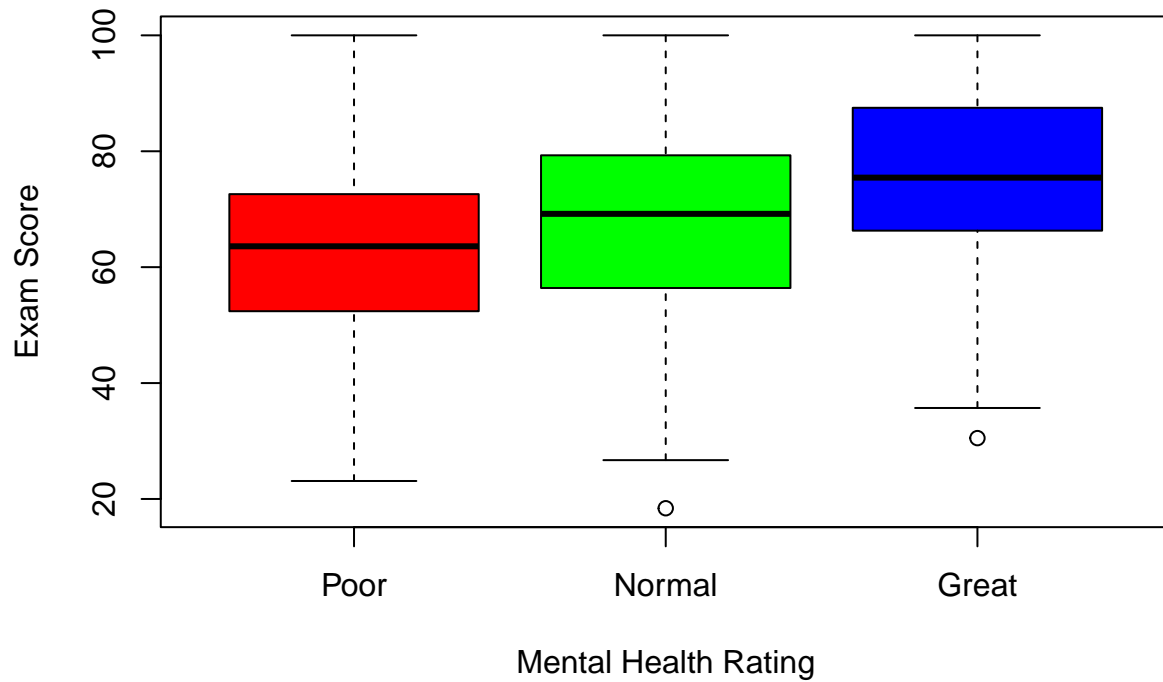
Probability Densities of Exam Scores based on Mental Health Qualit
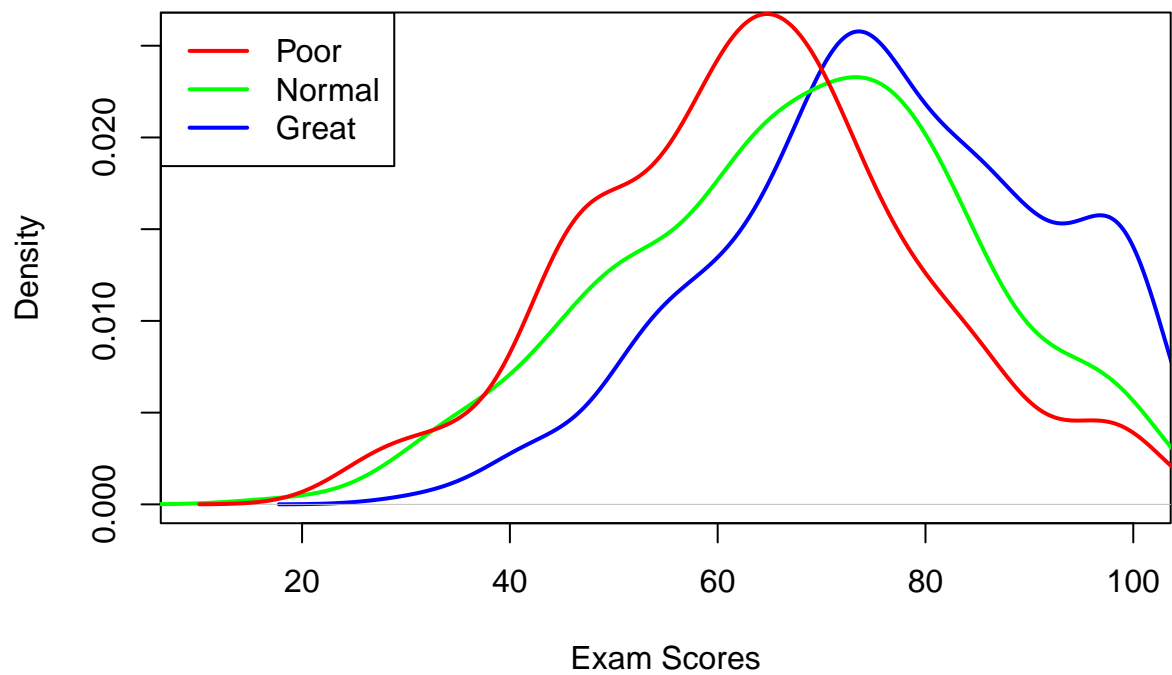
Parental Educational Level and Exam Scores

In this final analysis, I examine how a student's academic performance is influenced by the highest educational attainment of their parents. This factor is often considered a strong predictor of academic outcomes, based on the belief that parental education shapes values, resources, and support for learning.

A frequency table summarizing the distribution of students by their parents' highest education level is constructed.

```
##
##          None High School     Bachelor      Master
##            91          392          350         167
```

The table shows that the majority of students have parents whose highest education level is either a high school diploma or a bachelor's degree, which reflects typical educational attainment trends.

Summary statistics for the exam scores across each parental education category are shown:

Summary Statistics

```
## $None
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.20   59.80   71.00   70.03   81.20  100.00
##
## $`High School`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   23.10   58.27   69.95   69.55   82.30  100.00
##
## $Bachelor
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.40   58.67   71.70   70.27   81.88  100.00
##
## $Master
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   26.80   57.70   66.80   68.09   79.60  100.00
```

Standard Deviation

```
##          None High School     Bachelor      Master
##      16.56018    16.83271     17.28656    16.39533
```

Variance

```
##          None High School     Bachelor      Master
##      274.2396    283.3401     298.8251    268.8067
```

The descriptive statistics reveal some variation in exam score distributions across categories, but the differences in group means and medians are not pronounced. To determine whether these differences are statistically significant, I conduct an ANOVA (Analysis of Variance) test.

ANOVA TEST RESULTS

An ANOVA test is conducted to determine whether the differences in mean exam scores across the parental education level groups are statistically significant.

```
##                        Df Sum Sq Mean Sq F value Pr(>F)
## parental_education_level  3    559   186.3   0.653  0.581
## Residuals               996 284379   285.5
```

With a p-value of 0.581, the test fails to reject the null hypothesis that the group means are equal. This suggests that any differences in exam scores among the parental education groups are likely due to random variation rather than a true underlying effect. Thus, the highest level of parental education does not appear to significantly influence students' average exam scores.

Below are box plots and kernel density plots that provide a visual representation of this analysis:

Exam Scores Based on Parental Education Level of Students
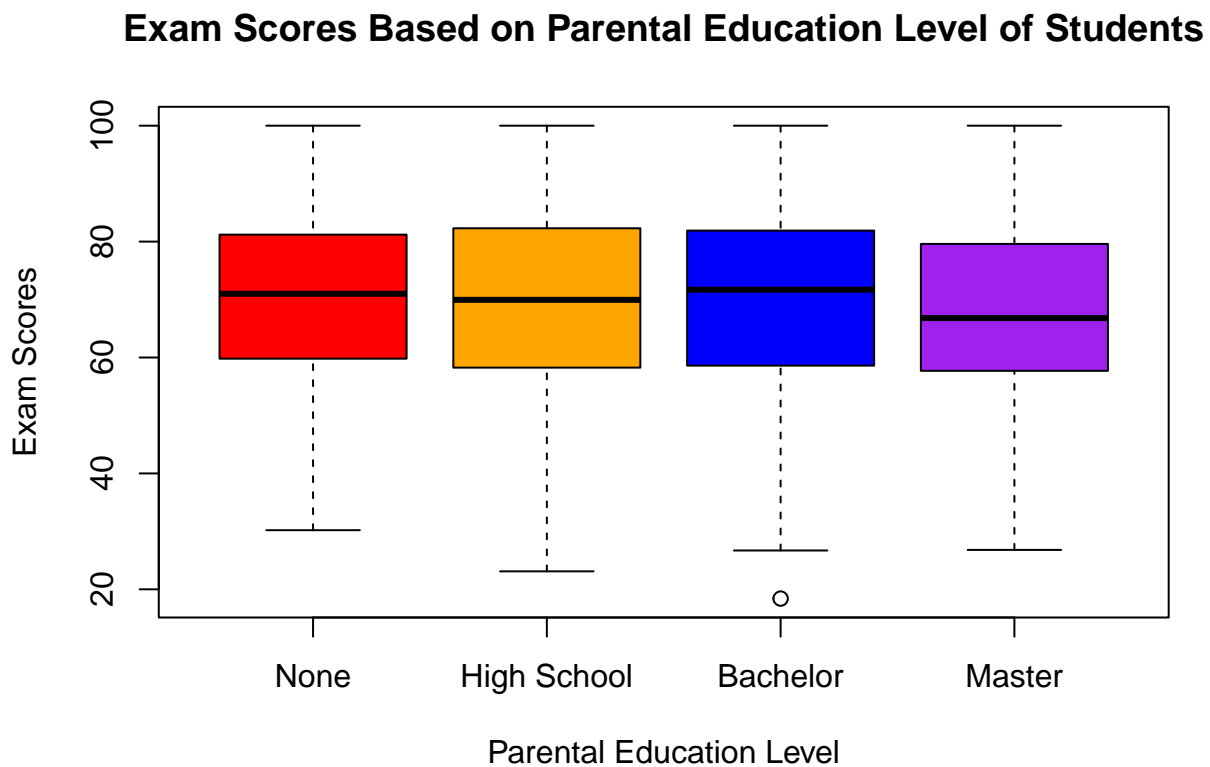
Probability Densities of Exam Scores Based on Parental Education Le
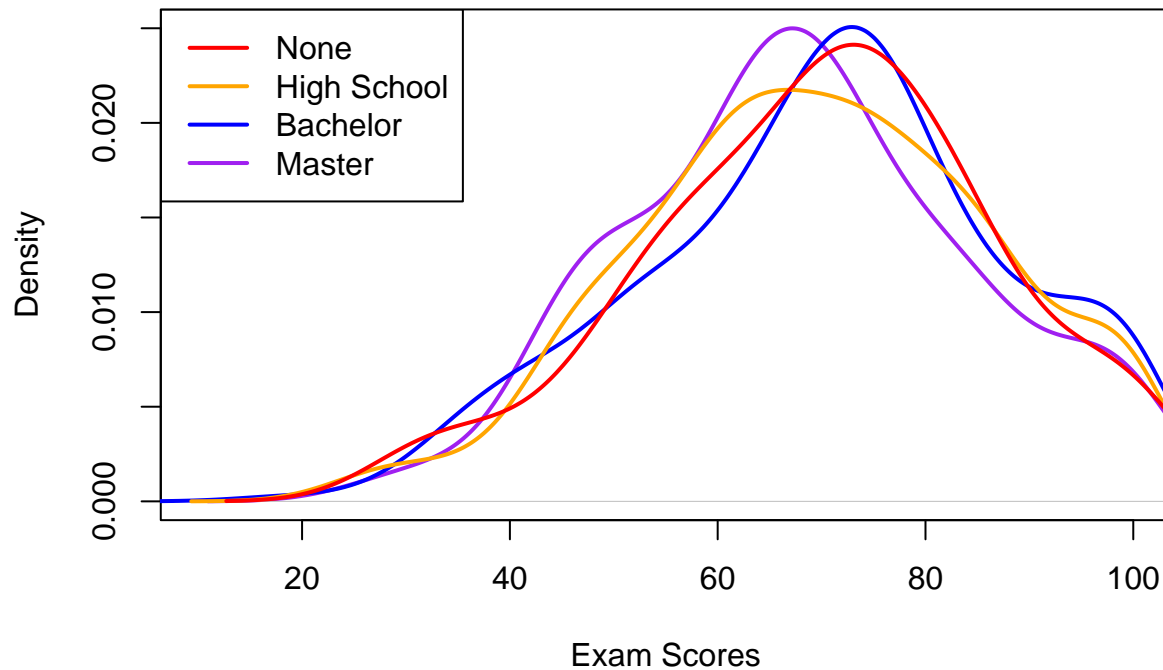
Findings

This study analyzed how various student-related traits influence academic performance, as measured by exam scores. The predictors examined included both numerical and categorical variables: mental health rating, study hours per day, sleep hours per day, social media hours per day, part-time job status, and highest parental education level.

Summary of Key Relationships

The analysis revealed the following patterns of association:

Traits Positively Associated with Exam Scores

- *Study Hours per Day*

  Showed the strongest positive correlation with exam scores. As expected, students who studied more consistently achieved higher average scores.

- *Mental Health Rating*

  Demonstrated a moderate positive correlation. Students reporting better mental health tended to perform better academically, suggesting a meaningful link between well-being and academic success.

- *Sleep Hours per Day*

  Displayed a weak positive correlation. While the effect was less pronounced, students with more sleep generally had higher average exam scores.

Trait Negatively Associated with Exam Scores

- ***Social Media Hours per Day***

  Showed a weak negative correlation with exam scores. Students who spent more time on social media tended to have slightly lower average scores, though the relationship was not strong.

Traits Not Significantly Associated with Exam Scores

- ***Part-time Job Status***

  The difference in average exam scores between students with and without part-time jobs was not statistically significant. This suggests that holding a part-time job does not materially impact academic performance.

- ***Parental Education Level***

  Contrary to common assumptions, no significant difference was found in exam scores across different parental education levels. This indicates that a student's academic outcomes are not strongly tied to the educational background of their parents—at least within this data set.