Exploring Determinants of Academic Success: A Multiple Regression Analysis of Factors Influencing Final Grades in Honors Biology 101

Ariel Rosario, Jr.

2024-03-10

Summary

This project embarks on an exploratory journey to uncover the factors that influence final grades in an Honors Biology 101 class at NotReal University. At its core, the study aims to dissect the complexities of academic performance, focusing on how certain measurable variables may interplay to shape student outcomes. Through the lens of statistical analysis, we investigate the impact of study habits, attendance, demographic factors, and personal discipline on the educational achievements within a challenging and rigorous academic environment.

To navigate through this academic inquiry, a series of variables have been meticulously selected for analysis. 'Study_Hours_Per_Week' quantifies the time students allocate to study outside of class—a reflection of their dedication and learning efforts. 'Classes_Missed' offers a window into the attendance patterns, positing the basic premise that presence in class is a fundamental component of academic success. 'Age' serves as a proxy for maturity and experience, providing perspective on whether these attributes correlate with academic performance. Lastly, 'Gender' is considered to understand if and how it plays a role in the educational attainment within this specific academic context.

As we distill the essence of these variables into actionable insights, two key findings emerge:

- Students who dedicated more hours to studying each week were found to have significantly higher final grades, underscoring the positive impact of study time on academic performance.
- A strong negative correlation was observed between the number of classes missed and final grades, indicating that regular attendance is crucial for success in Honors Biology 101.

These findings serve as pivotal points for educational strategies, implicating that both instructors and students alike might benefit from policies and practices that promote consistent study routines and class attendance.

Introduction

The quest to understand what drives academic success in the sciences leads us to a comprehensive analysis within the context of an Honors Biology 101 class. Here, we delve into the multifaceted nature of educational achievement, unraveling the myriad factors that could potentially influence a student's final grade. Armed with data and guided by inquiry, our study is designed to dissect the nuances of student performance and to examine the interrelationships between study patterns, classroom attendance, and various demographic indicators.

Our approach combines empirical scrutiny with statistical methods to assess how time invested in studying, frequency of class attendance, age, and gender correlate with academic outcomes. The analysis is tailored to identify the extent to which these selected variables predict final grades, providing insights that are expected to be instrumental in the development of strategies to enhance student performance. By illuminating the dynamics of these key factors, the project aims to contribute valuable information to educators, students, and policy makers in the realm of science education.

```r
library(tidyverse)
library(GGally)
library(lmtest)
library(readxl)
library(ggplot2)

Regression_Data <- read_excel("~/Desktop/Regression_Data.xlsx")

head(Regression_Data)
```

```
## # A tibble: 6 x 5
##   Final_Grade Study_Hours_Per_Week   Age Classes_Missed Gender
##         <dbl>                <dbl> <dbl>          <dbl>  <dbl>
## 1       100                     13    18              0      1
## 2        81                     13    22              3      1
## 3        89.7                   12    21              1      0
## 4       100                     12    21              0      1
## 5       100                     14    19              0      1
## 6        60.3                   11    18              5      1
```

```r
str(Regression_Data)
```

```
## tibble [50 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Final_Grade         : num [1:50] 100 81 89.7 100 100 ...
##  $ Study_Hours_Per_Week: num [1:50] 13 13 12 12 14 11 14 12 11 12 ...
##  $ Age                 : num [1:50] 18 22 21 21 19 18 18 19 20 18 ...
##  $ Classes_Missed      : num [1:50] 0 3 1 0 0 5 2 3 4 2 ...
##  $ Gender              : num [1:50] 1 1 0 1 1 1 0 0 1 1 ...
```

This tibble is showing us key pieces of information about 50 students across 5 different columns. We've got everything from their final grades, which have a pretty big range, to the hours they spend hitting the books each week. There's also a column that counts how many classes students miss, where it looks like the more classes skipped, the lower the grades tend to be. Then there's age, which doesn't seem to tell us much in this case, and a column for gender, noted simply as 0 or 1.

Data Manuplation

```r
Regression_Data$Gender <- ifelse(Regression_Data$Gender == 0, "Male", "Female")

Regression_Data$Gender <- as.factor(Regression_Data$Gender)
```

In the dataset, the 'Gender' variable initially used numeric coding, where 0 represented 'Male' and 1 represented 'Female.' To make the data more intuitive and reader-friendly, a transformation was performed: the numeric codes were converted into the actual character strings 'Male' and 'Female.' Following this, to facilitate easier analysis in R, these character strings were then converted into a factor, which is R's data structure for categorical variables. This conversion is particularly useful for statistical modeling and data visualization, as it ensures that R treats the Gender variable appropriately according to its categorical nature rather than as numeric data.

Univariate Analysis

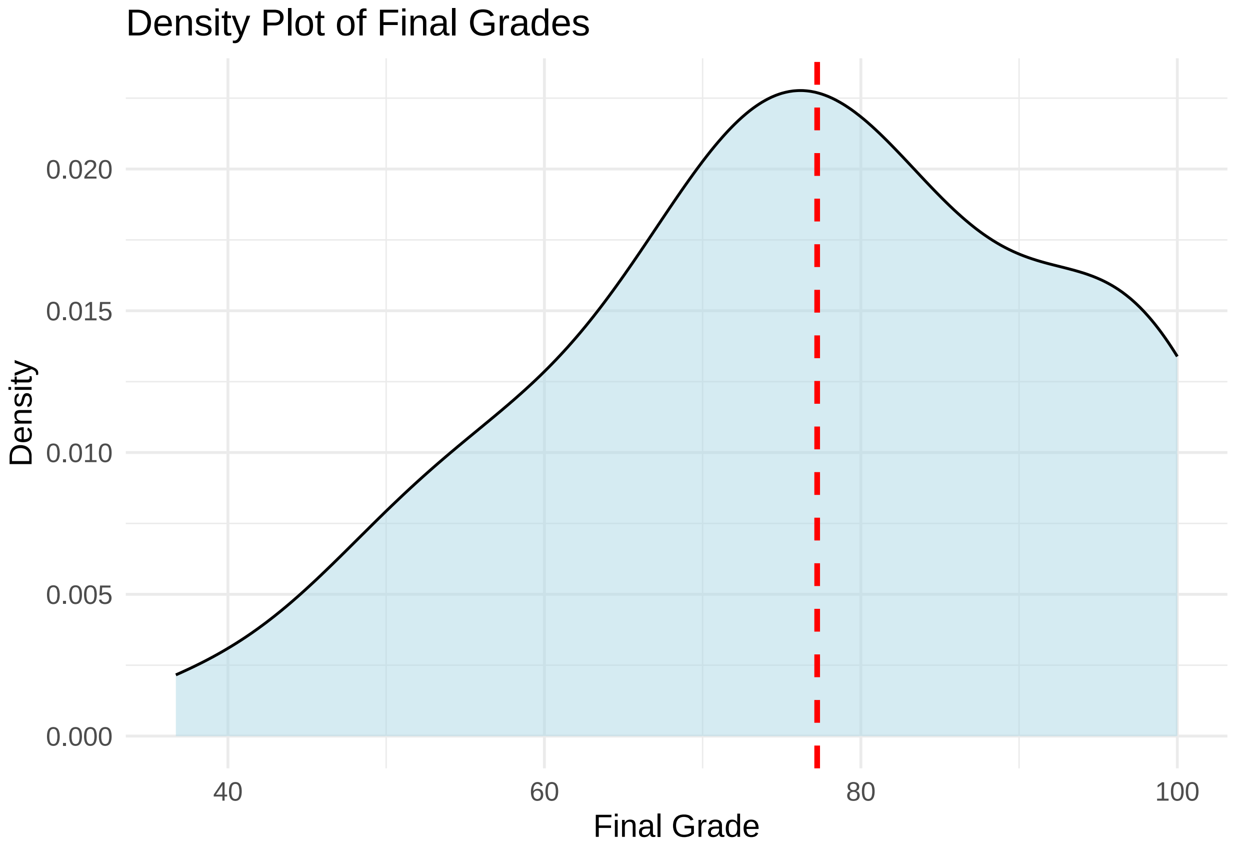
```
summary(Regression_Data)
```

```
##   Final_Grade     Study_Hours_Per_Week      Age        Classes_Missed
##  Min.   : 36.71   Min.   : 6.00        Min.   :18.00   Min.   :0.00
##  1st Qu.: 67.63   1st Qu.:11.00        1st Qu.:19.00   1st Qu.:1.00
##  Median : 77.24   Median :12.00        Median :20.00   Median :3.00
##  Mean   : 76.58   Mean   :12.10        Mean   :20.08   Mean   :3.14
##  3rd Qu.: 88.93   3rd Qu.:13.75        3rd Qu.:21.00   3rd Qu.:5.00
##  Max.   :100.00   Max.   :17.00        Max.   :22.00   Max.   :7.00
##     Gender
##  Female:27
##  Male  :23
##
##
##
##
```

In the dataset examining factors influencing final grades in an honors biology class, the Final Grade variable ranged from a minimum of 36.71 to a perfect maximum of 100, reflecting a wide spectrum of student performance. Study Hours Per Week varied from a low of 6 hours to a high of 17 hours, suggesting diverse study habits among the students. The Age of students in the class had a relatively narrow range, with the youngest being 18 years old and the oldest 22, indicating a sample of traditional college-age students. Classes Missed showed students with perfect attendance at one end of the spectrum (minimum of 0 classes missed) and a maximum of 7 classes missed, hinting at the challenges some students face in maintaining consistent class attendance.

```
median_final_grade <- median(Regression_Data$Final_Grade)

ggplot(Regression_Data, aes(x = Final_Grade)) +
  geom_density(fill = "lightblue", alpha = 0.5) +
  geom_vline(xintercept = median_final_grade, color = "red", linetype = "dashed", size = 1) +
  labs(title = "Density Plot of Final Grades",
       x = "Final Grade",
       y = "Density") +
  theme_minimal() +
  theme(text = element_text(size = 12))
```

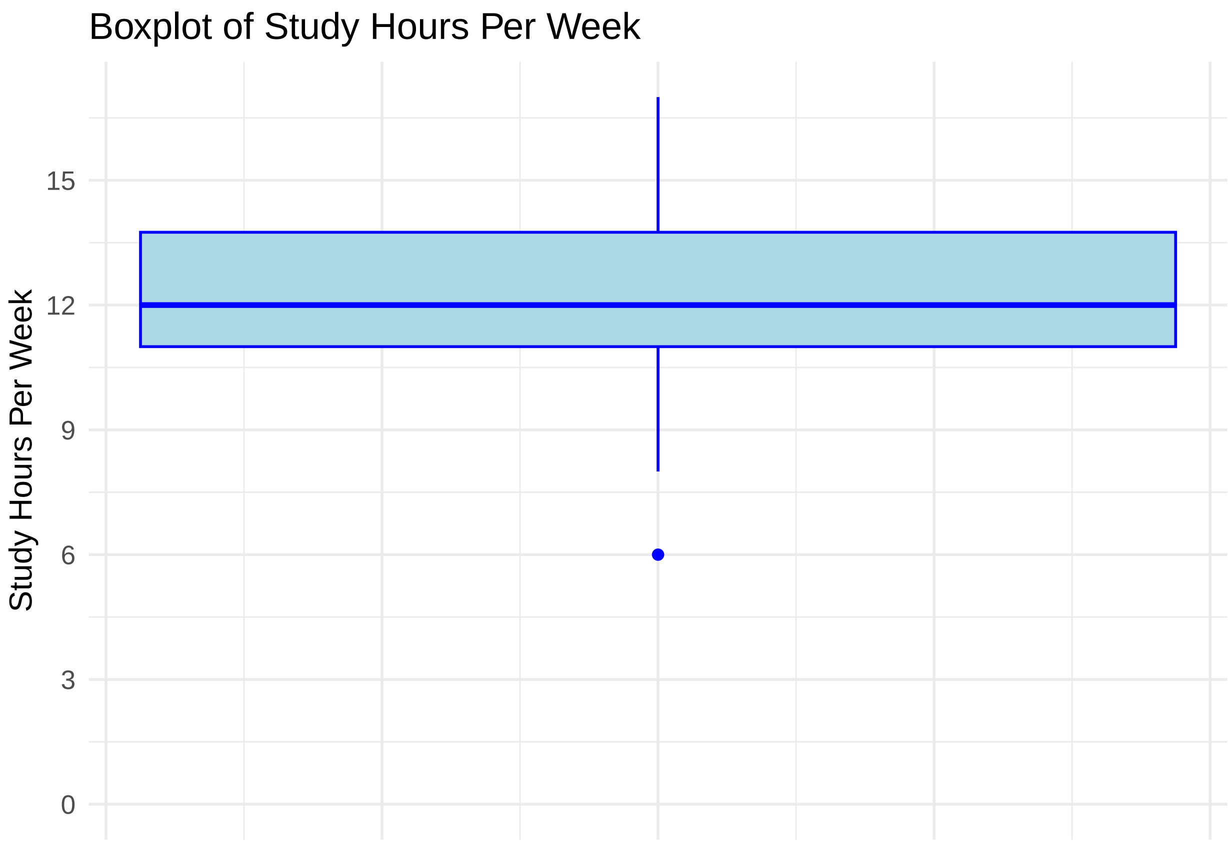


The summary statistics for Final_Grade in the dataset reveal a considerable range in student performance in an honors biology class. The lowest final grade recorded is 36.71, which suggests that at least one student faced significant challenges in the course. On the other end of the spectrum, the highest grade achieved is a perfect 100, indicating that the top-performing student fully met or exceeded the course expectations. The median grade is 77.24, which is slightly higher than the mean of 76.58, pointing towards a distribution that is not heavily skewed in either direction but might have a slight left skew given the median is greater than the mean.

The first quartile, or the 25th percentile, is 67.63, which indicates that a quarter of the students scored below this mark. Conversely, the third quartile is 88.93, which means that 75% of the students scored below this and 25% scored above it. This relatively high third quartile suggests that the top 25% of students performed quite well. The interquartile range, which is the range between the first and third quartiles, indicates the middle 50% of scores and reflects a spread of just over 21 points (from approximately 68 to 89), showing a diverse level of achievement within the class.

A valid assumption based on this data might be that the course has a comprehensive grading system that allows for a wide demonstration of skill levels, as evidenced by the full range of grades from the 30s to a perfect score. The data suggests that while there are students who excel, there are also those who may require additional support to reach their full potential, as indicated by the lower end of the grade spectrum. This spread could also reflect varying degrees of background knowledge, study habits, and possibly the efficacy of the instructional strategies employed in the course.

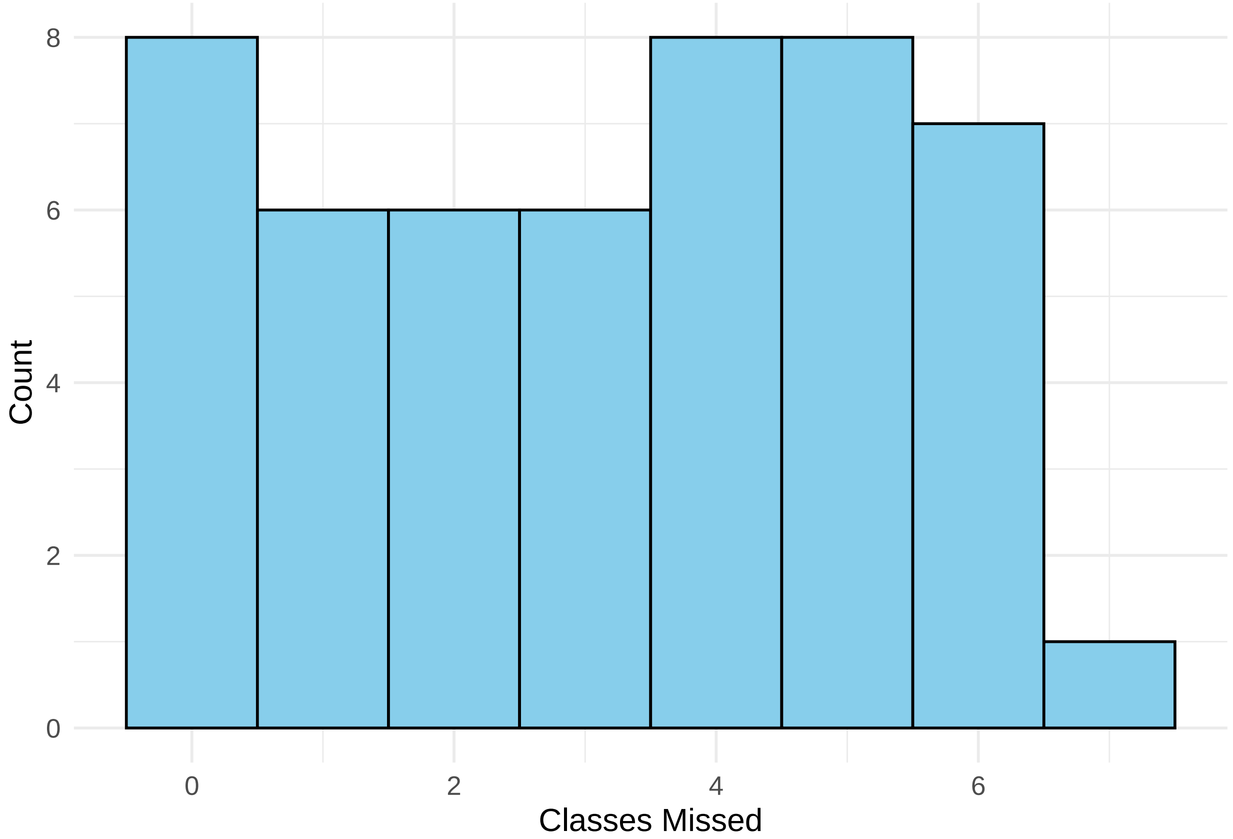
```
ggplot(Regression_Data, aes(y = Study_Hours_Per_Week)) +
  geom_boxplot(fill = "lightblue", color = "blue") +
  labs(title = "Boxplot of Study Hours Per Week",
       x = "",
       y = "Study Hours Per Week") +
  theme_minimal() +
  theme(text = element_text(size = 12),
        axis.title.x = element_blank(),
        axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  scale_y_continuous(breaks = seq(0, max(Regression_Data$Study_Hours_Per_Week, na.rm = TRUE), by = 3),
                     limits = c(0, max(Regression_Data$Study_Hours_Per_Week, na.rm = TRUE)))
```



The statistics for Study__Hours__Per__Week show that students in the honors biology class invested between a minimum of 6 and a maximum of 17 hours per week in their studies. The median value of 12 hours is very close to the mean of 12.10 hours, suggesting a fairly symmetrical distribution of study hours among students. The first quartile at 11 hours indicates that 25% of the students studied less than this amount per week, while the third quartile at 13.75 hours shows that 25% of the students spent more time studying, beyond the middle 50% of the class.

The close proximity of the mean and median, along with a modest range from the minimum to the maximum values, implies that while study habits vary, there is a general consistency in the amount of time students devote to their studies weekly. A valid assumption might be that students tend to follow a somewhat standardized routine in terms of weekly study time, possibly due to the structured nature of the class or external factors such as course load and extracurricular activities. The relatively tight spread of hours also suggests that the course demands a significant but manageable time commitment for studying, which most students seem to meet within a consistent range.

```
ggplot(Regression_Data, aes(x = Classes_Missed)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(x = "Classes Missed", y = "Count") +
  theme_minimal() +
  theme(text = element_text(size = 12))
```



The Classes_Missed variable indicates a range of attendance patterns among students in the honors biology class. The minimum number of classes missed is 0, reflecting perfect attendance by some students. At the higher end, the maximum number of classes missed is 7, suggesting that certain students may have faced challenges attending class regularly. The median number of classes missed stands at 3, closely aligned with the mean of 3.14, which points to a balanced distribution without extreme skewness.

The first quartile is at 1, indicating that 25% of students missed fewer than about one class, while the third quartile is at 5, meaning 75% of students missed fewer than five classes throughout the course. The interquartile range, which spans between the first and third quartiles, suggests that half of the class missed between approximately one and five classes. A reasonable assumption might be that attendance is a factor that has been well-maintained by the majority of the students, with only a few outliers missing more classes. This data could also suggest that while most students are able to attend regularly, a smaller segment may be struggling with attendance due to external factors or engagement with the course material.

Bivariate Analysis

```
cor(Regression_Data %>% select(-Gender))
```

```
##                      Final_Grade Study_Hours_Per_Week         Age
## Final_Grade           1.00000000           0.64289731  0.01979769
## Study_Hours_Per_Week  0.64289731           1.00000000  0.03044852
## Age                   0.01979769           0.03044852  1.00000000
## Classes_Missed       -0.83313681          -0.51852487 -0.01110569
##                      Classes_Missed
## Final_Grade             -0.83313681
## Study_Hours_Per_Week    -0.51852487
## Age                     -0.01110569
## Classes_Missed           1.00000000
```

The correlation matrix provides a snapshot of the relationships between the variables included in the study of factors affecting final grades in an honors biology class. The correlation between Final Grade and Study Hours Per Week is positive and moderately strong (r = 0.6428), indicating that as the number of study hours increases, final grades tend to increase as well. This suggests a significant association where students who invest more time in studying are likely to achieve higher grades. On the other hand, there is a very strong negative correlation between Final Grade and Classes_Missed (r = -0.8331), implying that as the number of classes missed increases, final grades tend to decrease significantly. This relationship highlights the critical impact of class attendance on students' academic performance.

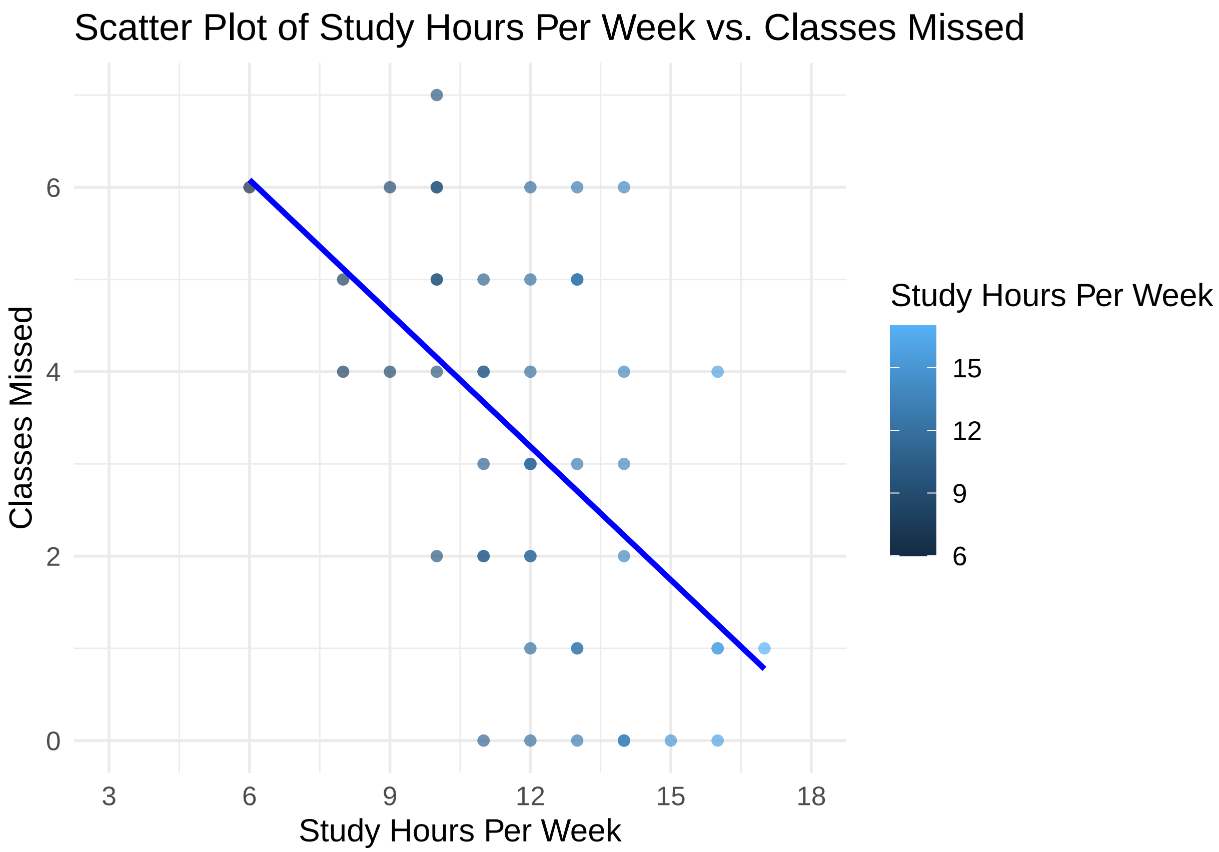
Other correlations in the matrix are less pronounced. Age shows a negligible correlation with Final_Grade (r = 0.0197), suggesting that within the scope of this analysis, the age of the students does not significantly influence their final grades. Similarly, the correlation between Age and Study Hours Per Week (r = 0.0304) is also very low, indicating that the amount of time students spend studying each week does not strongly relate to their age. Additionally, Classes_Missed has a moderate negative correlation with Study Hours Per Week (r = -0.5185), which could be interpreted as students who miss more classes tend to study less. However, the correlation between Age and Classes_Missed is nearly zero (r = -0.0111), indicating no meaningful relationship between the students' age and the number of classes they miss. These findings can inform educators and program designers about the relative importance of study habits and attendance over other factors like age in academic achievement.

```r
breaks <- seq(3, 18, by = 3)

max_hours <- max(Regression_Data$Study_Hours_Per_Week, na.rm = TRUE)

max_hours <- max(max_hours, 18)

ggplot(Regression_Data, aes(x = Study_Hours_Per_Week, y = Classes_Missed, color = Study_Hours_Per_Week)
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Scatter Plot of Study Hours Per Week vs. Classes Missed",
       x = "Study Hours Per Week",
       y = "Classes Missed",
       color = "Study Hours Per Week") +
  theme_minimal() +
  theme(text = element_text(size = 12)) +
  scale_x_continuous(
    breaks = seq(3, 18, by = 3),
    limits = c(3, 18)
  ) +
  scale_color_gradient(breaks = breaks, labels = breaks)
```



The scatter plot illustrates the relationship between the number of classes students missed and the hours they dedicated to study each week. There is a visible negative trend, as shown by the descending blue line, suggesting that students who spend more time studying tend to miss fewer classes. This could indicate a strong commitment to their academic work, as higher study hours might correlate with higher engagement and thus fewer absences. The spread of the data points also implies that while the trend is clear, individual student behaviors do vary, and other factors not displayed on this plot may influence both study habits and attendance rates. The range of study hours from the low end to the high end and the instances of class absences highlight the diversity in the students' approach to their academic responsibilities. This visualization could be used to support the argument that encouraging more consistent study habits may lead to better attendance in academic settings.

Multiple Regression Analysis

```
model <- lm(Final_Grade ~ Study_Hours_Per_Week + Age + Classes_Missed + Gender,
            data = Regression_Data)
summary(model)
```

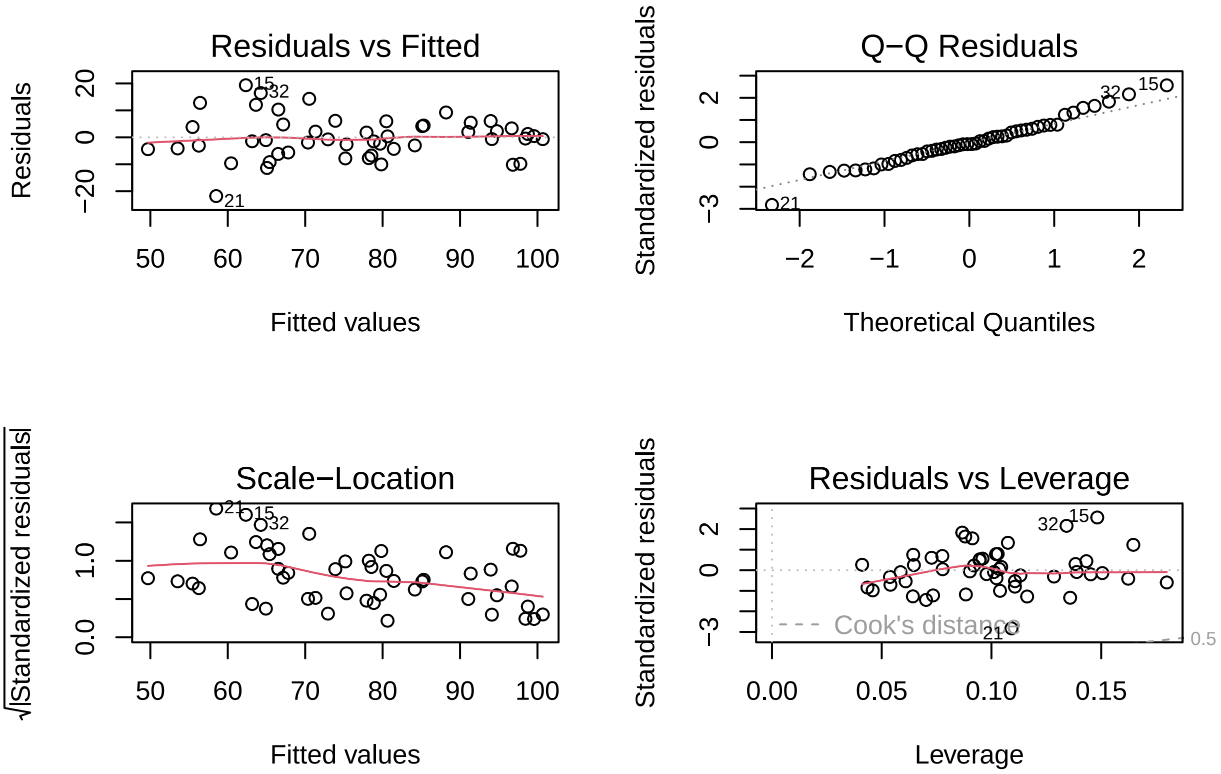
```
##
## Call:
## lm(formula = Final_Grade ~ Study_Hours_Per_Week + Age + Classes_Missed +
##     Gender, data = Regression_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.7931  -4.3734  -0.6718   4.3182  19.3243
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           70.3641    19.2361   3.658 0.000664 ***
## Study_Hours_Per_Week   2.2420     0.6135   3.655 0.000671 ***
## Age                   -0.1576     0.8736  -0.180 0.857648
## Classes_Missed        -5.1356     0.6468  -7.940 4.22e-10 ***
## GenderMale            -3.5238     2.3997  -1.468 0.148937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.169 on 45 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7454
## F-statistic: 36.86 on 4 and 45 DF,  p-value: 1.152e-13
```

In an analysis of factors influencing students' final grades in an honors biology class, a linear regression was conducted with variables including Study Hours Per Week, Age, Classes_Missed, and Gender. The model, which considered these variables to explain variations in Final Grade, was statistically significant. The regression analysis reveals that each additional hour of study per week is associated with an average increase of 2.2420 points in the final grade, holding other factors constant.

The coefficient for Classes Missed is particularly noteworthy, with a substantial and significant negative effect on the final grade: for each class missed, the model predicts a decrease of 5.1356 points, a finding strongly supported by a p-value of 4.22e-10. Meanwhile, Age was found to have a negligible effect on the final grade, with a coefficient of -0.1576, which was not statistically significant (p-value of 0.857648), indicating that age does not play a significant role in academic performance within this group of students. Similarly, the Gender variable (male) suggested a potential average decrease of 3.5238 points for males, but this did not reach statistical significance (p-value of 0.148937), suggesting that gender may not be a determining factor in final grades.

The overall model's fit was robust, with an R-squared value of 0.7662, indicating that approximately 76.62% of the variability in final grades can be explained by the independent variables included in the model. The adjusted R-squared of 0.7454 accounts for the number of predictors in the model and still suggests a strong fit. This model's predictive power is further underscored by the F-statistic of 36.86 with a highly significant p-value of 1.152e-13, confirming that the set of variables collectively has a significant effect on final grade outcomes. These results underscore the importance of attendance and study habits in academic achievement within the context of this honors biology class.

```r
par(mfrow = c(2, 2))
plot(model)
```



The diagnostic plots provide insight into the regression model's validity and the assumption underlying its results. The "Residuals vs Fitted" plot does not display any obvious patterns or systematic structure, suggesting that the linear model is appropriate for the data. Nonetheless, there are several outliers indicated by points far from the zero line, and a few influential observations are denoted by their case numbers, which could potentially affect the regression line's slope. The "Q-Q Plot" shows that residuals are mostly following the theoretical quantiles of a normal distribution, indicating that normality of residuals is largely upheld, though there are slight deviations at both ends of the distribution. The "Scale-Location" plot (also known as a Spread-Location plot) shows that residuals are spread equally along the ranges of predictors, which is a good sign of homoscedasticity.

Finally, the "Residuals vs Leverage" plot helps us to identify influential cases, and it appears that there aren't any points that have an unduly large influence on the model, as no points exceed the Cook's distance threshold, which would be indicative of potential leverage points. These diagnostics are essential for confirming the reliability of the regression analysis, indicating that while the model is generally well-fitted, the influence of outliers and leverage points should be further assessed to confirm these findings.

Conclusion

The conclusion of this statistical investigation into the determinants of final grades in an honors biology class underscores the crucial role of study habits and attendance. The significant regression model has elucidated that an additional hour spent studying per week can lead to an appreciable increase in the final grade, a testament to the value of diligent study practices. While the study also considered age and gender, these variables did not exhibit a substantial impact on the final grades within the demographic of this particular class. The robustness of the model is evidenced by a high R-squared value, which confidently attributes over three-quarters of the grade variability to the variables included in the model.

However, the analysis further reveals the particular importance of class attendance, with missed classes drastically reducing final grades. The substantial negative coefficient associated with Classes_Missed serves as a clear warning about the consequences of absenteeism. This insight is crucial for educational stakeholders who aim to improve student outcomes; it indicates that interventions designed to enhance student attendance could be particularly effective in raising academic performance.

Finally, the diagnostic plots reaffirm the validity of the linear regression model, displaying appropriate levels of homoscedasticity and no significant outliers that would undermine the model's integrity. The "Residuals vs Fitted" and "Q-Q Plot" collectively suggest that the model assumptions are satisfied. Although outliers are present, they do not appear to exert undue influence on the model, as confirmed by the "Residuals vs Leverage" plot. These results not only bolster the credibility of the current analysis but also provide a clear path for further research to explore additional factors that may influence academic success in similar settings.