Project1

Aditi Kumar

2025-10-29

Introduction

NOTE: After recording the video, I updated the report to include more descriptive captions and explanations for each graph to improve clarity.

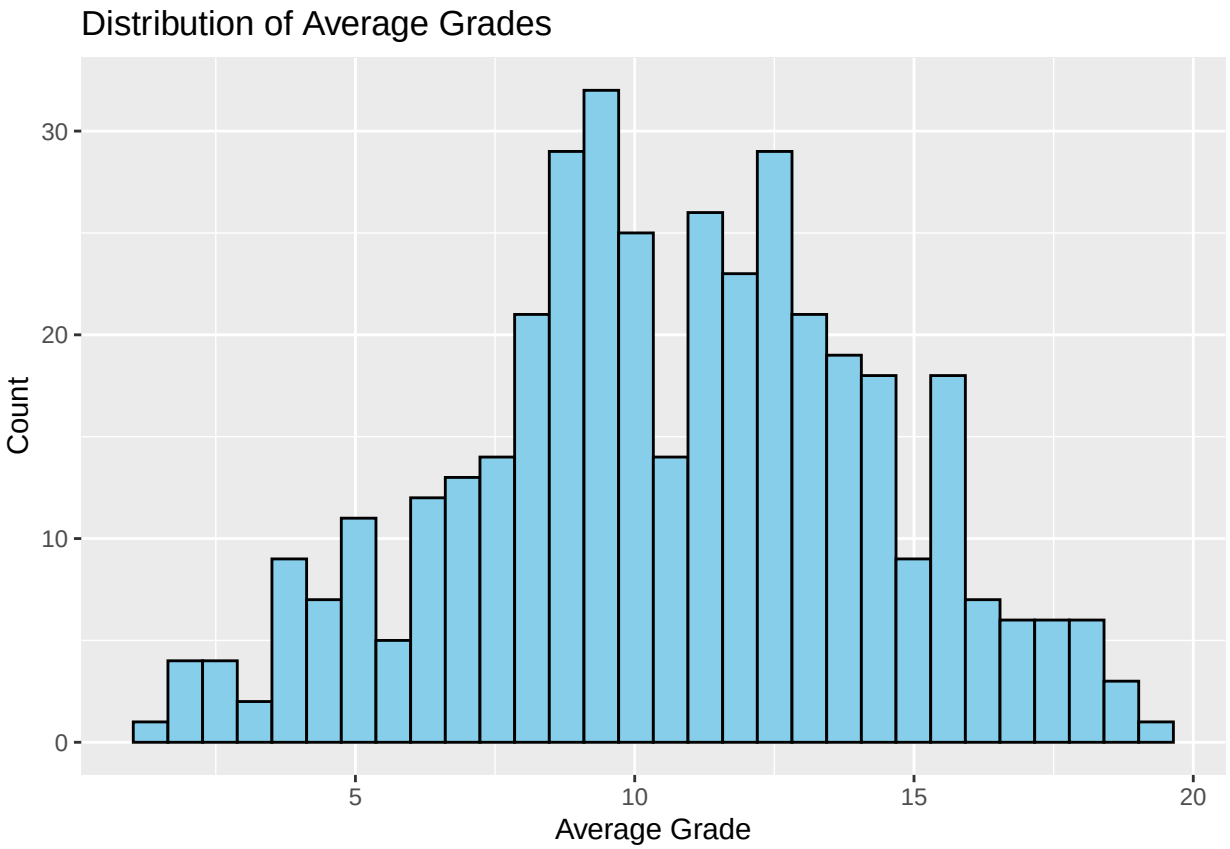
In this project, I analyzed the Student Performance dataset from the UCI Machine Learning Repository. The goal is to explore how different factors such as like study time, health, and family background, affect students' grades.

Dataset Source: UCI Machine Learning Repository, Student Performance dataset (Portuguese secondary school students). Source Link: https://archive.ics.uci.edu/dataset/320/student+performance

1. Grade Distribution

First, I wanted to check what the distribution of average grades are for all students.

```
ggplot(df, aes(x = avg_grade)) +
geom_histogram(bins = 30, fill = "skyblue", color = "black") +
labs(title = "Distribution of Average Grades", x = "Average Grade", y = "Count")
```

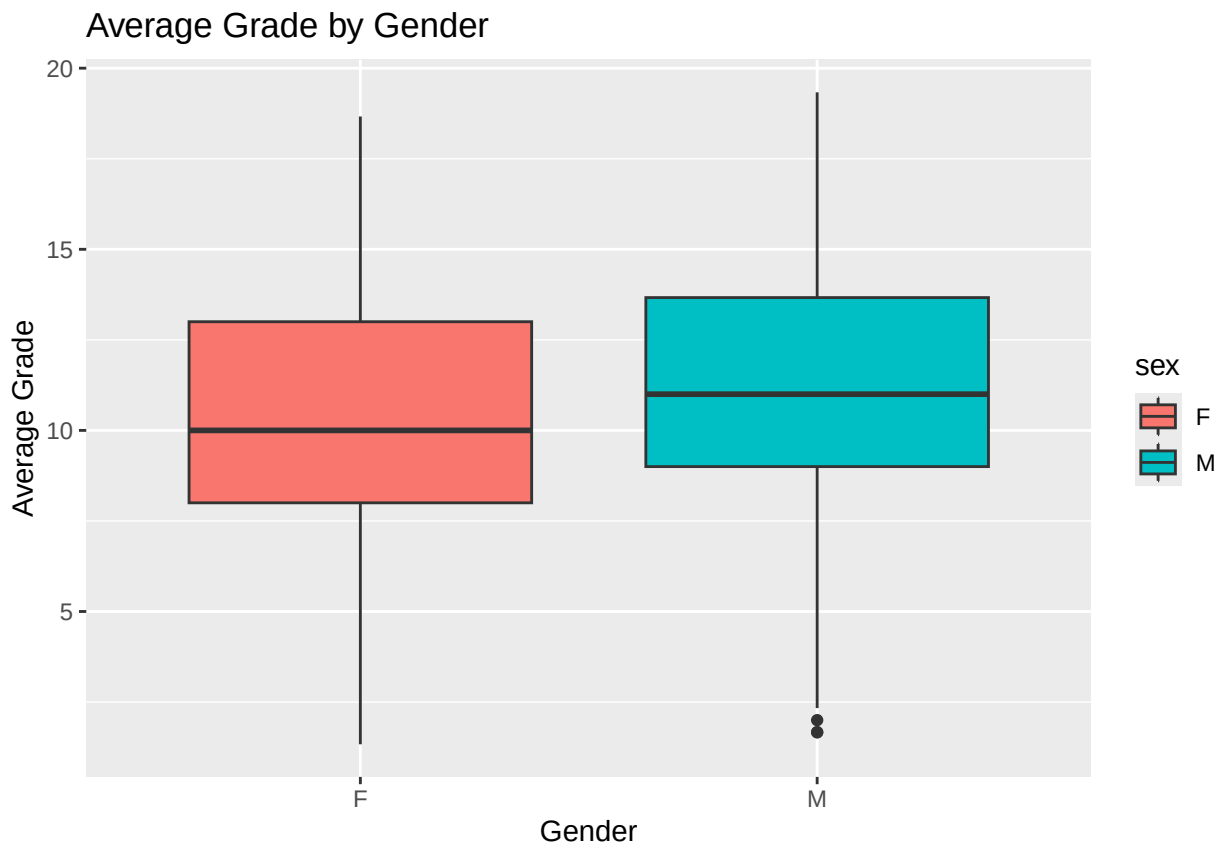


We can see how most students scored around the middle range, with fewer students at the extremes.

2. Average Grades by Gender

Next, I compared the average grade between male and female students.

```
ggplot(df, aes(x = sex, y = avg_grade, fill = sex)) +
geom_boxplot() +
labs(title = "Average Grade by Gender", x = "Gender", y = "Average Grade")
```

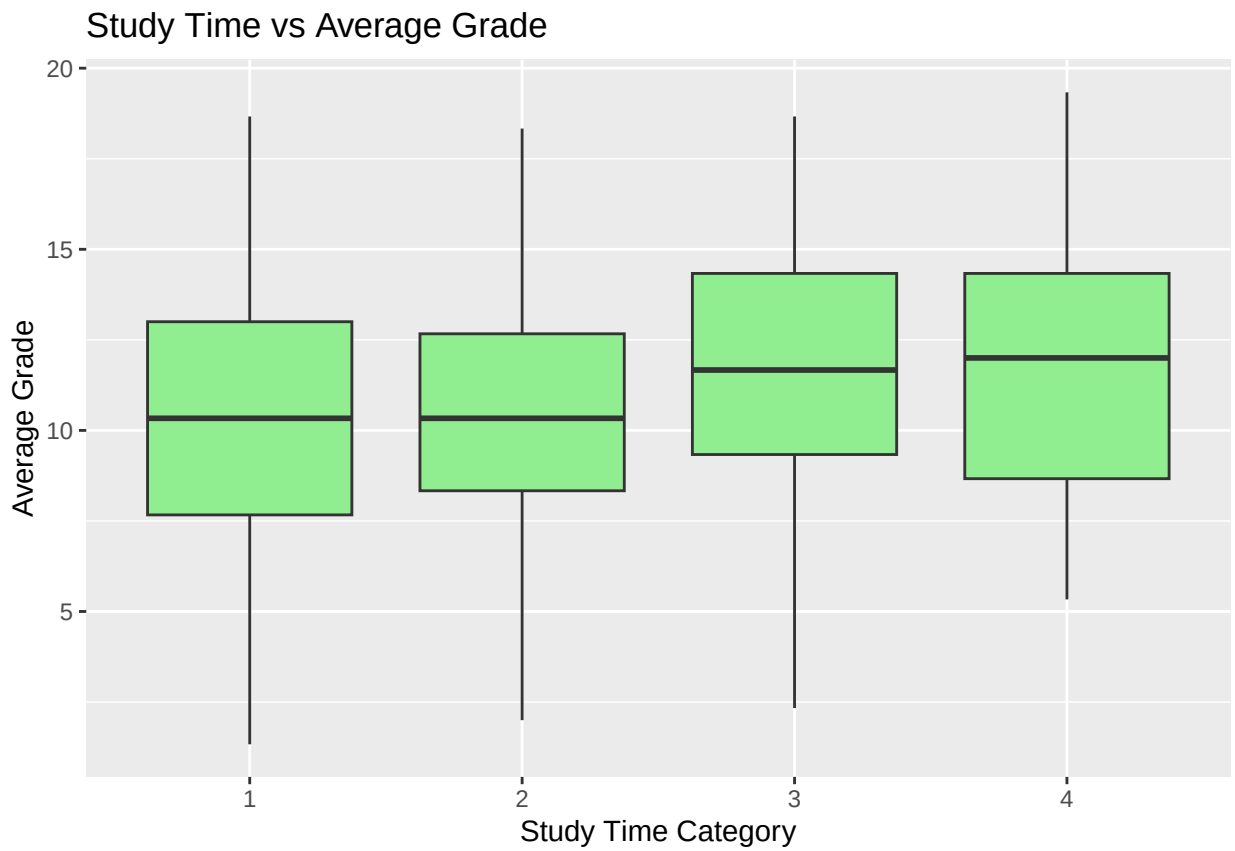


This shows whether gender has a clear impact on grades, but usually, the difference is small.

3. Study Time vs Average Grades

Third, I checked if students who study more tend to have better grades.

```
ggplot(df, aes(x = factor(studytime), y = avg_grade)) +
geom_boxplot(fill = "lightgreen") +
labs(title = "Study Time vs Average Grade", x = "Study Time Category", y = "Average Grade")
```



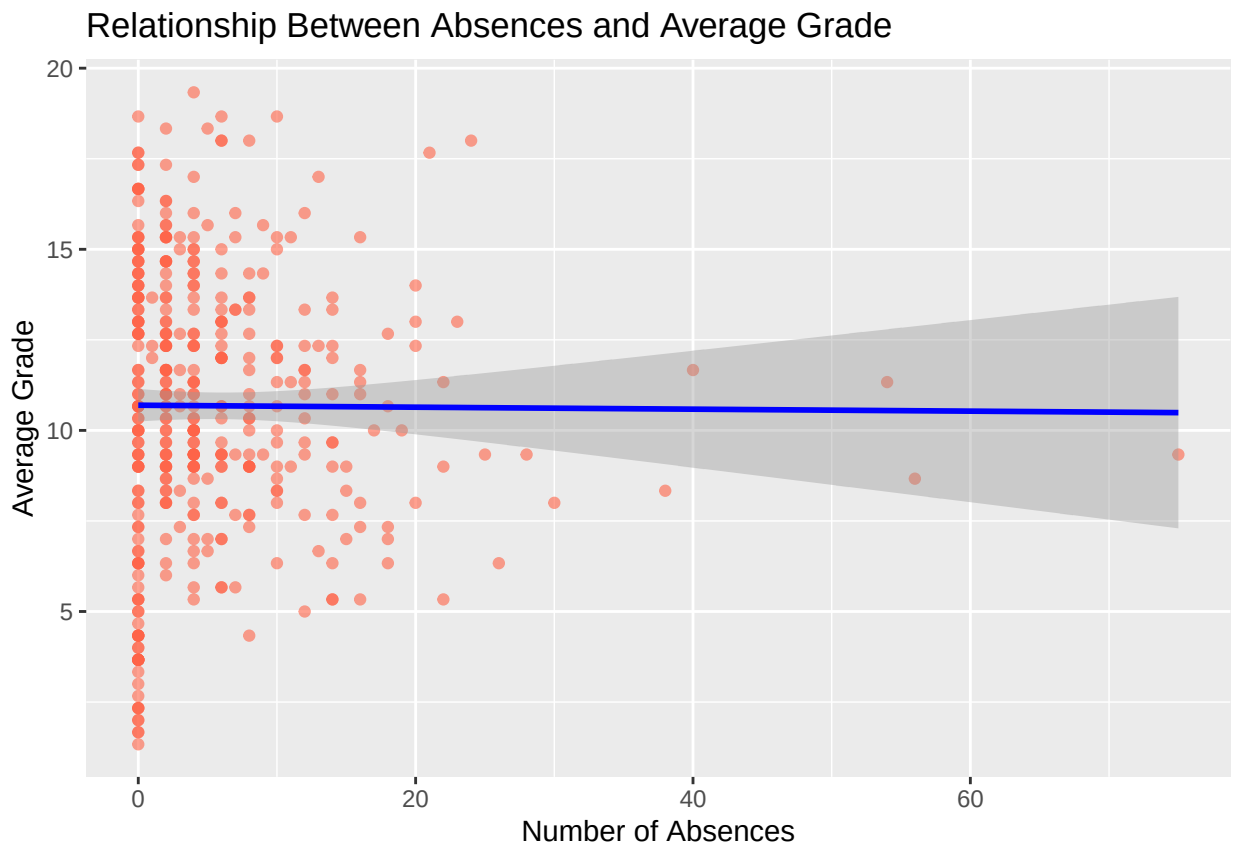
We can see that longer study times generally lead to better grade averages.

4. Absences vs Average Grades

Here, I checked to see if students who miss more classes tend to have lower average grades.

```
ggplot(df, aes(x = absences, y = avg_grade)) +
geom_point(color = "tomato", alpha = 0.6) +
geom_smooth(method = "lm", se = TRUE, color = "blue") +
labs(title = "Relationship Between Absences and Average Grade",
x = "Number of Absences",
y = "Average Grade")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



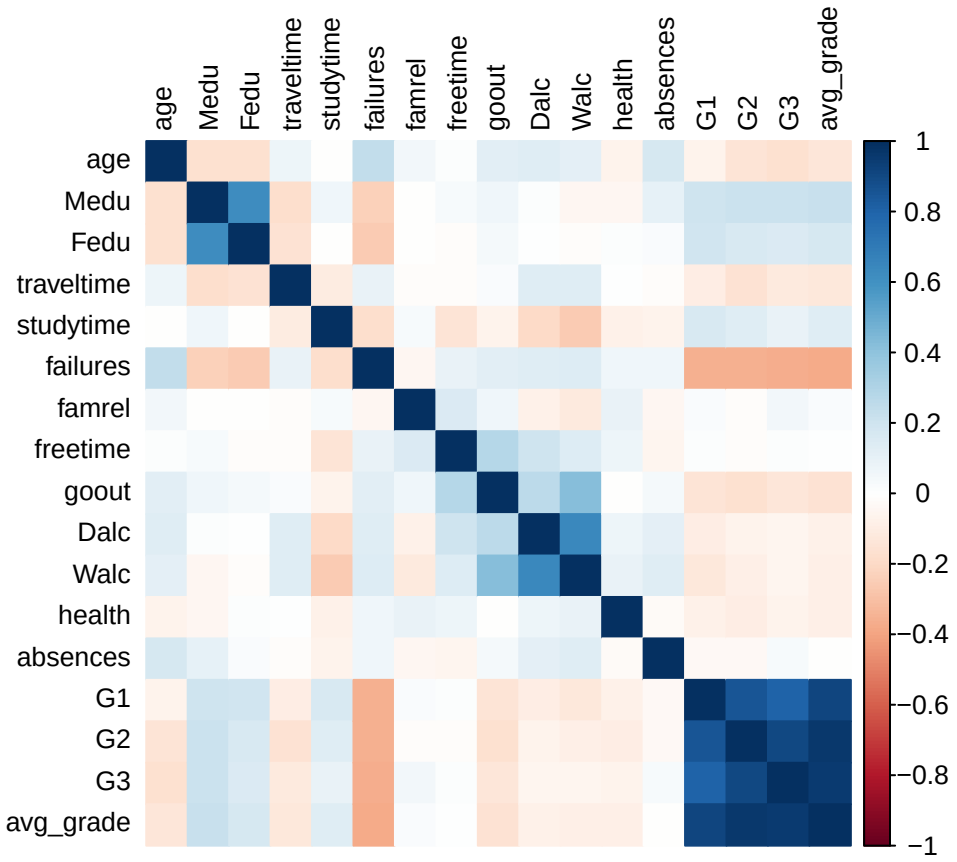
From this, we can see the general trend, as the number of absences goes up, the average grade usually goes down.

5. Correlation

Next, I looked at correlations between all numeric variables to see which factors might influence performance.

```
# pick only numeric columns
num_vars <- df[, sapply(df, is.numeric)]
corr_matrix <- cor(num_vars, use = "complete.obs")

# plot
corrplot(corr_matrix, method = "color", tl.col = "black", tl.cex = 0.8)
```



Darker colors show stronger relationships. For example, between G1, G2, and G3 grades.

#6. Parent's Education vs Average Grades I also checked if having school support or a mother with higher education makes a difference in grades.

```
# average grade by school support
school_support_avg <- aggregate(avg_grade ~ schoolsup, data = df, FUN = mean)
print(school_support_avg)
```

```
##   schoolsup avg_grade
## 1        no 10.875000
## 2       yes  9.359477
```

```
# average grade by mother's education
mother_edu_avg <- aggregate(avg_grade ~ Medu, data = df, FUN = mean)
print(mother_edu_avg)
```

```
##   Medu avg_grade
## 1    0  12.55556
## 2    1   9.19774
## 3    2  10.21683
## 4    3  10.45118
## 5    4  11.83969
```

We can see the average grade differences for each group.

7. Linear Model

Finally, I built a simple linear model to predict the average grade using selected variables.

```r
# Predict average grade using selected variables

model <- lm(avg_grade ~ studytime + failures + absences + health + sex, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = avg_grade ~ studytime + failures + absences + health +
##     sex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3933 -2.1783 -0.0255  2.4270  9.0649
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.19931    0.72662  14.037  < 2e-16 ***
## studytime    0.52758    0.21709   2.430 0.015539 *
## failures    -1.78887    0.23337  -7.665 1.44e-13 ***
## absences     0.01550    0.02146   0.722 0.470619
## health      -0.18743    0.12414  -1.510 0.131904
## sexM         1.22872    0.36235   3.391 0.000768 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.382 on 389 degrees of freedom
## Multiple R-squared:  0.1736, Adjusted R-squared:  0.163
## F-statistic: 16.34 on 5 and 389 DF,  p-value: 1.208e-14
```

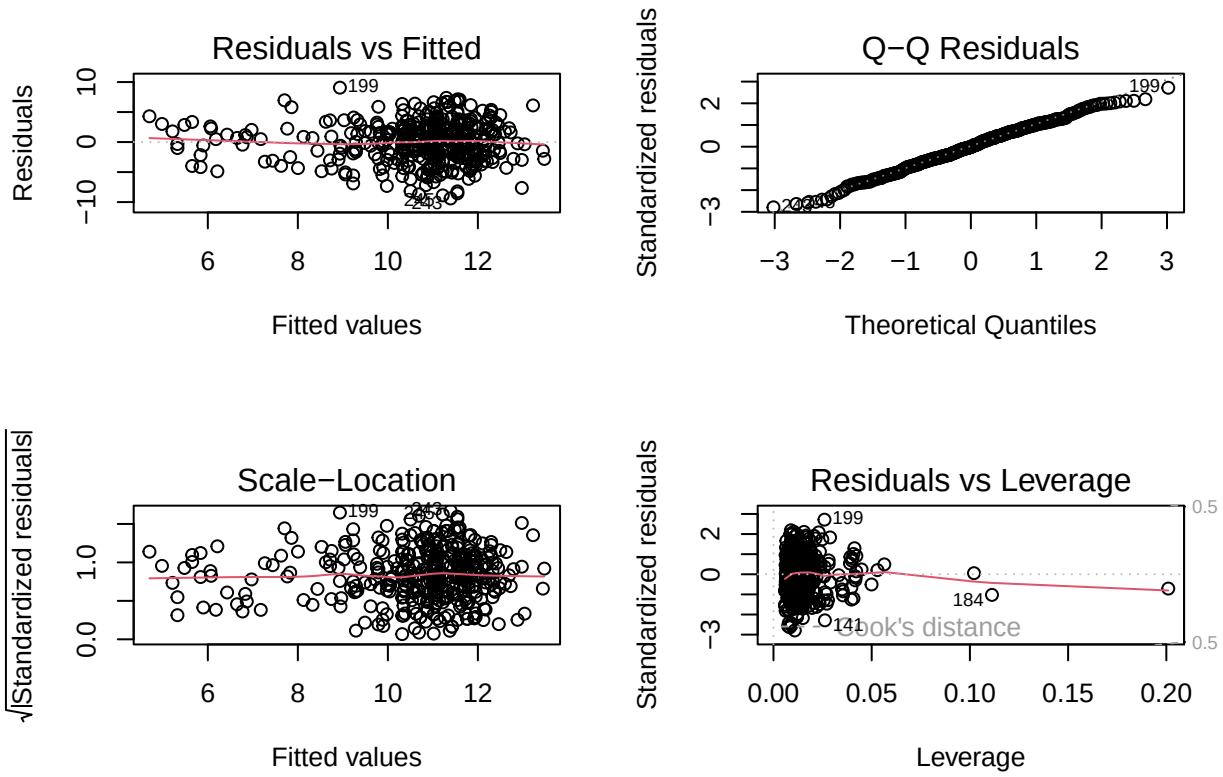
This shows which variables are most important for predicting student performance.

8.

I checked diagnostic plots to see if the model fits the data well.

```r
# Predict average grade using selected variables

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



9.

Conclusion

```
## Overall, students with higher study time and fewer failures tended to have better average grades.
## Health and absences also showed some effect, but gender and school support were less important.
```