REPORT

KAYLIE LE

Student Performance Analysis

May 24, 2025

Table of Content

Introduction and Key Findings

Understanding the key factors influencing student performance remains a central focus in educational research. This analysis explores how gender, diet quality, part-time job status, and weekly study hours relate to exam outcomes among university students. By integrating both ANOVA statistical modeling and visual exploration techniques, the analysis aims to identify which variables most effectively predict academic success.

Initial visualizations — including histograms, boxplots, and scatter plots — provided useful insights into the general distribution of exam scores across demographic and behavioral groups. These descriptive tools suggested a potential association between study time and exam performance, while indicating relatively consistent score distributions across gender, job status, and diet groups.

Together, the combination of exploratory visual analysis and formal statistical testing offers a comprehensive understanding of performance patterns. These findings highlight study time as a key predictor of academic outcomes, and suggest that lifestyle and demographic variables may be less influential in determining student success in this context.

In particular, through the process of combining data visualizations with ANOVA, I was able to draw several meaningful insights from the analysis:

- First and most clearly, study hours per week turned out to be the strongest factor associated with exam performance. Both the scatter plot and the ANOVA results confirmed a consistent upward trend: the more students studied, the higher their scores. Statistically, this relationship was highly significant ($F = 2135.882$, $p < 0.001$), which gave me solid confidence in the result.
- On the other hand, factors like gender, diet quality, and part-time job status didn't show any statistically significant effect on exam scores. All p-values were above 0.7, which matched what I observed visually — the boxplots showed very similar score distributions across these groups.
- What I found especially helpful was how the visualizations and the statistical test supported each other. The charts gave me a quick, intuitive understanding, and ANOVA helped confirm which patterns were actually meaningful.

In the end, among all the variables I looked at, study time clearly stood out as the most important contributor to exam success. While the other factors may still play a role in students' lives or well-being, they didn't show a direct impact on performance in this dataset.

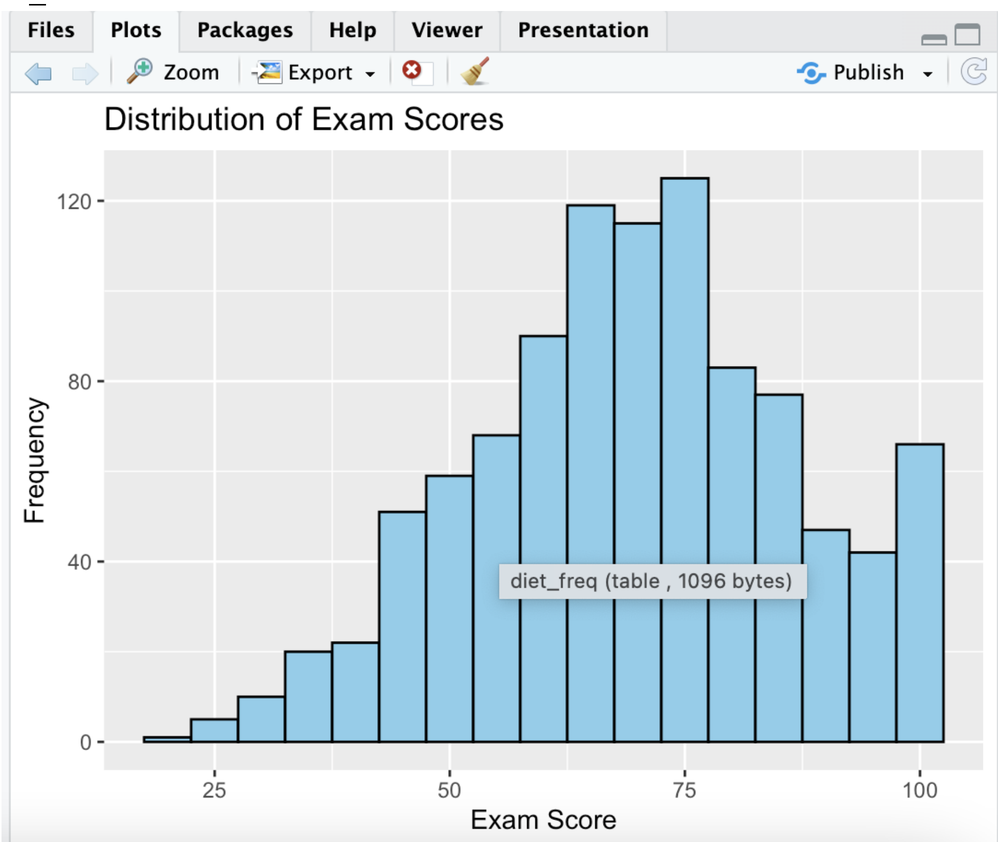
Visualizations

The reason for combining different visualizations was to clearly support the central goal of the analysis — to find out how gender, diet quality, part-time job status, and weekly study hours relate to exam outcomes among students.

To break this down, the analysis needed to explore both individual variable distributions and comparisons between groups — and this is why multiple types of visualizations were used:

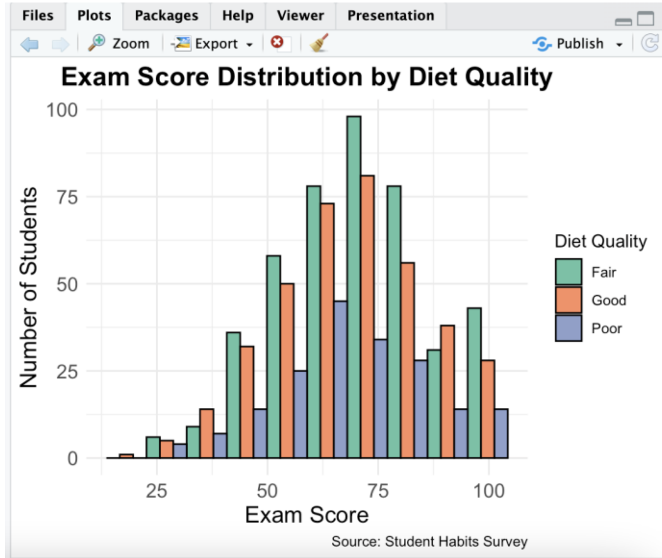
- Histograms were used to examine the overall distribution of continuous variables like exam_score and study_hours_per_week. This helped determine whether scores were normally distributed and whether any outliers or skewness might affect the analysis.
- Boxplots allowed quick visual comparison of exam scores across different groups, such as gender, diet quality, and job status.
- Scatter plots were chosen specifically to assess the relationship between study hours and exam scores, the only continuous-to-continuous relationship in the dataset.

Histogram of exam_score



The histogram shows that most students score between **60 and 85**, forming a near-normal distribution. A large concentration of scores appears around **70–75**, suggesting this is a common performance level.
Very few students score below 40 or above 95, indicating that **low and extremely high scores are rare**.

Histogram of exam_score by diet_quality



The histogram shows the distribution of exam scores across students with varying diet quality. While the overall shape of the distribution is similar for all three diet groups, students with "Fair" diets appear to dominate in terms of total count.

However, there is **no strong or consistent evidence** that diet quality significantly affects exam performance. All three groups tend to cluster within the 60–85 score range, suggesting that diet alone may not be a major determinant of academic success in this sample.

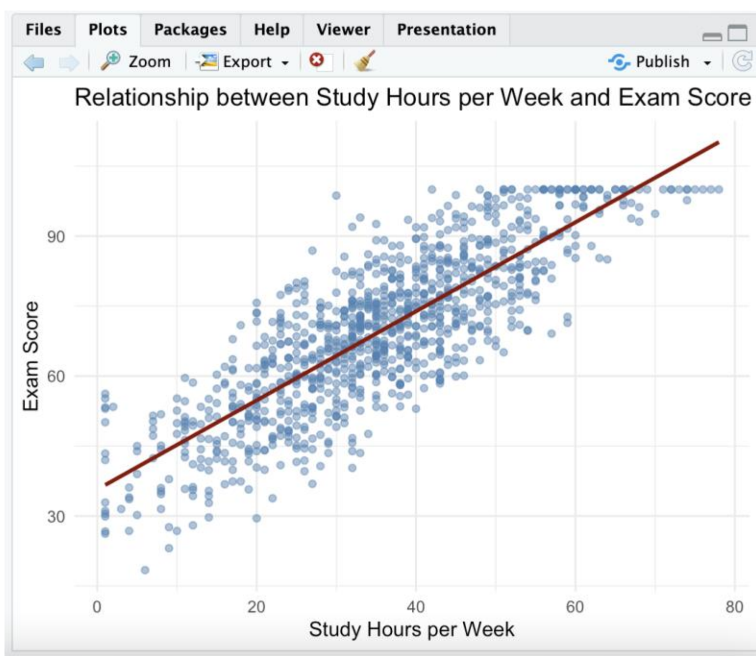
Boxplot of exam_score by gender

The boxplot reveals that exam scores across genders are **relatively similar in both median and spread**. All three gender groups—Female, Male, and Other—have median exam scores around 70 to 75, and score ranges that extend from roughly 20 to 100.

Boxplot of exam_score by part_time_job

The boxplot indicates that students with and without part-time jobs perform similarly on exams. The **median exam scores** are almost identical, and both groups show comparable distributions in terms of spread and range. Although there are slightly more **low-score outliers** in the "No Job" group, this difference is minimal and not statistically significant.
Overall, this visualization supports previous statistical findings that **having a part-time job does not meaningfully affect academic performance**.

Scatter Plot to visualize the relationship between study_hours_per_week and exam_score



The scatter plot clearly illustrates a **strong positive linear relationship** between the number of study hours per week and exam scores. Students who dedicate more time to studying tend to achieve higher results.
The linear regression line shows a consistent upward trend, and the relatively tight clustering of data around this line suggests that **study time is a reliable predictor of academic performance**.

How do you show and compare the results across multiple classes?
Although initial visualizations such as boxplots and histograms offered valuable insights into how exam scores differ across groups (e.g., gender, diet quality, and part-time job status), these visuals alone **cannot determine whether the differences are statistically significant**.
Therefore, a formal **Analysis of Variance (ANOVA)** was conducted to test whether the mean exam scores differ meaningfully across these categorical variables.
Since the goal is to assess score variation across multiple groups—such as **gender (Female, Male, Other)**, **diet quality (Good, Fair, Poor)**, and **part-time job status (Yes, No)**—a **one-way ANOVA** is the most appropriate statistical method.
It enables comparisons of **group means** while also incorporating a **numeric covariate**, study_hours_per_week. Moreover, ANOVA avoids the inflated error risk of running multiple individual t-tests, providing a more **reliable and statistically sound analysis** of which factors significantly influence exam performance.

```
> # ANOVA
> anova_model <- aov(exam_score ~ study_hours_per_week + gender + part_time_job
+                    + diet_quality, data = data)
> summary(anova_model)
                      Df Sum Sq Mean Sq  F value Pr(>F)
study_hours_per_week   1 194441  194441 2135.882 <2e-16 ***
gender                 2     38      19    0.210  0.811
part_time_job          1      2       2    0.023  0.880
diet_quality           2     59      29    0.323  0.724
Residuals            993  90398      91
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion

The ANOVA results show that **study hours per week** have a highly significant effect on exam scores (*F = 2135.882, p < 0.001*), indicating that the more time students spend studying, the higher their exam performance.

This result is also visually confirmed in the **scatter plot**, where a clear upward trend exists between weekly study time and exam scores.

In contrast, other variables — including **gender** (*p = 0.811*), **part-time job status** (*p = 0.880*), and **diet quality** (*p = 0.724*) — do not show statistically significant differences, as seen in both the **ANOVA table** and the corresponding **boxplots**, which reveal similar score distributions across groups.

Conclusion & Recommendation

This analysis highlighted weekly study hours as the most significant factor associated with students' academic performance. Among the variables examined — including gender, diet quality, and part-time job status — only study time showed a statistically significant impact on exam scores. The consistency between visual trends and statistical results reinforced the reliability of this finding. While demographic and lifestyle variables may still play a role in students' overall experience, they did not appear to directly influence academic outcomes in this dataset.

One of the strengths of the analysis was the integration of data visualization and statistical testing. Visual tools such as scatter plots and boxplots helped uncover patterns at a glance, while the application of ANOVA confirmed which of these patterns held statistical significance. This combination allowed for a more balanced and evidence-based interpretation of the results.

Based on these findings, it is recommended that students be encouraged to adopt consistent study routines, even if only modest increases in weekly study time are possible. For educators and academic advisors, the results suggest that supporting students' time management and study planning may be more effective in improving performance than focusing on factors such as gender or employment status.

Lastly, although this analysis provides clear insight into the impact of study time, it also points toward opportunities for future research. Further studies could explore how variables such as study quality, motivation, learning environment, or mental well-being interact with study habits to influence academic success in a more holistic way.

References

MarinStatsLectures. (2016). *Analysis of variance (ANOVA) in R* [Video]. YouTube. https://www.youtube.com/watch?v=qrP7evoNCy4

Naath, J. (2025). *Student Habits vs Academic Performance* [Dataset]. Kaggle. https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance

R Programming 101 (2021). *An introduction to data visualization using R programming* [Video]. YouTube. https://www.youtube.com/watch?v=HPJn1CMvtmI&list=PLtL57Fdbwb_C6RS0JtBojTNOMVlgpeJkS