Visualizing patterns using R:
Factors That Shape Student Performance

Jean Batista

Introduction – Developing the Story

In this project I explore how study habits, family background, and support programs interact to shape students' final mathematics grades. I use the UCI Student Performance dataset, which records grades and a range of demographic and educational variables for secondary-school students in Portugal. My goal is to develop a narrative around what factors help or hinder success, using a series of visualizations to reveal patterns and relationships. The story outlined below guides the choice of plots and the interpretation of the results.

Key Variables

Numeric variables

- **age** – student age in years
- **class_failures** – number of previous class failures
- **family_relationship** – quality of family relationship (1–5 scale)
- **free_time** – free time after school (1–5 scale)
- **social** – going out with friends (1–5 scale)
- **weekday_alcohol** – weekday alcohol consumption (1–5 scale)
- **weekend_alcohol** – weekend alcohol consumption (1–5 scale)
- **health** – current health status (1–5 scale)
- **absences** – number of school absences
- **grade_1**, **grade_2**, **final_grade** – grades from first, second and final periods (0–20 scale)

Categorical variables

- **sex** (F/M)
- **study_time** – weekly study time category ("<2 hours", "2–5 hours", "5–10 hours", ">10 hours")
- **mother_education**, **father_education** – parental education levels (none, primary, 5th–9th grade, secondary or higher education)
- **school_support**, **family_support**, **extra_paid_classes** – indicators of extra educational support
- **activities**, **nursery_school**, **higher_ed**, **internet_access**, **romantic_relationship** – further yes/no variables representing extracurricular activities, early education, aspirations, internet access and personal relationships

My analysis focuses on the final_grade as the outcome and uses the above variables to explore predictors of performance.

Week 5 Skills – Data Management with Pipes

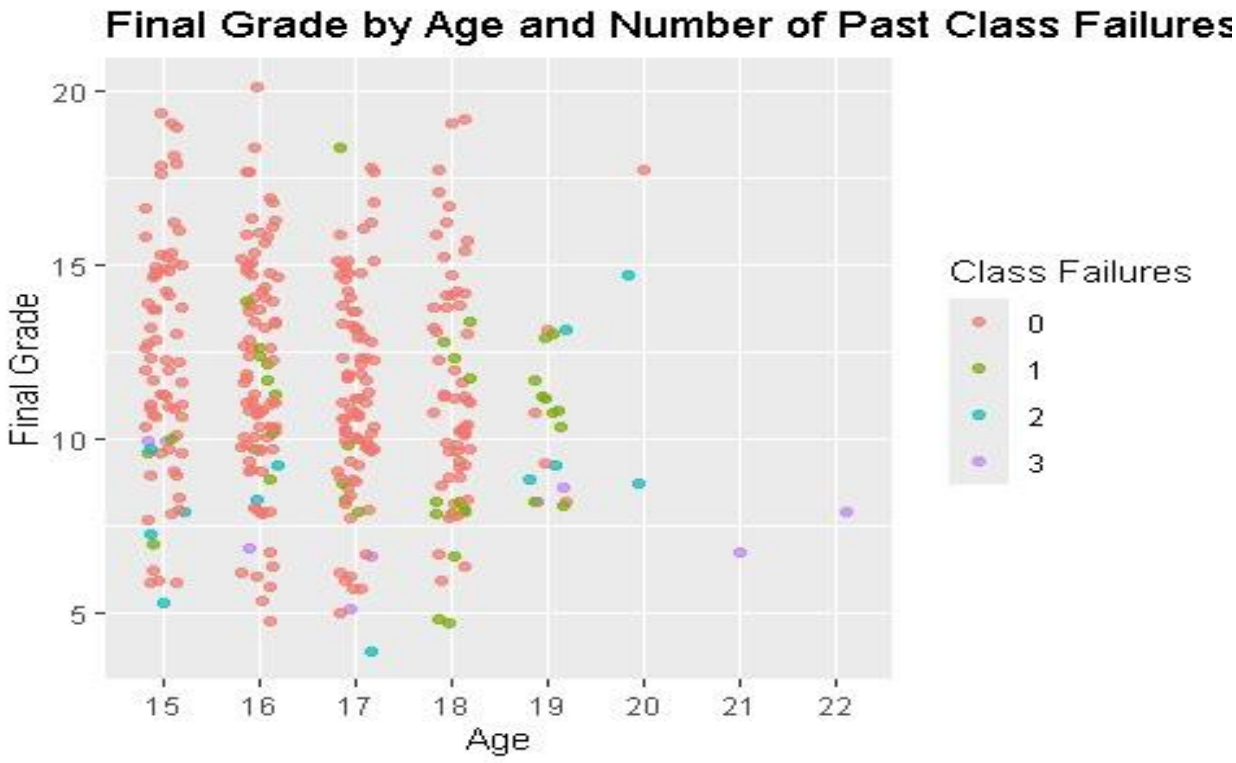
I began by cleaning the data. After examining the grade distributions, I noticed a concentration of zeros in **grade_1**, **grade_2** and **final_grade**. Investigation showed that these zeros correspond to dropouts – students who left the course and thus have missing grades. To handle these, I replaced 0s with missing values and created a `dropped_out` indicator. I also converted study time and other ordinal variables to ordered categories, renamed variables for clarity, removed duplicates, and checked for typos. All transformations were performed using a piped workflow of `mutate`, `na_if`, `if_else`, `select`, and `group_by` operations to streamline the process.

Week 6 Skills – Describing Variation and Covariation

With the cleaned data, I turned to visualizing variation (distributions) and covariation (relationships between variables). For variation, histograms, densities and boxplots reveal how grades and absences are distributed. For covariation, scatterplots, heatmaps and stacked bars show how multiple predictors interact. In each case I used summarization functions (e.g., `group_by` + `summarize`) to compute means and counts, and faceting or colour mapping to overlay categories. Below I present nine visualizations that build the narrative.

Visualizations and Discussion

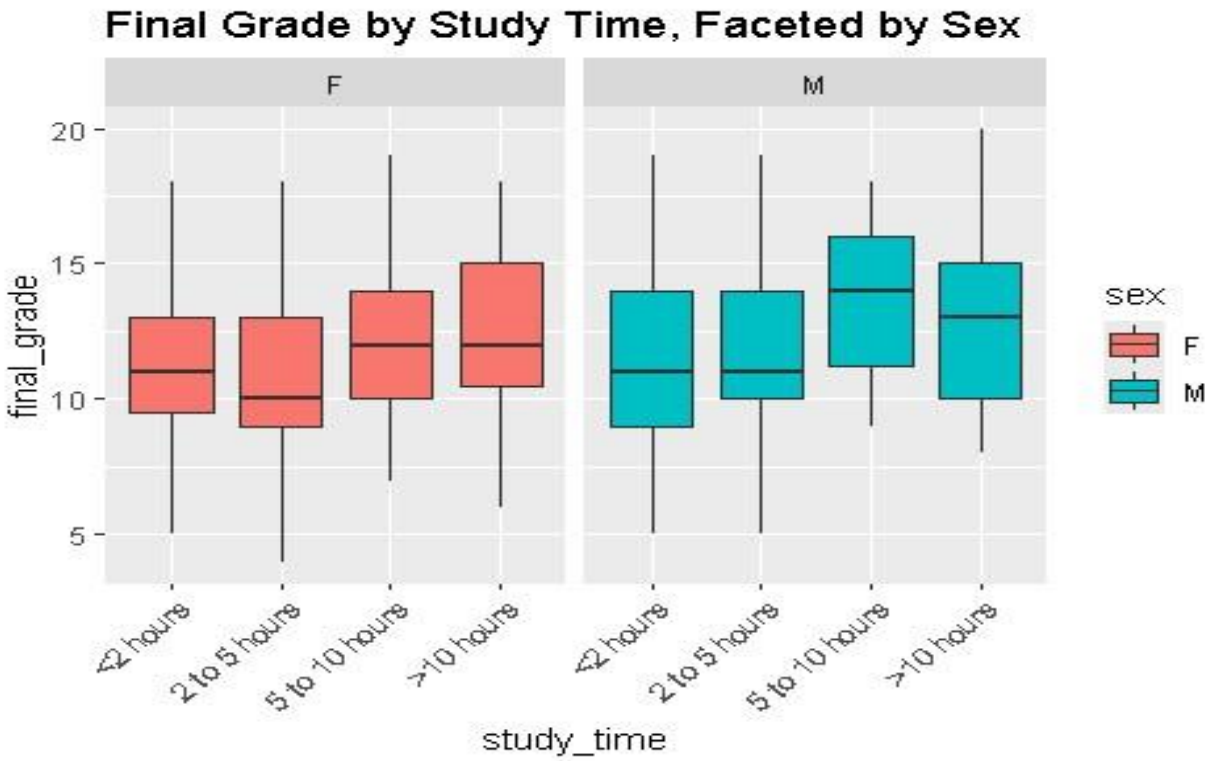
1 Final grade by age and past failures



Older students tend to have lower final grades, and students with more prior failures (darker color) perform worse. The scatter also shows few older students with many failures, consistent with repeaters. There are some students with class failures > 1& 2 even at 15 years old. Their final grade is very low.

2

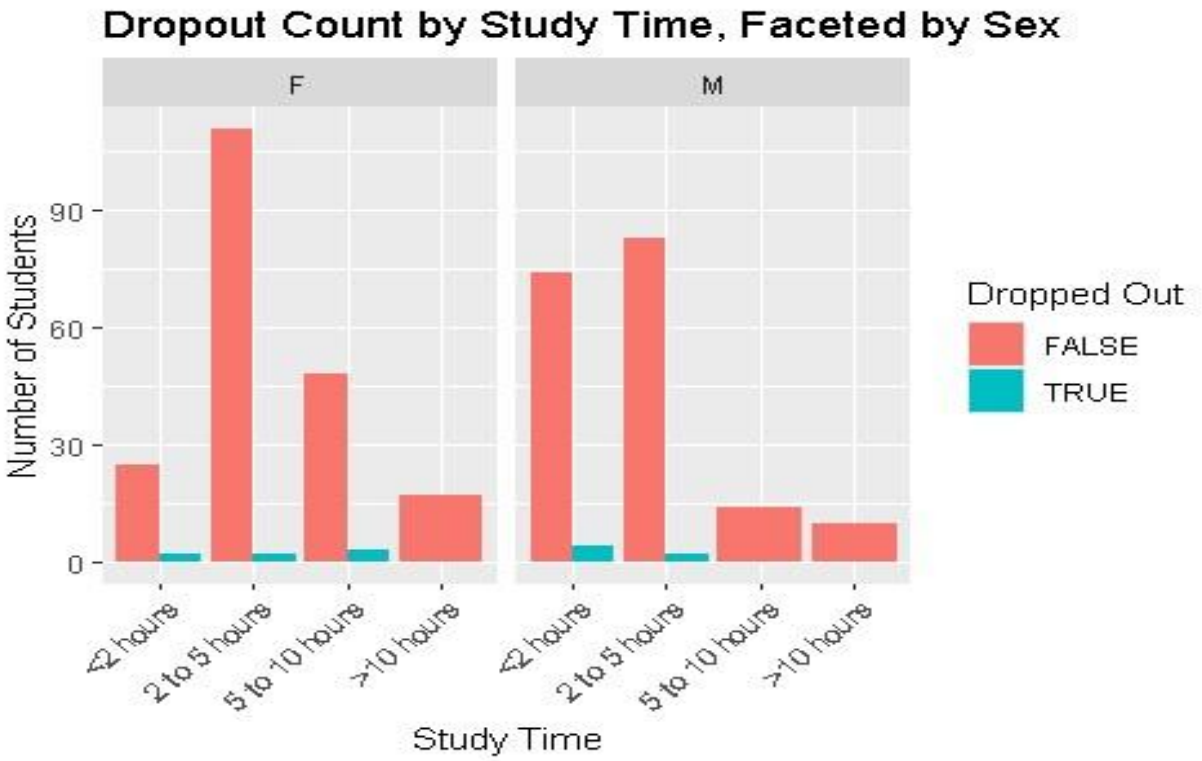
Final grade by study time and sex



Students who study more hours each week obtain higher grades. The positive gradient is similar for both sexes, although males show slightly more variation at high study times.

3

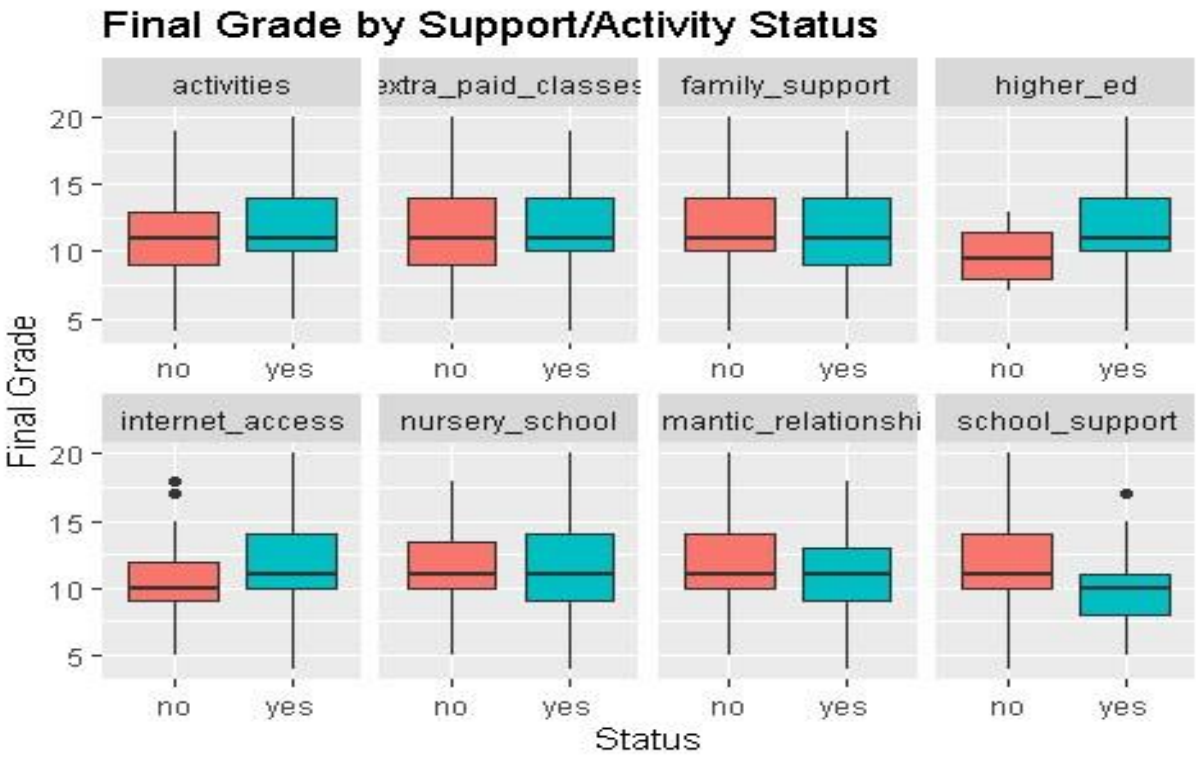
Dropout count by study time



Students who study less are more likely to drop out. Dropout cases cluster in the "<2 hours" and "2–5 hours" groups. Females students reflect better study habits.

4

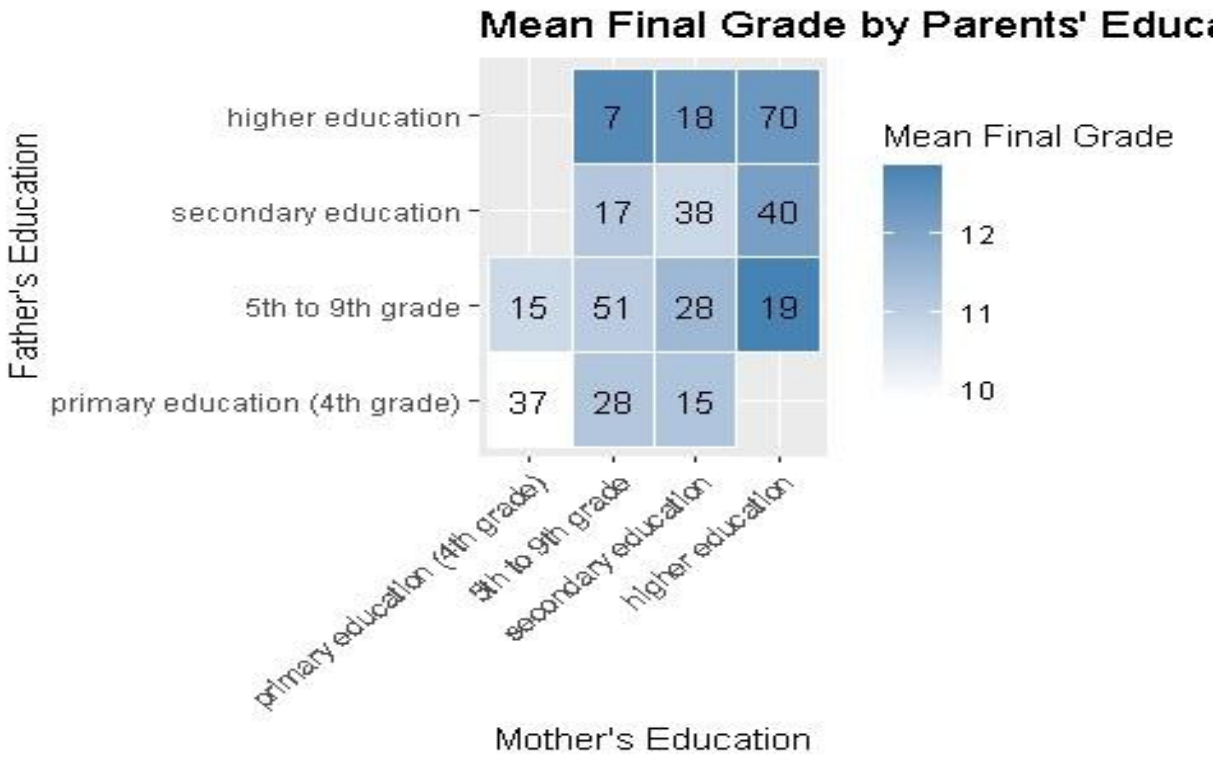
Final grade by support/activity variables



Boxplots across eight binary variables reveal different patterns. Having higher education aspirations and internet access is associated with higher grades. Students receiving school support have lower grades, indicating that support is targeted at struggling students.

Mean final grade by parents' education

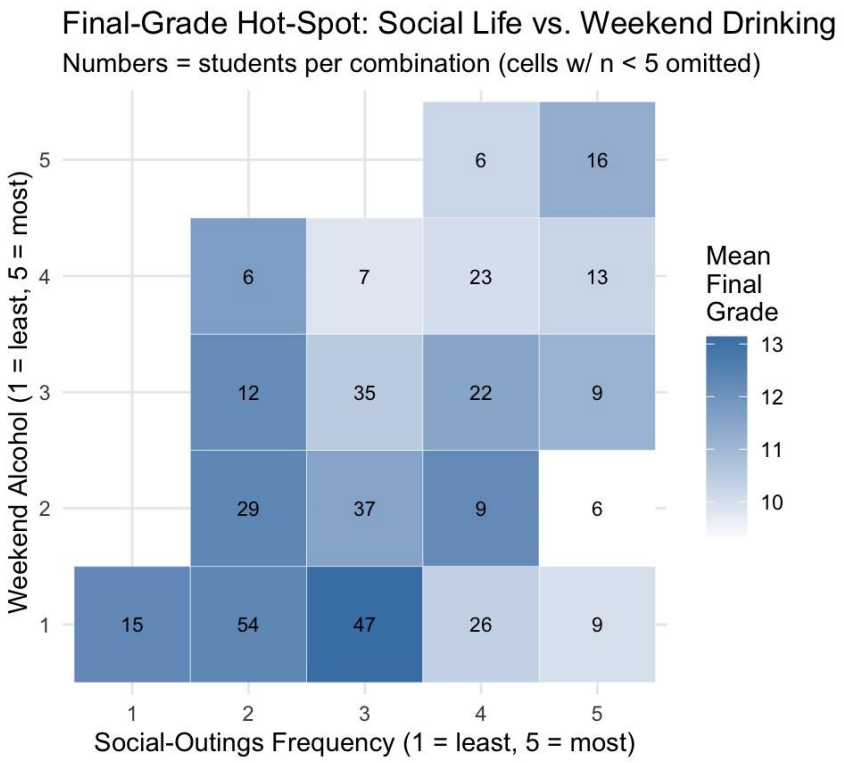
5



The heatmap shows mean grades only for cells with more than five students. Grades increase with parental education: the highest averages occur when both parents have higher education. Low-sample combinations are omitted to avoid misleading results.

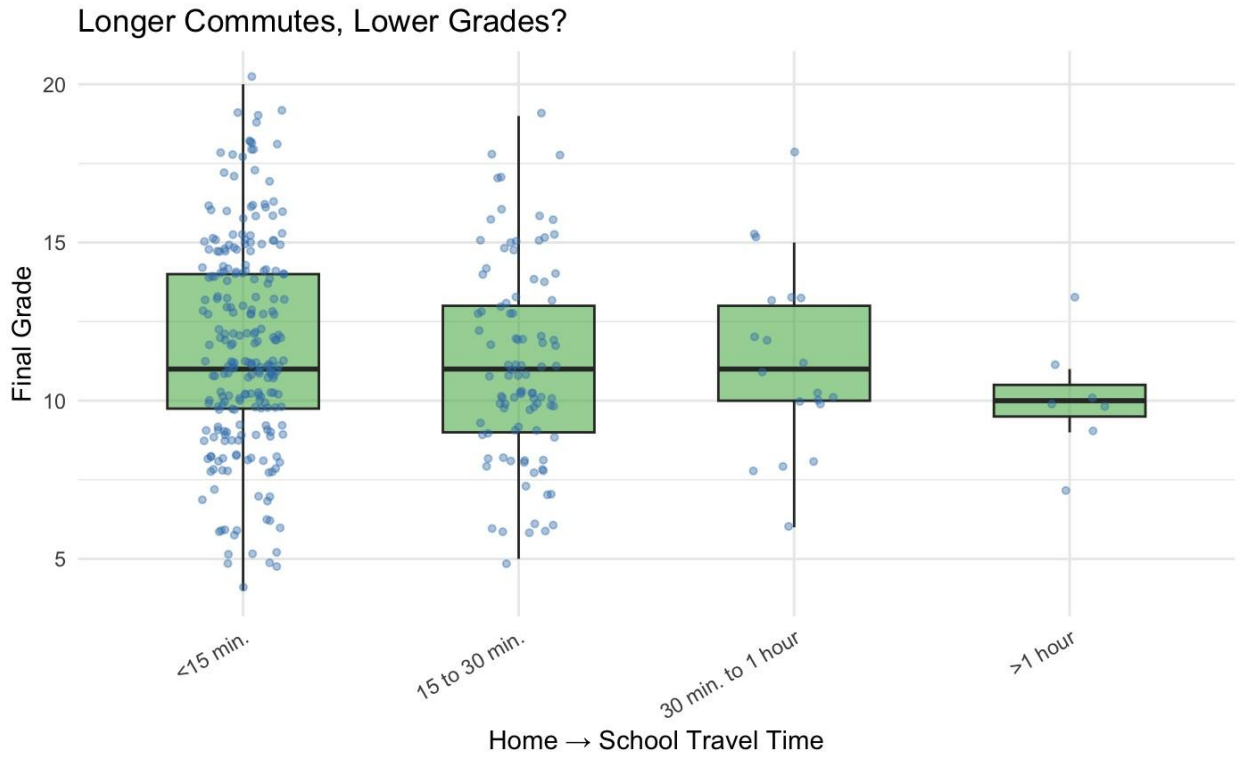
Mean final grade by social frequency and alcohol usage

6



As students attend more social outings, there are more likely to consume alcohol on the weekends. Both of these factors negatively impact the grade of the student. Students who do both drinking and going out are very likely to perform bad, while students who do only one activity at a low to moderate level can do fine.

Final grade by travel time.



Box plots illustrate that quick access to school has a modest positive effect on grades. However, the median is roughly the same across students who travel anywhere from <15 to 1 hour, indicating this factor alone do not guarantee high performance.

Conclusion

This analysis suggests that multiple factors including time investment in studying, parental background, and avoiding alcohol and frequent social outings are key drivers of student success. Support programs help but may not fully counteract disadvantages stemming from prior failures or low socioeconomic status. Internet access and paid classes, plus quick and reliable travel time to school offer only a small advantage. Understanding these patterns can

help educators target interventions more effectively and encourage habits that lead to improved academic outcomes.