Student Performance Visualization

Oskar Nesheim, August Nyheim, Håkon Refsvik

Abstract

Understanding the factors that influence student performance is crucial for educators and policy-makers. This project analyzes a secondary school students' math performance dataset sourced from Kaggle, using visual analytics to uncover key insights. The dataset includes demographic, familial, and lifestyle factors and their correlation with academic grades. Our approach combines interactive visualizations, such as scatterplots, heat maps, and histograms, to make complex data intuitive and accessible. Key findings include the strong influence of parental education, study habits, and reduced alcohol consumption on student grades, alongside the negative impact of long travel times and past failures. The visualizations effectively highlight trends for high-achieving students, though patterns for average or underperforming students were less distinct, suggesting additional unmeasured factors. Future work will involve expanding the dataset and incorporating qualitative data. This project aims to provide actionable insights for improving educational outcomes through data-driven decisions.

1 Introduction

Understanding what affects student performance in school has been studied for many years. Many factors, such as study habits, parental support, and social background contribute to how well students perform. However, identifying the most important factors can be difficult because educational data is often complex and hard to interpret.

In this project, we analyze a secondary school students' math performance dataset to uncover key factors that influence grades. Looking at the data independently can be overwhelming, so we use visual analytics to make the analysis more intuitive. By creating interactive visualizations, we aim to reveal patterns and connections in the data that are hard to see when the data is static but become more intuitive when you interact with the data. These insights will help highlight the most critical factors that impact academic success and provide useful information for educators, decision-makers and students themselves.

2 Related Work

Many studies have explored the factors that affect student performance, identifying key contributors such as parental involvement and substance use. A meta-analysis by Wilder (2014) found that when parents are more involved in their children's education, students tend to perform better academically. Similarly, Latendresse et al. (2010) showed that higher levels of parental involvement are strongly linked to better grades among adolescents.

However, substance use, particularly alcohol consumption, has been shown to negatively impact academic performance. For example, Singleton and Wolfson (2009) found that college students who consume more alcohol tend to have lower GPAs and face more academic challenges. This is further supported by reports from the National Institute on Alcohol Abuse and Alcoholism (NIAAA), which revealed that about 25% of college students experience academic problems caused by alcohol, such as missing classes or failing to complete assignments.

Dimensionality reduction techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) are commonly used to simplify large and complex datasets. Jolliffe and Cadima (2016) reviewed PCA and highlighted its effectiveness in reducing the complexity of data while retaining the most important information. Meanwhile, Van der Maaten and Hinton (2008) introduced t-SNE as a method for visualizing high-dimensional data in two or three dimensions. This method is particularly useful for finding patterns and clusters in the data.

Combining these techniques with interactive visualizations makes it easier to explore and understand complex datasets. Heer et al. (2010) emphasized the importance of interactive visualizations in making data more accessible and helping users gain deeper insights. By using these tools, researchers and

educators can better analyze the factors that influence student performance and find ways to improve academic outcomes.

3 Data and User Context

3.1 Dataset

In this project, we used the student performance dataset, sourced from Kaggle, as recommended in the project description. The dataset contains data on secondary school students' math performance. It was chosen for its relevance, as understanding factors influencing school performance is an important topic for both educators and students.

The dataset comprises 33 columns (features) of varying formats, including integers, strings, and boolean values. Of these, 30 features are used to predict math grades for the first, second, and third terms, scored on a scale of 0 to 20, where 0 is the worst and 20 is the best. The dataset includes 395 rows (students).

Key features include demographic information (e.g., gender, age, family size), parental background (e.g., education of mother and father), and lifestyle factors (e.g., family relations, health). These attributes offer a comprehensive view of factors that might contribute to academic performance. A full list of the attributes and their description can be found in the Appendix A.1

3.2 Intended user

The primary users of this visual analytics system are educators, school administrators, and policymakers. By uncovering the key factors that influence student performance, this system can guide interventions to improve academic outcomes. For instance, educators can identify students who may benefit from additional support, while policymakers can design programs to address broader trends in student achievement.

The project also resonates personally with us as students, offering insights into strategies that might improve our own academic performance. The findings could help students understand how various aspects of their lives, such as study habits or family dynamics, impact their grades.

4 Design process

In the initial stages of designing the visualization, we faced a challenge in selecting the most appropriate method due to the structure of our dataset. With a relatively small number of data points but a large number of attributes, we quickly realized that some common visualization techniques were not effective in conveying the relationships within the data.

Our first attempts included using boxplots, spider charts, and parallel coordinates. These tools, while valuable in other contexts, did not work well for this dataset. The boxplots, for example, struggled to communicate meaningful insights because the data was ordinal (ranging from 0 to 4 or 0 to 5) rather than continuous. This made it difficult to represent the scale in a way that made sense. The precision provided by boxplots (such as showing medians and quartiles) wasn't appropriate for ordinal data, where the differences between adjacent values weren't uniform.

Similarly, spider charts and parallel coordinates failed because, with a small number of data points, many of the lines ended up overlapping and obscuring each other, making it hard to distinguish meaningful patterns or clusters. This overcrowding in the charts reduced their effectiveness significantly, as it became challenging to see how data points were distributed across different attributes. We realized that visual clutter was a major issue, and the lines in these charts simply didn't represent the data clearly enough.

As we explored further, we realized the issue with heatmaps as well. Heatmaps seemed to be a promising option for visualizing ranked data (such as study time, health, or other similar attributes on a scale of 1-5), but they presented their own challenges. Initially, we wanted to make the heatmap interactive, allowing users to select specific bins or attributes and drill down for more insights. However, when the number of data points in a selected bin was too small, the heatmap no longer conveyed useful

information. The result was a visualization that looked too sparse to offer any real insights, which made the interactivity less effective.

At this point, we also recognized that the ultimate goal of the analysis was to understand how students' grades were influenced by different factors. Grades became the central focus of our visualization efforts, as they were what the user would be most interested in when analyzing the data. With this in mind, we wanted to incorporate color coding into our scatterplots, using the color of each point to represent the student's grade. This was a natural choice, as it allowed users to immediately see how grades were distributed in relation to various factors, such as study time or parental education. By making the color of the points correspond to grades, we gave the user a clear, intuitive way to focus on what they care about most, making it easier to spot clusters, trends, or outliers.

Ultimately, the challenge of managing a large number of attributes with a small set of data points led us to explore scatterplots with grade-based color coding. This approach allowed us to clearly identify patterns and relationships in the data while minimizing the visual clutter associated with other methods. It also provided a simple yet effective way to visualize the distribution of students' attributes and grades, which became the core of our design.

5 Prototype Description

Scatterplots with Grade-Based Coloring

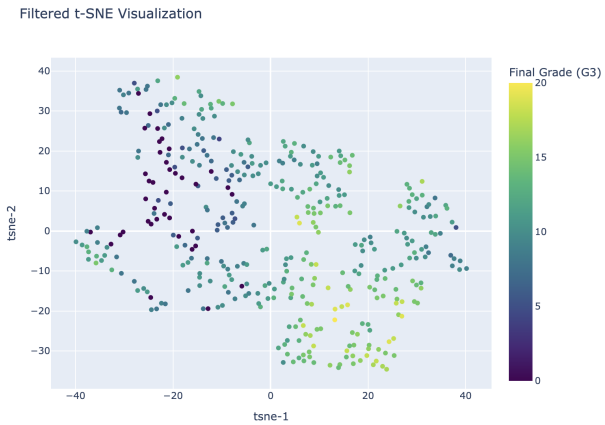

Figure 1: Scatterplot

We chose scatterplots as the primary visualization tool because they provide a clear way to explore the relationships between variables. To enhance their effectiveness, we incorporated color coding for the data points based on grades. This design makes it easy to spot clusters or trends in the data, such as groups of students with high or low grades. By visualizing grades as colors, users can quickly identify areas of interest, such as students with specific study times or levels of parental education. Furthermore, the scatterplots allow for highlighting specific clusters for further exploration, enabling users to interact with the data intuitively and discover deeper insights.

Heatmap

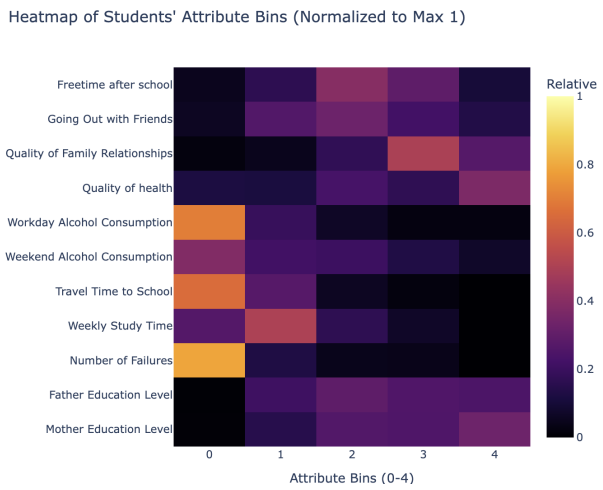

Figure 2: Heatmap

Heatmaps were used to represent data in which many attributes are ranked on a scale of 1-5, such as study time and health. This visualization effectively shows the distribution of the rankings within these bins. In Figure 2 above we can for example see how a large proportion of the students have 0 number of previous failures. Heatmaps prevent visual clutter associated with other techniques, such as parallel coordinates, which can become difficult to interpret when multiple attributes fall into the same range.

Histograms

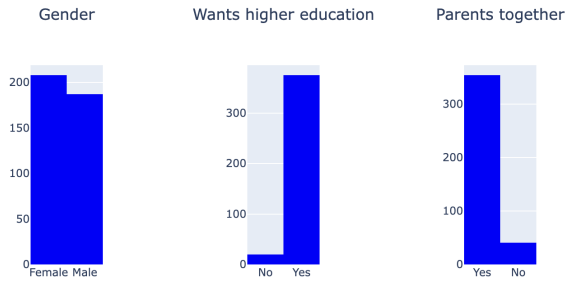

Figure 3: Histograms

We opted for histograms instead of heatmaps for visualizing binary data (e.g., male/female, true/false) because heatmaps appeared unintuitive in cases with only two cells per row. The histograms clearly illustrate the distribution of students across binary features, making it easier to interpret and analyze relationships. For example, these visualizations allow us to quickly see the proportions of male versus female students or determine which students want higher education.

6 Insights Discovered

Through visualizations and analysis, several key patterns emerged that helped uncover the factors influencing student performance. A significant insight was the relationship between parental education and student grades. Specifically, we discovered that students with lower grades tended to have parents with lower levels of education. In comparison, students with higher grades were more likely to have parents with higher education levels. This trend suggests that parental education is crucial in shaping student performance. It could reflect the influence of more educated parents in terms of providing better academic support, guidance, and access to resources that can contribute to academic success.

This finding highlights the importance of considering parental participation and education when addressing academic challenges. It also points to the need for targeted interventions to support students who

may not have the same level of educational support at home, which could help close the performance gap.

- Alcohol consumption: Students with lower alcohol consumption levels generally performed better academically, underscoring the potential impact of lifestyle choices on education. Both weekday and weekend alcohol consumption showed a negative correlation with grades. This was both prevalent in weekday and weekend alcohol consumption. However, weekday alcohol consumption showed the greatest negative correlation with good grades. This can be because students who drink alcohol during the weekdays will sleep worse, forget the things they have learned at school, and be subject to other negative consequences of alcohol.

- Past Failures: The number of previous failures significantly correlated with lower grades. This finding emphasizes the long-term effects of academic struggles, pointing to the need for interventions to prevent repeated failures. If a student fails a course it is very important for the school and/or parents to take action to prevent a downward spiral where the student ends up failing more courses. A single failure can plant a negative mindset in the students where they no longer believe they are not good enough.

- Study Time and Grades: Students who dedicated more weekly study hours tended to achieve higher grades. The difference between the lower and upper quartile was not striking, but a few more hours a week tended to correlate with higher grades. This highlights the critical role of effective time management and consistent study habits in academic performance.

- Travel time to school: Of the students that performed the best there seemed to be a trend where they spent less time traveling to school. In the upper quartile, 75 % of students spent less than 15 minutes traveling to school. In the lower quartile, this number was closer to 60 %. This trend might show how students that spend a long traveling to and from school have less free time to play and study.

7 Discussion and Future Work

Our design was effective in showcasing what factors were important in achieving good grades. What we had not anticipated was that when looking at the students with higher grades we saw clear trends that were prevalent. On the contrary, when looking at the average and below-average students it was harder to find clear patterns. This could mean that there may be a range of reasons why students do not perform that we have not been able to catch in this study.

The first step we would take to take this project further would be to get a larger and more diverse dataset. Getting a bigger dataset with more samples could lead to more accurate results. For example, in the heatmaps we created there seemed to be some limitation with regards to sparse data. Expanding the dataset with more students and additional variables could improve the robustness of insights. We could also incorporate some external data where we look at school infrastructure and teacher qualifications. Further, we could also look at a more qualitative approach where we took student surveys and talked to teachers.

References

Heer, J., Bostock, M., and Ogievetsky, V. (2010). A tour through the visualization zoo: A survey of powerful visualization techniques, from the obvious to the obscure. *Communications of the ACM*, 53(6):59–67.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

Latendresse, S. J., Rose, R. J., Viken, R. J., Pulkkinen, L., Kaprio, J., and Dick, D. M. (2010). Parental socialization and adolescents' alcohol use behaviors: Predictive disparities in parents' versus adolescents' perceptions of the parenting environment. *Journal of Clinical Child & Adolescent Psychology*, 38(2):232–244.

Singleton, R. A. and Wolfson, A. R. (2009). Alcohol consumption, sleep, and academic performance among college students. *Journal of Studies on Alcohol and Drugs*, 70(3):355–363.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Wilder, S. (2014). Effects of parental involvement on academic achievement: A meta-synthesis. *Educational Review*, 66(3):377–397.

A Appendix

A.1 Dataset Attributes

Attribute	Description
Feature Attributes	
school	Student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	Student's sex (binary: 'F' - female, 'M' - male)
age	Student's age (numeric: from 15 to 22)
address	Student's home address type (binary: 'U' - urban, 'R' - rural)
famsize	Family size (binary: 'LE3' - less or equal to 3, 'GT3' - greater than 3)
Pstatus	Parent's cohabitation status (binary: 'T' - living together, 'A' - apart)
Medu	Mother's education (numeric: 0: None, 1: Primary education (4th grade), 2: 5th to 9th grade, 3: Secondary education, 4: Higher education)
Fedu	Father's education (numeric: 0: None, 1: Primary education (4th grade), 2: 5th to 9th grade, 3: Secondary education, 4: Higher education)
Mjob	Mother's job (nominal: 'teacher', 'health', 'services', 'at_home', 'other')
Fjob	Father's job (nominal: 'teacher', 'health', 'services', 'at_home', 'other')
reason	Reason to choose this school (nominal: 'home', 'reputation', 'course', 'other')
guardian	Student's guardian (nominal: 'mother', 'father', 'other')
traveltime	Home-to-school travel time (numeric: 1: <15 min, 2: 15–30 min, 3: 30 min–1 hour, 4: >1 hour)
studytime	Weekly study time (numeric: 1: <2 hours, 2: 2–5 hours, 3: 5–10 hours, 4: >10 hours)
failures	Number of past class failures (numeric: n if 1 <= n < 3, else 4)
schoolsup	Extra educational support (binary: 'yes', 'no')
famsup	Family educational support (binary: 'yes', 'no')
paid	Extra paid classes within the course subject (binary: 'yes', 'no')
activities	Extra-curricular activities (binary: 'yes', 'no')
nursery	Attended nursery school (binary: 'yes', 'no')
higher	Plans for higher education (binary: 'yes', 'no')
internet	Internet access at home (binary: 'yes', 'no')
romantic	In a romantic relationship (binary: 'yes', 'no')
famrel	Quality of family relationships (numeric: 1 - very bad, 5 - excellent)
freetime	Free time after school (numeric: 1 - very low, 5 - very high)
goout	Frequency of going out with friends (numeric: 1 - very low, 5 - very high)
Dalc	Workday alcohol consumption (numeric: 1 - very low, 5 - very high)
Walc	Weekend alcohol consumption (numeric: 1 - very low, 5 - very high)
health	Current health status (numeric: 1 - very bad, 5 - very good)

Attribute	**Description**
absences	Number of school absences (numeric: from 0 to 93)
Course Grades	
G1	First period grade (numeric: 0 to 20)
G2	Second period grade (numeric: 0 to 20)
G3	Final grade (numeric: 0 to 20) (Output target)

Table 1: Descriptions of dataset attributes