Minh Huy Ngo

CIS.3389.001

Professor Aindrila Chakraborty

**Data Report**

**Introduction**

This data contains information for student achievement data for 2 Portuguese high schools. The data was collected using school reports and questionnaires and includes student grades, demographics, social, parent, and school-related features. The dataset is provided regarding performance in two distinct subjects: Mathematics and Portuguese language. The data set in use had already been cleaned up for ease of reading and use.

The primary aim of the report is to ascertain if there exists a correlation between student performance and their gender and age. Additionally, it seeks to delve deeper into the impact of their upbringing—factors like parental cohabitation status, residence, and family size—on a student's academic performance.

**Data Information**

The data dictionary is provided by the website:

1. school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2. sex - student's sex (binary: "F" - female or "M" - male)
3. age - student's age (numeric: from 15 to 22)
4. address\_type - student's home address type (binary: "Urban" or "Rural")
5. family\_size - family size (binary: "Less or equal to 3" or "Greater than 3")
6. parent\_status - parent's cohabitation status (binary: "Living together" or "Apart")
7. mother\_education - mother's education (ordinal: "none", "primary education (4th grade)", "5th to 9th grade", "secondary education" or "higher education")
8. father\_education - father's education (ordinal: "none", "primary education (4th grade)", "5th to 9th grade", "secondary education" or "higher education")
9. mother\_job - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other")
10. father\_job - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at\_home" or "other")
11. reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12. guardian - student's guardian (nominal: "mother", "father" or "other")
13. travel\_time - home to school travel time (ordinal: "<15 min.", "15 to 30 min.", "30 min. to 1 hour", or 4 - ">1 hour")
14. study\_time - weekly study time (ordinal: 1 - "<2 hours", "2 to 5 hours", "5 to 10 hours", or ">10 hours")
15. class\_failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. school\_support - extra educational support (binary: yes or no)
17. family\_support - family educational support (binary: yes or no)
18. extra\_paid\_classes - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher\_ed - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic\_relationship - with a romantic relationship (binary: yes or no)
24. family\_relationship - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. free\_time - free time after school (numeric: from 1 - very low to 5 - very high)
26. social - going out with friends (numeric: from 1 - very low to 5 - very high)
27. weekday\_alcohol - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. weekend\_alcohol - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)
31. grade\_1 - first period grade (numeric: from 0 to 20)
32. grade\_2 - second period grade (numeric: from 0 to 20)
33. final\_grade - final grade (numeric: from 0 to 20, output target)

The dataset was sourced from the "High School Student Performance & Demographics" post on Kaggle. It encompasses data from 395 students distributed with 187 males and 208 females across 33 distinct categories.

**Methods**

The exploration of this dataset will kick off with a comprehensive heatmap showcasing the correlations among the numerical variables. This map offers an overarching view of how each element within the dataset relates to one another.

Subsequently, a series of boxplots will be employed to visualize the impact of specific factors such as gender and age on students' final grades. These visuals aim to elucidate where students tend to encounter challenges academically, especially delineating how gender or age may influence final grades. Additionally, boxplots will highlight the relationship between the number of class failures and final grades, pinpointing where students often struggle, and a countplot will shed light on gender-based failure rates.

Moreover, a dive into the influence of external environments on student performance will follow suit. Similar visualizations will unravel how these external factors such as parent's cohabitation status, where they live, and the size of their family impact academic outcomes.

To fortify these findings, a comprehensive set of plots dedicated to final grades will delve deeper into the effect of study time on students' performance, aiming to ascertain whether study habits supersede other elements in influencing grades. This analysis will be stratified across three distinct categories: overall performance, students within the top 75%, and those falling below the average.

**Analysis**

A colorful squares with white text

Description automatically generated with medium confidence

The heatmap analysis reveals a rather weak correlation between a student's age and their final grade, indicated by an efficiency of -0.16. Interestingly, this negative correlation implies that as a student ages, their final grade tends to decrease. However, the data doesn't distinctly pinpoint the underlying reasons for this decline in grades as students grow older. It hints at a trend where older students tend to achieve lower final grades, but the precise factors influencing this phenomenon remain unknown.

A group of colored boxes

Description automatically generated with medium confidence

From the analysis of the boxplots, it's evident that males tend to exhibit slightly better performance compared to their female counterparts. The 50th percentile (median) for males falls slightly above the average grade threshold, whereas for females, the median grade aligns exactly with the average threshold. Additionally, the range between the maximum and minimum grades for males is higher than that for females.

Shifting focus to the number of class failures, it's evident that individuals with final grades below the average threshold tend to experience more than two class failures on average. This observation is quite self-explanatory as lower grades often coincide with a higher number of class failures.

Regarding age groups, students at the age of 19 showcase the poorest performance across all age brackets. Conversely, students at the age of 20 exhibit exceptional academic performance, with over 75% of them achieving grades above the average threshold. This highlights a distinct contrast between these age groups in terms of academic achievement.

A graph with numbers and bars

Description automatically generated with medium confidence

The examination of the boxplots reveals a disparity between the academic performance of males and females. It appears that a larger proportion of males achieve grades exceeding the 75th percentile compared to females. This suggests that gender might indeed play a role in determining grades.

Moreover, the analysis indicates a general trend where student grades decline with age. However, this observation comes with a caveat as the dataset's representation of students at the age of 20 is notably smaller (only encompassing 5 students) compared to those at the age of 19 and younger. Therefore, while it suggests a decline in grades as students get older, the conclusion might be influenced by the dataset's imbalanced distribution across different age groups.

A close-up of a graph

Description automatically generated

The data present a varied distribution concerning both age and gender among students performing below the average threshold. Notably, male students tend to receive grades at 0 more frequently than their female counterparts.

However, upon thorough observation, it becomes evident that using age and gender as indicators to gauge a student's performance might not be reliable. The correlation between these factors and academic performance is notably weak, lacking substantial evidence to conclusively support the notion that age and gender are determining factors influencing a student's performance.

A diagram of a group of boxes

Description automatically generated with medium confidence

A graph with blue and orange bars

Description automatically generated

A graph of different colored bars

Description automatically generated with medium confidence

Indeed, the consistency observed across all three boxplots regarding parents' cohabitation status and living situations hints at a potentially stronger correlation with a student's performance compared to age and gender. However, it's important to note that these indicators are minorities within the testing set.

To support this hypothesis, a more exhaustive analysis is crucial. Conducting deeper investigations and rigorous tests specifically focused on these variables against academic performance will be essential. Given their minority representation in the dataset, a thorough examination of parents' cohabitation status and living conditions in relation to student performance is necessary to validate and reinforce the idea that these factors significantly influence academic outcomes.

**Conclusion**

The data analysis uncovered diverse distributions concerning age and gender among students performing below average. Male students notably received more grades at 0 compared to females. However, relying solely on age and gender as performance indicators appeared unreliable, lacking substantial evidence to decisively influence a student's academic success.

Contrarily, parents' cohabitation status and living conditions displayed a consistent trend across boxplots, hinting at a potentially stronger correlation with student performance than age and gender. These factors, although minorities in the dataset, showed promise. A more exhaustive analysis focused on these variables is imperative to substantiate this hypothesis. Deep investigations and rigorous tests specifically tailored to parents' cohabitation status, living environment, and family size vis-à-vis student performance are crucial.

Overall, age and gender in isolation did not conclusively predict student performance. However, exploring environmental factors—parental cohabitation, living situations, and family size—holds potential significance. Further comprehensive analysis is needed to affirm their impact on academic outcomes. This underscores the need for nuanced examinations beyond age and gender to understand the intricate interplay between growing environments and student performance.

**Citation**

Myrick, Dillon. “High School Student Performance & Demographics.” *Kaggle*, 10 Nov. 2023, [www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics](http://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics).