Andy Ma, Haoliang (Jack) Hu, Truc Ly, Shawn Ha, Vi Tran
BAX 372/MIS 373 - Prof. Deepayan Chakrabarti
April 12th, 2023

Alcohol Consumption's Effects on Student Grades

Intro

Throughout much of human history, alcohol has been a part of people's lives. With so many forms of alcohol, there's a drink for every occasion, and of demographics that drink, college students are within the majority. According to the national survey on drug use and health in 2021, nearly half of all full-time college students in the U.S. drank alcohol within just a month prior to taking the survey. With so many students partaking in alcohol, a question came to mind:

What effect, if any, does alcohol have on the academic performance of students?

Rather than being a domestic issue, alcohol is consumed by students all across the world, and in some nations, people can legally drink as early as the age of 16. As such, to discover the link between students' grades and alcohol consumption, we've identified and run regressions through two datasets gathered from secondary schools in Portugal: One related to the performance of students in a quantitative course and the other related to performance in a qualitative course.

Due to American college culture being so closely associated with alcohol consumption, we hope to inform students of the potential risks associated with frequent drinking to their academic performance.

Methodology

Having looked through the datasets, we determined that the best way to utilize them was by combining the two. They each had the exact same attributes for us to work with, and so all we needed to do was add a column signifying which class, "portuguese" or "math", the row was affiliated with.

Our first attempt at working with the data was to visualize it; we needed to see the differences between the two datasets. To assist, we created two separate datasets detailing the final grades for the students taking Portuguese and the students taking math. We continued by creating several sets that would allow us to visualize the effects of alcohol consumption.

However, before assuming alcohol is the biggest factor in students' grades, we had to consider other factors that could contribute to academic performance. As such, we created various other bar graphs to reflect grades by other attributes, as our dataset has many to be worth with.

Our final step was running regressions, of which we did multiple to discover what attributes would be statistically significant for us to use. We created a column called “Final” for us to use; This column takes the final semester grade of the students and turns it into a binary value, with 0 being students that had a grade of 0-9 in “G3” and 1 being students with a grade of 10-20. We continued by selecting features to use in classification and finished by utilizing K-Nearest Neighbors on the merged dataset.

Visualizations/Trends

Conducting exploratory data analysis using the pandas and matplotlib libraries was a primary focus of the project. As such, we created numerous plots to visualize the attributes of interest in relation to course grades in both the math and Portuguese class subsets. We analyze the following plots within the context of descriptive statistics:

The first set of plots (see Figure 1 in the appendix) displays the distribution of “G3” or the final grade of students enrolled in both the math and Portuguese classes. The average grade achieved by students in the Portuguese section is about one point higher than that of the math subset, which can be explained by inherent differences in the difficulty of both courses. Portuguese classes may just seem more intuitive to students whose primary language is Portuguese, as well.

The second set of plots (see Figure 2) depicts the total count of students' drinking levels grouped by their age. The majority of students aged 16-19 cited low levels of alcohol consumption, which follows the general trend of similarly aged students in secondary education. A few outliers drink a large amount.

The third visualization (see Figure 3) shows the average grade achieved by students in math and Portuguese classes based on their daily alcohol consumption. Although there is no clear relationship between grades and alcohol consumption in the math portion, the Portuguese portion appears to indicate a slightly downward trend of grades by daily consumption of alcohol. Due to our dataset containing over 30 attributes, we decided it would be rash to conclude that a student's grades are affected solely by alcohol. As such, we continued to visualize data based on features that could also contribute to academic performance.

The fourth visualization (see Figure 4) displays the average grade achieved in both math and Portuguese sections by whether students' parents were "A," living apart, or "T," living together. Across the board, students in the Portuguese class seemed to perform better than their math counterparts when comparing average grades by both "A" and "T."

The fifth visualization (see Figure 5) is a quadruple set of bar graphs that displays students' grades based on the education level of their parents. Based on the data, we can infer that, on average, the higher a parent's education level, the better grades the corresponding student will achieve. However, there aren't any noticeable differences in performance between if that parent is a father or a mother.

The sixth visualization (see Figure 6) is a bar graph that gives a count of students who passed or failed based on whether they received family support. The proportions of students that failed, regardless of having family support or not, are practically identical. In addition, it is important to note from our analysis that the proportion of students who failed and had no family support is higher than that of the students who passed math or Portuguese and received no family support. This suggests that family support is crucial for a student's success.

Regression/Classification

For regression, we created models to predict two dependent variables of interest: “G3” and “studytime.” We used the subset of students in the Portuguese class to drive our regression models because it contains more records than the subset of math students, which could cater to more reliable results. The intuition for choosing some of the regressors for our two models for predicting the final grade and study time, such as “failures” and “higher,” were based on forward feature selection.

For the regression model with “G3” as the dependent variable, several attributes had significant p-values. Initially, we explored the possibility of including the “Final” variable within the regression model as a predictor because it dramatically increased the R-squared value, but we realized it does not make sense since it is highly correlated with “G3”.

For the regression model with “studytime” as the dependent variable, fewer attributes had significant p-values compared to our first. Although the R-squared value is also lower, it is interesting to note that the “reason” attribute had one significant subcategory: reputation. This suggests that students who choose to attend a school because of its reputation might be more inclined to increase their study time and succeed in class.

Other than using regression, we can also predict a student's performance by using classification. We decided to perform classification using K-Nearest Neighbor (KNN). KNN is a simple and effective algorithm used for classification and regression tasks. The main idea behind KNN is to classify a new data point by finding the K nearest data points in the training set and

assigning the class label of the majority of the K neighbors to the new data point. The distance between data points can be calculated using a variety of methods, such as Euclidean distance or Manhattan distance. We decided to use Euclidean distance.

Before we fitted our KNN model, we performed forward feature selection to determine the features that are most predictive of whether a student will pass or fail ("Final"), measured by improvement in R^2. Our result shows the top five features are the number of failures ("failures"), which school the student is from ("school"), if the student plans to pursue higher education ("higher"), weekly study time ("studytime"), and if the student has extra educational support ("schoolsup").

After fitting our KNN model using different combinations of the selected features, we found that a good combination is with the number of failures ("failures") and weekly study time ("studytime"). With a loop, we determined the optimal number of neighbors (K) is 3, which resulted in a test accuracy of around 80 percent.

By plotting the predictors we used for our KNN model (Figure 7), we can see there is a positive association between the number of failures ("failures") and having a failing grade ("Final") and a negative association between weekly study time ("studytime") and having a failing grade ("Final").

Conclusion

While there are some variables that can predict a student's academic success, such as alcohol consumption, the confidence is low. This may, in part, be due to the small sample size, which becomes more prevalent when considering the low percentage of students who frequently drink alcohol.

It's important to note that when this data was gathered, students were asked to fill out a survey which was then connected to grade information from the school, requiring personal information. It's expected that students, when data is not anonymized, may be less likely to answer questions such as those about alcohol consumption (the drinking age in Portugal is 18).

These results are still important for connecting grades to student behavior, whereas other studies often use GPA, which won't explain differences among courses, or standardized tests, which differ greatly from the traditional school experience. However, since the results are not strongly conclusive, we recommend further research using similar philosophy at a larger scale and possibly anonymized information, if possible.

Citations

Balsa, A. I., Giuliano, L. M., & French, M. T. (2011). The effects of alcohol use on academic achievement in high school. *Economics of education review*, *30*(1), 1–15. https://doi.org/10.1016/j.econedurev.2010.06.015

Betts, J. R., Zau, A., & Rice, L. (2003). *Determinants of student achievement: New evidence from San Diego*. Public Policy Institute of California.

U.S. Department of Health and Human Services. (2023, March). *College drinking*. National Institute on Alcohol Abuse and Alcoholism. Retrieved April 11, 2023, from https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/college-drinking

Appendix

Figure 1

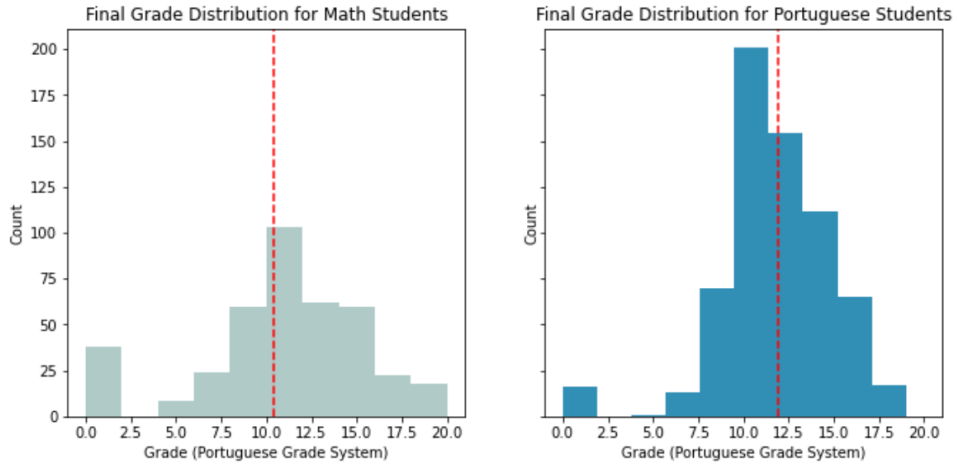


Figure 2

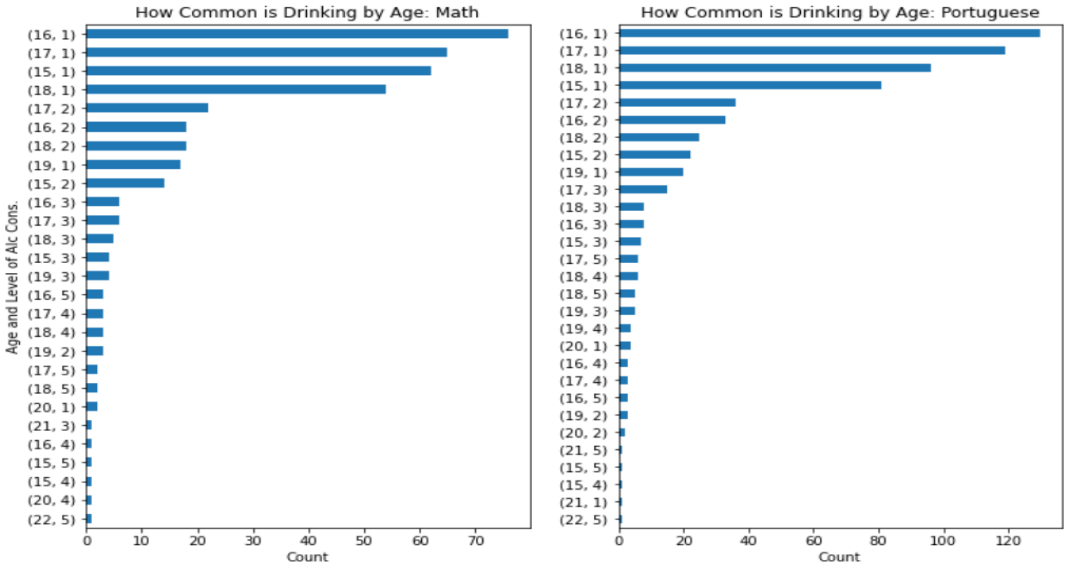


Figure 3

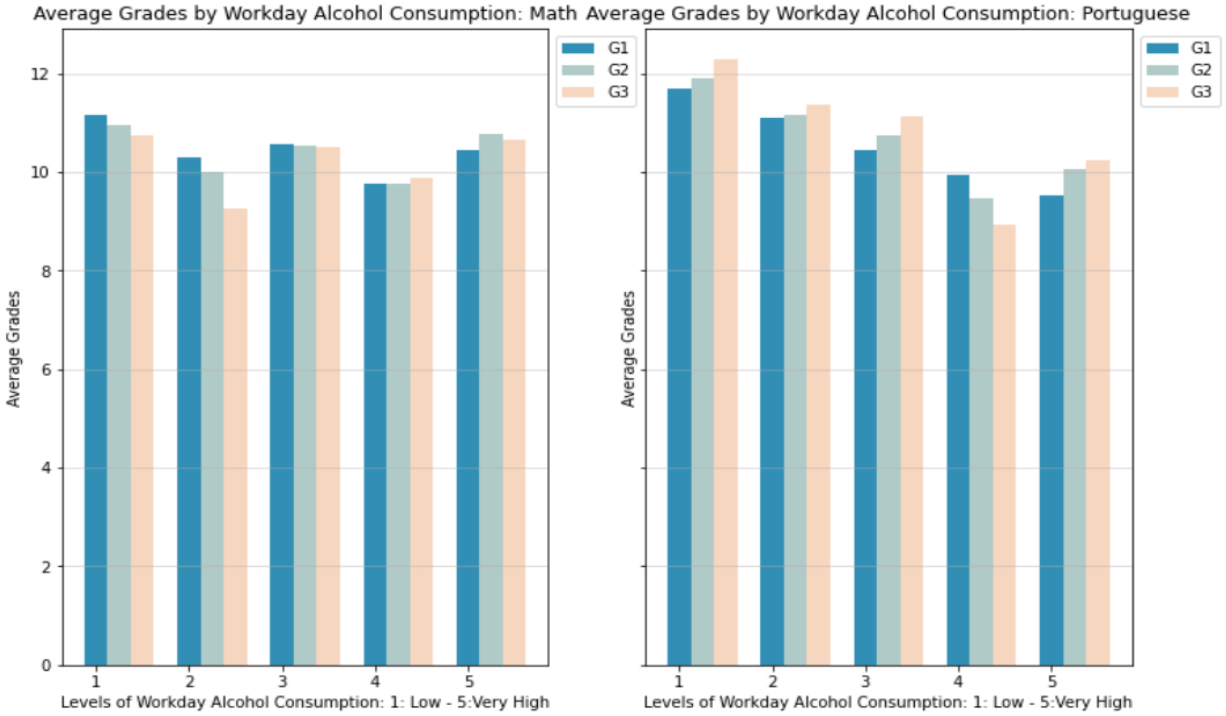


Figure 4

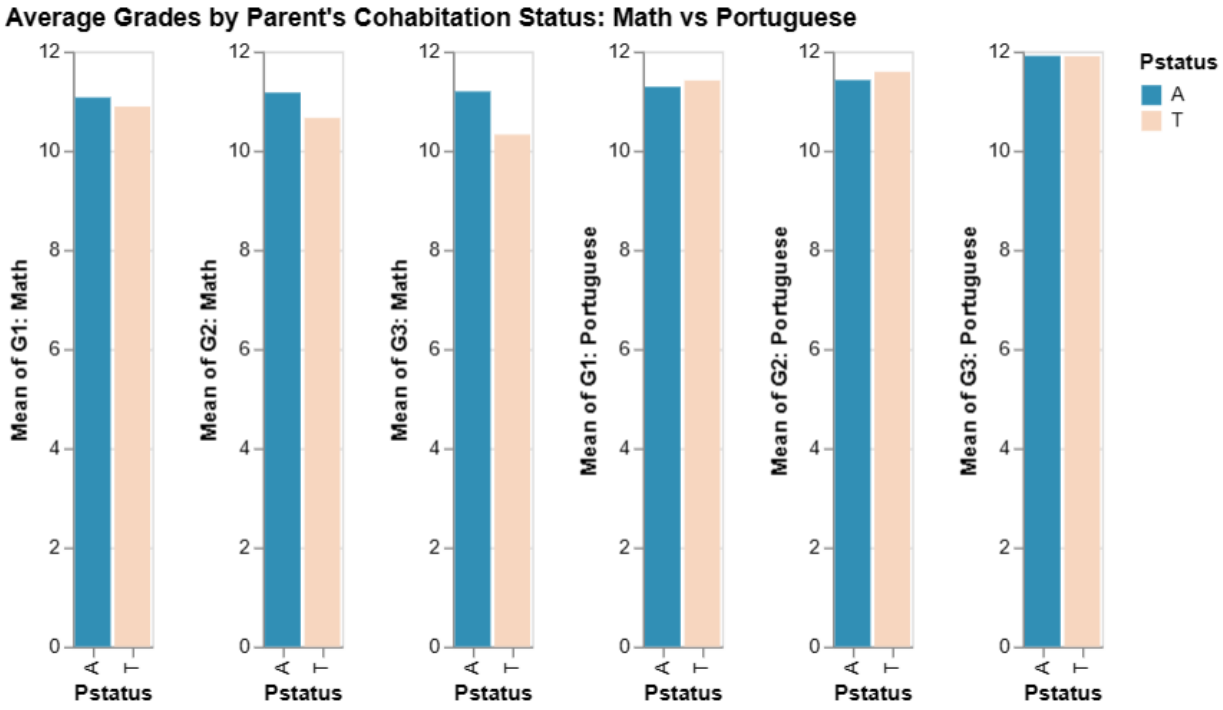


Figure 5

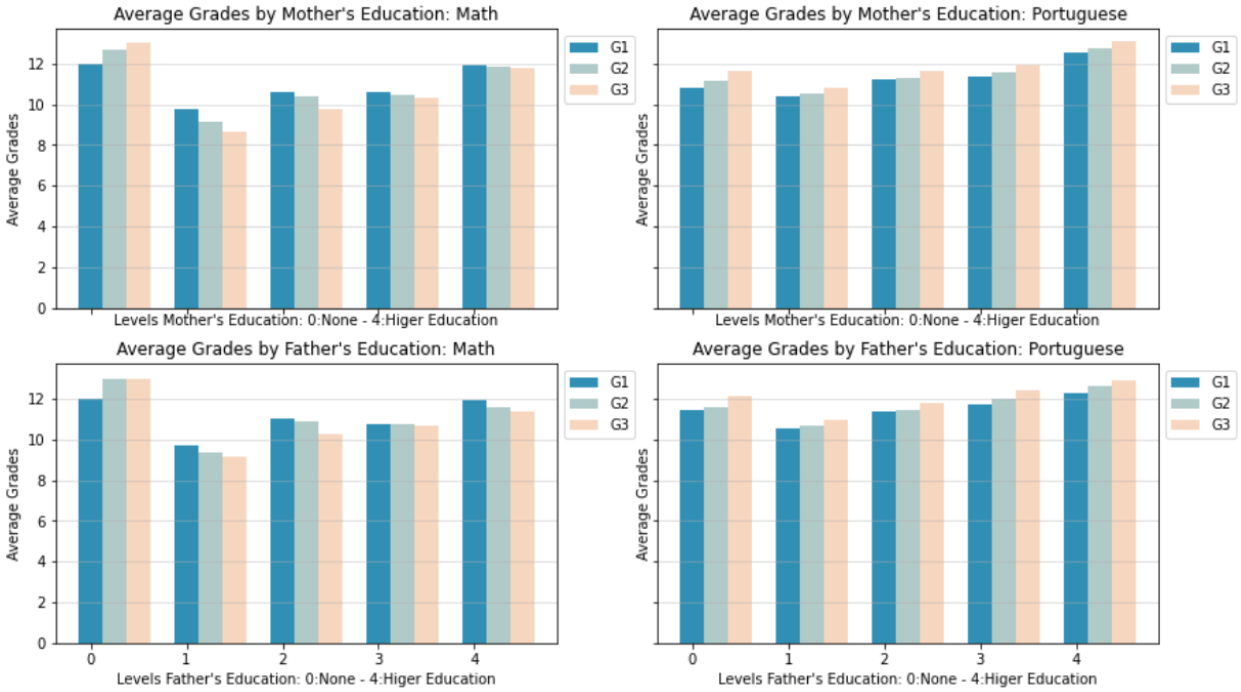


Figure 6

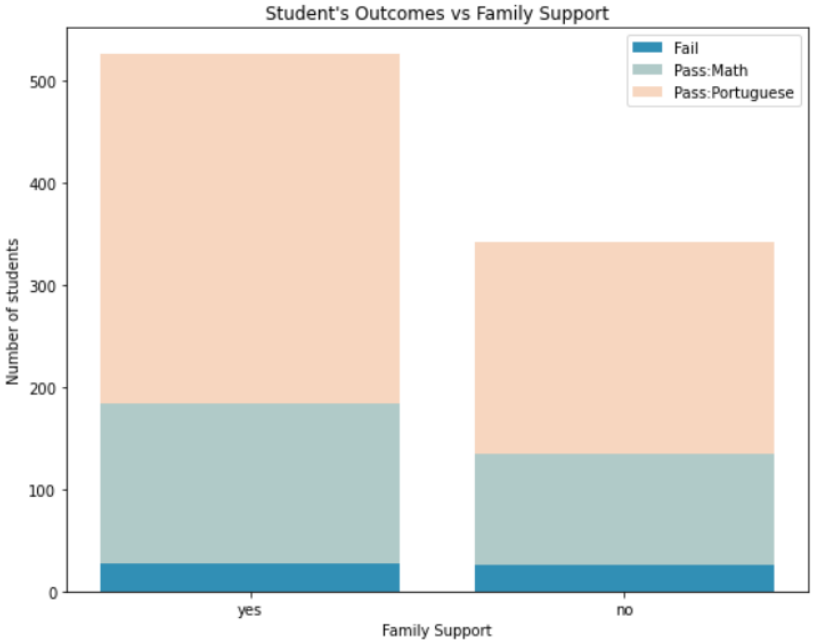
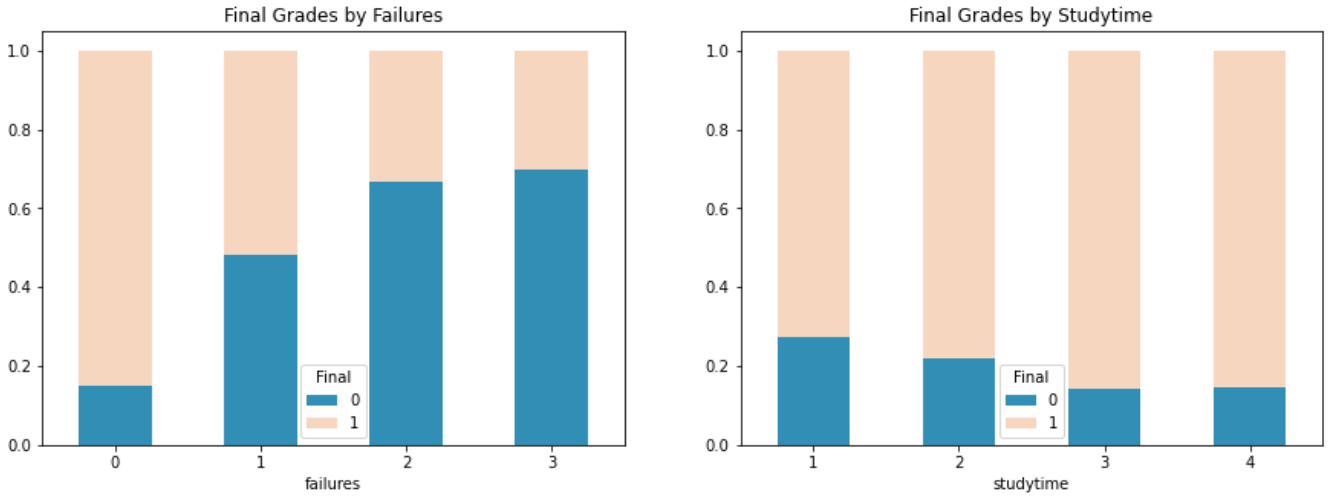


Figure 7



Link to Dataset: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption

Dataset Attributes:

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 - student's home address type (binary: 'U' - urban or 'R' - rural)

5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. Fjob - 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. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - 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 - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - 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)

These grades are related with the course subject, Math or Portuguese:

1. G1 - first period grade (numeric: from 0 to 20)
2. G2 - second period grade (numeric: from 0 to 20)
3. G3 - final grade (numeric: from 0 to 20, output target)