Maths Grades Analysis

```
##                         Name        NetID
## 1 Annabelle Griffith-Topps griffithtopp
## 2               Ana Klabjan     aklabjan
## 3             Jacob Larget      jlarget
## 4                 Bryan Li       bli378
## 5               Jake White     jtwhite4
## 6             Zi Hern Wong       zwong4
```


Abstract

Our data set was compiled in 2008 by two Portuguese researchers, and it describes the math grades of 395 high-school students at the end of each trimester, as well as some student demographic information. We used this data set to answer our statistical question: What variables have an effect on whether or not a student drops out?

We found that the variables that affected if a student dropped out with the highest statistical significance (alpha = 0.001) were: first trimester grades, receiving school support, and the number of absences. All were negatively correlated with dropping out. Next (alpha = 0.01), we found that weekend drinking and paying for tutoring negatively correlated with dropping out, and dating and previous failures positively correlated. Finally (alpha = 0.05), we found that one school performed better than the other.

Data Set Introduction

This data is important because it describes Portugal, which has a high failure rate in core classes compared to other European countries. In particular, the proportion of students leaving class early in Portugal is 40%, whereas the average in the rest of the European Union is 15%. This statistic is not an anomaly, so if we can identify the students who are most likely to drop out after first trimester, then teachers and school systems can proactively provide assistance to those students.

The data comes Paulo Cortez and Alice Silva, who work for the department of Information Systems/Algoritmi R&D Centre at the University of Minho. They created it by combining two data sets, one with student grades and another with student questionnaire responses.

Data Set Variable Descriptions

The following are the variables available in our data set. We have bolded those we might expect to have predictive power as explanatory variables.

Note: we have altered the provided grades by making them into percentages instead of a point system up to 20. We found this more intuitive to comprehend student achievement.

school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) We can train our model on one of the schools to use it to test the squared residuals on the other school. This will allow us to see whether our dataset will be appropriate for all schools or whether one school is just harder than the other.

sex - student's sex (binary: 'F' - female or 'M' - male) We know that girls tend to get higher grades than boys in primary school, so there could potentially be a correlation between sex and grades in secondary school.

age - student's age (numeric: from 15 to 22) We are not sure if age will matter but are thinking of including it anyway in case it is helpful for step-wise selection.

address - student's home address type (binary: 'U' - urban or 'R' - rural) Students might face different challenges living in the city (for instance, higher crime, louder environments, more places near their houses to do work), or living in a rural environment (further away from places like parks, grocery stores, libraries, after school jobs/programs)

famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) Smaller families might be able to help their children more individually and could help children understand concepts that they might struggle with.

Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart) Divorce can be stressful on children and have negative effects.

Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) Parents with higher education could influence their student's performance in school

Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) Parents with higher education could influence their student's performance in school

Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') If a student has a working mother, this could impact child performance. Additionally, if a mother is a teacher, the student might perform better and be discouraged to drop out. If a mother works in health care, she likely has a college education, so this might inspire the student to stick with their education and perform better. If a child has a stay-at-home mom, this could have a positive correlation with performance since the mother likely has more of a role in their child's education than a working mother. There might also be a correlation between student performance and working mothers.

Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') Society expects fathers to be the breadwinner for the family. A stay-at-home father might impact their student's performance positively and could be able to support their child with schoolwork and encourage them to stay in school. If a father is a teacher, this might encourage the student to do better in school and stay in school. If a father works in health care, this could encourage students to do well and stay in school.

reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') A student who chose a school because of the school's reputation or course preference likely wants to perform well and will stay in school longer. If the only reason that a student chose a school is because it is close to home, they might not feel as connected to the school and not enjoy their studies and perform well in school.

guardian - student's guardian (nominal: 'mother', 'father' or 'other') If a student's guardian is not their mother or father, this likely means that the student has grown up in a different dynamic than other students and could impact their performance. This variable doesn't provide us with much relevant information- a better factor, in our opinion, is seeing if a student's parents are married or divorced.

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) Longer travel times have general correlation with earlier bedtimes, which then might affect study time in the evening.

studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) Longer study times are generally correlated with higher grades because students are able to comprehend more of the material. Conversely, a student might overstudy and become less confident with the material or waste time when they have mastered the material.

failures - number of past class failures (numeric: n if 1<=n<3, else 4) If a student has failed in the past, they may be more likely to fail in the future

schoolsup - extra educational support (binary: yes or no) This is not clear enough for us. Does it mean a college tutoring service, scholarship or free and reduced lunch? If this variable was more clear, it could be more useful. For instance, if this variable indicates that a student is in a gifted and talented program, they likely perform well and won't drop out. If a student is in special education, this might have a negative correlation on grades and the student might be more likely to drop out. If this means that a student goes to peer tutoring or asks teachers for help, that might have a positive effect. Because this information is not accessible, we will not be using this variable.

famsup - family educational support (binary: yes or no) This binary is also unclear for similar reasons. This could mean that a student has a private tutor financed by the family, or it could mean that a student might ask their parents for help on a math problem for example. This is a very wide range, and the paid variable would make more sense and could be used effectively.

paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) Extra assistance comprehending the material may impact a student's performance in the course. If a student has access to a private tutor or a summer course, this could positively impact student performance.

activities - extracurricular activities (binary: yes or no) Students who are actively involved in school activities might tend to perform better than students who do not participate. They might benefit from a sense of community that could help them perform better. Conversely, this leaves less time for students to study so it could have a negative impact on performance.

nursery - attended nursery school (binary: yes or no) The years a student is in nursery are very important for development and this could have a lasting impact on student performance.

higher - wants to take higher education (binary: yes or no) Students who want to enroll in higher education generally try harder in school to get into a better college.

internet - Internet access at home (binary: yes or no) Lack of internet might have a negative impact on student performance. The internet can often be a useful resource for students to use (for instance, Khan Academy or Crash Course might be able to help students understand concepts more deeply).

romantic - with a romantic relationship (binary: yes or no) Students in a romantic relationship might have less time to spend on homework and studying. Additionally, they might tend to support each other academically.

famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent) Students who have a supportive relationship with their family might tend to do better, or if they have an unsupportive relationship, they might seek academic validation or want to do well to be able to go to a college with a good scholarship.

freetime - free time after school (numeric: from 1 - very low to 5 - very high) More free time may impact school results either positively or negatively because students may spend more of their free time studying or hanging out with friends

goout - going out with friends (numeric: from 1 - very low to 5 - very high) Students might prioritize spending time with friends rather than studying and doing homework and this could lead to a negative impact on grades.

Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) Students who drink alcohol during the week might struggle with addiction and this could lead to negative impact on grades.

Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) Creation of alcohol consumption habits over the weekend are known to work its way into the workweek, which can affect school performance.

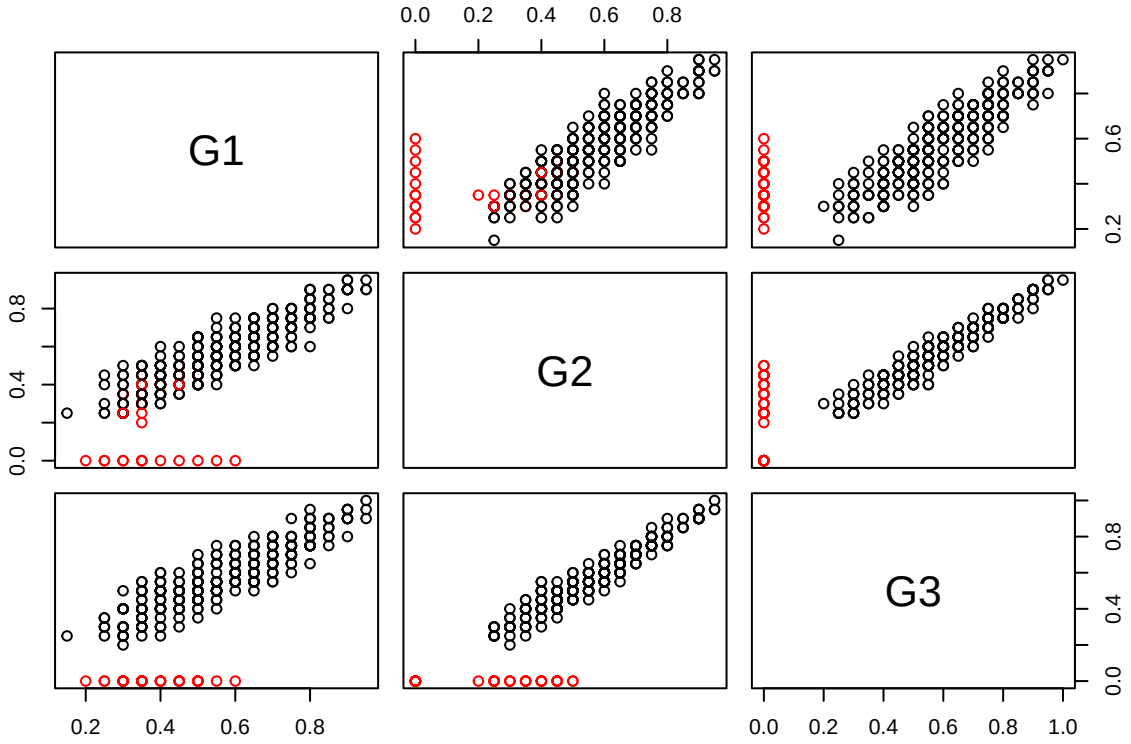
health - current health status (numeric: from 1 - very bad to 5 - very good) If a student has poor health, they might stay home from school more and have to miss class for appointments etc, this could lead to lower grades in theory.

absences - number of school absences (numeric: from 0 to 93) The number of absences can correlate with student performance.

G1 - the first trimester grades of the students. While this is likely correlated with drop out rate, it is not sufficient to determine it outright.

Discussion of Statistical Question

We started our research with some exploratory data analysis. As the grades for a given student are linked throughout the year, we thought it would be important to check that the grades do, in fact, have a positive correlation. However, when we ran a pairs plot on G1, G2 and G3, we noticed there was a distinct selection of data points that did not fit our expected outcome and instead formed perpendicular streaks. We were able to identify these streaks with the following coloring, where the students who completed the course in black and those who dropped out in red(meaning the G3 column was zero). While the distinction is apparent by the end of the year, our statistical question hopes to answer this question without all of the grade data so schools can proactively identify students who might at risk of dropping out.

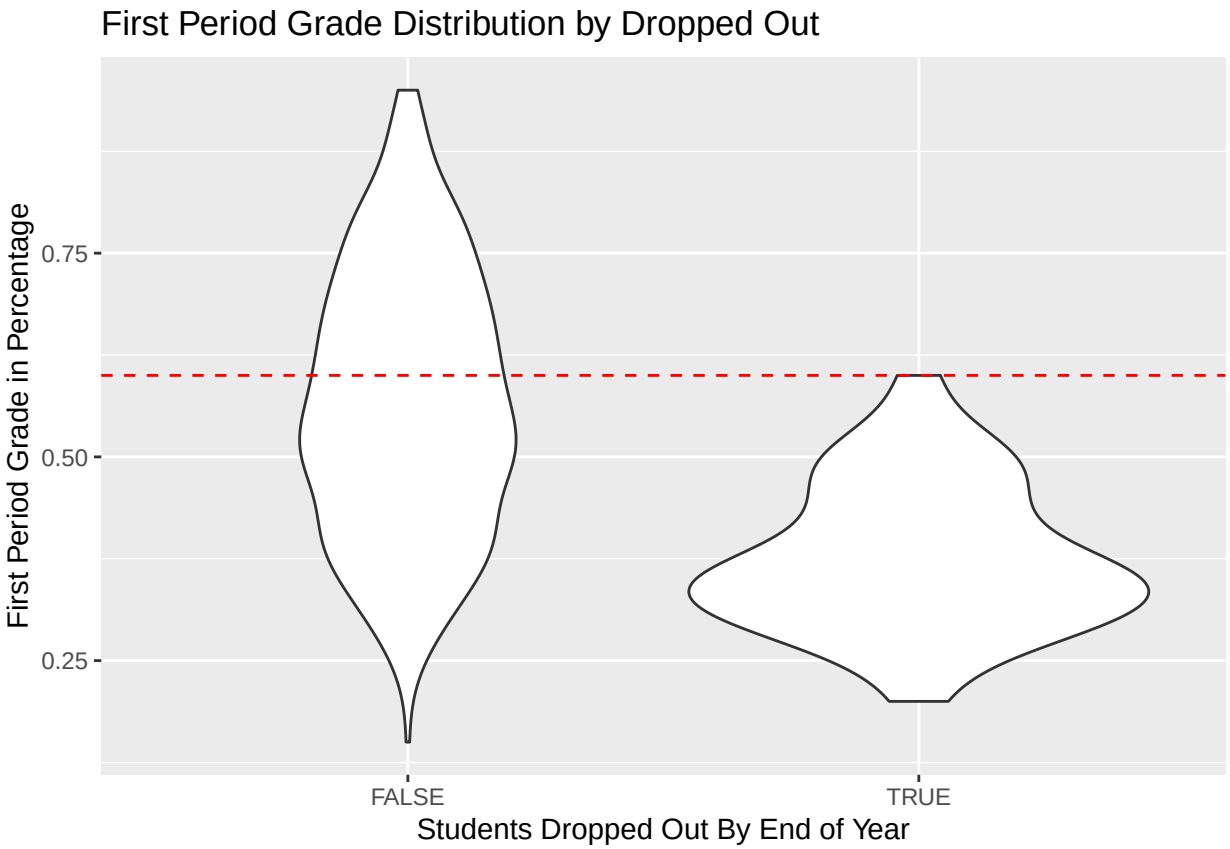


Running a frequency analysis on this coloring, we see that there was about a 10% drop out rate among our 395 students, which is very high. This is also good news for answering our statistical question because it means we have a meaningful enough sample size for each of our outcomes.

```
##   completedCourse count   freq
## 1           FALSE    38  9.62%
## 2            TRUE   357 90.38%
```

Analytical Methods

Compared to students who did not drop the class, we found a noticeable difference between first period grades. While only the highest scoring students among the dropped category were able to score 60%, the mean of all student's scores during the first period was 60%. Thus, we can conclude that students that dropped the class can be predicted to do considerably worse than those who did not drop the class. This seems pretty obvious, but perhaps there are other factors that could lead to this outcome.



In order to test whether G1 is the sole factor in determining drop out potential, we created a model with `dropped` as the response variable and G1 as the explanatory variable.

```
##       feature   estimate       pvalue significance
## 1 (Intercept)   2.431008 7.721083e-04          ***
## 2          G1 -10.238157 6.862660e-09          ***
```

We see that G1 is statistically significant at predicting if a student will drop out or not. However, our analysis of the data set shows us that the average drop out grade seems to be around 35% but when when putting that into our model we get that our predicted percentage of dropping out is 24%. We expected this value to be higher.

To see if adding other factors can be included for a better prediction, we created a model with `dropped` as response and everything else in our data set as the explanatory variables to see how this would affect the drop rate. At this point, we used step-wise selection via AIC. We found this method sufficient as it reduces the impact of over-fitting with its penalty of the number of variables and log-likelihood. With a reduction of AIC from 200.13 in our model with just G1 to 47.932 in our model with other factors, we recognize it as a meaningful improvement of our model, and through its significant values determine our critical factors.

```
##         feature    estimate       pvalue significance
## 1   (Intercept)  0.13130987 5.318468e-01
## 2            G1 -0.60476950 7.558588e-12          ***
## 3      schoolMS -0.11258688 1.231642e-02            *
## 4           age  0.02154715 6.915961e-02
## 5          Fedu  0.01738328 1.623677e-01
## 6    traveltime  0.02761489 1.522760e-01
## 7      failures  0.05352283 6.557871e-03           **
## 8  schoolsupyes -0.14558478 4.016022e-04          ***
## 9       paidyes -0.06950777 8.239160e-03           **
## 10 romanticyes  0.07854902 4.885342e-03           **
## 11         Walc -0.03099238 2.865658e-03           **
## 12     absences -0.01019219 2.769945e-09          ***
```

Summary of Findings and Future Research Potential

We discovered that the variables with the largest impact were variables that rely on grades. This makes logical sense, because the grades are recorded throughout the semester. The grades will be relatively consistent- a student could not go from failing to a perfect score. This also means that if a student has failures, this will impact grades. Our analysis proved this, which helps us show that our analysis makes sense. We discovered if a student had failed a class, they were more likely to drop out. If a student has higher grades, they were less likely to fail out. A surprising discovery an impactful factor in if a student is in a relationship, they are at high risk for dropping out. Maybe there is some merit in parents having rules about not dating in high school! Two other factors that impact if a student will drop out is if the student is receiving extra support from the school, and if a student takes extra paid classes within the course subject. Thankfully, these variables showed that if a student is receiving help, they are less likely to drop out.

If we had more time, resources or funding, there are a few ways we would expand this project and answer some further questions. First, an issue we found was because this sample was not very large and very specific. This means we could not draw accurate conclusions to apply our findings to other groups. Another issue is that some of the measures were too subjective. For instance, a few variables were measured by students answering a range from 1-5 about some of their habits instead of asking for more consistent measures across students. Another aspect to consider if the study was expanded is ensuring the curriculum is similar. For instance, if this study expanded to a variety of countries in Europe, then measuring grades in Portuguese might not be the best idea. We came up with a few questions that could be investigated further: How would our findings change if the study was expanded upon? Would different areas of the world have different results? If we could identify students that are at heightened risk for dropping out and provide them extra resources, would that be beneficial? Would that be ethical? What factors should educators be aware of when it comes to preventing students from dropping out? What are the most cost effective ways to prevent students from dropping out? What behaviors should parents be concerned about with their child that could increase their risk of dropping out?

We were able to come to a few conclusions, but not many based on our analysis. We tried a variety of methods to investigate our initial statistical question: what variables have an effect on whether a student will drop out or not? All in all, we conducted a through analysis, but could find some more impactful solutions if we had more resources and data.