Maths Grades Analysis

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


Abstract

Our data set describes 395 students, their math grades for a given semester, recorded on three occasions, as well as generic and consistent demographic information. With this data, we created models to answer our statistical question: (1) What variables have an effect on weather a student will drop out or not?

To_do: conclusion

Data Set Introduction

This data is important as in the case of our data Portugal they have a high failure rate in core classes compared to other European countries. If we can identify the students who are most likely to drop out after first quarter then teachers and the school systems can provide additional resources to them.

The data comes from schools in Portugal through school records as well as questionnaires. This data was collected by Paulo Cortez and Alice Silva who work for the department of Information Systems/Algoritmi R&D Centre at the University of Minho. It started off as two datasets, one with student grades and another with the student questionnaire responses. It was then combined before doing analysis. This study is significant as specifically in Portugal, the proportion of students leaving class early is 40% whereas the average in the rest of the European Union is 15%. This is just one example of a statistic where Portugal is near the bottom compared to the rest of the European Union. Pinpointing a few of these factors could help officials develop a solution to improve Portugal's educational performance.

The following are the variables available in our dataset. We have bolded those we expect to use as explanatory variables. We are bolding most of the variables because we want to resist the urge to determine the model because it is run. We anticipate being able to use forward and backward stepwise selection.

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 (nytimes.com/2019/02/07/opinion/sunday/girls-school-confidence.html#:~:text=From%20elementary%20schoo so potentially there could 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 stepwise 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') create for each its own coefficient 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.

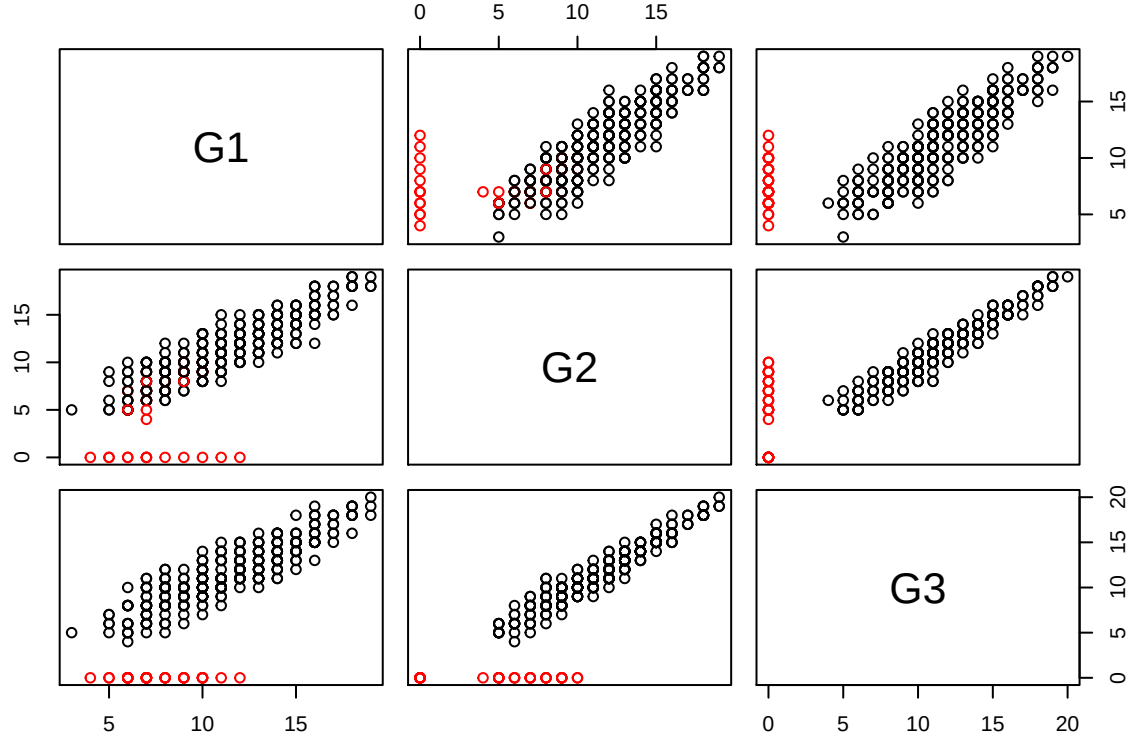
absences - number of school absences (numeric: from 0 to 93) The number of absences can correlate with student performance.

```
##   school sex age address famsize Pstatus Medu Fedu    Mjob    Fjob reason
## 1     GP   F  18       U     GT3       A    4    4 at_home teacher course
## 2     GP   F  17       U     GT3       T    1    1 at_home   other course
## 3     GP   F  15       U     LE3       T    1    1 at_home   other  other
```

```
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
```

Background

We started off our research with basic exploratory data analysis. We knew that the focus of our report would have a question that has a response variable related to grades when modeled. We ran a pairs plot on G1, G2 and G3 when we noticed there was a distinct pattern of several data points in a line that didn't fit our expected outcome: a strong positive linear correlation between G1, G2 and G3. We indicated these data points



in red. The coloring in the plot was determined by if the G3 column in our data set was equal to zero. The red points represent students who dropped out of the class before the end of the year. This trend lead us to the focus of our report: What variables have an effect on weather a student will drop out?

This question is important to ask since if we can identify which students are most likely to drop out then analysis can be done on future students and those students who are predicted to be most likely in danger of dropping out can receive access to additional resources and assistance. This is especially important to our data set when the Portuguese are trying to improve there student failure rate in core classes. Our data set shows that at the time of data collection there was about a 10% drop out rate which is very high.

```
## # A tibble: 2 x 3
##   completedCourse count freq
```

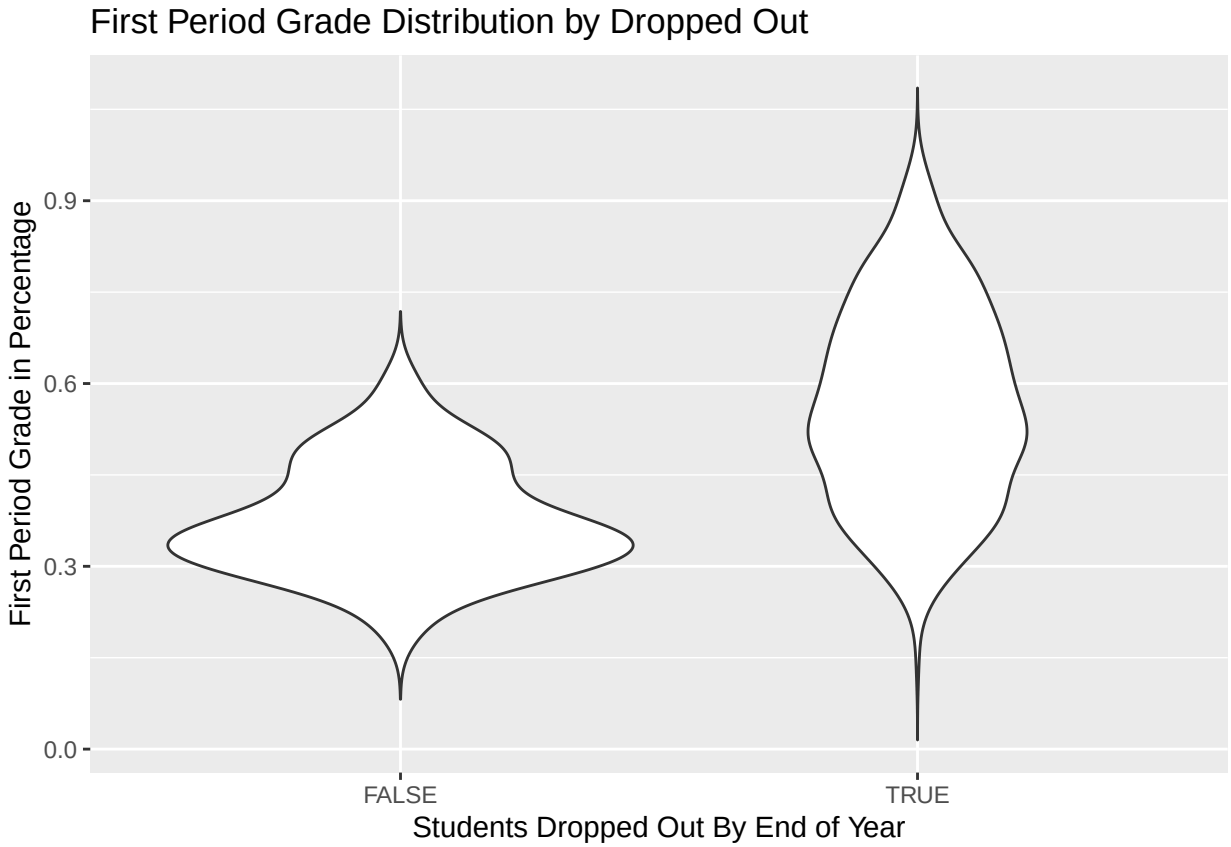
```
## * <lgl>           <int> <formttbl>
## 1 FALSE              38 9.62%
## 2 TRUE              357 90.38%
```

Report

We started off by transforming our data and adding two columns “dropped” and “completedCourse”. “dropped” and “completedCourse” were equivalent in meaning, whether a student ended with a final grade,G3, of zero indicating they dropped out and did not complete the course. G1, G2, and G3 were transformed into percentage values with 20 being the highest possible grade.

```
##   dropped completedCourse   G1   G2  G3 school sex age address famsize Pstatus
## 1       0            TRUE 0.25 0.30 0.3     GP   F  18       U     GT3       A
## 2       0            TRUE 0.25 0.25 0.3     GP   F  17       U     GT3       T
## 3       0            TRUE 0.35 0.40 0.5     GP   F  15       U     LE3       T
##   Medu Fedu    Mjob    Fjob reason guardian traveltime studytime failures
## 1    4    4 at_home teacher course   mother          2         2        0
## 2    1    1 at_home   other course   father          1         2        0
## 3    1    1 at_home   other  other   mother          1         2        3
##   schoolsup famsup paid activities nursery higher internet romantic famrel
## 1       yes     no   no         no     yes    yes       no       no      4
## 2        no    yes   no         no      no    yes      yes       no      5
## 3       yes     no  yes         no     yes    yes      yes       no      4
##   freetime goout Dalc Walc health absences
## 1        3     4    1    1      3        6
## 2        3     3    1    1      3        4
## 3        3     2    2    3      3       10
```

DELETE From these two graphs, we can see that among those who dropped at some point in the class, none of them were able to obtain a score higher than ~60%. Now, how does that relate to students who did not drop?



Compared to students who did not drop the class, we see 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 could there be other factors that lead to this outcome?

Now lets create a model with dropped as the response variable and G1 grade as the explanatory variable.

We see that G1 is statistically significant at predicting if a student will drop out or not. Based on the first graph we see the averaged drop out grade seems to be around 35% when putting that into our model we get that our predicted percentage of dropping out is 24%. I was expecting this value to be higher. Maybe other factors can be included for a better prediction. Also how does the distribution of percentages look for our data?

We created a model with dropped as response and everything else in our datas et as the explanatory variables.

TO-DO: TALK ABOUT WHAT WE DID(DON'T SHOW) AND WHY

This model has health, weekend drinking, past class failures, mother's education, age and school as being statistically significant when it comes to weather a student will drop out or not. This model has a degrees of freedom of 40 and AIC of 144.67 and RSS 64.67.

We improved the model by using stepwise selection according to AIC.

```
##
## Call:
## glm(formula = dropped ~ G1 + sex + age + famsize + Fedu + traveltime +
##     studytime + failures + schoolsup + paid + activities + higher +
##     romantic + famrel + goout + Dalc + Walc + health + absences,
##     family = "binomial", data = drop2)
##
```

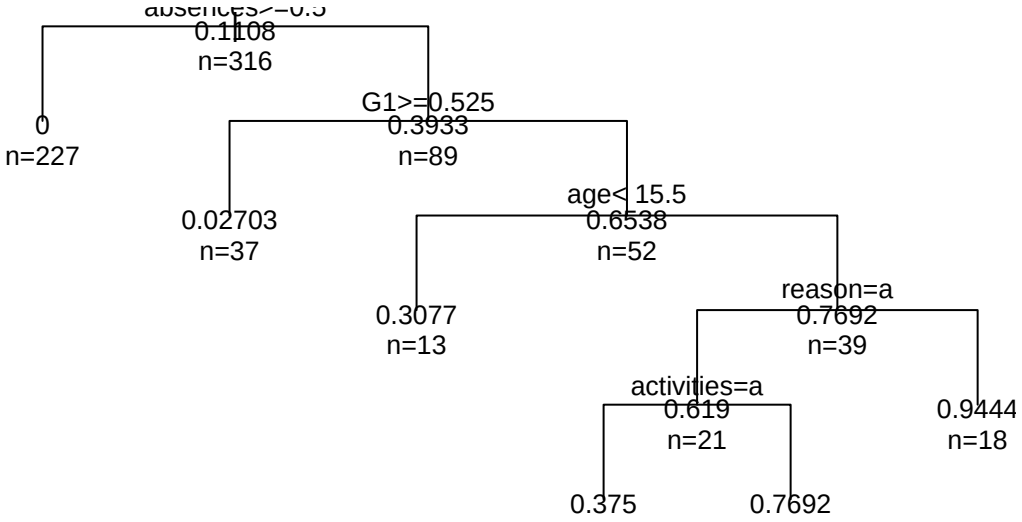
```
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.008581   0.000000   0.000000   0.000000   0.008790
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -15953      99875  -0.160    0.873
## G1              -55232     292194  -0.189    0.850
## sexM              1228      14027   0.088    0.930
## age               1588       9929   0.160    0.873
## famsizeLE3       -4059      27476  -0.148    0.883
## Fedu              2777      15628   0.178    0.859
## traveltime        1639       7845   0.209    0.835
## studytime         1824       7716   0.236    0.813
## failures          2253      12824   0.176    0.861
## schoolsupyes    -13457      72280  -0.186    0.852
## paidyes          -4309      18978  -0.227    0.820
## activitiesyes     3349      22294   0.150    0.881
## higheryes         9733      52986   0.184    0.854
## romanticyes       4058      23571   0.172    0.863
## famrel           -2199      13524  -0.163    0.871
## goout             1421       8464   0.168    0.867
## Dalc              4875      23588   0.207    0.836
## Walc             -2637      14388  -0.183    0.855
## health           -1874      10892  -0.172    0.863
## absences         -4884      28638  -0.171    0.865
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.5016e+02  on 394  degrees of freedom
## Residual deviance: 8.7109e-04  on 375  degrees of freedom
## AIC: 40.001
##
## Number of Fisher Scoring iterations: 25
```

This model has school, age, family size, past class failures, weekend drinking, attending nursery school and romantic partner as being statistically significant weather someone will drop out. This model has a degrees of freedom of 17 and AIC of 110.91 and RSS 110.91.

Based on our current knowledge we are not sure as to which model is better between the two as the first has a lower RSS although not by much so on the other hand the second model has better statistics from the extractAIC with a lower degrees of freedom and generalized Akaike Information Criterion. We are certain that school, age, past class failure and weekend drinking are all important factors in being able to predict if a student will drop out or not as they were marked as significant in both models we created. Health, mother's education, family size, attending nursery school and romantic partner were significant explanatory variables in one model but not the other. At this moment we aren't able to explain why some variable are significant in one model but not in the other. Further testing would need to be performed on this additional variables to determine if they are significant explanatory variables to predicting weather a student will drop out or not.

In conclusion, based on our analysis, all students who dropped out had a first grade of 60% or below with the median being around 30%. In addition to using G1 grade to determine if a student is likely to drop out school, age, past class failure and weekend drinking were found to also be significant in its influence weather a student is predicted to drop out.

Future steps with this question would be to create a decision tree using the 5 significant variables we identified. Could we categorize the students into buckets of likelihood to drop out?




```
## n= 316
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 316 31.1234200 0.11075950
##    2) absences>=0.5 227  0.0000000 0.00000000 *
##    3) absences< 0.5 89 21.2359600 0.39325840
##      6) G1>=0.525 37  0.9729730 0.02702703 *
##      7) G1< 0.525 52 11.7692300 0.65384620
##       14) age< 15.5 13  2.7692310 0.30769230 *
##       15) age>=15.5 39  6.9230770 0.76923080
##         30) reason=course 21  4.9523810 0.61904760
##           60) activities=no 8  1.8750000 0.37500000 *
##           61) activities=yes 13  2.3076920 0.76923080 *
##         31) reason=home,other,reputation 18  0.9444444 0.94444440 *
```

```
## [1] 0.05517349
```

Conclusion

TO-DO: summarize finding, further questions/exploration