

Project Document

Part I: Exploratory Data Analysis

About the Data

Our data was obtained via UCI's Machine Learning Repository. The data is a multivariate set designed to explore student performance tied to various predictors during a collection period from 2005-2006. The data is split into two sets: Mathematics (student-mat.csv) and Portuguese (student-por.csv). These are the two subjects where records of student's attending two public school's from the Alentejo region of Portugal performance (our outcome y) were recorded. Predictor variables include a range of demographic, social, health, and school related attributes.

The data was utilized by a paper published in 2008 titled "[Using data mining to predict secondary school performance](#)". The study's goal was to use BI/DM techniques to build a model that accurately predicted student performance given predictor variables that provided the best accuracy. Below, we will conduct an EDA exploring and cleaning this data set prior to conducting a replication of their study while critiquing their process and adding/removing anything we deem necessary to result in the best models for our given data and prior proposed research goal.

Loading our Libraries

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.4      v readr      2.1.5
v forcats    1.0.0      v stringr    1.5.1
v ggplot2    4.0.0      v tibble     3.2.1
v lubridate  1.9.3      v tidyr      1.3.1
v purrr      1.0.2
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(car)
```

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

```
recode
```

The following object is masked from 'package:purrr':

```
some
```

Loading our Data

```
# loading in both of our data sets, we will title them by their  
# subject  
math <- read.csv("student-mat.csv", sep = ";")  
portuguese <- read.csv("student-por.csv", sep = ";") # our data was seperated by semi colons  
# instead of the traditional comma
```

Understanding the Data Structure

Prior to inspecting and cleaning our data, it is important we fully encapsulate what each column, row, and value mean.

```
head(math)
```

	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
4	GP	F	15	U	GT3	T	4	2	health	services	home
5	GP	F	16	U	GT3	T	3	3	other	other	home
6	GP	M	16	U	LE3	T	4	3	services	other	reputation

	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities
1	mother		2	2	0	yes	no	no
2	father		1	2	0	no	yes	no
3	mother		1	2	3	yes	no	yes
4	mother		1	3	0	no	yes	yes
5	father		1	2	0	no	yes	yes
6	mother		1	2	0	no	yes	yes

	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
4	yes	yes	yes	yes	3	2	2	1	1	5
5	yes	yes	no	no	4	3	2	1	2	5
6	yes	yes	yes	no	5	4	2	1	2	5

	absences	G1	G2	G3
1	6	5	6	6
2	4	5	5	6
3	10	7	8	10
4	2	15	14	15
5	4	6	10	10
6	10	15	15	15

```
head(portuguese)
```

	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
4	GP	F	15	U	GT3	T	4	2	health	services	home
5	GP	F	16	U	GT3	T	3	3	other	other	home
6	GP	M	16	U	LE3	T	4	3	services	other	reputation
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	0	yes	no	no	no	
4	mother			1	3	0	no	yes	no	yes	
5	father			1	2	0	no	yes	no	no	
6	mother			1	2	0	no	yes	no	yes	
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	
4	yes	yes	yes	yes	3	2	2	1	1	5	
5	yes	yes	no	no	4	3	2	1	2	5	
6	yes	yes	yes	no	5	4	2	1	2	5	
absences G1 G2 G3											
1	4	0	11	11							
2	2	9	11	11							
3	6	12	13	12							
4	0	14	14	14							
5	0	11	13	13							
6	6	12	12	13							

Both of our data sets are structured the same with the same column and row variables as well as structured values. This will help make implementing any cleaning and modeling simpler.

Variables & Values

Referring back to the original paper, there are 33 columns of interest. Those variables are listed below alongside their values, with explanation and clarification as needed, below:

For a visual example, I have printed a random row to show how the values are presented as they are explained below

```
math[3, ]
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason
3	GP	F	15	U	LE3	T	1	1	at_home	other	other
guardian traveltime studytime failures schoolsup famsup paid activities											

3	mother		1		2		3	yes	no	yes		no
	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health		
3	yes	yes	yes	no	4	3	2	2	3	3		
	absences	G1	G2	G3								
3		10	7	8	10							

school - Binary values of either **GP** (Gabriel Pereira) or **MS** (Mousinho da Silveira) of which school a student attended.

sex - Binary values of either **F** (female) or **M** (male) regarding a students sex.

age - Numeric value of a students age from 15 - 22.

address - Binary values of either **U** (urban) or **R** (rural) regarding a students home address.

famsize - Binary values of either **LE3** (less than or equal to 3 family members) or **GT3** (greater than 3 family members).

Pstatus - Binary values of either **T** (parents are living together) or **A** (parents are living apart) for parents living status.

Medu - Leveled integer value of range 0-4 with 0 reflecting no education or below primary completion, 1 reflecting completion of primary education (up to 4th grade), 2 reflecting completion of 5-9th grade education, 3 reflecting completion of secondary education, and 4 reflecting higher education (college degree or higher) of a students' mother's education.

Fedu - Leveled integer value of range 0-4 with 0 reflecting no education or below primary completion, 1 reflecting completion of primary education (up to 4th grade), 2 reflecting completion of 5-9th grade education, 3 reflecting completion of secondary education, and 4 reflecting higher education (college degree or higher) of a students' father's education.

Mjob - Nominal values for a students' mother's job classified as **teacher**, **health** (any care related profession), **services** (any administrative or police related field), **at_home** (none), **other** (not stated).

Fjob - Nominal values for a students' father's job classified as **teacher**, **health** (any care related profession), **services** (any administrative or police related field), **at_home** (none), **other** (not stated).

reason - Nominal values for a student's reason for school selection as either **home** (close to home), **reputation**, **course** (valued courses provided), **other** (reason not stated).

guardian - Nominal values for who the primary caregiver of the student is as either **mother**, **father**, or **other**. Reason for why both parents cannot be listed is not stated.

traveltime - Leveled integer values representing travel time to school on a scale of 1-4, 1 reflecting <15 minutes, 2 reflecting 15-30 minutes, 3 reflecting 30 minutes to 1 hour, 4 reflecting >1 hour travel time.

studytime - Leveled integer values representing average weekly study time reported by the student on a scale of 1-4, 1 reflecting <2 hours, 2 reflecting 2-5 hours, 3 reflecting 5-10 hours, 4 reflecting >10 hours study time.

failures - Leveled integer values representing the number of classes a student has failed prior to enrolling in this course with a scale of 1-4, each reflecting the amount of courses failed, 4 being > or = 4 failed classes.

schoolsup - Binary value for either **yes** or **no** student receiving additional educational support outside of the course. Not specified if this is inclusive of in-school tutoring and/or support such as services for language gaps or speech development.

famsup - Binary value for either **yes** or **no** student receiving additional family educational support outside of the course (family members assist in helping the student with studying or homework). If **yes**, we are assuming a student receives help from family generally.

paid - Binary value for either **yes** or **no** student is paying for additional educational support for the course.

activities - Binary value for either **yes** or **no** student is participating in extra-curricular activities.

nursery - Binary value for either **yes** or **no** student attended nursery school in the past (equivalent to pre-school education in America).

higher - Binary value for either **yes** or **no** student wants to pursue higher education courses in the future.

internet - Binary value for either **yes** or **no** student has internet access at home.

romantic - Binary value for either **yes** or **no** student is currently in a romantic relationship.

famrel - Leveled integer values scaled from 1-5 for a students quality of family relationships, 1 being very bad and 5 being excellent.

freetime - Leveled integer values scaled from 1-5 for a students free time after school, 1 being very little free time and 5 being lots of free time.

goout - Leveled integer values scaled from 1-5 of how often a student goes out with freinds, 1 being not often and 5 being very often.

Dalc - Leveled integer values scaled from 1-5 of how often a student consumes alcohol on a weekday, 1 being not often and 5 being very often.

Walc - Leveled integer values scaled from 1-5 of how often a student consumes alcohol on a weekend, 1 being not often and 5 being very often.

health - Leveled integer values scaled from 1-5 of a students health status, 1 being bad and 5 being very good.

absences - Numeric values of the number of day absences the student has from the course so far, i.e. a value of 5 would mean the student has been absent from the class a total of 5 times.

G1 - Leveled integer values scaled from 0-20 of a students first period grade in the course(period is a trimester in American equivalency).

G2 - Leveled integer values scaled from 0-20 of a students second period grade in the course.

G3 - Leveled integer values scaled from 0-20 of a students third period grade in the course.

Classes & Values

Given the review of our variables and their values, we should expect (was gonna explain what classes i expect them to be and also suggest changing the G1-G3 to numeric values so we can use them unleveled)

```
str(math)
```

```
'data.frame':  395 obs. of  33 variables:
 $ school      : chr  "GP" "GP" "GP" "GP" ...
 $ sex         : chr  "F" "F" "F" "F" ...
 $ age         : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address     : chr  "U" "U" "U" "U" ...
 $ famsize     : chr  "GT3" "GT3" "LE3" "GT3" ...
 $ Pstatus     : chr  "A" "T" "T" "T" ...
 $ Medu        : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu        : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob        : chr  "at_home" "at_home" "at_home" "health" ...
 $ Fjob        : chr  "teacher" "other" "other" "services" ...
 $ reason      : chr  "course" "course" "other" "home" ...
 $ guardian    : chr  "mother" "father" "mother" "mother" ...
 $ traveltime  : int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime   : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures    : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup   : chr  "yes" "no" "yes" "no" ...
 $ famsup      : chr  "no" "yes" "no" "yes" ...
 $ paid        : chr  "no" "no" "yes" "yes" ...
 $ activities  : chr  "no" "no" "no" "yes" ...
 $ nursery     : chr  "yes" "no" "yes" "yes" ...
 $ higher      : chr  "yes" "yes" "yes" "yes" ...
 $ internet    : chr  "no" "yes" "yes" "yes" ...
 $ romantic    : chr  "no" "no" "no" "yes" ...
```



```

$ famrel      : int  4 5 4 3 4 5 4 4 4 5 ...
$ freetime    : int  3 3 3 2 3 4 4 1 2 5 ...
$ goout       : int  4 3 2 2 2 2 4 4 2 1 ...
$ Dalc        : int  1 1 2 1 1 1 1 1 1 1 ...
$ Walc        : int  1 1 3 1 2 2 1 1 1 1 ...
$ health      : int  3 3 3 5 5 5 3 1 1 5 ...
$ absences    : int  6 4 10 2 4 10 0 6 0 0 ...
$ G1          : int  5 5 7 15 6 15 12 6 16 14 ...
$ G2          : int  6 5 8 14 10 15 12 5 18 15 ...
$ G3          : int  6 6 10 15 10 15 11 6 19 15 ...

```

```
str(portuguese)
```

```

'data.frame':  649 obs. of  33 variables:
 $ school      : chr  "GP" "GP" "GP" "GP" ...
 $ sex         : chr  "F" "F" "F" "F" ...
 $ age         : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address     : chr  "U" "U" "U" "U" ...
 $ famsize     : chr  "GT3" "GT3" "LE3" "GT3" ...
 $ Pstatus     : chr  "A" "T" "T" "T" ...
 $ Medu        : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu        : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob        : chr  "at_home" "at_home" "at_home" "health" ...
 $ Fjob        : chr  "teacher" "other" "other" "services" ...
 $ reason      : chr  "course" "course" "other" "home" ...
 $ guardian    : chr  "mother" "father" "mother" "mother" ...
 $ traveltime  : int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime   : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ schoolsup   : chr  "yes" "no" "yes" "no" ...
 $ famsup      : chr  "no" "yes" "no" "yes" ...
 $ paid        : chr  "no" "no" "no" "no" ...
 $ activities  : chr  "no" "no" "no" "yes" ...
 $ nursery     : chr  "yes" "no" "yes" "yes" ...
 $ higher      : chr  "yes" "yes" "yes" "yes" ...
 $ internet    : chr  "no" "yes" "yes" "yes" ...
 $ romantic    : chr  "no" "no" "no" "yes" ...
 $ famrel      : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime    : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout       : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc        : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc        : int  1 1 3 1 2 2 1 1 1 1 ...

```

```

$ health      : int   3 3 3 5 5 5 3 1 1 5 ...
$ absences    : int   4 2 6 0 0 6 0 2 0 0 ...
$ G1          : int   0 9 12 14 11 12 13 10 15 12 ...
$ G2          : int  11 11 13 14 13 12 12 13 16 12 ...
$ G3          : int  11 11 12 14 13 13 13 13 17 13 ...

```

This gives us a general idea of what each column look, and the class which is all integers and character columns.

Cleaning Up

```
sum(is.na(math))
```

```
[1] 0
```

```
sum(is.na(portuguese))
```

```
[1] 0
```

The original data from the survey was processed and certain variables were excluded by the author of the paper due to lack of discriminative value. To verify that our data sets are clean, we check to see if there are any missing values.

Our general approach to this project involves replicating some of the models used in the paper. The paper would predict student success using the G3 score, in one of three forums: binary classification, classification with five levels, and regression on the 0-20 scale. To ease the replication process we will create two new columns to represent the forum we want our output to be in:

```

math <- math |>
  mutate(five_level=case_when(
    G3 > 15 ~ "I",
    G3 >= 14 ~ "II",
    G3 >= 12 ~ "III",
    G3 >= 10 ~ "IV",
    G3 < 10 ~ "V"
  )) |>
  mutate(pass_fail=case_when(
    G3 >= 10 ~ "Pass",

```

```

    G3<10 ~ "Fail"
  )) -> math2

portuguese |>
  mutate(five_level=case_when(
    G3 > 15 ~ "I",
    G3 >= 14 ~ "II",
    G3 >=12 ~ "III",
    G3 >=10 ~ "IV",
    G3 < 10 ~ "V"
  )) |>
  mutate(pass_fail=case_when(
    G3>=10 ~ "Pass",
    G3<10 ~ "Fail"
  )) -> portuguese2

math2$five_level<-factor(math2$five_level)
portuguese2$five_level<-factor(portuguese2$five_level)

portuguese2 <- portuguese2 %>%
  mutate(across(c(pass_fail, romantic, internet, higher, nursery, activities, paid, famsup, s

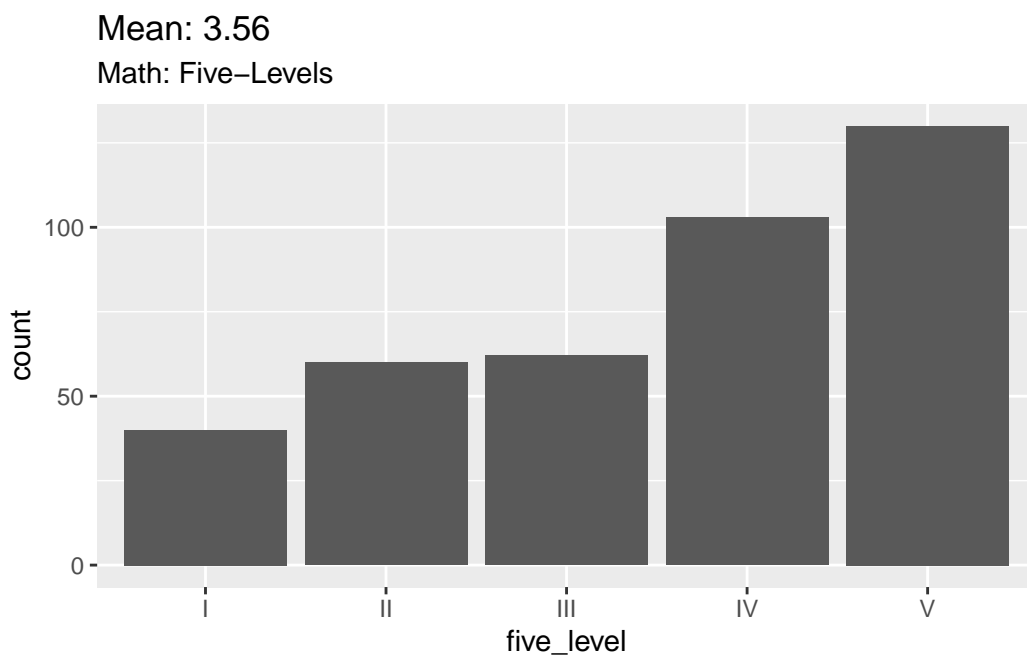
```

Correlations & Distribution

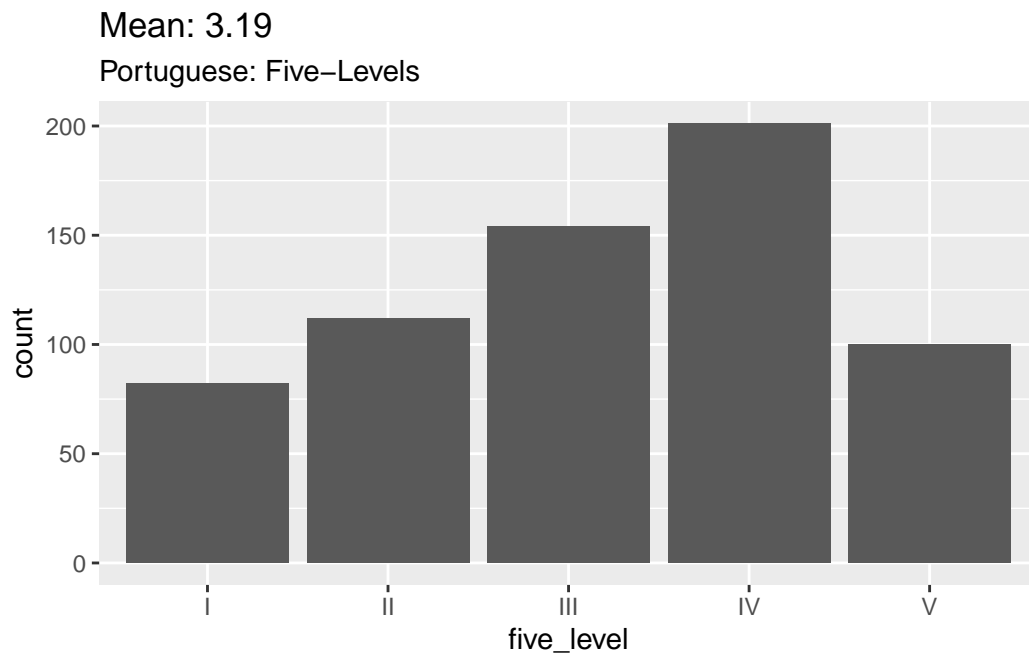
Means & Distributions

Here we exam visual distributions

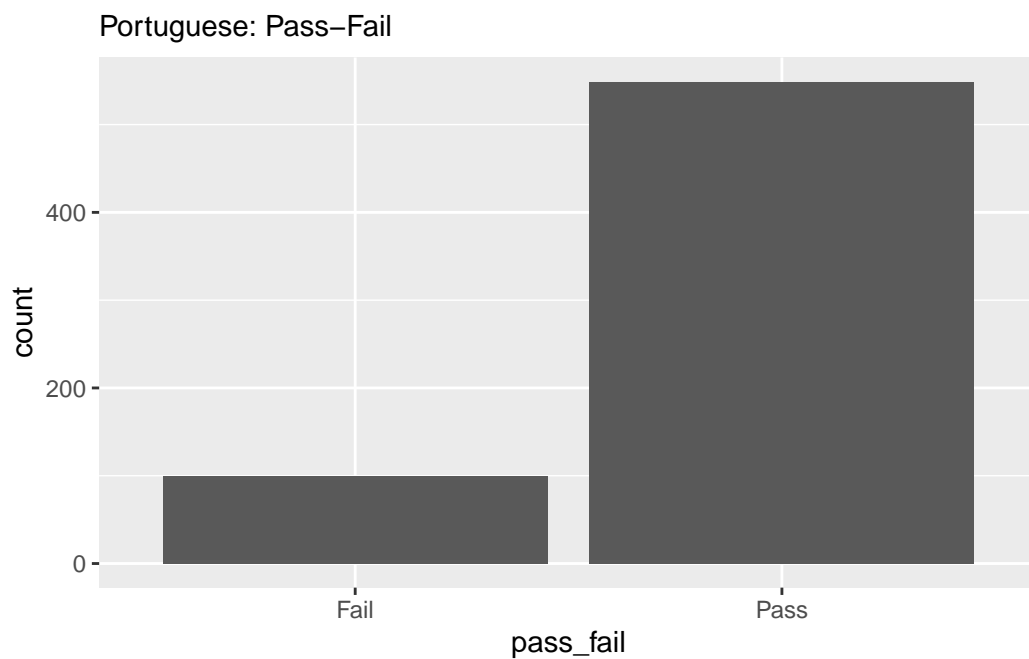
```
library(ggplot2)
ggplot(data=math2, aes(x=five_level))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(math2$five_level)), 2))) +
  labs(subtitle = "Math: Five-Levels")
```



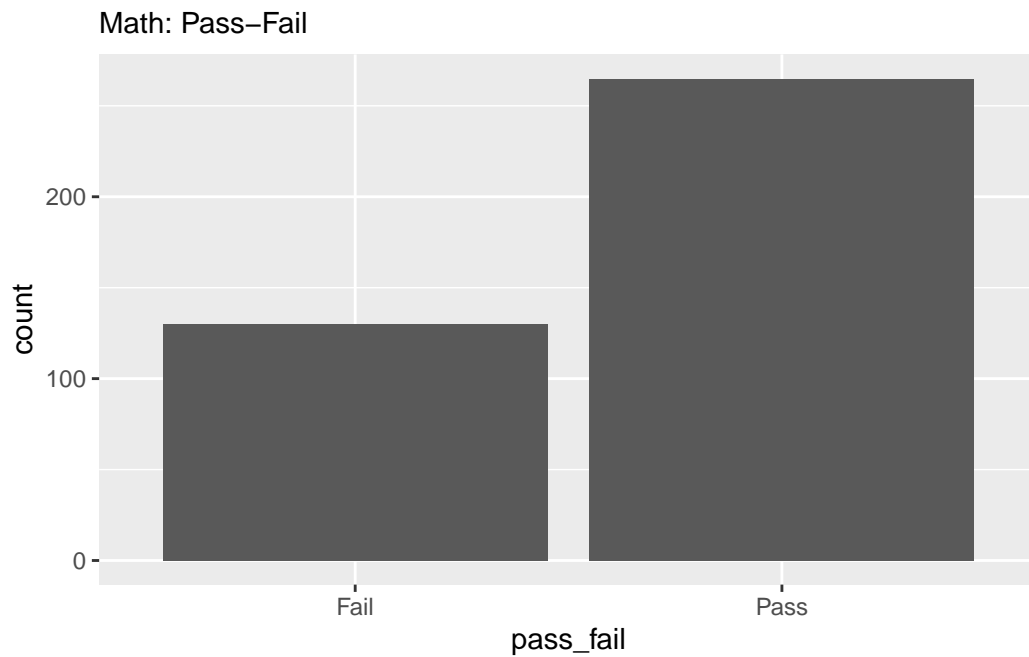
```
ggplot(data=portuguese2, aes(x=five_level))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(portuguese2$five_level)), 2))) +
  labs(subtitle = "Portuguese: Five-Levels")
```



```
ggplot(data=portuguese2, aes(x=pass_fail))+  
  geom_bar() +  
  labs(subtitle = "Portuguese: Pass-Fail")
```

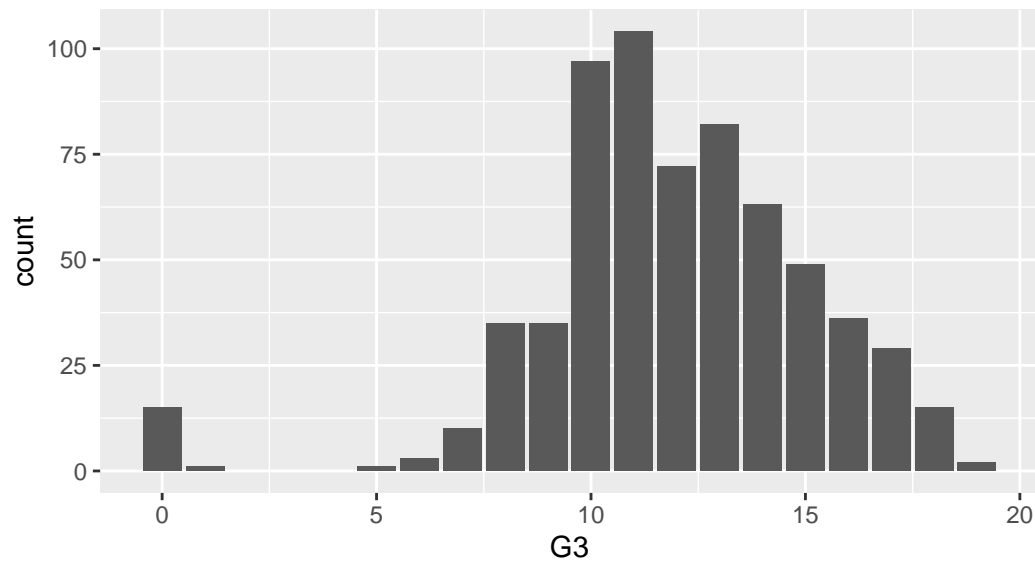


```
ggplot(data=math2, aes(x=pass_fail))+
  geom_bar() +
  labs(subtitle = "Math: Pass-Fail")
```



```
ggplot(data=portuguese2, aes(x=G3))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(portuguese2$G3)), 2))) +
  labs(subtitle = "Portuguese: G3")
```

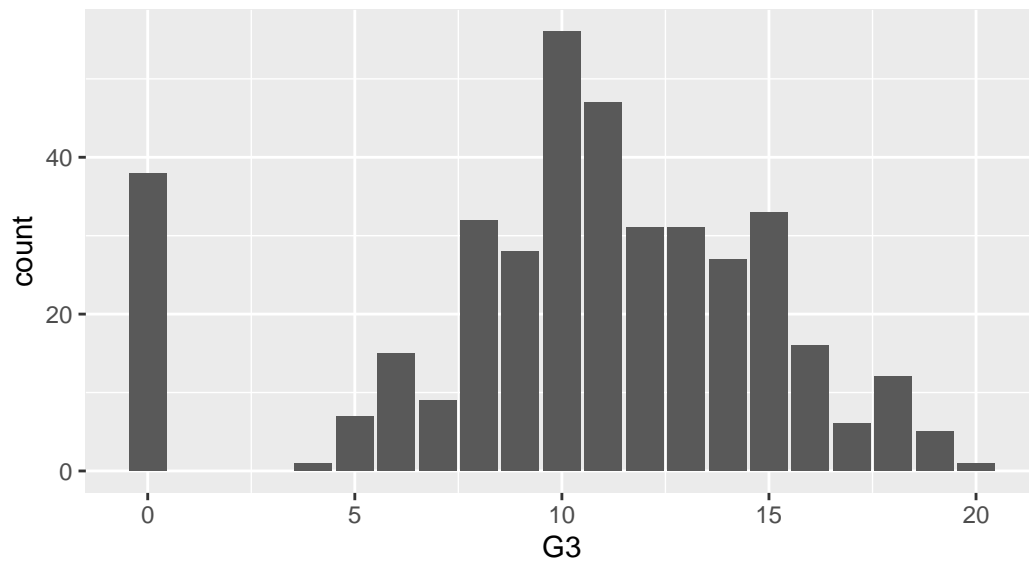
Mean: 11.91
Portuguese: G3



```
ggplot(data=math2, aes(x=G3))+  
  geom_bar() +  
  ggtitle(paste("Mean:", round(mean(  
    as.numeric(math2$G3)), 2))) +  
  labs(subtitle = "Math: G3")
```

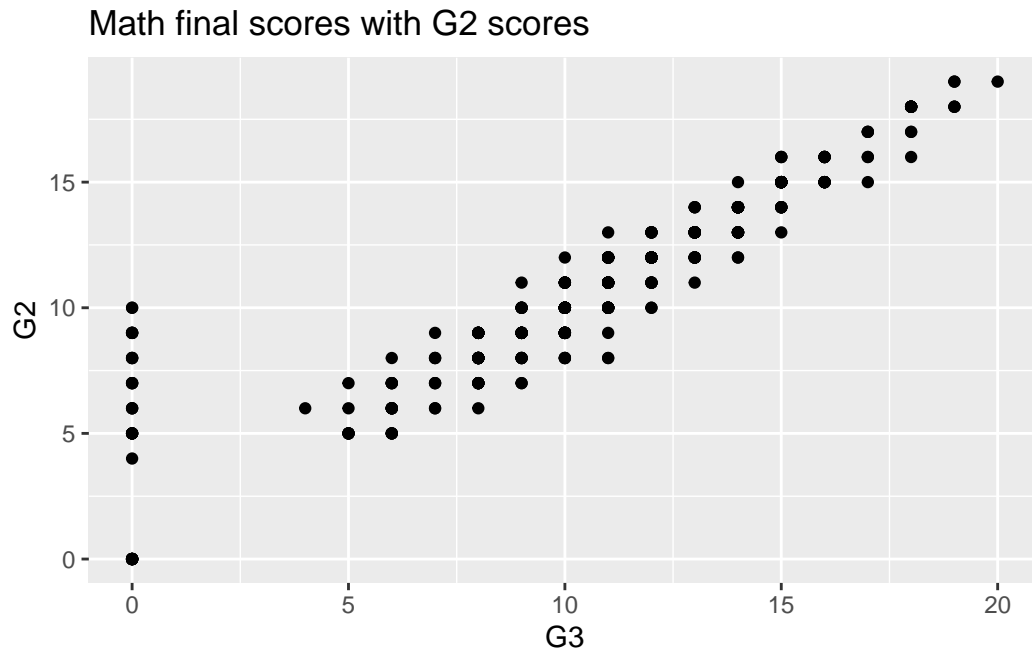
Mean: 10.42

Math: G3



Correlations & Plots

```
ggplot(data=math2, aes(x=G3, y=G2))+  
  geom_point()+  
  labs(title="Math final scores with G2 scores")
```

Clear correlation between second period grades and final grade, shows the in-balance of models that use G2 as a predictor vs those that don't. We will dive into collinearity assumptions with tests in the statistical analysis section.

Statistical Analysis

It's important to note any patterns or anomalies with our data. We will look at possible outliers and quickly summarize G3 (our predicted variable).

```
summary(portuguese2$G3)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	10.00	12.00	11.91	14.00	19.00

```
sd(portuguese2$G3)
```

```
[1] 3.230656
```

```
summary(math2$G3)
```

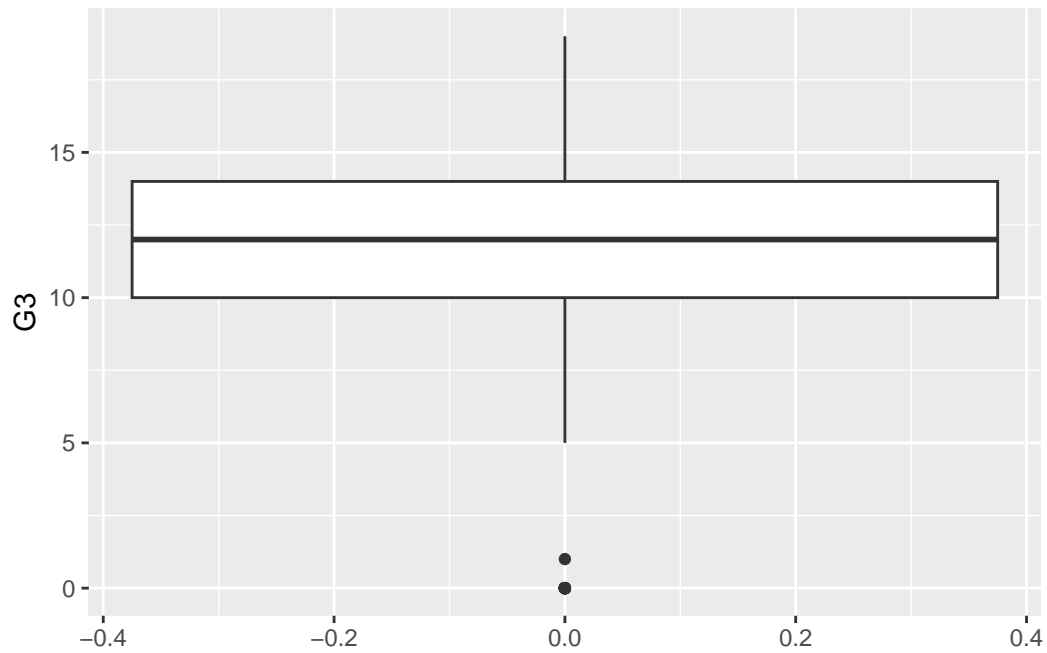
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	8.00	11.00	10.42	14.00	20.00

```
sd(math2$G3)
```

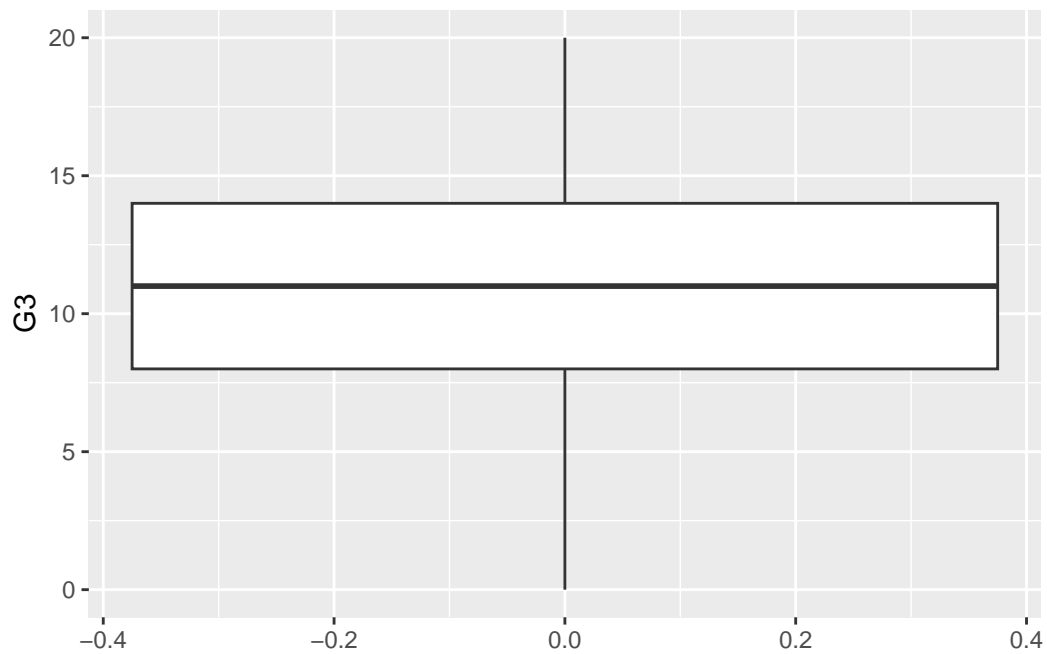
```
[1] 4.581443
```

It seems most students pass, with math scores being slightly lower on average.

```
ggplot(portuguese2, aes(y=G3)) + geom_boxplot()
```



```
ggplot(math2, aes(y=G3)) + geom_boxplot()
```



It seems our Portuguese course has two values that are outliers, but we will not remove them

as values due to their predictive ability for students who may fail a class. Also, tree-based models are not affected by outliers.

There are a few ways to test for collinearity with variables: VIF, visualization on a scatter plot, or using a pairwise approach and testing its correlation.

```
# testing using VIF
lm_for_VIF <- lm(G3 ~ G1 + G2, data=portuguese2)

vif(lm_for_VIF)
```

```
      G1      G2
3.971299 3.971299
```

A VIF score of 1 typically indicates no correlation with other predictors. A VIF of 10 is generally considered too high. However, it's also important to consider what kind of model we are creating. We are creating prediction models, so we would consider a value of about ~3.97 to be relatively moderate. Essentially, utilization of both predictors G1 and G2 in our model is not likely to cause issues with predicting our outcome, G3.

```
grades <- portuguese2[, c("G1", "G2", "G3")]

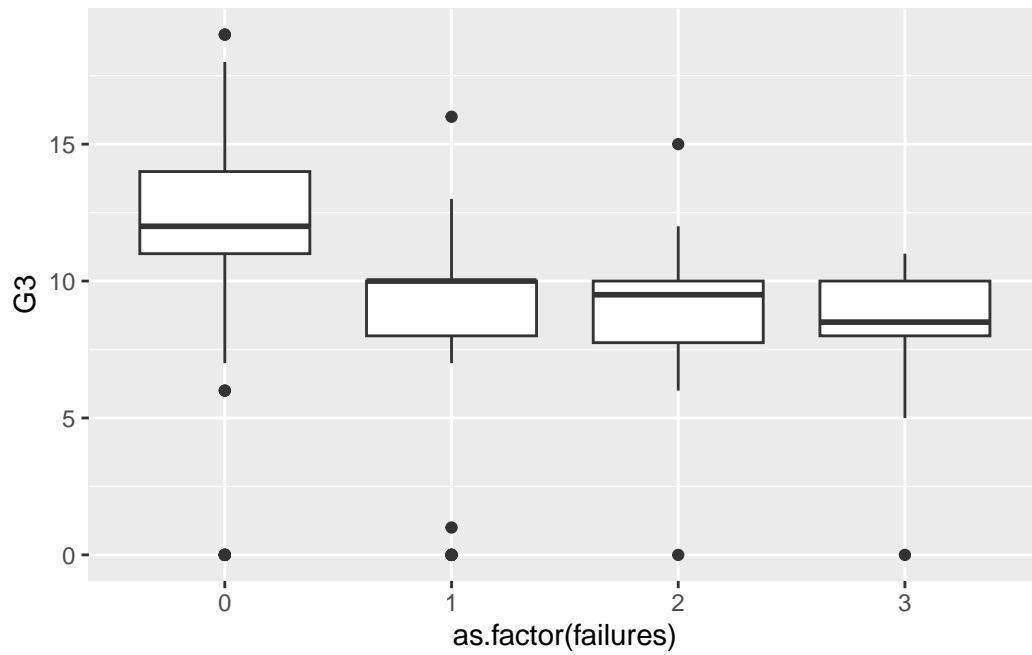
cor_matrix <- cor(grades, use = "complete.obs")
print(cor_matrix)
```

```
      G1      G2      G3
G1 1.0000000 0.8649816 0.8263871
G2 0.8649816 1.0000000 0.9185480
G3 0.8263871 0.9185480 1.0000000
```

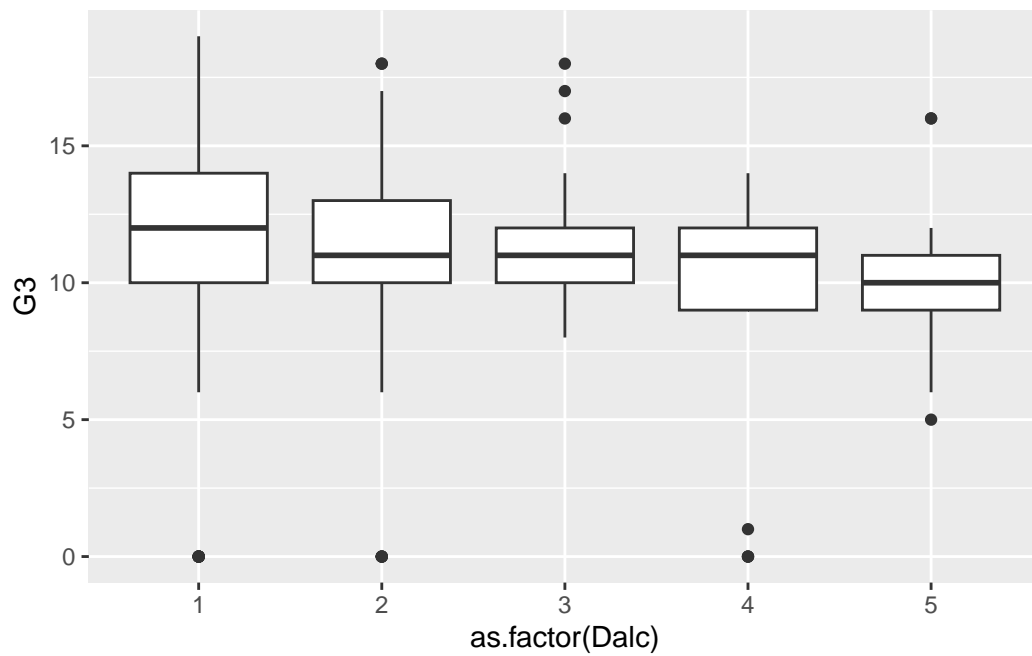
However, now that we have run a correlation matrix, it is displaying very strong correlation between our variables G1, G2, and G3. This confirms high collinearity among them, which would cause an increase in standard errors in our regression models.

Exploratory Graphs

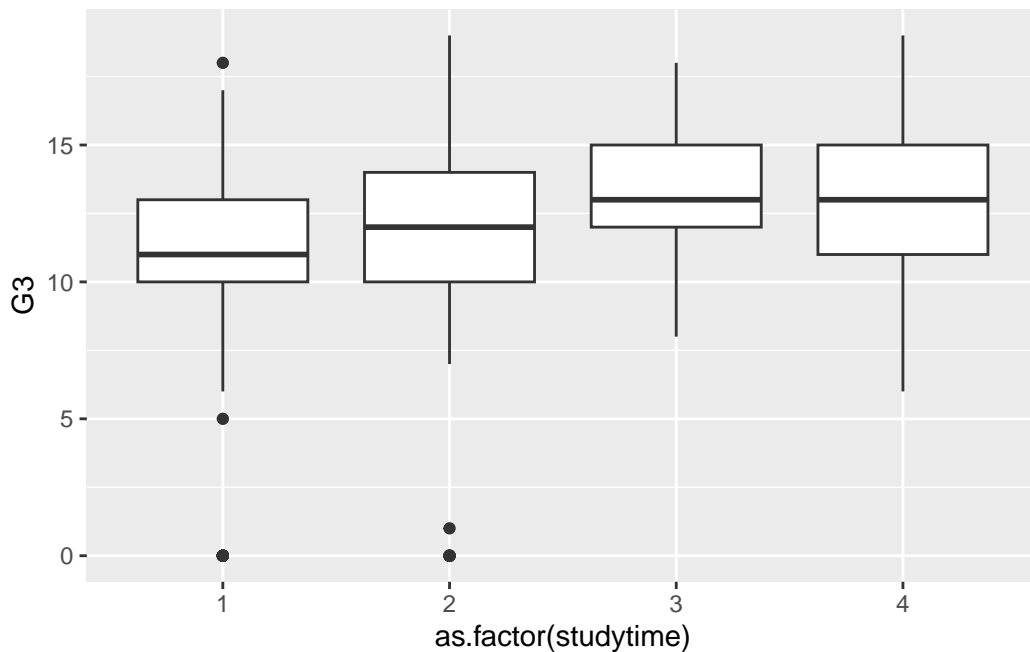
```
ggplot(data=portuguese2, aes(x=as.factor(failures), y=G3))+
  geom_boxplot()
```



```
ggplot(data=portuguese2, aes(x=as.factor(Dalc), y=G3))+
  geom_boxplot()
```



```
ggplot(data=portuguese2, aes(x=as.factor(studytime), y=G3))+
  geom_boxplot()
```



We wanted to look at the relationship between some predictor variables that we thought may have a strong relationship to G3 final grades. We do see small correlations like lower grades as alcohol consumption increases, and higher grades as study time increases.

Review of Plan to fit Models

Our general approach to this project will be to recreate some of the models created in the [paper](#) connected with this data set and additionally create some of our own. In the paper, the classification/regression methods tried to predict G3 (passing/failing Portuguese and Math) in 3 supervised approaches:

- Binary (Pass/Fail) - Pass is considered $G3 \geq 10$; else is Fail
- 5-level Classification
- Regression (as is current G3 column, scale 0-20)

The data was then modeled using 5 data mining algorithm:

- Neural Networks (NN) - E = 100 training epochs utilizing BFGS algorithm

- Support Vector Machines (SVM) - SMO algorithm utilized
- Decision Trees (DT) - node splitting utilized to reduce sum of squares
- Random Forest (RF) - default parameters, $T = 500$
- Naive Predictor (NV) - (1) baseline of $G3 = G2$, (2) $G3 = G1$, (3) most common class or mean

These 4 DM's were compared against the baseline naive predictor (NV) model. Additionally it was noted that 20 runs of 10-fold cross validation were applied to each configuration.

Each model was run with each of 3 input setups. The setups included (A), all variables minus $G3$, (B) all variables minus $G2$ and $G3$, and (C) neither $G2$, $G3$, or $G1$. This means that model (A) is utilizing the prediction power of $G1$ and $G2$ grades in their model for accurate prediction of $G3$ grades, where setup (B) utilizes solely $G1$ grades to predict $G3$, and setup (C) uses none of the trimester grades as a predictor for $G3$.

The reason the authors made this decision was due to the likelihood of high collinearity between $G1$, $G2$, and $G3$. The usage of Naive Predictors as three input configurations is to account for this potential (and likely) collinearity.

Furthermore, more pre-processing was established with nominal variables as well. The authors decided to transform them into a *1-of-C* encoding with all attributes being standardized to a 0 mean and a one standard deviation.

Modeling Procedures

According to the paper that we are replicating, their goal was to “give a simple description that summarizes the best DM models”. The authors used this model as it was collected, essentially creating a model that could be used to predict student outcomes once the student was in their third trimester of the class. We intend to work with the same desired prediction and usage of the variables.

In order to differentiate our model from theirs while still replicating part of the study, we wish to take the approach that the model is used prior to a student's choice to enroll in a class. Essentially, our model sees to predict a student's outcome in the class using variables that are known prior and during class enrollment so that a user could predict their grade before the third trimester.

We understand the choice of the authors to use an A-B-C subset method, however, given time constraint, we decided to solely select subset (A). Inclusion of G1 and G2 would indicate a model that can predict a student's grade for trimester 3 while considering trimester 1 and 2.

We will prioritize replicating two of their original models - Decision Tree (DT) and Random Forest (RF). However, we plan to add three of our models: Partial Least Squares (PLS), LASSO Regression (LR), and Linear Discriminant Analysis (LDA). PLS and LR will use continuous 1-20 outcomes while our LDA would utilize the binary (pass/fail) outcome. We will also add a simple Multiple Linear Regression (MLR) as a baseline model. All of such will be used to predict outcomes for the Portuguese course (which has more observations, 649 v. 395) for time constraint, and only reproduce models with input A, as we want to view prediction power with the utilization of G1 and G2 grades.

Note on data splitting:

The paper states: “To access the predictive performances, 20 runs of a 10- fold cross-validation (Hastie et al. 2001) (in a total of 200 simulations) were applied to each configuration. Under such scheme, for a given run the data is randomly divided in 10 subsets of equal size. Sequentially, one different subset is tested (with 10% of the data) and the remaining data used to fit the DM technique. At the end of this process, the evaluated test set contains the whole dataset, although 10 variations of the same DM model are used to create the predictions.”

We will do an initial 20-80 split and run cross-validation on most of our models.

In summary, we intend to replicate this study, with key differences: we intend to use solely setup (C), we intend to replicate Decision Tree and Random Forest from the original paper but replace the other models with... change

Partial Least Squares, Lasso and LDA, and examine solely the Portuguese course data for time constraint reasons.

Part 2: Model Fitting - Binary; Set Up C

If time allocates, we still intend to work towards replicating the models with five-level as well.

Modeling Set Up - Creating Five Level, and Pass/Fail

```
portuguese |>
  mutate(five_level=case_when(
    G3 > 15 ~ "I",
    G3 >= 14 ~ "II",
    G3 >=12 ~ "III",
    G3 >=10 ~ "IV",
    G3 < 10 ~ "V"
  )) |>
  mutate(pass_fail=case_when(
    G3>=10 ~ "Pass",
    G3<10 ~ "Fail"
  )) -> portuguese2

portuguese2$five_level<-factor(portuguese2$five_level)

portuguese2 <- portuguese2 %>%
  mutate(across(c(pass_fail, romantic, internet, higher, nursery, activities, paid, famsup, s

#portuguese2 <- portuguese2 |>
#  mutate(pass_fail = ifelse(pass_fail == "Pass", 1, 0))
#Commented out code that changed pass fail to 0 1 because our models need to read the respon

set.seed(627)
train.pct <- 0.8
Z <- sample(nrow(portuguese2), floor(train.pct*nrow(portuguese2)))
portuguese.data <- portuguese2[Z, ]
holdout.data <- portuguese2[-Z, ]
```

Note: we will use cross validation (as requested in the doc) and 20% testing 80% training data.

Logistic Regression (LR) with PCA - Binary Outcome

Below is our model, followed by assumptions to assure reliability.

```
# remember we are removing G3 and five_level since those are for other outcomes
library(dplyr)
lrpca <- glm(pass_fail ~ . -five_level -G3 -G2 -G1, data = portuguese.data,
             family = binomial)
```

Lets check and see how it is performing pre-PCA:

```
summary(lrpca)
```

Call:

```
glm(formula = pass_fail ~ . - five_level - G3 - G2 - G1, family = binomial,
     data = portuguese.data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.75513	2.81507	-0.623	0.53297	
schoolMS	-2.24567	0.40537	-5.540	3.03e-08	***
sexM	-0.24249	0.38527	-0.629	0.52908	
age	0.32909	0.15802	2.083	0.03729	*
addressU	0.16536	0.34922	0.474	0.63584	
famsizeLE3	0.43037	0.37651	1.143	0.25301	
PstatusT	0.02377	0.50428	0.047	0.96241	
Medu	-0.12384	0.20545	-0.603	0.54665	
Fedu	0.42658	0.21305	2.002	0.04526	*
Mjobhealth	-0.37835	0.76013	-0.498	0.61866	
Mjobother	0.02871	0.38531	0.075	0.94061	
Mjobservices	0.29942	0.51563	0.581	0.56145	
Mjobteacher	0.71178	0.86049	0.827	0.40814	
Fjobhealth	-2.31779	1.09859	-2.110	0.03488	*
Fjobother	-0.90121	0.64050	-1.407	0.15941	
Fjobservices	-1.46644	0.68504	-2.141	0.03230	*
Fjobteacher	-2.39491	1.18783	-2.016	0.04378	*
reasonhome	0.35360	0.43465	0.814	0.41593	
reasonother	0.31678	0.45758	0.692	0.48875	
reasonreputation	0.54396	0.51262	1.061	0.28863	
guardianmother	-0.66137	0.40632	-1.628	0.10359	
guardianother	0.19517	0.79965	0.244	0.80718	

traveltime	0.25544	0.22084	1.157	0.24741	
studytime	0.19951	0.21650	0.922	0.35677	
failures	-1.07777	0.22975	-4.691	2.72e-06	***
schoolsupyes	-0.86161	0.53200	-1.620	0.10532	
famsupyes	0.41928	0.32732	1.281	0.20021	
paidyes	-1.31506	0.57899	-2.271	0.02313	*
activitiesyes	0.39895	0.33580	1.188	0.23481	
nurseryyes	-0.33692	0.40088	-0.840	0.40066	
higheryes	1.39555	0.42529	3.281	0.00103	**
internetyes	-0.17765	0.36459	-0.487	0.62608	
romanticyes	-0.50700	0.31781	-1.595	0.11064	
famrel	-0.01271	0.15398	-0.083	0.93424	
freetime	-0.11001	0.15707	-0.700	0.48369	
goout	0.04354	0.15397	0.283	0.77735	
Dalc	0.02728	0.20017	0.136	0.89158	
Walc	-0.25789	0.16788	-1.536	0.12448	
health	-0.07436	0.12111	-0.614	0.53921	
absences	-0.10672	0.03618	-2.950	0.00318	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 459.55 on 518 degrees of freedom
 Residual deviance: 305.82 on 479 degrees of freedom
 AIC: 385.82

Number of Fisher Scoring iterations: 6

Our model is showing an residual deviance of 305.82, so our model is so and so on the portion of the variance in pass_fail. This was expected when removing G2 and G1.

Now lets utilize PCA's to create a more tuned model:

```
portuguese_X <- model.matrix(lrpca)[, -1]
pca <- prcomp(portuguese_X)
summary(pca)
```

Importance of components:

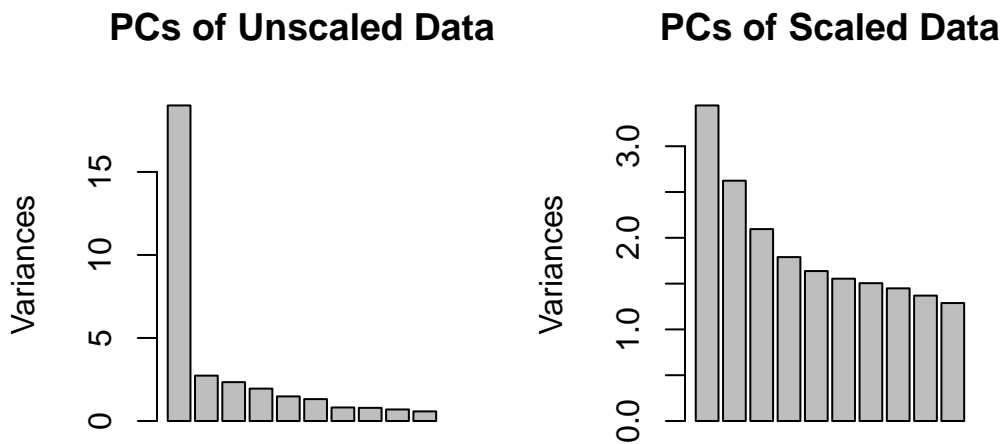
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	4.3595	1.6531	1.53037	1.39669	1.21770	1.14836	0.9032
Proportion of Variance	0.5195	0.0747	0.06402	0.05332	0.04053	0.03605	0.0223

Cumulative Proportion	0.5195	0.5942	0.65825	0.71158	0.75211	0.78816	0.8105
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Standard deviation	0.8909	0.83436	0.76149	0.69090	0.64606	0.62377	0.57140
Proportion of Variance	0.0217	0.01903	0.01585	0.01305	0.01141	0.01064	0.00893
Cumulative Proportion	0.8321	0.85118	0.86703	0.88008	0.89149	0.90213	0.91105
	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Standard deviation	0.51954	0.50800	0.49246	0.49097	0.47465	0.4686	0.44176
Proportion of Variance	0.00738	0.00705	0.00663	0.00659	0.00616	0.0060	0.00533
Cumulative Proportion	0.91843	0.92549	0.93212	0.93870	0.94486	0.9509	0.95620
	PC22	PC23	PC24	PC25	PC26	PC27	PC28
Standard deviation	0.41365	0.40021	0.38525	0.37632	0.36369	0.35652	0.3258
Proportion of Variance	0.00468	0.00438	0.00406	0.00387	0.00362	0.00347	0.0029
Cumulative Proportion	0.96088	0.96526	0.96931	0.97318	0.97680	0.98027	0.9832
	PC29	PC30	PC31	PC32	PC33	PC34	PC35
Standard deviation	0.29815	0.29328	0.28946	0.27588	0.25902	0.24547	0.22317
Proportion of Variance	0.00243	0.00235	0.00229	0.00208	0.00183	0.00165	0.00136
Cumulative Proportion	0.98561	0.98796	0.99025	0.99233	0.99416	0.99581	0.99717
	PC36	PC37	PC38	PC39			
Standard deviation	0.18951	0.18679	0.1487	0.10297			
Proportion of Variance	0.00098	0.00095	0.0006	0.00029			
Cumulative Proportion	0.99815	0.99911	0.9997	1.00000			

I want to look at both scaled and unscaled data to see if unscaled fits better (likely, given our differing ranges in scaled data)

```
portuguese_pcs <- prcomp(portuguese_X, scale=TRUE)

par(mfrow=c(1, 2))
plot(pca, main="PCs of Unscaled Data")
screeplot(portuguese_pcs, main="PCs of Scaled Data")
```



Very clearly were one variable carrying the most weight, so scaling is better here. We will now view the variance of our pca's to decide how many to include in our model:

```
library(pls)
```

Attaching package: 'pls'

The following object is masked from 'package:stats':

loadings

```
portuguese.data_01 <- portuguese.data |>
  mutate(pass_fail = ifelse(pass_fail == "Pass", 1, 0))

portuguese_pcr <- pcr(pass_fail ~ . -five_level -G3 -G2 -G1,
  data = portuguese.data_01, family = binomial,
  scale = TRUE, validation = "CV")

summary(portuguese_pcr)
```

Data: X dimension: 519 39
Y dimension: 519 1
Fit method: svdpc
Number of components considered: 39

VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps
CV	0.369	0.3573	0.3494	0.3476	0.3482	0.3477	0.3463
adjCV	0.369	0.3572	0.3492	0.3474	0.3481	0.3473	0.3459
	7 comps	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps
CV	0.3471	0.3465	0.3389	0.3394	0.3394	0.3400	0.3435
adjCV	0.3471	0.3467	0.3381	0.3387	0.3389	0.3393	0.3432
	14 comps	15 comps	16 comps	17 comps	18 comps	19 comps	20 comps
CV	0.3414	0.3394	0.3376	0.3385	0.3396	0.3391	0.3408
adjCV	0.3409	0.3385	0.3366	0.3378	0.3389	0.3390	0.3402
	21 comps	22 comps	23 comps	24 comps	25 comps	26 comps	27 comps
CV	0.3402	0.3405	0.3431	0.3436	0.3443	0.3444	0.3457
adjCV	0.3388	0.3395	0.3421	0.3427	0.3429	0.3432	0.3446
	28 comps	29 comps	30 comps	31 comps	32 comps	33 comps	34 comps
CV	0.3467	0.3462	0.3459	0.3424	0.3422	0.3385	0.3384
adjCV	0.3456	0.3451	0.3448	0.3411	0.3410	0.3371	0.3373
	35 comps	36 comps	37 comps	38 comps	39 comps		
CV	0.3399	0.3388	0.3390	0.3395	0.3397		
adjCV	0.3382	0.3374	0.3376	0.3380	0.3381		

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
X	8.834	15.56	20.94	25.52	29.72	33.71	37.57
pass_fail	7.030	11.53	12.75	12.79	13.56	14.19	14.20
	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps
X	41.28	44.79	48.09	51.23	54.28	57.25	59.97
pass_fail	14.53	18.42	18.51	18.52	18.67	18.75	19.55
	15 comps	16 comps	17 comps	18 comps	19 comps	20 comps	21 comps
X	62.59	65.08	67.50	69.82	72.06	74.27	76.37
pass_fail	20.55	21.08	21.19	21.21	21.23	21.81	22.84
	22 comps	23 comps	24 comps	25 comps	26 comps	27 comps	28 comps
X	78.42	80.38	82.26	83.97	85.66	87.30	88.88
pass_fail	22.90	23.00	23.00	23.43	23.54	23.54	23.75
	29 comps	30 comps	31 comps	32 comps	33 comps	34 comps	35 comps
X	90.43	91.87	93.19	94.45	95.63	96.69	97.68
pass_fail	23.82	24.11	25.46	25.76	27.26	27.39	28.15
	36 comps	37 comps	38 comps	39 comps			

X	98.63	99.31	99.74	100.0
pass_fail	28.21	28.37	28.39	29.1

STILL WORKING

Model 2: Step-wise AIC

Model 3: Decision Tree

```
set.seed(627)
library(tree)
tree_initial<-tree(pass_fail ~ . -G3-five_level-G2-G1, data=portuguese.data)
summary(tree_initial)
```

Classification tree:

```
tree(formula = pass_fail ~ . - G3 - five_level - G2 - G1, data = portuguese.data)
```

Variables actually used in tree construction:

```
[1] "failures" "school" "higher" "Walc" "absences"
[6] "famsup" "schoolsup" "famrel" "goout" "guardian"
[11] "activities" "Fedu" "Fjob" "Mjob" "Medu"
[16] "reason"
```

Number of terminal nodes: 22

Residual mean deviance: 0.4507 = 224 / 497

Misclassification error rate: 0.1175 = 61 / 519

Tuning tree:

```
set.seed(627)
cv.tree.model<- cv.tree(tree_initial, FUN=prune.misclass, K=10)
cv.tree.model
```

\$size

```
[1] 22 11 9 8 3 1
```

\$dev

```
[1] 83 83 82 83 80 88
```

\$k

```
[1] -Inf  0.0  0.5  1.0  1.8  6.0

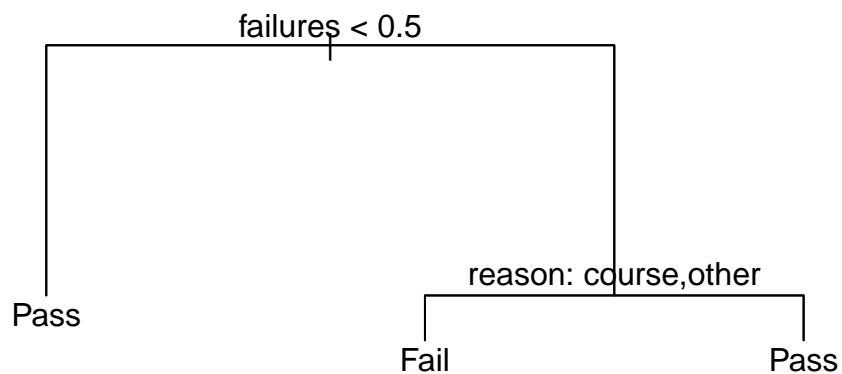
$method
[1] "misclass"

attr(,"class")
[1] "prune"          "tree.sequence"
```

```
plot(cv.tree.model$size, cv.tree.model$dev, type="b")
```



```
tuned_tree<-prune.misclass(tree_initial, best=3)
plot(tuned_tree)
text(tuned_tree, pretty=0)
```

Model 4: Random Forest

```
library(randomForest)
```

```
randomForest 4.7-1.2
```

Type `rfNews()` to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:dplyr':

`combine`

The following object is masked from 'package:ggplot2':

`margin`

```
set.seed(627)
randomForest(pass_fail ~ . -G3-five_level-G2-G1, data=portuguese.data, mtry=6, importance=TRUE)
RF6
```

```
Call:
randomForest(formula = pass_fail ~ . - G3 - five_level - G2 - G1, data = portuguese.data)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 6

OOB estimate of error rate: 15.41%
Confusion matrix:
      Fail Pass class.error
Fail   17   67  0.79761905
Pass   13  422  0.02988506
```

```
randomForest(pass_fail ~ . -G3-five_level-G2-G1, data=portuguese.data, mtry=5, importance=TRUE)
RF5
```

```
Call:
randomForest(formula = pass_fail ~ . - G3 - five_level - G2 - G1, data = portuguese.data)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 5

OOB estimate of error rate: 15.99%
Confusion matrix:
      Fail Pass class.error
Fail   11   73  0.86904762
Pass   10  425  0.02298851
```

```
set.seed(627)
randomForest(pass_fail ~ . -G3-five_level-G2-G1, data=portuguese.data, mtry=4, importance=TRUE)
RF4
```

```
Call:
randomForest(formula = pass_fail ~ . - G3 - five_level - G2 - G1, data = portuguese.data)
```

```

Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 4

```

```

OOB estimate of error rate: 15.61%
Confusion matrix:

```

```

      Fail Pass class.error
Fail    9   75   0.8928571
Pass    6  429   0.0137931

```

```

randomForest(pass_fail ~ . -G3-five_level-G2-G1, data=portuguese.data, mtry=7, importance=TRUE)
RF7

```

Call:

```

randomForest(formula = pass_fail ~ . - G3 - five_level - G2 - G1, data = portuguese.data)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 7

```

```

OOB estimate of error rate: 16.38%
Confusion matrix:

```

```

      Fail Pass class.error
Fail   13   71  0.84523810
Pass   14  421  0.03218391

```

Looking at the OOB estimate of error rate, and keeping in mind the ideal of $m = \sqrt{p}$, we tried making m 4, 5, 6, and 7 (we have 30 predictors in the model). The model with $m=8$ produces the smallest OOB error rate of 15.41.

looking at RF variable importance:

```

importance(RF6)

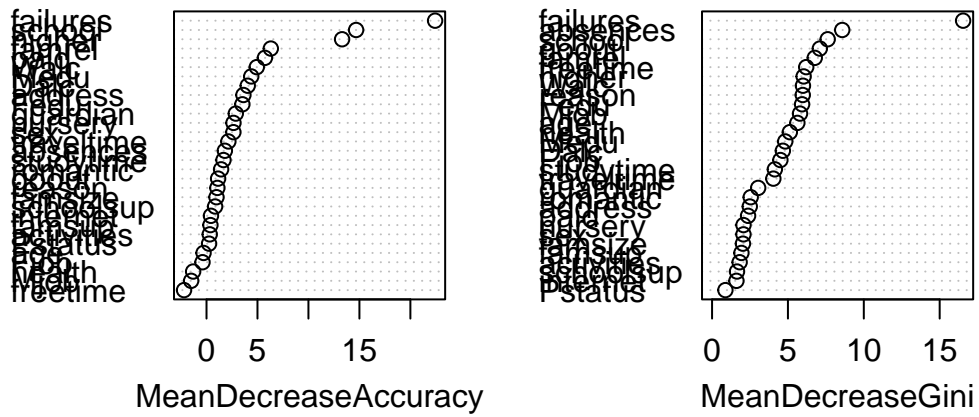
```

	Fail	Pass	MeanDecreaseAccuracy	MeanDecreaseGini
school	13.7114807	10.30657356	14.6768506	7.6201262
sex	0.4362858	2.48407103	2.6385787	2.0689599
age	-2.3624385	1.13445705	-0.2885577	5.6223881
address	2.0912671	2.92826929	3.6210346	2.4850924
famsize	1.6805549	0.34198811	0.9800217	2.0347374
Pstatus	0.3449406	0.03588579	0.2507111	0.8778668

Medu	0.6789849	4.50527497	4.3897110	4.8219339
Fedu	1.9519419	2.80316536	3.5198297	5.9326023
Mjob	-2.0217749	-0.52867112	-1.5057109	5.7963529
Fjob	-1.4502329	0.35206852	-0.3909861	4.5030425
reason	4.9084283	-1.83286420	1.0681412	5.9762324
guardian	0.8842241	2.96326664	2.8719778	3.0386220
traveltime	0.8135458	1.91265051	2.1449013	4.0386688
studytime	5.2450818	-1.13391799	1.6635819	4.1361001
failures	20.2476080	16.29799434	22.4263316	16.5449394
schoolsup	2.8855173	-0.20739444	0.8438599	1.6424743
famsup	0.8250592	0.05192372	0.3857237	1.9354232
paid	3.1860566	5.25500742	5.7292298	2.3733610
activities	0.3040239	0.32333543	0.3184525	1.8169551
nursery	1.1640589	2.32866591	2.6450890	2.0709782
higher	8.2603815	10.78585427	13.3000307	6.0349992
internet	-0.9734204	1.23929495	0.4373140	1.6069450
romantic	0.8178356	0.94714208	1.4321200	2.5469623
famrel	5.5423390	3.66719179	6.3073915	6.7670239
freetime	1.7281589	-3.28913038	-2.2019527	6.2058309
goout	1.9509136	0.00652265	1.1398874	7.1022976
Dalc	-1.1523620	5.05715203	4.0445543	4.6644272
Walc	0.9359551	5.02044989	4.9310812	5.9893684
health	-1.7409117	-0.50991774	-1.3092199	5.1247424
absences	-0.6257365	2.25550033	1.8006001	8.5779611

```
varImpPlot(RF6)
```

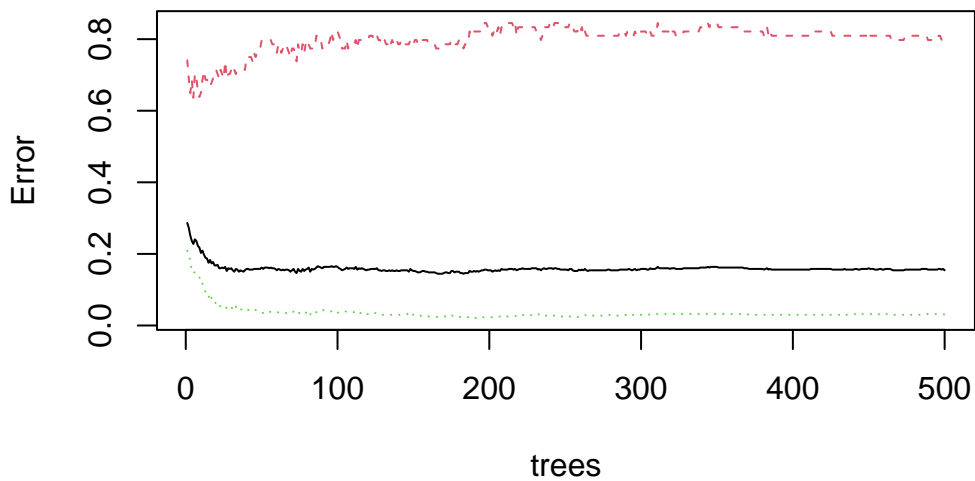
RF6



Failure seems to be very important in improving the accuracy, followed by school and absences.

```
plot(RF6)
```

RF6



Model 5: Support Vector Machine

more models if time

Using Testing data on models

Decision Tree Model - Binary Setup C

```
pred_val<-predict(tree_initial, newdata=holdout.data, type="class")
table(pred_val, holdout.data$pass_fail)
```

```
pred_val Fail Pass
      Fail     9   10
      Pass     7  104
```

```
mis.class<-mean(pred_val != holdout.data$pass_fail)
mis.class
```

```
[1] 0.1307692
```

```
pred_val2<-predict(tuned_tree, newdata=holdout.data, type="class")
table(pred_val2, holdout.data$pass_fail)
```

```
pred_val2 Fail Pass
      Fail     8    7
      Pass     8  107
```

```
mis.class2<-mean(pred_val2 != holdout.data$pass_fail)
mis.class2
```

```
[1] 0.1153846
```

Our initial tree model had a misclassification rate of around 13.08% on our testing data. We will tune it to see if that can be improved. Tuning the tree with cross validation brings our misclassification rate down to 11.54% which is better, but the model is more inflexible now. Note that reason is a categorical variable with 4 options, course and other being two of the options.

Random Forest Model - Binary Setup C