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 pubic 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

```r
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 becom
```

```r
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

```r
# 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.

```r
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         no
2   father          1         2        0        no    yes   no         no
3   mother          1         2        3       yes     no  yes         no
4   mother          1         3        0        no    yes  yes        yes
5   father          1         2        0        no    yes  yes         no
6   mother          1         2        0        no    yes  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
```

```r
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

```r
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)

```r
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 ...
```

```r
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

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

```
[1] 0
```

```r
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:

```r
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",
```

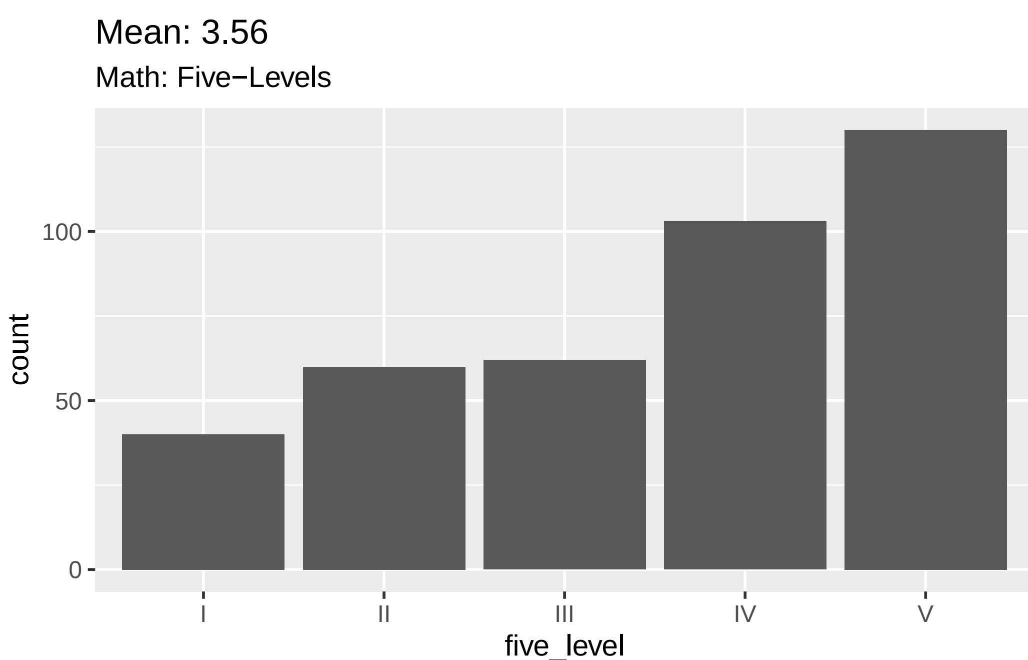
```
      G3<10 ~ "Fail"
    )) -> math2


portuguese <- 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)
```

Correlations & Distribution

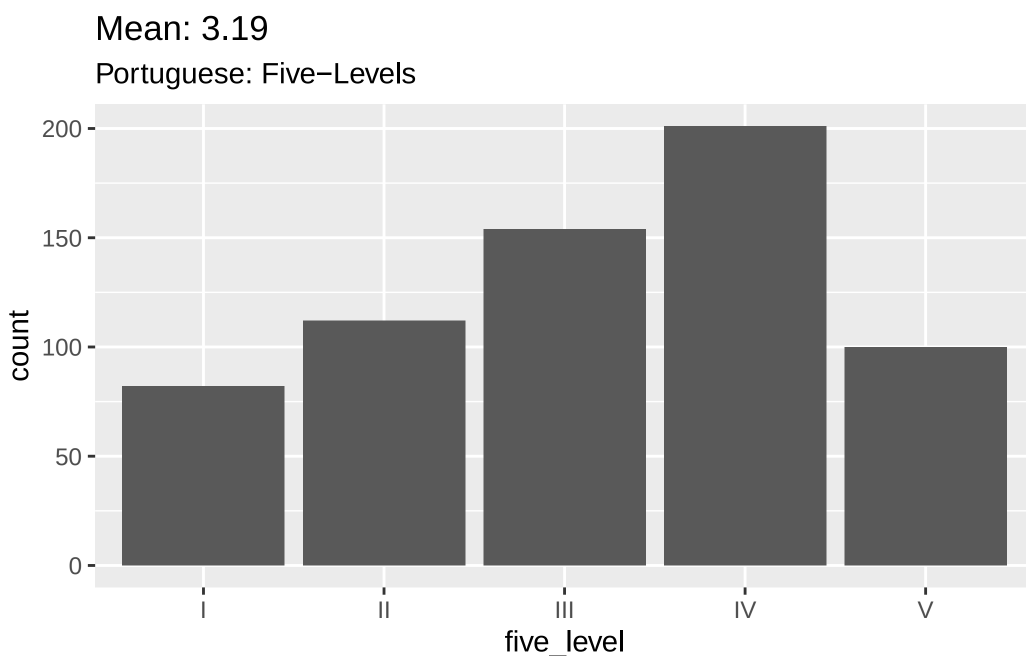
Means & Distributions

Here we exam visual distributions

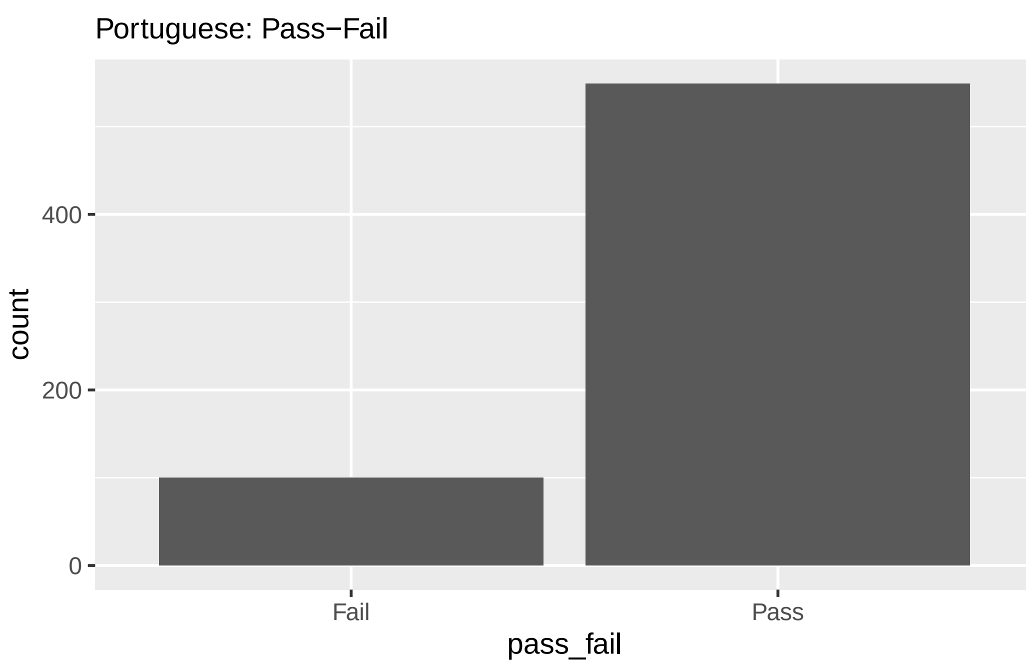
```
library(ggplot2)
ggplot(data=math2, aes(x=five_level))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(math2$five_level)), 2))) +
  labs(subtitle = "Math: Five-Levels")
```



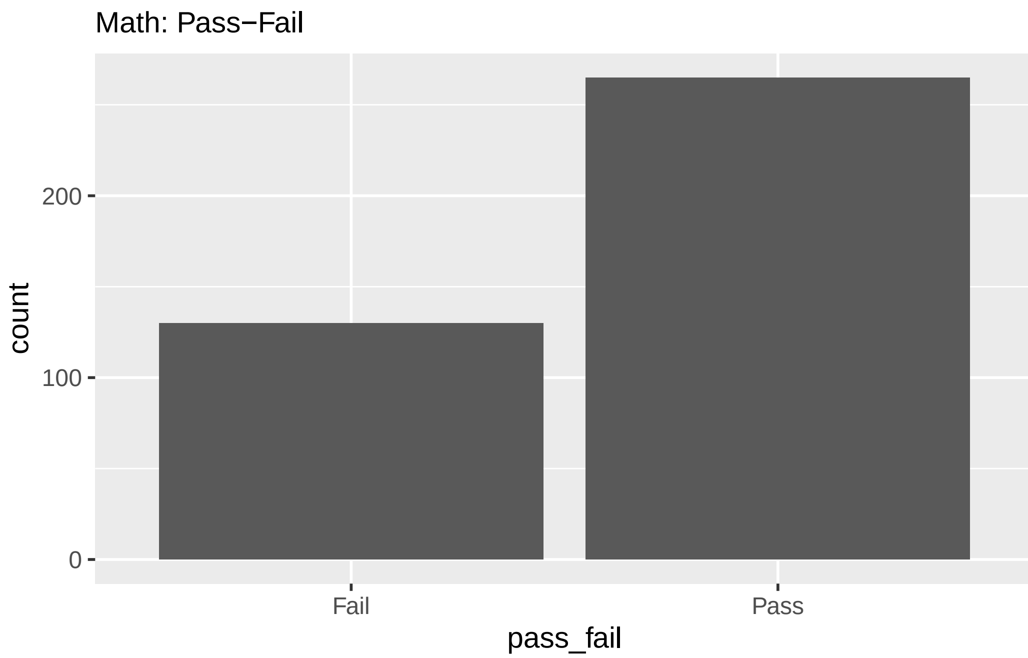

```
ggplot(data=portuguese2, aes(x=five_level))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(portuguese2$five_level)), 2))) +
  labs(subtitle = "Portuguese: Five-Levels")
```



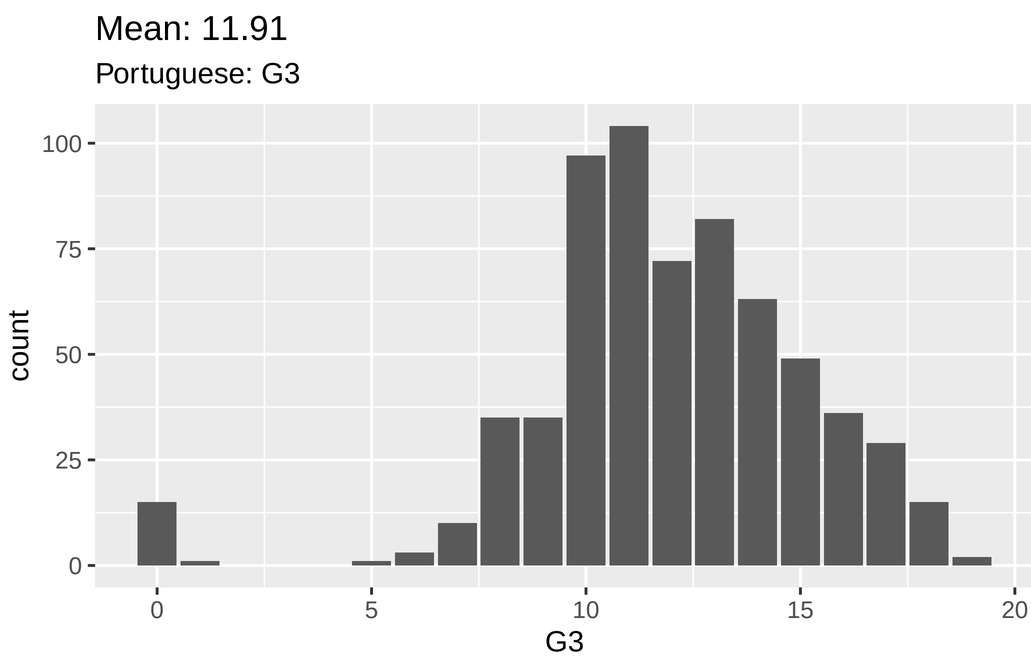

```
ggplot(data=portuguese2, aes(x=pass_fail))+
  geom_bar() +
  labs(subtitle = "Portuguese: Pass-Fail")
```



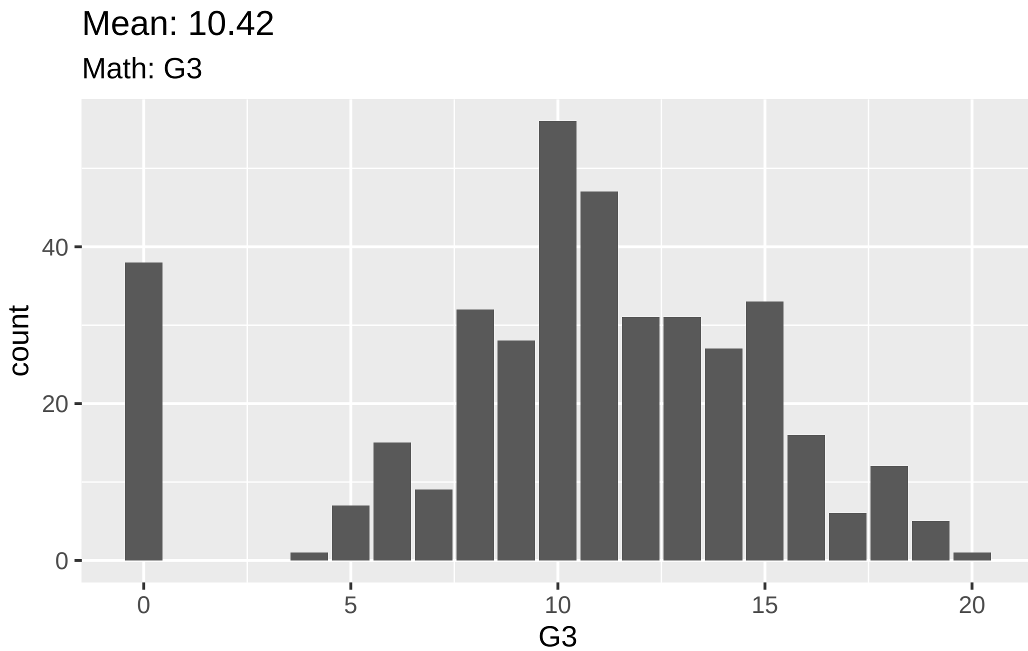
```
ggplot(data=math2, aes(x=pass_fail))+
  geom_bar() +
  labs(subtitle = "Math: Pass-Fail")
```




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

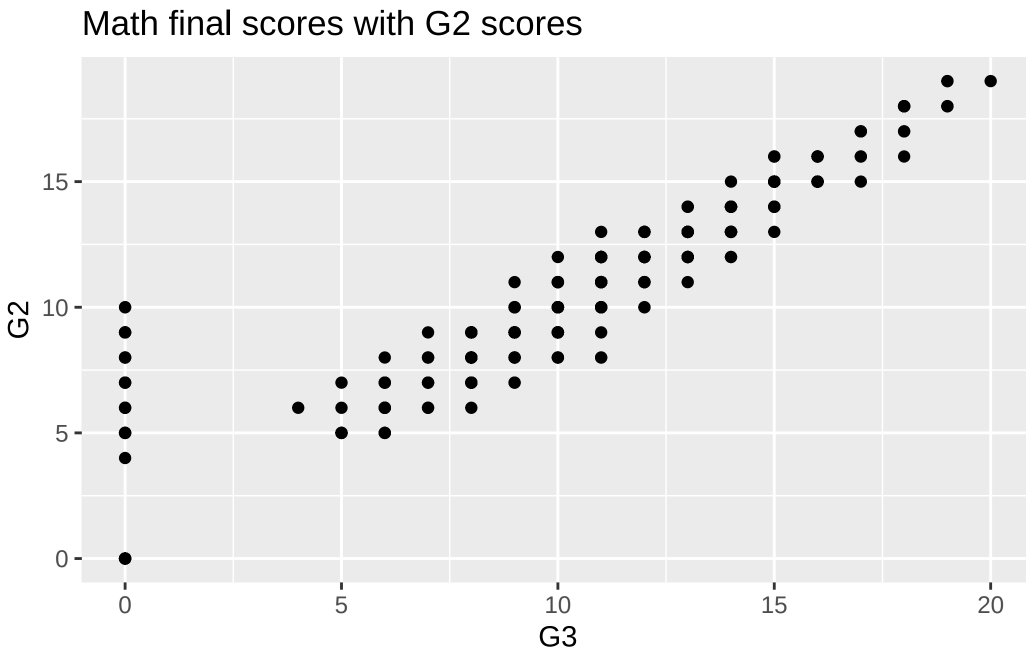



```r
ggplot(data=math2, aes(x=G3))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(math2$G3)), 2))) +
  labs(subtitle = "Math: G3")
```



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

Its 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)
```

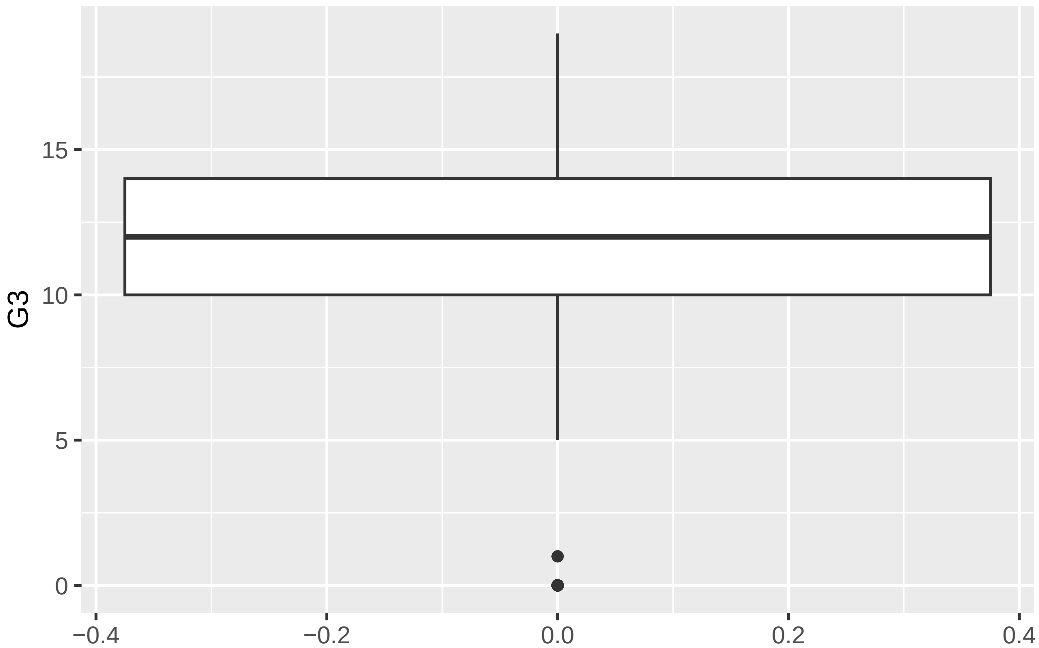
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   11.00   10.42   14.00   20.00
```

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

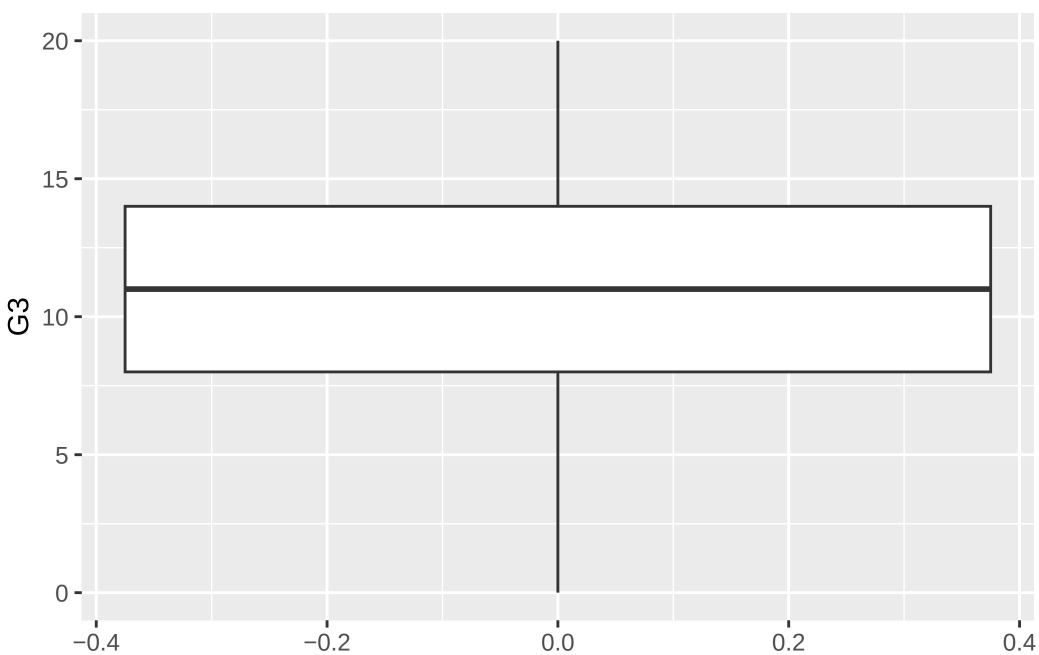
```
[1] 4.581443
```

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

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




```r
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 is typically indicates no correlation with other predictors. A VIF of 10 is generally considered too high. However, its 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.

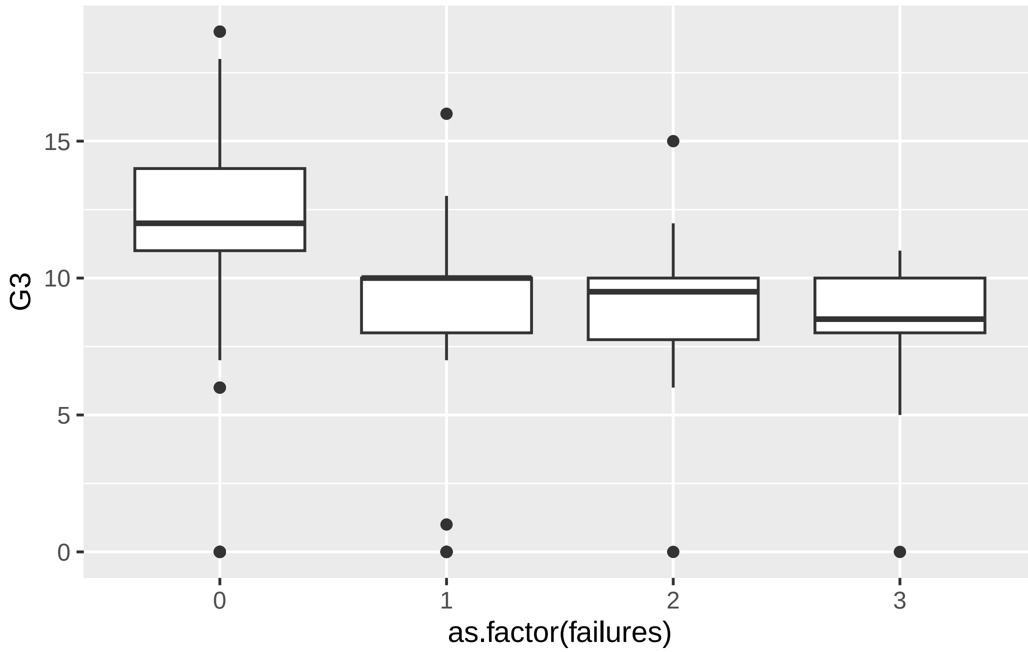
```
grades <- portuguese2[, c("G1", "G2", "G3")]

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
```

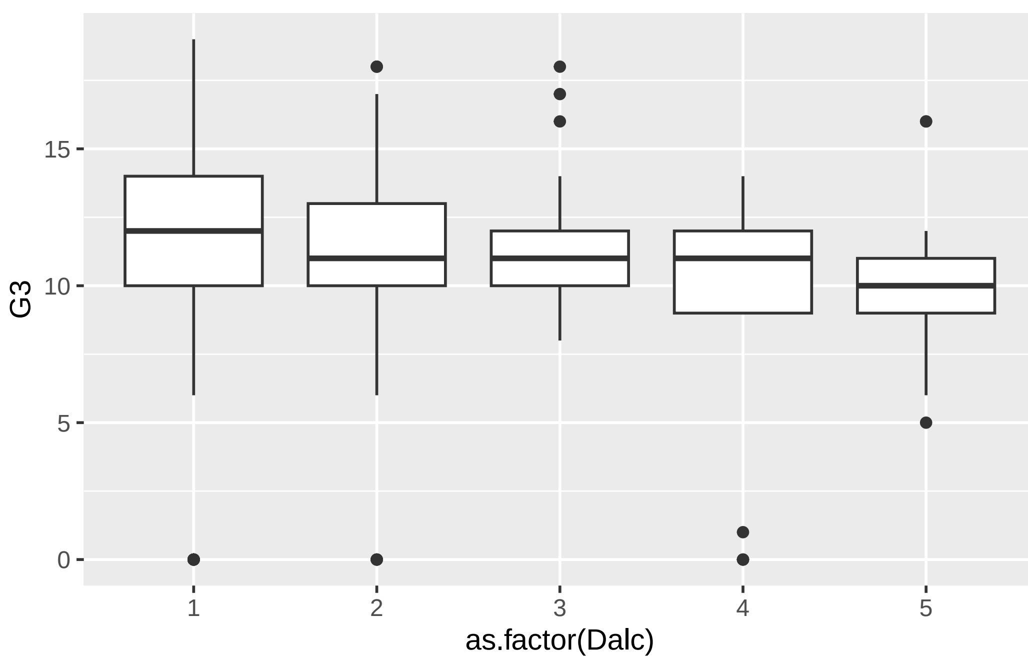
However, now that we have run a correlation matrix, it is displaying very strong correlation between our variables G1, G2, and G3. This confirms high collinearity among them, which would cause an increase in standard errors in our regression models.

Exploratory Graphs

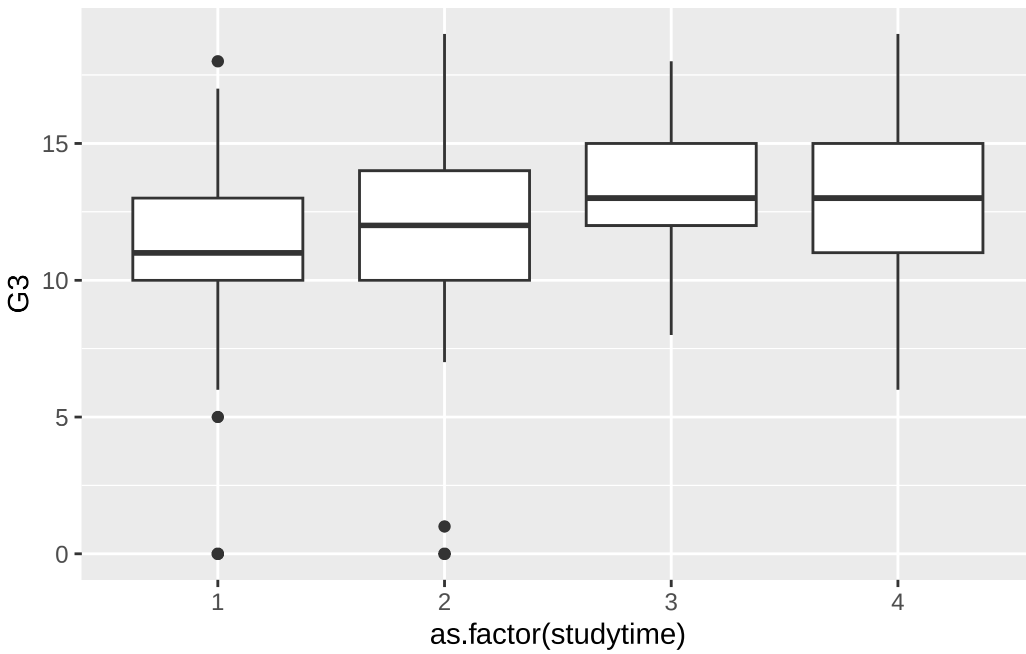
```
portuguese2 <- portuguese2 %>%
  mutate(across(c(pass_fail, romantic, internet, higher, nursery, activities, paid, famsup, s

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




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



```r
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 many of the models created in the paper connected with this data set. 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`>/=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 students choice to enroll in a class. Essentially, our model see's to predict a students 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 students 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."

To replicate the data we will not do an initial 20-80 split but run the 10-fold cross-validation on our models. This comes out to about 10% of the data utilized per fold, so 90% of our data will be used to train the models, and 10% will be used to test the data.

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

Part 2: Model Fitting

Multiple Linear Regression (MLR) - continuous 1-20

Prior to evaluating the performance of our models, we decided to create a "dummy" model for baseline comparison to efficiently evaluate the performance of our more complex models, such as Decision Tree or Random Forest models. While complex models provide deeper analysis on feature interactions and non-linearity considerations, they can be prone to over-fitting and require more computation and steps, which may lead to mistakes on our part. In order to prevent and check for mistakes such as these, we are creating this baseline model to compare performance, to assure our complex models are not performing *drastically* different.

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

```
library(mgcv)
```

```
Loading required package: nlme


Attaching package: 'nlme'


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

    collapse


This is mgcv 1.9-3. For overview type 'help("mgcv-package")'.
```

```
# we decided on a 90-10 split
set.seed(627)
train.pct <- 0.9
Z <- sample(nrow(portuguese), floor(train.pct*nrow(portuguese)))
portuguese.data <- portuguese[Z, ]
holdout.data <- portuguese[-Z, ]
# remember we are removing pass_fail and five_level since those are for categorical outcomes
MLR <- lm(G3 ~ . -five_level -pass_fail, data = portuguese.data)
```

Lets check and see how it is performing:

```
summary(MLR)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-8.7242 -0.5179  0.0268  0.6017  5.5840

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.855290   1.028470   0.832  0.40599
schoolMS         -0.198380   0.139067  -1.427  0.15430
sexM             -0.078799   0.127798  -0.617  0.53777
age               0.014339   0.052464   0.273  0.78471
addressU          0.100995   0.132616   0.762  0.44665
famsizeLE3        0.035963   0.125261   0.287  0.77414
PstatusT         -0.078465   0.175285  -0.448  0.65459
Medu             -0.091427   0.077051  -1.187  0.23591
Fedu              0.052332   0.069965   0.748  0.45480
Mjobhealth        0.203385   0.272777   0.746  0.45623
Mjobother        -0.090058   0.152370  -0.591  0.55474
Mjobservices      0.199580   0.186130   1.072  0.28408
Mjobteacher       0.256528   0.256653   1.000  0.31799
Fjobhealth       -0.471991   0.382160  -1.235  0.21734
Fjobother        -0.395790   0.230184  -1.719  0.08610 .
Fjobservices     -0.493370   0.241728  -2.041  0.04173 *
Fjobteacher      -0.630385   0.344924  -1.828  0.06816 .
reasonhome       -0.026885   0.144390  -0.186  0.85236
reasonother      -0.391026   0.184452  -2.120  0.03447 *
reasonreputation -0.160786   0.153193  -1.050  0.29439
guardianmother   -0.024747   0.135231  -0.183  0.85487
guardianother     0.277101   0.262576   1.055  0.29175
traveltime        0.134197   0.080182   1.674  0.09478 .
studytime         0.057879   0.070567   0.820  0.41247
failures         -0.252862   0.105493  -2.397  0.01687 *
schoolsupyes     -0.193024   0.188927  -1.022  0.30739
famsupyes         0.142095   0.115081   1.235  0.21746
paidyes          -0.245410   0.243216  -1.009  0.31341
activitiesyes     0.016989   0.114081   0.149  0.88167
nurseryyes       -0.115520   0.137857  -0.838  0.40242
higheryes         0.103865   0.196367   0.529  0.59707
internetyes       0.096685   0.139995   0.691  0.49009
romanticyes      -0.041976   0.115662  -0.363  0.71681
```

```
famrel           -0.025985   0.059759  -0.435  0.66386
freetime         -0.058359   0.057168  -1.021  0.30779
goout             0.002776   0.054697   0.051  0.95954
Dalc             -0.054413   0.078155  -0.696  0.48659
Walc             -0.043025   0.059950  -0.718  0.47326
health           -0.049970   0.039761  -1.257  0.20939
absences          0.015675   0.012707   1.234  0.21790
G1                0.115862   0.040531   2.859  0.00442 **
G2                0.890860   0.037380  23.833  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.281 on 542 degrees of freedom
Multiple R-squared:  0.8594,    Adjusted R-squared:  0.8487
F-statistic: 80.79 on 41 and 542 DF,  p-value: < 2.2e-16
```

Our model is showing an R squared of .8594, so our model is covering 86% of the variance in G3. This is relatively good. Our p-value indicates a significant model, lets check for MSE.

```r
predicted_MLR <- predict(MLR, newdata = portuguese.data)

mean((predicted_MLR - portuguese.data$G3)^2)
```

```
[1] 1.522204
```

```r
min(portuguese.data$G3)
```

```
[1] 0
```

```r
max(portuguese.data$G3)
```

```
[1] 19
```

The MSE for this model was ~1.52, which decently accurate. Within range of knowing the difference of a students passing or failing a class. If a student wanted strategic accuracy, a 1.5 difference in score prediction is reliable.

Logistic Regression (LR) - Binary Outcome

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

```r
# remember we are removing G3 and five_level since those are for other outcomes

# changing pass to 1 and fail to 0
library(dplyr)

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

LR <- glm(pass_fail ~ . -five_level -G3, data = portuguese_pass_fail,
          family = binomial)
```

Lets check and see how it is performing:

```r
summary(LR)
```

```
Call:
glm(formula = pass_fail ~ . - five_level - G3, family = binomial,
    data = portuguese_pass_fail)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -28.48599    6.72185  -4.238 2.26e-05 ***
schoolMS          -1.56028    0.67467  -2.313  0.02074 *
sexM               0.06716    0.72464   0.093  0.92616
age                0.64889    0.25848   2.510  0.01206 *
addressU          -0.22097    0.58161  -0.380  0.70399
famsizeLE3         0.06837    0.59887   0.114  0.90911
PstatusT          -0.63024    0.95561  -0.660  0.50956
Medu               0.02728    0.33673   0.081  0.93543
Fedu              -0.14445    0.31443  -0.459  0.64596
Mjobhealth        -0.70146    1.24378  -0.564  0.57277
Mjobother          0.26733    0.71880   0.372  0.70996
Mjobservices      -0.15240    0.93481  -0.163  0.87049
Mjobteacher        1.32605    1.34747   0.984  0.32507
Fjobhealth        -2.95089    2.10584  -1.401  0.16113
Fjobother         -2.20830    1.39691  -1.581  0.11391
Fjobservices      -2.00833    1.45201  -1.383  0.16662
```

```
Fjobteacher       -3.22446    2.05678  -1.568  0.11695
reasonhome         0.71920    0.77132   0.932  0.35111
reasonother        0.49300    0.78494   0.628  0.52996
reasonreputation   0.93148    0.88359   1.054  0.29179
guardianmother    -0.61906    0.70809  -0.874  0.38197
guardianother     -0.50708    1.24270  -0.408  0.68324
traveltime         0.21495    0.33175   0.648  0.51703
studytime          0.33811    0.37525   0.901  0.36757
failures          -0.20161    0.32473  -0.621  0.53470
schoolsupyes      -0.59862    0.87093  -0.687  0.49187
famsupyes          0.27838    0.53698   0.518  0.60417
paidyes           -2.00827    1.19552  -1.680  0.09299 .
activitiesyes      0.09891    0.55481   0.178  0.85851
nurseryyes         0.04745    0.61289   0.077  0.93830
higheryes          0.77565    0.63133   1.229  0.21922
internetyes       -0.22732    0.69190  -0.329  0.74250
romanticyes       -0.93117    0.62297  -1.495  0.13498
famrel            -0.12037    0.26828  -0.449  0.65367
freetime           0.17358    0.28528   0.608  0.54290
goout             -0.19376    0.26430  -0.733  0.46350
Dalc              -0.06010    0.33136  -0.181  0.85607
Walc              -0.25481    0.31150  -0.818  0.41334
health            -0.20705    0.21748  -0.952  0.34108
absences          -0.02104    0.05849  -0.360  0.71902
G1                 0.58453    0.20127   2.904  0.00368 **
G2                 1.96731    0.33793   5.822 5.83e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 512.07  on 583  degrees of freedom
Residual deviance: 132.68  on 542  degrees of freedom
AIC: 216.68

Number of Fisher Scoring iterations: 9
```

Our model is showing an residual deviance of 132.68, so our model is covering a large portion of the variance in pass_fail. This is good. Our AIC indicates a good fit.

Partial Least Squares - Continuous 1-20

```r
portuguese_clean <- portuguese.data[, sapply(portuguese.data,
                                             function(x) length(unique(x)) > 1)]

portuguese_lmF <- lm(G3 ~ . -five_level -pass_fail, data=portuguese_clean)
# matrix
portuguese.X <- model.matrix(portuguese_lmF)[, -1]
portuguese.pc <- prcomp(portuguese.X, scale=TRUE)
portuguese.pc
```

```
Standard deviations (1, .., p=41):
 [1] 1.9904215 1.6516559 1.4580622 1.3257497 1.2897365 1.2760604 1.2257861
 [8] 1.2076134 1.1911258 1.1464375 1.1080315 1.0977158 1.0764820 1.0578778
[15] 1.0162242 0.9974934 0.9821374 0.9755742 0.9548132 0.9478011 0.9111607
[22] 0.8925025 0.8783452 0.8706772 0.8468835 0.8228370 0.8211177 0.8036545
[29] 0.7872233 0.7788580 0.7450424 0.7237955 0.7019859 0.6782485 0.6552300
[36] 0.6291229 0.6012758 0.5161651 0.4126374 0.3535286 0.3148589

Rotation (n x k) = (41 x 41):
                          PC1          PC2           PC3         PC4
schoolMS          0.238332361 -0.045717355  0.3183701378 -0.09074875
sexM              0.033159783  0.300206653 -0.0738973676 -0.25312261
age               0.167084193  0.063059852 -0.1809879763  0.25709254
addressU         -0.172860927  0.061406205 -0.2134089683  0.16096771
famsizeLE3       -0.007409028  0.013240959 -0.0485838367 -0.06962434
PstatusT          0.001597842  0.045575945  0.1510633938 -0.14432233
Medu             -0.318037239  0.245179750 -0.0477087787  0.04533407
Fedu             -0.284190800  0.254935641 -0.0077174660  0.08804299
Mjobhealth       -0.138527075  0.068925257  0.0496333569  0.01006348
Mjobother         0.124397459 -0.232071619 -0.2242646601 -0.15840059
Mjobservices     -0.072314236  0.166459592  0.0949772417  0.16488306
Mjobteacher      -0.179382995  0.187883024  0.0002781272 -0.05387788
Fjobhealth       -0.110762210  0.076040574  0.0858136467  0.07658911
Fjobother         0.106537877 -0.219841211 -0.4222627956 -0.25858442
Fjobservices     -0.024772232  0.168956340  0.3684758284  0.16904241
Fjobteacher      -0.151680172  0.110460936  0.0549570943  0.07072669
reasonhome       -0.035187197 -0.006752480 -0.2263152685  0.05749473
reasonother       0.061927170  0.076482741  0.2747922768 -0.07704455
```

```
reasonreputation -0.127554055 -0.064406327 -0.0684933022 -0.01573658
guardianmother   -0.054764453 -0.007324826  0.0542257877 -0.22446865
guardianother     0.142700417  0.019637375 -0.2446848541  0.35293619
traveltime        0.211584589 -0.056228845  0.1090033675 -0.14464093
studytime        -0.162131203 -0.166174291 -0.0070558944  0.14855220
failures          0.253111985  0.103928261 -0.0644626275  0.24844256
schoolsupyes     -0.022813814 -0.062575911  0.0183487883  0.07349698
famsupyes        -0.078056355  0.028797601 -0.0010399468  0.19660940
paidyes          -0.017636623  0.131959284 -0.0029038745  0.01292467
activitiesyes    -0.071052962  0.123115084 -0.0191586210 -0.12319230
nurseryyes       -0.067868313  0.013792204  0.0904021890 -0.01165109
higheryes        -0.256325404 -0.098766105  0.0499528302 -0.12781430
internetyes      -0.167626775  0.155623310 -0.1959620195  0.02703176
romanticyes       0.066075421  0.013078419 -0.0457762883  0.19597194
famrel           -0.042099741  0.007601048  0.0221057704 -0.15000123
freetime          0.067743515  0.195040750 -0.0937614338 -0.18982690
goout             0.074196129  0.224198967 -0.1495421635 -0.19697434
Dalc              0.144928329  0.351376485 -0.1062132057 -0.09762165
Walc              0.110438135  0.365294173 -0.1233987536 -0.24046932
health            0.007141728  0.131438592 -0.0493860022 -0.08234479
absences          0.066646759  0.127813127 -0.2471691384  0.19272778
G1               -0.347000243 -0.161839925 -0.0903382220 -0.12080272
G2               -0.344205793 -0.165592118 -0.0952874667 -0.10604449
                           PC5           PC6          PC7           PC8
schoolMS          0.0434832732 -0.1572160933  0.157727330 -0.0228931210
sexM              0.0127724877  0.0881504520  0.058387185  0.1246938774
age              -0.1169573738 -0.1563818013  0.156755519 -0.2034036167
addressU          0.0308457710  0.2658081457 -0.033890915 -0.0104647109
famsizeLE3        0.0037332413  0.1173822415  0.268493525 -0.2689112874
PstatusT         -0.1928660753 -0.0446170288 -0.143094960  0.2758470010
Medu              0.1735739127 -0.1843774324  0.053466043 -0.0159999565
Fedu              0.1668635893 -0.1608873316  0.040100814  0.1185553256
Mjobhealth        0.0003943139 -0.2247593117 -0.036151096  0.1770522895
Mjobother        -0.0417227180  0.0263872824  0.088109687  0.2547434874
Mjobservices     -0.1425911296  0.2682629617 -0.283311258 -0.2475474922
Mjobteacher       0.2468313476 -0.1639891096  0.244769813 -0.1370094931
Fjobhealth        0.0410517126 -0.2060872435 -0.094929000  0.1896159759
Fjobother         0.1513503359 -0.1267566026 -0.123526916 -0.0617048885
Fjobservices     -0.3317629844  0.2882996775  0.105202758 -0.0107504882
Fjobteacher       0.2651891418 -0.1325919871  0.141236904  0.0522531509
reasonhome        0.1739200519  0.3747906335  0.036874365  0.2004231568
reasonother       0.0825232793 -0.0412260686  0.142797811  0.0965757520
reasonreputation -0.3177690647 -0.2455628177 -0.125227563 -0.1946428631
```

guardianmother	0.2506144986	0.1199877899	-0.154638620	-0.4521017206
guardianother	-0.1380372992	-0.1852148177	0.164645996	0.1770234958
traveltime	-0.0831644302	-0.1429169491	0.102782466	-0.0007841249
studytime	-0.2340909301	-0.1062351529	-0.039776645	-0.0345608508
failures	0.0570216569	-0.1312914267	-0.159800633	-0.0588426920
schoolsupyes	0.0877516001	0.1048172514	-0.351527696	0.1974564684
famsupyes	-0.0368872513	-0.1045855765	-0.215918835	0.1101158887
paidyes	0.1021907508	0.0288565122	-0.197297459	0.0921359756
activitiesyes	-0.1971161256	-0.1571829410	-0.191589429	-0.1442571383
nurseryyes	0.0859829527	-0.1024278269	-0.016324415	-0.2565746340
higheryes	-0.0352069206	0.0811605023	-0.017183837	0.1015758037
internetyes	-0.1397226642	0.0005017029	-0.066796569	-0.0112472902
romanticyes	-0.0290742792	-0.1626430575	0.161426952	-0.0909752610
famrel	-0.1373877238	-0.0760711195	-0.191731067	0.0087313282
freetime	-0.0863688490	-0.1795123055	-0.179263658	-0.0524606107
goout	-0.2099060529	-0.1076380223	-0.096096671	-0.0722147118
Dalc	-0.1443431760	0.0986891162	0.187530109	0.1212025684
Walc	-0.1588093697	0.1265880114	0.124881073	0.0475702340
health	0.1290499392	-0.0451917131	-0.140857411	0.1318333944
absences	0.0122790029	0.1347240744	0.008119456	-0.1597010340
G1	-0.2193959140	0.0271810566	0.218125222	0.0143166369
G2	-0.2157244583	0.0063054324	0.212139670	0.0020416767

	PC9	PC10	PC11	PC12
schoolMS	-0.040247584	0.033016439	0.0075791699	-0.181417483
sexM	-0.011357984	0.072169332	0.3392913555	0.159642179
age	0.107067180	0.189995153	0.0008270894	-0.089209438
addressU	-0.161489290	0.239573263	-0.1024957951	0.192961103
famsizeLE3	-0.329014914	-0.119960394	0.3652039563	-0.046618104
PstatusT	0.313049064	0.196517353	-0.1432174254	-0.008106820
Medu	0.012390488	0.031677928	-0.0059145524	-0.047189322
Fedu	0.021442260	-0.117144149	0.0083311291	-0.003480586
Mjobhealth	-0.432823586	0.265257157	-0.1076166834	-0.050809936
Mjobother	0.058163720	0.047754618	-0.0839761096	0.165186721
Mjobservices	0.005694122	-0.188458095	0.1278627337	-0.062287935
Mjobteacher	0.274433698	-0.011555179	0.0671266626	0.016374067
Fjobhealth	-0.467514031	-0.014998538	-0.0556403998	0.023563405
Fjobother	-0.024861059	-0.025461094	0.0279365842	-0.142796751
Fjobservices	0.078865071	0.160133790	0.0368935566	0.025591535
Fjobteacher	0.308817079	-0.245341048	-0.0629383056	0.243652331
reasonhome	0.082792055	0.144781151	-0.0204088771	-0.273059012
reasonother	-0.101674402	0.039076261	-0.1793808880	0.031389809
reasonreputation	-0.128070162	-0.199640347	0.0613294099	0.277092719
guardianmother	0.030092497	0.088510036	-0.2708133340	-0.080713638

```
guardianother     0.050192059 -0.044658750  0.2366609166 -0.053537060
traveltime        0.086212553 -0.310310877  0.0838545764 -0.187463004
studytime         0.038793248 -0.123093053 -0.0585546712 -0.162494110
failures          0.041638075  0.101729575  0.0770803088  0.039527447
schoolsupyes      0.020800641 -0.340181839 -0.0336825053  0.191218400
famsupyes         0.007390010 -0.223278055 -0.0952062563 -0.393979553
paidyes           0.094886147  0.058342318  0.1766768142 -0.445792480
activitiesyes     0.191361481  0.035406149  0.0872667439  0.088665167
nurseryyes       -0.085239208  0.089529644  0.0839812189 -0.190734539
higheryes         0.024617297 -0.138910400 -0.0173445418 -0.198941138
internetyes       0.096578517  0.180333616 -0.0186431419  0.017475332
romanticyes       0.081489487  0.166447075 -0.2639706199 -0.095656369
famrel            0.088373255  0.247617645  0.3194902760 -0.047651336
freetime          0.096489115  0.089765569 -0.0608693625  0.114789964
goout            -0.082897005 -0.006140091 -0.3202941308 -0.132074730
Dalc             -0.026614834 -0.235837967 -0.1193480689 -0.090507595
Walc             -0.130373981 -0.171640407 -0.1444925083 -0.082968231
health           -0.053655526  0.030094810  0.2989247315 -0.039096296
absences         -0.031834570 -0.173991860 -0.1736324184  0.043718241
G1                0.044860393 -0.007515132  0.0029925653 -0.086522777
G2                0.041979021 -0.003019317  0.0110873821 -0.097206509
                         PC13        PC14         PC15         PC16
schoolMS         -0.109600047  0.01743380  0.036746032  0.022693731
sexM              0.147973479 -0.13295651 -0.004037579  0.053322508
age               0.061577987  0.12998801 -0.082521844 -0.149427575
addressU         -0.089675294  0.03356181 -0.109177317  0.034429410
famsizeLE3       -0.221113554 -0.10425544 -0.056158778  0.009066564
PstatusT          0.206342397 -0.01694549  0.139380332 -0.041687943
Medu              0.057161245 -0.07404130  0.086024054 -0.099719926
Fedu             -0.036183681  0.01857368  0.069764623 -0.075201424
Mjobhealth        0.227695440 -0.05765355  0.159686096  0.286721693
Mjobother        -0.209486152 -0.20626261 -0.154217815 -0.223757919
Mjobservices      0.153205864  0.37131234 -0.053957907  0.109382715
Mjobteacher      -0.102724033 -0.13894178  0.061209069 -0.165724259
Fjobhealth       -0.096418674  0.01020884 -0.112892374 -0.103775080
Fjobother         0.194259925  0.09288563  0.034625616 -0.073830234
Fjobservices     -0.100866806 -0.15874380  0.027988809 -0.049220918
Fjobteacher      -0.070289282  0.15931711 -0.031528276  0.208308682
reasonhome       -0.195527574 -0.11498115  0.129705646  0.220713370
reasonother       0.046210145  0.28501509 -0.253820760 -0.335930239
reasonreputation  0.129260388 -0.24397484  0.096922459 -0.181421922
guardianmother    0.103647185 -0.07754337 -0.126057455 -0.112404758
guardianother    -0.034293431  0.05120887  0.005426405  0.039720417
```

```
traveltime        0.072670327  0.09468654  0.130063459  0.128844423
studytime        -0.102208805 -0.19006136 -0.320009465 -0.076040952
failures          0.146675538 -0.04008163  0.033179590  0.062791906
schoolsupyes     -0.239298095 -0.08710492 -0.104303362 -0.029135716
famsupyes        -0.133625398 -0.01026792  0.149456902 -0.079379712
paidyes           0.154408127 -0.25322126 -0.149727818 -0.120889941
activitiesyes     0.057770199 -0.31940763 -0.243619502  0.295308578
nurseryyes       -0.302940398 -0.14248717  0.114261019  0.062544928
higheryes         0.041633863 -0.04230945 -0.151560011 -0.041035719
internetyes       0.004663228  0.09368638  0.113257096 -0.297180194
romanticyes      -0.020856266 -0.06258691 -0.441136388  0.149204031
famrel           -0.298902452  0.25027600  0.062185331 -0.331590714
freetime         -0.381349505  0.13798706 -0.014176110  0.275133540
goout            -0.315029263  0.15568106  0.087417316  0.098503171
Dalc              0.005751837 -0.08209104 -0.117513376 -0.103794186
Walc              0.110034550 -0.07445116 -0.034365999 -0.028093932
health            0.046430515  0.26430394 -0.493535184  0.065888577
absences          0.086158806  0.05015528  0.123696313 -0.135321920
G1                0.061598408  0.18852728 -0.048845595  0.151284144
G2                0.065474641  0.20151993 -0.002122906  0.096691539
                         PC17         PC18         PC19         PC20
schoolMS          0.051410031  0.019447521 -0.108905554 -0.038338545
sexM              0.087992102  0.086098813  0.037829737 -0.011787792
age              -0.290256584  0.041127219  0.104783766 -0.128921470
addressU         -0.179715108  0.009144998 -0.050587540 -0.041734765
famsizeLE3        0.045121176 -0.021861878 -0.180979628  0.110595304
PstatusT         -0.288148558 -0.205139740  0.111029265  0.245665037
Medu              0.077955247  0.134599271 -0.029177234 -0.020475582
Fedu              0.051309908  0.103681895  0.118686732 -0.112534795
Mjobhealth        0.125753301  0.055817172 -0.042104809  0.003816575
Mjobother         0.109100180  0.059668390  0.248793126 -0.226685798
Mjobservices      0.041032310  0.141744852 -0.001111453  0.106899638
Mjobteacher      -0.187895286 -0.236931381 -0.236172179  0.141679857
Fjobhealth       -0.084570472 -0.166403077  0.155320631  0.106269227
Fjobother        -0.026845669  0.086391270 -0.127788020  0.182493355
Fjobservices      0.060281468 -0.041947378 -0.004744691 -0.122783122
Fjobteacher       0.086211649  0.063088974  0.154902783 -0.174810214
reasonhome        0.066056225 -0.001751497  0.055205667  0.122893395
reasonother      -0.009487591  0.383978749 -0.023964516  0.149969256
reasonreputation  0.058371028 -0.007195593  0.031171327 -0.126752051
guardianmother   -0.009602991 -0.069732352 -0.021440241 -0.176606207
guardianother    -0.036661459  0.116614457 -0.071493052  0.041197597
traveltime        0.122605715  0.013983026  0.176165178  0.165231297
```

studytime	-0.197876305	0.005995997	-0.137462501	0.013377851
failures	-0.154969062	0.109425618	0.002135703	-0.120314839
schoolsupyes	-0.145092294	0.259394167	-0.067104371	0.254580253
famsupyes	0.089075311	-0.333913093	-0.176046909	-0.153362414
paidyes	0.026531216	0.381923121	0.021766084	-0.274727240
activitiesyes	0.115152686	0.032166895	0.118073795	0.158160717
nurseryyes	-0.310815014	0.128786152	0.548700319	0.292788892
higheryes	0.138465787	-0.035654310	0.033110911	0.020700021
internetyes	0.174281993	0.029974376	-0.166347621	0.418591020
romanticyes	0.441492664	-0.038769800	0.013778989	0.266086939
famrel	0.299541691	-0.041466016	0.139648188	-0.070669573
freetime	-0.006259865	0.068235994	-0.286275580	-0.092889772
goout	-0.038903516	0.121269062	0.054784778	-0.177673916
Dalc	-0.095911324	0.013027387	-0.030549658	0.072967271
Walc	-0.127395839	-0.082337470	0.008101057	-0.023968490
health	-0.131786330	-0.443828205	0.145775227	-0.077824742
absences	0.258885676	-0.166404242	0.394517773	0.020668336
G1	-0.107411477	0.068894830	0.044868836	-0.067539301
G2	-0.107223460	0.116853115	0.069473445	-0.097962530

	PC21	PC22	PC23	PC24
schoolMS	-0.098115126	0.203815549	0.045205236	0.16619579
sexM	0.155485910	0.011222085	-0.068479564	-0.12882082
age	0.114022289	0.131021007	0.106379743	-0.25640953
addressU	-0.042538749	-0.132031019	-0.051168196	0.05796887
famsizeLE3	0.128705172	0.001683493	-0.281147742	-0.06700866
PstatusT	0.147444832	0.087530011	-0.152012464	0.01505100
Medu	-0.085074679	0.106555159	0.162680366	0.05254627
Fedu	-0.025756689	0.150597404	0.127317005	0.20070103
Mjobhealth	-0.107398584	-0.094451736	0.011942086	-0.34776580
Mjobother	-0.111610399	0.108246973	-0.092503616	0.11010306
Mjobservices	0.034922181	0.265274362	0.042884062	0.18090007
Mjobteacher	0.038414042	-0.215823022	0.149005035	-0.01440393
Fjobhealth	0.498522123	0.025510906	0.115299491	0.17430306
Fjobother	-0.092355516	0.018620287	0.016166205	0.05309114
Fjobservices	-0.176845357	-0.102875266	0.128274349	-0.09286778
Fjobteacher	0.098521876	0.160648728	-0.246938291	-0.19208604
reasonhome	0.230720438	0.119826488	0.122407461	-0.08770761
reasonother	-0.033416954	-0.351208034	-0.214732732	0.02044996
reasonreputation	-0.102272896	0.220398632	-0.006817478	-0.13449767
guardianmother	0.140949841	-0.044781665	0.054252871	-0.22655645
guardianother	-0.130326008	-0.198598664	0.135095520	0.06831749
traveltime	0.165970890	-0.136492719	0.212413470	-0.13710626
studytime	0.240446663	0.025063362	0.154860026	0.02493226

failures	0.067084999	-0.144305965	-0.040508405	-0.23502070
schoolsupyes	-0.040184884	-0.019595722	0.054268250	-0.40295832
famsupyes	-0.091833322	-0.127948942	-0.509060878	-0.08530025
paidyes	0.066711012	-0.015621718	-0.122973398	0.11255996
activitiesyes	0.088823673	-0.301164258	-0.127690577	0.25918644
nurseryyes	-0.311475649	0.044555320	-0.149448904	-0.04524203
higheryes	-0.276671413	-0.175590687	0.406558682	-0.16202226
internetyes	-0.141784041	0.152557185	-0.112509532	0.02195311
romanticyes	0.001062706	0.235142912	-0.072435981	-0.16121994
famrel	0.228431844	-0.039731337	0.058253119	-0.27391590
freetime	-0.022701208	-0.151169673	0.067480582	0.04303089
goout	-0.023934101	0.026739832	0.116515498	0.09034392
Dalc	0.045688033	0.082648656	-0.056266950	-0.13540753
Walc	-0.096160523	0.097539900	-0.004052077	-0.09488141
health	-0.309888936	0.041937613	0.032523023	-0.04163971
absences	0.046934920	-0.404839810	0.024492238	0.09096812
G1	0.055007870	-0.090095996	-0.131044381	-0.04658042
G2	0.083776824	-0.070033358	-0.122959244	-0.09230712

	PC25	PC26	PC27	PC28
schoolMS	0.067119532	-0.1035668818	-0.269161927	-0.0010125224
sexM	-0.335641662	-0.1235669739	0.101454927	0.1450434859
age	0.037223676	-0.0675092350	-0.281869385	0.0087278570
addressU	-0.295747060	0.3118188205	0.035206602	-0.3708632211
famsizeLE3	0.077428986	-0.1939033557	-0.254442435	-0.1770547906
PstatusT	0.076915835	-0.0676615600	-0.123986809	-0.0582809945
Medu	-0.109177788	-0.0266402455	-0.136371972	0.1726189923
Fedu	-0.022331921	-0.0614018278	-0.178320301	0.0797380227
Mjobhealth	-0.023957497	-0.0005943959	0.014533543	0.0608634460
Mjobother	-0.142358618	-0.2707894974	-0.074604385	-0.0244706944
Mjobservices	-0.102423288	-0.0217439350	-0.061916467	0.0978153401
Mjobteacher	0.109493303	0.1257207507	0.089829388	-0.0629072659
Fjobhealth	0.142003206	-0.1383989184	0.037197150	-0.1305357479
Fjobother	-0.064490050	0.1169710386	-0.061587303	0.1340212981
Fjobservices	0.033342056	-0.0968057019	0.071117211	0.0381074399
Fjobteacher	-0.113575228	0.0166670779	0.021978870	-0.1879751541
reasonhome	-0.000533315	-0.0522892354	-0.101212736	0.2384556914
reasonother	-0.132395979	0.0101767147	-0.051313240	0.1971641219
reasonreputation	0.096463952	0.0775318870	0.109526074	0.0697280965
guardianmother	0.010444382	-0.1503770821	0.042170736	-0.0113133880
guardianother	0.085221660	0.0976713202	0.059077907	-0.0176083832
traveltime	-0.270770778	-0.0592718215	0.247983915	-0.3265571397
studytime	-0.430901358	-0.0839601830	0.182363052	0.2244568644
failures	-0.143123074	-0.3121581059	-0.204257197	-0.1020132363

```
schoolsupyes     0.285291111 -0.0056658421 -0.128700474 -0.0144007126
famsupyes       -0.213669199  0.0098181219 -0.092402188  0.0750968109
paidyes          0.303137073  0.0221933231  0.292095434 -0.2614055742
activitiesyes    0.005085605  0.1359379066 -0.399884813 -0.0112832117
nurseryyes      -0.104763572  0.0144298104  0.176733746  0.1014504837
higheryes       -0.117263999 -0.1377445420 -0.303660374 -0.2106721081
internetyes     -0.042517274 -0.3928666368  0.121172311 -0.3162196582
romanticyes      0.104837705  0.1186086006  0.109403119 -0.0085148459
famrel          -0.002359616  0.3558489721 -0.081521334  0.0754608267
freetime         0.108744633 -0.3163765114  0.247643763  0.2229016760
goout           -0.041478652  0.0969457652 -0.107280030 -0.2790566039
Dalc             0.028512637  0.1296121828  0.015579072  0.1376512459
Walc             0.048135212  0.1438054194 -0.026840956  0.0002192873
health           0.073753588 -0.1101793401  0.086147351  0.0180109302
absences         0.184427610 -0.1819404939  0.044593508  0.1668011153
G1               0.145623622 -0.1061786553  0.027382031  0.0431236814
G2               0.201224938 -0.0874725218 -0.004084576  0.0333253149
                        PC29         PC30        PC31         PC32
schoolMS        -0.020844919 -0.074215648  0.14814291  0.213785773
sexM             0.138290149 -0.101586566  0.23217907 -0.238753132
age             -0.145920208 -0.020109333  0.15284496 -0.282128091
addressU         0.170525986 -0.149146872 -0.03394087 -0.030997753
famsizeLE3      -0.116419259 -0.236543579 -0.30224984 -0.152155229
PstatusT        -0.002776772 -0.372825987 -0.21060070 -0.219519429
Medu             0.100261988  0.120753016 -0.20747537 -0.079021406
Fedu             0.114128006  0.031449540 -0.32113819 -0.148647232
Mjobhealth      -0.282023479  0.003027667 -0.10277433  0.048629249
Mjobother        0.109171326  0.141979802 -0.04761868 -0.061185516
Mjobservices     0.120145311 -0.032861448  0.08491616 -0.036425139
Mjobteacher      0.101372227  0.018038017  0.04687649  0.044681045
Fjobhealth       0.176751614  0.078010413  0.27453583  0.005646111
Fjobother        0.073796594 -0.046929555 -0.11341950 -0.002505480
Fjobservices     0.031706323  0.167733615 -0.14251202 -0.101260681
Fjobteacher     -0.298351805 -0.153564607  0.13937893  0.155264081
reasonhome      -0.168540551  0.064654542  0.16573853 -0.051326814
reasonother     -0.102030817 -0.028529914  0.07546781 -0.143730068
reasonreputation -0.018817247 -0.141438518  0.18748656 -0.114848833
guardianmother  -0.042800475  0.225270762  0.11470746 -0.197553698
guardianother   -0.081311027  0.062345211  0.21926801 -0.214092738
traveltime       0.076999629  0.207454765 -0.18694635 -0.204351942
studytime       -0.299147951 -0.125119058 -0.21559713  0.200982936
failures         0.358083998  0.041903761 -0.13517597  0.423110479
schoolsupyes     0.024113923  0.129121446 -0.12467430 -0.118210169
```

```
famsupyes        0.151939344  0.087565407  0.14420847 -0.185578649
paidyes         -0.064170113 -0.169811353 -0.02412144  0.017190446
activitiesyes   -0.146696523  0.256957830  0.09944459  0.038385265
nurseryyes       0.087960430 -0.057717771  0.11179966  0.061972136
higheryes        0.107790134 -0.403631399  0.30460867  0.051036414
internetyes     -0.228731599  0.212308795  0.12557442  0.178487155
romanticyes      0.329026354 -0.116400624 -0.09942481 -0.108012049
famrel           0.019562820 -0.023975283 -0.08725974  0.216642268
freetime         0.132251664 -0.255435832  0.05038952 -0.055504226
goout           -0.161266718  0.029887930 -0.08641876 -0.119718988
Dalc             0.050811147  0.040279508  0.08466691  0.299895247
Walc             0.039766086  0.049089967 -0.03629935  0.131594249
health          -0.178953367  0.130384676 -0.12814039 -0.027856309
absences        -0.128496882 -0.263326270 -0.13297483  0.089821002
G1               0.175905543  0.121652726 -0.01384792  0.080432344
G2               0.150142210  0.115659208  0.01153883  0.060298913
                         PC33         PC34         PC35          PC36
schoolMS        -0.5629534439 -0.279140277 -0.121760438  1.993322e-01
sexM            -0.0354679150 -0.443103714 -0.173143134 -1.462153e-01
age             -0.1588161262  0.134163149 -0.013858079 -4.052246e-01
addressU        -0.4501491048  0.076433701 -0.021610995  1.196068e-01
famsizeLE3       0.1102441536  0.104752895  0.014797158  1.181095e-01
PstatusT         0.0091563335  0.024701597 -0.064654304  2.757131e-01
Medu            -0.1036009672 -0.006339043  0.004484547  1.379433e-01
Fedu            -0.0120970807  0.171097706 -0.095048780 -5.199441e-02
Mjobhealth      -0.0687923452  0.054952626 -0.084312431  2.051341e-02
Mjobother       -0.0193919804  0.086164813 -0.008042311  1.196532e-01
Mjobservices     0.0169000050  0.003393907  0.027008347  1.416723e-01
Mjobteacher     -0.0154235536 -0.192506754  0.217892228  5.536094e-05
Fjobhealth       0.0716415599  0.027316873  0.005934578  6.734159e-03
Fjobother       -0.0009821505 -0.003509859  0.007834446 -4.701739e-03
Fjobservices     0.0320141059 -0.004007247  0.052101072  1.027768e-02
Fjobteacher      0.0674490564  0.048869268 -0.009275635  1.086967e-01
reasonhome      -0.0920683785  0.058921038  0.452988965  1.840977e-01
reasonother      0.0158814057  0.035792297  0.331047445  9.397133e-02
reasonreputation -0.1804296713  0.027288798  0.438291376  2.266398e-01
guardianmother   0.0265632497  0.065803711 -0.289159034  3.905572e-01
guardianother    0.1702784920 -0.023552500 -0.281054384  5.225643e-01
traveltime      -0.2740505480  0.196543972  0.127071694  1.205795e-02
studytime       -0.0054226063 -0.153449996 -0.083999561  2.772200e-02
failures         0.1448613255 -0.074422284  0.238798020  1.814421e-01
schoolsupyes    -0.1813161746 -0.225112862 -0.122105399 -6.486280e-02
famsupyes       -0.0584882129 -0.020706698 -0.009863766 -7.505447e-02
```

```
paidyes          -0.0782797571  0.011718337  0.058313549 -6.253271e-02
activitiesyes    -0.0820339976  0.111835975  0.011792755 -5.578694e-02
nurseryyes        0.0680120949  0.029302744 -0.072189414  1.246626e-02
higheryes         0.1484492532  0.106814983 -0.014946174 -5.960523e-02
internetyes      -0.0406361715  0.006796176 -0.038423776 -6.391439e-02
romanticyes       0.0601849208 -0.128372888  0.028584929  1.113056e-03
famrel            0.0073047564  0.056262131 -0.088495404  9.563239e-02
freetime         -0.1139211866  0.309486071 -0.008848463 -4.471202e-02
goout             0.2182415517 -0.418643530  0.145210657  1.417293e-02
Dalc             -0.1212277623  0.310779342 -0.193497637  3.986885e-02
Walc              0.1208057402  0.012342087  0.040442161  5.589489e-02
health           -0.1105689384 -0.064397489  0.166688237  7.066719e-02
absences         -0.2168872173 -0.199216439 -0.076834345 -6.896782e-03
G1               -0.0390647305 -0.132140063  0.037852767  5.046113e-02
G2               -0.1248089193 -0.086263509 -0.060121797  6.537363e-03
                         PC37         PC38         PC39          PC40
schoolMS         -0.123536669 -0.011231905  0.035498067 -0.0377563275
sexM             -0.039177850  0.106438802 -0.015045352  0.0022057082
age               0.012315761 -0.113079993  0.050912472 -0.0725199492
addressU         -0.065670138  0.006989863 -0.015528585 -0.0135856274
famsizeLE3        0.066242754 -0.006040907 -0.007311901  0.0057608933
PstatusT          0.070019861  0.002597674 -0.011118212  0.0035237091
Medu              0.410647325 -0.326689428 -0.467985765 -0.0416383524
Fedu             -0.461952670  0.374642798  0.214369486 -0.0058975328
Mjobhealth        0.112020263 -0.086311114  0.393607585  0.0252344897
Mjobother         0.200553798 -0.244776691  0.371115355  0.0517851055
Mjobservices      0.195231323 -0.200501178  0.411689418  0.0589727788
Mjobteacher       0.131264065 -0.092959618  0.453717514  0.0762846603
Fjobhealth        0.027986063 -0.066988422  0.010513522 -0.0282849853
Fjobother        -0.018256794  0.086985828  0.156481019 -0.0464609019
Fjobservices      0.012567444  0.069731675  0.102244754 -0.0328090836
Fjobteacher      -0.002016395 -0.087421170  0.033350603 -0.0344565949
reasonhome       -0.095621160  0.072371862 -0.029380051  0.0208511450
reasonother      -0.069646906  0.024304885 -0.012346309  0.0202307009
reasonreputation -0.057195578  0.156446601 -0.030864531 -0.0054312998
guardianmother   -0.095394661  0.102245085  0.008498819 -0.0211924791
guardianother    -0.078357678 -0.012099779 -0.007721158 -0.0049770382
traveltime       -0.002467539 -0.046459324 -0.017479011 -0.0029902081
studytime        -0.100035230 -0.040115291  0.018256853  0.0286302999
failures         -0.052653951  0.100669531 -0.033370493  0.0324110679
schoolsupyes     -0.019439789 -0.062266586  0.014417164 -0.0328047964
famsupyes        -0.007283168 -0.031723617  0.015888453  0.0005086422
paidyes           0.038225928 -0.056924472  0.042988182 -0.0269211261
```

```
activitiesyes    -0.031649873 -0.019752128  0.013606515  0.0070361865
nurseryyes       -0.023089797 -0.001423544  0.008930329 -0.0010686509
higheryes         0.015011350  0.024386922 -0.033358896  0.0022010168
internetyes      -0.117472845  0.040835461 -0.047503554 -0.0112594908
romanticyes      -0.043883829  0.025134968  0.006509367  0.0288504733
famrel           -0.063567330 -0.067670034  0.016652965 -0.0362684767
freetime         -0.050913569 -0.109515410 -0.030951132 -0.0021633152
goout             0.182082220  0.209593409 -0.003791438  0.0266635571
Dalc              0.378349526  0.405539540  0.026713884 -0.0195312207
Walc             -0.475936476 -0.531650147 -0.039254738  0.0114273334
health            0.050050540  0.103849817 -0.074260515  0.0333588853
absences          0.018457664 -0.043208902  0.013508128 -0.0186950732
G1               -0.008934783  0.015954657  0.049134972 -0.6988045739
G2               -0.031260884  0.032809359 -0.070148859  0.6892736840
                         PC41
schoolMS          6.096892e-02
sexM             -2.448078e-03
age               2.278246e-02
addressU          4.114007e-02
famsizeLE3        8.882421e-03
PstatusT         -3.842567e-03
Medu              5.919281e-02
Fedu             -8.979443e-02
Mjobhealth       -7.430733e-02
Mjobother        -7.570811e-02
Mjobservices     -6.970905e-02
Mjobteacher      -8.443542e-02
Fjobhealth        2.802047e-01
Fjobother         6.301785e-01
Fjobservices      5.874770e-01
Fjobteacher       3.454312e-01
reasonhome       -1.148837e-02
reasonother      -2.659173e-02
reasonreputation -2.050855e-02
guardianmother   -2.884477e-02
guardianother    -1.103510e-02
traveltime       -2.991438e-02
studytime        -2.819711e-03
failures         -2.182786e-02
schoolsupyes      5.746018e-05
famsupyes         2.463426e-02
paidyes           1.837499e-02
activitiesyes     3.192286e-02
```

```
nurseryyes         1.407100e-02
higheryes         -4.278304e-03
internetyes       -1.894197e-02
romanticyes       -1.352088e-02
famrel            -3.742418e-02
freetime           2.509175e-02
goout             -1.420322e-02
Dalc              -3.152760e-02
Walc               1.580264e-02
health            -9.100684e-03
absences           2.262110e-02
G1                -7.024341e-02
G2                 6.125814e-02
```

```r
library(pls)
```

```
Attaching package: 'pls'


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

    loadings
```

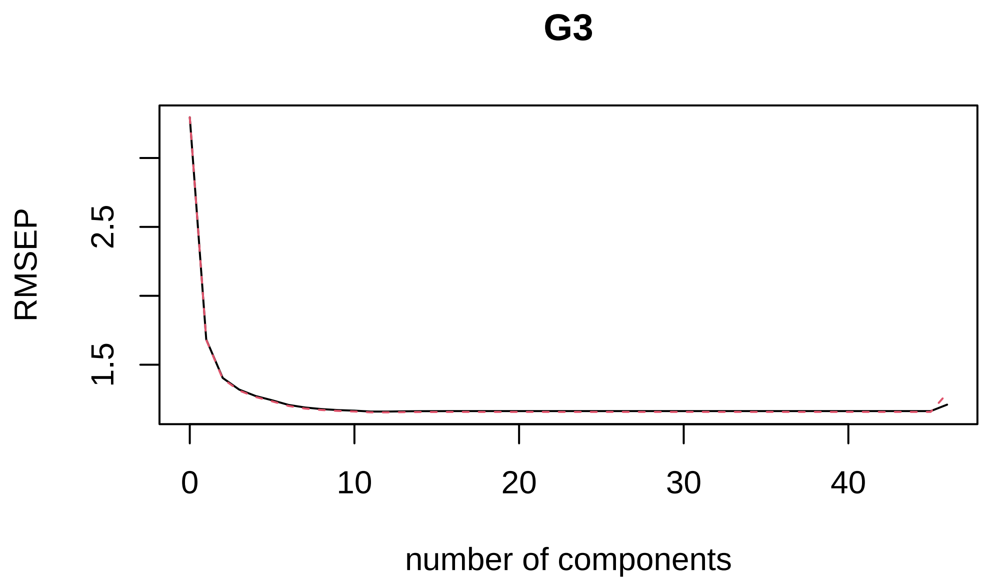
```r
PLS <- plsr(G3 ~ ., data = portuguese.data, scale = TRUE, validation = "CV")
summary(PLS)
```

```
Data:   X dimension: 584 46
    Y dimension: 584 1
Fit method: kernelpls
Number of components considered: 46

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           3.296    1.683    1.404    1.320    1.273    1.242    1.208
adjCV        3.296    1.681    1.398    1.313    1.266    1.234    1.201
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       1.189    1.178    1.171     1.167     1.160     1.160     1.162
adjCV    1.182    1.171    1.165     1.161     1.154     1.154     1.156
       14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
CV        1.163     1.163     1.163     1.163     1.163     1.163     1.163
```

```
adjCV      1.157      1.157      1.157      1.157      1.157      1.157      1.157
       21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps
CV         1.163      1.163      1.163      1.163      1.163      1.163      1.163
adjCV      1.157      1.157      1.157      1.157      1.157      1.157      1.157
       28 comps  29 comps  30 comps  31 comps  32 comps  33 comps  34 comps
CV         1.163      1.163      1.163      1.163      1.163      1.163      1.163
adjCV      1.157      1.157      1.157      1.157      1.157      1.157      1.157
       35 comps  36 comps  37 comps  38 comps  39 comps  40 comps  41 comps
CV         1.163      1.163      1.163      1.163      1.163      1.163      1.163
adjCV      1.157      1.157      1.157      1.157      1.157      1.157      1.157
       42 comps  43 comps  44 comps  45 comps  46 comps
CV         1.163      1.163      1.163      1.163      1.210
adjCV      1.157      1.157      1.157      1.157      1.292

TRAINING: % variance explained
    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X     10.40    14.84    18.88    22.15    25.00    27.66    29.82    32.27
G3    75.29    84.21    86.67    87.78    88.49    89.14    89.56    89.66
    9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X     34.92     37.17     39.18     40.76     42.78     45.26     47.41
G3    89.71     89.76     89.82     89.86     89.87     89.87     89.87
    16 comps  17 comps  18 comps  19 comps  20 comps  21 comps  22 comps
X      49.60     51.54     53.11     54.74     56.82     58.53     60.39
G3     89.87     89.87     89.87     89.87     89.87     89.87     89.87
    23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps
X      62.45     64.26     65.88     67.42     69.47     71.42     73.05
G3     89.87     89.87     89.87     89.87     89.87     89.87     89.87
    30 comps  31 comps  32 comps  33 comps  34 comps  35 comps  36 comps
X      75.03     76.95     78.61     80.29     81.87     83.70     85.30
G3     89.87     89.87     89.87     89.87     89.87     89.87     89.87
    37 comps  38 comps  39 comps  40 comps  41 comps  42 comps  43 comps
X      87.21     88.63     90.07     91.71     93.38     95.28     96.85
G3     89.87     89.87     89.87     89.87     89.87     89.87     89.87
    44 comps  45 comps  46 comps
X      98.26    100.00    100.02
G3     89.87     89.87     87.18
```

Lets plot it:

```r
validationplot(PLS)
```



It seems at 13 components the models variance for G3 doesn't get any higher. We prefer a model that's more simple, and given the additional components does not add predictive gain, we will stick with 13.

Let's put it to the test.

```
predicted_pls <- predict(PLS, newdata = portuguese.data, ncomp = 13)
# MSE
mean((predicted_pls - portuguese.data$G3)^2)
```

```
[1] 1.096997
```

A smaller MSE than our dummy model, this model is showing excellent prediction ability. An MSE of 1 is not likely to show any significant difference with grade predictability.

Model 2:

Recall we are only creating models to predict outcomes of portuguese and will not use G1 or G2 as predictors due to the clear prediction power of the variable and the colinearity. We will also be replicating the paper by using 20 runs of 10-fold cross-validation.

```
#library(rminer)
#portuguese2$pass_fail<-as.factor(portuguese2$pass_fail)

#K<-c("kfold", 10)
#DT<-mining(pass_fail~.-G1-G2-G3-five_level , data=portuguese2, model="dt", Runs=20, method=K
#print(mean(DT$error))
#savemining(DT,"DT-results")
```

This below was just trying to make more of them categorical, it doesn't seem to make a good difference in the outcome though so probably delete later.

```
#portuguese3 <- portuguese2 %>%
#  mutate(across(c(pass_fail, romantic, internet, higher, nursery, activities, paid, famsup,
```

```
#rpart(pass_fail~.-G1-G2-G3-five_level , data=portuguese2, model="rpart")
```