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

```
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 becor
```

```
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         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
```

```
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 |>
  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",
```

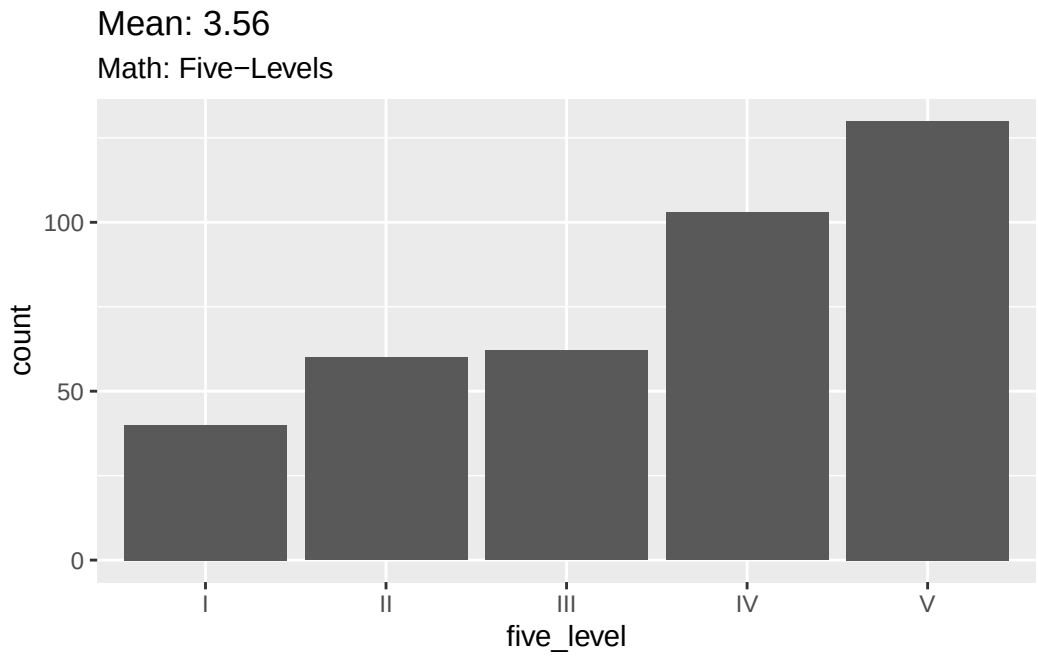
```
      G3<10 ~ "Fail"
    )) -> math2


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)
```

Correlations & Distribution

Means & Distributions

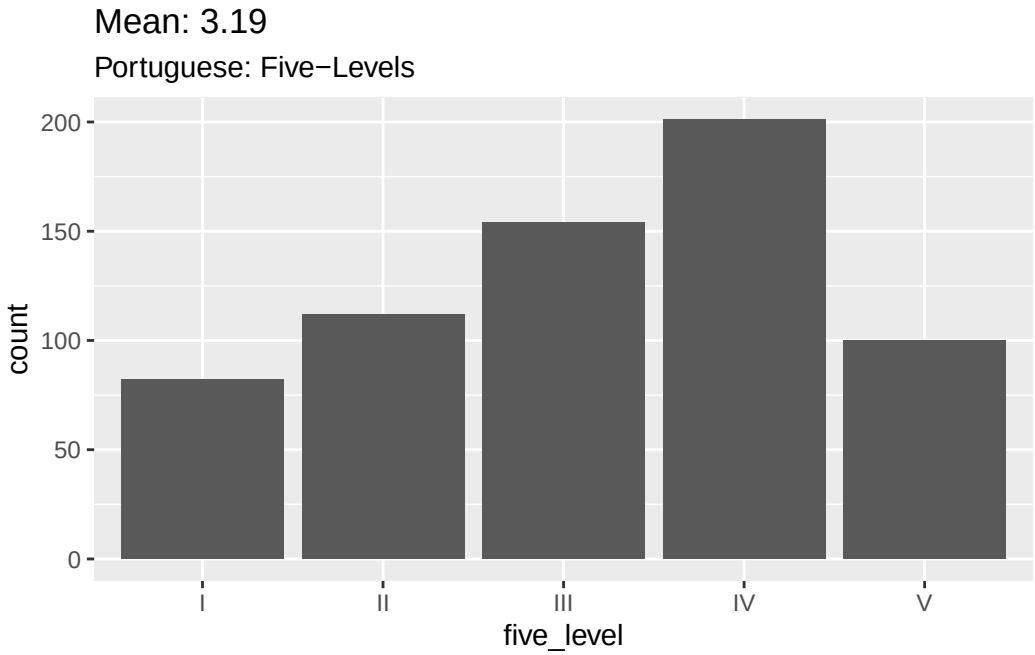
Here we exam visual distributions

```
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")
```

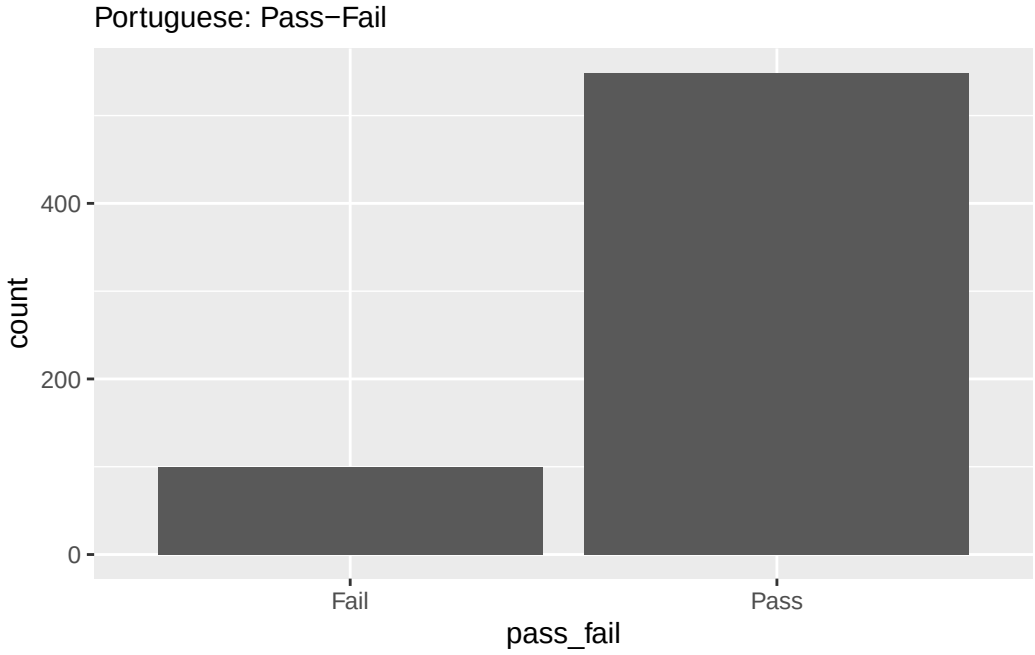



```
ggplot(data=portuguese2, aes(x=five_level))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(portuguese2$five_level)), 2))) +
  labs(subtitle = "Portuguese: Five-Levels")
```

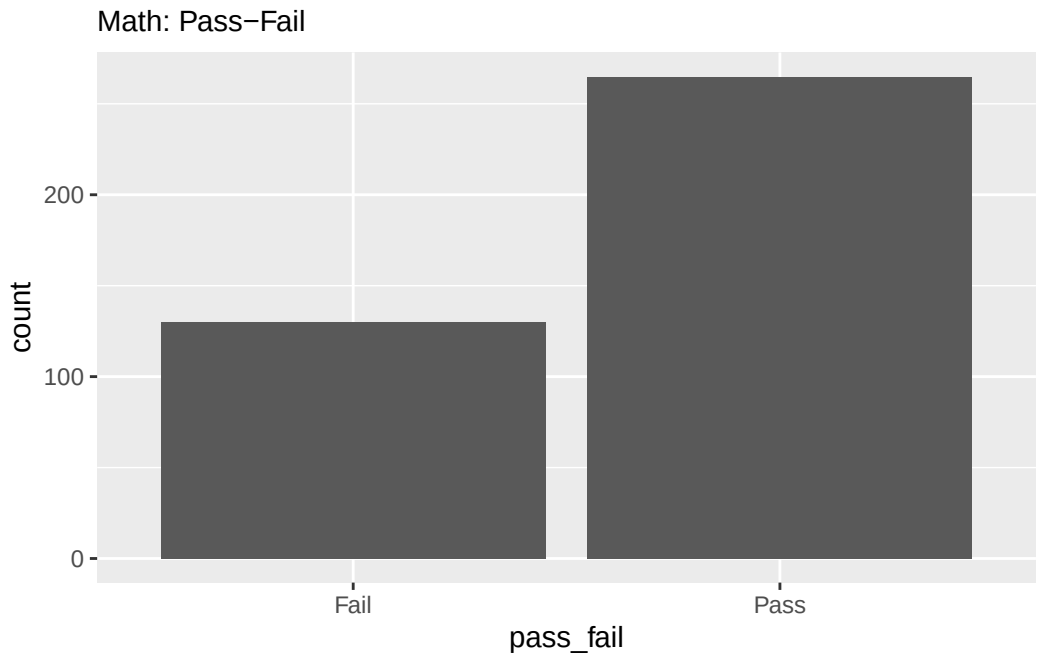
Mean: 3.19



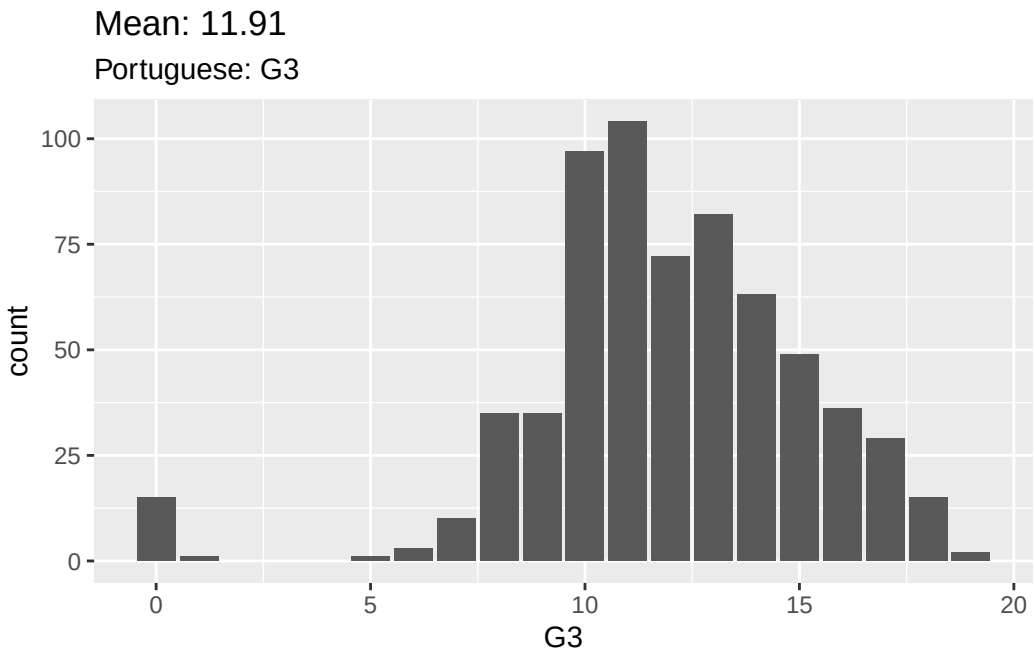

```
ggplot(data=portuguese2, aes(x=pass_fail))+
  geom_bar() +
  labs(subtitle = "Portuguese: Pass-Fail")
```



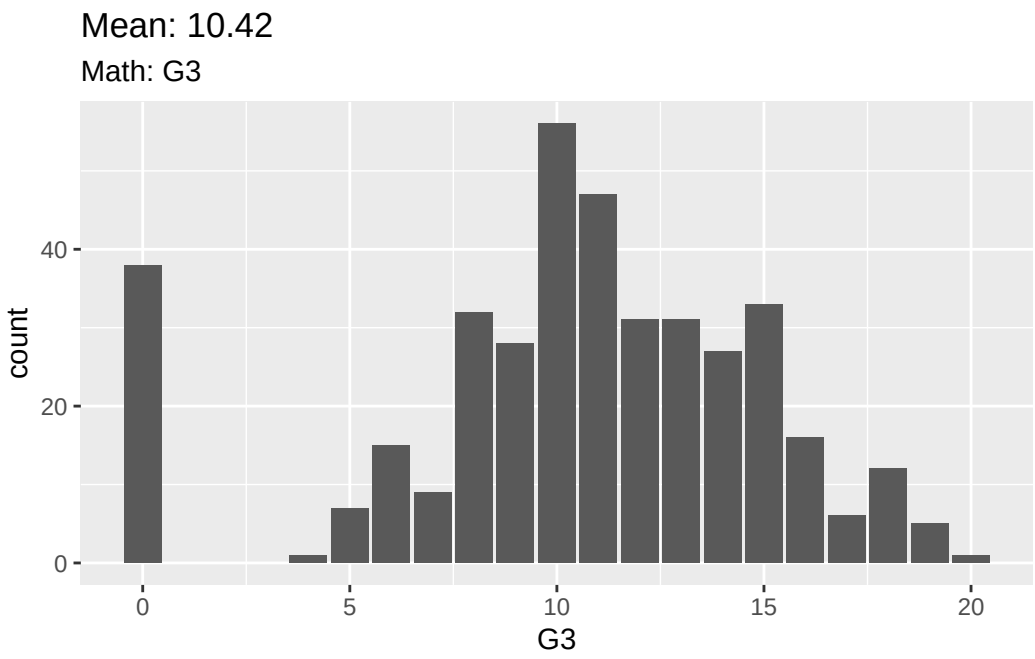
```
ggplot(data=math2, aes(x=pass_fail))+
  geom_bar() +
  labs(subtitle = "Math: Pass-Fail")
```




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

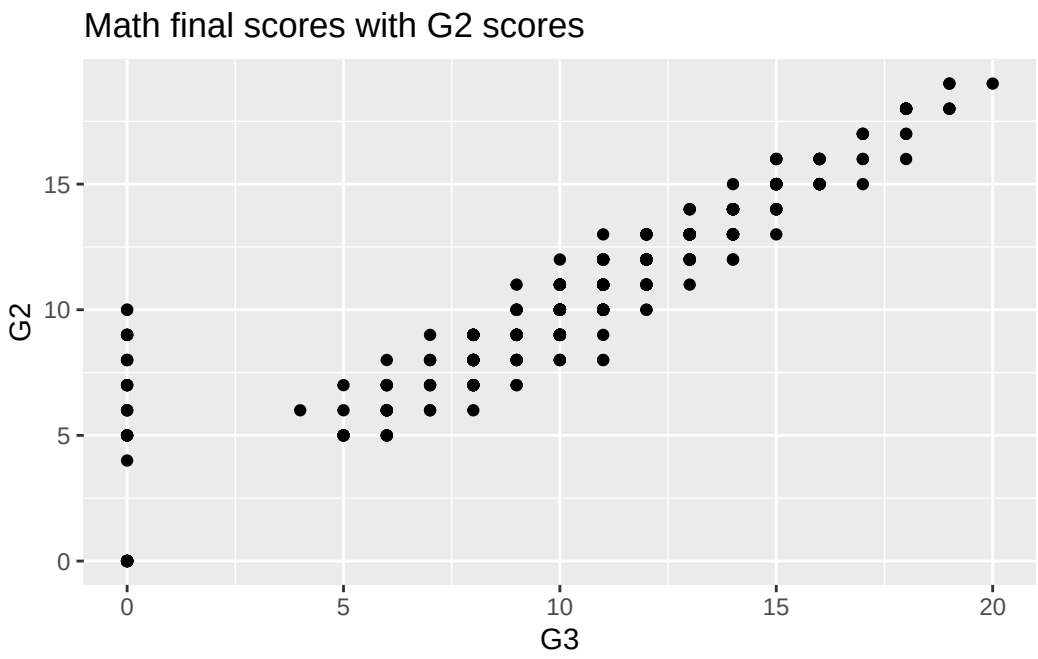



```
ggplot(data=math2, aes(x=G3))+
  geom_bar() +
  ggtitle(paste("Mean:", round(mean(
    as.numeric(math2$G3)), 2))) +
  labs(subtitle = "Math: G3")
```



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)
```

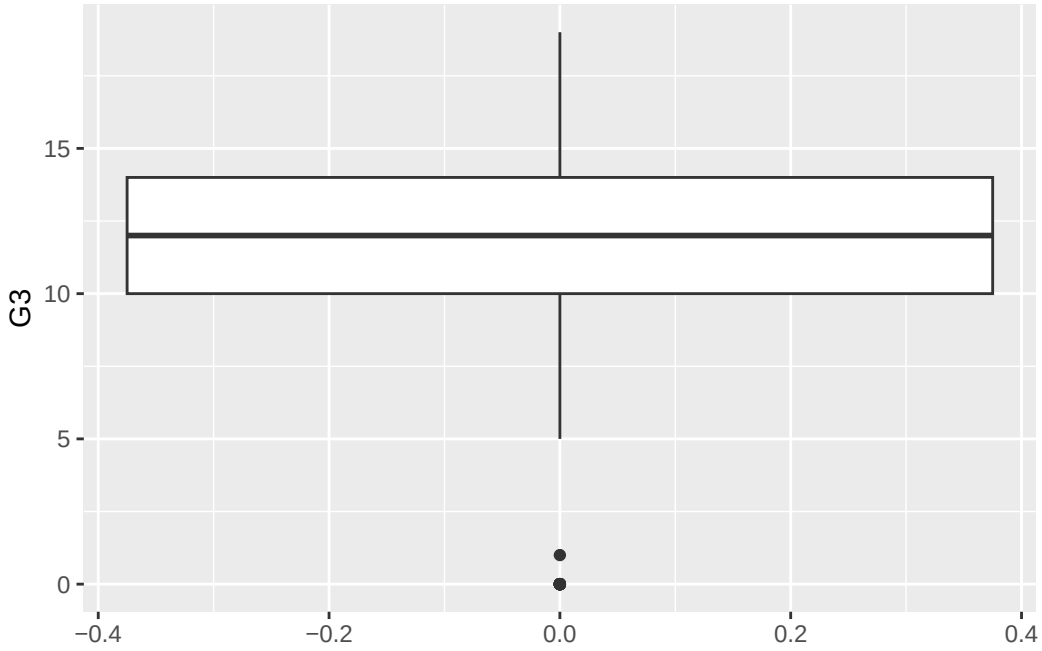
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   11.00   10.42   14.00   20.00
```

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

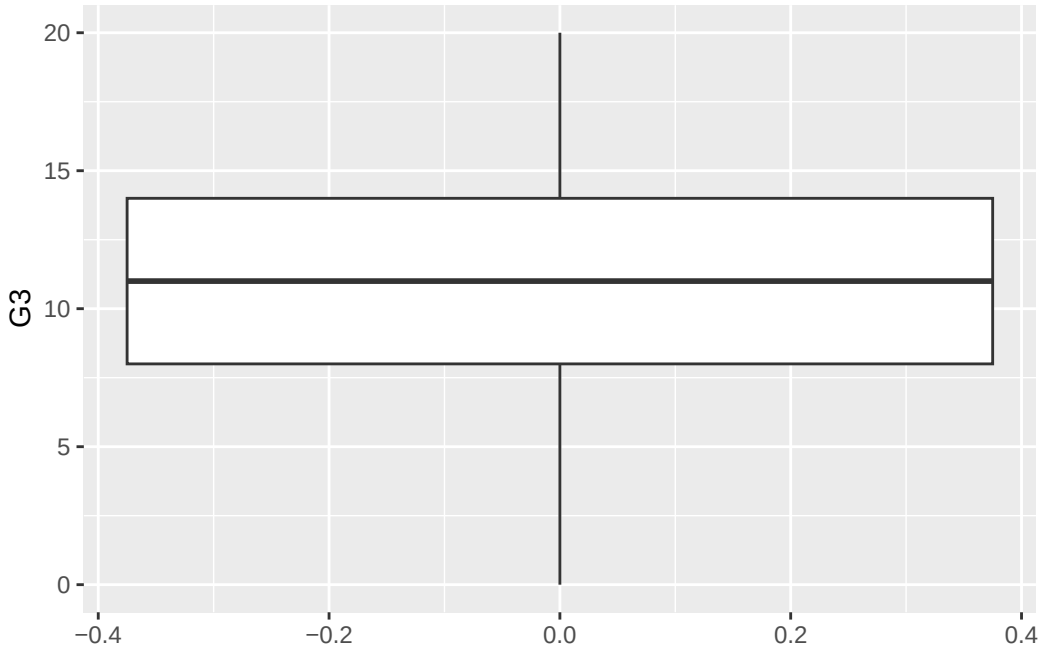
```
[1] 4.581443
```

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

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




```
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.

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". My colleague and I want to take a slightly different approach where we utilize this data to create a model signifying a different usage of the model. 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.

In order to differentiate our model from theirs while still replicating 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 to class enrollment so that a user could predict their grade before enrollment, potentially steering students towards or away from the class given their predicted outcome score (if our model proves statistically significant). This would mean removing variables G1 and G2 given their grades would not be recorded if the class was not yet taken. We would also be required to remove `studytime, schoolsup`, `paid`, `absences` as each of these variables are inclusive of information that can only be obtained while being currently enrolled in the class.

We understand the choice of the authors to use an A-B-C subset method, however, given our full removal of G2 and G1, we decided to solely select subset (C). Exclusion of G1 and G2 would indicate a model that can predict a students grade prior to finishing trimester 1.

We will prioritize replicating the models used to predict outcomes for the Portuguese course (which has more observations, 649 v. 395) for time constraint, and only reproduce models with input C, as we want to view prediction power without the utilization of G1 and G2 grades, due to our differing model usage.

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 will remove variables `studytime, schoolsup`, `paid`, `absences`, `G1`, and `G2`, and examine solely the Portuguese course data for time constraint reasons.