# Final Project: Data Insights to Improve school Education System Data Acquisition and Management (DATA 607)

Alexis Mekueko

The City University of New York School of Professional Studies
email: alexis.mekueko08@login.cuny.edu

12/9/2020

## Introduction

- Many students failed in school not because of thier intelligence.
- There are numerous factors that contribute to students success.
- In other words, students success in school relies upon on the ability of the school education system to take appropriate measures on these factors.
- These factors are : weekly studying time, extra-curricular activities, travel time to school, family educational support, student desire to pursue higher education, companionship, parents'job type, etc.
- Therefore, in this project, we interested in studying these factors to determine any correlation that could lead to students failure in a taken course.
- If none, then we would like to determine the factors which contribute to success.
- This is done in order for the school education system to keep track of success and to improve the factors that negatively impact students success.

Github Link: https://github.com/asmozo24/DATA607_Final_Project

Web link: https://rpubs.com/amekueko/702297

## Research question

- Do you students from Gabriel Pereira (GP) school do better in Math course than those from Mousinho da Silveira (MS) school?
- We could also explore the corelation between factors time and students performance. We could also verify a popular assumption out there.
- For instance, there are some studies out there suggesting that the amount of study time likely affects students' performance. Let's verify this assumption in this project.
- The question being, do students studying at least 10hrs weekly do well in Math course than those spending lesser time?

## Data Acquisition

### Data collection

Data is collected or made available by archive.ics.uci.edu: The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community

for the empirical analysis of machine learning algorithms. The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine. The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman, and this project is in collaboration with Rexa.info at the University of Massachusetts Amherst. Funding support from the National Science Foundation is gratefully acknowledged.

### Data source

We found some interesting dataset from -> data source: https://archive.ics.uci.edu/ml/machine-learning-databases/00320/. This data is about a study on students(395) taking math or/and portuguese language course. Each case represents a student at one of the two schools ("GP" - Gabriel Pereira or "MS" - Mousinho da Silveira). There are 395 observations in the given dataset. The data is pretty rich with a txt file that described all variables in the data. therefore there is no need to rename the column. The orignal data format is comma delimited and rendering from R was not easy. So, we used excel with one attemp to fix it. We are interested in the student taking Math course. with 33 variables.

- Data available –> https://github.com/asmozo24/DATA606_Project_Proposal

Using R to acquire data

```
# load the text file which has the description of all the variable.
variable_details <- read.delim("https://raw.githubusercontent.com/asmozo24/DATA607_Final_Project/main/s
student_math <- read.csv("https://raw.githubusercontent.com/asmozo24/DATA607_Final_Project/main/student
student_portuguese <- read.csv("https://raw.githubusercontent.com/asmozo24/DATA607_Final_Project/main/s
```

Using SQL to acquire data # {r, connection to sql} # # # # establishing the connection to SQL server to access db # con <- dbConnect(odbc(), # # server type # Driver = "SQL Server", # #server name # Server = "ATM\\ATMSERVER", # # this is one of the db I want to import # Database = "Final_ProjectDB", # UID = "Alex", # # password required # PWD = rstudioapi::askForPassword("Database password"), # ort = 1433) # # #dbListFields(con, "student_math") # #student_math <- dbReadTable(con, "student_math") # #barplot(table(Customer_Location$Country, color = rainbow())) # #

## Data Preparation / Data Wrangling

**Cleaning data** What is the structure of data?

```
glimpse(student_math)
```

```
## Rows: 395
## Columns: 33
## $ school     <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "G...
## $ sex        <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "...
## $ age        <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, ...
## $ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "...
## $ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", ...
## $ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "...
## $ Medu       <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3,...
## $ Fedu       <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2,...
## $ Mjob       <chr> "at_home", "at_home", "at_home", "health", "other", "ser...
## $ Fjob       <chr> "teacher", "other", "other", "services", "other", "other...
```

```
## $ reason     <chr> "course", "course", "other", "home", "home", "reputation...
## $ guardian   <chr> "mother", "father", "mother", "mother", "father", "mothe...
## $ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1,...
## $ studytime  <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1,...
## $ failures   <int> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3,...
## $ schoolsup  <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no",...
## $ famsup     <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "ye...
## $ paid       <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes...
## $ activities <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", ...
## $ nursery    <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "...
## $ higher     <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", ...
## $ internet   <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "ye...
## $ romantic   <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "...
## $ famrel     <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5,...
## $ freetime   <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5,...
## $ goout      <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5,...
## $ Dalc       <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2,...
## $ Walc       <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4,...
## $ health     <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5,...
## $ absences   <int> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, ...
## $ G1         <int> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 1...
## $ G2         <int> 6, 5, 8, 14, 10, 15, 12, 5, 18, 15, 8, 12, 14, 10, 16, 1...
## $ G3         <int> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, ...
```

```
#str(student_portuguese)
#print("Data frame is composed of character, boolean and numerical.")
view(student_math)
#head(student_math)
#print("Let's convert all chr type to factor and int type to numeric")
#student_math
#summary(student_math)
#dim(student_math)
dim(student_portuguese)
```

```
## [1] 649  33
```

```
#na1 <- is.na(student_math) # checking if there is a missing data in the dataset, return is yes
sum(is.na(student_math)) # file to big, checking the sum of all missing data (return is 0 missing data)
```

```
## [1] 0
```

```
sum(is.na(student_portuguese))
```

```
## [1] 0
```

## Explore Data

Let's take a look at the data frame...

```
# let's create a dummy variable for the purporse of data visualization
VarFunc <- function(df, x){
  for (i in 1:(nrow(df))) {
  if (x[i] < 20) {
    df$x[i] <- "All"
  }
  }
  return(df$x)
}

Var2Func <- function(df, x){
  for (i in 1:(nrow(df))) {
  if (x[i] < 20) {
    df$x[i] <- i
  }
  }
  return(df$x)
}


student_math$Var <-VarFunc(student_math, student_math$G1)
student_portuguese$Var <-VarFunc(student_portuguese, student_portuguese$G1)
student_math$Var2 <-Var2Func(student_math, student_math$G1)
#head(student_math)
#head(student_portuguese)
```

- Amount the 33 variables in the data frame, there are 03 variables (G1, G2 and G3) which represent the students's grades.
- These 03 variables are interesting as there are measures of students performances in the registered courses.

  G1: first period grade (numeric: from 0 to 20) G2: second period grade (numeric: from 0 to 20) G3: final grade (numeric: from 0 to 20)

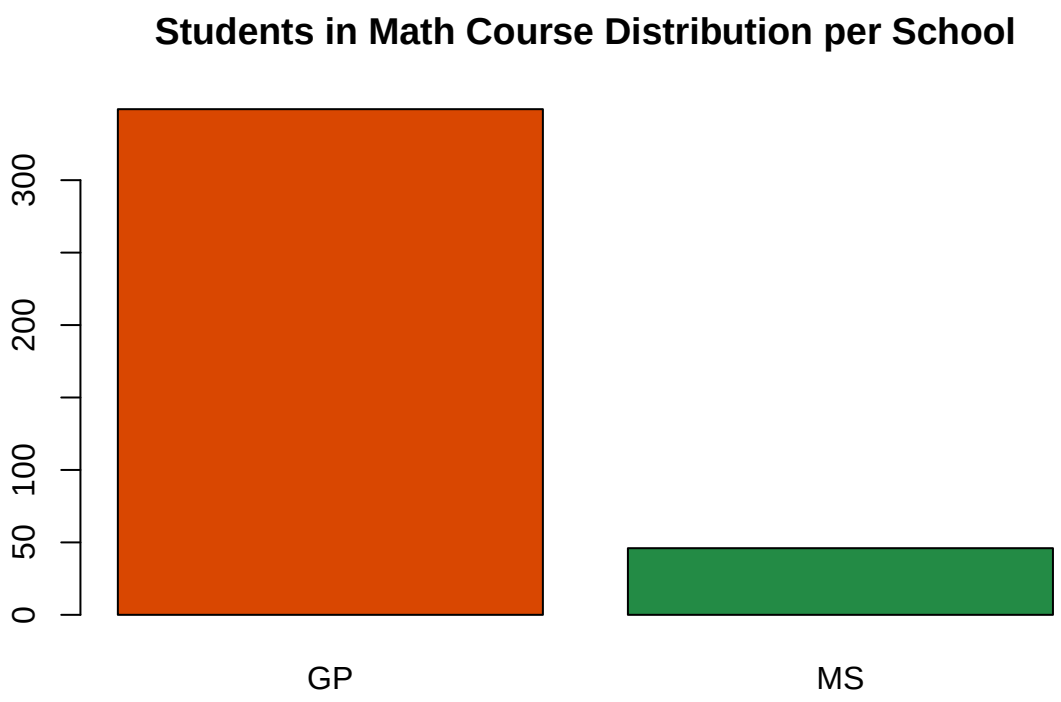
## Visualize students distribution per school in Math Course

```
describe(student_math$G3)
```

```
## student_math$G3
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      395        0       18    0.992    10.42    4.992      0.0      5.0
##      .25      .50      .75      .90      .95
##      8.0     11.0     14.0     15.6     17.0
##
## lowest :  0  4  5  6  7, highest: 16 17 18 19 20
##
## Value          0     4     5     6     7     8     9    10    11    12    13
## Frequency     38     1     7    15     9    32    28    56    47    31    31
## Proportion 0.096 0.003 0.018 0.038 0.023 0.081 0.071 0.142 0.119 0.078 0.078
##
## Value         14    15    16    17    18    19    20
```

```
## Frequency     27    33    16     6    12     5     1
## Proportion 0.068 0.084 0.041 0.015 0.030 0.013 0.003
```

```r
#summary(student_math$G3)
#print("Students taken Math course distribution from each school are: 88.4% students for Gabriel Pereir

barplot(table(student_math$school), main = "Students in Math Course Distribution per School", xlab = "G
```

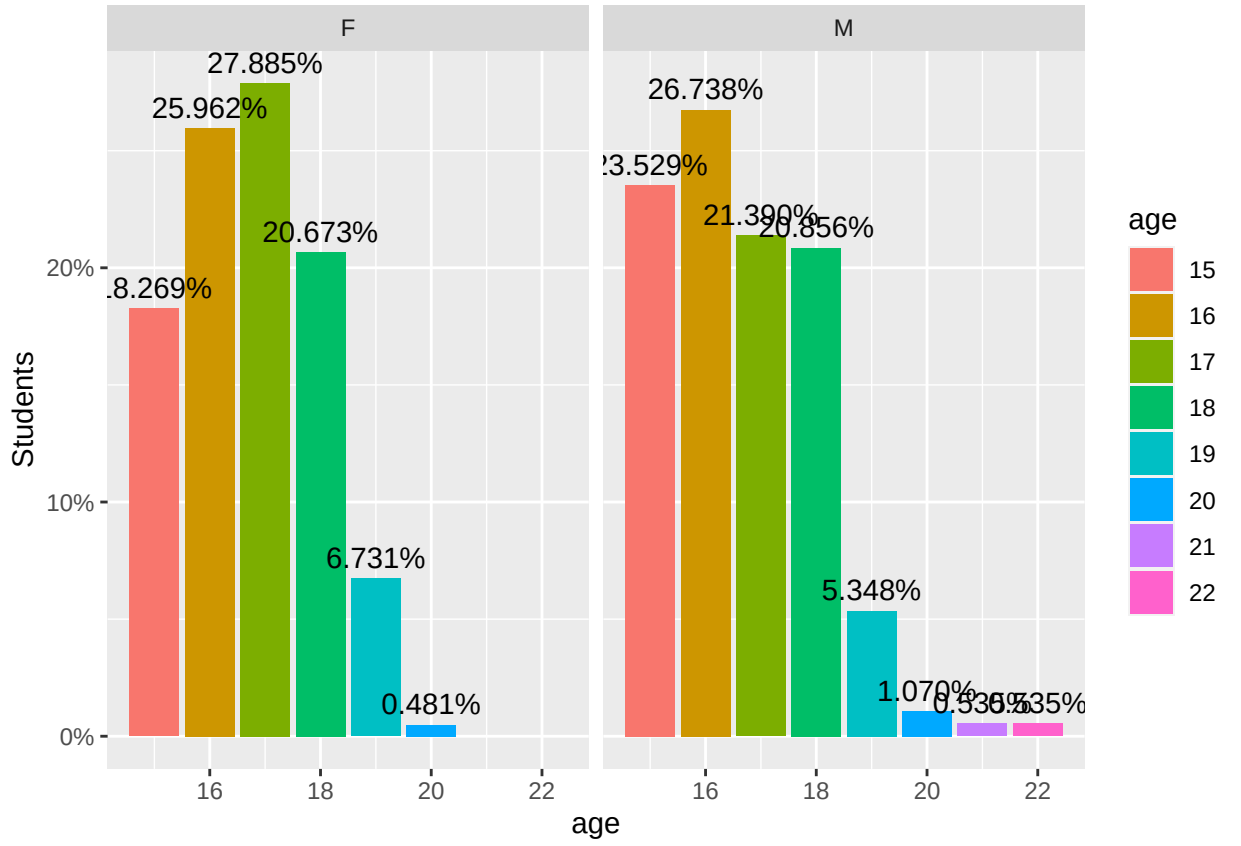



GP = Gabriel Pereira School, MS = Mousinho da Silveira School

```r
#boxplot(Var2~school, data = student_math, xlab = "GP = Gabriel Pereira School, MS = Mousinho da Silvei
```

- Students taken Math course distribution from each school are:
  - 88.4% students for Gabriel Pereira School
  - 11.6% students for Mousinho da Silveira School
- Age and sex distribution in the Math course

```r
ggplot(student_math, aes(x= age, group=sex)) +
    geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
    geom_text(aes( label = scales::percent(..prop..),
                   y= ..prop.. ), stat= "count", vjust = -.5) +
    labs(y = "Students", fill="age") +
    facet_grid(~sex) +
    scale_y_continuous(labels = scales::percent)
```




```
#ggplot2.barplot(data=student_math, xName="sex", yName='age', groupName="sex")


#ggplot(student_math, aes(x=age, y=age, fill=sex)) +   geom_bar(stat='identity', position='dodge')
```

**- First, we need to organize the data frame into two data frame that represents the two schools.**

```
gradeFunc <- function(df, grade){
  for (i in 1:(nrow(df))) {
  if (grade[i] < 7) {
    df$grade[i] <- "F"
  } else if (grade[i] < 10) {
    df$grade[i] <- "D"
  } else if (grade[i] < 14) {
    df$grade[i] <- "C"
  } else if (grade[i] < 18) {
    df$grade[i] <- "B"
  } else {
    df$grade[i] <- "A"
  }
  }
  return(df$grade)
```

```
}

#gradeFunc <- function(df){.. return df} ...now I can call
student_math$grade1 <- gradeFunc(student_math, student_math$G1)
student_math$grade2 <- gradeFunc(student_math, student_math$G2)
student_math$grade3 <- gradeFunc(student_math, student_math$G3)

student_portuguese$grade1 <- gradeFunc(student_portuguese, student_portuguese$G1)
student_portuguese$grade2 <- gradeFunc(student_portuguese, student_portuguese$G2)
student_portuguese$grade3 <- gradeFunc(student_portuguese, student_portuguese$G3)

view(student_math)

# le't organize data frame
student_math_MS <- student_math %>%
  filter ( school == "MS") # select(student_math, )
#head(student_math_MS)
student_math_GP <- student_math %>%
  filter ( school == "GP") # select(student_math, )
student_portuguese_MS <- student_portuguese %>%
  filter ( school == "MS") # select(student_math, )
student_portuguese_GP <- student_portuguese %>%
  filter ( school == "GP") # select(student_math, )
```

- Let's do summary on Math result 1 for students from Gabriel Pereira School

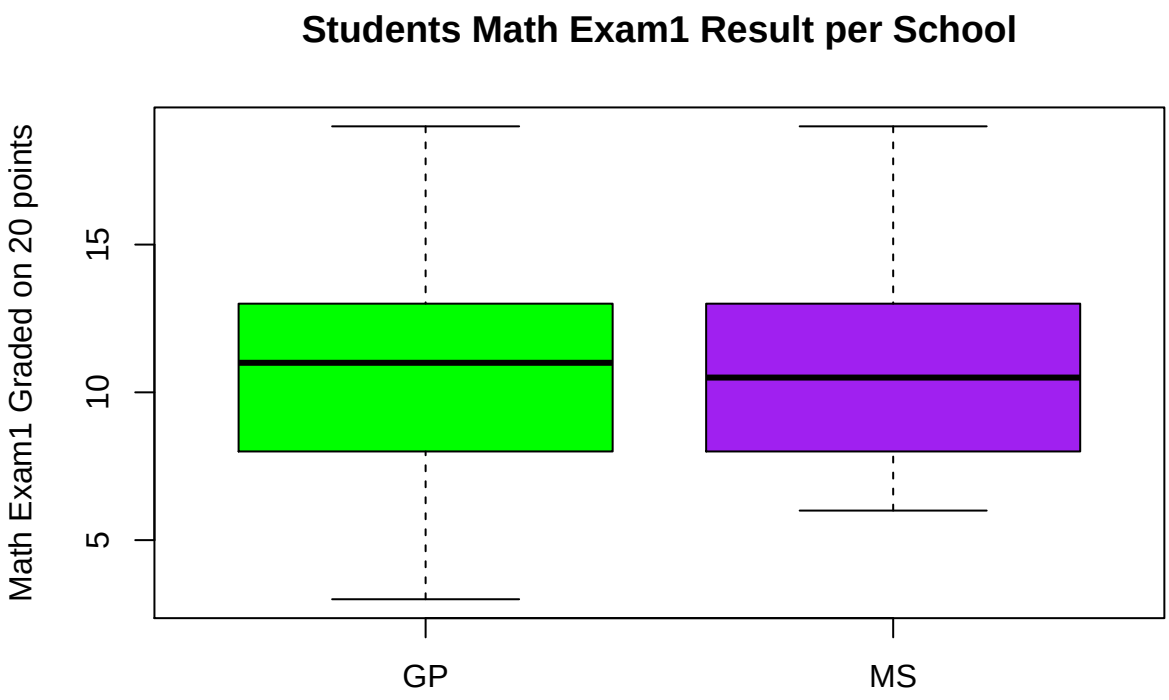
```
## student_math_GP$G1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      349        0       17    0.992    10.94    3.791        6        7
##      .25      .50      .75      .90      .95
##        8       11       13       16       16
##
## lowest :  3  4  5  6  7, highest: 15 16 17 18 19
##
## Value          3     4     5     6     7     8     9    10    11    12    13
## Frequency      1     1     7    19    32    35    30    45    34    32    27
## Proportion 0.003 0.003 0.020 0.054 0.092 0.100 0.086 0.129 0.097 0.092 0.077
##
## Value         14    15    16    17    18    19
## Frequency     27    21    21     8     7     2
## Proportion 0.077 0.060 0.060 0.023 0.020 0.006
```

- Let's see the mean, max for students from Gabriel Pereira School

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   11.00   10.94   13.00   19.00
```

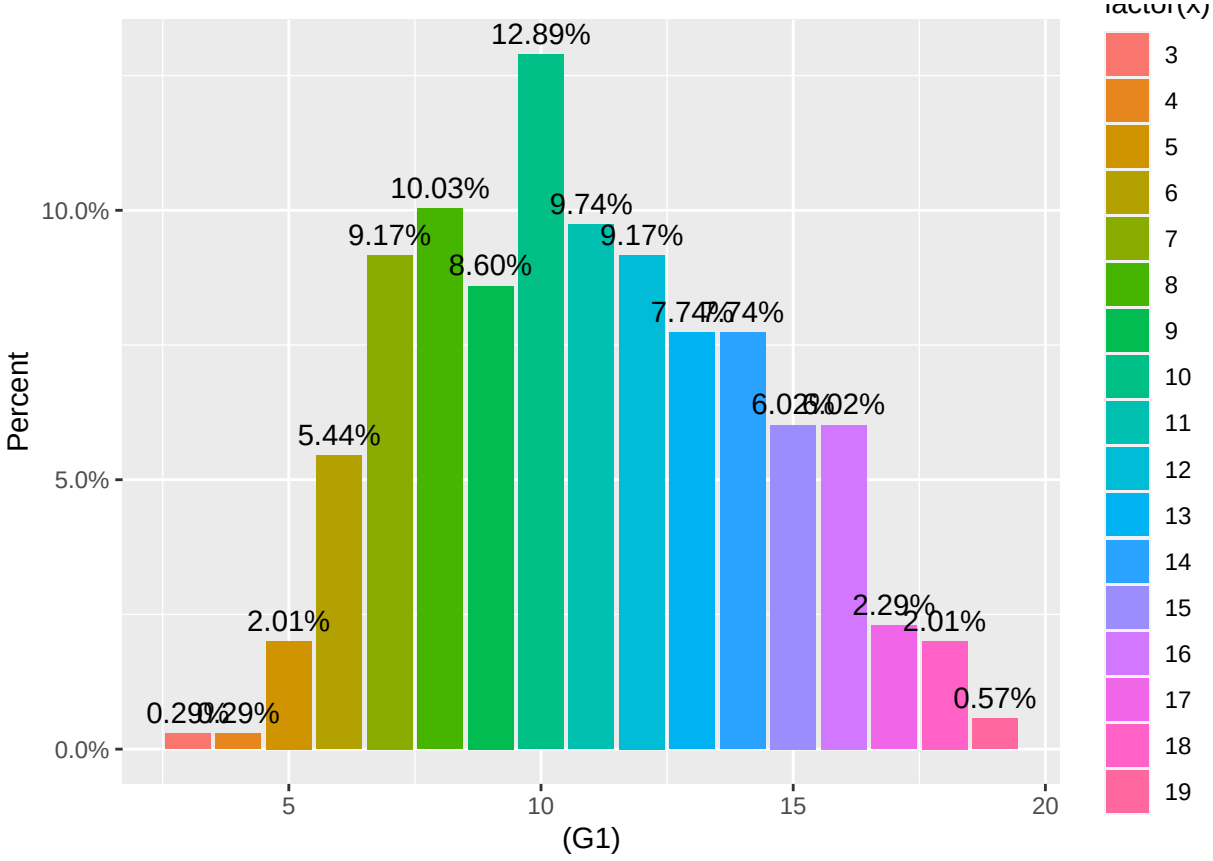
## Data Analysis

- Let's take a look at students performance on the Math Exam 1.
- We are interested in students performance in Math course

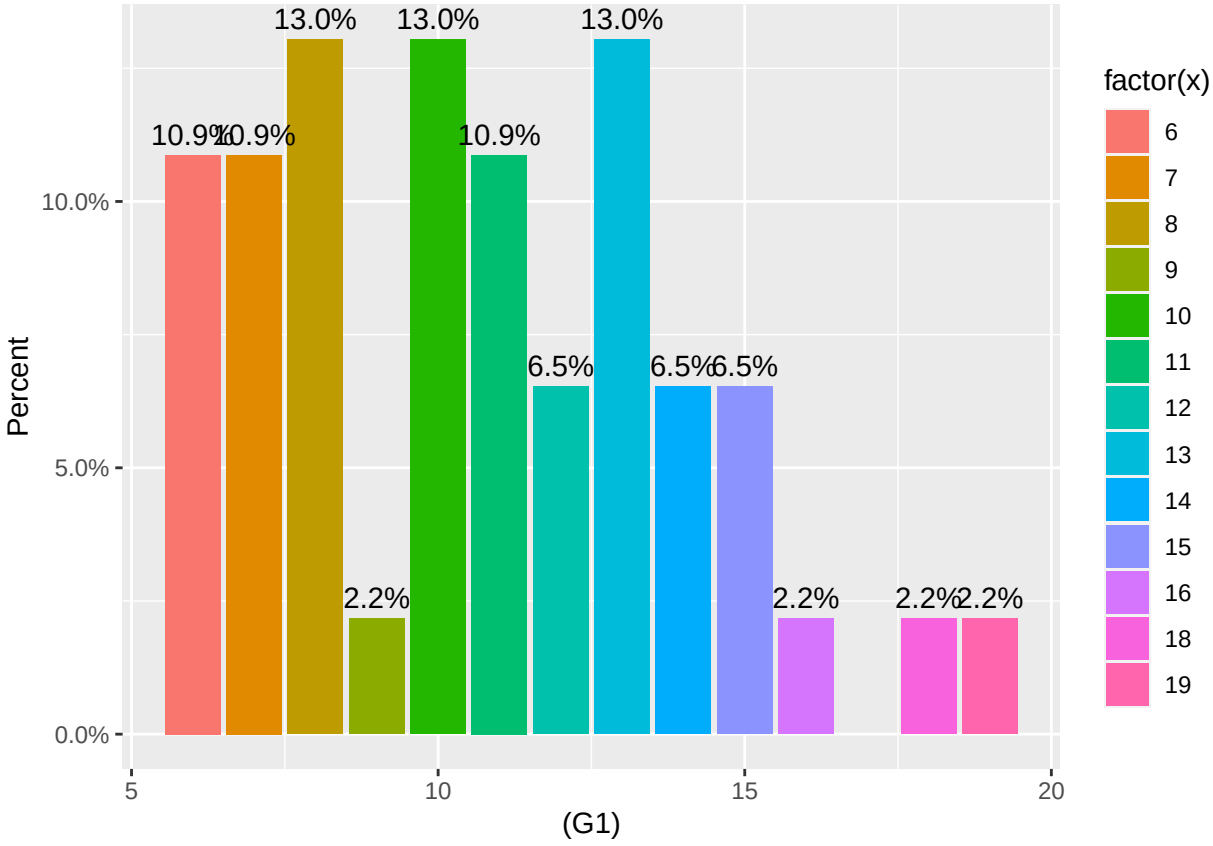



- Students Math Exam1 result distribution from Gabriel Pereira School

```
ggplot(student_math_GP, aes(x= (G1))) +
  geom_bar(aes(y = ..prop.., fill = factor(..x..)), stat="count") +
  geom_text(aes( label = scales::percent(..prop..),
                 y= ..prop.. ), stat= "count", vjust = -.5) +
  labs(y = "Percent") +
  scale_y_continuous(labels=percent)
```

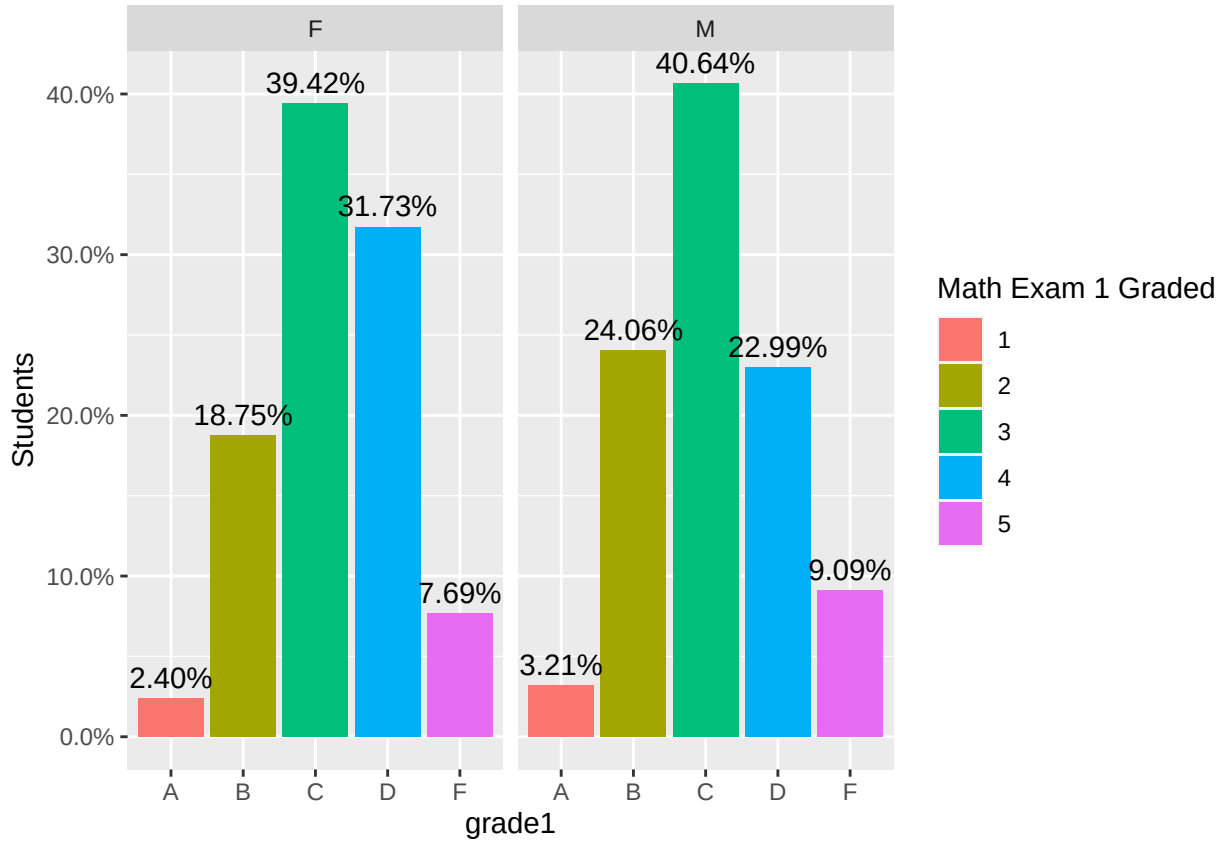



- Students Math Exam1 result distribution from Mousinho da Silveira School

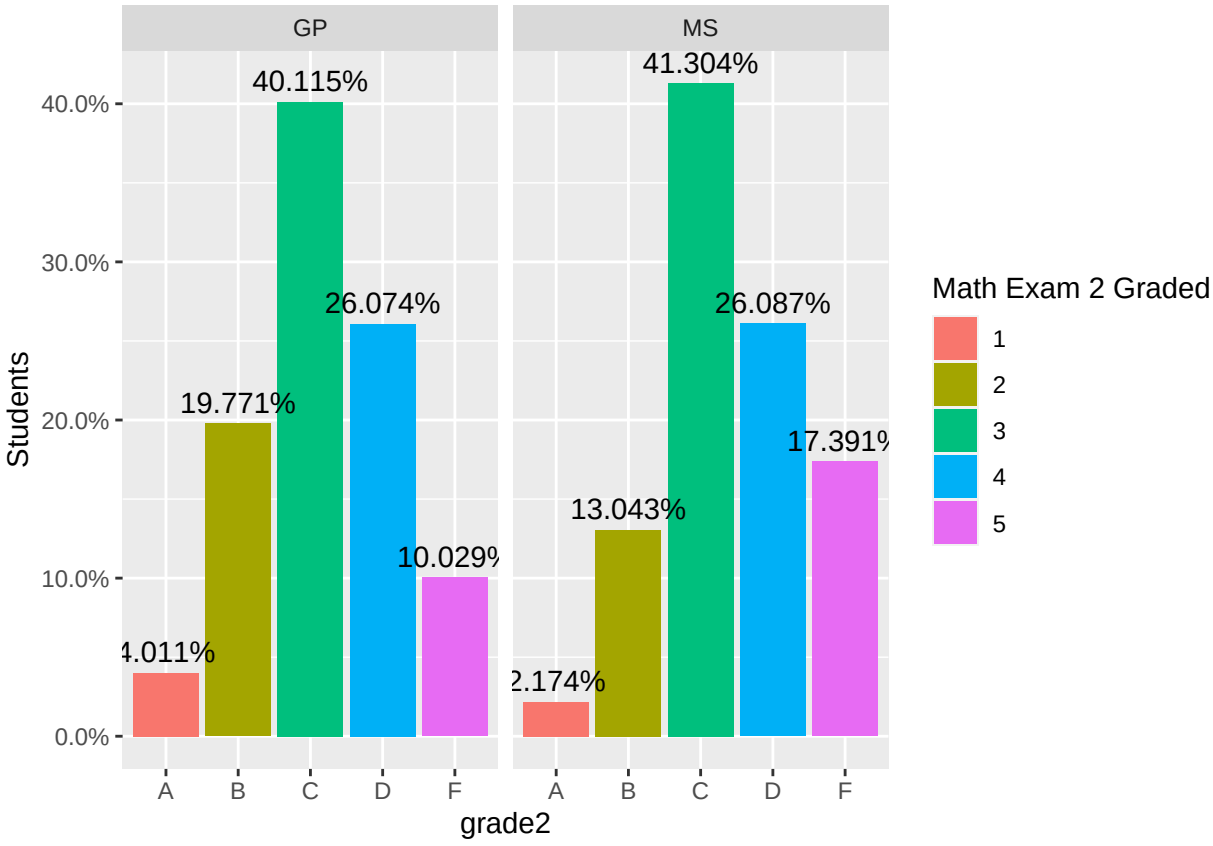



- Though. A better representation is graded letters

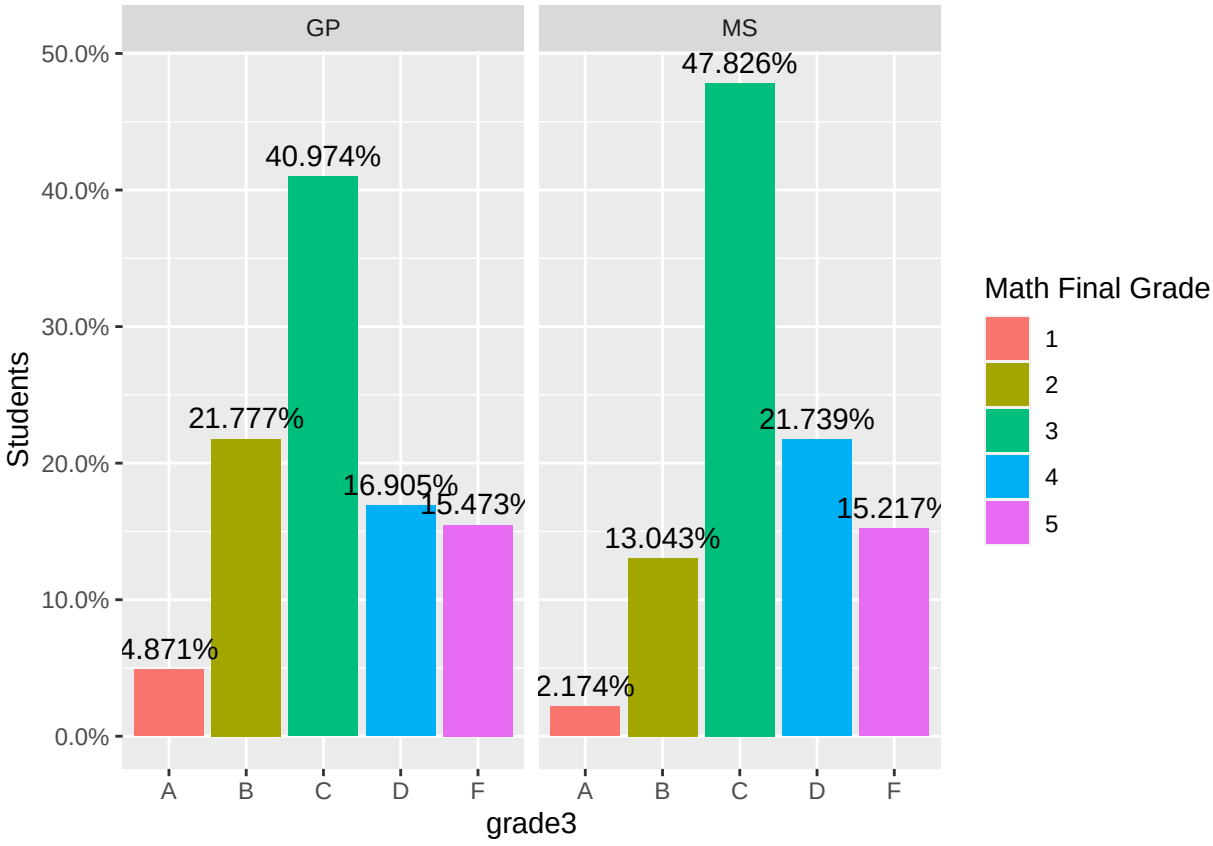
**- Let's see the math exam1 graded from the two schools**




- Let's see the math exam2 graded from the two schools



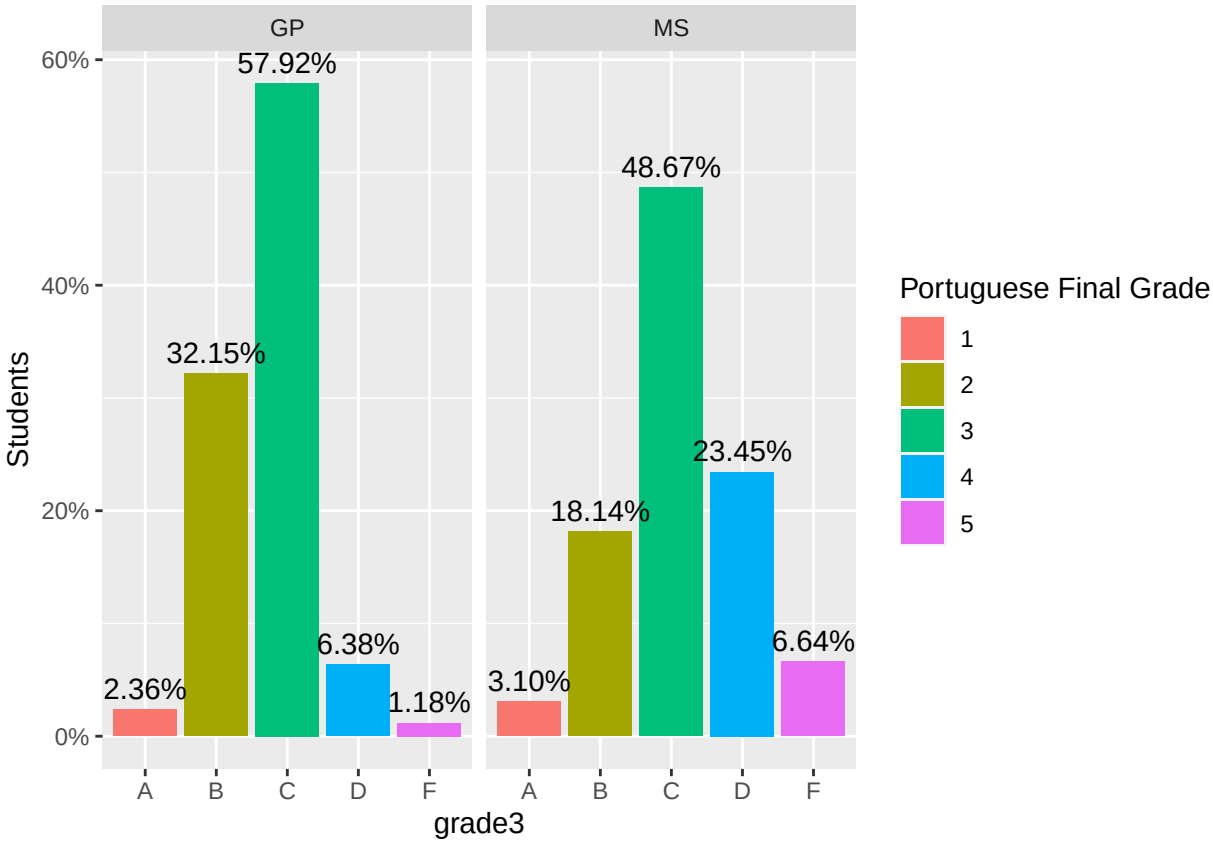

- Let's see the math final grade from the two schools




- The results from the two Math Exams don't look good.

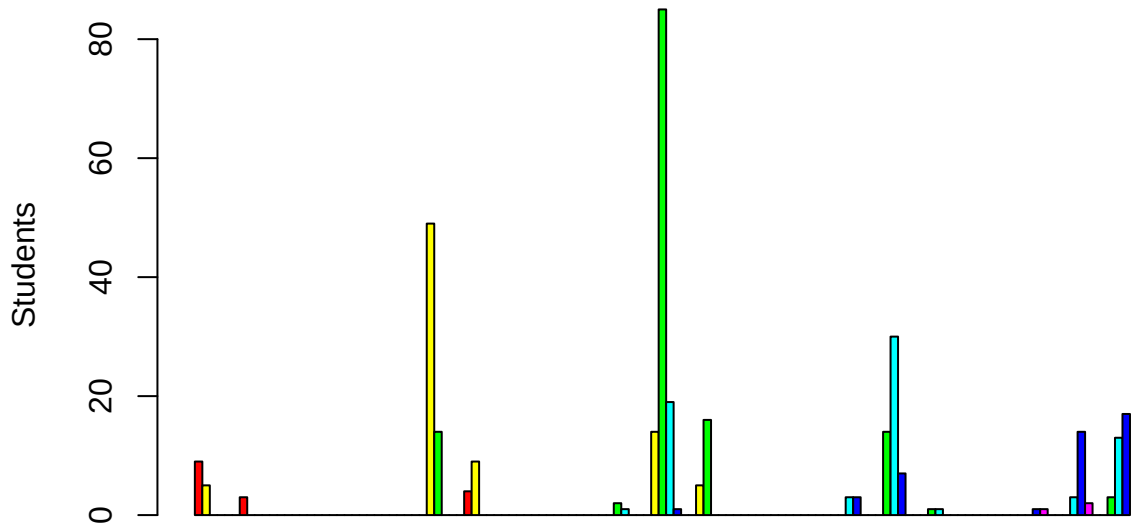
### Let's see the Portuguese Final grade from the two schools.

- Let's see the Portuguese Final grade from the two schools




- Let's see Multiple comparison or group barplots to show grade 1, 2 and 3 or G1, G2, G3
- To see overall performance trend from grade 1 to final grade

```
#Group all the grade (grade 1, grade 2, grade 3) in one plot
all_Math_Grade <- select(student_math_GP, grade1, grade2, grade3)
#all_Math_Grade2 <- table(all_Math_Grade$grade3)


#barplot(all_Math_Grade2, beside = TRUE, main = "Students Math Exam1 Graded Distribution from Gabriel P
#head(student_math_GP)



barplot(as.matrix(table(select(student_math_GP, grade1, grade2, grade3))), beside = TRUE, main = "Stude
```




```
#head(student_math_GP)



#barplot((select(student_math_GP, G1, G2, G3)), beside = TRUE, main = "Students Portuguese Exam1 Graded
#head(student_math_GP)
```

### - Stats summary from Gabriel Pereira School

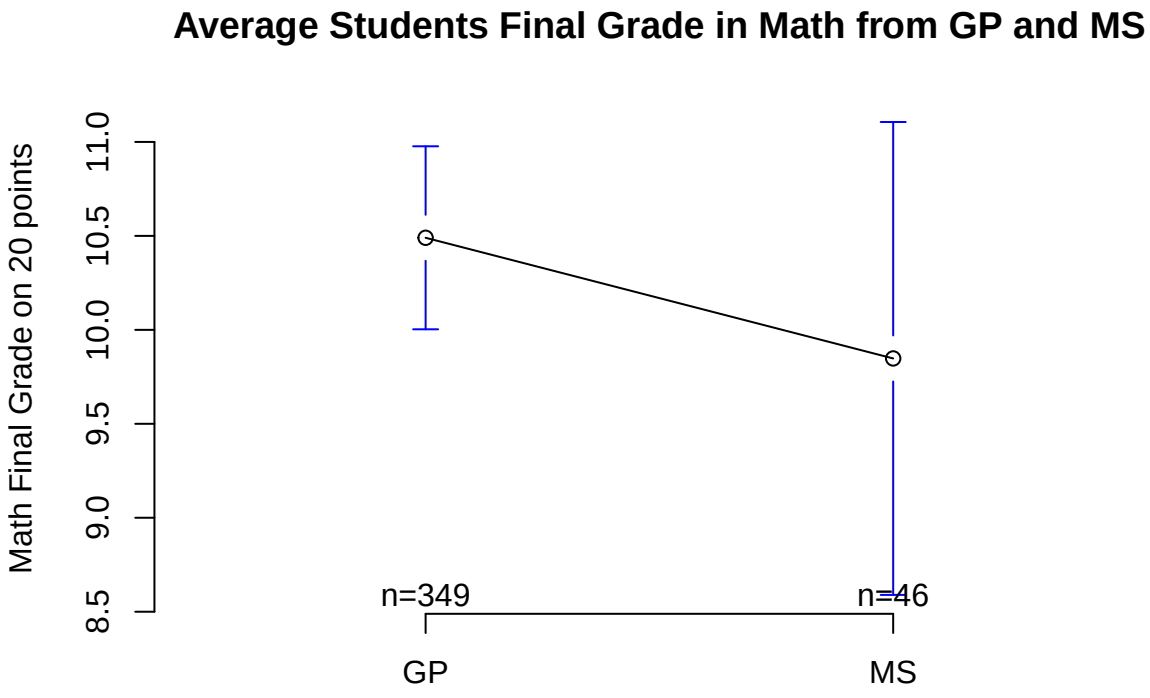
```
## student_math_GP$grade3
##        n  missing distinct
##      349        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency     17    76   143    59    54
## Proportion 0.049 0.218 0.410 0.169 0.155

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.49   14.00   20.00
```

- Stats summary from Mousinho da Silveira School

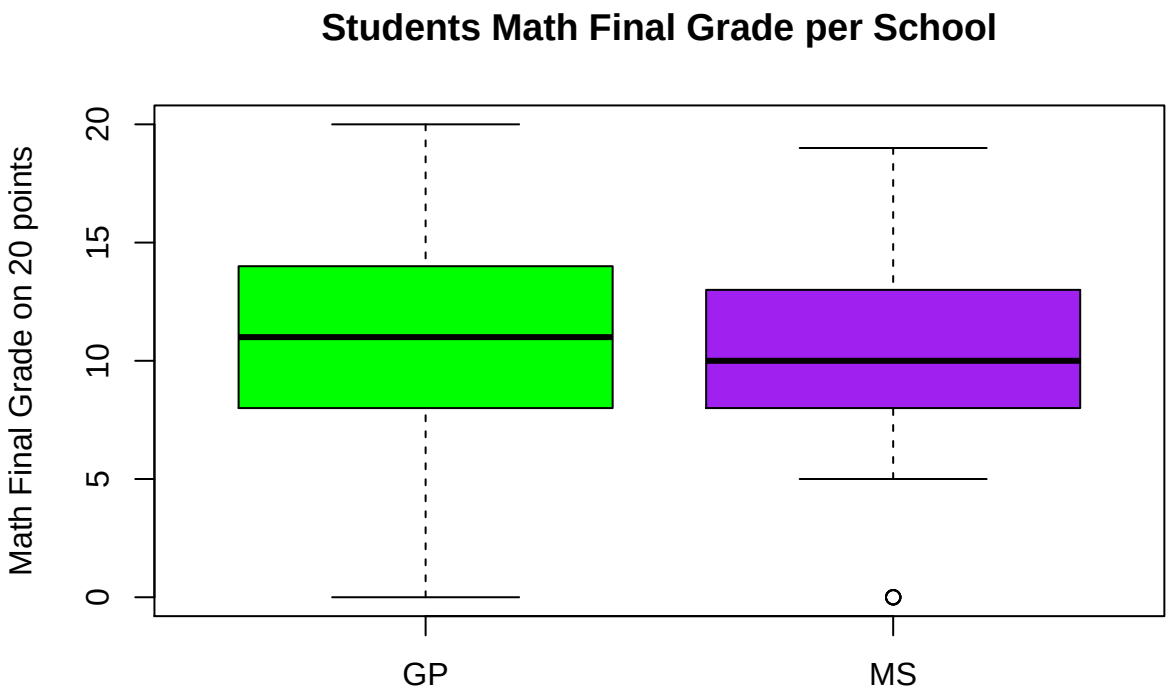
```
## student_math_MS$grade3
##        n  missing distinct
##       46        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency      1     6    22    10     7
## Proportion 0.022 0.130 0.478 0.217 0.152

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.848  12.750  19.000
```

**- Let's visualize the average score in Math course from the two schools.**

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter

## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
## a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```




GP = Gabriel Pereira School, MS = Mousinho da Silveira School

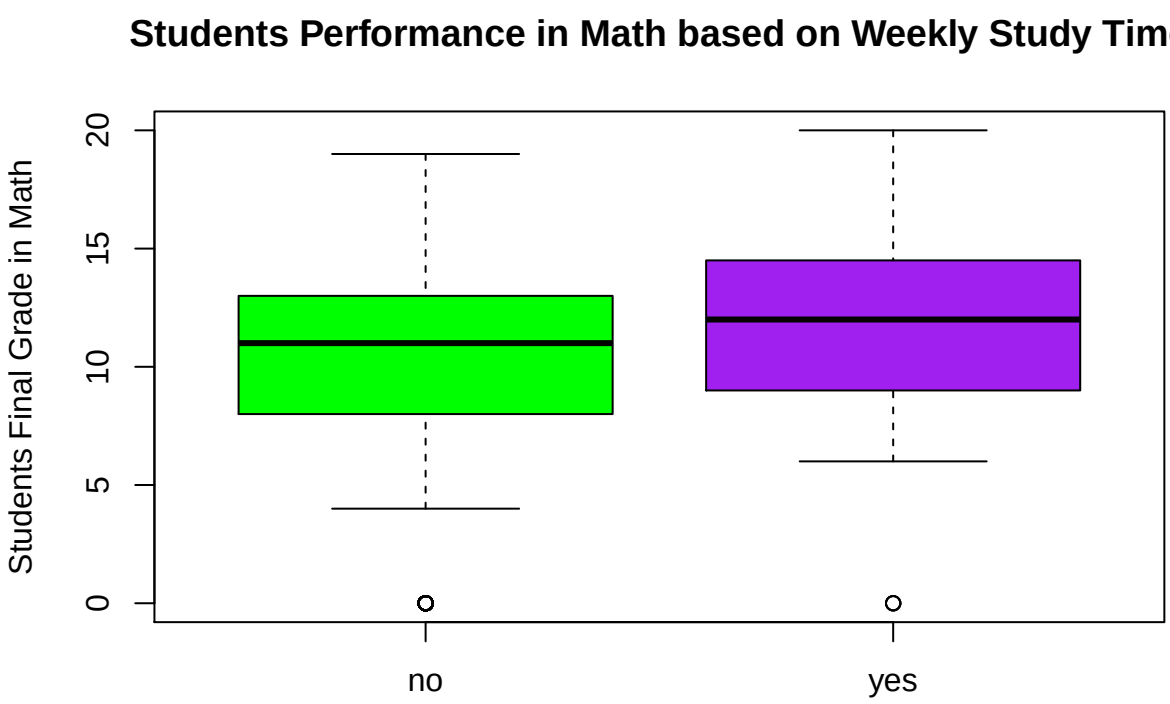



GP = Gabriel Pereira School, MS = Mousinho da Silveira School

### - Correlation between amount of study time and result

Conducting a hypothesis test to evaluate whether the average grade is different for those who study at least ten times a week than those who don't.

- H_null: there is no difference in the average grade for those who study at at least ten times a week than those who don't.
- H_alt: there is difference in the average grade for those who study at at least ten times a week than those who don't.
- case = students enrolled in Math course.
- sample is all students from both school (GP and MS).
- Let's see the difference between weekly study time and students final grade in Math




Students Weekly Study Time: Yes = student spent 10+hrs, No = student spent less than

- Let's see the final grade ratio between students who study 10+ hrs a week and those who don't in math course

```
#cat ("Let's see the final grade ratio between students who study 10+ hrs a week and those who don't in
student_math1 %>%
  group_by(studyTime10) %>%
  summarise(meanFinal_grade = mean(G3))
```

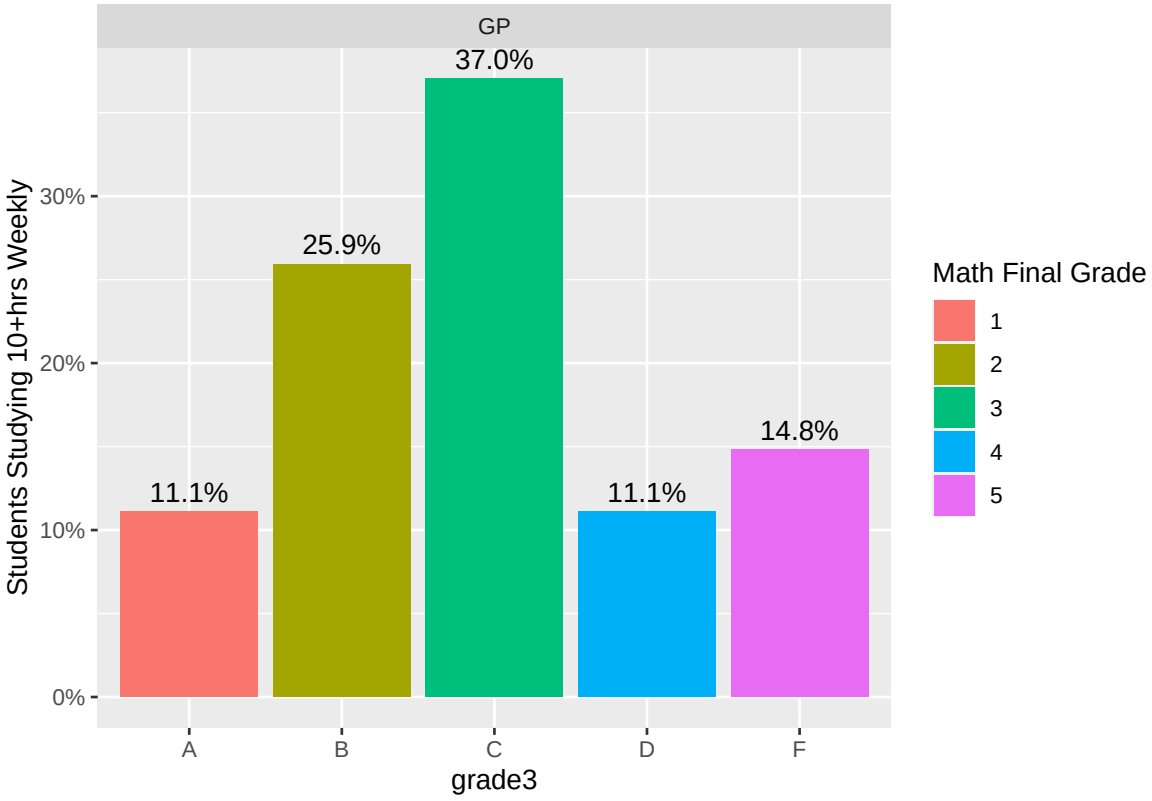
```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 2
##   studyTime10 meanFinal_grade
##   <chr>                 <dbl>
## 1 no                     10.4
## 2 yes                    11.3
```

```
study10plus <- student_math1 %>%
  filter(studyTime10 == "yes" )  # & G1 & G2 & G3

study10Less <- student_math1 %>%
  filter(studyTime10 == "no" )
```

**- Let's see the statical information about students final grade in Math based on 10+hrs weekly study time**

```
## study10plus$grade3
##        n  missing distinct
##       27        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency      3     7    10     3     4
## Proportion 0.111 0.259 0.370 0.111 0.148
```

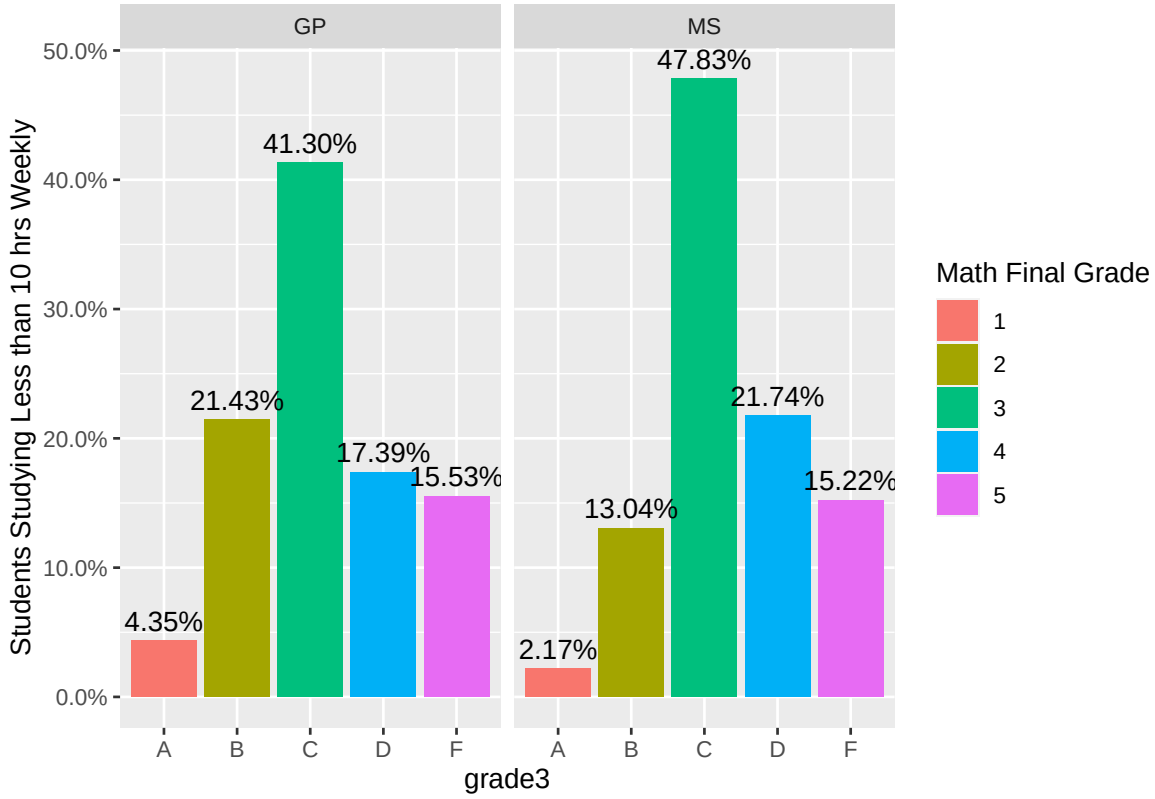
- Let's visualize the Math final grade distribution from the two schools based on 10+hrs weekly study time



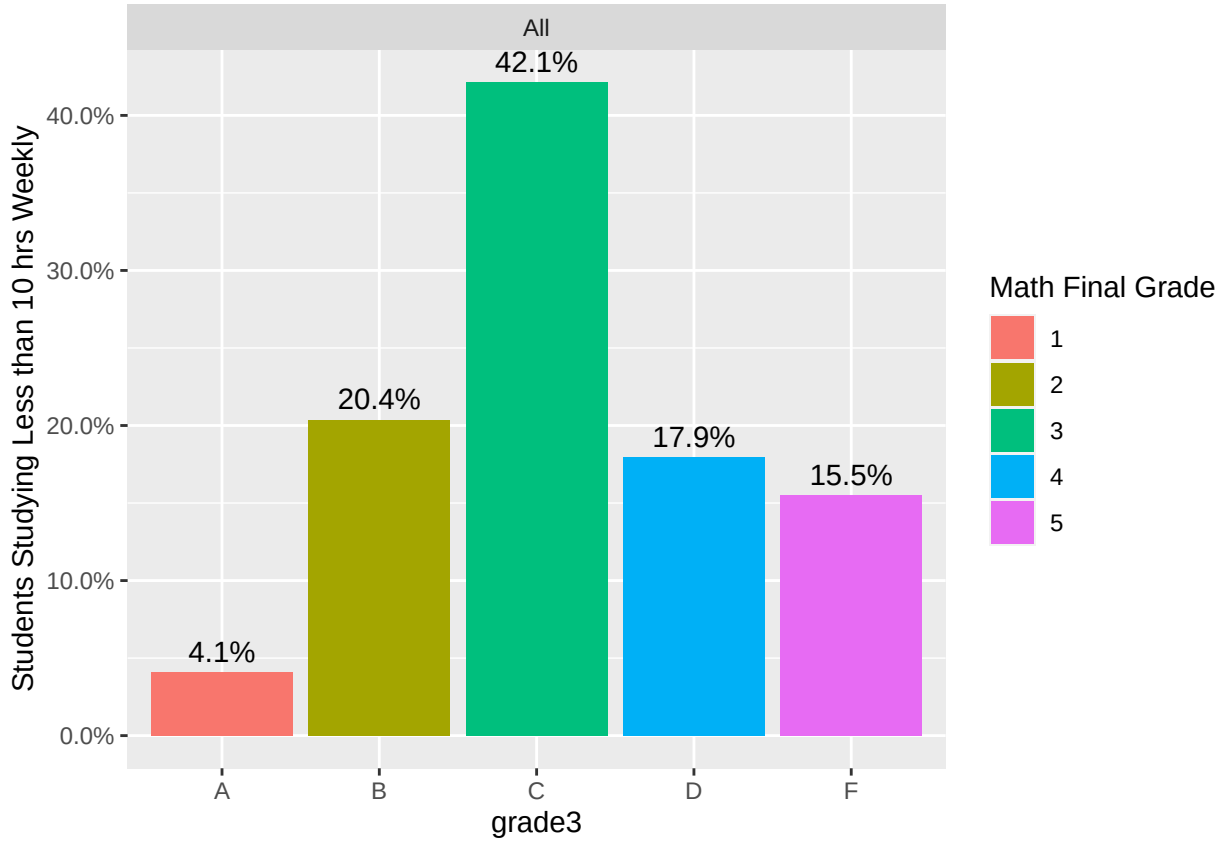

**- Statistical information about students final grade in Math based on less than 10hrs Weekly study time**

```
## study10Less$grade3
##        n  missing distinct
##      368        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency     15    75   155    66    57
## Proportion 0.041 0.204 0.421 0.179 0.155
```

- Let's visualize the math final grade distribution from the two schools based on 10+hrs weekly study time



**- Overall students performance in Math course from the two school**




- Computing the hypothesis test.

```
# This can be rewrite as a function

n_yes <- nrow(study10plus)
n_no <- nrow(study10Less)
df <- n_yes - 1
mean_no <- mean(study10Less$G3)
mean_yes <- mean(study10plus$G3)
sd_no <- sd(study10Less$G3)
sd_yes <- sd(study10plus$G3)
SE <- sqrt( (sd_yes^2)/n_yes + (sd_no^2)/n_no)
t_value <- qt(0.05/2, df, lower.tail = FALSE)
point_estimate <-  mean_yes - mean_no
lower_CI <- point_estimate - t_value * SE
upper_CI <- point_estimate + t_value * SE
lower_CI
```

```
## [1] -1.238795
```

```
upper_CI
```

```
## [1] 3.050792
```

```r
p_value <- 2*pt(t_value, df, lower.tail = FALSE)
p_value
```

```
## [1] 0.05
```

The p-value = 0.05 < alpha (0.1), thus we reject the null hypothesis. Thus, there is difference in the average grade for those who study at at least ten times a week than those who don't.

## Interpret Results

- In this study, there are 395 students both from Gabriel Pereira (GP) School and Mousinho da Silveira (MS) School.
- These students are enrolled in Math course of which 349 are from GP and 46 from MS.
- Based on the final grade in Math course, students from GP have a higher average grade than those from MS.
- Statistically, the average for students from GP in Math course is 10.49.
- Statistically, the average for students from MS in Math course is 9.85.
- The majority of students from both school received a "C" grade.
- Statiscally, 32.38% students from GP failed the Math course.
- Statistically, 36.96% students from MS failed the Math course.
- The analysis shows that students from both schools don't perform well in Math.
- The conducted test in this study has proved with 95% confidence interval that students who do studying at least 10hrs in a week do well in Math course than those who spent lesser time.
- Shockingly, there is no student from MS who studies at least 10hrs in a week.
- Overall, students from GP did better in Math course than those from MS.

## Benefits

- The interest in experimental study related to school will have the advantage to help schools' officials in decision making in term of improving school education system.
- This project is seeking to make the collected data about ("GP" - Gabriel Pereira or "MS" - Mousinho da Silveira) schools speaks or reveals useful information.
- This experiemental study aims to help school's officials in planning strategy for better school education system. Ultimately, I plan to become a consultant using my skills as data scientist in various domain of the society to present meaningful report to government entities, companies, and organizations to help them in decision making.
- So, this project will contribute to building skills necessary for one to be successful in data science.

## Challenges

- Adding percentage to a barplot (variable = non-numerical).
- How to perform multiple comparison or group barplots to show grade 1, 2 and 3 or G1, G2, G3.
- How to add mean on boxplot for all grades (G1, G2 and G3), or how to plot mean of two variables side by side for all grades (G1, G2 and G3).
- Issue with knit: in order to knit this project from Rmardown, we have to comment out the Rsql chunck code which works fine.
- Some times a function works, describe(), describeBy and later does not work.
- Struggled how to do a better project presentation in Rmarkdown.
- Dealing with slow computer during this project was little painful.

## References


1. https://fall2020.data606.net/assignments/labs/
2. file:///C:/Users/Petit%20Mandela/Documents/R/DATA606_Lab7/DATA606_Lab7/DATA606_Lab7.html
3. https://www.statisticshowto.com/least-squares-regression-line/
4. https://rcompanion.org/handbook/C_04.html
5. https://data-flair.training/blogs/t-tests-in-r/
6. https://rstatisticsblog.com/data-science-in-action/data-preprocessing/hypothesis-testing-in-r-with-examples-interpretations/
7. https://www.r-graph-gallery.com/all-graphs.html
8. http://www.sthda.com/english/wiki/ggplot2-barplot-easy-bar-graphs-in-r-software-using-ggplot2