The City University of New York School of Professional Studies

Statistics and Probability for Data Analytics (DATA 606)

Final Project: Data Insights to Improve school Education System

Alexis Mekueko

12/08/2020

Part 2 - 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/DATA606_Final_Project

Web link: https://rpubs.com/amekueko/697306 https://rpubs.com/amekueko/701875

Part 2a - 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.

Part 2b - 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?

Part 3 - Data

Part 3a - Data Acquisition

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.

Part 3b - 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 courses. 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 columns. 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.

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

Using R to acquire data

Part 4 - Data Preparation / Data Wrangling

Part 4a - Cleaning data

What is the structure of data?

```
## 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, ...
```

```
##   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
```

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

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

Part 5 - Explore Data

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

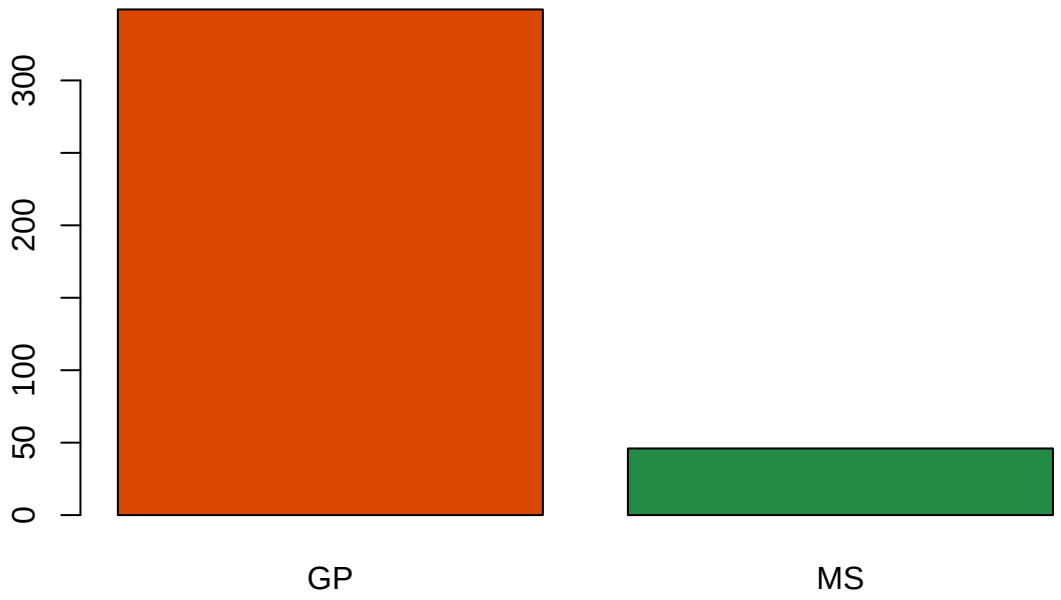
- Amount the 33 variables in the data frame, there are 03 variables (G1, G2 and G3) which represent the students's grades.
- These 03 variables are interesting as there are measures of students performances in the registered courses.

  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)

Let's keep in mind the research questions. Do students at "GP" - Gabriel Pereira school or "MS" - Mousinho da Silveira school perform well? If yes, what are the factors contributing to students's success? If no, what are the factors leading to students' poor performance? One way to go about these questions is to look at the 03 variables. These 03 variable can summary to one key element-That element is student's performance.

Let's take a closer look at these 03 variables. We might throw in a biais by neglecting the fact that there are two schools in the data frame. How significant is each school into the data frame.

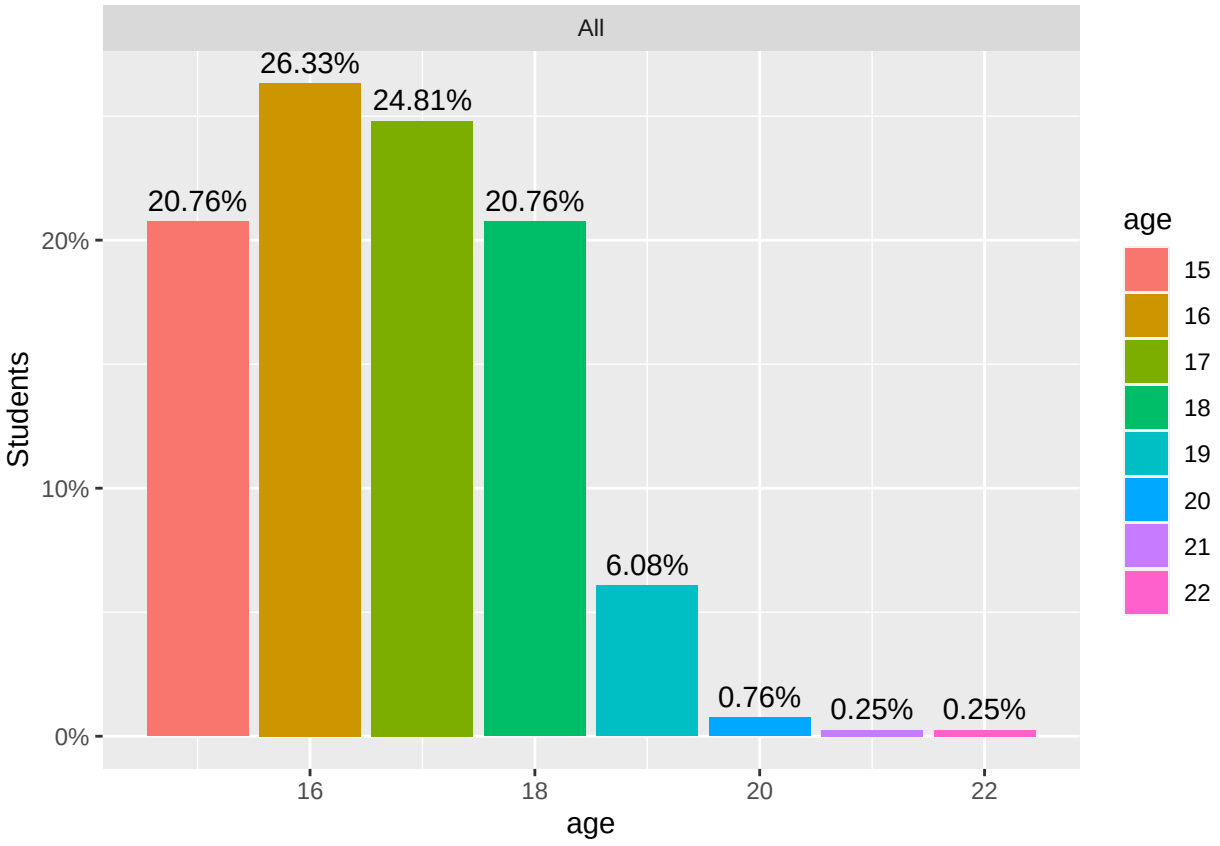
```
## student_math$G3
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      395        0       18    0.992    10.42    4.992      0.0      5.0
##      .25      .50      .75      .90      .95
##      8.0     11.0     14.0     15.6     17.0
##
## lowest :  0  4  5  6  7, highest: 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
```



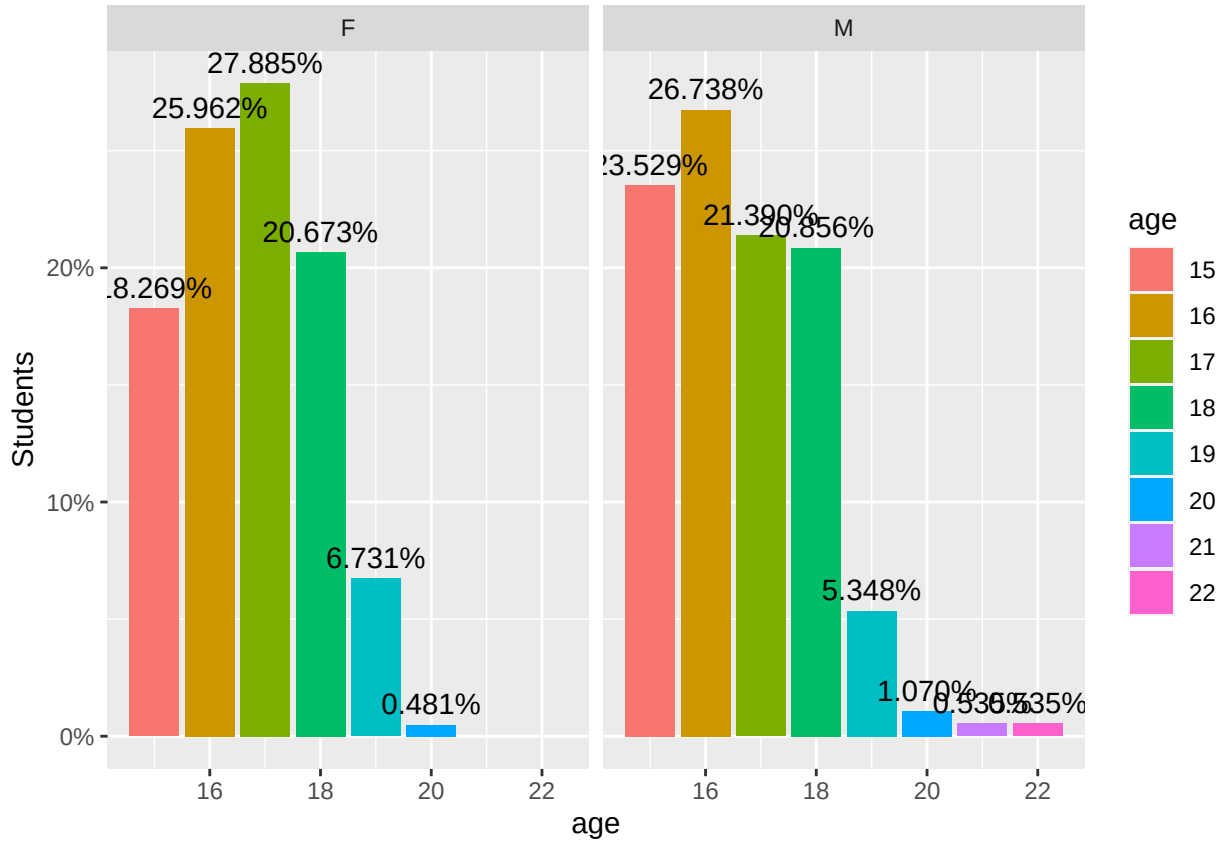
Students taken Math course distribution from each school are: * 88.4% students for Gabriel Pereira School * 11.6% students for Mousinho da Silveira School

- Actually, we could check the age distribution in the Math course

```
## student_math$sex
##        n  missing distinct
##      395        0        2
##
## Value          F     M
## Frequency    208   187
## Proportion 0.527 0.473
```



- Age and sex distribution in the Math course



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

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob reason
## 1     MS   M  18       R     GT3       T    3    2    other    other course
## 2     MS   M  19       R     GT3       T    1    1    other services   home
## 3     MS   M  17       U     GT3       T    3    3   health    other course
## 4     MS   M  18       U     LE3       T    1    3  at_home services course
## 5     MS   M  19       R     GT3       T    1    1    other    other   home
## 6     MS   M  17       R     GT3       T    4    3 services    other   home
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         1        1        no    yes   no         no
## 2    other          3         2        3        no     no   no         no
## 3   mother          2         2        0        no    yes  yes         no
## 4   mother          1         1        1        no     no   no         no
## 5    other          3         1        1        no    yes   no         no
## 6   mother          2         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1      no    yes      yes       no      2        5     5    5    5      5
## 2     yes    yes      yes       no      5        4     4    3    3      2
## 3     yes    yes      yes       no      4        5     4    2    3      3
## 4     yes     no      yes      yes      4        3     3    2    3      3
## 5     yes    yes      yes       no      4        4     4    3    3      5
## 6      no    yes      yes      yes      4        5     5    1    3      2
##   absences G1 G2 G3 Var Var2 grade1 grade2 grade3
## 1       10 11 13 13 All  350      C      C      C
```

```
## 2        8  8  7  8 All  351      D      D      D
## 3        2 13 13 13 All  352      C      C      C
## 4        7  8  7  8 All  353      D      D      D
## 5        4  8  8  8 All  354      D      D      D
## 6        4 13 11 11 All  355      C      C      C
```

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

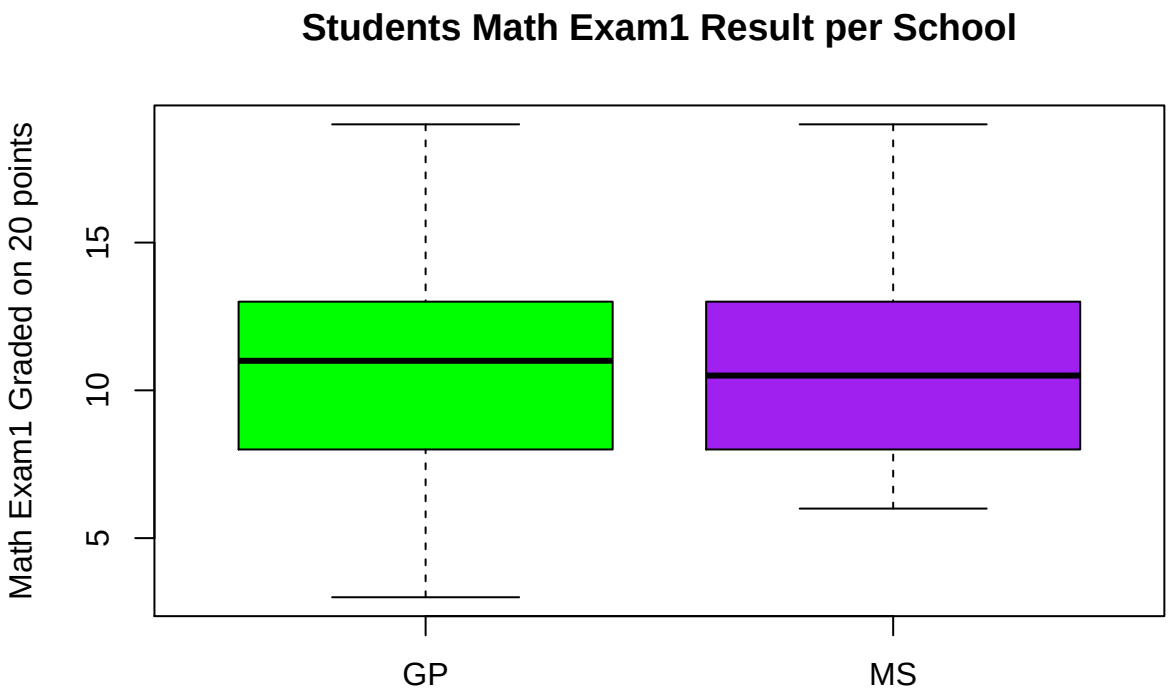
```
## student_math_GP$G1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      349        0       17    0.992    10.94    3.791        6        7
##      .25      .50      .75      .90      .95
##        8       11       13       16       16
##
## lowest :  3  4  5  6  7, 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
```

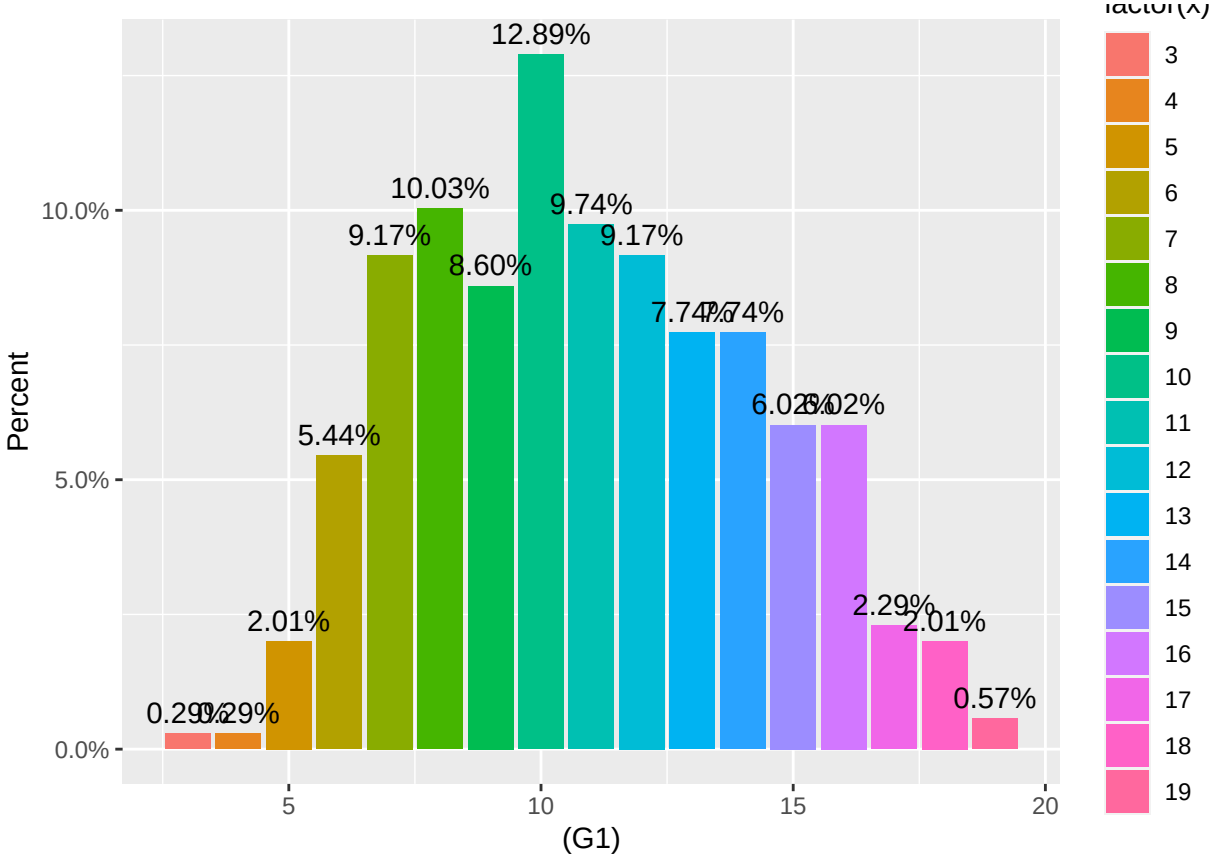
Part 6 - Data Analysis

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

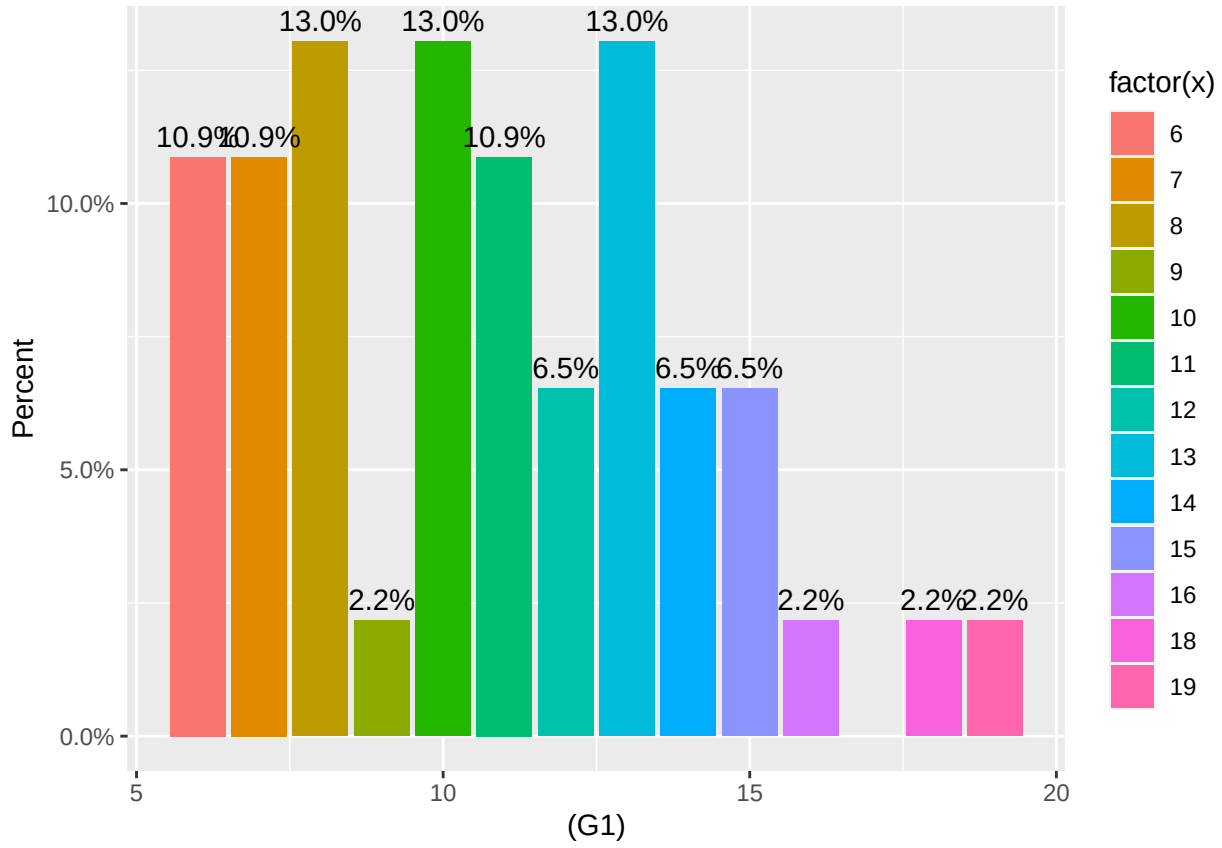


GP = Gabriel Pereira School, MS = Mousinho da Silveira School

```
##
##  Students performance in Math Exam 1 from Gabriel Pereira School
```

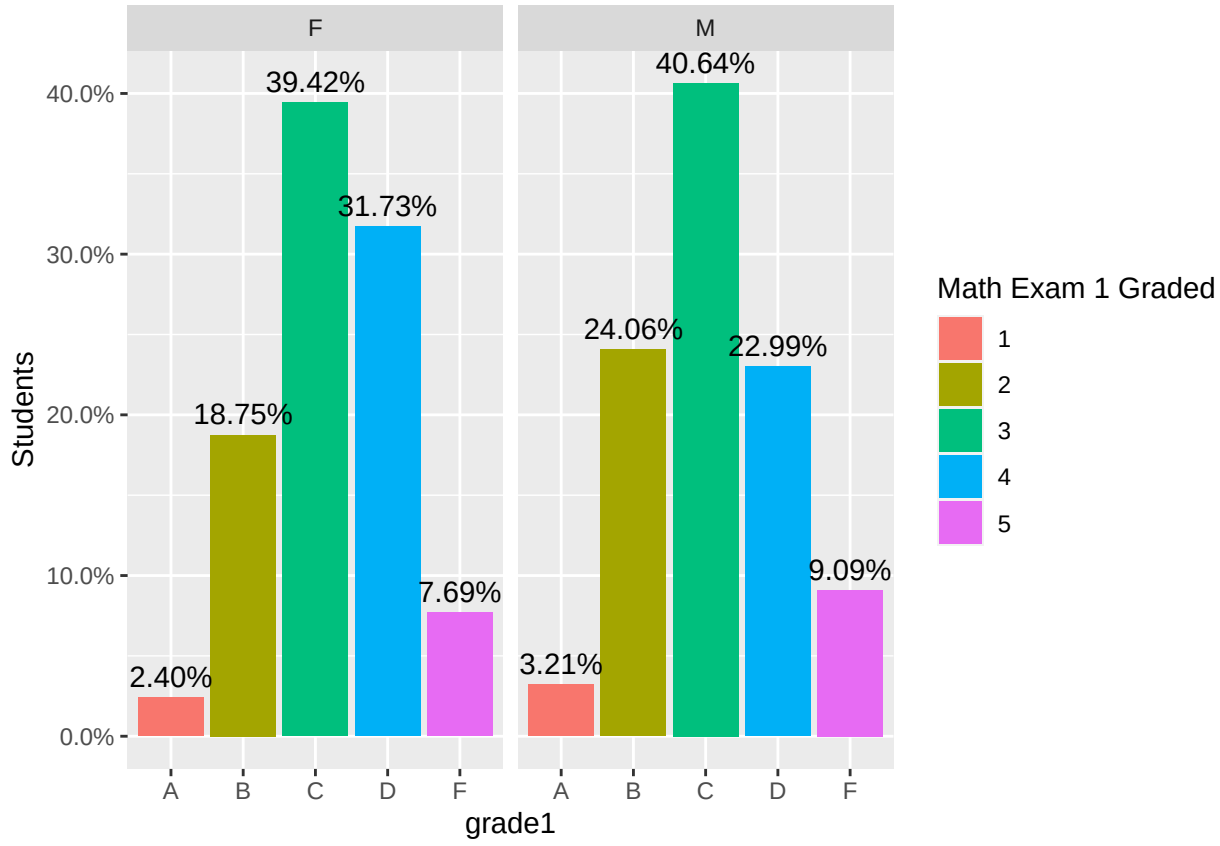


- Let's see students performance in Math Exam 1 from Mousinho da Silveira school.

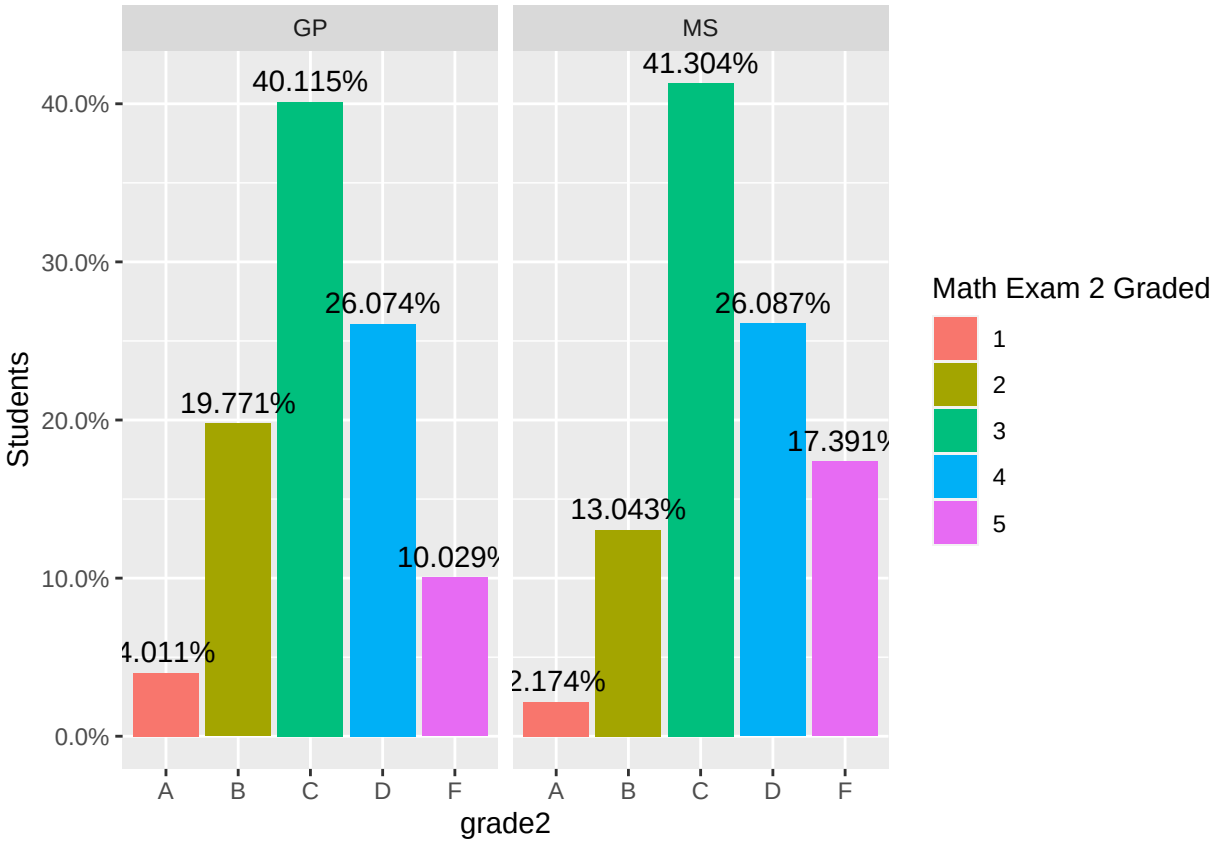



```
## A better representation is graded letters
```

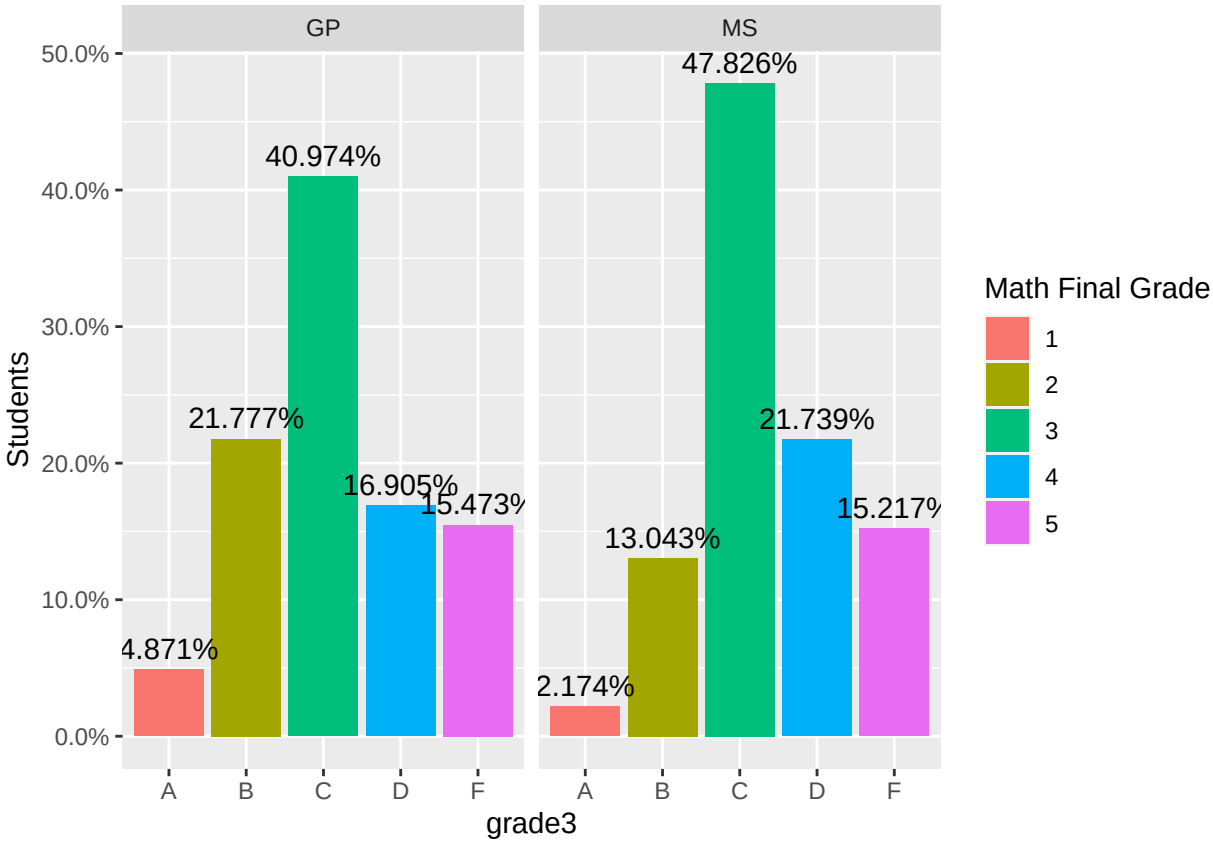
- Let's see the math exam1 graded from the two schools.



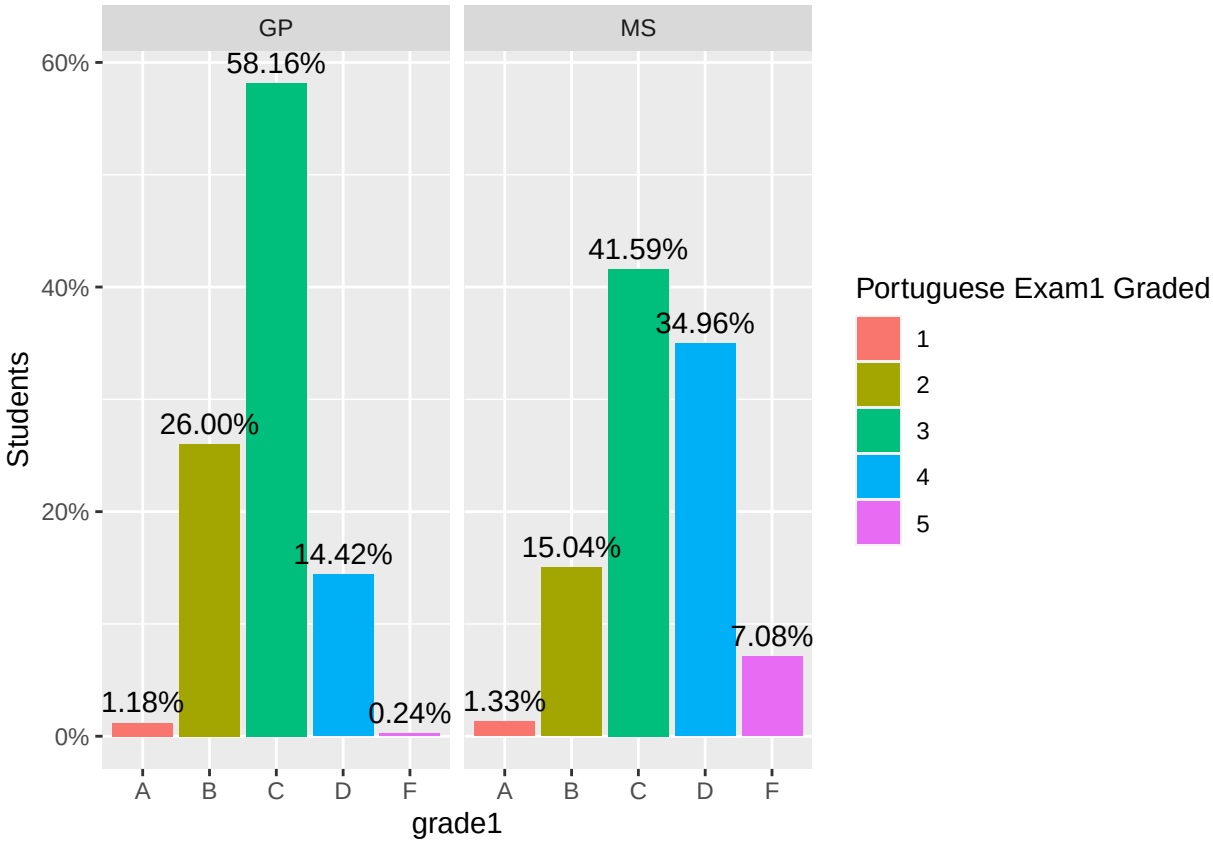
-Let's see the math exam2 graded from the two schools.



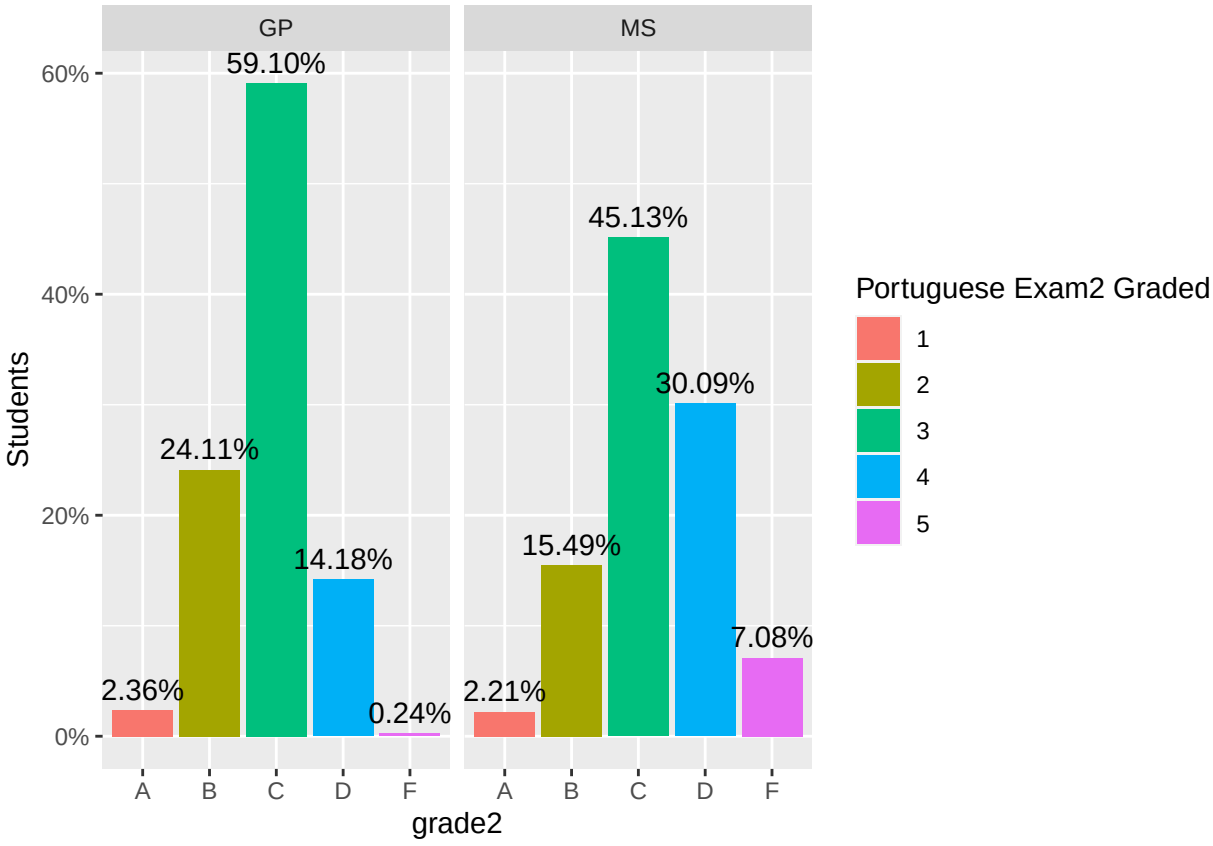
- Let's see the math final grade from the two schools.



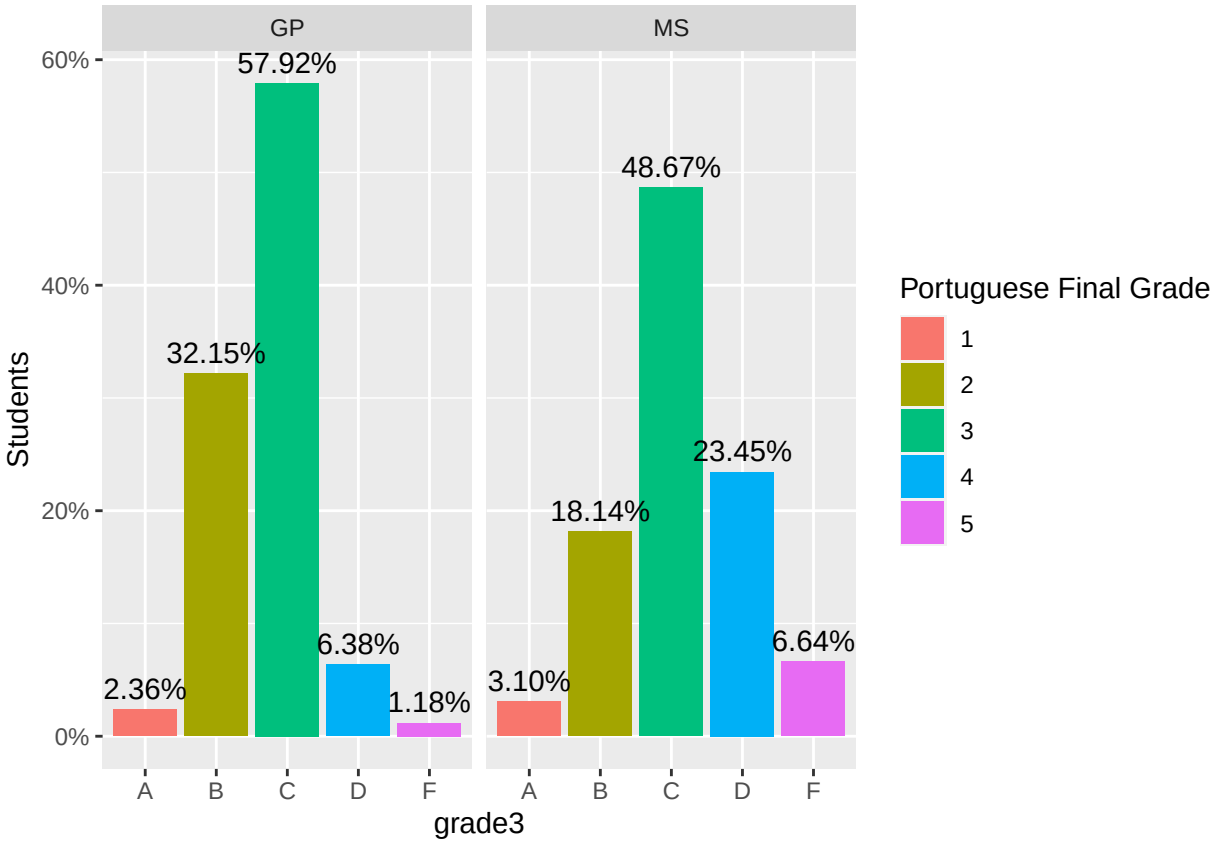
- Let's see the Portuguese Exam1 graded from the two schools



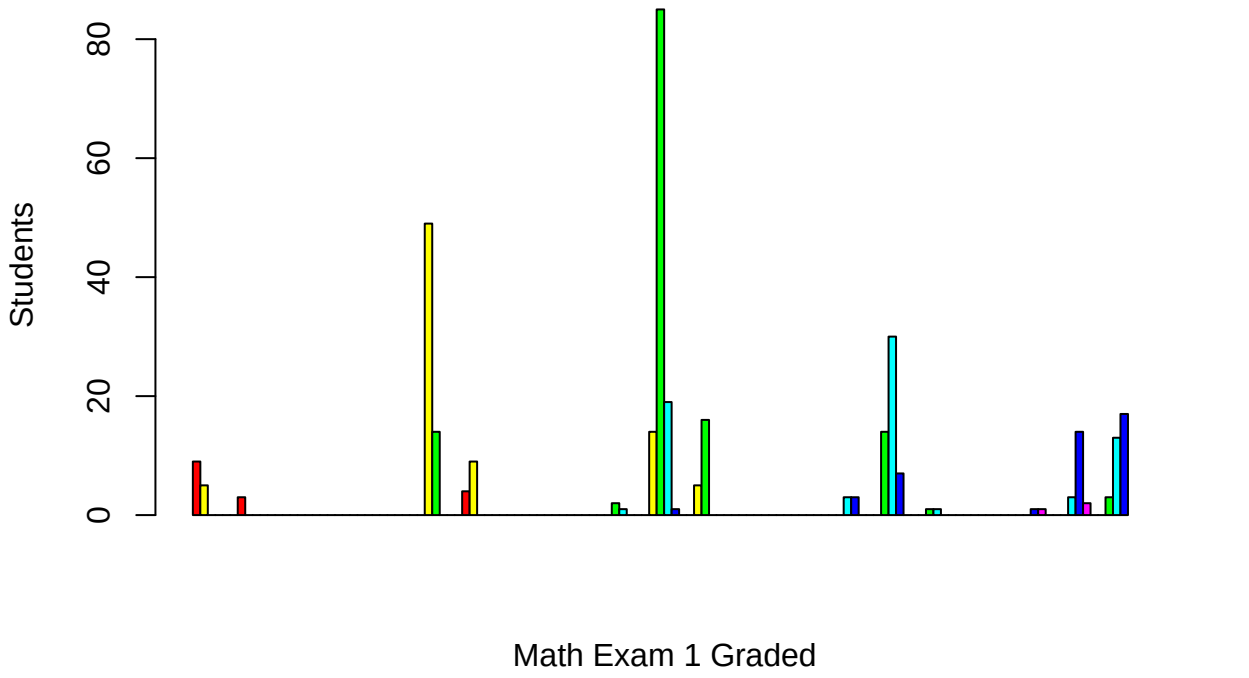
- Let's see the Portuguese Exam2 graded from the two schools.



- 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



- Let's take a look at the students performance

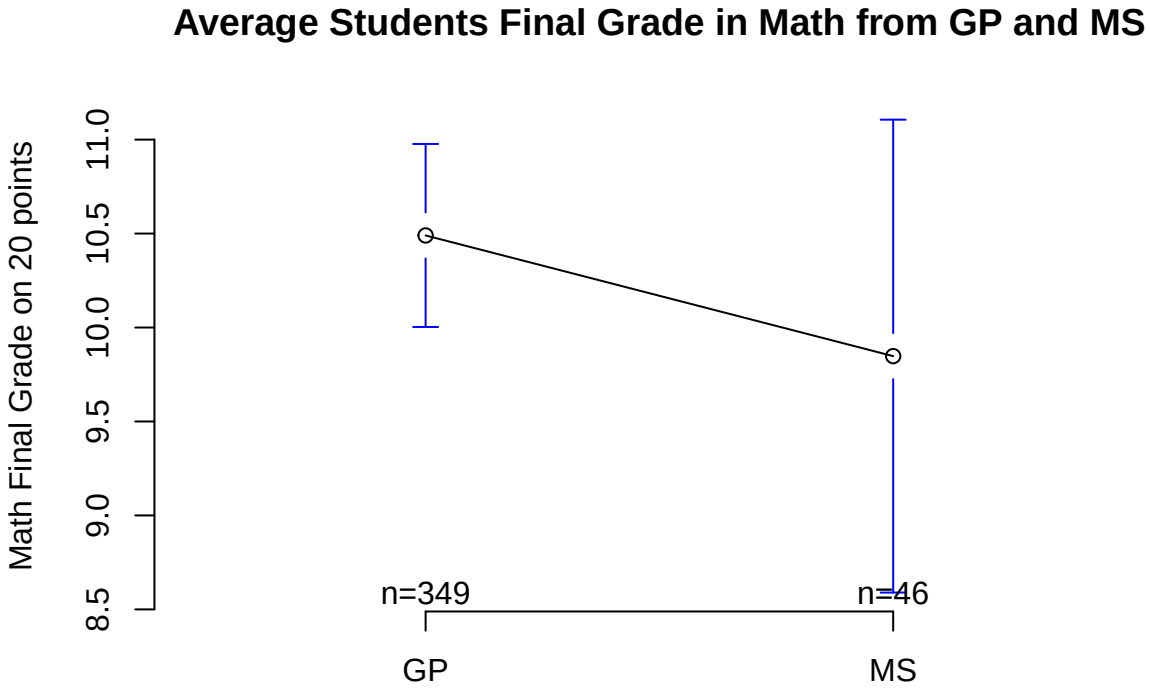
```
## student_math_GP$grade3
##        n  missing distinct
##      349        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency     17    76   143    59    54
## Proportion 0.049 0.218 0.410 0.169 0.155

## 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.
##    0.00    8.00   11.00   10.49   14.00   20.00
```

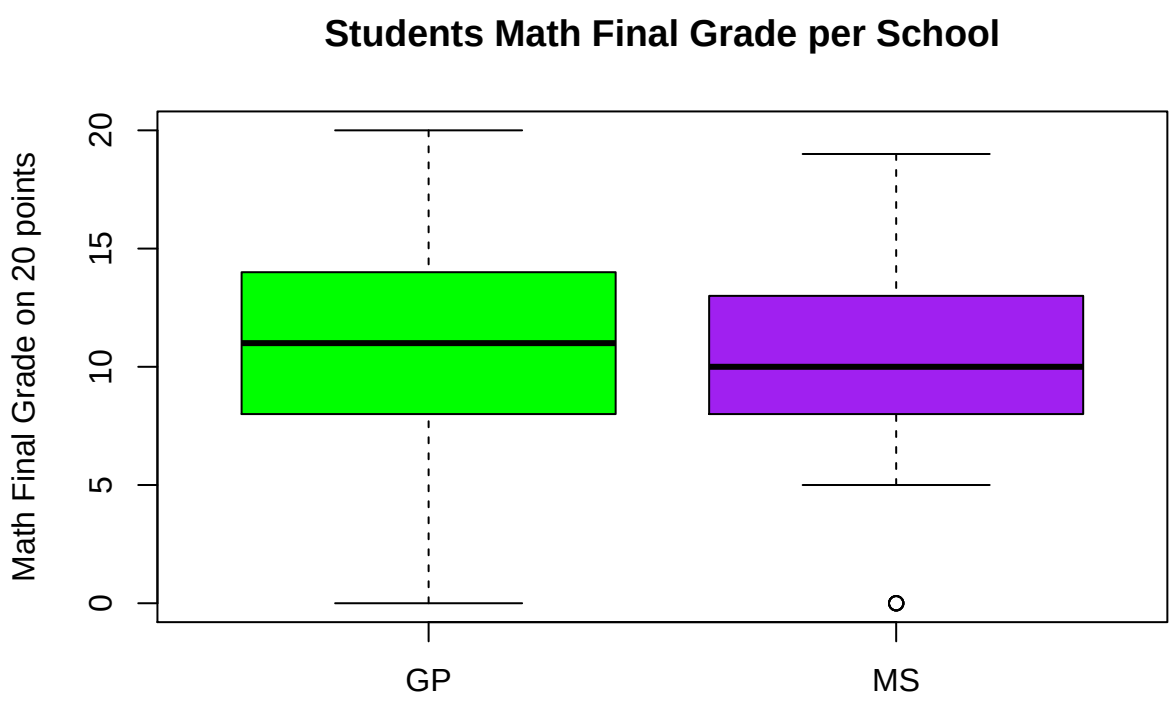
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   8.000  10.000   9.848  12.750  19.000
```

```
## 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
```

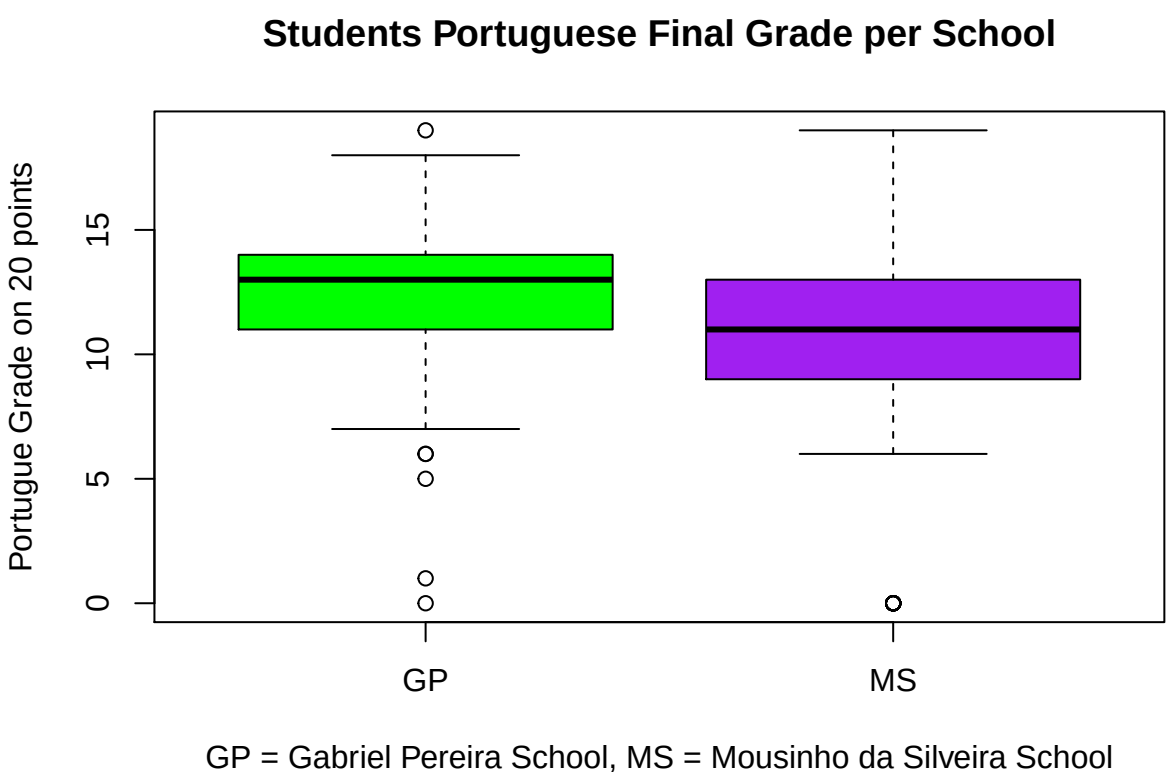
```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```





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

```
## student_portuguese_GP$grade3
##        n  missing distinct
##      423        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency     10   136   245    27     5
## Proportion 0.024 0.322 0.579 0.064 0.012

## student_portuguese_MS$grade3
##        n  missing distinct
##      226        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency      7    41   110    53    15
## Proportion 0.031 0.181 0.487 0.235 0.066
```



At this point, we can either explore students who did well in Math with a final grade of A or B to see if they do something than those who final grade below C. Alternative way is to explore students with the final grade of D or F.

Part 7 - Inference

Part 7a - Problem

- We have two interesting groups for our testing. Let's test the variable absence.
- There is a problem here: one is the two data frames (student_mathTop and student_mathBottom) which don't have the same dimension.
- Thus, the difference for variable absence will not be the same.
- One way to go about solving this issue is to pre-populate the variable (absence) with shorter lenght using insignificant number, but what if we want to test other variables as well.
- Second way, we can keep the original data frame "student_math", filter by grade.
- Then we assign a new variable which will be categorical of two type of students(T= top student [Final grade = A or B] and B = bottom student[Final grade = D or F]).
- This has the advantage of giving us more freedom in testing other variable on this new data frame.
- Before we impliment the above solution, let's use function summary to check the mean absence for each type of student.
- We are concerned that the difference in total number of students from each data frame might create a biais in this summary. But, at least summary will give us a glimpse of average absences.

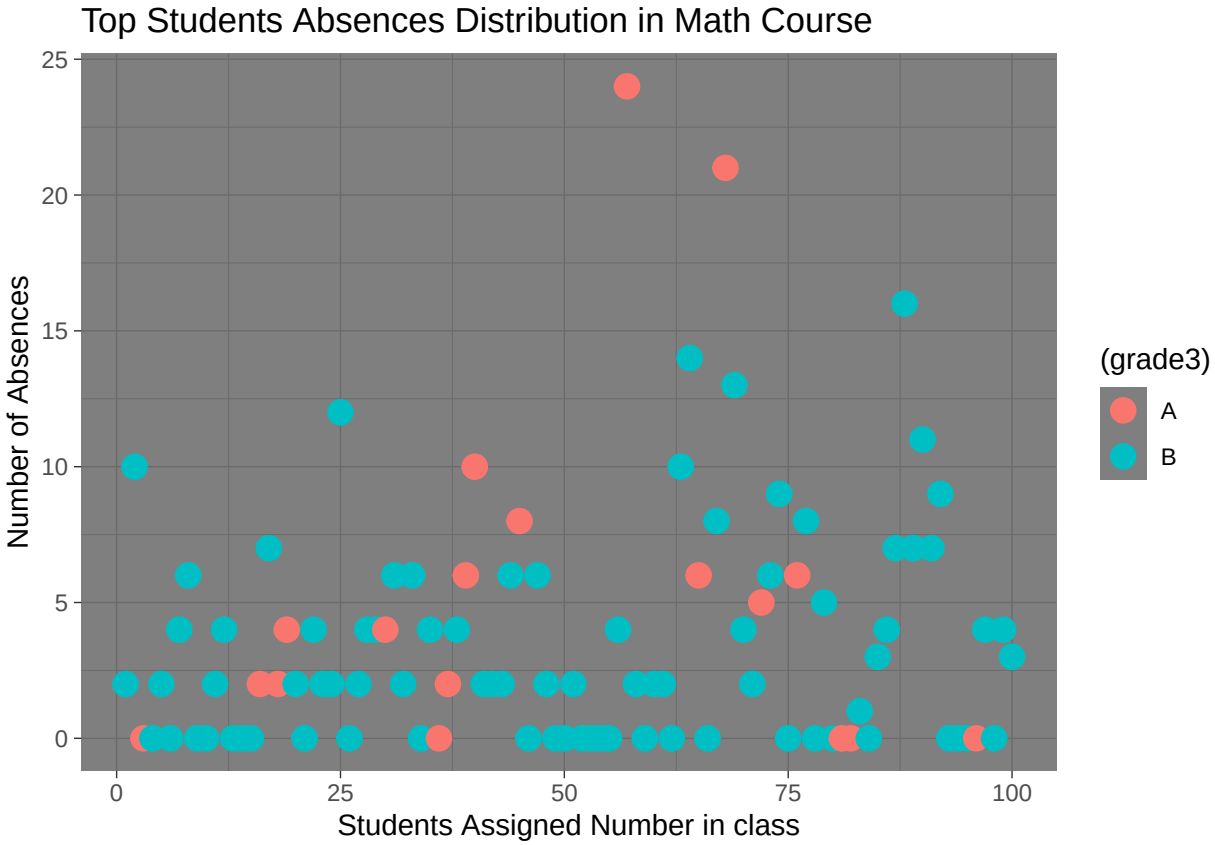
The average absence from the top students registered in Math course.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    2.00    3.78    6.00   24.00
```

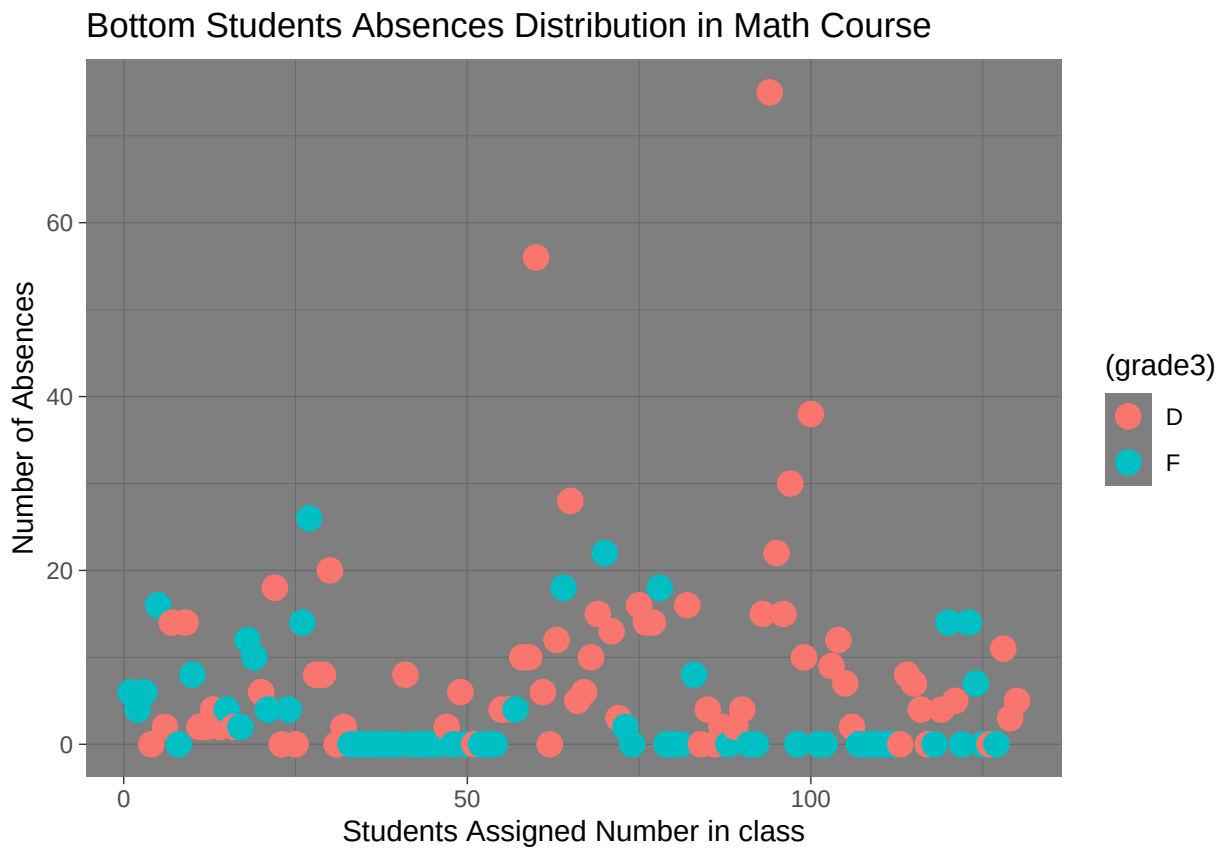
- The average absence from the Bottom students.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.000   6.762  10.000  75.000
```

- We can visualize this mean average by calling a scatter plot.
- To do this, we can call the function for dummy variable created above.

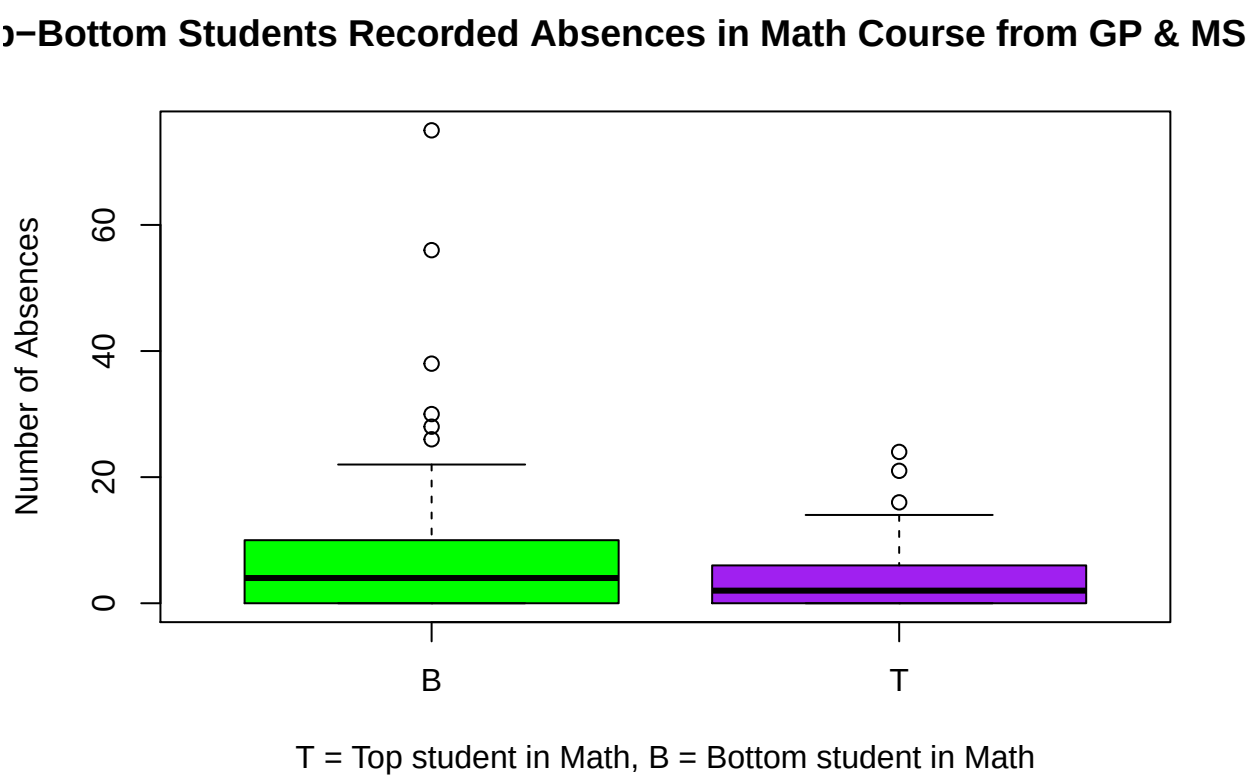


- Let's see the number of absences from the bottom students registered in Math Course.

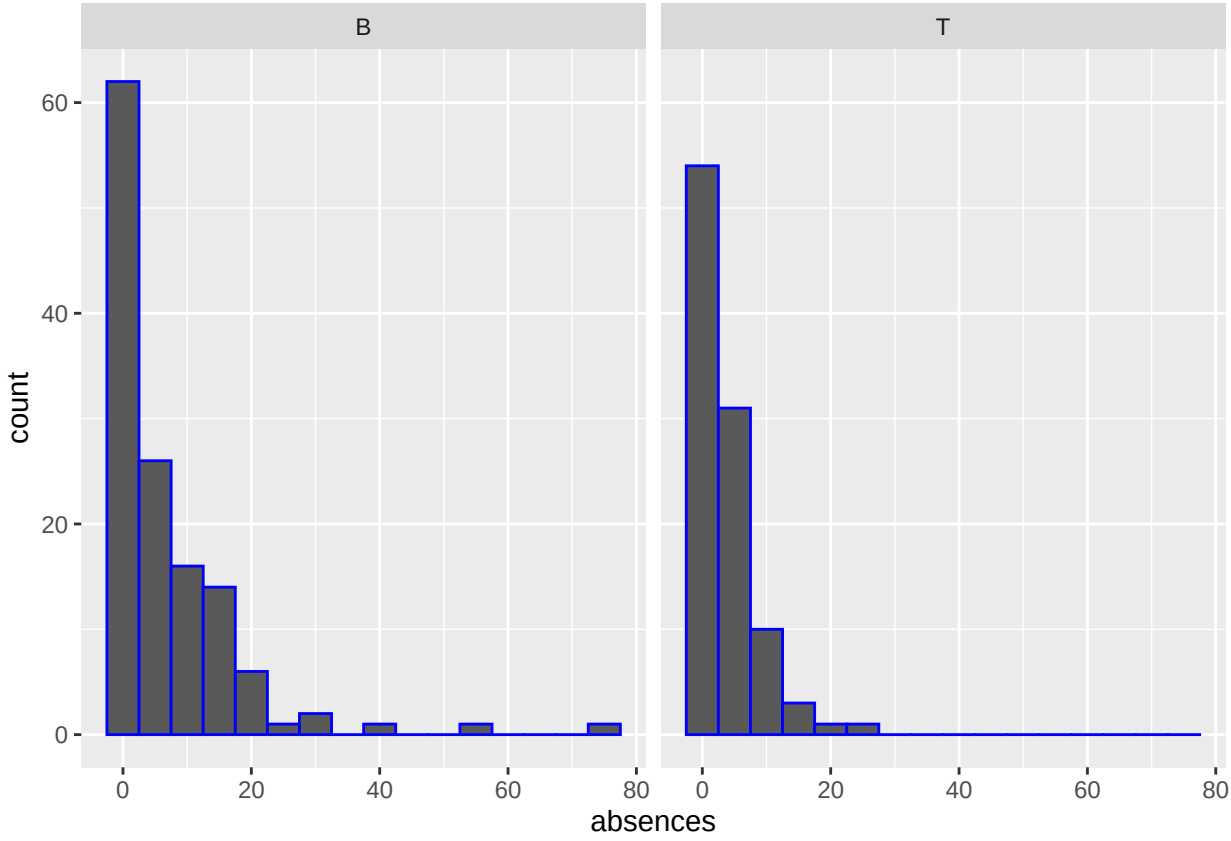


Now, let's create the two type of students.

- Question: is there a difference in absence between top and bottom students in Math course?


ɔ−Bottom Students Recorded Absences in Math Course from GP & MS
Number of Absences
0
20
40
60
B
T
T = Top student in Math, B = Bottom student in Math

Distributions of absences among top and bottom students registered in Math course.



- Right skewed!
- Let's calculate a 95% confidence interval for the average difference between number of absences for top and bottom students in Math course.
- Assumptions:

  Independence within groups. Independence between groups. Sample size/skew

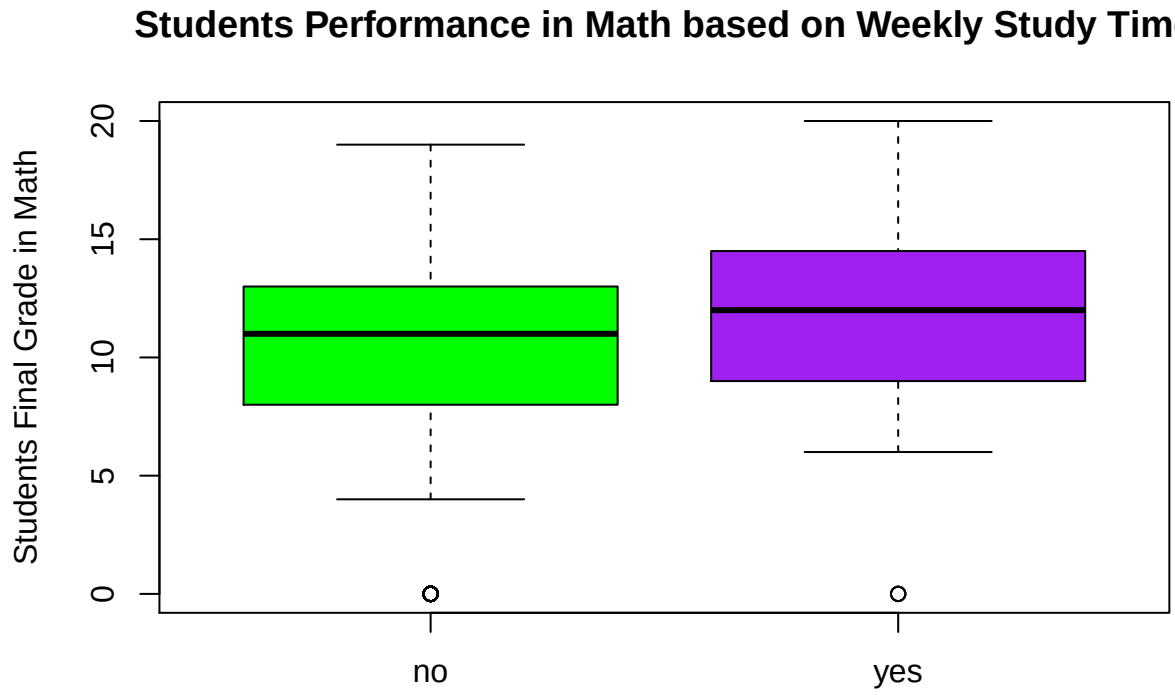
t-test

```
##
##  Welch Two Sample t-test
##
## data:  absences by TB
## t = 2.9118, df = 184.84, p-value = 0.004036
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9613824 5.0016945
## sample estimates:
## mean in group B mean in group T
##        6.761538        3.780000
```

Part 7b - 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 ration between students who study 10+ a week and those who don't in math course.

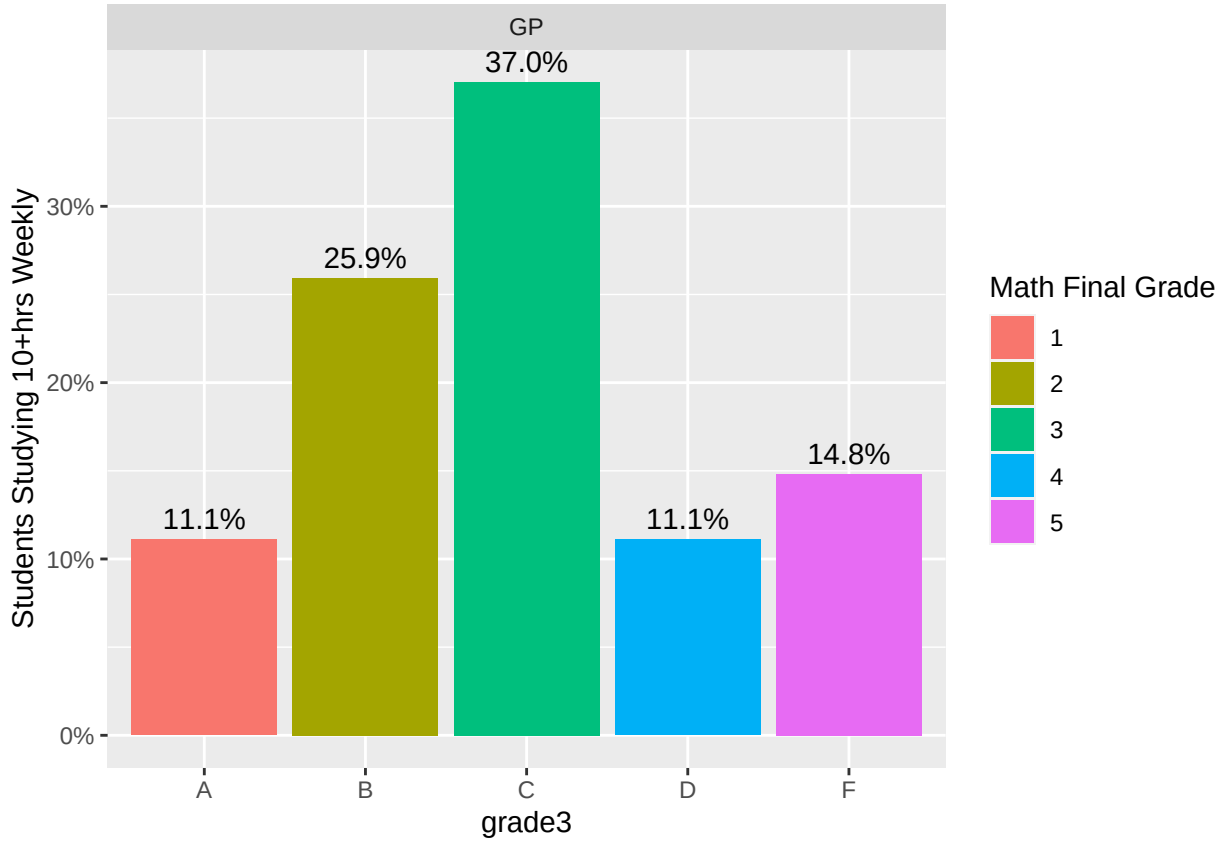
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

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

```
## study10plus$grade3
```

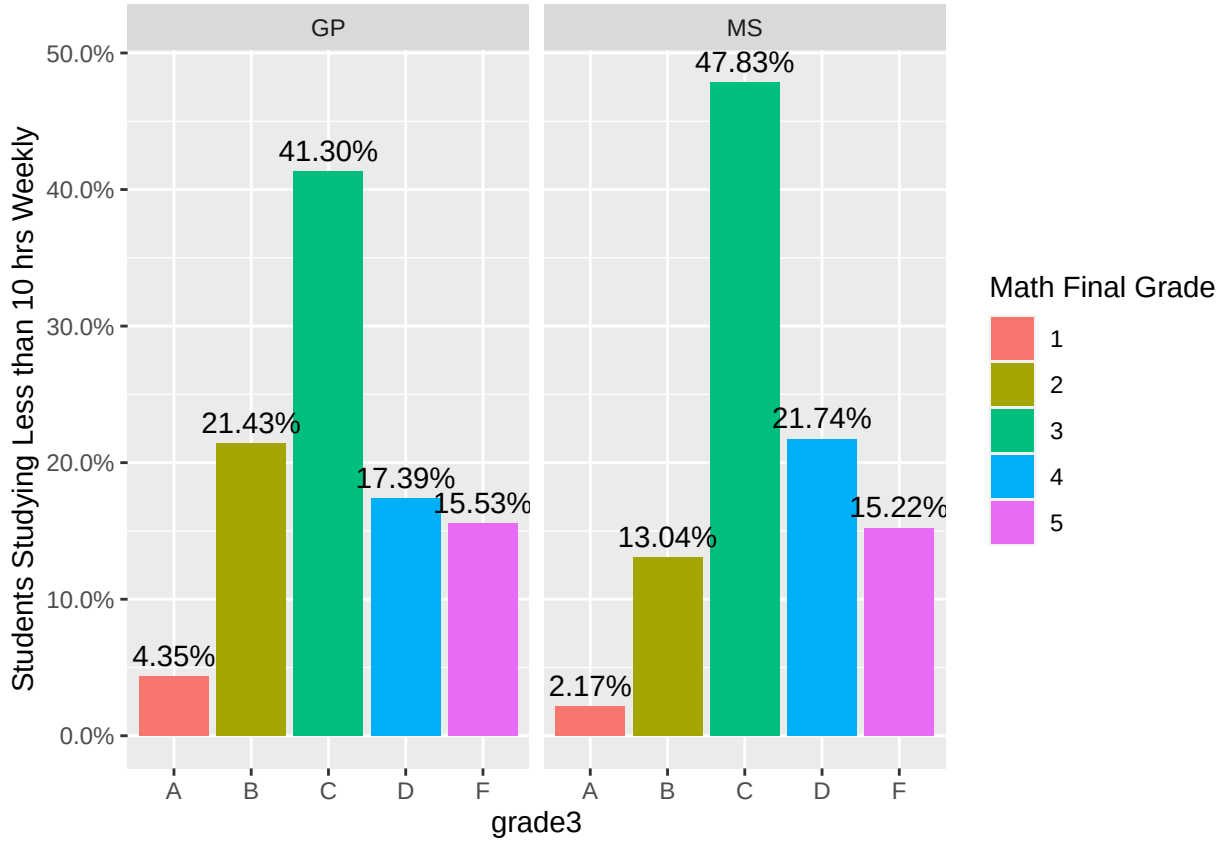
```
##        n  missing distinct
##       27        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency      3     7    10     3     4
## Proportion 0.111 0.259 0.370 0.111 0.148

##
## Let's see the math final grade distribution from the two schools based on 10+hrs weekly study time
```

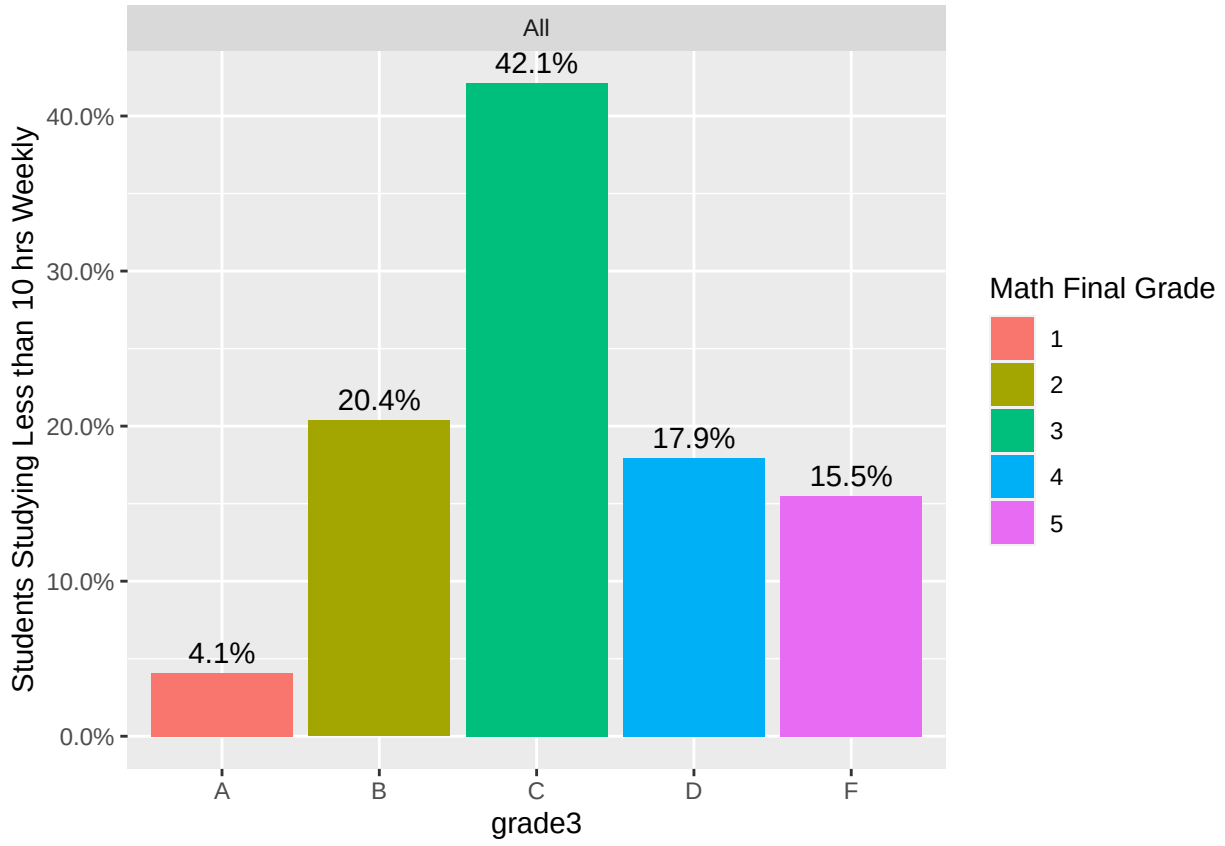


- Let's see the statical information about students final grade in Math based on less than 10hrs Weekly study time.

```
## 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 see the math final grade distribution from the two schools based on 10+hrs weekly study time
```





- Computing the hypothesis test.

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

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

```
## [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.

Part 8 - Conclusion

Part 8a - Findings

- 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 mean for students from GP in Math course is 10.49.
- Statistically, the mean 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.
- There is a significant difference in number of absences in Math course for top and bottom students.
- In order words, students who got a final grade of "A" or "B" in Math course were lesser absent than those who got fail the Math course(D and F grades).
- 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.

Part 8b - 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