The City University of New York School of Professional Studies

Data Acquisition and Management (DATA 607)

Final Project
Data Insights to Improve school Education System

Alexis Mekueko and DH Kim
email: alexis.mekueko08@login.cuny.edu

11/13/2020

Github Link: https://github.com/asmozo24/DATA607_Final_Project
Web link: https://rpubs.com/amekueko/697306

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

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

Grading the professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, “Beauty in the classroom: instructors’ pulchritude and putative pedagogical productivity” by Hamermesh and Parker found that instructors who are viewed to be better looking receive higher instructional ratings.

Here, you will analyze the data from this study in order to learn what goes into a positive professor evaluation.

Getting Started

Load packages

In this lab, you will explore and visualize the data using the tidyverse suite of packages. The data can be found in the companion package for OpenIntro resources, openintro.

Let's load the packages.

This is the first time we're using the GGally package. You will be using the ggpairs function from this package later in the lab.

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 factor. 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 corroletion that could lead to students failure. If none, then we would like to determine the factors which contribute for the most to success. This is done in order for the school education system to keep track of success and improve the factors that negatively impact students success.

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 speak or reveal 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.

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 some popular assumption out there. For instance, there are some studies out there suggesting that study time likely affects students performance. Let's verify that in this study. Do students studying at least 10hrs weekly do well in Math course than those spending lesser time?

1. 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, I used excel with one attemp to fix it. I am interested in the student taking Math course. with 33 variables. Data available –
> https://github.com/asmozo24/DATA606_Project_Proposal

Using R to acquire data

Using SQL to acquire data

2. Data Preparation / Data Wrangling

Cleaning 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, ...
```

```
## '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 ...
```

```
## [1] "Data frame is composed of character, boolean and numerical."
```

```
## [1] "Let's convert all chr type to factor and int type to numeric"
```

```
## [1] 395  33
```

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

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

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

```
- organizing data
- Tidying up data
```

3. Explore Data Let’s take a look at the data frame…

	school <chr>	… <chr>	… <int>	address <chr>	famsize <chr>	Pstatus <chr>	M… <int>	Fe… <int>
1	GP	F	18	U	GT3	A	4	4
2	GP	F	17	U	GT3	T	1	1
3	GP	F	15	U	LE3	T	1	1
4	GP	F	15	U	GT3	T	4	2
5	GP	F	16	U	GT3	T	3	3
6	GP	M	16	U	LE3	T	4	3

6 rows | 1-9 of 35 columns

	school <chr>	… <chr>	… <int>	address <chr>	famsize <chr>	Pstatus <chr>	M… <int>	Fe… <int>
1	GP	F	18	U	GT3	A	4	4
2	GP	F	17	U	GT3	T	1	1
3	GP	F	15	U	LE3	T	1	1
4	GP	F	15	U	GT3	T	4	2
5	GP	F	16	U	GT3	T	3	3
6	GP	M	16	U	LE3	T	4	3

6 rows | 1-9 of 35 columns

The data frame presents about 30 factors and 03 variables (G1, G2 and G3). These 03 variables are interesting as there are students's grades.
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)

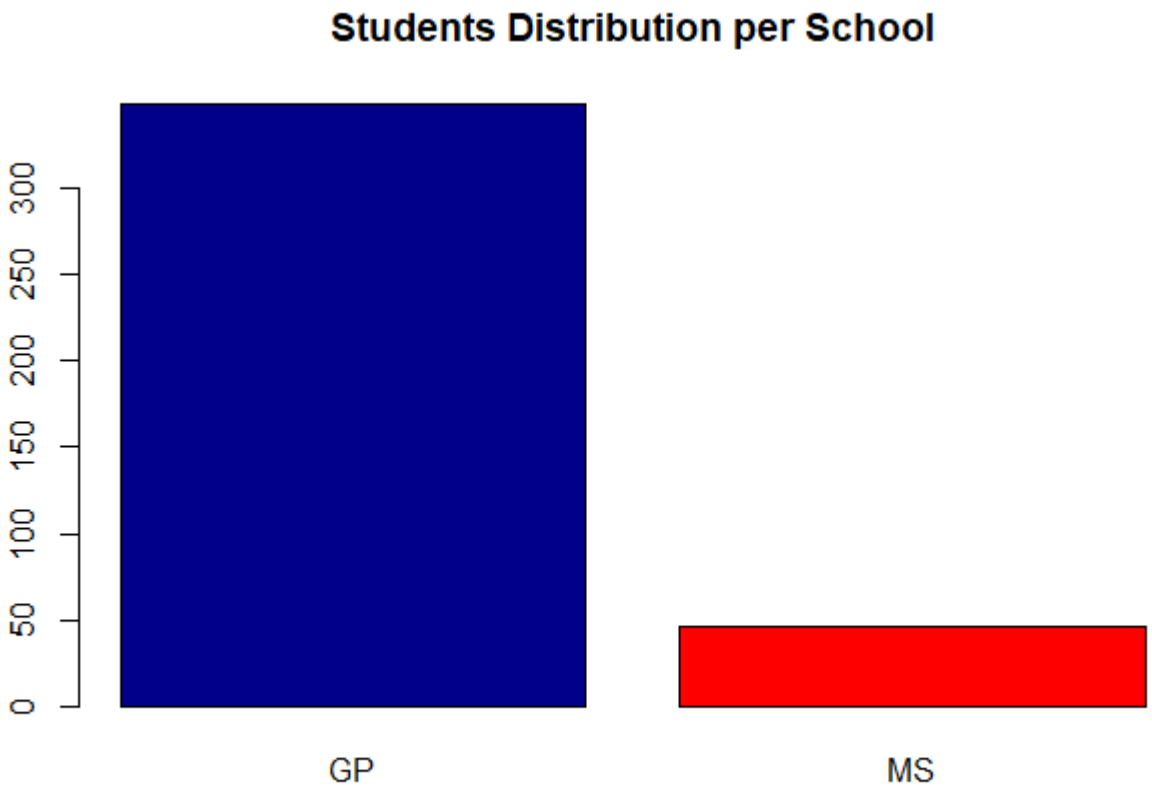
G1: first period grade (numeric: from 0 to 20) G2: second period grade (numeric: from 0 to 20) G3: final grade (numeric: from 0 to 20)

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.

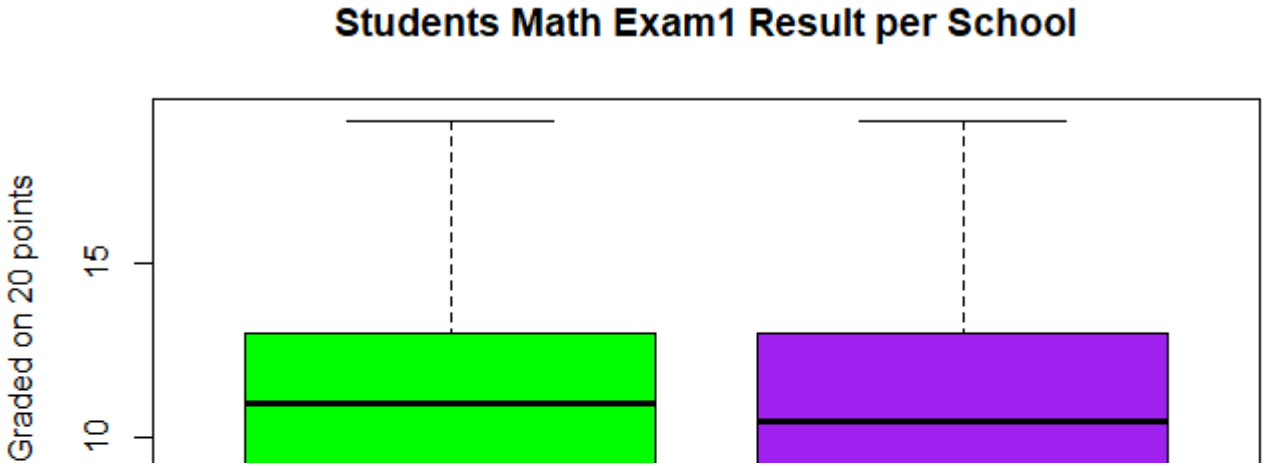
Let's take a closer look at these 03 variables. We might throw in a biais by neglecting the fact that there are two schools in the data frame. How significant is each school into the data frame.

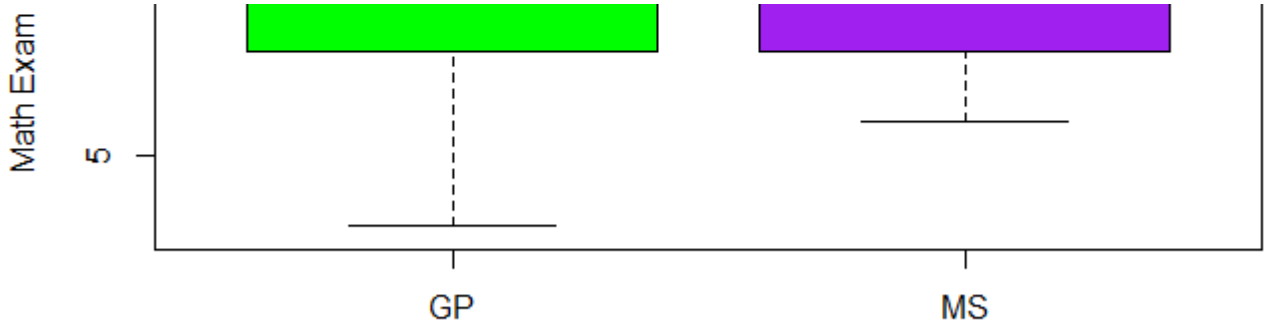
```
## [1] "Students dristribution from each school are: 88.4% students for Gabriel Pereira School
```

```
## student_math$school
##        n  missing distinct
##      395        0        2
##
## Value         GP    MS
## Frequency    349    46
## Proportion 0.884 0.116
```



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





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

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

	school <chr>	… <chr>	… <int>	address <chr>	famsize <chr>	Pstatus <chr>	Medu <int>	Fedu <int>
1	MS	M	18	R	GT3	T	3	2
2	MS	M	19	R	GT3	T	1	1
3	MS	M	17	U	GT3	T	3	3
4	MS	M	18	U	LE3	T	1	3
5	MS	M	19	R	GT3	T	1	1
6	MS	M	17	R	GT3	T	4	3

6 rows | 1-9 of 38 columns

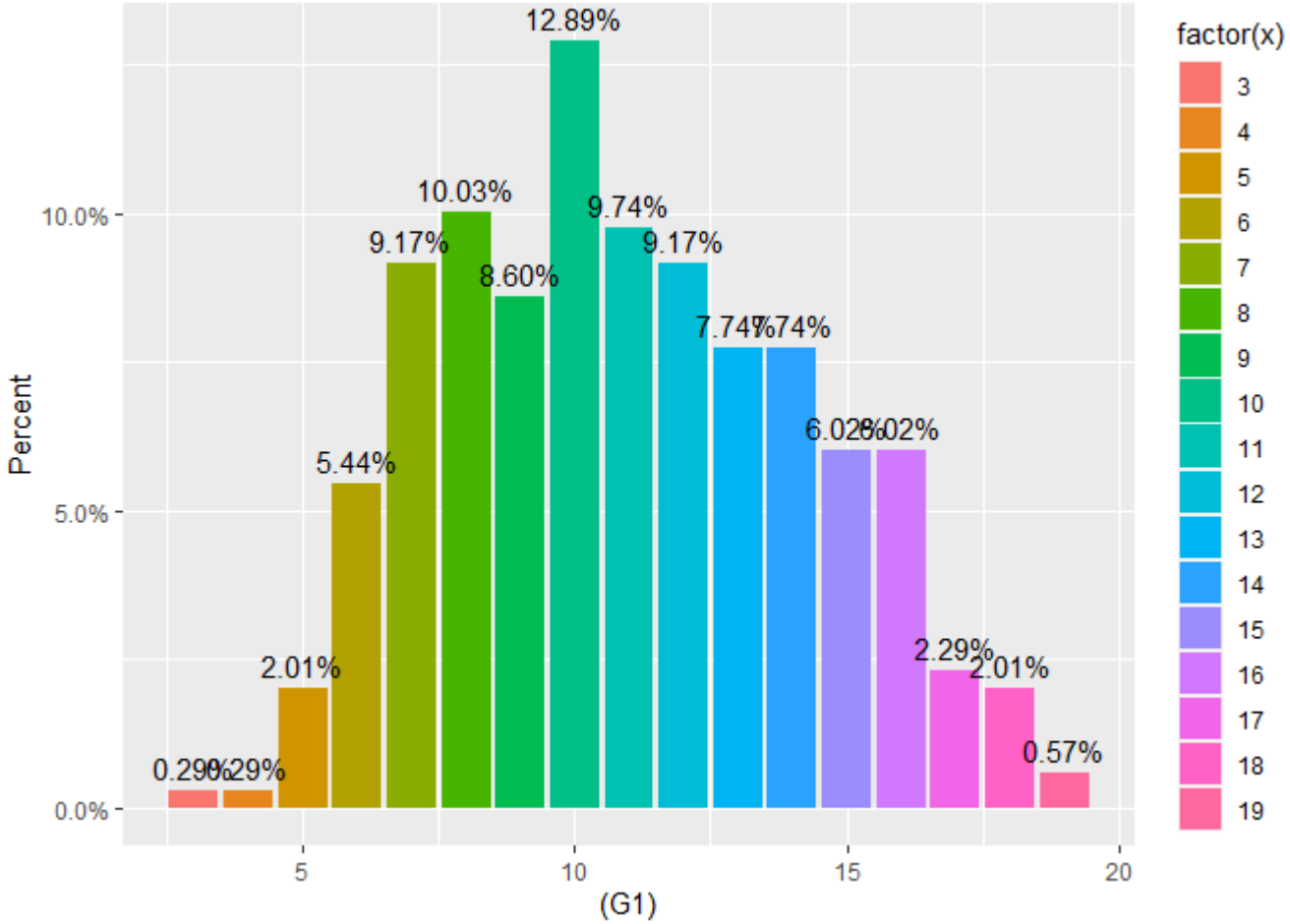
```
## Let's do summary on Math result 1 for students from Gabriel Pereira School
```

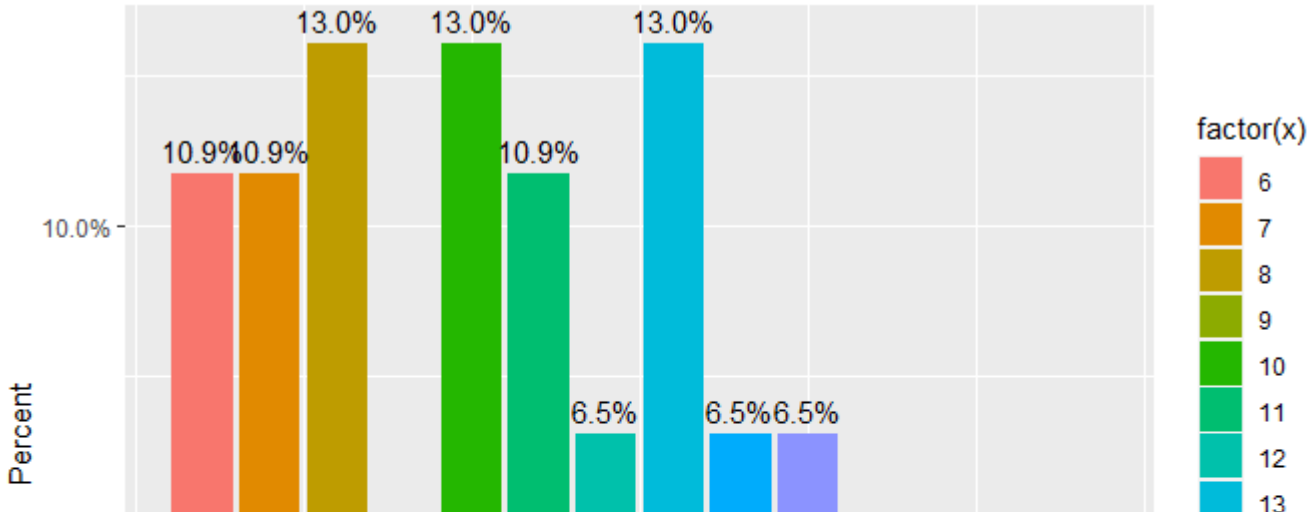
```
## student_math_GP$G1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      349        0       17    0.992    10.94    3.791        6        7
##      .25      .50      .75      .90      .95
##        8       11       13       16       16
##
## lowest :  3  4  5  6  7,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

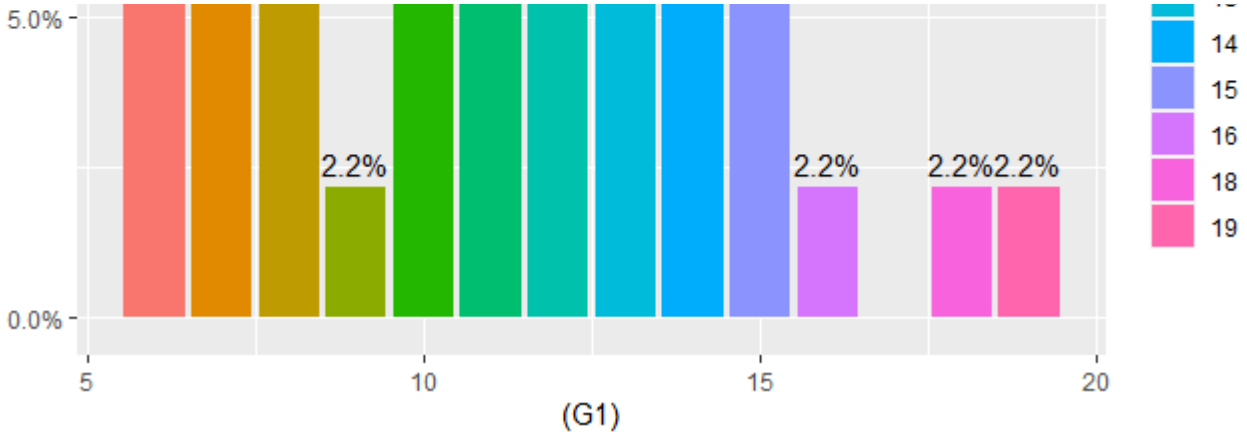
```
##
## Let's see the mean, max for students from Gabriel Pereira School
```

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

4. Data Analysis

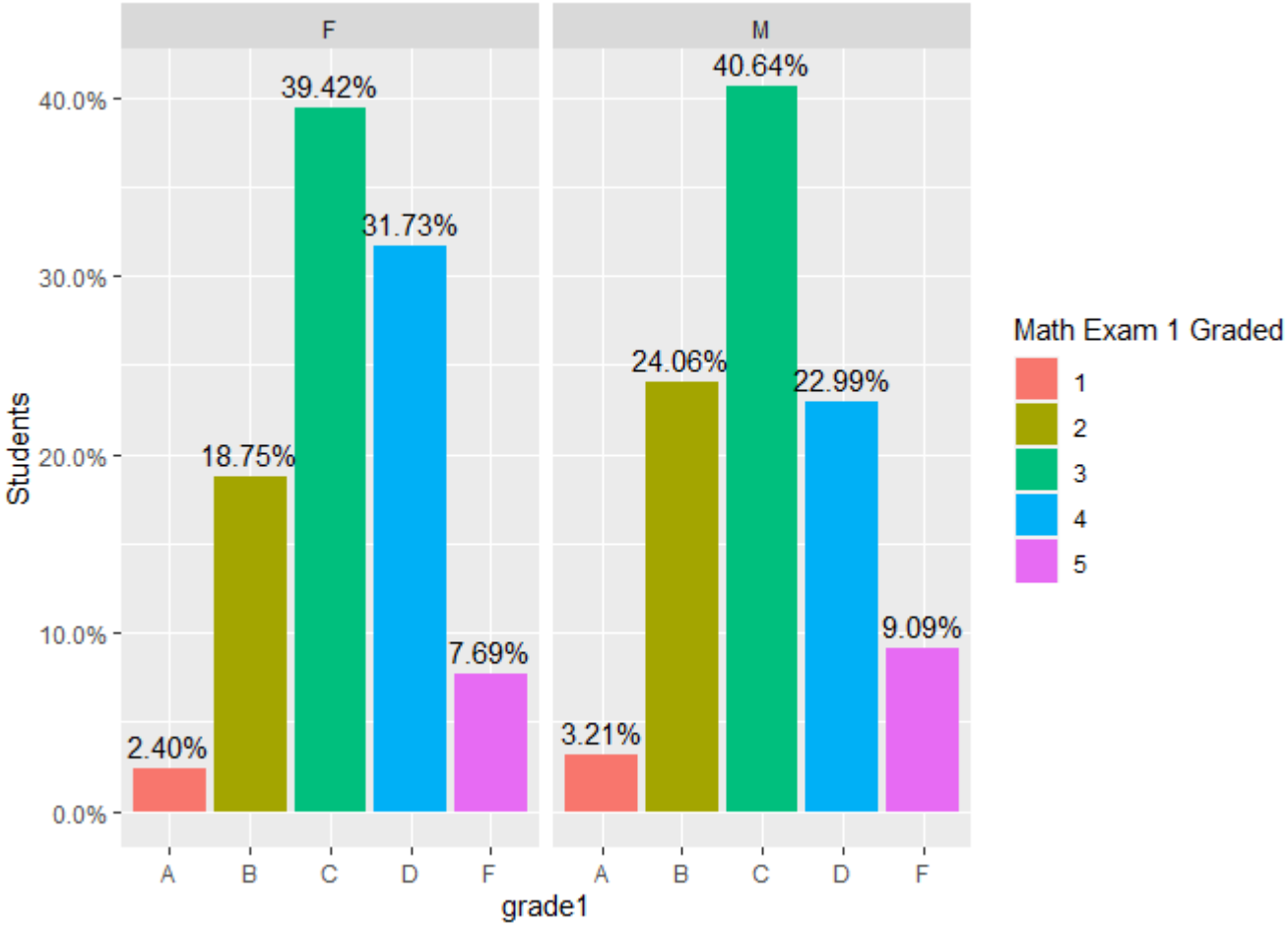

12.89%
10.03%
9.74%
9.17%
9.17%
8.60%
5.44%
2.01%
2.29%
2.01%
0.57%
10.0%
5.0%
0.0%
Percent
5
10
15
20
(G1)
factor(x)
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


13.0%
13.0%
13.0%
10.9%
10.9%
10.9%
6.5%
6.5%
6.5%
10.0%
Percent
factor(x)
6
7
8
9
10
11
12
13

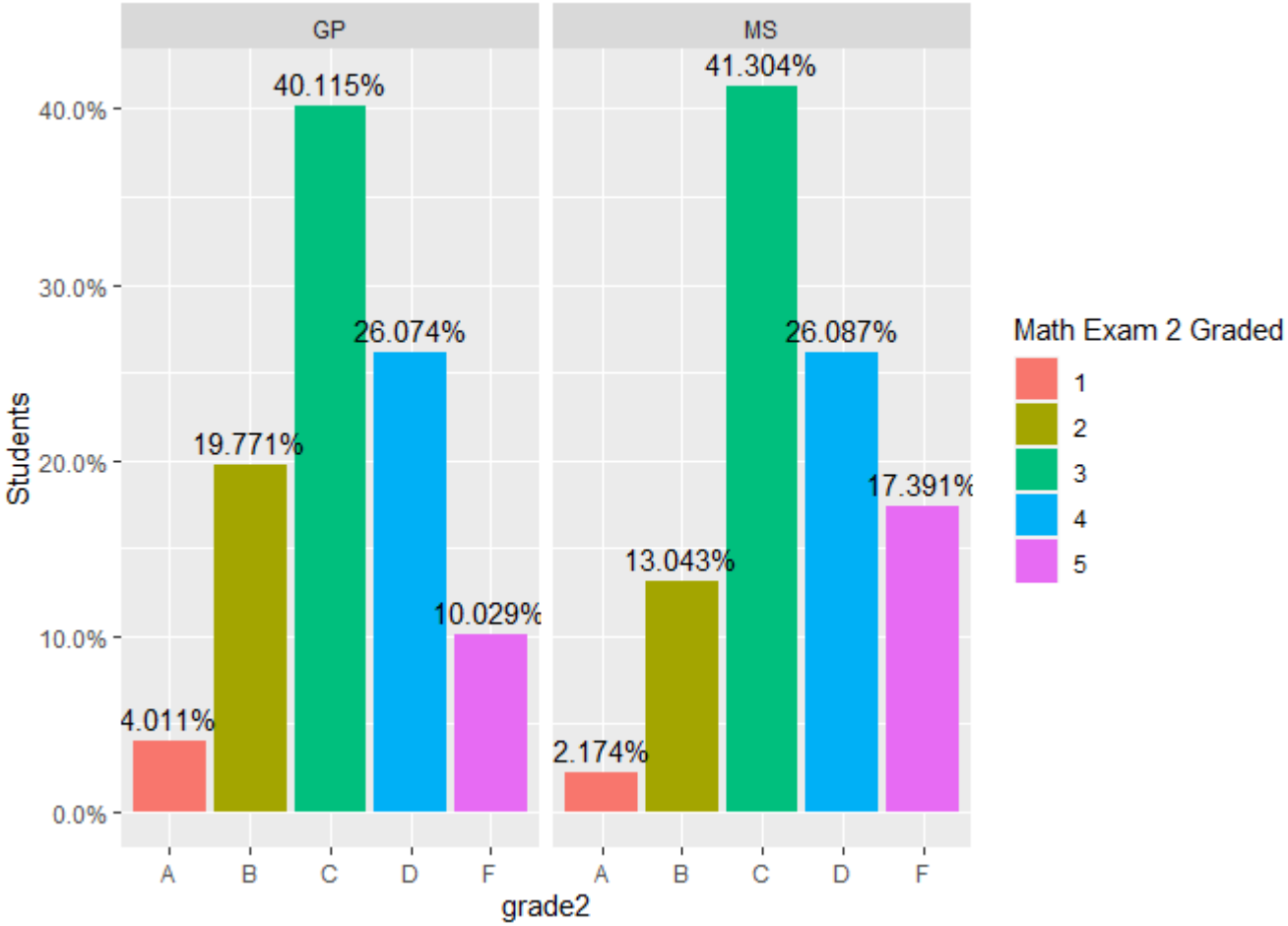



```
## A better representation is graded letters
```

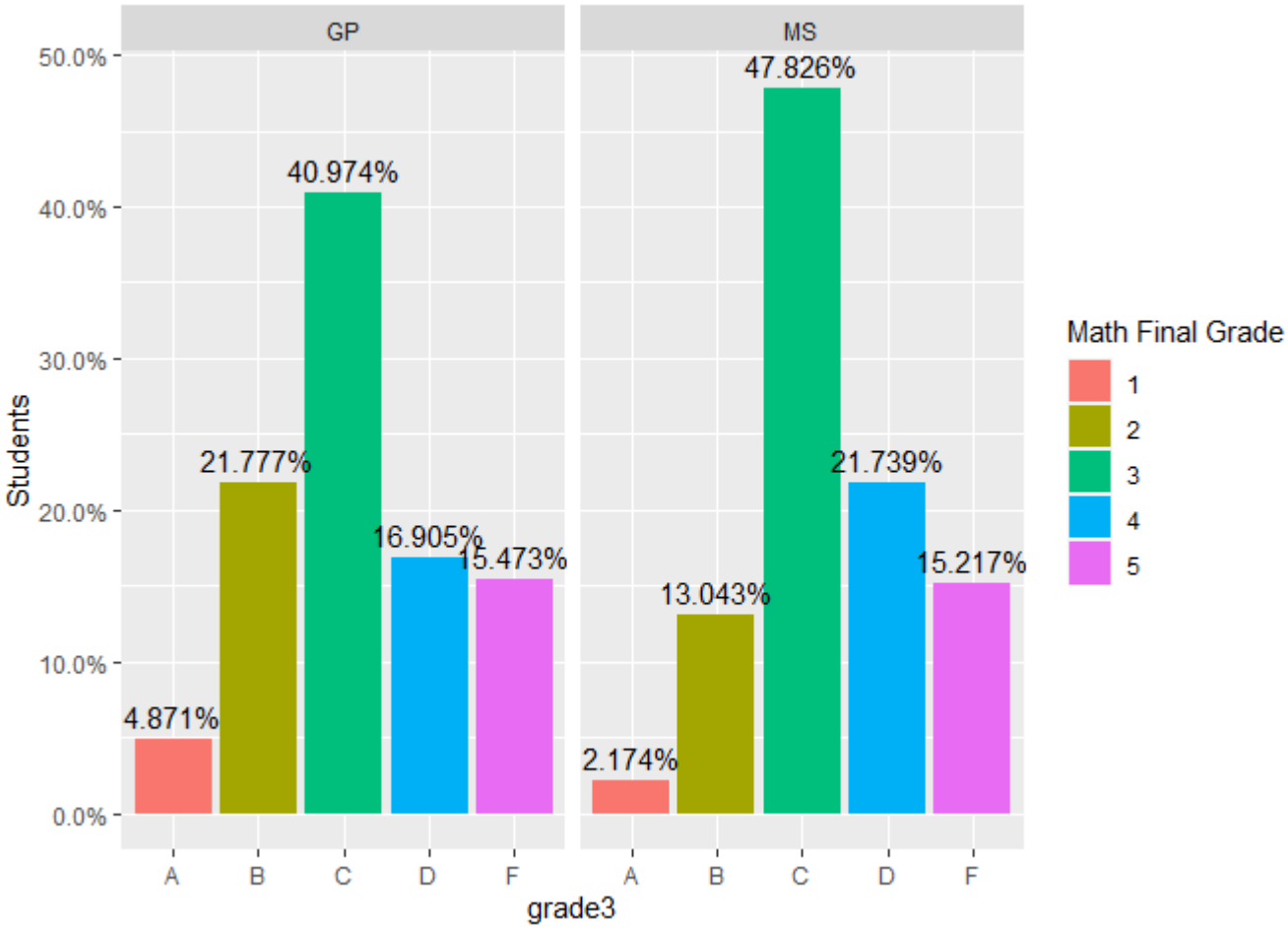
```
##
## Let's see the math exam1 graded from the two schools
```



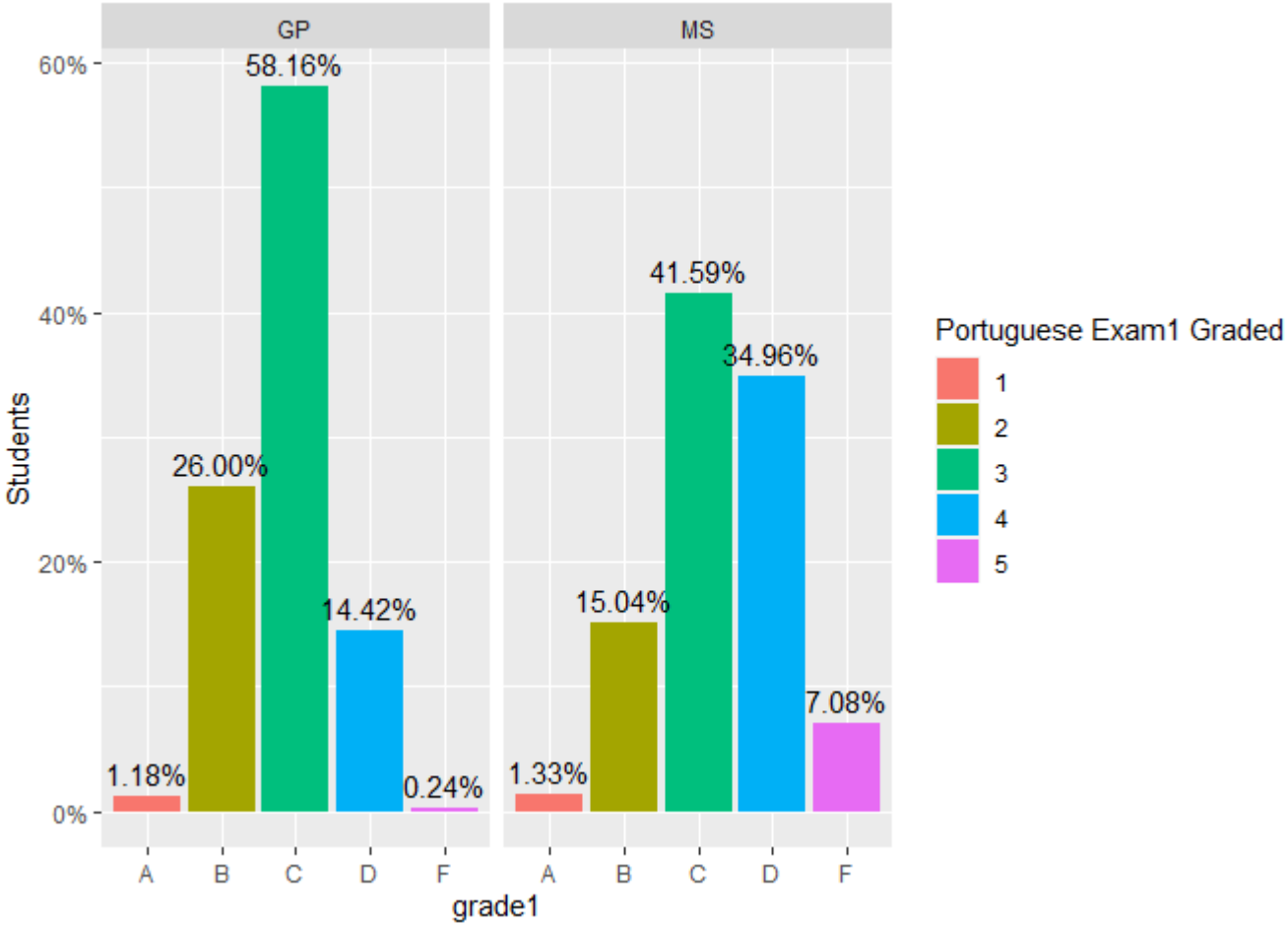

```
##
## Let's see the math exam2 graded from the two schools
```



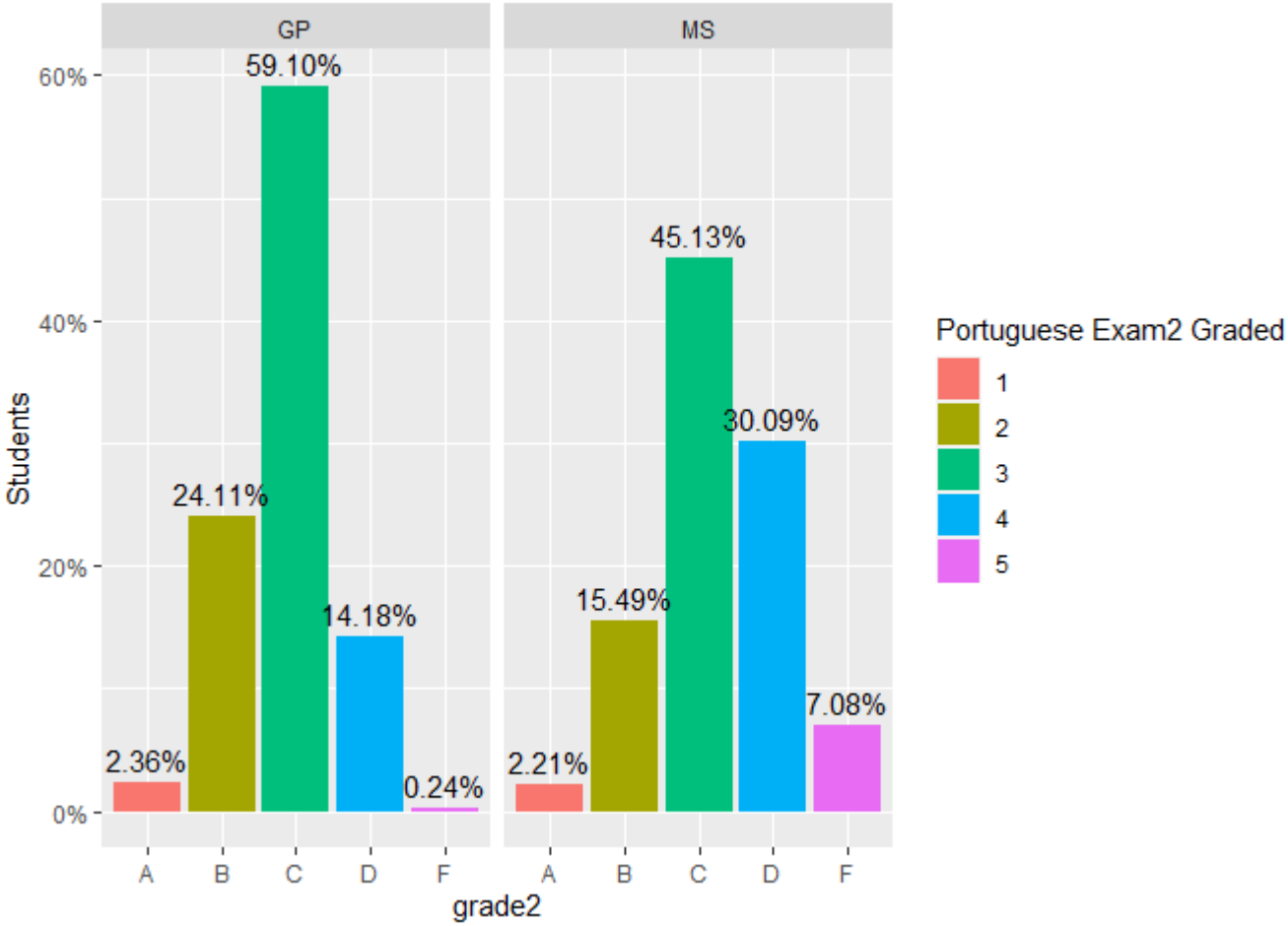

```
##
## Let's see the math final grade from the two schools
```



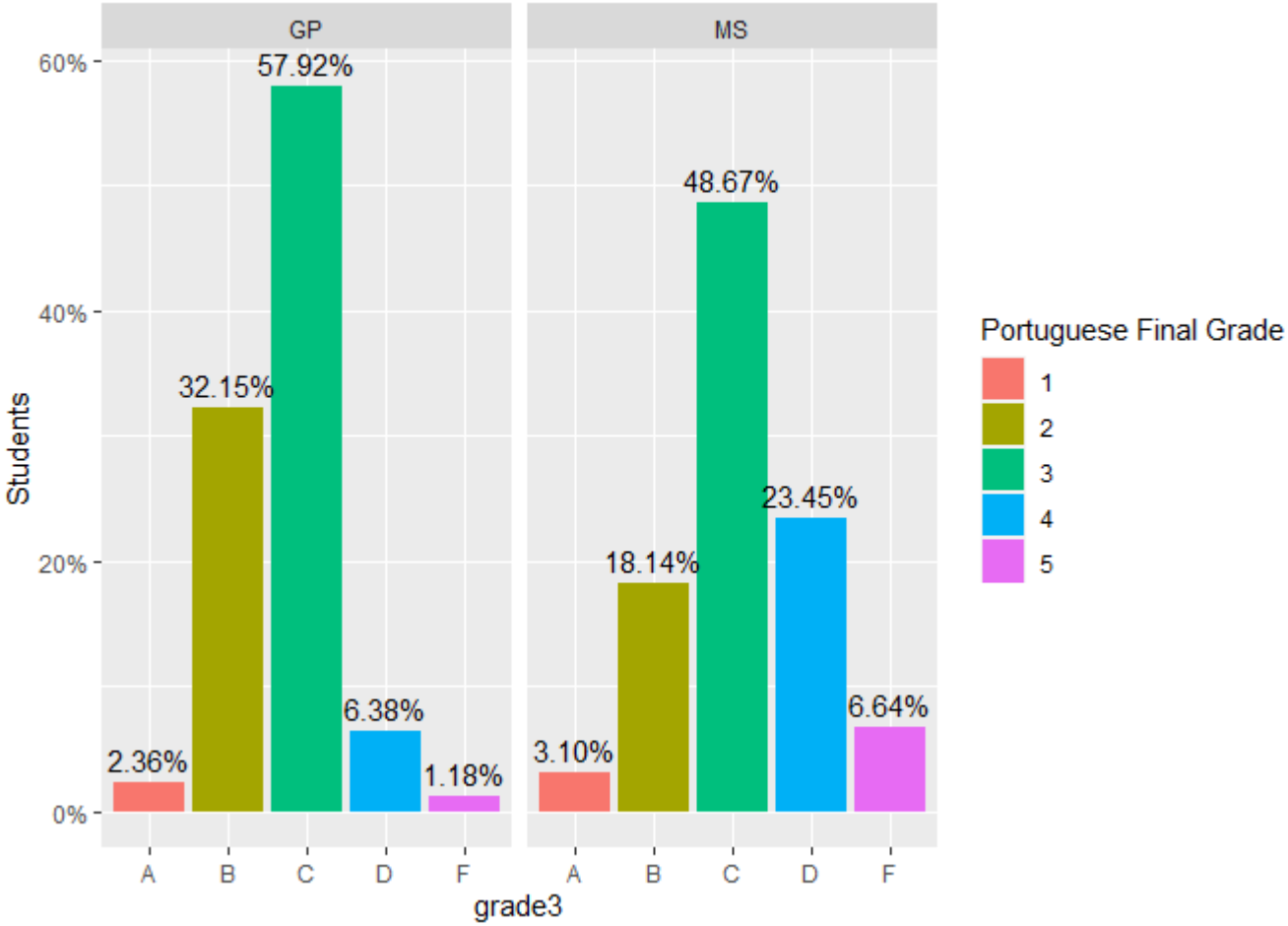

```
##
## Let's see the Portuguese Exam1 graded from the two schools
```



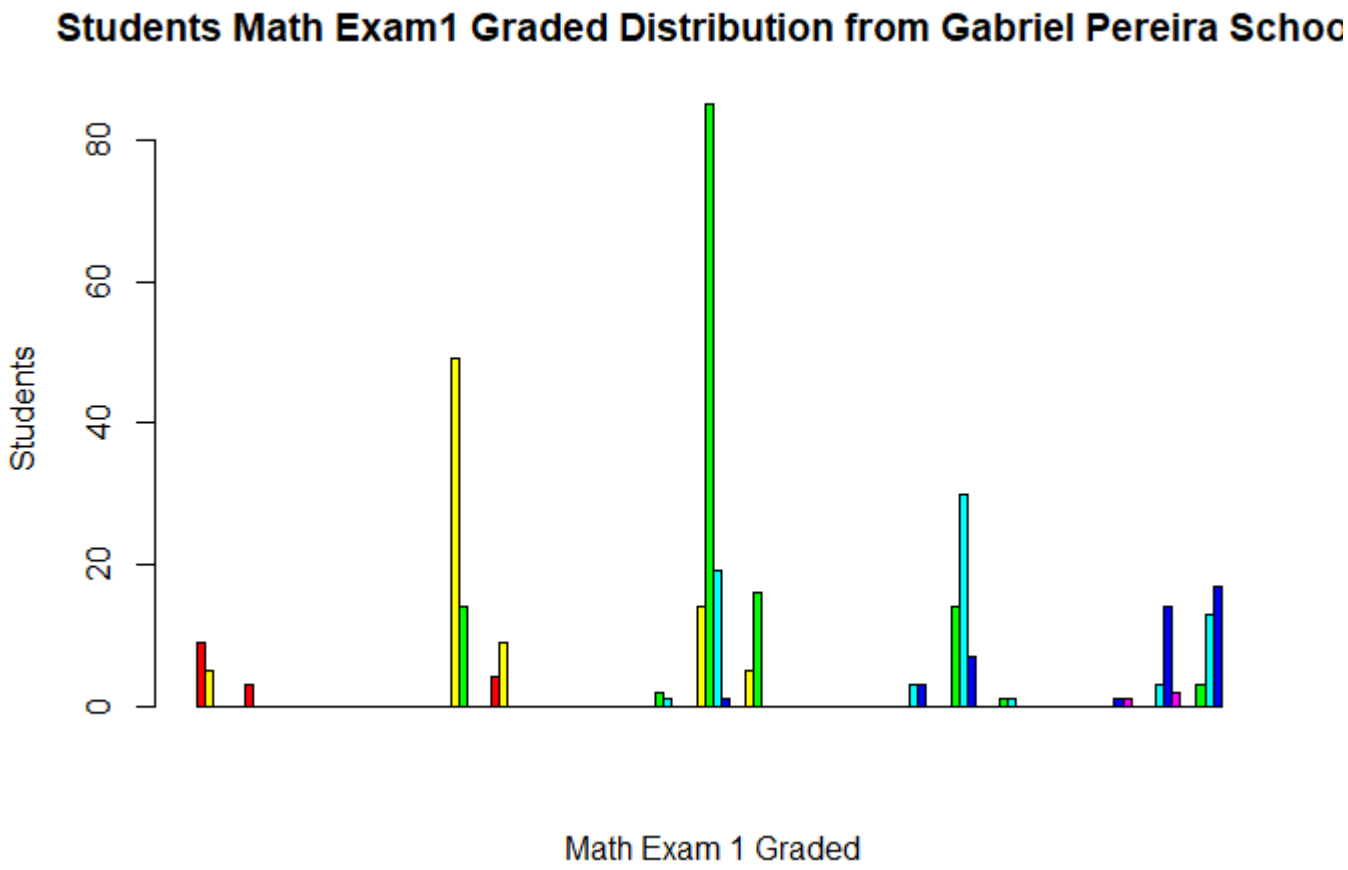

```
##
## Let's see the Portuguese Exam2 graded from the two schools
```




```
##
## 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




```
## 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
```

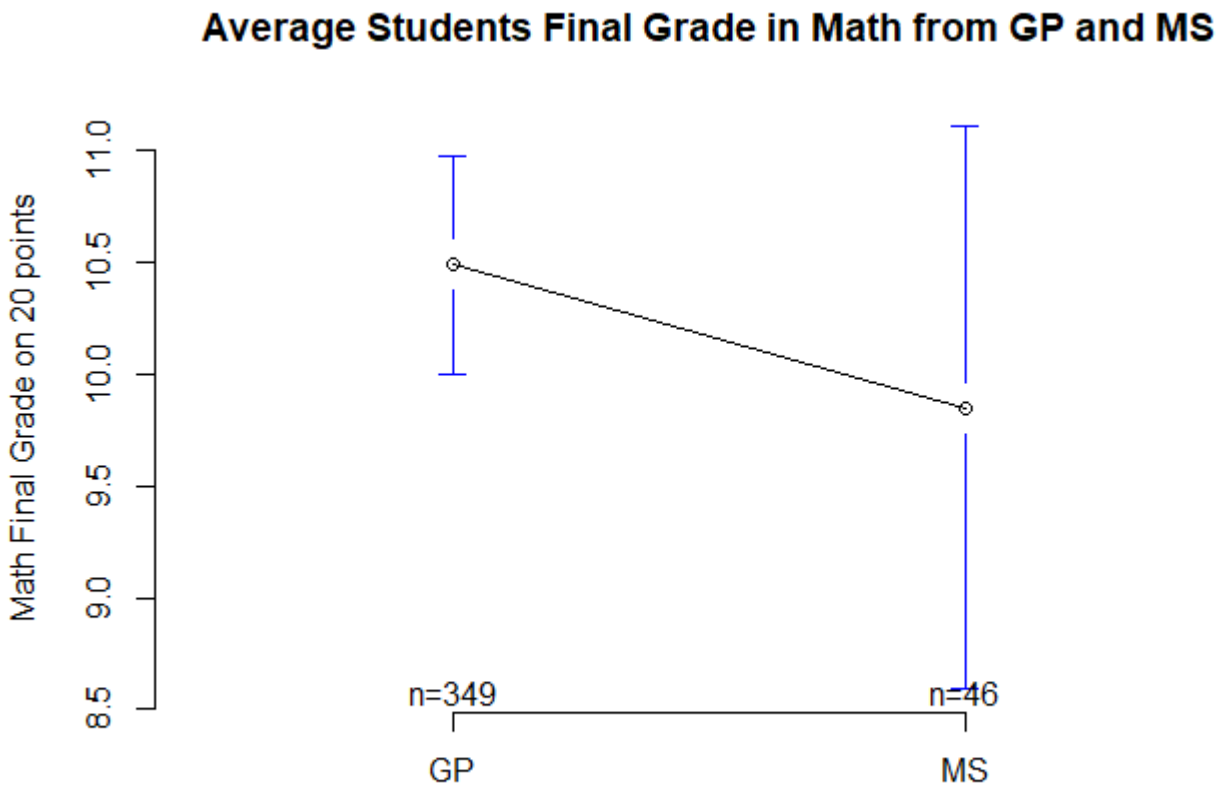
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    8.00   11.00   10.49   14.00   20.00
```

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

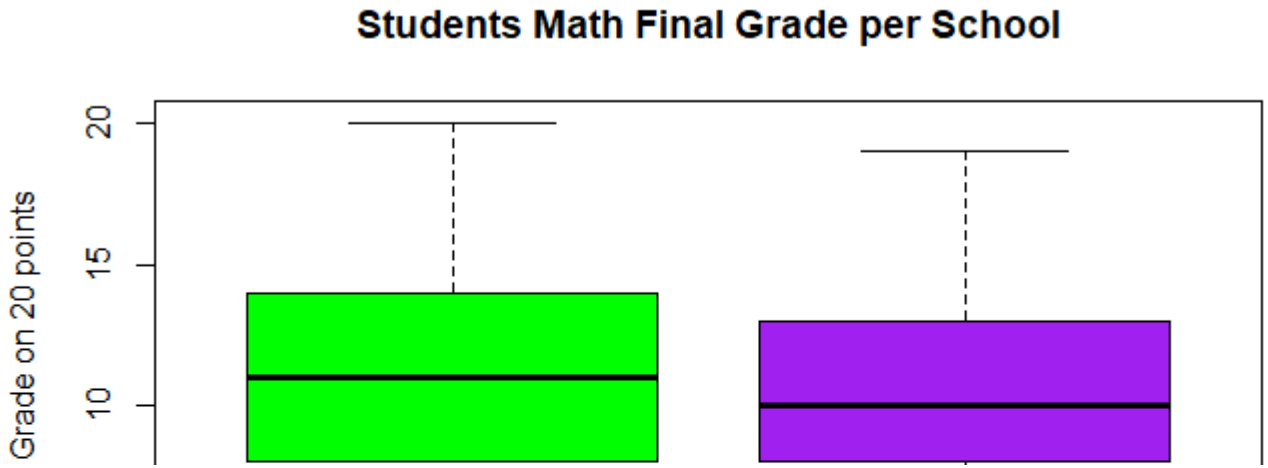
```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```

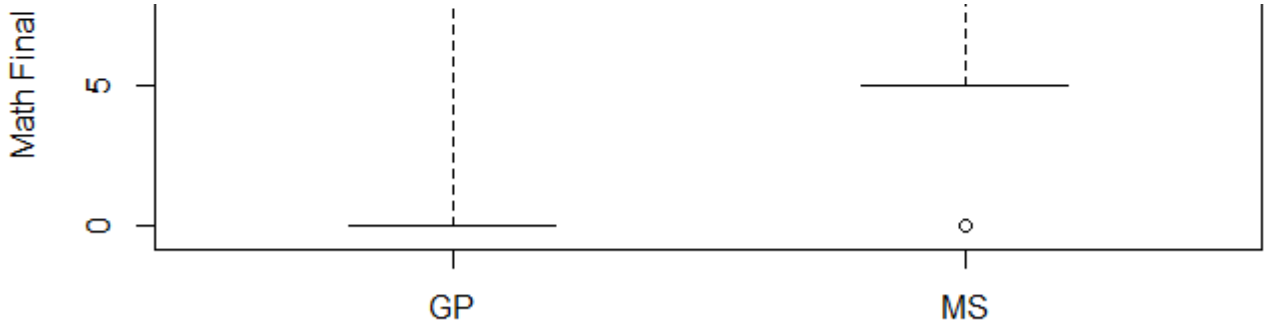
```
## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
## a graphical parameter
```

```
## 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

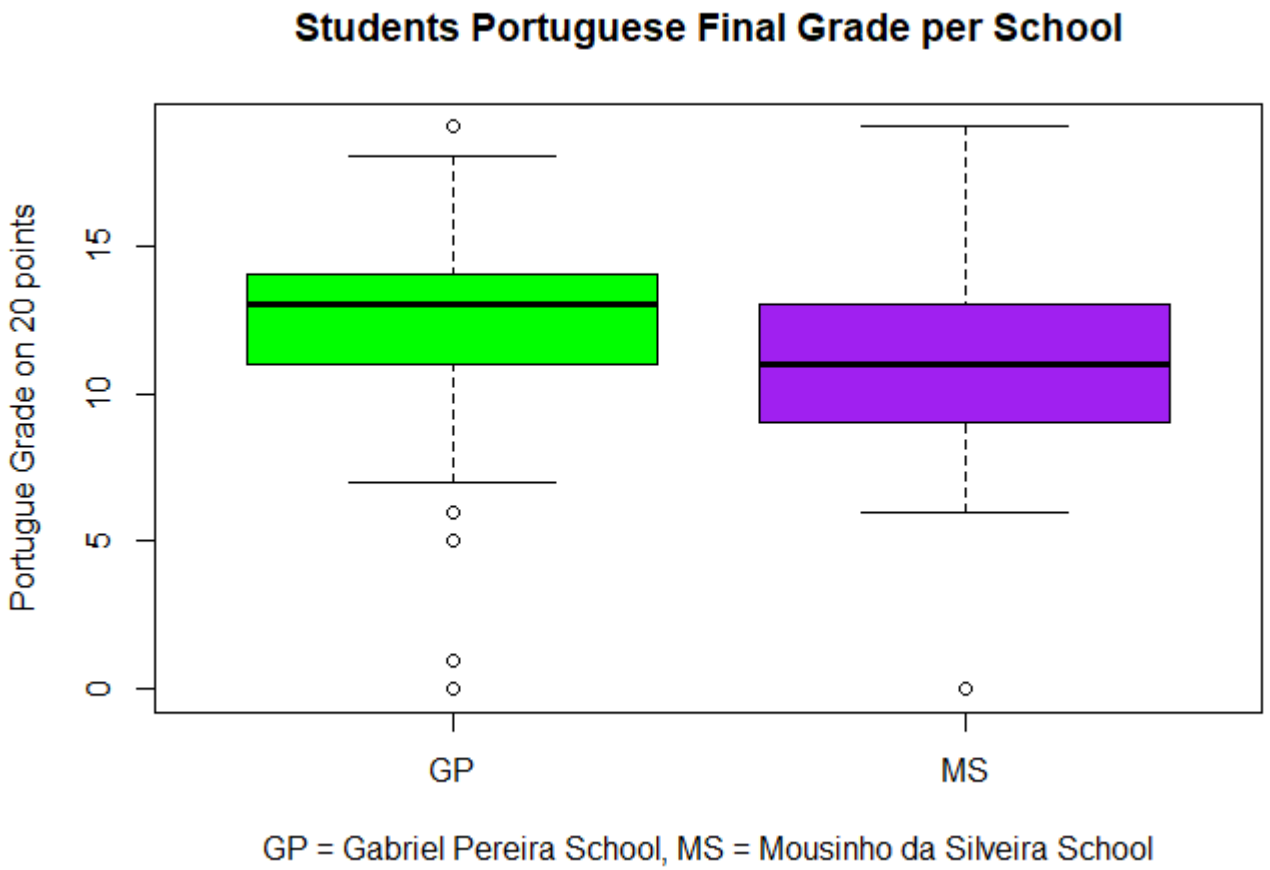




GP = Gabriel Pereira School, MS = Mousinho da Silveira School

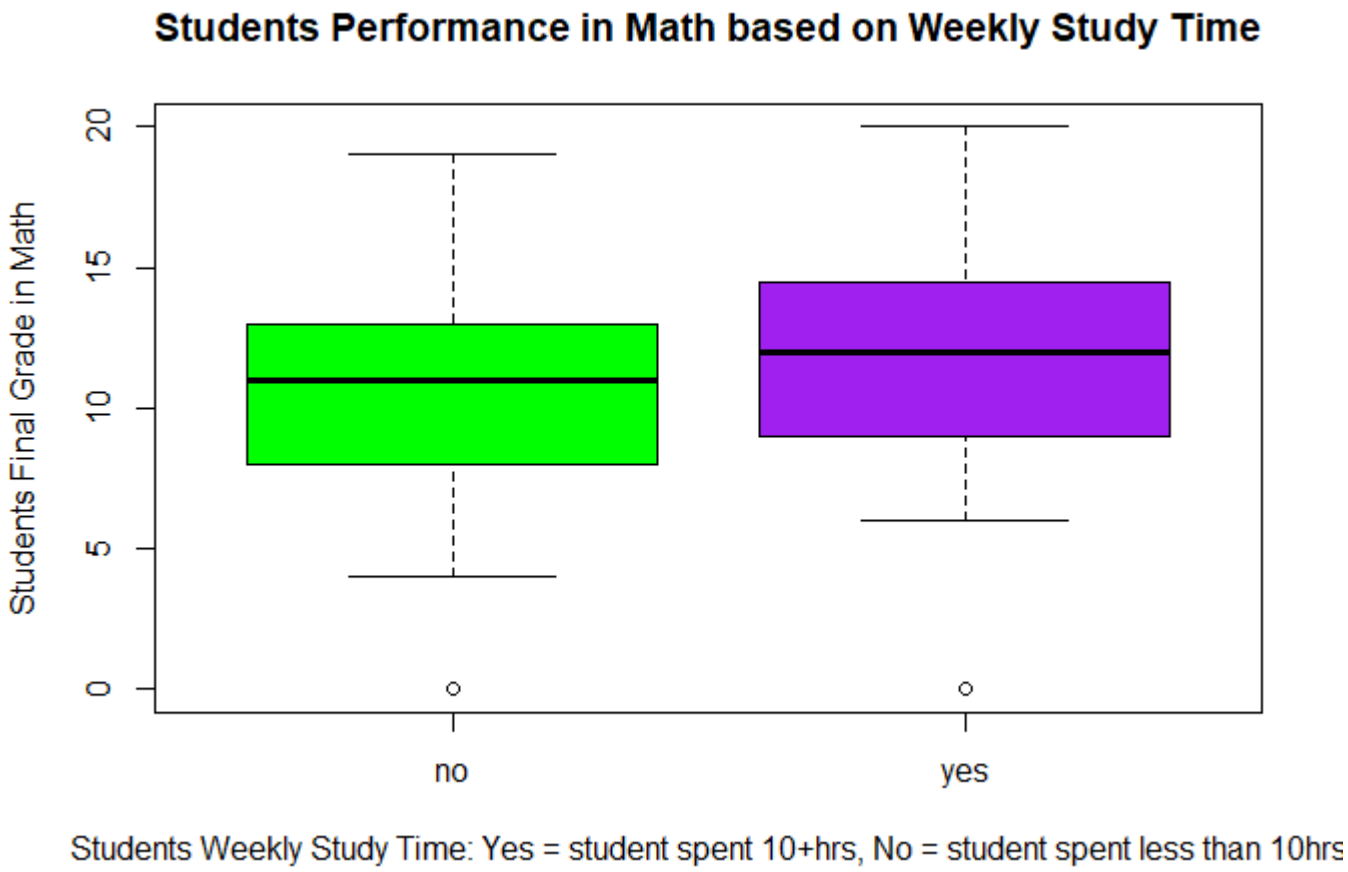
```
## student_portuguese_GP$grade3
##        n  missing distinct
##      423        0        5
##
## lowest : A B C D F, highest: A B C D F
##
## Value          A     B     C     D     F
## Frequency     10   136   245    27     5
## Proportion 0.024 0.322 0.579 0.064 0.012
```

```
## 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
```



Conduct a hypothesis test evaluating 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
```




```
## Let's see the final grade ration between students who study 10+ a week and those who don't
```

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

studyTime10 <chr>	meanFinal_grade <dbl>

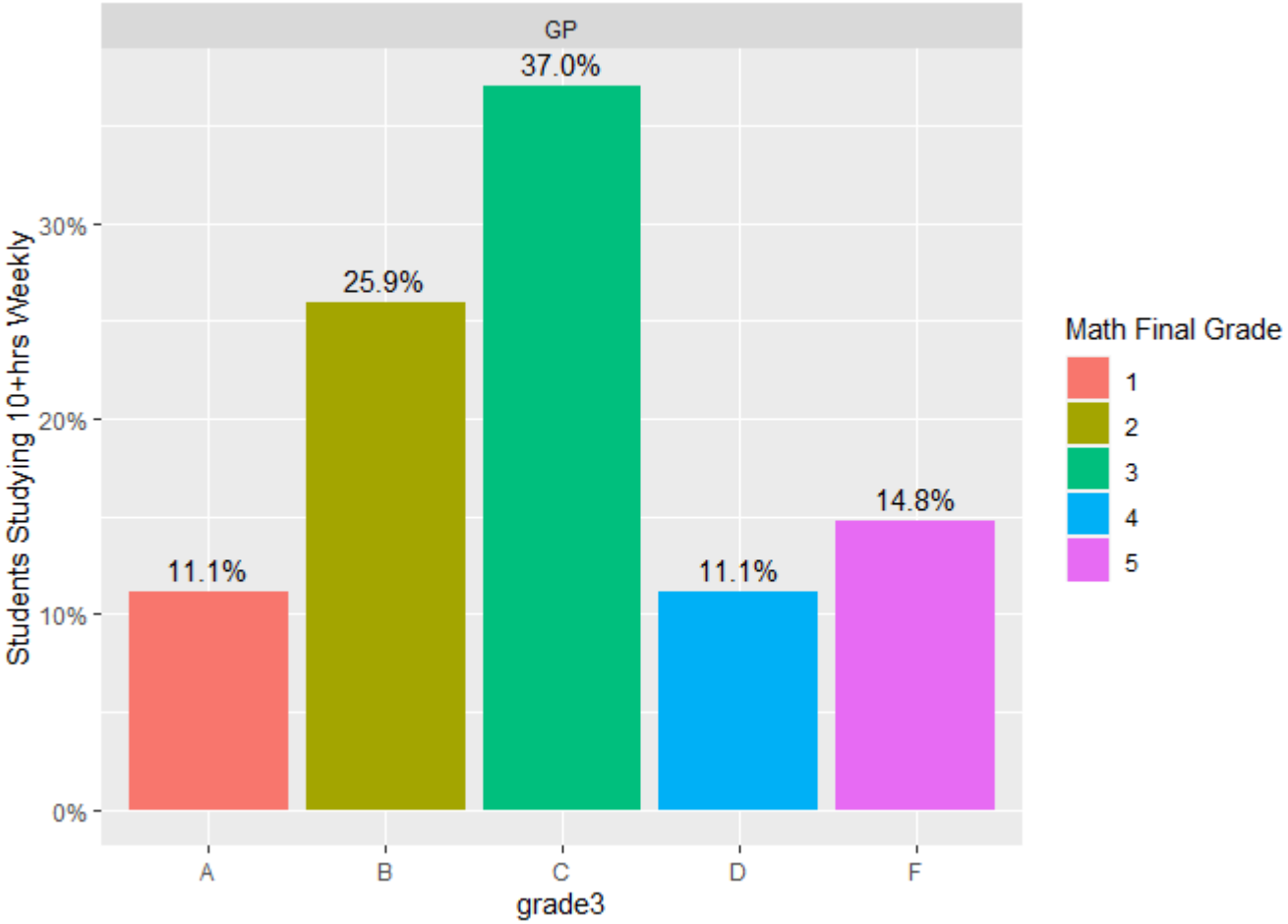
studyTime10 <chr>	meanFinal_grade <dbl>
no	10.35326
yes	11.25926

2 rows

```
## Let's see the statical information about students final grade in Math based on 10+hrs weekl
```

```
## 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 stu
```

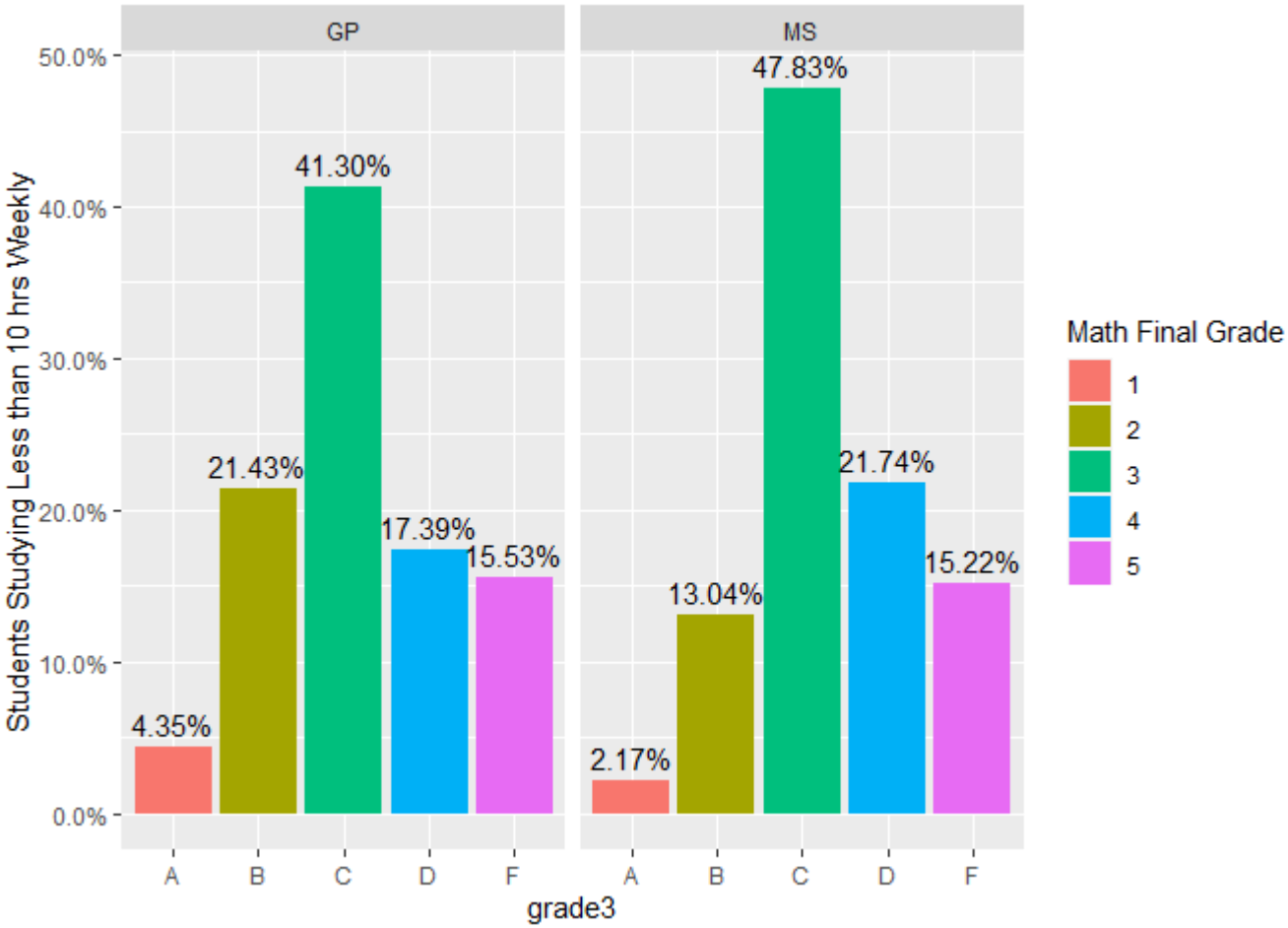


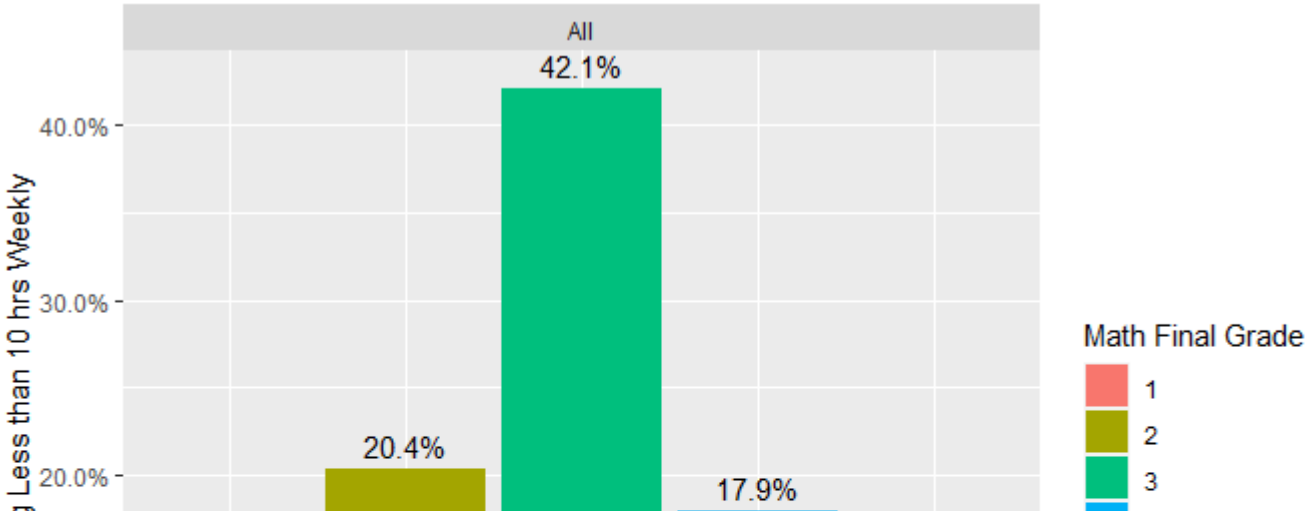

```
## Let's see the statical information about students final grade in Math based on less than 10
```

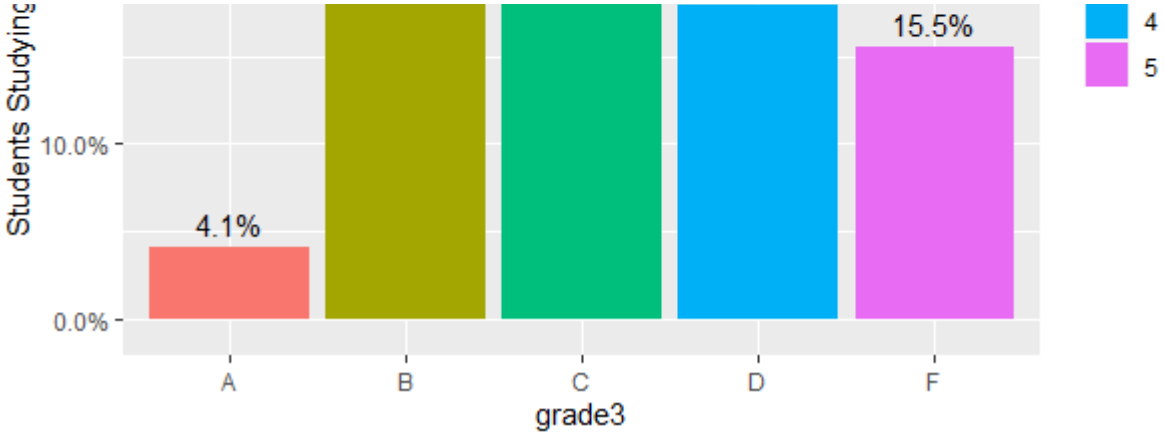
```
## study10Less$grade3
##        n  missing distinct
##      368        0        5
##
## lowest : A B C D F, highest: A B C D F
##
```

```
## 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 stu
```


GP
MS
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
Students Studying Less than 10 hrs Weekly
4.35%
21.43%
41.30%
17.39%
15.53%
2.17%
13.04%
47.83%
21.74%
15.22%
A
B
C
D
F
grade3
Math Final Grade
1
2
3
4
5


All
42.1%
40.0%
30.0%
20.0%
Less than 10 hrs Weekly
20.4%
17.9%
Math Final Grade
1
2
3




```
## [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.

sentiment analysis

```
## [1] "5"  "5"  "7"  "15" "6"  "15"
```

```
##  chr [1:395] "5" "5" "7" "15" "6" "15" "12" "6" "16" "14" "10" "10" "14" ...
```

5. 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 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 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.

6. 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)

References

https://fall2020.data606.net/assignments/labs/

file:///C:/Users/Petit%20Mandela/Documents/R/DATA606_Lab7/DATA606_Lab7/DATA606_Lab7.html

https://www.statisticshowto.com/least-squares-regression-line/

https://rcompanion.org/handbook/C_04.html

https://data-flair.training/blogs/t-tests-in-r/

https://rstatisticsblog.com/data-science-in-action/data-preprocessing/hypothesis-testing-in-r-with-examples-interpretations/