VIETNAM NATIONAL UNIVERSITY HO CHI MINH CITY

HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY

PROBABILITY AND STATISTICS

TEAMWORK PROJECT

CLASS CC05 --- HK 211

INSTRUCTOR: Nguyễn Tiến Dũng

Name	ID number	Score
Trần Nguyễn Minh Khôi	2052551	
Trương Nhã Chi	1852279	
Trần Thị Cẩm Giang	1952662	
Trương Thế Hào	2052973	

Ho Chi Minh City – 2021

CONTENT

1. Data types

- After having an overview of data types, we realized that there are some variables with inappropriate data types, so we have to change.

- Some columns here should be in the form of categorical values but in the dataset it is in character so we use the as.factor command to convert them to categorical values.

- After downloading Rstudio, my team downloaded the Tidyverse library. This is a very versatile library and helps a lot in graphing and data cleaning. It has many more convenient functions such as checking datasets... In short, it's kind of an upgrade.

```
# convert datatypes


grade_dataset <- grade_dataset %>%
  mutate(school = as.factor(school))
grade_dataset <- grade_dataset %>%
  mutate(sex = as.factor(sex))
grade_dataset <- grade_dataset %>%
  mutate(Mjob = as.factor(Mjob))
grade_dataset <- grade_dataset %>%
  mutate(Fjob = as.factor(Fjob))
grade_dataset <- grade_dataset %>%
  mutate(reason = as.factor(reason))
grade_dataset <- grade_dataset %>%
  mutate(guardian = as.factor(guardian))
grade_dataset <- grade_dataset %>%
  mutate(schoolsup = as.factor(schoolsup))
grade_dataset <- grade_dataset %>%
  mutate(famsup = as.factor(famsup))
grade_dataset <- grade_dataset %>%
  mutate(paid = as.factor(paid))
grade_dataset <- grade_dataset %>%
  mutate(activities = as.factor(activities))
grade_dataset <- grade_dataset %>%
  mutate(nursery = as.factor(nursery))
grade_dataset <- grade_dataset %>%
  mutate(higher = as.factor(higher))
grade_dataset <- grade_dataset %>%
  mutate(internet = as.factor(internet))
grade_dataset <- grade_dataset %>%
  mutate(romantic = as.factor(romantic))
```

summary(grade_dataset)
glimpse(grade_dataset)

- This is the result after cleaning is complete:

```
> summary(grade_dataset)
       X1          school   sex          age           address            famsize             Pstatus                Medu            Fedu      
 Min.   :  1.0   GP:349   F:208   Min.   :15.0   Length:395         Length:395         Length:395         Min.   :0.000   Min.   :0.000  
 1st Qu.: 99.5   MS: 46   M:187   1st Qu.:16.0   Class :character   Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000  
 Median :198.0                    Median :17.0   Mode  :character   Mode  :character   Mode  :character   Median :3.000   Median :2.000  
 Mean   :198.0                    Mean   :16.7                                                            Mean   :2.749   Mean   :2.522  
 3rd Qu.:296.5                    3rd Qu.:18.0                                                            3rd Qu.:4.000   3rd Qu.:3.000  
 Max.   :395.0                    Max.   :22.0                                                            Max.   :4.000   Max.   :4.000  

       Mjob           Fjob              reason        guardian     traveltime      studytime     failures schoolsup famsup     paid    
 at_home : 59   at_home : 20   course    :145   father: 90   Min.   :1.000   Min.   :1.000   0:312    no :344   no :153   no :214  
 health  : 34   health  : 18   home      :109   mother:273   1st Qu.:1.000   1st Qu.:1.000   1: 50    yes: 51   yes:242   yes:181  
 other   :141   other   :217   other     : 36   other : 32   Median :1.000   Median :2.000   2: 17                                  
 services:103   services:111   reputation:105                Mean   :1.448   Mean   :2.035   3: 16                                  
 teacher : 58   teacher : 29                                 3rd Qu.:2.000   3rd Qu.:2.000                                          
                                                             Max.   :4.000   Max.   :4.000                                          

 activities nursery   higher    internet  romantic      famrel         freetime         goout            Dalc             Walc      
 no :194    no : 81   no : 20   no : 66   no :263   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000  
                                                    Median :4.000   Median :3.000   Median :3.000   Median :1.000   Median :2.000  
                                                    Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291  
                                                    3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000  
                                                    Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  

     health         absences            G1              G2              G3       
 Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00  
 Median :4.000   Median : 4.000   Median :11.00   Median :11.00   Median :11.00  
 Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.72   Mean   :10.42  
 3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00  
                                                  NA's   :5                      
```

```
> glimpse(grade_dataset)
Rows: 395
Columns: 34
$ X1         <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32~
$ school     <fct> GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, G~
$ sex        <fct> F, F, F, F, F, M, M, F, M, M, F, F, M, M, M, F, F, F, M, M, M, M, M, M, F, F, M, M, M, M, M, M, M, M, M, F, M, M, F, ~
$ age        <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 15, 16, 16, 16, 17, 16, 15, 15, 16, 16, 15, 16, 15, 15, 16, 1~
$ address    <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U~
$ famsize    <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "LE3", "GT3", "GT3", "GT3", "LE3", "GT3", "GT3", "GT3", "GT3"~
$ Pstatus    <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T", "T", "T", "A", "T", "T", "T", "T", "T", "T", "T", "T", "T~
$ Medu       <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, 4, 4, 4, 4, 2, 2, 2, 2, 4, 3, 4, 4, 4, 4, 3, 3, 2, 4, 4, 3, ~
$ Fedu       <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, 3, 3, 4, 2, 2, 4, 2, 2, 2, 4, 4, 4, 4, 3, 3, 2, 3, 3, 4, 4, ~
$ Mjob       <fct> at_home, at_home, at_home, health, other, services, other, other, services, other, teacher, services, health, teacher~
$ Fjob       <fct> teacher, other, other, services, other, other, other, teacher, other, other, health, other, services, other, other, o~
$ reason     <fct> course, course, other, home, home, reputation, home, home, home, home, reputation, reputation, course, course, home, ~
$ guardian   <fct> mother, father, mother, mother, father, mother, mother, mother, mother, mother, mother, father, father, mother, other~
$ traveltime <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2, 1, ~
$ studytime  <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, 1, 2, 1, 2, 2, 3, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 1, 3, 3, 3, ~
$ failures   <fct> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ schoolsup  <fct> yes, no, yes, no, no, no, no, yes, no, no, no, no, no, no, no, no, no, yes, no, no, no, no, no, no, yes, no, no, no, ~
$ famsup     <fct> no, yes, no, yes, yes, yes, no, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, no, no, yes, no, yes, yes~
$ paid       <fct> no, no, yes, yes, yes, yes, no, no, yes, yes, yes, no, yes, yes, no, no, yes, no, no, yes, no, yes, no, no, yes, yes,~
$ activities <fct> no, no, no, yes, no, yes, no, no, no, yes, no, yes, yes, no, no, no, yes, yes, yes, yes, no, no, yes, yes, yes, no, n~
$ nursery    <fct> yes, no, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes~
$ higher     <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, ye~
$ internet   <fct> no, yes, yes, yes, no, yes, yes, no, yes, yes, yes, yes, yes, yes, yes, yes, yes, no, yes, yes, yes, yes, yes, yes, y~
$ romantic   <fct> no, no, no, yes, no, no, no, no, no, no, no, no, no, no, yes, no, no, no, no, no, no, no, no, no, no, no, no, no, no,~
$ famrel     <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, 3, 4, 5, 4, 5, 4, 1, 4, 2, 5, 4, 5, 4, 4, 5, 5, 3, 5, 2, 4, ~
$ freetime   <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, 1, 4, 4, 5, 4, 3, 2, 2, 2, 3, 4, 4, 3, 5, 3, 4, 5, 4, 4, 3, ~
$ goout      <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, 3, 1, 2, 1, 4, 2, 2, 2, 4, 3, 5, 2, 1, 2, 2, 3, 1, 3, 3, 2, ~
$ Dalc       <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 5, 3, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ Walc       <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, 3, 1, 1, 3, 4, 1, 3, 2, 4, 1, 5, 4, 1, 1, 1, 1, 1, 1, 1, 1, ~
$ health     <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, 5, 1, 5, 5, 5, 5, 5, 5, 1, 5, 5, 5, 5, 5, 2, 5, 5, 4, 5, 5, ~
$ absences   <dbl> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, 16, 4, 0, 0, 2, 0, 2, 14, 2, 4, 4, 16, 0, 0, 0, 0, 0, 0, 2, 7~
$ G1         <dbl> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 14, 13, 8, 6, 8, 13, 12, 15, 13, 10, 6, 12, 15, 11, 10, 9, 17,~
$ G2         <dbl> 6, NA, 8, 14, 10, NA, 12, 5, NA, 15, 8, 12, 14, 10, 16, 14, 14, 10, 5, 10, 14, 15, 15, 13, 9, 9, 12, 16, 11, 12, 11, ~
$ G3         <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 14, 14, 10, 5, 10, 15, 15, 16, 12, 8, 8, 11, 15, 11, 11, 12, ~
Warning message:
```

2. Range constraints:

- With this dataset, all values are within acceptable range, especially columns G1, G2, G3. G1 is the first period grade, G2 is the second period grade, and G3 is the final grade. These cells have the lowest value of 0 and the highest value of 20.

3. Calculate some stats

How do factors such as sex, study time, failures, G1, and G2 impact G3? We will figure it out by following these divided sections.

a. Sex

Firstly, the rate of male and female in this dataset is interesting and the command table is used to count the frequency. Since the focus is on sex statistics, we only refer to the sex column and save the frequency into variable sex_stats.

Next, to calculate the rate of male and female, we use the formula:

table(sex_stats)/length(sex_stats)

In which function length is used to count the total cases in sex. We have the following result:

```
> # Calculate some statistics
> sex_stats <- grade_dataset$sex

> table(sex_stats)
sex_stats
  F   M
208 187

> table(sex_stats)/length(sex_stats)
sex_stats
        F         M
0.5265823 0.4734177
```

b. Age

The next factor is age, specifically is mean, median, 1st quartile, 3rd quartile, min and max of the age variable. Moreover, we also calculate standard deviation and variance. Coding is no longer necessary in this part because mean, median, 1st quartile, 3rd quartile, min and max are known in the summary of dataset in the first place.

```
      age
 Min.   :15.0
 1st Qu.:16.0
 Median :17.0
 Mean   :16.7
 3rd Qu.:18.0
 Max.   :22.0
```

```
> age_stats <- grade_dataset$age
> sd(age_stats)
[1] 1.276043
> var(age_stats)
[1] 1.628285
```

c. Study time

The third factor is study time, we will also focus on statistics such as the age variable.

```
   studytime
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.035
 3rd Qu.:2.000
 Max.   :4.000
```

```
> study_time_stats <- grade_dataset$studytime
> sd(study_time_stats)
[1] 0.8392403
> var(study_time_stats)
[1] 0.7043244
```

d. Failures

The fourth one is the failed courses. This factor is considered as a numberical value instead of categorical value because the failures is not used to classify. The statistics are still the focus as same as other factors.

```
    failures
 Min.   :0.0000
 1st Qu.:0.0000
 Median :0.0000
 Mean   :0.3342
 3rd Qu.:0.0000
 Max.   :3.0000
```

```
failure_stats <- grade_dataset$failures
sd(failure_stats)
var(failure_stats)
```

```
> # 4 FAILURES_STATS
> failure_stats <- grade_dataset$failures

> sd(failure_stats)
[1] 0.743651

> var(failure_stats)
[1] 0.5530168
```

e. Absence

The fifth factor is about the absence of students. The statistics are:

```
    absences
 Min.   : 0.000
 1st Qu.: 0.000
 Median : 4.000
 Mean   : 5.709
 3rd Qu.: 8.000
 Max.   :75.000
```

```
> absences_stats <- grade_dataset$absences
> sd(absences_stats)
[1] 8.003096
> var(absences_stats)
[1] 64.04954
```

f. G1

As being mentioned about, G1 is the first period grade of students. Using code R will provide us the following statistics:

```
       G1
 Min.   : 3.00
 1st Qu.: 8.00
 Median :11.00
 Mean   :10.91
 3rd Qu.:13.00
 Max.   :19.00
```

```
> # 5 G1
> G1_stats <- grade_dataset$G1

> sd(G1_stats)
[1] 3.319195

> var(G1_stats)
[1] 11.01705
```

g. G2

G2 includes "NOT AVAILABLE" statistics could be because the students did not attend the exam so the grade was not recorded. Therefore, we need an extra step to filter out the NA value. There is one way to automatically remove them which is using the argument na.m = TRUE.

```
       G2
 Min.   : 0.00
 1st Qu.: 9.00
 Median :11.00
 Mean   :10.72
 3rd Qu.:13.00
 Max.   :19.00
 NA's   :5
```

```
> # 6 G2
> G2_stats <- grade_dataset$G2

> sd(G2_stats, na.rm = TRUE)
[1] 3.737868

> var(G2_stats, na.rm = TRUE)
[1] 13.97166
```

h. G3

G3 has some NA values as well so the filtering step is also needed in this part.

```
       G3
Min.   : 0.00
1st Qu.: 8.00
Median :11.00
Mean   :10.42
3rd Qu.:14.00
Max.   :20.00
```

```
> # 7 G3
> G3_stats <- grade_dataset$G3

> sd(G3_stats, na.rm = TRUE)
[1] 4.581443

> var(G3_stats, na.rm = TRUE)
[1] 20.98962
```

This is the statistics table of the included variables in our dataset:

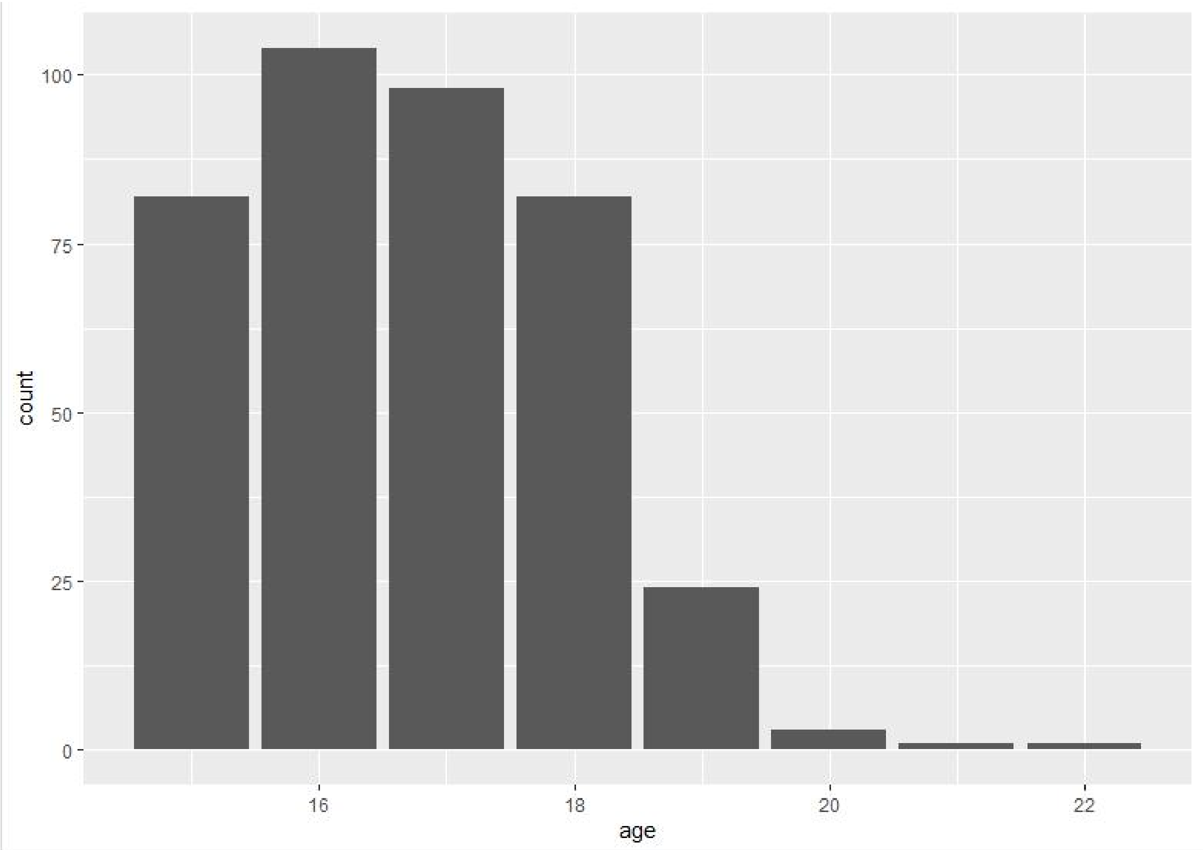
Variable	Min	1st Q	Median	Mean	3rd Q	Max	SD	Variance
Age	15	16	17	16	18	22	1.2760	1.6283
Study time	1	1	2	2.035	2	4	0.8392	0.7043
Failures	0	0	0	0.3342	0	3	0.7437	0.5530
Absences	0	0	4	5.709	8	75	8.0031	64.0495
G1	3	8	11	10.91	13	19	3.3192	11.0171
G2	0	9	11	10.72	13	19	3.7379	13.9717
G3	0	8	11	10.42	14	20	4.5814	20.9896

4. Graph

a. Age

- We are interested in the age variable, which is a graph of the age distribution of students.
- The age distribution chart from 15 to 22 years old.
- A bar chart would be appropriate to show that.

```
bar_age <- grade_dataset %>%
ggplot( aes(age)) +
  geom_bar(bins=30)
bar_age
```



- Therefore, based on this diagram, we can conclude that the majority of students' ages fall between 15 and 18 years old, most specifically at 16 and 17 years old, and 15 and 18 years old follow.

- The remaining people from 19 to 22 years old are much less than the above group. Therefore, the following conclusions can be drawn:

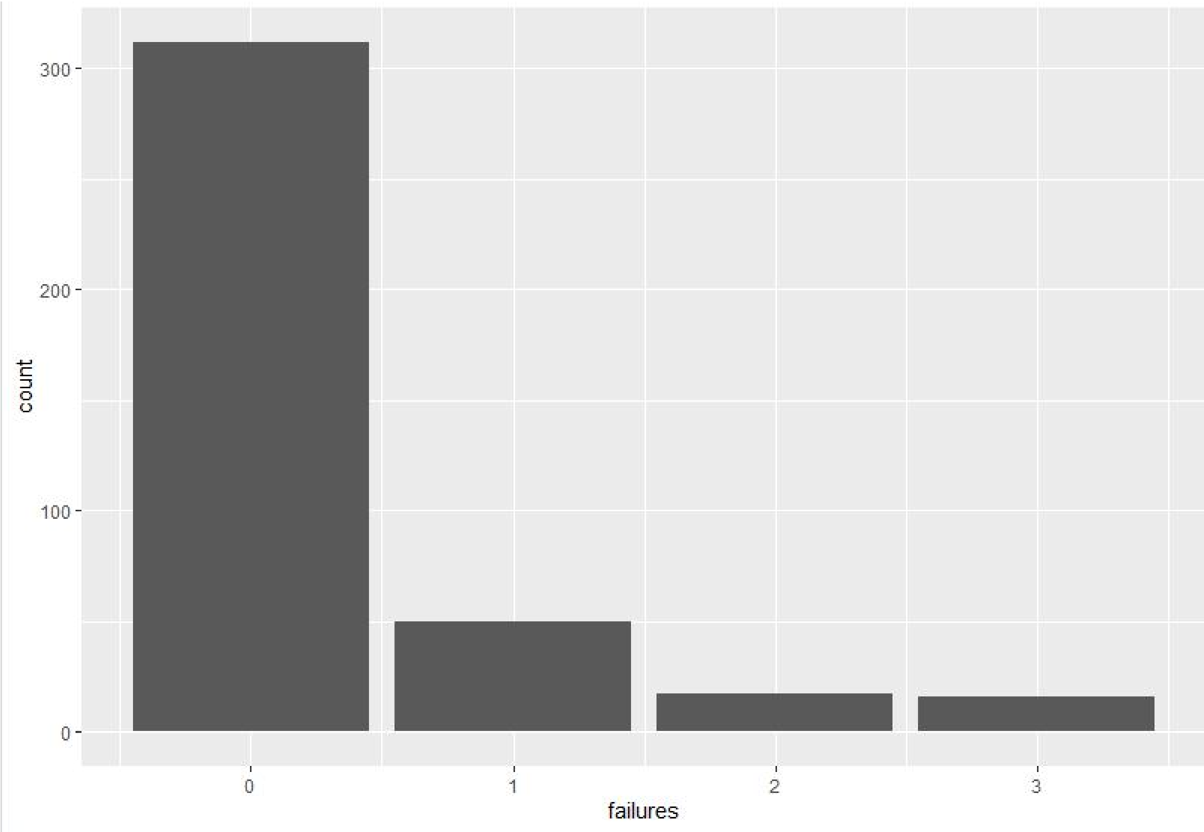
=> Result

This can only be high school. Because only high school has the age distribution from 15 to 18 as shown in the chart above.

b. Failures

- Besides, the failure rate of students is also quite important.

- The following chart shows the failure rate of students by the number of subjects including no subjects, 1 subject, 2 subjects and 3 subjects.



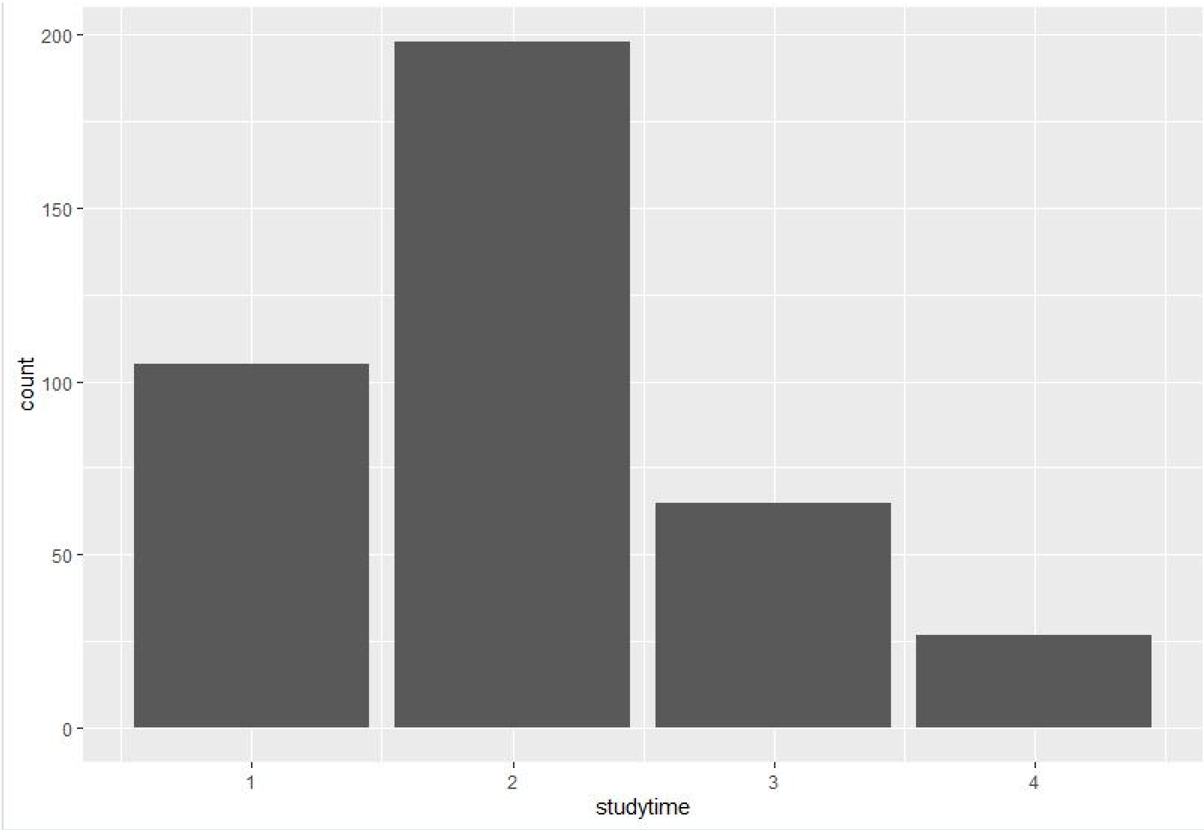
→ With the chart above we can see, most students pass the subject, the percentage of students who fail a subject accounts for 1/6 of the students who pass the subject and the percentage of students who fail 2 and 3 subjects is negligible.

c. Study time

- With the data table of the distribution of learning time, we see that there is not much expectation because there are only 4 possibilities: study for 1 hour, study for 2 hours, study for 3 hours and finally study for 4 hours.

- With the rate in descending order of 2 hours, 1 hour, 3 hours, 4 hours. With such a small range of values (specifically 4 values), it is not expected that it will be distributed in a normal, geometric or poisson manner.

```
bar_studytime <- grade_dataset %>%
  ggplot( aes(studytime)) +
  geom_bar(bins = 30)
bar_studytime
```

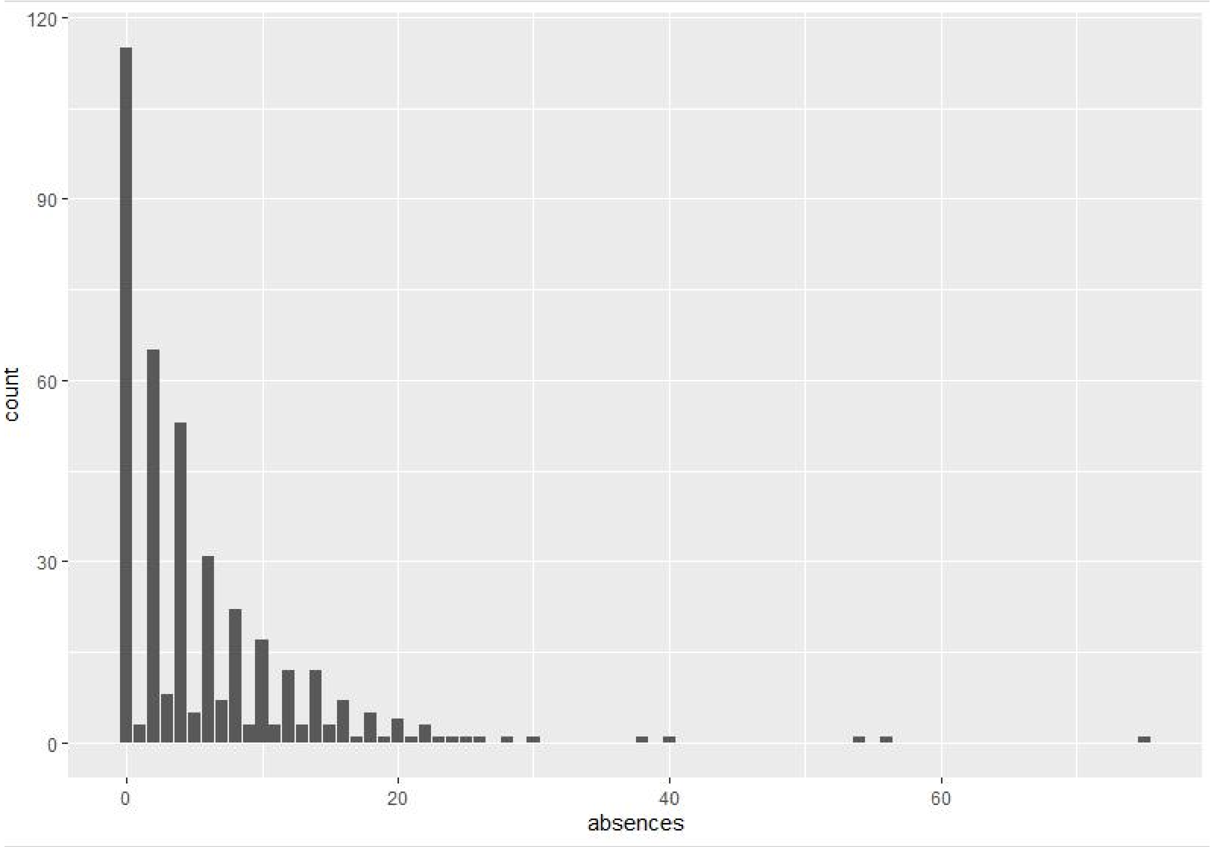


d. Absences

- With the variable absences, all the teachers would not expect it to have a normal distribution, but would expect it to have a geometric distribution (geometric distribution).

- It means most students attend full school without breaks. class or may miss school for a reasonable reason, but not more than 20% of the lessons.

```
bar_absences <- grade_dataset %>%
  ggplot( aes(absences)) +
  geom_bar(bins = 30)
bar_absences
```



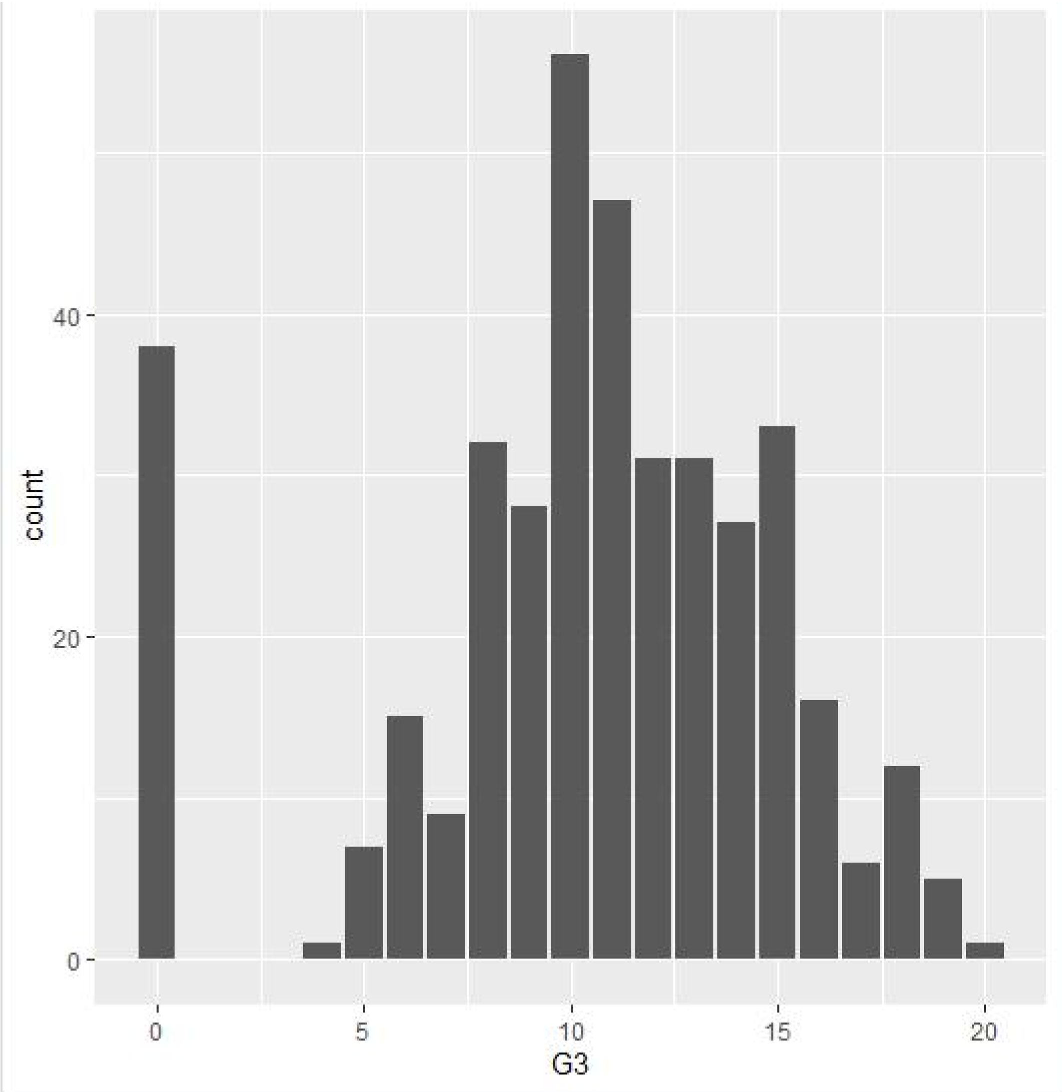
→ Fortunately, our chart is distributed according to the pattern we expect. It is a geometrical distribution instead of the normal distribution.

e. G3

- Grade G3 is the student's final score, the score we rely on to evaluate student learning outcomes. So we will pay more attention to this point than point G1, G2.

- We will be interested in the distribution graph of G3. Here, we temporarily assume G3 points will be distributed in a normal distribution, so we will draw a bar graph.

```
bar_G3 <- grade_dataset %>%
  ggplot( aes(G3)) +
  geom_bar(bins=30)
bar_G3
```



→ Because the dataset doesn't follow our assumption. That is according to the normal distribution.

- To show more information for viewers to understand more about this dataset, we will classify G3 scores as follows:

+ From 0 to 9: Fail
+ From 10 to 20: Pass

Thus, we need to do one more step to turn these two criteria into factor data for easy classification and graphing. And that method is called discretize a variable.

Discretize G3 variable

```
grade_dataset <- grade_dataset %>%
  mutate(classified = ifelse(G3 >= 10, "Pass", "Fail"))
```

- This means that in the original data set we create a new variable named classifier.

- In this variable, we pass it to the ifelse function, whoever has a G3 score greater than or equal to 10, specifically from 10 to 20 will pass, the rest of the people with G3 score from 0 to 10 will fail.

- Then, since this is a factor variable, we have to change the classified variable from character to factor and use the summary function.

```
grade_dataset <- grade_dataset %>%
  mutate(classified = as.factor(classified))
summary(grade_dataset)
```

```
classified
Fail:130
Pass:265
```

→ Thus,, in 395 students, we will have the rate of students who fail and pass the subject as follows:

```
classify_stats <- grade_dataset$classified
table(classify_stats)
table(classify_stats)/length(classify_stats)
```

```
> table(classify_stats)/length(classify_stats)
classify_stats
     Fail      Pass
0.3291139 0.6708861
```

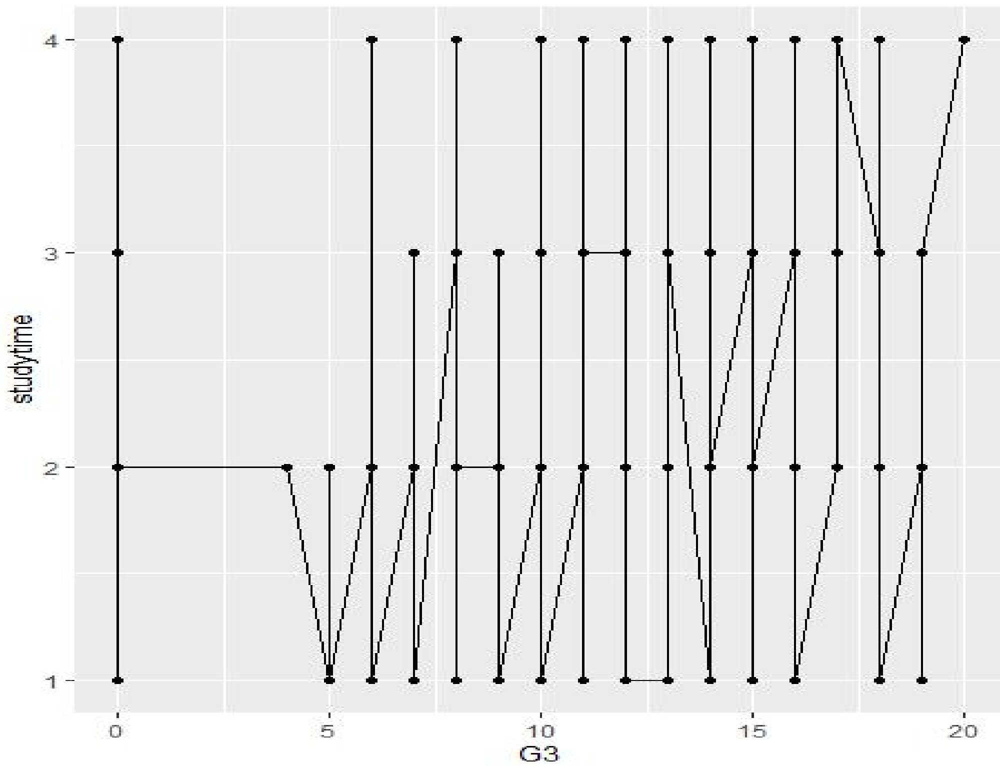
5. Linear regression

The main point is to explore the factors that have a strong influence on the G3 (which is the final score column). The work needed to be done is examining the factors Study time, Sex and Age, respectively. The Study time factor is mainly focused on because the popular opinion is that the harder you study the higher the score is.

a. Study time

As already being mentioned, most people's point of view is that study time has a direct effect on the result.

```
plot_G3_studytime <- grade_dataset %>%
  ggplot(aes(x=G3, y=studytime)) +
  geom_line()+
  geom_point()
plot_G3_studytime
```



Nonetheless, based on the survey, people who spend a lot of time studying somehow may not ensure a good grade in G3. Therefore we start filtering specific situations and surveys because there could be more factors impacting on G3 except for study time.

First of all, we filter out the situation of not being absent.

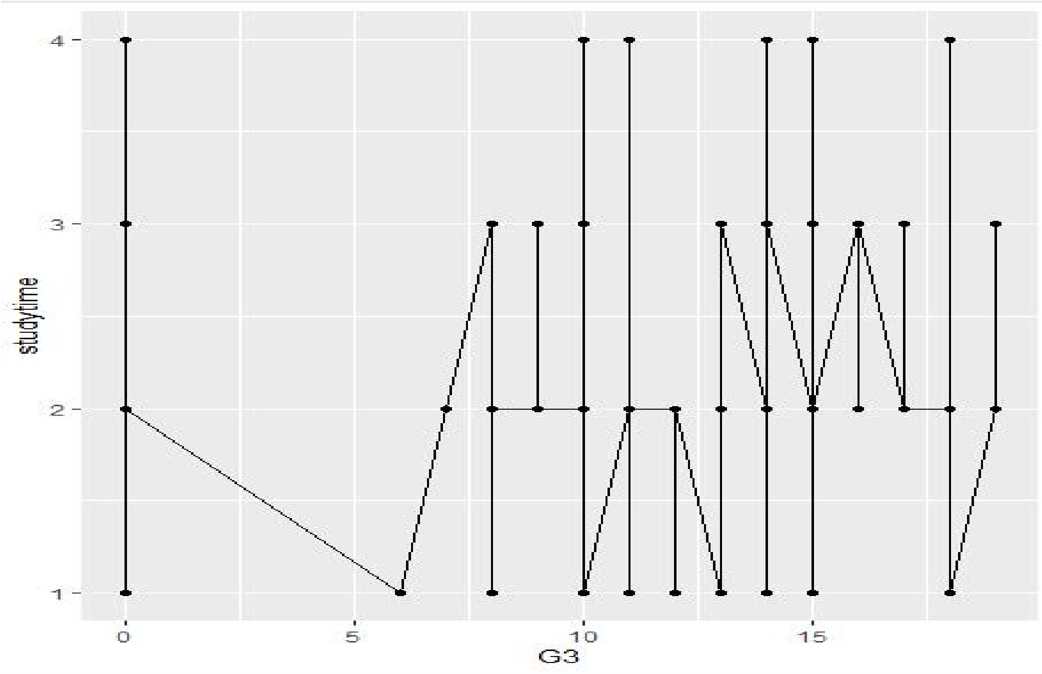
```
not_absences <- grade_dataset %>%
  filter(absences == 0)
```

traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
1	2	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	NA	19
1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15
1	2	0	no	yes	yes	no	yes	yes	yes	no	3	3	3	1	2	2	0	10	8	9
1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	16	16
1	2	0	no	no	no	no	yes	yes	yes	no	4	4	1	1	1	1	0	13	14	15
1	1	0	no	yes	yes	no	yes	yes	yes	no	5	4	2	1	1	5	0	12	15	15
2	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	4	2	4	5	0	13	13	12
1	2	0	no	yes	yes	no	no	yes	yes	no	5	4	2	3	4	5	0	9	11	12
2	2	0	no	yes	no	yes	yes	yes	yes	no	4	3	1	1	1	5	0	17	16	17
1	2	0	no	yes	no	yes	yes	yes	yes	yes	4	5	2	1	1	5	0	17	16	16
1	2	0	no	no	no	yes	no	yes	yes	no	5	3	2	1	1	2	0	8	10	12
1	1	0	no	yes	yes	no	no	yes	yes	no	5	4	3	1	1	5	0	12	14	15
2	1	0	no	yes	no	yes	yes	yes	no	no	3	5	1	1	1	5	0	8	7	6
1	1	0	yes	yes	no	no	yes	yes	yes	no	5	4	1	1	1	1	0	8	8	11
1	1	0	yes	yes	yes	no	yes	yes	yes	no	3	3	4	2	3	5	0	8	10	11
1	2	0	no	yes	yes	yes	yes	yes	yes	no	4	3	2	1	1	1	0	14	15	15
1	2	0	yes	no	no	yes	yes	yes	yes	yes	4	4	4	2	4	2	0	10	10	10
2	4	0	no	yes	yes	no	yes	yes	yes	no	4	3	2	1	1	5	0	13	15	15
1	4	0	no	no	no	no	yes	yes	yes	no	3	3	3	1	1	3	0	10	10	10

Showing 1 to 20 of 115 entries, 34 total columns

It is clear in the statistic table that the number decreased from 395 cases to only 115 cases after filtering. The next step is drawing a graph based on the filtered statistic table above.

```
plot_G3_with_not_absences <- not_absences %>%
  ggplot(aes(x=G3, y=studytime)) +
  geom_line()+
  geom_point()
plot_G3_with_not_absences
```



Hence, the possibility of passing courses of students is not assured by the factor study time with G3 and absences. However, there is another hypothesis to prove that spending a lot of time on studying may not ensure the efficiency of it. The hypothesis is that those who do not fail any course could be a factor of studying effectively. Let's filter this situation to get the specific numbers.

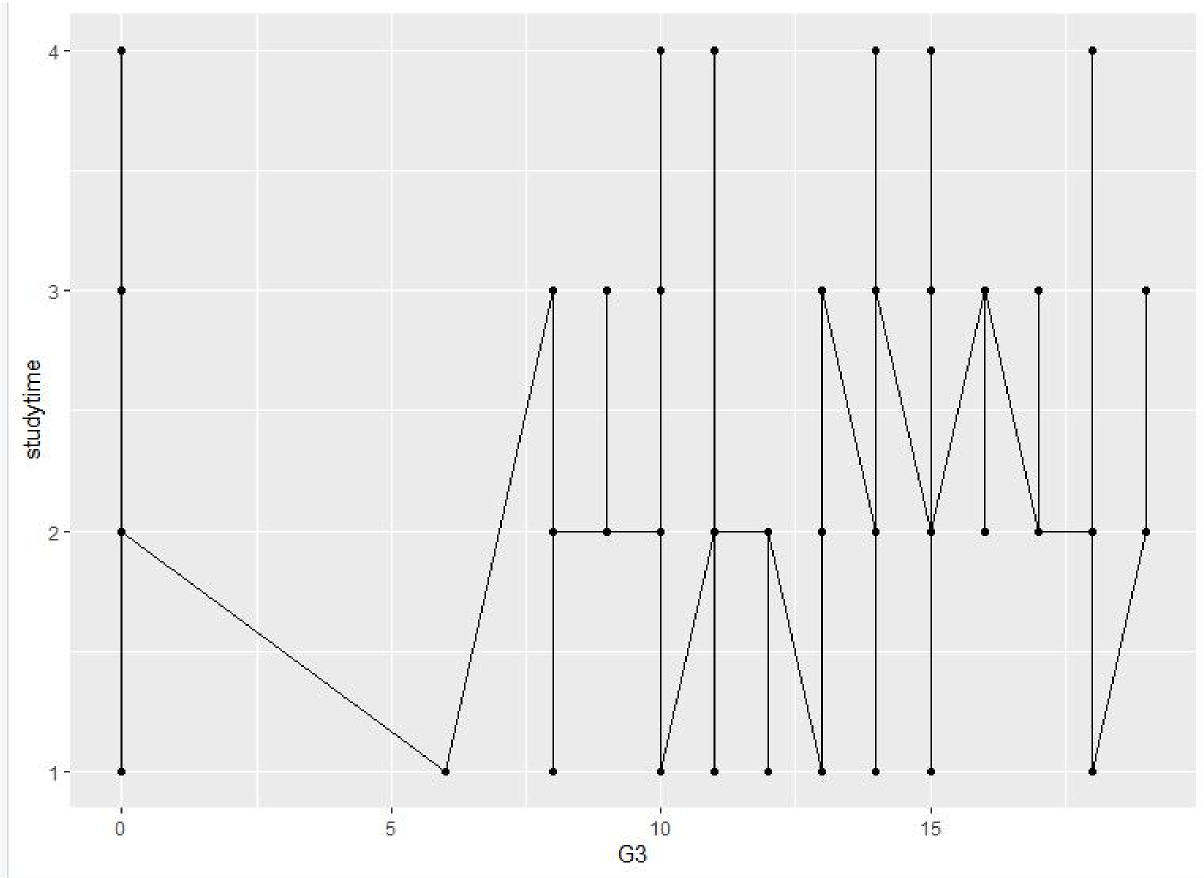
```
not_failures_in_absences <- not_absences %>%
  filter(failures == 0)
```

	X1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup
1	7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no
2	9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no
3	10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no
4	11	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no
5	15	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no
6	21	GP	M	15	U	GT3	T	4	3	teacher	other	reputation	mother	1	2	0	no
7	22	GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	0	no
8	24	GP	M	16	U	LE3	T	2	2	other	other	reputation	mother	2	2	0	no
9	31	GP	M	15	U	GT3	T	4	4	health	services	home	mother	1	2	0	no
10	32	GP	M	15	U	GT3	T	4	4	services	services	reputation	mother	2	2	0	no
11	33	GP	M	15	R	GT3	T	4	3	teacher	at_home	course	mother	1	2	0	no
12	34	GP	M	15	U	LE3	T	3	3	other	other	course	mother	1	2	0	no
13	35	GP	M	16	U	GT3	T	3	2	other	other	home	mother	1	1	0	no
14	36	GP	F	15	U	GT3	T	2	3	other	other	other	father	2	1	0	no
15	44	GP	M	15	U	GT3	T	2	2	services	services	course	father	1	1	0	yes
16	54	GP	F	15	U	GT3	T	4	4	services	services	course	mother	1	1	0	yes

Showing 1 to 16 of 89 entries, 34 total columns

It can be seen that the number of cases dropped from 115 to 89 after filtering and the plotting graph step will be taken based on those figures.

```
plot_G3_without_failures_in_absences <- not_failures_in_absences %>%
  ggplot(aes(x=G3, y=studytime)) +
  geom_line() +
  geom_point()
plot_G3_without_failures_in_absences
```

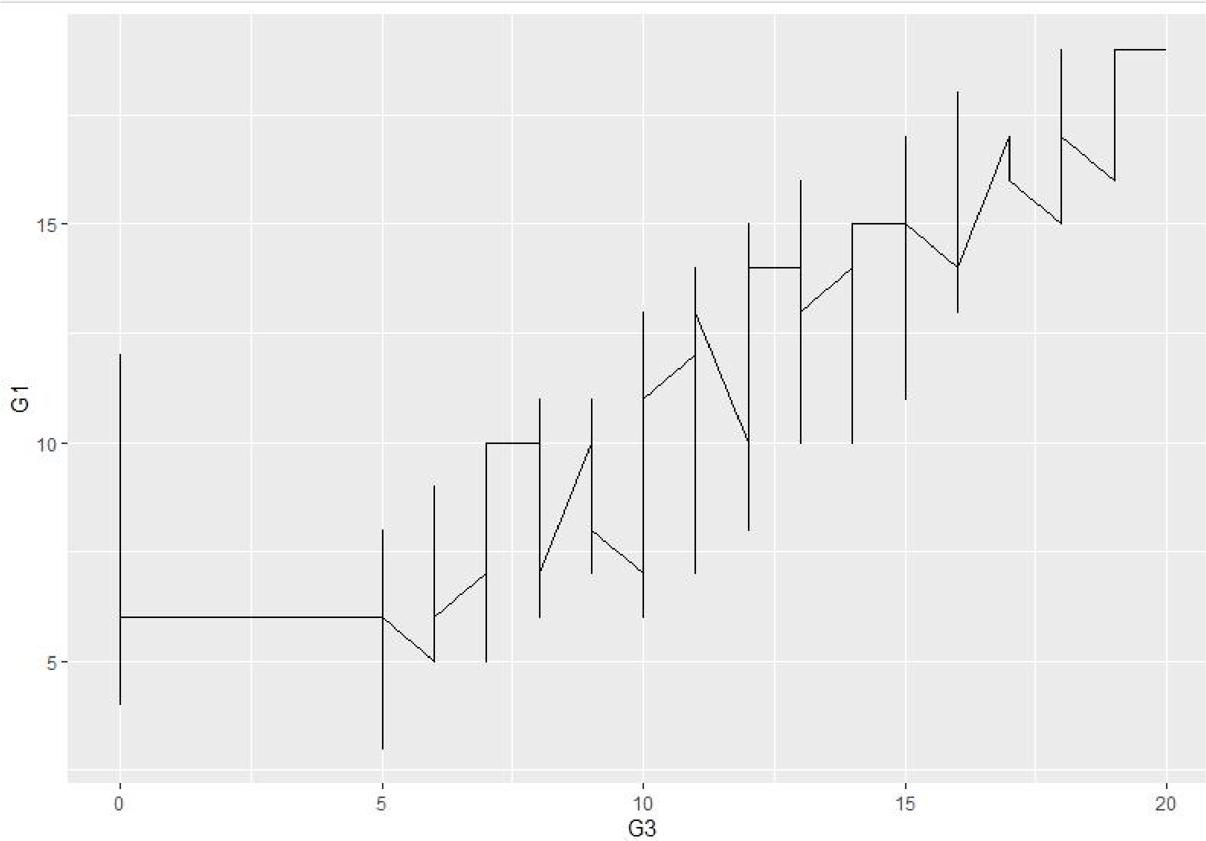


We can go to the conclusion that from this dataset after filtering twice with related factors such as absences and failures, study time is still not a factor to completely assure the possibility of passing courses of students.

b. G1, G2

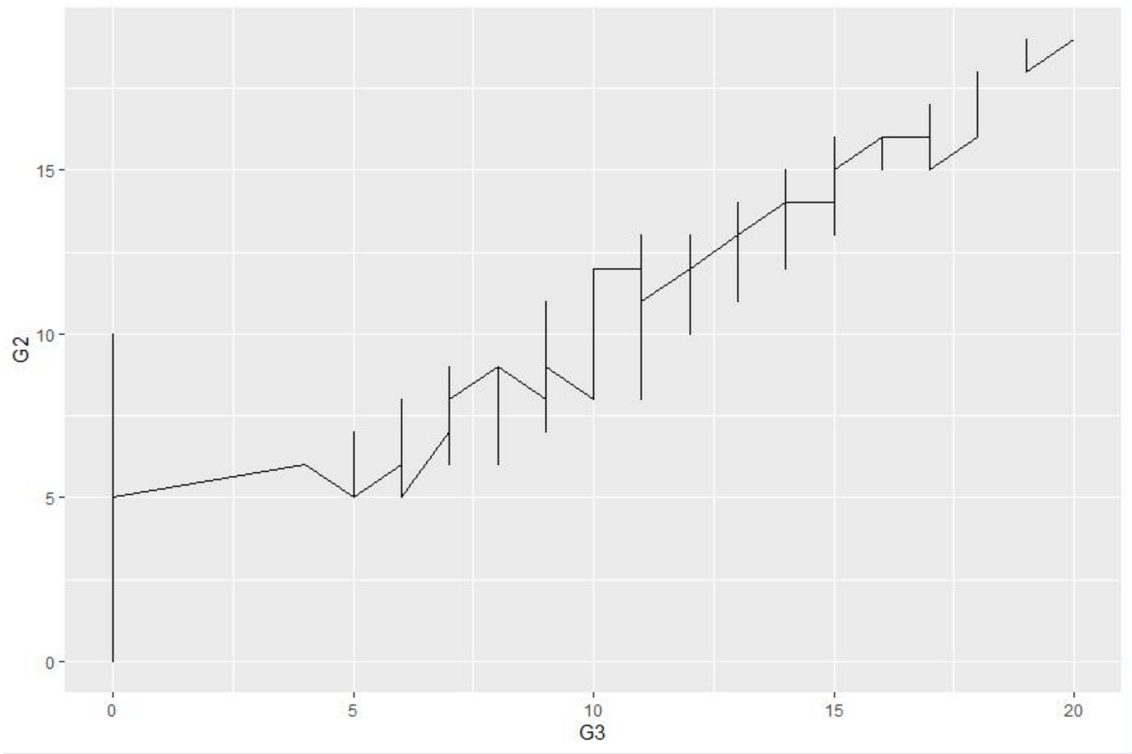
- However, when we perform a graph to show the relationship between G1 and G3, G2 and G3, there seems to be correlations between these two pairs of variables.

```
plot_g1_with_g3 <- grade_dataset %>%
  ggplot(aes(x=G3, y=G1)) +
  geom_line()
plot_g1_with_g3
```



- Through this graph, it is easy to see that G1 and G3 can be correlated with each other. The same thing happens between G2 and G3.

```
plot_g2_with_g3 <- grade_dataset %>%
  ggplot(aes(x=G3, y=G2)) +
  geom_line()
plot_g2_with_g3
```



- However, this graph is not enough to prove that G1, G2 can affect G3. That's why we need to use linear regression to check that.

c. Linear regression

• **Model 1**

- To be able to use linear regression, we need to have the following elements:

+ The first is to have a dataset.

+ The second is that one variable is influenced by the other variables.

+ The third is the variables that affect the variable we are interested in.

- We will use the function lm to test linear regression with the simultaneous effects of the following variables:

+ Sex

+ Age

+ Study time

+ Failures

+ Higher

+ Absences

+ G1

+ G2

```
result1= lm(data = grade_dataset, G3 ~ sex + age + studytime + failures + higher +
absences + G1 + G2)
summary(lresult1)
```

- And we have the following result:

```
Call:
lm(formula = G3 ~ sex + age + studytime + failures + higher +
    absences + G1 + G2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1217 -0.4473  0.3160  0.9743  3.6379 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.61310    1.51569   0.405 0.686068    
sexM         0.19679    0.20836   0.945 0.345511    
age         -0.15235    0.08108  -1.879 0.061000 .  
studytime   -0.13934    0.12477  -1.117 0.264810    
failures    -0.19862    0.14784  -1.344 0.179909    
higheryes    0.26384    0.47490   0.556 0.578836    
absences     0.04208    0.01233   3.413 0.000711 ***
G1           0.16637    0.05696   2.921 0.003701 ** 
G2           0.96039    0.04994  19.231  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.912 on 381 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8285,	Adjusted R-squared:  0.8249 
F-statistic: 230.1 on 8 and 381 DF,  p-value: < 2.2e-16
```

- summary ( ) function is used to produce result summaries of the results of linear model fitting functions.

- The result includes the entire thing and P-value (Pr) of each variable to check that should linear regression model is useful in this case. We also have the P-value (Pr) to test the hypothesis. The linear function formula is y=a+ b1x1 + b2x2 + b3x3 + b4x4 + b5x5 + b6x6 + b7x7 + b8x8. The significant level is chosen for 0.01.

a) Slope-intercept a :

• Null hypothesis : H_0^a : a = 0 (1)

• Alternative hypothesis : H_1^a : a ≠ 0 (2)

Pr (0.686068) > 0.01 (3)

From (1) , (2) , (3)

==> We fail to reject H_0 ==> a = 0

b) Coefficient b_1 (*sex* variable*) :*

• Null hypothesis : $H_0^{b_1}$: $b_1 = 0$ (1)

• Alternative hypothesis : $H_1^{b_1}$: $b_1 \neq 0$ (2)

Pr(0.345511) > 0.01 (3)

From (1) , (2), (3)

=> We fail to reject H_0 => $b_1 = 0$

c) Coefficient b_2 (*age* variable*) :*

• Null hypothesis : $H_0^{b_2}$: $b_2 = 0$ (1)

• Alternative hypothesis : $H_1^{b_2}$: $b_2 \neq 0$ (2)

Pr(0.061000) > 0.01 (3)

From (1), (2) , (3)

=> We fail to reject H_0 => $b_2 = 0$

d) Coefficient b_3 (*Study time* variable) :

• Null hypothesis : $H_0^{b_3}$: $b_3 = 0$ (1)

• Alternative hypothesis : $H_1^{b_3}$: $b_3 \neq 0$ (2)

Pr(0.264810) > 0.01 (3)

From (1), (2) , (3)

=> We fail to reject H_0 => $b_3 = 0$

e) Coefficient b_4 (*failures* variable) :

• Null hypothesis : $H_0^{b_4}$: $b_4 = 0$ (1)

• Alternative hypothesis : $H_1^{b_4}$: $b_4 \neq 0$ (2)

Pr(0.179909) > 0.01 (3)

From (1), (2) , (3)

=> We fail to reject H_0 => $b_4 = 0$

f) Coefficient b_5 (*higher* variable) :

• Null hypothesis : $H_0^{b_5}$: $b_5 = 0$ (1)

• Alternative hypothesis : $H_1^{b_5}$: $b_5 \neq 0$ (2)

Pr(0.578836) > 0.01 (3)

From (1), (2) , (3)

=> We can not reject H_0 => $b_5 = 0$

g) Coefficient b_6 (*absences* variable) :

• Null hypothesis : $H_0^{b_6}$: $b_6 = 0$ (1)

• Alternative hypothesis : $H_1^{b_6}$: $b_6 \neq 0$ (2)

Pr(0.000711) < 0.01 (3)

From (1), (2) , (3)

=> We fail to reject H_0 => $b_6 \neq 0$

h) Coefficient b_7 (G1 variable) :

• Null hypothesis : $H_0^{b_7}$: $b_7 = 0$ (1)

• Alternative hypothesis : $H_1^{b_7}$: $b_7 \neq 0$ (2)

Pr(0.003701) < 0.01 (3)

From (1), (2) , (3)

=> We can reject H_0 => $b_7 \neq 0$

i) Coefficient b_8 (G2 variable):

• Null hypothesis : $H_0^{b_8}$: $b_8 = 0$ (1)

• Alternative hypothesis : $H_1^{b_8}$: $b_8 \neq 0$ (2)

Pr(0.003701) < 0.01 (3)

From (1), (2) , (3)

=> We can reject H_0 => $b_8 \neq 0$

→ **Conclusion :** The input has influence on output, so the linear regression model can be used.

$R^2 = 0.8249$ (nearly 1) => This is the linear equation.

• **Model 2**

- With model 2, we are only interested in the following variables:

  + Absences

  + G1

  + G2

result2= lm(data = grade_dataset, G3 ~ G1 + G2)
summary(result2)

- And we have the following result:

```
> summary(result2)

Call:
lm(formula = G3 ~ absences + G1 + G2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.3471 -0.3582  0.3133  0.9811  3.9465 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.12101    0.34769  -6.100 2.57e-09 ***
absences     0.03660    0.01216   3.011  0.00277 ** 
G1           0.15971    0.05615   2.844  0.00469 ** 
G2           0.98711    0.04943  19.968  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.925 on 386 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.8238,	Adjusted R-squared:  0.8224 
F-statistic: 601.6 on 3 and 386 DF,  p-value: < 2.2e-16
```

- The result includes the entire thing and P-value (Pr) of each variable to check that should linear regression model is useful in this case. We also have the P-value (Pr) to test the hypothesis. The linear function formula is y=a+ b_1 x_1+b_2 x_2+b_3 x_3. The significant level is chosen for 0.01.

a) Slope-intercept a :

• Null hypothesis : H_0^a : a = 0 (1)

• Alternative hypothesis : H_1^a : a ≠ 0 (2)

Pr (2.57e-09) < 0.01 (3)

From (1) , (2) , (3)

==> We can reject H_0 ==> a ≠ 0

b) Coefficient b_1 (*absences* variable) :

• Null hypothesis : $H_0^{b_1}$: $b_1 = 0$ (1)

• Alternative hypothesis : $H_1^{b_1}$: $b_1 \neq 0$ (2)

Pr(0.00277) < 0.01 (3)

From (1) , (2), (3)

=> We can reject H_0 => $b_1 \neq 0$

c) Coefficient b_2 (*G1* variable) :

• Null hypothesis : $H_0^{b_2}$: $b_2 = 0$ (1)

• Alternative hypothesis : $H_1^{b_2}$: $b_2 \neq 0$ (2)

Pr(0.00469) < 0.01 (3)

From (1), (2) , (3)

=> We can reject H_0 => $b_2 \neq 0$

d) Coefficient b_3 (G2 variable) :

• Null hypothesis : $H_0^{b_3}$: $b_3 = 0$ (1)

• Alternative hypothesis : $H_1^{b_3}$: $b_3 \neq 0$ (2)

Pr(<2e - 16) < 0.01 (3)

From (1), (2) , (3)

=> We can reject H_0 => $b_3 \neq 0$

→ **Conclusion:** The input has influence on output, so the linear regression model can be used.

$R^2 = 0.8224$ (nearly 1) => This is the linear equation.

d. Anova

In this part, we will use the anova(result1,result2) function to compare the differences between considering all variables of data df and considering 3 variable included "absences", "G1", "G2".

We have the following results:

```
> anova(result1,result2)
Analysis of Variance Table

Model 1: G3 ~ sex + age + studytime + failures + higher + absences + G1 +
    G2
Model 2: G3 ~ absences + G1 + G2
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    381 1392.4
2    386 1430.7 -5   -38.297 2.0958 0.06521 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The explanation is that the significant level is chosen for **0.01**. We have:

1. Null hypothesis: $H_0: \text{result1} = \text{result2}$
2. Alternative hypothesis: $H_0: \text{result1} \neq \text{result2}$
3. Pr (0.06521) > 0.01

From (1), (2), (3), we come to the conclusion that it fails to reject H_0 so the differences are not significant. The "sex", "age", "studytime", "failures", "higher" variables can be eliminated => Using linear regression model of *result2*. Number of school absences (*absences*), first period grade (*G1*), second period grade (*G2*) also affects final grade (*G3*) significantly.

e. Prediction

After finishing the ANOVA, the *coef (result2)* function is used to find the coefficient $\mathrm{a}, \mathrm{b}_1, \mathrm{b}_2, \mathrm{b}_3$ of linear equation: $G3 = \mathrm{a} + \mathrm{b}_1 \times absences + \mathrm{b}_2 \times G1 + \mathrm{b}_3 \times G2$.

We have:

$\mathrm{a} = -2.1210142$ $\mathrm{b}_1 = 0.0366003$

$\mathrm{b}_2 = 0.1597066$ $\mathrm{b}_3 = 0.9871062$

The linear equation is $G3 = -2.1210142 + 0.0366003 \times absences + 0.1597066 \times G1 + 0.9871062 \times G2$. We can use this equation for prediction.

Then we have the following result:

```
> coef(result2)
(Intercept)    absences          G1          G2
 -2.1210142   0.0366003   0.1597066   0.9871062
```

The next step is using the *confict (result2,level=0.99)* function to compute confidence intervals for coefficients of linear equation. The confidence interval is chosen for **0.99**

As a result, we have the following information:

```
> confint(result2,level=0.99)
                   0.5 %      99.5 %
(Intercept) -3.021050414 -1.22097792
absences     0.005134539  0.06806605
G1           0.014355824  0.30505733
G2           0.859140873  1.11507145
```

The final step is using the function:

```
predict(result2,data.frame(absences=0,G1=10,G2=10), interval='confidence',level=0.99)
```

Breaking it down, we have the explanation for this function:

- *predict ( )*: predict an outcome value on the basis of one or multiple predictor variables.
- *data.frame( )*: create data frame for input data.
- *absences=0, G1=10, G2=10* : the example input data .
- *interval =‘confidence’*: the confidence interval reflects the uncertainty around the mean predictions.
- *level=0.99*: the confidence interval is chosen for 0.99

→ **The result is**:

```
> predict(result2,data.frame(absences=0,G1=10,G2=10), interval='confidence',level=0.99)
       fit      lwr      upr
1 9.347113 9.027274 9.666952
```

In which, *fit* means the predicted final grade for the new values of “absences”, ”G1”, ”G2”; *lwr* and *upr* are the lower and the upper confidence limits for the expected values, respectively.

The 99% confidence interval associated with a final grade is (9.027274, 9.666952). This means that, according to our model, a student with absences=0, G1=10 and G2=10 has, on average, a final grade between 9.027274 and 9.666952