INDIVIDUAL ASSIGNMENT

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

HAND OUT DATE: 5th April 2022

HAND IN DATE: 13th May 2022

STUDENT ID: TP067738

STUDENT NAME: LEE WOON XUN

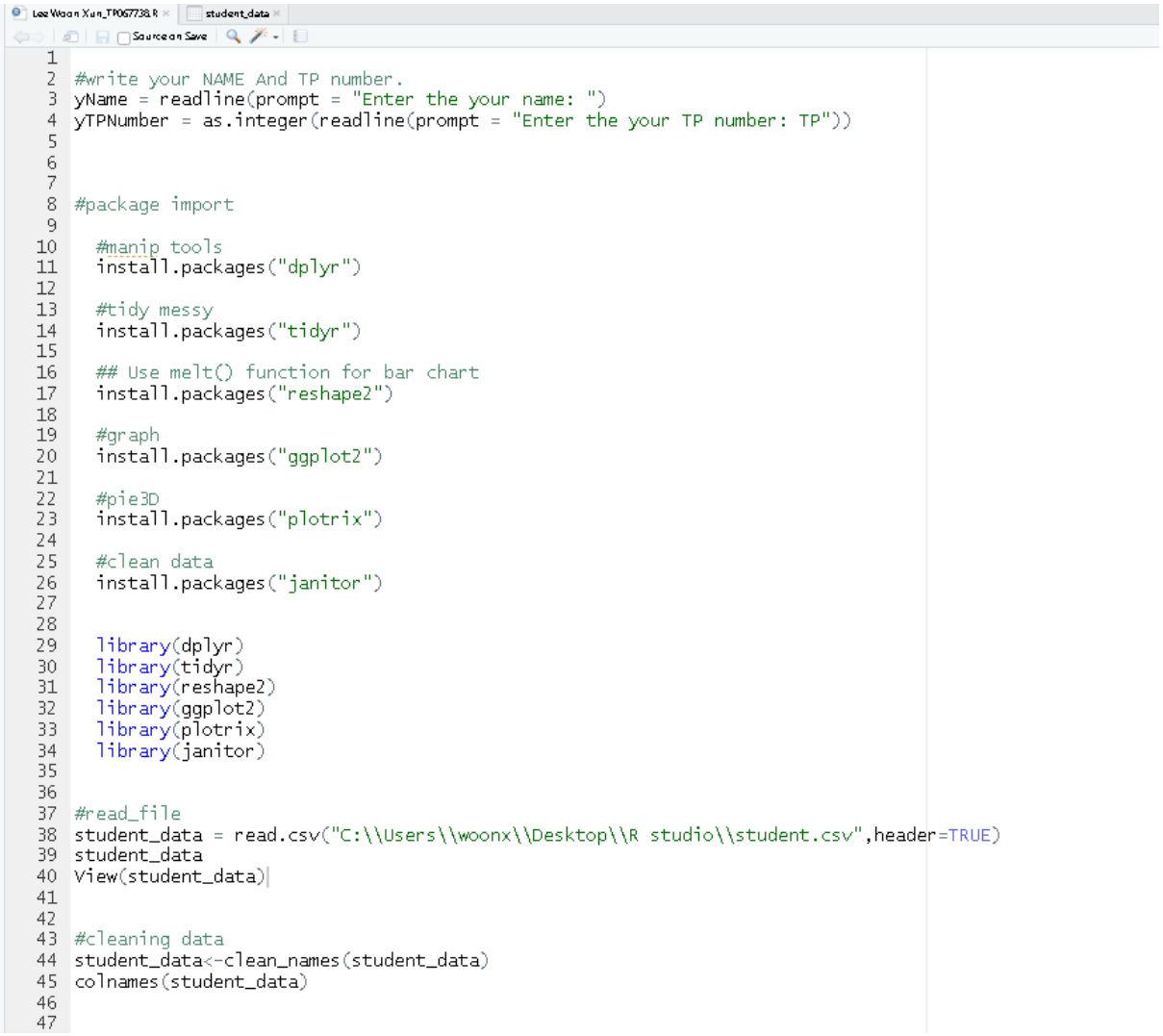
INTAKE: APU2F2202CS(CYB)

Introduction

In this project I will use the techniques and knowledge I have learned about R Language to analyse the data. The research object this time is a three-year final score of degree students marks at the end of their academic programs consisting of 33 columns with 16 integer datatype value, 9 string value, and 8 Boolean value. The data contains 32 columns and 396 rows, where the data contains their family background and their daily routine. I analyse the data and identify what factors affect students' grades. I analyse the data and identify what factors affect students' grades. Not only will I use the knowledge and techniques I've learned, but I'll also use features that allow me to fully present all the analysed data.

Table of Contents

Import Package, Read & Clean Data



```
1
2  #write your NAME And TP number.
3  yName = readline(prompt = "Enter the your name: ")
4  yTPNumber = as.integer(readline(prompt = "Enter the your TP number: TP"))
5
6
7
8  #package import
9
10   #manip tools
11   install.packages("dplyr")
12
13   #tidy messy
14   install.packages("tidyr")
15
16   ## Use melt() function for bar chart
17   install.packages("reshape2")
18
19   #graph
20   install.packages("ggplot2")
21
22   #pie3D
23   install.packages("plotrix")
24
25   #clean data
26   install.packages("janitor")
27
28
29   library(dplyr)
30   library(tidyr)
31   library(reshape2)
32   library(ggplot2)
33   library(plotrix)
34   library(janitor)
35
36
37 #read_file
38 student_data = read.csv("C:\\Users\\woonx\\Desktop\\R studio\\student.csv",header=TRUE)
39 student_data
40 View(student_data)
41
42
43 #cleaning data
44 student_data<-clean_names(student_data)
45 colnames(student_data)
46
47
```

Figure 1.0 Import Package, Read & Clean Data

When I start the analysis, I'll import my package tools (row 10 until 34) and used read function to import my package (rows 38 & 39). After that cleaning the data (rows 44) to change all column name to lower case.

Question 1: Which school's student total score better? Which school most suitable student to choose? Why?

```
49
50 # Question 1: which school's student total score better? which school most suitable student to choose? why?
51 Q1 = filter(student_data, g1&g2&g3 > 10)
```

Figure 2.0 Question 1 Global Filter Code.

Before starting the analysis, I would use this filter to filter all students who are above 10 in grades 1 to 3. Because the total of each grade is 20 score, so above 10 score mean already pass the grade.

Analysis 1-1 Total out how many total passes? (Included how many students in school, age, sex)

```
54
55 School = Q1 %>% group_by(school) %>% summarise(n()) %>% select(school)
56 TotalOfPass = Q1 %>% group_by(school) %>% summarise(TotalOfPass=n()) %>% select(-school)
57 TotalOfStudent = group_by(student_data,school) %>% summarise(TotalOfStudent=n()) %>% select(-school)
58
59
60 Q1A1DF <- data.frame(TotalOfPass, TotalOfStudent,School)
61 Q1A1DF
62
63 Q1A1DFMelt <- melt(Q1A1DF, id='school')
64 head(Q1A1DFMelt)
65
66 ggplot(Q1A1DFMelt, aes(x=school, y=value, fill=variable)) +
67   geom_bar(stat="identity", position="dodge") +
68   geom_text(aes(label=value),position = position_dodge(1),vjust = -0.5)
69 #position_doge(low=left, high=right), vjust= (low=move up, high=move down)
70
```

Figure 2.1.1 Count the two different school are student pass the entire grade.

From figure 2.1.1, I want to calculate the total number of passing grades for the entire grade and the total number of students in each school.

From rows 55 to 57, I just filter and get the data I want to use. I got the school type, the total number of students per school is eligible, and the total number of students per school. But I just want to take the number, so I drop it school name from rows 56 and 57. After when I filter the data, I store the result to data frame. And use melt() function to reshape the data frame.

Because I want display the bar chart with two variable, so I use melt() function to reshape, make my result from 'TotalOfPass' & 'TotalOfStudent' can store together in new a column call 'value'.

```
73
74 #Percentages
75 PGP = Q1A1DF[1,1]/ Q1A1DF[1,2]
76 PercentGP = PGP * 100
77
78 PMS = Q1A1DF[2,1]/ Q1A1DF[2,2]
79 PercentMS = PMS * 100
80
81 PercentagesOfQ1A1 = data.frame(GP= round(PercentGP, digits=2),MS= round(PercentMS,digits = 2))
82 PercentagesOfQ1A1
83
```

Figure 2.1.2 Calculate the percentages of the student pass from each school.

From figure 2.1.2 I extract the index of the data frame to calculate percentages of the student are passing entire grade and division the total student at each school.

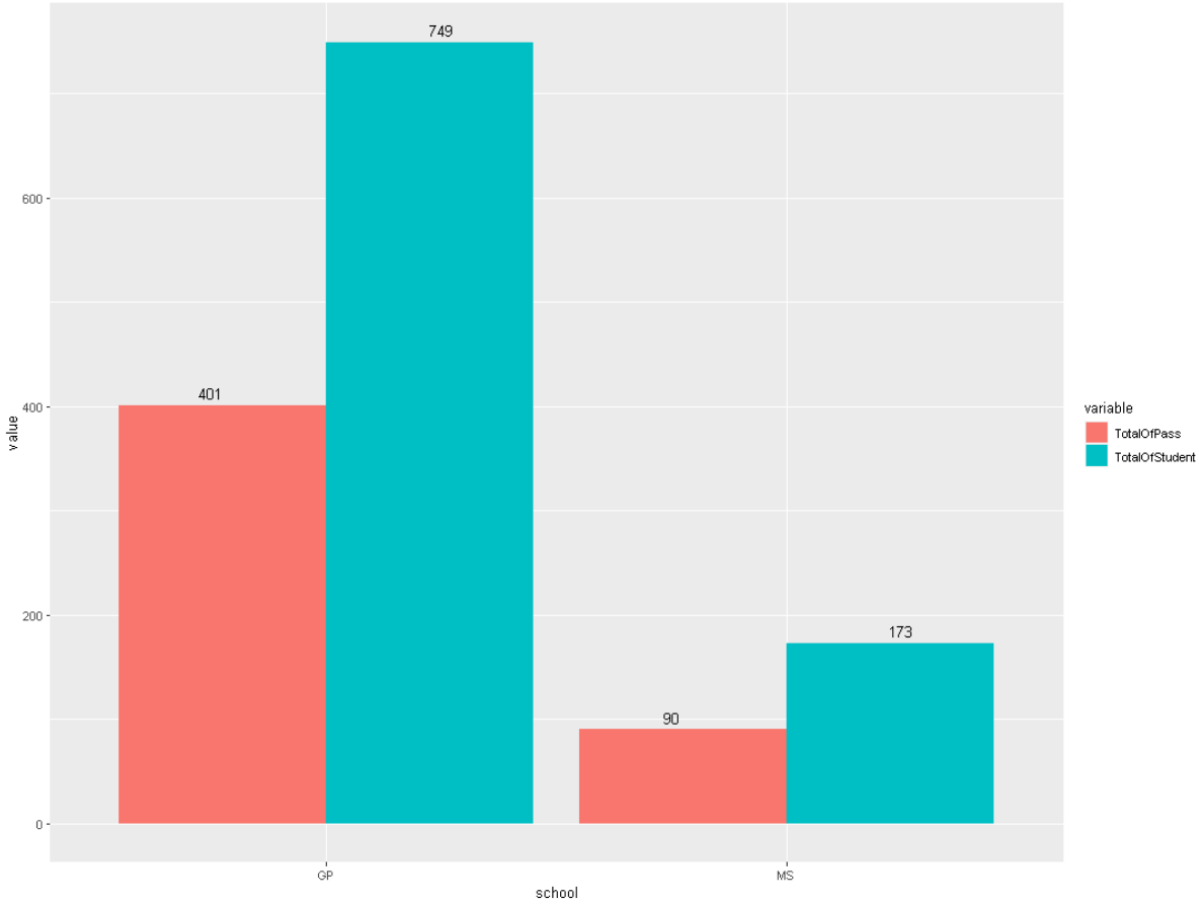
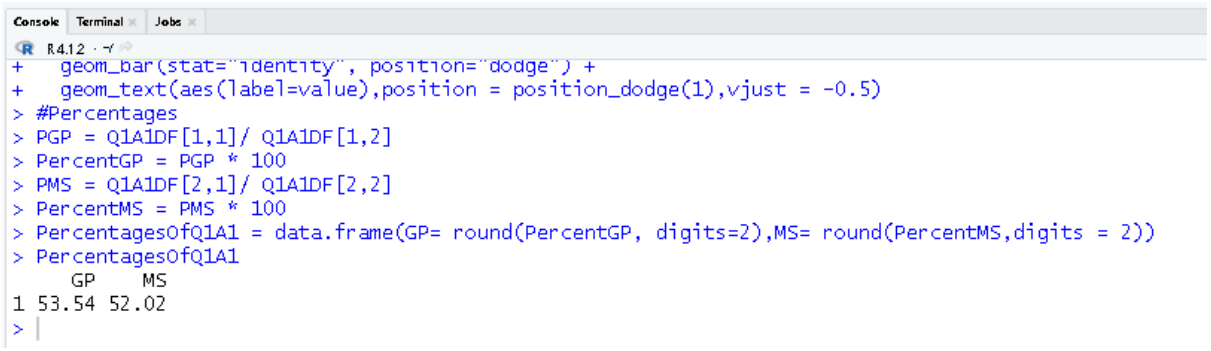


Figure 2.1.3 Bar Chart of present total pass entire grade student & total school of each school

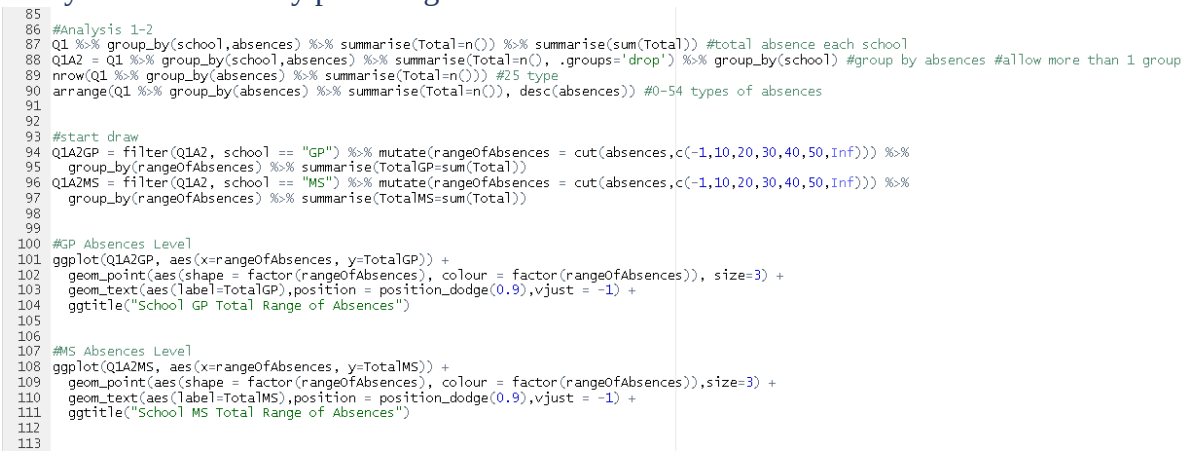


```
Console  Terminal  Jobs
R 4.1.2 · ~/
+   geom_bar(stat="identity", position="dodge") +
+   geom_text(aes(label=value),position = position_dodge(1),vjust = -0.5)
> #Percentages
> PGP = Q1A1DF[1,1]/ Q1A1DF[1,2]
> PercentGP = PGP * 100
> PMS = Q1A1DF[2,1]/ Q1A1DF[2,2]
> PercentMS = PMS * 100
> PercentagesOfQ1A1 = data.frame(GP= round(PercentGP, digits=2),MS= round(PercentMS,digits = 2))
> PercentagesOfQ1A1
     GP    MS
1 53.54 52.02
>
```

Figure 2.1.4 Result of the percentages of student pass from each school.

The pass rate for all grades in GP schools is 1.52% higher than the pass rate for all grades in MS schools through the calculated percentage rate. Although the student union of GP schools is much larger than that of MS schools, it does not affect the pass rate of students in all grades of GP schools.

Analysis 1-2 how many percentage absences from each school?



```
85
86  #Analysis 1-2
87  Q1 %>% group_by(school,absences) %>% summarise(Total=n()) %>% summarise(sum(Total)) #total absence each school
88  Q1A2 = Q1 %>% group_by(school,absences) %>% summarise(Total=n(), .groups='drop') %>% group_by(school) #group by absences #allow more than 1 group
89  nrow(Q1 %>% group_by(absences) %>% summarise(Total=n())) #25 type
90  arrange(Q1 %>% group_by(absences) %>% summarise(Total=n()), desc(absences)) #0-54 types of absences
91
92
93  #start draw
94  Q1A2GP = filter(Q1A2, school == "GP") %>% mutate(rangeOfAbsences = cut(absences,c(-1,10,20,30,40,50,Inf))) %>%
95    group_by(rangeOfAbsences) %>% summarise(TotalGP=sum(Total))
96  Q1A2MS = filter(Q1A2, school == "MS") %>% mutate(rangeOfAbsences = cut(absences,c(-1,10,20,30,40,50,Inf))) %>%
97    group_by(rangeOfAbsences) %>% summarise(TotalMS=sum(Total))
98
99
100 #GP Absences Level
101 ggplot(Q1A2GP, aes(x=rangeOfAbsences, y=TotalGP)) +
102   geom_point(aes(shape = factor(rangeOfAbsences), colour = factor(rangeOfAbsences)), size=3) +
103   geom_text(aes(label=TotalGP),position = position_dodge(0.9),vjust = -1) +
104   ggtitle("School GP Total Range of Absences")
105
106
107 #MS Absences Level
108 ggplot(Q1A2MS, aes(x=rangeOfAbsences, y=TotalMS)) +
109   geom_point(aes(shape = factor(rangeOfAbsences), colour = factor(rangeOfAbsences)),size=3) +
110   geom_text(aes(label=TotalMS),position = position_dodge(0.9),vjust = -1) +
111   ggtitle("School MS Total Range of Absences")
112
113
```

Figure 2.2.1 Coding filter absences.

From Figure 2.2.1 I mainly use some functions of Data Manipulate to calculate the data I want. For example, calculating abstracts from each school, identifying how many rows are in the data, and rearranging the order of the data using the arrange function.

Since the number of absences is widely spaced, I use the cut function to define the range. Absences range I divided into 0-10, 10-20, 20-30, all the way to 50-60, 60 to Infinity. Since 0 cannot be used as a starting point, I set the starting point to -1.

After defining the range of abstracts, draw a point chart.

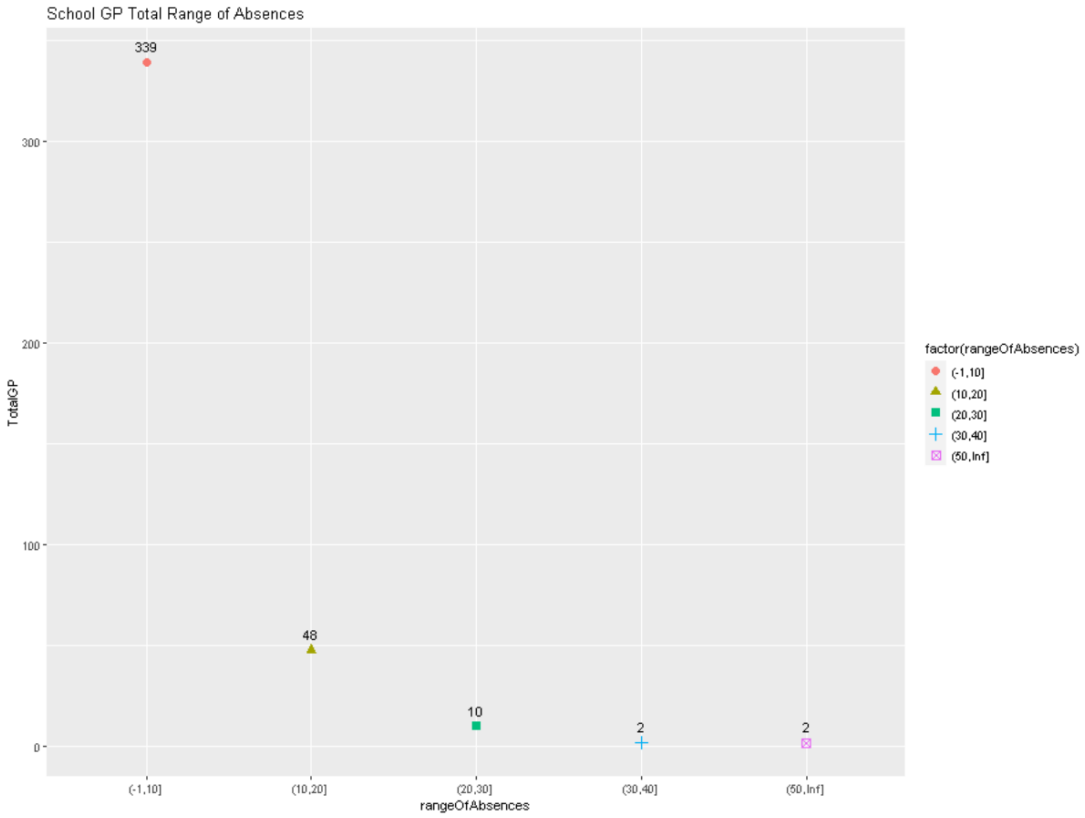


Figure 2.2.2 School GP total range of student absences.

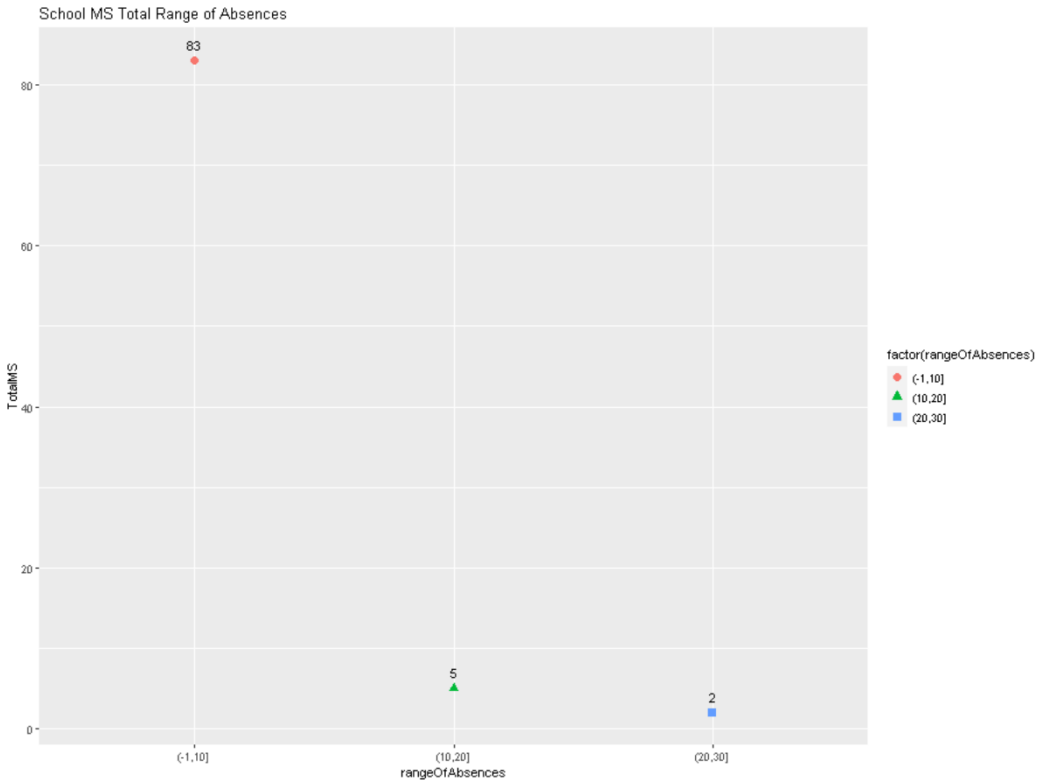


Figure 2.2.3 School MS Total range of student absences.

Let's start with the point chart of the GP school (Figure 2.2.2). The students of GP students occupy every position I locate. But we can immediately see from the chart that the number of absences above 20 is only 14, on the contrary, a full 387 students have an attendance rate of less than 20.

The attendance rate of the MS school (Figure 2.2.3) also appears to be on par with the performance of the GP school, with only 2 students in the MS school having more than 20 absences, and 88 students having less than 20 attendances.

From these two graphs, we can conclude that the attendance rate will affect the students' grades.

Analysis 1.3 how much percentages school support, and which types of students?

```
117 #Analysis 1-3
118 Q1A3Total = Q1 %>% group_by(schoolsup,school) %>% summarise(TotalSupport = n(), .groups = 'drop')  %>% select(TotalSupport)
119 Q1A3School = Q1 %>% group_by(schoolsup,school) %>% summarise(TotalSupport = n(), .groups = 'drop') %>% select(school)
120 Q1A3Support = Q1 %>% group_by(schoolsup,school) %>% summarise(TotalSupport = n(), .groups = 'drop') %>% select(schoolsup)
121
122
123 #draw bar chart
124 Q1A3DF = data.frame(Q1A3School, Q1A3Total,Q1A3Support)
125
126 ggplot(Q1A3DF, aes(x=school, y=TotalSupport, fill=schoolsup)) +
127   geom_bar(stat="identity", position="dodge") +
128   geom_text(aes(label=TotalSupport),position = position_dodge(1),vjust = -0.5)
129
130
131 #find out which type of student will got school support
132 check = filter(student_data, schoolsup == "yes")
133 nrow(check)
134
135 checkNo = filter(student_data, schoolsup == "no")
136 nrow(checkNo)
137
138
139 #reason 1
140 check %>% group_by(failures) %>% summarise(Total = n())
141
142   #Why no failures still got school support
143   check %>% group_by(failures) %>% filter(failures == 0) %>% group_by(higher) %>% summarise(n()) #this 0 is choose by they
144   check %>% group_by(failures) %>% filter(failures == 0) %>% group_by(paid) %>% summarise(n())
145   check %>% group_by(failures) %>% filter(failures == 0) %>% group_by(reason) %>% summarise(total=n()) %>%
146     mutate(Percent= (total/ sum(total) * 100 ))
147
148   #all student without not failures student
149   check %>% group_by(failures) %>% filter(failures != 0) %>% group_by(reason) %>% summarise(total=n()) %>%
150     mutate(Percent= (total/ sum(total) * 100 ))
151
152 #reason 2
153 check %>% group_by(paid) %>% summarise(n())
154
155 #reason 3
156 check %>% group_by(freetime) %>% summarise(total = n()) %>% mutate(Percent = total/sum(total) * 100)
157 check %>% group_by(freetime) %>% summarise(total = n()) %>% mutate(Percent = total/sum(total) * 100) %>%
158   summarise(compareA = Percent[1]+Percent[2], compareB = Percent[3]+Percent[4]+Percent[5])
159
```

Figure 2.3.1 Calculate Percentage of school support, and which type of student will get support.

Lines 118 to 129 of Figure 2.3.1, the same filtered data and calculations about grouping schools and schools supporting students and calculating how many students were supported and how many were not.

The resulting data is then displayed through a bar chart.

As for the row 132 below that is what I use to analyse what kind of students will get help from the school. Reason 1 I am analysing students without failures (row 143).

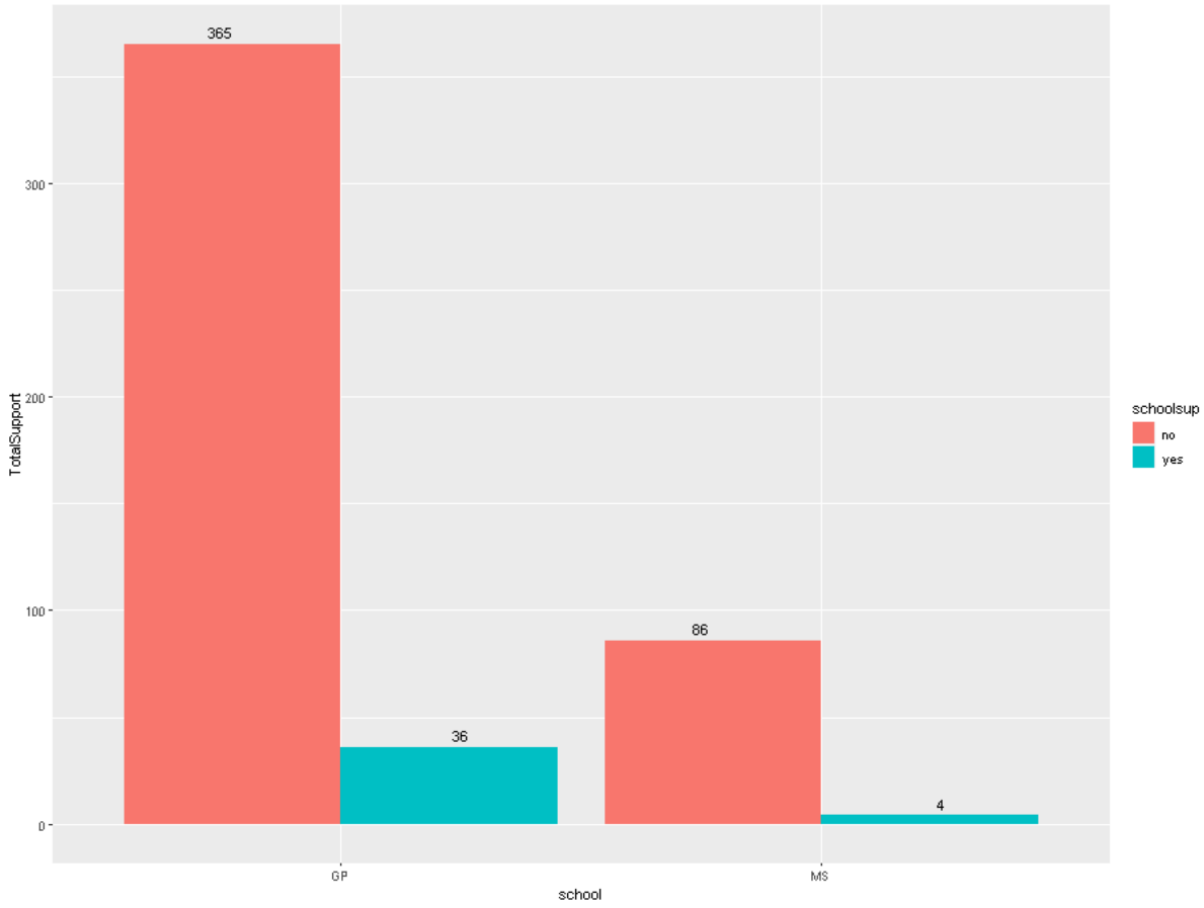


Figure 2.3.2 Total of school support by each school.

From the bar chart in Figure 2.3.2, we see that there are red and green. Red means not getting school support, while green means having school support. We can see that the GP school has 36 students who have school support, while the MS school has only 4 students. Most of the students who pass all grades do not receive school support.

```
R 4.1.2 · ~/
> #find out which type of student will got school support
> check = filter(student_data, schoolsup == "yes")
> nrow(check)
[1] 114
> checkNo = filter(student_data, schoolsup == "no")
> nrow(checkNo)
[1] 808
> #reason 1
> check %>% group_by(failures) %>% summarise(Total = n())
# A tibble: 4 x 2
  failures Total
     <int> <int>
1        0    89
2        1    16
3        2     4
4        3     5
>   #Why no failures still got school support
>   check %>% group_by(failures) %>% filter(failures == 0) %>% group_by(higher) %>% summarise(n()) #this 0 is choose by they
# A tibble: 1 x 2
  higher `n()`
  <chr>  <int>
1 yes       89
>   check %>% group_by(failures) %>% filter(failures == 0) %>% group_by(paid) %>% summarise(n())
# A tibble: 2 x 2
  paid  `n()`
  <chr> <int>
1 no       45
2 yes      44
>   check %>% group_by(failures) %>% filter(failures == 0) %>% group_by(reason) %>% summarise(total=n()) %>%
+     mutate(Percent= (total/ sum(total) * 100 ))
# A tibble: 4 x 3
  reason     total Percent
  <chr>      <int>   <dbl>
1 course        37   41.6
2 home          24   27.0
3 other          4    4.49
4 reputation    24   27.0
>   #all student without not failures student
>   check %>% group_by(failures) %>% filter(failures != 0) %>% group_by(reason) %>% summarise(total=n()) %>%
+     mutate(Percent= (total/ sum(total) * 100 ))
# A tibble: 4 x 3
  reason     total Percent
  <chr>      <int>   <dbl>
1 course         5      20
2 home           6      24
3 other          6      24
4 reputation     8      32
>
```

Figure 2.3.3 Result of reason 1 analyse which types of students will get school support.

```
R 4.1.2 · ~/
> #reason 2
> check %>% group_by(paid) %>% summarise(n())
# A tibble: 2 x 2
  paid  `n()`
  <chr> <int>
1 no       65
2 yes      49
> #reason 3
> check %>% group_by(freetime) %>% summarise(total = n()) %>% mutate(Percent = total/sum(total) * 100)
# A tibble: 5 x 3
  freetime total Percent
     <int> <int>   <dbl>
1        1    13   11.4
2        2     9    7.89
3        3    56   49.1
4        4    26   22.8
5        5    10    8.77
> check %>% group_by(freetime) %>% summarise(total = n()) %>% mutate(Percent = total/sum(total) * 100) %>%
+   summarise(compareA = Percent[1]+Percent[2], compareB = Percent[3]+Percent[4]+Percent[5])
# A tibble: 1 x 2
  compareA compareB
     <dbl>    <dbl>
1     19.3     80.7
>
```

Figure 2.3.4 Result of reason 2&3 analyse which types of students will get school support.

Of the 922 students analysed, only 114 students received school support.

First REASON 1 I checked these 114 failures for the students, only a few of them failed more than 1 lesson. The remaining 89 people passed the general subjects and still received school support. So, I made a special analysis for these 89 students. These 89 students all want to get a higher level of education. Among them, 51% of 45 students have no extra paid class. Then 68.6% of the students got school support because of course and reputation. The other 31.2% lived near the school or for other reasons.

Next is Figure 2.3.4 REASON 2 is for students with or without extra paid class, among which 65 students do not have extra paid class.

REASON 3 we can see Compare A and Compare B. Compare A is level 1 & 2 of free time, while Compare B is level of free time 3 & 4 & 5. The above data shows that most students have a lot of free time up to 80.67%.

In general, students who mainly get school support have the ability to get a higher education level and have extra time after school, followed by whether they have extra paid class.

Analysis 1.4 Distance from school affects grades.

```
164
165 #Analysis 1-4
166 Q1A4_1 = Q1 %>% group_by(address) %>% summarise(TotalAddress = n())
167
168 ggplot(Q1A4_1, aes(x=address, y=TotalAddress, fill=address)) +
169   geom_bar(stat="identity", position="dodge") +
170   geom_text(aes(label=TotalAddress),position = position_dodge(1),vjust = -0.5)
171
172
173 Q1A4_2 = Q1 %>% group_by(traveltime) %>% summarise(HomeToSchoolTime = n())
174
175 ggplot(Q1A4_2, aes(x=traveltime, y=HomeToSchoolTime)) +
176   geom_line() +
177   geom_text(aes(label=HomeToSchoolTime),position = position_dodge(1),vjust = 0)
178
```

Figure 2.4.1 Analysis address distance and the time to school.

Q1A4_1 from Figure 2.4.1 is mainly used to distinguish whether students live in urban areas or rural areas.

The other Q1A4_2 is to distinguish the time students travel from home to school.

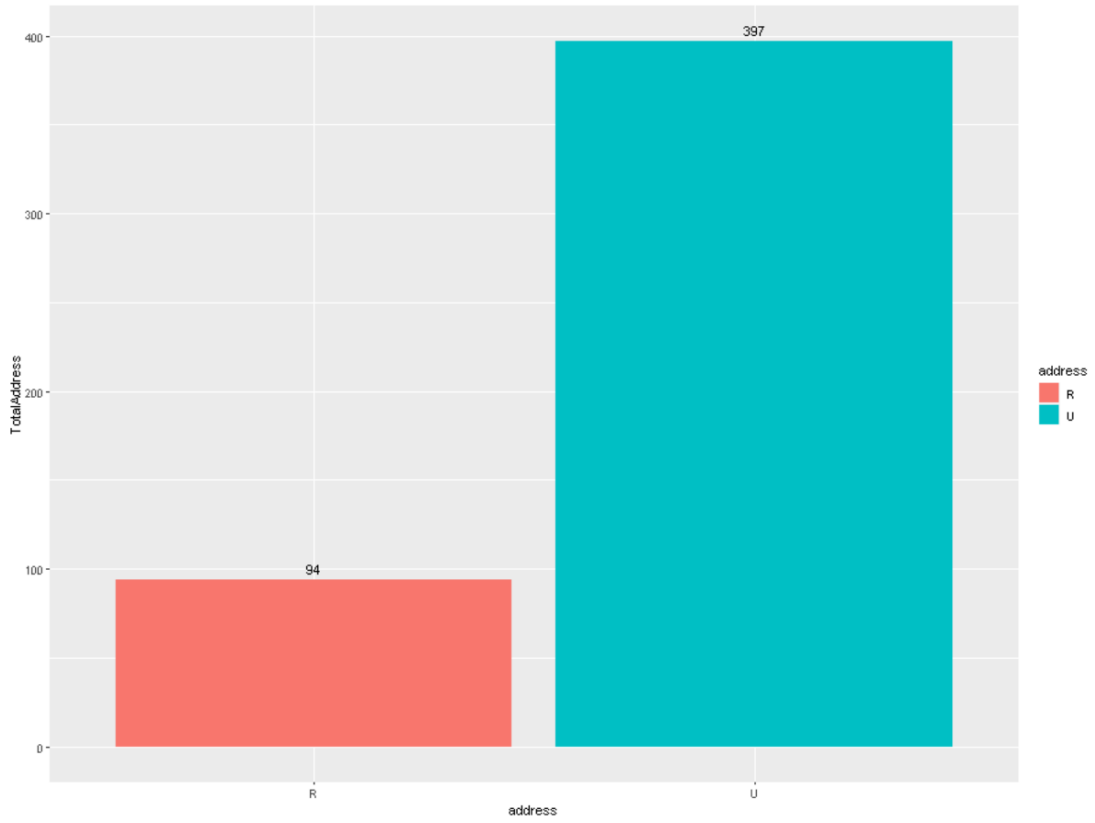


Figure 2.4.2 Bar Chart total of address, whether students live in urban or rural areas.

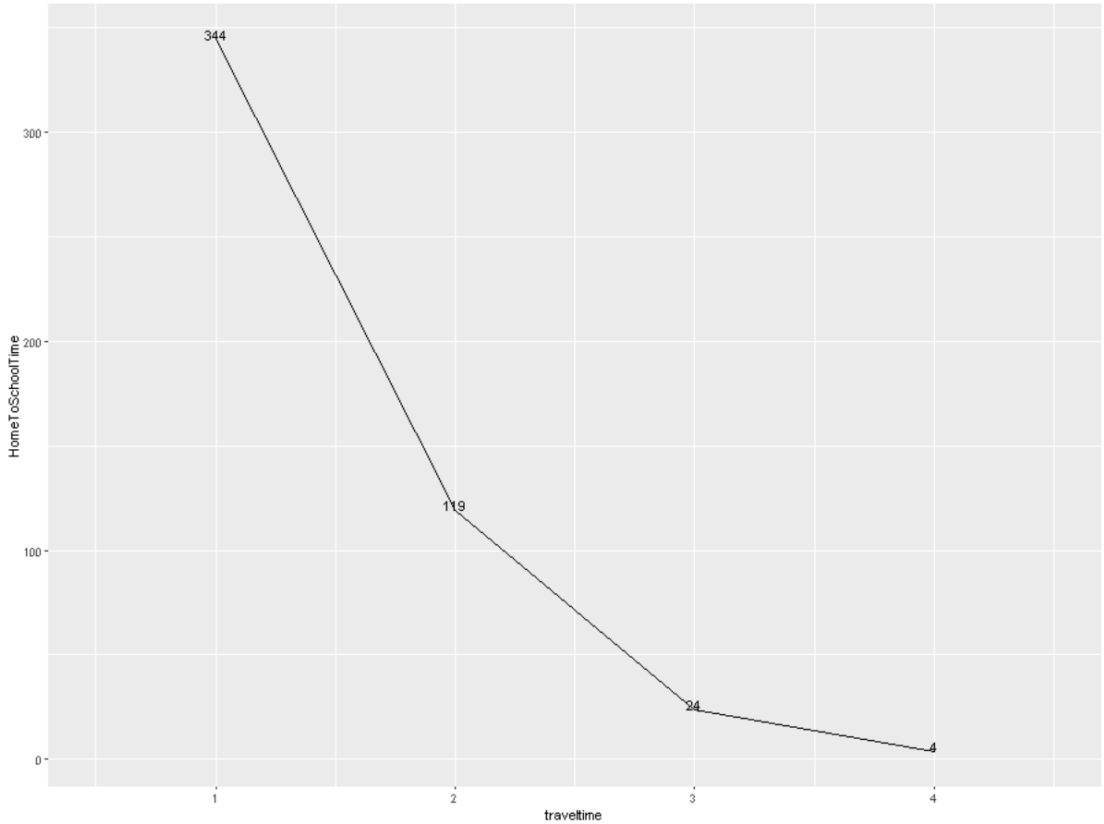


Figure 2.4.3 Line chart of home to school travel time.

From the bar chart in Figure 2.4.2 we see that 397 students live in urban areas, while 94 students live in rural areas.

Thus, the line chart in Figure 2.4.3 shows that only 28 students travelled from home to school principal for more than 3 hours.

It is not only because of personal reasons to choose which school to choose, but also we need to consider whether the students themselves live in urban or rural areas and consider the length of time they live to go to school.

Question 2: check how many percentages of students G1, G2 score no pass (no over 10), but G3 pass? All of them almost failed, and didn't get good scores in G3.

```
182
183 #Question 2
184 Q2 = filter(student_data, g1&g2 < 10, g3>=10 & g3<15)
185
```

Figure 3.0 Question 2 Global Filter

In question two, I want to find the students who fail in grade one and two, but pass on grade three (if a score over 10 means pass the grade). This global filter will be used for the whole question 2 analysis.

Analysis 2.1 Do family support and relationships both affect student achievement?

```
187
188 #Analysis 2-1
189 Q2A1 = Q2 %>% group_by(famrel, famsup) %>% summarise(Count = n())
190
191 ggplot(Q2A1, aes(x=famrel, y=Count, fill=famsup)) +
192   geom_bar(stat="identity", position="dodge") +
193   geom_text(aes(label=Count),position = position_dodge(1),vjust = -0.5)
194
195
```

Figure 3.1.1 Family Relation and Family education support coding

From figure 3.1.1 I will use data manipulation group_by function to group the family relationships , family education support of the students and count the number of each group.

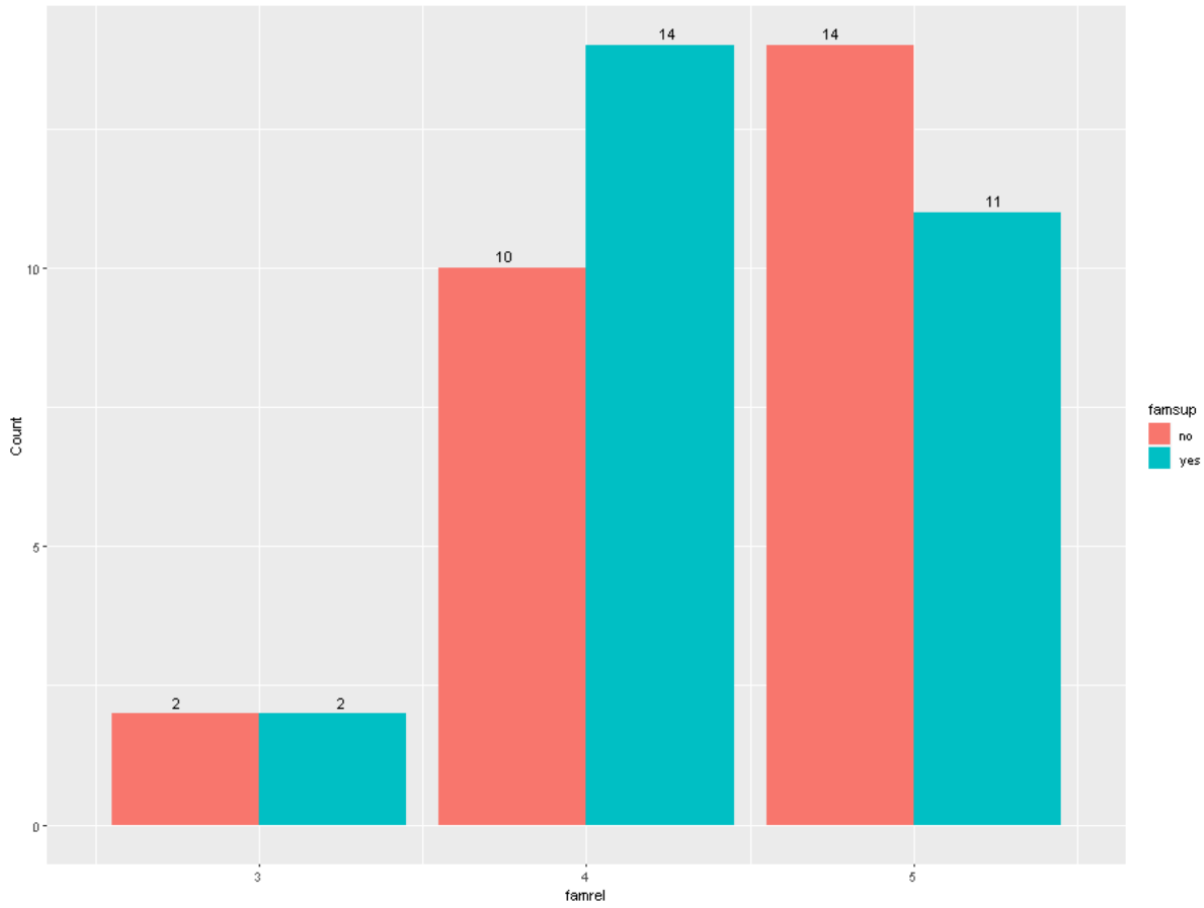


Figure 3.1.2 Count of the family education support and family relation.

Family relationships are important to students, because from bar chart data we can see only 4 students with family relationships are at level 3. This means that students with family are normal/ median relationships. And the other students have a high quality of family relationship, this can let students if any problem can ask the family for help.

Although 49% (26 of students) didn't get family education support, it still won't impact students pass the grade or getting the best score. Another 51% (27% of students) get family education support.

From this analysis we can get the conclusion that students with family relationships will impact a student's grade. But the family got given student education support or no, still not the major to impact a student's grade.

Analysis 2.2 Does not studying for a long time each week affect students' grades.

```
197
198 #Analysis 2-2
199 Q2A2 = Q2 %>% group_by(studytime) %>% summarise(CountTime = n())
200
201 ggplot(Q2A2, aes(x=studytime, y=CountTime, fill=studytime)) +
202   geom_bar(stat="identity", position="dodge") +
203   geom_text(aes(label=CountTime),position = position_dodge(1),vjust = -0.5)
204
```

Figure 3.2.1 Coding for count student's study time

In this analysis 2.2 the main reason I want to check student's study time, because we want to find out why students failed in grade one and two.

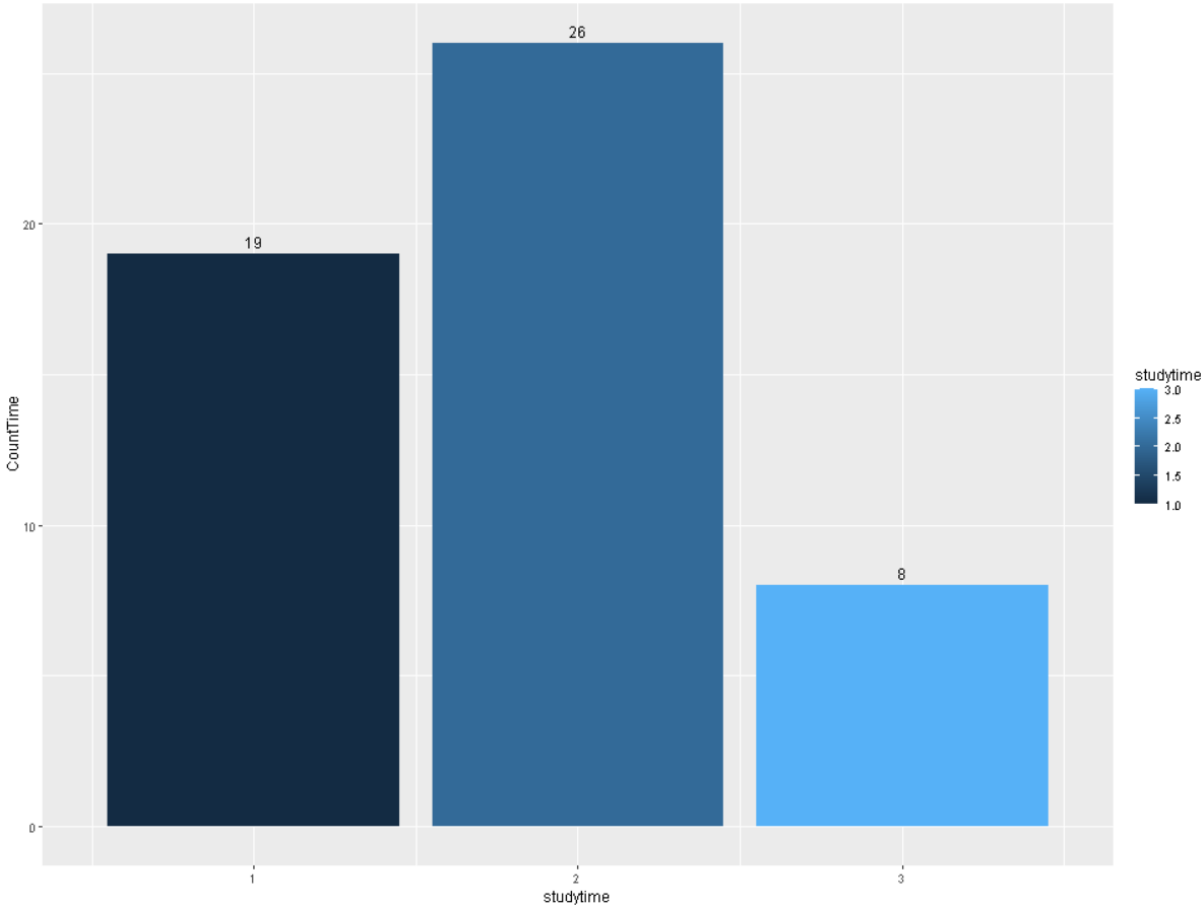


Figure 3.2.2 Count of the study time

Based on the above bar chart, we know 84. 9% (45 students) have less weekly study time. This is the main reason students fail in grade one and two. Because they only spend 1-2 hours weekly study or revision.

Analysis 2.3 Does the amount of alcohol a student drinks affect grades.

```
#Analysis 2-3
Q2A3DA = Q2 %>% group_by(dalc) %>% summarise(TotalDaily = n())

ggplot(data = Q2A3DA, aes(dalc, TotalDaily)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color="steelblue") +
  geom_text(aes(label=TotalDaily),position = position_dodge(1),vjust = -0.5) +
  labs(title = "Calculate daily alcohol consumption for students",
       subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
       y = "Count Of Students Number", x = "Daily Alcohol Consumption")


Q2A3WA = Q2 %>% group_by(walc) %>% summarise(TotalWeekly = n())

ggplot(data = Q2A3WA, aes(walc, TotalWeekly)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color="steelblue") +
  geom_text(aes(label=TotalWeekly),position = position_dodge(1),vjust = -0.5) +
  labs(title = "Calculate weekly alcohol consumption for students",
       subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
       y = "Count Of Students Number", x = "Weekly Alcohol Consumption")

```

Figure 3.3.1 daily alcohol and weekly alcohol

For this analysis I want to check if a student got drunk or not? And how many times for daily and weekly alcohol consumption. But I separate the part by daily and weekly.

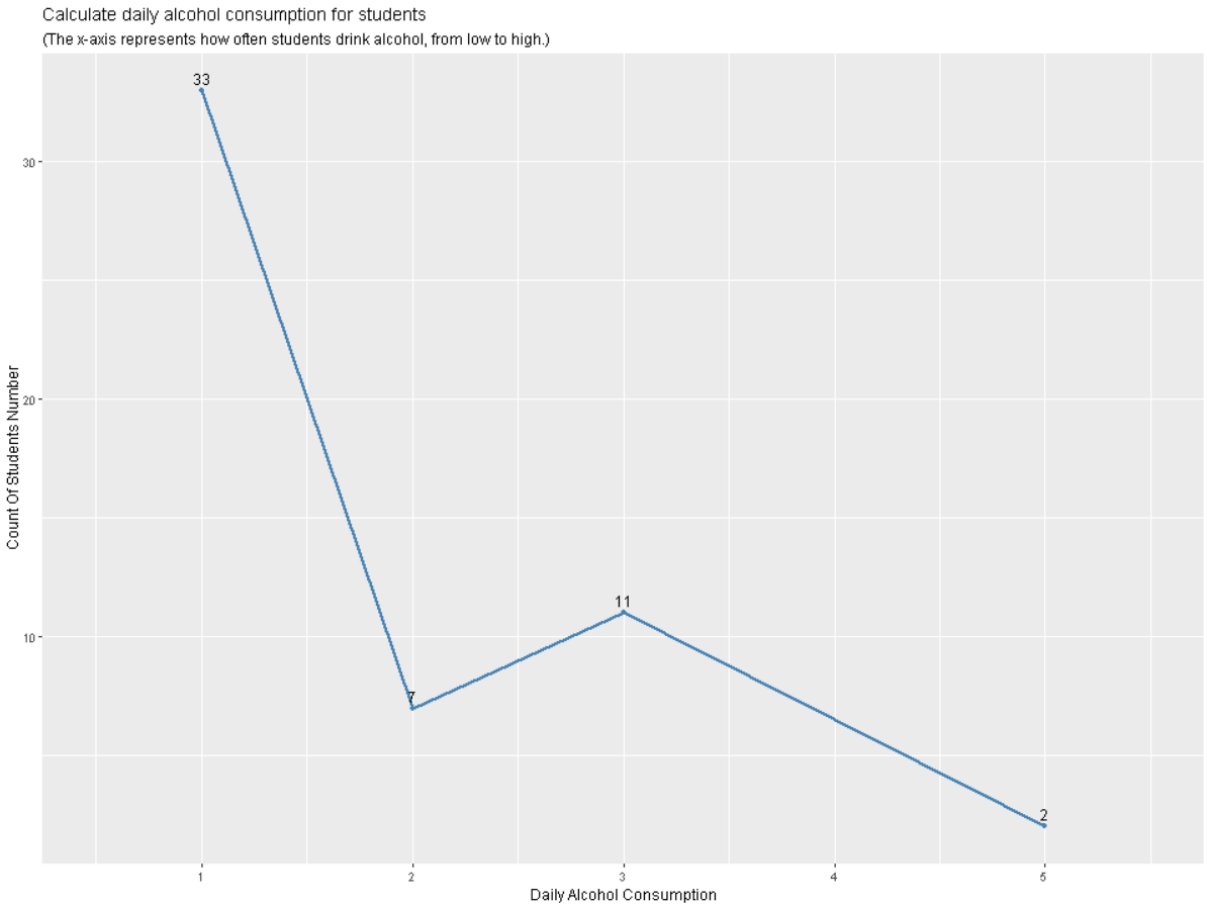


Figure 3.3.2 Line Chart Present the daily alcohol consumption for students.

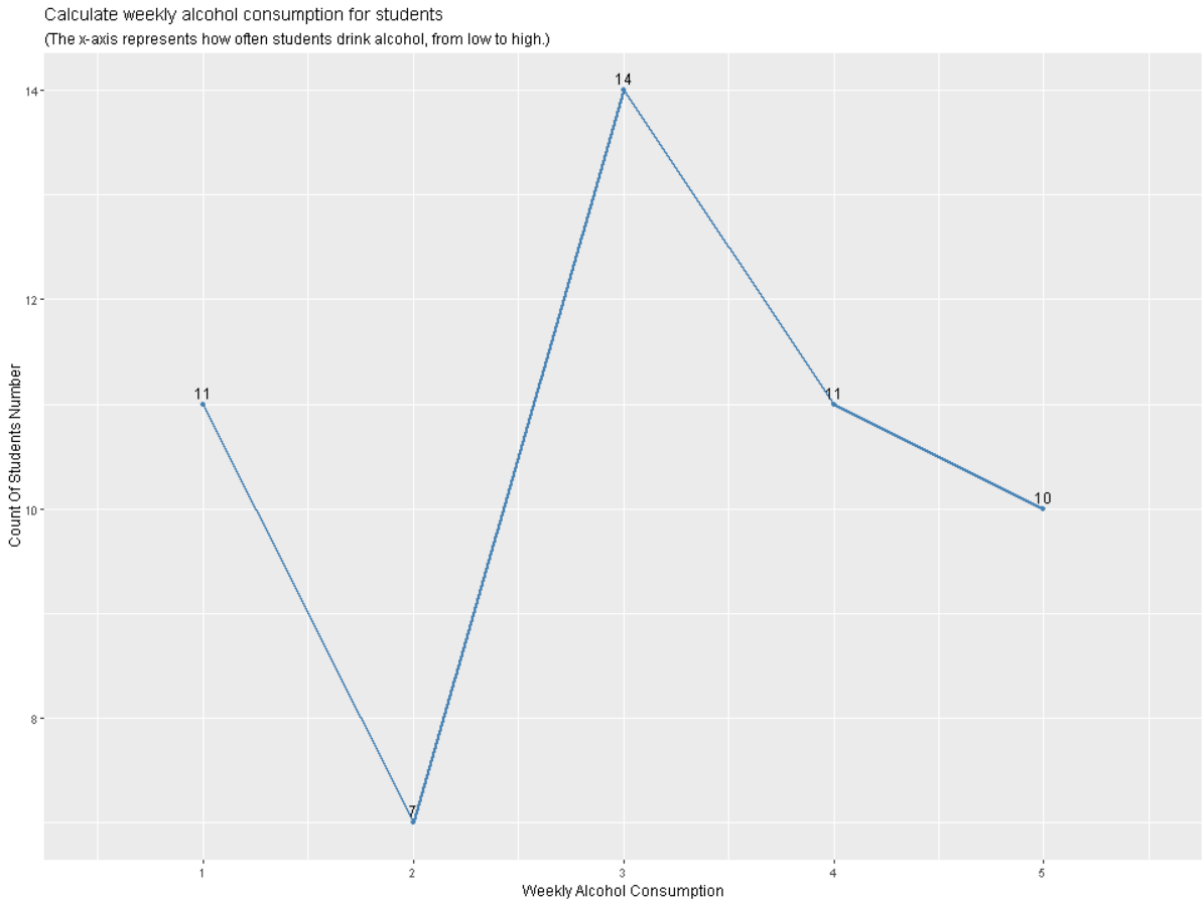


Figure 3.3.3 Line Chart Present the weekly alcohol consumption for students.

From figure 3.3.2 Line chart we can see a lot student's daily alcohol consumption stays at level 1, meaning they drink only 1 drink. Another 20 students drink alcohol over 2 or more times daily.

Figure 3.3.3 Weekly alcohol consumption line chart 35 students 66.04% are at the level 3 or more than 3. Because of this weekend, a lot of students will drink alcohol on the weekend.

Based on this kind of data, we can get the reason why students fail on grade one and two, but why they can pass on grade 3. Because a consistent weekday need focus on study so no drink a lot, and weekend got a lot free time can relax will drink a lot alcohol consumption.

Analysis 2.4 Students' addiction to the Internet affects their grades.

```
234
235 #Analysis 2-4
236 Q2A4 = Q2 %>% group_by(internet) %>% summarise(Total = n())
237 with(Q2A4,pie3D(Total,labels = Total, main = "Percentage Of Internet", explode = 0.2,
238                 col = rainbow(length(Total)), height = 0.17, labelcex = 1.8, labelcol = "black"))
239
```

Figure 3.4.1 Internet access at home

From this analysis 3.4 I want to check student access internet this reasons are impact a student's grade .And I use with function to write pie chart, because I filter it data type not allow direct write pie chart. So use with() to allow us to pass the arguments.

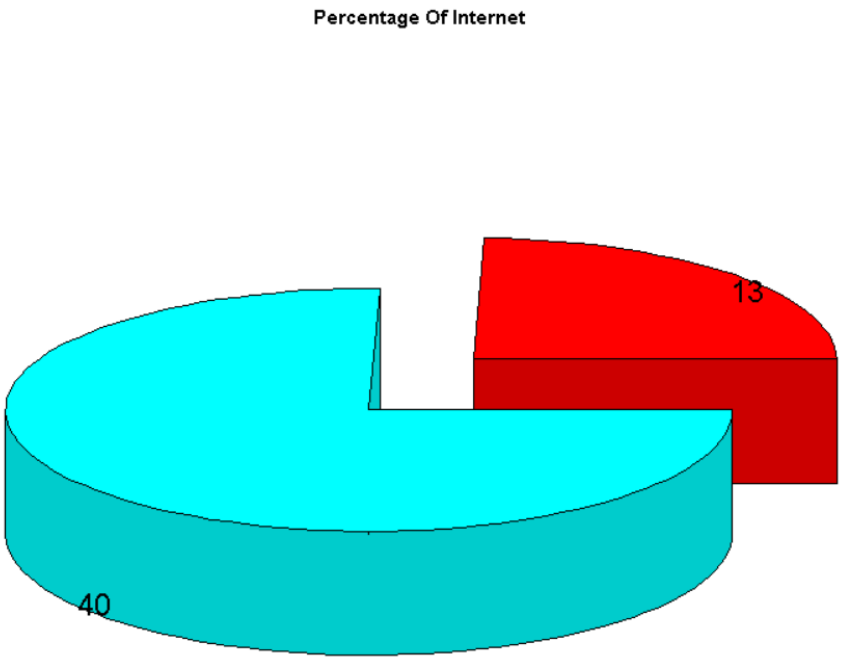


Figure 3.4.2 Percentage of Internet access at home.

From figure 3.4.2 pie chart we can see blue colour 40 students are access internet, other red colour 13 students are not access internet.

Analysis 2.5 Students who want a higher level of education will be more motivated to improve their grades.

```
241
242 #Features 2-1
243 Q2F1 = Q2 %>% group_by(higher) %>% summarise(Count=n())
244 with(Q2F1,pie3D(Count,labels = higher, main = "Percentage Of want take higher education level",
245                 explode = 0.2, col = hcl.colors(length(Count),"TealRose"), height = 0.17, labelcex = 1.2,theta = 0.9, labelcol = "black"))
246
```

Figure 3.5.1 student want to take higher education level

This analysis 3.5 I want to verify if students want to take a high education level, they will have more motive to improve their score. So when I group the data column getting yes or no, after counting by each other. Then draw the pie chart to present the result.

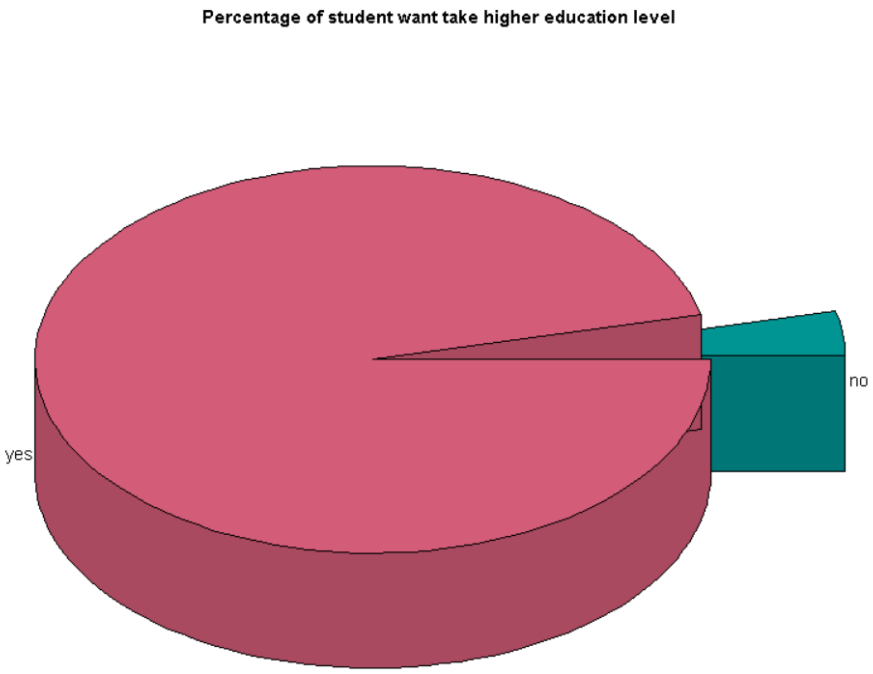


Figure 3.5.2 Percentage of student want higher education level

We can see there are a lot of students who want to take a higher education level, so they are more motivated to pass the grade three.

Analysis 2.6 Does the number of absences affect a student's grade.

```
#Features 2-2
Q2F2A2 = Q2 %>% group_by(absences) %>% mutate(rangeOfAbsences = cut(absences,c(-1,10,20,30,40,50,Inf))) %>%
  group_by(rangeOfAbsences) %>% summarise(TotalRange = n())

ggplot(Q2F2A2, aes(x=rangeOfAbsences, y=TotalRange)) +
  geom_point(aes(shape = factor(rangeOfAbsences), colour = factor(rangeOfAbsences)), size=3) +
  geom_text(aes(label=TotalRange),position = position_dodge(0.9),vjust = -1) +
  ggtitle("Total Range of Absences")

```

Figure 3.6.1 number of absences

In this analysis 3.6 I want to check absences, but there are a lot of different absences. So I group the number of absences, and then add a new column using the cut() function to separate the range of the absences. To allow us to easily separate the absence.

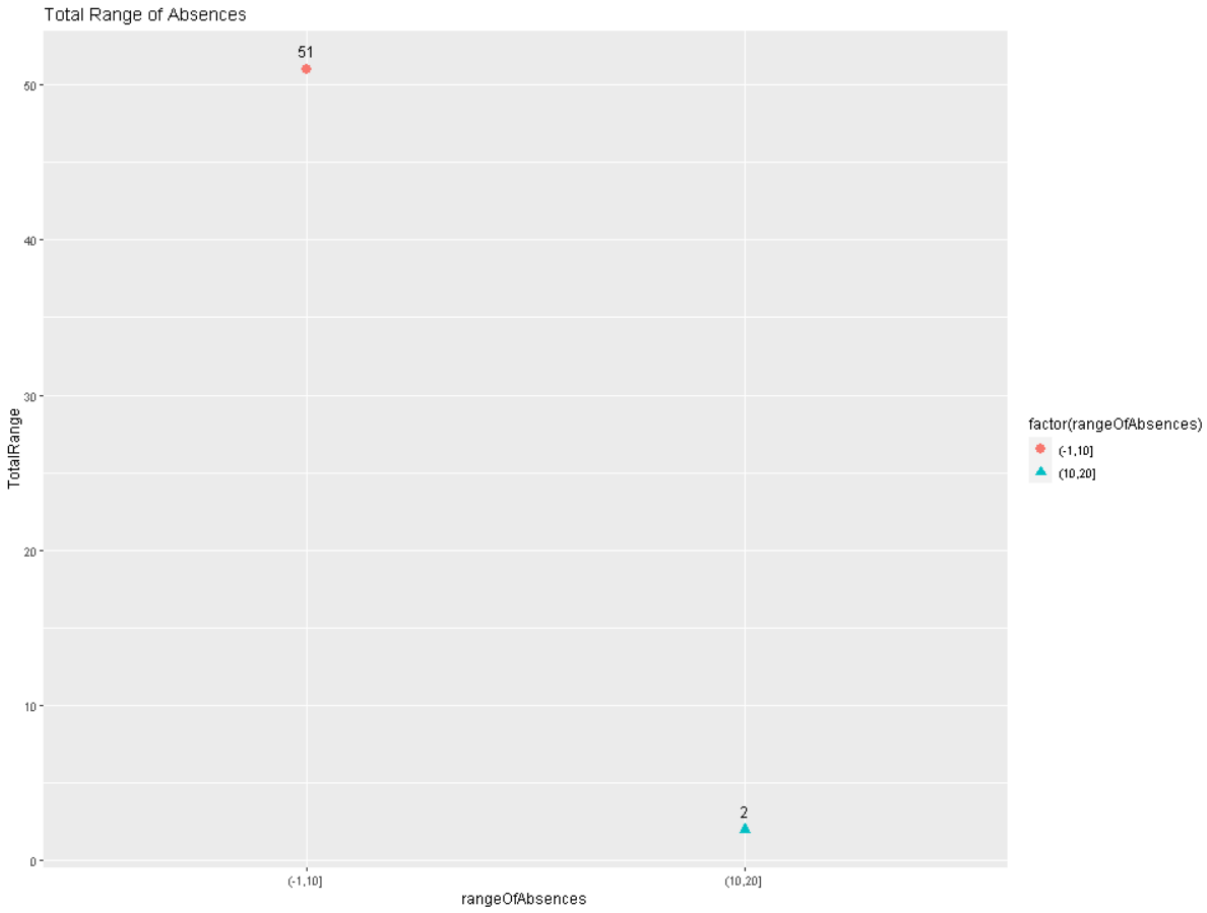


Figure 3.6.2 Total range of absences

From figure 3.6.2 we can see two ranges of absences, one is 0-9 times absences got 51 students in this range, another one is 10-20 times got 2 students only. So that is another reason why there students can pass grade three, cause they present class high percentages.

Analysis 2.7 Does Student Dating Affect Grades.

```
259
260 #Features Analysis 2-3
261 Q2F2A3 = Q2 %>% group_by(romantic) %>% summarise(Total = n())
262 with(Q2F2A3,pie3D(Total,labels = romantic, main = "Percentage Of romantic", explode = 0.2,
263                  col = rainbow(length(Total)), height = 0.17, labelcex = 1.5, theta = 0.8))
264
265
```

Figure 3.7.1 student with a romantic relationship

I want to verify if romantic relationships will impact students who pass the grade or get the best score.

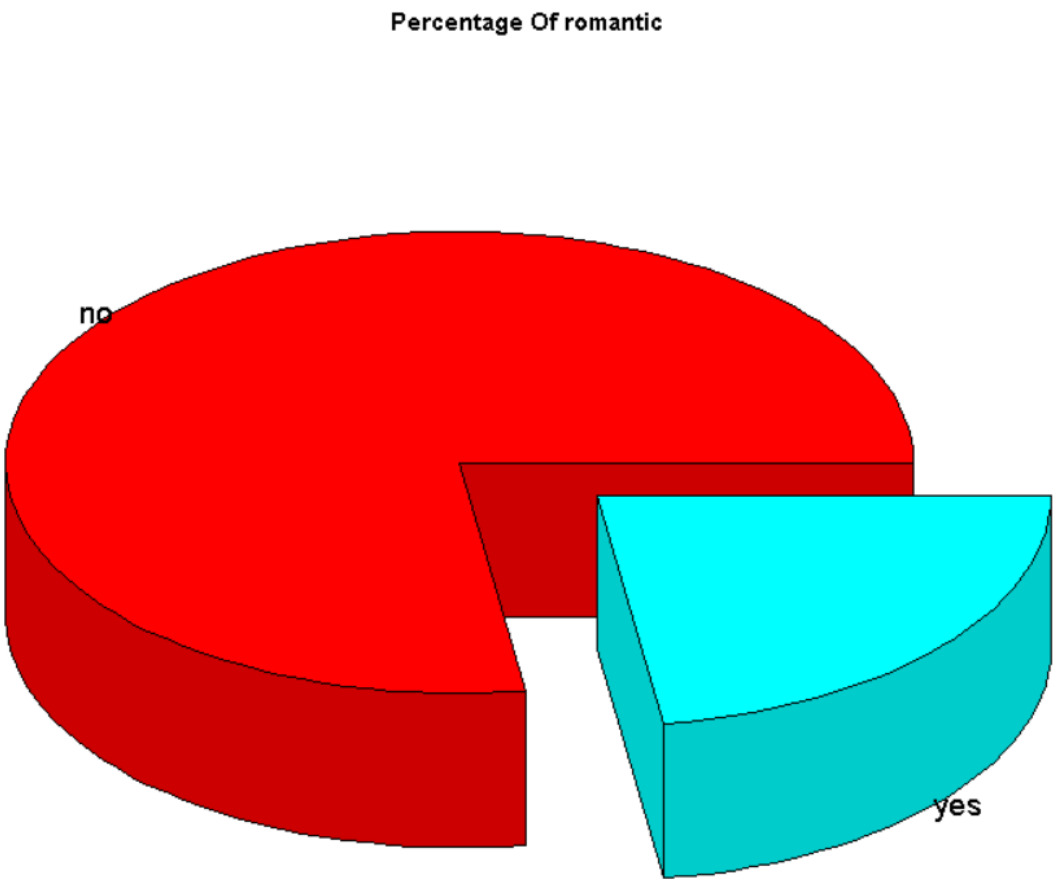


Figure 3.7.2 Percentages of student romantic relationship status.

Figure 3.7.2 shows that the majority of students are not in a relationship, and only a portion of the students are in a relationship. Therefore, romantic relationships are also a component that affects students' grades, allowing these students to pass several times in grade 3.

Question 3: check which type of student get best score in G1, G2, and G3? Why?

```
268
269 #Question 3
270 Q3 = filter(student_data, g1&g2&g3 >= 15)
271 nrow(Q3)
272
```

Figure 4.0 Question 3 Global Filter. I filter it every grade with a score equal and over 15 until 20 students.

Analysis 3.1 Course and reputation will have an impact on maintaining top grades.

```
274
275 #Analysis 3-1
276 Q3A1 = Q3 %>% group_by(reason) %>% summarise(Total = n())
277
278 with(Q3A1,pie3D(Total,labels = Total, main = "Total Best Student Reason To Choose This School",
279                 explode = 0.2, col = hcl.colors(length(Total), "Spectral"),
280                 height = 0.17, labelcex = 1.5,labelcol = "red", border="black",theta = 0.9))
281
```

Figure 4.1.1 Reason for choose school

In this analysis I want to find out if students' choice of course or reputation will improve and motivate them to increase their score.

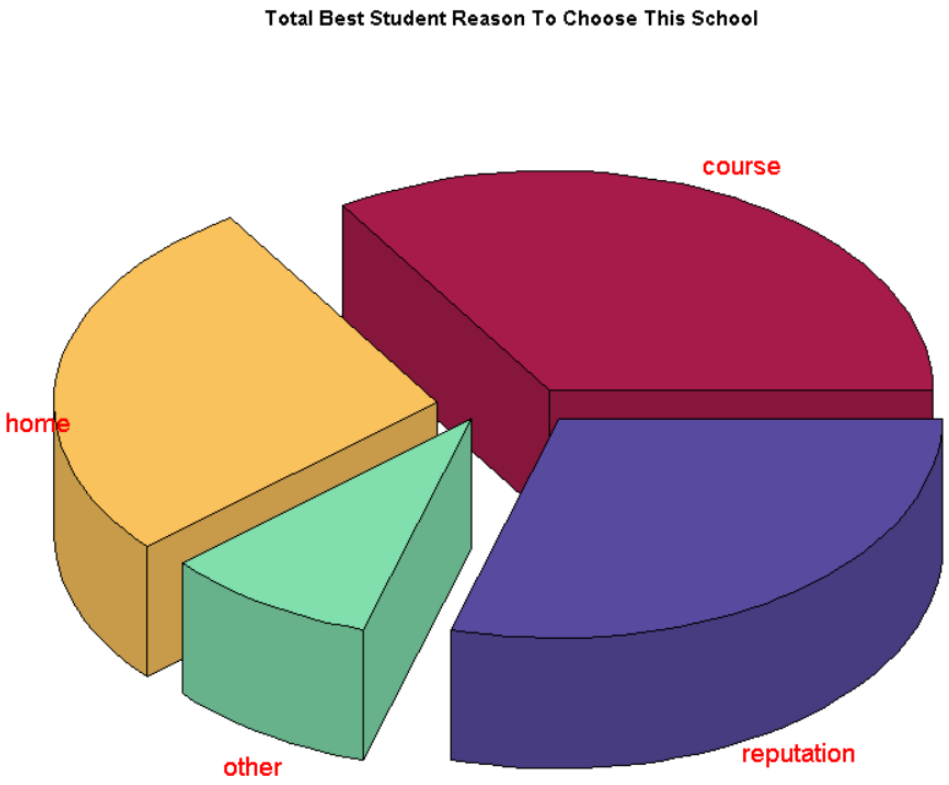


Figure 4.1.2 Total Best Student Reason to Choose This School.

From there all best scores students(every grade more than 15 score) most choose course or reputation, and the next is related home distances of school. The last is another reason to choose school.

Analysis 3.2 Does the number of absences affect a student's grade.

```
283
284 #Analysis 3-2
285 Q3A2 = Q3 %>% group_by(absences) %>% mutate(rangeOfAbsences = cut(absences,c(-1,10,20,30,40,50,Inf))) %>%
286   group_by(rangeOfAbsences) %>% summarise(TotalRange = n())
287
288 ggplot(Q3A2, aes(x=rangeOfAbsences, y=TotalRange, fill=rangeOfAbsences)) +
289   geom_bar(stat="identity", position="dodge") +
290   geom_text(aes(label=TotalRange),position = position_dodge(1),vjust = -0.5)
291
292
```

Figure 4.2.1 summary and calculate the absences.

I want checking absences to affect a student's grade, so I group the absences times and use cut() function to make the absences range.

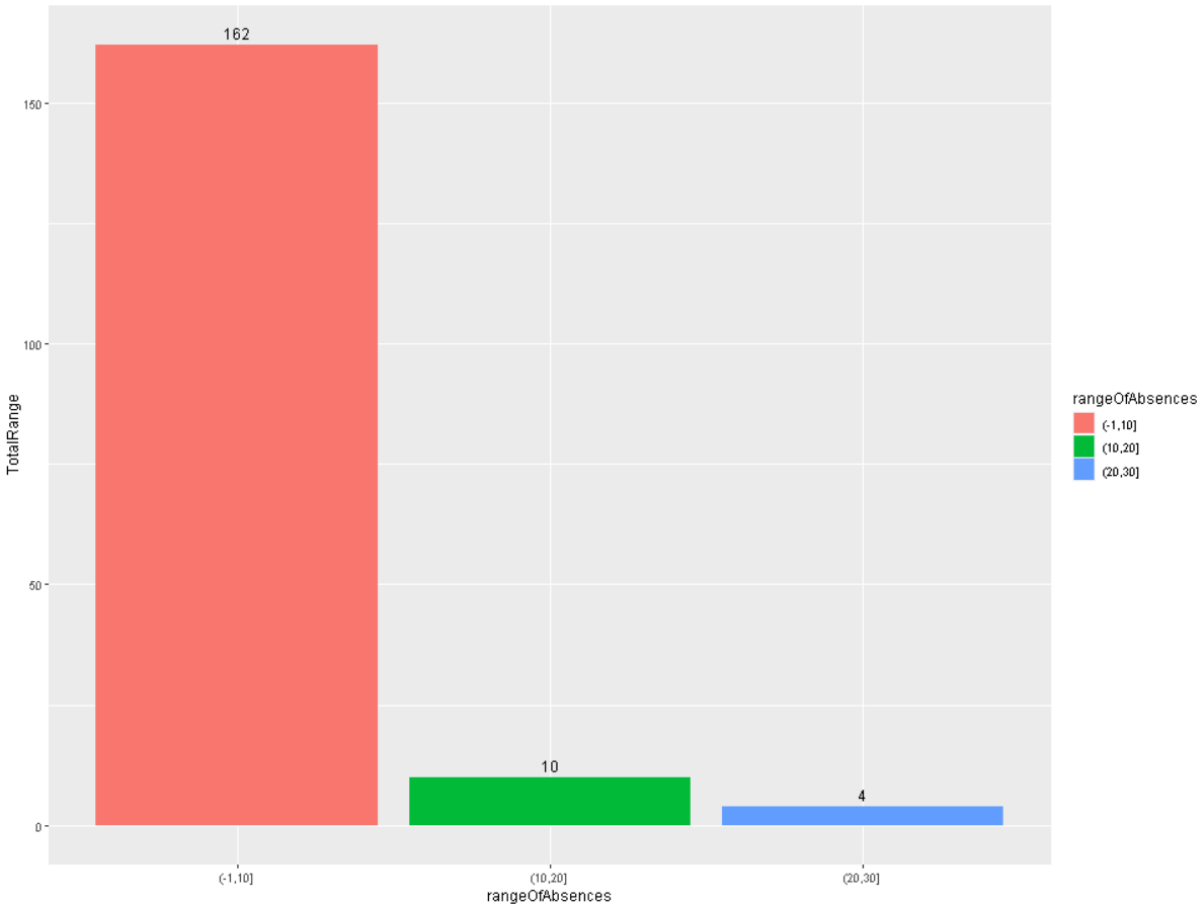


Figure 4.2.2 Total range of absences.

From figure 4.2.2 bar chart we can get the range and number of absences. From a range of [0,10] 162 students, another 14 students are between 10 to 30 times absent.

In conclusion, they all best score students not absences too many times, because they will miss some class or impact a student's score.

Analysis 3.3 Family relationship and support lead to better grades for students.

```
#Analysis 3-3
Q3A3 = Q3 %>% group_by(famrel, famsup) %>% summarise(Count = n())


ggplot(Q3A3, aes(x=famrel, y=Count, fill=famsup)) +
  geom_bar(stat="identity", position="dodge") +
  geom_text(aes(label=Count),position = position_dodge(1),vjust = -0.5) +
  scale_fill_manual("famsup", values = c("no" = "#FA8072", "yes" = "#228B22")) #light red & forest green color

```

Figure 4.3.1 family relationship and family education support.

From figure 4.1.1 I will use data manipulation group_by function to group the family relationships , family education support of the students and count the number of each group. This analysis allow me to make these can lead better grades for students or not.

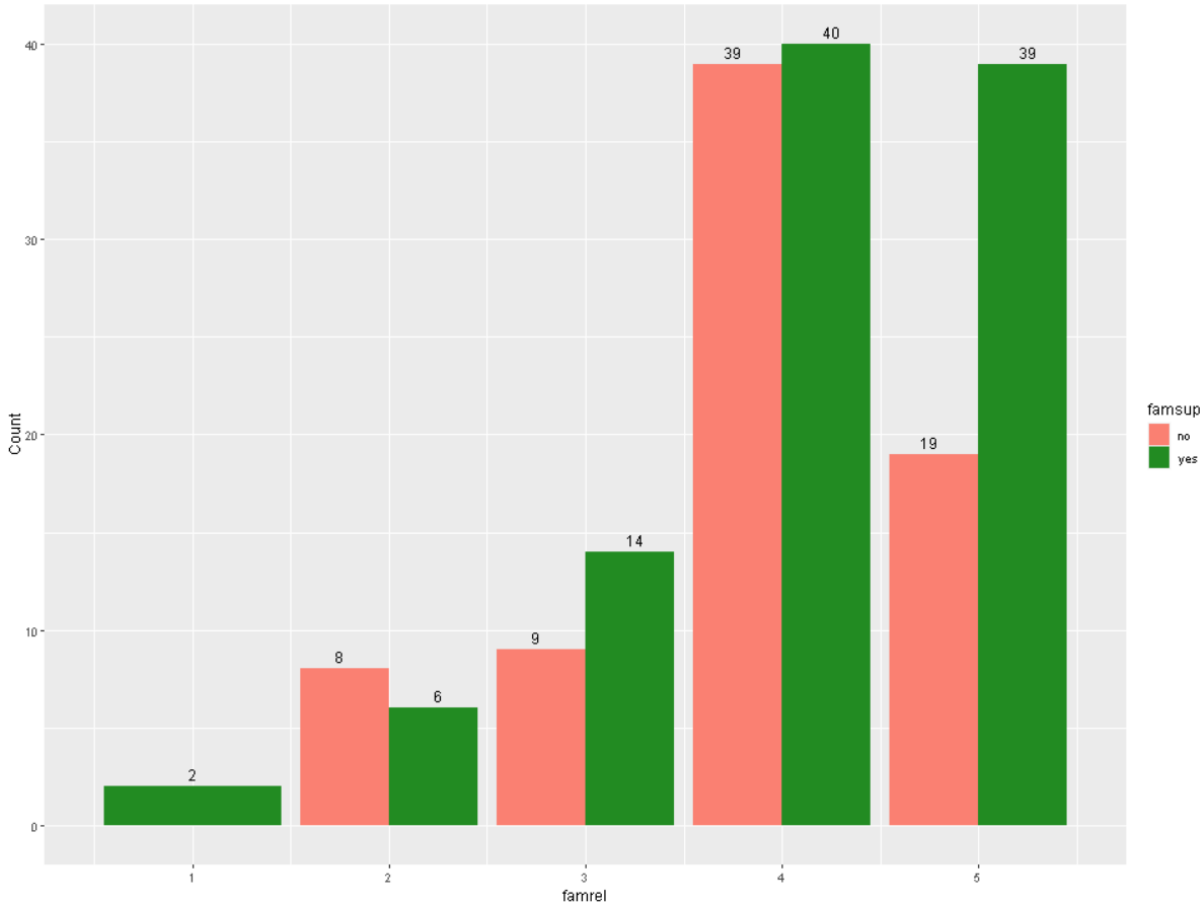


Figure 4.3.2 Bar Chart Count of the student with family relationship and family support.

From the bar chart present 5 different levels with family relationships, number of students and different kinds of color to present family education support or not.

At the family relationships levels 4-5 79 students got family support, and 58 students did not get the family education support.

The whole bar chart we can see got 101 student family education support, another 75 students didn’t get the family education support.

From this analysis we can get the conclusion that students with family relationships will impact a student's grade. But the family was given student education support or no, still not the major to impact a student's grade. And a lot of percentages are students who got family education support become best score students.

Analysis 3.4 Are high-achieving students still single?

```
304
305 #Analysis 3-4
306 Q3A4 = Q3 %>% group_by(romantic) %>% summarise(Count = n())
307
308 with(Q3A4,pie(Count,labels = romantic, main = "Count of Best Student Romantic Status",
309                col = hcl.colors(length(Count), "TealGrn"),clockwise = TRUE))
310
```

Figure 4.4.1 student with a romantic relationship.

This analysis I want to check it’s romantic relationship will impact a student's grade, and let students can’t focus on their study.

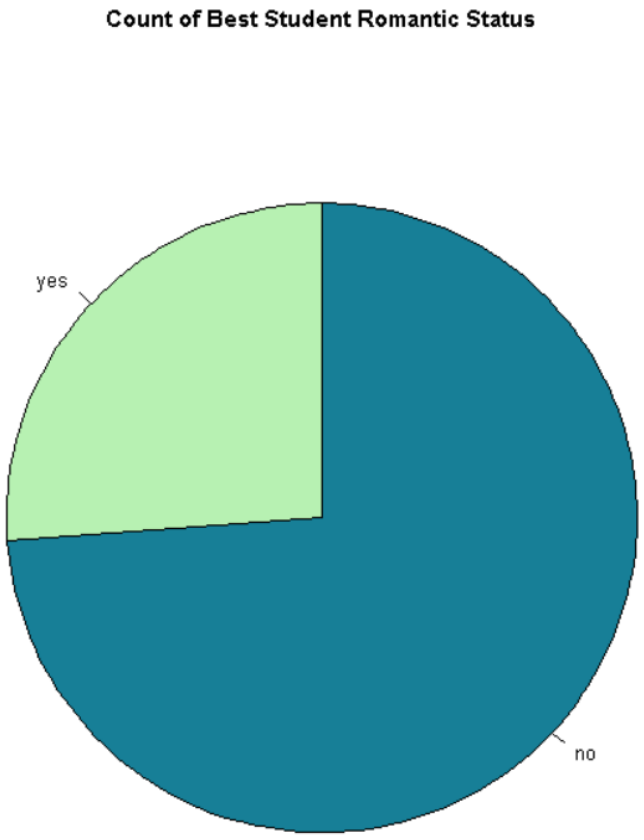


Figure 4.4.2 Count of best student romantic status.

Figure 4.4.2 is showing that 70% of students are not romantic, and only a separate part of students have a romantic relationship. So romantic relationships also are a component to impact a student's grade, based on global filter figure 4.0, the best score students most didn't have.

Analysis 3.5 Do high-achieving students necessarily need a good parental educational background?

```
312
313 #Analysis 3-5
314 Q3A5FE = Q3 %>% group_by(fedu) %>% summarise(TotalFatherEducationLevel = n())
315 Q3A5 = Q3 %>% group_by(medu) %>% summarise(TotalMotherEducationLevel = n()) %>% mutate(Q3A5FE)
316
317
318 Q3A5DF = data.frame(Q3A5)
319 Q3A5Melt = melt(Q3A5DF[,c("medu","TotalMotherEducationLevel","TotalFatherEducationLevel")], id.vars=1)
320
321 Q3A4Bar <- ggplot(Q3A5Melt, aes(x=medu, y=value, fill=variable)) +
322   geom_bar(stat="identity", position="dodge") +
323   geom_text(aes(label=value),position = position_dodge(1),vjust = -0.5)
324
325 print(Q3A4Bar + labs(title = "Count Of Parent Education Level", y = "Count", x = "Parent Education Level"))
326
```

Figure 4.5.1 Father & Mother education level

This analysis verified its best score students need to have a high quality parent education background. First I group the data and total the education background. And store it into a data frame to allow me to use the melt() function to reshape the data.

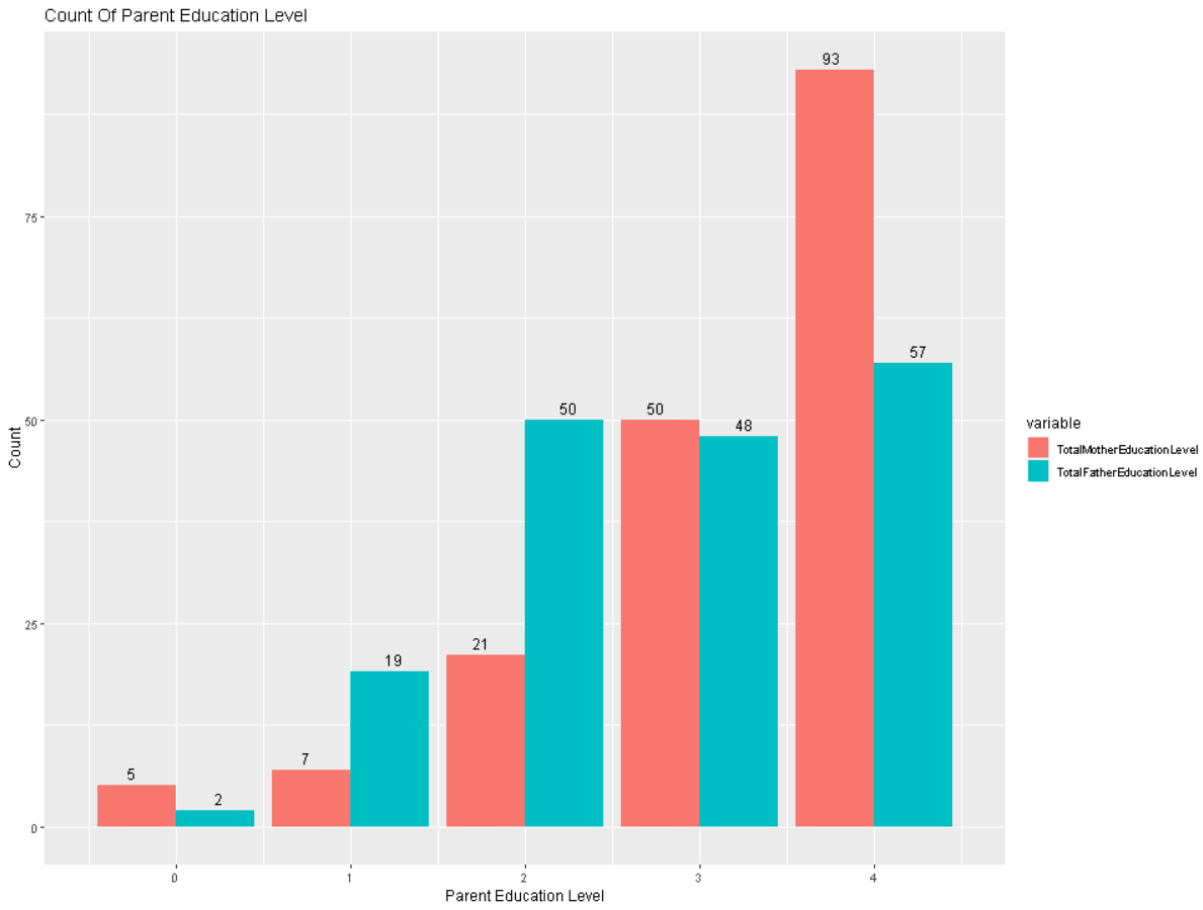


Figure 4.5.2 Count of Parent education Level

From figure 4.5.2 shown parent education level 0 to 5, 0 means didn't have any education background, 1 are present primary education (4th grade), level 2 are 5th to 9th grade, level 3 means secondary education and the last level 4 is higher education. And the blue color column is present father, another red color column is present mother.

Based on the above bar chart we can see that starting from level 2 parent education level, the total of the mother education level is over the father education level. The mother total 164, and the father total 156. But at the major, parent education background will be a major component of becoming a best student.

Analysis 3.6 Does the amount of alcohol a student drinks affect grades.

```
#Analysis 3-6
Q3A6DA = Q3 %>% group_by(dalc) %>% summarise(TotalDaily = n())

ggplot(data = Q3A6DA, aes(dalc, TotalDaily)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color="steelblue") +
  geom_text(aes(label=TotalDaily),position = position_dodge(1),vjust = -0.5) +
  labs(title = "Calculate daily alcohol consumption for students",
       subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
       y = "Count Of Students Number", x = "Daily Alcohol Consumption")



Q3A6WA = Q3 %>% group_by(walc) %>% summarise(TotalWeekly = n())

ggplot(data = Q3A6WA, aes(walc, TotalWeekly)) +
  geom_line(color = "steelblue", size = 1) +
  geom_point(color="steelblue") +
  geom_text(aes(label=TotalWeekly),position = position_dodge(1),vjust = -0.5) +
  labs(title = "Calculate weekly alcohol consumption for students",
       subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
       y = "Count Of Students Number", x = "Weekly Alcohol Consumption")


```

Figure 4.6.1 daily & weekly alcohol consumption.

From this analysis I want to check whether students still like to drink alcohol or not, and check their daily & weekly alcohol consumption.

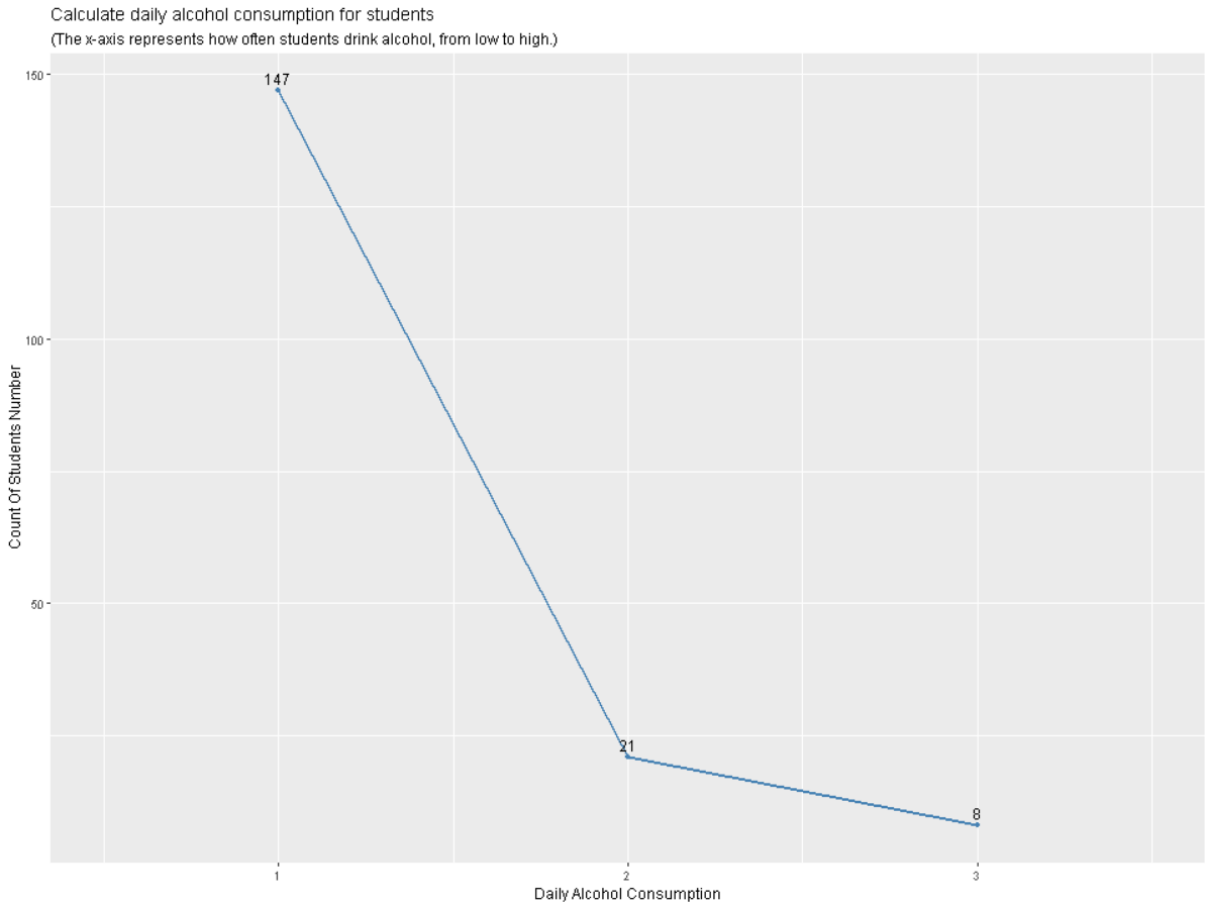


Figure 4.6.2 Calculate daily alcohol consumption for students.

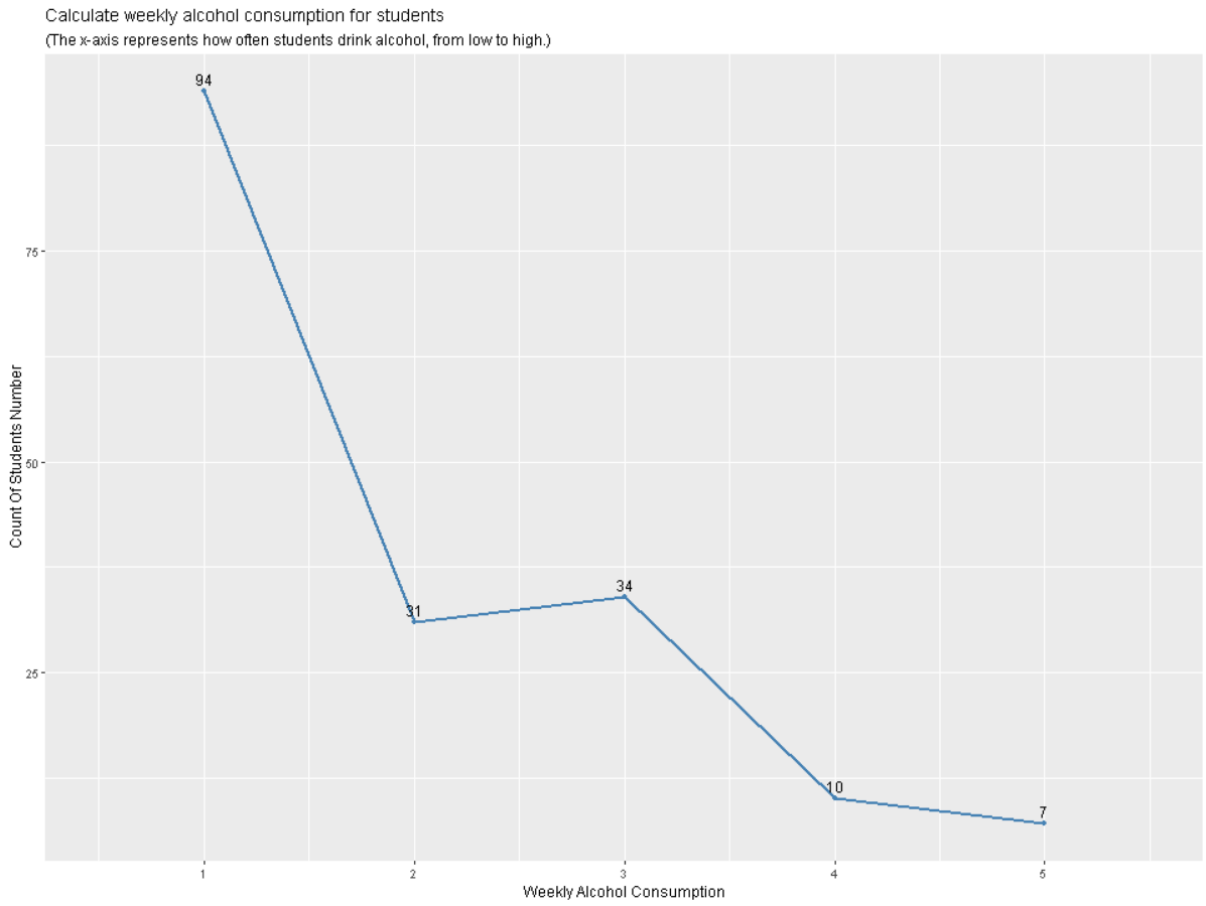


Figure 4.6.3 Calculate Weekly alcohol consumption for students.

From the line graph in Figure 4.6.2, we can see that the daily alcohol consumption of 147 students remained at level 1, which means that their daily alcohol consumption was not large. Another 209 students drank more than 2 levels of alcohol per day. Means they may have a drinking habit.

Figure 4.6.3 Weekly drinking line chart We can see that 94 students are at level 1, and the other 82 students are above the level of drinking level 2 or above, the higher the drink, the more.

Based on these data, we can draw the reasons why students pass all grades, because most of these students drink more frequently during the weekend. To make them more able to focus on their studies on weekdays, even on weekends, 94 students are level 1 drinkers.

Question 4: find out why G1, G2 pass but G3 failed a lot?

```
358
359 #Question 4
360 Q4 = filter(student_data, (g1&g2 >=10 & g3 < 10))
361 nrow(Q4) #19 students
362
```

Figure 5.0 Question 4 Global Filter.

This global filter was used by the whole question 5 analysis, discussing what reason students pass grade one and two, but fail on grade three.

Analysis 4.1 Students' addiction to the Internet affects their grades.

```
364
365 #Analysis 4-1
366 Q4A1 = Q4 %>% group_by(internet) %>% summarise(Count= n())
367
368 with(Q4A1,pie(Count,labels = internet, main = "Count of students accessing the Internet", col = Count))
369
370
```

Figure 5.1.1 student access internet.

I would like to verify it's student access internet because they fail on grade three. I also use simple data manipulation function to filter the result.

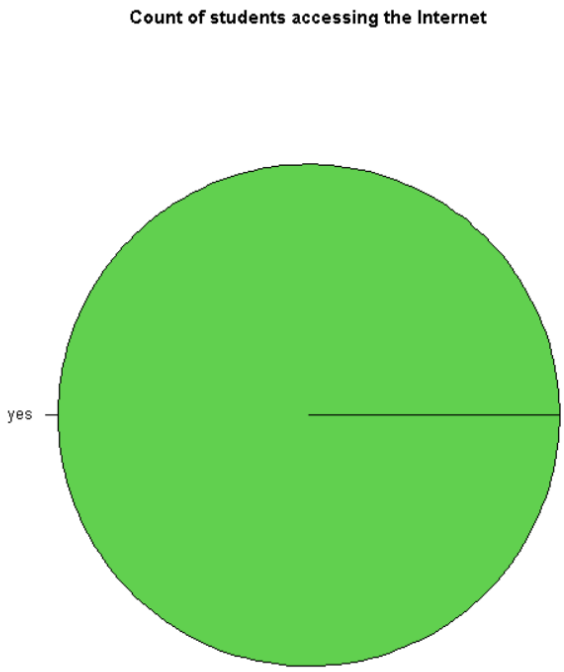


Figure 5.1.2 Count of students accessing the internet.

From figure 5.1.2 shown 100% students are accessing the internet cause they fail on grade three. Usually they access the internet and have no more time to study, improve their knowledge or do revision. So the main reason is that the internet can't focus on study.

Analysis 4.2 Do top students have extracurricular activities.

```
371
372 #Analysis 4-2
373 Q4A2 = Q4 %>% group_by(activities) %>% summarise(Count = n())
374
375 with(Q4A2,pie3D(Count,labels = activities, main = "Count of students extra-curricular activities",
376                 explode = 0.25, col = hcl.colors(length(Count), "Temps"), height = 0.17,
377                 labelcex = 1.5,labelcol = "black", border="black", theta = 0.9))
378
```

Figure 5.2.1 extra-curricular activities.

This analysis is the main reason to check students over join the extra-curricular activities to impact their ability to focus on study, because they fail on grade three.

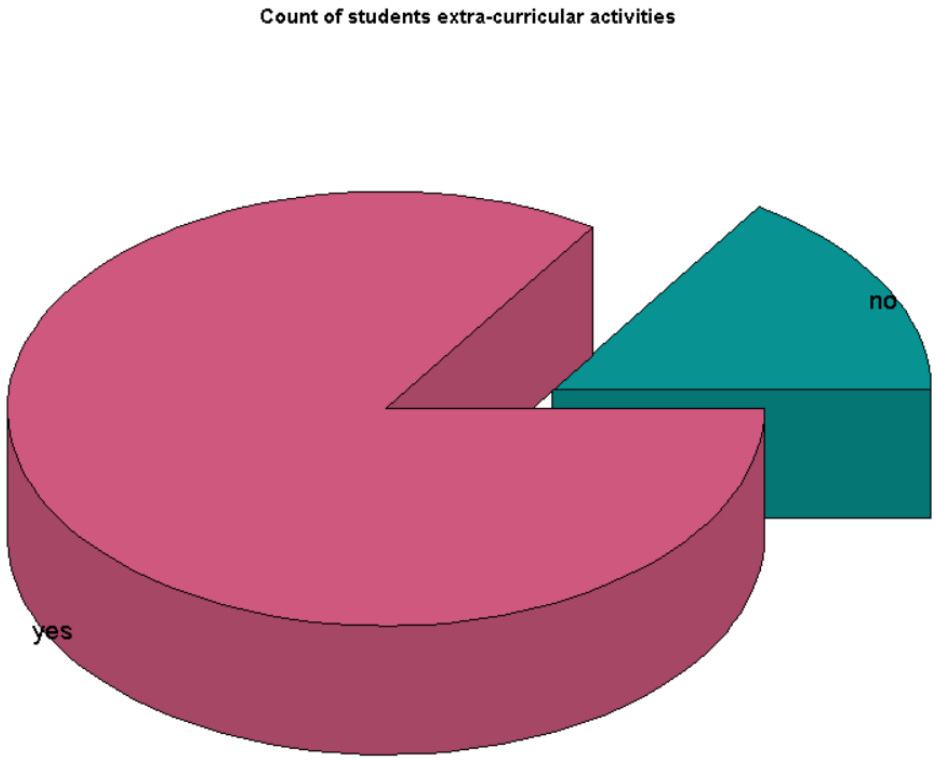


Figure 5.2.2 Count of student's extra-curricular activities.

From figure 5.2.2 shown us over 70% students are because over joining the extracurricular activities, cause they can't focus on study or do more time do to revision. In the end they failed in grade three.

Analysis 4.3 Does going out with friends often affect your grades?

```
379
380
381 #Analysis 4-3
382 Q4A3 = Q4 %>% group_by(goout) %>% summarise(Total=n())
383
384 ggplot(Q4A3, aes(x=goout, y=Total)) +
385   geom_bar(stat = "identity",width = 0.5,color="white",fill="orange")+
386   geom_text(aes(label=Total),position = position_dodge(1),vjust = -0.5) +
387   labs(title = "Count the number of times students and friends go out",
388        y = "Total number of students", x = "Number of time go out")
389
```

Figure 5.3.1 student going out with friends.

This analysis ensures that often going out with friends will affect a student's grade. In this also use same function, grouping and summaries.

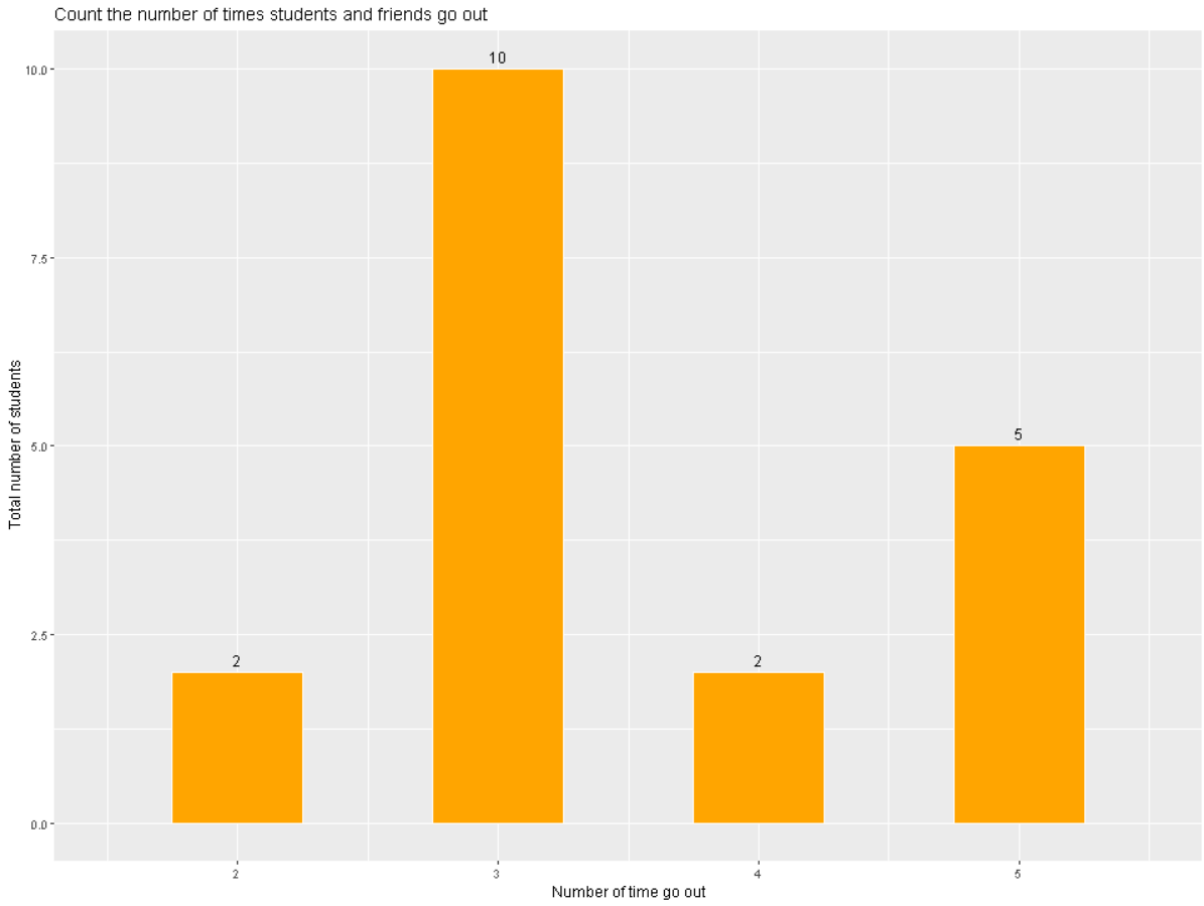


Figure 5.3.2 Count the number of times students and friends go out.

Based on above bar chart the number of time go out with friends actually are present the levels of number are from 1 until 5, means never until often go out with friends.

So we can see that starting from level 3 at the median, already 17 students are often going out with friends because they no longer focus on studying. There is a reason students fail in grade three.

Analysis 4.4 Does the amount of alcohol a student drinks affect grades.

```
392
393 #Analysis 4-4
394 Q4A4DA = Q4 %>% group_by(dalc) %>% summarise(TotalDaily = n())
395
396 ggplot(data = Q4A4DA, aes(dalc, TotalDaily)) +
397   geom_line(color = "steelblue", size = 1) +
398   geom_point(color="steelblue") +
399   geom_text(aes(label=TotalDaily),position = position_dodge(1),vjust = -0.5) +
400   labs(title = "Calculate daily alcohol consumption for students",
401        subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
402        y = "Count Of Students Number", x = "Daily Alcohol Consumption")
403
404
405
406 Q4A4WA = Q4 %>% group_by(walc) %>% summarise(TotalWeekly = n())
407
408 ggplot(data = Q4A4WA, aes(walc, TotalWeekly)) +
409   geom_line(color = "steelblue", size = 1) +
410   geom_point(color="steelblue") +
411   geom_text(aes(label=TotalWeekly),position = position_dodge(1),vjust = -0.5) +
412   labs(title = "Calculate weekly alcohol consumption for students",
413        subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
414        y = "Count Of Students Number", x = "Weekly Alcohol Consumption")
415
```

Figure 5.4.1 daily & weekly alcohol consumption.

From this analysis I want to verify it's daily or weekly alcohol consumption will affect student's grades.

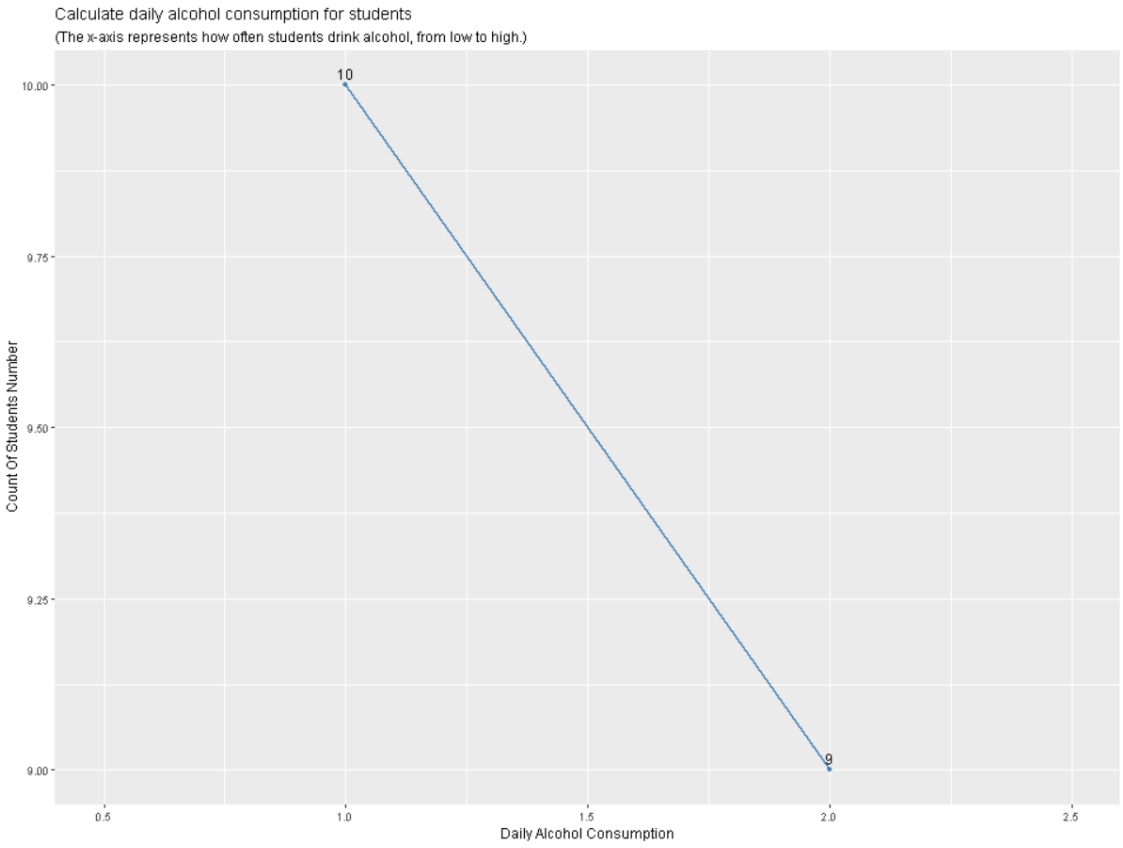


Figure 5.4.2 Calculate daily alcohol consumption for students.

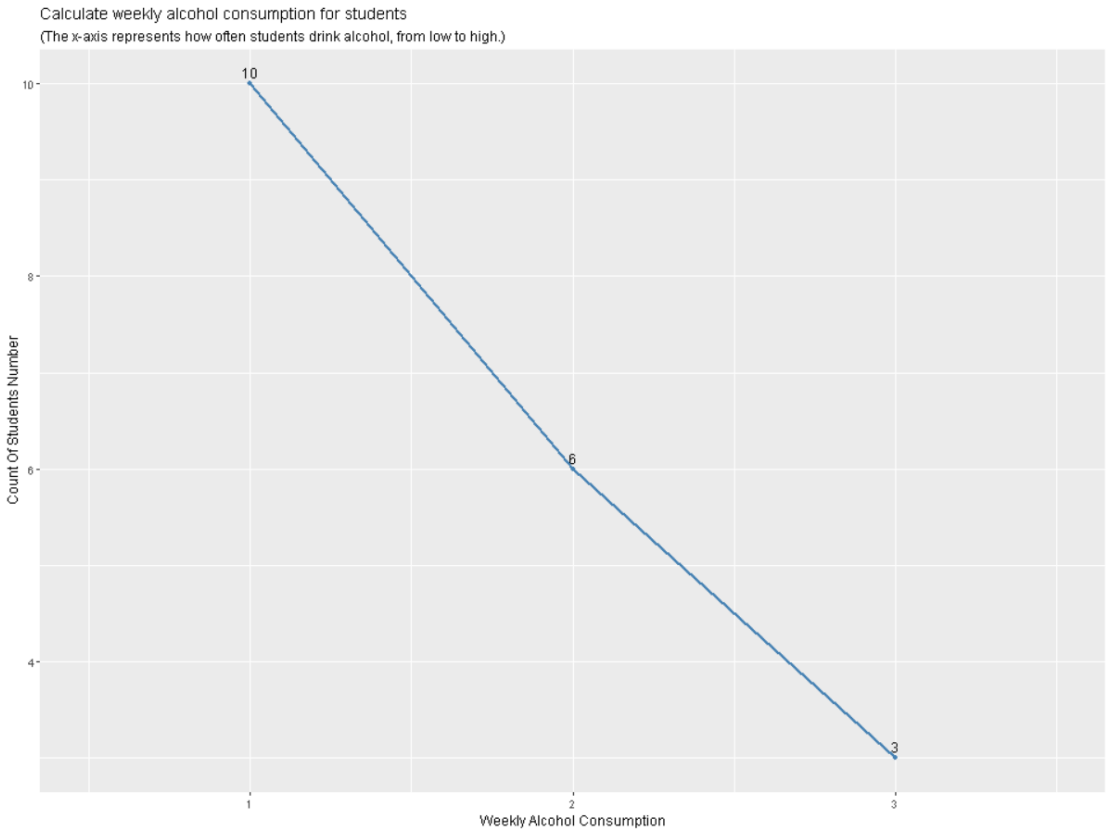


Figure 5.4.3 Calculate weekly alcohol consumption for students.

From the line graph in Figure 5.4.2 we can see that 10 students are in the daily drinking level of level 1, and there are 9 students in level 2.

Figure 5.4.3 Weekly Drinking Line Chart 10 students were at level 1, while the other 9 students were all drinking levels of level 2 or above.

In conclusion from these data, we know that daily and weekly alcohol consumption is more in the 1st category than in the other categories. But regardless of the amount of alcohol consumed daily or weekly, these students failed grade 3.

Analysis 4.5 Do students use their spare time to study to improve their grades.

```
418
419 #Analysis 4-5
420 Q4A5ST = Q4 %>% group_by(studytime) %>% summarise(CountStudyTime=n())
421 Q4A5FT = Q4 %>% group_by(freetime) %>% summarise(CountFreeTime = n())
422
423 Q4A5DF = data.frame(Q4A5ST, Q4A5FT)
424
425 ggplot(Q4A5DF, aes(y=CountStudyTime)) +
426   geom_line(aes(x = studytime), color = "darkred") +
427   geom_line(aes(x = freetime), color="steelblue", linetype="twodash")
428
```

Figure 5.5.1 study time and free time

In this algorithm, I will compare study time and free time. To check whether students are using free time to review their coursework.

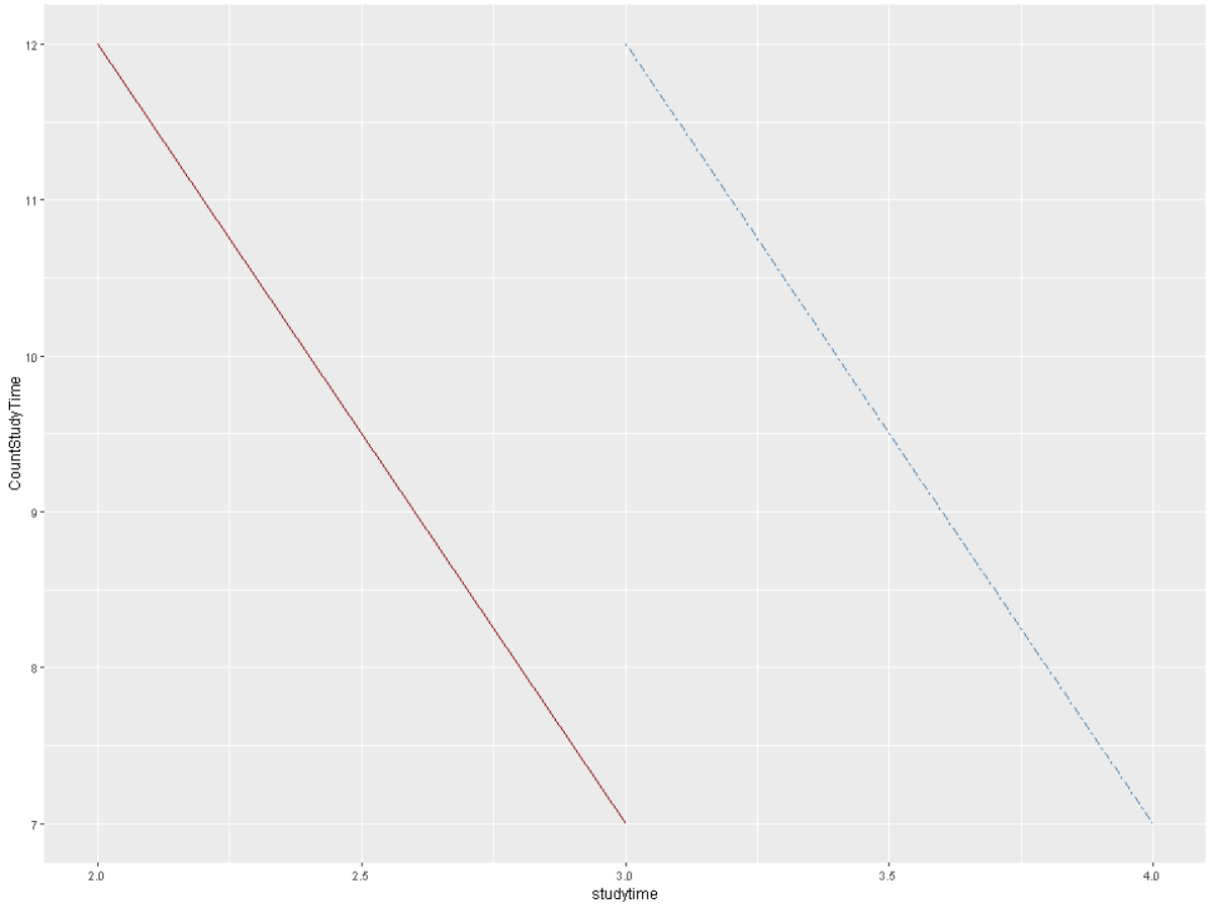


Figure 5.5.2 Line Chart Count Total of study time for students

From Figure 5.5.2 there are 2 lines, the dark red line represents the study time data, and the light blue dotted line represents the freetime. We see that this line is connected with 7 and 12, this is because according to the statistics, there are 12 students who have weekly study 2 hours in study time, and 7 students who have weekly study 3 hours. Then there are 12 students with 3 hours of freetime after school, and 7 students with 4 hours of freetime.

From the above data, we know that these students spend too little time for weekly study. On the contrary, they have a lot of free time after school and do not use this free time for study or revision.

Analysis 4.6 Is there any paid extra class that will affect my grades?

```
#Analysis 4-6
Q4A6 = Q4 %>% group_by(paid) %>% summarise(Count = n())

with(Q4A6,pie(Count,labels = paid, main = "Count of students paid extra class",
              col = hcl.colors(length(Count), "Purp"), clockwise = TRUE))

```

Figure 5.6.1 student paid extra class.

The above code I use to verify whether I need to pay extra class to get better grades.

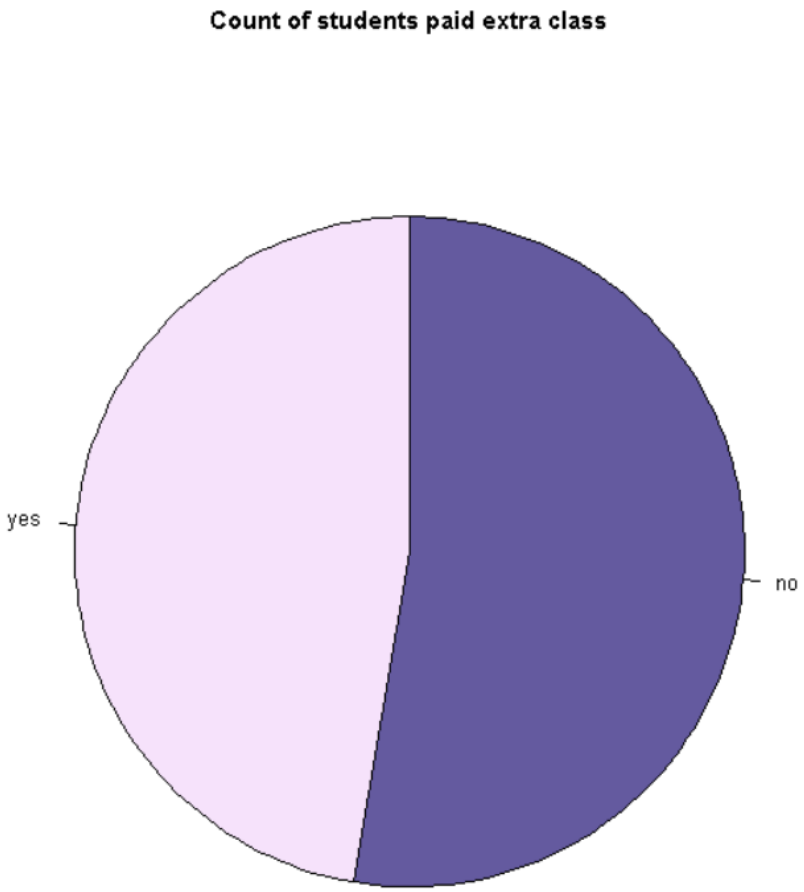


Figure 5.6.2 Count of students paid extra class.

From the results of the above pie chart, 52.63% of students do not have paid extra class, and it cannot be said that they can get good grades by paying extra class. But this could be one reason why these students failed grade 3.

Question 5: find out why G1, G2, G3 all fail?

```
444
445 #Question 5
446 Q5 = filter(student_data, g1&g2&g3 < 10)
447 nrow(Q5)
448
```

Figure 6.0 Question 5 Global Filter, find out which students fail on every grade.

Analysis 5.1 Students' long-term Internet access will affect their grades.

```
451
452 #Analysis 5-1
453 Q5A1Percent = round(Q5 %>% group_by(internet) %>% summarise(Count = n()) %>% summarise(Percent = Count/ sum(Count) * 100), digits = 2)
454 Q5A1 = Q5 %>% group_by(internet) %>% summarise(Count = n()) %>% mutate(Q5A1Percent)
455 Q5A1Lab = paste(Q5A1$internet, "\nTotal:", Q5A1$Count,"\n" ,Q5A1$Percent, "%")
456
457 with(Q5A1,pie3D(Count,labels = Q5A1Lab, main = "Count of students access inernet",
458               explode = 0.25, col = hcl.colors(length(Count), "Purp"), height = 0.17, labelcex = 1.4,
459               labelcol = "black", border="black", theta = 0.7))
460
```

Figure 6.1.1 Coding to filter out the internet will affect student's grades or not.

From this analysis, I want to find the percentages of students who access the internet. And in variable Q5A1Lab i use paste() function to store the labels message.

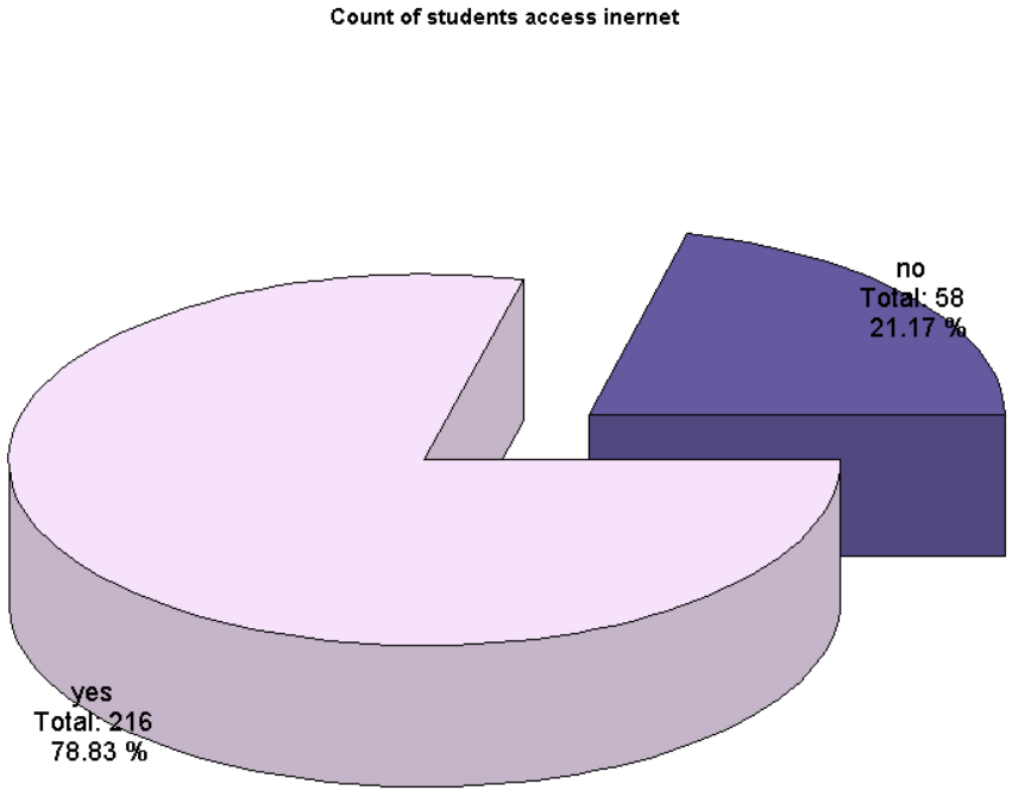


Figure 6.1.2 Count of students accessing the internet.

From Figure 6.1.2 shown total 216 students 78.83% have access to the internet, another 58 students 21.17 have no internet access. The data above looks like the results of these students

who failed in these grades were caused by long-term Internet access.Another 58 students failed because they did not have access to the Internet because of other factors.

Analysis 5.2 Students will fail every grade without paid extra class.

```
463
464 #Analysis 5-2
465 Q5A2Percent = round(Q5 %>% group_by(paid) %>% summarise(Count = n()) %>% summarise(Percent = Count/ sum(Count) * 100), digits = 2)
466 Q5A2 = Q5 %>% group_by(paid) %>% summarise(Count = n()) %>% mutate(Q5A2Percent)
467 Q5A2Lab = paste(Q5A2$paid, "\nTotal:", Q5A2$Count,"\n", Q5A2$Percent, "%")
468
469 with(Q5A2,pie(Count,labels = Q5A2Lab, main = "Count of students paid extra class",
470               col = hcl.colors(length(Count), "Greens"), clockwise = TRUE))
471
```

Figure 6.2.1 Count the percentages paid extra class and draw a pie chart.

This algorithm aims to count the percentages of students who got paid extra class, and also use paste() function to store the labels message. And the pie chart has no need to set the height, angle, theta anything, so I just set the clockwise to make the pie chart front on the median.

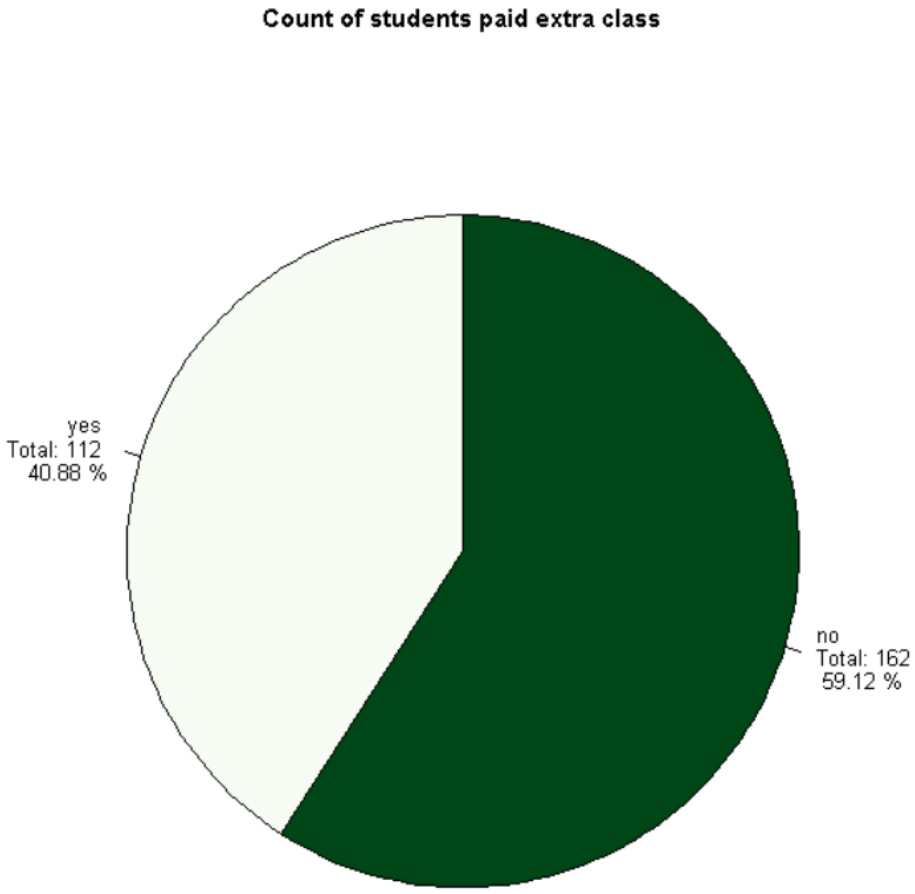


Figure 6.2.2 Count of students paid extra class.

From the results displayed in the pie chart, we can know that 59.12% of 162 students have no paid extra class, and 40.88% of 112 students have paid extra class.

These 162 students failed all grades because they did not have paid extra class. The other 112 students had paid extra class but still failed because they did not concentrate on their studies and other factors.

Analysis 5.3 Will students spend time reviewing, but going out with friends, will it affect their grades?

```
473
474 #Analysis 5-3
475 Q5A3ST = Q5 %>% group_by(studytime) %>% summarise(TotalStudyTime = n())
476 Q5A3GO = Q5 %>% group_by(goout) %>% summarise(TotalGoOut = n())
477
478 Q5A3ST[nrow(Q5A3ST)+1, ] <- 5
479
480 Q5A3DF = data.frame(Q5A3ST, Q5A3GO)
481 Q5A3DF[5,2] = 0
482
483 ggplot(Q5A3DF, aes(x=studytime)) +
484   geom_line(aes(y = TotalStudyTime), color = "darkred") +
485   geom_line(aes(y = TotalGoOut), color="steelblue", linetype="twodash")
486
```

Figure 6.3.1 Filter the data and compare study time and go out.

From Figure 6.3.1 I want to compare whether students spend time studying or going out with friends. Since going out with friends has one more row of data than study time, I can't make a graph. So I added a row of 5 data to the study time on row 478. After replacing the data with 0 and merging the data into the data frame, use the data frame to draw a line chart.

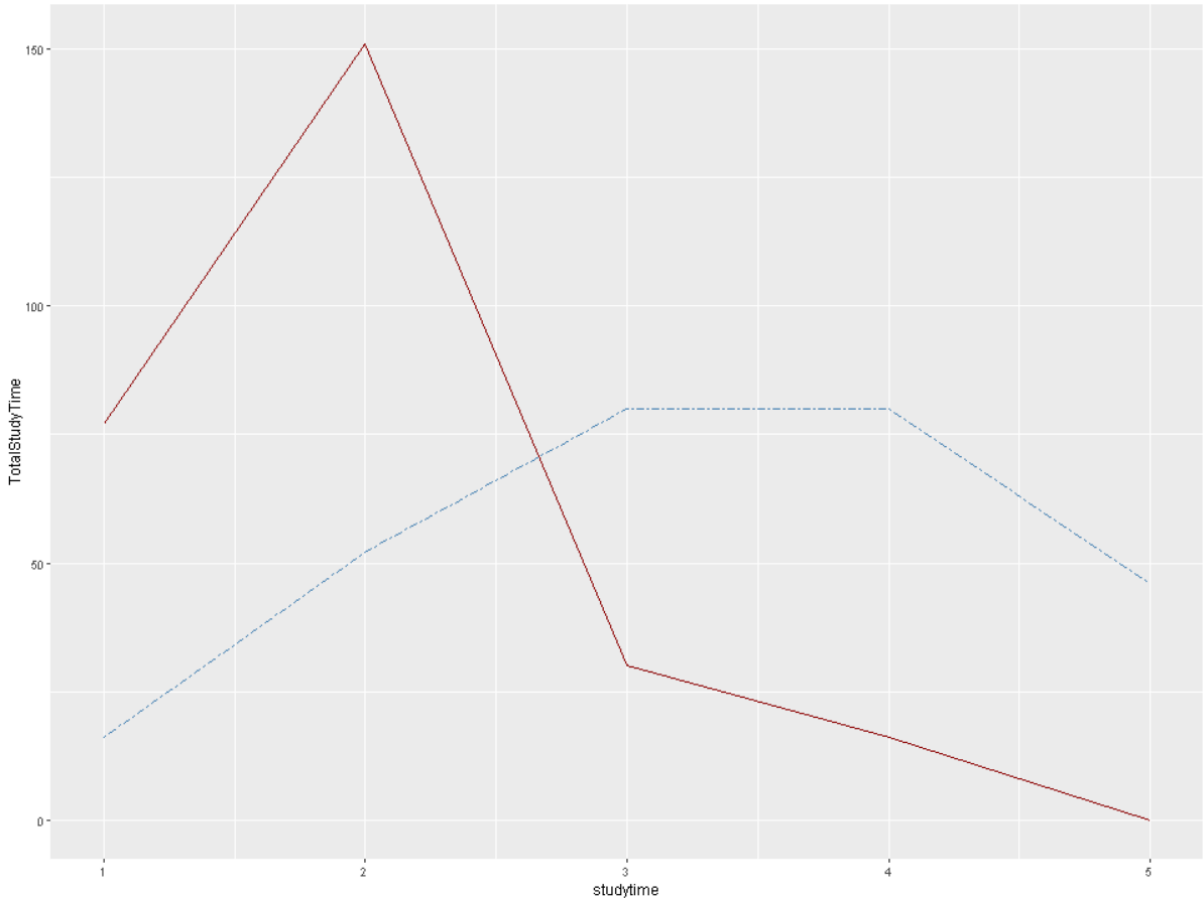


Figure 6.3.2 Line Chart Total study time and go out with friends.

490:14 (Top Level)

Console Terminal Jobs

R 4.1.2 · ~/

```
> Q5A3DF
  studytime TotalStudyTime goout TotalGoOut
1         1             77     1         16
2         2            151     2         52
3         3             30     3         80
4         4             16     4         80
5         5              0     5         46
>
```

Figure 6.3.3 Line Chart performs the data table.

From figure 6.3.2, there are 2 line charts, the darkred straight line represents the Total study time, and the light blue dotted line represents the Total goout. There is a clear trend and we can see that the red line keeps going down, while the light blue dotted line goes down until level 3 & 4.

We were able to learn that the number of students who spent no more than 3 hours in weekly study time was as many as 228 students. At level 3, the number of people who went out with friends for a long time was as high as 206. That's why these students fail the entire grade because they don't spend much time in study time. Instead, spend more time going out with friends.

Then figure 6.3.3 is the data executed by the above line chart. The last row index 5 of the study time was added manually so that this set of data could be the same as another row index.

Analysis 5.4 If the family size is large, will the lack of parental consideration affect the grades?

```
488
489 #Analysis 5-4
490 Q5A4Percent =round(Q5 %>% group_by(famsize) %>% summarise(TotalFamilySize = n()) %>%
491                      summarise(Percent = TotalFamilySize/ sum(TotalFamilySize) * 100), digits = 2)
492 Q5A4 = Q5 %>% group_by(famsize) %>% summarise(TotalFamilySize = n()) %>% mutate(Q5A4Percent)
493 Q5A4Lab = paste(Q5A4$famsize, "\nTotal:", Q5A4$TotalFamilySize,"\n", Q5A4$Percent, "%")
494
495 with(Q5A4,pie3D(Percent,labels = Q5A4Lab, main = "Count of students's family size",
496               explode = 0.4, col = hcl.colors(length(TotalFamilySize), "Greens"), height = 0.17, labelcex = 1.6,
497               labelcol = "red", border="black", theta = 0.9))
498
499
```

Figure 6.4.1 Calculate percentages of family size.

From this analysis, I want to verify that its family size is too large. Parents will lack consideration to affect students' grades. I calculate the percentages of a student's family size, after drawing the pie 3D.

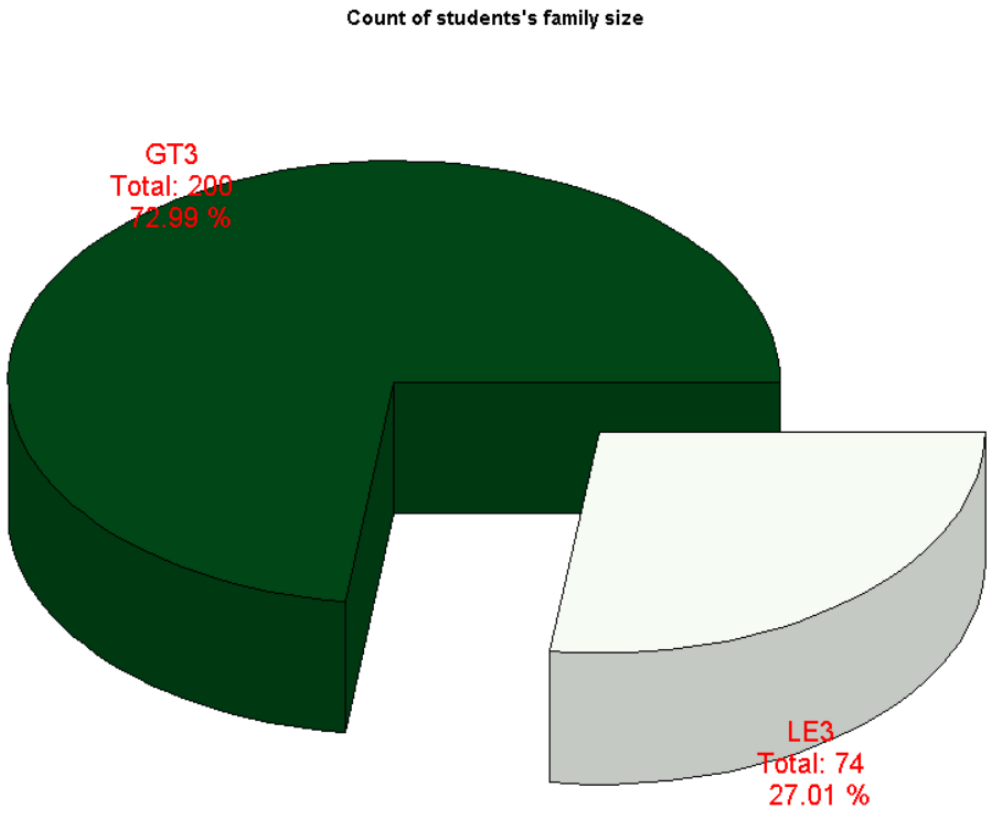


Figure 6.4.2 Count of student’s family size.

Figure 6.4.2 shows two different values of data, green color present GT3 means the family size is greater than 3 people, and white color present LE3 means the family size is less or equal to 3. GT3 got 200 students 72.99%, another LE3 74 students 27.01%.

From the above data, we know that most student families have more than 3 members, which often results in parents not being able to focus on educating their children. However, the other 27.01% of students' family members are equal to or less than 3, and they still fail, which requires other factors to be considered.

Analysis 5.5 Will student alcohol consumption affect grades?

```
501
502 #Analysis 5-5
503 Q5A5DA = Q5 %>% group_by(dalc) %>% summarise(TotalDaily = n())
504
505 ggplot(data = Q5A5DA, aes(dalc, TotalDaily)) +
506   geom_line(color = "steelblue", size = 1) +
507   geom_point(color="steelblue") +
508   geom_text(aes(label=TotalDaily),position = position_dodge(1),vjust = -0.5) +
509   labs(title = "Calculate daily alcohol consumption for students",
510        subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
511        y = "Count Of Students Number", x = "Daily Alcohol Consumption")
512
513
514
515 Q5A5WA = Q5 %>% group_by(walc) %>% summarise(TotalWeekly = n())
516
517 ggplot(data = Q5A5WA, aes(walc, TotalWeekly)) +
518   geom_line(color = "steelblue", size = 1) +
519   geom_point(color="steelblue") +
520   geom_text(aes(label=TotalWeekly),position = position_dodge(1),vjust = -0.5) +
521   labs(title = "Calculate weekly alcohol consumption for students",
522        subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
523        y = "Count Of Students Number", x = "Weekly Alcohol Consumption")
524
```

Figure 6.5.1 Calculate daily and weekly student’s alcohol consumption.

This analysis of alcohol consumption will fail every grade. I use a group and summarise function to filter and count the data by daily and weekly.

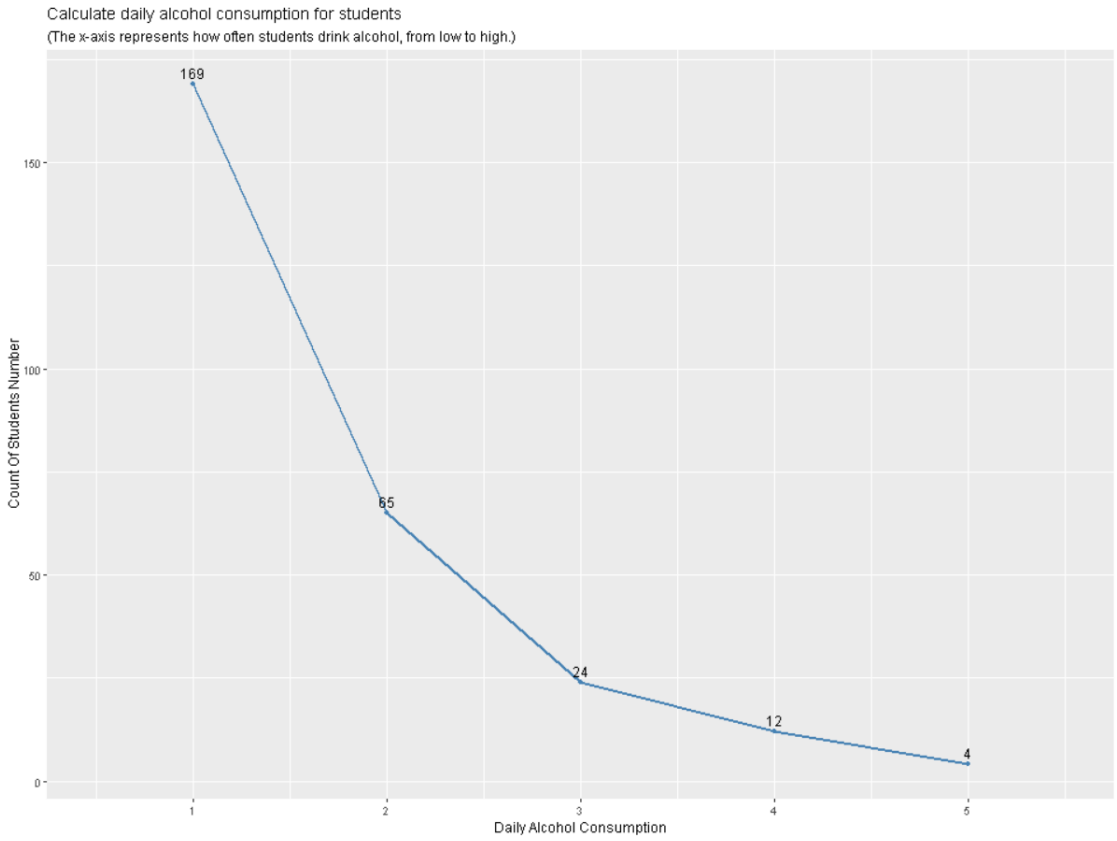


Figure 6.5.2 Calculate daily alcohol consumption for students.

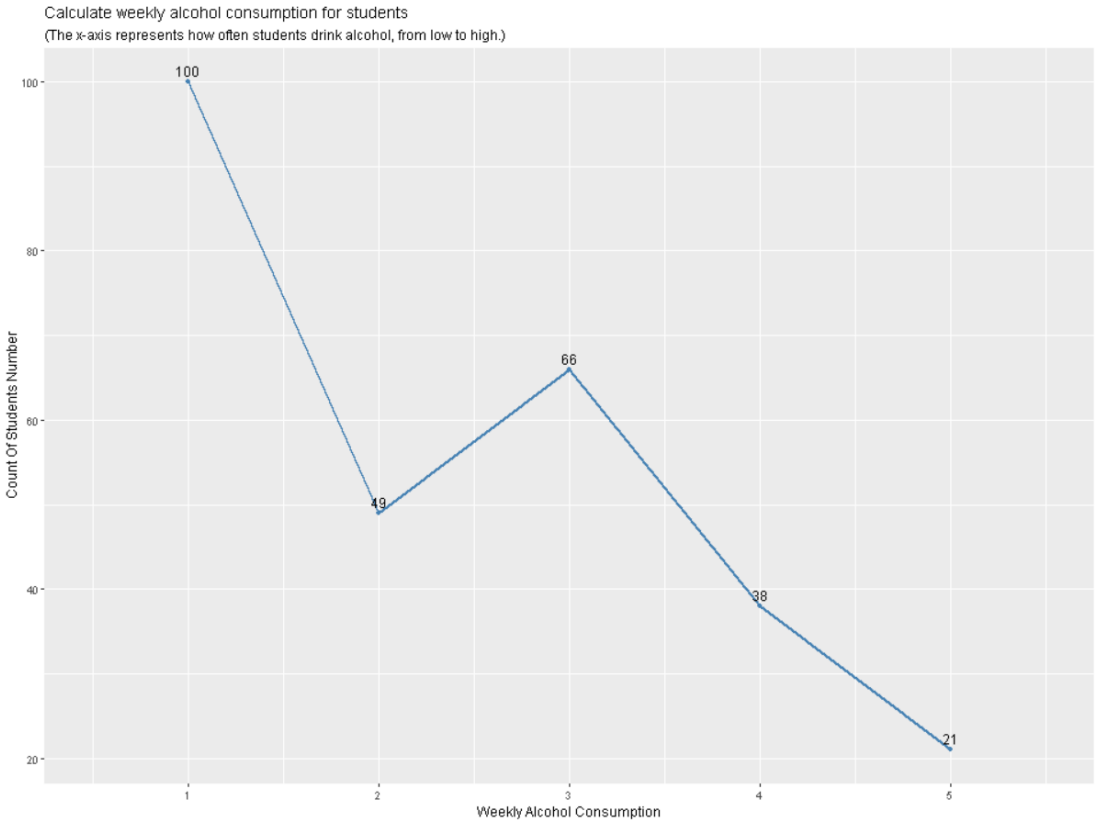


Figure 6.5.3 Calculate weekly alcohol consumption for students.

As can be seen from the line graph in Figure 6.5.2, 169 students maintained their daily alcohol consumption at level 1, which means that their daily alcohol consumption was not significant. Another 105 students drink more than 2 levels a day. Means they may have a drinking habit.

Figure 6.5.3 Weekly Drinking Line Chart We can see that 100 students were at level 1 and another 174 were drinking at level 2 or above, the higher the alcohol intake, the more.

From this data, we can deduce why students are failing grades because most of these students drink more frequently on weekends. There is no extra time and energy to focus on studying.

Analysis 5.6 Does the parent's educational background affect the child's grades?

```
528
529 #Analysis 5-6
530 Q5A6FE = Q5 %>% group_by(fedu) %>% summarise(TotalFatherEducationLevel = n())
531 Q5A6ME = Q5 %>% group_by(medu) %>% summarise(TotalMotherEducationLevel = n())
532
533 Q5A6FE[nrow(Q5A6FE)+1, ] <- 0
534
535 Q5A6DF = data.frame(Q5A6FE, Q5A6ME)
536 Q5A6Melt = melt(Q5A6DF[,c("medu","TotalMotherEducationLevel","TotalFatherEducationLevel")], id.vars=1)
537
538 ggplot(Q5A6Melt, aes(x=medu, y=value, fill=variable)) +
539   geom_bar(stat="identity", position="dodge") +
540   geom_text(aes(label=value),position = position_dodge(1),vjust = -0.5) +
541   labs(title = "Count Of Parent Education Level", y = "Count", x = "Parent Education Level")
542
543
```

Figure 6.6.1 Group and Calculate parent education background.

Does the analysis verify that students' parents' educational backgrounds affect them? First, I group the data and aggregate educational background. and store it into a dataframe so that I can use the melt() function to reshape the data.

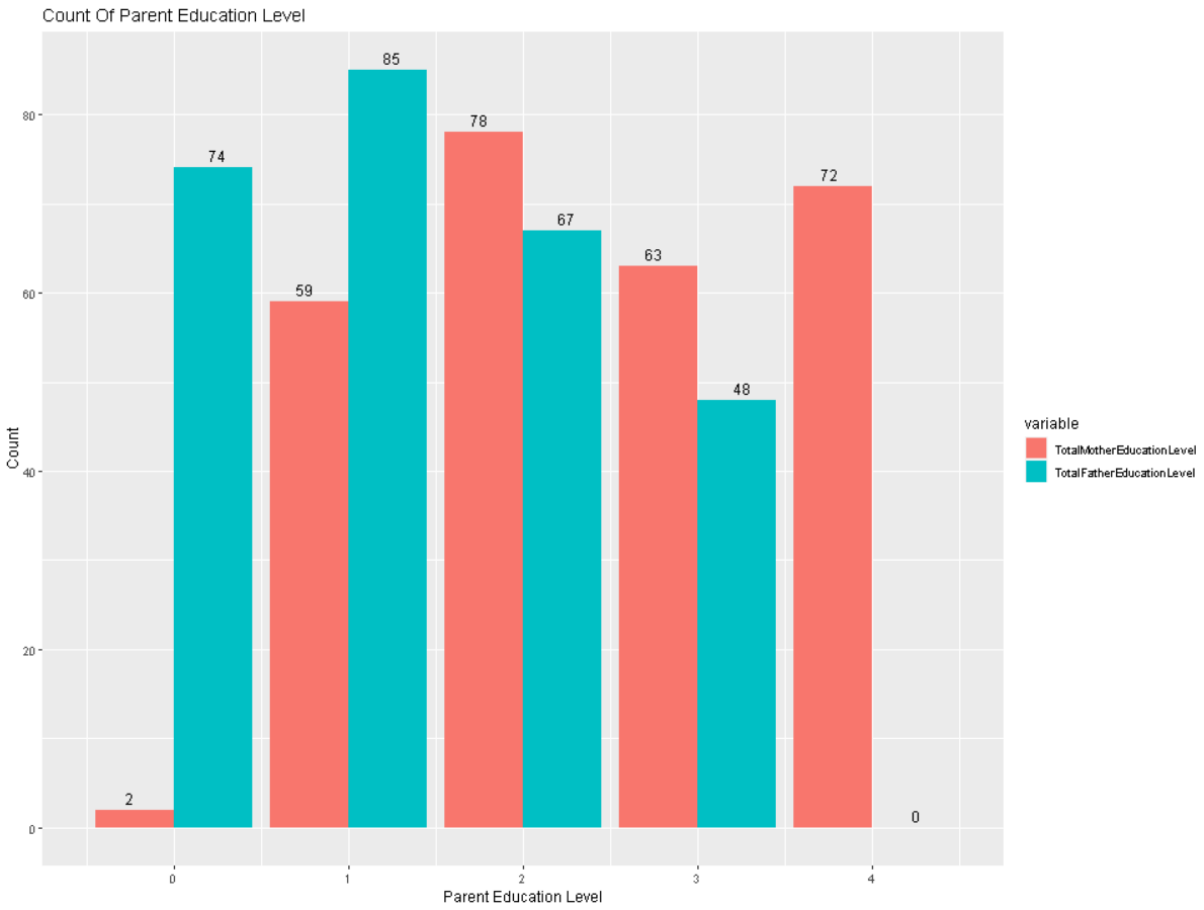


Figure 6.6.2 Count of parent education level.

The bar chart from Figure 6.6.2 shows that red represents total mother education, while blue represents total father education. Parent Education Level represents their respective highest educational level. 0 means no educational background, 1 is primary education (4th grade), 2 is 5th to 9th grade, 3 is secondary education and 4 is higher education.

From the above data, it can be seen that the 0-2 stage is a range of the father education level, and the 2-4 stage is a range of the mother education level. So we can know that these students who failed in all grades are influenced by their father's educational background. Just because the father's education background is relatively low, the highest father education level is in primary school. So that's one reason students fail all grades because fathers can't help their kids with schoolwork.

Analysis 5.7 If schools give students school education support, will students improve their grades?

```
545
546 #Analysis 5-7
547 Q5A7Percent =round(Q5 %>% group_by(schoolsup) %>% summarise(TotalSchoolSupport = n()) %>%
548                   summarise(Percent = TotalSchoolSupport/ sum(TotalSchoolSupport) * 100), digits = 2)
549 Q5A7 = Q5 %>% group_by(schoolsup) %>% summarise(TotalSchoolSupport = n()) %>% mutate(Q5A7Percent)
550 Q5A7Lab = paste(Q5A7$schoolsup, "\nTotal:", Q5A7$TotalSchoolSupport,"\n", Q5A7$Percent, "%")
551
552 with(Q5A7,pie3D(Percent,labels = Q5A7Lab, main = "Percentages of student got school education support",
553                 explode = 0.4, col = hcl.colors(length(TotalSchoolSupport), "OrYel"), height = 0.17, labelcex = 1.6,
554                 labelcol = "black", border="black", theta = 0.9))
555
```

Figure 6.7.1 calculates percentages of school education support.

In this algorithm to calculate percentages if a school is given student education support or not. I used round() function to remove all small points, and used paste() function to store labels messages to allow me to put in pie charts.

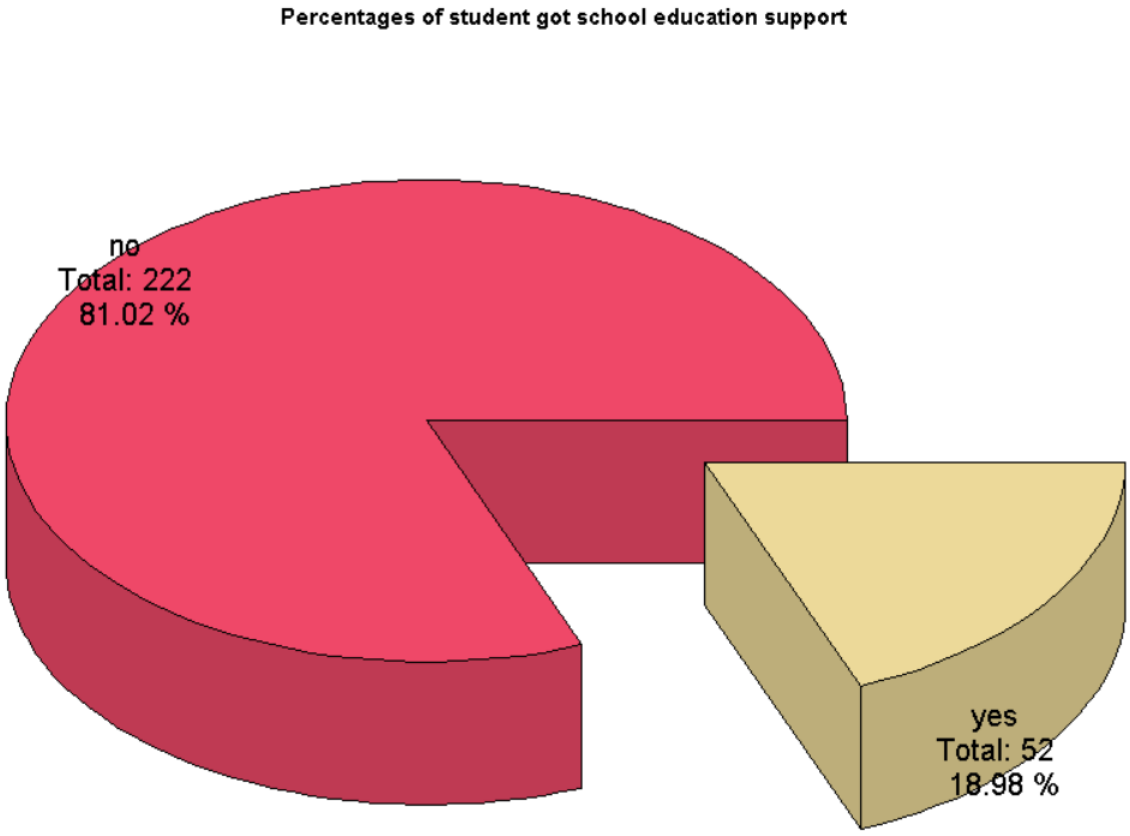


Figure 6.7.2 Percentages of students got school education support.

FFrom figure 6.7.2 shows that about 81.02% of the 222 students did not receive school education support, while only 52 students with 18.98% received school education support. When students do not absorb knowledge in the classroom, they can only review through self-study, paid extra class, or school education support. It's hard to improve grades when students don't even get school education support. This is not the only factor that affects students' grades and failing grades, but it is also one of the factors.

Question 6. How important is home discipline to a student's grades?

```
561
562 #Question 6
563 Q6 = filter(student_data, g1&g2&g3 > 10)
564 nrow(Q6) #491
565
```

Figure 7.0 Question 6 Global Filter will filter the students who pass every grade.

Analysis 6.1 Where do students prefer to choose a school?

```
567
568 #Analysis 6-1
569 Q6A1 = Q6 %>% group_by(address, reason) %>% summarise(Total= n(), .groups = 'drop')
570
571 ggplot(Q6A1, aes(x=address, y=Total, fill=reason)) +
572   geom_bar(stat="identity", position="dodge") +
573   geom_text(aes(label=Total),position = position_dodge(0.9),vjust = -0.5)
574
575
```

I want to find the reason for students preferring to choose a school, so I group the address and reason to summarize the result. After that drop the data to the bar chart.

Figure 7.1.1 group address & reason and total each reason.

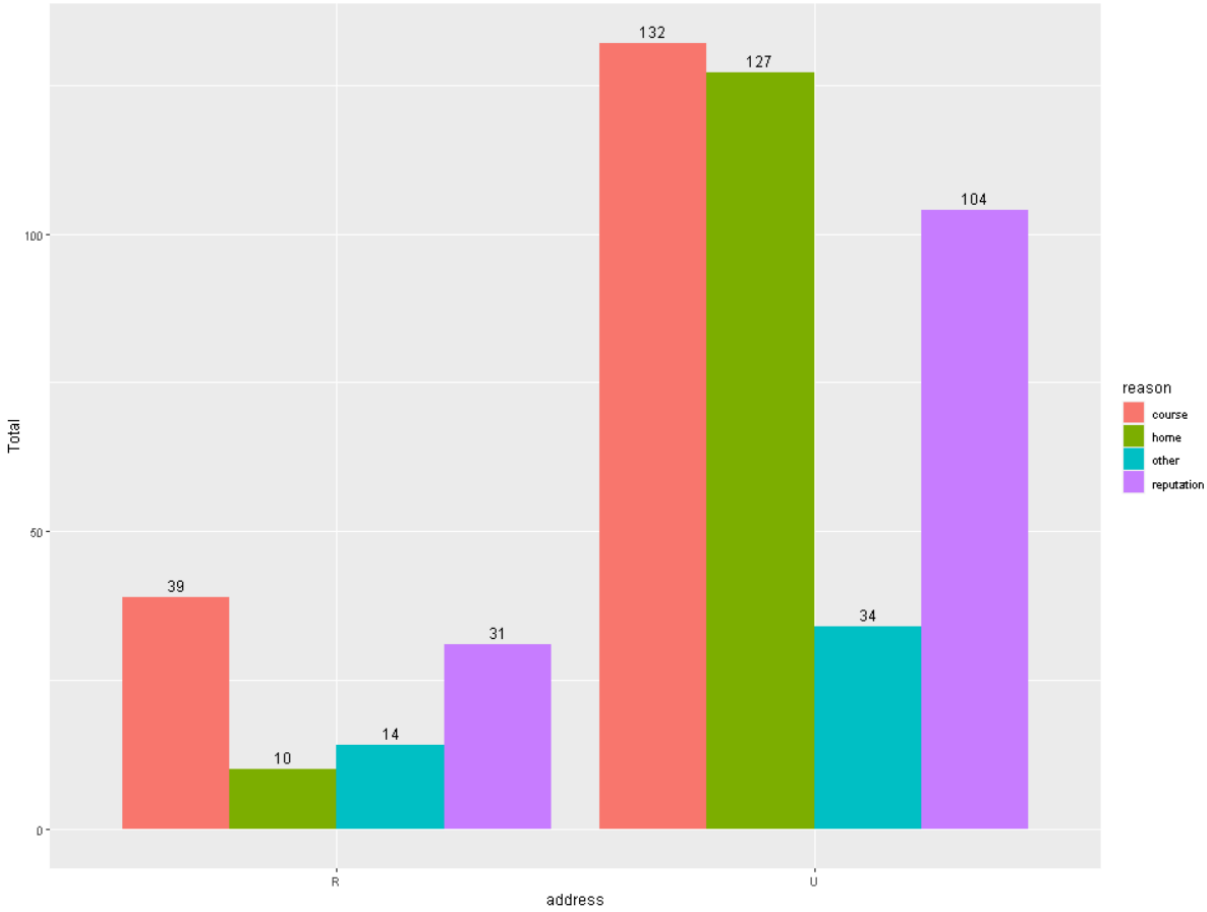


Figure 7.1.2 Count each address the total of chosen school reasons.

The bar chart shows different color columns, each color represents a different reason for choosing this school. And U for urban and r for rural where address is displayed below.

Regardless of urban or rural, we can clearly see that the students who choose the course as the reason are far ahead. Secondly, the reason for the rural is the reason for the reputation, and the reason for the urban is the reason for the home.

Course and home are balanced, some parents want their children to choose subjects they like, and some parents want their children to choose schools in a way that is close to home.

Students who choose Course and Reputation get good grades.

Analysis 6.2 Parents are absolutely opposed to a student's romantic relationship, isn't it?

```
576 #Analysis 6-2
577
578 Q6A2Percent =round(Q6 %>% group_by(romantic) %>% summarise(Total = n()) %>%  summarise(Percent = Total/ sum(Total) * 100), digits = 2)
579 Q6A2 = Q5 %>% group_by(romantic) %>% summarise(Total = n()) %>% mutate(Q6A2Percent)
580 Q6A2Lab = paste(Q6A2$romantic, "\nTotal:", Q6A2$Total,"\n", Q6A2$Percent, "%")
581
582 with(Q6A2,pie(Percent,labels = Q6A2Lab, main = "Total Percentages of student got romantic relationship",
583                 col = hcl.colors(length(Total), "Peach"), clockwise = TRUE))
584
```

Figure 7.2.1 This algorithm to calculate percentages of romantic and total number of students.

I will verify if the parental family discipline does not approve of the student having romantic relationships. Calculate the percentages of romantic relationship status used round() functions.

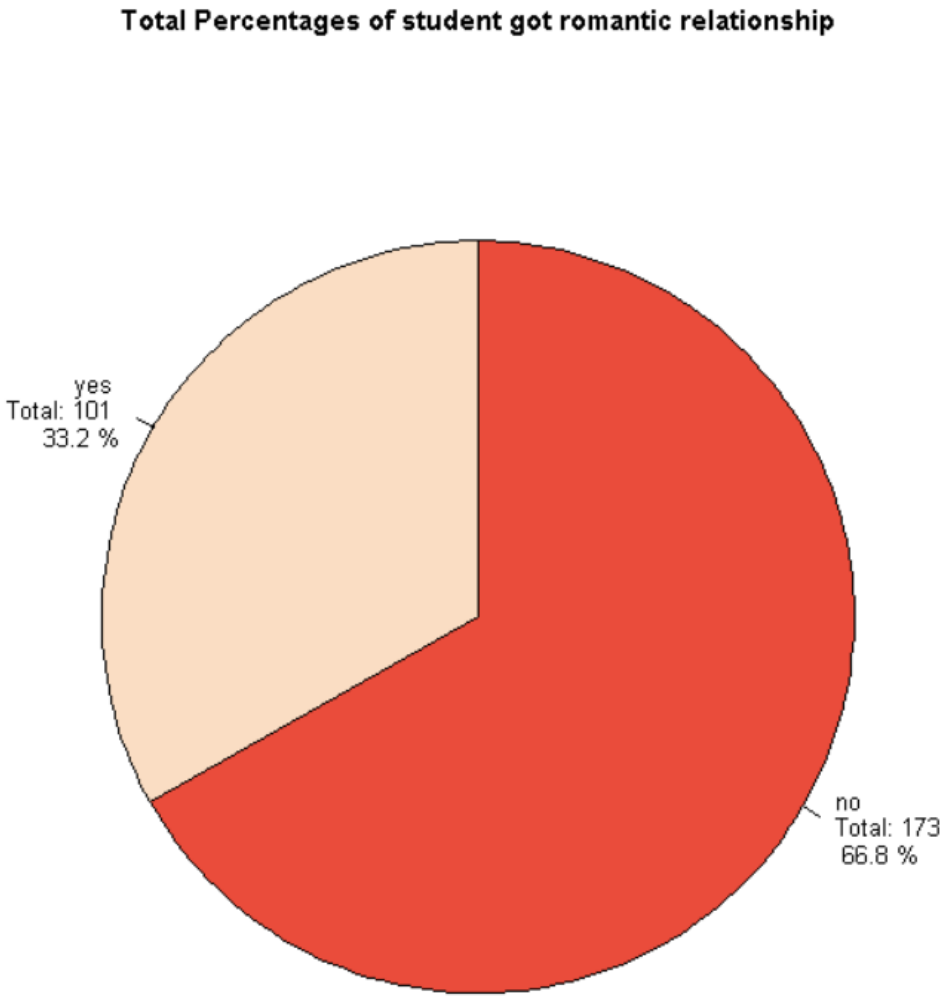


Figure 7.2.2 Total percentages of students got romantic relationships.

According to the data shown in the pie chart above, it is confirmed that most parents do not encourage students to have romantic relationships, and 66.8% of 173 students do not encourage students to have romantic relationships. Only 101 33.1% of students are romantic. Many parents sternly warn their children against romanticism on the way to school because it affects their grades and prevents them from focusing on schoolwork.

Analysis 6.3 Will parents strictly prohibit children and students from drinking alcohol?

```
587
588 #Analysis 6-3
589 Q6A3DA = Q6 %>% group_by(dalc) %>% summarise(TotalDaily = n())
590
591 ggplot(data = Q6A3DA, aes(dalc, TotalDaily)) +
592   geom_line(color = "steelblue", size = 1) +
593   geom_point(color="steelblue") +
594   geom_text(aes(label=TotalDaily),position = position_dodge(1),vjust = -0.5) +
595   labs(title = "Calculate daily alcohol consumption for students",
596        subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
597        y = "Count Of Students Number", x = "Daily Alcohol Consumption")
598
599
600
601 Q6A3WA = Q6 %>% group_by(walc) %>% summarise(TotalWeekly = n())
602
603 ggplot(data = Q6A3WA, aes(walc, TotalWeekly)) +
604   geom_line(color = "steelblue", size = 1) +
605   geom_point(color="steelblue") +
606   geom_text(aes(label=TotalWeekly),position = position_dodge(1),vjust = -0.5) +
607   labs(title = "Calculate weekly alcohol consumption for students",
608        subtitle = "(The x-axis represents how often students drink alcohol, from low to high.)",
609        y = "Count Of Students Number", x = "Weekly Alcohol Consumption")
610
611
```

Figure 7.3.1 I will divide 2 sections and 2 charts, daily and weekly. Both charts are displayed using Line charts.

This analysis mainly analyses whether parents will object to their children's drinking, and the amount of drinking they can accept. In the end, they thought that drinking alcohol could affect their studies.

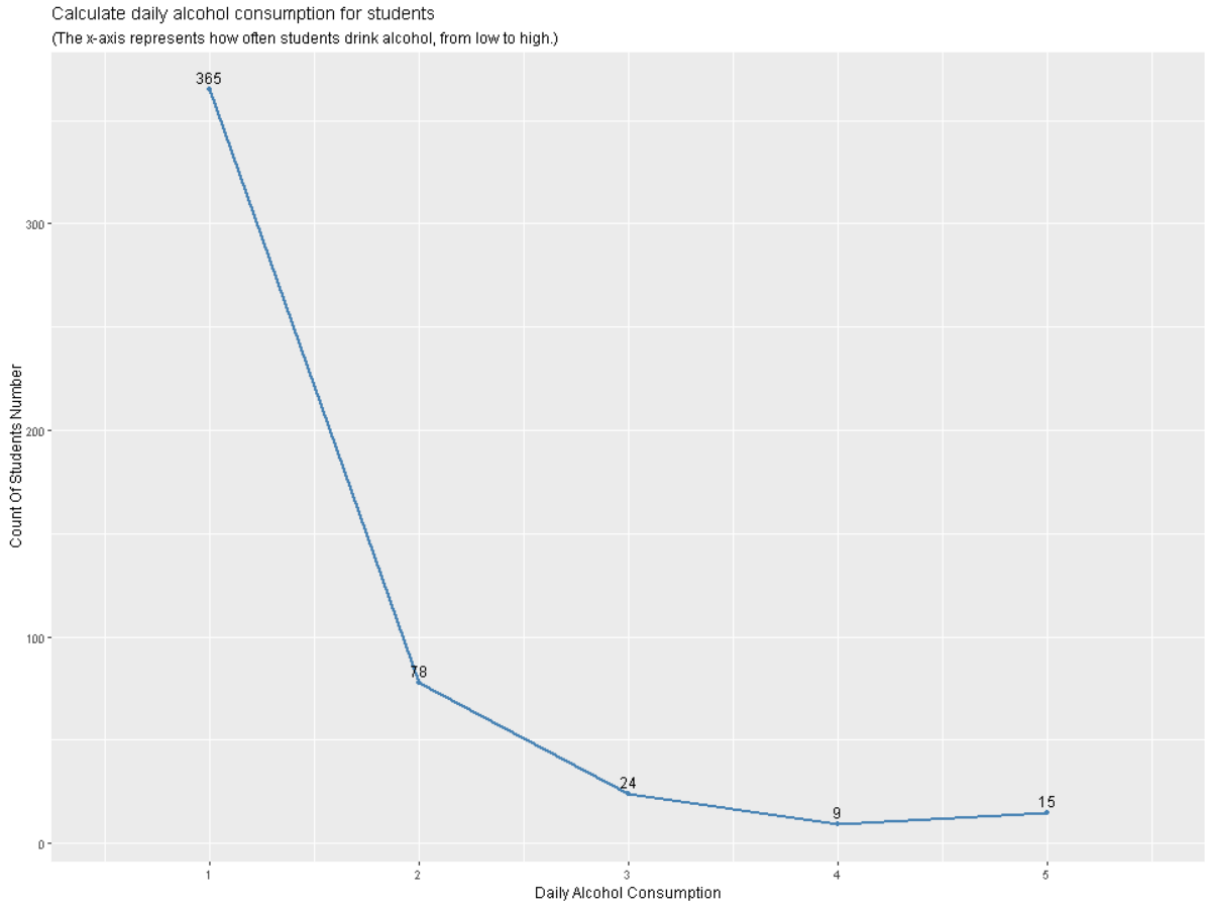


Figure 7.3.2 Calculate daily alcohol consumption for students.

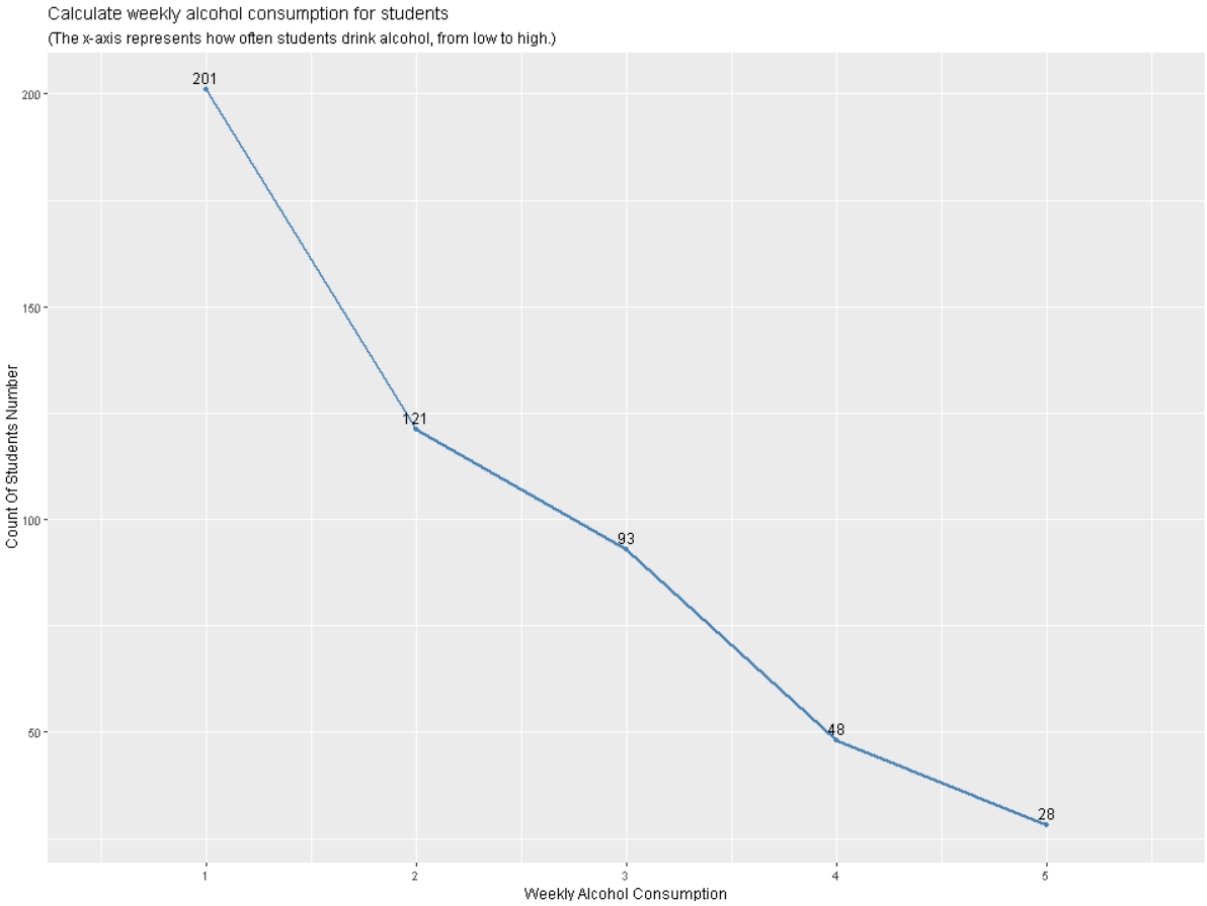


Figure 7.3.3 Calculate weekly alcohol consumption for students.

First of all, from the line chart figure 7.3.2 of daily alcohol, the most students are grade 1, with a total of 365 students. The number of alcohol consumption in x-axis represents the grade, from 1 to 5, from low to high alcohol consumption. The remaining 126 students are more than grade 2 and above.

Next is the weekly alcohol line chart figure 7.3.3. The top 2 students are at most grades 1 and 2. Only 168 students are above grade 3.

From the obtained data, we found that parents have strict control over daily alcohol consumption, while weekly alcohol consumption is more relaxed. This type of discipline helps the child's grades.

Analysis 6.4 Parents strictly control that their children must go to school and will not miss school.

```
612
613 #Analysis 6-4
614 Q6A4 = Q6 %>% group_by(absences) %>% mutate(rangeOfAbsences = cut(absences,c(-1,10,20,30,40,50,Inf))) %>%
615   group_by(rangeOfAbsences) %>% summarise(TotalRange = n())
616
617 ggplot(Q6A4, aes(x=rangeOfAbsences, y=TotalRange, fill=rangeOfAbsences)) +
618   geom_bar(stat="identity", position="dodge") +
619   geom_text(aes(label=TotalRange),position = position_dodge(1),vjust = -0.5)
620
```

Figure 7.4.1 This analysis I want to find out if parents will strictly control their children or not. And because there are so many different types of absences, I will make a range of absences using cut() functions.

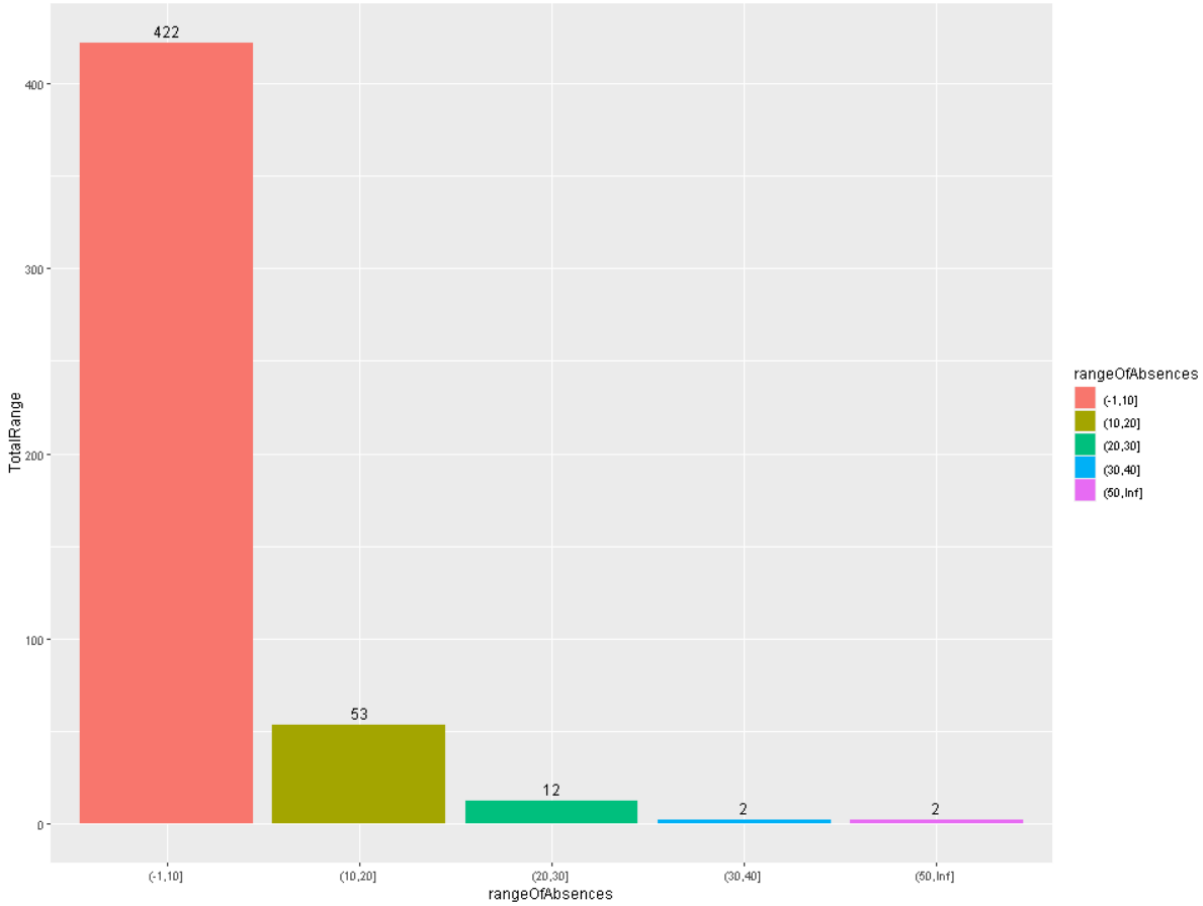


Figure 7.4.2 Total of absences times by range of absences.

From the bar chart, it can be seen that the number of absences of 422 students is 0-10 times, and the second most is 10-20 times of 53 students. The remaining few students have more than 20 absences or more.

Parents are also very concerned about the attendance rate of students and often monitor whether their children have absences. And absence is also one of the factors that affect students' grades, because students with absences miss any knowledge.

Analysis 6.5 Parents' educational background can be a driving force to encourage children.

```
623
624 #Analysis 6-5
625 Q6A5FE = Q6 %>% group_by(fedu) %>% summarise(TotalFatherEducationLevel = n())
626 Q6A5 = Q6 %>% group_by(medu) %>% summarise(TotalMotherEducationLevel = n()) %>% mutate(Q6A5FE)
627
628
629 Q6A5DF = data.frame(Q6A5)
630 Q6A5Melt = melt(Q6A5DF[,c("medu","TotalMotherEducationLevel","TotalFatherEducationLevel")], id.vars=1)
631
632  ggplot(Q6A5Melt, aes(x=medu, y=value, fill=variable)) +
633   geom_bar(stat="identity", position="dodge") +
634   geom_text(aes(label=value),position = position_dodge(1),vjust = -0.5) +
635   labs(title = "Count Of Parent Education Level", y = "Count", x = "Parent Education Level")
636
637
```

Figure 7.5.1 Calculate Parent education background level.

This analysis is to verify that parents' educational background has a lot of driving force or influence on children. I use the group function to find the data of each stage of the father and mother for statistics, and then put these two data together into the data frame. Since there are 4 columns, I reshape the father and mother columns into parent columns.

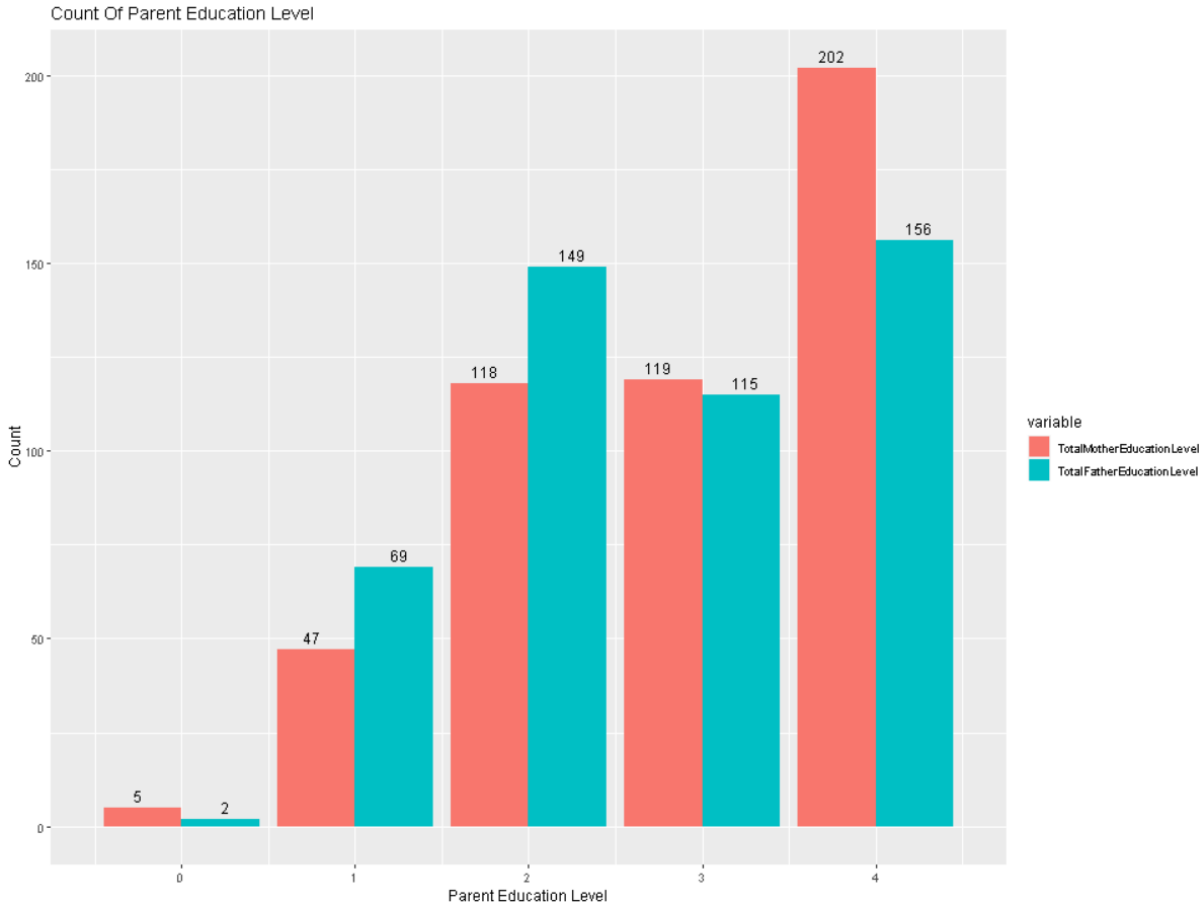


Figure 7.5.2 Count of Parent Education Level.

From Figure 7.5.2 shows that red represents total mother education, while blue represents total father education. Parent Education Level represents their respective highest educational level. 0 means no educational background, 1 is primary education (4th grade), 2 is 5th to 9th grade, 3 is secondary education and 4 is higher education.

From the above data, we can see that the 2-4 stage is the range of the father's educational level, and the 2-4 stage is the range of the mother's educational level. So we can know that parental education has obvious impetus for children, these students all pass, and their parents have a certain and very high educational background.

Addition Features – Melt() & facet_wrap()

```
656
657 #Extra Features
658
659
660 #separate the range
661 RangeOfGradeG3 = student_data %>% group_by(g3) %>% mutate(rangeOfG3 = cut(g3,c(-1, 5, 10, 15, 20))) %>% group_by(rangeOfG3) %>% summarise(TotalG3 = n())
662 RangeOfGradeG2 = student_data %>% group_by(g2) %>% mutate(rangeOfG2 = cut(g2,c(-1, 5, 10, 15, 20))) %>% group_by(rangeOfG2) %>% summarise(TotalG2 = n())
663 RangeOfGradeG1 = student_data %>% group_by(g1) %>% mutate(rangeOfGrade = cut(g1,c(-1, 5, 10, 15, 20))) %>% group_by(rangeOfGrade) %>% summarise(TotalG1 = n())
664
665
666 #store in data frame and reshape the data
667 RangeOfGrade = data.frame(RangeOfGradeG1,RangeOfGradeG2,RangeOfGradeG3)
668 RangeOfGradeMelt = melt(RangeOfGrade, id = c("rangeOfGrade","rangeOfG2","rangeOfG3"), value.name = "Total")
669
670
671 #draw multiple line in wrap function, include line & point & label text
672 ggplot(data = RangeOfGradeMelt, aes(x = rangeOfGrade, y = Total, group = 3, col = variable)) +
673   geom_line() +
674   geom_point() +
675   geom_text(aes(label=Total),position = position_dodge(1),vjust = -0.5) +
676   facet_wrap(facets = vars(variable)) +
677   theme_bw()
678
```

Figure 8.1 Addition Features to display every grade of student's score range.

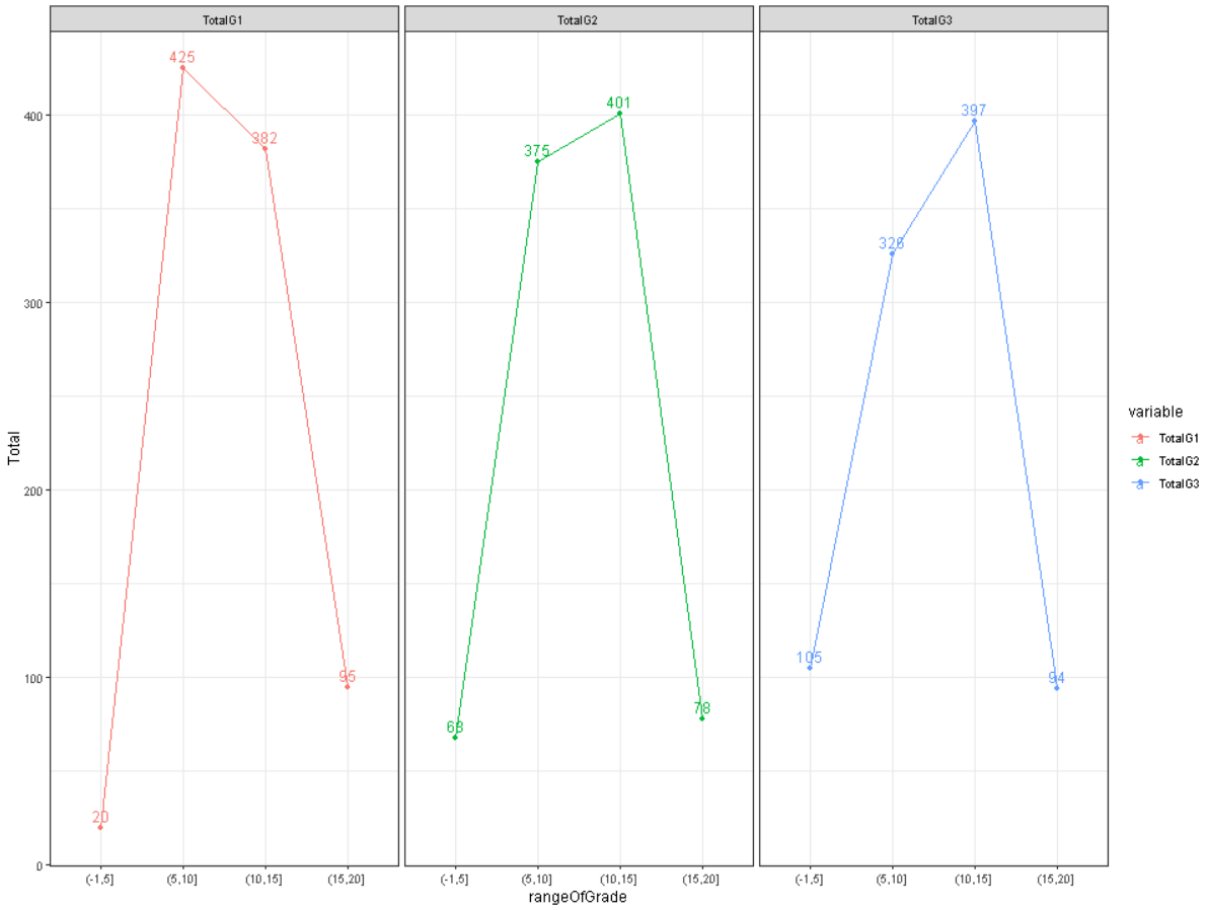


Figure 8.2 Display of multiple line chart in wrap function.

Conclusion

In general, students' grades are affected by different factors, such as family, friends, psychological factors and health factors. After analysing the scores of all grades for all students, the main factors that affect the scores are long Internet access, join extra-curriculum activities, weekly study time less, and some factors. There are also some factors that will give students the opportunity to improve, such as paid extra class, family relationship, education support, and even school support.

Another thing is R language, in this project I use data manipulation, data visualization, data exploration and data transformation to analyse each problem. R language is the best language for data calculus than other languages. And I myself have also improved my own analysis ability of problems in this project and made good use of R language to distinguish which data is more suitable for which problems.

References

1. Robinson(2016). *R melt() and cast() functions - Reshaping the data in R.* journaldev. https://www.journaldev.com/47883/r-melt-and-cast-function

2. Achim, Z. Paul, M. (2019). *HCL-Based Color Palettes in grDevices*. R developer page. https://developer.r-project.org/Blog/public/2019/04/01/hcl-based-color-palettes-in-grdevices/

3. Jim, L. *pie3D function in R*. R CHARTS. https://r-charts.com/part-whole/pie3d/

4. Flona, R. (2016). *TITLES AND AXES LABELS*. Environmental Computing. https://environmentalcomputing.net/graphics/ggplot/ggplot-labels/

5. Dans, T. *HTML Color Codes*. Dan's Tools. https://www.hexcolortool.com/#4eb19d