



INDIVIDUAL ASSIGNMENT

CT127 -3-2- PFDA

PROGRAMMING FOR DATA ANALYSIS

STUDENT NAME: Dipta Protim Guha

TP NUMBER: TP063351

INTAKE: FEBURARY 2022

HAND OUT DATE: 28th March 2022

HAND IN DATE: 9TH May 2022

INSTRUCTIONS TO CANDIDATES:

- 1 Submit your assignment at the administrative counter.**
- 2 Students are advised to underpin their answers with the use of references (cited using the Harvard Name System of Referencing).**
- 3 Late submission will be awarded zero (0) unless Extenuating Circumstances (EX) are upheld.**
- 4 Cases of plagiarism will be penalized.**
- 5 The assignment should be bound in an appropriate style (comb bound or stapled).**
- 6 Where the assignment should be submitted in both hardcopy and softcopy, the softcopy of the written assignment and source code (where appropriate) should be on a CD in an envelope/CD cover and attached to the hardcopy.**
- 7 You must obtain 50% overall to pass the module.**

Table of Contents

1.0 Introduction	4
1.1 Assumptions.....	4
2.0 Data Import, Pre-Processing, and cleaning.....	4
2.1 Data import.....	4
2.2 Data Exploration and Pre-Processing.....	5
2.2.1 View dataset in a table form	5
2.2.2 Take a sample of the dataset	6
2.2.3 View Column names	7
2.2.4 Number of rows and column in the dataset	7
2.2.5 Summary of Dataset.....	8
2.3 Data Cleaning	9
3.0 Question 1: Which gender group of students did well academically?	10
3.1 Analysis 1: Which gender group has a better average final grade score?	10
3.2 Analysis 2: How many male and female students are there?.....	11
3.3 Analysis 3: Average of all three grades for male and female students	12
3.4 Analysis 4: Which gender group scored more distinctions?	13
3.5 Question 1 Conclusion	14
4.0 Question 2: Students from which of the 2 schools performed well academically?	15
4.1 Analysis 1: How many students are there in each school?.....	15
4.2 Analysis 2: Which school students have a better average grade.	16
4.3 Analysis 3: Which school students have more failures in their final Grade?.....	17
4.4 Analysis 4: Which school students have more distinctions in their final Grade?	20
4.5 Question 2 conclusion.....	23
5.0 Question 3: Does alcohol consumption affect students' academic performance.....	24
5.1 Analysis 1: Does weekend alcohol consumption affect students' final grade?	24
5.2 Analysis 2: Does workday alcohol consumption affect students' final grade?	25
5.3 Analysis 3: Does workday alcohol consumption lead to health issues?	27
5.4 Analysis 4: Can health lead to absences which can lead to a drop in grade?	28
5.5 Question 3 conclusion.....	29
6.0 Question 4: Does romance affect academic performance?	30
6.1 Analysis 1: Relationship between romance and students final Grade	30

6.2 Analysis 2: Does romance cause students to go out more and study less which can lead to drops in their final grades?	31
6.3 Analysis 3: Does romance causes more absences for students and does that affect their final grades?	32
6.4 Question 4 Conclusion	34
7.0 Question 5: Does students' location from school affect their academic performance?	35
7.1 Analysis 1: Does students' location from school affect their Final Grade?	35
7.2 Analysis 2: Does students' location from school affect their travel time?	36
7.3 Analysis 3: Does students' location from school affect their travel time?	37
7.4 Question 5 conclusion.....	38
8.0 Extra features.....	39
8.1 Extra feature 1: tidyverse package	39
8.2 Extra feature 2: gridExtra package.....	39
8.3 Extra feature 3: Violin Graph	39
8.4 Extra feature 4: Density Graph	39
9.0 Conclusion	40
10.0 Appendix	41
11.0 References	51

1.0 Introduction

The following documentation is about exploration of a dataset which contains various information about three-year final scores of degree students' marks. The dataset also contains many other features that might have or might have not impacted the academic performance of these students. Some of these features are, the school the student goes to, gender of the student, area of their accommodation, family information, daily/weekly alcohol consumption and many more.

The dataset was stored in Microsoft excel in CSV format and was imported to RStudio to implement various preprocessing techniques to come up with appropriate analysis. The techniques include data exploration, manipulation, transformation, and visualization.

1.1 Assumptions

The coding for analysis of the dataset shall be conducted on RStudio and shall display appropriate outputs like tables, graphs, etc. Each analysis will have screenshot(s) of the source code, as well as the output and a small description of the analysis. Any additional features used to perform analysis shall also be documented in this reported with screenshots(s) of source code and output and its description.

For the student Grades column, the numbering is form 0-20. Any students who scored 16 or more was awarded distinction. Any student who scored less than 10, failed the class.

Both schools are assumed to be located in the urban area or city area.

2.0 Data Import, Pre-Processing, and cleaning

2.1 Data import

```
1 #Name and TP Number
2 #Dipta Protim Guha
3 #TP063351
4
5 data = read.csv("C:\\Users\\aritr\\Downloads\\student.csv", header=TRUE)
6 data
```

Figure 2.1.1

To import the data, in line 5, a variable called data was created and the reserved R keyword “read.csv” was used. This function has 2 parameters, the location of the file and header. The

location parameter is used to locate the file in the directory where it exists in the computer. The header function is used to determine if the first row of the dataset are column names or not. If the header is true, the first row will be column names; If false, the first row will be included as part of the information.

2.2 Data Exploration and Pre-Processing

Various pre-processing techniques were used to verify if the data import worked successfully.

Also, these techniques were used comprehend and get a better understanding of the dataset.

Before going ahead with the preprocessing techniques, a couple of R packages were installed and loaded to help perform various functions that are not readily available in RStudio.

```
10 #install packages
11 install.packages("dplyr")
12 install.packages("ggplot2")
13
14 #load Packages
15 library(ggplot2)
16 library(dplyr)
17
```

Figure 2.2.1.

2.2.1 View dataset in a table form

```
18 #view table form
19 View(data)
20
21
```

Figure 2.2.1.1

In line 19, the view keyword was used to view the dataset in a table form.

	index	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
1	1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother
2	2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father
3	3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother
4	4	GP	F	15	U	GT3	T	4	2	health	services	home	mother
5	5	GP	F	16	U	GT3	T	3	3	other	other	home	father
6	6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother
7	7	GP	M	16	U	LE3	T	2	2	other	other	home	mother
8	8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother
9	9	GP	M	15	U	LE3	A	3	2	services	other	home	mother
10	10	GP	M	15	U	GT3	T	3	4	other	other	home	mother

Showing 1 to 10 of 922 entries, 34 total columns

Figure 2.2.1.2

2.2.2 Take a sample of the dataset

21	head(data)	
22		
23		

23:1 (Top Level) R Script

```
R 4.1.2 · C:/Users/aritr/Desktop/R programming/R assignment/
> head(data)
```

	index	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
1	1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother
2	2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father
3	3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother
4	4	GP	F	15	U	GT3	T	4	2	health	services	home	mother
5	5	GP	F	16	U	GT3	T	3	3	other	other	home	father
6	6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother

	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic
1	2	2	0	yes	no	no	no	yes	yes	no	no
2	1	2	0	no	yes	no	no	no	yes	yes	no
3	1	2	3	yes	no	yes	no	yes	yes	yes	no
4	1	3	0	no	yes	yes	yes	yes	yes	yes	yes
5	1	2	0	no	yes	yes	no	yes	yes	no	no
6	1	2	0	no	yes	yes	yes	yes	yes	yes	no

	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
1	4	3	4	1	1	3	6	5	6	6
2	5	3	3	1	1	3	4	5	5	6
3	4	3	2	2	3	3	10	7	8	10
4	3	2	2	1	1	5	2	15	14	15
5	4	3	2	1	2	5	4	6	10	10
6	5	4	2	1	2	5	10	15	15	15

Figure 2.2.2.1

In line 21, the head keyword was used to view the first 6 rows of the dataset.

2.2.3 View Column names

```
23 names(data)
23:12 (Top Level) R S

R 4.1.2 · C:/Users/aritr/Desktop/R programming/R assignment/
> names(data)
[1] "index"      "school"     "sex"        "age"        "address"    "famsize"    "Pstatus"
[8] "Medu"       "Fedu"       "Mjob"       "Fjob"       "reason"     "guardian"   "traveltime"
[15] "studytime"  "failures"   "schoolsup"  "famsup"     "paid"       "activities" "nursery"
[22] "higher"     "internet"   "romantic"   "famrel"     "freetime"   "goout"      "Dalc"
[29] "walc"       "health"     "absences"   "G1"         "G2"         "G3"
>
>
>
```

Figure 2.2.3.1

In line 23, the names keyword was used to see all the column names of the dataset.

2.2.4 Number of rows and column in the dataset

```
27 #count the number of column and rows
28 nrow(data)
29 ncol(data)
29:11 (Top Level)

R 4.1.2 · C:/Users/aritr/Desktop/R programming/R assignment/
> nrow(data)
[1] 922
> ncol(data)
[1] 34
>
```

Figure 2.2.4.1

In lines 28 and 29, the reserved keywords, nrow and ncol were used to count the number of rows and columns in the dataset.

2.2.5 Summary of Dataset

```
31 summary(data)
```

R 4.1.2 · C:/Users/aritr/Desktop/R programming/R assignment/

```
> summary(data)
```

index	school	sex	age	address
Min. : 1.0	Length:922	Length:922	Min. :15.00	Length:922
1st Qu.:231.2	Class :character	Class :character	1st Qu.:16.00	Class :character
Median :461.5	Mode :character	Mode :character	Median :17.00	Mode :character
Mean :461.5			Mean :16.74	
3rd Qu.:691.8			3rd Qu.:18.00	
Max. :922.0			Max. :22.00	

famsize	Pstatus	Medu	Fedu	Mjob
Length:922	Length:922	Min. :0.000	Min. :0.000	Length:922
Class :character	Class :character	1st Qu.:2.000	1st Qu.:2.000	Class :character
Mode :character	Mode :character	Median :3.000	Median :2.500	Mode :character
		Mean :2.753	Mean :2.536	
		3rd Qu.:4.000	3rd Qu.:3.000	
		Max. :4.000	Max. :4.000	

Fjob	reason	guardian	traveltime	studytime
Length:922	Length:922	Length:922	Min. :1.000	Min. :1.000
Class :character	Class :character	Class :character	1st Qu.:1.000	1st Qu.:1.000
Mode :character	Mode :character	Mode :character	Median :1.000	Median :2.000
			Mean :1.457	Mean :2.037
			3rd Qu.:2.000	3rd Qu.:2.000
			Max. :4.000	Max. :4.000

Figure 2.2.5.1

Lastly, in line 31, the summary function was used to retrieve various information of each column in the dataset. The information includes the average, the median and mode value, etc.

2.3 Data Cleaning

As mentioned by the lecturer, the dataset is clean and there are no missing values. But the first column called “index” seems redundant as it is just a column that is counting the row number. That is why, it was decided to remove the column from the dataset.

```
34 #remove the index column from the dataset
35 data$index = NULL
36 data
```

35:18 (Top Level) R Script

R 4.1.2 · C:/Users/aritr/Desktop/R programming/R assignment/

```
Max. :19.00 Max. :19.00 Max. :20.00
> data$index = NULL
> data
```

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother
5	GP	F	16	U	GT3	T	3	3	other	other	home	father
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother
11	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother
12	GP	F	15	U	GT3	T	2	1	services	other	reputation	father
13	GP	M	15	U	LE3	T	4	4	health	services	course	father
14	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother
15	GP	M	15	U	GT3	A	2	2	other	other	home	other
16	GP	F	16	U	GT3	T	4	4	health	other	home	mother
17	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother
18	GP	F	16	U	GT3	T	3	3	other	other	reputation	mother
19	GP	M	17	U	GT3	T	3	2	services	services	course	mother

Figure 2.3.1

In line 35, the index column from the dataset was set to NULL. This resulted in the column, index, being deleted from the current dataset.

3.0 Question 1: Which gender group of students did well academically?

According to the dataset, students include both the gender group of male and females. A couple of analyses were conducted which gender group (male or female) did well academically.

3.1 Analysis 1: Which gender group has a better average final grade score?

Input

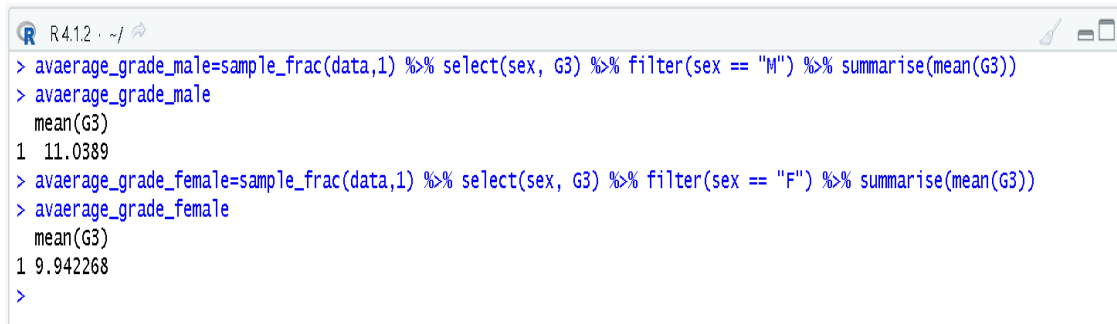
```
38 #analysis 3.1 which gender group has a better average final grade
39
40 #male
41 average_grade_male=select(data, sex, G3) %>% filter(sex == "M") %>% summarise(mean(G3))
42 average_grade_male
43
44 #female
45 average_grade_female=select(data, sex, G3) %>% filter(sex == "F") %>% summarise(mean(G3))
46 average_grade_female
```

Figure 3.1.1

In Figure 3.1, a variable called “average_grade_male” was created. The piping function was used which allows to take the output of one function and pass that as an argument for another function (Spanton, 2020). In this case, the select function was used to take only the “sex” and the “G3” column. Then, it was piped to the filter function. The filter function lets users retain the rows or columns which satisfy the condition set by the user. In this case, the filter function was used to set the condition of retaining all the male students. Lastly, the summarise function was piped to previously written code. The summarise function is a function of the dplyr package which allows users to perform basic math operations like find the total and average of columns and rows in a dataset. Here, it was used to find the average of the G3 column for only the students who are male.

Another variable was created called “average_grade_female”. This variable was defined similarly to the previously created “average_grade_male”. The only difference is that the sex column in the filter function had “F” instead of “M”.

Output



```
R R4.1.2 ~ /
> average_grade_male=sample_frac(data,1) %>% select(sex, G3) %>% filter(sex == "M") %>% summarise(mean(G3))
> average_grade_male
  mean(G3)
1 11.0389
> average_grade_female=sample_frac(data,1) %>% select(sex, G3) %>% filter(sex == "F") %>% summarise(mean(G3))
> average_grade_female
  mean(G3)
1 9.942268
>
```

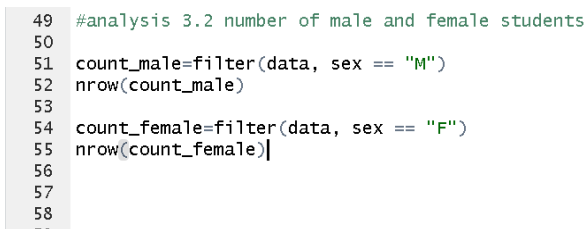
Figure 3.1.2

According to the output in figure 3.2.2, we can see both average scores of final grades for both males and females. It shows that mean value for male gender group is 11.0389 while 9.942268 for the female gender group. This indicates that males gender group scored better than the female gender group.

3.2 Analysis 2: How many male and female students are there?

The results of the first analysis, where only the average of G3 column was taken could have some inaccuracies. This is because, if the number of students in one gender group is drastically higher than the other gender group, the average value of G3 would not be as accurate. That is why it was decided to find out how many male and female students exist out of all the 922 students.

Input



```
49 #analysis 3.2 number of male and female students
50
51 count_male=filter(data, sex == "M")
52 nrow(count_male)
53
54 count_female=filter(data, sex == "F")
55 nrow(count_female)
56
57
58
59
```

Figure 3.2.1

In figure 3.2.1, at line 51, A variable called “count_male” was created and the filter function was used to add the condition of only male student only. Then, the nrow function was used with the

new created “count_male” variable to count the row numbers of all the male student. The same process was used to find the number of female students too in line 54.

Output

```
> count_male=filter(data, sex == "M")
> nrow(count_male)
[1] 437
>
> count_female=filter(data, sex == "F")
> nrow(count_female)
[1] 485
> |
```

Figure 3.2.2

In figure 3.2.2, We can see that the difference between the number of male and female students is very minimal as there only 48 female students more than male students. This concludes that the average of G3 column found in the previous was not affected heavily as the difference between and male and female students is not that high.

3.3 Analysis 3: Average of all three grades for male and female students

For the following analysis, it was decided to add all 3 grade columns (G1, G2, G3) to form a temporary third column and find the average of that temporary column for each respective gender group

Input

```
58 #analysis 3.3 average of all 3 grade columns combine for male and female students
59
60 average_grade_all_grades_male=sample_frac(data,1) %>% mutate(A11_Grades_out_of_60=G1+G2+G3) %>% select(sex, A11_Grades_out_of_60)%>%
61   filter(sex == "M") %>% summarise(mean(A11_Grades_out_of_60))
62 average_grade_all_grades_male
63
64 average_grade_all_grades_female=sample_frac(data,1) %>% mutate(A11_Grades_out_of_60=G1+G2+G3) %>% select(sex, A11_Grades_out_of_60) %>%
65   filter(sex == "F") %>% summarise(mean(A11_Grades_out_of_60))
66 average_grade_all_grades_female
67
```

Figure 3.3.1

In figure 3.3.1, the sample_frac function was used to take 100% of the dataset. Next, it was piped with the mutate function which added all the G1, G2, G3 columns to form a temporary column with the total score. Then, it was piped with the select function to select only the sex and newly created column and piped with the filter function. The filter function was used to set a condition

where the sex will only be the male students. Lastly, the code was piped with the summarise function helped finding the mean value for the newly created column for the male students only.

The same thing was done for the female students with the only difference of adding the female students only for the filter function.

Output

```
> average_grade_all_grades_male=sample_frac(data,1) %>% mutate(All_Grades_out_of_60=G1+G2+G3) %>% select(sex, All_Grades_out_of_60)%>%
+ filter(sex == "M") %>% summarise(mean(All_Grades_out_of_60))
> average_grade_all_grades_male
mean(All_Grades_out_of_60)
1 33.52632
> average_grade_all_grades_female=sample_frac(data,1) %>% mutate(All_Grades_out_of_60=G1+G2+G3) %>% select(sex, All_Grades_out_of_60) %>%
+ filter(sex == "F") %>% summarise(mean(All_Grades_out_of_60))
> average_grade_all_grades_female
mean(All_Grades_out_of_60)
1 30.94845
> |
```

Figure 3.3.2

According to the output in figure 3.3.2, it shows that male gender group has a better average score when all the scores from G1, G2, G3 were combined. This analysis still shows the male students did well academically compared to the female students.

3.4 Analysis 4: Which gender group scored more distinctions?

Another way, the question, which gender group did better academically is by finding out which gender group of students scored the most distinctions.

Input

```
70 #3.4 analysis Which gender group scored the most number of distinctions?
71 distinction_male= filter(data, sex == "M", G3 >= 16)
72 i=nrow(distinction_male)
73
74 distinction_female=filter(data, sex == "F", G3 >= 16)
75 q=nrow(distinction_female)
76
77 a=c(i,q)
78 b=c("male","female")
79 pie(a,b,radius = 1, main = "Most distinctions", col = c("red","blue"),clockwise = TRUE)|
80
```

Figure 3.4.1

In Figure 3.4.1, line 71, the filter function was used to retrieve only the students who are male and scored more than 16 in the G3 column. The code was used for the female students in line 74. Next, a variable was created which counted the number of rows for each male and female

students who scored distinctions. Then, the pie function was used to include the 2 newly created variables, along with labels to provide a visualization by means of a pie chart.

Output

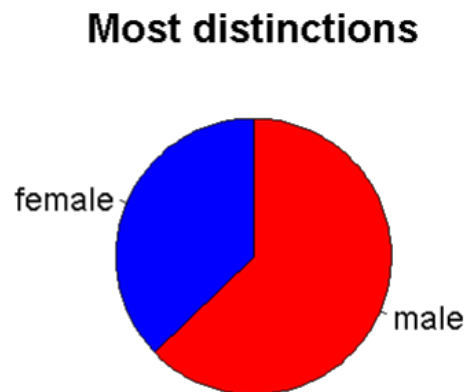


Figure 3.4.2

According to the output in figure 3.4.2, it shows that the pie chart is divided into 2 sections, each for male and female respectively. The male section covers more of the pie chart, compared to the female sections. This concludes that, the number of distinctions achieved by the male students are more than the female gender group.

3.5 Question 1 Conclusion

To answer the question of which gender group did well academically, 4 analyses were conducted. In all 4 analyses, it was discovered that the male group excelled more academically than the females. The average final score was higher for the male students, the average score of all 3 scores combined was higher for the male students and male students scored the most distinctions despite the number of female students being more. To conclude, the male students better than the female students academically.

4.0 Question 2: Students from which of the 2 schools performed well academically?

In the dataset, students belong to either of these 2 schools: Gabriel Pereira or Mousinho da Silveira. Various analyses were conducted to reveal which school students performed better academically.

4.1 Analysis 1: How many students are there in each school?

It was decided to find out how many students belong in each school to help in future analyses of student academic performances for each school.

Input

```
82 #4.1 Analysis 1: How many students are there in each school?
83
84 count_student_GP= filter(data, school == "GP")
85 nrow(count_student_GP)
86
87 count_student_MS= filter(data, school == "MS")
88 nrow(count_student_MS)
89
```

Figure 4.1.1

In figure 4.1.1, in line 84, a variable was created which stored the filter function, which was used with a condition for the school column for Gabriel Pereira. The nrow function was used to count the number of rows of the created variable. Some process was used for the other school, Mousinho da Silveira.

Output

```
> count_student_GP= filter(data, school == "GP")
> nrow(count_student_GP)
[1] 749
>
> count_student_MS= filter(data, school == "MS")
> nrow(count_student_MS)
[1] 173
> |
```

Figure 4.1.2

The output in figure 4.1.2 shows that the number of students in Gabriel Pereira is much higher than the students that belong to Mousinho da Silveira.

4.2 Analysis 2: Which school students have a better average grade.

Input

```
93 data %>% select(school, G1, G2, G3) %>% mutate(all_grades=G1+G2+G3) %>% subset(school == "GP") %>%  
94   summarise(mean(all_grades))  
95  
96 data %>% select(school, G1, G2, G3) %>% mutate(all_grades=G1+G2+G3) %>% subset(school == "MS") %>%  
97   summarise(mean(all_grades))
```

Figure 4.2.1

To find the average of all grades, firstly the dataset name was mentioned. Next, it was piped to the select function which only retrieved the school, G1, G2, G3 columns. Then, it was piped to the mutate function which added all three columns (G1,G2,G3) to create a temporary column called all_grades. Next, the code was piped to the subset function which helped put a condition where the school is only GP. Next, it was piped to the summarise function which helped to find the average of the newly created column, all_grades. The same process was conducted for the other school, MS.

```
> data %>% select(school, G1, G2, G3) %>% mutate(all_grades=G1+G2+G3) %>% subset(school == "GP") %>% summarise(me  
an(all_grades))  
  mean(all_grades)  
1      32.10013  
>  
> data %>% select(school, G1, G2, G3) %>% mutate(all_grades=G1+G2+G3) %>% subset(school == "MS") %>% summarise(me  
an(all_grades))  
  mean(all_grades)  
1      32.47399  
>
```

Figure 4.2.2

According to the output, the average grades of students from both of 2 schools are almost the same. Despite Gabriel Pereira having about 4 times the number of students in Mousinho da Silveira. Also, the average for both schools indicate that most number of students in both school barely passed academically since G1, G2, G3 columns were added to be out of 60; 32 average for both schools means that it is almost half of 60, which is about 50 percent.

A table view was also constructed to have a better analysis of how students' scores were in G3 for both school schools.

Input

```
100 table(data$school, data$G3)
101
```

Figure 4.2.3

The table function was used to select 2 columns, school and G3 from the dataset, data.

Output

```
> table(data$school, data$G3)
      0  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
GP  72  2 12 29 16 63 52 102 90 59 59 53 63 30 12 24  9  2
MS  14  0  5  7  5 11 13 28 21 11 13  9 19  9  2  2  3  1
```

Figure 4.2.4

According to the table view in figure 4.2.4, it shows that both schools have the highest number of students who scored 10 in their final Grade. This is also similar to the previously found average value for all 3 grades combined as 10 is 50% of 20 and G3 is out of 20.

4.3 Analysis 3: Which school students have more failures in their final Grade?

Input

```
106 GP=data %>% filter(school == "GP", G3<10) %>%
107   ggplot(aes(x=G3))+geom_histogram(colour="black",aes(fill=..count..))+
108   scale_fill_gradient("count", low = "green", high = "red")+
109   ggtitle("Student from Gabriel Pereira with score less than 10")
110
111
112 MS=data %>% filter(school == "MS", G3<10) %>%
113   ggplot(aes(x=G3))+geom_histogram(colour="black",aes(fill=..count..))+
114   scale_fill_gradient("count", low = "green", high = "red")+
115   ggtitle("Student from Mausinho da Silveira with score less than 10")
116
117 grid.arrange(GP,MS)
```

Figure 4.3.1

In Figure 4.3.1, the filter function was used to set 2 conditions where the school is MS and G3 is less than 10. Then, it was piped to the ggplot function to create a histogram. As histogram only

requires 1 continuous x value and no y value, the G3 column was chosen for the x axis. The color of the borders was set to black and the `scale_fill_gradient` function was used to provide changing colors based on frequency. Lastly, the `ggtitle` function was used to provide a title for the histogram.

The same process was repeated to create a histogram for the students from Mousinho da Silveira. School was set to MS instead of GP. Next, the `grid.arrange` function was used to combine the two histograms together into one plot.

Output

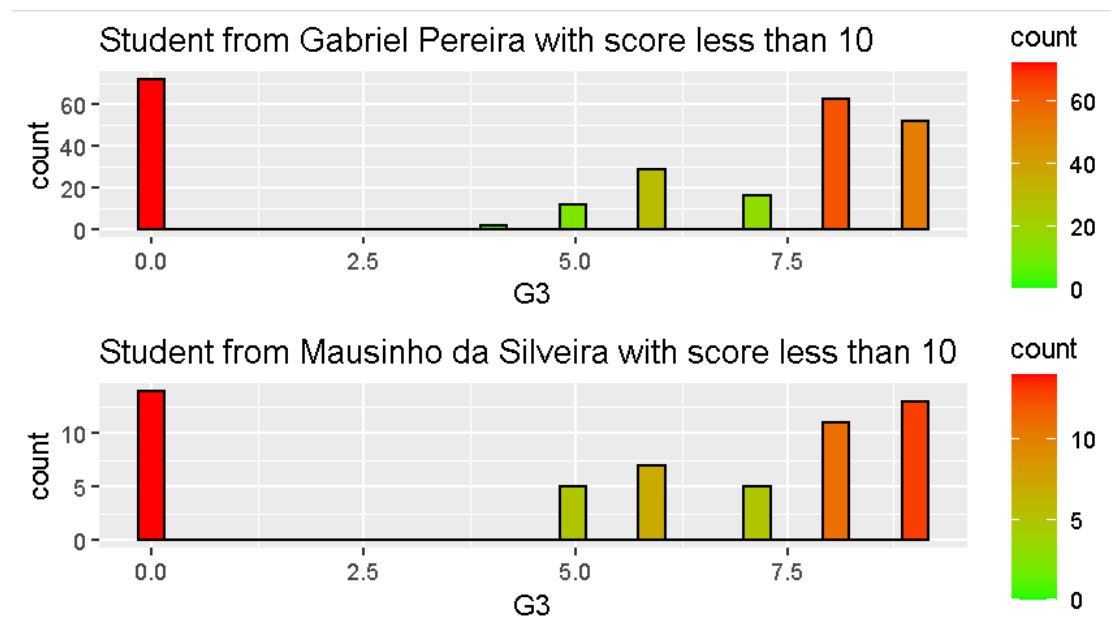


Figure 4.3.2

In the output, the first thing that can be noticed is that the count range for students from Gabriel Pereira is from 0 to 60 while it is only 0 to 10 for students from Mousinho da Silveira. Just by this, it can be understood that the number of students who failed Gabriel Pereira are much higher than students in Mousinho da Silveira. Therefore, Gabriel Pereira has more students who failed, but it has to be kept in mind that Gabriel Pereira has a lot more students than Mousinho da Silveira.

Exact number of students who failed and percentage of failures to students for each school

Input

```
120 GP_count=data %>% filter(school == "GP", G3 < 10)
121 MS_count=data %>% filter(school == "MS", G3 < 10)
122 nrow(GP_count)
123 nrow(MS_count)
124
125 GP_percent=246/749
126 round(GP_percent,digits = 2)
127
128 MS_percent=55/173
129 round(MS_percent,digits = 2)|
130
```

Figure 4.3.3

To count the exact number of students who failed in each school, a variable was created and the filter function was used to set school to GP and $G3 < 10$. Next the nrow function was used to call the created variable. This provided the number of students who failed in Gabriel Pereira. The same process was used to find the number of students from Mousinho da Silveira.

To find the percentage of students who failed, a variable was created and the number of students who failed in Gabriel Pereira was divided by the number of students in that school (the number of students in each school was found in the first analysis). Next, the round function was used to call the variable and round the output to 2 digits only. The same process was used for the other school.

Output

```

> GP_count=data %>% filter(school == "GP", G3 < 10)
> MS_count=data %>% filter(school == "MS", G3 < 10)
> nrow(GP_count)
[1] 246
> nrow(MS_count)
[1] 55
>
> GP_percent=246/749
> round(GP_percent,digits = 2)
[1] 0.33
>
> MS_percent=55/173
> round(MS_percent,digits = 2)
[1] 0.32
>

```

Figure 4.3.4

According to the output, 246 students from Gabriel Pereira and 55 students from Mousinho da Silverira failed. And the percentage of failed students to number of students in each school is almost the same. It is 33 percent for Gabriel Pereira students, and 32 percent for Mousinho da Silverira.

4.4 Analysis 4: Which school students have more distinctions in their final Grade?

Input

```

136 data %>% filter(G3>=16) %>% ggplot(aes(x=school))+
137   geom_bar(fill=c("red","blue"))+
138   ggtitle("Number of students who achieved distinction from each school")
139

```

Figure 4.4.1

In figure 4.4.1, the filter function was used to set the condition where G3 is more than or equal to 16. Then, it was piped to the ggplot function to create a bar graph with the x axis being the school column. The fill function was used to give different colors to each bars and the ggtitle function was used to give a title to the plot.

Output

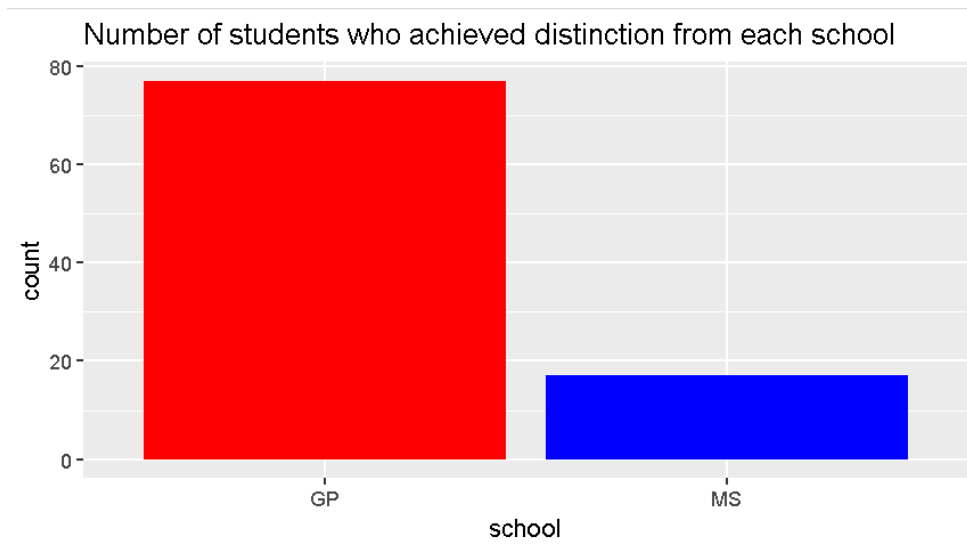


Figure 4.4.2

According to the output, it can be understood that students in Gabreil Pereira scored more distinctions than students in Mousinho da Silveira. But it must be kept into account that students in Gabriel Pereria is a lot more than in Mousinho da Silveira.

Exact number of students who received distinctions and percentage of distinctions to students for each school

Input

```

142 GP_count=data %>% filter(school == "GP", G3 >= 16)
143 MS_count=data %>% filter(school == "MS", G3 >= 16)
144 nrow(GP_count)
145 nrow(MS_count)
146
147 GP_percent=77/749
148 round(GP_percent,digits = 2)
149
150 MS_percent=17/173
151 round(MS_percent,digits = 2)
152

```

Figure 4.4.3

To count the exact number of students who got distinctions in each school, a variable was created, and the filter function was used to set school to GP and G3 >=10. Next the nrow function was used to call the created variable. This provided the number of students who got

distinction in Gabriel Pereira. The same process was used to find the number of students from Mousinho da Silveira.

To find the percentage of students who got distinction, a variable was created and the number of students who got distinction in Gabriel Pereira was divided by the number of students in that school (the number of students in each school was found in the first analysis). Next, the round function was used to call the variable and round the output to 2 digits only. The same process was used for Mousinho da Silveira.

Output

```
> data %>% filter(G3>=16) %>% ggplot(aes(x=school))+geom_bar(fill=c("red","blue"))+
ggtitle("Number of students who achieved distinction from each school")
> GP_count=data %>% filter(school == "GP", G3 >= 16)
> MS_count=data %>% filter(school == "MS", G3 >= 16)
> nrow(GP_count)
[1] 77
> nrow(MS_count)
[1] 17
>
> GP_percent=77/749
> round(GP_percent,digits = 2)
[1] 0.1
>
> MS_percent=17/173
> round(MS_percent,digits = 2)
[1] 0.1
> |
```

Figure 4.4.4

According to the output, 77 students from Gabriel Pereira and 17 students from Mousinho da Silverira received distinction. And the percentage of distinction students to number of students in each school is the same. It is 11 percent for students from both schools.

4.5 Question 2 conclusion

Firstly, it was found out the then number of students in Gabriel Pereira is much higher than the students in Mousinho da Silveira. Secondly, it was revealed that the mean value for both the schools, for all 3 grade columns combined, is almost the same. Next, it was discovered that although more students from Gabriel Pereira failed in the final grade than the other school, the failure to number of students percentage is the same as Mousinho da Silveira. Lastly, it was discovered that although the number of distinctions in Gabriel Pereira is also higher, the distinction students to the number of students percentage is also the same as Mousinho da Silveira. After all these findings, it can be concluded that students from neither school outperformed each other academically. Their academic prowess is almost the same.

5.0 Question 3: Does alcohol consumption affect students' academic performance

According to a study conducted by Karen Patte, Wei Qian and Scott Leatherdale, it was found that students who engage in regular binge drinking are less likely to attend classes, complete their homework and achieve better grades (Patte, Qian & Leatherdale, 2019). Various analyses from the dataset were conducted to confirm if it is in fact true that students' academic performance was affected because of alcohol consumption.

5.1 Analysis 1: Does weekend alcohol consumption affect students' final grade?

Input

```
158 data %>% ggplot(aes(x=walc, y=G3))+  
159   geom_point(aes(color=walc))+  
160   stat_smooth(method = lm)+  
161   labs(title = "Relationship between weekend alcohol consumption and students' final grade",  
162        x = "weekend alcohol consumption (1 - very low to 5 - very high", y="Final Grade")  
163
```

Figure 5.1.1

In figure 5.1.1, the code started with the dataset, data, and piped to the ggplot function to create a point graph. Inside the ggplot function, the aes function was used to assign the x and y axis. The x axis was assigned the Walc column while the y axis was assigned the G3 column. In line 159, the geom_point also has an aes function which was used to assign color for the column, Walc. This provided different shades of color from 1 to 5 for the Walc column in the dataset. The stat_smooth syntax was used to add a line in the point graph which helped visualize a positive or negative correlation between the x and y axis. Lastly, the labs function was used to give a title to the plot and add label for the x and y axis in the graph.

Output



Figure 5.1.2

In figure 5.1.2, the line in the middle indicates there is a negative correlation between final grades and weekend alcohol consumption. As the weekend alcohol consumption increases, the grades seem to decrease. While most of the points for weekend alcohol consumption of 1 and 2 are above 10 in G3 (above 10 means students passed), for weekend alcohol consumption of 3-5, the grades seemed to go down. Weekend alcohol consumption of 4, which is fairly high, takes the most hit as the graph indicates a lot of scores below 10 for the final grade.

5.2 Analysis 2: Does workday alcohol consumption affect students' final grade?

Input

```
165 #5.2 Analysis 2: Does daily alcohol consumption affect students' final grade?
166 data %>% ggplot(aes(x=Dalc, y=G3))+
167   geom_point(aes(color=Dalc))+facet_wrap(~Dalc)+
168   labs(title = "Relationship between daily alcohol consumption and students' final grade",
169     x= "daily alcohol consumption (1 - very low to 5 - very high", y="Final Grade")
170
```

Figure 5.2.1

In figure 5.2.1, the code started with the dataset, data, and piped to the ggplot function to create a point graph. Inside the ggplot function, the aes function was used to assign the x and y axis. The x axis was assigned the Dalc column while the y axis was assigned the G3 column. In line 159, the geom_point also has an aes function which was used to assign color for the column, Dalc. This provided different shades of color from 1 to 5 for the Walc column in the dataset. Next, the facet wrap function was used to divide the graph in to 5 different graphs based on the 1-5 range of the Dalc column. Lastly, the labs function was used to give a title to the plot and add label for the x and y axis in the graph.

Output

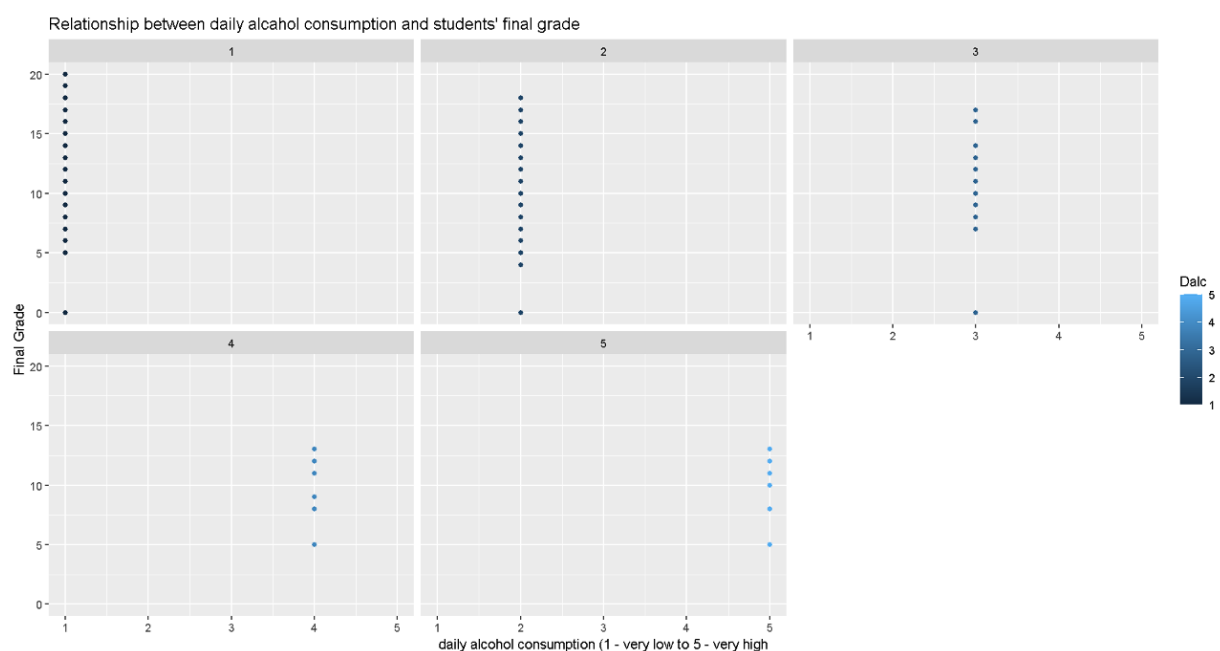


Figure 5.2.2

In figure 5.2.2, similar to the 1st analysis it, can be noticed that, as the workday alcohol consumption increases, the grades seem to decrease, but this time it is much worse than weekend alcohol consumption. The negative correlation is much higher, starting from 2 compared to the dip that was noticed from 3 in the weekend alcohol consumption. This time, workday alcohol consumption of both 4 and 5 took the most hit.

5.3 Analysis 3: Does workday alcohol consumption lead to health issues?

Input

```
175 data %>% ggplot(aes(x=Dalc, y=health))+geom_point(aes(color=Dalc))+facet_wrap(~Dalc)+  
176   labs(title = "Relationship between daily alcahol consumption and student health",  
177     x= "daily alcahol consumption (1 - very low to 5 - very high",  
178     y="Student health status (numeric: from 1 - very bad to 5 - very good)")  
179
```

Figure 5.3.1

In figure 5.3.1, the code started with the dataset, data and the ggplot function was used to create a point graph. In the ggplot function, the aes function was used to assign Dalc column to the x axis and health column to the y axis. Inside the geom_point function, the color function was used to assign the Dalc column which provided different shades of color from 1 to 5. Next, the facet_wrap function was used to divide the plot into 5 different graphs based on the column, Dalc. Lastly, the labs function was used to give a title to the plot and add label for the x and y axis in the graph.

Output

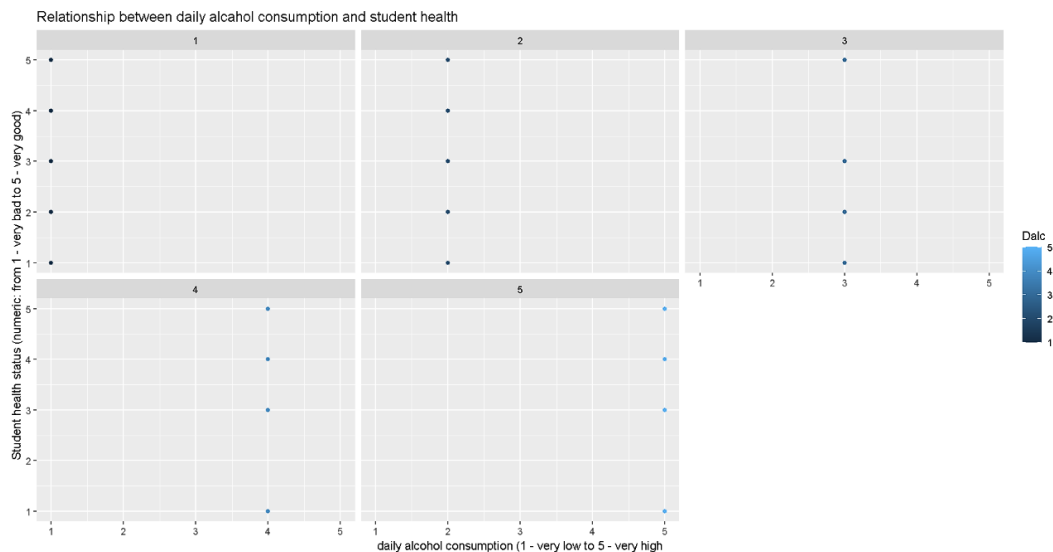


Figure 5.3.2

Based on the output in 5.3.2, it can be seen that there is some negative correlation between daily alcohol consumption and health. There is a point on 1 (very bad health) for both 4 and 5 of Dalc (5 being highest). Number 3 of Dalc seemed to be the most affected as there are dots on 1, 2, 3

for for health when Dalc is 3. After all these findings, it can be concluded that workday alcohol consumption does affect health of the students.

5.4 Analysis 4: Can health lead to absences which can lead to a drop in grade?

Input

```
175 data %>% ggplot(aes(x=absences, y=G3))+geom_point(aes(shape=factor(Dalc),color=factor(Dalc)))+  
176   labs(title = "Relationship between absences and students' final grade",  
177     x= "Absences", y="Final Grade")  
178
```

Figure 5.4.1

In figure 5.4.1, the code started with the dataset, data and the ggplot function was used to create a point graph. In the ggplot function, the aes function was used to assign absences column to the x axis and G3 column to the y axis. Inside the geom_point function, the color function was used to assign the Dalc column which provided different shades of color from 1 to 5. Also, the shape function was used to provide different shapes for the 1-5 range of the Dalc column. Lastly, the labs function was used to give a title to the plot and add label for the x and y axis in the graph.

Output

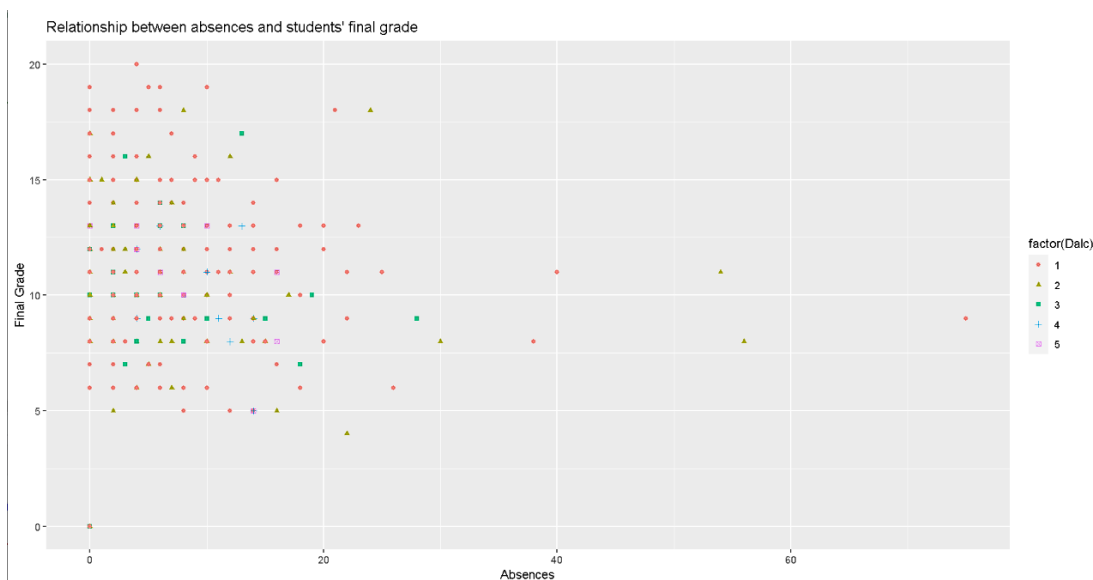


Figure 5.4.2

Based on the output in 5.4.2, it can be noticed that as the absences increase, the grades seem to decrease. Most of the effect can be seen between the range of 5 to 20 of the x axis (absences). Between the range of 5 to 20 of the x axis, there are a group of dots below 10 on the y axis (final grade). Also, the shapes from range of 1-5 of workday alcohol intake column can be seen between the range of 5-20 of the x axis and below 10 of y axis. Most shapes from the number 3 – 5 (high alcohol intake) of Dalc column can be visible between 10-20 of x axis and below 10 of y axis. Therefore, it can be concluded that workday alcoholism led to absences which cause grades of students to go down.

5.5 Question 3 conclusion

To analyze whether alcoholism affected students' grades in the dataset, various analyses were conducted. The first analysis was about if weekend alcohol intake affected students' grades and a negative correlation was noticed which led to the understanding that weekend alcohol intake did affect students' grades. For the 2nd analysis, workday alcohol was compared with students' grades and this time the effects were worst compared to the previous analysis as the effect on students' grades were higher. Next, if workday alcohol intake affected the students' health and there were some findings that indeed it did affect the health of the students. The last analysis was conducted to see if the health issues led to absences which caused drops in students' grades. The result proved that there is a negative correlation between health and students' final grades which was caused because of workday alcohol intake. After all these analyses, it can be concluded that alcoholism did hinder students' academic performance.

6.0 Question 4: Does romance affect academic performance?

In high school, romance between male and female can be quite fun but at the same time, it is also a form of distraction. Sometime being romantically involved with someone requires giving one's time to the other person, which most of the time is subtracted from the study time. That is why, a couple of analyses from the dataset were conducted to see if romance affects students' academic performance.

6.1 Analysis 1: Relationship between romance and students final Grade

Input

```
182 data %>%  
183   ggplot(aes(romantic, G3))+  
184   geom_boxplot(aes(fill=romantic))+  
185   labs(title = "Relationship between romance and students final Grade ",  
186        x= "Romantic involvement", y="Final Grade")
```

Figure 6.1.1

In figure 6.1.1, the code started with the dataset, data and the ggplot function was used to create a point box plot. In the ggplot function, the aes function was used to assign the romantic column and G3 column for the x and y axis respectively. Inside the boxplot function the aes function was used to use the fill function for the romantic column. This will enable the graph to fill all the parameters of the romantic column with different colors. Lastly, the labs function was used to assign graph titles and labels for x and y axis.

Output

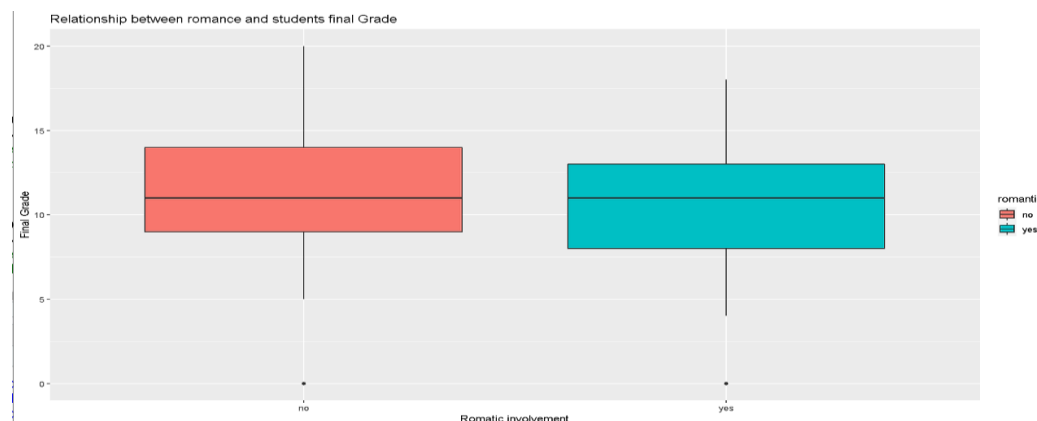


Figure 6.1.2

In the output in figure 6.1.2, it can be noticed just by glancing at the graph when the students are romantically involved, the grades seem to go down. If one was to take a closer look, it can be seen that the highest and lowest score of yes box plot is lower than no box plot. Even though the mean for the yes box plot is higher than the no box plot, the majority of the scores are above the mean line in the boxplot for the no box plot while majority of the scores are below the mean value for the yes box plot. The majority of the scores that are below the mean in the yes plot are below 10 which is fail.

6.2 Analysis 2: Does romance cause students to go out more and study less which can lead to drops in their final grades?

Input

```
193 romantic_table=data %>% select(romantic, goout, studytime, goout) %>% group_by(romantic) %>%  
194   summarise(mean(goout), mean(studytime))  
195  
196 View(romantic_table)
```

Figure 6.1.3

In figure 6.1.3, a new variable, romantic_table, was created. In the variable, the dataset, data was selected and piped to the select function which selects the romantic, goout, studytime and G3 columns. Next, the piping function was used to use the group by function which allowed the romantic column to be divided into yes and no. Lastly, the summarise function was used to find the mean of the goout, studytime and G3 columns. In line 196, the View function with the romantic_table variable in the parameter was used to view the data in table form.

Output

	romantic	mean(goout)	mean(studytime)	mean(G3)
1	no	3.097245	2.006483	10.829822
2	yes	3.081967	2.098361	9.718033

Figure 6.1.4

In the output in figure 6.1.4, although there is a correlation between the romantic and the G3 column as the average G3 column is lower for when romantic is yes, but there is no significant correlation between the romantic column and the average goout and studytime column. The mean values for goout and studytime column are almost identical when romantic is both yes and no. Therefore, it can be concluded that romance does not influence the frequency of students going out and their study time.

6.3 Analysis 3: Does romance causes more absences for students and does that affect their final grades?

Input

```
199 data %>%
200   ggplot(aes(romantic, absences))+
201   geom_violin(aes(fill=romantic))+ stat_summary(fun=mean, geom="point", size=2, color="yellow")+
202   labs(title = "Relationship between romance and absences ",
203        x= "Romantic involvement", y="Number of absences")
```

Figure 6.1.5

For the output in figure 6.1.5, the dataset, data was selected and piped to the ggplot function to create a violin graph. In the ggplot function, the aes function was used to assign the romantic column and absences column for the x and y axis respectively. Inside the boxplot function the aes function was used to use the fill function for the romantic column. This will enable the graph to fill all the parameters of the romantic column with different colors. Next, the stat_summary function was used to add a point in both violins which indicated the mean value of each violin respectively. Lastly, the labs function was used to assign graph titles and labels for x and y axis.

Output

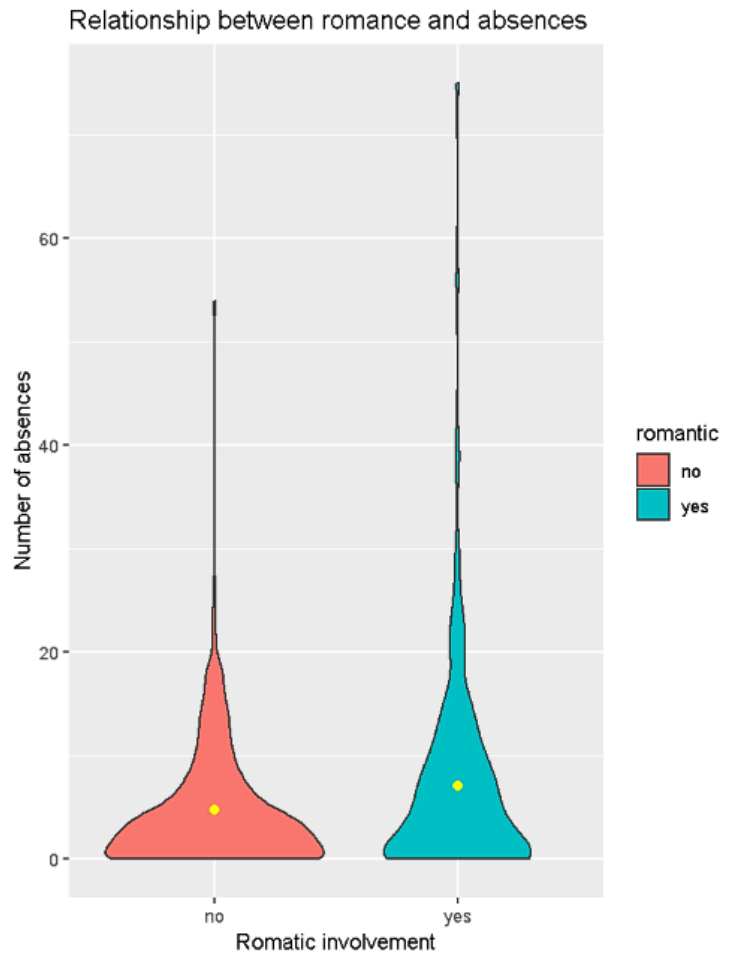


Figure 6.1.6

According to the output in figure 6.1.6, it can be comprehended that the average absences for students who are romantically involved are higher by looking at the two yellow points in the graph. Also, the height of the violin for the yes violin is higher reaching about 75 absences while it is 53 for the no violin. It is already understood from analysis 5.4 that absence does affect student performance. Therefore, it can be concluded that romance causes more absences for students, and it affects their academic performance.

6.4 Question 4 Conclusion

To find out if romance affects academic performance, various analyses were conducted. The first analysis was a direct comparison between students who are romantically involved and their final grade. The comparison proved that the final grade is affected when students are engaged in romance. The second analysis was conducted to find any relationship between romance and if it affects students' study time and the frequency of them going out. This analysis surprisingly did not show any correlation between romance and study time and frequency of going out as the values were almost the same for students who are in a romantic relationship and the students who are not. The last analysis was conducted to find out if romance led to an increased number of absences and it did and it is already known absences cause a drop in academic performance. After all 3 analyses, it can be concluded that romance does have a negative effect on students' academic performance.

7.0 Question 5: Does students' location from school affect their academic performance?

7.1 Analysis 1: Does students' location from school affect their Final Grade?

Input

```
206 data %>% ggplot(aes(x=G3))+  
207   geom_density(aes(fill=address))+  
208   labs(title = "Density graph of relationship between students' final grade and address",  
209         x= "Final Grade")
```

Figure 7.1.1

In figure 7.1.1, the dataset, data, was piped to the ggplot function to produce a density plot. In the ggplot function, the aes function was used to assign the G3 column to the x axis. In the geom_density function, the aes function was used to assign address to the fill parameter. This will help differentiate between the two types of student addresses, urban and rural. Lastly, the labs function was used to add a title and label to the x axis.

Output

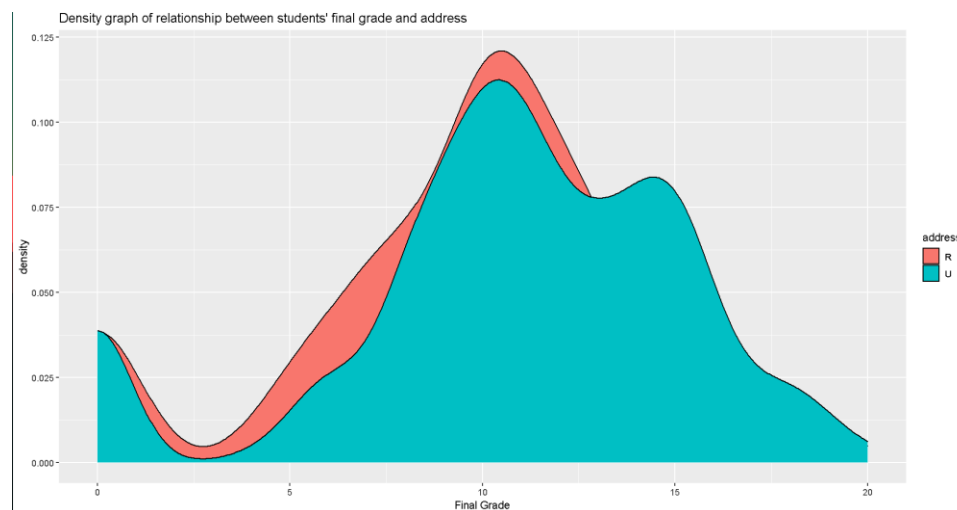


Figure 7.1.2

According to the output in figure 7.1.2, it can be understood right away that density of students in urban areas is much higher than students in rural areas. It also shows that most students who scored distinction (16 or more) are from urban areas. It can also be seen that the density for rural area students is most dense when it is below the score of 10 in the x axis, which is fail. These are

reasons why it can be said that students from rural areas performed better academically compared to students from rural areas.

7.2 Analysis 2: Does students' location from school affect their travel time?

Input

```
211 data %>% ggplot(aes(x=traveltime))+  
212   geom_density(aes(fill=address))+  
213   labs(title = "Density graph of relationship between students' travel time and address",  
214         x= "Travel time")  
215
```

Figure 7.2.1

In figure 7.2.1, the dataset, data was piped to the ggplot function to produce a density plot. In the ggplot function, the aes function was used to assign the traveltime column to the x axis. In the geom_density function, the aes function was used to assign address to the fill parameter. This will help differentiate between the two types of student addresses, urban and rural. Lastly, the labs function was used to add a title and label to the x axis.

Output

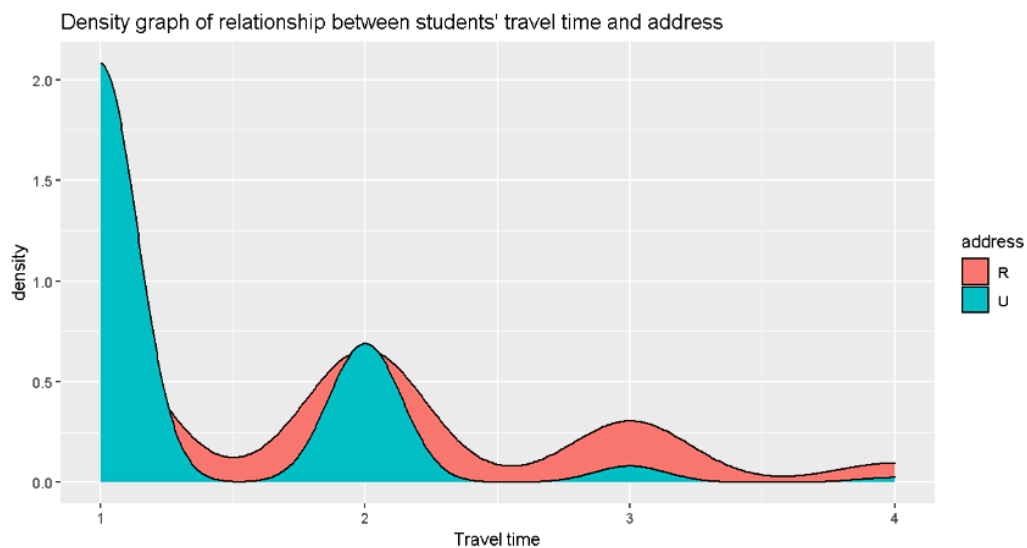


Figure 7.2.2

According to the output in figure 7.2.2, the density of urban area students can be seen as very dense when the travel time is 1 hour and slightly dense when it is 2 hours. For rural area

students, the density is very dense when travel time is between 2-4 hours as expected. As a result, it can be concluded that address does affect the travel time.

7.3 Analysis 3: Does travel time affect final grade?

Input

```
214 data %>% ggplot(aes(x=traveltime, y=G3))+geom_point(aes(color=traveltime))+  
215   stat_smooth(method = lm)+  
216   labs(title = "Relationship between travel time and students' final grade",  
217     x= "Travel Time", y="Final Grade")  
218
```

Figure 7.3.1

In figure 7.3.1, the code started with the dataset, data, and piped to the ggplot function to create a point graph. Inside the ggplot function, the aes function was used to assign the x and y axis. The x axis was assigned to the traveltime column while the y axis was assigned to the G3 column. The geom_point also has an aes function which was used to assign color for the column, traveltime. This provided different shades of color from 1 to 5 for the traveltime column in the dataset. The stat_smooth syntax was used to add a line in the point graph which helped visualize a positive or negative correlation between the x and y axis. Lastly, the labs function was used to give a title to the plot and add label for the x and y axis in the graph.

Output

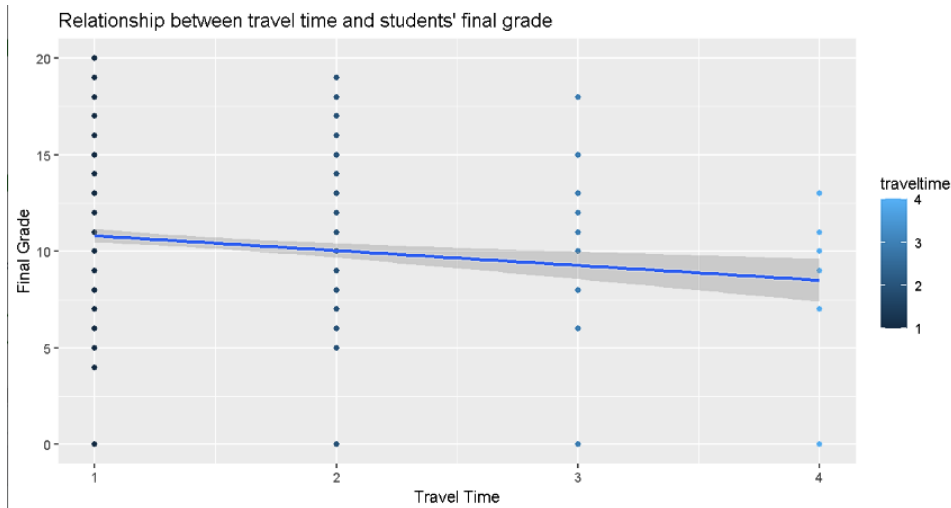


Figure 7.3.2

According to the graph it can be seen that there is a negative correlation between travel time and final grade. The line in the middle also indicated it. As the travel time increases, the grades seem to go down. The highest grade when the travel time is 4 hours is roughly 14 while it is 20 when it is 1 hour. After these findings it can be said that travel time does affect students' final grades.

7.4 Question 5 conclusion

To answer the question, does students' location affect their academic performance, various analyses were conducted. First, students' location and their final grade was compared which shows that students who live in urban areas did well in the final score than students living in rural areas. Next, students' location and their travel time was compared which showed that the students who lived in rural areas needed more travel time than students in urban areas. For the last analysis, students' travel time and final grade was compared which resulted in the scores dropping as the travel time is increased. After these 3 analyses, it can be concluded that students' location does affect their academic performance.

8.0 Extra features

8.1 Extra feature 1: tidyverse package

Instead of installing ggplot2 and dplyr packages separately, the tidyverse package was installed. The core tidyverse includes the packages that you're likely to use in everyday data analyses.

8.2 Extra feature 2: gridExtra package

The gridExtra package adds useful extensions to the grid system, with an emphasis on higher-level functions for working with grid graphic objects rather than the grid package's lower-level utilities for creating and editing specific lower-level elements of a plot. This package contains particularly useful functions for arranging and writing multiple grobs to a graphics device, as well as including tables in grid graphics objects.

8.3 Extra feature 3: Violin Graph

Violin plots are like box plots in that they display the data's kernel probability density at various values. As with standard box plots, violin plots typically include a marker for the data's median and a box indicating the interquartile range.

Violin graph was used for analyses 6.3.

8.4 Extra feature 4: Density Graph

A density plot depicts a numerical variable's distribution. It displays a variable's probability density function using a kernel density estimate. It is a smoothed version of the histogram with the same function.

The density plot was used for analyses 7.1 and 7.2 respectively.

9.0 Conclusion

The documentation was about exploring a dataset that contains various information about three-year final scores of degree students' marks. The dataset also includes many other characteristics that may or may not have influenced these students' academic performance. Some of these features include the student's school, gender, area of residence, family information, daily/weekly alcohol consumption, and many more.

The dataset was saved in CSV format in Microsoft Excel and imported into RStudio for various preprocessing techniques to produce appropriate analysis. Data exploration, manipulation, transformation, and visualization are among the techniques used.

10.0 Appendix

#Name and TP Number

#Dipta Protim Guha

#TP063351

#import dataset

```
data = read.csv("C:\\Users\\aritr\\Downloads\\student.csv", header=TRUE)
```

data

#install packages

```
install.packages("tidyverse")
```

```
install.packages("gridExtra")
```

#load Packages

```
library(tidyverse)
```

```
library(gridExtra)
```

#view table form

```
View(data)
```

#view first 6 rows of dataset

```
head(data)
```

```
#view all the column names
```

```
names(data)
```

```
#count the number of column and rows
```

```
nrow(data)
```

```
ncol(data)
```

```
#summary of dataset
```

```
summary(data)
```

```
#remove the index column from the dataset
```

```
data$index = NULL
```

```
data
```

```
#analysis 3.1 which gender group has a better average final grade
```

```
#male
```

```
average_grade_male=select(data, sex, G3) %>% filter(sex == "M") %>% summarise(mean(G3))
```

```
average_grade_male
```

```
#female
```

```
average_grade_female=select(data, sex, G3) %>% filter(sex == "F") %>%
```

```
summarise(mean(G3))
```

```
average_grade_female
```

```
#analysis 3.2 number of male and female students
```

```
count_male=filter(data, sex == "M")
```

```
nrow(count_male)
```

```
count_female=filter(data, sex == "F")
```

```
nrow(count_female)
```

```
#analysis 3.3 average of all 3 grade columns combine for male and female students
```

```
avaerage_grade_all_grades_male=sample_frac(data,1) %>%
```

```
mutate(All_Grades_out_of_60=G1+G2+G3) %>% select(sex, All_Grades_out_of_60)%>%
```

```
filter(sex == "M") %>% summarise(mean(All_Grades_out_of_60))
```

```
avaerage_grade_all_grades_male
```

```
avaerage_grade_all_grades_female=sample_frac(data,1) %>%
```

```
mutate(All_Grades_out_of_60=G1+G2+G3) %>% select(sex, All_Grades_out_of_60) %>%
```

```
filter(sex == "F") %>% summarise(mean(All_Grades_out_of_60))
```

```
avaerage_grade_all_grades_female
```

#3.4 analysis Which gender group scored the most number of distinctions?

```
distinction_male= filter(data, sex == "M", G3 >= 16)
```

```
i=nrow(distinction_male)
```

```
distinction_female=filter(data, sex == "F", G3 >= 16)
```

```
q=nrow(distinction_female)
```

```
a=c(i,q)
```

```
b=c("male","female")
```

```
pie(a,b,radius = 1, main = "Most distinctions", col = c("red","blue"),clockwise = TRUE)
```

#4.1 Analysis 1: How many students are there in each school?

```
count_student_GP= filter(data, school == "GP")
```

```
nrow(count_student_GP)
```

```
count_student_MS= filter(data, school == "MS")
```

```
nrow(count_student_MS)
```

#4.2 Analysis 2: Which school students have a better average grade.

```
data %>% select(school, G1, G2, G3) %>% mutate(all_grades=G1+G2+G3) %>% subset(school  
== "GP") %>%
```

```
summarise(mean(all_grades))
```

```
data %>% select(school, G1, G2, G3) %>% mutate(all_grades=G1+G2+G3) %>% subset(school  
== "MS") %>%
```

```
summarise(mean(all_grades))
```

```
table(data$school, data$G3)
```

#4.3 Analysis 3: Which school students have more failures in their final Grade?

```
GP=data %>% filter(school == "GP", G3<10) %>%
```

```
ggplot(aes(x=G3))+geom_histogram(colour="black",aes(fill=..count..))+
```

```
scale_fill_gradient("count", low = "green", high = "red")+
```

```
ggtitle("Student from Gabriel Pereira with score less than 10")
```

```
MS=data %>% filter(school == "MS", G3<10) %>%
```

```
ggplot(aes(x=G3))+geom_histogram(colour="black",aes(fill=..count..))+
```

```
scale_fill_gradient("count", low = "green", high = "red")+
```

```
ggtitle("Student from Mausinho da Silveira with score less than 10")
```

```
grid.arrange(GP,MS)
```

```
#Number of students and fail/students percentage.
```

```
GP_count=data %>% filter(school == "GP", G3 < 10)
```

```
MS_count=data %>% filter(school == "MS", G3 < 10)
```

```
nrow(GP_count)
```

```
nrow(MS_count)
```

```
GP_percent=246/749
```

```
round(GP_percent,digits = 2)
```

```
MS_percent=55/173
```

```
round(MS_percent,digits = 2)
```

```
#4.4 Analysis 4: Which school students have more distinctions in their final Grade?
```

```
data %>% filter(G3>=16) %>% ggplot(aes(x=school))+
```

```
  geom_bar(fill=c("red","blue"))+
```

```
  ggtitle("Number of students who achieved distinction from each school")
```

```
#Number of students and fail/students percentage.
```

```
GP_count=data %>% filter(school == "GP", G3 >= 16)
```

```
MS_count=data %>% filter(school == "MS", G3 >= 16)
```

```
nrow(GP_count)
```

```
nrow(MS_count)
```

```
GP_percent=77/749
```

```
round(GP_percent,digits = 2)
```

```
MS_percent=17/173
```

```
round(MS_percent,digits = 2)
```

#5.1 Analysis 1: Does weekend alcohol consumption affect students' final grade?

```
data %>% ggplot(aes(x=Walc, y=G3))+
```

```
  geom_point(aes(color=Walc))+
```

```
  stat_smooth(method = lm)+
```

```
  labs(title = "Relationship between weekend alcohol consumption and students' final grade",
```

```
        x= "weekend alcohol consumption (1 - very low to 5 - very high", y="Final Grade")
```

#5.2 Analysis 2: Does daily alcohol consumption affect students' final grade?

```
data %>% ggplot(aes(x=Dalc, y=G3))+
```

```
geom_point(aes(color=Dalc))+facet_wrap(~Dalc)+  
labs(title = "Relationship between daily alcohol consumption and students' final grade",  
x= "daily alcohol consumption (1 - very low to 5 - very high", y="Final Grade")
```

#5.3 Analysis 3: Does workday alcohol consumption lead to health issues?

```
data %>% ggplot(aes(x=Dalc, y=health))+geom_point(aes(color=Dalc))+facet_wrap(~Dalc)+  
labs(title = "Relationship between daily alcohol consumption and student health",  
x= "daily alcohol consumption (1 - very low to 5 - very high",  
y="Student health status (numeric: from 1 - very bad to 5 - very good")
```

#5.4 Analysis 4: Can health lead to absences which can lead to a drop in grade?

```
data %>% ggplot(aes(x=absences,  
y=G3))+geom_point(aes(shape=factor(Dalc),color=factor(Dalc)))+  
labs(title = "Relationship between absences and students' final grade",  
x= "Absences", y="Final Grade")
```

#6.1 Analysis 1: Relationship between romance and students final Grade

```
data %>%  
ggplot(aes(romantic, G3))+
```



```
geom_boxplot(aes(fill=romantic))+
labs(title = "Relationship between romance and students final Grade ",
      x= "Romantic involvement", y="Final Grade")
```

#6.2 Analysis 2: Does romance cause students to go out more and study less which can lead to drops in their final grades?

```
romantic_table=data %>% select(romantic, goout, studytime, G3) %>% group_by(romantic)
%>%
summarise(mean(goout), mean(studytime), mean(G3))
```

```
View(romantic_table)
```

#6.3 Analysis 3: Does romance causes more absences for students and does that affect their final grades?

```
data %>%
ggplot(aes(romantic, absences))+
geom_violin(aes(fill=romantic))+ stat_summary(fun=mean, geom="point", size=2,
color="yellow")+
labs(title = "Relationship between romance and absences ",
      x= "Romantic involvement", y="Number of absences")
```

#7.1 Analysis 1: Does students' location from school affect their Final Grade?

```
data %>% ggplot(aes(x=G3))+  
  geom_density(aes(fill=address))+  
  labs(title = "Density graph of relationship between students' final grade and address",  
        x= "Final Grade")
```

#7.2 Analysis 2: Does students' location from school affect their travel time?

```
data %>% ggplot(aes(x=traveltime))+  
  geom_density(aes(fill=address))+  
  labs(title = "Density graph of relationship between students' travel time and address",  
        x= "Travel time")
```

#7.3 Analysis 3: Does students' location from school affect their travel time?

```
data %>% ggplot(aes(x=traveltime, y=G3))+geom_point(aes(color=traveltime))+  
  stat_smooth(method = lm)+  
  labs(title = "Relationship between travel time and students' final grade",  
        x= "Travel Time", y="Final Grade")
```

11.0 References

- Spanton, R. (2020). An Introduction to the Pipe in R. Retrieved 29 April 2022, from <https://towardsdatascience.com/an-introduction-to-the-pipe-in-r-823090760d64#:~:text=What%20does%20the%20pipe%20do,a%20sequence%20of%20analysis%20steps.>
- Patte, K., Qian, W., & Leatherdale, S. (2019). Binge drinking and academic performance, engagement, aspirations, and expectations: a longitudinal analysis among secondary school students in the COMPASS study. Retrieved 9 May 2022, from <https://www.canada.ca/en/public-health/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-37-no-11-2017/binge-drinking-academic-performance-engagement-aspirations-expectations-longitudinal-analysis-secondary-school-students-compass-study.html>
- ggplot2 violin plot : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. (2018). Retrieved 11 May 2022, from <http://www.sthda.com/english/wiki/ggplot2-violin-plot-quick-start-guide-r-software-and-data-visualization>
- ggplot2 density plot : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA. (2019). Retrieved 11 May 2022, from <http://www.sthda.com/english/wiki/ggplot2-density-plot-quick-start-guide-r-software-and-data-visualization>
- Tidyverse packages. (2020). Retrieved 11 May 2022, from <https://www.tidyverse.org/packages/>
- Roger D. Peng, a. (2022). 4.5 The grid Package | Mastering Software Development in R. Retrieved 11 May 2022, from <https://bookdown.org/rdpeng/RProgDA/the-grid-package.html>

