INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APD2F2106CS(IS)

HAND OUT DATE: 5 JULY 2021

HAND IN DATE: 23 AUGUST 2021

WEIGHTAGE: 50%

Name	TP Number
Sia De Long	TP060810

Table of Contents

Introduction

A dataset of degree students' academic performance is given with various type of attributes related to the students' personal details, academic performance, family backgrounds and daily routines. The dataset will be consisting of total 33 column and 922 rows, so different technique will be use to conduct data exploration, data manipulation and data exploration in order to discover information and knowledge from it. The aim is to find the relationship between attributes and students' performance and it will be visualise using a suitable type of graph. In all these processes, R programming language will be used to implement in them on R studio platform because R is very useful and emphasising on statistical computing and graphics so it is definitely suited for the tasks (Debbie & Sara, 2017).

Assumptions

A few assumptions are made for this analysis in order to easier the process of analysis and also make the result turn out to be more sense. Some of them are made in the beginning of the analysis while some of them are made after found out some weird result is coming out after the analysis.

1. Health status includes physically and mentally such as stress.
2. Co-curricular activity is referring to sport clubs or liberal arts related clubs only instead of society service or school event.
3. Internet at home attribute has the difference in bandwidth speed but it is not stated in this dataset.
4. Extra educational support only refers to teaching

Install and Load Packages

First and foremost, there are a few directories or libraries need to be installed and imported before any analysis in the source code in order to enable some function in the source code.

```
# ======================== Install Packages ========================
install.packages("ggplot2")
install.packages("dplyr")
install.packages("cowplot")
```

Figure 1: Install Packages

Source code above showing the install.packages() function which is to download package from the internet to your computer. In this analysis, three packages will be used which are ggplot2, dplyr and cowplot, ggplot2 is mainly used to visualise all the graph in this analysis, dplyr is used for data manipulation and transformation and cowplot is used to generate display for multiple graphs at once to ease the process of data exploration and analysis in the source code interface. After all packages are downloaded successfully, the packages must be loaded to the library by using library() function as shown on the figure below.

```
# ======================== Load Libraries ========================
library(ggplot2)
library(dplyr)
library(cowplot)
```

Figure 2: Load Packages

Data Import

```
# ======================= Import Datasets ========================
# Ensure student.csv can be found in the files tab of Rstudio first
setwd("C:/Users/Sia De Long/Desktop/PFDA assignment")
raw = readLines("student.csv")
raw = gsub("\"", "", raw)
students = read.table(text = raw, sep = ';', quote = ';', header = T)
```

Figure 3: Import Datasets

As mentioned, a dataset is provided but it still needs to import to the R studio as a data frame variable for further calculation. Firstly, setwd() function is used to redirect the working directory for this R script, so the file path can be change according to different user. After making sure the student.csv file is also included in the working directory, raw will act as a character vector variable for storing the data temporary by using readLines(), so every lines in the csv file will be separated into single element in the vector instead of directly using read.csv(). The reason of doing this is because there are too many unnecessary double quotes in the data and will causing the vector treat numeric data as character data, hence the double quotes must be removed before move the whole data to the actual variable. For the purpose of removing double quotes, gsub() function will be used by mentioning quote with escape symbol (\) and replace it to empty character. Lastly, the raw will be read to students as data frame variable using read.table() while separator and quote is semicolon and setting the header to True, so column header and column data is separated perfectly and having a header on top of them. Before going to data pre-processing phase, an overview of the data can be displayed using str() for data type, summary() for average data as shown below and View() for data levels when hovering mouse to the column header.

```
# Check data
str(students)
summary(students)
View(students)
```

Figure 4: Check Data

```
> str(students)
'data.frame':   922 obs. of  33 variables:
 $ school    : chr  "GP" "GP" "GP" "GP" ...
 $ sex       : chr  "F" "F" "F" "F" ...
 $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
 $ address   : chr  "U" "U" "U" "U" ...
 $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
 $ Pstatus   : chr  "A" "T" "T" "T" ...
 $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
 $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
 $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
 $ Fjob      : chr  "teacher" "other" "other" "services" ...
 $ reason    : chr  "course" "course" "other" "home" ...
 $ guardian  : chr  "mother" "father" "mother" "mother" ...
 $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
 $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
 $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
 $ schoolsup : chr  "yes" "no" "yes" "no" ...
 $ famsup    : chr  "no" "yes" "no" "yes" ...
 $ paid      : chr  "no" "no" "yes" "yes" ...
 $ activities: chr  "no" "no" "no" "yes" ...
 $ nursery   : chr  "yes" "no" "yes" "yes" ...
 $ higher    : chr  "yes" "yes" "yes" "yes" ...
 $ internet  : chr  "no" "yes" "yes" "yes" ...
 $ romantic  : chr  "no" "no" "no" "yes" ...
 $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
 $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
 $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
 $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
 $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
 $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
 $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
 $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
 $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
 $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Figure 5: Students Data Type

```
> summary(students)
    school              sex                 age          address
 Length:922         Length:922         Min.   :15.00   Length:922
 Class :character   Class :character   1st Qu.:16.00   Class :character
 Mode  :character   Mode  :character   Median :17.00   Mode  :character
                                       Mean   :16.74
                                       3rd Qu.:18.00
                                       Max.   :22.00
   famsize            Pstatus              Medu            Fedu
 Length:922         Length:922         Min.   :0.000   Min.   :0.000
 Class :character   Class :character   1st Qu.:2.000   1st Qu.:2.000
 Mode  :character   Mode  :character   Median :3.000   Median :2.500
                                       Mean   :2.753   Mean   :2.536
                                       3rd Qu.:4.000   3rd Qu.:3.000
                                       Max.   :4.000   Max.   :4.000
     Mjob               Fjob              reason            guardian
 Length:922         Length:922         Length:922         Length:922
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


  traveltime      studytime        failures       schoolsup
 Min.   :1.000   Min.   :1.000   Min.   :0.0000   Length:922
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   Class :character
 Median :1.000   Median :2.000   Median :0.0000   Mode  :character
 Mean   :1.457   Mean   :2.037   Mean   :0.3319
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
 Max.   :4.000   Max.   :4.000   Max.   :3.0000
    famsup             paid            activities          nursery
 Length:922         Length:922         Length:922         Length:922
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
   higher            internet           romantic            famrel
Length:922         Length:922         Length:922         Min.   :1.000
Class :character   Class :character   Class :character   1st Qu.:4.000
Mode  :character   Mode  :character   Mode  :character   Median :4.000
                                                         Mean   :3.949
                                                         3rd Qu.:5.000
                                                         Max.   :5.000
   freetime         goout            Dalc            Walc           health
Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000
Median :3.000   Median :3.000   Median :1.000   Median :2.000   Median :4.000
Mean   :3.252   Mean   :3.092   Mean   :1.496   Mean   :2.293   Mean   :3.565
3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
   absences            G1              G2              G3
Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
Median : 4.000   Median :11.00   Median :11.00   Median :11.00
Mean   : 5.517   Mean   :10.94   Mean   :10.77   Mean   :10.46
3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
Max.   :75.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
```

Figure 6: Students Data Summary

These two figures showing that there are no wrong data type exists in this data set everything is as expected and there is no any null value to be replaced or removed.

Data Pre-processing

After everything has been make sure it is suitable for analysis, then the things left for data pre-processing is to change the data and header to a more suitable string name. Function names() is used to change every header name to suitable and clearer definition string while data will be modified using R property of array specify, so it will only make changes to every index that reach the condition as shown by figures below.

```
# Rename dataset header
names(students) = c("School_Name", "Sex", "Age", "Living_Area",
                    "Family_Size", "Cohabitation_Status", "Mother_Education",
                    "Father_Education", "Mother_Job_Type",  "Father_Job_Type",
                    "Reason_Choosing_School", "Guardian",
                    "Travel_Time", "Study_Time", "Past_Failures",
                    "School_Extra_EduSup", "Family_Extra_EduSup",
                    "Paid_Extra_Class", "Cocurricular_Activity",
                    "Attended_Nursery_School", "Wanting_Higher_Education",
                    "Internet_At_Home", "Couple_Relationship",
                    "Family_Relationship", "Free_Time", "Hang_Out",
                    "WeekDay_Alcohol_Consumed", "Weekend_Alcohol_Consumed",
                    "Health_Status", "School_Absences", "G1", "G2", "G3")
```

Figure 7: Rename Dataset Header

```
# Change suitable string for sex attribute
students$Sex[students$Sex == "M"] = "Male"
students$Sex[students$Sex == "F"] = "Female"

# Change suitable string for mother job type attribute
students$Mother_Job_Type[students$Mother_Job_Type == "teacher"] = "Teacher"
students$Mother_Job_Type[students$Mother_Job_Type == "health"] = "Health Care"
students$Mother_Job_Type[students$Mother_Job_Type == "services"] = "Civil Services"
students$Mother_Job_Type[students$Mother_Job_Type == "at_home"] = "Work At Home"
students$Mother_Job_Type[students$Mother_Job_Type == "other"] = "Other"

# Change suitable string for father job type attribute
students$Father_Job_Type[students$Father_Job_Type == "teacher"] = "Teacher"
students$Father_Job_Type[students$Father_Job_Type == "health"] = "Health Care"
students$Father_Job_Type[students$Father_Job_Type == "services"] = "Civil Services"
students$Father_Job_Type[students$Father_Job_Type == "at_home"] = "Work At Home"
students$Father_Job_Type[students$Father_Job_Type == "other"] = "Other"

# Change suitable string for family size attribute
students$Family_Size[students$Family_Size == "GT3"] = "Greater Than 3"
students$Family_Size[students$Family_Size == "LE3"] = "Less Than 3"

#Change suitable string for living area attribute
students$Living_Area[students$Living_Area == "R"] = "Rural Area"
students$Living_Area[students$Living_Area == "U"] = "Urban Area"

#Change suitable string for cohabitation status attribute
students$Cohabitation_Status[students$Cohabitation_Status == "A"] = "Living Apart"
students$Cohabitation_Status[students$Cohabitation_Status == "T"] = "Living Together"

#Change suitable string for reason choosing school attribute
students$Reason_Choosing_School[students$Reason_Choosing_School == "course"] = "Interest in Course"
students$Reason_Choosing_School[students$Reason_Choosing_School == "home"] = "Close to Home"
students$Reason_Choosing_School[students$Reason_Choosing_School == "other"] = "Other"
students$Reason_Choosing_School[students$Reason_Choosing_School == "reputation"] = "School Reputation"
```

Figure 8: Modify Data

Question 1: What is the personal attribute affecting student performance?

This question will be focusing on analysing and get a relationship between personal attribute and student performance so the result can help to predict an expected performance in academic for multiple kind of individual student. Therefore, some main attributes are getting targeted for this question including Sex, Age, Reason_Choosing_School, Study_Time", Wanting_Higher_Education and School_Absences. Some of the cause and inter-related attributes may also be found in the process of analysis, hence this question is worth to take a look into.

Data Exploration

Since View() function is used beforehand, so there is no need to use any other function to determine the range or levels of the data. Rather, some very direct and simple graphs are used to get some insight and idea about the brief relationship between each other, every attribute will be put in the same graph with G1, G2 and G3 in order to not miss any interesting relationship. Different graphs will be used depends on whether the data is in continuous or discrete, therefore in this question, stacked histogram, count graph and boxplot are used shown as below.

```
# Sex
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Sex)) +
  labs(title = "Frequency Distribution of G1 Result by Sex",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Sex)) +
  labs(title = "Frequency Distribution of G2 Result by Sex",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Sex)) +
  labs(title = "Frequency Distribution of G3 Result by Sex",
       x = "G3",
       y = "Number of Students")
```

Figure 9: Sex Attribute Exploration

```
# Age
ggplot(students, aes(x = G1, y = Age)) +
  geom_count() +
  labs(title = "Frequency Distribution of G1 Result by Age",
       x = "G1",
       y = "Age")
ggplot(students, aes(x = G2, y = Age)) +
  geom_count() +
  labs(title = "Frequency Distribution of G2 Result by Age",
       x = "G2",
       y = "Age")
ggplot(students, aes(x = G3, y = Age)) +
  geom_count() +
  labs(title = "Frequency Distribution of G3 Result by Age",
       x = "G3",
       y = "Age")
```

Figure 10: Age Attribute Exploration

```
# Reason_Choosing_School
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Reason_Choosing_School)) +
  labs(title = "Frequency Distribution of G1 Result by Reason Choosing School",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Reason_Choosing_School)) +
  labs(title = "Frequency Distribution of G2 Result by Reason Choosing School",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Reason_Choosing_School)) +
  labs(title = "Frequency Distribution of G3 Result by Reason Choosing School",
       x = "G3",
       y = "Number of Students")
```

Figure 11: Reason_Choosing_School Attribute Exploration

```
# Study_Time
ggplot(students, aes(x = G1, y = Study_Time)) +
  geom_boxplot(aes(group = Study_Time)) +
  labs(title = "Frequency Distribution of G1 Result by Study Time",
       x = "G1",
       y = "Study Time")
ggplot(students, aes(x = G2, y = Study_Time)) +
  geom_boxplot(aes(group = Study_Time)) +
  labs(title = "Frequency Distribution of G2 Result by Study Time",
       x = "G2",
       y = "Study Time")
ggplot(students, aes(x = G3, y = Study_Time)) +
  geom_boxplot(aes(group = Study_Time)) +
  labs(title = "Frequency Distribution of G3 Result by Study Time",
       x = "G3",
       y = "Study Time")
```

Figure 12: Study_Time Attribute Exploration

```
# Wanting_Higher_Education
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Wanting_Higher_Education)) +
  labs(title = "Frequency Distribution of G1 Result by Wanting Higher Education",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Wanting_Higher_Education)) +
  labs(title = "Frequency Distribution of G2 Result by Wanting Higher Education",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Wanting_Higher_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Wanting Higher Education",
       x = "G3",
       y = "Number of Students")
```

Figure 13: Wanting_Higher_Education Attribute Exploration

```
# School_Absences
ggplot(students, aes(x = G1, y = School_Absences)) +
  geom_count() +
  labs(title = "Frequency Distribution of G1 Result by School Absences",
       x = "G1",
       y = "School Absences")
ggplot(students, aes(x = G2, y = School_Absences)) +
  geom_count() +
  labs(title = "Frequency Distribution of G2 Result by School Absences",
       x = "G2",
       y = "School Absences")
ggplot(students, aes(x = G3, y = School_Absences)) +
  geom_count() +
  labs(title = "Frequency Distribution of G3 Result by School Absences",
       x = "G3",
       y = "School Absences")
```

Figure 14: School_Absences Attribute Exploration

Data Manipulation and Transformation

```
# Main data sets
Question1Data = students %>%
  mutate(Final_Result = round((G1+G2+G3)/60*100, digits = 2)) %>%
  select(Sex, Age, Reason_Choosing_School, Study_Time,
         Wanting_Higher_Education, School_Absences, Final_Result) %>%
  arrange(Final_Result)
# Arrange columns name
  Question1Data = Question1Data %>%
  select(order(colnames(Question1Data)))

View(Question1Data)
str(Question1Data)
summary(Question1Data)
```

Figure 15: Question 1 Data Manipulation and Transformation

A specified sub-dataset named Question1Data is created in this phase by using the dataset after pre-processing, piping is widely used right here to show a more readable source code. First of all, an overall performance in three years result is produced in new a column named Final_Result by using mutate() function with a simple mathematical formula, before create the column the calculation result will round up to 2 decimal place using round() function by giving desired decimal place to digits parameter. After that, every main attribute stated in data exploration will be selected so that no other column will be in the new dataset, then by using arrange() function the dataset will be sorted ascendingly following Final Result in rows. Lastly, select() function is used again with order() and colnames() function to sort the column this time with alphabet ascendingly. The fact that sorting rows and column is not executed together is because Final_Result is not mutate completely yet, so order() function cannot find the column name. In the end of data manipulation and data exploration, View(), summary() and str() function also can be used to display a more simple details of the data as shown below.

```
> str(Question1Data)
'data.frame':   922 obs. of  7 variables:
 $ Age                    : int  16 16 17 19 17 19 16 18 16 18 ...
 $ Final_Result           : num  6.67 6.67 8.33 8.33 8.33 8.33 10 10 10 10 ...
 $ Reason_Choosing_School  : chr  "Interest in Course" "Interest in Course" "Close to Home" "Clos
e to Home" ...
 $ School_Absences        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Sex                    : chr  "Female" "Female" "Male" "Male" ...
 $ Study_Time             : int  1 1 1 1 1 1 1 2 1 2 ...
 $ Wanting_Higher_Education: chr  "yes" "yes" "yes" "no" ...
```

Figure 16: Question 1 Data datatype

```
> summary(Question1Data)
      Age         Final_Result   Reason_Choosing_School School_Absences
 Min.   :15.00   Min.   : 6.67   Length:922             Min.   : 0.000
 1st Qu.:16.00   1st Qu.:41.67   Class :character       1st Qu.: 0.000
 Median :17.00   Median :53.33   Mode  :character       Median : 4.000
 Mean   :16.74   Mean   :53.62                          Mean   : 5.517
 3rd Qu.:18.00   3rd Qu.:66.67                          3rd Qu.: 8.000
 Max.   :22.00   Max.   :96.67                          Max.   :75.000
    Sex              Study_Time    Wanting_Higher_Education
 Length:922         Min.   :1.000   Length:922
 Class :character   1st Qu.:1.000   Class :character
 Mode  :character   Median :2.000   Mode  :character
                    Mean   :2.037
                    3rd Qu.:2.000
                    Max.   :4.000
```

Figure 17: Question 1 Data summary

Analysis 1: Determine the Relationship between Final Result and Age in Different Sex

Data Visualisation

```
# Population Pyramid Graph
studentFrequency = Question1Data %>%
  ggplot(aes(x = Age, fill = Sex)) +
  geom_bar(data = subset(students, Sex == "Male")) +
  geom_bar(data = subset(students, Sex == "Female"), aes(y=..count..*(-1)))  +
  scale_y_continuous(breaks=seq(-150,150,10),labels=abs(seq(-150, 150,10))) +
  coord_flip() +
  labs(title = "Number of Female and Male students in diferent age",
       x = "Age",
       y = "Frequency")
```

Figure 18: distribution of sex and age code

A population pyramid graph is created to display the distribution of sex and age in this dataset. For mappings, Age will be the x-axis and bar filled colour will be depend on sex. Female will be times -1 to the count frequency to make it to the opposite site, then scale_y_continuos() is to break y-axis to desired range and label it by using absolute function to make sure no negative number display on the graph. Lastly, coord_flip() is used to flip 90 degree of the graph so that the graph is a population pyramid and some labels is given using labs(). This graph is saved into studentFrequency variable for later used.

```
# Boxplot and Jitter plot
Final_Age_Sex = Question1Data %>%
  ggplot(aes(x = Age, y = Final_Result)) +
  geom_boxplot(aes(group = Age)) +
  geom_jitter(alpha = 0.4, aes(color = Sex)) +
  facet_wrap(~Sex) +
  labs(title = "Relationship Between Final Result And Age In Different Sex",
       x = "Age",
       y = "Final Result(%)")
```

Figure 19: relationship between final result and age in different sex code

A combination of boxplot and jitter plot is created to display the relationship between final result and age in different sex. For mappings, Age will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Age while jitter will be half transparent on top of boxplot by changing the alpha value and colour it followed by sex. Then, the graph will be separate to two by sex using facet_wrap() and some labels is given using labs(). This graph is saved into Final_Age_Sex variable for later used.

```
# Display both graph
plot_grid(studentFrequency, Final_Age_Sex, nrow = 2)
```

Figure 20: plot grid code

plot_grid() is used to display both studentFrequency and Final_Age_Sex at the same time with desired dimension to relating them together. Hence the visual as below.

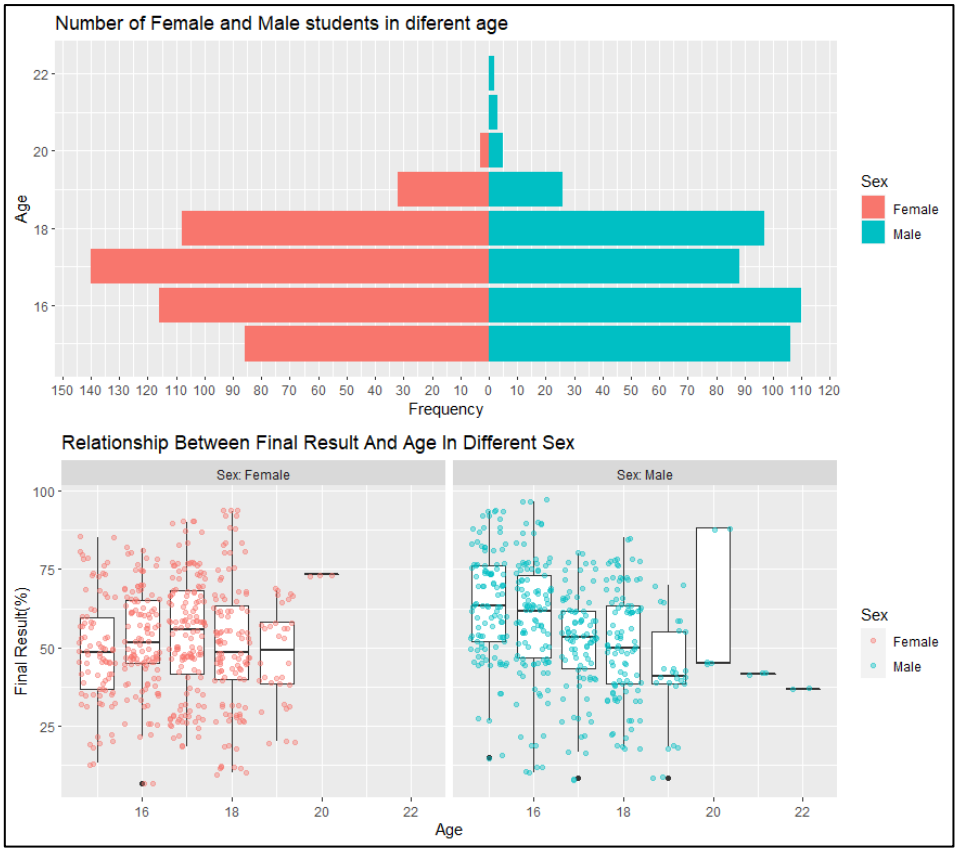

Figure 21: plot grid graph

The summary of the graph include:

1. Female students tend to produce higher performance in age of 17
2. Male students tend to produce higher performance in age of 15

Explanation:

we can see that the performance of male students is decreasing following the increasing in age which is not the case for female where all female in age of 20 get a high performance in academic, male in age of 20 also have exception of high performance but it is rare. Female in young age might suffer from puberty earlier than male according to the study (PROFESSOR ANDREW & KATHARINE, 2017), while it does not affect too much for male, turn out male have better performance at minimum age, then it will go down may be caused by playful. A sudden drop in age of 18 and 19 for female is because of the frequency of female student in that age decreased significantly.

Analysis 2: Determine the Relationship between Final Result and School Absences

Data Visualisation

```
# Correlation of age and absences (regression line)
ggplot(Question1Data, aes(x = Age, y = School_Absences)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relationship Between Age And Absences",
       x = "Age",
       y = "Absences to School") +
  theme(text=element_text(size = 16))
```

Figure 22: relationship between final result and age in different sex code

A regression line graph is created to display the relationship between final result and age in different sex. For mappings, Age will be the x-axis and School_Absences be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). Hence the visual as below.

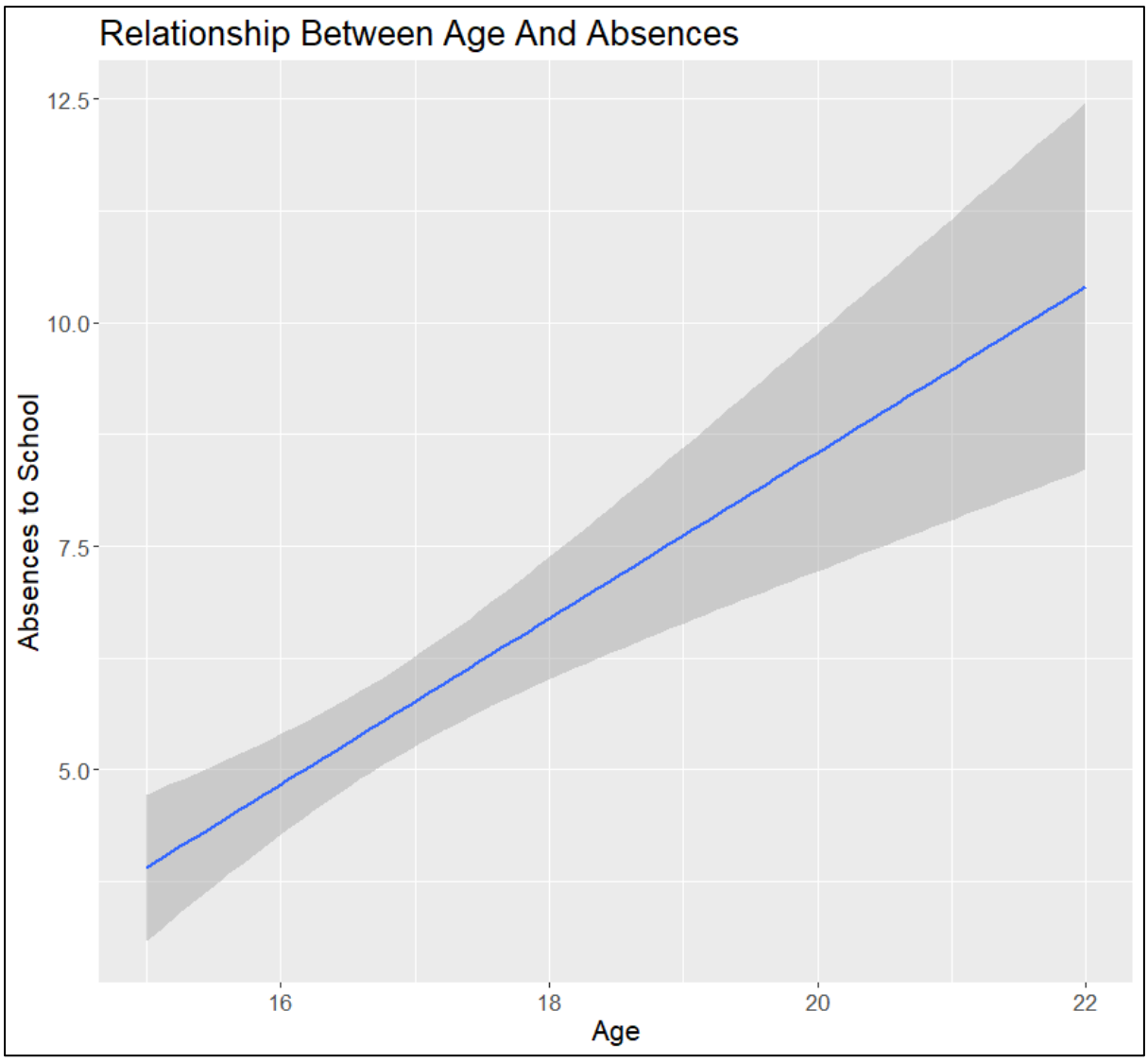

Figure 23: relationship between final result and age in different sex graph

The summary of the graph include:

1. The higher the student's age is, the greater number of absences to the school.
2. 17 years old absences day range is the smallest, so the prediction is more accurate.
3. 22 years old absences day range is the largest, so the prediction may be not accurate.

Explanation:

Some of the higher age students may have part time job aside from degree study and also may have more responsibility compared to younger students such as maturity, involvement, example, mentoring and watchfulness according to a study (Tim, 2017), so they will sometimes absence to the school.

```
# Density 2D graph
ggplot(Question1Data, aes(x = Final_Result, y = School_Absences)) +
  geom_density2d() +
  stat_density2d(aes(fill = stat(level)), geom="polygon") +
  labs(title = "Density of Final Result Affect by Frequency of Absences to School",
       x = "Final Result",
       y = "Density of Absences to School") +
  theme(text=element_text(size = 16))
```

Figure 24: density of final result affect by frequency of absences to school code

A density graph is created to display the density of final result affect by frequency of absences to school. For mappings, Final_Result will be the x-axis and School_Absences be the y-axis. Level colour is filled by overriding the area of every level to polygon. Then, the graph's label is given using labs(). Hence the visual as below.

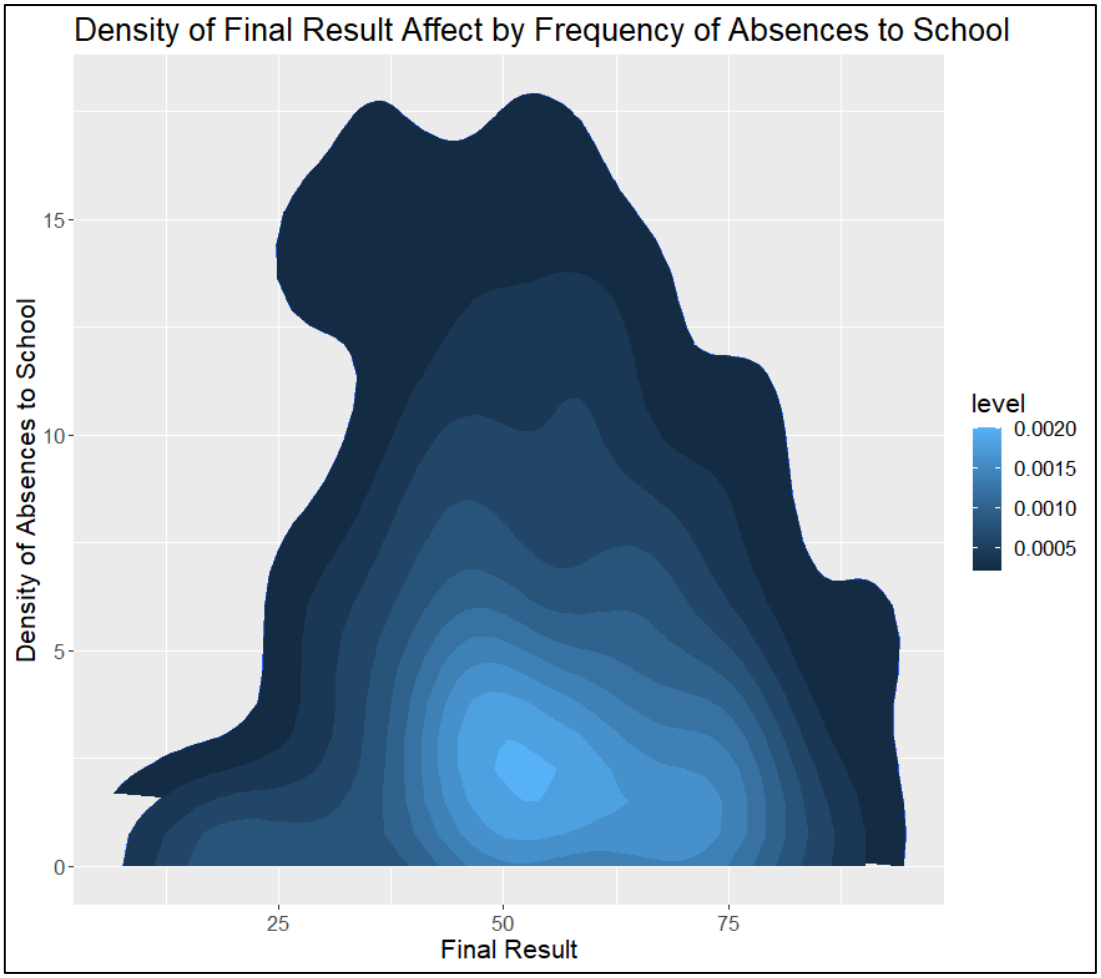

Figure 25: density of final result affect by frequency of absences to school graph

The summary of the graph include:

1. Around 50 marks have the most spread-out number of absences to the school.
2. Students with frequent number of absences to the school does not mean that student will fail the academic degree.
3. Most of the student absence to school around 2 days and get final result around 50 marks

Explanation:

When absences so many days and the final result still did not get to very low, that mean student can get the exam pass easily study own-self and without attend to school but to get another 50 marks or get to distinction, teacher will become a very important role for that such as cover much further and deeper topic. In this graph, we can conclude that school absences do not affect performance so much.

Analysis 3: Determine the Relationship between Final Result and Study Time

Data Visualisation

```
# Unstacked Bar graph
ggplot(Question1Data, aes(x = Study_Time)) +
  geom_bar(aes(fill = Sex), position = "dodge") +
  labs(title = "Number of Students in Every Level of Study Time",
       x = "Study Time",
       y = "Frequency") +
  facet_wrap(~Age, labeller = label_both) +
  theme(text=element_text(size = 16))
```

Figure 26: number of students in every level of study time code

An unstacked bar graph is created to display the number of students in every level of study time. For mappings, Study_Time will be the x-axis. Bar filled colour is depends on the sex and using dodge in position value to unstacked the bar. Then, the graph will be separate to eight by age using facet_wrap() and some label is given using labs(). Hence the visual as below.

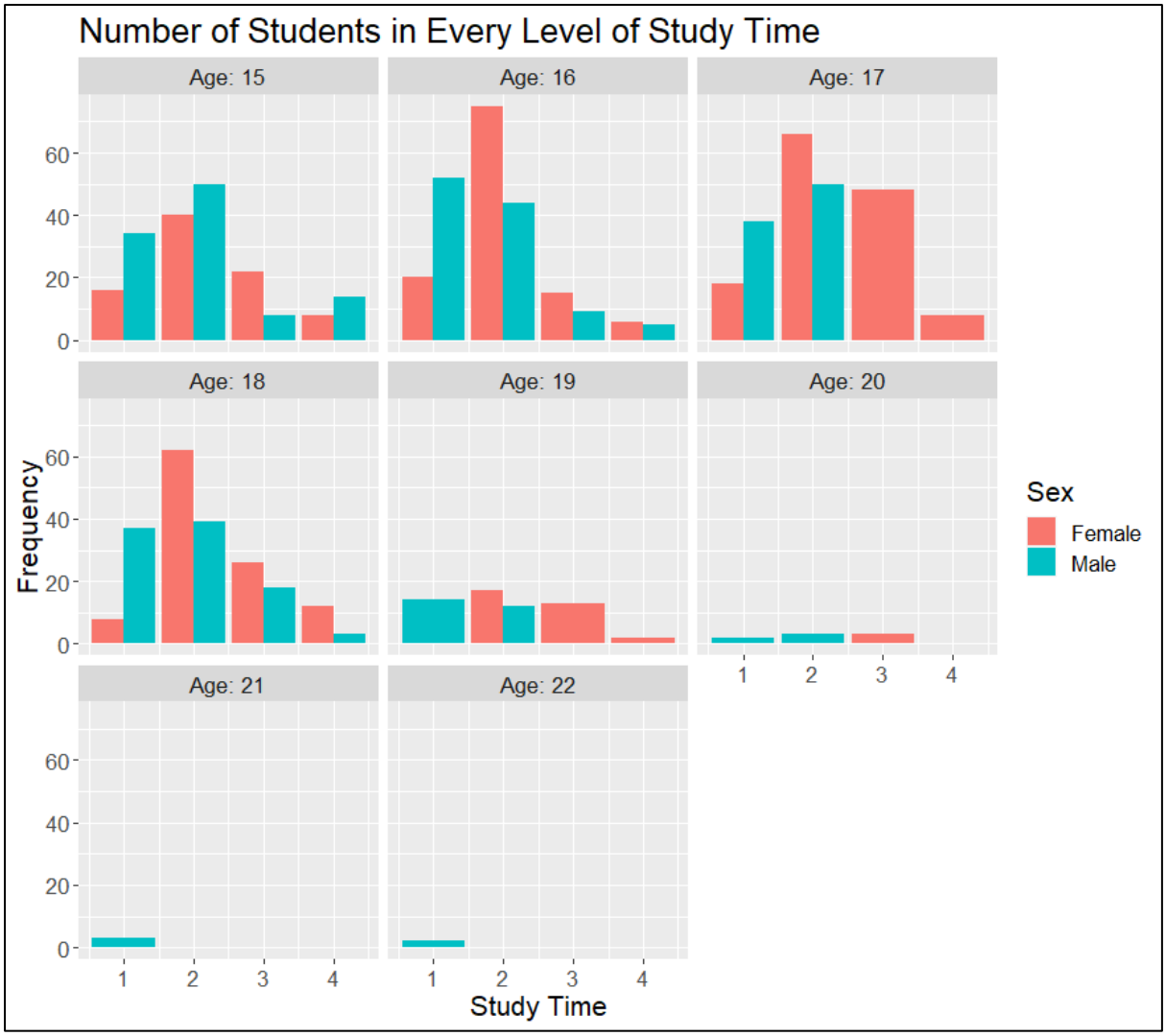

Figure 27: number of students in every level of study time graph

The summary of the graph include:

1. Male students study harder than female on the minimum age but the hardworking reduce as the age go higher.
2. Female students start to put more time on study when the age is higher.

Explanation:

Male students tend to be more playful so they cannot adhere to force themselves to study for a long period so the study time maintain a similar level in every age. However, female students put more time on study compared to male students except the minimum age.

```
# Correlation of study time and final result (regression line)
ggplot(Question1Data, aes(x = Study_Time, y = Final_Result)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relationship Between Final Result And Study Time",
       x = "Study Time",
       y = "Final Result") +
  scale_x_continuous(labels=c("low", "below average", "above average", "high"),
                     breaks=1:4) +
  theme(text=element_text(size = 16))
# -the more the study time is, the higher the performance in the exam
```

Figure 28: relationship between final result and study time code

A regression line graph is created to display the relationship between final result and study time. For mappings, Study_Time will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

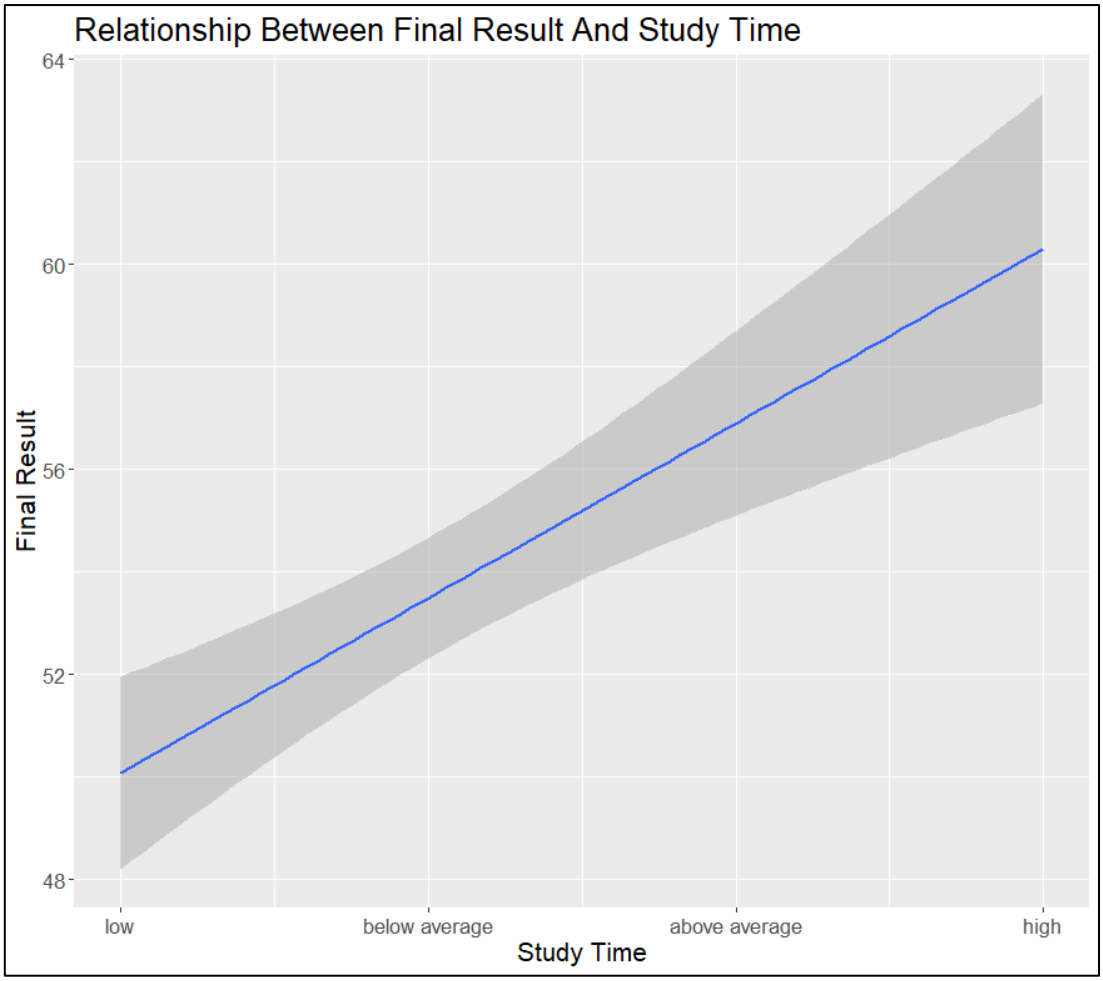

Figure 29: relationship between final result and study time graph

The summary of the graph include:

1. The more the study time is, the higher the performance in the exam.
2. Below average study time have the smallest range of final result, so the prediction is more accurate.
3. High study time have the largest range of final result, so the prediction may be not accurate.

Explanation:

When study time is high, the students can always revise to the topic and be ready to answer any questions. Not only that, the student also might discover something deeper and improve the understanding to the topic so they can perform a better result than other students. In this graph, we can observe that study time will improve performance significantly from the y-axis value range.

Analysis 4: Determine the Relationship between Final Result and Desired of Wanting Higher Education

Data Visualisation

```
# Boxplot and Count Plot
ggplot(Question1Data, aes(x = Wanting_Higher_Education, y = Final_Result)) +
  geom_boxplot(aes(group = Wanting_Higher_Education)) +
  geom_count(alpha = 0.2) +
  labs(title = "Relationship Between Final Result And Desired of Wanting Higher Education"
       x = "Desired of Wanting Higher Education",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 30: relationship between final result and desired of wanting higher education code

A combination of boxplot and count plot is created to display the relationship between final result and desired of wanting higher education. For mappings, Wanting_Higher_Education will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Wanting_Higher_Education while count will be half transparent on top of boxplot by changing the alpha value. Then, the graph's label is given using labs(). Hence the visual as below.

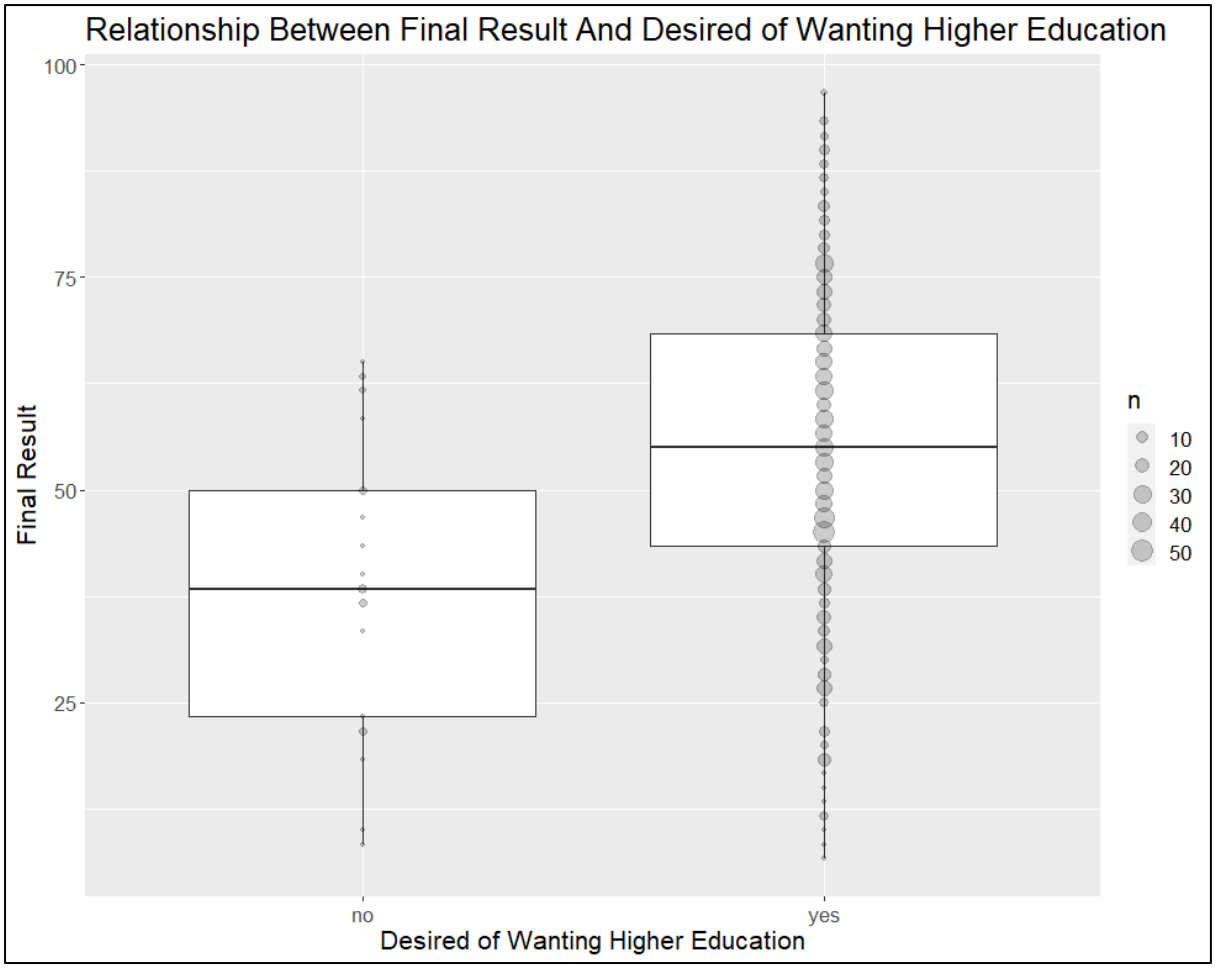

Figure 31: relationship between final result and desired of wanting higher education graph

The summary of the graph include:

1. If the student wants to get to a higher education level then it will definitely motivate that student to obtain a good result to obtain a better result or at least pass.
2. Most of the degree students have the desired of wanting higher education.

Explanation:

Normally, as long as the student can pass the previous education level then the opportunity for higher education is available for the student. Higher result will only obtain extra benefits, hence in this graph most of the students scored passing marks which mean the desired of wanting higher education do not affect student performance directly but will motivate them to pass the academic degree since higher education level is getting more and more important in 21st century according to a study (Vista College, 2021).

Analysis 5: Determine the Relationship between Final Result and Reason Choosing School

Data Visualisation

```
# Boxplot and Count Plot
ggplot(Question1Data, aes(x = Reason_Choosing_School, y = Final_Result)) +
  geom_boxplot(aes(group = Reason_Choosing_School)) +
  geom_count(alpha = 0.2) +
  labs(title = "Relationship Between Final Result And Reason Choosing School",
       x = "Reason Choosing School",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 32: relationship between final result and reason choosing school code

A combination of boxplot and count plot is created to display the relationship between final result and reason choosing school. For mappings, Reason_Choosing_School will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Reason_Choosing_School while count will be half transparent on top of boxplot by changing the alpha value. Then, the graph's label is given using labs(). Hence the visual as below.

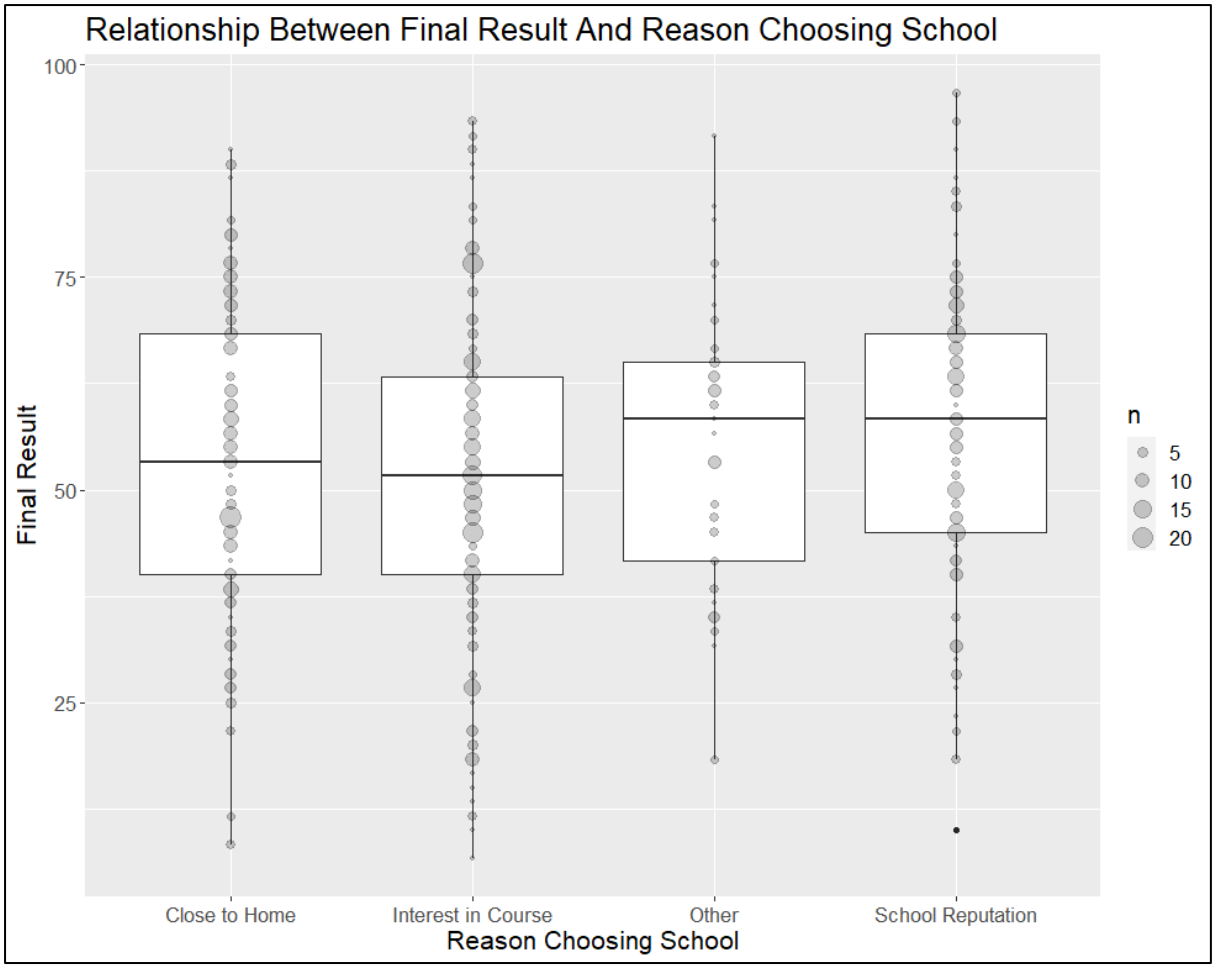

Figure 33: relationship between final result and reason choosing school graph

The summary of the graph include:

1. Other and School Reputation reason have the highest median.
2. Close to home have the highest third quarter and lowest first quarter.
3. Most students choose school with reason of interest in some particular course.
4. Close to Home, Interest in Course and School Reputation is having a close normal distribution in final result.
5. Other is having left skewness distribution in final result

Explanation:

If the school is close to home, then there will be much time saved from the transportation so the time can used to study more or sleep enough time to have fresh mind. Therefore, it has a similar average compared to interest in course but higher third quarter. In this graph, we can observe that there are not many differences between each reason, so reason choosing school only can slightly improve performance indirectly where close to home is the highest.

Question 1 Conclusion

In conclusion, Different sex will have higher performance in different age, it may cause by puberty, part time work or mature aspects. Absences does not affect to result too much but it is the opposite for study time. Besides that, self-motivation for students is important for student's performance.

For summarise, a table is created to shows the relation between each attributes and performance produced in this whole question analysis for a simple understanding of each studied attributes as shown below.

Attributes	Affect to Performance
Sex & Age	High
Reason_Choosing_School	Slightly
Study_Time	High
Wanting_Higher_Education	High
School_Absences	Indirectly

Attributes	Improve Performance
Close to home as the reason choosing school	Slightly
High study time	High
Want to have a higher education	High

In my opinion, sex and age cannot be changed easily, so it is not an attribute to be improved with but students can read more book related to the principles of the society to be more mature and also can try to balance the part time working and study time. Besides that, school can hold some campaigns to push students more and form a stronger self-motivation.

Question 2: Do the past result affect the following result?

This question will be focusing on analysing and get a relationship between past result and student performance and why if so, past result here indicating not only past failure attribute but also last year grade. Therefore, some main attributes are getting targeted for this question including Past_Failures, G1, G2, G3. Some of the cause and inter-related attributes may also be found in the process of analysis, hence this question is worth to take a look into.

Data Exploration

Since View() function is used beforehand, so there is no need to use any other function to determine the range or levels of the data. Rather, some very direct and simple graphs are used to get some insight and idea about the brief relationship between each other, every related attribute will be used to try to study with any of the past result in order to not miss any interesting relationship. Different graphs will be used depends on whether the data is in continuous or discrete, therefore in this question, count graph and boxplot are used shown as below.

```
# Past_Failures
ggplot(students, aes(x = G1, y = Past_Failures)) +
  geom_boxplot(aes(group = Past_Failures)) +
  labs(title = "Frequency Distribution of G1 Result by Past_Failures",
       x = "G1",
       y = "Past Failures")
ggplot(students, aes(x = G2, y = Past_Failures)) +
  geom_boxplot(aes(group = Past_Failures)) +
  labs(title = "Frequency Distribution of G2 Result by Past_Failures",
       x = "G2",
       y = "Past Failures")
ggplot(students, aes(x = G3, y = Past_Failures)) +
  geom_boxplot(aes(group = Past_Failures)) +
  labs(title = "Frequency Distribution of G3 Result by Past_Failures",
       x = "G3",
       y = "Past Failures")
```

Figure 34: Past_Failures Exploration

```
# G1
ggplot(students, aes(x = G2, y = G1)) +
  geom_count() +
  labs(title = "Frequency Distribution of G2 Result by G1",
       x = "G2",
       y = "G1")
ggplot(students, aes(x = G3, y = G1)) +
  geom_count() +
  labs(title = "Frequency Distribution of G3 Result by G1",
       x = "G3",
       y = "G1")
```

Figure 35: G1 Exploration

```
# G2
ggplot(students, aes(x = G3, y = G2)) +
  geom_count() +
  labs(title = "Frequency Distribution of G3 Result by G2",
       x = "G3",
       y = "G2")
```

Figure 36: G2 Exploration

Data Manipulation and Transformation

```
# Main data sets
Question2Data = students %>%
  mutate(Final_Result = round((G1+G2+G3)/60*100, digits = 2)) %>%
  select(Age, Sex, Past_Failures, G1, G2, G3, Reason_Choosing_School,
         Final_Result)%>%
  arrange(Final_Result)
# Arrange columns name
Question2Data = Question2Data %>%
  select(order(colnames(Question2Data)))

View(Question2Data)
str(Question2Data)
summary(Question2Data)
```

Figure 37: Question 2 Data Manipulation and Transformation

A specified sub-dataset named Question2Data is created in this phase by using the dataset after pre-processing, piping is widely used right here to show a more readable source code. First of all, an overall performance in three years result is produced in new a column named Final_Result by using mutate() function with a simple mathematical formula, before create the column the calculation result will round up to 2 decimal place using round() function by giving desired decimal place to digits parameter. After that, every main attribute stated in data exploration will be selected so that no other column will be in the new dataset, then by using arrange() function the dataset will be sorted ascendingly following Final Result in rows. Lastly, select() function is used again with order() and colnames() function to sort the column this time with alphabet ascendingly. The fact that sorting rows and column is not executed together is because Final_Result is not mutate completely yet, so order() function cannot find the column name. In the end of data manipulation and data exploration, View(), summary() and str() function also can be used to display a more simple details of the data as shown below.

```
> str(Question2Data)
'data.frame':   922 obs. of  8 variables:
 $ Age                  : int  16 16 17 19 17 19 16 18 16 18 ...
 $ Final_Result         : num  6.67 6.67 8.33 8.33 8.33 8.33 10 10 10 10 ...
 $ G1                   : int  4 4 5 5 5 5 6 6 6 6 ...
 $ G2                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ G3                   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Past_Failures        : int  2 2 3 3 3 3 0 0 0 0 ...
 $ Reason_Choosing_School: chr  "Interest in Course" "Interest in Course" "Close to Home" "Close
to Home" ...
 $ Sex                  : chr  "Female" "Female" "Male" "Male" ...
```

Figure 38: Question 2 Data datatype

```
> summary(Question2Data)
      Age         Final_Result         G1              G2              G3
 Min.   :15.00   Min.   : 6.67   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
 1st Qu.:16.00   1st Qu.:41.67   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
 Median :17.00   Median :53.33   Median :11.00   Median :11.00   Median :11.00
 Mean   :16.74   Mean   :53.62   Mean   :10.94   Mean   :10.77   Mean   :10.46
 3rd Qu.:18.00   3rd Qu.:66.67   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
 Max.   :22.00   Max.   :96.67   Max.   :19.00   Max.   :19.00   Max.   :20.00
 Past_Failures    Reason_Choosing_School     Sex
 Min.   :0.0000   Length:922             Length:922
 1st Qu.:0.0000   Class :character       Class :character
 Median :0.0000   Mode  :character       Mode  :character
 Mean   :0.3319
 3rd Qu.:0.0000
 Max.   :3.0000
```

Figure 39: Question 2 Data summary

```
# Focus on outliers data sets
students_outliers = Question2Data %>%
  filter(G2 == 0 & G3 == 0) %>%
  select(Reason_Choosing_School, Sex)

View(students_outliers)
summary(students_outliers)
```

Figure 40: students_outliers creation

A sub-dataset named student_outliers also being created from Question2Data which will focus on outliers for usage of graph later by using filter() function to have the condition, the outliers taken are only students who scored 0 marks in G2 and G3 academic year. It also reduce the column compared to main dataset using select() function and only choosing Reason_Choosing_School and Sex.

Analysis 1: Determine the Relationship between Student's Performance and Past Failure

Data Visualisation

```
# Correlation of age and past failures (regression line)
ggplot(Question2Data, aes(x = Age, y = Past_Failures)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between Age and Past Failures",
       x = "Age",
       y = "Past Failures") +
  theme(text=element_text(size = 16))
```

Figure 41: relation between age and past failures code

A regression line graph is created to display the relation between age and past failures. For mappings, Age will be the x-axis and Past_Failures be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). Hence the visual as below.

Figure 42: relation between age and past failures graph

The summary of the graph include:

1. The higher the age is, the more failures made in the past.
2. 17 years old have the smallest range of past failure, so the prediction is more accurate.
3. 22 years old have the largest range of final result, so the prediction may be not accurate.

Explanation:

As the age goes on, more and more experience will get from academic or works and some failures are having a high chance going to happened in the early phase of learning because everyone wanting to success in life according to a study (NICOLAS, 2016).

```
# Correlation of past failures and final result (regression line)
ggplot(Question2Data, aes(x = Past_Failures, y = Final_Result)) +
  geom_count() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between Past Failures and Final Result",
       x = "Past Failures",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 43: relation between past failures and final result code

A combination of count plot and regression line graph is created to display the relation between past failures and final result. For mappings, Past_Failures will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). Hence the visual as below.

Figure 44: relation between past failures and final result graph

The summary of the graph include:

1. Most of the students have 0 past failures.
2. The more failures made in the past, the lower the performance.

Explanation:

When a student having so many failures, there will be two reason how it affects performance. It is because of that is the style of that student study, least effort is put on learning so bad performances are done in the past and also current, else, some reason will be discovered in later analysis. In this graph, we can conclude that past failures related to final result significantly from the y-axis value range.

Analysis 2: Determine the Relationship between G1, G2 and G3

Data Visualisation

```
# Correlation of G1 and G2 (regression line)
G1G2 = Question2Data %>%
  ggplot(aes(x = G1, y = G2)) +
  geom_boxplot(aes(group = G1)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between G1 and G2",
       x = "G1",
       y = "G2") +
  theme(text=element_text(size = 16))
```

Figure 45: relation between G1 and G2 code

A regression line graph is created to display the relation between G1 and G2. For mappings, G1 will be the x-axis and G2 be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). This graph is saved into G1G2 variable for later used.

```
# Correlation of G2 and G3 (regression line)
G2G3 = Question2Data %>%
  ggplot(aes(x = G2, y = G3)) +
  geom_boxplot(aes(group = G2)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between G2 and G3",
       x = "G2",
       y = "G3") +
  theme(text=element_text(size = 16))
```

Figure 46: relation between G2 and G3 code

A regression line graph is created to display the relation between G2 and G3. For mappings, G2 will be the x-axis and G3 be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). This graph is saved into G2G3 variable for later used.

```
# Correlation of G1 and G3 (regression line)
G1G3 = Question2Data %>%
  ggplot(aes(x = G1, y = G3)) +
  geom_boxplot(aes(group = G1)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between G1 and G3",
       x = "G1",
       y = "G3") +
  theme(text=element_text(size = 16))
```

Figure 47: relation between G1 and G3 code

A regression line graph is created to display the relation between G1 and G3. For mappings, G1 will be the x-axis and G3 be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). This graph is saved into G1G3 variable for later used.

```
# Display three graph together
plot_grid(G1G2, G2G3, G1G3, ncol = 3)
```

Figure 48: plot grid code

plot_grid() is used to display all G1G2, G2G3 and G1G3 at the same time with desired dimension to relating them together. Hence the visual as below.

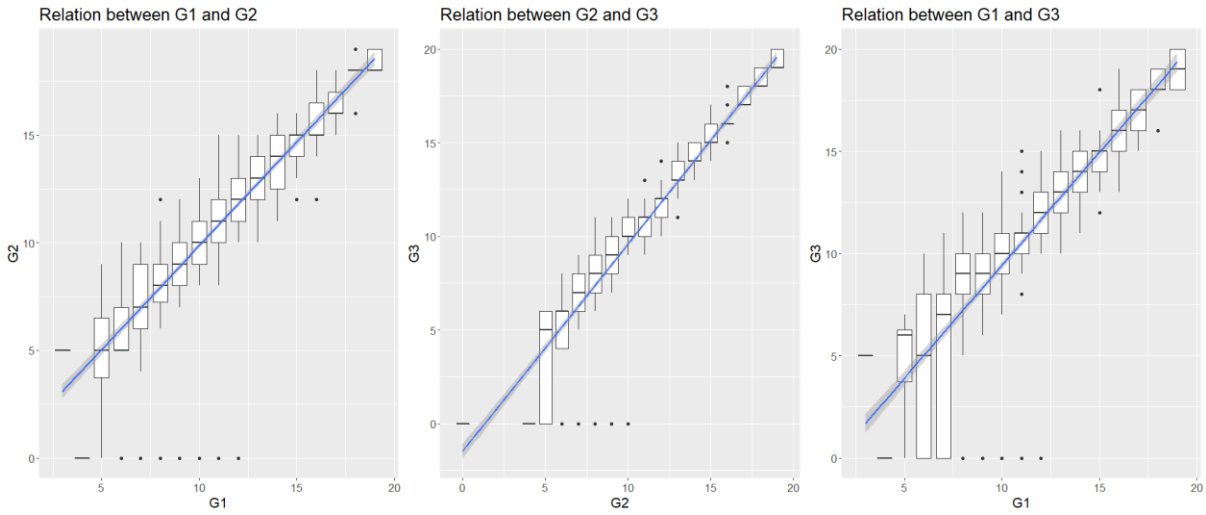

Figure 49: plot grid graph

The summary of the graph include:

1. the higher the past result is, the higher the following result.
2. Lower result in the past result tends to be more 0 marks outliers

Explanation:

This proved that a recent failure also having the same situation with past failures. Besides that, the worse the failure is, the worse the next performance is. In order to find out what is the reason the outliers exist; the next analysis is conducted.

Analysis 3: Determine the Relationship between Lower Limit Outliers and Reason Choosing School

Data Visualisation

```
# Focus on outliers
# Stacked Bar Graph
ggplot(students_outliers, aes(x = Reason_Choosing_School)) +
  geom_bar(aes(fill = Sex)) +
  labs(title = "Reason of Choosing School for those students who gave up",
       x = "Reason Choosing School",
       y = "Frequency of Students") +
  theme(text=element_text(size = 16))
```

Figure 50: reason of choosing school for those students who gave up code

A stacked bar graph is created to display the reason of choosing school for those students who gave up by using students_outliers dataset created before. For mappings, Reason_Choosing_Sex will be the x-axis. Bar filled colour is depends on the sex. Then, the graph's label is given using labs(). Hence the visual as below.

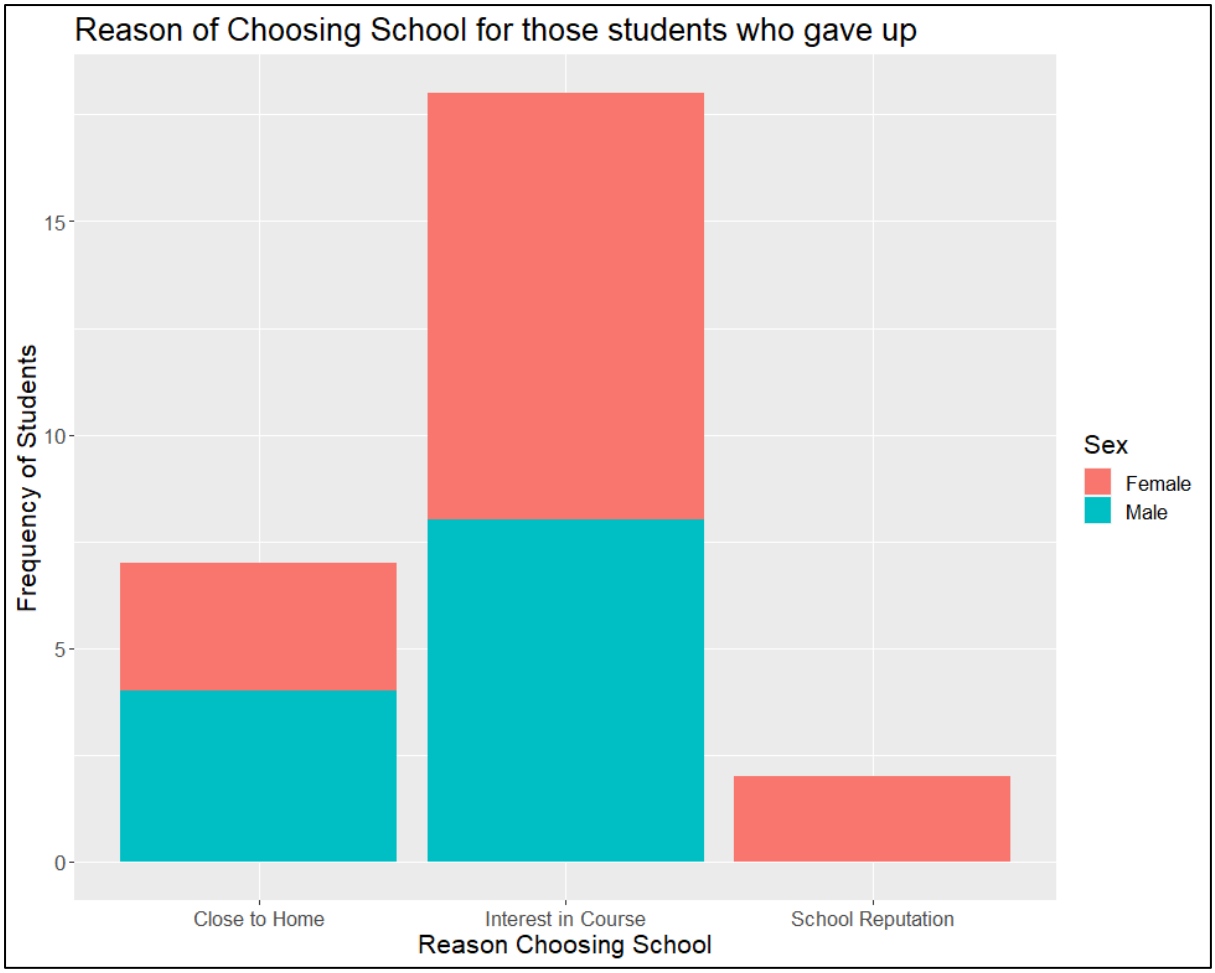

Figure 51: reason of choosing school for those students who gave up graph

The summary of the graph include:

1. Interest in course have the highest frequency no matter male or female.

Explanation:

This turned out to be weird, if a student interest in the course then the student will not give up the exam, hence we can say that the students will totally lose their confidence after the first exam does not go well especially the course is their favourite and interest one from study (betterhealth, 2014). When confidence is lost, it is very hard to concentrate anymore, so the performance remains bad significantly.

Question 2 Conclusion

The past result of a student will be affecting the following performance because of self-confidence lost, the more interest and dream of the student wanting for the course certificate, the more it will cause them to give up.

For summarise, a table is created to shows the relation between each attributes and performance produced in this whole question analysis for a simple understanding of each studied attributes as shown below.

Attributes	Affect to Performance
Past_Failures	High
G1	High
G2	High

Attributes	Affect to self-confidence
Interest in course as reason choosing school	High

In my opinion, students should try to let the past failures go and do not treat it as a bad memory instead students can get some experience and knowledge from the failures, so it will not affect to self-confidence and also improve the performance.

Question 3: What is the best living environment for students?

This question will be focusing on analysing and get a relationship between living environment and student performance so a recommended environment can be produced throughout the analysis for the students to improve performance. Therefore, some main attributes are getting targeted for this question including Living_Area, Internet_At_Home, Cohabitation_Status , Family_Relationship, Family_Size. Some of the cause and inter-related attributes may also be found in the process of analysis, hence this question is worth to take a look into.

Data Exploration

Since View() function is used beforehand, so there is no need to use any other function to determine the range or levels of the data. Rather, some very direct and simple graphs are used to get some insight and idea about the brief relationship between each other, every attribute will be put in the same graph with G1, G2 and G3 in order to not miss any interesting relationship. Different graphs will be used depends on whether the data is in continuous or discrete, therefore in this question, stacked histogram and boxplot are used shown as below. Family_Relationship shows something unexpected, so sum() function is used for conforming what is the cause.

```
# Living_Area
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Living_Area)) +
  labs(title = "Frequency Distribution of G1 Result by Living Area",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Living_Area)) +
  labs(title = "Frequency Distribution of G2 Result by Living Area",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Living_Area)) +
  labs(title = "Frequency Distribution of G3 Result by Living Area",
       x = "G3",
       y = "Number of Students")
```

Figure 52: Living_Area Exploration

```
# Internet_At_Home
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Internet_At_Home)) +
  labs(title = "Frequency Distribution of G1 Result by Internet At Home",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Internet_At_Home)) +
  labs(title = "Frequency Distribution of G2 Result by Internet At Home",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Internet_At_Home)) +
  labs(title = "Frequency Distribution of G3 Result by Internet At Home",
       x = "G3",
       y = "Number of Students")
```

Figure 53: Internet_At_Home Exploration

```
# Cohabitation_Status
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Cohabitation_Status)) +
  labs(title = "Frequency Distribution of G1 Result by Cohabitation Status",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Cohabitation_Status)) +
  labs(title = "Frequency Distribution of G2 Result by Cohabitation Status",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Cohabitation_Status)) +
  labs(title = "Frequency Distribution of G3 Result by Cohabitation Status",
       x = "G3",
       y = "Number of Students")
```

Figure 54: Cohabitation_Status Exploration

```
# Family_Relationship
ggplot(students, aes(x = G1, y = Family_Relationship)) +
  geom_boxplot(aes(group = Family_Relationship)) +
  labs(title = "Frequency Distribution of G1 Result by Family Relationship",
       x = "G1",
       y = "Family Relationship")
ggplot(students, aes(x = G2, y = Family_Relationship)) +
  geom_boxplot(aes(group = Family_Relationship)) +
  labs(title = "Frequency Distribution of G2 Result by Family Relationship",
       x = "G2",
       y = "Family Relationship")
ggplot(students, aes(x = G3, y = Family_Relationship)) +
  geom_boxplot(aes(group = Family_Relationship)) +
  labs(title = "Frequency Distribution of G3 Result by Family Relationship",
       x = "G3",
       y = "Family Relationship")
sum(students$Family_Relationship == 4)
sum(students$Family_Relationship == 5)
```

Figure 55: Family_Relationship Exploration

```
# Family_Size
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Family_Size)) +
  labs(title = "Frequency Distribution of G1 Result by Family Size",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Family_Size)) +
  labs(title = "Frequency Distribution of G2 Result by Family Size",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Family_Size)) +
  labs(title = "Frequency Distribution of G3 Result by Family Size",
       x = "G3",
       y = "Number of Students")
```

Figure 56: Family_Size Exploration

Data Manipulation and Transformation

```
# Main data sets
Question3Data = students %>%
  mutate(Final_Result = round((G1+G2+G3)/60*100, digits = 2)) %>%
  select(Sex, Living_Area, Internet_At_Home, Cohabitation_Status,
         Family_Relationship, Family_Size, Reason_Choosing_School,
         Final_Result) %>%
  arrange(Final_Result)
# Arrange columns name
Question3Data = Question3Data %>%
  select(order(colnames(Question3Data)))

View(Question3Data)
str(Question3Data)
summary(Question3Data)
```

Figure 57: Question 3 Data Manipulation and Transformation

A specified sub-dataset named Question3Data is created in this phase by using the dataset after pre-processing, piping is widely used right here to show a more readable source code. First of all, an overall performance in three years result is produced in new a column named Final_Result by using mutate() function with a simple mathematical formula, before create the column the calculation result will round up to 2 decimal place using round() function by giving desired decimal place to digits parameter. After that, every main attribute stated in data exploration will be selected so that no other column will be in the new dataset, then by using arrange() function the dataset will be sorted ascendingly following Final Result in rows. Lastly, select() function is used again with order() and colnames() function to sort the column this time with alphabet ascendingly. The fact that sorting rows and column is not executed together is because Final_Result is not mutate completely yet, so order() function cannot find the column name. In the end of data manipulation and data exploration, View(), summary() and str() function also can be used to display a more simple details of the data as shown below.

```
> str(Question3Data)
'data.frame':   922 obs. of  8 variables:
 $ Cohabitation_Status   : chr  "Living Apart" "Living Apart" "Living Together" "Living Together"
 ...
 $ Family_Relationship   : int  4 4 5 4 5 4 5 4 5 4 ...
 $ Family_Size           : chr  "Greater Than 3" "Greater Than 3" "Greater Than 3" "Greater Than
 3" ...
 $ Final_Result          : num  6.67 6.67 8.33 8.33 8.33 8.33 10 10 10 10 ...
 $ Internet_At_Home      : chr  "yes" "yes" "yes" "yes" ...
 $ Living_Area           : chr  "Urban Area" "Urban Area" "Urban Area" "Urban Area" ...
 $ Reason_Choosing_School: chr  "Interest in Course" "Interest in Course" "Close to Home" "Close
 to Home" ...
 $ Sex                   : chr  "Female" "Female" "Male" "Male" ...
```

Figure 58: Question 3 Data datatype

```
> summary(Question3Data)
 Cohabitation_Status Family_Relationship Family_Size         Final_Result
 Length:922          Min.   :1.000       Length:922         Min.   : 6.67
 Class :character    1st Qu.:4.000       Class :character   1st Qu.:41.67
 Mode  :character    Median :4.000       Mode  :character   Median :53.33
                     Mean   :3.949                          Mean   :53.62
                     3rd Qu.:5.000                          3rd Qu.:66.67
                     Max.   :5.000                          Max.   :96.67
 Internet_At_Home   Living_Area        Reason_Choosing_School     Sex
 Length:922         Length:922         Length:922             Length:922
 Class :character   Class :character   Class :character       Class :character
 Mode  :character   Mode  :character   Mode  :character       Mode  :character
```

Figure 59: Question 3 Data summary

```
# Focus on rural living area data sets
students_rural = Question3Data %>%
  filter(Living_Area == "Rural Area") %>%
  select(Family_Relationship, Living_Area, Reason_Choosing_School, Sex)

View(students_rural)
summary(students_rural)
```

Figure 60: student_rural creation

A sub-dataset named student_rural also being created from Question3Data which will focus on students who live at rural area by using filter() function to have the condition. It also reduce the column compared to main dataset using select() function and only choosing Family_Relationship, Living_Area, Reason_Choosing_School and Sex.

```
# Focus on specific family relationship data sets
students_famrel = Question3Data %>%
  filter(Family_Relationship == 4 | Family_Relationship == 5)

View(students_famrel)
summary(students_famrel)
```

Figure 61: student_famrel creation

A sub-dataset named student_famrel also being created from Question3Data which will focus on students with family relationship of 4 or 5 for usage of graph later by using filter() function to have the condition..

Analysis 1: Determine the Relationship between Final Result and Living Area

Data Visualisation

```
# Determine the relationship between Final Result and Living Area

# Boxplot
ggplot(Question3Data, aes(x = Living_Area, y = Final_Result)) +
  geom_boxplot(aes(group = Living_Area)) +
  labs(title = "Average of Final Result in Different Living Area",
       x = "Living Area",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 62: average of final result in different living area code

A boxplot is created to display the average of final result in different living area. For mappings, Living_Area will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Living_Area. Then, the graph's label is given using labs(). Hence the visual as below.

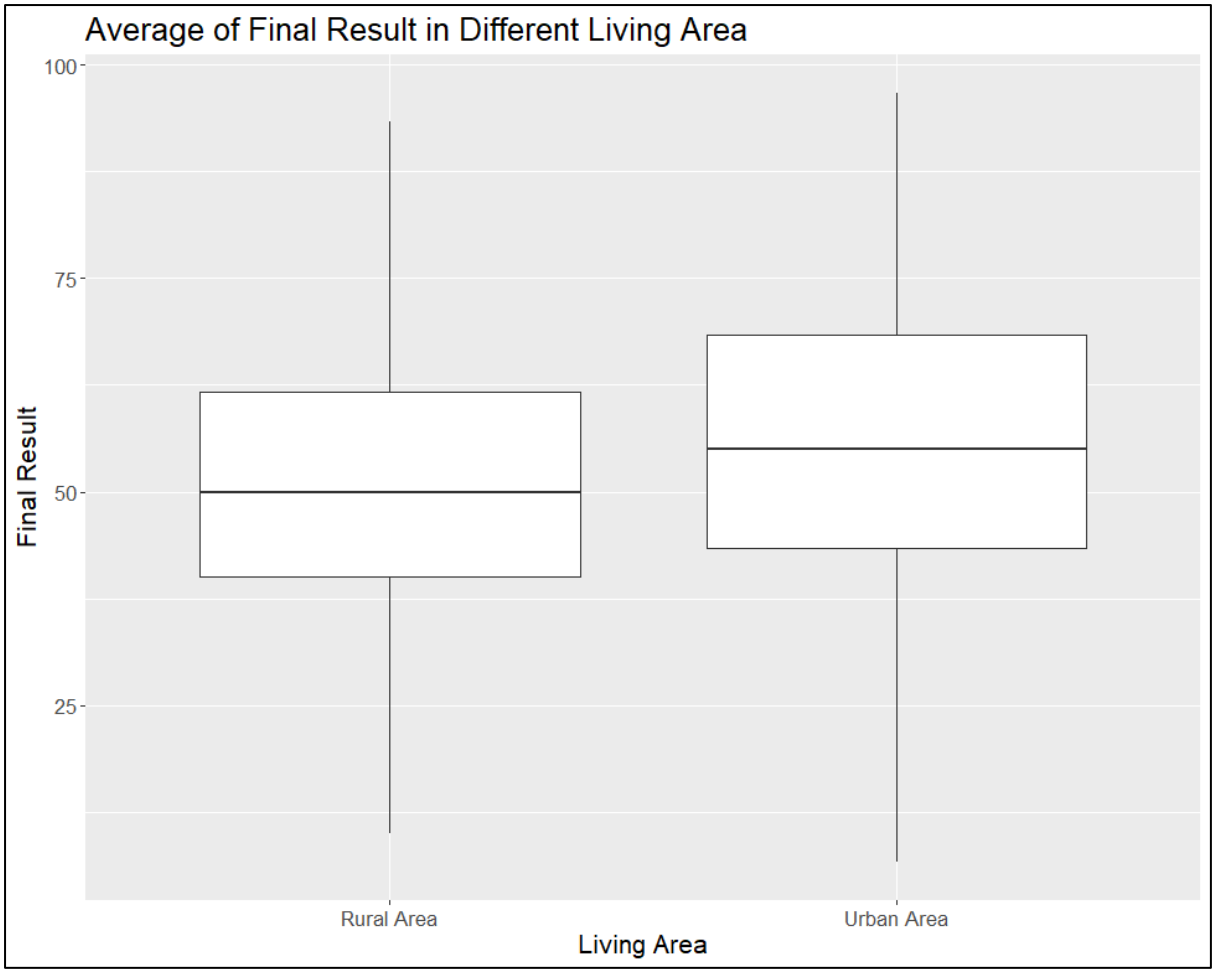

Figure 63: average of final result in different living area graph

The summary of the graph include:

1. Students living at urban area have slightly higher performance compared to rural area.

Explanation:

Urban area has more learning resources and facilities provided for students to study and increase the understanding of the course (Konuk, Turan, & Ardali, 2016). Since there is an attribute about internet availability in this dataset, it should be related to this result and will be analysed in the next analysis.

Analysis 2: Determine the Relationship between Living Area and Internet At Home

Data Visualisation

```
# Count plot
ggplot(Question3Data, aes(x = Living_Area, y = Internet_At_Home)) +
  geom_count() +
  labs(title = "Frequency of Students Have Internet at Home for Both Rural and Urban Area",
       x = "Living Area",
       y = "Internet at Home") +
  theme(text=element_text(size = 16))
```

Figure 64: frequency of students have internet at home for both rural and urban area code

A count plot is created to display the frequency of students have internet at home for both rural and urban area. For mappings, Living_Area will be the x-axis and Internet_At_Home be the y-axis. Then, the graph's label is given using labs(). Hence the visual as below.

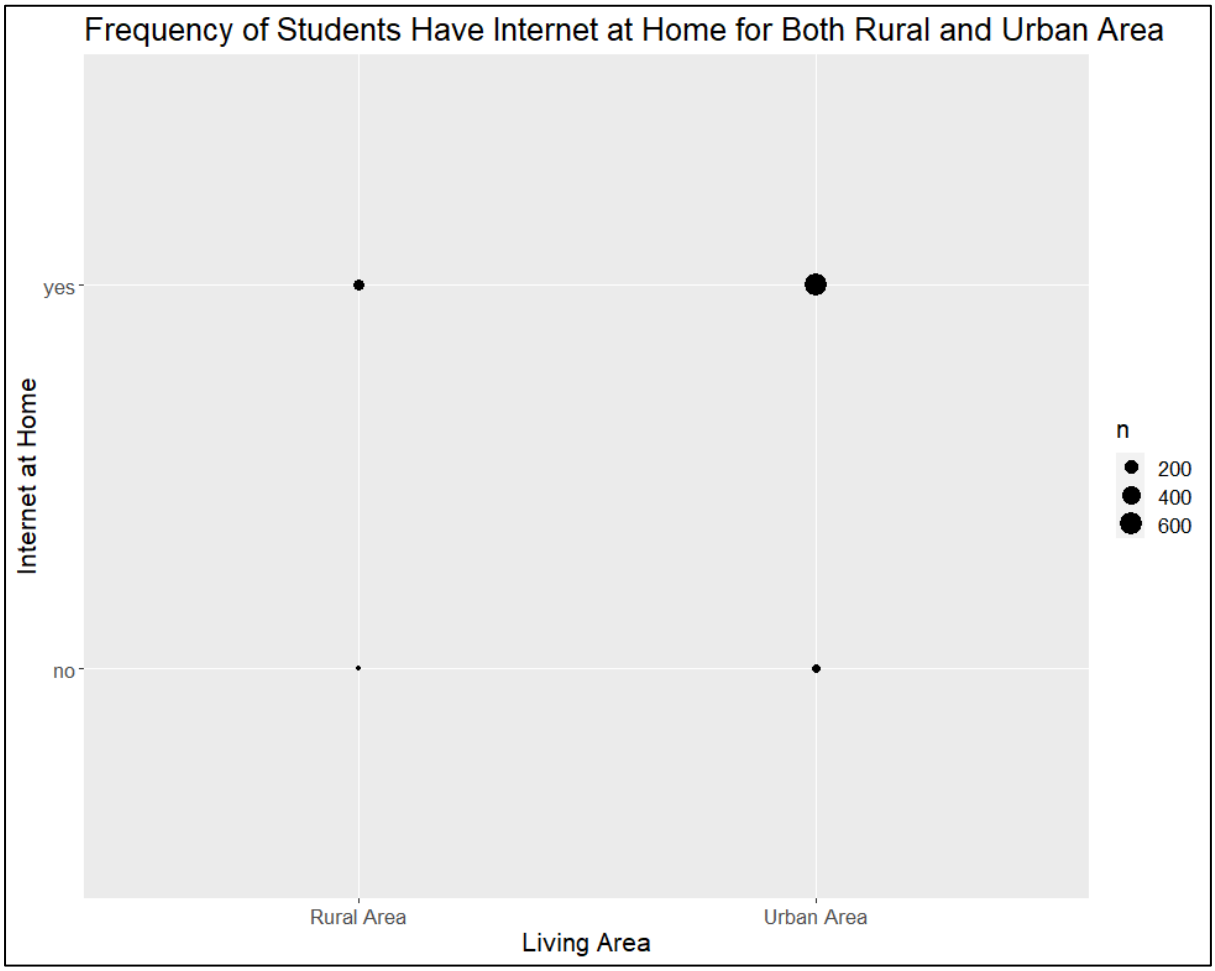

Figure 65: frequency of students have internet at home for both rural and urban area graph

The summary of the graph include:

1. Number of urban area students have internet at home is more than rural area students.
2. Distribution of students who have internet at home in rural area is better than urban area.

Explanation:

Although Number of urban area students have internet at home is more than rural area students, but number of urban area students do not have internet at home also higher than rural area, so the distribution in rural area is better. Since rural area facilities such as library are in short, so students at rural area usually rely on internet more to get educational resource. However, this does not correspond to real life situation such as this report (Steven , Edward, & Janet, 2014), hence an assumption is made where there is a difference in bandwidth speed but it is not stated in this dataset which mean rural area students will have a weak internet speed.

Analysis 3: Determine the Relationship between Reason Choosing School and Living Area

Data Visualisation

```
# Stacked Bar Graph
ggplot(students_rural, aes(x = Reason_Choosing_School)) +
  geom_bar(aes(fill = Sex)) +
  labs(title = "Number of Students Live in Rural Area with Different Reason Choosing School",
       x = "Reason Choosing School",
       y = "Frequncy of Students") +
  theme(text=element_text(size = 16))
```

Figure 66: number of students live in rural area with different reason choosing school code

A stacked bar graph is created to display the number of students live in rural area with different reason choosing school by using students_rural dataset created before. For mappings, Reason_Choosing_Sex will be the x-axis. Bar filled colour is depends on the sex. Then, the graph's label is given using labs(). Hence the visual as below.

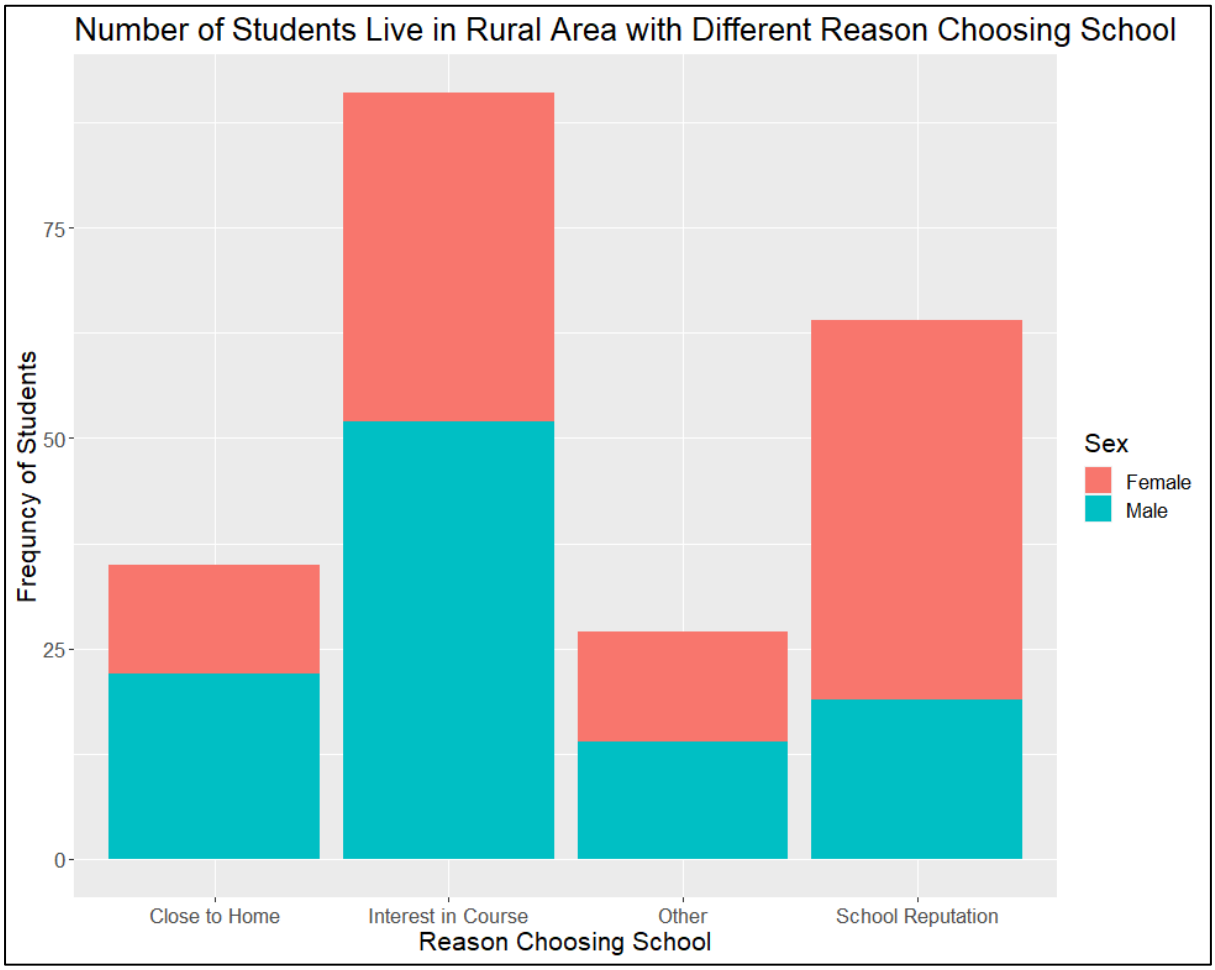

Figure 67: number of students live in rural area with different reason choosing school graph

The summary of the graph include:

1. Interest in course have highest frequency.
2. Female students will stay at rural area mostly because of the school reputation and interest in course.
3. Male students will stay at rural area mostly because of interest in course.

Explanation:

Male students will focus on interest in course when choosing a school, no matter the school is in rural area he will still stay at there. That goes the same to female students but female students will also go for rural area when the school reputation is high.

```
# Stacked Bar Graph
ggplot(students_rural, aes(x = Family_Relationship)) +
  geom_bar(aes(fill = Sex)) +
  labs(title = "Number of Students Live in Rural Area with Different Family Relationship",
       x = "Family Relationship",
       y = "Frequncy of Students") +
  scale_x_continuous(labels=c("Worst", "Bad", "Normal", "Close", "Closest"),
                     breaks=1:5) +
  theme(text=element_text(size = 16))
```

Figure 68: number of students live in rural area with different family relationship code

A stacked bar graph is created to display the number of students live in rural area with different family relationship by using students_rural dataset created before. For mappings, Family_Relationship will be the x-axis. Bar filled colour is depends on the sex. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

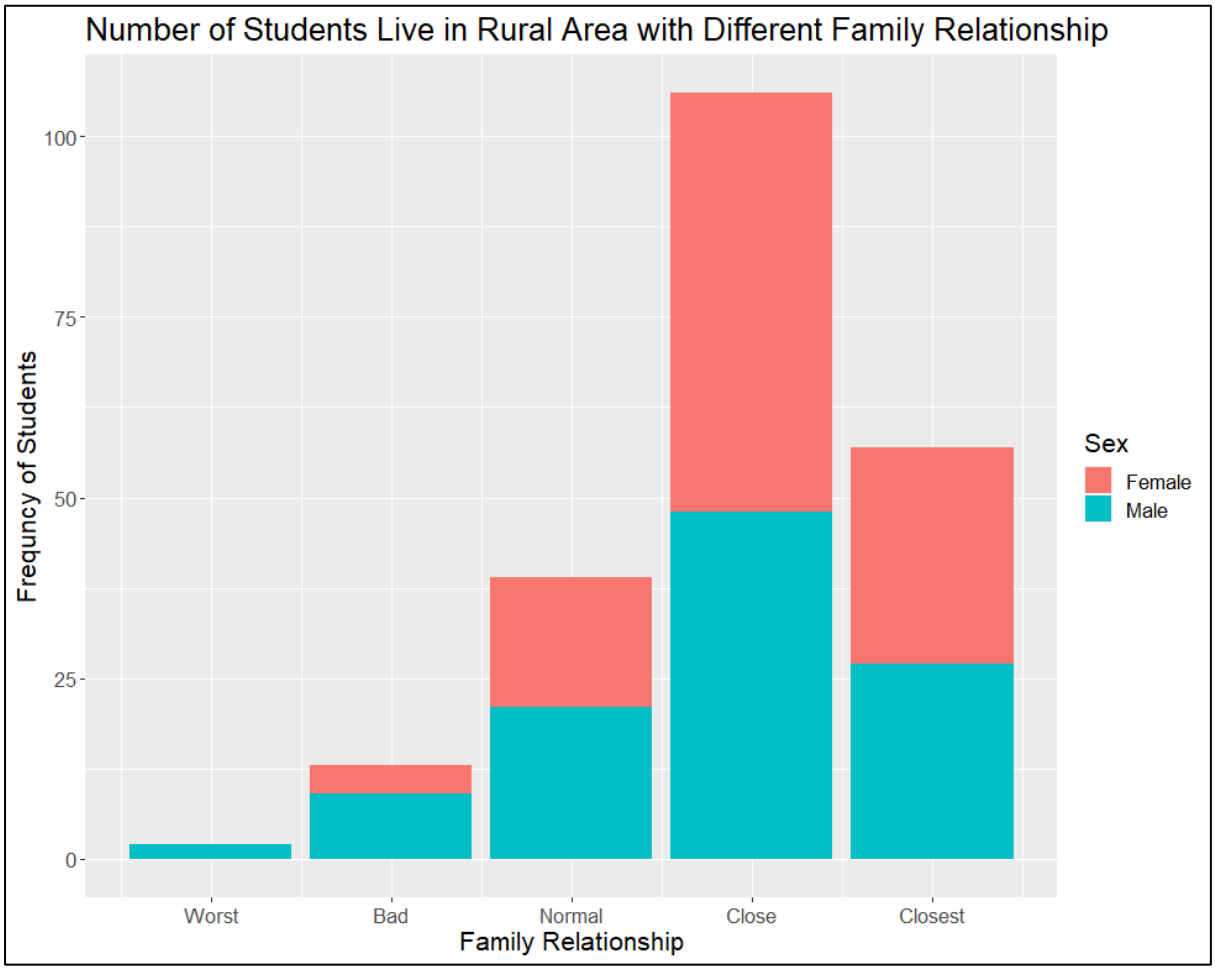

Figure 69: number of students live in rural area with different family relationship graph

The summary of the graph include:

1. Close family relationship has the highest frequency.
2. The closer the family relationship, the higher the chance the students will stay in rural area.

Explanation:

With close family relationship, the students will most likely do not want to stay apart from family, so the student will still stay in rural area for study. However, closest frequency shows lower than expectation, so a further graph is produced below to identify the cause.

```
# Reason why Closest relation is lower in graph compared to Close
# Bar graph
ggplot(students_famrel, aes(x = Family_Relationship)) +
  geom_bar(width = 0.5) +
  labs(title = "Ratio of Students in Close and Closest Family Relationship",
       x = "Family Relationship",
       y = "Frequncy of Students") +
  scale_x_continuous(labels=c("Close", "Closest"), breaks=4:5)+
  theme(text=element_text(size = 16))
```

Figure 70: ratio of students in close and closest family relationship code

A bar graph is created to display the ratio of students in close and closest family relationship by using students_famrel dataset created before. For mappings, Family_Relationship will be the x-axis. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

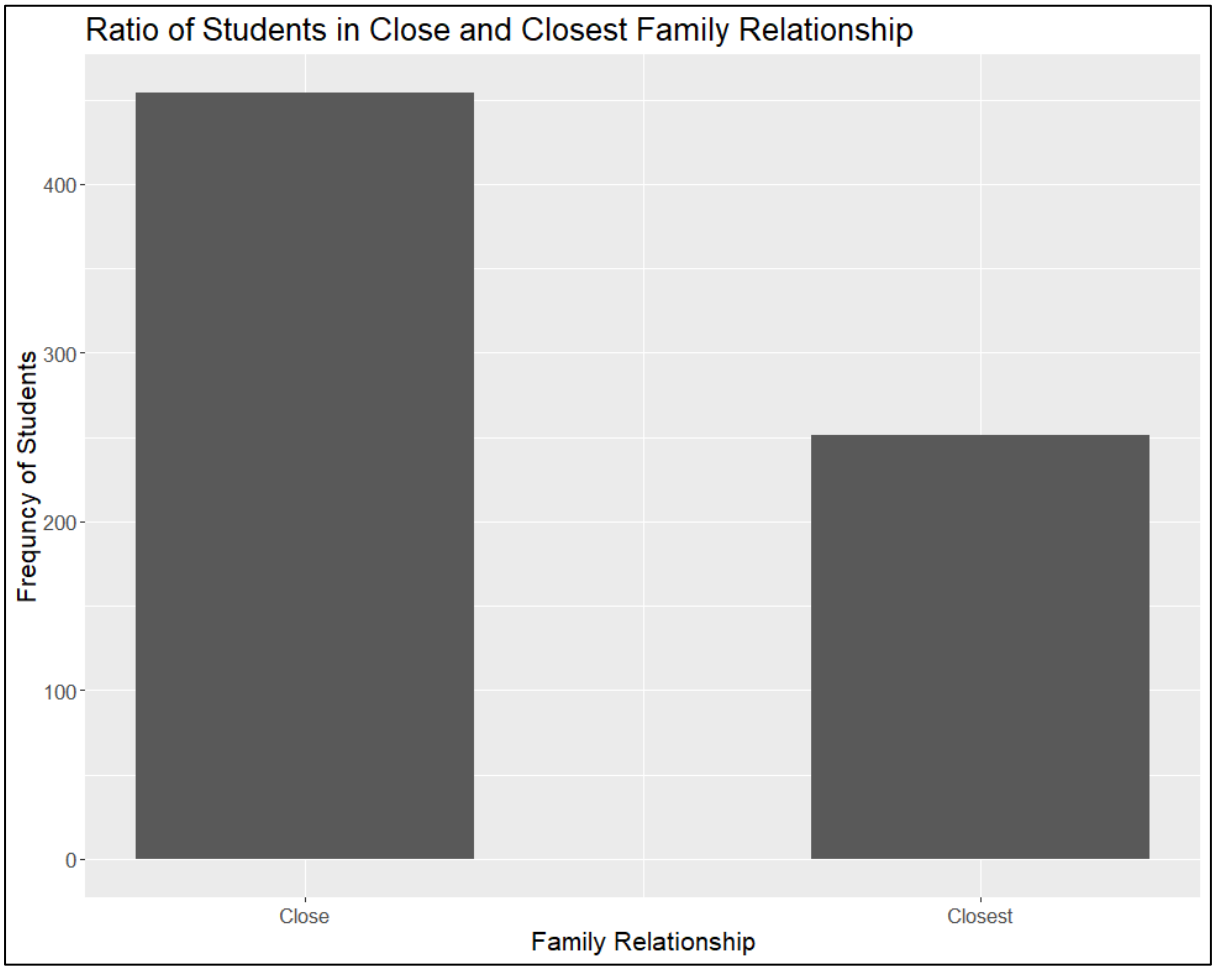

Figure 71: ratio of students in close and closest family relationship graph

The summary of the graph include:

1. Close family relationship has higher frequency than closest family relationship
2. The ratio of them is 2:1

Explanation:

This graph explains the cause in previous graph since the ratio is 2:1, because the previous graph also has about half frequency in closest relationship compared to close relationship. Since it is in bar graph, that is the limitation of it but it is still proved the information is correct.

Analysis 4: Determine the Relationship between Final Result and Family Relationship in Different Family Size

Data Visualisation

```
# Correlation between Family Relationship and Final Result(regression)
ggplot(Question3Data, aes(x = Family_Relationship, y = Final_Result)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between Final Result and Family Relationship in Different Family Size",
       x = "Family Relationship",
       y = "Final Result") +
  scale_x_continuous(labels=c("Worst", "Bad", "Normal", "Close", "Closest"),
                     breaks=1:5) +
  facet_grid(rows =  vars(Family_Size), cols = vars(Cohabitation_Status)) +
  theme(text=element_text(size = 16))
```

Figure 72: relation between final result and family relationship in different family size code

A regression line graph is created to display the relation between final result and family relationship in different family size. For mappings, Family_Relationship will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Lastly, the graph is separated to 2by2 using facet_grid() for Family_Size and Cohabitation_Status. Hence the visual as below.

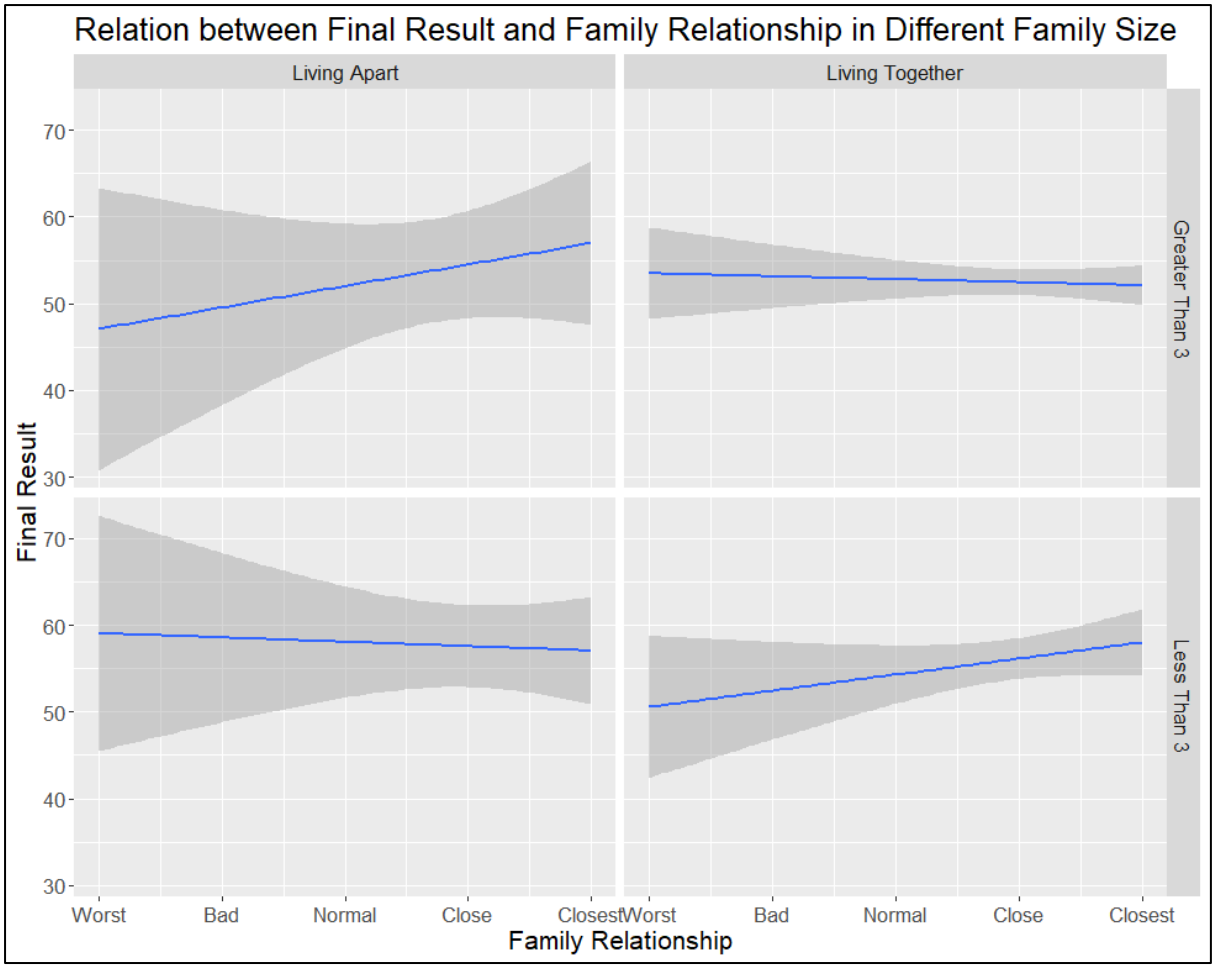

Figure 73: relation between final result and family relationship in different family size graph

The summary of the graph include:

1. Positive correlation:
    - Family Size Greater than 3 and Living Apart (strongest)
    - Family Size Less than 3 and Living Together
2. Negative correlation:
    - Family Size Greater than 3 and Living Together
    - Family Size Less than 3 and Living Apart (strongest)

Explanation:

Normally The closer the family relationship is, the higher the student performance but there is another attribute influences this result, if family size is greater than 3 it is better to living apart when the relationship is close because the environment for students is noisier and more unable to concentrate to study. In opposite, family size is less than 3 is better to living together so family member can talk with the student for any problem faced in academic.

Question 3 Conclusion

Living environment will affect student performance in many aspects. Student lives in urban area will be more facilities that can be utilizes for more understanding to the course for example availability of internet at home. Shelter environment also can affect student environment which involve of family aspect, if family member can give some help in appropriate time and not to make home environment become noisy then it can help the students on studying the course.

For summarise, a table is created to shows the relation between each attributes and performance produced in this whole question analysis for a simple understanding of each studied attributes as shown below.

Attributes	Affect to Performance
Living_Area	Slightly
Internet_At_Home	Slightly
Cohabitation_Status	High
Family_Relationship	High
Family_Size	High

Attributes	Improve performance
Live at urban area	Slightly
Have internet at home	Slightly
Family size greater than 3, living apart and closest family relationship	High

In my opinion, students should move to urban area for better performance no matter what because urban area is definitely more suitable environment for degree students as there are many resources provided.

Question 4: Health/Stress status or focus on study is more important?

This question will be focusing on heal/stress or study is more important to a student in term of performance and other aspect, so a good distribution of time can be produced and maximise the productivity of students. Therefore, some main attributes are getting targeted for this question including WeekDay_Alcohol_Consumed, Weekend_Alcohol_Consumed, Health_Status, Free_Time, Hang_Out, Cocurricular_Activity and Travel_Time. Some of the cause and inter-related attributes may also be found in the process of analysis, hence this question is worth to take a look into.

Data Exploration

Since View() function is used beforehand, so there is no need to use any other function to determine the range or levels of the data. Rather, some very direct and simple graphs are used to get some insight and idea about the brief relationship between each other, every attribute will be put in the same graph with G1, G2 and G3 in order to not miss any interesting relationship. Different graphs will be used depends on whether the data is in continuous or discrete, therefore in this question, stacked histogram and boxplot are used shown as below.

```
# WeekDay_Alcohol_Consumed
ggplot(students, aes(x = G1, y = WeekDay_Alcohol_Consumed)) +
  geom_boxplot(aes(group = WeekDay_Alcohol_Consumed)) +
  labs(title = "Frequency Distribution of G1 Result by WeekDay Alcohol Consumed",
       x = "G1",
       y = "WeekDay Alcohol Consumed")
ggplot(students, aes(x = G2, y = WeekDay_Alcohol_Consumed)) +
  geom_boxplot(aes(group = WeekDay_Alcohol_Consumed)) +
  labs(title = "Frequency Distribution of G2 Result by WeekDay Alcohol Consumed",
       x = "G2",
       y = "WeekDay Alcohol Consumed")
ggplot(students, aes(x = G3, y = WeekDay_Alcohol_Consumed)) +
  geom_boxplot(aes(group = WeekDay_Alcohol_Consumed)) +
  labs(title = "Frequency Distribution of G3 Result by WeekDay Alcohol Consumed",
       x = "G3",
       y = "WeekDay Alcohol Consumed")
```

Figure 74: WeekDay_Alcohol_Consumed Exploration

```
# Weekend_Alcohol_Consumed
ggplot(students, aes(x = G1, y = Weekend_Alcohol_Consumed)) +
  geom_boxplot(aes(group = Weekend_Alcohol_Consumed)) +
  labs(title = "Frequency Distribution of G1 Result by Weekend Alcohol Consumed",
       x = "G1",
       y = "Weekend Alcohol Consumed")
ggplot(students, aes(x = G2, y = Weekend_Alcohol_Consumed)) +
  geom_boxplot(aes(group = Weekend_Alcohol_Consumed)) +
  labs(title = "Frequency Distribution of G2 Result by Weekend Alcohol Consumed",
       x = "G2",
       y = "Weekend Alcohol Consumed")
ggplot(students, aes(x = G3, y = Weekend_Alcohol_Consumed)) +
  geom_boxplot(aes(group = Weekend_Alcohol_Consumed)) +
  labs(title = "Frequency Distribution of G3 Result by Weekend Alcohol Consumed",
       x = "G3",
       y = "Weekend Alcohol Consumed")
```

Figure 75: Weekend_Alcohol_Consumed Exploration

```
# Health_Status
ggplot(students, aes(x = G1, y = Health_Status)) +
  geom_boxplot(aes(group = Health_Status)) +
  labs(title = "Frequency Distribution of G1 Result by Health Status",
       x = "G1",
       y = "Health Status")
ggplot(students, aes(x = G2, y = Health_Status)) +
  geom_boxplot(aes(group = Health_Status)) +
  labs(title = "Frequency Distribution of G2 Result by Health Status",
       x = "G2",
       y = "Health Status")
ggplot(students, aes(x = G3, y = Health_Status)) +
  geom_boxplot(aes(group = Health_Status)) +
  labs(title = "Frequency Distribution of G3 Result by Health Status",
       x = "G3",
       y = "Health Status")
```

Figure 76: Health_Status Exploration

```
# Free_Time
ggplot(students, aes(x = G1, y = Free_Time)) +
  geom_boxplot(aes(group = Free_Time)) +
  labs(title = "Frequency Distribution of G1 Result by Free Time",
       x = "G1",
       y = "Free Time")
ggplot(students, aes(x = G2, y = Free_Time)) +
  geom_boxplot(aes(group = Free_Time)) +
  labs(title = "Frequency Distribution of G2 Result by Free Time",
       x = "G2",
       y = "Free Time")
ggplot(students, aes(x = G3, y = Free_Time)) +
  geom_boxplot(aes(group = Free_Time)) +
  labs(title = "Frequency Distribution of G3 Result by Free Time",
       x = "G3",
       y = "Free Time")
```

Figure 77: Free_Time Exploration

```
# Hang_Out
ggplot(students, aes(x = G1, y = Hang_Out)) +
  geom_boxplot(aes(group = Hang_Out)) +
  labs(title = "Frequency Distribution of G1 Result by Hang Out",
       x = "G1",
       y = "Hang Out")
ggplot(students, aes(x = G2, y = Hang_Out)) +
  geom_boxplot(aes(group = Hang_Out)) +
  labs(title = "Frequency Distribution of G2 Result by Hang Out",
       x = "G2",
       y = "Hang Out")
ggplot(students, aes(x = G3, y = Hang_Out)) +
  geom_boxplot(aes(group = Hang_Out)) +
  labs(title = "Frequency Distribution of G3 Result by Hang Out",
       x = "G3",
       y = "Hang Out")
```

Figure 78: Hang_Out Exploration

```
# Cocurricular_Activity
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Cocurricular_Activity)) +
  labs(title = "Frequency Distribution of G1 Result by Cocurricular Activity",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Cocurricular_Activity)) +
  labs(title = "Frequency Distribution of G2 Result by Cocurricular Activity",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Cocurricular_Activity)) +
  labs(title = "Frequency Distribution of G3 Result by Cocurricular Activity",
       x = "G3",
       y = "Number of Students")
```

Figure 79: Cocurricular_Activity Exploration

```
# Travel_Time
ggplot(students, aes(x = G1, y = Travel_Time)) +
  geom_boxplot(aes(group = Travel_Time)) +
  labs(title = "Frequency Distribution of G1 Result by Travel Time",
       x = "G1",
       y = "Travel Time")
ggplot(students, aes(x = G2, y = Travel_Time)) +
  geom_boxplot(aes(group = Travel_Time)) +
  labs(title = "Frequency Distribution of G2 Result by Travel Time",
       x = "G2",
       y = "Travel Time")
ggplot(students, aes(x = G3, y = Travel_Time)) +
  geom_boxplot(aes(group = Travel_Time)) +
  labs(title = "Frequency Distribution of G3 Result by Travel Time",
       x = "G3",
       y = "Travel Time")
```

Figure 80: Travel_Time Exploration

Data Manipulation and Transformation

```
# Main data sets
Question4Data = students %>%
  mutate(Final_Result = round((G1+G2+G3)/60*100, digits = 2),
         Daily_Alcohol = (WeekDay_Alcohol_Consumed + Weekend_Alcohol_Consumed)/2) %>%
  select(Sex, Study_Time, Daily_Alcohol, Health_Status, Free_Time,
         Hang_Out, Cocurricular_Activity, Travel_Time, Final_Result) %>%
  arrange(Final_Result)
# Arrange columns name
Question4Data = Question4Data %>%
  select(order(colnames(Question4Data)))

View(Question4Data)
str(Question4Data)
summary(Question4Data)
```

Figure 81: Question 4 Data Manipulation and Transformation

A specified sub-dataset named Question4Data is created in this phase by using the dataset after pre-processing, piping is widely used right here to show a more readable source code. First of all, an overall performance in three years result is produced in new a column named Final_Result by using mutate() function with a simple mathematical formula, before create the column the calculation result will round up to 2 decimal place using round() function by giving desired decimal place to digits parameter, a daily alcohol consumed also be created because the two attribute is bringing similar result. After that, every main attribute stated in data exploration will be selected so that no other column will be in the new dataset, then by using arrange() function the dataset will be sorted ascendingly following Final Result in rows. Lastly, select() function is used again with order() and colnames() function to sort the column this time with alphabet ascendingly. The fact that sorting rows and column is not executed together is because Final_Result is not mutate completely yet, so order() function cannot find the column name. In the end of data manipulation and data exploration, View(), summary() and str() function also can be used to display a more simple details of the data as shown below.

```
> str(Question4Data)
'data.frame':   922 obs. of  9 variables:
 $ Cocurricular_Activity: chr  "yes" "yes" "no" "no" ...
 $ Daily_Alcohol        : num  1 1 1.5 1 1.5 1 1 1.5 1 1.5 ...
 $ Final_Result         : num  6.67 6.67 8.33 8.33 8.33 8.33 10 10 10 10 ...
 $ Free_Time            : int  3 3 4 5 4 5 4 3 4 3 ...
 $ Hang_Out             : int  2 2 5 4 5 4 5 5 5 5 ...
 $ Health_Status        : int  5 5 5 4 5 4 3 3 3 3 ...
 $ Sex                  : chr  "Female" "Female" "Male" "Male" ...
 $ Study_Time           : int  1 1 1 1 1 1 1 2 1 2 ...
 $ Travel_Time          : int  2 2 1 1 1 1 1 2 1 2 ...
```

Figure 82: Question 4 Data datatype

```
> summary(Question4Data)
 Cocurricular_Activity Daily_Alcohol    Final_Result     Free_Time        Hang_Out
 Length:922            Min.   :1.000   Min.   : 6.67   Min.   :1.000   Min.   :1.000
 Class :character      1st Qu.:1.000   1st Qu.:41.67   1st Qu.:3.000   1st Qu.:2.000
 Mode  :character      Median :1.500   Median :53.33   Median :3.000   Median :3.000
                       Mean   :1.894   Mean   :53.62   Mean   :3.252   Mean   :3.092
                       3rd Qu.:2.500   3rd Qu.:66.67   3rd Qu.:4.000   3rd Qu.:4.000
                       Max.   :5.000   Max.   :96.67   Max.   :5.000   Max.   :5.000
 Health_Status       Sex              Study_Time     Travel_Time
 Min.   :1.000   Length:922         Min.   :1.000   Min.   :1.000
 1st Qu.:3.000   Class :character   1st Qu.:1.000   1st Qu.:1.000
 Median :4.000   Mode  :character   Median :2.000   Median :1.000
 Mean   :3.565                      Mean   :2.037   Mean   :1.457
 3rd Qu.:5.000                      3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :5.000                      Max.   :4.000   Max.   :4.000
```

Figure 83: Question 4 Data summary

Analysis 1: Determine the Relationship between Final Result and Health Status

Data Visualisation

```
# -Assume Health Status include physically and mentally(stress).
# Correlation between final result and health status (regression)
ggplot(Question4Data, aes(x = Health_Status, y = Final_Result)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between Final Result and Health Status",
       x = "Health Status",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 84: relation between final result and health status code

A regression line graph is created to display the relation between final result and health status. For mappings, Health_Status will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). Hence the visual as below.

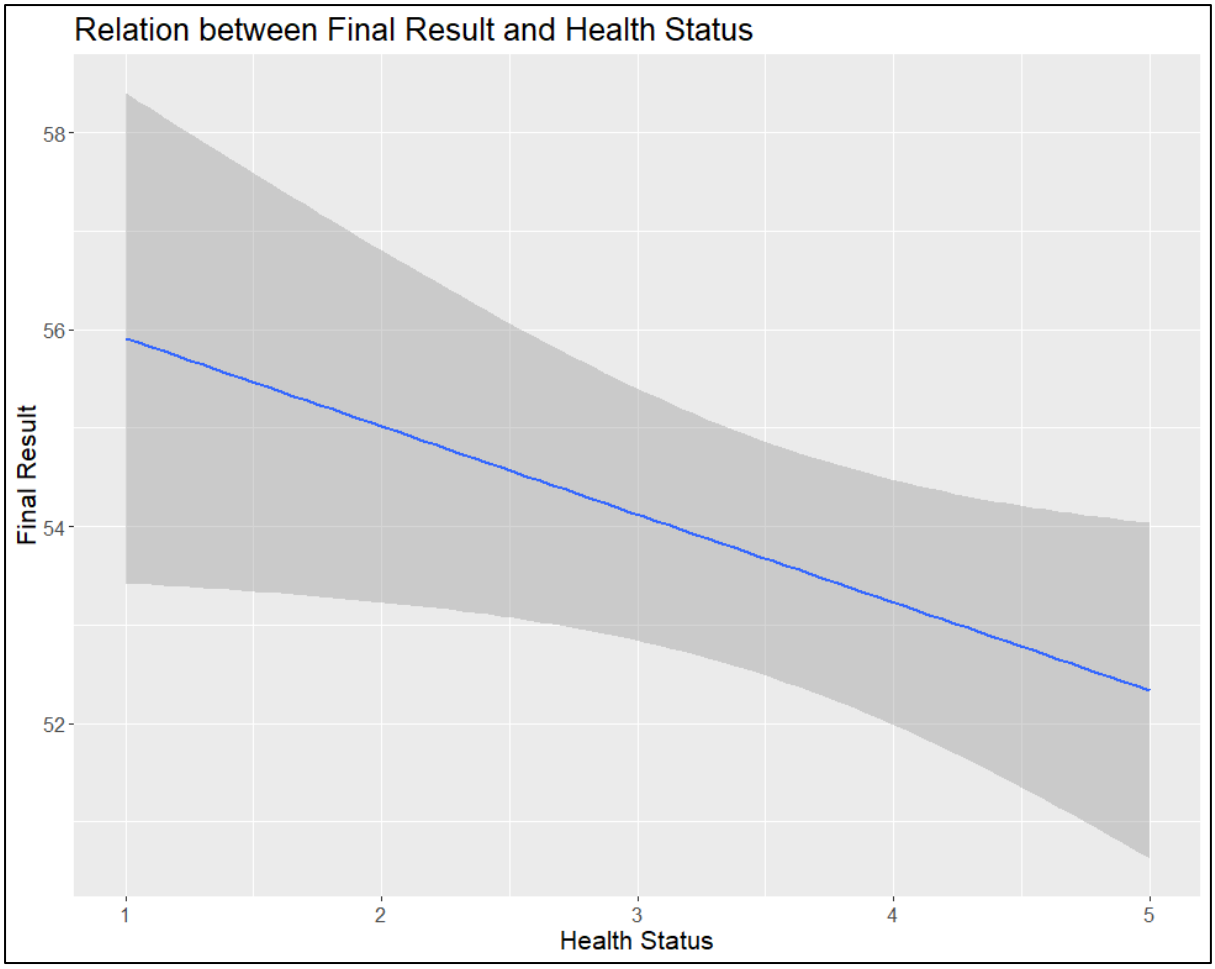

Figure 85: relation between final result and health status graph

The summary of the graph include:

1. The higher the health status, the lower the students' performance.
2. Health status is high means physical health is good but mental health is bad.
3. Health status level 3 have the smallest range of final result, so the prediction is more accurate.
4. Health status level 1 have the largest range of final result, so the prediction may be not accurate.

Explanation:

Normally, we expected a healthy student to have better performance but this graph shows in a completely opposite way. Therefore, an assumption is made where health status includes physically and mentally such as stress, then everything is making sense again. Obviously, if the students stress is very high and mentally damaged then it will definitely affect the student performance according to the study to mental health affect performance in work (CDC, 2019).

```
# Correlation between Study Time and health status (regression)
ggplot(Question4Data, aes(x = Study_Time, y = Health_Status)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between Study Time and Health Status",
       x = "Study Time",
       y = "Health Status") +
  theme(text=element_text(size = 16))
```

Figure 86: relation between study time and health status code

A regression line graph is created to display the relation between study time and health status. For mappings, Study_Time will be the x-axis and Health_Status be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). Hence the visual as below.

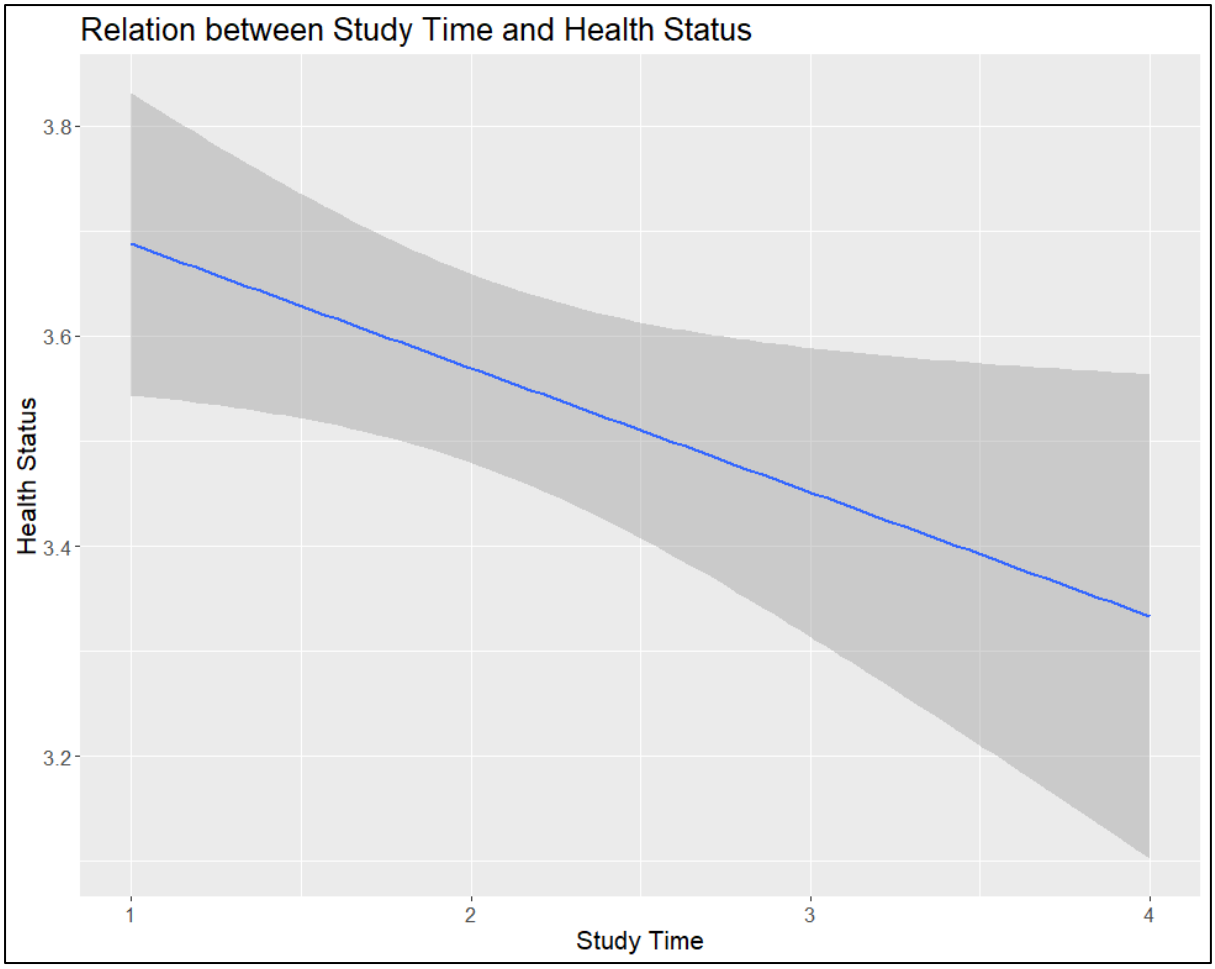

Figure 87: relation between study time and health status graph

The summary of the graph include:

1. The more time spent on study, the mental health of the student will be lower.
2. Study Time 2 have the smallest range of health status, so the prediction is more accurate.
3. Study Time 4 have the largest range of health status, so the prediction may be not accurate.

Explanation:

Some students may not be suitable to study in a long period of time, it will increase the stress of them or harm their mental health. Therefore, a suitable study time should be considered depends on student individual.

```
# Correlation between Daily Alcohol and health status (regression)
ggplot(Question4Data, aes(x = Daily_Alcohol, y = Health_Status)) +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relation between Daily Alcohol and Health Status",
       x = "Daily Alcohol",
       y = "Health Status") +
  theme(text=element_text(size = 16))
```

Figure 88: relation between daily alcohol and health status code

A regression line graph is created to display the relation between daily alcohol and health status. For mappings, Daily_Aclcohol will be the x-axis and Health_Status be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs(). Hence the visual as below.

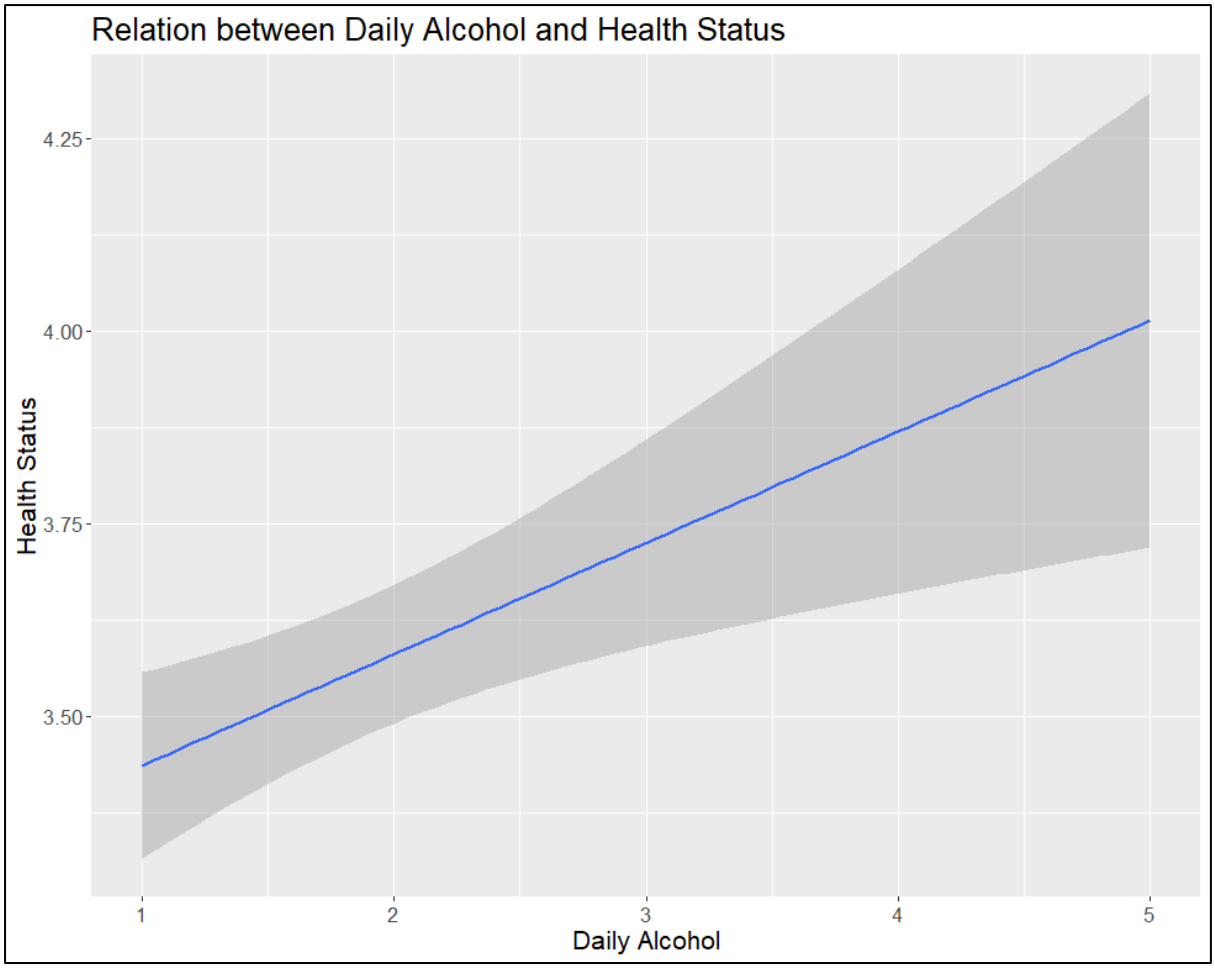

Figure 89: relation between daily alcohol and health status graph

The summary of the graph include:

1. The more daily alcohol consumed, the health status of the student improves.
2. Study Time 2 have the smallest range of health status, so the prediction is more accurate.
3. Study Time 4 have the largest range of health status, so the prediction may be not accurate.

Explanation:

We can say that student is too high pressure nowadays, hence drinking give them more relief on stress according to study (Michael, 1999) compared to harm to the body which mean the improvement in mental health outrun the harm cause by alcohol to the physical health.

Analysis 2: Determine the Relationship between Cocurricular Activity and Final Result

Data Visualisation

```
# -Assume co-curricular activity is referring to sport clubs or liberal arts
#  related clubs instead of society service or school event.
# Boxplot
ggplot(Question4Data, aes(x = Cocurricular_Activity, y = Final_Result)) +
  geom_boxplot() +
  facet_grid(~Sex) +
  labs(title = "Average of Final Result for Participation of Co-Curicullar",
       x = "Participation of Co-Curicullar",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 90: average of final result for participation of co-curricular code

A boxplot is created to display the average of final result for participation of co-curricular. For mappings, Cocurricular_Activity will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Cocurricular_Activity. Then, the graph will be separate to two by sex using facet_wrap() and some labels is given using labs(). Hence the visual as below.

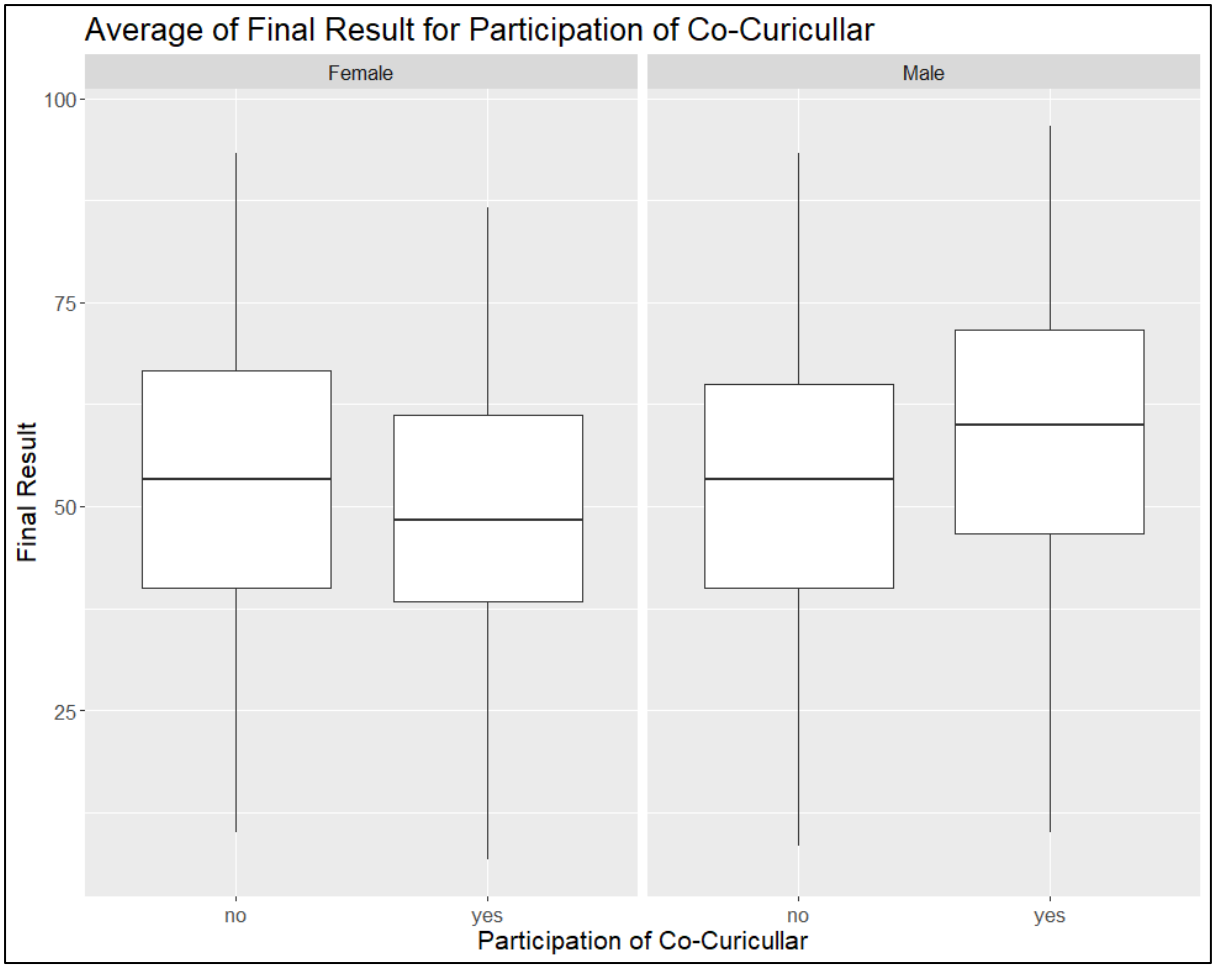

Figure 91: average of final result for participation of co-curricular graph

The summary of the graph include:

1. Male Students can get better performance if participate in co-curricular activities.
2. Female Students can get better performance if do not participate in co-curricular activities.

Explanation:

Majority of female student will choose for liberal arts related clubs. So, they also might need to consume some focus and memories knowledge from the clubs. Male student will have higher performance when participating in co-curricular because majority of female student will choose for sport clubs. So, it will be really useful for them to destress and have a fresh mind after some sport activities proved by a study (Kashif, Abdul Qayyum, & Muhammad, 2018).

Analysis 3: Determine the Relationship between Free Time, Hang Out, Travel Time and Study Time

Data Visualisation

```
# Heat Map
ggplot(Question4Data, aes(x = Free_Time, y = Hang_Out)) +
  geom_tile(aes(fill = Final_Result), color = "white") +
  scale_fill_gradient(low = "white", high = "steel blue") +
  labs(title = "Average Final Result with a Combination of Free Time and Hang Out",
       x = "Free Time",
       y = "Hang Out") +
  scale_x_continuous(labels = c("least", "less", "normal", "high", "highest"),
                     breaks = 1:5) +
  scale_y_continuous(labels = c("least", "less", "normal", "high", "highest"),
                     breaks = 1:5) +
  theme(text=element_text(size = 16))
```

Figure 92: average final result with a combination of free time and hang out code

A heat map is created to display the average final result with a combination of free time and hang out. For mappings, Free Time will be the x-axis and Hang_Out be the y-axis. The tile fill colour will be depending on Final_Result and using white colour line to separate the tiles, the fill colour is changed to from white to steel blue. Then, the graph's label is given using labs() while scale_x_continuous() and scale_y_continuous() used to change the label display to word, scale. Hence the visual as below.

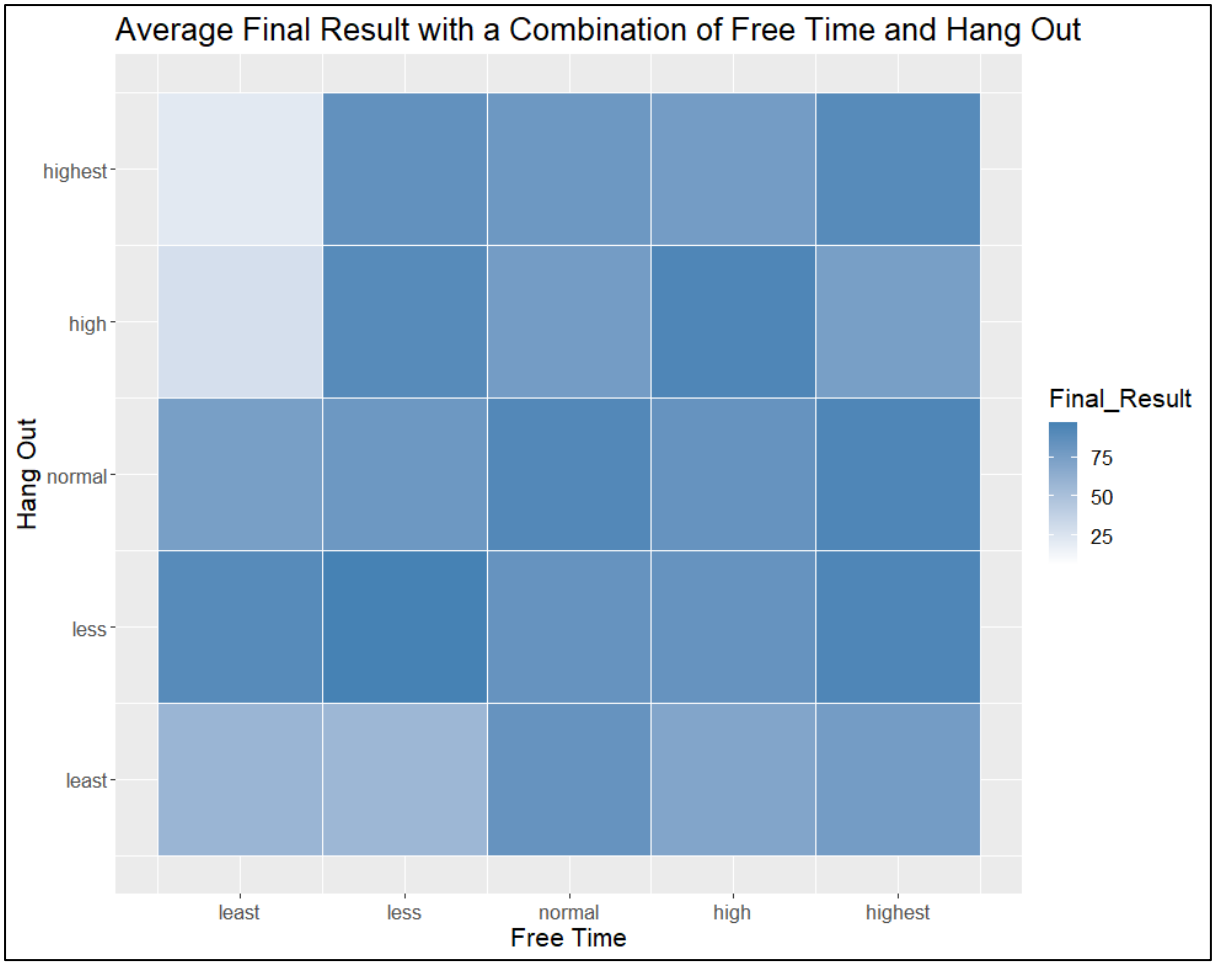

Figure 93: average final result with a combination of free time and hang out graph

Explanation:

if the free time after school is least then the student will have lower performance when they still spend much time on hang out with friends. Hence, the student should not hang out too much with friends when the free time is less.

```
# Heat Map
ggplot(Question4Data, aes(x = Free_Time, y = Travel_Time)) +
  geom_tile(aes(fill = Final_Result), color = "white") +
  scale_fill_gradient(low = "white", high = "steel blue") +
  labs(title = "Average Final Result with a Combination of Free Time and Travel Time",
       x = "Free Time",
       y = "Travel Time") +
  scale_x_continuous(labels = c("least", "less", "normal", "high", "highest"),
                     breaks = 1:5) +
  scale_y_continuous(labels = c("least", "less", "high", "highest"),
                     breaks = 1:4) +
  theme(text=element_text(size = 16))
```

Figure 94: average final result with a combination of free time and travel time code

A heat map is created to display the average final result with a combination of free time and travel time. For mappings, Free Time will be the x-axis and Travel_Time be the y-axis. The tile fill colour will be depending on Final_Result and using white colour line to separate the tiles, the fill colour is changed to from white to steel blue. Then, the graph's label is given using labs() while scale_x_continuous() and scale_y_continuous() used to change the label display to word, scale. Hence the visual as below.

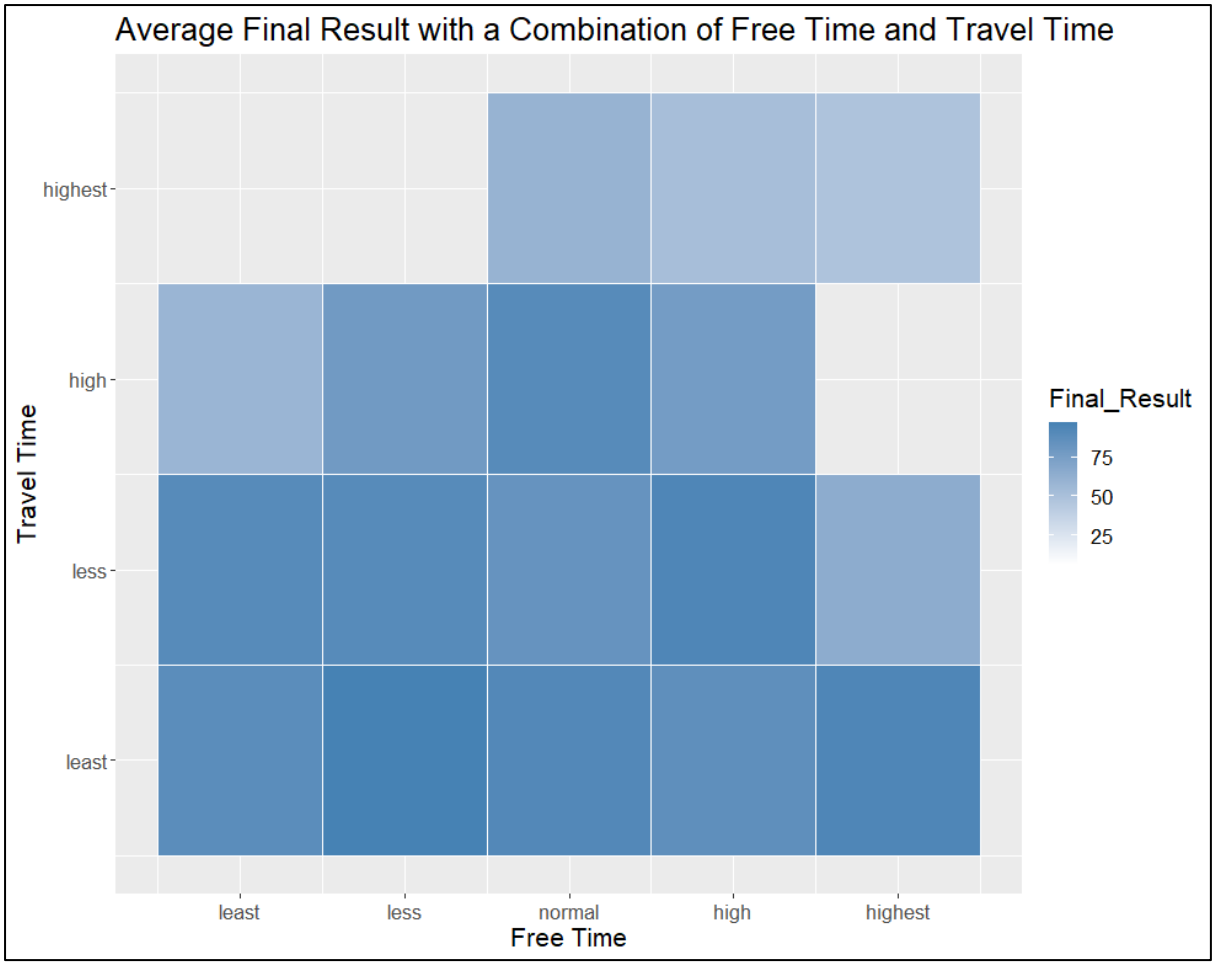

Figure 95: average final result with a combination of free time and travel time graph

Explanation:

From bottom half of travel time heat map, it shows that it is better in performance which mean travel once in a while will increase student performance because it is useful to destress students but if too often then the student performance will drop.

```
# Heat Map
ggplot(Question4Data, aes(x = Free_Time, y = Study_Time)) +
  geom_tile(aes(fill = Final_Result), color = "white") +
  scale_fill_gradient(low = "white", high = "steel blue") +
  labs(title = "Average Final Result with a Combination of Free Time and Study Time",
     x = "Free Time",
     y = "Study Time") +
  scale_x_continuous(labels = c("least", "less", "normal", "high", "highest"),
                     breaks = 1:5) +
  scale_y_continuous(labels = c("least", "less", "high", "highest"),
                     breaks = 1:4) +
  theme(text=element_text(size = 16))
```

Figure 96: average final result with a combination of free time and study time code

A heat map is created to display the average final result with a combination of free time and study time. For mappings, Free Time will be the x-axis and Study_Time be the y-axis. The tile fill colour will be depending on Final_Result and using white colour line to separate the tiles, the fill colour is changed to from white to steel blue. Then, the graph's label is given using labs() while scale_x_continuous() and scale_y_continuous() used to change the label display to word, scale. Hence the visual as below.

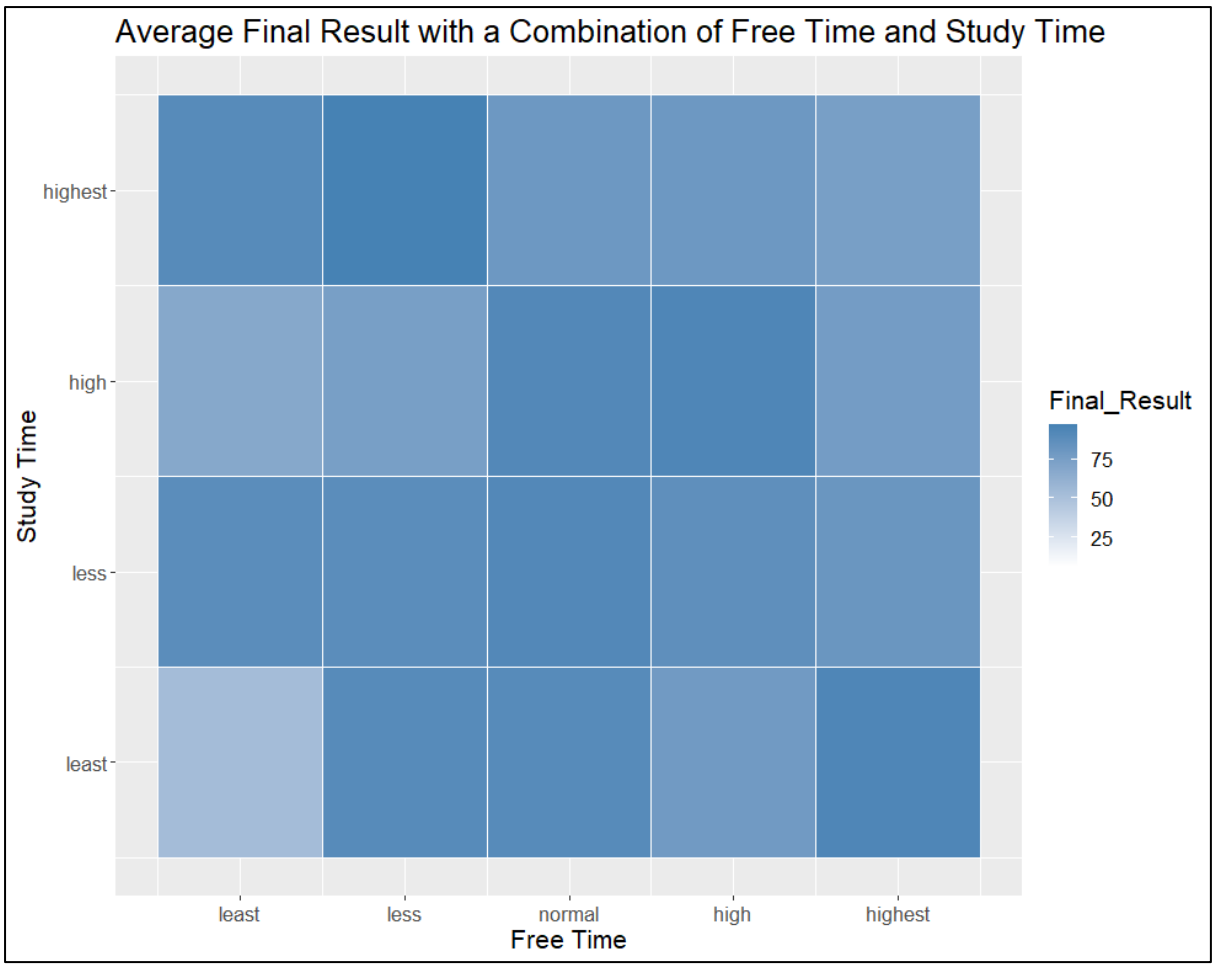

Figure 97: average final result with a combination of free time and study time graph

Explanation:

From study time heat map, the centre of the heat map shows more stable final result than the surroundings which mean balanced time management is more suitable for students for increase performance.

Question 4 Conclusion

Both health/stress status and study time are important to a student, if only one aspect is always focus on it will eventually decrease the student's performance and health status hence a balanced lifestyle is more suitable for students.

For summarise, a table is created to shows the relation between each attributes and performance produced in this whole question analysis for a simple understanding of each studied attributes as shown below.

Attributes	Affect to Performance
Daily_Alcohol	Slightly
Health_Status	Slightly
Free_Time	Indirectly
Hang_Out	Indirectly
Cocurricular_Activity	Slightly
Travel_Time	Indirectly

Attributes	Improve performance
High daily alcohol consumed	Slightly
Low health status	Slightly
Male participate in co-curricular/ Female no participate in co-corricular	Slightly

Question 5: Is fundamental and extra education a must for student?

This question will be focusing on analysing how important is fundamental and extra education for students and conclude that whether them are a must or not, so students can be more efficient on studying. Therefore, some main attributes are getting targeted for this question including School_Extra_EduSup, Family_Extra_EduSup, Paid_Extra_Class Attended_Nursery_School. Some of the cause and inter-related attributes may also be found in the process of analysis, hence this question is worth to take a look into.

Data Exploration

Since View() function is used beforehand, so there is no need to use any other function to determine the range or levels of the data. Rather, some very direct and simple graphs are used to get some insight and idea about the brief relationship between each other, every attribute will be put in the same graph with G1, G2 and G3 in order to not miss any interesting relationship. Only stacked histogram will be used in this question.

```
# School_Extra_EduSup
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = School_Extra_EduSup)) +
  labs(title = "Frequency Distribution of G1 Result by School Extra Educational Support",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = School_Extra_EduSup)) +
  labs(title = "Frequency Distribution of G2 Result by School Extra Educational Support",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = School_Extra_EduSup)) +
  labs(title = "Frequency Distribution of G3 Result by School Extra Educational Support",
       x = "G3",
       y = "Number of Students")
```

Figure 98: School_Extra_EduSup Exploration

```
# Family_Extra_EduSup
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Family_Extra_EduSup)) +
  labs(title = "Frequency Distribution of G1 Result by Family Extra Educational Support",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Family_Extra_EduSup)) +
  labs(title = "Frequency Distribution of G2 Result by Family Extra Educational Support",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Family_Extra_EduSup)) +
  labs(title = "Frequency Distribution of G3 Result by Family Extra Educational Support",
       x = "G3",
       y = "Number of Students")
```

Figure 99: Family_Extra_EduSup Exploration

```
# Paid_Extra_Class
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Paid_Extra_Class)) +
  labs(title = "Frequency Distribution of G1 Result by Paid Extra Class",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Paid_Extra_Class)) +
  labs(title = "Frequency Distribution of G2 Result by Paid Extra Class",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Paid_Extra_Class)) +
  labs(title = "Frequency Distribution of G3 Result by Paid Extra Class",
       x = "G3",
       y = "Number of Students")
```

Figure 100: Paid_Extra_Class Exploration

```
# Attended_Nursery_School
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Attended_Nursery_School)) +
  labs(title = "Frequency Distribution of G1 Result by Attended Nursery School",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Attended_Nursery_School)) +
  labs(title = "Frequency Distribution of G2 Result by Attended Nursery School",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Attended_Nursery_School)) +
  labs(title = "Frequency Distribution of G3 Result by Attended Nursery School",
       x = "G3",
       y = "Number of Students")
```

Figure 101: Attended_Nursery_School Exploration

Data Manipulation and Transformation

```
Question5Data = students %>%
  mutate(Final_Result = round((G1+G2+G3)/60*100, digits = 2)) %>%
  select(Sex, School_Extra_EduSup, Family_Extra_EduSup,
         Paid_Extra_Class, Attended_Nursery_School, Final_Result,
         Study_Time, Health_Status) %>%
  arrange(Final_Result)
# Arrange columns name
Question5Data = Question5Data %>%
  select(order(colnames(Question5Data)))

View(Question5Data)
str(Question5Data)
summary(Question5Data)
```

Figure 102: Question 5 Data Manipulation and Transformation

A specified sub-dataset named Question5Data is created in this phase by using the dataset after pre-processing, piping is widely used right here to show a more readable source code. First of all, an overall performance in three years result is produced in new a column named Final_Result by using mutate() function with a simple mathematical formula, before create the column the calculation result will round up to 2 decimal place using round() function by giving desired decimal place to digits parameter. After that, every main attribute stated in data exploration will be selected so that no other column will be in the new dataset, then by using arrange() function the dataset will be sorted ascendingly following Final Result in rows. Lastly, select() function is used again with order() and colnames() function to sort the column this time with alphabet ascendingly. The fact that sorting rows and column is not executed together is because Final_Result is not mutate completely yet, so order() function cannot find the column name. In the end of data manipulation and data exploration, View(), summary() and str() function also can be used to display a more simple details of the data as shown below.

```
> str(Question5Data)
'data.frame':   922 obs. of  8 variables:
 $ Attended_Nursery_School: chr  "no" "no" "yes" "yes" ...
 $ Family_Extra_EduSup    : chr  "yes" "yes" "yes" "yes" ...
 $ Final_Result           : num  6.67 6.67 8.33 8.33 8.33 8.33 10 10 10 10 ...
 $ Health_Status          : int  5 5 5 4 5 4 3 3 3 3 ...
 $ Paid_Extra_Class       : chr  "no" "no" "no" "no" ...
 $ School_Extra_EduSup    : chr  "no" "no" "no" "no" ...
 $ Sex                    : chr  "Female" "Female" "Male" "Male" ...
 $ Study_Time             : int  1 1 1 1 1 1 1 2 1 2 ...
```

Figure 103: Question 5 Data datatype

```
> summary(Question5Data)
 Attended_Nursery_School Family_Extra_EduSup  Final_Result   Health_Status
 Length:922              Length:922           Min.   : 6.67   Min.   :1.000
 Class :character        Class :character     1st Qu.:41.67   1st Qu.:3.000
 Mode  :character        Mode  :character     Median :53.33   Median :4.000
                                              Mean   :53.62   Mean   :3.565
                                              3rd Qu.:66.67   3rd Qu.:5.000
                                              Max.   :96.67   Max.   :5.000
 Paid_Extra_Class   School_Extra_EduSup     Sex                Study_Time
 Length:922         Length:922           Length:922         Min.   :1.000
 Class :character   Class :character     Class :character   1st Qu.:1.000
 Mode  :character   Mode  :character     Mode  :character   Median :2.000
                                                           Mean   :2.037
                                                           3rd Qu.:2.000
                                                           Max.   :4.000
```

Figure 104: Question 5 Data summary

Analysis 1: Determine the Relationship between Attended Nursery School, Final Result and Study Time

Data Visualisation

```
#Boxplot
ggplot(Question5Data, aes(x = Attended_Nursery_School, y = Final_Result)) +
  geom_boxplot(aes(group = Attended_Nursery_School)) +
  labs(title = "Average of Final Result for Attended Nursery School",
       x = "Attended Nursery School",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 105: average of final result for attended nursery school code

A boxplot is created to display the average of final result for attended nursery school. For mappings, Attended_Nursery_School will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Attended_Nursery_School. Then, the graph's label is given using labs(). Hence the visual as below.

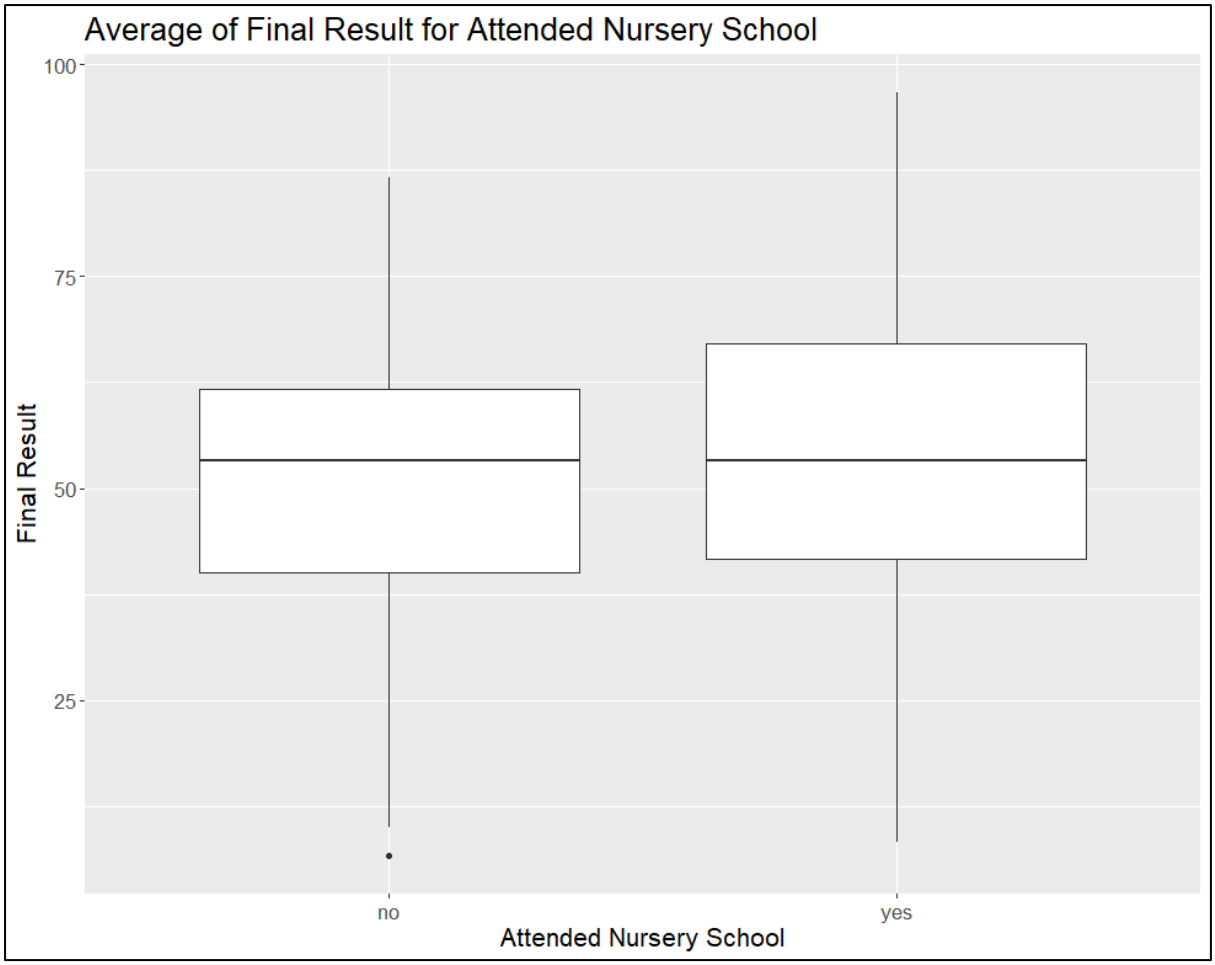

Figure 106: average of final result for attended nursery school graph

The summary of the graph include:

1. Students who attended nursery school before can get better performance.

Explanation:

Students who attended nursery school will act differently compared to students who did not. The reason will be studied in the next graph.

```
#Count plot
ggplot(Question5Data, aes(x = Attended_Nursery_School, y = Study_Time)) +
  geom_count() +
  labs(title = "Average of Study Time for Attended Nursery School",
       x = "Attended Nursery School",
       y = "Study Time") +
  theme(text=element_text(size = 16))
```

Figure 107: average of study time for attended nursery school code

A count plot is created to display the average of study time for attended nursery school. For mappings, Attended_Nursery_School will be the x-axis and Study_Time be the y-axis. Then, the graph's label is given using labs(). Hence the visual as below.

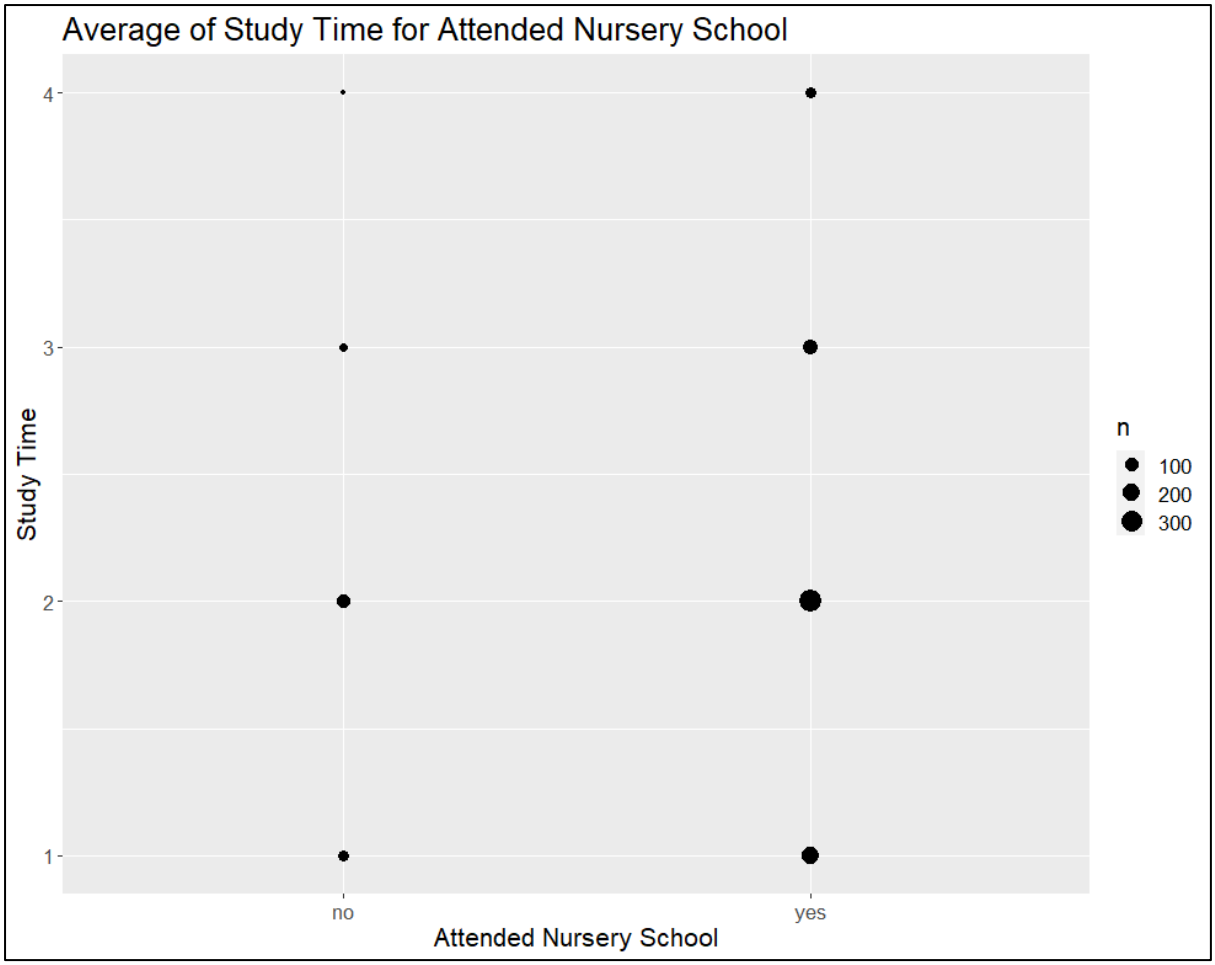

Figure 108: average of study time for attended nursery school graph

The summary of the graph include:

1. Students who attended nursery school will put more time on study.

Explanation:

Question from previous graph can be explained in this graph, students who attended nursery school will have a stronger awareness on importance of education because they start learning still young and cultivate a learning mindset referenced from a study (Bright Horizions, 2020). That is why the performance improved since study time increase according to analysis before.

Analysis 2: Determine the Relationship between Family and School Extra Educational Support with Final Result

Data Visualisation

```
# Heat map
ggplot(Question5Data, aes(x = Family_Extra_EduSup, y = School_Extra_EduSup)) +
  geom_tile(aes(fill = Final_Result)) +
  labs(title = "Average Final Result with a Combination of Family and School Extra Educational Support",
       x = "Family Extra Educational Support",
       y = "School Extra Educational Support") +
  theme(text=element_text(size = 16))
```

Figure 109: average final result with a combination of family and school extra educational support code

A heat map is created to display the average final result with a combination of family and school extra educational support. For mappings, Family_Extra_EduSup will be the x-axis and School_Extra_EduSup be the y-axis. The tile fill colour will be depending on Final_Result and using white colour line to separate the tiles. Then, the graph's label is given using labs(). Hence the visual as below.

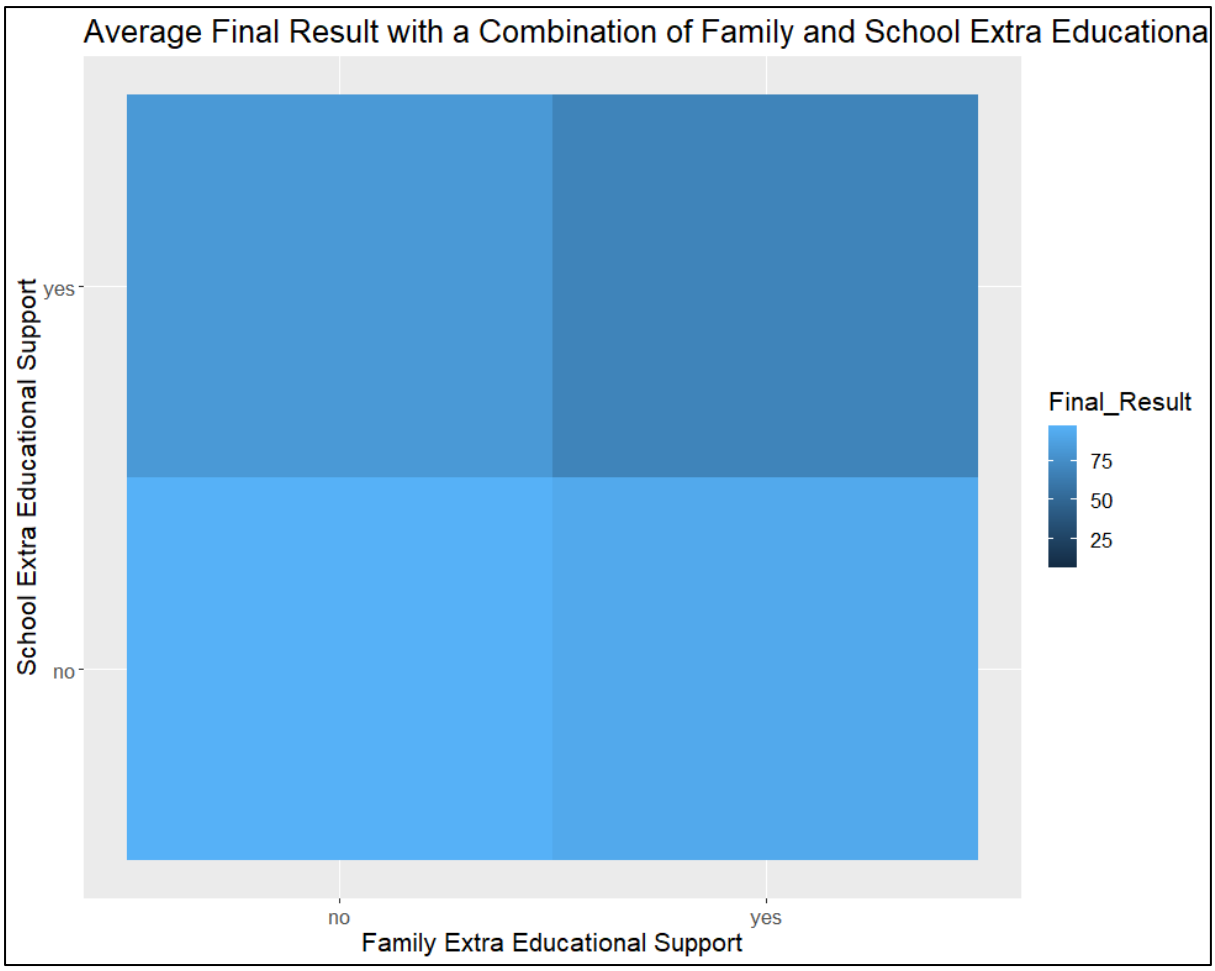

Figure 110: average final result with a combination of family and school extra educational support graph

The summary of the graph include:

1. Students who do not have both extra educational support from family and school have the highest performance.
2. Students who have both extra educational support from family and school have the lowest performance.

Explanation:

Most of the students do not like to have extra educational support because it will interrupt the student learning pace and destroy the interest of them to learn.

Analysis 3: Determine the Relationship between Paid Extra Class, Final Result and Health Status

Data Visualisation

```
# Boxplot
ggplot(Question5Data, aes(x = Paid_Extra_Class, y = Final_Result)) +
  geom_boxplot(aes(group = Paid_Extra_Class)) +
  labs(title = "Average of Final Result for Paid Extra Class",
       x = "Paid Extra Class",
       y = "Final Result") +
  theme(text=element_text(size = 16))
```

Figure 111: average of final result for paid extra class code

A boxplot is created to display the average of final result for paid extra class. For mappings, Paid_Extra_Class will be the x-axis and Final_Result be the y-axis. Boxplot will be grouped in Paid_Extra_Class. Then, the graph's label is given using labs(). Hence the visual as below.

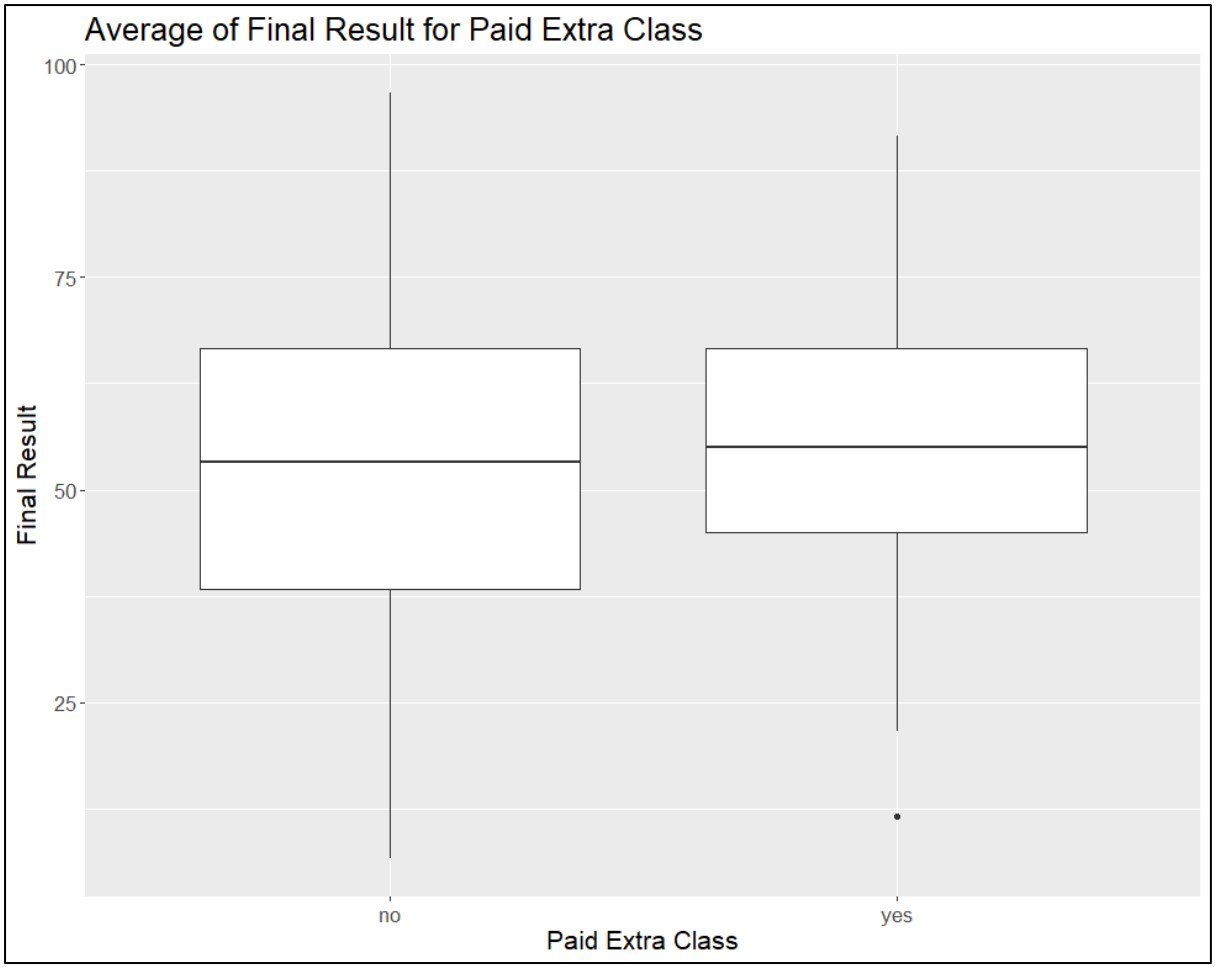

Figure 112: average of final result for paid extra class graph

The summary of the graph include:

1. Students who do not have paid extra class will have larger result range but higher for highest limit.
2. Students who do not have paid extra class will have smaller result range but there is outlier exists.

Explanation:

Apply for paid extra class only can shrink the average possible performance range, so the student has higher chance to pass the exam but some there is also some exception to it as there is outlier in the graph. It is because paid extra class is not suitable for all student learning pace, some students might have opposite impact from it.

Question 5 Conclusion

Fundamental and extra education is not a must for student because it will not change a student's performance so much but by having them will help some students on their study while some students may decrease their performance instead.

For summarise, a table is created to shows the relation between each attributes and performance produced in this whole question analysis for a simple understanding of each studied attributes as shown below.

Attributes	Affect to Performance
School_Extra_EduSup	Slightly
Family_Extra_EduSup	Slightly
Paid_Extra_Class	Slightly
Attended_Nursery_School	Slightly

Attributes	Improve performance
No school and family extra educational support	Slightly
Go for paid extra class	Slightly
Attended nursery school before	Slightly

In my opinion, family should not force student for any extra education, instead they should listen to student’s opinion and schedule a more suitable plan for an individual student so that it does not cause negative impacts.

Question 6: How family aspect related to student's performance?

This question will be focusing on identifying what is the family aspects that make impact to student performance so the result can help to predict an expected performance in academic for different kind of family background. Therefore, some main attributes are getting targeted for this question including Mother_Education, Father_Education, Mother_Job_Type, Father_Job_Type, Guardian. Some of the cause and inter-related attributes may also be found in the process of analysis, hence this question is worth to take a look into.

Data Exploration

Since View() function is used beforehand, so there is no need to use any other function to determine the range or levels of the data. Rather, some very direct and simple graphs are used to get some insight and idea about the brief relationship between each other, every attribute will be put in the same graph with G1, G2 and G3 in order to not miss any interesting relationship. Different graphs will be used depends on whether the data is in continuous or discrete, therefore in this question, stacked histogram and boxplot are used shown as below.

```
# Mother_Education
ggplot(students, aes(x = G1, y = Mother_Education)) +
  geom_boxplot(aes(group = Mother_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Mother Education",
       x = "G1",
       y = "Mother Education")
ggplot(students, aes(x = G2, y = Mother_Education)) +
  geom_boxplot(aes(group = Mother_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Mother Education",
       x = "G2",
       y = "Mother Education")
ggplot(students, aes(x = G3, y = Mother_Education)) +
  geom_boxplot(aes(group = Mother_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Mother Education",
       x = "G3",
       y = "Mother Education")
```

Figure 113: Mother_Education Exploration

```
# Father_Education
ggplot(students, aes(x = G1, y = Father_Education)) +
  geom_boxplot(aes(group = Father_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Father Education",
       x = "G1",
       y = "Father Education")
ggplot(students, aes(x = G2, y = Father_Education)) +
  geom_boxplot(aes(group = Father_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Father Education",
       x = "G2",
       y = "Father Education")
ggplot(students, aes(x = G3, y = Father_Education)) +
  geom_boxplot(aes(group = Father_Education)) +
  labs(title = "Frequency Distribution of G3 Result by Father Education",
       x = "G3",
       y = "Father Education")
```

Figure 114: Father_Education Exploration

```
# Mother_Job_Type
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Mother_Job_Type)) +
  labs(title = "Frequency Distribution of G3 Result by Mother Job Type",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Mother_Job_Type)) +
  labs(title = "Frequency Distribution of G3 Result by Mother Job Type",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Mother_Job_Type)) +
  labs(title = "Frequency Distribution of G3 Result by Mother Job Type",
       x = "G3",
       y = "Number of Students")
```

Figure 115: Mother_Job_Type Exploration

```
# Father_Job_Type
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Father_Job_Type)) +
  labs(title = "Frequency Distribution of G3 Result by Father Job Type",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Father_Job_Type)) +
  labs(title = "Frequency Distribution of G3 Result by Father Job Type",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Father_Job_Type)) +
  labs(title = "Frequency Distribution of G3 Result by Father Job Type",
       x = "G3",
       y = "Number of Students")
```

Figure 116: Father_Job_Type Exploration

```
# Guardian
ggplot(students, aes(x = G1)) +
  geom_histogram(binwidth = 5, aes(fill = Guardian)) +
  labs(title = "Frequency Distribution of G3 Result by Guardian",
       x = "G1",
       y = "Number of Students")
ggplot(students, aes(x = G2)) +
  geom_histogram(binwidth = 5, aes(fill = Guardian)) +
  labs(title = "Frequency Distribution of G3 Result by Guardian",
       x = "G2",
       y = "Number of Students")
ggplot(students, aes(x = G3)) +
  geom_histogram(binwidth = 5, aes(fill = Guardian)) +
  labs(title = "Frequency Distribution of G3 Result by Guardian",
       x = "G3",
       y = "Number of Students")
```

Figure 117: Guardian Exploration

Data Manipulation and Transformation

```
# Main data sets
Question6Data = students %>%
  mutate(Final_Result = round((G1+G2+G3)/60*100, digits = 2),
         Parent_Education = round((Mother_Education + Father_Education)/2), digits = 0)%>%
  select(Parent_Education, Mother_Education, Father_Education,
         Mother_Job_Type,  Father_Job_Type,
         Guardian, Final_Result) %>%
  arrange(Final_Result)
# Arrange columns name
Question6Data = Question6Data %>%
  select(order(colnames(Question6Data)))

View(Question6Data)
str(Question6Data)
summary(Question6Data)
```

Figure 118: Question 6 Data Manipulation and Transformation

A specified sub-dataset named Question6Data is created in this phase by using the dataset after pre-processing, piping is widely used right here to show a more readable source code. First of all, an overall performance in three years result is produced in new a column named Final_Result by using mutate() function with a simple mathematical formula, before create the column the calculation result will round up to 2 decimal place using round() function by giving desired decimal place to digits parameter, an average parent education also be created to have an overall view. After that, every main attribute stated in data exploration will be selected so that no other column will be in the new dataset, then by using arrange() function the dataset will be sorted ascendingly following Final Result in rows. Lastly, select() function is used again with order() and colnames() function to sort the column this time with alphabet ascendingly. The fact that sorting rows and column is not executed together is because Final_Result is not mutate completely yet, so order() function cannot find the column name. In the end of data manipulation and data exploration, View(), summary() and str() function also can be used to display a more simple details of the data as shown below.

```
> str(Question6Data)
'data.frame':   922 obs. of  7 variables:
 $ Father_Education: int  3 3 1 2 1 2 3 1 3 1 ...
 $ Father_Job_Type : chr  "other" "other" "other" "at_home" ...
 $ Final_Result    : num  6.67 6.67 8.33 8.33 8.33 8.33 10 10 10 10 ...
 $ Guardian        : chr  "other" "other" "mother" "mother" ...
 $ Mother_Education: int  3 3 2 3 2 3 4 2 4 2 ...
 $ Mother_Job_Type : chr  "other" "other" "other" "services" ...
 $ Parent_Education: num  3 3 2 2 2 2 4 2 4 2 ...
```

Figure 119: Question 6 Data datatype

```
> summary(Question6Data)
 Father_Education Father_Job_Type     Final_Result     Guardian         Mother_Education
 Min.   :0.000    Length:922         Min.   : 6.67   Length:922         Min.   :0.000
 1st Qu.:2.000    Class :character   1st Qu.:41.67   Class :character   1st Qu.:2.000
 Median :2.500    Mode  :character   Median :53.33   Mode  :character   Median :3.000
 Mean   :2.536                       Mean   :53.62                      Mean   :2.753
 3rd Qu.:3.000                       3rd Qu.:66.67                      3rd Qu.:4.000
 Max.   :4.000                       Max.   :96.67                      Max.   :4.000
 Mother_Job_Type    Parent_Education
 Length:922         Min.   :0.000
 Class :character   1st Qu.:2.000
 Mode  :character   Median :2.000
                    Mean   :2.705
                    3rd Qu.:4.000
                    Max.   :4.000
```

Figure 120: Question 6 Data summary

```
# Focus on student with mother as guardian data sets
students_mother = Question6Data %>%
  filter(Guardian == "mother")

View(students_mother)
summary(students_mother)
```

Figure 121: student_mother creation

A sub-dataset named student_mother also being created from Question6Data which will focus on students who having mother as their guardian by using filter() function to have the condition.

```
# Focus on student with father as guardian data sets
students_father = Question6Data %>%
  filter(Guardian == "father")

View(students_father)
summary(students_father)
```

Figure 122: student_father creation

A sub-dataset named student_father also being created from Question6Data which will focus on students who having father as their guardian by using filter() function to have the condition.

Analysis 1: Determine the Relationship between Parent Education and Final Result

Data Visualisation

```
# Count plot and Regression line
ggplot(Question6Data, aes(x = Parent_Education, y = Final_Result)) +
  geom_count() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relationship between Parent Education and Final Result",
       x = "Parent Education",
       y = "Final Result") +
  scale_x_continuous(labels = c("None", "4th Grade", "5th-9th Grade", "Secondary", "Higher"),
                     breaks = 0:4) +
  theme(text=element_text(size = 16))
```

Figure 123: relationship between parent education and final result code

A combination of count plot and regression line graph is created to display the relationship between parent education and final result. For mappings, Parent_Education will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

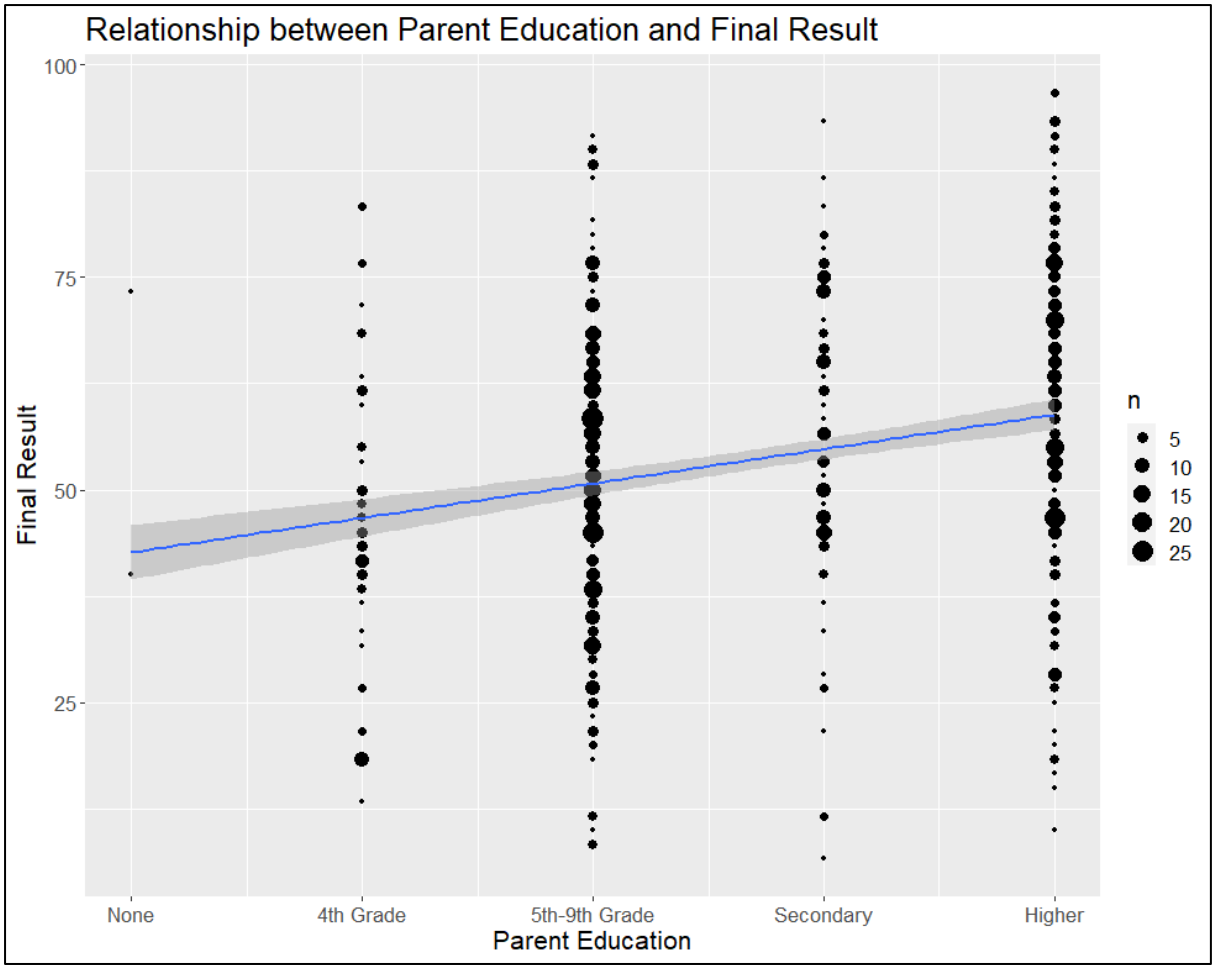

Figure 124: relationship between parent education and final result graph

The summary of the graph include:

1. Most of the parents have 5^{th}-9^{th} education level or higher education level.
2. The higher the parent education level is, the better the student performance.

Explanation:

According to a study (Abu Bakar, Mamat, & Ibrahim, 2017), parent education improve performance might because of it is from the genetic aspect, student who born from well-educated parents will be cleverer and more talented. Besides that, educated parents also might give them a very good young education, so that they know the importance of education and how useful it is.

Analysis 2: Determine the Relationship between Mother Education and Final Result

Data Visualisation

```
# Count plot and Regression line
ggplot(students_mother,aes(x = Mother_Education, y = Final_Result)) +
  geom_count() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relationship between Mother Education and Final Result",
       x = "Mother Education",
       y = "Final Result") +
  scale_x_continuous(labels = c("None", "4th Grade", "5th-9th Grade", "Secondary", "Higher"),
                     breaks = 0:4) +
  theme(text=element_text(size = 16))
```

Figure 125: relationship between mother education and final result code

A combination of count plot and regression line graph is created to display the relationship between mother education and final result by using students_mother dataset created before. For mappings, Mother_Education will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

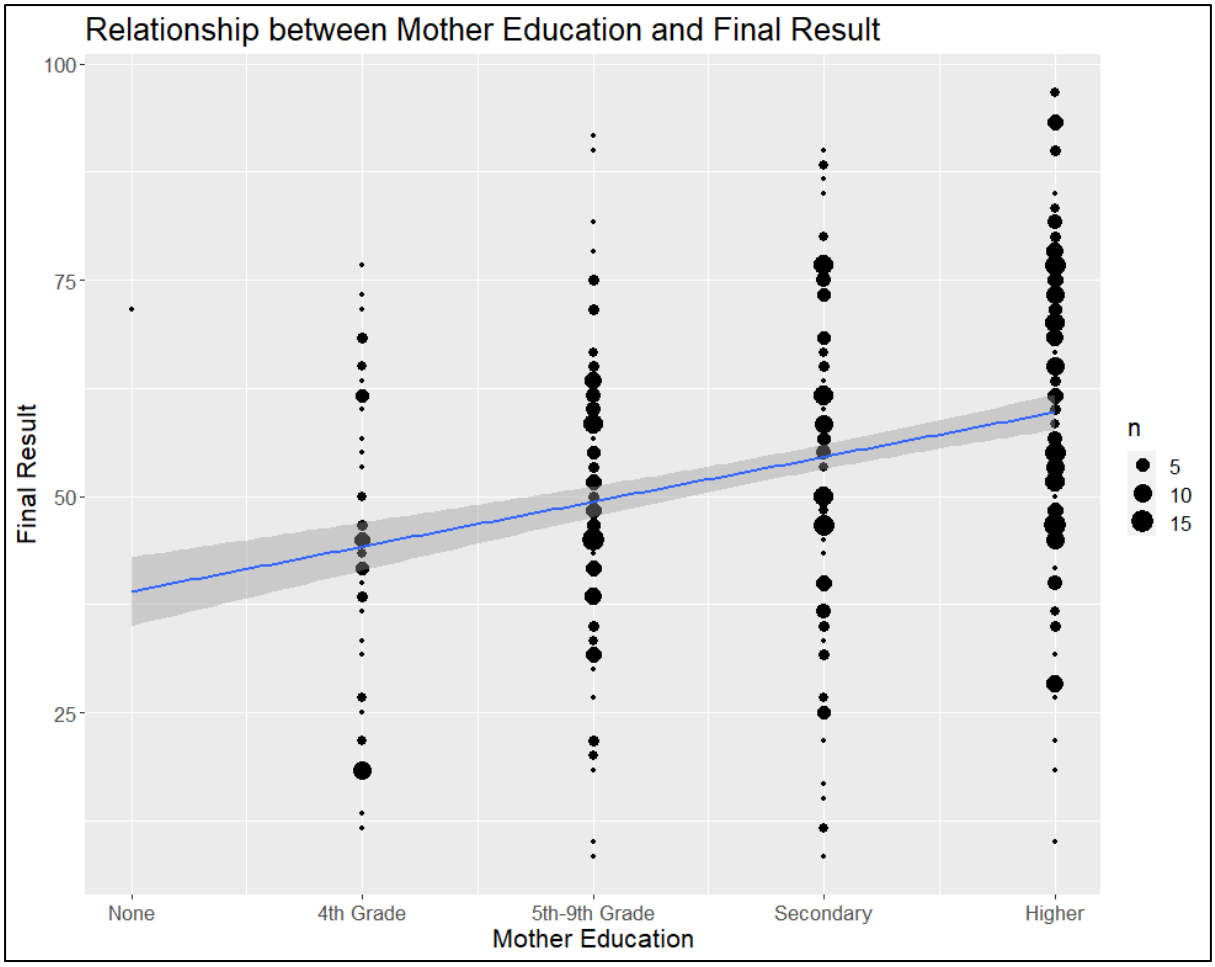

Figure 126: relationship between mother education and final result graph

The summary of the graph include:

1. Most of the mothers have higher education level.
2. The higher the mother education level is, the better the student performance.

Explanation:

As expected, higher education level will increase performance but this analysis is to used to compared to father as a guardian situation too, so a better explanation can be concluded in later analysis.

```
# Stacked Bar Graph
ggplot(students_mother,aes(x =Mother_Education )) +
  geom_bar(aes(fill = Mother_Job_Type)) +
  labs(title = "Distribution of Mother Job Type in Different Mother Education",
       x = "Mother Education",
       y = "Number of Mothers") +
  scale_x_continuous(labels = c("None", "4th Grade", "5th-9th Grade", "Secondary", "Higher"),
                     breaks = 0:4) +
  theme(text=element_text(size = 16))
```

Figure 127: relationship between mother education and final result code

A stacked bar graph is created to display the relationship between mother education and final result by using students_mother dataset created before. For mappings, Mother_Education will be the x-axis. Bar filled colour is depends on Mother_Job_Type. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

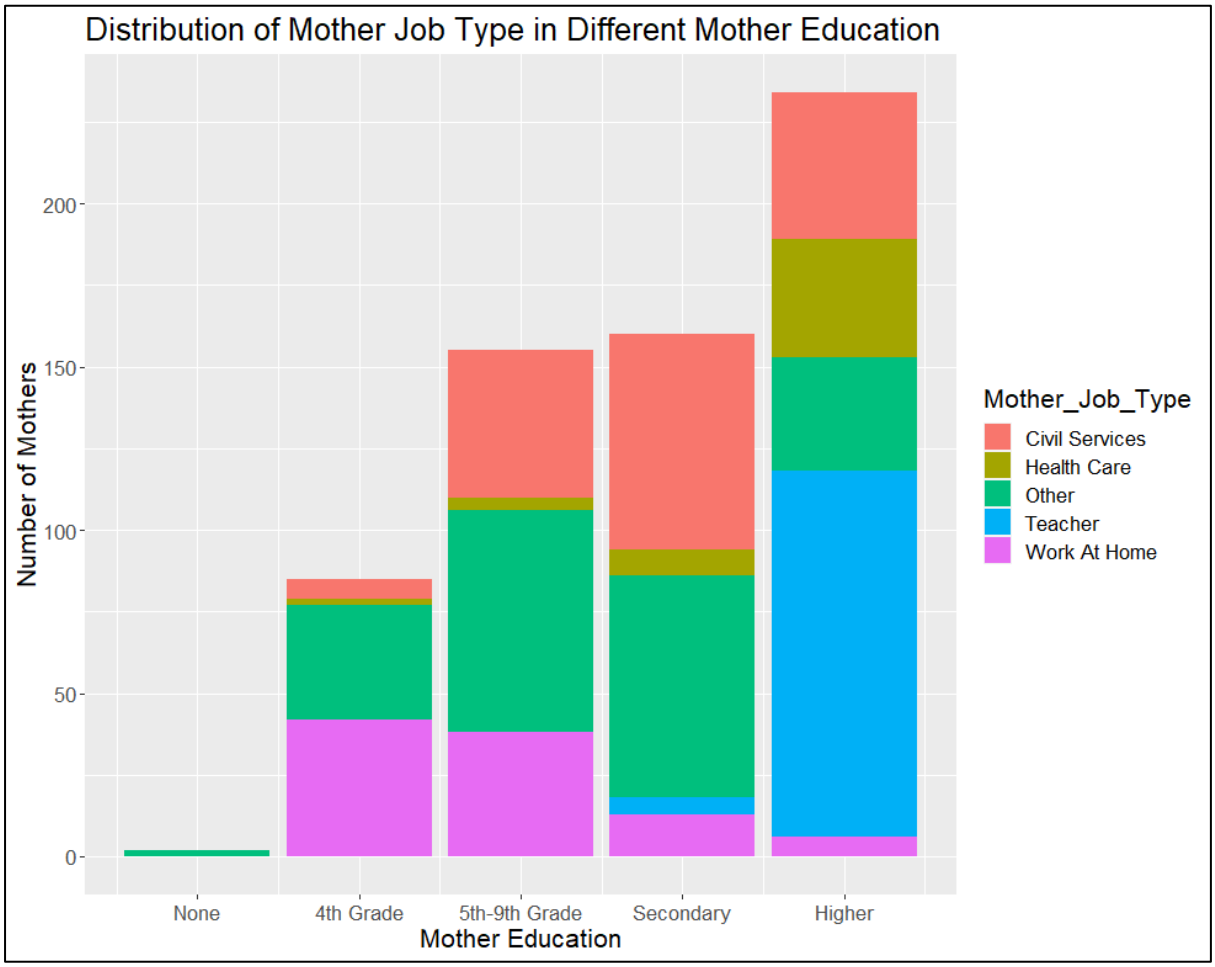

Figure 128: relationship between mother education and final result graph

The summary of the graph include:

1. Civil Services mostly is secondary
2. Health Care increasing
3. Other Mostly 5th-9th grade and secondary
4. Teacher increasing
5. Work at Home decreasing

Explanation:

Teacher is having higher education level compared to other job type because they need to teach students who are new to a subject.

Analysis 3: Determine the relationship between Father Education and Final Result

Data Visualisation

```
# Count plot and Regression line
ggplot(students_father,aes(x = Father_Education, y = Final_Result)) +
  geom_count() +
  geom_smooth(method = "lm", formula = y ~ x) +
  labs(title = "Relationship between Father Education and Final Result",
       x = "Father Education",
       y = "Final Result") +
  scale_x_continuous(labels = c("None", "4th Grade", "5th-9th Grade", "Secondary", "Higher"),
                     breaks = 0:4) +
  theme(text=element_text(size = 16))
```

Figure 129: relationship between father education and final result code

A combination of count plot and regression line graph is created to display the relationship between father education and final result by using students_father dataset created before. For mappings, Father_Education will be the x-axis and Final_Result be the y-axis. Method is being changed to lm for regression line and using formula y ~ x. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

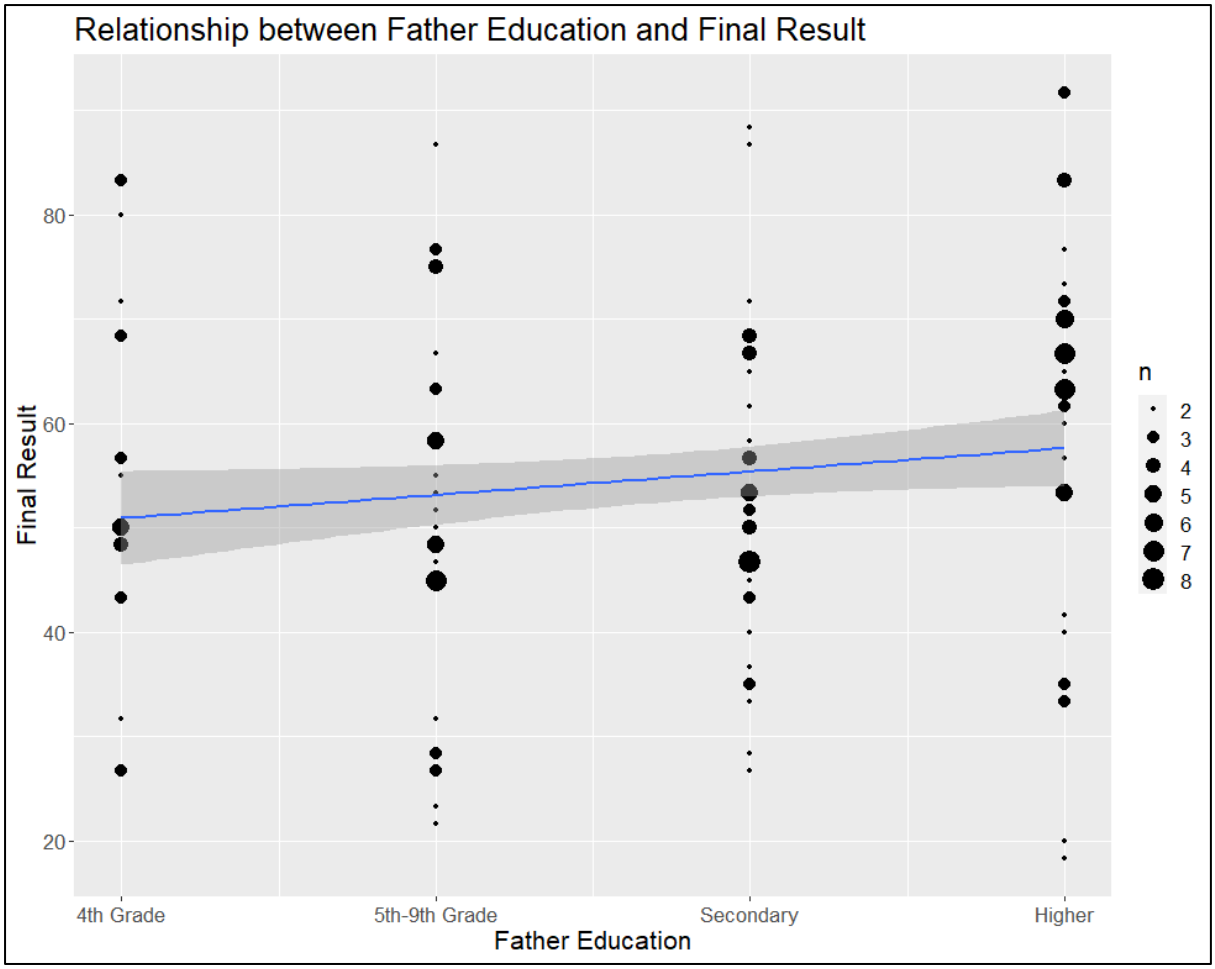

Figure 130: relationship between father education and final result graph

The summary of the graph include:

1. The higher the father education level is, the better the student performance.

Explanation:

First, we can observe that the lowest y-axis value is over passing mark which is different from mother as guardian situation where students might fail the exam. Therefore, we can say that when father is the guardian, it will be more strict for the student, so they will try fully for a passing mark while mother tends to be more worries about student emotion and opinion.

```
# Stacked Bar Graph
ggplot(students_father,aes(x =Father_Education )) +
  geom_bar(aes(fill = Father_Job_Type)) +
  labs(title = "Distribution of Father Job Type in Different Father Education",
       x = "Father Education",
       y = "Number of Fathers") +
  scale_x_continuous(labels = c("None", "4th Grade", "5th-9th Grade", "Secondary", "Higher"),
                     breaks = 0:4) +
  theme(text=element_text(size = 16))
```

Figure 131: relationship between father education and final result code

A stacked bar graph is created to display the relationship between father education and final result by using students_father dataset created before. For mappings, Father_Education will be the x-axis. Bar filled colour is depends on Father_Job_Type. Then, the graph's label is given using labs() and scale_x_continuous() used to change the label display to word. Hence the visual as below.

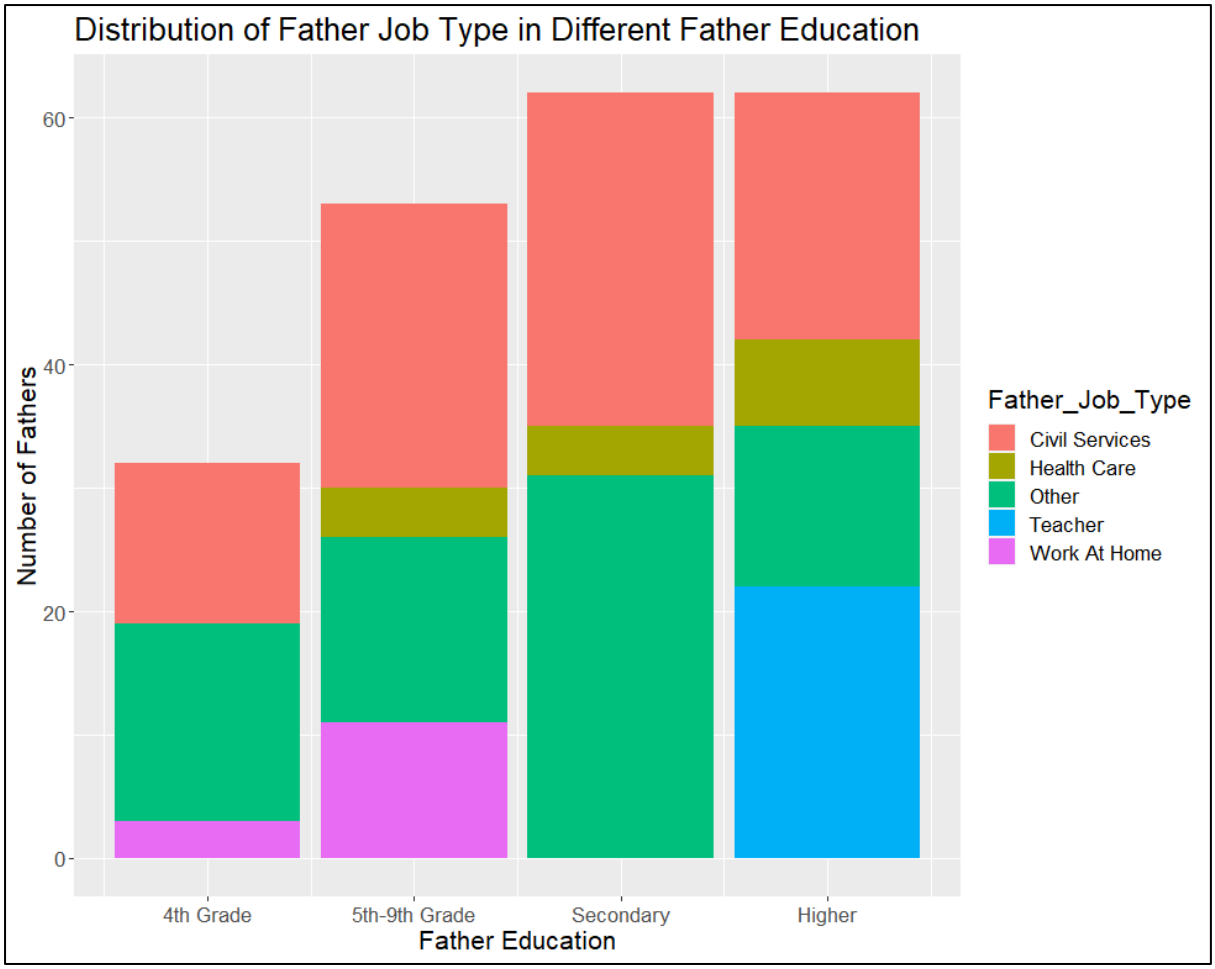

Figure 132: relationship between father education and final result graph

The summary of the graph include:

1. Civil Services mostly is secondary
2. Health Care increasing
3. Other Mostly secondary
4. Teacher only in Higher Education
5. Work At Home increasing

Explanation:

Every job type is increasing in frequency as the education level goes up, so in this dataset father tends to be more educated.

Question 6 Conclusion

Parent Education will be making changes on student's performance, it could be genetic or young education such as the parent will give the student mindset of how important education is in the society. When father is the guardian with high education, the performance range is more shrink and lower limit is higher compared to mother guardian with high education.

For summarise, a table is created to shows the relation between each attributes and performance produced in this whole question analysis for a simple understanding of each studied attributes as shown below.

Attributes	Affect to Performance
Parent_Education	High
Mother_Education	High
Father_Education	Slightly

Attributes	Improve performance
High parent education	High
High educated mother as guardian	High
High educated father as guardian	Slightly

Additional Features

1. Cleaning data – gsub()

One of the extra features sued is gsub() from base library for data cleaning, it is very useful especially when in this analysis dataset is given through an excel csv file, therefore there will be some unnecessary character exists within the same data slot by replacing a specified pattern to a suitable value. However, gsub() can only be used to modify vector data type, so for students dataset readLines() is used first to create vector instead of data frame. To demonstrate, a simple vector is created with some weird pattern exists within each data slot, then gsub() will be doing the job as shown below.

```
# Create a character vector
a = c("a1", "2a", "a3a","aaa4", "5aa")
class(Sampledf)

# Remove unnecessary character and convert it to numeric and put in a data frame
Modified_a = gsub("a", "", Sampledf)
SampleTable = read.table(text = Modifieddf)
str(SampleTable)
```

Figure 133: Cleaning data code

From the figure above, we can see that “a” pattern is randomly exists within the every data slot, hence by giving “a” as pattern and an empty character “” for gsub(), it will automatically replace pattern to empty character. As for the advantages of doing this, it includes data is automatically convert to suitable data type when creating a data frame to stored a clean data. If the data is alphanumeric, then most likely the data frame will treat that column as a character data type which will take extra step to convert it to suitable one. Data type before and after modified is shown as below.

```
> class(Sampledf)
[1] "character"
```

Figure 134: Sampledf datatype

```
> str(SampleTable)
'data.frame':   5 obs. of  1 variable:
 $ V1: int  1 2 3 4 5
```

Figure 135: SampleTable datatype

2. Display multiple inter-related graphs – plot_grid()

Besides that, plot_grid() function from cowplot library were also be used throughout this analysis as an addition feature for display multiple inter-related graphs. It will be very useful when exploring data in RStudio platform because when only observe one single graph, most of the time nothing is interesting and useful to generate some information, so this function will display all of them in desired dimension for deeper understanding and better interpret of students.csv dataset. To demonstrate, two simple data frame data are created and two data graphs together in the same time by using plot_grid() as shown below.

```
# Create sample data frame
df_a = data.frame(sample_row = c(1,2,1,1,2))
df_b = data.frame(sample_row = c(2,1,2,2,1))

# Require library "cowplot" & "ggplot2"
# Display both graph in desired dimension
plot_grid(ggplot(df_a, aes(x = sample_row)) +
            geom_bar(),
          ggplot(df_b, aes(x = sample_row)) +
            geom_bar(),
          nrow = 2,
          ncol = 1)
```

Figure 136: plot_grid code

Figure above shows that plot_grid can change dimension by giving desired dimension value to nrow and ncol. If those value are not given, plot_grid() will decide that itself. For addition information, plot_grid can be feed as much as graphs you want in the parameter, hence it is very useful for data exploration and visualization. The output is shown as below.

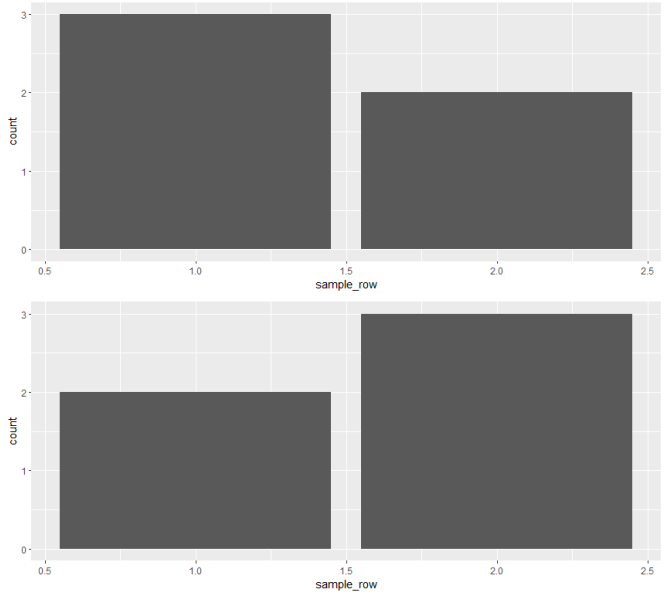

Figure 137: plot_grid graph

3. Density graph– geom_density2d()

Density2d graph is very useful when continuous data will be on the both axis in a graph. If count plot or boxplot are used then it is hard to see an overall view and the graph will be very messy, hence a density2d graph will be the most suitable one for this situation. For example, final result and absences in this analysis need to put in the same graph, then density2d graph is used. To demonstrate, diamonds datasets is used as shown below.

```
# Create data frame from diamonds dataset
a = diamonds

# Require library "ggplot2"
ggplot(a, aes(x = price, y = carat)) +
  geom_density2d()
```

Figure 138: Density code

Figure above shows that price will be the x-axis and carat be the y-axis, they are both continuous data, so density2d graph will be useful to display which combination having the most frequency in the dataset as shown below.

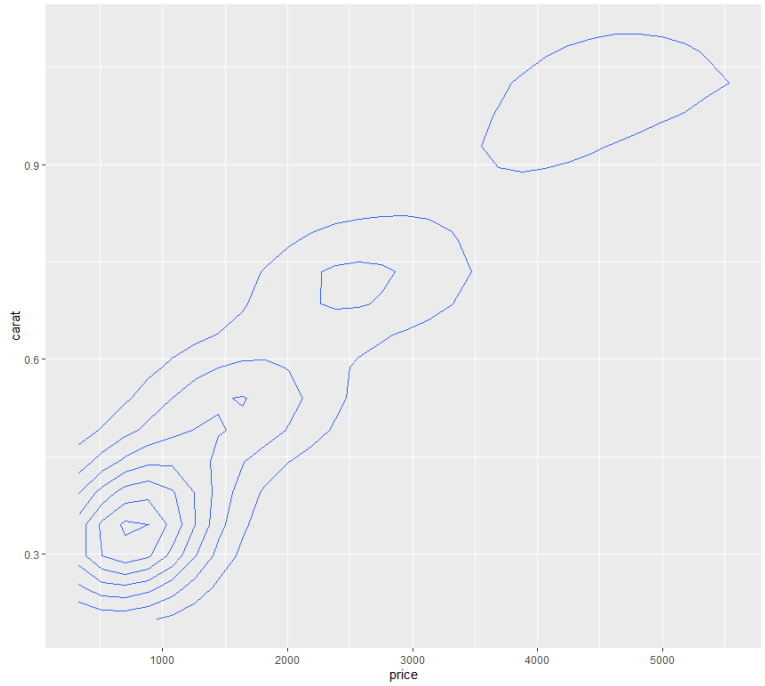

Figure 139: Density graph

4. Heat map – geom_tile()

Heat map is very useful when random pattern exists in data for both axis in a graph, while a third attribute is also involved in this graph. Hence a heat map will be the most suitable one for this situation and will be focus on interpret some pattern or special part on the map. For example, combination of free time and travel time in this analysis need to put in the same graph to study the relation between performance, then heat map is used. To demonstrate, diamonds datasets is used as shown below.

```
# Create data frame from diamonds dataset
a = diamonds

# Require library "ggplot2"
ggplot(a, aes(x = cut, y = clarity)) +
  geom_tile(aes(fill = depth))
```

Figure 140: Heat map code

Figure above shows that cut will be the x-axis and clarity be the y-axis, they are both having random pattern, so heat map graph will be useful to display which combination having the best depth in the dataset as shown below.

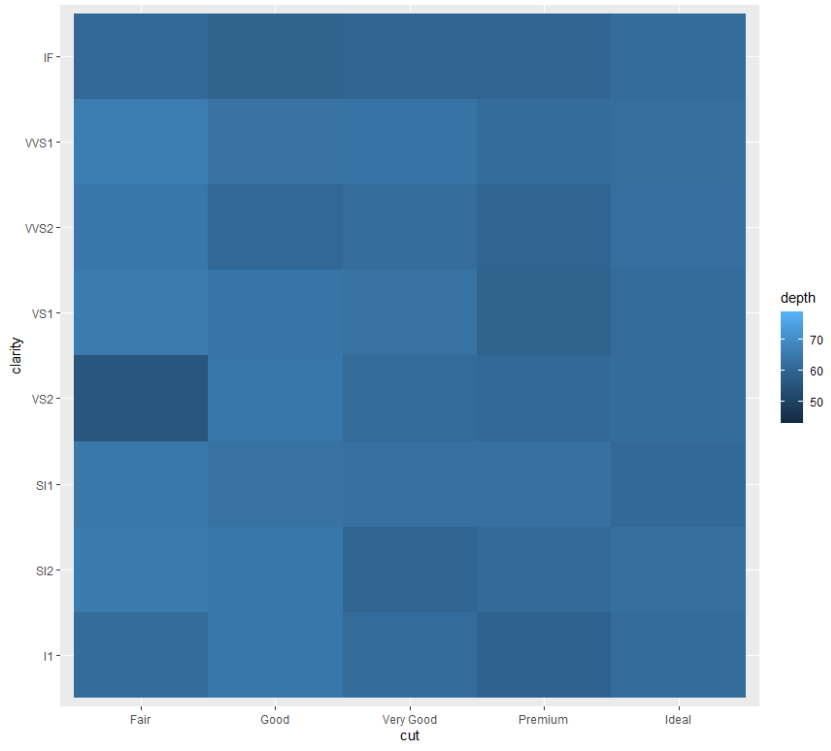

Figure 141: Heat map graph

5. Display categories in grid– facet_grid()

This is different from plot_grid() as well as facet_wrap() as it separate graph automatically follow with how many levels exists in the column not like plot_grid() and it can shows the graph in 2D grid which can take 2 column for the levels not like facet_wrap() which only accept 1 column and create a 1D grid. It is very useful when 2 or more inter-related attributes exists in a graph for deeper analysis. For example, Family_Size, Cohabitation_Status are inter-related in this analysis, hence facet_grid() function is used. To demonstrate, diamonds datasets is used as shown below.

```
# Create data frame from diamonds dataset
a = diamonds

# Require library "ggplot2"
ggplot(a, aes(x = carat, y = price)) +
  geom_point() +
  facet_grid(rows = vars(cut), cols = vars(clarity))
```

Figure 142: facet_grid code

Figure above shows that carat will be the x-axis and price be the y-axis, cut and clarity levels will be the one which separate graph into 2D grid as shown below. Although the grid will be messy in this example but if with less levels the grid will have a better illustration.

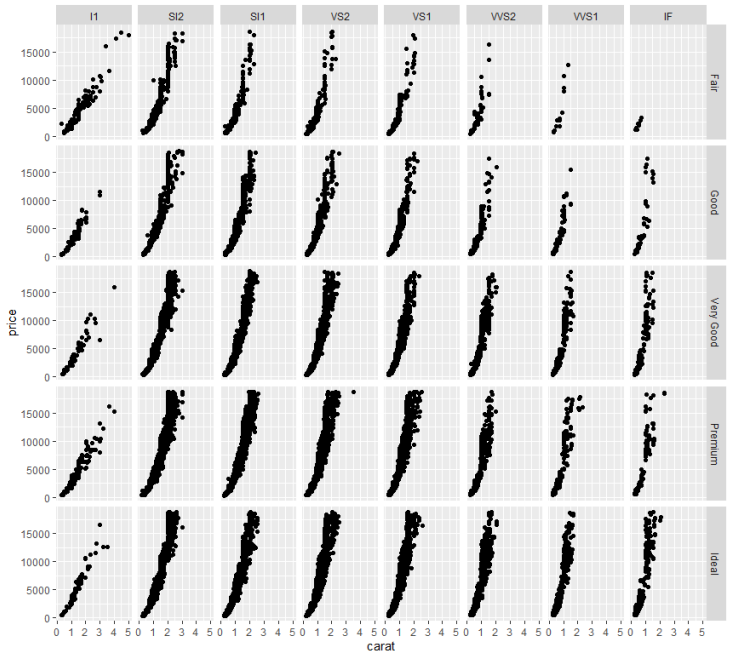

Figure 143: facet_grid graph

Conclusion

In a nutshell, there are many meaningful and effective in providing the information for the decision making discovered in the whole process of analysis. Some of the attributes are directly related to performance while some of them need to inter-related to each other to display out the relation. Hence, various type of techniques is widely used in data exploration, manipulation, transformation, and visualization as well as some addition features. I noticed that data analysis can go in so deep depend on the experience in this field, with a good depth of analysis some insight or most wanted wisdom can be discovered for a good usage in the future.

References

Abu Bakar, N., Mamat, I., & Ibrahim, M. (2017). Influence of Parental Education on Academic Performance of Secondary School Students in Kuala Terengganu. *International Journal of Academic Research in Business and Social Sciences*, Vol. 7, No. 8.

betterhealth. (2014, August 21st). *Better Health Channel.* Retrieved from Self esteem: https://www.betterhealth.vic.gov.au/health/healthyliving/self-esteem

Bright Horizions. (2020, october 15th). *Bright Horizions.* Retrieved from 8 WAYS CHILDREN CAN BENEFIT FROM A NURSERY ENVIRONMENT: https://www.brighthorizons.co.uk/family-resources/blog/2020/10/benefits-of-nursery-environment

CDC. (2019, April 10th). *Centers for Disease Control and Prevention.* Retrieved from Mental Health in the Workplace: https://www.cdc.gov/workplacehealthpromotion/tools-resources/workplace-health/mental-health/index.html

Debbie, Y., & Sara, J. (2017, February 28th). *Why You Should Become a UseR: A Brief Introduction to R.* Retrieved July 15th, 2021, from https://www.psychologicalscience.org/observer/why-you-should-become-a-user-a-brief-introduction-to-r

Kashif, R., Abdul Qayyum, C., & Muhammad, A. (2018). Relationship between Co-curricular Activities and Exam Performance: Mediating Role of Attendance. *Bulletin of Education and Research,* Vol. 40, No. 1 pp. 183-196.

Konuk, N., Turan, N., & Ardali, Y. (2016). THE IMPORTANCE OF URBANIZATION IN EDUCATION. *The Eurasia Proceedings of Educational & Social Sciences (EPESS),* Volume 5, Pages 232-236.

Michael, A. (1999). National Center for Biotechnology Information. *Does Drinking Reduce Stress?*, 23(4): 250–255. Retrieved from Does Drinking Reduce Stress?

NICOLAS, C. (2016, June 20th). *Inc.* Retrieved from Failure Doesn't Actually Exist. Here's WhyIf you believe in failure, you're missing the point.: https://www.inc.com/nicolas-cole/failure-doesnt-actually-exist-heres-why.html

PROFESSOR ANDREW, M., & KATHARINE, S. (2017, May 22nd). *Australian Psychological Society.* Retrieved from How puberty affects school performance: https://psychlopaedia.org/learning-and-development/puberty-affects-school-performance/

Steven , G., Edward, H., & Janet, E. (2014, October). *THE IMPACT OF THE INTERNET ON URBAN VITALITY: DOES CLOSENESS IN CYBER-SPACE SUBSTITUTE FOR URBAN SPACE?* Retrieved from University of Houston: https://www.uh.edu/~kohlhase/CraigHoangKohlhase_internet_WP_Oct_2014.pdf

Tim, C. (2017, January 23rd). *Challies.* Retrieved from Greater Age Brings Greater Responsibility: https://www.challies.com/articles/greater-age-brings-greater-responsibility/

Vista College. (2021, August 22nd). *Vista College.* Retrieved from The Importance of Higher Education in the 21st Century: https://www.vistacollege.edu/blog/resources/higher-education-in-the-21st-century/