PFDA Assignment Sample 2

Programming For Data Analysis (Asia Pacific University of Technology and Innovation)

GROUP ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CT127-3-2-PFDA

PROGRAMMING FOR DATA ANALYSIS

APD2F2309IT(MBT/CE/ISS)

(Group 4)

Group Members:	Abdulrahman Gamil Murshed Humadi (TP070609)
	Furas Sultan Ghaleb Mohammed (TP066989)
	Lee Wen Jing (TP071630)
	Lee Xin Yee (TP070654)
Hand out date:	04 October 2023
Hand in date:	08 December 2023
Lecturer:	Mary Ting

Table of Contents

1.0 Introduction

1.1 Data Description

In this report, our group is going to conduct research from the dataset "student_prediction.csv" given. We aimed to make data analysis on the dataset with the help of RStudio as our supporting tool. Below are the data description from the original dataset "student_prediction.csv":

Column Name	Attributes	Data Description
STUDENTID	Student ID	
AGE	Student Age	1: 18-21, 2: 22-25, 3: above 26
GENDER	Sex	1: female, 2: male
HS_TYPE	Graduated high-school type	1: private, 2: state, 3: other
SCHOLARSHIP	Scholarship type	1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full
WORK	Additional work	1: Yes, 2: No
ACTIVITY	Regular artistic or sports activity	1: Yes, 2: No
PARTNER	Do you have a partner	1: Yes, 2: No
SALARY	Total salary if available	1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410
TRANSPORT	Transportation to the university	1: Bus, 2: Private car/taxi, 3: bicycle, 4: Other
LIVING	Accommodation type in Cyprus	1: rental, 2: dormitory, 3: with family, 4: Other
MOTHER_EDU	Mother's education	1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.
FATHER_EDU	Father's education	1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.
X._SIBLINGS	Number of sisters/brothers (if available)	1: 1, 2: 2, 3: 3, 4: 4, 5: 5 or above
KIDS *(wrongly listed by owner)*	Parental status	1: married, 2: divorced, 3: died - one of them or both
MOTHER_JOB	Mother's occupation	1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other

FATHER_JOB	Father's occupation	1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other
STUDY_HRS	Weekly study hours	1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours
READ_FREQ	Reading frequency (non-scientific books/journals)	1: None, 2: Sometimes, 3: Often
READ_FREQ_SCI	Reading frequency (scientific books/journals)	1: None, 2: Sometimes, 3: Often
ATTEND_DEPT	Attendance to the seminars/conferences related to the department	1: Yes, 2: No
IMPACT	Impact of your projects/activities on your success	1: positive, 2: negative, 3: neutral
ATTEND	Attendance to classes	1: always, 2: sometimes, 3: never
PREP_STUDY	Preparation to midterm exams 1	1: alone, 2: with friends, 3: not applicable
PREP_EXAM	Preparation to midterm exams 2	1: closest date to the exam, 2: regularly during the semester, 3: never
NOTES	Taking notes in classes	1: never, 2: sometimes, 3: always
LISTENS	Listening in classes	1: never, 2: sometimes, 3: always
LIKES_DISCUSS	Discussion improves my interest and success in the course	1: never, 2: sometimes, 3: always
CLASSROOM	Flip-classroom	1: not useful, 2: useful, 3: not applicable
CUML_GPA	Cumulative grade point average in the last semester (/4.00)	1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49
EXP_GPA	Expected Cumulative grade point average in the graduation (/4.00)	1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49
COURSE.ID	Course ID	
GRADE	OUTPUT Grade	0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA

1.2 Hypothesis

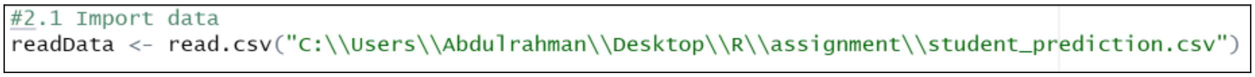
Students’ academic performance is affected by their engagement in classes, how much time they allocated for revision, and their family background and personal / economic circumstances.

1.3 Objectives

1. To examine the level of student engagement in academic sessions.
2. To investigate how much time students are allocating time for revision and preparation for exams and how this affects the CGPAs of the students.
3. To examine the relationship between students' family background and academic performance.
4. To investigate the impact of Students’ Personal Life and Economic Circumstances on Academic Performance.

2.0 Data Preparation

2.1 Data Import

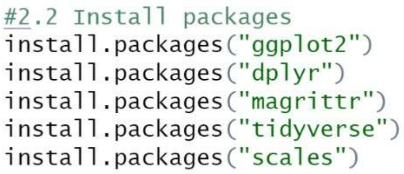


```
#2.1 Import data
readData <- read.csv("C:\\Users\\Abdulrahman\\Desktop\\R\\assignment\\student_prediction.csv")
```

Figure 1 Import data

Since the “student_prediction.csv” file is in csv file format, we use read.csv function to import the csv file dataset to the RStudio. In read.csv function, we entered the file path where the file is stored.

2.2 Install Packages

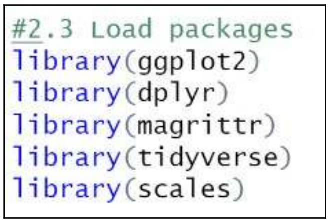


```
#2.2 Install packages
install.packages("ggplot2")
install.packages("dplyr")
install.packages("magrittr")
install.packages("tidyverse")
install.packages("scales")
```

Figure 2 Install packages

Packages need to be installed before we can access it. To install packages in RStudio, we can use the install.packages() with the name of the desired package inside the parenthesis.

2.3 Load Packages

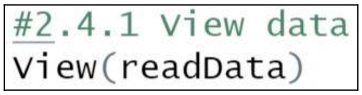


```
#2.3 Load packages
library(ggplot2)
library(dplyr)
library(magrittr)
library(tidyverse)
library(scales)
```

Figure 3 Load packages

We use library() -with the name of the desired package inside the parenthesis- to load the downloaded packages so we can use them immediately.

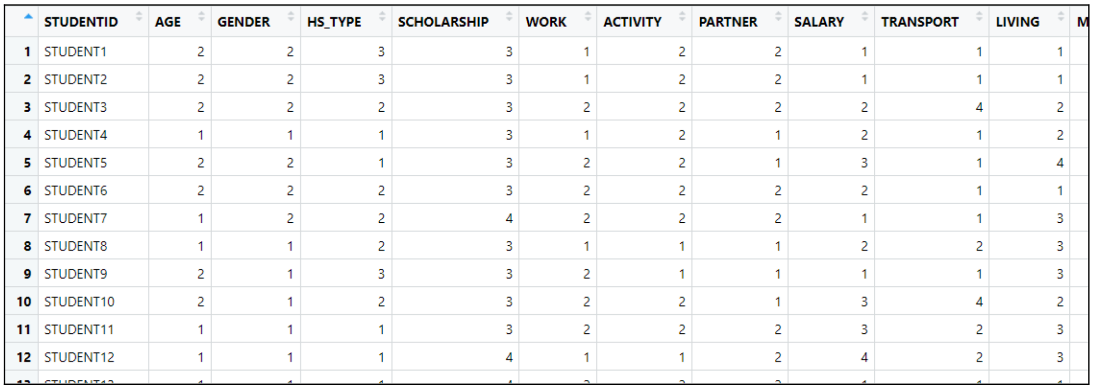
2.4 Data Exploration



```
#2.4.1 View data
View(readData)
```

Figure 4 View data

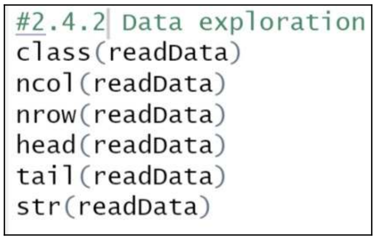
Firstly, we use View() function to view the dataset in RStudio. The output is as follow:



	STUDENTID	AGE	GENDER	HS_TYPE	SCHOLARSHIP	WORK	ACTIVITY	PARTNER	SALARY	TRANSPORT	LIVING	M
1	STUDENT1	2	2	3	3	1	2	2	1	1	1	
2	STUDENT2	2	2	3	3	1	2	2	1	1	1	
3	STUDENT3	2	2	2	3	2	2	2	2	4	2	
4	STUDENT4	1	1	1	3	1	2	1	2	1	2	
5	STUDENT5	2	2	1	3	2	2	1	3	1	4	
6	STUDENT6	2	2	2	3	2	2	2	2	1	1	
7	STUDENT7	1	2	2	4	2	2	2	1	1	3	
8	STUDENT8	1	1	2	3	1	1	1	2	2	3	
9	STUDENT9	2	1	3	3	2	1	1	1	1	3	
10	STUDENT10	2	1	2	3	2	2	1	3	4	2	
11	STUDENT11	1	1	1	3	2	2	2	3	2	3	
12	STUDENT12	1	1	1	4	1	1	2	4	2	3	

Figure 5 Output of View()

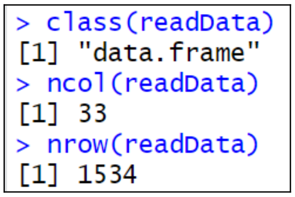
It shows a structured table which allows us to read the data easier.



```
#2.4.2 Data exploration
class(readData)
ncol(readData)
nrow(readData)
head(readData)
tail(readData)
str(readData)
```

Figure 6 More data exploration

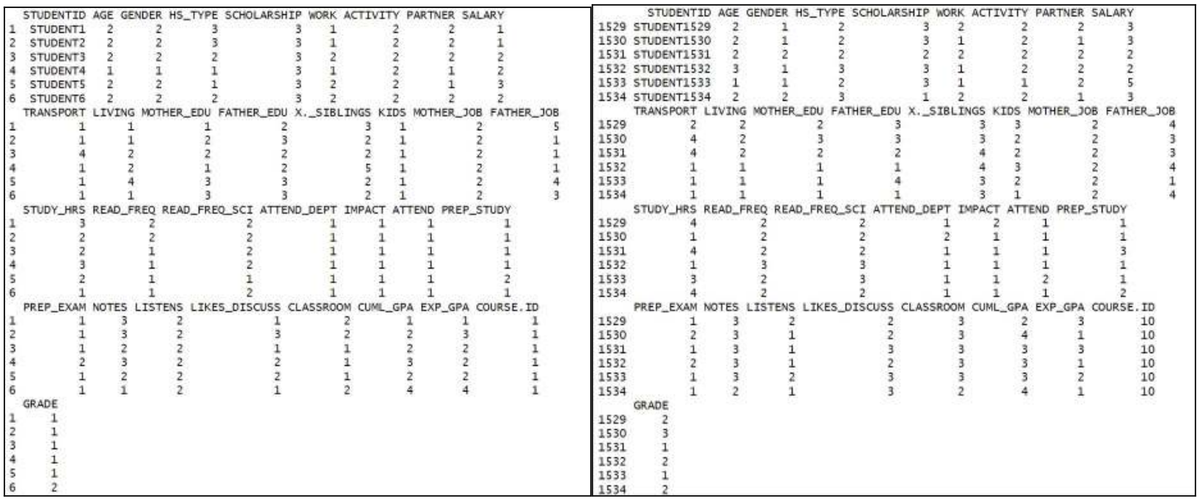
Additionally, other data exploration such as the figure above provides more different information about the dataset, allows us to have a better understanding of the dataset.



```
> class(readData)
[1] "data.frame"
> ncol(readData)
[1] 33
> nrow(readData)
[1] 1534
```

Figure 7 Output of class(), ncol(), and nrow()

We learned that the data type of the dataset is a data frame. It has 33 columns and 1534 rows, which means there are a total of 33 attributes and 1534 records.

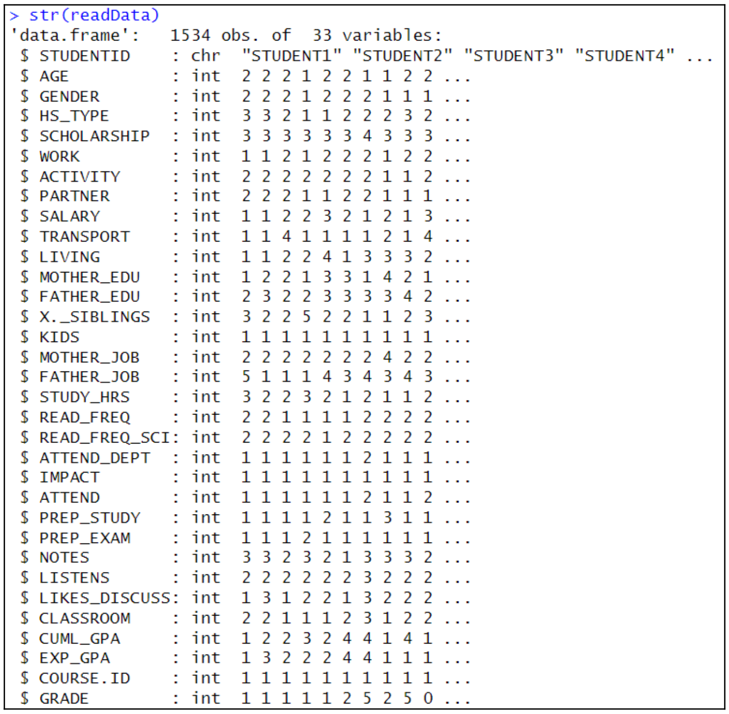


```
  STUDENTID AGE GENDER HS_TYPE SCHOLARSHIP WORK ACTIVITY PARTNER SALARY
1  STUDENT1   2      2       3           3    1        2       2      1
2  STUDENT2   2      2       3           3    1        2       2      1
3  STUDENT3   2      2       2           3    2        2       2      2
4  STUDENT4   1      1       1           3    1        2       1      2
5  STUDENT5   2      2       1           3    2        2       1      3
6  STUDENT6   2      2       2           3    2        2       2      2
  TRANSPORT LIVING MOTHER_EDU FATHER_EDU X._SIBLINGS KIDS MOTHER_JOB FATHER_JOB
1         1      1          1          2           3    1          2          5
2         1      1          2          3           2    1          2          1
3         4      2          2          2           2    1          2          1
4         1      2          1          2           5    1          2          1
5         1      4          3          3           2    1          2          4
6         1      1          3          3           2    1          2          3
  STUDY_HRS READ_FREQ READ_FREQ_SCI ATTEND_DEPT IMPACT ATTEND PREP_STUDY
1         3         2             2           1      1      1          1
2         2         2             2           1      1      1          1
3         2         1             2           1      1      1          1
4         3         1             2           1      1      1          1
5         2         1             1           1      1      1          2
6         1         1             2           1      1      1          1
  PREP_EXAM NOTES LISTENS LIKES_DISCUSS CLASSROOM CUML_GPA EXP_GPA COURSE.ID
1         1     3       2             1         2        1       1         1
2         1     3       2             3         2        2       3         1
3         1     2       2             1         1        2       2         1
4         2     3       2             2         1        3       2         1
5         1     2       2             2         1        2       2         1
6         1     1       2             1         2        4       4         1
  GRADE
1     1
2     1
3     1
4     1
5     1
6     2
```

```
      STUDENTID AGE GENDER HS_TYPE SCHOLARSHIP WORK ACTIVITY PARTNER SALARY
1529 STUDENT1529   2      1       2           3    2        2       2      3
1530 STUDENT1530   2      1       2           3    1        2       1      3
1531 STUDENT1531   2      2       2           2    2        2       2      2
1532 STUDENT1532   3      1       3           3    1        2       2      2
1533 STUDENT1533   1      1       2           3    1        1       2      5
1534 STUDENT1534   2      2       3           1    2        2       1      3
     TRANSPORT LIVING MOTHER_EDU FATHER_EDU X._SIBLINGS KIDS MOTHER_JOB FATHER_JOB
1529         2      2          2          3           3    3          2          4
1530         4      2          3          3           3    2          2          3
1531         4      2          2          2           4    2          2          3
1532         1      1          1          1           4    3          2          4
1533         1      1          1          4           3    2          2          1
1534         1      1          1          1           3    1          2          4
     STUDY_HRS READ_FREQ READ_FREQ_SCI ATTEND_DEPT IMPACT ATTEND PREP_STUDY
1529         4         2             2           1      2      1          1
1530         1         2             2           2      1      1          1
1531         4         2             2           1      1      1          3
1532         1         3             3           1      1      1          1
1533         3         2             3           1      1      2          1
1534         4         2             2           1      1      1          2
     PREP_EXAM NOTES LISTENS LIKES_DISCUSS CLASSROOM CUML_GPA EXP_GPA COURSE.ID
1529         1     3       2             2         3        2       3        10
1530         2     3       1             2         3        4       1        10
1531         1     3       1             3         3        3       3        10
1532         2     3       1             2         3        3       1        10
1533         1     3       2             3         3        3       2        10
1534         1     2       1             3         2        4       1        10
     GRADE
1529     2
1530     3
1531     1
1532     2
1533     1
1534     2
```

Figure 8 Output of head and tail

head() and tail() function show us the first and last 10 records of the dataset.



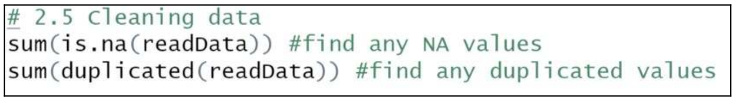
```
> str(readData)
'data.frame':	1534 obs. of  33 variables:
 $ STUDENTID    : chr  "STUDENT1" "STUDENT2" "STUDENT3" "STUDENT4" ...
 $ AGE          : int  2 2 2 1 2 2 1 1 2 2 ...
 $ GENDER       : int  2 2 2 1 2 2 2 1 1 1 ...
 $ HS_TYPE      : int  3 3 2 1 1 2 2 2 3 2 ...
 $ SCHOLARSHIP  : int  3 3 3 3 3 3 4 3 3 3 ...
 $ WORK         : int  1 1 2 1 2 2 2 1 2 2 ...
 $ ACTIVITY     : int  2 2 2 2 2 2 2 1 1 2 ...
 $ PARTNER      : int  2 2 2 1 1 2 2 1 1 1 ...
 $ SALARY       : int  1 1 2 2 3 2 1 2 1 3 ...
 $ TRANSPORT    : int  1 1 4 1 1 1 1 2 1 4 ...
 $ LIVING       : int  1 1 2 2 4 1 3 3 3 2 ...
 $ MOTHER_EDU   : int  1 2 2 1 3 3 1 4 2 1 ...
 $ FATHER_EDU   : int  2 3 2 2 3 3 3 3 4 2 ...
 $ X._SIBLINGS  : int  3 2 2 5 2 2 1 1 2 3 ...
 $ KIDS         : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MOTHER_JOB   : int  2 2 2 2 2 2 2 4 2 2 ...
 $ FATHER_JOB   : int  5 1 1 1 4 3 4 3 4 3 ...
 $ STUDY_HRS    : int  3 2 2 3 2 1 2 1 1 2 ...
 $ READ_FREQ    : int  2 2 1 1 1 1 2 2 2 2 ...
 $ READ_FREQ_SCI: int  2 2 2 2 1 2 2 2 2 2 ...
 $ ATTEND_DEPT  : int  1 1 1 1 1 1 2 1 1 1 ...
 $ IMPACT       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ ATTEND       : int  1 1 1 1 1 1 2 1 1 2 ...
 $ PREP_STUDY   : int  1 1 1 1 2 1 1 3 1 1 ...
 $ PREP_EXAM    : int  1 1 1 2 1 1 1 1 1 1 ...
 $ NOTES        : int  3 3 2 3 2 1 3 3 3 2 ...
 $ LISTENS      : int  2 2 2 2 2 2 3 2 2 2 ...
 $ LIKES_DISCUSS: int  1 3 1 2 2 1 3 2 2 2 ...
 $ CLASSROOM    : int  2 2 1 1 1 2 3 1 2 2 ...
 $ CUML_GPA     : int  1 2 2 3 2 4 4 1 4 1 ...
 $ EXP_GPA      : int  1 3 2 2 2 4 4 1 1 1 ...
 $ COURSE.ID    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ GRADE        : int  1 1 1 1 1 2 5 2 5 0 ...
```

Figure 9 Output of str()

str() function gives the information about the class of each column, as well as the name of the column. Except for STUDENTID which contains character data type, the rest of the attributes are in integer data type.

2.5 Data Cleaning

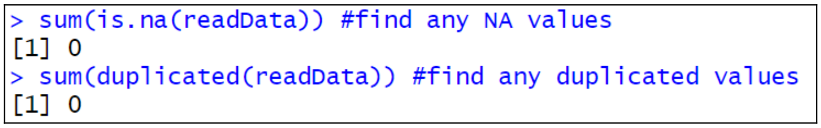
Before analyzing the data, we should first clean the data. Data cleaning is important because it avoids issues like data shortage and formatting discrepancies, emphasizing the importance of real-life data quality control in the industry (Amadebai, 2021).



```
# 2.5 Cleaning data
sum(is.na(readData)) #find any NA values
sum(duplicated(readData)) #find any duplicated values
```

Figure 10 Data cleaning

We did a basic verification with the code above. We use is.na() to check if there is any missing values, while duplicated() is used to find any duplicated values. In addition, we added sum() to the codes to find the amount of missing values and duplicate values.

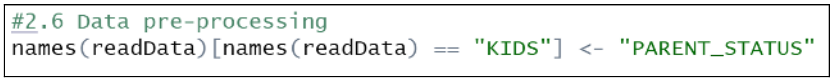


```
> sum(is.na(readData)) #find any NA values
[1] 0
> sum(duplicated(readData)) #find any duplicated values
[1] 0
```

Figure 11 Output of the sum of missing values and duplicate values

The result shows that there are no missing values or duplicated values in the dataset. Hence, data cleaning is not needed for the dataset.

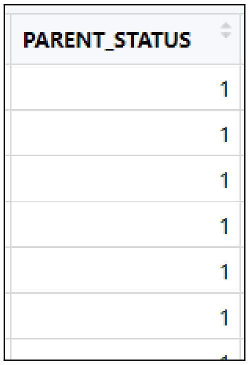
2.6 Data Pre-processing



```
#2.6 Data pre-processing
names(readData)[names(readData) == "KIDS"] <- "PARENT_STATUS"
```

Figure 12 Change column name

We modify the column name from “KIDS” to “PARENT_STATUS” to correspond to “Parental status” attribute. The modified table will look like below figure:

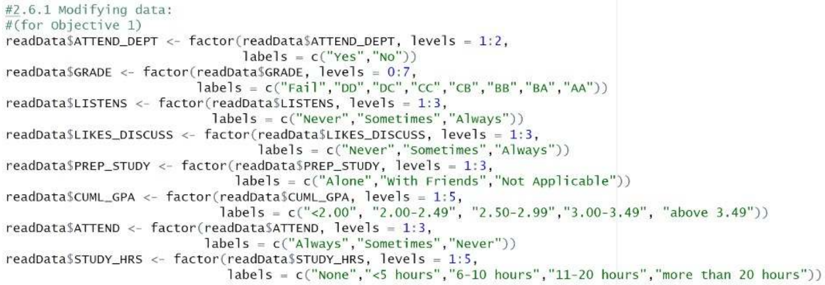


PARENT_STATUS
1
1
1
1
1
1

Figure 13 Output of modified column

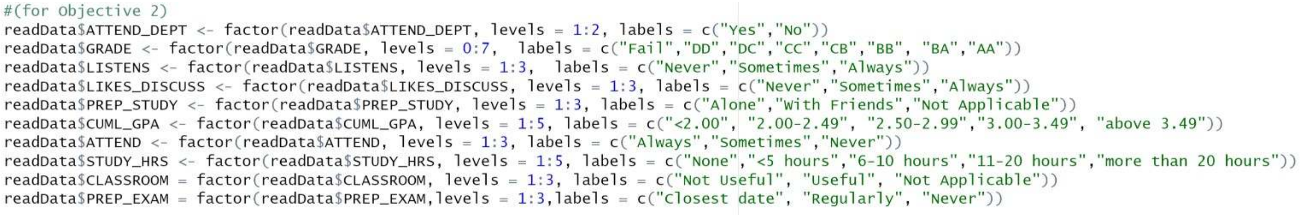
Before analyzing the data, related data should change into words so that we can understand what the data is about. Thus, data is modified from integer to readable words.

2.6.1 Modifying data



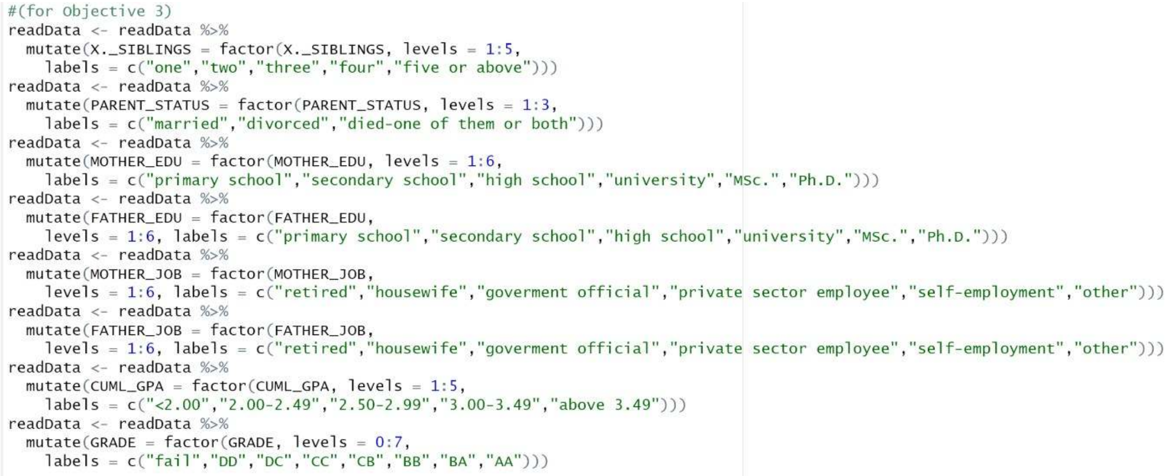
```
#2.6.1 Modifying data:
#(for Objective 1)
readData$ATTEND_DEPT <- factor(readData$ATTEND_DEPT, levels = 1:2,
                               labels = c("Yes","No"))
readData$GRADE <- factor(readData$GRADE, levels = 0:7,
                         labels = c("Fail","DD","DC","CC","CB","BB","BA","AA"))
readData$LISTENS <- factor(readData$LISTENS, levels = 1:3,
                           labels = c("Never","Sometimes","Always"))
readData$LIKES_DISCUSS <- factor(readData$LIKES_DISCUSS, levels = 1:3,
                                 labels = c("Never","Sometimes","Always"))
readData$PREP_STUDY <- factor(readData$PREP_STUDY, levels = 1:3,
                              labels = c("Alone","With Friends","Not Applicable"))
readData$CUML_GPA <- factor(readData$CUML_GPA, levels = 1:5,
                            labels = c("<2.00", "2.00-2.49", "2.50-2.99","3.00-3.49", "above 3.49"))
readData$ATTEND <- factor(readData$ATTEND, levels = 1:3,
                          labels = c("Always","Sometimes","Never"))
readData$STUDY_HRS <- factor(readData$STUDY_HRS, levels = 1:5,
                             labels = c("None","<5 hours","6-10 hours","11-20 hours","more than 20 hours"))
```

Figure 14 modify O1



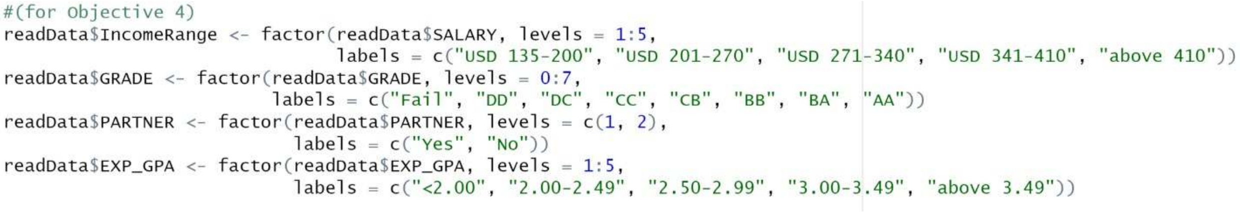
```
#(for Objective 2)
readData$ATTEND_DEPT <- factor(readData$ATTEND_DEPT, levels = 1:2, labels = c("Yes","No"))
readData$GRADE <- factor(readData$GRADE, levels = 0:7,  labels = c("Fail","DD","DC","CC","CB","BB", "BA","AA"))
readData$LISTENS <- factor(readData$LISTENS, levels = 1:3,  labels = c("Never","Sometimes","Always"))
readData$LIKES_DISCUSS <- factor(readData$LIKES_DISCUSS, levels = 1:3, labels = c("Never","Sometimes","Always"))
readData$PREP_STUDY <- factor(readData$PREP_STUDY, levels = 1:3, labels = c("Alone","With Friends","Not Applicable"))
readData$CUML_GPA <- factor(readData$CUML_GPA, levels = 1:5, labels = c("<2.00", "2.00-2.49", "2.50-2.99","3.00-3.49", "above 3.49"))
readData$ATTEND <- factor(readData$ATTEND, levels = 1:3, labels = c("Always","Sometimes","Never"))
readData$STUDY_HRS <- factor(readData$STUDY_HRS, levels = 1:5, labels = c("None","<5 hours","6-10 hours","11-20 hours","more than 20 hours"))
readData$CLASSROOM = factor(readData$CLASSROOM, levels = 1:3, labels = c("Not Useful", "Useful", "Not Applicable"))
readData$PREP_EXAM = factor(readData$PREP_EXAM,levels = 1:3,labels = c("Closest date", "Regularly", "Never"))
```

Figure 15 modify O2



```
#(for Objective 3)
readData <- readData %>%
  mutate(X._SIBLINGS = factor(X._SIBLINGS, levels = 1:5,
    labels = c("one","two","three","four","five or above")))
readData <- readData %>%
  mutate(PARENT_STATUS = factor(PARENT_STATUS, levels = 1:3,
    labels = c("married","divorced","died-one of them or both")))
readData <- readData %>%
  mutate(MOTHER_EDU = factor(MOTHER_EDU, levels = 1:6,
    labels = c("primary school","secondary school","high school","university","MSc.","Ph.D.")))
readData <- readData %>%
  mutate(FATHER_EDU = factor(FATHER_EDU,
    levels = 1:6, labels = c("primary school","secondary school","high school","university","MSc.","Ph.D.")))
readData <- readData %>%
  mutate(MOTHER_JOB = factor(MOTHER_JOB,
    levels = 1:6, labels = c("retired","housewife","goverment official","private sector employee","self-employment","other")))
readData <- readData %>%
  mutate(FATHER_JOB = factor(FATHER_JOB,
    levels = 1:6, labels = c("retired","housewife","goverment official","private sector employee","self-employment","other")))
readData <- readData %>%
  mutate(CUML_GPA = factor(CUML_GPA, levels = 1:5,
    labels = c("<2.00","2.00-2.49","2.50-2.99","3.00-3.49","above 3.49")))
readData <- readData %>%
  mutate(GRADE = factor(GRADE, levels = 0:7,
    labels = c("fail","DD","DC","CC","CB","BB","BA","AA")))
```

Figure 16 modify O3



```
#(for Objective 4)
readData$IncomeRange <- factor(readData$SALARY, levels = 1:5,
                               labels = c("USD 135-200", "USD 201-270", "USD 271-340", "USD 341-410", "above 410"))
readData$GRADE <- factor(readData$GRADE, levels = 0:7,
                         labels = c("Fail", "DD", "DC", "CC", "CB", "BB", "BA", "AA"))
readData$PARTNER <- factor(readData$PARTNER, levels = c(1, 2),
                           labels = c("Yes", "No"))
readData$EXP_GPA <- factor(readData$EXP_GPA, levels = 1:5,
                           labels = c("<2.00", "2.00-2.49", "2.50-2.99", "3.00-3.49", "above 3.49"))
```

Figure 17 modify O4

3.0 Data Analysis

3.1 Objective 1 (Lee Xin Yee TP070654)

The first objective is **to examine the level of student involvement in academic sessions**. The research questions are as follow:

1. Does attendance at seminars of department affect academic performance?

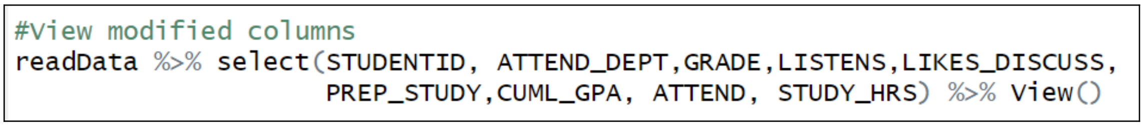
2. Is there a correlation between attendance at seminars of department and paying attention in classes? Do they affect students' academic performance?

3. Is there a correlation between discussion participation and grades?

The attributes that are used to study on student engagement included:

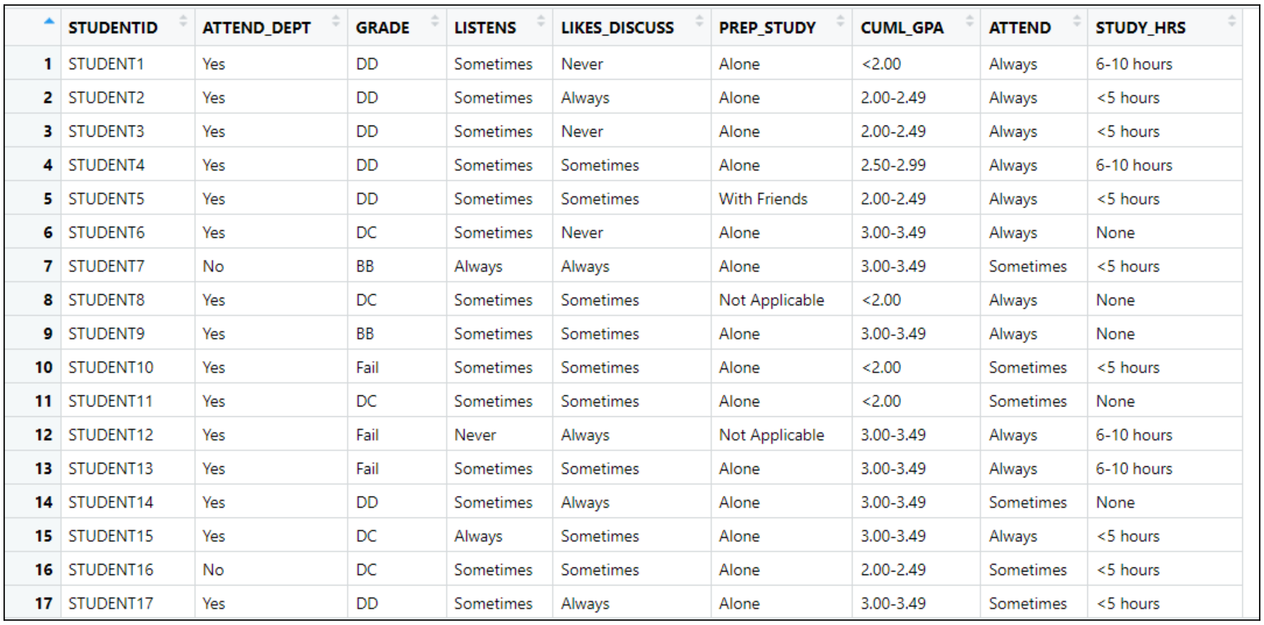
- Attendance to the seminars/conferences related to the department (ATTEND_DEPT)
- Output grade (GRADE)
- Listening in classes (LISTENS)
- Discussion in the course (LIKES_DISCUSS)
- Preparation to midterm exams 1 (PREP_STUDY)
- Cumulative GPA (CUML_GPA)
- Attendance to classes (ATTEND)
- Weekly study hours (STUDY_HRS)

To view the used attributes in this objective use View() to check the column.



```
#View modified columns
readData %>% select(STUDENTID, ATTEND_DEPT,GRADE,LISTENS,LIKES_DISCUSS,
                    PREP_STUDY,CUML_GPA, ATTEND, STUDY_HRS) %>% View()
```

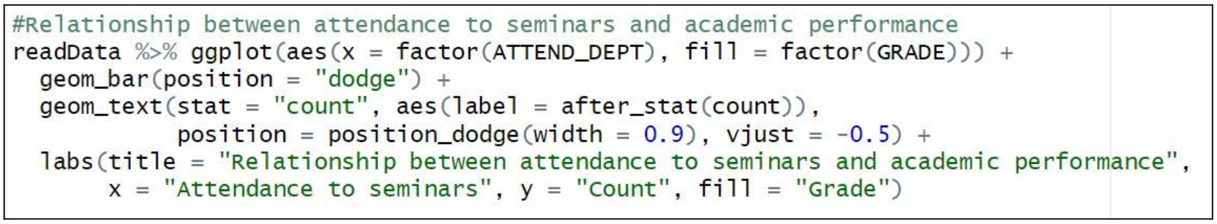
Figure 18 View modified columns



	STUDENTID	ATTEND_DEPT	GRADE	LISTENS	LIKES_DISCUSS	PREP_STUDY	CUML_GPA	ATTEND	STUDY_HRS
1	STUDENT1	Yes	DD	Sometimes	Never	Alone	<2.00	Always	6-10 hours
2	STUDENT2	Yes	DD	Sometimes	Always	Alone	2.00-2.49	Always	<5 hours
3	STUDENT3	Yes	DD	Sometimes	Never	Alone	2.00-2.49	Always	<5 hours
4	STUDENT4	Yes	DD	Sometimes	Sometimes	Alone	2.50-2.99	Always	6-10 hours
5	STUDENT5	Yes	DD	Sometimes	Sometimes	With Friends	2.00-2.49	Always	<5 hours
6	STUDENT6	Yes	DC	Sometimes	Never	Alone	3.00-3.49	Always	None
7	STUDENT7	No	BB	Always	Always	Alone	3.00-3.49	Sometimes	<5 hours
8	STUDENT8	Yes	DC	Sometimes	Sometimes	Not Applicable	<2.00	Always	None
9	STUDENT9	Yes	BB	Sometimes	Sometimes	Alone	3.00-3.49	Always	None
10	STUDENT10	Yes	Fail	Sometimes	Sometimes	Alone	<2.00	Sometimes	<5 hours
11	STUDENT11	Yes	DC	Sometimes	Sometimes	Alone	<2.00	Sometimes	None
12	STUDENT12	Yes	Fail	Never	Always	Not Applicable	3.00-3.49	Always	6-10 hours
13	STUDENT13	Yes	Fail	Sometimes	Sometimes	Alone	3.00-3.49	Always	6-10 hours
14	STUDENT14	Yes	DD	Sometimes	Always	Alone	3.00-3.49	Sometimes	None
15	STUDENT15	Yes	DC	Always	Sometimes	Alone	3.00-3.49	Always	<5 hours
16	STUDENT16	No	DC	Sometimes	Sometimes	Alone	2.00-2.49	Sometimes	<5 hours
17	STUDENT17	Yes	DD	Sometimes	Always	Alone	3.00-3.49	Sometimes	<5 hours

Figure 19 Output of modified columns

3.1.1 Does attendance at seminars of department affect academic performance?



```
#Relationship between attendance to seminars and academic performance
readData %>% ggplot(aes(x = factor(ATTEND_DEPT), fill = factor(GRADE))) +
  geom_bar(position = "dodge") +
  geom_text(stat = "count", aes(label = after_stat(count)),
            position = position_dodge(width = 0.9), vjust = -0.5) +
  labs(title = "Relationship between attendance to seminars and academic performance",
       x = "Attendance to seminars", y = "Count", fill = "Grade")
```

Figure 20 Relationship between attendance to seminars and academic performance

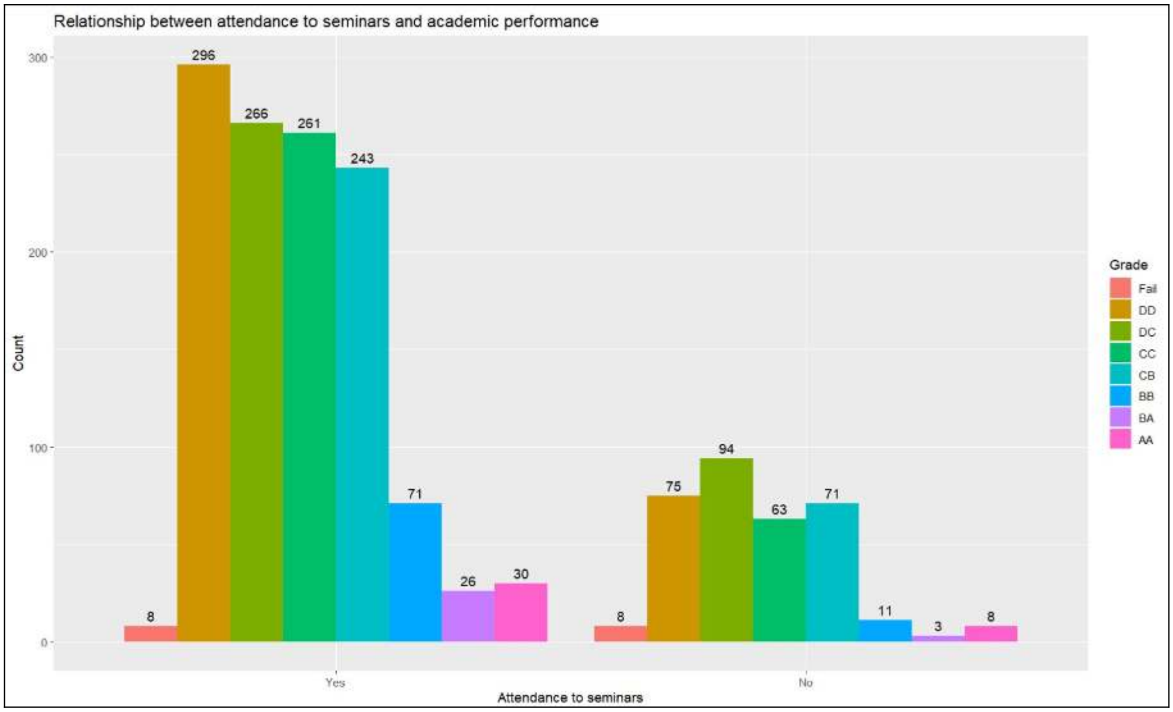
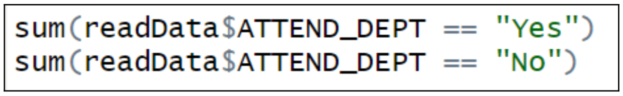


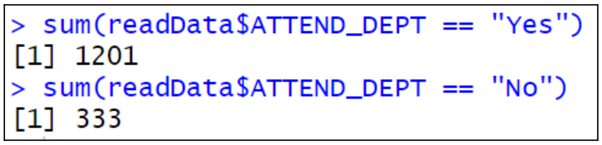
Figure 21 Bar plot of "Relationship between attendance to seminars and academic performance"

As from above, a bar plot is created in the analysis. It shows the frequencies of different grades for students who do and don't attend seminars of the department. Besides, geom_text is applied to know more precise count of every bar. There is a large difference in the number of "Yes" and "No" responses, and most of the students were present at the seminars. And from the "Yes" response, students' individual grades are also relatively satisfactory, such as the number of BB, BA, and AA grade is relatively high. To know more about the proportion of "Yes" and "No" responses, sum() is used to count the total number of students for respective responses.



```
sum(readData$ATTEND_DEPT == "Yes")
sum(readData$ATTEND_DEPT == "No")
```

Figure 22 Code to find the sum of both responses



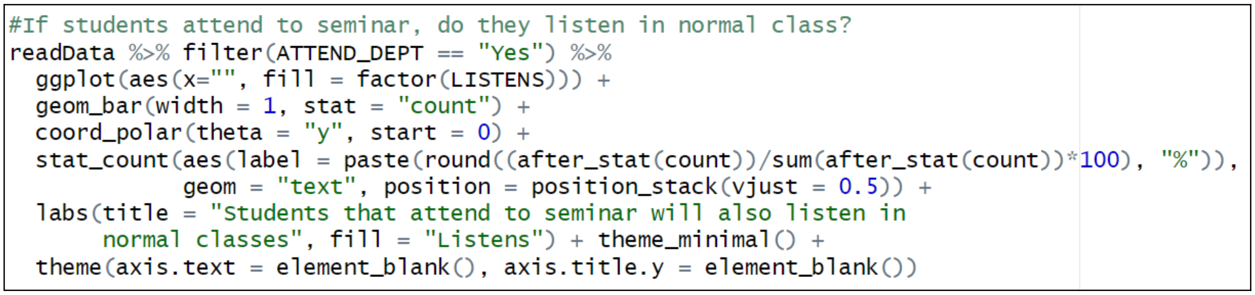
```
> sum(readData$ATTEND_DEPT == "Yes")
[1] 1201
> sum(readData$ATTEND_DEPT == "No")
[1] 333
```

Figure 23 Output of the code

The total number of students who attend the seminars is 1201 while those who don't attend are 333 students. As a result, although both responses have the same number of students who fail their grades, the proportions are different: one is 8 out of 1201 students and another is 8 out of 333 students (8/1201 is greater than 8/333). In addition, if we compare the proportion of the number of students who get AA grade, it also shows "Yes" response is slightly greater than "No" response (30/1201 is greater than 8/333). **In short, attendance at seminars will affect students' academic performance.** To support this result, further analysis in the next question is taken for higher accuracy.

3.1.2 Is there a correlation between attendance at seminars of department and paying attention in classes? Do they affect students’ academic performance?

Firstly, the analysis will start with studying **whether students who attend seminars will listen in their classes**.



```
#If students attend to seminar, do they listen in normal class?
readData %>% filter(ATTEND_DEPT == "Yes") %>%
  ggplot(aes(x="", fill = factor(LISTENS))) +
  geom_bar(width = 1, stat = "count") +
  coord_polar(theta = "y", start = 0) +
  stat_count(aes(label = paste(round((after_stat(count))/sum(after_stat(count))*100), "%")),
             geom = "text", position = position_stack(vjust = 0.5)) +
  labs(title = "Students that attend to seminar will also listen in
       normal classes", fill = "Listens") + theme_minimal() +
  theme(axis.text = element_blank(), axis.title.y = element_blank())
```

Figure 24 If students attend to seminar, do they listen in normal classes?

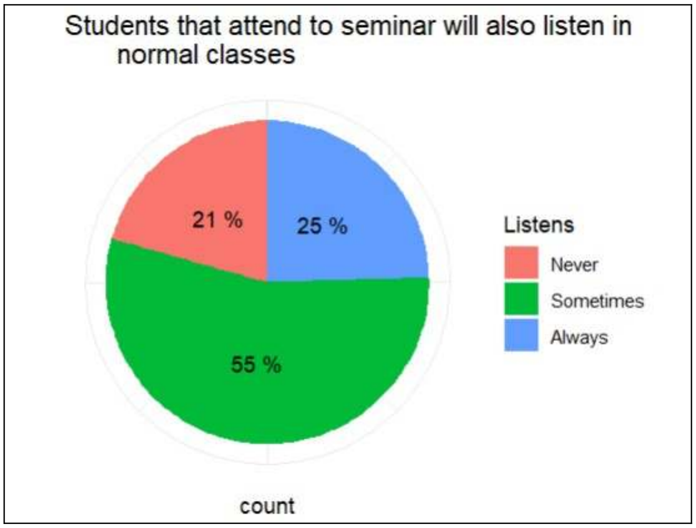
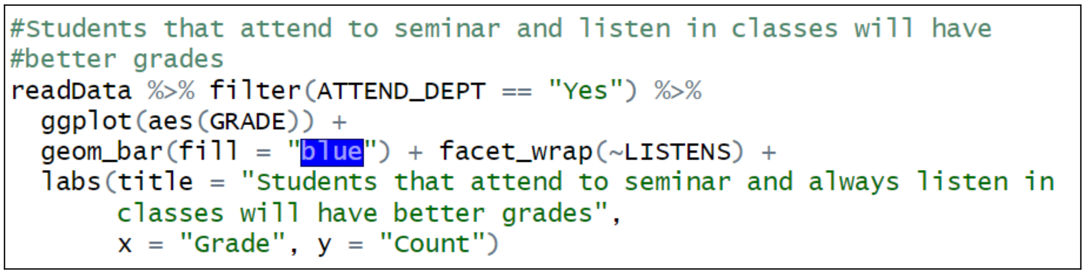


Figure 25 Pie Chart of “Students that attend to seminar will also listen in normal classes”

From the pie chart above, 25% of the students that choose to attend at related seminars will always listen during the classes, 55% which is more than half of the total number of students will sometimes listen in classes, and 21% of the students never listen in classes. Therefore, this shows that **since not all students will pay attention to the class, the students who attend the seminars may not necessarily pay attention.** So, this is one of the reasons why there is a significant difference in the grades of the individual students in attendance at seminars.

Secondly, the analysis will continue to find **the correlation of attendance to seminars, listening in classes, and output grade**. Filter() function is used to get the result of "Yes" responses only.



```
#Students that attend to seminar and listen in classes will have
#better grades
readData %>% filter(ATTEND_DEPT == "Yes") %>%
  ggplot(aes(GRADE)) +
  geom_bar(fill = "blue") + facet_wrap(~LISTENS) +
  labs(title = "Students that attend to seminar and always listen in
       classes will have better grades",
       x = "Grade", y = "Count")
```

Figure 26 Students that attend to seminar and listen in classes will have better grades

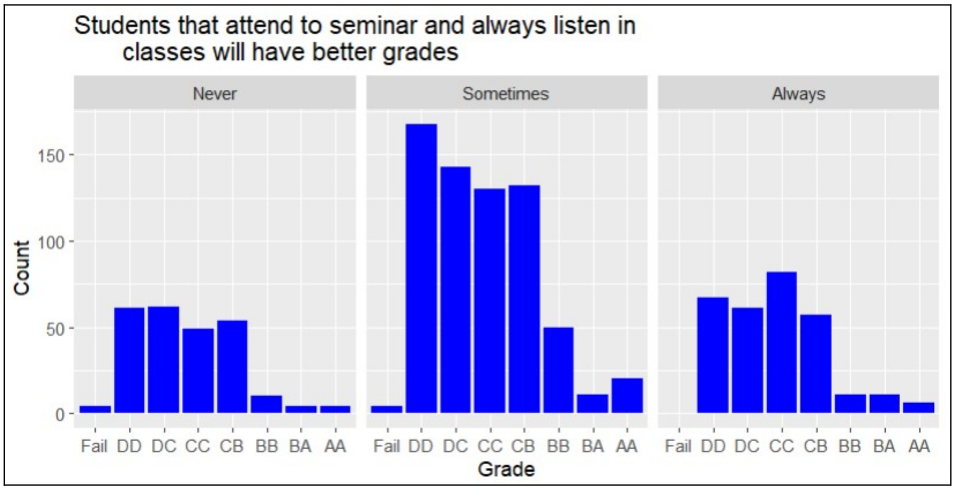


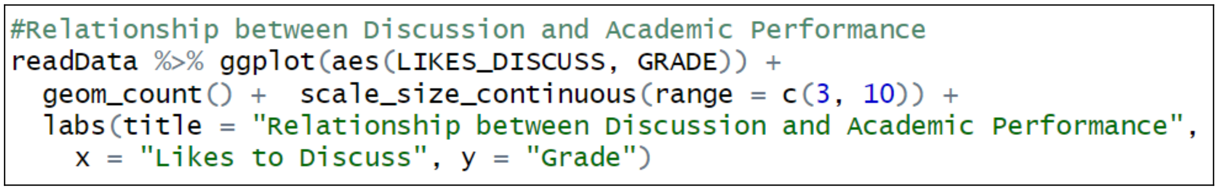
Figure 27 Facet bar plots of "Students that attend to seminar and listen in classes will have better grades"

By using faceting to show subplots, it clearly shows there are some differences among the three responses. The result not only shows that students who attend seminars and listen in classes have better grades, it also shows that none of the students fail when they meet both attributes. Hence, it proved that **attendance at seminars and listening to classes may affect students' academic performance**.

Overall, with the auxiliary data from "Listening in classes", the first two analyses have proved that **attendance at seminars related to the department does have a positive impact on students' academic performance**.

3.1.3 Is there a correlation between discussion participation and grades?

The second analysis will investigate students' participation in discussions and grades, as discussion is one of the most important engagements during academic studies.



```
#Relationship between Discussion and Academic Performance
readData %>% ggplot(aes(LIKES_DISCUSS, GRADE)) +
  geom_count() +  scale_size_continuous(range = c(3, 10)) +
  labs(title = "Relationship between Discussion and Academic Performance",
    x = "Likes to Discuss", y = "Grade")
```

Figure 28 Relationship between Discussion and Academic Performance

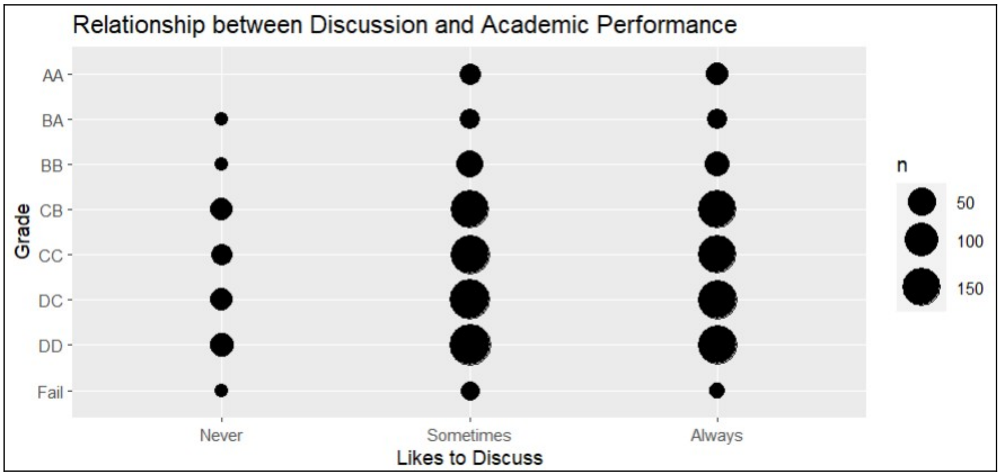
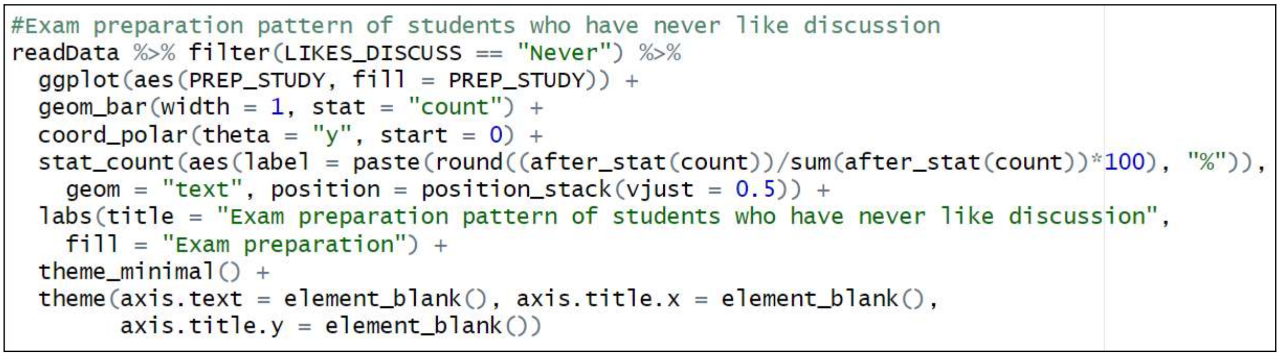


Figure 29 Dot plot of "Relationship between Discussion and Academic Performance"

The dot plot shows the **relationship between discussion and academic performance**. For better differentiate, scale_size_continuous() is used to customize the size of the points in the plot. The dot plot found out the majority of students somehow or always like discussion as the proportion of "Sometimes" and "Always" are bigger than "Never". Besides, students that never like discussion only contains "BA" as their highest grade while other students have achieved more "AA" grades in their responses. To find out the reason, analysis was made to know **how do student study if they never like discussion**.



```
#Exam preparation pattern of students who have never like discussion
readData %>% filter(LIKES_DISCUSS == "Never") %>%
  ggplot(aes(PREP_STUDY, fill = PREP_STUDY)) +
  geom_bar(width = 1, stat = "count") +
  coord_polar(theta = "y", start = 0) +
  stat_count(aes(label = paste(round((after_stat(count))/sum(after_stat(count))*100), "%")),
    geom = "text", position = position_stack(vjust = 0.5)) +
  labs(title = "Exam preparation pattern of students who have never like discussion",
    fill = "Exam preparation") +
  theme_minimal() +
  theme(axis.text = element_blank(), axis.title.x = element_blank(),
        axis.title.y = element_blank())
```

Figure 30 Exam preparation pattern of students who have never like discussion

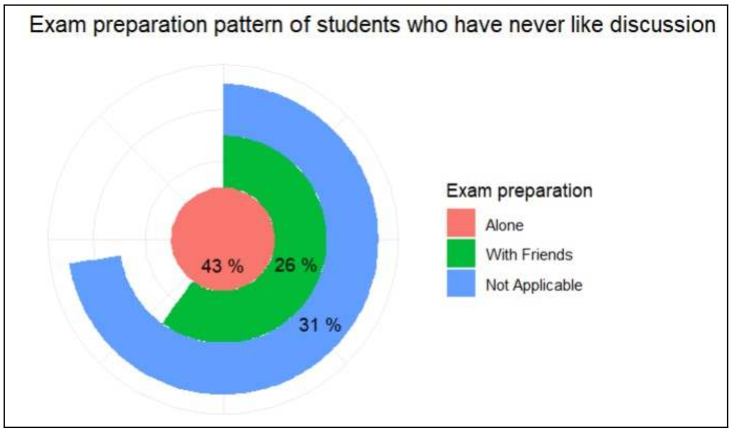
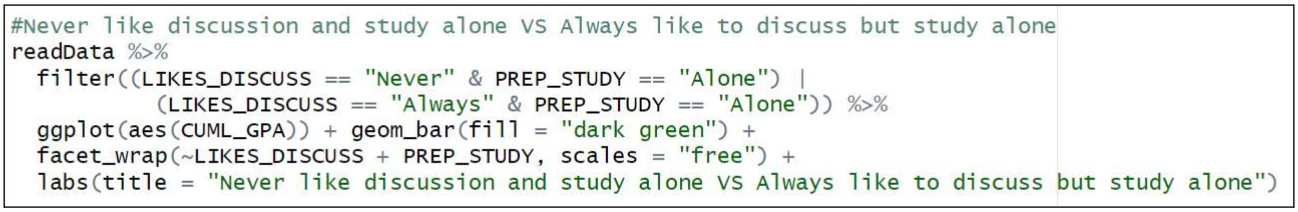


Figure 31 Polar plot of "Exam preparation pattern of students who have never like discussion"

The polar plot shows that most students study alone and some students will study with their friends for exams. This shows that **students never like to discuss doesn't mean they won't do revision with friends**. Thus, to verify the relationship between discussion and grade, comparison is made in the mode of one constant and one variable:



```
#Never like discussion and study alone VS Always like to discuss but study alone
readData %>%
  filter((LIKES_DISCUSS == "Never" & PREP_STUDY == "Alone") |
           (LIKES_DISCUSS == "Always" & PREP_STUDY == "Alone")) %>%
  ggplot(aes(CUML_GPA)) + geom_bar(fill = "dark green") +
  facet_wrap(~LIKES_DISCUSS + PREP_STUDY, scales = "free") +
  labs(title = "Never like discussion and study alone VS Always like to discuss but study alone")
```

Figure 32 Comparison of Never like discussion and study alone & Always like to discuss but study alone

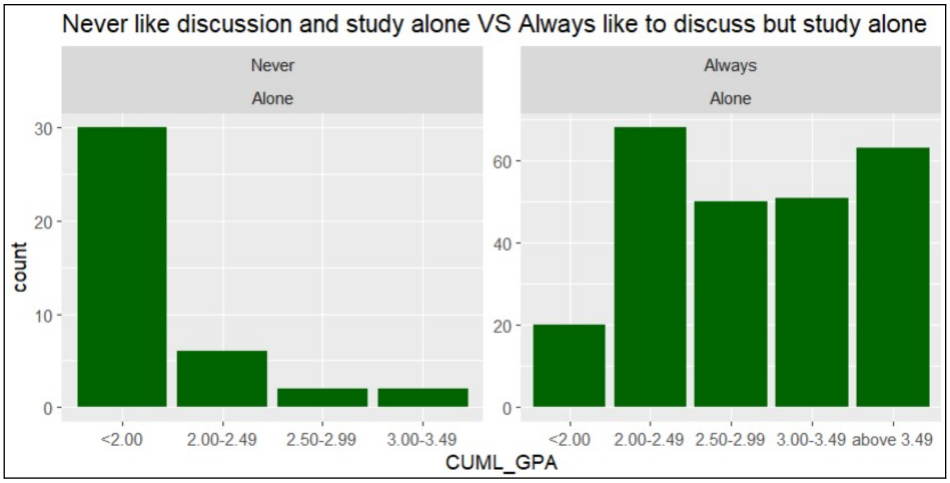


Figure 33 Facet bar plot of comparison

Constant is set for both plots which is study alone, the difference is one never likes discussion, one always likes to discuss. According to cumulative GPA, "<2.00" occupies the highest number in the former, while the latter is more average and mostly over 2.00 GPA. Hence, **discussion has been proven to be effective in improving student achievement**. Despite students studying alone in exams, discussions still played a big role in improving the overall performance. To conclude the second analysis, **there is a strong correlation between discussion participation and grades.**

Conclusion for Objective 1

To conclude the first objective, research demonstrates that student participation in academic sessions such as attendance in classes or seminars, discussion in class, and exam preparation with friends, does improve their academic performance. Besides that, the majority of students are aggressive in a variety of engagements, which indirectly illustrates the importance of student involvement in academic sessions.

3.2 Objective 2 (Lee Wen Jing, TP071630)

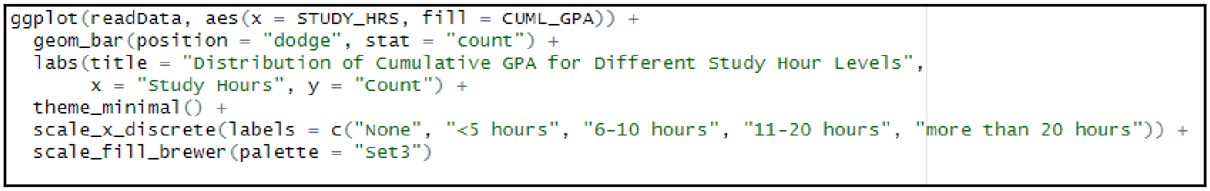
The next part of the analysis will focus on **how much time students allocated for revision**, both weekly and prior to the exams.

Here are the objective questions for this subsection:

1. Is there a relationship between the number of hours spent on self-study and the final results of the students?

2. How well are the students prepared for the midterm? How does this affect their CGPA?

3.2.1: Is there a relationship between the number of hours spent on self-study and the final results of the students?



```
ggplot(readData, aes(x = STUDY_HRS, fill = CUML_GPA)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Distribution of Cumulative GPA for Different Study Hour Levels",
       x = "Study Hours", y = "Count") +
  theme_minimal() +
  scale_x_discrete(labels = c("None", "<5 hours", "6-10 hours", "11-20 hours", "more than 20 hours")) +
  scale_fill_brewer(palette = "Set3")
```

Figure 34 creating a bar chart to show the relationship between study hours and CGPA

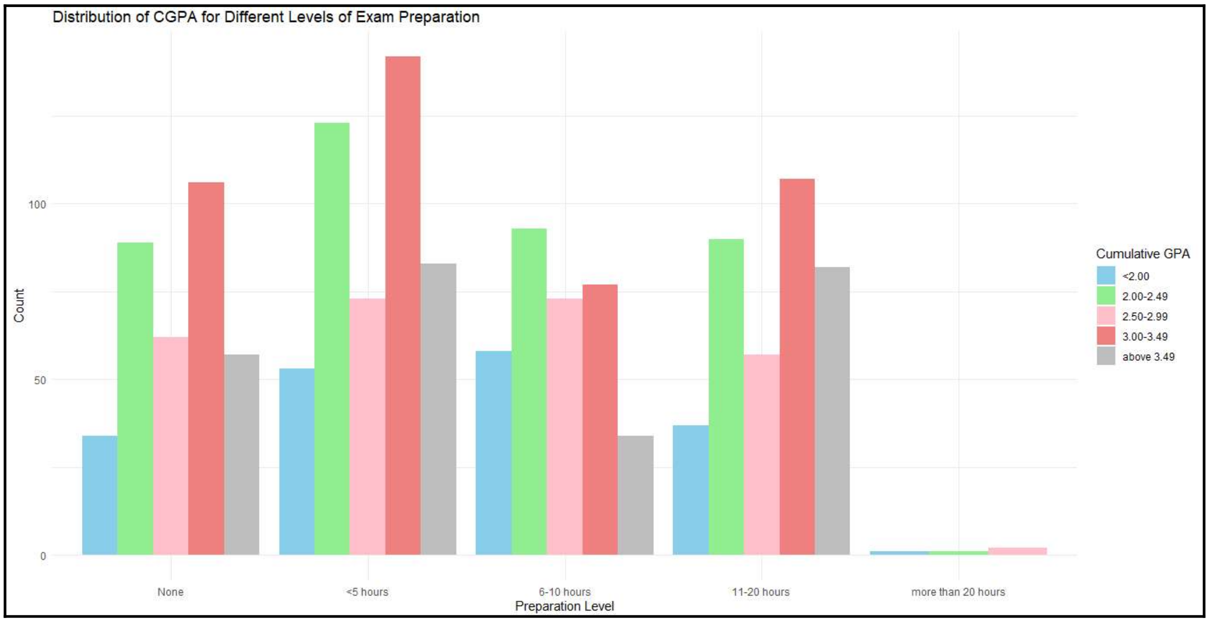
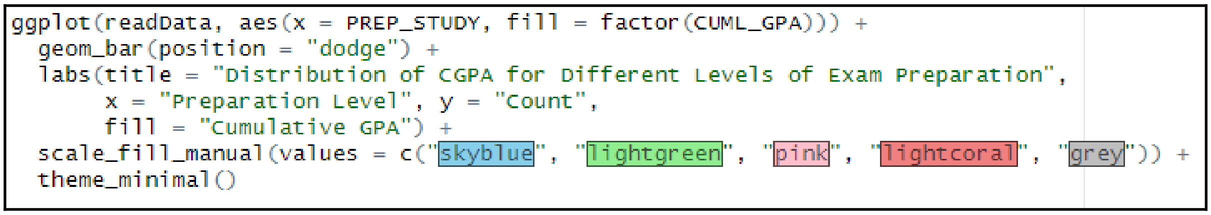


Figure 35 Bar chart generated from the code above

As can be seen from above, a bar chart is created to see if there is a relationship between our two variables. However, at first glance it seems there are no obvious patterns linking the two. In almost all categories of different weekly study hours, most students scored a CGPA between 3.00 and 3.49; with students who scored between 2.00 and 2.49 counting as second most in those categories as well. The only two different categories are the one with 6-10 weekly study hours and the one above 10 hours. In the first the CGPA of most students range from 2.00 to 2.49 while in the latter, the counts are too less to form a pattern for observation. From this bar chart we can conclude that **there are no visible patterns linking the two variables**, and this moves us forward to the next subsection to see how well the students have prepared for the exams and how this might be a factor influencing their CGPA.

3.2.2: How well are the students prepared for the midterm? How does this affect their CGPA?



```
ggplot(readData, aes(x = PREP_STUDY, fill = factor(CUML_GPA))) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of CGPA for Different Levels of Exam Preparation",
       x = "Preparation Level", y = "Count",
       fill = "Cumulative GPA") +
  scale_fill_manual(values = c("skyblue", "lightgreen", "pink", "lightcoral", "grey")) +
  theme_minimal()
```

Figure 36 Code to create a bar chart with PREP_STUDY against CUML_GPA

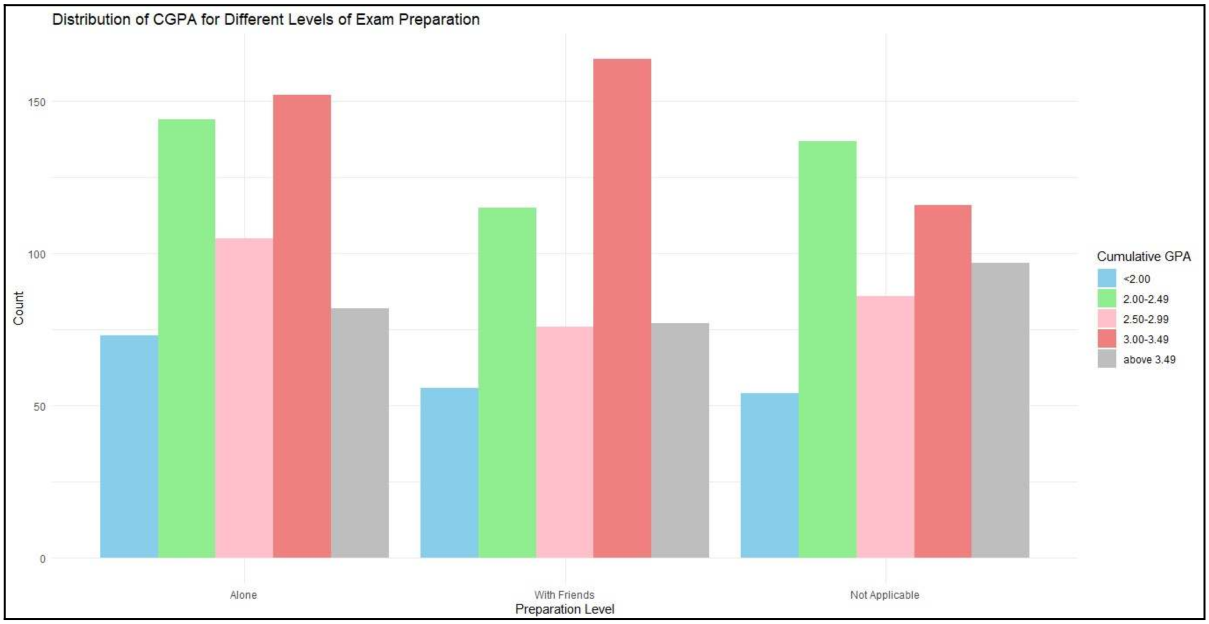
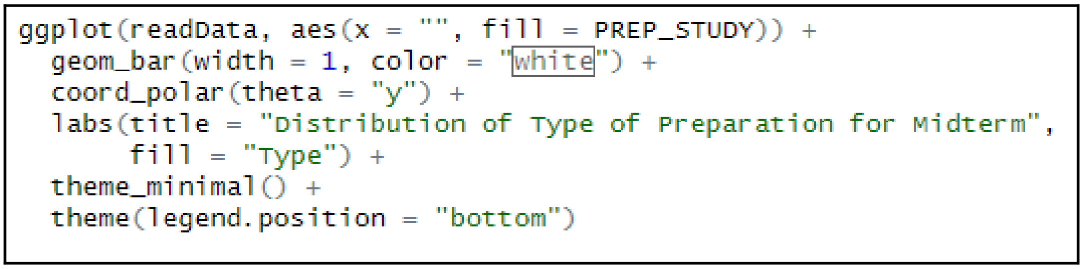


Figure 37 Bar chart created from the code above

From the chart above, again no visible pattern linking the two variables can be seen. Next I'm going to create a pie chart to investigate the proportion of students in each of these category:



```
ggplot(readData, aes(x = "", fill = PREP_STUDY)) +
  geom_bar(width = 1, color = "white") +
  coord_polar(theta = "y") +
  labs(title = "Distribution of Type of Preparation for Midterm",
       fill = "Type") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

Figure 38 Code to create a pie chart

The code generates a visualization resembling a pie chart using the ggplot2 package in R. Since ggplot2 doesn't have a direct pie chart option, a bar chart with polar coordinates is used. This chart illustrates the distribution of different types of preparation for midterm exams. Each section of the "pie" corresponds to a category in the 'PREP_STUDY' variable. The width of each bar is set to 1, and colours represent the various types of preparation. The chart is displayed in polar coordinates, transforming the bar chart into a pie chart-like format. The title is set as 'Distribution of Type of Preparation for Midterm', and the legend is positioned at the bottom of the chart for clarity.

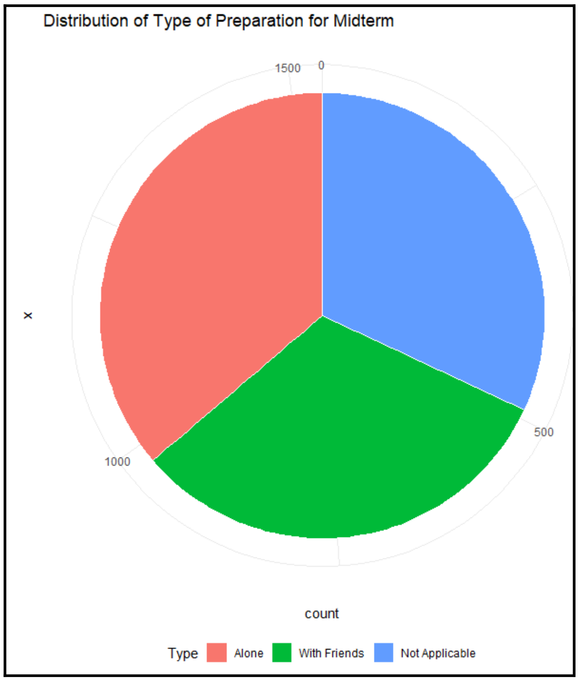
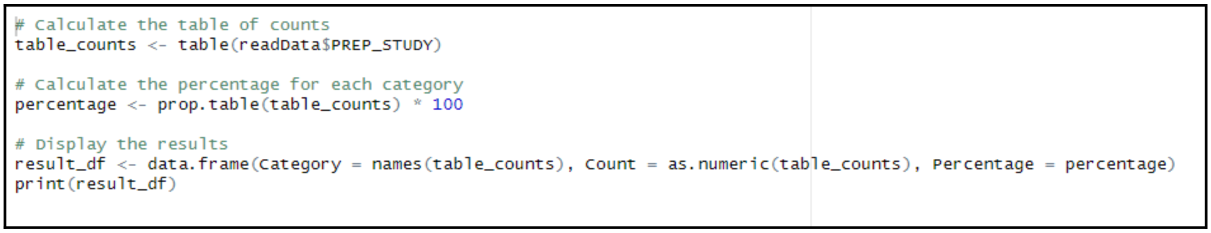
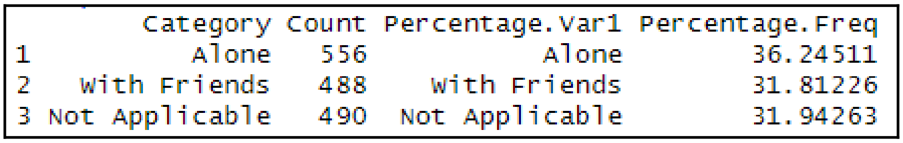


Figure 39 Visualization of pie chart generated



```
# Calculate the table of counts
table_counts <- table(readData$PREP_STUDY)

# Calculate the percentage for each category
percentage <- prop.table(table_counts) * 100

# Display the results
result_df <- data.frame(Category = names(table_counts), Count = as.numeric(table_counts), Percentage = percentage)
print(result_df)
```

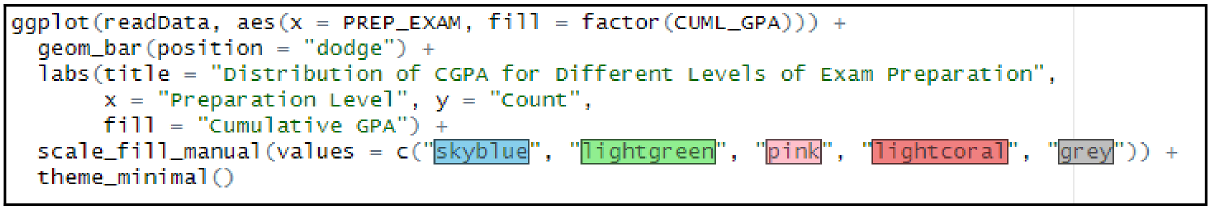
Figure 40 Code to calculate the percentage of each category



```
        Category Count Percentage.Var1 Percentage.Freq
1          Alone   556           Alone        36.24511
2   With Friends   488    With Friends        31.81226
3 Not Applicable   490  Not Applicable        31.94263
```

Figure 41 Percentages generate

From the pie chart and the calculation above, we can see that students are spread almost equally between the three categories, with 30-40% of students in each of them.



```
ggplot(readData, aes(x = PREP_EXAM, fill = factor(CUML_GPA))) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of CGPA for Different Levels of Exam Preparation",
       x = "Preparation Level", y = "Count",
       fill = "Cumulative GPA") +
  scale_fill_manual(values = c("skyblue", "lightgreen", "pink", "lightcoral", "grey")) +
  theme_minimal()
```

Figure 42 Code to create a bar chart with PREP_EXAM against CUML_GPA

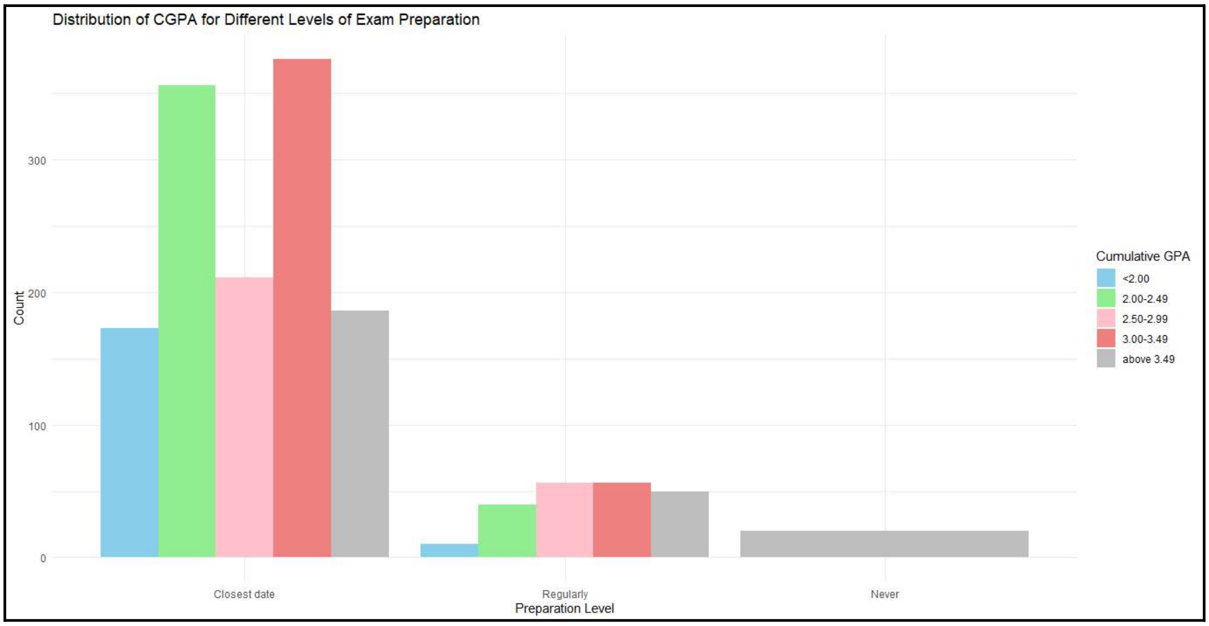


Figure 43 Bar chart generated

Next, a similar bar chart is created to see if there is a relationship between how regularly the students studied and their CGPA. As can be seen from the two bar charts above, even though most students studied alone and at the closest date possible, all students who never studied for the exams scored the highest range of CGPA, which is above 3.49.

Conclusion for objective 2:

From these charts a conclusion can be drawn: that **there is no relationship between the time allocated for revision and the final CGPS of the students**,

3.3 Objective 3 (Abdulrahman Gamil Murshed Humadi, TP070609)

The third objective is to examine the relationship between students' family background and academic performance.

The research on this objective will be done through answering the following questions:

Q1- What effect does the number of siblings have on the students' academic performance?

Q2- Whether the parents' relationships status has an effect students' academic performance or not?

Q3- - Does the academic qualification of parents affect students' academic performance?

The following attributes will be used to conduct this analysis:

X._SIBLINGS (Number of sisters/brothers (if available))

PARENT_STATUS (Parental status)

MOTHER_EDU (Mother's education)

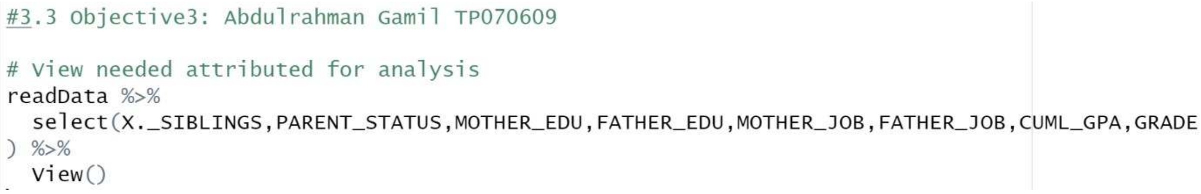
FATHER_EDU (Father's education)

MOTHER_JOB (Mother's occupation)

FATHER_JOB (Father's occupation)

CUML_GPA (Cumulative grade point average (in the last semester))

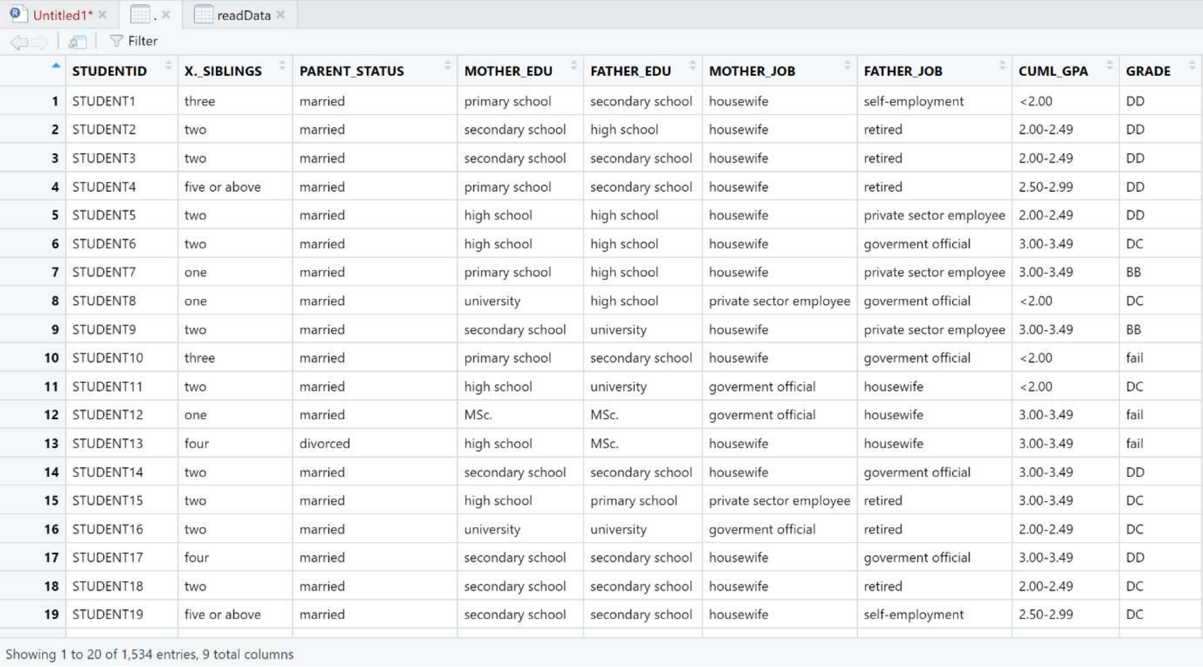
GRADE (OUTPUT Grade)



```
#3.3 Objective3: Abdulrahman Gamil TP070609

# View needed attributed for analysis
readData %>%
  select(X._SIBLINGS,PARENT_STATUS,MOTHER_EDU,FATHER_EDU,MOTHER_JOB,FATHER_JOB,CUML_GPA,GRADE
) %>%
  View()
```

Figure 44 view attributes code

Using the R script shown (Figure 44), it will view all the attributes that are going to be used for this analysis.

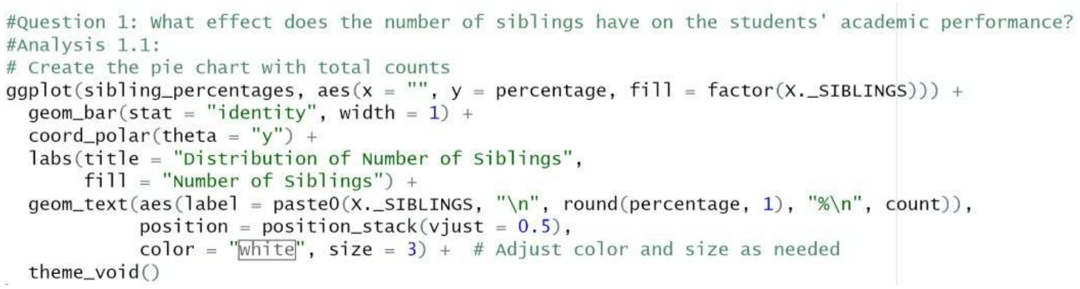


	STUDENTID	X._SIBLINGS	PARENT_STATUS	MOTHER_EDU	FATHER_EDU	MOTHER_JOB	FATHER_JOB	CUML_GPA	GRADE
1	STUDENT1	three	married	primary school	secondary school	housewife	self-employment	<2.00	DD
2	STUDENT2	two	married	secondary school	high school	housewife	retired	2.00-2.49	DD
3	STUDENT3	two	married	secondary school	secondary school	housewife	retired	2.00-2.49	DD
4	STUDENT4	five or above	married	primary school	secondary school	housewife	retired	2.50-2.99	DD
5	STUDENT5	two	married	high school	high school	housewife	private sector employee	2.00-2.49	DD
6	STUDENT6	two	married	high school	high school	housewife	goverment official	3.00-3.49	DC
7	STUDENT7	one	married	primary school	high school	housewife	private sector employee	3.00-3.49	BB
8	STUDENT8	one	married	university	high school	private sector employee	goverment official	<2.00	DC
9	STUDENT9	two	married	secondary school	university	housewife	private sector employee	3.00-3.49	BB
10	STUDENT10	three	married	primary school	secondary school	housewife	goverment official	<2.00	fail
11	STUDENT11	two	married	high school	university	goverment official	housewife	<2.00	DC
12	STUDENT12	one	married	MSc.	MSc.	goverment official	housewife	3.00-3.49	fail
13	STUDENT13	four	divorced	high school	MSc.	housewife	housewife	3.00-3.49	fail
14	STUDENT14	two	married	secondary school	secondary school	housewife	goverment official	3.00-3.49	DD
15	STUDENT15	two	married	high school	primary school	private sector employee	retired	3.00-3.49	DC
16	STUDENT16	two	married	university	university	goverment official	retired	2.00-2.49	DC
17	STUDENT17	four	married	secondary school	secondary school	housewife	goverment official	3.00-3.49	DD
18	STUDENT18	two	married	secondary school	secondary school	housewife	retired	2.00-2.49	DC
19	STUDENT19	five or above	married	secondary school	secondary school	housewife	self-employment	2.50-2.99	DC

Showing 1 to 20 of 1,534 entries, 9 total columns

Figure 45 Attributes used

3.3.1 What effect does the number of siblings have on the students' academic performance?



```
#Question 1: What effect does the number of siblings have on the students' academic performance?
#Analysis 1.1:
# Create the pie chart with total counts
ggplot(sibling_percentages, aes(x = "", y = percentage, fill = factor(X._SIBLINGS))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  labs(title = "Distribution of Number of Siblings",
       fill = "Number of Siblings") +
  geom_text(aes(label = paste0(X._SIBLINGS, "\n", round(percentage, 1), "%\n", count)),
            position = position_stack(vjust = 0.5),
            color = "white", size = 3) +  # Adjust color and size as needed
  theme_void()
```

Figure 46 Pie chart code: No of Siblings

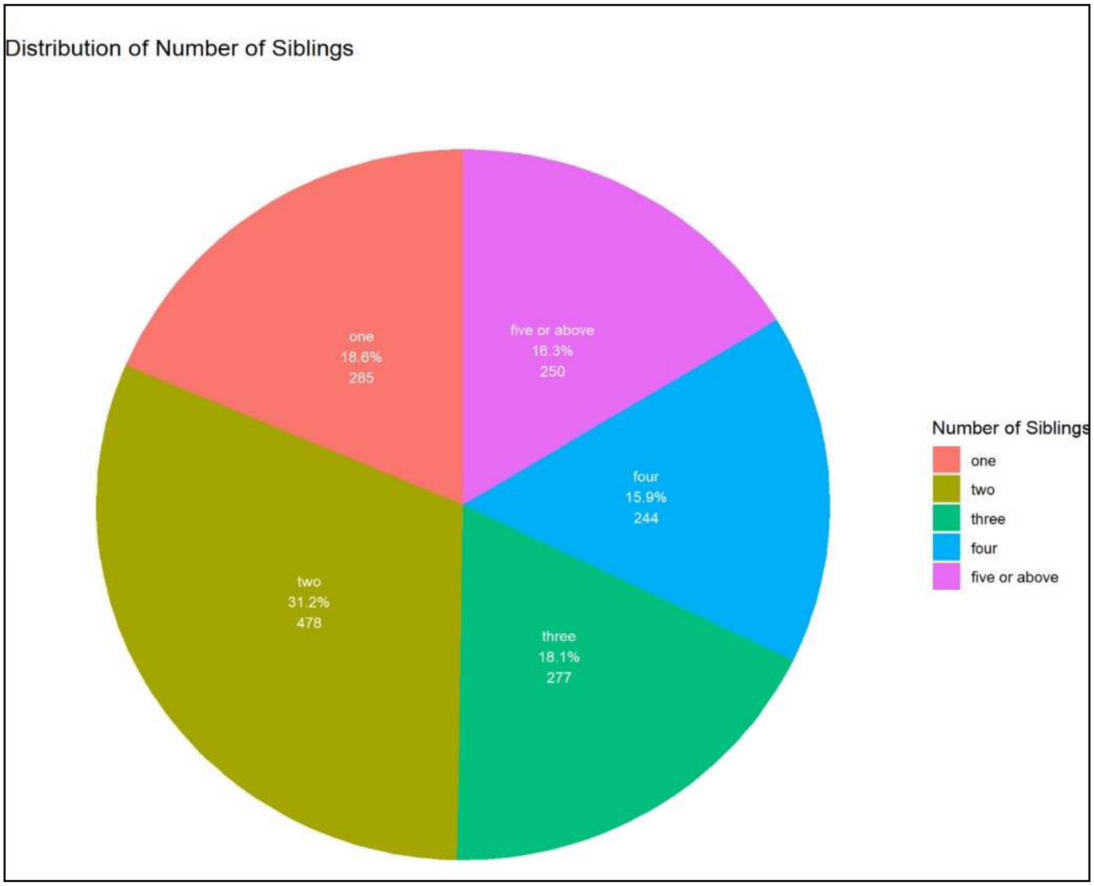
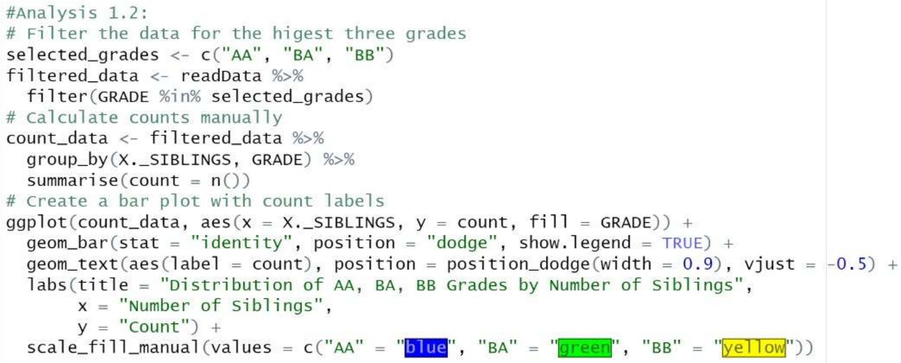


Figure 47 Pie chart visualization: No of Siblings

The pie chart above shows the percentages of students and their respective number of siblings. We can see that most of the students (31.2%) have two siblings, while students with four siblings correspond the least percentage (15.9%).



```
#Analysis 1.2:
# Filter the data for the higest three grades
selected_grades <- c("AA", "BA", "BB")
filtered_data <- readData %>%
  filter(GRADE %in% selected_grades)
# Calculate counts manually
count_data <- filtered_data %>%
  group_by(X._SIBLINGS, GRADE) %>%
  summarise(count = n())
# Create a bar plot with count labels
ggplot(count_data, aes(x = X._SIBLINGS, y = count, fill = GRADE)) +
  geom_bar(stat = "identity", position = "dodge", show.legend = TRUE) +
  geom_text(aes(label = count), position = position_dodge(width = 0.9), vjust = -0.5) +
  labs(title = "Distribution of AA, BA, BB Grades by Number of Siblings",
       x = "Number of Siblings",
       y = "Count") +
  scale_fill_manual(values = c("AA" = "blue", "BA" = "green", "BB" = "yellow"))
```

Figure 48 Bar chart code T3

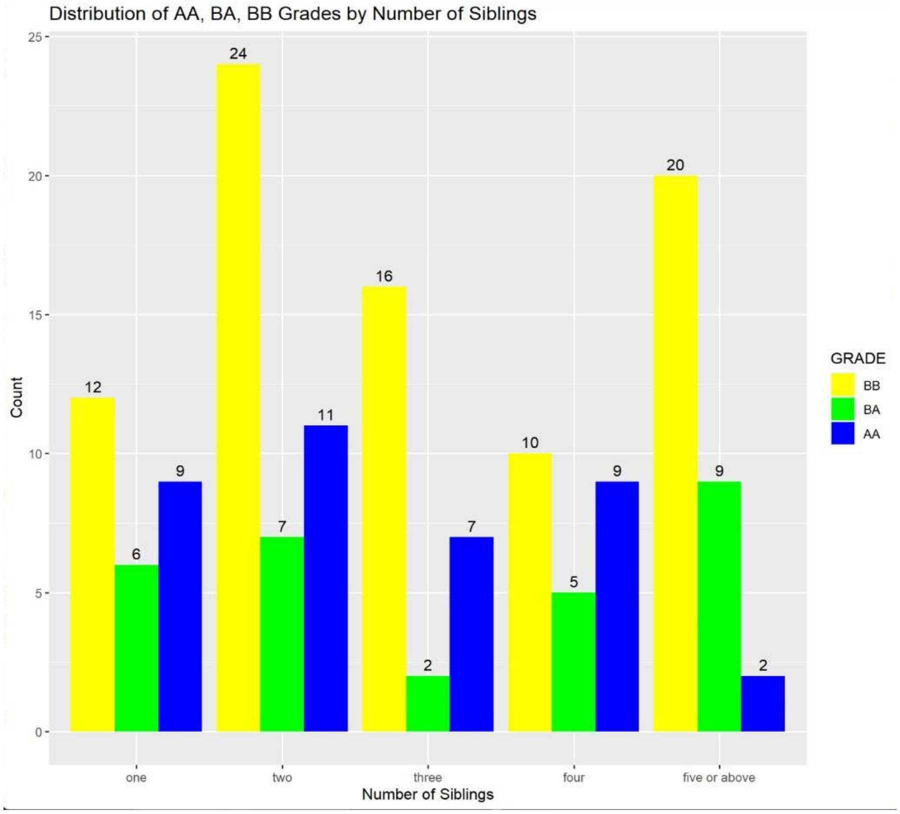


Figure 49 Bar chart visualization T3

Next, the above bar chart shows the distribution of the highest three grades by the number of siblings. If we start calculating the percentage of how many students got one of the highest three marks, we will come up with the following:

One sibling: 12+6+9/285 = 9.5%

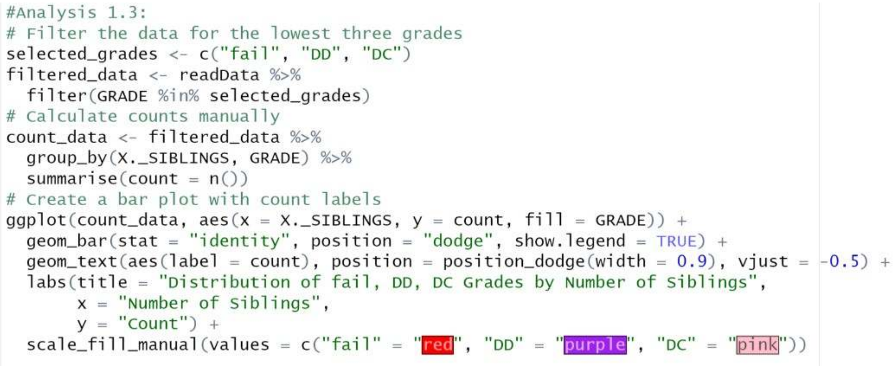
Two siblings: 24+7+11/478 = 8.8%

Three siblings: 16+2+7/ 277 = 6.9%

Four siblings: 10+5+9/244 = 9.8%

Five and above sibling: 20+9+2/250 = 12.4%

From the fluctuating percentages above while the number of siblings is increasing, it is concluded that the number of siblings doesn’t correspond to academic performance. To double check the conclusion of this analysis, the opposite will be done. Plotting the distribution of the lowest three grades then finding the percentages.



```
#Analysis 1.3:
# Filter the data for the lowest three grades
selected_grades <- c("fail", "DD", "DC")
filtered_data <- readData %>%
  filter(GRADE %in% selected_grades)
# Calculate counts manually
count_data <- filtered_data %>%
  group_by(X._SIBLINGS, GRADE) %>%
  summarise(count = n())
# Create a bar plot with count labels
ggplot(count_data, aes(x = X._SIBLINGS, y = count, fill = GRADE)) +
  geom_bar(stat = "identity", position = "dodge", show.legend = TRUE) +
  geom_text(aes(label = count), position = position_dodge(width = 0.9), vjust = -0.5) +
  labs(title = "Distribution of fail, DD, DC Grades by Number of Siblings",
       x = "Number of Siblings",
       y = "Count") +
  scale_fill_manual(values = c("fail" = "red", "DD" = "purple", "DC" = "pink"))
```

Figure 50 Bar chart code L3

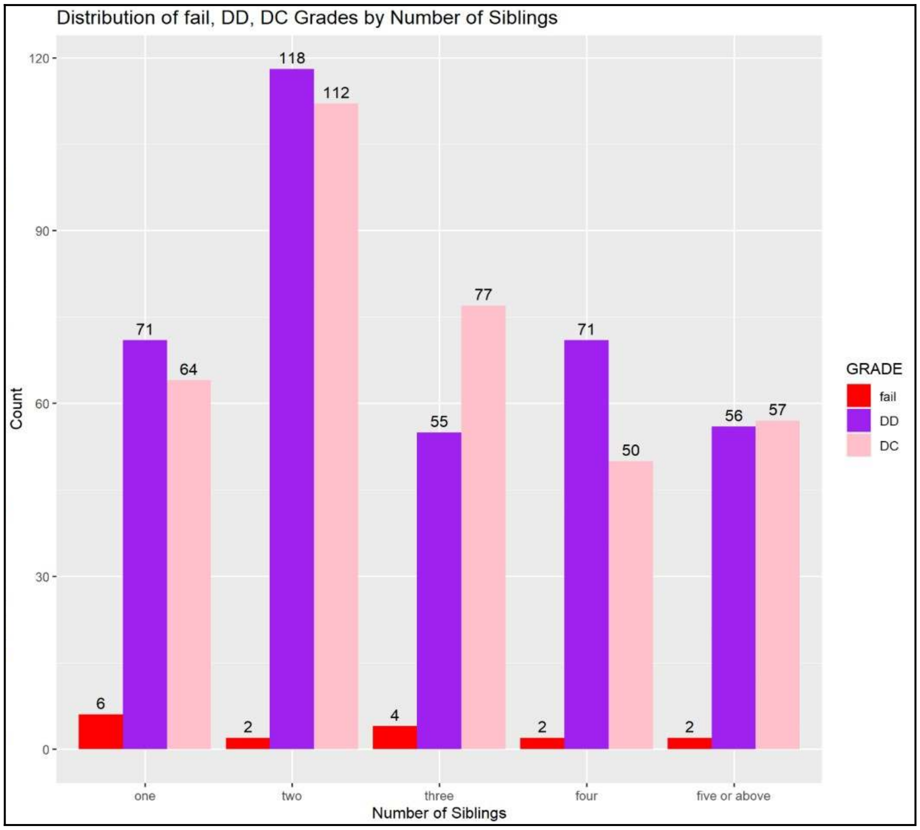


Figure 51 Bar chart visualization L3

One sibling: (6+71+64)/285 = 49.5%

Two siblings: (2+118+112)/478 = 48.5%

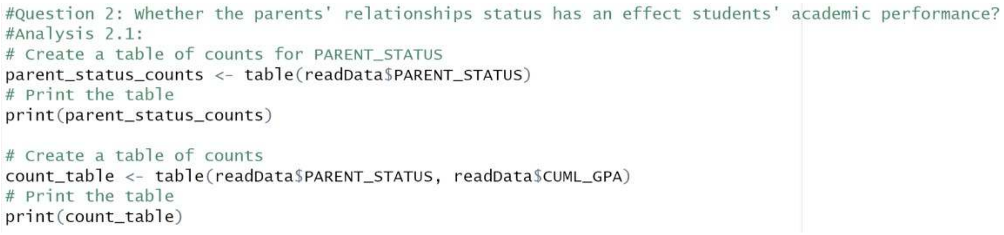
Three siblings: (4+55+77)/ 277 = 49.1%

Four siblings: (2+71+50)/244 = 50.4%

Five and above sibling: (2+56+57)/250 = 46.0%

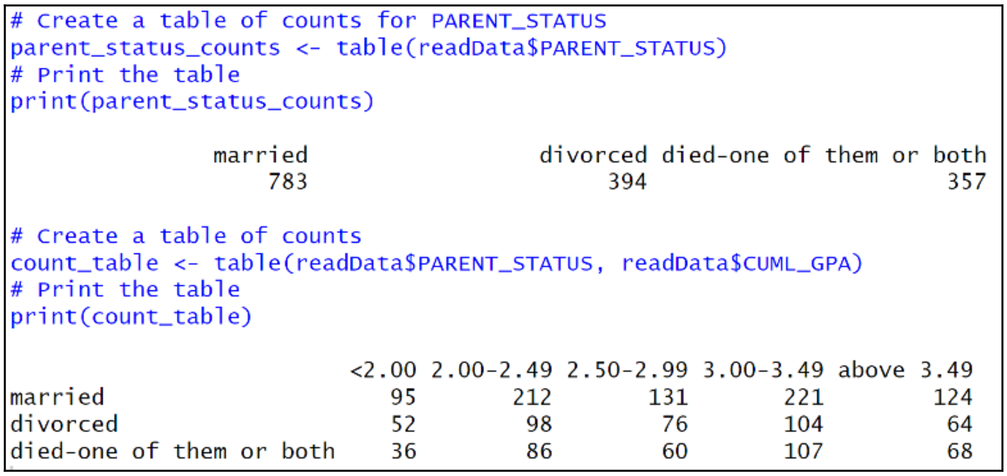
There is no pattern here too which supports the previous result of the analysis.

3.3.2 Whether the parents' relationships status has an effect students' academic performance or not?



```
#Question 2: Whether the parents' relationships status has an effect students' academic performance?
#Analysis 2.1:
# Create a table of counts for PARENT_STATUS
parent_status_counts <- table(readData$PARENT_STATUS)
# Print the table
print(parent_status_counts)

# Create a table of counts
count_table <- table(readData$PARENT_STATUS, readData$CUML_GPA)
# Print the table
print(count_table)
```

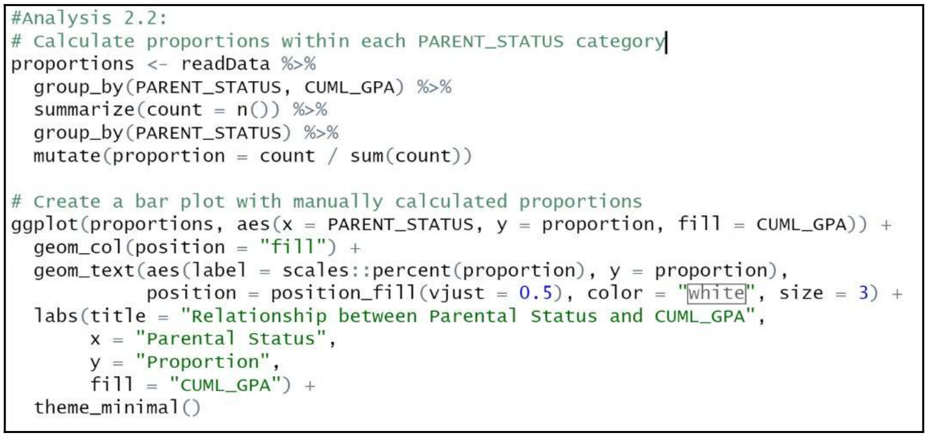
Figure 52 Count table code



```
# Create a table of counts for PARENT_STATUS
parent_status_counts <- table(readData$PARENT_STATUS)
# Print the table
print(parent_status_counts)

                married                     divorced died-one of them or both 
                    783                          394                      357 

# Create a table of counts
count_table <- table(readData$PARENT_STATUS, readData$CUML_GPA)
# Print the table
print(count_table)

                          <2.00 2.00-2.49 2.50-2.99 3.00-3.49 above 3.49
married                      95       212       131       221        124
divorced                     52        98        76       104         64
died-one of them or both     36        86        60       107         68
```

Figure 53 Count table

It is seen here in (Figure 53) first table, that approximately half of the parents' status are *married* (783, 51.0%), the other half consists of almost two quarters: *divorced* (394, 25.6%) and *died-one of them or both* (357, 23.4%). In the second table, it shows the distribution of parent's status with grades. Similar percentages are seen in every parent's status-CGPA combination, let's take 3.49 CGPA as an example: *married* (221, 51.2%) of this combination, while *divorced* (104, 24.1%) and *died-one of them or both* (107, 24.7%). This suggests that there is no relationship between parents' status and academic performance. A bar chart is made to visualize these percentages.



```
#Analysis 2.2:
# Calculate proportions within each PARENT_STATUS category
proportions <- readData %>%
  group_by(PARENT_STATUS, CUML_GPA) %>%
  summarize(count = n()) %>%
  group_by(PARENT_STATUS) %>%
  mutate(proportion = count / sum(count))

# Create a bar plot with manually calculated proportions
ggplot(proportions, aes(x = PARENT_STATUS, y = proportion, fill = CUML_GPA)) +
  geom_col(position = "fill") +
  geom_text(aes(label = scales::percent(proportion), y = proportion),
            position = position_fill(vjust = 0.5), color = "white", size = 3) +
  labs(title = "Relationship between Parental Status and CUML_GPA",
       x = "Parental Status",
       y = "Proportion",
       fill = "CUML_GPA") +
  theme_minimal()
```

Figure 54 Bar chart code Status vs CGPA

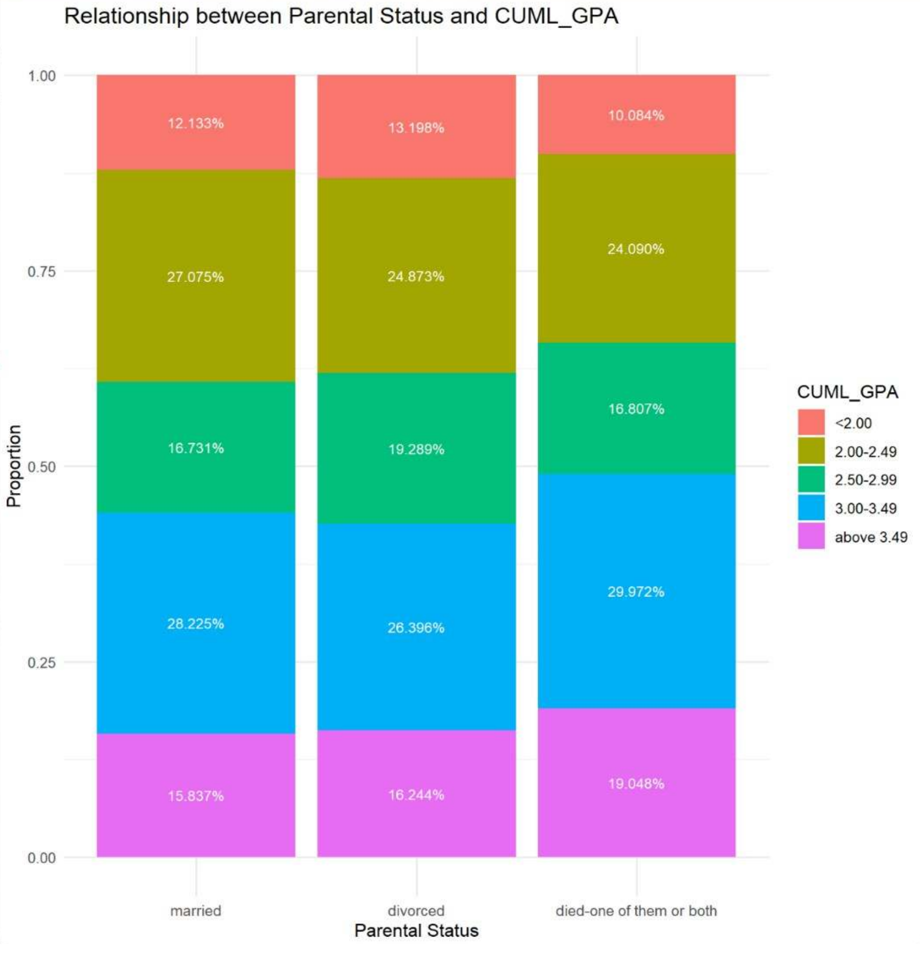
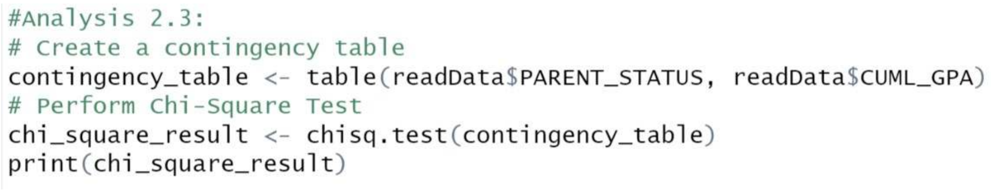


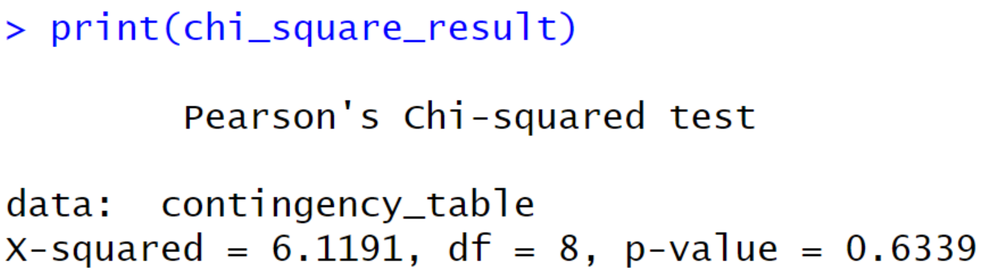
Figure 55 Bar chart code Status vs CGPA visualization

The bar chart in (Figure 55) shows the proportion of CGPA in each of the parents status. All the percentages are almost similar across the three categories.



```
#Analysis 2.3:
# Create a contingency table
contingency_table <- table(readData$PARENT_STATUS, readData$CUML_GPA)
# Perform Chi-Square Test
chi_square_result <- chisq.test(contingency_table)
print(chi_square_result)
```

Figure 56 Chi-test: Status vs CGPA

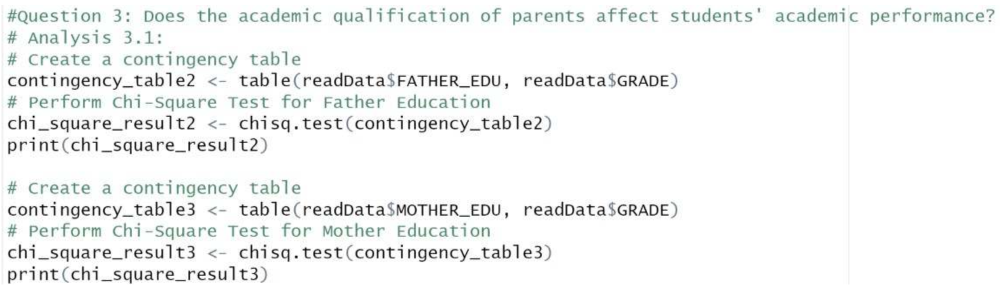


```
> print(chi_square_result)

        Pearson's Chi-squared test

data:  contingency_table
X-squared = 6.1191, df = 8, p-value = 0.6339
```

Figure 57 Chi-test results: Status vs CGPA

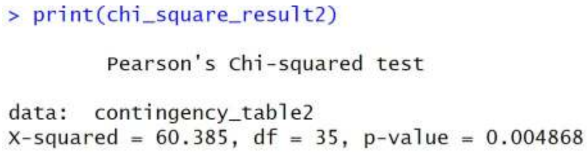
A chi-square test is done to double check the result of this analysis (Figure 57). The Null Hypothesis (Ho) in this case is: there is no association between parents' status and academic performance. Since the p-value (0.6339) is greater than the commonly chosen significance level of 0.05, so it does not reject the null hypothesis. Therefore, based on the data and the chi-square test, there is insufficient evidence to conclude that there is a significant association between parents' status and academic performance.

3.3.3 Does the academic qualification of parents affect students' academic performance?



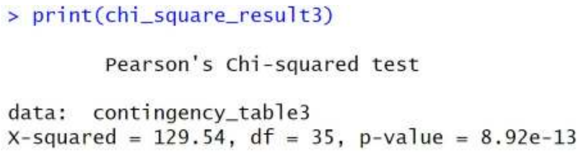
```
#Question 3: Does the academic qualification of parents affect students' academic performance?
# Analysis 3.1:
# Create a contingency table
contingency_table2 <- table(readData$FATHER_EDU, readData$GRADE)
# Perform Chi-Square Test for Father Education
chi_square_result2 <- chisq.test(contingency_table2)
print(chi_square_result2)

# Create a contingency table
contingency_table3 <- table(readData$MOTHER_EDU, readData$GRADE)
# Perform Chi-Square Test for Mother Education
chi_square_result3 <- chisq.test(contingency_table3)
print(chi_square_result3)
```

Figure 58 Chi-test: Father&Mother EDU vs Grade



```
> print(chi_square_result2)

        Pearson's Chi-squared test

data:  contingency_table2
X-squared = 60.385, df = 35, p-value = 0.004868
```

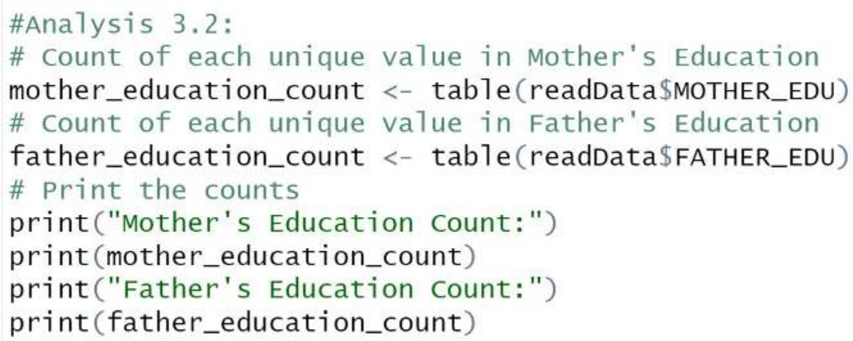
Figure 59 Chi-test result: Father EDU vs Grade



```
> print(chi_square_result3)

        Pearson's Chi-squared test

data:  contingency_table3
X-squared = 129.54, df = 35, p-value = 8.92e-13
```

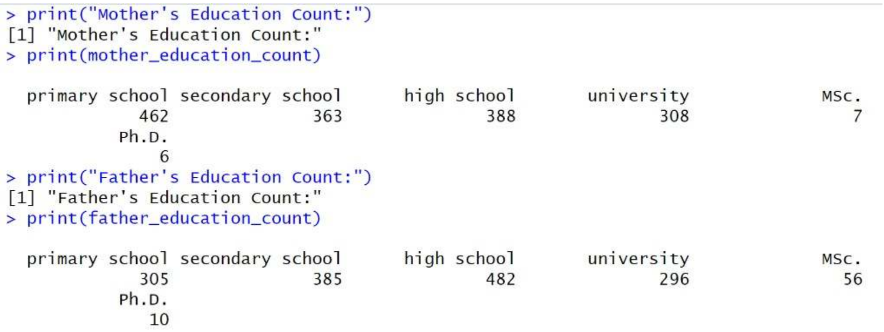
Figure 60 Chi-test result: Mother EDU vs Grade

To start this analysis, Chi-test is done on Father’s (Figure 59) and Mother’s (Figure 60) education level against grade. The Null Hypothesis (Ho) in both cases is: there is no association between Father’s / Mother’s Education and academic performance. Since the p-values (0.004868) and (8.92×10^-13) respectively, are less than the commonly chosen significance level of 0.05, so it does reject the null hypothesis. Therefore, the p-values provide evidence against the null hypothesis, suggesting that there is an association between parents' education levels and students' grades.



```
#Analysis 3.2:
# Count of each unique value in Mother's Education
mother_education_count <- table(readData$MOTHER_EDU)
# Count of each unique value in Father's Education
father_education_count <- table(readData$FATHER_EDU)
# Print the counts
print("Mother's Education Count:")
print(mother_education_count)
print("Father's Education Count:")
print(father_education_count)
```

Figure 61 Count: Mother&Father Categories

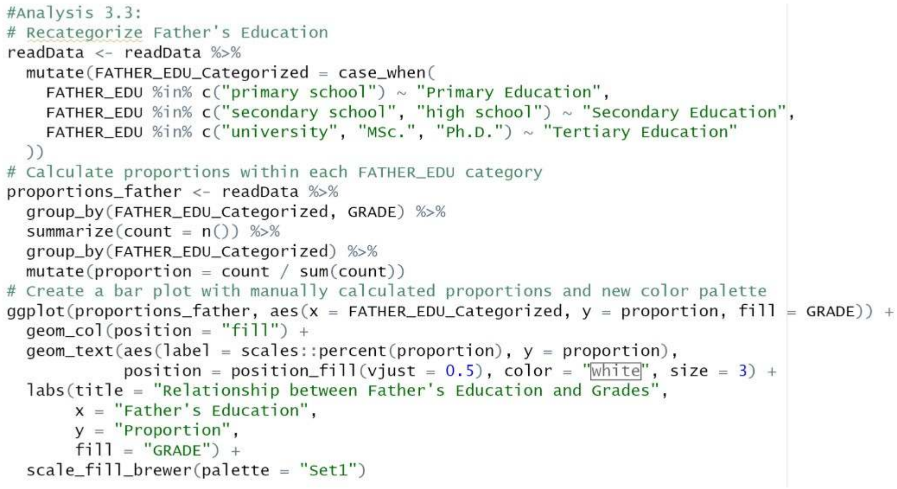


```
> print("Mother's Education Count:")
[1] "Mother's Education Count:"
> print(mother_education_count)

  primary school secondary school      high school       university              MSc. 
             462              363              388              308                 7 
           Ph.D. 
               6 
> print("Father's Education Count:")
[1] "Father's Education Count:"
> print(father_education_count)

  primary school secondary school      high school       university              MSc. 
             305              385              482              296                56 
           Ph.D. 
              10 
```

Figure 62 Count result: Mother&Father Categories

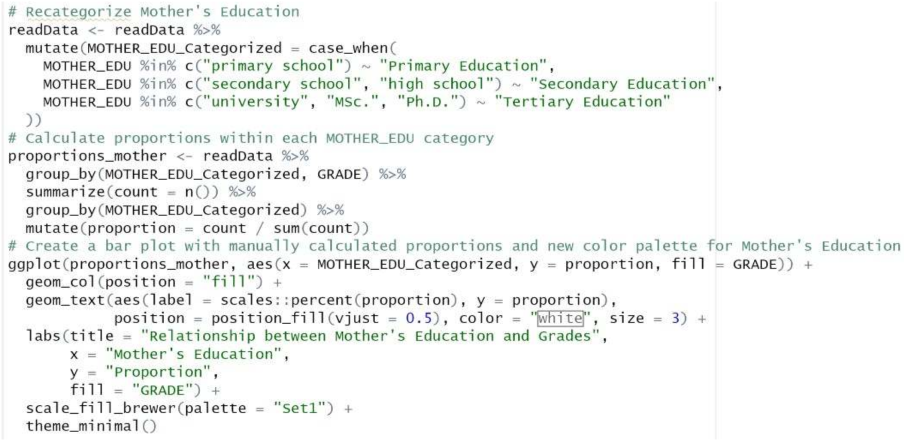
Further analysis will be done to confirm the above hypothesis testing. However, since the number of parents with a MSc. and Ph.D. is very low compared to the others, (Figure 62) the education levels will be divided to three sections so that the analysis is more accurate representation:

- Primary Education: consists of primary school.
- Secondary Education: consists of secondary school and high school.
- Tertiary Education: consists of university, MSc. and Ph.D.



```
#Analysis 3.3:
# Recategorize Father's Education
readData <- readData %>%
  mutate(FATHER_EDU_Categorized = case_when(
    FATHER_EDU %in% c("primary school") ~ "Primary Education",
    FATHER_EDU %in% c("secondary school", "high school") ~ "Secondary Education",
    FATHER_EDU %in% c("university", "MSc.", "Ph.D.") ~ "Tertiary Education"
  ))
# Calculate proportions within each FATHER_EDU category
proportions_father <- readData %>%
  group_by(FATHER_EDU_Categorized, GRADE) %>%
  summarize(count = n()) %>%
  group_by(FATHER_EDU_Categorized) %>%
  mutate(proportion = count / sum(count))
# Create a bar plot with manually calculated proportions and new color palette
ggplot(proportions_father, aes(x = FATHER_EDU_Categorized, y = proportion, fill = GRADE)) +
  geom_col(position = "fill") +
  geom_text(aes(label = scales::percent(proportion), y = proportion),
            position = position_fill(vjust = 0.5), color = "white", size = 3) +
  labs(title = "Relationship between Father's Education and Grades",
       x = "Father's Education",
       y = "Proportion",
       fill = "GRADE") +
  scale_fill_brewer(palette = "Set1")
```

Figure 63 Bar chart code: Father Edu vs Grade



```
# Recategorize Mother's Education
readData <- readData %>%
  mutate(MOTHER_EDU_Categorized = case_when(
    MOTHER_EDU %in% c("primary school") ~ "Primary Education",
    MOTHER_EDU %in% c("secondary school", "high school") ~ "Secondary Education",
    MOTHER_EDU %in% c("university", "MSc.", "Ph.D.") ~ "Tertiary Education"
  ))
# Calculate proportions within each MOTHER_EDU category
proportions_mother <- readData %>%
  group_by(MOTHER_EDU_Categorized, GRADE) %>%
  summarize(count = n()) %>%
  group_by(MOTHER_EDU_Categorized) %>%
  mutate(proportion = count / sum(count))
# Create a bar plot with manually calculated proportions and new color palette for Mother's Education
ggplot(proportions_mother, aes(x = MOTHER_EDU_Categorized, y = proportion, fill = GRADE)) +
  geom_col(position = "fill") +
  geom_text(aes(label = scales::percent(proportion), y = proportion),
            position = position_fill(vjust = 0.5), color = "white", size = 3) +
  labs(title = "Relationship between Mother's Education and Grades",
       x = "Mother's Education",
       y = "Proportion",
       fill = "GRADE") +
  scale_fill_brewer(palette = "Set1") +
  theme_minimal()
```

Figure 64 Bar chart code: Mother Edu vs Grade

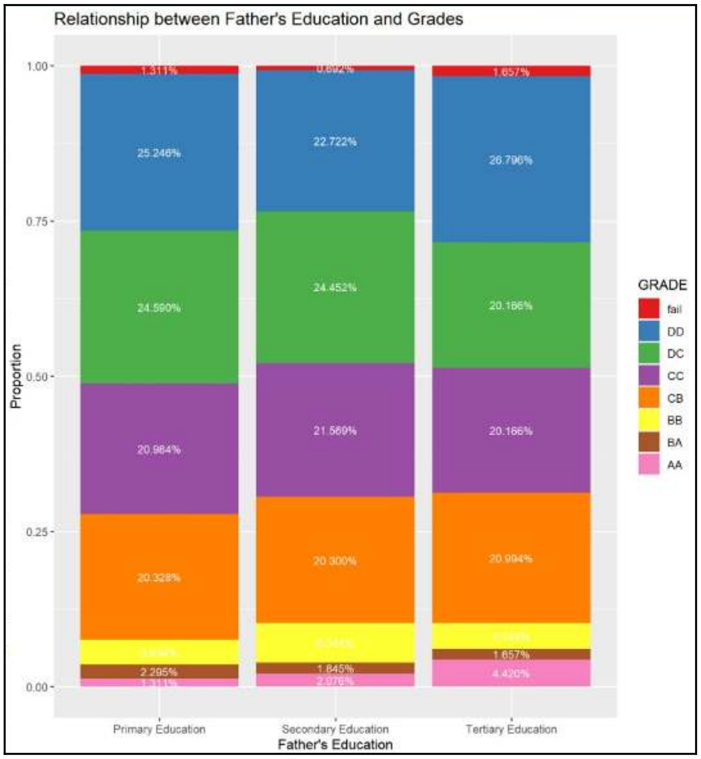


Figure 65 Bar chart visualization: Father Edu vs Grade

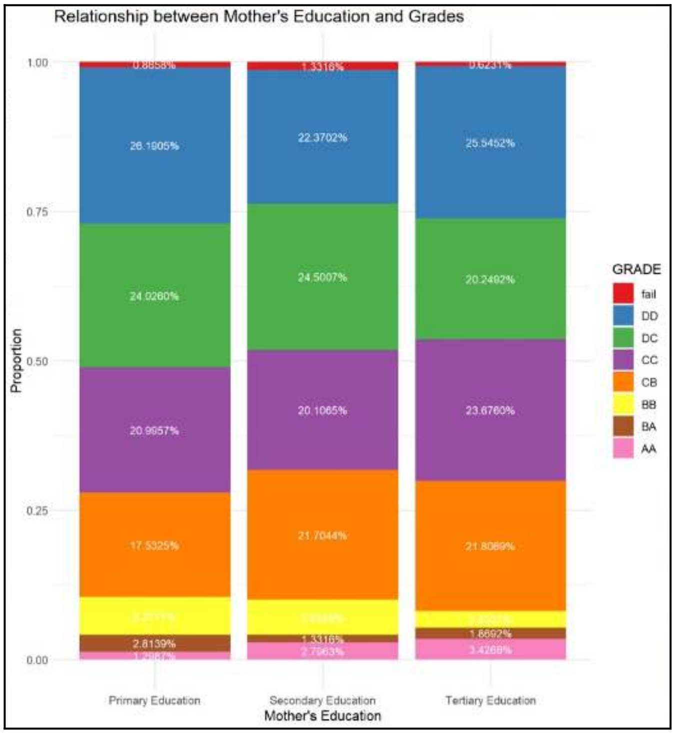


Figure 66 Bar chart visualization: Mother Edu vs Grade

Form the bar charts in (Figure 65&66), is plotted to show the proportions of the grades to the respective education categories. They show almost the same percentages across the three categories created, and there is no pattern in the proportions except for grade AA in both figures, they increase as the level of parents' education get higher. However, only this phenomenon with the small percentages jumps is not suggestive of a relationship between parents' level of education and academic achievement.

Conclusion for objective 3:

This was a thorough analysis done to examine the relationship between students' family background and academic performance, looking at the three different factors (number of siblings, parents' status, and parents' education level). It is concluded that: there is no relationship between students' family background and academic performance, because each factor of the student's family background on the analysis showed no relationship to the academic performance.

3.4 Objective 4 (Furas Sultan Ghaleb Mohammed, TP066989)

Investigating the Impact of Students’ Personal Life and Economic Circumstances on Academic Performance

This objective focuses on understanding how various aspects of students' personal life and their economics influence their academic performance. The research questions under this objective are designed to explore the relationship between personal life and economic factors and academic performance. The attributes used for this investigation include:

1- Total Salary Range (if available): Examining the impact of income levels on academic performance, reflecting financial security or challenges.

2- Partner Status: Investigating the effect of having a partner on grads.

3- Type of High School Graduation: Studying how different high school backgrounds (private, state, or other) affect current academic success.

3.4.1: Does the total salary range affect the student’s academic performance?

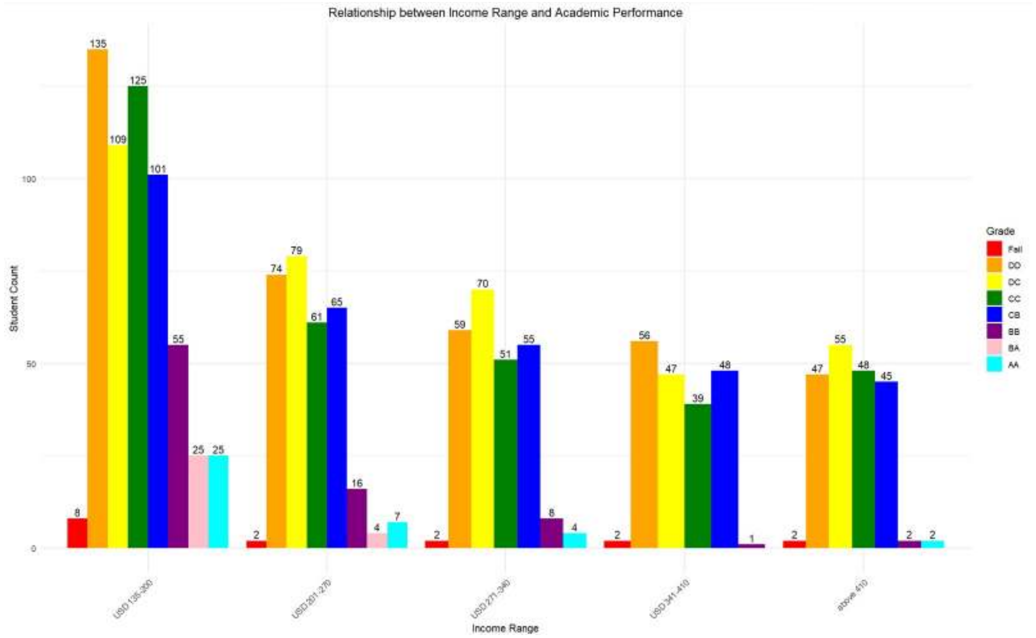
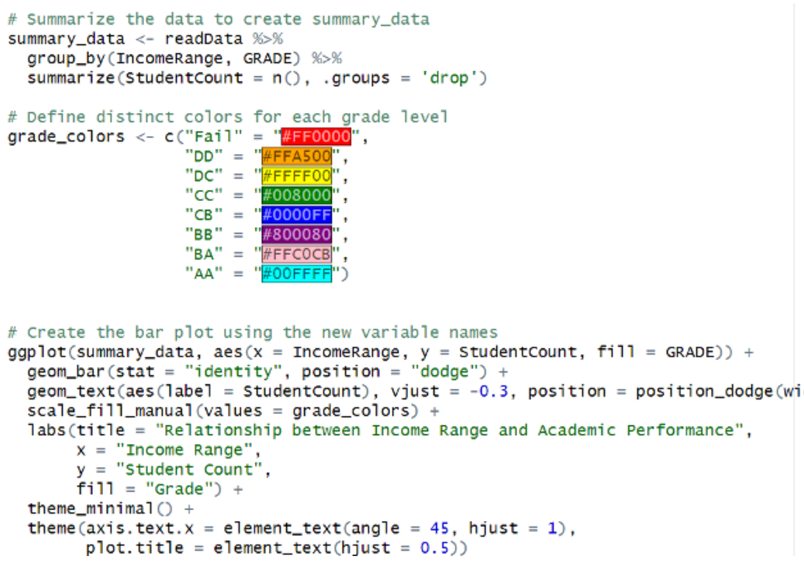


Figure 67 Bar Chart of Student Grades Across Different Income Ranges

The chart "Relationship between Income Range and Academic Performance" examines students' income (Salary) versus their grades. The "USD 135-200" income range category stands out in the bar chart as containing the majority of students who have achieved high grades. Despite being the lowest income bracket, it shows a significant count of students attaining the 'AA' grade, the highest academic performance level depicted. This suggests that while economic advantage may play a role in academic success, students in this income range are also reaching high levels of academic achievement. The substantial number of high achievers in this bracket could indicate that these students have access to quality education or possess a high level of academic resilience and dedication, regardless of their financial background.



```
# Summarize the data to create summary_data
summary_data <- readData %>%
  group_by(IncomeRange, GRADE) %>%
  summarize(StudentCount = n(), .groups = 'drop')

# Define distinct colors for each grade level
grade_colors <- c("Fail" = "#FF0000",
                  "DD" = "#FFA500",
                  "DC" = "#FFFF00",
                  "CC" = "#008000",
                  "CB" = "#0000FF",
                  "BB" = "#800080",
                  "BA" = "#FFC0CB",
                  "AA" = "#00FFFF")


# Create the bar plot using the new variable names
ggplot(summary_data, aes(x = IncomeRange, y = StudentCount, fill = GRADE)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = StudentCount), vjust = -0.3, position = position_dodge(width = 0.9), check_overlap = TRUE) +
  scale_fill_manual(values = grade_colors) +
  labs(title = "Relationship between Income Range and Academic Performance",
       x = "Income Range",
       y = "Student Count",
       fill = "Grade") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```

Figure 68: Bar Chart Generation Code Segment

3.4.2: Does the student without a partner status affect students' performance?

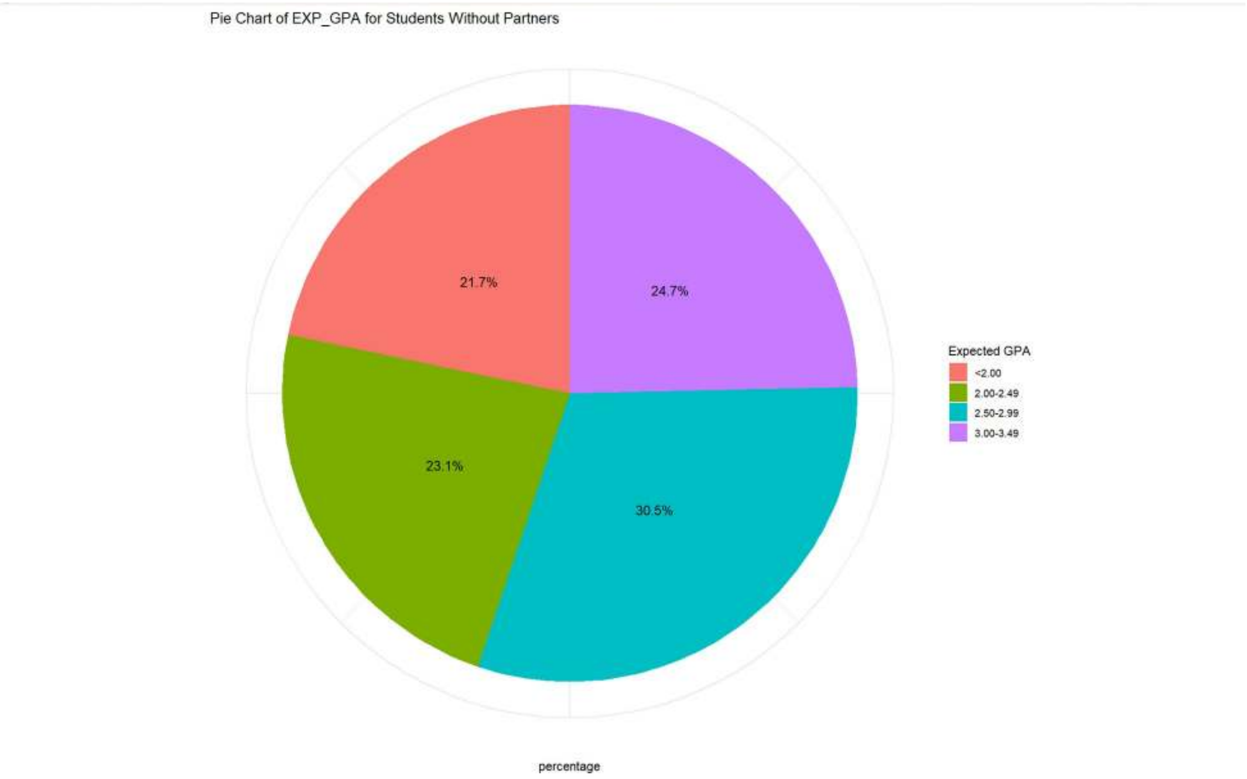
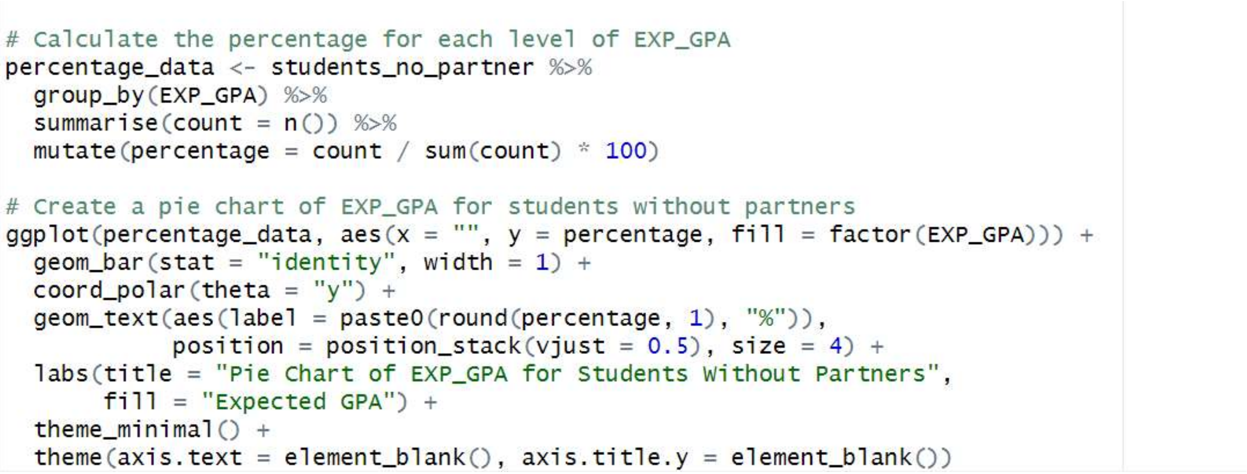


Figure 69: Pie Chart of Expected GPA for Single Students



```r
# Calculate the percentage for each level of EXP_GPA
percentage_data <- students_no_partner %>%
  group_by(EXP_GPA) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

# Create a pie chart of EXP_GPA for students without partners
ggplot(percentage_data, aes(x = "", y = percentage, fill = factor(EXP_GPA))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5), size = 4) +
  labs(title = "Pie Chart of EXP_GPA for Students Without Partners",
       fill = "Expected GPA") +
  theme_minimal() +
  theme(axis.text = element_blank(), axis.title.y = element_blank())
```

Figure 70: Pic Chart Generation Code Segment

The above screenshot is a pie chart titled "Pie Chart of EXP_GPA for Students Without Partners," which breaks down the expected GPA ranges for these students into four categories. The pie chart shows that the largest group of students without partners, 30.5%, is expected to have a GPA in the range of 2.50-2.99, followed by 24.7% expected to achieve a GPA between 3.00-3.49, 23.1% with a GPA between 2.00-2.49, and the smallest segment, 21.7%, expected to have a GPA below 2.00.

The segments are labeled with percentages that indicate the proportion of students in each GPA range:

- The red segment represents students with an expected GPA of less than 2.00, making up 21.7% of the total.

- The green segment corresponds to students with an expected GPA between 2.00 and 2.49, accounting for 23.1%.

- The blue segment denotes students with an expected GPA between 2.50 and 2.99, which is the largest group at 30.5%.

- The purple segment is for students with an expected GPA between 3.00 and 3.49, comprising 24.7%.

3.4.3: Does the student with a partner status affect students' performance?

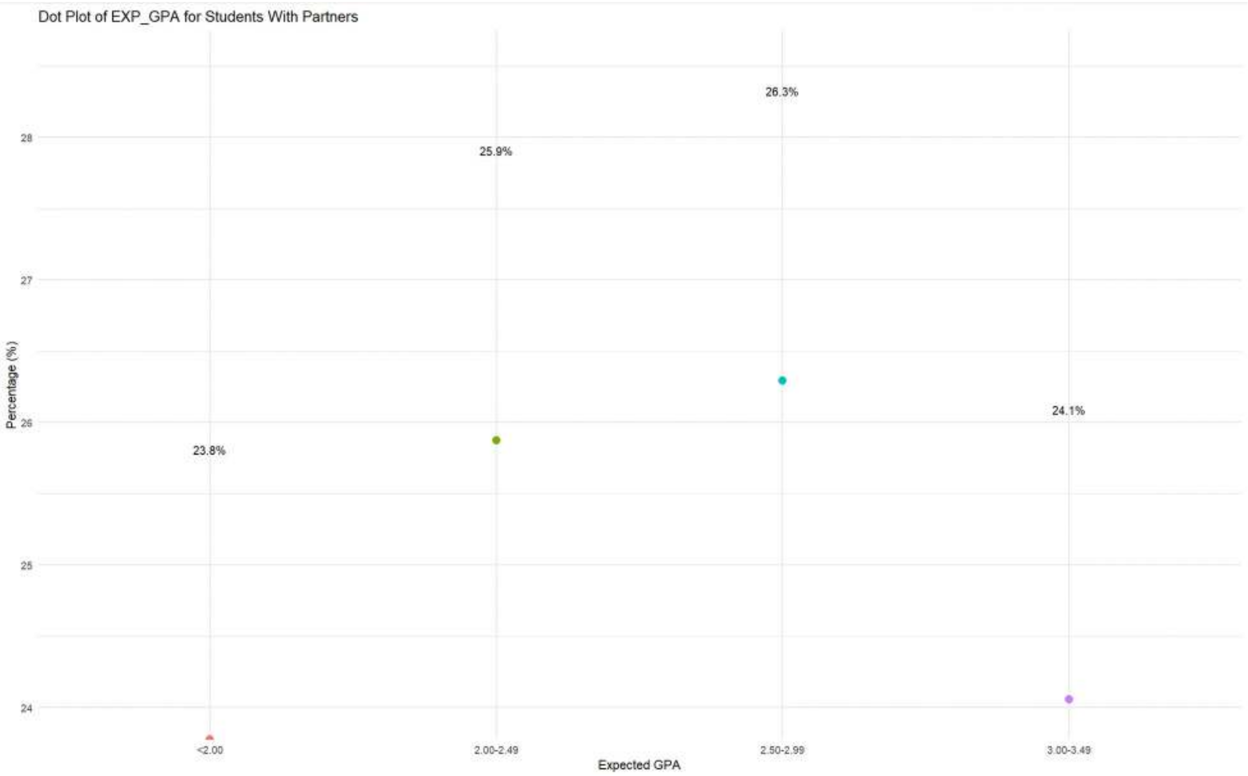
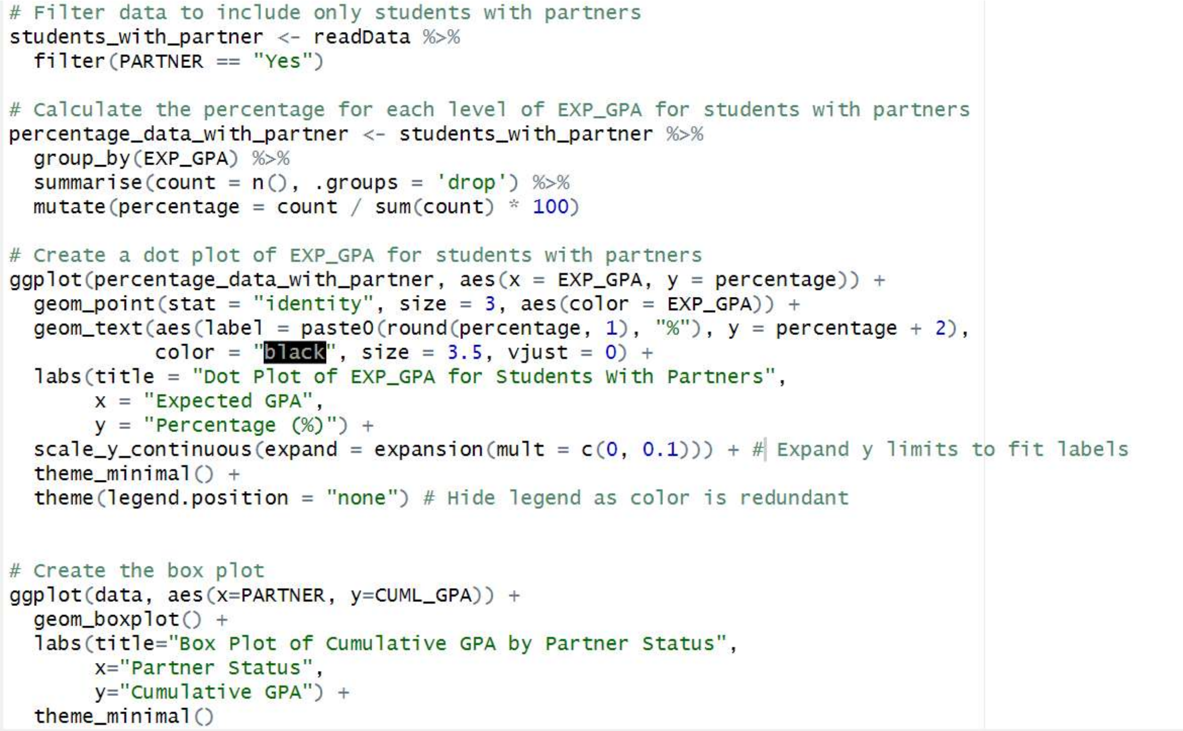


Figure 71: Dot Plot of Expected GPA for Students with Partners

The above graph displays a dot plot titled "Dot Plot of EXP_GPA for Students With Partners." This plot reveals the percentage of students with partners within the same GPA ranges. The distribution is fairly uniform across the different GPA categories, with 23.8% of students expected to have a GPA below 2.00, 25.9% within the 2.00-2.49 range, 26.3% in the 2.50-2.99 range, and 24.1% in the 3.00-3.49 range.

Upon comparing the two visual data representations, the pie chart of students without partners indicates a concentration in the mid-GPA ranges (2.50-2.99 and 3.00-3.49). In contrast, the dot plot for students with partners shows a more uniform distribution across the various GPA categories. This contrast may imply that students without partners are more frequently achieving GPAs in the moderate to high range.



```r
# Filter data to include only students with partners
students_with_partner <- readData %>%
  filter(PARTNER == "Yes")

# Calculate the percentage for each level of EXP_GPA for students with partners
percentage_data_with_partner <- students_with_partner %>%
  group_by(EXP_GPA) %>%
  summarise(count = n(), .groups = 'drop') %>%
  mutate(percentage = count / sum(count) * 100)

# Create a dot plot of EXP_GPA for students with partners
ggplot(percentage_data_with_partner, aes(x = EXP_GPA, y = percentage)) +
  geom_point(stat = "identity", size = 3, aes(color = EXP_GPA)) +
  geom_text(aes(label = paste0(round(percentage, 1), "%"), y = percentage + 2),
            color = "black", size = 3.5, vjust = 0) +
  labs(title = "Dot Plot of EXP_GPA for Students With Partners",
       x = "Expected GPA",
       y = "Percentage (%)") +
  scale_y_continuous(expand = expansion(mult = c(0, 0.1))) + # Expand y limits to fit labels
  theme_minimal() +
  theme(legend.position = "none") # Hide legend as color is redundant


# Create the box plot
ggplot(data, aes(x=PARTNER, y=CUML_GPA)) +
  geom_boxplot() +
  labs(title="Box Plot of Cumulative GPA by Partner Status",
       x="Partner Status",
       y="Cumulative GPA") +
  theme_minimal()
```

Figure 72: Coding Script for GPA Analysis by Partner Status

3.4.4: How does the type of high school graduation affect current academic performance?

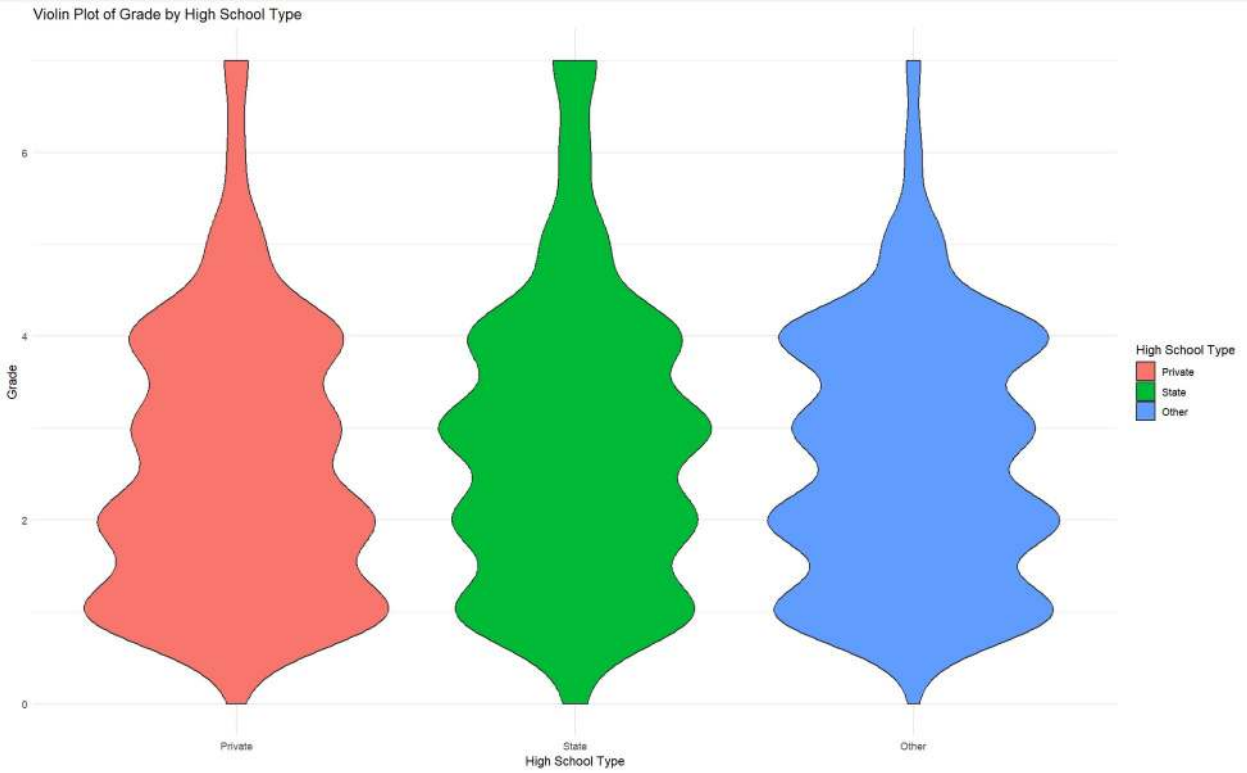
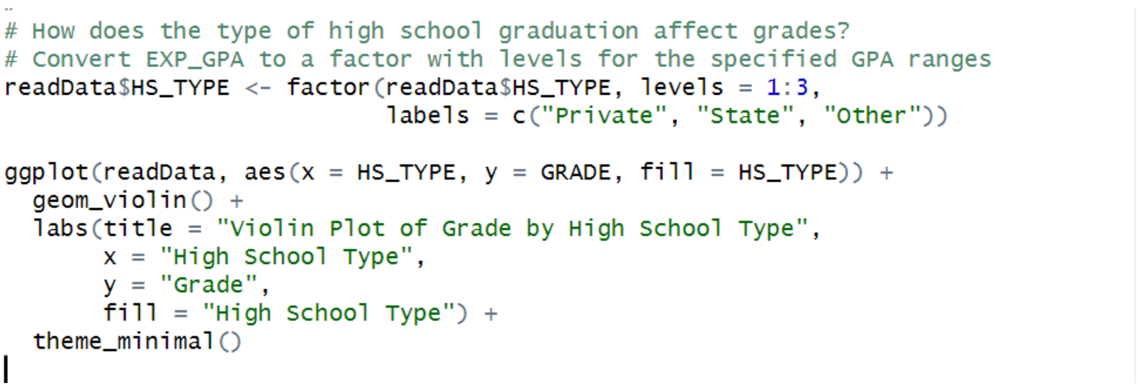


Figure 73: Violin Plot of Grades by High School Type



```
# How does the type of high school graduation affect grades?
# Convert EXP_GPA to a factor with levels for the specified GPA ranges
readData$HS_TYPE <- factor(readData$HS_TYPE, levels = 1:3,
                           labels = c("Private", "State", "Other"))

ggplot(readData, aes(x = HS_TYPE, y = GRADE, fill = HS_TYPE)) +
  geom_violin() +
  labs(title = "Violin Plot of Grade by High School Type",
       x = "High School Type",
       y = "Grade",
       fill = "High School Type") +
  theme_minimal()
```

Figure 74: High School Type Effect on Grades - Code Snippet

The violin plot presents a comparison of grade distributions across three different types of high schools: Private, State, and Other. In the plot, the Private high schools, represented by the red violin, show a broad distribution of grades with a concentration around grade 4, suggesting that most students from private institutions fall into this grade range, yet there's also a presence of higher achievers. The State high schools, depicted in green, exhibit a similar spread with a more pronounced focus in the mid-grade region, indicating a median level of academic performance among these students. The 'Other' high school category, shown in blue, has a more uniform and symmetric grade distribution with the bulk of students centered around the median grade. Notably, the 'Other' category also extends up to the highest grade, implying that while there may be less variability, students from this group can achieve the top grades. Overall, the plot reveals that while there are variations, all high school types have students achieving across the grade spectrum, with the possibility of reaching the highest academic performance levels.

Conclusion regards the Objective

This study sets out to investigate the effects of students' personal and financial circumstances on their academic achievement. The investigation concentrated on variables like partner status, type of high school graduation, and salary range. However, the results show that academic achievement is not greatly influenced by wage or the style of high school graduation. Although relationship status had some bearing, high school education and pay had less of an impact. This demonstrates the intricacy of variables influencing academic performance and raises the possibility that factors such as school background and personal financial situation may not have as much of an impact as previously believed.

4.0 Conclusion

In conclusion, the deep analysis undertaken to search for the impact of various factors on students' academic performance provided us with valuable insights from the data. We were looking for answers to our thorough questions to check whether our hypothesis was right. We divided the research into four (4) parts, so that each member of the team can focus on one objective. To close the curtains to our exploration, let us reflect on the results to our analysis.

There was a positive correlation between the level of students' engagement in academic sessions and academic performance, which highlights the importance of involvement and participation while learning. On the other hand, time granted for revision didn't have the same positive effect, it has no effect at all. This contests the notion that improving academic results is directly correlated with spending more time on revision.

In addition, family background may have very little effect on students' performance, but it is too small to notice and failed the Chi-testing almost every time. While economic circumstances of students didn't affect the academic performance, the partner status had a considerable impact on it.

Finally, academics success depends on many complex factors, which might be hard to analysis. However, knowing them will help in building effective strategies to help students in their academic journey. It is recommended that further investigation and study be done in these fields to improve our comprehension and guide evidence-based teaching methods.

5.0 Workload Matrix

	Abdulrahman Gamil Murshed Humadi (TP070609)	Furas Sultan Ghaleb Mohammed (TP066989)	Lee Wen Jing (TP071630)	Lee Xin Yee (TP070654)
1.0 Introduction	25%	25%	25%	25%
2.0 Data Preparation	25%	25%	25%	25%
3.0 Data Analysis	25%	25%	25%	25%
4.0 Conclusion	25%	25%	25%	25%
6.0 References	25%	25%	25%	25%
Documentation	25%	25%	25%	25%
Signature	Abdulrahman	Feras	WenJing	Xinyee

6.0 References

Amadebai, E. (2021, March 20). *The Importance Of Data Cleaning In Analytics Explained*. AnalyticsForDecisions. https://www.analyticsfordecisions.com/importance-of-data-cleaning/?expand_article=1

Yi, M. (2019, August 29). *A complete guide to pie charts*. Chartio. https://chartio.com/learn/charts/pie-chart-complete-guide/

(Yi, 2019)

R Bar Plot (With examples). (n.d.). https://www.datamentor.io/r-programming/bar-plot

(*R Bar Plot (With Examples)*, n.d.)

Matplotlib - Bar plot. (n.d.). https://www.tutorialspoint.com/matplotlib/matplotlib_bar_plot.htm

(*Matplotlib - Bar Plot*, n.d.)

All Graphics in R (Gallery) | Plot, Graph, Chart, Diagram, figure examples. (2023, August 1). Statistics Globe. https://statisticsglobe.com/graphics-in-r

(*All Graphics in R (Gallery) | Plot, Graph, Chart, Diagram, Figure Examples*, 2023)

Radečić, D. (2022, September 7). *Data cleaning in R: 2 R packages to clean and validate datasets*. Copyrights © appsilon.com. All Rights Reserved. https://appsilon.com/data-cleaning-in-r/

(Radečić, 2022)

Mulani, S. (2022, August 3). *Covariance and Correlation in R programming*. DigitalOcean. https://www.digitalocean.com/community/tutorials/covariance-and-correlation-in-r-programming

(Mulani, 2022)

Higher Education Students Performance Evaluation. (2021, December 21). Kaggle. https://www.kaggle.com/datasets/csafrit2/higher-education-students-performance-evaluation/data

(Higher Education Students Performance Evaluation, 2021)

7.0 Appendix:

The code is submitted with this word file in moodle.