



## GROUP ASSIGNMENT

CT127-3-1-PFDA

PROGRAMMING FOR DATA ANALYSIS

APD2F2309CS(CYB)

HAND OUT DATE : WEEK 3

HAND IN DATE : WEEK 10

LECTURER : MINNU HELEN JOSEPH

Name	TP Number
MUHAMAD AHMAD AL MUHDAR	TP070208
ABDULELAH HUSSEIN ABDULRAHMAN AL-KAF	TP069319
ABDULRAHMAN GAMIL MOHAMMED AHMED	TP071012
IBRAHEEM MOHAMMED IMADELDIN AWAD	TP070765

## TABLE OF CONTENTS

1.0 INTRODUCTION.....	5
1.1 Assumption.....	5
2.0 DATA PREPARATION.....	6
2.1 DATA IMPORT.....	6
2.2 DATA CLEANING.....	6
2.3 PRE-PROESSING.....	7
2.4 DATA EXPLORATION. ....	8
3.0 DATA ANAYSIS. ....	9
3.1 Question 1: What is the impact of reading scientific books on the final grade? (ABDULRAHMAN GAMIL MOHAMMED AHMED TP071012) .....	9
Analysis 3.1.1: Patterns and Proportions in Student Reading Choices. ....	11
Analysis 3.1.2: Ratio of Grades Based on the Numbers of the Students. ....	13
Analysis 3.1.3: Trends and Ratios in Students Grades of Reading Selections. ....	14
Analysis 3.1.4: The Impact of Reading on Grades. ....	15
Analysis 3.1.5: Chi Square Test to Examine the Connection Between Reading Scientific and Getting Better Marks.....	18
Analysis 3.1.6: Conclusion for Question 1. ....	19
3.1.7 Additional Features.....	19
3.2 Question 2: What is the impact of regular preparation for midterms on the Final Grade? (IBRAHEEM MOHAMMED IMADELDIN AWAD_TP070765_Cyber Security).....	22
Analysis 3.2.1: Types of preparation done by the students for the exam and their impact. (Without applying filters to the graphs and data).....	23
Analysis 3.2.2: Impact of regular preparation for midterms during the semester on final grade. ....	26

Analysis 3.2.3: Impact of external factors as well as regular preparation for midterms on final grades (Multivariate Analysis).....	29
Analysis 3.2.4: Chi Square Test for testing the relationship between regular preparation and achieving “AA” and “BA” grades.....	32
Analysis 3.2.5: Conclusion for Question 2.....	33
3.2.6 Additional Features.....	33
3.3 Question 3: What is the impact of actively taking notes on achieving higher grades? (ABDULELAH HUSSEIN ABDULRAHMAN AL-KAF   TP069319   Cyber Security) .....	35
Analysis 3.3.1: Impact of taking notes with grade.....	35
Analysis 3.3.2: Impact of taking notes with high grade .....	36
Analysis 3.3.3: Impact of taking notes and gender with high grade .....	38
Analysis 3.3.4: Impact of taking notes, gender and high school type with high grade .....	40
Analysis 3.3.5: Impact of taking notes, gender, high school and attendance with high grade .....	43
TESTING :.....	46
CONCLUSION OF QUESTION 3 .....	47
Extra features : .....	47
Question 4: the impact of students studying between 11 and 20 hours per week on achieving higher grade (MUHAMAD AHMAD AL MUHDAR TP070208) .....	48
Analysis 3.4.1: impact of studying 11 to 20 hours with grade .....	48
Analysis 3.4.2: impact of studying 11 to 20 hours with high grades .....	49
Analysis 3.4.3: impact of studying 11 to 20 hours and living with high grade .....	50
Analysis 3.4.4: impact of studying 11 to 20 hours, scholarship and HS type with high grade. ....	51
TESTING:.....	52
Conclusion of question 4: .....	53

4.0 CONCLUSION.....	53
EXTRA FEATURES: .....	53
5.0 WORKLOAD MATRIX. ....	55
6.0 REFERENCES.....	55

## **1.0 INTRODUCTION.**

RStudio is an integrated development environment for R, a programming language for statistical computing and it will be the program used to prove the hypothesis selected. To prove the hypothesis/assumption various statistical analysis methods will be used to analyze the information provided in the excel file named “student\_prediction” where a set of information regarding students and various factors affecting their final grades is available.

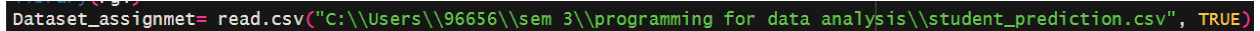
The goal of the investigation conducted is to identify based on an assumption, the relationship between 4 factors where the effect of each individual factor is studied in detail in regard to students achieving a high final grade which in this case is either an “AA” or a “BA” grade.

### **1.1 Assumption**

The hypothesis or assumption made indicates that 60% of students who read scientific books often, prepare regularly for midterms during the semester, always actively taking notes, and finally having study hours between 11 and 20 hours per week, end up achieving a high output grade which is either an “AA” grade or a “BA” final grade.

## 2.0 DATA PREPARATION.

### 2.1 DATA IMPORT.

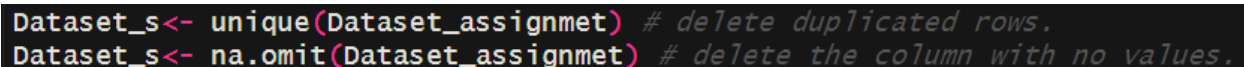


```
Dataset_assignmet= read.csv("C:\\Users\\96656\\sem 3\\programming for data analysis\\student_prediction.csv", TRUE)
```

Figure 1

In the screenshot provided it's clear that we used the `read` function to access a file with the ".csv" extension. To import our dataset, we utilized the `read.csv` function by indicating the file path (dataset), within quotation marks. After specifying the file path there's an option to input either TRUE or FALSE. If set to TRUE it assumes that the dataset already contains column names whereas setting it to FALSE instructs the function to use the first row of data as column names.

### 2.2 DATA CLEANING.



```
Dataset_s<- unique(Dataset_assignmet) # delete duplicated rows.  
Dataset_s<- na.omit(Dataset_assignmet) # delete the column with no values.
```

Figure 2

During the initial phase of data cleaning, we carried out two essential tasks to enhance the quality and redundancy of our dataset, which we have named {Dataset\_assignment}. First, we made a new dataset called "Dataset\_s" by removing any duplicate items using the "unique ()" method. By ensuring each item's uniqueness using this function, duplication and potential inaccuracies in subsequent analysis are avoided. In order to resolve the missing values in the dataset, we secondly generated a new dataset called "Dataset\_s" by using the "na.omit()" technique. By removing any rows with missing values, we were able to produce a dataset that was more thorough and trustworthy. By strengthening the dataset's dependability throughout the assignment, these preparatory actions facilitate the performance of trustworthy statistical analysis and visualizations.

## 2.3 PRE-PROESSING.

```

Dataset_s<- names(Dataset_assignmet) #names of the columns.
class(Dataset_assignmet) # show the datatype of Dataset_assignmet
# Rename the "kids" column to "Parental status" using names
colnames(Dataset_assignmet)[15] = "Parental_status"
# change the values:
Dataset_assignmet$GENDER <- factor(Dataset_assignmet$GENDER, levels=c(1,2), labels=c("female", "male"))
Dataset_assignmet$WORK <- factor(Dataset_assignmet$WORK, levels=c(1,2), labels=c("YES", "NO"))
Dataset_assignmet$ACTIVITY <- factor(Dataset_assignmet$ACTIVITY, levels=c(1,2), labels=c("YES", "NO"))
Dataset_assignmet$SALARY <- factor(Dataset_assignmet$SALARY, levels=c(1,2,3,4,5), labels=
c("135$ - 200$", "201$ - 270$", "271$ - 340$", "341$ - 410$", " >= 410$"))
Dataset_assignmet$TRANSPORT <- factor(Dataset_assignmet$TRANSPORT, levels=c(1,2,3,4), labels=c("BUS", "PRIVATE CAR/TAXI", "BICYCLE", "OTHER"))
Dataset_assignmet$GRADE <- factor(Dataset_assignmet$GRADE, levels=c(0,1,2,3,4,5,6,7), labels=c("FAIL", "DD", "DC", "CC", "CB", "BB", "BA", "AA"))
Dataset_assignmet$AGE <- factor(Dataset_assignmet$AGE, levels=c(1,2,3), labels=c("18-21", "22-25", " >=26"))
Dataset_assignmet$HS_TYPE <- factor(Dataset_assignmet$HS_TYPE, levels=c(1,2,3), labels=c("PRIVATE", "STATE", "OTHER"))
Dataset_assignmet$SCHOLARSHIP <- factor(Dataset_assignmet$SCHOLARSHIP, levels=c(1,2,3,4,5), labels=c("0%", "25%", "50%", "75%", "100%"))
Dataset_assignmet$PARTNER <- factor(Dataset_assignmet$PARTNER, levels=c(1,2), labels=c("YES", "NO"))
Dataset_assignmet$LIVING <- factor(Dataset_assignmet$LIVING, levels=c(1,2,3,4), labels=c("RENTAL", "DEORMITORY", "WITH FAMILY", "OTHER"))
Dataset_assignmet$MOTHER_EDU <- factor(Dataset_assignmet$MOTHER_EDU, levels=c(1,2,3,4,5,6), labels=c
("PRIMARY SCHOOL", "SECONDARY SCHOOL", "HIGH SCHOOL", "UNIVERSITY", "MSc.", "Bh.d."))
Dataset_assignmet$FATHER_EDU <- factor(Dataset_assignmet$FATHER_EDU, levels=c(1,2,3,4,5,6), labels=
c("PRIMARY SCHOOL", "SECONDARY SCHOOL", "HIGH SCHOOL", "UNIVERSITY", "MSc.", "Bh.d."))
Dataset_assignmet$X_SIBLINGS <- factor(Dataset_assignmet$X_SIBLINGS, levels=c(1,2,3,4,5), labels=c("1", "2", "3", "4", ">=5"))
Dataset_assignmet$Parental_status <- factor(Dataset_assignmet$Parental_status, levels=c(1,2,3,4), labels=
c("MARRIED", "DIVORCED", "DIED", "ONE OF THEM OR BOTH"))
Dataset_assignmet$MOTHER_JOB <- factor(Dataset_assignmet$MOTHER_JOB, levels=c(1,2,3,4,5,6), labels=
c("RETIRED", "HOUSEWIFE", "GOVERNMENT OFFICER", "PRIVATE SECTOR EMPLOYEE", "SELF-EMPLOYMENT", "OTHER"))
Dataset_assignmet$FATHER_JOB <- factor(Dataset_assignmet$FATHER_JOB, levels=c(1,2,3,4,5), labels=
c("RETIRED", "HOUSEWIFE", "GOVERNMENT OFFICER", "PRIVATE SECTOR EMPLOYEE", "OTHER"))
Dataset_assignmet$STUDY_HRS <- factor(Dataset_assignmet$STUDY_HRS, levels=c(1,2,3,4,5), labels=
c("NONE", "<5 HOURS", "6-10 HOURS", "11-20 HOURS", ">20"))
Dataset_assignmet$READ_FREQ <- factor(Dataset_assignmet$READ_FREQ, levels=c(1,2,3), labels=c("NONE", "SOMETIMES", "OFTEN"))
Dataset_assignmet$READ_FREQ_SCI <- factor(Dataset_assignmet$READ_FREQ_SCI, levels=c(1,2,3), labels=c("NONE", "SOMETIMES", "OFTEN"))
Dataset_assignmet$ATTEND_DEPT <- factor(Dataset_assignmet$ATTEND_DEPT, levels=c(1,2), labels=c("YES", "NO"))
Dataset_assignmet$IMPACT <- factor(Dataset_assignmet$IMPACT, levels=c(1,2,3), labels=c("POSITIVE", "NEGATIVE", "NEUTRAL"))

Dataset_assignmet$ATTEND <- factor(Dataset_assignmet$ATTEND, levels=c(1,2,3), labels=c("ALWAYS", "SOMETIMES", "NEVER"))
Dataset_assignmet$PREP_STUDY <- factor(Dataset_assignmet$PREP_STUDY, levels=c(1,2,3), labels=
c("ALONE", "WITH FRIENDS", "NOT APPLICABLE"))
Dataset_assignmet$PREP_EXAM <- factor(Dataset_assignmet$PREP_EXAM, levels=c(1,2,3), labels=
c("CLOSES DATE TO THE EXAM", "REGULARLY DURING THE SEM", "NEVER"))
Dataset_assignmet$NOTES <- factor(Dataset_assignmet$NOTES, levels=c(1,2,3), labels=c("NEVER", "SOMETIMES", "ALWAYS"))
Dataset_assignmet$LISTENS <- factor(Dataset_assignmet$LISTENS, levels=c(1,2,3), labels=c("NEVER", "SOMETIMES", "ALWAYS"))
Dataset_assignmet$LIKES_DISCUSS <- factor(Dataset_assignmet$LIKES_DISCUSS, levels=c(1,2,3), labels=c("NEVER", "SOMETIMES", "ALWAYS"))
Dataset_assignmet$CLASSROOM <- factor(Dataset_assignmet$CLASSROOM, levels=c(1,2,3), labels=c("NOT USEFUL", "USEFUL", "NOT APPLICABLE"))
Dataset_assignmet$CUMM_GPA <- factor(Dataset_assignmet$CUMM_GPA, levels=c(1,2,3,4,5), labels=
c("<2.00", "2.00-2.49", "2.50-2.99", "3.00-3.49", ">3.49"))
Dataset_assignmet$EXP_GPA <- factor(Dataset_assignmet$EXP_GPA, levels=c(1,2,3,4,5), labels=
c("<2.00", "2.00-2.49", "2.50-2.99", "3.00-3.49", ">3.49"))

```

Figure 3

Several steps were taken to enhance the structure and readability of the "Dataset\_assignmet data" set in the code area above. The column names were first extracted, and a new variable named "Dataset\_s" was established. The `class()` function was then used to display the datatype of the "Dataset\_assignmet" dataset, providing information regarding the dataset's underlying data structure.

Furthermore, we used the "colnames()" method to rename a certain column from "kids" to "Parental status". This step enhances the clarity and understanding of the column's content. Next, in order to obtain a standardized representation, we used the "factor()" function to recode the categorical variables in the dataset. Categories including gender, work status, activity participation, wage range, and others were recoded for consistency and ease of analysis.

## 2.4 DATA EXPLORATION.

```
ncol(Dataset_assignmet) # Get the number of columns in the dataset
nrow(Dataset_assignmet) # Get the number of rows in the dataset
head(Dataset_assignmet) # Display the top 6 rows of the dataset
head(Dataset_assignmet, 10) # Display the top 10 rows of the dataset
tail(Dataset_assignmet) # Display the last 6 rows of the dataset
tail(Dataset_assignmet, 10) # Display the last 10 rows of the dataset
names(Dataset_assignmet) # Get the column names of the dataset
summary(Dataset_assignmet) # Display a summary of the dataset (including min, Q1, median, mean, Q3, max)
min(Dataset_assignmet$AGE) # Get the minimum value in the "AGE" column
max(Dataset_assignmet$AGE) # Get the maximum value in the "AGE" column
mean(Dataset_assignmet$AGE) # Get the mean (average) value in the "AGE" column
```

```
DATA = Dataset_assignmet[, c(20, 25, 26, 18, 33)]
```

Figure 4

In this code section, we thoroughly explored the "Dataset\_assignmet" dataset. We obtained the number of columns and rows, displayed initial and concluding rows, and identified column names. The `summary()` function provided key statistics for each variable. For example, we specifically for "AGE," we extracted its minimum, maximum, and mean values. These operations collectively establish a foundational understanding of the dataset, paving the way for more in-depth analyses, and then create a dataset obtained from the main dataset to specify the only columns needed to do the analysis.



### 3.0 DATA ANALYSIS.

#### 3.1 Question 1: What is the impact of reading scientific books on the final grade? (ABDULRAHMAN GAMIL MOHAMMED AHMED TP071012) | Cyber Security

```
install.packages("plotrix")
library(plotrix)
install.packages("ggplot2")
library(ggplot2)
install.packages("tidyr")
library(tidyr)
install.packages("rgl")
library(rgl)
```

Figure 5

In the screenshot provided we observed the packages that will be utilized to enhance our analysis process. To install these packages we use the "install.packages" command. Specify the name of each package. We have obtained four packages. Activate them by using the "library" command followed by their names.

```
# Obj1: The impact of reading scientific books on achieving higher grade
nrow(DATA)      # Number of rows in the new dataset (1534)
ncol(DATA)      # Number of columns in the new dataset (5)

# Check unique values in the GRADE and READ_FREQ columns
unique(DATA$GRADE)      # Levels: FAIL DD DC CC CB BB BA AA
unique(DATA$READ_FREQ_SCI) # Levels: NONE SOMETIMES OFTEN
```

Figure 6

Once we finish cleaning and preparing the data our analysis begins by examining and validating our goals. The first objective we're investigating is whether reading books has an impact, on achieving grades. To gain insights into the datasets size we use the function to count its rows while the ncol function provides an overview of its structure by counting the columns. Next we apply the function to identify values in relevant columns like "GRADE" and "READ\_FREQ\_SCI." This

focused use of the function allows us to examine and confirm the values associated with grades and frequency of scientific book reading. This systematic exploration serves as a step in supporting or challenging the validity of our objective offering insights into how grades and reading frequencies are distributed within the dataset.

```
# Calculate the number of students in different combinations of READ_FREQ_SCI and GRADE
SOMETIMES_AA = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "AA",])
SOMETIMES_BA = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "BA",])
SOMETIMES_BB = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "BB",])
SOMETIMES_CB = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "CB",])
SOMETIMES_CC = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "CC",])
SOMETIMES_DC = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "DC",])
SOMETIMES_DD = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "DD",])
SOMETIMES_FAIL = nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES" & New_Sample_Data$GRADE == "FAIL",])

OFTEN_AA = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "AA",])
OFTEN_BA = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "BA",])
OFTEN_BB = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "BB",])
OFTEN_CB = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "CB",])
OFTEN_CC = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "CC",])
OFTEN_DC = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "DC",])
OFTEN_DD = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "DD",])
OFTEN_FAIL = nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN" & New_Sample_Data$GRADE == "FAIL",])

NEVER_AA = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "AA",])
NEVER_BA = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "BA",])
NEVER_BB = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "BB",])
NEVER_CB = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "CB",])
NEVER_CC = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "CC",])
NEVER_DC = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "DC",])
NEVER_DD = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "DD",])
NEVER_FAIL = nrow(DATA[DATA$READ_FREQ_SCI == "NONE" & New_Sample_Data$GRADE == "FAIL",])
```

Figure 7

In the illustration the provided code takes an approach by considering all possible scenarios that could affect the subsequent analysis. Each classification related to students reading habits whether its 'sometimes' 'often,' or 'never' is systematically paired with every grade value, including 'AA' 'BA,' 'BB,' 'CB,' 'CC,' 'DC' 'DD,' and 'FAIL.' This careful pairing results, in the creation of 24 variables, each representing a combination of reading frequencies and grades. As a result these variables encompass the possibilities in the dataset enabling a nuanced exploration of the connections between different reading habits and corresponding academic grades. By considering all combinations this method ensures a comprehensive and detailed examination that captures the intricate relationship, between students reading behaviors and their academic performance.

### Analysis 3.1.1: Patterns and Proportions in Student Reading Choices.

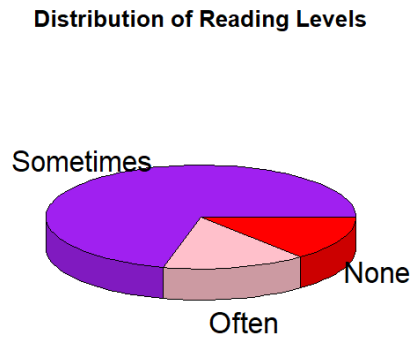


Figure 8

When we examine a 3D pie chart that shows how students distribute themselves based on their reading frequency of books or journals an interesting trend emerges. The largest group consists of students who read sometimes making up, than 70% of the total. This suggests that there is an inclination among students to engage in reading of scientific literature. On the hand those who read often make up a proportion around 15% while approximately 14% of students never engage in reading. The pie chart visually represents these proportions by allocating the segment to those who read sometimes the segment to those who read often and the red segment, to those who never read.

```
SOMETIMES = c(SOMETIMES_AA, SOMETIMES_BA, SOMETIMES_BB, SOMETIMES_CB, SOMETIMES_CC, SOMETIMES_DC, SOMETIMES_DD, SOMETIMES_FAIL)
OFTEN = c(OFTEN_AA, OFTEN_BA, OFTEN_BB, OFTEN_CB, OFTEN_CC, OFTEN_DC, OFTEN_DD, OFTEN_FAIL)
NONE = c(NEVER_AA, NEVER_BA, NEVER_BB, NEVER_CB, NEVER_CC, NEVER_DC, NEVER_DD, NEVER_FAIL)
ALL = c(sum(SOMETIMES), sum(OFTEN), sum(NONE))
# Define labels
LEVELS <- c("Sometimes", "Often", "None")
grades<- c("FAIL", "DD", "DC", "CC", "CB", "BB", "BA", "AA")
# Create 3D pie chart for levels of students
pie3D(ALL, labels = LEVELS, radius = 1, main = "Distribution of Reading Levels", col = c("purple", "pink", "red"))
```

Figure 9

In the code snippet provided we can see how a 3D pie chart is created to show the distribution of individuals, across reading levels; sometimes, often, and never. This visual representation effectively showcases the proportions of people with varying reading habits. It offers an understanding of how people in the analyzed dataset engage in reading.

```
# Count the number of rows for different READ_FREQ_SCI values
sometimes_count <- nrow(DATA[DATA$READ_FREQ_SCI == "SOMETIMES",]) # 1090
often_count <- nrow(DATA[DATA$READ_FREQ_SCI == "OFTEN",]) # 231
none_count <- nrow(DATA[DATA$READ_FREQ_SCI == "NONE",]) # 213

# Total number of students
total_students <- sometimes_count + often_count + none_count

# Calculate percentages
percentage_sometimes <- (sometimes_count / total_students) * 100
percentageOften <- (often_count / total_students) * 100
percentageNone <- (none_count / total_students) * 100
```

Figure (10)

```
> percentageNone
[1] 13.88527
> percentageOften
[1] 15.05867
> percentage_sometimes
[1] 71.05606
```

Figure 11

The visual representation aligns, with the calculations shown above two picsc. In these figures the code systematically computes the percentages for each reading level (never. This process involves tallying the occurrences of each level and then multiplying by 100 to present the distribution in an easily comprehensible way.

### Analysis 3.1.2: Ratio of Grades Based on the Numbers of the Students.

```
# Count the number of rows for different GRADE values
aa_count=nrow(DATA[DATA$GRADE == "AA",])
ba_count=nrow(DATA[DATA$GRADE == "BA",])
bb_count=nrow(DATA[DATA$GRADE == "BB",])
cb_count=nrow(DATA[DATA$GRADE == "CB",])
cc_count=nrow(DATA[DATA$GRADE == "CC",])
dc_count=nrow(DATA[DATA$GRADE == "DC",])
dd_count=nrow(DATA[DATA$GRADE == "DD",])
fail_count=nrow(DATA[DATA$GRADE == "FAIL",]) |
# Total number of students
total_students_grades <- aa_count + ba_count + bb_count + cb_count + cc_count + dc_count + dd_count + fail_count

# Calculate percentages
percentage_aa <- (aa_count / total_students_grades) * 100
percentage_ba <- (ba_count / total_students_grades) * 100
percentage_bb <- (bb_count / total_students_grades) * 100
percentage_cb <- (cb_count / total_students_grades) * 100
percentage_cc <- (cc_count / total_students_grades) * 100
percentage_dc <- (dc_count / total_students_grades) * 100
percentage_dd <- (dd_count / total_students_grades) * 100
percentage_fail <- (fail_count / total_students_grades) * 100

# Create pie chart for levels of grades
colors = c("red", "#4169E1", "pink", "#DC143C", "#DAA520", "#800080", "#008080", "#708090")
pie(ALL_grades, labels = grades, explode = 0.1, main = "Distribution of Grades", col = colors)
```

Figure 12

```
> percentage_aa
[1] 2.477184
> percentage_ba
[1] 1.890482
> percentage_bb
[1] 5.345502
> percentage_cb
[1] 20.46936
> percentage_cc
[1] 21.12125
> percentage_dc
[1] 23.46806
> percentage_dd
[1] 24.18514
> percentage_fail
[1] 1.043025
```

Figure 13

Distribution of Grades

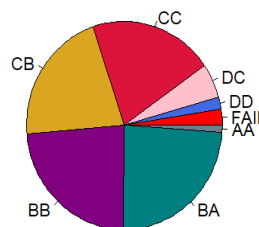


Figure 14

In the pie chart that follows it visually displays how grades are distributed among all students. Additionally, there is a code provided to calculate the percentages for each grade category. Specifically, the percentages for grades AA, BA, BB, CB, CC, DC, DD, and FAIL are 2.47%, 1.89%, 5.34%, 20.46%, 21.12%, 23.46%, 24.18% and 1.04%. Each grade is represented by a color; "" "#4169E1," "pink," "#DC143C," "#DAA520 " "#800080 " "#008080 " and "#708090." This pie chart and accompanying code provide a representation of how grades are distributed among the student population. To include a title in the diagram we used the "parameter," within the "pie" function.

### Analysis 3.1.3: Trends and Ratios in Students Grades of Reading Selections.

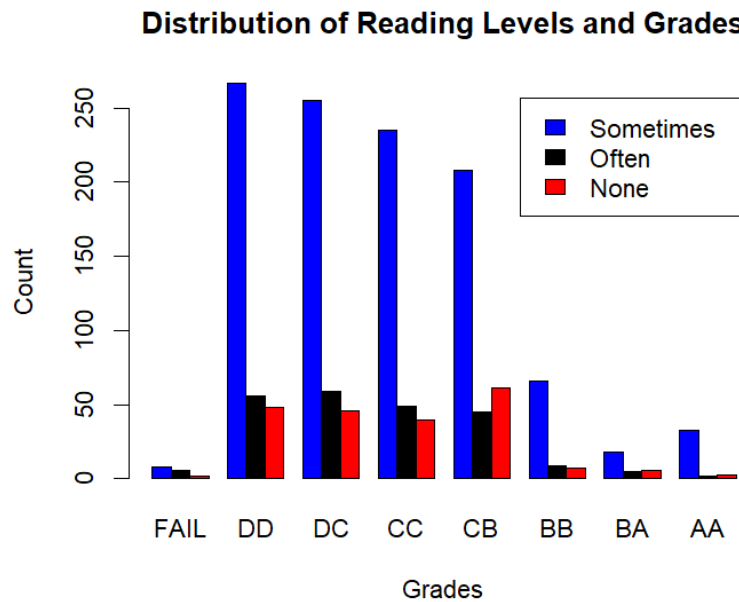


Figure 15

In this stacked bar chart, we can see how reading levels and grades are represented together. The blue bar shows students who read occasionally, the red bar represents those who never read, and the black bar represents students who frequently engage with books. The horizontal X axis shows the eighth grades while the vertical Y axis shows the number of students for each grade level. This visual depiction effectively captures how reading habits and academic performance relate across grade categories.

```
DisOften <- c(OFTEN_FAIL, OFTEN_DD, OFTEN_DC, OFTEN_CC, OFTEN_CB, OFTEN_BB, OFTEN_BA, OFTEN_AA)
DisSome <- c(SOMETIMES_FAIL, SOMETIMES_DD, SOMETIMES_DC, SOMETIMES_CC, SOMETIMES_CB, SOMETIMES_BB, SOMETIMES_BA, SOMETIMES_AA)
DisNone <- c(NEVER_FAIL, NEVER_DD, NEVER_DC, NEVER_CC, NEVER_CB, NEVER_BB, NEVER_BA, NEVER_AA)

# Create a stacked bar plot based on reading levels and grades
barplot(rbind(DisSome, DisOften, DisNone), beside = TRUE, col = c("blue", "black", "red"),
        legend.text = c("Sometimes", "Often", "None"),
        names.arg = c("FAIL", "DD", "DC", "CC", "CB", "BB", "BA", "AA"),
        xlab = "Grades", ylab = "Count", main = "Distribution of Reading Levels and Grades")
```

Figure 16

The code provided below generates the diagram above using three vectors: `DisSome` `DisOften` and `DisNone`. These vectors represent the number of students who read sometimes often and never, for each grade level. The stacked bar plot is created by combining these vectors using the

`rbind` function. The bars in the plot are colored blue, black and red to represent "" and "None" reading levels respectively. A legend is included to label each reading level while the x axis represents eight grades (FAIL, DD, DC, CC, CB, BB, BA, AA). The y axis indicates the student count. The plots title is set as "Distribution of Reading Levels and Grades " providing a representation of how students are distributed across different reading habits and grades.

#### Analysis 3.1.4: The Impact of Reading on Grades.

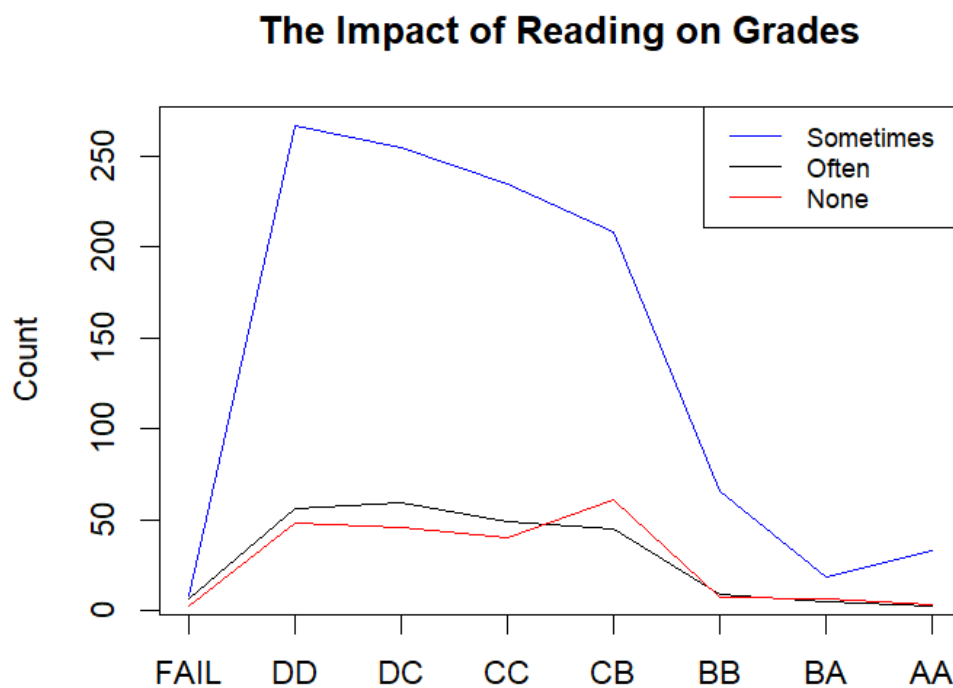


Figure 17

Below is a line plot diagram that showcases the impact of reading books on achieving grades. The blue line represents individuals who read occasionally, the black line represents those who read frequently, and the red line represents individuals who do not engage in reading. This visual representation effectively displays the patterns, in grade distribution based on varying reading habits.

```
# Create line plot for people who read sometimes distribution
plot(1:8, DisSome, type = "l", col = "blue", xaxt = "n", xlab = "", ylab = "Count", main = "The Impact of Reading on Grades")
axis(1, at = 1:8, labels = grades)

lines(1:8, DisOften, col = "black")

# Create line plot for people who do not read distribution
lines(1:8, DisNone, col = "red")
legend("topright", legend = c(LEVELS), col = c("blue", "black", "red"), lty = 1, cex = 0.8)
```

Figure 18

Above is a code snapshot that shows a line plot displaying the distribution of grades based on reading habits. The blue line represents individuals who read sometimes the black line corresponds to those who read often, and the red line shows students who never engage in reading. On the x axis you can see eight grades (FAIL, DD, DC, CC, CB, BB, BA, AA) while the y axis represents the count of students. This visualization helps us understand how reading frequency relates to performance. The title 'The Impact of Reading on Grades' captures the essence of this representation. Additionally customized labels for the x axis and a legend, with corresponding colors improve the clarity of the plot.



Figure 19



In the bar graph it is clear that there is a distinction, between individuals who have achieved high grades (AA, BA, BB) and those who have received low grades (CD, CC, DC, DD, FAIL). The orange color represents individuals with grades while the sky blue color represents those, with grades.

```
# Define the data
High_grades_Sometimes <- sum(SOMETIMES_AA, SOMETIMES_BA, SOMETIMES_BB)
High_grades_Often <- sum(OFTEN_AA, OFTEN_BA, OFTEN_BB)
High_grades_None <- sum(NEVER_AA, NEVER_BA, NEVER_BB)
Low_grades_Sometimes <- sum(SOMETIMES_CC, SOMETIMES_DD, SOMETIMES_FAIL, SOMETIMES_DC, SOMETIMES_CB)
Low_grades_Often <- sum(OFTEN_CC, OFTEN_DD, OFTEN_FAIL, OFTEN_DC, OFTEN_CB)
Low_grades_None <- sum(NEVER_FAIL, NEVER_CC, NEVER_DD, NEVER_CB, NEVER_DC)

# Create data vectors
HIGH <- c(High_grades_Sometimes, High_grades_None, High_grades_Often)
LOW <- c(Low_grades_None, Low_grades_Often, Low_grades_Sometimes)

MATRIX_FORMAT <- matrix(c(SOMETIMES,OFTEN,NONE),
                          nrow = 3,
                          ncol = 8,
                          byrow = TRUE,
                          list(c("SOMETIMES:", "OFTEN:", "NONE:"),
                               c(" AA ", " BA ", "BB ", " CB ", " CC ", " DC ", " DD ", " FAIL")))

data <- data.frame(
  LEVELS = c("NEVER", "SOMETIMES", "ALWAYS"),
  HIGH = c(High_grades_Sometimes, High_grades_None, High_grades_Often),
  LOW = c(Low_grades_None, Low_grades_Often, Low_grades_Sometimes)
)

# Reshape data to long format for ggplot2
MATRIX_FORMAT <- tidyr::gather(data, key = "Grade Level", value = "Count", -LEVELS)

ggplot(MATRIX_FORMAT, aes(x = LEVELS, y = Count, fill = `Grade Level`)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Reading Level", y = "Count", fill = "Grade Level") +
  ggtitle("Impact of Reading on Grades") +
  theme_minimal()
```

Figure 20

In the provided code screenshot, we are thoroughly. Visually represents the influence of reading habits on grades. We organize the data by calculating sums for combinations of reading frequencies and grades. To capture the counts of students with low grades based on their reading behaviors we utilize two vectors named 'HIGH' and 'LOW'.

For an overview of student counts corresponding to grade levels and reading frequencies we construct a matrix called 'MATRIX\_FORMAT'. Additionally, we created a data frame named 'data' to consolidate information about reading levels and the distribution of low grades.

To enhance visualization we reshape the data into a format using the 'tidyr::gather' function. Then with ggplot2 we generate a bar plot where the x axis represents reading levels the y axis represents student counts and bars are color coded according to grade level. The resulting visualization is titled "Impact of Reading on Grades". Provides insights into how reading habits correlate with academic performance. To ensure clarity, in presenting this representation we have chosen a theme that helps highlight patterns within the data.

### Analysis 3.1.5: Chi Square Test to Examine the Connection Between Reading Scientific and Getting Better Marks.

```
#create a table
Reading_table <- table(DATA$GRADE ,DATA$READ_FREQ_SCI)
Reading_table

# Create a new table excluding "Never" and "Often"
Reading_table <- Reading_table[, c("SOMETIMES", "OFTEN")]

# Perform Chi-square test
Reading_table <- chisq.test(Reading_table)

# Print the result
print(Reading_table)
```

Figure 21

	NONE	SOMETIMES	OFTEN
FAIL	2	8	6
DD	48	267	56
DC	46	255	59
CC	40	235	49
CB	61	208	45
BB	7	66	9
BA	6	18	5
AA	3	33	2

Figure 22

```
Pearson's Chi-squared test  
data: Reading_table  
X-squared = 11.861, df = 7, p-value = 0.1052
```

Figure 23

The code in Figure 21 above uses the "GRADE" and "READ\_FREQ\_SCI" variables in the dataset to create a contingency table ({Reading\_table}). After then, the table is filtered such that only the "SOMETIMES" and "OFTEN" categories are shown. The filtered table is subjected to a chi-square test, with printed results as shown in figure 22 and 23.

### Analysis 3.1.6: Conclusion for Question 1.

In summary, the Chi-square test findings, further evaluations, and graphical analyses demonstrate that students' scientific book reading habits had no impact on their final grades in Question 1. Additionally, the research emphasises the complexity of academic success by demonstrating how a range of factors, like students' sporadic book reading, affect their ability to attain high results.

### 3.1.7 Additional Features.

#### 1. explode():

- In a pie chart, the ``explode`` option modifies how far apart the separate slices are from the centre to highlight particular areas. You may use it to visually emphasise certain categories by underlining or demarcating them.

#### 2. axis():

- To add axis ticks and labels to a plot in R, use the ``axis`` function. For improved visualisation and interpretation, it enables customisation of the tick locations, labels, and other properties.

#### 3. lines():

One may add lines to an existing plot in R by using the ``lines`` function. In order to provide more information about trends or patterns, it is frequently used to overlay lines over scatter plots, line charts, or bar plots.

#### 4. legend():

- The R `legend()` function adds a legend to a plot that describes the many aspects it represents. It makes complicated visualisations easier to understand by enabling customisation of the legend's design, labels, and location.

#### 5. chisq.test():

- The R function `chisq.test()` is utilised to do the chi-square test of independence. It evaluates the relationship between categorical variables in contingency tables and determines if there is a significant difference between the observed and predicted frequencies.

#### 6. The ggplot2 Package's `ggplot()`:

- The `ggplot2` package contains the flexible function `ggplot()`, which is a strong R data visualisation tool. By employing a layered grammar of graphics approach, it makes it easier to create intricate and adaptable plots that provide flexibility in data representation.

#### 7. geom\_bar():

- To create bar charts, use the `geom_bar` `ggplot2` layer specifically designed for that purpose. It allows for customization of the bar's appearance and, when combined with additional layers, may provide complex visualisations.

#### 8. ggtitle():

- A plot's title may be added using the `ggplot2` function `ggtitle`. It allows you to offer a more comprehensible, descriptive label to the graphic depiction.

#### 9. "Assemble"::Tidyr():

- Data is converted from wide to long format using the `tidyr::gather()` function from the `tidyr` package. It is particularly useful for converting multi-column datasets into a format that makes them easier to examine and show.

#### 10. library(rgl) and install.packages("rgl"):

- Use ``install.packages("rgl")` to install the R package `rgl`, which makes 3D visualisation easier. The utilities for creating interactive 3D graphs are then made available by loading the package using `library(rgl)`.

11. ``install.packages("tidyr")` and ``library(tidyr)`:

- Similar to `rgl`, the `tidyr` package is a data manipulation tool that is especially helpful for reshaping and altering data for analysis. These scripts install and load the `tidyr` package.

12. `aes()`:

- An integral component of `ggplot2` is the `aes()` function, which maps variables to visual characteristics. It specifies the visual representation of variables in `ggplot` layers, including giving data points different colours or sizes.

13. `theme_minimal()`:

- The `ggplot2` function `theme_minimal()` establishes a minimalist plot style. It provides a tidy and well-organized

### 3.2 Question 2: What is the impact of regular preparation for midterms on the Final Grade? (IBRAHEEM MOHAMMED IMADELDIN AWAD\_TP070765\_Cyber Security)

The second question that will be answered using various data analysis techniques will be the impact of the regular preparation on the final grades of students. The impact of external factors such as whether the projects done by the student have a positive, neutral, or negative impact as well as the effect of students attending classes either always, sometimes, or never on their final grade which will be shown with corresponding graphs in this section as well.

```
install.packages("dplyr")
library(dplyr)
install.packages("plotrix")
library(plotrix)
install.packages("ggplot2")
library(ggplot2)
install.packages("plotly")
library(plotly)
install.packages("fmsb")
library(fmsb)
install.packages("gridExtra")
library(gridExtra)
```

Figure 24

Figure 21 shows all the package installed as well as all the libraries that have been used to perform the analysis for this section.

```
> Dataset_assignmet<- na.omit(Dataset_assignmet)
> Dataset_assignmet<- unique(Dataset_assignmet)
>
> nrow(Dataset_assignmet)
[1] 1534
```

Figure 25

Figure 22 shows that after omitting the null and duplicate rows, the number of remaining rows is 1534 rows after the data cleaning process has been completed.

### Analysis 3.2.1: Types of preparation done by the students for the exam and their impact. (Without applying filters to the graphs and data)

```
# Plotting for the second objective "impact of regular preparation for the midterms on achieving a higher grade"

# Plot a bar graph for all grades with count numbers
ggplot(Dataset_assignment, aes(x = GRADE, fill = PREP_EXAM)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Impact of Exam Preparation on All Grades",
    x = "Grade",
    y = "Count"
  ) +
  scale_fill_manual(
    values = c(
      "REGULARLY DURING THE SEM" = "lightblue",
      "CLOSES DATE TO THE EXAM" = "royalblue",
      "NEVER" = "darkblue"
    )
  ) +
  theme_minimal()
```

Figure 26

Figure 23 shows the code involved in making the stacked bar graph shown below by Figure 24 where the “ggplot” function is used to initialize the plotting process by laying the foundation for all upcoming layers, scales, and other plot-related additions. Following it, as parameters for the function, is the dataset that is going to be used for the plotting which in this case is “Dataset\_assignment” from which all the necessary columns as well as data related to the number of students and types of preparation done by them are going to be retrieved from. The second line shows the function “geom\_bar” which indicates that the function is specifically for creating a bar chart.

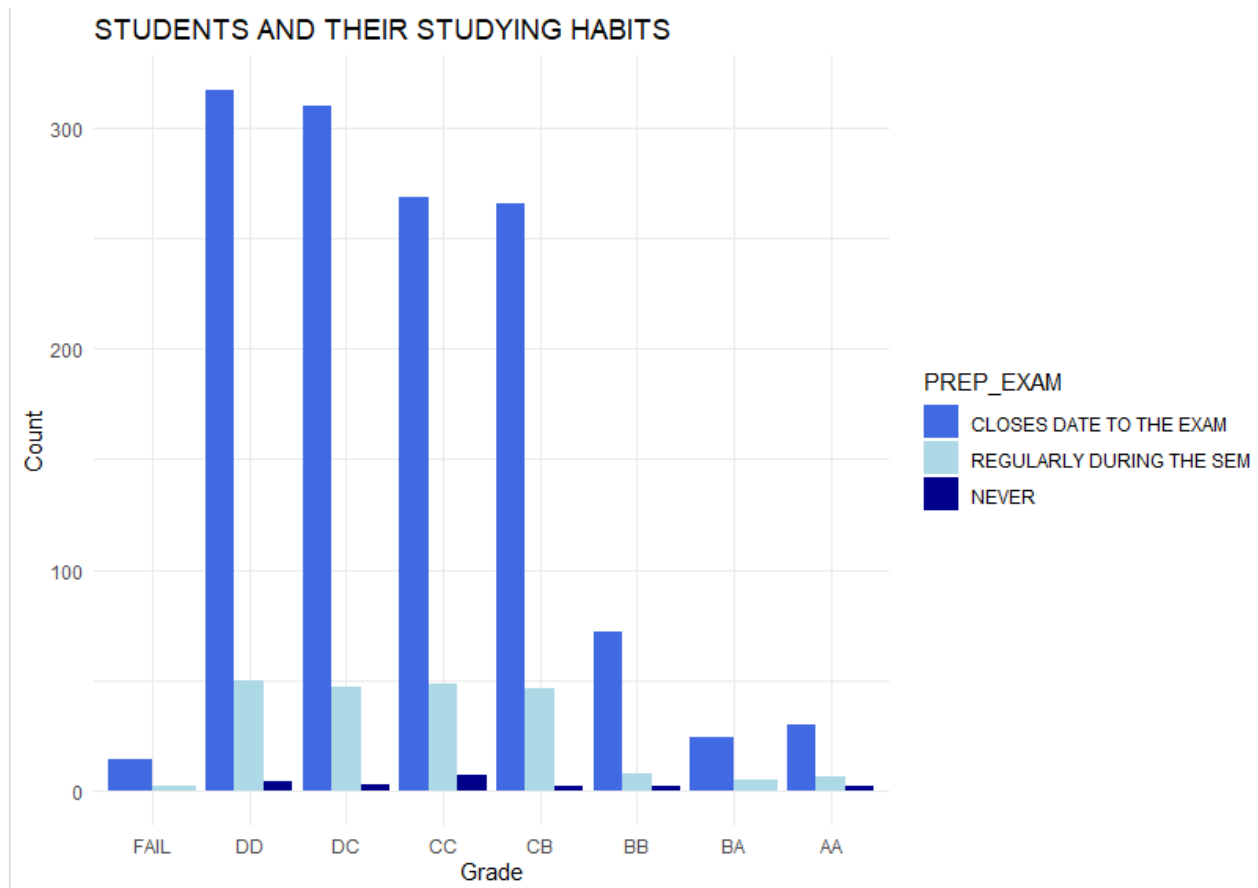


Figure 27

As shown by Figure 24, there are 3 categories of preparation which students categorized into. The first being those students who prepare closer to the date of the exam. The second category is for those students who prepare for the exams regularly during the semester. The final category is for those students who never prepare for the exams throughout the semester. As shown by the figure above, the largest number of students are in the category of preparing at a date that is closer to that of the exam. This applies to not only the students who on average achieved lower grades but also those who achieved higher grades, such as “BA” and “AA” grades.



```
# Calculate count for each GRADE
grade_counts <- table(Dataset_assignmet$GRADE)

# Create a pie chart
plot_ly(
  labels = names(grade_counts),
  values = grade_counts,
  type = "pie",
  marker = list(colors = c("purple", "red", "pink", "cyan", "yellow", "#fdbf6f", "#e6e6fa", "deeppink")),
  textinfo = "label+percent",
  hoverinfo = "label+percent",
  domain = list(x = c(0, 0.5), y = c(0, 1)),
  showlegend = FALSE
) %>%
layout(
  title = "PERCENTAGE OF STUDENTS IN EACH GRADE CATEGORY ",
  scene = list(
    aspectmode = "manual",
    aspectratio = list(x = 2, y = 2, z = 0.5)
  )
)
```

Figure 28

Figure 25 shows the code used to create the pie chart below using the function “plot\_ly” as shown in figure 26.

### PERCENTAGE OF STUDENTS IN EACH GRADE CATEGORY

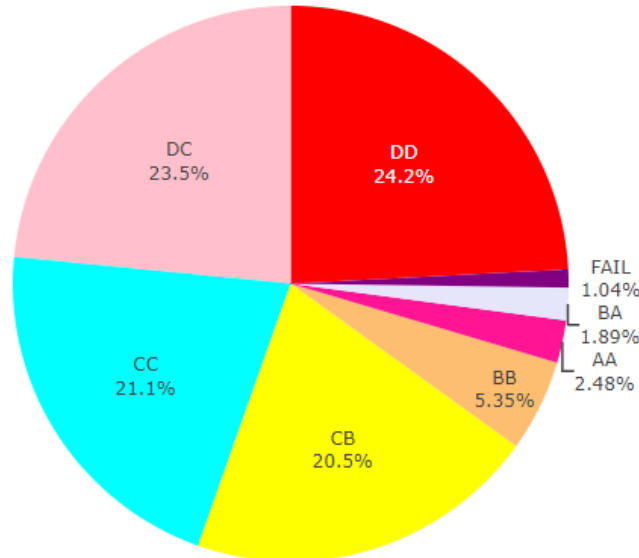


Figure 29

As shown by the pie chart in figure 26 the highest percentage of students are in the “DD” grade category with 24.2% of students having achieved that final grade. The grade category with the

second highest percentage is “DC” with 23.5% of students having achieved that grade and it is followed by “CC” grade with 21.1%, followed by “CB” grade with 20.5%, followed by “BB” grade with 5.35%, followed by “AA” grade with 2.48%, followed by “BA” grade with 1.89% and finally the “FAIL” grade with the lowest percentage of students having achieved that final result with a 1.04%.

### Analysis 3.2.2: Impact of regular preparation for midterms during the semester on final grade.

```
# Plot a line graph for the impact of EXAM preparation for the midterms on achieving a higher grade
ggplot(Dataset_assignmet, aes(x = GRADE, y = ..count.., color = PREP_EXAM, group = PREP_EXAM)) +
  geom_line(position = position_dodge(width = 0.5), stat = "count") +
  labs(
    title = "Impact of Exam Preparation on All Grades",
    x = "Grade",
    y = "Count"
  ) +
  scale_color_manual(
    values = c(
      "REGULARLY DURING THE SEM" = "purple",
      "CLOSES DATE TO THE EXAM" = "red",
      "NEVER" = "blue"
    )
  ) +
  theme_minimal()
```

Figure 30

Figure 27 shows the code written to produce the following line graph shown by figure 28. The same “ggplot” function is used but the function on the second line which is “geom\_line” which indicates that the graph that’s going to be generated is a line graph.

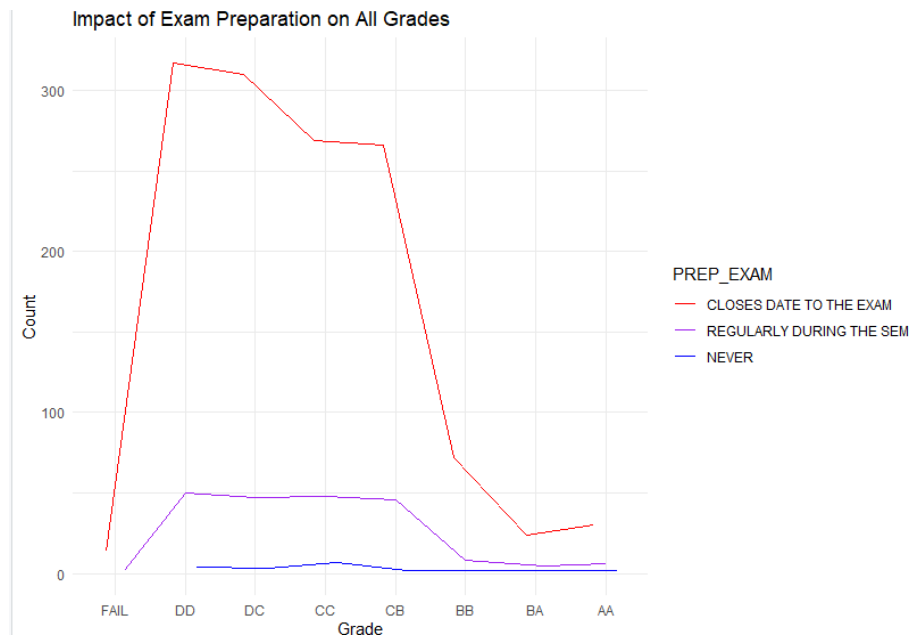


Figure 31

Figure 28 shows the impact of regular preparation on the final grades of various students. As observed by the graph. Grades that are categorized as “High” are “AA” grade which is the highest and “BA” grade. As shown by the line graph in Figure 28, a very small number of students that regularly prepare for the exams during the semester achieved a high grade. Figure 28 also shows that the largest number of students that achieved a high grade was done through studying closer to the date of the exam.

```
#BAR GRAPH where the impact of EXAM preparation is shown

# Filter the data set to include only "BA" and "AA" grades and count their occurrences
filtered_data <- Dataset_assignmet %>%
  filter(GRADE %in% c("BA", "AA")) %>%
  group_by(PREP_EXAM, GRADE) %>%
  summarise(count = n())

# Create a grouped bar graph showing the impact of exam preparation on "BA" and "AA" grades with count numbers
ggplot(filtered_data, aes(x = PREP_EXAM, y = count, fill = GRADE)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(
    aes(label = count),
    position = position_dodge(width = 0.9),
    vjust = -0.5,
    size = 3
  ) +
  labs(
    title = "Impact of Exam Preparation on 'BA' and 'AA' Grades",
    x = "Exam Preparation",
    y = "Count"
  ) +
  scale_fill_manual(values = c(
    "BA" = "deeppink",
    "AA" = "yellow"
  )) +
  theme_minimal()
```

Figure 32

Figure 29 shows the code used to display the graph in Figure 30 which focuses on the effect of 3 categories of exam preparation on students achieving a high final grade. Figure 29 begins by showing the filtering process where a variable called “filtered\_data” is created to filter data stored in the “Dataset\_assignment” variable based on the final grade where the condition is either “AA” or “BA” is the final grade. The data was also grouped into the “PREP\_EXAM” column and the “GRADE” column.

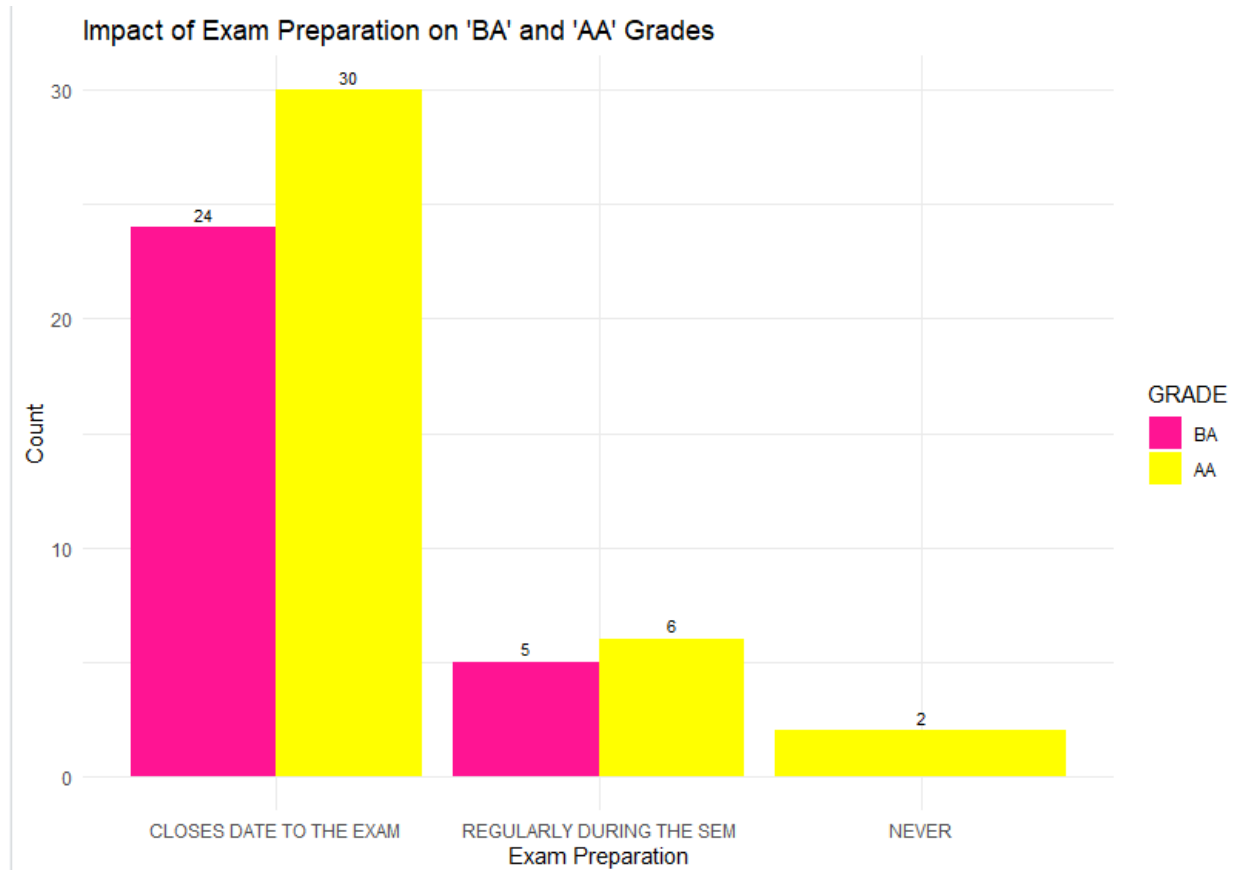


Figure 33

Figure 30 shows the bar graph showing the studying habits of those students who achieved “AA” and “BA” grades in a more focused and detailed manner when compared to Figure 24. As observed by the figure, there are only 2 students who achieved the “AA” grade without studying throughout the semester. The total number of students who achieved high grades while regularly preparing for the exams during the semester is 11 students and finally the largest number of students who have achieved high grades is 54 which was achieved by those students who study close to the exam

date. This can be explained by looking at 2 additional columns that can be considered as external factors, which are the “IMPACT” column and the “ATTEND” column.

### Analysis 3.2.3: Impact of external factors as well as regular preparation for midterms on final grades (Multivariate Analysis).

```
# Filter the data set to include only "BA" and "AA" grades and where the impact of preparation is close to the exam date and they always attend.
filtered_data <- Dataset_assignmet %>%
  filter(GRADE %in% c("AA", "BA") & PREP_EXAM == "CLOSES DATE TO THE EXAM" & ATTEND == "ALWAYS") %>%
  select(GRADE, PREP_EXAM, IMPACT, ATTEND)
view(filtered_data)
```

Figure 34

Figure 31 shows the filtered data with the added selection of the “EXTERNAL FACTORS” columns which are the “IMPACT” and “ATTEND” columns. As stated previously the impact column shows the impact of the projects in which the student took part on the final grade and the attend columns shows the student's attendance to classes.

```
# Define a blue color palette for grades
blue_palette <- c(
  "FAIL" = "navy",
  "DD" = "mediumblue",
  "DC" = "royalblue",
  "CC" = "dodgerblue",
  "CB" = "deeppskyblue",
  "BB" = "lightskyblue",
  "BA" = "lightblue",
  "AA" = "powderblue"
)

# Create the faceted bar chart
ggplot(Dataset_assignmet, aes(x = IMPACT, fill = GRADE)) +
  geom_bar(position = "dodge") +
  facet_grid(PREP_EXAM ~ ATTEND) +
  labs(
    title = "Grade Distribution by Impact, Attendance, and Exam Preparation",
    x = "Impact",
    y = "Count of Students",
    fill = "Grade"
  ) +
  scale_fill_manual(values = blue_palette) +
  theme_minimal()
```

Figure 35

Figure 33 shows the code for the Multivariate analysis done for “IMPACT”, “ATTEND”, and “PREP\_EXAM” columns against “GRADE” column. The colors get lighter as the grade improves.

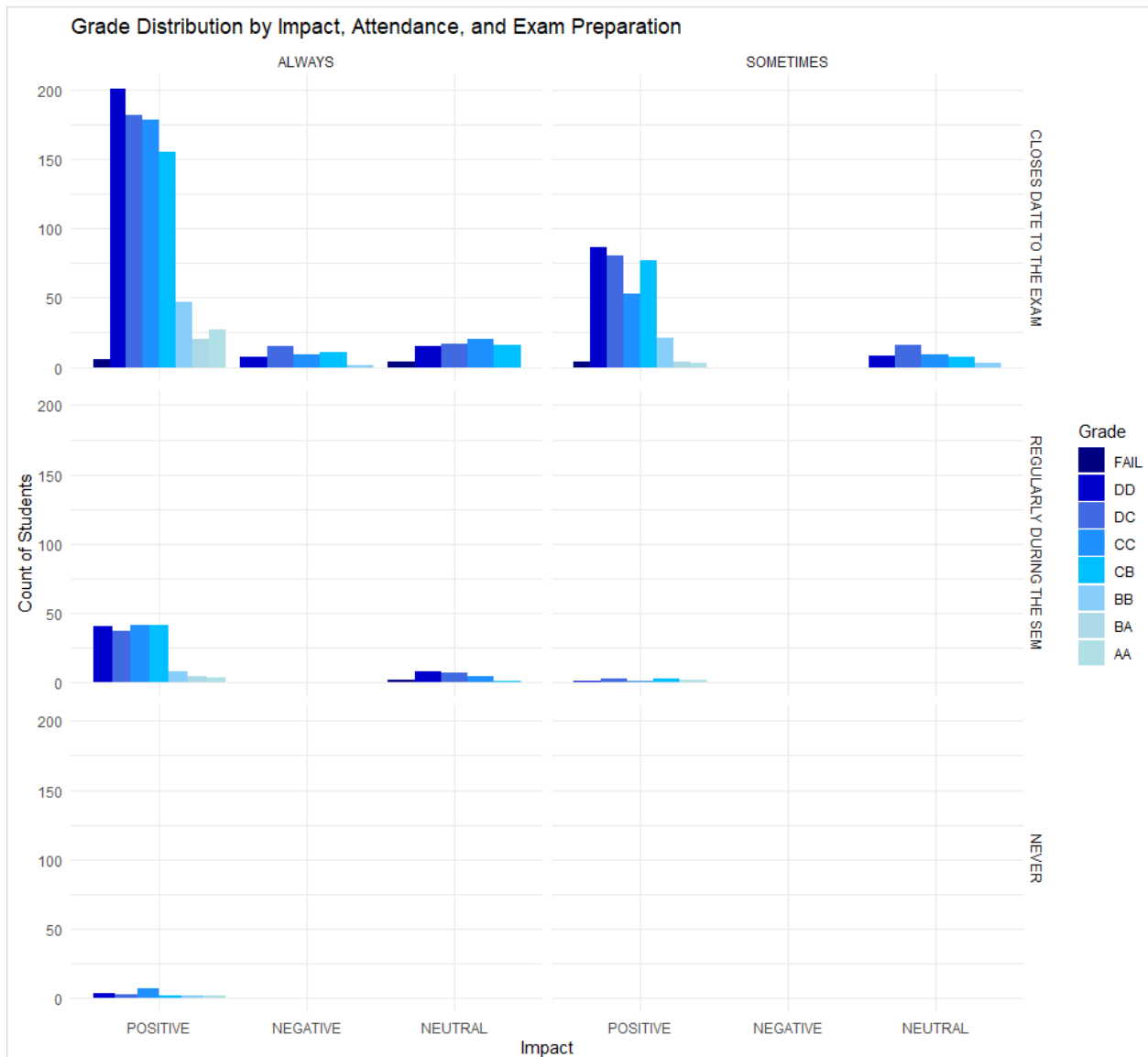


Figure 36

Figure 33 shows the Multivariate analysis result of code in figure 32. As observed in the graphs the number of students that always attended is much higher than those that attended sometimes. This directly translates to the number of students achieving “AA” and “BA” grades being much higher for those students who always attend. The grades are demonstrated using different shades with the student count being on the y axis of every graph. The x axis is categorized into 3 categories at the bottom which are positive, negative, and neutral impact of the project on the final grade. And the top shows the attendance of the students where it’s either always or sometimes as every student has attended a class at some point through the semester. When observing the faceted bar

chart, it can be identified that the greatest number of students who achieved the “AA” and “BA” grades are those students who prepare for the exam close to the date of the exam, have had projects that had a positive impact on their final grade, and finally have always attended classes. This leads to an understanding that there is either no relationship between regular preparation and achieving a high final grade or that the relation is there but it’s very insignificant.

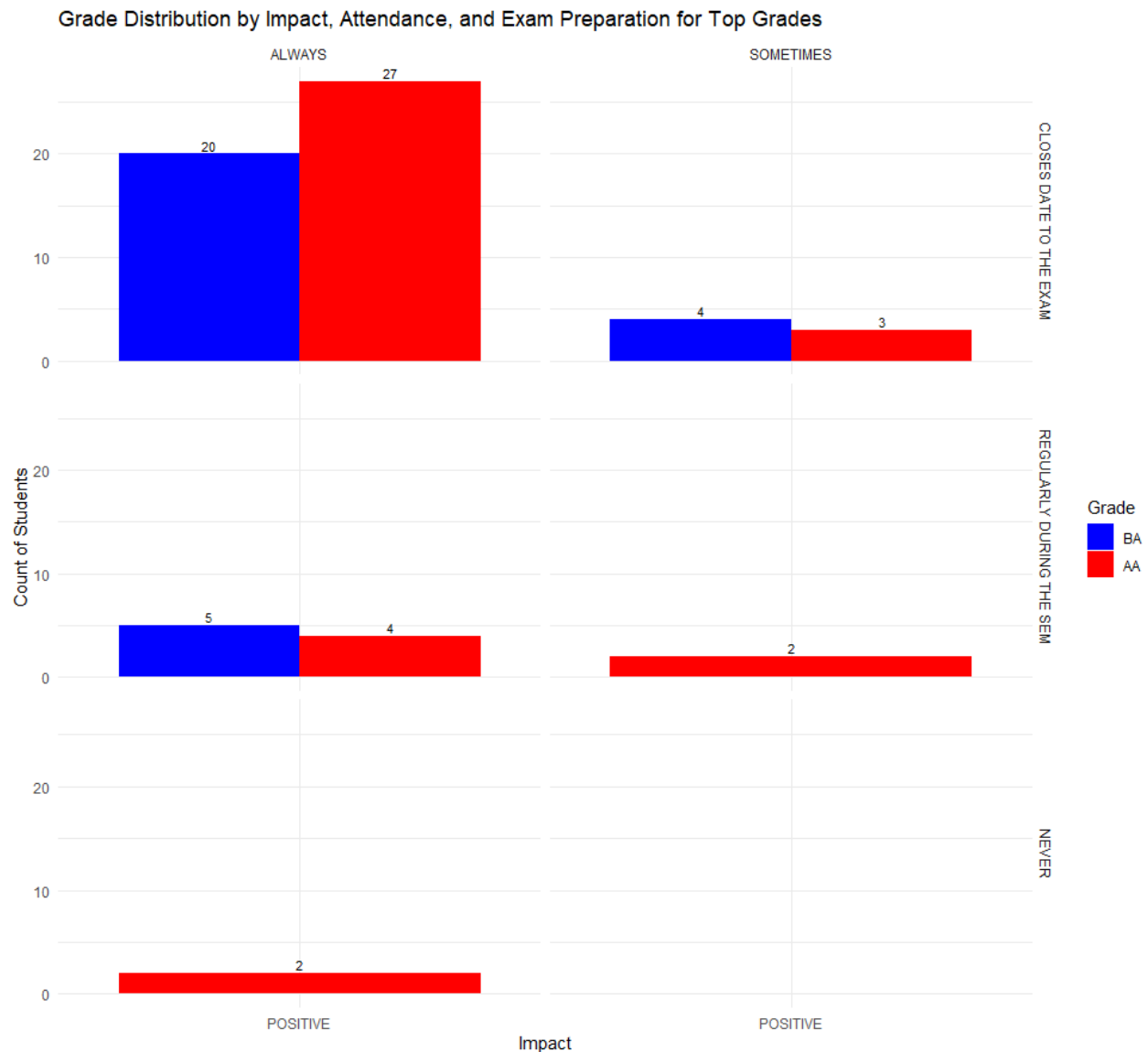


Figure 37

Figure 34 shows the explanation under figure 33 in more detail with regard those students whose final grade is “BA” and “AA”.

### Analysis 3.2.4: Chi Square Test for testing the relationship between regular preparation and achieving “AA” and “BA” grades.

The Chi-Square test is a statistical tool used to compare observed and anticipated data. This test may also be performed to see if it corresponds with our data's categorical variables. It aids in determining if a discrepancy between two category variables is the result of chance or of a link between them. (Biswal, 2023).

```
Exam_Prep_Table <- table(Dataset_assignmet$GRADE, Dataset_assignmet$PREP_EXAM)
print(Exam_Prep_Table)
# Remove Empty Values ('Never') since it will cause an error in Chi Square Test
Exam_Prep_Table <- Exam_Prep_Table[rowSums(Exam_Prep_Table) > 0, colSums(Exam_Prep_Table) > 0]
chisq.test(Exam_Prep_Table)
```

Figure 38

Figure 35 shows the snip of the code used to create the Chi Square Test observed in Figure 36.

```
      CLOSSES DATE TO THE EXAM REGULARLY DURING THE SEM NEVER
FAIL      14      2      0
DD      317      50      4
DC      310      47      3
CC      269      48      7
CB      266      46      2
BB       72       8      2
BA       24       5      0
AA       30       6      2
> # Remove Empty values ('Never') since it will cause an error in Chi Square Test
> Exam_Prep_Table <- Exam_Prep_Table[rowSums(Exam_Prep_Table) > 0, colSums(Exam_Prep_Table) > 0]
> chisq.test(Exam_Prep_Table)

      Pearson's Chi-squared test

data: Exam_Prep_Table
X-squared = 12.021, df = 14, p-value = 0.6046
```

Figure 39

When performing a Chi Square Test as shown by Figure 35, it is observed that the p-value equals 0.6046. In the context of Chi Square Test, the relationship is identified when the value of p is either less than or equal to 0.05. the p-value in Figure 35 shows that there is no relationship between Regularly studying during the semester and achieving a High GRADE.



### Analysis 3.2.5: Conclusion for Question 2.

To conclude, for Question 2, based on the Chi Square test as well as additional tests performed and information provided by various graphs, the final grade of students is found to be independent of the type of exam preparation done by student. It was also found that although students were studying closer to the date of the exam there were also additional factors affecting the final grades allowing them to achieve a high final grade and those factors were the impact of the project on the final grade as well as always attending the classes.

### 3.2.6 Additional Features.

1. **fmsb Library:**

This R package goes beyond the fundamental functionality found in regular scripts and is specifically designed for specialized charting and data processing applications.

2. **gridExtra Library:**

provides R utilities to help arrange a grid of plots in a methodical manner, improving the arrangement of several visual elements on a single page.

3. **factor() with Levels and Labels:**

In R, this function is important for translating numerical or textual vectors into categorized elements, specifying precise levels and labels, a key procedure in categorical data analysis and graphic representation.

4. **geom\_text():**

This function are essential parts of the ggplot2 suite and allow you to create a variety of R visualisations, such as line plots, bar charts, and textual annotations on graphs.

5. **aes():**

A function in R's ggplot2 package that determines the visual representation of data properties in a plot, such as axes, fills, and colors.

6. **scale\_fill\_manual(), scale\_color\_manual():**

To improve visual distinctiveness, these R functions are useful for manually adjusting colour schemes at various factor levels in a plot.

7. **theme\_minimal():**

ggplot2's aesthetic function, used to give graphical representations a clear, simple theme.

8. **plot\_ly():**

This function, which is a component of the R Plotly package, is essential for creating dynamic, interactive plots, including pie charts.

9. **layout():**

A versatile function in Plotly R that is used to adjust and customize plot layouts, such as positioning titles and axis labels.

10. **facet\_grid():**

An R function in ggplot2 that generates a structured matrix of panels depending on row and column variables supplied, allowing for thorough comparison visualizations.

11. **chisq.test():**

In R, the `chisq.test()` calls for the Chi-squared test which is used to evaluate the connection or independence between two categorical variables.

12. **summarise() from dplyr:**

Aggregation functions in R, with `n()` notably used for counting the count in each group, a key aspect of data summary in dplyr.

13. **geom\_bar(stat = "identity"):**

A subset of the `geom_bar` function in ggplot2, R, developed for bar plots when the bar heights are explicitly chosen by the data.

14. **position\_dodge():**

This function in ggplot2, R, when used with bar and line geometries, changes the placement of components to be side-by-side, essentially eliminating overlap in the visual result.

### 3.3 Question 3: What is the impact of actively taking notes on achieving higher grades? (ABDULELAH HUSSEIN ABDULRAHMAN AL-KAF | TP069319 | Cyber Security)

#### Analysis 3.3.1: Impact of taking notes with grade

In this analysis we will have a general view of the students who take notes and their impact on their grade.

```
library(ggplot2)
library(plotrix)
library(dplyr)

# Create a table of counts for NOTES and GRADE
notes_grade_counts <- table(Dataset_assignmet$NOTES, Dataset_assignmet$GRADE)

# Convert the table to a data frame
notes_grade_counts_df <- as.data.frame(notes_grade_counts)

# Rename columns for clarity
colnames(notes_grade_counts_df) <- c("Notes", "Grade", "Count")

# Plotting a grouped bar chart
ggplot(notes_grade_counts_df, aes(x = Notes, y = Count, fill = Grade)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  labs(title = "Impact of taking notes with grade",
       x = "Notes Taken",
       y = "Count of Students",
       fill = "Grade") +
  theme_minimal() +
  theme(
    text = element_text(size = 12),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
)
```

Figure 40

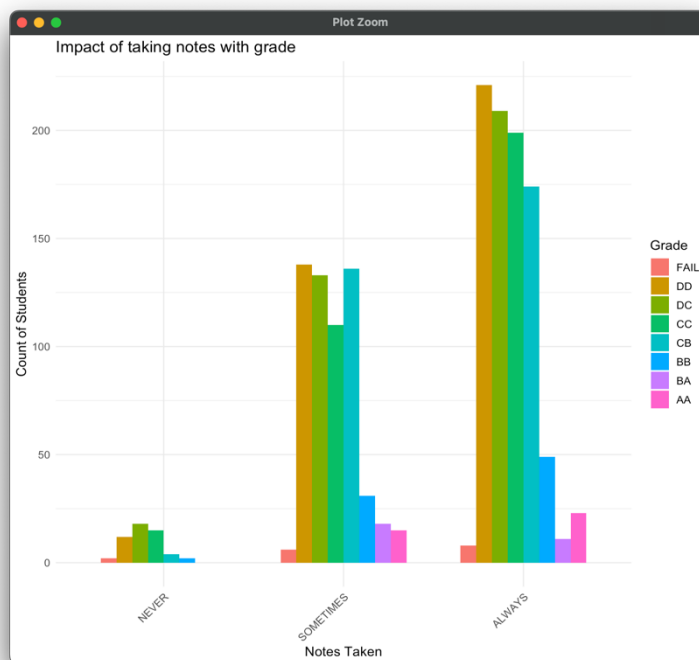


Figure 41

The figure illustrates the correlation between the students who take notes and their grades. It reveals that the students who 'ALWAYS' take notes make up a larger proportion of the population of students when compared to those who 'SOMETIMES' or 'NEVER' do so. Furthermore, the majority of the students who 'ALWAYS' take notes receive grades that are between 'DD' and 'CB,' which is considered to be a very poor mark.

In conclusion, students who ALWAYS, SOMETIMES, or NEVER take notes get a majority score that is between DD and CB, which is considered a very poor grade. This applies to students in all three categories.

### Analysis 3.3.2: Impact of taking notes with high grade

The primary goals of this analysis are to determine whether or not how well students who consistently take notes are well academically, as measured by grades ranging between BB to AA.

```
# Filter the dataset for BB, BA, and AA grades
high_grade_categories <- c("BB", "BA", "AA")
filtered_data <- Dataset_assignmet[Dataset_assignmet$GRADE %in% high_grade_categories, ]

# Calculate the counts for each note category within high grades
high_grade_notes_counts <- table(filtered_data$NOTES)

# Create a data frame
high_grade_notes_df <- as.data.frame(high_grade_notes_counts)

# Rename columns for clarity
colnames(high_grade_notes_df) <- c("Notes", "Count")

# Create labels with counts
labels_with_counts <- paste(high_grade_notes_df$Notes, "\n", "Students: ", high_grade_notes_df$Count)

# Plotting a 3D pie chart for notes distribution within high grades
library(plotrix)

pie3D(high_grade_notes_df$Count, labels = labels_with_counts, radius = 1,
      main = "Impact of Taking Notes on High Grades (BB, BA, AA)", col = c("red", "darkblue", "orange"))
```

Figure 42

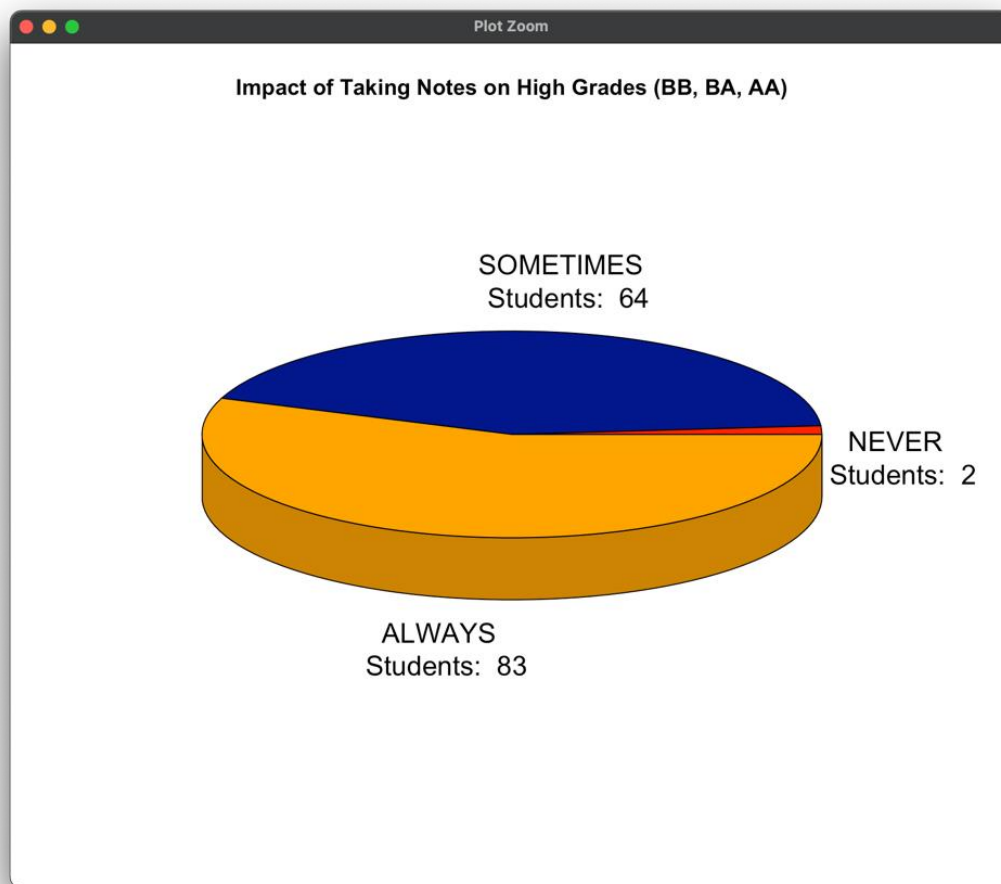


Figure 43

This figure represents the number of students who take notes and earn high grades of BB, BA, or AA. There are 83 students who ALWAYS take notes and score high grades, 64 students who SOMETIMES take notes and score high grades, and only 2 students who NEVER take notes who managed to earn high grades.

To summarize, the total number of students that get high grades regardless of whether they take notes or not is 149, which is clearly a relatively small number in comparison to the students who are included in the dataset, which totals somewhere around 1500 students.

### Analysis 3.3.3: Impact of taking notes and gender with high grade

The objective of this analysis is to investigate the connection between the manner in which students take notes, gender and their high grade. After gender-segregating the dataset, we conduct an analysis of the frequency distribution of note-taking within each gender group and their grades in order to find if they score high grade or not.

```
ggplot(Dataset_assignmet, aes(x = GENDER, fill = NOTES)) +  
  geom_bar(position = "stack", stat = "count") +  
  scale_fill_manual(values = c("NEVER" = "red", "SOMETIMES" = "yellow", "ALWAYS" = "green")) +  
  labs(title = "Note-Taking Distribution by Gender",  
        x = "Gender",  
        y = "Count") +  
  theme_minimal()  
  
# Assuming Dataset_assignmet is your data frame and 'GENDER' and 'GRADE' are factors  
  
# Load necessary library  
library(ggplot2)  
  
# Convert GENDER to categorical levels (if not done already)  
# Plotting a grouped bar chart for GENDER and GRADE  
ggplot(Dataset_assignmet, aes(x = GENDER, fill = GRADE)) +  
  geom_bar(position = "dodge", width = 0.7) +  
  labs(title = "Relationship between Gender and Grade",  
        x = "Gender",  
        y = "Count",  
        fill = "Grade") +  
  theme_minimal()
```

Figure 44

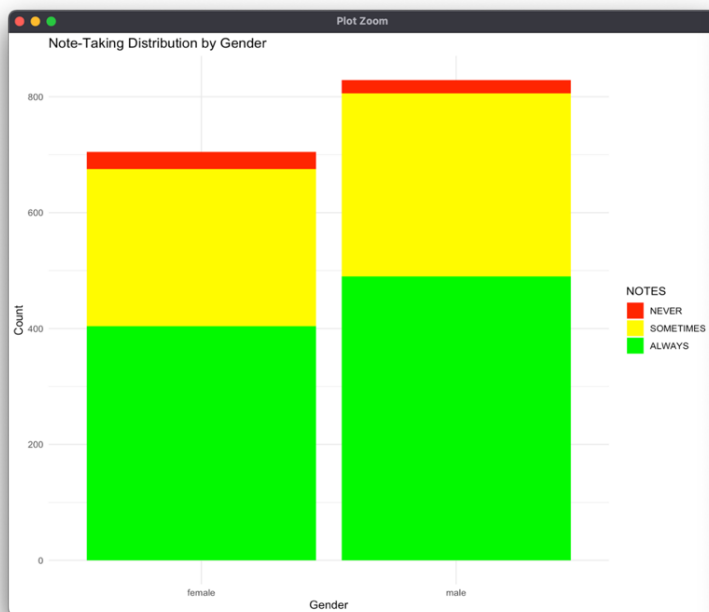


Figure 45

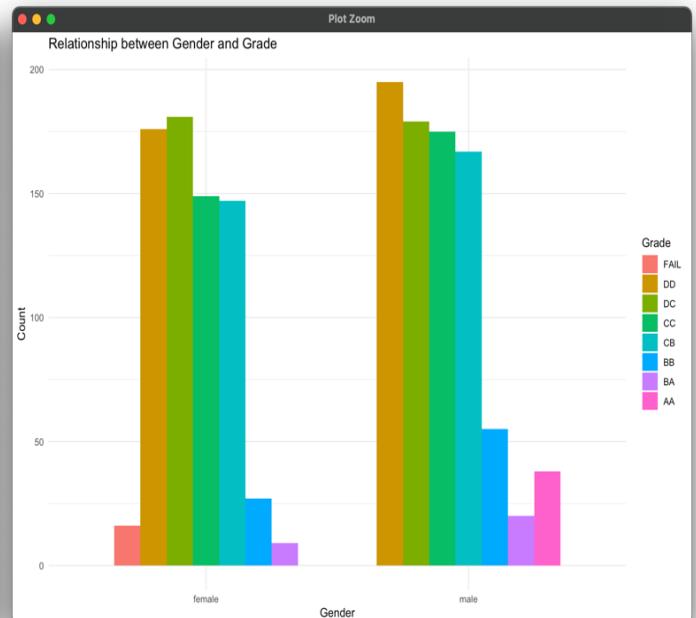


Figure 46

The first figure on the left demonstrates that male students make up the majority of note-takers, with over 800 students, while female students make up less than 800. The green part of the stacked bar indicates that male students take notes more frequently than female students do as shown in 'ALWAYS', and the yellow portion of the stacked bar indicates that male students take notes more occasionally than female students do as shown in 'SOMETIMES'. The red area of the stacked bar indicates that the female students are more than male students in 'NEVER' takes notes frequently. The second figure indicate the grades of the male and female students, as it shows most of the male students score very low grades its between DD and CB, its clearly shows a big difference with the high grades which is between BB and AA and it goes the same with the female students.

To conclude, this analysis indicate notable variations in note-taking practises across genders grade among the student body. Male students make up over 800 people, a much higher percentage of note-takers than female students, who make up fewer than 800 persons in the sample. Different trends can be seen in the stacked bar chart, the green segment which represents "ALWAYS" note-taking frequency, shows that male students are more likely than female students to take notes. In the other hand the analysis have approved that both male and female students most of them have scored a low grade which is between DD and CB.

#### Analysis 3.3.4: Impact of taking notes, gender and high school type with high grade

The purpose of this analysis is to look at the relationship between students' note-taking behaviors and compare the type of high school they attend with the high grade. The high school type of each category will be used to separate the dataset, and then the pattern of frequency of high school within their grades will be analyzed. Finding any common patterns or differences in grades among students who attended various kinds of high schools is the aim of this study.

```
# Plot the grouped bar plot for High School Type and Notes
ggplot(Dataset_assignmet, aes(x = HS_TYPE, y = ..count.., fill = NOTES)) +
  geom_bar(position = "dodge", stat = "count") +
  scale_fill_manual(values = c("NEVER" = "forestgreen", "SOMETIMES" = "maroon", "ALWAYS" = "cyan")) +
  labs(title = "Note-Taking Distribution by High School Type",
       x = "High School Type",
       y = "Count") +
  theme_minimal()

# Calculate counts of each grade within each high school type
grade_counts <- table(Dataset_assignmet$HS_TYPE, Dataset_assignmet$GRADE)

# Convert table to a data frame
grade_counts_df <- as.data.frame.matrix(grade_counts)
grade_counts_df$HS_TYPE <- rownames(grade_counts_df)

# Reshape data for plotting
library(reshape2)
grade_counts_melted <- melt(grade_counts_df, id.vars = "HS_TYPE")

# Plotting stacked bar chart for HS_TYPE and GRADE
ggplot(grade_counts_melted, aes(x = HS_TYPE, y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Relationship between High School Type and Grades",
       x = "High School Type",
       y = "Count",
       fill = "Grade") +
  theme_minimal()
```

Figure 47



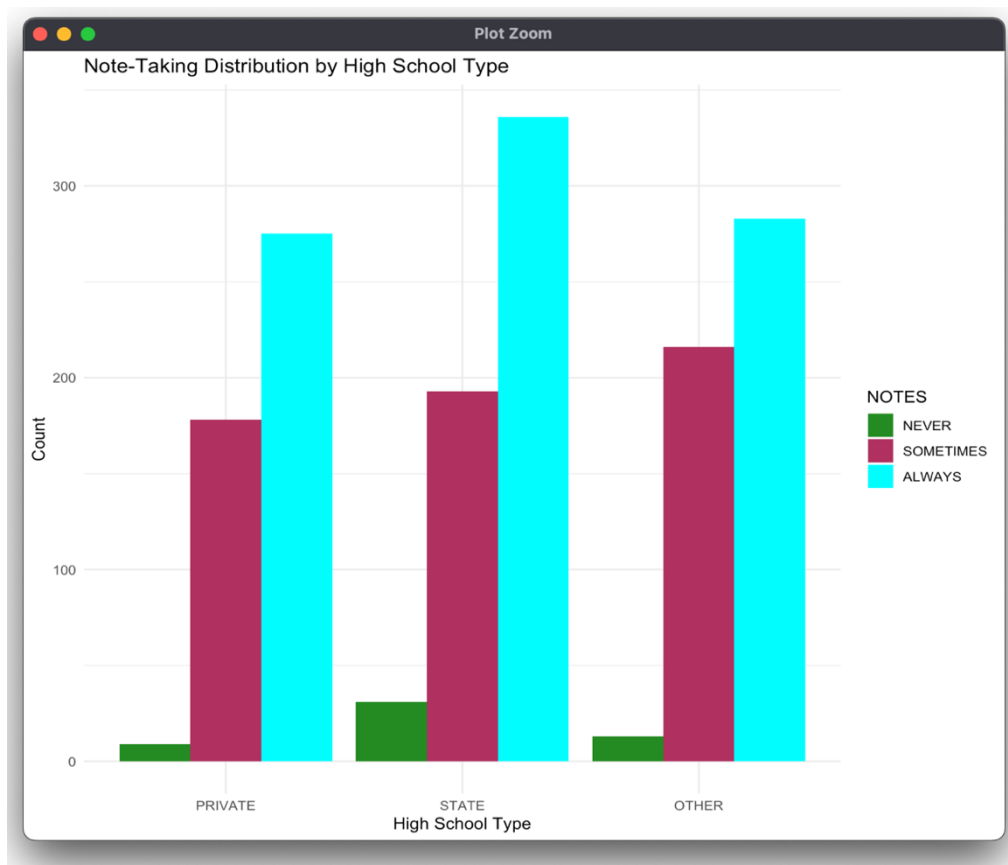


Figure 48

The above figure shows a grouped bar plot to show the number of students studying in a certain kind of high school as well as the number of students who are actively taking notes. As we can see, state high schools enroll the majority of students, with 'other' school types accounting for the second most common enrollment category. Private schools are also the least frequent kind of schooling establishment. Most students who actively take notes "ALWAYS" are enrolled in state schools, these schools typically have over 300 students. In return, most students who take notes "SOMETIMES" are enrolled in other types of high schools, additionally, state high school students are more likely to "NEVER" actively take notes.

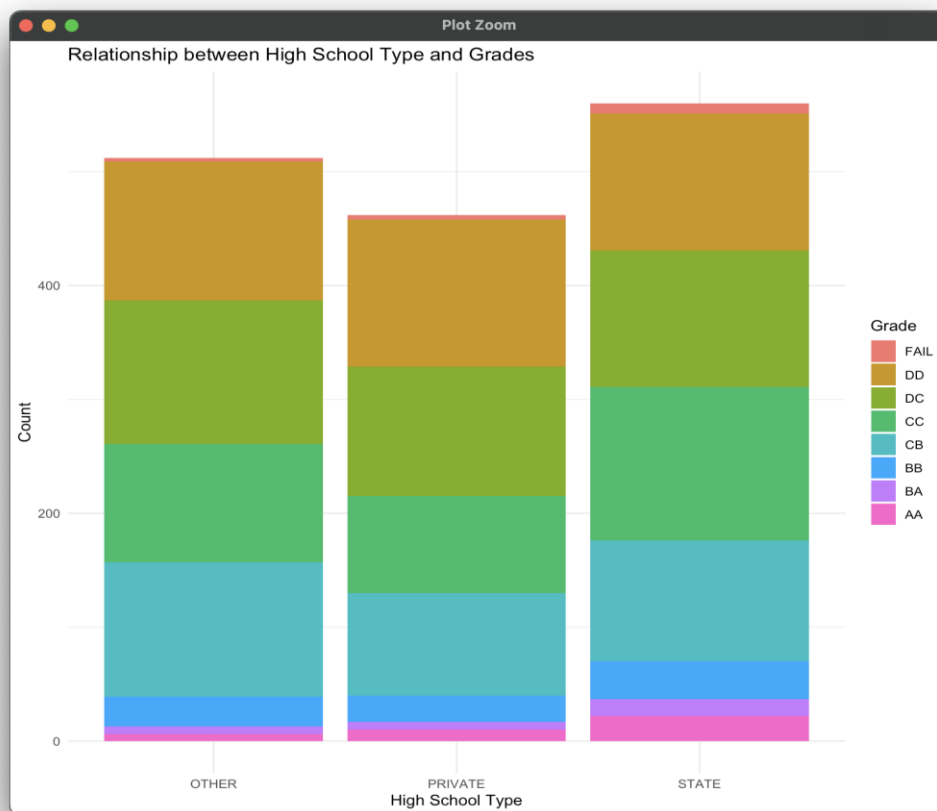


Figure 49

The figure above shows the types of high school and the students grade based on each school type, most of the students are enrolled in a state high school and most of the students grades are very low, most of them score between 'DD' and 'CB' and it's the same in the 'PRIVATE' and other type of high schools.

In Conclusion, Note-taking adheres to at different high schools and there grades provide data on the link between educational backgrounds and student note taking habits. Most students enrol in 'state' high schools, followed by "other" institutions, and fewest in private schools. Note taking is common in state high schools, where over 400 students "ALWAYS" take notes and many of them score very low grades. According to the analysis, students at private and other high schools are more likely to have many students who score low grades.

### Analysis 3.3.5: Impact of taking notes, gender, high school and attendance with high grade

The analysis focuses on examining the relationship between students' note-taking habits and their attendance in class. Students will be classified as either always, sometimes, or never attendees by dividing the dataset according to their attendance patterns. The distribution of note-taking frequencies among each attendance group will next be examined in more detail. Finding any noticeable trends or differences in note-taking practices that correlate with differences in attendance levels is the goal.

```
library(ggplot2)
# Calculate counts of notes and attendance
notes_attendance_counts <- table(Dataset_assignmet$NOTES, Dataset_assignmet$ATTEND)

# Convert the table to a data frame and create a factor for 'Notes'
notes_attendance_counts_df <- as.data.frame(notes_attendance_counts)
notes_attendance_counts_df$Notes <- factor(notes_attendance_counts_df$Var1,
                                           levels = c("NEVER", "SOMETIMES", "ALWAYS"))

# Rename columns for clarity
colnames(notes_attendance_counts_df) <- c("Notes", "Attendance", "Count")

# Plotting
ggplot(notes_attendance_counts_df, aes(x = Notes, y = Count, group = Attendance, color = Attendance)) +
  # Customize the line aesthetics
  geom_line(size = 1.5, linetype = "solid") +
  # Customize the point aesthetics
  geom_point(size = 3, shape = 16) +
  # Customize plot labels and title
  labs(title = "Relationship between Notes Taken and Attending Class",
       x = "Notes",
       y = "Number of students",
       color = "Attendance") +
  # Apply a minimal theme
  theme_minimal() +
  # Additional theme customization for better readability
  theme(
    text = element_text(size = 12), # Increase text size
    axis.text.x = element_text(angle = 45, hjust = 1) # Rotate x-axis text for better visibility
  )
```

Figure 50

```
# Filter data for high grades ('BB', 'BA', 'AA')
high_grade_data <- Dataset_assignmet[Dataset_assignmet$GRADE %in% c("BB", "BA", "AA"), ]

# Calculate counts of high grades within each attendance level
grade_counts <- table(high_grade_data$ATTEND, high_grade_data$GRADE)

# Convert table to a data frame
grade_counts_df <- as.data.frame.matrix(grade_counts)
grade_counts_df$ATTEND <- rownames(grade_counts_df)

# Plotting a line graph for ATTEND and high grades (BB, BA, AA)
ggplot(grade_counts_df, aes(x = ATTEND, y = `BB` + `BA` + `AA`, group = 1)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 3) +
  labs(title = "Relationship between Attendance and High Grades (BB, BA, AA)",
       x = "Attendance",
       y = "Count of (BB,BA,AA) Grades",
       color = "High Grades") +
  theme_minimal()
```

Figure 51

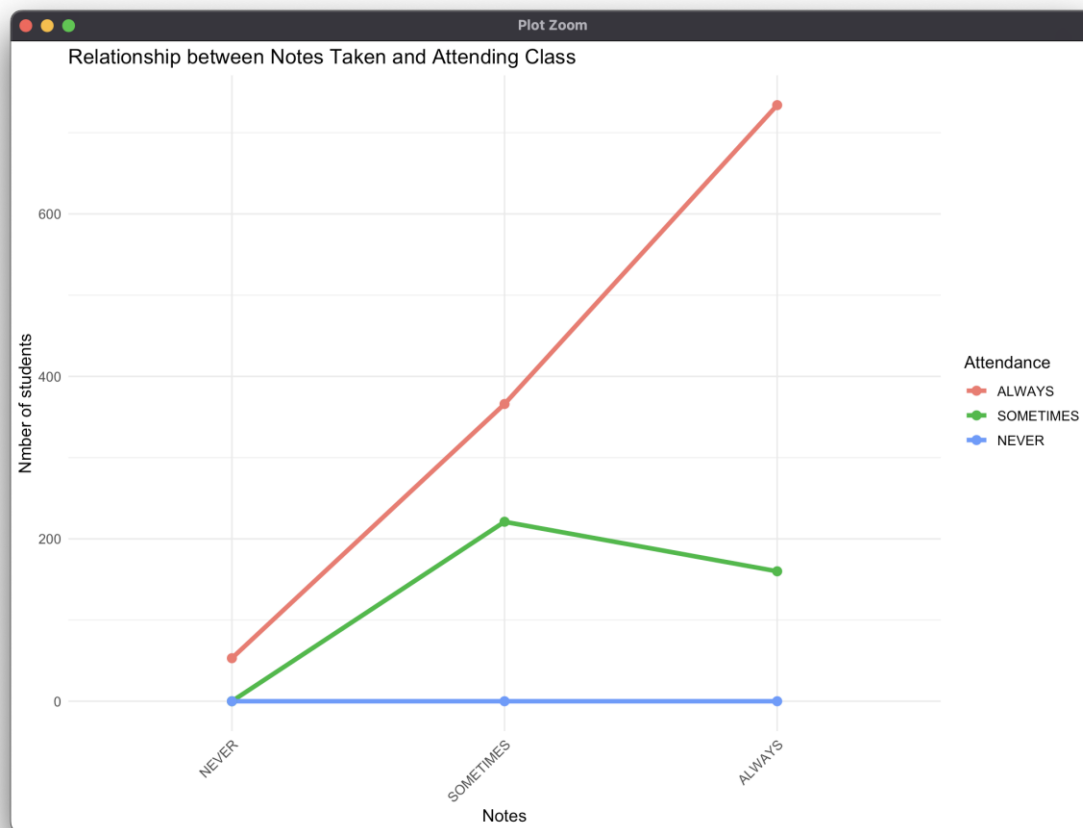


Figure 52

The relationship between a student's attendance and note-taking is displayed in the figure. The line graph that represented by the blue line, shows that students who never attend classes clearly "NEVER" take notes, as evidenced by the solid "0" that appears next to the phrases "NEVER," "SOMETIMES," and "ALWAYS." The red graph, which is positioned above the other line graphs, shows the largest amount of students who always attend. The majority of the students who "ALWAYS" attend classes also take notes "ALWAYS," which has the highest number of students—more than 700. On the other hand, the green line graph represents the students who "SOMETIMES" attend classes. This graph also reveals that most students take notes "SOMETIMES," with a number of more than 200 students.

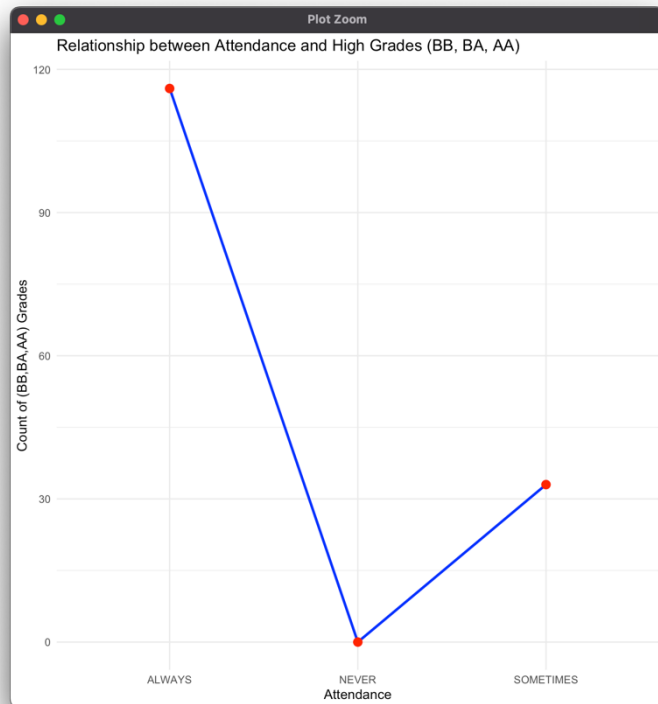


Figure 53

In this figure it shows the relationship between the students attendance and the high grades, in the line graph it shows the number of students who 'ALWAYS' attend is more than 100 students, and the students who 'NEVER' attend is 0, which means no one of them score high grade, and the students who 'SOMETIMES' attend, more than 30 of them who score high grades.

In summary, this analysis tries to determine if there is a significant correlation between taking notes and attending classes and relate it with high grade on a regular basis. It does this by comparing the numbers of consistent, infrequent, and non-takers across various attendance patterns. Gaining insight into the possible interactions between note-taking practises and classroom attendance with high grade may help determine how attendance practises affect students' grades patterns and level of participation in class. And it shows that very less students who score high grades.

**TESTING :**

```
# Load necessary library
library(gmodels) # For CrossTable function
install.packages("gmodels")
# Assuming 'Taking Notes' and 'Higher Grade Achievement' columns exist in Dataset_assignmet

# Create a contingency table between 'Taking Notes' and 'Higher Grade Achievement'
contingency_table <- table(Dataset_assignmet$NOTES, Dataset_assignmet$GRADE)

# Perform Chi-square test of independence
chi_sq_test <- chisq.test(contingency_table)

# Display the contingency table
print(CrossTable(Dataset_assignmet$NOTES, Dataset_assignmet$GRADE))

# Display the result of the Chi-square test
print(chi_sq_test)
```

Figure 54

Cell Contents

	N									
Chi-square contribution										
N / Row Total										
N / Col Total										
N / Table Total										

Total Observations in Table: 1534

	Dataset_assignmet\$GRADE									
Dataset_assignmet\$NOTES	FAIL	DD	DC	CC	CB	BB	BA	AA	Row Total	
NEVER	2	12	18	15	4	2	0	0	53	
	3.789	0.052	2.487	1.294	4.324	0.245	1.002	1.313		
	0.038	0.226	0.340	0.283	0.075	0.038	0.000	0.000	0.035	
	0.125	0.032	0.050	0.046	0.013	0.024	0.000	0.000		
	0.001	0.008	0.012	0.010	0.003	0.001	0.000	0.000		
SOMETIMES	6	138	133	110	136	31	18	15	587	
	0.002	0.111	0.164	1.577	2.089	0.005	4.294	0.014		
	0.010	0.235	0.227	0.187	0.232	0.053	0.031	0.026	0.383	
	0.375	0.372	0.369	0.340	0.433	0.378	0.621	0.395		
	0.004	0.090	0.087	0.072	0.089	0.020	0.012	0.010		
ALWAYS	8	221	209	199	174	49	11	23	894	
	0.188	0.106	0.003	0.548	0.442	0.031	2.060	0.033		
	0.009	0.247	0.234	0.223	0.195	0.055	0.012	0.026	0.583	
	0.500	0.596	0.581	0.614	0.554	0.598	0.379	0.605		
	0.005	0.144	0.136	0.130	0.113	0.032	0.007	0.015		
Column Total	16	371	360	324	314	82	29	38	1534	
	0.010	0.242	0.235	0.211	0.205	0.053	0.019	0.025		

Figure 56

x	y							
	FAIL	DD	DC	CC	CB	BB	BA	AA
NEVER	2	12	18	15	4	2	0	0
SOMETIMES	6	138	133	110	136	31	18	15
ALWAYS	8	221	209	199	174	49	11	23

Figure 55

The testing for our aim, which is to determine the impact of students who take notes on higher grades, is shown in the figure that can be seen above. It has calculated the rows, the columns, and the total, and it confirmed that the majority of students who take notes get a lower grade.

### CONCLUSION OF QUESTION 3

In conclusion, the objective of "what is the impact of actively taking notes on achieving a higher grade?" was to investigate this question. and according to our analysis, it is consistent with the hypothesis.

The majority of students who actively take notes score between CB to DD, which is the grade range that has been determined to be acceptable for students who take notes and who do not get high grades.

#### Extra features :

1. **library(gmodels)** : This library is used to create the CrossTable function for the testing
2. **theme\_minimal()** : This code is used to modify the appearance and style of the plot to achieve a minimalistic design.
3. **library(reshape2)**: This library in R provides functions to reshape, restructure, and transform data frames.
4. **scale\_fill\_manual()**: Is used to manually set the colors or fill for different groups or categories within a plot

## Question 4: the impact of students studying between 11 and 20 hours per week on achieving higher grade (MUHAMAD AHMAD AL MUHDAR TP070208) | Cyber Security

### Analysis 3.4.1: impact of studying 11 to 20 hours with grade

In this analysis we will check if study hours for the students affect grades, if they achieve better

```
# Filter dataset for all study hours
filtered_data <- Dataset_assignmet

# Create a grouped bar chart for study hours and grade
ggplot(filtered_data, aes(x = STUDY_HRS, fill = GRADE)) +
  geom_bar(position = "dodge", width = 0.7) +
  labs(title = "Relationship between Study Hours and Grades",
       x = "Study Hours",
       y = "Number of Students",
       fill = "GRADE") +
  theme_minimal()
```

grades or no depending on study hours.

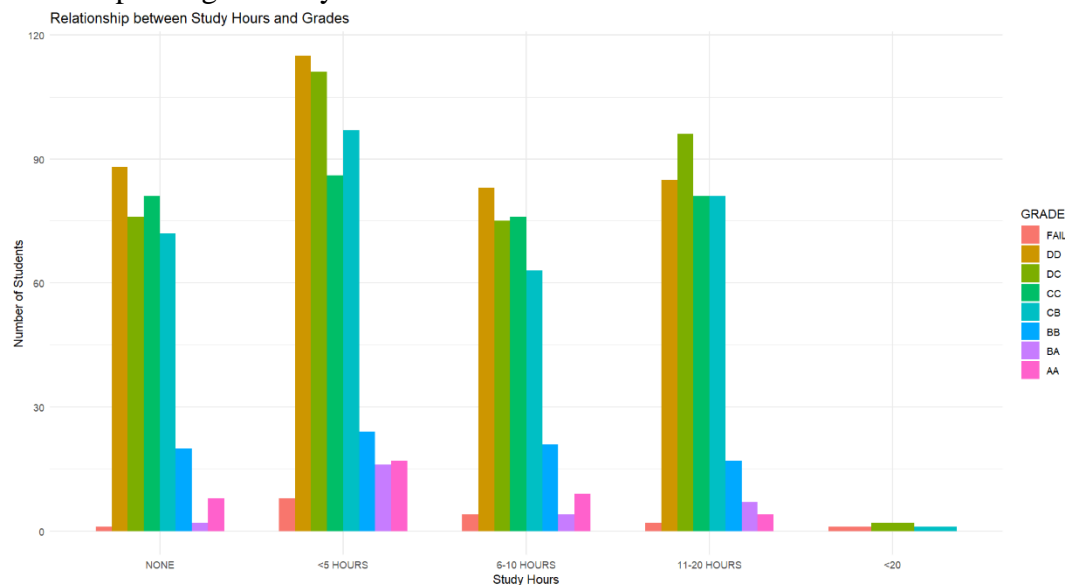


Figure 57

Figure 58

The graph above shows the number of students who study during different hours. It is important to understand that achieving good grades doesn't just rely on hard study. The data shows that students who study for 6-10 hours are typically achieving better grades than those who study for



11-20 hours. Furthermore, students who study for 5 hours or less are frequently getting higher grades than other students.

in conclusion, studying for more hours does not ensure getting better grades. Students who study less can achieve higher grades.

### **Analysis 3.4.2:** *impact of studying 11 to 20 hours with high grades*

Through this investigation, we hope to learn more about how study hours affect students' academic achievement. Specifically, we want to know if students who study for 11 to 20 hours a week achieve higher grades.

```
# Filter dataset for all study hours and higher grades
filtered_data <- subset(Dataset_assignment, GRADE %in% c("BB", "BA", "AA"))

# Create a grouped bar chart for study hours and higher grades
ggplot(filtered_data, aes(x = STUDY_HRS, fill = GRADE)) +
  geom_bar(position = "dodge", width = 0.7) +
  labs(title = "Relationship between Study Hours and Higher Grades",
       x = "Study Hours",
       y = "Number of Students",
       fill = "GRADE") +
  theme_minimal()
```

Figure 59

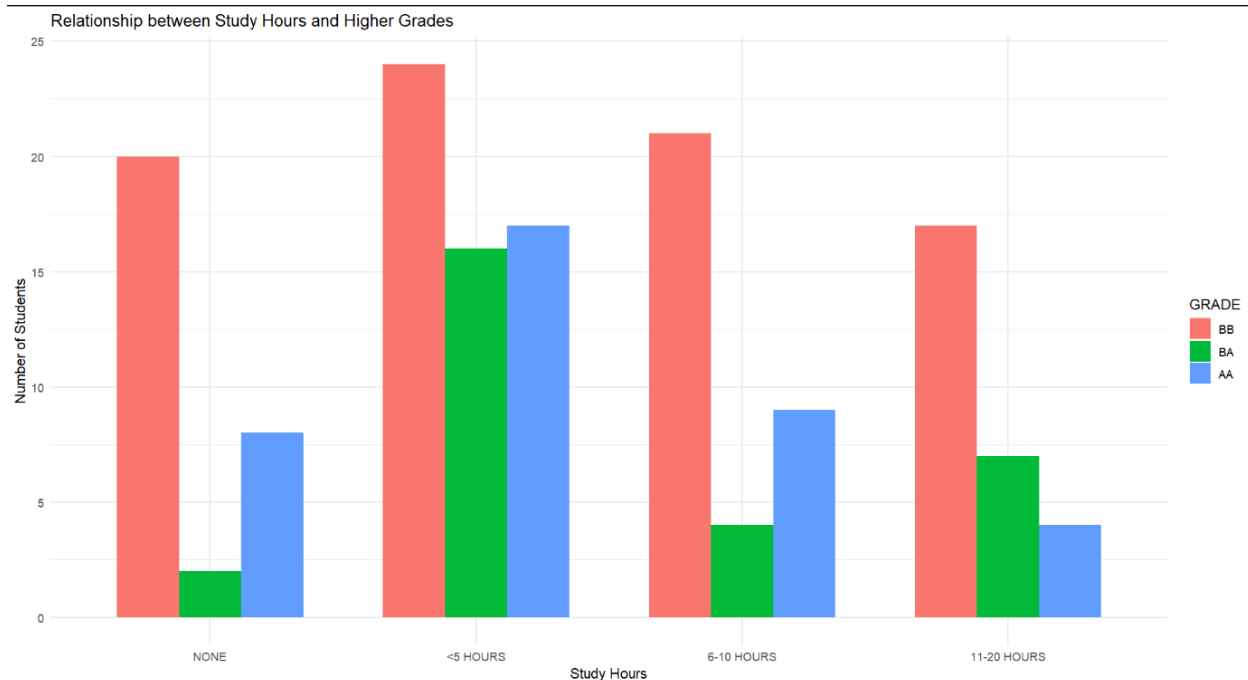


Figure 60

The graph indicates that study hours and grades are negatively correlated. It means that students who spend more study time are less likely to achieve better marks. That being said, the link is nonlinear, so a grade drop is not always the result of studying an extra hour.

In conclusion, studying longer does not guarantee a great grade. Students who studied for fewer hours had higher scores than others.

### **Analysis 3.4.3:** *impact of studying 11 to 20 hours and living with high grade*

The purpose of this analysis is to investigate the impact of study hours on students' grades, specifically the impact of studying 11 to 20 hours per week and living arrangements on achieving higher marks.

```
# Filter dataset for study hours from 11 to 20 hours and higher grades
filtered_data <- subset(Dataset_assignmet, STUDY_HRS == "11-20 HOURS" & GRADE %in% c("BB", "BA", "AA"))

ggplot(filtered_data, aes(x = STUDY_HRS, fill = LIVING)) +
  geom_bar(position = "stack", color = "white", alpha = 0.7, stat = "count") +
  facet_grid(. ~ GRADE) +
  labs(title = "Relationship between Study Hours (11-20 hours), Living Situation, and Grade",
       x = "Study Hours",
       y = "Number of Students",
       fill = "Living Situation") +
  theme_minimal()
```

Figure 61

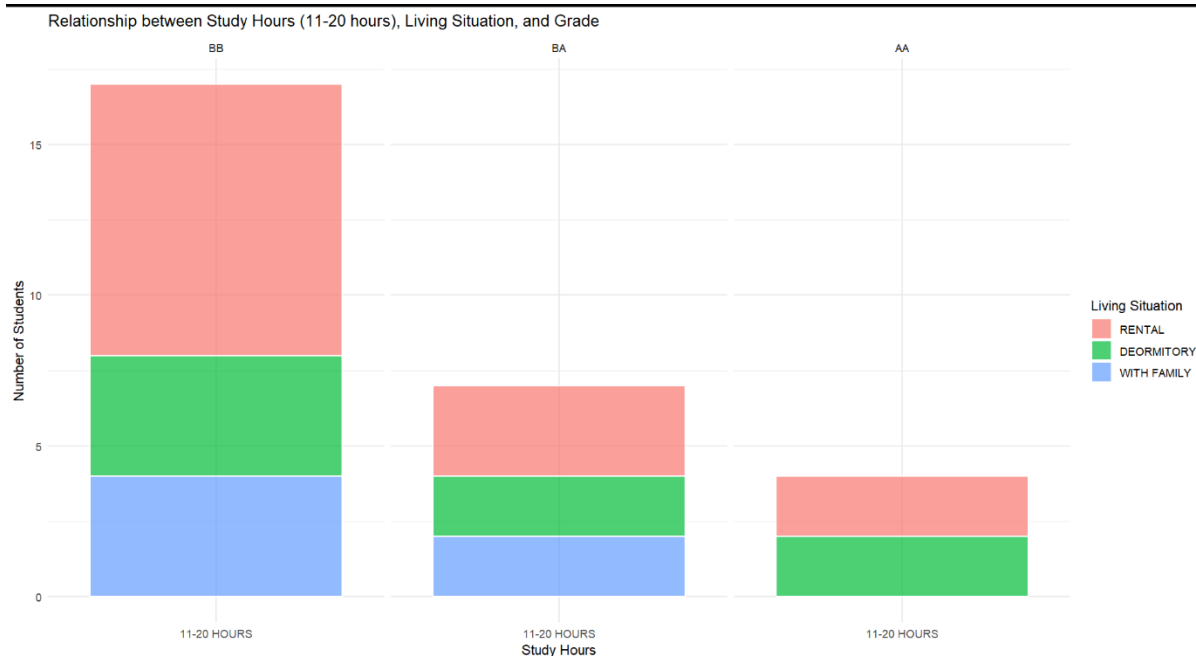


Figure 62

The graph illustrates how students' living situations affect their grades and how many hours they study. Compared to students who live in dorms or with family, there is a stronger positive correlation for students who rent a place to live. This shows that for students who are attempting to get good grades, living with family might be a more annoying living situation.

In summary, a student's living situation has an impact on their grade. Students who live in dorms or rent apartments and study from 11 to 20 hours a day typically earn better grades than those who live with their families.

#### Analysis 3.4.4: impact of studying 11 to 20 hours, scholarship and HS type with high grade.

In this analysis, our focus is to explore the impact of study hours, scholarship availability, and high school type on students' academic performance, specifically concerning the achievement of higher grades.

```
# Filter dataset for study hours from 11 to 20 hours and higher grades
filtered_data <- subset(Dataset_assignmet, STUDY_HRS == "11-20 HOURS" & GRADE %in% c("BB", "BA", "AA"))

# Create a grouped bar chart for study hours (11-20 hours), scholarship, HS type, and higher grades
ggplot(filtered_data, aes(x = STUDY_HRS, fill = GRADE)) +
  geom_bar(position = "dodge", width = 0.4) +
  geom_text(stat = "count", aes(label = ..count..), position = position_dodge(width = 0.4), vjust = -0.5) +
  facet_grid(SCHOLARSHIP ~ HS_TYPE) +
  labs(title = "Relationship between Study Hours (11-20 hours), Scholarship, HS Type, and Higher Grades",
        x = "Study Hours",
        y = "Number of Students",
        fill = "GRADE") +
  theme_minimal()
```

Figure 63

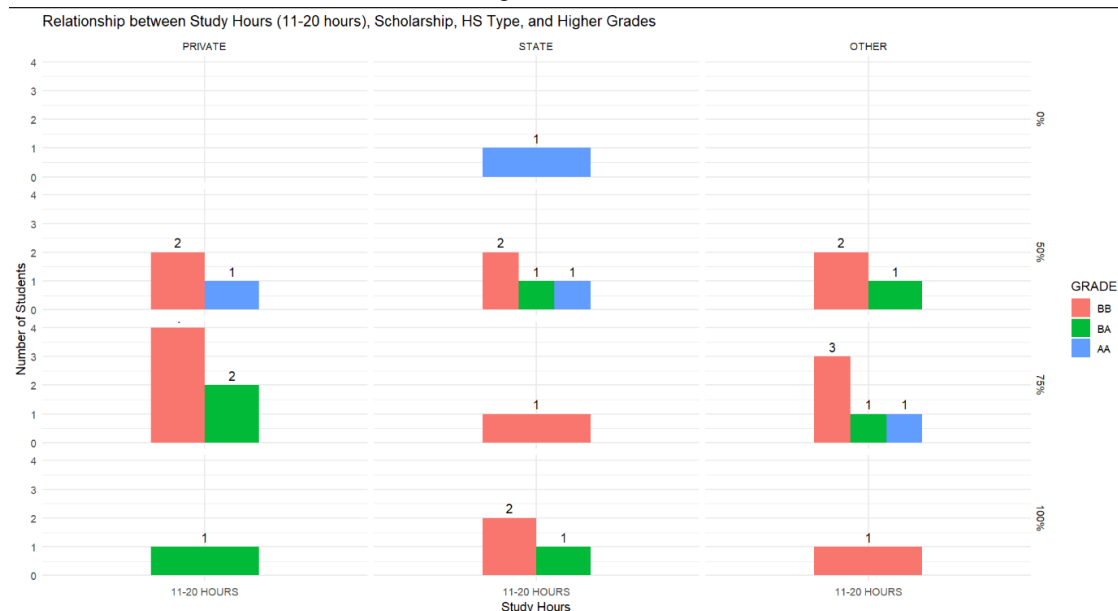


Figure 64

The graph indicates a negative relationship between grades and scholarship. Put differently, students who receive scholarships ranging from 50% to 75% typically have better grades than those who do not. However, compared to private and other schools, state school graduates typically receive higher grades.

In summary, a higher percentage scholarship does not guarantee that a student will receive good grades. Students attending state schools typically receive higher grades, even though private schools are somehow better than state and other educational institutions.

## TESTING:

```
# Load necessary library
library(gmodels) # For CrossTable function

# Assuming 'Studying 11-20 Hours' and 'Higher Grade Achievement' columns exist in Dataset_assignmet

# Create a contingency table between 'STUDY_HOURS' and 'Higher Grade Achievement'
contingency_table <- table(Dataset_assignmet$STUDY_HRS, Dataset_assignmet$GRADE)

# Perform Chi-square test of independence
chi_sq_test <- chisq.test(contingency_table)

# Display the contingency table
print(CrossTable(Dataset_assignmet$STUDY_HRS, Dataset_assignmet$GRADE))

# Display the result of the Chi-square test
print(chi_sq_test)
```

Figure 65

Dataset_assignmet\$STUDY_HRS	Dataset_assignmet\$GRADE								Row Total
	FAIL	DD	DC	CC	CB	BB	BA	AA	
NONE	1	88	76	81	72	20	2	8	348
	1.905	0.175	0.393	0.765	0.008	0.105	3.187	0.045	0.227
	0.003	0.253	0.218	0.233	0.207	0.057	0.006	0.023	
	0.062	0.237	0.211	0.250	0.229	0.244	0.069	0.211	
	0.001	0.057	0.050	0.053	0.047	0.013	0.001	0.005	
<5 HOURS	8	115	111	86	97	24	16	17	474
	1.889	0.001	0.001	1.990	0.000	0.071	5.529	2.355	0.309
	0.017	0.243	0.234	0.181	0.205	0.051	0.034	0.036	
	0.500	0.310	0.308	0.265	0.309	0.293	0.552	0.447	
	0.005	0.075	0.072	0.056	0.063	0.016	0.010	0.011	
6-10 HOURS	4	83	75	76	63	21	4	9	335
	0.073	0.048	0.166	0.389	0.453	0.534	0.860	0.059	0.218
	0.012	0.248	0.224	0.227	0.188	0.063	0.012	0.027	
	0.250	0.224	0.208	0.235	0.201	0.256	0.138	0.237	
	0.003	0.054	0.049	0.050	0.041	0.014	0.003	0.006	
11-20 HOURS	2	85	96	81	81	17	7	4	373
	0.919	0.301	0.818	0.062	0.283	0.433	0.000	2.972	0.243
	0.005	0.228	0.257	0.217	0.217	0.046	0.019	0.011	
	0.125	0.229	0.267	0.250	0.258	0.207	0.241	0.105	
	0.001	0.055	0.063	0.053	0.053	0.011	0.005	0.003	
<20	1	0	2	0	1	0	0	0	4
	22.010	0.967	1.200	0.845	0.040	0.214	0.076	0.099	0.003
	0.250	0.000	0.500	0.000	0.250	0.000	0.000	0.000	
	0.062	0.000	0.006	0.000	0.003	0.000	0.000	0.000	
	0.001	0.000	0.001	0.000	0.001	0.000	0.000	0.000	
Column Total	16	371	360	324	314	82	29	38	1534
	0.010	0.242	0.235	0.211	0.205	0.053	0.019	0.025	

Figure 66

Cell Contents
N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total
Total Observations in Table: 1534

Figure 67

The graph above examines what occurs when students study 11-20 hours per week. It counts the number of students who fall into each category. According to the findings, most students who study in this range score lower grades.

### **Conclusion of question 4:**

In conclusion, the graph's data and chi-square's test findings call into question the assumption that more study hours necessarily result in higher grades. In contradiction to popular belief, intense study efforts do not guarantee academic success.

### **EXTRA FEATURES:**

1. `gmodels` library: This serves as a kind of toolkit that provides us with specific tools to perform tests and verify items in our data. It assists in identifying any relationships or patterns.
2. `Theme minimal()`: Consider this to be our charts' decorator. It makes them look simple and clean, without too many decorations.
3. `reshape2` library: This unique collection of tools aids in data organizing. This helps us organise data that is sometimes messy so that we can better understand it.
4. `manual_fill_scale()`: Consider an example in which you wish to select the colours for a picture by yourself rather than depending on the computer to do so. You can use this function to select the colours you want to use for the different parts of your chart.

## 4.0 CONCLUSION.

To conclude, based on the extensive analysis done and as shown by all the graphs as well as the multiple Chi Square tests that have been done, the hypothesis which indicates that 60% of students who read scientific books often, prepare regularly for midterms during the semester, always actively taking notes, and having study hours between 11 and 20 hours per week, end up achieving a high output grade which is either an “AA” grade or a “BA” grade has been proven to be a false hypothesis. Each factor was investigated individually and in detail. As a result, the impact of external factors was investigated and several external factors such as the positive impact of projects on final grades as well as always attending classes and its impact on the final grade of students. This has proved that various factors claimed by the hypothesis to be related, were either not related or slightly related but not to the point where it would cause a large impact on the final grade of students.

## 5.0 WORKLOAD MATRIX.

Component:	MUHAMAD AHMAD AL MUHDAR	IBRAHEEM MOHAMMED_IMAD ELDIN AWAD	ABDULRAHMAN GAMIL MOHAMMED AHMED	ABDULELAH_HUSSEIN ABDULRAHMAN_AL-KAF	Total
Data preparation	25%	25%	25%	25%	100%
Explore the dataset and work on objectives	25%	25%	25%	25%	100%
Conclusion, introduction, and assumption	25%	25%	25%	25%	100%

## 6.0 REFERENCES.

- Biswal, A. (2023, October 11). *What is a Chi-Square Test? Formula, Examples & Application.* Simplilearn.com. <https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test>
- Pedamkar, R. (2023, March 22). *Graphs in R.* educba.com. <https://www.educba.com/graphs-in-r/>