

**GROUP ASSIGNMENT**

**TECHNOLOGY PARK MALAYSIA**

**CT127-3-2-PFDA**

**PROGRAMMING FOR DATA ANALYSIS**

**LECTURER NAME** **:** **NEOH HOO THYE**

**HAND OUT DATAE** **:** **2 OCTOBER 2023**

**HAND IN DATE** **:** **21 NOVEMBER 2023**

**WEIGHTAGE** **:** **50%**

|  |  |  |  |
| --- | --- | --- | --- |
| **Student ID** | **Student Name** | **Signature** | **Role** |
| TP070586 | Mohammad Abu Huzaifa | Huzaifa | Group Leader |
| TP066079 | Mansimran Kaur Sohi |  | Group Member |
| TP066035 | Janani A/P Munaindy | Jay | Group Member |

**Instruction To Candidates:**

1. **Students are advised to underpin their answers with the use of references (cited using the American Psychological Association (APA) referencing).**
2. **Late submission will be awarded zero (0) unless extenuating Circumstances (EC) are upheld.**
3. **Cases of plagiarism will be penalized.**
4. **Submit the assignment to APU Learning Management System.**
5. **You must obtain 50% overall to pass this module.**

# ABSTRACT

This study explores the relationship between students' living arrangements and academic performance, with a focus on those residing with their families. Grounded in several assumptions, the research suggests that familial living provides a supportive, stable environment conducive to academic success. It assumes parents play a vital role in offering encouragement and guidance, and financial stability enables students to concentrate on studies. Cultural values, social support, access to resources, routine, motivation, and personal responsibility are believed to collectively contribute to academic achievement in family-living settings. It is acknowledged, however, that these assumptions may not universally apply due to individual variations and diverse cultural backgrounds.

The primary objective of this detailed analysis is to identify patterns influencing students' academic performance. The study delves into various aspects of students' personal lives, including lifestyle choices, academic and familial influences, and demographics. Specifically, the investigation aims to assess the hypothesis that students living with their families achieve higher academic success. To achieve this, the research examines the potential impact of parental education levels, cultural beliefs, family support, and financial stability on students' educational outcomes. The study also scrutinizes students' study methods, class attendance, reading habits, and exam preparation, along with the influence of contemporary instructional practices. Predictive studies utilizing indicators like cumulative grade point averages provide a comprehensive evaluation of academic progress.

The aim of this research is to provide insightful information for policymakers and educational institutions. By understanding the intricate interactions between specific circumstances and broader systemic variables, the study aims to assist in creating an academic environment that fosters student achievement. This approach ensures that recommendations are grounded in a thorough understanding of the complex factors influencing students' academic success.

[ABSTRACT 2](#_Toc151495036)

[1.0 Introduction 5](#_Toc151495037)

[1.1 Data Description 5](#_Toc151495038)

[1.2 Assumptions 7](#_Toc151495039)

[1.3 Objectives 7](#_Toc151495040)

[2.0 Methods and Tools 8](#_Toc151495041)

[3.0 Data Preparation 8](#_Toc151495042)

[3.1 Data Import 8](#_Toc151495043)

[3.2 Data Exploration 9](#_Toc151495044)

[3.2.1 Print The Data Types 9](#_Toc151495045)

[3.2.2 Print First Six Rows 10](#_Toc151495046)

[3.2.3 Print Last n Number of Rows 10](#_Toc151495047)

[3.2.4 Print All Column Names 11](#_Toc151495048)

[3.2.5 Print Number of Rows & Columns 11](#_Toc151495049)

[3.2.6 Explore Structure of Data Frame Columns 13](#_Toc151495050)

[3.2.7 Calculate Descriptive Statistics 13](#_Toc151495051)

[3.3 Data Cleaning/Pre-processing 14](#_Toc151495052)

[3.3.1 Checking Missing Values 14](#_Toc151495053)

[3.3.2 View All Column Name 15](#_Toc151495054)

[3.3.3 Rename Columns Name 15](#_Toc151495055)

[3.3.4 Format Missing Values 16](#_Toc151495056)

[3.3.5 Remove Rows With Missing Values 17](#_Toc151495057)

[3.3 Validation 17](#_Toc151495058)

[3.3.1 Validating the use of R Packages 17](#_Toc151495059)

[3.3.2 VALIDATING MISSING VALUES 18](#_Toc151495060)

[4.0 Data Analysis 19](#_Toc151495061)

[4.1 How does the student’s attendance affect their output grade? (TP070586) 19](#_Toc151495062)

[4.1.1 Analysis-1: Univariate 19](#_Toc151495063)

[4.1.2 Analysis-2: Bivariate 20](#_Toc151495064)

[4.1.3 Analysis-3: Multivariate 22](#_Toc151495065)

[4.1.4 Analysis-4: Chi-Square Test 23](#_Toc151495066)

[4.1.5 Analysis-5: Correlation(Polychoric) 24](#_Toc151495067)

[4.2 How does student’s taking notes affect their output grade? (TP066079) 25](#_Toc151495068)

[4.2.1 Analysis 1: Univariate 25](#_Toc151495069)

[4.2.2 Analysis 2: Bivariate 26](#_Toc151495070)

[4.2.3 Analysis 3: Multivariate 27](#_Toc151495071)

[4.2.4 Analysis 4: Chi-Square Test 29](#_Toc151495072)

[4.2.5 Analysis 5: Correlation (polychoric) 30](#_Toc151495073)

[4.3 How does reading journals can affect student's grades? (TP066035) 31](#_Toc151495074)

[4.3.1 Analysis-1: Univariate 31](#_Toc151495075)

[4.3.2 Analysis-2: Bivariate 31](#_Toc151495076)

[4.3.3 Analysis-3: Multivariate 32](#_Toc151495077)

[4.3.4 Analysis-4: Chi-Square Test 34](#_Toc151495078)

[4.3.5 Analysis-5: Correlation Analysis 35](#_Toc151495079)

[5.0 Extra Features 37](#_Toc151495080)

[5.1 Decision Tree 37](#_Toc151495081)

[5.2 KNN 38](#_Toc151495082)

[5.3 NB 40](#_Toc151495083)

[6.0 Conclusion 41](#_Toc151495084)

[6.1 Interpret the result 41](#_Toc151495085)

[6.2 Recommendation 42](#_Toc151495086)

[6.3 Limitation and future direction 42](#_Toc151495087)

[Workload Matrix 43](#_Toc151495088)

[References 43](#_Toc151495089)

# Introduction

## 1.1 Data Description

This collection of data includes student information from a Cyprus-based organization that may be used to identify unreported problems and merits in students' performance on the final test of the semester. The dataset carries approximately 1535 student’s information under 33 segments. This extensive and diverse dataset offers a great starting point for investigating relationships, seeing trends, and deriving insightful knowledge about the elements that support academic achievement for students.

The dataset provides a comprehensive knowledge of students' academic achievement, lifestyle choices, educational backgrounds, and demographics by encapsulating a multitude of characteristics. The information in the report includes things like age, gender, kind of high school graduation, scholarships awarded, further work done, and participation in regular athletic or creative endeavours. It also examines aspects such as family dynamics, parental education levels, parental employment, and the number of siblings in addition to personal details. It provides insight on how students study, including how long they spend studying each week, how often they read both scientific and non-scientific sources, whether they attend conferences and seminars, and how they prepare for tests. The collection expands to include housing and travel choices as well as other elements of academic life. It also assesses how much a student's projects and activities contribute to their total performance, attendance in lectures and seminars, and use of contemporary teaching techniques like the flip classroom. Metrics measuring academic success are also important; these include predicted CGPA at graduation and average CGPA from the previous semester and mention the current semester output grade what they convey in the exams.

**The structure of the dataset:**

* **Age:** Student Age.
* **Gender:** Sex.
* **HS Type:** Graduated high-school type.
* **Scholarship:** Scholarship type.
* **Work:** Additional work.
* **Activity:** Regular artistic or sports activity.
* **Partner:** Whether he/she has a partner or not.
* **Salary:** Overall salary if it is accessible.
* **Transport:** The University’s transport.
* **Living:** A different kind of lodging or accommodation
* **Mother Education:** Mother education.
* **Father Education:** Father education.
* **Siblings:** Number of brothers/sisters (if available).
* **Marital status:** Marital status of the student.
* **Mother Job:** Mother occupation.
* **Father Job:** Father occupation.
* **Study HRS:** Weekly study hours.
* **Read FREQ:** Regularly reading (non-scientific books/journals).
* **Read FREQ SCI:** Regularly reading (scientific books/journals).
* **Attend DEPT:** Participation in department-related seminars and conferences.
* **Impact:** The effect of your efforts and pursuits on your achievements.
* **Attend:** Participation in class.
* **PREP Study:** The way of taking preparation to midterm exams (alone or group study).
* **PREP Exam:** Allocating time for studying for midterm examinations.
* **Notes:** Engaging in class note-taking.
* **Listens:** Paying attention in lecture class.
* **Attending Discussion:** Interest to attend discussion.
* **Classroom:** Flip-classroom.
* **CGPA:** CGPA of the last semester.
* **Expected CGPA:** Expected CGPA in the graduation.
* **Course ID:** Course ID.
* **Grade:** The result in the current semester.

## 1.2 Assumptions

Students living with their families tend to perform better academically, based on a few assumptions. Living with family is seen as supportive for studying, providing a stable and caring environment. The hypothesis also assumes that parents play a significant role, offering encouragement and guidance in academic pursuits. Financial stability is considered a factor, as it allows students to focus more on studies without worrying about money. Cultural values related to family living are thought to influence how students approach their studies. Additionally, a combination of factors like social support, access to resources, routine, motivation, and personal responsibility is seen as contributing to the academic success of students living with their families. It's important to note that these ideas may not apply universally, as individual situations and cultural backgrounds can influence the relationship between living arrangements and academic performance.

## 1.3 Objectives

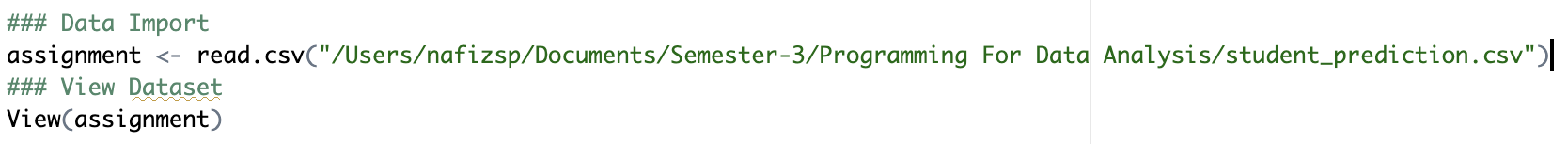
This detailed analysis delves into the variety of aspects that contribute to students' overall academic performance. The purpose of this research is to identify patterns in different aspects of the personal lives of students, such as the choices they make regarding their lifestyle, the academic influences they are subjected to, the familial influences they are subjected to, and their demographics. The research endeavours to shed light on the dark world of academic achievement by dissecting the connections that exist between a variety of factors, such as age, gender, educational experience, financial support, and other variables. The primary purpose of the inquiry is to determine whether or not the hypothesis that students who live with their families have higher academic achievement is supported by the data. In order to accomplish this, it is necessary to investigate the potential consequences, on the educational outcomes of students, of the levels of parental education, cultural beliefs, family support, and financial stability. The research also investigates the methods that students use to study, such as the number of hours that they devote to this activity on a weekly basis, the frequency with which they read journals, whether or not they attend classes or discussions, and how they prepare for exams. In addition to these features, the study investigates the impact that a variety of factors, such as participation in class discussions, attendance, and the application of contemporary instructional practices, have on academic attainment. Predictive studies, which make use of indicators like cumulative grade point average from previous semesters and predicted CGPA at graduation, make it possible to conduct a more thorough evaluation of academic progress. The end goal is to provide insightful information that will enable politicians and educational institutions to establish an atmosphere that fosters academic achievement by acknowledging the intricate interactions between specific situations and more general systemic variables. This will be accomplished by providing information that will aid them.

# 2.0 Methods and Tools

The dataset will explored and analysed using an integrated development environment for the R programming language called RStudio .For this project, we used RStudio to analyse the dataset which is student\_prediction.csv. For analysing, we used different types of packages such as ‘ggplot2’, ‘dplyr’, ‘plyr’, ‘polycor’, ‘scales’, ‘remotes’, ‘hadley’, ‘vcd’, ‘rpart.plot’, ‘caret’, ‘lattice’ etc. And then we use lot of function to more visualise the data such as ‘library’, ‘geom\_bar’, ‘geom\_point’, ‘geom\_box’, ‘plot’, ‘aes’, ‘labs’, ‘facet\_wrap’, ‘polychor’, ‘hist’, ‘mosaic’, ‘print’, ‘summary’ etc.

# 3.0 Data Preparation

## 3.1 Data Import



*Figure 3.1.1: Source Code of Data Import*

In figure 3.1.1, the .csv file is imported and stored in the variable of ‘student\_prediction.csv’ where ‘assignment’ is the name assigned to the variable that will store the data read from the CSV (Comma-Separated Values) file. This ‘<-’ is the assignment operator, used to set the variable on the left to the output of the ‘read.csv’ function. A function in R called ‘read.csv’ is used to read data from a CSV file. ‘/Users/nafizsp/Downloads/PFDA Assignment-2309/student\_prediction.csv’ is the file path specifying the location of the CSV file to be read. ‘View(assignment) is the line of code opens a data viewer for the variable ‘assignment’. The data frames in R can be seen in a spreadsheet-like format using the 'View' function. It allows you to interactively explore the contents of the data. Figure 3.1.2 is the view of the dataset in RStudio.

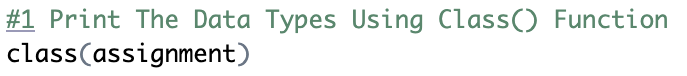
A screenshot of a computer

Description automatically generated

*Figure 3.1.2: View the Dataset in RStudio*

## 3.2 Data Exploration

### 3.2.1 Print The Data Types



*Figure 3.2.1.1: Source Code of Print Data Types*

In figure 3.2.1.1, the ‘class()’ function is used to determine the class of an object. In this Case, it is applied to the object ‘assignment’ to find out what type of data it contains. That is the outcome of the data kinds, as shown in Figure 3.2.1.2.

A close up of a word

Description automatically generated

*Figure 3.2.1.2: The result of Data Types*

### 3.2.2 Print First Six Rows



*Figure 3.2.2.1: Source Code of Print First Six Rows*

The 'head()' method is used to show the first rows of a data structure in Figure 3.2.2.1. The system shows the top six rows by default if no number is supplied in the brackets. Here, it’s applied to the object ‘assignment’ to display the initial rows of the data frame. That is the outcome of the first six rows, as shown in Figure 3.2.2.2.

A table of numbers and words

Description automatically generated with medium confidence

*Figure 3.2.2.2: The result of First Six Rows*

### 3.2.3 Print Last n Number of Rows



*Figure 3.2.3.1: Source Code of Print Last n Number of Rows*

Figure 3.2.3.1 shows how to get the final few rows of a data structure using the 'tail()' method. In this case, it’s applied to the object ‘assignment’, and ‘5’ specifies that the last five rows of the data frame will be displayed. Figure 3.2.3.2 shows, that is the result of the last five rows of data.

A table of numbers and words

Description automatically generated with medium confidence  
*Figure 3.2.3.2: The result of Last n Number of Rows*

### 3.2.4 Print All Column Names



*Figure 3.2.4.1: Source Code of Print All Columns*

In figure 3.2.4.1, the ‘names()’ function is used to retrieve the names of columns within a data frame. Applied to the object ‘assignment’, this function fetches and displays the column names present in the data frame. Figure 3.2.4.2 shows, that is the result of all the columns.

A close up of a text

Description automatically generated

*Figure 3.2.4.2: The result of All Columns*

### 3.2.5 Print Number of Rows & Columns



*Figure 3.2.5.1: Source Code of Print Number of Rows & Columns*

In figure 3.2.5.1, the ‘dim()’ function in R stands for “dimension”. It’s primarily used to retrieve the dimensions (the number of rows and columns) of the object like matrices and data frames. The ‘assignment’ represents the object whose dimensions are being queried. When ‘dim()’ is applied to a data frame like ‘assignment’, it returns a vector containing two elements: the quantity of rows and the quantity of columns, correspondingly. Figure 3.2.5.2 shows, that is the result of the number of rows and columns.

A close up of a number

Description automatically generated

*Figure 3.2.5.2:The result of Number of Rows and Columns*

  
*Figure 3.2.5.3: Source Code of Print Rows Count*

In figure 3.2.5.3, the ‘nrow(assignment)’ is a function call to ‘nrow()’ with ‘assignment’ as its argument. The ‘nrow()’ function in R specifically designed to count and return the number of rows present in the specified data frame, which in this case is ‘assignment’. Figure 3.2.5.4 shows, that is the result of the number of rows.



*Figure 3.2.5.4: The result of Rows Count*



*Figure 3.2.5.5: Source Code of Columns Count*

In figure 3.2.5.5, the ‘ncol()’ is an R function used to count the number of columns in a data structure, such as a data frame. The ‘ncol(assignment)’ is specifically using the ‘ncol()’ function to determine the number of columns within the object ‘assignment’. Figure 3.2.5.6 shows, that is the result of the number of columns.



*Figure 3.2.5.6: The result of Columns Count*

### 3.2.6 Explore Structure of Data Frame Columns

  
*Figure 3.2.6.1: Source Code of Explore Structure of Data Frame Columns*

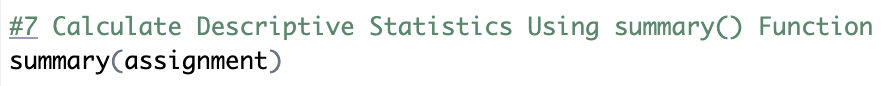
In figure 3.2.6.1, the ‘str()’ is the function of R used to display the structure of an object. It stands for “structure”. And, ‘assignment’ is the object whose structure I want to examine. Figure 3.2.6.2 shows, that is the structure of data frame columns.

A table of numbers and symbols

Description automatically generated with medium confidence

*Figure 3.2.6.2: The result of Structure of Data Frame Columns*

### 3.2.7 Calculate Descriptive Statistics



*Figure 3.2.7.1: Source Code of Calculate Descriptive Statistics*

In figure 3.2.7.1, the ‘str()’ is the function in of used to generate summary statistics or information about an object’s contents. And, the ‘assignment’ is the object whose summary I want to obtain. After execute the ‘summary(assignment)’, Figure 3.2.7.2 displays the outcome of descriptive statistics calculation.

A screenshot of a computer

Description automatically generated

*Figure 3.2.7.2: The result of Calculate Descriptive Statistics*

## 3.3 Data Cleaning/Pre-processing

### 3.3.1 Checking Missing Values

A close-up of a white background

Description automatically generated

*Figure 3.3.1.1: Source Code of Checking Missing Values*

The 'is.na()' method detects whether any values in the 'assignment' object are missing (NA) and produces a logical vector in which 'TRUE' indicates missing values and 'FALSE' indicates non-missing values, as shown in figure 3.3.1.1. The ‘sum()’ function calculates the sum of ‘TRUE’ values, representing the identified missing values through ‘is.na()’. Figure 3.3.1.2 shows, the outcome of verifying missing values.

A close-up of a logo

Description automatically generated

*Figure 3.3.1.2: The result of verifying Missing Values*

### 3.3.2 View All Column Name

A close-up of a white background

Description automatically generated

*Figure 3.3.2.1: Source Code of View All Columns Name*

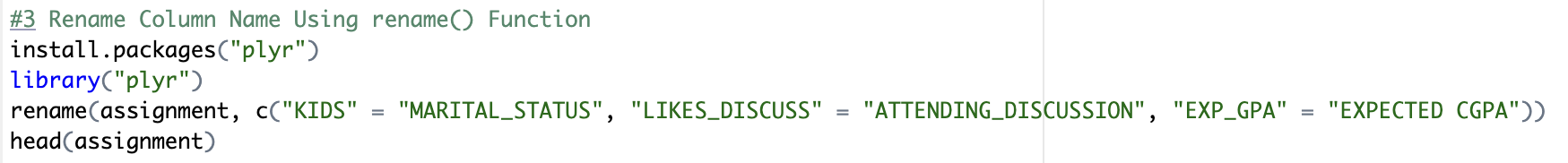
In figure 3.3.2.1, the ‘names()’ is the function of R used to obtain the column names (variable names) of a data frame. ‘assignment’ is the object assumed to be a data frame whose column names I want to retrieve. Figure 3.3.2.2 shows, that is the result of all columns name.

A close up of text

Description automatically generated

*Figure 3.3.2.2: The result of All Columns Name*

### 3.3.3 Rename Columns Name



*Figure 3.3.3.1: Source Code of Rename Columns Name*

In figure 3.3.3.1, ‘install.packages(“plyr”)’ is the command that installs the ‘plyr’ package if it’s not already installed in your R environment. Packages in R contain sets of functions and data sets that can be used for the various purposes. ‘library(“plyr”)’ is the line that loads the ‘plyr’ package into the current R session. The ‘library()’ function is used to make functions from the specified package available for use in the current R session. ‘rename(assignment, c(“KIDS” = “MARITAL\_STATUS”, “LIKES\_DISCUSS” = “ATTENDING\_DISCUSSION”, “EXP\_GPA” = “EXPECTED\_CGPA”))’ is the line part where column names in the assignment data frame being renamed. The ‘rename()’ function from the ‘plyr’ package is used to rename columns in a data frame. Within ‘rename()’, the first argument is the data frame (‘assignment’ in this case). The second argument is a vector specifying the renaming scheme. It uses the format ‘c(“old\_column\_name” = “new\_column\_name”,….). The ‘head(assignment)’ is used to display the first few rows of the ‘assignment’ data frame after the column names have been renamed. The ‘head()’ function in R used to view the initial rows of a data structure. Figure 3.3.3.2 shows, that is the result of renamed column names.

A white background with black numbers

Description automatically generatedA white sheet with black numbers

Description automatically generated A white background with black numbers

Description automatically generated

*Figure 3.3.3.2: The result of Rename Columns Name*

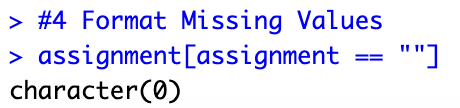
### 3.3.4 Format Missing Values

A close-up of a sign

Description automatically generated

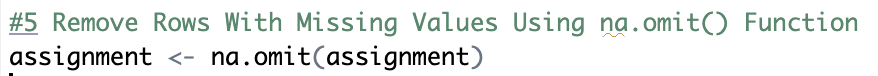
*Figure 3.3.4.1: Source Code of Format Missing Values*

In Figure 3.3.4.1, ‘assignment’ is the object being referenced, likely a data frame, where the aims to identify empty strings. The ‘[assignment == “”]’ is a part of the performs a logical comparison within the ‘assignment’ object. The ‘assignment == “”’ checks every element of the ‘assignment’ data frame to see if it’s an empty string (‘“”’). Figure 3.3.4.2 shows, that is the result of missing values.



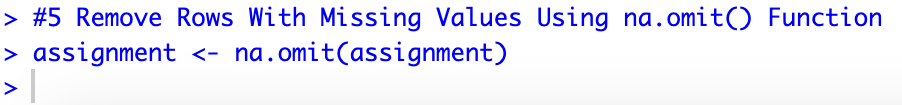
*Figure 3.3.4.2: The result of Missing Values*

### 3.3.5 Remove Rows With Missing Values



*Figure 3.3.5.1: Source Code of Remove Rows With Missing Values*

In Figure 3.3.5.1, is line of code performs the removal of rows containing missing values (NA) within the ‘assignment’ data frame. The ‘na.omit()’ is a function in R used to omit/remove rows with missing values (NA) from a data frame. ‘assignment’ is the data frame to which this operation is applied. The same data frame is reassigned with the rows containing missing values removed. Figure 3.3.5.2 shows, that is the result of remove rows with missing values.



*Figure 3.3.5.2: The result of Remove Rows With Missing Values*

## 3.3 Validation

### 3.3.1 Validating the use of R Packages

**A white background with black text

Description automatically generated**

*Figure 3.3.1.1: Validating “plyr” package in the R code*

The required R Packages are successfully installed. Below is the source code for above result:

A screen shot of a computer program

Description automatically generated

*Figure 3.3.1.2: Source Code*

### 3.3.2 VALIDATING MISSING VALUES

A close-up of a logo

Description automatically generated

*Figure 3.3.2.1: Validating Missing Values*

The outcome of validating missing values is displayed in the image above. After the missing values were first verified, missing values=0L was the outcome. “NA” was then used in place of the "0L." After replacing “0L” with “NA”, the result shows that our dataset contains 0 missing values (NA) which also means that 0 items in the dataset that were initially recorded as 0L have been changed to NA. Since many statistical and machine learning techniques may not handle missing data properly, handling missing values is essential for data analysis. The source code used to create the aforementioned result is shown below:

A computer screen shot of a white background

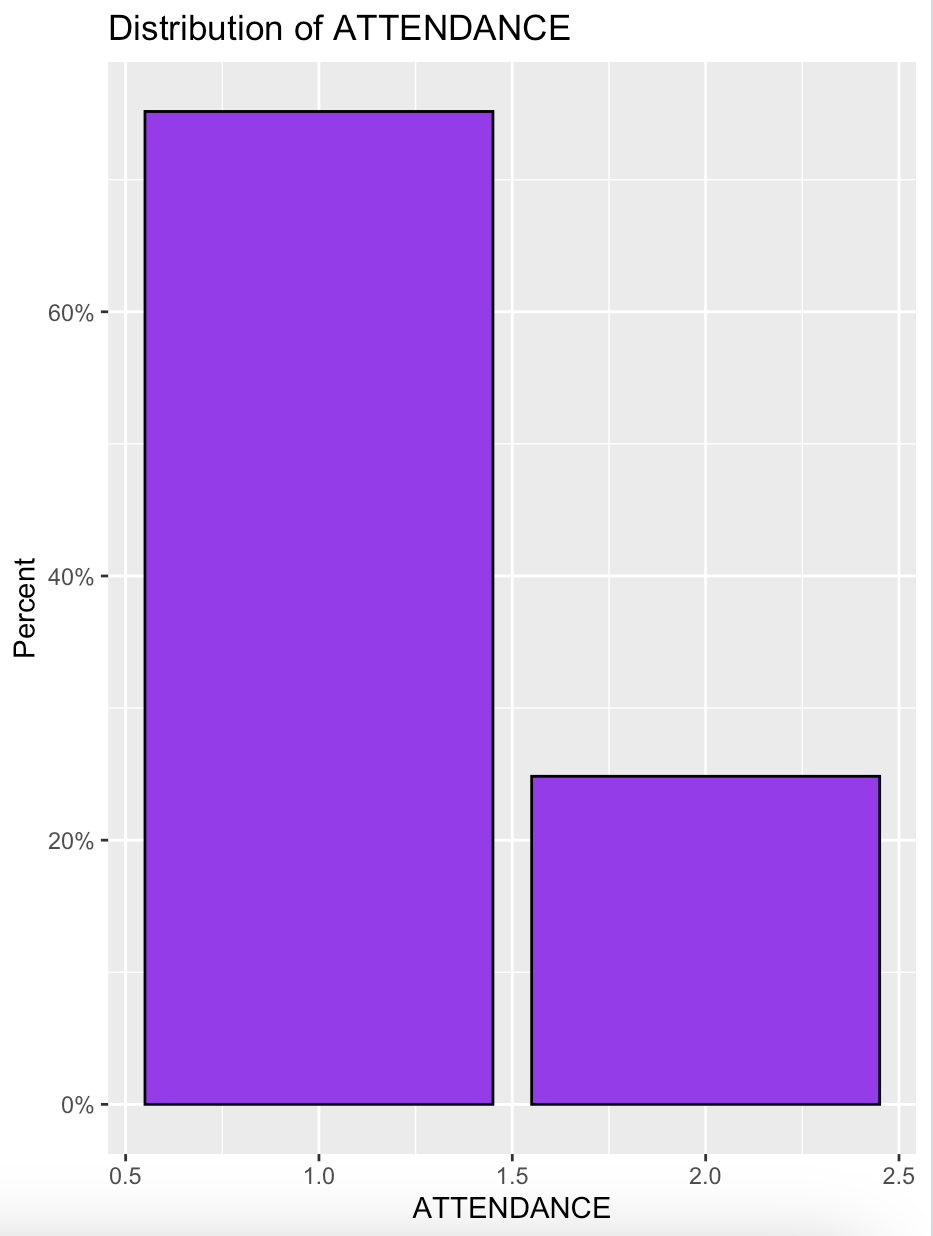
Description automatically generated

*Figure 3.3.2.2: Source Code*

# 4.0 Data Analysis

## 4.1 How does the student’s attendance affect their output grade? (TP070586)

### 4.1.1 Analysis-1: Univariate



*Figure 4.1.1.1: Distribution of ATTENDANCE (UNIVARIATE)*

The attendance distribution is shown as a percentage on the graph. The student's attendance rate is shown on the x-axis, while the proportion of students who have a certain attendance rate is shown on the y-axis.

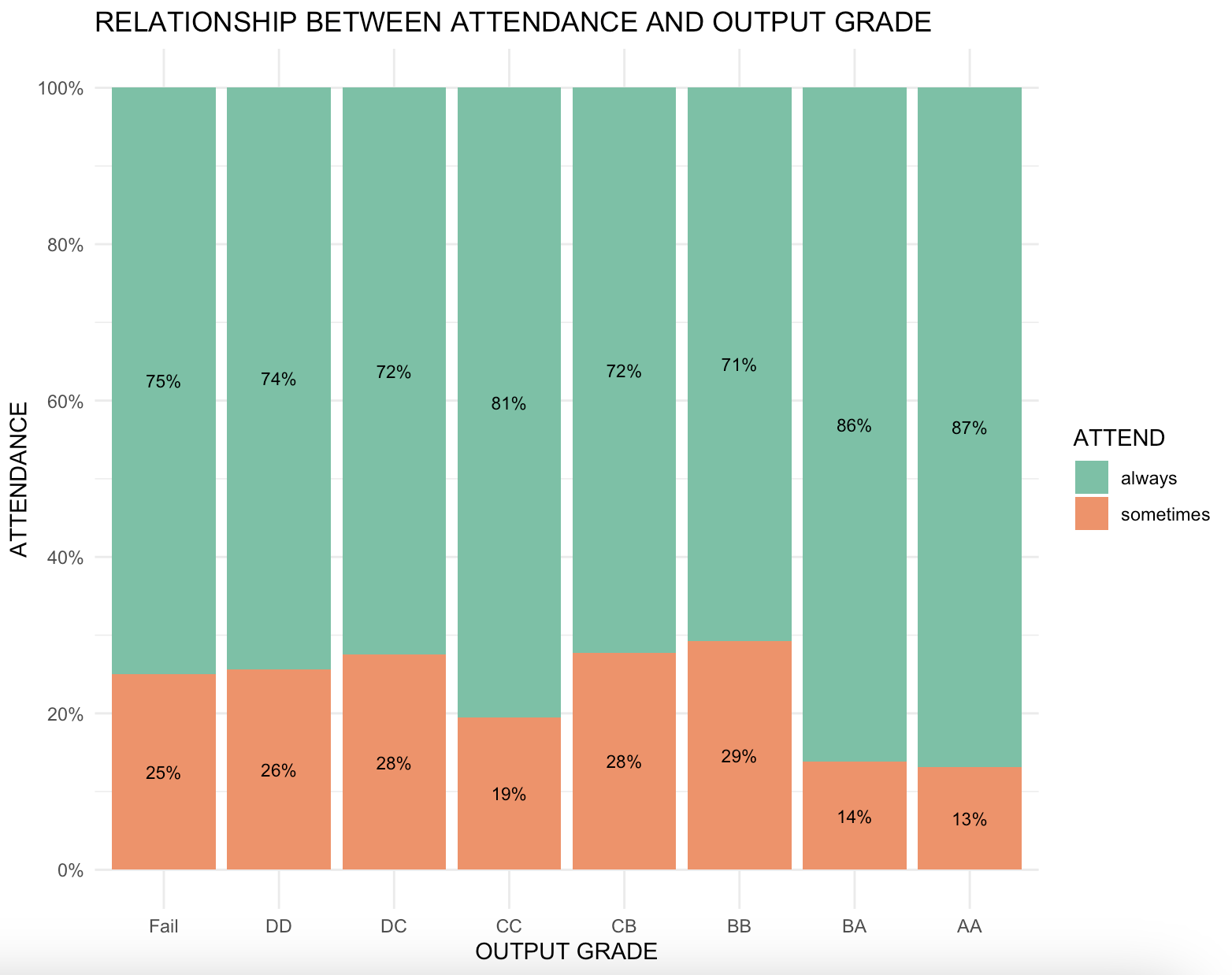
Additionally, the data demonstrates that most students attend between 60% and 100% of classes. The graph's apex, at 75%, indicates that more pupils than any other rate have attendance rates of 75%. Additionally, the graph demonstrates that a tiny percentage of pupils have very high and extremely low attendance rates. The graph's tails extend to 100% and 0% of students in attendance, however the proportion of pupils at these extremes is rather small. The proportion of students who are meeting or surpassing attendance requirements as well as the percentage of students who are not reaching expectations may both be seen on the graph.

A computer screen shot of a program code

Description automatically generated

*Figure 4.1.1.2: Source Code of Univariate Analysis*

### 4.1.2 Analysis-2: Bivariate



*Figure 4.1.2.1: Relationship between attendance and output grade (Bivariate)*

The association between output grade and attendance is seen in the graph. The proportion of students that attend class is shown by the x-axis, and the percentage of students who get a certain output grade is represented by the y-axis.

Additionally, the graph demonstrates a significant positive link between output grade and attendance. This indicates that there is a tendency for output grade to rise along with attendance. This is probably due to the fact that frequent class attendees are more likely to retain the information, finish their assignments, and take part in discussions. The graph also demonstrates that as attendance rises from 0% to 50%, it does so at a faster rate than it does from 50% to 100% in terms of output grade improvement. This implies that for difficult students in particular, attendance is crucial. The graph may be used to comprehend how crucial attendance is to the academic achievement of students. The graph may be used, for instance, to demonstrate to parents and students how attendance affects output grade. Students that have poor attendance rates and are thus at danger of falling may also be identified using the graph.

A screenshot of a computer code

Description automatically generated

*Figure 4.1.2.2: Source Code of Bivariate analysis*

### 4.1.3 Analysis-3: Multivariate

A screenshot of a graph

Description automatically generated

*Figure 4.1.3.1: Relationship between GRADE, ATTEND, and LIVING (Multivariate)*

The link between two variables—attendance and output grade—is shown as a scatter plot on the graph. Attendance is shown on the x-axis, while output grade is plotted on the y-axis. Individual students are represented by the dots on the graph, and each dot's location indicates the student's and output grade.

The output grade and attendance have a positive link, as shown in the graph. This indicates that there is a general tendency showing better output grades for students who attend class more often. There is considerable variation in the statistics, however, since some students with high attendance rates may also have poor output grades, and vice versa. You may use the graph to see how a set of students' output grade and attendance relate to each other. The graph, for instance, may be used to determine which students have poor attendance and are thus at danger of failing. For students at varying achievement levels, the graph may also be used to examine how attendance impacts the final grade.

A computer screen shot of a code

Description automatically generated

*Figure 4.1.3.2: Source Code of Multivariate Analysis*

### 4.1.4 Analysis-4: Chi-Square Test

A screenshot of a diagram

Description automatically generated

*Figure 4.1.4.1: Chi-test between Attendance and output grade*

The image is a particular kind of table that displays the correlation between two category variables. The two variables in this instance are the outcome grade and attendance. The output grade and attendance have a substantial correlation, as the table demonstrates. Compared to students who consistently attend class, those who miss substantially more are likely to fail or get a DD grade. Pupils who consistently show up for class are more likely to get a CC grade or above. A statistical test that may be used to evaluate the degree of correlation between two category variables is the chi-squared test. This contingency table's chi-squared test is extremely significant (p < 0.001), indicating a statistically significant association between attendance and output grade.

A white background with black text

Description automatically generated

*Figure 4.1.4.2: Result of Chi-Square Test*

This is the Chi-Squared Test result. This finding raises the possibility that the factors under study do not significantly correlate with one another. The null hypothesis cannot be strongly refuted due to the p-value of 0.06184, which is more than the traditional significance limit of 0.05. To get a firm conclusion, further research or a bigger sample size may be required since it's still very near to the threshold.

A close-up of a computer code

Description automatically generated

*Figure 4.1.4.3: Source Code of Chi-Square Test*

### 4.1.5 Analysis-5: Correlation(Polychoric)

A close-up of a number

Description automatically generated

*Figure 4.1.5.1: Result of Correlation (Polychoric)*

The function polychor(x, y) yielded a value of -0.03101246, which indicates a weakly negative correlation between the variables x and y. It shows that there is a tiny tendency for one variable to drop somewhat when the other grows, although the link is weak. The results show that the polychoric correlation is -0.03101246, a rather low number that suggests no significant positive connection exists between the output grade and attendance.

A computer code with text

Description automatically generated

*Figure 4.1.5.2: Source Code of Correlation (Polychoric)*

## 4.2 How does student’s taking notes affect their output grade? (TP066079)

### 4.2.1 Analysis 1: Univariate

A graph with a bar graph

Description automatically generated

*Figure 4.2.1.1: Taking Notes Distribution (Univariate)*

From the graph, one can conclude that the highest frequency of students taking notes is between 750 and 1000. The graph peaks at 900 which means the number of students that always take notes is most of the time at this frequency. The graph also shows that there is a moderate and low frequency of student taking notes. The tails of the graph stretch out to 0% and 100% attendance, but the percentage of students at these extremes is very low which is around 33%. Certainly, this graph can be used to show the relationship between taking and output grade where the x-axis represents the frequency of students who take notes during class, while the y-axis represents the percentage of students who achieve a given output grade.



*Figure 4.2.1.2: Source Code of Univariate Analysis*

### 4.2.2 Analysis 2: Bivariate

A graph of different colored squares

Description automatically generated

*Figure 4.2.2.1: Relationship between taking notes and output grade (Bivariate)*

The graph also shows a strong positive correlation between taking notes and output grade. This means that when the frequency of taking notes increases, the output grade also increases. This is highly possible because taking notes encourages students to process information, extract key points and summarize content, promoting a deeper understanding of the subject.

The graph also shows that the students who take notes sometimes has the greatest improvement in their output grades as compared to those who always take notes and never at all. With a majority of 65% scoring A’s and B’s. On the other hand, students who always take notes have a similar percentage margin of scoring every grade. For example, the percentage of student scoring AA, BB, CC, and DD are the same which is 60% while, the students who never take notes have no improvement or success rate in scoring high grades. This graph can be used to encourage students to start somewhere by taking notes. This is because making important as opposed to every day daily notes can help them understand the subject better rather than taking very detailed notes that can cause confusion and uninteresting for them to read.

A computer code with green and white text

Description automatically generated

*Figure 4.2.2.2: Source Code of Bivariate Analysis*

### 4.2.3 Analysis 3: Multivariate

A screen shot of a chart

Description automatically generated

*Figure 4.2.3.1: Scatter plot Matrix of NOTES, LIVING, and GRADE (Multivariate)*

The scatter- plot portrays the relationship between two variables: students taking notes and output grade. The x-axis represents students taking notes, and the y-axis represents output grade. The dots on the graph represents student’s place of living, with the position of each dot representing the student’s and output grade.

The graph shows a positive correlation between taking notes and output grade. This means that there is a general trend for students who take notes in class more regularly to have higher output grades. However, there is also some variability in the data, with some students who always take notes but still having low output grades, and vice versa. The graph can be used to understand the relationship between taking notes and output grade for a group of students. For example, the graph can be used to identify students who are at risk of failing because they never take notes during class at all.

A white background with black text

Description automatically generated

*Figure 4.2.3.2: Source Code of Multivariate Analysis*

### 4.2.4 Analysis 4: Chi-Square Test

A screenshot of a computer

Description automatically generated

*Figure 4.2.4.1: Chi-test between taking notes and output grade*

The picture is a type of table that shows the relationship between two categorical variables which are taking notes and output grade. From the table, there is a strong relationship between taking notes and output grade. Students who never take notes class are much more likely to fail get a DD grade than students who sometimes or always take notes. Whereas, Students who sometimes or always take notes in class have a higher probability to get a BB grate or higher. The chi-squared test is a statistical test that can be used to assess the strength of the relationship between two categorical variables. The chi-squared test for this contingency table is highly significant (p < 0.001). This proves that the relationship between taking, and output grade is statistically significant.

A white background with black text

Description automatically generated

*Figure 4.2.4.2: Result of Chi-Square test*

This is the Chi- Squared Test result which suggests that there is a significant association between the analysed variables. The value of p is 0.0246 that is smaller than the conventional significance level of 0.05. This indicates that there is a full proof evidence to reject the null hypothesis.

A close-up of words

Description automatically generated

*Figure 4.2.4.3: Source Code of Chi-Square Test*

### 4.2.5 Analysis 5: Correlation (polychoric)

A close-up of a number

Description automatically generated

*Figure 4.2.5.1: Result of Correlation (polychoric)*

The result 0.0007648886 from the function polychor(x, y) suggests a strong positive correlation between variables x and y. It indicates that as one variable increases, the other tends to decrease. The polychoric correlation turns out to be 0.0007648886, which is a high value, and indicates that there is a strong positive association between the taking notes and output grades.

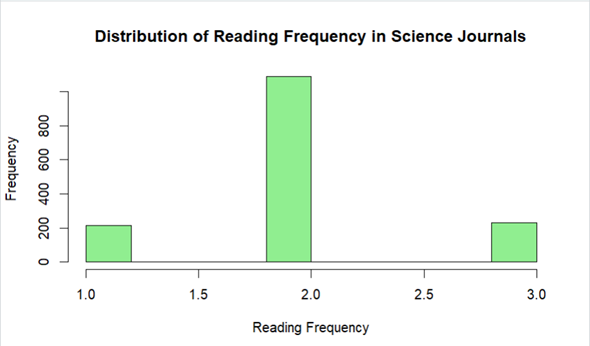
A computer code with text

Description automatically generated with medium confidence

*Figure : Source Code of Correlation (Polychoric)*

## 4.3 How does reading journals can affect student's grades? (TP066035)

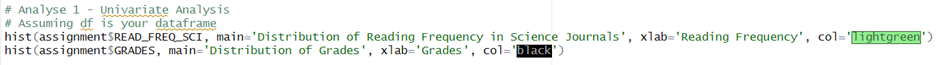
### 4.3.1 Analysis-1: Univariate



*Figure 4.3.1.1: Distribution of Reading Frequency in Science Journal*

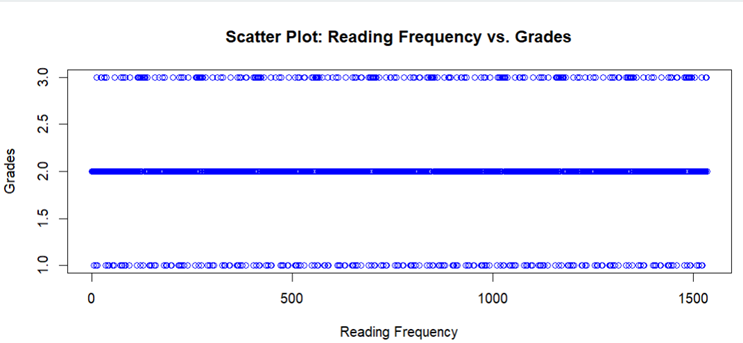
The above histogram shows the frequency of students reading scientific journals. Based on the histogram, the highest frequency is at 2 compared to the others which means that the greatest number of students have a reading frequency equal to 2.

Below is the source code used to plot the above histogram:



*Figure 4.3.1.2: Source Code of Univariate Analysis*

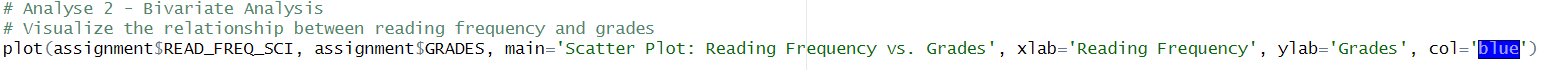
### 4.3.2 Analysis-2: Bivariate



*Figure 4.3.2.1: Scatter plot: Reading Frequency vs Grades*

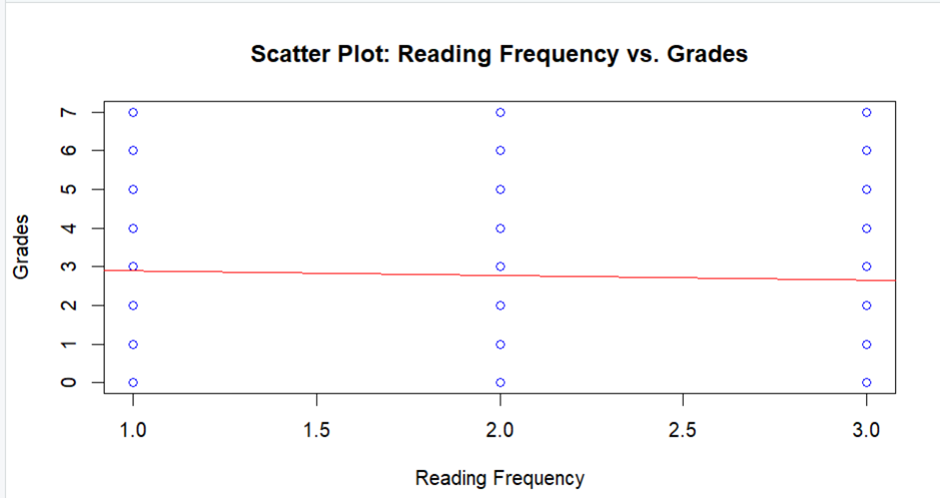
The accompanying scatterplot illustrates the correlation between students' reading habits and their final grades. The significance of students' reading habits on their final grades is made clear. Each student is represented by a point in the plot, which has reading frequency on the x-axis and grades on the y-axis. A scatter plot can be used to visually examine the relationship between reading habits and academic performance to reveal any trends or patterns. If a trend or pattern does emerge, it could point to a connection between how often you read and how well you do in school.

The source code used to plot the relationship is given below:



*Figure 4.3.2.2: Source Code of Bivariate Analysis*

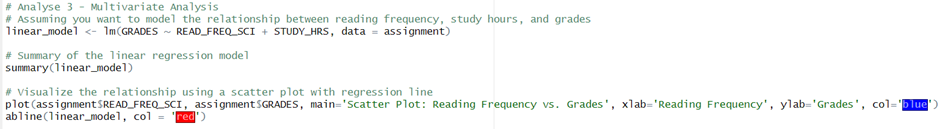
### 4.3.3 Analysis-3: Multivariate



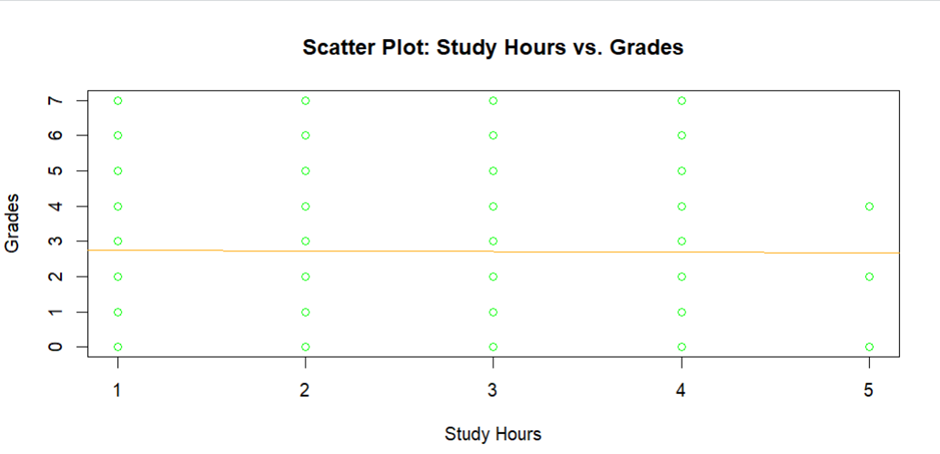
*Figure 4.3.3.1: Scatter plot: Reading Frequency vs Grades*

The related scatter plot illustrates the relationship between the grades students receive and the amount of time they spend reading. These lines of code construct a scatter plot in which reading frequency is plotted along the x-axis, grades are plotted along the y-axis, and each point represents an observation from the dataset. The regression line that is added to the scatter plot illustrates the model that the linear regression analysis creates to describe the association that exists between the amount of time spent reading and grades. The red regression line serves as a visual representation of both the direction and the strength of the linear relationship that exists between the variables.

Below is the source code used to plot the above relationship:



*Figure 4.3.3.2: Source Code of figure 4.3.3.1*



*Figure 4.3.3.3: Scatter plot: Study Hours vs Grades*

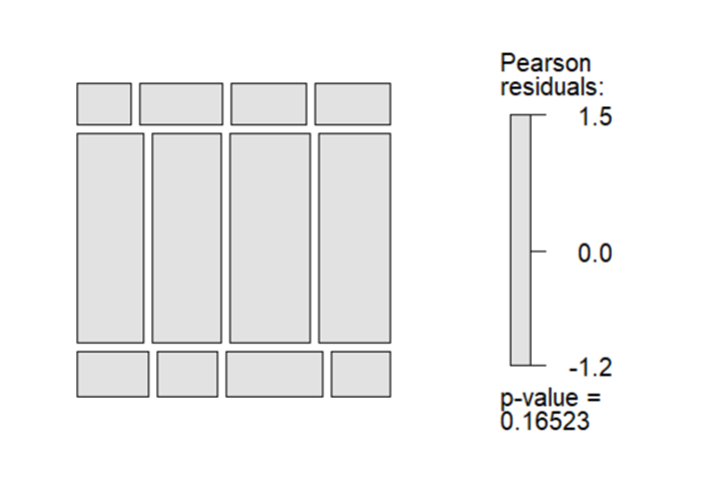
The third variable, which is the correlation between students' study hours and grades, is graphically represented here by a scatter plot. Each point on the plot represents the observation of a single student, with the y-coordinate of the point indicating the student's grades and the x-coordinate of the point indicating the number of study hours the student put in. The color green is used to represent the data points in the figure. The best-fitting linear model for the association between study hours and grades is then presented as a regression line on the scatter plot in orange, giving a visual portrayal of the trend in the data.

The source code below was used to form the plot:

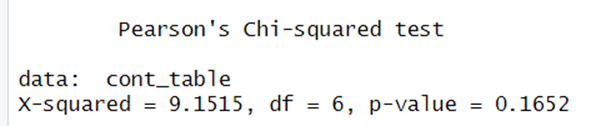


*Figure 4.3.3.4: Source Code of figure 4.3.3.3*

### 4.3.4 Analysis-4: Chi-Square Test



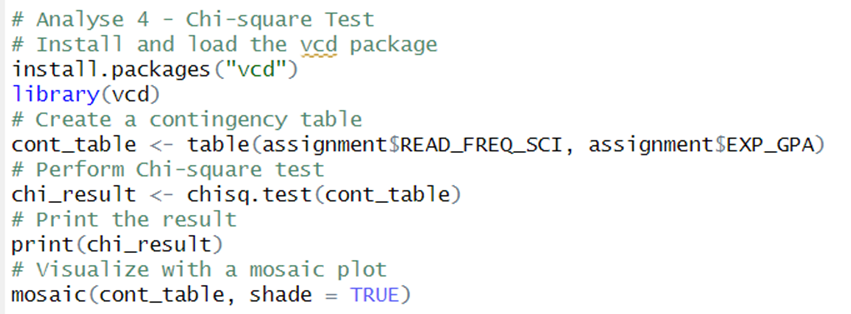
*Figure 4.3.4.1: Figure of Chi-Square test*



*Figure 4.3.4.2: Result of Chi-Square Test*

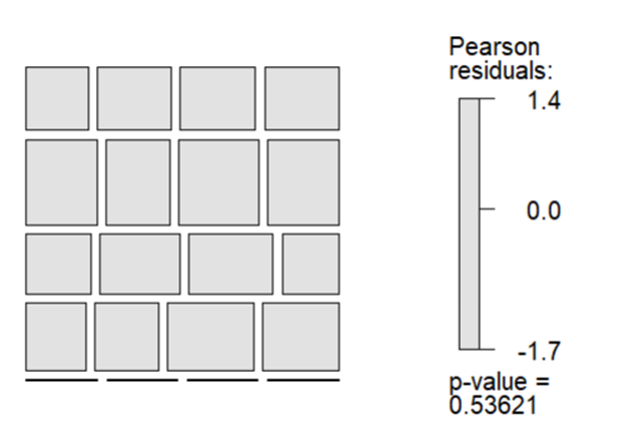
This test investigates whether or whether there is a correlation between the expected GPA and the amount of time spent reading scientific publications, both of which are categorical variables. To determine whether or not the two variables do, in fact, have a meaningful relationship with one another, a statistical test known as the chi-square test is carried out. This output includes the test statistic (X squared), the degrees of freedom (df), and the p-value. This is the statistic that will be tested. In the context of the chi-square test, X squared refers to the sum of the squares of the differences that exist between the actual counts and the predicted counts. A higher X-squared value indicates a greater degree of dissimilarity between the observed frequencies and the expected frequencies, and the degrees of freedom are determined by the number of categories present in the variables that are being analysed. In this particular scenario, there are six degrees of freedom. When analysing statistical hypotheses, the p-value is a crucial factor to consider. To put it another way, it gives an indication of the likelihood of witnessing the data in the case where there is no true link between the variables. In this particular scenario, the p-value comes in at 0.1652. When the p-value is greater, it indicates that there is less convincing evidence to reject the null hypothesis. The conventional level of significance, often known as alpha, is set at 0.05. If the p-value is less than the alpha threshold, the null hypothesis is rejected. If it is higher, you won't be able to reject the null hypothesis no matter what. Because the p-value in this scenario is more than 0.05, it is that you do not have sufficient evidence to definitively exclude the possibility that the null hypothesis is correct. According to the results of this test, there may not be a substantial association between the amount of time spent reading scientific journals and the overall grade point average that is anticipated.

The source code to form the above outcome is below:

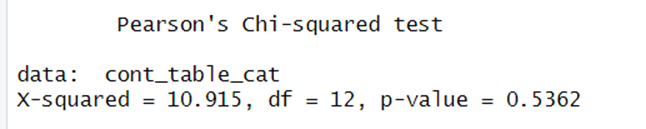


*Figure 4.3.4.3: Source Code of Chi-Square Test*

### 4.3.5 Analysis-5: Correlation Analysis



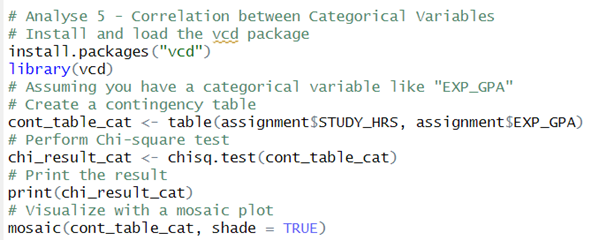
*Figure 4.3.5.1: Figure of Correlation Analysis*



*Figure 4.3.5.2: Result of Correlation Analysis*

The findings of the chi-square test, which are provided above, provide an analysis of the association between the amount of time spent studying and the anticipated grade point average. The result of the test is therefore (10.915). It determines the extent to which the observed frequencies and the predicted frequencies in the contingency table diverge from one another. A larger X-squared number is indicative of a greater difference between the two sets of data. The number of categories present in the variables that are the subject of the study is what determines the 12 degrees of freedom. In this particular scenario, the expected grade point average is lower than one and is calculated as the product of the number of different subject areas represented by the study hours. The p-value of 0.5362 indicates that the evidence is not sufficient to reject the null hypothesis. This is evidenced by the fact that the significance level of 0.05, which is typically accepted, is higher than the p-value. It is clear from this that there is not a statistically significant association between the total number of hours spent studying and the expected grade point average as a consequence of taking this exam.

Below is the given source code to get the above result:



*Figure 4.3.5.3: Source Code of Correlation Analysis*

# 5.0 Extra Features

## 5.1 Decision Tree

A screenshot of a computer

Description automatically generated

*Figure 5.1.1: Extra Feature Using Decision Tree*

The diagram shows a decision tree with a number of branches and nodes. The nodes are labelled with features or attribute values, and the branches are labelled with decisions or conclusions. The tree is used to classify data, and it works by starting at the root node and following the branches until a leaf node is reached. The leaf node contains the predicted classification for data point. The extra feature of this decision tree is that it uses a random subset of features at each node to make a decision. This is done to improve the accuracy of the tree and prevent it from overfitting the training data. Here is the more detailed explanation of the diagram:

* The root node is labelled with the feature READ\_FREQ\_SCI. This feature represents the frequency at which the term "sci" appears in the data.
* The left branch from the root node is labelled with the value >= 2. This means that if the frequency of "sci" is greater than or equal to 2, the data point is classified as DD.
* The right branch from the root node is labelled with the value < 2. This means that if the frequency of "sci" is less than 2, the data point is sent to the next node in the tree.
* The next node in the tree is labelled with the feature DC. This feature represents the number of decimal places in the data point.
* The left branch from the DC node is labelled with the value >= 1. This means that if the data point has at least one decimal place, it is classified as BB.
* The right branch from the DC node is labelled with the value < 1. This means that if the data point has no decimal places, it is sent to the next node in the tree.
* The next node in the tree is labelled with the feature CC. This feature represents the number of characters in the data point.
* The left branch from the CC node is labelled with the value >= 40. This means that if the data point has at least 40 characters, it is classified as AA.
* The right branch from the CC node is labelled with the value < 40. This means that if the data point has less than 40 characters, it is classified as BA.

The decision tree continues in this way until all of the data points have been classified. For example, at the root node, the tree could use a random subset of features such as READ\_FREQ\_SCI, DC, and CC to make a decision. This would make the tree more robust to noise in the data and would prevent it from overfitting the training data.

A close-up of a math equation

Description automatically generated

*Figure 5.1.2: Source Code of Extra Feature (Decision Tree)*

## 5.2 KNN

A screenshot of a computer

Description automatically generated

*Figure 5.2.1: Result of Extra Feature using KNN*

The picture shows the result of a KNN model that was trained on a dataset of 1534 samples and 4 predictors. The model was trained to classify into 8 classes: ‘Fail’, ‘DD’, ‘DC’, ‘CC’, ‘CB’, ‘BB’, ‘ BA’, and ‘AA’. The model was evaluated using 10-fold cross-validation. This means that the dataset was randomly divided into 10 folds, and th model was trained on 9 folds and tested on the remaining fold. This process was repeated 10 times, and the average accuracy and Kappa score were calculated. The best model was found to have a k value of 7. This means that the model used the 7 nearest neighbours of each sample to make a classification. The accuracy of this model is 0.2321, and the Kappa score is 0.0056. Overall, the KNN model achieved a moderate level of accuracy on this dataset. The Kappa score, which takes into account the agreement between the model’s predictions and the actual classifications, was also low. This suggests that the model may not be generalizable to other datasets. Here are some additional details about the results:

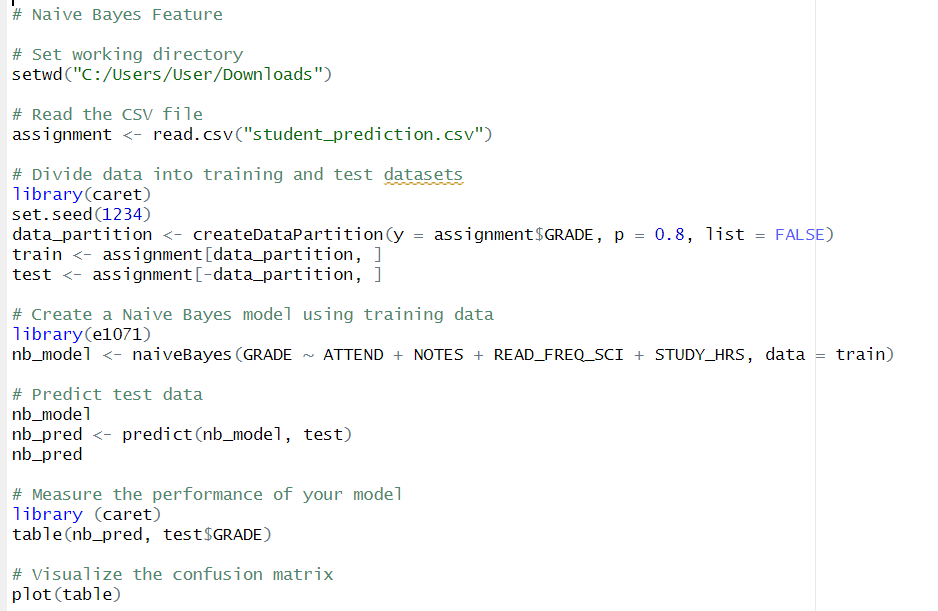
* The model was trained on a relatively large dataset of 1534 samples.
* The model was trained to classify samples into 8 classes.
* The model was evaluated using 10-fold cross-validation.
* The best model was found to have a k value of 7.
* The accuracy of the best model was 0.2321.
* The Kappa score of the best model was 0.0056.

A computer code with text

Description automatically generated with medium confidence

*Figure 5.2.2: Source Code of Extra Feature (KNN)*

## 5.3 NB



*Figure: Source Code*

In order to predict student grades based on some factors like attendance (ATTEND), notes (NOTES), reading frequency in science journals (READ\_FREQ\_SCI), and study hours (STUDY\_HRS), the aforementioned R source code has been applied with the Naive Bayes classification model. A confusion matrix is used to assess the model's performance once the dataset has been split into training and test sets. Below is the result:

A graph with a line

Description automatically generated

*Figure: Result*

# 6.0 Conclusion

## 6.1 Interpret the result

The purpose of the case study that was given to us for our assignment was to anticipate the grades that students will receive as well as their cumulative grade point average by analysing the factors that contribute to it. Throughout the paperwork, graphs were created to investigate the links between a number of different variables and the student's grade. This was done with the goal of determining which factor has a more significant bearing on the student's overall academic performance.

The factors that were discovered to have a greater influence than the others were the amount of time spent studying, attendance, the practice of taking notes, and the reading of journals. This is due to the fact that studies have shown that students who put in more effort in their academic work by studying for longer periods of time, taking notes, and participating in all of their classes and discussions are more likely to succeed academically. In addition to the factors that were discussed before, the case study places an emphasis on the role that familial influences have in determining academic achievement This is likely due to the fact that family life provides an atmosphere that is both secure and supportive.

The research also highlights the application of advanced statistical approaches, such as linear regression models, to anticipate the grades and cumulative grade point averages of students based on a number of different factors. This demonstrates a complete strategy for interpreting the complex relationships between variables that influence academic achievements, which ultimately results in a prediction model that is more accurate and sophisticated.

A Chi-square test is incorporated into the study so that more investigation may be conducted into the connections that exist between category variables. This statistical test offers illuminating information on the relationships between many aspects of students' expectations and behaviours in a variety of settings.

The thoroughness of the case study, which makes use of a variety of statistical approaches and takes into account a variety of variables, improves the nuance and accuracy of the findings produced about the prediction of students' grades and CGPA. All things considered, the case study takes into account a variety of variables and makes use of a variety of statistical techniques. The emphasis placed on statistical analysis and visual representations ensures that the variables that influence academic achievement will be investigated in depth and will be supported by empirical evidence.

## 6.2 Recommendation

In conclusion, the big collection of information from the Cyprus-based organization is really helpful for understanding the good and not-so-good parts of how students do in their studies. By looking at lots of different things like age, lifestyle choices, and school backgrounds, the organization can learn detailed insights that go deeper than just the basic facts. Studying how students learn, their family situations, and the effects of new teaching methods gives a full picture of what helps students do well in school. This understanding can be used to make specific plans that deal with problems and build on strengths, improving how students perform in their final tests. As the organization moves forward, using what they find in this information can help them make smart decisions based on evidence, making their approach to helping students succeed in school more effective and personalized. This collection of data is a valuable tool for making well-informed and impactful improvements.

## 6.3 Limitation and future direction

While the comprehensive dataset from the Cyprus-based organization offers valuable insights into students' academic performance, it is essential to acknowledge certain limitations. The data may not capture all possible factors influencing student success, and there may be unaccounted variables that play a role in academic outcomes. Additionally, the findings are context-specific to the Cyprus educational system and may not be universally applicable. Future directions for research could involve expanding the scope to include a more diverse set of institutions or collaborating with other organizations to compare findings. Furthermore, incorporating longitudinal data could provide a deeper understanding of the evolving dynamics influencing student achievement over time. Despite these limitations, the dataset remains a powerful tool for informed decision-making, and ongoing research can refine strategies to better address the complexities of student success.

# Workload Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| Task Component | Mohammad Abu Huzaifa (TP070586) | Mansimran Kaur Sohi (TP) | Janani A/P Munaindy (TP) |
| ABSTRACT | 10% | 80% | 10% |
| Introduction | | | |
| Data Description | 100% | 0% | 0% |
| Assumption | 0% | 100% | 0% |
| Objectives | 0% | 0% | 100% |
| Methods & Tools | 80% | 10% | 10% |
| Data Preparation | | | |
| Data Import | 80% | 10% | 10% |
| Data Exploratory | 80% | 10% | 10% |
| Data Cleaning/pre-processing | 80% | 10% | 10% |
| Validation | 10% | 10% | 80% |
| Individual Part | | | |
| Data Analysis | 100% | 100% | 100% |
| Extra Features | 100% | 100% | 100% |
| Conclusion | | | |
| Interpret the Result | 10% | 10% | 80% |
| Recommendation | 10% | 80% | 10% |
| Limitation and Future Direction | 10% | 80% | 10% |

# References

ZACH. (2022, April 13). *How to Perform Exploratory Data Analysis in R (With Example)*. STATOLOGY. <https://www.statology.org/exploratory-data-analysis-in-r/>

*Exploratory Data Analysis: Techniques, Best Practices & Applications*. simplilearn. (2023, May 31). <https://www.simplilearn.com/exploratory-data-analysis-article>

finnstats. (2021, April 4). *How to clean the datasets in R?*. R-bloggers. <https://www.r-bloggers.com/2021/04/how-to-clean-the-datasets-in-r/>

Ebner, J. (2022, July 19). *How to Use geom\_smooth in R*. SHARP SIGHT. <https://www.sharpsightlabs.com/blog/geom_smooth/>

haaf, sandy. (2019, April 2). *Easy multi-panel plots in R using facet\_wrap() and facet\_grid() from ggplot2*. ZevRoss. <http://zevross.com/blog/2019/04/02/easy-multi-panel-plots-in-r-using-facet_wrap-and-facet_grid-from-ggplot2/>

Kosourova, E. (2022, June 15). *How to Write Functions in R (with 18 Code Examples)*. DATAQUEST. <https://www.dataquest.io/blog/write-functions-in-r/>

Lathiya, K. (2023, November 12). *How to Use sum() Function in R*. R-LANG. <https://r-lang.com/sum-in-r/>

avcontentteam. (2022, March 17). *A Comprehensive Comparison Among Data Science Bootcamps vs Degree vs Online Courses*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2022/03/a-comprehensive-comparison-among-data-science-bootcamps-vs-degree-vs-online-course/>

Naveen. (2023, March 3). *Data Manipulation in R*. Intellipaat. <https://intellipaat.com/blog/tutorial/r-programming/data-manipulation-in-r/>

Willems, K. (2022, December). *Pipes in R Tutorial For Beginners*. datacamp. <https://www.datacamp.com/tutorial/pipe-r-tutorial>

Shah, K. (2023, April 3). *Exploratory Analysis Using Univariate, Bivariate, and Multivariate Analysis Techniques*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/>

ZACH. (2021, September 27). *How to Calculate Correlation Between Categorical Variables*. STATOLOGY. <https://www.statology.org/correlation-between-categorical-variables/>