Final Project Report

Student Performance Analysis

IS302 Semester 2 2023 (Broncos Group)

Team members

Ravneel Sewak - S11199333

Roska Takayawa - S11187423

Liu Ying - S11188808

SUPERVISOR: Dr Karuna Reddy.

Table of Contents:

Introduction

In today's educational system, the idea of student performance includes not only gaining knowledge and outstanding educational performance but also learners' overall well-being and pleasure. The interaction of internal and environmental factors that affect students' academic success is complex, involving anything from personal habits, family backgrounds, extracurricular activities, and the kind of high school they attended. For education, organizations, and people, it is crucial to grasp the complex nature of these influences. The main purpose of this study is to identify the major factors that impact the overall performance of a student.

Our overarching goal for this assignment is to identify how details such as age, gender, and the type of high school attended, as elements related to parental education, the individual's academic pursuits, and other contributing aspects like extracurricular activities, affect a student’s performance, either by themselves or by combinations of different factors. To achieve our results, we utilized a variety of statistical approaches in our analysis, such as data visualization strategies, descriptive statistics, and the application of prediction models like multiple linear regression and ordinal logistic regression.

There are plenty of benefits that the analysis provides for students, politicians, and educational institutions. It makes it easier for educational institutions to provide individualized programs and support services, which raises graduation rates and promotes evidence-based decision-making. Well-informed education policies help policymakers achieve increased equity in education by addressing inequities. Conversely, students acquire knowledge that enables them to make well-informed decisions, enhance their well-being, and access essential support services. In the end, the study provides stakeholders with doable tactics to improve general well-being and academic performance, thereby making the educational environment more fruitful for all parties.

To get our dataset ready to be analyzed, we will first focus on cleaning and preparing the data. To improve the dataset's analytical readiness and consistency, cleaning methods included renaming columns, formatting variables appropriately, and fixing errors.

The study will then focus on the three different hypotheses that we have, which are Students who graduated from private high schools will have, on average, higher final GPAs compared to those from state or other types of high schools; students who engage in regular artistic or sports activities will have, on average, similar GPAs compared to those who do not participate in these activities; and students with parents who have higher levels of education (e.g., university, MSc, or PhD) will have, on average, higher expected GPAs upon graduation. It looks at important connections between particular attributes and student output grades. The research's main concerns centre on the possible effects of factors on academic performance, including the kind of high school attended, participation in extracurricular activities like athletics or the arts, and parental education.

The analyses of our exploratory tests and predictive models that follow take place as follows, to determine how particular distinct variables might affect the ordinal nature of the output grades, ordinal logistic regression is used. Multiple linear regression is used in parallel to assess the correlation between output grades and cumulative grade point averages.

With everything looked at, this research takes an in-depth approach to understanding and revealing the complex factors influencing the performance of students. Through the utilization of statistical techniques and data-driven techniques, our goal is to provide significant insights into the various factors that impact students' academic performance.

Research Questions

Research Question 1: Does the type of high school (private, state, or other) that students graduate from significantly affect their final GPA?

Hypothesis 1: Students who graduated from private high schools will have, on average, higher final GPAs compared to those from state or other types of high schools.

Research Question 2: Do students who engage in regular artistic or sports activities have higher or lower GPAs compared to students who do not participate in these activities?

Hypothesis 2: Students who engage in regular artistic or sports activities will have, on average, similar GPAs compared to those who do not participate in these activities.

Research Question 3: How does the level of parents' education (mothers', and fathers' separately) correlate with a student's expected cumulative GPA upon graduation?

Hypothesis 3: Students with parents who have higher levels of education (e.g., University, MSc, or PhD) will have, on average, higher expected GPAs upon graduation.

Literature Review

With the ever-changing needs of both society and students, education is a dynamic field that is always changing. There are many factors other than standard academic measurements that affect a student's educational performance and using machine learning and data analysis techniques can help extract meaningful information to help us better understand these factors that affect a student's academic performance. Therefore, in this literature review, we are going to be using machine learning to help us find these factors and conjure up a conclusion on their academic success.

A possible factor that could affect student performance is attending different types of high schools could accumulating different final GPAs. A study done by Hoxby, 1994 states that students in public schools tend to have better productivity when students are exposed to competition and have. There are also studies that state that private school graduates achieve better grades, (Frenette & Chan, 2015) stated that attending private schools increased the likelihood that children would have peers with parents who had college degrees and that they had more socioeconomic traits that are favourable to academic achievement. However, there are also studies that state that the school type a person attends has no meaningful effect on academic performance (Harry, 2008). In addition to that factor, extracurricular activities play an important role in affecting a student's academic performance. According to Reeves(2008), students who took part in three or four extracurricular activities during the academic year had dramatically better grades compared to those who didn't. Students who did not partake in any activity were most likely to get into drugs and just bad habits(Adachi-Mejia et al., 2014).

Moving on, another factor that could play an important role in a student's academic performance is their background specifically with their parents, if they had higher qualifications or not. Students whose parents had earned at least a bachelor's degree were expecting their children to finish college compared to students whose parents had just graduated from high school or had less than a high school diploma (Lippman, 2008).

In conclusion, there have been previous studies on this topic with varying results. This study of literature has brought attention to a few aspects that affect students' academic performance such as the kind of school a student graduates from, extracurricular activities and family background.

Data Source and Preprocessing

With the problem statement that we have mentioned, we have drawn a plan that combines data analysis with strict statistical methodology to address the complicated issue of identifying the variables that affect a student's academic success.

Data Source:

Data Description

The data was obtained from Kaggle and here is a hyperlink to our dataset: Student Performance. The original purpose of this dataset(Student Performance) was to explore the factors that might influence students' academic outcomes. It seeks to identify patterns and relationships between various students' attributes and their final grades, which are represented as Grade Point Averages(GPA) in ordinal categories. The dataset aims to provide insights into how different factors contribute to or impact student success.

Collection Details:

The precise collection details, including the data collection method and the organization responsible for collecting the data, may not be explicitly stated within the dataset. However, it is common for educational datasets like this to be generated through surveys, administrative records, or a combination of data sources. The data may represent students from various educational institutions over a specified period.

Number of Variables:

The "Student Performance" dataset consists of multiple variables, each capturing different aspects of students' characteristics and academic performance. Based on the information you provided earlier, there are 32 variables in this dataset. These variables encompass a range of factors, including demographic information, educational background, and behaviours, which are believed to be relevant to student performance.

Peculiarities of the Source Data:

Missing Values: One common peculiarity in datasets related to educational performance is the presence of missing values. Missing data may occur due to various reasons, such as students not providing certain information or administrative errors. It's essential to handle missing values appropriately during data analysis, either through imputation techniques or by excluding observations with missing data.

Ordinal Nature of GPA: The GPA variable in the dataset is recorded as ordinal categories (e.g., 1, 2, 3, etc.) rather than continuous values. This ordinal nature of the data requires appropriate statistical techniques like ordinal logistic regression to analyse the relationships effectively.

Categorical Variables: The dataset includes both nominal and ordinal categorical variables. Nominal variables, such as "Graduated high-school type" and "Regular artistic or sports activity," require the creation of dummy variables to represent different categories in the analysis.

Data Preprocessing:

Packages Required

```
library(readxl)
library(tidyverse)
library(ggplot2)
library(broom)
library(MASS)
library(tidymodels)
library(pROC)
library(olsrr)
```

readxl: used to easily import data from Excel files into R for analysis.

tidyverse: used to help clean up the data and make it more consistent with the values.

ggplot2: used to allow us to create visually appealing and customizable graphs compared to the default standard graphs set by R.

broom: used to convert complex model outputs into a tidy format for easier analysis.

MASS: has a comprehensive set of tools that are for modern applied statistics.

tidymodels: smoothens the entire modelling process from data preparation to evaluation.

pROC: was used to evaluate

olsrr: used to conduct the assumption test

Data Preprocessing

Firstly the dataset was loaded into R and we viewed the data so that we could better understand and familiarize ourselves with it.

```
data <- read_csv('studentdata.csv') # Student performance Data into a variable
data
head(data)
str(data)
View(data) # Have a look at the entire dataframe
class(data) # To check if it is a dataframe or tibble
glimpse(data) # look at the structure of the data
```

```
> glimpse(df) #look at the structure of data
Rows: 145
Columns: 33
$ `STUDENT ID`  <chr> "STUDENT1", "STUDENT2", "STUDENT3", "STUDENT4", "STUDENT5", "STUDE…
$ `1`           <dbl> 2, 2, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, 3, 2, 1, 2, 1, 1, 1, 1, …
$ `2`           <dbl> 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1, 2, 2, 2, 2, …
$ `3`           <dbl> 3, 3, 2, 1, 1, 2, 2, 2, 3, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, …
$ `4`           <dbl> 3, 3, 3, 3, 3, 3, 4, 3, 3, 3, 3, 4, 4, 5, 4, 3, 5, 3, 4, 3, 5, 5, …
$ `5`           <dbl> 1, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, …
$ `6`           <dbl> 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, …
$ `7`           <dbl> 2, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, …
$ `8`           <dbl> 1, 1, 2, 2, 3, 2, 1, 2, 1, 3, 3, 4, 1, 1, 3, 1, 1, 1, 3, 2, 1, 1, …
$ `9`           <dbl> 1, 1, 4, 1, 1, 1, 1, 2, 1, 4, 2, 2, 1, 1, 4, 1, 1, 1, 1, 2, 4, 4, …
$ `10`          <dbl> 1, 1, 2, 2, 4, 1, 3, 3, 3, 2, 3, 3, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, …
$ `11`          <dbl> 1, 2, 2, 1, 3, 3, 1, 4, 2, 1, 3, 5, 3, 2, 3, 4, 2, 2, 2, 3, 3, 2, …
$ `12`          <dbl> 2, 3, 2, 2, 3, 3, 3, 3, 4, 2, 4, 5, 5, 2, 1, 4, 2, 2, 2, 3, 3, 2, …
$ `13`          <dbl> 3, 2, 2, 5, 2, 2, 1, 1, 2, 3, 2, 1, 4, 2, 2, 2, 4, 2, 5, 3, 3, 4, …
$ `14`          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, …
$ `15`          <dbl> 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 3, 3, 2, 2, 4, 3, 2, 2, 2, 2, 2, 2, …
$ `16`          <dbl> 5, 1, 1, 1, 4, 3, 4, 3, 4, 3, 2, 2, 2, 3, 1, 1, 3, 1, 5, 4, 3, 3, …
$ `17`          <dbl> 3, 2, 2, 3, 2, 1, 2, 1, 1, 2, 1, 3, 3, 1, 2, 2, 2, 2, 5, 4, 4, 3, …
$ `18`          <dbl> 2, 2, 1, 1, 1, 1, 2, 2, 2, 2, 1, 3, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, …
$ `19`          <dbl> 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 3, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, …
$ `20`          <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, …
$ `21`          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, 1, 1, 1, 2, 1, 1, …
$ `22`          <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 2, 1, 1, 2, 1, 2, 2, 2, 2, 1, 1, 1, …
$ `23`          <dbl> 1, 1, 1, 1, 2, 1, 1, 3, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ `24`          <dbl> 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, …
$ `25`          <dbl> 3, 3, 2, 3, 2, 1, 3, 3, 3, 2, 2, 3, 2, 3, 2, 3, 3, 2, 3, 3, 3, 3, …
$ `26`          <dbl> 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 1, 2, 2, 3, 2, 2, 2, 1, 2, 1, 1, …
$ `27`          <dbl> 1, 3, 1, 2, 2, 1, 3, 2, 2, 2, 2, 3, 2, 3, 2, 2, 3, 2, 3, 2, 2, 3, …
$ `28`          <dbl> 2, 2, 1, 1, 1, 2, 3, 1, 2, 2, 2, 3, 3, 3, 1, 3, 3, 2, 3, 3, 3, 3, …
$ `29`          <dbl> 1, 2, 2, 3, 2, 4, 4, 1, 4, 1, 1, 4, 4, 4, 4, 2, 4, 2, 3, 2, 4, 3, …
$ `30`          <dbl> 1, 3, 2, 2, 2, 4, 4, 1, 3, 2, 1, 3, 2, 2, 4, 2, 3, 2, 3, 3, 4, 3, …
$ `COURSE ID`   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ GRADE         <dbl> 1, 1, 1, 1, 1, 2, 5, 2, 5, 0, 2, 0, 0, 1, 2, 2, 1, 2, 2, 3, 1, 1, …
```

Since there were no missing variables in the dataset we moved on to cleaning the data which included subsetting and factorising the data.

Cleaning the data

Firstly, to clean the data we changed the column names to their proper forms that were given on Kaggle since the titles were just numbers.

```
#change column name
colnames(data) <- c("STUDENT ID", "Student Age", "Sex",
                    "Graduated high-school type", "Scholarship type",
                    "Additional work", "Regular artistic or sports activity",
                    "Do you have a partner", "Total salary if available",
                    "Transportation to the university", "Accommodation type in Cyprus",
                    "Mothers education", "Fathers education", "Number of sisters/brothers",
                    "Parental status", "Mothers occupation", "Fathers occupation",
                    "Weekly study hours", "Reading frequency", "Reading frequency (scientific
books/journals)",
                    "Attendance to the seminars/conferences related to the department",
                    "Impact of your projects/activities on your success",
                    "Attendance to classes", "Preparation to midterm exams 1",
                    "Preparation to midterm exams 2", "Taking notes in classes",
                    "Listening in classes" ,"Discussion improves my interest and success in the
course",
                    "Flip-classroom", "Cumulative grade point average in the last semester
(/4.00)",
                    "Expected Cumulative grade point average in the graduation (/4.00)",
                    "Course ID", "OUTPUT Grade")
```

```
> glimpse(df)
Rows: 145
Columns: 33
$ `STUDENT ID`                                                  <chr> "STUDENT1", "STUDENT2", "STUDENT3", "STUDENT4", "STUDENT5", "STUDENT6", "STUDENT7", "STUDENT8", "STUDEN…
$ `Student Age`                                                 <dbl> 2, 2, 2, 1, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, 3, 2, 1, 2, 1, 1, 1, 1, 2, 3, 2, 2, 2, 1, 3, 2, 2, 3, 2, 2, 1…
$ `Sex`                                                         <dbl> 2, 2, 2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 2…
$ `Graduated high-school type`                                  <dbl> 3, 3, 2, 1, 1, 2, 2, 2, 3, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 1, 2, 3, 2, 2, 2, 2, 1…
$ `Scholarship type`                                            <dbl> 3, 3, 3, 3, 3, 3, 4, 3, 3, 3, 3, 4, 4, 5, 4, 3, 5, 3, 4, 3, 5, 5, 3, 2, 3, 3, 3, 3, 3, 4, 5, 3, 3, 3, 3…
$ `Additional work`                                             <dbl> 1, 1, 2, 1, 2, 2, 2, 1, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2…
$ `Regular artistic or sports activity`                         <dbl> 2, 2, 2, 2, 2, 2, 2, 1, 1, 2, 2, 1, 2, 2, 1, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 1, 2, 2, 2, 1, 2, 2, 2, 2…
$ `Do you have a partner`                                       <dbl> 2, 2, 2, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 1, 2, 1…
$ `Total salary if available`                                   <dbl> 1, 1, 2, 2, 3, 2, 1, 2, 1, 3, 3, 4, 1, 1, 3, 1, 1, 1, 3, 2, 1, 1, 1, 5, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2…
$ `Transportation to the university`                            <dbl> 1, 1, 4, 1, 1, 1, 1, 2, 1, 4, 2, 2, 1, 1, 4, 1, 1, 1, 1, 2, 4, 4, 1, 1, 1, 1, 1, 1, 4, 4, 1, 1, 1, 1, 1…
$ `Accommodation type in Cyprus`                                <dbl> 1, 1, 2, 2, 4, 1, 3, 3, 3, 2, 3, 3, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 1, 1, 1, 2, 1, 1, 2, 2, 2, 2, 1, 1, 1…
$ `Mothers education`                                           <dbl> 1, 2, 2, 1, 3, 3, 1, 4, 2, 1, 3, 5, 3, 2, 3, 4, 2, 2, 2, 3, 3, 2, 1, 1, 3, 1, 4, 1, 2, 3, 1, 1, 2, 1, 3…
$ `Fathers education`                                           <dbl> 2, 3, 2, 2, 3, 3, 3, 3, 4, 2, 4, 5, 5, 2, 1, 4, 2, 2, 2, 3, 3, 2, 1, 4, 1, 4, 4, 1, 2, 3, 1, 2, 2, 2, 3…
$ `Number of sisters/brothers`                                  <dbl> 3, 2, 2, 5, 2, 2, 1, 1, 2, 3, 2, 1, 4, 2, 2, 2, 4, 2, 5, 3, 3, 4, 4, 3, 3, 3, 4, 4, 4, 2, 5, 4, 2, 5, 3…
$ `Parental status`                                             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ `Mothers occupation`                                          <dbl> 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 3, 3, 2, 2, 4, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 4, 2, 2, 2, 2, 2…
$ `Fathers occupation`                                          <dbl> 5, 1, 1, 1, 4, 3, 4, 3, 4, 3, 2, 2, 2, 3, 1, 1, 3, 1, 5, 4, 3, 3, 4, 1, 4, 4, 4, 3, 1, 1, 4, 3, 1, 3, 4…
$ `Weekly study hours`                                          <dbl> 3, 2, 2, 3, 2, 1, 2, 1, 1, 2, 1, 3, 3, 1, 2, 2, 2, 2, 5, 4, 4, 3, 3, 3, 1, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2…
$ `Reading frequency`                                           <dbl> 2, 2, 1, 1, 1, 1, 2, 2, 2, 2, 1, 3, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1…
$ `Reading frequency (scientific books/journals)`               <dbl> 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 3, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 3, 2, 2, 1…
$ `Attendance to the seminars/conferences related to the department` <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ `Impact of your projects/activities on your success`          <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ `Attendance to classes`                                       <dbl> 1, 1, 1, 1, 1, 1, 2, 1, 1, 2, 2, 1, 1, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2…
$ `Preparation to midterm exams 1`                              <dbl> 1, 1, 1, 1, 2, 1, 1, 3, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1…
$ `Preparation to midterm exams 2`                              <dbl> 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ `Taking notes in classes`                                     <dbl> 3, 3, 2, 3, 2, 1, 3, 3, 3, 2, 2, 3, 2, 3, 2, 3, 3, 2, 3, 3, 3, 3, 3, 3, 3, 2, 2, 3, 3, 3, 3, 2, 3, 2, 1…
$ `Listening in classes`                                        <dbl> 2, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 1, 2, 2, 3, 2, 2, 2, 1, 2, 1, 1, 1, 2, 1, 1, 3, 1, 2, 2, 3, 3, 2, 3, 2…
$ `Discussion improves my interest and success in the course`   <dbl> 1, 3, 1, 2, 2, 1, 3, 2, 2, 2, 2, 3, 2, 3, 2, 2, 3, 2, 3, 3, 2, 2, 3, 2, 3, 3, 3, 3, 3, 2, 3, 2, 3, 2, 2…
$ `Flip-classroom`                                              <dbl> 2, 2, 1, 1, 1, 2, 3, 1, 2, 2, 2, 3, 3, 3, 1, 3, 3, 2, 3, 3, 3, 3, 3, 3, 2, 2, 2, 1, 3, 2, 3, 3, 1, 2, 1…
$ `Cumulative grade point average in the last semester (/4.00)` <dbl> 1, 2, 2, 3, 2, 4, 4, 1, 4, 1, 1, 4, 4, 4, 4, 2, 4, 2, 3, 2, 4, 3, 3, 3, 4, 1, 2, 2, 5, 4, 5, 4, 2, 2, 4…
$ `Expected Cumulative grade point average in the graduation (/4.00)` <dbl> 1, 3, 2, 2, 2, 4, 4, 1, 3, 2, 1, 3, 2, 2, 4, 2, 3, 2, 3, 3, 4, 3, 3, 3, 4, 2, 1, 1, 4, 3, 4, 3, 3, 3, 1…
$ `Course ID`                                                   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ `OUTPUT Grade`                                                <dbl> 1, 1, 1, 1, 1, 2, 5, 2, 5, 0, 2, 0, 0, 1, 2, 2, 1, 2, 2, 3, 1, 1, 3, 1, 2, 3, 1, 1, 3, 5, 5, 3, 1, 2, 2…
```

Categorising and Factoring

After we were done changing the columns to make it more understandable with what we were dealing with it was time to factorise the data which will make it easier to plot our graphs.

```
#Convert the variables
df$`Student Age`<-factor(df$`Student Age`,c("1", "2", "3"), labels =
c("18-21", "22-25", "above 26"))
df$Sex <- factor(df$Sex,c("1", "2"), labels = c( "Female","Male"))
df$`Graduated high-school type`<-factor(df$`Graduated high-school type`,c("1",
"2", "3"),labels=c("private", "state", "other"))
df$`Additional work`<- factor(df$`Additional work`,c("1", "2"), labels = c(
"Yes","No"))
df$`Regular artistic or sports activity`<-factor(df$`Regular artistic or
sports activity`,c("1", "2"), labels = c( "Yes","No"))
df$`Do you have a partner`<-factor(df$`Do you have a partner`,c("1", "2"),
labels = c( "Yes","No"))
df$`Transportation to the university`<-factor(df$`Transportation to the
university`,c("1", "2", "3", "4"), labels = c("Bus", "Private car/taxi"
,"bicycle","Other"))
df$`Accommodation type in Cyprus`<-factor(df$`Accommodation type in
Cyprus`,c("1" ,"2", "3", "4"), labels = c("rental", "dormitory"," with
family", "Other"))
df$`Mothers education`<-factor(df$`Mothers education`,c("1", "2", "3", "4"
,"5" ,"6"), labels = c("primary school", "secondary school", "high school",
"university", "MSc.", "Ph.D."))
df$`Fathers education`<-factor(df$`Fathers education`,c("1",
"2","3","4","5","6"),labels = c("primary school","secondary school","high
```

```
school","university", "MSc.", "Ph.D."))
df$`Parental status`<-factor(df$`Parental status`,c("1","2","3"),labels =
c("married","divorced","died - one of them or both"))
df$`Mothers occupation`<-factor(df$`Mothers
occupation`,c("1","2","3","4","5","6"),labels =
c("retired","housewife","government officer", "private sector
employee","self-employment","other"))
df$`Fathers occupation`<-factor(df$`Fathers
occupation`,c("1","2","3","4","5"),labels = c("retired","government officer",
"private sector employee","self-employment","other"))
df$`Reading frequency`<-factor(df$`Reading frequency`,c("1","2","3"),labels =
c("None","Sometimes","Often"))
df$`Reading frequency (scientific books/journals)`<-factor(df$`Reading
frequency (scientific books/journals)`,c("1","2","3"),labels =
c("None","Sometimes","Often"))
df$`Attendance to the seminars/conferences related to the department`<-
factor(df$`Attendance to the seminars/conferences related to the
department`,c("1", "2"), labels = c( "Yes","No"))
df$`Impact of your projects/activities on your success`<- factor(df$`Impact of
your projects/activities on your success`,c("1", "2","3"), labels = c(
"positive","negative","neutral"))
df$`Attendance to classes`<- factor(df$`Attendance to classes`,c("1",
"2","3"), labels = c( "always","sometimes","never"))
df$`Preparation to midterm exams 1`<-factor(df$`Preparation to midterm exams
1`,c("1","2","3"),labels = c("alone","with friends","not applicable"))
df$`Preparation to midterm exams 2`<-factor(df$`Preparation to midterm exams
2`,c("1","2","3"),labels = c("closest date to the exam","regularly during the
semester","never"))
df$`Taking notes in classes`<-factor(df$`Taking notes in
classes`,c("1","2","3"),labels=c("never","sometimes","always"))
df$`Listening in classes`<-factor(df$`Listening in
classes`,c("1","2","3"),labels=c("never","sometimes","always"))
df$`Discussion improves my interest and success in the
course`<-factor(df$`Discussion improves my interest and success in the
course`,c("1","2","3"),labels=c("never","sometimes","always"))
df$`Flip-classroom`<-factor(df$`Flip-classroom`,c("1","2","3"),labels = c("not
useful","useful","not applicable"))
df$`OUTPUT Grade`<-factor(df$`OUTPUT
Grade`,c("0","1","2","3","4","5","6","7"),labels =
c("Fail","DD","DC","CC","CB","BB", "BA", "AA"))
```

Once all the variables were changed it was time to save them into another CSV file for better analysis.

```
write.csv(data, file = "performance.csv", row.names = FALSE)
```

And now we will have a look at the final summary of the entire dataset:

```
> summary(df)
  STUDENT ID        Student Age         Sex      Graduated high-school type Scholarship type Additional work Regular artistic or sports activity Do you have a partner
 Length:145         Min.   :1.000   Min.   :1.0   Min.   :1.000              Min.   :1.000    Min.   :1.000   Min.   :1.0                         Min.   :1.000
 Class :character   1st Qu.:1.000   1st Qu.:1.0   1st Qu.:2.000              1st Qu.:3.000    1st Qu.:1.000   1st Qu.:1.0                         1st Qu.:1.000
 Mode  :character   Median :2.000   Median :2.0   Median :2.000              Median :3.000    Median :2.000   Median :2.0                         Median :2.000
                    Mean   :1.621   Mean   :1.6   Mean   :1.945              Mean   :3.572    Mean   :1.662   Mean   :1.6                         Mean   :1.579
                    3rd Qu.:2.000   3rd Qu.:2.0   3rd Qu.:2.000              3rd Qu.:4.000    3rd Qu.:2.000   3rd Qu.:2.0                         3rd Qu.:2.000
                    Max.   :3.000   Max.   :2.0   Max.   :3.000              Max.   :5.000    Max.   :2.000   Max.   :2.0                         Max.   :2.000
 Total salary if available Transportation to the university Accommodation type in Cyprus Mothers education Fathers education Number of sisters/brothers Parental status
 Min.   :1.000             Min.   :1.000                   Min.   :1.000                 Min.   :1.000     Min.   :1.000     Min.   :1.000             Min.   :1.000
 1st Qu.:1.000             1st Qu.:1.000                   1st Qu.:1.000                 1st Qu.:1.000     1st Qu.:2.000     1st Qu.:2.000             1st Qu.:1.000
 Median :1.000             Median :1.000                   Median :2.000                 Median :2.000     Median :3.000     Median :3.000             Median :1.000
 Mean   :1.628             Mean   :1.621                   Mean   :1.731                 Mean   :2.283     Mean   :2.634     Mean   :2.807             Mean   :1.172
 3rd Qu.:2.000             3rd Qu.:2.000                   3rd Qu.:2.000                 3rd Qu.:3.000     3rd Qu.:3.000     3rd Qu.:4.000             3rd Qu.:1.000
 Max.   :5.000             Max.   :4.000                   Max.   :4.000                 Max.   :6.000     Max.   :6.000     Max.   :5.000             Max.   :3.000
 Mothers occupation Fathers occupation Weekly study hours Reading frequency Reading frequency (scientific books/journals)
 Min.   :1.000      Min.   :1.000      Min.   :1.0        Min.   :1.000     Min.   :1.000
 1st Qu.:2.000      1st Qu.:2.000      1st Qu.:2.0        1st Qu.:2.000     1st Qu.:2.000
 Median :2.000      Median :3.000      Median :2.0        Median :2.000     Median :2.000
 Mean   :2.359      Mean   :2.807      Mean   :2.2        Mean   :1.945     Mean   :2.014
 3rd Qu.:2.000      3rd Qu.:4.000      3rd Qu.:3.0        3rd Qu.:2.000     3rd Qu.:2.000
 Max.   :5.000      Max.   :5.000      Max.   :5.0        Max.   :3.000     Max.   :3.000
 Attendance to the seminars/conferences related to the department Impact of your projects/activities on your success Attendance to classes Preparation to midterm exams 1
 Min.   :1.000                                                    Min.   :1.000                                    Min.   :1.000        Min.   :1.000
 1st Qu.:1.000                                                    1st Qu.:1.000                                    1st Qu.:1.000        1st Qu.:1.000
 Median :1.000                                                    Median :1.000                                    Median :1.000        Median :1.000
 Mean   :1.214                                                    Mean   :1.207                                    Mean   :1.241        Mean   :1.338
 3rd Qu.:1.000                                                    3rd Qu.:1.000                                    3rd Qu.:1.000        3rd Qu.:2.000
 Max.   :2.000                                                    Max.   :3.000                                    Max.   :2.000        Max.   :3.000
 Preparation to midterm exams 2 Taking notes in classes Listening in classes Discussion improves my interest and success in the course Flip-classroom
 Min.   :1.000                  Min.   :1.000           Min.   :1.000       Min.   :1.000                                          Min.   :1.000
 1st Qu.:1.000                  1st Qu.:2.000           1st Qu.:2.000       1st Qu.:2.000                                          1st Qu.:1.000
 Median :1.000                  Median :3.000           Median :2.000       Median :2.000                                          Median :2.000
 Mean   :1.166                  Mean   :2.545           Mean   :2.055       Mean   :2.393                                          Mean   :1.807
 3rd Qu.:1.000                  3rd Qu.:3.000           3rd Qu.:3.000       3rd Qu.:3.000                                          3rd Qu.:2.000
 Max.   :3.000                  Max.   :3.000           Max.   :3.000       Max.   :3.000                                          Max.   :3.000
 Cumulative grade point average in the last semester (/4.00) Expected Cumulative grade point average in the graduation (/4.00)   Course ID      OUTPUT Grade
 Min.   :1.000                                               Min.   :1.000                                                    Min.   :1.000   Min.   :0.000
 1st Qu.:2.000                                               1st Qu.:2.000                                                    1st Qu.:1.000   1st Qu.:1.000
 Median :3.000                                               Median :3.000                                                    Median :3.000   Median :3.000
 Mean   :3.124                                               Mean   :2.724                                                    Mean   :4.131   Mean   :3.228
 3rd Qu.:4.000                                               3rd Qu.:3.000                                                    3rd Qu.:7.000   3rd Qu.:5.000
 Max.   :5.000                                               Max.   :4.000                                                    Max.   :9.000   Max.   :7.000
```

Exploratory Data Analysis

Now that the data has been cleaned up and we have a fair understanding of what the data types are and so on, it is time to analyze it and see if there are any trends within the dataset before we move on to our Substantive Data Analysis (SDA).

General Analysis of Students' Grades

Firstly we will do a general analysis of the students' grades. We used the ggplot2 package to create this bar graph with the grades achieved and the number of students that achieved the grade.

```
# Summary the number of students in each grade
Grade_Counts <- df %>%
  group_by(`OUTPUT Grade`) %>%
  summarise(Count = n())
ggplot(Grade_Counts, aes(x =`OUTPUT Grade` , y = Count)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Student Grades", x = "grades", y = "couts") +
  theme_minimal()#plot the summary grades diagram of students
```

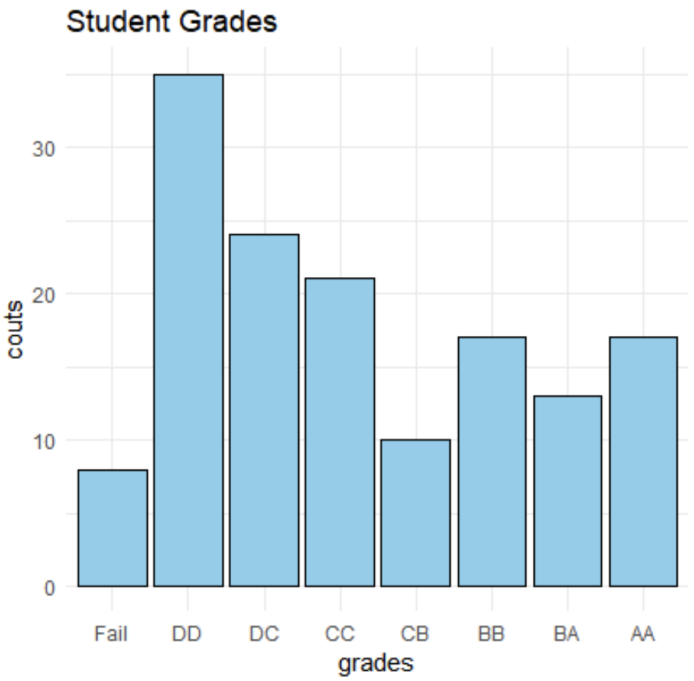


Figure 1: Summary of students' grades

The bar graph above shows us that the most often obtained grade was 'DD,' which corresponds to a 1.0 on a 4.0 scale, translating to a range of 0-49%. The least frequent grade, on the other hand, was 'F,' with around 8 pupils obtaining a score of 0 on both a 4.0 and a 100% scale. To conclude, it can be seen that the failing grades (DD, DC and DC) have the most students falling in those ranges whereas the top few grades (CB, BB, BA and AA) had the lowest.

Research Question & Hypothesis One

Does the type of high school (private, state, or other) that students graduate from significantly affect their final GPA?

The following figures that will be shown below are to answer our first hypothesis that is, students who graduated from private high schools will have, on average, higher final GPAs compared to those from state or other types of high schools.

```
# Graph for the Final GPA vs Graduated High School Type (Figure 2)
ggplot(data, aes(x = `OUTPUT Grade`, fill = `Graduated high-school type`)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Final GPA VS Graduated High School Type",
       x = "GPA",
       y = "Count",
       fill ="High School Type" ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2")
```

```
# The following code is to plot the graph that shows the output of AA & BA Grade
Ratios by High School Type (Figure 3)
AABA_count <- data %>%
  filter(`OUTPUT Grade` == "AA" | `OUTPUT Grade` == "BA") %>%
  group_by(`Graduated high-school type`)%>%
  summarise(Count = n())
#Calculate the number of students that graduate from different type of high
school
School_counts <- data %>%
  group_by(`Graduated high-school type`) %>%
  summarise(Total_Count = n())

#calculate the ratio of students who get A's who graduated from different school
AABA_Ratio <- left_join(AABA_count, School_counts, by = NULL) %>%
  mutate(Ratio = Count / Total_Count)

ggplot(AABA_Ratio, aes(x = `Graduated high-school type`, y = Ratio, color =
`Graduated high-school type`)) +
  geom_point(size = 3) +
  geom_text_repel(aes(label = round(Ratio, 4)),
                  box.padding = 0.1)+
  labs(title = "AA & BA Grade Ratio by High School Type",
       x = "High School Type",
       y = "AA & BA Grade Ratio",
       color = "High School Type") +
  theme_minimal()
```

```
# To plot the graph for the ratio grade distribution by high school type (Figure
4)
# calculate the number of students of different grades by different school
```

```
grade_counts <- data %>%
  group_by(`Graduated high-school type`, `OUTPUT Grade`) %>%
  summarise(Count = n())

# calculate all ratios of different grades by different school
grade_ratios <- grade_counts %>%
  group_by(`Graduated high-school type`) %>%
  mutate(Total = sum(Count),
         Ratio = Count / Total)


ggplot(grade_ratios, aes(x = `OUTPUT Grade`, y = Ratio, color = `Graduated
high-school type`, group = `Graduated high-school type`)) +
  geom_point(size = 3) +
  geom_line(aes(group = `Graduated high-school type`), size = 1) +
  geom_text_repel(aes(label = round(Ratio, 4)),
                  box.padding = 0.4, ) +
  labs(title = "Grade Distribution by High School Type",
       x = "Grade",
       y = "Ratio",
       color = "High School Type") +
  theme_minimal()
```

Graph Output:

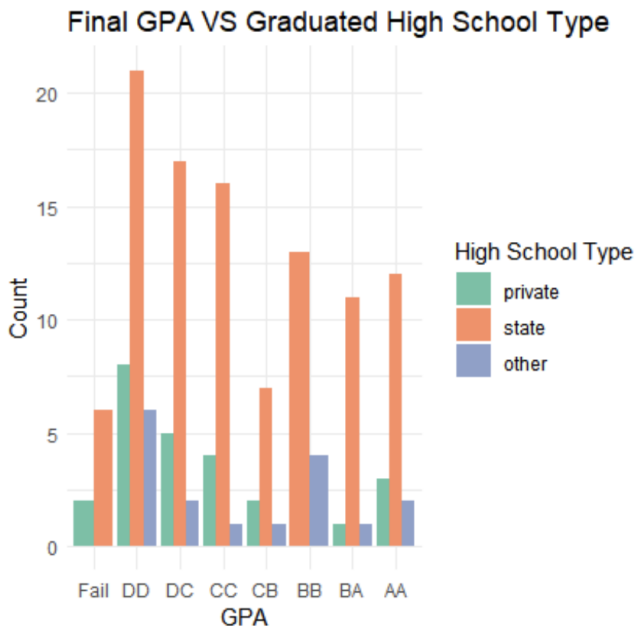


Figure 2: Summary distribution of students' GPAs that graduated from different high schools.

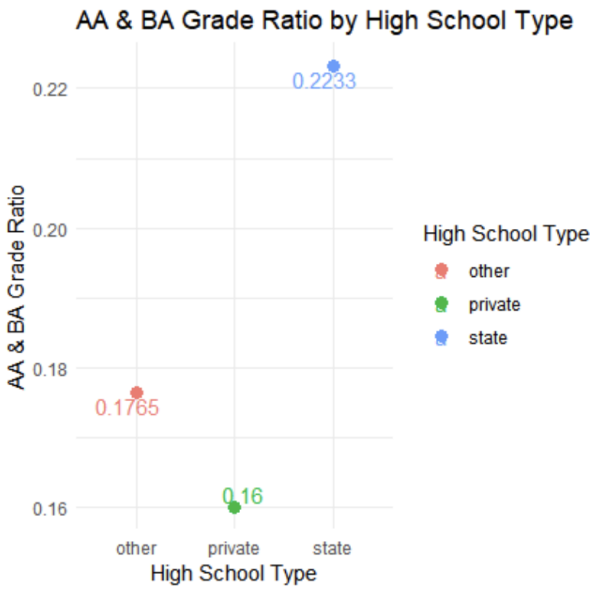


Figure 3: AA & BA Grade Ratio by High School Type

Figure 4: Ratio grade distribution by high school type

From our analysis, we can conclude that our hypothesis is disproven. Students who have attended state or other high schools had a higher final GPA compared to private high schools. Our group's assumptions as to why the results are as so; the diverse student populations that state high schools serve create a rich learning environment with a range of viewpoints and experiences that can enhance their academic performance. These schools often put in place specific structures of support and motivation that push students to aim for greater academic achievement, which may raise their GPAs. Furthermore, because a high grade in one school may not be the same in another, differences in grading schemes between state and private high schools may result in discrepancies in how students are regarded to have performed academically.

Research Question & Hypothesis Two

Do students who engage in regular artistic or sports activities have higher or lower GPAs compared to students who do not participate in these activities?

The following figures will be shown below to answer our second hypothesis, that is, students who engage in regular artistic or sports activities will have, on average, similar GPAs compared to those who do not participate in these activities.

```
# Summary plot of student GPA that engage in regular artistic or sports
activities or not
ggplot(data, aes(x = `OUTPUT Grade`, fill = `Regular artistic or sports
activity`)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Final GPA VS regular artistic or sports activity",
       x = "GPA",
       y = "Count",
       fill ="regular artistic or sports activity" ) +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")
```

```
sports_counts <- data %>%
  group_by(`Regular artistic or sports activity`, `OUTPUT Grade`) %>%
  summarise(Count = n())

# Calculate ration of student GPAs that engage in regular artistic or sports
activities or not
sports_ratios <- sports_counts %>%
  group_by(`Regular artistic or sports activity`) %>%
  mutate(Total = sum(Count),
         sportsRatio = Count / Total)

ggplot(sports_ratios, aes(x = `OUTPUT Grade`, y = sportsRatio, color = `Regular
artistic or sports activity`, group = `Regular artistic or sports activity`)) +
  geom_point(size = 3) +
  geom_line(aes(group = `Regular artistic or sports activity`), size = 1) +
  geom_text_repel(aes(label = round(sportsRatio, 4)),
                  box.padding = 0.4, ) +
  labs(title = "Grade Distribution by engage in regular artistic or sports
activities",
       x = "Grade",
       y = "Ratio",
       color = "regular artistic or sports activities") +
  theme_minimal()
```

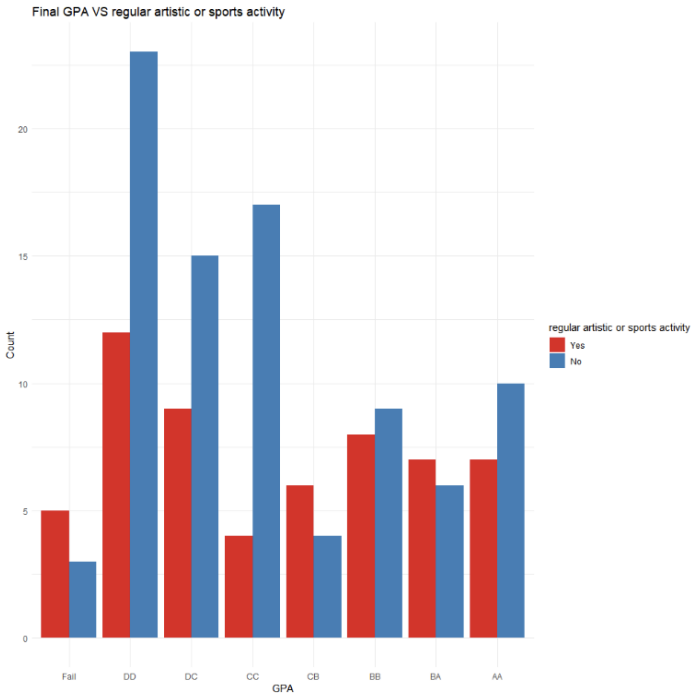


Figure 4: Bar graph of students who participated in regular artistic or sports activities compared to those who don't.

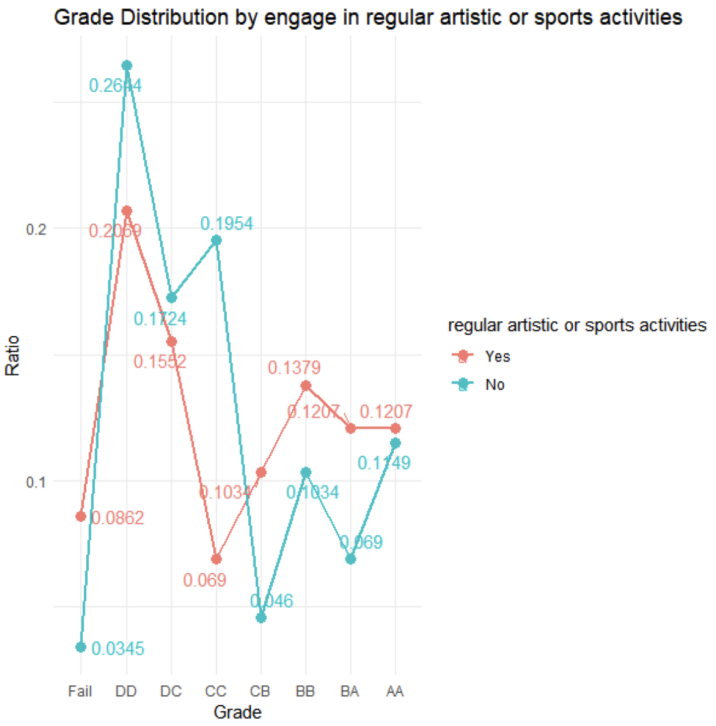


Figure 5: Line graph of the grade distribution among students who engage in regular artistic or sports activities.

From our analysis, we can conclude that our hypothesis was proven to be true. Students who engage in regular artistic or sports activities have better rates of getting good grades compared to students who do not engage in regular artistic or sports activities. Our group's assumptions as to why the results are as so; participating in athletics or the arts develops self-discipline and dedication, traits that easily convert to academic endeavors and build a stronger work ethic which improves grades. Participating in these activities also relieves stress and promotes better mental health, both of which are linked to higher marks in school. Sports and the arts require a great deal of focus and concentration, which frequently leads to increased attention spans that carry over into schoolwork and improve school performance by promoting improved knowledge retention. Moreover, engaging in sports regularly improves brain health and cognitive function, which may boost school performance through improved brain function.

Research Question & Hypothesis Three

How does the level of parents' education (mothers', and fathers' separately) correlate with a student's expected cumulative GPA upon graduation?

The following figures that will be shown below are to answer our third hypothesis, that is, Students with parents who have higher levels of education (e.g., University, MSc, or PhD) will have, on average, higher expected GPAs upon graduation.

```
#summary plot of each parent
ggplot(data, aes(x = `OUTPUT Grade`, fill = `Mothers education`)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Final GPA affect by mother education level",
       x = "GPA",
       y = "Count",
       fill ="mother education" ) +
  theme_minimal()

ggplot(data, aes(x = `OUTPUT Grade`, fill = `Fathers education`)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Final GPA affect by father education level",
       x = "GPA",
       y = "Count",
       fill ="Father education" ) +
  theme_minimal()

data$`Mothers education`
data$`Fathers education`
data$`OUTPUT Grade`

# Combine mother's and father's education levels into a new variable
data$ParentsEducation <- paste(data$`Mothers education`, data$`Fathers education`,
sep="&")

# Convert the Parents Education variable to a factor
data$ParentsEducation <- as.factor(data$ParentsEducation)

ggplot(data, aes(x = `OUTPUT Grade`  , fill = ParentsEducation)) +
  geom_bar(position = "dodge", stat = "count") +
  labs(title = "Final GPA affect by parents education level",
       x = "GPA",
       y = "Count",
       fill ="GPA" ) +
  theme_minimal()

# Calculate counts for AA and BA grades based on parents' education level
PEgrades_count <- data %>%
  filter(`OUTPUT Grade` %in% c("AA", "BA")) %>%
  group_by(ParentsEducation, `OUTPUT Grade`) %>%
  summarise(Count = n())

# Plotting the bar chart
ggplot(PEgrades_count, aes(x = `OUTPUT Grade`, y = Count, fill = ParentsEducation)) +
  geom_bar(position = "dodge", stat = "identity") +
  labs(title = "AA & BA Grade Count by Parents' Education Level",
       x = "GPA",
       y = "Count",
```

```
      fill = "Parents Education level") +
theme_minimal()
```

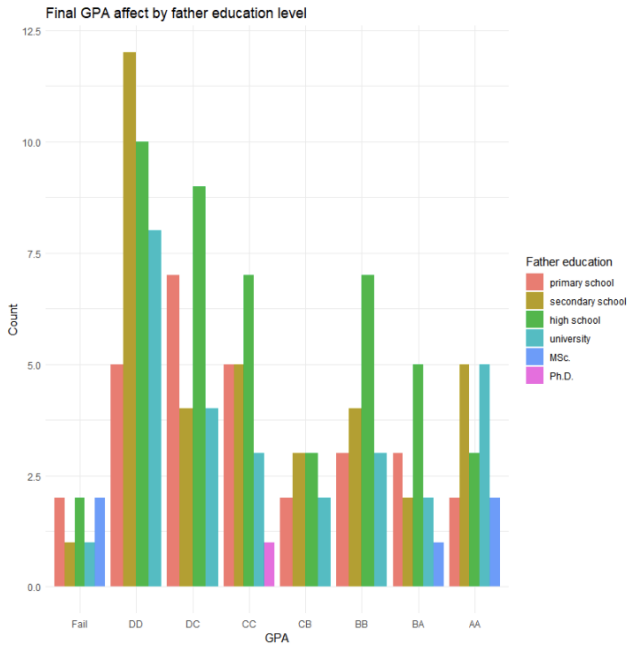


Figure 6: Bar graph of father's education level with final GPA

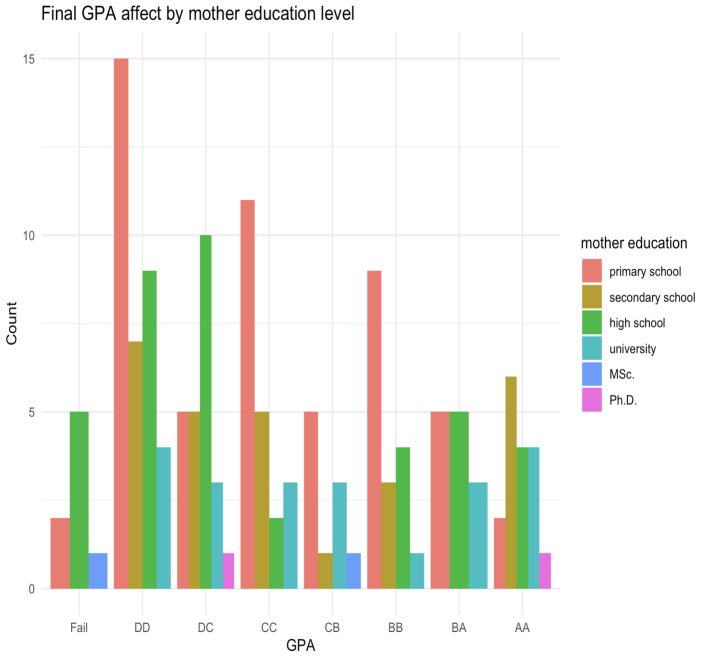


Figure 7: Bar graph of mothers' educational level with final GPA.

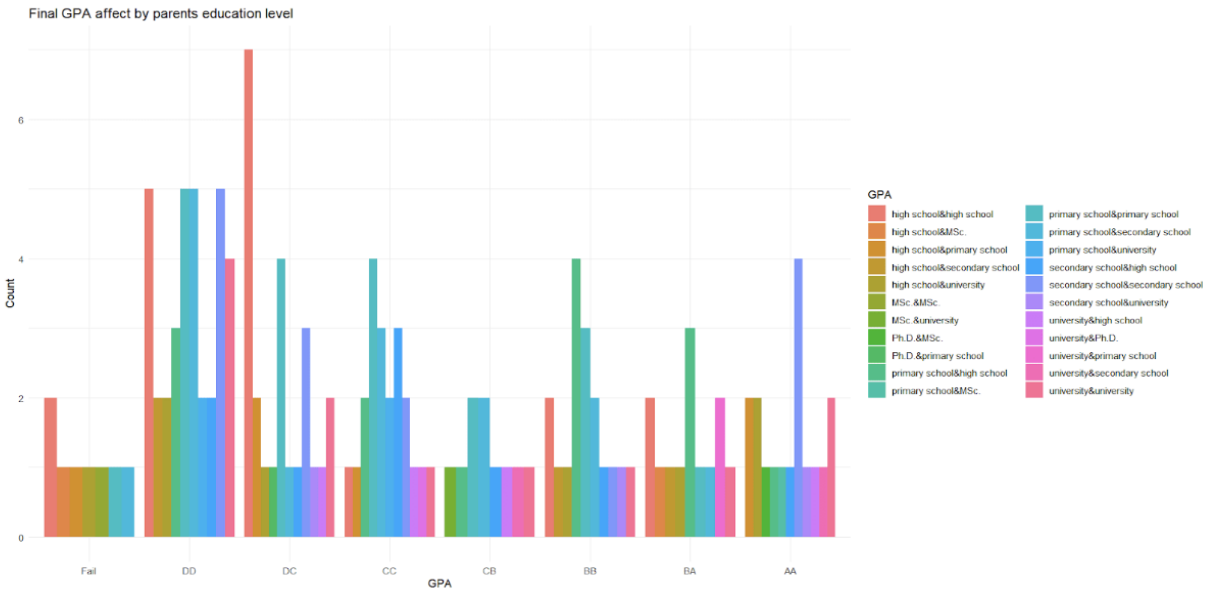


Figure 8: Both parent's education levels are combined and shown in the bar graph

From our analysis, we can conclude that our hypothesis was proven to be true. Students who have parents who have higher education are more likely to have students with higher GPAs. Our group's assumptions as to why the results are as so; higher-educated parents are effective role models for their children, and they may encourage them to strive for comparable

accomplishments by seeing their parents involved in challenging academic or professional endeavors. These parents often become more interested in their kids' schooling, helping out with homework, and providing advice, which has a good impact on the kids' academic achievement. Furthermore, these kids may put in more effort in their academics because of the greater expectations placed on them by their families and the desire for them to succeed in school and their careers.

Substantive Data Analysis

Ordinal Logistic Regression

Firstly to predict the GPA of the students we are using Ordinal Logistic regression. Before we begin with this machine learning model firstly we need to have an overview summary of the dataset.

```
summary(data)
```

After understanding the dataset more we moved on to choosing relevant variables that will be used to help with our prediction.

Using the chi-sq test to find the significant relation for the variables

```
# Chi-Squared Test
summary(data)
chisq.test(data$`OUTPUT Grade`, data$`Student Age`)
chisq.test(data$`OUTPUT Grade`, data$Sex)
chisq.test(data$`OUTPUT Grade`, data$`Cumulative grade point average in the last
semester (/4.00)`)
chisq.test(data$`OUTPUT Grade`, data$`Expected Cumulative grade point average in
the graduation (/4.00)`)
chisq.test(data$`OUTPUT Grade`, data$`Scholarship type`)
chisq.test(data$`OUTPUT Grade`, data$`Impact of your projects/activities on your
success`)

chisq.test(data$`OUTPUT Grade`, data$`Weekly study hours`)#worth further
investigation

# example for no significant relation
chisq.test(data$`OUTPUT Grade`, data$`Transportation to the university`)
chisq.test(data$`OUTPUT Grade`, data$`Additional work`)
```

We performed Pearson's Chi-squared test with our selected variables and came up with the following outputs:
NOTE: In all the outputs of the Chi-squared test the approximation may be incorrect.

Student Age:
`data:  data$`OUTPUT Grade` and data$`Student Age``

```
X-squared = 28.954, df = 14, p-value = 0.0106
```

Sex:

```
data:  data$`OUTPUT Grade` and data$Sex
X-squared = 26.233, df = 7, p-value = 0.0004575
```

Cumulative GPA in the last semester (/4.00):

```
data:  data$`OUTPUT Grade` and data$`Cumulative grade point average in the last
semester (/4.00)
X-squared = 66.836, df = 28, p-value = 5.088e-05
```

Expected Cumulative GPA in the graduation (/4.00):

```
data:  data$`OUTPUT Grade` and data$`Expected Cumulative grade point average in
the graduation (/4.00)`
X-squared = 33.339, df = 21, p-value = 0.042
```

Scholarship Type:

```
data:  data$`OUTPUT Grade` and data$`Scholarship type`
X-squared = 46.401, df = 28, p-value = 0.01583
```

Impact of your projects/activities on your success:

```
data:  data$`OUTPUT Grade` and data$`Impact of your projects/activities on your
success`
X-squared = 25.599, df = 14, p-value = 0.02909
```

Weekly Study Hours:

```
data:  data$`OUTPUT Grade` and data$`Weekly study hours`
X-squared = 39.83, df = 28, p-value = 0.06847
```

Transportation to the University:

```
data:  data$`OUTPUT Grade` and data$`Transportation to the university`
X-squared = 27.131, df = 21, p-value = 0.1666
```

Additional Work:

```
data:  data$`OUTPUT Grade` and data$`Additional work`
X-squared = 9.6447, df = 7, p-value = 0.2096
```

From our results, we have concluded that the statistically significant variable is the expected cumulative GPA in graduation followed by the cumulative GPA in the last semester which had a highly significant relationship. Secondly, we have concluded that student age, sex, scholarship type, and the impact of a person's projects/activities on one's success are significant relationships. The insignificant variables are transportation to the university and additional work a person is doing followed by weekly study hours a student does which could do some further investigation or analysis.

Once we had received which variables have a significant relationship it was time to do a final analysis by combining those significant variables of our data sets to check which ones affected students' performance more:

```
######ordinal logistic regression model to predict GPA #2 ######
# Select relevant columns for prediction
features1 <- data[c("Student Age","Sex",
                    "Scholarship type",
                    "Impact of your projects/activities on your success",
                    "Expected Cumulative grade point average in the graduation
(/4.00)",
                    "Cumulative grade point average in the last semester
(/4.00)",
                    "Weekly study hours",
                    "Reading frequency",
                    "Additional work",
                    "Transportation to the university",
                    "Accommodation type in Cyprus",
                    "Taking notes in classes")]

output_grade_ordinal <- data$`OUTPUT Grade` # Output Grade is ordinal (e.g.,
"AA", "BA", "CB", "Fail")

# Train the ordinal logistic regression model
ordinal_model <- polr(output_grade_ordinal ~ ., data = features1, method =
"probit")

ordinal_model
summary(ordinal_model)
```

```
> ordinal_model
Call:
polr(formula = output_grade_ordinal ~ ., data = features1, method = "probit")

Coefficients:
                                                  `Student Age`22-25                                                   `Student Age`above 26
                                                         -0.05334357                                                             -0.90013251
                                                             SexMale                                                  `Scholarship type`25%
                                                          0.70539466                                                             -0.09859978
                                                `Scholarship type`50%                                                  `Scholarship type`75%
                                                         -0.60938856                                                             -0.07598645
                                               `Scholarship type`Full        `Impact of your projects/activities on your success`negative
                                                         -0.81101028                                                             -0.13474709
       `Impact of your projects/activities on your success`neutral `Expected Cumulative grade point average in the graduation (/4.00)`
                                                         -0.70721155                                                              0.02116424
     `Cumulative grade point average in the last semester (/4.00)`                                        `Weekly study hours`
                                                          0.27310394                                                             -0.15630180
                                       `Reading frequency`Sometimes                                       `Reading frequency`Often
                                                          0.41122610                                                              0.95991100
                                            `Additional work`No                 `Transportation to the university`Private car/taxi
                                                          0.35018146                                                             -0.03993696
                          `Transportation to the university`bicycle                  `Transportation to the university`Other
                                                          0.83766639                                                             -1.06223988
                           `Accommodation type in Cyprus`dormitory                `Accommodation type in Cyprus`with family
                                                          0.59976825                                                              0.17219489
                               `Accommodation type in Cyprus`Other                      `Taking notes in classes`sometimes
                                                         -0.96512116                                                              0.16002723
                                   `Taking notes in classes`always
                                                          0.18625497

Intercepts:
   Fail|DD      DD|DC      DC|CC      CC|CB      CB|BB      BB|BA      BA|AA
-0.7958333  0.5232323  1.0860109  1.5828357  1.8368328  2.3254870  2.8164593

Residual Deviance: 504.1405
AIC: 564.1405

> summary(ordinal_model)

Re-fitting to get Hessian

Call:
polr(formula = output_grade_ordinal ~ ., data = features1, method = "probit")

Coefficients:
                                                                      Value Std. Error  t value
`Student Age`22-25                                                -0.05334    0.22493 -0.23716
`Student Age`above 26                                             -0.90013    0.44853 -2.00687
SexMale                                                            0.70539    0.21680  3.25368
`Scholarship type`25%                                             -0.09860    1.26358 -0.07803
`Scholarship type`50%                                             -0.60939    1.09788 -0.55506
`Scholarship type`75%                                             -0.07599    1.11938 -0.06788
`Scholarship type`Full                                            -0.81101    1.12999 -0.71771
`Impact of your projects/activities on your success`negative      -0.13475    0.55297 -0.24368
`Impact of your projects/activities on your success`neutral       -0.70721    0.33801 -2.09229
`Expected Cumulative grade point average in the graduation (/4.00)` 0.02116    0.14220  0.14883
`Cumulative grade point average in the last semester (/4.00)`      0.27310    0.09902  2.75819
`Weekly study hours`                                              -0.15630    0.11049 -1.41460
`Reading frequency`Sometimes                                       0.41123    0.25024  1.64336
`Reading frequency`Often                                           0.95991    0.35591  2.69704
`Additional work`No                                                0.35018    0.21277  1.64580
`Transportation to the university`Private car/taxi                -0.03994    0.29511 -0.13533
`Transportation to the university`bicycle                          0.83767    1.14892  0.72909
`Transportation to the university`Other                           -1.06224    0.30747 -3.45476
`Accommodation type in Cyprus`dormitory                            0.59977    0.24792  2.41916
`Accommodation type in Cyprus`with family                          0.17219    0.33775  0.50984
`Accommodation type in Cyprus`Other                               -0.96512    1.12762 -0.85589
`Taking notes in classes`sometimes                                 0.16003    0.52397  0.30541
`Taking notes in classes`always                                    0.18625    0.52401  0.35544

Intercepts:
        Value   Std. Error t value
Fail|DD -0.7958  1.2662    -0.6285
DD|DC    0.5232  1.2551     0.4169
DC|CC    1.0860  1.2555     0.8650
CC|CB    1.5828  1.2560     1.2603
CB|BB    1.8368  1.2576     1.4606
BB|BA    2.3255  1.2656     1.8375
BA|AA    2.8165  1.2727     2.2130

Residual Deviance: 504.1405
AIC: 564.1405
```

Given our outputs above, the probit link function was utilised to simulate the connections between several predictor variables and an ordinal response variable. The computed coefficients clarify these connections by illuminating how variations in the variables affect the probability of falling into certain ordinal answer groups. For example, there is a larger log-odds of obtaining a higher category for males as opposed to females. The intercepts represent the log odds of switching between neighbouring categories. The residual deviation,

which was 504.1405, and the AIC, which had a value of 564.1405, were used to evaluate the model's goodness of fit. Better model fit is indicated by lower values of these measures. Our larger study goals are furthered by this analysis, which provides us with useful information for comprehending and forecasting factors impacting our ordinal response variable.

```
# Data frame of estimated coefficients
tidy(ordinal_model)

# Performance metrics on training data
glance(ordinal_model)

performance_test<- data

predict_data<-predict(ordinal_model, new_data = performance_test)

predict_data
```

```
> # Data frame of estimated coefficients
> tidy(ordinal_model)

Re-fitting to get Hessian

# A tibble: 30 × 5
   term                                                          estimate std.error statistic coef.type
   <chr>                                                            <dbl>     <dbl>     <dbl> <chr>
 1 `Student Age`22-25                                             -0.0533     0.225   -0.237  coefficient
 2 `Student Age`above 26                                          -0.900      0.449   -2.01   coefficient
 3 SexMale                                                         0.705      0.217    3.25   coefficient
 4 `Scholarship type`25%                                          -0.0986     1.26    -0.0780 coefficient
 5 `Scholarship type`50%                                          -0.609      1.10    -0.555  coefficient
 6 `Scholarship type`75%                                          -0.0760     1.12    -0.0679 coefficient
 7 `Scholarship type`Full                                         -0.811      1.13    -0.718  coefficient
 8 `Impact of your projects/activities on your success`negative   -0.135      0.553   -0.244  coefficient
 9 `Impact of your projects/activities on your success`neutral    -0.707      0.338   -2.09   coefficient
10 `Expected Cumulative grade point average in the graduation (/4.00)`  0.0212     0.142    0.149  coefficient
# ℹ 20 more rows
# ℹ Use `print(n = ...)` to see more rows
> # Performance metrics on training data
> glance(ordinal_model)
# A tibble: 1 × 7
    edf logLik   AIC   BIC deviance df.residual  nobs
  <int>  <dbl> <dbl> <dbl>    <dbl>       <dbl> <dbl>
1    30  -252.  564.  653.     504.         115   145
> performance_test<- data
> predict_data<-predict(ordinal_model, new_data = performance_test)
> predict_data
  [1] DD   DD   DD   DD   CC   AA   DD   CC   DD   DD   DD   CC   DD   DD   DD   DD   DD   CC   CC   DD   DD   CC   DD   AA   CC   BB   DD   DD
 [30] BB   AA   DD   DD   DD   CC   CC   BB   BB   CC   DD   CC   DD   DD   CC   BB   CC   BB   AA   CC   DD   CC   DD   DD   DD   DD   CC   BB   AA
 [59] DD   CC   DD   DD   CC   BB   DD   DD   AA   CC   CC   DD   AA   DD   AA   AA   AA   AA   AA   AA   DD   BB   DD   CC   AA   DD   CC   CC   AA
 [88] DD   AA   CC   DD   AA   DD   CC   AA   CC   AA   AA   AA   DD   AA   AA   AA   AA   DD   AA   AA   CC   DD   AA   DD   CC   DD   Fail DD   DD
[117] DD   DD   DD   CC   DD   Fail DD   DD   DD   DD   DD   DD   DD   DD   DD   BB   DD   DD   DD   DD   DD   DD   DD   CC   CC   CC   CC   DD
Levels: Fail DD DC CC CB BB BA AA
```

In this section of our analysis, using the probit link function, we provide the results of an ordinal regression model in this portion of our study. There are two main components to the results. First off, a thorough understanding of the predicted coefficients for each of the predictor variables is provided in the 'Coefficients and Model Summary' section. We may identify the impacts of variables like age, gender, and kind of research on transitions between ordinal categories by using these coefficients, which give crucial insights into the direction and degree of their influence on the ordinal response variable. A collection of summary statistics for evaluating the overall performance of the model is also provided in the 'Model Performance Metrics' section. The model's balance between fit and complexity is measured by metrics such as AIC, BIC, and deviation; lower AIC values suggest a more effective trade-off. Moreover, we provide the predicted ordinal categories for every observation in the

model's predictions for a test dataset. This data gives us a thorough grasp of the model's efficacy and a solid basis on which to base our analysis's data-driven judgements.

Multiple Linear Regression

The next machine learning model we used to predict our dataset was Multiple Linear Regression. Firstly before moving on to any of the analyses, we had to create a new dataset because we wanted to keep the values in number for building the linear regression.

```
#######Multiple linear regression #########
#create new data-set because we want keep the values in numbers for building the
linear regression
data2<-read_csv('studentdata.csv')
colnames(data2) <- c("STUDENT ID", "Student Age", "Sex",
                     "Graduated.high.school.type","Scholarship type",
                     "Additional work","Regular.artistic.or.sports.activity",
                     "Do you have a partner","Total salary if available",
                     "Transportation to the university","Accommodation type in
Cyprus",
                     "Mothers.education","Fathers.education","Number of
sisters/brothers",
                     "Parental status","Mothers occupation","Fathers occupation",
                     "Weekly study hours","Reading frequency","Reading frequency
(scientific books/journals)",
                     "Attendance to the seminars/conferences related to the
department",
                     "Impact of your projects/activities on your success",
                     "Attendance to classes","Preparation to midterm exams 1",
                     "Preparation to midterm exams 2","Taking notes in classes",
                     "Listening in classes","Discussion improves my interest and
success in the course",

"Flip-classroom","Cumulative.grade.point.average.in.the.last.semester",
                     "Expected Cumulative grade point average in the graduation
(/4.00)",
                     "Course ID","OUTPUT.Grade")

# Perform linear regression
linear_model <- lm(`OUTPUT.Grade` ~ `Graduated.high.school.type` +
                     `Regular.artistic.or.sports.activity` +
                     `Mothers.education` +
                     `Fathers.education` +
                     `Cumulative.grade.point.average.in.the.last.semester`,
                   data = data2)

# Summary of the linear regression model
summary(linear_model)
```

```
> # Summary of the linear regression model
> summary(linear_model)

Call:
lm(formula = OUTPUT.Grade ~ Graduated.high.school.type + Regular.artistic.or.sports.activity +
    Mothers.education + Fathers.education + Cumulative.grade.point.average.in.the.last.semester,
    data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0414 -1.5092 -0.5494  1.3544  4.7206

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       0.2956     1.1401   0.259    0.796
Graduated.high.school.type                        0.5062     0.3299   1.535    0.127
Regular.artistic.or.sports.activity              -0.2513     0.3581  -0.702    0.484
Mothers.education                                 0.1286     0.1594   0.806    0.421
Fathers.education                                 0.1325     0.1691   0.783    0.435
Cumulative.grade.point.average.in.the.last.semester   0.5464     0.1343   4.067 7.93e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.089 on 139 degrees of freedom
Multiple R-squared:  0.1277,    Adjusted R-squared:  0.09633
F-statistic:  4.07 on 5 and 139 DF,  p-value: 0.001769
```

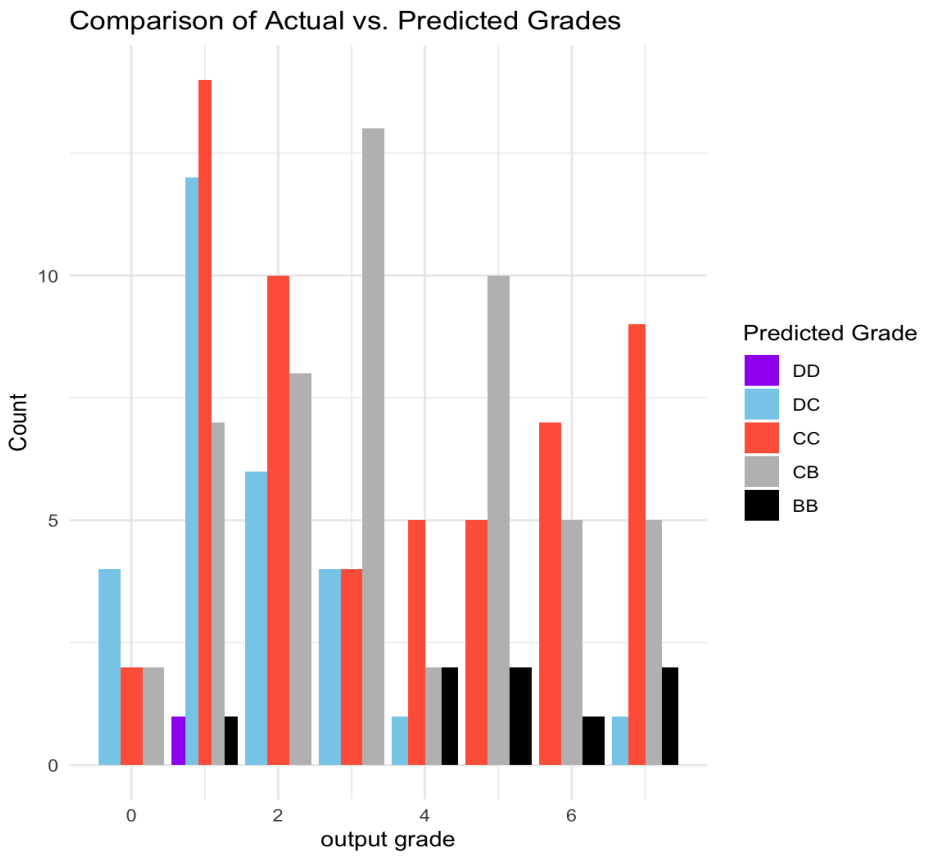
From the results we received the math equation for the prediction is:
(0.2956+0.5062*Graduated.high.school.type -0.251*Regular.artistic.or.sports.activity +
0.12*Mothers.education + 0.1*Fathers.education +
0.5*Cumulative.grade.point.average.in.the.last.semester)

For each unit increase in the variable "Graduated high school type", and the rest remain constant, the predicted output grade increases by 0.5062 points.
#However, this coefficient is not statistically significant at the conventional levels (p-value = 0.127), meaning it might not have a significant impact on the output grade. For each unit increase in the variable "Regular artistic or sports activity", the predicted output grade decreases by 0.2513 points. However, this coefficient is also not statistically significant (p-value = 0.484). For each unit increase in the variables "Mothers education" and "Fathers education", the predicted output grade increases by 0.1286 and 0.1325 points, respectively. However, neither of these coefficients is statistically significant (p-values = 0.421 and 0.435). For each unit increase in the variable "Cumulative grade point average in the last semester", the predicted output grade increases by 0.5464 points. This coefficient is statistically significant (p-value < 0.001), indicating a strong positive impact of the cumulative GPA on the output grade.

The multiple R-squared value (0.1277) indicates that approximately 12.77% of the variance in the output grade is explained by the predictors in the model. The p-value associated with the F-statistic (p-value < 0.001) suggests that the overall model is statistically significant, meaning at least one of the predictors has a significant effect on the output grade.

```
# Create a new data-frame for prediction
prediction_data <- data.frame('Graduated.high.school.type' =
data2$`Graduated.high.school.type`,
                              'Regular.artistic.or.sports.activity' =
data2$`Regular.artistic.or.sports.activity`,
                              'Mothers.education' = data2$`Mothers.education`,
                              'Fathers.education' = data2$`Fathers.education`,

'Cumulative.grade.point.average.in.the.last.semester' =
data2$`Cumulative.grade.point.average.in.the.last.semester`,
                              'OUTPUT.Grade'= data2$OUTPUT.Grade)
```

```
# Add predicted values to the prediction data and round to the nearest whole
number
prediction_data$predicted_grade <- round(predict(linear_model, newdata =
prediction_data), 0)

#Plot
ggplot(prediction_data, aes(x = OUTPUT.Grade, fill = factor(predicted_grade))) +
  geom_bar(position = "dodge") +
  labs(title = "Comparison of Actual vs. Predicted Grades",
       x = "output grade",
       y = "Count",
       fill = "Predicted Grade") +
  scale_fill_manual(values = c("0" = "red", "1" = "purple", "2" = "skyblue", "3"
= "tomato", "4" = "grey", "5" = "black", "6" = "pink", "7" = "brown"),
                    breaks = c("0", "1", "2", "3", "4", "5", "6", "7"),
                    labels = c("Fail", "DD", "DC", "CC", "CB", "BB", "BA",
"AA"),
                    name = "Predicted Grade") +
  theme_minimal()
```



Upon getting the graph above (Bar-graph of "Comparison of Actual vs Predicted Grades") we wanted to do a comparison between the actual grades from the dataset and the grades that we predicted.

```
# Create a data frame for comparison
compare_data <- data.frame(
  actual_Grade = factor(prediction_data$OUTPUT.Grade),
  predicted_Grade = factor(prediction_data$predicted_grade)
)

# Create a contingency table
compare_table <- table(compare_data$actual_Grade, compare_data$predicted_Grade)

# Print the contingency table
print(compare_table)
```

```
> print(compare_table)

   1  2  3  4  5
0  0  4  2  2  0
1  1 12 14  7  1
2  0  6 10  8  0
3  0  4  4 13  0
4  0  1  5  2  2
5  0  0  5 10  2
6  0  0  7  5  1
7  0  1  9  5  2
```

In the comparison table above the numbers from 1-5 horizontally are the predicted grades of our data set and 0-7 vertically are the actual grades of the dataset. Those students who get 7 means that they have an AA grade. We have predicted that only 1 student would get a DC grade, 9 students would get a CC grade, 5 students would get a CB grade and 2 got a BB grade.

```
########correlation test#######
cor.test(data2$OUTPUT.Grade,data2$Cumulative.grade.point.average.in.the.last.semester)
#moderate positive corr
cor.test(data2$OUTPUT.Grade,data2$`Additional work`)#weak positive
cor.test(data2$OUTPUT.Grade,data2$`Student Age`)#weak negative
cor.test(data2$OUTPUT.Grade,data2$`Do you have a partner`)#weak negative
cor.test(data2$OUTPUT.Grade,data2$`Parental status`)#weak positive
cor.test(data2$OUTPUT.Grade,data2$`Accommodation type in Cyprus`)#weak positive
cor.test(data2$OUTPUT.Grade,data2$`Transportation to the university`)#weak negative
cor.test(data2$OUTPUT.Grade,data2$Mothers.education)#weak positive
cor.test(data2$OUTPUT.Grade,data2$`Fathers occupation`)#weak negative
cor.test(data2$OUTPUT.Grade,data2$`Mothers occupation`)#weak negative
```

For the correlation tests that were done the only higher significant variable we found was the last semester GPA.

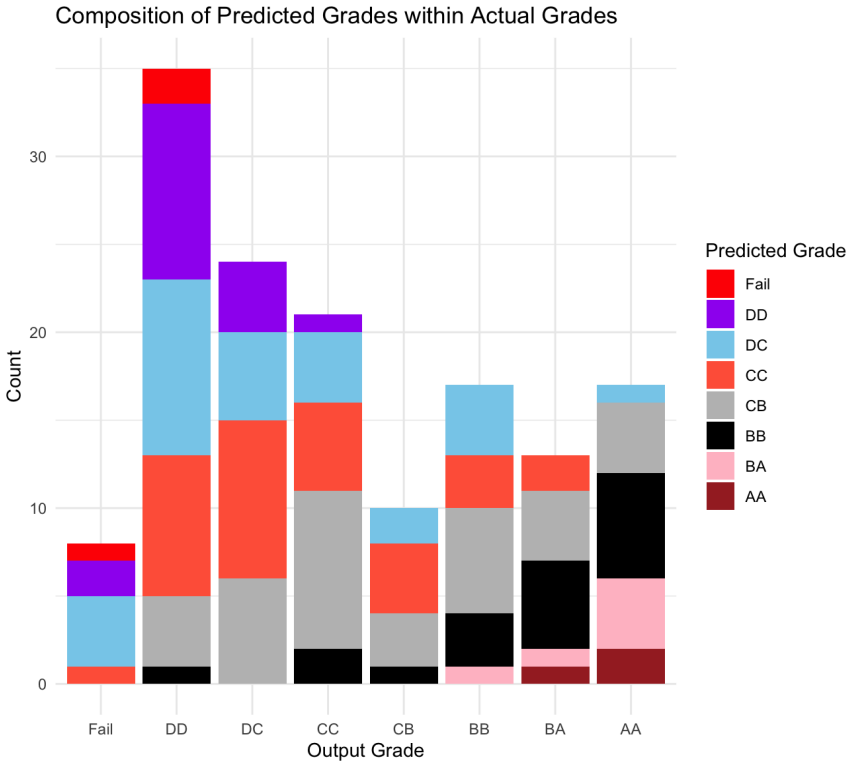
In the next stage of our analysis using multiple linear regression we are now applying the variables that we found to have a positive relationship.

```
########Multiple linear regression model 2######
data2$Sex <- factor(data2$Sex,c("1", "2"), labels = c( "Female","Male"))
linear_model2 <- lm(`OUTPUT.Grade` ~ `Student Age` +
                      `Sex`+
                      `Additional work` +
                      `Weekly study hours`+
                      `Transportation to the university`+
                      `Accommodation type in Cyprus`+
                      `Preparation to midterm exams 1` +
                      `Preparation to midterm exams 2`+
                      `Reading frequency` +
                      `Student Age`*`Reading frequency`+
                      `Additional work`*`Listening in classes`*`Taking notes in
classes`+
                      `Preparation to midterm exams 1`*`Preparation to midterm
exams 2`+
                      `Cumulative.grade.point.average.in.the.last.semester`,
                    data = data2)
# Summary of the linear regression model
summary(linear_model2)
```

```
> # Summary of the linear regression model
> summary(linear_model2)

Call:
lm(formula = OUTPUT.Grade ~ `Student Age` + Sex + `Additional work` +
    `Weekly study hours` + `Transportation to the university` +
    `Accommodation type in Cyprus` + `Preparation to midterm exams 1` +
    `Preparation to midterm exams 2` + `Reading frequency` +
    `Student Age` * `Reading frequency` + `Additional work` *
    `Listening in classes` * `Taking notes in classes` + `Preparation to midterm exams 1` *
    `Preparation to midterm exams 2` + Cumulative.grade.point.average.in.the.last.semester,
    data = data2)

Residuals:
   Min     1Q Median     3Q    Max
-3.567 -1.263 -0.241  1.241  4.715

Coefficients:
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                        15.5228     7.8019   1.990 0.048799 *
`Student Age`                                                       0.6520     1.1362   0.574 0.567069
SexMale                                                             1.5994     0.3406   4.696 6.83e-06 ***
`Additional work`                                                 -13.7933     4.5701  -3.018 0.003079 **
`Weekly study hours`                                               -0.2784     0.1935  -1.439 0.152768
`Transportation to the university`                                 -0.4065     0.1525  -2.665 0.008708 **
`Accommodation type in Cyprus`                                      0.3535     0.2236   1.581 0.116359
`Preparation to midterm exams 1`                                    0.6923     0.6868   1.008 0.315374
`Preparation to midterm exams 2`                                    1.7491     1.0168   1.720 0.087862 .
`Reading frequency`                                                 2.3532     0.8886   2.648 0.009127 **
`Listening in classes`                                             -7.0525     3.0962  -2.278 0.024424 *
`Taking notes in classes`                                          -9.6538     2.8124  -3.433 0.000809 ***
Cumulative.grade.point.average.in.the.last.semester                 0.3744     0.1276   2.935 0.003971 **
`Student Age`:`Reading frequency`                                  -0.7649     0.5649  -1.354 0.178096
`Additional work`:`Listening in classes`                            5.4963     2.0076   2.738 0.007081 **
`Additional work`:`Taking notes in classes`                         6.3022     1.7498   3.602 0.000453 ***
`Listening in classes`:`Taking notes in classes`                    3.5450     1.2028   2.947 0.003821 **
`Preparation to midterm exams 1`:`Preparation to midterm exams 2`  -0.5272     0.4844  -1.088 0.278545
`Additional work`:`Listening in classes`:`Taking notes in classes` -2.4261     0.7709  -3.147 0.002058 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.759 on 126 degrees of freedom
Multiple R-squared:  0.4397,    Adjusted R-squared:  0.3596
F-statistic: 5.493 on 18 and 126 DF,  p-value: 2.495e-09
```

For the linear_model2 there was no strong evidence to prove that the variables had positive relationships and that the model is too weak to rely on.

```
# Create a new data-frame for prediction
prediction_data2 <-
data.frame('Cumulative.grade.point.average.in.the.last.semester' =
data2$`Cumulative.grade.point.average.in.the.last.semester`,
                                'OUTPUT.Grade'= data2$OUTPUT.Grade)
view(prediction_data2)

# Add predicted values to the prediction data and round to the nearest whole
number
prediction_data2$predicted_grade <- round(predict(linear_model2, newdata =
data2), 0)

#Plot
ggplot(prediction_data2, aes(x = factor(OUTPUT.Grade, levels = c("0", "1", "2",
"3", "4", "5", "6", "7")), fill = factor(predicted_grade))) +
  geom_bar() +
  labs(title = "Composition of Predicted Grades within Actual Grades",
       x = "Output Grade",
       y = "Count",
       fill = "Predicted Grade") +
  scale_fill_manual(values = c("0" = "red", "1" = "purple", "2" = "skyblue", "3"
= "tomato",
                               "4" = "grey", "5" = "black", "6" = "pink", "7" =
"brown"),
                    breaks = c("0", "1", "2", "3", "4", "5", "6", "7"),
                    labels = c("Fail", "DD", "DC", "CC", "CB", "BB", "BA",
"AA"),
                    name = "Predicted Grade") +
  scale_x_discrete(labels = c("0" = "Fail", "1" = "DD", "2" = "DC", "3" = "CC",
"4" = "CB", "5" = "BB", "6" = "BA", "7" = "AA")) +
  theme_minimal()
```

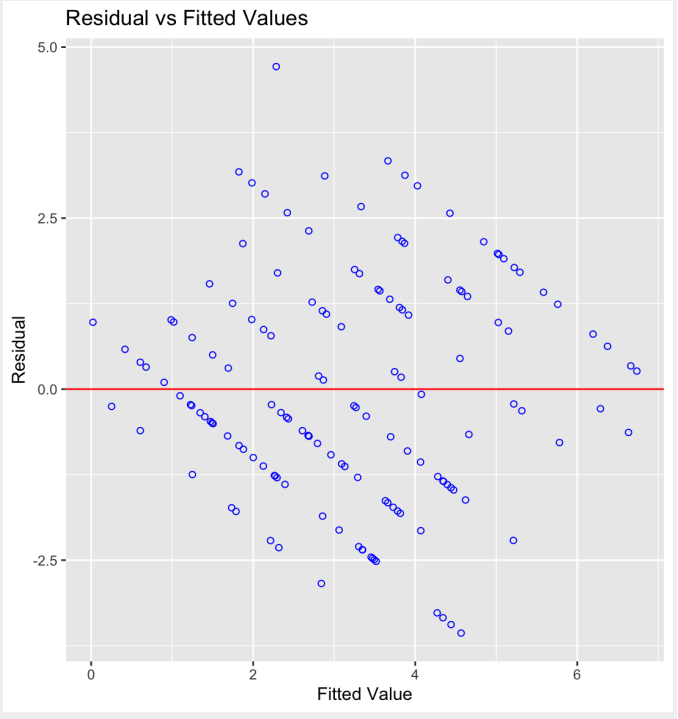


```
#change levels
prediction_data2$`OUTPUT.Grade`<-factor(prediction_data2$`OUTPUT.Grade`,c("0","1
","2","3","4","5","6","7"),labels =
c("Fail","DD","DC","CC","CB","BB","BA","AA"))
prediction_data2$predicted_grade<-factor(prediction_data2$`predicted_grade`,c("0
","1","2","3","4","5","6","7"),labels =
c("Fail","DD","DC","CC","CB","BB","BA","AA"))
#Filter the data for grades from Fail to AA
filtered_data <- subset(prediction_data2, predicted_grade %in%
c("Fail","DD","DC","CC","CB","BB","BA","AA"))

# Create a text-based comparison
comparison_table <- table(filtered_data$OUTPUT.Grade,
filtered_data$predicted_grade)
print(comparison_table)
```

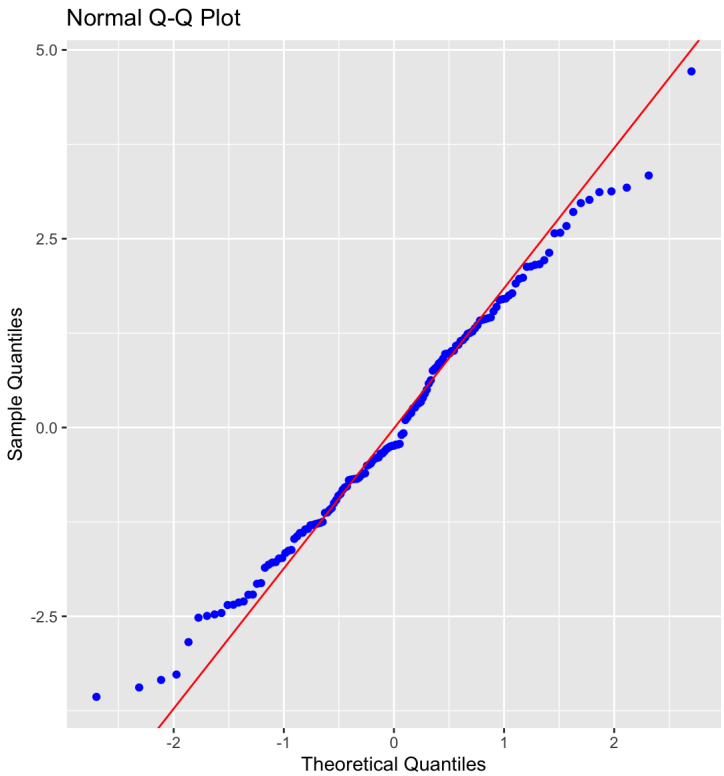
```
> print(comparison_table)

     Fail DD DC CC CB BB BA AA
Fail    1  2  4  1  0  0  0  0
DD      2 10 10  8  4  1  0  0
DC      0  4  5  9  6  0  0  0
CC      0  1  4  5  9  2  0  0
CB      0  0  2  4  3  1  0  0
BB      0  0  4  3  6  3  1  0
BA      0  0  0  2  4  5  1  1
AA      0  0  1  0  4  6  4  2
```

Here are the four assumptions our analysis must follow:

```
########Assumptions######
# Linearity Assumption
# Scatter plots for each predictor variable against the residuals
library(olsrr)#load packages
ols_plot_resid_fit(linear_model2)
```

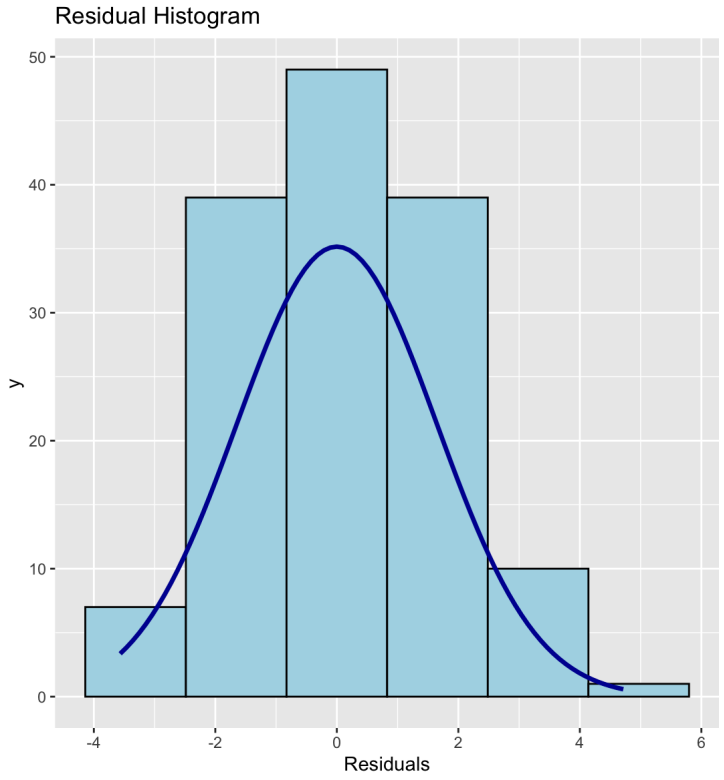



```
#NORMALITY
ols_plot_resid_qq(linear_model2)
```



As can be seen by the figure above the data is not normally distributed and the points do not fall along the 45-degree reference line which indicates that the quantiles of the data are not the same as the quantiles of the normal distribution. The plot shows that the data is skewed to the left, with more extreme values in the left tail than in the right tail. This suggests that the median of the data is less than the mean

```
ols_plot_resid_hist(linear_model2)
```



The majority of residuals falling close to zero, suggests that the model is making good predictions on average. However, there are a few outliers in the histogram, with some residuals falling far from zero. This suggests that the model is not perfect and that there are some cases where it is making large errors

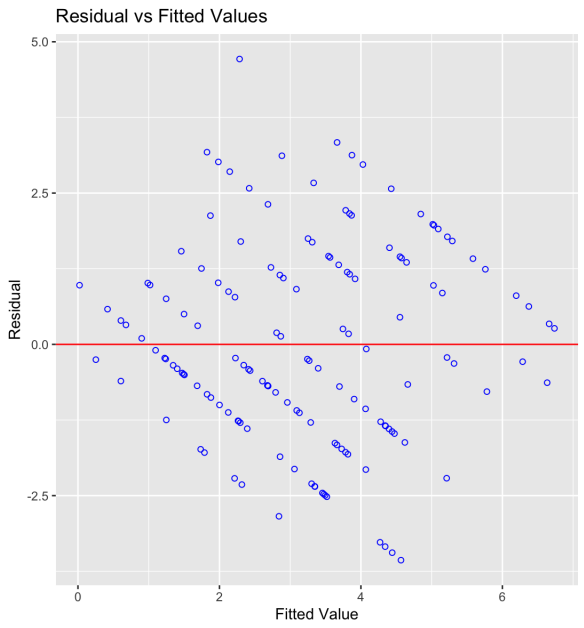
```
ols_test_normality(linear_model2)#test fail to reject the null hypothesis
```

```
> ols_test_normality(linear_model2)#test fail to reject the null hypothesis
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9898         0.3764
Kolmogorov-Smirnov        0.0767         0.3616
Cramer-von Mises         10.4663         0.0000
Anderson-Darling          0.4581         0.2604
-----------------------------------------------
```

According to our ols normality test Shapiro-Wilk shows no strong evidence to conclude that the residuals are not normally distributed. Kolmogrov-Smirnov shows that there is no significant departure from normality and Cramer-von Mises shows that there is a significant

departure from normality. Lastly, Anderson-Darling shows that the data does not significantly deviate from a normal distribution.

```
#HETEROSCEDASTICITY
ols_plot_resid_fit(linear_model2)

ols_test_breusch_pagan(linear_model2)
```




```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

                 Data
-----------------------------------------
Response : OUTPUT.Grade
Variables: fitted values of OUTPUT.Grade

       Test Summary
----------------------------
DF            =    1
Chi2          =    1.210193
Prob > Chi2   =    0.2712939
```

The variance of the residuals is constant (homoskedasticity). The variance of the residuals is not constant (heteroskedasticity). The p-value (0.2712939) associated with the Chi-square statistic is greater than the typical significance level (such as $\alpha = 0.05$). Therefore, do not have enough evidence to reject the null hypothesis. Based on the Breusch-Pagan test, there is no significant evidence to suggest that the variance of the residuals is not constant across

different levels of the independent variables. Hence, you do not have sufficient grounds to conclude that heteroskedasticity is present in your regression model.

```
ols_test_breusch_pagan(linear_model2, rhs = TRUE)
```

```
> ols_test_breusch_pagan(linear_model2, rhs = TRUE)

 Breusch Pagan Test for Heteroskedasticity
 -----------------------------------------
 Ho: the variance is constant
 Ha: the variance is not constant


Data
 ---------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------------------------------------
-----------
 Response : OUTPUT.Grade
 Variables: `Student Age` SexMale `Additional work` `Weekly study hours` `Transportation to the university` `Accommodation
type in Cyprus` `Preparation to midterm exams 1` `Preparation to midterm exams 2` `Reading frequency` `Listening in classes
` `Taking notes in classes` Cumulative.grade.point.average.in.the.last.semester `Student Age`:`Reading frequency` `Addition
al work`:`Listening in classes` `Additional work`:`Taking notes in classes` `Listening in classes`:`Taking notes in classes
` `Preparation to midterm exams 1`:`Preparation to midterm exams 2` `Additional work`:`Listening in classes`:`Taking notes
in classes`

        Test Summary
 ----------------------------
 DF            =    18
 Chi2          =    23.8172
 Prob > Chi2   =    0.1611082
```

There isn't enough evidence to conclude that there is a significant association between the categorical variables being analyzed in the chi-squared test.

```r
#Heteroskedasticity
ols_test_f(linear_model2)
#do not have sufficient evidence to conclude that the variance of errors is not
homogenous across the levels
#of the independent variables
ols_test_f(linear_model2, rhs = TRUE)
```

```
> #Heteroskedasticity
> ols_test_f(linear_model2)

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: fitted values of OUTPUT.Grade

      Test Summary
 -------------------------
 Num DF     =    1
 Den DF     =    143
 F          =    1.566822
 Prob > F   =    0.2127128
> #do not have sufficient evidence to conclude that the variance of errors is not homogenous across the levels
> #of the independent variables
> ols_test_f(linear_model2, rhs = TRUE)

 F Test for Heteroskedasticity
 -----------------------------
 Ho: Variance is homogenous
 Ha: Variance is not homogenous

 Variables: `Student Age` SexMale `Additional work` `Weekly study hours` `Transportation to the university` `Accommodation
type in Cyprus` `Preparation to midterm exams 1` `Preparation to midterm exams 2` `Reading frequency` `Listening in classes
` `Taking notes in classes` Cumulative.grade.point.average.in.the.last.semester `Student Age`:`Reading frequency` `Addition
al work`:`Listening in classes` `Additional work`:`Taking notes in classes` `Listening in classes`:`Taking notes in classes
` `Preparation to midterm exams 1`:`Preparation to midterm exams 2` `Additional work`:`Listening in classes`:`Taking notes
in classes`

       Test Summary
 -------------------------
 Num DF     =    18
 Den DF     =    126
 F          =    1.897907
 Prob > F   =    0.02148223
```

There isn't enough evidence to conclude that there is a significant association between the categorical variables being analyzed in the chi-squared test.

Summary

To summarize our report, we have learned the different factors that contribute to student academic performance. An in-depth analysis was conducted to investigate student performance towards their final GPAs. After rigorous analysis, it has been discovered that the models that were employed in this study did not yield substantial evidence to support the initial hypotheses or provide strong conclusions.
Despite the use of chi-square and correlation tests, the only highly significant variable was the 'cumulative average credit of the previous semester', indicating a lack of significant patterns in the dataset explored, thus limiting our analyses. The linear regression model yielded a multiple R-square value of 0.4397 indicating limited explanatory power of about 43.97 per cent of the variation in students' performance based on the selected characteristics.

References

Adachi-Mejia, A.M. *et al.* (2014) *The relative roles of types of extracurricular activity on smoking and drinking initiation among tweens*, *Academic pediatrics*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4096144/ (Accessed: 31 October 2023).

Frenette, M. and Chan, P.C.W. (2015) *Academic outcomes of public and private high school students: What ... - ed*, *Academic Outcomes of Public and Private High School Students: What Lies Behind the Differences?* . Available at: https://files.eric.ed.gov/fulltext/ED585228.pdf (Accessed: 31 October 2023).

Harry, L. (2016) *The effects of school type on academic performance-evidence from the ...*, *THE EFFECTS OF SCHOOL TYPE ON ACADEMIC PERFORMANCE-EVIDENCE FROM THE SECONDARY ENTRANCE ASSESSMENT EXAM IN TRINIDAD*. Available at: https://files.eric.ed.gov/fulltext/ED586316.pdf (Accessed: 31 October 2023).

Hoxby, C.M. (1994) *Does competition among public schools benefit students and taxpayers?*, *NBER*. Available at: https://www.nber.org/papers/w4979 (Accessed: 31 October 2023).

Lippman, L. *et al.* (2008) *Parent expectations and planning for College Statistical Analysis Report*, *Parent Expectations and Planning for College Statistical Analysis Report*. Available at: https://files.eric.ed.gov/fulltext/ED501131.pdf (Accessed: 31 October 2023).

Reeves, D. (2008) *(PDF) the learning leader / the extracurricular advantage - researchgate*, *The Learning Leader / The Extracurricular Advantage*. Available at: https://www.researchgate.net/publication/242115973_The_Learning_Leader_The_Extracurricular_Advantage (Accessed: 31 October 2023).

Group Contribution

Student ID	Name	Contribution	Signature
S11188808	Liu Ying	40%	LYing
S11187423	Roska Takayawa	40%	RTakayawa
S11199333	Ravneel Sewak	20%	RSewak