Final Project Proposal

IS-680/IT-634: Introduction to Data Mining and Visualization

Group Members: Carol Liang, James Louis, Vaibhav Mayekar, Marissa McCaffery, John Melo, & Aeraj Arib Baig Mirza

Topic Introduction

For years it has been widely known that teacher engagement and motivation, parental engagement, and financial status has a significant impact on student academic performance. It has also been proven that an increase of usage and access to technology has had a positive impact on student performance worldwide. According to a study conducted in Portugal, students who have easier and higher rates of internet access typically have better academic outcomes (Nunes et al, 2023).

Although current research has proven the fact that the internet, parent, and teacher involvement have helped students achieve better outcomes, the current research does not speak specifically to whether parent engagement has an effect on students' study habits. Current studies also do not answer whether study habits vary by students from different financial backgrounds and if different study habits affect academic performance.

Our research will provide insight on which study habits are more effective in providing positive academic outcomes to students of a lower financial background. This research will also determine whether there is a correlation between specific study habits and academic performance, and whether this varies by financial status. Finally, this research will determine the degree to which parental engagement varies among the different financial status groups.

Dataset Description

Dataset Source: ***Student Behavior:*** *Understanding Students: Behaviors, Preferences, and Expectations*

https://www.kaggle.com/datasets/gunapro/student-behavior

The dataset utilized for this project has been sourced from Kaggle and encompasses the information of 235 university students. This dataset includes their responses to inquiries about various aspects, including their financial and academic backgrounds, as well as personal preferences related to their academic pursuits and life in general.

Within this dataset, you will find the following variables, with its details provided below:

1) *Certification course*: (Yes/No) if the student has completed any certification course or not.

2) *Gender*: (Male/Female) the gender of the student.

3) *Department*: (BCA/B.com Accounting and Finance/B.com ISM/Commerce) the department of field or study the student is enrolled in.

4) *Height*: the height of the student in centimeters.

5) *Weight*: the weight of the student in kilograms.

6) *10th grade mark*: (From 1-100 range) student's 10th grade score obtained.

7) *12th grade mark*: (From 1-100 range) student's 12th grade score obtained.

8) *College mark*: (From 1-100 range) student's college or university score obtained.

9) *Hobbies*: (Cinema/ Sports/ Reading books/ Video games) 4 options of habits or interest were provided for the student who took the survey to select.

10) *Daily studying time*: the amount of time that each student spends studying each day, the following time range was provided for the students to select.

a. 0-30 minutes
b. 30-60 minutes
c. 1-2 hours
d. 2-3 hours
e. 3-4 hours
f. More than 4 hours

11) *Prefer to study in*: student's preferred study environment, options including the following:

a. Moring
b. Night
c. Anytime

12) *Salary expectation*: the expected number of salaries of the student.

13) '*Do you like your degree?' question*: (Yes/No) indicates if the student likes the currently pursuing degree or not.

14) *Willingness to pursue a career based on their degree*: (From 0% - 100% range): indicates if the student wants to work in the field that they are currently studying for in the future.

15) *Social media & video*: the time that students engage with social media platforms each day. The following time range was provided to the students to select on the survey.

a. 0 minute
b. 1-30 minutes
c. 30 – 60 minutes
d. 1 – 1.3 hours
e. 1.3 – 2 hours
f. More than 2 hours

16) *Traveling time*: the travel time of students commuting to their school every day. The following time range was provided to the students to select on the survey.

a. 0 minute
b. 30 – 60 minutes
c. 1 – 1.3 hours
d. 1.3 – 2 hours
e. 2-2.3 hours
f. More than 3 hours

17) *Stress level*: the perceived stress level of the student. The following options were provided to the student to select.

a. Good
b. Fabulous
c. Bad
d. Awful

18) *Financial status*: the current financial status of the student. The following options were provided for the student to select.

a. Good
b. Fabulous
c. Bad
d. Awful

19) *Part-time*: (Yes/No): indicates if the student has a part-time job or not.

Dependent Variables and Independent Variables

Dependent Variable:

- **College Mark**: This represents the student's academic performance in their college or university, serving as the primary focus of the analysis to understand their achievements.

Independent Variables:

- **Financial Status**: Students from more financially stable backgrounds might have better access to better resources and support, positively impacting academic performance and leading to higher college marks.
- **Social Media & Video**: Excessive engagement with social media and video platforms might potentially detract from study time and focus, leading to lower college marks.
- **Certification Course** : The completion of certification courses might positively correlate with higher college marks, indicating that students with additional certifications exhibit better academic performance.
- **Stress Level**: Higher stress levels might negatively impact focus and academic performance, potentially resulting in lower college marks.

Through analysis of our dataset, our research will aim to address how financial status and study habits jointly influence academic performance.

```
# import Libraries
library(dplyr)
library(ggplot2)
library(tidyverse)

# Read the CSV file
df  <- read.csv("/content/Student_Behaviour.csv")
data <- read.csv("/content/Student_Behaviour.csv", header = TRUE)
```

A data.frame: 235 × 21

Certification_Course	Gender	Department	Height.CM.	Weight.KG.	Tenth_Mark	Twelfth_Mark	college_Mark	hobbies	X	⋯	Daily_Study_Time	prefer.to.study.in	salary.expectation	Do.you.like.your.degree.	willingness.to.pursue.a.career.based.on.their.degree	social.media...video	Travelling.Time	Stress.Level	Financial.Status	part.time.job
<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<lgl>	⋯	<int>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
No	Male	BCA	100	58	79.0	65.00	80	Video Games	NA	⋯	30	Morning	40000	No	50%	1.30 - 2 hour	30 - 60 minutes	Bad	Bad	No
No	Female	BCA	90	40	70.0	80.00	70	Cinema	NA	⋯	60	Morning	15000	Yes	75%	1 - 1.30 hour	0 - 30 minutes	Bad	Bad	No
Yes	Male	BCA	159	78	69.5	61.00	55	Cinema	NA	⋯	120	Anytime	13000	Yes	50%	More than 2 hour	30 - 60 minutes	Awful	Bad	No
Yes	Female	BCA	147	20	70.0	59.00	58	Reading books	NA	⋯	120	Anytime	1500000	No	50%	1.30 - 2 hour	0 - 30 minutes	Bad	good	No
No	Male	BCA	170	54	40.0	65.00	30	Video Games	NA	⋯	60	Morning	50000	Yes	25%	1.30 - 2 hour	30 - 60 minutes	Good	good	No
Yes	Female	BCA	139	33	90.0	75.00	70	Cinema	NA	⋯	60	Night	20000	Yes	50%	30 - 60 Minute	0 - 30 minutes	Bad	good	No
Yes	Male	BCA	165	50	70.0	63.00	3	Reading books	NA	⋯	120	Morning	15000	Yes	75%	1 - 1.30 hour	1 - 1.30 hour	Good	good	No
No	Male	BCA	152	43	61.6	61.67	75	Sports	NA	⋯	120	Morning	25000	Yes	50%	1 - 1.30 hour	1 - 1.30 hour	Good	good	No
No	Male	BCA	190	85	88.2	67.50	60	Video Games	NA	⋯	30	Morning	20000	Yes	75%	1.30 - 2 hour	1 - 1.30 hour	Bad	Bad	No
No	Male	BCA	150	84	60.0	65.00	70	Video Games	NA	⋯	30	Anytime	20000	Yes	75%	30 - 60 Minute	30 - 60 minutes	Good	good	No
No	Male	BCA	99	50	75.0	70.00	60	Video Games	NA	⋯	120	Morning	18000	Yes	75%	1 - 1.30 hour	0 - 30 minutes	Good	Bad	No
No	Female	BCA	152	51	60.0	56.00	65	Cinema	NA	⋯	180	Night	20000	Yes	75%	1 - 1.30 hour	2 - 2.30 hour	Bad	Bad	No
Yes	Female	BCA	158	64	59.0	54.00	80	Reading books	NA	⋯	60	Morning	18000	Yes	75%	1 - 30 Minute	30 - 60 minutes	Bad	good	No
No	Male	BCA	160	52	50.0	55.00	50	Cinema	NA	⋯	30	Anytime	17	Yes	75%	30 - 60 Minute	1.30 - 2 hour	Good	Awful	No
Yes	Male	BCA	160	83	89.5	69.70	70	Cinema	NA	⋯	60	Anytime	60000	Yes	75%	1 - 30 Minute	30 - 60 minutes	Good	good	No
Yes	Male	BCA	156	68	54.0	48.00	60	Video Games	NA	⋯	60	Anytime	18000	Yes	75%	1 - 1.30 hour	more than 3 hour	Good	Bad	No
Yes	Male	BCA	153	52	83.0	76.00	75	Sports	NA	⋯	120	Anytime	15000	Yes	75%	1.30 - 2 hour	30 - 60 minutes	Good	good	Yes
Yes	Male	BCA	162	60	62.0	61.30	53	Cinema	NA	⋯	30	Night	0	Yes	100%	1 - 1.30 hour	1 - 1.30 hour	Bad	Bad	No
No	Male	BCA	150	60	60.0	65.00	60	Cinema	NA	⋯	30	Anytime	30000	Yes	100%	1 - 30 Minute	1 - 1.30 hour	Bad	Bad	No
Yes	Female	BCA	109	42	60.0	70.00	88	Cinema	NA	⋯	60	Anytime	20000	Yes	100%	1 - 30 Minute	30 - 60 minutes	Good	good	No
Yes	Male	BCA	160	50	65.0	65.00	50	Sports	NA	⋯	60	Anytime	10000	Yes	75%	30 - 60 Minute	1 - 1.30 hour	Bad	Bad	Yes
No	Male	BCA	154	51	65.0	70.00	50	Cinema	NA	⋯	120	Morning	20000	Yes	50%	1 - 30 Minute	1 - 1.30 hour	Good	Bad	No
No	Female	BCA	142	42	78.0	60.00	60	Cinema	NA	⋯	120	Anytime	25000	Yes	75%	1 - 1.30 hour	1 - 1.30 hour	Bad	good	No
No	Male	BCA	180	90	75.0	73.80	65	Video Games	NA	⋯	120	Morning	18000	Yes	50%	1 - 1.30 hour	30 - 60 minutes	Good	good	No
No	Male	BCA	167	85	75.0	60.00	80	Cinema	NA	⋯	30	Anytime	20000	Yes	50%	1.30 - 2 hour	2 - 2.30 hour	Awful	good	No
No	Male	BCA	90	89	60.0	50.00	50	Sports	NA	⋯	60	Morning	18000	Yes	75%	30 - 60 Minute	30 - 60 minutes	Bad	good	Yes
Yes	Male	BCA	175	73	90.0	90.00	80	Cinema	NA	⋯	120	Anytime	30000	Yes	50%	0 Minute	1 - 1.30 hour	Good	Bad	No
No	Male	BCA	151	69	73.0	85.00	55	Sports	NA	⋯	30	Morning	15000	Yes	100%	1 - 1.30 hour	30 - 60 minutes	Good	good	No
No	Female	BCA	145	39	89.0	77.00	65	Sports	NA	⋯	60	Night	20000	Yes	100%	0 Minute	1 - 1.30 hour	Awful	Bad	Yes
No	Female	BCA	99	40	94.0	94.00	80	Reading books	NA	⋯	60	Night	15000	Yes	75%	1 - 1.30 hour	30 - 60 minutes	Good	Bad	No
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Yes	Male	Commerce	161	55	89.0	79.0	70	Sports	NA	⋯	60	Morning	10000	Yes	75%	30 - 60 Minute	30 - 60 minutes	Bad	Bad	No
Yes	Female	Commerce	146	39	95.6	86.0	85	Cinema	NA	⋯	120	Anytime	13000	Yes	75%	30 - 60 Minute	0 - 30 minutes	Good	good	No
Yes	Male	Commerce	155	70	60.0	60.0	60	Sports	NA	⋯	180	Night	15000	Yes	50%	More than 2 hour	1 - 1.30 hour	Good	good	Yes
No	Male	Commerce	120	80	81.0	80.0	70	Cinema	NA	⋯	60	Night	20000	Yes	50%	More than 2 hour	0 - 30 minutes	Good	Fabulous	Yes
Yes	Male	Commerce	160	70	84.0	76.0	70	Video Games	NA	⋯	60	Anytime	15000	Yes	75%	More than 2 hour	1 - 1.30 hour	Bad	Bad	Yes
Yes	Female	Commerce	152	32	80.0	60.0	70	Sports	NA	⋯	120	Night	15	Yes	75%	1 - 1.30 hour	30 - 60 minutes	Bad	good	No
No	Male	Commerce	168	75	75.0	65.0	75	Cinema	NA	⋯	120	Morning	22	Yes	75%	1 - 1.30 hour	1 - 1.30 hour	Good	good	No
Yes	Female	Commerce	152	64	90.4	80.0	80	Cinema	NA	⋯	120	Anytime	15000	Yes	75%	30 - 60 Minute	1 - 1.30 hour	Good	good	No
No	Male	Commerce	156	45	85.0	75.0	70	Sports	NA	⋯	300	Anytime	12000	Yes	75%	1 - 1.30 hour	2 - 2.30 hour	Good	Bad	Yes
Yes	Female	Commerce	151	55	93.8	83.5	100	Reading books	NA	⋯	120	Morning	15000	Yes	75%	1 - 1.30 hour	30 - 60 minutes	Good	good	Yes
No	Female	Commerce	165	62	78.0	75.0	90	Sports	NA	⋯	180	Anytime	15000	Yes	100%	30 - 60 Minute	0 - 30 minutes	Good	good	No
Yes	Female	Commerce	160	60	75.0	85.0	90	Cinema	NA	⋯	180	Anytime	20000	Yes	75%	30 - 60 Minute	30 - 60 minutes	Good	good	No
Yes	Female	Commerce	153	58	85.0	74.0	75	Cinema	NA	⋯	30	Anytime	20000	Yes	75%	More than 2 hour	0 - 30 minutes	Awful	good	No
Yes	Male	Commerce	175	75	88.0	85.0	90	Sports	NA	⋯	60	Morning	20000	Yes	75%	1 - 30 Minute	30 - 60 minutes	Good	good	No
No	Male	Commerce	161	54	90.0	73.0	80	Cinema	NA	⋯	60	Morning	20000	Yes	100%	30 - 60 Minute	more than 3 hour	Bad	Bad	No
Yes	Male	Commerce	160	60	85.0	80.0	80	Cinema	NA	⋯	60	Anytime	20000	Yes	100%	1 - 30 Minute	30 - 60 minutes	Awful	Bad	No
Yes	Female	Commerce	153	45	83.0	73.0	75	Cinema	NA	⋯	30	Morning	20000	Yes	75%	30 - 60 Minute	0 - 30 minutes	Good	good	No
Yes	Male	Commerce	160	50	85.0	80.0	60	Sports	NA	⋯	60	Night	100	Yes	75%	1 - 1.30 hour	0 - 30 minutes	Good	Bad	No
Yes	Male	Commerce	170	65	82.0	73.0	80	Cinema	NA	⋯	30	Anytime	16000	Yes	50%	1.30 - 2 hour	0 - 30 minutes	fabulous	Bad	No
Yes	Female	Commerce	152	65	97.0	80.0	100	Reading books	NA	⋯	240	Morning	15000	Yes	100%	30 - 60 Minute	0 - 30 minutes	Good	good	Yes
Yes	Female	Commerce	164	43	80.0	89.0	95	Reading books	NA	⋯	180	Morning	15000	Yes	100%	1 - 30 Minute	0 - 30 minutes	Good	good	No
Yes	Female	Commerce	167	65	93.0	85.0	90	Reading books	NA	⋯	120	Night	15000	Yes	75%	30 - 60 Minute	30 - 60 minutes	Good	good	No
Yes	Male	B.com ISM	167	55	55.0	57.0	60	Sports	NA	⋯	60	Anytime	17000	Yes	25%	0 Minute	0 - 30 minutes	Awful	Bad	No
Yes	Male	B.com ISM	155	68	75.0	75.0	65	Sports	NA	⋯	120	Anytime	500	Yes	100%	1.30 - 2 hour	1 - 1.30 hour	Good	good	No
No	Female	B.com ISM	153	60	60.0	70.0	70	Sports	NA	⋯	120	Anytime	15000	Yes	75%	1 - 1.30 hour	0 - 30 minutes	Good	Bad	No
Yes	Male	B.com ISM	170	76	72.0	67.0	65	Video Games	NA	⋯	60	Morning	7000	Yes	50%	More than 2 hour	30 - 60 minutes	Bad	Bad	No
Yes	Male	B.com ISM	172	52	72.0	70.0	76	Cinema	NA	⋯	180	Anytime	25000	Yes	75%	30 - 60 Minute	30 - 60 minutes	Good	good	No
Yes	Female	BCA	139	33	90.0	75.0	70	Cinema	NA	⋯	60	Night	20000	Yes	50%	30 - 60 Minute	2.30 - 3 hour	Bad	good	No
Yes	Female	Commerce	153	58	85.0	74.0	75	Cinema	NA	⋯	30	Anytime	20000	Yes	75%	More than 2 hour	2.30 - 3 hour	Awful	good	No
No	Female	B.com ISM	155	39	45.0	45.0	50	Sports	NA	⋯	240	Night	10	No	75%	30 - 60 Minute	2.30 - 3 hour	fabulous	good	No

```
# Display the first few rows
head(df)
```

A data.frame: 6 × 22

	Certification.Course	Gender	Department	Height.CM.	Weight.KG.	X10th.Mark	X12th.Mark	college.mark	hobbies	daily.studing.time	⋯	Do.you.like.your.degree.	willingness.to.pursue.a.career.based.on.their.degree	social.media...video	Travelling.Time	Stress.Level	Financial.Status	part.time.job	Encoded.Stress.Level	Numeric.Stress.Level	Binary.Stress.Level
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	⋯	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<ord>	<dbl>	<fct>
1	No	Male	BCA	100	58	79.0	65	80	Video Games	0 - 30 minute	⋯	No	50%	1.30 - 2 hour	30 - 60 minutes	Bad	Bad	No	1	1	High Stress
2	No	Female	BCA	90	40	70.0	80	70	Cinema	30 - 60 minute	⋯	Yes	75%	1 - 1.30 hour	0 - 30 minutes	Bad	Bad	No	1	1	High Stress
3	Yes	Male	BCA	159	78	69.5	61	55	Cinema	1 - 2 Hour	⋯	Yes	50%	More than 2 hour	30 - 60 minutes	Awful	Bad	No	0	0	High Stress
4	Yes	Female	BCA	147	20	70.0	59	58	Reading books	1 - 2 Hour	⋯	No	50%	1.30 - 2 hour	0 - 30 minutes	Bad	good	No	1	1	High Stress
5	No	Male	BCA	170	54	40.0	65	30	Video Games	30 - 60 minute	⋯	Yes	25%	1.30 - 2 hour	30 - 60 minutes	Good	good	No	2	2	Low Stress
6	Yes	Female	BCA	139	33	90.0	75	70	Cinema	30 - 60 minute	⋯	Yes	50%	30 - 60 Minute	0 - 30 minutes	Bad	good	No	1	1	High Stress

```r
# Display the structure
str(df)
```

```
'data.frame':   235 obs. of  21 variables:
 $ Certification_Course                              : chr  "No" "No" "Yes" "Yes" ...
 $ Gender                                            : chr  "Male" "Female" "Male" "Female" ...
 $ Department                                        : chr  "BCA" "BCA" "BCA" "BCA" ...
 $ Height.CM.                                        : num  100 90 159 147 170 139 165 152 190 150 ...
 $ Weight.KG.                                        : num  58 40 78 20 54 33 50 43 85 84 ...
 $ Tenth_Mark                                        : num  79 70 69.5 70 40 90 70 61.6 88.2 60 ...
 $ Twelfth_Mark                                      : num  65 80 61 59 65 ...
 $ college_Mark                                      : num  80 70 55 58 30 70 3 75 60 70 ...
 $ hobbies                                           : chr  "Video Games" "Cinema" "Cinema" "Reading books" ...
 $ X                                                 : logi  NA NA NA NA NA NA ...
 $ daily.studing.time                                : chr  "0 - 30 minute" "30 - 60 minute" "1 - 2 Hour" "1 - 2 Hour" ...
 $ Daily_Study_Time                                  : int  30 60 120 120 60 60 120 120 30 30 ...
 $ prefer.to.study.in                                : chr  "Morning" "Morning" "Anytime" "Anytime" ...
 $ salary.expectation                                : int  40000 15000 13000 1500000 50000 20000 15000 25000 20000 20000 ...
 $ Do.you.like.your.degree.                          : chr  "No" "Yes" "Yes" "No" ...
 $ willingness.to.pursue.a.career.based.on.their.degree: chr  "50%" "75%" "50%" "50%" ...
 $ social.medai...video                              : chr  "1.30 - 2 hour" "1 - 1.30 hour" "More than 2 hour" "1.30 - 2 hour" ...
 $ Travelling.Time                                   : chr  "30 - 60 minutes" "0 - 30 minutes" "30 - 60 minutes" "0 - 30 minutes" ...
 $ Stress.Level                                      : chr  "Bad" "Bad" "Awful" "Bad" ...
 $ Financial.Status                                  : chr  "Bad" "Bad" "Bad" "good" ...
 $ part.time.job                                     : chr  "No" "No" "No" "No" ...
```

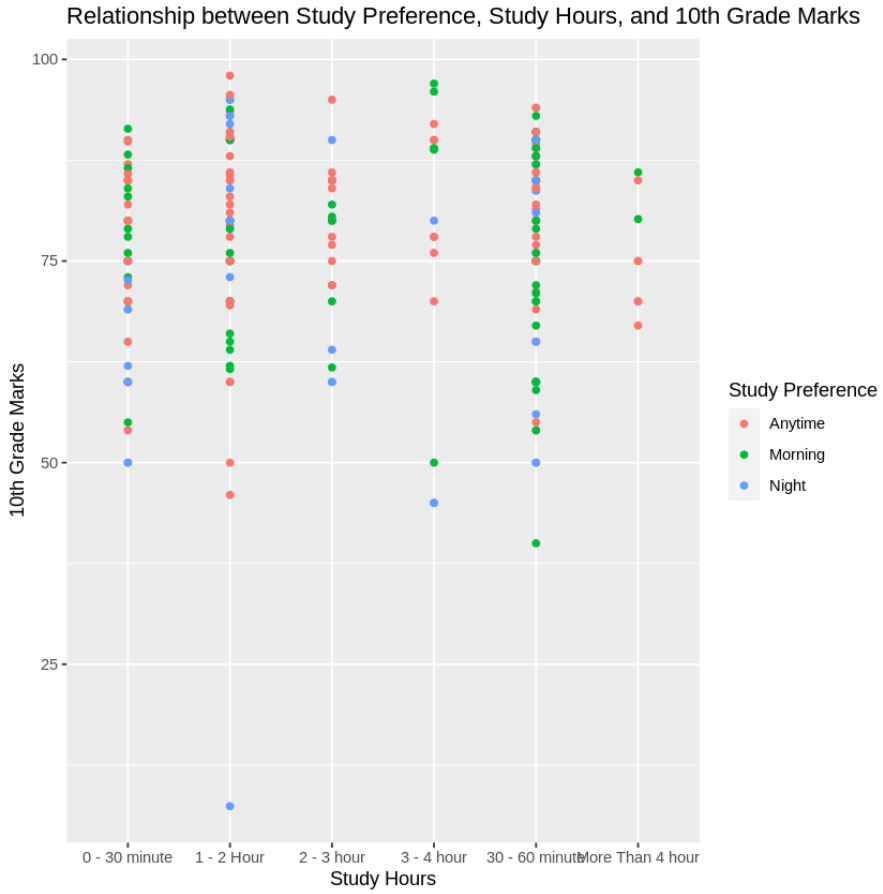
```r
# summary of the  data
summary(df)
```

```
 Certification_Course    Gender           Department          Height.CM.
 Length:235          Length:235         Length:235         Min.   :  4.5
 Class :character    Class :character   Class :character   1st Qu.:152.0
 Mode  :character    Mode  :character   Mode  :character   Median :160.0
                                                           Mean   :157.4
                                                           3rd Qu.:170.0
                                                           Max.   :192.0
   Weight.KG.      Tenth_Mark     Twelfth_Mark    college_Mark
 Min.   : 20.0   Min.   : 7.40   Min.   :45.00   Min.   :  1.00
 1st Qu.: 50.0   1st Qu.:70.00   1st Qu.:60.00   1st Qu.: 60.00
 Median : 60.0   Median :80.00   Median :69.00   Median : 70.00
 Mean   : 60.8   Mean   :76.84   Mean   :68.78   Mean   : 70.66
 3rd Qu.: 70.0   3rd Qu.:86.25   3rd Qu.:76.00   3rd Qu.: 80.00
 Max.   :106.0   Max.   :98.00   Max.   :94.00   Max.   :100.00
   hobbies             X           daily.studing.time Daily_Study_Time
 Length:235         Mode:logical   Length:235         Min.   : 30.0
 Class :character   NA's:235       Class :character   1st Qu.: 60.0
 Mode  :character                  Mode  :character   Median : 60.0
                                                      Mean   :101.6
                                                      3rd Qu.:120.0
                                                      Max.   :300.0
 prefer.to.study.in salary.expectation Do.you.like.your.degree.
 Length:235         Min.   :      0    Length:235
 Class :character   1st Qu.:  15000    Class :character
 Mode  :character   Median :  20000    Mode  :character
                    Mean   :  32482
                    3rd Qu.:  25000
                    Max.   :1500000
 willingness.to.pursue.a.career.based.on.their.degree social.medai...video
 Length:235                                           Length:235
 Class :character                                     Class :character
 Mode  :character                                     Mode  :character


 Travelling.Time    Stress.Level       Financial.Status   part.time.job
 Length:235         Length:235         Length:235         Length:235
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

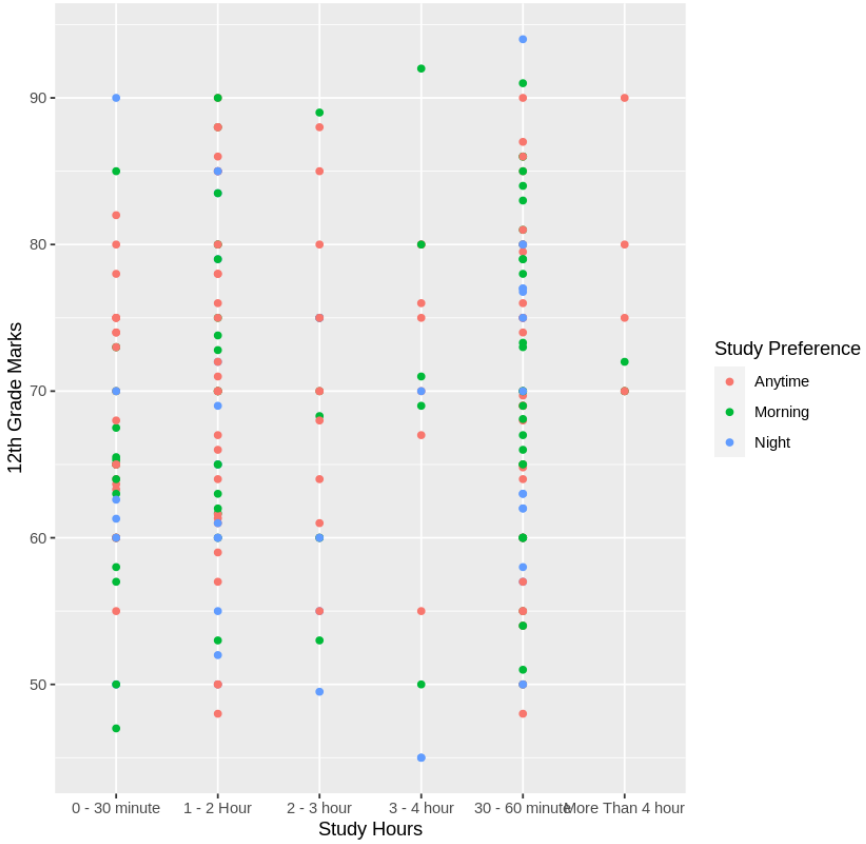
1A. How do different study habits relate to academic performance? Does this relationship vary by financial status? Does the amount of study time impact performance?

```
# Load library
library(ggplot2)

# Explore the relationship between study preference, study hours, and
10th grade marks
ggplot(df, aes(x = daily.studing.time, y = X10th.Mark, color =
prefer.to.study.in)) +
  geom_point() +
  labs(title = "Relationship between Study Preference, Study Hours, and
10th Grade Marks", x = "Study Hours", y = "10th Grade Marks", color =
"Study Preference")

# Explore the relationship between study preference, study hours, and
12th grade marks
ggplot(df, aes(x = daily.studing.time, y = X12th.Mark, color =
prefer.to.study.in)) +
  geom_point() +
  labs(title = "Relationship between Study Preference, Study Hours, and
12th Grade Marks", x = "Study Hours", y = "12th Grade Marks", color =
"Study Preference")

# Explore the relationship between study preference, study hours, and
college marks
ggplot(df, aes(x = daily.studing.time, y = college.mark, color =
prefer.to.study.in)) +
  geom_point() +
  labs(title = "Relationship between Study Preference, Study Hours, and
College Marks", x = "Study Hours", y = "College Marks")

# Explore the relationship between financial status and 10th grade
marks
ggplot(df, aes(x = Financial.Status, y = X10th.Mark, fill =
Financial.Status)) +
  geom_boxplot() +
  labs(title = "The Relationship between Financial Status and 10th
Grade Marks", x = "Financial Status", y = "Marks")

# Explore the relationship between financial status and 12th grade
marks
ggplot(df, aes(x = Financial.Status, y = X12th.Mark, fill =
Financial.Status)) +
  geom_boxplot() +
  labs(title = "The Relationship between Financial Status and 12th
Grade Marks", x = "Financial Status", y = "Marks")
```

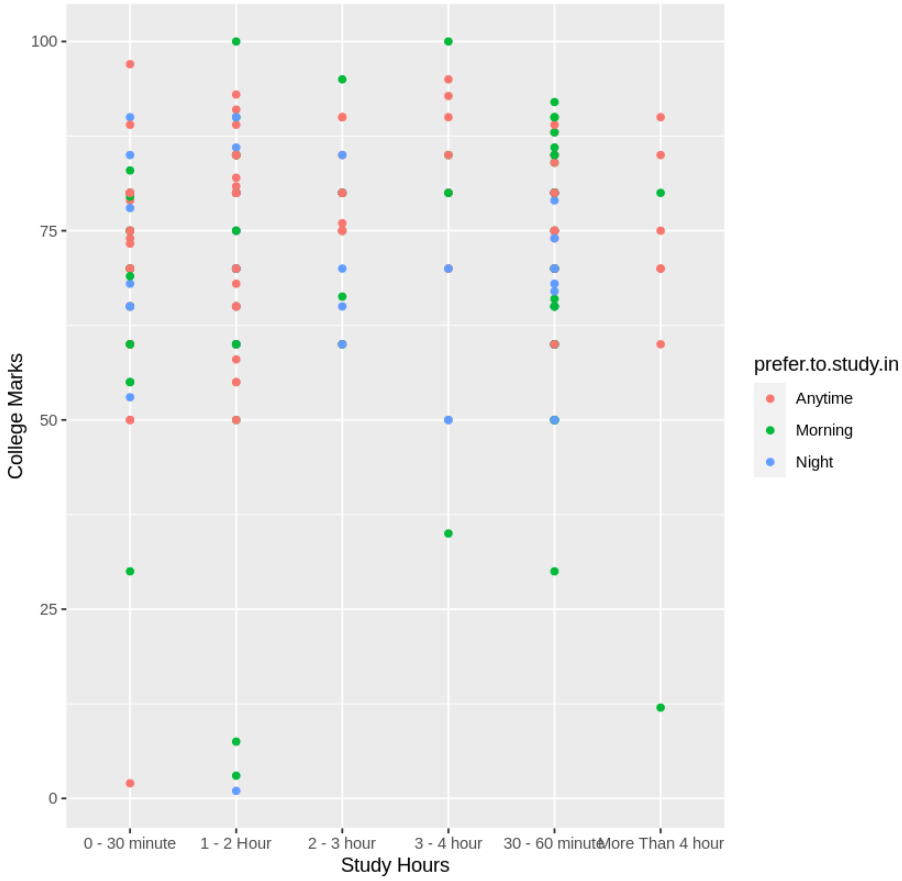
```r
# Explore the relationship between financial and college marks
ggplot(df, aes(x = Financial.Status, y = college.mark, fill =
Financial.Status)) +
  geom_boxplot() +
  labs(title = "The Relationship between Financial Status and College
Marks", x = "Financial Status", y = "Marks")
```



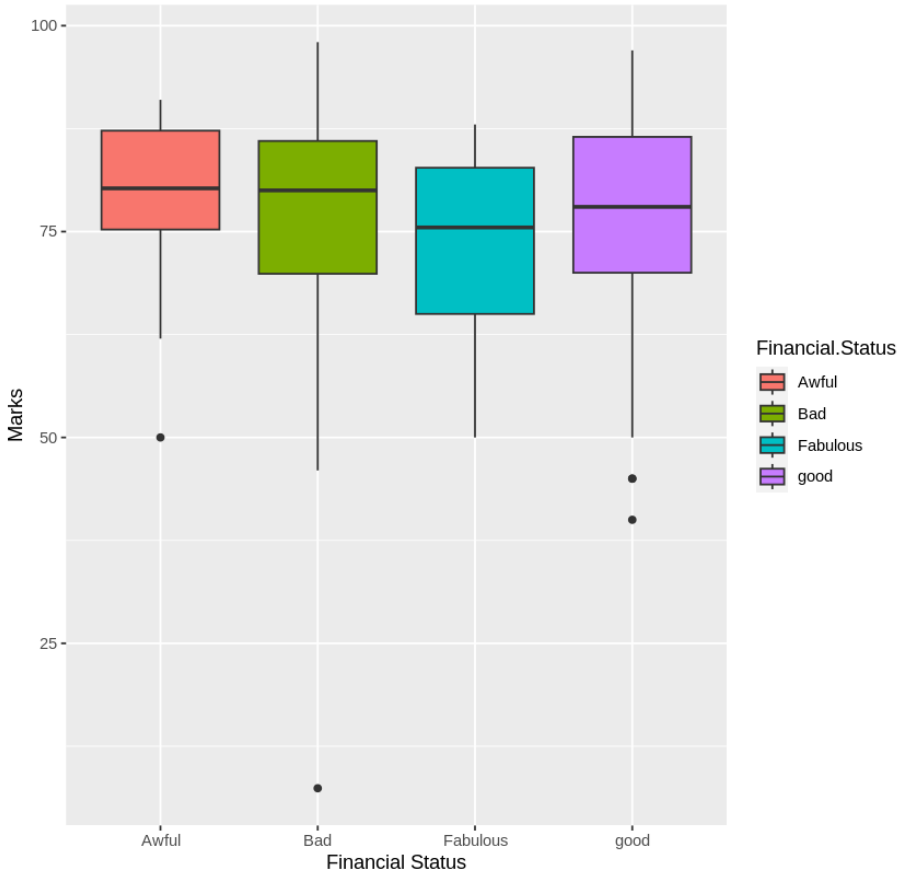
Relationship between Study Preference, Study Hours, and 12th Grade Marks



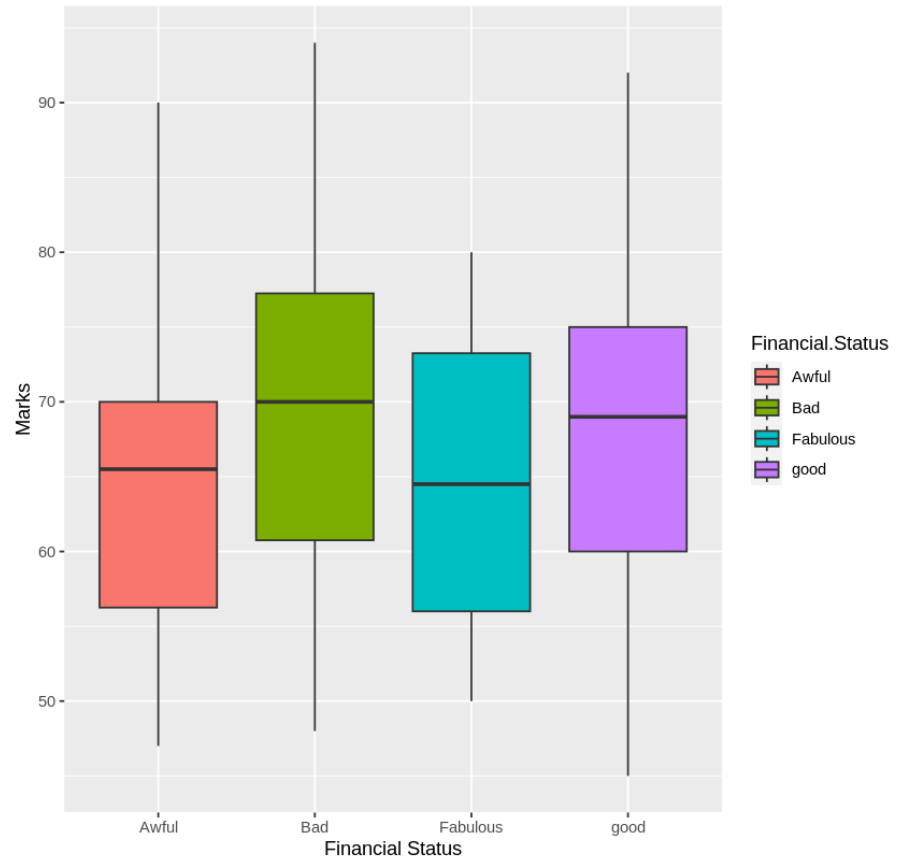
Relationship between Study Preference, Study Hours, and College Marks

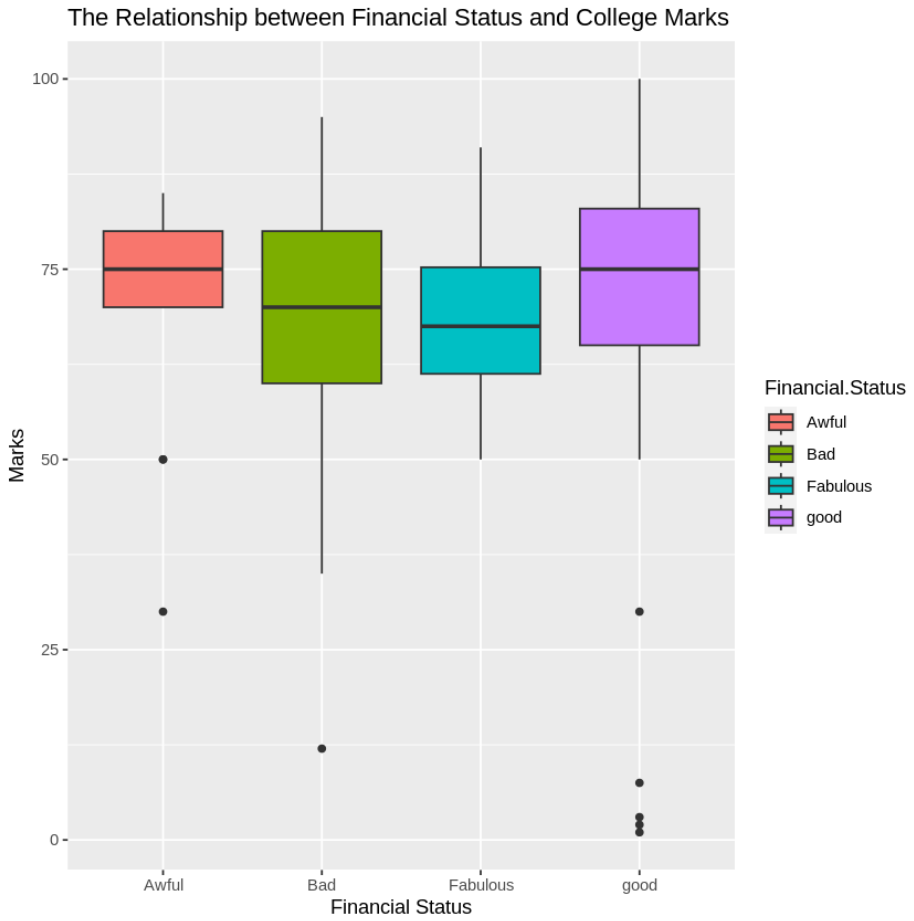


The Relationship between Financial Status and 10th Grade Marks



The Relationship between Financial Status and 12th Grade Marks



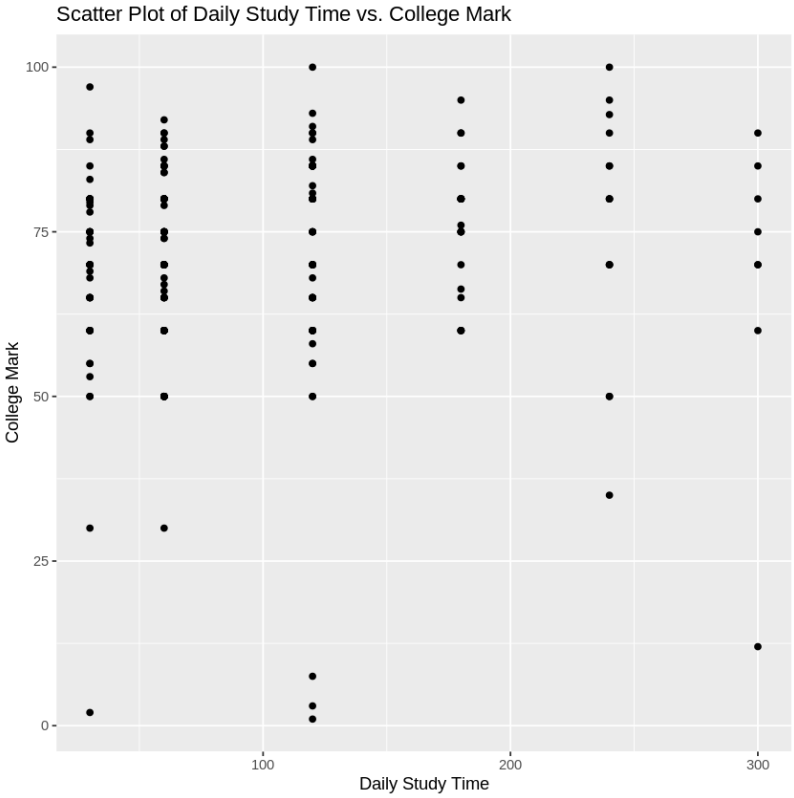


1B. Does the amount of study time impact performance?

```
# Set the path for the dataset
file_path <- "/content/Student_Behaviors.csv"

# Read the CSV file into a data frame
my_data <- read.csv(file_path, header = TRUE)

#Load the necessary libraries for plotting
library(ggplot2)

# Step 1: To visualize the relationship between the college grade and
the each student's daily studying time
    #Create a scatterplot
  ggplot(my_data, aes(x = Daily_Study_Time, y = college_Mark)) +
    geom_point() +
    labs(x = "Daily Study Time", y = "College Mark", title = "Scatter
Plot of Daily Study Time vs. College Mark")

  #Calculate the correlation between college mark and daily studying
time
  correlation <- cor(my_data$Daily_Study_Time, my_data$college_Mark)
  # Print the correlation coefficient
```

```
print(correlation)

  # Use regression analysis to find the relationships betwwen the
college mark and student daily studying time
  lm_model <- lm(college_Mark ~ Daily_Study_Time, data = my_data)
  # Print the summary of the regression model
  summary(lm_model)
```

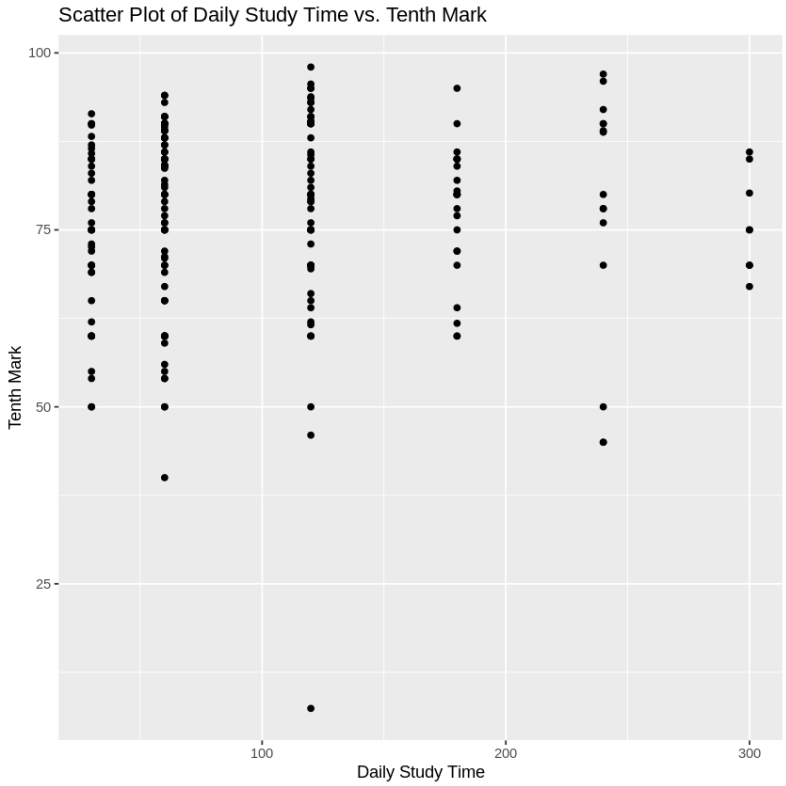
```
[1] 0.0938276

Call:
lm(formula = college_Mark ~ Daily_Study_Time, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-70.048  -9.151   0.849  10.217  28.952

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.51817    1.80711  37.916   <2e-16 ***
Daily_Study_Time  0.02108    0.01466   1.439    0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.69 on 233 degrees of freedom
Multiple R-squared:  0.008804,  Adjusted R-squared:  0.00455
F-statistic: 2.069 on 1 and 233 DF,  p-value: 0.1516
```




```
# Step 2: To visualize the relationship between the 10th grade and the
each student's daily studying time
    #Create a scatterplot
  ggplot(my_data, aes(x = Daily_Study_Time, y = Tenth_Mark)) +
```

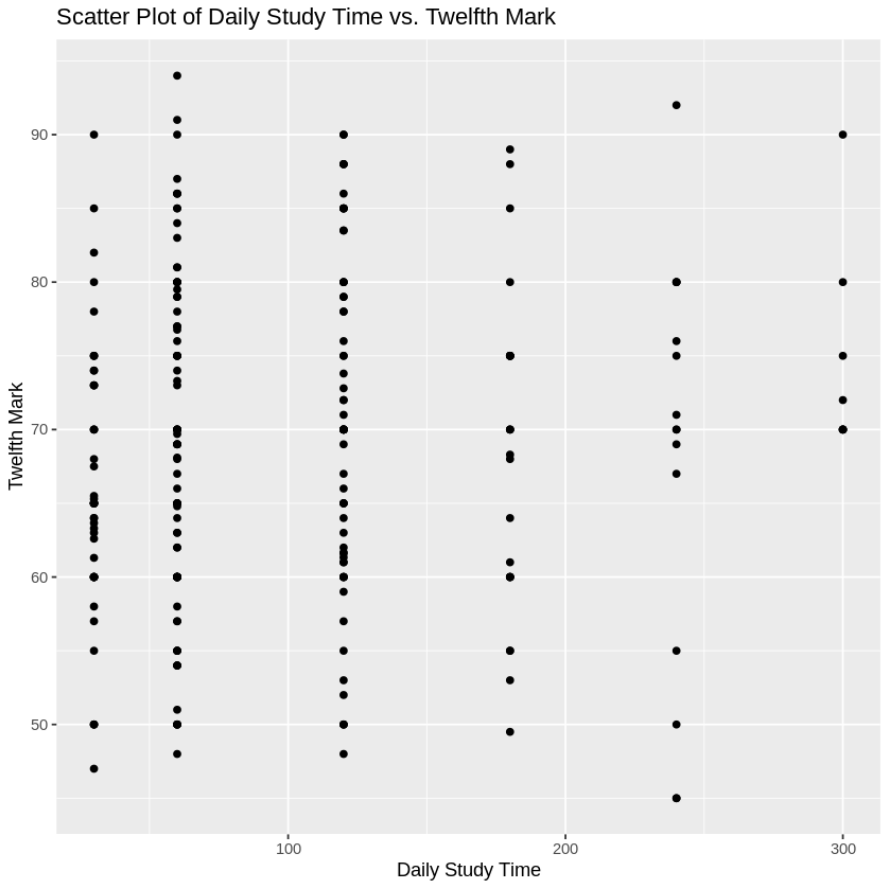
```
    geom_point() +
    labs(x = "Daily Study Time", y = "Tenth Mark", title = "Scatter
Plot of Daily Study Time vs. Tenth Mark")

  #Calculate the correlation between 10th grade mark and daily studying
time
  correlation <- cor(my_data$Daily_Study_Time, my_data$Tenth_Mark)
  # Print the correlation coefficient
  print(correlation)

  # Use regression analysis to find the relationships betwwen the 10th
grade mark and student daily studying time
  lm_model <- lm(Tenth_Mark ~ Daily_Study_Time, data = my_data)
  # Print the summary of the regression model
  summary(lm_model)
```

```
[1] 0.05127898

Call:
lm(formula = Tenth_Mark ~ Daily_Study_Time, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-69.618  -7.018   2.408   9.993  20.982

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      75.870532   1.504224  50.438   <2e-16 ***
Daily_Study_Time  0.009561   0.012199   0.784    0.434
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.06 on 233 degrees of freedom
Multiple R-squared:  0.00263,   Adjusted R-squared:  -0.001651
F-statistic: 0.6143 on 1 and 233 DF,  p-value: 0.434
```




```
# Step 3: To visualize the relationship between the 12th grade and the each student's daily studying time
    #Create a scatterplot
  ggplot(my_data, aes(x = Daily_Study_Time, y = Twelfth_Mark)) +
    geom_point() +
    labs(x = "Daily Study Time", y = "Twelfth Mark", title = "Scatter Plot of Daily Study Time vs. Twelfth Mark")

  #Calculate the correlation between 12th grade mark and daily studying time
  correlation <- cor(my_data$Daily_Study_Time, my_data$Twelfth_Mark)
  # Print the correlation coefficient
  print(correlation)

  # Use regression analysis to find the relationships betwwen the 12th grade mark and student daily studying time
  lm_model <- lm(Twelfth_Mark ~ Daily_Study_Time, data = my_data)
  # Print the summary of the regression model
  summary(lm_model)
```

```
[1] 0.09562254

Call:
lm(formula = Twelfth_Mark ~ Daily_Study_Time, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-25.8628  -8.0568  -0.0568   7.5720  25.8462

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     67.25075    1.26560  53.137   <2e-16 ***
Daily_Study_Time  0.01505    0.01026   1.466    0.144
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.99 on 233 degrees of freedom
Multiple R-squared:  0.009144,  Adjusted R-squared:  0.004891
F-statistic:  2.15 on 1 and 233 DF,  p-value: 0.1439
```

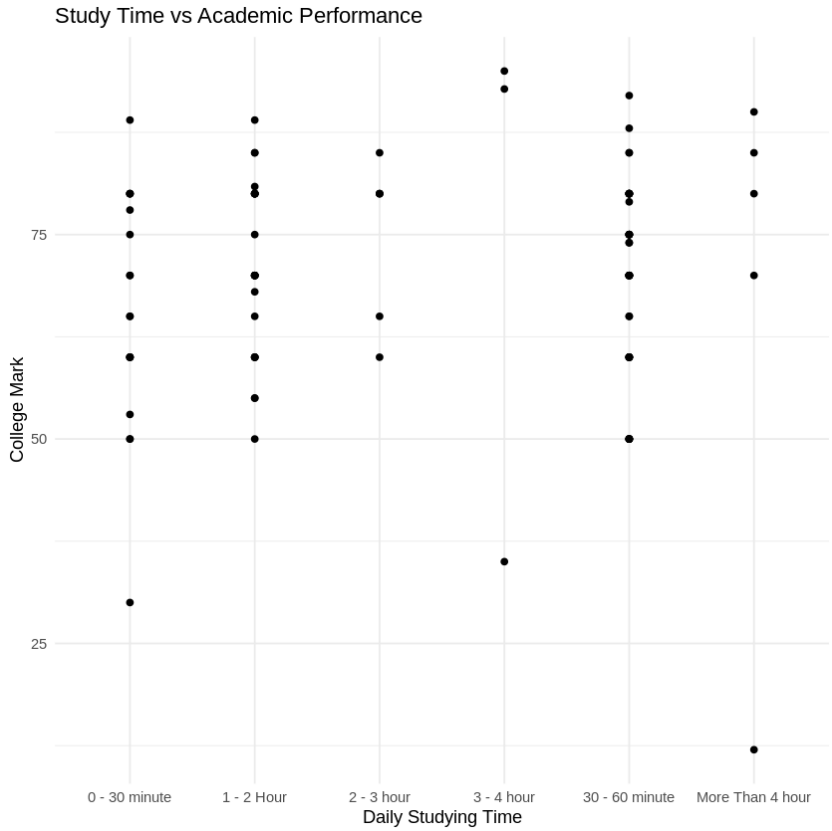


2. Are there specific study habits that are more beneficial for students with limited financial resources in improving their academic performance?

```r
# Load necessary libraries
library(dplyr)
library(ggplot2)
```

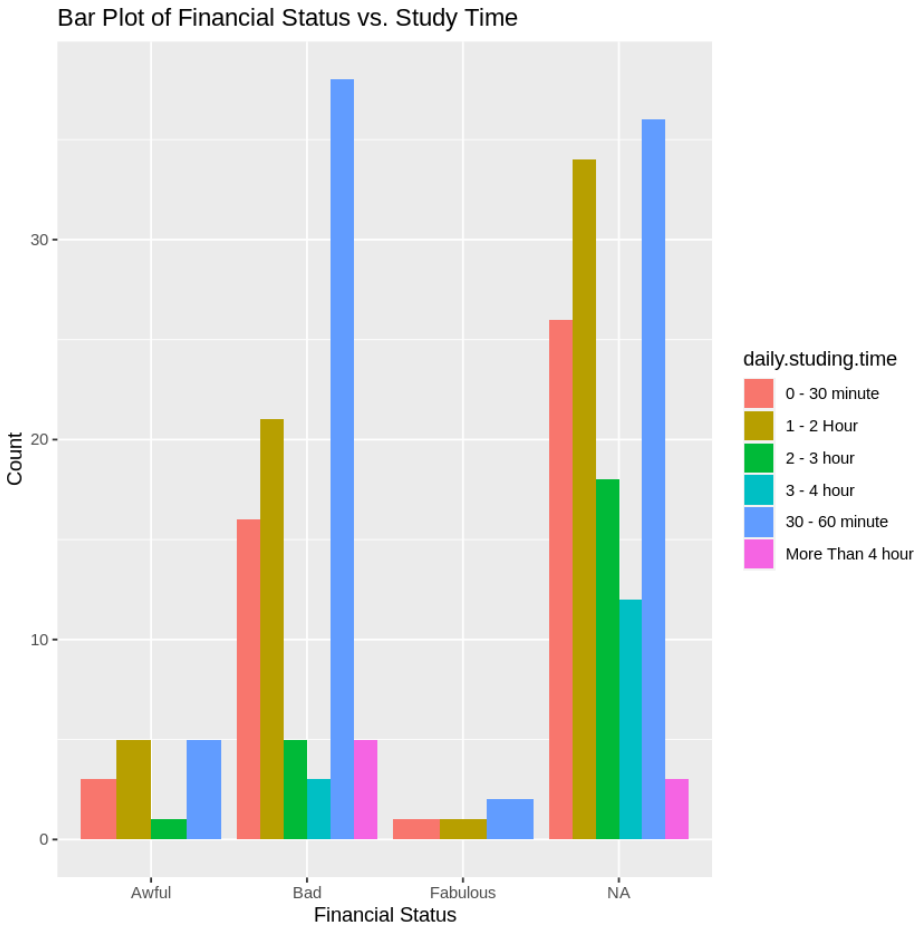
```r
# Step 1: Filter students with limited financial resources
limited_financial_students <- filter(data, Financial.Status == "Bad" |
Financial.Status == "Awful")

# Step 2: Analyze study habits and academic performance using linear
regression
lm_results <- lm(college.mark ~ daily.studing.time + prefer.to.study.in
+ hobbies, data = limited_financial_students)

# Display summary of the linear regression model
summary(lm_results)

# Step 3: Visualize the relationship between daily studying time and
academic performance
ggplot(limited_financial_students, aes(x = daily.studing.time, y =
college.mark)) +
  geom_point() +
  labs(title = "Study Time vs Academic Performance",
       x = "Daily Studying Time",
       y = "College Mark") +
  theme_minimal()
```

```
Call:
lm(formula = college.mark ~ daily.studing.time + prefer.to.study.in +
    hobbies, data = limited_financial_students)

Residuals:
    Min      1Q  Median      3Q     Max
-52.034  -8.313   1.577   8.924  25.314

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      69.1696     3.9619  17.459   <2e-16 ***
daily.studing.time1 - 2 Hour      3.7344     4.3057   0.867   0.3881
daily.studing.time2 - 3 hour      9.7709     6.7219   1.454   0.1495
daily.studing.time3 - 4 hour      6.4476     8.8943   0.725   0.4704
daily.studing.time30 - 60 minute  3.3399     3.8939   0.858   0.3933
daily.studing.timeMore Than 4 hour -0.1161   7.2516  -0.016   0.9873
prefer.to.study.inMorning        -6.7093     3.1179  -2.152   0.0341 *
prefer.to.study.inNight          -4.4543     4.0365  -1.103   0.2727
hobbiesReading books              1.6896     4.5234   0.374   0.7096
hobbiesSports                     0.8859     3.3918   0.261   0.7945
hobbiesVideo Games                0.2424     4.3804   0.055   0.9560
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.94 on 91 degrees of freedom
Multiple R-squared:  0.07913,   Adjusted R-squared:  -0.02207
F-statistic: 0.7819 on 10 and 91 DF,  p-value: 0.6459
```



3. Does financial status correlate with study activity?

```
data <- data %>%
  mutate(Financial.Status = ifelse(Financial.Status %in% c("NA", "NA ",
NA), "Good", as.character(Financial.Status)))

data$Financial.Status <- factor(data$Financial.Status, levels =
c("Awful", "Bad", "Good", "Fabulous"))

library(ggplot2)

ggplot(data, aes(x = Financial.Status, fill = daily.studing.time)) +
  geom_bar(position = "dodge") +
  labs(x = "Financial Status", y = "Count") +
  ggtitle("Bar Plot of Financial Status vs. Study Time")
```

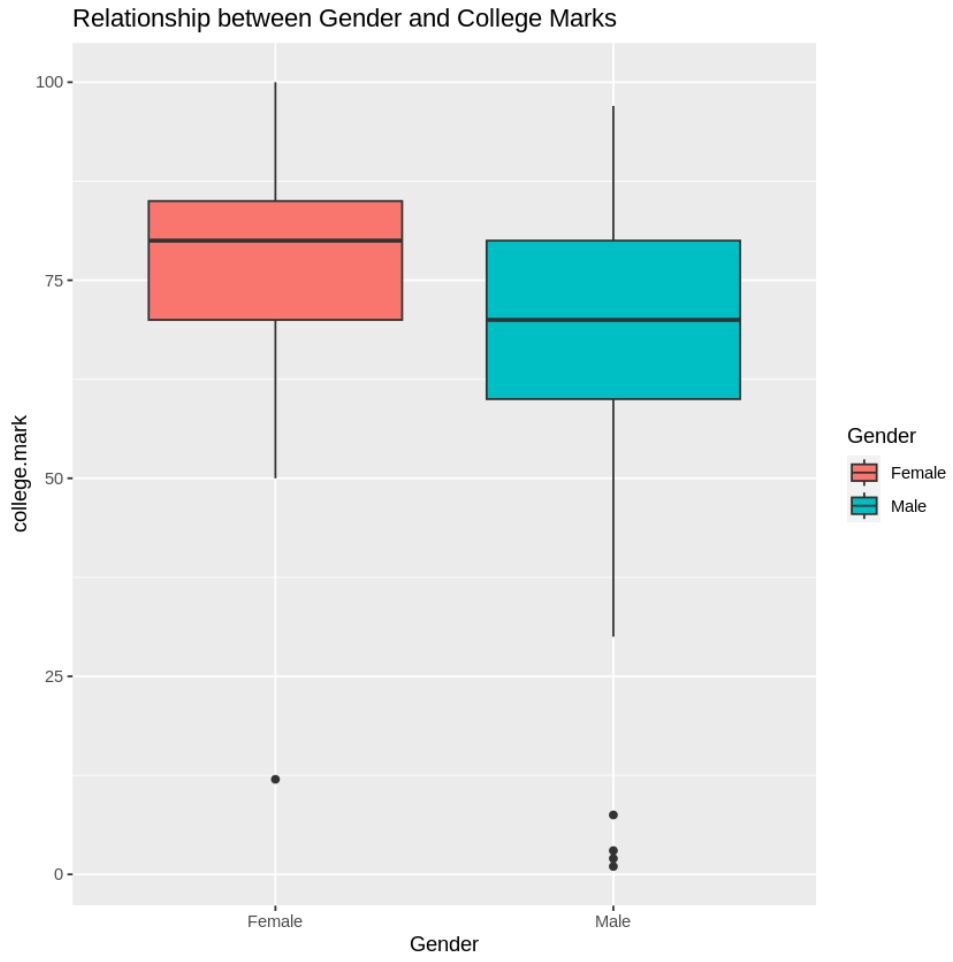


Financial status does not correlate with study activity. As shown from the bar plot, financial status' described as "bad" or "good" show the highest levels of study activity from students. The "Awful" and "Fabulous" financial status categories, the worst and best financial standings respectively, show the lowest levels of study activity.

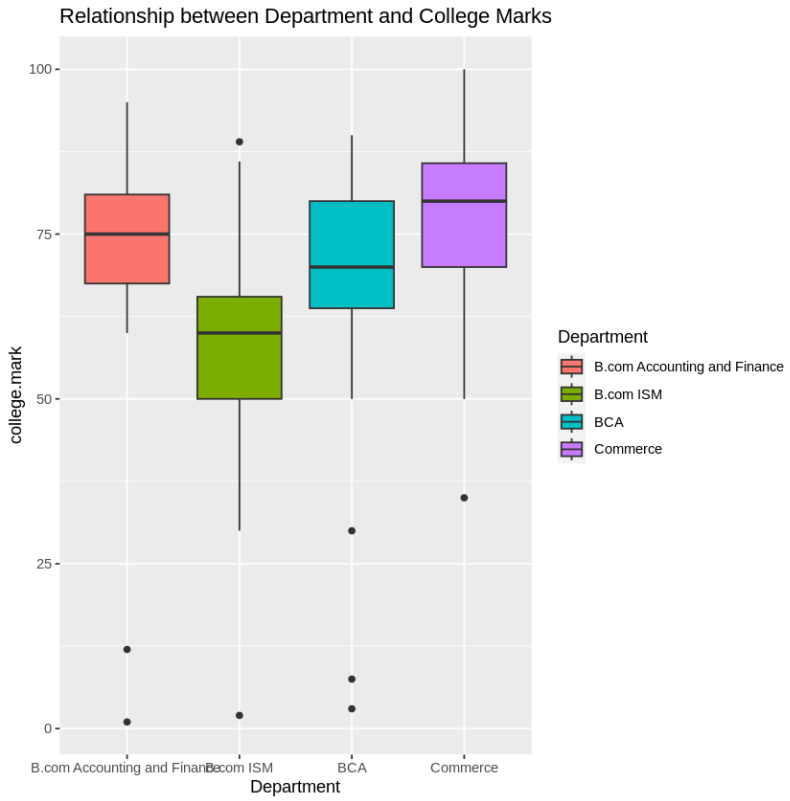
4. What is the relationship between student demographics and academic performance?

```
# Explore the relationship between gender and academic performance
ggplot(df, aes(x = Gender, y = college.mark, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Relationship between Gender and College Marks")

# Explore the relationship between department and academic performance
ggplot(df, aes(x = Department, y = college.mark, fill = Department)) +
  geom_boxplot() +
  labs(title = "Relationship between Department and College Marks")

# Explore the relationship between 12thmark and academic performance
ggplot(df, aes(x = X12th.Mark, y = college.mark)) +
  geom_point() +
```

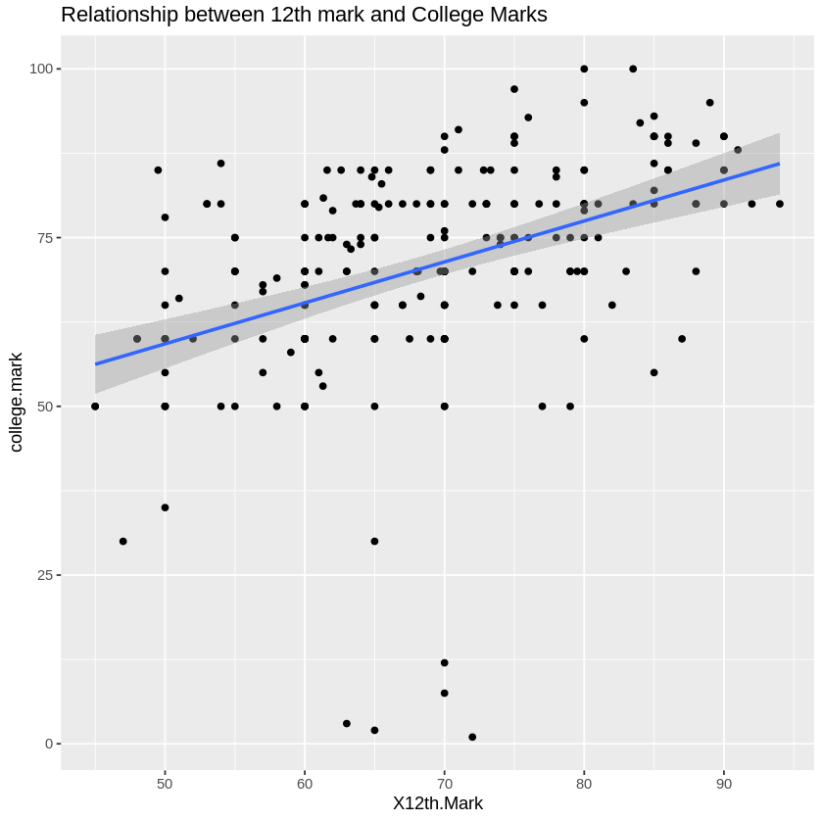
```
  geom_smooth(method = "lm") +
  labs(title = "Relationship between 12th mark and College Marks")

# Explore the relationship between financial status and academic
performance
ggplot(df, aes(x = Financial.Status, y = college.mark)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Relationship between Financial Status and College
Marks")
```

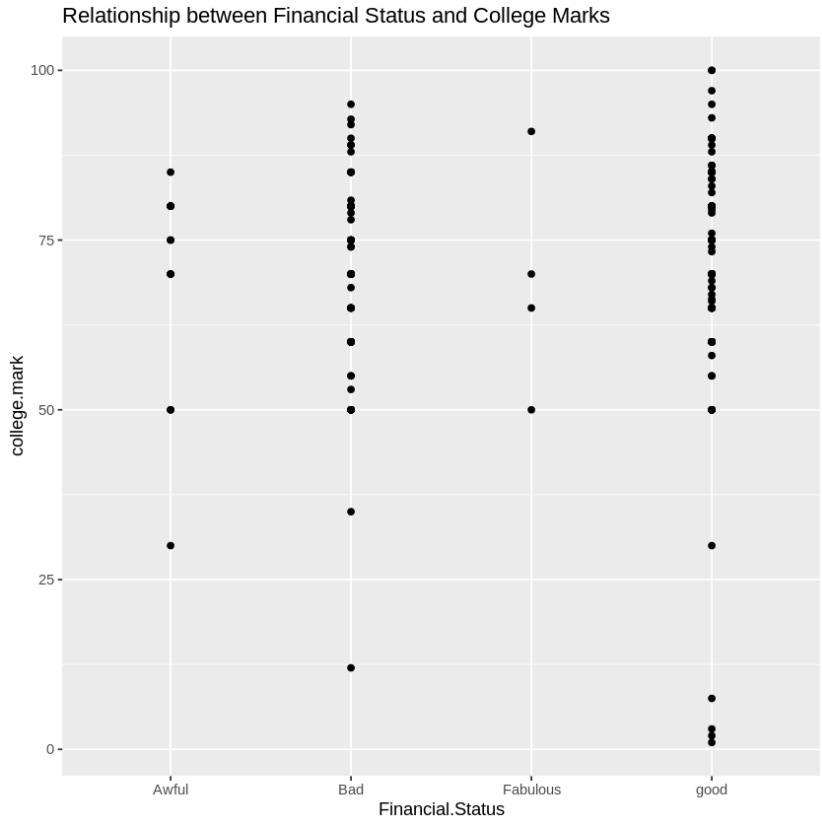


`geom_smooth()` using formula = 'y ~ x'



`geom_smooth()` using formula = 'y ~ x'





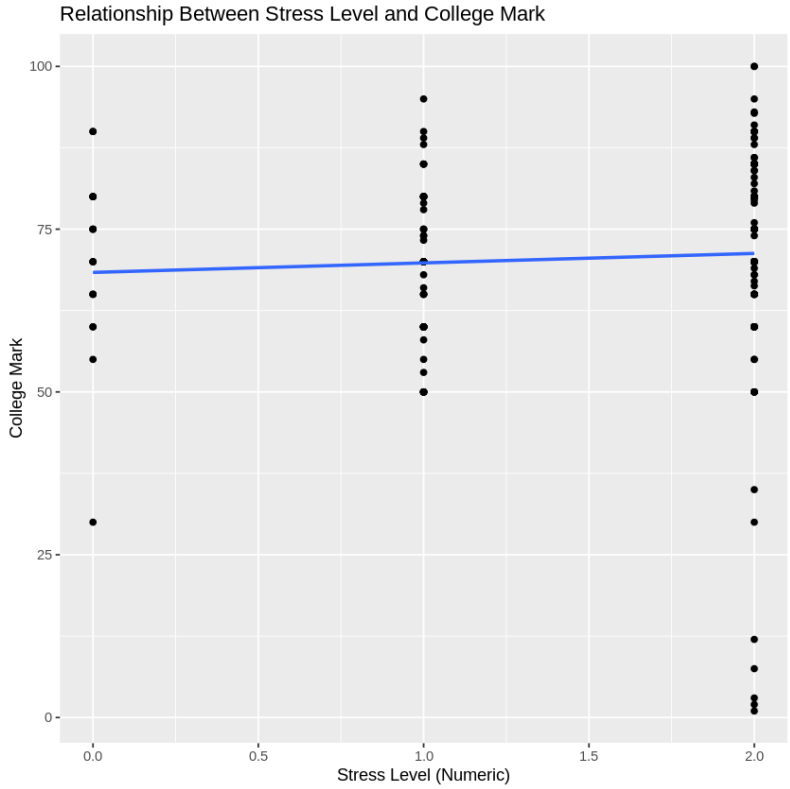
5. To what extent do college marks influence student stress levels?

```
# Converting Stress.Level to an ordered factor and then to numeric
df$Encoded.Stress.Level <- factor(df$Stress.Level,
                                  levels = c("Awful", "Bad", "Good",
"Fabulous"),
                                  labels = c(0, 1, 2, 3),
                                  ordered = TRUE)
df$Numeric.Stress.Level <-
as.numeric(as.character(df$Encoded.Stress.Level))

# Building a linear regression model
model <- lm(Numeric.Stress.Level ~ college.mark, data = df)

# Reviewing the model summary
summary(model)
```

```
Call:
lm(formula = Numeric.Stress.Level ~ college.mark, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5749 -0.5253  0.4375  0.4747  0.6456

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.351913   0.199707   6.769 1.14e-10 ***
college.mark 0.002477   0.002762   0.897    0.371
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.649 on 222 degrees of freedom
  (11 observations deleted due to missingness)
Multiple R-squared:  0.003612,  Adjusted R-squared:  -0.0008766
F-statistic: 0.8047 on 1 and 222 DF,  p-value: 0.3707
```

```r
ggplot(df, aes(x = Numeric.Stress.Level, y = college.mark)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Stress Level (Numeric)", y = "College Mark") +
  ggtitle("Relationship Between Stress Level and College Mark")
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 11 rows containing non-finite values (`stat_smooth()`).”
Warning message:
“Removed 11 rows containing missing values (`geom_point()`).”
```



Random Forest Model

```
install.packages("randomForest")
library(randomForest)

# Change 'Stress.Level' into a binary factor
df$Binary.Stress.Level <- ifelse(df$Stress.Level %in% c("Awful",
"Bad"), "High Stress", "Low Stress")

# Convert the binary outcome into a factor
df$Binary.Stress.Level <- factor(df$Binary.Stress.Level, levels =
c("Low Stress", "High Stress"))
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
# logistic regression model
model <- glm(Binary.Stress.Level ~ college.mark, data = df, family =
binomial)

# Summary of the model
summary(model)
```

```
Call:
glm(formula = Binary.Stress.Level ~ college.mark, family = binomial,
    data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.171604   0.613331   0.280    0.780
college.mark -0.009986   0.008525  -1.171    0.241

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 309.76  on 234  degrees of freedom
Residual deviance: 308.39  on 233  degrees of freedom
AIC: 312.39

Number of Fisher Scoring iterations: 4
```

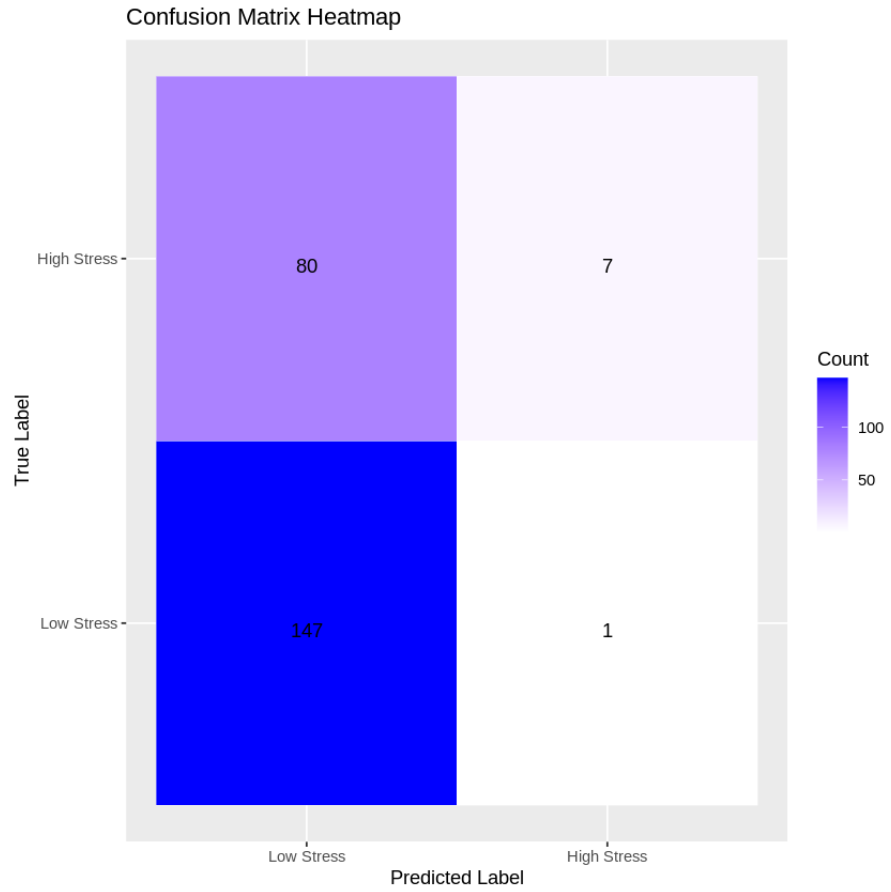
```
# Random Forest model
set.seed(123)
rf_model <- randomForest(Binary.Stress.Level ~ college.mark, data = df,
ntree = 500)

# Print the model summary
print(rf_model)
```

```
Call:
 randomForest(formula = Binary.Stress.Level ~ college.mark, data = df,      ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 48.94%
Confusion matrix:
            Low Stress High Stress class.error
Low Stress         118          30   0.2027027
High Stress         85           2   0.9770115
```

```r
# Predict on the training data
rf_predictions <- predict(rf_model, df)

# Confusion Matrix to see the accuracy
confusionMatrix <- table(df$Binary.Stress.Level, rf_predictions)
print(confusionMatrix)
```

```
             rf_predictions
              Low Stress High Stress
  Low Stress         147           1
  High Stress         80           7
```

```r
# Convert the confusion matrix to a data frame for plotting
confusionMatrix_df <- as.data.frame(as.table(confusionMatrix))

# Rename columns to match the variable names
names(confusionMatrix_df) <- c("TrueLabel", "PredictedLabel", "Freq")


# Create Heat Map
ggplot(confusionMatrix_df, aes(x = PredictedLabel, y = TrueLabel, fill
= Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "white", high = "blue") +
  geom_text(aes(label = sprintf("%0.0f", Freq)), vjust = 1) +
  labs(x = "Predicted Label", y = "True Label", fill = "Count") +
  ggtitle("Confusion Matrix Heatmap")
```



Our study reveals little correlation between college grades and stress levels among students. Our analysis included regression and Random Forest models. We found that college grades only account for approximately 0.36% of the variation in stress levels, indicating a weak connection. Random Forest model showing a 47.23% out-of-bag error rate. This suggests that grades are not a strong predictor of stress levels in students.

For predictive analysis, our plan involves leveraging advanced statistical models to forecast and comprehend the intricate relationships among educational factors such as study habits, financial status, and academic performance.

Here are the key steps in our predictive analysis plan:

Feature Selection: Identify key variables that significantly contribute to academic performance, stress levels, and other relevant outcomes. This will involve assessing the importance of each variable based on statistical methods.

Data Preprocessing: Clean and preprocess the data to address missing values and outliers, ensuring the reliability of the dataset.

Model Selection: Choose appropriate predictive modeling techniques, such as regression analysis, to consider the multifaceted nature of the academic experience.

Cross-Validation: Implement cross-validation techniques to identify underlying patterns.

Evaluation Metrics: Define suitable evaluation metrics to gauge the performance of the models.

Interpretation: Interpret the results of the predictive models and validate their accuracy and reliability, considering the various factors identified in the conclusion.

After observing how factors such as study habits, financial status, and academic performance influence one another, our findings highlight that daily studying time alone does not significantly impact academic success. Instead, a comprehensive mix of diverse factors, such as gender, department, 12th-grade marks, and financial status, all contribute to the typical but nuanced academic experience.

Although we observed some positive correlations between daily studying time and academic performance, the relationship between financial status and study activity contradicts that this was achieved alone. Furthermore, despite the lack of correlation between study habits and academic grades, students from lower financial backgrounds displayed higher median marks. Additionally, our investigation into the link between college grades and stress levels suggested a weak connection. College grades accounted for a mere 0.36% of the variation in stress levels, with a Random Forest model indicating a 47.23% out-of-bag error rate, highlighting the limited predictive power of grades in anticipating stress levels among students.

Conclusively, our research suggests that the conventional perspectives on the determinants of academic performance may not offer a complete understanding. There are many factors tied to academic experience that might not be considered wholly when determining educational methods and standards. A continued analysis might entail a comparison of educational methods between schools and whether or not they consider these factors.

References

Nunes, Catarina, et al. “Determinants of Academic Achievement: How Parents and Teachers Influence High School Students’ Performance.” *Heliyon*, U.S. National Library of Medicine, 3 Feb. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC9950838/.

M, Guna. “Student Behavior.” *Kaggle*, 24 June 2023, www.kaggle.com/datasets/gunapro/student-behavior.