STA 141A Final Project
Group 40

Understanding Factors Influencing College Student Performance: Insights and Recommendations for Student Success Initiatives

Emma Lam: emklam@ucdavis.edu
Aidan Jun asjun@ucdavis.edu
Lola Siu (leader) lsiu@ucdavis.edu
Merry Sutijono msutijono@ucdavis.edu

1. Introduction

In today's rapidly changing educational landscape, there's a growing recognition of the various factors that influence student success in college. While traditional metrics like grades and test scores remain important, educators and policymakers increasingly understand that a student's academic performance is shaped by a complex interplay of various factors. From personal characteristics and study habits to socio-economic background and extracurricular activities, numerous factors contribute to a student's performance in college. Understanding these factors and their relative importance can provide valuable insights for educators to enhance student outcomes and promote holistic development.

To delve deeper into this issue, we turn our attention to the "Student Attitude and Behavior" dataset, which is sourced from a large cohort of university students in India. This dataset offers a rich repository of information on academic performance, personal factors such as motivation and stress levels, study habits, extracurricular involvement, and more. Through rigorous analysis of this dataset using statistical techniques such as correlation analysis, logistic regression, multiple linear regression, and cluster analysis, we aim to uncover the key factors influencing student marks in college. By identifying significant predictors and discerning underlying patterns, we seek to provide insights that can inform educators on how they can involve and support students in their academic success.

2. Key Questions

- Question 1: How do prior academic performance and study habits correlate with student's marks in college?
- Question 2: How do personal factors such as motivation and stress levels impact a student's marks?
- Question 3: How does the level of engagement with hobbies and social media platforms impact a student's academic performance?

3. Data and Methodology

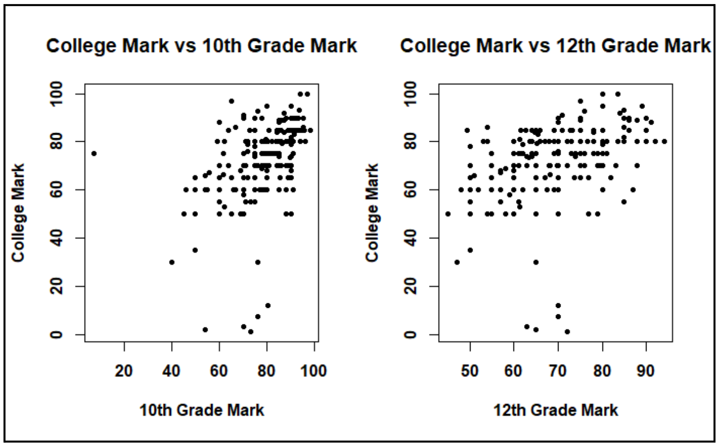
Our data was obtained from https://www.kaggle.com/datasets/susanta21/Student-attitude-and-behavior. This dataset covers a range of 19 factors, including certification courses, gender, department, height, weight, academic performance in previous high school grades and college, hobbies, study habits, career aspirations, social media/video usage, travel time, stress levels, and financial status. This data was gathered via a Google form from 236 university students in India. It's important to note that the data's limitations include its specific focus on Indian students, potentially limiting its generalizability to other college student populations. Moreover, the reliance on self-reported Google Form answers could also introduce the possibility of bias in the data. Despite these drawbacks, the dataset still provides valuable insights that can offer strategies to support student success, both within and beyond the context of Indian students.

Our methodology for addressing the research questions concerning student marks in college revolves around a comprehensive analysis of the "Student Attitude and Behavior" dataset. We deploy a series of statistical techniques to investigate the relationships between various factors and academic performance, providing valuable insights for educators

2.1 Correlation between prior academic performance and college marks (Lola)

Introduction:

In this study, we investigated the relationship between students' prior academic performance in 10th and 12th grade and their subsequent college marks. The study aimed to explore how a strong academic foundation can impact students' success in college by equipping them with essential skills, knowledge, and habits for achieving their educational goals and overcoming challenges.



According to the scatter plots, there appears to be a positive linear relationship between the marks that the students obtained from their 10th Grade and 12th Grade, and the marks that the students obtained in College. This indicates that students who perform well academically during their 10th Grade and 12th Grade tend to continue their academic success in their college years. However, there seem to be a few outliers, the most obvious of which is from the “10th Grade Mark” scatter plot, the point where the 10th-grade mark of a student is ~5 when their college mark is ~75.

Method:

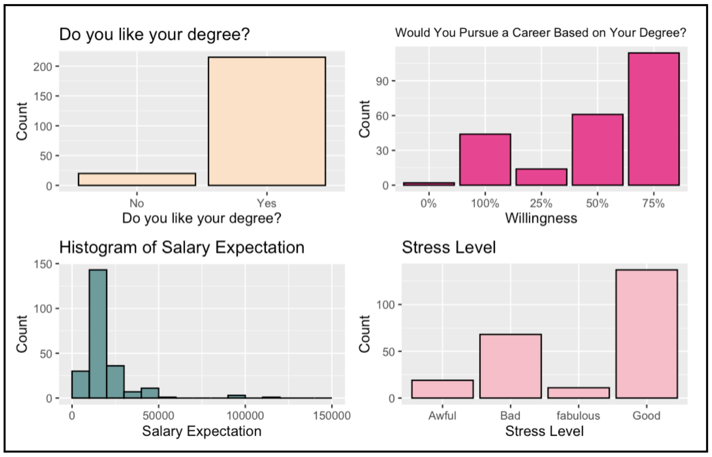
To determine the significance of the two variables, we fit linear regression models for the two separately. According to the analysis, the coefficient for Mark_10 is 0.56155, which indicates that for each additional mark obtained in 10th Grade, students’ college mark is estimated to increase by about 0.56155. Having a p-value of 4.62e-14, we can conclude that students’ marks obtained in 10th Grade are a significant predictor of college marks at the significance level α = 0.05. The coefficient of Mark_12 is 0.60640, meaning that for every additional mark obtained in 12th Grade, students’ college mark is estimated to increase by about 0.60640. The p-value for Mark_12 is 1.03e-11, we can confirm the importance of the students’ marks obtained in 12th Grade in the model.

Findings:

Evidence from regression models shows that students' academic performance in 10th and 12th grade significantly predicts their college marks. The positive linear relationships between students’ prior academic performance also indicate that performance in earlier stages of education is a reliable indicator of future academic success in college.

2.2 Correlation Between Personal Factors and College Marks

To conduct summary statistics, we used bar graphs to illustrate the relationships between personal factors such as: opinion on degree, opinion on future career based on degree, salary expectation, and stress level.



2.2.1 College Marks vs Salary Expectation (Aidan)

Introduction:

In today's competitive job market, students often correlate their academic performance with their future salary expectations. Understanding the relationship between college marks and salary expectations can provide insights into a student's career aspirations and the perceived value of academic success. This section investigates how college marks correlate with student's salary expectations post-graduation. In order to do that, we computed the average salary expectations and grouped them according to different ranges of college marks (Table 2.3.1).

Table 2.1

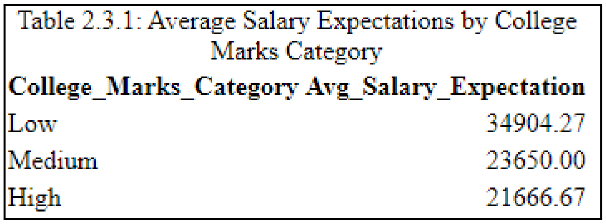
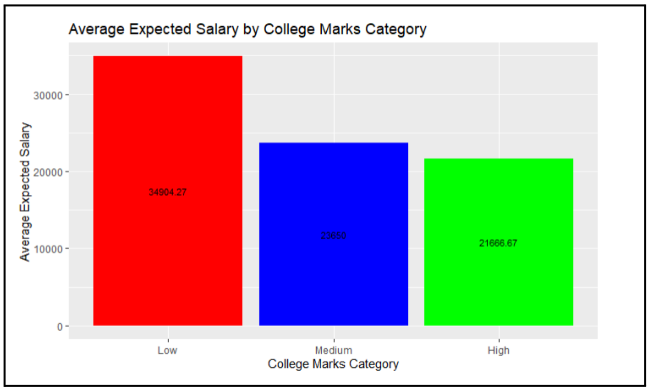


Table 2.3.1: Average Salary Expectations by College Marks Category

College_Marks_Category	Avg_Salary_Expectation
Low	34904.27
Medium	23650.00
High	21666.67

Plot 2.1



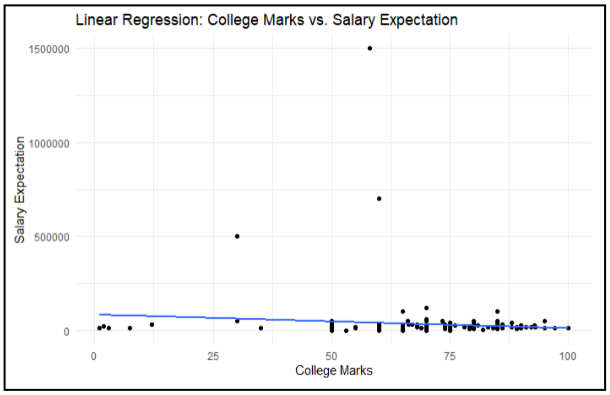
The table displays the average salary expectations categorized by different ranges of college marks (e.g., low, medium, and high). Low marks are marks that are lower than or equal to 80. Medium marks are marks that are greater than 80 but less than or equal to 90. High marks are marks that are greater than 90.

Plot 2.3.1 shows the distribution of average salary expectations across different ranges of college marks. This helps us understand whether students with higher academic performance anticipate higher salaries after graduation. Recognizing this relationship can inform both educational strategies and career counseling services, guiding students to align their academic efforts with their career goals. We can see from the plot above that college students with lower academic performance expect the highest salaries after graduation ($34,904.27), while those with higher academic performance expect the lowest salaries after graduation ($21,666.67).

Method:

To analyze the correlation between college marks and salary expectations, we performed a linear regression analysis. The predictor was college marks, and the dependent variable was salary expectations. Linear regression allows us to assess the strength and direction of the relationship between academic performance and salary expectations.

Linear Regression Plot



Findings:

The linear regression analysis results indicate that the relationship between college marks and salary expectations is not statistically significant at the conventional significance level of 0.05. The coefficient estimate for college marks is -729.2, indicating that, on average, for each unit increase in college marks, the salary expectation decreases by $729.2. However, this coefficient is not statistically different from zero (p = 0.1152). The adjusted R-squared value of 0.00637 suggests that only about 0.64% of the variation in salary expectations can be explained by variations in college marks. The scatter plot with the regression line is displayed, showing the trend in the data. However, there is insufficient evidence to conclude a significant relationship between college marks and salary expectations.

2.2.2 College Marks vs Willingness to pursue a career based on their degree (Aidan)

Introduction:

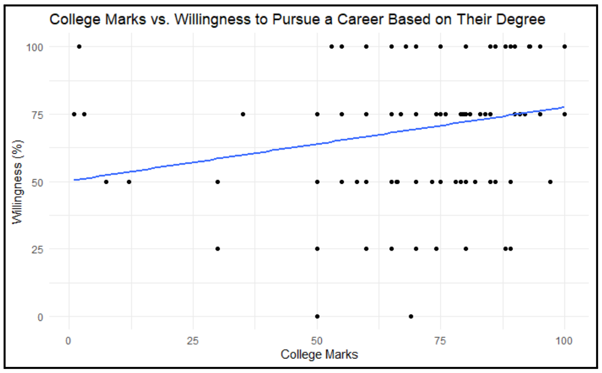
Understanding the factors that influence a student's decision to pursue a career based on their degree is crucial for educators and policymakers. This insight can help tailor educational programs to better align with student's career aspirations and academic performance. In this analysis, we aim to explore the relationship between college marks and the willingness to pursue a career related to their degree. Specifically, we investigate whether higher academic performance is associated with a stronger inclination to pursue a career in their field of study.

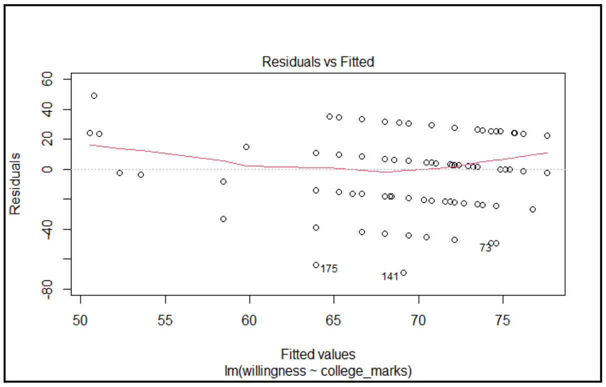
Method:

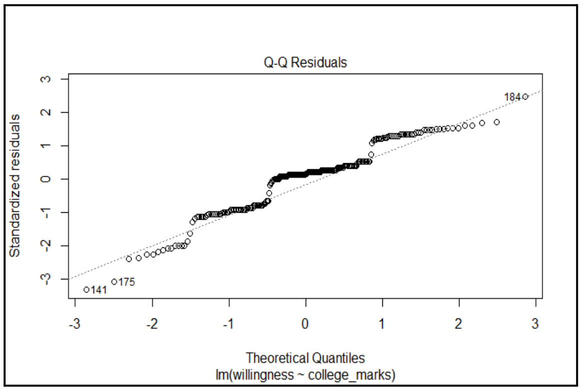
To analyze the relationship between college marks and the willingness to pursue a career based on their degree (measured as a percentage), we used multiple linear regression. This method allows us to model the dependent variable (willingness to pursue a career based on their degree) as a function of the independent variable (college marks). First we fit a multiple linear regression model and then we visualized the results using a scatter plot with a regression line.

Multiple Linear Regression Model Formula:
Willingness = $\beta_0 + \beta_1$ x College Marks + Ɛ

Scatter plot with Regression Line







Findings:

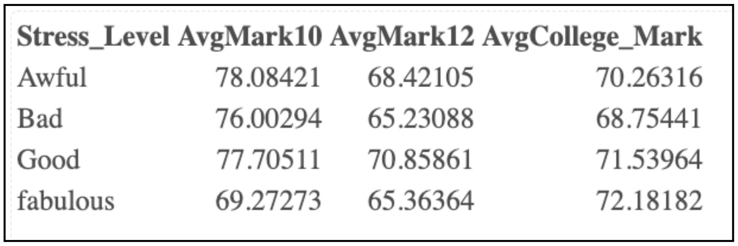
The multiple linear regression analysis provided the following results. The intercept represents the predicted willingness percentage when college marks are zero. In this case, the intercept is 50.24345. The positive coefficient for college marks (0.27358) indicates that higher college marks are associated with a higher willingness to pursue a career based on their degree. The R-squared value (0.04127) represents the proportion of variance explained by the model. In this case, the model indicates that approximately 4.13% of the variability in willingness to pursue a career based on their degree can be explained by college marks. This relatively low R-squared value suggests that college marks are not the sole predictor of willingness, and other factors likely play a significant role. The scatterplot shows a positive trend, indicating that as college marks increase, willingness to pursue a career based on their degree tends to increase as well. As we can see from the residuals vs fitted plot, there appears to be a curve pattern, suggesting heteroscedasticity or non-linearity. Also, the normal Q-Q plot shows that the points tend to deviate from the reference line, suggesting that the residuals are not normally distributed. Given the patterns in the residual plots and Q-Q plot, it's clear that the model is not fully meeting the assumptions of linear regression and should go through a process of transformations and other regression techniques in order to work towards a more appropriate model that better fits the data.

2.2.3 College Marks v.s. Stress Level (Emma)

Introduction:

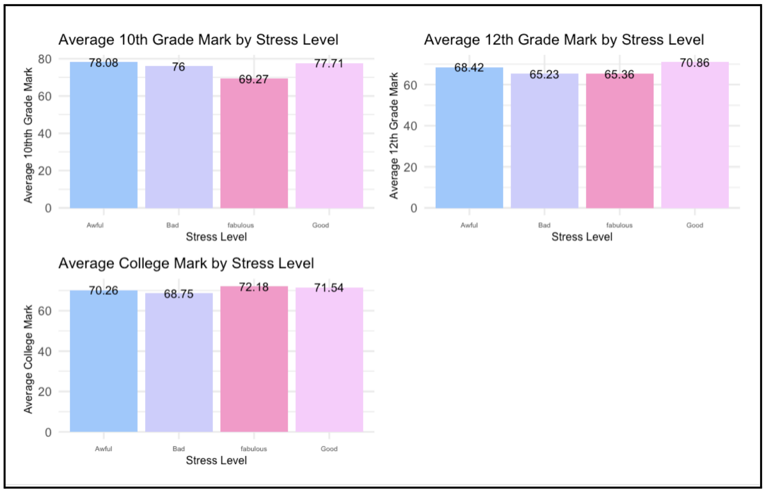
In academic settings, students often face significant levels of stress associated with deadlines, academic performance expectations, pressure, and more. We want to explore the possibilities of correlation between higher marks and higher stress levels. For some exploratory data analysis, we first computed the averages of marks achieved in different educational years and grouped them according to the 4 reported stress levels: "bad", "awful", "good", and "fabulous" (Table 2.1)

Table 2.2



Stress_Level	AvgMark10	AvgMark12	AvgCollege_Mark
Awful	78.08421	68.42105	70.26316
Bad	76.00294	65.23088	68.75441
Good	77.70511	70.85861	71.53964
fabulous	69.27273	65.36364	72.18182

Plot 2.2

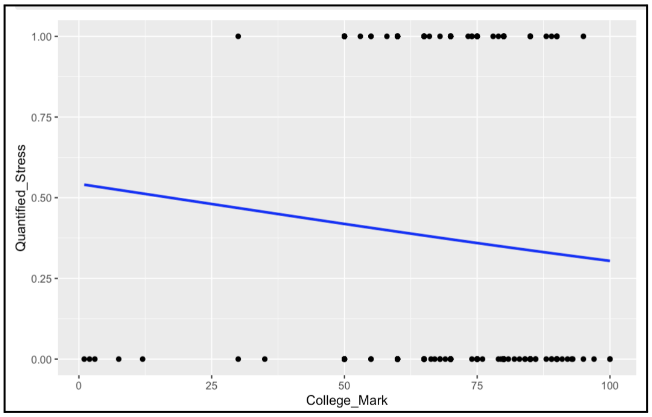


To visualize this data, Plot 2.1 shows 3 plots: one for average 10th grade mark, one for average 12th grade mark, and a third for average college mark, each categorized by the four stress levels. Having explored the relationship between stress levels and academic performance across different academic years, we now pivot our investigation to focus specifically on college marks. This allows us to delve deeper into understanding whether higher college marks correlate with experiencing stress. Recognizing the relationship between academic performance and student stress can be impactful in determining ways to foster a healthy balance between academic rigor and student mental health and well-being.

Method:

To quantify stress levels, we converted the variable stress levels into a binary variable. The dataset contains four possible levels for stress levels, including: “bad”, “awful”, “good”, and “fabulous”. We categorized “good” and “fabulous” into 0, indicating no stress, and “awful” and “bad” as 1, indicating stress. Subsequently, we utilized logistic regression, a supervised classification technique, to examine how college marks relate to quantified stress levels. The logistic regression model allied us to assess the probability of experiencing stress based on college marks.

Plot 2.3



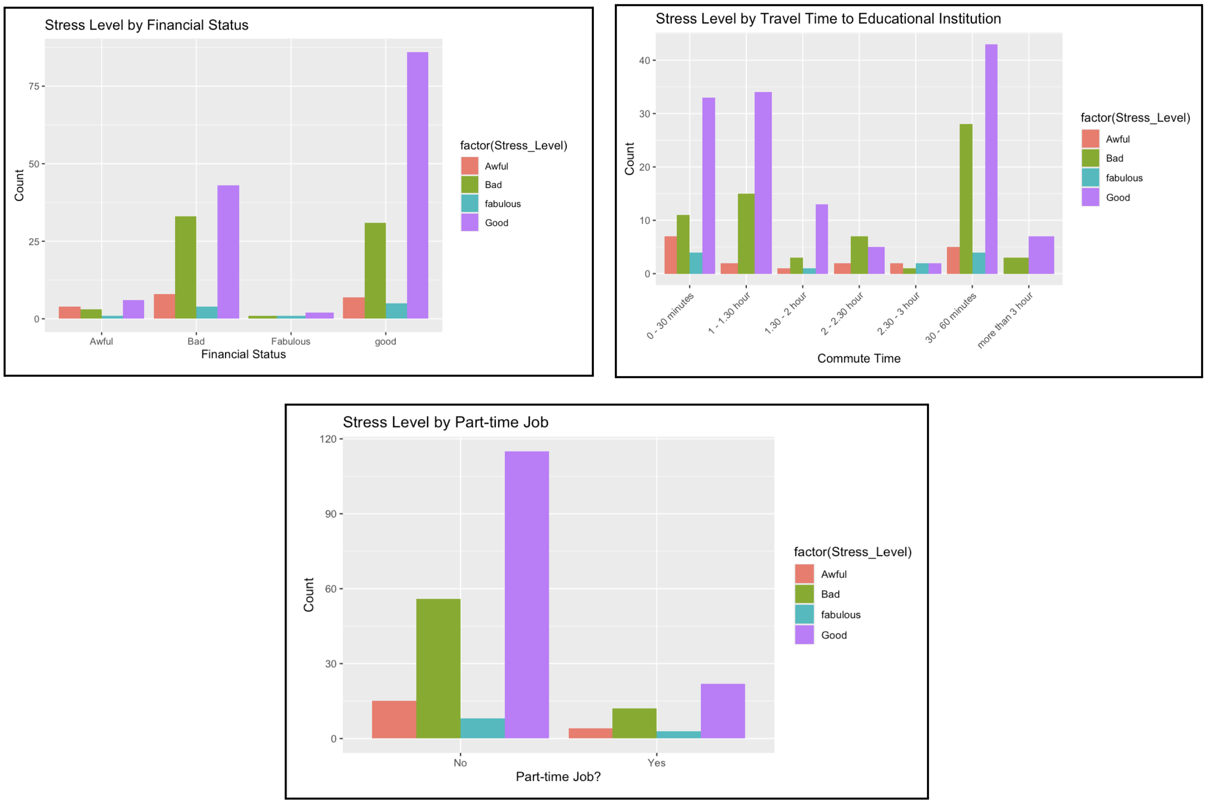
Findings:

Based on the model, the estimated coefficient for college marks is -0.009986, and the p-value for the coefficient is 0.241. Because this value is greater than the significance level of α = 0.05, we fail to reject the null hypothesis. This means that there is no statistically significant evidence of a relationship between college marks and experiencing stress. It is possible that other factors not included in the model may better explain variations in stress levels among individuals.

2.2.4 Stress Levels v.s. Personal Factors (Emma)

Introduction: Acknowledging the presence of external personal factors, such as financial pressures, responsibilities, and travel time to educational institutions, is crucial when assessing the impact on academic stress levels. These factors can significantly influence a student's well-being and academic performance. We first visualize these relationships between personal factors and stress levels through bar graphs.

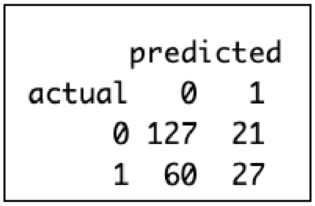
Plot 2.4



Method:

Logistic regression, a supervised classification technique, was conducted to examine the relationships between three personal factors (job, financial status, travel time) and experiencing stress. The logistic regression model predicted the probability of experiencing stress based on the personal factors. A confusion matrix was also calculated to evaluate the accuracy of the model in predicting stress.

Figure 2.2



```
       predicted
actual   0   1
     0 127  21
     1  60  27
```

Findings:

The logistic regression analysis revealed that the only significant predictor of stress was a travel time between 2-2.30 hours. Students with a high travel time in this range demonstrated a statistically significant higher likelihood of experiencing stress, highlighting that travel time may contribute to challenges in managing academic responsibilities. In the scope of this study, other factors like financial status and job were not significant. Additionally, the calculated confusion matrix provided insights into the predictive abilities and errors made in the model. The accuracy rate was around 65.53% indicates a moderate accuracy level for correctly classified stress. However, we must acknowledge the inherent limitations of this study which leverages self-reported data, increasing the possibilities of bias and inaccuracies. Further investigation with larger scale and more diverse datasets may be necessary to accurately explore and validate these questions of student stress predictors.

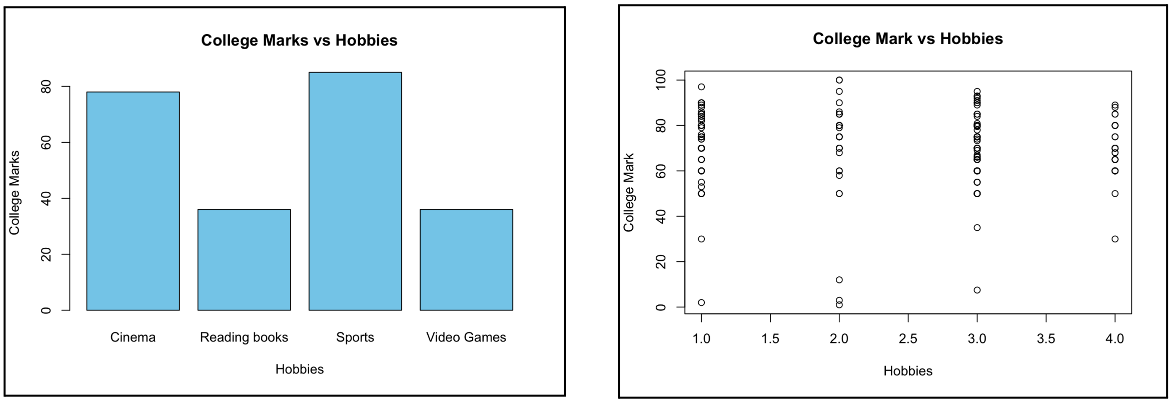
2.3 Correlation between student's engagement with society and college marks

2.3.1 College Marks vs Hobbies (Merry)

Introduction:

It should be noted that students will have time to prioritize their interests outside of academics such as being involved in their hobbies. The dataset we are working on includes hobbies that students usually choose to do outside of school such as playing video games, going to the cinemas, reading books, and playing sports. We will visualize the two variables college marks and hobbies and see if they correlate with each other using a bar graph and a scatter plot as provided below.

Plot 2.3.1 (Bar Plot (left) and Scatter Plot (right))



Method:

In order for us to see how the college marks of students correlate with their hobbies, we would proceed and carry on with the logistic regression method. We obtained a summary based on our dataset which was based on our two variables: college marks and hobbies. Residuals on the summary reveals that the minimum is -70.910, median is 1.306, and maximum is 28.972. The p-value is 0.4194 .

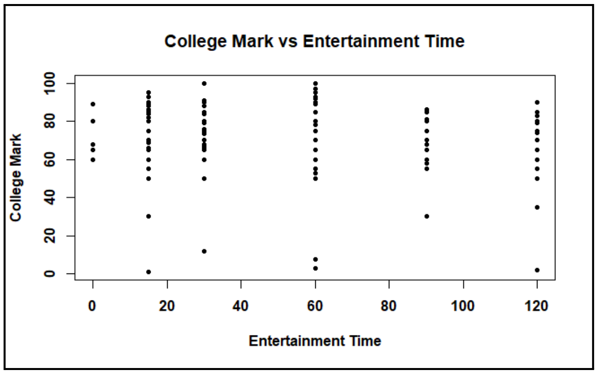
Findings:

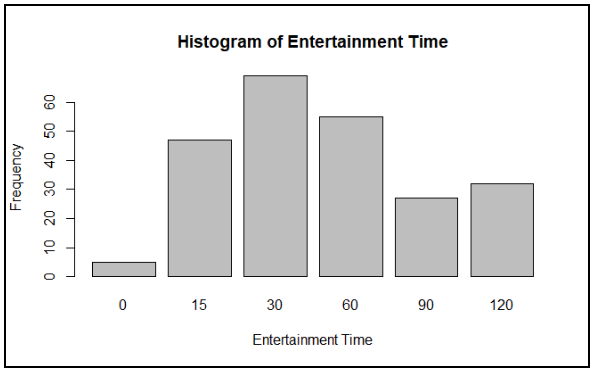
After analyzing the correlation between the college marks and the hobbies these students have spent their time with we can conclude that as the p-value is 0.4194 which is greater than α = 0.05, we will fail to reject the null hypothesis. Hence, there is no statistically significant evidence of a relationship between college marks and partaking in hobbies.

2.3.2 College Marks vs Engagement with Social Media (Lola)

Introduction:

In contrast to other variables examined in this study, time spent on social media and video platforms is often cited as a negative factor affecting academic performance. Increased engagement with entertainment is commonly associated with poorer academic outcomes. The duration of students' engagement with social media and video platforms was categorized as follows: '0 minutes' = '0', '1-30 minutes' = '15', '30-60 minutes' = '30', '1-1.5 hours' = '60', '1.5-2 hours' = '90', and 'more than 2 hours' = '120'. The corresponding scatter plot and histograms are shown below:





The scatter plot shows a possible relationship between students' marks obtained at college with the amount of time the students spend on entertainment platforms. There does not appear to be a significant linear correlation between college marks and entertainment. Within the data presented, we can identify that the majority of points cluster around study times of '1 - 30 Minute', '30 - 60 minute', and '1 - 1.30 hour'', with college marks generally in the range of 50-100. Outliers are also observed in the plot.

The histogram showcases the distribution of the daily studying time among the participated college students and appears to be slightly right-skewed. The mode is at '30 - 60 minute', with a frequency of about 70. The most common study time periods are '30 - 60 minutes' and '1 - 1.30 hours'.

Method:

We fit the entertainment time in the linear regression model and the results of the model fitting process show that the coefficient of Social is 0.005204, meaning that for each additional minute spent on entertainment, students' college mark is estimated to increase by about 0.005204. As the p-value for Social is 0.856, we can conclude that the preferred study time is not statistically significant at $\alpha = 0.05$. The adjusted R-squared value for Social is -0.00415, it indicates that not only that the entertainment time spent is ineffective in explaining the variance in college marks but also that the model will perform worse than a model without this variable.

Findings:

The analysis of the relationship between students' engagement on social media and video platforms and their college marks reveals that these factors are not significant predictors of academic performance in college. Furthermore, the low and negative adjusted R-squared value for the entertainment time variable indicates that it explains less than 1% of the variability in college marks in this model. This suggests that the inclusion of this variable actually diminishes the model's predictive performance compared to a model without predictors. Therefore, despite the common perception that time spent on entertainment negatively impacts academic performance, it does not substantially account for the variations in college grades among the sampled students.

Conclusion (Merry)

Our final project focuses on three research questions which focuses on the correlation of some of the variables based on the "Student Attitude and Behavior" dataset which have resulted in several key findings.

Firstly, we have compared the college marks of students with their academic performance and see how these two variables correlate with each other. Based on our findings, it can be indicated that there is a positive linear relationship, which signifies that the students' study habits or hours that they have spent on their college courses indeed affect how each student performs in college. Moreover, the findings based on the college marks and preferred time of study indicates that study habits do not substantially explain the differences in college grades among the sampled students.

Secondly, the first part of the second research question focuses on the correlation between college marks and salary expectation. Our findings on the relationship between these two variables suggests that there is insufficient evidence that the college marks will affect the salary of the students once they enter the workforce. The second part revealed patterns in the residual plots and Q-Q plot have revealed that the model is not fully meeting the assumptions of linear regression and should go through transformations and other regression techniques in order to work towards a more appropriate model that better fits the data. Our other findings between the relationship of the higher college marks and experiencing stress suggests that there is no statistically significant evidence of a relationship between higher college marks and experiencing stress. There might be a possibility in which other factors not included in the model may better explain variations in stress levels among individuals. Hence, the correlation between stress levels and personal factors have revealed that further investigation may be necessary in order for us to provide an answer to our second research question.

Thirdly, our findings based on the third research question suggests that the correlation between college marks and hobbies are not related because there is no statistically significant evidence of a relationship between college marks and partaking in hobbies. In addition to that, the relationship between students' engagement on social media and video platforms and their college marks are not significant predictors of academic performance in college.

Appendix

2.1 Correlation between prior academic performance and college marks

```
library(ggplot2)
library(dplyr)
library(gridExtra)
filename <- "Student Attitude and Behavior.csv"
file_path <- file.choose()
if (basename(file_path) == filename) {
  data <- read.csv(file_path, header = TRUE)
} else {
  cat("Selected file does not match expected filename.\n")
}
head(data)
# Data Transform
names(data) <- c('Certification_Course','Gender','Department','Height','Weight',
                 'Mark_10','Mark_12','College_Mark', 'Hobbies', 'Daily_Study_Time',
                 'Study_Spot', 'Salary_Expectation', 'Degree_Opinion', 'Opinion_Future_Career',
                 'Social_Media_Engagement', 'Travel_Time', 'Stress_Level', 'Financial_Status',
                 'Job')
# Scatter Plots
par(mfrow = c(1, 2))
plot(data$Mark_10, data$College_Mark,
     main = "College Mark vs 10th Grade Mark", ylab = "College Mark",
     xlab = "10th Grade Mark", pch = 19, font = 1, font.lab = 2,
     cex = 0.75)
plot(data$Mark_12, data$College_Mark,
     main = "College Mark vs 12th Grade Mark", ylab = "College Mark",
     xlab = "12th Grade Mark", pch = 19, font = 1, font.lab = 2,
     cex = 0.75)
par(mfrow = c(1, 1))
# Regression Models
the.model_1.1 <- lm(College_Mark ~ Mark_10, data = data)
the.model_1.2 <- lm(College_Mark ~ Mark_12, data = data)
summary(the.model_1.1)
summary(the.model_1.2)
```

2.2 Correlation between personal factors and college marks

```
## exploratory data visualization
par(mfrow = c(2, 2))
## Do you like your degree?
plot1 = ggplot(data, aes(x = factor(Degree_Opinion))) +
  geom_bar(fill = "bisque", color = "black") +
  labs(title = "Do you like your degree?",
       x = "Do you like your degree?",
       y = "Count")
##  Willingness to pursue a career based on their degree
plot2 = ggplot(data, aes(x = factor(Opinion_Future_Career))) +
  geom_bar(fill = "deeppink", color = "black") +
  labs(title = "Would You Pursue a Career Based on Your Degree?",
       x = "Willingness",
       y = "Count")+
  theme(plot.title = element_text(size = 9.5))
## Salary Expectations
breaks <- seq(0, 150000, by = 10000)  # Adjust the range and width of bins as needed
# Create the histogram
plot3= ggplot(data, aes(x = Salary_Expectation)) +
  geom_histogram(bins = length(breaks) - 1, fill = "cadetblue", color = "black", breaks = breaks) +
  labs(title = "Histogram of Salary Expectation",
       x = "Salary Expectation",
       y = "Count")
## Stress Level
plot4 = ggplot(data, aes(x = factor(Stress_Level))) +
  geom_bar(fill = "pink", color = "black") +
  labs(title = "Stress Level",
       x = "Stress Level",
       y = "Count")

#arrange plots
grid.arrange(plot1, plot2, plot3, plot4, nrow = 2, ncol = 2)
```

2.2.1 College Marks vs Salary Expectation

Clean Data:

```
library(dplyr)
library(car)

#Check for missing values and omit NAs
missing_values <- colSums(is.na(data))
data_no_NA <- na.omit(data)

#Check to make sure there are indeed no missing values
missing_values_check <- colSums(is.na(data_no_NA))
missing_values_check

colnames(data) <- c('Certification_Course', 'Gender', 'Department', 'Height', 'Weight', 'Mark_10', 'Mark_12', 'College_Mark',
'Hobbies', 'Daily_Study_Time', 'Study_Spot', 'Salary_Expectation', 'Degree_Opinion', 'Opinion_Future_Career',
'Social_Media_Engagement', 'Travel_Time', 'Stress_Level', 'Financial_Status', 'Job')
```

Table 2.1

```
College_Mark <- data$College_Mark
Salary_Expectation <- data$Salary_Expectation
data1 <- data.frame(
  college_marks = College_Mark,
  salary_expectation = Salary_Expectation
)

data1 <- data1 %>%
  mutate(College_Marks_Category = cut(college_marks,
                            breaks = c(-Inf, 80, 90, Inf),
                    labels = c("Low", "Medium", "High")))

Average_Salary <- data1 %>%
  group_by(College_Marks_Category) %>%
  summarise(Avg_Salary_Expectation = mean(salary_expectation))

print(Average_Salary)

library(knitr)
kable(Average_Salary, caption = "Table 2.3.1: Average Salary Expectations by College Marks Category")
```

Plot 2.1

```
library(ggplot2)

ggplot(data = Average_Salary, aes(x = College_Marks_Category, y = Avg_Salary_Expectation, fill =
College_Marks_Category)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(Avg_Salary_Expectation, 2)), position = position_stack(vjust = 0.5), color = "black", size =
3) +
  labs(title = "Average Expected Salary by College Marks Category",
       x = "College Marks Category",
       y = "Average Expected Salary") +
  scale_fill_manual(values = c("Low" = "red", "Medium" = "blue", "High" = "green")) +
  theme(legend.position = "none")
```

Linear Regression Plot

```
lm_model <- lm(salary_expectation ~ college_marks, data = data1)

summary(lm_model)

ggplot(data1, aes(x = college_marks, y = salary_expectation)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "College Marks", y = "Salary Expectation") +
  ggtitle("Linear Regression: College Marks vs. Salary Expectation") +
  theme_minimal()
```

2.2.2 College Marks vs Willingness to pursue a career based on their degree

```
## Scatterplot and Residual/Q-Q Plot
College_Mark <- data$College_Mark
Willingness <- data$Opinion_Future_Career
willingness_numeric <- as.numeric(gsub("%", "", Willingness))
data2 <- data.frame(
  college_marks = College_Mark,
  willingness = willingness_numeric
)
library(ggplot2)
library(dplyr)

model <- lm(willingness ~ college_marks, data = data2)
```

```
summary(model)

ggplot(data2, aes(x = college_marks, y = willingness)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "College Marks vs. Willingness to Pursue a Career Based on Their Degree",
    x = "College Marks",
    y = "Willingness (%)"
  ) +
  theme_minimal()

plot(model)
qqnorm(residuals(model))
qqline(residuals(model))
```

2.2.3 College Marks v.s. Stress Level

Table 2.2

```
library(knitr)
mark_by_stress = data %>%
                  group_by(Stress_Level) %>%
                  summarize(AvgMark10 = mean(Mark_10),
                           AvgMark12 = mean(Mark_12),
                           AvgCollege_Mark = mean(College_Mark))
kable(mark_by_stress, format = "markdown")
mark_by_stress
```

Plot 2.2

```
##VISUALIZATION
# average 10th grade mark by stress level plot
plot_10 = ggplot(mark_by_stress) +
  geom_col(aes(x = Stress_Level , y = AvgMark10, fill = Stress_Level)) +
  geom_text(aes(x = Stress_Level , y = AvgMark10, label = round(AvgMark10, 2)), vjust = 0.5, size = 3) +
  scale_fill_manual(values=c("#99CCFF",
                           "#CCCFFF",
                           "#FF99CC",
                           "#FFCCFF")) +
  theme_minimal() +
  labs(title = 'Average 10th Grade Mark by Stress Level',
     x = "Stress Level",
     y = "Average 10thth Grade Mark") +
  theme(title = element_text(size =9),
    axis.title = element_text(size = 8),
    axis.text.x = element_text(hjust = 1, size =5),
    legend.position = "none")

# average 12th grade mark by stress level plot
plot_12 = ggplot(mark_by_stress) +
  geom_col(aes(x = Stress_Level , y = AvgMark12, fill = Stress_Level)) +
  scale_fill_manual(values=c("#99CCFF",
                          "#CCCFFF",
                          "#FF99CC",
                          "#FFCCFF")) +
  geom_text(aes(x = Stress_Level , y = AvgMark12, label = round(AvgMark12, 2)), vjust = 0.5, size = 3) +
  theme_minimal() +
  labs(title = 'Average 12th Grade Mark by Stress Level',
       x = "Stress Level",
       y = "Average 12th Grade Mark") +
  theme(title = element_text(size =9),
        axis.title = element_text(size = 8),
        axis.text.x = element_text(size = 5),
        legend.position = "none")

# average college mark by stress level plot
college_plot = ggplot(mark_by_stress) +
  geom_col(aes(x = Stress_Level , y = AvgCollege_Mark, fill = Stress_Level)) +
  scale_fill_manual(values=c("#99CCFF",
                          "#CCCFFF",
                          "#FF99CC",
                          "#FFCCFF")) +
  geom_text(aes(x = Stress_Level , y = AvgCollege_Mark, label = round(AvgCollege_Mark, 2)), vjust = 0.5, size = 3) +
  theme_minimal() +
  labs(title = 'Average College Mark by Stress Level',
       x = "Stress Level",
       y = "Average College Mark") +
  theme(title = element_text(size =9),
        axis.title = element_text(size = 8),
        axis.text.x = element_text(size = 5),
        legend.position = "none")

# arrange the plots
grid.arrange(plot_10, plot_12, college_plot, ncol = 2)

## Logistic Regression
# To convert stress into binary variable:
data$Quantified_Stress = ifelse(data$Stress_Level %in% c("fabulous", "Good"), 0, 1)
## logistic regression
college_logreg = glm(Quantified_Stress ~College_Mark, data = data, family = binomial)
## Summary of the logistic regression model
summary(college_logreg)
```

Plot 2.3

```
ggplot(data, aes(x=College_Mark, y=Quantified_Stress)) + geom_point() +
      stat_smooth(method="glm", color="blue", se=FALSE,
                method.args = list(family=binomial))
```

2.2.4 Stress Levels v.s. Personal Factors

Plot 2.4

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

```
# How does having a part time job affect stress levels?
ggplot(data, aes(x = factor(Job), fill = factor(Stress_Level))) +
  geom_bar(position = "dodge") +
  labs(title = "Stress Level by Part-time Job",
       x = "Part-time Job?",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

## How does financial status affect stress levels?
ggplot(data, aes(x = factor(Financial_Status), fill = factor(Stress_Level))) +
  geom_bar(position = "dodge") +
  labs(title = "Stress Level by Financial Status",
       x = "Financial Status",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# How does travel time/commute to educational institutions affect stress levels?
ggplot(data, aes(x = factor(Travel_Time), fill = factor(Stress_Level))) +
  geom_bar(position = "dodge") +
  labs(title = "Stress Level by Travel Time to Educational Institution",
       x = "Commute Time",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

## Logistic Regression #2
# To quantify stress levels, convert into binary variable
data$Quantified_Stress = ifelse(data$Stress_Level %in% c("fabulous", "Good"), 0, 1)
## logistic regression
logreg = glm(Quantified_Stress ~ Job + Financial_Status + Travel_Time, data = data, family = binomial)
# Summary of the logistic regression model
summary(logreg)
## Predict probabilities
prob <- predict(logreg, type = "response")
## Assign predicted quantified stress levels
predicted <- ifelse(prob < 0.5, "Low Stress", "High Stress")

## confusion matrix
predicted = predict(logreg, data, type = 'response')
predicted_class = ifelse(predicted>=0.5, 1, 0)
confusion_matrix = table(actual = data$Quantified_Stress, predicted = predicted_class)
confusion_matrix
## calculate accuracy rate
accuracy = sum(diag(confusion_matrix))/sum(confusion_matrix)
accuracy
```

2.3.1 College Mark vs Hobbies

```
hobby_counts <- table(data$Hobbies)
hobby_df <- as.data.frame(hobby_counts)
# Bar Plot
barplot(hobby_df$Freq,
        names.arg = hobby_df$Var1,
        main = "College Marks vs Hobbies",
        xlab = "Hobbies",
        ylab = "College Marks",
        col = "skyblue")
unique(data$Hobbies)
data <- data %>%
  mutate(B = case_when(
    Hobbies == 'Cinema' ~ 1,
    Hobbies == 'Reading books' ~ 2,
    Hobbies == 'Sports' ~ 3,
    Hobbies == 'Video Games' ~ 4,
    TRUE ~ NA_integer_
  ))
data$B <- as.numeric(data$B)
# Scatter Plot
plot(data$B,data$College_Mark,main = "College Mark vs Hobbies",
     ylab = "College Mark",xlab = "Hobbies")
#Regression
the.model <- lm(College_Mark ~ Hobbies, data = data)
summary(the.model)
```

2.3.2 College Marks vs Engagement with Social Media

```
# Data Transform
data <- data %>%
  mutate(Social = case_when(
    Social_Media_Engagement == '0 Minute' ~ '0',
    Social_Media_Engagement == '1 - 30 Minute' ~ '15',
    Social_Media_Engagement == '30 - 60 Minute' ~ '30',
    Social_Media_Engagement == '1 - 1.30 hour' ~ '60',
    Social_Media_Engagement == '1.30 - 2 hour' ~ '90',
    Social_Media_Engagement == 'More than 2 hour' ~ '120'
  )) %>%
  select(-Social_Media_Engagement)
data$Social <- as.numeric(data$Social)
unique_values2 <- unique(data$Social)
social_counts <- table(data$Social)
social_time_df <- as.data.frame(social_counts)
social_time_df$Var1 <- factor(social_time_df$Var1, levels = unique(social_time_df$Var1))
# Scatter Plot
plot(data$Social, data$College_Mark,
     main = "College Mark vs Entertainment Time", ylab = "College Mark",
     xlab = "Entertainment Time", pch = 19, font = 2, font.lab = 2, cex = 0.75)
# Bar Plot
barplot(social_time_df$Freq,
        names.arg = social_time_df$Var1,
        main = "Histogram of Entertainment Time",
        xlab = "Entertainment Time",
        ylab = "Frequency")
# Regression
```

```
the.model_2.1 <- lm(College_Mark ~ Social, data = data)
summary(the.model_2.1)
```