MATH 408 Project - Students' Social Media Addiction

Euro Kim

Introduction

In 2025, social media presence is as influential as ever. In this project, I will be looking at a data set I found on Kaggle titled: "Students' Social Media Addiction".

Link: https://www.kaggle.com/datasets/adilshamim8/social-media-addiction-vs-relationships

I was interested in this topic because I see not only myself but many other people spending a lot of time online, often affecting their academic performance. With this data set, I will look into specifically undergraduate students in the top 3 populated countries, and try to analyze how average daily social media presence affects academic performance and mental health.

In this project, we will:

- Filter undergraduate students in the top 3 most populated countries
- Find descriptive summaries
- Do two sample t-tests
- Find potential correlation between social media usage and mental health

Data Preparation

```r
library(dplyr)
library(ggplot2)
library(knitr)


df <- read.csv("C:/Users/kimeu/Downloads/MATH408 Project/Students Social Media Addiction.csv")

# Convert variables
df <- df %>%
  mutate(
    Country = factor(Country),
    Academic_Level = factor(Academic_Level),
    Academic_Affected_num = ifelse(Affects_Academic_Performance == "Yes", 1, 0)
  )

# Filter undergraduate students only
df_undergrad <- df %>% filter(Academic_Level == "Undergraduate")

# Identify top 3 countries by sample size
top3 <- df_undergrad %>% count(Country, sort = TRUE) %>% pull(Country) %>% as.character() %>% .[1:3]

kable(data.frame(Top_3_Countries = top3),
      caption = "Top Three Countries by Sample Size")
```

Table 1: Top Three Countries by Sample Size

Top_3_Countries
USA
France
Spain

```r
df_top3_undergrad <- df_undergrad %>%
  filter(Country %in% top3) %>%
  droplevels()

country_summary <- df_top3_undergrad %>%
  count(Country)

knitr::kable(
  country_summary,
  caption = "Count of Students in Top 3 Countries",
  col.names = c("Country", "Count"))
```

Table 2: Count of Students in Top 3 Countries

Country	Count
France	24
Spain	22
USA	35

Descriptive Summary

```r
library(knitr)

df_top3 <- df_top3_undergrad %>%
  group_by(Country) %>%
  summarise(
    Samples = n(),
    daily_usage = round(mean(Avg_Daily_Usage_Hours, na.rm = TRUE), 2),
    mental_score = round(mean(Mental_Health_Score, na.rm = TRUE), 2),
    prop_academic_affected = round(mean(Academic_Affected_num, na.rm = TRUE), 2)
  )

kable(df_top3, caption = "Summary Mean Statistics for Top 3 Countries")
```

Table 3: Summary Mean Statistics for Top 3 Countries

Country	Samples	daily_usage	mental_score	prop_academic_affected
France	24	3.97	6.88	0.08
Spain	22	4.67	5.86	1.00
USA	35	7.10	4.74	1.00

This table shows a simple summary of the different categories we are interested in. Grouped by the top 3 countries, we can see the number of samples, the daily average time on social media in hours, the mean mental score, and the proportion of students reporting that there is academic impact.

Two Sample T Test

Since the Affects_Academic_Performance column is Yes/No, we test whether undergraduates who use social media more have a higher probability of reporting academic problems. We do a two-sample t-test comparing average usage hours between students who say “Yes” vs “No” at alpha level 0.05.

Hypothesis

$H_0 : \mu_{Yes} = \mu_{No}$
$H_a : \mu_{Yes} \neq \mu_{No}$

```
t_usage_academic <- t.test(
  Avg_Daily_Usage_Hours ~ Affects_Academic_Performance,
  data = df_top3_undergrad,
  var.equal = FALSE
)
t_usage_academic
```

```
##
##  Welch Two Sample t-test
##
## data:  Avg_Daily_Usage_Hours by Affects_Academic_Performance
## t = -9.9227, df = 76.111, p-value = 2.291e-15
## alternative hypothesis: true difference in means between group No and group Yes is not equal to 0
## 95 percent confidence interval:
##  -2.745733 -1.827764
## sample estimates:
##  mean in group No mean in group Yes
##          3.845455          6.132203
```

We compared the average daily social media usage hours between students who reported that social media does affect their academic performance (Yes) and those who reported it does not (No).

The two-sample t-test showed a **statistically significant difference** between the two groups: t = -9.92

- Mean usage (No): 3.85 hours

- Mean usage (Yes): 6.13 hours

- 95% CI for the mean difference: [-2.75, -1.83]

Because the confidence interval does not include 0 and the p-value is extremely small (0), we conclude that **students who say social media affects their academics spend significantly more time on social media** compared to those who say it does not.

Simple Linear Regression (Covered in Class)

To see whether social media usage is correlated with students’ mental health, we will use a simple linear regression equation where mental health score is predicted by average daily social media usage hours.

A simple regression equation is:

$$\text{Mental Health Score} = \beta_0 + \beta_1(\text{Usage Hours}) + \varepsilon$$

In this equation,

- β_0 is the expected mental health score for a student with 0 social media usage
- β_1 is the change in mental health score for each additional hour of daily social media use (slope)

```r
linear_equation <- lm(Mental_Health_Score ~ Avg_Daily_Usage_Hours, data = df_top3_undergrad)
summary(linear_equation)
```

```
## 
## Call:
## lm(formula = Mental_Health_Score ~ Avg_Daily_Usage_Hours, data = df_top3_undergrad)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.21765 -0.41143  0.09377  0.47093  0.95190 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             8.5395     0.2400   35.58   <2e-16 ***
## Avg_Daily_Usage_Hours  -0.5190     0.0418  -12.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6061 on 79 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6569 
## F-statistic: 154.2 on 1 and 79 DF,  p-value: < 2.2e-16
```

The regression equation shows a **negative relationship** between daily social media usage and mental health score. The slope estimate is $\beta_1 = -0.519$, meaning that for each hour of social media use, the mental health score decreases on average by about 0.52 points.

The model shows $R^2 = 0.6612$. This means approximately 66% of the variability in mental health scores among these students can be explained by differences in daily social media usage

Overall, the findings suggest that **higher social media usage is strongly associated with poorer mental health**. Students who spend more time on social media tend to have lower mental health scores, and this conclusion is both statistically significant and practically meaningful.