MY WORK- Students Social Media Addiction Analysis

Nam N Lai - Data 101 - September 21, 2025

1. Data selection and preprocess of dataset:
    a. Surfing Dataset on Kaggle suitable to perform the article (based on the number of downloads, comments,character,sufficient sample size ...)
    b. Eyeballing through the dataset to check for undefined, null, NaN variables to repair for preprocessing
    c. Review and understand the basic functions of Data R101 from the link to start handling dataset with R
    d. Download .csv file and import dataset to R studio
2. My analytical approach and reasoning:
    a. Started with basic summaries to understand data structure and identify potential issues:
        i. Used basic function: nrow(), ncol(), name(),... to understand structure
        ii. Use cut() to categorize some number or continuous variable(Addicted_Score, Avg_Daily_Usage_Hours, Mental_Health_Score) to more convenience for visualization or plots(boxplot effectively) , also to make comparisons clearer
        iii. Applied summary() to numerical columns to check for outliers or unusual distributions
        iv. Used table() on categorical variables to see distribution balance
    b. Pattern discovery process:
        i. Pattern 1 discovery: Lost Sleep to Scrolling

1. Initially as the thinking of a normal person, I think usage time vs sleep hour have reverse relation, so I try to draw the scatterplot with R by plot() function + abline for regression line
2. After that, I realized that there were a lot of dots distributions far from the linear. So I decide to explore the mystery behind it.
3. From this scatterplot, I observe from the x-axis and y-axis and then use subset () functions with condition, which I observed
4. I can eventually observe the Attribute from them and draw in common from them

ii. Pattern 2 discovery: "Conflict Fuels Addiction"

1. Initially as the inertia of a normal person, I think the level of social network addiction will be proportional to the time of use should use the tapply() function to see the dependent variables
2. Then I started with basic scatter plot: Usage_Hours vs Addicted_Score
3. I observe and identified that some students had low usage (3-4h) but high addiction scores
4. My hypothesis is that there must be another factor beyond just time spent
5. Then I try to draw boxplot between Addiction Level vs Avg usage hour and the outliner of plot prove me my hypothesis is understandable
6. I imported csv file to Chat GPT and prompted, which helps me to clarify the issue: addiction high but usage hour low, color on the conflict >3 (because the conflict of each student just from

1-5). I also ask LLM to explain the function plot() to make sure that i understand clearly the problem

7. Then I check again by trying to create a boxplot of Conflicts Over Social Media vs Addiction Level to find the relation of them.

iii. Pattern 3 discovery: "Platform of Doom and High School Student"

1. I also have my initial question are all platforms equally addictive?
2. I decided to use plotbox to visualize the addicted score vs platform, it can also help me to avoid outliners
3. Then I continue use tapply () function to check the relation of Mental_Health_Score and these platform because I believe it can help me to find the involvement of Mental_Health_Score and addicted score
4. Finally, it give me the answer that student with high addicted score have low Mental_Health_Score and I check again by scatter plot and abline ()
5. Next, I decided to find more insights about Academic Levels for more attractive articles, I create boxplot and check again by tapply() and make sure that high school is the most addicted level.
6. I create subset () of hs (high school level), then i draw boxplot of hs: Addiction Score based on Platform and Usage Hours vs Academic Level and then conclude the insight in the article

3. Validation: How I verified my findings:

a. Used both scatter plots and boxplots for same relationships, check again the consumption by tapply()

b. I always consider between mean value and median of plotbox, which can help me to avoid the value of outliners

c. I also try to analyze subset separately and eliminate uncertainty results when there are too few people used in that social networking platform

4. Prompt that I used:

   a. Explain the syntax and parameters of plot() function in R. When should I use scatter plots versus other plot type

   b. Explain the cut() function syntax for creating categorical variables from continuous ones.

   c. What's the difference between c(), min(), max(), and mean() functions?

   d. How does nrow(), ncol(), and str() help me understand my dataset structure?

   e.

   f. How do I choose appropriate break points for addiction levels?

   g. How do I interpret the relationship between two continuous variables like usage hours and sleep hour

   h. What does abline(lm( )) do in R scatter plots? How I interpret a negative slope regression line

   i. What does subset() function do in R?

   j. What's the difference between correlation and the regression line I'm seeing

   k. What R function help me subset data to examine specific group like sleep 8+ hours despite 5 hours usage?

   l. I want to create a scatter plot showing the relationship between daily social media usage hours (x-axis) and addiction scores (y-axis), I need to color the points based on conflict levels, red points for students with conflicts >= 3, and grey points for conflicts < 3 and then explain the function for me

m. How does tapply work?

n. When should I calculate group means versus creating boxplots to compare groups

o. I want to compare addiction scores across 8 different social media platforms. Should I use boxplot(), barplot(), or another visualization?

p. How is this different from tapply() with mean()? When should I count versus average?

q. What's the difference between barplot(table(x)) and barplot(tapply(y,x,mean))?

r. I have three academic levels (High School, Undergraduate, Graduate) and want to compare their addiction scores

s. I want to show that conflicts contribute to addiction even with low usage.

t. How do I create a legend in R plots?

u. When I have three related variables (usage hours, addiction score, conflict level), what's the best way to visualize their relationship using basic R plotting functions?

…

→ I mainly use GPT Prompt sentences to analyze and choose the appropriate plot types to perform relationships. Sometimes when I was tangled and did not know how to write the Syntax of Ham, I asked the GPT chat but still asked to explain to make sure I understood that function in the best way. After running the code, I have to review this code to work properly with its variables

4. Resource used:

- [Tutorial of Data101]

- The Chat GPT 5.0 version prompt
- The Dataset Student’s Social Media Addiction of Adil Shamim on Kaggle
- The note of Data101