HW3 -Final Data 101 report

Nam N Lai - Data 101- November 28 2025
Dataset: Students_Social_Media_Addiction.csv (N = 705)

1. Languages of Data: Translating Raw Chaos into Meaning
“ Who Scrolls, Who Sleeps, Who Struggles? The Story Hidden in 705 Rows”

- **Dataset description & variable types**
    - This dataset tracks **705 students** from different countries and education levels, focusing on how they use social media, how much they sleep, and how they feel about their mental health and academic performance. It contains **13 original variables**, which can be separated into numeric and categorical types.
    - **Numeric variables:** Student_ID, Age, Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, Mental_Health_Score, Conflicts_Over_Social_Media, Addicted_Score
    - **Categorical variables:** Gender, Academic_Level, Country, Most_Used_Platform, Affects_Academic_Performance, Relationship_Status

    -> Using summary() on the numeric variables, I find that students spend on average about **4.9 hours per day** on social media and sleep around **6.9 hours per night**. Addiction scores (Addicted_Score) range roughly from **2 to 9**, suggesting that the sample includes both light users and highly addicted users

- Distributions using table(), subset(), and tapply()
    - First, table(Gender) shows that the sample is almost perfectly **gender-balanced**, with **353 females** and **352 males**. table(Academic_Level) reveals that most participants are **Undergraduate** students, with smaller groups of **High School** and **Graduate** students. This indicates that the dataset is heavily centered on university students but still includes some younger and older learners. table (Most_Used_Platform) shows that Instagram (249 students, about 35%) is the dominant platform in this sample, followed by TikTok (154 students, about 22%) and Facebook (123 students, about 17%). Messaging apps like WhatsApp (54 students, about 8%) form a middle tier, while platforms such as KakaoTalk, LINE, VKontakte, WeChat, Snapchat, LinkedIn, Twitter, and YouTube each account for only a small fraction of the sample.The Affects_Academic_Performance table shows that a clear majority of students believe social media impacts their studies: **453 out of 705 students (about 64%)** answered “Yes”, while only **252 (about 36%)** answered “No”
    - Next, I examine addiction and platform use across groups:
        - Using tapply(Addicted_Score, Academic_Level, mean), I find that:
            - **High School** students tend to have the **highest average addiction scores**.
            - **Graduate** students have slightly lower average addiction levels.

        - Undergraduates fall somewhere in between.
    - Using tapply(Avg_Daily_Usage_Hours, Most_Used_Platform, mean), I see that:
        - Platforms such as **WhatsApp** and **TikTok** emerge as the **heaviest time sinks**.
        - Their average daily usage is **above the overall mean** of roughly 4.9 hours.
    - Finally, subset() helps me zoom in on specific groups that will be important later:
        - subset(social, Academic_Level == "Undergraduate") isolates **undergraduate students only**
        - subset(social, Addicted_Score >= 8) focuses on students with **very high addiction scores**

- **Data Transformation and New Risk-Focused Variables**
    - Addiction_Level groups students into three categories according to their Addicted_Score:
        - *Low* for scores ≤ 4
        - *Moderate* for scores between 5 and 7
        - *High* for scores ≥ 8.

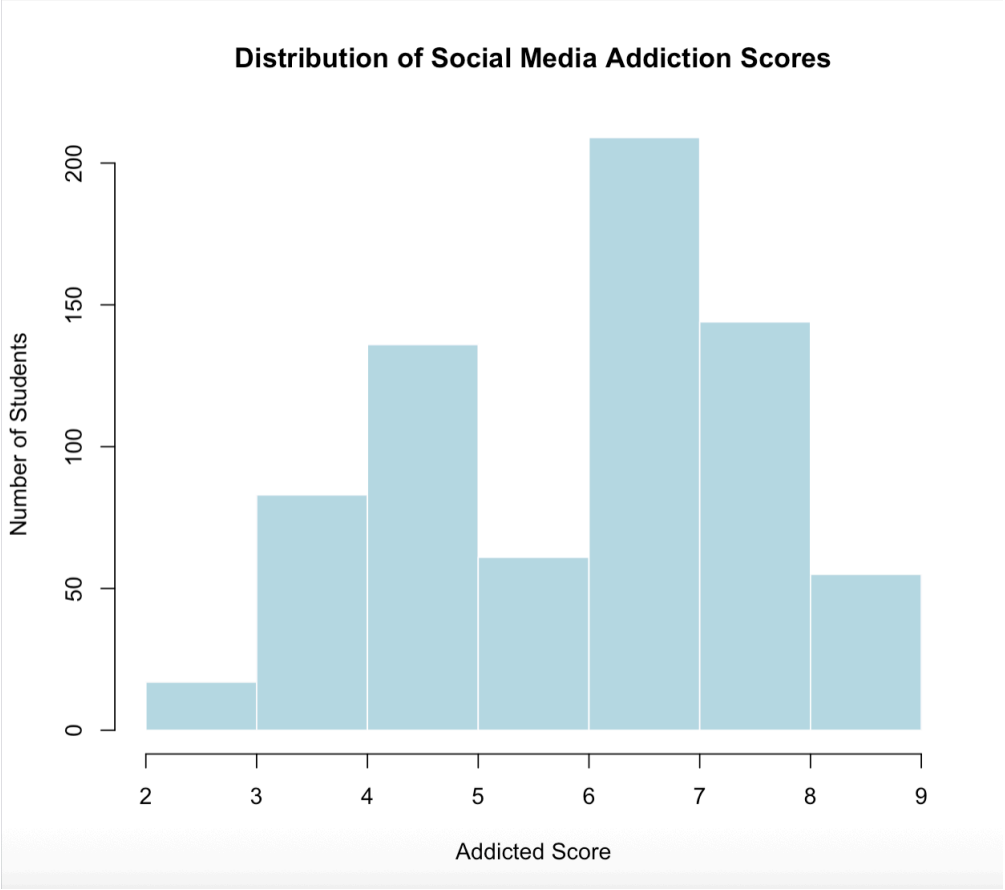
          This categorical variable turns the continuous addiction scale into interpretable levels that are easy to compare across academic groups and platforms.
    - High_Addiction is a binary flag that equals 1 when a student's Addicted_Score is at least 8 and 0 otherwise. This indicator focuses attention on the most extreme cases of social media addiction and will be useful for chi-square tests and classification-type models.
    - Sleep_Deprived is defined as 1 if a student reports fewer than 7 hours of sleep per night (Sleep_Hours_Per_Night < 7) and 0 otherwise. This variable identifies students who may not be getting enough rest.
    - Heavy_User equals 1 for students who use social media for 5 or more hours per day (Avg_Daily_Usage_Hours ≥ 5) and 0 otherwise. This variable flags very intensive social media users.
    - Usage_to_Sleep_Ratio is a continuous variable calculated as Avg_Daily_Usage_Hours / Sleep_Hours_Per_Night. Instead of looking at usage and sleep separately, this ratio shows how many hours of social media use correspond to one hour of sleep, highlighting students whose screen time may be competing directly with their rest
    - Affects_Binary is a binary variable created from Affects_Academic_Performance, where 1 represents students who answered "Yes" and 0 represents those who answered "No". This makes academic

impact easier to use as an outcome in regression models and as an event in Bayesian reasoning

- High_Risk_Study_Impact equals 1 if a student is both highly addicted (High_Addiction = 1) and reports that social media affects their academic performance (Affects_Academic_Performance == "Yes"). Otherwise, it equals 0. This variable pinpoints the students with the most concerning profiles: those who are not only heavy users but also recognize that their social media habits are harming their studies

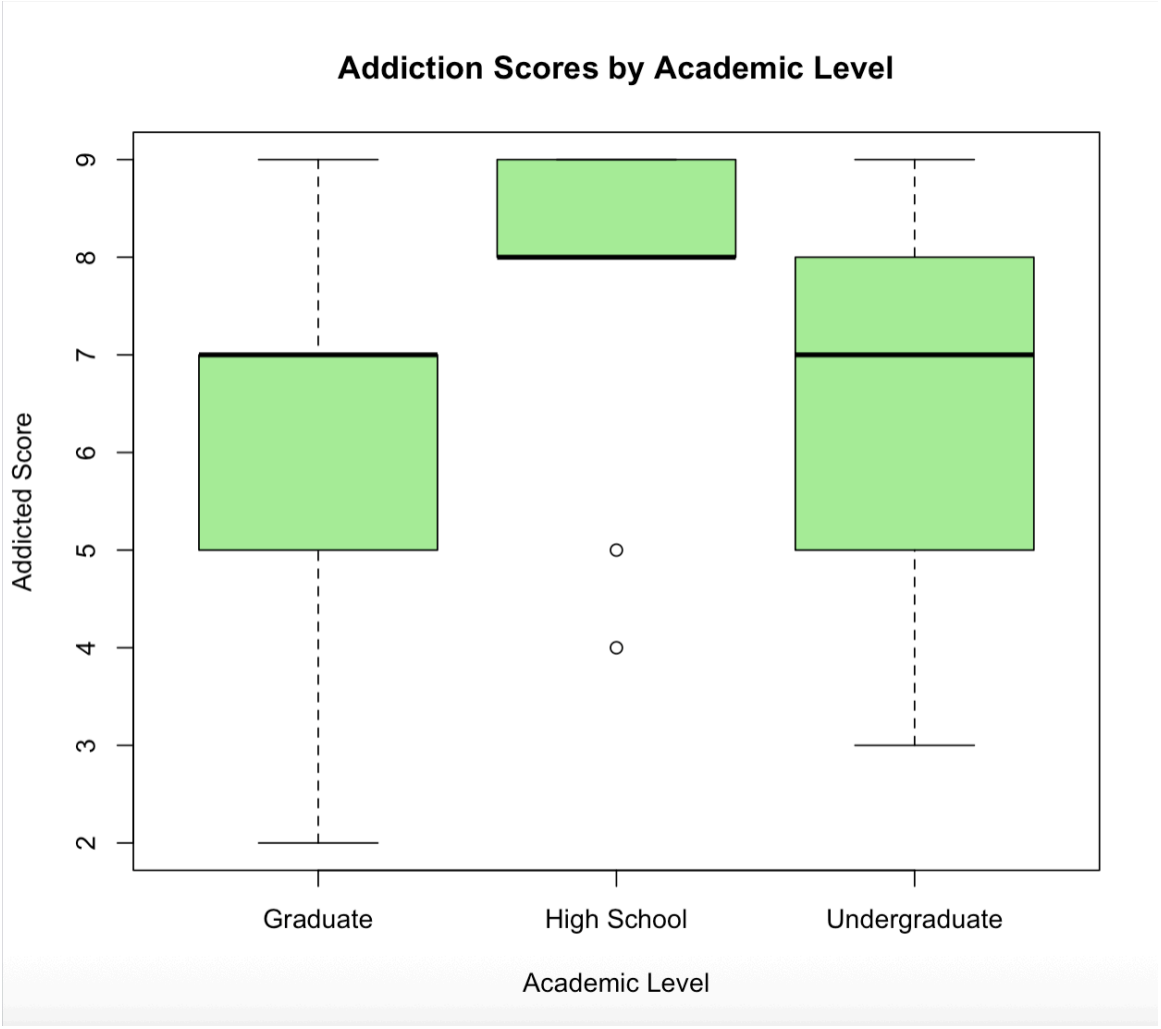
2. Exploratory Data Analysis: Seeing the Invisible

"More Scroll, Less Sleep? Four Plots that Expose Students' Habits"



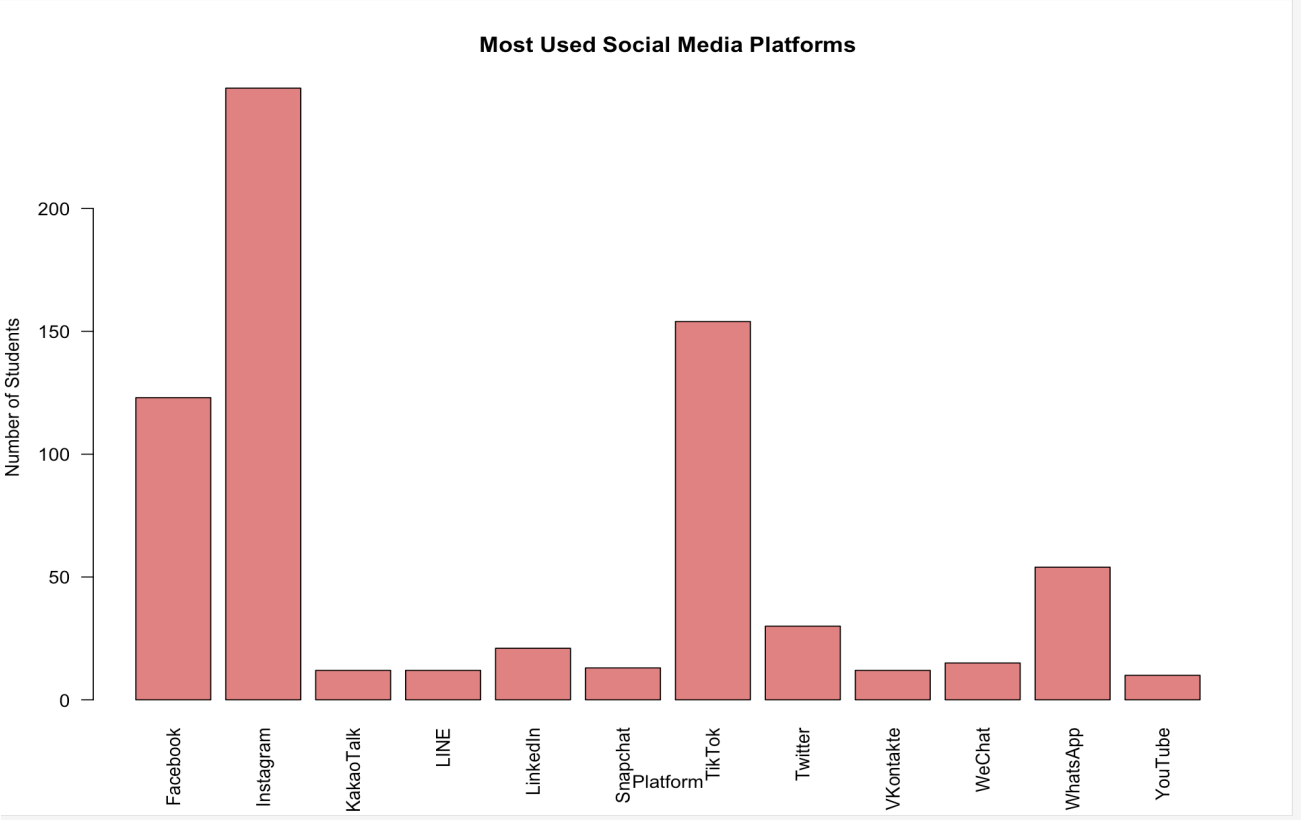
Plot 1 – Histogram of social media addiction scores

The first plot is a histogram of Addicted_Score for all 705 students. The distribution is clearly shifted toward the higher end of the scale: most scores fall between 5 and 8, with the median around 7. Only a tiny fraction of students have very low scores (for example, there is only 1 student with a score of 2), while a large share have scores of 7 or 8. Overall, about two-thirds of the sample have Addicted_Score of 6 or higher, which suggests that moderate to high levels of social media addiction are the norm rather than the exception in this dataset



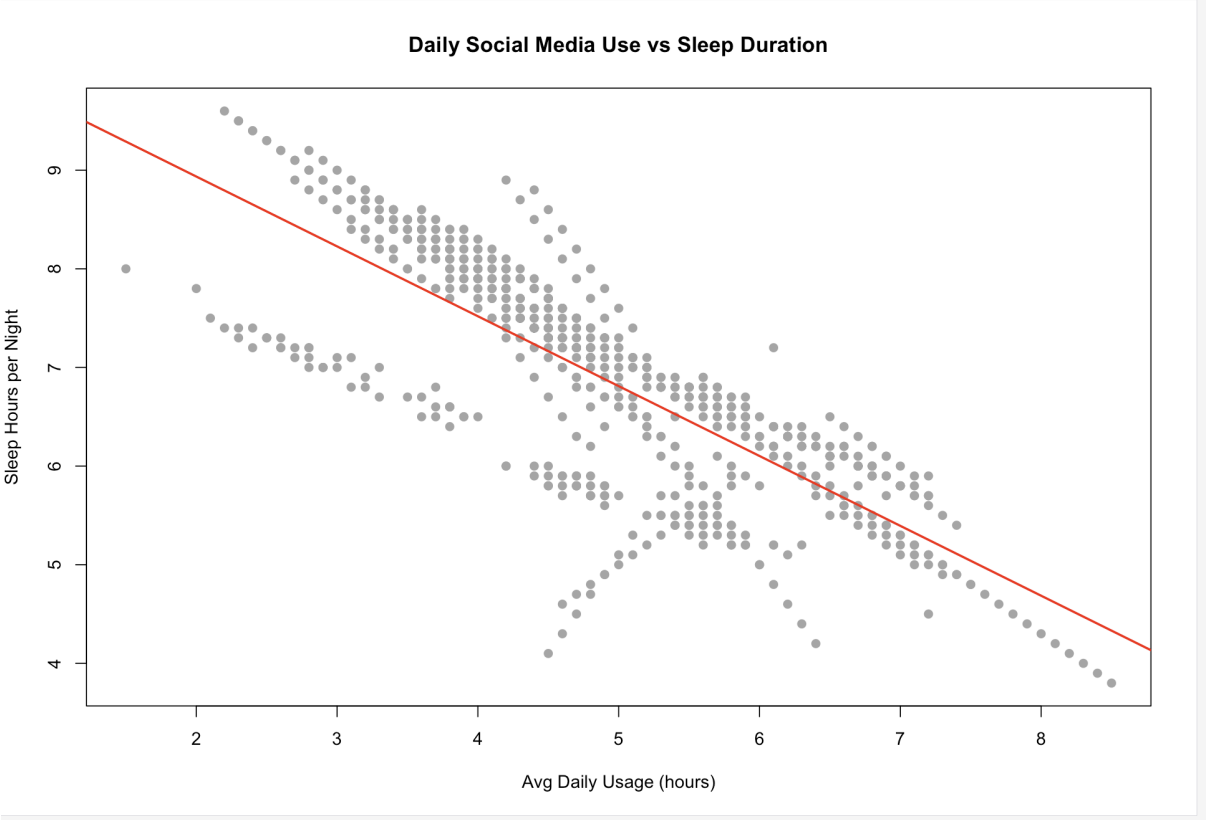
Plot 2 – Boxplot of addiction scores by academic level

The second plot compares Addicted_Score across Academic_Level using a boxplot. The boxes show that High School students have the highest typical addiction scores(~8), with their median close to the top of the scale. Undergraduate students have slightly lower scores on average, while Graduate students tend to have the lowest addiction levels among the three groups. The spread within each group is also visible: undergraduates and high school students have a wide range of scores, whereas graduate students are somewhat more concentrated around the middle-high range. This suggests that younger students might be more vulnerable to high social media addiction.



Plot 3 – Barplot of the most used platforms

The third plot is a barplot of Most_Used_Platform. The bars reveal that the sample is dominated by a few major platforms: Instagram (249 students, about 35%), TikTok (154, about 22%), and Facebook (123, about 17%). Messaging apps like WhatsApp (54 students) form a smaller but still noticeable group, while other platforms such as Twitter, LinkedIn, WeChat, Snapchat, LINE, KakaoTalk, VKontakte, and YouTube each have relatively few users (typically between 10 and 30 students). This creates a clear “long tail” pattern: most students cluster around a small set of highly popular apps, while many niche platforms are used by only a small minority

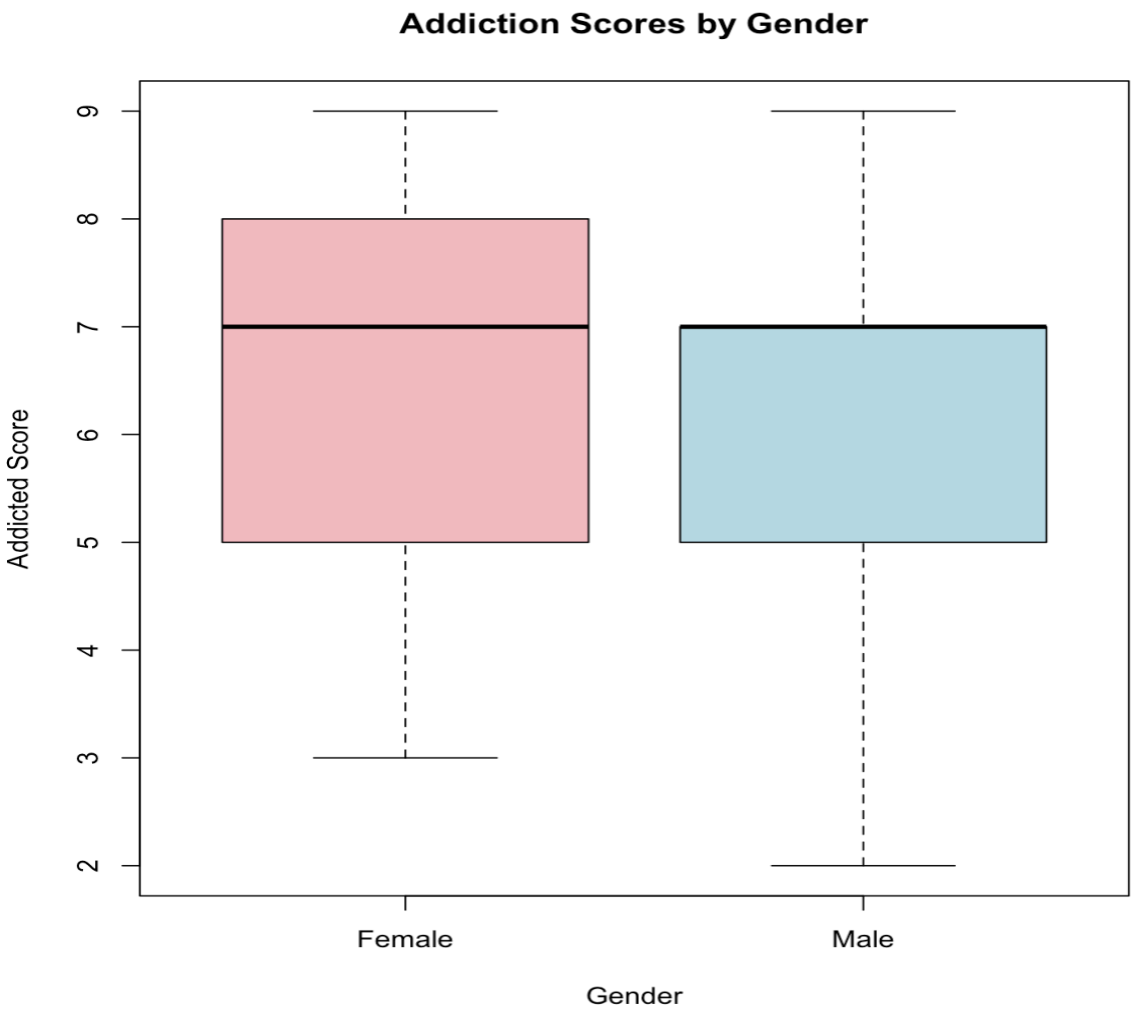


Plot 4 – Scatterplot of daily usage vs. sleep duration

The fourth plot is a scatterplot of Avg_Daily_Usage_Hours on the x-axis and Sleep_Hours_Per_Night on the y-axis, with a red regression line added. The cloud of points shows a strong downward trend: students who spend more time on social media each day tend to report fewer hours of sleep. The fitted line slopes downward, confirming this negative relationship. For example, many students using social media for around 6–8 hours per day cluster in the lower range of sleep (around 5–6 hours), while those with more moderate usage often report closer to 7–8 hours of sleep. This visual evidence supports the idea that heavier social media use may be closely linked to reduced sleep duration.However, the scatterplot also shows some **exceptions**: a few students report **low social media use but still sleep very little**, and others use social media heavily yet still manage relatively high sleep duration.

-> Together, these four plots show a consistent story: most students in the dataset are moderately to highly addicted to social media, younger students tend to have higher addiction levels, a few platforms dominate students' attention, and heavier daily usage is associated with less sleep. These visual patterns motivate the more formal hypothesis tests and models in the later sections of the project.

3. Fooled by Data: When Randomness Plays Tricks
"Are Women Really More Addicted? A Gender Gap That Randomness Can Easily Fake"



To see how easily I could be fooled by random variation, I looked at a very common question: are women more addicted to social media than men? At first glance, the data seemed to say "yes." Using tapply(Addicted_Score, Gender, mean), I found that female students had a slightly higher average addiction score than male students. A simple boxplot of Addicted_Score by Gender showed the box for females shifted a bit upward, and it was tempting to interpret this as a real gender difference.

The observed numbers look like this: female students have an average addiction score of about m_F, male students have about m_M, and the difference

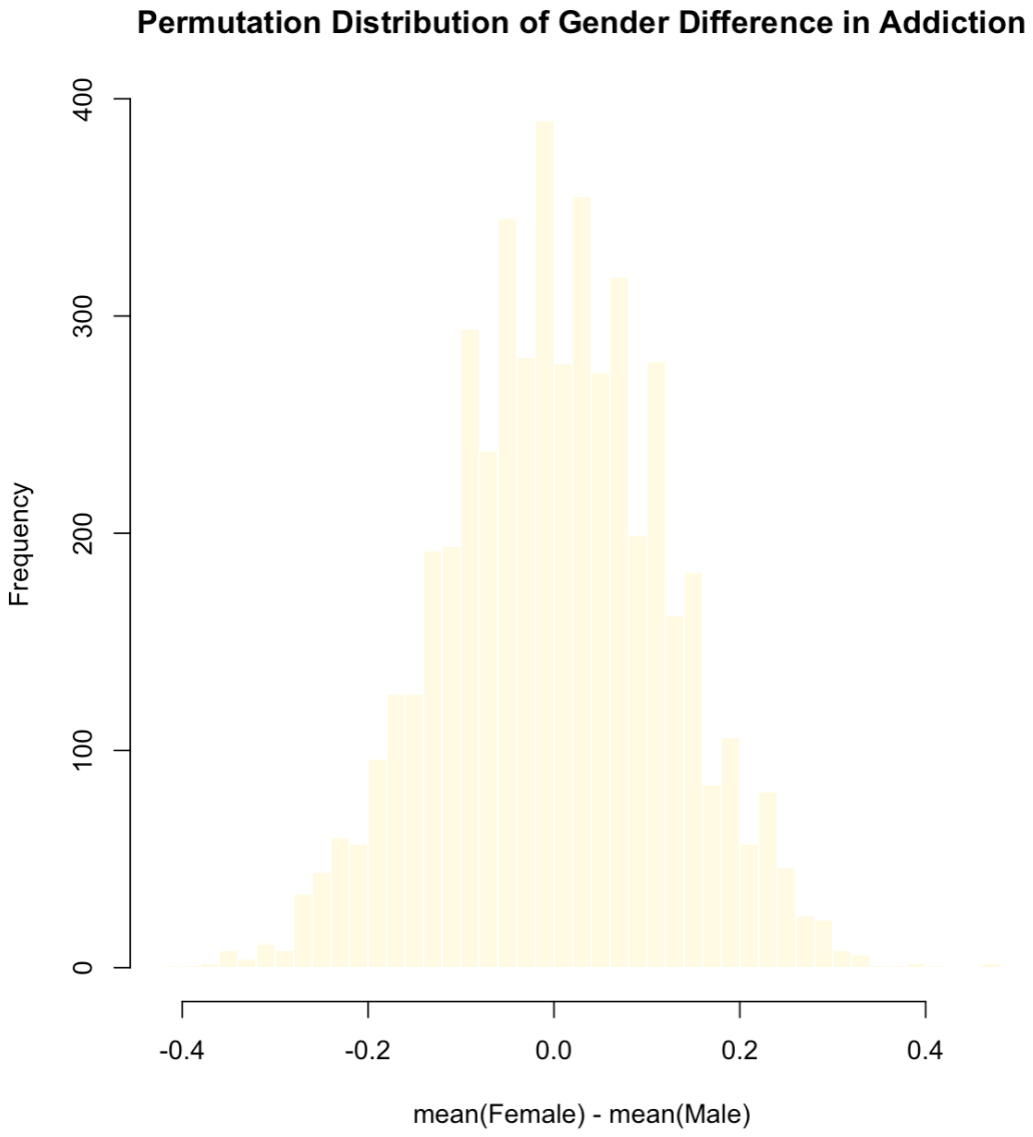
$$\text{obs\_diff} = \text{mean female} - \text{mean male} \approx 0.16$$

is positive. Even though 0.16 is not a huge gap, the boxplot makes it look like there is a systematic tendency for women to be more addicted than men.

To check whether this small difference could simply be due to randomness, I performed a permutation-style randomization. I kept all the addiction scores fixed, but I randomly shuffled the gender labels across students 5,000 times. In each simulation, I recomputed the difference

in mean addiction score between the “female” and “male” labels under this random assignment. This produced 5,000 values of

$$\text{diff\_sim} = \text{mean female(perm)} - \text{mean male(perm)}$$



Plotting a histogram of diff_sim gave me the permutation distribution of the gender difference under the null hypothesis that gender and addiction are unrelated. The distribution is centered very close to 0 and spreads out symmetrically in both directions, showing the kinds of gender gaps that could appear purely from random mixing of scores and labels.

I then added a vertical red line at the observed difference obs_diff ≈ 0.16. The red line sits comfortably inside the main bulk of the histogram, not in the extreme tails. When I computed the permutation p-value,

$$\text{p\_perm} = \Pr(|\text{diff\_sim}| \geq |\text{obs\_diff}|) \approx 0.19,$$

I found that about 19% of the random shuffles produced a gender gap as large as, or larger than, the one seen in the real data. In other words, a difference of 0.16 points between male and female addiction scores is not remotely unusual under pure randomness.

The misleading pattern I almost trusted was the idea that "women are clearly more addicted than men" based on a small gap in the boxplot. The permutation analysis showed me that this apparent gender effect could easily be generated by random variation alone. This section reminded me that even clean-looking plots and small differences in group means can be illusions, and that without a low p-value or a clearly extreme statistic, I should be very cautious about turning a visual impression into a strong claim about real differences between groups.

4. Central Limit Theorem & Confidence Intervals
"Instagram Time Is Almost Clockwork: 95% CI Puts Daily Use Between 4.72 and 5.03 Hours"

For the confidence-interval and sampling section, I focused on one concrete quantity: the average daily social-media usage among students whose most-used platform is **Instagram**. After subsetting to Instagram-primary users and removing missing values, I obtained a sample of **n = 249** students. Their average daily usage is **4.872 hours**, with a standard deviation of **1.238 hours**, so individual behavior is quite variable

Using the standard normal critical value z=1.96 for 95% confidence, the standard error of the mean is $s / \sqrt{n} = 1.238 / \sqrt{249} \approx 0.078$ hours. This leads to a margin of error of about **0.154** hours and a 95% confidence interval of **[4.719, 5.026]** hours per day. Interpreted in context, if I were to repeat this sampling process many times, about 95% of the intervals built this way would contain the true population mean for Instagram-primary students. The interval is very narrow, which is only about 0.31 hours wide, so my estimate is quite precise: it strongly suggests that Instagram is not an occasional distraction, but a stable daily habit of roughly **five hours** for this group

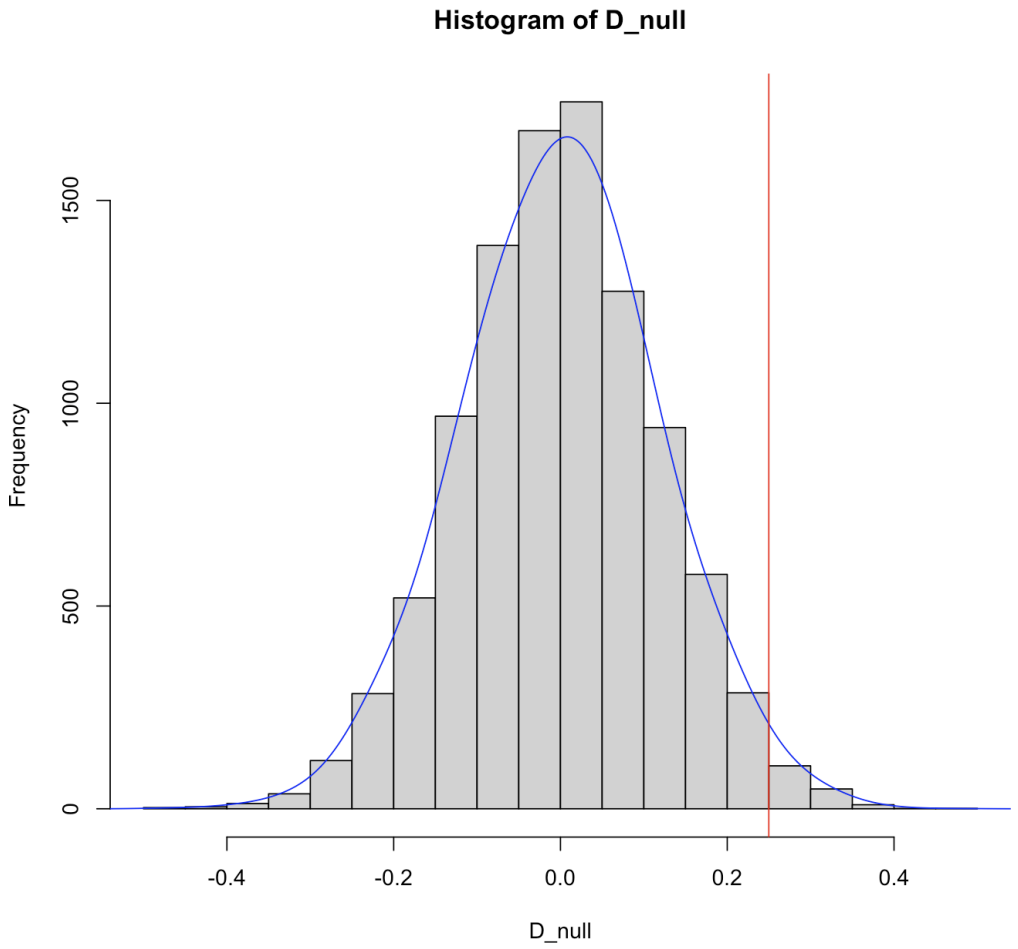
5. Hypothesis Testing: When Suspicion Meets Statistics

"Undergrads Edge Out Grads on Addiction: A 0.25-Point Lead That Both Tests Agree On"

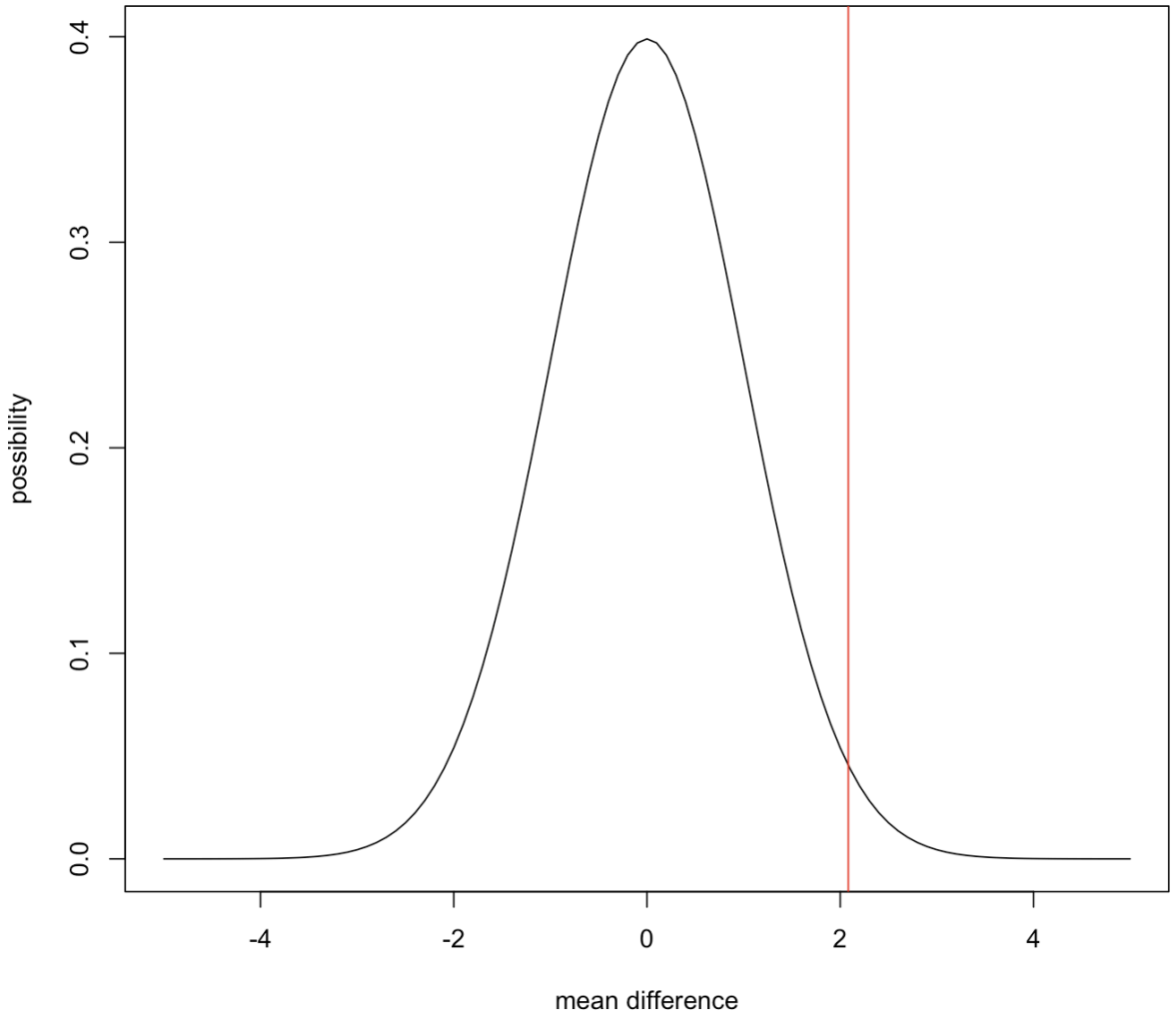
$\mathbf{H_0}$: The average value of Addicted_Score is the same when Academic_Level = 'Undergraduate' and when Academic_Level = 'Graduate'
$\mathbf{H_1}$: The average value of Addicted_Score is greater when Academic_Level = 'Undergraduate' than when Academic_Level = 'Graduate'

To compare addiction levels across academic stages, I tested whether **undergraduate students** score higher on the Addicted_Score scale than **graduate students**. The null hypothesis states that both groups have the same mean addiction score, while the one-sided

alternative claims that undergraduates are more addicted. Using simple group means, I found that undergraduates score about **0.25 points higher** on average than graduates (obs_diff ≈ 0.2498)



Permutation p(one-tailed) = **0.0167 (10,000 permutations)**



z-value: **2.083**; p(one-tailed) = **0.0186**

I then ran two complementary tests using the same code: a **z-test** and a **permutation test**. The z-test produced a z-value of **2.083** and a one-tailed p-value of **0.0186**. The permutation test, which repeatedly shuffles the academic-level labels 10,000 times to simulate the null, gave a nearly identical one-tailed p-value of **0.0173**

Because both p-values are well below 0.05, I reject the null hypothesis at the 5% significance level and conclude that undergraduates are more addicted to social media than graduates in this dataset. The effect size is modest but consistent across two very different methods, suggesting that younger students, who have less academic experience and possibly weaker time-management skills, are genuinely more vulnerable to social-media addiction rather than this being a random fluctuation

6. Hypothesis Testing: When Suspicion Meets Statistics

“Academic Stage and Academic Harm Are Not Independent: High School Students Stand Out”

H_0: Academic_Level and Affects_Academic_Performance are independent
H_1 : Academic_Level and Affects_Academic_Performance are independent are associated

Next, I examined whether **academic level** is related to whether students feel social media affects their academic performance. The two categorical variables are Academic_Level (Graduate, High School, Undergraduate) and Affects_Academic_Performance (Yes/No). The null hypothesis states that these two variables are independent; the alternative states that they are associated.

- Independence (Null Hypothesis): expected of Academic_Level and Affects_Academic_Performance

	No	Yes
Graduate	116.170213	208.82979
High School	9.651064	17.34894
Undergraduate	126.178723	226.82128

- Observation:

	No	Yes
Graduate	126	199
High School	2	25
Undergraduate	124	229

The chi-square test of independence on the 3×2 contingency table yields a test statistic of **X^2 = 10.79** with **2 degrees of freedom** and a p-value of about **0.0045**, clearly below the 0.05 threshold. Thus, I reject the null hypothesis and conclude that academic level and perceived academic impact are not independent. Inspecting the observed vs expected counts shows that **high school students are the main drivers of this association**: under independence, we would expect about 17.35 high school students to say “Yes” and 9.65 to say “No,” but in reality **25** say “Yes” while only **2** say “No.”

All expected counts are above 5, so the chi-square approximation is valid. The pattern suggests that social media’s academic impact is especially pronounced among high school students, who may face stronger peer pressure and have less developed self-regulation than

undergraduates and graduates. In contrast, the university-level groups are much closer to their expected counts, reinforcing the idea that early academic stages are the most vulnerable.

7. Multiple Hypothesis Testing: The False Discovery Jungle
"Is TikTok more addictive than other major social-media platforms?"

I computed the mean Addicted_Score for each of the 12 platforms in the dataset and then considered all pairwise comparisons. With 12 platforms, there are 12*(12-1)/2= 66 possible pairs, so controlling the family-wise error rate at 5% via the Bonferroni method leads to a very strict threshold of α Bonf=0.05/66 ≈ 0.00076

Within this larger family, I explicitly tested 10 one-sided hypotheses of the form "TikTok has a higher mean addiction score than platform X" using permutation tests with 10,000 random label shuffles each.

The results split sharply into two groups. For eight platforms: **Instagram, Facebook, YouTube, Twitter, WeChat, LINE, LinkedIn, and VKontakte**, the permutation p-values were effectively **0.0000**, meaning that in 10,000 simulated datasets under the null, none produced a TikTok–X difference as large as the one observed. All eight of these p-values are far below 0.05 and still comfortably below the Bonferroni-corrected cutoff of 0.00076, so these gaps remain statistically significant even after accounting for the fact that many platform pairs could have been compared. In other words, TikTok users appear systematically more addicted than users of these eight platforms, and this conclusion is robust to a strong multiple-testing correction

	platform1	platform2	p_perm
1	Instagram	TikTok	0.0000
2	Facebook	TikTok	0.0000
3	YouTube	TikTok	0.0000
4	WhatsApp	TikTok	0.5780
5	Twitter	TikTok	0.0000
6	Snapchat	TikTok	0.5908
7	WeChat	TikTok	0.0000
8	LINE	TikTok	0.0000
9	LinkedIn	TikTok	0.0000
10	VKontakte	TikTok	0.0000

The remaining two comparisons, however **WhatsApp vs TikTok** and **Snapchat vs TikTok** tell a different story. Their permutation p-values are about **0.578** and **0.591**, respectively, which are not only non-significant at the usual 0.05 level but also vastly larger than the Bonferroni threshold. These large p-values indicate that the observed differences between TikTok and WhatsApp or Snapchat are completely compatible with random variation; in fact, the mean addiction scores for WhatsApp and Snapchat are slightly higher than that of TikTok in this sample, so a one-sided test in the direction "TikTok > X" has no support. After correction, the only safe conclusion is that TikTok stands out as more addictive than many platforms (such as Instagram, Facebook, and YouTube), but **not uniquely**: WhatsApp and Snapchat sit in a similarly high-addiction range and cannot be statistically ruled out as "just as bad" as TikTok.

8. Bayesian Reasoning: Updating Beliefs

"TikTok as a Red Flag: One Platform Pushes Academic-Impact Odds from 64% to 94%"

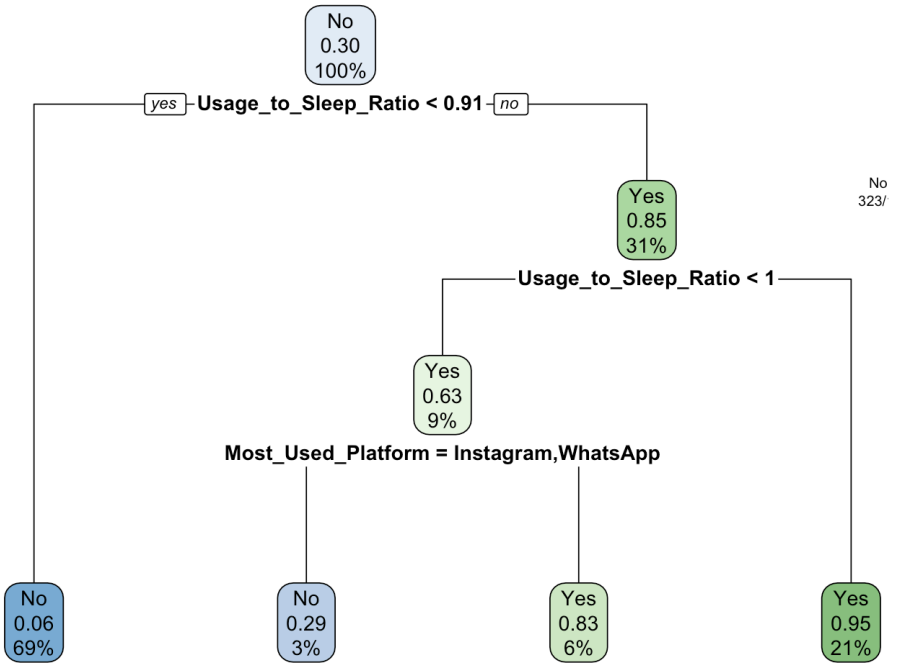
For the Bayesian part of my analysis, I treated "being academically affected by social media" as my initial belief and "using TikTok as the most-used platform" as new evidence. Before conditioning on any platform, the prior probability that a randomly chosen student reports academic impact is about 0.643, corresponding to prior odds of roughly 1.80 : 1 in favor of "Yes." I then computed how informative TikTok usage is. Among students who say they are affected, 31.8% use TikTok, whereas among those who say they are not affected, only 4.0% do. This yields a likelihood ratio of about 8.01, meaning TikTok usage is eight times more common among affected students than unaffected ones. Multiplying this LR by the prior odds produces posterior odds of about 14.4 : 1, which correspond to a posterior probability of approximately 93.5% that whose main platform is TikTok reports academic impact

Simply, before I know anything about the platform, I would guess that about two-thirds of students feel academically affected by social media. After learning that a student's main platform is TikTok, that belief jumps to roughly 94%. The large likelihood ratio shows that TikTok is a highly diagnostic signal: it does not prove causation, but it strongly tilts the odds toward academic problems. Bayesian updating thus turns a broad concern ("social media might hurt grades") into a much sharper statement about a specific platform

9. Prediction Models: From Correlation to Prediction

"Can a Tree Predict Who Gets Hooked? Turning Sleep and Screen Time into Warnings."

For the prediction part of my project, I built a classification tree using rpart() to predict which students fall into the high-addiction group. I defined the binary target High_Addiction_F as “Yes” when a student’s Addicted_Score was at least 8 and “No” otherwise. As predictors, I combined demographic variables (Gender, Academic_Level), behavioral variables (Most_Used_Platform, Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, Affects_Academic_Performance), and several engineered features (Usage_to_Sleep_Ratio, Sleep_Deprived, Heavy_User). After removing rows with missing values, I obtained 705 complete cases and randomly split them into a 70% training set and a 30% testing set using set.seed(123) for reproducibility.



The decision tree was trained with method = "class" and mild regularization (minsplit = 20, cp = 0.01) to avoid overfitting. On the training data, the model correctly classified 93.1% of students, while on the held-out testing data it achieved an accuracy of 91.0%. The small gap between training and testing accuracy suggests that the tree captures real patterns in the data rather than simply memorizing the training set. In other words, a simple tree with a few intuitive rules is already able to predict high addiction quite reliably.

The plot of the fitted tree reveals how the model “thinks.” The root split is on the Usage_to_Sleep_Ratio: students who spend less than about 0.91 hours of social media per hour of sleep are sent to a leaf that almost always predicts “No,” meaning they are very unlikely to be highly addicted. On the right-hand side, where the ratio is higher, the tree

further splits at Usage_to_Sleep_Ratio < 1, creating a region of moderate imbalance where some students are predicted “Yes” and others “No.” Deeper down, Most_Used_Platform becomes important: among students with high usage-to-sleep ratios, those who mainly use Instagram or WhatsApp are routed into leaves with very high probabilities of “Yes,” while students with lower ratios or different platforms are more likely to be classified as non-addicted. Overall, the tree formalizes a common sense story: heavy screen time relative to sleep, combined with certain high-intensity platforms, is a strong warning sign of high social-media addiction.

10. Association Rules & Lift: The Hidden Recipes

“Heavy Use + TikTok → High Addiction: Hidden Recipes in Student Data.”

For the association-rules section, I treated each student as a “basket” of binary items describing their digital habits: whether they are a heavy user (≥ 5 hours of social media per day), sleep-deprived (< 7 hours of sleep), primarily a TikTok, Instagram, or WhatsApp user, whether they report academic impact, and whether they belong to the high-addiction group (Addicted_Score ≥ 8). I converted these flags into a transaction object and ran the apriori() algorithm with a minimum support of 5%, minimum confidence of 60%, and rule length between 2 and 3. I then focused on rules whose lift exceeded 1, which indicate a positive association between the left-hand side (LHS) “if” pattern and the right-hand side (RHS) outcome.

Several strong and interpretable rules emerged. One of the most striking is {Heavy_User, TikTok_User} → {High_Addiction}.This rule has a support of about 9.8%, meaning nearly one in ten students both use social media heavily, primarily on TikTok, and fall into the high-addiction group. Its confidence is roughly 75%, so three out of four heavy TikTok users are classified as highly addicted. Because the overall probability of high addiction in the dataset is only about 28%, the lift of this rule is about 2.66—being both a heavy user and a TikTok user more than doubles the odds of high addiction compared to a random student.

A second rule, {Heavy_User, Sleep_Deprived} → {High_Addiction}, is also powerful. It covers about 26% of the sample and has a confidence of roughly 63%, giving a lift of around 2.24. This means that students who both spend many hours on social media and sleep less than seven hours per night are more than twice as likely to be highly addicted as the average student. Even the simpler rule {TikTok_User} → {High_Addiction} has a confidence of about 51% and a lift of 1.79, showing that TikTok users alone already skew toward the high-addiction end of the scale.

Finally, the rule {High_Addiction} → {Affects_Academic_Performance = Yes} has support around 28% and a confidence of 100% in this dataset: every highly addicted student reports that social media affects their academic performance. Because only about 64% of all students say that their academics are affected, this rule has a lift of roughly 1.56. In plain language, the association rules uncover a clear “recipe” for trouble: heavy, TikTok-centered

usage combined with lack of sleep strongly predicts high addiction, and high addiction almost always comes with self-reported academic problems. These patterns do not prove causation, but they show that certain combinations of habits are far from random noise, they are statistically consistent red flags.

11. Reflection: What I Learned & Why It Matters

- **One paragraph on the dataset's biggest insight:**

The biggest thing I learned from this project is that "social media addiction" in this sample is not just about using social media a lot; it's about how that usage combines with sleep, platform choice, and school performance. When I split students into Low, Moderate, and High addiction, I saw that a surprisingly large chunk fall into the moderate range, but the high-addiction group is special: almost every highly addicted student reports that their academics are affected, and many are both heavy users ($\geq$ 5 hours/day) and sleep-deprived ($<$ 7 hours). The decision tree and association rules hammered this home. A simple tree using just the usage-to-sleep ratio and platform choice already predicted high addiction with about 91% accuracy on the test set, and rules like {Heavy_User, TikTok_User} $\rightarrow$ High_Addiction and {Heavy_User, Sleep_Deprived} $\rightarrow$ High_Addiction had lifts well above 2. That means these "recipes" for trouble more than double the odds of addiction compared to an average student. I also expected TikTok to be the lone villain, but multiple-testing results showed a small cluster of high-risk platforms: TikTok is clearly worse than many apps, yet WhatsApp and Snapchat sit in a similarly dangerous zone. The data ended up telling a more nuanced story than "phones are bad", it pointed to specific combinations of habits that consistently show up when students feel their grades are suffering.

- **One paragraph on the dataset's biggest insight:**

The idea that surprised me most was how strongly the chi-square tests showed that "who you are" and "where you scroll" are not independent of addiction and academic problems. Before running the tests, I expected small, vague differences across groups, but the contingency tables told a much sharper story. When I looked at Academic_Level versus academic impact, high school students clearly had the highest proportion of "Yes" responses, meaning they were more likely than undergraduates and graduate students to say that social media affects their performance. The chi-square statistic was large enough that the null hypothesis of "no relationship" between level and impact was very hard to believe. A similar pattern appeared when I grouped students by their most-used platform and looked at the share who were highly addicted or reported academic problems. TikTok users showed noticeably higher proportions than most other platforms, and the differences were too large to be explained by random variation alone. Seeing those differences of proportions quantified, made the patterns feel real in a way that simple bar charts never did. It turned my loose intuition (maybe high school students and TikTok users are worse

off) into a clear statistical message: in this dataset, certain groups really do carry a heavier burden of addiction and academic impact than others.

12. Make a data tour as GPT chat with data set attached just like one in class

I designed the following GPT metaprompt to create an interactive “data tour” through my social media addiction dataset. The idea is that GPT acts as a guide: it shows each question, computes the answer in R, and pauses for clarification before moving on. Below is the full metaprompt I would use with the attached CSV file:

Make an interactive tour through the following list of prompts below, using the attached dataset "Students_Social_Media_Addiction.csv".

“

Behavior rules:

- After each answer have the explain clearly the meaning of result and the purpose of that result

- Always show the current prompt text to the user FIRST (in quotes).

- After showing the prompt, run the corresponding analysis in R on the attached dataset and show the result

(numbers, tables, and plots as appropriate).

- After each prompt is answered, ask the user: "Any questions so far?"

- If the user says "proceed" or "continue", move to the next prompt on the list.

- If the user asks a follow-up question, answer it, then ask again "Any questions so far?" and only move on

when they say "proceed" / "continue".

- Never skip a prompt. Always go in order.

Assume the dataset has at least the following raw variables:

- Addicted_Score, Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, Mental_Health_Score,

Conflicts_Over_Social_Media, Age

- Gender, Academic_Level, Country, Most_Used_Platform, Relationship_Status, Affects_Academic_Performance

Assume the data frame is called social.

Now walk through the following prompts in sequence:

---

1. Dataset Overview, Variable Types, and New Variables

Prompt 1:

"Load the dataset 'Students_Social_Media_Addiction.csv' into an R data frame called social, show the first few rows with head(), and display the structure with str(). How many rows are there?"

Prompt 2:

"Convert the following variables to factors: Gender, Academic_Level, Country, Most_Used_Platform, Affects_Academic_Performance, and Relationship_Status. Then use sapply() to list which variables are numeric and which are categorical (factors)."

Prompt 3:

"Using summary(), summarize the key numeric variables: Age, Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, and Addicted_Score. Report at least the mean, min, max, and quartiles for each."

Prompt 4:

"Using table(), report the counts for Gender, Academic_Level, Most_Used_Platform, and Affects_Academic_Performance. Comment on whether the sample is balanced by gender and how many students say social media affects their academic performance."

Prompt 5:

"Using subset(), create two subsets: (1) undergraduates only (Academic_Level == 'Undergraduate'), and (2) students with Addicted_Score >= 8 (high addiction). For each subset, show how many students there are and the mean Addicted_Score."

Prompt 6:

"Using tapply(), compute (1) the mean Addicted_Score for each Academic_Level, and (2) the mean Avg_Daily_Usage_Hours for each Most_Used_Platform. Which academic level and which platforms have the highest averages?"

Prompt 7:

"Create the engineered variables in the same order as in my project: (1) Addiction_Level (Low for Addicted_Score <= 4, Moderate for 5–7, High for >= 8), (2) High_Addiction (1 if Addicted_Score >= 8, else 0), (3) Sleep_Deprived (1 if Sleep_Hours_Per_Night < 7, else 0), (4) Heavy_User (1 if Avg_Daily_Usage_Hours >= 5, else 0), (5) Usage_to_Sleep_Ratio = Avg_Daily_Usage_Hours / Sleep_Hours_Per_Night, (6) Affects_Binary (1 if Affects_Academic_Performance == 'Yes', else 0), and (7) High_Risk_Study_Impact (1 if High_Addiction == 1 and Affects_Academic_Performance == 'Yes', else 0). For each new variable, show table() or summary() so we can see its distribution."

Prompt 8:

"Summarize the dataset: how many students are there in total, how many numeric variables, and how many categorical (factor) variables after all transformations?"

---

2. Exploratory Data Analysis: Plots

Prompt 9:

"Draw a histogram of Addicted_Score with a reasonable number of bins. Describe the overall shape and how common high addiction scores are."

Prompt 10:

"Create a boxplot of Addicted_Score by Academic_Level. Which academic level appears to have the highest median and the widest spread of addiction scores?"

Prompt 11:

"Make a barplot of the counts for Most_Used_Platform. Highlight in your description where TikTok, Instagram, and WhatsApp stand in terms of popularity."

Prompt 12:

"Create a scatterplot of Avg_Daily_Usage_Hours versus Sleep_Hours_Per_Night with a fitted regression line. Comment on the direction of the relationship and mention that there are also a few students who use social media relatively little but still sleep very little."

---

3. Fooled by Data: Randomness and Permutation

Prompt 13:

"Compare the mean Addicted_Score for male vs female students. Then run a permutation test for the difference in means (female minus male) with at least 5000

permutations. Report the observed difference, the permutation p-value, and whether this gap could easily arise by random chance."

Prompt 14:

"Now focus only on students whose Most_Used_Platform is TikTok or Instagram. Compute the observed difference in mean Addicted_Score (TikTok minus Instagram). Then run a permutation test with at least 5000 permutations using shuffled platform labels. Show the distribution of simulated differences, the observed difference as a vertical line, and the permutation p-value."

---

4. Central Limit Theorem & Confidence Intervals

Prompt 15:

"For Instagram users, repeatedly draw random samples of size n = 50 with replacement from Avg_Daily_Usage_Hours at least 1000 times. For each sample, compute the sample mean and plot the histogram of these sample means. Use this to illustrate the Central Limit Theorem and describe how the distribution of means looks more bell-shaped than the raw data."

Prompt 16:

"Using a z-based interval, compute the 95% confidence interval for the mean Avg_Daily_Usage_Hours among Instagram users. Show the sample size, sample mean, sample standard deviation, standard error, z-score (1.96), margin of error, and the final interval."

Prompt 17:

"Recompute the margin of error and confidence interval for Instagram usage at 90% confidence ($z \approx 1.645$) and at 50% confidence ($z \approx 0.674$). Compare how the margin of error shrinks as the confidence level decreases."

---

5. Hypothesis Testing: Mean Differences (Undergrad vs Grad)

Prompt 18:

"Using a one-sided z-test, test the hypothesis that Undergraduate students have a higher mean Addicted_Score than Graduate students. Report the sample means for each group, the observed difference, the z statistic, the p-value, and your conclusion at alpha = 0.05."

Prompt 19:

"Run a one-sided permutation test for the same hypothesis (mean Undergrad > mean Grad) using at least 10000 permutations of the Academic_Level labels. Report the permutation p-value and comment on whether it agrees with the z-test conclusion."

---

6. Independence & Difference of Proportions (Chi-Square)

Prompt 20:

"Construct a contingency table of Academic_Level vs Affects_Academic_Performance (Yes/No) and run a chi-square test of independence. Show the observed counts, expected counts, chi-square statistic, p-value, and a one-sentence interpretation about whether academic level and reported academic impact are related."

Prompt 21:

"Construct a 2x2 table of High_Addiction (0/1) vs Affects_Academic_Performance (Yes/No). Compute the proportion in each cell and run a chi-square test. How strong is the association between being highly addicted and reporting that social media affects academic performance?"

Prompt 22:

"Create a contingency table of Most_Used_Platform vs High_Addiction (0/1) and run a chi-square test of independence. Which platforms appear to have especially high proportions of highly addicted students?"

---

7. Multiple Hypothesis Testing & Bonferroni

Prompt 23:

"Compute the mean Addicted_Score for each Most_Used_Platform using tapply(). If there are 12 platforms, compute the total number of possible platform pairs (12 * 11 / 2 = 66) and the Bonferroni-corrected alpha = 0.05 / 66."

Prompt 24:

"Run permutation tests comparing TikTok to at least five other platforms (for example Instagram, Facebook, YouTube, WhatsApp, Twitter), each time testing whether the mean Addicted_Score for TikTok is higher than for the other platform. Report the p-values for each comparison."

Prompt 25:

"Using the Bonferroni-corrected alpha from before, mark which of the TikTok-vs-other-platform comparisons are significant before correction ($p < 0.05$) and which remain significant after correction ($p <$ alpha_Bonferroni). Summarize which platform gaps are strong enough to survive multiple-testing correction."

---

8. Bayesian Reasoning: Updating Beliefs

Prompt 26:

"Compute the prior probability and prior odds that a randomly chosen student reports Affects_Academic_Performance == 'Yes'."

Prompt 27:

"Now update this belief for TikTok users. Compute: (1) the probability that a student uses TikTok given Affects == 'Yes', (2) the probability that a student uses TikTok given Affects == 'No', (3) the likelihood ratio, (4) the posterior odds and posterior probability that a TikTok-primary student reports academic impact. Explain in plain language how much TikTok usage shifts your belief."

---

9. Prediction Model: Decision Tree

Prompt 28:

"Create a factor target variable High_Addiction_F with levels 'No' (High_Addiction == 0) and 'Yes' (High_Addiction == 1). Build a decision tree using rpart() to predict High_Addiction_F from Gender, Academic_Level, Most_Used_Platform, Avg_Daily_Usage_Hours, Sleep_Hours_Per_Night, Sleep_Deprived, Heavy_User, Usage_to_Sleep_Ratio, and Affects_Academic_Performance. Use a 70/30 train-test split with set.seed(123)."

Prompt 29:

"Report the training and testing accuracy of the decision tree, then plot the tree. Describe which variables appear near the root (for example Usage_to_Sleep_Ratio or Heavy_User) and how they help split high vs non-high addiction."

---

10. Association Rules & Lift

Prompt 30:

"Treat Heavy_User, Sleep_Deprived, TikTok_User (Most_Used_Platform == 'TikTok'), WhatsApp_User (Most_Used_Platform == 'WhatsApp'), High_Addiction, and Affects_Binary as binary items. Compute support, confidence, and lift for the rule {Heavy_User, TikTok_User} -> {High_Addiction}, and interpret what the lift means in terms of risk compared to an average student."

Prompt 31:

"Compute support, confidence, and lift for the rule {High_Addiction} -> {Affects_Binary == 1}. How much more likely is it that a highly addicted student reports academic impact compared to a randomly chosen student?"

End of prompt list.

Remember: for each prompt, show the text first, then run it, show the result, and ask 'Any questions so far?' before moving on.

“

Link of demo version for meta prompt: Link