Final Project - Data Science

Yonatan Schwartz & Gali Shaany

Introduction

Social media has become an integral part of students' daily lives, offering platforms for communication, entertainment, and information sharing. However, excessive use can lead to problematic behaviors, potentially impacting academic performance, sleep quality, and mental health. This work aims to characterize user profiles on social media among students and to examine whether these profiles remain consistent throughout the questionnaire, as well as whether there is any tendency toward self-assessment bias.

The Student Social Media & Relationships dataset contains anonymized records of students' social media behaviors and related life outcomes. The Dataset presents self-evaluation of students who use social media platforms. Yet, behind these self-reported figures and subjective evaluations lies an intriguing puzzle: can we truly trust students to accurately portray their digital habits, or are there hidden patterns suggesting discrepancies between perceived and actual behavior? By diving deeper into the underlying structure of this data, this study not only seeks to reveal the nuances of online self-perception but also explores whether subtle biases or distortions emerge when users reflect on their own social media use. The findings may surprise you, raising critical questions about how students perceive themselves and which survey questions contradict others.

Data

Structure of the Dataset

We obtained this dataset from Kaggle, which comprises 705 subjects (students) and 13 variables, each representing different aspects of a student's demographic background, social media habits, and psychological well-being. The Data was collected via a one-time online survey administered on early 2025.

We recommend downloading the dataset from our GitHub page to get the Jupiter notebook up and running.

Data Cleaning

Several cleaning steps were taken to prepare the dataset for analysis:

- Categorical variables were ordered by our constraints.
- When necessary, variables exceeding a certain threshold were filtered out to establish a reliable basis for the results.

Table 1- Variables and their type. The main variables are **bolted**.

Group	Vairable	Type	Explanation
Basic student information	Age	Integer	Subjects' age
	Student ID	Integer	Students given an ID from the survey
	Gender	Categorical	Male / Female
	Education	Categorical	High school, Undergraduate, graduated
	Country	Categorical	Subjects' Country of origin
	Relationship Status	Categorical	Single, In Relationship, Complicated
Usage-related variables	**Avg Daily Usage Hour**s	Float	Average hours per day on social media
	Most Used Platform	Categorical	Instagram, Facebook, TikTok, etc.
Behavioral and impact indicators	**Sleep Hours Per Night**	Float	Average nightly sleep hours
	Academic Performance	Boolean	Self-reported impact on academics (Yes/No)
	Addiction Level	Integer	Social Media Addiction Score (1 = low to 10 = high)
	Conflicts Over Social Media	Integer	Number of relationship conflicts due to social media
	Mental Health Score	Integer	Self-rated mental health (1 = poor to 10 = excellent)

These main variables serve as the basis for understanding behavioral patterns and potential signs of social media addiction.

EDA- Exploratory Data Analysis

To remain concise and present only data directly related to the research questions, we will include Only a selected portion of the EDA is performed in the [projects' Jupiter notebook](#).

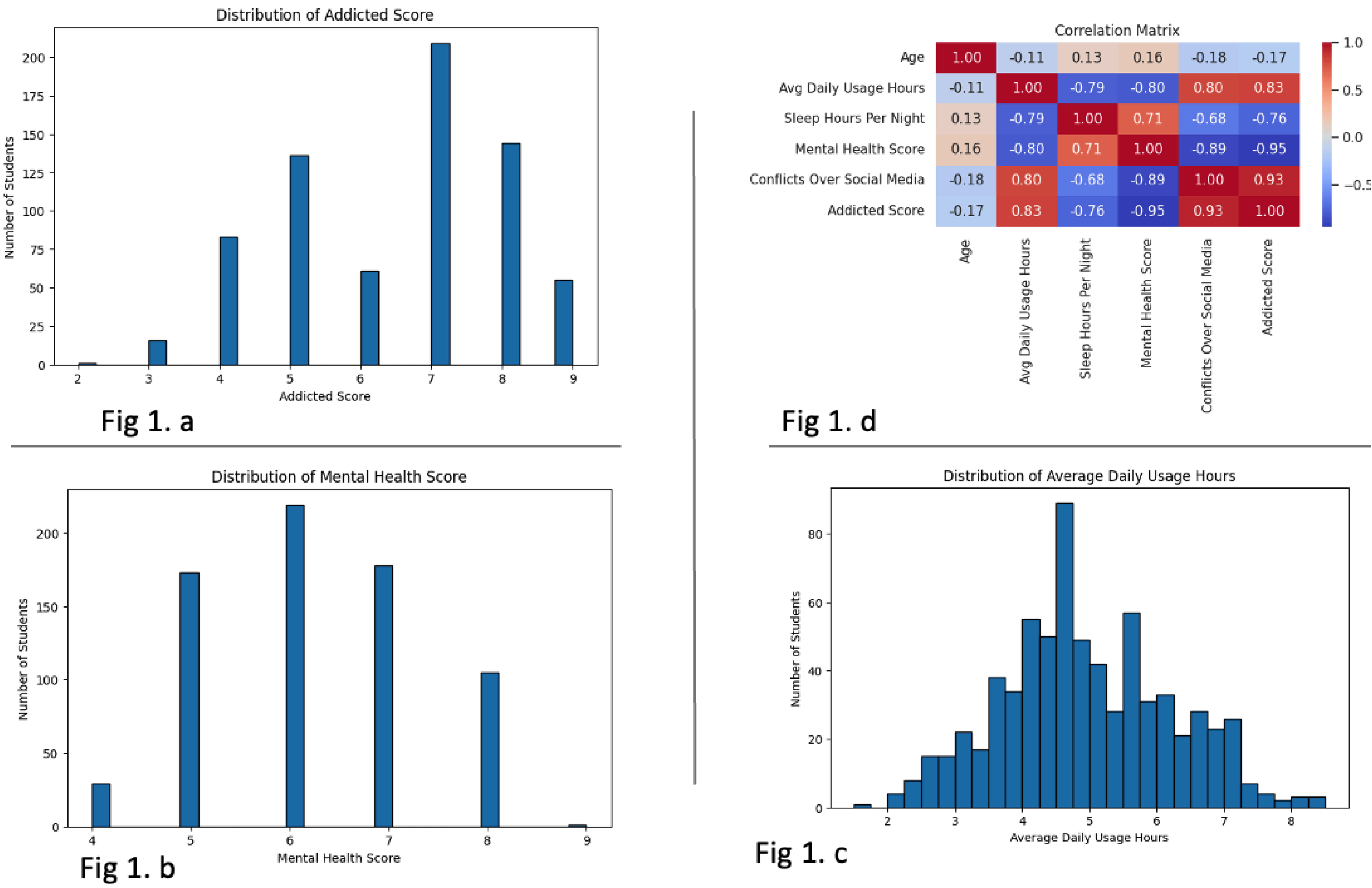

Fig.1 - Distributions of selected EDA outputs

By examining and reviewing the EDA, we were able to think critically and generate new questions and hypotheses.

Initially, we aimed to identify which variables best explain the addiction score and whether there are hidden relationships among the predictors.

Later, we considered the fact that the dataset is subjective in nature, which led us to question whether we could detect inconsistencies or a tendency among respondents to misjudge themselves.

Questions

1. **Do personal characteristics such as age, gender, sleep duration, and relationship status influence the social media addiction score?**

Analysis Q1

This question aims to examine whether and how the students' personal circumstances influence their addiction score. Variables such as age, gender, sleep duration, and relationship status may have an impact on social media use and addiction.
To answer this question, we applied multiple linear regression with the Addicted score as the dependent variable and Age, Gender, Sleep hours per night, and relationship status as independent variables.

Results Q1

The results of the multiple linear regression analysis indicated an R^2 value of 0.601, suggesting that approximately 60% of the variance in social media addiction scores can be explained by the selected variables. Age, sleep hours per night, and relationship status were identified as statistically significant predictors ($p < 0.05$), with sleep hours per night emerging as the most influential factor, demonstrating a strong negative association (coefficient = -1.088).
The Lasso regression analysis further supported these findings, confirming sleep hours per night as the primary predictor of social media addiction. In contrast, the influences of age, gender, and relationship status were entirely eliminated after regularization.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Addicted Score   R-squared:                       0.601
Model:                           OLS   Adj. R-squared:                  0.599
Method:                Least Squares   F-statistic:                     263.4
Date:               Sat, 26 Jul 2025   Prob (F-statistic):          5.35e-138
Time:                       14:27:39   Log-Likelihood:                -1001.8
No. Observations:                705   AIC:                             2014.
Df Residuals:                    700   BIC:                             2036.
Df Model:                          4
Covariance Type:           nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                    15.2852      0.652     23.443      0.000      14.005      16.565
Age                      -0.0888      0.031     -2.824      0.005      -0.151      -0.027
Gender_encoded            0.0878      0.087      1.007      0.314      -0.083       0.259
Sleep Hours Per Night    -1.0881      0.034    -31.647      0.000      -1.156      -1.021
Relationship_encoded      0.2781      0.066      4.223      0.000       0.149       0.407
==============================================================================
Omnibus:                      10.094   Durbin-Watson:                   2.201
Prob(Omnibus):                 0.006   Jarque-Bera (JB):                9.900
Skew:                         -0.260   Prob(JB):                      0.00708
Kurtosis:                      2.740   Cond. No.                         378.
==============================================================================
```

Figure 1- Multiple linear regression summary for question number 1.

Conclusion Q1

Sufficient sleep has been shown to significantly reduce the risk of social media addiction among students. Consequently, future interventions should prioritize strategies aimed at improving students' sleeping habits.

2. **Does health status, comprising mental health, sleep duration, interpersonal conflicts, and academic impact, affect the addiction score?**

Analysis Q2

This question inquires whether a person's health status affects their social media addiction score. We conducted a multiple linear regression using the following four predictors:

- **Mental Health Score** (higher = better mental health)
- **Sleep Hours Per Night**
- **Conflicts Over Social Media** (1 = Yes, 0 = No)
- **Affects Academic Performance** (1 = Yes, 0 = No)

All predictors were standardized or numerically encoded where needed. The dependent variable was the Addiction Score.

Results Q2

The regression model is significant and explained 95.3% of the variance in addiction scores (Adjusted $R^2 = 0.953$), indicating a good model fit. All four predictors were statistically significant ($p < 0.001$)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Q("Addicted Score")   R-squared:                       0.953
Model:                             OLS   Adj. R-squared:                  0.953
Method:                  Least Squares   F-statistic:                 1.419e+04
Date:                 Sat, 19 Jul 2025   Prob (F-statistic):               0.00
Time:                         15:29:47   Log-Likelihood:                -249.19
No. Observations:                  705   AIC:                             502.4
Df Residuals:                      703   BIC:                             511.5
Df Model:                            1
Covariance Type:             nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        6.4369      0.013    495.316      0.000       6.411       6.462
Health_Index     0.9999      0.008    119.130      0.000       0.983       1.016
==============================================================================
Omnibus:                       61.734   Durbin-Watson:                   2.104
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              139.628
Skew:                           0.500   Prob(JB):                     4.79e-31
Kurtosis:                       4.938   Cond. No.                         1.55
==============================================================================
```

Fig 2- Multiple linear regression summary for question number 2

Conclusion Q2

Health status significantly affects the addiction score.
Specifically, worse mental health, less sleep, conflicts caused by social media, and academic disruption are all associated with higher levels of addiction.
The strongest predictor was mental health, with a substantial negative relationship:
Students with lower mental health scores showed meaningfully higher levels of social media addiction.
These findings suggest that interventions targeting mental well-being, sleep hygiene, and mitigating the impact of social media on academics and relationships may help reduce problematic social media use.

Side question- Does gender and academic level influence students' addiction scores and mental health status?

Table 2: Patterns of Addiction and Mental Health Across Educational Stages and Gender

Academic Level	Gender	Addiction Score	Mental Health Score
High School	Male & Female	Very high (around 8), no noticeable gender difference	Centered around 5, low variation, no noticeable gender difference
Undergraduate	Male	Slightly higher than females, a tendency toward higher addiction	Slightly higher than females, indicating better reported mental health
	Female	Slightly lower than males	Wider spread, more variability in mental health
Graduate	Male & Female	Lower than undergraduate and high school; no strong gender gap	Higher and more varied scores, suggesting better mental health

The data reveals that academic level strongly influences both addiction and mental health scores. High school students exhibit the highest addiction levels and the lowest mental health scores, with minimal gender differences. Undergraduate students show intermediate addiction levels, where males tend to have slightly higher addiction and better reported mental health compared to females. Graduate students display the lowest addiction scores and the highest, most varied mental health scores, with no significant gender gap. Overall, gender differences are subtle and mainly observable at the undergraduate level.

3. Main question: Are respondents consistent in their answers throughout the questionnaire, and is there any tendency toward self-assessment bias?

Analysis Q3

The main question helped us assess whether respondents were consistent throughout the questionnaire and to identify which variables may reflect a tendency toward inaccurate self-assessment. In this section, we defined three student profiles, each focusing on the reliability of a different variable.

- Profile 1 assesses the respondents' reliability regarding their perception of the impact of social media on their **academic performance**.
- Profile 2 examines the reliability of the **addiction score** as reported by the respondents in relation to other variables.
- Profile 3 assesses the reliability of the respondents' self-reported **mental health score** in relation to other variables.

The complete profiles' components can be found in the project's Jupiter notebook.

For each profile, we selected variables that represent and influence the examined variable. Thresholds were determined based on the average values of the respective variables. It is important to note that the choice of thresholds carries significant weight and may bias the results.

In addition, we examined the relationship between average daily usage and addiction score using a box plot to determine whether there is a correlation between them and to observe the range of self-reported values. This is done to assess if subjects with similar usage hours tend to score themselves in the same range.

Results Q3

Table 3: Results of contradictions in the profiles

	Underestimate		Overestimate	
Profile	Subjects	Presecntage	Subjects	Presecntage
1- Academic	0	0	39	5.53
2- Addiction	4	0.57	0	0
3- Mental Health	0	0	0	0

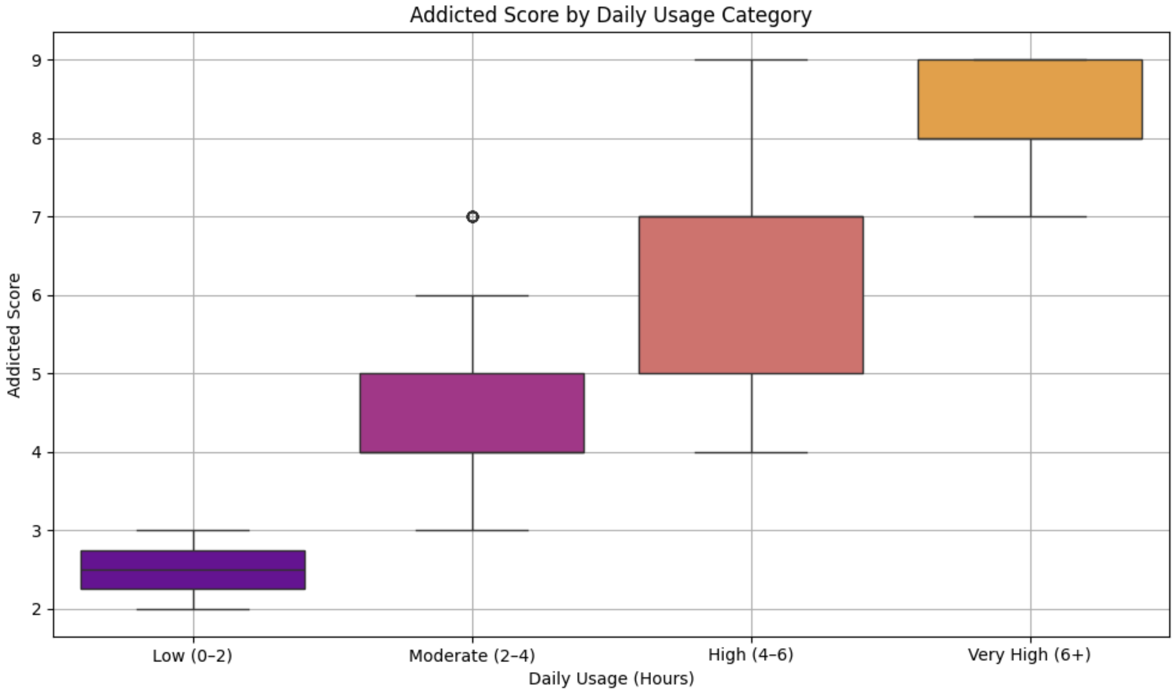

Fig 3- Addicted score by Daily usage boxplot graph

Based on table 3, 5.53% of respondents in the academic profile exhibited overestimation, while no cases of underestimation were found. In the addiction profile, 0.57% of respondents showed underestimation, with no cases of overestimation. No contradictions were identified in the mental health profile.

The boxplot graph demonstrates the distribution of self-reported addiction scores across different categories of daily usage hours. The median addiction score increases with higher usage categories, and the variability within each group is visually apparent. These results provide an overview of the patterns of self-assessment and the relationship between usage and reported addiction levels within the dataset.

Conclusion Q3

Based on the two analyses we conducted—a check for logical contradiction in respondents' answers throughout the questionnaire and a box plot illustrating the relationship between daily usage and addiction scores. It can be concluded that, in front of our test, participants were mostly consistent in their responses. There were not many exceptional cases in which respondents tended to overstate or understate their situation. Moreover, it cannot be stated with certainty that the structured profiles necessarily explain the limited contradictions observed. There may be additional factors, not captured by the questionnaire, that influence academic performance, addiction score, and mental health.

Discussion

Our findings reveal clear patterns, but also highlight several limitations of the data used. First, the regression models explain a large share of the variation in addiction scores, approximately 60% when considering basic personal factors and more than 95% when including health-related variables. Within those models, sleep hours and mental health rating stand out as the strongest predictors, whereas age, gender, and relationship status play a more minor role. Even so, we must emphasize again that every key variable in this project is self-reported. Students who feel tired or stressed may also rate their social-media use more harshly, which could make the relationships appear stronger than they are.

Second, the internal consistency checks revealed only a small number of contradictions (less than 6% in any profile). This suggests that most respondents answered the survey in a consistent manner, yet it does not prove that their answers are accurate. Because the study is cross-sectional and relies on a single questionnaire, we cannot determine whether heavy social-media use causes lower mental health or a decrease in other variables.

Finally, our sample comprises 705 students from a single online survey conducted in early 2025. This limits our ability to generalize the results to other populations or periods. Future research should combine self-reports with objective measures, such as app-usage logs, wearable sleep trackers, and follow-up surveys conducted over several months. Such data would help confirm whether the patterns we observe here hold and clarify the direction of cause and effect. Until then, the results point to practical steps, especially encouraging better sleep habits and supporting mental well-being, that universities can start testing right away, while keeping in mind the limits of the evidence

Conclusion

In sum, the analysis paints a consistent yet preliminary picture: students who sleep less and report lower mental health also exhibit markedly higher social-media addiction scores, while demographic factors play a comparatively minor role. The small fraction of internal contradictions suggests that respondents were largely consistent; however, the exclusive reliance on self-reported, cross-sectional data tempers confidence in both causality and accuracy. Until passive usage logs, objective sleep metrics, and follow-ups are incorporated, our results should be viewed as directional signals rather than definitive proof. Even so, they point toward practical interventions, such as longer sleep and mental-health support, that universities can pilot immediately while collecting richer evidence to validate and refine these initial insights.