Project Proposal 1

Team

Our team brings together a diverse set of skills in data science, programming, analysis, and visualization, ensuring a well-rounded approach to this project. **Caden Schoonveld** specializes in Python, SQL, R, and data visualization, contributing to data processing, analysis, and clear presentation of insights. **Matthew Sparkman** brings expertise in full-stack web development, with experience in SQL, database design, and data visualization, allowing for effective data analysis and a structured presentation of findings. **Court Swenson** has a strong background in data analysis, reporting, and programming, ensuring accurate interpretation and execution of technical aspects. **Destiny Voyles-Perez**, a financial analyst, excels in data cleaning, statistical analysis, and visualization, playing a key role in ensuring data accuracy and generating insightful visual representations. **Haluh Saleh**, a data science student, brings proficiency in Python, R, Excel, and Tableau, supporting data manipulation, exploratory analysis, and modeling. **Sandeep Kang** (Team Leader) is a data science student with extensive experience in Python, R, and strategic problem-solving from his background in the pharmaceutical industry. As team leader, he will oversee project coordination and research design, ensuring a structured and effective workflow. With complementary strengths in data science, analytics, and visualization, our team is well-equipped to analyze social media trends and their impact on mental health effectively.

Project Introduction: Social Media and Mental Health

Social media has become an integral part of modern life, shaping how individuals communicate, work, and perceive themselves. While it provides numerous benefits, such as enhanced connectivity and access to information, its impact on mental health remains widely debated. Some individuals experience increased stress, anxiety, and decreased self-esteem due to excessive social media use, while others find support and self-expression through these platforms. This project examines the relationship between social media usage patterns and mental health indicators, using a dataset that includes variables such as time spent on social media, engagement habits, self-comparison, distraction levels, sleep issues, and feelings of depression. By analyzing survey responses, we aim to uncover how different forms of social media engagement, particularly purposeless or passive use, correlate with mental health outcomes such as anxiety, depression, distraction, and sleep quality. Given the rising concerns over social media addiction and its psychological effects, this research is both timely and essential.

Motivation: Practical and Theoretical Importance

From a practical standpoint, mental health professionals, educators, and policymakers need reliable data to assess the risks and benefits of social media use. This study provides valuable insights into social media habits, such as excessive or purposeless use, the emotional effects of self-comparison and validation-seeking, and cognitive and behavioral impacts, including sleep disruption and concentration difficulties. By addressing these aspects, our research can inform strategies that promote healthier digital habits and mental well-being.

From a theoretical perspective, while existing studies explore the psychological effects of social media, key questions remain unanswered. For example, how do different patterns of social media use lead to varying mental health outcomes? How does social media distraction relate to concentration difficulties and sleep disturbances? What role does self-comparison play in shaping self-esteem and emotional well-being? By analyzing self-reported survey data, this study contributes a data-driven perspective to the ongoing discussion on social media's psychological impact.

Identified Problem and Research Gap

Despite increasing awareness of social media's mental health risks, several research gaps remain. Many studies focus on overall screen time without distinguishing between purposeful engagement and mindless scrolling. Additionally, few integrate multiple mental health factors—such as self-comparison, validation-seeking, distraction, and sleep issues—into a comprehensive analysis.

A key limitation in existing research is the reliance on highly correlated independent variables to predict mental health outcomes. While these studies confirm known relationships, they fail to uncover causal links or new social media impact patterns. More quantitative, survey-based approaches are needed to assess specific behaviors—such as checking social media without purpose or feeling restless when not using social media—and their connection to depression, anxiety, and fluctuating interest in daily activities.

This study addresses these gaps by analyzing self-reported engagement habits, emotional responses, and well-being. Using statistical and machine learning techniques, we aim to identify meaningful patterns that improve mental health risk prediction.

Research Question

To guide our study, we propose the following research question:

How do different patterns of social media usage—especially time spent and purposeless use—impact mental health indicators such as anxiety, depression, attention span, and sleep quality?

By answering this question, we aim to contribute meaningful insights that help individuals, mental health professionals, and policymakers develop strategies to foster a healthier relationship with social media.

The Literature Review

This review examines three scholarly articles that explore the impact of social media on mental health through different methodological approaches. These methods include surveys from self-reporting, leveraging machine learning, and large-scale data analysis of trends. Understanding how these different methodologies analyze social media's impact on mental health is increasingly necessary for developing reliable insights into these unseen patterns.

One approach to the issue is the application of machine learning techniques to predict mental health outcomes through user engagement on social media sites. Purohit et al. (2023) in their study *Analyzing the Impact of Social Media Usage on Mental Health: A Machine Learning Approach* analyzed a dataset of 481 people by applying logistic regression and Gaussian Naïve Bayes, with more than 94% accuracy to find the correlations between social media usage and symptoms of ADHD, Anxiety, Depression, and Self Esteem issues. Their findings indicate that certain online behaviors serve as indicators of mental health struggles.

Yunus (2022) in *Understanding the Link Between Social Media Use and Mental Health Issues* utilizes a behavioral approach that looks at the existing factors and trends leading to mental health problems, such as social comparison, fear of missing out (FOMO), and cyberbullying. Yunus highlights adolescents, young adults and those with pre-existing mental health conditions as the demographics most impacted by negative interactions online. Research increasingly supports that active engagement on social media correlates to higher risk of depression because of the potential for negative feedback and self-comparison. The research also supports a positive correlation between the amount of time spent on social media and ADHD symptoms, poor sleep patterns and an increase in depressive symptoms. This research uses the framing of theoretical methods and builds a foundation for developing policy and guidelines to help address these problems.

A third study *Assessing Social Media's Impact on Mental Health: A Gaussian Naive Bayes* Approach (2024) used multiple models (linear regression, Gaussian Naive Bayes, random forest, etc.) to evaluate the same dataset as Purohit et al. The visualizations in this study offer compelling evidence supporting the findings of the other studies. These three studies show the evolving methodologies in researching social media's impact on mental health. While machine learning models allow for individual level predictions, large-scale data analysis can show population level insights, and behavioral psychology research can identify the underlying reasoning behind these outcomes. A combination of these approaches would serve to best develop a comprehensive understanding of the impact social media has, while also addressing any shortcomings of each approach.

Data Resources

The dataset utilized in this study was sourced from Kaggle (https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health/data?select=smmh.csv), a reputable platform widely used in the fields of data science and machine learning. Specifically, the dataset is titled "Social Media and Mental Health" and was created by Souvik Ahmed, with contributions from Muhesena Nasiha Syeda. According to the metadata and author notes, the survey was conducted in Dhaka, the capital of Bangladesh—an important detail that provides cultural and demographic context for interpreting the results.

The data was gathered through a structured, self-reported survey that focused on indicators of mental health and patterns of social media use. It consists of 481 entries, with only one column containing 30 missing values, reflecting high data integrity. The survey employed quantitative Likert-scale questions addressing areas such as distraction, depression, sleep disturbances, validation-seeking behavior, and self-comparison.

Originally compiled for a university-level project on statistics and machine learning, the dataset was designed to support predictive analysis of mental health outcomes based on social media behavior. In addition to being well-organized and clearly labeled, it has been cited in other academic studies related to machine learning and behavioral psychology, highlighting its credibility and relevance. These qualities make it a robust and valuable resource for analyzing the relationship between social media usage and mental health.

Description of the Data

Preliminary Data Analysis

The dataset exhibits a well-organized structure, comprising numeric, categorical, and timestamp data types. The numeric variables span thirteen columns, each containing responses to Likert-scale questions rated from 1 to 5. These variables assess key aspects of mental health, including worry, restlessness, distraction, depression, and sleep disturbances. The categorical variables capture demographic and behavioral characteristics such as gender, occupation status, relationship status, average time spent on social media, and most frequently used platforms. Additionally, one timestamp column records the exact date and time of each survey submission.

Missing Values

In terms of data completeness, the dataset demonstrates excellent integrity, with no missing values across all 481 observations. Although an initial inspection suggested that the column concerning organizational affiliation might contain missing entries, further analysis confirmed that all variables are fully populated. As a result, no data imputation or row deletion was necessary, allowing the full dataset to be retained for analysis.

Data Cleaning Steps

Several essential cleaning steps were undertaken to prepare the data for statistical analysis. Column names were reformatted to enhance clarity and consistency by removing punctuation and spacing, making them compatible with programming environments such as R or Python. Categorical variables, particularly gender, were reviewed for inconsistencies or redundancies. Although a diverse range of gender identities was recorded, the analysis was simplified by focusing on the two most common categories—male and female—for basic inferential purposes. Likert-scale responses were retained in numeric form to support the calculation of means, standard deviations, and other descriptive metrics. The timestamp column was preserved in its original format but was not central to the present phase of analysis.

Descriptive Statistics and Inference

The descriptive statistics revealed key patterns in participant demographics and mental health responses. The mean age of respondents was approximately 26.1 years, with a standard deviation of 9.9, indicating a sample skewed toward younger individuals. Nearly all participants reported daily use of social media, with platforms like Facebook, Instagram, and YouTube cited most frequently. A substantial portion of respondents indicated spending more than five hours per day on these platforms.

Mental health indicators showed moderate levels of concern across the sample. The average score for worry was 3.56, for depression 3.26, and for sleep disturbances 3.20. These values, accompanied by standard deviations ranging from 1.2 to 1.3, suggest a fairly consistent psychological impact among participants, with moderate variability in responses.

Statistical Inference

To assess whether gender was associated with different levels of reported depression, a Welch Two-Sample *t*-test was conducted using R Programming. The results indicated a statistically significant difference, with a *p*-value of 0.0102. Female respondents reported a higher mean depression score (3.39) compared to males (3.08). Given that the *p*-value is below the conventional significance level of 0.05, we reject the null hypothesis and infer that gender may influence reported levels of depression within the sample. While this test does not account for potential confounding variables or involve more complex modeling, it provides initial evidence

that demographic factors like gender could play a meaningful role in psychological outcomes related to social media use.
Beyond group comparisons, we also performed regression analysis using time spent and purposeless use as predictors of mental health metrics. Results indicated small but significant effects, particularly highlighting purposeless use as a strong predictor of depression and distraction.

Hypothesis and Goals

Time spent on social media is associated with various mental health indicators. We hypothesize that both the amount of time and how we use social media adversely affect our mental health outcomes.

There are several factors that may influence this relationship, such as age, gender, and employment status. Additionally, the self-reported nature of the dataset introduces potential biases individuals may underreport or overestimate their social media usage and mental health symptoms.

With this study, our goal is to determine whether usage patterns are more strongly associated with negative mental health outcomes. With this dataset, we will use time spent on social media and age to determine potential correlations with reported mental health outcomes. If a clear relationship exists, it could help inform discussions on digital well-being and platform design. Conversely, if no strong correlation is found, this may suggest that external factors beyond social media play a more significant role in shaping mental health outcomes.

Research Design and Methods

This study used a quantitative approach to investigate the relationship between social media use and mental health, leveraging a structured survey dataset containing responses on psychological well-being, engagement habits, and demographics. We performed a comprehensive data transformation process to prepare the data for analysis. Key variables like time spent on social media, originally ordinal, were converted into both categorical labels for boxplots and numeric values for regression and correlation modeling. Participants who reported not using social media were assigned a distinct “Not Use SM” label to differentiate them in the visualizations.

Likert-scale responses for anxiety, depression, distraction, sleep issues, and purposeless use were retained as numeric values to support statistical modeling. Demographic variables such as age, gender, and employment status were encoded appropriately to allow for inclusion in extended regression models. We then applied linear regression models—both simple (with two predictors) and extended (with demographic controls)—to examine the associations between usage patterns and mental health indicators, supported by visualizations such as histograms, boxplots, scatterplots with trendlines, and a correlation heatmap to contextualize the findings.

Visualizations and Analysis

The analytical process was conducted entirely in R, incorporating both descriptive and inferential statistical techniques to investigate the relationships between social media engagement and mental health outcomes. Visualization played a central role in communicating patterns, distributions, and model-based relationships observed in the data.

Boxplots were used to explore the variation in depression and anxiety scores across different time spent categories. These plots revealed clear upward trends in median scores, indicating that higher durations of social media use correspond to elevated symptoms. Notably, individuals categorized as "Not Use SM" consistently exhibited the lowest median scores for both depression and anxiety, suggesting a protective effect against psychological distress associated with abstaining from social media use.

Figure 1: Boxplot – Depression Score by Time Spent on Social Media

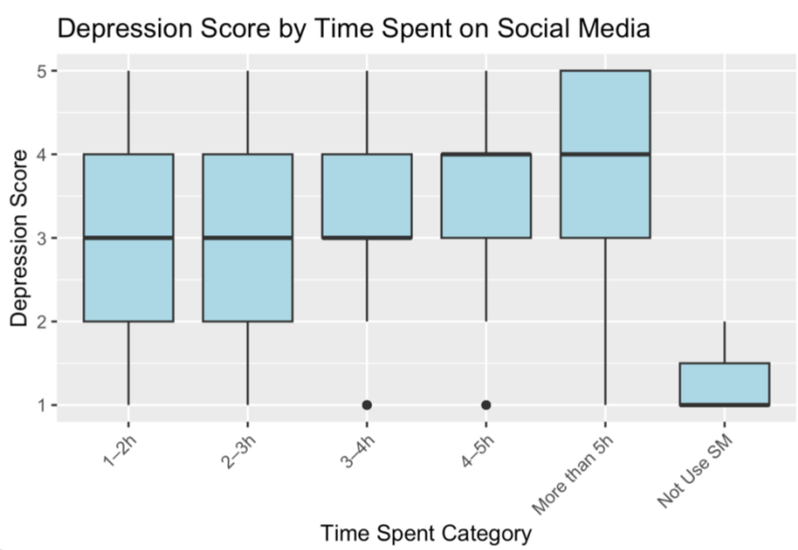
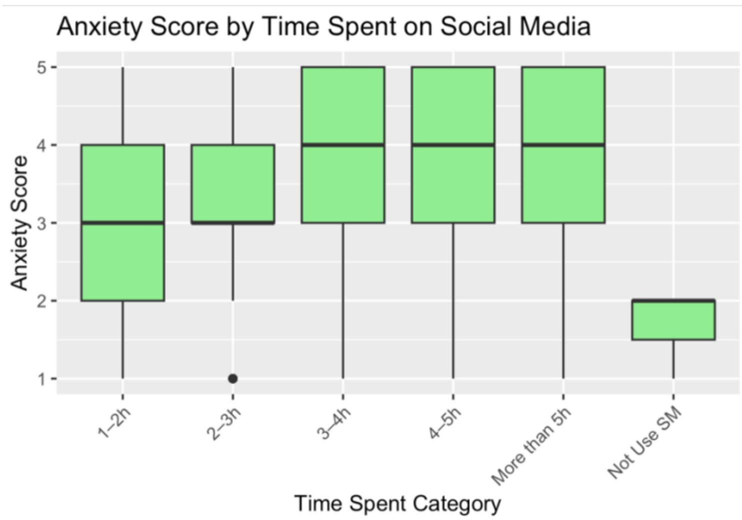


Figure 2: Boxplot – Anxiety Score by Time Spent on Social Media



Histograms provided insight into the overall distribution of both time spent and depression scores. The distribution of depression scores was centered around the mid-to-high range (scores of 3–5), indicating moderate psychological distress across the sample. Time spent on social media skewed heavily toward higher usage categories (2–5+ hours), reflecting the widespread prevalence of extended engagement among users.

Figure 3: Histogram – Distribution of Depression Scores

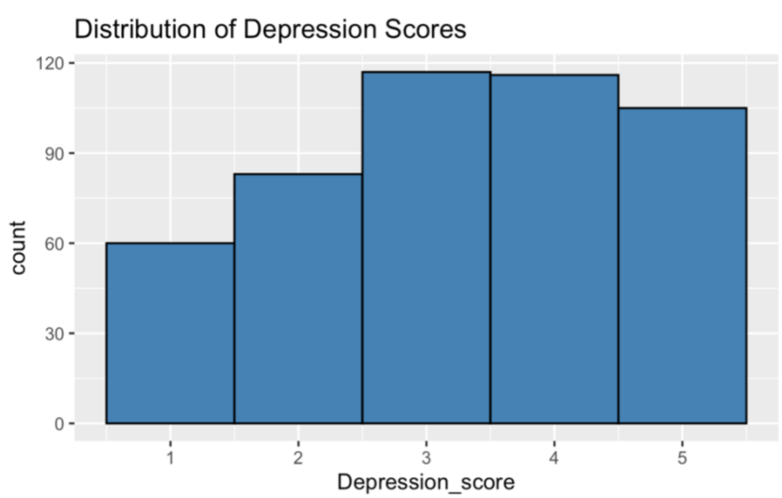
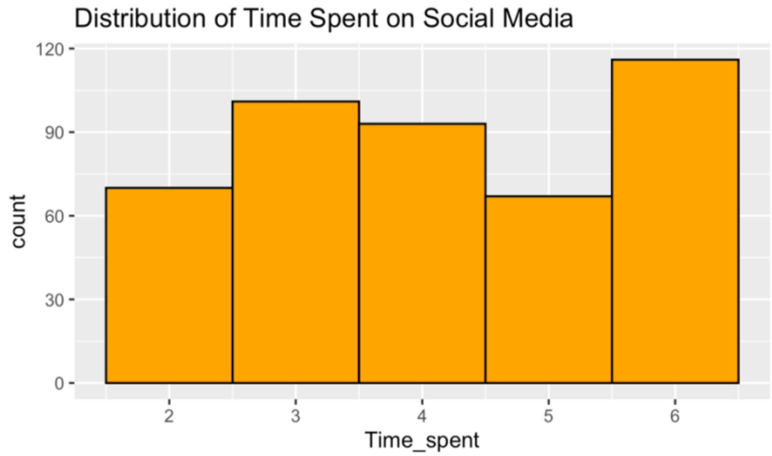


Figure 4: Histogram – Distribution of Time Spent on Social Media



Scatterplots with linear regression lines were used to visualize the continuous relationships between time spent on social media and mental health outcomes such as anxiety and depression. The positive slopes observed in both plots confirm the results of the regression models, showing that increased usage is associated with higher reported symptoms. Although the spread of data points indicates moderate variance, the trend lines reinforce the predictive relationships identified statistically.

Figure 5: Scatterplot – Time Spent vs. Anxiety Score

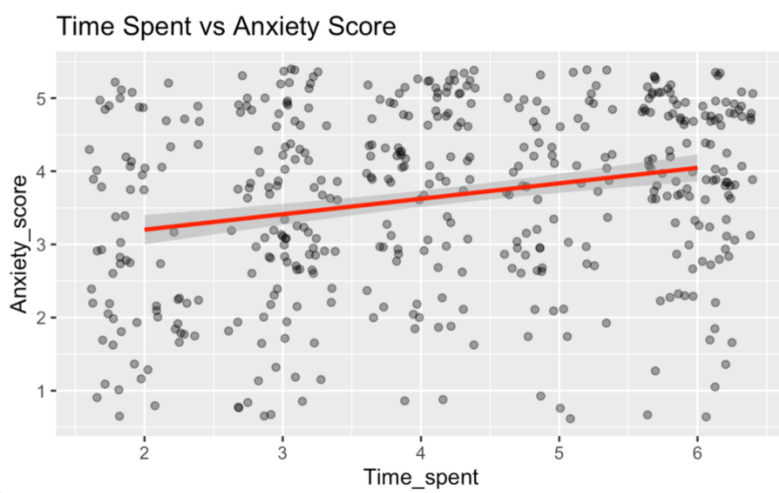
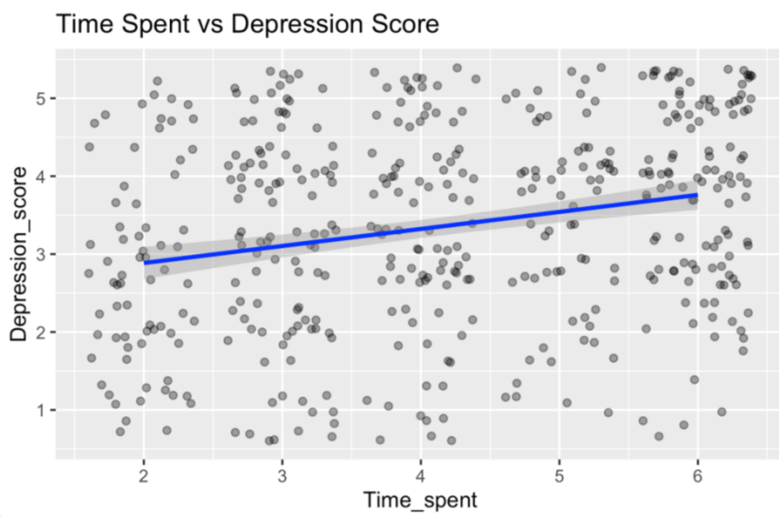
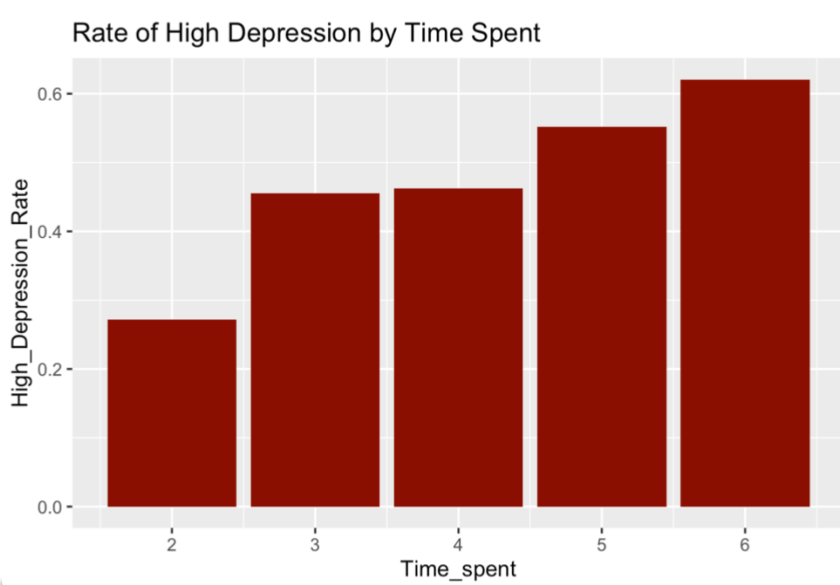


Figure 6: Scatterplot – Time Spent vs. Depression Score



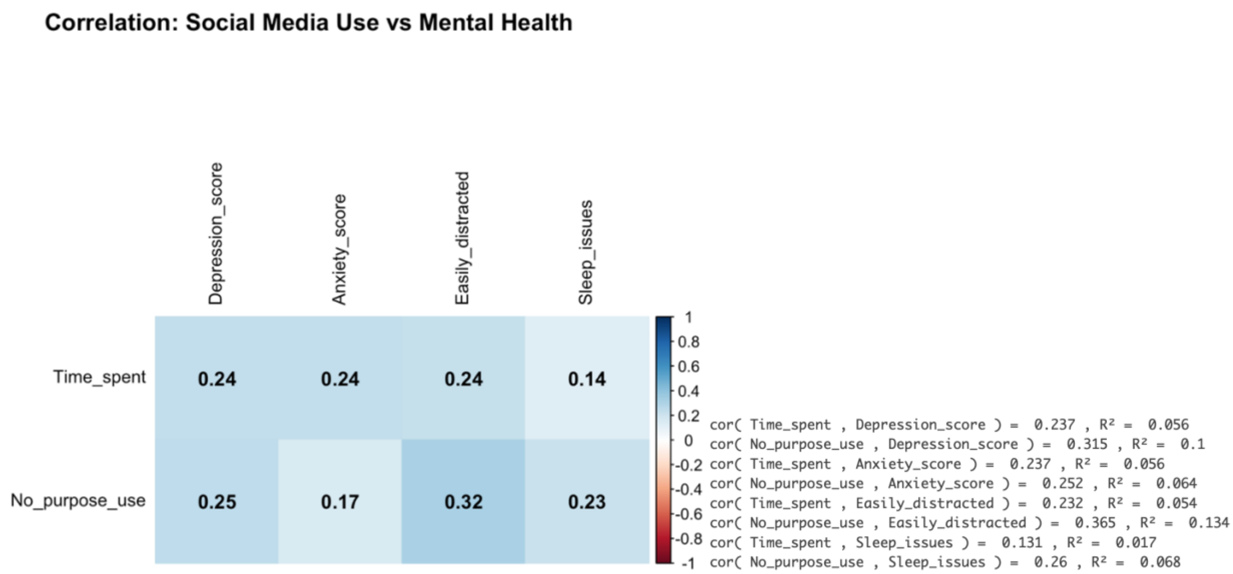
To assess the proportion of individuals experiencing high levels of depression (defined as a depression score of 4 or greater), a bar chart was created. This visualization showed a monotonic increase in the proportion of high-depression cases with greater time spent on social media, climbing from approximately 26% to over 60%. This supports a potential dose-response relationship, where more exposure correlates with greater psychological distress.

Figure 7: Bar Chart – Rate of High Depression by Time Spent



The correlation analysis was visualized using a heatmap of correlation coefficients between key independent variables (time spent and purposeless use) and dependent variables (depression, anxiety, distraction, and sleep issues). The strongest correlation was observed between purposeless use and distraction ($r = 0.365$), followed by purposeless use and depression ($r = 0.315$). Time spent had a weaker but still statistically significant correlation with depression and anxiety ($r \approx 0.24$). These results reinforce the regression outcomes and support the hypothesis that purposeless social media engagement is a stronger predictor of mental health challenges than duration alone.

Figure 8: Heatmap – Correlation Matrix of Social Media Use vs. Mental Health Indicators



Regression modeling confirmed the relationships identified through visualization and correlation. Simple linear regression was used to predict depression, anxiety, distraction, and sleep issue scores from both time spent and purposeless use. In all models, purposeless use was statistically significant, with p-values well below 0.001 in most cases. For example, in the model predicting distraction, purposeless use had the highest coefficient of 0.29 ($p < 0.001$), while time spent had a smaller but still significant coefficient. These findings suggest that passive or unintentional engagement with social media may have a more adverse psychological impact than overall time online.

Figure 9: Model Summary Tables – Regression Output for Depression, Anxiety, Distraction, Sleep

```
$Anxiety

Call:
lm(formula = Anxiety_score ~ Time_spent + No_purpose_use, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1558 -0.7880  0.1448  0.9668  1.9241 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.47457    0.23241  10.647  < 2e-16 ***
Time_spent      0.17804    0.04188   4.251  2.6e-05 ***
No_purpose_use  0.12261    0.05662   2.166   0.0309 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.193 on 442 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.06611,	Adjusted R-squared:  0.06189 
F-statistic: 15.65 on 2 and 442 DF,  p-value: 2.724e-07
```

```
$Depression

Call:
lm(formula = Depression_score ~ Time_spent + No_purpose_use, 
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9790 -0.9790  0.1328  1.0210  2.3736 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.82962    0.23662   7.732 7.19e-14 ***
Time_spent      0.15754    0.04264   3.695 0.000248 ***
No_purpose_use  0.24083    0.05764   4.178 3.54e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.215 on 442 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.09181,	Adjusted R-squared:  0.0877 
F-statistic: 22.34 on 2 and 442 DF,  p-value: 5.71e-10
```

```
$Distraction

Call:
lm(formula = Easily_distracted ~ Time_spent + No_purpose_use, 
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.06756 -0.71620 -0.01504  0.93244  2.33477 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.90474    0.20472   9.304  < 2e-16 ***
Time_spent      0.11712    0.03689   3.175   0.0016 ** 
No_purpose_use  0.29202    0.04987   5.856 9.28e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.051 on 442 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.122,	Adjusted R-squared:  0.118 
F-statistic: 30.71 on 2 and 442 DF,  p-value: 3.242e-13
```

```
$Sleep

Call:
lm(formula = Sleep_issues ~ Time_spent + No_purpose_use, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7667 -1.2778  0.2333  1.2803  2.6292 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.95188    0.27367   7.132 4.06e-12 ***
Time_spent      0.06992    0.04931   1.418    0.157    
No_purpose_use  0.27905    0.06667   4.186 3.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.405 on 442 degrees of freedom
  (36 observations deleted due to missingness)
Multiple R-squared:  0.0546,	Adjusted R-squared:  0.05032 
F-statistic: 12.76 on 2 and 442 DF,  p-value: 4.086e-06
```

Together, these visualizations and models provide consistent and statistically grounded evidence that both time spent and purposeless engagement with social media are associated with elevated mental health symptoms. Visual storytelling combined with empirical modeling enhances the interpretability and persuasiveness of these findings.

In combination, these visual and statistical tools provide strong evidence that social media use—particularly frequent and purposeless engagement, is linked with increased risk of mental health challenges.

Conclusions

This study aimed to explore the relationship between social media usage patterns and mental health indicators, with a focus on depression, anxiety, attention span, and sleep issues. Our proposal was grounded in the growing concern about digital well-being and hypothesized that both the amount of time spent on social media and how individuals engage with it, particularly through purposeless use, would be associated with negative mental health outcomes.

Through the analysis of a self-reported dataset, we identified statistically significant relationships between social media use and several psychological health measures. Notably, purposeless social media use emerged as a stronger predictor of mental health issues than time spent alone, particularly concerning distraction and depression. These findings highlight the importance of evaluating not only the duration of social media engagement but also the quality and intent behind it. Our research design was deliberately chosen to capture these behavioral nuances through Likert-scale measures and targeted usage questions, rather than broad platform usage metrics. This design allowed us to generate deeper insight into user experiences and their psychological effects.

The implications of this study are both practical and theoretical. Practically, the findings support the development of digital interventions aimed at reducing passive or purposeless scrolling behaviors. This could inform social media platform design, mental health education, and personal digital wellness strategies. Theoretically, the results contribute to existing literature by reinforcing the argument that not all social media use is equal—intentionality matters. Methodologically, our study demonstrates the value of using behavioral self-report data and multiple outcome variables to assess mental health holistically.

However, the study is not without limitations. The reliance on self-reported data introduces potential response bias, including underreporting or overreporting of both mental health symptoms and usage habits. The cross-sectional nature of the data also limits our ability to make causal claims. Additionally, some demographic variables and platform-specific usage patterns were not explored in detail, which could offer further insight. Future research could address these limitations by incorporating longitudinal data, experimental methods, or integrating biometric or observational data for greater accuracy and depth.

In summary, this study provides meaningful evidence that social media usage patterns, especially aimless or unstructured use, are modestly but consistently associated with negative mental health outcomes. These insights have value for researchers, clinicians, educators, and designers alike, as we seek to build a healthier relationship between users and the digital platforms that shape daily life.

References

Purohit, S., Mudgal, R., Vats, S., Rana, P., Shah, O. A., & Verma, A. (2024). Analyzing the impact of social media usage on Mental Health: A Machine learning approach. *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/icccnt61001.2024.10724267

Nanditha, S., Joy, H. K., Kounte, M. R., Sridevi, R., Jayarani, A. E. A., & Vanusha, D. (2024). Assessing Social Media's Impact on Mental Health: A Gaussian Naive Bayes Approach. *IEEE*, 791–797. https://doi.org/10.1109/ictacs62700.2024.10840779

Yunus, M. (2023). Understanding the link between social media use and mental health issues. *Journal of Clinical Research and Applied Medicine*, *3*(3), 31–36. https://doi.org/10.5530/jcram.3.3.7

Social media and mental health. (2023, July 18). Kaggle. https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health/data?select=smmh.csv