Vhuhwavho Todani D

INST414

Sprint 2

QSSR track

Data Acquisition & Description

https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health

This dataset which attempts to collect data to research possible correlations between social media usage and mental health is fully accessible. The data in this dataset was gathered through a survey. The survey uses participants from a university and was conducted in 2022 spanning across 1 month (4/18 - 5/22) with 2 more inputs across 6 additional months. During this period a total of 481 responses were received.

Variable inventory:

Name	Data type	Description	Relevance	Missing data percentage
Timestamp	Datetime	Time of response submission		0%
Age ★	int/numerical	Age of respondent	Allows observation of any differences or trends in responses by age	0%
Gender	categorical	Gender of respondant (male,female, Nononbinary)	Allows us to measure any notable difference in reported mental	0%

			health by gender.	
Relationship status	Categorical	Relationship status of respondent (Married,single, in a relationship)	Allows observing if there are notable gaps between relationship status and responses.	0%
Social Media usage	Categorical	Asks if respondent uses social media (yes/no)	Could enable observation of differences between people who do and don't use social media often.	0%
Average daily social media usage ★	categorical	Gives range of hours user uses social media daily	Allows us to scale how much time is spent on social media and how it might correlate with other variables.	0%
Social media distraction★	int/numerical	Scale of respondents distraction by social media level from 1 to 5		0%

How easily are you distracted?	int/numerical	Scale of how easily respondent feels distracted from 1 to 5		0%
Do you find it difficult to concentrate on things?	int/numerical	Scale on how hard respondent finds it to concentrate from 1 to 5		0%
How often do you compare yourself to others on social media?★	int/numerical	Scale how often respondents compares themselves to others from 1 to 5	Comparison is often sighted as an influence on mental health.	0%
How often does your interest in daily activities fluctuate?	int/numerical	Scales how often user feels changes in interest in daily activities from 1 to 5	Changes in daily activity interest could also be a possible marker for mental health differences.	0%
How often do you feel depressed or down?★	int/numerical	Scales how often respondent feels depressed	This allows us to get a read on a respondents mental health	0%

		from 1 to 5	and see if there are any interesting relationships.	
How often do you experience issues with sleep?	int/numerical	Scale how often respondent has sleep issues from 1 to 5		0%

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/1WWCA5
This secondary data source will allow us to observe correlations between average screen time and whether or not a child needed mental health treatment. After cleaning and filtering, the final size of the set was 21,772 rows which is around half the original size

Data Quality Assessment & Cleaning

The data in the dataset has overall functional quality for building observations However it does have areas to be changed/altered to allow easier handling of relevant information. Missing values were present in the columns asking for organizational affiliations. Although this would be an issue in a lot of cases, this column will be removed in the final dataset as its information is not relevant for the purpose of this research.

There is also another issue in the ‘what social media platforms do you commonly use?’ column, it lists multiple values in one column which makes it more tedious to run proper analysis on than if the data was already properly split. This column will also not be present in the final dataset as the research focuses on social media/technology usage without any focus on a specific platform and another column which simply asks if the user uses social media already gives the needed information.

Naming conventions were introduced as the columns for the primary data set were renamed to increase readability as column names in the original set were long since they were formatted as questions. This change also allows handling code to be simpler. For example “Average daily social media usage“ was renamed to "average_daily_usage".

The main change in the secondary data set is a simple filter which only includes entries where the age of an individual is 10 years or older.
Summary:

- Both datasets had their column names renamed to enable easier coding and documentation.
- The primary dataset had organizational affiliation and platform columns dropped.
- The secondary data set had a filter added which reduced the dataset to only include entries where age was 10 or higher
- The “treatment not received column” was dropped from the secondary dataset

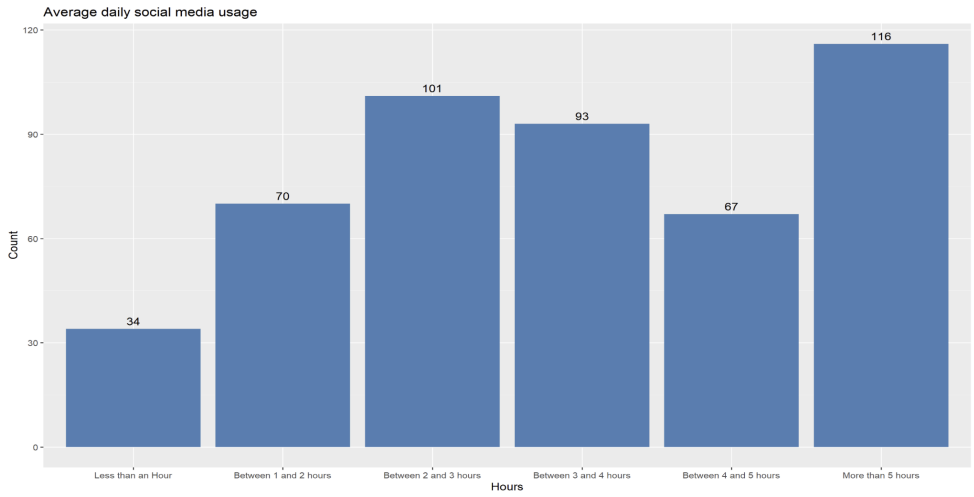
Data cleaning decisions may introduce bias or influence results as things such as filters may not consider the groups of people being removed, for example it could be that adding an age filter ends up disproportionately removing women as the women in the data happen to be younger.

Exploratory Data Analysis

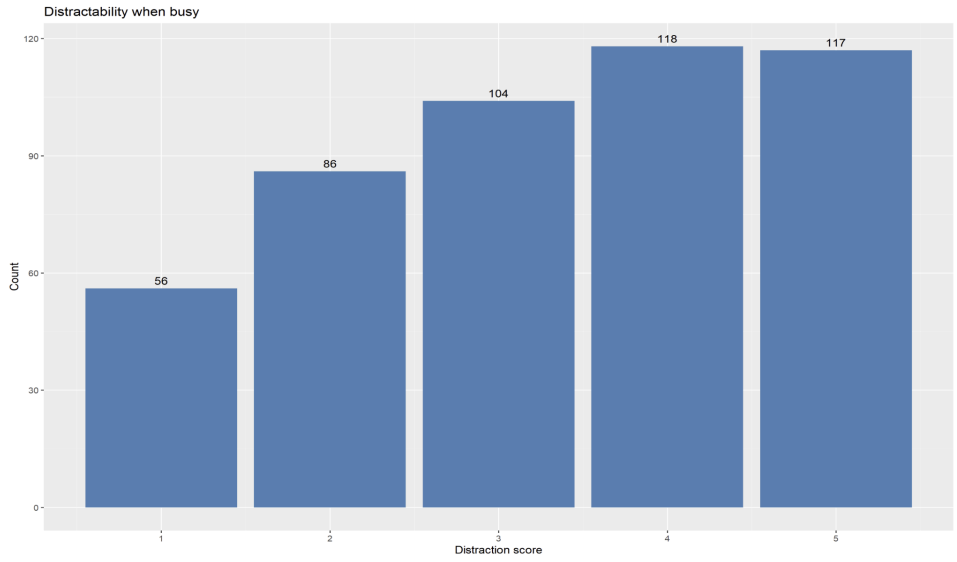
Univariate analysis

There are a total of 481 entries and of those 481, 478 responded yes and only 3 responded no. This indicates that the majority (99.38%) of respondents make use of social media, this falls in line with the assumption that a large portion of people who make use of the internet and digital devices make use of some type of social media platform.

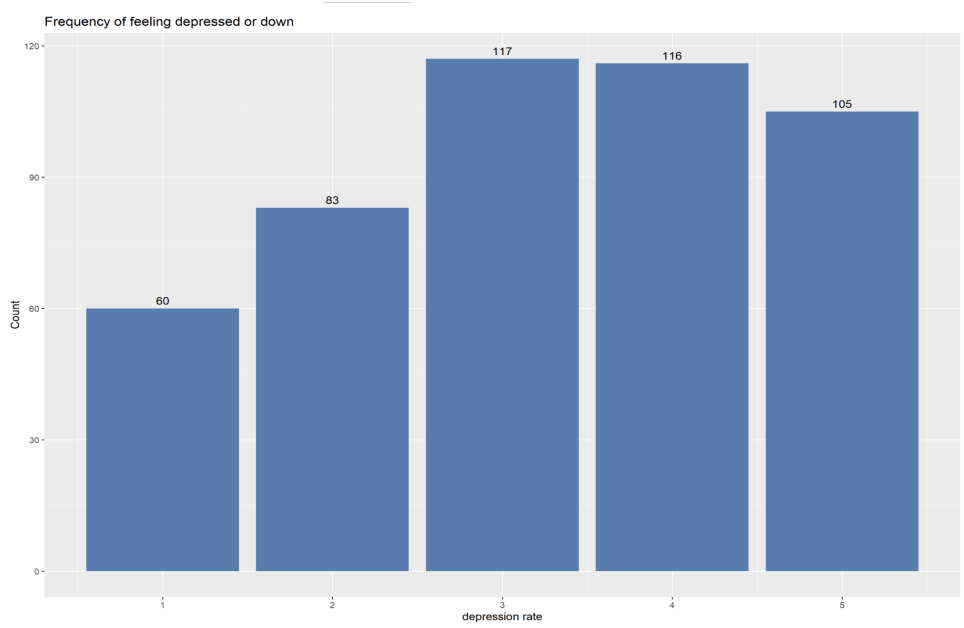
Data on average daily social media usage also indicates that respondents tend to spend a good portion of time on social media. As seen on the visualization, 377 (78%) respondents reported using social media 2 or more hours per day and visually most of the input skews towards the right.



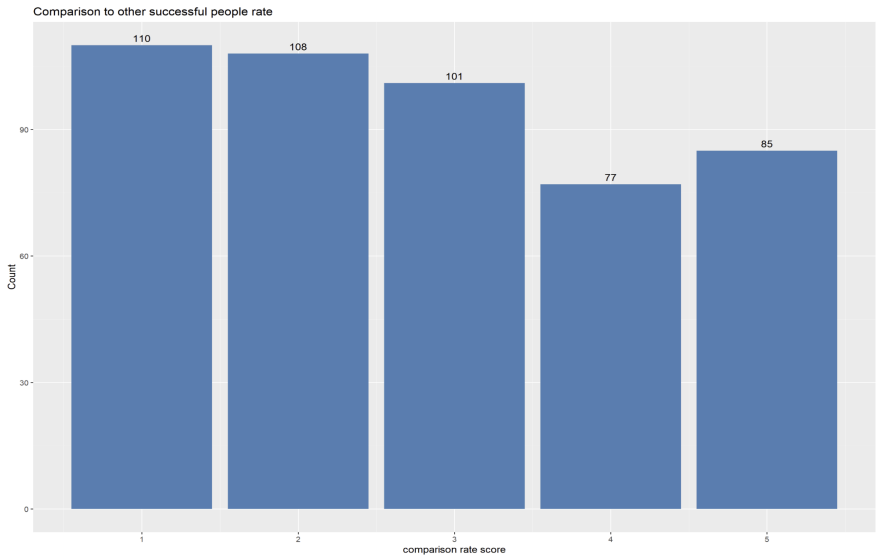
Using a scale of 1 to 5, respondents reported how often they feel distracted by social media when doing something. The graph shows a skew towards the right with a mean score of 3.32. This indicates that on average, social media did function as a notable type of distraction for most respondents.



Respondents also showed notable rates of feeling depressed or down often. It is rated on a scale of 1 to 5 and the graph shows a right skew with a mean of 3.256 and median of 3. 338 (70%) of respondents gave a score of 3 or higher indicating that feeling depressed or down is fairly common.



Respondents showed different trends when rating how often they compare themselves to successful people through social media. The resulting graph skewed to the left with a mean of 2.832



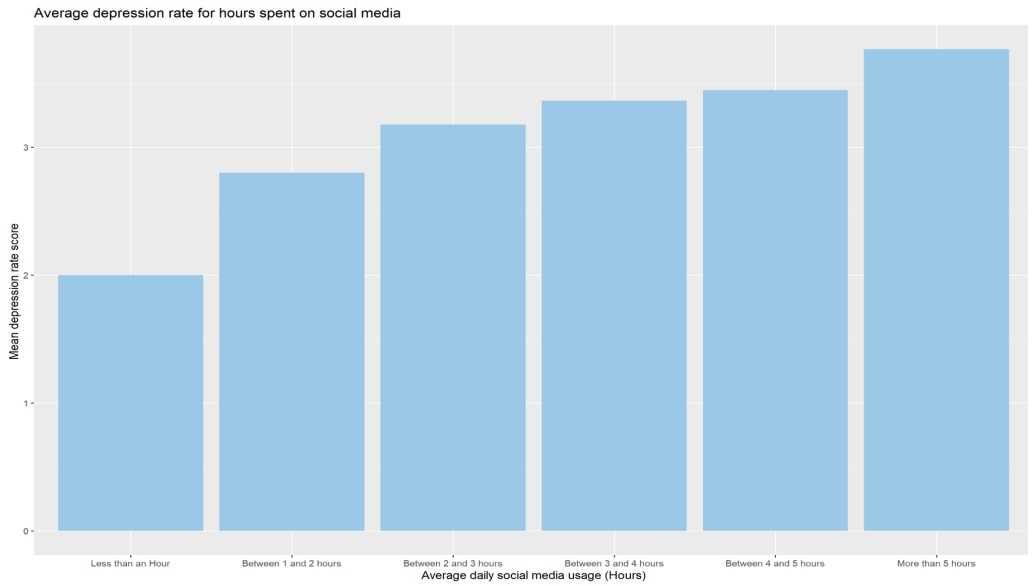
Multivariate analysis

An important area for analysis is if there is any notable correlation or trends between the amount of hours respondents spend on social media and their reported scores. Using an ANOVA test with an alpha level of 0.05, the difference between averages of each social media usage hour range group can be compared.

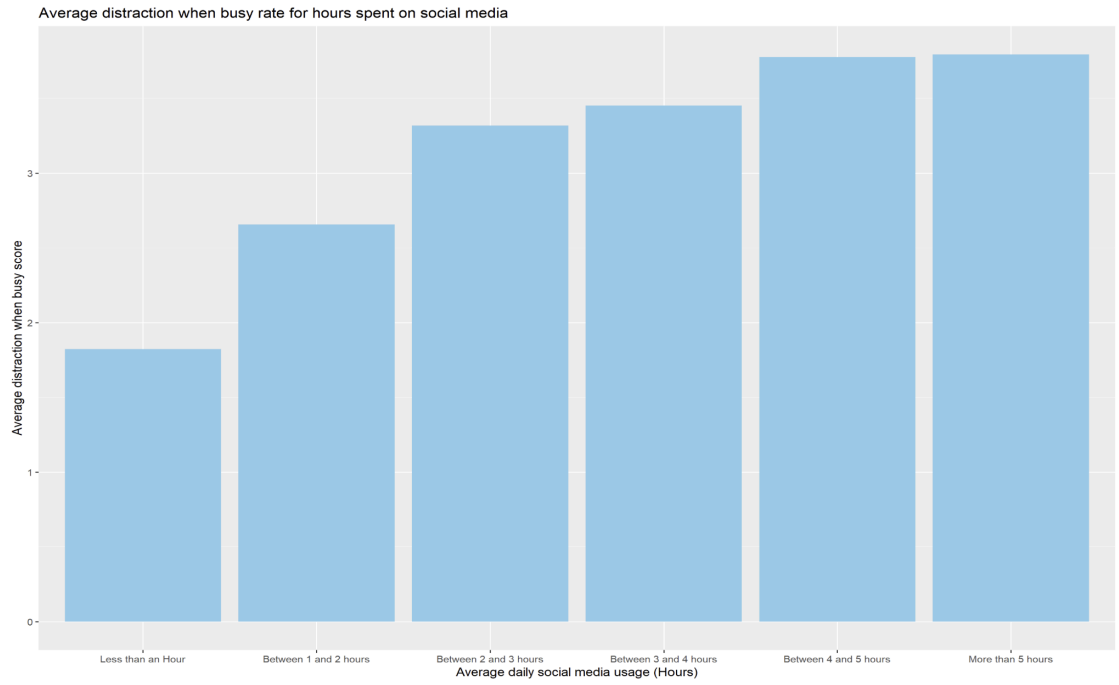
Null Hypothesis: There is no difference between average depression rates between groups.

Alpha hypothesis: There is a difference between average depression rates between groups.

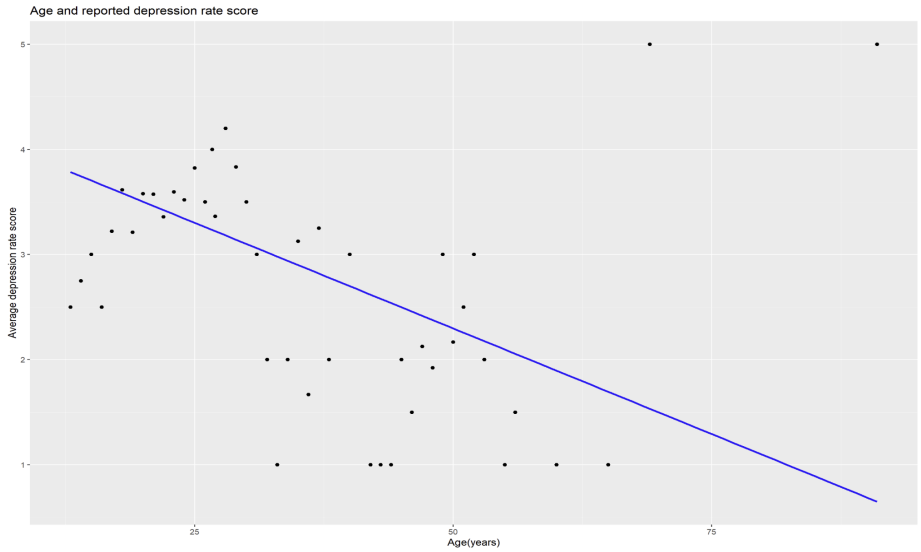
An anova test brings back a p value of 2e-16 which is significant and less than 0.05 so we reject the null hypothesis, indicating that there is a difference in response between groups. This graph also illustrates this as the mean depression score sees an increase as the number of hours on social media increases and is skewed left. This allows us to infer that social media usage does play a role in mood.



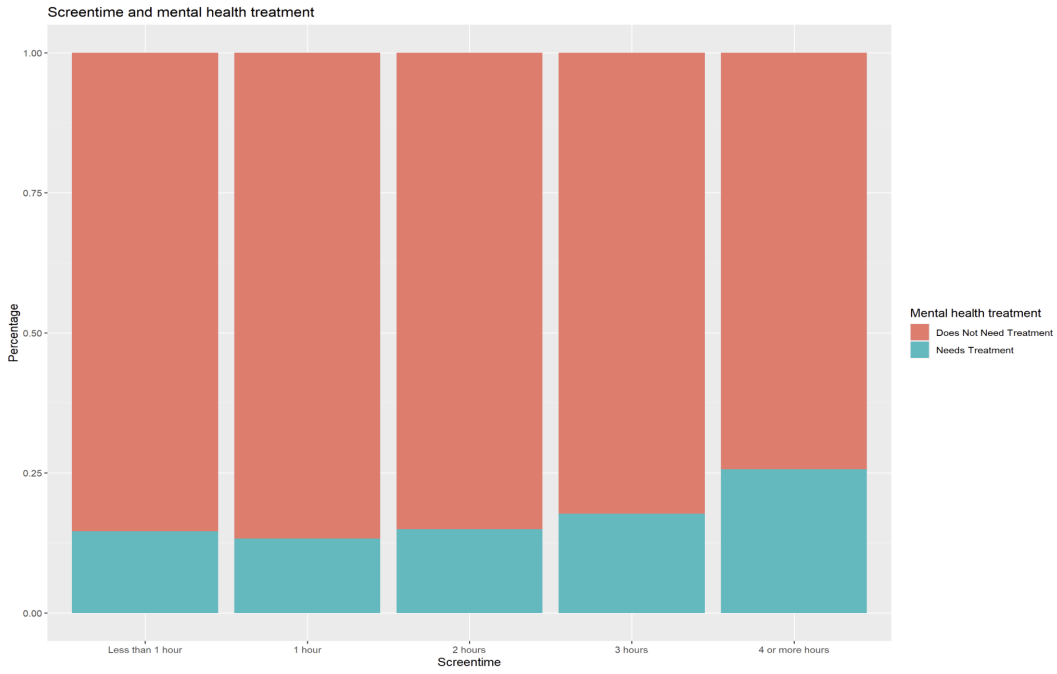
A similar trend is notable when analyzing the rate of distraction when busy for each group, there is a general increase as hours on social media increases, however the rate is very similar after 4 or more hours.



Another interesting aspect of this data is the trends in responses by age. This scatterplot visualizes a negative relationship between age and depression rate score which carries onto other responses as well. This could indicate that younger people might lean more towards making use of social media and being influenced by it more than older groups.



The secondary data set also has information regarding social media hours and the need for mental health treatment. Running a chi square test with an alpha of 0.05, a low p value of 2.2e-16 is returned so we reject the null hypothesis and conclude that there is a difference between need for mental health treatment and screentime hours. Similarly to the primary dataset, this bar plot visualizes a general increase in the proportion of youth who needed mental health treatment in relation to their average screentime.



Surprising findings

The data for average social media hours was higher than I expected as the group with the largest amount of responses was one indicating that they use 5 or more hours of social media a day. Responses in comparison rate were also lower than I expected, social media is often referenced in a negative way to include issues such as constantly comparing yourself to others however results indicated that the data skewed left and the score of 1 out of 5 was the modal group.
The relationship between individuals' age and their responses was also surprising as it showed comparison to not be as notable as you'd expect, older people tending to give lower scores to things like depressed or down scores or being distracted by social media when busy. Though it can be hard to make conclusions based just on this as there are a lot of factors relating to age that could be leading the relationship.

Data limitations

A limitation of these data sets is that it is difficult to get a proper analysis of people who don't use social media, correlations can be found with low hour usage but overall not even 1% of responders indicated not using social media so the pool is too small to try to make any reasonable conclusions. In addition certain other groups having low counts such as gender identities that are not Male/Female make it hard to reliably make conclusions.

Refined Problem Statement & Analytical Plan

After exploring this data, my research question has expanded. Originally it was just focused on how social media/technology impacts mental health but I would also like to introduce the influence of information such as age ,gender and relationship status as well as how technology/social media impacts habits. Ultimately it would introduce thinking about how the respondents demographics differed in response.

In future analytical attemptsI would like to employ similar techniques that were used such as the Anova and chi square test as I want to see if there are any other interesting differences between the reported score for groups based on things like age and gender. After observing the ages, I would also like to place more consideration into the influence of age as some confounders such as income level or individual's position in life could be anchoring these results. For robustness, I will have to consider the response sizes when looking at different groups as the dataset has 500 entries but some groups within that could have very low sample sizes.

The main obstacle was lacking data on people who don't use social media in the primary dataset and a similar case in the secondary dataset. This shifted me to focus more on the differences between each group during sprint 3 and also informed the updated research question and data points to analyze.

Progress Tracking & Next Steps

During this phase, I established the final primary and secondary dataset, conducted uni and multivariate EDA for the main sets of data that relate to the original research questions. Areas of concern were also uncovered and an updated/broader research question was put in place. In sprint 3, and 4 I want to increase analyzing data that deals with the relationship between variables. This will also come with addressing an issue I uncovered later on with the data on gender where non binary was listed in multiple different ways which made unnecessary extra groups which I'll address. I will also solidify if it's worthwhile including when analyzing gender based trends as only a tiny portion Identified as anything other than Male/Female. Another task would be updating visualizations to be more interesting and effectively labeled.