**Project Description**

**Topic:**

Internet usage and its effects on mental health

**Why it matters:**  
The widespread use of online interactions, social media, and digital platforms connects individuals but also exposes them to new stressors. Studying the impact of internet use on mental health is crucial for several reasons. Firstly, it helps identify risks like cyberbullying and social comparison. Secondly, it enables the development of interventions to mitigate negative effects and promote positive online experiences. Lastly, in the rapidly evolving digital landscape, such research allows for adapting mental health support strategies to meet the changing needs of individuals.

**Research Question:**

What are the effects of daily screen time on mental health at an individual level in the United States?

**Motivation:**

In this digitally connected era, marked by an unprecedented surge in screen time, individuals immerse themselves in a myriad of online activities, be it for work, socializing, or leisure. The pervasive nature of this digital shift has fundamentally altered our perceptions and interactions with the world. Notably, COVID-19 with its widespread adoption of remote work and increased reliance on digital platforms for various activities, has accelerated this trend. In the midst of this transformative period in history, it becomes essential to investigate the ramifications of heightened screen time on mental health.

**Description of Datasets:**

The Research and Development Survey (RANDS) is a series of cross-sectional surveys initiated in 2015 using probability-sampled commercial survey panels. These surveys have been integral for methodological research at the National Center for Health Statistics (NCHS), focusing on question-response patterns and statistical methodology for survey estimate calibration. Seven rounds of surveys have been conducted, spanning from 2015 to 2022 and covering various health-related topics. The data from RANDS 1 (2015) and RANDS 2 (2016) are utilized as pre-COVID-19 data in this study[[1]](#footnote-1).

Additionally, Pew Research Center's American Trends Panel (ATP) proved valuable for individual-level data on this subject. The ATP is a nationally representative panel survey initiated in 2014, covering diverse topics including technology use, social trends, and political views. This dataset consists of 102 waves, with relevant waves for this research being wave 72 and wave 93[[2]](#footnote-2).

In summary, RANDS offers targeted health-related data pre-COVID-19, while ATP provides a broader range of data, including internet usage trends during the pandemic. Both datasets contribute to understanding different aspects of individuals' behaviors and conditions, providing valuable insights for research and analysis.

**Data Processing**

**Overview:**

It is important to note that after 2016, RANDS had a change in funding. Because of this, internet usage was no longer measured. If there wasn’t a change in funding, and questions pertaining to internet usage were still present in the survey, I would have selected data closer to the start of COVID-19. With that, a search began for internet usage data elsewhere. Many other sources provided only percentage changes, and most data was only available on a country-level basis, which was insufficient for this research's objective. This made it difficult to combine any COVID-19 data with the RANDS data. Strategies for merging these datasets will be discussed later in this section of the paper.

**Pre-COVID-19 Data:**

Cleaning and manipulating the two pre-COVID datasets were rather straight forward. The two original datasets comprised of over 100 variables. To minimize noise and computation time, the data was subset to contain only the necessary variables. This resulted in using 12 variables. Additionally, a year column was created to distinguish the data between the two datasets. This was done to compare the demographics between the two years to ensure that the populations were similar, making it appropriate to merge the data. Given that both tables had the same columns and response types, the function *rbind()* was used to combine the data.

To make the pre-COVID data comparable to the COVID data, age ranges were created to match the format of the COVID data. The function *cut()* was used to standardize the data. The codebook[[3]](#footnote-3) for the pre-COVID data indicated that participants could report their internet usage as either daily or weekly. To handle this, vectorized operations and indexing were employed to selectively update elements in the 'AWEBOFNO\_N' column based on a condition in the 'AWEBOFNO\_F' column. Specifically, when 'AWEBOFNO\_F' equaled 2, the corresponding values in 'AWEBOFNO\_N' were divided by 7 to standardize the data to a daily level.

If the participants did not answer a question, “.” was recorded as their response. Given that only a small percentage of participants chose to not respond to certain questions, their data was removed from the dataset. This was done by using vectorized operations and logical indexing to filter the rows of the data frame based on two conditions: having complete cases and not containing the character ".". When looking at the demographic boxplots, it can be observed that there are multiple outliers in some of the responses. It was decided to not manipulate or address these values considering there were only a handful of these points, and the dataset was large enough to not be influenced by these values. The final data manipulation involved recoding variables numerically to enhance readability in plots. In retrospect, more efficient approaches, like leveraging functions from 'tidymodels,' such as recipes, could have been employed.

**COVID DATA:**

Cleaning, manipulating, and combining COVID datasets posed significant challenges. The RANDS dataset lacked internet usage reportings due to changes in funding, necessitating the exploration of alternative sources. The Pew Research Center's American Trends Panel (ATP) offered valuable data but sifting through its 102 waves for internet usage information during COVID-19 was a laborious task, ultimately utilizing only two waves (wave 72 and wave 93). These were the only two waves that had pertinent information on internet usage. Unlike pre-COVID data, the ATP datasets recorded internet usage categorically rather than numerically, complicating the comparison between pre-COVID and COVID internet use, a topic I will address later.

Cleaning and manipulating the two RANDS COVID datasets were very similar to how things were conducted with the pre-COVID RANDS datasets. Data was subset only pulling the necessary variables. The addition of the ‘YEAR’ column was added to ensure that when comparing the two datasets that they have similar demographics, ensuring the validity and reliability of my later analyses. Data was also recoded for ease of readability to the reader. Again, more efficient methods could have been used. The function *rbind()* was used to merge the two datasets together, as well as the function *cut()* to create age ranges.

Cleaning and manipulating the two ATP datasets was also straightforward. Only the necessary columns were subset into a new table. The ‘YEAR’ column was added to both tables to again ensure similar demographics. *Rbind()* was used to merge the two datasets. The columns ‘F\_AGECAT’, ‘F\_GENDER’, and ‘F\_EDUCCAT’ were renamed to match the columns in the merged RANDS COVID dataset. The columns for education level and gender were recoded to match the responses in the RANDS dataset (i.e., “A man” was recoded to “Male”).

To amalgamate the ATP merged dataset with the merged RANDS dataset, the *left\_join()* function was utilized, focusing on the columns 'AGE,' 'GENDER,' 'EDUC,' and 'YEAR.' These columns encapsulate crucial information about demographic characteristics and the temporal dimension of the data. The objective was to align records between the ATP and RANDS datasets based on matching demographic and temporal attribute

**Findings**

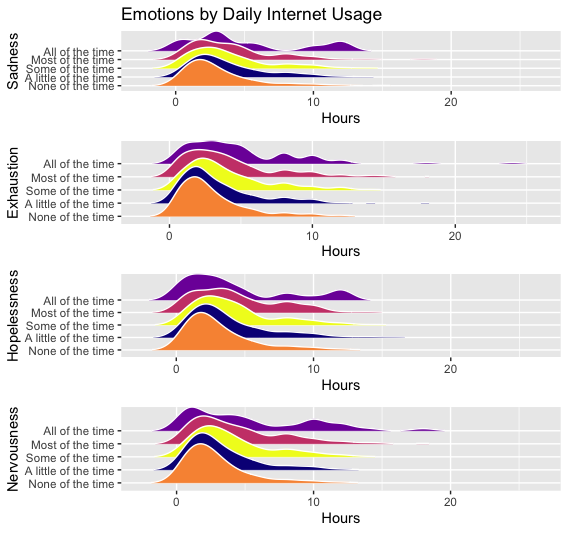
**A graph of different colored bars

Description automatically generatedPre-COVID:**

This plot illustrates the parallels in internet usage patterns between male and female respondents. Notably, full-time female students exhibited the highest interquartile range (IQR) for internet usage, a trend mirrored by their male counterparts with a similar IQR spread. The most discernible difference arises in the median of internet usage between full-time working women and men, with the former displaying a slightly higher median. However, this discrepancy is not significant enough to compromise the validity of the results. Furthermore, the plot reveals consistent patterns between employment types and internet usage, reinforcing the credibility of the data due to the remarkable demographic similarities. It's worth noting the presence of outliers across each employment category. Despite attempting a log transformation, the resulting plot did not exhibit substantial improvements.



This plot highlights the consistency in daily internet usage across age groups and education levels between 2015 and 2016. Notably, participants aged 18-29 exhibited the highest internet usage among all age groups, irrespective of their educational backgrounds. While this observation is noticeable, it is not as drastic as differences observed in other age groups. Overall, there is a uniformity in internet usage time across all education levels, with no single educational category standing out significantly from the others. This further reinforces the credibility of the data.



This plot provides an overview of the correlation between participants' reported emotions and their internet usage. Emotions were assessed on a scale of six levels, with participants responding to questions like 'During the past 30 days, how often did you feel...'. The 'Don't know' response was omitted from the plot due to its insignificance. Notably, the emotions of sadness and exhaustion exhibit a moderate increase with higher levels of internet usage. Conversely, the emotions of hopelessness and nervousness follow a concave distribution, intensifying up to the 'Some of the time' category with increasing internet usage, and subsequently becoming less intense beyond this point.

In conclusion, these diverse plots shed light on various aspects of the relationship between internet usage and demographic factors, emotional experiences, and temporal dynamics. The exploration of internet usage patterns among male and female respondents, as well as across different employment types, not only identifies nuanced trends but also underscores the credibility of the data due to the consistent patterns observed. Likewise, the examination of daily internet usage across age groups and education levels reveals a notable uniformity, strengthening the reliability of the dataset. Finally, the correlation between reported emotions and internet usage provides valuable insights, uncovering intriguing patterns such as the concave distribution of hopelessness and nervousness. Together, these analyses contribute to a comprehensive understanding of the multifaceted interplay between internet usage and various demographic and emotional factors.

**A group of colorful squares

Description automatically generated with medium confidenceCOVID:**

The demographic profiles of the two datasets closely align, demonstrating remarkable similarities. Gender distribution is nearly identical in both datasets, boasting an equitable representation of male and female participants. While the RANDS dataset skews towards a higher count of individuals aged 18-29, the ATP data exhibits a greater presence in the 50-64 age range. Nevertheless, these variations are relatively minor, reinforcing the overall congruence in age distribution. Furthermore, both datasets share comparable educational backgrounds among participants. Given these consistent demographic parallels, the decision to merge the two datasets appears well-founded and substantiated.

**A graph of different colored bars

Description automatically generated**

The outcomes depicted above indicate a lack of substantial impact of internet usage on participants' emotions. Specifically, when examining feelings of being nervous or on edge, the majority of participants reported minimal to no experience of these emotions irrespective of their internet usage. Similarly, for emotions related to depression or hopelessness, most participants did not report such feelings, regardless of their frequency of internet usage. In the case of feeling worried or anxious, the prevalent pattern reveals that participants tended to experience these emotions a few times a year. Notably, a higher frequency of feeling worried or anxious was observed when participants reported almost constant or several times a day internet usage.

In conclusion, the comprehensive examination of demographic profiles and emotional responses within the two datasets illuminates alignment and consistency. The gender distribution, with only minor variations in age representation, underscores the overall congruence in demographic characteristics between the RANDS and ATP datasets. The findings further reveal a nuanced relationship between internet usage and emotional states. Despite limited emotional impact observed on average, certain trends emerged, particularly with heightened feelings of worry or anxiety associated with more frequent internet use. Notably, it's worth highlighting that the increased data usage during the COVID-19 era did not show significant correlations with adverse mental health factors on a broad scale.

**Limitations:**

Analyzing the datasets and their respective visualizations reveals certain limitations that must be acknowledged in the interpretation of the findings. Firstly, the reliance on self-reported data introduces the potential for response bias. Participants may underreport or overreport their internet usage or emotional states, impacting the accuracy of the conclusions drawn.

Furthermore, the age distribution in the datasets may introduce a limitation. The age groupings employed might oversimplify the diverse experiences within each category, potentially masking nuanced trends specific to narrower age ranges. Additionally, the absence of granularity in the education level variable may limit the depth of insights, as participants with diverse educational backgrounds are grouped into broad categories.

A noteworthy methodological difference between the datasets lies in the measurement of internet usage. The pre-COVID data records internet usage numerically, providing a quantitative measure of the extent of participants' online engagement. In contrast, the COVID-19 dataset measures internet usage categorically, assigning participants to distinct usage groups. This shift from numerical to categorical measurement introduces a challenge when attempting to directly compare internet usage patterns before and during the COVID-19 pandemic. The transition between these measurement approaches may impact the ability to draw precise conclusions about changes in internet usage behaviors over time. Additionally, this methodological difference underscores the importance of considering the unique characteristics and limitations of each dataset independently, as the categorization may introduce a level of abstraction that affects the reliability and comparability of the findings.

In summary, a critical consideration of these limitations is essential for a nuanced understanding of the datasets and their visualizations. These constraints highlight the need for caution in generalizing findings and emphasize the importance of contextualizing results within the broader scope of data collection methodologies and potential biases.

**AI Disclosure:**

I used chatGPT to help me create my plots. As a start, I used the website that was recommended in class[[4]](#footnote-4) to figure out where I wanted to take this project. Once I figured out the direction I wanted to go, I used chatGPT to help me with design specifics. For example, with one of the pre-COVID plots, I knew I wanted to create a boxplot displaying internet usage against age and employment type. I honestly had no clue how to execute this idea besides knowing that I wanted to *ggplot2().* I told chatGPT my idea for the boxplot and it spit out code necessary to execute this. Some of the specifics I wanted included created a legend that showed each employment type. I had to tailor the code chatGPT provided to my project, and the process was repeated for all the other graphs.

1. RANDS DATASETS: <https://www.cdc.gov/nchs/rands/r1probsample.htm> [↑](#footnote-ref-1)
2. ATP DATASETS: <https://www.pewresearch.org/american-trends-panel-datasets/> [↑](#footnote-ref-2)
3. <https://www.cdc.gov/nchs/rands/files/RANDS1_codebook.pdf> [↑](#footnote-ref-3)
4. <https://www.data-to-viz.com/> [↑](#footnote-ref-4)