Analyzing the impact of internet usage on depression and anxiety with Python

Adam Sam  
 DATA-6150-01  
 Wentworth Institute of Technology  
 Boston, MA USA  
 sama1@wit.edu

ABSTRACT

This study examines the impact of internet and social media usage on the global prevalence of depression and anxiety. As global internet connectivity rises, questions arise about its correlation with mental health trends, particularly among adolescents. This study explores whether the adoption of internet technology correlates with higher rates of reported depression. Using public datasets from The Institute for Health Metrics and Evaluation (IHME) and Our World in Data (OWID), the analysis employs Python tools like Pandas and NumPy for exploratory data analysis, SciPy-Stats for regression modeling, and MatPlotLib for visualization. Findings reveal a minimally negative correlation between the average global time spent on social media and anxiety prevalence, but a minimally positive correlation with depression. For internet usage, there is a weak positive correlation with depression among youths less than 20 years of age.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*DATA-6150-01, November 2024, Boston, MA USA*

© 2024 Copyright held by the owner/author(s).

KEYWORDS

Internet, social media, depression, anxiety, Python

1 Introduction

The use of the internet and social media has become an integral part of modern life, largely reshaping our communications, work, and culture. However, this transformation in society has raised concerns about its potential effects on mental health, particularly on the impressionable youth. Understanding the underlying causes of depression and anxiety is vital for developing effective prevention strategies.

Due to the relatively recent adoption of internet and social media technologies, research in this topic is not definitive. Studies in this topic suggest a strong link between excessive internet and social media usage and mental health issues, especially among the youth. For instance, an article of the Journal of Abnormal Child Psychology found that acts of social comparison and reassurance seeking performed on social media led to depressive symptoms among adolescents [4]. Another article of the Adolescent Research Review journal found that exposure to beauty ideals propagated through social media among young adults, both male and female, has shown an increase in body dysmorphia [1]. These concerning findings have led healthcare organizations such as the American Psychological Association (APA) and the United States Surgeon General to issue health advisories warning the issues that may arise with youth social media usage [2].

With the purpose of contributing to the topic, this study explores whether the widespread adoption of internet technology and increased social media consumption is associated with rising rates of depression and anxiety both globally, and in the United States. Additionally, individual age groups will be analyzed to determine whether the issue is widespread, or more localized to a particular demographic. By leveraging datasets from the Institute for Health Metrics and Evaluation (IHME) and Our World in Data (OWID), the research analyzes trends in depression prevalence worldwide to answer the proposed questions.

2 Data

To answer the questions presented by this topic the following datasets will be used:

1. Select data from the 2021 Global Burden of Disease study by the IHME that records various measures such as percentage of population with anxiety and major depressive disorders, from 1990 to 2021 [3]
2. Dataset provided by OWID showing the global share of population using the Internet, from 1990 to 2021 [7]
3. Dataset provided by Statista highlighting average time spent on social media in minutes from 2014 to 2024 [6]
4. Select data from the same source of Dataset 1 that exclusively selects the same data for the United States [3]
5. Internet usage distribution across the United States by age group provided by the Pew Research Center from year 2000 to 2024, excluding the years 2017, 2020, and 2022 [5]

2.1 Source of datasets

Datasets 1 and 4 were obtained from the IHME’s website under the condition that it is used for non-profit purposes. The IHME is an independent public research institute based in the University of Washington and trusted as a reputable source. It is a subset of data selected from a larger dataset that compiles findings from the 2021 Global Burden of Disease study.

Dataset 2 comes from data collected by the World Bank that was processed by OWID. The OWID are non-profit organization that selects credible data to present.

Dataset 3 includes data from We Are Social, a social media marketing agency, and is prepared by Statista, a reputable organization specializing in data gathering.

Dataset 5 comes from the Pew Research Center, a reliable, nonprofit, and unbiased fact tank based in Washington D.C.

2.2 Characteristics of the datasets

A screenshot of a computer

Description automatically generated

Figure 1: Excerpt of the datasets; pictured left is anxiety/depressive disorder prevalence; pictured right is internet usage percentage.

All datasets are stored in Excel files. Dataset 1 is split into two sheets – anxiety and depressive disorders prevalence worldwide, and by individual country; pictured in Figure 1 is the individual country variant. Datasets 1 and 2 are joined on the basis of country name, while Dataset 3 is joined solely by the year.

Due to discrepancies in the naming conventions between the datasets, proper data cleaning was performed to ensure the similarity of country names by taking the difference of each named country in the “location” column and renaming accordingly. All data cleaning and pre-processing were performed using Microsoft Excel.

Another caveat with the data formatted was the differing bin sizes between Datasets 4 and 5. Dataset 5 used wider bins for age groups than Dataset 4. Because of this, multiple smaller bins for Dataset 4 had to be combined by averaging their values. Additionally, since the age bins of both datasets do not perfectly align, this unfortunately introduces some uncertainty to the accuracy of the data.

One important action that was taken regarding the dataset was the choice of excluding data set in the years after the 2020 COVID-19 pandemic while modeling the data. This period saw an unexpected spike in global depression and anxiety rates. Since this event happened recently, it was decided that it was appropriate to omit these years to benefit the accuracy of the model.

3 Methodology

This section outlines the methods and models employed to analyze the data and interpret the results. Each step of the methodology is explained in the following sections, including the rationale for choosing these specific tools and adjustments made to enhance the results.

3.1 Cleaning and interpreting data in Excel

The initial step involved ensuring that the field names across all datasets were properly aligned using Microsoft. The assumption behind this process is that properly structured datasets with consistent field names minimize errors during subsequent data analysis. Corrections were performed manually, with the assistance of Excel formulas – specifically using the “not equals” operator to check differences between cells. This was especially useful when comparing country names, since each dataset had slightly differing names for countries. This step was chosen to establish a readable and uniform foundation for the remaining data analysis process.

3.2 Importing data into Python

The cleaned datasets were then imported into the Python notebook using the read\_excel() function from the Pandas library. Pandas provides an intuitive and efficient way to load structured data from Excel files into data frames. Pandas was selected as a library solely for this reasoning.

3.3 Preparing data frames using Pandas joining

To form specialized views of the data frames, data manipulation operations such as slicing, masking, and joining were employed from the Pandas library. The join() function in particular was used to join tables based on similar field names to construct a unified dataset for analysis, as seen in Figure 2.

A table of numbers and letters

Description automatically generated

Figure 2: A combined data frame generated using Pandas.

3.4 Fitting model using SciPy

To effectively analyze trends within the data, a linear regression model was employed using the SciPy library to analyze the relationship between variables, specifically the depression and anxiety prevalence variables analyzed with the internet connectivity and social media minutes spent variables. The scipy.stats.linregress() function returns a model, along with critical statistics such as p-values and r-squared values, which are essential for evaluating model significance and fit.

A linear regression was chosen for its simplicity and ability to provide detailed insights related to its significance (e.g., p-values, r-squared). This however, comes with disadvantages – one being it’s incompatibility with non-linear data, which is discussed in the Results section of this paper.

3.5 Data visualization with MatPlotLib

The MatPlotLib library was used to visualize the data frames and model outcomes. This library was chosen for its flexibility and ability to integrate seamlessly with Pandas and SciPy outputs. Scatter plots with regression lines were created to enhance the interpretability of results. Hexagonal heatmaps were supplementally used in situations where a scatterplot has too many data points to visualize effectively. Bar plots were utilized to demonstrate focused comparisons, such as a comparison between a selected list of countries.

4 Results

The analysis examined the relationship between internet usage, social media activity, and mental health outcomes globally. The results provided slight insights into the correlations between these factors and mental health conditions such as anxiety and depression which are presented in the following sections.

4.1 Correlation with internet usage

A screen shot of a graph

Description automatically generated

Figure 3: Global anxiety prevalence and internet usage. The points represent years (1990 - 2019). The red line is the linear regression model.

A graph of a graph with red dots

Description automatically generated

Figure 4: Global anxiety prevalence and internet usage among teenagers aged 15-19. The y-axis represents the share of population within the age group with anxiety.

The findings revealed no significant impact of internet usage on anxiety prevalence among youths globally across the years. Globally (Figure 3), the correlation found for all standardized ages was non-existent, with an R-squared value of near zero and a high p-value. For teenagers aged 15-19 (Figure 4), there was a negative trend, though there was still a rather low R-squared value of 0.18. The p-value was found to be 0.017, which indicates there may be some significance that warrants further investigation. The non-linearity of these figures signifies that mental health is a complex matter, with many factors contributing to the outcome. Because of this, a linear regression model is not ideal for this type of data. Additionally, the lack of correlation in these figures may be explained by the ranging time periods (1990-2019) – other factors that may contribute to anxiety, such as wars, disease, changing cultures, and major economic changes, are not constant, and change with time, adding further to the complexity of the matter.

A graph with a red line and blue dots

Description automatically generated

Figure 5: Global depression prevalence and internet usage.

A graph with red dots

Description automatically generated

Figure 6: Global depression prevalence and internet usage.

The findings for depression (Figures 5 and 6) showed slightly improved significance and model fitness. Compared to anxiety prevalence, teens aged 15-19 showed a positive correlation, while globally, the trend remained negative. The p-values have greatly decreased, meaning that internet connectivity may be a higher significance for depression than anxiety. However, these figures still suffer the same flaws of complexity and non-linearity introduced by the change in time.

A graph of a graph of a graph

Description automatically generated with medium confidence

Figure 7: Global anxiety prevalence and internet usage in the year 2019. The points represent countries.

A graph of a graph with blue dots

Description automatically generated with medium confidence

Figure 8: Global anxiety prevalence and internet usage of those less than 20 years of age.

A screenshot of a video game

Description automatically generated

Figure 9: Hexagonal heatmap of Figure 8.

When only a set year is focused upon, for example, the year 2019, the significance greatly rises, with the p-value being near zero. This data (Figures 7 and 8) assumes a more linear form, meaning that a linear regression model is more appropriate. The r-squared values range from 0.2 to 0.3, which represents a weak correlation. The higher r-squared value for those aged less than 20 suggests that the correlation is stronger among the youths. A heatmap (Figure 9) was generated to assist visualizing Figure 8, in which a slight upward trend can be seen.

A red line with blue dots

Description automatically generated

Figure 10: Global depression prevalence and internet usage in the year 2019.

A screen shot of a graph

Description automatically generated

Figure 11: Global depression prevalence and internet usage of those less than 20 years of age.

A screenshot of a video game

Description automatically generated

Figure 12: Hexagonal heatmap of Figure 11.

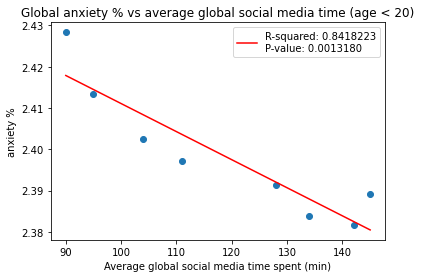
Depression correlations (Figures 10, 11, and 12) had similar results to anxiety, although with slightly worse linear fits. Unexpectedly, however, the prevalence for all ages switched to a downward trend.

4.2 Correlation with social media usage

A graph with a red line and blue dots

Description automatically generated

Figure 13: Global anxiety prevalence and average global social media time spent in minutes. The y-axis represents the share of population with anxiety disorders.

 Figure 14: Global anxiety prevalence and average global social media time spent in minutes among those aged less than 20.

When narrowed down to the specific activity of browsing social media, the linear fit improves drastically (Figure 13). The significance is particularly high for young people (Figure 14), with a p-value of 0.0013. There appears to be a negative correlation between anxiety and social media time spent. This is as expected, since social media interaction is a social activity after all.

A graph of a graph

Description automatically generated with medium confidence

Figure 15: Global depression prevalence and average global social media time spent in minutes.

A graph of a line with blue dots

Description automatically generated

Figure 16: Global depression prevalence and average global social media time spent in minutes among those aged less than 20.

Depression appeared to be strongly correlated with social media time spent, with a near perfect linear fit for all ages, (Figure 15) and an acceptable fit for young people (Figure 16). Both had significant p-values as well. Upon closer inspection of Figure 16, however, the increase in depression prevalence appears negligible, with only a 0.006% increase from beginning to end, meaning that the influence may not be as strong.

4.3 Comparing age groups in the United States

A screenshot of a graph

Description automatically generated

Figure 17: Share of U.S population among age groups that use the internet, have depressive disorders, and have anxiety disorders in 2019.

A graph with a line going up

Description automatically generated

Figure 18: Increase in internet usage, anxiety, and depression over time in the U.S.

A graph with a line drawn on it

Description automatically generatedFigure 19: Increase in internet usage, and anxiety, and depression prevalence among those under 20 over time in the U.S.

To better understand the differences between age groups, four distinct age groups in the U.S were analyzed (Figures 17, 18, 19). Findings show that the youngest age group (age 18-29) were the most connected to the internet and experienced the most anxiety and depression. Following along the x-axis, as age increases, internet connectivity decreases, along with anxiety and depression prevalence. Anxiety and depression prevalence were closely tied together, with negligible differences within age groups. Overall, younger people in the U.S were more likely to use the internet and experience depression or anxiety.

5 Discussion

One major limitation of this project was the scarcity of publicly available mental health data, particularly historical data from pre-internet times. This restriction made it difficult to establish a robust foundation for evaluating the changes in mental health trends due to rising internet and social media usage. The lack of detailed social media usage data, age distribution data in particular, was another limitation towards effectively answering the questions imposed by this project.

Future studies should prioritize obtaining richer and more comprehensive datasets, potentially through collaborations with organizations specializing in mental health research and world governments.

The use of a linear regression model to analyze the relationships between variables was not an optimal choice for this project. While it provides interpretability, linear regression assumes a linear relationship, which does not adequately capture the complexities and potential non-linearities in the data.

Future analyses should consider employing more appropriate models that can handle non-linear relationships, such as polynomial regression to handle non-linear data. Another possibility could be utilizing machine learning models along with algorithms such as decision trees, random forests, or neural networks could uncover complex interactions between variables.

Future work for this project should also include additional variables such as:

* demographic factors (age, gender, socioeconomic status)
* lifestyle factors (sleep, exercise, screen time)
* environmental and social influences (family dynamics, peer interactions).

To effectively include these variables, the scope of the analysis should scale down from an international, to national perspective.

6 Conclusion

This study explored the relationship between internet usage, social media activity, and the prevalence of depression and anxiety, particularly among youths under 20 years of age. The findings are as follows:

**Anxiety**:

When analyzed over a timeframe (1990-2019), there was no significant correlation between internet usage and anxiety prevalence. However, when focusing on a singular year (2019), a weak positive correlation was observed in all ages, suggesting that in specific contexts, internet usage may be slightly associated with anxiety prevalence. The global average time spent on social media saw a minimally negative correlation with anxiety prevalence.

**Depression**:

A weak positive correlation was found between internet usage and depression among youths, indicating a slight increase in depression prevalence with rising internet activity. A minimally positive correlation was also observed between the average global time spent on social media and depression.

These results contribute to the ongoing debate of mental health impacts imposed by internet technologies, particularly upon younger populations. By identifying weak but present correlations and significances, this study puts focus on the importance of responsible internet and social media use. The findings could help promote awareness about healthy internet habits minimizing negative online experiences and help promote online policy implementations that support mental well-being.

While the correlations identified in this study are modest, they raise important questions about how digital environments shape mental health outcomes. Future studies with more comprehensive datasets and advanced modeling techniques will help refine these insights further solidify our understanding of mental health.

REFERENCES

[1] Giulia Fioravanti, Sara Bocci Benucci, Giulia Ceragioli, and Silvia Casale. 2022. How the exposure to beauty ideals on social networking sites influences Body Image: A systematic review of experimental studies. Adolescent Research Review 7, 3 (January 2022), 419–458. DOI:http://dx.doi.org/10.1007/s40894-022-00179-4

[2] How Social Media Affects Your Teen’s Mental Health: A Parent’s Guide. Retrieved from <https://www.yalemedicine.org/news/social-media-teen-mental-health-a-parents-guide>

[3] IHME, Global Burden of Disease (2024) – with major processing by Our World in Data. “Anxiety disorders prevalence” [dataset]. IHME, Global Burden of Disease, “Global Burden of Disease - Mental Health Prevalence” [original data]. Retrieved November 24, 2024 from https://ourworldindata.org/grapher/anxiety-disorders-prevalence

[4] Jacqueline Nesi and Mitchell J. Prinstein. Using social media for social comparison and feedback-seeking: gender and popularity moderate associations with depressive symptoms. *Springer US*. Retrieved from <https://doi.org/10.1007/s10802-015-0020-0>

[5] Pew Research Center. 2024. Internet, Broadband Fact Sheet. Pew Research Center. Retrieved November 25, 2024 from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

[6] We Are Social, & DataReportal, & Hootsuite. 2024. Daily time spent on social networking by internet users worldwide from 2012 to 2024 (in minutes) [Graph]. In Statista. Retrieved November 25, 2024, from https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

[7] World Bank (2023) – with minor processing by Our World in Data. “Share of the population using the Internet – International Telecommunication Union” [dataset]. World Bank, “World Bank World Development Indicators” [original data]. Retrieved November 24, 2024 from https://ourworldindata.org/grapher/share-of-individuals-using-the-internet

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00