Increasing Frequency of Depression for Americans throughout 2020

Eli Asimow

Abstract:

In 2020, The United States has found itself in contention over how to balance the dangers of a pandemic with the threats of an extended lockdown. Arguments have been made across the political spectrum for both sides. One key point against a continuous lockdown is the effect quarantine has on mental health. This study suggested a trend in 2020, a year made up of pandemic stressors and lockdown isolation, of increased frequencies for symptoms of depression. Using NYU's Hadoop system, a census of the American people for symptoms of depression was analyzed over the eight month period between April and December. The results indicate which demographics' mental health have been most and least affected by the year. 18-19 year olds saw the highest increase in depression frequency. The analysis also procured a particular trend in education demographics, with a definitive positive correlation between level of education and increase in depression.

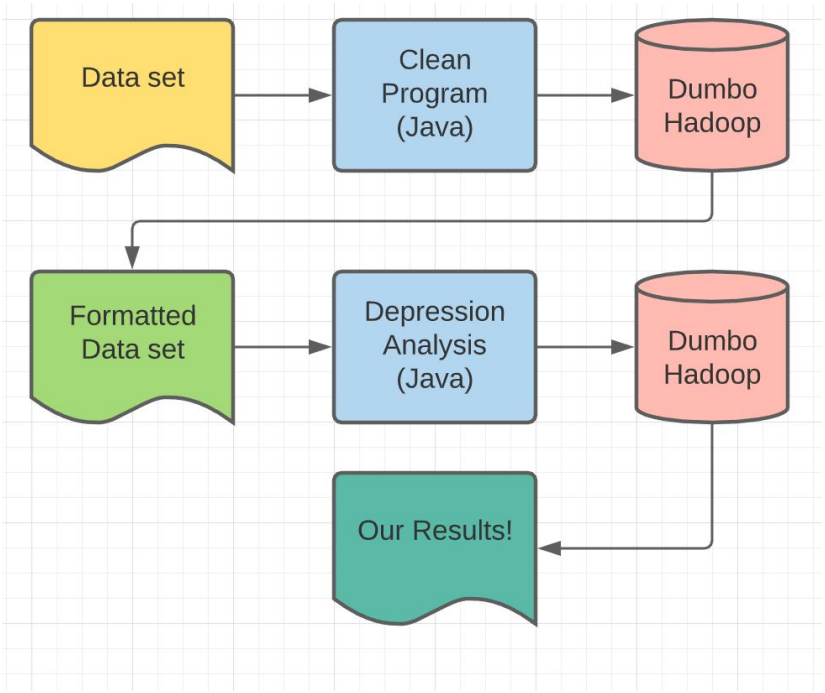
Keywords: Mental health, Depression, Covid, Social isolation, Hadoop, 2020.

Introduction:

I went into this research of American's mental health with a frame of mind already set by a non stop torrent of terrifying news. 2020 has been a disastrous year for the well being of The United States. Even ignoring the 300,000 who have died directly from Covid, the secondary effects of this pandemic have changed the daily lives of nearly everyone. Many have now been in almost perpetual lockdown for over eight months. Social gatherings, exercise, and new experiences, all important for maintaining a human's well-being, are made much more difficult by a withdrawal to a single living space. Furthermore, our country experienced a massive

recession in the beginning months of the pandemic, adding an additional fiscal stress to many families.

All this has formed an underlying assumption, common among many, that depression has become a major issue in our country. By studying a census of the American people, this study hoped to validate or disprove that assumption. The analytic used provided hard statistics for demographics of age, ethnicity, education, and location that will further advance our knowledge of how Covid-19 has hurt the American people.



In the above diagram, the data flow process for this research is visualized. It begins with the data set, a CSV file of symptom frequency among different demographics across different weeks.

Through Java code written for Hadoop, the CSV file is cut down to only contain the relevant information. This process includes removing entries for mental illnesses other than depression, and values for unwanted information such as “Phase” and “Quartile Range.” The entries are then sorted by demographic, and have their day in the year calculated from the starting date of their two week range.

The formatted CSV is then run through Hadoop again, this time using our Depression Analysis Code. This code generates a best fit line slope for each demographic, finds the date of the highest frequency of symptoms, and finds the date of the lowest frequency of symptoms. These stats are then outputted in one final results text file.

Motivation:

Depression is a terrible mental illness. It can tear apart ambitions, destroy relationships, and even lead to suicide. And yet, for as dangerous an illness as it is, conversation of the subject is still largely stigmatized. For many, discussing the mental health dangers of quarantine means delegitimizing the very real threat that a deadly pathogen poses. Most visibly, president Donald Trump, who has repeatedly down played the deadly nature of Covid-19, has used mental health as a talking point and justification for a looser approach to mandated lock down. Thus, the subject of depression in quarantine becomes politicized.

My hope is that, with real data on the subject, direct help can be provided to those most affected, without lessening the mandated social isolation needed to keep this virus under control. I fully expected to find an increase in depression across most demographics of Americans. Afterall, studies of mental health in Wuhan, China tied the stresses of the pandemic to high odds for depression and anxiety. In their paper, published in April of 2020, the Chinese team led by Junling Gao observed a correlation (OR = 1.72, 95%CI: 1.31–2.26) between social media exposure to Covid-19 and symptoms of depression or anxiety. My research strove to prove a similar correlation in America, utilizing the Household Pulse survey from the U.S Census Bureau. In their paper, the Chinese researchers used their results to argue that the government must "pay more attention to mental health problems, … while [also] combating a public health emergency."[1] I sought to, results permitting, make a similar argument to our public officials. If we could find out how severe a problem depression has become in America, and the demographics which it most affects, our country will be all the more equipped to provide the American people the support that they need. That support, when given, will lead to saved lives in a pandemic that has already taken far too many.

---

[1] *Gao J, Zheng P, Jia Y, Chen H, Mao Y, Chen S, et al. (2020) Mental health problems and social media exposure during COVID-19 outbreak. PLoS ONE 15(4): e0231924. https://doi.org/10.1371/journal.pone.0231924*

Related Work:

There are already several research papers published on the subject of mental health during Covid-19:

Junling Gao,Pinpin Zheng,Yingnan Jia,Hao Chen,Yimeng Mao,Suhong Chen,Yi Wang,Hua Fu, and Junming Dai [*Mental health problems and social media exposure during COVID-19 outbreak*] studied the effect of social media exposure (SME) in Covid-19 on symptoms of anxiety and depression. The research, using online surveys for rapid assessment, deduced a positive correlation between SME and symptoms of anxiety and depression in China. 4827 participants were surveyed between January 31 and February 2, 2020. This is extremely relevant to this research paper due to the similar subject, motivation, and survey data set used. See their research here:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231924

Samantha K Brooks, PhD Rebecca K Webster, PhD Louise E Smith, PhD Lisa Woodland, MSc Prof Simon Wessely, FMedSci Prof Neil Greenberg, FRCPsych Gideon James Rubin, PhD, et Al. [*The psychological impact of quarantine and how to reduce it: rapid review of the evidence*] researched the psychological effects of the various stressors of a pandemic, including quarantine, infection fears, and financial loss. Using 24 different data sets on quarantine, including surveys, interviews, and observations, the researchers observed that serious long term psychological harm can occur during quarantine. They used these results to argue that quarantining of individuals should be ended as soon as it is safe to do so.
See their research here:
https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30460-8/fulltext

Dataset

Household Pulse survey, "Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days."

This data set, provided through the CDC and constructed by the U.S Census Bureau, was the main data used for this research paper. A survey of the American people from April 23rd to the present, the data set includes values for different mental illnesses, representing the frequency of their symptoms among different demographics for two week periods. These demographics include each State, age ranges of 10 years starting at 18, education, and different ethnicities.

The survey was given to a random selection of the American people, sent to them via text message and email. The questionnaire is a modified version of the two-item Patient Health Questionnaire (PHQ-2) and the two-item Generalized Anxiety Disorder (GAD-2) scale. Over all, 3,838 data points have been made. It is safe to assume that that count has increased even further in recent weeks, as the survey continues past the point of the paper's publication. The dataset met the standards for data presentation set by the NCHS, the National Center for Health Statistics.

After eliminating non relevant information, the data schema used in this research fell into four categories:

Group, String

- Which category of people this depression / anxiety value is assigned to for the week.
- Examples: By Education, State, Gender, and Age

SupGroup, String

- Which sub group of the group category that set is.
- Examples: the subgroups for age are ~10 year stretches, such as 18-29. The subgroups for State are the specific state, such as California.

Time Period, String

- Which two weeks this value was recorded during.
- Example: Apr 23 -May 5

Value, float

- Number out of 100 for recorded frequency of depression symptoms from the questionnaire given.

Analytic Stages:

Note, all code was in the Java language, and ran through NYU's Dumbo Hadoop system.

Ingestion Phase:

The ingestion phase was rather straightforward. Due to the universal nature of the dataset's CSV file format, and the dataset's small size of only 500 KB, the file could simply be manually copy pasted into NYU's Dumbo system.

Cleaning, Profiling Phase:

There were several parts to the cleaning phase of the data set. There were many inconsistencies in formatting, including additional commas in categories such as "Non-Hispanic Asian, single race" that needed to be accounted for. Furthermore, many columns included unnecessary data, such as Time Period, which was always 1, and Phase, which was only -1 in blank entries. Such columns were eliminated such that only the relevant values of the data schema remained. Then, the date was converted from a string (Ex: Feb 3) to a day in the year, (Ex: 34). These dates were then combined with the symptom frequency value to form a tuple, separated by a comma. Following this, the data set was now cleaned and ready to be analyzed.

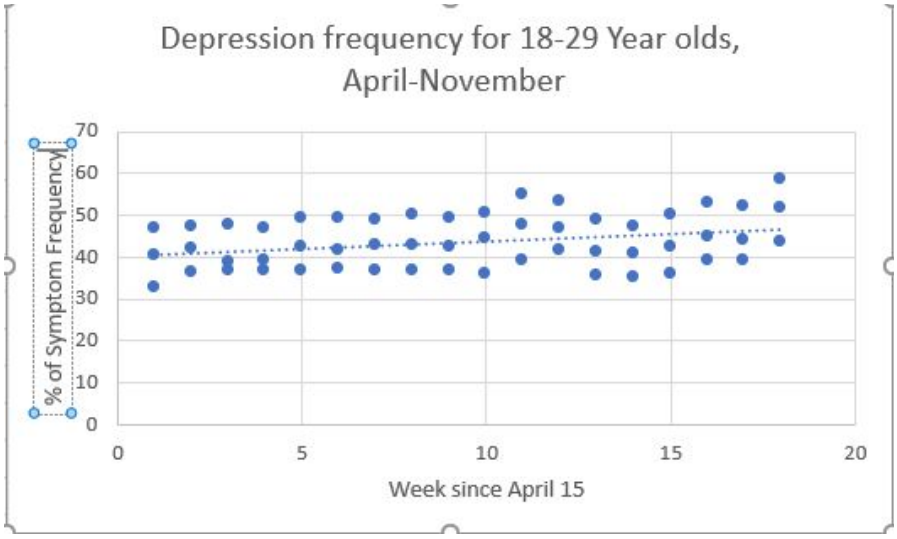
Analytics Phase:

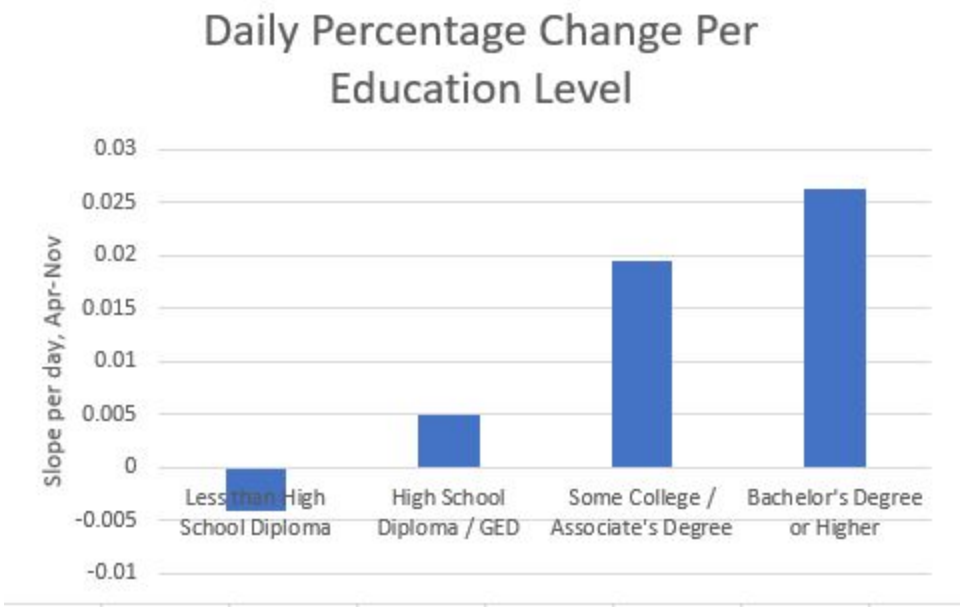
From the cleaned dataset, pairs of keys and values were made from each line in MapReduce's Mapping phase. The key would be the category and subcategory of the demographic together as a string. The value would be the symptom frequency and day tuple pair, still passed as a string. Each line would also map a general key to their value, for later calculating of overall trends. In the reduce phase, a loop ran through each tuple value of the demographic, separating them into a float for date and float for value, appending them to their separate lists. A lowest value and highest value index were also tracked through this loop. Then, in a "best fit line

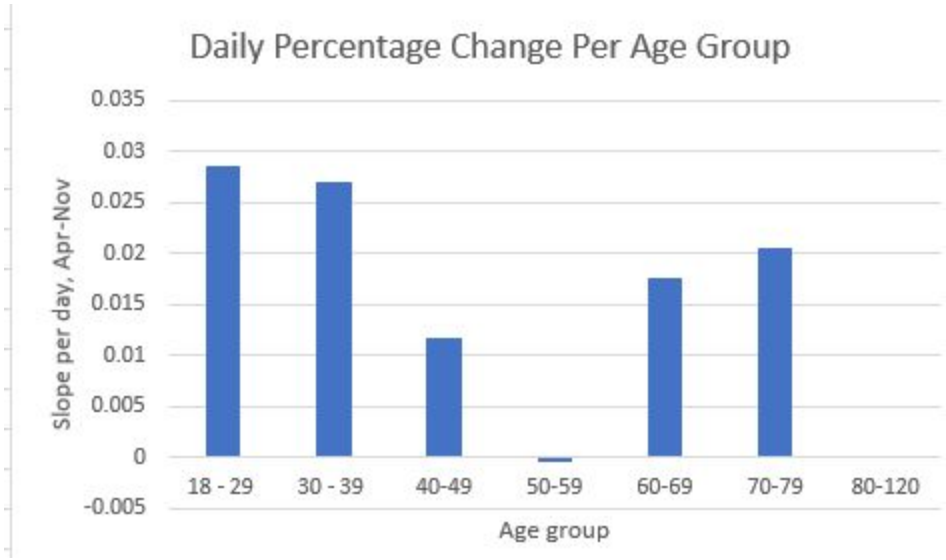
slope" method, a slope for the value of depression frequency was calculated over the survey's time period. With all the data obtained, a final result was printed with the format:

Category increased/decreased by percentage points _X_ per day since April 23. Their highest value was on day _D1_ with percentage _H_, and their lowest value was on day _D2_ with percentage _L_.

Results:




Daily Percentage Change Per Education Level
Slope per day, Apr-Nov
0.03
0.025
0.02
0.015
0.01
0.005
0
-0.005
-0.01
Less than High School Diploma
High School Diploma / GED
Some College / Associate's Degree
Bachelor's Degree or Higher


Daily Percentage Change Per Age Group
Slope per day, Apr-Nov
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
-0.005
18 - 29
30 - 39
40-49
50-59
60-69
70-79
80-120
Age group

Conclusion:

In conclusion, this paper used NYU's Dumbo Hadoop system to evaluate the effect 2020 has had on all types of Americans. By calculating average daily increases in depression for each demographic, the hope was that I could find which groups in particular needed the most mental health assistance. The results showed some fascinating trends across different education levels and generations. Generally, depression has certainly increased in frequency, with an overall calculated daily percentage increase of 0.0165% since April. Though this number is small, when applied to the 233 days since the survey began, it suggests an increase in depression symptoms' frequency of roughly 4%. This number shouldn't be taken too literally given the inconsistent nature of depression symptoms, but it still suggests a serious problem brewing during quarantine.

In specific categories, this issue is even more pronounced. The 18-29 year olds showed the highest single date value of any demographic, surpassing all 50 states, all ethnicity categories, and all other measured demographics. 58.7% of 18-29 aged surveyees reported positive to symptoms of depression in the two week period starting on October 29th. Looking at the bar graph comparing them to other age groups, we see that they have been disproportionately affected by this year's stressors.

At the education level we see a clean correlation with the rate of depression symptoms' increase. Those without a high school education are one of the few demographics to actually decline in value since April. Seemingly the more education one has, the more this year has deteriorated their mental state. Although fascinating, this trend is a little worrying. It suggests that education, and the critical thinking skills and knowhow that it develops, has coincided with a likelihood for serious mental illness as a result of 2020.

Further analytics could be drawn from this data set for specific results. In particular, given more time I would like to develop a polynomial best fit line algorithm, such that we could see which months saw the greatest increase in depression frequency. In doing so, one could see if months of heavy lock down coincided with higher rates of depression.

These results help to contextualize who exactly has hurt the most mentally from some 8 months of quarantine. The indirect duress that Covid-19 has caused this country should not be overlooked. I hope that this paper can be used to argue for better mental health support infrastructure across the board. More programs should be made with outreach to our youngest,

showing them that help is there if they need it. Funding for programs like the National Suicide Prevention Lifeline and the Substance Abuse and Mental Health Services Administration need to reflect our country's growing need for them.