**The Impact of Covid-19 on the Influx of Stress Relief Behavior**

by Lydia Halter (4000 Level)

Rensselaer Polytechnic Institute

Abstract and Introduction

Eating disorders (or “EDs”) are extremely serious, and sometimes fatal, affecting tens of millions of Americans (mainly women). Most eating disorder onsets occur between the ages of 12 and 25 so the majority of people with some sort of ED are in high school or college. This makes sense as this is the time where major stressors such as social standard pressure, schoolwork, college applications, etc. are coming into a field of view. Being one such person within this age range, and the fact that myself and many other women I know have dealt with a variety of such disorders, this topic is of personal significance, the end goal of which is to hopefully propose a way to be more proactive in preventing, or at best lessening, ultimate spikes in ED cases.

Eating disorders are under the umbrella of emotional disorders and often go hand-in-hand with other such disorders, namely depression and anxiety, and is often a method of control. Understanding this helps us to realize why the onset of the Covid-19 pandemic has had the impact that it did on bulimia searches. Additionally, throughout my research, I’ve expanded my focus from not just bulimia but to the global search cases of alcohol due to their stress relieving nature, hypothesizing that the pandemic caused a spike in these searches as well.

Data Description and Exploratory Data Analytics

Initially, searching for extensive data on bulimia was definitely challenging. The few datasets I found were either completely different from the sub-topics I was looking for or were far too qualitative to perform statistical data analysis.

Eventually, I landed on the Google Trends site, where it allows you to put in any topic (or multiple if you choose) and see the search rates over a period of time of your choosing. You can also filter by country/state/region/city, category, and type of search (i.e. image, news, etc.). It provides a csv of a column of dates and the rate of searches by percentage over that period (100 indicating the time of highest search in that period). While this wasn’t exactly a fool-proof way of acquiring statistics for eating disorder cases, I believe the information it does provide outweighs the inhibiting technicalities. As ED cases increase, searches of that topic and those related generally tend to increase not only by the individuals affected but also by their family and friends. Well after I collected my data on bulimia, I used Google Trends to collect my data on alcohol as well.

My third dataset, however, was acquired from statistica.com. It contains the data for [almost] every global country with variables such as “confirmed cases,” “cases in the last 7 days,” and “death per million (total),” the last is my variable of focus. This was used in conjunction with the other two data sets for my second model of k-means clustering analysis. However, in my first model (was actually two separate linear regression models, one for bulimia and the other for alcohol), this dataset was not implemented.

Analysis

Because I would be using my Covid-19 statistics data in the second model, I thought it best to only include the data for countries that were found across all three datasets. Many countries included in the Covid dataset were unfortunately excluded in both my bulimia and alcohol datasets. This is because Google Trends only collects data in countries where the browser is accessible, so a good handful of countries presented no data. Taking note of these, I also excluded said countries in my Covid dataset, allowing for a total of 127 countries to remain.

My exploratory data analysis for my bulimia dataset includes boxplots and histograms. For all three datasets, the countries are in order of greatest to least “deaths per million.” Taking note of this, I initially created eight boxplots (figures 1-8): two representing every “quarter” of this so-called list to see if there were any patterns, trends, or differences as the deaths per million decreased. What I found was that the data means (i.e. average search rate) began fairly low, then increased in the middle, and decreased again. For my histograms, I created six (figures 9-14): two first world countries, two second world, and two third world. Interestingly enough, these findings also produced a pattern of sorts. All graphs were shown to be unimodal, and the first world countries had a slight skew to the right and an average between 30-50%. The second world countries showed a similar display, but the center, at 10-20%, was a little further left, thus had more dramatic skews to the right. The final third world country histograms showed an even further progression to the left with an central mode of 0-10%, whose skews were incredibly dramatic.

I also did some exploratory analysis in Excel for both Google Trend datasets, and I used the same method for both. I first selected two time periods before and after the onset of Covid-19 (around January 2020). Then, under my data, I used the AVERAGE function to calculate the mean of the pre-Covid block, and under it calculate the mean of the post-Covid block. Under that, I took the difference of the post-average and pre-average to see how that specific country’s searches (for either bulimia or alcohol) increased or decreased as a result of the pandemic. Using the autofill/drag feature, I quickly ascertained these same values for all 127 countries. Finally, averaging out the differences allowed me to see the overall global effect. I did this for a few time blocks and found that generally, both datasets resulted in a positive change. However, I did notice that the average difference for alcohol searches were much higher than that of bulimia. This is actually one of the reasons why I eventually included my alcohol dataset as an official part of this study; shifting my focus from one aspect of the aftereffects of Covid-19 to two seemingly disconnected topics not only increased the topic’s interest but allowed for more conclusive models to be applicable.

Additionally, I was able to create some preliminary scatterplots, adding to them their linear regression equations and R-squared values. This was in part out of curiosity and in part a clarification that I was on the right track, and my data was presenting itself as I had expected it to. I also made some residual scatterplots to visualize how often a country fell below the x-axis versus above (where ideally most if not all would be above, hinting to an overall increase of search rates.

Switching the focus to potential error, uncertainty, and bias, there are a few notable things to point out. Firstly, the luxury of choosing my before and after periods has its pros and cons. The pros being that since the spikes in bulimia and alcohol were different, I shouldn’t be consistent with my times selected between the two datasets, and that I am able to navigate to the best (subjectively) possible encapsulation of search averages before and after Covid. The cons, on the other hand, are that there may be a much better representation of such periods pre- and post-, and this freedom also gives way to a potential bias in choosing which time periods reflect the best for a greater average of pre- and post- differences.

Model Development and Application of Models

*Model 1: Linear Regression*

As previously stated, this research involved two models, the first regarding linear regression. Starting with my bulimia dataset, I began by separating my data into two subsets: pre.df which encloses all 127 countries from September 29, 2019 to December 28, 2019, and post.df which again encloses all countries but from February 16, 2020 to January 9, 2021. Taking the average of these two subsets I created two new vectors, bul\_pre and bul\_post, reflecting this step. From here, I introduced a new data frame “means” to combine bul\_pre and bul\_post.

The next step was to create my linear model equation, including the R-squared value. Underneath, using ggplot, I was able to create a scatterplot of post-Covid bulimia search means per country (bul\_post) versus pre-Covid bulimia search means per country (bul\_pre). My results (figure 15) showed a very strong, linear trend. The equation came to y = 0.87x + 0.35, and my R-squared value came to 0.803. What’s interesting to note is that as the pre- and post- means increased, the residuals of my data got smaller and smaller. The fact that this trend is so strong helps to verify the premise that all countries seem to be behaving the same way.

Now looking at the slope, we can generalize our data to say that as a country’s pre-Covid average increases, their post-Covid average increases at close to the same rate, but even moderate increases are significant. The y-intercept value is pretty close to zero compare to the range of the y-values, so we can’t draw too notable a significance here.

Regarding the alcohol dataset, I also began with reading in the data and creating subsets, however, the time spans have changed. We have alc\_pre.df enclosing all 127 countries from September 29, 2019 to December 28, 2019, and alc\_post.df enclosing all 127 countries from December 29, 2019 to March 28, 2020. Two new vectors denoted alc\_pre and alc\_post were created containing the averages of these subsets per country, and again a new data frame, alc\_means, combined the two. Using the same linear equation function as before, I set my model equation and R-squared value. Using ggplot, I visualized my scatterplot including the trendline, equation, and R-squared value (figure 16) which produced similar results as with my bulimia data.

The equation came out to be y = 0.95x – 5.7 and the R-squared value was 0.832. Both were relatively close to the first model, but what was significantly different was that the residuals stayed mostly the same throughout the length of the trendline. If anything, they were slightly larger toward the lower center of the line. But again, we can justify that the countries on average were behaving very close to the same way. We can also reuse the previous generalization that as a country’s pre-Covid average increases, their post-Covid average increases at nearly the same rate since the slope is very close to 1. Looking at the y-intercept, however, is a different subject this time as we have a slightly more significant value, and it’s negative. While this in and of itself “hurts” the hypothesis that the post-Covid alcohol search averages were greater than pre-Covid averages, this is a very small effect and likely has no real ground for influence.

Unlike with my bulimia data, I took my analysis one step further. I created a vector for the differences in search averages (post-Covid minus pre-Covid) and printed a residuals scatterplot (figure 17). The plot showed an astounding majority of the points (i.e. countries) not just above the x-axis, or zero-line, but significantly above versus those below the line. Compared to a residual graph of the same time periods exactly one year earlier, this is significantly more positive regarding the y-axis. The plot of one year prior shows a relatively even distribution above and below the x-axis, illustrating that there is likely no confound causing the data to behave this way.

*Model 2: K-Means Clustering Analysis*

To begin construction of my second model using k-means clustering analysis, I first created a new csv in excel imported as “gen.” This is a data frame with the rows as the 127 countries alongside three variables: average bulimia search differences, average alcohol search differences, and deaths per million.

Using this new dataset, I went on to construct my pca function (figure 18). The “elbow” clearly presented itself at 2, so I chose to use two clusters for my analysis. Next, I set the seed to be 50, and created my k-means function. Filtering out the clusters from this new development, I created a new data frame with the country names in one column and either TRUE or FALSE in the second column, TRUE meaning that the country is in cluster 1. Finding the country number where the FALSEs stop and the TRUEs begin, I was able to subset my data into cluster 1 (c1) and cluster 2 (c2).

Initially, my plan was to produce a list-like result of the countries per cluster being either first world, or second/third world countries. Initially hypothesizing that this is how they would eventually be clustered, I manually began searching and marking which countries in my two clusters were first and second/third world. I soon found that there was a very clear mix of all three in each cluster, so I knew I wouldn’t get the results of my hypothesis.

Turning instead to a country-highlighting world map website, I wasn’t sure the patterns I was going to see, if I saw any at all, but decided to create two maps, each highlighting the countries of their clusters. To my surprise, there were distinct regional and continental differences far beyond what I had expected, so I decided to find a way to use R to plot these maps.

After going through countless tutorials, both video and written, I finally found a method that suited 1) the type of map I was intending to create and 2) the packages that my R-Studio edition allowed for installation. I installed many, including tidyverse, dplyr, and maps, and loaded through the library() function those plus several others including ggplot2, mapdata, and ggmap.

First, I created the data for a whole world map (gen\_map) but excluding the countries that weren’t in my datasets (most of them were from Africa). I was able to do this by using the function map\_data() and specifying the countries desired, or “region,” as the country names from my cluster data. I then created the data for two new maps, one for each of my clusters. Using the same function and ‘world’ data previously imported, I specified my regions to be c1 and c2, naming the variables c1\_map and c2\_map, respectively. I finally transitioned these to actual maps using ggplot and specifying the global dimensions using latitude and longitude.

My first map printed (figure 19) represents all 127 of the countries in my dataset. The second (figure 20) represents ONLY the countries in cluster 1. To even further distinguish, these countries are outlined in red. The third and final map (figure 21) represents the cluster 2 countries alone, outlined in blue.

Returning to my previous statement of how distinct the regional and continental differences were, it is much more clearly visualized with these three maps. As evidenced by figure 20, cluster 1 contains Canada, some of the Caribbean and northern South America, northern Europe, Africa, the Middle East, Asia, and Oceania. Cluster 2 contains the USA, most of South America, and central Europe (including Russia). For these near entire continents to have been distinctly separated by merely three variables of seemingly insignificant influence (except maybe deaths per millions) tells a lot about how these variables are connected.

I am confident with all my results. The strong linear regressions of both my bulimia and alcohol data suggest nonrandom behavior and combined with a third variable has created a strong method of distinction between two sets of varying regions.

Conclusions and Discussion

To summarize, both my linear regression models proved very strong, each over the 0.8 threshold, and my k-means clustering showed strong distinctions regionally and continentally. That aside, I am both grateful and disappointed that my focus shifted in the middle of my research. My initial goal was to not only examine bulimia search averages pre- and post-Covid, but to also search that of other eating disorders and additionally topics on the tangent of such disorders. For example, when bulimia searches spiked in the USA shortly after the Covid-19 pandemic began, other searches including ipecac and mouthwash also spiked. These two factors are related to the EDs anorexia nervosa and bulimia nervosa in that ipecac is sometimes taken to induce vomiting and mouthwash is frequently used among individuals with either of these disorders to keep up oral care. Knowing this, my intention was to cluster factors such as these, be they known or unknown in their relations (or lack thereof) to different eating disorders. The purpose of this would be to strategize which topics need more get-help services or hotlines featured on related browsers or websites as a preemptive measure for largescale traumatic events such as Covid.

The upside, however, was in a way going one step back to then go two steps forward. Branching away from the closed topic of eating disorders, I moved to generalize methods of stress relief or stress managing/dealing, specifically focusing on potentially negative or unhealthy methods. Continuing the research, I would implement the type of clustering I had initially intended to do, and broaden my scope of tangible features related to such methods to then, in turn, figure a way to implement preemptive action in society.

Bibliography

Rai, Bharatendra. *Geographic Plots in R | 1. World & USA Map - YouTube*. https://www.youtube.com/watch?v=lta5Mu1nryQ.

“Interactive Visited Countries Map.” *AmCharts*, https://www.amcharts.com/visited\_countries/.

Best, Raynor de. “Covid-19 Deaths per Capita by Country.” *Statista*, 25 Apr. 2022, https://www.statista.com/statistics/1104709/coronavirus-deaths-worldwide-per-million-inhabitants/.

*Google Trends*, Google, https://trends.google.com/trends/?geo=US.

MYaseen208MYaseen208 21.1k3434 gold badges141141 silver badges279279 bronze badges, et al. “Add Regression Line Equation and R^2 on Graph.” *Stack Overflow*, 1 June 1959, https://stackoverflow.com/questions/7549694/add-regression-line-equation-and-r2-on-graph.

Finnstats. “Cluster Analysis in R: R-Bloggers.” *R*, 20 Apr. 2021, https://www.r-bloggers.com/2021/04/cluster-analysis-in-r/.

Basille, Mel Moreno and Mathieu. “Mel Moreno and Mathieu Basille.” *r*, 25 Oct. 2018, https://r-spatial.org/r/2018/10/25/ggplot2-sf.html.

“World Map of Visited Countries in R.” *Stats and R*, https://statsandr.com/blog/world-map-of-visited-countries-in-r/.