Team No: 202

Track No.: 2.

Leveraging Covid and Trends Data to Inform University Policy on Anxiety

Mantautas Rimkus, Paul Chong, Nathan Ryder, Ian Taylor

Link to GitHub Repository: https://github.com/MantautasRimkus/StanfordDatathonCSU

Introduction

University students have been found to experience mental health disorders at a disproportionate rate. One meta-analysis found the average rate of depression in university students to be 30.6%, when the general US population is estimated to have a 9% average rate. The age range of most university students, 18-22, is host to large social and mental developments and is one of the more common time periods for mental disorders like anxiety to manifest. Adding the pressures of school and high-impact life choices, it is understandable that university students experience more than their fair share of mental health issues.

The year 2020 and the disease Covid-19 added further difficulty to the lives of those who are college aged. Students who returned or were admitted to universities in fall 2020 were forced to learn new styles of learning and make sacrifices in their social lives. Universities transitioned to partially or fully online curriculum, adding distance between student and staff communities. Unsurprisingly, measures for anxiety have increased under the duration of the Covid-19 pandemic. This report examines the survey-reported changes in anxiety for college-aged persons in the United States. Data on Covid-19 infection rates are linked with the mental health surveys via a fixed effects model. Google search trends are also tested for correlations with weekly anxiety measures.

University administration is concerned with student mental health. This study examines relationships of age-group specific anxiety surveys with data on Covid cases or Google search trends. Notable relationships may be useful for the monitoring the mental health of university student populations.

Data

The US Household Pulse Survey began in April 2020 as an attempt to monitor potential effects of the Covid pandemic on the US population. The survey asked questions on education, employment, food security, health, and other household characteristics that may be susceptible to change during a health emergency. There have been three phases to the survey, spanning from April 2020 to March 2021 with semi-weekly data available. For this study, Table 2a: "Symptoms of Anxiety Experienced in the Last 7 Days, by Select Characteristics" was gathered for each week of the survey. This table tracked the frequency that anxiety was experienced for the survey week, subjects responded with either "Not at all", "Several days", "More than half the days", or "Nearly every day". Additional variables, such as age, sex, race, education, etc. were recorded, and the survey results were split into the 50 states as well as overall US .

The anxiety survey data posed a few difficulties. First, the survey "weeks" were at irregular intervals and an inconsistent number of days. The irregularity of the intervals can be seen by some gaps in observations in Figure @ref{fig:usanx}. Second, the survey method used the Census Bureau's Master Address File to send emails and texts to very large samples. This method had high non-response rates and is likely subject to considerable bias, although this is a common feature of any survey data. Third, the age ranges recorded in this

survey begin with "18-29". To demonstrate that this age range is highly relevant to university administration, data was obtained from the National Center for Education Statistics. Figure 1 shows that ~30% of 18-29 year-olds in the United States are in higher education and that of those in higher education in the US, ~75% are between the ages 18 and 29. Thus the mental health characteristics of the age bracket 18-29 are will apply to the vast majority of university students.

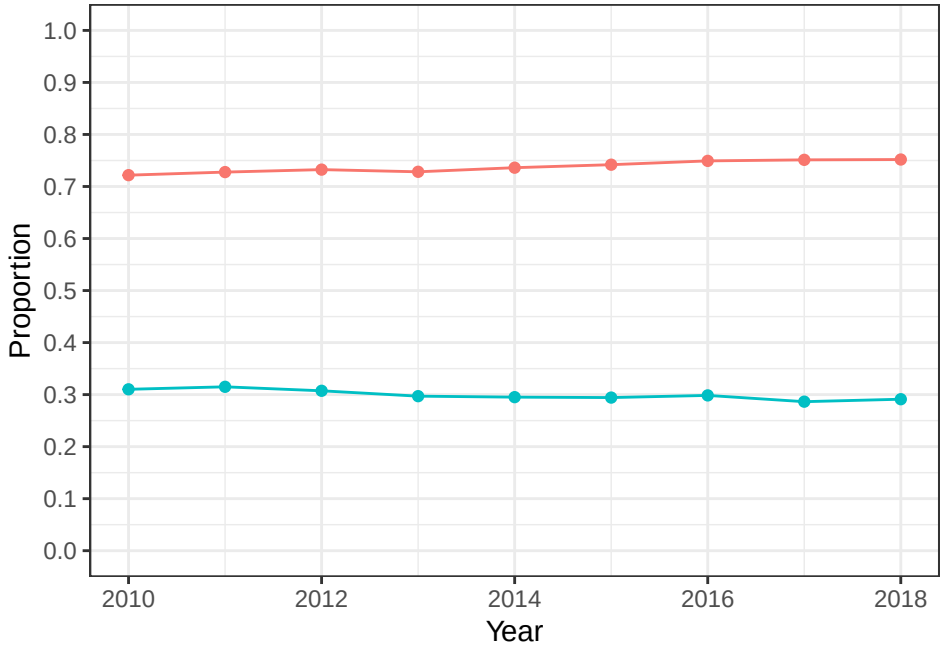


Figure 1: The proportion of US higher education students who are between ages 18-29 (blue) and the proportion of US 18-29 year olds who are in higher education (red).

To make a single value metric for anxiety levels, the survey estimates for the last two responses in the survey ("More than half the days" and "Nearly every day") were summed and divided by the sum of the first two responses("Not at all" and "Several days"). The resulting metric is the ratio of survey respondents who felt anxious more than half the time divided by respondents who felt anxious less than half the time. Furthermore, the anxiety survey data came with standard error estimates for each response. Using error propagation formulas as in, upper-bound standard error estimates could be produced. Figure 2 shows the anxiety ratio metric for 18-29 year-olds plotted over the survey period.

The Household Pulse Survey occurred in three phases, which can be seen in Figure 2 as gaps in observed time points. The anxiety ratio for 18-29 year-olds had an overall upwards trend for the year period, and the jumps in summer and winter may be related to spikes in Covid infections. In November 2020 a large increase in anxiety can be easily related to the 2020 presidential election.

As described in the technical documentation (LINK), the weeks of the Household Pulse Survey do not represent strict intervals of 7 days. For example, "Week 1" refers to data that was collected between April 23rd and May 5th, 2020, (13 days), but Week 9 refers to data that was collected between June 25th and June 30th, 2020 (6 days). The time points s_i will be defined as the end date of each Week collection.

Notice that the s_i are sparse and lead to unequal spaces between data points. To ensure univariate differences between time points, and be able to implement the models described later, interpolation was used with new time points $t_1, \ldots, t_{47}$ representing Sundays between May 05, 2020, and March 29, 2021. To calculate each X_{t_i}, where $s_{j-1} < t_i \leq s_j$ the formula :

$$X_{t_i} = \frac{d(s_{j-1}, t_i)}{d(s_{j-1}, s_j)} X_{s_j} + \frac{d(t_i, s_j)}{d(s_{j-1}, s_j)} X_{s_{j-1}} \text{ was used, where } d \text{ is difference in days.}$$

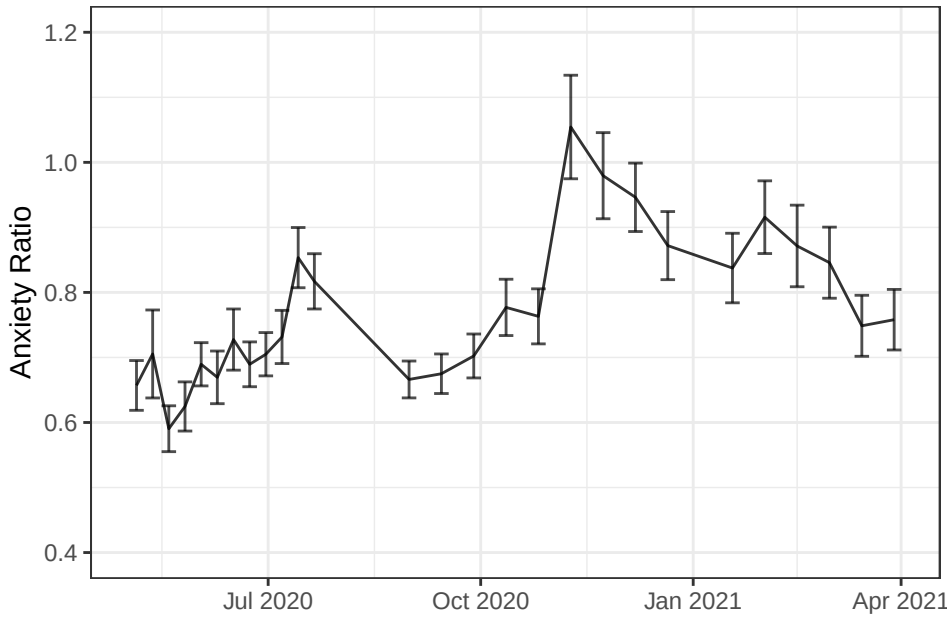


Figure 2: Anxiety ratio (more than half the time)/(less than half the time) of 18-29 year-olds plotted over survey period with standard error estimates at each observation.

Figure 3 depicts raw data $X_{s_1}, \ldots, X_{s_2 7}$ and interpolated data $X_{t_1}, \ldots, X_{t_j}$ for anxiety ratios in Colorado. Figure 4 shows the interpolated anxiety ratios for all states with Colorado and California emphasized.

Modeling

Predicting Anxiety from COVID Cases Using a Panel Model

We are searching for a way for school administrators to anticipate increased anxiety among their students. In 2020 and today, anecdotally, COVID is a significant source of anxiety. We used a panel model to find the predictiveness of covid cases on anxiety in 18-29 year olds.

Panel models are used for regression when the response and predictor variables are longitudinal. In this case, we treat US states as entities and have longitudinal values for both anxiety and new COVID cases.

For states $s = 1, \ldots, 50$ and time periods $t = 1, \ldots, T$, we use the model

$$\text{anx}_{s,t} = \alpha_s + \beta_1 \text{anx}_{s,t-1} + \beta_2 \text{COVID}_{s,t} + \varepsilon_{s,t}$$

This model allows for state-specific intercepts (α_s), autocorrelation in anxiety within a state (β_1), and an effect of COVID cases on anxiety (β_1).

Alternatively, we can have a random effect for the state-specific intercepts,

$$\begin{aligned} \text{anx}_{s,t} &= \alpha + a_s + \beta_1 \text{anx}_{s,t-1} + \beta_2 \text{COVID}_{s,t} + \varepsilon_{s,t} \\ a_s &\sim N(0, \sigma_a^2) \end{aligned}$$

The `plm` package in `R` allows us to fit both of these models and perform a hypothesis test to determine which fits our data better.

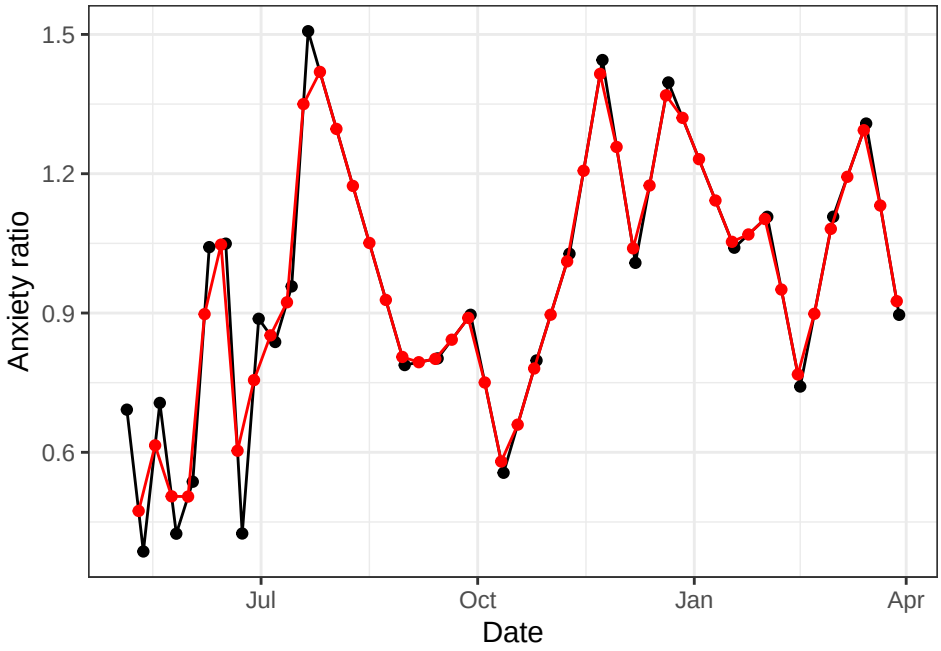


Figure 3: Anxiety Ratio in Colorado. Black - original data points, Red - interpolated data with equal time spaces

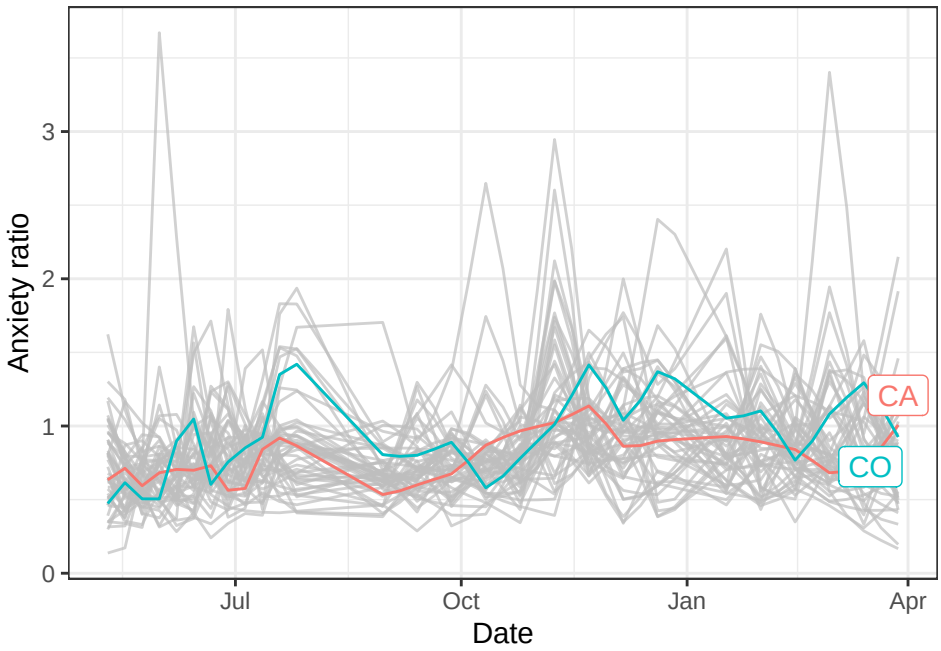


Figure 4: Interpolated anxiety ratios for all states with emphasis for Colorado and California

Because COVID cases can vary largely by state due to the size of the state, we were careful in how we constructed the $\text{COVID}_{s,t}$ covariate. We used the log ratio of new cases in sequential weeks as the COVID covariate:

$$\text{COVID}_{s,t} = \log\left(\frac{\text{New cases in state } s \text{ and week } t}{\text{New cases in state } s \text{ and week } t-1}\right).$$

This value is appealing for three reasons. First, it places all states on the same scale by only considering the ratio of subsequent weeks. Second, by using a ratio of new covid cases it handles the exponential growth nature of the pandemic. And third, it works well with an autoregressive model. A positive value means the pandemic is currently accelerating in a state, possibly resulting in increasing anxiety relative to week $t-1$. Similarly, a negative value means the pandemic is slowing down, possibly resulting in decreasing anxiety relative to week $t-1$.

Results

First, we have the output from the fixed effects model:

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = anx ~ LogCasesRatio + lag(anx), data = paneldata,
##     model = "within", index = c("State", "Date"))
##
## Unbalanced Panel: n = 50, T = 44-46, N = 2298
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -0.929705 -0.107738 -0.026498  0.082389  2.783689
##
## Coefficients:
##               Estimate Std. Error t-value  Pr(>|t|)
## LogCasesRatio 0.040212   0.015592   2.579  0.009971 **
## lag(anx)      0.696476   0.015199  45.824 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    192.43
## Residual Sum of Squares: 99.396
## R-Squared:      0.48346
## Adj. R-Squared: 0.47173
## F-statistic: 1051.1 on 2 and 2246 DF, p-value: < 2.22e-16
```

We can see that both the lagged anxiety variable and LogCasesRatio are significant predictors. The value of the coefficient with $\text{COVID}_{s,t}$ shows that if new cases were to double from one week to the next, we expect our response anxiety odds to increase by 0.0279.

The `plm` package also allows us to run a Hausman hypothesis test to see whether a fixed effects or random effects model is a better fit for our data.

```
##
##  Hausman Test
##
## data:  anx ~ LogCasesRatio + lag(anx)
## chisq = 71.979, df = 2, p-value = 2.344e-16
## alternative hypothesis: one model is inconsistent
```

The small p-value indicates that the null hypothesis of a random effects model should be rejected, and so our results from the fixed effects model are a better fit.

Associating Google Trends Data with Anxiety Surveys

The search engine Google makes its search data available through Google Trends at https://trends.google.com/trends/. Using the package `gtrendsr`, relative search popularity over time of a specified keyword can be obtained for a variety of regions. Search popularity of the keywords "lockdown" and "covid" was gathered for the study period at US and state levels. Figure 5 shows the search popularity (0-100) of the word "covid" in the US from April 2020 to March 2021.

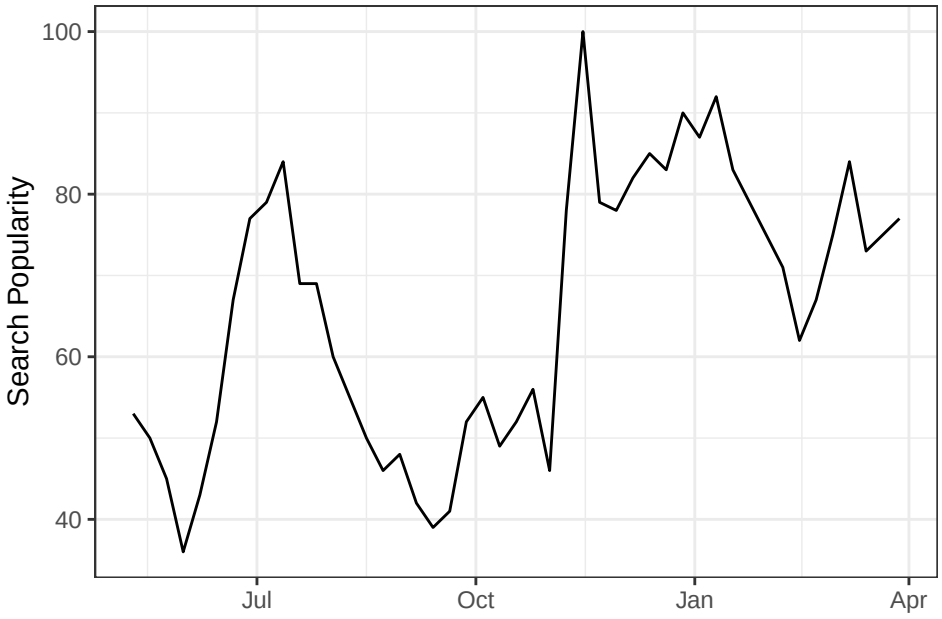
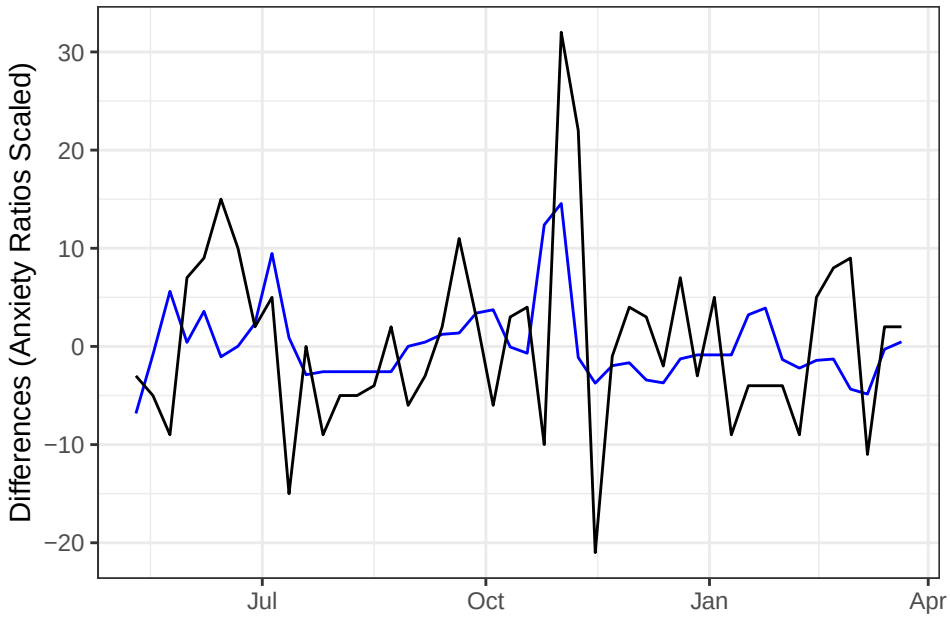


Figure 5: US Google Trends search popularity of "covid" over time.

Once again, a generally increasing pattern is noted, with a spike in November which may correspond with the 2020 elections. To compare search trends with anxiety data, we will instead find correlation with the first order differences of each data set. This will prevent spurious correlation of unit-root processes. The correlation was not very significant, even though the the trends do match by eye.




```
##
##  Pearson's product-moment correlation
##
## data:  anx_diffs$USdiffs and covidint_diffs$hitdiffs
## t = 1.8689, df = 44, p-value = 0.0683
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02074294  0.52050889
## sample estimates:
##       cor
## 0.2711881
```

We further examined the search trends data across the 50 states. Correlations calculated between the first order differences in google trends and anxiety data are as follows:

As it was seen in the given map, there may be some spatial-temporal trends between the difference in anxiety ratio and differences in Google Search trends for Covid. For the sake of this project, the joint significance of correlations of Colorado (where the team is based) and of California was tested using a permutation test. Each statistic T was calculated as

$$T = corr^2_{\text{California}} + corr^2_{\text{Colorado}}$$

The sampling distribution was derived using permutation sampling, wherein each iteration of each state's differences in anxiety ratio is assigned to a randomized state's differences in google trends. A sampling distribution was then obtained as $T_1, \ldots, T_{1000}$.

We obtained $p = 0.12$, thus the hypothesis of no significance cannot be rejected. Notice that taking into account the nature of data (noisy, interpolation), etc. One can argue that with $p - value$ around significance 0.1 there may be potential for more research in this area.

Discussion

The anxiety changes in 18–29-year-olds is important for universities, as this age group makes up a large majority of college students. However, due to the lack of university public data on student anxiety levels, the

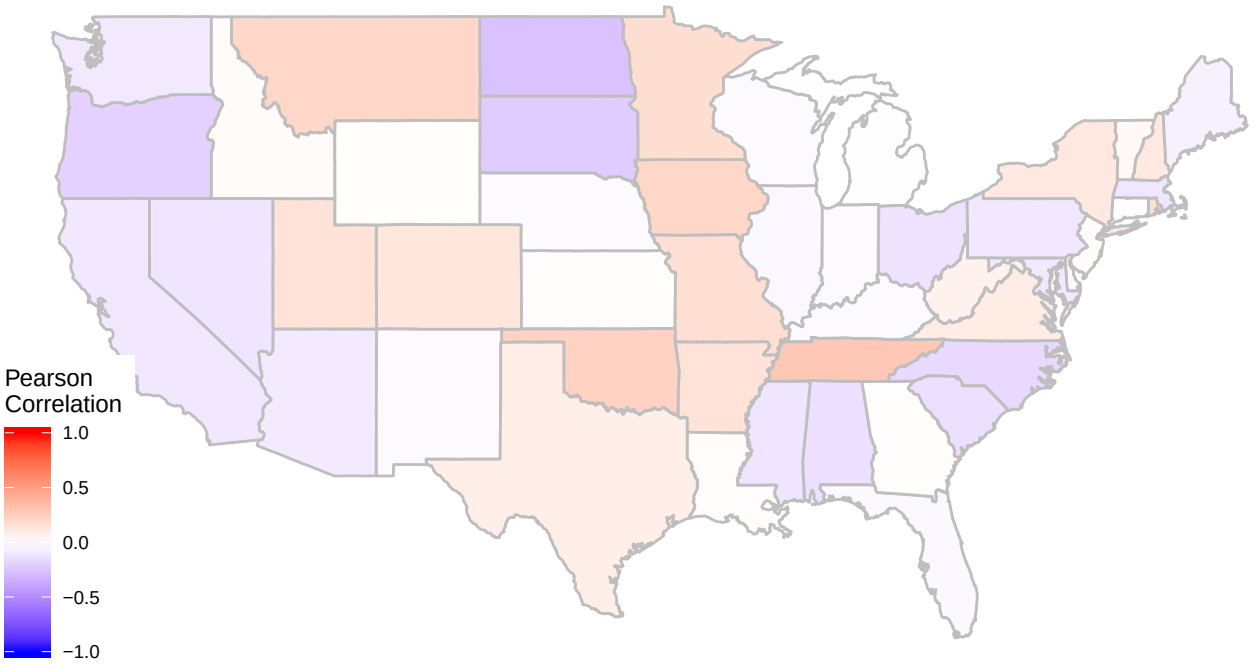


Figure 6: Using the keyword “covid”.

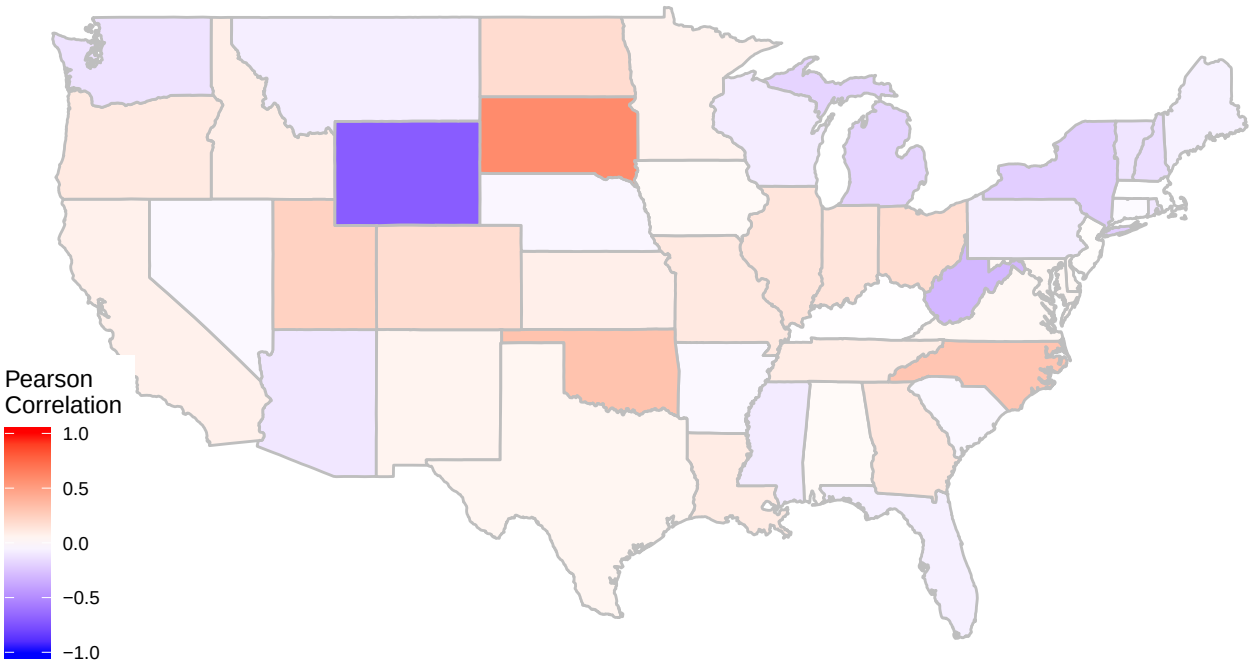


Figure 7: Using the keyword “lockdown”.

finest level of analysis can be done only at the state level.

It is shown that the anxiety ratio of 18–29 year-olds in the United States went through significant perturbations and tended to increase. Such changes might negatively affect the performance and achievements of students.

The panel data analysis shows that there are possible links between changes in covid numbers and anxiety ratios. Although the direct causal link between variables might be not found, some intermediate variables, like stress about the current situation or the uncertainty about the future, possibly could help to explain such a link.

Additionally, the analysis between changes in google searches and changes in anxiety did not produce a significant relationship but showed to have potential. The "eye test" for some clusters of states suggests that some universities can use Google search data to get a general overview of what might be going on in the state and infer about their students.

The methodology described in this report can be easily expanded and should be utilized by universities. A possible extension would be to use a network analysis to find a relationship between COVID and anxiety for neighboring states. This relationship could allow universities to anticipate increases in anxiety earlier. Universities could also collect their own data about student anxiety to establish a more concrete relationship between the statewide anxiety of 18–29-year-olds and the anxiety of their students. This analysis could also be applied to more mental health outcomes such as depression.

R code for data collection, cleaning, interpolation where necessary, and analysis are available in the supplemental GitHub repository.

Bibliography

1. Brey, Cristobal de, Thomas D. Snyder, Anlan Zhang, and Sally A. Dillow. "Digest of Education Statistics, 2019." NCES, February 25, 2021. https://nces.ed.gov/ipeds/Search?query=Age&query2=Age&resultType=all&page=1&sortBy=date_desc&collectionYears=2019-20&collectionYears=2018-19&overlayDigestTableId=201035.
2. Centers for Disease Control and Prevention, COVID-19 Response. COVID-19 Case Surveillance Public Data Access, Summary, and Limitations https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36
3. Croissant Y, Millo G (2008). "Panel Data Econometrics in R: The plm Package." Journal of Statistical Software, 27(2), 1–43. doi: 10.18637/jss.v027.i02.
4. Fields JF, Hunter-Childs J, Tersine A, Sisson J, Parker E, Velkoff V, Logan C, and Shin H. Design and Operation of the 2020 Household Pulse Survey, 2020. U.S. Census Bureau https://www.census.gov/programs-surveys/household-pulse-survey/data.html
5. Ibrahim, Ahmed K., Shona J. Kelly, Clive E. Adams, and Cris Glazebrook. "A Systematic Review of Studies of Depression Prevalence in University Students." Journal of Psychiatric Research 47, no. 3 (March 1, 2013): 391–400. https://doi.org/10.1016/j.jpsychires.2012.11.015.
6. Ku, H.H. "Notes on the Use of Propagation of Error Formulas." Journal of Research of the National Bureau of Standards: Engineering and Instrumentation 70C, no. 4 (1966): 263–73.
7. Lijster, Jasmijn M. de, Bram Dierckx, Elisabeth M.W.J. Utens, Frank C. Verhulst, Carola Zieldorff, Gwen C. Dieleman, and Jeroen S. Legerstee. "The Age of Onset of Anxiety Disorders." Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie 62, no. 4 (April 2017): 237–46. https://doi.org/10.1177/0706743716640757.
8. Philippe Massicotte and Dirk Eddelbuettel (2021). gtrendsR: Perform and Display Google Trends Queries. R package version 1.4.8. https://CRAN.R-project.org/package=gtrendsR
9. trends.google.com. "Google Trends." 2021. http://trends.google.com/trends?geo=US