Winter 12-5-2022

Critically Observing the Challenges and Changes: an Analysis on Covid-19’s Impact with an Emphasis on Students in Higher Education

Landon Perkins

CRITICALLY OBSERVING THE CHALLENGES AND CHANGES: AN ANALYSIS ON COVID-19'S IMPACT WITH AN EMPHASIS ON STUDENTS IN HIGHER EDUCATION

By
Landon Perkins

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of the requirements of the Sally McDonnell Barksdale Honors College.

Oxford
November 2022

Approved by

Dr. Kristin Davidson
Advisor: Dr. Kristin Davidson

Timothy Holston
Reader: Dr. Timothy Holston

Dr. Phillis George
Reader: Dr. Phillis George

To all my family and friends, who have always supported me and driven me to become the person I am today.

ACKNOWLEDGEMENTS

To Dr. Kristin Davidson, for being a great mentor and advisor as I worked my way through this process and being a great influence to challenge me throughout the project.

To Dr. Timothy Holston, for working with me and driving me to meet all the set goals for this project as well as being a great resource for whenever questions arose.

To Dr. Charlie Walter, for providing me with consistent feedback and questions and always keeping me on the right track.

To Dr. Timothy Yenter, for helping to guide me through this process and always being available to help whenever there were issues or questions.

ABSTRACT

This project involves comparing different visualizations related to COVID-19 and higher education in order to determine key impacts of the COVID-19 pandemic on students in higher education, as well as higher education as a whole. The main metrics used to determine the impact were mental health indicators for anxiety or depressive disorders, enrollment numbers by control type (public, private non-profit, or private for-profit) and state for 2020 and 2021, and state mandate lift dates for a variety of mandates implemented across the United States. These metrics were analyzed both individually and against each other to determine if they had any effect on each other. The project finds that mental health indicators for higher education demographics do not generally follow the state of COVID-19 case trends, but rather it was determined by circumstances of the start of the pandemic that led to a peak in mental health indicators. In addition, an immediate impact from COVID-19 can be seen in a vast majority of universities in the United States. Overall enrollment in higher education saw a substantial decrease in the United States. Despite this, private non-profit schools were not impacted like public schools, and some of the largest private non-profit schools actually saw a large increase in enrollment per capita compared to other types of universities.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: INTRODUCTION

Problem Statement

The COVID-19 pandemic has drastically impacted almost every aspect of peoples' lives since the start of the pandemic in 2020. Every business, system, and facet of life had to adapt in a variety of ways due to how infectious and rapidly spreading the virus was. Because of the recency of the pandemic, a lot of data has been collected, but that data is scattered across multiple groups around the world and has not been generally brought together to analyze the effects in a variety of factors with empirical data. In addition, the recentness of the events of the pandemic leaves a lot of questions in regard to the data. Some data was incomplete as it has only been two full years since the start of the pandemic, and data collection was not always optimized throughout different times of the collection period. Some present day data is still not available for the public and because of this comparing the present situation to pre-pandemic and during the pandemic is not yet possible. While some early studies and surveys have been completed, there is still a lot of room for examination and analysis of the impacts COVID-19 has had from a data perspective rather than from experiential based perspectives.

Define the Problem

The problem currently present is that the COVID-19 pandemic was a recent world event that needs more empirical analysis from a data science perspective. For this project, the COVID-19 pandemic is defined to have a started in March of 2020 and continues until September 2022. COVID-19 has had a far-reaching influence on most aspects of life for people, the impact of which will be felt for a long time. There is a need for critical analysis of data to pinpoint key factors that were greatly affected by the pandemic. There are great sources of information on the pandemic like the Centers for Disease Control and Prevention (CDC), but the information reported focuses largely on individual aspects and not merged with other data from the same time frame to look for COVID-19's impact on various of aspects of life. Some key points to be observed are key spikes in the pandemic such as the rises of the Alpha, Delta, and Omicron variants. It should also be observed if there are any trends in data related to mental health as well as academic statistics such as enrollment for students in higher education. While higher education has received some focus in regard to viewing COVID-19's impact, there is still a need for studies to be completed as more data becomes available and is merged from a variety of fields like academic institutions, academic resource providers, aggregate state data, and general pandemic information available.

Background Information

Because of the recency of the COVID-19 pandemic, the majority of data collection and presentation has been done through health organizations like the CDC or other parties whose line of work was directly impacted or controlled by the pandemic. Because of this, there are only a few main sources for aggregate count data or other empirical measures and now a lot of work has been done with this data besides presenting it. The rest of the data collection in regard to COVID-19 has been mostly from a variety of parties completing surveys pertaining to particular facets of individuals' lives and how those individuals are impacted by or interact with COVID-19. While there is other data being collected and reported, the two types listed are the majority of data being used in studies pertaining to COVID-19 currently, and by combining these two types of data along with other data from a variety of sources (like the U.S. Department of Education and third-party providers), COVID-19's impact can be analyzed. A lot of recent data for the most current information has also not been publicly released yet, so it is not possible to create a comprehensive look at the pandemic from the beginning in 2020 to the present where lingering effects are still felt. In addition, another point of consideration is that the data is potentially being split between a variety of sectors without awareness of what all information is available. If this is happening it would prevent a full representation of the situation.

Possible Solutions

Possible solutions to understanding the impacts of COVID-19 utilize two main approaches. The first is experience from the pandemic and an understanding of the events that transpired in order to cause changes in society as well as adapt to the circumstances. The second is a more empirical approach: looking at data collected before and during the pandemic to understand key impacts. This project merges both solutions—taking experiences from the pandemic with data collected to analyze and better understand the effects of the pandemic.

In terms of the analysis and visualization used for this project the two main choices for a data science project like this are Python or R. While Python is very powerful and versatile for data science, it’s primarily used more for machine learning, deep learning, and web-based applications that involve data science. On the other hand, R is more useful for statistical analysis and learning. R is also much more known for data exploration, experimentation, and visualization.

With the strengths of both languages in mind, R was chosen for this project. The main reason R is used is that it provides a lot of tools to help simplify the process and provide more freedom to the user than Python. Another reason in favor of R is that this project focuses more on standalone analysis and processing which does not require the vast amount of resources Python offers. Instead, the more refined statistical tools in R can be used. Another major advantage of R being utilized is the ability to plot interactive choropleths with GeoJSON data which is a very powerful tool for this project. One major drawback of R is the data limit that is placed on projects, but the data found and used for this project did not come near this limit meaning R was a valid choice.

Purpose of the Project

The main point of this project is to analyze the effects of the COVID-19 pandemic specifically on the higher education sphere. This will be paired with key visualizations from a multitude of sources collected to try and gain a clearer understanding of key impacts directly involved with professional staff and their field of work. This information could also be used to help inform future decision making for prospective or current students in a post COVID-19 higher education system when so many fundamental aspects were changed. In addition, the data aggregation completed for this project will be used to help understand the current status and understanding of data collection throughout the pandemic.

Chapter 2: REVIEW OF LITERATURE

Studies on COVID-19's Effect on Higher Education

The major challenge of developing studies on the impact of COVID-19 currently is the recency of all of the events regarding the pandemic and how not enough time has passed for a lot of information to be brought together for insights. This hinders studies and research as some of the most useful data is on the post-pandemic years and not enough time has passed for a majority of sources to release the most current data they have for the first time periods that could potentially be considered post-pandemic. In its current state, most of the data analysis compares data collected during the pandemic to before 2020 rather than to post-pandemic data. Despite these issues, studies can still be completed by looking at data from the first two years of the pandemic as well as by completing surveys to get results from a snapshot of time.

While a vast majority of studies looked for more general impacts from COVID-19, some also decided to investigate specific areas of interest and how they responded to the pandemic. One example of this is a study conducted during the first wave of the COVID-19 pandemic, the peak of lockdown restrictions, that was distributed online and covered aspects of life for higher education students like academic, social, and emotional circumstances and feelings. The survey gained 31,212 responses and then looked at the relationship between socio-demographic characteristics and elements of student life. The results showed that when given a choice of emotions to pick from to describe their emotional life over 50 percent of general respondents picked between bored, anxious, hopeful, and frustrated, and for personal circumstances over 50 percent of general respondents designated that they were dealing with study issues and personal finances (Aristovnik et al., 2021). These results show how at the start of the COVID-19 pandemic there was a noticeable amount of people in higher education who could feel the immediate impacts of the pandemic on their schooling experience as classes rapidly shifted online and other virtual styles.

Another study looked at the mental health implications for people in school during the lockdown period of the pandemic. This survey was held in India and received 1,182 responses across multiple age groups (primarily within higher education age groups). The survey contained questions asking about the time spent on online classes and self-study, the methods of learning used, sleeping habits, and daily fitness routines before looking at the effects of these habits on life factors like mental health. The research found that a variety of coping mechanisms were utilized to deal with stress and anxiety that were in-part caused by the drastic shift in learning methods and efforts caused by the pandemic. The authors suggest that online learning should continue to be invested in and that there needs to be more emotional support for students to help handle stressors and other mental health factors (Chaturvedi et al., 2021). Like the previous study mentioned, the primary methodology involved surveys in the early stages of the COVID-19 pandemic and, as such, lack relevant data to the discussion that can be added as the pandemic continued, but they both support similar findings about how quickly the pandemic showed an impact on those in higher education.

For a slightly different perspective the survey completed by Nicole Johnson, George Veletsianos, and Jeff Seaman is a relevant source (2020). The group surveyed 897 faculty and administrators across 672 U.S. institutions and found that almost all reported facilities rapidly shifted to emergency teaching and learning methods. They discovered that this shift was regardless of prior online teaching experience and no matter the level of experience new teaching methods were used. The greatest weakness was found in the lack of student support and access to resources for digital materials and working from home. The study gives some insight into the professional side of the pandemic during the first wave and showcases that issues in higher education were much larger than just one variable like students.

For research done on the key impact points for higher education, a case study was completed at Peking University in China that focused on the principles for online learning for higher education. Six teaching strategies were presented to provide a snapshot view of the online teaching experience early in the pandemic. Each strategy was evaluated to discover the impact of different teaching styles on student engagement and learning. The study found that there were five key principles for online education: high relevance between online instructional material and learning, effective delivery of information, appropriate support from staff, high-quality participation in an online medium, and a contingency plan to deal with unexpected incidents (Bao, 2020). These findings support the push to look at the impacts of COVID-19 on higher education and give a snapshot view of early studies done on higher education and how it was shaken up by the COVID-19 pandemic. Specifically, the snapshot gave an idea of how teaching scenarios would have to adapt to accommodate for the non-traditional classroom setting.

Another interesting factor to consider for mental health in higher education is the technostress, or stress caused by working with technology on a daily basis, that is present due to switching to a majority technology-based learning environment. One study conducted had 189 students and 172 educators and looked at three antecedents of technostress and its effect on the participants' performance in regard to remote classrooms and learning. The results of this study found that work-home conflict and overload were major contributors to technostress. The study also discovered that the effects of technostress on both students and educators was noticeable and distinct within both groups studied. The most noticeable difference is that "techno-distress in students explains almost twice as much of the variance in performance as educators (37.0%

versus 20.6%), showing that young people are more affected by technology-mediated study than their educators" (Bravo-Adasme & Cataldo, 2022). This study further shows evidence that the COVID-19 pandemic was directly related to an increase in mental health indicators for those in higher education, which is a focal point of this project.

Overall, the current literature available in regard to the COVID-19 pandemic and higher education primarily focuses on the first wave of the pandemic due to the time period of this writing, the desire to quickly get results to help understand the problems that arose from the pandemic causing a drastic shift in education methods, and the potential to determine solutions to the key issues that arose as the world began to further adapt under the circumstances brought about by the pandemic. This leaves a drastic need to continue looking into the lingering effects of the pandemic as well as later points in the pandemic past the first wave, points which have largely been neglected thus far. In addition, the current research on education is split by education level as well so there is an even greater spread of data that dilutes the amount of research for education regarding COVID-19.

Data Availability for COVID-19

While studies on COVID-19 have been mostly conducted during the first wave of the pandemic in 2020, data collection has continued to grow beyond the first wave and continued with daily, weekly, or periodic updating for most of the major sources for information used in this project. The CDC provides public datasets to be used regarding many statistics related to COVID-19. These datasets cover data ranging from vaccine distribution allocations for each of the main distributors (Janssen, Pfizer, and Moderna) to cases and deaths by state over time to variant proportions of the SARS-CoV-2 virus (Data | Centers for Disease Control and Prevention, n.d.). The presence of a major figure like the CDC who has led the discussion in regard to COVID-19 was beneficial as it provided a useful tool to find a majority of the COVID-19 data required for this project. Another major benefit is that the data provided by the CDC was not limited by not having data up to the current period of time.

Despite the ease of access to aggregate data for metrics like cases, deaths, vaccinations, and mental health, any more specific or area dependent metrics are not stored with the CDC and have to be collected and maintained by other groups. For higher education related datasets specifically there are a few providers that were pertinent to the discussion present for this project. The first of these is a dataset that contains education technology engagement data from over 200 school districts in 2020. Specifically, the dataset collects load page events from apps, web apps, extensions, eBooks, hardware, software, and services used in educational institutions across all school district levels (Learnplatform COVID-19 Impact on Digital Learning, n.d.). While not used in this project, this is valuable data to gain insight into what the most successful tools were for handling online education early in the COVID-19 pandemic. With this knowledge, it could be determined what tools could be improved upon for continued use, as well as potentially expose new solutions to educators who used a less popular alternative for their online learning environment.

Another useful dataset not maintained by the CDC is a record of the state mandates implemented in the United States. This dataset, which is maintained by the Boston University School of Public Health, records closures, shelter-in-place orders, physical distancing closures, reopenings, and more (Julia Raifman et al., 2021). While it does not record guidance or recommendations, having a distinct source for mandates and directives that apply to the entire state for each state in the United States is a very beneficial tool to have available. This allows for the impact of COVID-19 to be examined as it can provide context and help build a timeline of events for individual states as well as the United States as a whole. It was maintained from early 2020 through 2022, the time period that covers the entirety of the period in which state mandates were in place for the COVID-19 pandemic. This provides a complete collection for mandate data, which is a beneficial resource to be available when looking at the time period encompassed by the pandemic.

In regard to data availability for COVID-19 there is a plethora of information and general aggregate data available about the virus throughout the pandemic; however, when it comes to drilling down into specific information for a topic like higher education in tandem with COVID-19 the amount of data becomes drastically less. While there is some overlap in areas like business and economics, there is a lot of room for growth for a majority of other areas. There is also a distinct lack of data for some areas of research regarding COVID-19 past early 2020 with the first wave of the pandemic that can hopefully be further researched as more private institutions and groups provide their information to be analyzed in tandem with the vast amount of standalone COVID-19 data currently available.

Chapter 3: RESEARCH APPROACH AND PROCEDURES

Preliminary Ideas

With the premise of creating a dashboard to help visualize data related to COVID-19 with regard to higher education, multiple concepts and choices had to be completed early on in the research process for the project. When the project began, the goal was to find a list of key areas impacted by the pandemic and the datasets that match those areas to evaluate and visualize, but throughout the research process it was discovered that there were many limitations on the amount of public data that could be quickly retrieved and used so some concessions had to be made. While looking at prior research to see what had already been done in the field, it was found that a majority of studies looked into surveys primarily. So while there was good documentation on snapshot moments from the pandemic, there was no research taking aggregate data for higher education institutions and merging with COVID-19 data to provide insights implying the data would have to be aggregated for this project.

Statistics like enrollment and mental health indicators were the first to be researched as potential areas of impact and had relevant data, but with both some immediate issues arose that had to be addressed when wrangling the data. Other aspects that were initially looked into include online engagement with learning tools and taking survey responses to filter subset groups and find new results, but both topics were ruled out of analysis due to a lack of time or lack of complete data for the time period being looked at for this project. Because of finding early issues with many of the aspects being researched it was determined that the best course of action would be to focus on getting strong datasets for COVID-19 in the United States, general aggregate data for higher education institutions, and data on mental health indicators in the United States to start with and then grow from that subset as time permitted. From this decision, a lot of the focus for the project was shifted to highlight two main factors in higher education: mental health and enrollment. Each factor would be examined to find its relation to mandates enforced by individual states. Simple visualizations to be paired with each factor were also determined at this time to help guide preliminary coding and attempting to find early insights.

Generating Questions and Visualizations

Before selecting exactly which dataset would be used for this project, driving questions had to be generated. It was also necessary to determine what visualizations would be best to help answer each question. This process guided research as having specific questions led to knowing what kind of data was needed to answer each question and it could be determined if there was current publicly available data to answer a specific question.

One example of this process could be seen with the question “How much did student enrollment numbers change during COVID-19, and did some states change more than others?” In order to answer this question, a column chart was the visualization selected due to the ease of comparison between states and control types, specifically with the years being the x-axis and enrollment being the y-axis. Therefore, all that would be needed to visualize this question and get the desired answer was a dataset that contained university enrollment for 2020 and 2021 by university with a column designating the state that university is located in and a column designating the university control type. From there the wrangling and generating of visualizations for different states and controls would be able to be completed in R.

Another question that involved a similar process was the question “Is there any kind of difference between how quickly different states returned to 'normal' operations?” The first concept to be determined was how to measure a state returning to 'normal.' After some preliminary research it was determined that looking at state mandates placed and lifted regarding the COVID-19 pandemic would be the best way to quantify returning to 'normal' as it was a part of the current data available and removed subjectivity as much as possible. After determining the qualifier, it was decided that for this question multiple visualizations would be beneficial for getting the desired answer. A choropleth, or a statistical map colored by a variable value, would be useful to compare the average date a state lifted COVID-19 related mandates, and a scatterplot would be useful for looking at each individual mandate. Since a dataset was already found with state mandate information in research, the only other necessary dataset for this visualization would be a dataset with general COVID-19 information at a state level for additional information as a popup on the choropleth.

Some questions asked during this research process had to be dropped from the project due to a lack of publicly available data or time for the project. One such question is “Was there a difference in the proportion between in-state and out-of-state enrollment for schools due to COVID-19?” While this is a good question that would provide insight into how the higher education environment was affected by the pandemic, it could not be answered due to a lack of available aggregate data. Universities report in-state and out-of-state enrollment yearly on an individual basis, but it is not aggregated into a publicly available dataset like a lot of other college statistics so it would require an enormous effort to gather this data individually in the relatively short time frame this project persisted in.

Dataset Selection

When selecting data for this project, the main source for COVID-19 data was the CDC. The CDC provides public data sets on a plethora of points related to COVID-19 within the United States broken down at a state level, but the main pieces of data used from the CDC were from two main datasets. The first dataset contained the number of cases and deaths resulting from COVID-19 at a state level being reported daily, and the second dataset contained mental health indicators for a variety of subgroups of the population broken up by different time periods of data collection. These two datasets were the foundational datasets for this project as they provided the baseline that most other data would be compared to. Both datasets paired together nicely to create a time series showcasing the ebb and flow of COVID-19 throughout the United States since early 2020. While not as integral to the early development of the project, vaccination data was also pulled indirectly from the CDC to be used in state-level comparisons with the choropleth map present in the project (*State-by-State Data on COVID-19 Vaccinations in the United States*, n.d.).

Another large contributor of data for this project was the College Scorecard, a website controlled by the U.S. Department of Education that maintains datasets with comprehensive institution-level data for most universities in the United States. This was used for the project because it provided a breakdown of key information points for each university that would be useful for placing universities into a variety of groups (such as by state, control type, predominant degree type, etc.). This dataset also served as a solid base for connecting all university related datasets due to having the largest number of universities of all the datasets used and the unique identification information associated with each institution. The dataset also contains a vast amount of financial data for a number of variables including average cost, total debt, monthly loan payments, and more, but these were left out of the discussion due to a lack of time and difficulty to use these statistics in the discussion on key impacts.

The Integrated Postsecondary Education Data System (IPEDS) was also a plentiful resource for gathering data on higher education institutions. The system maintains information on institutional characteristics, enrollment, admissions, test scores, and much more. While some of this data overlaps with the College Scorecard, the IPEDS provides much smaller datasets and is friendlier to work with, so it was chosen for some of the data like enrollment statistics. One of the few issues with this system is that for enrollment numbers the methodology and process for recording enrollment was drastically overhauled starting in the 2020-2021 academic school year, so enrollment data from before 2019 could not be considered at a widespread scale for all universities or states. For this project the maintainers of the IPEDS were contacted about this discrepancy in the data but there was nothing that could be done to change it retroactively in the timeframe of this project. It is possible that Fall 2020 is a return to normal enrollment, but this could not be observed in this project due to a lack of complete data. This led to some struggles in the analysis process, but since the focus is on the impact of COVID-19 comparing 2020-2021 enrollment to 2021-2022 enrollment is sufficient for the scale of this project.

Dataset Wrangling

In preparing the data for this project, multiple datasets had to be wrangled to similar values in order to be merged together. Some datasets, like the College Scorecard and IPEDS datasets, this was a simple process due to sharing common identifiers like a university ID which designates the specific institution represented on a row. For other datasets, the process was not as simple and a large amount of modification had to be done to make the datasets compatible.

One common variable that had to be examined and wrangled was the variable representing the state in a dataset. For some datasets this would be the full state name, others the two-letter abbreviation, and some a state code. To join two separate datasets there had to be a common state value in order to merge the datasets, which required the creation of a custom csv file containing the full name, two letter abbreviation, and miscellaneous data about the state. First the original dataset would be merged with the custom dataset and then the variable names would be renamed to match the final dataset to be merged. This process had to be completed for each dataset that involved state level data to complete the visualization and analysis processes.

Another specific example of issues while data wrangling comes from creating a cohesive dataset for the time series for mental health indicators and COVID-19 data present in the final dashboard for this project. The first task was to calculate the total number of cases and deaths in the United States for each day in the dataset. This task was accomplished by grouping dates and then taking the sum of the cases and deaths for that date. The next step was to calculate the total for date ranges that lined up with the date ranges present in the mental health indicators dataset, a process that involves manually breaking by the date range intervals for a two-year span. Afterwards the mental health dataset could be wrangled to fit with the COVID-19 data by using responses that fall into a subgroup that corresponds with higher education demographics by age range or designating as having completed some college. From here, both dataset values were placed on a zero to one scale so the visualization could be more useful. Finally, the data could be merged for visualization.

In addition to this dataset, the final visualization includes vertical bars designating the average state facemask mandate start and lift date which had an entirely separate wrangling process. Specifically, the start date average was obtained by taking the average date for implementation of a face mask mandate and the lift date average was obtained by taking the average date for implementation for each phase of lifting mask mandates amongst all states. This process required pivoting the dataset to be longer as well as converting each date to its numeric POSIXct representation, calculating the mean, and then converting back to a date again. This is a process was repeated for each mandate and then averaged to get the United States average mask mandate. The key difference for each individual mandate is that both a numeric representation and a date representation were stored in the custom dataframe for different visualization purposes, and each visualization had to be renamed to the proper description of the mandate. An example of this is converting the column “FM_END” to “End General Mask Mandate” so the data was able to be properly interpreted in a visualization. One unique condition in the mandate dataset is that South Dakota never enforced any state mandates, instead opting for only releasing recommendations and guidance. The solution used for this project was to pull the average of recommendation and guidance dates to determine a start and lift time. These dates were pulled manually for this project to provide every state a value for the visualization. Overall the mandates are used in three different visualizations, so a large portion of the data wrangling was formatting the dates of each mandate to fit each visualization whether it be by taking averages, leaving it as a numeric date, or taking a subset of the mandates.

Chapter 4: FINDINGS, ANALYSIS, AND PRODUCT DEVELOPMENT

Technologies Used

R was chosen as the main programming language and basis for this project. The main reason to use R is that it is very potent for data exploration and visualization by providing a lot of tools to help simplify the process and provide more freedom due to the vast number of libraries available all with the goal of making visualizations simple. Another reason to use R is that the project focuses more on standalone analysis and processing so the more refined statistical tools in R can be used to help with the vast amount of data wrangling required to get the data ready for visualization as well as make a dashboard for the analysis to be presented and visualizations to have interactions with. Some of these tools are the potent libraries available in R like tidyverse for helping overall with data wrangling and visualization, lubridate for handling date data present in a lot of the datasets used in this project, leaflet for providing a base for interactive choropleth maps, and shiny for making it possible to take R visualizations and make them more dynamic on an interactive web app. One negative aspect that was considered with R was the data size limit, but the scope of the project and the amount of data looked at in research did not exceed the limit. This meant it was safe to use R.

In addition to the base language itself, the development environment was also important for a variety of reasons. RStudio is the main IDE for R and provides a lot of useful features for both development and deployment of the project such as making it easy to export a project for version control or quickly test code at multiple stages, so it was chosen to be the base for the project to run on. Because of the functionality provided between base R and RStudio the development and deployment environment could be the same which made the version control and overall handling of the project very simple.

Design and Implementation

The first step in the design process was to first bring in all of the data that would be used for visualizations and to determine what wrangling had to be completed for each visualization. This was completed using the “read_csv” method provided by the readr package, which comes packaged with the tidyverse library. Once read into R, the data could be mutated and manipulated to fit whatever visualization is desired. One example of this is from the state mandate dataset where in addition to handling the data to have a date or numeric representation for each mandate, the dataset was also pivoted to make it so that each mandate classification became a value in each row observation rather than a column for each state. Originally each column was a mandate and it had the date for each applicable state, but after pivoting the mandates became another value in a row that was paired with the date value for a state, making the table much longer as opposed to wider. The figure below gives a visual example where the left image was before wrangling and pivoting and the right image was after.

STATE	POSTCODE	FIPS	STEMERG
State	State Abbreviation	FIPS Code	State of emergency issued
category	n/a	n/a	state_of_emergency
type	note	note	start
unit	text	attribute	date
Alabama	AL	1	3/13/2020
Alaska	AK	2	3/11/2020

state	mean	Mandate	Date
Alabama	2020-07-27	End State of Emergency	2021-07-06
Alabama	2020-07-27	Reopen Day Cares	2020-05-23
Alabama	2020-07-27	End Stay at Home/Shelter in Place	2020-04-30
Alabama	2020-07-27	Reopen Businesses	2020-04-30
Alabama	2020-07-27	End Mask Mandate for Vax. People	2021-04-09
Alabama	2020-07-27	End General Mask Mandate	2021-04-09

Figure 1: State Mandate Pivoting (Left: Before, Right: After)

The process of wrangling and determining the necessary columns was completed one question at a time and was guided by the questions generated in the research portion of the project. After reading in the data and completing the wrangling through data mutation and manipulation as described in the dataset wrangling section, the next step was to begin creating the basic visualization that would be present in the final dashboard.

The process of visualizing is simple in R due to the power of ggplot2, a package in the tidyverse library. Ggplot2 is a versatile tool that gives the basic layout for column charts, scatter plots, line charts, and bar charts, which is a majority of what the visualizations for this project entails. Generally the versatility of ggplot2 made the visualization process for most of the questions manageable due to the great amount of customization that can be done for each individual plot.

One specific use case of ggplot2 in this project was for the mental health indicators line chart due to the complexities of plotting multiple graphs on the same plot. Ggplot2 allows for an accessible way to edit the details of a plot from the axis range to making a custom legend which was very useful as it remedied a broken legend when overlaying multiple plots on the same chart. In that visualization, the two vertical bars representing the average date of state face mask mandates placed and lifted were values calculated separately from the primary dataset used for the visualization, so the custom legend was made to make the visualization work properly. Unlike other visualizations in the project, the only part that could change is the date range or the vertical bars as a preset selection, so minimal editing had to be done to the visualization in order to make it fit with the interactive dashboard.

A majority of the other visualizations presented in the dashboard followed a similar process to the mental health indicators visualization of customization, but with an extra step. Each of the other plots included a subsection that could be broken down into for the interactive dashboard, so each individual setup had to be tested to confirm functionality. For example, the enrollment visualizations can change the scope of what is being looked at: all the states as a total, each state individually at the same time, or just one specific state. This allowed for greater ability to analyze the data of one state or all states for a direct comparison for the United States. In addition, the control version allows for further specificity by looking at each control type: public, private non-profit, and private for-profit. Each control type can be looked at individually if desired, or as a group for each state allowing for further analysis.

The state mandates graph also has a similar level of specificity by allowing for designation of a desired state as well as a mandate group: business related mandates or face mask related mandates. A unique issue presented in this visualization that had to be handled was if there were no mandates. South Dakota never released state mandates that fit the qualifications for the dataset used in this project so there had to be error handling for if there were no state mandates available to be plotted. This was handled in a dynamic manner that will be discussed with the server implementation but is an interesting issue that came up in the development process.

One plot that could not be completed with ggplot2 was the interactive choropleth map that showcases state-level COVID-19 data. For an interactive choropleth, leaflet was the best available solution. Leaflet is a JavaScript library that is able to be used within R for the purpose of interactive maps that can be paired with powerful tools and GeoJSON data to create much cleaner maps than any that are possible using the tools available in ggplot2 or other choropleth implementations. In addition to leaflet, Mapbox, a free map and location data provider used by major companies like The New York Times, Toyota, The Weather Company, and more was utilized as the tile provider for the visualization. With the customization options in leaflet this allowed for the ability to highlight specific states and popup key information. This is functionality that is doable to a lesser extent on its own with leaflet but greatly simplified and cleaned up when Mapbox is used as well. After making a free account on the Mapbox website an API key was given that allowed for the use of the enhanced map data. From there the choropleth could be made that colors by the average date each state lifted mandates and creates a popup that contains the total number of COVID-19 related cases, deaths, vaccinations, and the average state mandate lift date.

For the implementation of this project, all of the code for the dashboard is in one R script that can be divided into a few key points: reading in and wrangling data, establishing the user interface (UI) for the dashboard, establishing the server with reactive functionality to generate the visualizations, and the final piece to run all of the code and launch the dashboard. The entire script builds off of the shiny framework for a web application. The UI and server are functions that interact directly to pass information between the web app and the code to establish the choices available to the user and the styling of each visualization.

Establishing the UI was doable with Shiny as all the individual layers were very split and able to be parsed through. The page is established as a fluid page with a title panel, sidebar panel, and main panel. The sidebar panel controls the user input and interactivity, which allows for choosing which visualization to see as well as date range selection, selecting a state, or other variable changes that vary depending on the specific visualization selection. These are done using conditional panels to confirm which specific plot is selected and only show the appropriate fields for that plot. The date range input is a standard format selection that provides a calendar to select dates as well as allowing for the user to type the date in, and the state selection, mandate selection, and control selection all use a dropdown select from preset options for their user input. For the main panel, there is the leaflet choropleth map providing general state insights and then the bottom plot, the user selected plot, which is also paired with a description of the plot.

Creating the server function involved the creation of a reactive function named makePlot that generates the bottom plot depending on the user input. The makePlot function looks for the user inputs and then generates the plot according to the subset group of specifications for that plot. For example, on the enrollment by control plot the server first looks at what state value is selected, then what control type is selected, and subsequently properly filters the dataset to narrow down to the desired subset. This process occurs for each plot and eventually the final plot is returned from the function where the plot is rendered using renderPlot. While generating the plot, dynamic error handling had to be implemented to confirm that a non-empty plot was being created. In the example of state mandates, the data is filtered based on the user input and then checked to confirm whether or not there are any values to plot. If there are no values an error message is rendered, otherwise the plot is generated like normal. After rendering the plot, the choropleth is rendered the exact same way every time the reactive section executes by using renderLeaflet on a leaflet plot with Mapbox tiles as described earlier. Finally, the description for the bottom plot is pulled from a csv stored in the data folder and rendered as text at the bottom of the page.

Testing was completed with three main approaches: regression testing, alpha testing, and beta testing. Regression testing was performed after each update to all functionality to make sure no previous functionality was broken with the implementation of new functionality. Alpha testing was performed as the project nears completion to determine the ease of use for users by navigating the dashboard and selecting all the options to confirm simplicity. Beta testing was implemented after alpha testing to confirm alpha testing results by telling users to complete basic navigation of the dashboard such as selecting a visualization and date range to observe behaviors and actions and make sure the dashboard is user friendly and simple to navigate and use. The majority of testing went smoothly due to the back-end nature of the project, and the only issues present were known issues at the time of testing that were later fixed or unique edge cases that attempted to generate a plot with no data points. One such example was with the state mandates plot, attempting to specify a state and date range where that state did not lift any mandates would cause an error when generating the plot and as such had to have unique error handling implemented.

While the development and implementation of the project ran into setbacks due to issues with publicly available data and a general lack of time, development largely went smoothly due to a steady process by first generating questions and finding data, determining the desired visualizations, creating the visualizations independently, and finally binding everything together into the final build with a web application to generate interactive visualizations. Overall, five main visualizations were produced for observing the effect of COVID-19 on higher education and were used to see what areas were impacted the most based on the publicly available data at the time.

Findings and Analysis

Each visualization created for this project was able to provide some insights or reaffirm some common beliefs around COVID-19 and its impacts, and this section will discuss those for each visualization.

The choropleth created for a state view of COVID-19 provides the average state mandate lift date for each state, as well as total state information about cases, deaths, and vaccinations related to COVID-19. This visualization was able to confirm that states like California and New York were more reluctant to remove mandates, which when looking at enrollment between 2020 and 2021 aspects like mandates in place could impact enrollment for certain individuals.

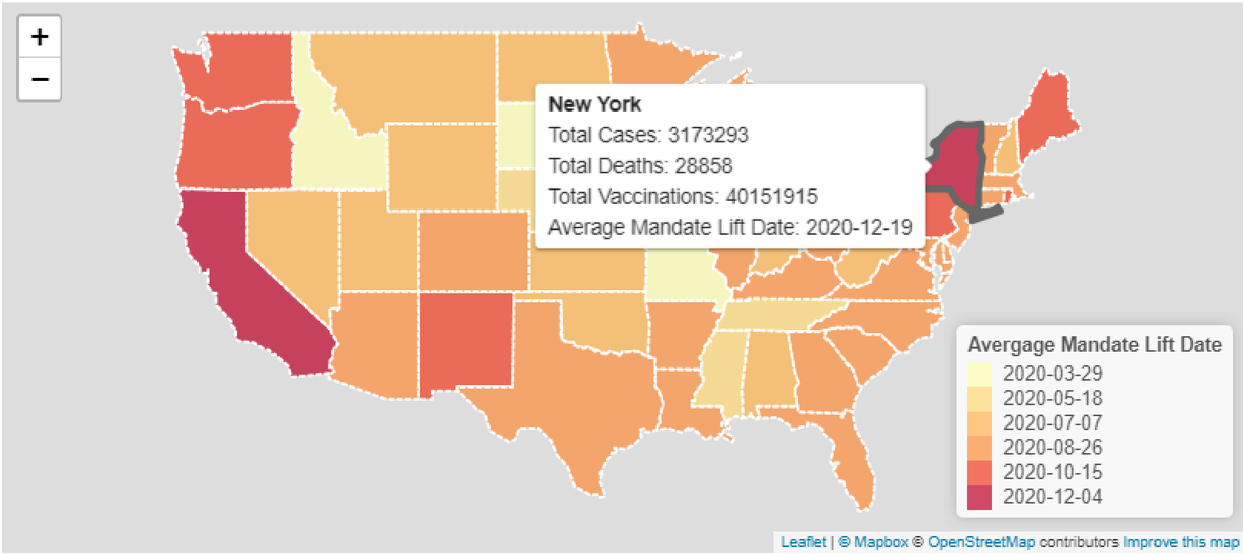

Figure 2: State View of COVID-19 Choropleth

The enrollment visualization tells a different story than the choropleth as California and New York's decrease in enrollment per capita was relatively similar to that of most other states, who almost universally saw a decrease in enrollment. The notable exceptions to this are New Hampshire and Utah, who have an extremely high enrollment per capita that actually showed a substantial increase in enrollment between 2020 and 2021.

Figure 3: Total Higher Education Enrollment 2020-2021

Utah and New Hampshire had a relatively regular average mandate lift date with Utah's average being 07-23-2020 and New Hampshire's being 08-16-2020, so something else caused the drastic increase in enrollment when everyone else decreased. Part of this can be seen by breaking down the total enrollment per capita into subsets based on the control of the university. Here a clear substantive increase can be seen in private higher education enrollment for New Hampshire in particular, which Utah shows a smaller but still large growth compared to almost every other state in the United States. This is the only control type to show such a drastic increase in enrollment as the public enrollment largely mirrors the total enrollment for most states and for-profit institutions only saw a relatively small increase in West Virginia and Arizona.

Figure 4: Higher Education by Control 2020-2021: Private Non-Profit

Figure 5: Higher Education by Control 2020-2021: Public

Figure 6: Higher Education by Control 2020-2021: Private For-Profit

These visualizations showcase that major private universities in Utah and New Hampshire were the major contributors to increasing enrollment. For both states there are notable private universities that could contribute to this increase, Brigham Young University in Utah and Dartmouth College in New Hampshire. Brigham Young University had a total enrollment of 34,802 for 2021 while Dartmouth College had a total enrollment of 6,761. There is a large population disparity between these two universities, but these differences are accounted for by looking at the enrollment per capita as Utah has a population of 3.338 million and New Hampshire has a population of 1.389 million. The overall smaller population makes the growth have a larger impact per capita there than in states like California, a state with a population of 39.24 million. Despite these circumstances it is still impressive to see such an increase for Utah and New Hampshire in particular amongst an overwhelming majority of states who showed decreased enrollment per capita. One potential explanation could be that students who normally would leave the state for school instead needed to be close to home and went in-state, but without data for in-state and out-of-state enrollment this cannot be determined. With that said, on the whole private non-profit universities seem to be the schools least affected by the pandemic while private for-profit and public universities especially saw a decrease in enrollment due to COVID-19.

The choropleth map, in addition to looking deeper into enrollment per capita, led to looking at individual mandates and states. The choropleth shows that the west coast and northeast regions were generally later than the South or Midwest of the United States when it comes to lifting state mandates. The most shocking part of this is that some of the earliest mandate lifts came as early as April 2020. By looking at the average for all states of each individual mandate, some of the reasoning for the early average per state can be seen. When broken down to just business mandates, the vast majority had an average lift date before July of 2020, and these mandates were a majority of what made up the average mandate lift date for each state so they brought the date earlier. When looking at mask mandates they also generally fall within the first half of 2021, which helps to balance out so that most states have an average mandate lift date between July and September of 2020.

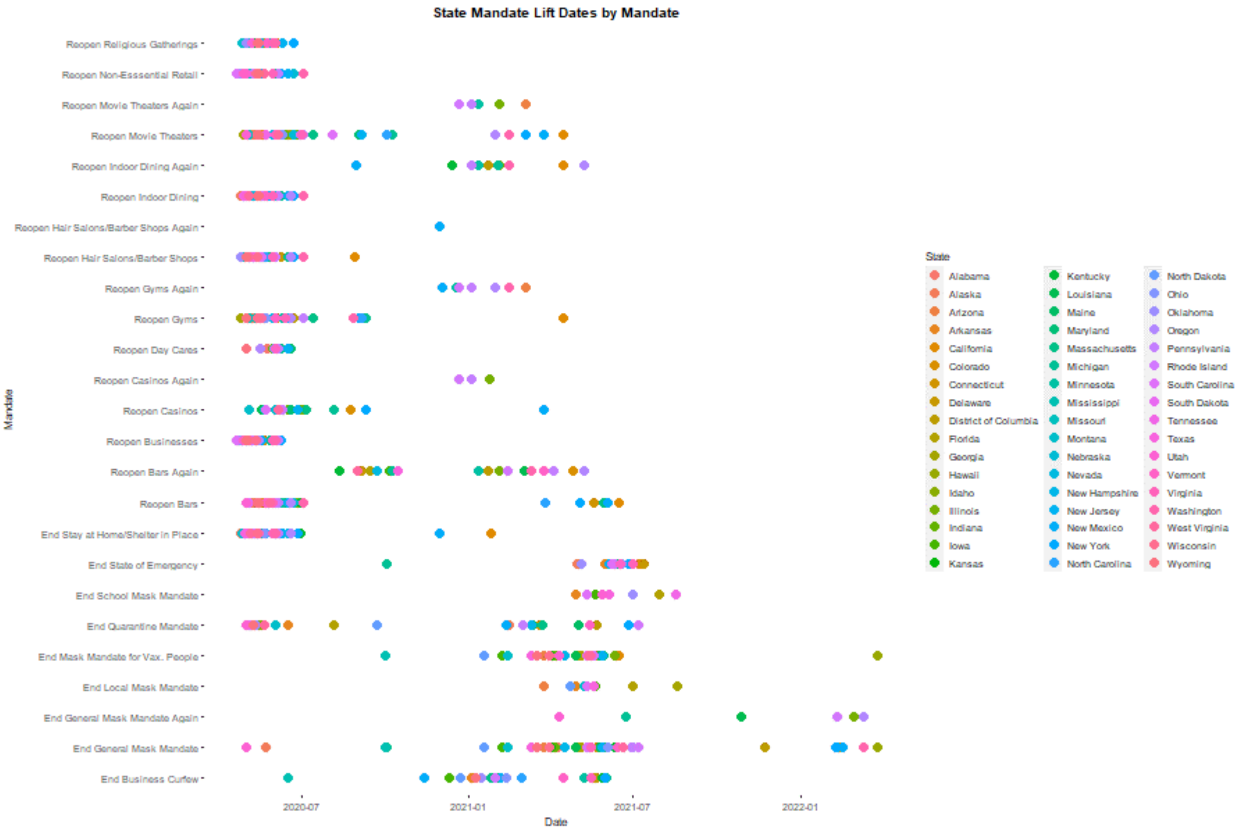

Figure 7: State Mandate Lift Dates by Mandate

When specifically looking at the states with the earliest average there are some interesting points to be made. South Dakota never officially released any state mandates, only guidance and recommendations, so they are not present on the scatterplot and their average is based on guidance given by the state. Other states like Idaho and Missouri who lifted on average in May 2020 lifted almost all of their business mandates at the same time and never instituted mask mandates, so their average was far earlier than states that would implement mask mandates.

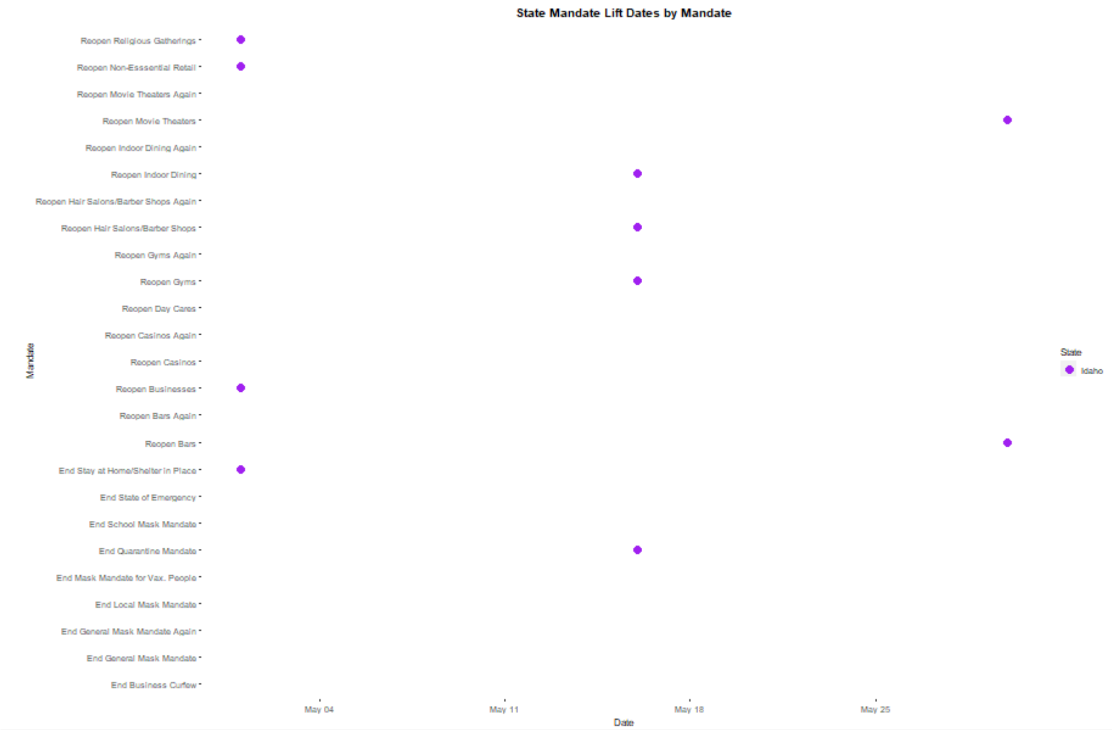

Figure 8: State Mandate Lift Dates by Mandate - Idaho

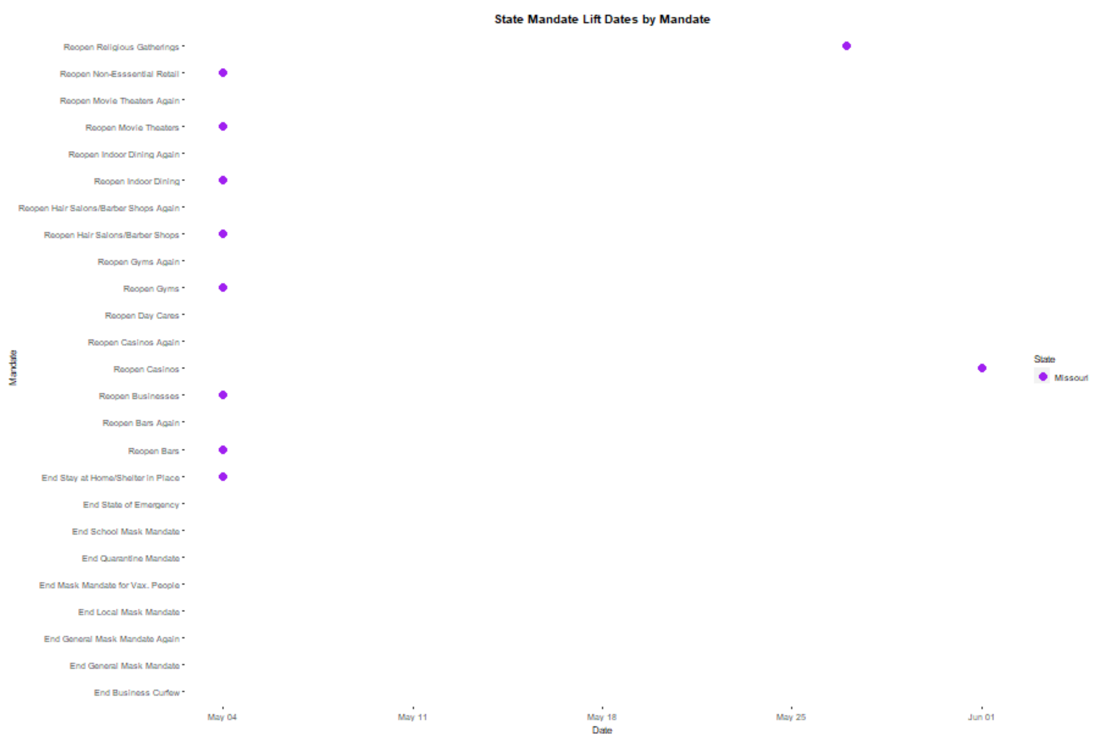

Figure 9: State Mandate Lift Dates by Mandate - Missouri

One final factor to be considered with the visualizations created for this project is how mental health fared for those in higher education demographics. The visualization looks at how the mental health indicators reported to the CDC fluctuated throughout the pandemic, and the plot is overlaid with COVID-19 cases, deaths, and vertical bars that show the average mask mandate lift date for the United States as these were determined to have the largest impact on the higher education sphere. All of the data was scaled to fit on a zero to one scale to all be visible on the same visualization, and this helps to clearly see how the mental health indicators responded to the trends seen in COVID-19 cases.

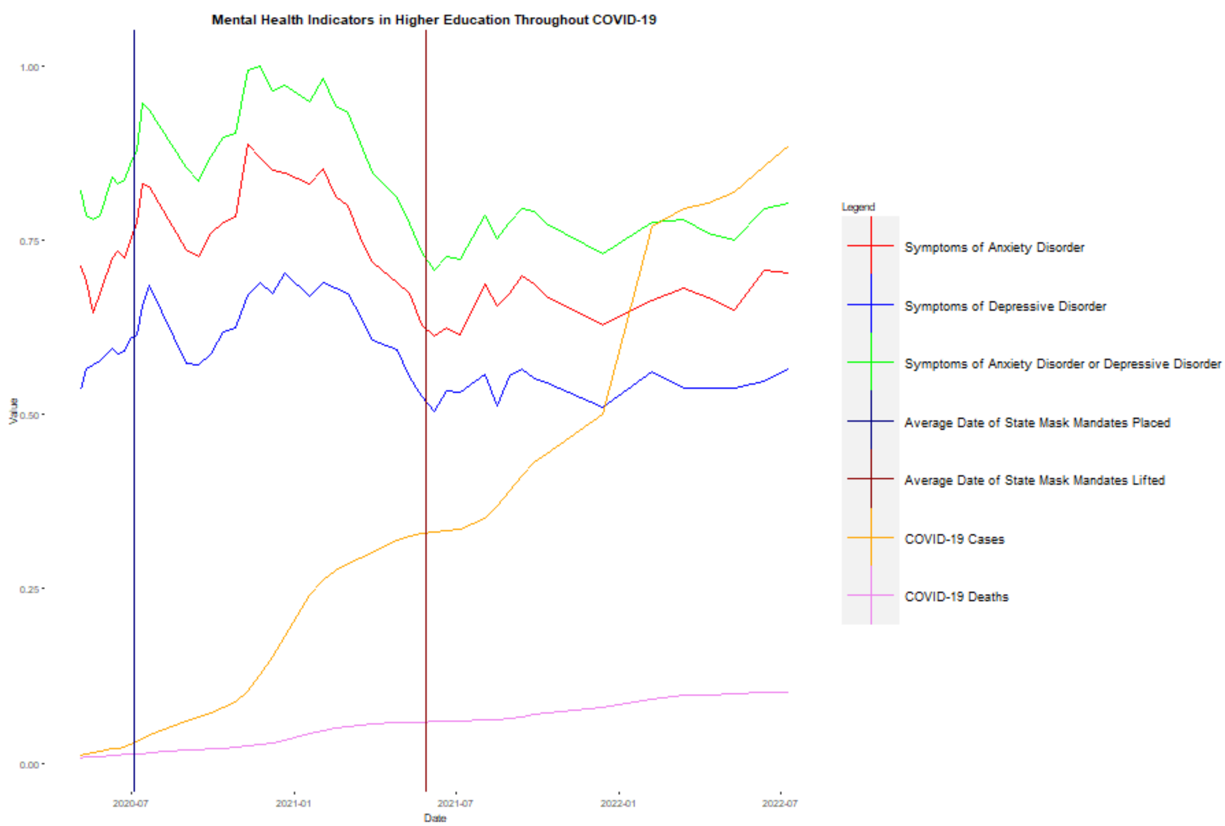

Figure 10: Mental Health Indicators in Higher Education Throughout COVID-19

The most notable finding from this visualization is that mental health indicators for anxiety order were always higher than depressive disorder, and that the peak of mental health indicators is encapsulated entirely by the placement and lifting of state mask mandates. This is the time frame of July 2020 to June 2021, which is largely considered to be the entirety of the first wave of COVID-19 and part of the second wave. This logically makes sense as this was the start of a worldwide pandemic that led to every facet of life being closed off or drastically adapted to follow mandates placed by states and guidance by the CDC. The time was filled with a large amount of uncertainty for students in higher education specifically due to a drastic shift in the methodology of learning and location. These rapid changes were shown in research to cause anxiety symptoms like trouble concentrating or having a feeling of nervousness, restlessness, or uneasiness. In addition the isolation caused by the pandemic could be amplified in students as social activities were completely halted as many universities forced students out of dormitories. This had a decent probability of leading to increased loneliness which in-turn could lead to mental health indicators for both anxiety or depressive disorders as the individual's mood would be affected by world events and a lack of social interaction.

These explanations also help make sense of the decrease in mental health indicators as the pandemic went on. People were able to find and utilize alternative methods of social interaction online or in-person with social distancing as state mandates became less and less restrictive. This culminated in the state face mask mandates being lifted, which was a shift to more normalcy for many people. While universities like the University of Mississippi continued to enforce a mask mandate on-campus and the host city of Oxford, MS did as well, they were far less restrictive than early in the fall 2020 semester when students first came back to campus. This shift in mandates and guidance allowed for the social interaction missing during the peak of the first wave and mental health indicators.

The mental health indicators do not appear to be directly related to the number of COVID-19 cases or rate of infection and are instead based upon the understanding of how the cases impact the world of higher education students. This is evidenced by the extreme jump in cases that began at the start of 2022, the start of the most drastic increase in COVID-19 cases due to the Omicron variant. Students' mental health indicators did not show a drastic shift like they did in response to the first massive increase from the Alpha variant in early 2020. The research done for this project leads to the conclusion that there was no massive swing in mental health indicators at this point since a majority of states had lifted mandates related to COVID-19. Because state mandates had been largely lifted, many more social activities were able to be participated in and higher education institutions were operating relatively similarly to before the pandemic with the major differences being mask mandates for some institutions, generally smaller class sizes, and an increased presence of online classes for students to take. Since the day-to-day events for students were not impacted by the sharp increase in cases like it was during the first wave of the pandemic, mental health indicators did not reflect the massive spike in cases like they did during the first wave of the pandemic.

Chapter 5: CONCLUSION

This project took a variety of the currently available data related to COVID-19 and tried to find key insights on how higher education was impacted by the pandemic. With the results of the project more insight has been gained into the effects that COVID-19 had on higher education from an empirical perspective, specifically by looking into mental health and enrollment data from 2020 and 2021.

The results from the data used in this project suggest that mental health indicators for higher education demographics do not generally follow the state of COVID-19 case trends, but rather were determined by circumstances of the start of the pandemic. This was evidenced by the sharpest increase in cases throughout the pandemic, early 2022 with the Omicron variant, not being matched with an increase in mental health indicators proportionally similar to the first wave of the pandemic. State mandate data was also observed in relation to mental health indicators, and it was found that mental health indicators were at their highest when mask mandates were being the most strictly enforced and saw a decrease as mask mandates began to be lifted and social activities were able to open up more.

In regard to enrollment data, the results showed that overall enrollment in higher education saw a substantial decrease in the United States, but private non-profit schools were not impacted like public schools. Some of the largest non-profit schools actually saw a large increase in enrollment per capita compared to other types of universities. Due to this the project concludes that an immediate impact from COVID-19 can be seen in a vast majority of universities in the United States. More enrollment data is needed for future years to see if there is an extended decrease in enrollment from the pandemic, but it is extremely unlikely.

The evaluation methods used in this project were able to provide explanations for phenomena observed throughout the COVID-19 pandemic and can be used to be better informed about how higher education students responded to mostly unprecedented world events. This achieves the primary goal of this project, to complete a study on COVID-19 and its impacts on higher education that is based on data analysis rather than survey results, helping to contribute to the discourse on the topic and provide value. The results tend to show that there will be some lasting impacts from the COVID-19 pandemic, but for future decision making more information that is not currently publicly available is necessary.

Chapter 6: FUTURE WORK

With the goal of future decision making, one improvement in future research would be finding a way to implement previous aggregate enrollment data from before 2019 and include enrollment data from 2022 and beyond. The addition of this enrollment data would help to provide a better understanding of how drastic the decrease seen in enrollment from 2020 to 2021 was relative to other major events that led to a decrease in enrollment across the United States.

One concern with research on higher education and COVID-19 is that not many surveys were conducted after the first wave of the pandemic, so if there was research to form after the pandemic that asked participants to reflect on their experiences it could be beneficial to compare those results to that of the first wave studies. Looking into measurements of success within students who graduated during the pandemic could also be an indicator of long-lasting impacts from the pandemic.

For improving research there is potential to increase other metrics to observe the difference in education before and after COVID-19. One such example would be to look at engagement data with higher education tools, websites, and services. This aspect was considered in the research for this project, but analysis could not be completed due to time constraints. In addition, looking at other metrics of university success such as student retention, graduation rates for students who began school during the pandemic, grades for specific disciplines within universities, and more could be beneficial but are not publicly available or aggregated.

Another concern with the results of this research is that it could not be as beneficial for future decision making. Future research would attempt to find new data from a variety of fields that come together to prove or disprove the long-term effects of COVID-19 expounded upon in this project. In addition, research could look to provide guidance for future decision making if people find a similar situation arises or there is a need to look at what the key impacts were and what the most effective strategies employed during the pandemic were. The research could also be more applicable in future research to narrow down to a specific university for all the metrics used to observe a case study of how one particular higher education institution reacted to the pandemic, looking specifically at where it succeeded and where it struggled.

REFERENCES

Aristovnik, A., Keržič, D., Ravšelj, D., Tomaževič, N., & Umek, L. (2021). Impacts of the Covid-19 Pandemic on Life of Higher Education Students: Global Survey Dataset from the First Wave. Data in brief, 39, 107659. https://doi.org/10.1016/j.dib.2021.107659

Bao, W. (2020). Covid -19 and Online Teaching in Higher Education: A Case Study of Peking University. Human Behavior and Emerging Technologies, 2(2), 113–115. https://doi.org/10.1002/hbe2.191

Bravo-Adasme, N., & Cataldo, A. (2022). Understanding techno-distress and its influence on educational communities: A two-wave study with multiple data samples. Technology in Society, 70, 102045. https://doi.org/10.1016/j.techsoc.2022.102045

Chaturvedi, K., Vishwakarma, D. K., & Singh, N. (2021). COVID-19 and its Impact on Education, Social Life and Mental Health of Students: A Survey. Children and Youth Services Review, 121, 105866. https://doi.org/10.1016/j.childyouth.2020.105866

Data | Centers for Disease Control and Prevention. (n.d.). Data.CDC.Gov. Retrieved August 28, 2022, from https://data.cdc.gov/

Johnson, N., Veletsianos, G., & Seaman, J. (2020). U. S. Faculty and Administrators' Experiences and Approaches in the Early Weeks of the Covid-19 Pandemic. Online Learning, 24(2), 6–21. https://eric.ed.gov/?id=EJ1260365

Learnplatform covid-19 impact on digital learning. (n.d.). Retrieved September, 2022, from https://kaggle.com/competitions/learnplatform-covid19-impact-on-digital-learning

Raifman, Julia, Nocka, Kristen, Jones, David, Bor, Jacob, Lipson, Sarah, Jay, Jonathan, … Galea, Sandro. COVID-19 US State Policy Database. Ann Arbor, MI: Inter-university

Consortium for Political and Social Research [distributor], 2021-03-29. https://doi.org/10.3886/E119446V75

Sahu, P. (2020). Closure of universities due to coronavirus disease 2019 (COVID-19): Impact on education and mental health of students and academic staff. Cureus, 12(4). https://doi.org/10.7759/cureus.7541

State-by-state data on COVID-19 vaccinations in the United States. (n.d.-b). Our World in Data. Retrieved October 3, 2022, from https://ourworldindata.org/us-states-vaccinations

Toquero, C. M. (2020). Challenges and opportunities for higher education amid the covid-19 pandemic: The philippine context. Pedagogical Research, 5(4). https://eric.ed.gov/?id=EJ1263557

Viner, R. M., Russell, S. J., Croker, H., Packer, J., Ward, J., Stansfield, C., Mytton, O., Bonell, C., & Booy, R. (2020). School closure and management practices during coronavirus outbreaks including COVID-19: A rapid systematic review. The Lancet Child & Adolescent Health, 4(5), 397–404. https://doi.org/10.1016/S2352-4642(20)30095-X