Milestone 4: Data Critique

A5: Chinmaya V, Christy Y, Daniel J, Muthu P, Simon Z, Zack P

In our research, we analyzed a dataset on COVID-19 cases amongst colleges across the U.S., as well as a dataset on COVID-19 cases and deaths at prison facilities across the U.S. Both were published by and collected from the New York Times.

### **DATASET A: COVID-19 DATA ON COLLEGES** <https://github.com/nytimes/covid-19-data/blob/master/colleges/colleges.csv>

### What are the original sources?

1. The New York Times began surveying American colleges and universities for their article *Tracking Coronavirus Cases at U.S. Colleges and Universities* which visualizes percent changes in COVID-19 cases between 2020 and 2021 within colleges due to the infection outbreaks in college communities that outpaced the national average at the start of the pandemic. The content types and the article focus on a summary view of data, highlighting comparisons of 2021 case counts to 2020 case counts within each institution individually.

### How was the data generated?

1. Since late July, the New York Times has been conducting a rolling survey of COVID-19 cases among students and employees of universities across the nation, including every four-year public institution and every private college that competes in N.C.A.A. sports. However, with no national standard, colleges determined their own rules and practices for how to tally infections. Case numbers mostly originate from publicly published data, open record requests, and responses to New York Times’s surveys. However, some institutions declined to provide information or only provided partial information. Refer to the Processing section of the About page to learn more about how we remedied some of the shortcomings in this dataset.

### What information is included in your dataset?

* 1. Each entry has identifying attributes of **college**, which represents the name of the college, as well as **ipeds\_id**, which is the Integrated Postsecondary Education Data System ID number for each college. The dataset also includes a **date** attribute which describes the date of the last update. Every value for the date column is ‘2021-04-28’ which indicates that the dataset is well kept and updated consistently. There are also location-related attributes of **state**, **county**, and **city** of where each college resides. Most importantly, there is the attribute **cases**, which represents the total number of reported confirmed positives and probable COVID-19 cases among university students and employees, as well as **cases\_2021**, which represents the number of reported confirmed of reported confirmed positives and probable COVID-19 cases among university students and employees starting January 1, 2021. Lastly, there is the **notes** attribute which is used to indicate whether there was a medical unit associated with the institution which contributes to the case counts.

### What information, events, or phenomena can your dataset illuminate?

1. This data all pertains to the pandemic that began early 2020 and continues up to April 28, 2021. The information here specifically pertains to 1,949 colleges and universities across the United States, including every public 4-year institution and every private school that competes in NCAA sports. Given this, we can likely glean information related to how schools in different regions fared during the pandemic and get an idea how these individual institutions responded to the crisis. In conjunction with our dataset on COVID-19 cases in prisons, we can compare and contrast the distinctions of preventative measures taken between the college and prison systems.

### What can your dataset not reveal?

1. Our dataset does not illuminate whether each college campus operated fully in-person, fully remote, or a hybrid of the two. In addition, the case data is presented in summary form only, providing 2020 cases vs 2021 cases until April 28th, 2021. Also, the summary numbers do not indicate how many cases stem from students, employees, or medical professionals when applicable. In addition, the dataset lacks information on the financial health of the colleges or the population density of their respective communities, which significantly impacts an institution’s ability to control cases on campus. Many college reports also do not explain whether repeat cases (people who are infected multiple times) are counted as one or multiple instances. Also, the dataset’s authors at New York Times mention some colleges subtract the number of recoveries from their total case count, but the dataset does not reveal which colleges do so. Furthermore, health outcomes of recoveries and deaths are not reported, so we are unable to deduce how effective healthcare at colleges are.

### What information is left out of the spreadsheet?

1. Overall, because the dataset only offers a summary view of the COVID-19, it lacks a more personal aspect of understanding individual situations of the people who work or attend each college. For example, it’s unclear to what extent students and employees are satisfied with the level of sanitation protocol and healthcare at each institution. Also, since the data is generalized into aggregate counts, it is impossible to determine if there is a trend in infection based on personal ideologies, race, gender, age, or wealth across colleges that is true for prison facilities as well and therefore more indicative of health outcomes than prison or college environments. Furthermore, comparing personal accounts of living situations and changes during COVID-19 between colleges and prisons may be even more illuminating than numbers.

### Give your account of the ideological effects of the way in which your sources have been divided into data (your dataset’s ontology). If your dataset were your only source, what information would be left out?

1. Although the New York Times is a fairly reliable and neutral source, due to the lack of national standard in reporting cases, the process of data generation introduces a large potential for method bias. By choosing to publish infection counts or respond to a newspaper company with global readership, the institutions may realize that their response or published data may sway their public image. For example, higher case counts reflect poorly on the institution’s ability to implement and maintain health and safety protocols, provide adequate health care for their students and employees, and adapt to the pandemic overall. Therefore, it is likely that colleges choose definitions for what counts as a case and for how to consider repeat cases and recoveries using methods that keep their total case count as low as possible. Additionally, schools that are better funded with more access to resources and schools that have more scientific beliefs on the virus may have opted to test more students for COVID-19 in comparison to poorer schools or schools that are skeptical of the pandemic.

### 

### **DATASET B: COVID-19 DATA ON PRISON FACILITIES**

<https://github.com/nytimes/covid-19-data/blob/master/prisons/facilities.csv>

### What are the original sources?

1. The New York Times sought to collect data on prison populations because they are a particularly vulnerable group of people who are at a far higher risk of contracting COVID-19 virus than the general public. Their collected data was then used for their interactive article, *Incarcerated and Infected: How the Virus Tore Through the U.S. Prison System*, which compares the much higher infection rate in prisons compared to the national and global rate, as well as recounts personal stories of people who lived and died in prison facilities during the pandemic. The content types and the article focus on a summary view of data, reporting only annual totals. However, a distinction is made between officer and inmate cases and deaths in the dataset, which reveals a motivation by the New York Times to analyze these groups’ data in isolation and to compare their data.

### How was the data generated?

1. From March 2020 until the end of March 2021, the New York Times collected data about COVID-19 infections and related deaths at a total of 2,805 prison facilities including, state and federal prisons; local, regional, and reservation jails; immigration detention centers; juvenile detention facilities; and those under custody of the U.S. Marshals Service. Some data was published on websites managed by state and federal prison systems as well as the Immigration and Customs Enforcement. However, since there was no uniform national reporting system for COVID-19 in correctional systems, some facilities stopped releasing data without explanation. When not publicly available, the New York Times relied on direct inquiries, public records requests, coroners’ reports, and numbers announced at news conferences and meetings of state or county officials, while always defaulting to the most complete available data. Refer to the Processing section of the About page to learn more about how we remedied some of the shortcomings in this dataset.

### What information is included in your dataset?

1. The dataset includes identifier attributes of **nyt\_id** and **facility\_name** to uniquely identify each data entry. There are also location-specific attributes including **facility\_city**, **facility\_county**, **facility\_county\_fips**, **facility\_state** that indicates the city, county, county FIPS code, and state of each facility, as well as **facility\_lng** and **facility\_lat** which represents the longitude and latitude of each facility respectively. Furthermore, there is a **facility\_type** attribute that describes the type of prison facility. Most entries are listed as state prisons and state jails, though other facility types include state rehabilitation center, state work camp, state juvenile detention, low-security work release, state halfway house, federal halfway house, detention center, and federal prison. There is also population data of each facility with **latest\_inmate\_population**, representing the most recent number of inmates at the facility, as well as **max\_inmate\_population\_2020**, representing the highest reported occupancy at the facility, which indicates that the dataset takes into consideration the constant release and transfer of inmates from March 2020 through March 2021. Most importantly, there are the COVID-19 specific attributes **total\_inmate\_cases**, **total\_inmate\_deaths**, **total\_officer\_cases**, and **total\_officer\_deaths**, which aggregates the total count of confirmed COVID-19 infections and deaths since the beginning of the pandemic through the end of March 2021, distinguishing between inmates and correctional officers at each facility. The distinction between officer and inmate cases indicates an attention to detail for detailed, comprehensive data since although both groups interact in similar environments, they experience varying degrees of freedom, movement, resources, and personal space to adapt to the necessary social distancing to protect one’s health. Finally, there is a last attribute **note** which indicates any notes for outlying interpretation of data on the facility, including state facilities that housed inmates from other states.

### What information, events, or phenomena can your dataset illuminate?

1. This dataset mainly focuses on the death and cases of a total of 2,640 facilities within the State Prison System, and it also splits death and cases into two categories: inmate and officer. So maybe we could investigate the relationship between these two groups. In addition, we are likely to gain the geographical statistics insight from this dataset.

### What can your dataset not reveal?

1. The latest inmate population does not always reflect the true population at the time. In addition, the case data is presented in summary form, only providing total numbers of cases and deaths for inmates and officers that occurred anytime between March 2020 through the end of March 2021, without detailing the timeline of when the cases or deaths occurred (spread out over the year vs. concentrated within a specific timeframe). Also, inmates reported that some who had COVID-19 symptoms were never tested due to lack of testing capabilities, especially at the start of the pandemic. People who were infected with COVID-19 virus in multiple instances are regarded as one case. Furthermore, if an infected inmate was transferred, the infection was only recorded at the institution of origination, not at the destination institution.

### What information is left out of the spreadsheet?

1. The summary numbers provided by the dataset do not paint a complete picture of what it was truly like to work and live in a correctional facility during a global pandemic. For example, personal accounts, images, or videos of the sanitary conditions as well as any evidence of new health and safety protocols that were introduced and effectively practiced at each facility would help to illuminate the conditions and potential hazards of interacting within a prison facility. Furthermore, it would be more eye-opening to understand the availability of essential health resources such as masks, hand sanitizer and soap, and infection testing, as well as whether ease of access varied between officers and inmates.

### Give your account of the ideological effects of the way in which your sources have been divided into data (your dataset’s ontology). If your dataset were your only source, what information would be left out?

1. While New York Times is a rather reputable and neutral newspaper, the step of data generation is likely to be influenced by bias and personal ideologies. For example, prison facilities in more conservative areas are more likely to regard and prioritize inmate health less. As a result, it is possible that inmate case counts for those facilities are much greater of an undercount than their officer counterparts because officers were prioritized higher when resources and testing were limited. Also, if the New York Times had asked more questions in their surveys to retrieve a more comprehensive picture of the pandemic within each facility, they risk less facilities willing to participate and take the time to respond to their surveys, thereby missing more information, especially because their alternative methods of data collection were much more involved and laborious.

In addition to the two primary datasets on COVID-19 cases in colleges and prison facilities, we also used an additional supplemental dataset for one data visualization, similarly published by the New York Times at [U.S. States’ Live COVID-19 Data](https://github.com/nytimes/covid-19-data/blob/master/live/us-states.csv). It is frequently updated and records the most recent number of COVID-19 cases and deaths by state in the U.S. This third dataset was used in order to compare COVID-19 case and death rates of each college and prison institution with the general population for each institution’s state.