Milestone 4: Data Critique Guidelines

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

**Data Set A**: <https://github.com/nytimes/covid-19-data/blob/master/colleges/colleges.csv>

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

# What information is included in your dataset?

From the README listed with the CSV, The New York Times says:

1. Features:
   * **date**: The date of the last update.
   * **state**: The state where the college is located.
   * **county**: The county where the college is located.
   * **city**: The city where the college is located.
   * **ipeds\_id**: The Integrated Postsecondary Education Data System (IPEDS) ID number for the college.
   * **college**: The name of the college or university.
   * **cases**: The total number of reported Covid-19 cases among university students and employees in all fields, including those whose roles as doctors, nurses, pharmacists or medical students put them at higher risk of contracting the virus, since the beginning of the pandemic.
   * **cases\_2021**: The total number of newly reported Covid-19 cases since Jan. 1, 2021 only. notes: Specific methodological notes that apply to the institution, for example if the count includes cases from a medical unit, and if there is a possibility that duplicate cases have been counted due to the manner in which the institution reports data.
2. Features:
   * **nyt\_id**: A unique identifier we may use to match future data to this data set.
   * **facility\_name**: The name of the facility.
   * **facility\_type**: The type of facility.
   * **facility\_city**: The city where this facility is located.
   * **facility\_county**: The county where this facility is located.
   * **facility\_county\_fips**: The county FIPS code of the county.
   * **facility\_state**: The state where the facility is located.
   * **facility\_lng and facility\_lat**: The longitude and latitude of the facility.
   * **latest\_inmate\_population**: The most recent number of inmates at the facility.
   * **max\_inmate\_population\_2020**: The maximum number of inmates at the facility reported at any time from March 2020 through March 2021.
   * **total\_inmate\_cases**: Total number of cases of Covid-19 reported among inmates from the beginning of the pandemic through the end of March 2021.
   * **total\_inmate\_deaths**: Total number of inmates who were reported to have died of Covid-19 from the beginning of the pandemic through the end of March 2021.
   * **total\_officer\_cases**: Total number of cases of Covid-19 reported among correctional officers working at the facility from the beginning of the pandemic through the end of March 2021.
   * **total\_officer\_deaths**: Total number of correctional officers who worked at the facility who were reported to have died of Covid-19 from the beginning of the pandemic through the end of March 2021.
   * **note**: Any notes important for the interpretation of data on this facility.

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

1. This data all pertains to the pandemic that began early 2020 and continues now. The information here specifically pertains to colleges and universities and in the United States. 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 second data set, we can compare and contrast the distinctions of preventative measures taken between the college and prison systems.
2. This data set mainly focuses on the death and cases in the State Prison System, and it also split 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 data set.

# What can your dataset not reveal?

1. Our data set ignores the fact that a large portion of college students studied remotely. Some sources repeat cases, so we are unsure which cases are unique and which are repeated. Can you think of a reason why a case may be repeated versus having each entry be a unique case? What is your approach to address this issue?
2. We can’t determine the timeline of each facility, which means the trend of the case increasing or decreasing is hard to tell. Can you be a bit more detailed and specific here about the data you have and don’t have? You have case data presented in summary form only, correct? The data does not provide the individual records that make up the case totals and therefore it’s impossible for your project to consider case numbers over time for this dataset because there is not time information present.

# How the data was generated?

1. The colleges self-reported these cases; the cases cover those of students, faculty, staff members and other college workers. Since there is no national standard to report cases, the NYT adds that colleges may report their findings in different ways, so it is safe to exercise caution when comparing them.
2. 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.

# What are the original sources?

1. The New York Times has been surveying American colleges and universities — including every four-year public institution and every private college that competes in

N.C.A.A. sports on a rolling basis.

1. When data was not publicly published, New York Times relied on direct inquiries, public record requests, coroners’ reports, and news conferences and meetings of county or state officials for infection cases and mortality statistics, always using the most complete data available.

# What information is left out of the spreadsheet?

1. We are unsure about the health outcomes after being diagnosed with COVID, so we are unable to deduce whether colleges are actually providing better healthcare compared to prisons. We also don’t have information on the death toll from COVID cases.
2. The latest inmate population does not always reflect the true population at the time. In addition, the spreadsheet does not detail the timeline of when the cases or deaths occurred (spread out over time vs. concentrated within one month). 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, and inmates with multiple infections are regarded as one case. Furthermore, if an infected inmate was transferred, the infection was not recorded at the destination institution.

# 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. We think that the NYT is fairly neutral because they conducted surveys. However, ideologies may have played a role in how individual schools may have reported cases. For instance, a school that may have more scientific beliefs about COVID may have opted to test more students in comparison to a school that is skeptical of the pandemic.
2. While the data collection methodology relied on fairly reputable and neutral sources, the data generation step is more likely to be influenced by ideologies. Prison facilities in more conservative areas are more likely to regard and prioritize inmate health less. As a

result, it is possible that officer case counts are less of an undercount because they were prioritized higher when testing resources were limited.

As DHers, be want to be aware that there’s always a level of implicit bias that we need to content with when data is assembled: <https://www.jstor.org/stable/20439056>. This is we why we consider the ontology of the dataset. What perspectives are embedded in the way the data is collected and assembled? In the case of self-reported data, there are even more factors to consider: <https://www.jstor.org/stable/20650304>. You’ve started to unpack this a bit here, but what are additional considerations? Would it take longer or be more difficult for NYT to ask for more information? Would they risk people participating? Can they get that comparative information elsewhere? In that case, what does the information collected tell you about what they thought was most important?

Good work here – do think about how you would rearrange this information for your website. You will want to walk your reader carefully through what data you have and how you’ve transformed it. Think of the 3 parts of a DH project sources, process, and presentation. The data critique should provide your reader with a good understanding of where the data came from, but also how you transformed it based on your own project goals. There’s not much here on how you are going to manage the “messiness” of your data – think about that and make sure you share those transformations explicitly in your final data critique. Also, think carefully about how you are framing bringing these two data sets together. These are two very vulnerable populations, but they are also very different cultures. How might that weigh on the data and your interpretation of it?

Preliminary grade: 85/100