Lab 3 - Context and Historical Background

Name

Date

Instructions and Overview

This week you are going to do some background research on the dataset that you will be working with throughout the quarter - assessing how the data you plan to analyze got (or gets) produced, who produced it, what it represent (and also what it might ignore), and what socio-political contexts might affect their scope.

Please respond to **at least 7** of the questions listed below for the dataset that you will be analyzing this quarter. **Note that everyone should respond to the last question.** As you respond to these questions, you should be sure to examine the data documentation, the web source where you retrieved the data, the website of the organization that collected or published the data, and any laws/press releases/news articles that reference the data. **All responses should be in complete sentences, in your own words, and be sure to cite all sources.** Citing your sources is not only important for academic integrity. Please keep in mind that these datasets are also new to me. I'm learning these datasets with you, and if you don't cite your sources then it will be very difficult for me to offer you feedback.

As you work through these questions, I will also be responding to them in order to introduce you to the one of the datasets that will serve as an example in our labs throughout this quarter: the Covid-19 Case Counts Dataset.

Check it out below:

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
cases_raw <- read.csv("https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_da

#Do not worry about this line of code for now. Since the cases data gets appended every day with a new
cases <-
  cases_raw %>%
  mutate(Total.Cases =
           cases_raw %>%
           select(starts_with("X")) %>%
           rowSums()
         ) %>%
  select(Province.State, Country.Region, Total.Cases)
```

```r
cases %>% head()
```

```
##   Province.State      Country.Region Total.Cases
## 1                        Afghanistan     9173475
## 2                            Albania     4529846
## 3                            Algeria    11901443
## 4                            Andorra      792297
## 5                             Angola     1576943
## 6                Antigua and Barbuda       25066
```

We will talk more about the values encoded in this dataset in later labs. For this week, however, we will consider the context where the dataset comes from, how it was produced, and how it have been used and interpreted by data scientists and decision makers alike.

Datasets

Before you start, note that very rarely is data documentation thorough enough to address **all** of the questions below. You can expect that you will likely not be able to answer all questions. As you work through the lab, consider how well you believe the dataset to be documented and how the documentation could be improved. Also remember that **everyone should respond to the last question for their dataset.**

Data Context and Historical Background

In this class, we have been talking about the importance of understanding the social, political, and economic contexts under which data gets produced. These forces play a crucial role in implicating how variables in a dataset get defined, how numbers get reported, and how we interpret the features and quantities represented in data. Consider data about the spread of Covid-19. If we were take the numbers reported in this dataset at face value without examining the contexts under which that data was produced, we could seriously be undermining the extent of the spread. Please watch a *short segment* of the following video to learn more about this.

Who collects the data? Who aggregates the data? Who publishes the data? Why are these groups involved in the production of the data?

Covid-19 Case Counts

The Covid-19 Case Counts are published on GitHub by a team of researchers at Johns Hopkins Center for Systems Science and Engineering, led by Professor Lauren Garnder. The Johns Hopkins team aggregates the data from the World Health Organization, the U.S. Centers for Disease Control and Prevention, the European Center for Disease Prevention and Control, the National Health Commission of the People's Republic of China, 1point3acres, Worldometers.info, BNO, DYX, state and national health departments, and media reports. These organizations aggregate the data from national authorities, who aggregate the data from state and local health departments, who aggregate the data from various testing centers. Because of this long chain of data transfer, local data is likely to be more accurate and up-to-date than data reported in this dataset. See: GitHub Page

Now, respond to this question for your dataset.

```
Fill response here.
```

According to what procedures is the data collected?

Covid-19 Case Counts

Covid-19 case count data is collected through laboratory testing that checks for the presence of Covid-19 in respiratory specimens, saliva, and sometimes blood samples. Two kinds of tests are available for COVID-19:

viral tests and antibody tests.

- A viral test tells you if you have a current infection.
- An antibody test might tell you if you had a past infection.

These testing procedures must be approved for diagnostic purposes by government officials. Positive tests are reported to state and local health departments.

See: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html

Respond to this question for your dataset.

```
Fill response here.
```

How long has the data been collected? Have there been any major changes to how the data gets collected since its inception?

Covid-19 Case Counts

The Johns Hopkins research team first published a dashboard presenting this data on January 21, 2020. While at first they were manually checking reported case count numbers on a daily basis, on February 1, 2020, they transitioned to a semi-automated process for checking case counts across sources. While this sped up aggregation, it also introduced new opportunities for error in the dataset, as well as changes to the structure of the dataset that other organizations were already relying on. With case count data streaming in from a variety of sources in each country, the Johns Hopkins team has to continuously make decisions about which data to trust, and at times has to retroactively change reported numbers when new evidence is reported.

In addition, the actual testing of patients, which is how this data is initially collected, has changed considerably since January 2020. as a result of changing federal regulations restricting testing and supply chain shortages.

Respond to this question for your dataset.

```
Fill response here.
```

Where is the data collected? How often is the data collected?

Covid-19 Case Counts

This data is collected globally and is aggregated on a daily basis. Since the data is published on GitHub, members of the GitHub community regularly request for new countries to be represented in the data as new cases are reported. See: GitHub Page

Respond to this question for your dataset.

```
Fill response here.
```

Are there any laws that mandate the data to be collected? If so, what are they?

Covid-19 Case Counts

There are currently not laws that specifically mandate the aggregation of the data reported in this dataset. However, there are laws that implicate their data collection methods. For instance, in the US, the Families First Coronavirus Response Act mandates that employer-sponsored health plans provide coverage for Covid-19 testing.

Respond to this question for your dataset.

```
Fill response here.
```

Have any social movements pushed for the data to be collected? How have these groups exerted influence over the data collection?

Covid-19 Case Counts

Absolutely! Demands across the globe for governments to lift restrictions on testing have played an important role in the numbers that we see reported in this data. One area where we are seeing activism play a big role in the numbers reported in this data is in communities advocating for testing of incarcerated individuals, since social distancing and sanitation is much less accessible in prisons. See this article, for instance. Healthcare workers have also protested demanding more regular testing. As we know the case counts reported in this data are dependent on testing so these social movements play an important role in the numbers that ultimately get reported.

Respond to this question for your dataset.

```
Fill response here.
```

What technologies, algorithms, or other datasets are needed to produce this dataset?

Covid-19 Case Counts

The Johns Hopkins research team relies on other data streams across the globe in order to aggregate this data. In the process, they have to use judgment to decide which sources of data are reputable and what to do when conflicting numbers are reported across streams. Case count data is also dependent on the availability of testing equipment, and shortages of testing materials and agents can have a significant impact on the numbers reported in the data.

Respond to this question for your dataset.

```
Fill response here.
```

List any expenses that you can anticipate being involved in producing this data. Who pays for these expenses?

Covid-19 Case Counts

With many stakeholders and technologies involved this data production, the cost of producing this data is enormous. First, we should consider the work of the individuals composing the research team at Johns Hopkins. They have graciously devoted an incredible amount of time to make this a public resource, and while the work they've put into it likely cannot be reduced to a standard work week, the costs of their human labor should be factored into the cost of this data production.

See: Kaiser, Jocelyn. 2020. " 'Every Day Is a New Surprise.' Inside the Effort to Produce the World's Most Popular Coronavirus Tracker." Science | AAAS. April 6, 2020. https://www.sciencemag.org/news/2020/04/every-day-new-surprise-inside-effort-produce-world-s-most-popular-coronavirus-tracker.

Then, we should consider all of the infrastructure needed to support data storage and flows - servers to host data, web platforms where case counts are shared, in-country communications channels to get data from local and state departments to national health authorities. At every stage of data transfer, there are salaries to be paid for individuals that maintain this data storage and networking infrastructure.

Then, there are the tests: We need to consider the human labor costs of all of the healthcare staff responsible for administering tests, the lab technicians responsible for analyzing tests, the public health authorities responsible for overseeing testing, the legal authorities responsible for setting testing mandates, the packaging and shipping workers responsible for distributing supplies and transporting specimens to labs, and the individuals employed in the various industries where the testing supplies are manufactured. I'm sure that we can continue to grow this list.

Tests themselves cost money. In the US, the Families First Coronavirus Response Act aims to ensure that individuals with or without insurance will not have to pay for these tests. Instead insurers will be required to pay for tests of those they insure, while public funds will cover the cost of tests for those without insurance. Out-of-pocket fees may still incur when tests are performed in out-of-network labs. The cost of the actual tests depends on the kind of testing being performed.

Respond to this question for your dataset.

```
Fill response here.
```

How long has this dataset been public? Under what circumstances was this dataset made public? Through what channels has the dataset been made available to the public?

Covid-19 Case Counts

The Johns Hopkins team published this data on January 22, 2020 - both through a public GitHub repository and through a data dashboard visualizing the spread.

Respond to this question for your dataset.

```
Fill response here.
```

Do any institutions leverage the dataset to support decision-making? Which institutions, and which decisions?

Covid-19 Case Counts

The production of data about confirmed cases has been a significant part of the public discourse around the spread of the virus. Why has data been so central to this virus? The reasons are almost too many to count and seem to multiply each day. Here are just a few:

- In helping to indicate the degree to which the virus has spread throughout a community, data about confirmed cases has helped federal regulators, states, counties, and local authorities discern when they need to establish protocols for social distancing and business closures.
- Data has been an important component of the public relations strategy, helping to communicate the urgency to communities to practice social distancing and to honor shelter-in-place protocols.
- Data about confirmed cases has helped to indicate which hospitals are likely to face shortages in beds, respirators, and personal protective equipment. This helps authorities predict how they may need to redistribute resources and hospitals make decisions about purchasing, healthcare worker schedules, and how to reconfigure hospital set-ups for patient intake.
- Data about the Covid-19 confirmed cases has helped policy-makers determine when they need to pass legislation to protect citizens from exorbitant healthcare costs.
- Banks rely on data about confirmed cases to determine the extent to which gross domestic product will fall as a result of the virus and what impact it will have on the economy.
- Your university is relying on data about confirmed cases to plan for the longevity of remote teaching and whether institutional events will need to be cancelled.

This list could be considerably expanded.

Respond to this question for your dataset.

```
Fill response here.
```

How has this dataset been talked about in the media?

Covid-19 Case Counts

There are many different organizations tracking confirmed cases of Covid-19, but the Johns Hopkins dataset is often referenced as the most up-to-date and comprehensive. It has primarily been *applauded* in the media by a number of sources. Just a few are listed below:

- New York Magazine
- NPR
- Science Magazine

Respond to this question for your dataset.

```
Fill response here.
```

(REQUIRED) Which people, places, things, or issues may be over-represented in the data? Why? Which people, places, things, or issues may be under-represented in the data? Why?

Covid-19 Case Counts

We know that this data only reflects confirmed cases of those individuals that have been tested for the virus and cannot account for those that have not been tested. But who is getting tested? Some academic and journalistic sources have indicated that:

- testing can be scarce in rural communities.
- undocumented immigrants may choose not to get tested to avoid deportation or detainment in centers where the virus is known to be spreading

Under these conditions, confirmed cases in these communities are likely to be much higher than what is reported in the data.

Also see: Li, Siyue, et al. "Internet use, risk awareness, and demographic characteristics associated with engagement in preventive behaviors and testing: cross-sectional survey on COVID-19 in the United States." Journal of medical Internet research 22.6 (2020): e19782.

Respond to this question for your dataset.

```
Fill response here.
```