Data Collection And Cleaning

Justine Huynh

**Source**

This data came from the Covid Tracking Project website (<https://covidtracking.com/data>). This project records the cases, deaths, hospitalizations, PCR tests, antibody tests, and much more in the face of the Covid-19 pandemic. Specifically, the data for this project mostly records new cases, new deaths (per state, per race) every few days, with few records of hospitalization rates and testing rates. The data resembles a timeseries dataset because there exists a date variable.

**License**

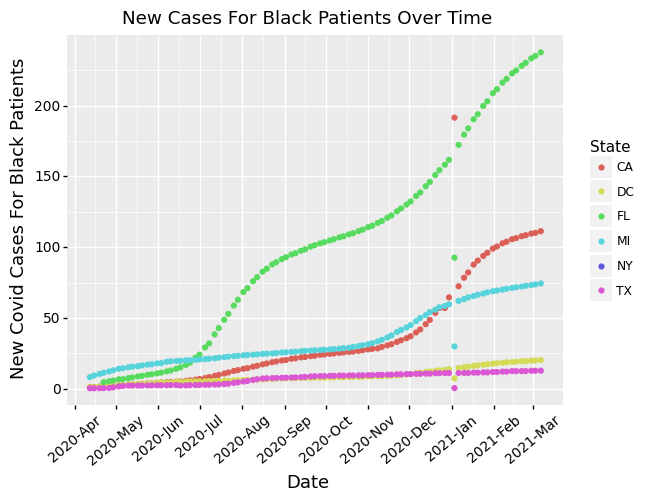
There are no licenses on the webpage, as this dataset (and other Covid-related dataset) is publically available.

**Collection Methods**

The Covid Tracking team did not explicitly explain how they collected the data, though they mentioned the data was federally available and “relied on hundreds of volunteers who entered data manually on a daily and weekly basis for a year.” The webpage had separate hyperlinks of CSV data for each state, so I had to scrape through the webpages to collect data.

**Biases/Sampling**

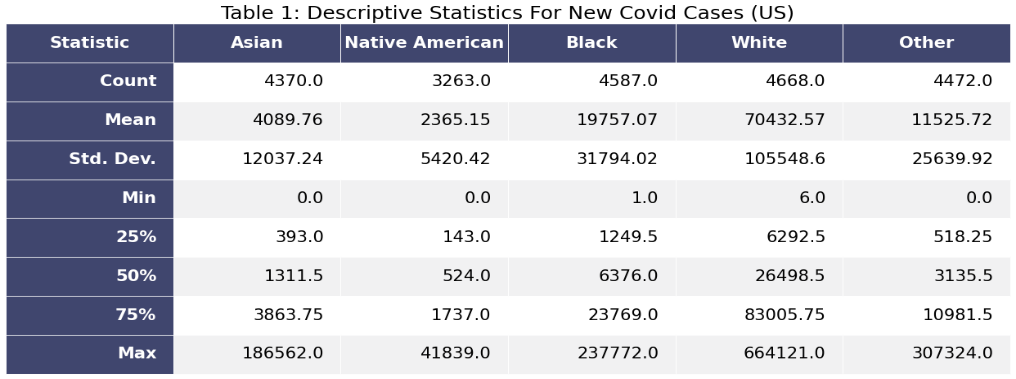
Because there exists a date variable, autocorrelation errors exist from time. The errors from Alaska one day could closely mimic the errors from Alaska the next couple days, as both recordings occurred from the same state in a similar time period. Furthermore, the data may be too small. Although the data I have scraped has about 4900 rows, it is divided by 50 states (and Washington, DC). Specifically, the data has about 95 rows per state, which I fear may not be a big enough sample per state and increase variance. Additionally, a quick glance at the missingness matrix reveals a lot of NAs or non-entries in many columns. Not all the states released hospitalization and test data (like Alabama). Not every state revealed new cases and new deaths for multi-racial people, which could be because the state forces the multi-racial patient to identify with one race or sadly neglected that patient. Such omission of data could severely alter the information I could analyze or display to the readers.



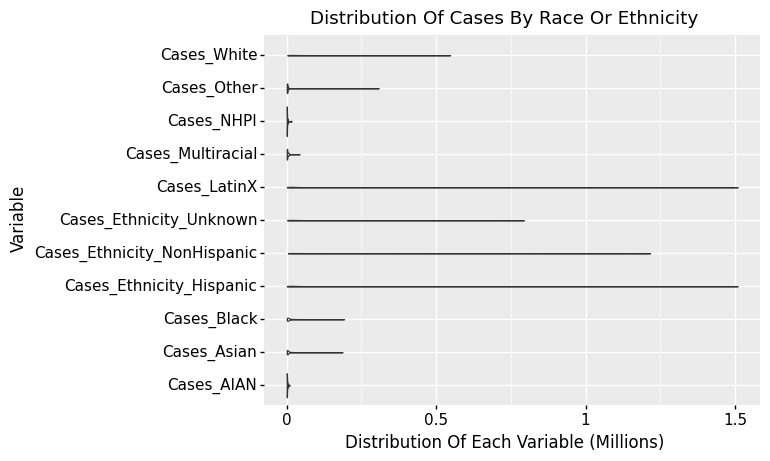
I also noticed that for many states, entries for the beginning of January of 2021 often severely under-counted the new cases for namely Black patients, probably because people are too busy understandably celebrating the New Year during that time to care about people of color, which can bias some of my findings. This implies that racism still permeates the data collection.

**Descriptive Statistics**

The dataset has about 4900 rows and 50 columns. Most of the columns are numeric data that depicts new cases and new deaths by race. The date column, although originally presented as an integer, was later casted as a date object.

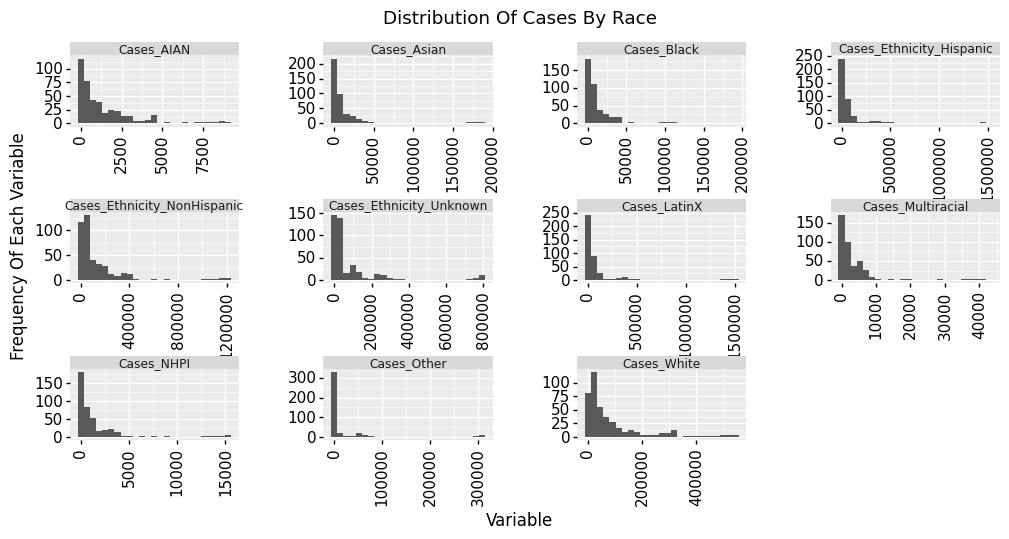


Detailed here is the descriptive statistics for new Covid cases (aggregating all states) for each race in the US. Note that Native Americans have the fewest records of cases and on average, lowest cases. Whites, on the other hand, have the highest mean Covid cases.



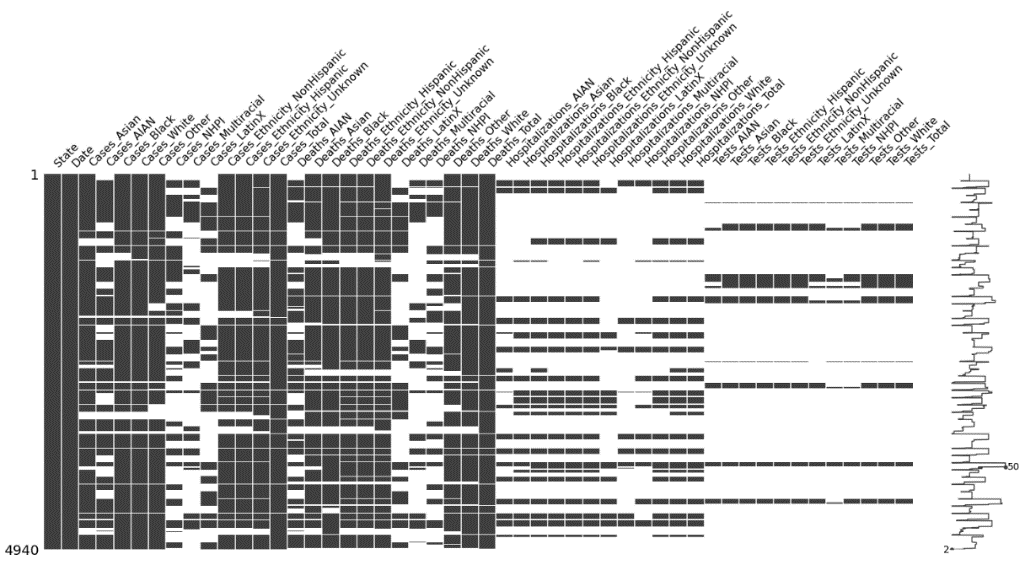
According to the violin plot (after adding ethnicity and other races), people who are Hispanic or LatinX suffer from the highest Covid cases, some as high as 1.5 million new cases. The violin plot is so thin that it implies the distribution is more stretched out, with most of the cases at the bottom end of the violin. It’s worthy to note that some of these races had about 50% of their entries at NA, so the 1.5 million new cases may be outliers, albeit somber ones.

**Data Cleaning Methods & Identified Issues**



For most of the cases, I notice that they are mostly right-skew, with a tail near the high-end of cases. There is a larger density in the lower-number cases.

As mentioned above, more than 50% of the hospitalization and testing variables had NAs or missing entries, so I unfortunately had to delete them.



According to the missingness matrix, there are unfortunately many columns with a substantial amount of NAs. The hospitalizations and testing variables consist of more than 70% NAs. Furthermore, there are at least 50% NAs for the cases with Native Americans, Native Hawaiian/Pacific Islander, Multiracial, and LatinX races.

Further, the date column was not in its natural datetime format; instead, the original data collectors saved each date as an integer. Therefore, I recast the column into a datetime column.

**Data Dictionary of Key Features**

I believe these variables (namely deaths and cases by race) will be important in this Covid project:

Cases\_Asian: number of new cases among Asian patients

Cases\_Black: number of new cases among Black patients

Cases\_White: number of new cases among White patients

Cases\_LatinX: number of new cases among LatinX patients

Deaths\_Asian: number of new deaths among Asian patients

Deaths\_Black: number of new deaths among Black patients

Deaths\_White: number of new deaths among White patients

State: the US state (where the observation was recorded)

Date: the date (in *year, month, day* formatting) the observation was made

Works Cited

*The Data,* The Covid Tracking Project, 7 Mar 2021. <https://covidtracking.com/data>. Accessed 21 September 2021