**6.1: Sourcing Open Data**

**Data source**

Summary:

*Number of COVID-19 Confirmed, Death and Recovered cases every day across the globe.*

Collection Details:

Coronavirus COVID-19 Global Cases data collected by the [Center for Systems Science and Engineering (CSSE)](https://coronavirus.jhu.edu/map.html) at Johns Hopkins University; the Census American Community Survey; the Department of Health and Human Services; and the Bureau of Labor and Statistics all of which are credible/trustworthy, updated regularly government and educational based institutions.

Content:

* full\_grouped.csv - Day to day country wise no. of cases (contains County/State/Province level data)
* **covid*19*clean\_complete.csv - Day to day country wise no. of cases (not Including County/State/Province level data) The file contains the cumulative count of confirmed, death and recovered cases of COVID-19 from different countries from 22nd January 2020**
* country*wise*latest.csv - Latest country level no. of cases
* day\_wise.csv - Day wise no. of cases (not Including country level data)
* usa*county*wise.csv - Day to day county level no. of cases
* worldometer\_data.csv - Latest data from <https://www.worldometers.info/>
* Time period - The dataset starts from 22nd January to 27th July 2020

Context & Why I chose this dataset:

* A novel strain of coronavirus — SARS-CoV-2 — was first detected in December 2019 in Wuhan, a city in China’s Hubei province with a population of 11 million, after an outbreak of pneumonia without an obvious cause. The virus has now spread to over 200 countries and territories across the globe and was characterized as a pandemic by the World Health Organization.
* The number of COVID-19 positive cases in the United States currently stands at 75.3 million\* as of 02/02/2022. Being from a medical background, I want to show what our current status in fighting this novel pandemic that has taken so many lives in the U.S. where there have been 890K\* deaths and much more worldwide since the start of the pandemic.

\*ReSources:

<https://www.nytimes.com/interactive/2021/us/covid-cases.html>

<https://www.kaggle.com/imdevskp/corona-virus-report>

**Data Profile**

* The data set contains 49068 rows and 10 columns.
* Columns: Province/State, Country/Region, Latitude, Longitude, Date, Confirmed (COVID-19 cases), Deaths (due to COVID-19), Recovered (from COVID-19 infection), Active (COVID-19 cases), and WHO region.
* There are 187 unique countries and 78 unique province/states enlisted in the dataset.
* The dataset time period is from 22nd January to 27th July 2020.

**Data wrangling:**

* To be more concise, I renamed ‘Province/State' column to 'State' and the 'Country/Region' to 'Country’.
* Since the column "State" includes a lot of NaN, and it's needless for this for this particular analysis as we will focus on the world, this column was dropped.
* After removing column “State” the dataset contains 49068 rows and 9 columns.

**Consistency Checks:**

* There are no mixed data types resulted.
* There are no missing values resulted.
* There are no duplicates values resulted.
* Studying the descriptive statistical results, we see that the minimum of active cases is [-14], we cannot have a negative number of cases that are active. This minimum negative value and others like it come from the equation: [Active Case = confirmed - deaths - recovered]. For example, if the confirmed cases have a value lower than the amount recovered, you will get a negative value. In simpler terms, if a person has recovered, tested negative for the COVID-19 virus, they are essentially no longer active, decreasing the likelihood of transmissibility and transmission of the virus to others, making them a part of the recovered pool.
* Solution: Since the negative values only make up 18 of the 49,068 rows, a very small portion of our data, and for the purposes of our final analysis focusing on confirmed cases, deaths and recovered only, we will replace the negative values with NaN.

**Basic descriptive statistics:**

* 49068 rows, 9 columns.
* Confirmed cases:
  + Mean – 16885
  + Minimum – 0
  + Maximum – 4290259
* Deaths:
  + Mean – 884
  + Minimum – 0
  + Maximum – 148011
* Recovered:
  + Mean – 7916
  + Minimum – 0
  + Maximum – 1846641
* Active:
  + Mean – 8088
  + Minimum – 0
  + Maximum – 2816444

**Limitations and ethical considerations**

The limitations in this dataset include the ‘NaN’ values. These values are included in a different U.S. focused data set if needed but for the purposes of this analysis are not needed. Since there no PPI, or private patient information like patient name, medical record numbers or home address, there are no ethical concerns at this time. Reporting error may occur as data is uploaded daily and some data points may have been missed due to lack of reporting on time.

**Key Questions**

* Are COVID-19 infections increasing or decreasing worldwide?
  + Based on these results we can determine if treatment and current prevention methods are working or if changes need to be made.
* Where in the world are the highest number of confirmed cases?
  + Based on these results we can see where medical efforts need to be placed and if need be, isolation and immigration restrictions need to be discussed to prevent spread.
* Mortality and recovery rate among various countries?
  + Again, this will show us where medical efforts need to be implemented to prevent spread and save lives for the future.