**DASC 5433: Big Data Analytics**

**Project Report of a**

**Big Data Analytics on COVID – 19**

Team:

Akshay Kumar Puvvada 2250348

Monika Bondala 2250349

Haswanth Tummala 2249790

Table of Contents

[Introduction 3](file:///C:\Users\akkpu\projects\hadoop\docker-hadoop\project\Reports\Big%20Data%20Analytics%20on%20Covid.docx#_Toc146301145)

[Problem Statement 3](file:///C:\Users\akkpu\projects\hadoop\docker-hadoop\project\Reports\Big%20Data%20Analytics%20on%20Covid.docx#_Toc146301146)

[Objectives 4](file:///C:\Users\akkpu\projects\hadoop\docker-hadoop\project\Reports\Big%20Data%20Analytics%20on%20Covid.docx#_Toc146301147)

[Preliminary Literature Review 4](file:///C:\Users\akkpu\projects\hadoop\docker-hadoop\project\Reports\Big%20Data%20Analytics%20on%20Covid.docx#_Toc146301148)

[Methodology 5](file:///C:\Users\akkpu\projects\hadoop\docker-hadoop\project\Reports\Big%20Data%20Analytics%20on%20Covid.docx#_Toc146301149)

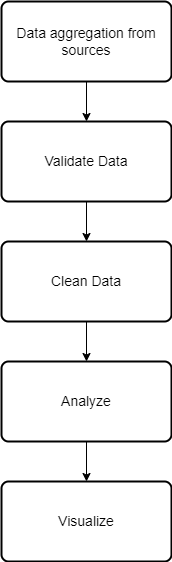
[References: 6](file:///C:\Users\akkpu\projects\hadoop\docker-hadoop\project\Reports\Big%20Data%20Analytics%20on%20Covid.docx#_Toc146301150)

# Introduction

The COVID-19 pandemic has presented an unprecedented global health challenge, necessitating innovative approaches for monitoring, understanding, and mitigating its impact. Big Data Analytics has emerged as a pivotal tool in this context, offering a comprehensive framework for collecting, processing, and analyzing vast and diverse datasets related to COVID-19. This abstract outlines the key components of a Big Data Analytics project aimed at addressing the challenges posed by the pandemic.

The foundation of this project rests upon the meticulous collection and integration of diverse data sources. These encompass medical databases from various government organizations. The integration process ensures data accuracy, consistency, creating a robust and dynamic data ecosystem.

Standardize and integrate data from various sources to ensure consistency and compatibility. Identify and rectify data quality issues, such as missing values and outliers, to ensure the accuracy of analyses. Implement a robust data management system that can handle the sheer volume and variety of COVID-19 data. Feature Engineering to create relevant features and variables from raw data to facilitate analysis. Create interactive dashboards and data visualizations to communicate findings effectively to decision-makers and the public. Utilize geospatial visualizations to track the spread of the virus at regional and global scales. Continuously adapt the analytics framework to address evolving challenges in the pandemic. Collaborate with epidemiologists, healthcare professionals, and policymakers to make data-driven decisions. The essential components of a Big Data Analytics project for COVID-19, highlighting the importance of data collection, integration, analysis, visualization, and ethical considerations in gaining comprehensive insights and informing effective responses to the ongoing pandemic.



# Problem Statement

The main problem in case of pandemics in such a large mass of people is to determine where the health care is needed most, which category of people need more attention and which geographical area will be most affected at particular time. Also, to collect the enormous amount of information for big data analytics we need to extract the lots of data from different government bodies, websites, organizations. Data needs to be validated, redundant data has to be cleaned.

The following problems are major based on the review of supporting health care workers, pharma companies and analyzing risk factor score for a particular person.

* Age factor of a person
* Type of variant affected for the person, based on some reports Delta variant of Covid-19 affected young generation rather than old generation people compared to other variants
* To gather and analyze the information related to rate of infection at different geographical areas, their travel history etc.

Misinformation and vaccine reluctance have also impacted vaccination rates unevenly, which can affect the spread of the pandemic and efforts to build herd immunity. In addition to investing in standardized data collection and reporting systems, healthcare infrastructure, and public education campaigns to promote accurate information about COVID-19 and vaccines, addressing these challenges requires collaboration and coordination across national and international levels.

# Objectives

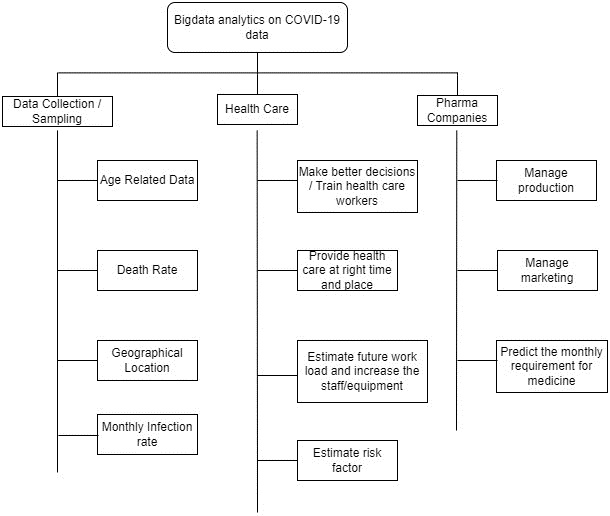
The goal of the project is to use the huge and variant amount of data, this data can be used to apply big data analytics in various areas like diagnosing according to symptoms, estimating the risk score with the rate of infection, to help take better decision for health care workers and also give a proper estimate to pharmaceutical companies to manage their production, marketing insights, improving the pace of clinical trials. accordingly. The main objectives of the project is as follows.

* To collect the covid-19 (SARS-COV-19) related data, from at least 5 countries with the time frame ranging 2019-2022. The data collected is to be analyzed on monthly basis to predict
* To find the pattern for risk factor, based on the symptoms, age and geographical factors. Though there will be different aspects for analyzing the risk factor like premedical condition for a particular patient, lifestyle habits, type of covid variant person infected with, health care facilities available at that area.
* To provide future predictions in terms of health care sector, to make better decisions for increasing the number of health care workers, PPE kits, oxygen cylinders at a particular geographical area at a particular time. This helps to provide better health care at right time and right place.
* To predict the risk score for a particular person based on symptoms, age and geographical area so that health care professional can take better decision for providing the appropriate medicine and also determine patient priority based on the risk factor through the data three factors collected.

# Preliminary Literature Review

The COVID-19 pandemic has prompted an unprecedented surge in research efforts, resulting in a vast and rapidly evolving body of literature. This preliminary literature review aims to provide an overview of key themes and findings related to leveraging Big Data Analytics for comprehensive insights into the pandemic.

                Early in the pandemic, researchers emphasized the need for diverse data sources. Studies (Zhu et al., 2020) highlighted the significance of integrating data from healthcare records, genomic sequencing, and social media to capture the multifaceted nature of the pandemic. Data integration challenges, including standardization and cleansing, were recognized as critical for robust analysis (Wang et al., 2020). Machine learning and predictive modeling emerged as powerful tools for forecasting COVID-19 trends. Research (Kumar et al., 2020) demonstrated the effectiveness of models in predicting case numbers, hospitalizations, and mortality rates. Time series analysis (Liu et al., 2020) was commonly employed to capture temporal patterns. Researchers emphasized the importance of effective data visualization for communication. Interactive dashboards (Bilal et al., 2021) and geospatial visualizations (Dong et al., 2020) were instrumental in conveying complex information to the public and policymakers. This preliminary literature review underscores the multidisciplinary nature of research related to COVID-19 and Big Data Analytics. It showcases the evolution of analytical techniques and the growing importance of data integration, ethical considerations, and interdisciplinary collaboration in tackling the ongoing pandemic. Future research in this field is expected to continue to shape our understanding of the virus and guide public health responses.



# Methodology

For preliminary research we have made a literature review for understanding Covid-19 disease, its variants, how it affects different people with premedical health conditions ,age groups(young adults, children ,old people)rent age group The methodology employed in this Big Data Analytics project for comprehensively addressing the COVID-19 pandemic involves a structured and iterative process encompassing data acquisition, pre-processing analysis.

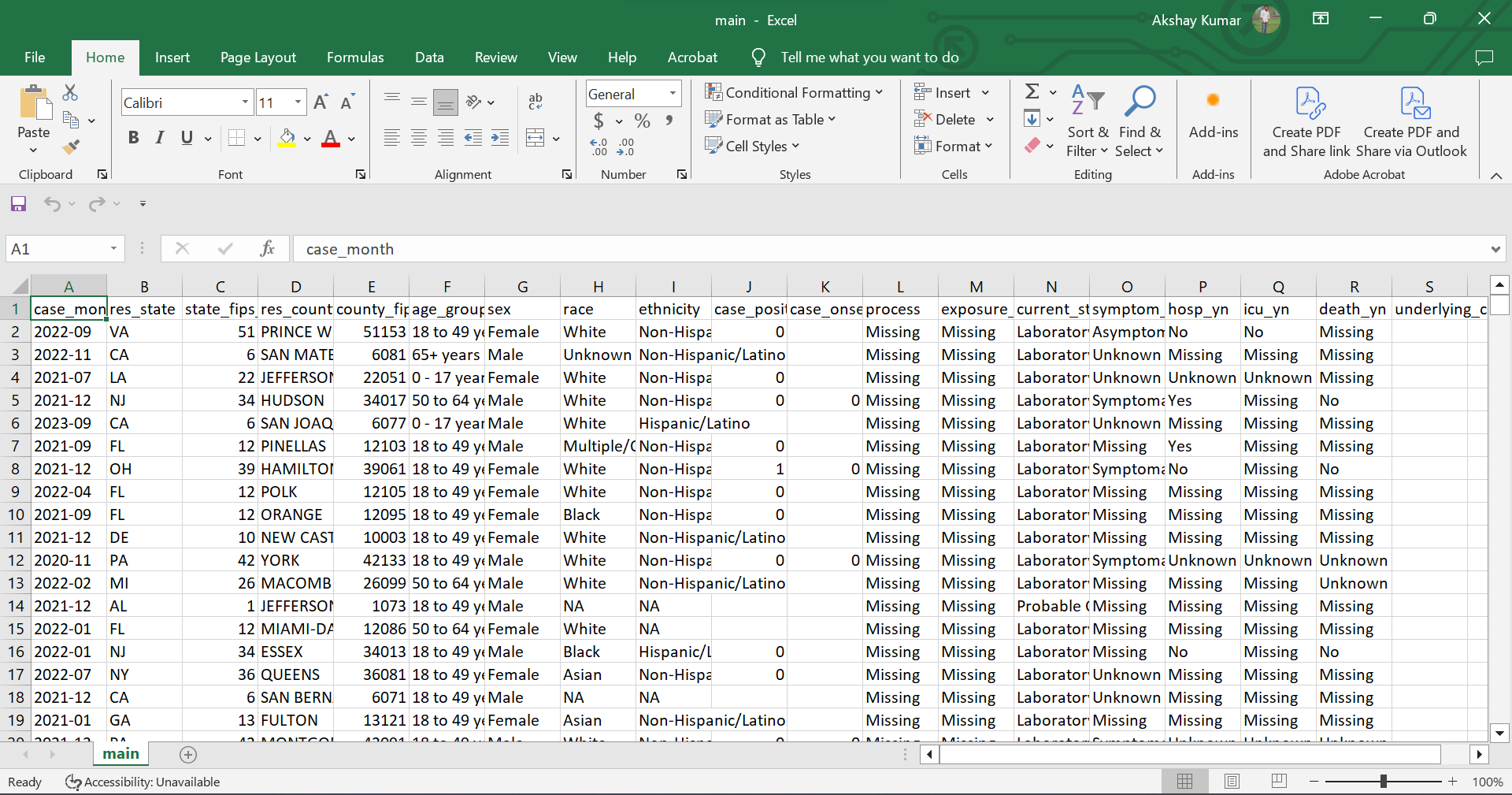
          We will be analyzing the data with time-range 2019-2022 , to predict the risk factor. The project begins with the systematic collection of COVID-19-related data from diverse sources, based on symptoms, age and geographical factors. These sources provide a holistic view of the pandemic's impact, ensuring the availability of comprehensive and up-to-date information. Hadoop’s processing model, MapReduce Programming paradigm we believe it is suitable for parallelizing and processing large-scale data. It can be used for various analytics tasks, such as data aggregation, filtering, and transformation. This structured methodology serves as a robust framework for leveraging Big Data Analytics to gain comprehensive insights into the COVID-19 pandemic, contributing to our understanding of the virus and supporting informed responses at local, regional, and global levels.

Steps followed for analysis

* Data Collection
* Data Cleaning
* Push data into HDFS
* Write mapper functions
* Write reducer functions
* Analyze the summarized results from reducer functions.

## Data collection

The data is downloaded from CDC – Center For Disease Control (<https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/ge62-5fe5>) 13.3 GB.



Key points of data

* Data has been categorized with age groups as follows

0-17 years

18-49 years

50-64 years

65+ years

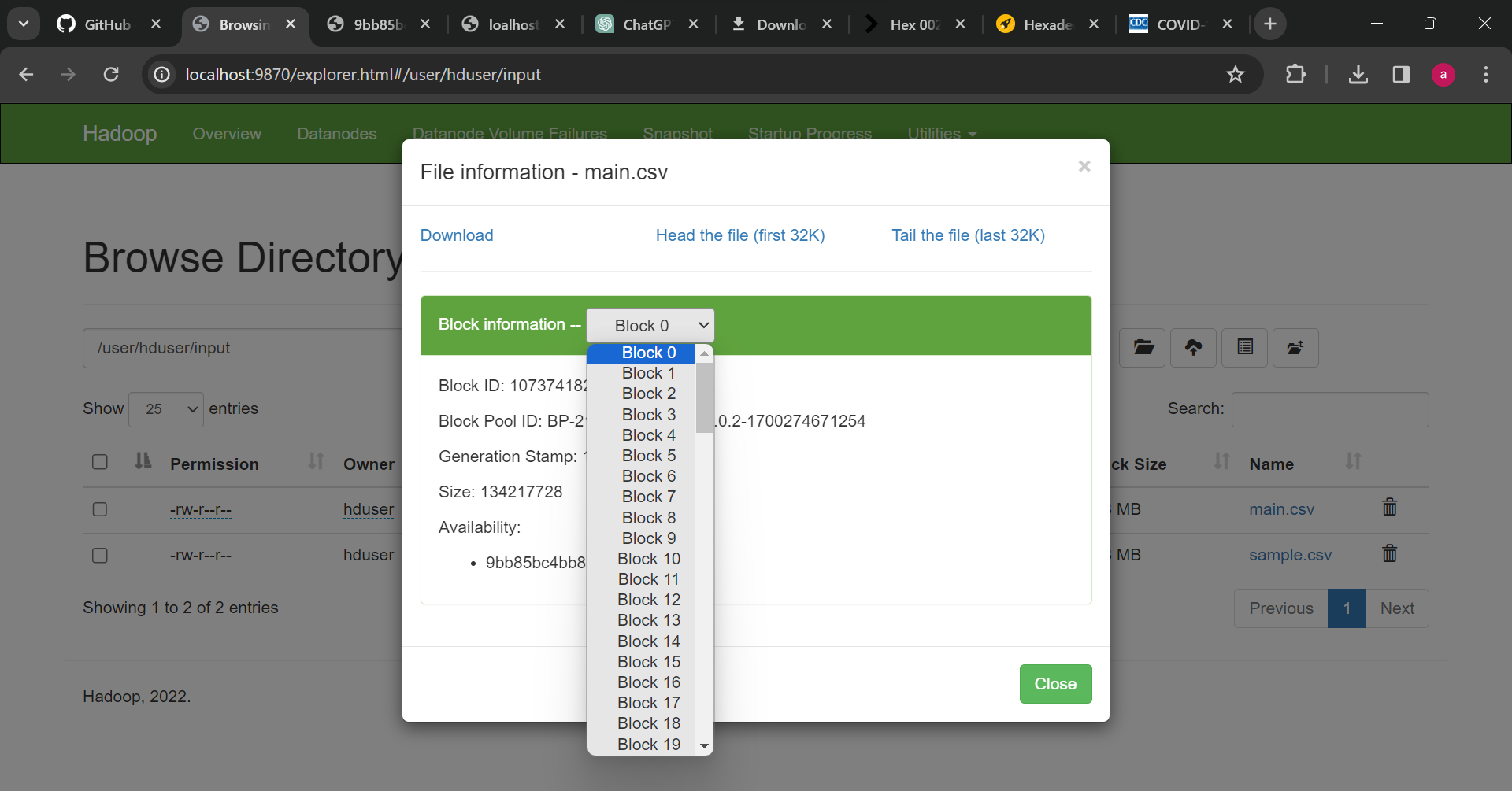
And the data that is missing or not available

* Data related to states
* Whether dead or not

We have used only mentioned columns for analysis

## Data in Hadoop

Data has been uploaded to Hadoop Distributed File System the total 13.3 GB data has been divided to 106 blocks.



## Mapper functions

Following are mapper functions for extracting the data we care about and form key value pairs.

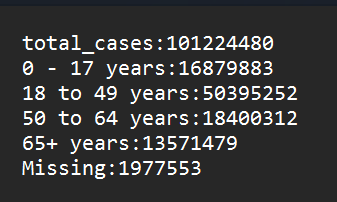
* mapper.py for forming county\tage\_group pair
* dead\_mapper.py for forming age\_group:is\_dead pair
* state\_mapper.py for forming state: is\_dead

## Reducer functions

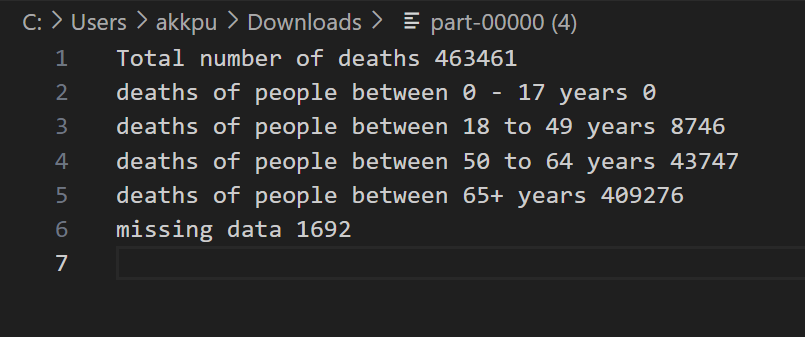
* reducer.py for aggregating data related to age group
* dead\_reducer.py to aggregate data related to dead and age\_group
* state\_reducer.py to aggregate data related to state
* state\_dead\_reducer.py to aggregate data related to death in different states

## Analysis achieved in this project

* Number of cases registered as per age group



* Number of deaths as per age group



* Number of covid cases in different states

{'AK': 302397, 'AL': 1710458, 'AR': 1021595, 'AZ': 2525978, 'CA': 12489810, 'CO': 1805662, 'CT': 728115, 'DC': 179229, 'DE': 330578, 'FL': 5927547, 'GA': 2707870, 'GU': 54224, 'HI': 331040, 'IA': 503717, 'ID': 529487, 'IL': 4212196, 'IN': 2104148, 'KS': 952290, 'KY': 1739148, 'LA': 1668554, 'MA': 2292748, 'MD': 1655280, 'ME': 327523, 'MI': 3004872, 'MN': 1819165, 'MO': 1821721, 'MS': 477972, 'MT': 267982, 'NA': 1704, 'NC': 3543183, 'ND': 294373, 'NE': 531004, 'NH': 383080, 'NJ': 3192933, 'NM': 701436, 'NV': 819245, 'NY': 7214608, 'OH': 3528174, 'OK': 772176, 'OR': 977833, 'PA': 3613663, 'PR': 178400, 'RI': 341591, 'SC': 1960618, 'SD': 302659, 'TN': 2557214, 'TX': 9172843, 'UT': 1112638, 'VA': 2363374, 'VI': 24635, 'VT': 143988, 'WA': 2000719, 'WI': 1800516, 'WV': 11002, 'WY': 189564}

* Deaths in each state

{'AK': 109, 'AL': 5189, 'AR': 3789, 'AZ': 22987, 'CA': 72269, 'CO': 4639, 'CT': 5407, 'DC': 712, 'FL': 50381, 'GA': 897, 'IA': 907, 'ID': 1900, 'IL': 23782, 'IN': 6271, 'KS': 4368, 'KY': 5986, 'LA': 2616, 'MA': 16640, 'MD': 2908, 'ME': 507, 'MI': 18872, 'MN': 6117, 'MO': 8499, 'MS': 1027, 'MT': 1319, 'NC': 5387, 'ND': 1031, 'NE': 222, 'NH': 1120, 'NJ': 20755, 'NM': 2576, 'NV': 10624, 'NY': 42682, 'OH': 22517, 'OK': 3920, 'OR': 1540, 'PA': 23112, 'PR': 4893, 'RI': 605, 'SC': 5385, 'SD': 81, 'TN': 9428, 'TX': 21050, 'UT': 1339, 'VA': 6030, 'VT': 44, 'WA': 5278, 'WI': 5400, 'WY': 344}

# References:

* <https://www.who.int/health-topics/coronavirus#tab=tab_1>
* <https://www.worldometers.info/coronavirus/>
* <https://ourworldindata.org/coronavirus>
* An Analysis of COVID-19 Mortality During the Dominancy of Alpha, Delta, and Omicron in the USA:<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10125879/#:~:text=It%20is%20believed%20that%20the,rapid%20reversal%20of%20that%20trend>.
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8037067/#:~:text=The%20spread%20of%20the%20global,and%20pharmaceutical%20industry%20%5B38%5D>.
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8577168/>