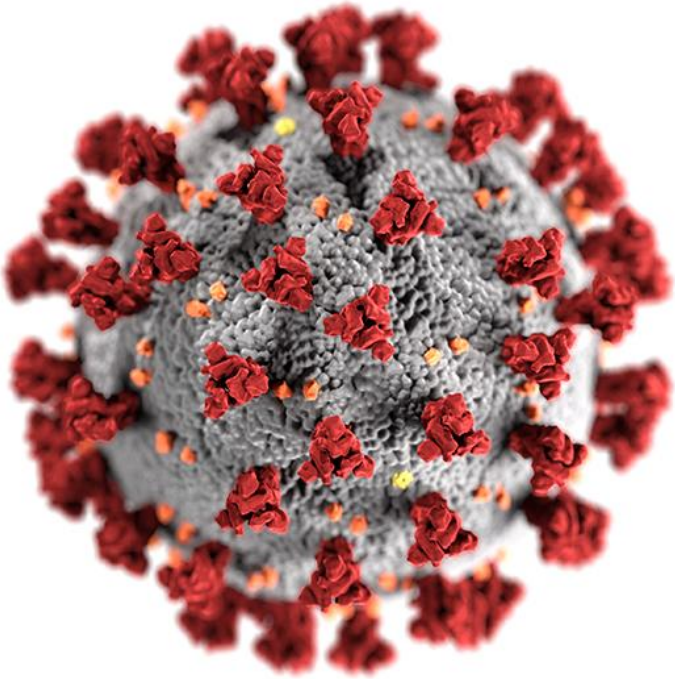
Open Dataset Research: Covid-19 Vaccine

User Guide and Analysis



INST 490 (iSchool - Open Dataset - Project #3)
March 2023
Tao Lin, Tsion Demissie, May Kyaw, Ebaide Akhigbe,
and Victor Ameche Aldridge

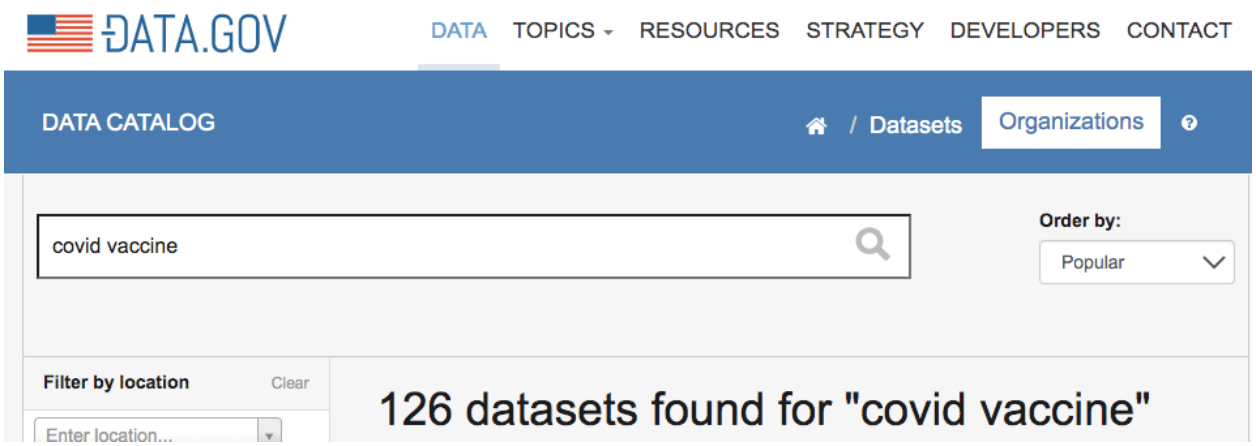
Table of Contents

Introduction to Database

The data is accessible by using the link: [Dataset-Catalog]

The COVID-19 pandemic has generated significant disturbances to our daily lives, as well as the global economy. As a result, scientists and researchers around the globe have been working tirelessly to formulate a vaccine that can assist us to combat this virus. The COVID-19 vaccine is a scalding subject right now, and it is critical to have access to reliable data and information about it. Luckily, the DATA.GOV database is an incredible resource for those seeking details about the COVID-19 vaccine.DATA.GOV is a platform that permits access to hundreds of thousands of datasets from assorted government agencies in the United States. The portal was developed as part of the Open Government Initiative, which strives to facilitate clarity, accountability, and public participation in government operations.

This dataset is a collection of data associated with the COVID-19 vaccine. The data is carefully compiled by the United States government working with its various branches, such as the Centers for Disease Control and Prevention (CDC) and the Department of Health & Human Services (HHS). The data collection includes a wide range of sources from statewide to federal datasets, accumulating data from all over the US. Utilizing the search function inputting the keywords of "COVID-19 vaccine," users are able to locate the information on the vaccine's adverse effects, distribution, and efficacy.



One of the benefits of utilizing this dataset is that it permits you to produce visualizations that can assist in pinpointing trends throughout time. These visualizations can be helpful in indicating key indicators of COVID-19 vaccine distribution and efficacy. Furthermore, the dataset has various aspects of ethnicity, age, and demographic, which can be factors that influence the spread of the virus. Thus, this dataset can enable researchers and policymakers to comprehend the affinity between vaccination and the spread of COVID-19. It is crucial to

mention that there are constraints to this dataset, as it only includes 126 datasets correlated to the COVID-19 vaccine. Furthermore, the data is restricted to what has been gathered by the US government and its entities. Nevertheless, the dataset is still useful for anyone seeking more details about the COVID-19 vaccine.

The DATA.GOV database is an incredible resource for those desiring more knowledge concerning the COVID-19 vaccine. In this user guide, we intentionally outline and illustrate step-by-step procedures on how to operate and modify data on the Data.gov website to equip the end user with a substantial understanding of what the data is all about.

Types of Datasets

Data.gov provides an exceptional prospect for researchers, analysts, and the general public to access and dissect data from an extensive spectrum of sources. The datasets are composed of eclectic topics, including, education, climate, finance, health, and energy, among others. Data.gov subsidizes various dataset publication formats such as JSON, XML, CSV, and many more. All the datasets are available to the public, the data is obtainable in a variety of formats that are supported by the platform and it is also more manageable to operate with the data for day-to-day users.

Different types of files supported

The most common file types that are supported in our dataset (in order of most common to least)

- CSV
    - Stands for Comma Separated Values, a simple file format utilized to hold tabular data.
    - Each row symbolizes a single data record, and each column represents a data field.
    - Can be readily imported or exported from different software such as Excel, MySQL, and Python.
    - Can retain thousands of cells of data on different COVID Vaccine statistics.
- RDF
    - Stands for Resource Description Framework, ideal for modeling metadata.
    - Utilized to represent resources and their associations with one another.
    - Can be employed for comprehensive expression, semantic web, and correlated data.
    - Supports triples, which involve a subject, predicate, and object.
- JSON
    - Stands for JavaScript Object Notation, a light data format.
    - Utilized to accumulate and exchange data between various methods.
    - Compromised with key-value pairs, arrays, and nested objects.

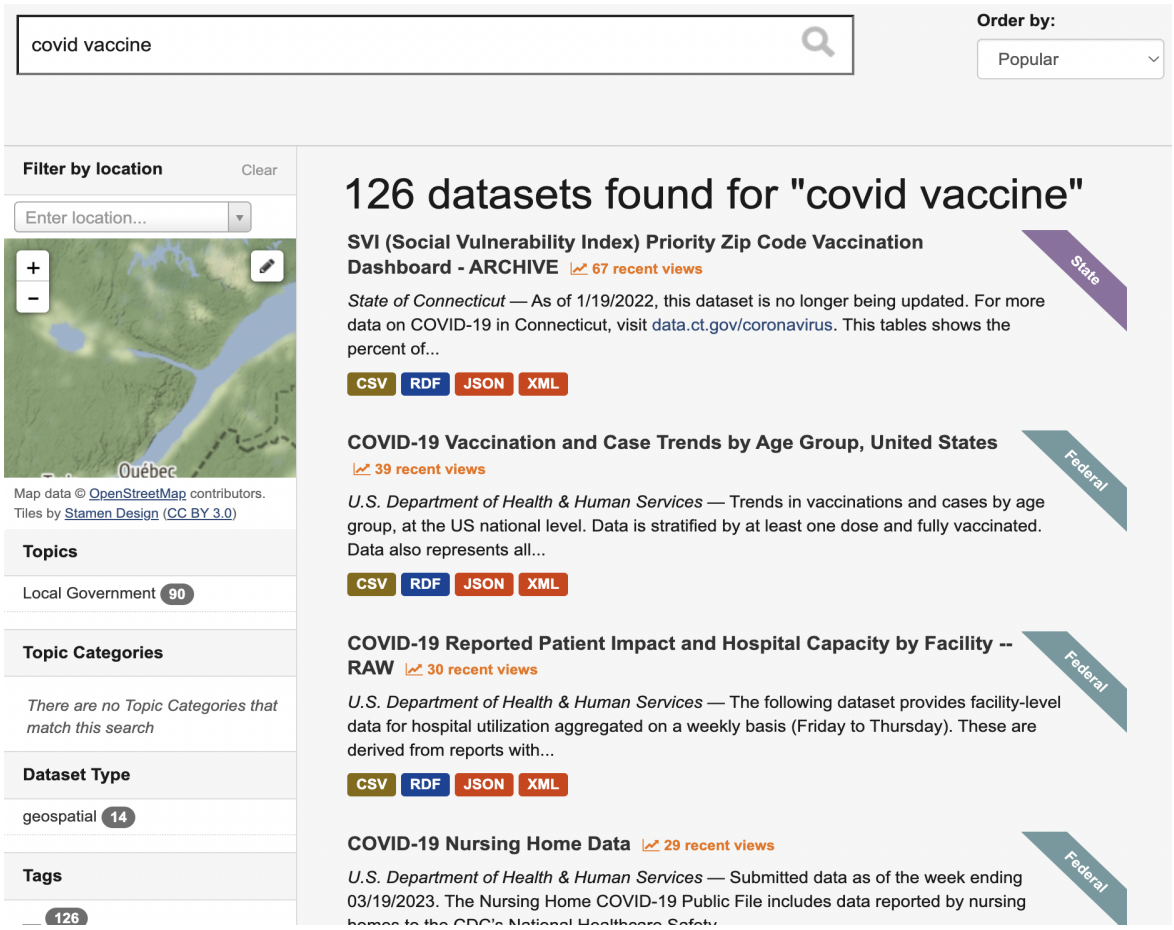
    - Is able to be operated with a variety of programming languages such as JavaScript, Python, and Ruby.
- XML
    - Stands for eXtensible Markup Language, a markup language employed for accumulating and fascinating data.
    - Utilized to represent structured data and their connections to one another.
    - Contains tags, attributes, and values.
    - Is able to be utilized for data relations, document authoring, and web services about COVID Vaccines.
- HTML
    - Stands for HyperText Markup Language, a markup language used for assembling web pages.
    - Utilized to format content and illustrate its formation.
    - Consists of tags, attributes, and content.
    - Is able to be used with CSS and JavaScript to develop vigorous and interactive web pages.
    - In this dataset, HTML files with divert you to a website that will carry the COVID Vaccine data that you are looking for.
- Often the datasets can be seen to have multiple file types for one of the topics. This can be witnessed in the picture below which has an HTML, HTML ArcGIS GeoServices REST API, CSV, and GeoJSON.

Data.gov previously had an array of files that is supported by a number of files.
These are the list of the files supported on the platform four our databases include:

- CSV
- RDF
- JSON
- XML
- HTML
- ZIP
- GeoJson
- ArcGSI GeoServices REST API

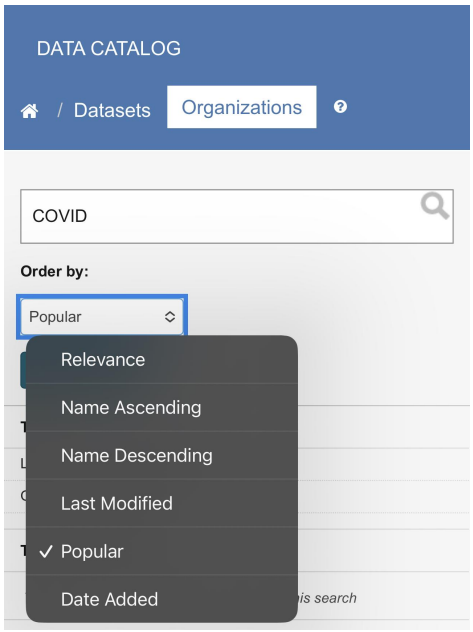
Filters and Searching for Data

The “Filters” section of a website allows a user to narrow down their search results. This make it easier to navigate their search results and reach their desired source at a faster and more precise rate. In this case, after the user accesses the Data.gov website, and they type their keywords into the search bar, a list of search results will appear along with the

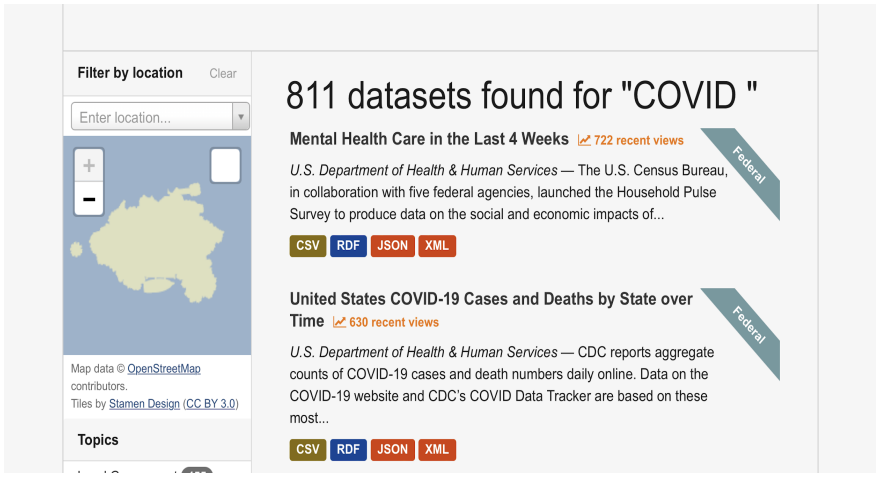


filters section of the website. On the Data.gov website, the filters bar will appear on the left-hand side of the screen as pictured above. The filters that can be altered by the user includes: location, topics, topic categories, dataset type, tags, format organization, publishers, and bureaus. The white numbers in the grey bubbles next to each filter choice represent the number of results that meet the criteria of the filter. For example, if the user wanted to filter the Dataset Type to geospatial, their search results would go from 126 dataset results to 14 dataset results. It is important to keep in mind that the choices under each filter category vary with each search inserted in the search bar. For example, the list of publishers to filter through when searching Covid Vaccine Datasets will be different from US Housing Crisis Datasets.

Users can also order their data set results from relevance, name ascending, name descending, last modified popular, or added data set.



After that users are able to choose their datasets by location:



There are also other relevant categories that users are able to use to their liking. This is meant to further help users find the dataset that they're looking for. The different filter options is as follows::

- Topics
- Topic categories
- Dataset types
- Tags
- Format
- Organization types
- Organizations

- Publisher
- Bureaus

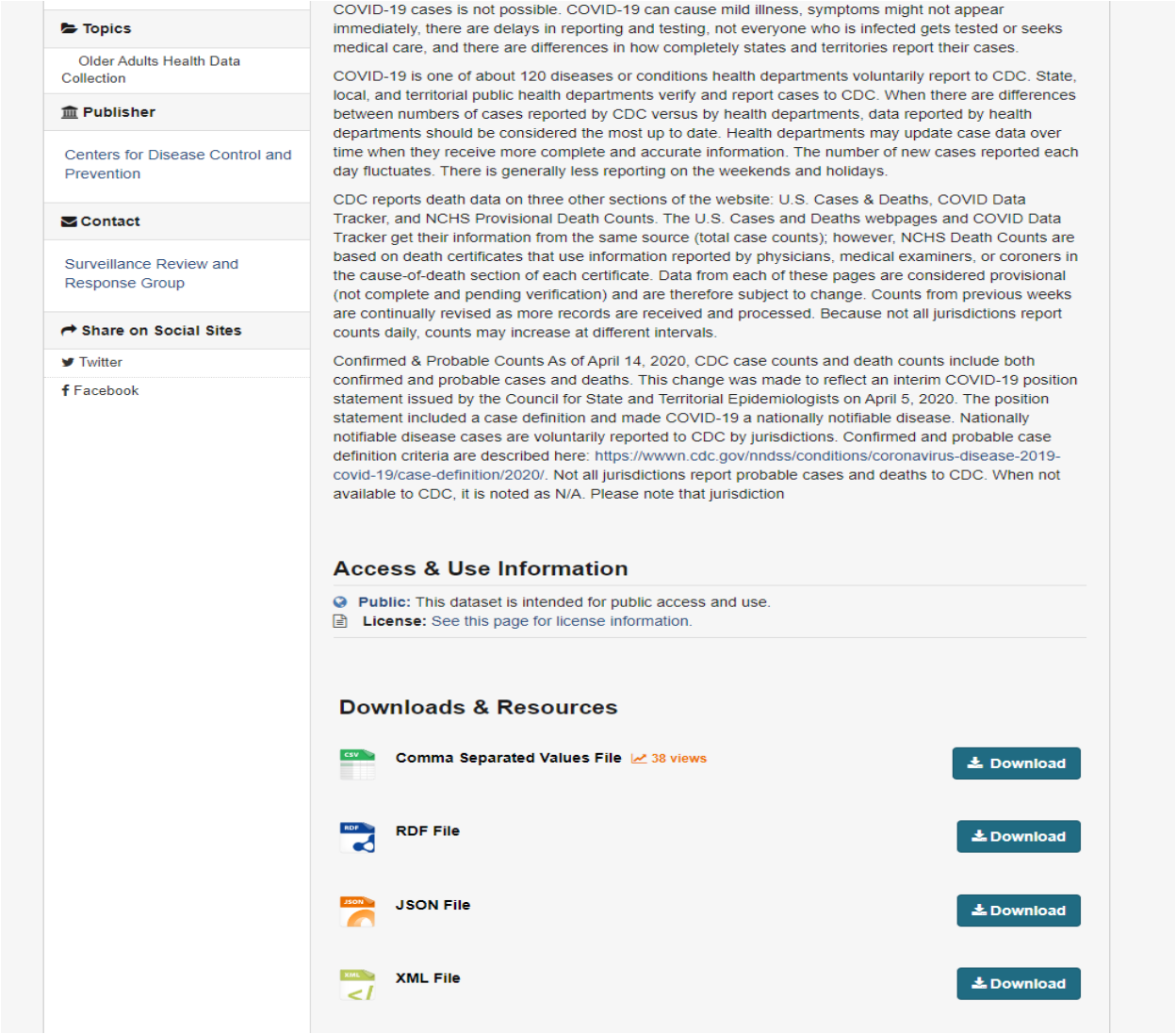
Exporting Data:

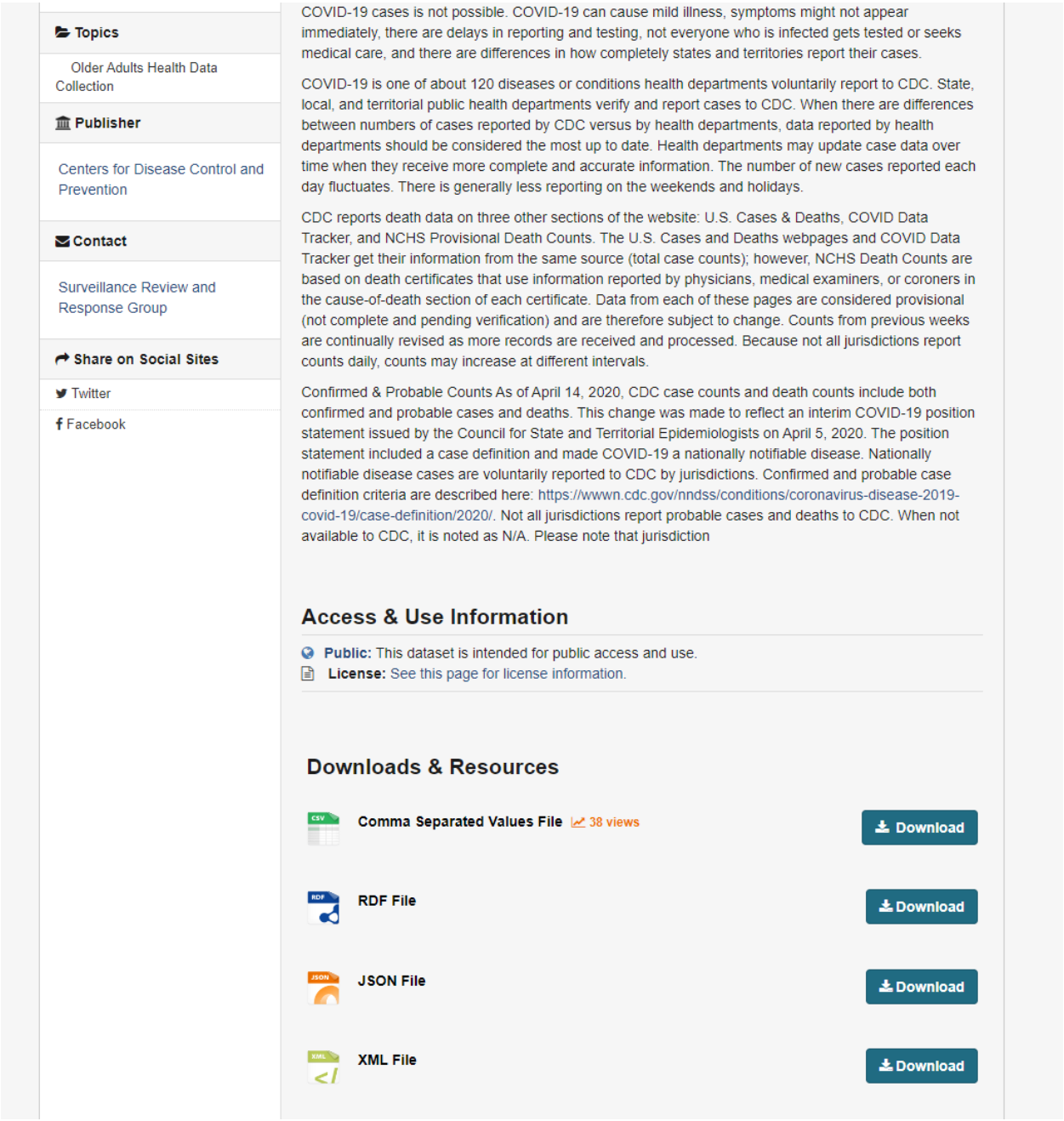
[Downloading the Data Sets:]

- How to download from the dataset

In order to download the data set you will have to navigate to the download and resource page from there you have several different options to select.

- Each data collection has a set of data that can be opened with programs such as CSV formatted data and Excel.
- Data can also be downloaded from the Dataset
- Each data provides different variations of files that can be downloaded

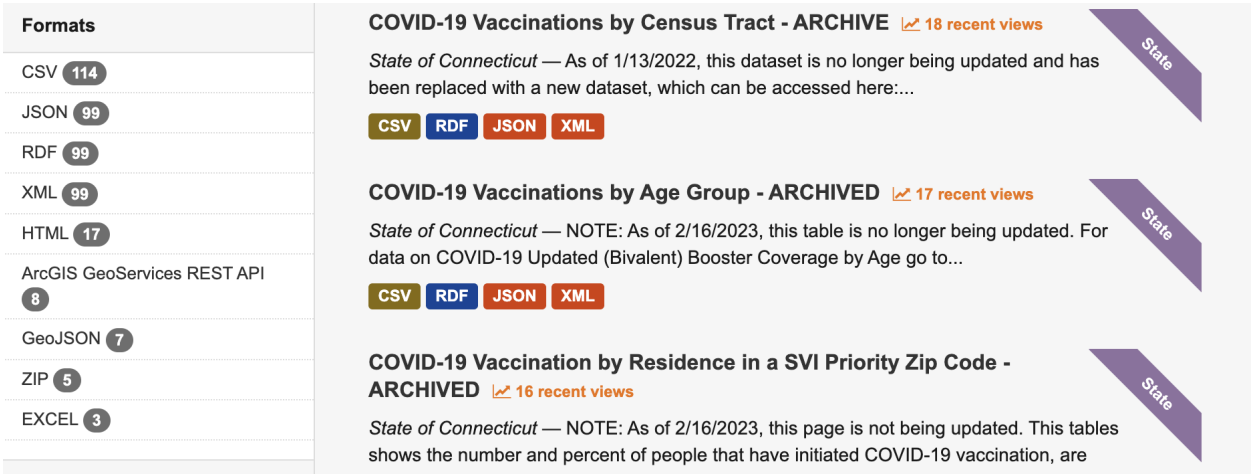




- Possible programs they can use to analyze the datasets
  - XLS files
  - Zipped
  - CSV files
  - HTML
  - ArcGIS
  - Data.gov Map preview

What Are The Formatting Options?

- CSV
- XML
- JSON
- RDF
- HTML
- ArcGIS GeoServices REST API
- GeoJSON
- ZIP
- EXCEL



Which Do I Choose?

When downloading data sets from data.gov, it's important to be aware of the different file formats that are available. All of the file formats available on data.gov can be opened in various text editing applications, including popular options like VSCode and Atom. However, if you plan to use Python for data analysis, you'll want to import the data into a Pandas data frame or other appropriate data structure.

To import a CSV file into a Pandas data frame, you can use the read_csv function provided by the Pandas library. To import a JSON file, you can use the read_json function, and to import an XML file, you can use the ElementTree library. ZIP files can be extracted using file extraction tools like WinZip or 7-Zip, while XLS files can be opened and manipulated in Microsoft Excel.

Additionally, there are numerous other Python libraries that can be employed for data analysis, including NumPy, Matplotlib, Seaborn, and Plotly. These libraries provide powerful tools for manipulating and visualizing data and can be especially useful when working with large data sets.

If you're interested in creating data visualizations, there are numerous tools available beyond Python libraries as well. Tableau is a popular one and can connect to various data sources. When working with Tableau, it is recommended to use a file format that can easily be imported, such as a CSV or Excel file. These file formats are easy to work with and can be quickly imported, allowing you to start analyzing and visualizing your data right away

Overall, when selecting a file format for your data set, it's important to consider your data analysis needs, technical abilities, and whether or not you'll need to create visualizations. By utilizing the appropriate software tools and libraries, such as Python libraries, Tableau, or more you can efficiently analyze and visualize your data to gain meaningful insights.

File Conversion:

- File conversion refers to the transfer of a file from one format to another. It is typically used to manipulate the data into a different type so it can be used on a different program
- One easy method is using conversion websites
    - Cloudconvert is very useful in converting over 200 different types of file formats.
    - Convertio is very useful in converting images, documents, spreadsheets, ebooks, audio, etc
- Another method for converting your data is using the built-in Excel conversion tool.
    - To import HTML, XML, or text scripts into an Excel file: Go to the data tab, select Get & Transform Data group, and click on the type of file you are using.
- Another method is using the window to change the extension
    - Open Windows File Explorer and select a file to change, Replace the old file extension with a new one to change your file, and Push the "Enter" button on your keyboard to save your changes

COVID-19 Vaccine Metadata

After researching information about the COVID-19 Vaccine Datasets on DATA.GOV, we assessed and survey the COVID-19 Vaccine related datasets. This product displays a clear picture of the available dataset results as of March 04, 2023. The results contained a total of 144 datasets.

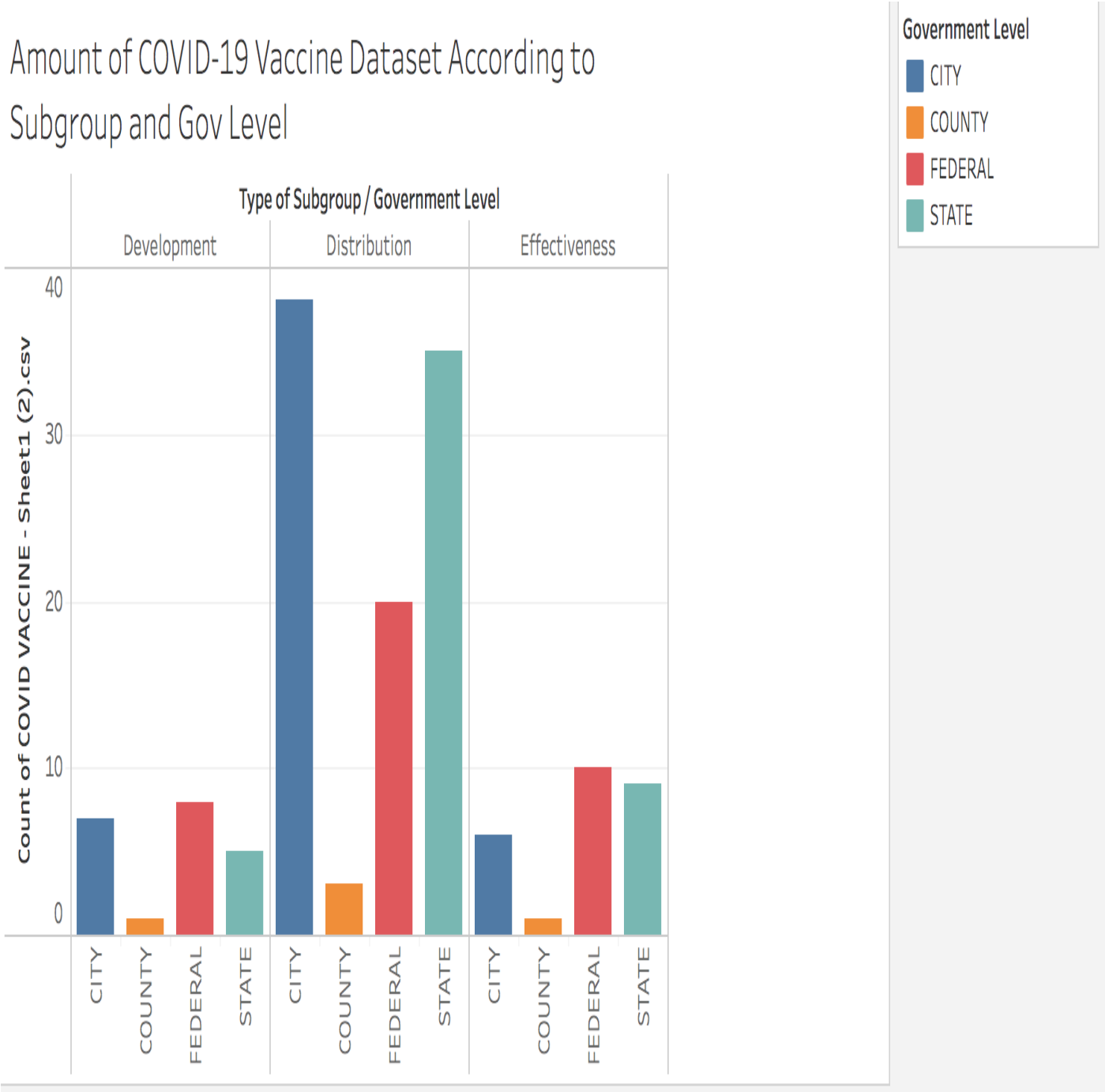
The data collection process consisted of our team going through each of the datasets manually and verifying the fields such as Title of Dataset, Type of Dataset, Link, Government Level, Date Released, Date Updated, Vaccination SubGroup, and Agency. Our collection is available on a Google Sheet which is a cloud base application, that can be accessed and exported in various formats for the user's needs.

The organization of the Spreadsheet:

Variable	**Operational Definition**
Title	The specific name of the dataset
File type of Dataset	A list of the various file types that is available for the dataset
Link	The URL or web address where the dataset can be accessed.
Government Level	The level of government accountable for compiling and publicizing the dataset.
Date Released	The date when the dataset was initially made available to the public.
Date Updated	The date when the dataset was last updated or modified.
Vaccination SubGroup	The specific subgroup within the dataset analyzed in relation to vaccination. ● Vaccine Development: ○ Includes datasets that supply information on the various types of COVID-19 vaccines being created and the rate of COVID-19 infection in areas. ● Vaccine Distribution: ○ Includes datasets that supply information on the various types of COVID-19 vaccines that are distributed to different regions, countries, and populations. ● Vaccine Effectiveness: ○ Includes datasets that supply information on the effectiveness of COVID-19 vaccines in preventing Covid infections, hospitalizations, and deaths.

Agency	The government agency is liable for accumulating and publicizing the dataset.

Metadata Data Visualization



Generating A Visualization Depicting The Data

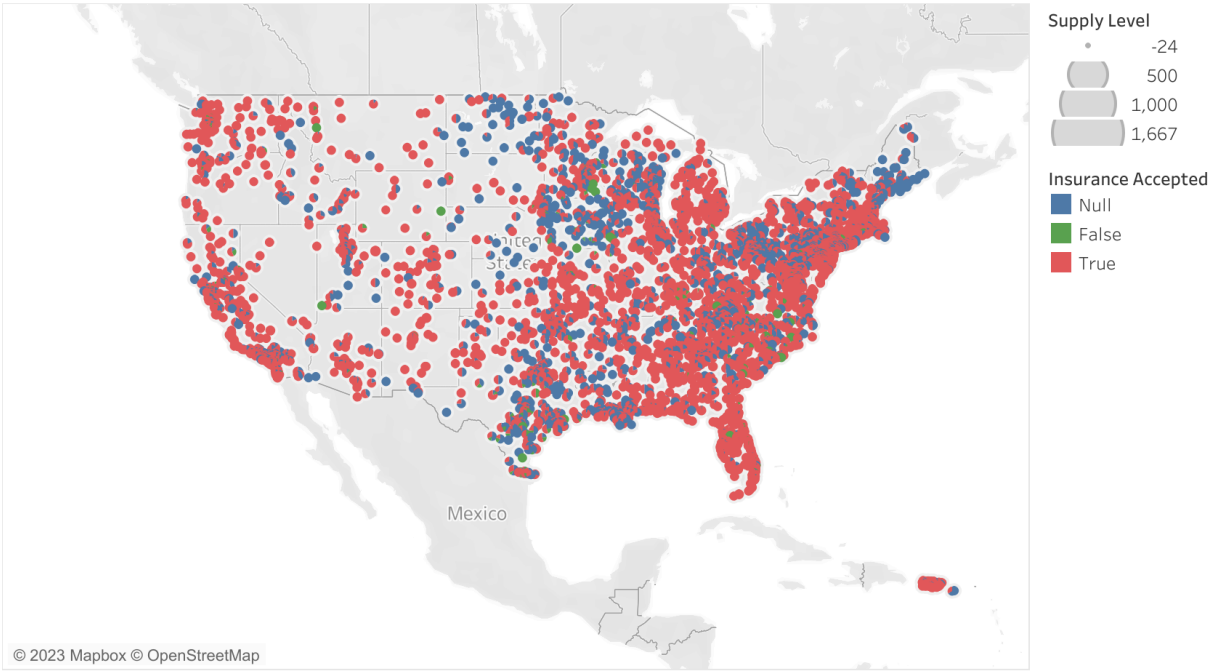
[Vaccines.gov: COVID-19 vaccinating provider locations]

COVID-19 vaccination providers in the United States

- Basic of dataset
    - Number of entries: 450,000+
    - File Format: CSV

A extensive exhibition of data that furnishes information on the sites of different vaccination providers all around the United States. It comprises data on the different kinds of vaccines offered, the availability of appointments, and the number of doses dispensed at each correlating location. The data is updated on a frequent basis to assure that it stays authentic and up-to-date. This dataset is a critical instrument for anyone pursuing to get vaccinated with the COVID-19 vaccine. It equips individuals with a rich of details on the availability of vaccines in distinct locations, which can be utilized to schedule appointments and confirm that people are able to get vaccinated as soon as possible. The dataset also presents a beneficial understanding of the distribution of the COVID-19 vaccine all across the nation, which can be helpful to inform public health policy and decision-making. It supplies extensive data on the availability of vaccines, which can be utilized to assist people plan their vaccinations and ensure that they are able to get vaccines in a timely manner.

National COVID-19 Vaccination Location of Supply Level and Acceptance of Insurance

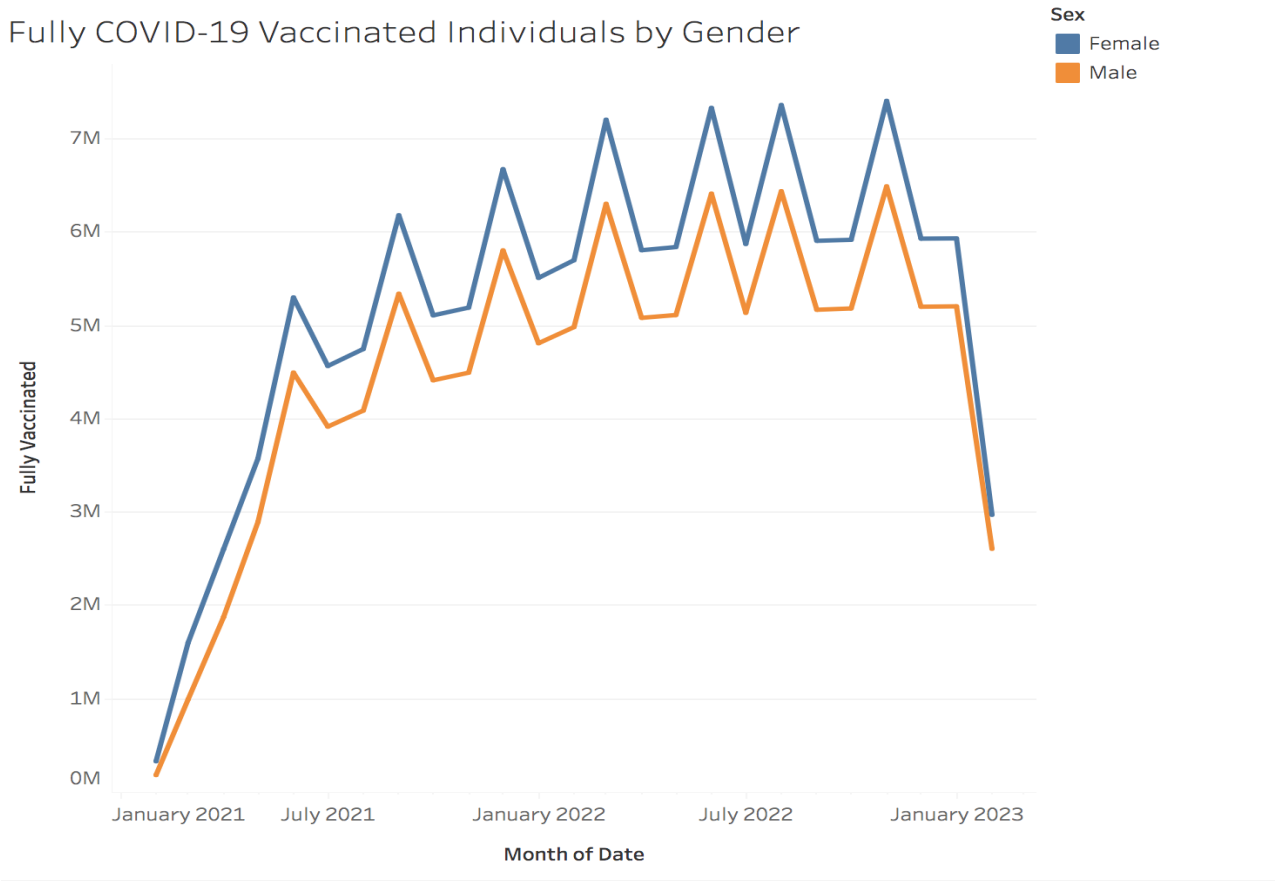


[COVID-19 Vaccination Age and Sex Trends in the United States, National, and Jurisdictional]

COVID-19 vaccination rates in the United States

- Basic of dataset
    - Number of entries: 517, 9 Tables
    - File Format: CSV

The Centers for Disease Control and Prevention (CDC) has been accumulating data on COVID-19 vaccination endeavors. The dataset supplies an abundance of information on the vaccination endeavors in the United States. The dataset comprises information on the number of individuals vaccinated by age and sex, as well as the number of people who have obtained one or both doses of the vaccine. The data is offered at the federal level, as well as at the state and territorial levels. One of the fundamental acuities from this dataset is the variation in vaccination rates by sex and age. The dataset also supplies details on the number of individuals who have acquired one or both doses of the vaccine. This knowledge can be utilized to advise public health policy and to assure that everyone who desires the vaccine can obtain it.

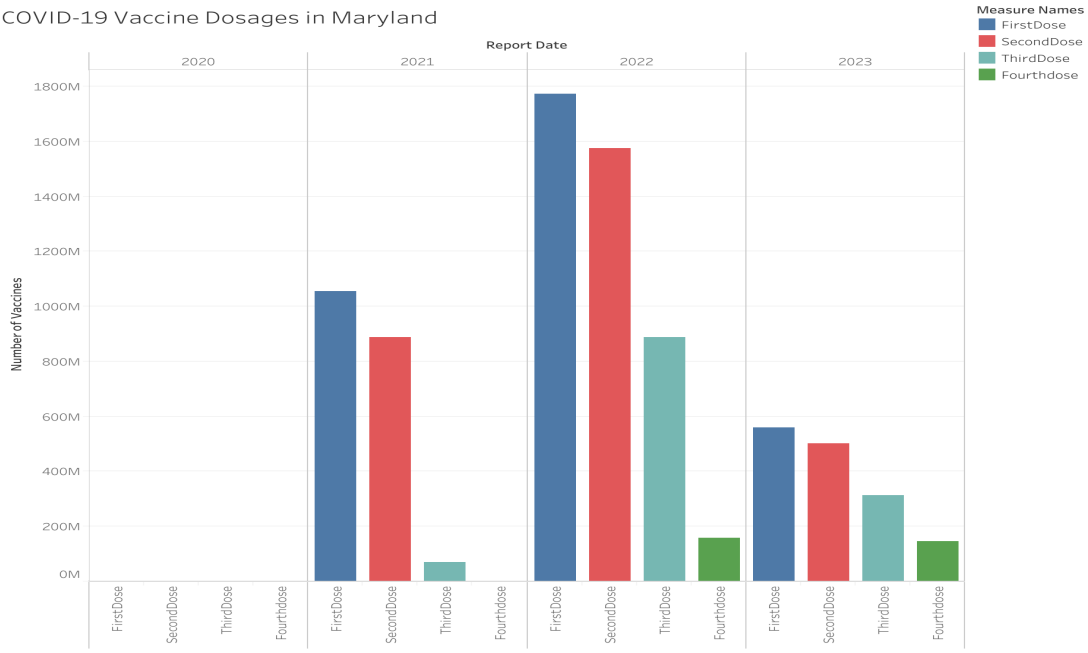


MD COVID-19 Total Vaccinations by Age - First, Second, Single, and Third Dose

COVID-19 vaccinations by age group in the state of Maryland

- Basic of dataset
    - Number of entries: 843
    - File Format: CSV

The dataset "MD COVID-19 Total Vaccinations by Age - First, Second, Single, and Third Dose" is a vital resource for researchers and policymakers. The dataset contains data about the first, second, third, and fourth doses of the vaccine. This data can be utilized to examine vaccination trends within various age groups and to consider the significance of the vaccination campaign. Researchers and policymakers are able to utilize this dataset to observe vaccination advancement and pinpoint regions where additional measures are required to boost vaccination rates. This dataset can furthermore be employed to analyze vaccination rates among various states and regions, which can assist to determine triumphant procedures for improving vaccination rates. The dataset even incorporates data regarding the age groups that have obtained the most vaccinations, which can be utilized to target vaccination actions toward distinct age groups. The dataset also delivers information on the diverse types of doses distributed, such as the first, second, third, and fourth doses. This data can be employed to dissect the significance of different vaccines and determine which vaccines are most influential in different age groups. The data can be utilized to pinpoint areas where more action is necessary to improve vaccination rates, target vaccination efforts toward specific age groups, and explore the significance of various vaccines.



Limitations

1. Inconsistent data formats: Datasets within data.gov can be provided in different formats, such as CSV, JSON, or XML. This can make it challenging to work with multiple datasets simultaneously, as additional data processing may be required to harmonize the formats. Whereas some Data sets don't have download
2. Varying data quality: The quality of datasets in data.gov can vary, depending on the data source and the processes used to collect and maintain the data. This may affect the reliability and accuracy of the datasets and the subsequent analyses.
3. Limited metadata: The metadata provided for each dataset in data.gov may be insufficient or inconsistent, making it difficult to understand the context and meaning of the data. This can hinder the effective use and interpretation of the data.
4. Accessibility: Some datasets in data.gov may require users to have specific software or tools to access or process the data, which could create barriers to entry for users with limited resources or technical expertise.

To address these limitations, it's important to be aware of the potential biases and challenges when working with data from data.gov. Additional efforts may be required to clean, process, and validate the data before conducting any analysis to ensure the quality and accuracy of the results.

Restraints

- Accessible issues due to possible lack of support for certain files
- Lack of detail about the topic due to the filter we applied to the search
- The data set might become outdated after some time because there is new information coming out on COVID-19.

Frequently Asked Questions (FAQs)

What other kinds of data can I research on this platform?

Data.gov harbors thousands of datasets on various different research topics from Covid-19 as discussed in this document, to child birth rates in the United States or Climate change effects in the Arctic to Food Shortage data in Eastern European countries. The scope of this data website is quite vast, meaning, any user could research almost anything.

What data visualization tools are compatible with the datasets on the platform?

The data visualization tools that are compatible with the datasets available on Data.gov are Google Charts, Tableau, Grafana, Chartist. js, FusionCharts, Datawrapper, Infogram, ChartBlocks, and D3.

Does Data.gov follow accessibility protocol?

Data.gov is a follower of the WCAG 2.1, Level AA which is recognized internationally and complies with Section 508 of the Rehabilitation Act, which ensures that the websites of American federal agencies are accessible to people with disabilities. Meaning, there are website alternatives to meet the needs of people with disabilities, such as text alternatives, content without text, using descriptive labels and headings, and ensuring that all content can be accessed using the keyboard alone.

Why was COVID-19 chosen as the example topic for this analysis?

Covid-19 was chosen as the example in this document, because it is a recent and relevant occurrence that affects many people around the world. The Covid-19

Why is CSV the most popular dataset format?

CSV is the most popular dataset format because it can be used in Excel, SQL, Python, Tableau, and more! It is also versatile as it can hold large amounts of data. This makes it adaptable and easy to use.

How could one use these datasets?

The datasets can be used for general research, vaccination research by individuals, and general use by medical professionals

Version History

Version	Month and Year
1.0	February 2023
2.0	March 2023