**Data Wrangling**

**Learning Objectives**

* Learn how to acquire and compile data through data collection
* Develop the skills to keep projects organized by building local file structures and keeping GitHub repo versions updated
* Learn how to review data types, build data profiles, and begin to develop an understanding of the feature dimensions of your data
* Be able to handle problems in your data such as null values, duplicates, and formatting issues

**Work to Complete**

In this unit, you'll:

* Work through three mini-projects:
  + An API mini-project
  + A NASA exercise
  + A text data handling exercise
* Submit ideas and a project proposal for your second capstone
* Wrangle data for your second capstone

Data wrangling is a crucial step in the process of preparing data so that you can identify insights, perform EDA, and build models. By wrangling your data at the beginning of any data science project, you'll set yourself up to be able to explore your data efficiently. You were introduced to data wrangling concepts while working on the guided capstone. This unit takes a deeper dive into each of these concepts and introduces new scenarios and constraints that are important for you to understand. At the end of this unit, you'll put your data wrangling skills to use by kicking off your second capstone. This capstone will feature prominently in your data science portfolio. Employers will expect to see proof of your ability to collect and clean data. Your portfolio will play an integral role in demonstrating your abilities.

**Capstone Two: Project Ideas & Proposal**

To kick off this unit, we'd like you to start planning out your second capstone project. For this project, you'll work through the entirety of the Data Science Method, starting with problem identification. First, you'll identify three possible project ideas. Once you've talked to your mentor about your ideas, you'll expand your idea into a complete project proposal. At the end of this unit, you'll take on the data wrangling step of this capstone.

1

**Capstone Two - Project Ideas**

Save

2 - 4 Hours

55 Points

For the first step of your second capstone, you'll use Google Dataset Search to find three datasets and brainstorm project ideas related to those datasets. Please note that you can also combine two or more datasets to answer a single project question.

Review the[**rubric**](https://docs.google.com/document/d/1_oSYbMy8zeUvld6DSAVtwfev21hYWeb3VdLdtR10of0/edit)to understand how your work will be assessed.

**Submission Steps**

1. Read [**these instructions**](https://www.springboard.com/archeio/download/dff20f432d924ac6a412b03c70c861f9/) to get started.
2. Write a description of three capstone project ideas. Your ideas can be broad and high-level. The descriptions should address the problem and identify the data you’ll use. You don’t need to talk about specific methods and techniques.
3. Submit a Google Doc link here. Please set sharing permissions to “anyone with a link can comment” so that your mentor can provide feedback. Please do not submit .pdfs, .ppts, or markdowns.
4. Review your ideas with your mentor during your next call.
5. (Optional) Post your idea, including a title and description, to the student community for feedback in the #ask-for-feedback Slack channel.
6. After collecting feedback, choose one idea and get ready for the next step: creating a project proposal.

**Data Science Career Track**

**Capstone: Project Ideas Rubric**

**Learning Objectives**

* Identify three potential projects with associated data to support the projects
* Practice using the Google Dataset Search tool

|  |  |
| --- | --- |
| **Criteria** | **Meets Expectations** |
| Completion | * A short description of each of the three ideas as a single Google doc has been submitted. |
| Process and understanding | * The submission contains three high-level ideas with links to the appropriate data that could be used to support the idea. * The ideas are framed around real-world business problems. |

Data Science Career Track  
Capstone Two: Project Ideas  
Project Overview  
For your second capstone, you’ll work through the entirety of the Data Science Method as you  
move through each step of the project — to get started, you simply need a few good ideas.  
Welcome to the ideation phase of your capstone! For this step of the project, you’ll use Google  
Dataset Search and other open data sources to identify up to three project ideas. Google  
Dataset Search combines the prowess of Google’s search engine with the nuance of Google  
Scholar. If you’d like to learn more about this search engine, click here.  
When identifying a dataset to work with, here are a few things to consider:  
● If you’re coming from a different professional background, you may want to search for a  
dataset that will bridge the gap between your previous experience and data science.  
● While we love Airbnb and Yelp, we don’t recommend that you use their datasets, as these  
datasets are used frequently in aspiring data science portfolios. Find a dataset that will  
set you apart from the crowd.  
● Note that Google’s Dataset Search returns paid resources as well as free ones. You may  
want to direct your search to the free datasets.  
● If you’re having a hard time coming up with ideas, check out this set of datasets to get  
your creative juices flowing.  
Project Steps  
Estimated Time: 2-4 Hours  
1. Explore datasets from Google Data Search. Some other search engines include:

● Quandl, US Government Open Data, UCI Machine Learning Repository, Kaggle  
competitions, Mode Analytics  
● Data is Plural is an email list that sends newly released and interesting datasets  
2. Find three datasets of interest and turn those into project idea drafts. You can also  
combine two or more datasets to answer a single question.  
For your initial project ideas, please:  
● Include a short description of each idea. The description should briefly discuss  
the problem and the data you’ll use to solve it. At this point, there’s no need to  
outline specific methods and techniques.  
● Post your idea, including the title and description, to the community and solicit  
feedback from both other students.  
One word of caution:  
Be sure to choose your datasets wisely; check for a reputable source and detailed data  
documentation or metadata to describe how it was collected and the appropriate level of  
context required for using it responsibly.  
Student Capstone Examples  
To find inspiration, please check out our collection of awesome student Capstone Projects.

 [ce](https://www.springboard.com/blog/category/career-advice/)

 [Cybersecurity](https://www.springboard.com/blog/category/cybersecurity/)

 [Data Science](https://www.springboard.com/blog/category/data-science/)

 [Coding](https://www.springboard.com/blog/category/software-engineering/)

 [Sales](https://www.springboard.com/blog/category/sales/)

 [Marketing](https://www.springboard.com/blog/category/business-and-marketing/)



Top of Form

Bottom of Form

* [Blog](https://www.springboard.com/blog)
* >
* [Data Science](https://www.springboard.com/blog/category/data-science/)
* >
* **15 Free Data Sets for Your Next Project or Portfolio**

# 15 Free Data Sets for Your Next Project or Portfolio

Sakshi GuptaSakshi Gupta | 8 minute read | June 29, 2022



In this article

* [What Is a Data Set?](https://www.springboard.com/blog/data-science/free-public-data-sets-data-science-project/#h0)
* [Free Data Sets To Analyze](https://www.springboard.com/blog/data-science/free-public-data-sets-data-science-project/#h1)
* [Data Set FAQs](https://www.springboard.com/blog/data-science/free-public-data-sets-data-science-project/#h2)

If you’re early in your career as a data scientist, you might want to consider taking on some personal projects. There are two reasons why.

Firstly, it’s a way for you to test yourself. You’ve probably spent many months working through data science theory and studying different approaches to analyzing data. But how do you know you’ve actually gained useful real-world skills? You can do that by choosing a problem that seems interesting to you and unleashing your newfound analytical skills to solve it.

Another important reason you should build projects is so that you have something to put in your data science portfolio. Recruiters prefer to look at candidates’ portfolios instead of reading long statements of purpose or lists of data science classes they’ve taken. A [portfolio](https://www.springboard.com/blog/data-science/data-science-portfolio/) evinces that you’ve got practical skills and the ability to take projects from conception to completion.

Now if you want to work on a data science project, then you can’t do that without the data. If you’re wondering where you can source data from, we’ve got you covered. We’re going to c

But before we get into those resources, let’s take a look at what a data set is.

## What Is a Data Set?



You’ve probably figured out partially what a data set is from what it’s called. It is, of course, a set of data points. But along with that, it’s also important to remember a few other characteristics that a data set must exhibit.

A data set is always formed of related data. Let’s say you have a data set about a housing subsidy program. In that case, the data would include data points relating to the prices of houses over time, the demographics of the buyers, the areas where these programs are run, and so on. All of these data points are related and therefore would constitute a data set.

Secondly, the data in a data set is always discrete. Each record is independent and can take the form of only a finite value.

Data sets are most commonly stored in a tabular format. Every column in the table corresponds to a specific category of information. The rows are the data values that fall under that specific category.

For example, assume that you have a database on stock prices during a certain period. Some of the columns that you would have in this data set are company name, company stock price, change in stock price year on year, and so on. As you can see, the values that would be entered in this table would be related, discrete, and structured.

Table

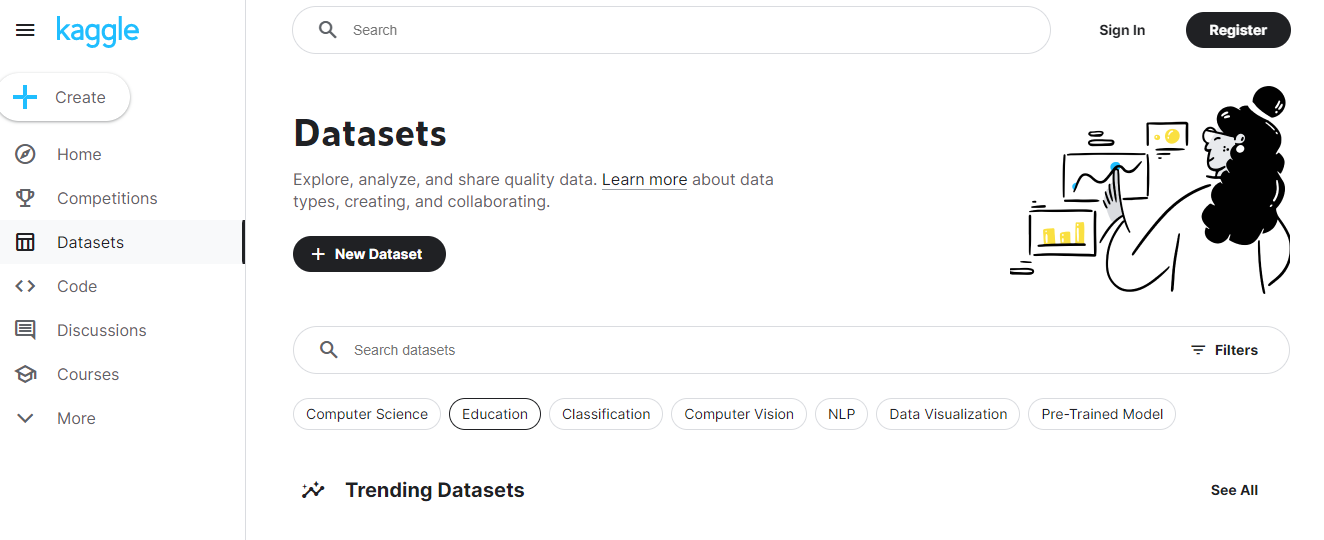
Description automatically generated with medium confidence

## Free Data Sets To Analyze

Now that we know what a data set is, we can move on to looking at some of the best public data sets that are out there. These data sets have been sourced from government agencies, private companies, and public institutions. All of the data available in them is structured, so you don’t have to worry about [cleaning data](https://www.springboard.com/blog/data-analytics/data-cleaning/).

### Free General Data Sets

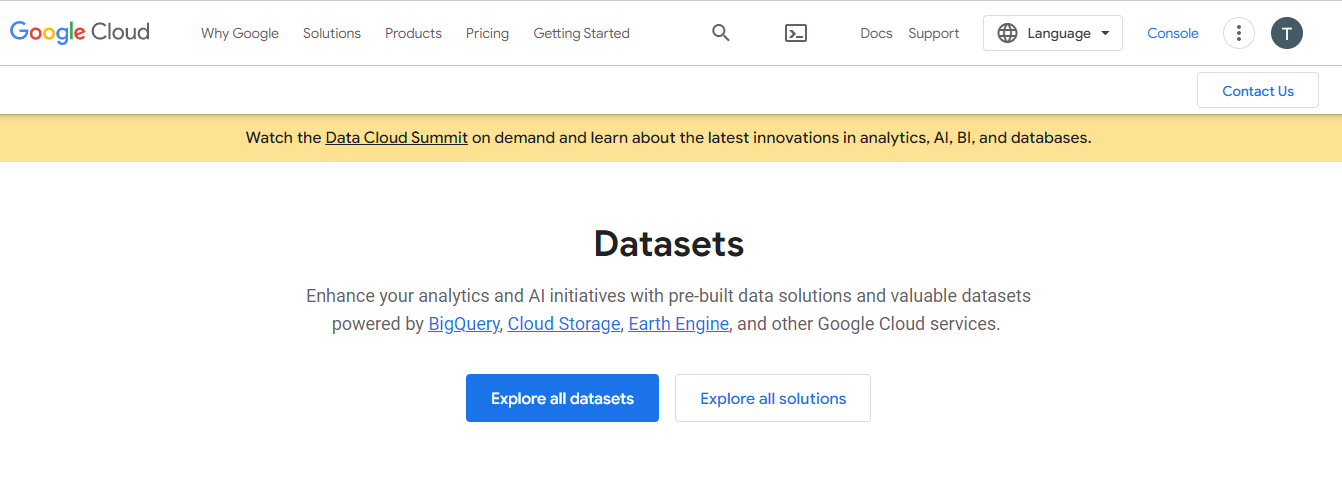
#### [Kaggle](https://www.kaggle.com/datasets)



Kaggle is a community that has been built specifically for [data scientists](https://www.springboard.com/blog/data-science/what-does-a-data-scientist-do/) and [machine learning engineers](https://www.springboard.com/blog/data-science/machine-learning-job-description/). The goal is to have a place where members can work on Kaggle data problems together and access data sets so they can regularly practice [data analysis](https://www.springboard.com/blog/data-analytics/data-analysis-methods-and-techniques/).

Kaggle has something to offer for data scientists across levels, whether that’s a simple data set for students or something advanced for a data scientist looking to work on their artificial intelligence chops. The platform is also known for hosting regular competitions where you can go up against other data scientists to solve real-world problems posted by companies.

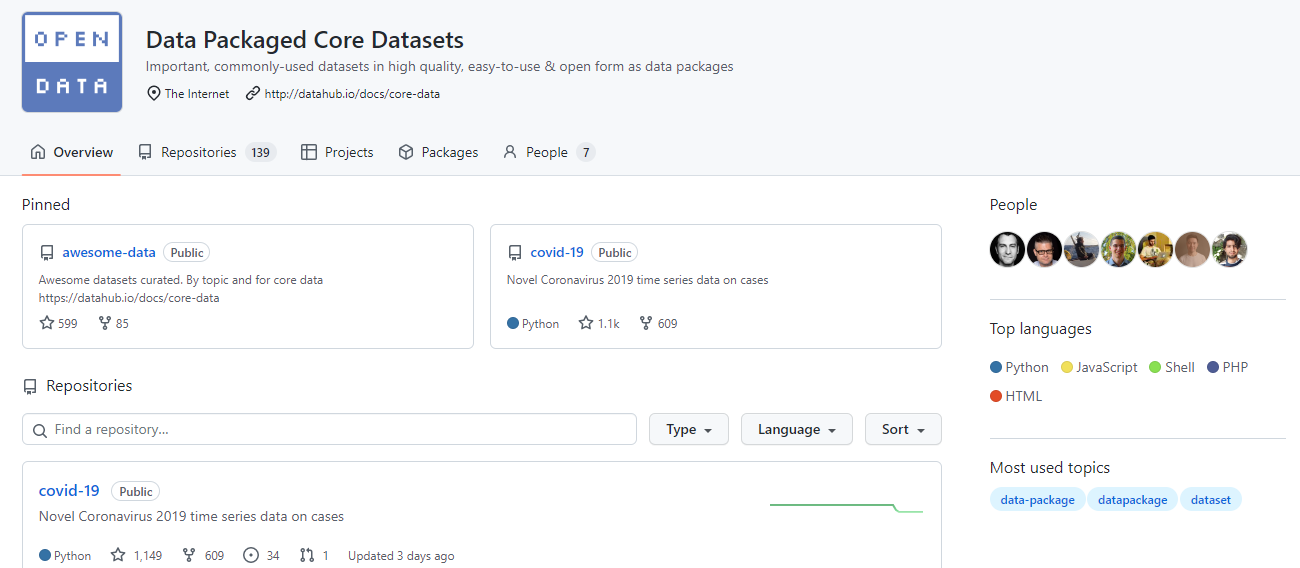
#### [Google](https://cloud.google.com/datasets)



The Google Cloud marketplace comes with a website that offers data sets that have been sourced from various Google products. So if you want an excellent data set from services like Google Trends, Google Patents Research, and Community Mobility Reports, this is where you can find it.

Google also offers a collection of repositories from commercial and public data sets. You can conduct your analyses on Google Cloud or download the data sets and use your own tools for the job.

#### [Github](https://github.com/datasets)



You probably think of [Github](https://www.springboard.com/blog/software-engineering/github-for-dummies/) as a version control tool, but did you know that they also offer a wide variety of data sets that you can use for your personal projects? These are all available for free and you can quickly port the data into your project when you need it.

Let’s take this [glacier mass balance data set](https://github.com/datasets/glacier-mass-balance), for example. This fantastic data set provides information on the mass of reference glaciers across the world. You can use this and similar data sets to conduct analyses on a wide range of topics.

### Free Government Data Sets

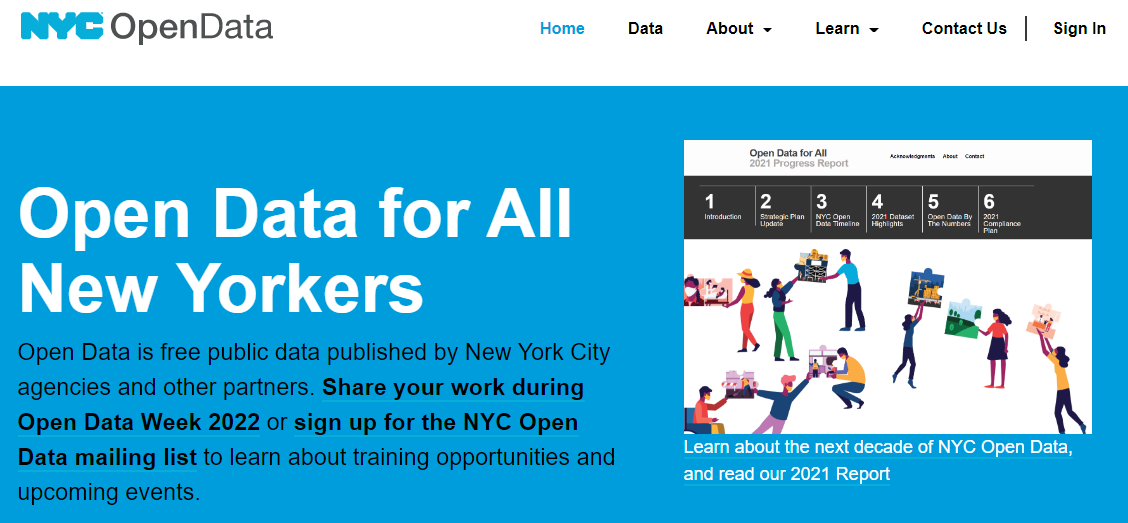
#### [Data.gov](https://data.gov)



Data.gov is where all of the American government’s public data sets live. You can access all kinds of data that is a matter of public record in the country. The main categories of data available are agriculture, climate, energy, local government, maritime, ocean, and older adult health.

Along with giving access to this collection of repositories for free, the website also has various resources for data scientists. You can use it to learn more about [data analysis tools](https://www.springboard.com/blog/data-analytics/data-analytics-tools/), data management frameworks, and case studies of projects taken up by data scientists who work in government.

#### [NYC Open Data](https://opendata.cityofnewyork.us)



What Data.gov does at the federal level, NYC Open Data does for New York City. This website is a collection of repositories that offer data sourced from various public institutions that govern the city.

The main categories of data available here are business, city government, education, environment, and health. You can also browse data sets compiled by different agencies, such as the [Financial Information Services Agency (FISA)](https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Financial+Information+Services+Agency+%28FISA%29) or the [Mayor’s Office of Climate Policy and Programs (CPP)](https://data.cityofnewyork.us/browse?Dataset-Information_Agency=Mayor%27s+Office+of+Climate+Policy+and+Programs+%28CPP%29).

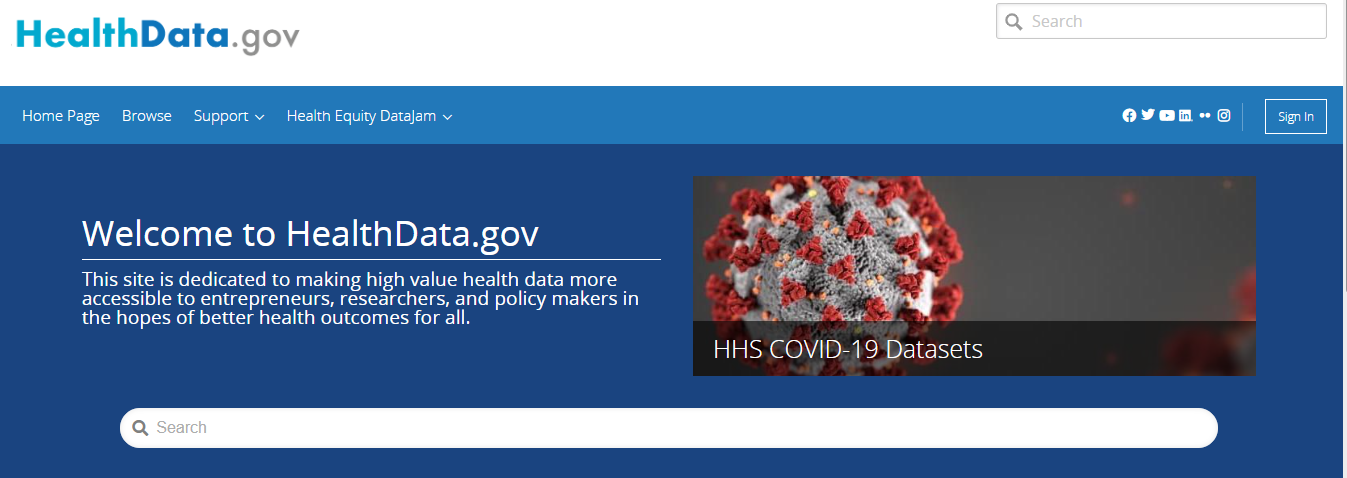
#### [Data.europa.eu](https://data.europa.eu/en)



This is the official portal for all of the public data that is offered by the European Union. The scope of the available data is broken down into national data, European data, and international data. You can find a detailed data set for just about any aspect of European life here, covering economic indicators, law enforcement agencies, health care institutions, and more.

### Free Health Data Sets

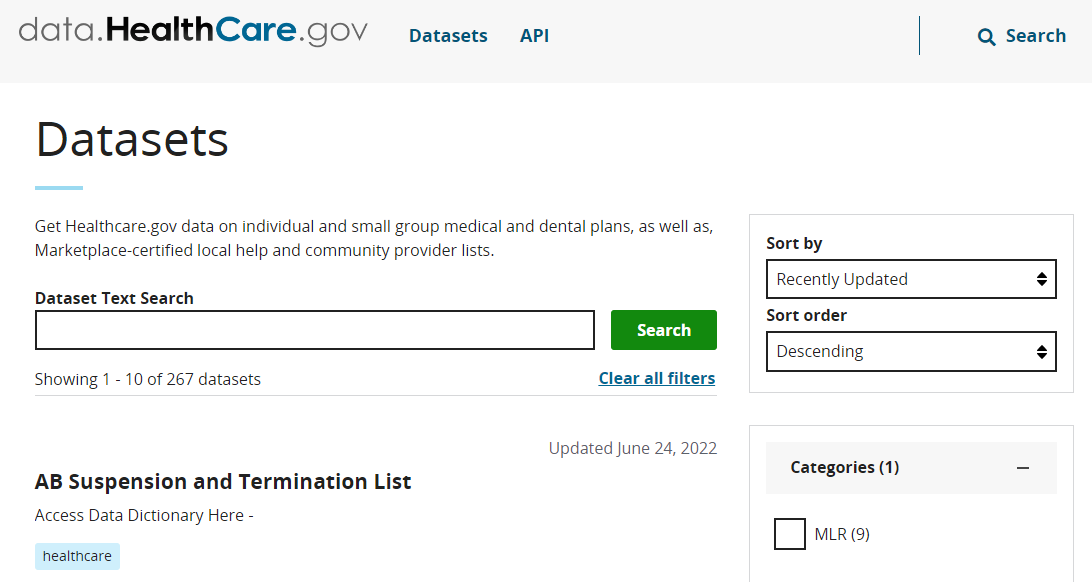
#### [Healthdata.gov](https://healthdata.gov)



Healthdata.gov is a repository of freely available healthcare data from the US government. It is managed by the U.S. Department of Health and Human Services Office.

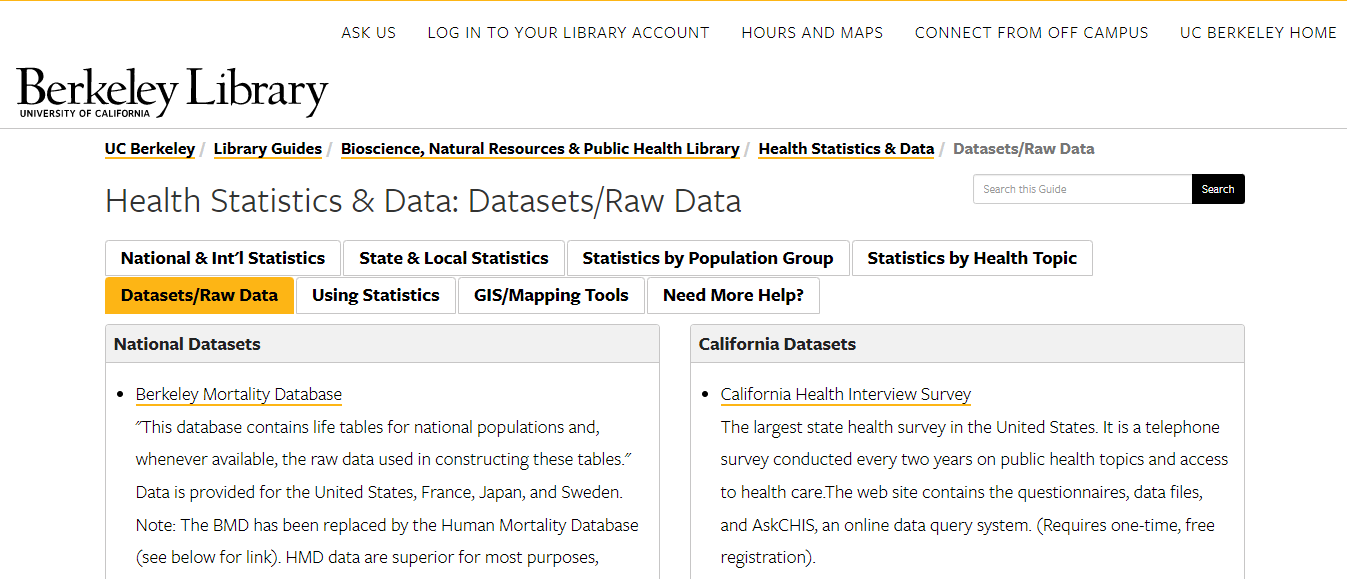
This website is a treasure trove for anyone interested in healthcare data. You can find public data sets on everything ranging from cancer incidence to COVID-19 prevalence and impact. Working on these data sets can be especially helpful if you plan on getting a [data science job in healthcare](https://www.springboard.com/blog/data-science/data-science-in-healthcare/).

#### [Healthcare.gov](https://data.healthcare.gov/datasets)



This is a federal website managed by the U.S. Centers for Medicare & Medicaid Services. The data sets available on this website are specifically geared towards medical and dental plans for groups and individuals. There’s also an API with clear documentation in case you want to source your data directly into a web application.

#### [Health Statistics & Data](https://guides.lib.berkeley.edu/publichealth/healthstatistics/rawdata)



The Berkeley Library Health Statistics and Data website provides free access to a large variety of data sets. That includes data sets that are both nationwide statistics and specific to California state.

#### Get To Know Other Data Science Students

[](https://www.springboard.com/success/jonathan-orr)

Jonathan Orr

Data Scientist at Carlisle & Company

[](https://www.springboard.com/success/meghan-thomason)

Meghan Thomason

Data Scientist at Spin

[](https://www.springboard.com/success/karen-masterson)

Karen Masterson

Data Analyst at Verizon Digital Media Services

### Free Environment Data Sets

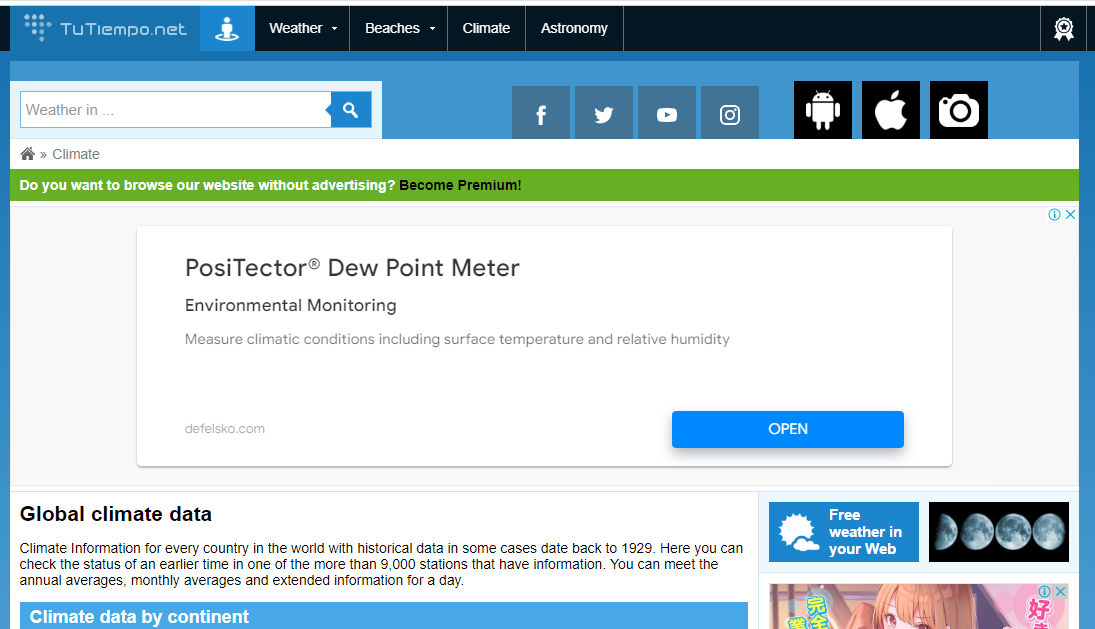
#### [US Climate Data](https://www.ncdc.noaa.gov/cdo-web/)



The National Centers for Environmental Information offers its climate data for free through these public data sets. The goal of the undertaking is to make global climate data available for analysis and study.

The public data sets available on this website constitute a cross-section of data across months, seasons, and years. You can get information on things like temperature, wind, precipitation, and other climate data here. The site also offers specialized tools that you can use to access this climate data.

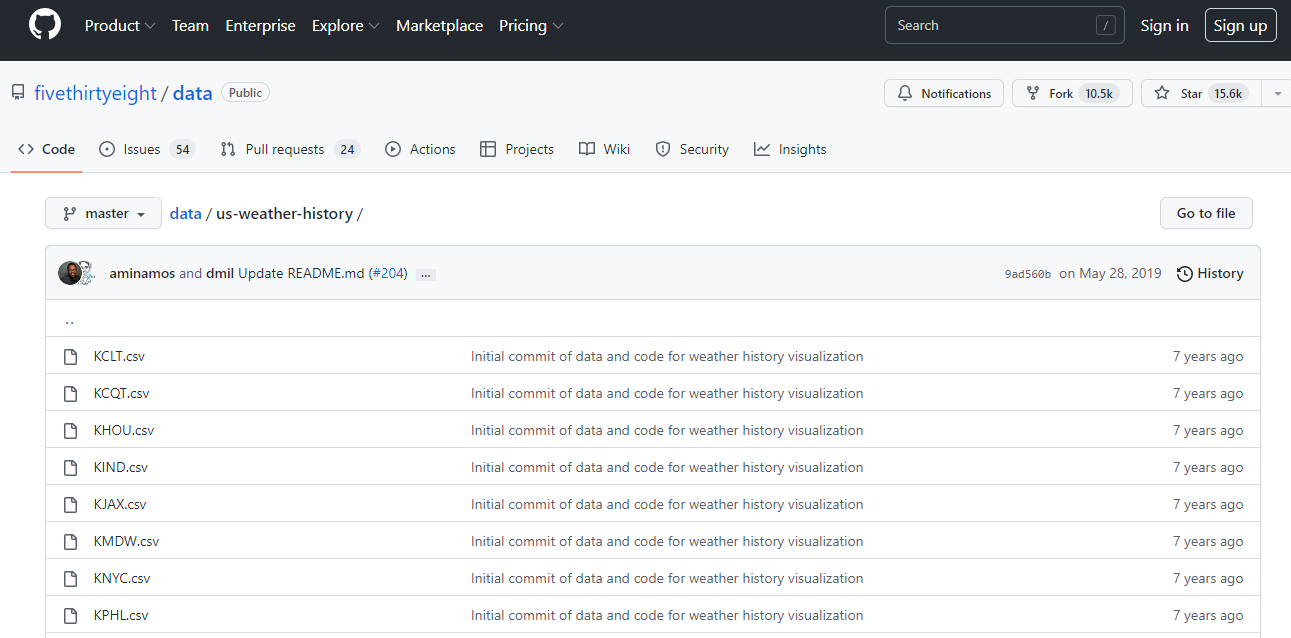
#### [Global Climate Data](https://en.tutiempo.net/climate)



If you want to do a [data science project](https://www.springboard.com/blog/data-science/data-science-projects/) on climate data, then this website offers just about every kind of data set that you could possibly need. This website by Tutiempo Network contains public data sets with climate data for every country on the planet. Some of this data goes back to the first half of the 20th century.

The data on this website is sourced from over 9,000 weather stations. It is easy to break the available data sets down by continent or country if you want to focus your analyses on one particular region.

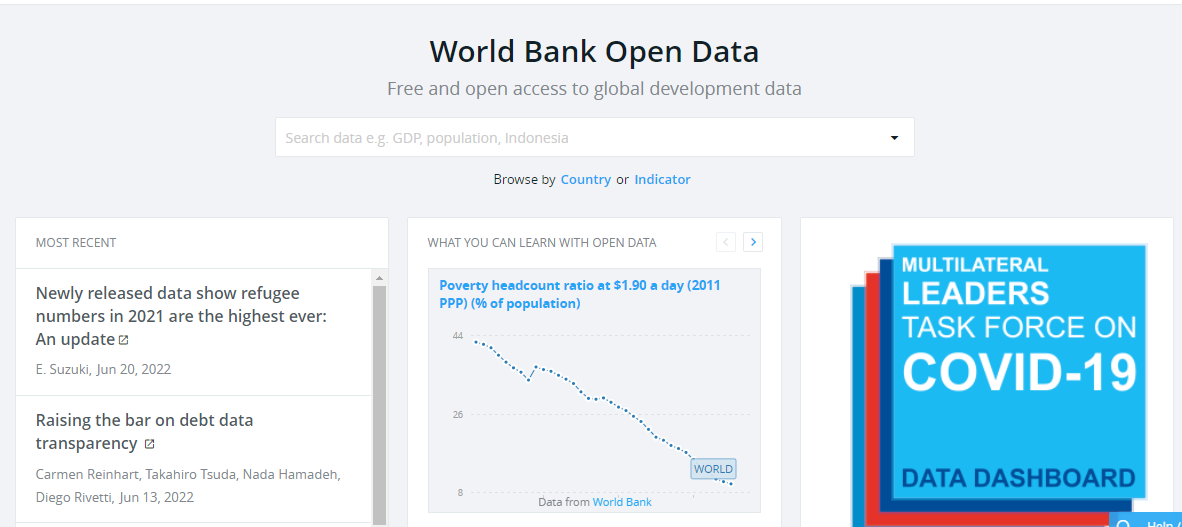
#### [US Weather History](https://github.com/fivethirtyeight/data/tree/master/us-weather-history)



Five Thirty-Eight—the website known for its data journalism stories—used this US Weather History data repository to produce its 2015 story [What 12 Months of Record-Setting Temperatures Looks Like Across the US](https://fivethirtyeight.com/features/what-12-months-of-record-setting-temperatures-looks-like-across-the-u-s/). Analyzing this data set is a good way to understand how [data science](https://www.springboard.com/blog/data-science/data-science-definition/) connects with [storytelling](https://www.springboard.com/blog/data-science/data-storytelling-what-it-is-and-how-your-company-can-use-it/). You can use the story as inspiration to work on your [data visualization skills](https://www.springboard.com/blog/data-analytics/7-types-of-data-visualizations-and-how-to-use-them/).

### Free Economic Data Sets

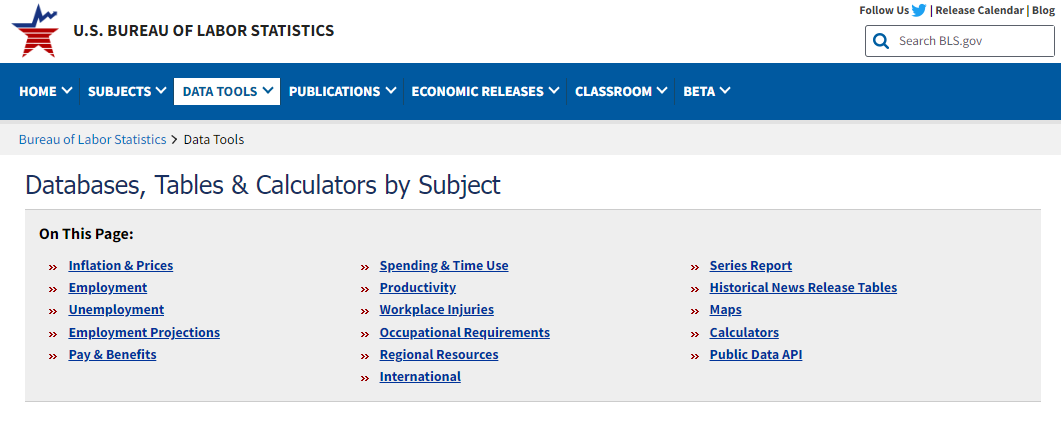
#### [World Bank Open Data](https://data.worldbank.org)



This is a public website with data offered by the World Bank. Due to the nature of this institution, you know that you’re going to get access to economic data from across every continent on the planet.

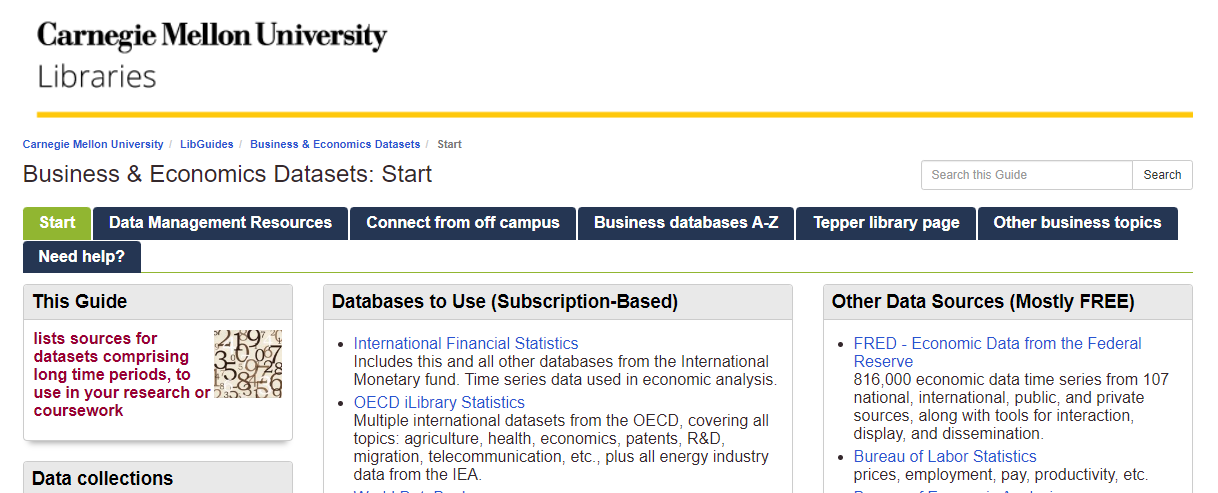
Each data page allows you to download data in bulk in a CSV file and other file formats. There is also an API using which you can access this data to analyze or display on your own tool.

#### [US Employment and Labor Data](https://www.bls.gov/data/)



This is a website with public data that pertains to employment levels and labor information for the United States. You can access data that covers things like inflation and prices, workplace injuries, productivity, employment benefits, etc.

#### [Business and Economics Data Sets](https://guides.library.cmu.edu/datasets)



These are business-related data sets that are made available by the Carnegie Mellon library. You can peruse data that pertains to all kinds of national and international economic information. Some examples include economic data from the federal reserve, data from the International Labor Organization, and data from the World DataBank.

## Data Set FAQs

We’ve got answers to your most frequently asked questions about data sets.

### How Do I Know if a Free Data Set Is Complete?

You can make sure that the data you source is complete by choosing reliable sources for your data sets. Always go with data that has been made available by governments, reputed private companies, and public institutions.

### Are All Data Sets Free?

Not all data sets are free. Some require users to pay for access to download the data or to use an API that gives access to the data.

### Can You Make Your Own Data Set?

Yes, you can build your own data set by sourcing data from various sources like social media sites, online directories, and so on.