1. FiveThirtyEight

FiveThirtyEight is an interactive news and sports site that has some incredible data visualizations (which you should totally check out). They make a lot of their data open to the public, meaning you can download and play with the source data yourself! Here are some examples:

- Airline Safety — contains information on accidents from each airline.
- US Weather History — historical weather data for the US.
- Study Drugs — data on who's taking Adderall in the US.

2. BuzzFeed News

BuzzFeed makes the data sets, analysis, libraries, tools, and guides used in its articles available on GitHub. Check them out to learn from some of the best! Here are some examples:

- Federal Surveillance Planes — contains data on planes used for domestic surveillance.
- Zika Virus — data about the geography of the Zika virus outbreak.
- Firearm Background Checks — data on background checks of people attempting to buy firearms.

3. Socrata

Socrata hosts cleaned open source data sources ranging from government, business, and education data sets. Here are some examples:

- White House Staff Salaries — data on the salary of each White House staffer in 2010.
- Radiation Analysis — data on which milk products in the US were radioactive based on location.
- Workplace Fatalities by US State — the number of workplace deaths across the US.

4. Awesome-Public-Datasets on GitHub

This github hosts a library of awesome, public datasets! They are all sorted by category and link you straight to the hosting website. Here are some examples:

- Global Climate Data — climate information for every country in the world with historical data (some date back to 1929).
- Heart Rate Time Series Data — two series of data contains 1800 evenly-spaced measurements of instantaneous heart rate from a single subject.
- Plane Crash Database — plane crash data dating from 1929.

5. Google Public Datasets

Google lists all of the data sets on a page. Google has a cloud hosting service called Google Cloud Platform (GCP), and you can query using a tool called BigQuery to explore these datasets. You'll need to sign up for a GCP account, but the first 1TB of queries you make are free! Here are some examples:

- US Name Data Set — contains all names from social security card applications from births that occur after 1879.
- Major League Baseball Data — data includes pitch-by-pitch data for Major League Baseball (MLB) games in 2016.

6. UCI Machine Learning Repository

University of California Irvine hosts 440 data sets as a service to the machine learning community. These data sets are nice because most of them are squeaky clean and ready for modeling! Here are some examples:

- Iris Data Set — the most famous pattern recognition dataset.
- Wine Data Set — using chemical analysis to determine the origin of wine.

- Forest Fires — try to predict the burn area of forest fires using this dataset.

7. Academic Torrents

Academic Torrents is a site that is geared around sharing the data sets from scientific papers. It has tons of interesting data sets. You can browse the data sets directly on the site and download them. Here are some examples:

- Enron Emails — a set of many emails from executives at Enron, a company that famously went bankrupt.
- Student Learning Factors — a set of factors that measure and influence student learning.
- News Articles — contains news article attributes and a target variable.

8. Quandl

Quandl is a repository of economic and financial data. Some of the datasets are free, while others are up for purchase. Here are some examples:

- Entrepreneurial Activity by Race and Other Factors — contains data from the Kauffman foundation on entrepreneurs in the US.
- Chinese Macroeconomic Data — indicators of Chinese economic health.
- US Federal Reserve Data — US economic indicators from the Federal Reserve.

9. Jeremy Singer-Vine

Jeremy Singer-Vine collects awesome data sets across multiple sources. If you're interested in getting data sets straight to your inbox, you should consider signing up for his newsletter.

10. Google Dataset Search

Type of data: Miscellaneous
Data compiled by: Google
Access: Free to search, but does include some fee-based search results
Sample dataset: **Global price of coffee, 1990-present**

It seems we turn to Google for everything these days, and data is no exception. Launched in 2018, Google Dataset Search is like Google's standard search engine, but strictly for data. While it's not the best tool if you prefer to browse, if you have a particular topic or keyword in mind, it won't disappoint. Google Dataset Search aggregates data from external sources, providing a clear summary of what's available, a description of the data, who it's provided by, and when it was last updated. It's an excellent place to start.

Prefer to watch this information over reading it? **Check out this video on dataset resources**, presented by our very own in-house data scientist, Tom!

11. Kaggle

Type of data: Miscellaneous
Data compiled by: Kaggle
Access: Free, but registration required
Sample dataset: **Daily temperature of major cities**

Like Google Dataset Search, Kaggle offers aggregated datasets, but it's a community hub rather than a search engine. Kaggle launched in 2010 with a number of machine learning competitions, which subsequently solved problems for the likes of NASA and Ford. It has since evolved into a renowned open data platform, offering cloud-based collaboration for data scientists, as well as educational tools for teaching artificial intelligence and **data analysis**

techniques…plus, of course, tonnes of great datasets covering almost any topic you can imagine.

12. Data.Gov

Type of data: Government
Data compiled by: US Federal Government
Access: Free, no registration required
Sample dataset: **Lobster Report for Transshipment and Sales**

In 2015, the US Government made all its data publicly available. With over 200,000 datasets covering everything from climate change to crime, you can lose yourself in the database for hours. For a government website, it has some surprisingly user-friendly search functions, including the ability to drill down by geographical area, organization type, and file format. Search results are also clearly labeled at federal, state, county, and city levels. If you're interested in more general data about the US population, you can also check out the **US Census Bureau**, offering a rich selection of data about US citizens, their geography, education, and population growth.

13. Datahub.io

Type of data: Mostly business and finance
Data compiled by: Datahub
Access: Mostly free, no registration required
Sample dataset: **Average mass of glaciers since 1945**

The goal of many data analysts is to help drive savvy business decisions. As such, using economic or business datasets for your portfolio project might be worth considering. While Datahub covers a variety of topics from climate change to

entertainment, it mainly focuses on areas like stock market data, property prices, inflation, and logistics. Because many of the data on the portal are updated monthly (or even daily) you'll always have something fresh to work with, as well as data that covers broad timescales.

14. UCI Machine Learning Repository

Type of data: Machine learning
Data compiled by: University of California Irvine
Access: Free, no registration required
Sample dataset: **Behavior of urban traffic in Sao Paulo, Brazil**

Generalized repositories are great if you're happy to browse. But if you're seeking something more niche, why not specialize? Enter the UCI Machine Learning Repository. Launched thirty years ago by the University of California Irvine, don't let the 90s vibe mislead you—the UCI repository has a strong reputation among students, teachers, and researchers as the go-to place for machine learning data. Datasets are clearly categorized by task (i.e. classification, regression, or clustering), attribute (i.e. categorical, numerical), data type, and area of expertise. This makes it easy to find something that's suitable, whatever machine learning project you're working on.

15. Earth Data

Type of data: Earth science
Data compiled by: NASA
Access: Free, no registration required
Sample dataset: **Environmental conditions during fall moose hunting season in Alaska, 2000-2016**

If you think space is awesome (let's face it, space is awesome!) look no further than Earth Data. Publicly available since 1994, this repository provides access to all of NASA's satellite observation data for our little blue planet. As you can imagine, there's plenty to peruse, from weather and climate measurements to atmospheric observations, ocean temperatures, vegetation mapping, and more. If Earth-based data isn't your thing, NASA's **Planetary Data System** takes things a step further with data from interplanetary missions, such as the Cassini probe (which orbited Saturn from 2004 to 2017). Who knows, you might even make a scientific discovery…

16. CERN Open Data Portal

Type of data: Particle Physics
Data compiled by: CERN
Access: Free, no registration required
Sample dataset: **Higgs candidate collision events from 2011 and 2012**

Want to demonstrate your ability to work with highly complex datasets? Head to the CERN Open Data Portal. It offers access to over two petabytes of information, including datasets from the Large Hadron Collider particle accelerator. Frankly, these data aren't for the faint of heart but if you're interested in particle physics, they're worth checking out. While even the names of these datasets are pretty complex, each entry has a helpful breakdown of what's included, as well as related datasets, and how to go about analyzing them. In many cases, they even provide sample code to get you started (thanks, CERN!)

17. Global Health Observatory Data Repository

Type of data: Health
Data compiled by: UN World Health Organization
Access: Free, no registration required
Sample dataset: **Polio immunization coverage estimates by region**

The Global Health Observation data repository is the UN WHO's gateway to health-related statistics from across the globe. If you're looking to break into the healthcare industry (a key focus for many data scientists, especially in the area of machine learning), these datasets are a good option for your portfolio. Covering everything from malaria to HIV/AIDS, antimicrobial resistance, and vaccination rates, the portal even has a nice little feature that lets you preview data tables before downloading them. Not strictly necessary, but definitely nice to have!

18. BFI film industry statistics

Type of data: Entertainment and film
Data compiled by: British Film Institute
Access: Free, no registration required
Sample dataset: **Weekend box office figures from 2001-present**

If you're looking for some data that are a bit more digestible, the next few should be right up your street. First off: the British Film Institute industry statistics. Throughout the year, the BFI accrues and releases data on everything from UK box office figures, to audience demographics, home entertainment, movie production costs, and more. The best part though is their annual statistical yearbook. This breaks down the year's data with some excellent statistical analysis and visual reports—great if you're new to data analytics and want to check your work against the real thing.

19. NYC Taxi Trip Data

Type of data: Transport
Data compiled by: New York City Taxi and Limousine Commission
Access: Free, no registration required
Sample dataset: Take your pick!

This is a weirdly fascinating one…since 2009, the NYC Taxi and Limousine Commission has been accruing transport data from across New York City. Find datasets covering pick-up/drop-off times and locations, trip distances, fares, rate and payment types, passenger counts, and more. It's pretty interesting to compare the differences in figures from 2009 to the present day, especially within such a small geographic area. The site also provides some additional tools, including user guides, taxicab zone maps, data dictionaries (for explaining the spreadsheet labels), and annual industry reports. All very intuitive and quite a helpful guide if you're new to data analytics.

20. FBI Crime Data Explorer

Type of data: Crime and drugs
Data compiled by: Federal Bureau of Investigation
Access: Free, no registration required
Sample dataset: **Homicide offense counts in Point Pleasant, 2008-2018**

If you're fascinated by crime, the FBI Crime Data Explorer is the one for you. It provides a broad collection of crime statistics from a variety of state organizations (universities and local law enforcement) and government (on a local, regional, and state-level). Pull data on hate crimes, officer assaults, homicides, and more. Like the last couple of entries on our list, it also includes some **helpful user guides** to

support data navigation. Each dataset also has some pretty nice visual breakdowns and analysis, so you can see if it has the features you’re looking for before downloading it.