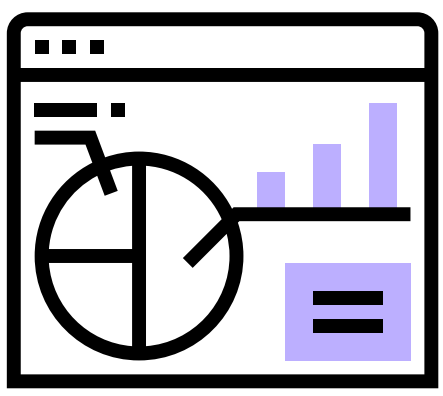
MACHINE LEARNING INDIA

WHERE TO FIND DATASETS

20+ fantastic repositories!



First of all, what is a dataset?

A dataset, or data set, is simply **a collection of data.**

The simplest and most common format for datasets you'll find online is a **spreadsheet** or **CSV format** — a single file organized as a table of rows and columns.

Datasets can consist of images, videos, audio files, numercial data or textual data, and are stored in different formats. Also, they don't have to be just one file. Sometimes a dataset can be a zip file or folder containing multiple data tables with related data.

How are datasets created?

Different datasets are created in different ways. Some of them are **machine-generated** data. Some comprise of data that's been collected via **surveys**. Some may be data that's recorded from **human observations**. Some may be data that's been **scraped** from websites or pulled via **APIs**.

Whenever you're working with a dataset, it's important to consider: **how was this dataset created? Where does the data come from?** Don't jump right into the analysis; take the time to first understand the data you are working with.

Where to find datasets?

1. FiveThirtyEight: FiveThirtyEight is an interactive news and sports site that has some incredible data visualizations. They makes a lot of their data open to the public, meaning you can download and play with the source data yourself!
Link: https://data.fivethirtyeight.com/

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!
Link: github.com/BuzzFeedNews

3. Kaggle: Kaggle is a place where you can learn, practice, and fine-tune your data science, analytics skills. They have tons of open, public data, and allow users of the platform to share code so you can learn best practices within the data space.
Link: https://www.kaggle.com/datasets

4. Socrata: Socrata hosts cleaned open source data sources ranging from government, business, and education data sets.
Link: https://opendata.socrata.com/

5. Awesome Public Datasets: This Github hosts a library of awesome, public datasets! They are all sorted by category and link you straight to the hosting website.
Link: github.com/awesomedata/

6. Google Public Datasets: Google lists all of the data sets on a page. On Google Cloud Platform (GCP), you can query using BigQuery to explore these datasets. You'll need to sign up for a **GCP account** an only the first **1TB** of queries you make are **free**!
Link: cloud.google.com/bigquery/public-data/

7. UCI Machine Learning Repository: University of California Irvine hosts 440 data set as a service to the ML community. These data sets are nice, squeaky clean, and are ready for modeling!
Link: http://archive.ics.uci.edu/ml/index.php

8. Quandl: Quandl is a repository of economic and financial data. Some of the datasets are free, while others are up for purchase.
Link: https://www.quandl.com/search

9. Data.gov: Data.gov allows you to download and explore data from **US government** agencies. Data can range from government budgets to climate data. The data is very **well documented** so you should have an easy time to navigate the sources.
Link: https://www.data.gov/

10. Academic Torrents: Academic Torrents is a site that is geared around sharing the data sets from scientific papers. You can browse the data sets directly on the site, and download the ones you like!
Link: http://academictorrents.com/browse.php

11. AWS Public Data Sets: Amazon has a page that lists all of the data sets for you to browse. You'll need an **AWS** account, and Amazon also gives a free access tier for new accounts.
Link: https://aws.amazon.com/datasets

Some other repositories:

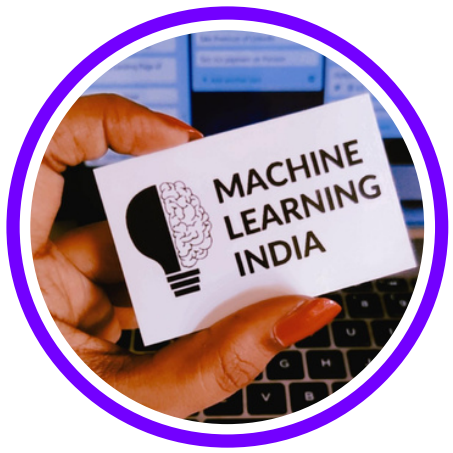
- Jeremy Singer-Vine
- Wikipedia Data Sets
- Data.world Data Sets
- World Bank Data Sets
- Reddit - /r/datasets
- NASA Datasets
- Twitter Dataset via Twitter API
- Github Dataset via Github API
- CERN Open Data Portal
- Global Health Observatory Data Repository

Notable references:

- 11 websites to find free, interesting datasets: **interviewqs.com**.
- 21 Places to Find Free Datasets for Data Science Projects: **www.dataquest.io**.

Important note:

The links to these resources will be put up on our Telegram. Channel ID: **@machinelearning24x7**.



@ml.india

@ml_india_

bit.ly/mli-linkedin

@machinelearning24x7

Find our content valuable?

Show your support by following us, sharing our content and commenting below! A **HUGE** shout-out to **Shahshwat Kothari**, our patron, for making this post possible. ♥

Become a **patron** on: patreon.com/machinelearningindia

Link in bio!

Like. **Comment.** **Share.**