Datasets

Explore, analyze, and share quality data

Types of Datasets

Kaggle supports a variety of dataset publication formats, but we strongly encourage dataset publishers to share their data in an accessible, non-proprietary format if possible. Not only are open, accessible data formats better supported on the platform, they are also easier to work with for more people regardless of their tools.

This page describes the file formats that we recommend using when sharing data on Kaggle Datasets. Plus, learn why and how to make less well-supported file types as accessible as possible to the data science community.

Supported File Types

CSVs

The simplest and best-supported file type available on Kaggle is the “Comma-Separated List”, or CSV, for tabular data. CSVs uploaded to Kaggle should have a header row consisting of human-readable field names. A CSV representation of a shopping list with a header row, for example, looks like this:

id,type,quantity

0,bananas,12

1,apples,7

CSVs are the most common of the file formats available on Kaggle and are the best choice for tabular data.

On the Data tab of a dataset, a preview of the file's contents is visible in the data explorer. This makes it significantly easier to understand the contents of a dataset, as it eliminates the need to open the data in a Notebook or download it locally.

CSV files will also have associated column descriptions and column metadata. The column descriptions allows you to assign descriptions to individual columns of the dataset, making it easier for users to understand what each column means. Column metrics, meanwhile, present high-level metrics about individual columns in a graphic format.

"The Complete Pokemon Dataset" is an example of a great CSV-type Dataset.

JSON

While CSV is the most common file format for "flat" data, JSON is the most common file format for "tree-like" data that potentially has multiple layers, like the branches on a tree:

{[{'id': 0, 'type': 'bananas', 'quantity': 12}, {'id': 1, 'type': 'apples', 'quantity': 7}]}

For JSON files, the Data tab preview will present an interactive tree with the nodes in the JSON file attached. You can click on individual keys to open and collapse sections of the tree, exploring the structure of the dataset as you go along. JSON files do not support column descriptions or metrics.

You can filter the Datasets listing by File Type to show all datasets containing JSON files.

SQLite

Kaggle supports database files using the lightweight SQLite format. SQLite databases consist of multiple tables, each of which contains data in tabular format. These tables support large datasets better than CSV files do, but are otherwise similar in practice.

The Data tab represents each table in a database separately. Like CSV files, SQLite tables will be fully populated by “Column Metadata” and “Column Metrics” sections.

“European Soccer Database” is an example of a great SQLite-type Dataset.

Archives

Although not technically a file format per se, Kaggle also has first-class support for files compressed using the ZIP file format as well as other common archive formats like 7z.

Compressed files take up less space on disk than uncompressed ones, making them significantly faster to upload to Kaggle and allowing you to upload datasets that would otherwise exceed the Dataset size limitations.

Archives are uncompressed on our side so that their contents are accessible in Notebooks without requiring users to unzip them. Archives do not currently populate previews for individual file contents, but you can still browse the contents by file name.

As a result, we recommend that you only upload your dataset as an archive if the dataset is large enough, is made up of many smaller files, or is organized into subfolders. For instance, ZIPs and other archive formats are a great choice for making image datasets available on Kaggle.

"Chest X-Ray Images (Pneumonia)" is an example of a dataset made of archived images.

BigQuery

Kaggle also supports special BigQuery Datasets. BigQuery is a "big data" SQL store invented by Google. Many massive public datasets, like all the code in GitHub and the complete history of the Bitcoin blockchain, are available publically through the Google BigQuery Public Datasets initiative. Some of these are in turn also available as Kaggle Datasets!

BigQuery Datasets are special in many ways. Because they are multi-terabyte datasets hosted on Google's servers they cannot be uploaded or downloaded. Within Notebooks, instead of loading the files from disk, you interact with the dataset by writing SQL fetch queries within either the Google BigQuery Python library or Kaggle's bq_helper library. And, due to

the large size of the datasets involved, there is a quota of 5 TB of data scanned per user per 30-days.

Some resources for understanding how to use BigQuery:

- Getting Started with Big Query
- Beyond Queries: Exploring the Bigquery API

"USA Names Data" is an example of a BigQuery-type Dataset. Here are some helpful Notebooks for learning more about BigQuery: "SQL Scavenger Hunt Handbook, Getting Started with BigQuery", and "Beyond Queries: Exploring the BigQuery API".

Other File Formats

The file formats listed in the section above are the ones best supported and most common on the Kaggle format. This doesn't mean that other types of files can't be uploaded; any file you can think of can be uploaded. Other formats are just less well-supported: they may not have previews or any of the other data explorer components available. They will also likely be less familiar with Kaggle users, and hence, less accessible.

If you can convert your file into one of the formats above (the simpler the better), we highly recommend doing so. For example, Excel spreadsheets are a proprietary format that should be uploaded as CSV files instead. Your users will thank you!

However, there are nevertheless use cases for alternative data formats. We do encourage uploads in speciality data formats like NPZ, image file formats like PNG, and complex hierarchical data formats like HDF5. But, when doing so, we suggest also uploading a Notebook discussing what

and where the files are, how to work with them, and demonstrating how to get started with the dataset. Reproducible code samples can go a long way towards making your data files accessible to the data science world!

Searching for Datasets

Datasets is not just a simple data repository. Each dataset is a community where you can discuss data, discover public code and techniques, and create your own projects in Notebooks. You can find many different interesting datasets of all shapes and sizes if you take the time to look around and find them!

The latest and greatest from Datasets is surfaced on Kaggle in several different places.

Newsfeed

When you're logged into your Kaggle account, the Kaggle homepage provides a live newsfeed of what people are doing on the platform. New Datasets uploaded by people you follow and hot Datasets with lots of activity will show up here. By browsing down the page you can check out all the latest updates from your fellow Kagglers.

You can tweak your news feed to your liking by following other Kagglers. To follow someone, go to their profile page and click on "Follow User". Content posted and upvotes made by users you have followed will show up more prominently.

The same is true of other users who choose to follow you. Post high-quality content and you will soon find other users following along with what you are doing!

Datasets Listing

A more structured way of accessing datasets is accessible from the “Datasets” tab in the main menu bar.

Datasets are grouped by different categories: "Trending Datasets", "Popular Datasets", "Recently Viewed Datasets" and a few other rotating categories.

At the bottom of this page, you can click on the "Explore all public datasets" button to get a list view of all datasets. The list is sorted by “Hotness” by default. “Hotness” is what it sounds like: a way of measuring the interestingness and recency of datasets on the platform. Datasets which score highly in Hotness, and thus appear highly in this list, are usually either recently released Datasets that have been marked Reviewed and are scoring highly in engagement, or “all-time” greats that have been consistently popular on the platform for a long time.

Other methods of sorting are by Most Votes, New, Updated and Usability.

Other filtering options, available from the navigation bar, are Sizes (Small, Medium, or Large), File types (CSV, SQLite, JSON, BigQuery), Licenses (Creative Commons, GPL, Other Database, Other), and Tags (described in the next section).

You can also use the listing to view your own Datasets (“Your Datasets”), or to look at datasets you have previously bookmarked ("Bookmarks").

Finally, a Datasets-specific search bar is available here. This is often the fastest way to find a specific dataset that you are looking for.

Tags and Tag Pages

Tags are the most advanced of the searching options available in the Datasets listing page. Tags are added by dataset owners to indicate the topic of the Dataset, techniques you can use (e.g., “classification”), or the type of the data itself (e.g., “text data”). You can navigate to tag pages to browse more content sharing a tag either by clicking on a tag on a Dataset, or by clicking on the “Tags” dropdown in the site header.

Searching by tags allow you to search for Datasets by topical area. For example, if you are interested in animal shelter data you might try a search with the tag “animals”; if you are interested in police records a search with “crime” would do the trick.

Tag pages include a section listing the most popular pages with the given tag, making them a great way of searching for datasets by content.

Creating a Dataset

It’s easy to create a dataset on Kaggle and doing so is a great way to start a data science portfolio, share reproducible research, or work with

collaborators on a project for work or school. You have the option to create private datasets to work solo or with invited collaborators or publish a dataset publicly to Kaggle for anyone to view, download, and analyze.

Navigating the Dataset Interface

To publish a private or public dataset, start by navigating to the Datasets listing. There you will find a New Dataset button. Click on it to open the New Dataset modal.

The required “bare minimum” fields for uploading a dataset to Kaggle in descending order are:

- The **Title** is the name of the Dataset – e.g. what will appear in the listing when searching or browsing.
- The **URL** is the link the Dataset will live at. The slug will first auto-populate and mimic your Title. However, you can hover over the slug to change it right away.
- Finally, you may upload data from one of four sources:

- **Your local machine** - upload files/folders via drag and drop or by selecting them in your file browser. To speed up file/folder uploads, try uploading them as a ZIP archive; the contents will be unzipped on our side to make them accessible in Notebooks.
- **Remote Files** - enter list of public URL(s) which identify files to be imported into dataset
- **Github Repository** - enter URL to github repository whose files will be imported into dataset
- **Notebook Outputs** - use inbuilt search to explore publicly available files produced from Kaggle’s large repository of public Notebooks

To make your dataset more useful for your collaborators and the community it is recommended you update the following settings:

- The Sharing menu controls the Dataset's visibility. Datasets may be Private (visible only to you and your collaborators, and to Kaggle for purposes consistent with the Kaggle Privacy Policy) or Public (visible to everyone). The default setting is Private.
- The Licence is the license the dataset is released under (relevant for public datasets). If the license you need doesn't appear in the dropdown, select the "Other (specified in description)" option and be sure to provide information on the license when writing the dataset description (in the next step). Below is a list of common licenses.

Common Licenses

- **Creative Commons**
  - CC0: Public Domain
  - CC BY-NC-SA 4.0
  - CC BY-SA 4.0
  - CC BY-SA 3.0
  - CC BY 4.0 (Attribution 4.0 International)
  - CC BY-NC 4.0 (Attribution-NonCommercial 4.0 International)
  - CC BY 3.0 (Attribution 3.0 Unported)
  - CC BY 3.0 IGO (Attribution 3.0 IGO)
  - CC BY-NC-SA 3.0 IGO (Attribution-NonCommercial-ShareAlike 3.0 IGO)
  - CC BY-ND 4.0 (Attribution-NoDerivatives 4.0 International)
  - CC BY-NC-ND 4.0 (Attribution-NonCommercial-NoDerivatives 4.0 International)
- **GPL**
  - GPL 2
  - LGPL 3.0 (GNU Lesser General Public License 3.0)
  - AGPL 3.0 (GNU Affero General Public License 3.0)

- FDL 1.3 (GNU Free Documentation License 1.3)
- **Open Data Commons**
- Database: Open Database, Contents: Database Contents
- Database: Open Database, Contents: © Original Authors
- PDDL (ODC Public Domain Dedication and Licence)
- ODC-BY 1.0 (ODC Attribution License)
- **Community Data License**
- Community Data License Agreement - Permissive - Version 1.0
- Community Data License Agreement - Sharing - Version 1.0
- **Special**
- World Bank Dataset Terms of Use
- Reddit API Terms
- U.S. Government Works
- EU ODP Legal Notice
    - Owner allows you to specify the dataset Owner if you belong to any Organizations. You may assign ownership to yourself or to any Organizations you are a member of (see the section “Creating and using organizations” to learn more about this feature).

      Once you have provided the required information alongside your data source, click on “Create Dataset” and your dataset will start processing. Once the dataset is finished processing, you will be taken to your new dataset’s home page.

      Note that if your dataset is very large (multiple gigabytes in size), processing may take a while, up to several minutes. Feel free to navigate away from the browser window whilst processing is inflight as it will continue in the background.

Your datasets has now been created! However, for truly great Datasets, the work doesn't stop there. Once you have specified the required fields there are a few other things you should do in order to maximize your dataset's usefulness to the community or your collaborators:

- Upload a cover image. We recommend using unsplash.com for shareable, high resolution images.
- Add a subtitle to the dataset. This is a short bit of text explaining in slightly more detail what is in it. This subtitle will appear alongside the title in the search listings.
- Add tags. Tags help users find datasets on topics they are interested in by making them easier to find.
- Add a description. The description should explain what the dataset is about in long-form text. A great description is extremely useful to Kaggle community members looking to get started with your data.
- Publish a public Notebook. Use Notebooks to show community members or your collaborators how to get started with the data. This can be something simple like an exploratory data analysis or a more complex project reproducing research using the data.

A few examples of well-formatted datasets are "CS:GO Competitive Matchmaking Data", "Yelp Dataset", "1.6 million UK traffic accidents", and "Fashion MNIST".

Creating Datasets from Various Connectors

As outlined above, in addition to uploading files from your local machine, you can also create Datasets from various data sources including GitHub, remote URLs (any public file hosted on the web), and Notebook output

files. These are each icons that can be found in the Dataset Upload Modal sidebar.

GitHub and Remote File Datasets

Datasets created from a GitHub repository or hosted (remote) files are downloaded directly from the remote server to Kaggle's cloud storage and, therefore, will consume none of your local network's bandwidth. This makes the remote files connector a convenient solution for creating datasets from large files.

When a dataset is created from a github repository or hosted file, the publisher is able to set up automatic interval updates from the dataset's Settings tab. Here's an example stock market dataset that updates daily.

Don't want to wait for a refresh? No problem! Click the Update button within the "..." dropdown in the dataset menu header to sync your dataset immediately.

Notebook Output File Datasets

Creating a dataset from a Notebook's output files will let you create reproducible data pipelines. To create a dataset from a Notebook's output files, click on the icon in the uploader and search for your Notebook. Alternatively, you can click "Create Dataset" from the Output tab on your rendered Notebook. Then, select the files you want to use in your dataset.

Limitations

It's worth noting that for user experience and technical simplicity, a dataset can be created and versioned from exclusively one data source.

That is, data sources currently can not be mixed and matched in any given dataset (for example, a dataset created from a GitHub repository can't also include files uploaded from your local machine). If you would like to use various different data sources in a Notebook you can create multiple datasets and add them both to said Notebook.

The usual technical specifications for dataset creation apply to connectors too. See the Technical Specifications section for more information.

Updating Dataset Using JSON Config

For advanced users, you may find it easier to update key parameters of your dataset by specifying the details as JSON configuration. To do this, navigate to your dataset and click Settings, followed by “JSON Config” in the menu of options on the left.

You can update any of the settings you would normally edit through the datasets user interface, such as title, collaborators, licenses, keywords and more. For a reference to the schema you can use for updating dataset settings, you can look at our documentation for the relevant actions within the Public API.

Please note, there are some subtle differences between the Public API schema and the schema supported in the JSON Config settings UI. They are as follows:

- **id** is omitted as it cannot be changed after dataset creation
- **resources** is omitted as you cannot change the uploaded files using this UI

- The **isPrivate** is an added boolean option that allows users to change the privacy of their datasets (note: public datasets can NOT be made private)
- **collaborators** is an added array of objects with shape `{ "username": string; "role": "read" | "write"}` that can be used to specify dataset collaborators

Collaborating on Datasets

Dataset collaboration is a powerful feature. It allows multiple users to co-own and co-maintain a private or publicly shared dataset. For example, you can invite collaborators to view and edit a private dataset to work together on preparing it before changing its visibility to public.

When uploading a Dataset you may choose either yourself or any Organization you are a part of as the Owner of that Dataset. If you select yourself, that Dataset will be created with yourself as the Owner. If you select an Organization, that Organization will be the Owner of the dataset, and every other user in the Organization (including yourself) will be added as a Collaborator with editing privileges (if you are unfamiliar with Organizations, you may also want to read the section "Creating and using organizations").

This means that Organizations are an easy way to manage access to datasets or groups of datasets.

Inviting Collaborators

Alternatively, you may manage Collaborators directly. To do so, go to any dataset you own and navigate to Settings > Sharing. There, use the search box to find and add other users as Dataset collaborators.

If your Dataset is private, you may choose between giving Collaborators either viewing privileges (“Can view”) or editing privileges (“Can edit”). If your Dataset is public, Collaborators can only be added with editing privileges (“Can edit”), as anyone can view it already.

When you add a collaborator, they will receive a notification via email.

“Data Science for Good: Kiva Crowdfunding” is a great example of a Collaborative Dataset.

Using Notebooks with Dataset Collaborators

Using Notebooks, Kaggle’s interactive code editing and execution environment, is a powerful way to work with your collaborators on a Dataset. You might want to work with collaborators to write public Notebooks that help familiarize other users with your dataset. Or you may want to keep all of your code private among your collaborators as you work on privately shared projects together.

Notebooks you create are private by default, and their sharing settings are distinct from the sharing settings on your Dataset. That is, your Dataset collaborators won’t automatically see your private Notebooks. Here’s

what that means and how you can productively use sharing settings on Datasets and Notebooks together:

- You can make public Notebooks on a private Dataset which will allow anyone to view your Notebook, but not the underlying private data source.
- If you want to add view or edit collaborators to a private Notebook (whether the dataset is private or public), you can do so by adding users via Options > Sharing on the Notebook.

Resources for Starting a Data Project

There are many resources available online to help you get started working on your open data project.

Using Datasets

- Getting Started on Kaggle video tutorials: Just started on Kaggle? Not sure what is where and why? Here are our very own Kaggle team tutorials to orient you quickly on navigating the Kaggle platform and creating your own datasets and Notebooks
- A Guide to Open Data Publishing: This article includes the key ingredients to an open data project.
- Web scraping data in Python: A tutorial showing you how to scrape data with BeautifulSoup. It goes over the same code used to create the Craft Beers dataset published on Kaggle.

- Making Kaggle the Home of Open Data: Ben's post shares instructions for publishing your open data project on Kaggle and how you can explore others' datasets.
- Creating an Organization: If you're publishing data from an organization, you can create an organization profile first. Then you just select the organization profile from the dropdown near your avatar when publishing (https://www.kaggle.com/datasets/new).
- Open Data Spotlights: This series highlights some of the best open data projects on Kaggle.
- Have requests or want to discuss data collection, cleaning, or other aspects of open data projects? Post away in the Datasets Discussion forum on Kaggle.

Using Notebooks

- Getting Started on Kaggle video tutorials: Just started on Kaggle? Not sure what is where and why? Here are our very own Kaggle team tutorials to orient you quickly on navigating the Kaggle platform and creating your own datasets and Notebooks
- Kaggle Learn is a great place to start getting hands on with data science and machine learning techniques using Notebooks.
- Does open data make you happy? An introduction to Kaggle Notebooks: Learn how to use Notebooks to explore any combination of datasets published on Kaggle.
- Seventeen Ways to Map Data in Notebooks: A collection of mini-tutorials by Kaggle users for Python and R users.

Analysis

- How to Get Started with Data Science in Containers: One of our data scientists, Jamie Hall, explains how and why Docker containers are at the heart of Notebooks – reproducible analysis.
- Approaching (Almost) Any Machine Learning Problem by Kaggle Grandmaster Abhishek Thakur: Exactly what it says – a great tutorial.

Other

- Kaggle Datasets Twitter: The new account features newly featured datasets plus open data news.
- Collecting & Using Open Data: A blog by Kaggler MLWave recommended by Triskelion.

Technical Specifications

Kaggle Datasets allows you to publish and share datasets privately or publicly. We provide resources for storing and processing datasets, but there are certain technical specifications:

- 200GB per dataset limit
- 200GB max private datasets (if you exceed this, either make your datasets public or delete unused datasets)
- A max of 50 top-level files (if you have more, use a directory structure and upload an archive)

When you upload a dataset we apply certain processing steps to make the dataset more usable.

- A complete archive is created so the dataset can be easily downloaded later
- Any archives (e.g., ZIP files) that you upload are uncompressed so that the files are easily accessible in Notebooks (directory structure is preserved)
- Data types for tabular data files are automatically detected (e.g., geospatial types)
- Column-level metrics are calculated for tabular data which are viewable on the data explorer on the dataset's "Data" tab

When publishing datasets, you might also want to consider the technical specifications of Notebooks if you intend to use (or encourage other Kaggle users to use) Notebooks to analyze the data.