

# How to Load Kaggle Datasets Directly into Google Colab?

[BEGINNER](#)[DATA SCIENCE](#)[DATABASE](#)[PROJECT](#)[RESOURCE](#)

This article was published as a part of the [Data Science Blogathon](#)

## Introduction

Almost every data science aspirant uses Kaggle. It houses datasets for every domain. You can get a dataset for every possible use case ranging from the entertainment industry, medical, e-commerce, and even astronomy. Its users practice on various datasets to test out their skills in the field of Data Science and Machine learning.

The Kaggle datasets can have varying sizes. Some datasets can be as small as under 1MB and as large as 100 GB. Also, some of the Deep learning practices require GPU support that can boost the training time. Google Colab is a promising platform that can help beginners to test out their code in the cloud environment.

In this article, I will explain how to load the datasets directly from Kaggle to Google Colab notebooks.



Image by Author (Made in [Canva](#))

## Step 1: Select any dataset from Kaggle

The first and foremost step is to choose your dataset from Kaggle. You can select datasets from competitions too. For this article, I am choosing two datasets: One random dataset and one from the active competition.

Research Prediction Competition

Google Smartphone Decimeter Challenge

Improve high precision GNSS positioning and navigation accuracy on smartphones

Google

Google · 461 teams · 2 months to go (a month to go until merger deadline)

\$10,000

Prize Money

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

My Submissions

Submit Prediction

Data Description

This challenge provides data from a variety of instruments useful for determining a phone's position: signals from GPS satellites, accelerometer readings, gyroscope readings, and more.

As this challenge's design is focused on post-processing applications such as lane-level mapping, future data along a route will be

Screenshot from [Google Smartphone Decimeter Challenge](#)

Dataset

The Complete Pokemon Images Data Set

A collection of 898 images of all the Pokemons taken from the Pokedex database.

Rohan Asokan

• updated 11 days ago (Version 1)

Data

Tasks

Code (1)

Discussion

Activity

Metadata

Download (11)

Usability 9.4

License CC0: Public Domain

Tags arts and entertainment

Description

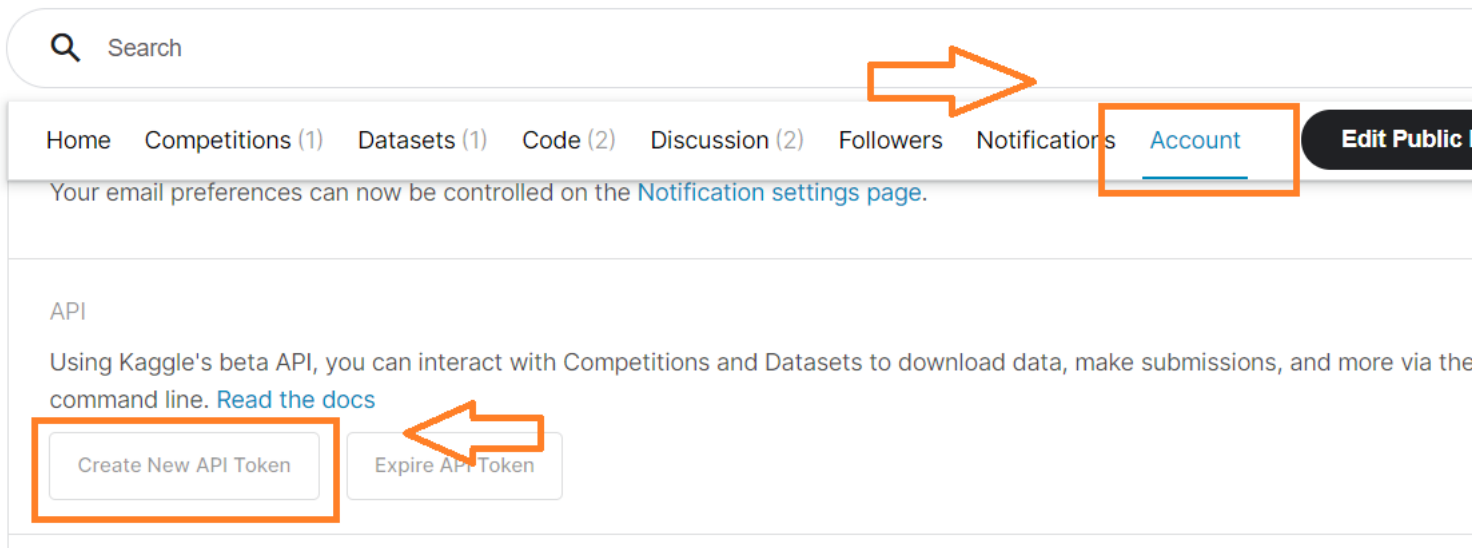
Context

As necessity is the mother of all invention, this data set was a necessity for a personal project on matching Pokemon images. No other data was updated and hence this data set was born.

Screenshot from [The Complete Pokemon Images Data Set](#)

## Step 2: Download API Credentials

To download data from Kaggle, you need to authenticate with the Kaggle services. For this purpose, you need an API token. This token can be easily generated from the profile section of your Kaggle account. Simply, navigate to your Kaggle profile and then,



Click the Account tab and then scroll down to the API section (Screenshot from Kaggle profile)

A file named “kaggle.json” will be download which contains the username and the API key.

This is a one-time step and you don't need to generate the credentials every time you download the dataset.

## Step 3: Setup the Colab Notebook

Fire up a Google Colab notebook and connect it to the cloud instance (basically start the notebook interface). Then, upload the “kaggle.json” file that you just downloaded from Kaggle.

Screenshot from Colab interface

Now you are all set to run the commands need to load the dataset. Follow along with these commands:

Note: Here we will run all the Linux and installation commands starting with "!". As Colab instances are Linux-based, you can run all the Linux commands in the code cells.

## 1. Install the Kaggle library

```
! pip install kaggle
```

## 2. Make a directory named ".kaggle"

```
! mkdir ~/.kaggle
```

## 3. Copy the "kaggle.json" into this new directory

```
! cp kaggle.json ~/.kaggle/
```

## 4. Allocate the required permission for this file.

```
! chmod 600 ~/.kaggle/kaggle.json
```

The colab notebook is now ready to download datasets from Kaggle.

All the commands needed to set up the colab notebook

## Step 4: Download datasets

Kaggle host two types of datasets: Competitions and Datasets. The procedure to download any type remains the same with just minor changes.

### Downloading Competitions dataset:

```
! kaggle competitions download <name-of-competition>
```

Here, the name of the competition is not the bold title displayed over the background. It is the slug of the competition link followed after the "/c/". Consider our example link:

"<https://www.kaggle.com/c/google-smartphone-decimeter-challenge>"

"google-smartphone-decimeter-challenge" is the name of the competition to be passed in the Kaggle command. This will start downloading the data under the allocated storage in the instance:

The output of the command (Notebook screenshot)

### Downloading Datasets:

These datasets are not part of any competition. You can download these datasets by:

```
! kaggle datasets download <name-of-dataset>
```

Here, the name of the dataset is the “user-name/dataset-name”. You can simply copy the trailing text after “www.kaggle.com/”. Therefore, in our case,

“https://www.kaggle.com/arenagrenade/the-complete-pokemon-images-data-set”

It will be: “arenagrenade/the-complete-pokemon-images-data-set”

The output of the command (Notebook screenshot)

In case you get a dataset with a zip extension, you can simply use the unzip command of Linux to extract the data:

```
! unzip <name-of-file>
```

## Bonus Tips

### Tip 1: Download Specific Files

You just saw how to download datasets from Kaggle in Google Colab. It is possible that you are only concerned about a specific file and want to download only that file. Then you can use the “-f” flag followed

by name of the file. This will download only that specific file. The “-f” flag works for both competitions and datasets command.

Example:

```
! kaggle competitions download google-smartphone-decimeter-challenge -f baseline_locations_train.csv
```

The output of the command (Notebook screenshot)

You can check out Kaggle API [official documentation](#) for more features and commands.

## Tip 2: Load Kaggle Credentials from Google Drive

In step 3, you uploaded the “kaggle.json” when executing the notebook. All the files uploaded in the storage provided while running the notebook are not retained after the termination of the notebook.

It means that you need to upload the JSON file every time the notebook is reloaded or restarted. To avoid this manual work,

1. Simply upload the “kaggle.json” to your Google Drive. For simplicity, upload it in the root folder rather than any folder structure.
2. Next, mount the drive to your notebook:

### Steps to Mount Google Drive

3. The initial command for installing the Kaggle library and creating a directory named “.kaggle” remains the same:

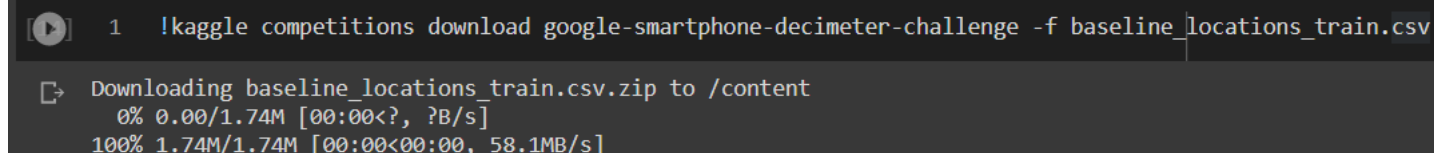
```
! pip install kaggle
```

```
! mkdir ~/.kaggle
```

4. Now, you need to copy the “kaggle.json” file from the mounted google drive to the current instance storage. The Google drive is mounted under the “./content/drive/MyDrive” path. Just run the copy command as used in Linux:

```
!cp /content/drive/MyDrive/kaggle.json ~/.kaggle/kaggle.json
```

Now you can easily use your Kaggle competitions and datasets command to download the datasets. This method has the added advantage of not uploading the credential file on every notebook re-run.

A screenshot of a Google Colab terminal window. The first line shows a command: `!kaggle competitions download google-smartphone-decimeter-challenge -f baseline_locations_train.csv`. The second line shows the progress of downloading the file: `Downloading baseline_locations_train.csv.zip to /content`. The third line shows the progress bar: `0% 0.00/1.74M [00:00<?, ?B/s]`. The fourth line shows the completion: `100% 1.74M/1.74M [00:00<00:00, 58.1MB/s]`.

```
1 !kaggle competitions download google-smartphone-decimeter-challenge -f baseline_locations_train.csv
Downloading baseline_locations_train.csv.zip to /content
0% 0.00/1.74M [00:00<?, ?B/s]
100% 1.74M/1.74M [00:00<00:00, 58.1MB/s]
```

Downloading dataset after configuring API key

## Benefits of Using Google Colab

Google Colab is a great platform to practice data science questions. One of the major benefits of the Colab is the free GPU support. Data science aspirants, in the beginning, are short of computation resources, and therefore using Google Colab solves their hardware problems. The Colab notebooks run on Linux instances and therefore, you can run all the usual Linux commands and interact with the kernel more easily.

The RAM and disk allocation are more than enough for practice datasets but if your research requires more compute power, you can opt for the paid program “Colab pro”.

## About the Author

Hi, I am Kaustubh Gupta, a Python Developer capable of Web Scraping, Automations, Data Science, a bit of Backend Web Development with knowledge of CSS and Bootstrap, Android App Developer in Python. I explore everything that can use Python. I am currently mastering Machine learning algorithms along with real-world applications involving a mixture of all tech stacks.

If you have **any doubts, queries, or potential opportunities**, then you can reach out to me via

1. LinkedIn – [in/kaustubh-gupta/](#)
2. Twitter – [@Kaustubh1828](#)
3. GitHub – [kaustubhgupta](#)
4. Medium – [@kaustubhgupta1828](#)

***The media shown in this article are not owned by Analytics Vidhya and are used at the Author's discretion.***

---

Article Url - <https://www.analyticsvidhya.com/blog/2021/06/how-to-load-kaggle-datasets-directly-into-google-colab/>



**[kaustubh1828](#)**