



**WOMEN IN DATA SCIENCE**  
@ GOOGLE



# WiDS@Google 2023 Datathon

## Workshop Preparation

### Welcome to WiDS@Google 2023 Datathon Workshop!

To make sure you have a good experience in this workshop, we have put together some “how-to” instructions to get you ready for the workshops. We are looking forward to seeing you in the workshop!

<b>WiDS@Google 2023 Datathon</b>	<b>1</b>
<b>Workshop Preparation</b>	<b>1</b>
<b>Where are the workshop materials? How to use github?</b>	<b>2</b>
<b>How to use Google Colab?</b>	<b>3</b>
How to access Python notebooks from our github repository through Google Colab:	3
Where can I find the data for this workshop:	7
How to access csv files (data) on Google Colab notebook:	8
<b>What is Kaggle and how to access the datathon challenge?</b>	<b>8</b>
What is Kaggle?	8
How to join the datathon?	8
<b>Having questions about the tutorials during the speaker session?</b>	<b>11</b>
<b>Having questions about the event?</b>	<b>12</b>

## Where are the workshop materials? How to use github?

We have uploaded the workshop materials to the [github repository](#). If you are not familiar with github, you can take a look at the instructions below.

In short, Github is a code platform for version control and collaboration. Check out this [doc](#) if you want to learn more about github.

Here's the [github repository](#) for the workshop. Within the repository, you will find all the materials to be used for this workshop.

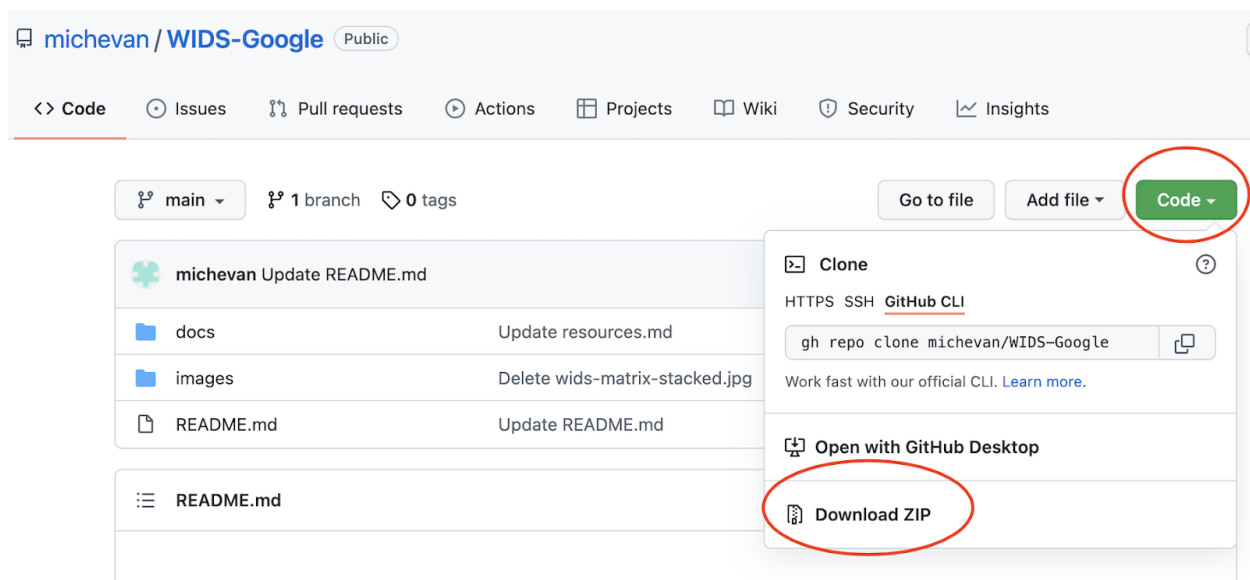
**Under the github repository, you will be able to see two main files:**

*(Note that this is still work in progress. We will make sure all the finalized materials get uploaded Feb 5 EOD US eastern time)*

- **Notebooks:** wids\_datathon\_2023\_code\_demo.ipynb: We will use the python notebook to do a code walkthrough and explain how to tackle this year's datathon problem
- **Slides:**
  - [WiDS datathon] 2023 WiDS@Google Datathon Workshop Event Intro.pdf
  - WiDS@Google2023 Datathon Workshop Instructions.pdf (this document)

If you have a github repository, you can choose to clone our [repository](#). Here are some [instructions](#) on how to clone a repository.

If you do not have a github account, you can just download all the materials to your local PC. Click on **code** and then click on **download ZIP** to download all the materials.

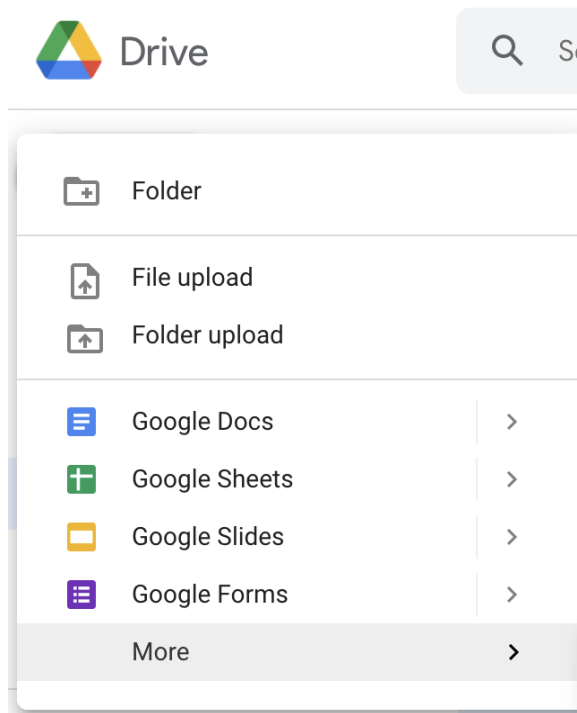


## How to use Google Colab?

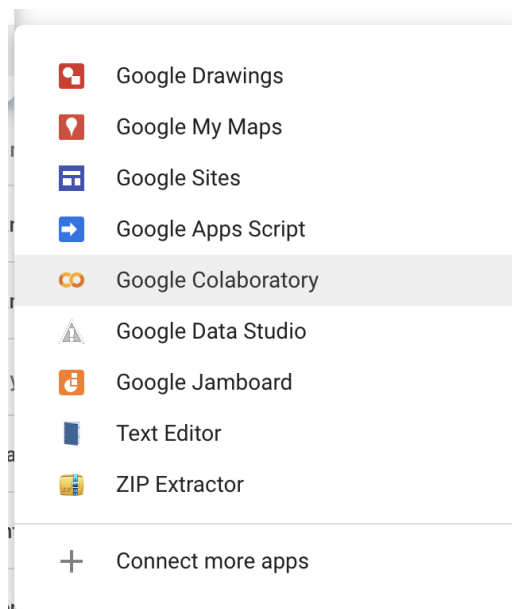
We will use Google Colab for a python code demo. Also, all the python notebooks we are sharing with you are developed in the [Google Colab environment](#). It is very similar to Jupyter lab, for those who are familiar with that, but Colab includes more features for people to share work and collaborate online. In case you are not familiar with Google Colab, you can check out the following instructions.

How to access Python notebooks (.ipynb) from our [github repository](#) through Google Colab:

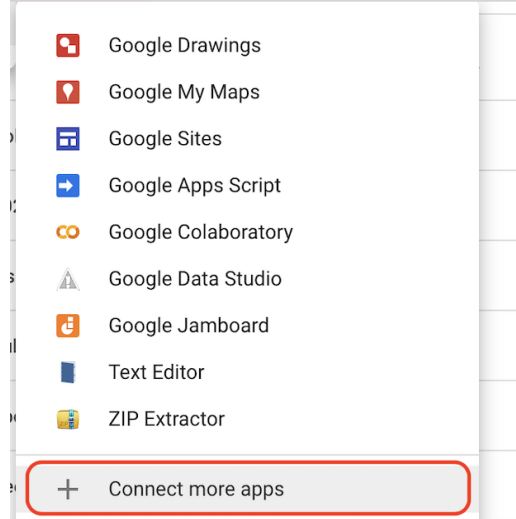
Go to the Google drive and click on **More**.



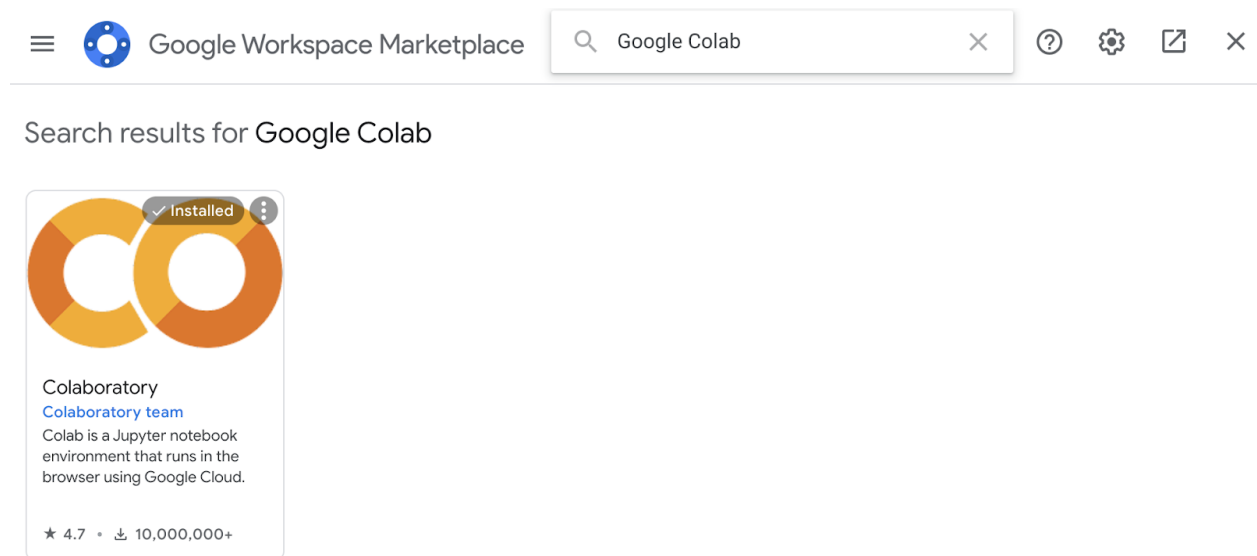
Go to **Google Colaboratory**



**Note:** In case you cannot find **Google Colaboratory**: Go to “**Connect more apps**”



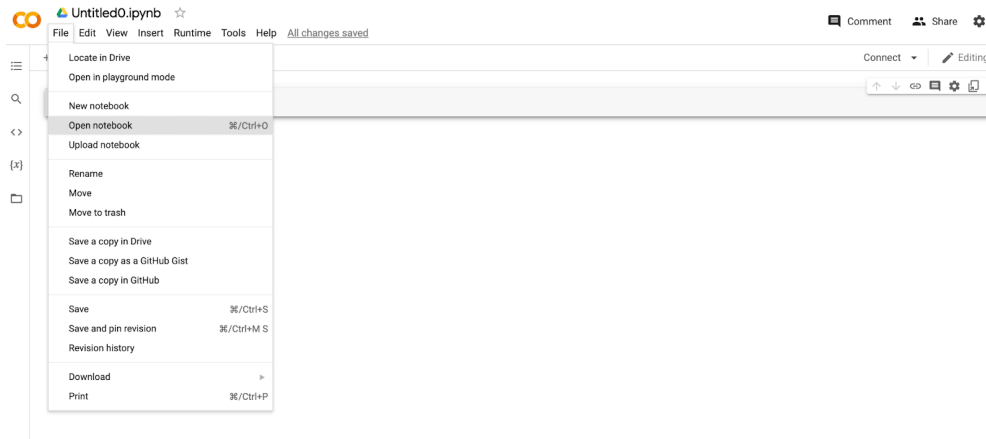
Clicking on “**Connect more apps**” will lead you to the **Google Workspace Marketplace**, by typing “**Google Colab**” on the search bar, you will find the **Colaboratory** as shown below. You can click on it and install Google colab



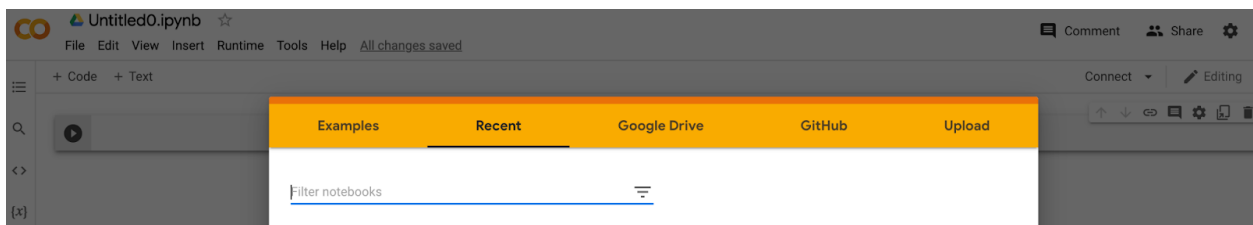
After installing Google Colaboratory, you can follow the steps above to access it.

Clicking on **Google Colaboratory** will lead you to the brand new colab notebook as below.

Click on “**Open notebook**”.



Here you can choose to open the colab notebooks that we have prepared for you in a few different ways:







To open the colab directly from our github account, you can click on the “**GitHub**” tab and copy-paste the URL links of the colab notebooks as below and search for the notebook.




ExamplesRecentGoogle DriveGitHubUpload

Enter a GitHub URL or search by organization or user☐ Include private repos

https://github.com/michevan/WIDS-Google/blob/main/data/WiDS%20-%20Regression%20Analysis.ipynb

repository. branch.   
michevan/WIDS-Google main 

Path

 [data/WiDS - Regression Analysis.ipynb](#) 

Cancel

**Note:** You can copy and paste the URL of the notebook under the [notebook](#) folder to access the notebook as well

By clicking on the notebook, you basically have made a copy of the python notebook from the github account to your local environment.



Next time you come to Google drive, you can find the notebook you have copied directly from google drive.

[Here's](#) a helpful beginner video tutorial on how to use Google colab

## Where can I find the data for this workshop:

During the speaker session, we will cover how to use python to approach this year's weather forecasting problem; you can download the data [here](#).

## How to access csv files (data) on Google Colab notebook:

In the workshop, you can access the data in .csv files in 2 ways:

- For the Datathon challenge, you can refer to the code in this [notebook](#) to use python to download data directly from Kaggle and load the data into pandas dataframe. To download the data, you need to make sure you have created an account on Kaggle, joined the competition, and accepted the competition's terms and conditions. (check out the section below on how to use Kaggle)

- Another alternative is to access the csv files through your google drive

As mentioned above, you can download the data (csv files) from Kaggle and upload the files to your google drive.

To download the data for the datathon, click [here](#)

After uploading the files to your google drive, you just need to follow [this video tutorial](#) to load the file from google drive to your pandas dataframe.

## What is Kaggle and how to access the datathon challenge?

### What is Kaggle?

**Kaggle** is a data science community hosting hackathon competitions and sharing public data sources for data science practitioners to solve data science challenges. This is a helpful [post](#) about what Kaggle is and why you should use it if you are not familiar with it.

### How to join the datathon?

#### Step 1: Join the competition

In order to join the competition, you will need to create an account on [kaggle.com](#). If you do not have an account, you can use your email address to create one. It is very straightforward.

After logging into the account, go to the [WiDS datathon challenge](#). Click on the “**Join Competition**” button on the right, then you will be able to access the data.

Also, you will need to fill out the form [here](#) to be able to attend the competition.

# WiDS Datathon 2023

## Adapting to Climate Change by Improving Extreme Weather Forecasts

213 teams · a month to go (a month to go until merger deadline)

[Overview](#)
[Data](#)
[Code](#)
[Discussion](#)
[Leaderboard](#)
[Rules](#)

[Join Competition](#)

Overview

Description

Evaluation

FAQ

Datathon Timeline

Tutorials And Resources

Prizes

WiDS Datathon Partners

Team Formation Tips & Tools

### Women in Data Science (WiDS Datathon) 2023

In advance of the Women in Data Science (WiDS) Stanford Conference to be held on March 8, 2023, we invite you to build a team, hone your data science skills, and join us for the 6th Annual WiDS Datathon focused on social impact. In this year's datathon challenges participants will use data science to improve longer-range weather forecasts to help people prepare and adapt to extreme weather events caused by climate change.

The WiDS Datathon encourages women worldwide to hone their data science skills, creating a supportive environment for women to connect with others in their community who share their interests. Data scientists of all levels are invited to participate in the datathon, including beginners.

[REGISTER HERE](#) to compete in the WiDS Datathon. All participants must register to participate in the challenge.

### Background on the challenge

Extreme weather events are sweeping the globe and range from heat waves, wildfires and drought to hurricanes, extreme rainfall and flooding. These weather events have multiple impacts on agriculture, energy, transportation, as well as low resource communities and disaster planning in countries across the globe.

Accurate long-term forecasts of temperature and precipitation are crucial to help people prepare and adapt to these extreme weather events. Currently, purely physics-based models dominate short-term weather forecasting. But these models have a limited forecast horizon. The availability of observational data offers an opportunity for data scientists to improve sub-seasonal forecasts by blending physics-based

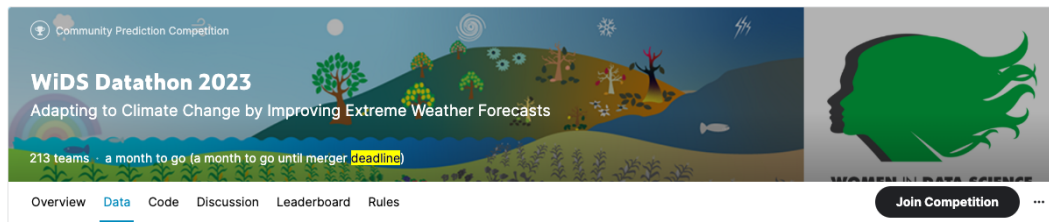
Step 2: Go over the overview of the challenge under the “Overview” tab.

Step 3: Check out the [rules](#) of the competition.

One important thing to remember is that the competition can only allow teams of up to 4 members. Within the team, at least 2 of the team members should be women.



Step 4: Go to the [Data tab](#) to learn more about the data including the data dictionary.



## Dataset Description

### Data Overview

The WiDS Datathon 2023 focuses on a prediction task involving forecasting sub-seasonal temperatures (temperatures over a two-week period, in our case) within the United States. We are using a pre-prepared dataset consisting of weather and climate information for a number of US locations, for a number of start dates for the two-week observation, as well as the forecasted temperature and precipitation from a number of weather forecast models (we will reveal the source of our dataset after the competition closes). Each row in the data corresponds to a single location and a single start date for the two-week period. Your task is to predict the arithmetic mean of the maximum and minimum temperature over the next 14 days, for each location and start date.

You are provided with two datasets:

1. `train_data.csv`: the training dataset, where `contest-tmp2m-14d__tmp2m`, the arithmetic mean of the max and min observed temperature over the next 14 days for each location and start date, is provided
2. `test_data.csv`: the test dataset, where we withhold the true value of `contest-tmp2m-14d__tmp2m` for each row.

To participate in the Datathon, you will submit a solution file containing the predicted values of `contest-tmp2m-14d__tmp2m` for each row in the test dataset. The predicted values you submit will be compared against the observed values for the test dataset and this will determine your standing on the Leaderboard during the competition as well as your final standing when the competition closes.

You are also provided with an example of a solution file prepared for submission.

#### Files

3 files

#### Size

651.4 MB

#### Type

csv

Scroll down to the bottom of the [webpage](#) and click on **download all** to download the data.

- `cancm30`, `cancm40`, `ccsm30`, `ccsm40`, `cfsv20`, `gfdlflora0`, `gfdlflorb0`, `gfdl0`, `nasa0`, `nmme0mean`: most recent forecasts from weather models

#### Target

- `contest-tmp2m-14d__tmp2m`: the arithmetic mean of the max and min observed temperature over the next 14 days for each location and start date, computed as  $(\text{measured max temperature} + \text{measured mini temperature}) / 2$

^ View less

**sample\_solution.csv** (802.82 kB)



#### Competition Rules



To see this data you need to agree to the [competition rules](#).  
By clicking "I understand and agree" you agree to be bound to these rules.

I understand and agree

#### Data Explorer

651.4 MB

- sample\_solution.csv
- test\_data.csv
- train\_data.csv

#### Summary

- 3 files
- 493 columns

Download All

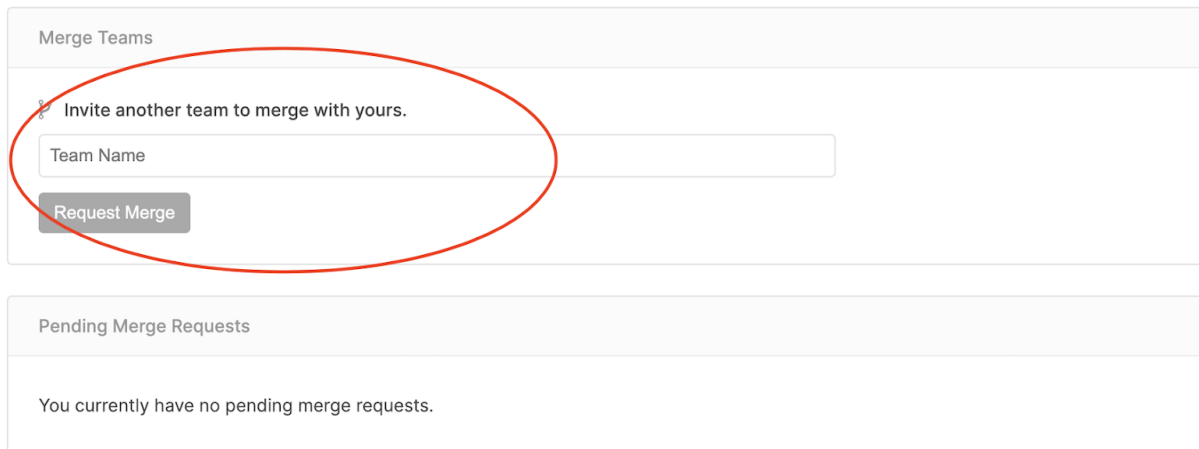
```
> kaggle competitions download -c widsdatathon2023
```



You will have 3 files in the download folder: **train.csv**, **test.csv** and **sample\_solution.csv**. Use the `train.csv` to train your model and use the trained model to make predictions on the `test.csv`. Submit the predictions on the `test.csv` by using `sample_solutions.csv` as the template.

## Step 5: Invite members to your team for collaboration

Make sure your team members/friends have also joined the competition. Go to the team tab [here](#). Add the names of your team members here and request merge. Once your team members have accepted the invitation, then you can collaborate and submit the predictions as a team.



Merge Teams

Invite another team to merge with yours.

Team Name

Request Merge

Pending Merge Requests

You currently have no pending merge requests.

## Step 6: Connect and learn from the Kaggle community

The [Discussion](#) and [Code](#) tabs are great resources to learn from how others are tackling the problem.

## Step 7: Submit your predictions

Click on submit predictions button to submit your predictions



## Step 6: Check out the leaderboard

After the submission, you can go to the [leaderboard](#) to see how your model performs compared with the rest of the community and reiterate.

## Having questions about the tutorials during the speaker session?

Since we have a lot of content to cover during the speaker session, the speakers might not be able to address your questions during the workshop.

However, during the hands-on exercise session, you will be able to connect with other attendees and mentors. Please do not hesitate to ask your questions to the mentors. Even though they are not the ones who put together the tutorials, they will be able to answer more generic questions around data science & machine learning and share best practices.

## Having questions about the event?

If you have specific questions about the workshop/instructions, you can contact our WiDS@Google2023 planning team through this year: [widsgoogle2023@google.com](mailto:widsgoogle2023@google.com).

We hope you find this instruction doc helpful and enjoy the event on Friday!

**Yuka Abe (WiDS Ambassador 2023)**