



WOMEN IN DATA SCIENCE
@ GOOGLE



WiDS@Google 2022 Datathon

Workshop Preparation

Welcome to WiDS@Google 2022 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!

WiDS@Google 2022 Datathon	1
Workshop Preparation	1
Where are the workshop materials? How to use github?	2
How to use Google Colab?	3
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
What is Kaggle and how to access the datathon challenge?	10
What is Kaggle?	10
How to join the datathon?	10
Having questions about the tutorials during the speaker session?	13
Having questions about the event?	

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.

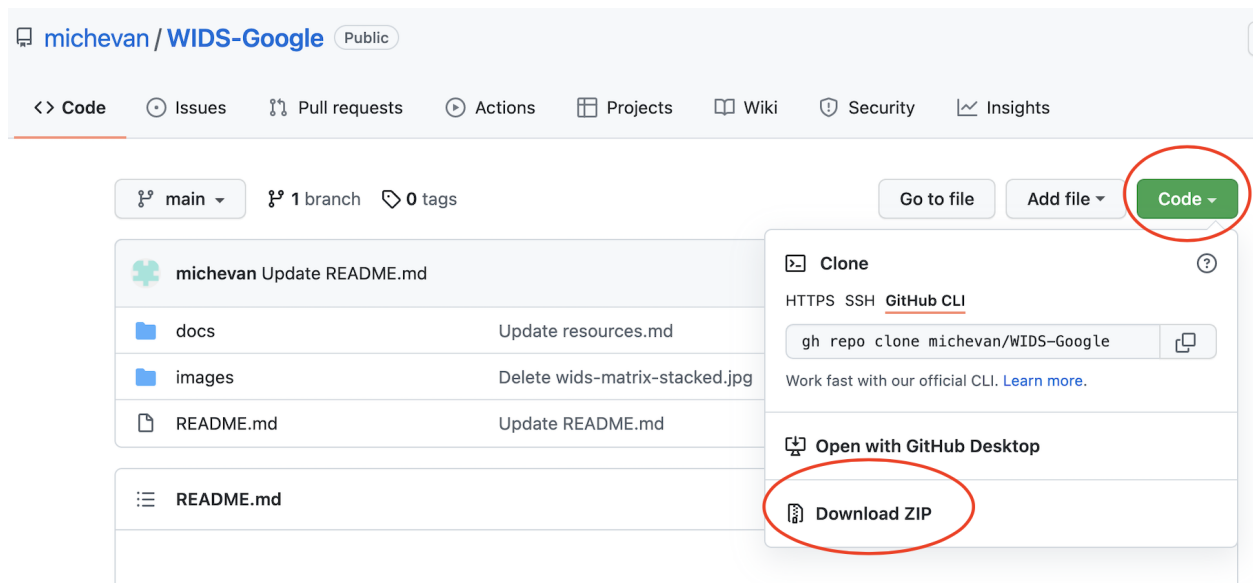
Under the github repository, you will be able to see two main folders:

(Note that this is still work in progress. We will make sure all the materials get uploaded Feb 16th EOD)

- [Colab notebooks](#):
 - WiDS_Regression_Analysis.ipynb (python code demo for regression problem)
 - Template_datathon_no_solution.ipynb (exercise notebook with no solution)
 - Template_datathon_with_solution.ipynb (solution notebook for hands-on exercise)
- [Slides](#)
 - [WiDS datathon] Intro to DS and ML.pdf
 - [WiDS Datathon] Data Science Code Walkthrough.pdf

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.



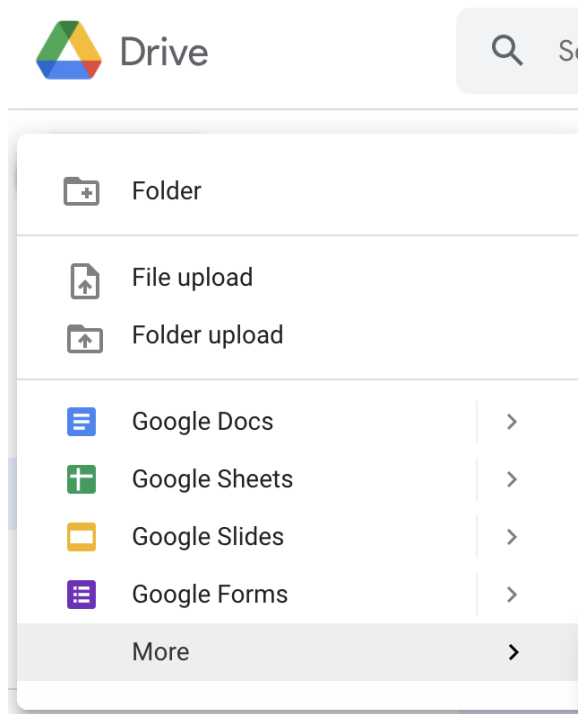
How to use Google Colab?

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

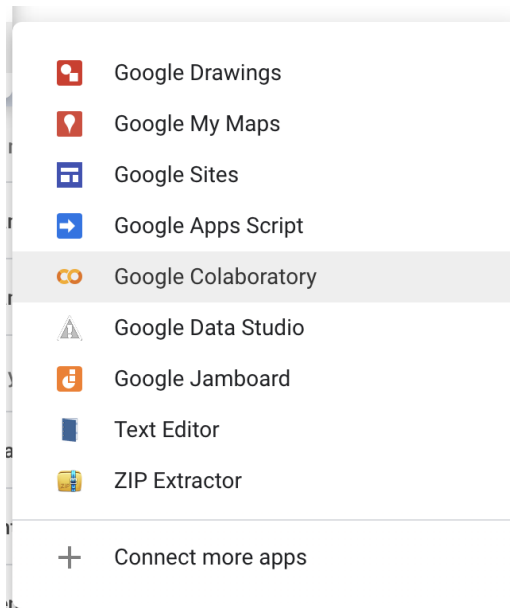
work and collaborate online. In case you are not familiar with Google Colab, you can check out the following instructions.

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

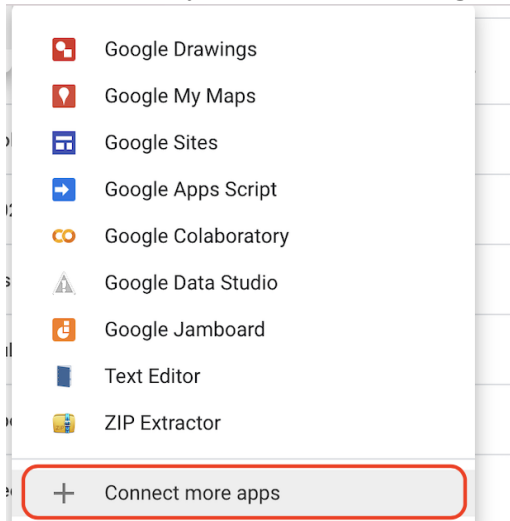
Go to the Google drive and click on **More**.



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

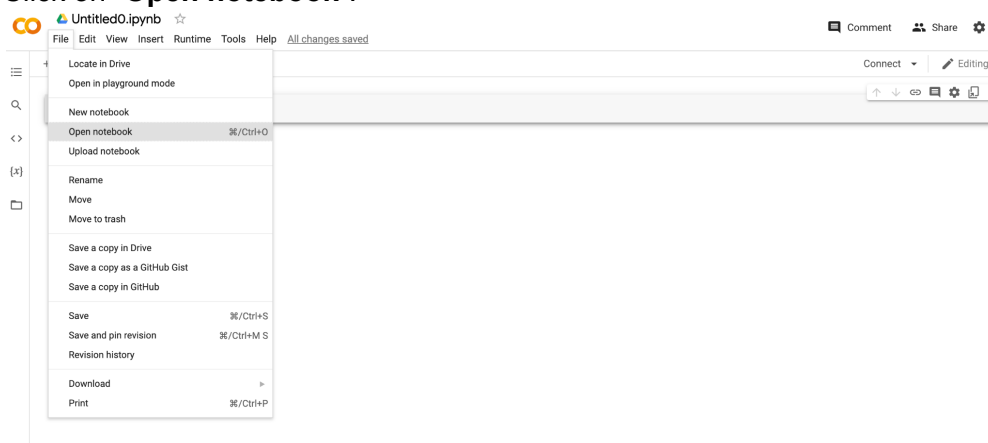
Search results for Google Colab



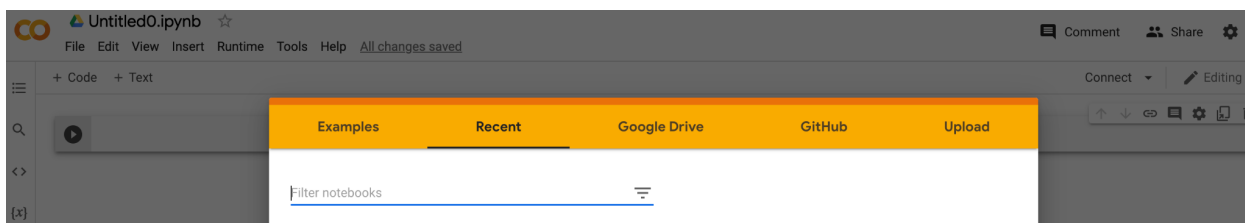
After installing Google Colaboratory, you can follow the steps above to access it.

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.

Examples Recent Google Drive **GitHub** Upload

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 URLs of the notebooks under the [colab_notebook](#) to access the different notebooks we have prepared for you

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 solve a typical house price prediction regression problem, you can download the data [here](#).

For the hands-on exercise, we will use the data provided by this WiDS datathon. You can find 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 3 main 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 python demo regression problem, click here to [get house price prediction data](#).

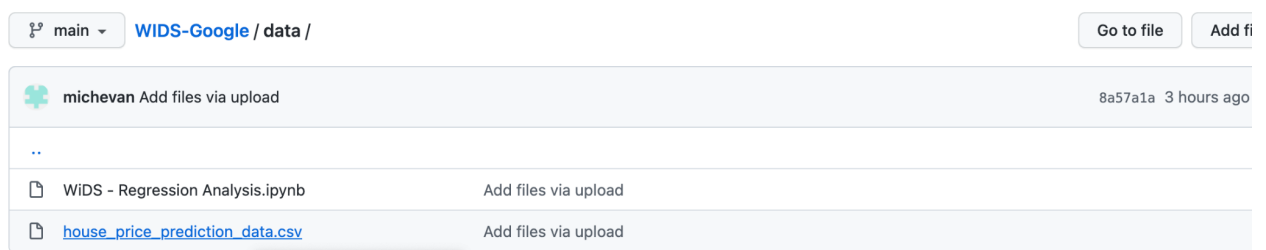
To download the data for the datathon, click [here](#)

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.

- The third way is to access the files directly from the github repository (if you have one)

For instance, if we have a file called house_price_prediction_data.csv file on a github repo as below:

Click on the house_price_prediction_data.csv file:



Click on the “**Raw**” button:

main
WIDS-Google / data / house_price_prediction_data.csv
Go to file

michevan Add files via upload
Latest commit 85911c3 3 hours ago
History

1 contributor

1460 Lines (1460 sloc) | 442 KB
Raw
Blame

Search this file...

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Condition2	BldgType	HouseStyle
1	1461	20	RH	80	11622	Pave	NA	Reg	Lvl	AllPub	Inside	Gtl	NAmes	Feedr	Norm	1Fam	1Story
2	1462	20	RL	81	14267	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	NAmes	Norm	Norm	1Fam	1Story
4	1463	60	RL	74	13830	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story
5	1464	60	RL	78	9978	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	2Story
6	1465	120	RL	43	5005	Pave	NA	IR1	HLS	AllPub	Inside	Gtl	StoneBr	Norm	Norm	TwnhsE	1Story
7	1466	60	RL	75	10000	Pave	NA	IR1	Lvl	AllPub	Corner	Gtl	Gilbert	Norm	Norm	1Fam	2Story
8	1467	20	RL	NA	7980	Pave	NA	IR1	Lvl	AllPub	Inside	Gtl	Gilbert	Norm	Norm	1Fam	1Story

Copy the url at the address bar:

← → ↻
raw.githubusercontent.com/michevan/WIDS-Google/main/data/house_price_prediction_data.csv

Apps
My Stuff | Homep...
TechSpot - Googli...
Moma Home
go/shortlinks - Go...
go/helpful@home
Global employeeme...
Employee handbo...
US Benefits
»
Reading List

Id,MSSubClass,MSZoning,LotFrontage,LotArea,Street,Alley,LotShape,LandContour,Utilities,LotConfig,LandSlope,Neighborhood,Condition1,Condition2,BldgType,HouseStyle,OverallQual,OverallC
ond,YearBuilt,YearRemodAdd,RoofStyle,RoofMatl,Exterior1st,Exterior2nd,MasVnrType,MasVnrArea,ExterQual,ExterCond,Foundation,BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1,BsmtFinSF1,Bsmt
FinType2,BsmtFinSF2,BsmtUnfSF,TotalBsmtSF,Heating,HeatingQC,CentralAir,Electrical,1stFlrSF,2ndFlrSF,LowQualFinSF,GrLivArea,BsmtFullBath,BsmtHalfBath,FullBath,HalfBath,BedroomAbvGr,Ki
tchenAbvGr,KitchenQual,TotRmsAbvGrd,Functional,Fireplaces,FireplaceQu,GarageType,GarageYrBlt,GarageFinish,GarageCars,GarageArea,GarageQual,GarageCond,PavedDrive,WoodDeckSF,OpenPorchS
F,EnclosedPorch,3SsnPorch,ScreenPorch,PoolArea,PoolQC,Fence,MiscFeature,MiscVal,MoSold,YrSold,SaleType,SaleCondition
1461,20,RH,80,11622,Pave,NA,Reg,Lvl,AllPub,Inside,Gtl,NAmes,Feedr,Norm,1Fam,1Story,5,6,1961,1961,Gable,CompShg,VinylSd,VinylSd,None,0,TA,TA,CBlock,TA,TA,No,Rec,468,LwQ,144,270,882,Ga
sa,TA,Y,SBkr,896,0,0,896,0,0,1,0,2,1,TA,5,Typ,0,NA,Atchd,1961,Unf,1,730,TA,TA,Y,140,0,0,0,120,0,NA,MnPrv,NA,0,6,2010,WD,Normal
1462,20,RL,81,14267,Pave,NA,IR1,Lvl,AllPub,Corner,Gtl,NAmes,Norm,Norm,1Fam,1Story,6,6,1958,1958,Hip,CompShg,Wd Sdng,Wd
Sdng,BrkFace,108,TA,TA,CBlock,TA,TA,No,ALQ,923,Unf,0,406,1329,Gasa,TA,Y,SBkr,1329,0,0,1329,0,0,1,1,3,1,Gd,6,Typ,0,NA,Atchd,1958,Unf,1,312,TA,TA,Y,393,36,0,0,0,0,NA,NA,Gar2,12500,6,
2010,WD,Normal
1463,60,RL,74,13830,Pave,NA,IR1,Lvl,AllPub,Inside,Gtl,Gilbert,Norm,Norm,1Fam,2Story,5,5,1997,1998,Gable,CompShg,VinylSd,VinylSd,None,0,TA,TA,PConc,Gd,TA,TA,GLQ,791,Unf,0,137,928,Gasa
,Gd,Y,SBkr,928,701,0,1629,0,0,2,1,3,1,TA,6,Typ,1,TA,Atchd,1997,Fin,2,482,TA,TA,Y,212,34,0,0,0,0,NA,MnPrv,NA,0,3,2010,WD,Normal
1464,60,RL,78,9978,Pave,NA,IR1,Lvl,AllPub,Inside,Gtl,Gilbert,Norm,Norm,1Fam,2Story,6,6,1998,1998,Gable,CompShg,VinylSd,VinylSd,BrkFace,20,TA,TA,PConc,TA,TA,No,GLQ,602,Unf,0,324,926,G
asa,Ex,Y,SBkr,926,678,0,1604,0,0,2,1,3,1,Gd,7,Typ,1,Gd,Atchd,1998,Fin,2,470,TA,TA,Y,360,36,0,0,0,0,NA,NA,NA,0,6,2010,WD,Normal
1465,120,RL,43,5005,Pave,NA,IR1,HLS,AllPub,Inside,Gtl,StoneBr,Norm,Norm,TwnhsE,1Story,8,5,1992,1992,Gable,CompShg,HdBoard,HdBoard,None,0,Gd,TA,PConc,Gd,TA,TA,No,ALQ,263,Unf,0,1017,1280,
Gasa,Ex,Y,SBkr,1280,0,0,1280,0,0,2,0,2,1,Gd,5,Typ,0,NA,Atchd,1992,Rfn,2,506,TA,TA,Y,0,82,0,0,144,0,NA,NA,NA,0,1,2010,WD,Normal
1466,60,RL,75,10000,Pave,NA,IR1,Lvl,AllPub,Corner,Gtl,Gilbert,Norm,Norm,1Fam,2Story,6,5,1993,1994,Gable,CompShg,HdBoard,HdBoard,None,0,TA,TA,PConc,Gd,TA,TA,No,Unf,0,Unf,0,763,763,Gasa,G
d,Y,SBkr,763,892,0,1655,0,0,2,1,3,1,TA,7,Typ,1,TA,Atchd,1993,Fin,2,440,TA,TA,Y,157,84,0,0,0,0,NA,NA,NA,0,4,2010,WD,Normal
1467,20,RL,NA,7980,Pave,NA,IR1,Lvl,AllPub,Inside,Gtl,Gilbert,Norm,Norm,1Fam,1Story,6,7,1992,2007,Gable,CompShg,HdBoard,HdBoard,None,0,TA,Gd,PConc,Gd,TA,TA,No,ALQ,935,Unf,0,233,1168,Gasa
,Ex,Y,SBkr,1187,0,0,1187,1,0,2,0,3,1,TA,6,Typ,0,NA,Atchd,1992,Fin,2,420,TA,TA,Y,483,21,0,0,0,0,NA,GdPrv,Shed,500,3,2010,WD,Normal
1468,60,RL,63,8402,Pave,NA,IR1,Lvl,AllPub,Inside,Gtl,Gilbert,Norm,Norm,1Fam,2Story,6,5,1998,1998,Gable,CompShg,VinylSd,VinylSd,None,0,TA,TA,PConc,Gd,TA,TA,No,Unf,0,Unf,0,789,789,Gasa,Gd
Y,SBkr,789,676,0,1465,0,0,2,1,3,1,TA,7,Typ,1,Gd,Atchd,1998,Fin,2,393,TA,TA,Y,0,75,0,0,0,0,NA,NA,NA,0,5,2010,WD,Normal
1469,20,RL,85,10176,Pave,NA,Reg,Lvl,AllPub,Inside,Gtl,Gilbert,Norm,Norm,1Fam,1Story,7,5,1990,1990,Gable,CompShg,HdBoard,HdBoard,None,0,TA,TA,PConc,Gd,TA,Gd,GLQ,637,Unf,0,663,1300,Gas
A,Gd,Y,SBkr,1341,0,0,1341,1,0,1,1,2,1,Gd,5,Typ,1,Po,Atchd,1990,Unf,2,506,TA,TA,Y,192,0,0,0,0,0,0,NA,NA,NA,0,2,2010,WD,Normal
1470,20,RL,70,8400,Pave,NA,Reg,Lvl,AllPub,Corner,Gtl,NAmes,Norm,Norm,1Fam,1Story,4,5,1970,1970,Gable,CompShg,Plywood,Plywood,None,0,TA,TA,CBlock,TA,TA,No,ALQ,804,Rec,78,0,882,Gasa,TA
,Y,SBkr,882,0,0,882,1,0,1,0,2,1,TA,4,Typ,0,NA,Atchd,1970,Fin,2,525,TA,TA,Y,240,0,0,0,0,0,0,NA,MnPrv,NA,0,4,2010,WD,Normal

Paste to replace the url highlighted in yellow with the copied url and use the following code to load the to the pandas data frame.

```
import pandas as pd

url =
'https://raw.githubusercontent.com/michevan/WIDS-Google/main/data/house_price_prediction_data.csv'

df = pd.read_csv(url)
```


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.


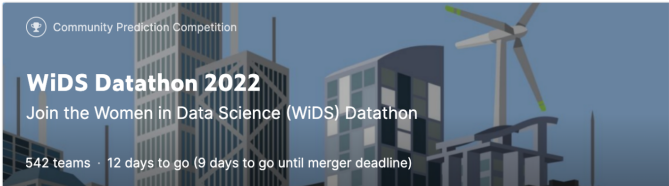
After logging into the account, go to the [WiDS datathon challenge](#). Click on the “**Join Competition**” button on the right, then you will be able to access the data.

Community Prediction Competition

WiDS Datathon 2022

Join the Women in Data Science (WiDS) Datathon

542 teams · 12 days to go (9 days to go until merger deadline)



WOMEN IN DATA SCIENCE
STANFORD UNIVERSITY

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

[Join Competition](#) ...

Overview

Description

Evaluation

FAQ

Datathon Phase II: Excellence In Research Award

Datathon Timeline

Tutorials And Resources

Prizes

WiDS Datathon

WiDS Datathon 2022

In advance of the Women in Data Science (WiDS) Worldwide Conference to be held on March 7, 2022, we invite you to build a team, hone your data science skills, and join us for the 5th Annual WiDS Datathon focused on social impact.

This year's WiDS Datathon, organized by the WiDS Worldwide team, [Stanford University](#), [Harvard University IACS](#), and the [WiDS Datathon Committee](#), will address an important way to mitigate the effects of climate change with a focus on energy efficiency. The WiDS Datathon Committee is partnering with experts from many disciplines at [Climate Change AI \(CCAI\)](#), [Lawrence Berkeley National Laboratory \(Berkeley Lab\)](#), [US Environmental Protection Agency \(EPA\)](#), and [MIT Critical Data](#). This year's datathon is open until February 26, 2022. Winners will be announced at the WiDS Conference via livestream, reaching a community of 100,000+ data enthusiasts across more than 85 countries.

[REGISTER HERE](#) (this is a required step to participate in the WiDS Datathon).

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.

WIDS Datathon 2022

Join the Women in Data Science (WiDS) Datathon

542 teams · 12 days to go (9 days to go until merger deadline)

WOMEN IN DATA SCIENCE
STANFORD UNIVERSITY

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

My Submissions

Submit Predictions

...

Data Description

Data Overview

The WIDS Datathon 2022 focuses on a prediction task involving roughly 100k observations of building energy usage records collected over 7 years and a number of states within the United States. The dataset consists of building characteristics (e.g. floor area, facility type etc), weather data for the location of the building (e.g. annual average temperature, annual total precipitation etc) as well as the energy usage for the building and the given year, measured as Site Energy Usage Intensity (Site EUI). Each row in the data corresponds to the a single building observed in a given year. Your task is to predict the Site EUI for each row, given the characteristics of the building and the weather data for the location of the building.

You are provided with two datasets: (1) the training dataset where the observed values of the Site EUI for each row is provided and (2) the test dataset where we withhold the observed values of the Site EUI for each row. To participate in the Datathon, you will submit a solution file containing the predicted Site EUI values for each row in the test dataset. The predicted values you submit will be compared against the observed Site EUI 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.

Note: During the competition the leaderboard is calculated with approximately 51% of the test data. After the competition closes, the final standings will be computed based on the other 49%. As such, *the final leaderboard standings may be different than those during the competition.*

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

Overview

Data

Code

Discussion

Leaderboard

Rules

Team

My Submissions

Submit Predictions

...

75782	0.0
75783	0.0
75784	0.0
75785	0.0
75786	0.0
75787	0.0
75788	0.0
75789	0.0
75790	0.0
75791	0.0
75792	0.0
75793	0.0
75794	0.0
75795	0.0

Summary

3 files

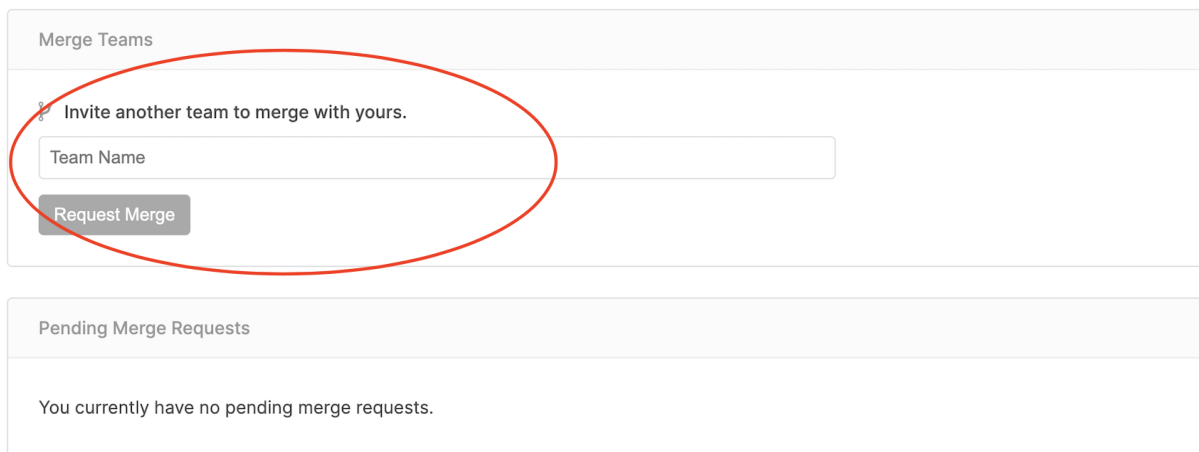
129 columns

Download All

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

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

Step 7: Submit your predictions

Go to the submit tab [here](#) to follow the instructions and 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 your hosts [Yuka](#) or [Sumedha](#) via LinkedIn.

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

Yuka Abe (WiDS Ambassador 2022)