

Hands-on Activity: Introduction to Kaggle

TOTAL POINTS 2

1.



Activity overview

By now, you've learned a lot about different data types and data structures. In this activity, you will work with datasets from **Kaggle**, an online community of people passionate about data. To start this activity, you'll create a Kaggle account, set up a profile, and explore Kaggle notebooks.

Every data analyst has a data community that they rely on for help, support, and inspiration. Kaggle can help you build your own data community.

Kaggle has millions of users in all stages of their data career, from beginners to data scientists with decades of experience. The Kaggle community brings people together to develop their data analysis skills, share datasets and interactive notebooks, and collaborate on solving real-life data problems.

Check out this [brief introductory video](#) to learn more about Kaggle.

By the time you complete this activity, you will be able to use many of Kaggle's key features. This will enable you to create notebooks and browse data, which is important for completing and sharing data projects in your career as a data analyst.

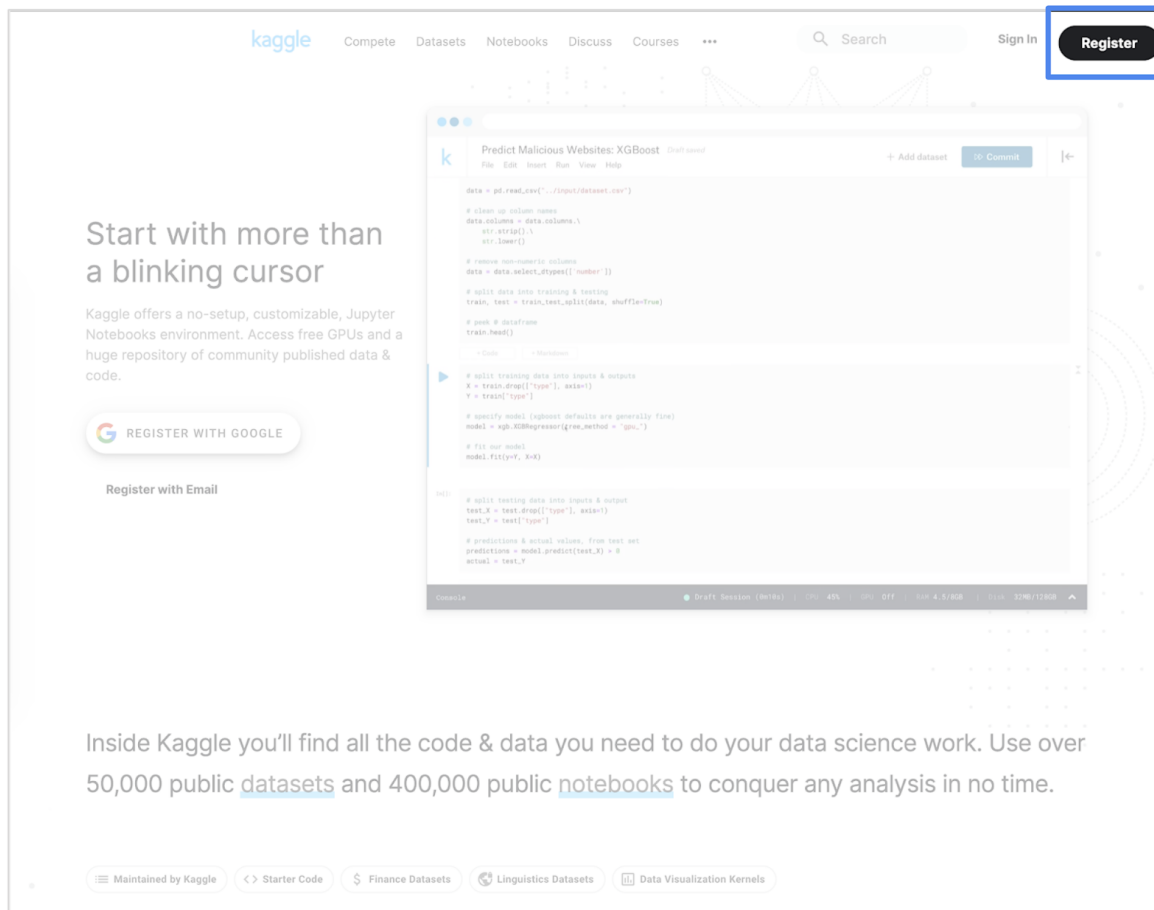
Create a Kaggle account

To get started, follow these steps to create a Kaggle account.

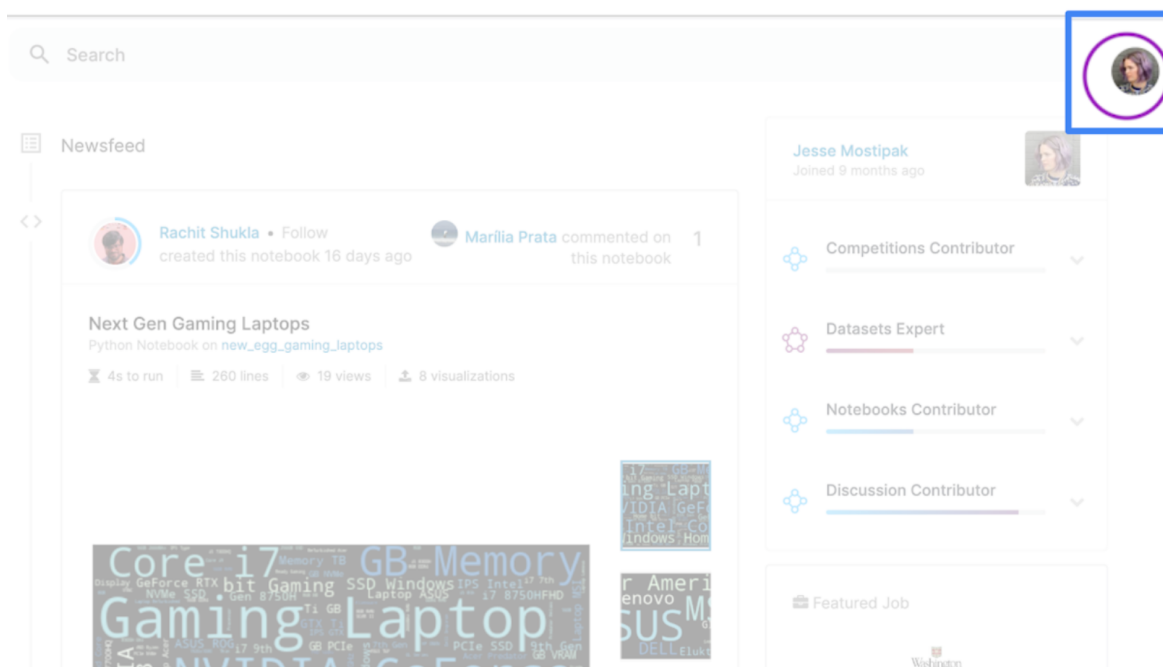
- **Note:** Kaggle frequently updates its user interface. The latest changes may not be reflected in the screenshots, but the principles in this activity remain the same. Adapting to changes in software updates is an essential skill for data analysts, and we encourage you to practice troubleshooting. You can also reach out to your community of learners on the discussion forum for help.

1. Go to kaggle.com

2. Click on the **Register** button at the top-right of the Kaggle homepage. You can register with your Google credentials or your personal email address.

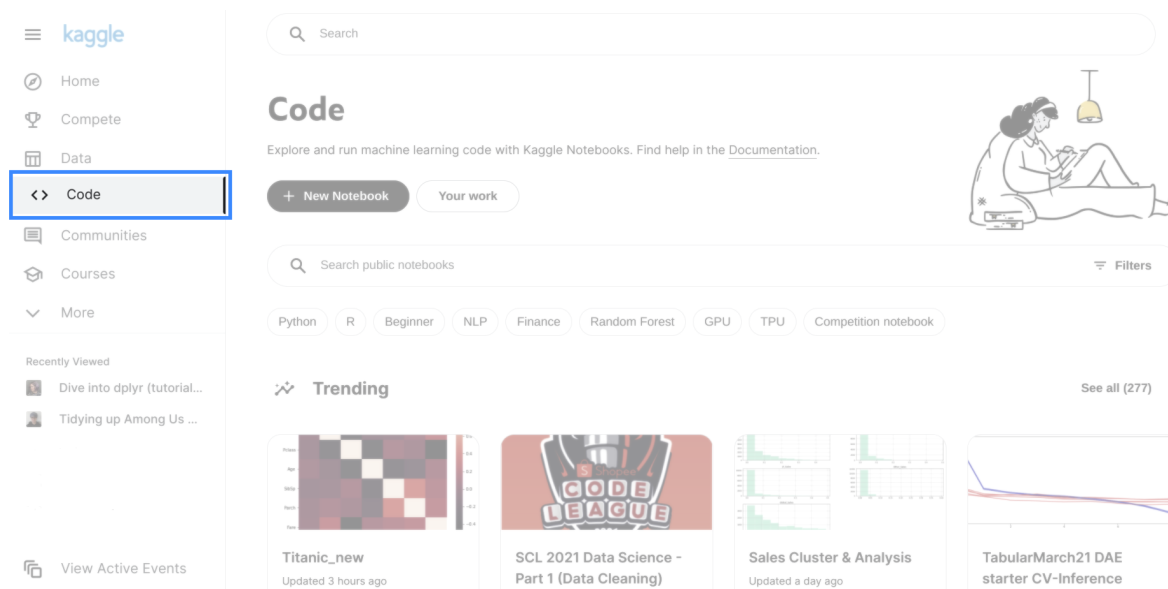


3. Once you're registered and logged in to Kaggle, click on the **Account** icon at the top-right of your screen. In the menu that opens, click the **Your Profile** button.



Step 1: Go to the Code home page

First, go to the **Navigation** bar on the left side of your screen. Then, click on the **Code** icon. This takes you to the Code home page.



Step 2: Review Kagglers contributions

On the Code home page, you'll notice links to notebooks created by other Kagglers.

To begin, feel free to scroll through the list and click on notebooks that interest you. As you explore, you may come across unfamiliar terms and new information: That's fine! Kagglers come from diverse backgrounds and focus on different areas of data analysis, data science, machine learning, and deep learning.

Step 3: Narrow your search

Once you're familiar with the Code home page, you can narrow your search results by typing a word in the search bar or by using the filter feature.

For example, type **Beginner** in the search bar to show notebooks tagged as beginner-friendly. Or, click on the **Filter** icon, the triangle shape on the right side of the search bar. You can filter results by tags, programming language, output, and other options. Filter to **Datasets** to show notebooks that use one of the tens of thousands of public datasets available on Kaggle.

Step 4: Review suggested notebooks

If you're looking for specific suggestions, check out the following notebooks:

- [gganimate](#) by Meg Risdal
- [Getting staRted in R](#) by Rachael Tatman
- [Writing Hamilton Lyrics with TensorFlow/R](#) by Ana Sofia Uzsoy
- [Dive into dplyr \(tutorial #1\)](#) by Jesse Mostipak

Spend some time checking out a couple of notebooks to get an idea of the work that Kagglers share online—and that you'll be able to create by the time you've finished this course!

Edit a notebook

Now, take a look at a specific notebook: [Dive into dplyr \(tutorial #1\)](#) by Jesse Mostipak. Follow these steps to learn how to edit notebooks:

1. Click on the link to open up the notebook. It contains the dataset you'll work with later on.
2. Click on the **Copy and Edit** button at the top-right to make a copy of the notebook in your account. Now, the notebook appears in **Edit** mode. Edit mode lets you make changes to the notebook if you want.

The screenshot shows the Kaggle notebook editor interface. At the top, the notebook title is "Dive into dplyr (tutorial #1)" with a "Draft saved" status. On the right, there are buttons for "Share", "Save Version", and a version count of "13". Below the title bar is a menu bar with "File", "Edit", "View", "Run", "Add-ons", and "Help". A toolbar contains icons for adding, running, and saving code cells. The main content area displays the notebook's text, which includes an introduction to the `dplyr` package, a list of learning objectives, and a workflow section. The right sidebar contains panels for "Data" (showing input and output files), "Settings" (with options for Language, Environment, Accelerator, and Internet), and "Code Help" (with a search bar).

Introduction: why dplyr?

There are a lot of amazing packages in the [Tidyverse](#), but `dplyr` is hands-down my absolute favorite package. I use `dplyr` when I'm cleaning and exploring my dataset, and what I particularly love is that after I get a good handle on my dataset with `dplyr`, I can feed the various manipulations I've created into the `ggplot2` package for visualization.

This tutorial is for anyone interested in learning the basics of the `dplyr` package. We'll be focusing on data exploration and manipulation, building off of the examples in the `dplyr` package documentation using the [Palmer Penguins](#) dataset.

By the end of this notebook, you'll be able to:

- Demonstrate what each of the main five `dplyr` functions does
- Use the pipe operator `%>%` to chain together multiple `dplyr` functions

What I've learned

-

I still have questions about

-

My analytical workflow

We won't be covering *all* of the steps in my workflow in this tutorial, but in general I follow these steps:

1. Set up the programming environment by loading packages
2. Import my data
3. Check out my data
4. Explore my data
5. Model my data
6. Communicate what I've learned

3. Take a moment to explore the Edit mode of the notebook.

Working with datasets in notebooks

In this notebook, you'll find the data in a box labeled **Data** at the top-right of your screen. In the box, there's an input folder with the title: **palmer-archipelago-antarctica-penguin-data**. Follow these instructions to explore the datasets and learn more about the data within them:

Dive into dplyr (tutorial #1)

Draft saved

File Edit View Run Add-ons Help

+ > >> Run All

Draft Session (1m)

Share

Save Version 13

Introduction: why dplyr?

There are a lot of amazing packages in the `Tidyverse`, but `dplyr` is hands-down my absolute favorite package. I use `dplyr` when I'm cleaning and exploring my dataset, and what I particularly love is that after I get a good handle on my dataset with `dplyr`, I can feed the various manipulations I've created into the `ggplot2` package for visualization.

This tutorial is for anyone interested in learning the basics of the `dplyr` package. We'll be focusing on data exploration and manipulation, building off of the examples in the `dplyr` package documentation using the `Palmer Penguins` dataset.

By the end of this notebook, you'll be able to:

- Demonstrate what each of the main five `dplyr` functions does
- Use the pipe operator `%>%` to chain together multiple `dplyr` functions

penguins_iter.csv

../input/palmer-archipelago-antarctica-penguin-data/penguins_iter.csv

Copy

10 of 17 columns

studyName	Sample Num	Species	Region	Island	Stage	Individual ID	Clutch-Size	Date Egg	Culmen Le...
PAL0708	1	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N1A1	Yes	11/11/07	39.1
PAL0708	2	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N1A2	Yes	11/11/07	39.5
PAL0708	3	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N2A1	Yes	11/16/07	40.3
PAL0708	4	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N2A2	Yes	11/16/07	
PAL0708	5	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N3A1	Yes	11/16/07	36.7
PAL0708	6	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N3A2	Yes	11/16/07	39.3
PAL0708	7	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N4A1	No	11/15/07	38.9
PAL0708	8	Adelie Penguin (Pygoscelis adellae)	Anvers	Torgersen	Adult, 1 Egg Stage	N4A2	No	11/15/07	39.2

Data

+ Add data

input (62.92 KB)

palmer-archipelago-antarctica-penguins_iter.csv

penguins_size.csv

output

/kaggle/working

Settings

Language R

Environment Preferences

Accelerator None

Internet On

Code Help

Find Code Help

Search for examples of how to do things

2. Click on the other .csv file. This opens a second tab with the second dataset.

3. Take a moment to check out each dataset.

4. Sort the data in each column by clicking on the **horizontal bars** to the right of each column name.

5. Click on the button that says **10 of 17 columns** to change the columns that are visible in the table.

In the dropdown menu, there's a checkmark next to the name of each column that appears in the table. Checking or unchecking one of these boxes will change what data is presented.

Congratulations! You've explored several ways to interact with the dataset. This will help you get familiar with the Kaggle interface. You can save the notebook you worked in for future reference. Coming up, you'll learn more about other ways you can use Kaggle.