

✔ **Congratulations! You passed!**

Grade received **100%** To pass 100% or higher

[Go to next item](#)

Hands-on Activity: Introduction to Kaggle

Total points 2

1.



1 / 1 point

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.

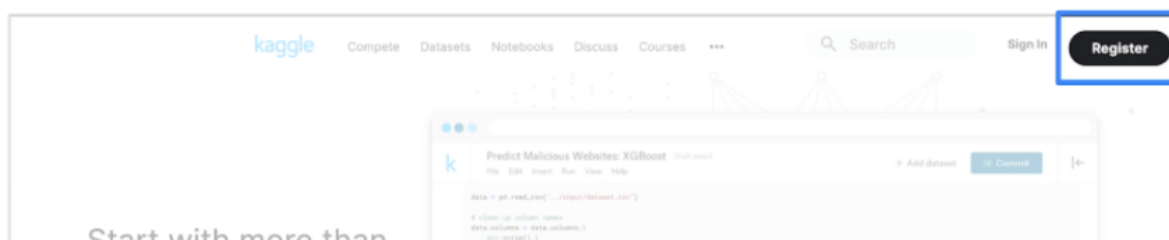
Create a Kaggle account

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.



Start with more than a blinking cursor

Kaggle offers a no-setup, customizable, Jupyter Notebooks environment. Access free GPUs and a huge repository of community published data & code.

 REGISTER WITH GOOGLE

Register with Email

```

def train():
    # remove non-numeric values
    data = data.select_dtypes(include=[np.number])

    # split data into training & testing
    train, test = train_test_split(data, shuffle=True)

    # pick X data/features
    train.head()

    # split training data into inputs & outputs
    X = train.drop("type", axis=1)
    Y = train["type"]






    # specify model (highest defaults are generally fine)
    model = xgb.XGBRegressor(grow_method="gpu_")

    # fit our model
    model.fit(X, Y)

    # split testing data into inputs & outputs
    test_X = test.drop("type", axis=1)
    test_Y = test["type"]

    # predictions & actual values, from test set
    predictions = model.predict(test_X)
    actuals = test_Y
  
```


Inside Kaggle you'll find all the code & data you need to do your data science work. Use over 50,000 public [datasets](#) and 400,000 public [notebooks](#) to conquer any analysis in no time.


 Maintained by Kaggle
  Starter Code
  Finance Datasets
  Linguistics Datasets
  Data Visualization Kernels

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.

Search


Newsfeed

 Rachit Shukla • Follow
created this notebook 16 days ago

 Marilia Prata commented on this notebook 1

Next Gen Gaming Laptops
Python Notebook on [new_egg_gaming_laptops](#)

4s to run | 260 lines | 19 views | 6 visualizations



Jesse Mostipak
Joined 9 months ago

- Competitions Contributor
- Datasets Expert
- Notebooks Contributor
- Discussion Contributor

Featured Job

Washington University in St. Louis is hiring
Applications Developer III
St. Louis, Missouri, United States

4. On your profile page, click on the **Edit Profile** button. Enter any information you'd like to share with the Kaggle community. Your profile will be public, so only enter the information you're comfortable sharing.

Search

Jesse Mostipak
Community Advocate at Kaggle
Dallas, Texas, United States
Joined 9 months ago · last seen in the past day
Followers 113
Following 22
<https://www.jessemaegan.com/>

Edit Profile

Home Competitions Datasets Code Discussion Followers Notifications Account

Competitions Contributor	Datasets Expert	Notebooks Contributor	Discussion Contributor
Unranked	Unranked	Unranked	Unranked
0	1	0	2
0	1	0	3
0	3	2	39
Five Million User... 4 months ago Top 18%	Hotel booking d... 9 months ago	Dive into dplyr (... a month ago	New video: wha... 7 months ago
3 rd of 16	1138 votes	21 votes	32 votes
Housing Prices ... 8 years to go Top 7%	Animal Crossing... 6 months ago	Starter Noteboo... 6 months ago	New video: wha... 7 months ago
3,475 th of 52375	94 votes	5 votes	25 votes
	Caribou Locatio... 5 months ago	Palmer Penguin... 4 months ago	Opportunity to l... 2 months ago
	22 votes	4 votes	9 votes

5. If you want some inspiration, check out the profile of [Kaggle's Community Advocate, Jesse Mostipak](#)!

Explore Kaggle notebooks

Now that you've created an account and set up your profile, you can check out some notebooks on Kaggle. Kagglers use **notebooks** to share datasets and data analyses.

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.

Search

Code
Explore and run machine learning code with Kaggle Notebooks. Find help in the [Documentation](#).

+ New Notebook Your work

Search public notebooks Filters

Python R Beginner NLP Finance Random Forest GPU TPU Competition notebook

Trending See all (277)



Step 2: Review Kaggle 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:

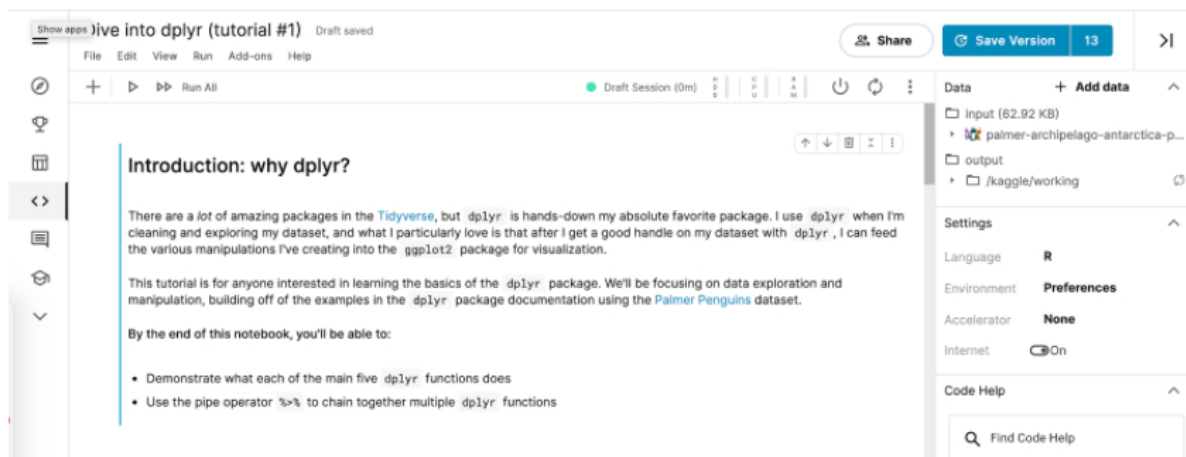
- [gganimate](#) by Meg Risdal
- [Getting started in R](#) by Rachael Tatman
- [Writing Hamilton Lyrics with TensorFlow/R](#) by Ana Sofia Uzsoy
- [Dive into dplyr \(tutorial #1\)](#) by Jesse Mostipak

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.



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

Search for examples of how to do things

This notebook is private. If you want to share your work, you can choose to make it public. When you copy and edit another Kaggle's work, always make meaningful changes to the notebook before publishing it. That way, you're not misrepresenting someone else's work as your own.

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

Some of this may seem unfamiliar—and that's just fine. By the end of this course, you'll know how to create a notebook like this from scratch!

Working with datasets in notebooks

Now, you can check out the data!

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:

1. Click on this title. Two .csv files appear: **penguins_lter.csv** and **penguins_size.csv**. Click on one of them. At the bottom of the notebook, you'll now find an interactive data table with all the information from the dataset.

Dive into dplyr (tutorial #1) Draft saved

File Edit View Run Add-ons Help

Run All

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_lter.csv

Copy 10 of 17 columns

studyName	Sample No.	Species	Region	Island	Stage	individual ID	Clutch Co.	Date Egg	Culmen l.e.
PAU.0708	1	Adelie Penguin (Pygoscelis adeliae)	Antarctica	Torgersen	Adult, 1 Egg Stage	NIA1	Yes	11/11/07	39.1
PAU.0708	2	Adelie Penguin (Pygoscelis adeliae)	Antarctica	Torgersen	Adult, 1 Egg Stage	NIA2	Yes	11/11/07	39.5
PAU.0708	3	Adelie Penguin (Pygoscelis adeliae)	Antarctica	Torgersen	Adult, 1 Egg Stage	NIA1	Yes	11/16/07	40.3
PAU.0708	4	Adelie Penguin (Pygoscelis adeliae)	Antarctica	Torgersen	Adult, 1 Egg Stage	NIA2	Yes	11/16/07	
PAU.0708	5	Adelie Penguin (Pygoscelis adeliae)	Antarctica	Torgersen	Adult, 1 Egg Stage	NIA1	Yes	11/16/07	36.7
PAU.0708	6	Adelie Penguin (Pygoscelis adeliae)	Antarctica	Torgersen	Adult, 1 Egg Stage	NIA2	Yes	11/16/07	38.3

Search for examples of how to do things

PAU.0708	7	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen	Adult, 1 Egg Stage	MAA1	No	11/15/07	38.9
PAU.0708	8	Adelie Penguin	Anvers	Torgersen	Adult, 1 Egg Stage	MAA2	No	11/15/07	38.2

- Click on the other .csv file. This opens a second tab with the second dataset.
- Take a moment to check out each dataset.
- Sort the data in each column by clicking on the **horizontal bars** to the right of each column name.
- 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.

Confirmation and reflection

Which statements are true about the two penguin datasets in the **Dive into dplyr (tutorial #1)** notebook? Select all that apply.

☒ **penguins_size.csv** has 7 columns.

☒ **Correct**

The **penguins_size.csv** has 7 columns. In **penguins_lter.csv**, the highest value in the column Sample Number is 152. To learn about the penguin datasets, you used an interactive notebook's data viewing feature. Going forward, you can use interactive notebooks to examine and describe data. This is an important skill that will help you complete data projects in the future.

☐ In **penguins_lter.csv**, the column Individual ID cannot be sorted.

☐ In both datasets, the number of columns is the same.

☒ In **penguins_lter.csv**, the highest value in the column Sample Number is 152.

☒ **Correct**

The **penguins_size.csv** has 7 columns. In **penguins_lter.csv**, the highest value in the column Sample Number is 152. To learn about the penguin datasets, you used an interactive notebook's data viewing feature. Going forward, you can use interactive notebooks to examine and describe data. This is an important skill that will help you complete data projects in the future.

- In this activity, you've learned a lot about data types and data structures. Using what you've learned so far, consider your experience with datasets and the two penguins datasets. In the text box below, write 2-3 sentences (40-60 words) in response to each of the following questions:

1 / 1 point

- Using all of the information you learned while exploring in Kaggle, how would you thoroughly describe these datasets to someone else?
- How do you think sharing interactive notebooks online can help you develop your data analysis skills?

Using all of the information you learned while exploring in Kaggle, how would you thoroughly describe these datasets to someone else?

I would describe it as a nice way to start programming

How do you think sharing interactive notebooks online can help you develop your data analysis skills?
It will help me by creating a programming routine.

✓ **Correct**

Congratulations on completing this hands-on activity! You worked with notebooks and used different datasets in Kaggle. A strong response would include that online resources like Kaggle help data analysts accomplish many important tasks. Beyond that, consider the following:

Data analysts use a variety of resources to complete data analysis projects. For instance, an analyst could use Kaggle notebooks to host projects in a portfolio. This is important for practicing and demonstrating your skills, as well as getting feedback from more experienced data analysts on your work.