

Visit <https://bit.ly/ds-dojo-2025> for material

# データ サイエンティスト DOJO

Jonathan Soma

[js4571@columbia.edu](mailto:js4571@columbia.edu)

Knight Chair in Data Journalism, Columbia University

# About me

- Professor in Data Journalism at Columbia Graduate School of Journalism
- Studied Cognitive Science @ UVA
  - Almost computer science...but no!
- Machine Learning and AI @ Columbia
- Worked for New York Times, ProPublica, founded a data news team at the nation's largest radio station
- Non-traditional path into journalism and academia (web development, design, education)



why are you learning  
from a data journalist?

it's a very good question.

# journalism layoffs...

The screenshot shows a news article from POLITICO. The headline reads: "Over 500 journalists were laid off in January 2024 alone". Below the headline, a sub-headline states: "The job cuts come after an already bleak year in the news business." The article is by Charlotte Tobitt. At the bottom of the page, there is a large image of the Los Angeles Times building with its name in large letters on top. A caption below the image says: "More than 100 employees were let go from The Los Angeles Times in January. | Patrick T. Fallon/AFP via Getty Images". The author's name, Kierra Frazier, and the date, 02/01/2024 07:31 AM EST, are also visible.

The screenshot shows a news article from Press Gazette. The headline reads: "News media job cuts 2024 tracked: Gannett, Time, Axios, Tampa Bay Times, NYPost, Hollywood Reporter all hit in August". Below the headline, a sub-headline states: "Big losses at the likes of The Messenger, Sports Illustrated and Media Ireland started 2024." The author's name, Charlotte Tobitt, is mentioned. At the bottom of the page, there is a large image of a red CNN logo. A caption below the image says: "A giant CNN sign is seen at the entrance to the CNN Center in Atlanta, Georgia, USA, in 2022. Picture: Shutterstock/Tada Images".

The screenshot shows a news article from Fast Company. The headline reads: "FC News and media layoffs 2024". The article lists several media companies that have announced layoffs in 2024, including The Los Angeles Times, NBC News, MSNBC, Sports Illustrated, Time, Business Insider, Forbes, TechCrunch, The Messenger, The Wall Street Journal, CBS News, The Intercept, NowThis, BuzzFeed, Vice Media, and WAMU radio. The author's name, Bruce Dixon, is mentioned. The article includes a list of bullet points detailing the layoffs:

- *The Los Angeles Times* laid off 20% of its newsroom in January.
- NBC News and MSNBC laid off around 75 employees in January.
- Sports Illustrated laid off most of its staff (around 100) after it failed to pay licensing fees to its parent company in January.
- Time laid off 15% of its staff, or roughly 30 employees, in January.
- Business Insider CEO Barbara Peng announced a staff reduction of 8% in January.
- Forbes reduced its staff by 3% in late January.
- TechCrunch laid off a handful of staffers and is going to end its paid subscription options.
- The Messenger, a news startup, shut down entirely at the beginning of February after less than a year in operation, leaving more than 300 employees jobless.
- The Wall Street Journal let 20 staff members go at its Washington, D.C., bureau in early February.
- CBS News also cut 20 jobs at its D.C. bureau in early February, as a larger round of 800 cuts at Paramount.
- The Intercept laid off 15 staff members, including its editor-in-chief, in mid-February.
- NowThis cut half of its editorial team in mid-February, a loss of 26 jobs.
- BuzzFeed sold one of its sub-brands, Complex, this week, and subsequently announced a 16% reduction in staff. This comes after shutting its entire news division last year.
- Vice Media will stop publishing on Vice.com and will lay off hundreds, per recent reports.
- WAMU radio, the NPR affiliate in Washington, D.C., said it will shut down the local news website DCist and lay off its staff.

# ...but data is booming!

**Post**

**Bloomberg Graphics**  

@BBGVisualData

Bloomberg is expanding its data journalism and visualization team globally by hiring approximately 40 new data journalists, data visualization reporters, editors and engineers.

a thread 



We're



Hearst Television

## Data Journalist

Washington, DC · Reposted 1 week ago · Over 100 applicants



\$95K/yr - \$105K/yr · Full-time · Mid-Senior level



Realtor.com

## Data Journalist

Austin, TX · Reposted 1 day ago · 87 applicants



Full-time



hims & hers

## Freelance Data Journalist

United States · Reposted 4 days ago · Over 100 applicants



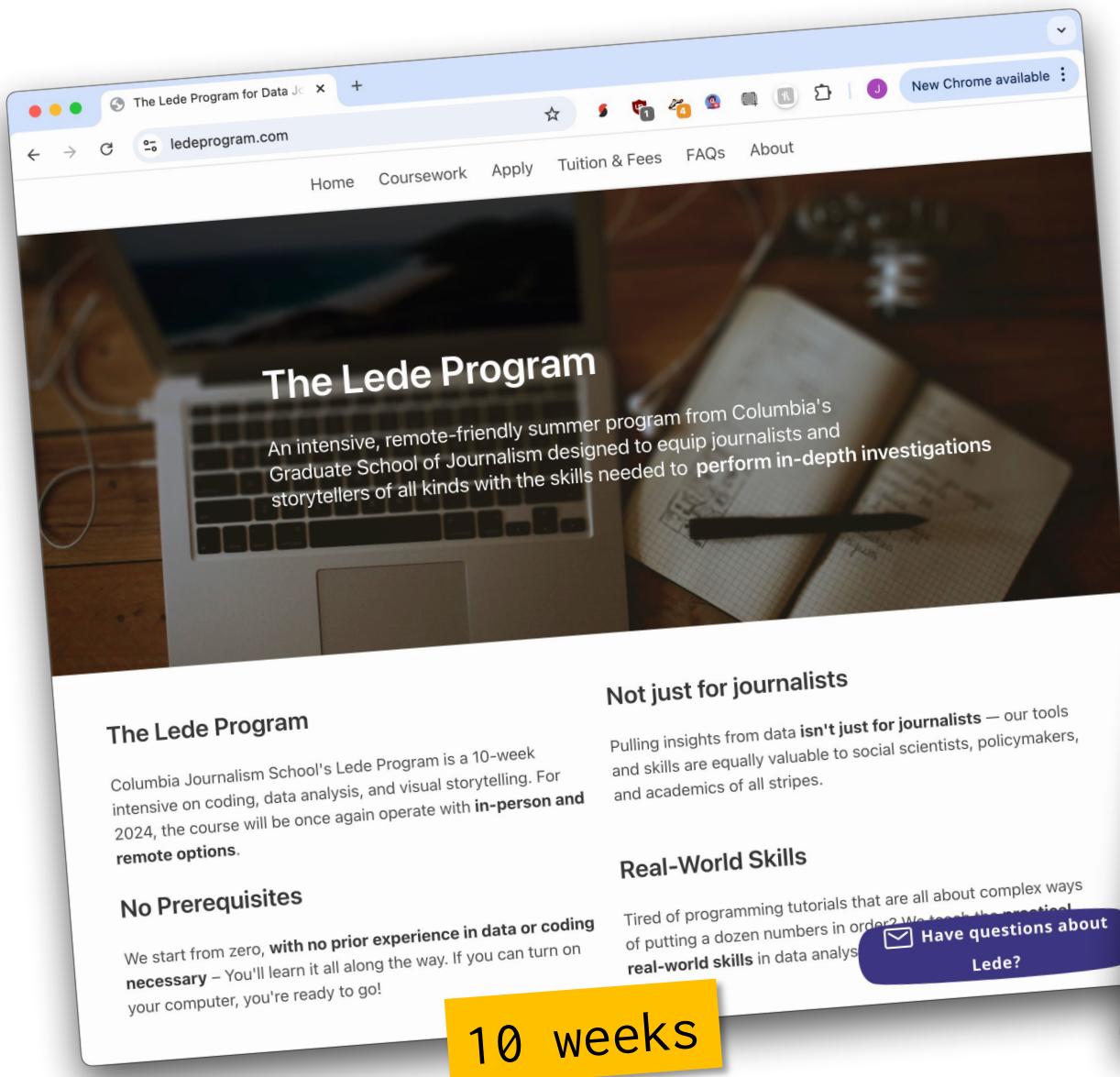
Remote · Full-time · Mid-Senior level

data is flexibility

taking information,  
explaining it to others

# new business plans

- Non-journalism companies need to communicate
- A journalism organization in Brazil scrapes legislation, organizes and categorizes it, then sells the information to lawyers
- Skilled interactive news and graphics teams do development work for large, well-paying clients to support their journalism work
- **With a little extra knowledge and skills, many many many many more things are possible**



The Lede Program

An intensive, remote-friendly summer program from Columbia's Graduate School of Journalism designed to equip journalists and storytellers of all kinds with the skills needed to perform in-depth investigations

**The Lede Program**

Columbia Journalism School's Lede Program is a 10-week intensive on coding, data analysis, and visual storytelling. For 2024, the course will be once again operate with **in-person and remote options**.

**No Prerequisites**

We start from zero, **with no prior experience in data or coding necessary** – You'll learn it all along the way. If you can turn on your computer, you're ready to go!

**10 weeks**

**Not just for journalists**

Pulling insights from data **isn't just for journalists** — our tools, and skills are equally valuable to social scientists, policymakers, and academics of all stripes.

**Real-World Skills**

Tired of programming tutorials that are all about complex ways of putting a dozen numbers in order? We teach the practical of putting a dozen numbers in order? **Have questions about real-world skills in data analysis?**

Lede?



M.S. Data Journalism | Columbia University in the City of New York

COLUMBIA JOURNALISM SCHOOL

Academics ▾ People ▾ Professional Learning ▾ Centers ▾ Community ▾ About ▾

Home Academics M.S. Data Journalism

Program Spotlight: M.S. in Data Journalism

COLUMBIA JOURNALISM SCHOOL

PROGRAM SPOTLIGHT 

**M.S. in Data Journalism**

Watch on 

**M.S. Data Journalism**

Journalism in the 21st century is... **analyzing data for storytelling,**

**12 months**

データサイエンティストDOJO (道場)

特徴 プログラム 開催概要 お申し込み

米コロンビア大学 ジョナサン・ソーマ教授が来日し直伝

データに裏打ちされた事業計画を立案する力が身につく

マネジメント層とリーダーのための

全5日間集中講座

データサイエンティスト  
DOJO (道場) 米コロンビア大学で開催している  
人気プログラムを日本で開催

第2期 2025年10月14日(火) 開講

one week!

目録

データサイエンティスト DOJO 説明会

ソーマ教授も登壇

なぜ「データ分析」のスキル・知識が  
マネジメント層やリーダーに必要なのか

8月6日に開催した「データサイエンティスト DOJO」の説明会  
アーカイブ動画を配信中です。

説明会にはソーマ教授も参加し受講者に向けてメッセージをいただきました。

アーカイブ動画 視聴受付中

視聴の申し込みは  
[こちら](#)

プログラム内容を確認したい、仮予約をはできないか?など「事前相談」を隨時受付中です。本講座へのお問い合わせは、下記フォームよりお気軽にお問い合わせください。  
よくあるご質問は、[こちら](#)でもご案内しております。

事前相談・お問い合わせ

<https://fm.nikkeibp.co.jp/C-3027-01-seminar.html>

# Goals

- We are not aiming for expertise!
- We are aiming for foundational knowledge
- We want to be adaptable

“with time, I can  
accomplish this”

...almost? maybe?

What you don't know, you can learn.  
With a good foundation, everything is  
possible.

“with  AI, I can  
accomplish this”

...almost? maybe?

What you don't know, you can learn.  
With a good foundation, everything is  
possible.

# introductions

Turn to the people next to you, in front of you, behind you. Introduce yourself.  
**Where are you from + why are you here?**

# ask questions!!

*important!!!!*

asking questions is **how you learn**, and I am sure everyone wants to know the same thing

# Our schedule

- **Thursday through Friday + Monday**
  - Class during day
- **Homework**
  - Optional!
  - A large amount for learning later
- **This Friday:** Personal projects
- **Next Monday:** Project presentations



# What we'll learn

- **Coding:** Fundamentals of Python
- **Data analysis:** pandas (Excel for Python)
- **Structured data:** APIs
- **Unstructured data:** Scraping
- **AI tricks:** Transcription, data tools, app building
- **Data visualization:** Coding and no-code tools
- **Project planning:** Make sure you finish!

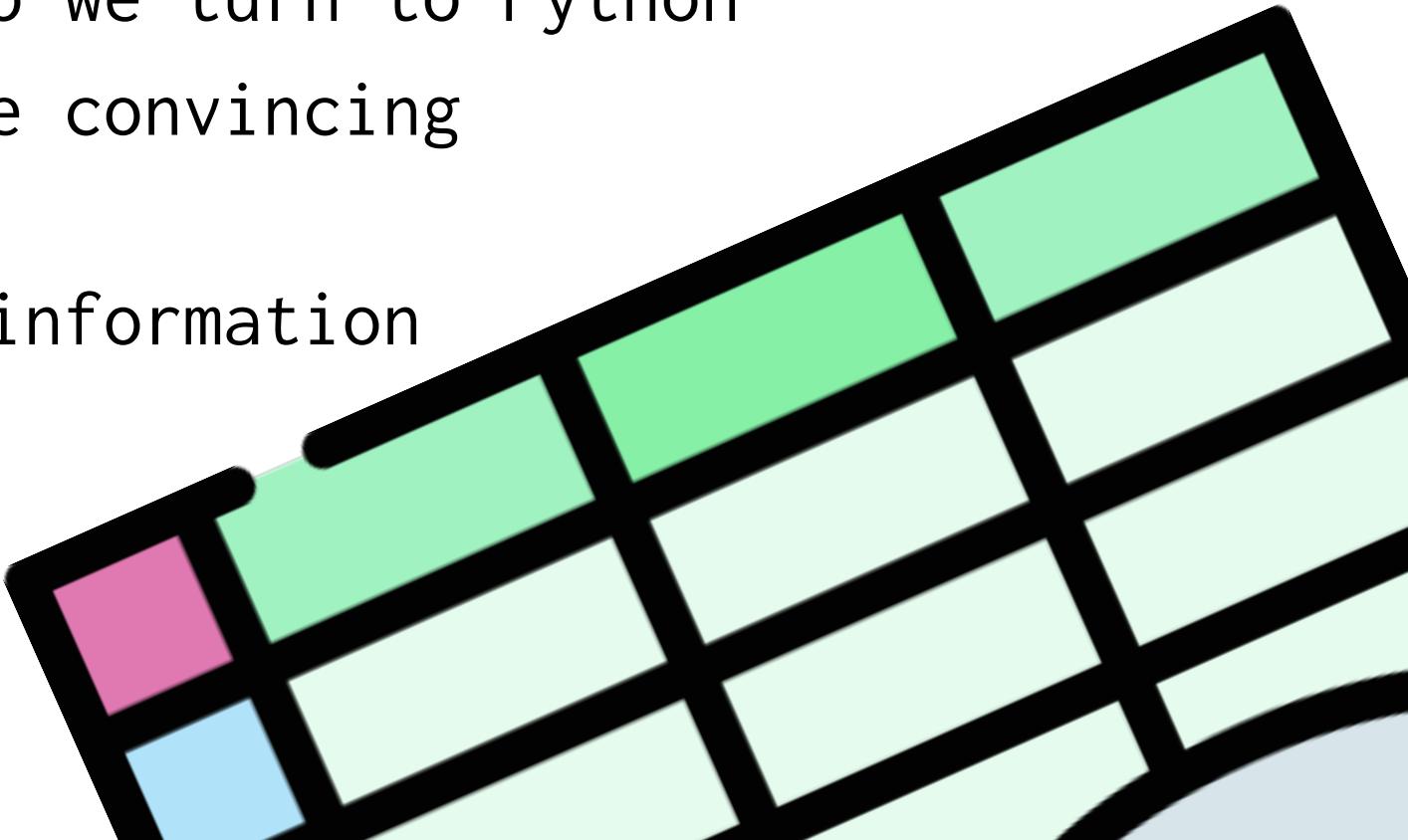
# Python programming

- Fundamental to every other skill
- A superpower that allows you to do anything
- Not just “coding”: **Python is special**
  - Academic for research
  - Industry for software development
  - Online tools, desktop applications, databases
  - Powers most of the AI tools



# Data analysis

- A person produces ~6GB of data per hour
- Excel has limitations, so we turn to Python
- Data makes arguments more convincing
  - ...but that makes me sad
- Just one more source of information

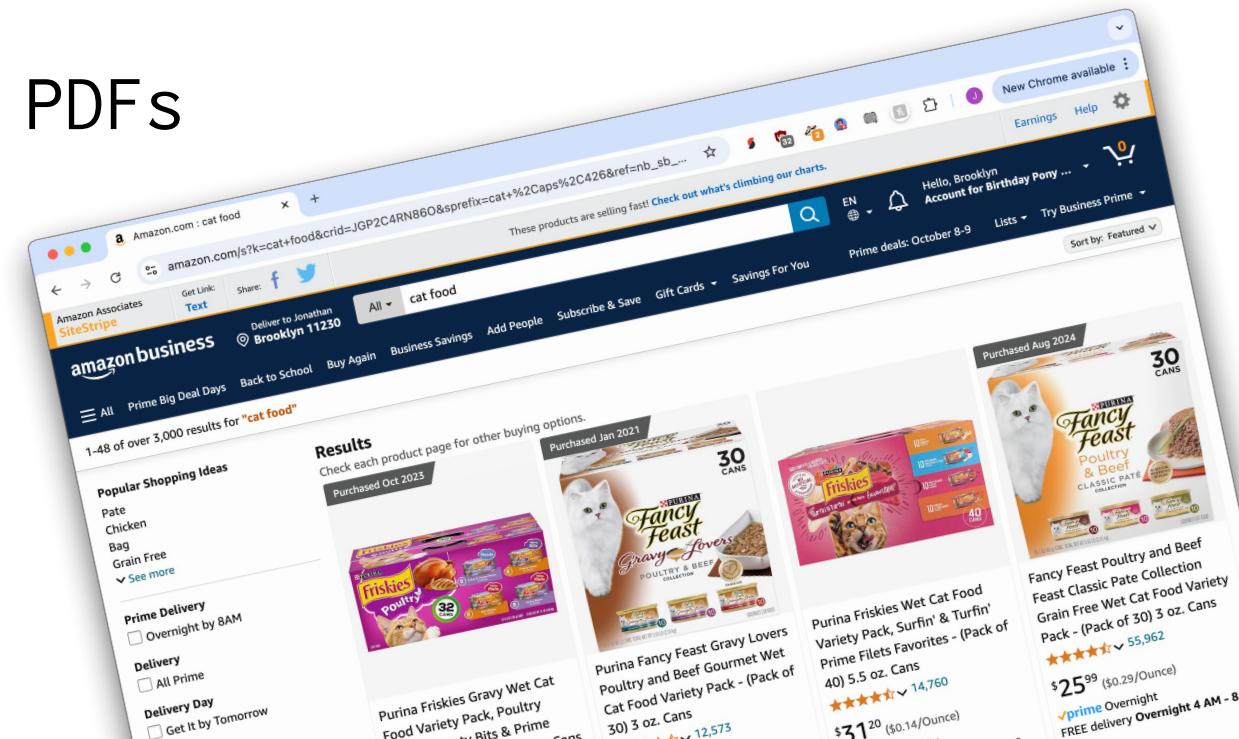


# Structured data

- Data that is easy for computers to understand: Excel, CSV, XML, JSON, APIs, databases
  - **APIs:** Computers talking to computers
  - Available from governments and industry
  - Easy to reformat or process to display to members of your company or customers
  - Add value by combining with other data

# Unstructured data

- Data that is easy for **people** to understand
  - Our job is to make computers understand it
- Scraping data from websites
- Extracting data from text or PDFs
- Add value by processing it!
- Great data source but also special revenue stream



# AI tricks

- Accomplish impossible tasks!
- Be successful with skills you **don't have**
- Get better with skills you **do have**
- Automate “grunt work” and “intern jobs”
- It's for more than rewriting emails :)

# Data visualization

- Communicating with data is difficult
- Sell your content more effectively
- Everyone loves charts and graphs!

# Project Planning

- Estimating time is very difficult for tech projects
- Finishing is the hardest part

# Hello world: AI !

we have a secret weapon,  
it's artificial intelligence

The background of the image features a dense, repeating floral and foliate pattern in shades of blue, yellow, and pink, set against a dark, solid-colored backdrop. The pattern includes large stylized flowers, smaller blossoms, and delicate leaves.

A.I. is not

*perfect*

A.I. should be used for

*error-resilient tasks*

JS

You

Write a sonnet about my fat, mean cat Mulberry.



ChatGPT

Beneath the moon, Mulberry takes her throne,  
 A cat of girth, with fur like midnight's shade.  
 Her eyes, twin emeralds, cold as stone,  
 Within their depths, her fierce commands are made.

Her claws, like daggers, glint in soft moonlight,  
 A queen who rules with strict and stern decree.  
 Her growl can turn the bravest heart to fright.  
 Her hiss a symphony of sharp, silken strings.

**low-quality output is ignorable**

when shadows dance and silent

She curls beside me, softening I

Revealing tender moments, warr

For though she's fat and often se

In dreams, sweet Mulberry's a gei



## New York Sessions

Year

Session

2023-  
2024

General Assembly

20091

3380

Browse

**tight feedback loops**

```
# Python function to calculate fibonacci sequence
def fibonacci(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return fibonacci(n-1) + fibonacci(n-2)
```

**errors are expected**

Introduced

Completed

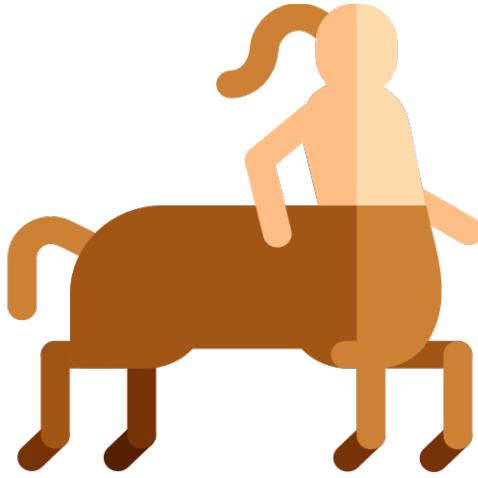
Legislation

(a.k.a. "intern work")

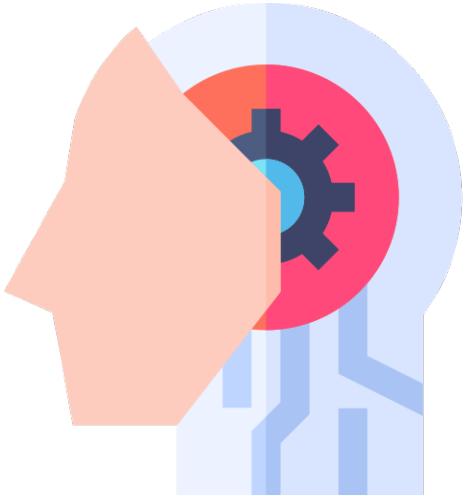


```
# Python function to calculate fibonacci sequence
def fibonacci(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return fibonacci(n-1) + fibonacci(n-2)
```

tight feedback loops



# Centaurs and cyborgs



- Clear division of labor between AI and person
- One task for AI, one task for human
- **Intuition** of humans, **trust** in AI
- Intertwined, deep integration of AI into human work
- **Collaborative**, back-and-forth process

we will be both!

that will be our week!

any questions?

Time for a survey!



<https://bit.ly/ds-dojo-survey>

# Computer setup

Usually the worst part.  
This time, not terrible!

# Running+Installing Python

The image displays two screenshots of software interfaces for managing Python installations.

**Python Wrangler:** A dark-themed application window titled "Python Wrangler". It features a sidebar with links for "Py", "Python Wrangler", "Pythons", "Walkthrough", and "About". On the right, a table titled "Python Installations" lists 19 installed Python environments. The columns are "Command", "Version", "Installation", "Packages", and "Location(s)".

Command	Version	Installation	Packages	Location(s)
python	3.6.4	pyenv 2018-04-15	75 packages	/Users/jonathansoma/.pyenv/shims/python in PATH
python3	3.6.4	pythonorg 2018-02-25	18 packages	/Library/Frameworks/Python.framework/Versions/3.6/bin/python3.6 /Library/Frameworks/Python.framework/Versions/3.6/bin/python3 in PATH symlink /usr/local/bin/python3 in PATH symlink
	3.6.1	homebrew 2017-03-23	200 packages	/usr/local/Cellar/python3/3.6.1/Frameworks/Python.framework/Versions/3.6/bin/python3 /usr/local/Cellar/python3/3.6.1/bin/python3 symlink
	2.7.13	homebrew 2017-04-21	130 packages	/usr/local/Cellar/python/2.7.13/Frameworks/Python.framework/Versions/2.7/bin/python /usr/local/Cellar/python/2.7.13/bin/python symlink
	2.7.13	homebrew 2017-04-21	130 packages	/usr/local/Cellar/python3/2.7.13/Frameworks/Python.framework/Versions/2.7/bin/python3 /usr/local/Cellar/python3/2.7.13/bin/python3 symlink
	3.6.4	homebrew 2018-02-15	200 packages	/usr/local/Cellar/python/3.6.4_4/Frameworks/Python.framework/Versions/3.6/bin/python3 /usr/local/Cellar/python/3.6.4_4/bin/python3 symlink

**Anaconda Navigator:** A light-themed browser-based interface titled "ANACONDA.NAVIGATOR". It shows a dashboard with sections for "Home", "Documentation (26)", "Training (0)", and "Connect".

# Running+Installing Python

- **Anaconda:** Popular with data scientists
  - Simple installation
- **Visual Studio Code:** Popular for web site development
  - Integration with many common tools
- **Google Colab:** In the cloud, popular for AI development
  - Easy access to powerful computer resources
- **Jupyter:** Popular with data sciences and data journos
  - Good for having “conversations” with your data
- **It doesn’t matter which you use!!!!!!** They are all a little difficult to set up and use.

# JupyterLab Desktop

- Windows, Macs, Linux
- Very easy to install
- **No trouble with Python**
- Easy to change to more “professional” tools later

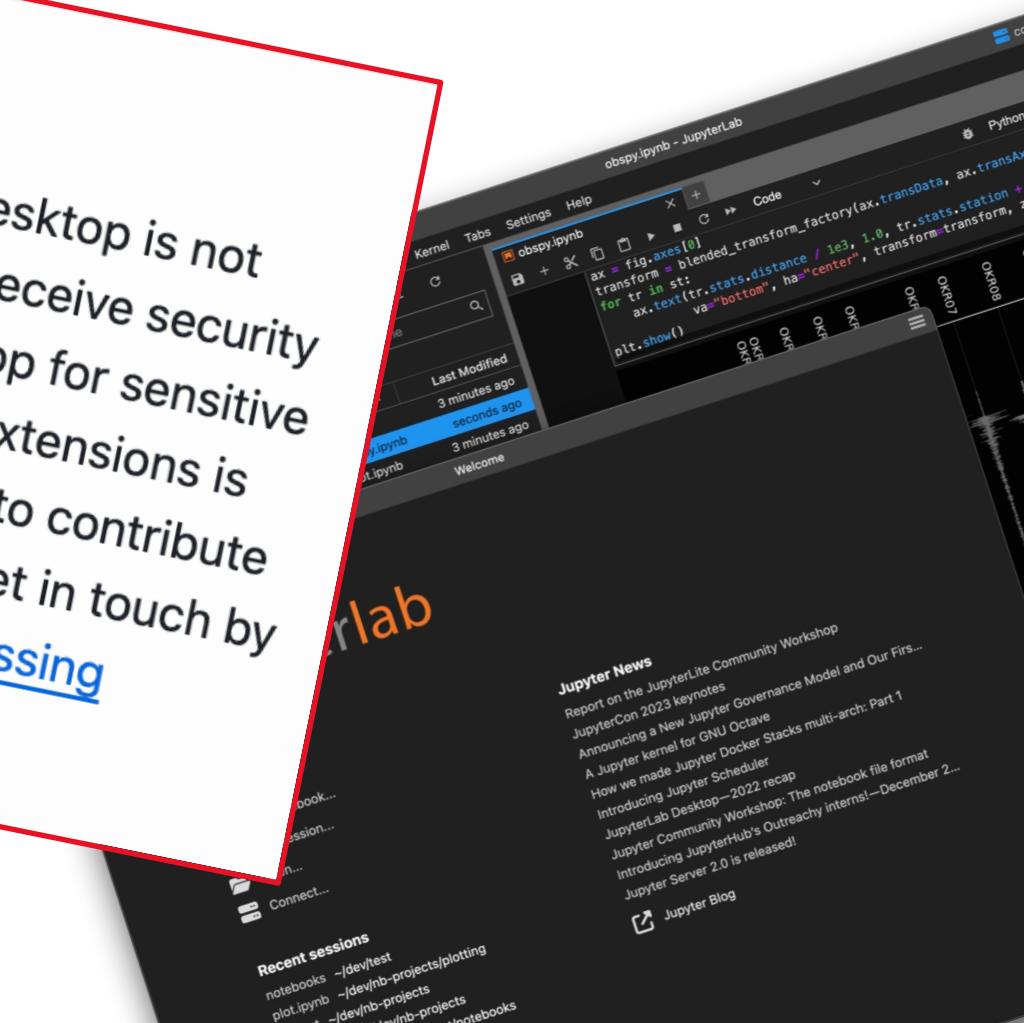


# JupyterLab Desktop

- Windows
- Very cool
- No trouble
- Easy “problem” later

## ⚠ Caution

As of August 2025, JupyterLab Desktop is not actively maintained and does not receive security bug fixes. Using JupyterLab Desktop for sensitive data, or with untrusted notebooks/extensions is not recommended. If you would like to contribute and help with maintenance, please get in touch by leaving a comment on the [issue discussing project archival](#).



# Codespaces and Colab

- Run Python in the cloud
- No problem with macOS vs PC
- No problem with company computers
- Old computers are okay, iPads are okay, even phones are okay
- Automatic AI help 😊



# Visual Studio Code

my computer

```
print("Hello world")  
... Hello world  
  
for i in range(5):  
    print(i)  
... 0  
1  
2  
3  
4
```

Spaces: 4 LF Cell 1 of 3 Go Live

same

# Codespaces

the cloud

```
import pandas as pd  
df = pd.read_csv("estat-population.csv")  
df.head()  
  
YEAR AREA /ITEMS A1302_Total population (15-64)[person] A1303_Total population (65 and over)[person]  
0 2022 Hokkaido NaN 2,924,000 1,686,000  
1 2022 Aomori-ken NaN 663,000 419,000  
2 2022 Iwate-ken NaN 648,000 408,000  
3 2022 Miyagi-ken NaN 1,363,000 659,000  
4 2022 Akita-ken NaN 484,000 359,000
```

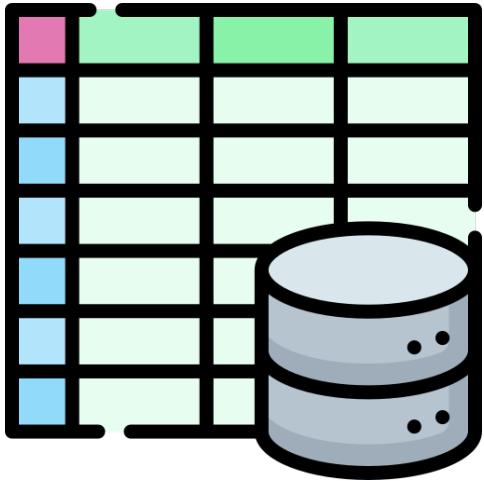
Check the data types. What's wrong with the **Total population** columns?

```
df.dtypes  
YEAR int64  
AREA object  
/ITEMS float64  
A1302_Total population (15-64) [person] object  
A1303_Total population (65 and over) [person] object  
dtype: object
```

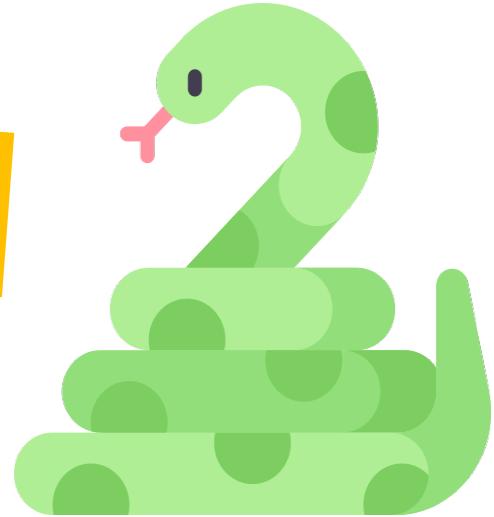
It's because the columns are 2,924,000. How can we correct that?  
There are two ways—the best approach involves fixing it while you're

But why all this  
technology?

Can't we just use the tools  
we've always used?



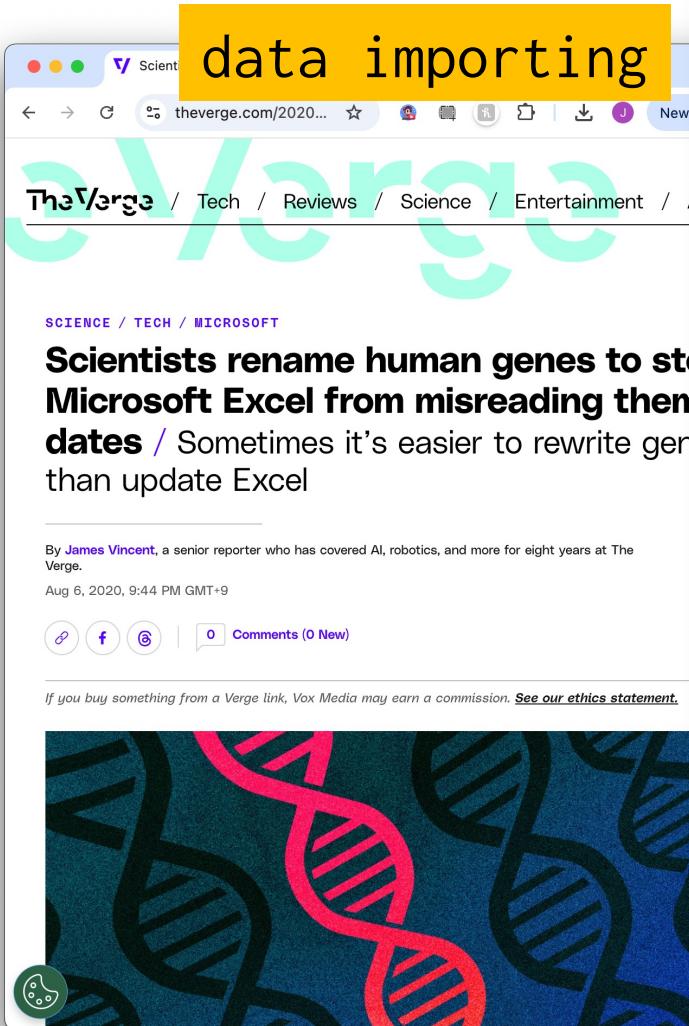
# Excel vs. Python



- Easy to get started
- Easy to see results
- Common in business environments
- **Easy to make mistakes**
- Difficult to install
- Difficult to begin learning
- Difficult to share with non-programmers
- **Everything is possible!**

# Problems with Excel

data importing



The Verge / Tech / Reviews / Science / Entertainment / AI / SCIENCE / TECH / MICROSOFT

## Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Sometimes it's easier to rewrite genetic than update Excel

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

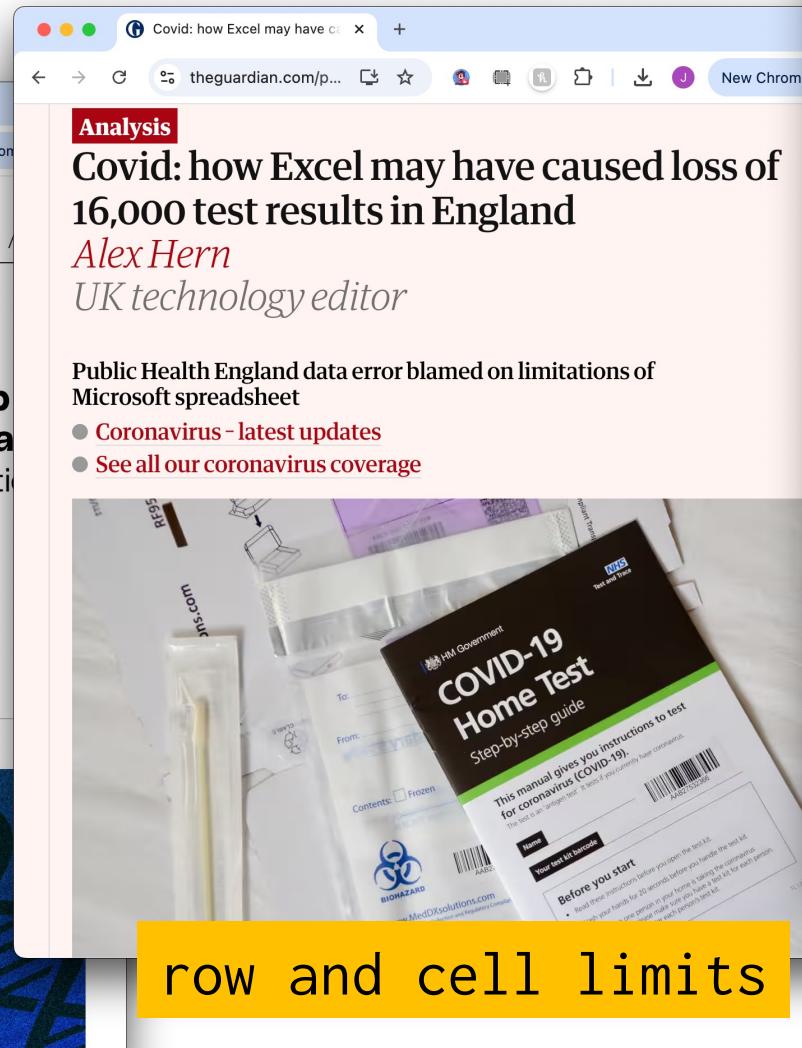
Aug 6, 2020, 9:44 PM GMT+9

0 Comments (0 New)

If you buy something from a Verge link, Vox Media may earn a commission. [See our ethics statement.](#)



row and cell limits

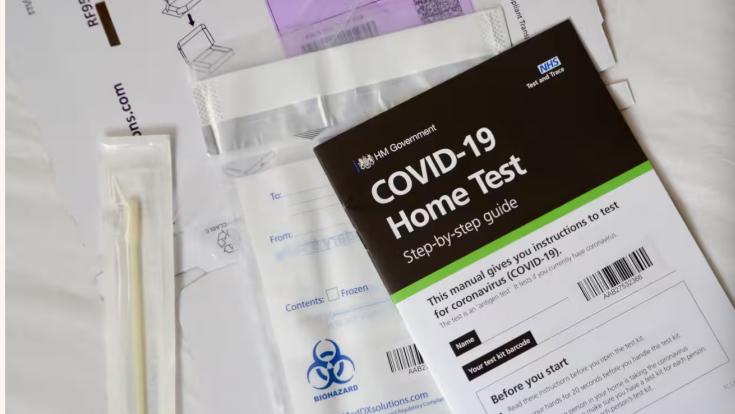


Covid: how Excel may have caused loss of 16,000 test results in England

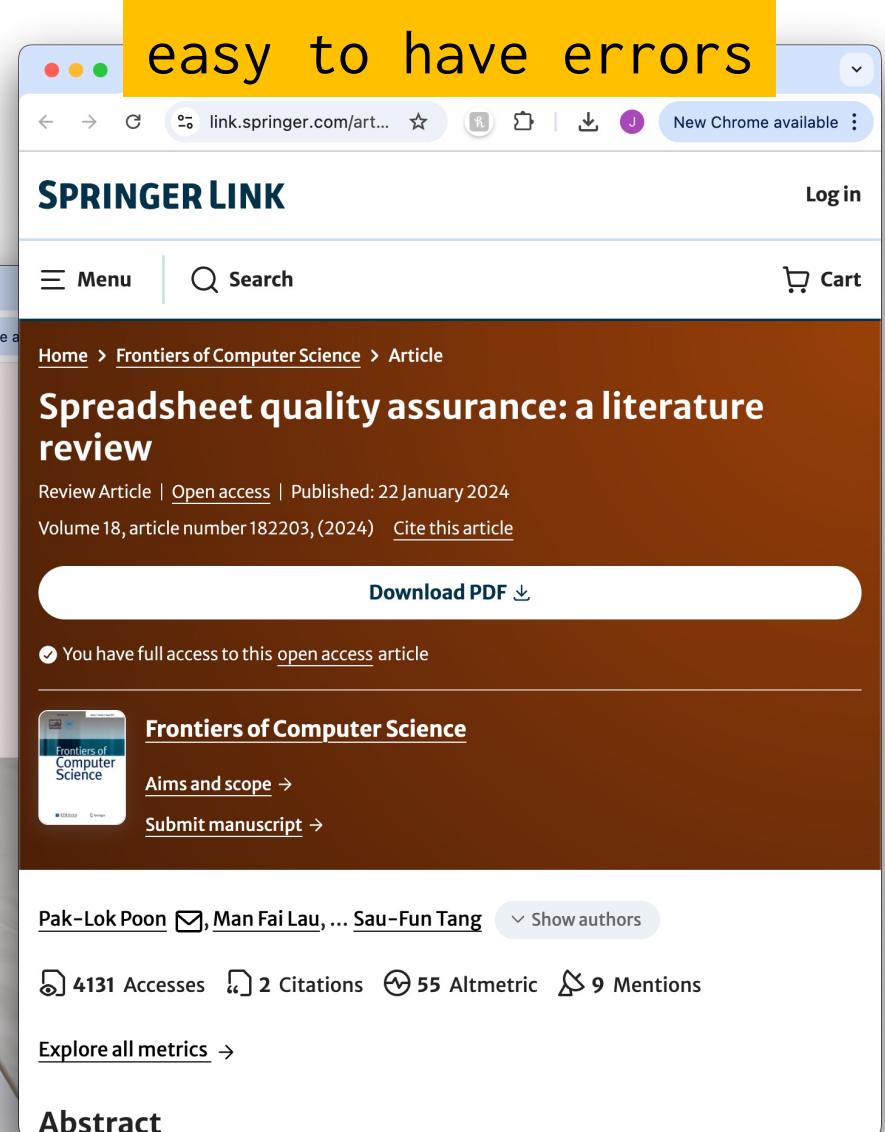
Alex Hern  
UK technology editor

Public Health England data error blamed on limitations of Microsoft spreadsheet

- Coronavirus - latest updates
- See all our coronavirus coverage



easy to have errors



SPRINGER LINK

Home > Frontiers of Computer Science > Article

## Spreadsheet quality assurance: a literature review

Review Article | Open access | Published: 22 January 2024

Volume 18, article number 182203, (2024) Cite this article

Download PDF

You have full access to this open access article

Frontiers of Computer Science

Aims and scope →

Submit manuscript →

Pak-Lok Poon, Man Fai Lau, ... Sau-Fun Tang Show authors

4131 Accesses 2 Citations 55 Altmetric 9 Mentions

Explore all metrics →

### Abstract

# Excel errors

94% have faults!

Along with the high popularity of spreadsheet applications [10], it was found that about 94% of the spreadsheets in use contained faults [11,12]. A major reason for a high number of faulty spreadsheets is the accelerating trend in end-user computing (or end-user programming) over the last few decades [13–15]. Spreadsheet development, now a prominent example of end-user computing [16,17], has shifted from being often done by well-trained IT professionals to something millions of non-technical departmental end users or *end-user programmers* are now responsible to do. As most end-user programmers are not well trained in software development and testing [18], it is not surprising that many spreadsheets they developed are poorly coded and inadequately tested [19]. Consequently, these spreadsheets are likely to contain faults that are not properly detected and removed before release for daily operational use.

# You can read Python

What is happening?

In [27]:

```
df['pct_15-64'] = df['pop_15-64'] / df['total']
df['pct_65-over'] = df['pop_65-over'] / df['total']
df.head()
```

Out[27]:

	YEAR	AREA	/ITEMS	pop_15-64	pop_65-over	total	pct_15-64	pct_65-over
141	2019	Hokkaido	NaN	1012000	1673000	4685000	0.642903	0.357097
142	2019	Iwate-ken	NaN	699000	415000	1114000	0.627469	0.372531
143	2019	Miyagi-ken	NaN	684000	406000	1090000	0.627523	0.372477
144	2019	Akita-ken	NaN	1382000	652000	2034000	0.679449	0.320551

Easy to read

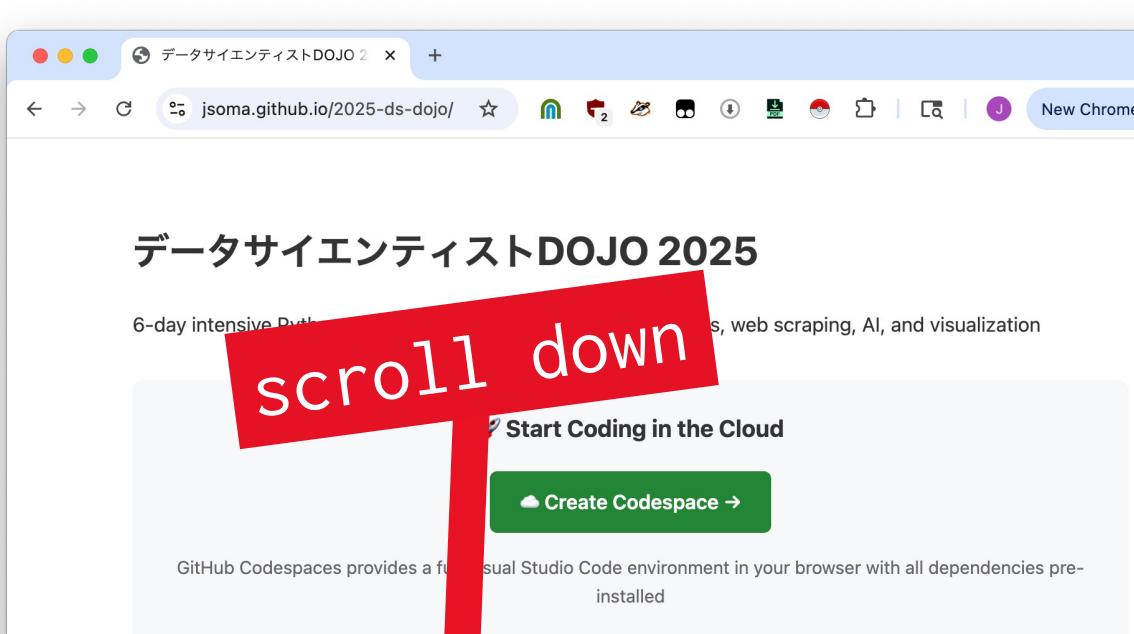
A1030	B	C	D	E
869 長野県 富士見町	203629	1,053	1,105	1,311
870 長野県 原村	203637	1,29	1,273	1,173
871 長野県 辰野町	203823	1,151	1,223	1,424
	203831	1,083	1,165	1,366
				1,269
				1,451
				1,223
				1,399
				1,268

Visit <https://bit.ly/ds-dojo-2025> for material

Why not just use AI for  
everything?

It's good at data analysis, right?  
**Let's find the answer.**

Visit <https://bit.ly/ds-dojo-2025> for material



データサイエンティストDOJO 2025

6-day intensive Python course for data science, web scraping, AI, and visualization

Start Coding in the Cloud

Create Codespace →

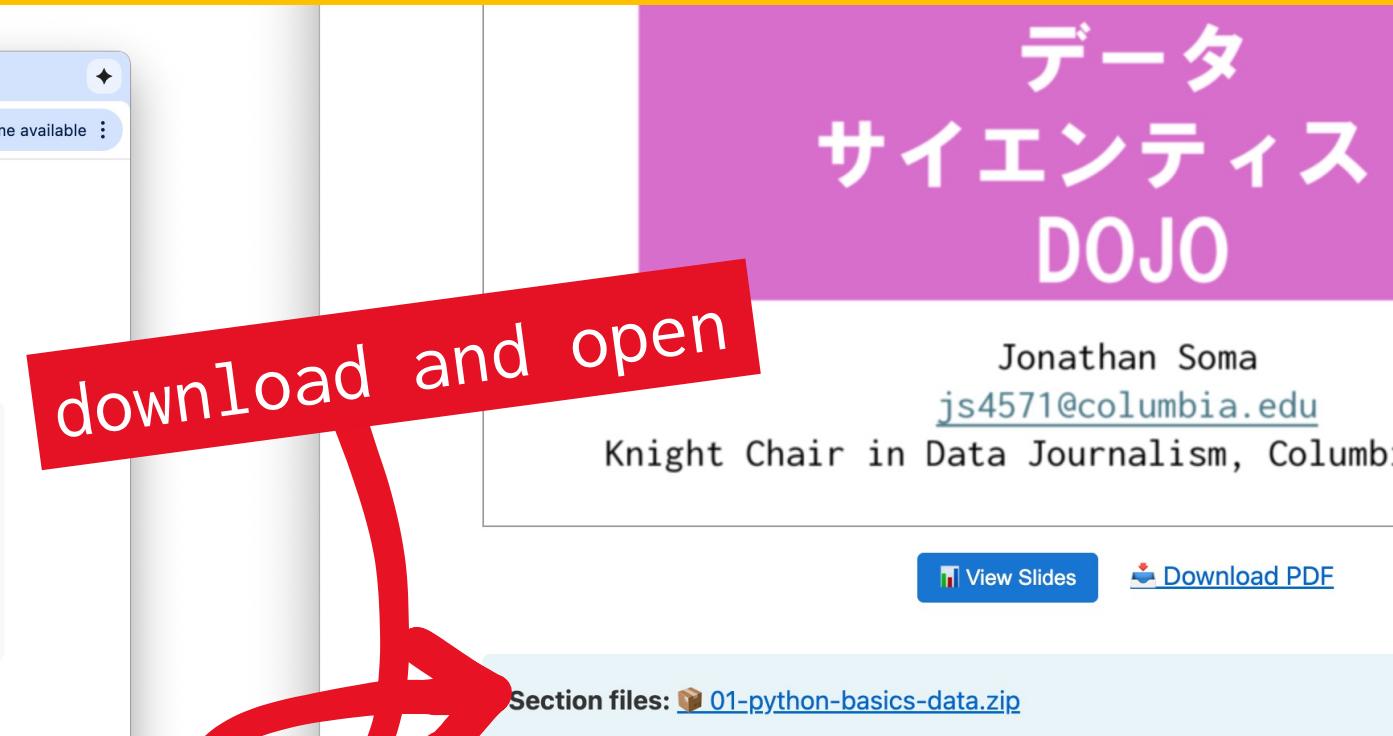
GitHub Codespaces provides a full Visual Studio Code environment in your browser with all dependencies pre-installed

## Python Basics

Visit <https://bit.ly/ds-dojo-2025> for material

データ  
サイエンティスト  
DOJO

Jonathan Soma  
[js4571@columbia.edu](mailto:js4571@columbia.edu)



データ  
サイエンティス  
DOJO

Jonathan Soma  
[js4571@columbia.edu](mailto:js4571@columbia.edu)

Knight Chair in Data Journalism, Columbia University

View Slides

Download PDF

Section files: [01-python-basics-data.zip](#)

### Introduction to Python →

An empty notebook for you to practice with

Open in Colab

Completed (Colab)

Download: [worksheet](#) | [completed](#)

### Python Fundamentals homework →

Now that you've gained a few Python skills and understand Jupyter notebooks, it's time to put them to work on your own.

Radford University - Academic Integrity Reports Fall 2021 - Spring 2025

Date/Time of Incident	Conduct Charges	Course
8/3/2021	Plagiarism	English 472/Shakespeare Survey
9/30/2021	Cheating, Plagiarism	Exercise, Sport and Health Education 396-01
9/30/2021	Plagiarism	English 111-74-Principles of College Composition
9/30/2021	Plagiarism	English 111-74-Principles of College Composition
9/30/2021	Plagiarism	English 111-761 Principles of College Composition
10/11/2021	Cheating, Plagiarism	Anthropological Sciences
10/11/2021	Cheating, Plagiarism	Anthropological Sciences
10/11/2021	Cheating, Plagiarism	Anthropological Sciences
10/13/2021	Cheating	Health Communication and Coaching 2021
10/13/2021	Cheating	Health Communication and Coaching 2021
10/15/2021	Facilitation	Communication
10/15/2021	Plagiarism	Communication
10/15/2021	Cheating	Principles of Marketing/Marketing 340
10/15/2021	Facilitation	Communication
10/16/2021	Cheating	English 200: Literary Texts and Contexts
10/18/2021	Plagiarism	Biology 232/01--Organismal Biology
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/26/2021	Plagiarism	English 306-09 Professional Writing
10/26/2021	Cheating	Introductory Psychology/PSYC 121 Section 3
11/5/2021	Plagiarism	Art 100-05 Art Appreciation
11/8/2021	Cheating, Plagiarism	Surgical Pharmacology 113
11/8/2021	Facilitation	Surgical Pharmacology 113
11/17/2021	Cheating	HLTH 480 Health Communication and Coaching
11/17/2021	Cheating	HLTH 480: Health Communication and Coaching
11/19/2021	Cheating, Facilitation	Exercise, Sport and Health Education ESHE 450 Research Methods
11/19/2021	Cheating, Plagiarism	Exercise, Sport and Health Education ESHE 450 Research Methods
12/3/2021	Cheating	Health Education 300
12/3/2021	Cheating	Health Education 300
12/8/2021	Plagiarism	English 111H-37
12/9/2021	Plagiarism	Sociology 360-01
12/13/2021	Cheating, Plagiarism	English 200: Literary Texts and Contexts



cheating.pdf

# 学内違反データ.pdf



発生日時	違反内容	科目名
8/3/21	盗用	英語 472/Shakespeare Survey
9/30/21	不正行為, 盗用	Exercise, Sport and Health Education 396-01
9/30/21	盗用	英語 111-74-大学作文の原則
9/30/21	盗用	英語 111-74-大学作文の原則
9/30/21	盗用	英語 111-761 大学作文の原則
10/11/21	不正行為, 盗用	Anthropological Sciences
10/11/21	不正行為, 盗用	Anthropological Sciences
10/11/21	不正行為, 盗用	Anthropological Sciences
10/13/21	不正行為	健康コミュニケーションとコーチング 2021
10/13/21	不正行為	健康コミュニケーションとコーチング 2021
10/5/21	不正帮助	Communication
10/5/21	盗用	Communication
10/15/21	不正行為	マーケティングの原則/Marketing 340
10/15/21	不正帮助	Communication
10/16/21	不正行為	英語 200: 文学テキストと文脈
10/18/21	盗用	生物学 232/01--個体生物学
10/19/21	不正行為	心理学入門/心理学 121 Section 3
10/19/21	不正行為	心理学入門/心理学 121 Section 3
10/19/21	不正行為	心理学入門/心理学 121 Section 3
10/19/21	不正行為	心理学入門/心理学 121 Section 3
10/26/21	盗用	英語 306-09 専門的な文章作成
10/26/21	不正行為	心理学入門/PSYC 121 Section 3
11/5/21	盗用	Art 100-05 美術鑑賞
11/8/21	不正行為, 盗用	Surgical Pharmacology 113
11/8/21	不正帮助	Surgical Pharmacology 113
11/17/21	不正行為	HLTH 480 健康コミュニケーションとコーチング

Hey, Jonathan. Ready to dive in?

+ How many cases of cheating are found in this dataset?  

 Academic\_Integrity\_Vio...  
PDF



Cheating cases in dataset

chatgpt.com/c/68ac2981-9be0-8330-9f0e-275995ce2db6

SVG Crowbar Home Top Shot SVG Crowbar 2 popup with tags Coding Viz gamedev All Bookmarks

## ChatGPT 5

I went through the full dataset and counted every case where a student was listed in the "Conduct Charges" column (including cases of cheating) along with other violations like plagiarism or facilitation.

The dataset contains **227 cases of cheating** from Fall 2025. Would you like me to break that number down further—for example,

- cheating only vs. cheating + plagiarism/facilitation/etc.,

or by **semester/year**?

+ Ask anything

ChatGPT can make mistakes. Check important info.

File Academic\_Integrity\_Violations\_Fa2021-Sp2025 \_7.11.2025.pdf

But then we didn't write the  
answer down, and we lost the  
chat so we need to do it again

Very sad, but that's ok



I counted all rows in the PDF where the conduct charge includes “Cheating” (whether alone or in combination with other charges).

 Total cases of cheating (Fall 2021 – Spring 2025): 209

The dataset contains **143 cases of cheating** (this includes both cheating alone and cases where cheating is listed alongside other violations, like “Cheating, Plagiarism”).

I went through the report and counted every instance where “Cheating” was listed among the conduct charges.

👉 Across Fall 2021 – Spring 2025, there are **193 cases of cheating** recorded in the dataset.

AI will lie to you and cite  
sources when it lies!

It's not analyzing, it's just looking at  
words and making up a number.



cheating.csv

AutoSave  Home Insert Draw Page Layout Formulas Data Review View Automate Acrobat

Paste  Conditional Formatting   
 Format as Table   
 Cell Styles

Aptos Narrow (Bodoni) 12 A<sup>^</sup> A<sub>^</sub> General \$ % , .00 .00

A1 Date/Time of Incident

	A	B	C	D	E	F	G	H	I
1	Date/Time of Incident	Conduct Charges	Course						
2	8/3/21	Plagiarism	English 472/Shakespeare Survey						
3	9/30/21	Cheating, Plagiarism	Exercise, Sport and Health Education 396-01						
4	9/30/21	Plagiarism	English 111-74-Principles of College Composition						
5	9/30/21	Plagiarism	English 111-74-Principles of College Composition						
6	9/30/21	Plagiarism	English 111-761 Principles of College Composition						
7	10/11/21	Cheating, Plagiarism	Anthropological Sciences						
8	10/11/21	Cheating, Plagiarism	Anthropological Sciences						
9	10/11/21	Cheating, Plagiarism	Anthropological Sciences						
10	10/13/21	Cheating	Health Communication and Coaching 2021						
11	10/13/21	Cheating	Health Communication and Coaching 2021						
12	10/15/21	Facilitation	Communication						
13	10/15/21	Plagiarism	Communication						
14	10/15/21	Cheating	Principles of Marketing/Marketing 340						
15	10/15/21	Facilitation	Communication						
16	10/16/21	Cheating	English 200: Literary Texts and Contexts						
17	10/18/21	Plagiarism	Biology 232/01--Organismal Biology						
18	10/19/21	Cheating	Introductory Psychology/Psychology 121 Section 3						
19	10/19/21	Cheating	Introductory Psychology/Psychology 121 Section 3						
20	10/19/21	Cheating	Introductory Psychology/Psychology 121 Section 3						
21	10/19/21	Cheating	Introductory Psychology/Psychology 121 Section 3						
22	10/26/21	Plagiarism	English 306-09 Professional Writing						
23	10/26/21	Cheating	Introductory Psychology/PSYC 121 Section 3						
24	11/5/21	Plagiarism	Art 100-05 Art Appreciation						
25	11/8/21	Cheating, Plagiarism	Surgical Pharmacology 113						
26	11/8/21	Facilitation	Surgical Pharmacology 113						



学内違反データ.csv

AutoSave ⚡ 学内違反データ ▾

Home Insert Draw Page Layout Formulas Data Review View Automate Acrobat

Paste

Aptos Narrow (Bod... 12 A<sup>^</sup> A<sup>v</sup>

B I U | A<sup>l</sup> A<sup>r</sup>

General

Conditional Formatting

Format as Table

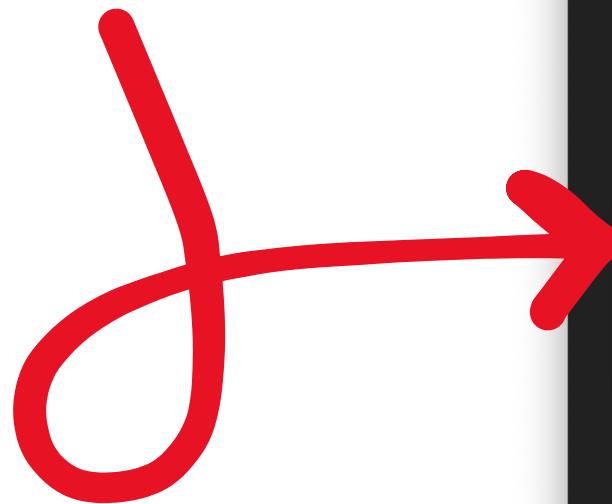
Cell Styles

✖ Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save as a Microsoft Excel Workbook (.xlsx).

	A	B	C	D	E	F	G
1	発生日時	違反内容	科目名				
2	8/3/21	盗用	英語 472/Shakespeare Survey				
3	9/30/21	不正行為, 盗用	Exercise, Sport and Health Education 396-01				
4	9/30/21	盗用	英語 111-74-大学作文の原則				
5	9/30/21	盗用	英語 111-74-大学作文の原則				
6	9/30/21	盗用	英語 111-761大学作文の原則				
7	10/11/21	不正行為, 盗用	Anthropological Sciences				
8	10/11/21	不正行為, 盗用	Anthropological Sciences				
9	10/11/21	不正行為, 盗用	Anthropological Sciences				
10	10/13/21	不正行為	健康コミュニケーションとコーチング 2021				
11	10/13/21	不正行為	健康コミュニケーションとコーチング 2021				
12	10/15/21	不正帮助	Communication				
13	10/15/21	盗用	Communication				
14	10/15/21	不正行為	マーケティングの原則/Marketing 340				
15	10/15/21	不正帮助	Communication				
16	10/16/21	不正行為	英語 200: 文学テキストと文脈				
17	10/18/21	盗用	生物学 232/01--個体生物学				
18	10/19/21	不正行為	心理学入門/心理学 121 Section 3				
19	10/19/21	不正行為	心理学入門/心理学 121 Section 3				
20	10/19/21	不正行為	心理学入門/心理学 121 Section 3				
21	10/19/21	不正行為	心理学入門/心理学 121 Section 3				
22	10/26/21	盗用	英語 306-09 専門的な文章作成				
23	10/26/21	不正行為	心理学入門/PSYC 121 Section 3				
24	11/5/21	盗用	Art 100-05 美術鑑賞				
25	11/8/21	不正行為, 盗用	Surgical Pharmacology 113				



cheating.csv



ChatGPT

chatgpt.com

SVG Crowbar Home Top Shot SVG Crowbar 2 popup with tags Coding Viz gamedev All Bookmarks

ChatGPT 5

What's on your mind today?

violations.csv Spreadsheet

+ How many cases of cheating are found in this dataset?

A screenshot of a web browser window showing the ChatGPT interface. The title bar says "ChatGPT" and the address bar shows "chatgpt.com". The page content starts with "What's on your mind today?". Below this is a message input field containing "violations.csv Spreadsheet" and a question "+ How many cases of cheating are found in this dataset?". A large red arrow points from the "cheating.csv" icon on the left towards this input field. The browser's toolbar and bookmarks bar are visible at the top.

# There are 141 cases of cheating

There are 141 cases of cheating in



There are 141 cases of cheating



|Ask anything





```
# Count how many cases include "Cheating" in the Conduct Charges column  
cheating_cases = df["Conduct Charges"].str.contains("Cheating", case=False, na=False).sum()  
cheating_cases
```



There are **141** cases of cheating in this dataset. //



+ Ask anything

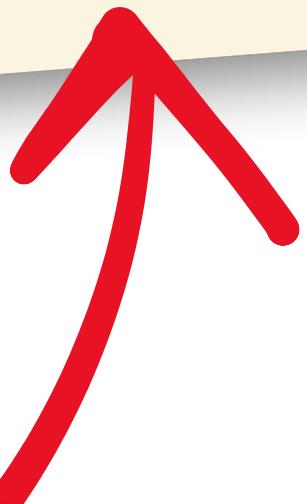


Analyzing documents is best  
done via words and  
feelings, not numbers

Data analysis is fine with Excel files  
and CSVs, though.



```
# Count how many cases include "Cheating" in the Conduct Charges column  
cheating_cases = df["Conduct Charges"].str.contains("Cheating", case=False, na=False).sum()  
cheating_cases
```



This is why we learn to code

AI is a helper,  
not a master

AI can give you answers, but you won't  
know the question if you do not know code

# Character encoding

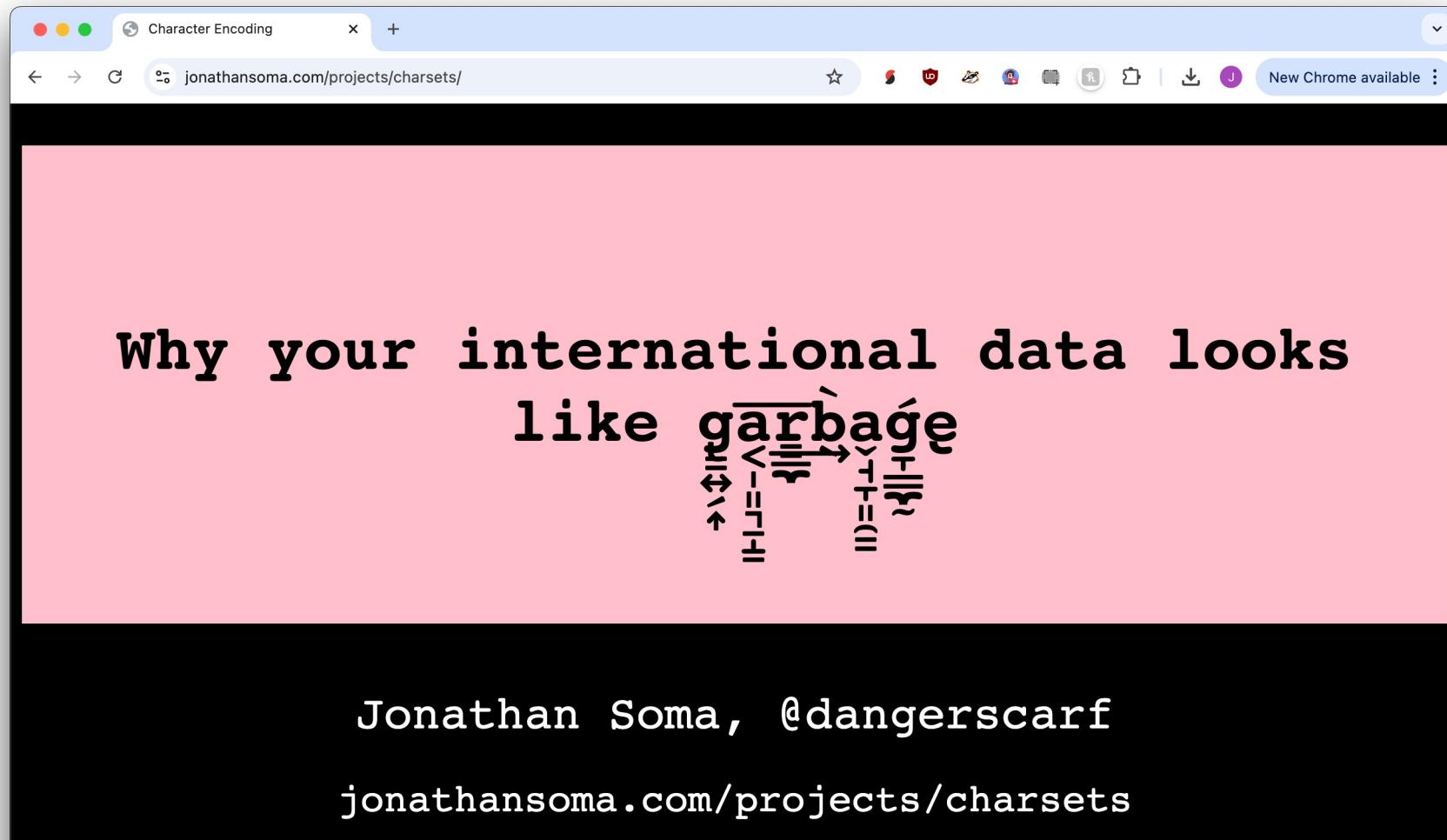
Features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Save As...

	A	B	C	D	E	F	G
1	name	price	tax	image_url	tags		
2	CORN	298	•298%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
3	YUZU CITRON	358	•358%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
4	BROCCOLI	398	•398%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
5	CUCUMBER 1PC	128	•128%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
6	CUCUMBERS 4PCS	498	•498%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
		138	•138%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
		298	•298%tax	https://www.	PERISHABLE	ÄÜΣËÎµ	
		398	•398%tax		PERISHABLE	ÄÜΣËÎµ	ORGANIC

**✖️** **⌚** products\_output.csv

name	price	tax	image_url	tags
CORN	298	¥298 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/0607133157_629ed4bd3cb1b.jpeg">https://www.national-azabu.net/upload/save_image/0607133157_629ed4bd3cb1b.jpeg</a>	PERISHABLE 冷藏
YUZU CITRON	358	¥358 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/1227165628_5fe83e2c165a4.jpeg">https://www.national-azabu.net/upload/save_image/1227165628_5fe83e2c165a4.jpeg</a>	PERISHABLE 冷藏
BROCCOLI	398	¥398 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/18_s.jpg">https://www.national-azabu.net/upload/save_image/18_s.jpg</a>	PERISHABLE 冷藏
CUCUMBER 1PC	128	¥128 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/230051000000_s.jpg">https://www.national-azabu.net/upload/save_image/230051000000_s.jpg</a>	PERISHABLE 冷藏
CUCUMBERS 4PCS	498	¥498 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/2_s.jpg">https://www.national-azabu.net/upload/save_image/2_s.jpg</a>	PERISHABLE 冷藏
CARROT 1PC	138	¥138 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/230034000000_s.jpg">https://www.national-azabu.net/upload/save_image/230034000000_s.jpg</a>	PERISHABLE 冷藏
CARROTS 3PCS	398	¥398 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/16_s.jpg">https://www.national-azabu.net/upload/save_image/16_s.jpg</a>	PERISHABLE 冷藏
ORGANIC CARROTS BIG PACK 1KG	398	¥398 (8%tax)	<a href="https://www.national-azabu.net/upload/save_image/16_s.jpg">https://www.national-azabu.net/upload/save_image/16_s.jpg</a>	PERISHABLE ORGANIC



<https://jonathansoma.com/projects/charsets/>

Visit <https://bit.ly/ds-dojo-2025> for material

# データ サイエンティスト DOJO

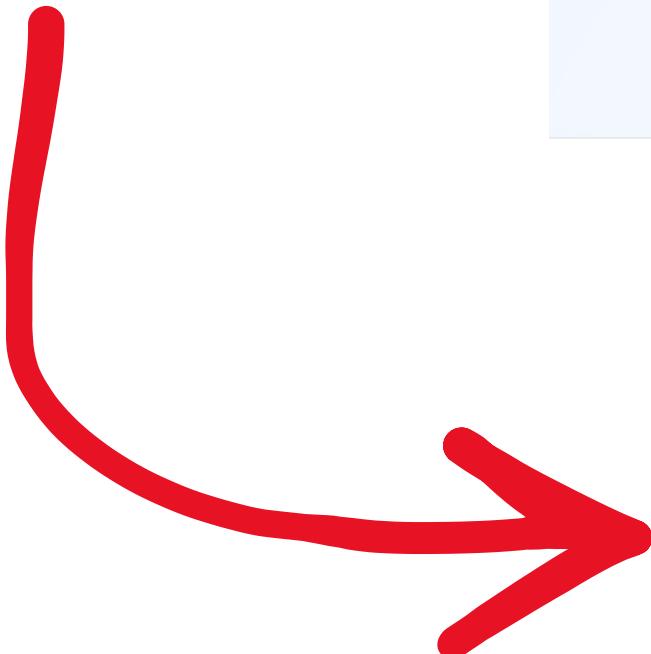
Jonathan Soma

[js4571@columbia.edu](mailto:js4571@columbia.edu)

Knight Chair in Data Journalism, Columbia University

# データサイエンティストDOJO 2025

6-day intensive Python workshop for covering data analysis, APIs, web scraping, AI, and visualization



## 🚀 Start Coding in the Cloud

GitHub Codespaces lets you code right in your browser — no installs, no setup. Your files and tools live safely in the cloud.

☁️ new codespaces

Create a fresh cloud workspace for this course on the  
codespaces branch — nothing to install.

📁 open existing

If you already have a Codespace for this repo, reopen it and pick  
up where you left off.

## Start Coding in the Cloud

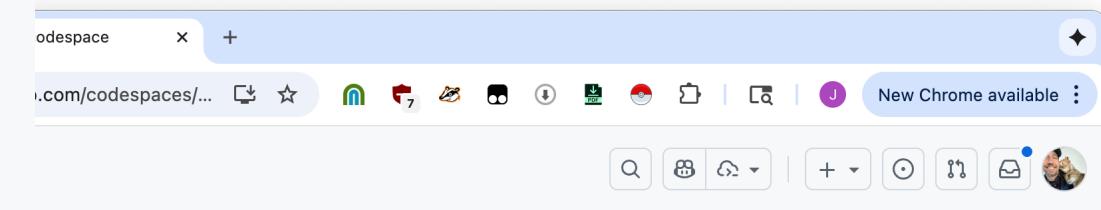
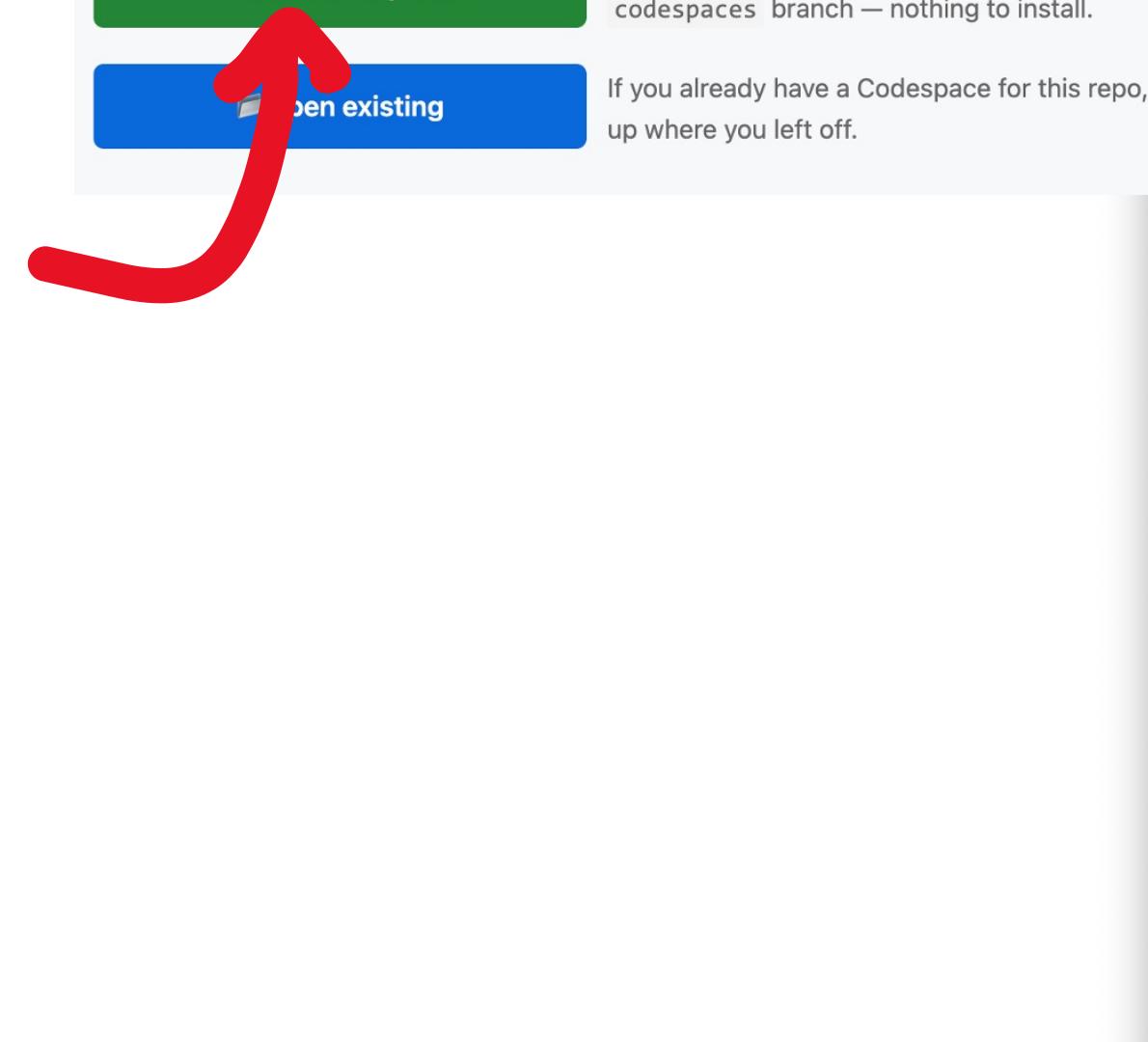
GitHub Codespaces lets you code right in your browser — no installs, no setup. Your files and tools live safely in the cloud.

 new codespaces

Create a fresh cloud workspace for this course on the `codespaces` branch — nothing to install.

 Open existing

If you already have a Codespace for this repo, reopen it and pick up where you left off.



### Create a new codespace

#### Repository

To be cloned into your codespace

jsoma/2025-ds-dojo ▾

 Codespace usage for this repository is paid for by jsoma.

#### Branch

This branch will be checked out on creation

 codespaces ▾

#### Region

Your codespace will run in the selected region

Southeast Asia ▾

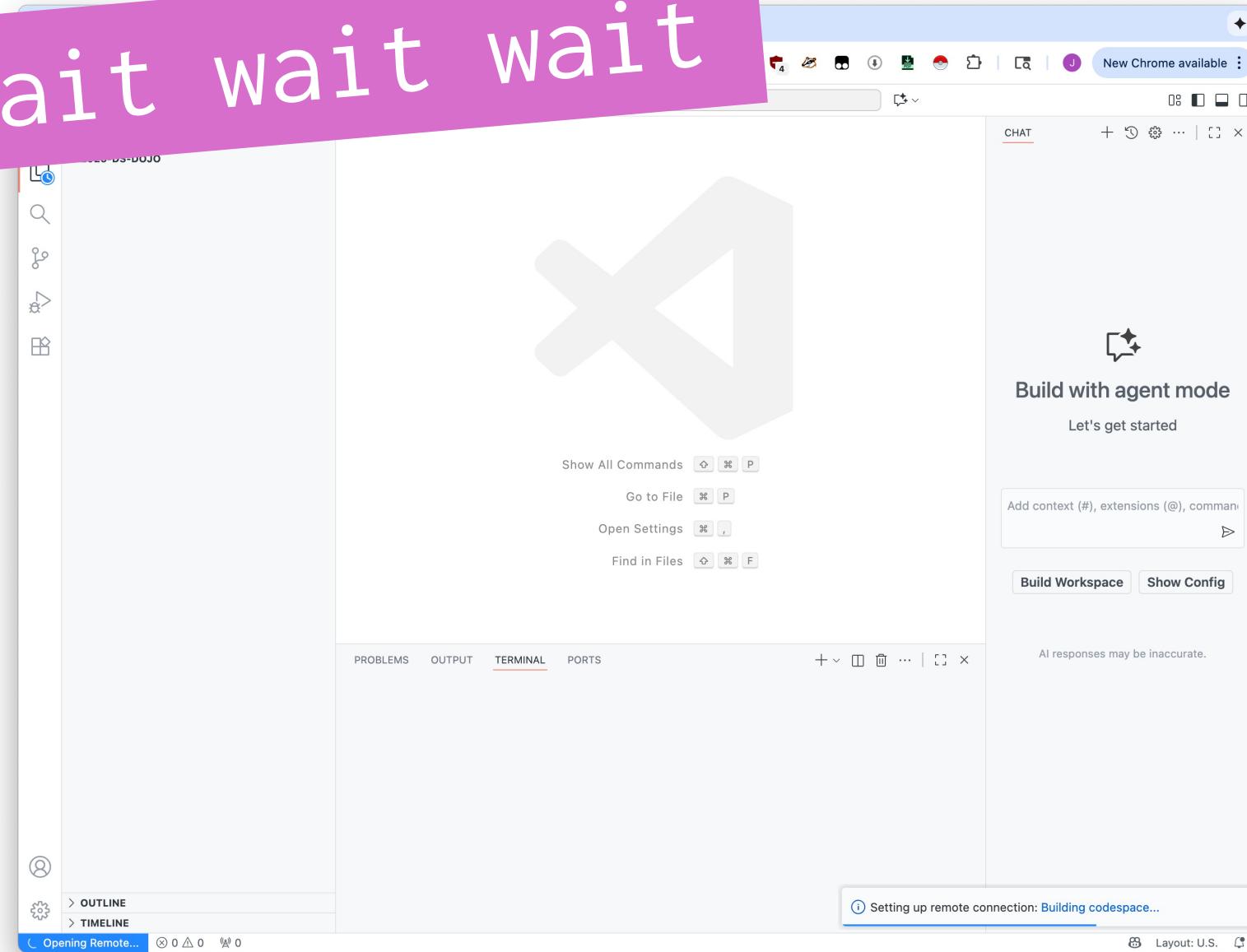
#### Machine type

Resources for your codespace

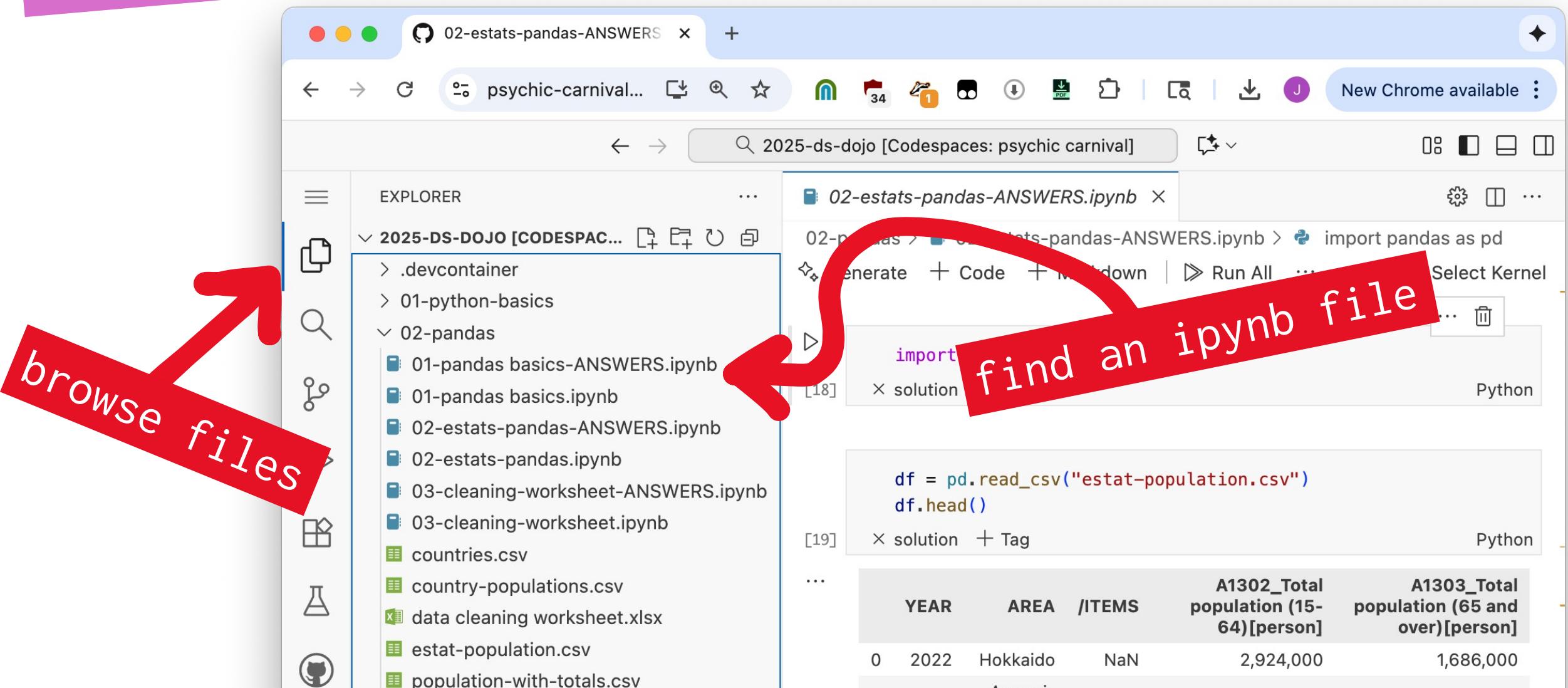
2-core ▾

**Create codespace**

# wait wait wait wait



# Open a notebook



# Run the code



New Chrome av

2025-ds-dojo [Codespaces: psychic carnival]



RER

DS-DOJO [CODESPAC...]



02-estats-pandas-ANSWERS.ipynb x

vcontainer

python-basics

pandas

-pandas basics-ANSWERS.ipynb

-pandas basics.ipynb

2-estats-pandas-ANSWERS.ipynb

2-estats-pandas.ipynb

3-cleaning-worksheet-ANSWERS.ipynb

3-cleaning-worksheet.ipynb

countries.csv

country-populations.csv

+ General + Code + Markdown ▶ Run All ...



...

import pandas as pd

[18] x solution + Tag

df = pd.read\_csv("estat-population.csv")  
df.head()

[19] x solution + Tag

...

YEAR

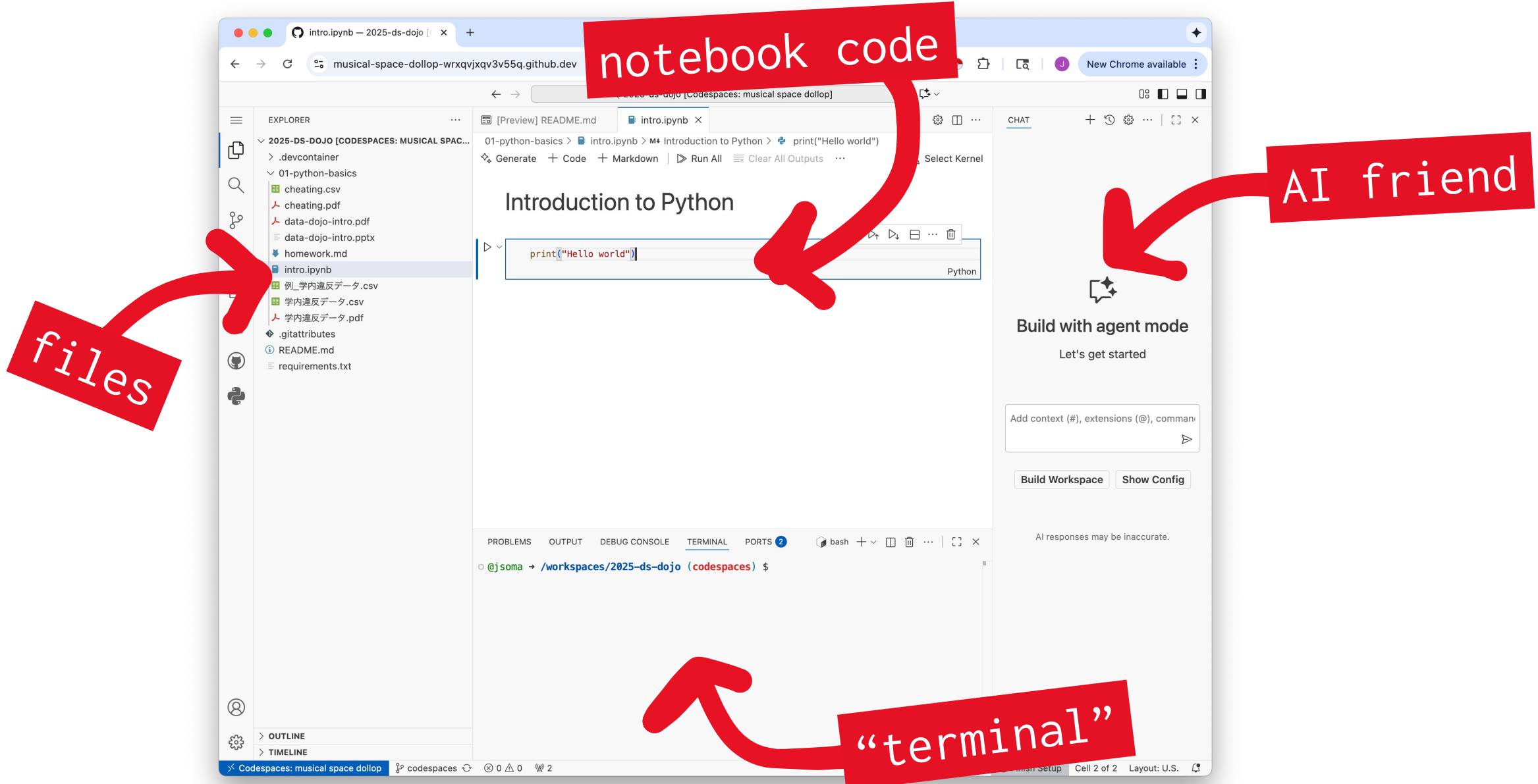
AREA

A1302\_Total  
population (15)

A1303

# Write new code

A screenshot of a Jupyter Notebook interface. At the top, there is a navigation bar with back and forward arrows, a search bar containing "2025-ds-dojo [Codespaces: musical space dollop]", and a tab for "[Preview] README.md". Below the navigation bar, there is a sidebar showing a file tree: "01-python-basics" is expanded, showing "intro.ipynb > m→ Introduction to Python". Underneath the file tree are buttons for "Generate", "Code", "Markdown", "Run All", "Clear All Outputs", and more. A large red arrow points to the "Code" button. The main content area shows a cell with the text "# This is a cell in a notebook".



EXPLORER

Browse files



Search files



Plugins



2025-DS-DOJO [CODESPAC...]



.devcontainer

01-python-basics

cheating.csv

cheating.pdf

data-dojo-intro.pdf

data-dojo-intro.pptx

homework.md

intro.ipynb

M

例\_学内違反データ.csv

学内違反データ.csv

学内違反データ.pdf

.gitattributes

README.md

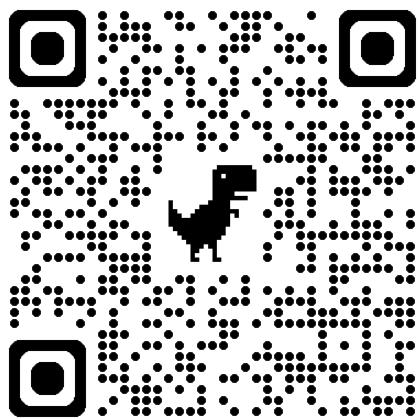
requirements.txt

# Python basics

I hope it is review, but if  
not: that's okay!

# A tiny Python quiz

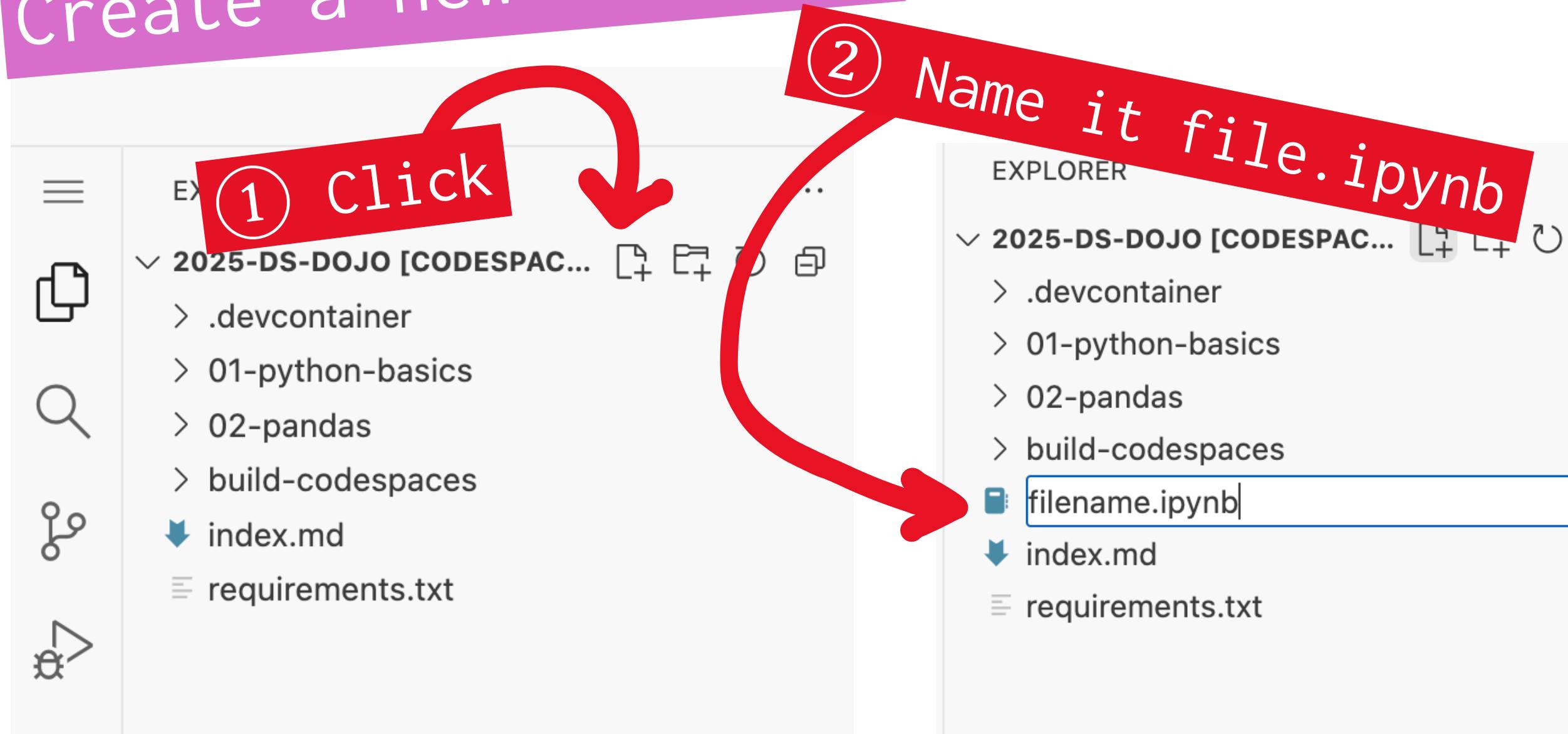
Don't run the  
code, just think  
and try to answer  
the questions



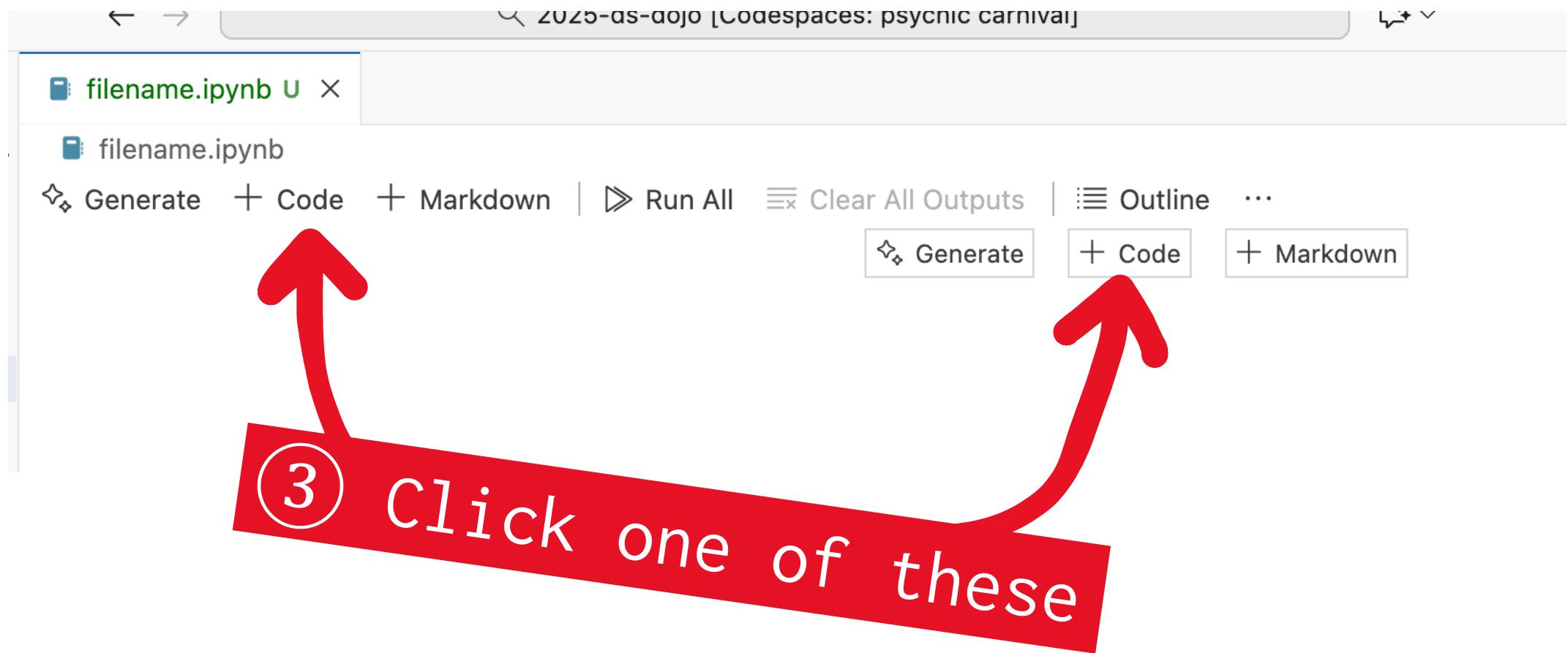
Maybe easier on your  
computer than phone?

<https://bit.ly/dojo25-python-quiz>

# Create a new notebook



# Add a new cell



let's do some work!

time to try out some Python

# Building Products

nytimes.com/interactive/2014/upshot/dialect-quiz-map.html

THE UPSHOT The New York Times GIVE THE TIMES Account

TheUpshot

## How Y'all, Youse and You Guys Talk

By [Josh Katz](#) and [Wilson Andrews](#) Dec. 21, 2013

Share full article

What does the way you speak say about where you're from?  
Answer all the questions below to see your personal dialect map.

QUESTION 1 OF 25

How would you address a group of two or more people?

- you all
- yous / youse
- you lot
- you guys
- you 'uns
- yinz
- you
- other
- y'all

### Your Map

See the pattern of your dialect in the map below. Three of the most similar cities are shown.

Least similar Most similar

Show least similar SHARE YOUR MAP: [f](#) [t](#) [g](#)

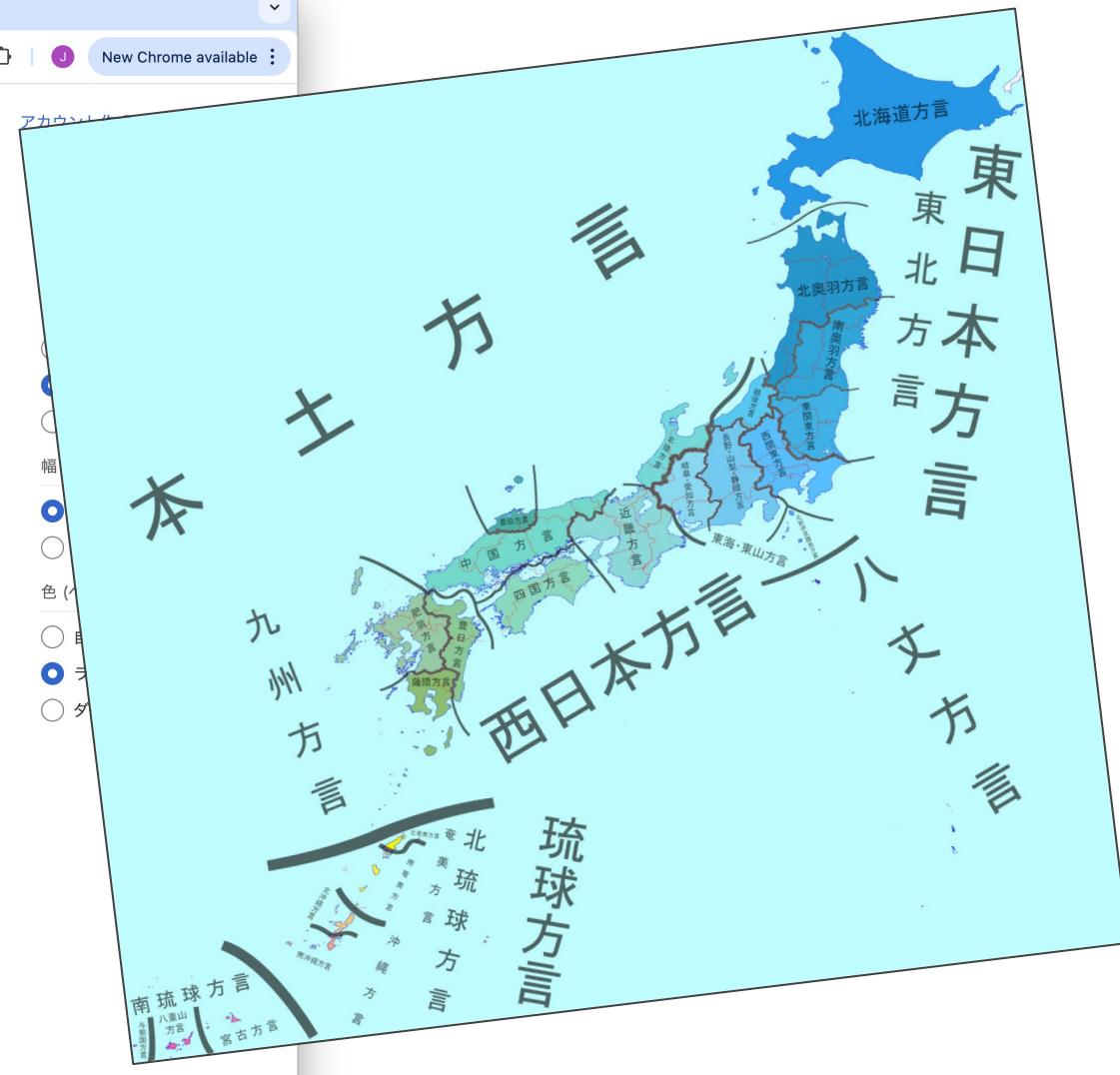
Yonkers

Washington DC

Arlington

# Can you build it?

The screenshot shows a Google Chrome browser window with the Japanese Wikipedia article on "Dialects of Japanese" (日本語の方言) open. The page content includes a sidebar with a map of Okinawa and its dialects.



# claude.ai

The screenshot shows the Claude AI web interface. At the top, there's a pink header bar with the text "claude.ai". Below it is a browser window with the URL "claude.ai/new". The main content area is titled "Claude" and features a "Professional Plan" button. A large, friendly greeting "Good morning, j" is displayed with a small orange starburst icon. Below the greeting, a text input field asks "How can Claude help you today?". Underneath this, a section titled "Claude 3.5 Sonnet" is visible. A "Add content" button is present, along with three cards: "Provide stakeholder perspective", "Generate interview questions", and "Polish your prose". There's also a "Use a project" dropdown and a "View all" link for recent chats. At the bottom left, there's a small circular icon with "JS" and a square icon.

claude.ai/new

Claude

Professional Plan

Good morning, j

How can Claude help you today?

Claude 3.5 Sonnet

Add content Use a project

Provide stakeholder perspective Generate interview questions Polish your prose

Your recent chats Show View all →

JS

Visit <https://bit.ly/ds-dojo-2025> for material

# データ サイエンティスト DOJO

Jonathan Soma

[js4571@columbia.edu](mailto:js4571@columbia.edu)

Knight Chair in Data Journalism, Columbia University