

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

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...

POLITICO

MEDIA

Over 500 journalists were laid off in January 2024 alone

The job cuts come after an already bleak year in the news business.

Los Angeles Times

More than 100 employees were let go from The Los Angeles Times in January. | Patrick T. Fallon/AFP via Getty Images

By KIERRA FRAZIER
02/01/2024 07:31 AM EST

[Facebook](#) [Twitter](#) [Email](#) [More](#)

Publishers | September 4, 2024

News media job cuts 2024 tracked: **Guardian, Time, Axios, Tampa Bay Times, NYPost, Hollywood Reporter all hit in August**

Big losses at the likes of The Messenger, Sports Illustrated and Media Ireland started 2024.

By Charlotte Tobitt

A giant CNN sign is seen at the entrance to the CNN Center in Atlanta, Georgia, U.S., February 2022. Picture: Shutterstock/Tada Images

FC News and media layoffs 2024

announced layoffs in the media space

FAST COMPANY

- *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

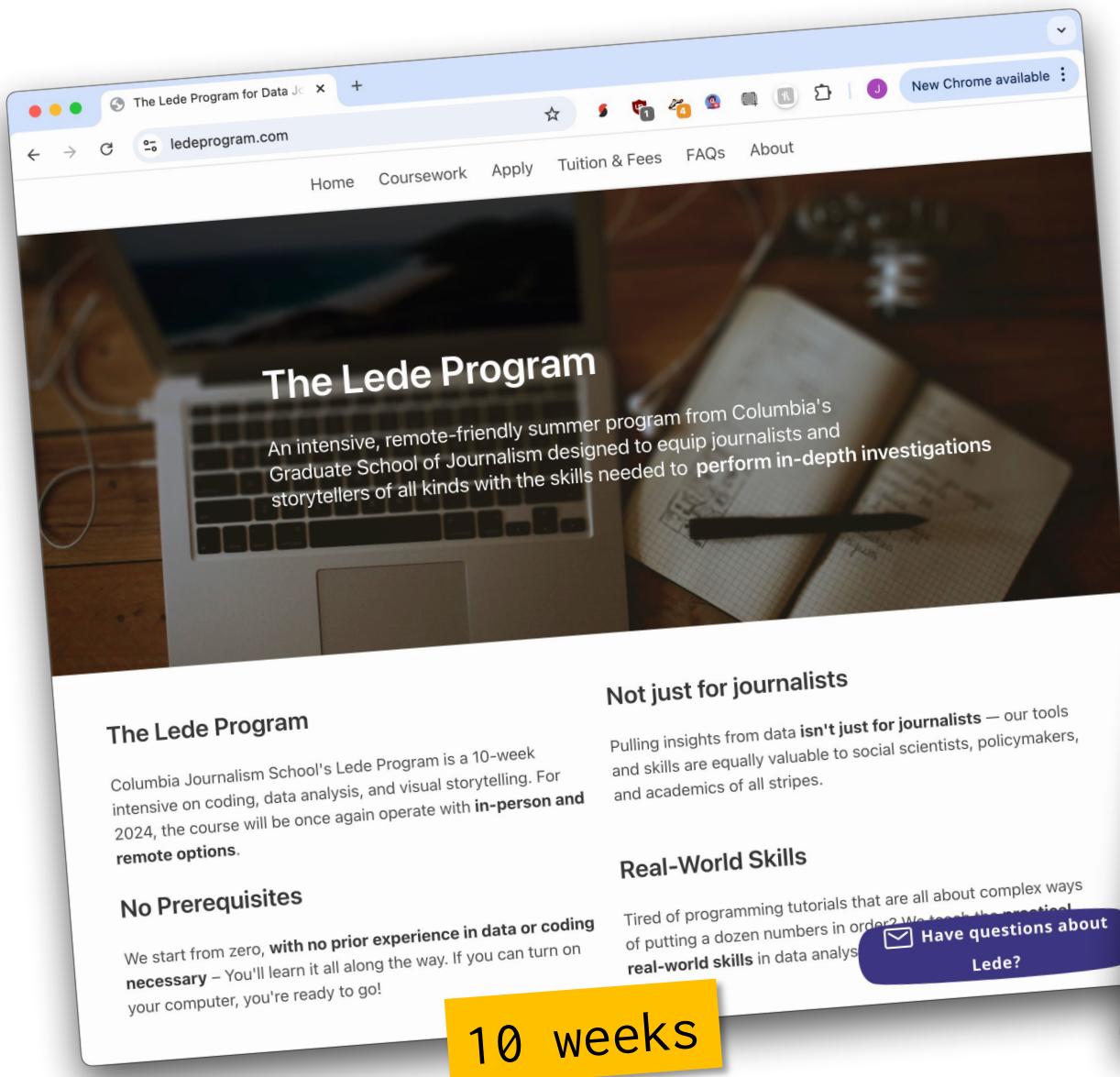
Remote · Full-time · Mid-Senior level

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

nikkeibp.co.jp/seminar/nb/ddojo/?n_cid=emsl_...

特徴 プログラム 開催概要 お問い合わせ 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

データサイエンティスト
DOJO(道場)

全6日間集中講座

2024年9月23日(月)開講

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

one week!

事業計画や中期経営計画の立案で、
過去の経験や勘に頼ってはいませんか？

データの裏付けがない計画は、実効力がありません。プロジェクトリーダー自らが、適切な仮説を打ち立て、データの意味を読み解き、深い洞察力を身に着ける必要があります。データサイエンティストとしてのスキルを学ぶにとどまらず、データに裏打ちされた仮説を検証する思考回路を要請するのが、この研修の主たる目的です。「仮説を立てる」→「データを集める」→「データを分析する」→「ビジュアライゼーションによって、理解と洞察力を深化する」→「チームで共有し、次なる仮説を立てる」このサイクルを繰り返すことによって、より精度の高い事業計画や中期経営計画の立案が可能になります。

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.

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:** Working on 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!

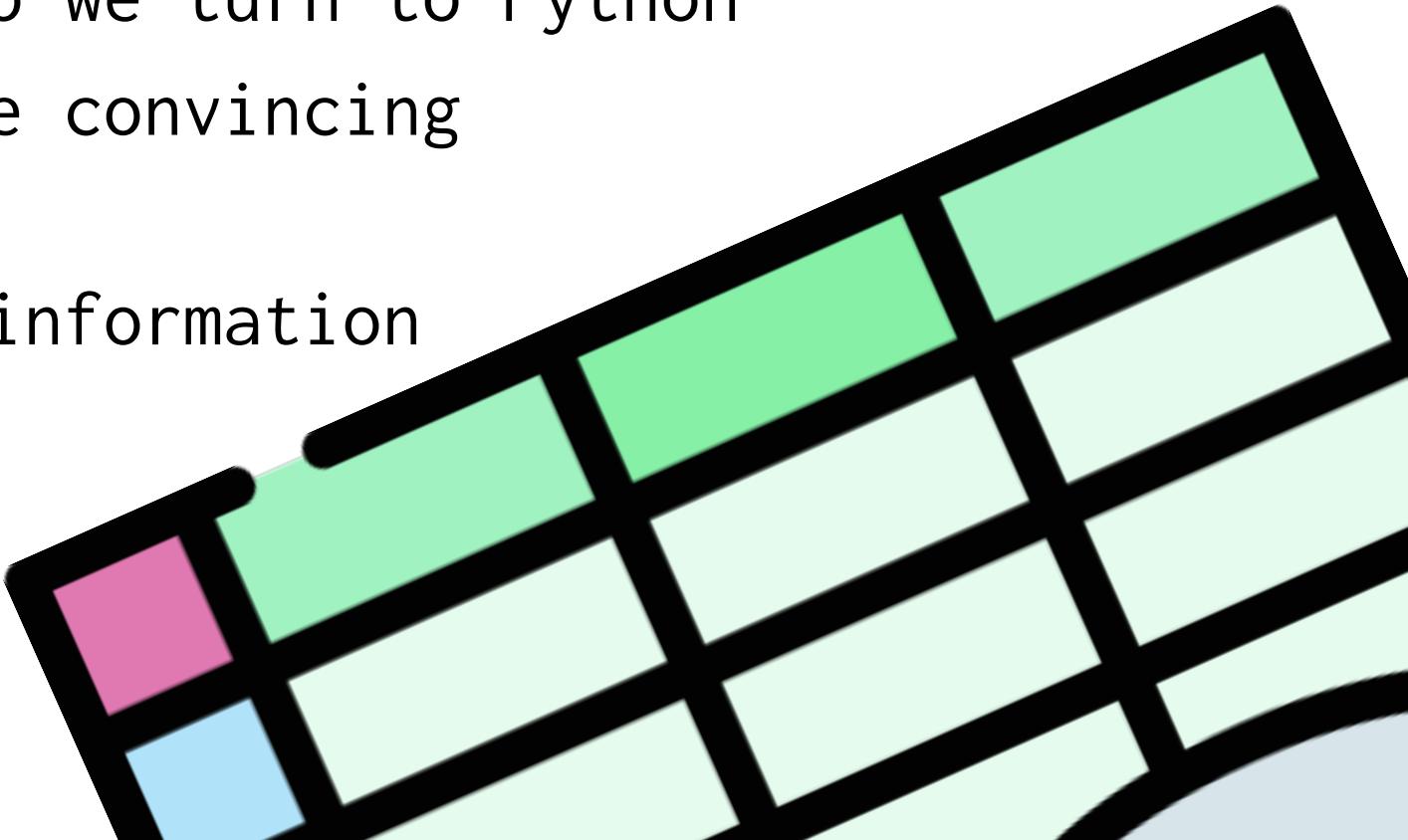
Python programming

- 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



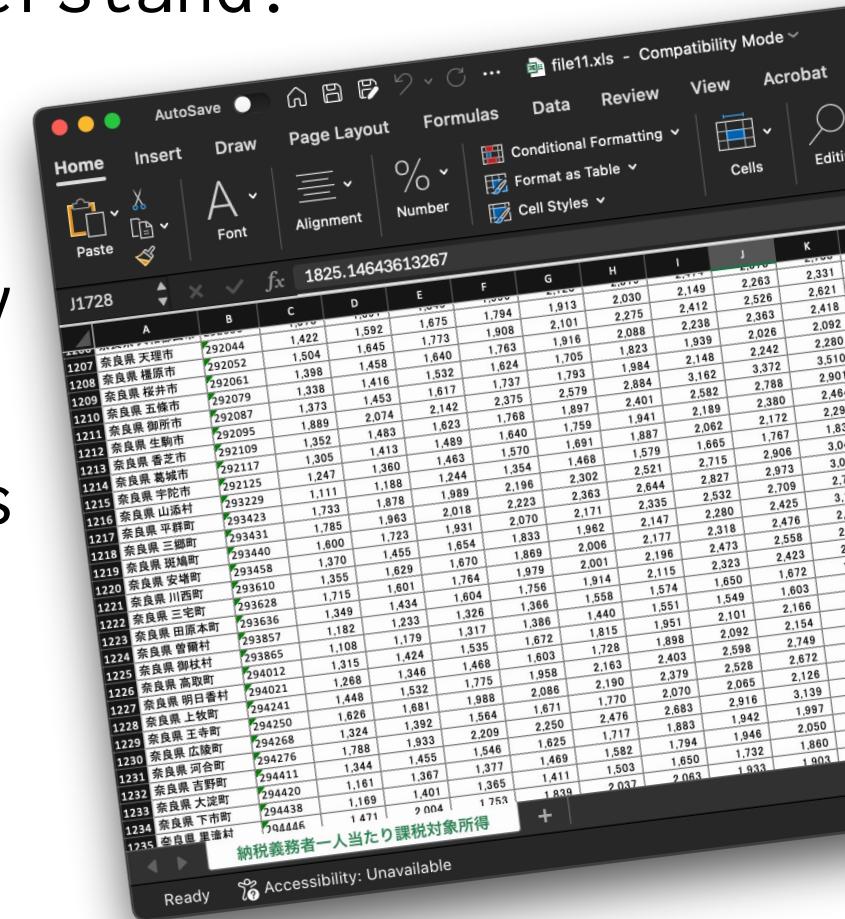
Data analysis

- 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

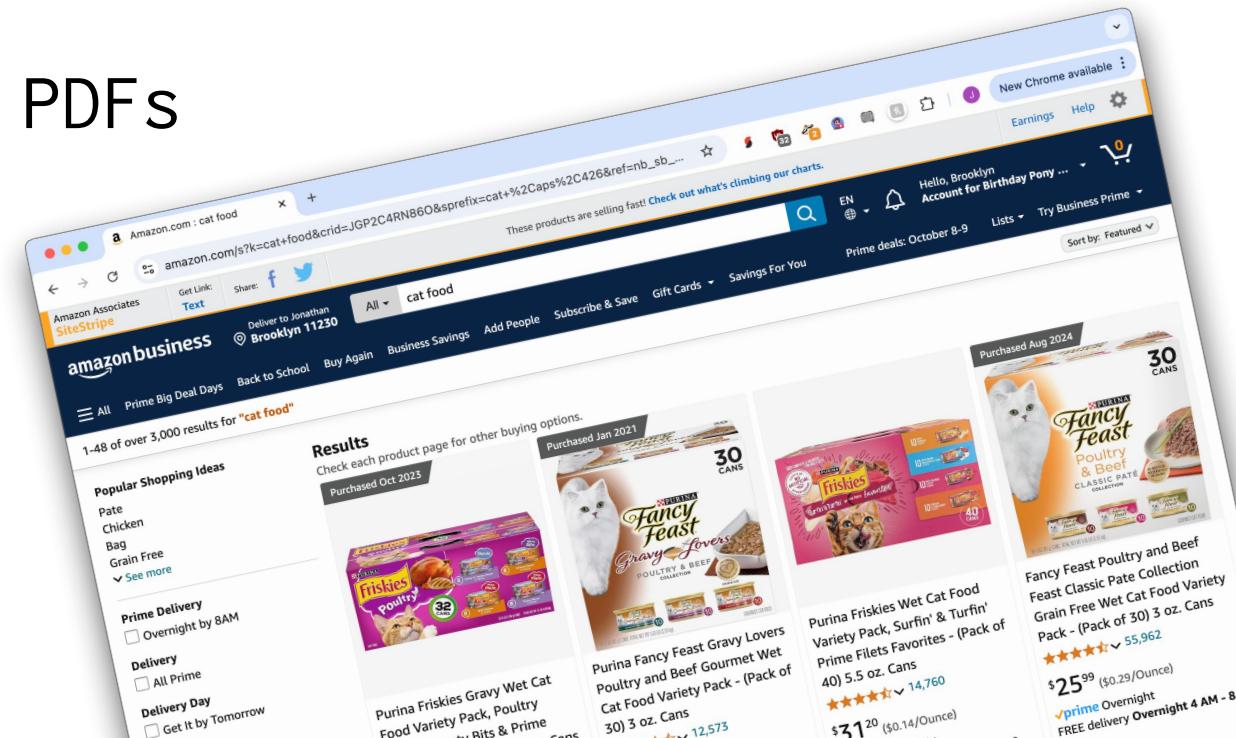


A screenshot of Microsoft Excel showing a table of structured data. The table has columns labeled A through K. Column A contains Japanese place names, and columns B through K contain numerical values. The Excel ribbon at the top shows tabs for Home, Insert, Draw, Page Layout, Formulas, Data, Review, View, and Acrobat. The status bar at the bottom right indicates "Accessibility: Unavailable".

	J1728	A	B	C	D	E	F	G	H	I	J	K
1207	奈良県 天理市	292044	1.422	1.592	1.675	1.794	1.913	2.030	2.149	2.263	2.331	
1208	奈良県 檜原市	292052	1.504	1.645	1.773	1.908	2.101	2.275	2.412	2.526	2.618	
1209	奈良県 桜井市	292061	1.398	1.458	1.840	1.763	1.916	2.088	2.238	2.363	2.418	
1210	奈良県 五條市	292079	1.338	1.416	1.532	1.624	1.705	1.823	1.939	2.026	2.026	2.092
1211	奈良県 御所市	292087	1.373	1.453	1.617	1.737	1.793	1.884	2.148	2.242	2.372	2.418
1212	奈良県 生駒市	292095	1.889	2.074	2.142	2.375	2.579	2.884	3.162	3.372	3.510	
1213	奈良県 奈芝市	292109	1.352	1.483	1.623	1.768	1.897	2.401	2.582	2.788	2.907	
1214	奈良県 墓城市	292117	1.305	1.413	1.489	1.640	1.759	1.941	2.189	2.380	2.468	
1215	奈良県 宇陀市	292125	1.247	1.360	1.463	1.570	1.665	1.759	1.887	2.062	2.172	2.288
1216	奈良県 山添村	292129	1.111	1.188	1.244	1.354	1.468	1.579	1.665	1.767	1.853	1.883
1217	奈良県 平群町	293423	1.733	1.878	1.989	2.196	2.302	2.383	2.464	2.521	2.593	
1218	奈良県 三郷町	293431	1.785	1.963	2.018	2.223	2.383	2.435	2.532	2.625	2.709	
1219	奈良県 斎場町	293440	1.604	1.723	1.931	2.070	2.171	2.335	2.532	2.625	2.709	
1220	奈良県 安堵町	293458	1.370	1.455	1.654	1.833	1.962	2.147	2.280	2.476	2.558	
1221	奈良県 川西町	293610	1.355	1.629	1.670	1.869	2.006	2.177	2.318	2.423	2.506	
1222	奈良県 三宅町	293628	1.715	1.801	1.764	1.979	2.001	2.196	2.473	2.558	2.643	
1223	奈良県 田原本町	293636	1.349	1.434	1.604	1.756	1.914	2.115	2.323	2.425	2.512	
1224	奈良県 曽爾村	293857	1.182	1.233	1.326	1.366	1.558	1.574	1.650	1.734	1.812	
1225	奈良県 御杖村	293865	1.108	1.179	1.317	1.388	1.440	1.551	1.549	1.603	1.663	
1226	奈良県 高取町	294012	1.315	1.424	1.535	1.672	1.815	1.951	2.101	2.166	2.214	
1227	奈良県 明日香村	294021	1.268	1.346	1.468	1.603	1.728	1.898	2.092	2.154	2.214	
1228	奈良県 上牧町	294241	1.448	1.532	1.775	1.958	2.163	2.403	2.598	2.749	2.872	
1229	奈良県 王寺町	294250	1.626	1.881	1.988	2.088	2.190	2.379	2.528	2.672	2.816	
1230	奈良県 広陵町	294268	1.324	1.392	1.584	1.671	1.770	2.070	2.216	2.365	2.516	
1231	奈良県 河合町	294276	1.788	1.933	2.209	2.250	2.476	2.683	2.916	3.139	3.346	
1232	奈良県 吉野町	294411	1.344	1.455	1.546	1.625	1.717	1.883	1.942	2.050	2.160	
1233	奈良県 大淀町	294420	1.161	1.367	1.377	1.469	1.582	1.794	1.946	2.050	2.160	
1234	奈良県 下市町	294438	1.169	1.401	1.365	1.411	1.503	1.650	1.732	1.833	1.903	
1235	奈良県 里塙村	294446	1.471	2.004	1.753	2.037	2.063	2.183	2.283	2.383	2.480	

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!



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

```
# 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

errors are expected

Introduced

Completed

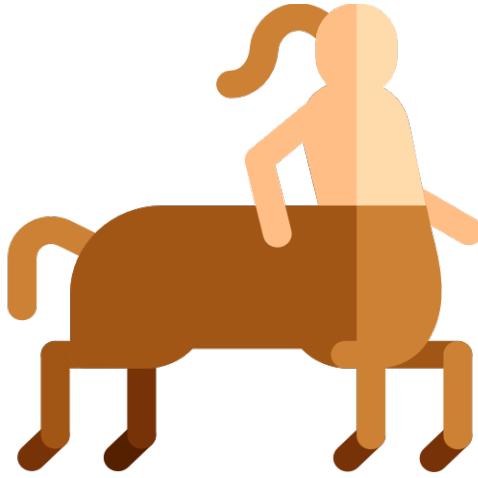
Legislation

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

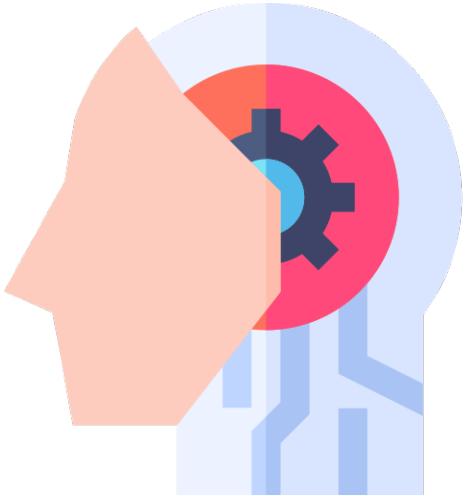


```
# Python function to calculate fibonacci sequence
def fibonacci(n):
    if n == 0:
        return 0
    elif n == 1:
        return 1
    else:
        return fibonacci(n-1) + fibonacci(n-2)
```

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>

Software Installation

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 application window titled "ANACONDA NAVIGATOR". It has tabs for "Home", "Documentation (26)", "Training (0)", and "Connect". The main area shows a list of installed Python environments, which is identical to the one in Python Wrangler.

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

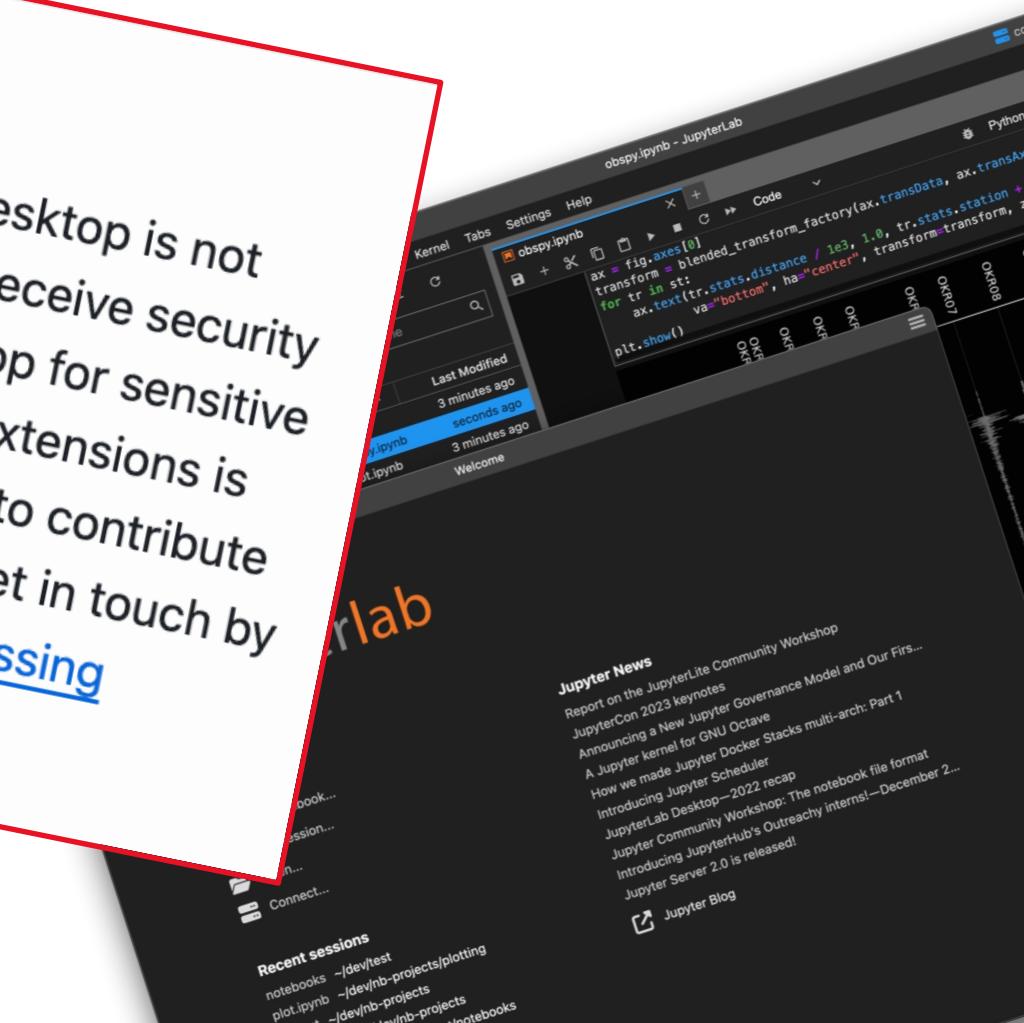


JupyterLab Desktop

- 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

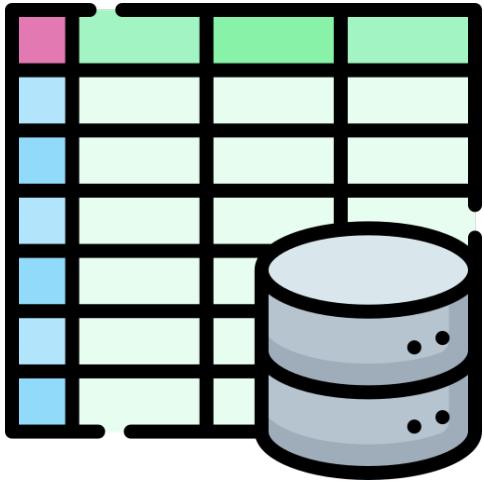
the cloud

```
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



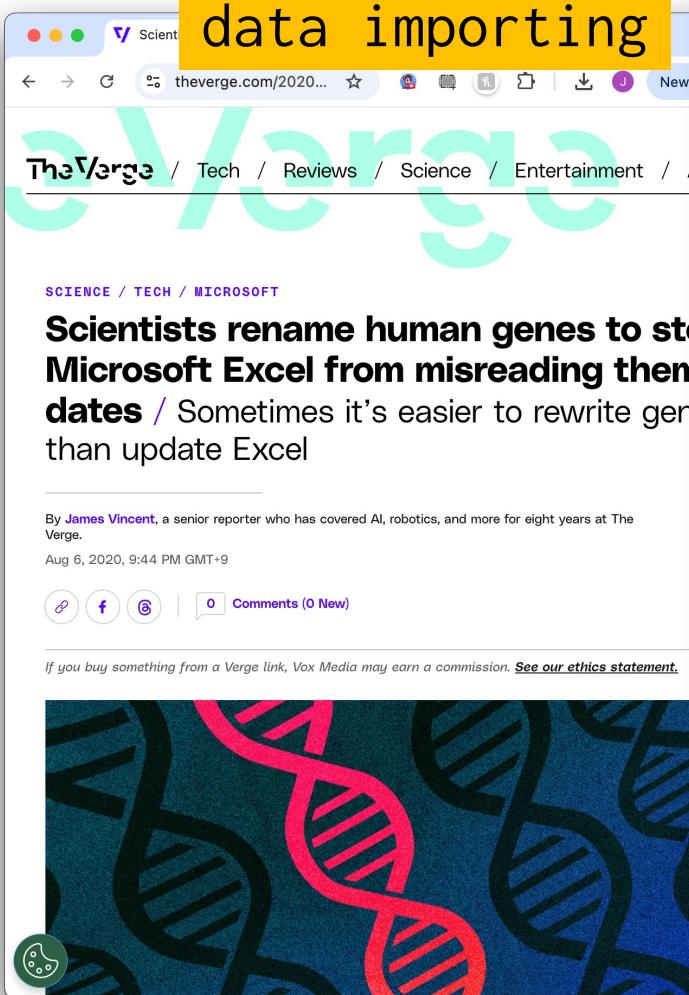
Excel vs. Python



- 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.

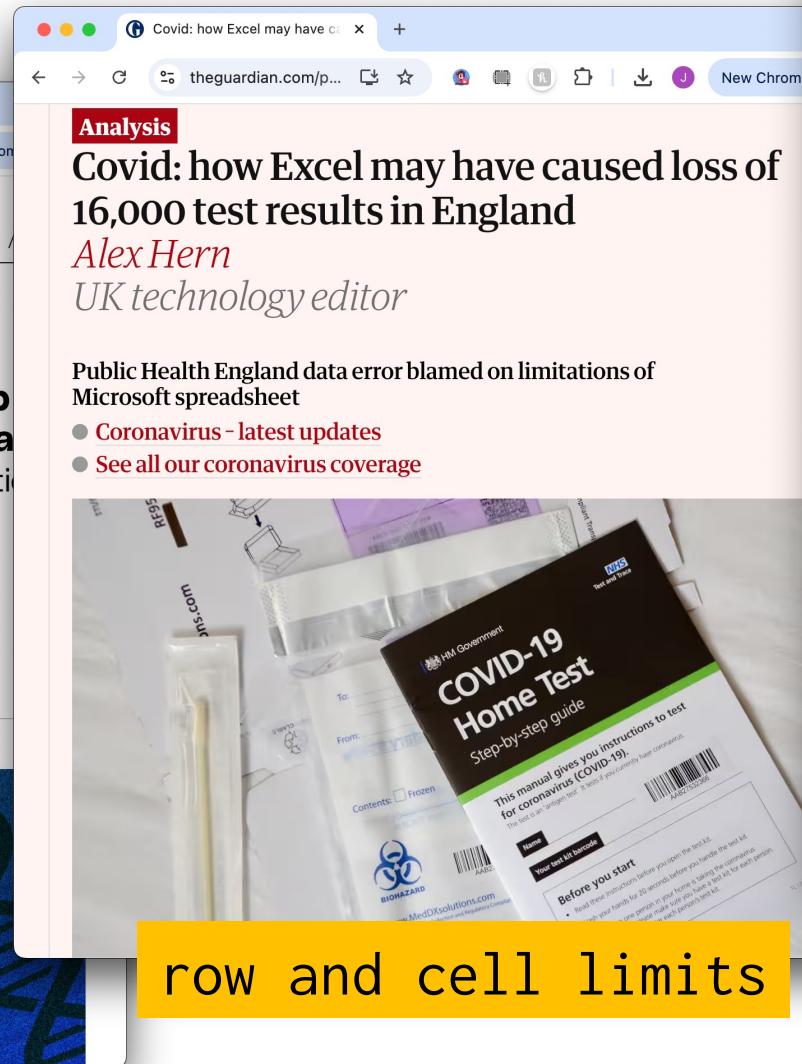
Aug 6, 2020, 9:44 PM GMT+9

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

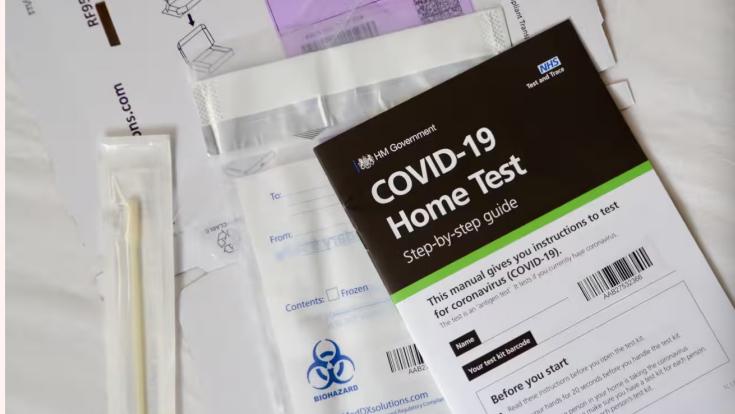


Covid: how Excel may have caused loss of 16,000 test results in England

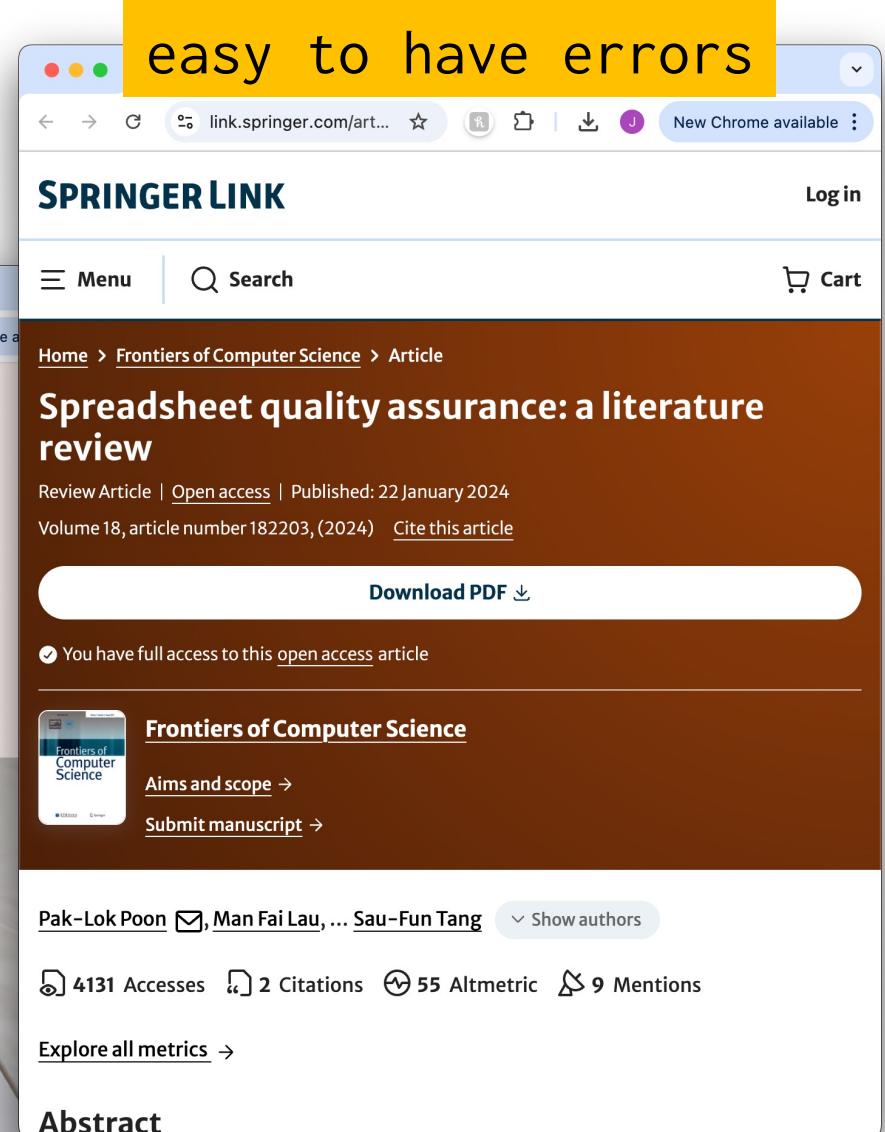
Alex Hern
UK technology editor

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

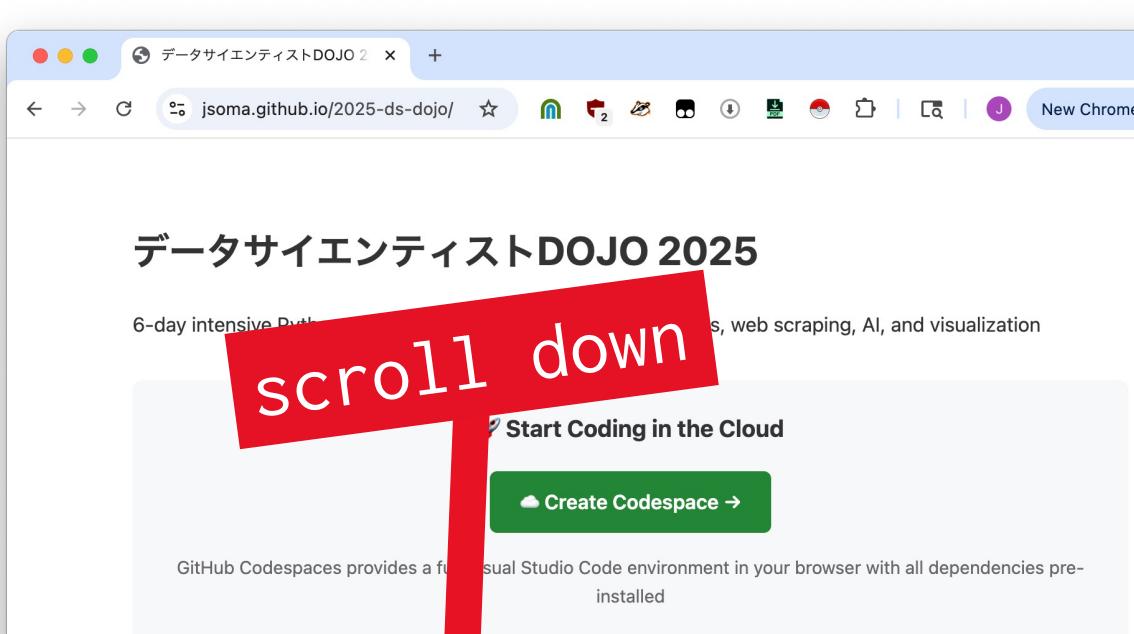
Easy to read

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

But 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 →

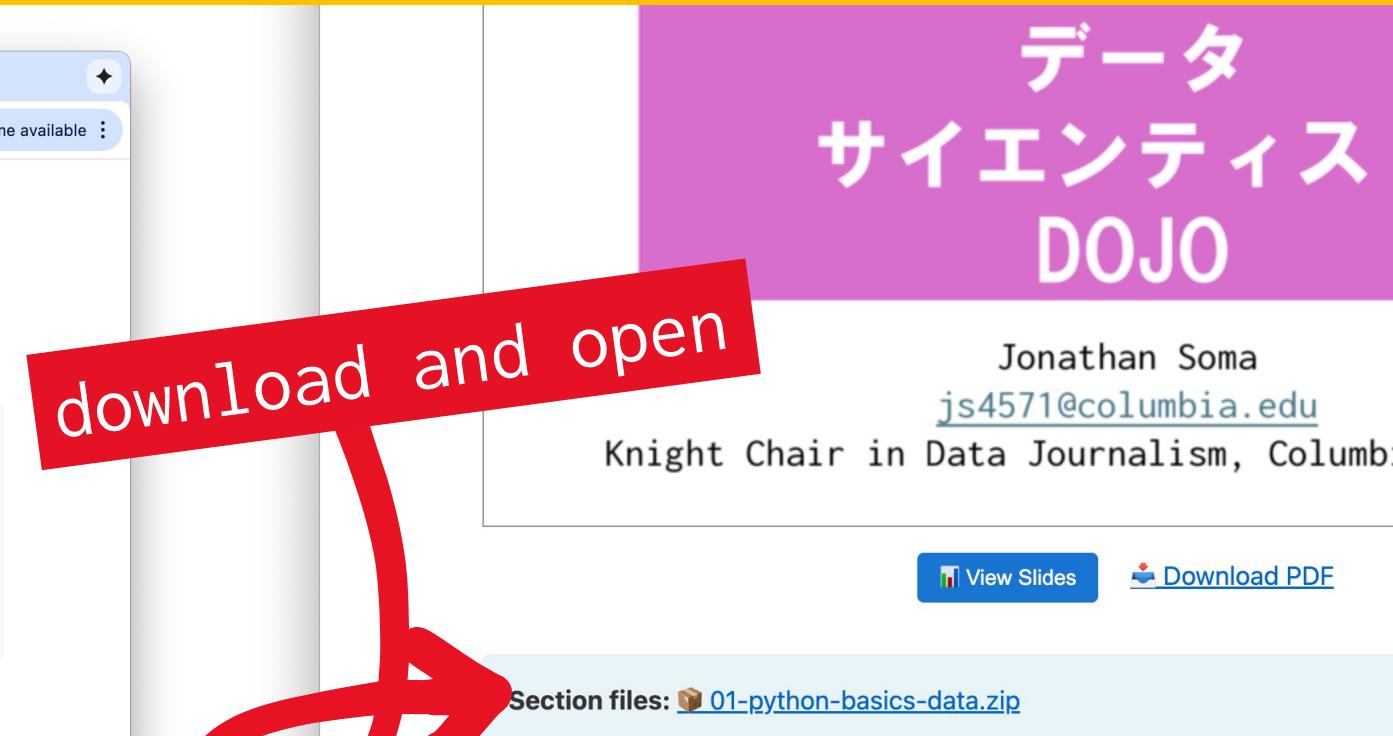
GitHub Codespaces provides a full Visual Studio Code environment in your browser with all dependencies pre-installed

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



データ
サイエンティス
DOJO

Jonathan Soma
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

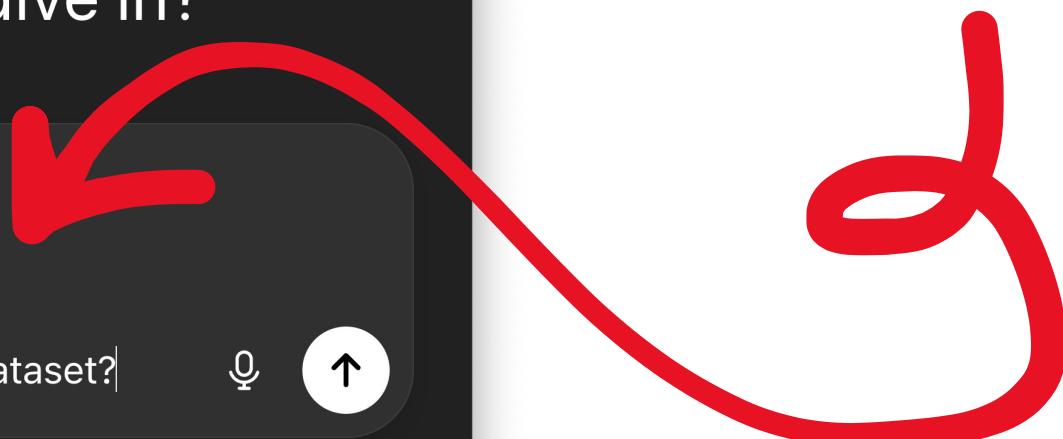


発生日時	違反内容	科目名
8/3/21	盗用	英語 472/Shakespeare Survey
9/30/21	不正行為, 盗用	Exercise, Sport and Health Education 396-01
9/30/21	盗用	英語 111-74-大学作文の原則
9/30/21	盗用	英語 111-74-大学作文の原則
9/30/21	盗用	英語 111-761 大学作文の原則
10/11/21	不正行為, 盗用	Anthropological Sciences
10/11/21	不正行為, 盗用	Anthropological Sciences
10/11/21	不正行為, 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**?

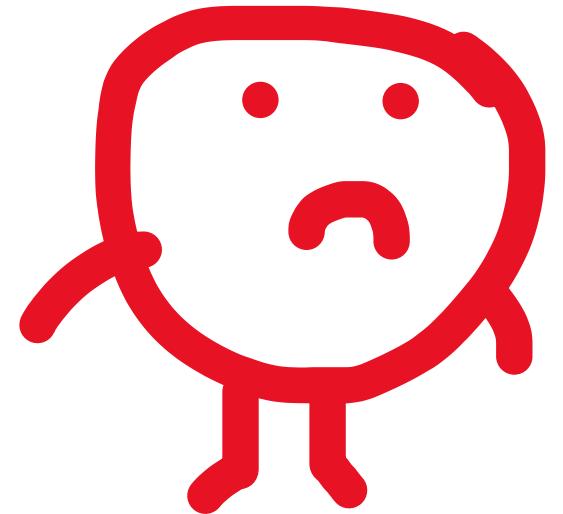
+ Ask anything

ChatGPT can make mistakes. Check important info.

File Academic_Integrity_Violations_Fa2021-Sp2025 _7.11.2025.pdf

But then we didn't write it 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.

**It will lie to your face and cite
sources while it does so.**

It's not analyzing, it's just looking at the words and predicting 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[^] A_^ 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[^] A^v

B I U | A^l A^r

General

Conditional Formatting

Format as Table

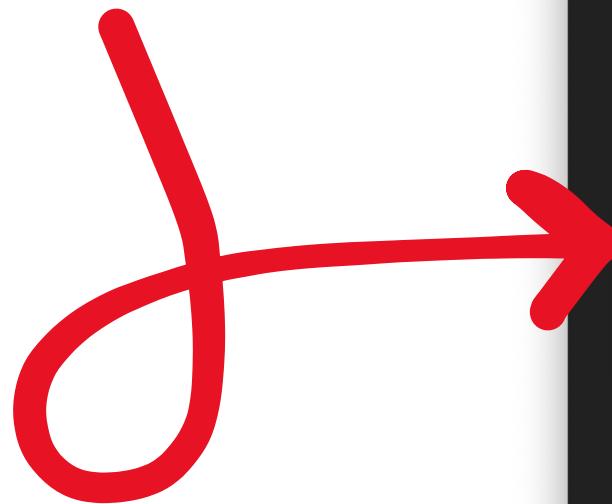
Cell Styles

✖ 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? 0 ↑

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.

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

Knight Chair in Data Journalism, Columbia University

The screenshot shows two overlapping web pages. The background page is the 'Data Scientist DOJO 2025' landing page, which includes a 'Create Codespace' button. The foreground page is a GitHub 'Create new codespace' interface, which also has a 'Create codespace' button. A large red arrow points from the GitHub button to the landing page button.

oma.github.io/2025-ds-dojo/

タサイエンティストDOJO 2025

Comprehensive Python workshop for covering data analysis, APIs, web scraping, AI, and visualization

Start Coding in the Cloud

Create Codespace →

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

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

データ サイエンティスト DOJO

Create new codespace

github.com/codespaces/...

Codespaces

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

Terms Privacy Security Status Community Docs Contact Manage cookies Do not share my personal information

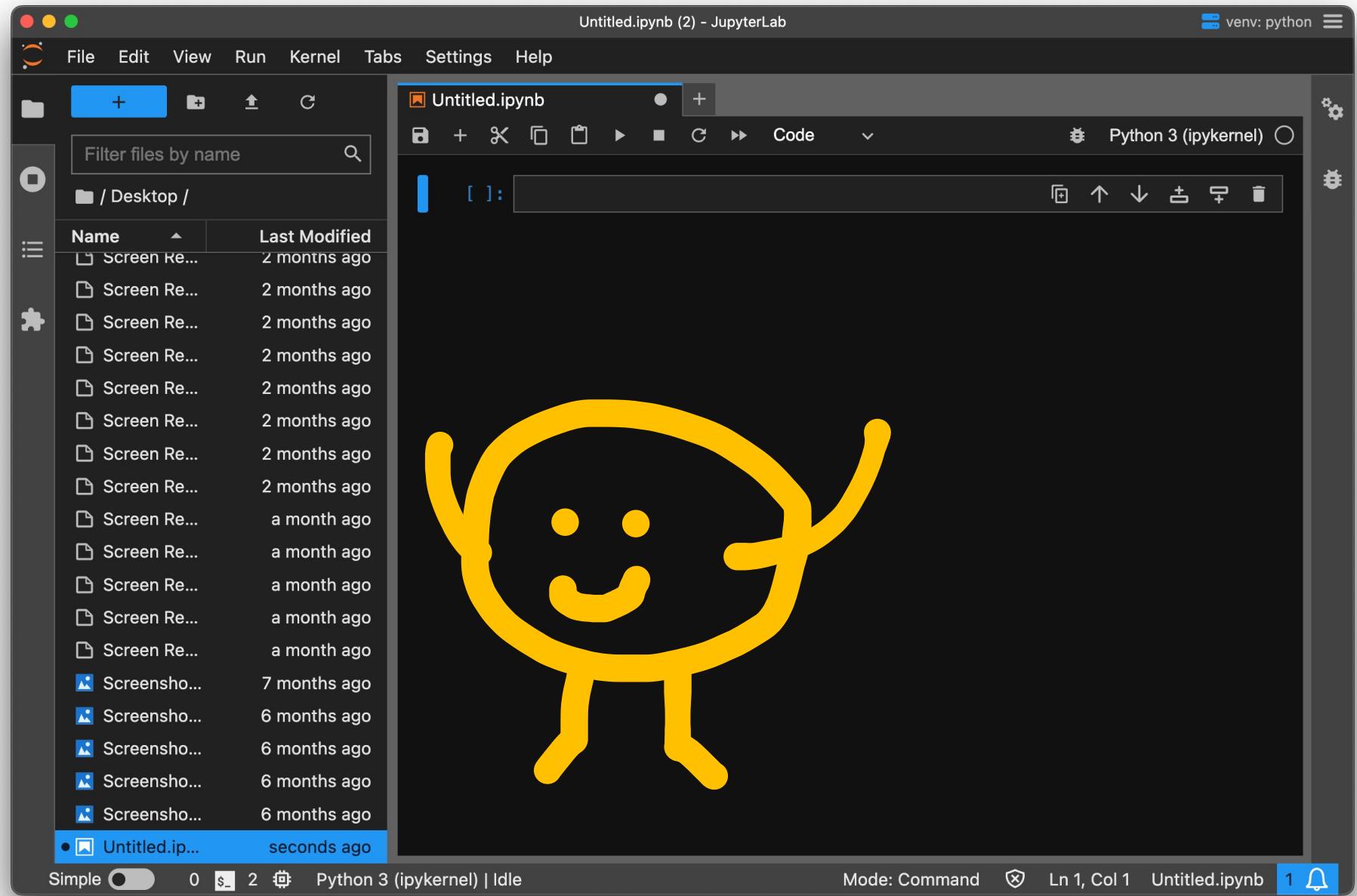
© 2025 GitHub, Inc.

Create a new notebook

① click “New session”

② Double-click Desktop

③ click Python 3
under “Notebook”



Untitled.ipynb

[]: `print("Hello world")`

Python 3 (ipykernel)

Press Play or Shift+Enter/Return



You should see this

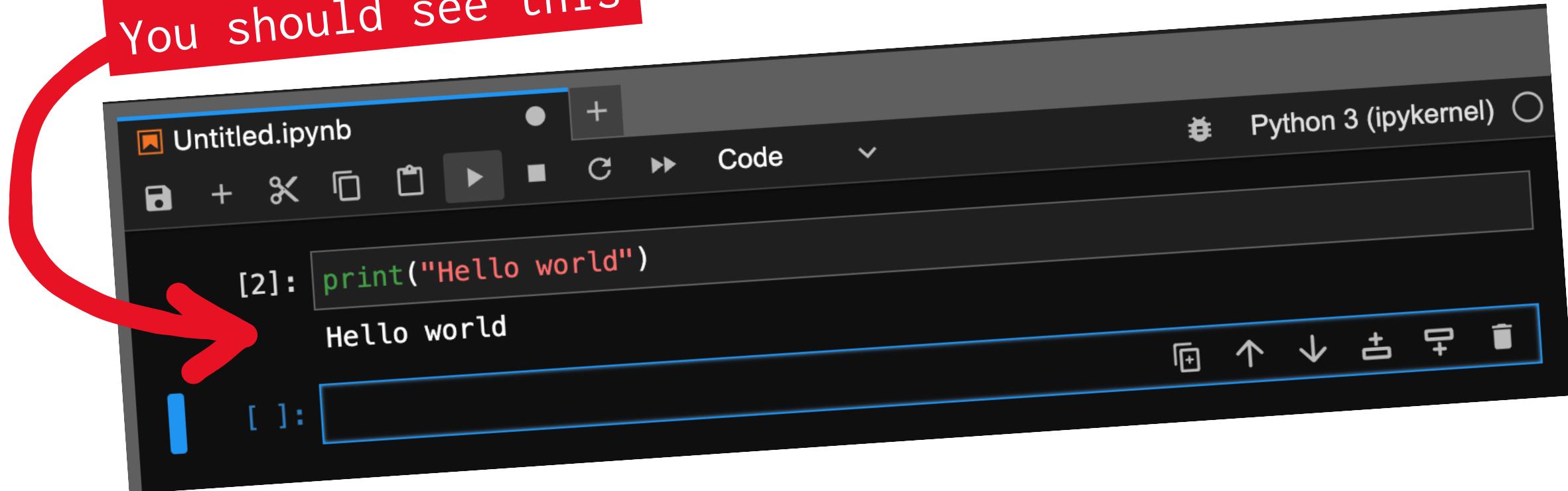
Untitled.ipynb

[2]: `print("Hello world")`

Hello world

[]:

Python 3 (ipykernel)



Home Insert Draw

Comments

Share

Character encoding

Cell Formatting

Table

Cell Styles

Insert

Delete

Format

Σ

A

Y

Filter

Sort &

Find &

Add-ins

Analyze

Data

Create PDF

and share link

Save As...

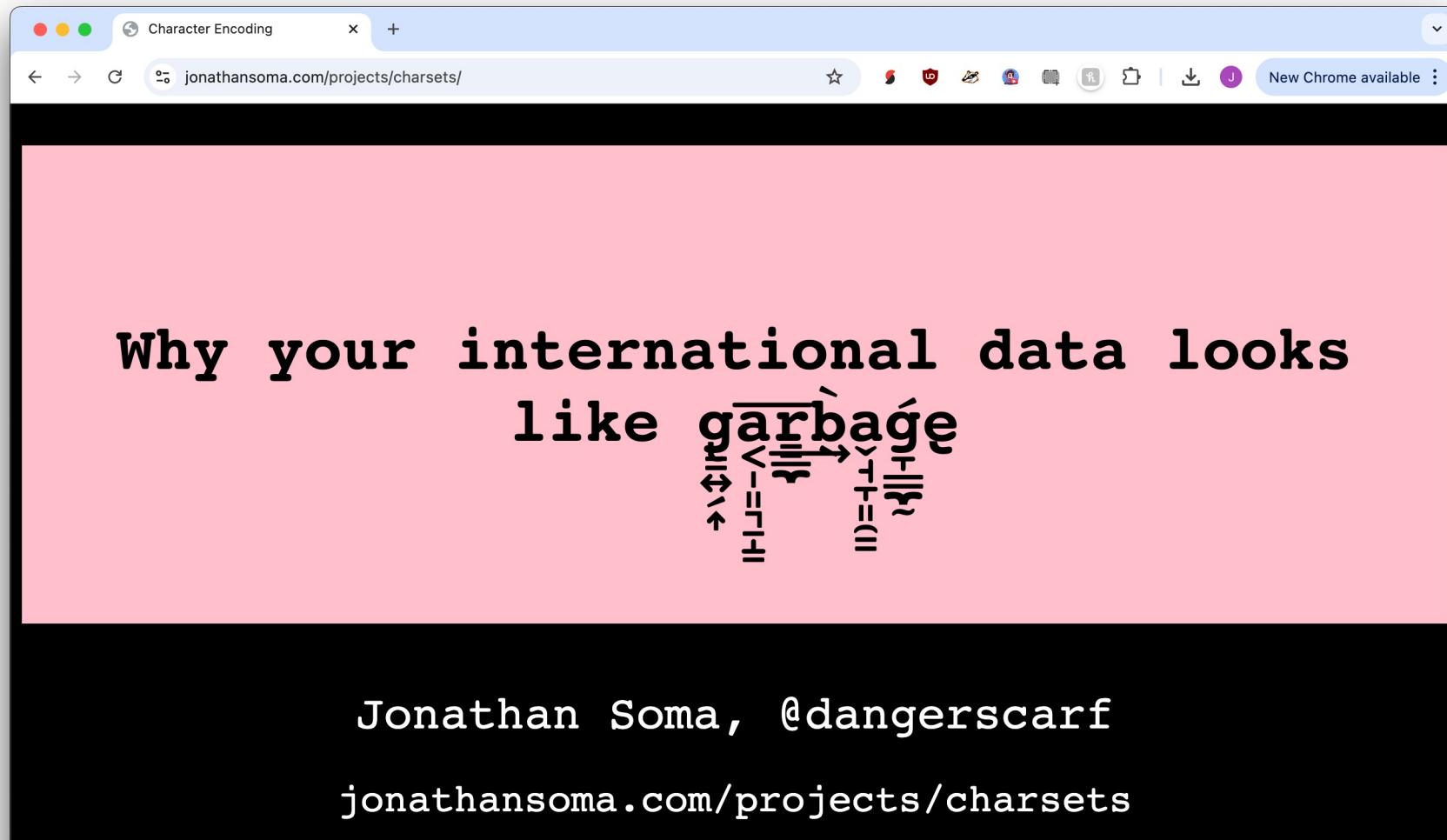
	A	B	C	D	E	F	G
1	name	price	tax	image_url	tags		
2	CORN	298	¥298 (8%tax)	https://www.national-azabu.net/upload/save_image/0607133157_629ed4bd3cb1b.jpeg	PERISHABLE 冷藏		
3	YUZU CITRON	358	¥358 (8%tax)	https://www.national-azabu.net/upload/save_image/1227165628_5fe83e2c165a4.jpeg	PERISHABLE 冷藏		
4	BROCCOLI	398	¥398 (8%tax)	https://www.national-azabu.net/upload/save_image/18_s.jpg	PERISHABLE 冷藏		
5	CUCUMBER 1PC	128	¥128 (8%tax)	https://www.national-azabu.net/upload/save_image/230051000000_s.jpg	PERISHABLE 冷藏		
6	CUCUMBERS 4PCS	498	¥498 (8%tax)	https://www.national-azabu.net/upload/save_image/230034000000_s.jpg	PERISHABLE 冷藏		
		138	¥138 (8%tax)	https://www.national-azabu.net/upload/save_image/2_s.jpg	PERISHABLE 冷藏		
		398	¥398 (8%tax)	https://www.national-azabu.net/upload/save_image/16_s.jpg	PERISHABLE 冷藏		

products_output.csv



Open with Numbers

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



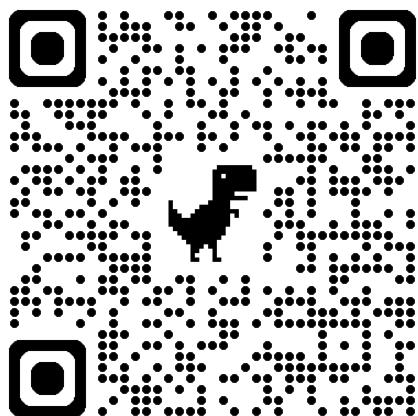
<https://jonathansoma.com/projects/charsets/>

Python basics

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

A tiny Python quiz

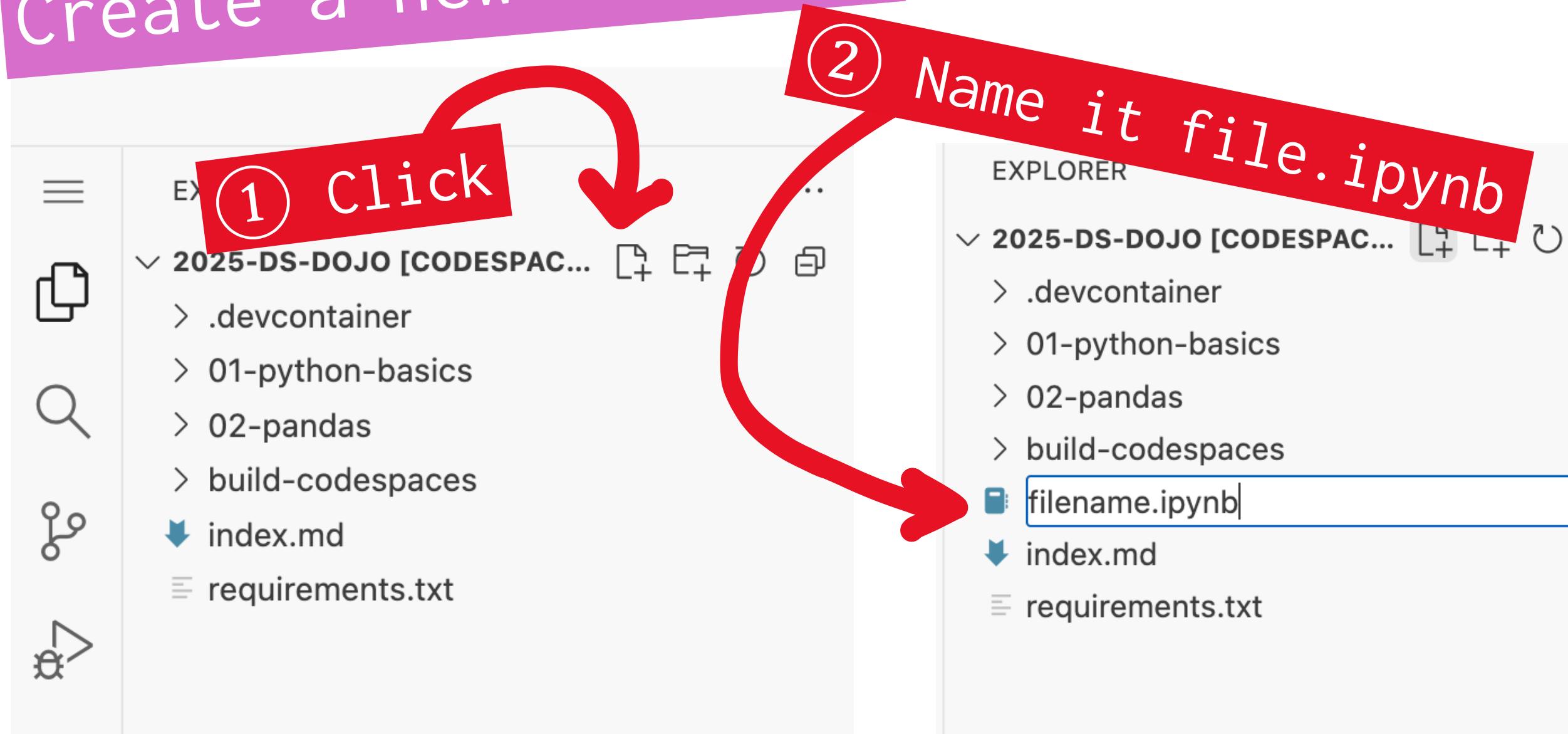
Don't run the
code, just think
and try to answer
the questions



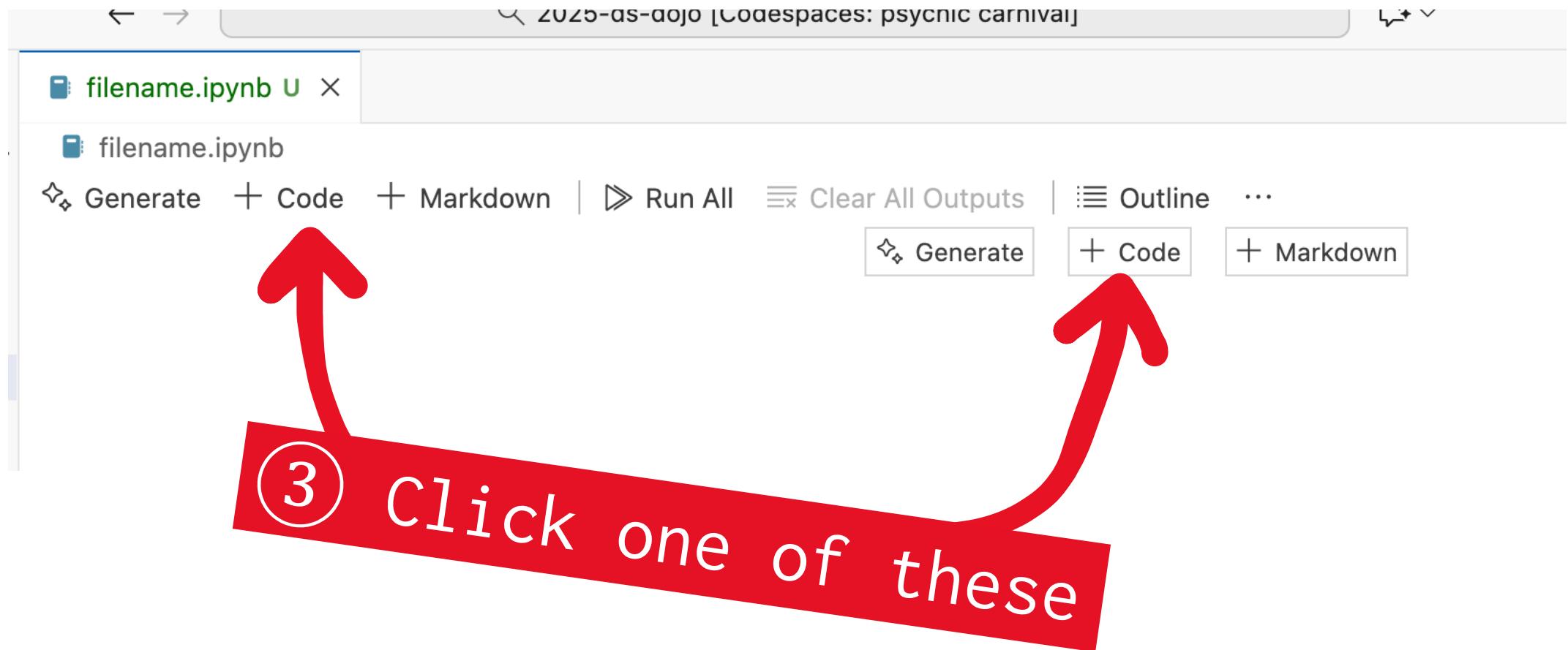
Maybe easier on your
computer than phone?

<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?

Can you build it? This slide presents two challenges related to Japanese dialects:

Challenge 1: Wikipedia Page

The screenshot shows a Japanese Wikipedia page titled "日本語の方言" (Dialects of the Japanese Language). The page content discusses the geographical distribution and characteristics of Japanese dialects. A sidebar on the left provides navigation links for topics like regional distribution, mutual understanding, and historical notes. A note at the top right indicates that the page has been moved (転送) from the Korean Wikipedia page for the topic.

Challenge 2: Map of Japanese Dialects

A large map of Japan illustrating the distribution of various dialects. The map is divided into several main regions labeled in large, bold, black characters:

- 東日本方言 (Tō-nihon hōgen) - Kanto-Khantō (North-Central)
- 西日本方言 (Nishi-nihon hōgen) - Chūgoku-Kansai (Central-West)
- 九州方言 (Kyūshū hōgen) - Kyūshū (South)
- 琉球方言 (Ryūkyū hōgen) - Ryūkyū Islands (South-East)

Each region is further subdivided into smaller dialect areas, each color-coded. A legend at the bottom left defines the colors used for different vowel sounds:

だ	だ + ん
じや	じや + ん
や	や + ん
じゃ	じゃ + ん
や	や + ん
ぢや	ぢや + ん
ぢ	ぢ + ん



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, preceded by a small orange starburst icon. Below the greeting is a text input field asking "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" button 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

Knight Chair in Data Journalism, Columbia University