

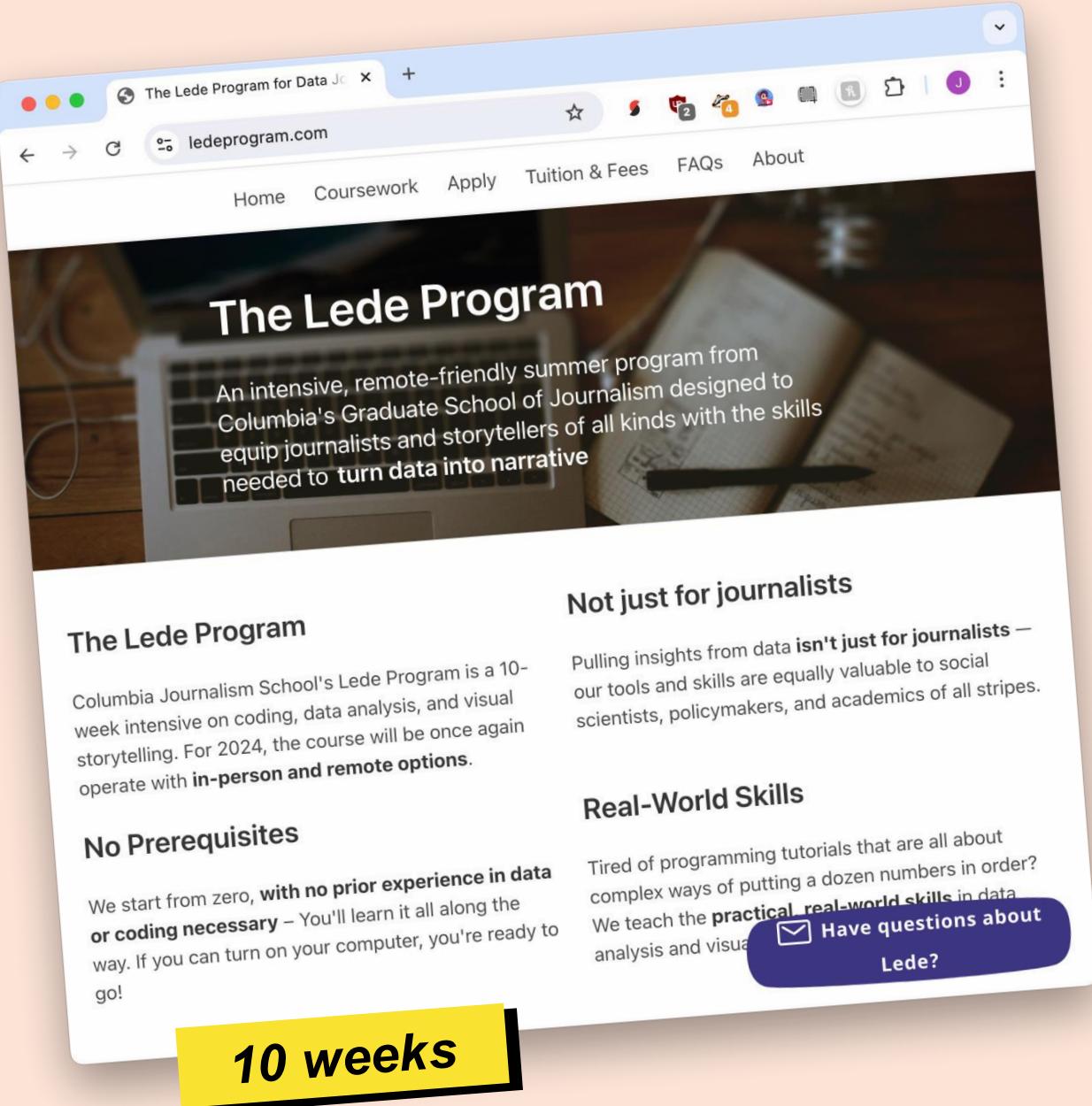
# Advanced Data Journalism

## Part I: Structured Data

Jonathan Soma

Columbia Journalism School

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



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 turn data into narrative

**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**

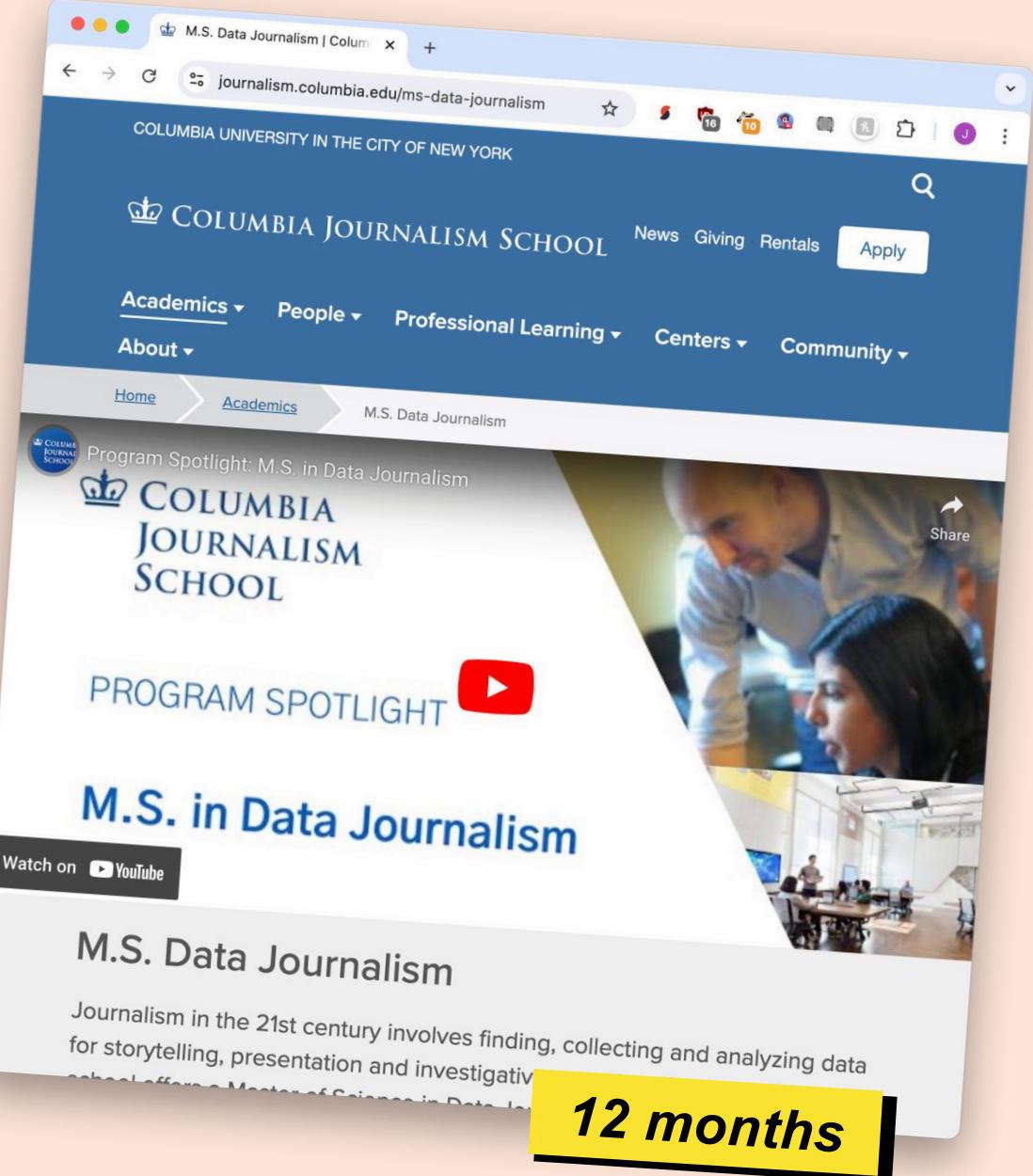
**Not just for journalists**

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, real-world skills** in data analysis and visual

**Lede?**



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 involves finding, collecting and analyzing data for storytelling, presentation and investigation

**12 months**

# What is data journalism?

Nothing special, honestly.

# Krispy Kreme Bets on Big-Box Stores to Stay Fresh

Executives are racing to boost revenues at the struggling doughnut maker as sales slump and shares tumble.

▶ Listen to this article · 9:31 min [Learn more](#)

Share full article



179



Shares of Krispy Kreme, which has been making doughnuts since 1937, have dropped 66 percent over the past year. Scott Olson/Getty Images

The past three decades have been a roller-coaster ride for the Charlotte, N.C., company, aiming to keep investors on a sugar high as it works out how to expand while remaining true to its heritage of serving fresh doughnuts.

Yet its shares have plunged 66 percent in the past year and currently trade around \$3.60, a little more than the cost of a chocolate-iced, cream-filled doughnut in New York City. The company's stock is one of the largest shorts in the market, meaning many investors are betting it could fall even farther.

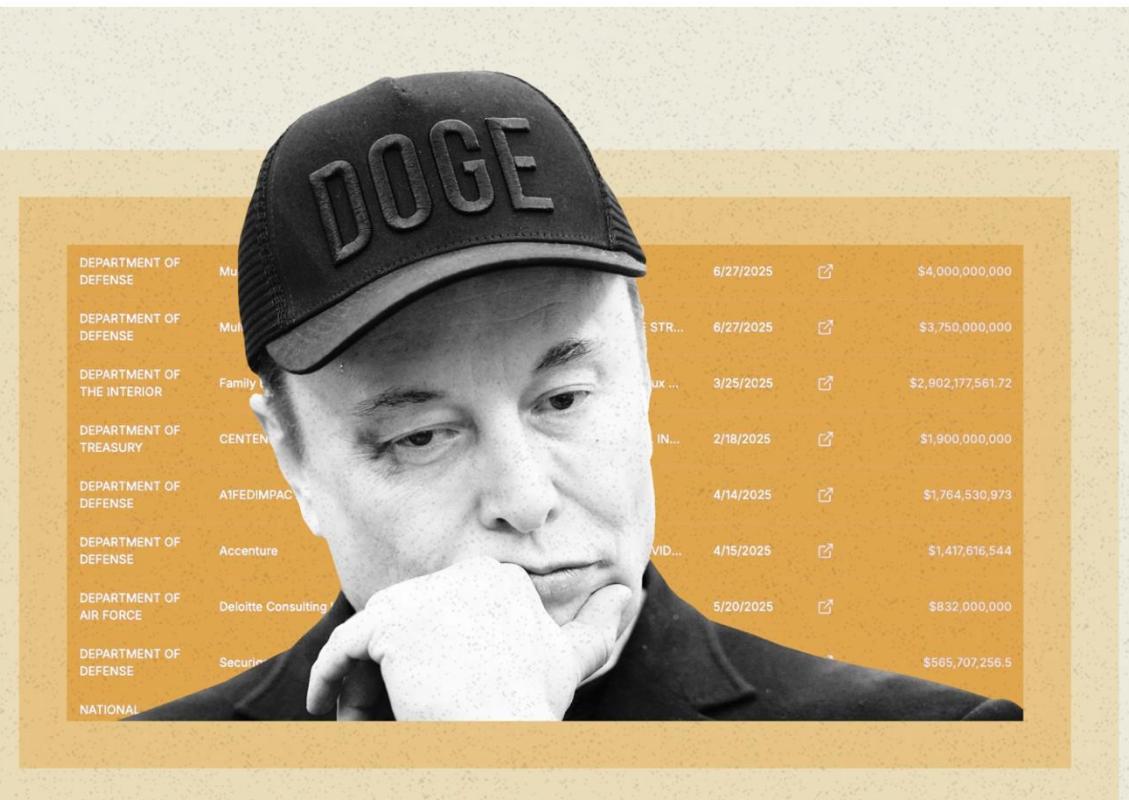
Revenue for the quarter ending in June slipped 13.4 percent. The company said it had lost \$441 million, compared with a loss of \$5 million in the same quarter last year. The drop was largely due to an accounting charge of \$407 million, reflecting the falling value of the chain.

# POLITICO

EUROPE

## Just how much has DOGE exaggerated its numbers? Now we have receipts.

A POLITICO analysis of DOGE data reveals the organization saved less than 5 percent of its claimed savings from nearly 10,100 contract terminations.



## In the West, Lightning Grows as a Cause of Damaging Fires

By John Schwartz and Veronica Penney Oct. 23, 2020

Wildfires in the West caused by lightning have been growing bigger and occurring more frequently. If the weather extremes already brought by climate change are any indication, other parts of the country will start paying a price, too.



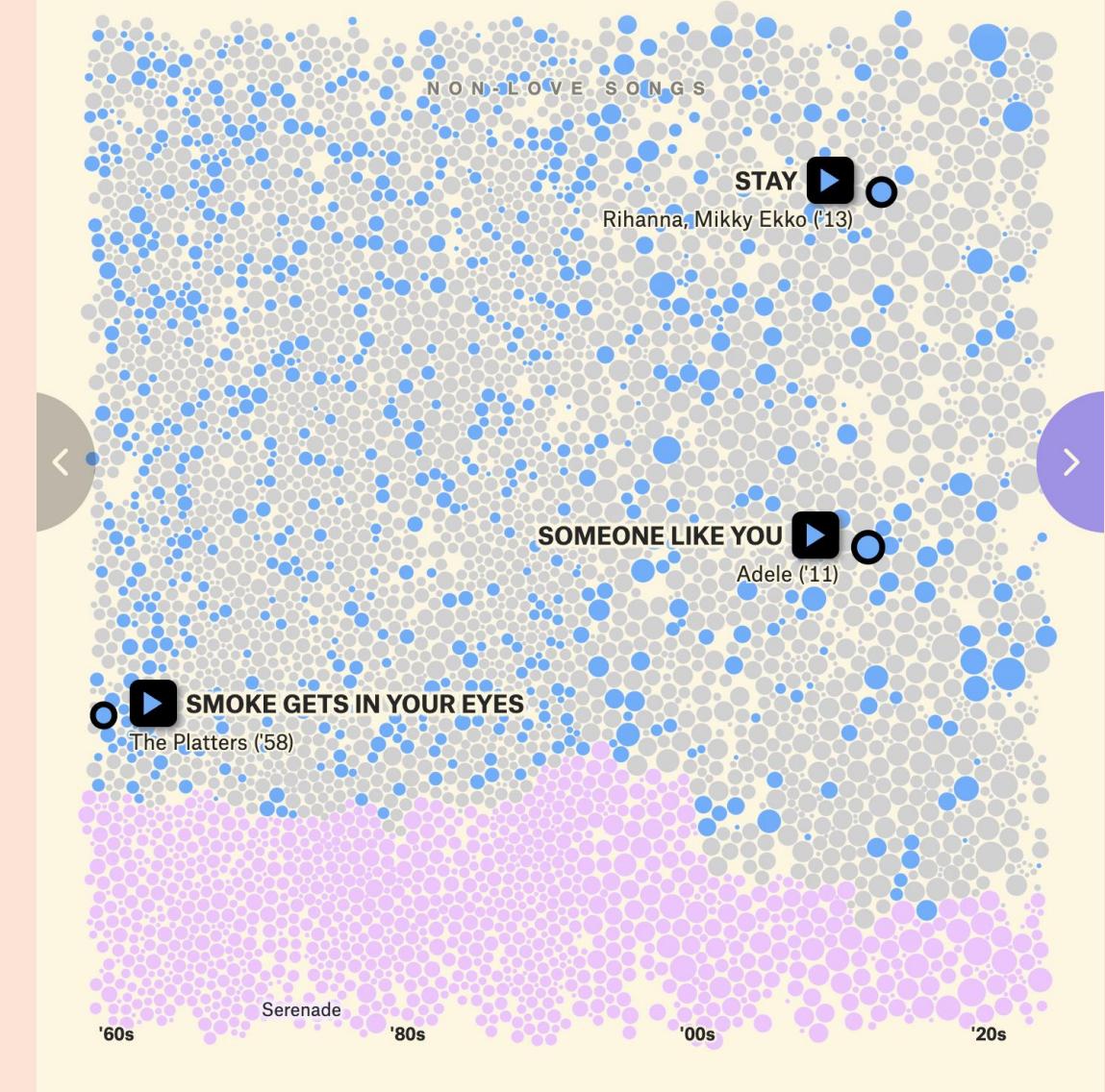
## Deaths in custody and in police operations by nationality between 2020 and 2022

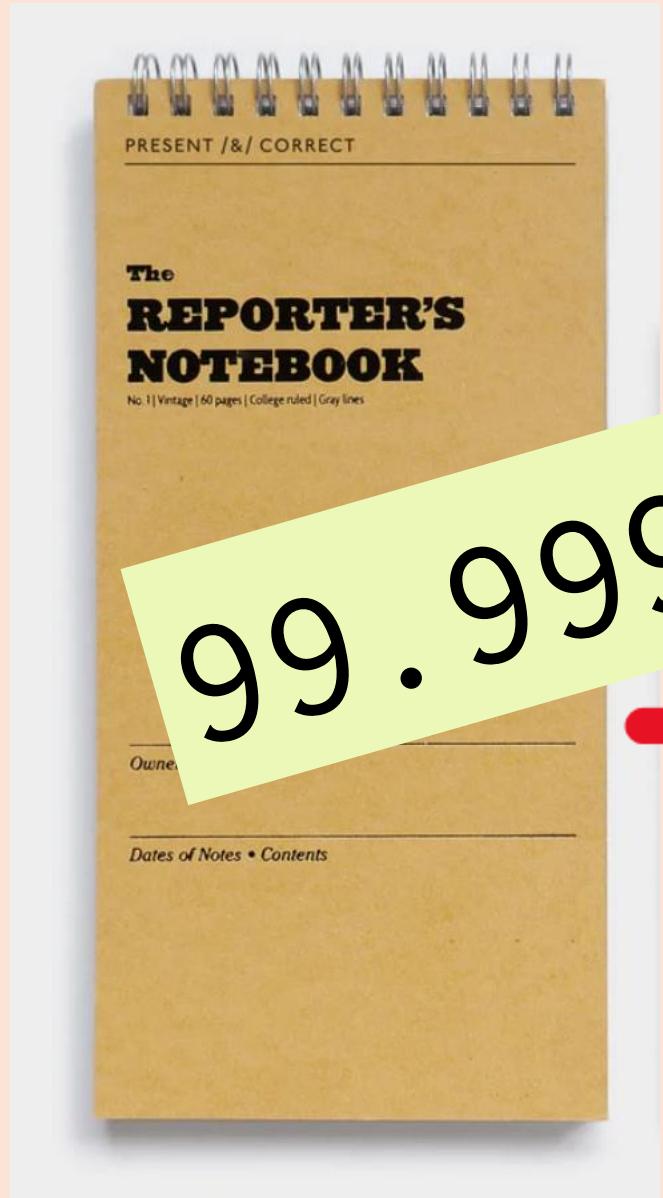


Note: Belgium, Luxembourg, Malta, Croatia, Poland, Lithuania, Bulgaria, Greece, and Cyprus failed to reply to

Chart: Civio

But what happens if you love them, but they just... don't? Maybe you broke up, or maybe it's just unrequited. Let's call this type of love song **Heartache**.

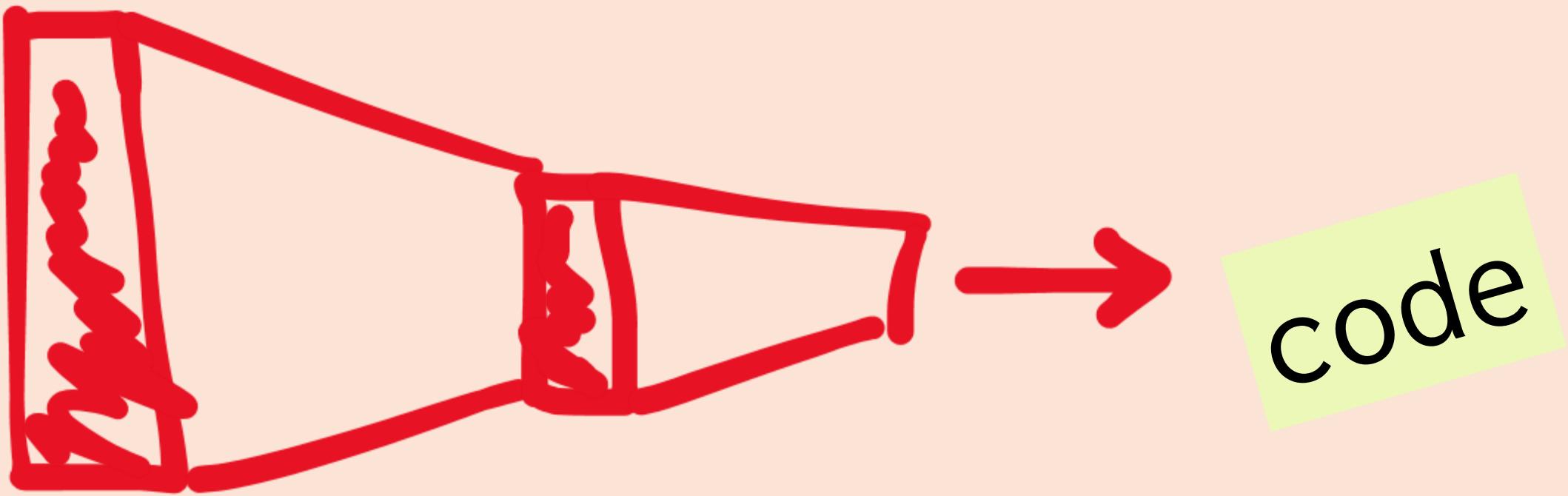




99.999%

A red curved arrow starts from the bottom right of the notebook image and points towards the top right corner of the Google Sheets window. The Google Sheets window displays a spreadsheet titled "Untitled spreadsheet - Google Sheets". The spreadsheet contains a table with columns labeled A, B, C, D, E, F, and G. Column A lists dates, column B lists agencies, column C lists recipients, column D lists values, column E lists savings, column F lists links, and column G lists descriptions. The descriptions in column G are truncated versions of the full URLs, such as "RESEARCHERS WILL PERFORM STUDIES" and "THE PROJECT GOALS ARE TO 1) EVALUATE".

A	B	C	D	E	F	G
1	date	agency recipient				
2	5/8/2025	Environmental P North Carolina A&T State University	1340704	216661	<a href="https://usaspending.gov">https://usaspending.gov</a>	RESEARCHERS WILL PERFORM STUDIES
3	5/8/2025	Environmental P The University of North Carolina at Chapel Hill dba Office of Contracts and	1894823	78653	<a href="https://usaspending.gov">https://usaspending.gov</a>	RESEARCHERS WILL USE LONGITUDINAL
4	5/8/2025	Environmental P GEORGIA TECH RESEARCH CORP	1000000	476179	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE PROJECT GOALS ARE TO 1) EVALUATE
5	5/8/2025	Environmental P The University of North Carolina at Chapel Hill dba Office of Contracts and	799833	73805	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:
6	5/8/2025	Environmental P University of Wisconsin - Madison	998049	567356	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE PURPOSE OF THIS PRO
7	5/8/2025	Environmental P MICHIGAN STATE UNIVERSITY	1600000	1561741	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE GOALS OF THIS PROJ
8	5/8/2025	Environmental P The Ohio State University	999995	119525	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE GOAL OF THE RESEARCH IS TO DEC
9	5/8/2025	Environmental P University of Wisconsin - Madison	1362435	842529	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE PURPOSE OF THE PRO
10	5/8/2025	Environmental P The Little Village Environmental Justice Organization	196034	0	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE PURPOSE OF THIS AME
11	5/7/2025	Environmental P Water Research Foundation	1497907	700941	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE PURPOSE OF THE PROPOSED RESEA
12	5/7/2025	Environmental P Oregon State University	2350211	2064835	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE PURPOSE OF THE PROJ
13	5/7/2025	Environmental P TEXAS A&M ENGINEERING EXPERIMENT STATION	1600000	1555041	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE GOALS OF THIS PROJE
14	5/7/2025	Environmental P Kansas Department of Health & Environment	1000000	957738	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE AGREEMENT PROVIDES
15	5/7/2025	Environmental P Regents of the University of California, The	549940	134504	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE PURPOSE OF PROPOSED RESEARCH IS
16	5/7/2025	Environmental P University of California, Davis	799660	237460	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:DESCRIPTION:
17	5/7/2025	Environmental P University of California, Los Angeles	798825	341961	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE GOAL OF THE PROPOSED
18	5/7/2025	Environmental P The Regents of the University of CA - Riverside	576683	478108	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THIS PROJECT WILL STRATEG
19	5/7/2025	Environmental P The Regents of the University of CA - Berkeley	1999998	946636	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE PURPOSE OF THIS PRO
20	5/7/2025	Environmental P The University of Texas at Dallas	24999	105	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:RECIPIENT PROPOSES TO EXA
21	5/7/2025	Environmental P University of Texas at Austin	36322	22267	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE GOAL OF THE RESEA
22	5/7/2025	Environmental P The Water Research Foundation	3245999	1803049	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THIS PROJECT AIMS TO ENHA
23	5/7/2025	Environmental P The Rand Corporation	1349997	952166	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE GOAL OF THE RESEA
24	5/7/2025	Environmental P Physicians, Scientists, and Engineers for Sustainable and Healthy Energy Inc	1350000	538622	<a href="https://usaspending.gov">https://usaspending.gov</a>	DESCRIPTION:THE GOAL OF THIS RESEA
25	5/7/2025	Environmental P Oregon State University	1984400	99712	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE PROJECT PROPOSES TO STUDY THE TO
26	5/6/2025	Environmental P Vanderbilt University	790352	0	<a href="https://usaspending.gov">https://usaspending.gov</a>	THIS PROJECT WILL FILL A FUNDAMENTAL GA
27	5/6/2025	Environmental P University of Florida	798193	0	<a href="https://usaspending.gov">https://usaspending.gov</a>	THIS PROJECT AIMS TO UNDERSTAND HOW FI
28	5/6/2025	Environmental P Emory University	1345022	654617	<a href="https://usaspending.gov">https://usaspending.gov</a>	THE GOAL OF THIS PROJECT IS TO UNDERSTA



**0.0001% of 0.0001%**

We're going to focus on that tiny  
sliver of coding

What is “structured data?”

Also nothing fancy: spreadsheets. Rows and columns.

A	B	C	D	E
e	agency	recipient	value	savings
5/8/2025	Environmental P	North Carolina A&T State University	1340704	21666
5/8/2025	Environmental P	The University of North Carolina at Chapel Hill dba Office of Contracts and	1894823	78653
5/8/2025	Environmental P	GEORGIA TECH RESEARCH CORP	1000000	476179
5/8/2025	Environmental P	The University of North Carolina at Chapel Hill dba Office of Contracts and	799833	73803
5/8/2025	Environmental P	University of Wisconsin - Madison	998049	567356
5/8/2025	Environmental P	MICHIGAN STATE UNIVERSITY	1600000	156174
5/8/2025				95
5/8/2025				85
5/8/2025				34
5/7/2025				70094
5/7/2025				2064835
5/7/2025		INEERING EXPERIMENT STATION	1600000	155504
5/7/2025		In of Health & Environment	1000000	957738
5/7/2025	Environmental P	Regents of the University of California, The	549940	134504
5/7/2025	Environmental P	University of California, Davis	799660	237460
5/7/2025	Environmental P	University of California, Los Angeles	798825	34196
5/7/2025	Environmental P	The Regents of the University of CA - Riverside	576683	478108

## 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
12/12/2021	Cheating, Plagiarism	English 200: Literary Texts and Contexts

Academic\_Integrity\_Violati...025.pdf



# Why don't we just use AI?

Well, we will. It does two things:  
works perfectly and lies to your face.

<https://bit.ly/birn-2025-data>

download this



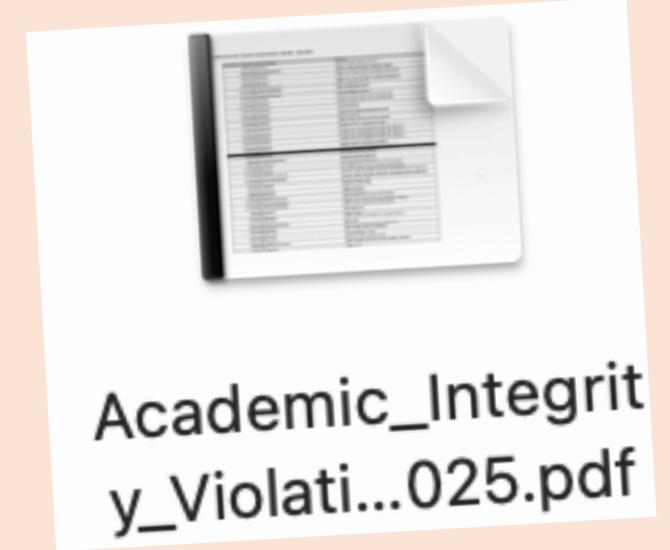
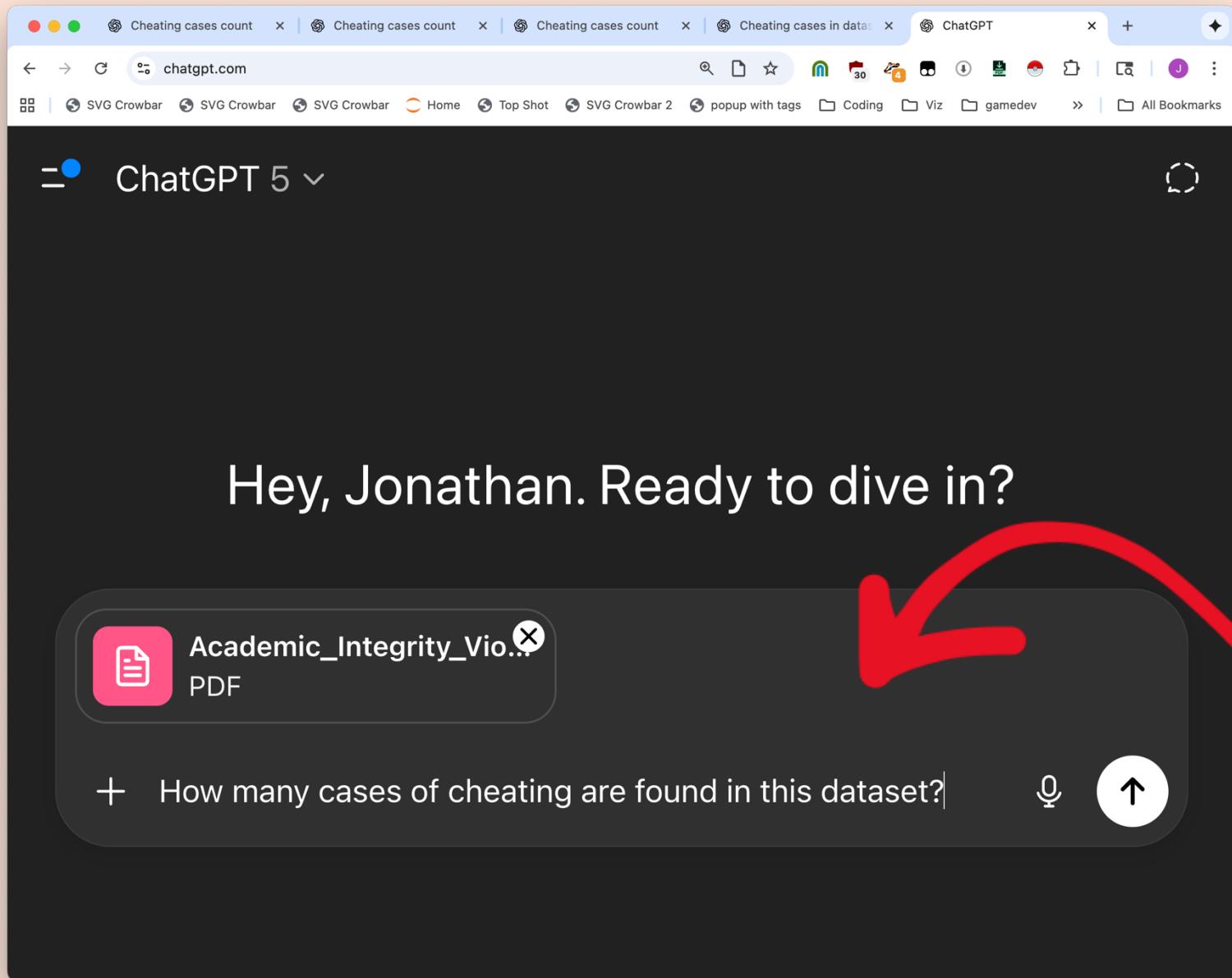
## Why can't we just use AI for everything???

Exploring what AI can and can't do when working with structured data

📦 Data: [00-cheating-data.zip](#)

### Links:

- [ChatGPT](#) Your AI best friend
- [MuckRock](#) A website for filing (and collecting) FOI responses



Cheating cases in dataset

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

SVG Crowbar 42 SVG Crowbar 2 Home Top Shot SVG Crowbar 2 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 with other violations like plagiarism or facilitation).

The dataset contains \_\_\_\_\_ cases of cheating in 2025.

Would you like me to break that number down further—for example,

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

or by semester/year?

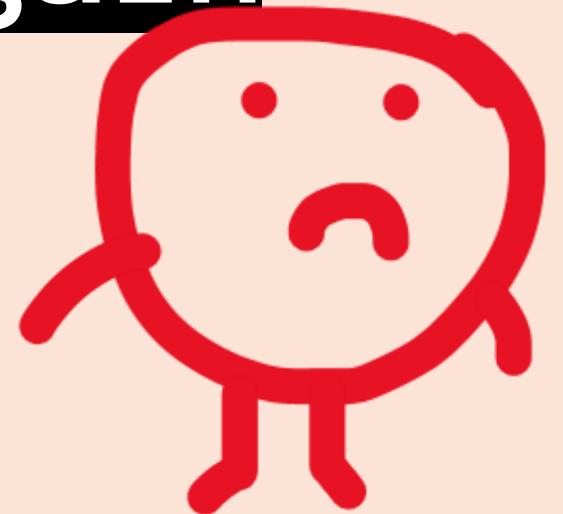
+ Ask anything

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**

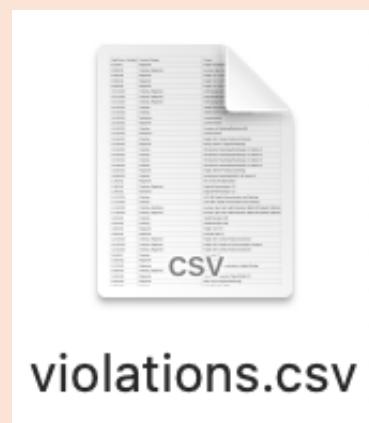
The dataset contains **143 cases of cheating** (this includes both cheating alone and cases where cheating is listed alongside other violations, like “Cheating, Plagiarism”).

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.

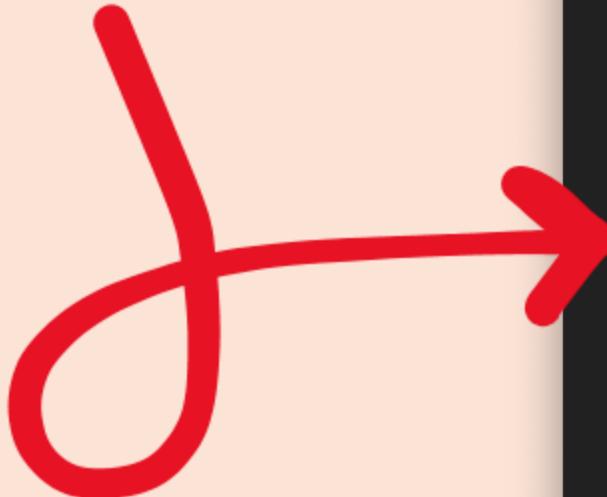
**It will lie to your face  
and cite sources while it  
does so.**

We'll talk more about why later

A screenshot of Microsoft Excel showing the "violations.csv" data imported into a spreadsheet. The data is displayed in a table with columns A through I. Column A contains the row numbers and the header "Date/Time of Incident". Column B contains the incident descriptions. Column C contains the course names. The data is as follows:

	Date/Time of Incident	Course
1	8/3/21	Conduct Charges
2	9/30/21	Plagiarism
3	9/30/21	Cheating, Plagiarism
4	9/30/21	Plagiarism
5	9/30/21	Plagiarism
6	9/30/21	Plagiarism
7	10/11/21	Cheating, Plagiarism
8	10/11/21	Cheating, Plagiarism
9	10/11/21	Cheating, Plagiarism
10	10/13/21	Cheating
11	10/13/21	Cheating
12	10/15/21	Facilitation
13	10/15/21	Plagiarism
14	10/15/21	Cheating
15	10/15/21	Facilitation
16	10/16/21	Cheating
17	10/18/21	Plagiarism
18	10/19/21	Cheating
19	10/19/21	Cheating
20	10/19/21	Cheating
21	10/19/21	Cheating
22	10/26/21	Plagiarism
23	10/26/21	Cheating
24	11/5/21	Plagiarism
25	11/8/21	Cheating, Plagiarism
26	11/8/21	Facilitation

	Date/Time of Incident	Course
1	8/3/21	Conduct Charges
2	9/30/21	Plagiarism
3	9/30/21	Cheating, Plagiarism
4	9/30/21	Plagiarism
5	9/30/21	Plagiarism
6	9/30/21	Plagiarism
7	10/11/21	Cheating, Plagiarism
8	10/11/21	Cheating, Plagiarism
9	10/11/21	Cheating, Plagiarism
10	10/13/21	Cheating
11	10/13/21	Cheating
12	10/15/21	Facilitation
13	10/15/21	Plagiarism
14	10/15/21	Cheating
15	10/15/21	Facilitation
16	10/16/21	Cheating
17	10/18/21	Plagiarism
18	10/19/21	Cheating
19	10/19/21	Cheating
20	10/19/21	Cheating
21	10/19/21	Cheating
22	10/26/21	Plagiarism
23	10/26/21	Cheating
24	11/5/21	Plagiarism
25	11/8/21	Cheating, Plagiarism
26	11/8/21	Facilitation



CSV

violations.csv

ChatGPT 5

What's on your mind today?

violations.csv  
Spreadsheet

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

This image shows a screenshot of a Mac OS X desktop with a light beige background. On the left, there is a white rectangular area containing a CSV file icon (a sheet of paper with data) and the text "violations.csv". A large, hand-drawn style red arrow originates from the CSV file icon and points towards the ChatGPT input field on the right. The right side of the image shows a dark-themed ChatGPT window. The window title bar says "ChatGPT" and the address bar says "chatgpt.com". Below the address bar, the URL "chatgpt.com" is visible. The main content area of the window has a dark background with the text "What's on your mind today?" in white. Below this, there is a message card with a green icon representing a spreadsheet, the text "violations.csv", and the word "Spreadsheet". At the bottom of the message card, there is a question "How many cases of cheating are found in this dataset?" followed by a microphone icon and an upload icon.

# 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



# What's the role of code in an AI world?

- Knowing what's possible instead of only knowing what the AI knows
- You're smarter than AI (...in some ways)
- It's teamwork
- Communication with other human beings

Your editor is never going  
to read a ChatGPT  
conversation and trust it.

You'll probably lose the link anyway.

READ ▾ LISTEN ▾ WATCH ▾

Reveal

ABOUT US SHOP DONATE



KEPT OUT

# For people of color, banks are shutting the door to homeownership

by Aaron Glantz and Emmanuel Martinez February 15, 2018



# KEPT OUT



also, giant data

An official website of the United States government [Here's how you know](#)

**FFIEC**

HOME FILING DATA BROWSER ▾ DATA PUBLICATION ▾ TOOLS ▾ DOCUMENTATION ▾

## HMDA Data Publication

The HMDA data and reports are the most comprehensive publicly available information on mortgage market activity. The data and reports can be used along with the [Census](#) demographic information for data analysis purposes. Available below are the data and reports for HMDA data collected in or after 2017. For HMDA data and reports for prior years, visit <https://www.ffiec.gov/hmda/hmdaproducts.htm>.

For information about changes to HMDA Publications visit the [Updates and Notes](#) page.

Select a year

[2024](#) [2023](#) [2022](#) [2021](#) [2020](#) [2019](#) [2018](#) [2017](#)

---

**Dynamic Datasets**

**[Modified Loan/Application Register \(LAR\)](#)**  
The modified LAR provides loan-level data for an individual financial institution, as modified by the Bureau to protect applicant and borrower privacy.

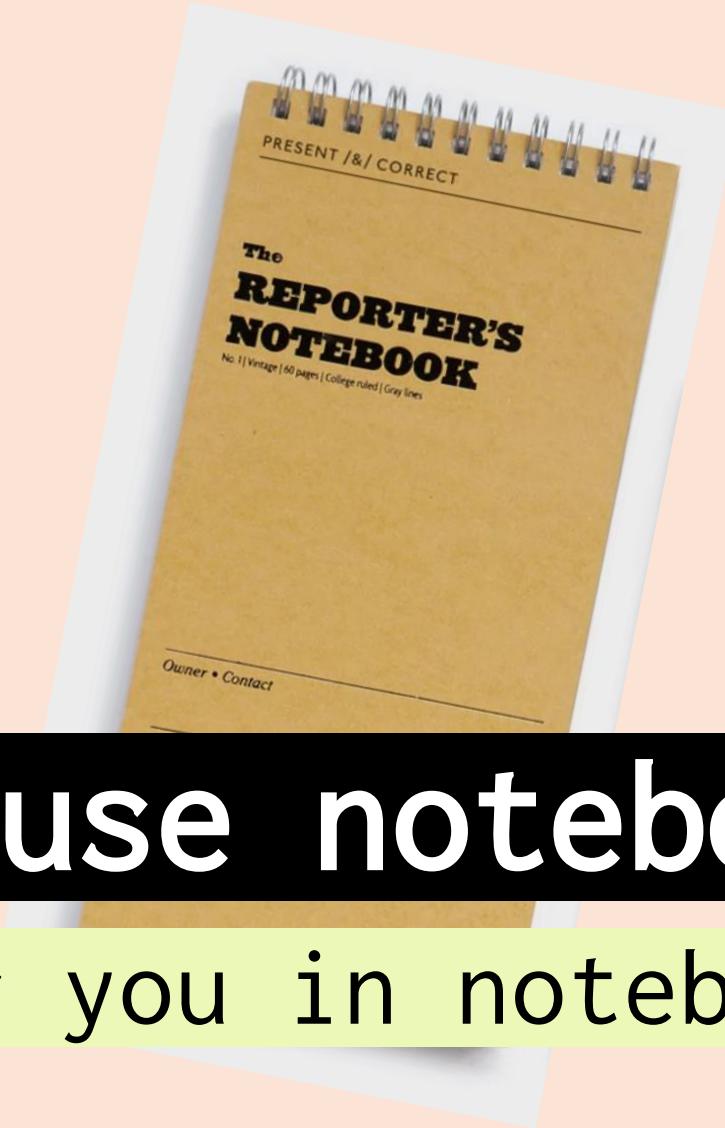
→ **Update Frequency:** Upon Institution resubmission

**[Dynamic National Loan-Level Dataset](#)**

**[Snapshot National Loan-Level Dataset](#)**  
The snapshot files contain the national HMDA datasets as of a fixed date for all HMDA reporters, as modified by the Bureau to protect applicant and borrower privacy.

→ **Data Freeze Date:** May 19, 2025

→ **Update Frequency:** Does not update. Generated a few months after filing deadline



# Data people use notebooks

(AI can write for you in notebooks too)

Reveal Mortgage Analysis - L x +

colab.research.google.com/github/littlecolumns/ds4j-notebooks/blob/master/reveal-mortgages/notebook... 🔍 ⭐

SVG Crowbar 2 Home Top Shot SVG Crowbar 2 popup with tags Coding Viz gamedev bread lede

All Bookmarks

Share J

CO PRO Reveal Mortgage Analysis - Logistic Regression using statsmodels formulas.ipynb

File Edit View Insert Runtime Tools Help

Connect ▾

Commands + Code + Text ▶ Run all ▾ Copy to Drive

## Deal with categorical variables

Let's go ahead and take a look at our categorical variables:

- Applicant sex (male, female, na)
- Applicant race
- Mortgage agency
- Co-applicant (yes, no, unknown)

Before we do anything crazy, let's use the codebook to turn them into strings.

- Tip: We already did this with the `co_applicant` column, you only need to do the rest
- Tip: Just use `.replace`

```
▶ mortgage.applicant_sex = mortgage.applicant_sex.replace({  
    1: 'male',  
    2: 'female',  
    3: 'na'  
})  
mortgage.applicant_race = mortgage.applicant_race.replace({  
    1: 'native_amer',  
    2: 'asian',  
    3: 'black',  
    4: 'hawaiian',  
    5: 'other'
```

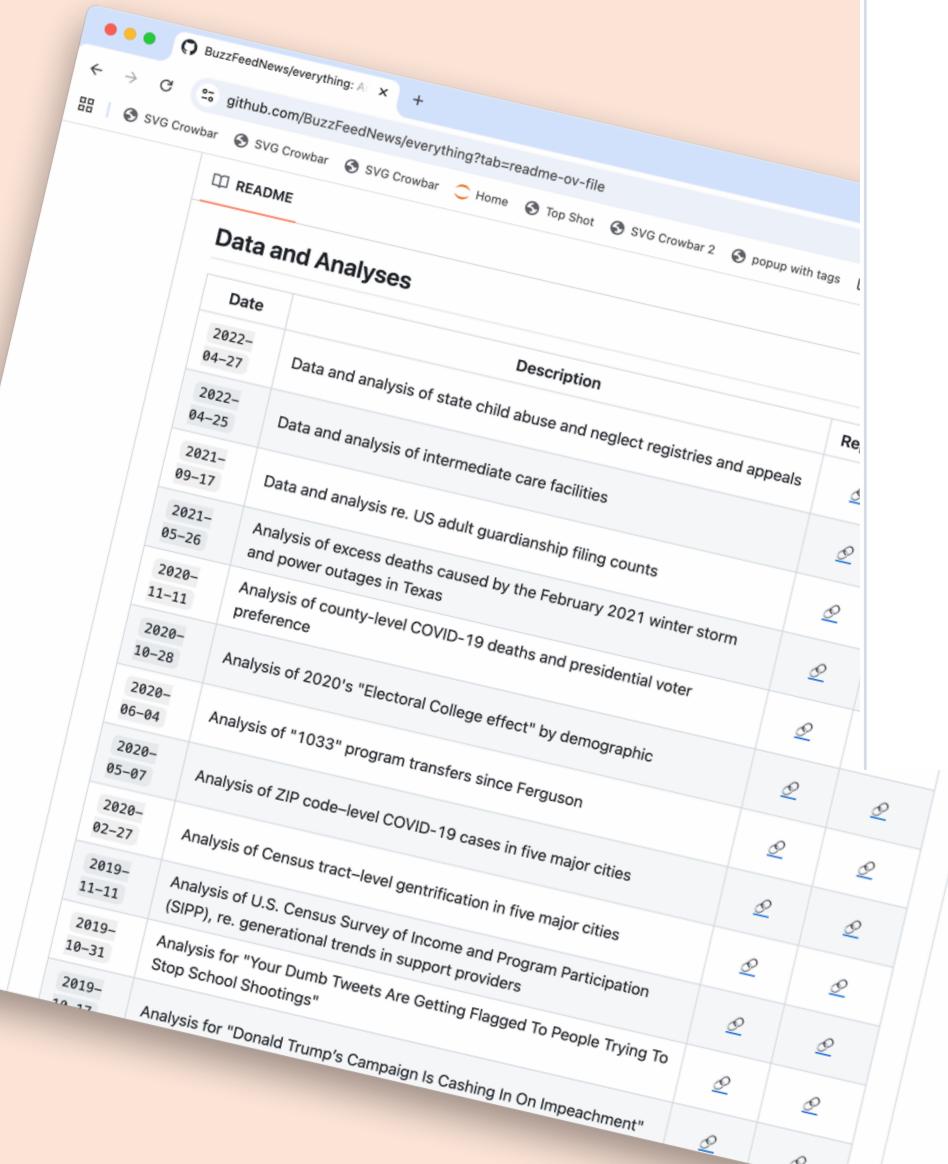
## 📄 README

# BuzzFeedNews/everything

An index of all our open-source data, analysis, libraries, tools, and guides.

## Table of Contents

- [Data and Analyses](#)
- [Standalone Datasets](#)
- [Libraries and Tools](#)
- [Guides](#)



BuzzFeedNews/everything: A X +

2022-04-registries / notebooks / substantiations / CA\_subs.ipynb ↑ Top

Preview Code Blame

```
return row.apply( lambda x: x/total )
```

In [7]:

```
yearly_race = (
    df
    .groupby(['year', "race"])
    .sum()
    .unstack()
    .apply( percent, axis = 1)
)

yearly_race
```

Out [7]:

		count					
year	race	asian	black	hispanic	other	pac	white
2009		0.017784	0.129772	0.488668	0.071739	0.014508	0.277529
2010		0.023724	0.132140	0.498675	0.068125	0.016103	0.261233
2011		0.022403	0.146665	0.496541	0.070762	0.016310	0.247319
2012		0.020894	0.145136	0.480004	0.089418	0.014940	0.249607
2013		0.021145	0.132895	0.482005	0.093124	0.015964	0.254866
2014		0.023552	0.136214	0.451695	0.108611	0.011851	0.268077
2015		0.021110	0.144731	0.445386	0.119143	0.014073	0.255557
2016		0.027168	0.139125	0.445887	0.110614	0.014629	0.262577
2017		0.022707	0.130237	0.443745	0.110900	0.011427	0.280984
2018		0.030452	0.140226	0.342246	0.166964	0.014112	0.306001
2019		0.026297	0.135039	0.170576	0.253305	0.014925	0.399858
2020		0.028760	0.138300	0.112683	0.266698	0.014616	0.438944
2021		0.023730	0.138807	0.079168	0.258085	0.018690	0.481520

In [8]:

```
# average yearly race
yearly_race.mean().to_frame("")
```

Out [8]:

2019-10-17 Analysis for "Donald Trump's Campaign Is Cashing In On Impeachment"

It's Like A Leech On Me": Child Abuse Registries Punish Unsuspecting Parents Of Color

Millions of parents have been placed on these lists, often for the vague offense of "neglect." The consequences can last for decades. A BuzzFeed News investigation.

Scott Pham  
BuzzFeed News Reporter

Posted on April 27, 2022 at 12:31 pm

X f e View All 101 Comments

When Nzinga Terrell-Brown took a job as a teacher's assistant in 2018, she thought it was the start of a new life. For years, Terrell-Brown, a college graduate with a degree in English, had worked in daycare centers and group homes, carrying the dream of one day becoming a teacher. Now, she hoped, she was on her way.

Less than three months later, she was fired.

Subscribe to BuzzFeed Daily Newsletter ▾

## We Trained A Computer To Search For Hidden Spy Planes. This Is What It Found.

From planes tracking drug traffickers to those testing new spying technology, US airspace is buzzing with surveillance aircraft operated for law enforcement and the military.



Peter Aldhous  
BuzzFeed News Reporter

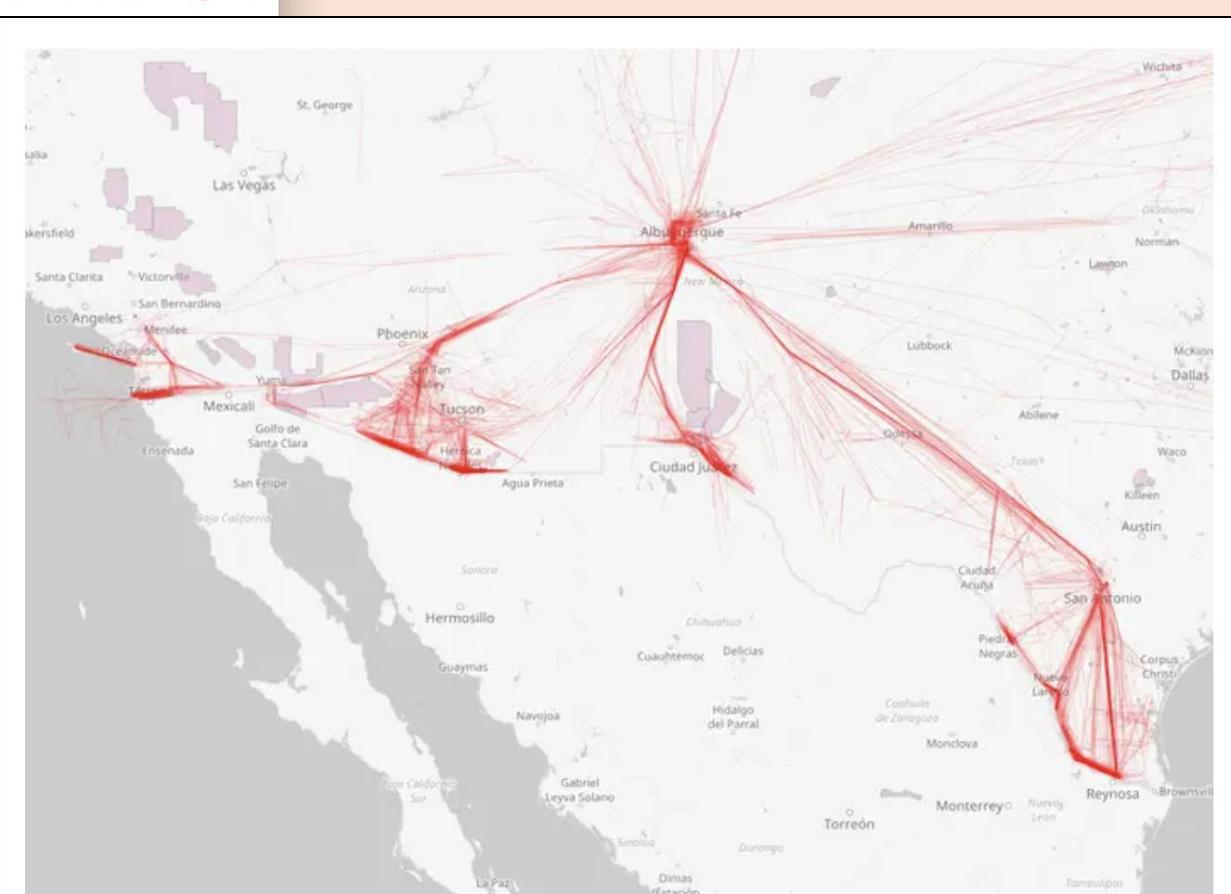
Updated on August 8, 2017 at 4:47 pm  
Posted on August 7, 2017 at 6:33 pm



[View Comments](#)

A secret spy plane operated by the US Marshals hunted drug cartel kingpins in Mexico. A military contractor that tracks terrorists in Africa is also flying surveillance aircraft over US cities. In two stories published last week, BuzzFeed News revealed the activities of aircraft that their operators didn't want to discuss.

These discoveries came not from tip-offs from anonymous sources, but by training a computer to recognize known spy planes, then setting it loose on



Flights by Global Geo Mapping planes near the US-Mexico border, January 2015 to July 2017.

Peter Aldhous / BuzzFeed News / Via flightradar24.com

2017-  
08-07

## Data and analysis for "BuzzFeed News Trained A Computer To Search For Hidden Spy Planes. This Is What We Found."



## Feature engineering

Using the same data, we had [previously reported](#) on flights of spy planes operated by the FBI and the Department of Homeland Security (DHS), and reasoned that it should be possible to train a machine learning algorithm to identify other aircraft performing similar surveillance, based on characteristics of the aircraft and their flight patterns.

First we filtered the data to remove planes registered abroad, based on their `adshex` code, common commercial airliners, based on their `type`, and aircraft with fewer than 500 transponder detections.

Then we took a random sample of 500 aircraft and calculated the following for each one:

- `duration` of each flight segment recorded by Flightradar24, in minutes.
- `boxes` Area of a rectangular bounding box drawn around each flight segment, in square kilometers.

Finally, we calculated the following variables for each of the aircraft in the larger filtered dataset:

- `duration1`, `duration2`, `duration3`, `duration4`, `duration5` Proportion of flight segment durations for each plane falling into each of five quantiles calculated from `duration` for the sample of 500 planes. The proportions for each aircraft must add up to 1; if the durations of flight segments for a plane closely matched those for a typical plane from the sample, these numbers would all approximate to 0.2; a plane that mostly flew very long flights would have large decimal fraction for `duration5`.
- `boxes1`, `boxes2`, `boxes3`, `boxes4`, `boxes5` Proportion of bounding box areas for each plane falling into each of five quantiles calculated from `boxes` for the sample of 500 planes.
- `speed1`, `speed2`, `speed3`, `speed4`, `speed5` Proportion of speed values recorded for the aircraft falling into each of five quantiles recorded for `speed` for the sample of 500 planes.
- `altitude1`, `altitude2`, `altitude3`, `altitude4`, `altitude5` Proportion of altitude values recorded for the aircraft falling into each of five quantiles recorded for `altitude` for the sample of 500 planes.
- `steer1`, `steer2`, `steer3`, `steer4`, `steer5`, `steer6`, `steer7`, `steer8` Proportion of steer values for each aircraft falling into bins set manually, after observing the distribution for the sample of 500 planes, using the breaks: -180, -25, -10, -1, 0, 1, 22, 45, 180.
- `flights` Total number of flight segments for each plane.
- `squawk_1` Squawk code used most commonly by the aircraft.
- `observations` Total number of transponder detections for each plane.
- `type` Aircraft manufacturer and model, if identified, else unknown.

## Machine learning, using random forest algorithm

For the machine learning, we selected the [random forest](#) algorithm, popular among data scientists for classification tasks. (See [this tutorial](#) for background on running the random forest in R.)

As training data, drawn from `planes_features.csv`, we used 97 fixed-wing FBI and DHS planes from our previous story, given a `class` of `surveil`, and a random sample of 500 other planes, given a `class` of `other`.

Data identifying these planes is in the file `train.csv`.

```
# load required packages
library(readr)
library(dplyr)
library(randomForest)

# load planes_features data
planes <- read_csv("data/planes_features.csv")

# convert type to integers, as new variable type2, so it can be used by the random
# forest algorithm
planes <- planes %>%
  mutate(type2=as.integer(as.factor(type)))

# load training data and join to the planes_features data
train <- read_csv("data/train.csv") %>%
  inner_join(planes, by="adshex")
```

We then trained the random forest algorithm using this data.

```
# set seed for reproducibility of model fit
set.seed(415)

# train the random forest
```

Misinformation on TikTok: How 'Documented' Examined Hundreds of Videos in Different Languages

Author: Lam Thuy Vo  
GRANTEE

JOURNALIST RESOURCE | JANUARY 10, 2025

Translate page with Google

Home / Resources

pulitzercenter.org/misinformation-tiktok-how-documented-examined-hu...

All Bookmarks

## tiktok-analysis-pipeline / notebooks /



lamthuyvo initial commit

### Name



..



00-scraper-tiktok-links-extraction.ipynb



00-scraper-yt-dlp-tiktok-downloader.ipynb



01-autotranscribe-whisper-solution.ipynb



02-topics-clustering-gensim-solution.ipynb

# We only have two goals

- Learn to run Python
- Learn one fundamental tool (pandas)
- That's it!!!!

# Two worst parts of a data journalist's life

~~Installing software~~, and every day  
after that

<https://bit.ly/birn-2025-data>

## Data analysis basics with pandas

Pandas is the most common tool that programmers use for analyzing data. And if that wasn't good enough for you: AI uses it, too!

click this



Live coding worksheet



Completed version



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



Data: [01-pandas-data.zip](#)

**The reason why you do data  
journalism is scale.**

Number of rows, number of files, number  
of differences, across time, etc.

Everything uses pandas.

Everyone uses pandas. ChatGPT

uses pandas, it can help you

with anything.

The screenshot shows a web browser window for the iMEDD:content website. The header includes the iMEDD logo and language options (EA, EN). Below the header is a navigation bar with links for STORIES, OPEN DATA, APPLICATIONS, and TOOLS & PRACTICES. A large image of a residential building is displayed above the main content area. The main article title is "Residential property accounts for 38% of personal assets ‘lost’ in auctions linked to banks". The date of publication is 14.06.2021, and the author is Thanasis Troboukis, Kelly Kiki. The text discusses the volume of auctions, the complexity of debt repayment, and the involvement of banks in foreclosures. A large paragraph begins with "As of 12 June 2021, a total of 91 744 auctions had been posted on the online auction platform [eauction.gr](#), following the country's launch of the service in September 2017. The iMEDD Lab has looked into almost all auctions posted on the platform up to 31 May 2021. Of these, 51.2% have been completed, 41.8% are suspended and 1.2% have been canceled, while 5205 auctions (5.8%) have already been posted and are scheduled to take place by February 2022. For reference, according to publicly available data up to May 2021, more than 2100 posted auctions are scheduled for June 2021 and more than 1300 posts concern auctions scheduled for July 2021."

#### DATA ANALYSIS

## Residential property accounts for 38% of personal assets “lost” in auctions linked to banks

14.06.2021

Thanasis Troboukis, Kelly Kiki

In the last 3.5 years, more than 46 000 auctions have taken place via the online auction service. In 66% of cases, multiple auctions were required to settle a single debt, and in half of the cases, debt repayment was not possible, based on the starting price. Banks, against individuals, are involved in speeding up the vast majority of foreclosures. Auctions soared in between 2020 lockdowns.

**A**s of 12 June 2021, a total of 91 744 auctions had been posted on the online auction platform [eauction.gr](#), following the country's launch of the service in September 2017. The iMEDD Lab has looked into almost all auctions posted on the platform up to 31 May 2021. Of these, 51.2% have been completed, 41.8% are suspended and 1.2% have been canceled, while 5205 auctions (5.8%) have already been posted and are scheduled to take place by February 2022. For reference, according to publicly available data up to May 2021, more than 2100 posted auctions are scheduled for June 2021 and more than 1300 posts concern auctions scheduled for July 2021.

**As of May 31, 2021, 46 198 online auctions had been completed, representing a total**

The screenshot shows a web browser window for the iMEDD:content website. The header includes the iMEDD logo and language options (EA, EN). Below the header is a navigation bar with links for STORIES, OPEN DATA, APPLICATIONS, and TOOLS & PRACTICES. The main content area starts with a quote: "Special purpose vehicles and debtors' debt recovery up to now, as well as an estimate on the percentage of debts that may have been repaid – considering on one hand, the total claims of the hasteners and on the other hand the starting prices in a sample of completed auctions." This is followed by a section titled "The data" which explains the data source ([eauction.gr](#)), processing tools (Python, Apache TIKA, OWASP), and anonymization methods (hashing, salting, bcrypt). A "Embed" button is visible at the bottom right of the content area.

<https://bit.ly/birn-2025-data>

## Extracting Bid Data from PDFs

Working with data in the real world is an awful, awful experience. Let's work on some spreadsheets about Kosovo's privatisation efforts.

click this  
→



Live coding worksheet



Completed version



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



Data: [02-bids-data.zip](#)

### Links:

- [Bid Reports page](#)
- [Natural PDF](#) A Python tool for analyzing PDFs

Ajencia Kosovare e Privatizit 

[pak-ks.org...](http://pak-ks.org/)

WEBMAIL | SHQIP SRPSKI ENGLISH

AGJENCIJA KOSOVARE E PRIVATIZIMIT  
KOSOVSKA AGENCJA ZA PRIVATIZACIJU  
PRIVATISATION AGENCY OF KOSOVO

Menya ▾

Home / Resultatet e Ofertimit

## Resultatet e Ofertimit

-  Shitia - 74
-  Shitia - 73
-  Shitia - 72
-  Shitia - 71
-  Shitia - 70
-  Shitia - 69
-  Shitia - 67
-  Shitia - 68

< 1 2 3 4 5 6 7 8 9 10 >

» SHITJA E ASETEVE NE LIKUIDIM

Shitjet Paraprake të Aseteve

Resultatet e Ofertimit

Vendimet e Shitjes

example-bid.pdf Page 1 of 3

REZULTATET E OFERTAVE - AKP SHITJA PERMES LIKUIDIMIT 33

Tabela

Data e Ofertimit: 28.06.2017

Nr.	Njësité	AKP ID	Çmimi më i Lartë	Tre Ofertuesit me çmim më të lartë					
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi
1	Njësia nr.01: Agrokultura Toka në Gjilan (Lot L)	GJI004	€15,127	L106	€15,127	0	€0	0	€0
2	Njësia nr.02: Agrokultura Toka në Gjilan (Lot M)	GJI004	€0	0	€0	0	€0	0	€0
3	Njësia nr.03: Pasuria Bujqësore Toka në Blbaj (Lot B)	GJI011	€69,611	L76	€69,611	L34	€58,000	L63	€12,222
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€111,000	L33	€111,000	L58	€66,666	L55	€13,501
5	Njësia nr.05: Pasuria Bujqësore Toka në Pojadë (Lot B)	GJI011	€3,892	L57	€3,892	0	€0	0	€0
6	Njësia nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)	GJI011	€16,666	L60	€16,666	L28	€16,550	0	€0
7	Njësia nr.07: Qëndresa Lokali 1 Kamenicë	GJI101	€169,000	L21	€169,000	L15	€125,559	L72	€91,153
8	Njësia nr.08: Qëndresa Lokali 2 Kamenicë	GJI101	€51,499	L16	€51,499	L19	€51,000	0	€0
9	Njësia nr.09: Qëndresa Lokali 3 Kamenicë	GJI101	€28,153	L71	€28,153	L47	€22,200	L17	€21,599
10	Njësia nr.10: KB Gjilobocica Toka dhe Objekti	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	Njësia nr.11: KB Novobërdë Toka dhe Objekte	GJI142	€5,730	L29	€5,730	0	€0	0	€0
12	Njësia nr.12: KB Bashkimi Toka dhe Objekti në Mirash	GJI008	€10,213	L108	€10,213	0	€0	0	€0
13	Njësia nr.13: Agrokultura Toka në Gjilan (Lot E)	GJI004	€21,000	L104	€21,000	L99	€20,250	0	€0
14	Njësia nr.14: Agrokultura Toka në Shillovë (Lot B)	GJI004	€15,556	L25	€15,556	L103	€15,100	0	€0
15	Njësia nr.15: Pasuria Bujqësore Toka në Sojevë	GJI011	€0	0	€0	0	€0	0	€0
16	LARGUAR NGA SHITJA		€0	0	€0	0	€0	0	€0
17	Njësia nr.17: Pasuria Bujqësore Toka në Gërlincë (Lot A)	GJI011	€33,333	L96	€33,333	L82	€31,500	L36	€24,200
18	Njësia nr.18: Agromarava Toka në Sllatinë e Poshtme (Lot B)	GJI035	€0	0	€0	0	€0	0	€0
19	Njësia nr.19: Produkti Tokë në Baks	MIT008	€0	0	€0	0	€0	0	€0
20	Njësia nr.20: Produkti Tokë në Brojë 1	MIT008	€0	0	€0	0	€0	0	€0
21	Njësia nr.21: Produkti Tokë në Brojë 2	MIT008	€0	0	€0	0	€0	0	€0
22	Njësia nr.22: Produkti Tokë në Brojë 3	MIT008	€1,888	L50	€1,888	L51	€1,491	0	€0
23	Njësia nr.23: Produkti Tokë në Brojë 4	MIT008	€200	L09	€200	0	€0	0	€0
24	Njësia nr.24: Elan Parcels 01001-0 Muhxazeri i Epërm/Studime e Epërme	MIT012	€0	0	€0	0	€0	0	€0
25	Njësia nr.25: Elan Parcels në Vërnicë	MIT012	€0	0	€0	0	€0	0	€0
26	Njësia nr.26: Preluzha Parcels 00280-0	MIT092	€5,000	L83	€5,000	0	€0	0	€0
27	Njësia nr.27: Preluzha Parcels 00290-0	MIT092	€2,000	L84	€2,000	0	€0	0	€0
28	Njësia nr.28: Lokali Afarisë nr.2	PRN011	€599,999	L98	€599,999	L105	€115,200	L110	€100,000
29	Njësia nr.29: Tokë në Vojnik	MIT008	€0	0	€0	0	€0	0	€0
30	Njësia nr.30: Tokë në Klinë e Epërmë 1	MIT008	€7,120	L13	€7,120	L11	€6,200	0	€0
31	Njësia nr.31: Elan Tokë në Dubovc	MIT012	€0	0	€0	0	€0	0	€0
32	Njësia nr.32: Ngastrat 0150-0, 0151-0 dhe 0159-0	MIT092	€5,000	L85	€5,000	0	€0	0	€0
33	Njësia nr.33: Ngastrat 0118-0	MIT092	€2,000	L86	€2,000	0	€0	0	€0
34	Njësia nr.34: Ngastrat 0007-0 dhe 0014-0	MIT092	€22,100	L121	€22,100	L87	€12,000	0	€0
35	Njësia nr.35: Stacioni i Veterinës në Soçanicë	MIT095	€6,700	L64	€6,700	0	€0	0	€0
36	Njësia nr.36: Zyra e Jugobankës në Mitrovicë	MIT301	€51,111	L128	€51,111	L130	€49,900	L129	€44,444
37	Njësia nr.37: KB Bec Prona në Lipovec	PEJ003	€15,100	L07	€15,100	0	€0	0	€0
38	Njësia nr.38: KB Bec Toka Bujqësore në Shishmon të Bokës	PEJ003	€12,000	L61	€12,000	0	€0	0	€0
39	Njësia nr.39: KB Ponoshec Toka Bujqësore në Smolicë	PEJ026	€3,190	L22	€3,190	0	€0	0	€0
40	Njësia nr.40: NSH Produkt Dugani në Asilan Çeshme	PEJ014	€0	0	€0	0	€0	0	€0

# DON'T FORGET

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

The image shows a dual-screen setup. The left screen displays the official Tabula website ([tabula.technology](https://tabula.technology)). The right screen shows the Tabula application window, which is a graphical interface for extracting tables from PDF files.

**Tabula Website (Left Screen):**

- Header:** Tabula: Extract Tables from PDF
- Section: Latest Version: Tabula 1.2.1 (June 4, 2018)**
  - Text: Tabula 1.2.1 fixes several bugs in the user interface. Read about all the changes in the release notes.
  - Text: Download Tabula below, or [on the release notes page](#).
  - Text: Special thanks to our OpenCollective backers.
  - Text: If you find Tabula useful in your work, please consider supporting us on OpenCollective!
- Section: How Can Tabula Help Me?**
  - Text: If you've ever tried to do anything with data provided in a PDF, you know how painful it is — there's no easy way to copy-and-paste that data into a spreadsheet.
  - Text: Tabula allows you to extract that data into a CSV or JSON file, directly from the PDF, using a simple, easy-to-use interface. Tabula works with almost any PDF.
- Section: Who Uses Tabula?**
  - Text: Tabula is used to power investigative reporting at news organizations like [ProPublica](#), [The Times of London](#), [Foreign Policy](#), [The New York Times](#) and the [St. Paul \(MN\) Pioneer Press](#).
  - Text: Grassroots organizations like [SchoolCuts.org](#) rely on Tabula to turn PDF reports into human-friendly public resources.
  - Text: And researchers of all kinds use Tabula to turn PDF reports into CSVs, and JSON files for use in analysis and database applications.
- Call to Action:** View the Project on GitHub

**Tabula Application Window (Right Screen):**

- Header:** Select Tables | Tabula
- Toolbar:** My Files, My Templates, About, Help, Source Code, Templates, Clear All Selections, Autodetect Tables, Preview & Export Extracted Data.
- Content Area:** A large table titled "Radford University - Academic Integrity Reports Fall 2021 - Spring 2025". The table has columns for "Date/Time of Incident", "Conduct Charges", and "Course". Numerous rows of data are listed, such as "9/1/2021 Cheating, Plagiarism English 111: Shakespeare Survey", "9/30/2021 Cheating, Plagiarism English 111: Principles of College Composition", etc. The entire table is highlighted with a red dashed selection box.
- Left Panel:** Four smaller preview windows labeled 1, 2, 3, and 4, each showing a different section of the PDF document with a red dashed selection box.
- Bottom Right:** A button labeled "Repeat this Selection".

We could use Tabula.

But we want to scale.

```
from natural_pdf import PDF
```

```
pdf = PDF("example-bid.pdf")
page = pdf.pages[0]
page.show()
```



REZULTATET E OFERTAVE - AKP SHITJA PERMES LIKUIDIMIT 33										Data e Ofertimit: 28.06.2017	
Nr.	Njësitë	AKP ID	Çmimi më i Lartë	Tre Ofertuesit me çmim më të lartë							
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi		
1	Njësie nr.01: Agrokultura Toka në Gilian (Lot L)	GJ004	€15,127	L106	€15,127	0	€0	0	€0		
2	Njësie nr.02: Agrokultura Toka në Gilian (Lot M)	GJ004	€0	0	€0	0	€0	0	€0		
3	Njësie nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)	GJ011	€69,611	L76	€69,611	L34	€58,000	L63	€12,222		
4	Njësie nr.04: Pasuria Bujqësore Toka në Ferizaj	GJ011	€1,111,000	L133	€1,111,000	L58	€666,666	L55	€513,501		
5	Njësie nr.05: Pasuria Bujqësore Toka në Pojadë (Lot B)	GJ011	€3,892	L57	€3,892	0	€0	0	€0		
6	Njësie nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)	GJ011	€16,666	L60	€16,666	L28	€16,550	0	€0		
7	Njësie nr.07: Qëndresa Lokali 1 Kamenicë	GJ101	€169,000	L21	€169,000	L15	€125,559	L72	€91,153		
8	Njësie nr.08: Qëndresa Lokali 2 Kamenicë	GJ101	€51,499	L16	€51,499	L19	€51,000	0	€0		
9	Njësie nr. 09: Qëndresa Lokali 3 Kamenicë	GJ101	€28,153	L71	€28,153	L47	€22,200	L17	€21,599		
10	Njësie nr.10: KB Gillobocica Toka dhe Objekti	GJ141	€28,700	L38	€28,700	0	€0	0	€0		
11	Njësie nr.11: KB Novobërdë Toka dhe Objekte	GJ142	€5,730	L29	€5,730	0	€0	0	€0		
12	Njësie nr.12: KB Bashkimi Toka dhe Objekti në Mirash	GJ008	€10,213	L108	€10,213	0	€0	0	€0		
13	Njësie nr.13: Agrokultura Toka në Gilian (Lot E)	GJ004	€21,000	L104	€21,000	L99	€20,250	0	€0		
14	Njësie nr.14: Agrokultura Toka në Shillovë (Lot B)	GJ004	€15,556	L25	€15,556	L103	€15,100	0	€0		
15	Njësie nr.15: Pasuria Bujqësore Toka në Sojevë	GJ011	€0	0	€0	0	€0	0	€0		
16	LARGUAR NGA SHITJA		€0	0	€0	0	€0	0	€0		
17	Njësie nr.17: Pasuria Bujqësore Toka në Gërlincë (Lot A)	GJ011	€33,333	L96	€33,333	L82	€ 31,500	L36	€ 24,200		
18	Njësie nr.18: Agromorava Toka në Sllatina e Poshtme (Lot B)	GJ035	€0	0	€0	0	€0	0	€0		
19	Njësie nr.19: Produkti Tokë në Baks	MIT008	€0	0	€0	0	€0	0	€0		
20	Njësie nr.20: Produkti Tokë në Brojë 1	MIT008	€0	0	€0	0	€0	0	€0		
21	Njësie nr.21: Produkti Tokë në Brojë 2	MIT008	€0	0	€0	0	€0	0	€0		
22	Njësie nr.22: Produkti Tokë në Brojë 3	MIT008	€1,888	L50	€1,888	L51	€ 1,491	0	€0		
23	Njësie nr.23: Produkti Tokë në Brojë 4	MIT008	€200	L09	€200	0	€0	0	€0		
24	Njësie nr.24: Elan Parcela 01001-0 Muhaxheri i Epërm/Studime e Epërme	MIT012	€0	0	€0	0	€0	0	€0		
25	Njësie nr.25: Elan Parcela në Vërnicië	MIT012	€0	0	€0	0	€0	0	€0		
26	Njësie nr.26: Prelluzha Parcels 00280-0	MIT092	€5,000	L83	€5,000	0	€0	0	€0		
27	Njësie nr.27: Prelluzha Parcels 00290-0	MIT092	€2,000	L84	€2,000	0	€0	0	€0		
28	Njësie nr.28: Lokali Afarist nr.2	PRN011	€599,999	L98	€599,999	L105	€ 115,200	L110	€ 100,000		
29	Njësie nr.29: Tokë në Vojnik	MIT008	€0	0	€0	0	€0	0	€0		
30	Njësie nr.30: Tokë në Klinë e Epërme 1	MIT008	€7,120	L13	€7,120	L11	€ 6,200	0	€0		
31	Njësie nr.31: Elan Tokë në Dubovc	MIT012	€0	0	€0	0	€0	0	€0		
32	Njësie nr.32: Ngastrat 0150-0, 0151-0 dhe 0159-0	MIT092	€5,000	L85	€5,000	0	€0	0	€0		
33	Njësie nr.33: Ngastra 0118-0	MIT092	€2,000	L86	€2,000	0	€0	0	€0		
34	Njësie nr.34: Ngastrat 0007-0 dhe 0014-0	MIT092	€22,100	L121	€22,100	L87	€ 12,000	0	€0		
35	Njësie nr.35: Stacioni i Veterinës në Soçanicë	MIT095	€6,700	L64	€6,700	0	€0	0	€0		
36	Njësie nr.36: Zyra e Jugobankës në Mitrovicë	MIT101	€51,111	L128	€51,111	L130	€ 49,900	L129	€ 44,444		
37	Njësie nr.37: KB Bec Prona në Lipovec	PEJ003	€15,100	L07	€15,100	0	€0	0	€0		
38	Njësie nr.38: KB Bec Toka Bujqësore në Shishmon të Bokës	PEJ003	€12,000	L61	€12,000	0	€0	0	€0		
39	Njësie nr. 39: KB Ponoshec Toka Bujqësore në Smolnicë	PEJ026	€3,190	L22	€3,190	0	€0	0	€0		
40	Njësie nr.40: NSH Produkt Dygani në Asllan Çeshme	PEJ014	€0	0	€0	0	€0	0	€0		

page.find('rect[fill~="yellow"]').below().show()

### REZULTATET E OFERTAVE - AKP SHITJA PERMES LIKUIDIMIT 33

**Tabela**

Data e Ofertimit: 28.06.2017

Nr.	Njësitë	AKP ID	Çmimi më i Lartë	Tre Ofertuesit me çmim më të lartë					
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi
1	<a href="#">Njësia nr.01: Agrokultura Toka në Gilian (Lot L)</a>	GJI004	€15,127	L106	€15,127	0	€0	0	€0
2	<a href="#">Njësia nr.02: Agrokultura Toka në Gilian (Lot M)</a>	GJI004	€0	0	€0	0	€0	0	€0
3	<a href="#">Njësia nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)</a>	GJI011	€69,611	L76	€69,611	L34	€58,000	L63	€12,222
4	<a href="#">Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj</a>	GJI011	€1,111,000	L133	€1,111,000	L58	€666,666	L55	€513,501
5	<a href="#">Njësia nr.05: Pasuria Bujqësore Toka në Pojatë (Lot B)</a>	GJI011	€3,892	L57	€3,892	0	€0	0	€0
6	<a href="#">Njësia nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)</a>	GJI011	€16,666	L60	€16,666	L28	€16,550	0	€0
7	<a href="#">Njësia nr.07: Qëndresa Lokali 1 Kamenicë</a>	GJI101	€169,000	L21	€169,000	L15	€125,559	L72	€91,153
8	<a href="#">Njësia nr.08: Qëndresa Lokali 2 Kamenicë</a>	GJI101	€51,499	L16	€51,499	L19	€51,000	0	€0
9	<a href="#">Njësia nr. 09: Qëndresa Lokali 3 Kamenicë</a>	GJI101	€28,153	L71	€28,153	L47	€22,200	L17	€21,599
10	<a href="#">Njësia nr.10: KB Glloboçica Toka dhe Objekti</a>	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	<a href="#">Njësia nr.11: KB Novobërda Toka dhe Objekte</a>	GJI142	€5,730	L29	€5,730	0	€0	0	€0
12	<a href="#">Njësia nr.12: KB Bashkimi Toka dhe Objekti në Mirash</a>	GJI008	€10,213	L108	€10,213	0	€0	0	€0
13	<a href="#">Njësia nr.13: Agrokultura Toka në Gilian (Lot E)</a>	GJI004	€21,000	L104	€21,000	L99	€20,250	0	€0
14	<a href="#">Njësia nr.14: Agrokultura Toka në Shillovë (Lot B)</a>	GJI004	€15,556	L25	€15,556	L103	€15,100	0	€0
15	<a href="#">Njësia nr.15: Pasuria Bujqësore Toka në Sojevë</a>	GJI011	€0	0	€0	0	€0	0	€0
16	<b>LARGUAR NGA SHITJA</b>		€0	0	€0	0	€0	0	€0
17	<a href="#">Njësia nr.17: Pasuria Bujqësore Toka në Gërlincë (Lot A)</a>	GJI011	€33,333	L96	€33,333	L82	€ 31,500	L36	€ 24,200
18	<a href="#">Njësia nr.18: Agromorava Toka në Sillatina e Poshtme (Lot B)</a>	GJI035	€0	0	€0	0	€0	0	€0
19	<a href="#">Njësia nr.19: Produkti Tokë në Baks</a>	MIT008	€0	0	€0	0	€0	0	€0
20	<a href="#">Njësia nr.20: Produkti Tokë në Brojë 1</a>	MIT008	€0	0	€0	0	€0	0	€0
21	<a href="#">Njësia nr.21: Produkti Tokë në Brojë 2</a>	MIT008	€0	0	€0	0	€0	0	€0
22	<a href="#">Njësia nr.22: Produkti Tokë në Brojë 3</a>	MIT008	€1,888	L50	€1,888	L51	€ 1,491	0	€0
23	<a href="#">Njësia nr.23: Produkti Tokë në Brojë 4</a>	MIT008	€200	L09	€200	0	€0	0	€0
24	<a href="#">Njësia nr.24: Elan Parcels 01001-0 Muhaxheri i Epërm/Studime e Epërmë</a>	MIT012	€0	0	€0	0	€0	0	€0
25	<a href="#">Njësia nr.25: Elan Parcels në Vërnicë</a>	MIT012	€0	0	€0	0	€0	0	€0
26	<a href="#">Njësia nr.26: Prelluzha Parcels 00280-0</a>	MIT092	€5,000	L83	€5,000	0	€0	0	€0
27	<a href="#">Njësia nr.27: Prelluzha Parcels 00290-0</a>	MIT092	€2,000	L84	€2,000	0	€0	0	€0
28	<a href="#">Njësia nr.28: Lokali Afarist nr.2</a>	PRN011	€599,999	L98	€599,999	L105	€ 115,200	L110	€ 100,000
29	<a href="#">Njësia nr.29: Tokë në Vojnik</a>	MIT008	€0	0	€0	0	€0	0	€0
30	<a href="#">Njësia nr.30: Tokë në Klinë e Epërme 1</a>	MIT008	€7,120	L13	€7,120	L11	€ 6,200	0	€0
31	<a href="#">Njësia nr.31: Elan Tokë në Duhavë</a>	MIT012	€0	0	€0	0	€0	0	€0

```
df = page.find('rect[fill~=yellow]').below().extract_table().to_df(header=False)
df.head()
```

0	1	2	3	4	5	6	7	8	9	10	11
0	1	Njësia nr.01: Agrokultura Toka në Gjilan (Lot L)	GJI004	€15,127	L106	€15,127	<NA>	0	€0	<NA>	0
1	2	Njësia nr.02: Agrokultura Toka në Gjilan (Lot M)	GJI004	€0	0	€0	<NA>	0	€0	<NA>	0
2	3	Njësia nr.03: Pasuria Bujqësore Toka në Bibaj ...	GJI011	€69,611	L76	€69,611	<NA>	L34	€58,000	<NA>	L63
3	4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€1,111,000	L133	€1,111,000	<NA>	L58	€666,666	<NA>	L55
4	5	Njësia nr.05: Pasuria Bujqësore Toka në Pojatë...	GJI011	€3,892	L57	€3,892	<NA>	0	€0	<NA>	0

0	1	2	3	4	5	6	7	8	9	10	11	
0	1	Njësia nr.01: Agrokultura Toka në Gjilan (Lot L)	GJI004	€15,127	L106	€15,127	<NA>	0	€0	<NA>	0	€0
1	2	Njësia nr.02: Agrokultura Toka në Gjilan (Lot M)	GJI004	€0	0	€0	<NA>	0	€0	<NA>	0	€0
2	3	Njësia nr.03: Pasuria Bujqësore Toka në Bibaj ...	GJI011	€69,611	L76	€69,611	<NA>	L34	€58,000	<NA>	L63	€12,222
3	4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€1,111,000	L133	€1,111,000	<NA>	L58	€666,666	<NA>	L55	€513,501
4	5	Njësia nr.05: Pasuria Bujqësore Toka në Pojatë...	GJI011	€3,892	L57	€3,892	<NA>	0	€0	<NA>	0	€0

€15,127

Nr.	Njësitetë	AKP ID	Çmimi më i Lartë	Tre Ofertuesit me çmim më të lartë					
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi
1	Njësia nr.01: Agrokultura Toka në Gjilan (Lot L)								
2	Njësia nr.02: Agrokultura Toka në Gjilan (Lot M)								
3	Njësia nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)								
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj								
5	Njësia nr.05: Pasuria Bujqësore Toka në Pojatë (Lot B)								
6	Njësia nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)								

Nr.	Njësitetë	AKP ID	Çmimi më i Lartë	Tre Ofertuesit me çmim më të lartë					
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi
41	Njësia nr.01: NSH Deva ngrastja nr.705 ZX Deva	PEN092	€0	€0	€0	€0	€0	€0	€0
42	Njësia nr.45: KB Malishevë Ngrastja nr. 24/1-në Dranëshiq	PEN172	€0	€0	€0	€0	€0	€0	€0
43	Njësia nr.45: Dritë e KLL-në Sheshin e Republikës	PEN173	€24,000	€24,000	€24,000	€24,000	€24,000	€24,000	€24,000
44	Njësia nr.44: FAP Fëmije Sadica Fabrika e Jelqo	PEN174	€0	€0	€0	€0	€0	€0	€0
45	Njësia nr.45: KB Shëpjet - Tokë Bujqësore në Shkopet 1	PEN175	€0	€0	€0	€0	€0	€0	€0
46	Njësia nr.47: NSH Emri Durci - Dritë nr. 3 afér Gjimnazit, Gjakovë	PEN176	€0	€0	€0	€0	€0	€0	€0
47	Njësia nr.48: KB Bistrica Ngrastja nr. 59-1-në Dranc	PEN177	€0	€0	€0	€0	€0	€0	€0
48	Njësia nr.49: KB Bistrica Tokë Bujqësore në Kamenicë 2	PEN178	€0	€0	€0	€0	€0	€0	€0
49	Njësia nr.50: NT Korrikus Tokë Bujqësore në Berat e Peje	PEN179	€0	€0	€0	€0	€0	€0	€0
50	Njësia nr.51: Dritë e Këndi Lokal Afërsit Mitrovicë	PEN180	€0	€0	€0	€0	€0	€0	€0
51	Njësia nr.52: Llogara Shëpjetorja nr. IV	PEN181	€0	€0	€0	€0	€0	€0	€0
52	Njësia nr.53: EUP e Këndi Lokal nr. Gjilan	PEN182	€0	€0	€0	€0	€0	€0	€0
53	Njësia nr.54: Vitoq Comerci - Tokë Bujqësore në Zonë Industriale - LANGUARINA SHITJA	PEN183	€0	€0	€0	€0	€0	€0	€0
54	Njësia nr.55: Gazi Cevdetinës Tokë Bujqësore Aferatës së Dantani	PEN184	€0	€0	€0	€0	€0	€0	€0
55	Njësia nr.56: Tokë Bujqësore Mirzë e Edderme V	PEN185	€0	€0	€0	€0	€0	€0	€0
56	Njësia nr.57: Tokë Bujqësore Mirzë e Edderme VI	PEN186	€0	€0	€0	€0	€0	€0	€0
57	Njësia nr.58: Tokë Bujqësore Mirzë e Edderme VII	PEN187	€0	€0	€0	€0	€0	€0	€0
58	Njësia nr.59: Tokë Bujqësore Mirzë e Edderme VIII	PEN188	€0	€0	€0	€0	€0	€0	€0
59	Njësia nr.60: KB Durrësi - Tokë Bujqësore Mirzë e Edderme VIII	PEN189	€0	€0	€0	€0	€0	€0	€0
60	Njësia nr.61: KB Përmet - Tokë Bujqësore Mirzë I	PEN190	€0	€0	€0	€0	€0	€0	€0
61	Njësia nr.62: KB Përmet Tokë Bujqësore Lushnjë	PEN191	€0	€0	€0	€0	€0	€0	€0
62	Njësia nr.63: Këndja Libratore e Pashkës e Edderme I	PEN192	€0	€0	€0	€0	€0	€0	€0
63	Njësia nr.64: KB Besa Tokë Bujqësore Tiranës I	PEN193	€0	€0	€0	€0	€0	€0	€0
64	Njësia nr.65: KB Besa Tokë Bujqësore Tiranës II	PEN194	€0	€0	€0	€0	€0	€0	€0
65	Njësia nr.66: KB Durrësi - Tokë Bujqësore Tiranës II	PEN195	€0	€0	€0	€0	€0	€0	€0
66	Njësia nr.67: KB Durrësi - Tokë Bujqësore Domnicë e Pashkës I	PEN196	€0	€0	€0	€0	€0	€0	€0
67	Njësia nr.68: KB Durrësi - Tokë Bujqësore Domnicë e Pashkës II	PEN197	€0	€0	€0	€0	€0	€0	€0
68	Njësia nr.69: KB Durrësi - Tokë Bujqësore Domnicë e Pashkës III	PEN198	€0	€0	€0	€0	€0	€0	€0
69	Njësia nr.70: KB Durrësi - Tokë Bujqësore Domnicë e Pashkës IV	PEN199	€0	€0	€0	€0	€0	€0	€0

In a Jupyter notebook have the code below that extracts a table

```
from natural_pdf import PDF
```

```
pdf = PDF("example-bid.pdf")  
page = pdf.pages[0]  
table_header = page.find("rect[fill~='yellow']"  
table_header.below().extract_table().to_df(header=None)
```

I have this code that extracts the column headers

```
column_names = (
    table_header
    .find_all('text:not(:contains(Tre Ofer))', overlap='center')
    .dissolve(padding=5)
    .extract_each_text(newlines=False, order='ltr')
)
print("Headers are", column_names)
```

Headers are ['Nr.', 'Njësitë', 'AKP ID', 'Çmimi më i Lartë', 'Ofertuesi', 'Çmimi', 'Ofertuesi', 'Çmimi', 'Ofertuesi', 'Çmimi']

your problems

`pdf = PDF("example-bid.pdf")`

your problems

the problems are

- 1) Some of the columns are missing all data, we need to remove them before we assign column names
  - 2) The 'Ofertuesi', 'Çmimi', header names are repeated because it the top bidder, second bidder, third bidder.
  - 3) It's multiple pages of PDFs, not just one, and we need to combine them.
  - 4) The bids are in euros but we need them in 'real' currency to analyze them.

*your request*

can you give me the code to fix all of these problems?

0	1	2	3	4	5	6	7	8	9	10	11
0	1	Njësia nr.01: Agrokultura Toka në Gjilan (Lot L) GJI004 €15,127 <NA> 0 €0 <NA> 0 €0									
L106	€15,127	<NA>	0	€0	<NA>	0	€0				
1	2	Njësia nr.02: Agrokultura Toka në Gjilan (Lot M) GJI004 €0 <NA> 0 €0 <NA> 0 €0									
2	3	Njësia nr.03: Pasuria Bujqësore Toka në Bibaj ... GJI011 €69,611 L76 €69,611 <NA> L34 €58,000 <NA> L63 €12,221									
3	4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj GJI011 €1,111,000 L133 €1,111,000 <NA> L58 €666,666 <NA> L513,501									
4	5	Njësia nr.05: Pasuria Bujqësore Toka në Pojatë... GJI011 €3,892 <NA> 0 €0 <NA> 0 €0									
L57	€3,892	<NA>	0	€0	<NA>	0	€0				

In the code below I am scraping a table from the first page of a PDF in a Jupyter notebook. It works, but I need some improvements:

1. The bidding price columns are money with euro and commas symbols in them. Clean them up so I can analyze them.
2. Some of the columns don't have any data in them. Remove those columns (ONLY those NA ones - not zeroes, zeroes are ok)
3. I want the nice column names in the dataframe, but you need to remove the "bad" columns before you assign names
4. The bidder and the bid price column names have duplicate names for the first, second, and third place bidders. Add numbers after the column name to keep it organized
5. There are multiple pages or sections in the PDF. Create one dataframe. As you do this

The code should be "safe" with the data - if anything unexpected happens, provide a warning or an error to show up so we don't lose data. Here is my current code:

```
from natural_pdf import PDF

pdf = PDF("example-bid.pdf")
page = pdf.pages[0]

table_header = page.find('rect[fill~=yellow]')

bad_text = table_header.find('text:contains(Tre Ofert)')
if bad_text:
    bad_text.exclude()
    ↓
column_names = (
```

```
column_names = (
    table_header
    .find_all('text')
    .dissolve(vertical=True)
    .extract_each_text(newlines=False, order='ltr')
)
```

```
print("Columns are", column_names)
```

```
df = table_header.below().extract_table().to_df(header=None)
df.head()
```

# embrace confusion

```
Columns are ['Njësia', 'Njësia', 'Njësia', 'Çmimi më i Lartë', 'Ofertuesi',
'Çmimi', 'Ofertuesi', 'Çmimi', 'Ofertuesi', 'Çmimi']
```

```
0 1 2 3 4 5 6 7 8 9 10 11
```

```
0 1 Njësia nr.01: Agrokultura Toka në Gjilan (Lot L) GJI004 €15,127
L106 €15,127 <NA> 0 €0 <NA> 0 €0
```

```
1 2 Njësia nr.02: Agrokultura Toka në Gjilan (Lot M) GJI004 €0 0
€0 <NA> 0 €0 <NA> 0 €0
```

```
2 3 Njësia nr.03: Pasuria Bujqësore Toka në Bibaj ... GJI011
€69,611 L76 €69,611 <NA> L34 €58,000 <NA> L63 €12,222
```

```
3 4 Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj GJI011
€1,111,000 L133 ↓ 1,111,000 <NA> L58 €666,666 <NA> L55
€513,501
```

Name	Date Modified	Size	Kind
0C8EE446-2572-48...-DA2E41B49247.pdf	Nov 2, 2017 at 10:07 AM	502 KB	PDF Document
0D8822D5-8E33-4...-4DA76323384A.pdf	Jul 23, 2025 at 2:12 PM	309 KB	PDF Document
1E15DB7F-47E6-4A...-0B56BDC0308E.pdf	May 8, 2019 at 3:56 PM	599 KB	PDF Document
2B8B3DD2-7A05-4...-3ED255C7F788.pdf	May 28, 2025 at 1:50 PM	181 KB	PDF Document
2F551774-A6CF-4D...-8421BB412E8A.pdf	Nov 2, 2017 at 10:07 AM	86 KB	PDF Document
3E175E76-4F07-4C...-7-C310251C5239.pdf	Aug 21, 2019 at 3:34 PM	55 KB	PDF Document
4B177C46-2CAF-4F...-81ADA9F2EA3E.pdf	Jan 10, 2018 at 9:40 AM	603 KB	PDF Document
5AD5D3DE-5EA7-4...-4210CAA97590.pdf	Nov 2, 2017 at 10:07 AM	487 KB	PDF Document
5AE7104C-8724-49...-871BD6A11C7F.pdf	Apr 24, 2024 at 3:40 PM	37 KB	PDF Document
6B7EFF38-0A3E-43...-A817E26CE859.pdf	Nov 2, 2017 at 10:07 AM	51 KB	PDF Document
6CE0359E-6D13-49...-C296A79D48E5.pdf	Nov 8, 2018 at 3:48 PM	445 KB	PDF Document
6D945485-6F65-4...-4-91B2C70A64E7.pdf	Nov 2, 2017 at 10:07 AM	58 KB	PDF Document
7A770171-5E96-4F8...-A601DE0F40DD.pdf	Oct 18, 2023 at 4:46 PM	30 KB	PDF Document
9A02D67A-C4A9-4...-ABC0C8813DFF.pdf	Mar 15, 2023 at 3:11PM	38 KB	PDF Document
9CA182A2-C4B4-4...-2289CF8B31BB.pdf	Nov 2, 2017 at 10:07 AM	16 KB	PDF Document
9E70B6EF-4DB2-45...-5B38CA828E80.pdf	Nov 2, 2017 at 10:07 AM	73 KB	PDF Document
9F338E1D-B999-4E...-C934EFFCD345.pdf	Nov 2, 2017 at 10:07 AM	14 KB	PDF Document
15A8700C-E26C-42...-E03180FE0A5E.pdf	Nov 27, 2024 at 3:12 PM	237 KB	PDF Document
38C3475F-3D0E-4...-1B83EB53CBD2.pdf	Dec 7, 2022 at 3:41PM	417 KB	PDF Document
052C52EF-7825-4C...-27554DD1938C.pdf	Sep 5, 2018 at 12:52 PM	560 KB	PDF Document
62B523FB-5DF9-47...-6007F70A8476.pdf	Apr 18, 2025 at 9:24 AM	113 KB	PDF Document
64D0BF1C-B88C-41...-F5167AEE6BBE.pdf	Nov 2, 2017 at 10:07 AM	16 KB	PDF Document
73C02FA3-EAA1-44...-97166333FA6F.pdf	Nov 2, 2017 at 10:07 AM	17 KB	PDF Document
73CBC8C0-7FBF-4...-8-85F62140489E.pdf	Nov 3, 2022 at 8:30 AM	275 KB	PDF Document
074F95C2-19B5-45...-2600D905B547.pdf	Jun 16, 2022 at 1:54 PM	191 KB	PDF Document
77AF7BA1-A384-4D...-63B5A3230E78.pdf	Nov 2, 2017 at 10:07 AM	120 KB	PDF Document
86A703F9-6313-49...-7650CA369476.pdf	Jun 28, 2024 at 12:24 PM	437 KB	PDF Document

Dropbox > Soma > Curriculum > 2025-birn > structured-data > bid-pdfs

75 items, 520.28 GB available

I have a collection of pdfs in the "bid-pdfs" folder, I want to combine them all and save them as one file.

Some of them have an extra column, the last one is "number of bidders." We can save that and keep it empty for the ones that don't have it.

If a PDF page has some other number of columns just flag it and keep going.

Give me a progress bar while the computer is working through this.

When all you have is a  
hammer everything looks  
like a nail.

but there are many things that can be  
solved with a hammer.

