

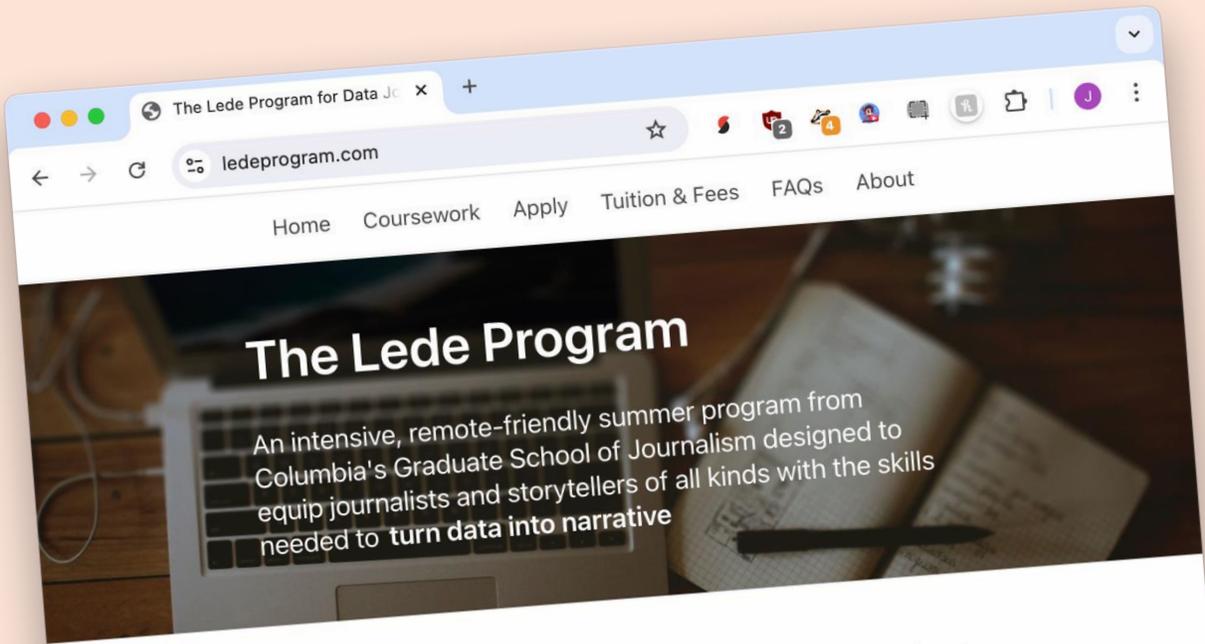
# Advanced Data Journalism

## Part I: Structured Data

Jonathan Soma

Columbia Journalism School

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



### 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!

### Not just for journalists

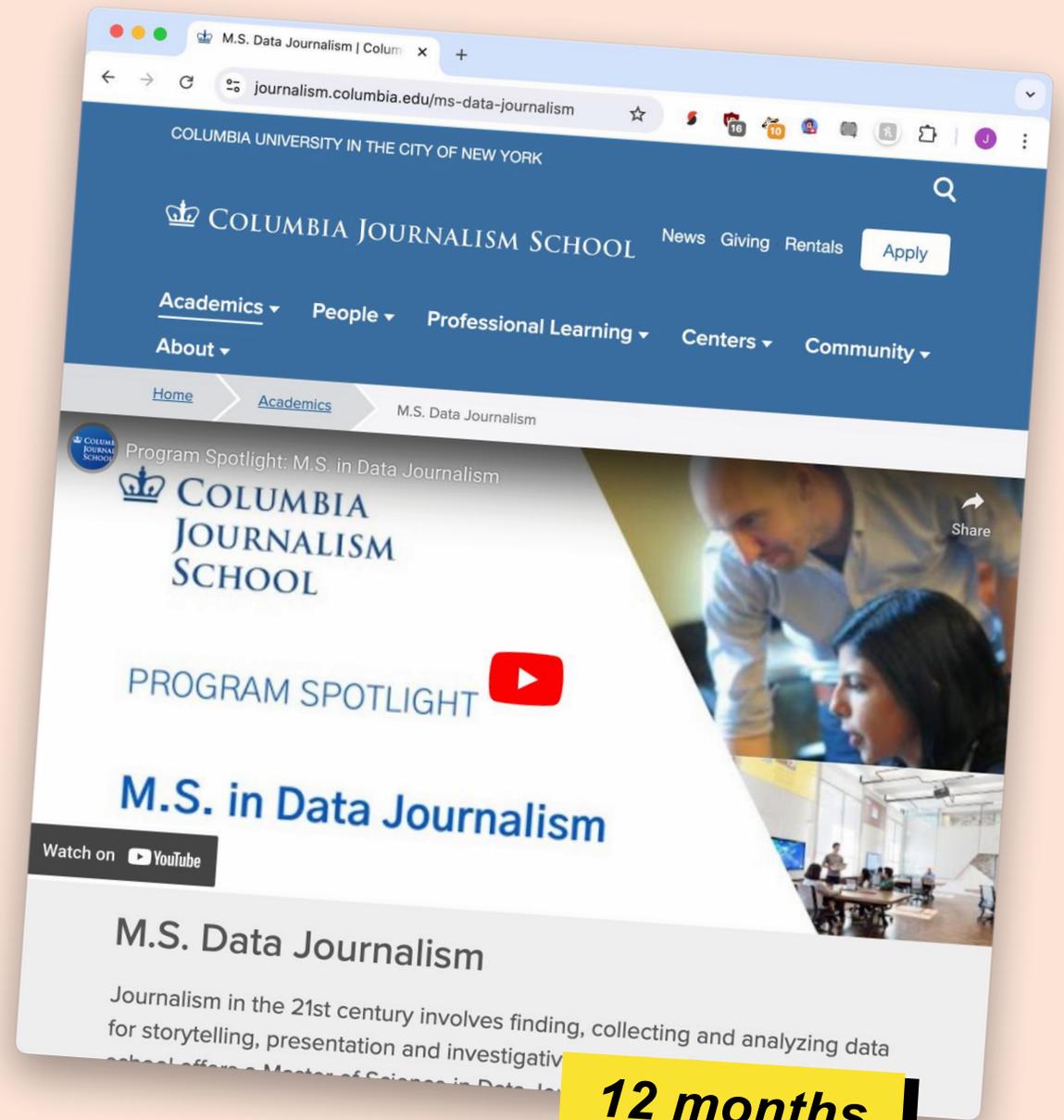
Pulling insights from data **isn't just for journalists** – our tools and skills are equally valuable to social scientists, policymakers, and academics of all stripes.

### 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 storytelling.

 [Have questions about Lede?](#)

**10 weeks**



### M.S. Data Journalism

Journalism in the 21st century involves finding, collecting and analyzing data for storytelling, presentation and investigative reporting.

**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.

# 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

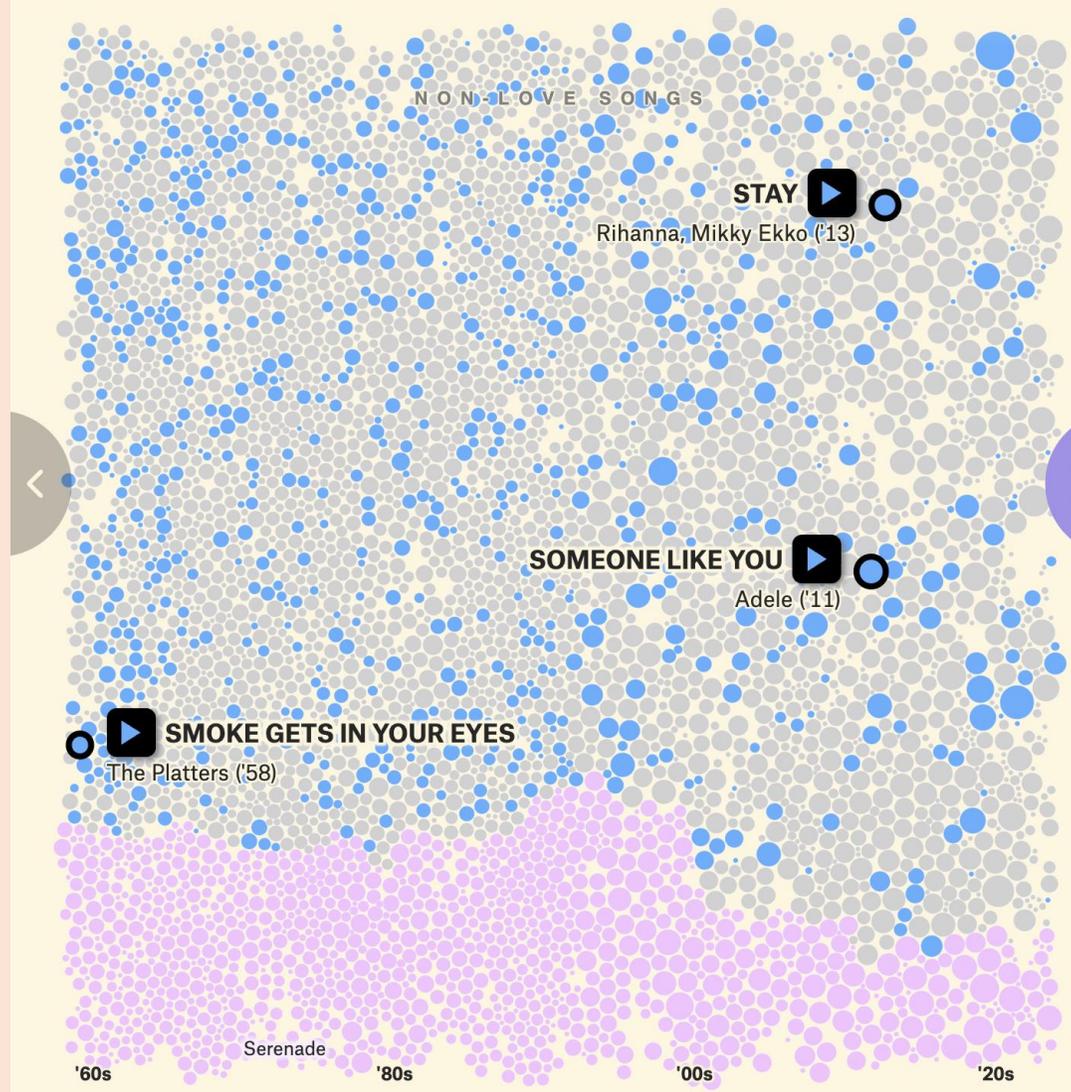
Foreigner National No data

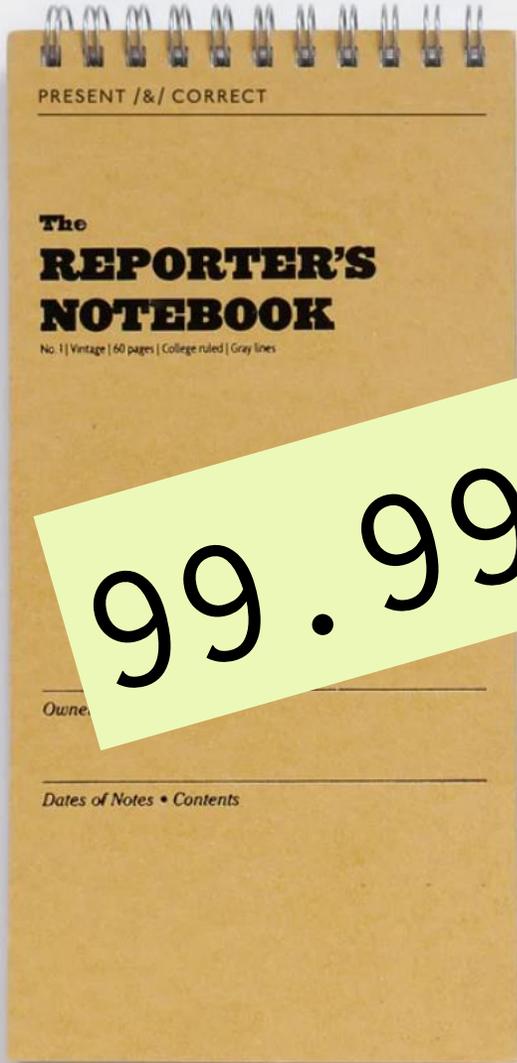


Note: Belgium, Luxembourg, Malta, Croatia, Poland, Lithuania, Bulgaria, Greece, and Cyprus failed to reply to

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**.

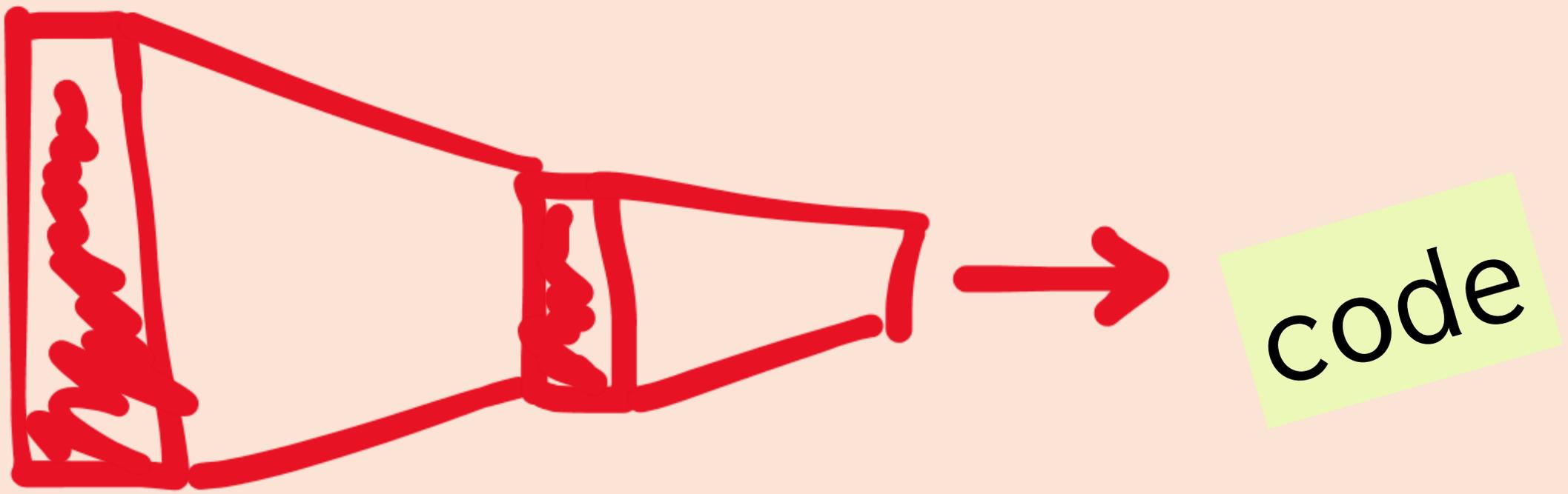




99.9999%

99.9999%

1	A	B	C	D	E	F	G
1	date	agency	recipient	value	savings	link	description
2	5/8/2025	Environmental P	North Carolina A&T State University				
3	5/8/2025	Environmental P	The University of North Carolina at Chapel Hill dba Office of Contracts and	1340704	216661	<a href="https://usaspending">https://usaspending</a>	RESEARCHERS WILL PERFORM STUDIES
4	5/8/2025	Environmental P	GEORGIA TECH RESEARCH CORP	1894823	78653	<a href="https://usaspending">https://usaspending</a>	RESEARCHERS WILL USE LONGITUDINAL
5	5/8/2025	Environmental P	The University of North Carolina at Chapel Hill dba Office of Contracts and	1000000	476179	<a href="https://usaspending">https://usaspending</a>	THE PROJECT GOALS ARE TO 1) EVALUAT
6	5/8/2025	Environmental P	University of Wisconsin - Madison	799833	73805	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:
7	5/8/2025	Environmental P	MICHIGAN STATE UNIVERSITY	998049	567356	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE PURPOSE OF THIS PRO
8	5/8/2025	Environmental P	The Ohio State University	1600000	1561741	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE GOALS OF THIS PROE
9	5/8/2025	Environmental P	University of Wisconsin - Madison	999995	119525	<a href="https://usaspending">https://usaspending</a>	THE GOAL OF THE RESEARCH IS TO DECR
10	5/8/2025	Environmental P	The Little Village Environmental Justice Organization	1362435	842529	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE PURPOSE OF THE PRO.
11	5/7/2025	Environmental P	Water Research Foundation	196034	0	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE PURPOSE OF THIS AME
12	5/7/2025	Environmental P	Oregon State University	1497907	700941	<a href="https://usaspending">https://usaspending</a>	THE PURPOSE OF THE PROPOSED RESEAF
13	5/7/2025	Environmental P	TEXAS A&M ENGINEERING EXPERIMENT STATION	2350211	2064835	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE PURPOSE OF THE PROJ
14	5/7/2025	Environmental P	Kansas Departmetn of Health & Environment	1600000	1555041	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE GOALS OF THIS PROJEC
15	5/7/2025	Environmental P	Regents of the University of California, The	1000000	957738	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE AGREEMENT PROVIDES
16	5/7/2025	Environmental P	University of California, Davis	549940	134504	<a href="https://usaspending">https://usaspending</a>	THE PURPOSE OF PROPOSED RESEARCH IS
17	5/7/2025	Environmental P	University of California, Los Angeles	799660	237460	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:
18	5/7/2025	Environmental P	The Regents of the University of CA - Riverside	798825	341961	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:
19	5/7/2025	Environmental P	The Regents of the University of CA - Berkeley	576683	478108	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE GOAL OF THE PROPOSED
20	5/7/2025	Environmental P	The University of Texas at Dallas	1999998	946636	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THIS PROJECT WILL STRATEG
21	5/7/2025	Environmental P	University of Texas at Austin	24999	105	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE PURPOSE OF THIS PROE
22	5/7/2025	Environmental P	The Water Research Foundation	36322	22267	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:RECIPIENT PROPOSES TO EXA
23	5/7/2025	Environmental P	The Rand Corporation	3245999	1803049	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THIS PROJECT AIMS TO ENHAN
24	5/7/2025	Environmental P	Physicians, Scientists, and Engineers for Sustainable and Healthy Energy Inc	1349997	952166	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE GOAL OF THE RESEARCH
25	5/7/2025	Environmental P	Oregon State University	1350000	538622	<a href="https://usaspending">https://usaspending</a>	DESCRIPTION:THE GOAL OF THIS RESEARCH
26	5/6/2025	Environmental P	Vanderbilt University	1984400	99712	<a href="https://usaspending">https://usaspending</a>	THE PROJECT PROPOSES TO STUDY THE TO
27	5/6/2025	Environmental P	University of Florida	790352	0	<a href="https://usaspending">https://usaspending</a>	THIS PROJECT WILL FILL A FUNDAMENTAL GA
28	5/6/2025	Environmental P	Emory University	798193	0	<a href="https://usaspending">https://usaspending</a>	THIS PROJECT AIMS TO UNDERSTAND HOW F
29				1345022	654617	<a href="https://usaspending">https://usaspending</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

A	B	C	D	E
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	7865
5/8/2025	Environmental P	GEORGIA TECH RESEARCH CORP	1000000	476175
5/8/2025	Environmental P	The University of North Carolina at Chapel Hill dba Office of Contracts and	799833	7380
5/8/2025	Environmental P	University of Wisconsin - Madison	998049	56735
5/8/2025	Environmental P	MICHIGAN STATE UNIVERSITY	1600000	156174
5/8/2025	Environmental P		95	11952
5/8/2025	Environmental P		35	84252
5/8/2025	Environmental P		34	
5/7/2025	Environmental P			70094
5/7/2025	Environmental P			206483
5/7/2025	Environmental P	INEERING EXPERIMENT STATION	1600000	155504
5/7/2025	Environmental P	tn of Health & Environment	1000000	95773
5/7/2025	Environmental P	Regents of the University of California, The	549940	13450
5/7/2025	Environmental P	University of California, Davis	799660	23746
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	47810

# What is "structured data?"

Also nothing fancy: spreadsheets. Rows and columns.

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\_Violations\_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:

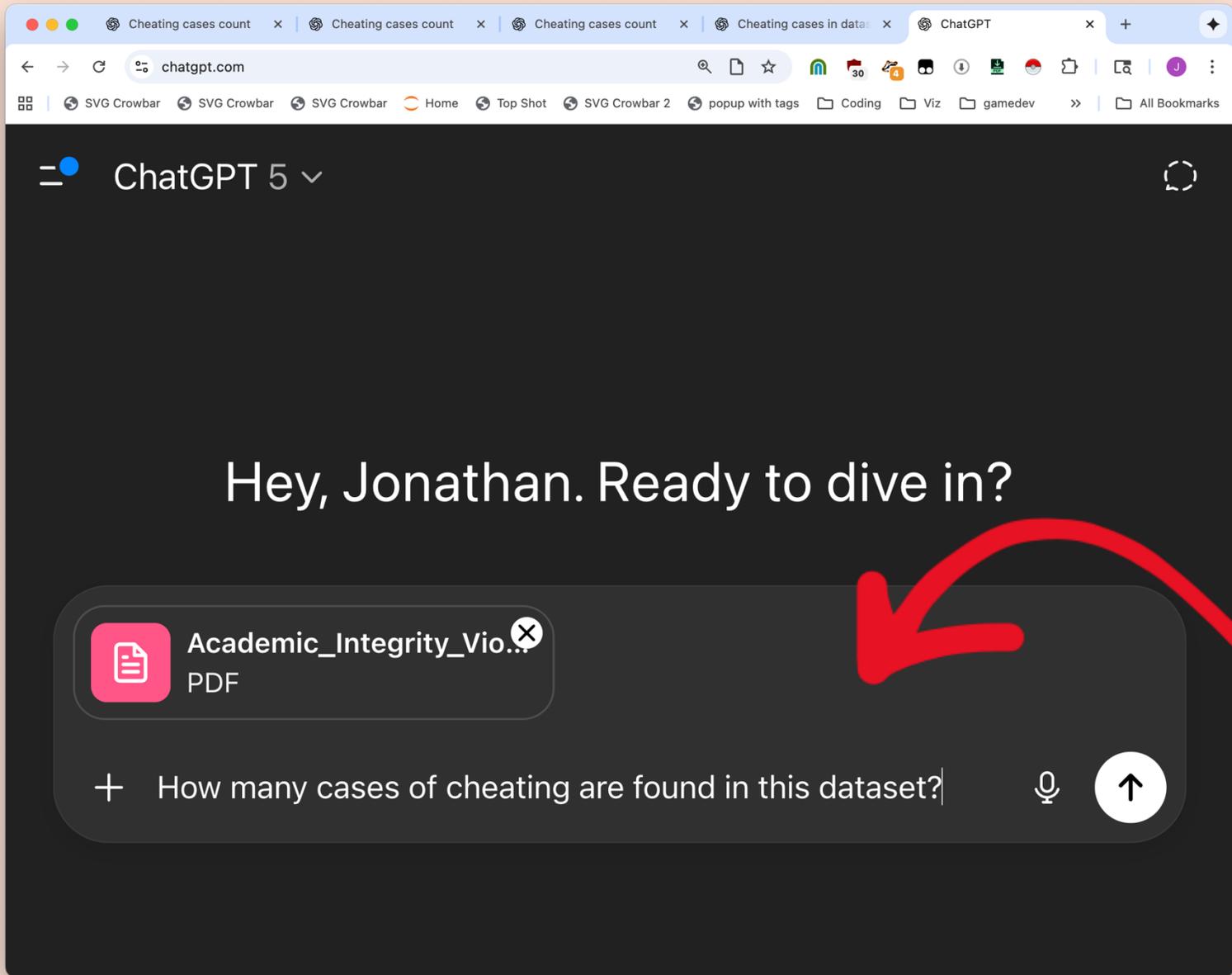
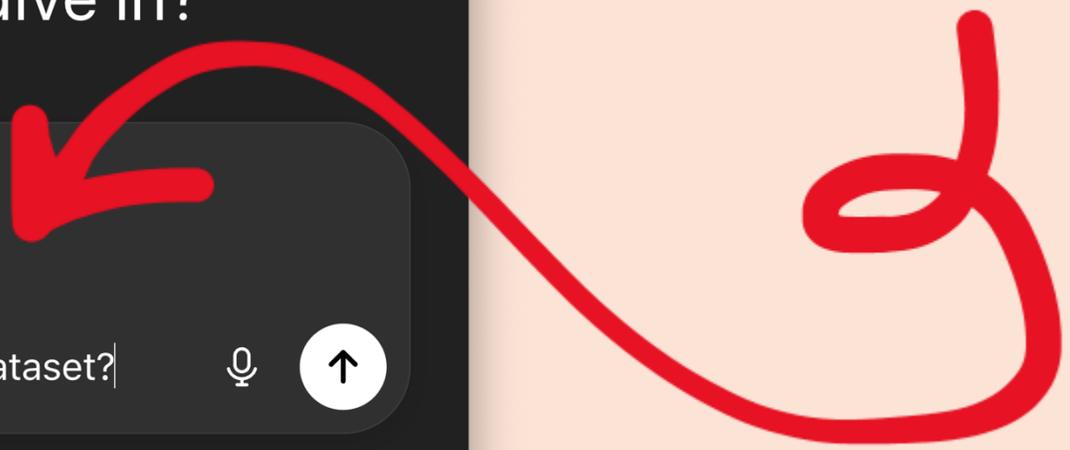
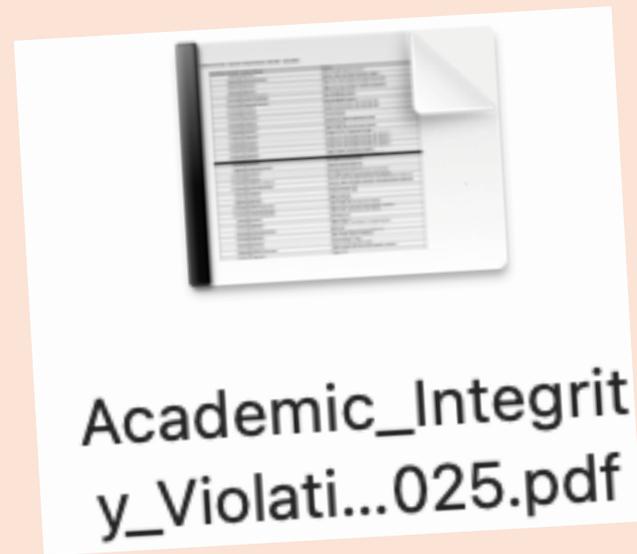
- [ChatGPT](#) Your AI best friend
- [MuckRock](#) A website for filing (and collecting) FOI responses

ChatGPT 5

Hey, Jonathan. Ready to dive in?

Academic\_Integrity\_Vio... PDF

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

A screenshot of the ChatGPT web interface. The browser tabs at the top show multiple instances of 'Cheating cases count' and one 'ChatGPT' tab. The address bar shows 'chatgpt.com'. The main chat area has a dark background. At the top left, it says 'ChatGPT 5'. Below that, a message from the AI says 'Hey, Jonathan. Ready to dive in?'. In the input area, a document icon is shown with the filename 'Academic\_Integrity\_Vio...' and 'PDF' below it. Below the document icon, the user has typed the question 'How many cases of cheating are found in this dataset?'. There are icons for adding more documents, voice input, and sending the message.

ChatGPT 5

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

✓ The dataset contains **100** cases of cheating in the dataset from Fall 2021 to Spring 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

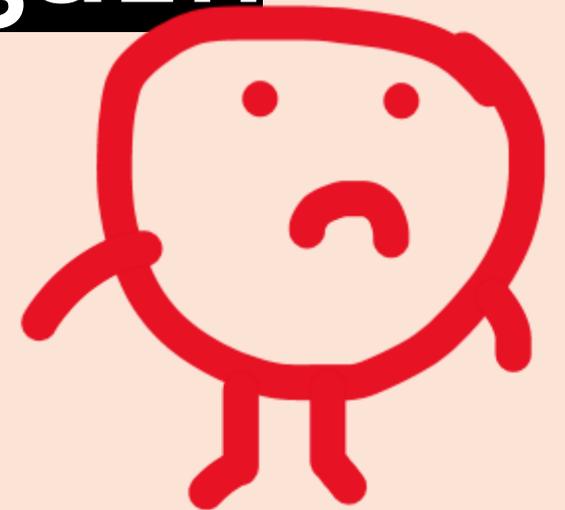
ChatGPT can make mistakes. Check important info.

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



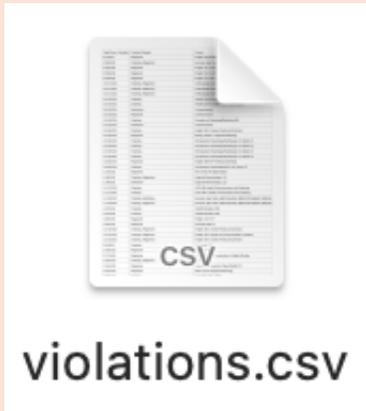
violations.csv

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

Paste Aptos Narrow (Bod... 12 General Conditional Formatting Format as Table Cell Styles

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						



ChatGPT

chatgpt.com

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

ChatGPT 5

What's on your mind today?

violations.csv  
Spreadsheet

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

↑

A thick, hand-drawn red arrow originates from the 'violations.csv' icon on the left and points towards the 'violations.csv' file card in the chat interface.

There are 141 cases of cheating

There are 141 cases of cheating recorded in this dataset

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.

KEPT OUT

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

by Aaron Glantz and Emmanuel Martinez February 15, 2018



**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

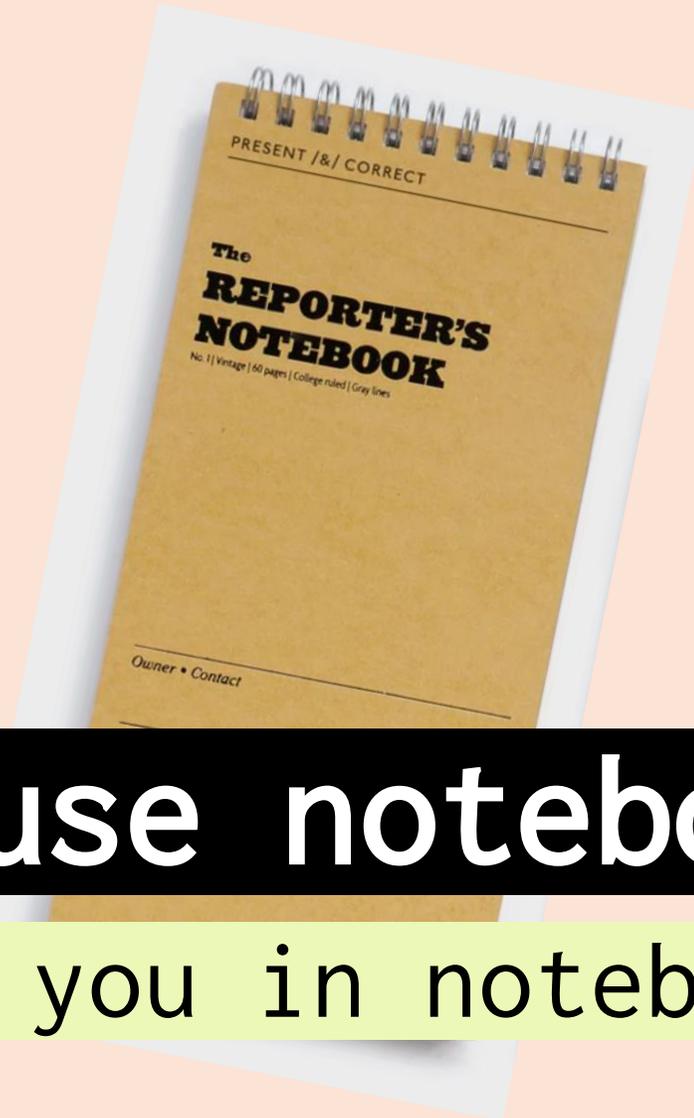
The HMDA data and reports are the most comprehensive publicly available information on mortgage market activity. The data and reports can be used along with the [Census](#) demographic information for data analysis purposes. Available below are the data and reports for HMDA data collected in or after 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	Static Datasets
<p><b><a href="#">Modified Loan/Application Register (LAR)</a></b></p> <p>The modified LAR provides loan-level data for an individual financial institution, as modified by the Bureau to protect applicant and borrower privacy.</p> <p>→ Update Frequency: Upon Institution resubmission</p>	<p><b><a href="#">Snapshot National Loan-Level Dataset</a></b></p> <p>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.</p> <p>→ Data Freeze Date: May 19, 2025</p> <p>→ Update Frequency: Does not update. Generated a few months after filing deadline</p>
<p><b><a href="#">Dynamic National Loan-Level Dataset</a></b></p>	



Data people use notebooks

(AI can write for you in notebooks too)

## 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',
```

# 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](#)

Date	Description	Re
2022-04-27	Data and analysis of state child abuse and neglect registries and appeals	
2022-04-25	Data and analysis of intermediate care facilities	
2021-09-17	Data and analysis re. US adult guardianship filing counts	
2021-05-26	Analysis of excess deaths caused by the February 2021 winter storm and power outages in Texas	
2020-11-11	Analysis of county-level COVID-19 deaths and presidential voter preference	
2020-10-28	Analysis of 2020's "Electoral College effect" by demographic	
2020-06-04	Analysis of "1033" program transfers since Ferguson	
2020-05-07	Analysis of ZIP code-level COVID-19 cases in five major cities	
2020-02-27	Analysis of Census tract-level gentrification in five major cities	
2019-11-11	Analysis of U.S. Census Survey of Income and Program Participation (SIPP), re. generational trends in support providers	
2019-10-31	Analysis for "Your Dumb Tweets Are Getting Flagged To People Trying To Stop School Shootings"	
2019-10-17	Analysis for "Donald Trump's Campaign Is Cashing In On Impeachment"	

```

2022-04-registries / notebooks / substantiations / CA_subs.ipynb
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]:
          race  asian  black  hispanic  other  pac  white  count
year
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("")

```

# "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



[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.

SCIENCE

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.



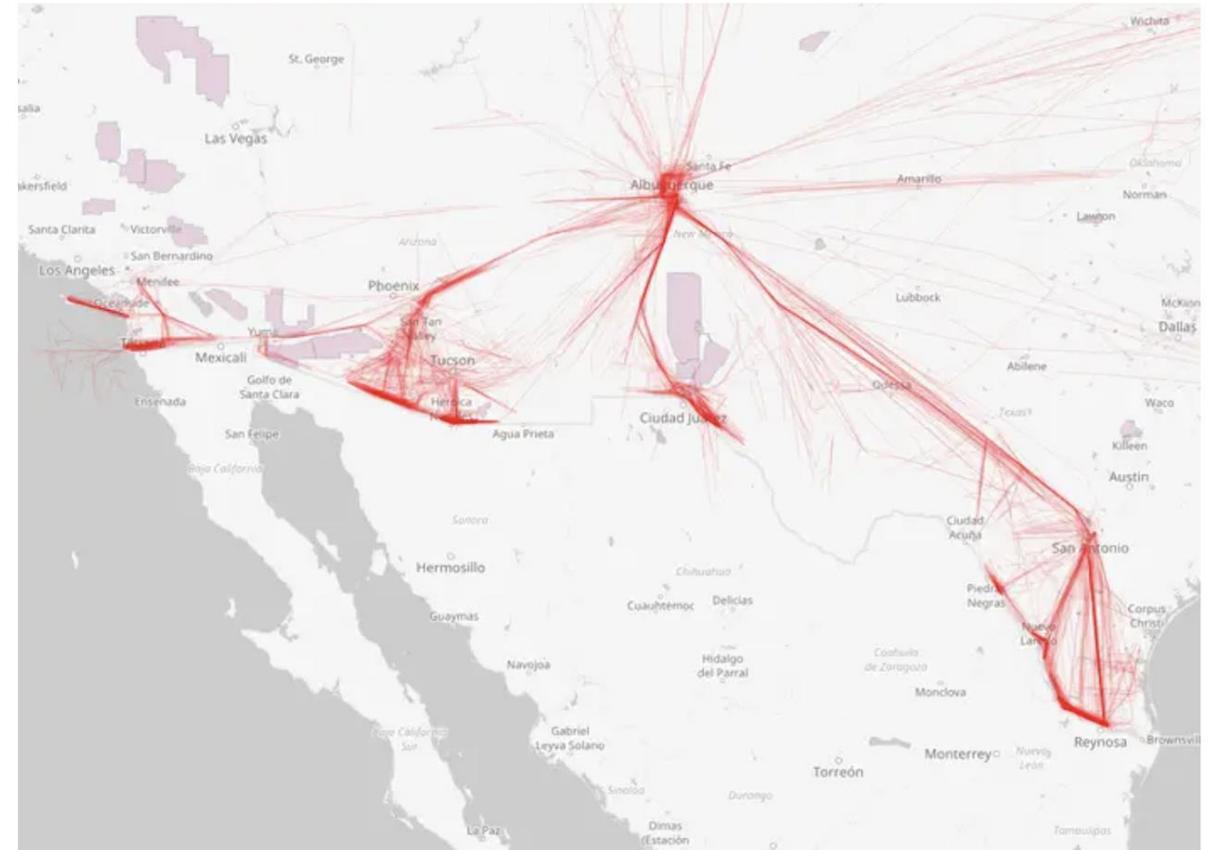
**Peter Aldhous**  
BuzzFeed News Reporter

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](http://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

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

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

Translate page with Google

Home / Resources

JOURNALIST RESOURCE | JANUARY 10, 2025

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

Author:



Lam Thuy Vo  
GRANTEE

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!



[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.

Residential property accounts

lab.imedd.org/en/pleistiriasmoi-katoikies-to-38-to...

EA EN An initiative of iMEDD

iMEDD: content STORIES OPEN DATA APPLICATIONS TOOLS & PRACTICES



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.

As of 12 June 2021, a total of 91 744 auctions had been posted on the online auction platform [eauction.gr](http://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**

Auctions in Greece

Building the application for

lab.imedd.org/e...

EA EN An initiative of iMEDD

iMEDD: content

Special purpose vehicles and companies set up mainly in speeding up auctions, 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.

### The data

All the analyses set out in the application are based on publicly available information displayed in the online auction platform [eauction.gr](http://eauction.gr). The iMEDD Lab has studied almost all auctions whose display in the platform dates from September 2017, while it continues to study new auctions made available daily.

These data are processed using [Python](#) programming language. [Apache TIKA](#) library is used for extracting available data from published .pdf files. At the final stage of data processing and before being stored in a database, data anonymization with the method of “hashing” in accordance with the recommendations of the Open Web Application Security Project ([OWASP](#)) on the [storage of codes or sensitive data](#) (see “Peppering”, “Salting” and [bcrypt](#)). The iMEDD Lab’s application does not share any kind of demographic or personal information about hasteners, debtors or other parties involved in the cases concerned. The only non-anonymized data shared by the application are the company names of banks, special purpose vehicles and credit servicing firms involved in speeding up auctions.

</> Embed

<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

Agjencia Kosovare e Privatizimit

pak-ks.org/...

AGJENCIA KOSOVARE E PRIVATIZIMIT  
KOSOVSKA AGENCIJA ZA PRIVATIZACIJU  
PRIVATISATION AGENCY OF KOSOVO

Menyja

Home / Rezultatet e Ofertimit

## Rezultatet e Ofertimit

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

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

» SHITJA E ASETEVE NE LIKUIDIM

Shitjet Paraprake të Aseteve

Rezultatet 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.	Njësitë	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 Gillan (Lot L)	GJI004	€15,127	L106	€15,127	0	€0	0	€0
2	Njësia nr.02: Agrokultura Toka në Gillan (Lot M)	GJI004	€0	0	€0	0	€0	0	€0
3	Njësia nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)	GJI011	€69,611	L76	€69,611	L34	€38,000	L63	€12,222
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€1,111,000	L133	€1,111,000	L58	€666,666	L55	€313,501
5	Njësia nr.05: Pasuria Bujqësore Toka në Pojatë (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 Gllboçica Toka dhe Objekti	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	Njësia nr.11: KB Novobërda Toka dhe Objekti	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 Gillan (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	<b>LARGUAR NGA SHITJA</b>		€0	0	€0	0	€0	0	€0
17	Njësia nr.17: Pasuria Bujqësore Toka në Gërlicë (Lot A)	GJI011	€33,333	L96	€33,333	L82	€31,500	L36	€24,200
18	Njësia nr.18: Agromorava Toka në Sllatina 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 Parcela 01001-0 Muhaxheri i Epërm/Studime e Epërme	MIT012	€0	0	€0	0	€0	0	€0
25	Njësia nr.25: Elan Parcela në Vërmicë	MIT012	€0	0	€0	0	€0	0	€0
26	Njësia nr.26: Prelluzha Parcela 00280-0	MIT092	€5,000	L83	€5,000	0	€0	0	€0
27	Njësia nr.27: Prelluzha Parcela 00290-0	MIT092	€2,000	L84	€2,000	0	€0	0	€0
28	Njësia nr.28: Lokali Afarist nr.2	PRN011	€399,999	L98	€399,999	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 Epërme 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: Ngastra 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	MIT101	€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 Ponohec Toka Bujqësore në Smolicë	PEJ026	€3,190	L22	€3,190	0	€0	0	€0
40	Njësia nr.40: NSH Produkt Dyqani në Asllan Ceshme	PEJ014	€0	0	€0	0	€0	0	€0

# DON'T FORGET

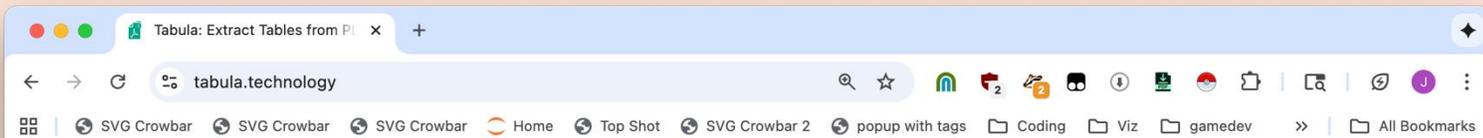
I counted all rows in the PDF where the conduct charge includes "Cheating" (whether alone or in combination with other charges).

 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



## Tabula



Tabula is a tool for liberating data tables locked inside PDF files.

[View the Project on GitHub](#)  
tabulapdf/tabula

Download for **Windows** | Download for **Mac** | [View source on GitHub](#)

Current Version: **1.2.1**

Other Versions: [pre-releases & archives](#)

**Need help?** Open an [issue on Github](#).

**Donate:** Help support this project by [backing us on OpenCollective](#).

We'd love to hear from you! Say hi on Twitter at [@TabulaPDF](#)

## Latest Version: Tabula 1.2.1

June 4, 2018

Tabula 1.2.1 fixes several bugs in the user interface. You can read about all the changes in the [release notes](#).

Download Tabula below, or [on the release notes](#).

Special thanks to our [OpenCollective backers](#). If you find Tabula useful in your work, please contribute.

## How Can Tabula Help Me?

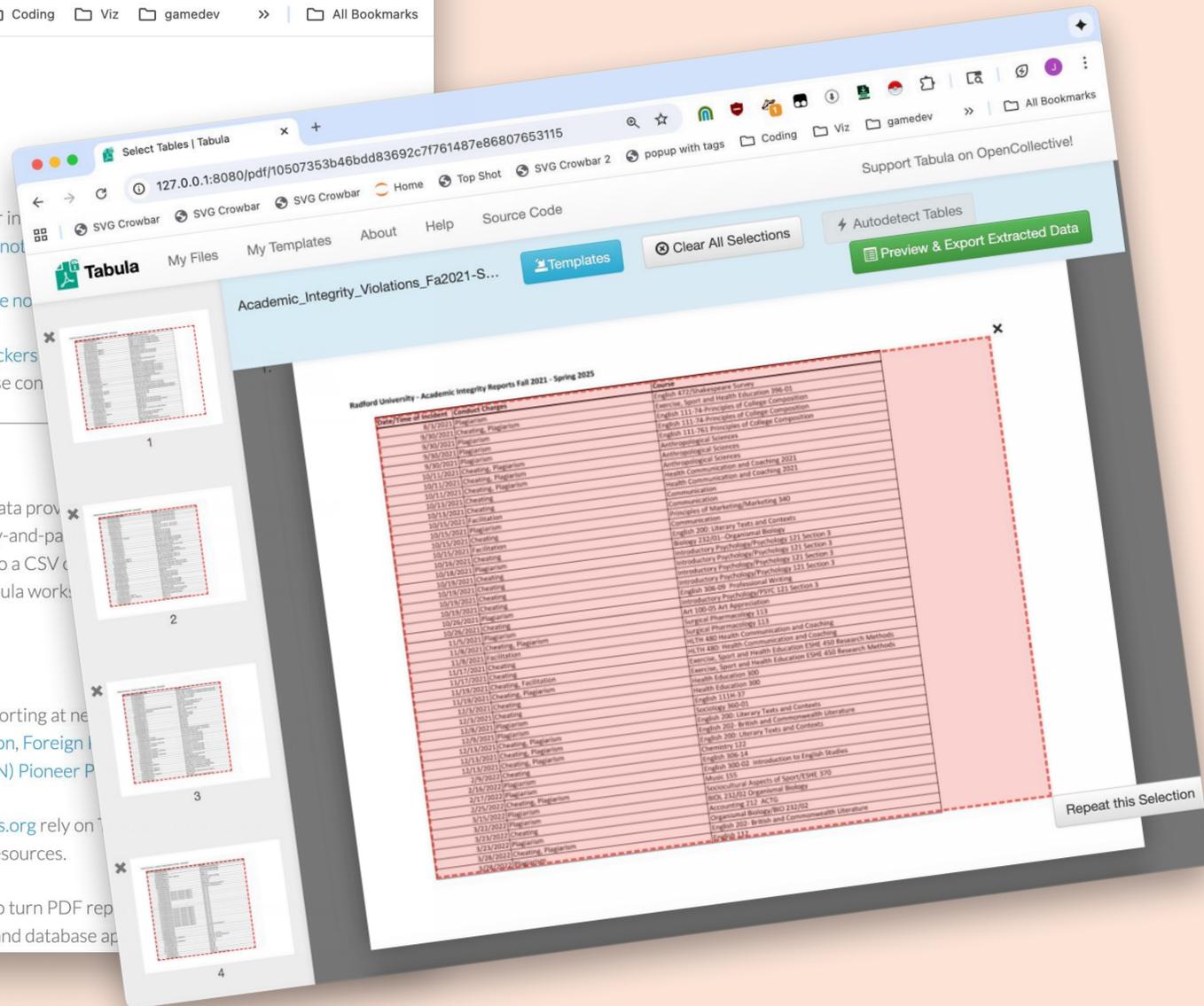
If you've ever tried to do anything with data provided in a PDF, you know how painful it is — there's no easy way to copy-and-paste the data. Tabula allows you to extract that data into a CSV or JSON file using a simple, easy-to-use interface. Tabula works on Windows, Mac, and Linux.

## Who Uses Tabula?

Tabula is used to power investigative reporting at news organizations including [ProPublica](#), [The Times of London](#), [Foreign Correspondence](#), [The New York Times](#) and the [St. Paul \(MN\) Pioneer Press](#).

Grassroots organizations like [SchoolCuts.org](#) rely on Tabula to turn PDF documents into human-friendly public resources.

And researchers of all kinds use Tabula to turn PDF reports into CSVs, and JSON files for use in analysis and database applications.



**We could use Tabu1a.**

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

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 Gijlan (Lot L)</a>	GJI004	€15,127	L106	€15,127	0	€0	0	€0
2	<a href="#">Njësia nr.02: Agrokultura Toka në Gijlan (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 Gllobocica Toka dhe Objekti</a>	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	<a href="#">Njësia nr.11: KB Novobërda Toka dhe Objekti</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 Gijlan (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 Gëriticë (Lot A)</a>	GJI011	€33,333	L96	€33,333	L82	€ 31,500	L36	€ 24,200
18	<a href="#">Njësia nr.18: Agromorava Toka në Sllatina 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 Parcela 01001-0 Muhaxheri i Epërm/Studime e Epërme</a>	MIT012	€0	0	€0	0	€0	0	€0
25	<a href="#">Njësia nr.25: Elan Parcela në Vërrnicë</a>	MIT012	€0	0	€0	0	€0	0	€0
26	<a href="#">Njësia nr.26: Prelluzha Parcela 00280-0</a>	MIT092	€5,000	L83	€5,000	0	€0	0	€0
27	<a href="#">Njësia nr.27: Prelluzha Parcela 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 Voinik</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 Dubovc</a>	MIT012	€0	0	€0	0	€0	0	€0
32	<a href="#">Njësia nr.32: Ngastrat 0150-0, 0151-0 dhe 0159-0</a>	MIT092	€5,000	L85	€5,000	0	€0	0	€0
33	<a href="#">Njësia nr.33: Ngastra 0118-0</a>	MIT092	€2,000	L86	€2,000	0	€0	0	€0
34	<a href="#">Njësia nr.34: Ngastrat 0007-0 dhe 0014-0</a>	MIT092	€22,100	L121	€22,100	L87	€ 12,000	0	€ 0
35	<a href="#">Njësia nr.35: Stacioni i Veterinës në Socanicë</a>	MIT095	€6,700	L64	€6,700	0	€0	0	€0
36	<a href="#">Njësia nr.36: Zyra e Jugobankës në Mitrovicë</a>	MIT101	€51,111	L128	€51,111	L130	€ 49,900	L129	€ 44,444
37	<a href="#">Njësia nr.37: KB Bec Prona në Lipovec</a>	PEJ003	€15,100	L07	€15,100	0	€0	0	€0
38	<a href="#">Njësia nr.38: KB Bec Toka Bujqësore në Shishmon të Bokës</a>	PEJ003	€12,000	L61	€12,000	0	€0	0	€0
39	<a href="#">Njësia nr. 39: KB Ponoshec Toka Bujqësore në Smolicë</a>	PEJ026	€3,190	L22	€3,190	0	€0	0	€0
40	<a href="#">Njësia nr.40: NSH Produkt Dyqani në Asllan Çeshme</a>	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 Gijlan (Lot L)</a>	<a href="#">GJI004</a>	€15,127	L106	€15,127	0	€0	0	€0
2	<a href="#">Njësia nr.02: Agrokultura Toka në Gijlan (Lot M)</a>	<a href="#">GJI004</a>	€0	0	€0	0	€0	0	€0
3	<a href="#">Njësia nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)</a>	<a href="#">GJI011</a>	€69,611	L76	€69,611	L34	€58,000	L63	€12,222
4	<a href="#">Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj</a>	<a href="#">GJI011</a>	€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>	<a href="#">GJI011</a>	€3,892	L57	€3,892	0	€0	0	€0
6	<a href="#">Njësia nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)</a>	<a href="#">GJI011</a>	€16,666	L60	€16,666	L28	€16,550	0	€0
7	<a href="#">Njësia nr.07: Qëndresa Lokali 1 Kamenicë</a>	<a href="#">GJI101</a>	€169,000	L21	€169,000	L15	€125,559	L72	€91,153
8	<a href="#">Njësia nr.08: Qëndresa Lokali 2 Kamenicë</a>	<a href="#">GJI101</a>	€51,499	L16	€51,499	L19	€51,000	0	€0
9	<a href="#">Njësia nr. 09: Qëndresa Lokali 3 Kamenicë</a>	<a href="#">GJI101</a>	€28,153	L71	€28,153	L47	€22,200	L17	€21,599
10	<a href="#">Njësia nr.10: KB Gllboçica Toka dhe Objekti</a>	<a href="#">GJI141</a>	€28,700	L38	€28,700	0	€0	0	€0
11	<a href="#">Njësia nr.11: KB Novobërda Toka dhe Objekte</a>	<a href="#">GJI142</a>	€5,730	L29	€5,730	0	€0	0	€0
12	<a href="#">Njësia nr.12: KB Bashkimi Toka dhe Objekti në Mirash</a>	<a href="#">GJI008</a>	€10,213	L108	€10,213	0	€0	0	€0
13	<a href="#">Njësia nr.13: Agrokultura Toka në Gijlan (Lot E)</a>	<a href="#">GJI004</a>	€21,000	L104	€21,000	L99	€20,250	0	€0
14	<a href="#">Njësia nr.14: Agrokultura Toka në Shillovë (Lot B)</a>	<a href="#">GJI004</a>	€15,556	L25	€15,556	L103	€15,100	0	€0
15	<a href="#">Njësia nr.15: Pasuria Bujqësore Toka në Sojevë</a>	<a href="#">GJI011</a>	€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 Gërlicë (Lot A)</a>	<a href="#">GJI011</a>	€33,333	L96	€33,333	L82	€31,500	L36	€24,200
18	<a href="#">Njësia nr.18: Agromorava Toka në Sllatina e Poshtme (Lot B)</a>	<a href="#">GJI035</a>	€0	0	€0	0	€0	0	€0
19	<a href="#">Njësia nr.19: Produkti Tokë në Baks</a>	<a href="#">MIT008</a>	€0	0	€0	0	€0	0	€0
20	<a href="#">Njësia nr.20: Produkti Tokë në Brojë 1</a>	<a href="#">MIT008</a>	€0	0	€0	0	€0	0	€0
21	<a href="#">Njësia nr.21: Produkti Tokë në Brojë 2</a>	<a href="#">MIT008</a>	€0	0	€0	0	€0	0	€0
22	<a href="#">Njësia nr.22: Produkti Tokë në Brojë 3</a>	<a href="#">MIT008</a>	€1,888	L50	€1,888	L51	€1,491	0	€0
23	<a href="#">Njësia nr.23: Produkti Tokë në Brojë 4</a>	<a href="#">MIT008</a>	€200	L09	€200	0	€0	0	€0
24	<a href="#">Njësia nr.24: Elan Parcela 01001-0 Muhaxheri i Epërm/Studime e Epërme</a>	<a href="#">MIT012</a>	€0	0	€0	0	€0	0	€0
25	<a href="#">Njësia nr.25: Elan Parcela në Vërnice</a>	<a href="#">MIT012</a>	€0	0	€0	0	€0	0	€0
26	<a href="#">Njësia nr.26: Prelluzha Parcela 00280-0</a>	<a href="#">MIT092</a>	€5,000	L83	€5,000	0	€0	0	€0
27	<a href="#">Njësia nr.27: Prelluzha Parcela 00290-0</a>	<a href="#">MIT092</a>	€2,000	L84	€2,000	0	€0	0	€0
28	<a href="#">Njësia nr.28: Lokali Afarist nr.2</a>	<a href="#">PRN011</a>	€599,999	L98	€599,999	L105	€115,200	L110	€100,000
29	<a href="#">Njësia nr.29: Tokë në Vojnik</a>	<a href="#">MIT008</a>	€0	0	€0	0	€0	0	€0
30	<a href="#">Njësia nr.30: Tokë në Klinë e Epërme 1</a>	<a href="#">MIT008</a>	€7,120	L13	€7,120	L11	€6,200	0	€0
31	<a href="#">Njësia nr.31: Elan Tokë në Dubeve</a>	<a href="#">MIT012</a>	€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	€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

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.	Njësiti	AKP ID	Çmimi më i Lartë	Të Ofertuesit me çmim më të Lartë							
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi		
1	<a href="#">Njësia nr.01: Agrokultura Toka në Gjilan (Lot L)</a>				0		€0		0		€0
2	<a href="#">Njësia nr.02: Agrokultura Toka në Gjilan (Lot M)</a>				0		€0		0		€0
3	<a href="#">Njësia nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)</a>				0		€0		0		€0
4	<a href="#">Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj</a>				0		€0		0		€0
5	<a href="#">Njësia nr.05: Pasuria Bujqësore Toka në Pojatë (Lot B)</a>				0		€0		0		€0
6	<a href="#">Njësia nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)</a>				0		€0		0		€0



Nr.	Njësiti	AKP ID	Çmimi më i Lartë	Të Ofertuesit me çmim më të Lartë					
				Ofertuesi	Çmimi	Ofertuesi	Çmimi	Ofertuesi	Çmimi
41	Njësia nr.41: Nën Deva ngastrat nr.705 25 Dvrit	PK002	0	0	0	0	0	0	0
42	Njësia nr.42: Nën Malleshajti ngastrat nr. 247-1 në Drenashtje	PK012	0	0	0	0	0	0	0
43	Njësia nr.43: FAP Rrethi i Gjakovës në Shkup	PK013	0	0	0	0	0	0	0
44	Njësia nr.44: FAP Rrethi i Gjakovës në Shkup	PK014	0	0	0	0	0	0	0
45	Njësia nr.45: FAP Rrethi i Gjakovës në Shkup	PK015	0	0	0	0	0	0	0
46	Njësia nr.46: FAP Rrethi i Gjakovës në Shkup	PK016	0	0	0	0	0	0	0
47	Njësia nr.47: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK017	0	0	0	0	0	0	0
48	Njësia nr.48: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK018	0	0	0	0	0	0	0
49	Njësia nr.49: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK019	0	0	0	0	0	0	0
50	Njësia nr.50: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK020	0	0	0	0	0	0	0
51	Njësia nr.51: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK021	0	0	0	0	0	0	0
52	Njësia nr.52: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK022	0	0	0	0	0	0	0
53	Njësia nr.53: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK023	0	0	0	0	0	0	0
54	Njësia nr.54: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK024	0	0	0	0	0	0	0
55	Njësia nr.55: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK025	0	0	0	0	0	0	0
56	Njësia nr.56: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK026	0	0	0	0	0	0	0
57	Njësia nr.57: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK027	0	0	0	0	0	0	0
58	Njësia nr.58: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK028	0	0	0	0	0	0	0
59	Njësia nr.59: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK029	0	0	0	0	0	0	0
60	Njësia nr.60: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK030	0	0	0	0	0	0	0
61	Njësia nr.61: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK031	0	0	0	0	0	0	0
62	Njësia nr.62: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK032	0	0	0	0	0	0	0
63	Njësia nr.63: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK033	0	0	0	0	0	0	0
64	Njësia nr.64: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK034	0	0	0	0	0	0	0
65	Njësia nr.65: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK035	0	0	0	0	0	0	0
66	Njësia nr.66: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK036	0	0	0	0	0	0	0
67	Njësia nr.67: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK037	0	0	0	0	0	0	0
68	Njësia nr.68: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK038	0	0	0	0	0	0	0
69	Njësia nr.69: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK039	0	0	0	0	0	0	0
70	Njësia nr.70: Nën Emini Duraku - Dymari nr. 3 afër Gjimnastit, Gjakovë	PK040	0	0	0	0	0	0	0

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

what you're doing (and how)

```
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)
```

your current code

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

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's 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 so we can 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
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 €12,200
3 4 Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj GJI011 €1,111,000 L133 €1,111,000 <NA> L58 €666,666 <NA> L59 €513,501
4 5 Njësia nr.05: Pasuria Bujqësore Toka në Pojatë... GJI011 €3,892 L57 €3,892 <NA> 0 €0 <NA> 0 €0
```

the output

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 tables in the PDF. I want to get 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 ['Nr.', 'Njësia', 'Akti i D.', '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
```

< > bid-pdfs ☰ »

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 3:11 PM	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 3:41 PM	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
86A703E9-6313-49...-765CCA369476.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.

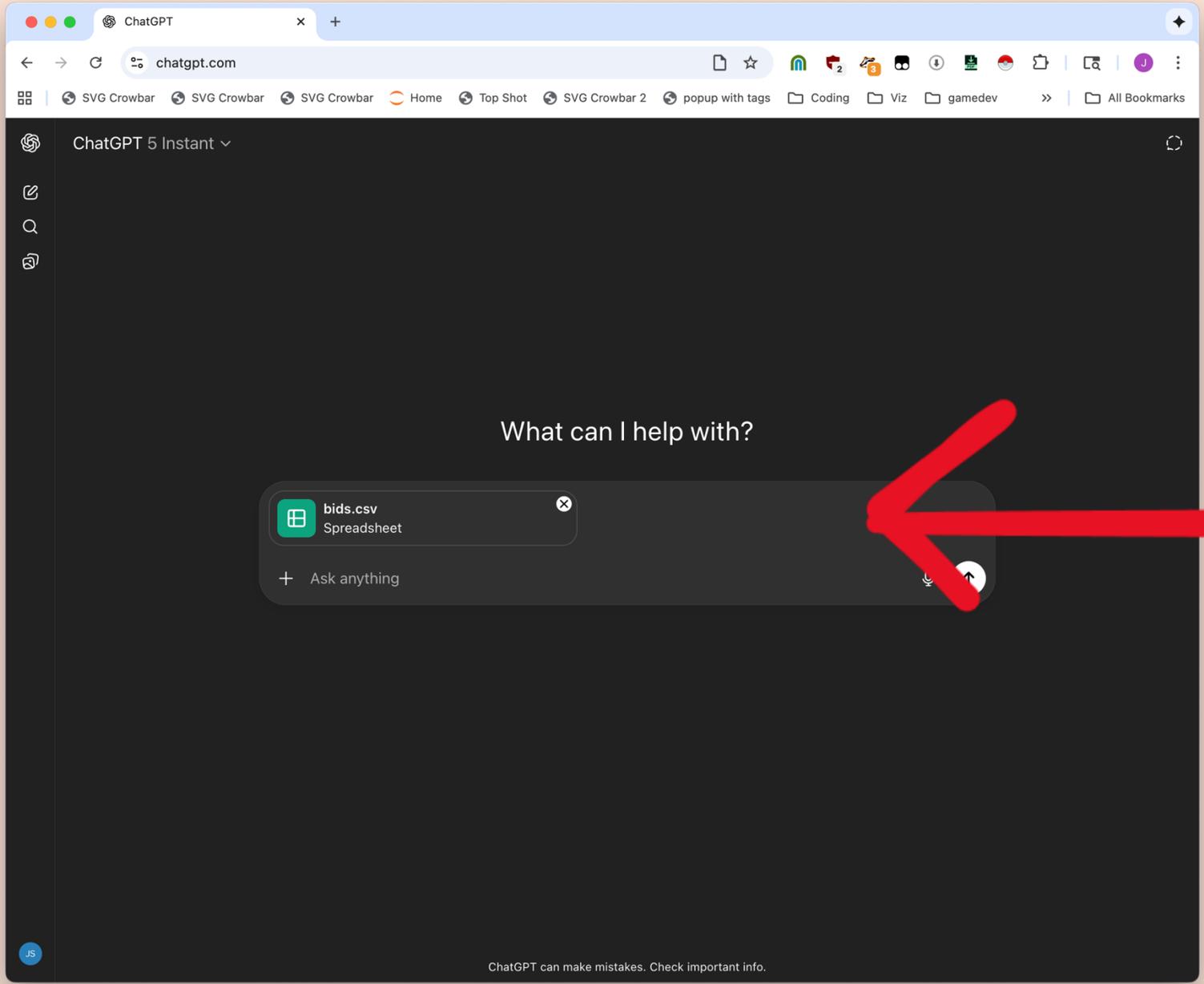
Some of them have an extra column, the last one is "number of bidders." We can save that and keep it empty for the ones that don't have it.

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.



What can I help with?

bids.csv  
Spreadsheet

+ Ask anything

ChatGPT can make mistakes. Check important info.