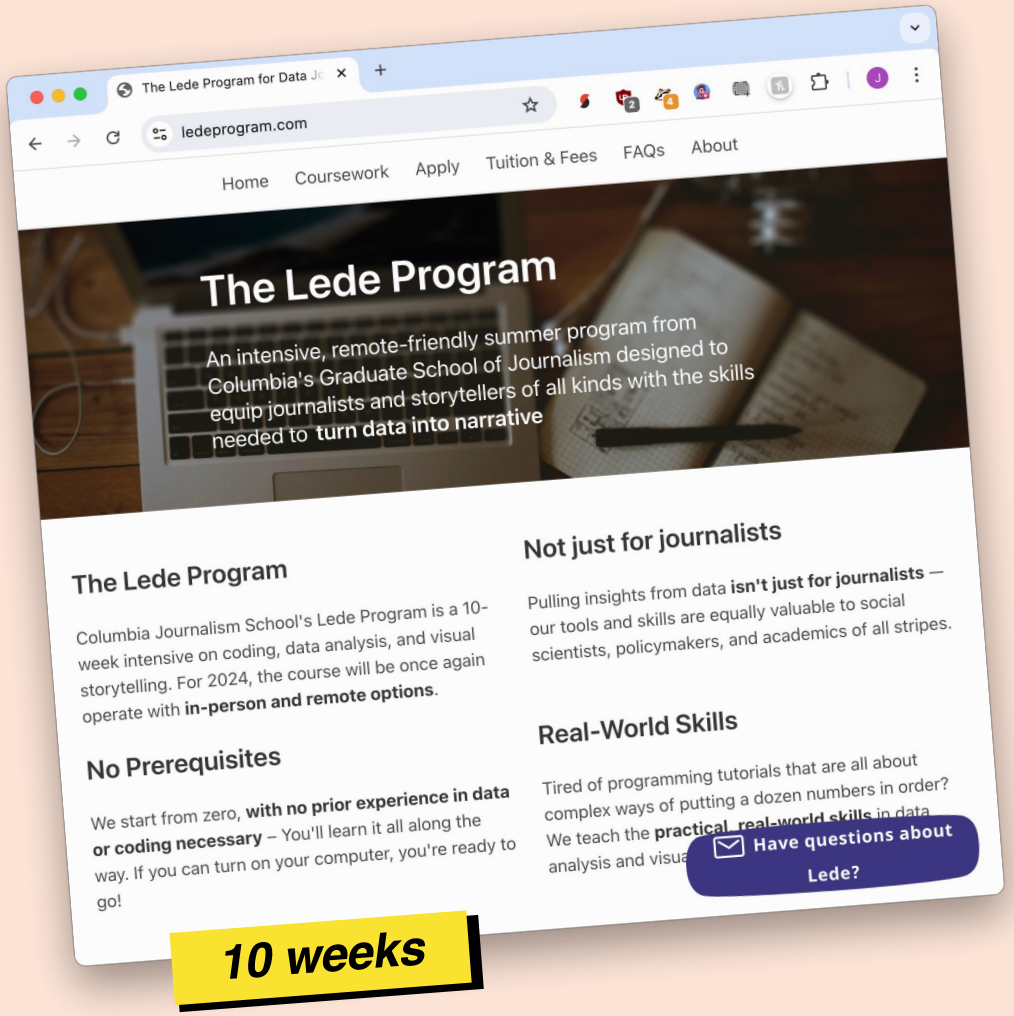
Advanced Data Journalism
Part I: Structured Data

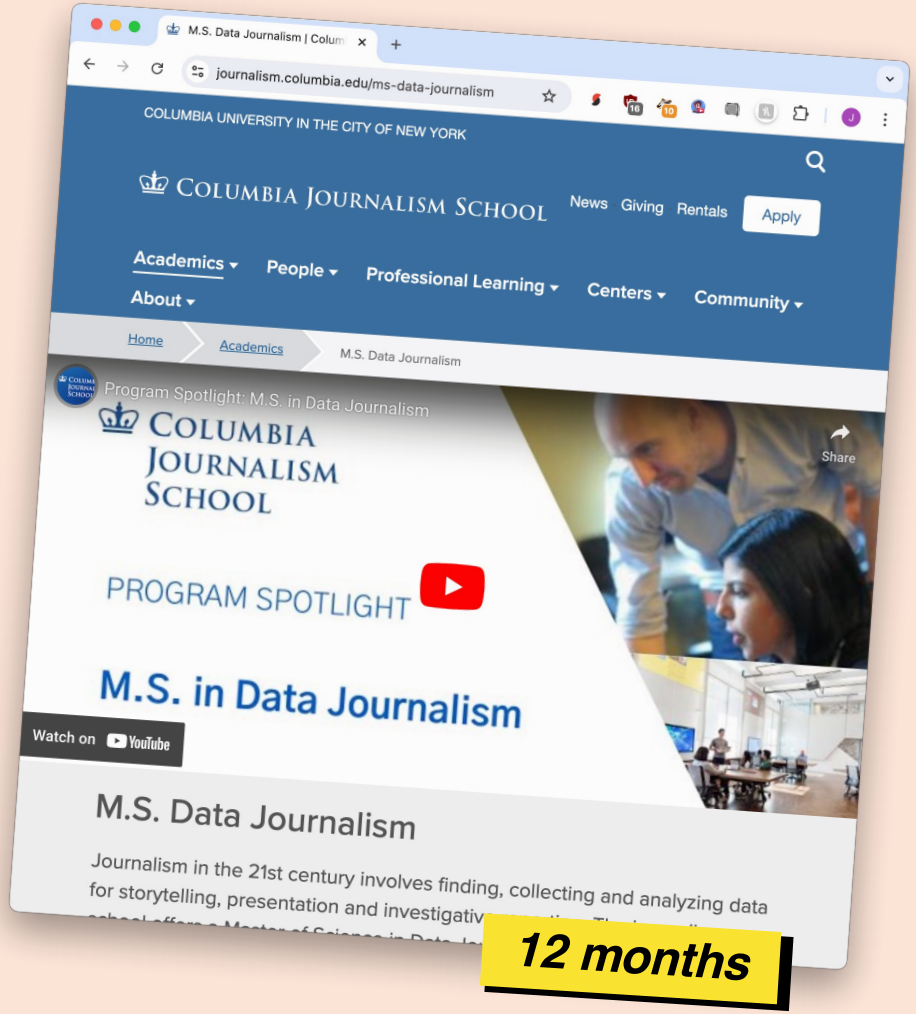
Jonathan Soma

Columbia Journalism School

js4571@Columbia.edu



10 weeks



12 months

What is data journalism?

Nothing special, honestly.

The New York Times

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.

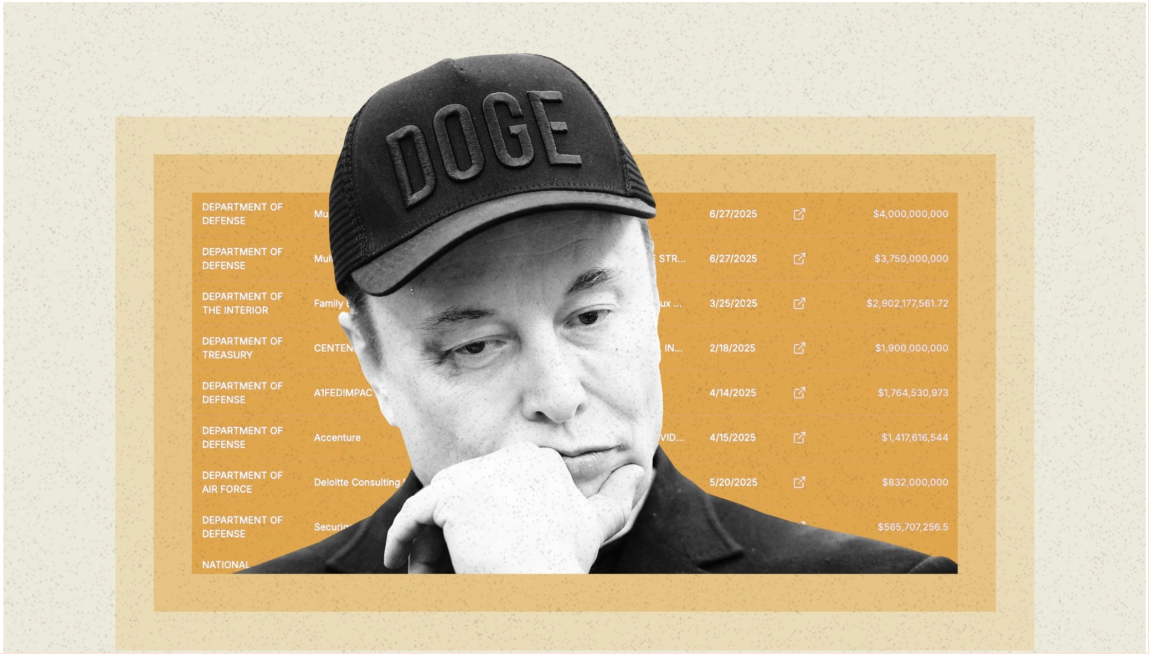
Revenue for the quarter ending in June slipped 13.4 percent. The company said it had lost $441 million, compared with a loss of $5 million in the same quarter last year. The drop was largely due to an accounting charge of $407 million, reflecting the falling value of the chain.

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.

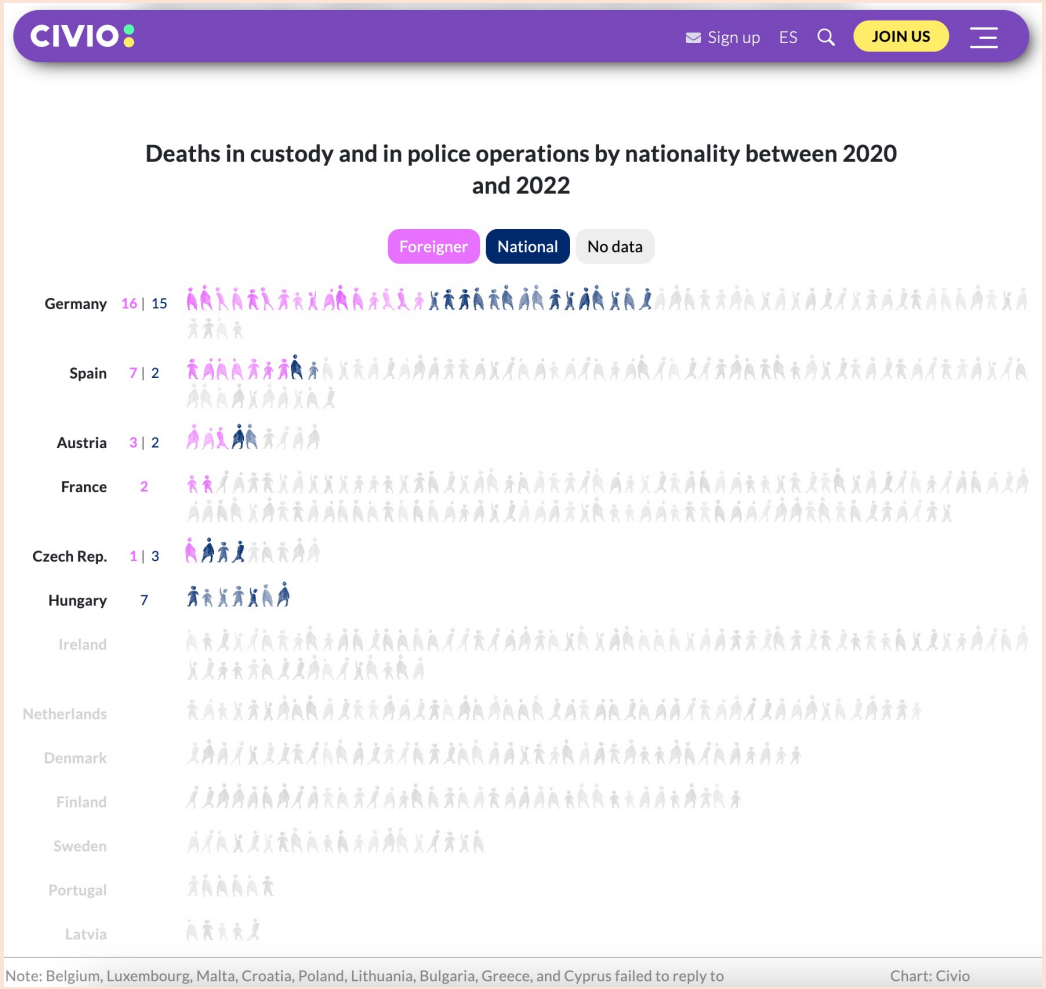


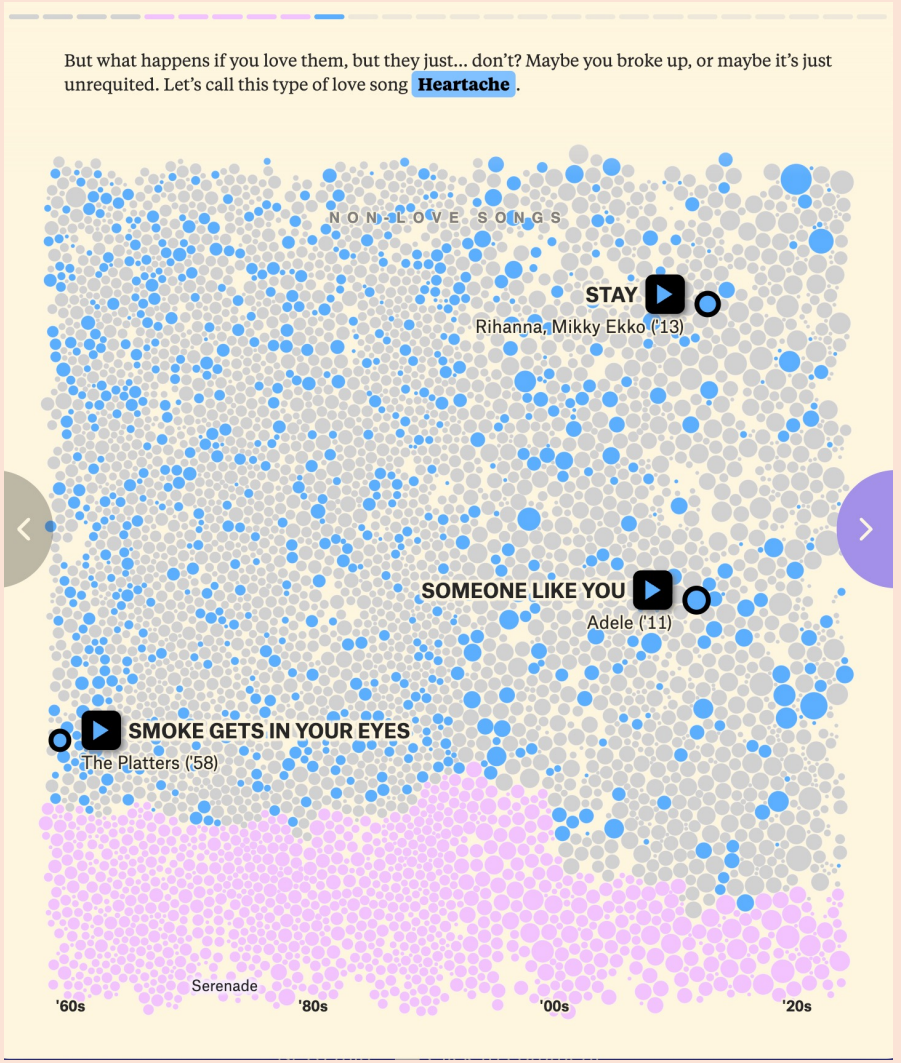
The New York Times

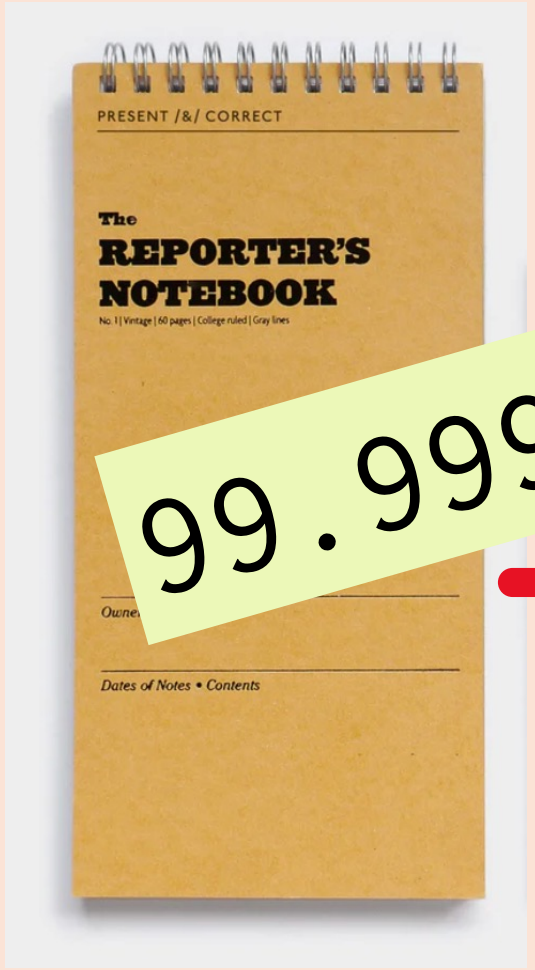
In the West, Lightning Grows as a Cause of Damaging Fires

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.


CIVIO
Sign up
ES
JOIN US
Deaths in custody and in police operations by nationality between 2020 and 2022
Foreigner
National
No data
Germany 16 | 15
Spain 7 | 2
Austria 3 | 2
France 2
Czech Rep. 1 | 3
Hungary 7
Ireland
Netherlands
Denmark
Finland
Sweden
Portugal
Latvia
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.
NON-LOVE SONGS
STAY
Rihanna, Mikky Ekko ('13)
SOMEONE LIKE YOU
Adele ('11)
SMOKE GETS IN YOUR EYES
The Platters ('58)
Serenade
'60s
'80s
'00s
'20s



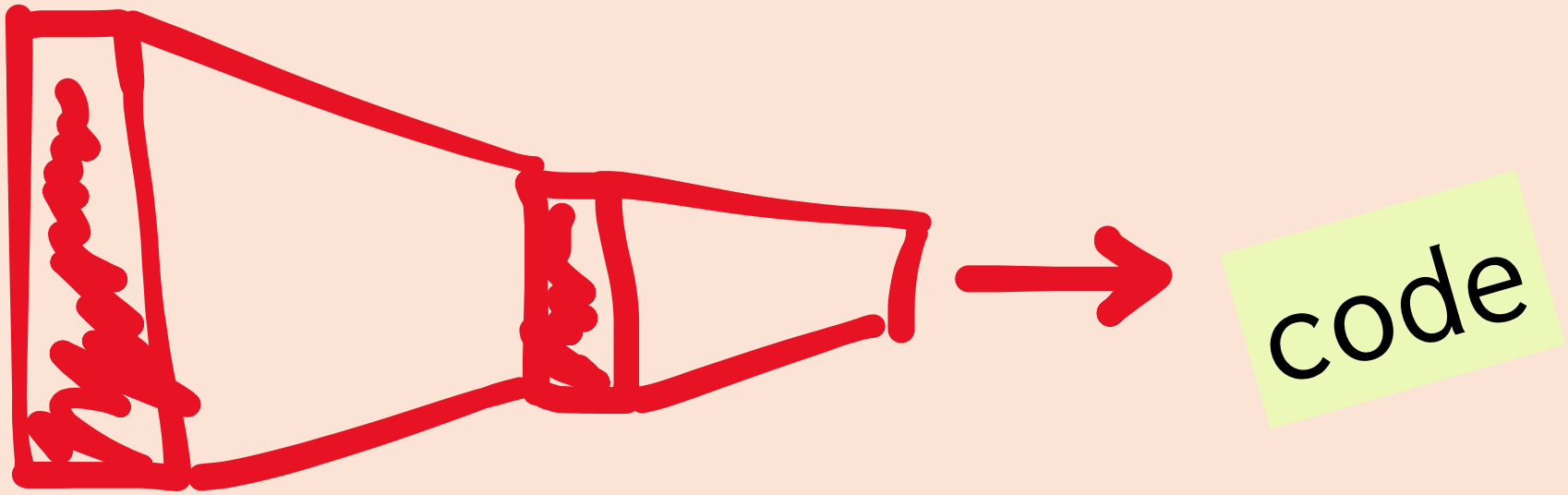
99.999%

99.999%

Untitled spreadsheet

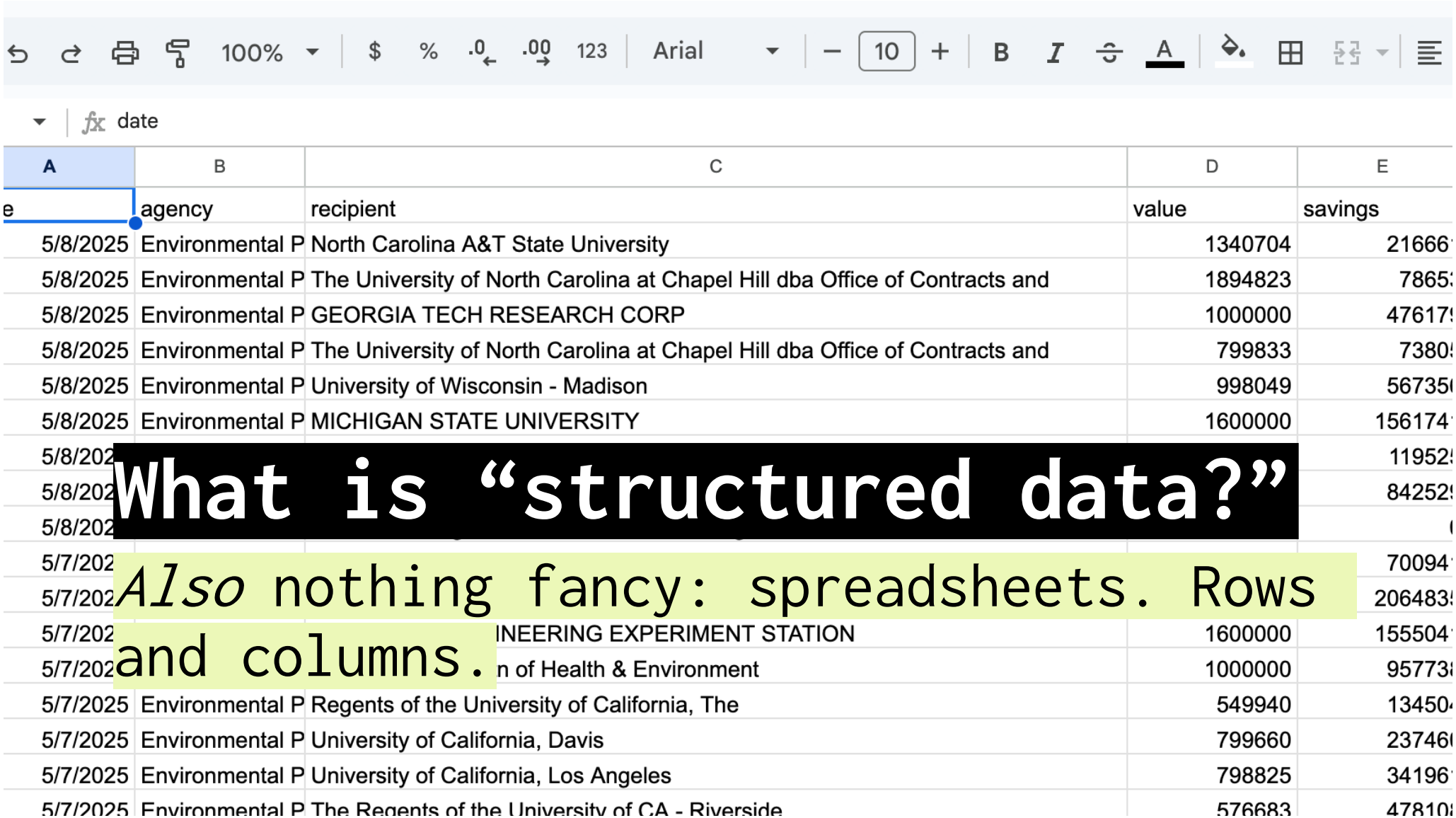
	A	B	C	D	E	F	G
1	date	agency	recipient	value	savings	link	description
2	5/8/2025	Environmental P	North Carolina A&T State University	1340704	216661	https://usaspendin	RESEARCHERS WILL PERFORM STUDIES
3	5/8/2025	Environmental P	The University of North Carolina at Chapel Hill dba Office of Contracts and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

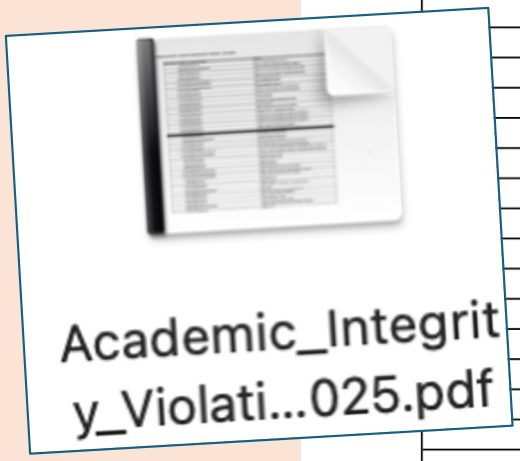
GRANTS



0.0001% of 0.0001%

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


100%
$
%
123
Arial
10
B
I
A
date
A
B
C
D
E
e
agency
recipient
value
savings
5/8/2025 Environmental P North Carolina A&T State University 1340704
5/8/2025 Environmental P The University of North Carolina at Chapel Hill dba Office of Contracts and 1894823
5/8/2025 Environmental P GEORGIA TECH RESEARCH CORP 1000000
5/8/2025 Environmental P The University of North Carolina at Chapel Hill dba Office of Contracts and 799833
5/8/2025 Environmental P University of Wisconsin - Madison 998049
5/8/2025 Environmental P MICHIGAN STATE UNIVERSITY 1600000
What is “structured data?”
Also nothing fancy: spreadsheets. Rows and columns.
5/7/2025 Environmental P Regents of the University of California, The 549940
5/7/2025 Environmental P University of California, Davis 799660
5/7/2025 Environmental P University of California, Los Angeles 798825



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

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

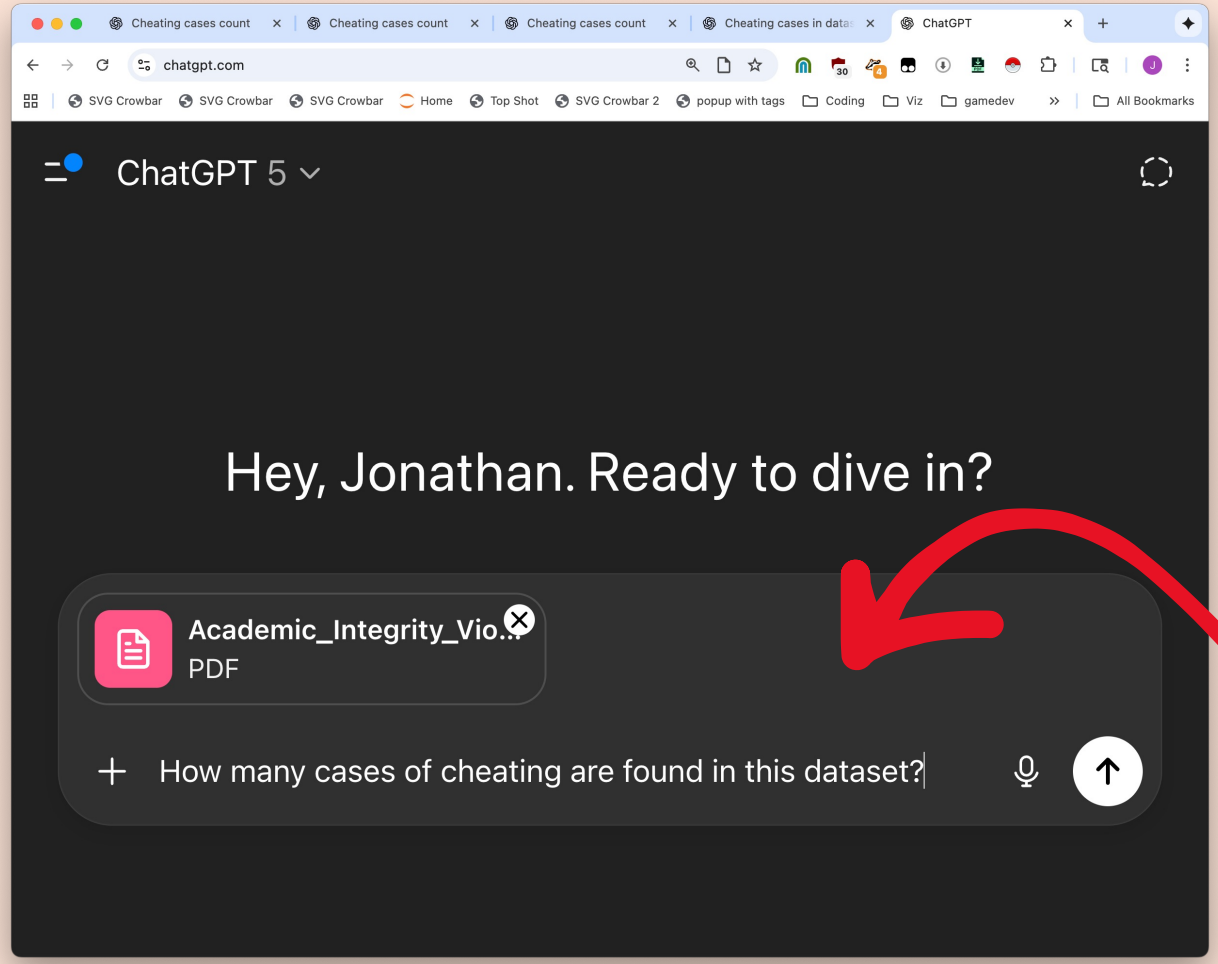
Why can't we just use AI for everything???

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

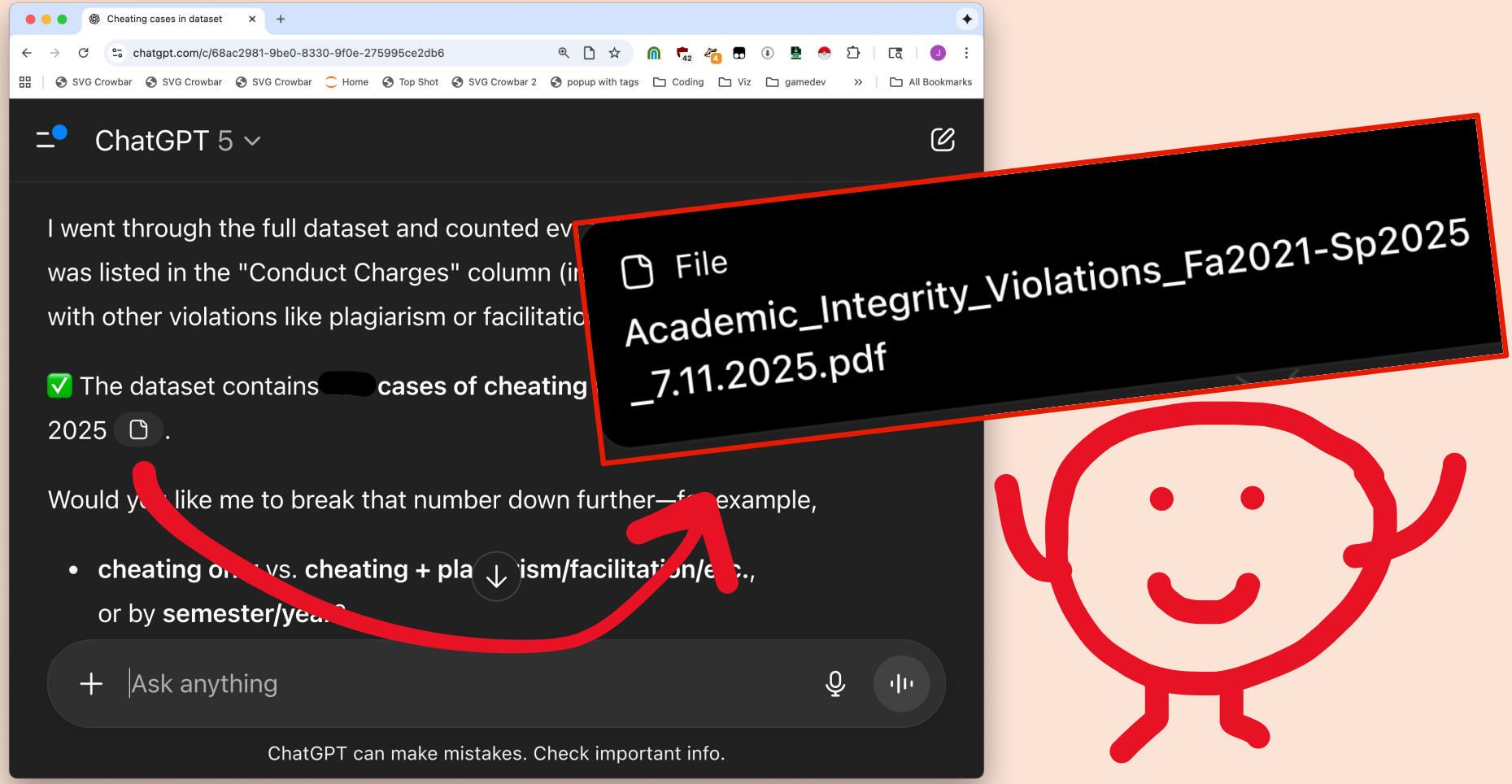
📦 Data: 00-cheating-data.zip

Links:

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


Cheating cases count
Cheating cases count
Cheating cases count
Cheating cases in data
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
Hey, Jonathan. Ready to dive in?
Academic_Integrity_Vio...
PDF
How many cases of cheating are found in this dataset?


Academic_Integrit
y_Violati...025.pdf


Cheating cases in dataset
chatgpt.com/c/68ac2981-9be0-8330-9f0e-275995ce2db6
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
Top Shot
SVG Crowbar 2
popup with tags
Coding
Viz
gamedev
All Bookmarks
ChatGPT 5
I went through the full dataset and counted ev
was listed in the "Conduct Charges" column (i
with other violations like plagiarism or facilitatio
✅ The dataset contains [redacted] cases of cheating
2025 .
Would y like me to break that number down further—f example,
cheating o vs. cheating + pla ism/facilitati n/e c.,
or by semester/yea
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**

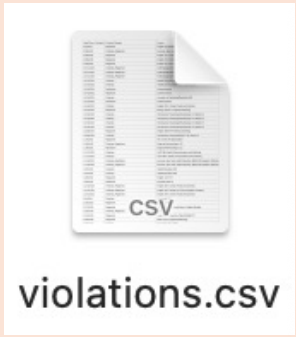
The dataset contains **143 cases of cheating** (this includes both cheating alone and cases where cheating is listed alongside other violations, like "Cheating, Plagiarism").

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



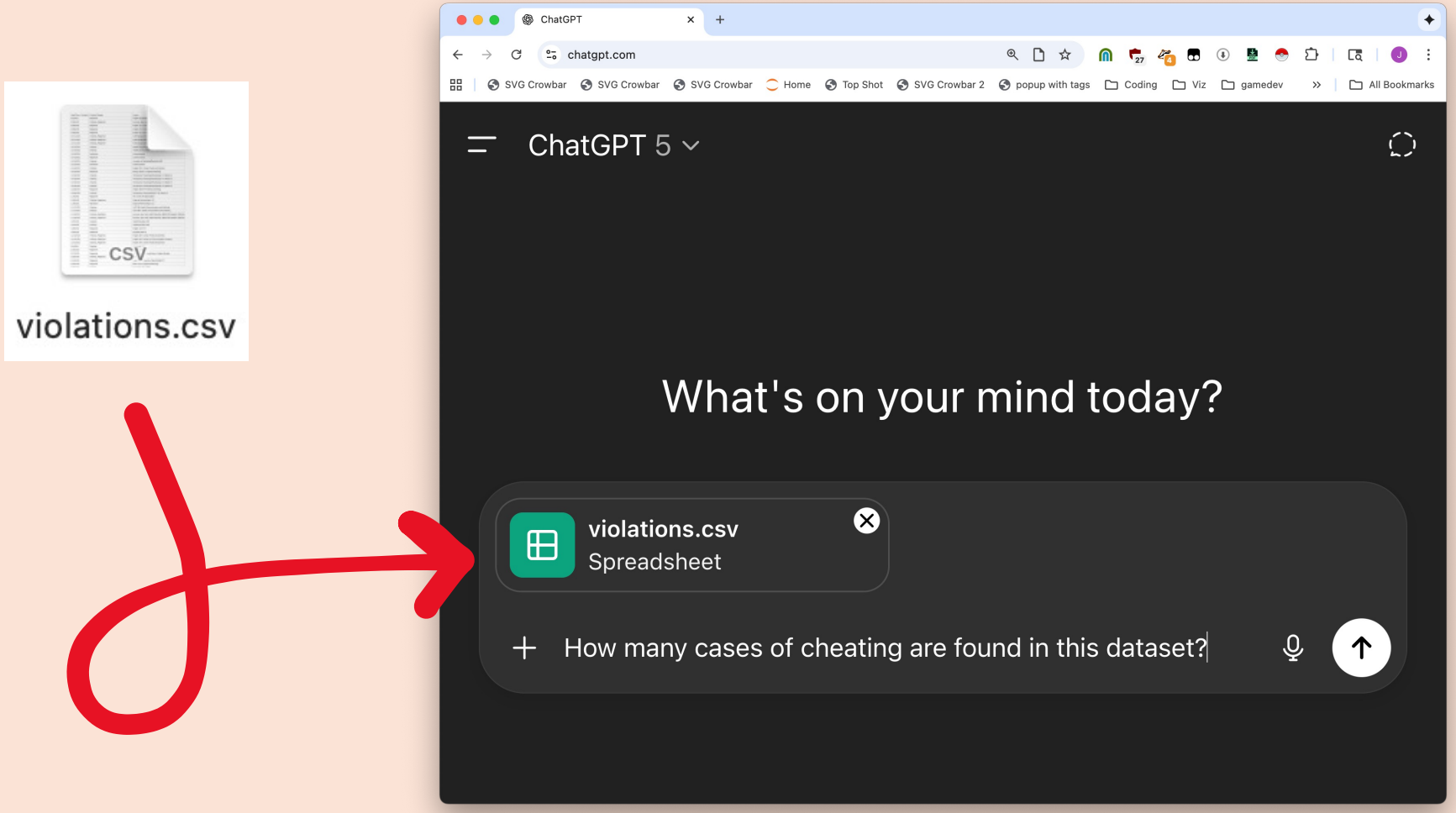
AutoSave violations

Home Insert Draw Page Layout Formulas Data Review View Automate Acrobat

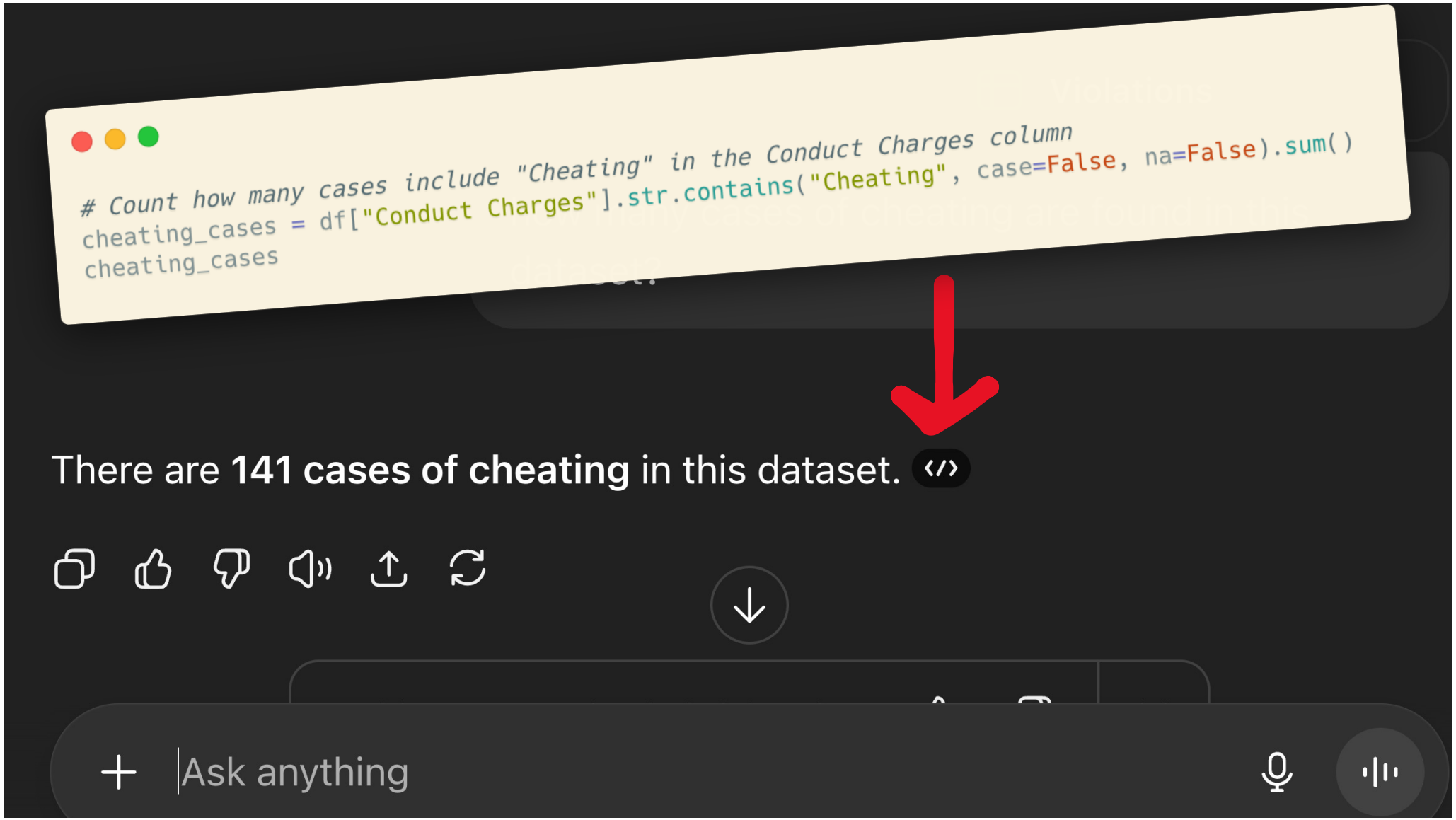
Paste Aptos Narrow (Bod... 12 General Conditional Formatting Format as Table Cell Styles

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						


violations.csv
CSV
violations.csv
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?


There are 141 cases of cheating
There are 141 cases of cheating
There are 141 cases of cheating
There are 141 cases of cheating
Ask anything


Violations
# 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

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

Static Datasets

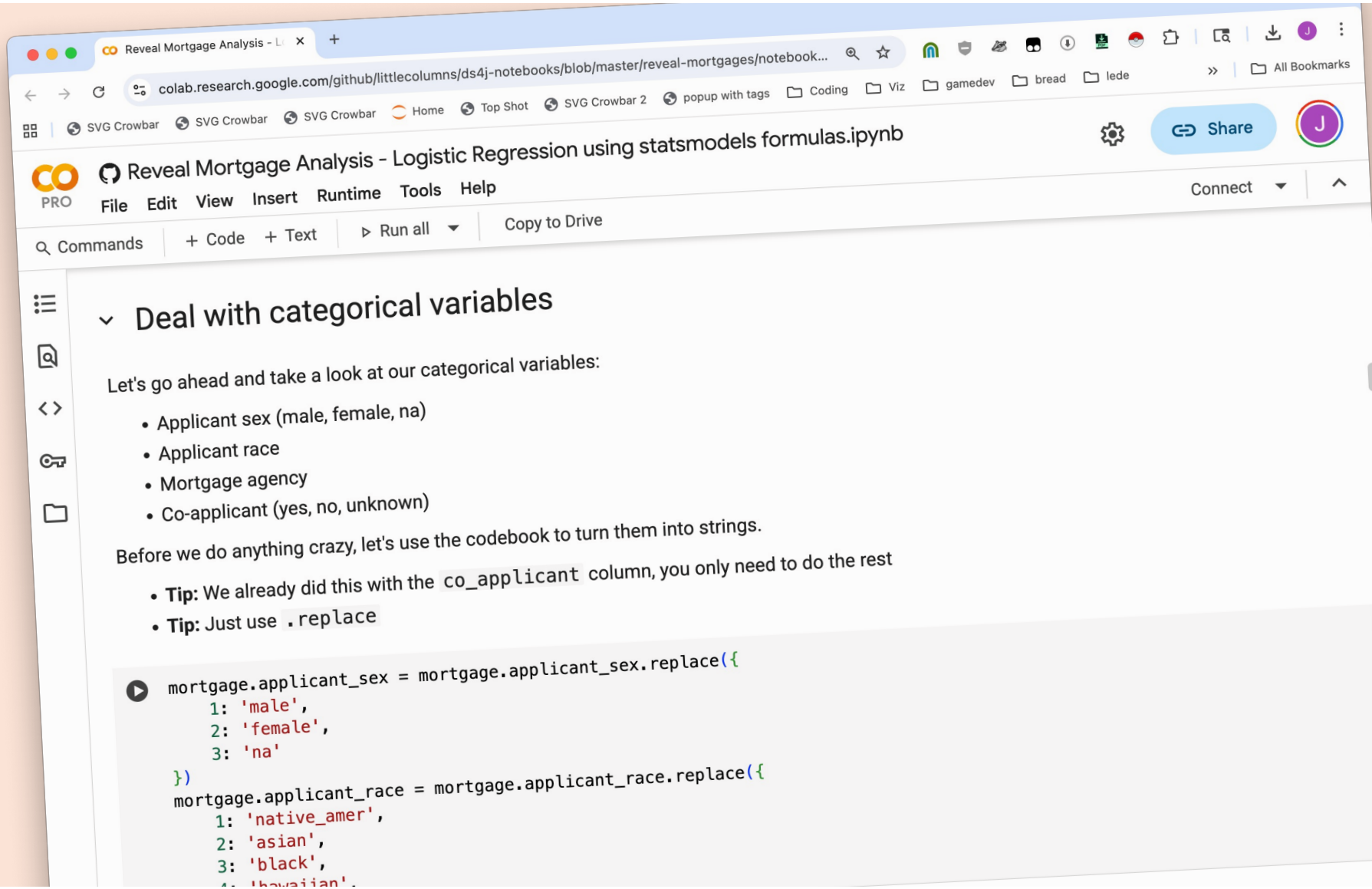
Snapshot National Loan-Level Dataset
The snapshot files contain the national HMDA datasets as of a fixed date for all HMDA reporters, as modified by the Bureau to protect applicant and borrower privacy.

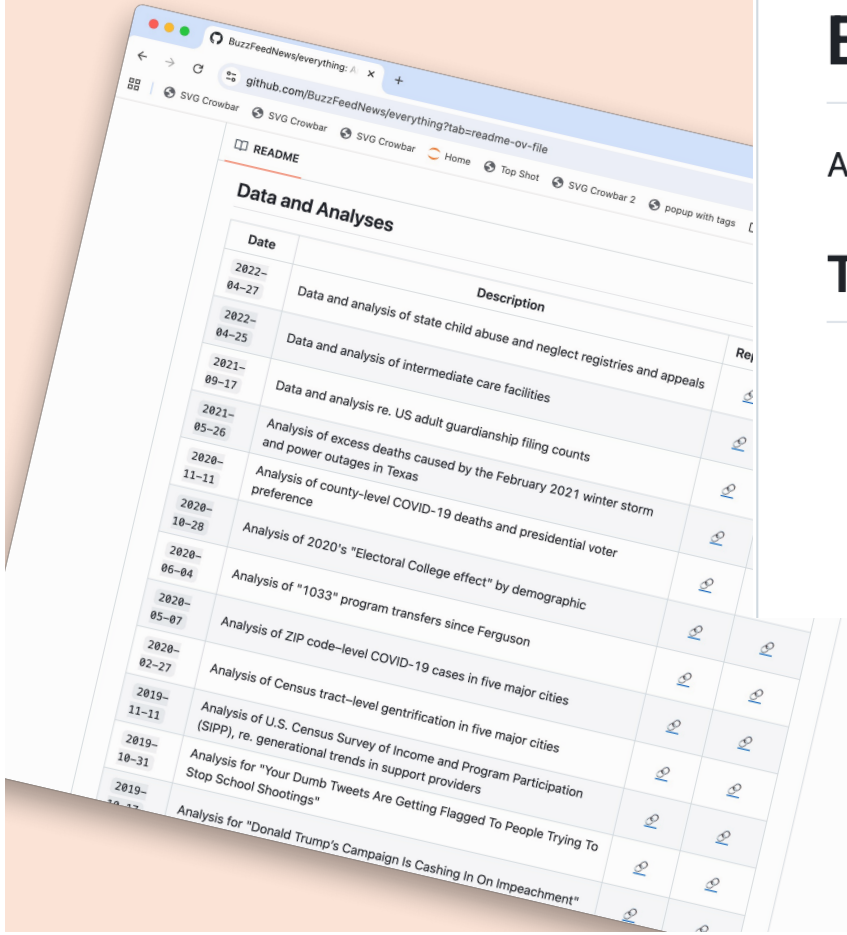
→ **Data Freeze Date:** May 19, 2025

→ **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
colab.research.google.com/github/littlecolumns/ds4j-notebooks/blob/master/reveal-mortgages/notebook...
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
Top Shot
SVG Crowbar 2
popup with tags
Coding
Viz
gamedev
bread
lede
All Bookmarks
Reveal Mortgage Analysis - Logistic Regression using statsmodels formulas.ipynb
Share
PRO
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',

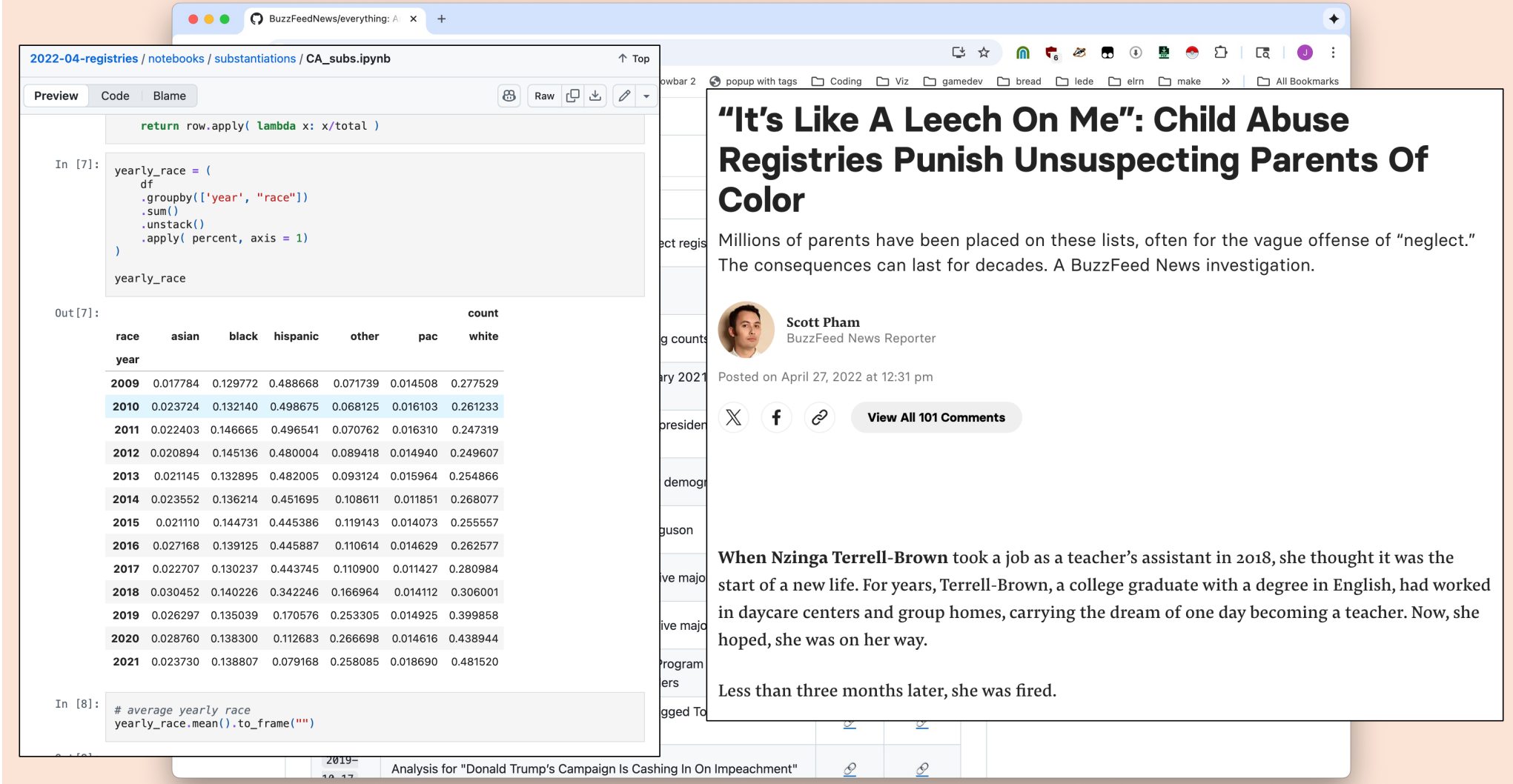


README

BuzzFeedNews/everything

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

Table of Contents



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]:

	count					
race	asian	black	hispanic	other	pac	white
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.

We Trained A Computer To Se

buzzfeednews.com/article/peteraldhous/hidden-spy-planes

SVG Crowbar | SVG Crowbar | SVG Crowbar | Home | Top Shot | SVG Crowbar 2 | popup with tags | Coding

BuzzFeed News

SIGN IN ABOUT US GOT A

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.

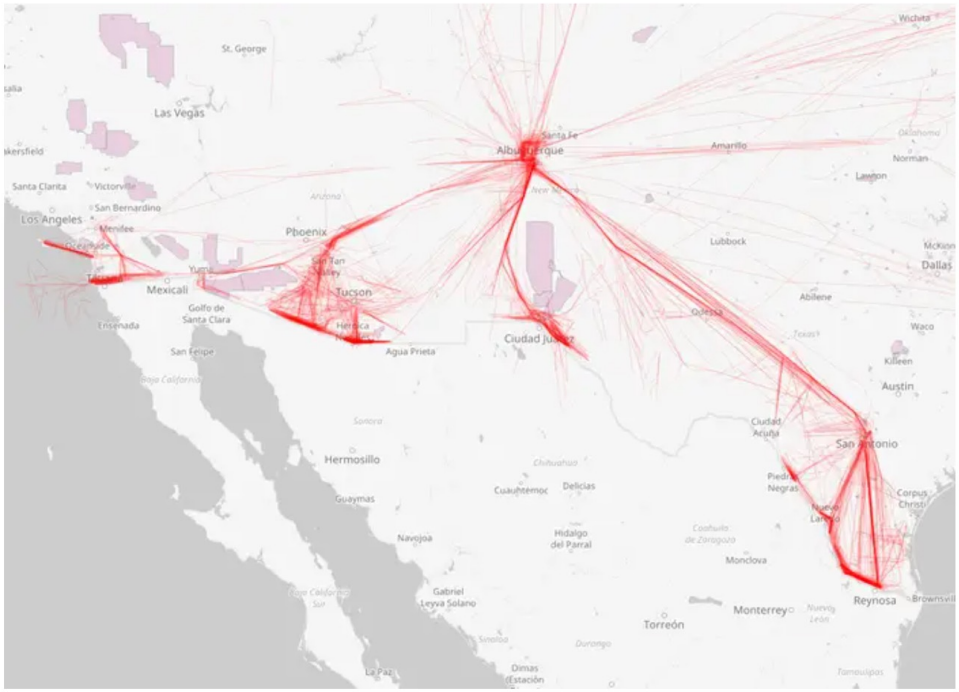
Peter Aldhous
BuzzFeed News Reporter

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



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 algorthim 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 manufacter 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.)

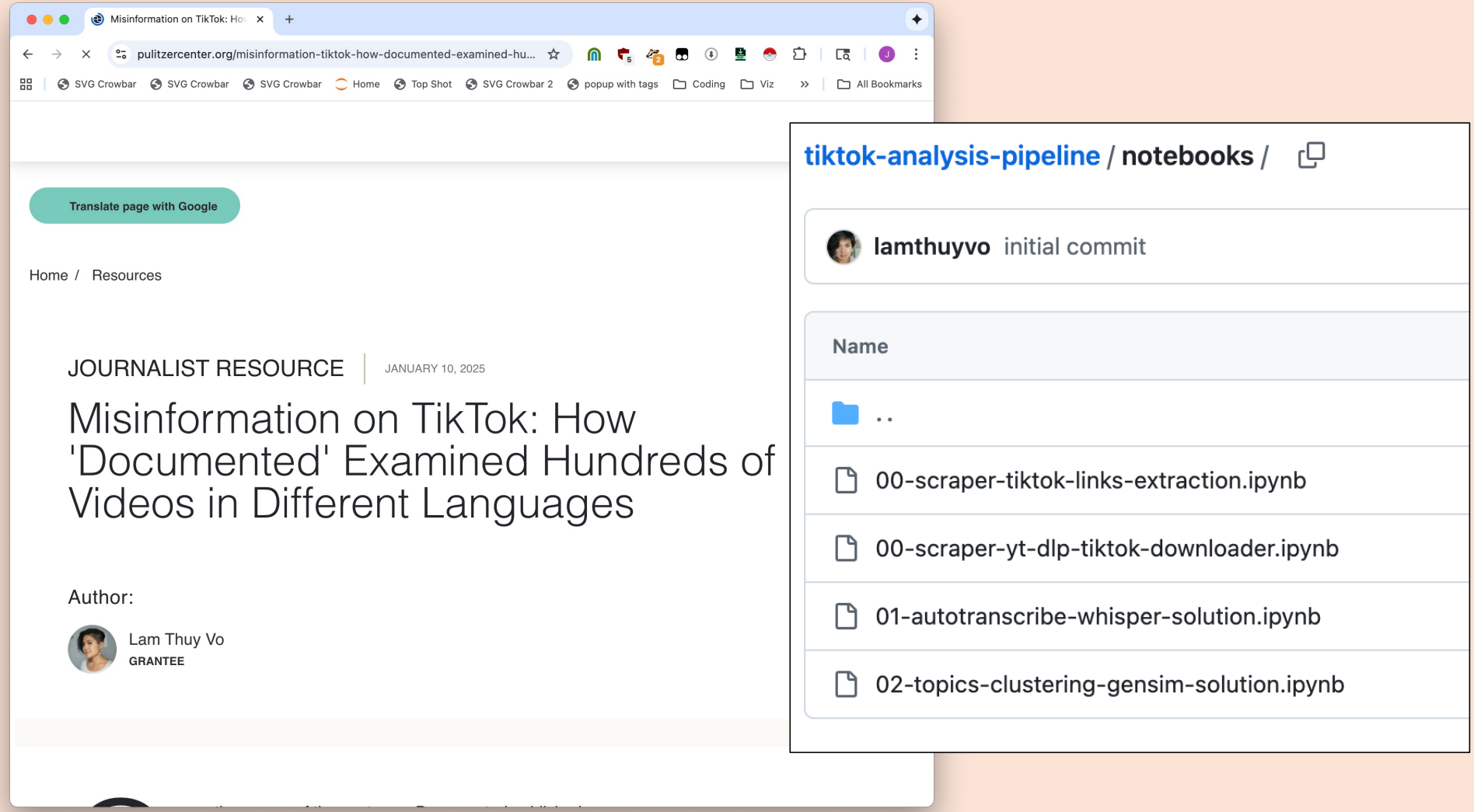
As training data, drawn from `planes_features.csv`, we used 97 fixed-wing FBI and DHS planes from our previous story, given a `class` of `surveil`, and a random sample of 500 other planes, given a `class` of `other`.

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: Ho
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

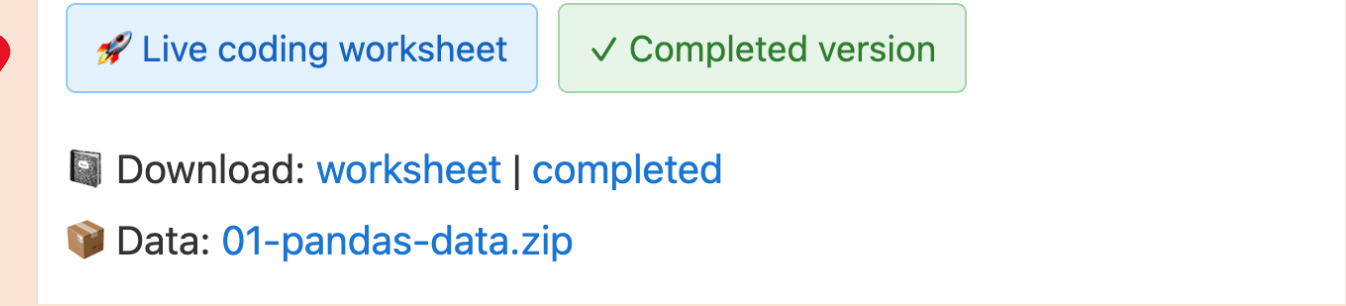
- Learn to run Python
- Learn one fundamental tool (pandas)
- That’s it!!!!

Two worst parts of a data journalist's life

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

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

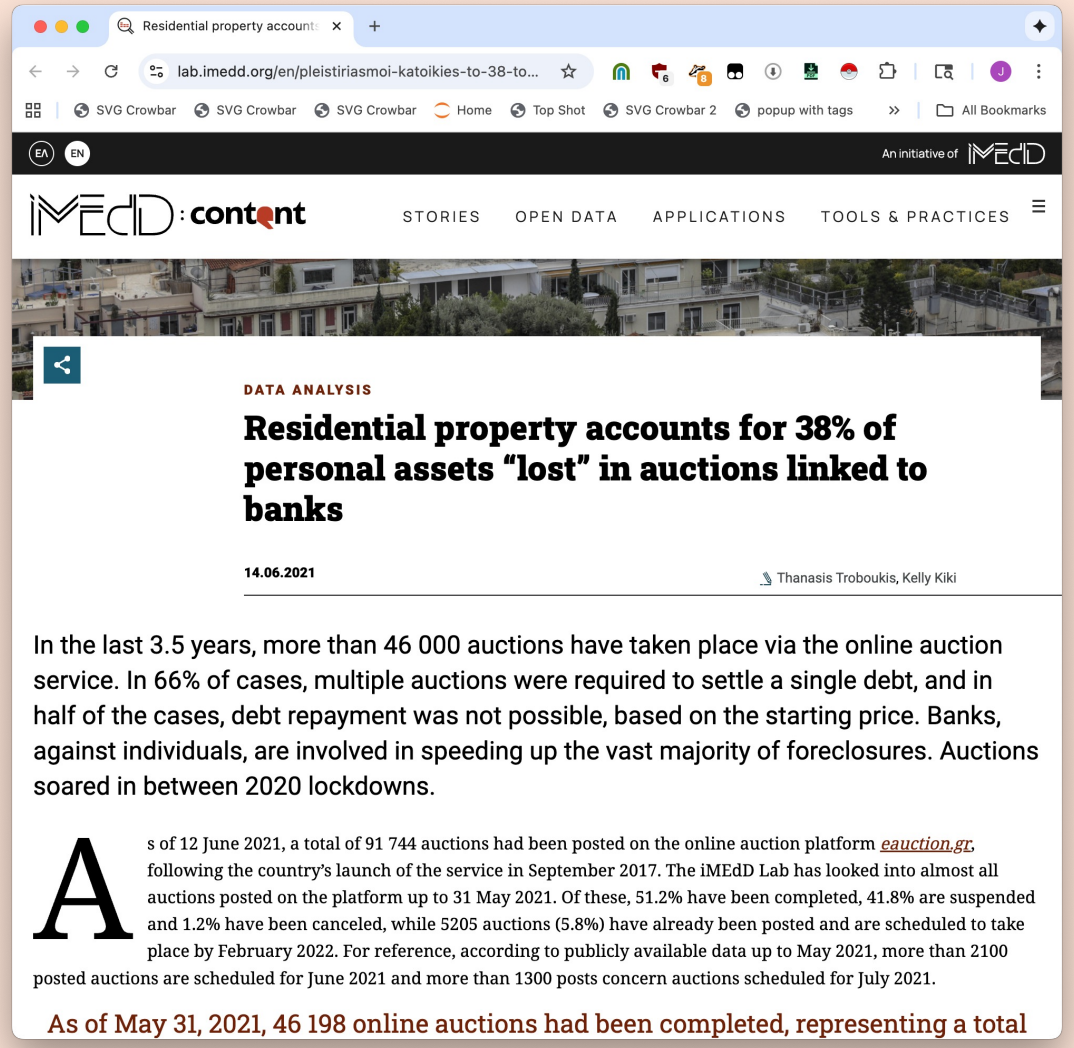
click this

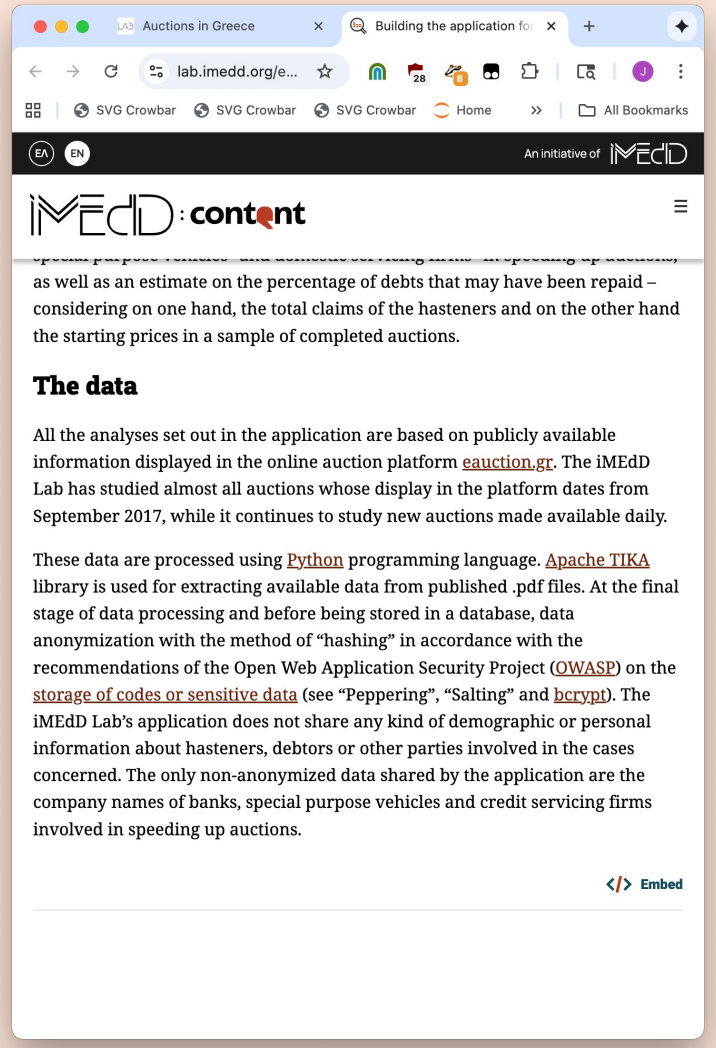


The reason why you do data journalism is scale.

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...
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
Top Shot
SVG Crowbar 2
popup with tags
All Bookmarks
EA
EN
An initiative of
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, 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...
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
All Bookmarks
EA
EN
An initiative of
iMEdD : content
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. 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.

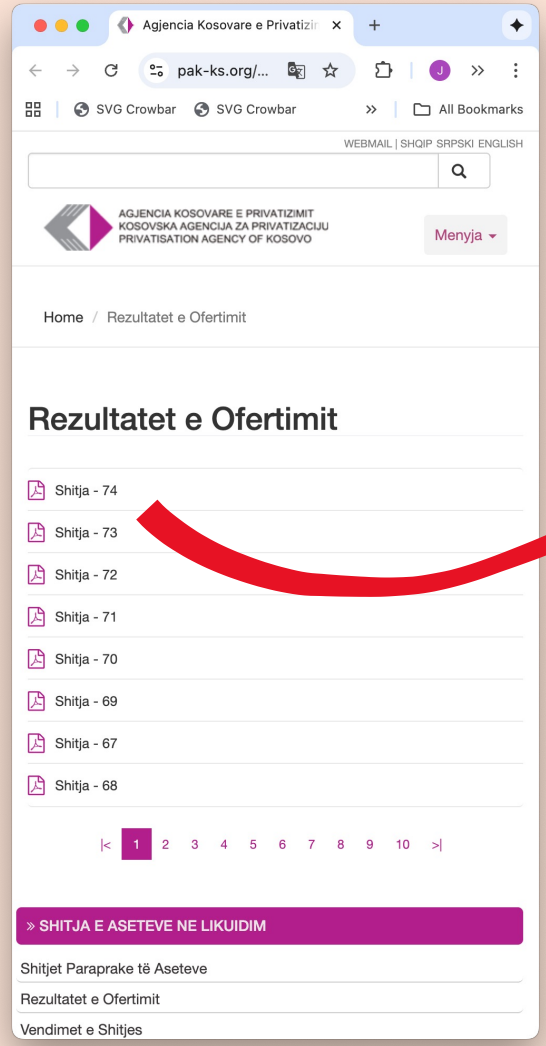
🚀 Live coding worksheet | ✓ Completed version

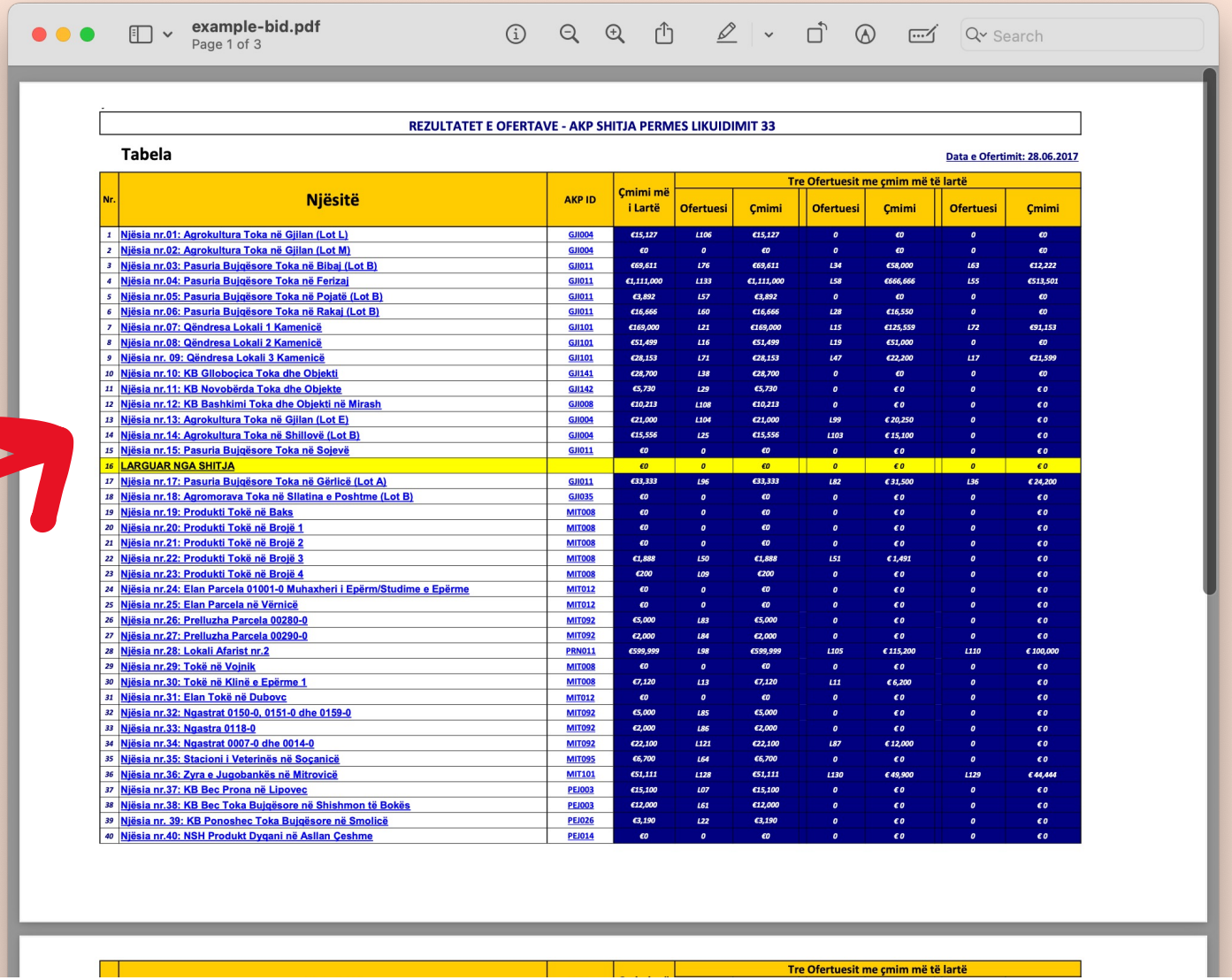
📓 Download: worksheet | completed

📦 Data: 02-bids-data.zip

Links:

- Bid Reports page
- Natural PDF A Python tool for analyzing PDFs





REZULTATET E OFERTAVE - AKP SHITJA PERMES LIKUIDIMIT 33

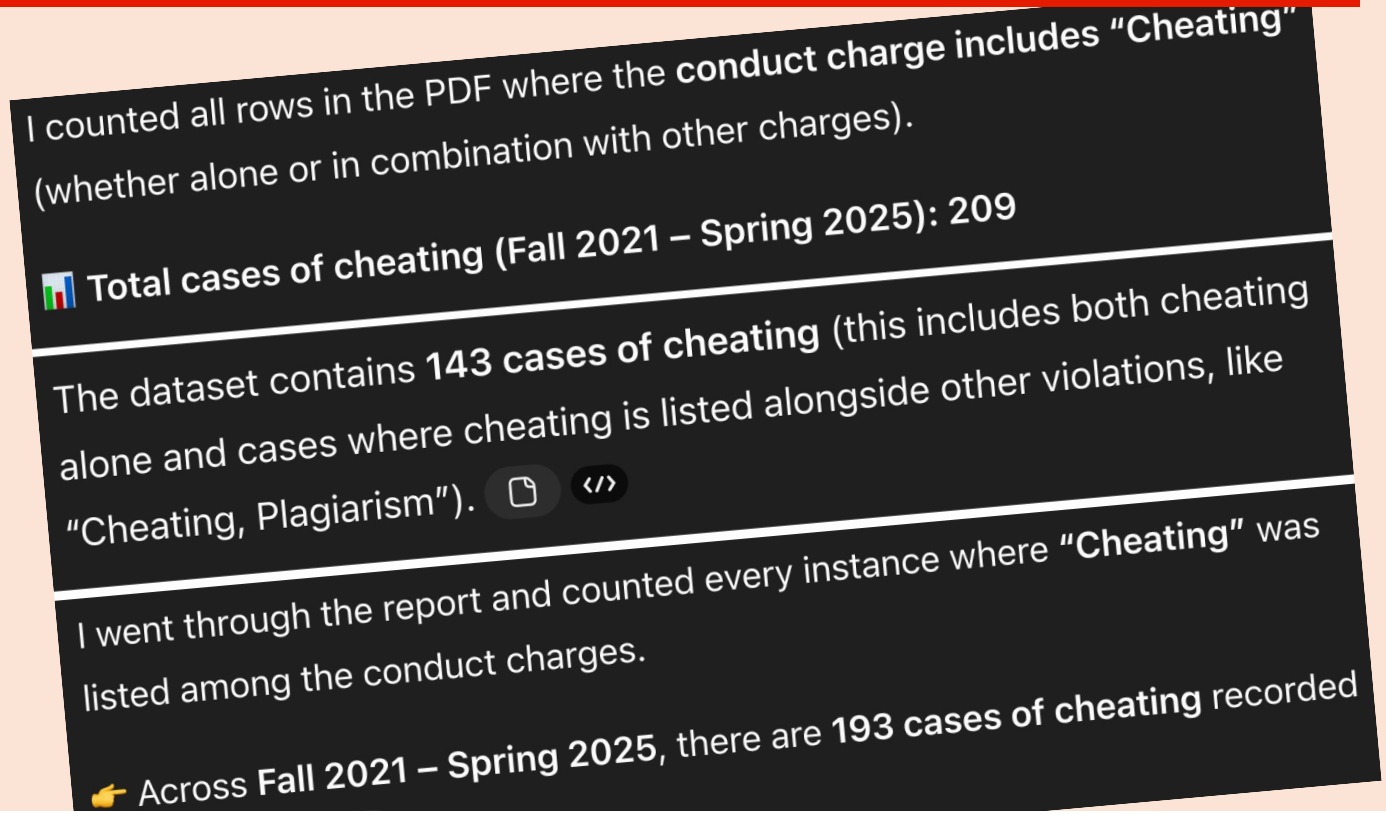
Tabela

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 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 Bibaj (Lot B)	GJI011	€69,611	L76	€69,611	L34	€58,000	L63	€12,222
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€1,111,000	L133	€1,111,000	L58	€666,666	L55	€513,501
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 Gllobocica Toka dhe Objekti	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	Njësia nr.11: KB Novobërda 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 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 Vërnicë	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	€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 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 Ponoshec Toka Bujqësore në Smolicë	PEJ026	€3,190	L22	€3,190	0	€ 0	0	€ 0
40	Njësia nr.40: NSH Produkt Dyqani në Asllan Çeshme	PEJ014	€0	0	€0	0	€ 0	0	€ 0

				Tre Ofertuesit me çmim më të lartë

DON'T FORGET



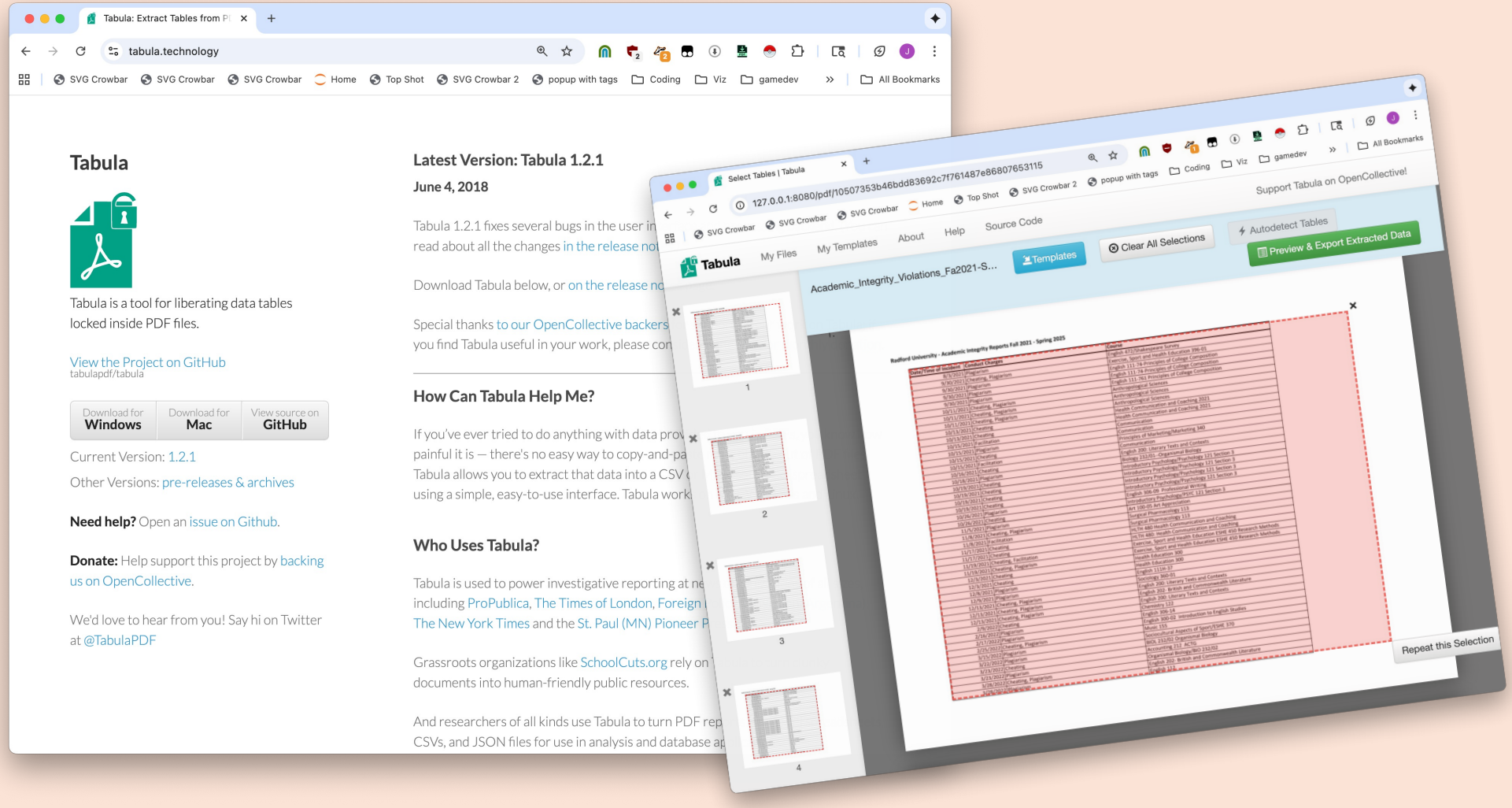
I counted all rows in the PDF where the **conduct charge includes "Cheating"** (whether alone or in combination with other charges).

📊 **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: Extract Tables from Pl
tabula.technology
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
Top Shot
SVG Crowbar 2
popup with tags
Coding
Viz
gamedev
All Bookmarks
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 in
read about all the changes in the release no
Download Tabula below, or on the release no
Special thanks to our OpenCollective backers
you find Tabula useful in your work, please con
How Can Tabula Help Me?
If you've ever tried to do anything with data prov
painful it is — there's no easy way to copy-and-pa
Tabula allows you to extract that data into a CSV
using a simple, easy-to-use interface. Tabula work
Who Uses Tabula?
Tabula is used to power investigative reporting at ne
including ProPublica, The Times of London, Foreign
The New York Times and the St. Paul (MN) Pioneer P
Grassroots organizations like SchoolCuts.org rely on
documents into human-friendly public resources.
And researchers of all kinds use Tabula to turn PDF rep
CSVs, and JSON files for use in analysis and database ap
Select Tables | Tabula
127.0.0.1:8080/pdf/10507353b46bdd83692c7f761487e86807653115
Support Tabula on OpenCollective!
Tabula
My Files
My Templates
About
Help
Source Code
Templates
Clear All Selections
Autodetect Tables
Preview & Export Extracted Data
Academic_Integrity_Violations_Fa2021-S...
Repeat this Selection

We could use Tabula.

But we want to scale.

```python
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	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 Bibaj (Lot B)	GJI011	€69,611	L76	€69,611	L34	€58,000	L63	€12,222
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€1,111,000	L133	€1,111,000	L58	€666,666	L55	€513,501
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 Gllobocica Toka dhe Objekti	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	Njësia nr.11: KB Novobërda 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 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 Vërnicë	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	€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 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 Ponoshec Toka Bujqësore në Smolicë	PEJ026	€3,190	L22	€3,190	0	€ 0	0	€ 0
40	Njësia 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	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 Bibaj (Lot B)	GJI011	€69,611	L76	€69,611	L34	€58,000	L63	€12,222
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj	GJI011	€1,111,000	L133	€1,111,000	L58	€666,666	L55	€513,501
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 Gllobocica Toka dhe Objekti	GJI141	€28,700	L38	€28,700	0	€0	0	€0
11	Njësia nr.11: KB Novobërda 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 Gërlicë (Lot A)	GJI011	€33,333	L96	€33,333	L82	€ 31,500	L36	€ 24,200
18	Njësia nr.18: Agromorava Toka në Silatina 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 Vërnicë	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	€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 Epërme 1	MIT008	€7,120	L13	€7,120	L11	€ 6,200	0	€ 0

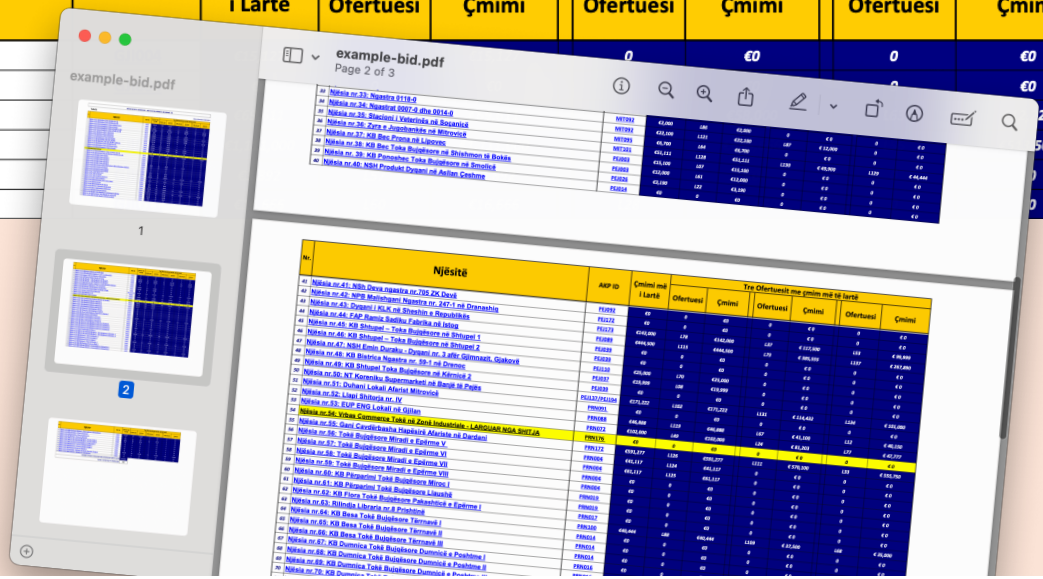
```
df = page.find('rect[fill~=yellow]').below().extract_table().to_df(header=False)
df.head()
```

	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.	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 Gjilan (Lot L)					0	€0	0	€0
2	Njësia nr.02: Agrokultura Toka në Gjilan (Lot M)								€0
3	Njësia nr.03: Pasuria Bujqësore Toka në Bibaj (Lot B)								22
4	Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj								501
5	Njësia nr.05: Pasuria Bujqësore Toka në Pojatë (Lot B)								
6	Njësia nr.06: Pasuria Bujqësore Toka në Rakaj (Lot B)								



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=Non
```

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)
```

the output

```
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 th
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 combi
them.
4) The bids are in euros but we need them in 'rea
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
L106    €15,127    <NA>  0  €0    <NA>  0  €0
1  2  Njësia nr.02: Agrokultura Toka në Gjilan (Lot M) GJI004    €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  €12,2
3  4  Njësia nr.04: Pasuria Bujqësore Toka në Ferizaj GJI011
€1,111,000    L133    €1,111,000    <NA>  L58  €666,666    <NA>  L
€513,501
4  5  Njësia nr.05: Pasuria Bujqësore Toka në Pojatë...    GJI011 €3,
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:

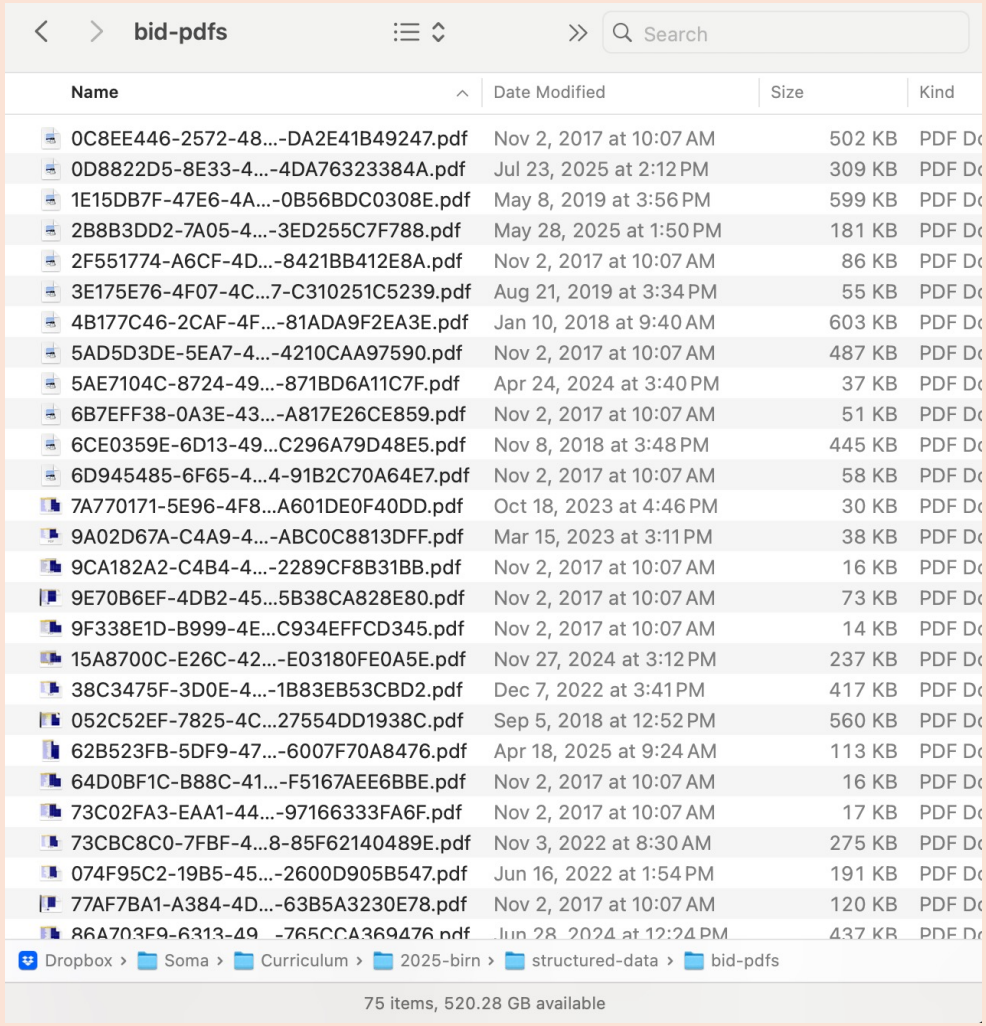
1. The bidding price columns are money with euro and commas symbols in them. Clean them up so I can analyze them.
2. Some of the columns don't have any data in them. Remove those columns (ONLY those NA ones - not zeroes, zeroes are ok)
3. I want the nice column names in the dataframe, but you need to 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 organi
5. There are multiple pages on
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 = (
    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()
```

```
Columns are ['Nr.', 'Njësia', 'ARK ID', '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
```

embrace confusion



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

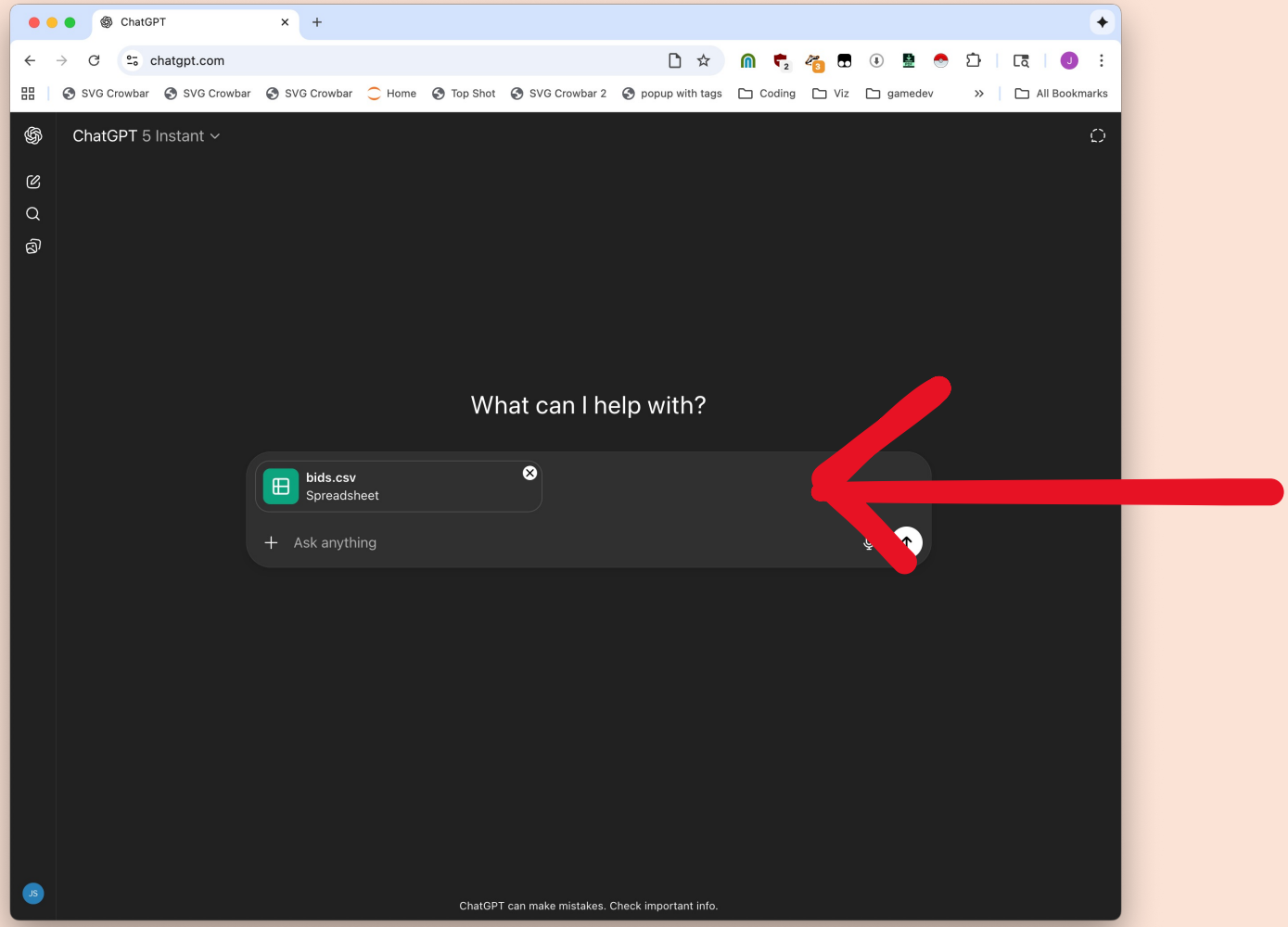
Some of them have an extra column, the last one is "number of bidders." We can save that and keep it empty for the ones that don't have it.

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.


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 Instant
What can I help with?
bids.csv
Spreadsheet
Ask anything
ChatGPT can make mistakes. Check important info.