Tabula: Getting Data from PDFs

...

I hate PDFs and so should you.

Data Jail

Download this file:

https://wmata.com/about/transit-police/upload/Blotter-June-17-2024.pdf

MetroTransit Police

Bus - Rail - Parking Lot - Metro Facilities - Others

Metro Transit Police Blotter

For : Monday, June 17, 2024

Date	Time	Disposition	Location	Rail Station	Call Type	Disposition Comments
06/17/2024	03:16:00	REPORT	1911 Southern Ave	SOAV	ROBBERY ARMED	ARMED ROBBERY
06/17/2024	07:05:00	ARREST	650 Maryland Ave SW	-	FOLLOW UP	WARRANT SERVICE / RECOVERED GUN
06/17/2024	07:09:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	THEFT	THEFT
06/17/2024	07:34:00	REPORT	730 Kenilworth Ave Ne(295 Side)	MINA	SUBJ WWEAPON	THREATS
06/17/2024	07:43:00	SUMMONS/CITATION	625 9th St NW	GALP	SUBJECT STOP	FTP
06/17/2024	08:06:00	REPORT	4704 Old Soper Rd	BRAV	TRAFFIC STOP	DRIVING ON SUSP LICENSE
06/17/2024	08:27:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:32:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:38:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:44:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:51:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:52:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:54:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:55:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	09:06:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	ASSIST PATRON	HARASSMENT
06/17/2024	09:08:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP

The goal: get tabular data out of a PDF

PDFs often have data that we can use to make comparisons. For example, the tables in this document.

If we had this as data, we could ask some more questions of it: what kinds of calls at which locations, for instance.

Date	Time	Disposition	Location	Rail Station	Call Type	Disposition Comments
06/17/2024	03:16:00	REPORT	1911 Southern Ave	SOAV	ROBBERY ARMED	ARMED ROBBERY
06/17/2024	07:05:00	ARREST	650 Maryland Ave SW	-	FOLLOW UP	WARRANT SERVICE / RECOVERED GUN
06/17/2024	07:09:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	THEFT	THEFT
06/17/2024	07:34:00	REPORT	730 Kenilworth Ave Ne(295 Side)	MINA	SUBJ WWEAPON	THREATS
06/17/2024	07:43:00	SUMMONS/CITATION	625 9th St NW	GALP	SUBJECT STOP	FTP
06/17/2024	08:06:00	REPORT	4704 Old Soper Rd	BRAV	TRAFFIC STOP	DRIVING ON SUSP LICENSE
06/17/2024	08:27:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:32:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:38:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:44:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:51:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:52:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:54:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:55:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	09:06:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	ASSIST PATRON	HARASSMENT
06/17/2024	09:08:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP

Sure looks like a spreadsheet to me!

Setup

Download the latest version for your operating system from https://tabula.technology/, then unzip it and put it in your Applications folder.

On a Mac, you may see this warning when launching Tabula for the first time: “Tabula is an app downloaded from the internet. Are you sure you want to open it?” Click Open.

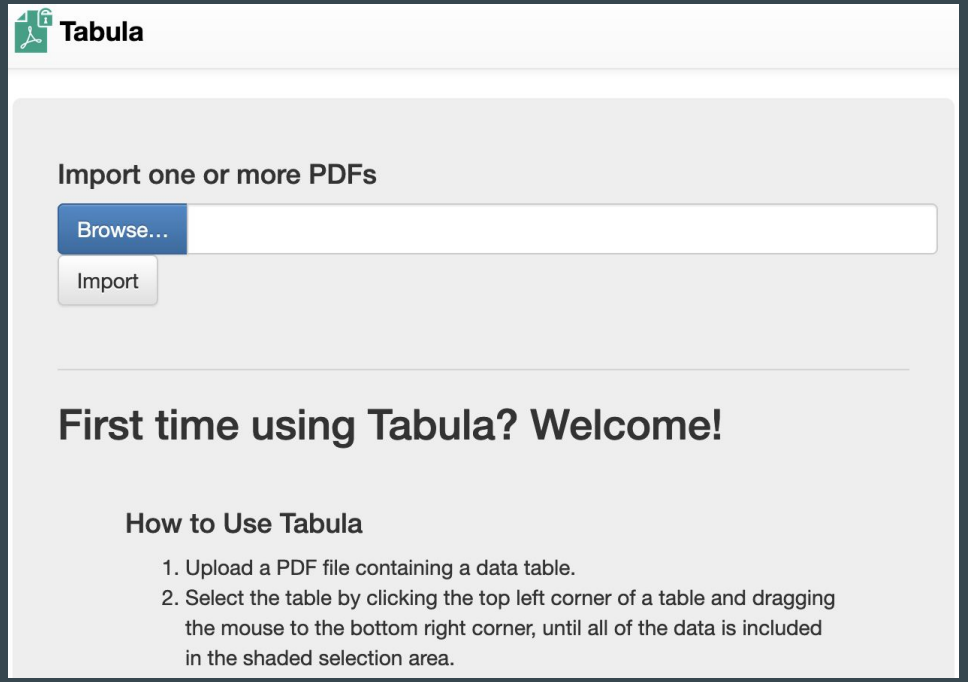


Running Tabula

When you launch Tabula, it will open a browser window at http://127.0.0.1:8080/. You might have to hit refresh, but you should see this.

The idea is that you upload a PDF to Tabula and then can extract data from tables in it.

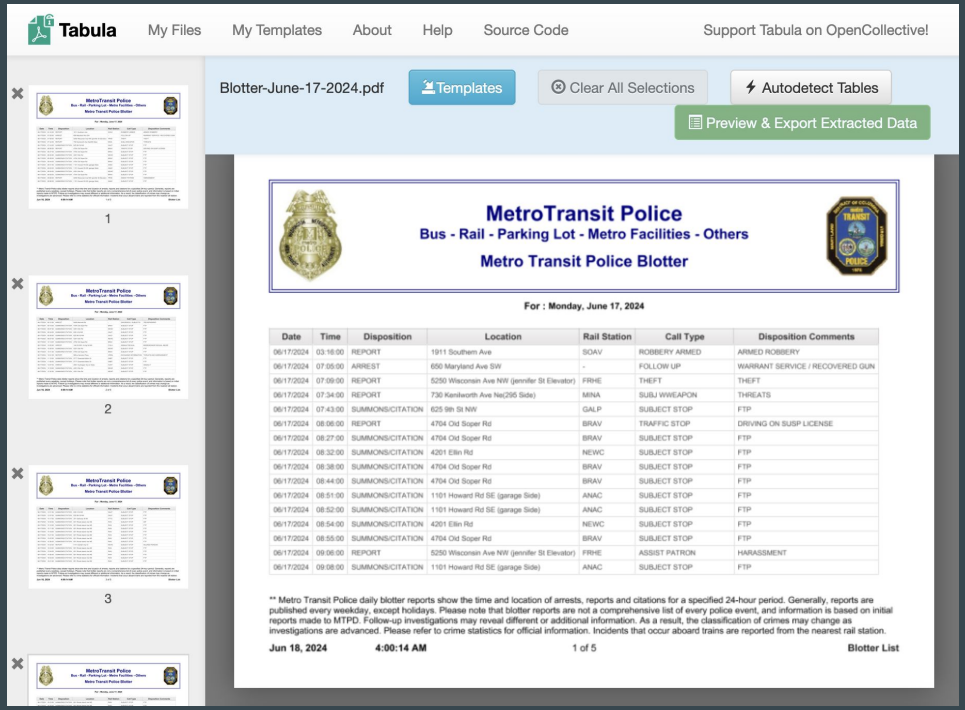
Let's upload our PDF:



Check the results

Tabula will take a minute or so to process a PDF, maybe longer depending on your computer.

You'll then get a browser window with pages from the PDF on the left and each page larger on the right.



Find our table and draw a box around it

Go to page 1, where our table begins, and then using your cursor click and draw a box around the table. Err on the side of more room, not getting it as tight to the border as possible.

If it looks good, click the green “Preview & Export” button.



MetroTransit Police

Bus - Rail - Parking Lot - Metro Facilities - Others

Metro Transit Police Blotter

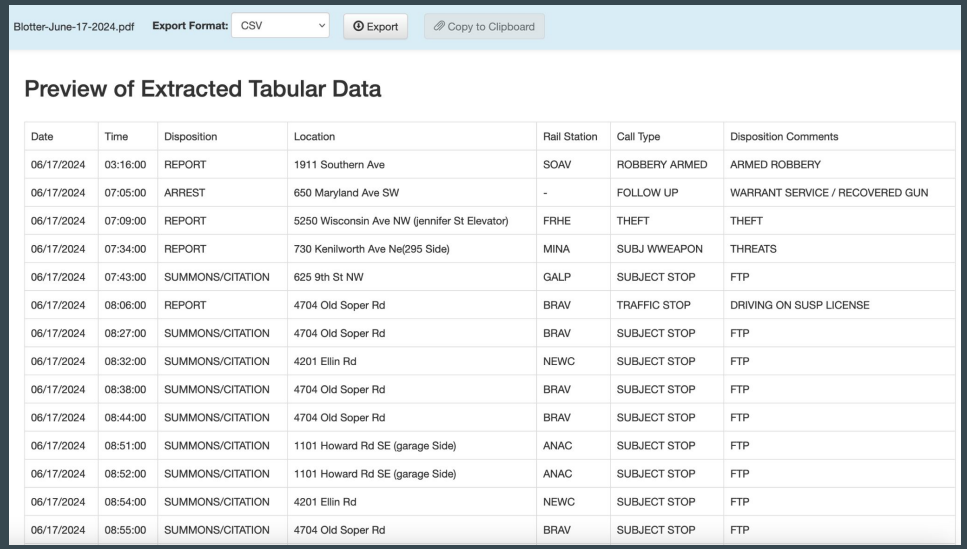
For : Monday, June 17, 2024

Date	Time	Disposition	Location	Rail Station	Call Type	Disposition Comments
06/17/2024	03:16:00	REPORT	1911 Southern Ave	SOAV	ROBBERY ARMED	ARMED ROBBERY
06/17/2024	07:05:00	ARREST	650 Maryland Ave SW	-	FOLLOW UP	WARRANT SERVICE / RECOVERED GUN
06/17/2024	07:09:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	THEFT	THEFT
06/17/2024	07:34:00	REPORT	730 Kenilworth Ave Ne(295 Side)	MINA	SUBJ WWEAPON	THREATS
06/17/2024	07:43:00	SUMMONS/CITATION	625 9th St NW	GALP	SUBJECT STOP	FTP
06/17/2024	08:06:00	REPORT	4704 Old Soper Rd	BRAV	TRAFFIC STOP	DRIVING ON SUSP LICENSE
06/17/2024	08:27:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:32:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:38:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:44:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:51:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:52:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:54:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:55:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	09:06:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	ASSIST PATRON	HARASSMENT
06/17/2024	09:08:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP

Check the results

Depending on how you drew the box, your results might look different from this, but they should be pretty similar.

Tabula has two methods of extraction; you should try them both! When satisfied, hit Export button to get a CSV file.

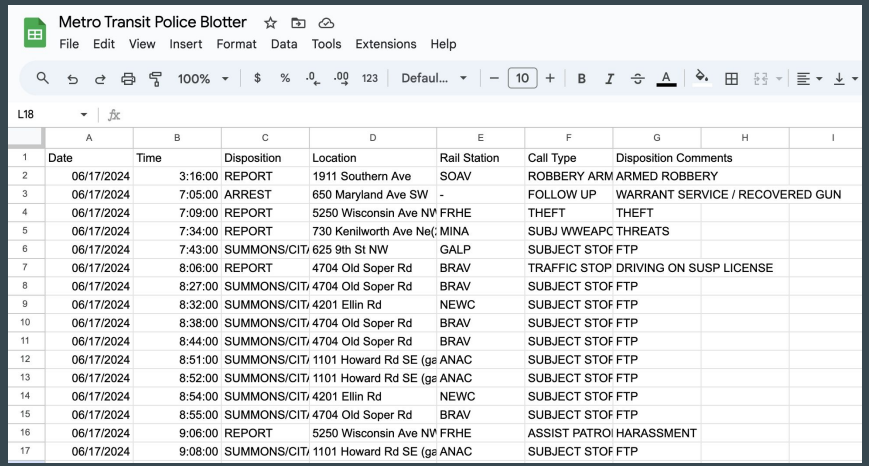


Date	Time	Disposition	Location	Rail Station	Call Type	Disposition Comments
06/17/2024	03:16:00	REPORT	1911 Southern Ave	SOAV	ROBBERY ARMED	ARMED ROBBERY
06/17/2024	07:05:00	ARREST	650 Maryland Ave SW	-	FOLLOW UP	WARRANT SERVICE / RECOVERED GUN
06/17/2024	07:09:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	THEFT	THEFT
06/17/2024	07:34:00	REPORT	730 Kenilworth Ave Ne(295 Side)	MINA	SUBJ WWEAPON	THREATS
06/17/2024	07:43:00	SUMMONS/CITATION	625 9th St NW	GALP	SUBJECT STOP	FTP
06/17/2024	08:06:00	REPORT	4704 Old Soper Rd	BRAV	TRAFFIC STOP	DRIVING ON SUSP LICENSE
06/17/2024	08:27:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:32:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:38:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:44:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:51:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:52:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:54:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:55:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP

Exercise

Upload the CSV file into Google Sheets (you can open a new browser window and type **sheet.new** to start. You might need to clean things up a little bit once you upload the CSV file into Sheets - some PDFs are worse than others.

Here's what mine looks like:



Metro Transit Police Blotter

File Edit View Insert Format Data Tools Extensions Help

L18

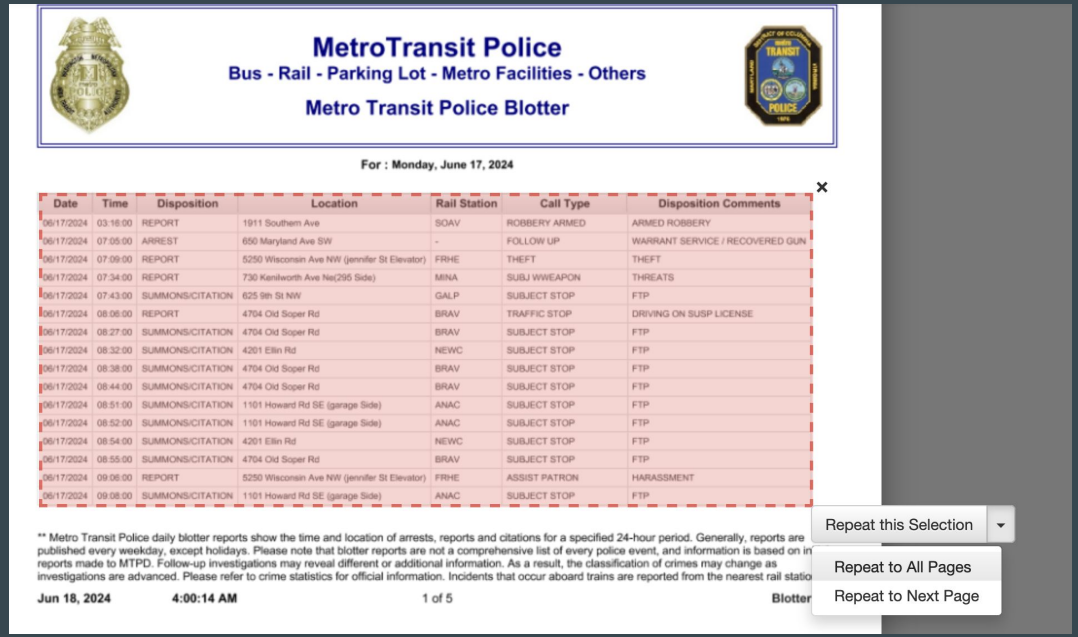
	A	B	C	D	E	F	G	H	I
1	Date	Time	Disposition	Location	Rail Station	Call Type	Disposition Comments		
2	06/17/2024	3:16:00	REPORT	1911 Southern Ave	SOAV	ROBBERY ARM	ARMED ROBBERY		
3	06/17/2024	7:05:00	ARREST	650 Maryland Ave SW	-	FOLLOW UP	WARRANT SERVICE / RECOVERED GUN		
4	06/17/2024	7:09:00	REPORT	5250 Wisconsin Ave NW	FRHE	THEFT	THEFT		
5	06/17/2024	7:34:00	REPORT	730 Kenilworth Ave Ne(	MINA	SUBJ WWEAPC	THREATS		
6	06/17/2024	7:43:00	SUMMONS/CIT	625 9th St NW	GALP	SUBJECT STOF	FTP		
7	06/17/2024	8:06:00	REPORT	4704 Old Soper Rd	BRAV	TRAFFIC STOP	DRIVING ON SUSP LICENSE		
8	06/17/2024	8:27:00	SUMMONS/CIT	4704 Old Soper Rd	BRAV	SUBJECT STOF	FTP		
9	06/17/2024	8:32:00	SUMMONS/CIT	4201 Ellin Rd	NEWC	SUBJECT STOF	FTP		
10	06/17/2024	8:38:00	SUMMONS/CIT	4704 Old Soper Rd	BRAV	SUBJECT STOF	FTP		
11	06/17/2024	8:44:00	SUMMONS/CIT	4704 Old Soper Rd	BRAV	SUBJECT STOF	FTP		
12	06/17/2024	8:51:00	SUMMONS/CIT	1101 Howard Rd SE (ga	ANAC	SUBJECT STOF	FTP		
13	06/17/2024	8:52:00	SUMMONS/CIT	1101 Howard Rd SE (ga	ANAC	SUBJECT STOF	FTP		
14	06/17/2024	8:54:00	SUMMONS/CIT	4201 Ellin Rd	NEWC	SUBJECT STOF	FTP		
15	06/17/2024	8:55:00	SUMMONS/CIT	4704 Old Soper Rd	BRAV	SUBJECT STOF	FTP		
16	06/17/2024	9:06:00	REPORT	5250 Wisconsin Ave NW	FRHE	ASSIST PATRO	HARASSMENT		
17	06/17/2024	9:08:00	SUMMONS/CIT	1101 Howard Rd SE (ga	ANAC	SUBJECT STOF	FTP		

Moar Data!

That first page isn't the only data in the PDF - we can use Tabula to export all of the tables.

Go back and choose "Repeat this Selection", then "Repeat to All Pages".

Export and import again.



MetroTransit Police

Bus - Rail - Parking Lot - Metro Facilities - Others

Metro Transit Police Blotter

For : Monday, June 17, 2024

Date	Time	Disposition	Location	Rail Station	Call Type	Disposition Comments
06/17/2024	03:16:00	REPORT	1911 Southern Ave	SOAV	ROBBERY ARMED	ARMED ROBBERY
06/17/2024	07:05:00	ARREST	650 Maryland Ave SW	-	FOLLOW UP	WARRANT SERVICE / RECOVERED GUN
06/17/2024	07:09:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	THEFT	THEFT
06/17/2024	07:34:00	REPORT	730 Kenilworth Ave Ne(295 Side)	MINA	SUBJ WWEAPON	THREATS
06/17/2024	07:43:00	SUMMONS/CITATION	625 9th St NW	GALP	SUBJECT STOP	FTP
06/17/2024	08:06:00	REPORT	4704 Old Soper Rd	BRAV	TRAFFIC STOP	DRIVING ON SUSP LICENSE
06/17/2024	08:27:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:32:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:38:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:44:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	08:51:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:52:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP
06/17/2024	08:54:00	SUMMONS/CITATION	4201 Ellin Rd	NEWC	SUBJECT STOP	FTP
06/17/2024	08:55:00	SUMMONS/CITATION	4704 Old Soper Rd	BRAV	SUBJECT STOP	FTP
06/17/2024	09:06:00	REPORT	5250 Wisconsin Ave NW (jennifer St Elevator)	FRHE	ASSIST PATRON	HARASSMENT
06/17/2024	09:08:00	SUMMONS/CITATION	1101 Howard Rd SE (garage Side)	ANAC	SUBJECT STOP	FTP

** Metro Transit Police daily blotter reports show the time and location of arrests, reports and citations for a specified 24-hour period. Generally, reports are published every weekday, except holidays. Please note that blotter reports are not a comprehensive list of every police event, and information is based on in reports made to MTPD. Follow-up investigations may reveal different or additional information. As a result, the classification of crimes may change as investigations are advanced. Please refer to crime statistics for official information. Incidents that occur aboard trains are reported from the nearest rail statio

Jun 18, 2024 4:00:14 AM 1 of 5 Blotter

Repeat this Selection

Repeat to All Pages

Repeat to Next Page

Now we can ask questions

After removing duplicate header rows, we can make a pivot table to count the call types, for instance. Or the nearest rail station. Or the disposition.

You get the idea. Every dataset can spark multiple questions.

Call Type	COUNTA of Loc
	0
ASSAULT	1
ASSAULTSEXUAL	1
ASSIST PATRON	1
DISORDERLY SUBJECTS	1
EXCHANGE INFORMATION	1
FOLLOW UP	1
PROPERTYDESTRUCTION	2
ROBBERY ARMED	1
SUBJ WWEAPON	1
SUBJECT STOP	51
THEFT	3
TRAFFIC STOP	1
Grand Total	**65**

What are the stories here?

Also, what if you collected this every day? What if by “you” I meant “a computer”?

What if you automatically got a Slack message summarizing each day’s report and suggesting possible ideas based on the data?

All of that is doable.

Your turn

Pick another PDF and try extracting it into a CSV file like we just did.

Try to add a calculation or pivot table to make it more useful, and write up a few sentences describing the best idea or story from it.