

SUMMARY. DATA MODULE ONE WORKSHOP

Jomec, June 10

Key points before today (in the three hours of preparatory work)

- CSV files, organising Excel, sourcing work, filtering, sorting
- Exporting csvs, Basic charts
- Simple calculations and formulas in Excel
- Pivot tables (very slightly)
- Reading: getting the data *you* need / history / methodology

Today's main areas

- Getting stories
- Getting data for journalism
- ~~Picturing & Numbers~~ (we didn't get to this! It will feature in the June 14+ material)
- Our group, resources and continuing module one

Random questions that came up

- Is there a strict deadline for working through the material?

No. The goal is just to get through everything by the end of the month.

- Why didn't the "COUNT IF" formula in Walkthrough 2 work?

=COUNTIF(B2:B195, "*university*")

Because someone (→ Aidan) originally wrote the example with single commas ('*university*') instead of double ("*university*"). This frequently causes problems in Sheets and Excel and " " is usually the best option.

- Why is the data in the [Obama commutation CSV](#) weird?

Either because of how it was collected or because it was designed for another format (not .CSV) the data is laid out strangely. It's basically a person then a separate row for each piece of data, then another person with rows for their data etc.

Name Kendrick
Offence Cocaine
Date 2012

Name Dale
Offence Methamphetamine
Date 2011

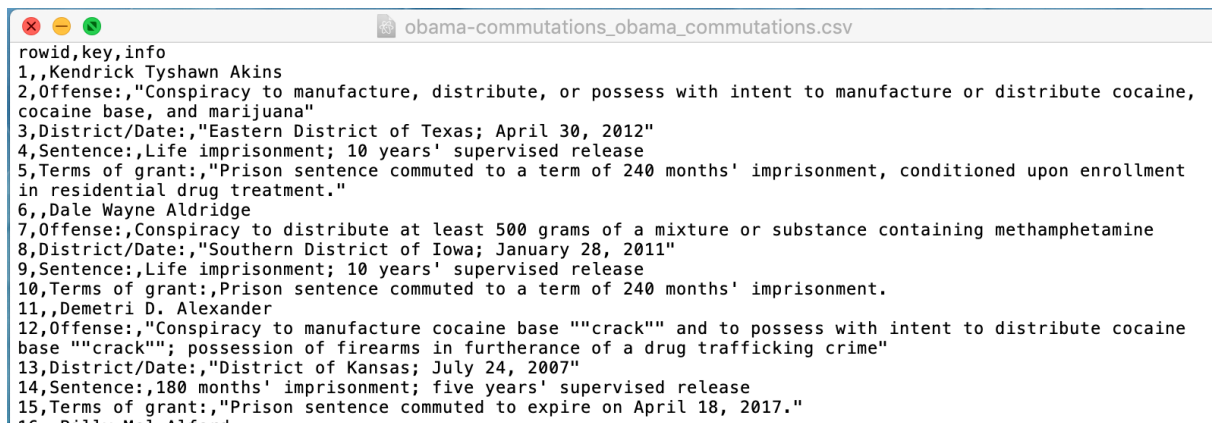
etc

What would be useful would be a column type layout in the CSV:

Name, Offence, Date
Kendrick, Cocaine, 2012
Dale, Methamphetamine, 2011
etc

So it's not possible to have it laid out in a spreadsheet in the latter format. We won't fix it (now), but we'll note the following points:

— If something's odd, opening the file locally to look at the csv (our very first task in Walkthrough 1) lets us check if it's Excel, the file, or a mad computer that's generating the weird layout. When we do that we see that the data is indeed laid out a bit oddly and it's not us or Excel:



```
rowid,key,info
1,,Kendrick Tyshawn Akins
2,Offense:,"Conspiracy to manufacture, distribute, or possess with intent to manufacture or distribute cocaine, cocaine base, and marijuana"
3,District/Date:,"Eastern District of Texas; April 30, 2012"
4,Sentence:,Life imprisonment; 10 years' supervised release
5,Terms of grant:,"Prison sentence commuted to a term of 240 months' imprisonment, conditioned upon enrollment in residential drug treatment."
6,,Dale Wayne Aldridge
7,Offense:,"Conspiracy to distribute at least 500 grams of a mixture or substance containing methamphetamine
8,District/Date:,"Southern District of Iowa; January 28, 2011"
9,Sentence:,Life imprisonment; 10 years' supervised release
10,Terms of grant:,"Prison sentence commuted to a term of 240 months' imprisonment."
11,,Demetri D. Alexander
12,Offense:,"Conspiracy to manufacture cocaine base ""crack"" and to possess with intent to distribute cocaine base ""crack""; possession of firearms in furtherance of a drug trafficking crime"
13,District/Date:,"District of Kansas; July 24, 2007"
14,Sentence:,180 months' imprisonment; five years' supervised release
15,Terms of grant:,"Prison sentence commuted to expire on April 18, 2017."
16,,Billie Mae Alexander
```

— It's possible to transpose an entire (normal) dataset in Excel (which means to flip the rows and columns as in the example below): Copy it all, then PASTE / TRANSPOSE. It works quite well with the flytipping data as an example (below). To transpose every five rows as in the Obama data will also be possible but we'll skip that problem for now (since it will be much more complicated).

	A	B	C	D	E	F	G
1		Animal carcass	Green	Vehicle parts	White goods	Other electrical	Tyres
2	2006-07	281	2,462	1,046	2,564	1,057	1,360
3	2007-08	333	2,901	815	2,868	1,496	1,641
4	2008-09	302	2,939	588	2,216	1,561	1,621
5	2009-10	213	2,374	470	1,809	1,740	1,360
6	2010-11	100	1,943	513	1,622	1,336	1,141
7	2011-12	94	1,739	341	1,161	830	1,151
8	2012-13	110	1,610	236	726	676	811
9	2013-14	72	1,197	247	1,574	708	591
10	2014-15	70	1,278	184	1,827	889	551
11	2015-16	56	1,219	269	1,782	980	481
12	2016-17	52	1,080	367	2,168	1,072	651
13	2017-18	75	1,158	338	2,082	924	641
14	2018-19	86	1,011	274	2,040	786	541
15	2019-20	95	1,305	254	1,777	643	461

- Is it possible to sort (A → Z etc) in a *single* spreadsheet column while leaving everything else in place and thereby cause a disaster?

Technically yes but it's unlikely. As long as you have selected a cell somewhere inside the data, Excel will assume you want to work with *all* the data. To do a single column sort you'd have to select all of one column, only one column, and you'd get a warning:

The screenshot shows an Excel spreadsheet with a 'Sort Warning' dialog box open. The dialog box contains the following text: 'Data outside your current selection won't be sorted. What do you want to do?'. There are two radio button options: 'Expand the selection' (which is selected) and 'Continue with the current selection'. At the bottom of the dialog are 'Cancel' and 'Sort' buttons. The spreadsheet data in the background is as follows:

	A	B	C	D	E	F	G
1		2006-07	2007-08	2008-09	2009-10	2010-11	2011-12
2	Clinical	122	97				
3	Animal carcass	281	333				
4	Chemical drugs	205	268				
5	Asbestos	236	168				
6	Vehicle parts	1,046	815				
7	Tyres	1,360	1,641				
8	Other electrical	1,057	1,496				
9	Black bags - clothing	2,332	941				
10	White goods	2,564	2,868				
11	Other community	2,105	2,238				
12	Other waste	1,921	2,229	2,446	1,896	1,764	1,209
13	Green	2,462	2,901	2,939	2,374	1,943	1,739
14	Construction	3,394	3,885	3,120	2,652	2,861	2,978

Note You can always double check with spot checks - pick a row and get a reading in a column before you sort, then check they're still aligned after the Sort.

- What useful Excel moves did Aidan notice we hadn't covered?

The resize zoom on the bottom right makes everything more readable

You can freeze a first column just like a first row (Tab View / Freeze first column)

You can hide one or more columns (Select the column then Right click / Hide)

Case study. Oscars 2018

Source <https://www.imdb.com/list/ls021363441/>

BBC story <https://www.bbc.co.uk/news/entertainment-arts-43146027>

Key points

- Ideas

Case study. Fly tipping south Wales

Source

https://docs.google.com/spreadsheets/d/1AAI1ZIHhRPXo4pnugCDYQImHP5J4Pj9zYI_wO8RJ6SA/edit?usp=sharing

Key points

- Google Sheets (lighter alternative to Excel. Handles only about 40,000 rows)
You can make a *copy* if you have a Google account open (otherwise where would it be copied to?)
File types available for *downloading* the Sheets file: XLSX / ODS / CSV / TSV
- Data types here are: text, number
- Personal experience → dataset (we can start with a question, *then* the data)
- Who are the people we need to chase? 1. Adds to the story 2. Tests our analysis
- What is the data we might add to this? (prosecutions, Council refuse collection etc.)
- Methodology of data collection (We know x number of “incidents” were recorded - what *don't* we know when we look at the source that published the data?)

Story angles: Bastien, Bradshaw & NY Times

Bradshaw's angles

<https://onlinejournalismblog.com/2020/08/11/here-are-the-7-types-of-stories-most-often-found-in-data/>

<https://onlinejournalismblog.com/2020/08/12/3-more-angles-most-often-used-to-tell-data-stories-explorers-relationships-and-bad-data-stories/>

Bastien ([We do data](#))

Every variable can be tried as a (potential) angle

New York Times

<https://drive.google.com/drive/folders/1FOLQKiQdVX2Wr5Z2YXw5bel6S9ECATg0>

Key points

- Bradshaw 1 Scale / Change / Ranking / Geographical variation
- Bradshaw 2 User lookup & interactivity / Networks / Missing or bad
- Bastien Every variable as an angle?
- The 7 NY Times angles
- Can we have user lookup on a dataset with no context?

Dataset for story - LFB

<https://docs.google.com/spreadsheets/d/1tM-0ediE8oeZEgJXhW1TmlY-wW1otmwUfX9oxomc3tc/edit?usp=sharing>

Key points

- We are interviewing the dataset
- More data types here (including datetime and geographical)
- More “variables” (≈ columns) , more “observations” (≈ rows)
- Sorting *and* Filtering are what we do when shopping for shoes online

Getting data. 17 case studies

https://drive.google.com/file/d/1LaFsk_NIMoS0jtJdY6o2RgG8oN3Ut2lj/view?usp=sharing

Key points

- *Wide* range of topics. We (mostly) didn't know these *sources* existed this morning
- Questions we ask. What is the: data / form and shape / publisher / quality?
- What data is *not* available?

Getting data for journalism

<https://drive.google.com/file/d/1S0zIJpic-NC9f6FO3ramfO9CryaVSNYD/view?usp=sharing>

Key points

- BBC, BuzzFeed, Propublica, 538, Pudding. They publish the data, and often methods
- They are finding their *own* data
- Sources of data then are broadly:
 - “Published”:
 - Government or the ONS (for example)
 - Commercial (PA or Bloomberg for example)
 - FOI (or at least made available, eventually)
 - Collected:
 - Data collection (sending someone out with ‘a clipboard’)
 - Crowdsourcing
 - Scraping
 - Sensors
 - Missing: What’s not there? (or who’s not there?)

Final key points

- Email newsletters on the website ([Resources / Data Journalism Community](#)) are worth reviewing to see if you find them helpful. It can be a great way to see what journalists are doing and how (or maybe you get enough email already)
- Similarly, the Monitoring and Collections pages [here](#) might be interesting
- Problems? Crises? Misbehaving data? odonnella4@cardiff.ac.uk

Next

- More material to work on from Monday 14th (<https://aodhanlutetiae.github.io/dj/>)
- Drop-in zoom sessions this month if you have a question (feel free to ignore otherwise). Zoom links are in my email of June 3rd

BBC Data Module 1 drop-in 17/6

Time: Jun 17, 2021 02:00 PM London

BBC Data Module 1 drop-in 24/6

Time: Jun 24, 2021 02:00 PM London