

CAJ #24 Data Journalism 101 “Cheat sheet”

Useful links, formulas, tips, sources and additional educational materials that you can use long after the session is done

Useful formulas and functions are highlighted in **yellow**

Data sources

Federal

- [StatCan](#)
- [Canada Open Data](#)
- [Lobbying registry](#)
- [Procurement](#)
- [Registered charity tax returns](#)
- And many more!

Provincial

- [Open Data Ontario](#)
- [Open Data Quebec](#)
- [Open Data Alberta](#)
- [Open Data Manitoba](#)
- Saskatchewan
 - [GeoData](#)
 - [Sask. Bureau of Statistics](#)
- [Open Data B.C.](#)
- [Open Data N.S.](#)
- [Open Data N.B.](#)
- [Open Data PEI](#)
- [Open Data NL](#)
- [Open Data Yukon](#)
- [Open Data NWT](#)
- [Nunavut Statistics](#)
- You can also look at department-specific data or databases

Municipal datasets

- Too many to list all of them!
- Look up city + open data and see what you find!

Private sources


- [Google](#)

- [Statista](#) (note: this is paid service)

Setting up your workspace

- **Sheet.new** - type into your browser in chrome and it will automatically open a new Google Sheet
- Importing data into sheets. You have two main options:
 - Download as a csv from the site, import into sheets**
 - Copy paste (if possible, can cause formatting issues)**
 - Import Formulas**
 - **ImportHTML** (imports a table from a HTML page)
 - <https://support.google.com/docs/answer/3093339?hl=en>
 - Here's a step by step video for how to use the formula: <https://www.youtube.com/watch?v=3wV9JtRRLdQ>
 - Example:
 - Say we want a list of all of Canada's foreign affairs ministers pulled from this webpage: [https://en.wikipedia.org/wiki/Minister_of_Foreign_Affairs_\(Canada\)](https://en.wikipedia.org/wiki/Minister_of_Foreign_Affairs_(Canada))
 - `=IMPORTHTML("https://en.wikipedia.org/wiki/Minister_of_Foreign_Affairs_(Canada)", "table", 3)`
 - With this formula, you must include the URL, type and index
 - **URL:** Link you want to pull from
 - **Type:** either table or list
 - **Index:** the number of the table on the page. In the above example, the list of ministers is the third table on the page, so I wrote 3.
 - Make sure to include quotation marks around both words!!
 - Paste the formula in A1 of your spreadsheet and it will populate for you
 - This table includes unnecessary numbers and portraits. So if you want to select only the name and the term, you can choose specific columns using the query/select function.
 - `=query(IMPORTHTML("https://en.wikipedia.org/wiki/Minister_of_Foreign_Affairs_(Canada)", "table", 3), "Select Col3, Col4, Col5")`
 - **Query** indicates that you want to get some but not all of the info from the table
 - **Select:** operator to choose which columns you want to import into your sheet
 - **ImportData** (imports a csv or tsv dataset from a url)
 - <https://support.google.com/docs/answer/3093335?sjid=592386596448455588-NC>
 - This is great, but it is not common for urls to contain just csv data

Demo 1 - Fires in Toronto

Practice Data:  Clean Fire Incidents Data (EDITED) **


****NOTE:** This data has been changed for instructional purposes. Any reporting on this data should be based on the original dataset, which can found here:

open.toronto.ca/dataset/fire-incidents/

If you intend to follow along, or practice using this data, please make your own copy of the dataset and work from there

- **ACTIVITY:** Make a pivot table (or pivot tables) to answer:
 - Which wards saw the most civilian casualties over the five years of data?
 - In which year did the most civilians die or get injured in fires?
 - What are other questions you can ask of this data?

Having trouble with the above? Here's a step-by-step guide walking you through using google sheets to complete this analysis:

 Data demo walkthrough CAJ24

Math formulas every journalist should know

Rate change over time

From old value to new

Formula:

$$= (\text{New value} - \text{Old value}) / \text{Old value} * 100$$

E.g. crime is up 20%

Per capita analysis

Comparing events amongst populations

Formula:

$$= (\text{Event} / \text{Population}) \times \text{Per unit}$$

PER UNIT is usually 100,000

E.g. cancer rates are 546 per 100,000 for women and 670 per 100,000 for men in Canada

Comparison rates

Dividing rates by one another

Formula:

= Rate 1 / Rate 2

E.g. comparing per capita incarceration rates by race

Google Sheets/Excel tips

Formulas

- adding two variables =A1+A2
- summing several variables = SUM(A2:A12)
- percentages = A1/A2 (and then format as %)
- summing columns = SUM(A2:N2)
- rate formula= A1/A2 * per unit
- average = AVERAGE(A2:N2)
- median = MEDIAN(A2:N2)

The Google Sheets help function is your friend!

Save time with Google Sheets keyboard shortcuts:

- You don't have to drag to select thousands of rows! Highlight all values in a column with:
 - ⌘ + Shift + down arrow key
- Find and replace values
 - ⌘ + Shift + h
- Fill range
 - ⌘ + Enter
- Fill down
 - ⌘ + d
- Fill right
 - ⌘ + r
- Paste values without any special formatting
 - ⌘ + Shift + v
- Looking for others? Here's a complete list:
 - <https://support.google.com/docs/answer/181110?hl=en&co=GENIE.Platform%3DDesktop>

Don't make these mistakes..

- CLEAN YOUR DATA! 95 percent of data analysis is data cleaning, we promise you. Look over your data, make sure it is not missing values and has been standardized
 - CAJ this year is having a data cleaning panel for anyone who is interested
 - [OpenRefine](#) is considered the industry standard data cleaning tool
- Make sure what you think data represents actually represents it.
 - Many open data files contain data dictionaries to help you better understand, read them.
- Always make sure to offer proper context to your data. Data is not everything, it's the beginning point of the story, never the end.
 - Speak with experts for statistical and historical input.
 - Do per capita analysis to evaluate how these trends are affecting different populations. But also don't rely solely on per capita analysis, statistical outliers are a possibility
- Keep it simple, don't use too many numbers in a story.
 - Our brains process numbers differently than they do letters, and switching between the two a lot discourages readers from continuing.
 - Focus on a couple key ones and back them up with a responsible narrative that contextualizes them with expert opinion

Learn how to code!

Coding can be a way to elevate what you're able to do. It lets you manage large datasets that don't fit in excel/sheets and scrape data from webpages

But coding is hard. Where do I begin?

- There's lots of free online courses for journalists to learn how to code:
 - <https://datajournalism.com/watch/python-for-journalists>
 - <https://coding-for-journalists.readthedocs.io/en/latest/>
- The first thing most agree upon though is you'll need to do is learn how to create a coding "environment"
 - [Google Collab](#) is free, but it requires you to be connected to the internet
 - A great one we recommend is [Anaconda](#), which is used by a lot of data scientists
- Web scraping is a great way to learn how to code because it's project-oriented and relatively straightforward
 - Python web scraping guide for journalists: <https://project.journalism.torontomu.ca/jrn-305-2021/2021/04/22/a-simple-python-web-scraping-guide-for-journalists/>
- Coding games are a great way to learn because they keep it fun!
 - <https://www.freecodecamp.org/news/12-free-coding-games-to-learn-programming-for-beginners/>

Other helpful tips/resources to know:

- There's such a wealth of information around data journalism. Literally any question you have someone else has encountered it, I promise you
 - Google and [Stack Overflow](#) are your friends
- Other conferences like IRE and NICAR post all their tipsheets too, take advantage of them!
 - <https://www.ire.org/training/conferences/nicar-2024/nicar24-tipsheets-audio/>