Data Journalism with R and the Tidyverse

Code, data and visuals for storytellers

Derek Willis and Sean Mussenden, based on earlier work by Matt Waite & Sarah Co

2025-01-16

Table of contents

1 Introduction

If you were at all paying attention in pre-college science classes, you have probably seen this equation:

```
d = rt or distance = rate*time
```

In English, that says we can know how far something has traveled if we know how fast it's going and for how long. If we multiply the rate by the time, we'll get the distance.

If you remember just a bit about algebra, you know we can move these things around. If we know two of them, we can figure out the third. So, for instance, if we know the distance and we know the time, we can use algebra to divide the distance by the time to get the rate.

```
d/t = r or distance/time = rate
```

In 2012, the South Florida Sun Sentinel found a story in this formula.

People were dying on South Florida tollways in terrible car accidents. What made these different from other car fatal car accidents that happen every day in the US? Police officers driving way too fast were causing them.

But do police regularly speed on tollways or were there just a few random and fatal exceptions?

Thanks to Florida's public records laws, the Sun Sentinel got records from the toll transponders in police cars in south Florida. The transponders recorded when a car went through a given place. And then it would do it again. And again.

Given that those places are fixed – they're toll plazas – and they had the time it took to go from one toll plaza to another, they had the distance and the time.

It took high school algebra to find how fast police officers were driving. And the results were shocking.

Twenty percent of police officers had exceeded 90 miles per hour on toll roads. In a 13-month period, officers drove between 90 and 110 mph more than 5,000 times. And these were just instances found on toll roads. Not all roads have tolls.

The story was a stunning find, and the newspaper documented case after case of police officers violating the law and escaping punishment. And, in 2013, they won the Pulitzer Prize for Public Service.

All with simple high school algebra.

1.1 Modern data journalism

It's a single word in a single job description, but a Buzzfeed job posting in 2017 is another indicator in what could be a profound shift in how data journalism is both practiced and taught.

"We're looking for someone with a passion for news and a commitment to using data to find amazing, important stories — both quick hits and deeper analyses that drive conversations," the posting seeking a data journalist says. It goes on to list five things BuzzFeed is looking for: Excellent collaborator, clear writer, deep statistical understanding, knowledge of obtaining and restructuring data.

And then there's this:

"You should have a strong command of at least one toolset that (a) allows for filtering, joining, pivoting, and aggregating tabular data, and (b) enables reproducible workflows."

This is not the data journalism of 20 years ago. When it started, it was a small group of people in newsrooms using spreadsheets and databases. Data journalism now encompases programming for all kinds of purposes, product development, user interface design, data visualization and graphics on top of more traditional skills like analyzing data and writing stories.

In this book, you'll get a taste of modern data journalism through programming in R, a statistics language. You'll be challenged to think programmatically while thinking about a story you can tell to readers in a way that they'll want to read. They might seem like two different sides of the brain – mutually exclusive skills. They aren't. I'm confident you'll see programming is a creative endeavor and storytelling can be analytical.

Combining them together has the power to change policy, expose injustice and deeply inform.

1.2 Installations

This book is all in the R statistical language. To follow along, you'll do the following:

1. Install the R language on your computer. Go to the R Project website, click download R and select a mirror closest to your location. Then download the version for your computer.

2. Install R Studio Desktop. The free version is great.

Going forward, you'll see passages like this:

```
install.packages("tidyverse")
```

That is code that you'll need to run in your R Studio. When you see that, you'll know what to do.

1.3 About this book

This book is the collection of class materials originally written for Matt Waite's Data Journalism class at the University of Nebraska-Lincoln's College of Journalism and Mass Communications. It has been substantially updated by Derek Willis and Sean Mussenden for data journalism classes at the University of Maryland Philip Merrill College of Journalism, with contributions from Sarah Cohen of Arizona State University.

There's some things you should know about it:

- It is free for students.
- The topics will remain the same but the text is going to be constantly tinkered with.
- What is the work of the authors is copyright Matt Waite 2020, Sarah Cohen 2022 and Derek Willis and Sean Mussenden 2023.
- The text is Attribution-NonCommercial-ShareAlike 4.0 International Creative Commons licensed. That means you can share it and change it, but only if you share your changes with the same license and it cannot be used for commercial purposes. I'm not making money on this so you can't either.

- As such, the whole book – authored in Quarto – in its original form is open sourced on Github. Pull requests welcomed!

1.4 What we'll cover

- Google Sheets
- Public records and open data
- R Basics
- Replication
- Data basics and structures
- Aggregates
- Mutating

- Working with dates
- Filters
- Cleaning I: Data smells
- Cleaning II: Janitor
- Cleaning III: Open Refine
- Cleaning IV: Pulling Data from PDFs
- Joins
- Basic data scraping
- Getting data from APIs: Census
- Visualizing for reporting: Basics
- Visualizing for reporting: Publishing
- Geographic data basics
- Geographic queries
- Geographic visualization
- Text analysis basics
- Basic statistics
- Writing with and about data
- Data journalism ethics

2 Learn a new way to read

Getting started in data journalism often feels as if you've left the newsroom and entered the land of statistics, computer programming and data science. This chapter will help you start seeing data reporting in a new way, by learning how to study great works of the craft as a writer rather than a reader.

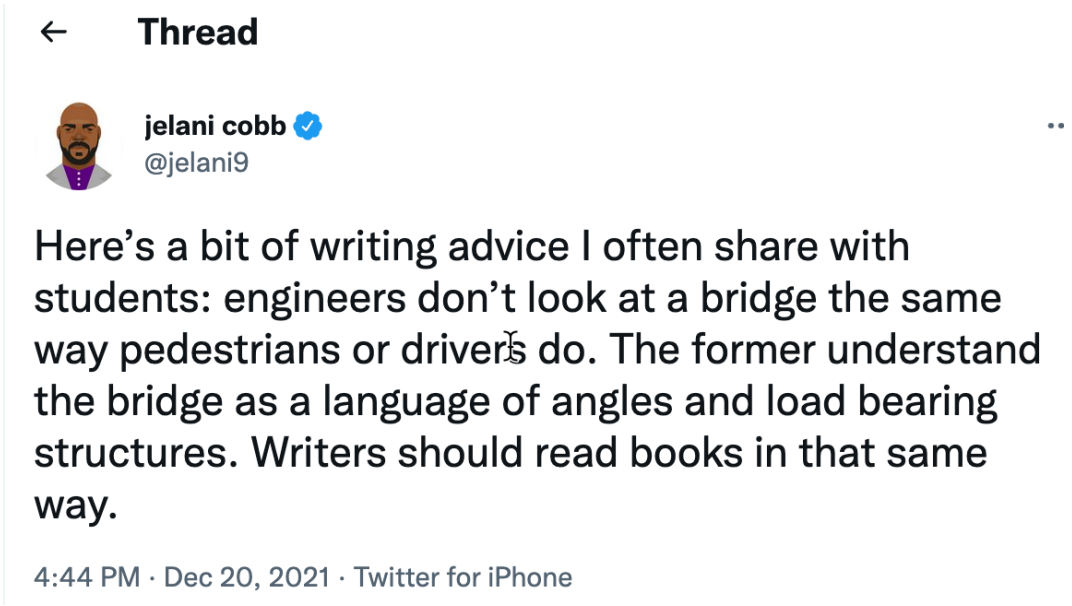


Figure 2.1: jelani cobb

Jelani Cobb tweeted, "an engineer doesn't look at a bridge the same way pedestrians or drivers do." They see it as a "language of angles and load bearing structures." We just see a bridge. While he was referring to long-form writing, reporting with data can also be learned by example – if you spend enough time with the examples.

Almost all good writers and reporters try to learn from exemplary work. I know more than one reporter who studies prize-winning journalism to hone their craft. This site will have plenty of examples, but you should stay on the lookout for others.

2.1 Read like a reporter

Try to approach data or empirical reporting as a reporter first, and a consumer second. The goal is to triangulate how the story was discovered, reported and constructed. You'll want to think about why *this* story, told this way, at this time, was considered newsworthy enough to publish when another approach on the same topic might not have been.

What were the questions?

In data journalism, we often start with a tip, or a hypothesis. Sometimes it's a simple question. Walt Bogdanich of The New York Times is renowned for seeing stories around every corner. Bogdanich has said that the prize-winning story "A Disability Epidemic Among a Railroad's Retirees" came from a simple question he had when railway workers went on strike over pension benefits – how much were they worth? The story led to an FBI investigation and arrests, along with pension reform at the largest commuter rail in the country.

The hypothesis for some stories might be more directed. In 2021, the Howard Center for Investigative Journalism at ASU published "Little victims everywhere", a set of stories on the lack of justice for survivors of child sexual assault on Native American reservations. That story came after previous reporters for the center analyzed data from the Justice Department showing that the FBI dropped most of the cases it investigated, and the Justice Department then only prosecuted about half of the matters referred to it by investigators. The hypothesis was that they were rarely pursued because federal prosecutors – usually focused on immigration, white collar crime and drugs – weren't as prepared to pursue violent crime in Indian Country.

When studying a data-driven investigation, try to imagine what the reporters were trying to prove or disprove, and what they used to do it. In journalism, we rely on a mixture of quantitative and qualitative methods. It's not enough to prove the "numbers" or have the statistical evidence. That is just the beginning of the story. We are supposed to ground-truth them with the stories of actual people and places.

Go beyond the numbers

It's easy to focus on the numbers or statistics that make up the key findings, or the reason for the story. Some reporters make the mistake of thinking all of the numbers came from the same place – a rarity in most long-form investigations. Instead, the sources have been woven together and are a mix of original research and research done by others. Try to pay attention to any sourcing done in the piece. Sometimes, it will tell you that the analysis was original. Other times it's more subtle.

But don't just look at the statistics being reported in the story. In many (most?) investigations, some of the key people, places or time elements come directly from a database.

When I was analyzing Paycheck Protection Program loan data for ProPublica, one fact hit me as I was looking at a handful of sketchy-looking records: a lot of them were from a single county in coastal New Jersey. It turned out to be a pretty good story.

Often, the place that a reporter visits is determined by examples found in data. In this story on rural development funds, all of the examples came from an analysis of the database. Once the data gave us a good lead, the reporters examined press releases and other easy-to-get sources before calling and visiting the recipients or towns.

2.2 Reading tips

You'll get better at reading investigations and data-driven work over time, but for now, remember to go beyond the obvious:

- Where might the reporters have found their key examples, and what made them good characters or illustrations of the larger issue? Could they have come from the data?
- What do you think came first – a narrative single example that was broadened by data (naively, qualitative method), or a big idea that was illustrated with characters (quantitative method)?
- What records were used? Were they public records, leaks, or proprietary data?
- What methods did they use? Did they do their own testing, use statistical analysis, or geographic methods? You won't always know, but look for a methodology section or a description alongside each story.
- How might you localize or adapt these methods to find your own stories?
- Pick out the key findings (usually in the nut graf or in a series of bullets after the opening chapter): are they controversial? How might they have been derived? What might have been the investigative hypothesis? Have they given critics their due and tried to falsify their own work?
- How effective is the writing and presentation of the story? What makes it compelling journalism rather than a dry study? How might you have done it differently? Is a video story better told in text, or would a text story have made a good documentary? Are the visual elements well integrated? Does the writing draw you in and keep you reading? Think about structure, story length, entry points and graphics all working together.
- Are you convinced? Are there holes or questions that didn't get addressed?

2.3 Analyze data for story, not study

As journalists we'll often be using data, social science methods and even interviewing differently than true experts. We're seeking stories, not studies. Recognizing news in data is one of the hardest skills for less experienced reporters new to data journalism. This list of potential newsworthy data points is adapted from Paul Bradshaw's "Data Journalism Heist".

LAST UPDATED ON 2015-06-10

- Compare the claims of powerful people and institutions against facts – the classic investigative approach.
- Report on *unexpected* highs and lows (of change, or of some other characteristic)

- Look for outliers – individual values that buck a trend seen in the rest
- Verify or bust some myths
- Find signs of distress, happiness or dishonesty or any other emotion.
- Uncover *new* or *under-reported* long-term trends.
- Find data suggesting your area is *the same* or *different* than most others of its kind.

Bradshaw also did a recent study of data journalism pieces: "Here are the angles journalists use most often to tell the stories in data", in Online Journalism Blog. I'm not sure I agree, only because he's looking mainly at visualizations rather than stories, but they're worth considering.

2.4 Exercises

- If you're a member of Investigative Reporters and Editors, go to the site and find a recent prize-winning entry (usually text rather than broadcast). Get a copy of the IRE contest entry from the Resources page. Try to match up what the reporters said they did and how they did it with key portions of the story.
- The next time you find a good data source, try to find a story that references it. If your data is local, you might look for a story that used similar data elsewhere, such as 911 response times or overdose deaths. But many stories use federal datasets that can easily be localized. Look at a description of the dataset and then the story to see how the data might have been used.

3 Newsroom math

Statistics are people with the tears washed off

- Paul Brodeur

Jo Craven McGinty, then of The New York Times, used simple rates and ratios to discover that a 6-story brick New Jersey hospital was the most expensive in the nation. In 2012, Bayonne Medical Center "charged the highest amounts in the country for nearly one-quarter of the most common hospital treatments," the Times story said.

To do this story, McGinty only needed to know the number of the procedures reported to the government and the total amount each hospital charged. Dividing those to find an average price, then ranking the most common procedures, led to this surprising result.

3.1 Why numbers?

Using averages, percentages and percent change is the bread and butter of data journalism, leading to stories ranging from home price comparisons to school reports and crime trends. It may have been charming at one time for reporters to announce that they didn't "do" math, but no longer. Instead, it is now an announcement that the reporter can only do some of the job. You will never be able to tackle complicated, in-depth stories without reviewing basic math.

The good news is that most of the math and statistics you need in a newsroom isn't nearly as difficult as high school algebra. You learned it somewhere around the 4th grade. You then had a decade to forget it before deciding you didn't like math. But mastering this most basic arithmetic again is a requirement in the modern age.

In working with typical newsroom math, you will need to learn how to:

- Overcome your fear of numbers
- Integrate numbers into your reporting
- Routinely compute averages, differences and rates
- Simplify and select the right numbers for your story

While this chapter covers general tips, you can find specific instructions for typical newsroom math in this Appendix A

3.2 Overcoming your fear of math

When we learned to read, we got used to the idea that 26 letters in American English could be assembled into units that we understand without thinking – words, sentences, paragraphs and books. We never got the same comfort level with 10 digits, and neither did our audience.

Think of your own reaction to seeing a page of words. Now imagine it as a page of numbers.

Instead, picture the number "five". It's easy. It might be fingers or it might be a team on a basketball court. But it's simple to understand.

Now picture the number 275 million. It's hard. Unfortunately, 275 billion isn't much harder, even though it's magnitudes larger. (A million seconds goes by in about 11 days but you may not have been alive for a billion seconds – about 36 years.)

The easiest way to get used to some numbers is to learn ways to cut them down to size by calculating rates, ratios or percentages. In your analysis, keep an eye out for the simplest *accurate* way to characterize the numbers you want to use. "Characterize" is the important word here – it's not usually necessary to be overly precise so long as your story doesn't hinge on a nuanced reading of small differences. (And is anything that depends on that news? It may not be.)

Here's one example of putting huge numbers in perspective. Pay attention to what you really can picture - it's probably the $21 equivalent.

The Chicago hedge fund billionaire Kenneth C. Griffin, for example, earns about $68.5 million a month after taxes, according to court filings made by his wife in their divorce. He has given a total of $300,000 to groups backing Republican presidential candidates. That is a huge sum on its face, yet is the equivalent of only $21.17 for a typical American household, according to Congressional Budget Office data on after-tax income. *"Buying Power", Nicholas Confessore, Sarah Cohen and Karen Yourish, The New York Times, October 2015*

Originally the reporters had written it even more simply, but editors found the facts so unbelievable that they wanted give readers a chance to do the math themselves. That's reasonable, but here's an even simpler way to say it: "earned nearly $1 billion after taxes…He has given $300,000 to groups backing candidates, the equivalent of a dinner at Olive Garden for the typical American family , based on Congressional Budget Office income data." (And yes, the reporter checked the price for an Olive Garden meal at the time for four people.)

3.3 Put math in its place

For journalists, numbers – or facts – make up the third leg of a stool supported by human stories or anecdotes, and insightful comment from experts. They serve us in three ways:

- ***As summaries***. Almost by definition, a number counts something, averages something, or otherwise summarizes something. Sometimes, it does a good job, as in the average height of Americans. Sometimes it does a terrible job, as in the average income of Americans. Try to find summaries that accurately characterize the real world.
- ***As opinions***. Sometimes it's an opinion derived after years of impartial study. Sometimes it's an opinion tinged with partisan or selective choices of facts. Use them accordingly.
- ***As guesses***. Sometimes it's a good guess, sometimes it's an off-the-cuff guess. And sometimes it's a hopeful guess. Even when everything is presumably counted many times, it's still a (very nearly accurate) guess. Yes, the "audits" of presidential election results in several states in 2021 found a handful of errors – not a meaningful number, but a few just the same.

Once you find the humanity in your numbers, by cutting them down to size and relegating them to their proper role, you'll find yourself less fearful. You'll be able to characterize what you've learned rather than numb your readers with every number in your notebook. You may even find that finding facts on your own is fun.

3.4 Going further

Tipsheets

- Steve Doig's "Math Crib Sheet"
- Appendix A: Common newsroom math, adapted from drafts of the book *Numbers in the Newsroom*, by Sarah Cohen.

Reading and viewing

- "Avoiding Numeric Novocain: Writing Well with Numbers," by Chip Scanlan, Poynter.com
- T. Christian Miller's "Writing the data-driven story"
- A viral Twitter thread:

3.5 Exercises

- Imagine that someone gave you $1 million and you could spend it on anything you want. Write down a list of things that would add up to about that amount. That should be easy. Now, imagine someone gave you $1 billion and you could spend it on whatever you want, but anything left over after a year had to be returned. How would you spend it? (You can give away money, but it can't be more than 50% of a charity's annual revenues. So you can't give 10 $100 million gifts!) See how far you get trying to spend it. A few homes, a few yachts, student loan repayments for all of your friends? You've hardly gotten started.

4 Defining "Data"

> data / de .tə/ :
>
> information in an electronic form that can be stored and used by a computer, or information, especially facts or numbers, collected to be examined and >considered and used to help decision-making
>
> – Cambridge Dictionary – sort of [1]

4.1 The birth of a dataset

Most journalism uses data collected for one purpose for something entirely different. Understanding its original uses – what matters to the people who collected it, and what doesn't – will profoundly affect its accuracy or usefulness.

Trace data and administrative records

In "The Art of Access", David Cullier and Charles N. Davis describe a process of tracking down the life and times of a dataset. Their purpose is to make sure they know how to request it from a government agency. The same idea applies to using data that we acquire elsewhere.

Understanding how and why data exists is crucial to understanding what you, as a reporter, might do with it.

Anything you can systematically search or analyze could be considered one piece of of data. As reporters, we usually deal with data that was created in the process of doing something else – conducting an inspection, delivering a tweet, or scoring a musical. In the sciences, this flotsam and jetsom that is left behind is called "digital trace data" if it was born digitally.

In journalism and in the social sciences, many of our data sources were born during some government process – a safety inspection, a traffic ticket, or the filing of a death certificate. These administrative records form the basis of much investigative reporting and they are often the subject of public records and FOIA requests. They were born as part of the government doing its job, without any thought given to how it might be used in another way. In the sciences, those are often called "administrative records".

[1] I flipped the order of these two definitions!

This trace data might be considered the first part of the definition above – information that can be stored and used.

Here's how Chris Bail from Duke University describes it.

Data collected and curated for analysis

Another kind of data is that which is compiled or collected specifically for the purpose of studying something. It might collected in the form of a survey or a poll, or it might be a system of sampling to measure pollution or weather. But it's there because the information has intrinsic value AS information.

The video suggests a hard line between trace data and custom data. In practice, it's not that clear. Many newsrooms may curate data published in other sources or in administrative records, such as the Washington Post's police shooting dataset. In other cases, the agencies we are covering get already-compiled data from state and local governments.

This type of data might be considered the second type in the definition – tabular information that is used for decision-making.

4.2 Granular and aggregated data

One of the hardest concepts for a lot of new data journalists is the idea of *granularity* of your data source. There are a lot of ways to think about this: individual items in a list vs. figures in a table; original records vs. compilations; granular data vs. statistics.

Generally, an investigative reporter is interested in getting data that is as close as possible to the most granular information that exists, at least on computer files. Here's an example, which might give you a little intuition about why it's so important to think this way:

When someone dies in the US, a standard death certificate is filled out by a series of officials - the attending physician, the institution where they died and even the funeral direcor.

Click on this link to see a blank version of the standard US death certificate form – notice the detail and the detailed instructions on how it is supposed to be filled out. [2]

A good reporter could imagine many stories coming out of these little boxes. Limiting yourself to just to COVID-19-related stories: You could profile the local doctor who signed the most COVID-19-related death certificates in their city, or examine the number of deaths that had

[2]You should do this whenever you get a dataset created from administrative records. That is, track down its origin and examine the pieces you were given and the pieces that were left out; look at what is written in free-form vs what is presented as a check box. You may need a copy of the template that an agency uses to collect the information, but many governments make these available on their websites or are willing to provide them without a fuss.

COVID as a contributing, but not underlying or immediate, cause of death. You could compare smoking rates in the city with the number of decedents whose tobacco use likely contributed to their death. Maybe you'd want to know how long patients suffered with the disease before they died. And you could map the deaths to find the block in your town most devastated by the virus.

Early in the pandemic, Coulter Jones and Jon Kamp examined the records from one of the few states that makes them public, and concluded that "Coronavirus Deaths were Likely Missed in Michigan, Death Certificates Suggest"

But you probably can't do that. The reason is that, in most states, death certificates are not public records and are treated as secrets. [3]. Instead, state and local governments provide limited statistics related to the deaths, usually by county, with no detail. That's the difference between granular data and aggregate data. Here are some of the typical (not universal) characteristics of each:

Granular	Aggregate
Intended for some purpose other than your work	Intended to be presented as is to the public
Many rows (records), few columns (variables)	Many columns (variables), few rows (records)
Requires a good understanding of the source	Explanatory notes usually come with the data
Easy to cross-reference and compile	Often impossible to repurpose
Has few numeric columns	May be almost entirely numerical
Is intended for use in a database	Is intended for use in a spreadsheet

We often have to consider the trade-offs. Granular data with the detail we need - especially when it involves personally identifiable information like names and addresses - can take months or years of negotiation over public records requests, even when the law allows it. It's often much easier to convince an agency to provide summarized or incomplete data. Don't balk at using it if it works for you. But understand that in the vast majority of cases, it's been summarized in a way that's lost information that could be important to your story.

4.3 Nouns

That brings us to one of the most important things you must find out about any data you begin to analyze: What "noun" does each row in a tabular dataset represent? In statistics, they might be called *observations* or *cases*. In data science, they're usually called *records*. Either

[3] See "Secrecy in Death Records: A call to action", by Megain Craig and Madeleine Davison, Journal of Civic Information, December 2020

way, every row must represent the same thing – a person, a place, a year, a water sample or a school. And you can't really do anything with it until you figure out what that is.

In 2015, Sarah Cohen did a story at The New York Times called "More Deportation Follow Minor Crimes, Records Show". The government had claimed it was only removing hardened criminals from the country, but our analysis of the data suggested that many of them were for minor infractions.

In writing the piece, they had to work around a problem in our data: the agency refused to provide them anything that would help us distinguish individuals from one another. All the reporters knew was that each row represented one deportation – not one person! Without a column, or *field* or a *variable* or an *attribute* for an individual – say, name and date of birth, or some scrambled version of an their DHS number – they had no way to even estimate how often people were deported multiple times. If you read the story, you'll see the very careful wording, except when they had reported out and spoken to people on the ground.

4.4 Further reading

- "Basic steps in working with data", the Data Journalism Handbook, Steve Doig, ASU Professor. He describes in this piece the problem of not knowing exactly how the data was compiled.
- "Counting the Infected" , Rob Gebellof on The Daily, July 8, 2020.
- "Spreadsheet thinking vs. Database thinking", by Robert Kosara, gets at the idea that looking at individual items is often a "database", and statistical compilations are often "spreadsheets".
- "Tidy Data", in the Journal of Statistical Software (linked here in a pre-print) by Hadley Wickham , is the quintessential article on describing what we think of as "clean" data. For our purposes, much of what he describes as "tidy" comes when we have individual, granular records – not statistical compilations. It's an academic article, but it has the underlying concepts that we'll be working with all year.

4.5 Exercises

- The next time you get a government statistical report, scour all of the footnotes to find some explanation of where the data came from. You'll be surprised how often they are compilations of administrative records - the government version of trace data.

5 Introduction

Some people consider using spreadsheets the table stakes for getting into data journalism. It's relatively easy to see what you're doing and you can easily share your work with your colleagues. In fact, pieces of the Pulitzer-Prize winning COVID-19 coverage from The New York Times was compiled using an elaborate and highly tuned set of Google spreadsheets with dozens of contributors.

This guide uses Google Sheets, although you should be able to do these exercises with Excel on the Mac or Windows. Excel on the Mac is pretty good, bu Excel in Windows is very different – it has much more capability for working with large and more complex data, and provides better tuning for import and other operations. There is a table that compares keystrokes for Apple desktops, laptops and Windows machines for Excel at the bottom of An Excel Refresher

5.1 Tutorials

Spreadsheets in the form of Google Sheets or Excel are used in almost every workplace in America. This section covers most of what you need in the newsroom, which is a different set of skills than in other businesses.

- An Excel Refresher : Start over with good habits
- Sorting and filtering to find stories : The first step of interviewing data
- Grouping with pivot tables: Aggregating, and the super power of spreadsheets
- Formulas in Excel: Percents, sums, and other basic computations used in newsrooms.

5.2 Practice exercises

- Practice with "notice of claims" from Phoenix: Filtering and pivot table practice using claims made against the city of Phoenix 2010-2020.

6 An Excel Refresher

Spreadsheets are everywhere, so it's worth re-learning how to use them well. Reporters usually use spreadsheets in three ways:

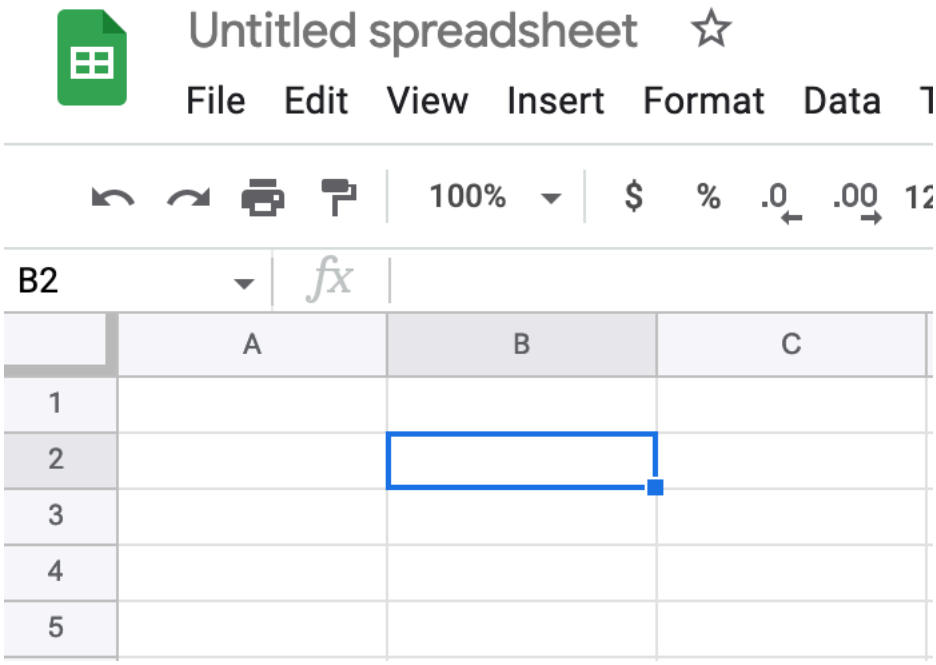
- To create original databases of events for sorting, filtering and counting. Examples include a long-running court case; the details of each opioid death in a city; a list of police shootings and their documents; or even a list of your own public records requests or contact log.
- To use data created by others for fast, simple analysis and data cleanup. Many government agencies provide their information in spreadsheet form, but they often require some rejiggering before you can use them.
- To perform simple, straightforward analysis on data and share with team members. This is becoming less common as more reporters learn programming languages, but it's still common in newsrooms to share data, especially through Google Sheets.

(This guide is done using a Mac. Windows machines will be a little different, mainly because you'll have more choices in most menus. The Mac CMD key is the same as the Windows CNTL key.)

Some reporters flinch at typing in 30 or 100 entries into a spreadsheet. You shouldn't. If you learn to take notes in a structured way, you'll always be able to find and verify your work. If you try to calculate a sum of 30 numbers on a calculator, you'll have to type them all in at least twice anyway. Also, getting used to these easy tasks on a spreadsheet keeps your muscles trained for when you need to do more.

6.1 Re-learning Sheets from the ground up

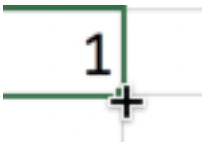
6.1.1 The spreadsheet grid



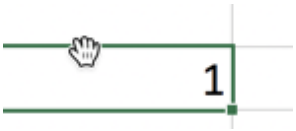
When you start up a spreadsheet (try entering sheet.new into your browser!), you'll see letters across the top and numbers down the side. If you ever played Battleship, you'll recognize the idea – every little square, or cell, is referenced by the intersection of its column letter and row number:

B2 is the cell that is currently active. You can tell because it's outlined in the sheet and it's shown on the upper left corner.

6.1.2 Mouse shapes



The Copy Tool, or the thin black cross. When you see this, you'll copy anything that's selected. This can be good or bad.



The Evil Hand. (In Windows, this is the Evil Arrow). If you use this symbol, you will MOVE the selection to a new location. This is very rarely a good idea or something you intend.

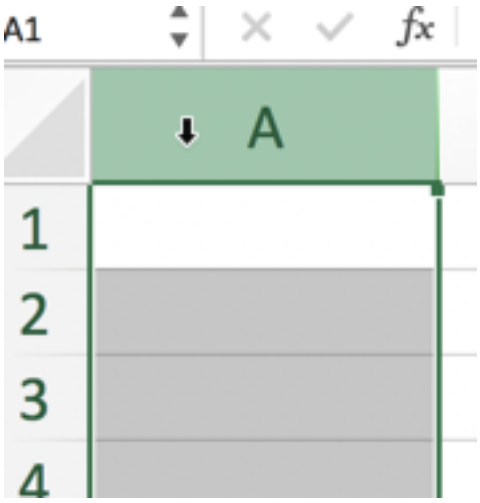
6.1.3 Selecting cells and ranges

Spreadsheets act only on the cells or regions you have selected. If you begin typing, you'll start entering information into the currently selected cell.

To select: Hold the BFWPS over the cell and clice ONCE – *not twice*. Check the formula bar to make sure you've selected what you think you've got. You can also look at the bottom right of your spreadsheet for more information.

You'll often work with *ranges* of cells in formulas. These are defined by the corners of the area you want to work on – often a column of information. In the example below, the range is A1:B6, with the ":" referring to the word "through".

To select a group of cells and act on them all at once: Hover the BFWPS over one corner, click ONCE and drag to the diagonal corner. Make sure the Evil Hand is nowhere to be seen. The entire area will be shaded in except for the currently selected cell. Look at the upper right corner to see how many rows and columns you selected.



To select a column or row : Hover the cursor over the letter at the top of the column. For a row, hover it over the row number in the margin

6.1.4 Reading the screen

The areas of the spreadsheet have different visual clues, and learning to read them will make your life much easier.

6.1.5 Entering data

Select the cell and start typing. The information you type won't be locked into the cell until you hit the Return / Enter key, or move your selection to another cell. Hit "Escape" to cancel the entry.

You can't do a lot of things while you're editing, so if you have a lot of greyed out menu items, look at your formula bar to see if you are still editing a cell.

If you're having trouble getting to a menu item or seeing the result of your work, try hitting "Escape" and try again. You may not have actually entered the information into the sheet.

6.1.6 Locking in headings

As your spreadsheet grows vertically with more rows, you'll want to be able to see the top all the time. When it grows horizontally with more columns, you'll probably want to see columns in the left, such as names. This is called "Freezing Panes" – you freeze part of the page so it stays in place when you move around.

In Google Sheets, this is done via the View -> Freeze menu:

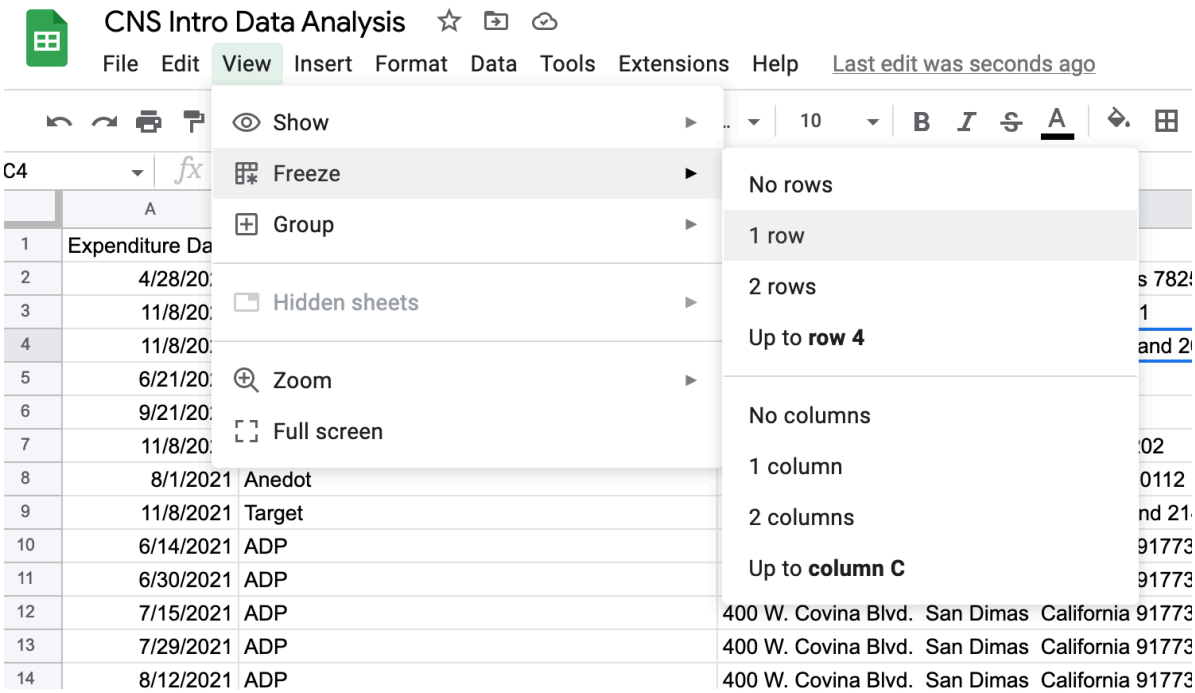


Figure 6.1: freeze panes

6.1.7 Formatting tricks

- Use the buttons or the format dialog box to make numbers easier to read.
- If a column is filled with a lot of text, select the column and look on the Home ribbon next to the formatting area for "Wrap Text". This means that when you double-click to widen a column, it will get taller, not wider. This is good when you need to save valuable real estate on the screen.

6.2 Getting started with a dataset

SLOW DOWN! Don't do anything until you understand what you have in front of you and can predict what your next mouse click will do to it.

Most data we encounter was created by someone else for some purpose other than ours. This means that you can't assume anything. It may not be complete. It may be inaccurate. It may mean something completely different than it appears at first blush.

6.2.1 First steps

- Document where you got the spreadsheet and how you can get back to the original.
- Read anything you can about what it contains. Look for documentation that comes with the data.
- Save the original into a safe place with its original name and metadata. Work on a copy.
- If the spreadsheet shows #### instead of words or numbers, widen your columns. If it shows 7E-14 or something like that, format them as numbers, not "General".
- Check your corners – look at the top left and bottom right. Is the data all in one area? Are there footnotes or other non-data sections mixed in? We're going to want to fix that later.

6.2.2 Interview your data

6.2.2.1 Headings

The most fraught part of data reporting is understanding what each *column* actually means. These often have cryptic, bureaucratic names. You may need to go back to the source of the data to be sure you actually understand them.

If your data doesn't have any headings, that's going to be your first priority. In effect, you'll need to build what we call a *data dictionary* or *record layout* if one hasn't been provided. Many reporters create these as a page in a dataset.

6.2.2.2 Unit of analysis

A *unit of analysis* refers to the items that are listed in the rows of your dataset. Ideally, every row should be at the same unit of analysis – a person, an inspection, or a city, for example. Summaries should be separated by a blank row, or moved to a different sheet. Think of this as the noun you'd use to describe every row.

6.2.2.3 Row numbers

The data was probably given to you in some sort of natural sort order. Different computer systems sort differently – some are case-sensitive, others are not. It may depend on when and where the data as created! The order of the data may even depend on a column you don't have. If you don't do something now, you'll never be able to get back to the original order, which could have meaning for both the agency and for fact-checking.

7 Sorting and filtering to find stories

7.1 A sorting miracle

When Stephen Neukam - who was sitting in this class a year ago - wanted to find out who was funding candidates for Maryland's open governor's seat this year, he downloaded data from the State Board of Elections that listed contributions to the wide array of hopefuls seeking to replace Larry Hogan in Annapolis.

He wasn't sure at first what he was looking for, so he started the way that many reporters do with data: by sorting and filtering. Were there outliers in the list of contributions, and which candidates were getting their money from unusual (non-Maryland) sources?

Neukam quickly found his story: in the race to be governor, Maryland candidates, and in particular Wes Moore, a first-time Democratic candidate, were raising millions of dollars from out of state donors.

The story, "Millions in out-of-state donations help fuel high-profile Maryland Democratic governor candidates" helped explain where candidates were going to fund one of the most contested primaries in recent history (Moore ended up winning).

7.2 Sorting and filtering as a reporting tool

Sorting and filtering can:

- Narrow your focus to specific items that you want to examine in your story.
- Show you rows containing the highest and lowest values of any column. That can be news or it can be errors or other problems with the data.
- Let you answer quick "how many?" questions, with a count of the rows that match your criteria. (In the next lesson, you'll see that pivot tables, or group-by queries, are much more powerful for this in most cases.)

7.3 Example data

::: {.alert .alert-info } - Data from the State Board of Elections for use in this tutorial - Documentation from the SBOE's site :::

- The data for this is from the Maryland State Board of Elections's Campaign Finance Database. There are a couple of caveats:
- It includes money raised as of Jan. 12, 2022, which covers all of 2021.
- These are self-reported by campaigns, and subject to amendment in case of errors or omissions.

The original data download link for Wes Moore's contributions is https://github.com/stephenneukam/CNS_Annapolis/raw/main/Campaign_finance/Moore_ContributionsList.csv. Download it to your computer and then, in a browser, type sheet.new to create a new Google Sheet. From there, use File -> Import and choose "Upload" and select the file on your computer. Click the "Import Data" button when it appears. Then give your sheet a name, like "Wes Moore Contributions".

It's a good example set for us because it's been used as the basis of Neukam's story and it has at least one of each *data type* that we plan to deal with in Google Sheets or Excel. And, critically, the first row contains headers, not data. Always have headers, even if you have to add them.

7.4 Understanding data types

When you open the spreadsheet, the first thing to notice is its *granularity*. Unlike Census or budget spreadsheets, this is a list capturing specific characteristics of each contribution. Each column has the same *type* of data from top to bottom. Those types are:

- **Text**. Text or "character" columns can come in long or short form. When they are standardized (the values can contain only one of a small list of values), they're called "categorical". If they're more free-form, they're might be called "free text". The computer doesn't know the difference, but you should. The Post data has examples of both. In spreadsheets, text is left-justified (they move toward the left of the cell and will line up vertically at the beginning)
- **Numbers**. These are pure numbers with no commas, dollar signs or other embellishments. In Google Sheets these can be formatted to *look* like numbers in different ways, but underneath they're just numbers. Adding up a column of numbers that has a word in it or has missing values will just be ignored. It will trip up most other languages. These are right-justified, so the last digit is always lined up vertically.

- **Logical**: This is a subset of text. It can take one of only two values – yes or no, true or false. There is no "maybe".
- **Date and times**: These are actual dates on the calendar, which have magical properties. Underneath, they are a number. In Google Sheets and Excel, that number is the number of days since Jan. 1, 1900.[1] They can also have time attached to them, which is a fraction of a day. What this means is that the number 44,536.5 is really Dec. 6, 2021 at noon. In Sheets, you use a format to tell the spreadsheet how you want to see the date or time, just the way you look at dollar values with commas and symbols. (If you get a spreadsheet with a lot of dates of 1/1/1900, it means there is a 0 in that column, which is sometimes a fill-in for "I don't know.")

Here's a picture of a date that is shown in a variety of formats.

Unformatted	Formatted values					
As a number	"Short date"	"Long date"	Time	Date & mil. time	Month	Day of the week
44540.87431	12/10/21	Friday, December 10, 2021	8:59:00 PM	12/10/21 20:59	Dec. 2021	Friday

Figure 7.1: date formats

All of these are the same, underlying value – the number at the left. Notice that all of these are right-justified.

This means that when you see "Friday, December 10", the computer sees 44540.87431. When you put the dates in order, they won't be alphabetized with all of the Fridays shown together. Instead, they'll be arranged by the actual date and time.

It also means that you can compute 911 response times even when it crosses midnight, or or compute the someone's age today given a date of birth. Keeping actual calendar dates in your data will give it much more power than just having the words. (Sheets and Excel use the 1st of the month as a stand-in for an actual date when all you know is the month and year.)

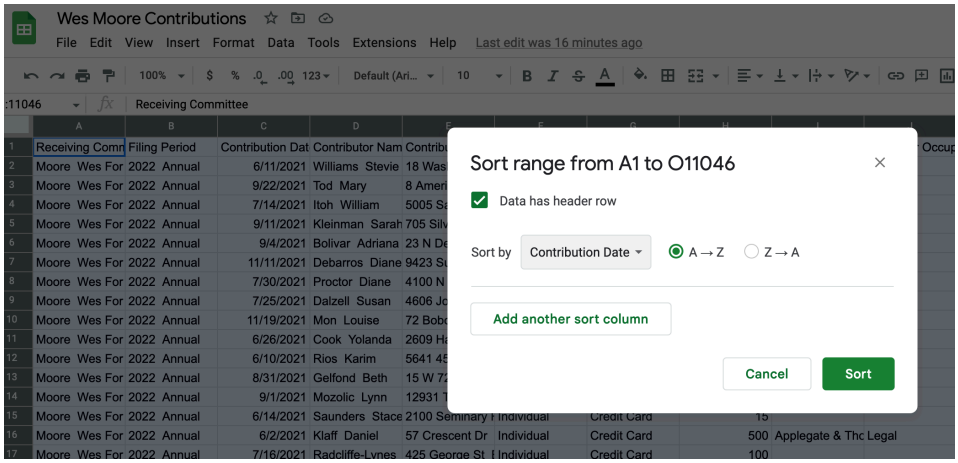
7.4.1 Sorting rows

Sorting means rearranging the rows of a data table into a different order. Some reporters take a conceptual shortcut and call this "sorting columns". That thinking will only get you into trouble – it lets you forget that you want to keep the rows in tact while changing the order in which you see them. In fact, in other languages it's called "order by" or "arrange" by one or more columns – a much clearer way to think of it.

To sort in Google Sheets, first highlight the entire sheet by clicking on the button above the first row and to the left of the first column. So, between the 1 and the A. Then, look for

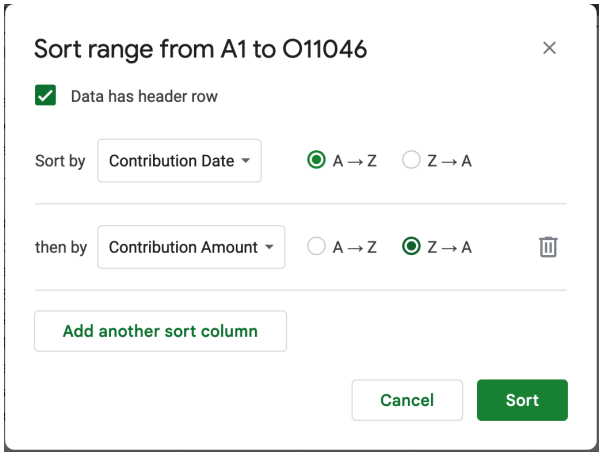
[1] Each language deals with dates and times a little differently. We'll see how R does it later on. But just know that dates can be tricky because of these differences and time is even more tricky

the sort options under the Data drop-down menu at the top of your screen, and choose “Sort Range” and then “Advanced range sorting options”. Trust me, this is how you want to do it. Check the box that says “Data has header row” and Sort by Contribution Date. In this case, sorting from A to Z gives you a list of the contributions in chronological order.



Adding fields to the sort

Adding more columns to the sort box tells Sheets what to do when the first one is the same or tied. For example, sorting first by date then by amount gives you a list that shows all of the contributions by date in sequence:



7.4.2 Filtering

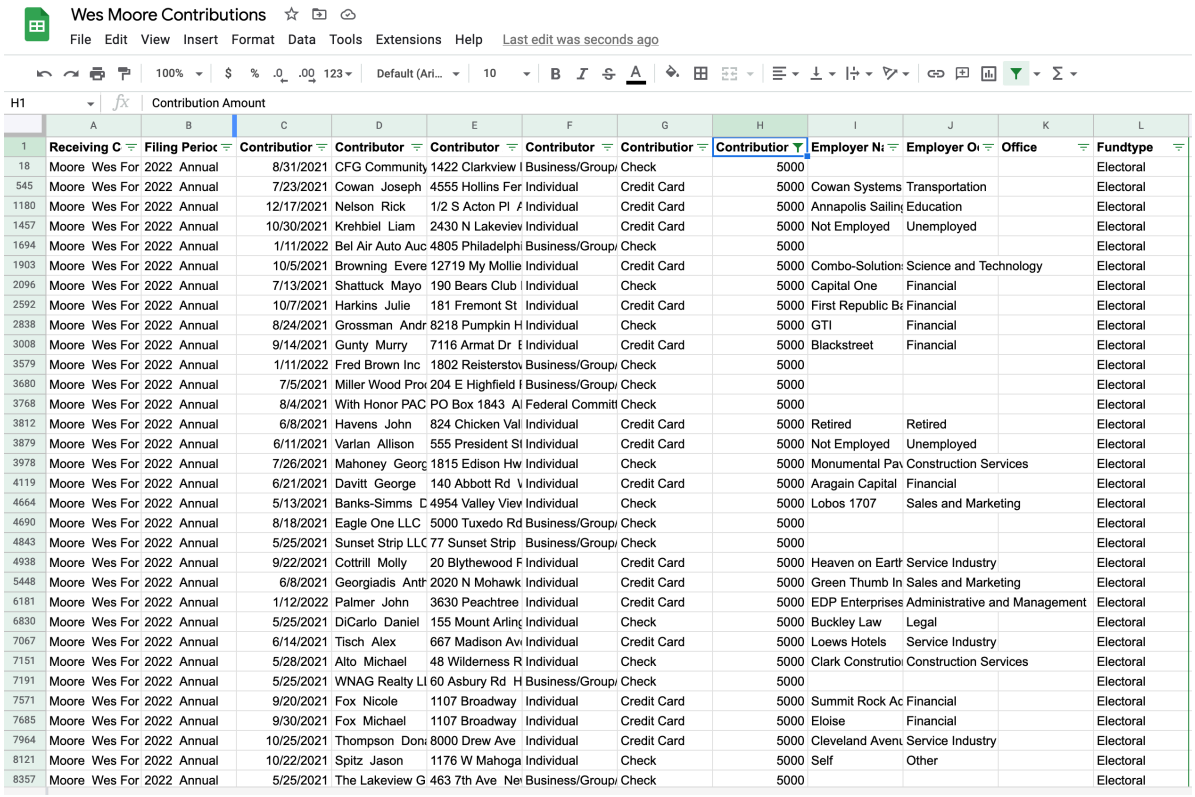
Filtering means picking out only some of the rows you want to see based on a criteria you select in a column. Think of it as casting a fishing net – the more filters you add, the fewer

fish will be caught.

To turn on filters in Google Sheets, go to Data -> Create a filter. It will add small down arrows to each column in the first row, another reason why headers are crucial. You can filter by multiple columns, and each filter you select adds more conditions, narrowing your net.

To find contributions of $5,000, use the drop-down menu under `Contribution Amount` to select it and hit "OK".

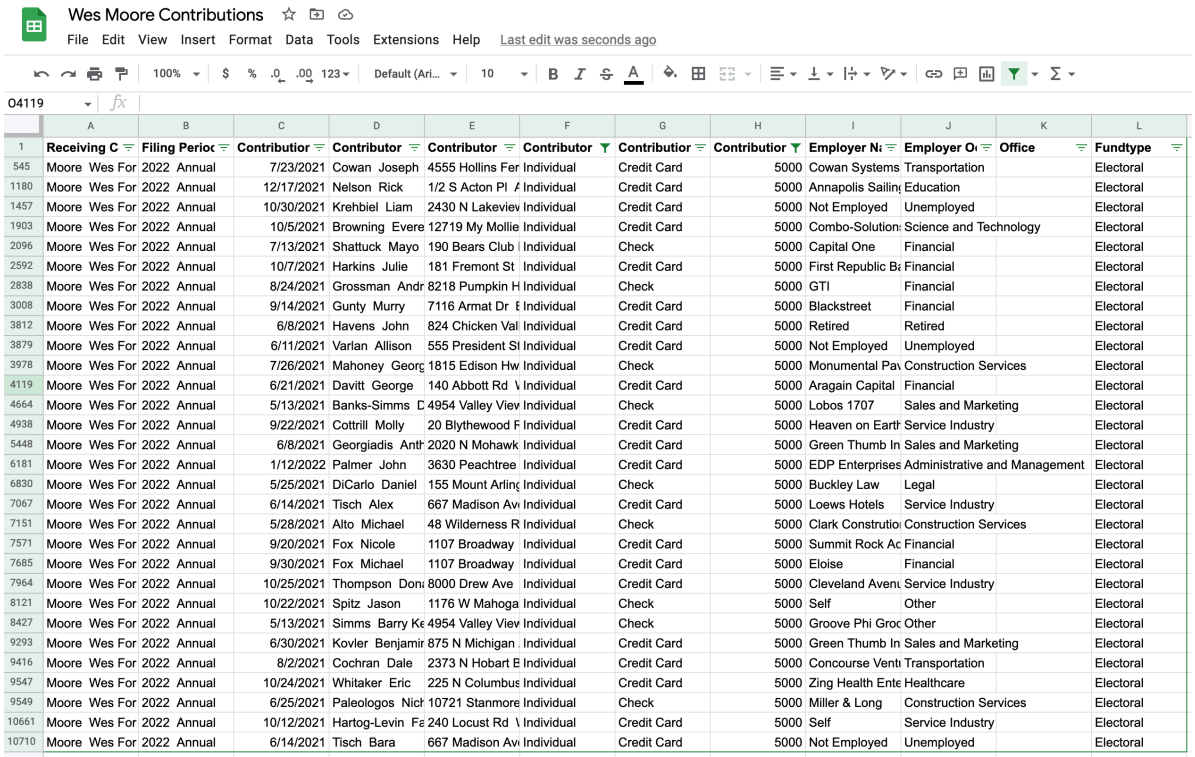
When you do this, notice that the drop-down arrow has turned into a solid green funnel and that any rows that don't match your filter are hidden.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Receiving C	Filing Perioc	Contribution	Contributor	Contributor	Contributor	Contributior	Contribution	Employer Na	Employer Oc	Office	Fundtype
18	Moore Wes For	2022 Annual	8/31/2021	CFG Community	1422 Clarkview I	Business/Group/	Check	5000				Electoral
545	Moore Wes For	2022 Annual	7/23/2021	Cowan Joseph	4555 Hollins Fer	Individual	Credit Card	5000	Cowan Systems	Transportation		Electoral
1180	Moore Wes For	2022 Annual	12/17/2021	Nelson Rick	1/2 S Acton Pl A	Individual	Credit Card	5000	Annapolis Sailin	Education		Electoral
1457	Moore Wes For	2022 Annual	10/30/2021	Krehbiel Liam	2430 N Lakeviev	Individual	Credit Card	5000	Not Employed	Unemployed		Electoral
1694	Moore Wes For	2022 Annual	1/11/2022	Bel Air Auto Auc	4805 Philadelphi	Business/Group/	Check	5000				Electoral
1903	Moore Wes For	2022 Annual	10/5/2021	Browning Evere	12719 My Mollie	Individual	Credit Card	5000	Combo-Solution	Science and Technology		Electoral
2096	Moore Wes For	2022 Annual	7/13/2021	Shattuck Mayo	190 Bears Club I	Individual	Check	5000	Capital One	Financial		Electoral
2592	Moore Wes For	2022 Annual	10/7/2021	Harkins Julie	181 Fremont St	Individual	Credit Card	5000	First Republic Ba	Financial		Electoral
2838	Moore Wes For	2022 Annual	8/24/2021	Grossman Andr	8218 Pumpkin H	Individual	Check	5000	GTI	Financial		Electoral
3008	Moore Wes For	2022 Annual	9/14/2021	Gunty Murry	7116 Armat Dr E	Individual	Credit Card	5000	Blackstreet	Financial		Electoral
3579	Moore Wes For	2022 Annual	1/11/2022	Fred Brown Inc	1802 Reisterstov	Business/Group/	Check	5000				Electoral
3680	Moore Wes For	2022 Annual	7/5/2021	Miller Wood Proc	204 E Highfield I	Business/Group/	Check	5000				Electoral
3768	Moore Wes For	2022 Annual	8/4/2021	With Honor PAC	PO Box 1843 Al	Federal Committ	Check	5000				Electoral
3812	Moore Wes For	2022 Annual	6/8/2021	Havens John	824 Chicken Val	Individual	Credit Card	5000	Retired	Retired		Electoral
3879	Moore Wes For	2022 Annual	6/11/2021	Varlan Allison	555 President St	Individual	Credit Card	5000	Not Employed	Unemployed		Electoral
3978	Moore Wes For	2022 Annual	7/26/2021	Mahoney Georg	1815 Edison Hw	Individual	Check	5000	Monumental Pav	Construction Services		Electoral
4119	Moore Wes For	2022 Annual	6/21/2021	Davitt George	140 Abbott Rd V	Individual	Credit Card	5000	Aragain Capital	Financial		Electoral
4664	Moore Wes For	2022 Annual	5/13/2021	Banks-Simms D	4954 Valley Viev	Individual	Check	5000	Lobos 1707	Sales and Marketing		Electoral
4690	Moore Wes For	2022 Annual	8/18/2021	Eagle One LLC	5000 Tuxedo Rd	Business/Group/	Check	5000				Electoral
4843	Moore Wes For	2022 Annual	5/25/2021	Sunset Strip LLC	77 Sunset Strip	Business/Group/	Check	5000				Electoral
4938	Moore Wes For	2022 Annual	9/22/2021	Cottrill Molly	20 Blythewood F	Individual	Credit Card	5000	Heaven on Earth	Service Industry		Electoral
5448	Moore Wes For	2022 Annual	6/8/2021	Georgiadis Anth	2020 N Mohawk	Individual	Credit Card	5000	Green Thumb In	Sales and Marketing		Electoral
6181	Moore Wes For	2022 Annual	1/12/2022	Palmer John	3630 Peachtree	Individual	Credit Card	5000	EDP Enterprises	Administrative and Management		Electoral
6830	Moore Wes For	2022 Annual	5/25/2021	DiCarlo Daniel	155 Mount Arling	Individual	Check	5000	Buckley Law	Legal		Electoral
7067	Moore Wes For	2022 Annual	6/14/2021	Tisch Alex	667 Madison Av	Individual	Credit Card	5000	Loews Hotels	Service Industry		Electoral
7151	Moore Wes For	2022 Annual	5/28/2021	Alto Michael	48 Wilderness R	Individual	Check	5000	Clark Construtio	Construction Services		Electoral
7191	Moore Wes For	2022 Annual	5/25/2021	WNAG Realty LL	60 Asbury Rd H	Business/Group/	Check	5000				Electoral
7571	Moore Wes For	2022 Annual	9/20/2021	Fox Nicole	1107 Broadway	Individual	Credit Card	5000	Summit Rock Ac	Financial		Electoral
7685	Moore Wes For	2022 Annual	9/30/2021	Fox Michael	1107 Broadway	Individual	Credit Card	5000	Eloise	Financial		Electoral
7964	Moore Wes For	2022 Annual	10/25/2021	Thompson Dona	8000 Drew Ave	Individual	Credit Card	5000	Cleveland Aven	Service Industry		Electoral
8121	Moore Wes For	2022 Annual	10/22/2021	Spitz Jason	1176 W Mahoga	Individual	Check	5000	Self	Other		Electoral
8357	Moore Wes For	2022 Annual	5/25/2021	The Lakeview G	463 7th Ave Ne	Business/Group/	Check	5000				Electoral

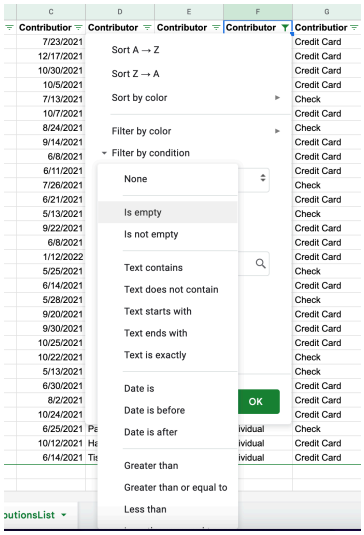
This method works for small-ish and simple-ish columns. If your column has more than 10,000 different entries, such as names or addresses, only the first 10,000 will be considered. We only caught these for stories when someone did a fact-check using a different method of filtering. If your column has a lot of distinct entries, use option that says "Choose One", and then use the "Contains" option. Better yet, don't use filtering for counting things at all.

Add more filters to narrow down your list of cases even more. For example, if you wanted to see $5,000 contributions from individuals, you would choose "Individual" under Contributor Type:



Different kinds of filters

There are several options under the filter drop-down menu under “Filter by condition”, and you also can type values into a search box to try and filter that way (the latter option is best for text columns). There also is a “Filter by color” option. My opinion: don’t do this.



FAQ

How do I turn off all of my filters

In the data tab, chose "Remove filter" to remove all of the filters.

Where is the button to filter *columns*?

Sometimes you don't want to see all of your columns – there are too many and they're getting confusing. There is no column filter in Sheets or Excel (You'll see how to filter, or "Select", columns from a dataset in R later.)

Instead, you can hide the columns you don't want to see. When columns and rows are hidden, they generally won't copy to a new sheet.

8 Formulas in Google Sheets

The quick review of math in Google Sheets uses the City of Phoenix's budgeted spending for the 2018 fiscal year, compared with previous years.

Make a copy of the data file to follow along

You should get into the habit of creating unique identifiers, checking your corners and looking for documentation before you ever start working with a spreadsheet. These habits were covered in Data journalism in the age of replication and on a Google Sheets refresher.

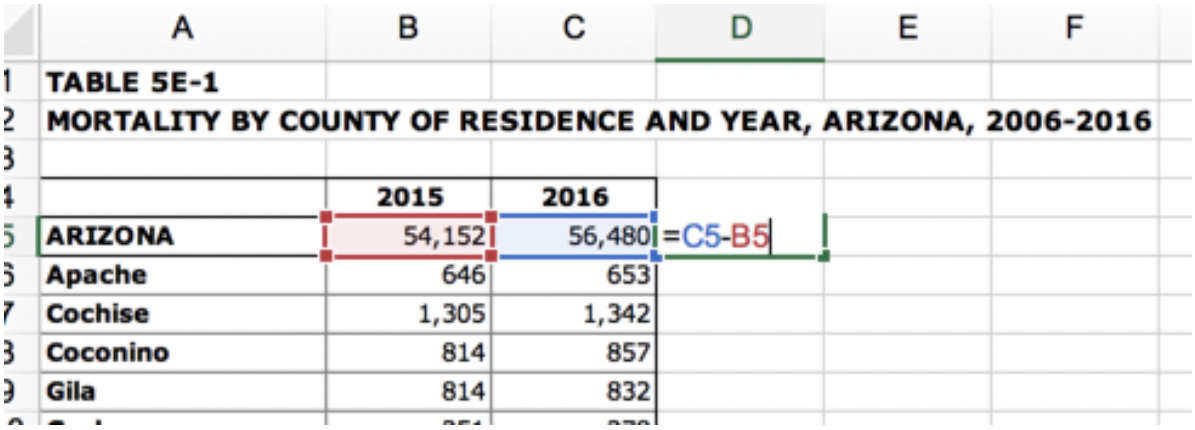
8.1 Formulas in spreadsheets

Whether you use Google sheets or Excel, remember that every formula begins with the equals sign (=). Rather than the values you want to work with in the formula, you'll use *references* to other cells in the sheet.

The easiest formulas are simple arithmetic: adding, subtracting, multiplying and dividing two or more cells. You'll just use simple operators to do this:

operator	symbol	example
addition	+	=A2+B2
subtraction	-	=A2-B2
multiplication	*	=A2*B2
division	/	=A2/B2

Here's what a spreadsheet looks like while editing some simple arithmetic:



	A	B	C	D	E	F
1	TABLE 5E-1					
2	MORTALITY BY COUNTY OF RESIDENCE AND YEAR, ARIZONA, 2006-2016					
3						
4		2015	2016			
5	ARIZONA	54,152	56,480	=C5-B5		
6	Apache	646	653			
7	Cochise	1,305	1,342			
8	Coconino	814	857			
9	Gila	814	832			

Figure 8.1: formula

The other kind of formula is a *function*. A function is a command that has a name, and requires *arguments* – usually the cell addresses or the range of addresses that it will act on. Every programming language has functions built in and many have extensions, or packages or libraries, that add even more as users find things they want to do more efficiently. You begin using a function the same way you begin a formula – with an = sign. Here are three common functions that create summary statistics for the numbers contained in a *range* of addresses. A range is a set of cells defined by its corner cell address: the top left through the bottom right.

You'll usually use them on a single column at a time.

Formula	What it does
=SUM(start:finish)	Adds up the numbers between `start` and `finish`
=AVERAGE(start:finish)	Computes the mean of the numbers
=MEDIAN(start:finish)	Derives the median of the numbers

…where "start" means the first cell you want to include, and finish means the last cell. Use the cell address of the first number you want to include , a colon, then the cell address of the last number you want to include. You can also select them while you're editing the formula.

Here's an example of adding up all of the rows in a list by county:

	A	B	C
1		year_2015	year_2016
2	Apache	646	653
3	Cochise	1,305	1,342
4	Coconino	814	857
5	Gila	814	832
6	Graham	251	278
7	Greenlee	69	52
8	La Paz	254	270
9	Maricopa	28,945	30,311
10	Mohave	3,024	3,181
11	Navajo	907	1,010
12	Pima	9,241	9,492
13	Pinal	2,968	2,991
14	Santa Cruz	294	301
15	Yavapai	2,918	2,955
16	Yuma	1,427	1,506
17	Unknown	275	449
18			
19		=SUM(B2:B17)	

Figure 8.2: formula