Data Journalism with R and the Tidyverse

Code, data and visuals for storytellers

Matt Waite & Sarah Cohen (original authors); updated by Sean Mussenden, Rob We

8/29/2022

Table of contents

1 Introduction

Welcome to data journalism. The main goal of this course is to expand your ability to report and tell stories using data. You will use these tools to discover trends in data, like what Rachell Sanchez-Smith found with the sharp jump in COVID-19 cases in Arkansas children. You will learn how to create and publish graphics and maps. It's hard work but it is a lot of fun and very rewarding.

We have some basic goals for you to reach in this class. By the end of the semester, we want you to have basic proficiency and independence with data analysis. We want you to be able to write about data clearly, using the Associated Press style as a benchmark. We want you to be able to find and download a dataset, clean it up, visualize it.

The skills you will learn in the coming weeks are in high demand in journalism and beyond. Examine this BuzzFeed job description from 2017:

"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.

"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."

You will learn these skills in this book. You'll get a taste of modern data journalism through Google Sheets and 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. Combining them together has the power to change policy, expose injustice and deeply inform.

1.1 Installations

This book begins with a basic review of Google Sheets and then shifts to the R statistical language. To follow along, you'll do the following:

1. Install the R language on your computer. Go to the this website, click download R based on your operating system. If that link somehow doesn't work, check R Project website and find a different location.
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 common software packages in your R Studio.

1.2 About this book

This book is the collection of class materials compiled by various data journalism professors around the country: Matt Waite at the University of Nebraska-Lincoln's College of Journalism and Mass Communications and Sarah Cohen of Arizona State University. This version was edited by Derek Willis, Sean Mussenden and Rob Wells at the University of Maryland Philip Merrill College of Journalism.

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, Sean Mussenden and Rob Wells 2022.
- 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!

1.3 What we'll cover

- Spreadsheets
- R Basics
- Replication, Data Diary
- Data basics and structures
- Aggregates
- Mutating

- Working with dates
- Filters
- Data cleaning techniques, Janitor
- 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
- 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.

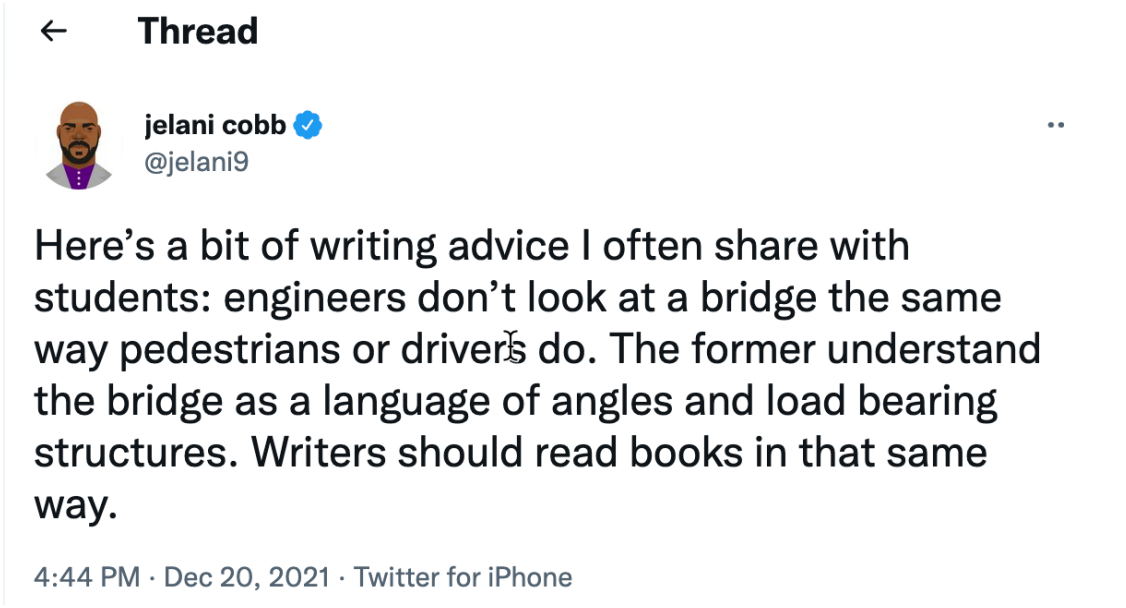


Figure 2.1: jelani cobb

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.

Sarah Cohen at Arizona State University analyzed the Paycheck Protection Program loan data for ProPublica and found a handful of sketchy-looking records 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.

3 Newsroom math

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, an except from Sarah Cohen's outstanding book, Numbers in the Newsroom. It's worth getting your own copy if you don't already have one.

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.
- A viral Twitter thread:

4 Census Data

We will be using data from the U.S. Census for many of these exercises. Here's a quick rundown of the origins and inner-workings of this important dataset.

Each decade, the Census Bureau counts every person living in the United States and the five U.S. territories. This is known as the Decennial Census and it is used to apportion seats in the U.S. House of Representatives, among other things.

In addition, the Census Bureau conducts ongoing survey of communities, known as the American Community Survey or ACS that provides more timely information about social, economic, housing, and demographic data every year. You can find information on housing, small business ownership, population profiles and much more. The ACS uses an annual sample size of about 3.5 million addresses and is collecting information daily.

Congressional lawmakers look at the ACS data to determine distribution of federal spending, among other things. Read more about census data here

Here's something important to know about ACS results: Data are pooled across a calendar year. So the ACS numbers reflect data collected over a period of time. By contrast, the Decennial Census is a single point-in-time count of the population.

We will be using ACS data to examine trends in wealth in Baltimore neighborhoods because this survey provides detail not available yet in the Decennial Census.

You can access Census Data through multiple ways. The U.S. Census Bureau offers data.census.gov. Using this site requires some training and patience. Here is a good place to start, a presentation the Census Bureau staff made to the Investigative Reporters and Editors conference in 2019.

Another useful resource is censusreporter.org, a site not affiliated with the Census Bureau that's designed to make it easier to navigate and retrieve the ACS data.

When we use R, we will use a software library called tidycensus that makes it very easy to retrieve Census data from the Census API, or application programming interface, basically a raw data feed optimized for R, python and similar programs. Stay tuned on that later this semester.

Data journalist Paul Overberg, now with The Wall Street Journal, compiled this useful guide about terminology when dealing with Census data.

5 Spreadsheets

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, which allows students to do the exercises regardless of their computer operating system, Mac, Windows or Linux. The exercises can be easily adapted to Microsoft Excel, which can handle larger datasets and has more options for pivot tables and other more advanced functions. However, Google Sheets have their own advanced functions for scraping websites or importing non-tabular file formats like JSON.

Most of the screen shots and instructions are created with a MacOS Monterey. Some come from earlier Mac versions, but are largely the same now. Windows users should replace any instructions for using the CMD- key with the CTL- key. There is a table that compares keystrokes for Apple desktops, laptops and Windows machines for Excel at the bottom of Spreadsheet Refresher

Tutorials

Spreadsheets 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.

- Spreadsheet 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
- Spreadsheet Formulas: Percents, sums, and other basic computations used in newsrooms.

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 Spreadsheet Refresher

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

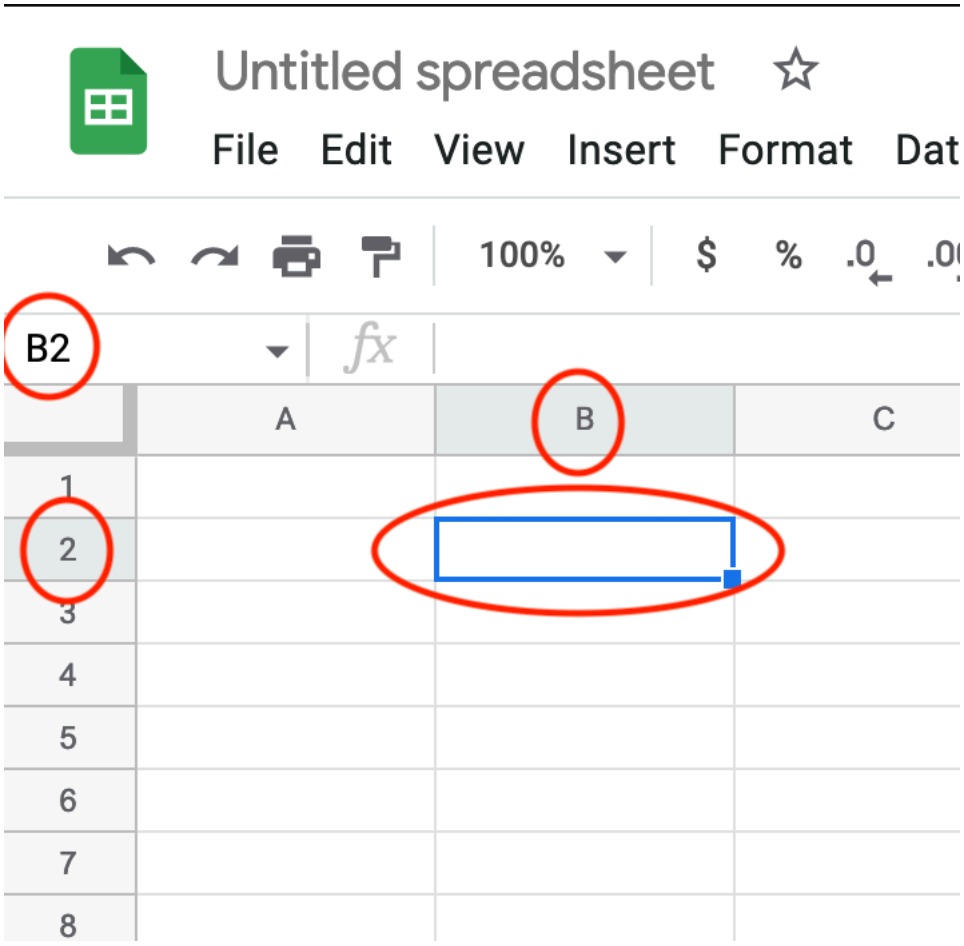
- To create original databases of events for sorting, filtering and counting. Examples include a long-running court case; the details of each opioid death in a city; a list of police shootings and their documents; or even a list of your own public records requests or contact log.
- To use data created 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 will Google Sheets since the program is available to anyone regardless of operating system. Google Sheets are easy to share for reporting teams.

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 Excel from the ground up

The spreadsheet grid

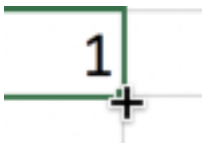


::: {.pl-2rem .float-right .col-md-4}

When you start up a spreadsheet, 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.

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. 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.

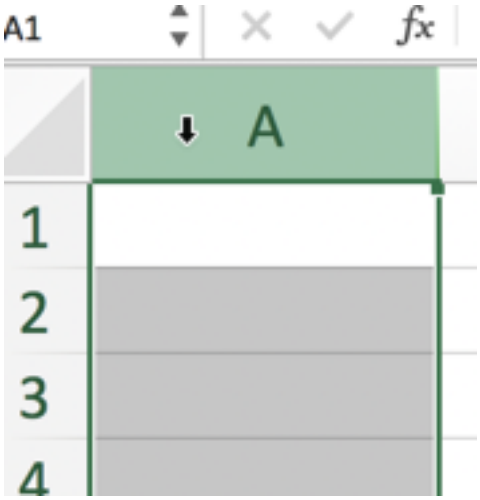
Selecting cells and ranges

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 cursor over the cell and click 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:C6, with the ":" referring to the word "through".

To select a group of cells and act on them all at once: Hover the cursor 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

Reading the screen

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

This image shows some key areas on the screen when you're just viewing the sheet:

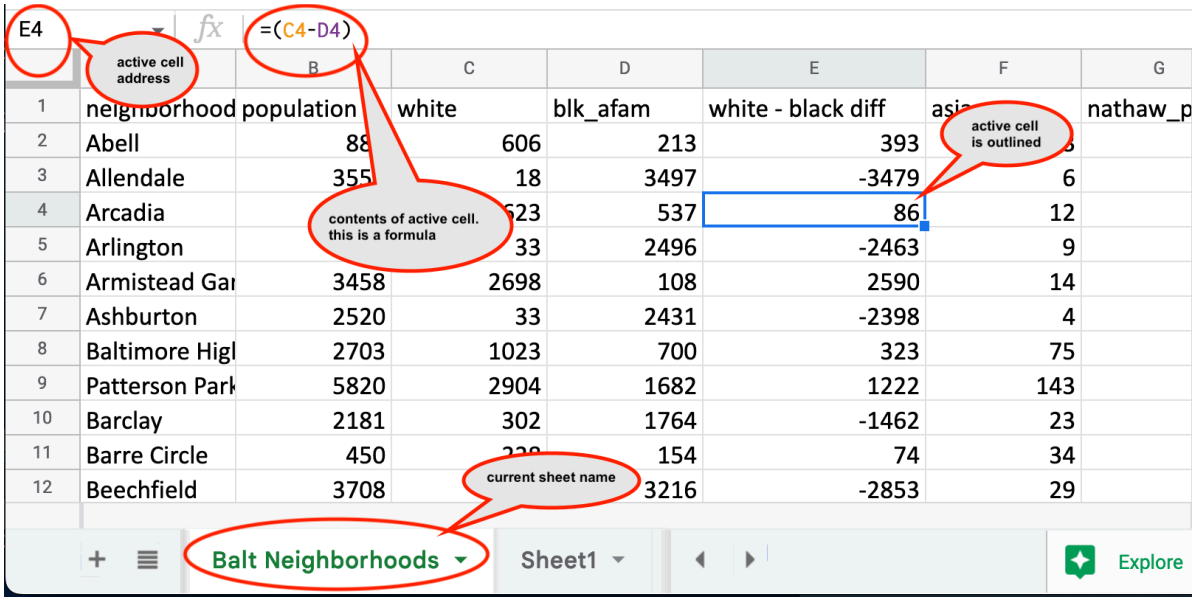
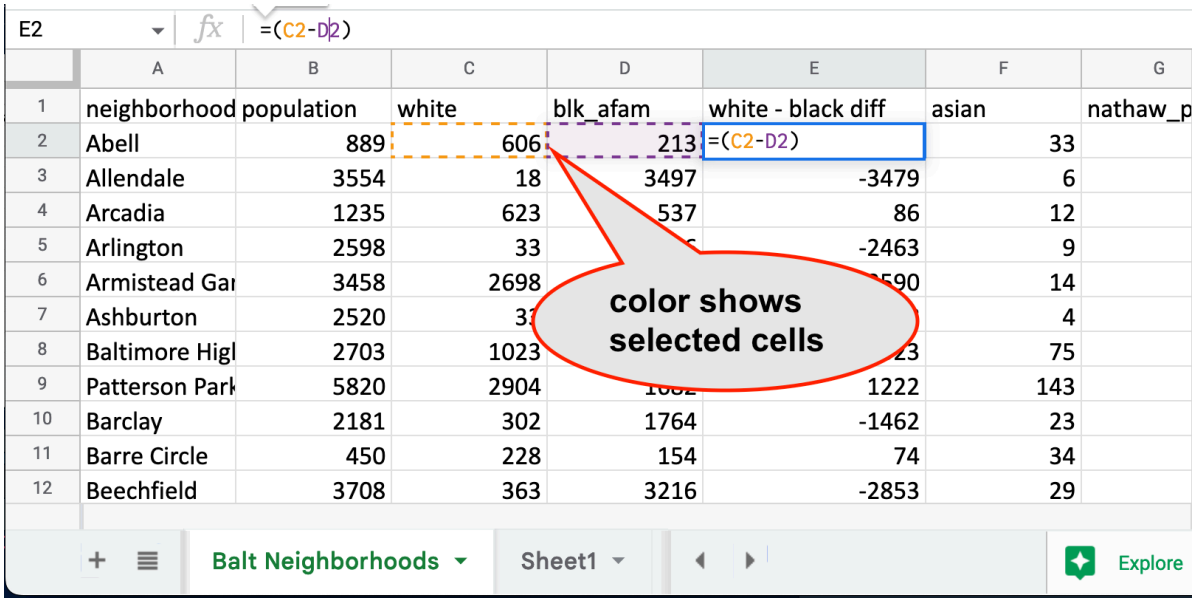


Figure 6.1: ready

This is how it changes when you're editing



	A	B	C	D	E	F	G
1	neighborhood	population	white	blk_afam	white - black diff	asian	nathaw_p
2	Abell	889	606	213	=(C2-D2)	33	
3	Allendale	3554	18	3497	-3479	6	
4	Arcadia	1235	623	537	86	12	
5	Arlington	2598	33		-2463	9	
6	Armistead Gar	3458	2698		590	14	
7	Ashburton	2520	3			4	
8	Baltimore Higl	2703	1023		23	75	
9	Patterson Park	5820	2904		1222	143	
10	Barclay	2181	302	1764	-1462	23	
11	Barre Circle	450	228	154	74	34	
12	Beechfield	3708	363	3216	-2853	29	

Figure 6.2: editing

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.

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.

Select View in the menu, then Freeze, then the number of rows to freeze. Select 1 row. As you now scroll down the sheet, your headings will remain but you can see the data as you move deeper into the sheet.

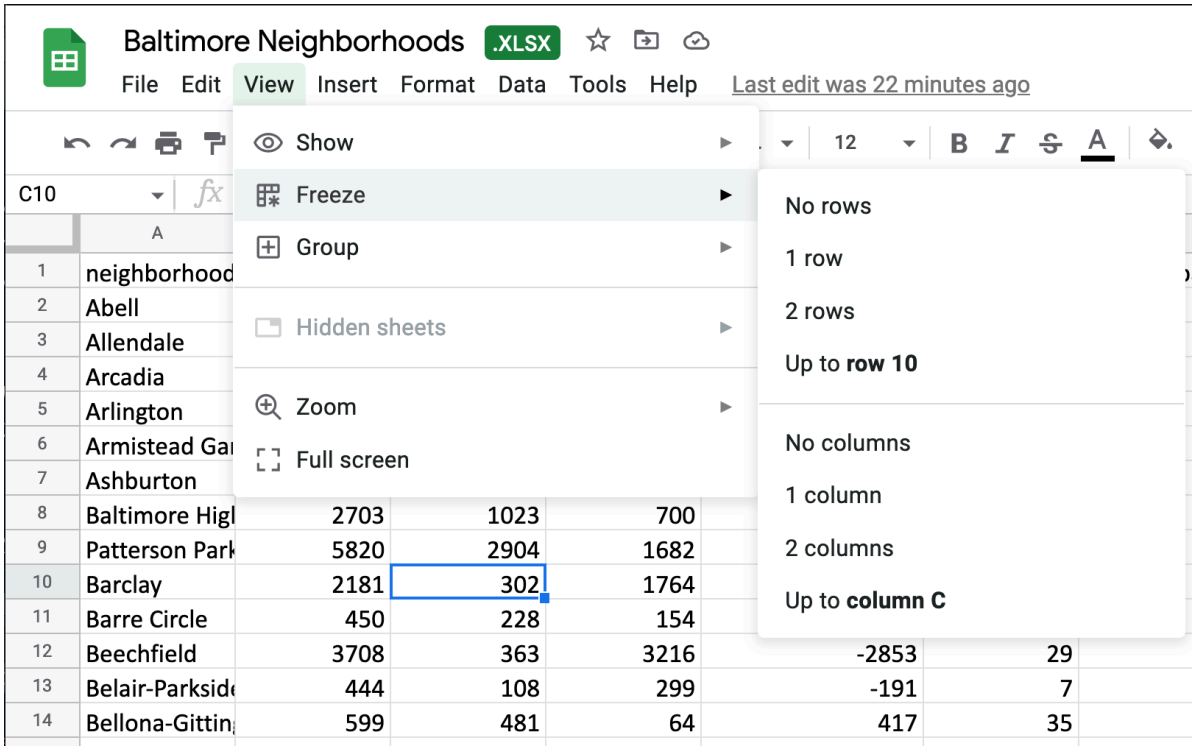


Figure 6.3: freeze panes

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 Format, then Wrapping from the menu to 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.

First steps

- Document where you got the spreadsheet and how you can get back to the original. Create a new tab (click the + sign in the lower left), name it Data Dictionary, copy the URL of your source data and any other notes about it. Make this your regular practice. It will save time and stress on deadline.
- 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.

Interview your data

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.

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.

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.

6.3 Video walkthrough

These first steps, along with adding an ID row, are shown here. You can follow along with the same dataset.

Getting started with Google Sheets

[Getting started with Google Sheets](https://www.youtube.com/embed/1hGoYzmkhfc)

6.4 Keyboard shortcuts

Google Sheets keyboard shortcuts can be found in the menu: Help, then Keyboard Shortcuts.

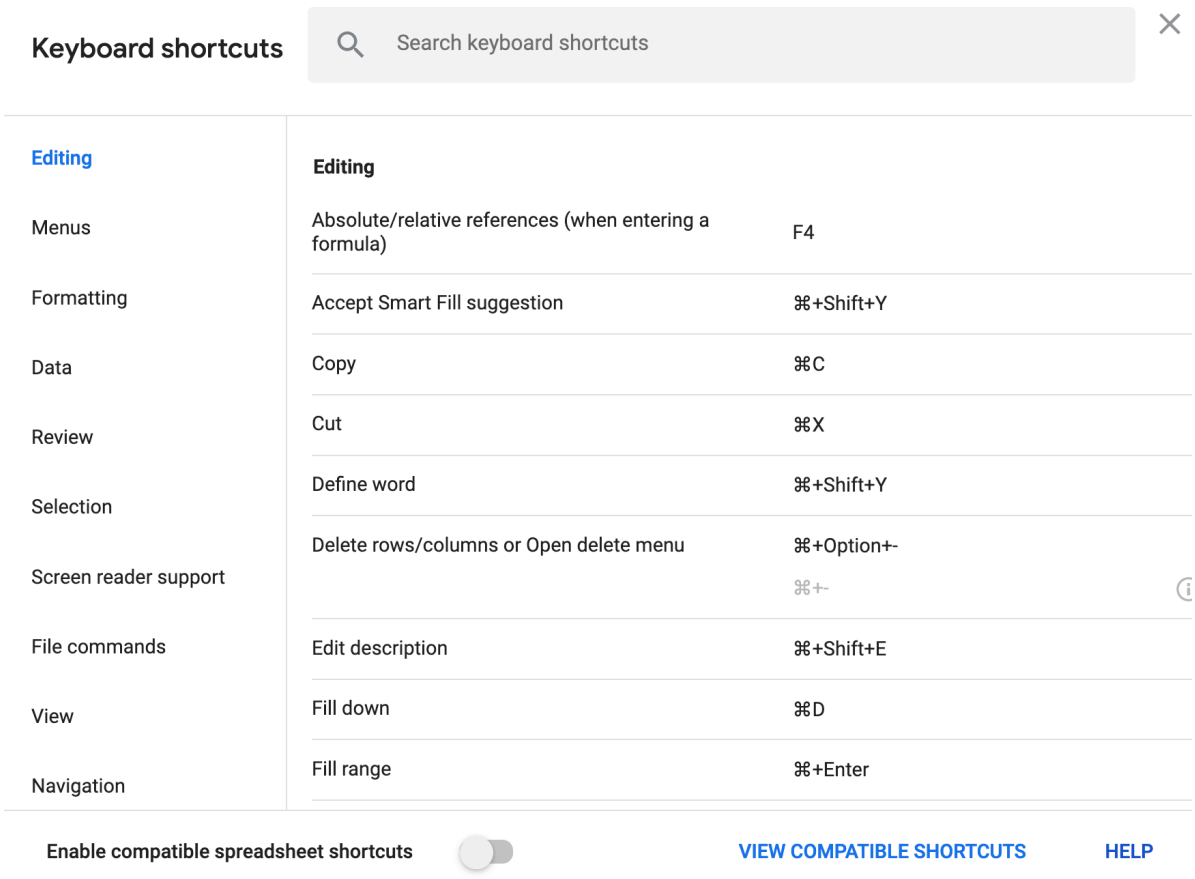


Figure 6.4: Keyboard shortcuts

7 Sorting and filtering to find stories

7.1 A sorting miracle

After police in Ferguson, Mo., killed Michael Brown in 2014, advocates and journalists began examining the racial and ethnic gaps between police departments and the communities they served.

The New York Times found a 7-year-old survey conducted by the Justice Department that allowed it to compare the data for major cities in a standalone graphic that it published later that year.

When newer data reflecting departments' makeup in 2012 was released a year later, Matt Apuzzo and Sarah Cohen hoped it would show some differences. It didn't. So we were left trying to find news in the data that was clearly of public interest.

Cohen matched up the demographics of police departments with their cities and then started sorting, filtering and Googling. Could there be news in the outliers on the list? Which departments most closely represented their communities? Which ones had unusually large gaps?

Cohen quickly stumbled on telling anecdote to frame the story: Inkster, Mich. had one of the least representative departments in the country, and had recently hired a new police chief to help mend the department's fraught relationship with its largely African-American community. Where had he come from? Selma, Ala., one of the most representative police departments in the nation. Interviews with the chief, William T. Riley III, suggested one reason for some cities' disparities: there was no state or federal money to pay for training new police officers.

The story, "Police Chiefs, Looking to Diversity Forces, Face Structural Hurdles" helped explain the persistent gap between the makeup of police in some areas and the communities they served.

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.

Figure 7.1: Chief William T. Riley III. Credit: Laura McDermott for The New York Times

- 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

Data from the Washington Post police shootings database for use in this tutorial - Documentation from the Post's github site :::

- The data for this and several other chapters is the Washington Post's public data collection of police shootings in the U.S. It includes the nation's best guess about each fatal police shooting since 2015. There are a couple of caveats:
- It excludes deadly police interactions other than shooting a firarem at the suspect. Any strangulation, car crashes, Tasers without guns or other methods are excluded.
- It is based primarily on news reports and the results public records requests so it often contains the story as told by police. We know that many of those reports are sugar-coated at best, and lies at worst.
- The Post says this is a list of fatal shootings, but doesn't say what happens if more than one person is killed. The 2019 shooting of D'Angelo Brown & Megan Rivera in West Memphis is shown as two rows[1] in the data even though it was one event. So each row might be considered a shooting "victim", a "suspect" or a shooting "fatality" rather than a "shooting".

The screenshots in this tutorial may not match exactly to what you get on the Washington Post data. This tutorial used data current to Aug. 3, 2022.

It's a good example set for us because it's been used as the basis of many stories, it has at least one of each *data type* that we plan to deal with in Google Sheets, and it is well documented on the Post's github site.

7.4 Get the data into Google Sheets

- Download the [police shooting data from the Washington Post](https://github.com/washingtonpost/data-police-shootings/releases/download/v0.1/fatal-police-shootings-data.csv
- Open Google Sheets. File | Import | Upload | Select the downloaded file "fatal-police-shootings-data.csv" . After it uploads, select the green "Import Data" button.

[1]Finding these is something that's pretty hard in a spreadsheet but will be really easy in R.

•

7.5 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 fatality 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, as we'll see in the computing section, these can be formatted to *look* like numbers we care about , 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 in Google Sheets. 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, that number is the number of days since Jan. 1, 1900.[2] They can also have time attached to them, which in Google Sheets is a fraction of a day. What this means is that the number 44,536.5 is really Dec. 6, 2021 at noon. In Google 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 7.2: date formats

[2]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

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. (Google Sheets uses the 1st of the month as a stand-in for an actual date when all you know is the month and year.)

7.5.1 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.

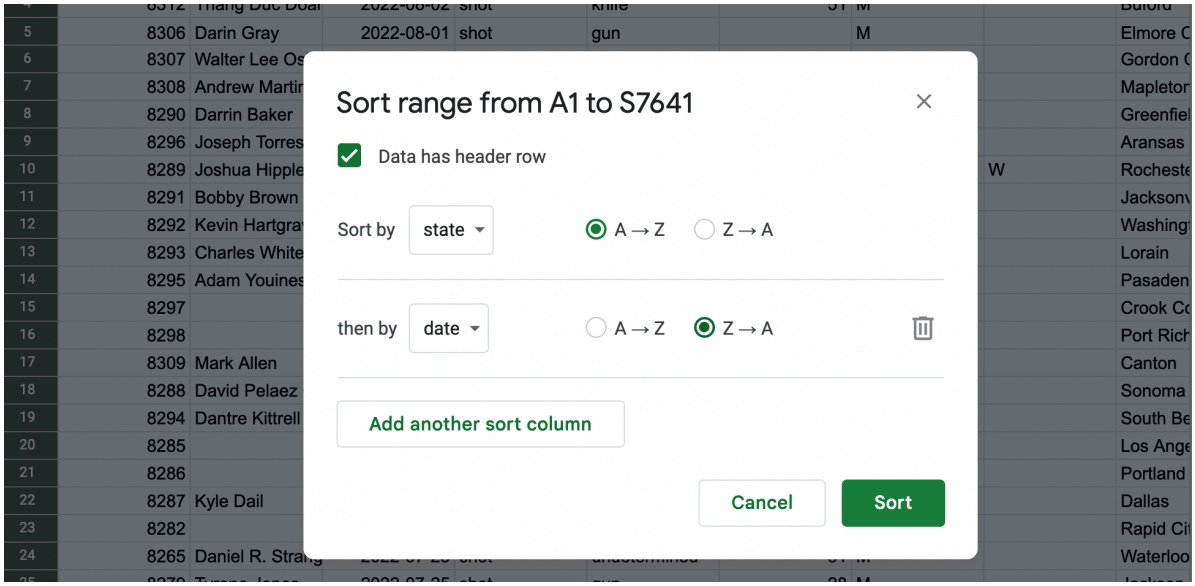
In Google Sheets, look for the sort options under the Data tab at the top of your screen. In this case, sorting from oldest to newest gives you a list of the fatalities in chronological order, including the time of day.

To sort your data:

- Make a copy of your data. Left click on the "fatal-police-shootings-data" tab, select Duplicate
- Select your data by clicking the box above Row 1 and to the left of Column A
- Select Data | Sort Range | Advanced Range Sorting Options
- Click "Data has header row" and then select date from the Sort by dialog box. Select Z –> A
- Select Sort

Adding fields to the sort

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



7.5.2 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 activate filters in Google Sheets, from the Menu:

- Data | Filter Views | Create a New Filter View
- You'll see little triangles next to the column headings.

Click the "armed" heading. You will see options for various weapons. All are selected by default with a check mark. To select just "ax", click on clear and then select "ax." The sheet how is filtered to just weapons using an ax. To remove the filter, repeat the steps and "select all" and the entire sheet is displayed again.

Each filter you select adds more conditions, narrowing your net.

To find fatalities that involved a firearm with a Taser, use the drop-down menu under `manner_of_death` select "shot and Tasered".

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.

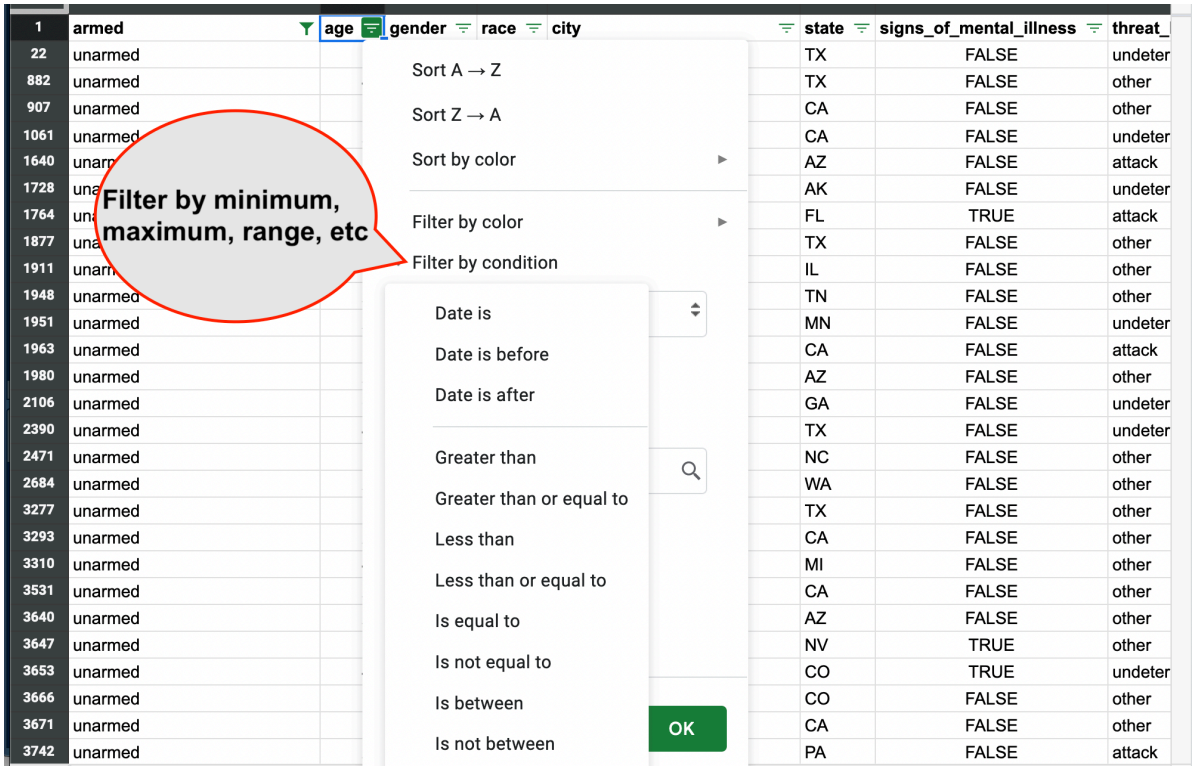
Add more filters to narrow down your list of cases even more. For example, the New York Times ran a series of stories in 2021 about unarmed people shot by police. One story was about those who were fleeing by car. Here's one way to get a preliminary list of those cases:

1. Remove any filter you already have on.
2. Turn on the filters again if you turned them off.
3. Choose "unarmed" under `armed` and "car" under `flee`.

(Of course, the Times didn't stop there in trying to find more cases and teasing out more of them from this and other data. But this is a start.)

Different kinds of filters

There are several filter options. You can filter by various conditions. For numerical data, you can set a minimum or maximum value or a range of values. This is useful for dates to specify a certain time period. For text, you can filter if a word contains a few letters, useful to capture spelling variations.



8 Spreadsheet Formulas

The quick review of math in Google Sheets uses the City of Baltimore's 2022 budget, compared with previous years.

Get the Google Sheet 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 Replication and the data diary and on an Excel refresher .

8.1 Formulas in spreadsheets

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.

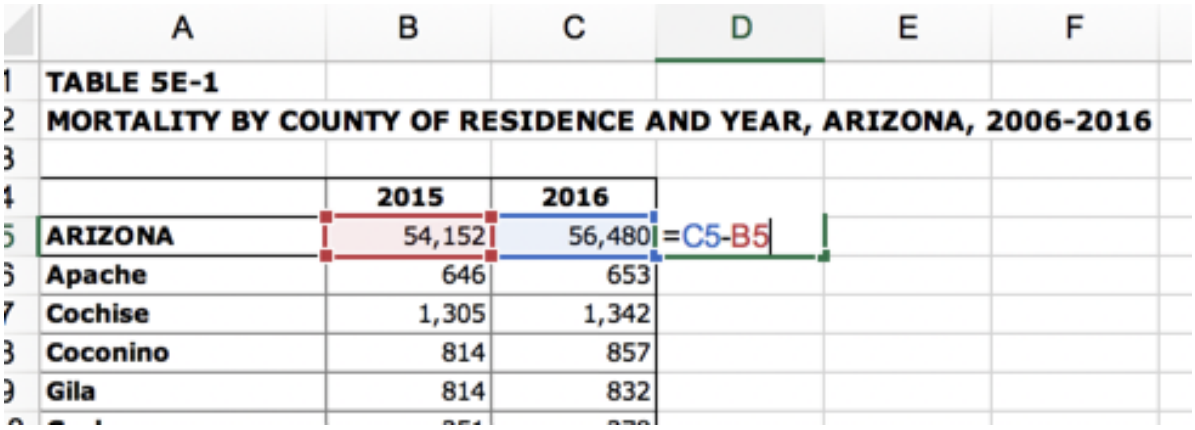
The easiest formulas are simple arithmetic: adding, subtracting, multiplying and dividing two or more cells. You'll just use simple operators to do this:

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:

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.



	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

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:

8.2 Common spreadsheet arithmetic

The budget document shows three years' of data: The actual spending in the fiscal year that ended in 2016; the spending that was estimated for the end of fiscal year 2017; and the proposed spending for fiscal year 2018. The first page of the document shows these amounts for broad spending categories.

You may want to widen the columns and format the numbers before you start:

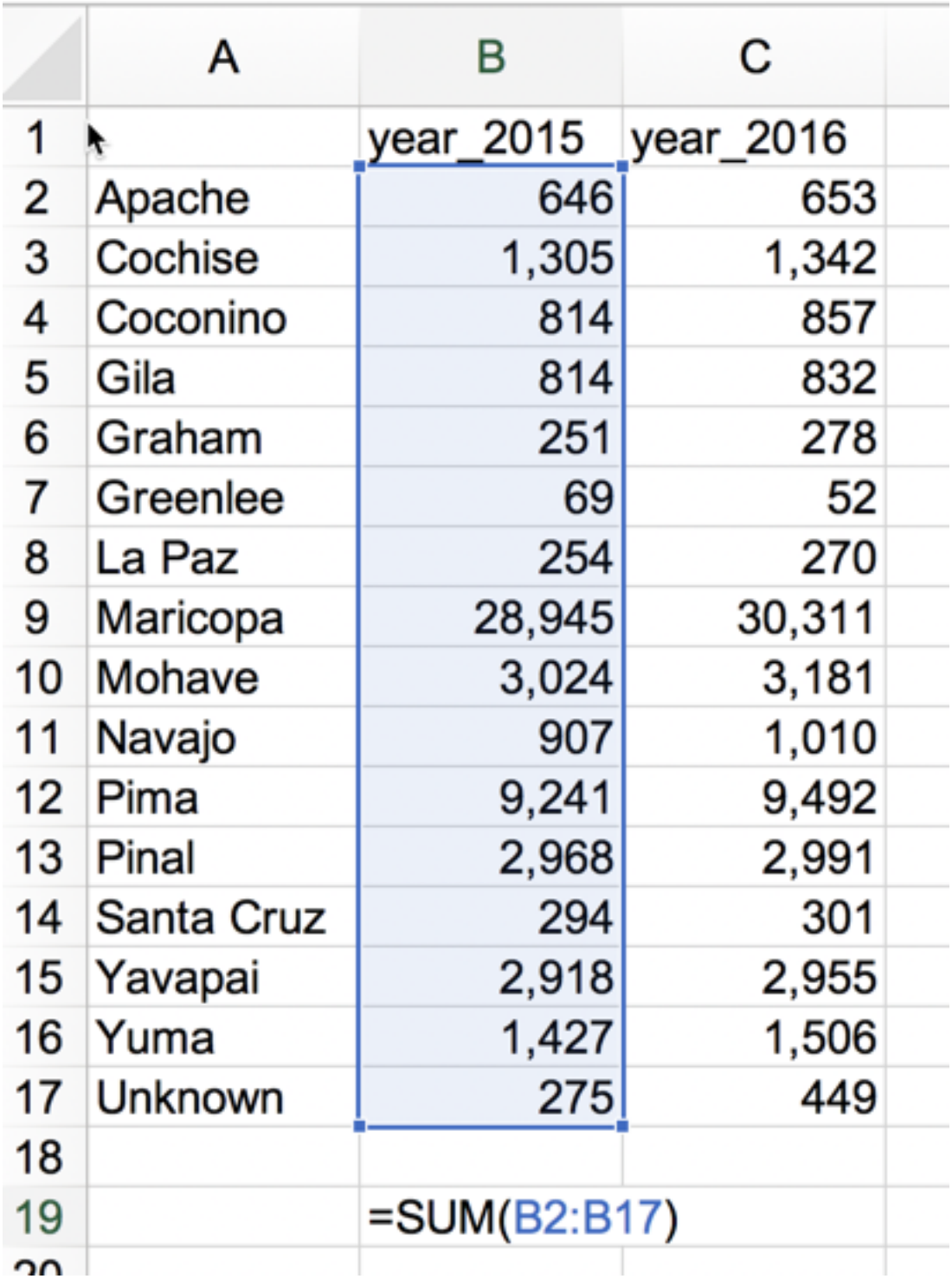
8.2.1 Check the government's math with SUM

Our first job is to make sure the government has provided us data that adds up. To do that, we'll SUM all of the departments' spending.

To add up the numbers from FY 2020 Actual, enter the following formula in cell C16, just below the number provided by the government:

```
=SUM(C2:C13)
and hit the enter key
```


Copy that formula to the right. Notice how the formula changes the addresses that it is using as you move to the right – it's adjusted them to refer to the current column.



	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

8.2.2 Change in spending

The increase or decrease in projected spending from 2017 to 2018 is just the difference between the two values, beginning in cell F3

```
new-old, or  =E2-D2
```

When you copy it down, note how the references to each row also adjusted. In line 3, it's E3-D3, and so on. Excel and other spreadsheets assume that, most of the time, you want these kinds of adjustments to be made.

8.2.3 Percent change

We can't tell the *rate* of growth for each department until we calculate the percent change from one year to another. Now that we already have the change, the percent change is easy. The formula is:

```
( new - old ) / old

.. or just scream "NOO"
```

The new-old is already in column F, so all that's left is to divide again. In grade school, you also had to move the decimal place over two spots, since the concept of percent change is "out of 100". Excel formats will do that for you.

Remember, it's always (new-old)/old , **NOT** the big one minus the little one. Doing it correctly, the answer could be negative, meaning the value fell.

B	C	D	E	F	G
	Act 2016	Est 2017	Budget 2018	change from 2017	percent chan
rnment	$103,679	$111,601	$129,653	$18,052	=F2/D2
	$874,560	$930,155	$1,032,609	$102,454	
ce	$53,367	$61,104	$63,272	$2,168	
on	$541,951	$605,421	$650,439	$45,018	
levelopment	$176,318	$207,237	$239,428	$32,191	
nrichment	$243,029	$259,893	$268,072	$8,179	
al services	$378,827	$429,960	$450,858	$20,898	

Figure 8.3: formula

When you're done, you can format the answer as a percentage to get it into whole numbers.

It's also worth comparing the picture you get by looking at raw numbers vs. percentages. It's instructive that federal grants are up 308%.

8.2.4 Parts of a whole: percent of total

We'd also like to know what portion of the total spending is eaten up by each department. To do that, we need the percent of total.

In our case, let's use the total that the government gave us. In practice, you'd have to decide what to do if your figures didn't match those provided by officials. You can't assume that the total is wrong – you could be missing a category, or there could be a mistake in one of the line items.

The formula for percent of total is:

```
category / total
```

Here's a good trick with spreadsheets when you need to divide against a fixed total. You don't have to type in each formula one by one, though. Instead, you'll use anchors, known in spreadsheets as "absolute references". Think of a dollar sign as an anchor or stickpin, holding down the location of part of your formula. If you put the stickpin before the letter in the formula, it holds the column in place. If you put it before the number, it holds the row in place. If you put it in both places, it holds the cell in place.

In this case, we want to see what percentage property taxes, income taxes, etc. are of the total revenues in FY22, which is $4,331,049,486 (cell E14). Let's figure it out.

- Left click column F, insert column to the left. Name it Pct of Total
- Create formula to divide property taxes into total revenues: =(E2/E14)
- Modify forumula so it will anchor to the E14 as you move down the spreadsheet =(E2/E14)
- Copy formula down to F13

8.3 While we're at it: two kinds of averages

Although it doesn't make a lot of sense in this context, we'll go ahead and calculate the *average* or *mean* size of each department, and then calculate the *median* size.

Simple average, or mean

A simple average, also known as the mean, is skewed toward very high or very low values. Its formula is

```
sum of pieces / # of pieces that were summed
```

But in Google Sheets, all we need is the word AVERAGE:

```
=AVERAGE(C2:C9)
```

Median

In Google Sheets, you can get the median of a list of numbers by just using the formula, MEDIAN()

```
= MEDIAN(C2:C9)
```

Doing simple calclutions like this on data that is provided to you by the government lets you ask better questions when you get an interview, and may even convince officials to talk with you. There's a big difference between asking them to tell you what the budget numbers are, and asking them to explain specific results!

8.4 FAQs

Should I use average or median?

It depends. Averages are easier to explain but can be misleading. Usually, if they're very different, median will be a better representation of the typical person, city or department. Averages in these cases are more like totals.

My percents are small numbers with decimal points

Use the format as a % button to move the decimal point over two places and insert the percentage symbol.

9 Grouping with pivot tables

In the wake of a police shooting in 2016, reporter Mitch Smith obtained a list of traffic stops from the St. Anthony Police Department in Minnesota. He was writing a story on Philandro Castile's death and was running out of time. He wanted to answer a simple question: Were minority motorists more likely to be stopped in St. Anthony than whites?

Rob Gebeloff made a quick pivot table to answer the question. That night, Smith wrote:

> In each of the three small suburbs patrolled by the St. Anthony police, less than 10 percent of the population is black. But data released by the city on Tuesday showed that a far higher percentage of the people ticketed or arrested by St. Anthony officers were African-American.
>
> Last year, around 19 percent of those cited by St. Anthony police were black, as were roughly 41 percent of people arrested by the department, a review of the city's data showed. Those percentages do not include the large number of defendants whose race was unknown.

Summarizing a list of items in a spreadsheet is done using pivot tables. In other languages, it's considered "aggregating" or "grouping and summarizing". Think of pivot tables and grouping as answering the questions, "How many?" and "How much?" They are particularly powerful when your question also has the words "the most" or the "the least" or "of each". Some examples:

- Which *Zip Code* had *the most* crimes?
- What *month* had *the least* total rainfall?
- *How much* did *each candidate* raise last quarter?
- In playing cards, *how many* of *each suit* do I have in my hand?
- On average, are *Cronkite students taller or shorter* than in other schools?

Confusing grouping with sorting or arranging

Many reporters confuse this summarization with "sorting". One reason is that this is how we express the concept in plain language: "I want to sort Skittles by color."

But in data analysis, sorting and and grouping are very different things. *Sorting* involves shuffling a table's rows into some order based on the values in a column. In other languages,

this is called *arranging* or *ordering*, much clearer concepts. *Grouping*, which is what pivot tables do, is a path to aggregating and computing summary statistics such as a count (the number of items), sum (how much they add up to), or average for category. It means "make piles and compute statistics for each one."

When to use filter vs. pivot tables

Something that trips up beginners is a desire to see details and totals at the same time, which is more difficult than it sounds.

A filter is used to *display* your selected items as a list. You'll get to see all of the detail and every column. As a convenience, Google Sheets shows you how many items are in that filtered list. That's great when you want to just look at them, or get more information about them. For instance, if you had a list of crimes by ZIP code, you might just want to see the list in your neighborhood – where, exactly, were they? When did they happen? Was it at night or the morning? What crimes happened on which blocks?

A pivot table is used when you *just want to see summaries* – does my ZIP code have more crime than others? Are robberies more common than car theft in my Zip code, and how does that compare to others?

In practice, you'll go back and forth between summary and detail. They're both important, just different.

9.1 Tutorial

::: {.alert .alert-info .opacity-2} We will continue to use data from the Washington Post police shootings database for this tutorial.

First, let's modify this spreadsheet to include the descriptions of race and ethnicities: A for Asian, B for Black, etc.

- Select Column I, city, and insert a new column to the left. Name it race_ethicity
- Create a filter. Select race, filter for A
- Type Asian in Column I. Copy Asian down the entire column so every A in column H corresponds with Asian in Column I
- Repeat: B = Black. H = Hispanic. W = White, non-Hispanic, N= Native American, O=Other, blanks=Unknown.

Follow this video to see the process.

Setting up the pivot table

From the main menu on Google Sheets, choose *Insert*, then *Pivot table*, then New sheet.

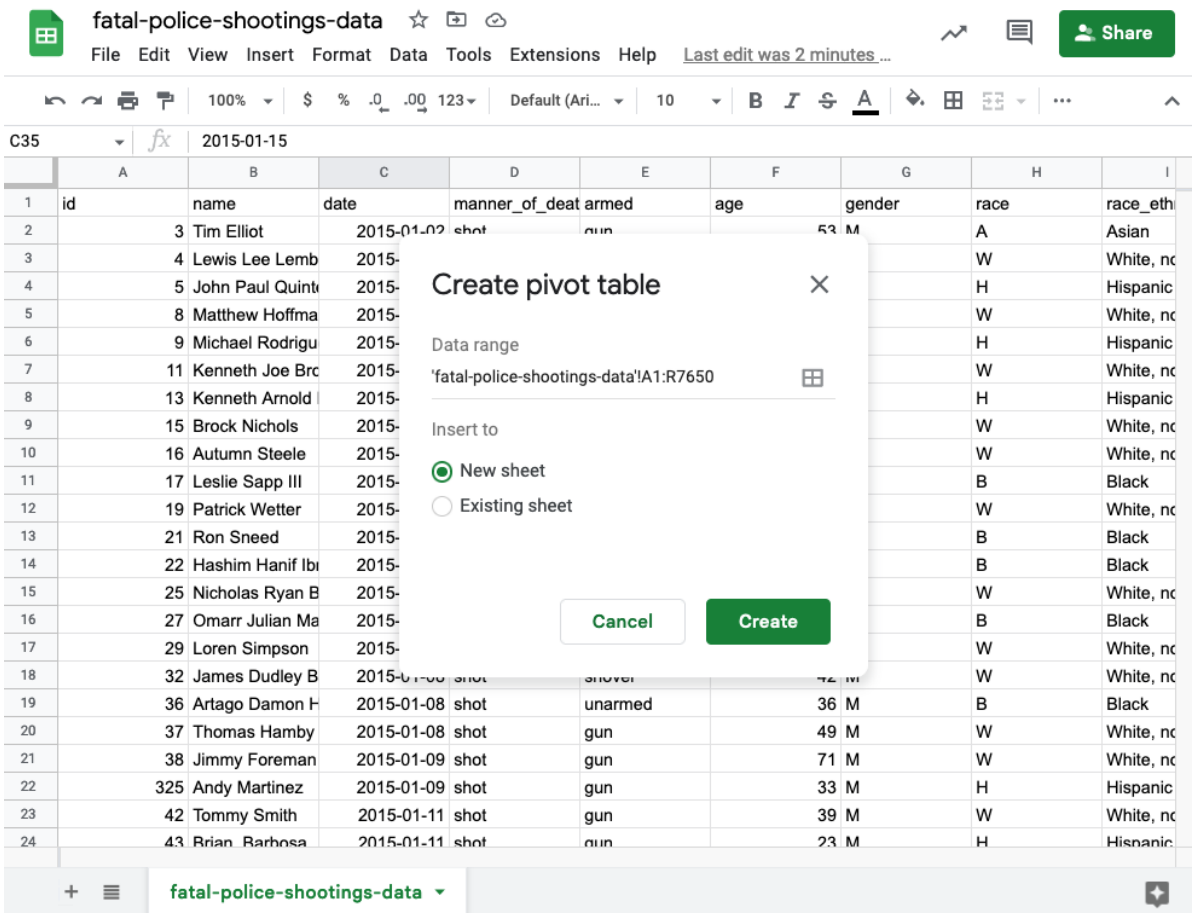


Figure 9.1: insert menu

Next, you will see the Pivot Table editor. Here's what it looks like:

Counting , or "how many"?

For Rows, select Add and then race_ethnicity. For values, select Add and then state. You will now see all of the race and ethnicity totaled.

We're totalling on state because it's good to have something that's always filled out into the Values area (`state` is a safe one in this data).

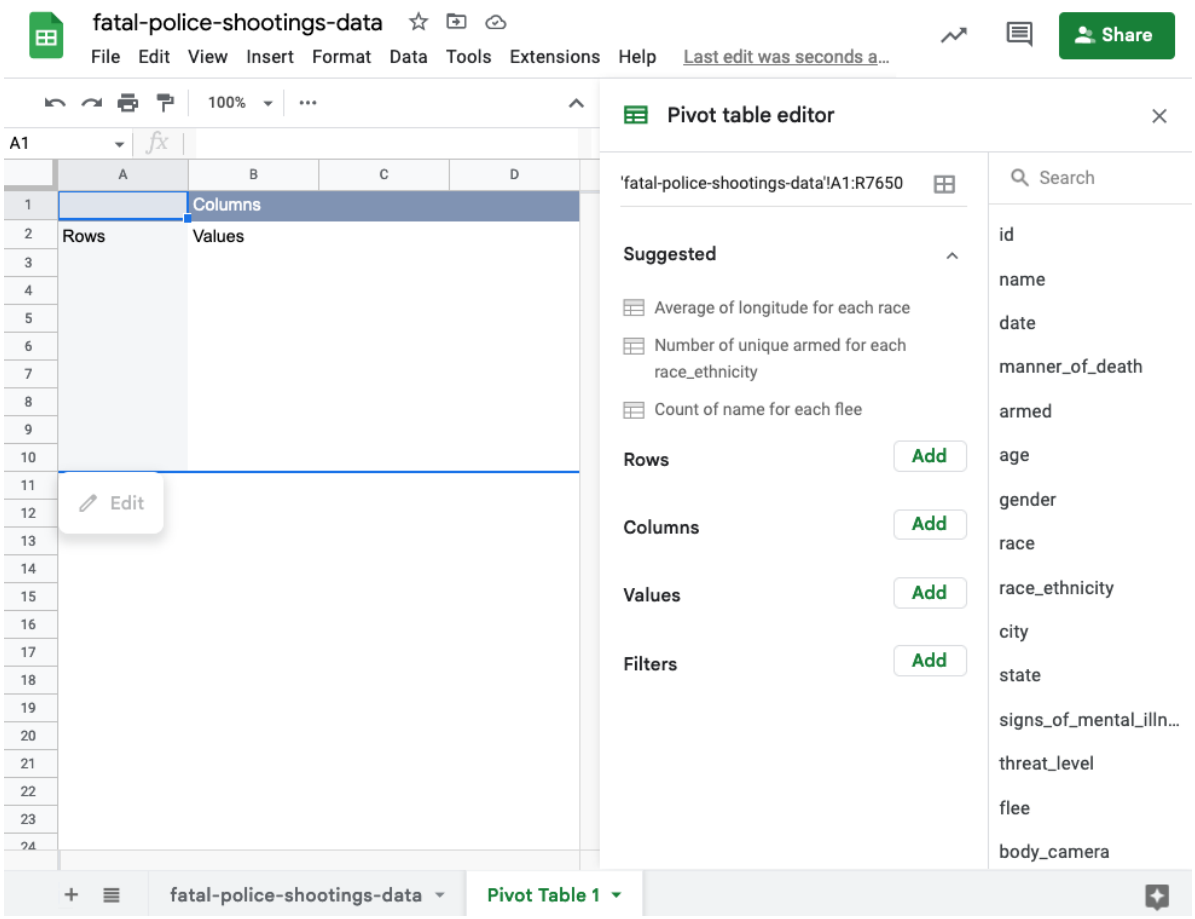
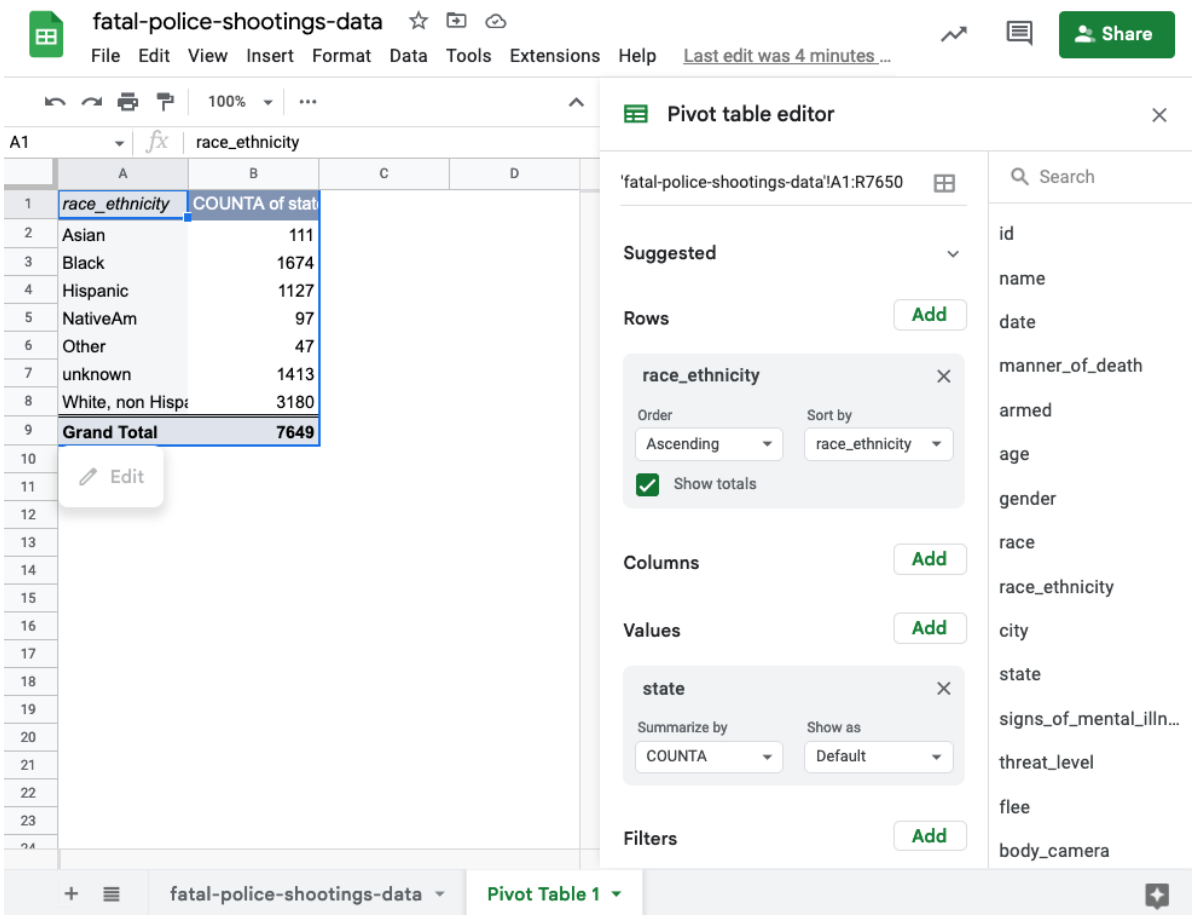


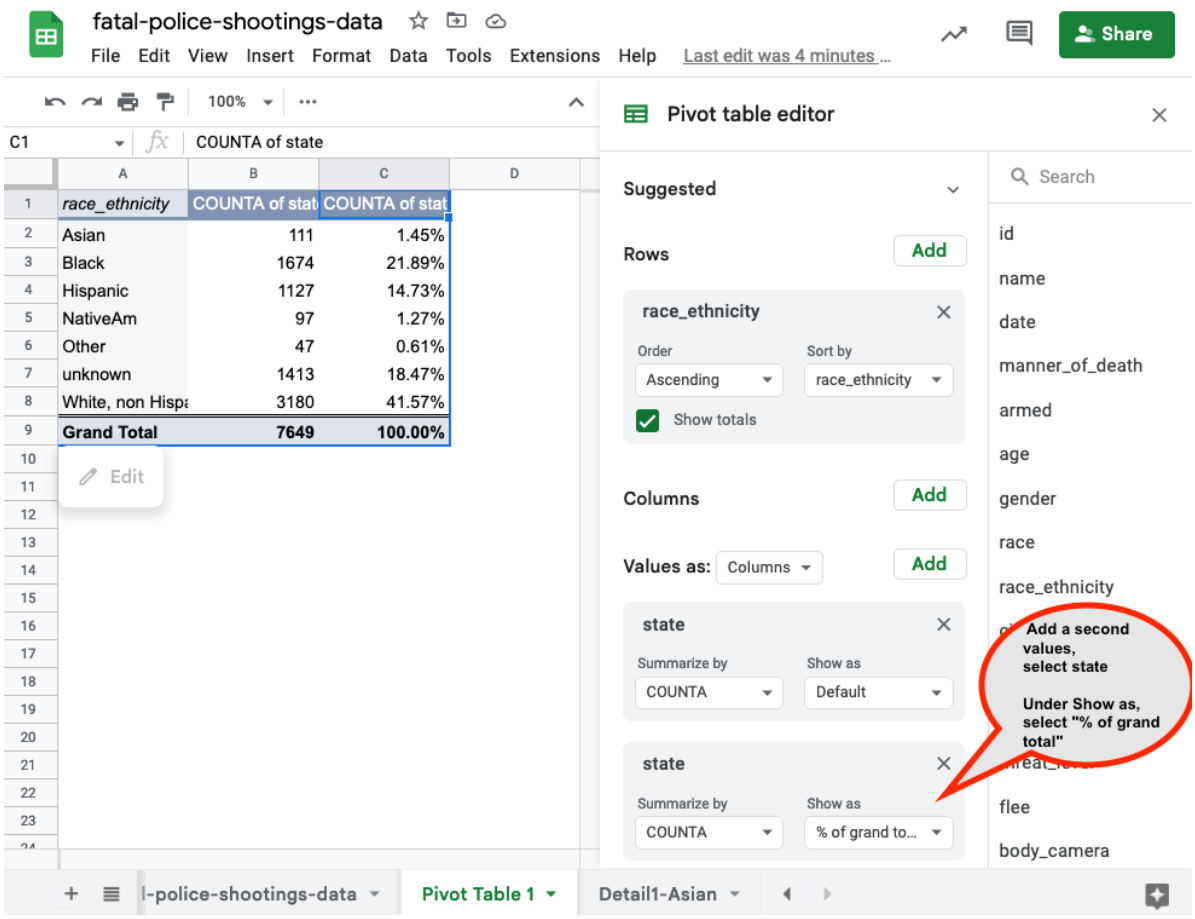
Figure 9.2: pivot menu



Percents of total

It's hard to compare raw numbers unless they're really small. Instead, we'd like to know what *percent* of fatalities by ethnicity. To get a "Percent of Column total", do the following:

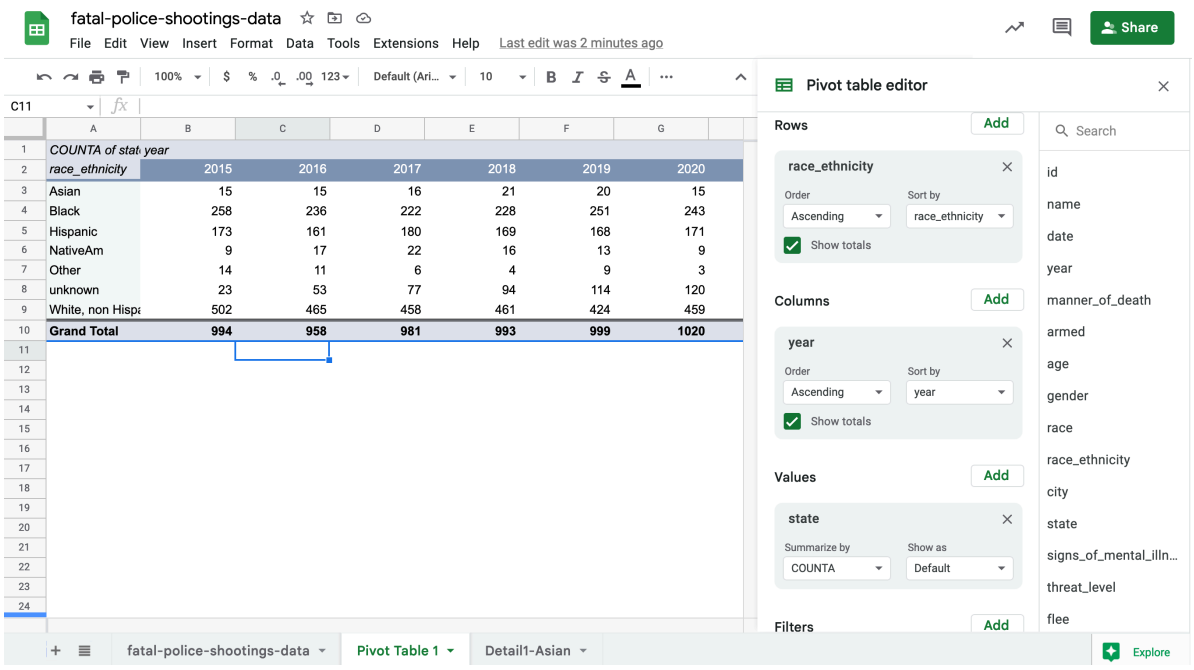
- Add a second values, select state
- Under Show as, select "% of grand total"



More variables

Suppose you'd like to see the number of fatalities by year, with the years across the top and the `ethnicity` down the sides. Add a year variable to columns

- Remove the percent of total column
- Select Columns, then year



Even more variables

Say you wanted to see each city's total shootings by year. Which one had the most last year, and which one had the most overall?

This is actually really hard in a pivot table, because there are cities with the same names in different states. It means you'd need to have a pivot table with TWO columns down the side, and one across the top. Here's an attempt at solving the problems:

- First, Rows, add state
- Rows, add city
- Values, add state, CountA is the default
- Columns, add year

The problems is we can't sort by the combination of city and state. But it does help answer the question on some level.

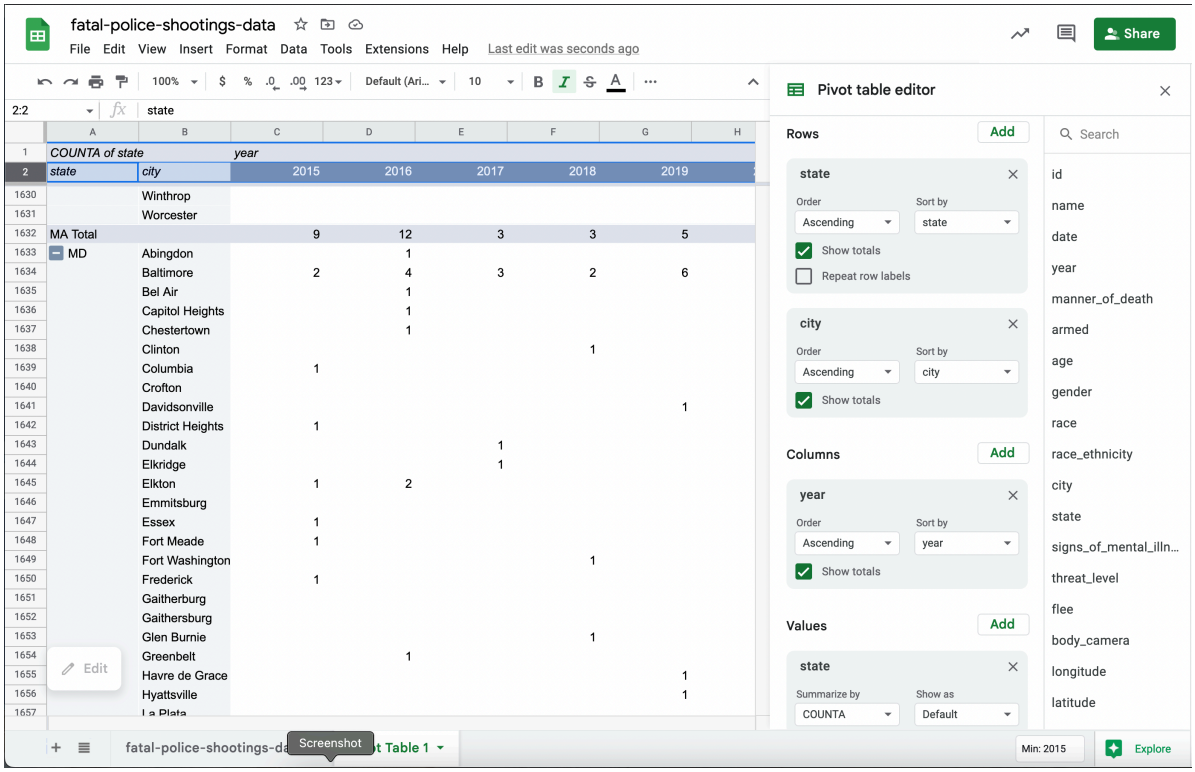


Figure 9.3: More Variables!

9.2 FAQ

I have too many columns

If you want two sets of statistics – say, number of fatalities and percent of fatalities – across the top, it can get very wide and confusing very quickly. One alternative is to change it into more of a vertical rectangle by dragging the "Values" element from the columns to the rows on the right. (This only shows up when you have two calculations being made.)

I want to sort by percents, not numbers

You can't.

Things aren't adding up

You have to be super careful about which column you use to Count things – it has to always be filled out (there can't be any blanks). Go through the filters and find one that doesnt have (Blanks) at the bottom to be sure.

Its a crazy number!

You might have dragged a numeric column into the "Values" area. Check to see if it says "Count" or "Sum". Change it to "Count" if it has something else on it, unless you wanted to add up that column.

This is so frustrating - I can't get what I want

Right? It's time to go to a programming language!

10 Cleaning data with Google Sheets

This chapter will demonstrate some of the bedrock data cleaning skills using Google Sheets, techniques that can be used in Excel and other spreadsheets. We will normalize and clean data by deleting rows, stripping whitespace, making characters lowercase or uppercase. In addition, you will learn to split text to columns, a very handy tool for splitting up dates.

Make a copy of your data before cleaning)

We will use a version of the Washington Post police shooting data to conduct these exercises.

10.0.1 Text to columns

We want to split up the date field into day, month and year. Currently, the format is 2015-01-02. Luckily, the fields all share a common separator, a hyphen, and we can ask Google Sheets to split all according to the hyphen. Other common separators are commas and spaces.

First steps when modifying data: make a backup copy! - Left click on the tab "Police Shootings to Clean" - Select duplicate - Rename "Copy of Police Shootings to Clean" to "Original Police Shootings to Clean." Do not touch this version.

Time to split text to columns. I am extra paranoid (for good reason) and so I always duplicate a date field before modifying it. Duplicate the date column (click on Column C, left click, copy, then Insert column to left, select new blank Column C and paste), save the copy as date-original.

- Select date column
- Select Data | Split text to columns
- See a dialog box: Separator. Select Custom and type in a dash - and enter. You now have the date field chopped up to year, month and day. Rename column E for month and column f for day.

10.0.2 Normalizing

Scroll down the race_ethnicity column and you will see a number of different categories for the same thing: white, White, non Hispanic and Black, African Am. To see all the variations of categorical variables, create a filter and check the different variables

This presents a big problem when you are trying to group and summarize based on these variable names. See this chart

We see white totals 44 and White, non Hispanic total 3,136. We want those to be together – the total is 3,180 – because they are the same thing. Also note that African Am totals 29 and Black totals 1,645, and we would want to combine those as well.

Let's fix it!

Before changing any data, let's work with a copy of the column. - Select race_ethnicity (Column k), left click, copy - Left click on Column K, insert column to right, paste - Rename as race_ethnicity2

Renaming variables. We will rename all "white" as "White, non Hispanic" - Filter race_ethnicity (Column K) to white - in race_ethnicity2, write "White, non Hispanic" in the first column and copy down the list

See how this process works

10.0.3 Lowercase or Uppercase character conversion

Create a filter and notice two variations on Native American: NativeAm and nativeam. You can resolve these differences easily by converting all to Upper or Lower case text using the =UPPER or =LOWER functions.

- To convert NativeAm to lower case, filter on race_ethnicity (Column K) for NativeAm.
- In race_ethnicity2 (Column L), insert a blank column, and type the function =LOWER(K67) and hit enter.

The result should be nativeam as the first entry in race_ethnicity2.

See this example

10.0.4 White space

One obnoxious feature of spreadsheet data is the invisible "white space" or hidden character or carriage returns that can impede your ability to group and summarize variables. Look at the age column. See how some numbers are flush left while most are flush right. The flush left data has hidden white space. You can fix this by clicking on individual cells and deleting the space around the number or you can do it with a function.

- Select age (Column H), left click on Column H, insert column to right, rename as age2
- In cell I2, type =TRIM(H2) and enter. Copy the formula down.

Note how all of the values have been normalized.

These are some of the basic go-to tools for data cleaning in Google Sheets, which can be adapted to Excel, R and other programming languages.