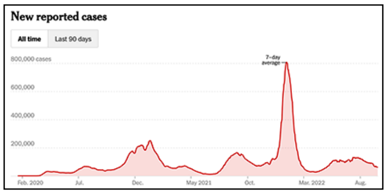
2.1 Collect Datasets & Generate Graphs step-by-step

Canvas assignment page: https://sfsu.instructure.com/courses/17190/assignments/251100

NYT data repository: https://github.com/nytimes/covid-19-data
NYT coronavirus page (US): https://www.nytimes.com/interactive/2021/us/covid-cases.html

Below are the steps to download datasets from the NYT public repository of covid data (Github). Also, how to cut and paste data from NYT homepage, and to make some data visualizations from those datasets using Tableau.

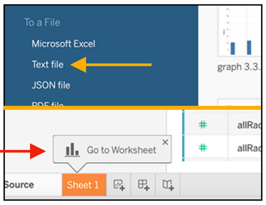
The first visualization is an area graph of covid cases in the US since Jan. 21, 2020. The NYT version looks like the image below.



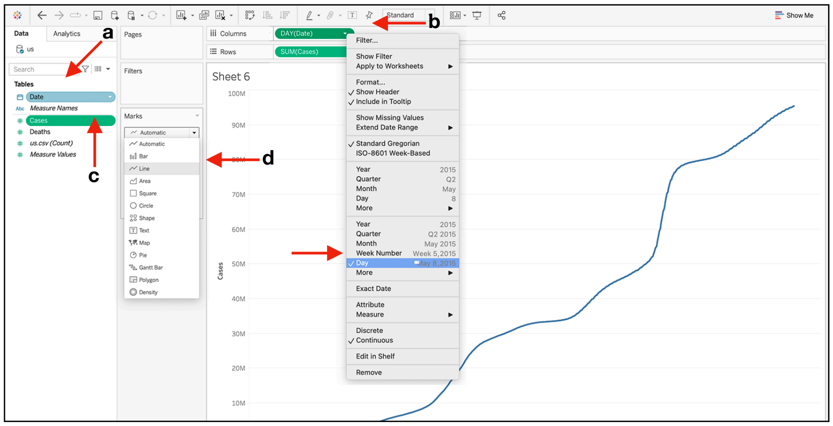
Download the datasets from NYT's Github repository: https://github.com/nytimes/covid-19-data

Click green button (Code) —> Download ZIP.

Unzip and save the archive, it includes a file named us.csv which contains three columns: date, cases, deaths. The numbers are cumulative, so each cell is the total of the day's cases or deaths plus the tally from the day before.



In Tableau, import the us.csv text file; then start a new sheet.



NOTE: screenshots are from Fall 2022. Your Spring 2023 graphs will look a bit different (for example the line will have 5 extra months with a hump on Jan. 2023.

(1) a) Double-click date; **b)** Right-click Date to change date type from Year to Day; **c)** Double click cases; **d)** Change Mark type from Automatic to Line.

The Cases line is now visualized, but because the values are cumulative, the line only goes up. In the next step, you apply a calculation where each value subtracts from it the previous value. This will be the exact value for each individual day and the line will be an accurate representation of the data.

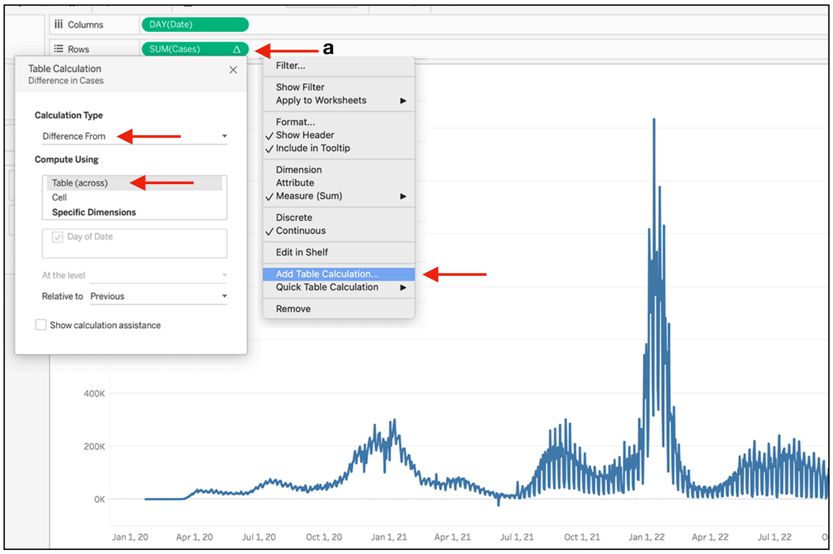


Table Calculation
Difference in Cases

Calculation Type: Difference From

Compute Using: Table (across)

Relative to: Previous

☐ Show calculation assistance

(2) a) Right-click Cases —> Add Table Calculation.

The first, default table calculation is Type: Difference From and Compute Using: Table (across). This is actually the calculation we want (subtract the previous day value from the current value).

The line is now accurate but because there are big variations in the number of cases from day to day, and because there are more than 1,000 days in the graph (more than 2.5 years), the line is hard to read as it jumps back and forth from peak to valley.

In the next step, change the Mark type from Line to Bar. Later, you will change back to line, but bars, with a separate line on top, was how the NYT graph started, so we will replicate it. Later, the bars became unwieldy to show the full length of the pandemic, and so now NYT only uses bars for the last 90 days of data.

Note: If needed, Play Video symbols are linked to extra video content. Datasets and layouts might be different.

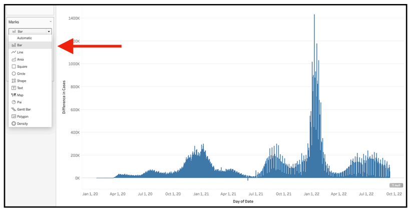
US cases bars and 7-day average line (Tableau)

US cases bars and 7-day average line (Excel)

California cases bars and 7-day average line (Tableau)

SF County cases bars and 7-day average line (Tableau)

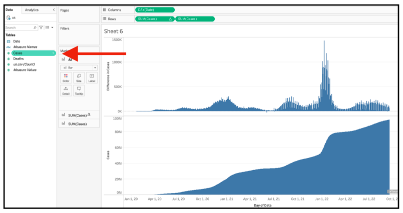
SF County deaths bars and 7-day average line (Tableau)



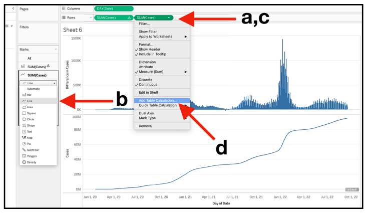
(3) Change Mark type from Line to Bar.

Because there are so many bars, one for each day for more than 1,000 days, they are now all bunched together and look like an area graph. It's OK for now. The next, multiple-step operation is to add a 7-Day Moving Average line on top of the bars. This will be a smoother line showing the weekly average of daily cases. The line will show the seven-day trend while the bars under the line will still show each daily count.

Note: the colors in the images might not match your own colors.



(4) Double-click Cases again. This will add a second bar chart below the first. Again, it looks solid because there are lots of bars. It also only goes up like in the first line graph. In the next steps you will change this second bar chart into a 7-Day Moving Average line.



(5) a) Select the second Cases button in the Rows field; **b)** Change Marks type from Bar to Line; **c)** Right-click Cases button —> **d)** Add Table Calculation.

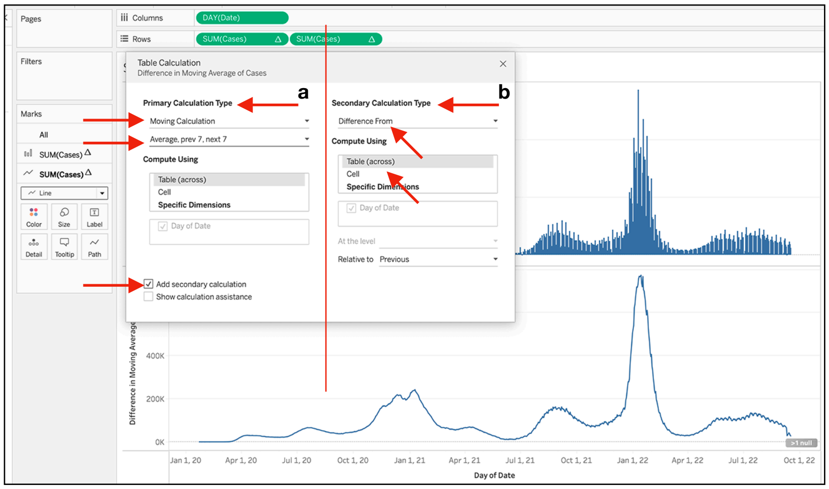


Table Calculation
Difference in Moving Average of Cases

Primary Calculation Type
Moving Calculation
Average, prev 7, next 7
Compute Using
Table (across)
Cell
Specific Dimensions
☑ Day of Date

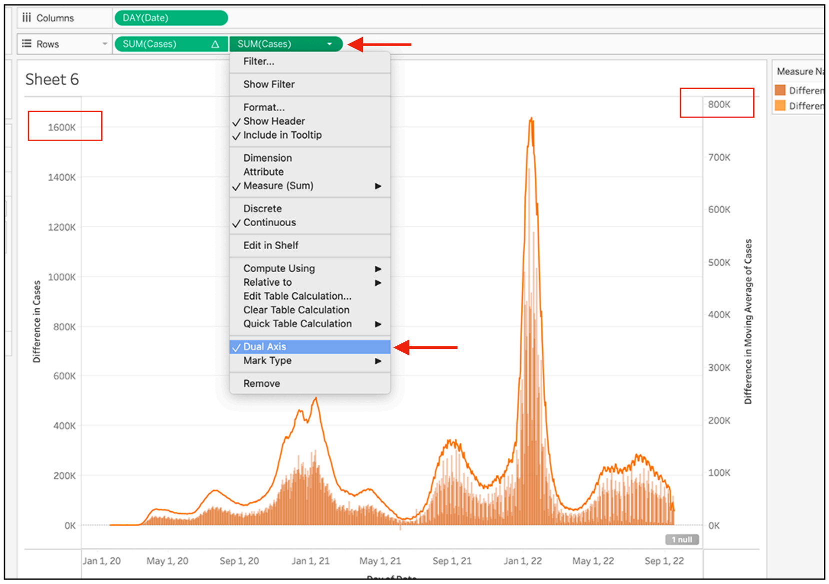
☑ Add secondary calculation
☐ Show calculation assistance

Secondary Calculation Type
Difference From
Compute Using
Table (across)
Cell
Specific Dimensions
☑ Day of Date
At the level
Relative to Previous

(6) In this step the average calculation is set up first:

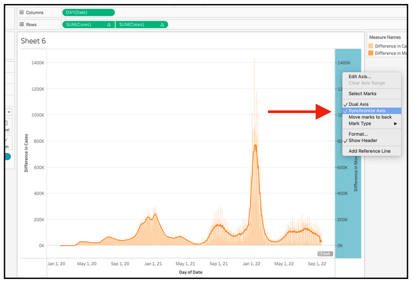
a) Under Primary Calculation Type, select Moving Calculation from the Calculation Type drop-down menu (menu not shown), then select Average from the Summarize Values drop-down menu (menu not shown); then Previous Values: 7; Next Value: 7. Then, check the Add secondary calculation box.

b): The default secondary calculation type is what you want: Difference From. The average line is rendered in the lower pane.

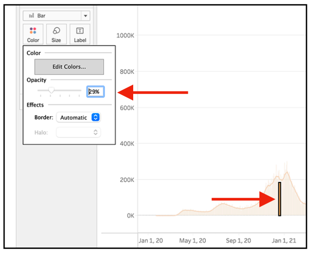


(7) Right-click Cases —> Dual Axis.

Line and bars now overlap (the colors may change and also be different for you). Note that the line does not appear to be an average somewhere between the tips of the tall and the short bars. That's because the two vertical axes (bars and line axes) are now on two different scales: bars at about 1600K and line at about 800K. So the two need to be "synchronized". (in next step).

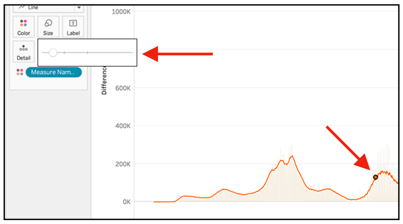


(8) Right-click right axis —> Synchronize Axis. Now the 7-Day Average line overlaps the bars.

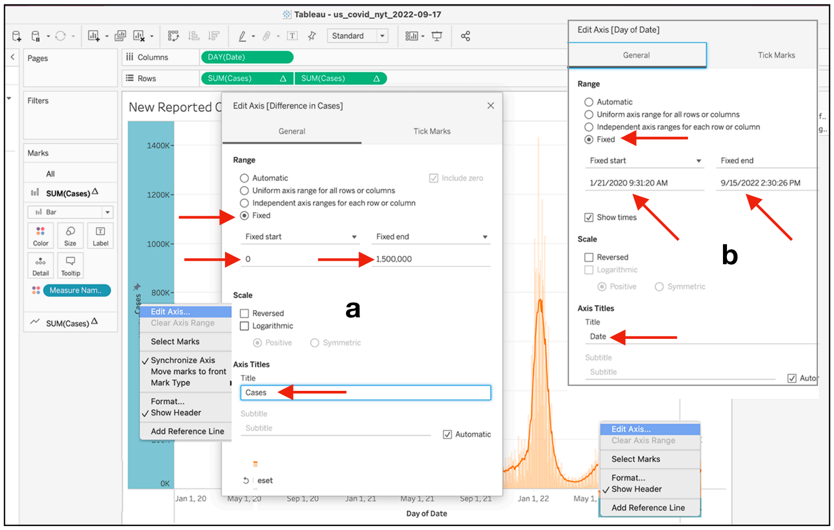


Color
Opacity: 29%
Border: Automatic

(9) Edit the colors and the size of the bars and the line. Select a bar; change the color by clicking Edit Colors and choosing a different color in the color palette (not shown) or you can change the opacity of the current color. Click Size if you want to change the thickness of the bars, however here they are already the smallest possible size.



(10) Edit the color and size (thickness) of the line. Select the line. Use the Color and Size tools to modify the line. In this design based on the NYT graph, you want the bars a very light orange, and the line a bit thick and in a dark orange color.



Columns: DAY(Date)
Rows: SUM(Cases) Δ SUM(Cases) Δ

Edit Axis [Difference in Cases]
General | Tick Marks
Range
○ Automatic ☑ Include zero
○ Uniform axis range for all rows or columns
○ Independent axis ranges for each row or column
● Fixed
Fixed start: 0
Fixed end: 1,500,000
Scale
□ Reversed
□ Logarithmic
● Positive ○ Symmetric
Axis Titles
Title: Cases
Subtitle ☑ Automatic

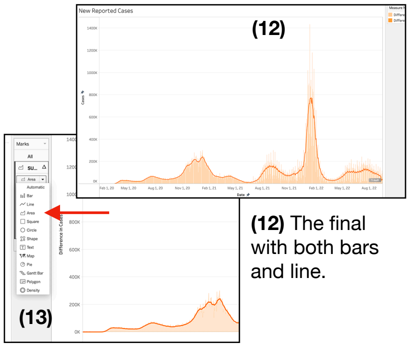
Edit Axis [Day of Date]
General | Tick Marks
Range
○ Automatic
○ Uniform axis range for all rows or columns
○ Independent axis ranges for each row or column
● Fixed
Fixed start: 1/21/2020 9:31:20 AM
Fixed end: 9/15/2022 2:30:26 PM
☑ Show times
Scale
□ Reversed
□ Logarithmic
● Positive ○ Symmetric
Axis Titles
Title: Date
Subtitle ☑ Auto

(11) Edit the vertical and horizontal axes.

a) Right-click the vertical axis (Cases): Range: Fixed —> Start: 0; End: 1,500,000; Axis Title: Cases.

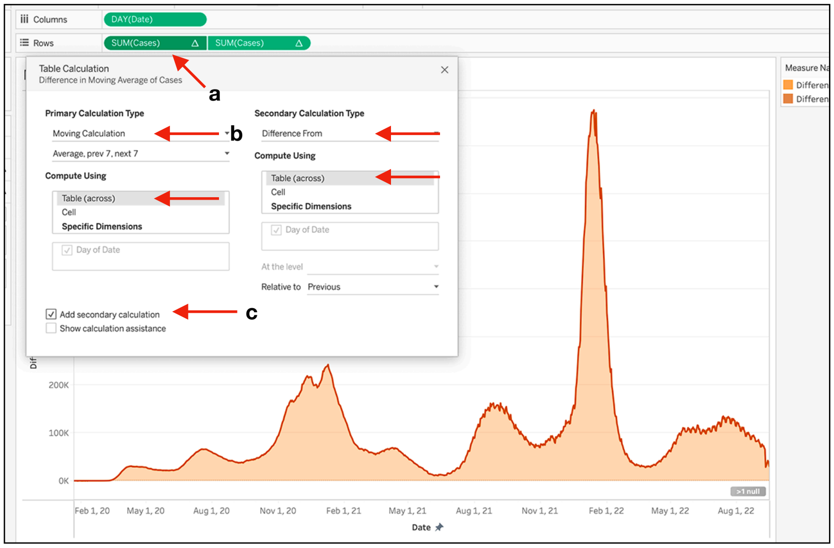
Setting the origin at zero will also remove any “data anomaly”, if for some reason there are any negative values that result in bars pointing downward.

b) Right-click the horizontal axis (Dates): Range: Fixed —> Start: 1/21/2020; End: 2/27/2023; Axis Title: Date. Note: your end date may be different later.



(12) The final with both bars and line.

(13) To match the area/line chart from NYT (see picture on first page) you need to first change the Bar mark to Area.



Columns: DAY(Date)
Rows: SUM(Cases) Δ SUM(Cases) Δ

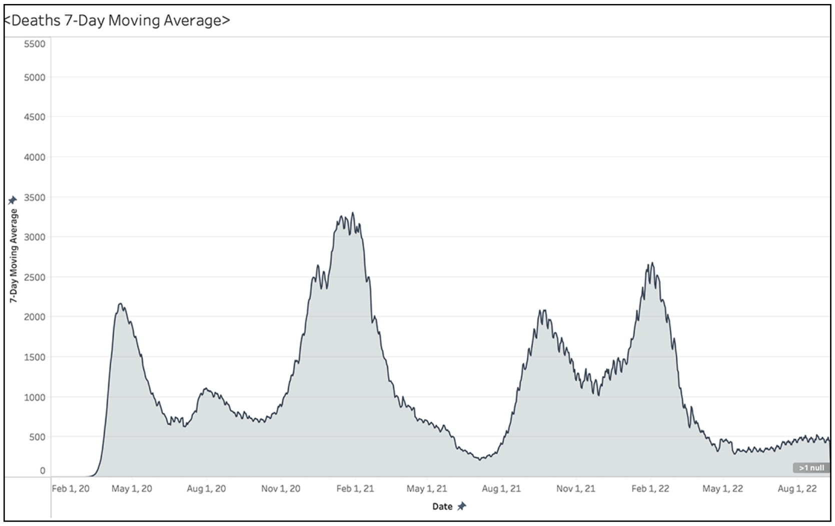
Table Calculation
Difference in Moving Average of Cases

Primary Calculation Type: Moving Calculation
Average, prev 7, next 7
Compute Using: Table (across) / Cell / **Specific Dimensions**
☑ Day of Date

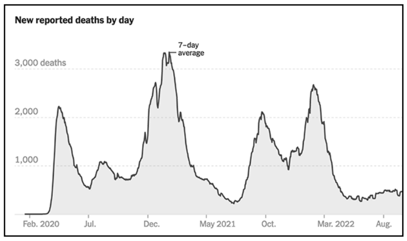
Secondary Calculation Type: Difference From
Compute Using: Table (across) / Cell / **Specific Dimensions**
☑ Day of Date
At the level
Relative to: Previous

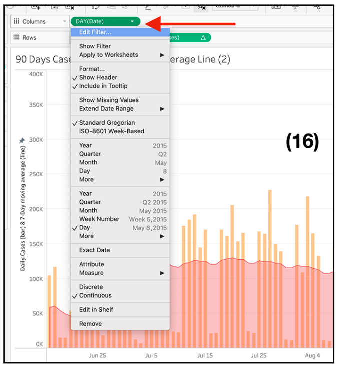
☑ Add secondary calculation
□ Show calculation assistance

(14) Then, smooth the edges of the area by applying the same average calculation as the line.
a) Right-click the left Cases button.
b) Change the existing Difference calculation (not shown) to the 7-Day average as for the line. See Fig. 6a and text for detailed steps. (the area will temporarily go only up again).
c) Add secondary Calculation —> Difference from Table across. The edge of the light area will now match the darker line above it.



(15) You can try to create the same area chart as #14 but using the Deaths variable instead of Cases. Here I used a light gray for the area and a dark gray for the line, trying to match the same chart on the NYT website, reprinted below:





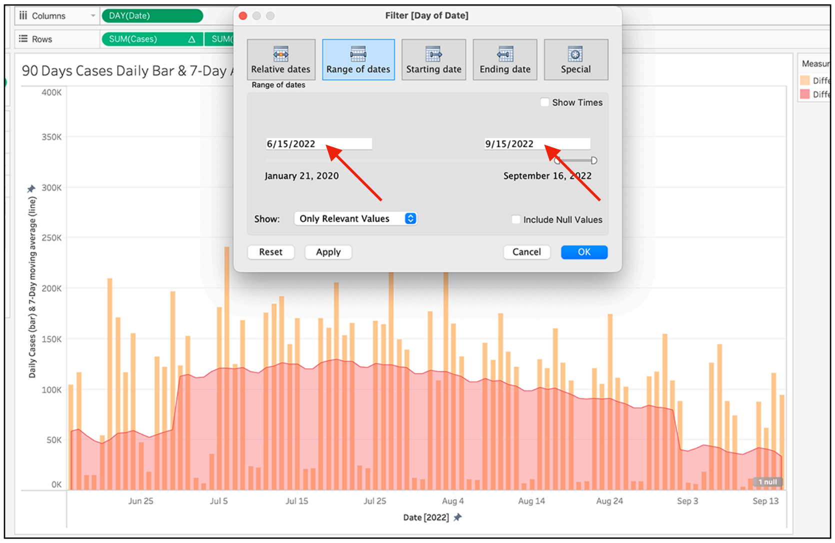
(16)

Cases in last 90 days

In this variation, showing cases for only the last 90 days, follow the same steps from #1 to #12, but before #3 … apply the date filter below:

(16) To apply the filter, right-click Day/Date in the Columns field.

Apply a filter to Day/Date to restrict the time frame to start at Dec. 1, 2022 and end at Feb. 28, 2023.
(Screenshot may look slightly different).


Filter [Day of Date]

Range of dates

6/15/2022 9/15/2022

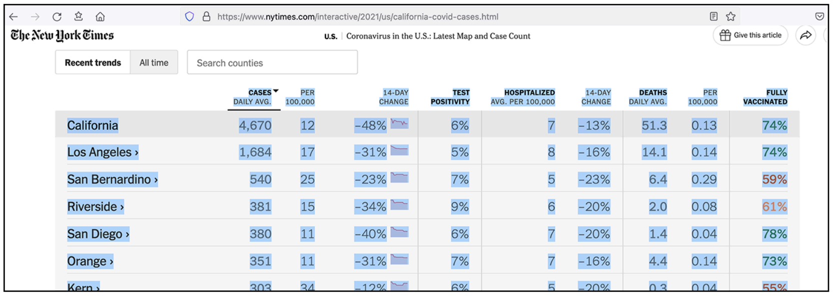
January 21, 2020 September 16, 2022

Show: Only Relevant Values

(17) Filter the Day/Date. Use the slider or type the desired dates in the left and right fields: here I used June 15, 2022 and September 15, 2022.

Change the dates to a later date when you actually do this graph, using an updated dataset. Note: the pics show the Filter steps, however these were done before all the other steps that resulted in the final chart shown here.

California counties choropleth map (colored map) showing average daily covid cases per 100,000 in last seven days (exact dates on day of web scrape).

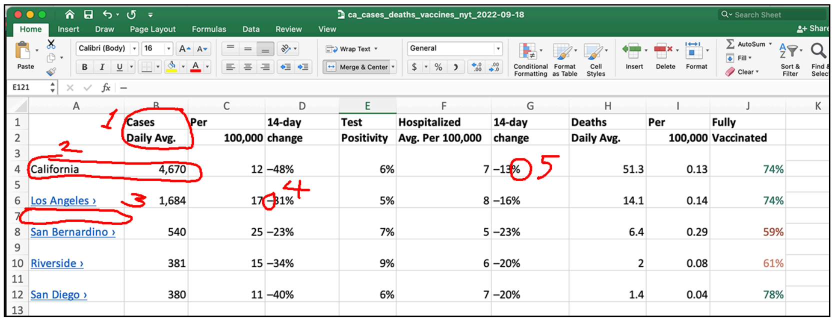

https://www.nytimes.com/interactive/2021/us/california-covid-cases.html
The New York Times
U.S. | Coronavirus in the U.S.: Latest Map and Case Count
Give this article
Recent trends
All time
Search counties

	CASES DAILY AVG.	PER 100,000	14-DAY CHANGE	TEST POSITIVITY	HOSPITALIZED AVG. PER 100,000	14-DAY CHANGE	DEATHS DAILY AVG.	PER 100,000	FULLY VACCINATED
California	4,670	12	−48%	6%	7	−13%	51.3	0.13	74%
Los Angeles ›	1,684	17	−31%	5%	8	−16%	14.1	0.14	74%
San Bernardino ›	540	25	−23%	7%	5	−23%	6.4	0.29	59%
Riverside ›	381	15	−34%	9%	6	−20%	2.0	0.08	61%
San Diego ›	380	11	−40%	6%	7	−20%	1.4	0.04	78%
Orange ›	351	11	−31%	7%	7	−16%	4.4	0.14	73%
Kern ›	303	34	−12%	6%	5	−20%	0.3	0.04	55%

Click on link below:
https://www.nytimes.com/interactive/2021/us/california-covid-cases.html

On webpage, click Show All button to expand the table and include all counties. **(18)** Select all table information including column headers. Copy and paste the data into a new Excel file.

NOTE: for this type of cut-and-paste and in general, I highly recommend that you use an external mouse as an input device, instead of the trackpad on your laptop.

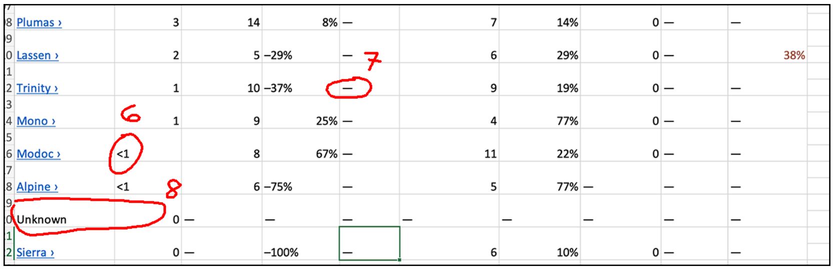


	A	B	C	D	E	F	G	H	I	J	K
1		Cases	Per	14-day	Test	Hospitalized	14-day	Deaths	Per	Fully	
2		Daily Avg.	100,000	change	Positivity	Avg. Per 100,000	change	Daily Avg.	100,000	Vaccinated	
3											
4	California	4,670	12	−48%	6%	7	−13%	51.3	0.13	74%	
5											
6	Los Angeles ›	1,684	17	−31%	5%	8	−16%	14.1	0.14	74%	
7											
8	San Bernardino ›	540	25	−23%	7%	5	−23%	6.4	0.29	59%	
9											
10	Riverside ›	381	15	−34%	9%	6	−20%	2	0.08	61%	
11											
12	San Diego ›	380	11	−40%	6%	7	−20%	1.4	0.04	78%	
13											

(19) Save the file after pasting the data into Excel. Note the following items that need work:

1. Header names are in two rows; combine info into just one row, taking care to clearly label each column.
2. Delete the California row.
3. Note empty rows between data rows. Delete all empty rows.
4. The minus signs are dashes but should be hyphens. Find/replace all in text editor.
5. In text-editor, delete all percentage symbols (%).

Note: add "county" label in top-left cell. Also, add an extra column (not shown) titled "state" and write "California" in every cell.

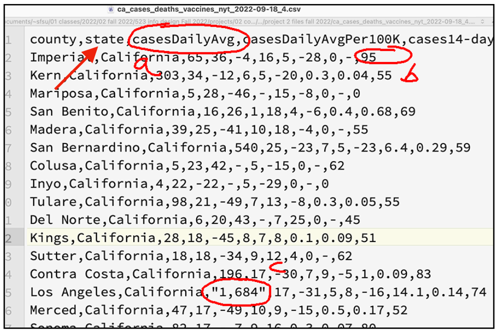


8	Plumas ›	3	14	8%	—	7	14%	0	—	—
9										
0	Lassen ›	2	5	−29%	—	6	29%	0	—	38%
1										
2	Trinity ›	1	10	−37%	—	9	19%	0	—	—
3										
4	Mono ›	1	9	25%	—	4	77%	0	—	—
5										
6	Modoc ›	<1	8	67%	—	11	22%	0	—	—
7										
8	Alpine ›	<1	6	−75%	—	5	77%	—	—	—
9										
0	Unknown	0	—	—	—	—	—	—	—	—
1										
2	Sierra ›	0	—	−100%	—	6	10%	0	—	—

(20) Scroll down in the file to note additional problematic characters:

6. Change data with "<" symbols; write "0.5"
7. A different dash that should be an hyphen; find/replace all.
8. Delete "Unknown" row.

Save the file as CSV, then reopen it in Excel to delete the rows, but reopen it in text-editor to do find/replace. Save as CSV.

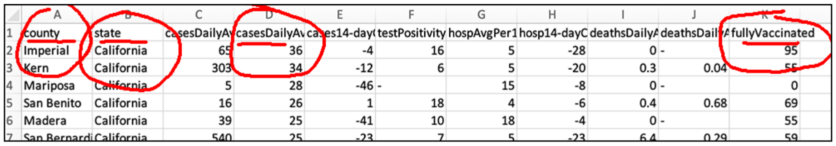

ca_cases_deaths_vaccines_nyt_2022-09-18_4.csv
cuments/~sfsu/01 classes/2022/02 fall 2022/523 info design Fall 2022/projects/02 co.../.../project 2 files fall 2022/ca_cases_deaths_vaccines_nyt_2022-09-18_4....
a
b
c

```
county,state,casesDailyAvg,casesDailyAvgPer100K,cases14-day
Imperial,California,65,36,-4,16,5,-28,0,-,95
Kern,California,303,34,-12,6,5,-20,0.3,0.04,55
Mariposa,California,5,28,-46,-,15,-8,0,-,0
San Benito,California,16,26,1,18,4,-6,0.4,0.68,69
Madera,California,39,25,-41,10,18,-4,0,-,55
San Bernardino,California,540,25,-23,7,5,-23,6.4,0.29,59
Colusa,California,5,23,42,-,5,-15,0,-,62
Inyo,California,4,22,-22,-,5,-29,0,-,0
Tulare,California,98,21,-49,7,13,-8,0.3,0.05,55
Del Norte,California,6,20,43,-,7,25,0,-,45
Kings,California,28,18,-45,8,7,8,0.1,0.09,51
Sutter,California,18,18,-34,9,12,4,0,-,62
Contra Costa,California,196,17,-30,7,9,-5,1,0.09,83
Los Angeles,California,"1,684",17,-31,5,8,-16,14.1,0.14,74
Merced,California,47,17,-49,10,9,-15,0.5,0.17,52
Sonoma,California,82,17,-,7,9,16,0.3,0.07,80
```

(21) This view in BBEdit (text editor) shows: New State column (arrow)

a. Simplified header names (no spaces)
b. All percentage symbols (%) have been removed.
c. Note quotation marks if data point is 1,000 or higher. This is OK here as Tableau will interpret the data correctly.

Saving the file as Excel or CSV should not matter here. Tableau will work with both.

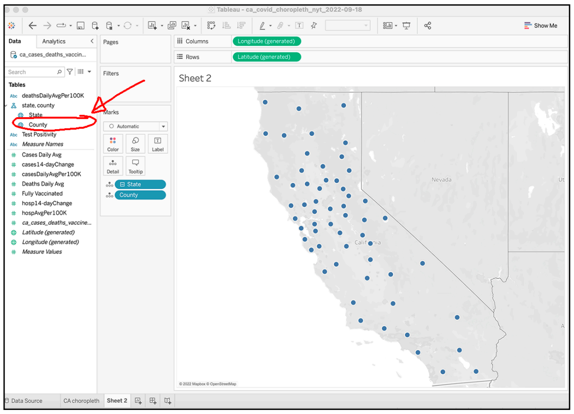


	A	B	C	D	E	F	G	H	I	J	K
1	county	state	casesDailyA	casesDailyAv	cases14-day(	testPositivity	hospAvgPer1	hosp14-dayC	deathsDailyA	deathsDail	fullyVaccinated
2	Imperial	California	65	36	-4	16	5	-28	0	-	95
3	Kern	California	303	34	-12	6	5	-20	0.3	0.04	55
4	Mariposa	California	5	28	-46	-	15	-8	0	-	0
5	San Benito	California	16	26	1	18	4	-6	0.4	0.68	69
6	Madera	California	39	25	-41	10	18	-4	0	-	55
7	San Bernardi	California	540	25	-23	7	5	-23	6.4	0.29	59

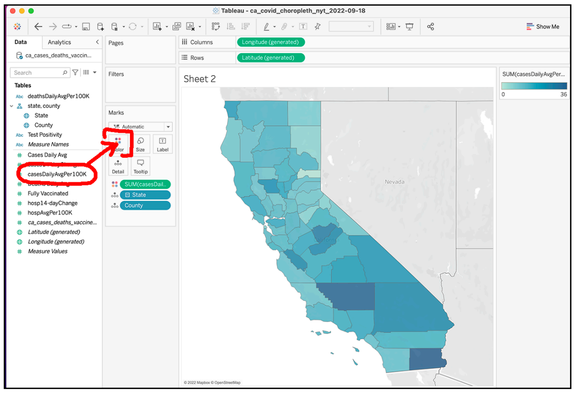
(22) The same file opened in Excel, showing the County header label, the additional State column, the other simplified header names. The percentage symbols (%) have been removed. Save the file and connect to it from Tableau.

US cases per 100K choropleth map (Tableau)

California deaths per 100K choropleth map (Tableau)

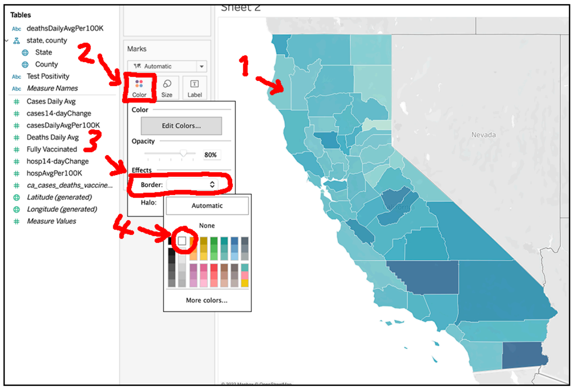


(23) After connecting to the file in Tableau, start a new sheet and double-click County in the Tables pane. Sometimes not every county gets visualized by a dot and some counties might be missing. If this happens (you will notice it later if there are unfilled color areas), double-click State as well.



(24) Drag casesDailyAvgPer100K (or the corresponding name in your file) to the color symbol. Each county now is colored with a shade corresponding to its number of cases (darker color = more cases).

Note: Map colors will reflect most recent data used.

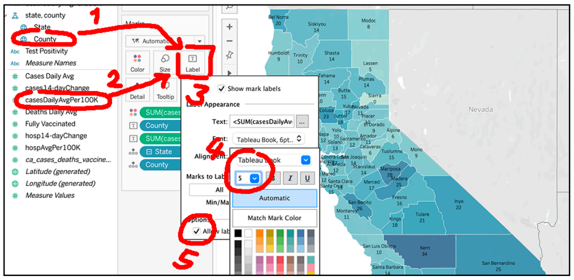


(25)

1. Click on a county.
2. Click Color.

3-4. Border —> White

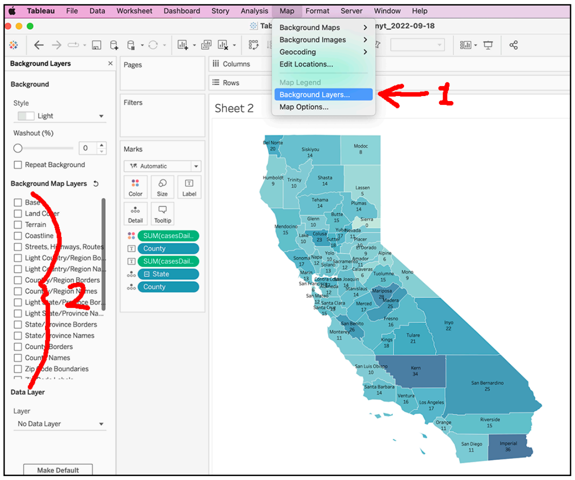
This will make the county borders white, helping to separate them from each other.



(26)

1. Drag County to Label.
2. Drag casesDaily… to Label.
3. Click Label
4. Label type size: 5 pts
5. Check "Allow labels to overlap…"

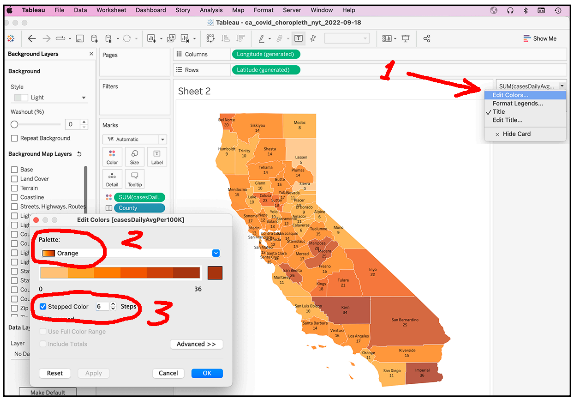
This will add the county names and the cases number to the map. Note: if "California" gets added and repeated with the county names, delete "State" under the marks pane and tools.



(27) Remove map background elements:

1. Map —> Background Layers
2. Uncheck all Background Map Layers.

The map is much cleaner now.

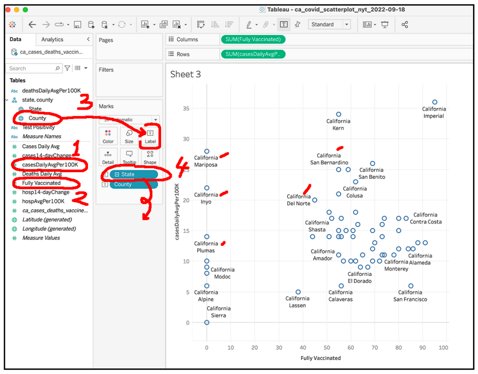


(28) Change the color to orange and add steps:

1. In color key at right, click Edit Colors…
2. In pop-up menu, select a different color scale.
3. Check Stepped Color; 6 steps.

Save the Tableau file and Print/Save the graph as PDF.

California deaths per 100K choropleth map (Illustrator edits)

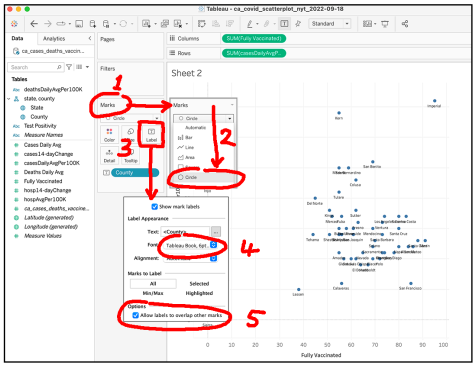


CA cases and vaccinations scatterplot

Start a new Tableau file and connect to the same dataset used for the California choropleth map (#s 18–28).

(29) Steps:

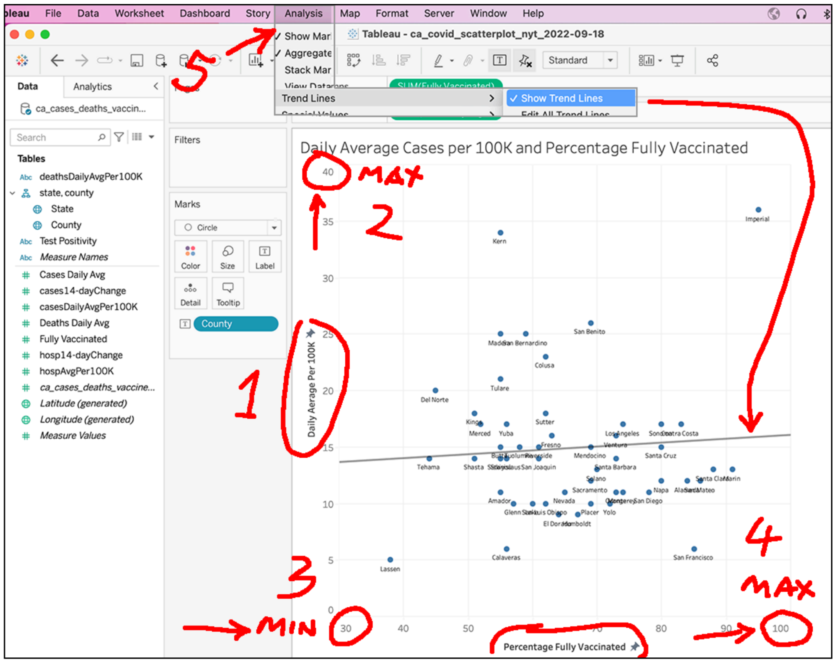
1. Double-click casesDaily…
2. Double-click Fully Vaccinated
3. Drag County to Label
4. Delete State item under tools if “California” label is being repeated next to each county name on the scatterplot (Right-click to delete).

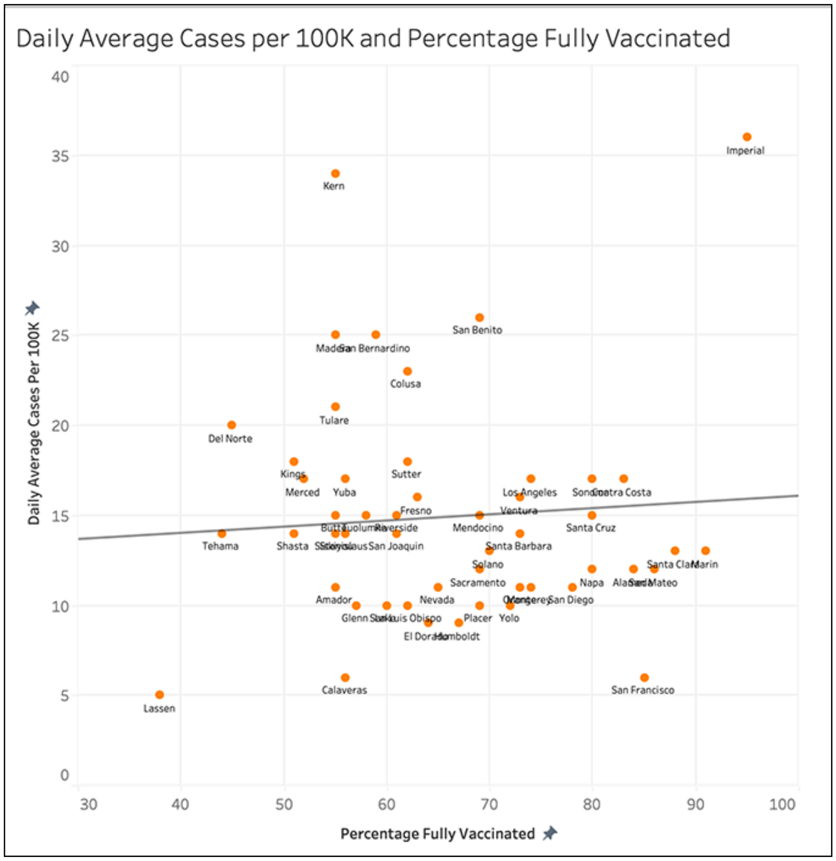


(30) Steps:

1. Right-click Marks and change from automatic to…
2. Circle (solid dot)
3. Edit Label
4. Type size —> 6 pts
5. Check “Allow label to overlap…"

California counties scatterplot cases & vaccines (Tableau)





(31) Steps:

1. Right-click and edit vertical axis: Simplify axis label wording
2. Fixed axis: Max: 40
3. Right-click and edit horizontal axis: Min: 30 (percent)
4. Max: 100 (percent) and add "Percentage" to axis label
5. Add a trend line: Analysis —> Show trend lines

(32) Change colors if desired. Right-click color and edit colors (tools not shown).

Save Tableau file and Print/Save as PDF.