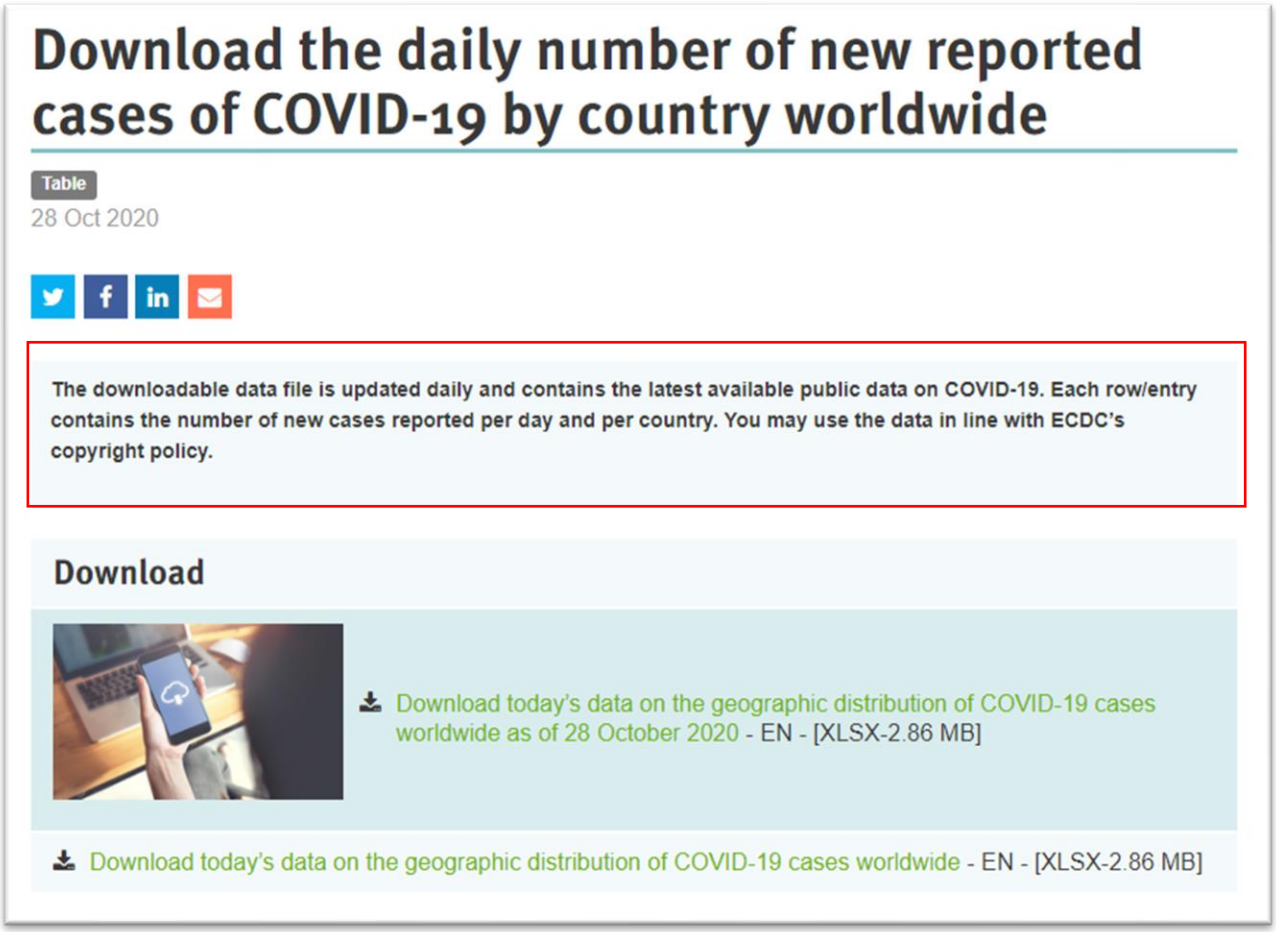
Downloading COVID-19 data and analysing it using Excel

Go to the European Centre for Disease Prevention and Control and download the Excel file that contains the most recent data on the worldwide, geographic distribution of COVID-19 infections.

Before we download data from a portal like this one, always read about the file to make sure you understand what you're getting. Typically, this information is contained in a "readme" file or data dictionary that can be downloaded or read on the site.

In this case, the explanation – also called a "nerd box" by data-journalists – is one of the first things you notice on the website that was last accessed on Oct 28.



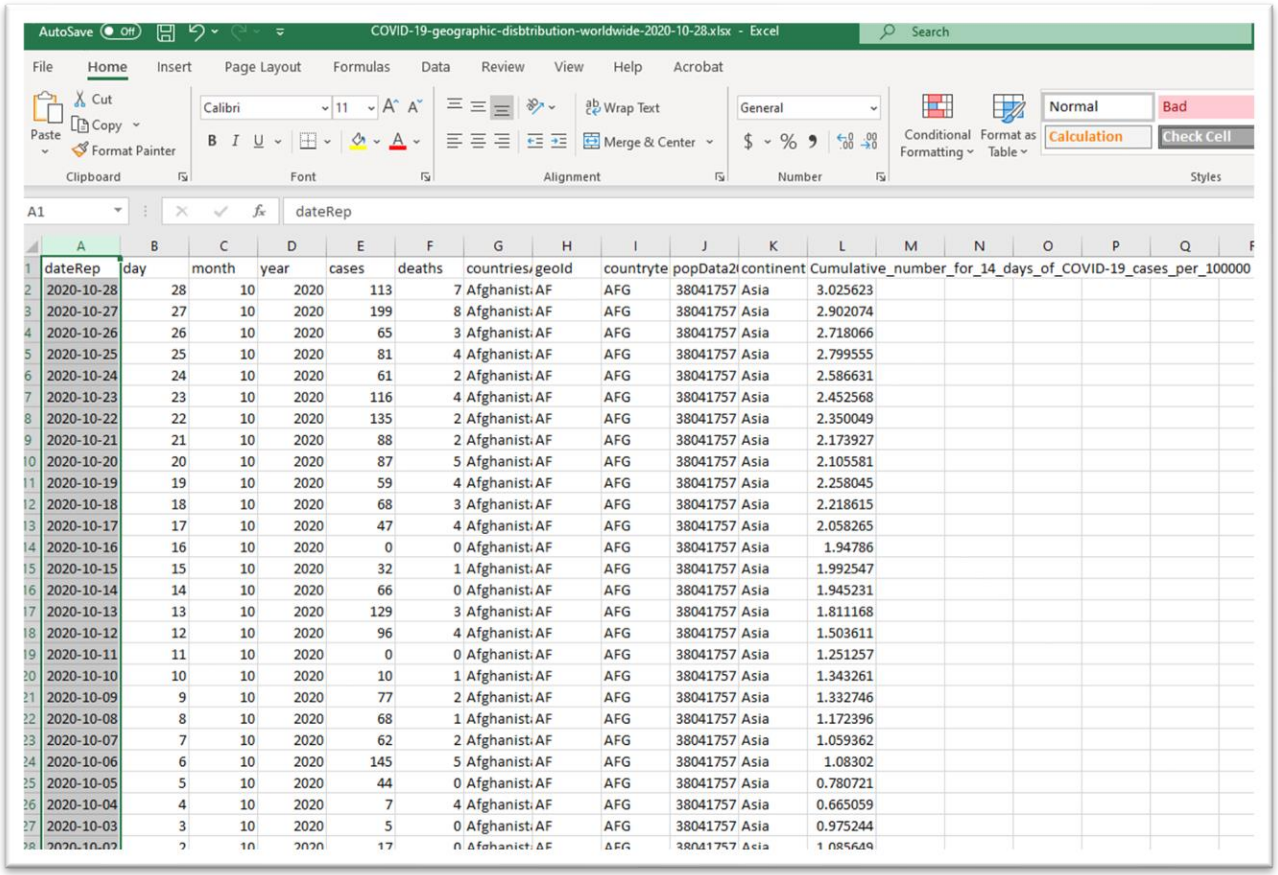
From the explanation, we learn that each row of information in the spreadsheet represents a daily count, snapshots of infections and deaths.

This explanation is crucial if we are to make sense of the table we're about the download.

After you have read the explanation, download the file.

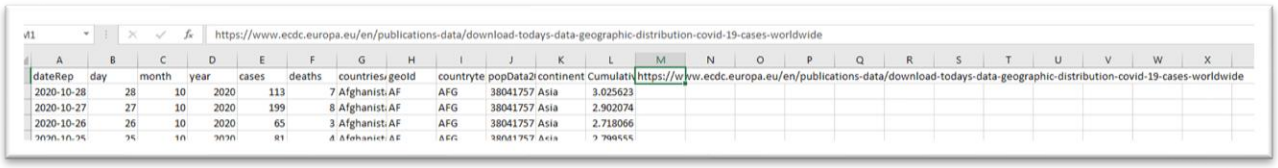
Be sure to save it in a folder you have created for this exercise.

Open the Excel file, which will look similar this earlier version.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	dateRep	day	month	year	cases	deaths	countries	geoId	countryte	popData2	continent	Cumulative_number_for_14_days_of_COVID-19_cases_per_100000					
2	2020-10-28	28	10	2020	113	7	Afghanist	AF	AFG	38041757	Asia	3.025623					
3	2020-10-27	27	10	2020	199	8	Afghanist	AF	AFG	38041757	Asia	2.902074					
4	2020-10-26	26	10	2020	65	3	Afghanist	AF	AFG	38041757	Asia	2.718066					
5	2020-10-25	25	10	2020	81	4	Afghanist	AF	AFG	38041757	Asia	2.799555					
6	2020-10-24	24	10	2020	61	2	Afghanist	AF	AFG	38041757	Asia	2.586631					
7	2020-10-23	23	10	2020	116	4	Afghanist	AF	AFG	38041757	Asia	2.452568					
8	2020-10-22	22	10	2020	135	2	Afghanist	AF	AFG	38041757	Asia	2.350049					
9	2020-10-21	21	10	2020	88	2	Afghanist	AF	AFG	38041757	Asia	2.173927					
10	2020-10-20	20	10	2020	87	5	Afghanist	AF	AFG	38041757	Asia	2.105581					
11	2020-10-19	19	10	2020	59	4	Afghanist	AF	AFG	38041757	Asia	2.258045					
12	2020-10-18	18	10	2020	68	3	Afghanist	AF	AFG	38041757	Asia	2.218615					
13	2020-10-17	17	10	2020	47	4	Afghanist	AF	AFG	38041757	Asia	2.058265					
14	2020-10-16	16	10	2020	0	0	Afghanist	AF	AFG	38041757	Asia	1.94786					
15	2020-10-15	15	10	2020	32	1	Afghanist	AF	AFG	38041757	Asia	1.992547					
16	2020-10-14	14	10	2020	66	0	Afghanist	AF	AFG	38041757	Asia	1.945231					
17	2020-10-13	13	10	2020	129	3	Afghanist	AF	AFG	38041757	Asia	1.811168					
18	2020-10-12	12	10	2020	96	4	Afghanist	AF	AFG	38041757	Asia	1.503611					
19	2020-10-11	11	10	2020	0	0	Afghanist	AF	AFG	38041757	Asia	1.251257					
20	2020-10-10	10	10	2020	10	1	Afghanist	AF	AFG	38041757	Asia	1.343261					
21	2020-10-09	9	10	2020	77	2	Afghanist	AF	AFG	38041757	Asia	1.332746					
22	2020-10-08	8	10	2020	68	1	Afghanist	AF	AFG	38041757	Asia	1.172396					
23	2020-10-07	7	10	2020	62	2	Afghanist	AF	AFG	38041757	Asia	1.059362					
24	2020-10-06	6	10	2020	145	5	Afghanist	AF	AFG	38041757	Asia	1.08302					
25	2020-10-05	5	10	2020	44	0	Afghanist	AF	AFG	38041757	Asia	0.780721					
26	2020-10-04	4	10	2020	7	4	Afghanist	AF	AFG	38041757	Asia	0.665059					
27	2020-10-03	3	10	2020	5	0	Afghanist	AF	AFG	38041757	Asia	0.975244					
28	2020-10-02	2	10	2020	17	0	Afghanist	AF	AFG	38041757	Asia	1.085649					

One of the first things we will do is copy the URL for the centre's site and paste it into the table's first available cell.

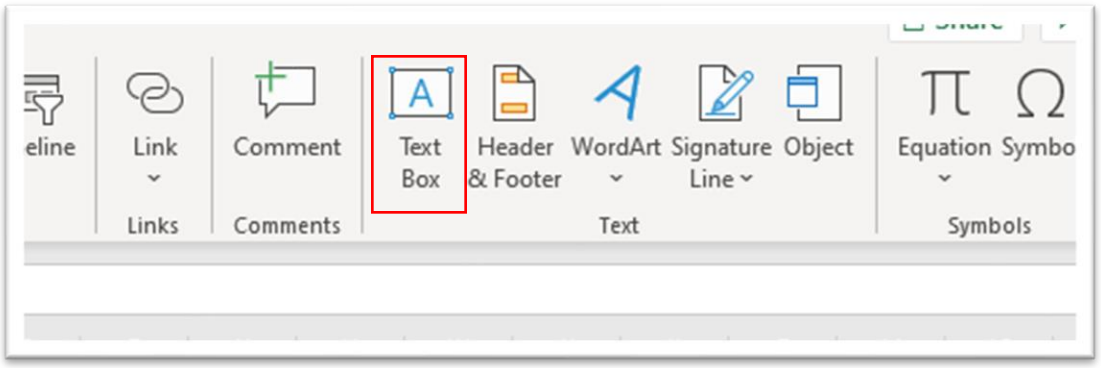


This is good practice that ensures you always know the origin of the data.

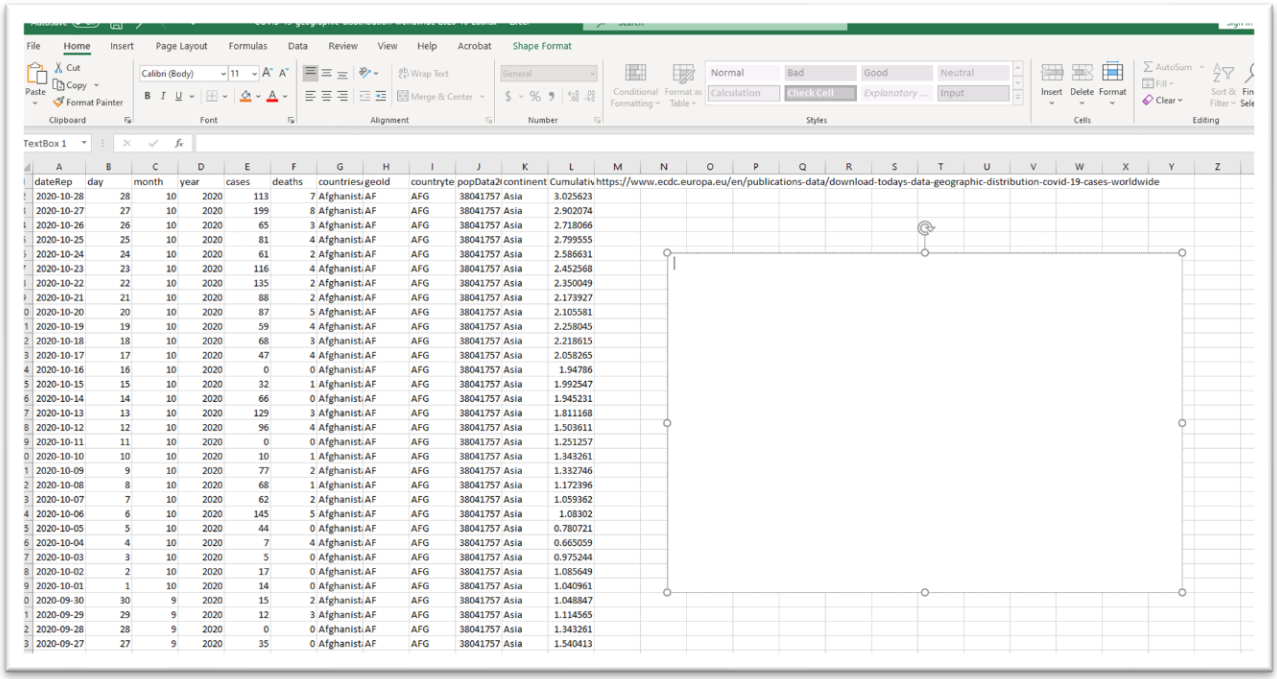
And remember that "nerd box", let's copy and embed that explanation into this worksheet.

Click "Insert" on the menu across the top and scroll across to your right to the "Text section".

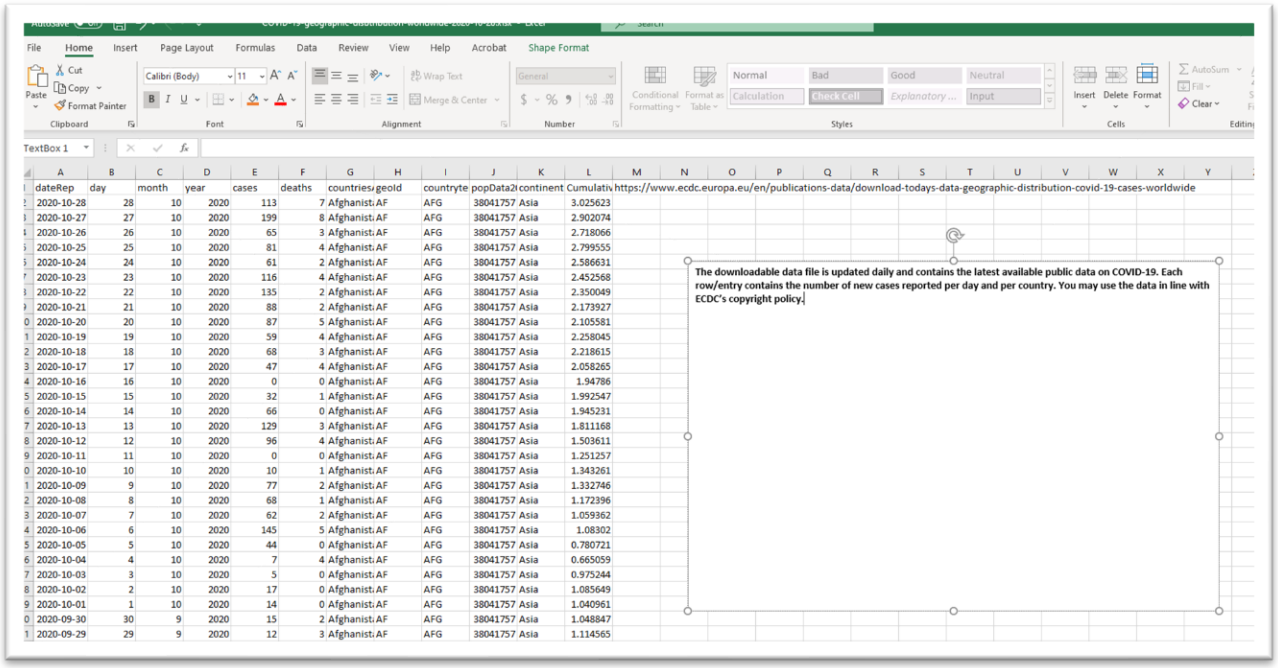
Hover your mouse over the “Text Box” icon.



Place your mouse – which turns into a cross-like icon -- anywhere to the right of the table, hold the mouse down with your index finger and draw a box large enough to contain the nerd box explanation.

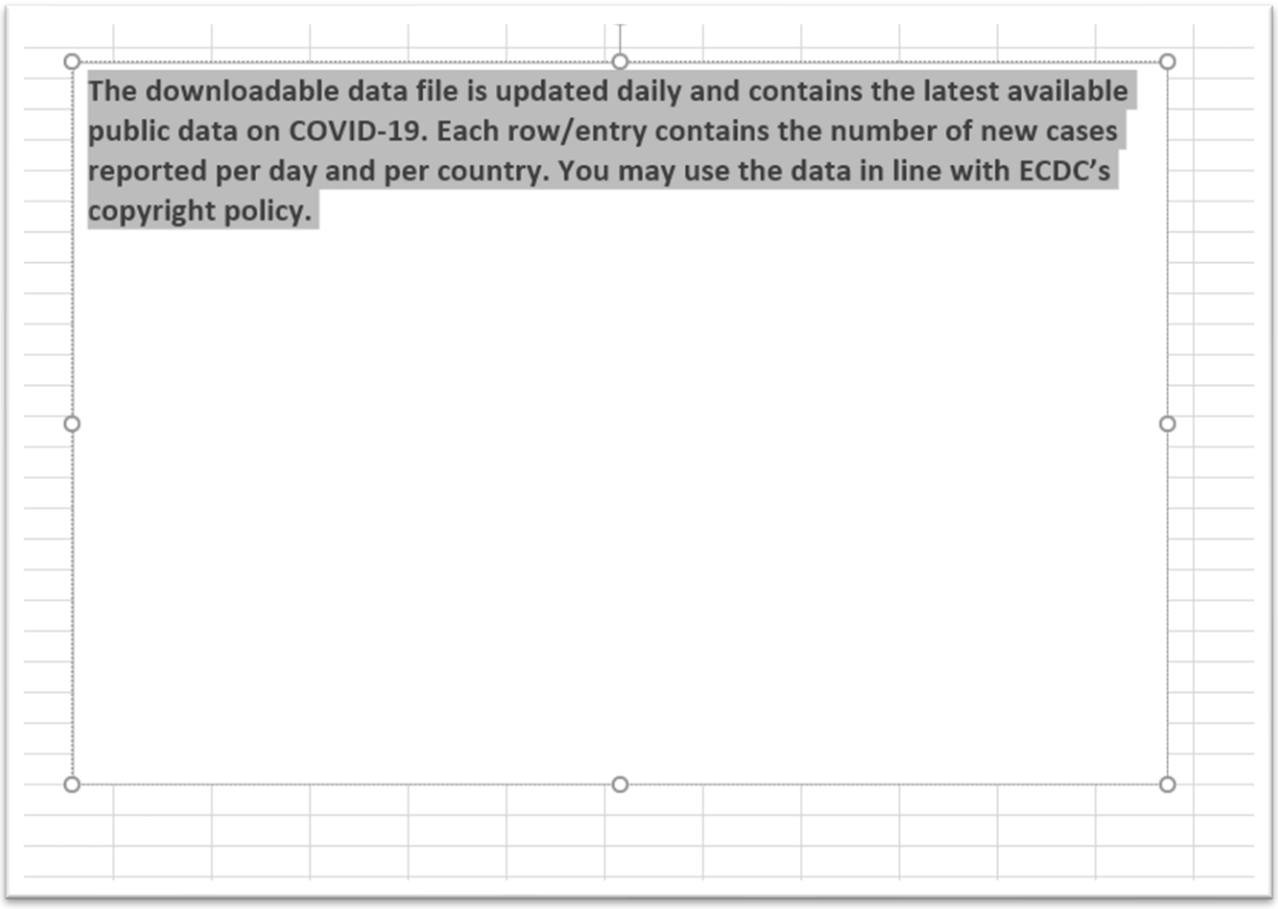


Paste the explanation.

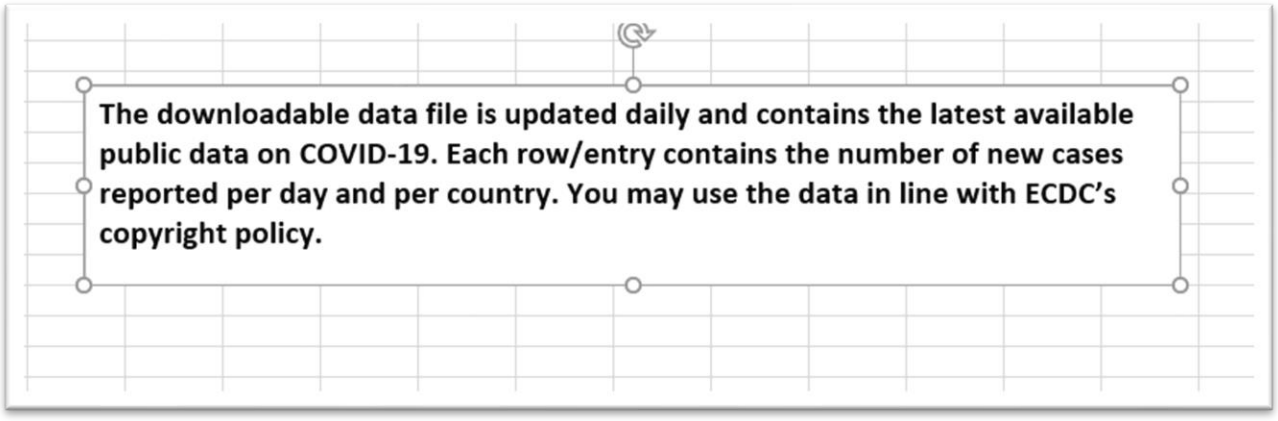


Our text box is akin to a Word document in our spreadsheet which can be expanded to fit as much text as you would like. They are handy, especially when you're sharing information with colleagues that might require some quick context. Feel free to increase the point size by highlighting the paragraph and increasing the font size, just like you

would in a Word document.

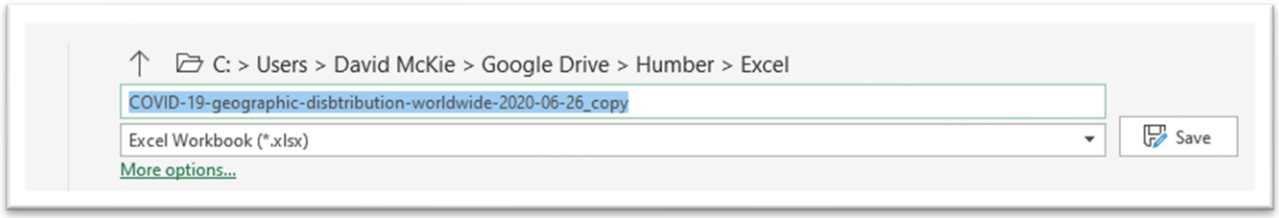


You can even make the box smaller by placing your cursor over small circle in the middle of the bottom boundary and moving it upwards.



Now you have the website's URL and a quick explainer.

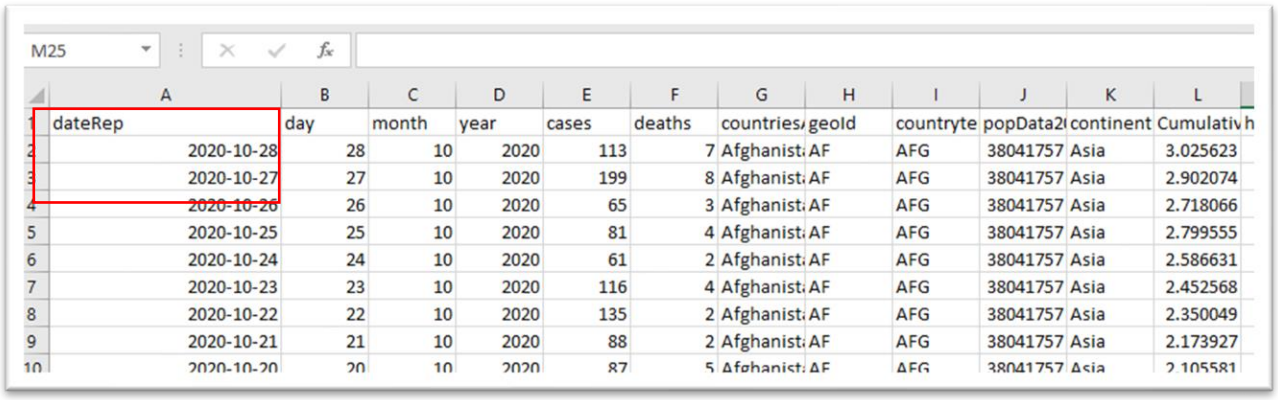
Save this table as a back-up, a crucial step that must be followed to avoid potential tears if disaster ever strikes.



Now we will take a few minutes to analyze what the information in the table can tell us, and just as importantly what it can't.

Let's begin with what it can tell us.

Adjust the column widths to read the column titles.



M25

	A	B	C	D	E	F	G	H	I	J	K	L
1	dateRep	day	month	year	cases	deaths	countries/	geoId	countryte	popData2(	continent	Cumulativ h
2	2020-10-28	28	10	2020	113	7	Afghanist	AF	AFG	38041757	Asia	3.025623
3	2020-10-27	27	10	2020	199	8	Afghanist	AF	AFG	38041757	Asia	2.902074
4	2020-10-26	26	10	2020	65	3	Afghanist	AF	AFG	38041757	Asia	2.718066
5	2020-10-25	25	10	2020	81	4	Afghanist	AF	AFG	38041757	Asia	2.799555
6	2020-10-24	24	10	2020	61	2	Afghanist	AF	AFG	38041757	Asia	2.586631
7	2020-10-23	23	10	2020	116	4	Afghanist	AF	AFG	38041757	Asia	2.452568
8	2020-10-22	22	10	2020	135	2	Afghanist	AF	AFG	38041757	Asia	2.350049
9	2020-10-21	21	10	2020	88	2	Afghanist	AF	AFG	38041757	Asia	2.173927
10	2020-10-20	20	10	2020	87	5	Afghanist	AF	AFG	38041757	Asia	2.105581

Notice that the numbers in a number of columns -- A, B, C, D, E, F and J -- are all justified to the right. This means that Excel has correctly interpreted them as numbers. This is important because spreadsheets can only perform calculations – summing, counting, averages, etc. -- on numbers. The content in columns G,H, I and K are left-justified, meaning they are text. If your numbers are left-justified, which can happen when downloading data from the internet, you would have to do a bit of reformatting or cleaning. Fortunately, this data set is clean, tidy, and good to go.

The dates in column A are sorted in descending order with the most recent period at the top. As this table was downloaded on Wednesday Oct. 28, 2020, that is the day of the most recent entry in cell A2.

The three cells to the right, represent the components of the date: day, month, and year. Again, this is excellent for analysis because we may want to group days or certain months to determine the virus's most deadly period.

Sticking with numbers, you will also notice that column J contains the country's population figure. This will come in handy when we perform calculations such as deaths rate which allow for country-to-country comparisons.

There is also a lot this table cannot tell us: Where are the country's hotspots? What parts of the population are being the most affected? And are the number of tests affecting the cases and deaths? What about neighbourhoods and schools which are now the subject of a lot of worry as B.C., Ontario and Quebec are caught in the feared second wave which, while not as deadly as the first wave, is still concerning.

You won't find these answers in the nerd box or the table we have just downloaded. If nothing is available on the website, you may have to try to connect with the individual responsible for the data. The rules governing open-data policies work off the premise that data should be open by default. Governments typically provide mere subsets of larger datasets they possess. So, if something appears to be missing, there is a chance it may be available for the asking. Don't be shy.

Another key question is who updates and verifies the data? What government agencies send their death and infection numbers to the centre? Indeed, there's checklist of questions that should be answered

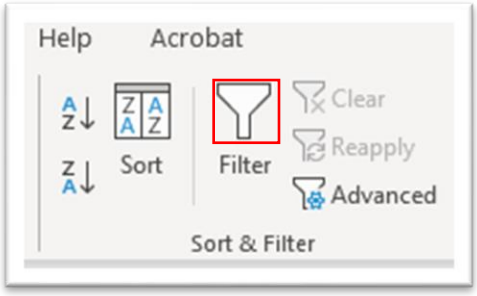
before we go much further, which you'll find on page 29 of The Data Journalist.

Notice how much legwork we have done BEFORE any data manipulation. Taking time to understand your data is crucial. Data journalists also call this process interviewing your data, the same way you would a human source.

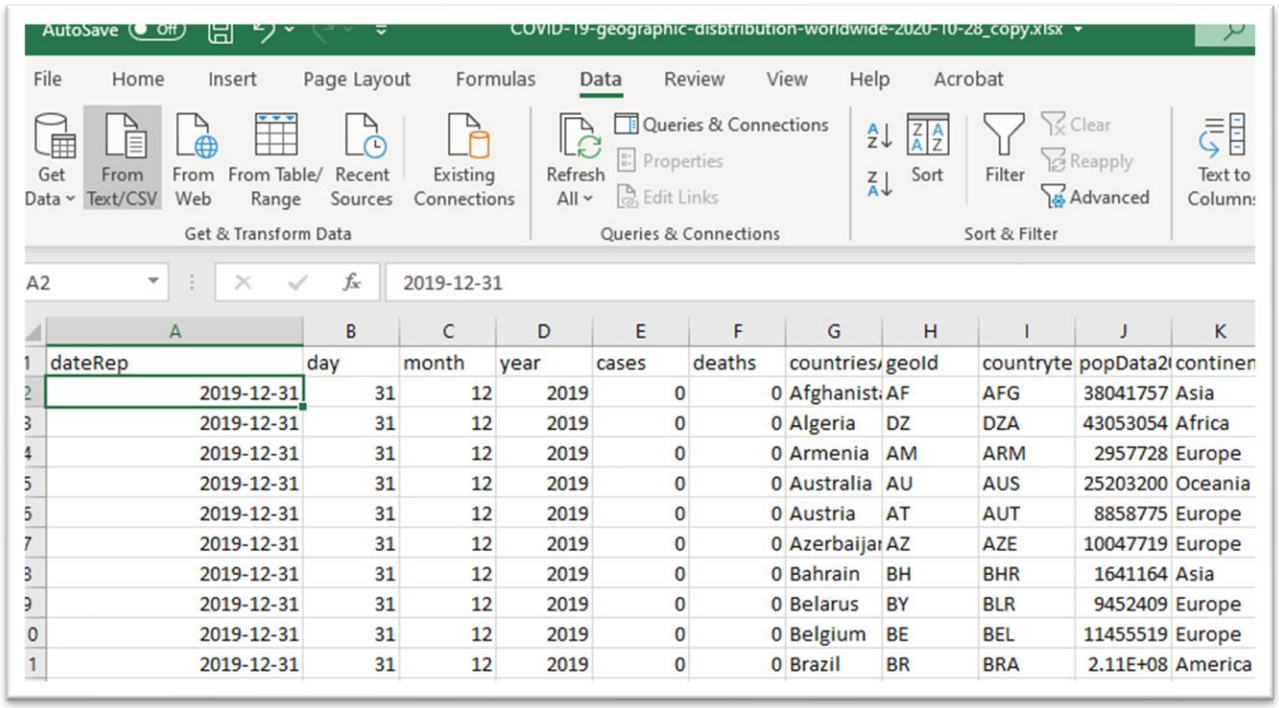
Sorting key columns of information is one of the first things we do during this interviewing process.

Sorting

The dates in column A are already in descending order. However, if we wanted to sort in ascending order to determine when the centre began collecting the COVID-19 information, we would highlight cell A2, go to "Data" on your menu at the top, and then scroll over to the "Sort & Filter" section.



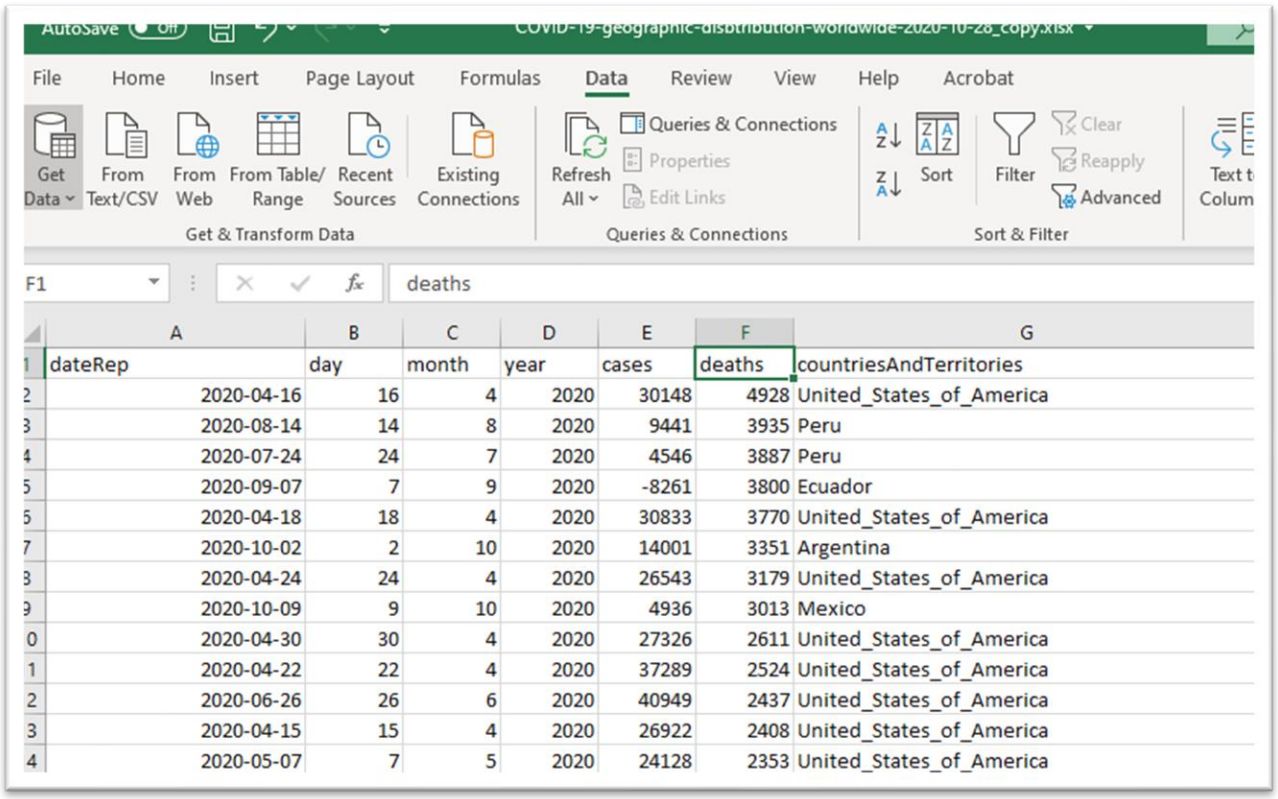
A to Z is descending order.



	A	B	C	D	E	F	G	H	I	J	K
1	dateRep	day	month	year	cases	deaths	countries/	geoId	countryte	popData2(	continer
2	2019-12-31	31	12	2019	0	0	Afghanist	AF	AFG	38041757	Asia
3	2019-12-31	31	12	2019	0	0	Algeria	DZ	DZA	43053054	Africa
4	2019-12-31	31	12	2019	0	0	Armenia	AM	ARM	2957728	Europe
5	2019-12-31	31	12	2019	0	0	Australia	AU	AUS	25203200	Oceania
6	2019-12-31	31	12	2019	0	0	Austria	AT	AUT	8858775	Europe
7	2019-12-31	31	12	2019	0	0	Azerbaijar	AZ	AZE	10047719	Europe
8	2019-12-31	31	12	2019	0	0	Bahrain	BH	BHR	1641164	Asia
9	2019-12-31	31	12	2019	0	0	Belarus	BY	BLR	9452409	Europe
10	2019-12-31	31	12	2019	0	0	Belgium	BE	BEL	11455519	Europe
11	2019-12-31	31	12	2019	0	0	Brazil	BR	BRA	2.11E+08	America

We can see that the data started coming into the centre on December 31.

Now let's sort the "deaths" column in descending order.



	A	B	C	D	E	F	G
1	dateRep	day	month	year	cases	deaths	countriesAndTerritories
2	2020-04-16	16	4	2020	30148	4928	United_States_of_America
3	2020-08-14	14	8	2020	9441	3935	Peru
4	2020-07-24	24	7	2020	4546	3887	Peru
5	2020-09-07	7	9	2020	-8261	3800	Ecuador
6	2020-04-18	18	4	2020	30833	3770	United_States_of_America
7	2020-10-02	2	10	2020	14001	3351	Argentina
8	2020-04-24	24	4	2020	26543	3179	United_States_of_America
9	2020-10-09	9	10	2020	4936	3013	Mexico
10	2020-04-30	30	4	2020	27326	2611	United_States_of_America
11	2020-04-22	22	4	2020	37289	2524	United_States_of_America
12	2020-06-26	26	6	2020	40949	2437	United_States_of_America
13	2020-04-15	15	4	2020	26922	2408	United_States_of_America
14	2020-05-07	7	5	2020	24128	2353	United_States_of_America

The U.S. experienced the deadliest day on April 16. Again, this number is an aggregate. For instance, we are still lacking more precise information such as the states in which these deaths occurred, age group, gender, etc.

Still, the aggregate number is a good starting point for research and storytelling.

We can also sort column E to conduct the same kind of analysis for cases.

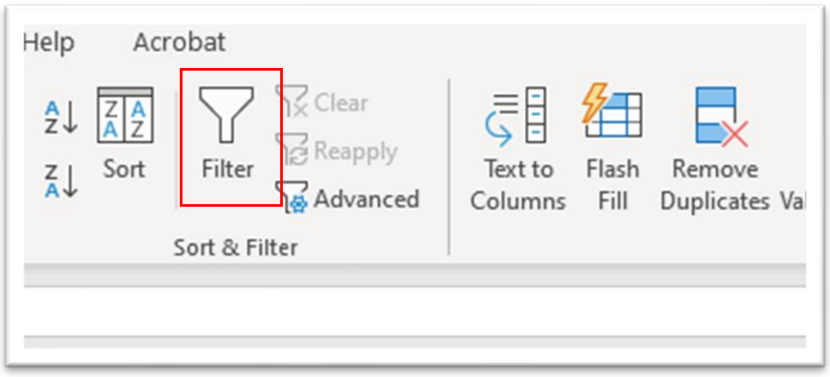
Sorting is useful when dealing with columns containing numbers, because we can go from most to least and vice-versa. It's only useful to sort columns with text if there's a reason to organize information alphabetically, like country names in column G.

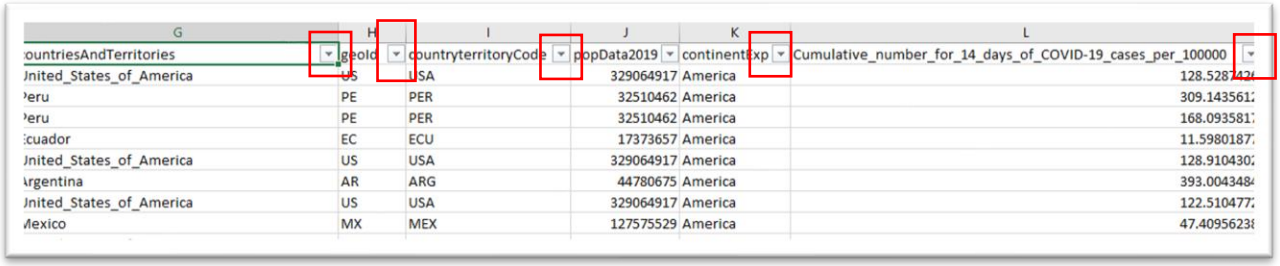
We have reached the limit of sorting.

It’s time to take a slightly deeper dive into the data by using our filter.

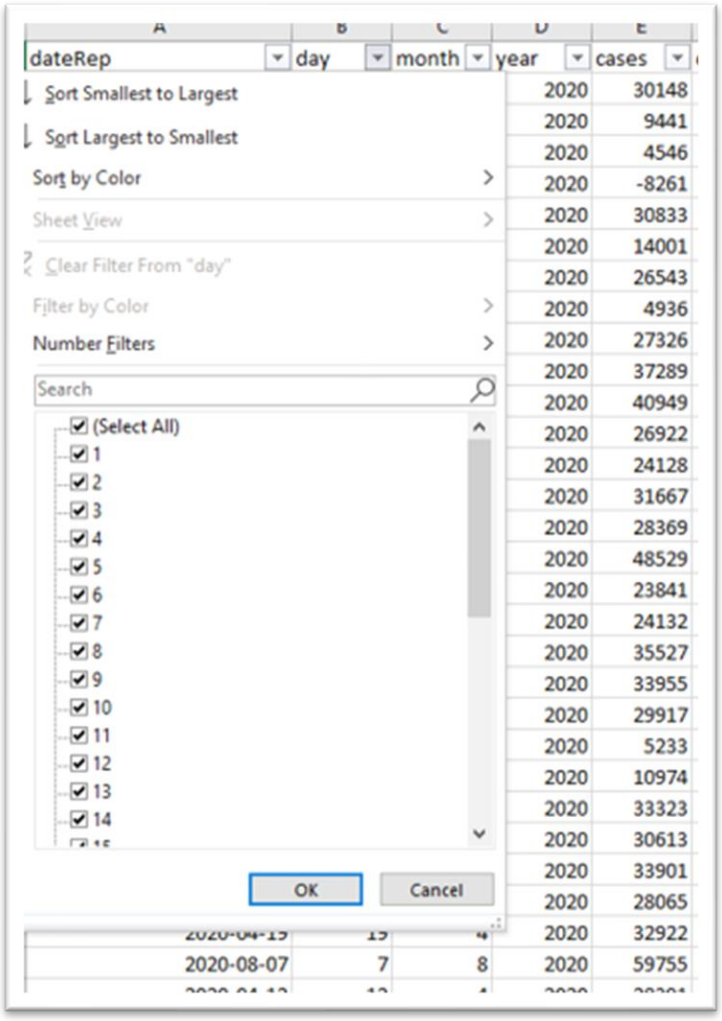
Filtering

Click any cell in the first row and select the filter icon in the menu’s “Sort & Filter” section.

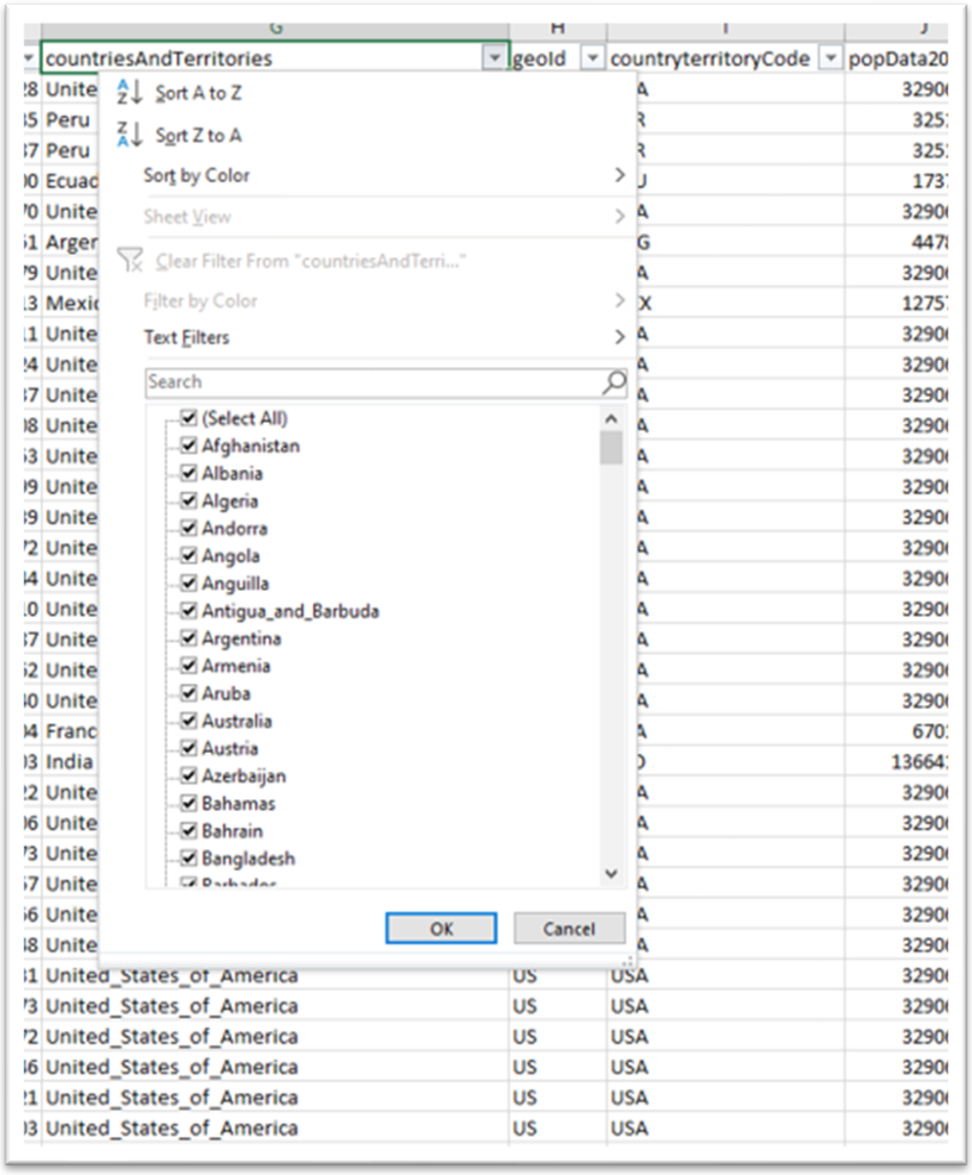




Now each column has an arrow that when selected displays the numbers in ascending order.

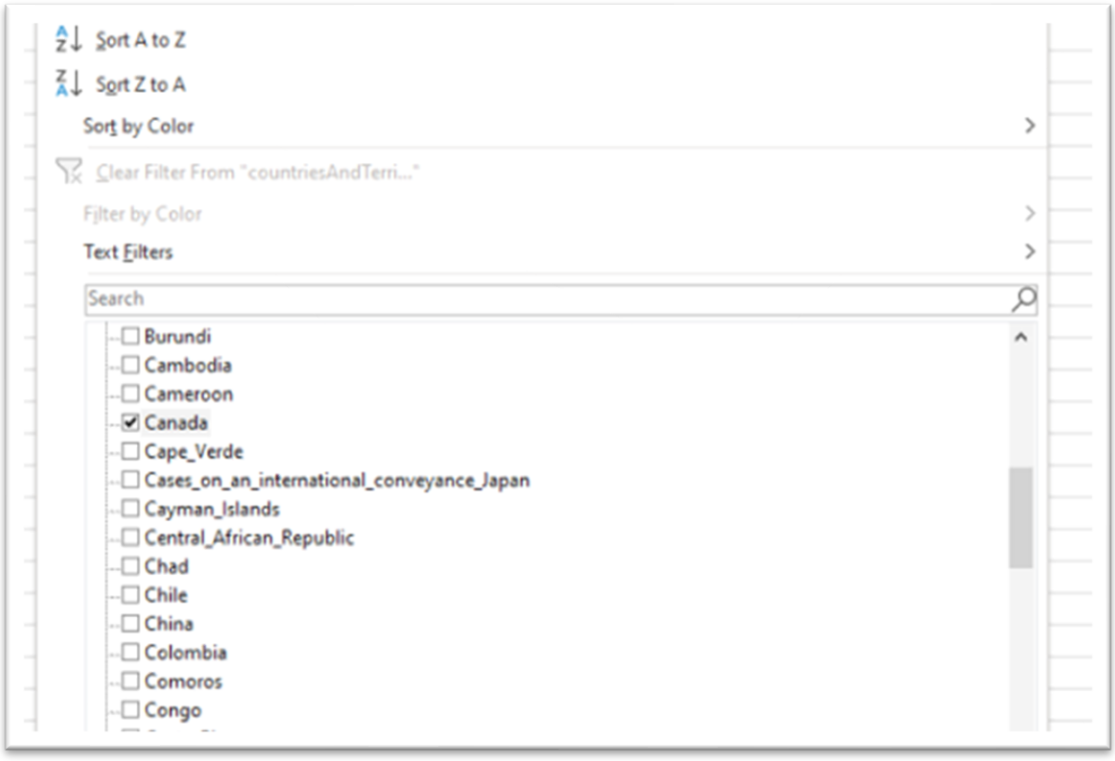


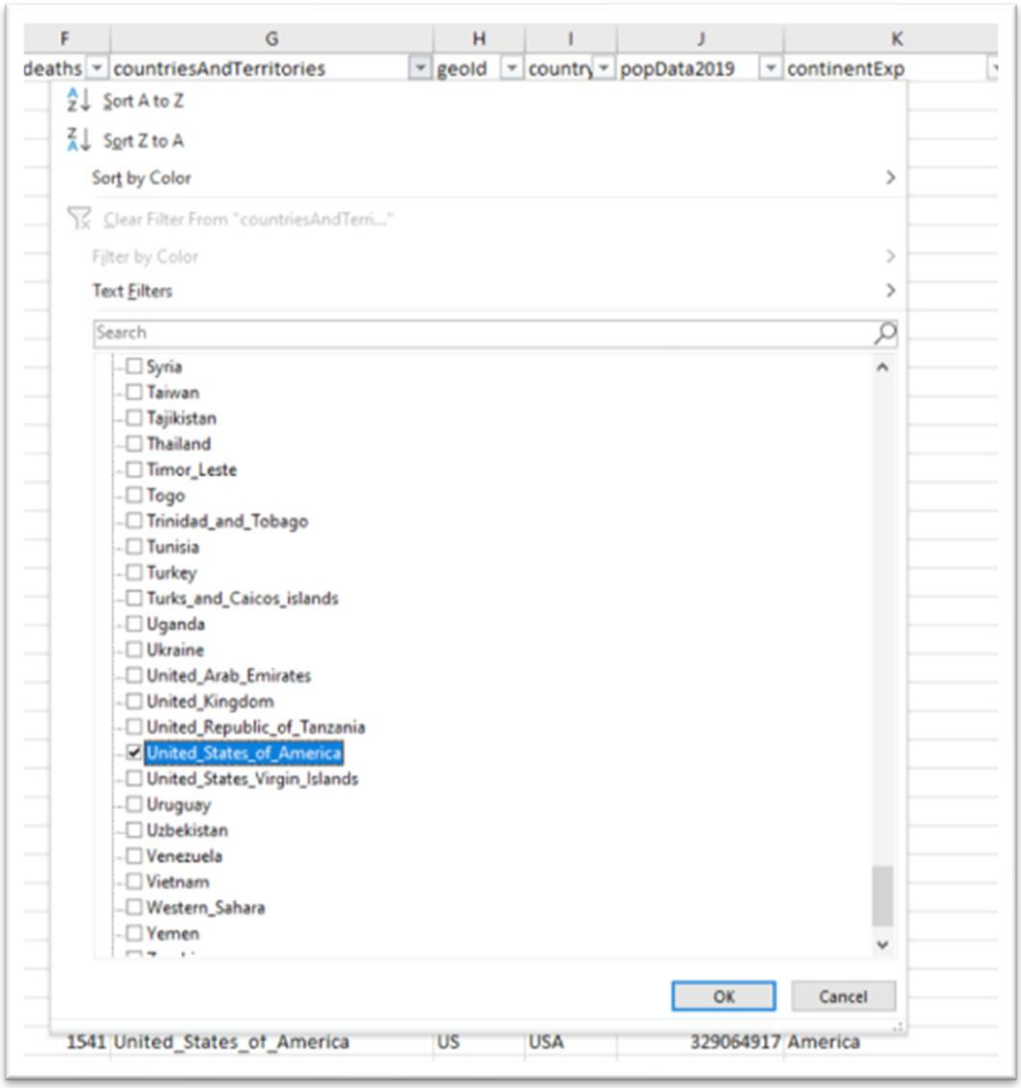
and text in ascending order.

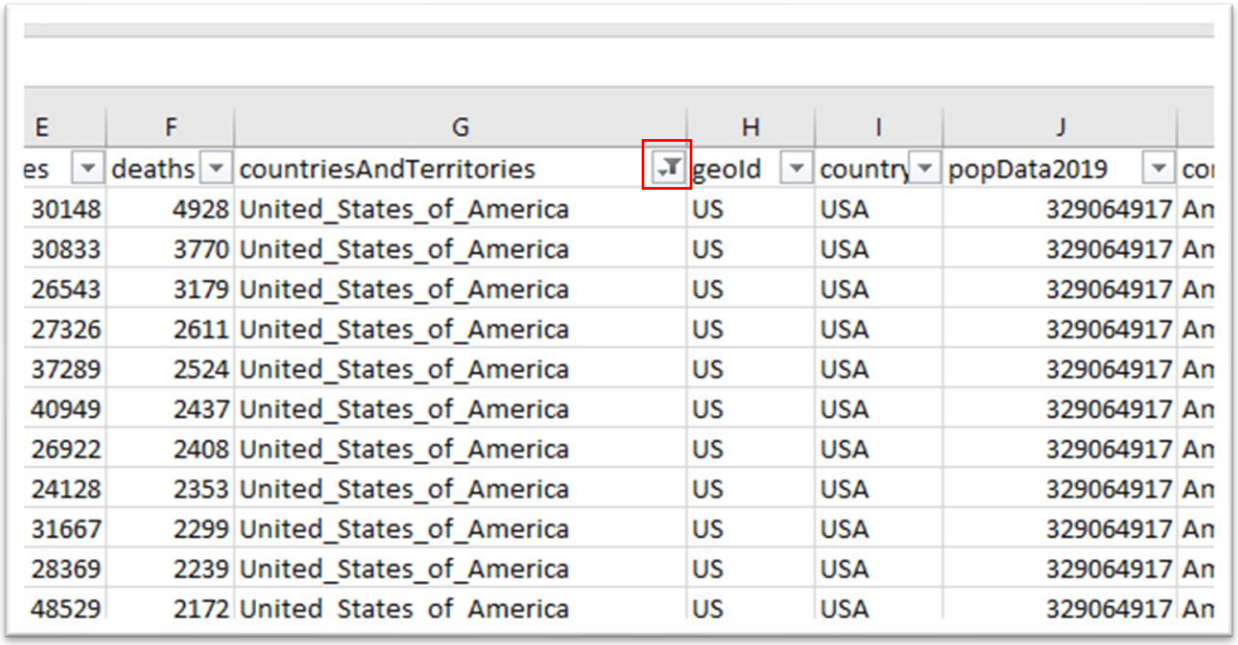


Filtering allows us to only select certain countries such as Canada and the United States.

This can be accomplished by activating the filter in column G, clearing the selection and choosing Canada and the United States.



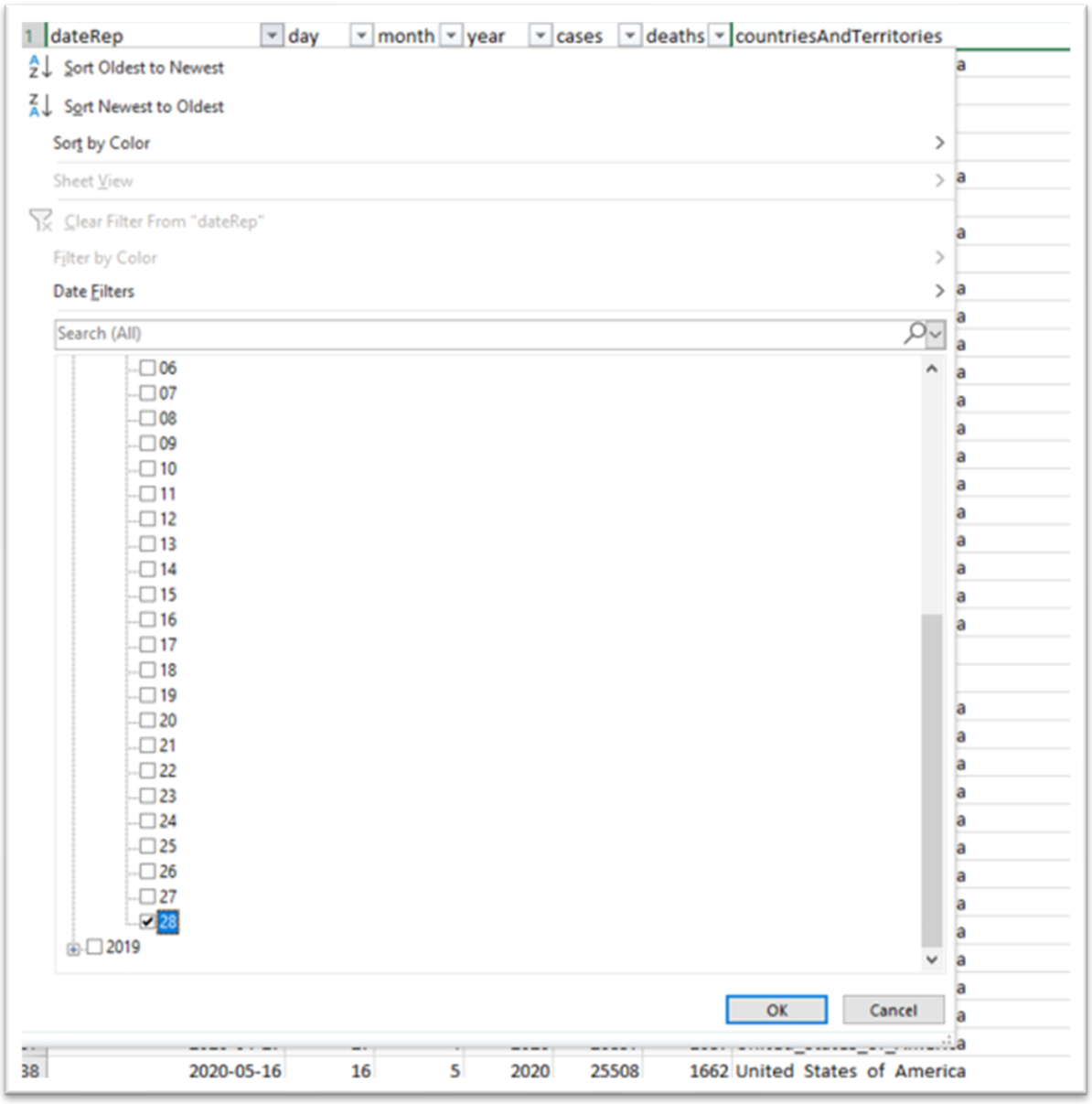

F
G
H
I
J
K
deaths
countriesAndTerritories
geoId
countr
popData2019
continentExp
Sort A to Z
Sort Z to A
Sort by Color
Clear Filter From "countriesAndTerri..."
Filter by Color
Text Filters
Search
Syria
Taiwan
Tajikistan
Thailand
Timor_Leste
Togo
Trinidad_and_Tobago
Tunisia
Turkey
Turks_and_Caicos_islands
Uganda
Ukraine
United_Arab_Emirates
United_Kingdom
United_Republic_of_Tanzania
United_States_of_America
United_States_Virgin_Islands
Uruguay
Uzbekistan
Venezuela
Vietnam
Western_Sahara
Yemen
OK
Cancel
1541 United_States_of_America US USA 329064917 America



E	F	G	H	I	J	
es	deaths	countriesAndTerritories	geoId	country	popData2019	co
30148	4928	United_States_of_America	US	USA	329064917	An
30833	3770	United_States_of_America	US	USA	329064917	An
26543	3179	United_States_of_America	US	USA	329064917	An
27326	2611	United_States_of_America	US	USA	329064917	An
37289	2524	United_States_of_America	US	USA	329064917	An
40949	2437	United_States_of_America	US	USA	329064917	An
26922	2408	United_States_of_America	US	USA	329064917	An
24128	2353	United_States_of_America	US	USA	329064917	An
31667	2299	United_States_of_America	US	USA	329064917	An
28369	2239	United_States_of_America	US	USA	329064917	An
48529	2172	United_States_of_America	US	USA	329064917	An

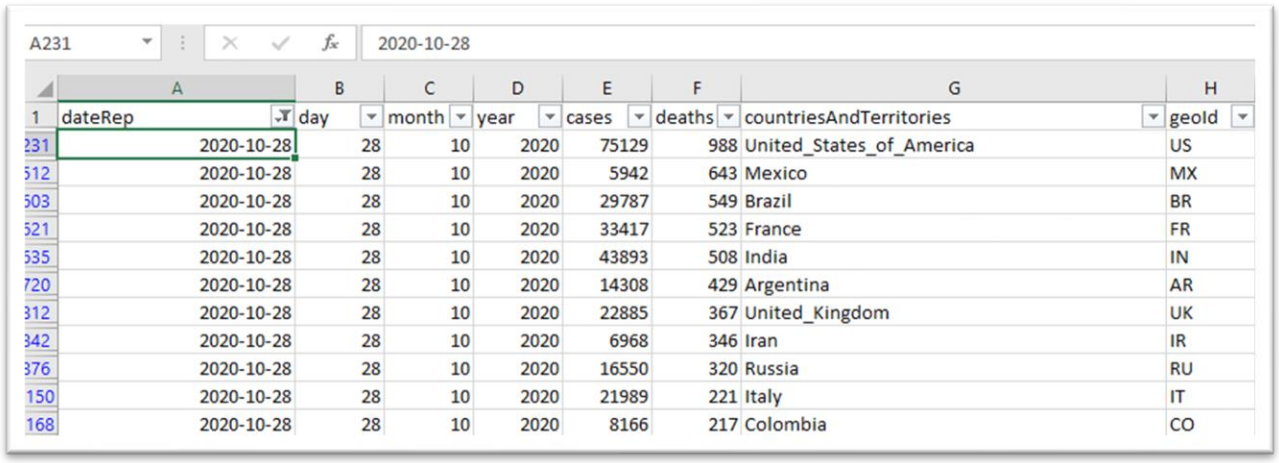
The funnel icon in column G1 tells us that “countriesAndTerritories” is now filtered.

We can filter the date column for the most recent day.



Filtering by day is trickier, because we have to drill down to the actual day by clicking the “+” signs next to the Years and months in order to

get to days, as you can see in the screen grab above.



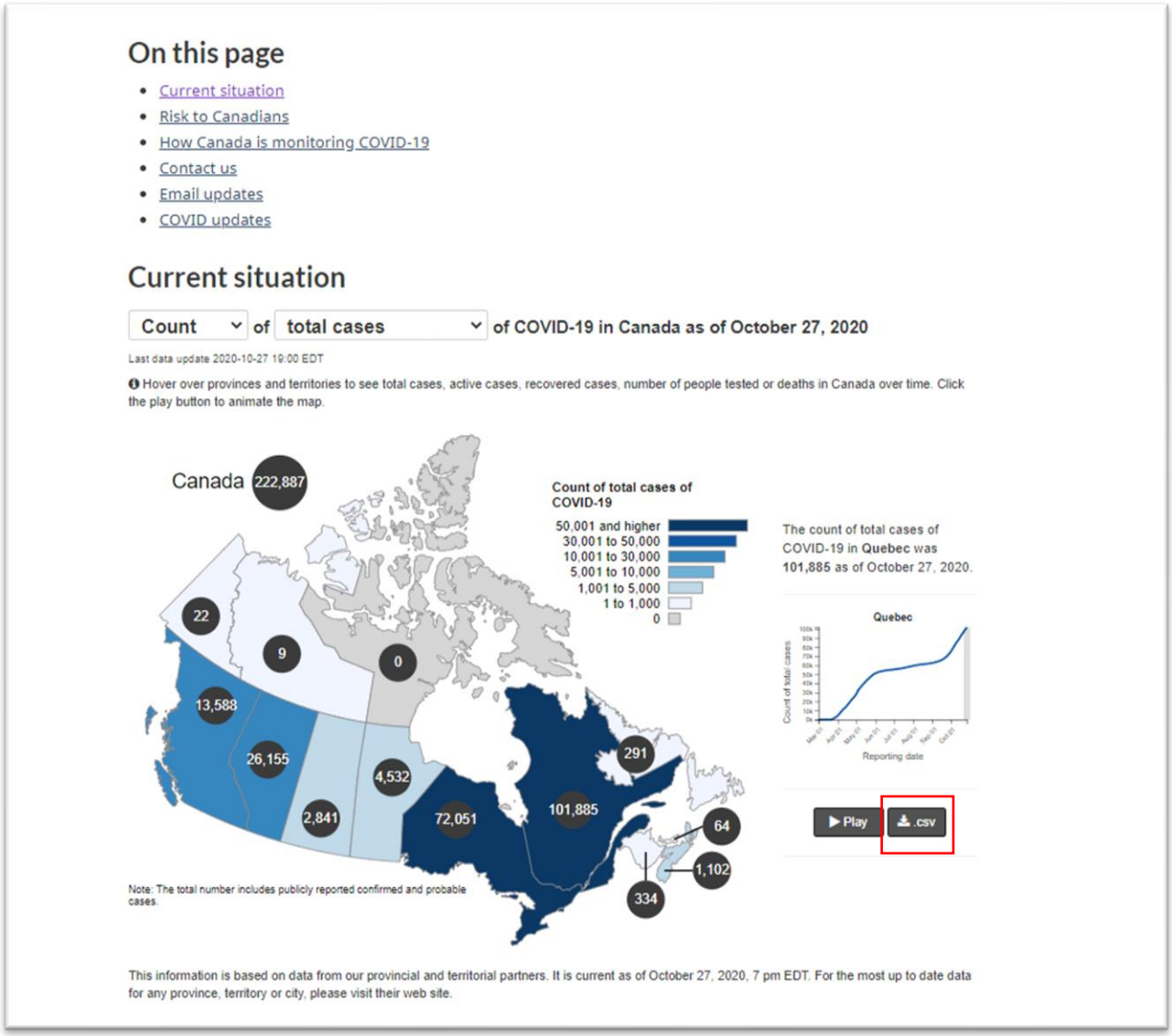
A231 | 2020-10-28

	A	B	C	D	E	F	G	H
1	dateRep	day	month	year	cases	deaths	countriesAndTerritories	geoId
231	2020-10-28	28	10	2020	75129	988	United_States_of_America	US
512	2020-10-28	28	10	2020	5942	643	Mexico	MX
503	2020-10-28	28	10	2020	29787	549	Brazil	BR
521	2020-10-28	28	10	2020	33417	523	France	FR
535	2020-10-28	28	10	2020	43893	508	India	IN
720	2020-10-28	28	10	2020	14308	429	Argentina	AR
312	2020-10-28	28	10	2020	22885	367	United_Kingdom	UK
342	2020-10-28	28	10	2020	6968	346	Iran	IR
376	2020-10-28	28	10	2020	16550	320	Russia	RU
150	2020-10-28	28	10	2020	21989	221	Italy	IT
168	2020-10-28	28	10	2020	8166	217	Colombia	CO

Now we have a snapshot for the United States on Oct 28, 2020, six days before a U.S. presidential election in which COVID-19 has become a major issue. Of course, raw numbers only tell part of the story. Using a death rate would be better, which is possible to calculate because we have the population numbers in column J.

We will cover rates in the next tutorial.

You can practice the steps in this tutorial with data from the federal government's COVID-19 website.



Download the “csv” file.