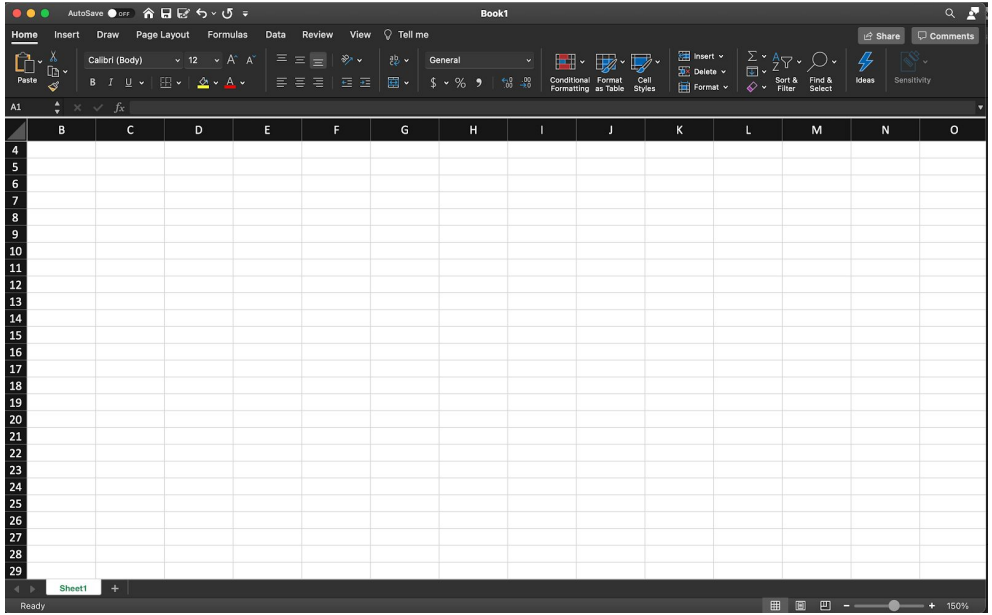
Data Analysis Fundamentals : Excel

Instructor: Sarah

UNIVERSITY OF TORONTO

What is Excel?

- Microsoft Excel is an application that is used for recording, analyzing and visualizing data. It is in the form of a spreadsheet



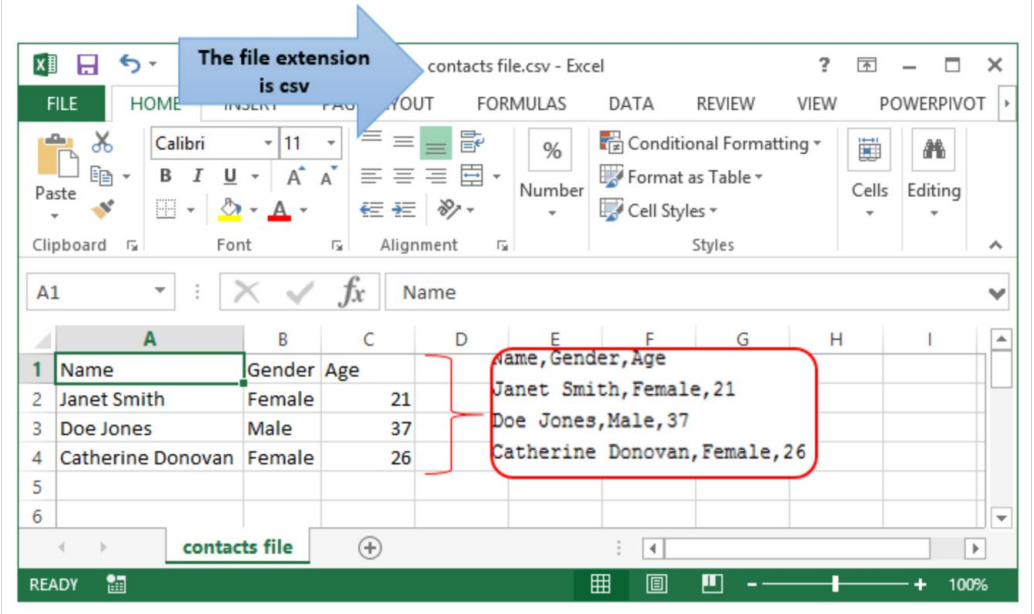
Why use Excel?

- Allows you to visualize data and present it in the form of charts, tables and data bars
- Proves to be a great platform to perform various mathematical calculations on large datasets
- Various features in excel like searching, sorting, filtering make it easier for you to play with data
- Provides good security for your data. Excel files have the feature of password-protection, this way your information is safe

.xls/.xlsx vs .csv files

Both .xls and .csv files help store data in tabular format

- CSV files does not have a limit of saving data while Excel files have the limitation of storing the data sets
- CSV files are faster and also consumes less memory whereas Excel consumes more memory while importing data
- CSV files can be opened with any text editor in windows while Excel files can't be opened with text editors



future tense

An Outdated Version of Excel Led the U.K. to Undercount COVID-19 Cases

BY WHITNEY TESI

OCT 07, 2020 • 3:47 PM

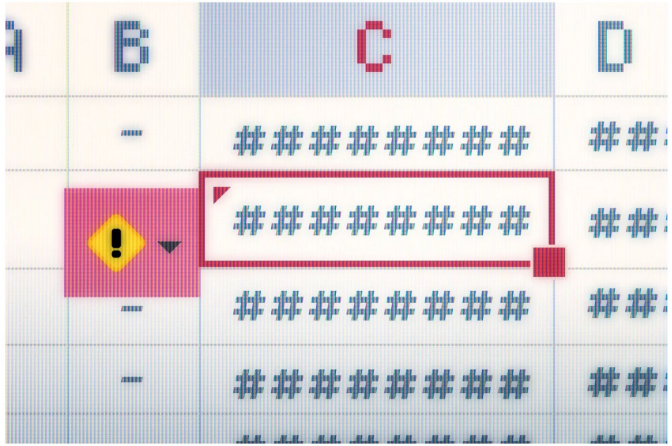


Photo illustration by Natalie Matthews-Ramo/Slate. Photo by Thomaslenne/iStock/Getty Images Plus.

Over the weekend, the U.K. saw its number of COVID-19 cases spike in large part because of a Microsoft Excel error.

Between Sept. 25 and Oct. 2, 15,841 cases went unreported in the government database. According to Public Health England, the files with positive results sent by the National Health Service's test-and-trace system exceeded the maximum size. Because of the error, which was discovered Friday, nearly 48,000 people who had had contact with those who tested positive weren't traced. All told, it could have put many lives at risk.

And you thought you had had frustrating experiences with Excel.

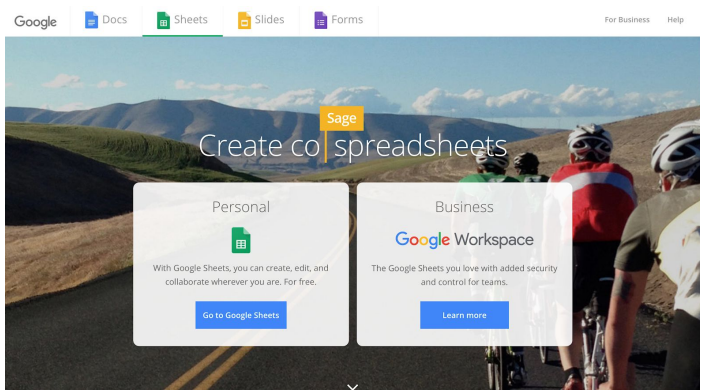
According to the BBC, the error was caused by the fact that Public Health England developers stored the test results in the file format known as .XLS. These .XLS files were then sent to the NHS after uploading to a central system. .XLS is an outdated file format, however, and each spreadsheet can have only 65,000 rows. By contrast, the .XLSX file format, which was first released in 2007, allows for more than 1 million rows. Because of the limited number of rows, each spreadsheet could contain about 1,400 cases, leaving excess cases off the file altogether. Although the issue was reportedly fixed by splitting the files into smaller batches, many are slamming Public Health England. "Why are critical databases in a national pandemic posted on Excel spreadsheets?" Jonathan Ashworth, the Labour Party's shadow health secretary, said. "Why aren't they using specialist data-based software?"

This isn't the first time that misuse of Excel has caused a massive error. Although the platform can look simple, one basic error can end up creating ghastly results for entire companies. Here are six other supremely costly spreadsheet mistakes.

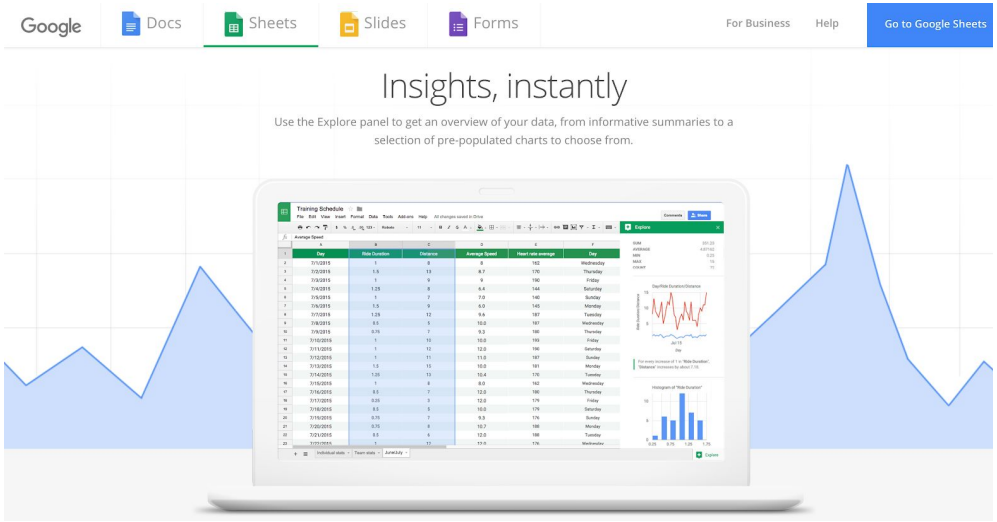
https://slate.com/technology/2020/10/u-k-covid-19-spike-caused-by-microsoft-excel-error.html

Google Sheets

Google Sheets is a spreadsheet program included as part of the free, web-based Google Docs Editors suite offered by Google.

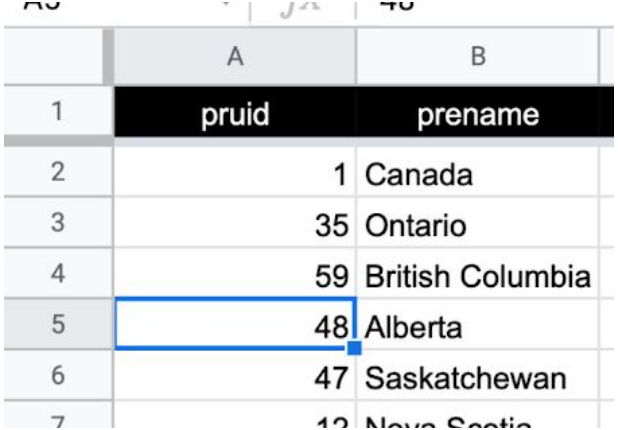


https://www.google.ca/sheets/about/



Familiarization with Spreadsheets

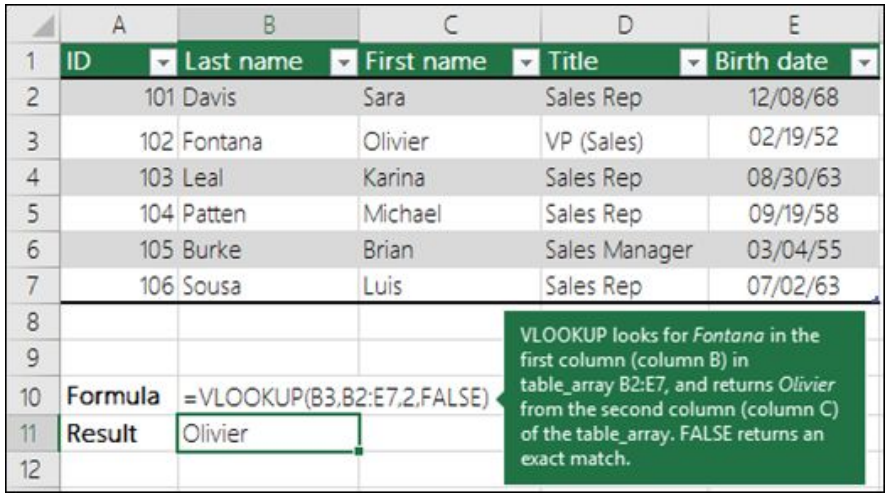
- Spreadsheets are displayed in a grid layout
- Column headings = letters
  - To highlight an entire Column, click on any of the letters
- Numbers = rows
  - To highlight an entire Row, click on any of the numbers
- A Cell is a letter combined with a number
  - So if you combine the A column with Row 5,
    you get Cell A5



	A	B
1	pruid	prename
2	1	Canada
3	35	Ontario
4	59	British Columbia
5	48	Alberta
6	47	Saskatchewan

Basic Functions in Excel

- Functions are predefined formulas and are already available in Excel
- Basic functions:
  - Multiply, Divide, Add and Subtract
- Most often used functions:
  - SUM
  - AVERAGE
  - MAX, MIN
  - COUNT, IF
  - VLOOKUP



	A	B	C	D	E
1	ID	Last name	First name	Title	Birth date
2	101	Davis	Sara	Sales Rep	12/08/68
3	102	Fontana	Olivier	VP (Sales)	02/19/52
4	103	Leal	Karina	Sales Rep	08/30/63
5	104	Patten	Michael	Sales Rep	09/19/58
6	105	Burke	Brian	Sales Manager	03/04/55
7	106	Sousa	Luis	Sales Rep	07/02/63
8					
9					
10	Formula	=VLOOKUP(B3,B2:E7,2,FALSE)			
11	Result	Olivier			
12					

VLOOKUP looks for *Fontana* in the first column (column B) in table_array B2:E7, and returns *Olivier* from the second column (column C) of the table_array. FALSE returns an exact match.

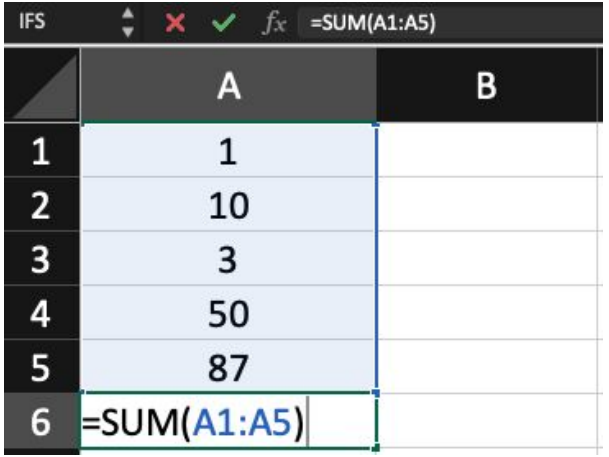
SUM, AVERAGE

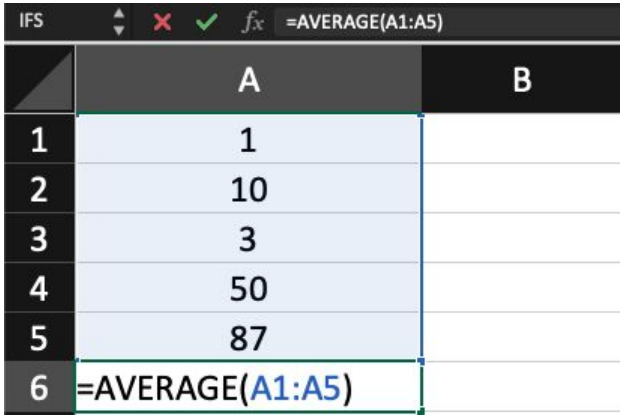
=SUM(A1:A5)

: for example, if the range A1:A5 contains numbers, the formula returns the sum of those numbers

=AVERAGE(A1:A5)

: for example, if the range A1:A5 contains numbers, the formula returns the average of those numbers





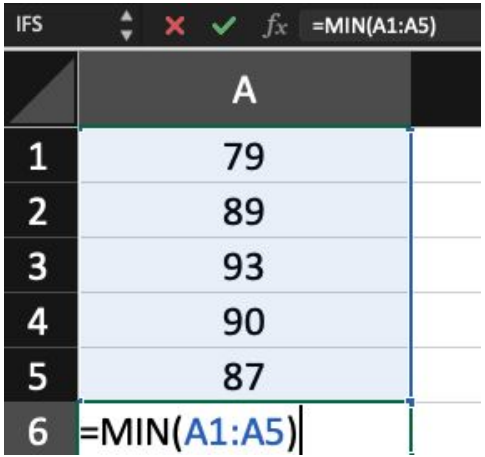
MIN, MAX

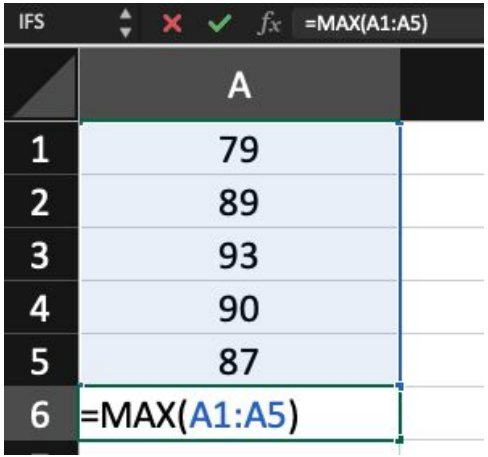
=MIN(A1:A5)

: for example, if the range A1:A5 contains numbers, the formula returns the min of those numbers

=MAX(A1:A5)

: for example, if the range A1:A5 contains numbers, the formula returns the max of those numbers





COUNT, IF

=COUNT(A1:A5)

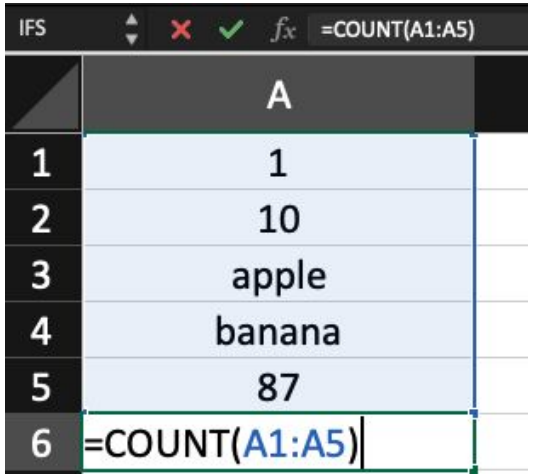
: counts the number of cells that contain numbers in cells A1 through A5

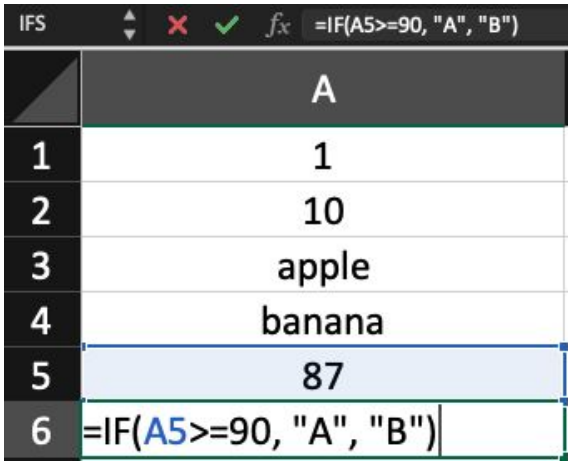
=IF([logical_test],[value_if_true],[value_if_false])

: returns [value_if_true] if [logical_test] is true, and [value_if_false] if [logical_test] is false

examples

=IF(A5>=90, “A”, “B”)





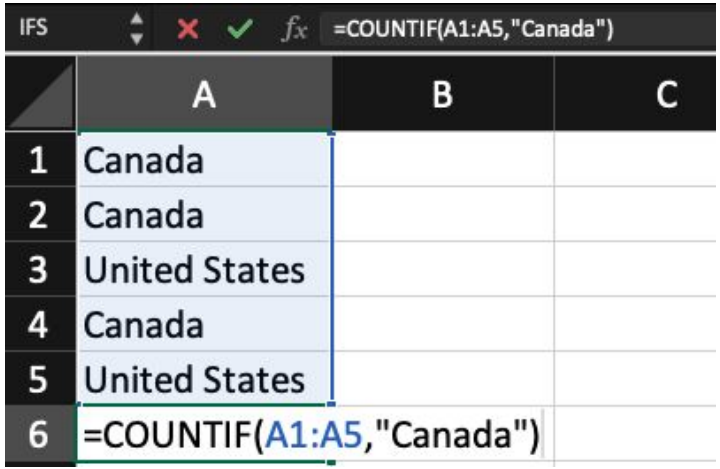
COUNTIF

=COUNTIF(range, criteria)

=COUNTIF(where you want to look?, what do you want to look for?)

=COUNTIF(A1:A5,"Canada")

: count cells that match criteria



Q: how can we make this function more dynamic?

=COUNTIF(A1:A5,"Canada")

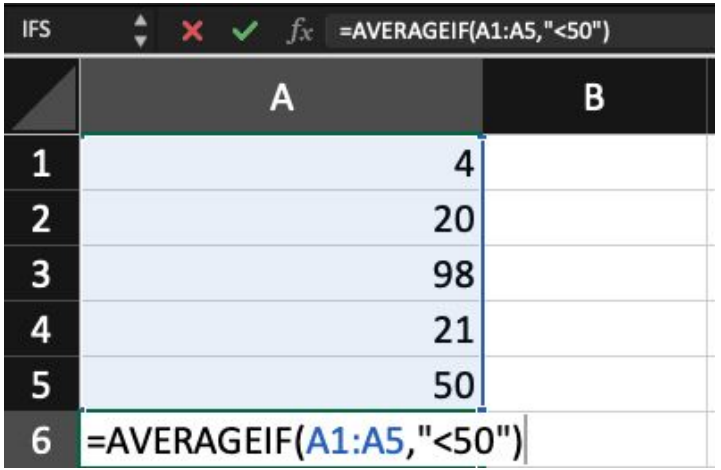
AVERAGEIF

=AVERAGEIF(range, criteria)

=AVERAGEIF(where do you want to look?, what is the criteria that you are looking for)

=AVERAGEIF(A1:A5,"<50")

: get the average of numbers that meet criteria

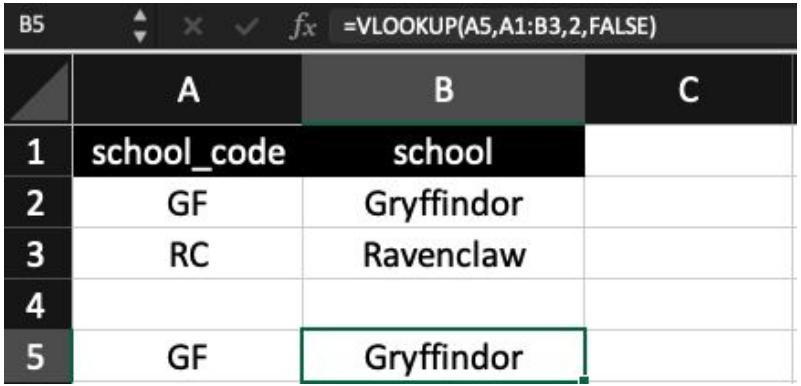


VLOOKUP

=VLOOKUP(lookup_ value, table_array, column_index_num, [match_type])

=VLOOKUP(what you want to look at,
where you want to look for it,
the column number in the range containing the value to return,
return an approximate or approximate match - True, exact match - False)

: when you need to find things in a table or a range by row

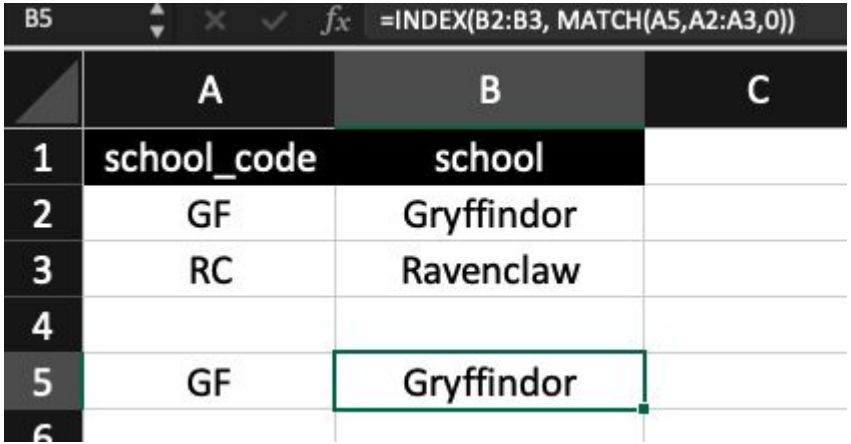


INDEX MATCH

=INDEX(return_array, MATCH(lookup_value, lookup_array, [match_type]))

=INDEX(what you want to return, MATCH(what you want to look at, where you want to look for it, return less than:1, exact match:0, greater than:-1))

:use two functions INDEX and MATCH to find things in a table or a range by row

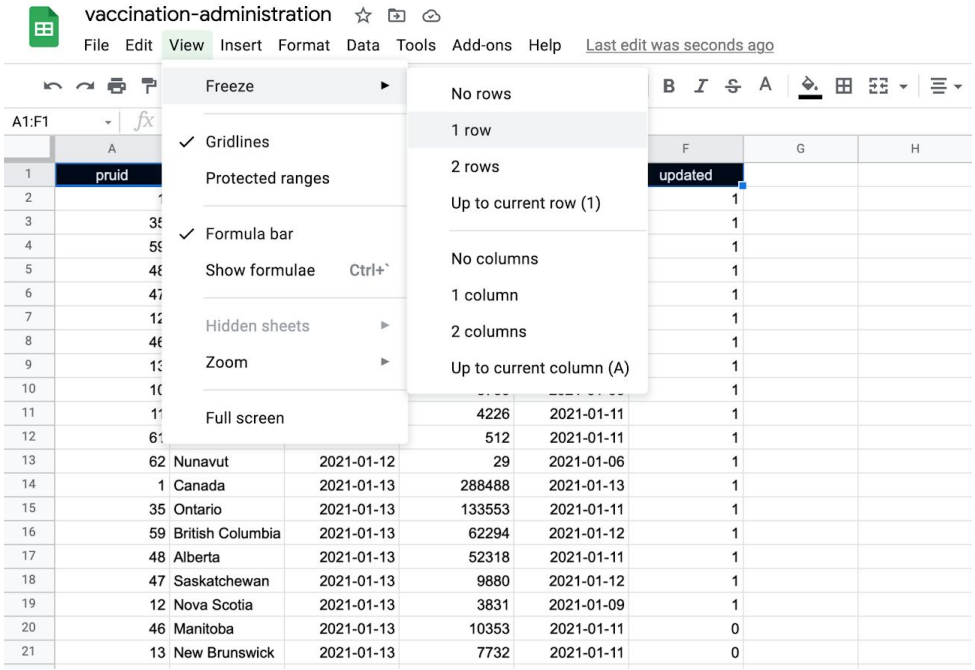


Freeze Panes in Excel

To see the data better, it is nice to freeze the top row to make it visible at the top no matter how far you go down

1. View
   -> Freeze 1 Row

cf. you can also use the freeze pane or freeze first column function as well



	pruid					updated
2	1					1
3	35					1
4	59					1
5	48					1
6	47					1
7	12					1
8	46					1
9	13					1
10	10					1
11	11			4226	2021-01-11	1
12	61			512	2021-01-11	1
13	62	Nunavut	2021-01-12	29	2021-01-06	1
14	1	Canada	2021-01-13	288488	2021-01-13	1
15	35	Ontario	2021-01-13	133553	2021-01-11	1
16	59	British Columbia	2021-01-13	62294	2021-01-12	1
17	48	Alberta	2021-01-13	52318	2021-01-11	1
18	47	Saskatchewan	2021-01-13	9880	2021-01-12	1
19	12	Nova Scotia	2021-01-13	3831	2021-01-09	1
20	46	Manitoba	2021-01-13	10353	2021-01-11	0
21	13	New Brunswick	2021-01-13	7732	2021-01-11	0