Creating new variables by calculation in Excel

effini

THE DATA LAB
value from data

Data Skills FOR WORK

Data Education in Schools

Learning intentions

We will be learning how to create new variables in Excel, specifically to,

- understand what it means to **create a new variable by performing calculations** using existing data
- how to create simple new variables by performing a calculation in Excel
- understand the **concept of conditional statements** for creating new variables.

Definition

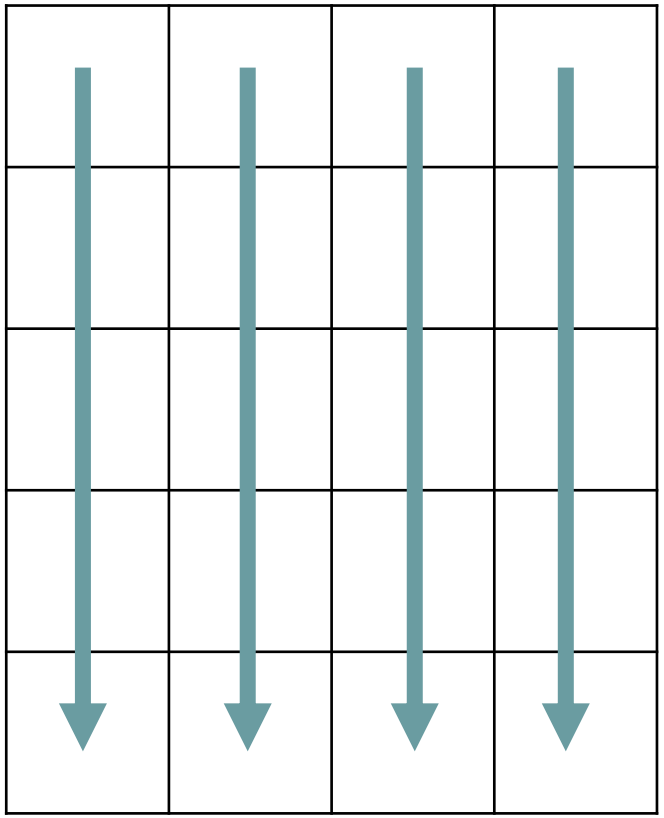
Variable

A column of data that only contains one data type

Variables

In data science **columns of data** are often referred to as **variables**.

Each variable only **contains one data type** such as numbers, text, date or time.

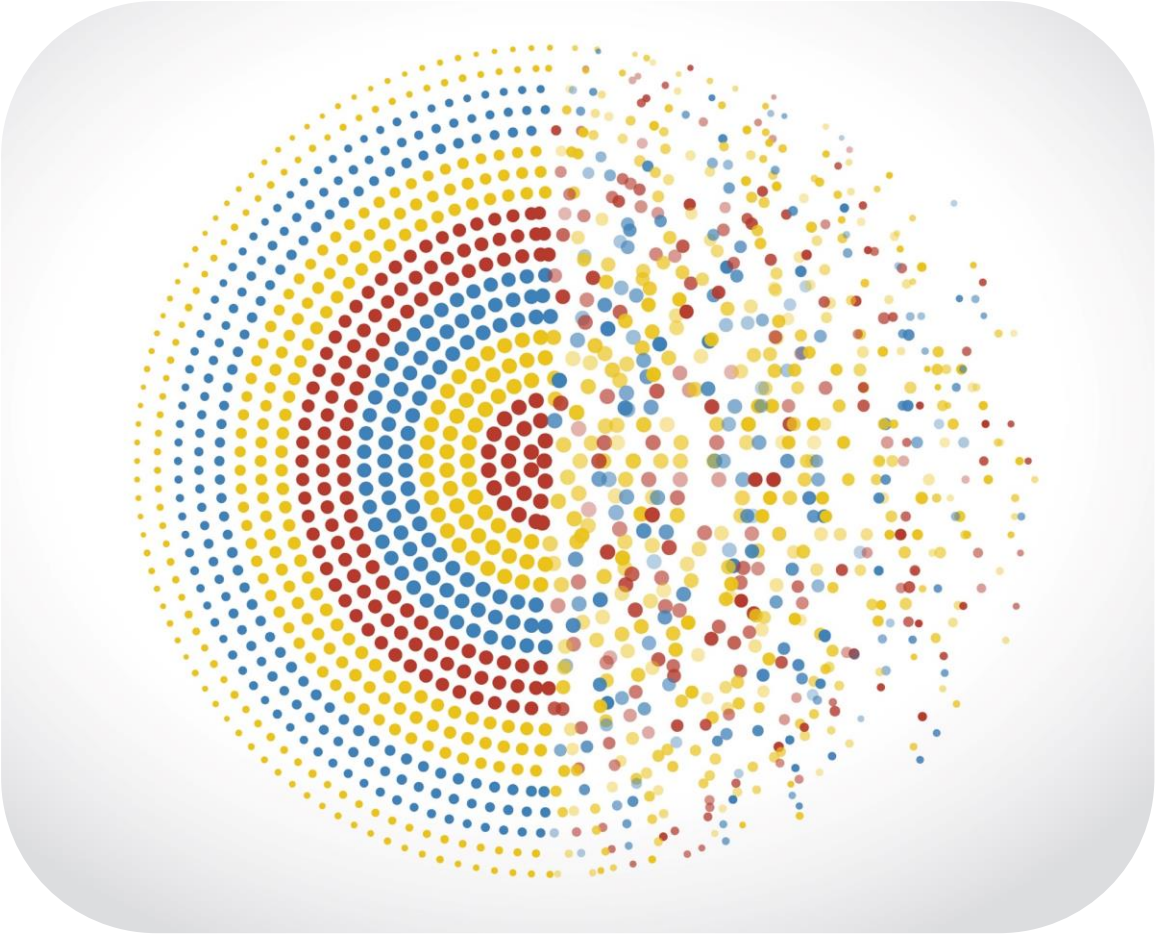


Background

When a data scientist receives a dataset, it is rarely in a format that will allow them to start analysis right away. It often needs to be **manipulated first.**

Variables (columns) can be manipulated in a dataset by **reordering** or **selecting** them.

You can also manipulate datasets by **creating new variables of data** in the dataset.



Why this is important?

Some benefits of creating new variables of data are,

Allows you to **understand more** about the data

Allows you to **perform calculations** such as adding, subtracting, dividing

You can **compare different** data items

Example

What are the variables in this dataset?

event	time	runner
Marathon (men)	2:01:39	Eliud Kipchoge
Marathon (women)	2:14:04	Brigid Kosgei

The **variables** are,

event, time, runner

Your turn…

What do you think the **variables** are in this dataset?

event	athlete	nationality	date	location
Long jump (men)	Mike Powell	USA	30 Aug 1991	Tokyo
Long jump (women)	Heike Drechsler	Germany	13 Feb 1988	Vienna
Triple jump (men)	Jonathan Edwards	United Kingdom	7 Aug 1995	Gothenburg
Triple jump (women)	Yulimar Rojas	Venezuela	21 Feb 2020	Madrid

Variable

A column of data that only contains one data type

Your turn…

What are the **variables** in this dataset?

The variables are

- event
- athlete
- nationality
- date
- location

event	athlete	nationality	date	location
Long jump (men)	Mike Powell	USA	30 Aug 1991	Tokyo
Long jump (women)	Heike Drechsler	Germany	13 Feb 1988	Vienna
Triple jump (men)	Jonathan Edwards	United Kingdom	7 Aug 1995	Gothenburg
Triple jump (women)	Yulimar Rojas	Venezuela	21 Feb 2020	Madrid

Next steps

Complete **questions 1 to 4**
in **section 1** of the
‘Creating new calculated variables in Excel’
workbook

Definition

Create by calculation

To perform calculations on one or more data items in a dataset to make a new variable

Creating new variables

Some of the most common calculations performed on columns of data to **create new variables** are,

- Adding
- Subtracting
- Multiplying
- Dividing
- Average (mean, mode or median)

A	B	A+B	A-B	A/B
5	1	6	4	5
10	2	12	8	5

Creating new variables

You can also create new variables by comparing data items such as,

- What is the **minimum** of the variables?
- What is the **maximum** of the variables?
- Deciding if a variable is **greater than, less than** or **equal** to a number

A	B	Minimum of A and B	Maximum of A and B	Is column A greater than 11?
5	1	1	5	FALSE
10	2	2	10	FALSE
15	3	3	15	TRUE
20	4	4	20	TRUE

Show me…

Add VAT of 20% to these items. (price_inc_VAT = price x 1.2)

item	price	price_inc_VAT
Hat	£5.00	£6.00
Top	£10.00	£12.00
Jumper	£15.00	£18.00
Socks	£2.50	£3.00

This is a new variable in the dataset

Show me…

What is the **maximum distance** that the athletes threw their javelins over their 3 attempts?

name	throw_1	throw_2	throw_3
Jack	25.6	32.8	45.5
Jill	46.2	33.5	21.0

Maximum of each row

max
45.5
46.2

Example

How long did it take these people to run 5km?

name	start_time	end_time
Layla	12:00	12:35
Jamie	10:30	11:00
Greg	11:00	11:45
Kim	09:30	09:59

Example

How long did it take these people to run 5km?

time_to_run = end_time – start_time

Name	start_time	end_time	time_to_run
Layla	12:00	12:35	00:35
Jamie	10:30	11:00	00:30
Greg	11:00	11:45	00:45
Kim	09:30	09:59	00:29

We have created a new variable **time_to_run** by calculating the difference between the start and end times.

Your turn…

Imagine you own a gym and want to work out how long your staff worked for you.

How do you think you would create a new variable **years_employed?**

name	year_started	year_left	years_employed
S. Philips	1998	2008	
J. King	2012	2020	
P. McDonald	2017	2022	
L. Whyte	2019	2021	

Your turn…

Imagine you own a gym and want to work out how long your staff worked for you.

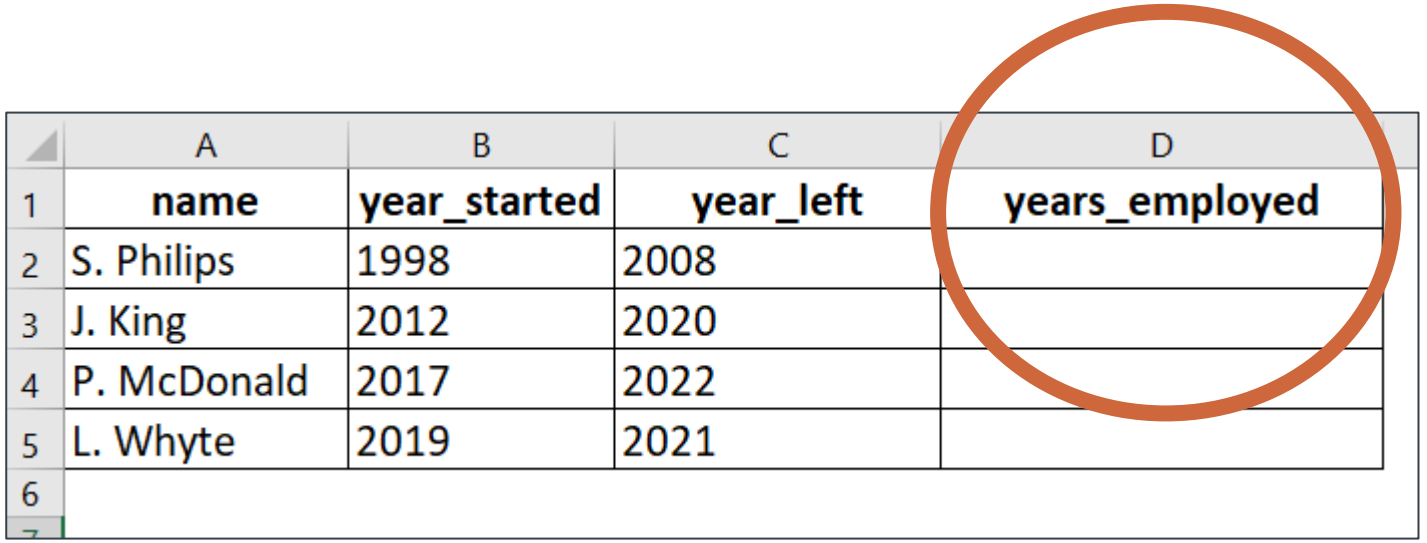
New variable **years_employed = year_left – year_started**

name	year_started	year_left	years_employed
S. Philips	1998	2008	10
J. King	2012	2020	8
P. McDonald	2017	2022	5
L. Whyte	2019	2021	2

Name the new variable

Step 1.

Type in the **name of the new variable** in the column heading where you are going to **create** the new data.

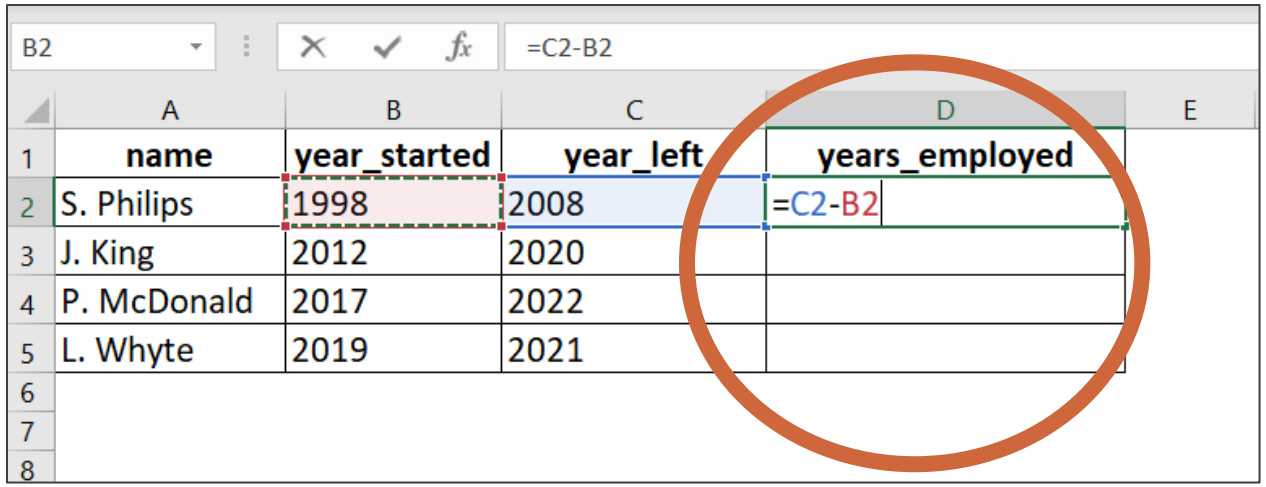


	A	B	C	D
1	**name**	**year_started**	**year_left**	**years_employed**
2	S. Philips	1998	2008	
3	J. King	2012	2020	
4	P. McDonald	2017	2022	
5	L. Whyte	2019	2021	
6				

Calculation

Step 2.

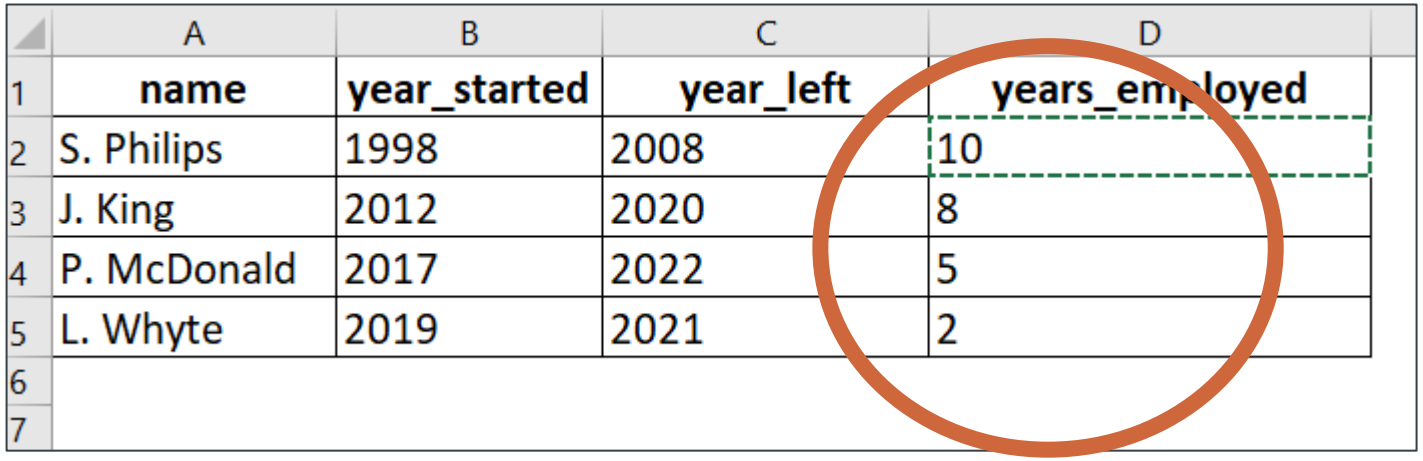
In the first row of the new variable, **type in the calculation** you will use to create the new column of data.



Create in Excel

Step 3.

Copy the calculation you have just typed into the first row, and paste into the remaining rows of the new variable.



	A	B	C	D
1	**name**	**year_started**	**year_left**	**years_employed**
2	S. Philips	1998	2008	10
3	J. King	2012	2020	8
4	P. McDonald	2017	2022	5
5	L. Whyte	2019	2021	2
6				
7				

Create in Excel

Here are some examples of calculations you can use in Excel to create new variables.

Calculation	Formula
Adding	=sum(A,B,C,D…)
Subtracting	=B-A
Multiple	=product(A,B,C,D…)
Dividing	=B/A
Maximum	=max(A,B,C,D….)
Minimum	=min(A,B,C,D,…)
Average (mean)	=average(A,B,C,D,…)

For more details of the formulas in Excel, see https://support.microsoft.com/en-us/excel

Next steps

Complete **questions 1 to 11**
in **section 2** of the
‘Creating new calculated variables in Excel’
workbook

Definition

Conditional

Performs one action if the condition is True and a different action if the condition is False

Conditional statements

IF *something is true*

THEN *this*

ELSE *that*

Conditional statements are useful when comparing data items.

The statement can compare strings (such as names, colours, places) or numbers.

Show me…

IF *it is raining*

THEN *take an umbrella*

ELSE *leave umbrella at home*

Show me…

IF *driving to Edinburgh*

THEN *turn left*

ELSE *turn right*

Your turn…

What do you think you could write in the IF statement gaps to show that,

“you need to charge up your phone when the battery power is less than 5%”?

IF *[blank]*

THEN *[blank]*

ELSE *[blank]*

Your turn…

What do you think you could write in the IF statement gaps to show that,

“you need to charge up your phone when the battery power is less than 5%”?

IF *battery_power<5%*

THEN *plug in*

ELSE *don’t*

Your turn…

What do you think you could write in the IF statement gaps to show that,

“you can cross the road when the light turns green”?

IF *[blank]*

THEN *[blank]*

ELSE *[blank]*

Your turn…

What do you think you could write in the IF statement gaps to show that,

“you can cross the road when the light turns green”?

IF *light is green*

THEN *walk*

ELSE *wait*

Show me…

By using conditional formulas (or IF statements) you can categorise or compare data items.

Animal	Is it green?
Peacock	True
Zebra	False
Tiger	False
Frog	True
Puffin	False

IF *animal is green*

THEN *True*

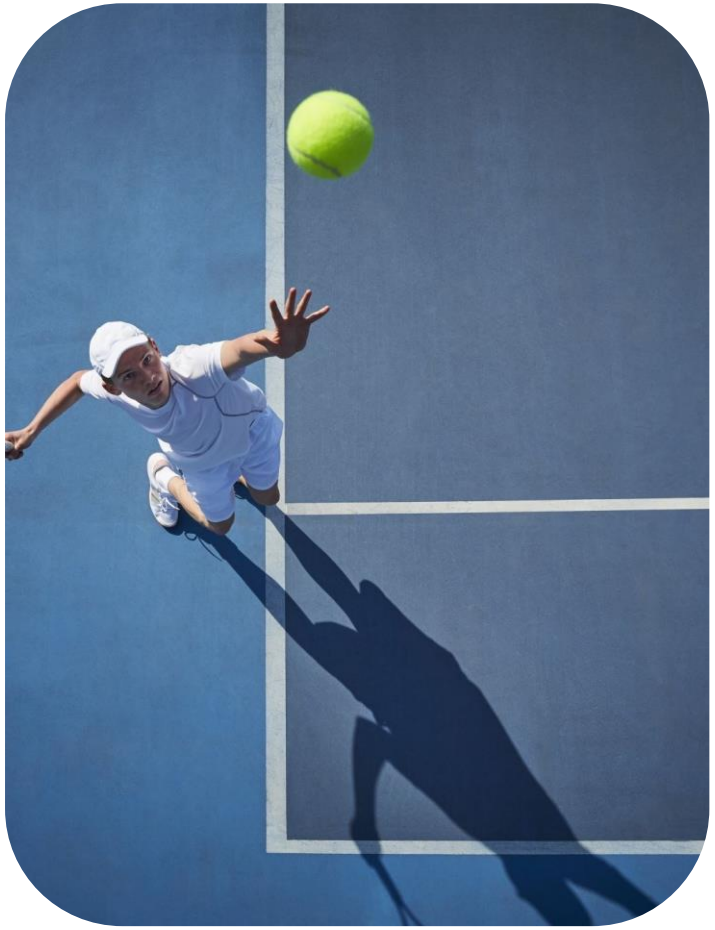
ELSE *False*

Example

Which of these people have won **more than 10** Grand Slam Tennis titles?

Create a new variable a value of Yes if the tennis player has won more than 10 titles, otherwise its No.

name	grand_slam_titles
Billie Jean King	8
Steffi Graf	22
Pete Sampras	14
Bjorn Borg	11



Example

Which of these people have won **more than 10** Grand Slam Tennis titles?

name	grand_slam_titles	more_than_10_titles
Billie Jean King	8	No
Steffi Graf	22	Yes
Pete Sampras	14	Yes
Bjorn Borg	11	Yes

IF **Grand_slam_titles** >10

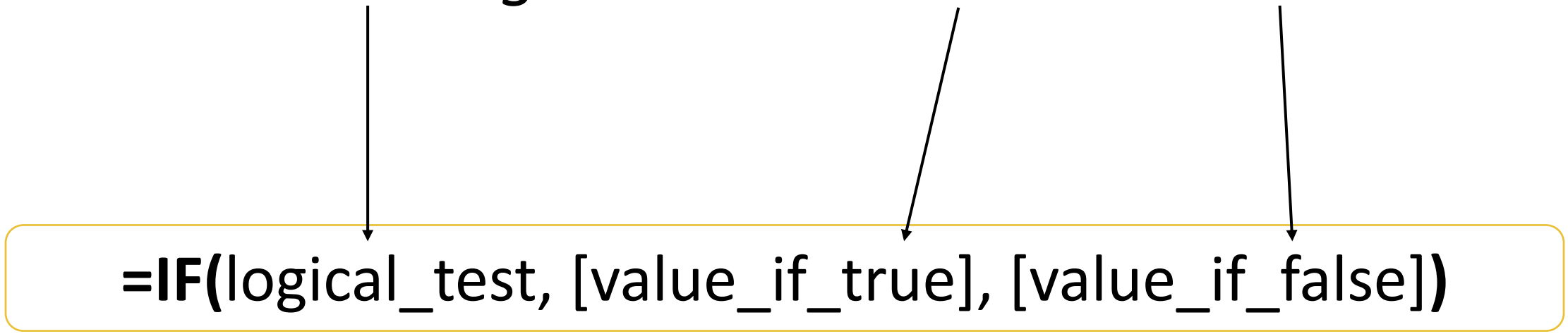
THEN *Yes*

ELSE *No*

IF statement in Excel

We are now going to use the IF statement in Excel to create the new variable.

IF *something is true* THEN *this* ELSE *that*



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

Logical Test

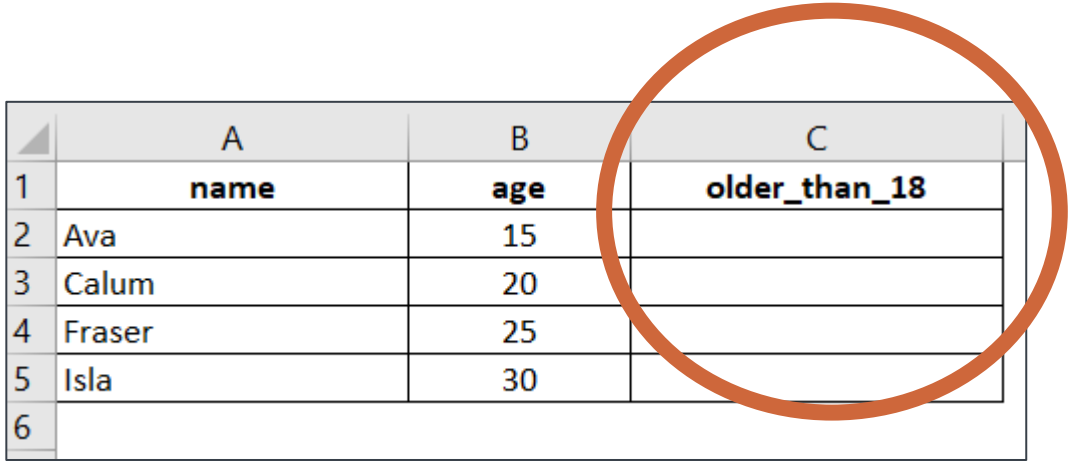
Here are some examples of the logical tests you can use in the IF statement.

Logical_test	
Equals	variable_1=variable_2
Less than	variable_1<variable_2
More than	variable_1>variable_2

Name the new variable

Step 1.

Type in the **name of the new variable** in the column heading where you are going to **create** the new data.

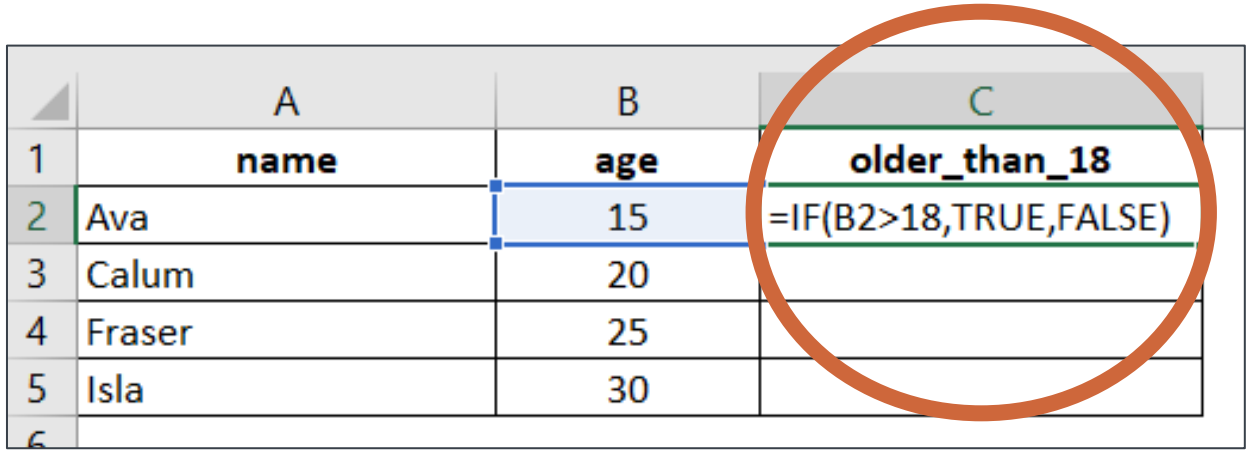


	A	B	C
1	**name**	**age**	**older_than_18**
2	Ava	15	
3	Calum	20	
4	Fraser	25	
5	Isla	30	
6			

Calculation

Step 2.

In the first row of the new variable, **type in the calculation** you will use to create the new column of data. In this case it is an IF statement.

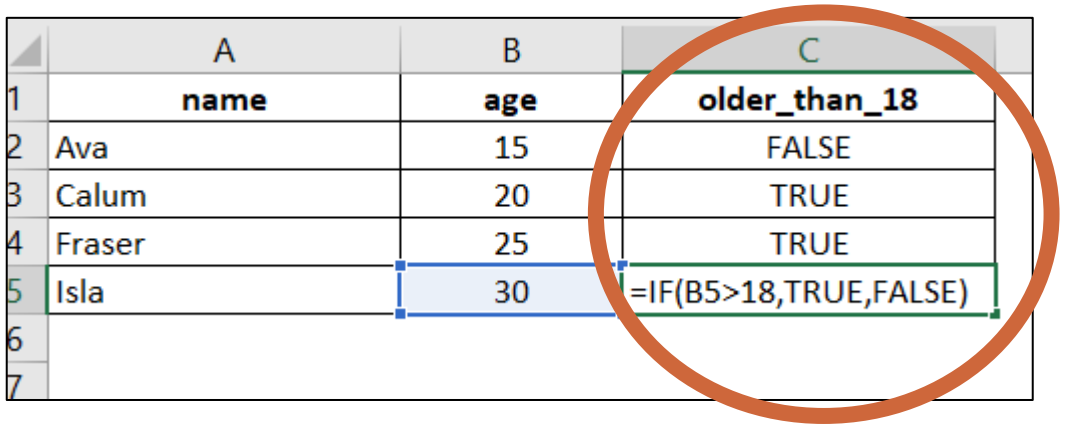


	A	B	C
1	**name**	**age**	**older_than_18**
2	Ava	15	=IF(B2>18,TRUE,FALSE)
3	Calum	20	
4	Fraser	25	
5	Isla	30	
6			

IF statement in Excel

Step 3.

Copy the calculation you have just typed into the first row, and paste into the remaining rows of the new variable.



	A	B	C
1	**name**	**age**	**older_than_18**
2	Ava	15	FALSE
3	Calum	20	TRUE
4	Fraser	25	TRUE
5	Isla	30	=IF(B5>18,TRUE,FALSE)
6			
7			

Text in formulas

If you are going to use text in formulas, you need to wrap the text in quotes (e.g. “Text”).

The only exception to that is using TRUE or FALSE, which Excel automatically understands.

Next steps

Complete **questions 1 to 8**
in **section 3** of the
‘Creating new calculated variables in Excel’
workbook

Missing values

When creating new variables there are some situations that might cause issues in your dataset.

Some of the most common issues are caused by:

- **Missing values**
- **Zero (0) values**

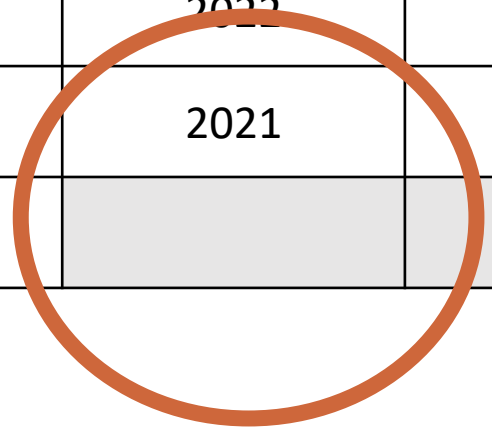
Missing values

Datasets quite often will have missing or blank information.

In this example, **R. Evans is still working for the company** so doesn't have a year they left.

Can you think of any solutions for dealing with this missing value so you can calculate years_employed?

name	year_started	year_left	years _employed
S. Philips	1998	2008	
J. King	2012	2020	
P. McDonald	2017	2022	
L. Whyte	2019	2021	
R. Evans	2016		



Missing values

Some possible solutions are,

- Put a **value such as NA** (Not Available), **NaN** (Not a Number) **or NULL** to show that an answer couldn't be calculated
- **Remove the row** and only calculate based on people who have left.

name	year_started	year_left	years _employed
S. Philips	1998	2008	10
J. King	2012	2020	8
P. McDonald	2017	2022	5
L. Whyte	2019	2021	2
R. Evans	2016		NA

Handling zero (0) values

When creating a new variable you might come across a row where you are **trying to divide by 0.**

Like with a missing value you can:

- Put a **value such as NA, NaN or NULL** to show that an answer couldn't be calculated

- **Remove the row**

average_price = total_sales/number_sold

item	total_sales	number_sold	average_price
trainers	£30	3	£10.00
top	£15	2	£7.50
shorts	£10	4	£2.50
socks	£0	0	!can't divide by 0

Documentation best practice

When creating new variables it is best practice to write down **why** you are creating it and **the calculation/formula used.**

This can be within the file you are using to create it or as a separate document.

Learning checklist

I can *describe* how to create a new variable by performing calculations.

I can *create* new variables in Excel by performing calculations.

I can *describe* the concept of conditional formulas.

I can *use* conditional formulas in Excel.

How you can use this lesson

You are free to:

- **Share** – copy and redistribute the material in any medium or format
- **Adapt** – remix, transform and build upon the material

Under the following terms:

- **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** — You may not use the material for commercial purposes.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Created by effini in partnership with Data Education in Schools, The Data Lab and Data Skills for Work, with funding from the Scottish Government.

effini

THE DATA LAB
value from data

Data Skills FOR WORK

Data Education in Schools

Alternative format

If you require this document in an alternative format, such as large print or a coloured background, please contact

hello@effini.com

or

4th Floor, The Bayes Centre
47 Potterrow
Edinburgh
EH8 9BT

effini

Data Education in Schools

Skills Development Scotland