Summarising datasets in Excel

Learning intentions

We will be learning to summarise datasets in Excel, specifically to

- calculate the **total, count**, **min/max** and **average** of rows data
- **group rows of data** based on logical criteria
- perform **calculations on grouped data**

Background

When trying to solve a problem in data science understanding the data you have is fundamental.

Rows of data can be **filtered** and **sorted** to help you understand your data.

In this lesson we will look at how you can also **summarise and group rows of data.**

Why this is important?

Some benefits of grouping and summarising data are,

Makes the data **easier to work** with

Focus on the important messages

Allows you to **simplify** your dataset

Helps you **describe** your data, e.g., What is the total? How many rows to you have?

Definition

Summarise

To condense the rows in a dataset (often to a single value) by performing a calculation on the data items within a variable.

Summarising data

In the similar way that you can perform calculations on columns of data, you can **perform calculations on rows of data**.

The most common calculations performed on rows of data are,

- Count (number of rows)
- Total
- Average

Show me…

The new rows created by summarising are often shown as a separate dataset.

Original dataset

animal	age_yrs
Lion	10
Tiger	2
Elephant	15
Penguin	3
Parrot	5

Summarise

Summarised dataset

summary	age_yrs
Count	5
Total	35
Average (mean)	7

Show me…

When you are summarising a dataset you can **select the required variables** and then **summarise them.**

month	number_sold	price	income
Jan	6	41	246
Feb	5	27	135
Mar	4	46	184
Apr	6	28	168
May	2	41	82

Summarise

total_number_sold	total_income
23	851

Example

This dataset shows the test results for 5 pupils. Summarise the test results in this dataset by calculating,

- Count
- Average (mean)
- Maximum test score
- Minimum test score

pupil_ID	test_1	test_2	test_3
GH1254	50%	36%	72%
SE1547	45%	64%	94%
DM4758	90%	48%	78%
KL4758	32%	93%	52%
PM4575	85%	86%	92%

Example

Summarise the test results in this dataset.

Original dataset

pupil_ID	test_1	test_2	test_3
GH1254	50%	36%	72%
SE1547	45%	64%	94%
DM4758	90%	48%	78%
KL4758	32%	93%	52%
PM4575	85%	86%	92%

Summarised dataset

summary	test_1	test_2	test_3
Count	5	5	5
Average	60%	65%	78%
Maximum	90%	93%	94%
Minimum	32%	36%	52%

Your turn…

What do you think the **count, maximum and minimum** would be in this dataset?

ocean	depth_m
Pacific	3,970
Atlantic	3,646
Indian	3,741
Arctic	1,205

Your turn…

What do you think the **count, maximum and minimum** would be in this dataset?

ocean	depth_m
Pacific	3,970
Atlantic	3,646
Indian	3,741
Arctic	1,205

Summarise

summary	depth_m
Count	4
Maximum	3,970
Minimum	1,205

Summarise in Excel

We are now going to look at summarising datasets in Excel.

The table below shows the formulas we will need for this.

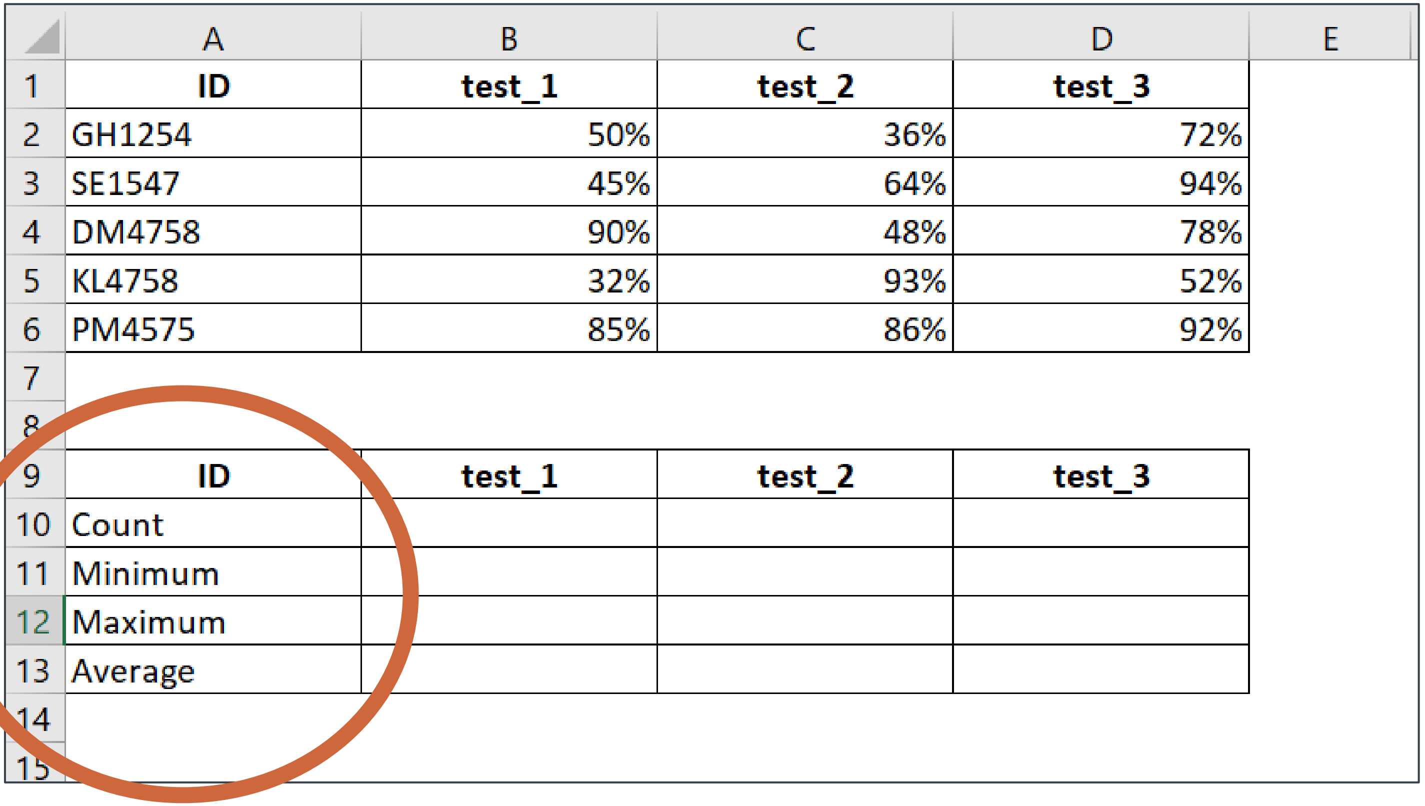
Calculation	Formula
Total	=sum(A,B,C,D…)
Maximum	=max(A,B,C,D….)
Minimum	=min(A,B,C,D,…)
Average (mean)	=average(A,B,C,D,…)
Count	=count(A,B,C,D,…)

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

Summarising in Excel

Step 1.

Create a new dataset with the variable headings you have selected and **row labels for summary types** you will calculate.

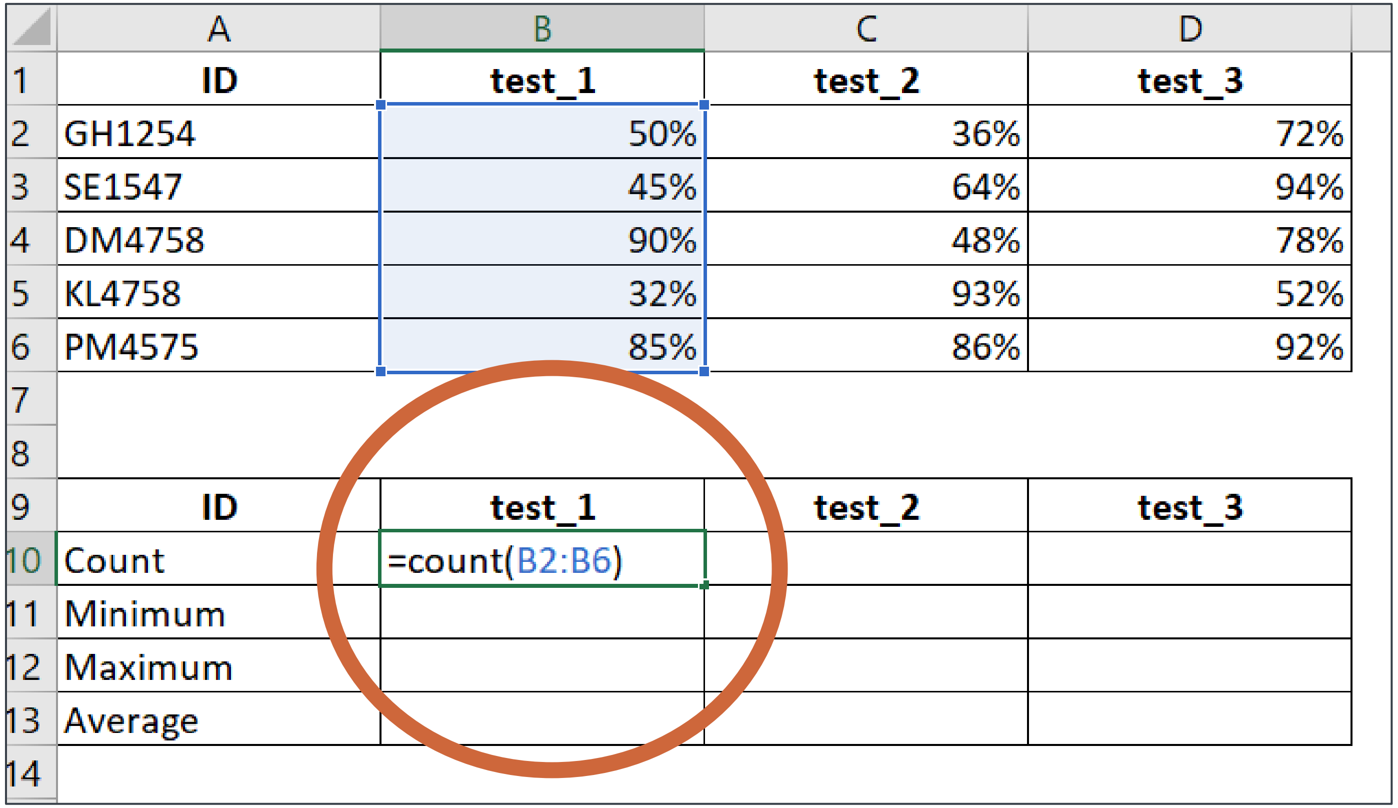


	A	B	C	D	E
1	**ID**	**test_1**	**test_2**	**test_3**	
2	GH1254	50%	36%	72%	
3	SE1547	45%	64%	94%	
4	DM4758	90%	48%	78%	
5	KL4758	32%	93%	52%	
6	PM4575	85%	86%	92%	
7					
8					
9	**ID**	**test_1**	**test_2**	**test_3**	
10	Count				
11	Minimum				
12	Maximum				
13	Average				
14					
15					

Summary formulas in Excel

Step 2.

Type in the calculation you will use to summarise the data.

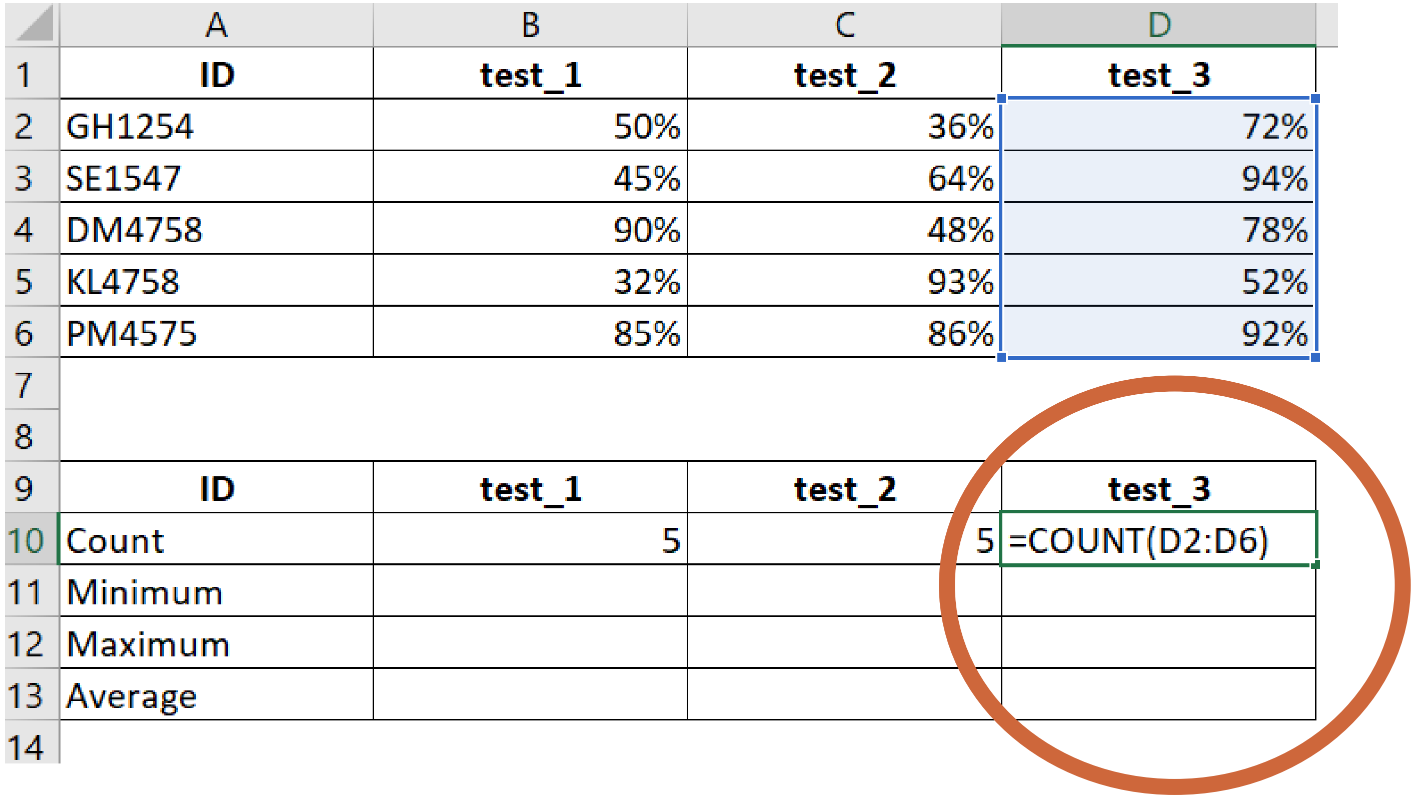


	A	B	C	D
1	ID	test_1	test_2	test_3
2	GH1254	50%	36%	72%
3	SE1547	45%	64%	94%
4	DM4758	90%	48%	78%
5	KL4758	32%	93%	52%
6	PM4575	85%	86%	92%
7				
8				
9	ID	test_1	test_2	test_3
10	Count	=count(B2:B6)		
11	Minimum			
12	Maximum			
13	Average			
14				

Copy formulas

Step 3.

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

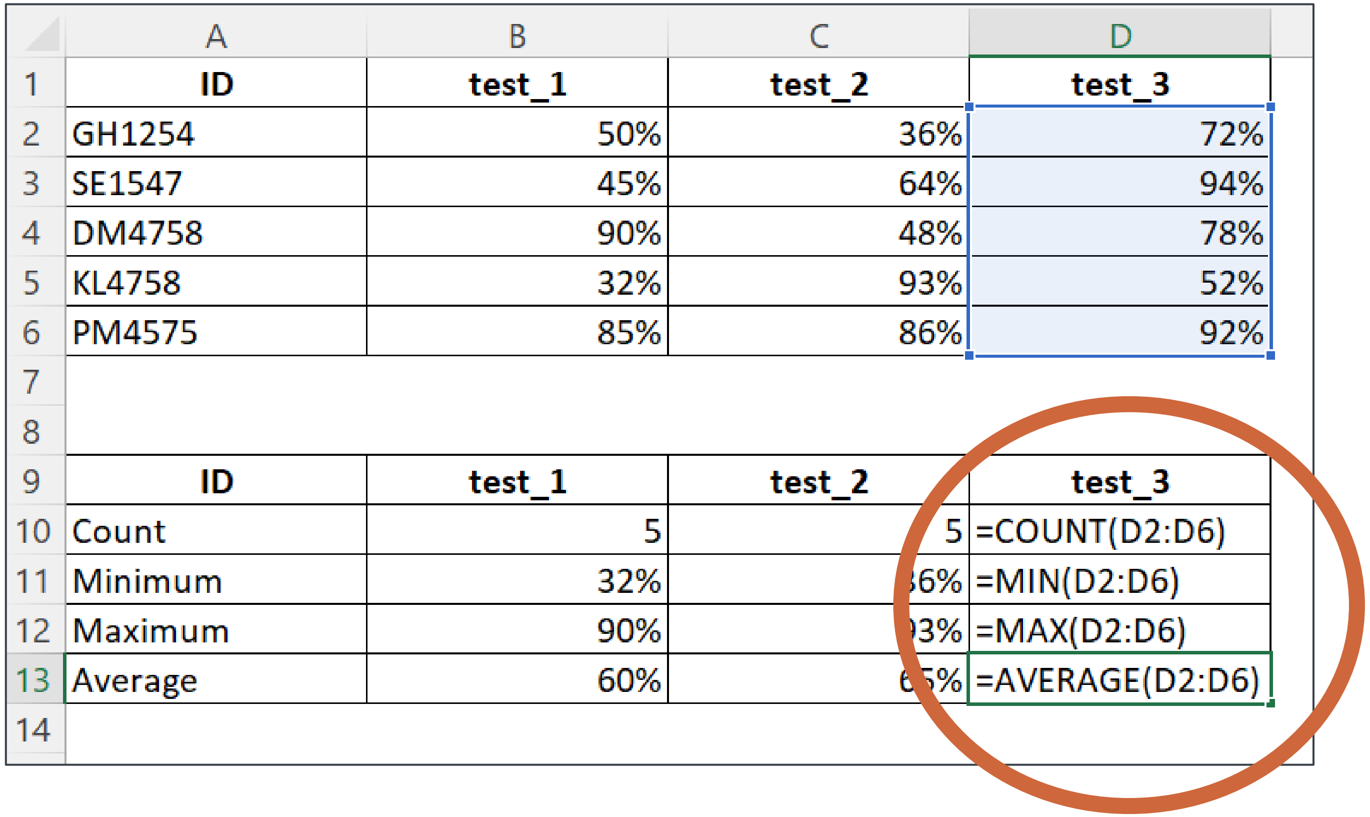


	A	B	C	D
1	**ID**	**test_1**	**test_2**	**test_3**
2	GH1254	50%	36%	72%
3	SE1547	45%	64%	94%
4	DM4758	90%	48%	78%
5	KL4758	32%	93%	52%
6	PM4575	85%	86%	92%
7				
8				
9	**ID**	**test_1**	**test_2**	**test_3**
10	Count	5	5	=COUNT(D2:D6)
11	Minimum			
12	Maximum			
13	Average			
14				

Summarise in Excel

Step 4.

Repeat the process for any other summary calculations you need.



	A	B	C	D
1	**ID**	**test_1**	**test_2**	**test_3**
2	GH1254	50%	36%	72%
3	SE1547	45%	64%	94%
4	DM4758	90%	48%	78%
5	KL4758	32%	93%	52%
6	PM4575	85%	86%	92%
7				
8				
9	**ID**	**test_1**	**test_2**	**test_3**
10	Count	5	5	=COUNT(D2:D6)
11	Minimum	32%	36%	=MIN(D2:D6)
12	Maximum	90%	93%	=MAX(D2:D6)
13	Average	60%	65%	=AVERAGE(D2:D6)
14				

Next steps

Complete **questions 1 to 6** in **section 1** of the 'Summarising datasets in Excel' workbook.

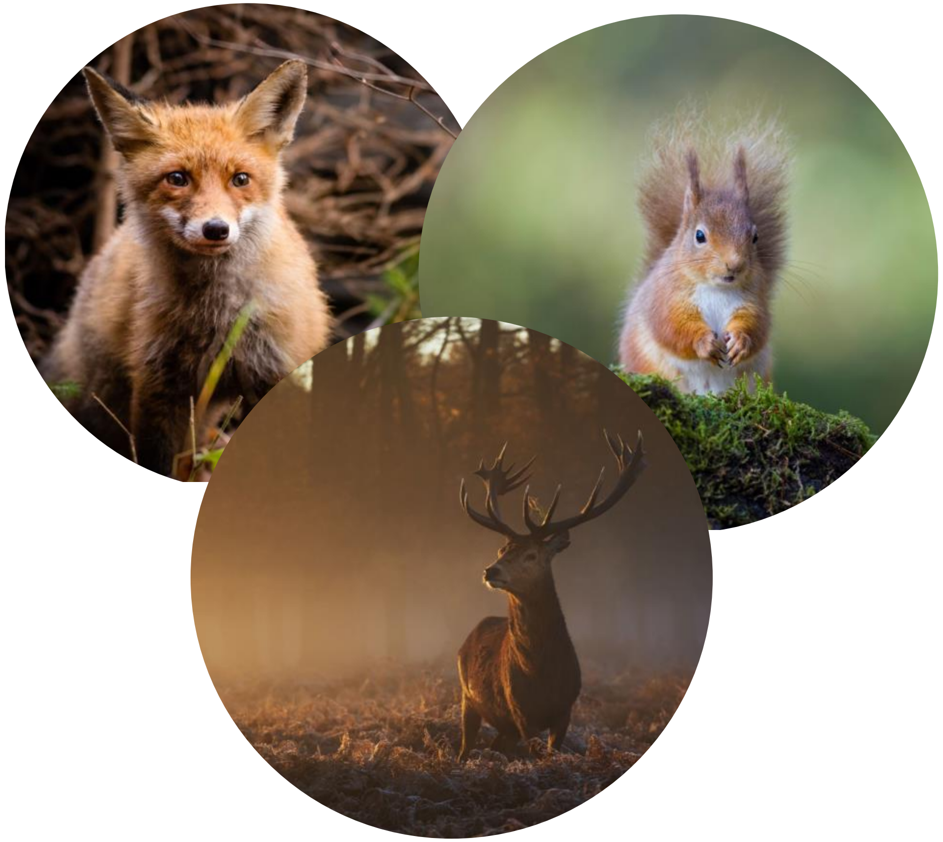
Definition

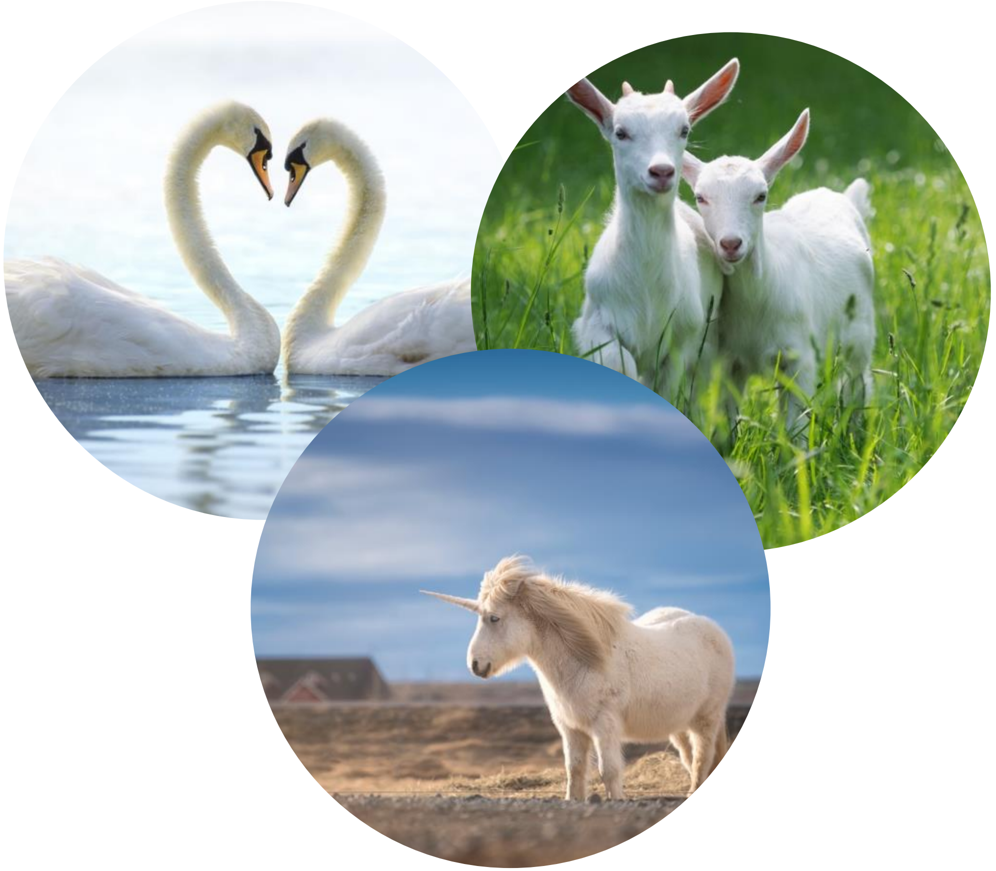
Group

To split a dataset into sets of rows based on some criteria.

Show me…

These animals have been **grouped by colour**





Grouping data

In the last section we looked at summarising all the rows in a dataset.

Now we are going to look at how to split datasets into sets of rows then **summarise these groups of rows**.

In data science, it is more usual (and useful) to summarise grouped data.

Grouping data

When grouping data it can help to think about the following questions,

What **data** do I have?

What do you **need** from the data?

What **criteria** do I need to group my data by?

Grouping data

What **data** do I have?

What do you **need** from the data?

What **criteria** do I need to group my data by?

Details of people visiting different venues in a town,

Month,

Venue,

Number of visitors

“I need to know the **total number of visitors** in **each month**”

Grouped **by month** then calculate the **total**

Grouping data

"I need to know the **total number of visitors** in **each month**"

month	venue	number_visitors
January	Café	300
January	Ice cream shop	50
January	Restaurant	2,500
February	Café	200
February	Ice cream shop	40
February	Restaurant	1,000

Group by **month**
Then **summarise**

month	total_number_visitors
January	2,850
February	1,240

Your turn…

You need to calculate the **total_sales** grouped by different criteria.

Can you think of some criteria you could use to group this dataset? e.g. group by price

product	colour	type	price	sales
Apple	Pink	Fruit	£1.00	£53.00
Banana	Yellow	Fruit	£0.50	£40.50
Carrot	Orange	Vegetable	£0.50	£37.00
Dragon fruit	Pink	Fruit	£1.00	£15.00
Pepper	Yellow	Vegetable	£0.50	£12.50

Your turn…

product	colour	type	price	sales
Apple	Pink	Fruit	£1.00	£53.00
Banana	Yellow	Fruit	£0.50	£40.50
Carrot	Orange	Vegetable	£0.50	£37.00
Dragon fruit	Pink	Fruit	£1.00	£15.00
Pepper	Yellow	Vegetable	£0.50	£12.50

It could be grouped by:

- Colour
- Type
- Price

colour	total_sales
Pink	£68.00
Yellow	£53.00
Orange	£37.00

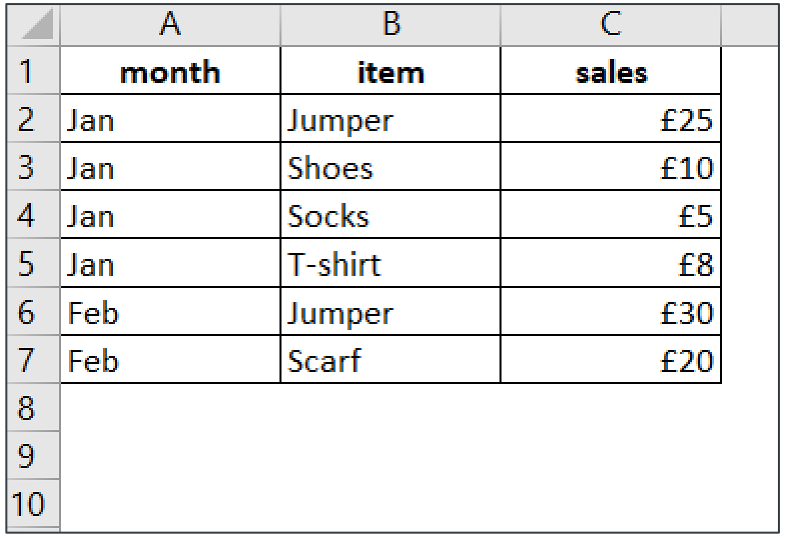
type	total_sales
Fruit	£108.50
Vegetable	£49.50

price	total_sales
£1.00	£68.00
£0.50	£90.00

How to group a dataset in Excel

We are now going to look at **how to group this dataset in Excel** and calculate,

- Number of sales by month (Count)
- Total sales by month (Sum)

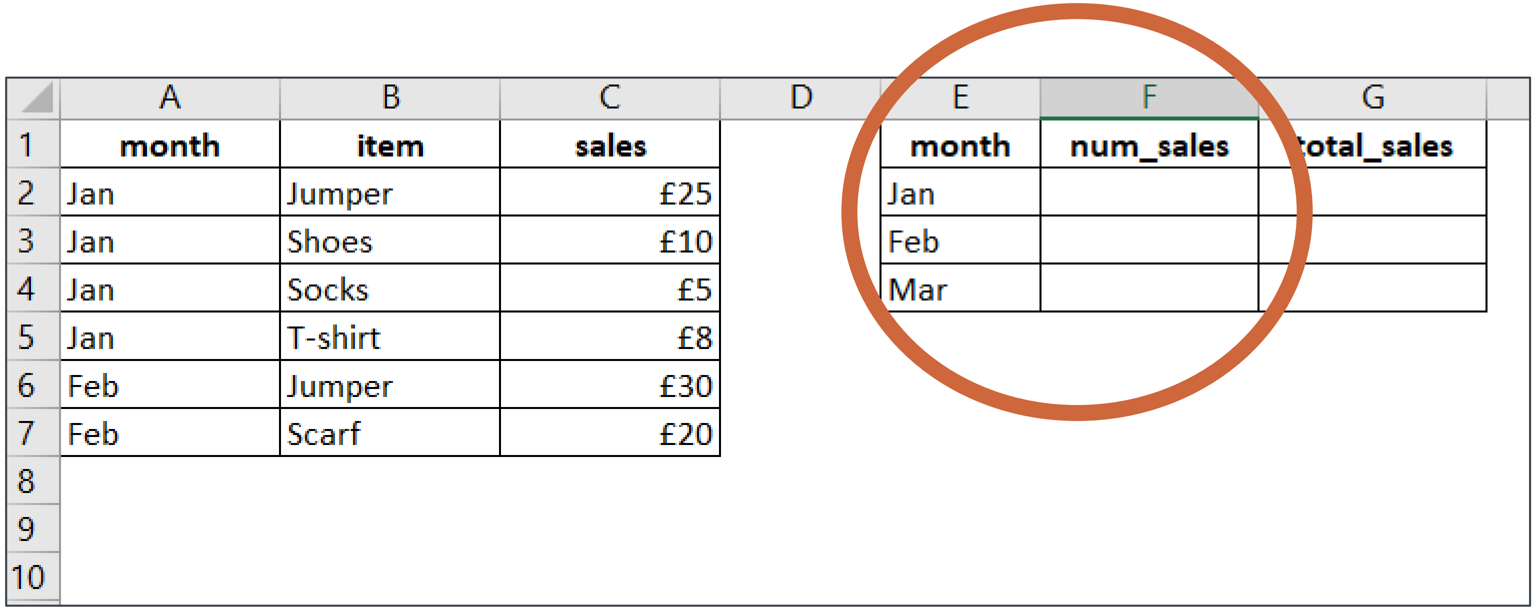


	A	B	C
1	**month**	**item**	**sales**
2	Jan	Jumper	£25
3	Jan	Shoes	£10
4	Jan	Socks	£5
5	Jan	T-shirt	£8
6	Feb	Jumper	£30
7	Feb	Scarf	£20
8			
9			
10			

Setting up a grouped dataset in Excel

Step 1.

Create a new dataset with the variable headings and the data items you want to group by, in this case by **month**.



	A	B	C	D	E	F	G
1	**month**	**item**	**sales**		**month**	**num_sales**	**total_sales**
2	Jan	Jumper	£25		Jan		
3	Jan	Shoes	£10		Feb		
4	Jan	Socks	£5		Mar		
5	Jan	T-shirt	£8				
6	Feb	Jumper	£30				
7	Feb	Scarf	£20				
8							
9							
10							

COUNTIF

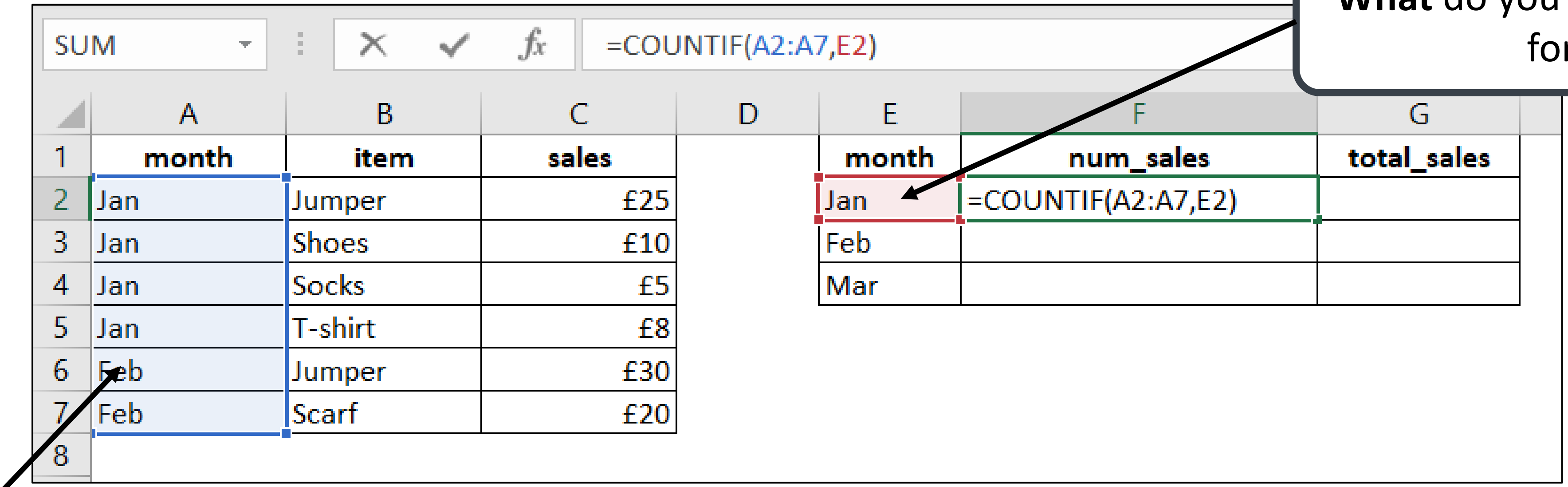
Step 2.

As we want to count the number of sales by month, we can use an Excel function COUNTIF.

=countif(Where do you want to look?, What do you want to look for?**)**

COUNTIF

=countif(Where do you want to look?, What do you want to look for?**)**



SUM =COUNTIF(A2:A7,E2)

	A	B	C	D	E	F	G
1	**month**	**item**	**sales**		**month**	**num_sales**	**total_sales**
2	Jan	Jumper	£25		Jan	=COUNTIF(A2:A7,E2)	
3	Jan	Shoes	£10		Feb		
4	Jan	Socks	£5		Mar		
5	Jan	T-shirt	£8				
6	Feb	Jumper	£30				
7	Feb	Scarf	£20				
8							

What do you want to look for?

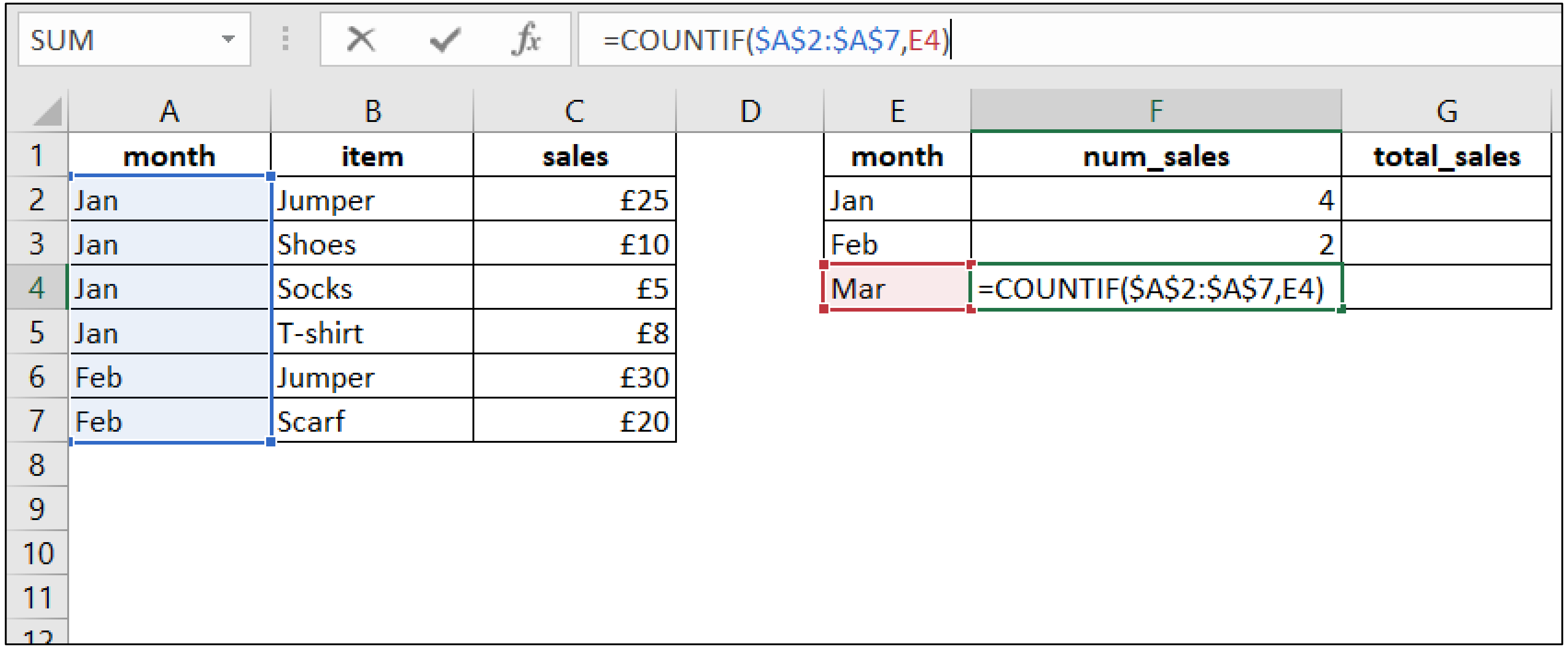
Where do you want to look?

This formula looks for the string “Jan” in the cells A2 to A7 and returns the number of times it finds it.

COUNTIF

Step 3.

Copy the formula and paste it into the remaining rows.



SUM

=COUNTIF($A$2:$A$7,E4)

	A	B	C	D	E	F	G
1	**month**	**item**	**sales**		**month**	**num_sales**	**total_sales**
2	Jan	Jumper	£25		Jan	4	
3	Jan	Shoes	£10		Feb	2	
4	Jan	Socks	£5		Mar	=COUNTIF($A$2:$A$7,E4)	
5	Jan	T-shirt	£8				
6	Feb	Jumper	£30				
7	Feb	Scarf	£20				
8							
9							
10							
11							

There were
4 items sold in January,
2 in February
and 0 in March.

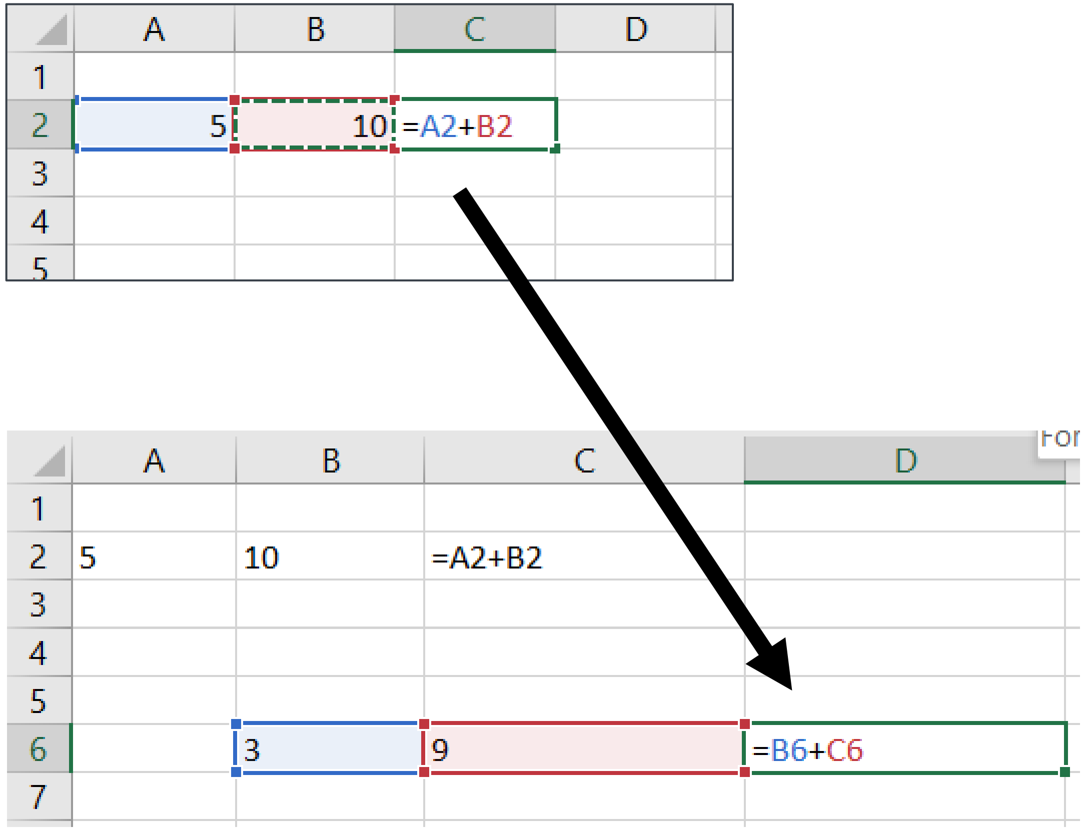
Relative vs. absolute cell references

When you type a formula that uses cell names (e.g. A2) in Excel it is looking at **where the cells are** compared to the formulas.

In this example, the formula is adding the 2 cells to its left.

When you copy the formula to another part of the sheet it is still adding the 2 cells to its left.

These are called **relative cell references**.



	A	B	C	D
1				
2	5	10	=A2+B2	
3				
4				
5				

	A	B	C	D
1				
2	5	10	=A2+B2	
3				
4				
5				
6		3	9	=B6+C6
7				

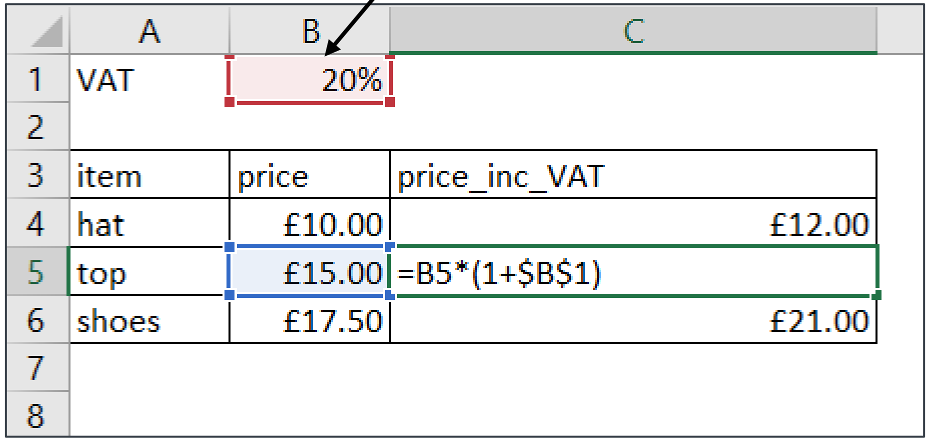
Relative vs. absolute cell references

Relative cell references is the standard way Excel uses formulas, and most of the time is what you will need.

However there are some times when you want specific cell(s) to be used in a formula regardless of where you copy the formula too.

These are called **absolute cell references.**

The cell containing the VAT % is fixed and needs to be used in lots of different formulas.



	A	B	C
1	VAT	20%	
2			
3	item	price	price_inc_VAT
4	hat	£10.00	£12.00
5	top	£15.00	=B5*(1+$B$1)
6	shoes	£17.50	£21.00
7			
8			

Making cells absolute references

By **adding dollar signs ($)** before the column letter and/or row number, it tells Excel not to change these references when copying and pasting the formula.

It makes the **cell reference absolute.**

=countif(A2:A7,E2)

↓

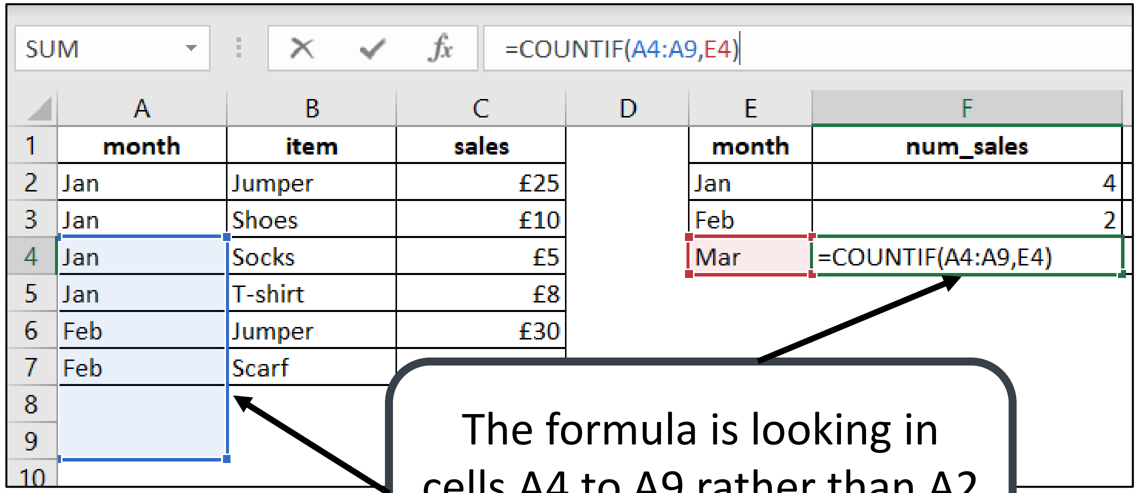
=countif(**$**A**$**2:**$**A**$**7,E2)

See ‘Switch between relative, absolute and mixed references’ for more details.

Microsoft support: switch between relative absolute and mixed references

Impact of the $ signs in a formula

Without $ signs in the formula

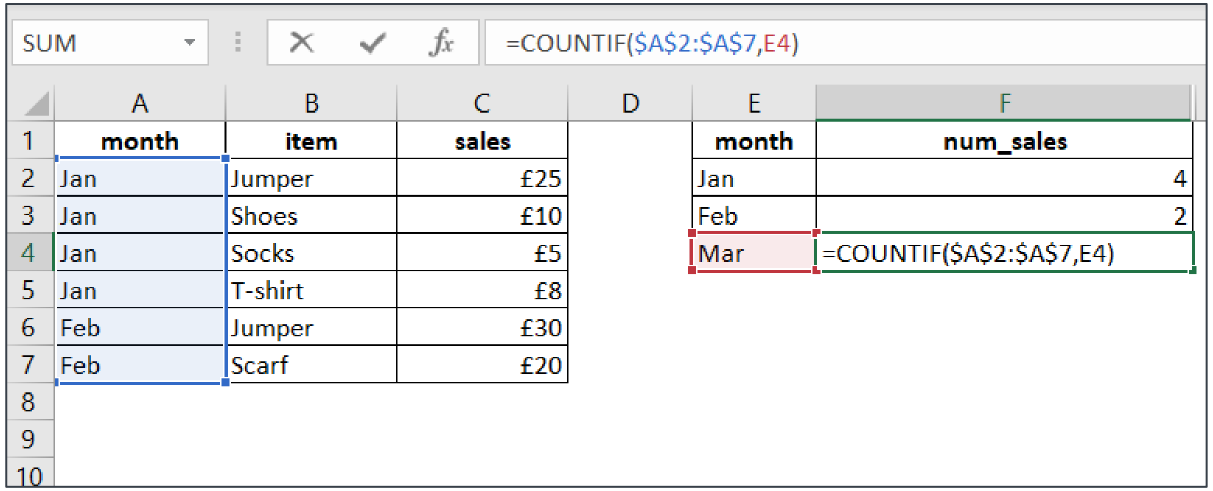


=COUNTIF(A4:A9,E4)

	A	B	C	D	E	F
1	**month**	**item**	**sales**		**month**	**num_sales**
2	Jan	Jumper	£25		Jan	4
3	Jan	Shoes	£10		Feb	2
4	Jan	Socks	£5		Mar	=COUNTIF(A4:A9,E4)
5	Jan	T-shirt	£8			
6	Feb	Jumper	£30			
7	Feb	Scarf				
8						
9						
10						

The formula is looking in cells A4 to A9 rather than A2 to A7.

With $ signs in the formula



=COUNTIF($A$2:$A$7,E4)

	A	B	C	D	E	F
1	**month**	**item**	**sales**		**month**	**num_sales**
2	Jan	Jumper	£25		Jan	4
3	Jan	Shoes	£10		Feb	2
4	Jan	Socks	£5		Mar	=COUNTIF($A$2:$A$7,E4)
5	Jan	T-shirt	£8			
6	Feb	Jumper	£30			
7	Feb	Scarf	£20			
8						
9						
10						

By adding the $ signs into the formulas, when the formula has been copied it is still referencing the cells we want.

Next steps

Complete **questions 1 to 9** in **section 2** of the ‘Summarising datasets in Excel’ workbook.

More on grouping

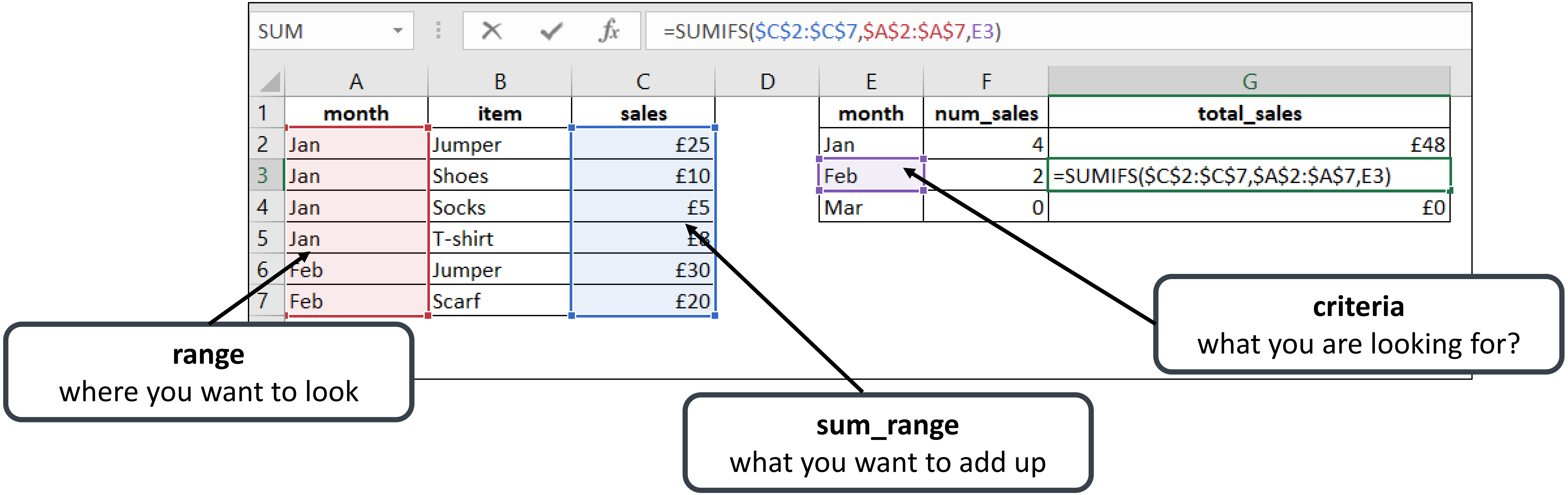
As well as counting by a group you can calculate the total, average, maximum or minimum in a group.

Calculation	**Formula**
Total	=sumifs(sum_range, range, criteria)
Average (mean)	=averageifs(average_range, criteria_range, criteria)
Maximum	=maxifs(max_range, criteria_range, criteria)
Minimum	=minifs(max_range, criteria_range, criteria)

SUMIFS

In this dataset we are looking to create a variable **total_sales** grouped by **month**.

=sumifs(**sum_range**, **range**, **criteria**)



SUM

=SUMIFS($C$2:$C$7,$A$2:$A$7,E3)

	A	B	C	D	E	F	G
1	**month**	**item**	**sales**		**month**	**num_sales**	**total_sales**
2	Jan	Jumper	£25		Jan	4	£48
3	Jan	Shoes	£10		Feb	2	=SUMIFS($C$2:$C$7,$A$2:$A$7,E3)
4	Jan	Socks	£5		Mar	0	£0
5	Jan	T-shirt	£8				
6	Feb	Jumper	£30				
7	Feb	Scarf	£20				

range
where you want to look

sum_range
what you want to add up

criteria
what you are looking for?

Next steps

Complete **questions 1 to 8** in **section 3** of the ‘Summarising datasets in Excel’ workbook.

Summarising with missing values

Datasets can often **contain missing or blank values.** This can cause issues when you are summarising them.

How you handle these missing values can impact on your final results.

Summarising with missing values

If a dataset has missing values you can,

- Put a **value such as NA** (Not Available)**, NaN** (Not a Number) or **NULL**
- **Remove the row**

With both of these options you need to careful when you go on to summarise the dataset.

ID	test_1	test_2	test_3
GH1254	50%	36%	72%
SE1547	45%	64%	94%
DM4758	90%	48%	78%
KL4758	32%	93%	52%
PM4575	85%	NA	NA

Show me…

Option 1: Put a **value such as NA, NaN or NULL**

ID	test_1	test_2	test_3
GH1254	50%	36%	72%
SE1547	45%	64%	94%
DM4758	90%	48%	78%
KL4758	32%	93%	52%
PM4575	85%	NA	NA
Average	**60%**	**60%**	**74%**

Option 2: **Remove row** containing ID PM4575

ID	test_1	test_2	test_3
GH1254	50%	36%	72%
SE1547	45%	64%	94%
DM4758	90%	48%	78%
KL4758	32%	93%	52%
Average	**54%**	**60%**	**74%**

The two options have produced different answers for **test_1.**

Summarising with missing values

Before you decide which option to use when handling missing values, it helps to think about these questions.

Why is there a missing value?

If you remove the row will you lose information?

Is the missing value likely to be added in later?

Example

Calculate the **average price** of these video games.

video_game	**price**
Super Mario Odyssey	39.99
Call of Duty: Modern Warfare	27.99
Mario Kart 8 Deluxe	39.99
Hades	?
Overcooked! All You Can Eat	22.85

Example

For this dataset, the row with the missing data has been **removed** before the average has been calculated.

No information about the price of the games has been lost by removing the row.

video_game	price
Super Mario Odyssey	39.99
Call of Duty: Modern Warfare	27.99
Mario Kart 8 Deluxe	39.99
~~Hades~~	~~?~~
Overcooked! All You Can Eat	22.85

video_game	average_price
All excluding "Hades"	32.71

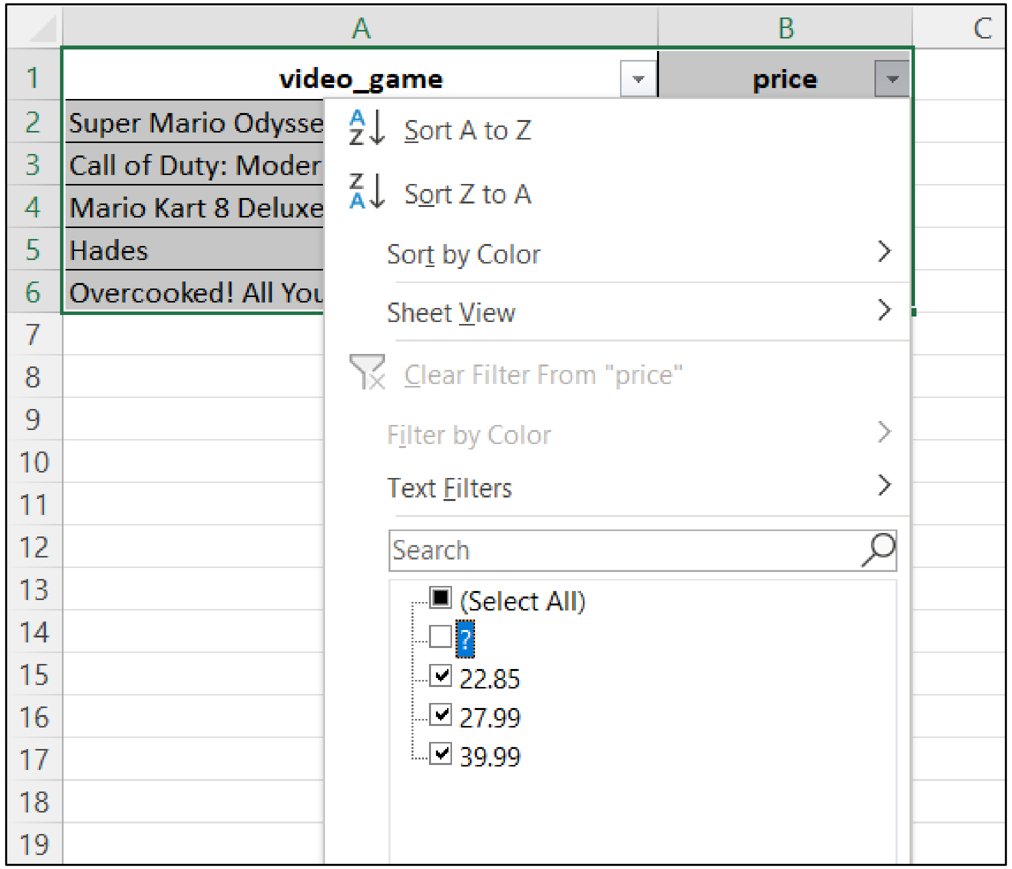
Removing rows from a dataset in Excel

Step 1.

To remove a row of data from a dataset in Excel, you need to follow the same steps as for **filtering data.**

Reminder: to turn on filters on dataset you need to press,

Windows	Ctrl + Shift + L
Mac	Command + Shift +L



	A	B	C
1	video_game	price	
2	Super Mario Odysse		
3	Call of Duty: Moder		
4	Mario Kart 8 Deluxe		
5	Hades		
6	Overcooked! All You		

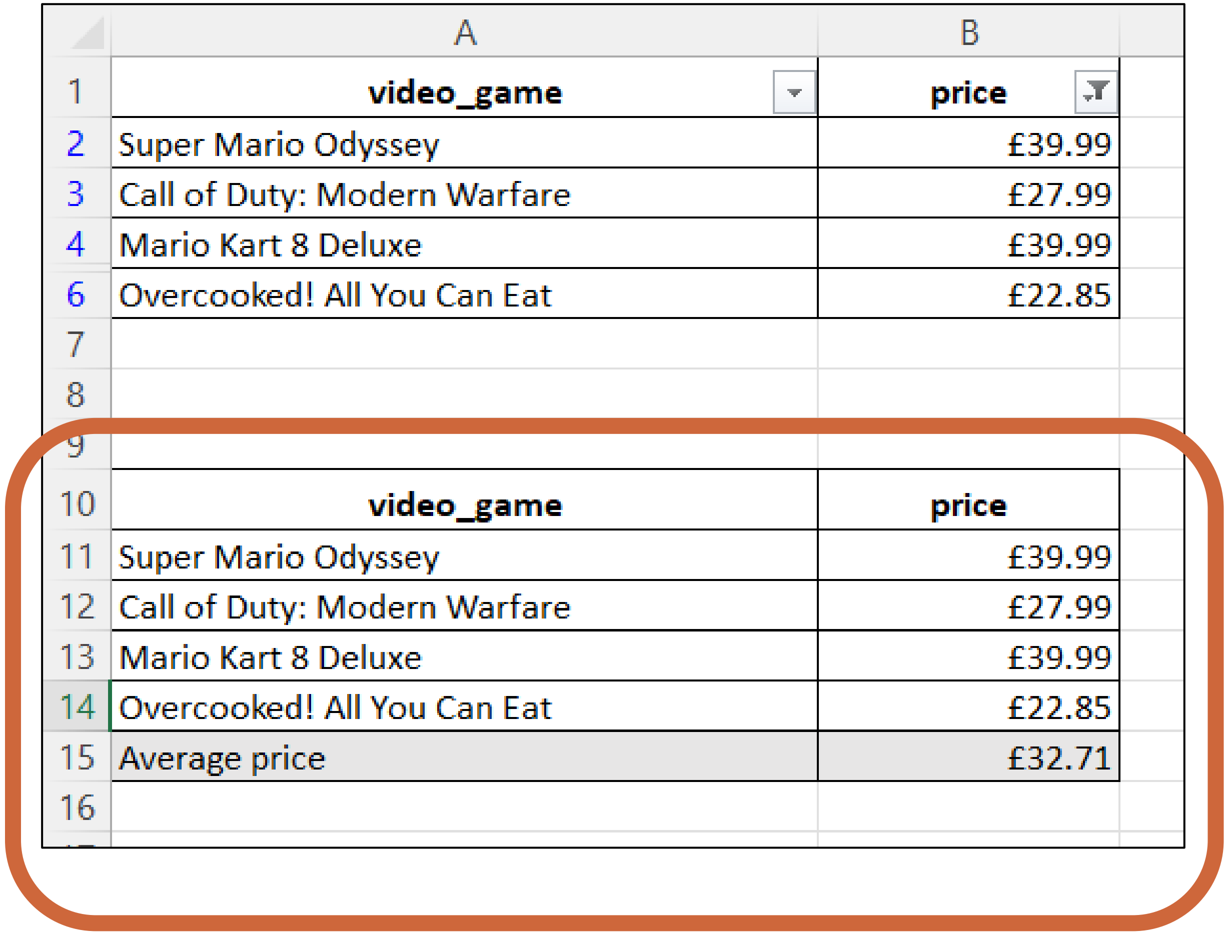
Removing rows from a dataset in Excel

Step 2.

Once you have turned on the filter, you need to select just the **rows that don't have missing values.**

Then **copy and paste** the filtered rows into a new part of the workbook.

You can now **summarise** the dataset.



	A	B
1	**video_game**	**price**
2	Super Mario Odyssey	£39.99
3	Call of Duty: Modern Warfare	£27.99
4	Mario Kart 8 Deluxe	£39.99
6	Overcooked! All You Can Eat	£22.85
7		
8		
9		
10	**video_game**	**price**
11	Super Mario Odyssey	£39.99
12	Call of Duty: Modern Warfare	£27.99
13	Mario Kart 8 Deluxe	£39.99
14	Overcooked! All You Can Eat	£22.85
15	Average price	£32.71
16		

Learning checklist

I can *describe* how to summarise rows of data.

I can *describe* how to group rows of data based on logical criteria.

I can *group* and *summarise* rows of data in Excel.

How you can use this lesson

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