Reshaping datasets

Learning intentions

We will be learning to what it means to reshape data, specifically,

- why we **reshape** data
- understand what is meant by **long data**
- understand what is meant by **wide data**

Background

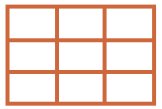
The look of a dataset is often determined by how the data items are collected.

However, it may not be the best form for the analysis you want to perform on it.

In this lesson we will look at **wide and long datasets** and then how we **reshape** them.

Why we reshape datasets?

Some of the reasons we need to reshape the layout of a dataset are,



Some software packages **require data in a certain layout**

To make datasets **easier to view** and analyse

The layout of a dataset can impact on how easy it is to **summarise or group** the data

Show me…

This is an example of data exported from a Microsoft Forms survey.

	A	B	C	F	G	H	I	J	K
1	ID	Start time	Completion time	What is today's date	What type of datasets are we	What type of dataset do you n	How happy are you with this le	Data science is for everyone	Data science is important in m
2	1	12/9/21 8:09:02	12/9/21 8:09:27	12/9/2021	Wide	Long	4		
3	2	12/9/21 8:09:29	12/9/21 8:09:48	2/23/2022	Wide	Long, but it's wide at the momen	5		
4	3	12/9/21 8:09:50	12/9/21 8:10:19	2/12/2022	Long	Long, but needs cleaned	5		
5	4	12/9/21 8:12:54	12/9/21 8:12:58						
6	5	12/9/21 8:13:00	12/9/21 8:13:29	11/30/2022	Wide	Long, with ID, key columns	5	Strongly Agree	Agree
7									
8									

- Every question is in its own column and every response is in a separate row.
- If the survey has more than a few questions the output becomes **very wide and difficult to view**.
- **Changing the layout of the data could make it easier** to plot graphs, group the data or to calculate summary statistics (e.g. min/max/count)

Definition

Long data

Contains an ID column, one or more key columns, and a value column, with every row containing a single observation

Show me…

This is an example of a **long dataset** from Traffic Scotland. Every row contains a single observation.

Live Traffic Information

Type	Date	Location	Description
	26 Oct 21 - 16:00	Multiple	Bridge Wind Restrictions Forecast
	27 Oct 21 - 00:00	Multiple	Yellow - Weather Warnings
	26 Oct 21 - 19:34	A702 A721 - Candy Mill, Northbound & Southbound	Closure
	25 Oct 21 - 05:12	A814 Finnieston - M8 slips, Westbound	Closure
	24 Oct 21 - 07:22	A814 Hayburn Street interchange, Eastbound	Closure
	03 Jun 21 - 00:00	National	COVID19: Latest Travel Information
	17 Mar 21 - 09:56	Glasgow	UN Climate Change Conference of the Parties (COP26) in Glasgow
	22 Mar 21 - 08:17	Rest and Be Thankful	A83 Rest and Be Thankful - Latest Information
	27 Oct 21 - 07:16	M80 J5 Auchenkilns - between slips, Northbound	Accident
	27 Oct 21 - 05:00	Dumfries & Galloway, Any Direction	Surface water

Source: https://trafficscotland.org/

Show me…

This is an example of a **long dataset** from NASA that shows meteorite landings. Every row contains a single observation.

id	nametype	recclass	mass (g)
1	Valid	L5	21
2	Valid	H6	720
6	Valid	EH4	107,000
10	Valid	Acapulcoite	1,914
370	Valid	L6	780
379	Valid	EH4	4,239
390	Valid	LL3-6	910
392	Valid	H5	30,000
[illegible]	[illegible]	[illegible]	[illegible]

Source: https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh/

Show me…

This is a **long** dataset that shows the medals won by nations.

nation	medal_type	number
Sweden	Gold	145
Australia	Gold	147
France	Gold	212
Italy	Gold	206
Australia	Gold	147
France	Gold	212
Sweden	Silver	170
Australia	Silver	163
France	Silver	241
Italy	Silver	178
Australia	Silver	163
France	Silver	241
Sweden	Bronze	179
Australia	Bronze	187
France	Bronze	263
Italy	Bronze	193
Australia	Bronze	187
France	Bronze	263

Each row contains a single observation for each nation and medal type.

Show me…

This is a **long** dataset that shows the temperature from different weather stations in different years.

weather_station	year	temperature
Lerwick	1960	7.5
Armagh	1960	9.4
Eastbourne	1960	10.8
Lerwick	1980	7.0
Armagh	1980	9.0
Eastbourne	1980	10.6
Lerwick	2000	7.5
Armagh	2000	9.9
Eastbourne	2000	11.4
Lerwick	2020	8.0
Armagh	2020	10.7
Eastbourne	2020	12.8

There are a lot of repeated data items.

Structure of a long dataset

id	key	value
1	A	0.50
2	A	0.14
3	A	0.21
1	B	1.23
2	B	5.26
3	B	4.23
1	C	9
2	C	8
3	C	8
1	D	100
2	D	200
3	D	300

In long datasets there are at least 3 columns,

- **id** – labels the observation
- **key** – name of the data item
- **value** – contains the value of the data item

It is possible to have more than one key column.

Note: ID isn't always displayed in reports to users, like in some of the examples we've just seen, but it will be stored

Show me…

purpose_visit	year	month	visitors
Holiday	2020	Mar	584,000
Business	2020	Mar	299,000
Visit friends/relatives	2020	Mar	382,000
Misc	2020	Mar	181,000
Holiday	2020	Feb	891,000
Business	2020	Feb	751,000
Visit friends/relatives	2020	Feb	700,000
Misc	2020	Feb	170,000
Holiday	2020	Jan	1,125,000
Business	2020	Jan	610,000
Visit friends/relatives	2020	Jan	1,115,000
Misc	2020	Jan	186,000

This is a **long** dataset that shows the number of overseas visitors to the UK.

It has 2 **key** columns, **year** and **month**.

Why use long data?

Some benefits for using long data are,

Makes it easier to **group** and **summarise** data

Some **graphs can only be plotted** when the dataset is long

Easier at handling irregular and/or **missing data items**

Definition

Wide data

Each different data variable is in a separate column

Show me…

This is an example of a wide dataset from the Met Office.

Each row contains values for different categories e.g. temperature or cloud cover. To add information for another date/time, you would need to add another column.

Thursday at	06:00	07:00	08:00	09:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00
Chance of precipitation	10%	50%	10%	20%	20%	20%	20%	20%	40%	20%	10%	10%	10%	10%	10%	10%	10%
Temperature °C	13°	13°	13°	13°	13°	13°	14°	14°	14°	14°	14°	14°	13°	13°	12°	12°	12°
Feels like temperature (°C)	12°	12°	12°	12°	13°	13°	13°	13°	13°	13°	13°	12°	12°	12°	11°	11°	11°

Source: https://www.metoffice.gov.uk/weather/forecast/gcvwr3zrw#?date=2021-10-28

Show me…

This is a **wide** dataset that shows the temperature from weather stations.

weather_station	temp1960	temp1980	temp2000	temp2020
Lerwick	7.5	7.0	7.5	8.0
Armagh	9.4	9.0	9.9	10.7
Eastbourne	10.8	10.6	11.4	12.8

The headings contain important information about the data.

This additional information is “lost” in the column headings.

If you wanted to **add a measurement** (i.e. from another year), you would need to add on an **additional column**.

Why use wide data?

Some benefits for using wide data are,

no repetition of data items

Easier to calculate **time between date points**

Easier to **calculate simple statistics** such as average or total on a **single column**

Your turn…

Do you think this dataset is **wide** or **long**?

location	wind_speed_monday	wind_speed_tuesday
Thornhill	10.25	9.73
Kingsbarns	13.54	6.86
Dalkeith	8.01	7.80
Killin	6.25	8.70
Kirkcudbright	5.45	5.09

Your turn…

Do you think this dataset is **wide** or **long**?

location	wind_speed_monday	wind_speed_tuesday
Thornhill	10.25	9.73
Kingsbarns	13.54	6.86
Dalkeith	8.01	7.80
Killin	6.25	8.70
Kirkcudbright	5.45	5.09

This dataset is **wide**.

Each data variable
(e.g. wind_speed_monday,
wind_speed_tuesday)
is in a separate column.

Your turn…

location	day	wind_speed
Thornhill	Monday	10.25
Kingsbarns	Monday	13.54
Dalkeith	Monday	8.01
Killin	Monday	6.25
Kirkcudbright	Monday	5.45
Thornhill	Tuesday	9.73
Kingsbarns	Tuesday	6.86
Dalkeith	Tuesday	7.80
Killin	Tuesday	8.70
Kirkcudbright	Tuesday	5.09

This is how the same data would look like in a **long** dataset.

- **id** – location
- **key** – day
- **value** – wind_speed

Next steps

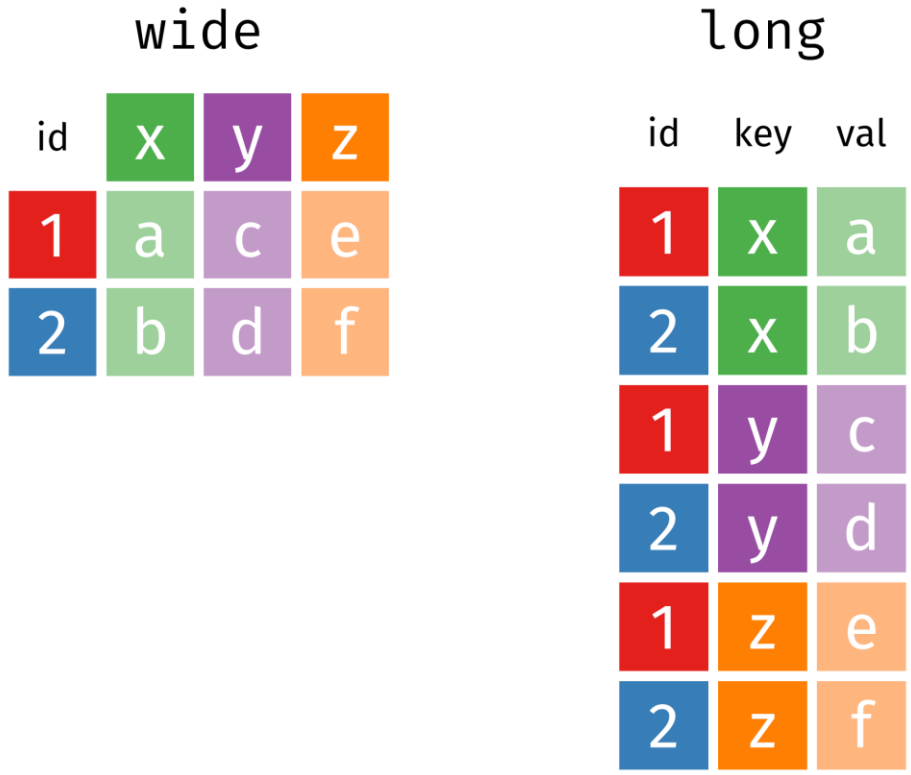
Complete **questions 1 to 6**
in **section 1** of the
‘Reshaping datasets in Excel’ workbook.

Reshaping data

As we have seen there are benefits of both wide and long datasets.

The analysis you want to perform on the dataset will determine the shape you need your data in.

We are now going to look at **reshaping datasets.**



Source: Garrick Aden-Buie's (@grrrck) Tidy Animated Verbs

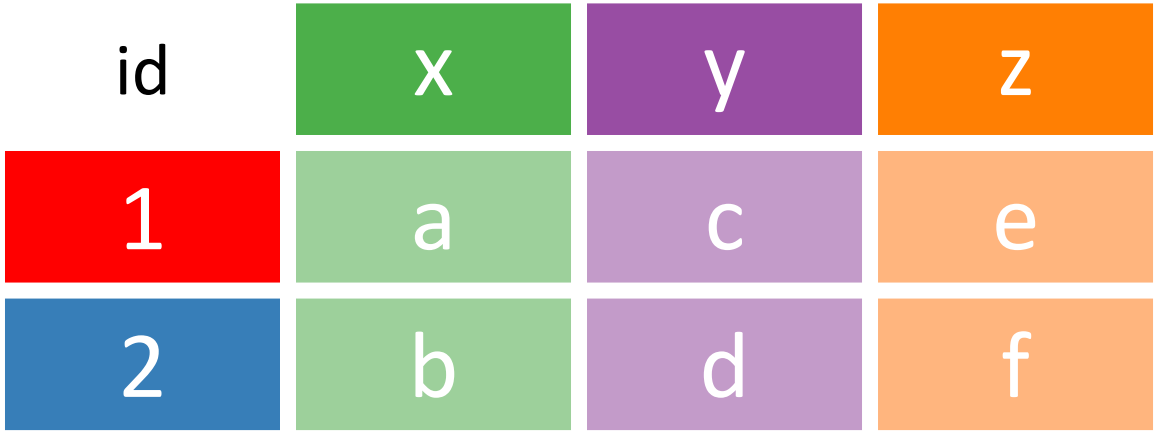
Definition

Reshape data

To convert data from wide to long or vice versa

Wide to long reshaping

This is our wide dataset, we need to reshape it into a long dataset.



Wide to long reshaping

When reshaping a dataset, the **id column remains the same**.

The long dataset needs at least 2 new columns, **key** and **value**

id	x	y	z
1	a	c	e
2	b	d	f

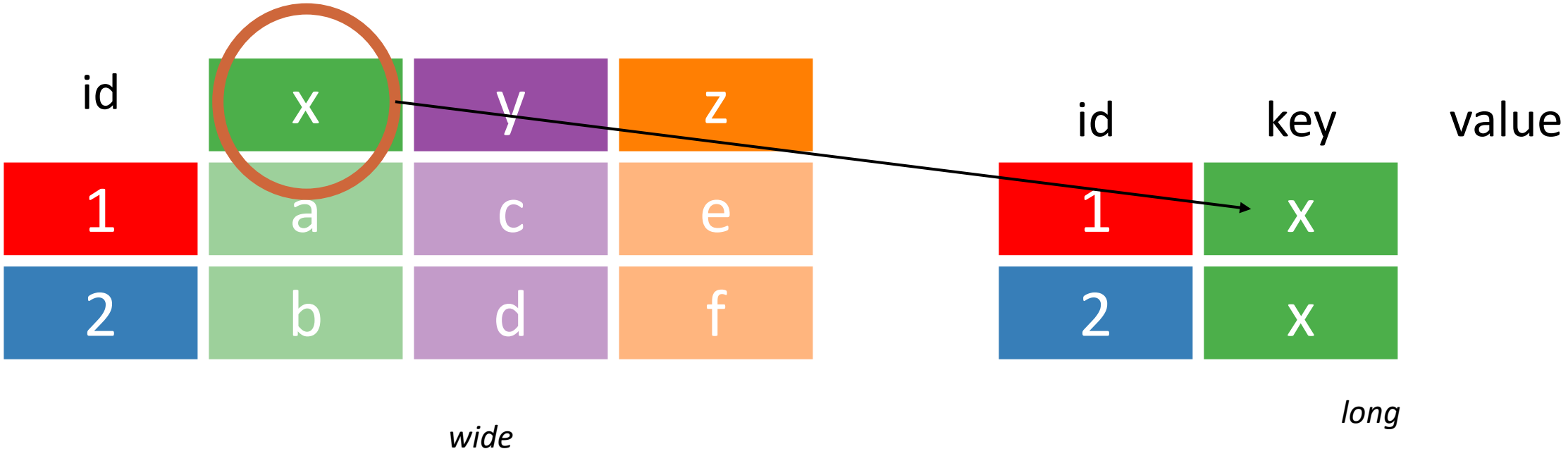
wide

id	key	value
1		
2		

long

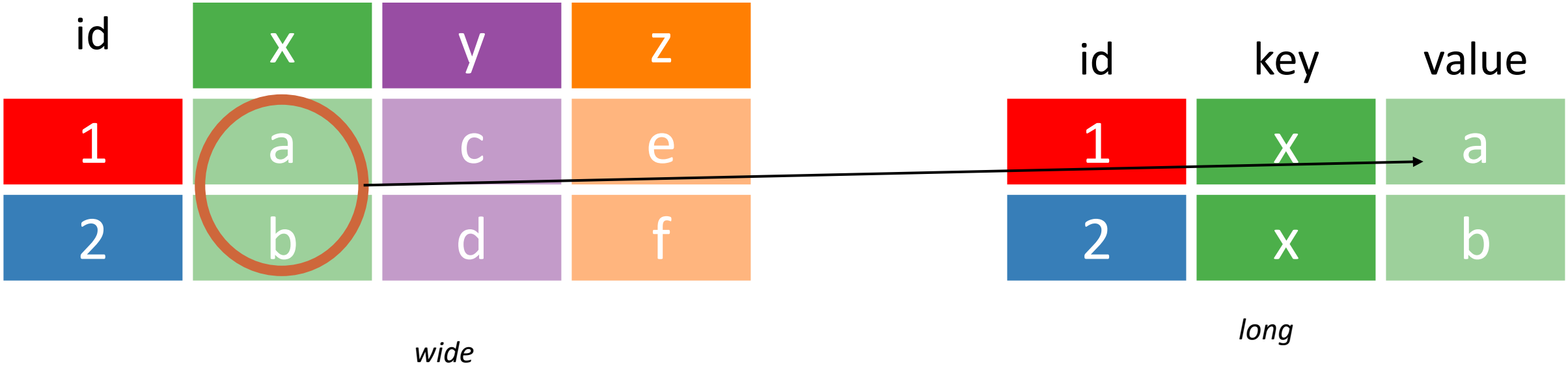
Wide to long reshaping

The information in the headers becomes the data items in the **key** variable.



Wide to long reshaping

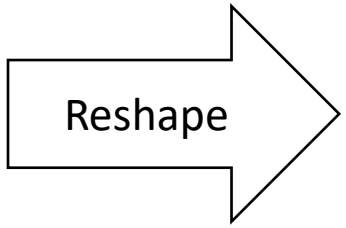
The data items from the **x** column have been placed next to the x **key**.



Wide to long reshaping

The process is repeated for the rest of the columns in the wide dataset.

id	x	y	z
1	a	c	e
2	b	d	f



id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

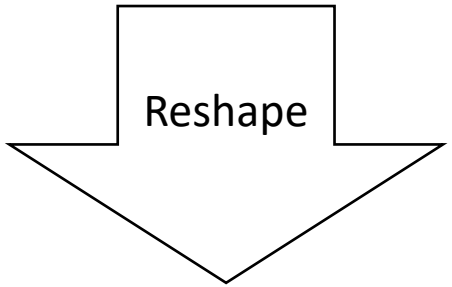
Example

Reshape this **wide** dataset into a **long** dataset.

name	class	test_1	test_2	test_3
Ross McDonald	4F	90%	85%	75%
Chloe Warwick	8G	85%	92%	93%
Finn Marcus	1P	75%	56%	80%
Erin Malcolm	5R	91%	95%	90%

Example

name	class	test_1	test_2	test_3
Ross McDonald	4F	90%	85%	75%
Chloe Warwick	8G	85%	92%	93%
Finn Marcus	1P	75%	56%	80%
Erin Malcolm	5R	91%	95%	90%



name	class	test	result
		1	
		2	
		…	

The columns **name** and **class** will still be in the long dataset.

There are now 2 new columns,

- **test** (the key column)
- **result** (the value column)

Example

This dataset is now **long**.

name	class	test	result
Ross McDonald	4F	1	90%
Chloe Warwick	8G	1	85%
Finn Marcus	1P	1	75%
Erin Malcolm	5R	1	91%
Ross McDonald	4F	2	85%
Chloe Warwick	8G	2	92%
Finn Marcus	1P	2	56%
Erin Malcolm	5R	2	95%
Ross McDonald	4F	3	75%
Chloe Warwick	8G	3	93%
Finn Marcus	1P	3	80%
Erin Malcolm	5R	3	90%

The **test number** is now in the **column as a data item** rather than part of the heading name

The name and class information is **repeated** many times in the dataset

There is a **new column** that contains all the results

Your turn…

You need to work out the **time taken** for each of the races in Leven, Stirling, Dumfries and Glasgow.

It would be easier to calculate if the dataset was **wide**.

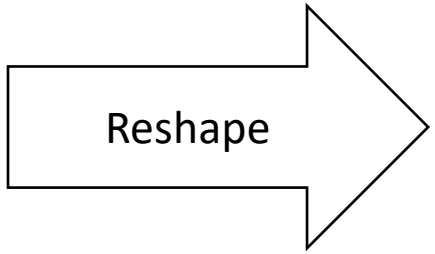
What do you think the **column headings** could be in your new dataset if you **reshaped this dataset**?

Reminder: the data items in the **key** column in a **long** dataset will becomes the **headings** in a **wide** dataset.

race	time_type	time
Leven	start	10:00
Stirling	start	09:30
Dumfries	start	10:15
Glasgow	start	10:30
Leven	end	13:04
Stirling	end	20:08
Dumfries	end	12:02
Glasgow	end	16:30

Your turn…

race	time_type	time
Leven	start	10:00
Stirling	start	09:30
Dumfries	start	10:15
Glasgow	start	10:30
Leven	end	13:04
Stirling	end	20:08
Dumfries	end	12:02
Glasgow	end	16:30



race	start_time	end_time
Leven	10:00	13:04
Stirling	09:30	20:08
Dumfries	10:15	12:02
Glasgow	10:30	16:30

The column headings could be,

- **race**
- **start_time**
- **end_time**

Your turn…

Now the dataset has been reshaped you can calculate the

time_taken = end_time – start_time

race	start_time	end_time	time_taken
Leven	10:00	13:04	03:04
Stirling	09:30	20:08	10:38
Dumfries	10:15	12:02	01:47
Glasgow	10:30	16:30	06:00

Next steps

Complete **questions 1 to 6**
in **section 2** of the
‘Reshaping datasets in Excel’ workbook.

Learning checklist

I can *describe* the difference between long and wide data

I can *identify* when to use long and wide data

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