Combining datasets

Version: 1.0

effini

THE DATA LAB

Learning intentions

We will be learning **how to combine datasets**, specifically

- what we mean by **combining datasets**
- to add rows to a dataset by **appending**
- to add columns to a dataset by **joining**
- understand common **types of joins**

Background

When a data analyst is given a dataset to analyse, **most of their time is spent manipulating** it to allow them to conduct the analysis.

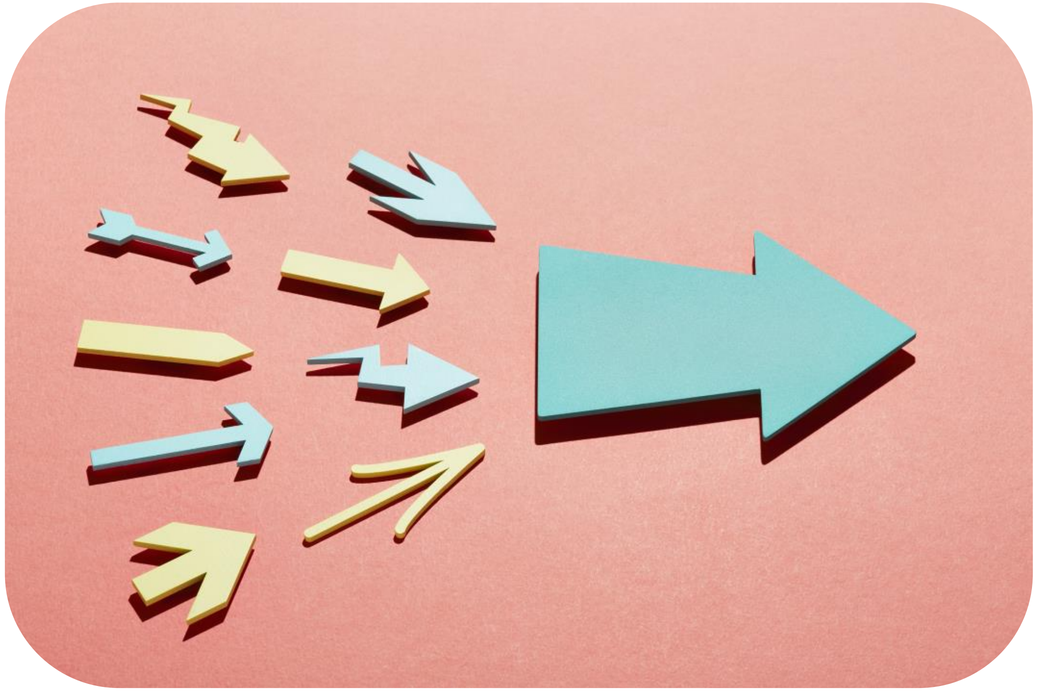
In previous lessons, we have looked at manipulating datasets by selecting columns and filtering rows.

In this lesson we will look at manipulating datasets by **combining data** from more than one dataset.

Combining multiple datasets

Up until now all the data manipulation we have done has been on a single dataset. However it is usual for an analyst to undertake analysis just using a single dataset.

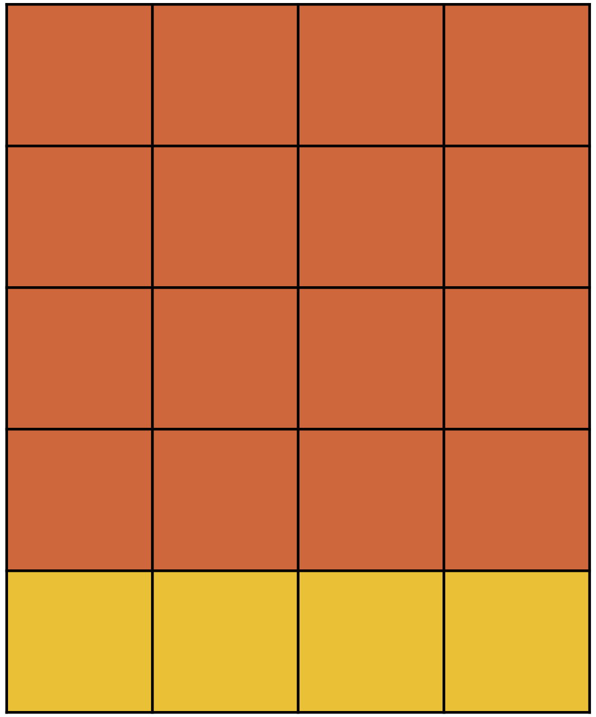
Therefore combining datasets is something that commonly needs to be done.



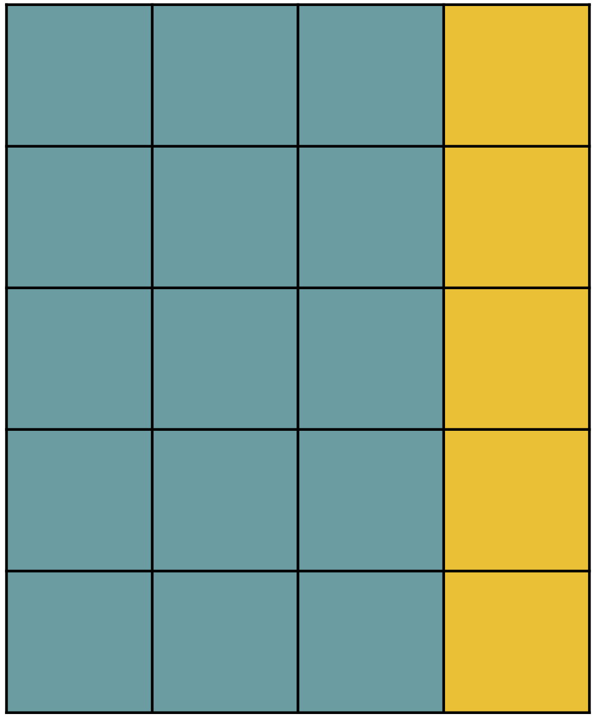
Combining datasets

Data items can be combined into a dataset by either,

…adding rows



…or adding columns



Why it’s important to combine datasets?

Allows you to **fill in gaps** in your dataset

Improves decision-making by seeing all your data together

To **add newer data** to an existing dataset

To **supplement** the dataset you have with other data

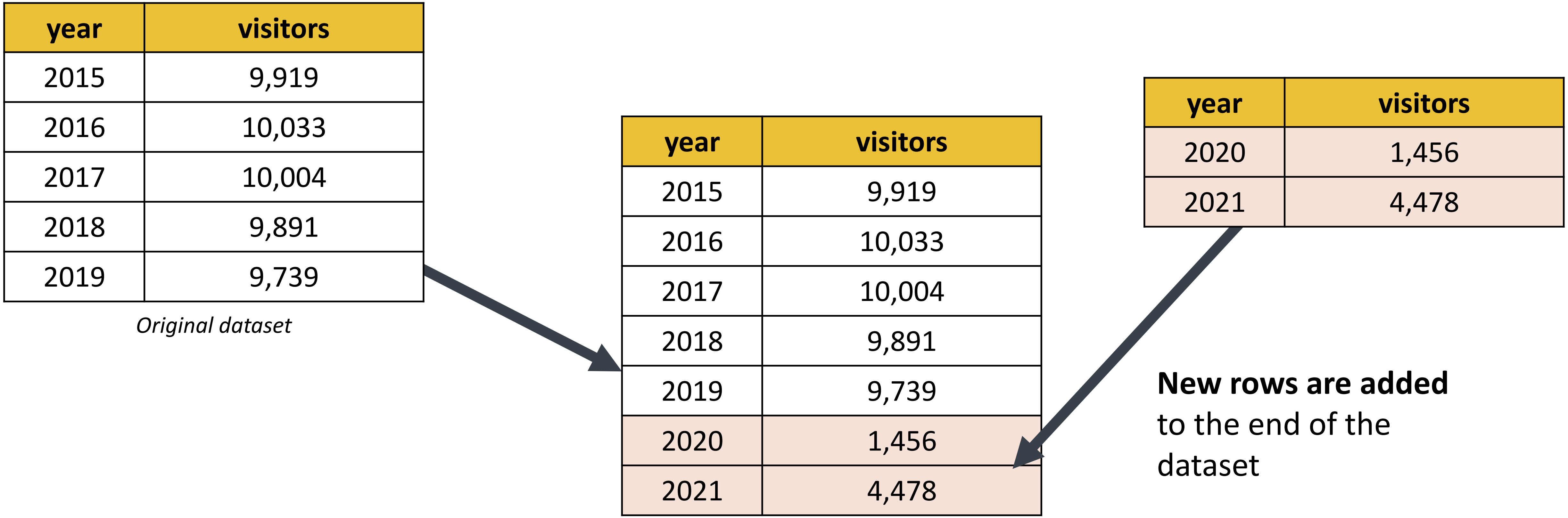
Definition

Append

To add rows to the end of a dataset

Show me...adding rows

A museum has a dataset that shows the number of visitors, they have **appended the most recent visitor numbers** to their dataset.



year	visitors
2015	9,919
2016	10,033
2017	10,004
2018	9,891
2019	9,739

Original dataset

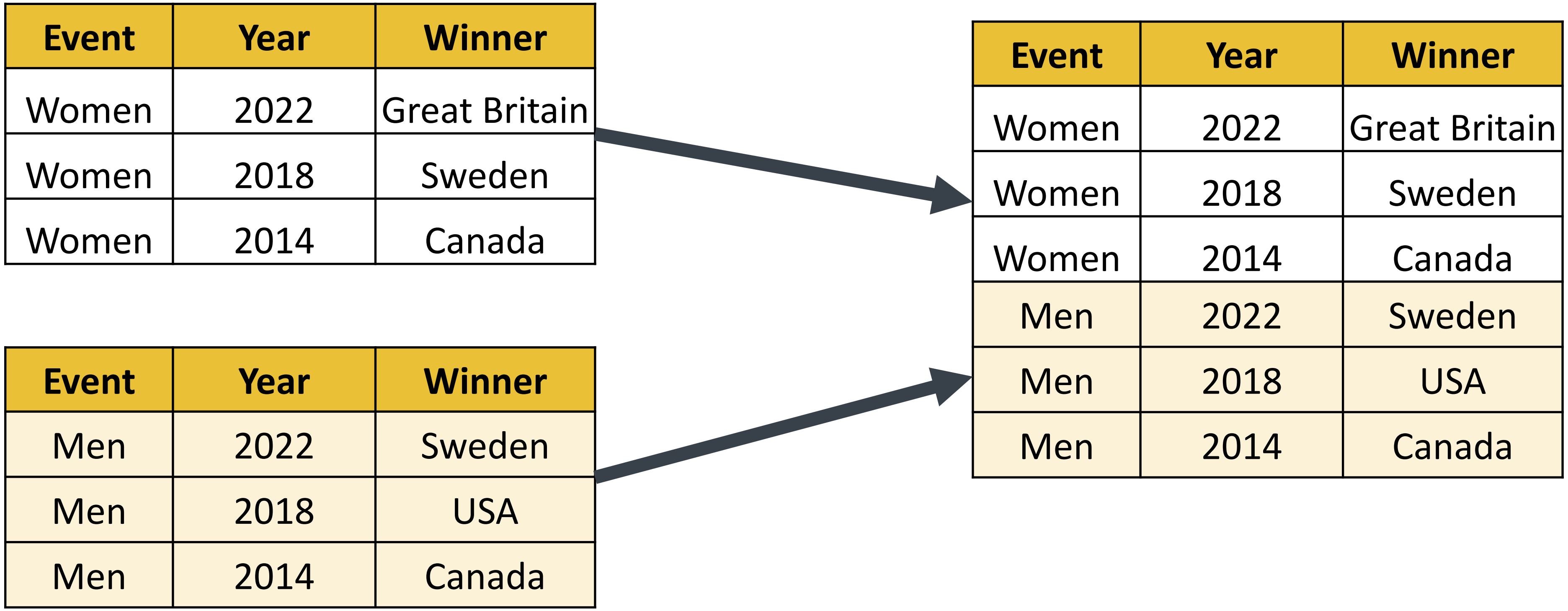
year	visitors
2020	1,456
2021	4,478

year	visitors
2015	9,919
2016	10,033
2017	10,004
2018	9,891
2019	9,739
2020	1,456
2021	4,478

New rows are added to the end of the dataset

Example

The gold medal winners of the Winter Olympics Curling are held in two datasets. They have been **combined into one dataset by appending the rows**.



Event	Year	Winner
Women	2022	Great Britain
Women	2018	Sweden
Women	2014	Canada

Event	Year	Winner
Men	2022	Sweden
Men	2018	USA
Men	2014	Canada

Event	Year	Winner
Women	2022	Great Britain
Women	2018	Sweden
Women	2014	Canada
Men	2022	Sweden
Men	2018	USA
Men	2014	Canada

Your turn...

If we append the second dataset to the first dataset, how **many rows will there be in the final dataset**?

River	Continent	Length
Nile	Africa	6690
Congo	Africa	4371
Niger	Africa	4167
Zambezi	Africa	2693

River	Continent	Length
Amazon	South America	6387
Parana	South America	3998

Your turn…

Now the dataset has been appended, **there are 6 rows** in the final dataset.

River	Continent	Length
Nile	Africa	6690
Congo	Africa	4371
Niger	Africa	4167
Zambezi	Africa	2693
Amazon	South America	6387
Parana	South America	3998

Checks before appending a dataset

Before appending rows you need to make sure that both datasets have exactly the same columns. This means they have,

- **Same number** of columns
- Columns are in the **same order**
- Columns contain the **same data**

Sometimes columns in one of the datasets may need to be created, removed or reordered before you can append the rows.

Example… number of columns

We need to append the bottom dataset to the top dataset. However they **don’t have the same number of columns**.

name	street	postcode
Edinburgh Castle	Castlehill	EH1 2NG
Stirling Castle	Castle Wynd	FK8 1EJ
Kelvingrove Museum	Argyle Street	G3 8AG
Falkland Place	East Port	KY15 7BY

name	postcode
Glenfinnan Viaduct	PH37 4LT
Greyfriars Bobby	EH1 2QE

Photo by Jörg Angeli on Unsplash

Example

To allow us to append the datasets, we need to **add an empty column** called “street” into the bottom dataset. They now have the same number of columns.

name	street	postcode
Edinburgh Castle	Castlehill	EH1 2NG
Stirling Castle	Castle Wynd	FK8 1EJ
Kelvingrove Museum	Argyle Street	G3 8AG
Falkland Palace	East Port	KY15 7BY

name	street	postcode
Glenfinnan Viaduct	BLANK	PH37 4LT
Greyfriars Bobby	BLANK	EH1 2QE

The new column is filled with blank or empty data items

Example

The rows can now be **appended to the dataset**. The final dataset has 6 rows.

name	street	postcode
Edinburgh Castle	Castlehill	EH1 2NG
Stirling Castle	Castle Wynd	FK8 1EJ
Kelvingrove Museum	Argyle Street	G3 8AG
Falkland Palace	East Port	KY15 7BY
Glenfinnan Viaduct	BLANK	PH37 4LT
Greyfriars Bobby	BLANK	EH1 2QE

Example…order of columns

These datasets contain the depth of oceans and they need to be appended.

They have the **same number of columns** but they are **not in the same order**.

ocean	depth_m
Pacific	3,970
Atlantic	3,646

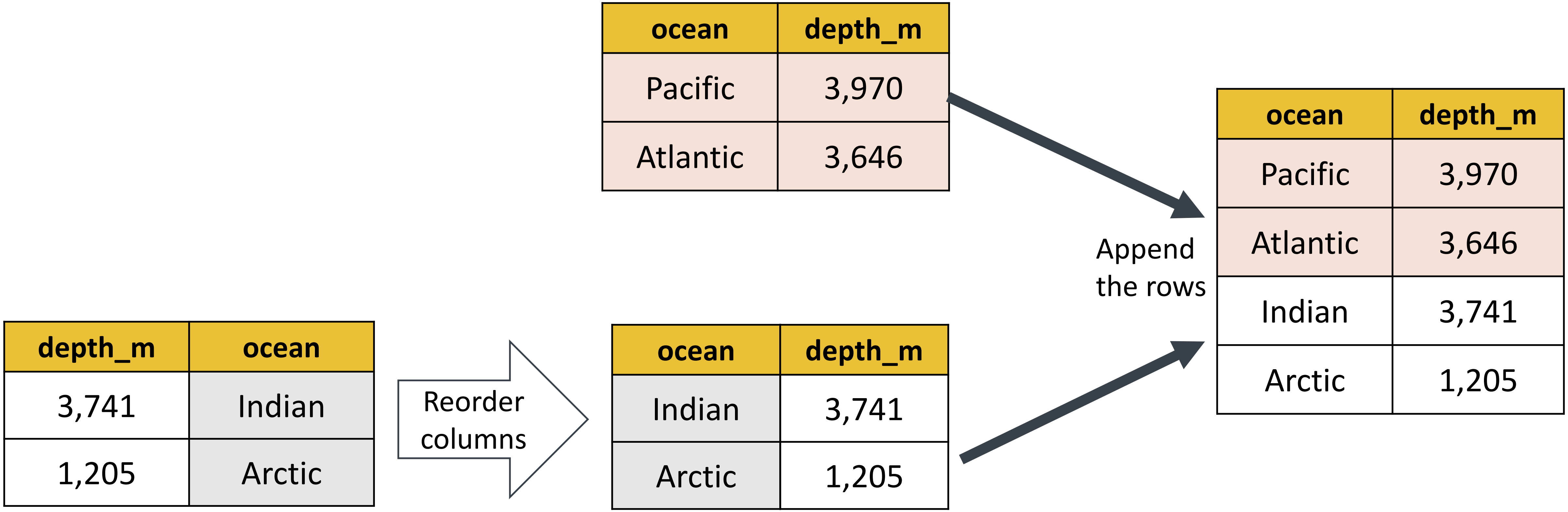
depth_m	ocean
3,741	Indian
1,205	Arctic

Example…order of columns

Reorder

To change the order of columns in a dataset

Before appending the rows, the columns in the bottom dataset have been **reordered**.



ocean	depth_m
Pacific	3,970
Atlantic	3,646

depth_m	ocean
3,741	Indian
1,205	Arctic

Reorder columns

ocean	depth_m
Indian	3,741
Arctic	1,205

Append the rows

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

Your turn…

What changes would need to be made to the bottom dataset so the rows can be appended to the top dataset?

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

sales	product	price
£15.00	Dragon fruit	£1.00
£12.50	Pepper	£0.50

Your turn...

What changes would need to be made to the bottom dataset so the rows can be appended to the top dataset?

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

product	colour	type	price	sales
Dragon fruit	BLANK	BLANK	£1.00	£15.00
Pepper	BLANK	BLANK	£0.50	£12.50

1. **Two empty columns** (colour and type) have been created.

2. The columns have been **reordered** to match the order of the top dataset.

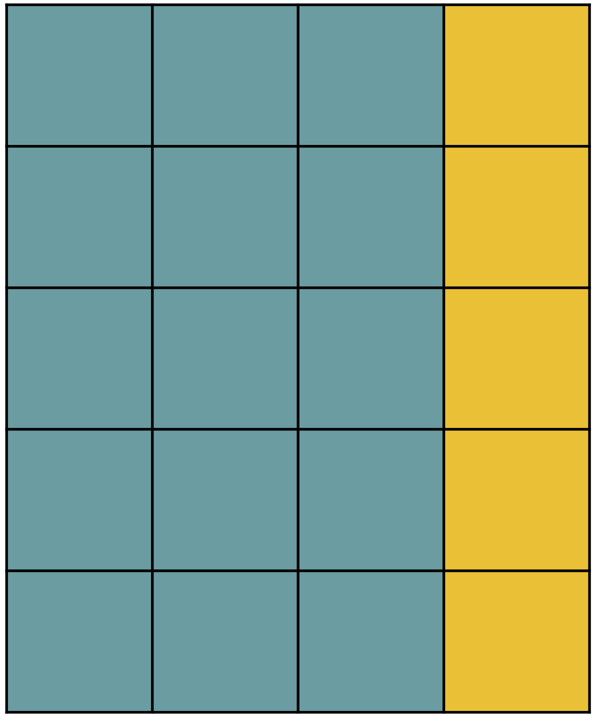
Next steps

Complete **questions 1 to 10**
in **section 1** of the
‘Combining datasets’ workbook.

Combining datasets by adding columns

We have looked at combining datasets by appending rows.

We are now going to look at **combining datasets by adding columns**.

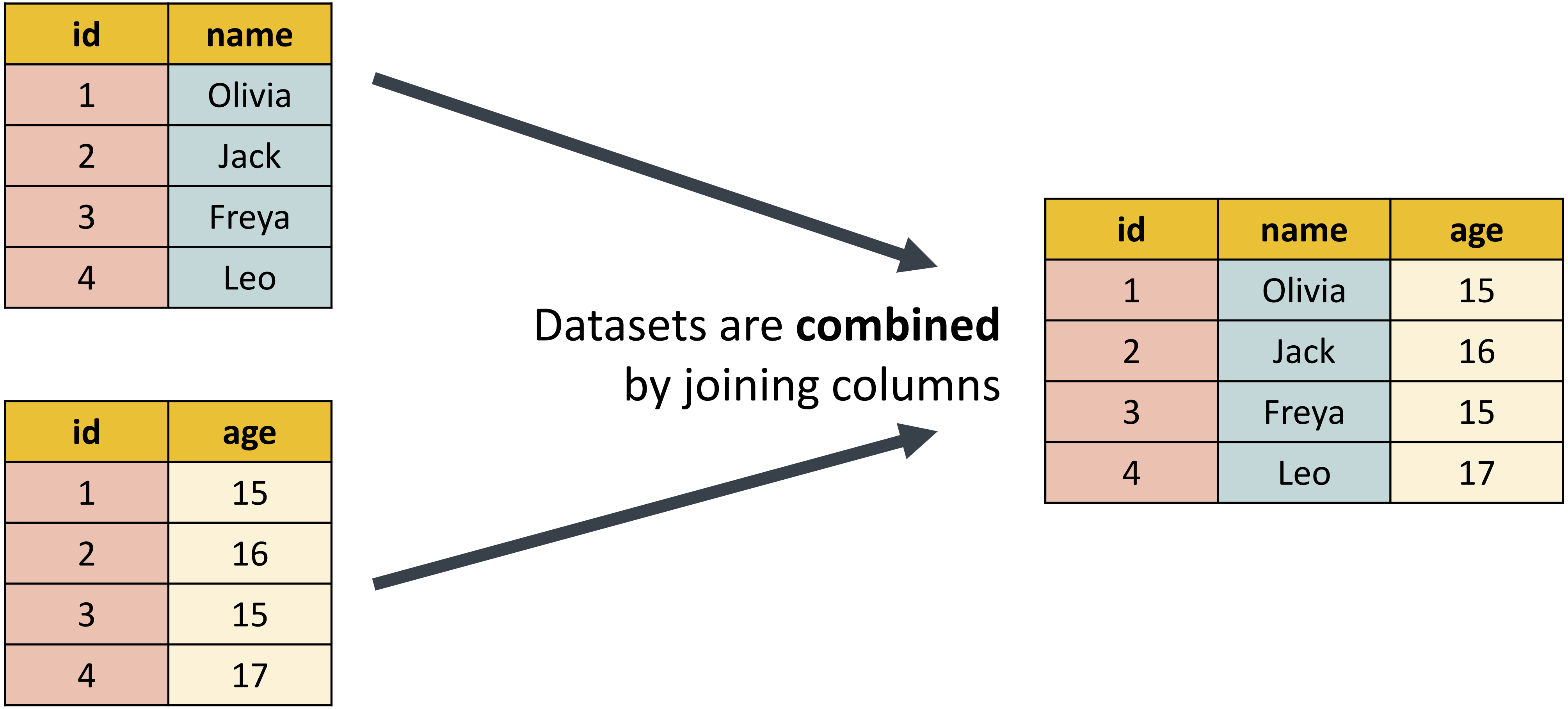


Definition

Join datasets

To add columns from one dataset to another dataset

Show me...



id	name
1	Olivia
2	Jack
3	Freya
4	Leo

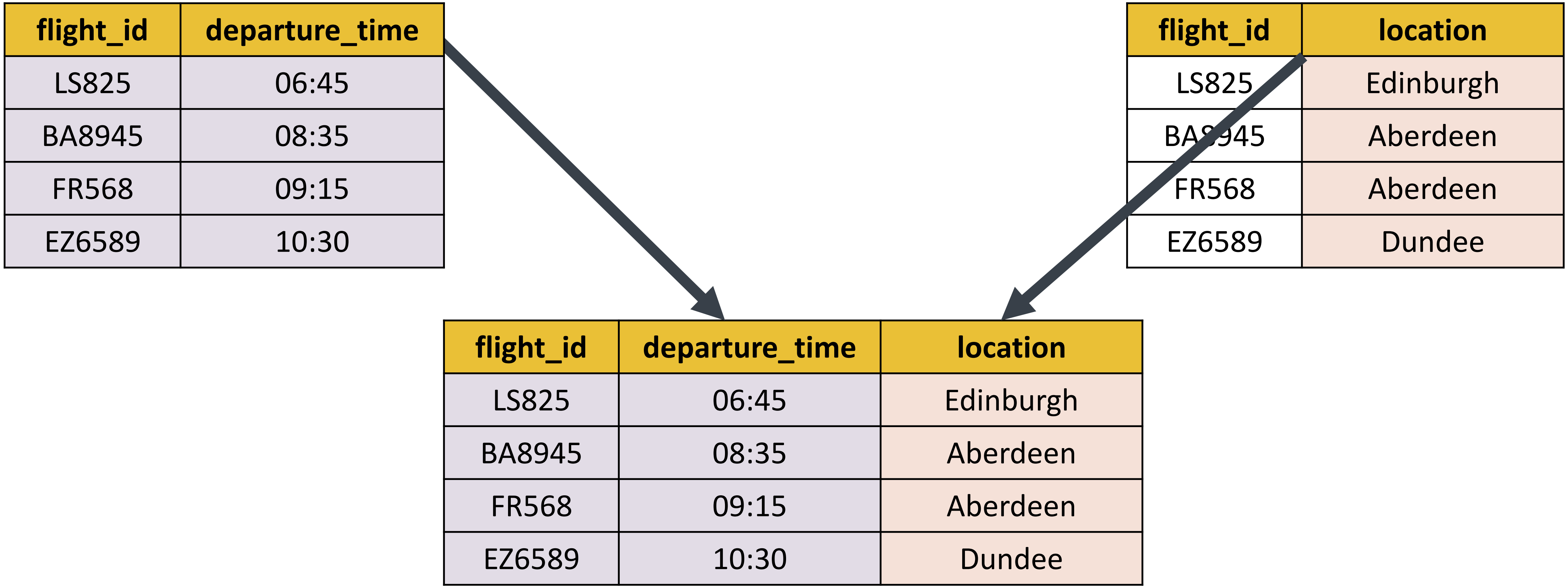
id	age
1	15
2	16
3	15
4	17

Datasets are **combined** by joining columns

id	name	age
1	Olivia	15
2	Jack	16
3	Freya	15
4	Leo	17

Example

You’ve been asked to find out **what time the flight leaves from Dundee**. However the information for the airport location and the flight times is stored in 2 different datasets. Therefore **you need to join the datasets**.



flight_id	departure_time
LS825	06:45
BA8945	08:35
FR568	09:15
EZ6589	10:30

flight_id	location
LS825	Edinburgh
BA8945	Aberdeen
FR568	Aberdeen
EZ6589	Dundee

flight_id	departure_time	location
LS825	06:45	Edinburgh
BA8945	08:35	Aberdeen
FR568	09:15	Aberdeen
EZ6589	10:30	Dundee

Definition

Key

Column(s) that the datasets have common

Show me… key columns

When joining data, the datasets need to have at least one column in common that contains the same information.

Name	HomeTown
Mike	Falkirk
Freya	Dumfries
Isla	Inverness
Gail	Wick

ID	Name	Event
1	Mike	Long jump
2	Freya	Pole vault
3	Isla	Pole vault
4	Gail	Hammer throw

These columns contain the same information.

Properties of a key column

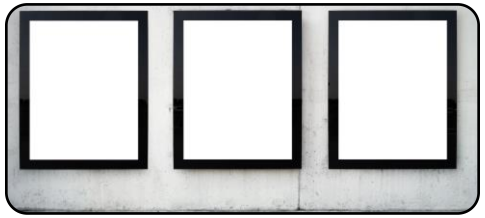
The data items in the key columns (as well as appearing in both datasets) need to be,

Unique values

- Each row much contain a unique value, which means there are no duplicates

Complete values

- All the rows need to contain values, with no missing or incomplete data items



Non-NULL values

- There cannot be any empty or blank data items

Show me…

These datasets both contain information related to planets. The **planet** column is the key.

These are the **key** columns

planet	diameter_km
Mercury	4,879
Venus	12,104
Earth	12,742
Mars	6,779
Jupiter	139,820
Saturn	116,460
Uranus	50,724
Neptune	46,244

planet	number_moons
Mercury	0
Venus	0
Earth	1
Mars	2
Jupiter	95
Saturn	83
Uranus	27
Neptune	14

Your turn…

What is the **key column** in these datasets?

forest_id	habitat
46841	Upland oakwood
48472	Upland birchwood
55076	Wet woodland
51517	Upland oakwood

forest_id	maturity
46841	Mature
48472	Young
55076	Mixed
51517	Mature

Your turn…

What is the **key column** in these datasets?

forest_id	habitat
46841	Upland oakwood
48472	Upland birchwood
55076	Wet woodland
51517	Upland oakwood

forest_id	maturity
46841	Mature
48472	Young
55076	Mixed
51517	Mature

forest_id is the key column for these datasets.

It is the column that the datasets have in common.

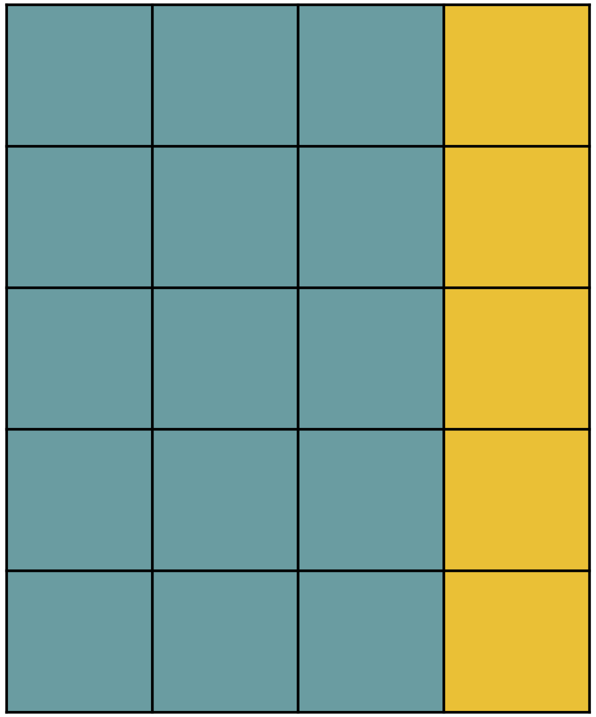
Next steps

Complete **questions 1 to 6**
in **section 2** of the
'Combining datasets' workbook.

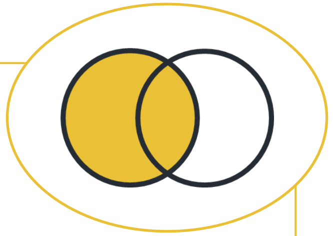
Joining columns to a dataset

There are 4 ways of joining datasets. They are,

- **Left** join
- **Right** join
- **Inner** join
- **Outer** join



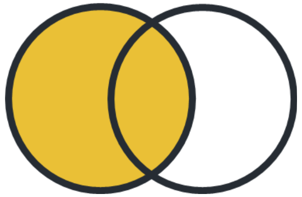
Definition



Left join

Returns all the values from the left dataset and any matching records from the right dataset

Show me…left join



We need to combine these two datasets using a **LEFT** join using the **KEY** column "Name".

Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

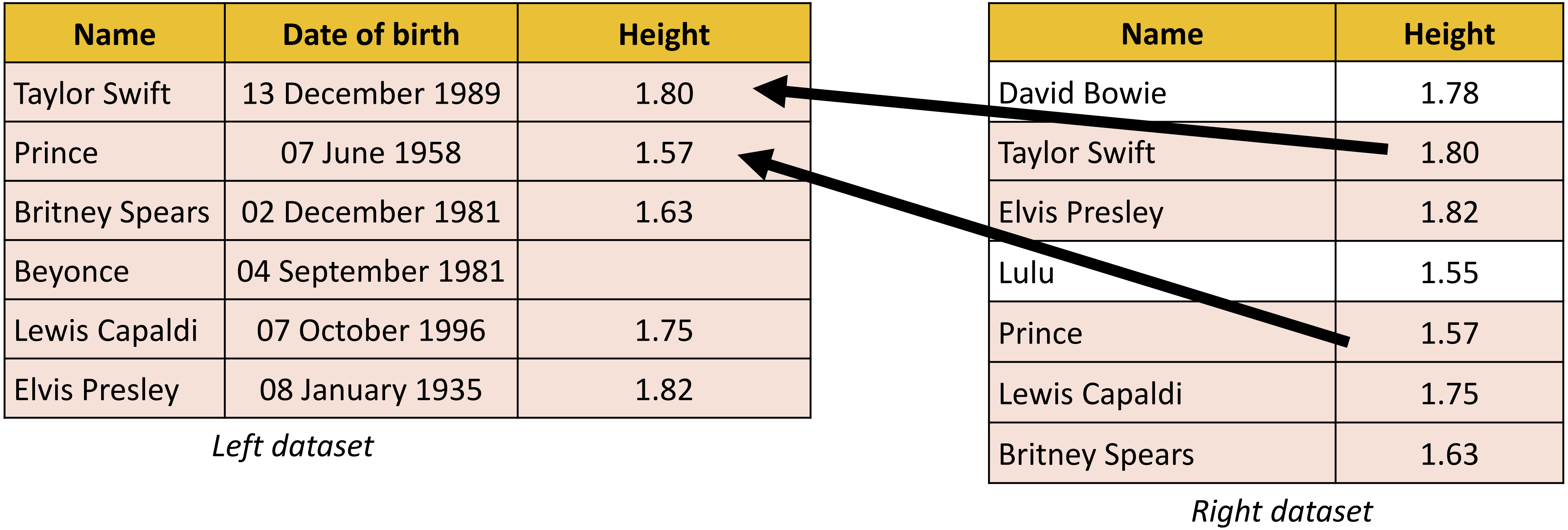
Left dataset

Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Right dataset

Show me…left join

For a left join, all the columns from the left dataset are kept and columns from the right dataset are added.



Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Beyonce	04 September 1981	
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82

Left dataset

Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Right dataset

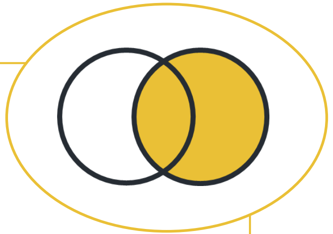
Show me…left join

This is the final dataset after it has been joined.

Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Beyonce	04 September 1981	
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82

Beyonce’s height was not in the right dataset, so we now have a gap in the height column.

Definition



Right join

Returns all the values from the right dataset and any matching records from the left dataset

Show me...right join

We are combining these two datasets again. The key is common is still **Name**, but this time using a **RIGHT** join.

Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Left dataset

Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Right dataset

Show me…right join

This time the data from the left dataset is **added into the right dataset**,

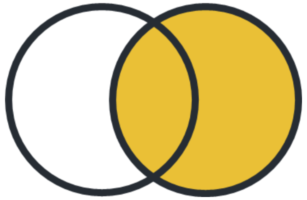
Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Left dataset

Name	Date of birth	Height
David Bowie		1.78
Taylor Swift	13 December 1989	1.80
Elvis Presley	08 January 1935	1.82
Lulu		1.55
Prince	07 June 1958	1.57
Lewis Capaldi	07 October 1996	1.75
Britney Spears	02 December 1981	1.63

Right dataset

Show me…right join



All of the data from the right dataset is kept, and any matching data from the left dataset is added in.

Name	Date of birth	Height
David Bowie		1.78
Taylor Swift	13 December 1989	1.80
Elvis Presley	08 January 1935	1.82
Lulu		1.55
Prince	07 June 1958	1.57
Lewis Capaldi	07 October 1996	1.75
Britney Spears	02 December 1981	1.63

David Bowie’s date of birth is missing in the left dataset, so we now have a gap in the date of birth column.

Left vs. right join

The choice of type of join will change how your final dataset looks.

Left join

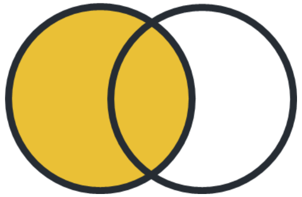
Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Beyonce	04 September 1981	**MISSING**
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82

Right join

Name	Date of birth	Height
David Bowie	**MISSING**	1.78
Taylor Swift	13 December 1989	1.80
Elvis Presley	08 January 1935	1.82
Lulu	**MISSING**	1.55
Prince	07 June 1958	1.57
Lewis Capaldi	07 October 1996	1.75
Britney Spears	02 December 1981	1.63

Your turn...

We are going to **LEFT join** these datasets,
how many rows will the final dataset have?



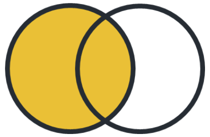
author	book
Muriel Spark	The Prime of Miss Jean Brodie
J. R. R. Tolkien	The Lord of the Rings
Irvine Welsh	Trainspotting
J. R. R. Tolkien	The Hobbit
J.K. Rowling	Harry Potter and the Chamber of Secrets

Left dataset

author	publisher
Muriel Spark	Macmillan
J. R. R. Tolkien	Allen & Unwin
Irvine Welsh	Secker & Warburg

Right dataset

Your turn...



Using a LEFT join, there are **5 rows** in the final dataset.
This is the same number of rows as the left dataset.

author	book	publisher
Muriel Spark	The Prime of Miss Jean Brodie	Macmillan
J. R. R. Tolkien	The Lord of the Rings	Allen & Unwin
Irvine Welsh	Trainspotting	Secker & Warburg
J. R. R. Tolkien	The Hobbit	Allen & Unwin
J.K. Rowling	Harry Potter and the Chamber of Secrets	MISSING

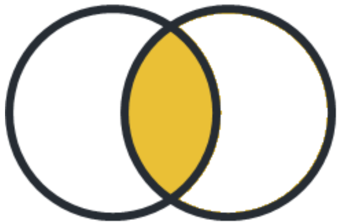
Final dataset

Definition

Inner join

Returns data items whenever there are matching values in both datasets

Show me…



We are going to combine these datasets again, but this time using an **INNER** join.

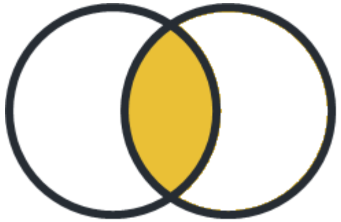
Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Left dataset

Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Right dataset

Show me…inner join



The final dataset only includes **data items that appear in both** the left and right datasets.

Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82

Definition

Outer (full) join

No information is lost, since it merges any data in either dataset

Show me…

We are going to combine these datasets again, but this time using an **OUTER** join.

Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Left dataset

Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Right dataset

Show me… outer join

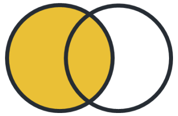
Using an outer join means all the information from both datasets ends up in the final dataset.

Name	Date of birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Beyonce	04 September 1981
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Beyonce	04 September 1981	
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82
David Bowie		1.78
Lulu		1.55

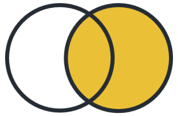
Name	Height
David Bowie	1.78
Taylor Swift	1.80
Elvis Presley	1.82
Lulu	1.55
Prince	1.57
Lewis Capaldi	1.75
Britney Spears	1.63

Comparing the joins



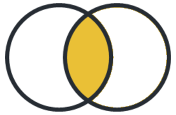
Left join

Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Beyonce	04 September 1981	
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82



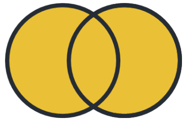
Right join

Name	Date of birth	Height
David Bowie		1.78
Taylor Swift	13 December 1989	1.80
Elvis Presley	08 January 1935	1.82
Lulu		1.55
Prince	07 June 1958	1.57
Lewis Capaldi	07 October 1996	1.75
Britney Spears	02 December 1981	1.63



Inner join

Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82



Outer join

Name	Date of birth	Height
Taylor Swift	13 December 1989	1.80
Prince	07 June 1958	1.57
Britney Spears	02 December 1981	1.63
Beyonce	04 September 1981	
Lewis Capaldi	07 October 1996	1.75
Elvis Presley	08 January 1935	1.82
David Bowie		1.78
Lulu		1.55

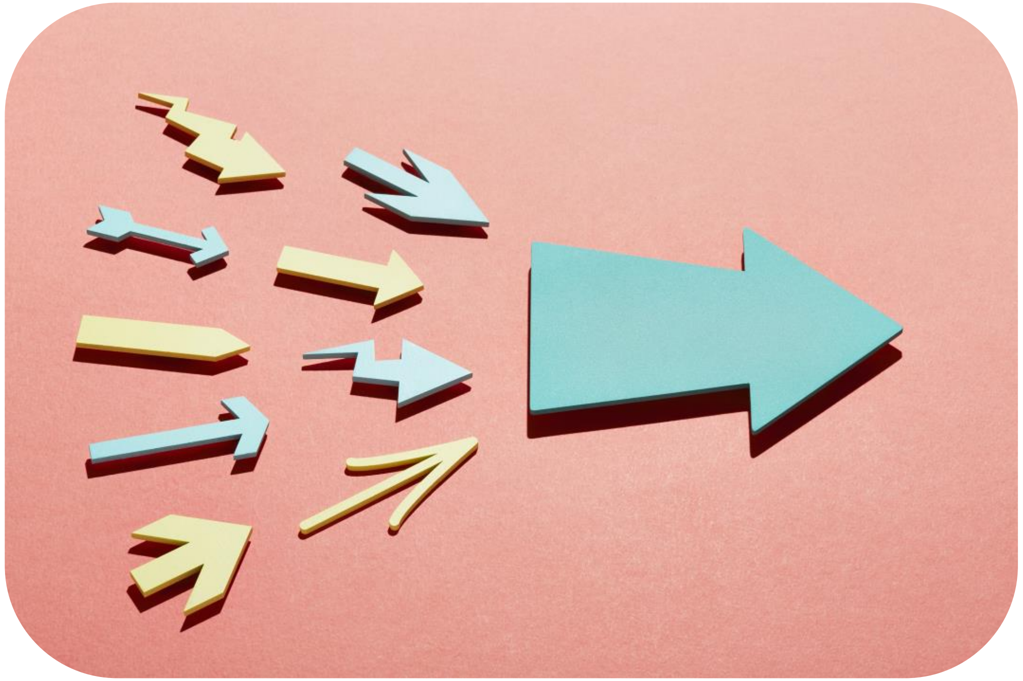
Next steps

Complete **questions 1 to 9**
in **section 3** of the
‘Combining datasets’ workbook.

What happens if combining goes wrong…

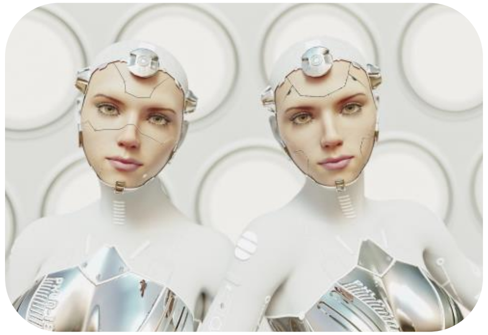
When manipulating datasets there is the possibility that the final dataset may not look as you expect.

We are going to look at some of **common issues that can arise when combining datasets.**



Common causes of joining issues

Sometimes when you join datasets, the final dataset doesn’t look as you expect. Here are some common issues that can cause problems when joining datasets.



Duplicate rows

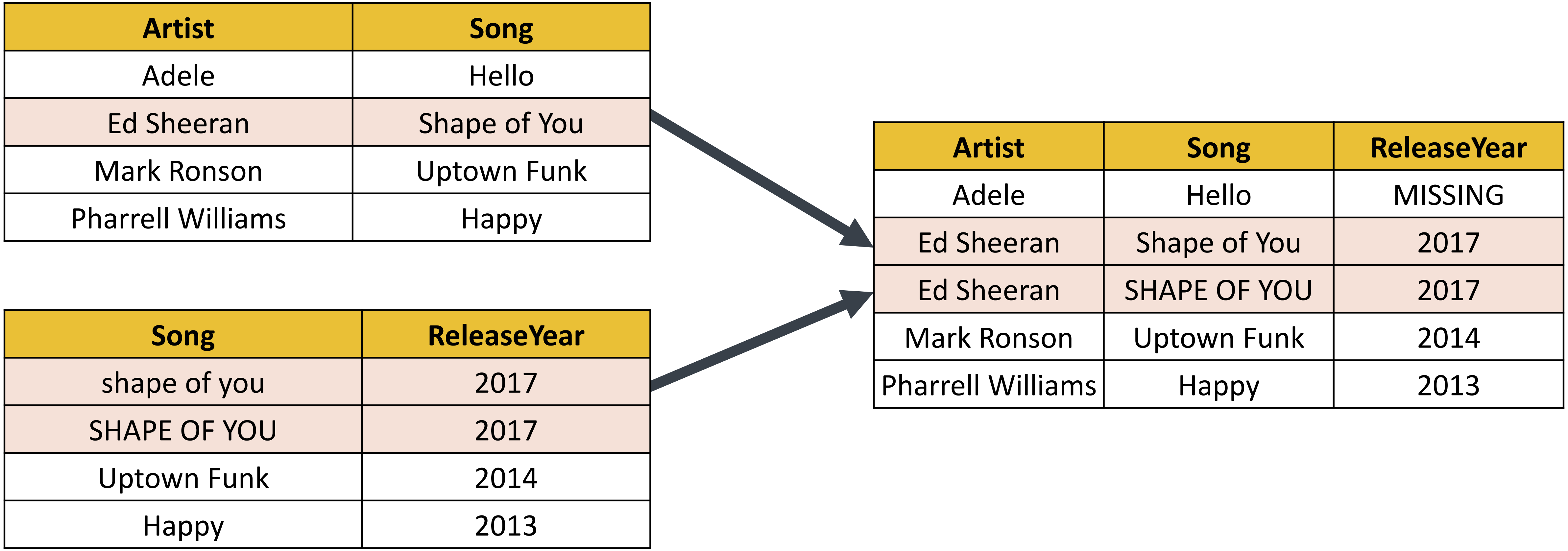
Extra spaces between text

Data types not matching

Incorrect join type used

Show me… duplicate rows

If one of the datasets has duplicate rows, when you join them you will end up with extra rows that you were not expecting.



Artist	Song
Adele	Hello
Ed Sheeran	Shape of You
Mark Ronson	Uptown Funk
Pharrell Williams	Happy

Song	ReleaseYear
shape of you	2017
SHAPE OF YOU	2017
Uptown Funk	2014
Happy	2013

Artist	Song	ReleaseYear
Adele	Hello	MISSING
Ed Sheeran	Shape of You	2017
Ed Sheeran	SHAPE OF YOU	2017
Mark Ronson	Uptown Funk	2014
Pharrell Williams	Happy	2013

Show me… extra spaces

When joining these datasets, the data items for the movie “Mary Poppins” would not be combined as there are extra spaces between ‘Mary’ and ‘Poppins’ in the bottom dataset.

movie	year
Toy Story	1995
Up	2009
Mary Poppins	1964
The Lion King	1994
Moana	2016

movie	budget_$_million
Toy Story	70
Up	175
Mary Poppins	6
The Lion King	45
Moana	150

movie	year	budget_$_million
Toy Story	1995	70
Up	2009	175
Mary Poppins	1964	MISSING
The Lion King	1994	45
Moana	2016	150

Show me… data types

These datasets below contain customer information and **phone_number** is the key column. However in the top dataset the key column data is stored as a string and in the bottom it is stored as an integer.

The datasets would not be able to be combined as the **data items are in different data types**.

customer_name	phone_number
Frank Terry	07700 900531
Paula Oliver	07700 900597
Reagan Hudson	07700 900391
Ashley Hawthorne	07700 900981

Stored as a string

phone_number	cost_month
7700900531	21.99
7700900597	31.99

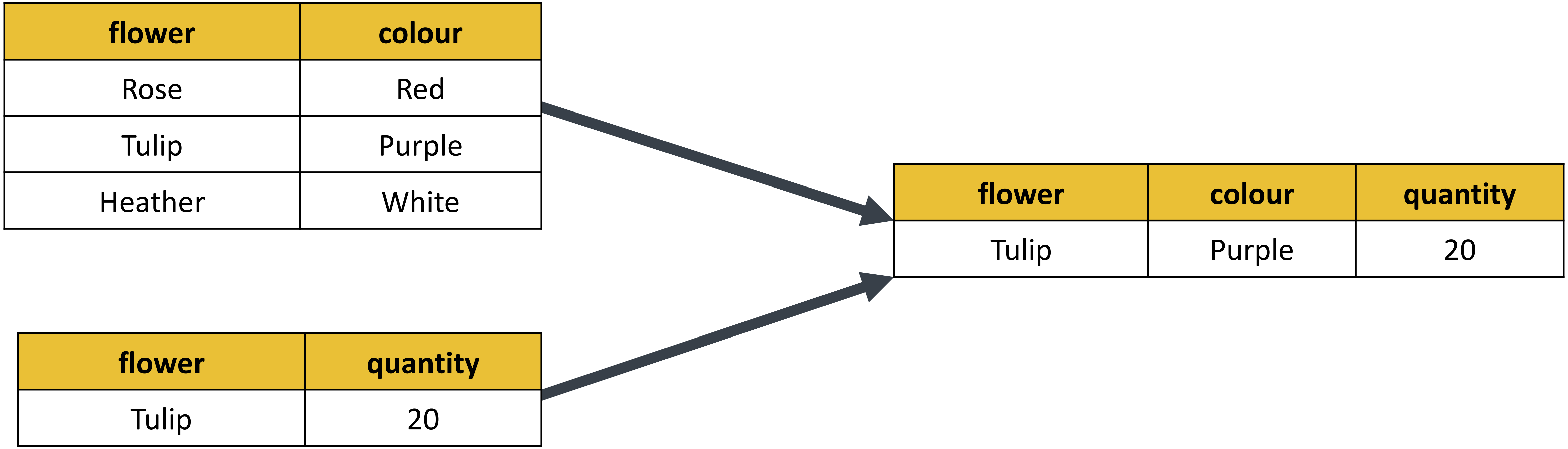
Stored as an integer

customer_name	phone_number	cost_month
Frank Terry	07700 900531	MISSING
Paula Oliver	07700 900597	MISSING
Reagan Hudson	07700 900391	MISSING
Ashley Hawthorne	07700 900981	MISSING

Show me… wrong join type

These datasets have been joined, and we were expecting 3 rows, however the final dataset only has 1.

They were meant to use a LEFT join, but used an INNER join instead.



flower	colour
Rose	Red
Tulip	Purple
Heather	White

flower	quantity
Tulip	20

flower	colour	quantity
Tulip	Purple	20

Joining datasets checklist

Before joining data it is useful to work through this checklist,

- ☑ Have you identified the **key** column(s)?
- ☑ Are they data items in the key column(s) in the **same data type**?
- ☑ Have you checked for **duplicate rows**?
- ☑ Do you know how **many rows** you expect in the final dataset?
- ☑ Do you know how **many columns** you expect in the final dataset?
- ☑ Do you expect **gaps/missing data items** in your final dataset?

Next steps

Complete **questions 1 to 4**
in **section 4** of the
‘Combining datasets’ workbook.

Learning checklist

I can *describe* how to append rows to a dataset

I can *describe* how to join columns to a dataset

I can *explain* the difference between the types of joins (left, right, inner, outer)

I can *explain* what are the common causes of issues when combining datasets

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 The Data Lab.

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