Caring for your data

Version: 1.0

effini

Data Education in Schools

Skills Development Scotland

Learning intentions

We will be looking at **caring for data**, specifically

- Why you should care for your data

- What are the different data types that need cared for

- How to create a data dictionary

Background

Data is an asset that needs to be cared for in the same way as any other valuable item in an organisation.

Data that is not cared for can become less valuable, or even cause damage.

In this lesson we will look at **why we should care for data** and **how to create a data dictionary to accurately document your data** as part of caring for your data.

Why it is important to care for data?

To stop data from being **accidentally changed or deleted**

To keep it **accurate and of a high quality**

To make sure the data is **fair and unbiased**

Show me….

GP surgeries need to care for the data they hold so they can accurately,

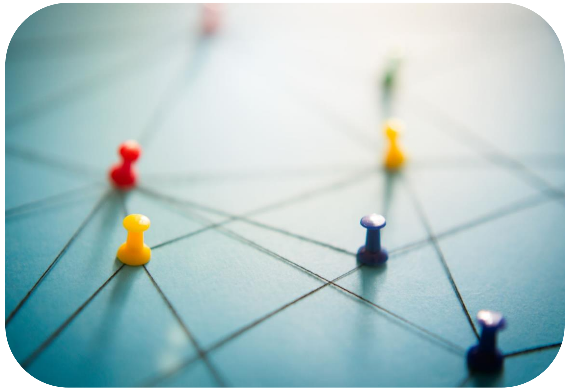
- Review patients medical history
- Prescribe the correct medication
- Invite patients to vaccination appointments

Different types of datasets

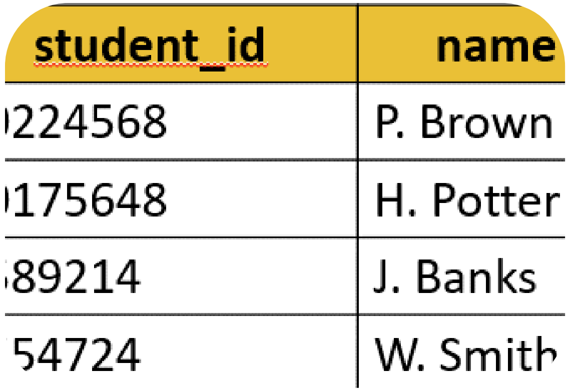
All data needs to be cared for, but different dataset types need different levels of care.

There are 4 different types of datasets,

Metadata



Reference data



student_id	name
224568	P. Brown
175648	H. Potter
89214	J. Banks
54724	W. Smith

Main data

Transactional data

Definition

Metadata

Data about the data

Show me…

The **metadata** for a photograph would be,

- Date and time of when the photo was taken
- Details of the camera settings
- Geotagging (where the photo was taken)

Definition

Reference data

Data used by other data sources such as a lookup table or a list.

Show me…

These are examples of reference datasets.

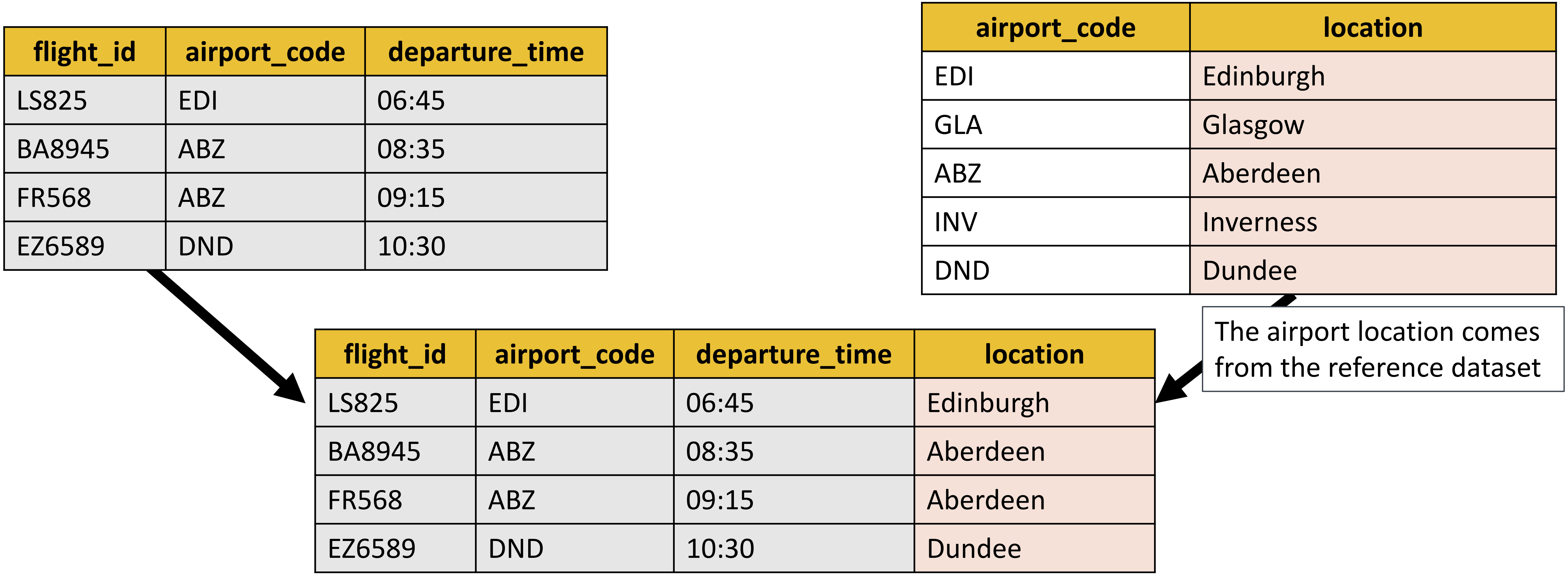
Reference data is often **tightly controlled** to reduce data quality issues and rarely changes.

course_number	unit_title
J2HN	Data Citizenship
J2G2	Data Science
HY2C	Computer Programming
H9E2	Data Security

airport_code	location
EDI	Edinburgh
GLA	Glasgow
ABZ	Aberdeen
INV	Inverness
DND	Dundee
BRR	Barra
PIK	Glasgow Prestwick
SYY	Stornoway
PSL	Perth

Example

Reference datasets are used to link information to other datasets.



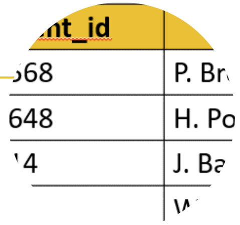
flight_id	airport_code	departure_time
LS825	EDI	06:45
BA8945	ABZ	08:35
FR568	ABZ	09:15
EZ6589	DND	10:30

airport_code	location
EDI	Edinburgh
GLA	Glasgow
ABZ	Aberdeen
INV	Inverness
DND	Dundee

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

The airport location comes from the reference dataset

Definition



nt_id	
68	P. Br
648	H. Po
4	J. Ba

Main data

Dataset that contains the core information that is important to an organisation

Show me…

The European Space Agency would store information about their rockets in the main datasets such as,

- Name of rocket
- Fuel source
- Size
- Year built
- Planned destination (e.g. moon)

Show me…

Main data contains core information that used by other datasets.

As it is used by other datasets it needs to be accurate and up to date.

student_id	first_name	last_name	house
2022456	Harry	Potter	G
2017564	Hermione	Granger	G
6589214	Draco	Malfoy	S
5654724	Luna	Lovegood	R
4547121	Ron	Weasley	G
7787454	Cedric	Diggory	H

Main data vs. reference data

The biggest difference between main and reference data is that reference data is tightly controlled and is rarely changes. Main data is added to and kept up to date.

Main dataset

student_id	first_name	last_name	house
2022456	Harry	Potter	G
2017564	Hermione	Granger	G
6589214	Draco	Malfoy	S
5654724	Luna	Lovegood	R
4547121	Ron	Weasley	G
7787454	Cedric	Diggory	H

Reference dataset

house	name	animal
G	Gryffindor	Lion
H	Hufflepuff	Badger
S	Slytherin	Snake
R	Ravenclaw	Eagle

Your turn....

This dataset contains a list of countries and their capitals. Is this dataset a **reference** or **main** dataset?

country	capital
Austria	Vienna
Brazil	Brasilia
Canada	Ottawa
Denmark	Copenhagen
Egypt	Cairo
Finland	Helsinki
Greece	Athens

Your turn....

This is a **reference dataset**. It would be used by other datasets, and it would be tightly controlled and would not change regularly.

country	capital
Austria	Vienna
Brazil	Brasilia
Canada	Ottawa
Denmark	Copenhagen
Egypt	Cairo
Finland	Helsinki
Greece	Athens

Definition

Transactional data

Data that records events, normally with datetime information

Show me…

Transactional data will be updated regularly.

date	student_id	lunch
23/9/2022	41145	Steak pie
24/9/2022	41145	n/a
25/9/2022	41145	Fish and chips
23/9/2022	12451	Steak pie
24/9/2022	12451	Cheese roll
25/9/2022	12451	Salad
23/9/2022	NULL	Baked potato and cheese

Example

For a library to find out the names of the books that are currently borrowed, they would need the transactional and reference datasets.

Transactional dataset

book_id	date_borrowed	date_returned
K142	4/8/2022	5/8/2022
K142	6/8/2022	16/8/2022
K142	17/8/2022	1/9/2022
C474	5/7/2021	9/12/2021
C474	6/2/2022	14/5/2022
H587	5/2/2020	NULL
M234	6/9/2022	1/11/2022
M234	5/11/2022	6/12/2022

Reference dataset

book_id	title
K142	To Kill a Mockingbird
C474	A Tale of Two Cities
H587	The Hobbit
W909	Charlotte’s Web
M234	Matilda

Your turn…

Can you decide whether these datasets are main, transactional or reference datasets?

Dataset 1?

student_id	first_name	last_name	house
2022456	Harry	Potter	G
2017564	Hermione	Granger	G
6589214	Draco	Malfoy	S

Dataset 2?

Class_id	description
1T	1st year - Transfiguration
2T	2nd year – Transfiguration
1DDA	1st year – Defence Against the Dark Arts
4P	4th year - Potions

Dataset 3?

Class_id	Date	Student_id	Attended
1T	5/12/1991	2022456	Y
1T	5/12/1991	2017564	Y
1T	5/12/1991	6589214	Y
1T	5/12/1991	1245758	N
1T	6/12/1991	2022456	Y
1T	6/12/1991	2017564	Y
1T	6/12/1991	6589214	Y
1T	6/12/1991	1245758	Y

Your turn…

Can you decide whether these datasets are main, transactional or reference datasets?

Dataset 1 = Main

student_id	first_name	last_name	house
2022456	Harry	Potter	G
2017564	Hermione	Granger	G
6589214	Draco	Malfoy	S

Dataset 2 = Reference

Class_id	description
1T	1st year - Transfiguration
2T	2nd year – Transfiguration
1DDA	1st year – Defence Against the Dark Arts
4P	4th year - Potions

Dataset 3 = Transactional

Class_id	Date	Student_id	Attended
1T	5/12/1991	2022456	Y
1T	5/12/1991	2017564	Y
1T	5/12/1991	6589214	Y
1T	5/12/1991	1245758	N
1T	6/12/1991	2022456	Y
1T	6/12/1991	2017564	Y
1T	6/12/1991	6589214	Y
1T	6/12/1991	1245758	Y

Your turn…

Below is a dataset that contains the latitude and longitude of UK postcodes.
Is this dataset,

a) **Metadata** - *data about the data*
b) **Reference data** – *tightly controlled data used by other datasets*
c) **Main data** - *links datasets together*
d) **Transactional data** - *data that records events*?

postcode	latitude	Longitude
EH1 2NG	55.948612	-3.200833
FK9 4TW	56.1229	-3.9456
KW1 4YT	58.6373	-3.0689
G3 8YW	55.8603	-4.2874

Your turn…

The dataset is contains **reference data**.

The data would be used by other datasets and will be tightly controlled and rarely changed.

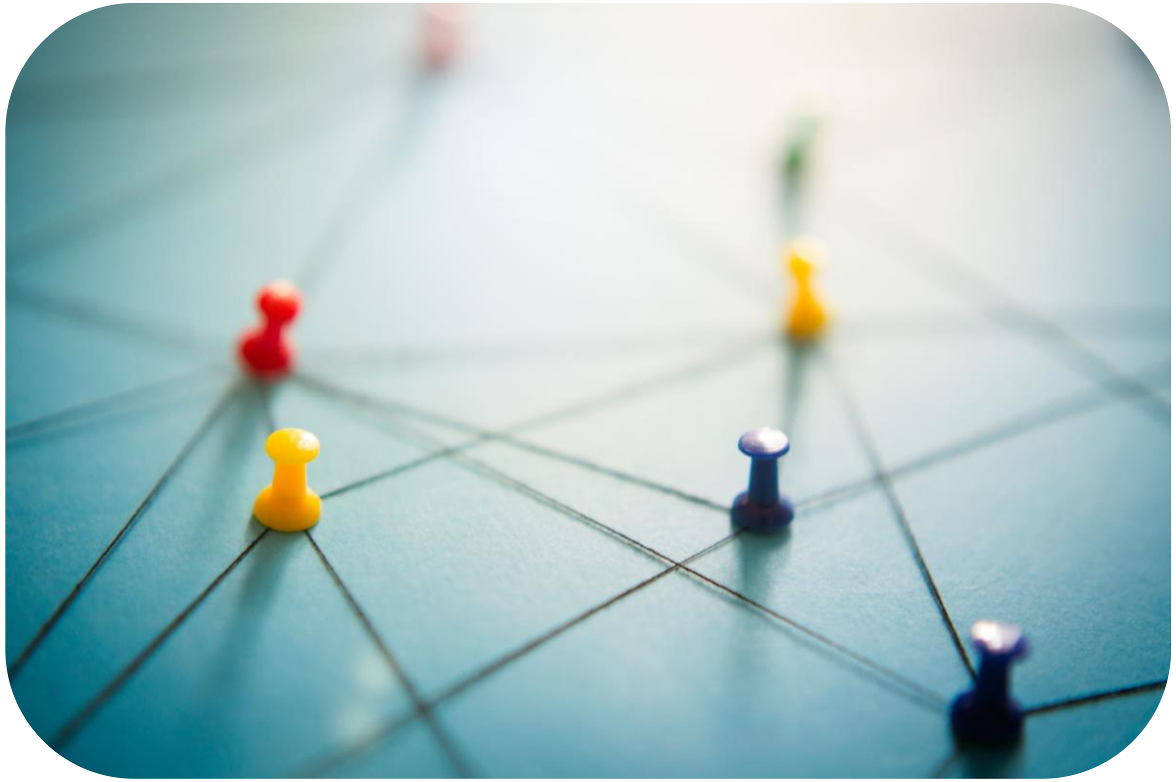
postcode	latitude	Longitude
EH1 2NG	55.948612	-3.200833
FK9 4TW	56.1229	-3.9456
KW1 4YT	58.6373	-3.0689
G3 8YW	55.8603	-4.2874

Caring for reference data

All data needs to be cared for but **reference data needs the most care**.

Reference data is important as it used by other datasets.

It needs to be of high quality and tightly controlled so that it is not accidently changed.



Show me…

Imagine this **reference dataset** was used by an online shop, but it hasn't been cared for. It now has 2 products with the same ID, but different descriptions.

Product_id	Description
A123	White chocolate bar, 500g
B452	Dark mint biscuit, 100g
B452	Strawberry and cream bar, 1kg
Z101	Peach yogurt, 6 pack

This means customers trying to order a product might end up with the wrong item.

Your turn…

Imagine this **reference dataset** is being used to tell patients where to go for their hospital appointments, however it has not been cared for and now contains errors.

What could be the **consequences to the patients of this uncared for dataset**?

hospital_name	town	Postcode
Queen Margaret	Dunfermline	KY12 0SU
Borders General Hospital	Melrose	TD6 9BS
Western General	Edinburgh	XXXXXXXXX
Ninewells	Dundee	DD2 1UB
Perth Royal Infirmary	Perth	PH1 !NX
Perth Royal Infirmary	Perth	PH1 1NX

Your turn…

What could be the consequences to the patients of the dataset not being cared for?

- Patients could **miss their appointment** as they don't know where to do go.
- **Patients might get stressed** due to the incorrect information.

hospital_name	town	Postcode
Queen Margaret	Dunfermline	KY12 0SU
Borders General Hospital	Melrose	TD6 9BS
Western General	Edinburgh	XXXXXXXXX
Ninewells	Dundee	DD2 1UB
Perth Royal Infirmary	Perth	PH1 !NX
Perth Royal Infirmary	Perth	PH1 1NX

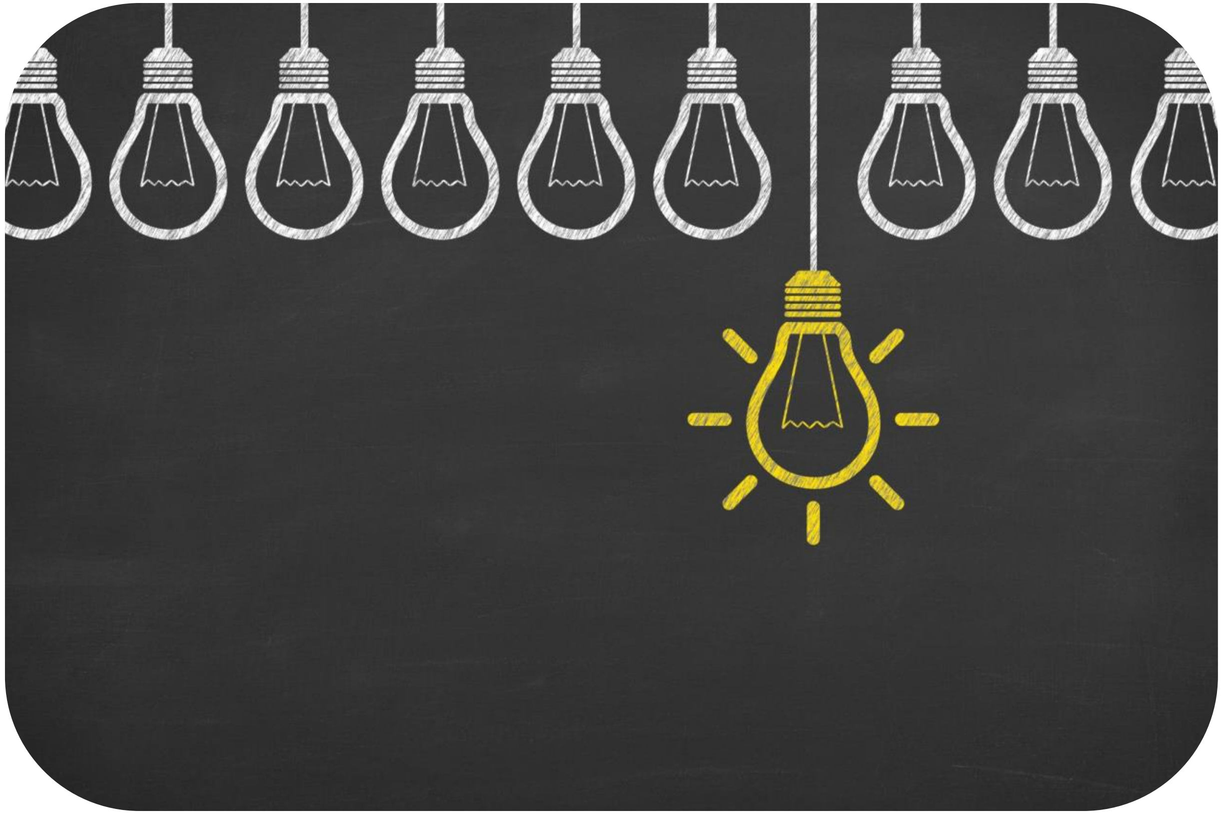
Next steps

Complete **questions 1 to 9** in **section 1** of the ‘Caring for data’ workbook.

Metadata and data dictionary

Without metadata it would be very difficult to work with any dataset.

A **data dictionary** is one of the most important pieces of metadata.



Definition

Data dictionary

the names, definitions and attributes of the elements in a dataset

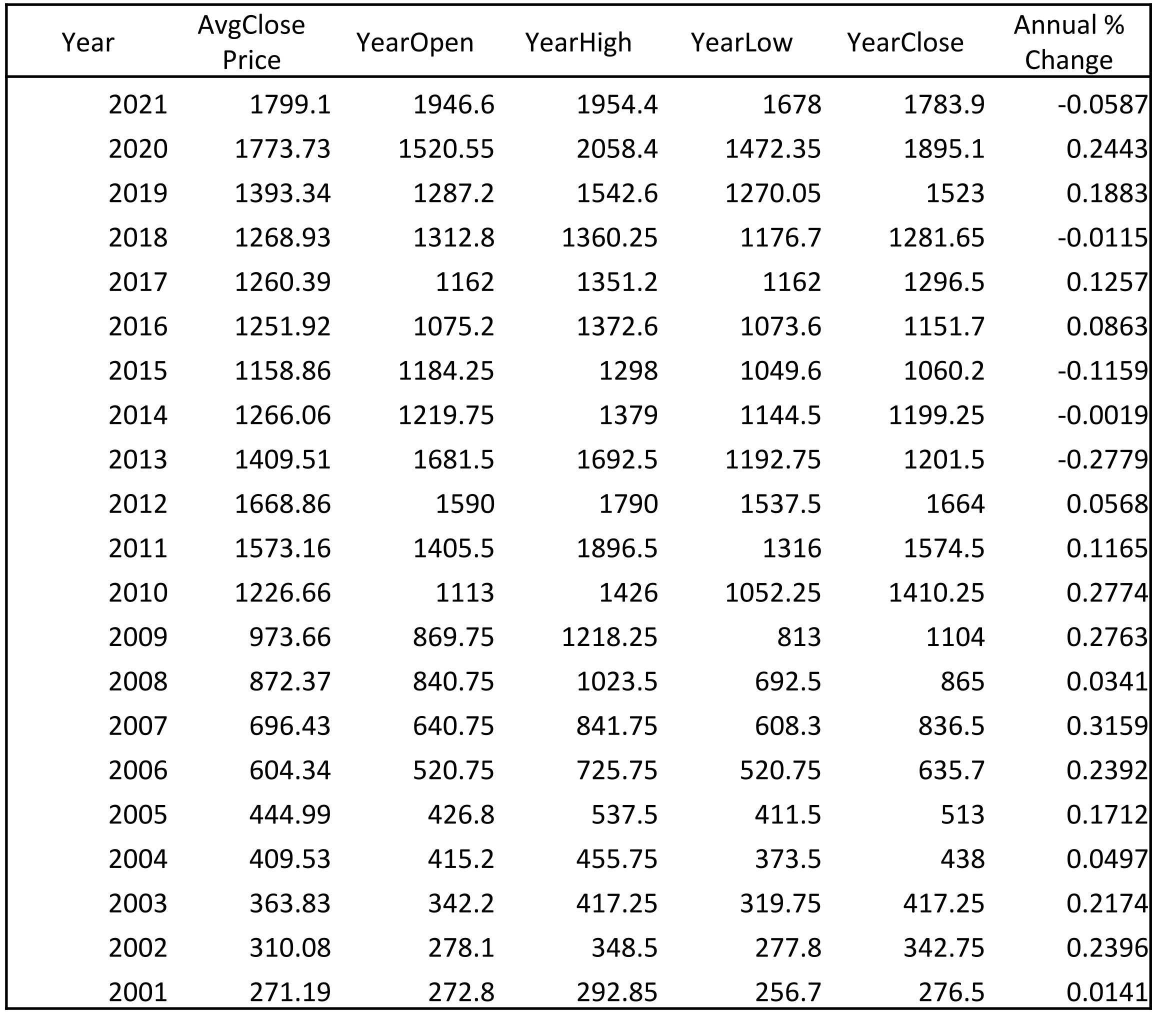
Show me…

This dataset contains the historic value of gold from Kaggle (www.kaggle.com)

The associated **data dictionary** describes what is held in each of the columns.

This contains data files of gold historical data (USD).

Year:	Year of observation
AvgClosePrice:	The average close price in the year
YearOpen:	Opening price in the year
YearHigh:	Highest price in the year
YearLow:	Lowest price in the year
YearClose:	Closing price in the year
Annual%Change:	Percent change of the previous and current year price



Year	AvgClose Price	YearOpen	YearHigh	YearLow	YearClose	Annual % Change
2021	1799.1	1946.6	1954.4	1678	1783.9	-0.0587
2020	1773.73	1520.55	2058.4	1472.35	1895.1	0.2443
2019	1393.34	1287.2	1542.6	1270.05	1523	0.1883
2018	1268.93	1312.8	1360.25	1176.7	1281.65	-0.0115
2017	1260.39	1162	1351.2	1162	1296.5	0.1257
2016	1251.92	1075.2	1372.6	1073.6	1151.7	0.0863
2015	1158.86	1184.25	1298	1049.6	1060.2	-0.1159
2014	1266.06	1219.75	1379	1144.5	1199.25	-0.0019
2013	1409.51	1681.5	1692.5	1192.75	1201.5	-0.2779
2012	1668.86	1590	1790	1537.5	1664	0.0568
2011	1573.16	1405.5	1896.5	1316	1574.5	0.1165
2010	1226.66	1113	1426	1052.25	1410.25	0.2774
2009	973.66	869.75	1218.25	813	1104	0.2763
2008	872.37	840.75	1023.5	692.5	865	0.0341
2007	696.43	640.75	841.75	608.3	836.5	0.3159
2006	604.34	520.75	725.75	520.75	635.7	0.2392
2005	444.99	426.8	537.5	411.5	513	0.1712
2004	409.53	415.2	455.75	373.5	438	0.0497
2003	363.83	342.2	417.25	319.75	417.25	0.2174
2002	310.08	278.1	348.5	277.8	342.75	0.2396
2001	271.19	272.8	292.85	256.7	276.5	0.0141

Why are data dictionaries important?

Saves time figuring out what the data means

Provides details of **how variables have been created** (e.g. through calculation)

Helps ensures everyone is **using the datasets consistently** with the same definitions

Your turn….

This dataset contains information about trees. However **you don’t have a data dictionary** for this dataset.

What **problems might you have when trying to use this dataset** without a data dictionary?

ObjectID	LocationOrTagNo	AgeGroup	Name1	Name2	DBH
17232	y5756	Semi-mature	Prunus spp.	Cherry spp	90+
17766	y4134	Middle Aged	Sorbus aucuparia	Rowan	10-20
17449	y4130	Mature	Prunus x hillieri 'Spire'	Hillier's Cherry	50-60
17238	y4124	Mature	Prunus laurocerasus	Common Laurel	90+
12350	y4123	Semi-mature	Cupressocyparis leylandii	Leyland Cypress	10-20

Your turn....

What is the difference between the two variables that contain names? Which one should you use?

What does the variable DBH mean?

ObjectID	LocationOrTagNo	AgeGroup	Name1	Name2	DBH
17232	y5756	Semi-mature	Prunus spp.	Cherry spp	90+
17766	y4134	Middle Aged	Sorbus aucuparia	Rowan	10-20
17449	y4130	Mature	Prunus x hillieri 'Spire'	Hillier's Cherry	50-60
17238	y4124	Mature	Prunus laurocerasus	Common Laurel	90+
12350	y4123	Semi-mature	Cupressocyparis leylandii	Leyland Cypress	10-20

How to create a data dictionary

If you are creating a new dataset you also need to create a data dictionary at the same time. It should include,

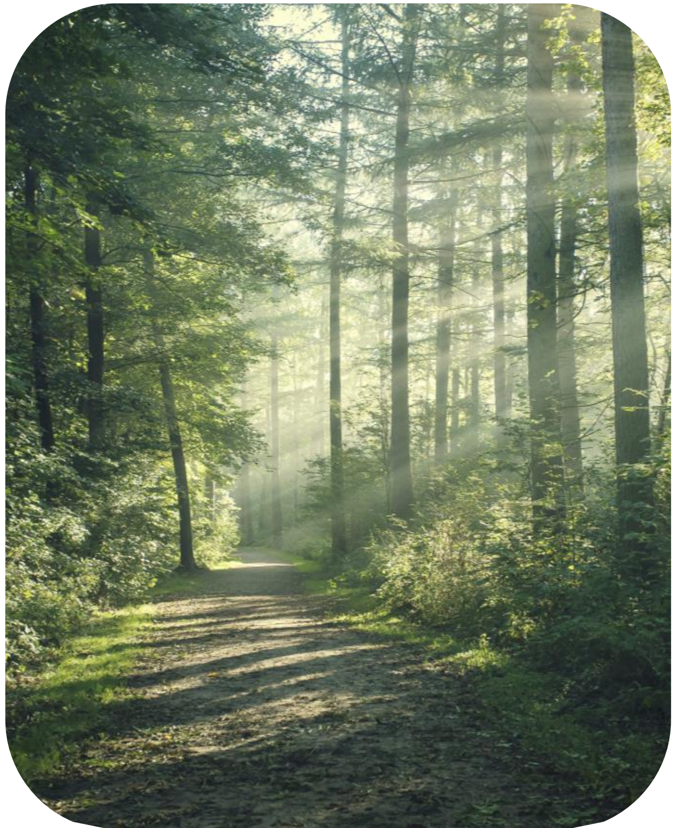
- **Name** of each variable
- **Description** of the variable
- The **variable type** e.g. integer, text, datetime

Example

Here is the data dictionary for the tree dataset we have just looked at.

We are now able to see that difference between Name1 and Name2 and what DBH means.

variable_name	description	type
ObjectID	Primary key for the tree	Text
LocationOrTagNo	Location of the tree	Text
AgeGroup	Age of tree e.g. mature, middle-aged	Text
Name1	Latin name of the tree	Text
Name2	Common name of the tree	Text
DBH	Diameter at Breast Height (4.5 feet above the ground)	Text



Next steps

Complete **questions 1 to 8**
in **section 2** of the
‘Caring for data’ workbook.

Learning checklist

I can *explain* that data needs to be cared for.

I can *describe* the different type of datasets.

I can *create* a data dictionary for a given dataset.

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 and Skills Development Scotland.

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