Caring for data
(Answers)

Worksheet section	Contents
1	Caring for datasets
2	Create a data dictionary

Version: 1.0

This lesson has been created by effini in partnership with Data Education in Schools and Skills Development Scotland.

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

1. Caring for datasets

1) Fill in the missing gaps in the different types of datasets

Metadata   Data about the   data

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

Main data   Dataset that contains the   core   information that is important to an organisation.

Transactional data Data that records   events   normally with datetime information

Section 1.2 (rephrase)

2) Why is important to care for data?

> 1. To stop data from being accidentally changed or deleted.
> 2. To keep it accurate and of a high quality
> 3. To make sure the data is fair and unbiased.

3) Why does reference data need to be cared for the most?

> Reference data needs to be cared for the most as it is used by other datasets. It needs to be of high quality and tightly controlled.

Section 1.3 (apply)

For each of these datasets state the type of dataset and describe why.

4)

planet	type	size_km
Mercury	Terrestrial	2,440
Venus	Terrestrial	6,052
Earth	Terrestrial	6,371
Mars	Terrestrial	3,390
Jupiter	Gas giant	69,911
Saturn	Gas giant	58,232
Uranus	Ice giant	25,362
Neptune	Ice giant	24,622

Dataset type   Reference

Reason for choice   Information that would be used by other data sources that would rarely change.

5)

patient_id	name	DOB
AQ1245	F. Copper	25/08/1998
QW1254	M. Reid	19/12/1998
ER7854	L. Power	12/02/2001
FT6954	T. Young	22/10/2001
GT3564	D. Whyte	23/03/2004
PL1289	M. Trent	22/09/2005

1. Caring for datasets

Dataset type	Main

Reason for choice	It is core information that an organisation needs

6)

date	number_visitors
05/04/2023	5124
06/04/2023	6475
07/04/2023	4578
08/04/2023	210
09/04/2023	1234
10/04/2023	2546

Dataset type	Transactional

Reason for choice	Dataset contains record of events

7)

Dataset name:	Public Toilets and Comfort Schemeds
Description:	The location of Public toilets and comfort schemes within the Perth and Kinross area.
Maintainer:	Perth & Kinross Council
Created:	18-Aug-22
Last updated:	12-Sep-22
Language:	English
Source of data:	https://data.pkc.gov.uk/dataset/public-toilets-and-comfort-schemes

Dataset type	Metadata

Reason for choice	It's the data about the dataset

Section 1.4 (active)

8) Public Health Scotland shares open data for the public to use.
Using their website (link below) find an example of reference dataset. Fill in the name and description of the dataset below.
https://www.opendata.nhs.scot/dataset

Dataset name:	Example dataset: Accident & Emergency Sites

Description:	Sites are classed as either Emergency Departments (larger A&E services that typically provide a 24-hour consultant led service) or other sites including minor injuries units, small hospitals and health centres in rural areas that carry out Emergency Department related activity and are GP or nurse led. They may or may not be open 24 hours. All sites have been open from before the A&E data mart started collecting data in June 2007 unless stated otherwise.

9) What would happen if this reference data was not cared for?

Example answer: if someone was using this to find out which hospital to visit in an emergency and it was incorrectly classified they might end up going to the wrong place - not what you'd want in an emergency.

2. Create a data dictionary

1) Why are data dictionaries important?

> 1. Saves time figuring out what the data means
> 2. Provides details of how variables have been created
> 3. Helps ensures everyone is using the datasets consistently with the same definitions.

2) What are the 3 pieces of information about a variable should you include in a data dictionary?

> 1. Name of the variable
> 2. Description of the variable
> 3. The variable type

Section 2.2 (apply)

For each of these datasets, fill in the gaps for the associated data dictionary.

3)

planet	type	size_km
Mercury	Terrestrial	2,440
Venus	Terrestrial	6,052
Earth	Terrestrial	6,371
Mars	Terrestrial	3,390
Jupiter	Gas giant	69,911
Saturn	Gas giant	58,232
Uranus	Ice giant	25,362
Neptune	Ice giant	24,622

Data dictionary

variable_name	description	type
planet	Name of the plant in our solar system	string
type	The type of the planet. Either terrestrial, gas giant or ice giant.	string
size_km	The diameter of the planet in km	integer

4)

ice_cream	price_per_scoop
Chocolate	1.20
Strawberry	1.20
Vanilla	1.00
Mint Choc	1.25
Raspberry	1.20
Irn Bru flavour	1.30

Data dictionary

variable_name	description	type
ice_cream	Flavour of ice cream available in a shop.	string
price_per_scoop	Price in £s of each scoop of ice cream	floating point

2. Create a data dictionary

5)

date	number_visitors
05/04/2023	5124
06/04/2023	6475
07/04/2023	4578
08/04/2023	210
09/04/2023	1234
10/04/2023	2546

Data dictionary

variable_name	description	type
date	Date (DD/MM/YYYY)	datetime
number_visitors	Number of visitors to the park per day	integer

6)

statues	height	location
Scott Monument	61	Edinburgh
Statue of Liberty	93	New York
Little Mermaid	1	Copenhagen
Christ the Redeemer	38	Rio de Janeiro
Nelson's column	52	London
The Kelpies	30	Falkirk

Data dictionary

variable_name	description	type
statues	Names of famous statues	string
height	Maximum height in m	integer
location	City/town of the statues	string

7)

time	temperature
09:00	7
10:00	8
11:00	8
12:00	9
13:00	6
14:00	6
15:00	5
16:00	5
17:00	5
18:00	5

Data dictionary

variable_name	description	type
time	Time in hours of the measurements	time
temperature	Air temperature in degrees celcius	integer

2. Create a data dictionary

8) In section 1 question 7, you found a dataset from the Public Health Scotland website. Fill in the data dictionary for your chosen dataset.

> Note for teacher
> This is an example answer, the learners may choose another dataset.

Dataset name:

Example dataset: Accident & Emergency Sites

Description:

Sites are classed as either Emergency Departments (larger A&E services that typically provide a 24-hour consultant led service) or other sites including minor injuries units, small hospitals and health centres in rural areas that carry out Emergency Department related activity and are GP or nurse led. They may or may not be open 24 hours. All sites have been open from before the A&E data mart started collecting data in June 2007 unless stated otherwise.

Data dictionary

variable_name	description	type
_id	unique identify	integer
HB	Health board	string
TreatmentLocationName	Name of the hospital	string
TreatmentLocationCode	Location of hospital	string
CurrentDepartmentType	Type of department	string
FileType	File type	string
Comments	Comments	string
Status	Whether the department is open or closed	string