Importance of data quality (Answers)

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. Importance of high quality data

Section 1.1 (recall)

1) Fill in the gaps in the definition of **high quality data**.

Data that is good enough to be used for the task it is intended for.

2) Why is it important to have high quality data?

1. You can trust your results of your analysis
2. Can help you make accurate decisions
3. You won't need to clean the dataset
4. Be able to use data that is fair and accurate

Section 1.2 (apply)

Can you explain why you think you could (or could not) confidently use these datasets to complete the tasks described?

3) Task: Calculate the time taken to run 5km in these different locations.

Dataset:

race	start_time	end_time
Leven	10:00	10:35
Stirling	09:30	10:02
Dumfries	10:15	10:38
Glasgow	10:30	11:01

Confidently use the dataset to complete the task?

Yes, the dataset is good enough calculate the task.

4) Task: Find out the area of the largest ocean on Earth.

Dataset:

ocean	area_km_2
Pacific	TBC
Atlantic	85,133,000
Indian	70,560,000
Southern	21,960,000
Arctic	15,558,000

Confidently use the dataset to complete the task?

No, the area for the Pacific ocean is missing so you are not able to find out which ocean is the largest.

5) Task: Find out the average number of gold medals won at this competition

Dataset:

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

Confidently use the dataset to complete the task?

No, there are duplicate rows that would make you doubt your results.

5) Task: How old is Andy Murray?

Dataset:

FirstName	LastName	Age
A	Murray	35
Andrew	Murray	36

Confidently use the dataset to complete the task?

No, the dataset does not allow you to make accurate decisions to complete the task.

Section 1.3 (active)

6) The building of the Sick Kids hospital is Edinburgh was delayed by a data quality issue in a spreadsheet. Read over this article from the BBC website and then answer the questions below.

Spreadsheet error led to Edinburgh hospital opening delay by Andrew Picken.
https://www.bbc.co.uk/news/uk-scotland-edinburgh-east-fife-53893101

How many air changes per hour should the critical care rooms had?

10 times per hour

What was the name of the spreadsheet that contained the error?

The environmental matrix

What did the investigation believe caused the error in the spreadsheet?

Human error when copying the requirements from the generic ventilation to the critical care room detail.

Was the environmental matrix spreadsheet "good enough" to complete the intended task?

No, it didn't allow them to make accurate decisions.

2. Identifying high quality data

Section 2.1 (recall)

1) Fill in the missing words in the 6 dimensions of quality data

1. Completeness How non-blank or populated the data is

2. Timeliness How up to date the data is

3. Uniqueness Data is not recorded more than once

4. Validity That data is in the correct format, type and range

5. Accuracy How data represents the real-world

6. Consistency Data matches if two copies of the same information are compared

Section 2.2 (apply)

Review these datasets against the 6 dimensions of quality data.

Fill in the grey boxes with Yes or No.

2)

planet	type	size_km
Mercury	Terrestrial	05/09/1906
Venus	Terrestrial	6,052
Earth	Terrestrial	6,371
Mars	Terrestrial	3,390
Jupiter	Gas giant	6.99E+04
Saturn	Gas giant	5.82E+04
Uranus	Ice giant	2.54E+04
Neptune	Ice giant	2.46E+04

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
Yes	Unknown	Yes	No	Unknown	Unknown

3)

animal	speed_km/h
Cheetah	121
Golden eagle	319
Golden eagle	320
Lion	81
Lion	81
Peregrine falcon	389
Rock dove	149
Swordfish	97

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
Yes	Unknown	No	Yes	Unknown	No

2. Identifying high quality data

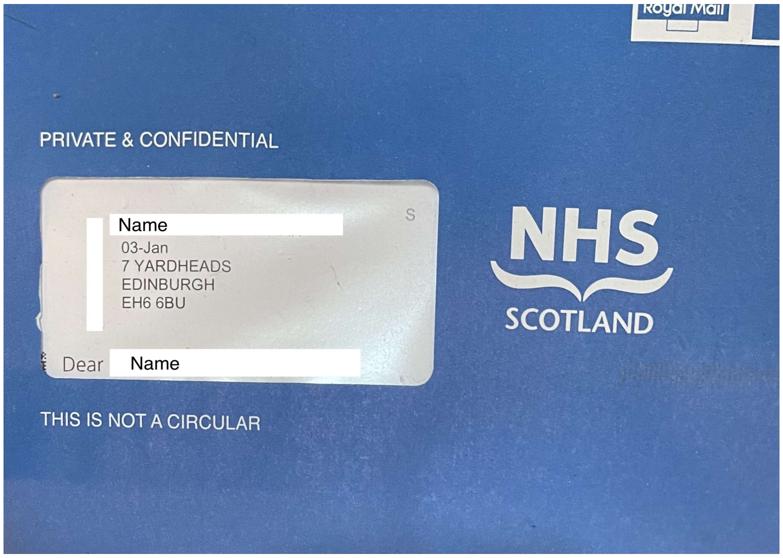
4)

mountain	range	height_m
Everest	Himalaya	#REF!
K2	Baltoro	8,611
Nanaga Parbat	Himalaya	8,091
Broad Peak	Baltoro	8,051
Changtse	Himalaya	7,543

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
No	Unknown	Yes	Yes	Unknown	Unknown

Section 2.3 (rephase)

5) Below is a letter where the address has a data quality issue.



PRIVATE & CONFIDENTIAL

Name
03-Jan
7 YARDHEADS
EDINBURGH
EH6 6BU

Dear Name

THIS IS NOT A CIRCULAR

Which part of the address has a data quality issue?

The first line address has a data quality issue (03-Jan).

Which of the 6 dimensions of quality data has caused the issue?

Validity - in the wrong format.

3. Improving the quality of data

Section 3.1 (apply)

For each of these datasets, describe how you could improve the quality of data.

1)

planet	type	size_km
Mercury	Terrestrial	05/09/1906
Venus	Terrestrial	6,052
Earth	Terrestrial	6,371
Mars	Terrestrial	3,390
Jupiter	Gas giant	6.99E+04
Saturn	Gas giant	5.82E+04
Uranus	Ice giant	2.54E+04
Neptune	Ice giant	2.46E+04

How could you improve the **validity** of this dataset?

Change the format of the size_km variable so they are correct.

2)

animal	speed_km/h
Cheetah	121
Golden eagle	319
Golden eagle	320
Lion	81
Lion	81
Peregrine falcon	389
Rock dove	149
Swordfish	97

How could you improve the **uniqueness** of this dataset?

Delete the duplicate rows.

How could you improve the **consistency** of this dataset?

Check the speed of the Golden eagle in another source.

3)

mountain	range	height_m
Everest	Himalaya	#REF!
K2	Baltoro	8,611
Nanaga Parbat	Himalaya	8,091
Broad Peak	Baltoro	8,051
Changtse	Himalaya	7,543

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
No	Unknown	Yes	Yes	Unknown	Unknown

How could you improve the **completeness** of this dataset?

Add in the height of Everest or remove the row.