Importance of data quality

Worksheet section	Contents
1	Importance of data quality
2	Identifying high quality data
3	Improving the quality of a dataset

Version: 1.0

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

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

Data that is [   ] to be used for the task it is intended for.

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

[   ]

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?

[   ]

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?

[   ]

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

1. Importance of high quality data

Confidently use the dataset to complete the task?

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?

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?

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

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

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

2. Identifying high quality data

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

1. Completeness How _____ the data is

2. Timeliness How _____ the data is

3. Uniqueness Data is not recorded more than _____

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

5. Accuracy How data represents the _____

6. Consistency Data matches if _____ 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
	Unknown			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
	Unknown			Unknown	

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
	Unknown			Unknown	Unknown

Section 2.3 (rephase)

5) Below is a letter where the address has a data quality issue.
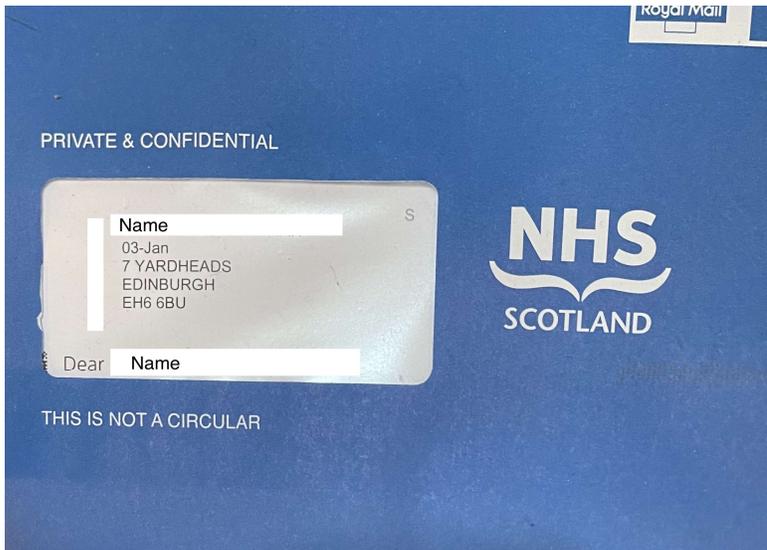
Which part of the address has a data quality issue?

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

3. Improving the quality of data

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?

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?

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

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
	Unknown			Unknown	Unknown

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