

Importance of data quality

Version: 1.0



Learning intentions

We will be looking at the **quality of data**, specifically

- Why it is important to have high quality data
- How to assess the quality of a dataset
- How to improve the quality of a dataset

Background

A common problem data scientists come across is that the dataset is not of a high enough quality for them to complete the task they are working on.

In this lesson, we will look at what is **meant by high quality data** and how to identify if your dataset is high quality.



Definition



High quality data

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

High quality data vs. “perfect” data

Data does not have to be perfect for it to be of high quality.

The important thing is for it to be **fit for purpose**.

As long as the data is good enough to allow you to accurately complete the task it is intended for, it is high quality data.



Why it's important to have high quality data?



You can **trust your results** of your analysis



Can help you make **accurate decisions**



You won't need to **clean the dataset**



Be able to use data that is **fair** and **accurate**

Example

Man offered Covid vaccine after error lists him as 6.2cm tall.

“A man in his 30s with no underlying health conditions was offered a Covid vaccine after an NHS error mistakenly listed him as just 6.2cm in height.

Liam Thorp was told he qualified for the jab because his measurements gave him a body mass index of 28,000.”



Show me...



Task: Identify people with high BMI (= weight/height²) and invite them for a Covid vaccine.

Dataset:

name	weight_kg	height_cm	bmi
Lee	80	160.0	31.25
Kim	74	175.0	24.16
Rorv	95	181.0	29.00
Liam	110	6.2	28,616.02

Can we confidently use the dataset to complete the task?

No, the data is not good enough to undertake analysis and trust the results.

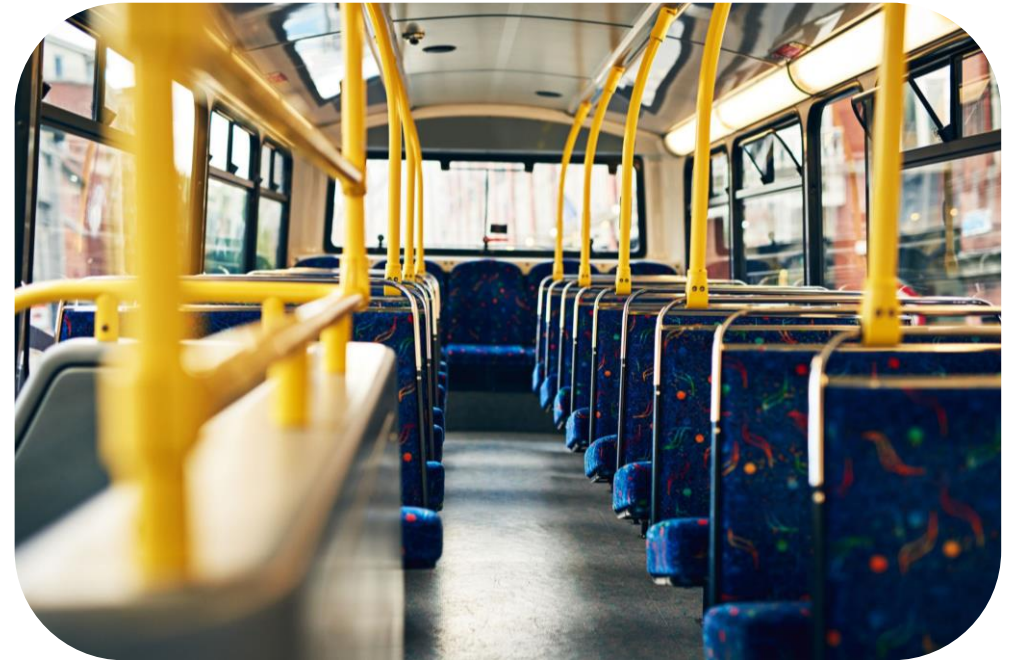
Show me...



Task: Which day of the week do most people use the bus?

Dataset:

day	num_passagers
Monday	1,546
Tuesday	1,785
Wednesday	NULL
Thursday	1,689
Friday	2,578
Saturday	986
Sunday	321



Can we confidently use the dataset to complete the task?

No, there is missing data on Wednesday. We need data for everyday to be able to complete the task accurately.

Show me...



Task: To find out the approximate average height of people in Scotland.

Dataset: Height of people in Scotland from 2010.

Can we confidently use the dataset to complete the task?

Yes, the dataset is not up to date as the average height is unlikely to have changed much for since then.

It is **not perfect but is good enough** for the intended task.

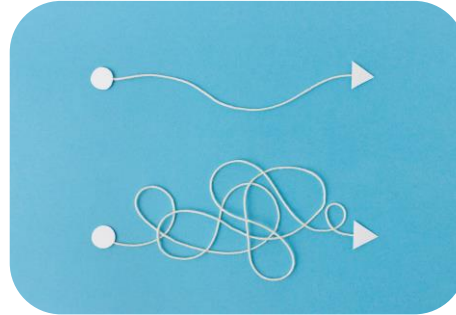


id	date_collected	height_m
345	1/4/2010	1.70
147	5/4/2010	1.76
2558	20/1/2010	1.66
1475	5/8/2010	1.89

Impacts of low-quality data



Analysis needs to be **redone**



People will come up with **work-arounds**



Companies will work in an **inefficient way**



Low job satisfaction
from excessive manual
processes



Customers will get frustrated from
incorrect information



Opportunity cost of **missed sales**



Compliance costs or
finances for incorrect reporting



Reputational costs
from **loss of trust** in the
organisation

Next steps

Complete **questions 1 to 6**
in **section 1** of the
'Importance of data quality' workbook.

How to spot high quality data

There are many ways data can be incorrect and therefore cause the dataset to be of low quality.

The different ways that data can be incorrect are organised into six different areas called the **Dimensions of Data Quality**.



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 - identifies duplicates

6 dimensions of quality data



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

Show me...



The dataset is not high quality as the average temperature column is mostly blank.

town	population	average_temp
Selkirk	4,540	NULL
St Andrews	18,410	NULL
Edinburgh	527,620	15
North Berwick	7,840	NULL
Callander	3,080	NULL
Cumbernauld	52,290	NULL

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✗	Unknown	✓	✓	Unknown	Unknown

Show me...



The dataset is not high quality as the mountain Ben More has a duplicate row.

mountain	height_m
Ben Nevis	1,344
Ben Lawers	1,214
Ben More	1,174
Ben More	1,174
Ben Cruachan	1,127
Schiehallion	1,083

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✓	Unknown	✗	✓	Unknown	Unknown

Show me...



These datasets contain the prices of items in a bakery. Bread has a different price in the two datasets, so they are not of high quality.

item	price
Cake	£2.50
Biscuit	£1.35
Bread	£1.99
Pie	£5.99
Bridie	£2.50

item	price
Sausage roll	£1.99
Biscuit	£1.35
Bread	£9.99
Iced bun	£0.50
Pie	£5.99

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✓	Unknown	✓	✓	Unknown	✗

Show me...



This dataset contains the age of celebrities. It is a high quality dataset.

name	date_of_birth	age_oct_2022
Andy Murray	15 May 1987	35
Chris Hoy	23 Mar 1976	46
Lorraine Kelly	30 Nov 1959	62
Laura Muir	09 May 1993	29
Ewan McGregor	31 Mar 1971	51
Hannah Rankin	21 Jul 1990	32

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✓	✓	✓	✓	Unknown	Unknown

Show me...



This dataset contains the date of births of celebrities, however they are not in a valid date format.

name	date_of_birth
Tom Hanks	20645
Marilyn Monroe	9649
Emma Watson	32978
Kate Winslet	27672
Robbie Coltrane	18352

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✓	Unknown	✓	✗	Unknown	Unknown

Your turn....



Is this dataset high quality?

world_ranking	mens_rugby_team
1	Ireland
2	#REF!
3	South Africa
4	New Zealand
5	#REF!
6	#REF!

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
?	?	?	?	?	?

Your turn....



Is this dataset high quality?

No, it is not complete as there are missing values in dataset.

world_ranking	mens_rugby_team
1	Ireland
2	#REF!
3	South Africa
4	New Zealand
5	#REF!
6	#REF!

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✗	Unknown	✓	✓	Unknown	Unknown

Next steps

Complete **questions 1 to 5**
in **section 2** of the
'Importance of data quality' workbook.

Improving the quality of the data

When improving the quality of the dataset you need focus on **getting it good enough** for the intended purpose and **not perfect**.



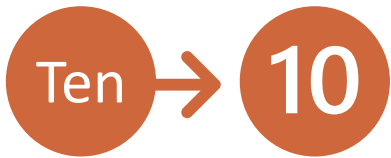
Some examples of ways to improve a dataset



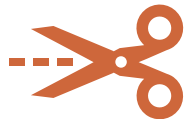
Remove duplicate rows



Check the individual data items are accurate



Change the format of the data



Remove/replace missing values

Your turn....



How could you improve the quality of this dataset?

mountain	height_m
Ben Nevis	1,344
Ben Lawers	1,214
Ben More	1,174
Ben More	1,174
Ben Cruachan	1,127
Schiehallion	1,083

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✓	Unknown	✗	✓	Unknown	Unknown

Your turn....



How could you improve the quality of this dataset? By **removing the duplicate rows**

mountain	height_m
Ben Nevis	1,344
Ben Lawers	1,214
Ben More	1,174
Ben More	1,174
Ben Cruachan	1,127
Schiehallion	1,083



mountain	height_m
Ben Nevis	1,344
Ben Lawers	1,214
Ben More	1,174
Ben Cruachan	1,127
Schiehallion	1,083

Completeness	Timeliness	Unique	Validity	Accuracy	Consistency
✓	Unknown	✓	✓	Unknown	Unknown

Next steps

Complete **questions 1 to 3**
in **section 3** of the
'Importance of data quality' workbook.

Learning checklist

I can *explain* what is meant by high quality data

I can *evaluate* the quality of a dataset

I can *describe* how to improve the quality of a 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.

© 2022. This work is licensed under a [CC BY-NC-SA 4.0 license](#).

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**



**Skills
Development
Scotland**