Reshaping datasets (Answers)

Worksheet section	Contents
1	Wide and long datasets
2	Process for reshaping

This lesson has been created by Effini in partnership with Data Education in Schools, The Data Lab and Data Skills for Work, with funding from the Scottish Government.

1. Wide and long datasets

Section 1.1

1) Fill in the gaps of these descriptions of the columns in a **long** dataset.

id	labels the	observation
key	name	of the data item
value	contains the value of the	data item

2) Select the correct definition of a **wide** dataset.

- ☐ Each different data variable is in a separate row
- ☑ Each different data variable is in a separate column

Section 1.2

3) Label these datasets as wide or long datasets

wide

Country	population_1900	population_2000
Scotland	4,437,000	5,063,000
Canada	5,301,000	31,689,000
Greece	2,504,070	11,120,000
New Zealand	802,200	4,028,000

long

building_or_statue	measurement	value
Scott Monument	height	61
The Glasgow Tower	height	127
Shanghai Tower	height	632
Empire State building	height	381
Nelson's column	height	52
The Shard	height	310

wide

ID	test_1	test_2
GF15421	45%	41%
SD14582	21%	95%
WS12452	33%	93%
AW5248	96%	62%

long

name	address_part	value
A. Horn	town	Selkirk
G. Davies	town	Crail
V. Corrs	town	Wigtown
R. Jones	town	Lanark
A. Horn	street	Ashby Close
G. Davies	street	Avon Street
V. Corrs	street	Law View
R. Jones	street	Birks Street

1. Wide and long datasets

Section 1.3

4) Explain why this dataset from Traffic Scotland is wide.

Destination	Current Journey Time	Typical Journey Time	Delay
M8 J19 Anderston	55 mins	54 mins	Less than 1 minute
M8 J1 Hermiston Gait (WB)	11 mins	12 mins	No delay
M8 J25A Braehead	59 mins	59 mins	Less than 1 minute
M8 J28 Glasgow Airport	1 hour, 2 mins	1 hour, 2 mins	Less than 1 minute
M8 J8 Baillieston	45 mins	45 mins	Less than 1 minute

https://trafficscotland.org/journeytimes/list/

Each different data variable is in a separate columns. i.e. destination, journey time, delay are all separate columns.

5) What are the **id**, **key** and **value** columns in this **long** dataset? This dataset shows Data Science and STEM Salaries from top companies.

timestamp	company	title	totalyearly...
8/17/2021 8:28:57	Adobe	Product Designer	300000
8/17/2021 8:26:21	HSBC	Software Engineer	159000
8/17/2021 8:24:56	Cisco	Software Engineer	154000
8/17/2021 8:22:17	Fidelity Investments	Software Engineer	98000
8/17/2021 8:16:36	Amazon	Product Manager	241000
8/17/2021 7:55:47	AT&T	Data Scientist	175000
8/17/2021 7:50:25	Raytheon Technologies	Software Engineer	95000

id

timestamp

key (or keys)

company and title

value

totalyearly

Source: https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries

6) Can you give a benefit of wide dataset and a benefit of long dataset?

Benefit of wide data	Benefit of long data
No repetition of data items	Makes it easier to group and summarise data
Easier to calculate time between date points	Easier at handling irregular and/or missing data items

1. Wide and long datasets

Section 1.4

7) Can you go online and find an example of a wide and long dataset? Copy and paste them below and explain why they are wide or long.

Wide

Explain why it is wide

e.g. Each different data variable is in a separate column. There is important information in the header row (e.g. round 1, round 2)

Paste your dataset here.

Note for teachers: example dataset below

Source: https://www.bbc.co.uk/sport/golf/leaderboard

Portugal Masters

Date:	Course:	Par:	Yards:	Prize Money:
4 - 7 November 2021	Dom Pedro Victoria Golf Course	71	7,191	€1,500,000

	Player	Total	Thru	Round 1	Round 2	Round 3	Round 4	Strokes
1	Thomas Pieters	-19	F	68	64	65	68	265
2	Nicolai Hoejgaard	-17	F	67	69	67	64	267
2	Lucas Bjerregaard	-17	F	67	65	69	66	267
2	Matthieu Pavon	-17	F	68	64	65	70	267
5	Matthew Jordan	-13	F	70	68	67	66	271
5	Nino Bertasio	-13	F	61	69	74	67	271
7	Francesco Laporta	-12	F	70	66	69	67	272
8	Min-Woo Lee	-11	F	68	68	71	66	273
8	Adria Arnaus	-11	F	65	67	73	68	273
8	Victor Perez	-11	F	72	68	68	65	273
8	Richard Bland	-11	F	70	65	69	69	273

Long

Explain why it is long

e.g. There are id, key(s) and value columns.

Paste your dataset here.

Note for teachers: example dataset below

Source: https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh

mass (g)	fall	year	reclat	reclong
21	Fell	1880	50.775000	6.083330
720	Fell	1951	56.183330	10.233330
107,000	Fell	1952	54.216670	-113.000000
1,914	Fell	1976	16.883330	-99.900000
780	Fell	1902	-33.166670	-64.950000
4,239	Fell	1919	32.100000	71.800000

2. Process for reshaping datasets

Section 2.1

1) Explain what it means to reshape a dataset

To convert data from wide to long or visa versa.

Section 2.2

2a) Calculate how many years these TV shows ran for in the wide and long datasets.

Long

TV_show	date_type	date
friends	first show	1994
friends	last show	2004
Game of Thrones	first show	2011
Game of Thrones	last show	2019
The Big Bang Theory	first show	2007
The Big Bang Theory	last show	2019
M*A*S*H	first show	1972
M*A*S*H	last show	1983

TV_show	years
friends	10
Game of Thrones	8
The Big Bang Theory	12
M*A*S*H	11

Wide

TV_show	first_show	last_show	years
Game of Thrones	2011	2019	8
M*A*S*H	1972	1983	11
The Big Bang Theory	2007	2019	12
friends	1994	2004	10

2b) Which did format did you find easier to calculate how long these TV shows ran for and why?

The answer will be based on the opinion of the learner. Things that could be covered are, easier to copy formulas, easier to read/understand the data, easier to add additional information to the dataset.

2. Process for reshaping datasets

3a) Summarise this dataset in the wide and long format to work out the **total_population** by year and **% urban population** by year

Long

population_type	year	world_population
urban	2020	4,378,993,944
rural	2020	3,415,804,795
urban	2000	2,868,307,513
rural	2000	3,275,186,310
urban	1980	1,754,201,029
rural	1980	2,703,802,485
urban	1960	1,023,845,517
rural	1960	2,011,104,231

year	total_population	%_urban
2020	7,794,798,739	56%
2000	6,143,493,823	47%
1980	4,458,003,514	39%
1960	3,034,949,748	34%

Wide

population_type	1960	1980	2000	2020
rural	2,011,104,231	2,703,802,485	3,275,186,310	3,415,804,795
urban	1,023,845,517	1,754,201,029	2,868,307,513	4,378,993,944
Total_population	3,034,949,748	4,458,003,514	6,143,493,823	7,794,798,739
% urban	34%	39%	47%	56%

Source:
https://www.worldometers.info/world-population/world-population-by-year/

3b) Which did format did you find easier to summarise and why?

The answer will be based on the opinion of the learner. Things that could be covered are, easier to copy formulas, easier to read/understand the data, easier to additional information to the dataset.

2. Process for reshaping datasets

Section 2.3

4) This dataset is being reshaped from wide to long. Can you fill in the gaps in the long dataset?

planet	distance_from_sun	radius_km	mass_vs_earth
Jupiter	748,000,000	69911	1321
Saturn	1,483,000,000	58232	95.159

planet	measurement	value
Jupiter	distance_from_sun	748,000,000
Saturn	distance_from_sun	1,483,000,000
Jupiter	radius_km	69,911
Saturn	radius_km	58,232
Jupiter	mass_vs_earth	1,321
Saturn	mass_vs_earth	95

5) This dataset is being reshaped from long to wide. Can you fill in the gaps in the wide dataset?

song	information	value
Happy	artist	Pharrell
Happy	year released	2013
Umbrella	artist	Rihanna
Umbrella	year released	2008
Smells Like Teen Spirit	artist	Nirvana
Smells Like Teen Spirit	year released	1991
Don't Stop Believin'	artist	Journey
Don't Stop Believin'	year released	1981
Sweet Home Alabama	artist	Lynyrd Skynyrd
Sweet Home Alabama	year released	1974
She Loves You	artist	The Beatles
She Loves You	year released	1963
Jailhouse Rock	artist	Elvis Presley
Jailhouse Rock	year released	1957

song	artist	year_released
Happy	Pharrell	2013
Umbrella	Rihanna	2008
Smells Like Teen Spirit	Nirvana	1991
Don't Stop Believin'	Journey	1981
Sweet Home Alabama	Lynyrd Skynyrd	1974
She Loves You	The Beatles	1963
Jailhouse Rock	Elvis Presley	1957

2. Process for reshaping datasets

Section 2.4

6) Find the weather forecast in a location any where in the world that interests you and put the data into a wide and long dataset.

This should include,

- forecast for 4 days or times
- temperature
- wind speed
- change of rain (precipitation)

Here are some websites where you can find weather forecasts

https://www.bbc.co.uk/weather
https://www.metoffice.gov.uk/
https://weather.com/en-GB/
https://www.accuweather.com/

Wide

forecast				

Long

	forecast	

2. Process for reshaping datasets

Note for teachers: the datasets below are examples of types of information the learners could produce. This is the weather forecast for the North Pole.

Wide

forecast	13/11/2021	14/11/2021	15/11/2021	16/11/2021
temperature	-17	-21	-29	-26
wind speed	1	2	1	2
change of rain	10%	20%	20%	22%

Long

date	forecast	value
13/11/2021	temperature	-17
13/11/2021	wind speed	1
13/11/2021	change of rain	10%
14/11/2021	temperature	-21
14/11/2021	wind speed	2
14/11/2021	change of rain	20%
15/11/2021	temperature	-29
15/11/2021	wind speed	1
15/11/2021	change of rain	20%
16/11/2021	temperature	-26
16/11/2021	wind speed	2
16/11/2021	change of rain	22%