Tidy data in Excel

Brendan Clarke, NHS Education for Scotland, brendan.clarke2@nhs.scot

28/06/2024

KIND Learning Network

Welcome

- this session is for Excel beginners
- we'll get going properly at 9.35
- this is a mainly-practical session, and you'll need Excel of some sort to follow along
- if you can't access the chat, you might need to join our Teams channel: **tinyurl.com/kindnetwork**

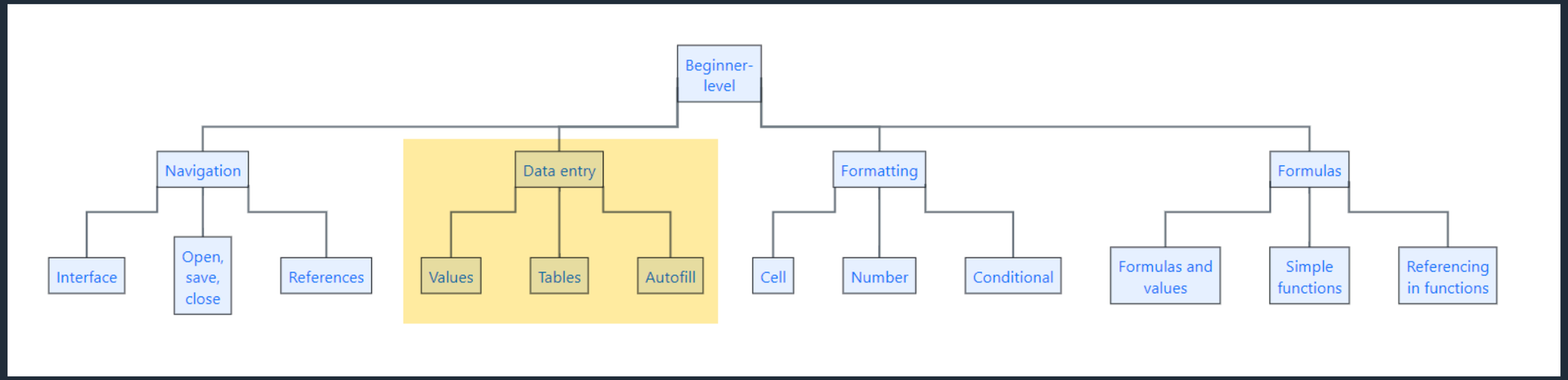
KIND Learning Network

The KIND network

- a social learning space for staff working with **k**nowledge, **i**nformation, and **d**ata across health, social care, and housing in Scotland
- we offer social support, free training, mentoring, community events, …
- Teams channel / mailing list

KIND Learning Network

Where does this fit in?



KIND Excel beginner skill tree

- for this session, you'll need to be familiar with the Excel basics (getting around in Excel, opening/saving/closing files, and a little bit of A1 referencing)
- we're going to dodge formatting and formulas as much as possible today

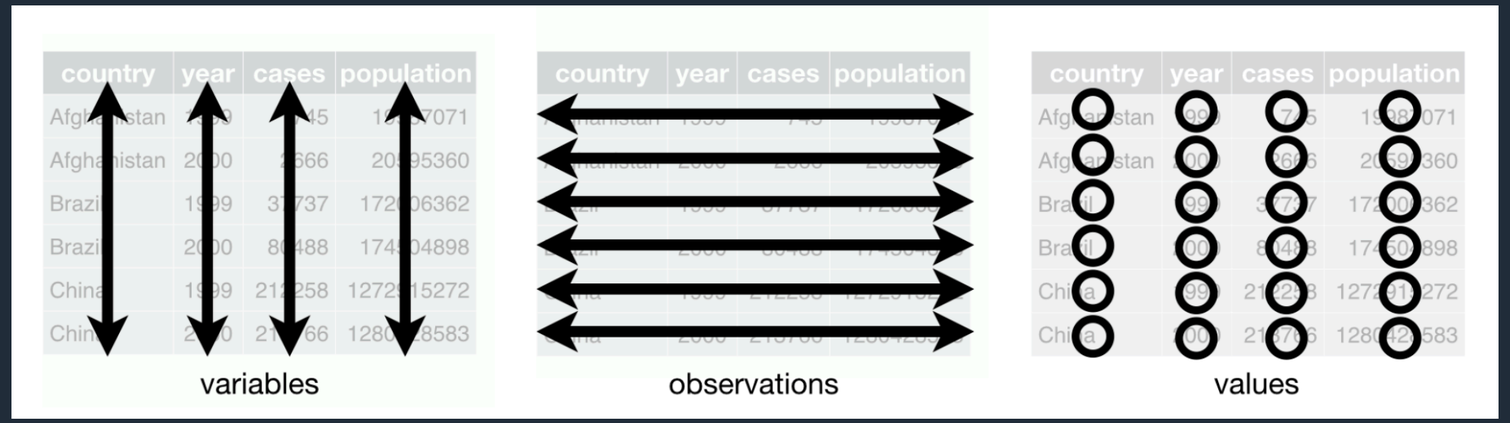
KIND Learning Network

Session outline

- understanding tidy data
- a word of warning
- a practical introduction to making tidy data:
  - values
  - tables
  - autofill
- exercises and demos

KIND Learning Network

Understanding tidy data



R4DS Figure 5.1: The following three rules make a dataset tidy: variables are columns, observations are rows, and values are cells, via CC BY-NC-ND 3.0 US

KIND Learning Network

A word of warning

- tidying data can be very slow and complicated
- in Excel, there are lots of advanced tools that can speed things up
  - PowerQuery especially
- this is a beginner's session, so we'll avoid the more fancy tools
- **but** if your process takes lots of manual work, it's definitely worth exploring alternative ways of working

Values

- `values` is the word we use to describe each bit of information in an Excel spreadsheet. Some examples:
  - a date, like `2024-06-28`
  - a number, like `11.2`
  - a name, like `NHS Grampian`
  - a cost, like `£12.50`
  - …
- each value should have its own cell

Entering values

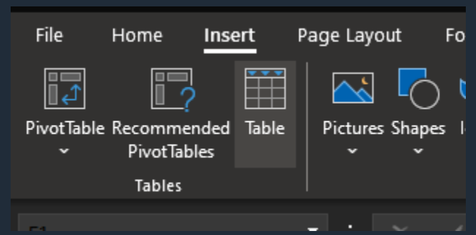
- how would you enter this data: `2018, 2019, 2020, 2021, 2022`?
- please now:
  - open Excel
  - start a new workbook
  - add a column header `year` in cell `A1`
  - then add each of those five values in the five cells underneath (down to `A6`)

More values

- we're going to be using some birthrate data from the NRS for this session. We'll start by adding some birth rate data
- this is given as births per 1,000 women in five year age brackets. We'll start with 25-29 year old mothers
- please add the header `25-29yrs` in cell `B1`
- here are the values for our five years: `73.4, 71, 66.8, 69.6, 66.7`

Tables

- you should keep your new data in a **table**
- `Insert` > `Table`

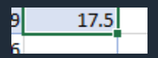


- tables allow you to sort and filter your data
- they also act as a useful ‘container’ (or **data structure**) for your data

Extending tables

30-34yrs
90.9
88.6
83.4
85.9
84.4

Autofill

- one last way of adding values: autofill
- drag again to make a new empty column, and label it `difference`
- in `D2` (the first ‘proper’ cell), copy this formula: `=C2-B2`. This will calculate the difference in birth rates between the two columns
- finally, click the small green corner of that newly-filled cell to autofill the column 

Back to tidy data

- we've now got some data with:
  - each value in a cell
  - each variable in a column
    - here, this is a maternal age bracket
  - each observation in a row
    - here, this is a year
- we could work through and extend this data by hand, but we'll now switch over to some supplied data to save all the typing
  - errors are common in manual data-entry
  - if you can import data, that's usually better than re-typing it

Many values per cell

- We often find useful data with more than one value per cell
- this can be helpful for humans

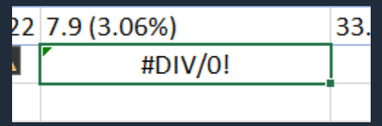
	15-19	20-24	25-29
09	24.3 (7.03%)	63.8 (18.47%)	93.8 (27.15%)
10	22.9 (6.68%)	59.9 (17.48%)	92.6 (27.02%)
11	21.1 (6.23%)	57.9 (17.1%)	90.2 (26.64%)
12	19.7 (5.89%)	55.2 (16.51%)	90.9 (27.19%)
13	17.9 (5.56%)	52.8 (16.39%)	85.7 (26.6%)

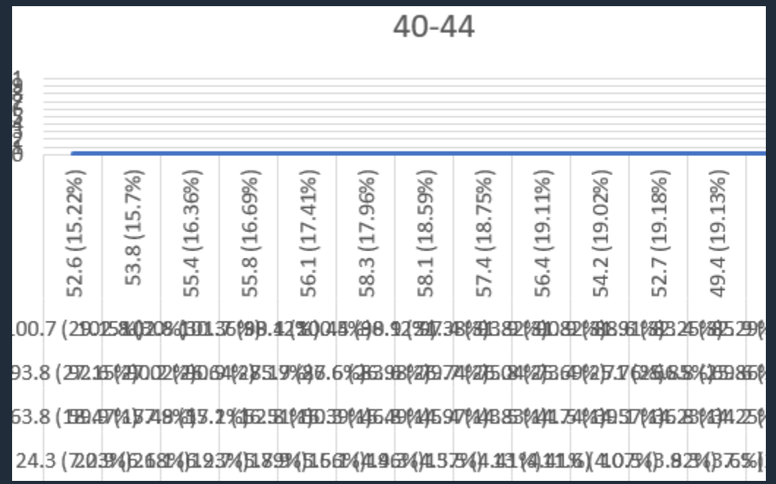
- Excel can’t do anything with this data

Exercise one: many values per cell

- find the `Exercise one` sheet in the exercise file
- try calculating an average for each of the groups
- or, if you're more confident, try plotting the data

Nothing works!



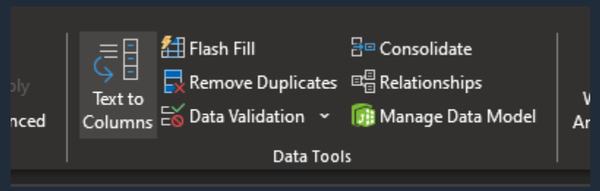


KIND Learning Network

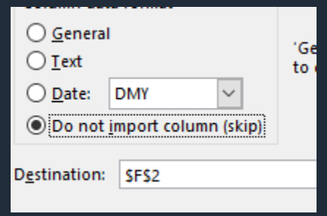
What’s the solution?

Text to Columns

- select a column
- in the `Data` tab of the ribbon, you should find the `Text to Columns` tool

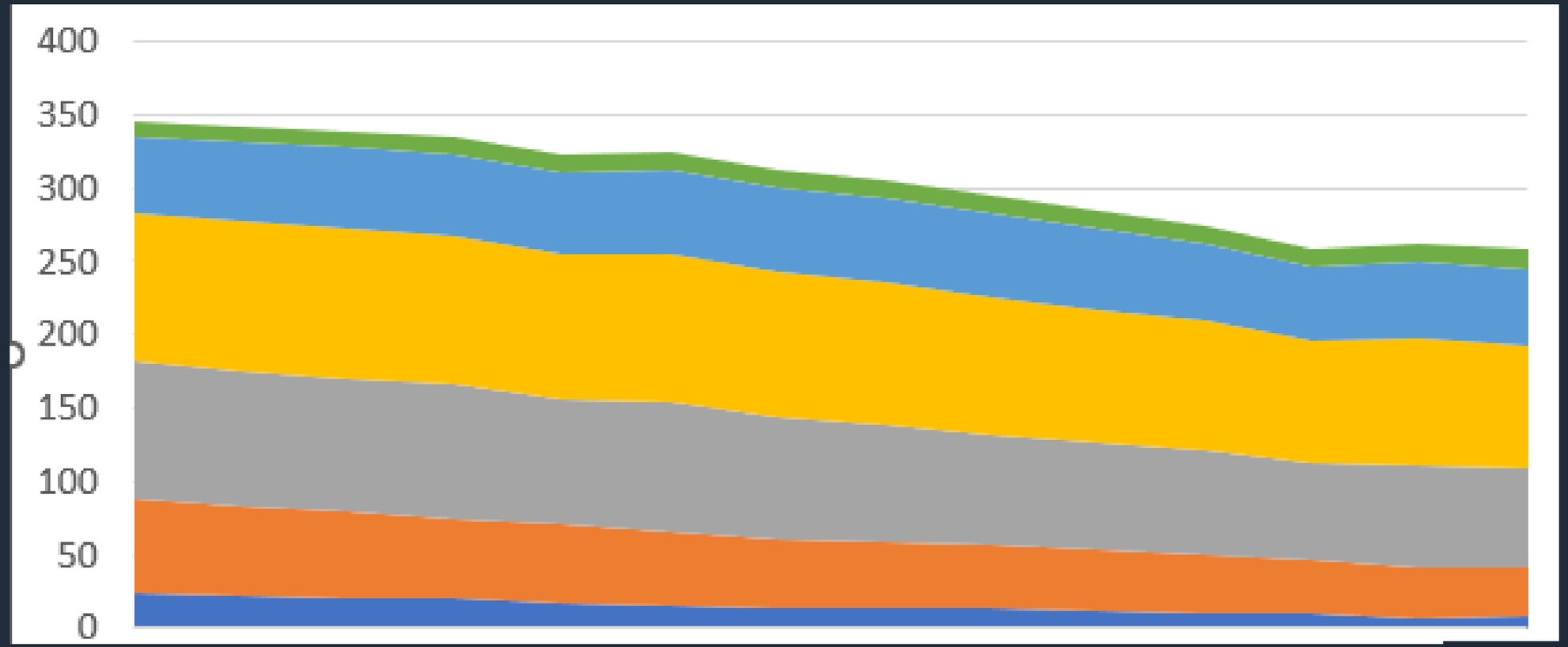


- note that you can keep, or remove, the percentage column. We'll **skip** it, to keep things simple



KIND Learning Network

Try working with that data again



We can plot the data now

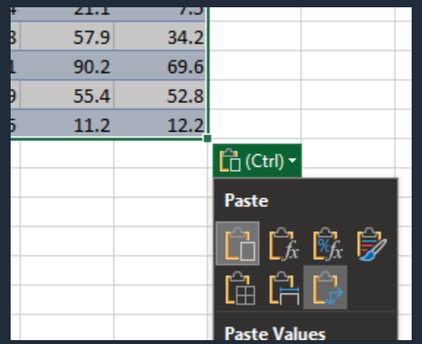
KIND Learning Network

Exercise two: transposed data

- sometimes you'll find data where the columns and rows have been flipped

	A	B	C	D	E	F	G	H
1	group	1951	1961	1971	1981	1991	2001	2011
2	15-19	19.6	33.7	47.7	30.5	33.3	28.4	21.1
3	20-24	128.6	179.4	163.5	112.3	82.3	57.8	57.9
4	30-34	147.3	188.9	164.4	131.3	116.5	85.1	90.2
5	35-39	59.4	56.7	36.5	20.8	26.8	36.9	55.4
6	40-44	17	16.1	9.2	3.9	4	6.5	11.2
7								

- that's slow to fix by hand, but luckily you can **transpose it**, which swaps rows and columns
- select your data, and copy/paste into a new cell
- then use the transpose option

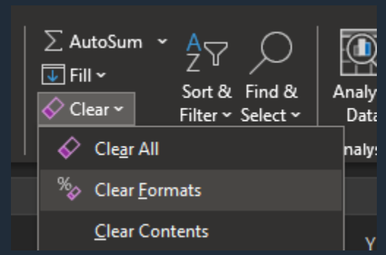


KIND Learning Network

Exercise two: transposed data

- you might need to fix formatting and labels:

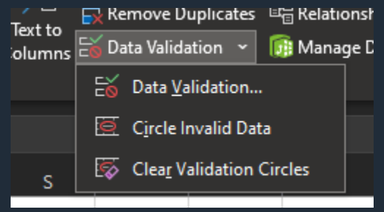
group	15-19	20-24	30-34	35-39	40-44
1951	19.6	128.6	147.3	59.4	17
1961	33.7	179.4	188.9	56.7	16.1
1971	47.7	163.5	164.4	36.5	9.2



- the `Clear formats` option might help this
- if you run into trouble, please note that transposing only works on data that is not in a table

Demo one: validation and really messy data

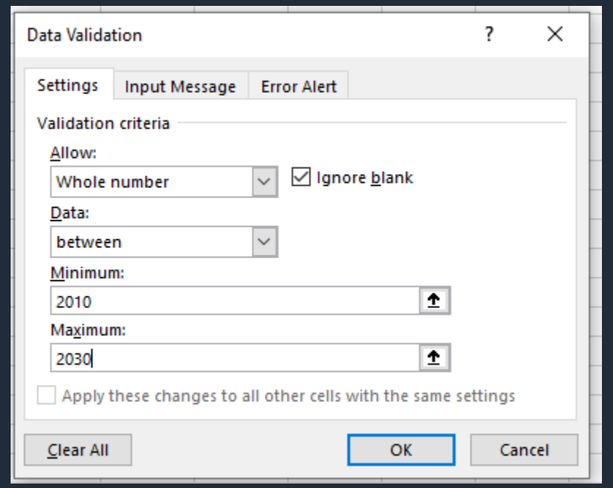
- one of the most time-consuming bits of tidying is checking your values
- we'll briefly demonstrate the **data validation** tool



- this allows you to describe what format you think your data should be in, and then highlights anything that doesn't fit

Demo one: validation and really messy data

- we select the `years` column
- then open the data validation tool
- then set appropriate validation options, so **Whole number** between 2010 and 2030



KIND Learning Network

Demo one: validation and really messy data

- then select `Circle Invalid Data`

	A	B	C
1	year	25-29yer	30-34yr
2	201 8	73.4	9o.9
3	2019	71	88.6
4	20	66.8	83 4
5	2021	69.6	85.9
6	2022	66,7	84.4
7			

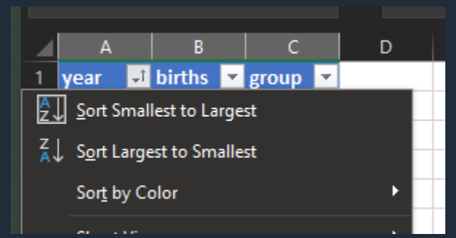
- we can now go through and fix anything circled

Demo two: reshaped data

- there are several ways of reshaping data that's not in a tidy format
- we'll look at the manual way here, but - as it's horrible - I'd be keen to encourage you to investigate Power Query or Pivot Tables to reshape if this is a regular part of your working day. PQ takes < 10 seconds, PT not much longer

Manual reshape

- sort the data by year



- then copy and paste blocks of data, making sure to keep the years aligned

	A	B	C	D	E	F	G	H	I
1	year	births	group						
2	1951	19.6	15-19	1951	128.6	20-24	1951	147.3	30-34
3	1961	33.7	15-19	1961	179.4	20-24	1961	188.9	30-34
4	1971	47.7	15-19	1971	163.5	20-24	1971	164.4	30-34
5	1981	30.5	15-19	1981	112.3	20-24	1981	131.3	30-34
6	1991	33.3	15-19	1991	82.3	20-24	1991	116.5	30-34
7	2001	28.4	15-19	2001	57.8	20-24	2001	85.1	30-34
8	2011	21.1	15-19	2011	57.8	20-24	2011	80.2	30-34

Manual reshape

- make sure you then copy the age brackets to label the column

	40-44	
1	17	40-44
1	16.1	40-44
1	9.2	40-44
1	3.9	40-44
1	4	40-44

- then delete the spare years columns, and the age brackets

	A	B	C	D	E	F
1	year	15-19	20-24	30-34	35-39	40-44
2	1951	19.6	128.6	147.3	59.4	17
3	1961	33.7	179.4	188.9	56.7	16.1
4	1971	47.7	163.5	164.4	36.5	9.2
5	1981	30.5	112.3	131.3	20.8	3.9

Forthcoming Excel sessions

Session	Date	Area	Level
Excel tables	10:00-10:30 Mon 1st July 2024	Excel	🌶 :beginner-level
Formulas in Excel	15:00-16:00 Wed 3rd July 2024	Excel	🌶 :beginner-level
Lambda formulas in Excel	13:00-13:30 Mon 15th July 2024	Excel	🌶🌶 : intermediate-level
Lookups in Excel	13:00-14:30 Thu 1st August 2024	Excel	🌶🌶 : intermediate-level

KIND Learning Network

Session	Date	Area	Level
Relative, absolute, mixed, structured, and R1C1 references in Excel	15:00-16:00 Thu 8th August 2024	Excel	🌶🌶 : **intermediate-level**

KIND Learning Network

Feedback

Feedback link

Please give us one minute of your time. We add feedback comments to our training pages, because we think this is the most useful resource for people looking for specific training that suits their needs