Introduction to data science in health and social care

Week 5

Brittany Blankinship | 19 October 2022

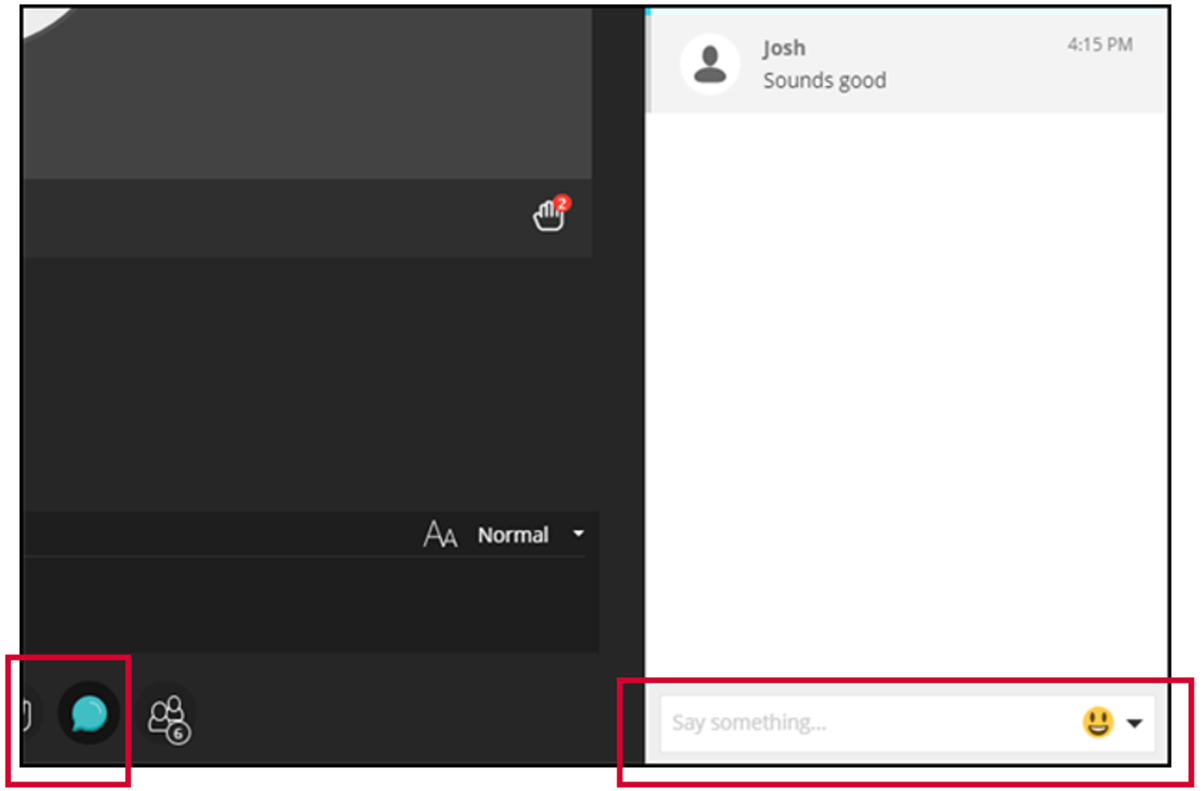
Audio check

Can you hear the presenter talking?

Please type **yes** or **no** in the “Text chat area”

If you can’t hear:

- Check your Audio/Visual settings in the Collaborate Panel
- Try signing out and signing back into the session
- Type into the chat box and a moderator will try to assist you

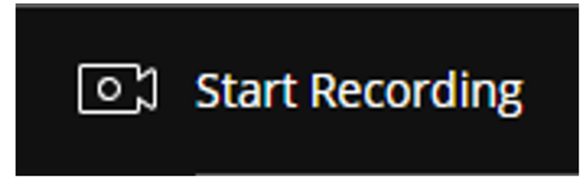


ddi.hsc.talent@ed.ac.uk

Supported by

Recording

This session will now be recorded. Any further information that you provide during a session is optional and in doing so you give us consent to process this information.



These sessions will be stored by the University of Edinburgh for one year and published for 30 days after the event. Schools or Services may use the recordings for up to a year on relevant websites.

By taking part in a session, you give us your consent to process any information you provide during it.

ddi.hsc.talent@ed.ac.uk

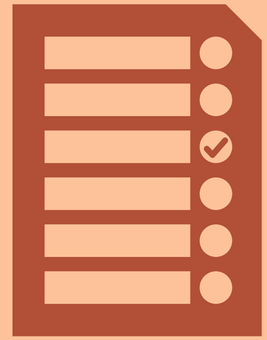
Supported by

Introduction to data science in health and social care

Week 5

Brittany Blankinship | 19 October 2022

Agenda



- Course/assessment updates
- Review of some key functions and principles
- Paired programming exercises
- Questions

Graded Discussion Post

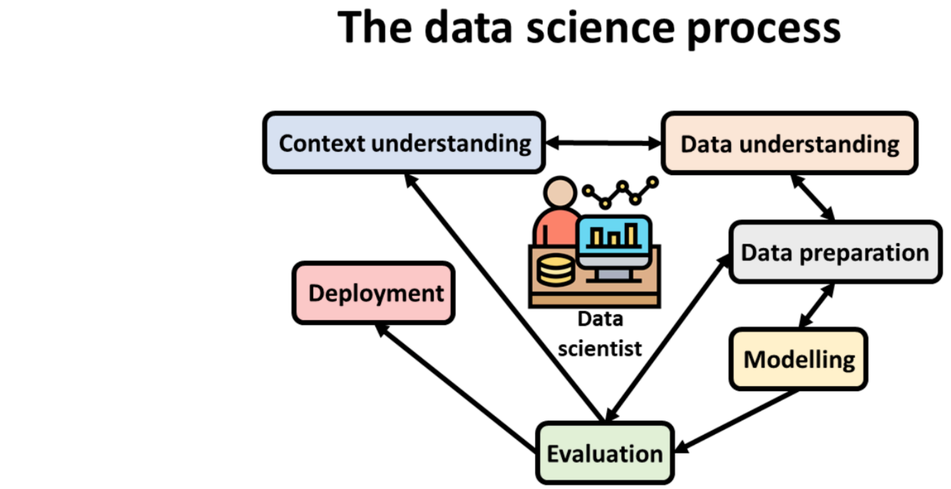
- Due: **21 October 12:00 UK time**

- Reply to a peer's response to receive full marks
- Can view other posts after submitting yours
  - Submit your response by creating a thread

Instructions

Please craft a response to the following discussion board prompt:

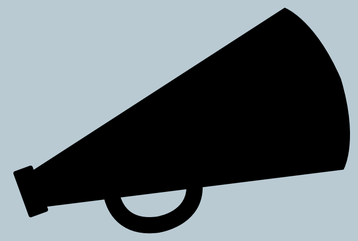
Select a topic within health and social care that interests you, how might you use data science or data driven innovation to improve or impact the systems for the better? Link your response to the Data Science Process discussed in Week 2.



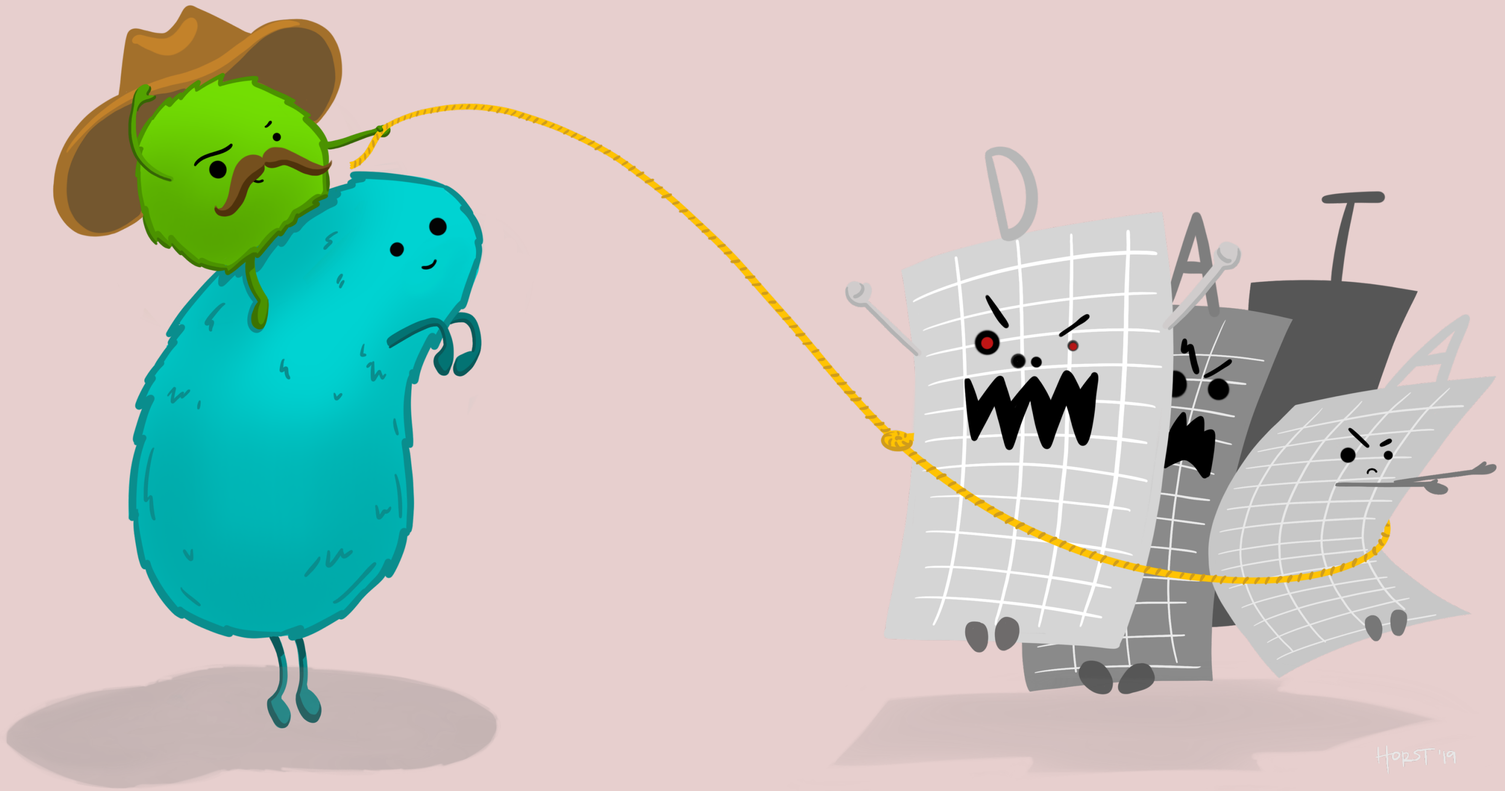
No live sessions next week

- Reading week 24 October to 30 October
  - Week 6 materials will be available on this Friday if you wish to get ahead
  - Use this week to catch up, revise what we have covered, and connect with your datathon team!
- Week 6 will start 31 October with our next live tutorial on 2 November

Mid-course feedback survey



Wrangling

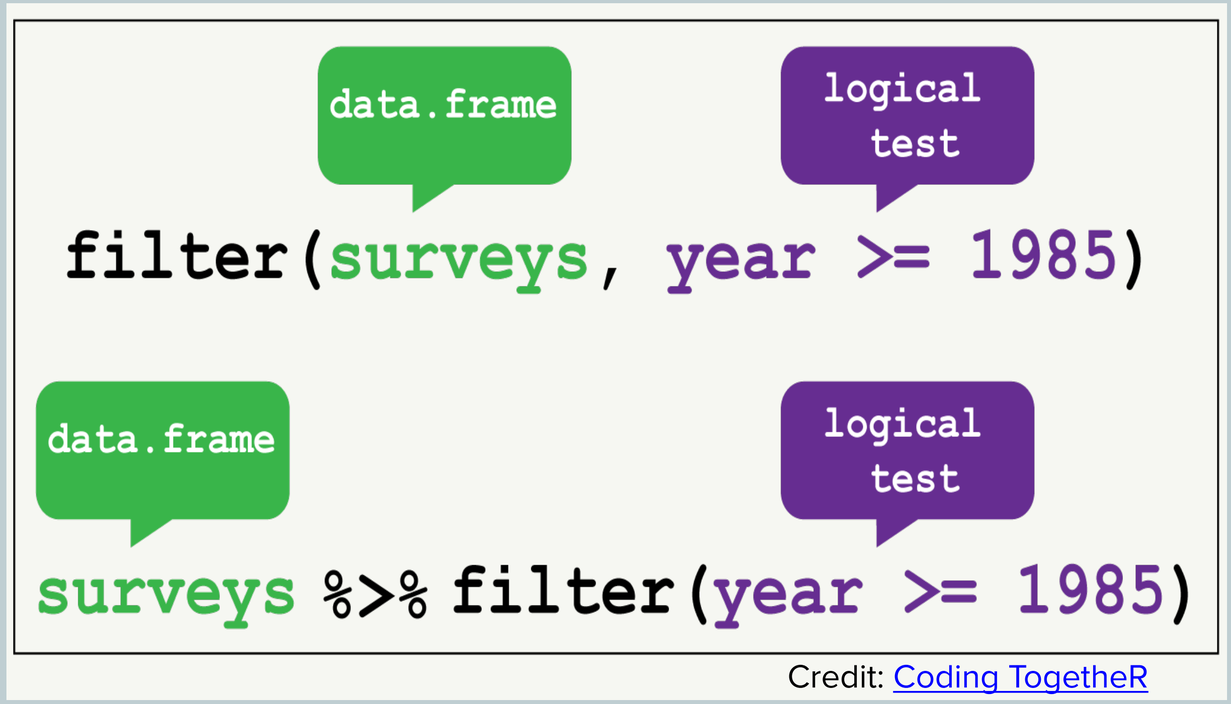


Artwork by Allison Horst

Subsetting Data (observations)

🌟`filter()`

Extract rows of existing data that meeting logical conditions



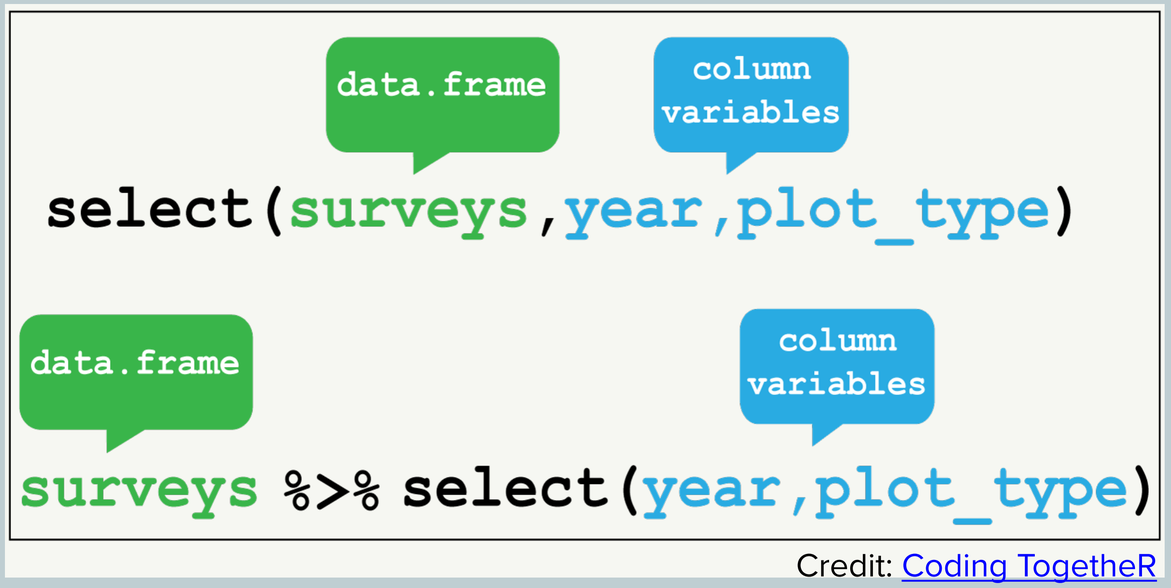
Credit: Coding TogetheR

For more check out this RStudio Data Manipulation video from Garrett Grolemund https://www.youtube.com/watch?v=Zc_ufg4uW4U

Subsetting Data (variables)

🌟`select()`

Select columns by name



Credit: Coding TogetheR

Logical Operators

<	Less than
>	Greater than
==	Equal to
<=	Less than equal to
>=	Greater than equal to
!=	Not equal to
%in%	Group membership

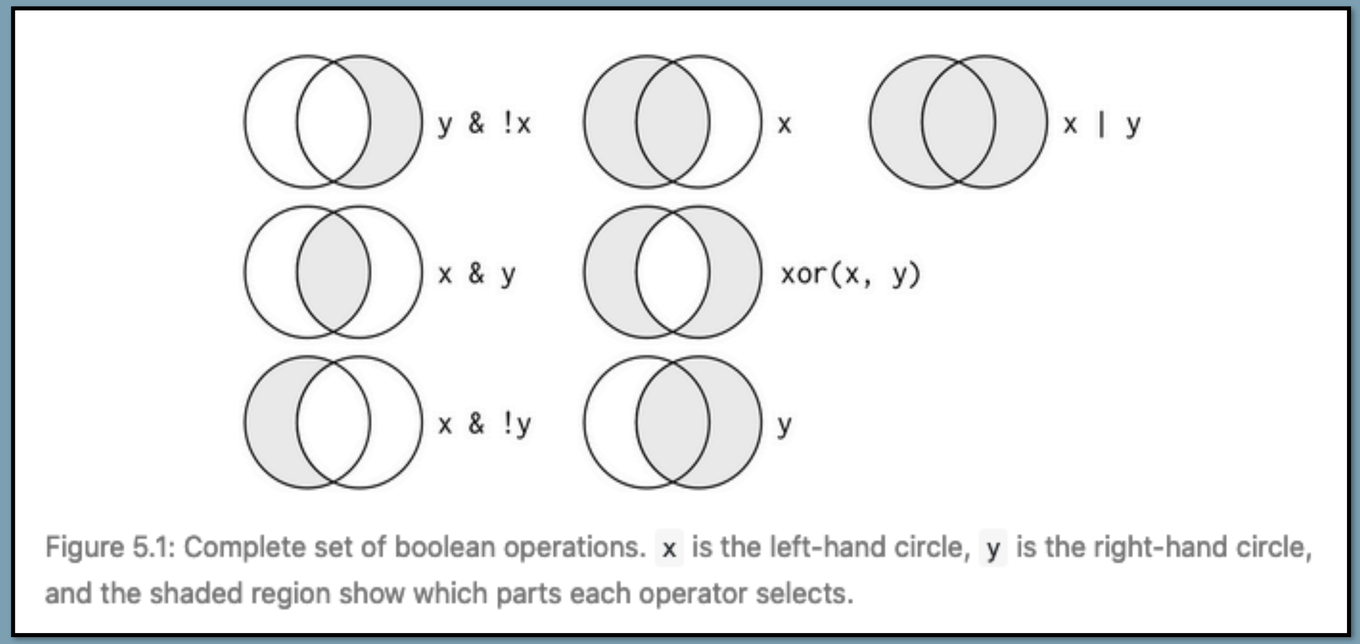


Figure 5.1: Complete set of boolean operations. `x` is the left-hand circle, `y` is the right-hand circle, and the shaded region show which parts each operator selects.

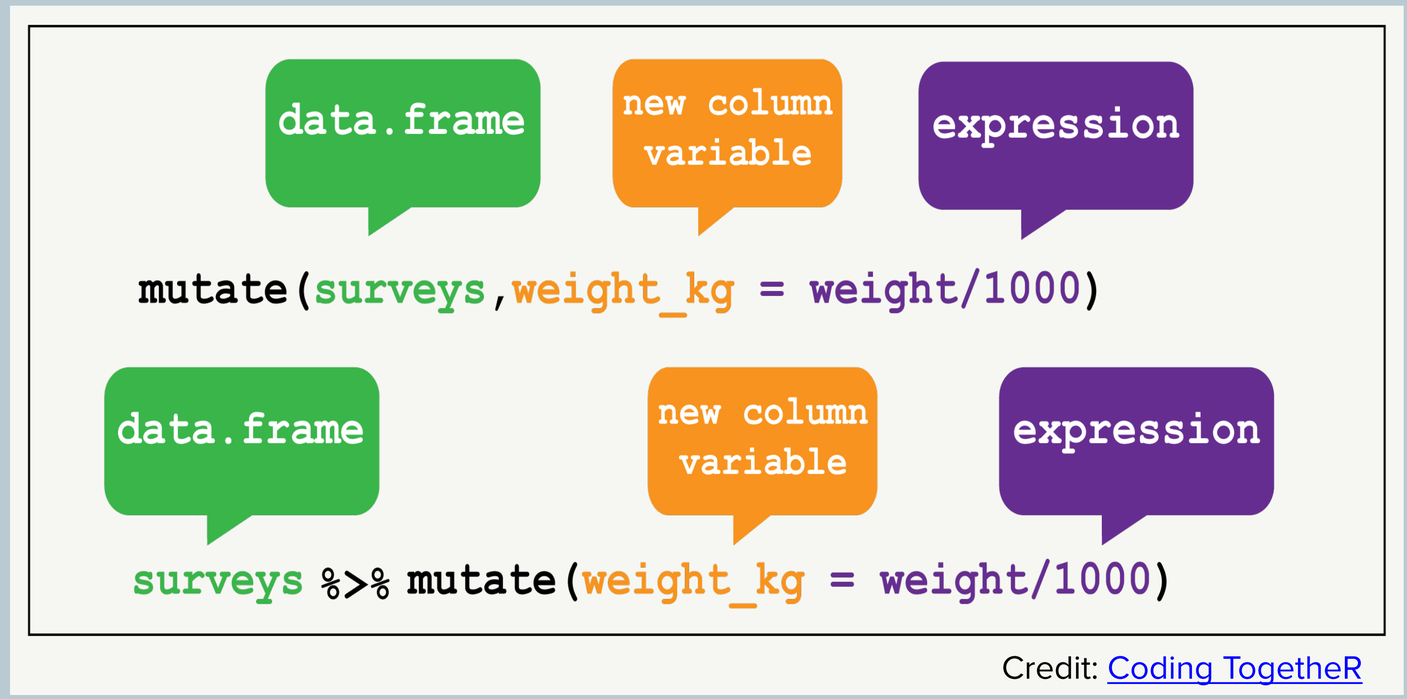
Source: R for Data Science book, Figure 5.1

New Variables

🌟`mutate()`

Compute and append one or more new columns – changes an existing column or adds a new one

Original columns remain after being passed to mutate



Credit: Coding TogetheR

For more check out this RStudio Data Manipluation video from Garrett Grolemund https://www.youtube.com/watch?v=Zc_ufg4uW4U

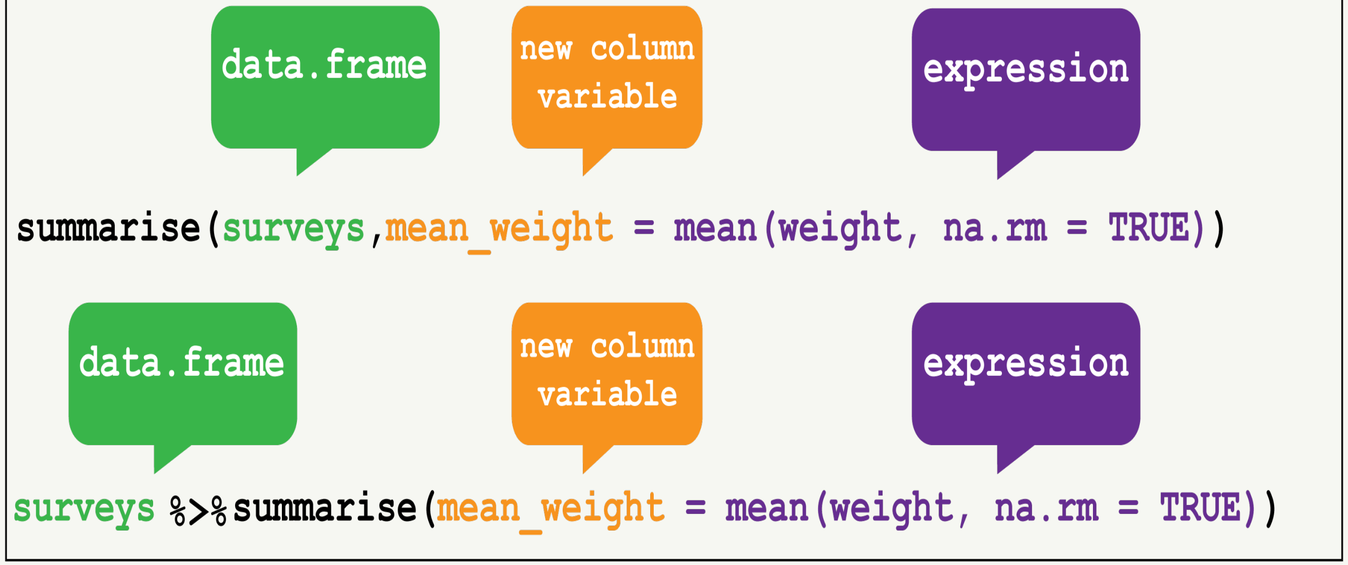
Summarise

🌟`summarise()`

Summarise data into single row of values

Drops variables not selected or or created inside it

- drops columns after calculating the new one



Credit: Coding TogetheR

For more check out this RStudio Data Manipluation video from Garrett Grolemund https://www.youtube.com/watch?v=Zc_ufg4uW4U

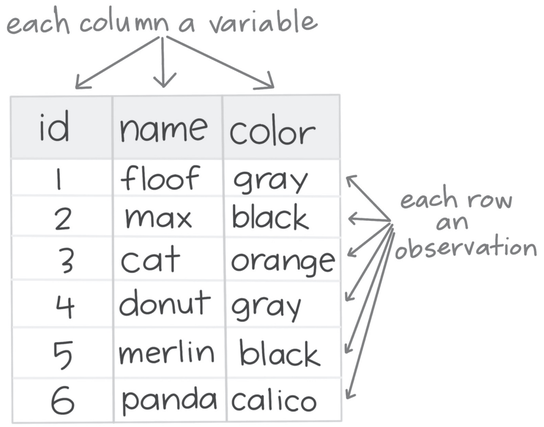
Data formats & Tidy Data

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

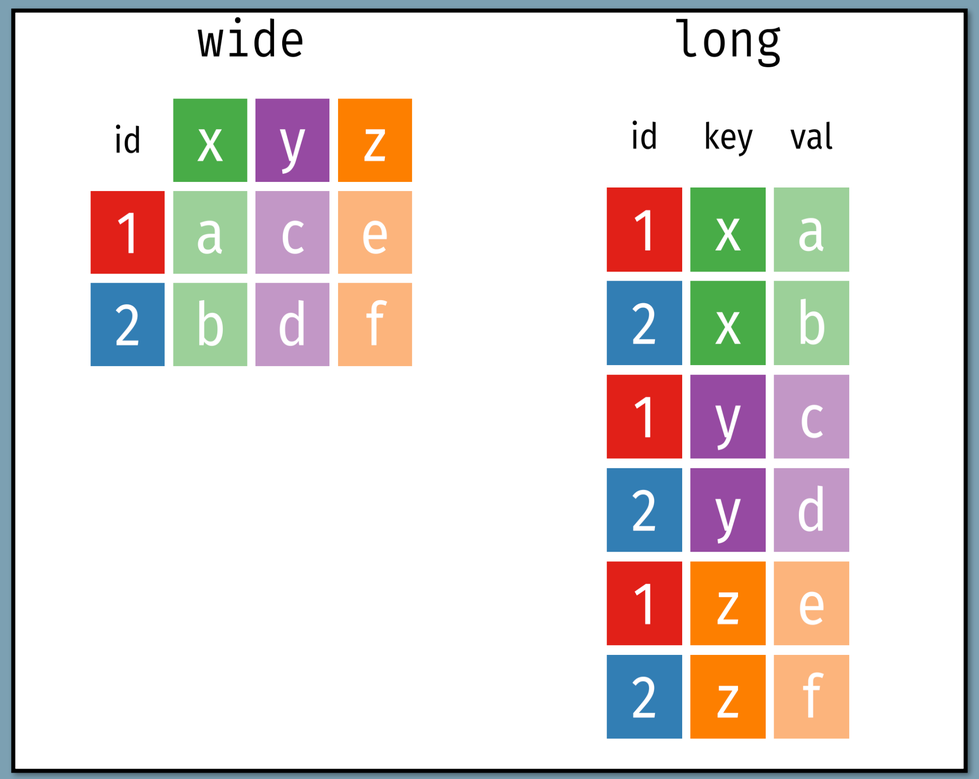
Ways data can become untidy:

- Column headers contain values, rather than names
- Multiple variables are stored in a single column
- Variables are stored in both rows and columns
- Multiple observational types are stored in a single table
- A single observational unit is stored in multiple tables

Wickham, H. (2014). Tidy data. *Journal of statistical software*, *59*(1), 1-23.

For more on tidy data see the above paper & Chapter 12 of the R for Data Science Book

Data formats & Tidy Data



"Long" format

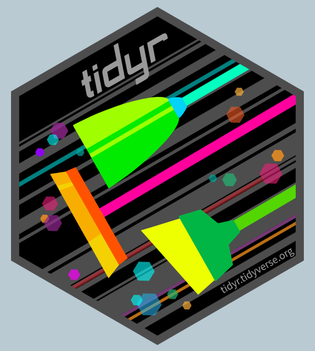
country	year	metric
x	1960	10
x	1970	13
x	2010	15
y	1960	20
y	1970	23
y	2010	25
z	1960	30
z	1970	33
z	2010	35

"Wide" format

country	yr1960	yr1970	yr2010
x	10	13	15
y	20	23	25
z	30	33	35

Wide format = generally untidy, but found in many datasets & useful for tables

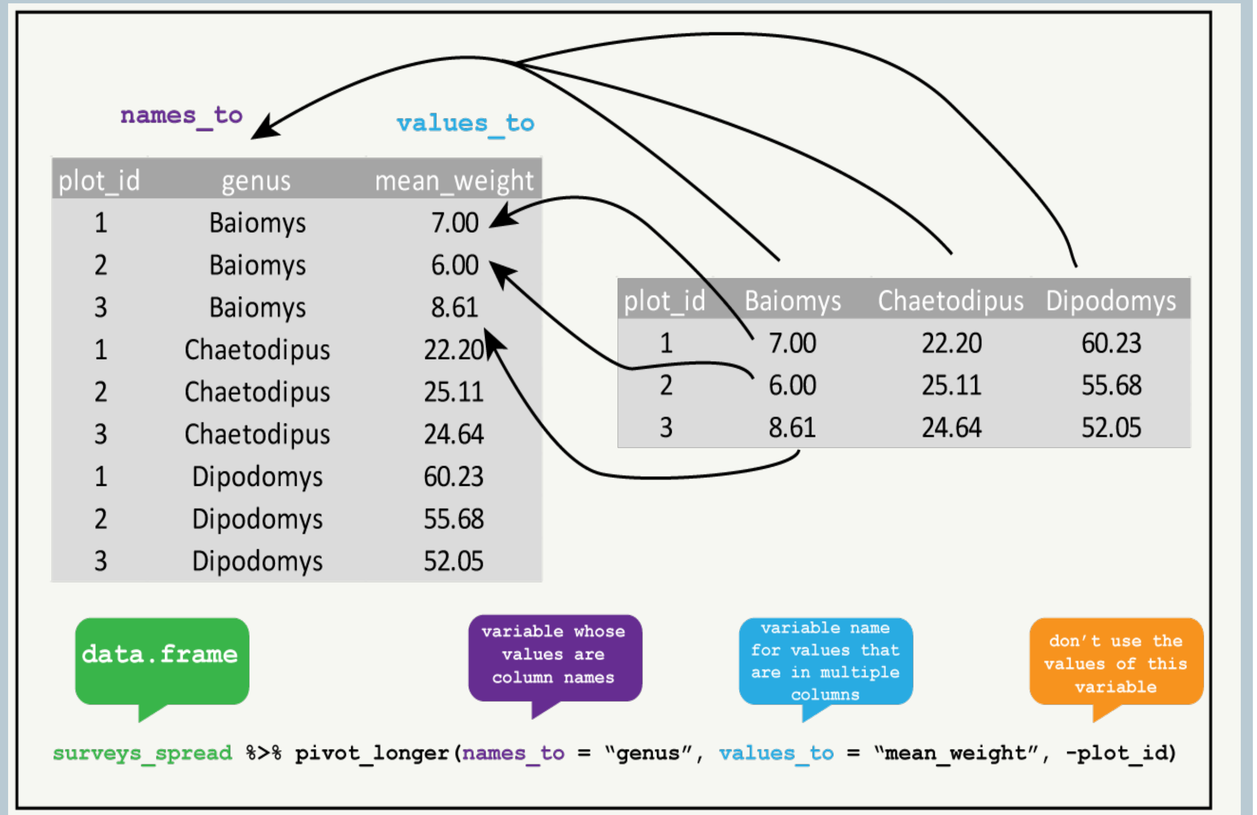
Transforming Data (wide to long)



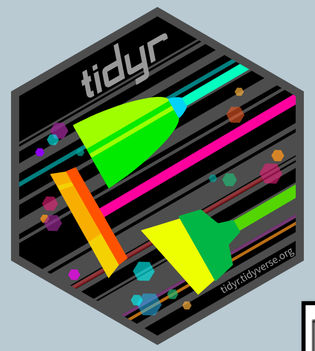
🌟`pivot_longer()`

Requires:

1. `data` = data you want to pivot
2. `names_to` = name of the column you want to create to capture condition, requires a character string
3. `values_to` = name of column you want to contain data values, requires a character string
4. `column X:column Y` = range of columns that you have and want to pivot longer, or that you do not want to pivot



Credit: Alistair Bailey's Coding TogetheR



Transforming Data

⭐ `pivot_longer()` = wide to long

country	1999	2000	2001	2002
Angola	800	750	925	1020
India	20100	25650	26800	27255
Mongolia	450	512	510	586

Pivot data longer

```
data %>%
  pivot_longer(
    cols = 1999:2002,
    names_to = "year",
    values_to = "cases"
  )
```

country	year	cases
Angola	1999	800
Angola	2000	750
Angola	2001	925
Angola	2002	1020
India	1999	20100
India	2000	25650
India	2001	26800
India	2002	27255
Mongolia	1999	450
Mongolia	2000	512
Mongolia	2001	510
Mongolia	2002	586

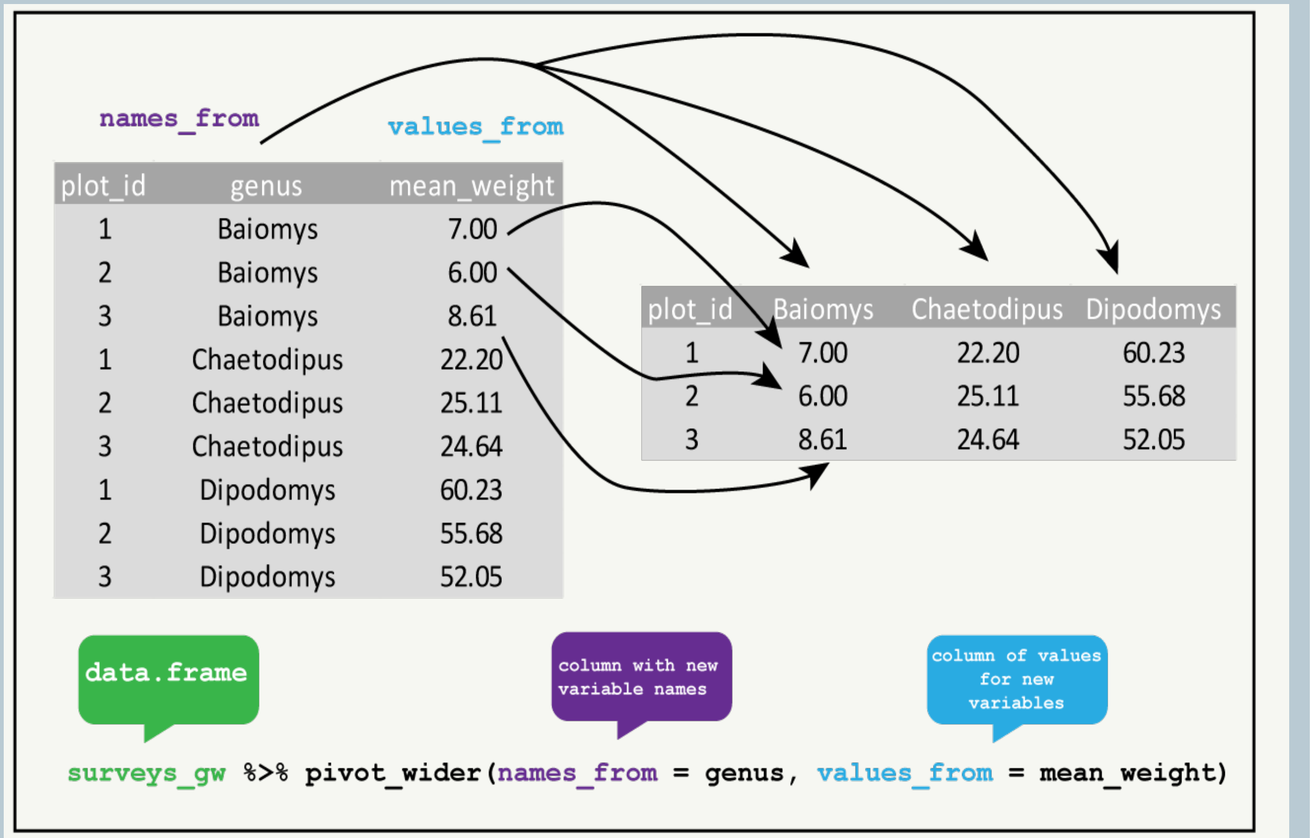
Credit: Epidemiologist R Handbook

Transforming Data (long to wide)

🌟`pivot_wider()`

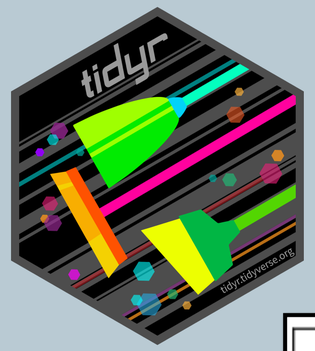
Requires:

1. `data` = data you want to pivot
2. `names_from` = name of column you want to end up in several columns
3. `values_from` = name of column that currently contains data values



Credit: Alistair Bailey's Coding TogetheR

For more check out this RStudio Data Wrangling video from Garrett Grolemund https://www.youtube.com/watch?v=1ELALQlO-yM - however includes the now superseded functions `gather()` & `spread()`



Transforming Data

🌟 `pivot_wider()` = long to wide

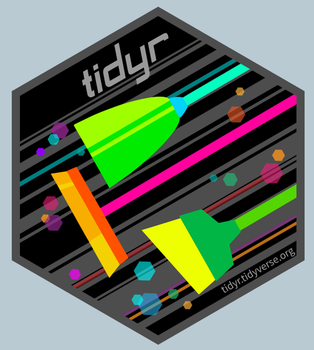
country	year	cases
Angola	1999	800
Angola	2000	750
Angola	2001	925
Angola	2002	1020
India	1999	20100
India	2000	25650
India	2001	26800
India	2002	27255
Mongolia	1999	450
Mongolia	2000	512
Mongolia	2001	510
Mongolia	2002	586

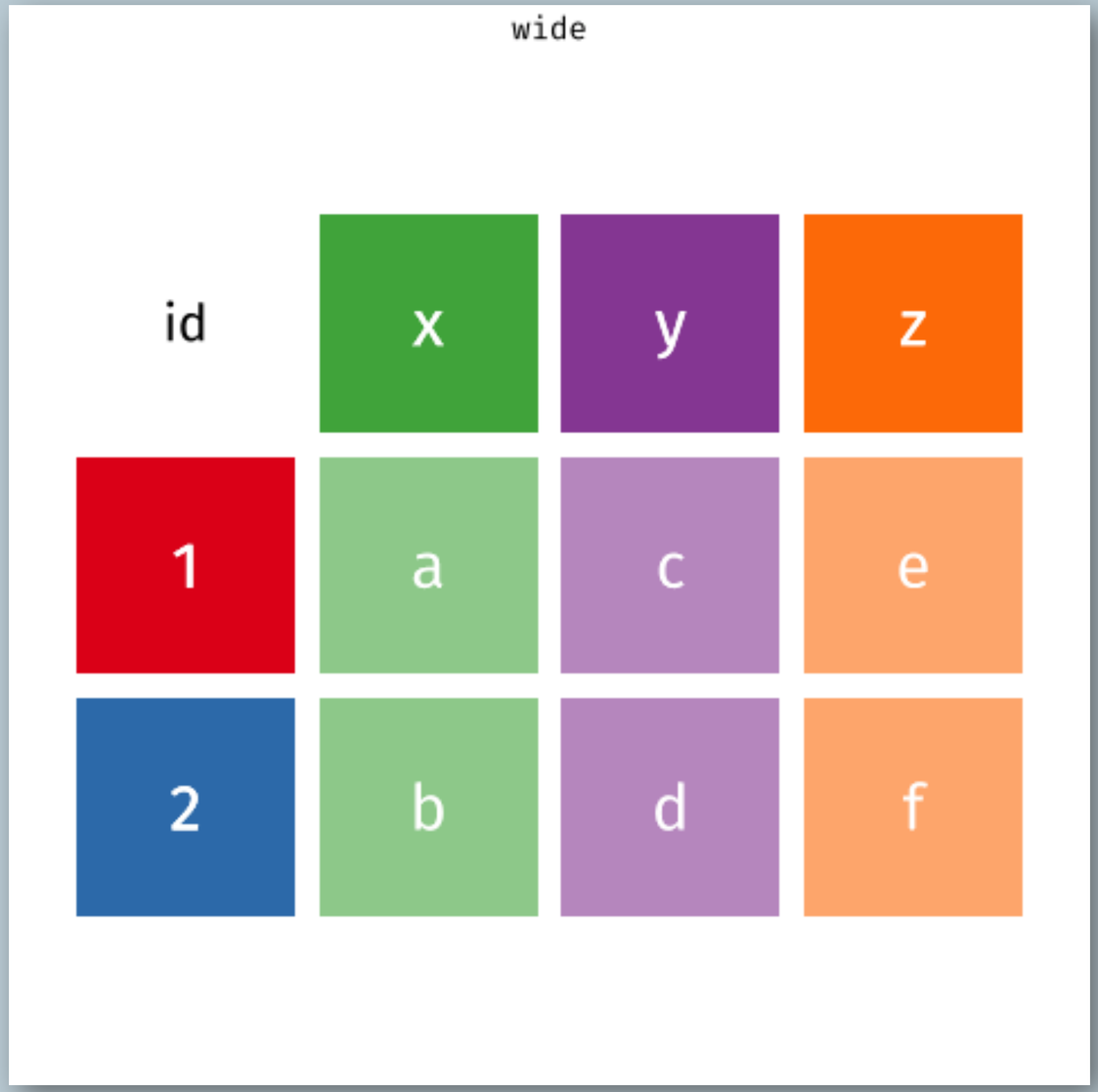
country	1999	2000	2001	2002
Angola	800	750	925	1020
India	20100	25650	26800	27255
Mongolia	450	512	510	586

Pivot data wider

```
data %>%
  pivot_wider(
    names_from = "year",
    values_from = "cases"
  )
```

Credit: Epidemiologist R Handbook

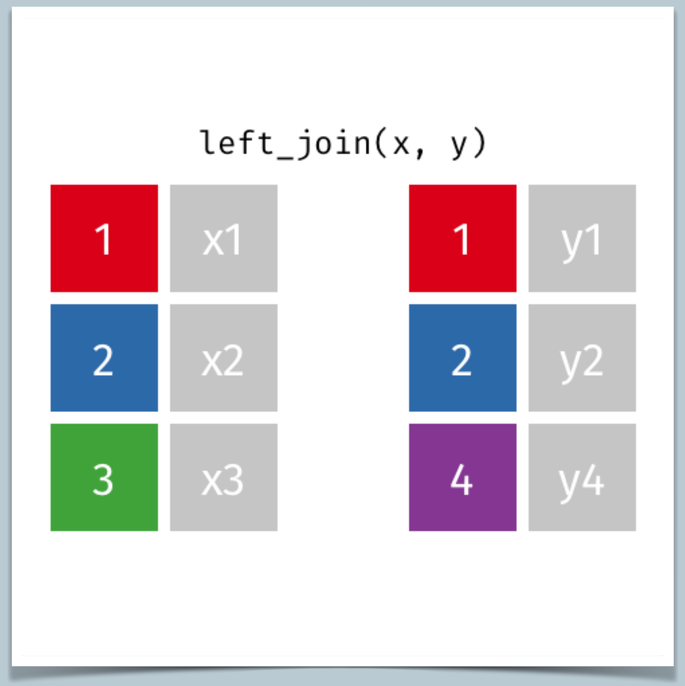


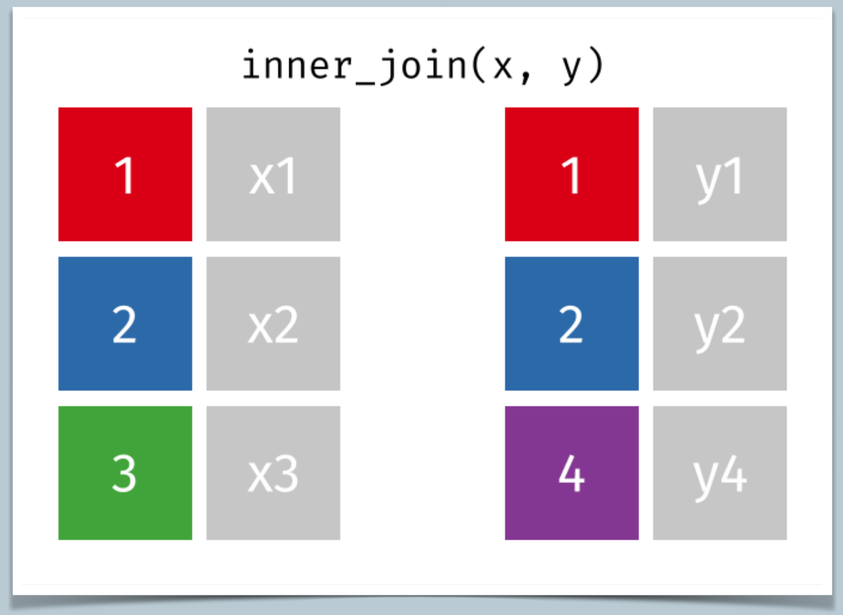


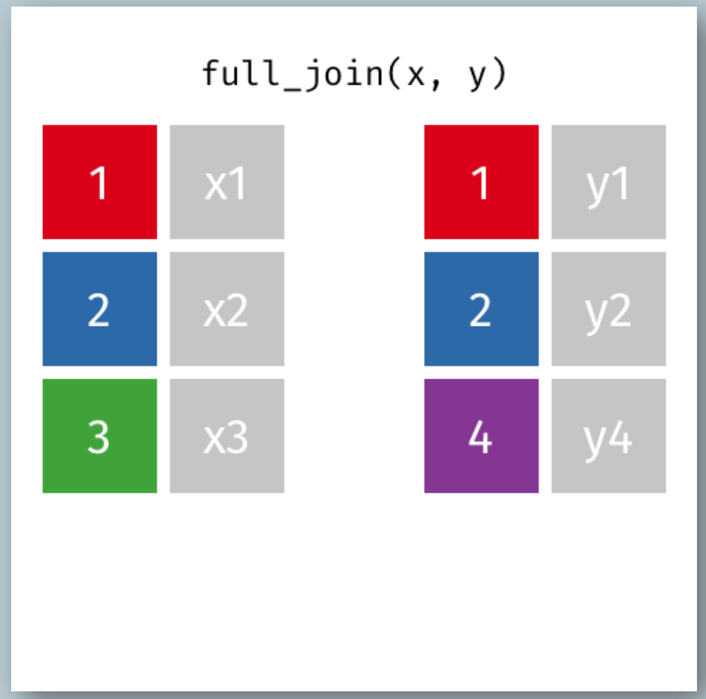
Credit: Garrick Aden-Buie & Mara Averick

Joins

- `left_join()`
- `inner_join()`
- `full_join()`







For a fun explanation using The Beatles & Rolling Stones see Nic Crane's tweet: https://bit.ly/3qfhBb9

Credit: Garrick Aden-Buie
https://www.garrickadenbuie.com/project/tidyexplain/

Paired programming

We have 2 documents

1. data wrangling

2. data storytelling interactive document

Questions?