Workshop 4:
The tidyverse and beyond

- It's the end of base R as you know it

Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork

Recall: The tidyverse package

A bundle of ~20 individual R packages

The six main ones are loaded when the tidyverse package is called

```
> library(tidyverse)

Loading tidyverse: tidyr
Loading tidyverse: readr

Loading tidyverse: dplyr
```

Data tidying

Data import

Data manipulation

Importing data with readr

- Used to read plain text rectangular files into R (e.g. csv)

- read_csv is the equivalent of read.csv in base R

- readr has a number of advantages over base R import function
    - ~10X faster
    - produces tibbles
    - doesn’t convert character vectors to factors
    - more reproducible (readr code on your computer is likely to work on another computer)

Importing data from stata

and other stats packages

- Stata, SPSS and SAS file formats can be imported **and** exported using the tidyverse package "haven"

Importing data with other packages

- Very often the data you handle may not be your own
    - collaborators
    - government
    - online databases
    - publications
    - webpages

5481 Datasets

102 Publishers

Search Datasets...

Advanced Search

IRELAND'S **OPEN DATA** PORTAL

Promoting innovation and transparency through the publication of Irish Public Sector data in open, free and reusable formats.

Explore Datasets

- Agriculture, Fisheries, Forestry & Food
- Arts, Culture and Heritage
- Justice, Legal System, and Public Safety
- Economy and Finance
- Education and Sport
- Energy
- Environment
- Government and Public Sector
- Health
- Housing and Zoning
- Population and Society
- Science and Technology
- Regions and Cities
- Transport

Worksheet A

Open script1_ws4_working_with_readr.R

Data transformation with tidyr

- In this part of the workshop we will look at using some key tidyr functions:
    - separate()
    - gather()
    - spread()
    - unite()

Worksheet B

Open script2_ws4_working_with_tidyr.R

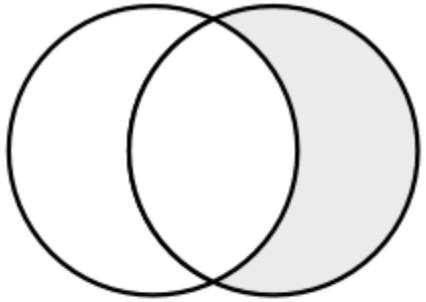
Data transformation with dplyr

- For the last part of the workshop, we will look at using some key dplyr functions:
    - filter()
    - select()
    - mutate()

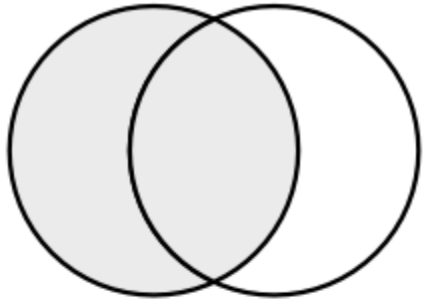
- Next week
    - join
    - summarise() or summarize()
        [depending on your grammatical upbringings!!]

    - exploratory data analysis
    - missing data

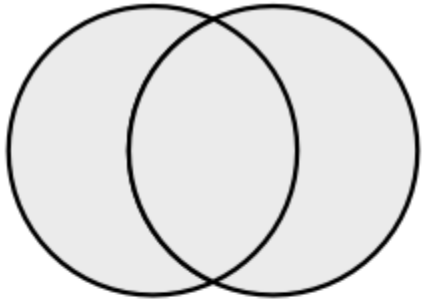
Logical operators and conditional subsetting



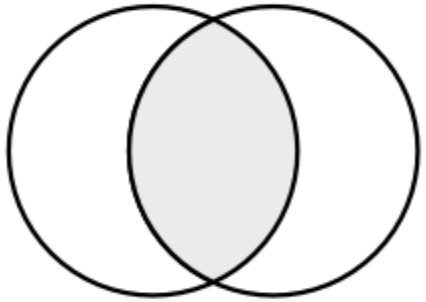
`y & !x`



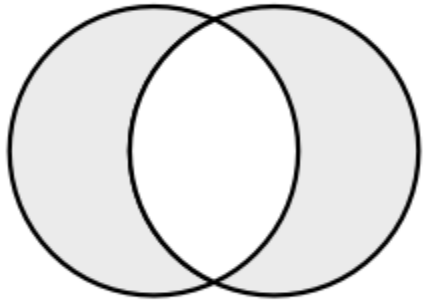
`x`



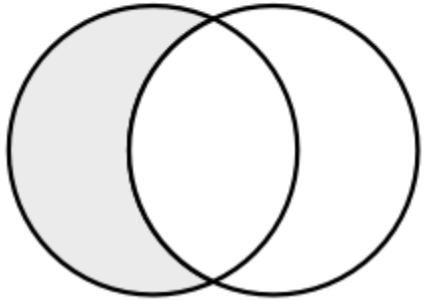
`x | y`



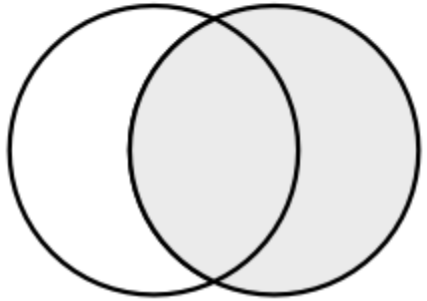
`x & y`



`xor(x, y)`



`x & !y`



`y`

- **`&` -> AND**
- **`|` -> OR (inclusive)**
- **`!` -> NOT**
- **`==` -> EQUAL**
- **`!=` -> NOT EQUAL**

Image: Hadley Wickham R for Data Science Chapter 3

Worksheet C
Open script3_ws4_working_with_dplyr.R

A word of caution

- Computers use finite precision arithmetic
- Therefore all numbers are an approximation
- For this reason, instead of using == for numeric searches, use near()
- See lines 11-15 of the worksheet