

# Data Analysis and Visualisation in R: Day 2

Instructor: Philip (Pip) Grylls

Helpers: Philly Broadbent

# Learning Objectives

- Factors
  - Strings v. factors
- Tidyverse
  - dplyr
    - select
    - filter
    - mutate
    - summarise
    - Arrange
    - count
  - tidyr
    - spread
    - gather
- Pipes
  - %>%

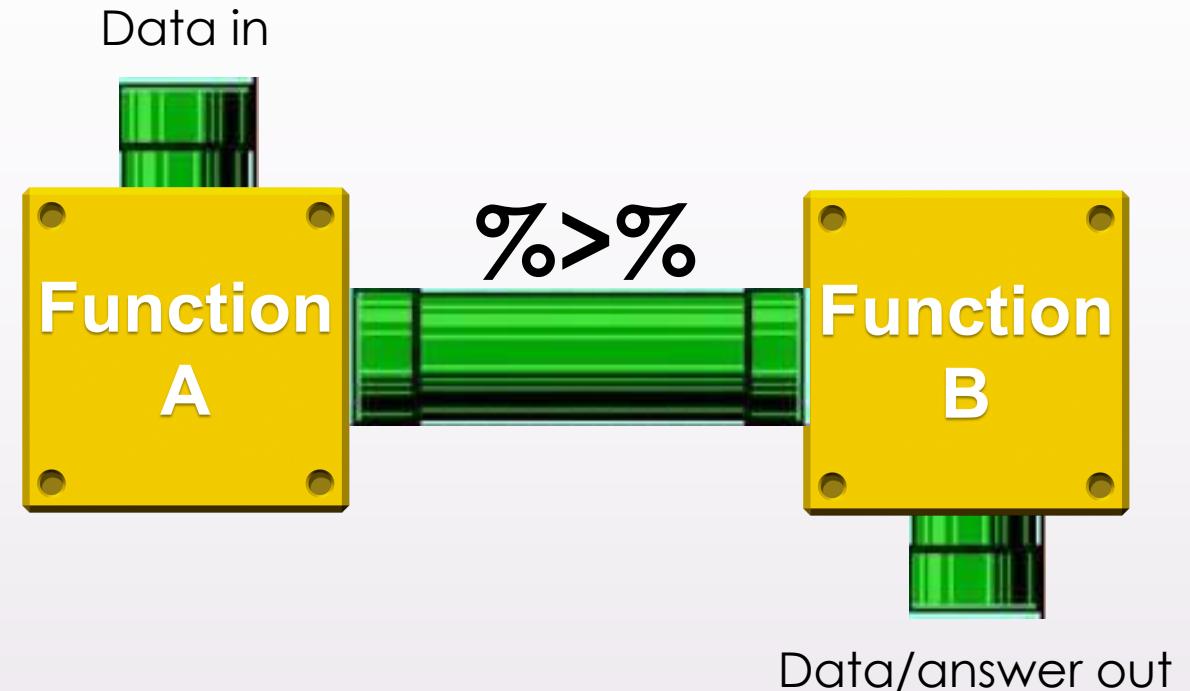


## Factors

- ▶ Suitable for:
  - ▶ Categorical data
- ▶ Useful for:
  - ▶ Data ordering
  - ▶ Data grouping
  - ▶ Renaming data

# Pipes

- Pipes take the output of one function and pass it to the next.



The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.



# dplyr

dplyr is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges:

- ▶ `mutate()` adds new variables that are functions of existing variables
- ▶ `select()` picks variables based on their names.
- ▶ `filter()` picks cases based on their values.
- ▶ `summarise()` reduces multiple values down to a single summary.
- ▶ `arrange()` changes the ordering of the rows.

# tidyverse

## gather()

Wide ~> Long

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K



country	year	cases
A	1999	0.7K
B	1999	37K
C	1999	212K
A	2000	2K
B	2000	80K
C	2000	213K

## Spread()

Long ~> Wide

country	year	type	count
A	1999	cases	0.7K
A	1999	pop	19M
A	2000	cases	2K
A	2000	pop	20M
B	1999	cases	37K
B	1999	pop	172M
B	2000	cases	80K
B	2000	pop	174M
C	1999	cases	212K
C	1999	pop	1T
C	2000	cases	213K
C	2000	pop	1T

country	year	cases	pop
A	1999	0.7K	19M
A	2000	2K	20M
B	1999	37K	172M
B	2000	80K	174M
C	1999	212K	1T
C	2000	213K	1T



Good luck  
and have fun.

If you finish this section take a look at the cheat sheets and explore what else you can do within the tidyverse and feel free to let us know what you find in the wrap up!

Remember to use the teams groups to help each other and reach out to us for help!