THE UNIVERSITY OF
MELBOURNE

Skills for impactful data visualisation

Andrew Perfors

School of Psychological Sciences
Complex Human Data Hub
2022 ACNS ECR Webinar

THE UNIVERSITY OF MELBOURNE

What is data visualisation all about?

THE UNIVERSITY OF
MELBOURNE

What is data visualisation all about?

THE UNIVERSITY OF MELBOURNE

What is data visualisation all about?

What is data visualisation all about?

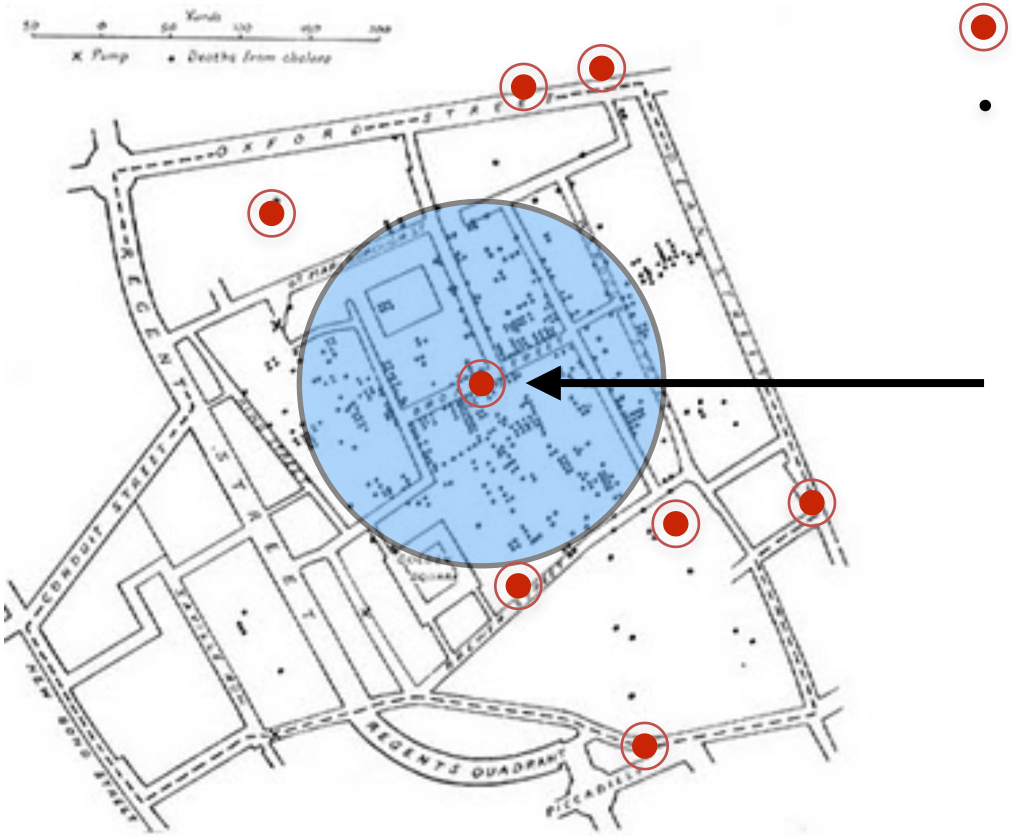
Data visualisation is discovery

Example 1. Stopping a cholera outbreak

Person	Age	Occupation	Family size	Address	Health	Cholera?
Mary Smith	12	child	8	7 Cross St	good	yes
Robert Plank	48	unemployed	5	12 King St	fair	yes
John WIlliams	7	child	12	16 Main St	good	no
Henry Locke	23	dockworker	9	24 King St	poor	yes
Elizabeth Gates	3	child	5	32 Banks St	poor	no
Jane Potter	29	homemaker	7	35 Cross St	fair	no

John Snow, 1854

Data visualisation is discovery



● Water pump

• Cholera case

Remove the handle from this pump

John Snow, 1854

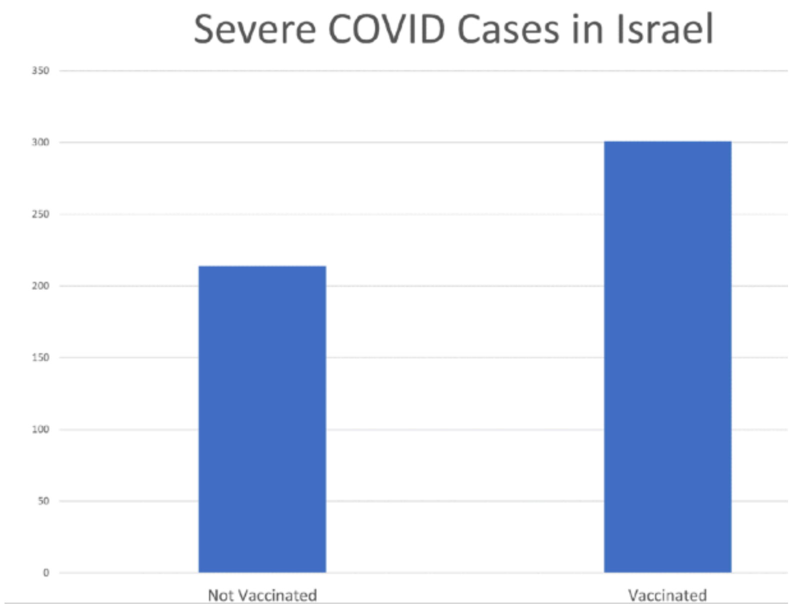
THE UNIVERSITY OF
MELBOURNE

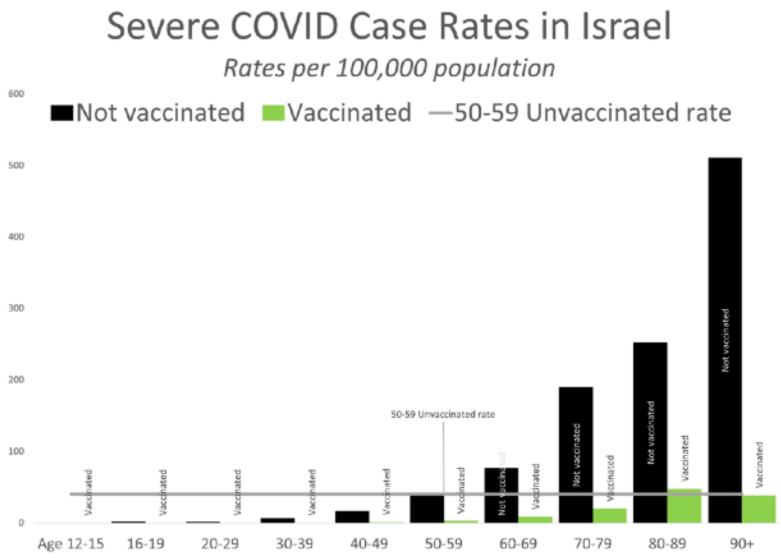
Data visualisation is discovery

Example 2. Covid and vaccination

More vaccinated people at risk of severe covid?

… but not at all when you break it down by age


Severe COVID Cases in Israel


Severe COVID Case Rates in Israel

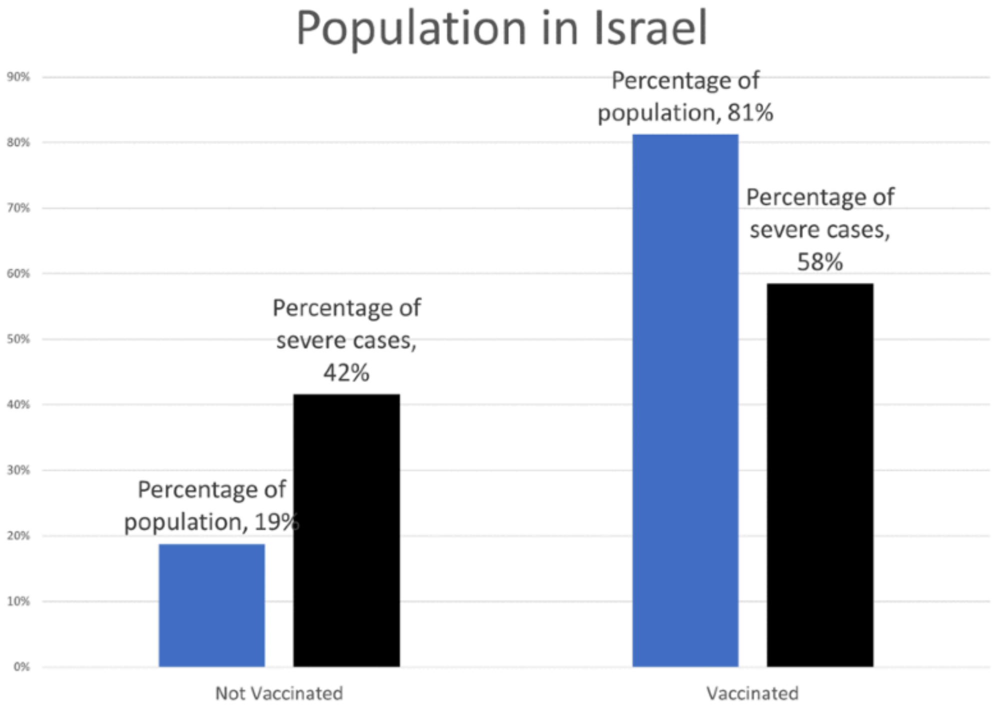
Rates per 100,000 population

Source: https://marypatcampbell.substack.com/p/covid-and-simpsons-paradox-why-so published 19 August 2021, accessed 3 Dec 2021.

Source: https://marypatcampbell.substack.com/p/covid-and-simpsons-paradox-why-so published 19 August 2021, accessed 3 Dec 2021.

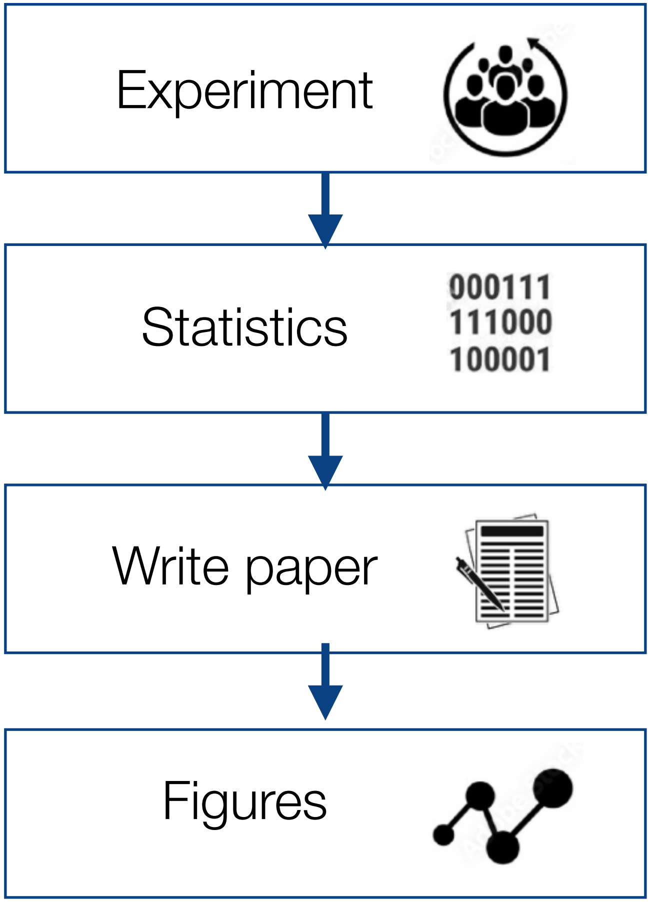
Data visualisation is discovery

More people were vaccinated in the first place
AND
The most vulnerable (i.e., oldest) were more likely to be vaccinated



Data visualisation is discovery

A common workflow



Experiment

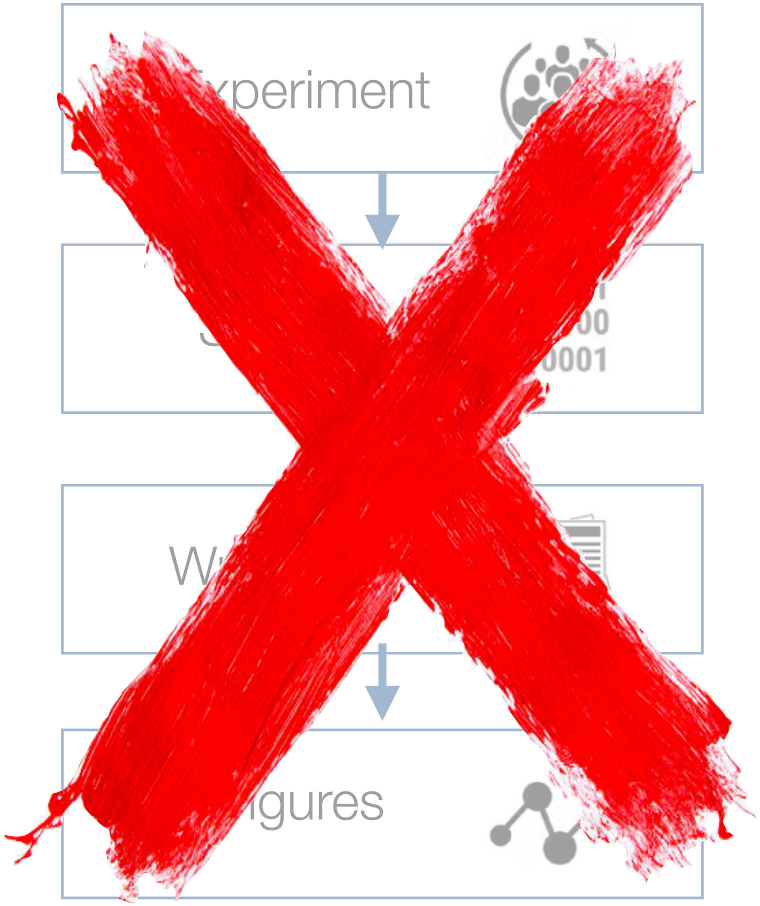
Statistics

Write paper

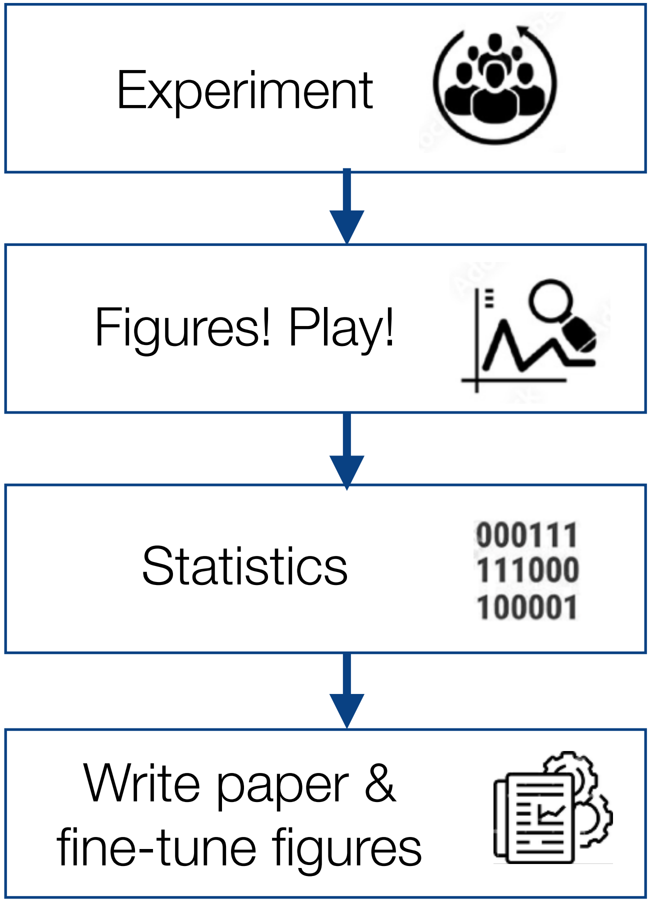
Figures

Data visualisation is discovery

A common workflow



A *good* workflow



Experiment

Figures! Play!

Statistics

000111
111000
100001

Write paper &
fine-tune figures

Skills for good data visualisation

- Technical
  - Good tools combine ease & power
  - In R, ggplot & tidyverse are great
- Active, critical, aesthetic
  - Guided by scientific questions
  - Some tips to get you started

Example: Zombie apocalypse

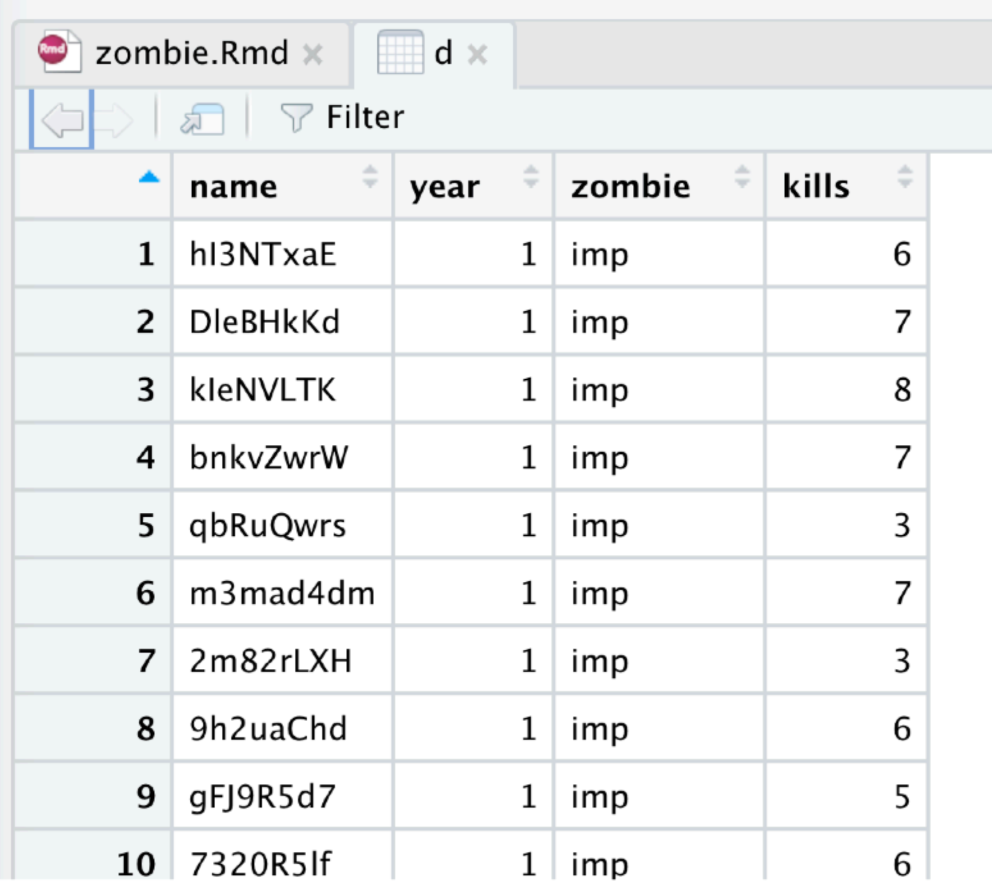
5 years ago,
zombies started
taking over the
world…

How dangerous are they?

Are we making any headway?

Example: Zombie apocalypse

At great peril to our lives, we have started tracking them, and now have five years of data

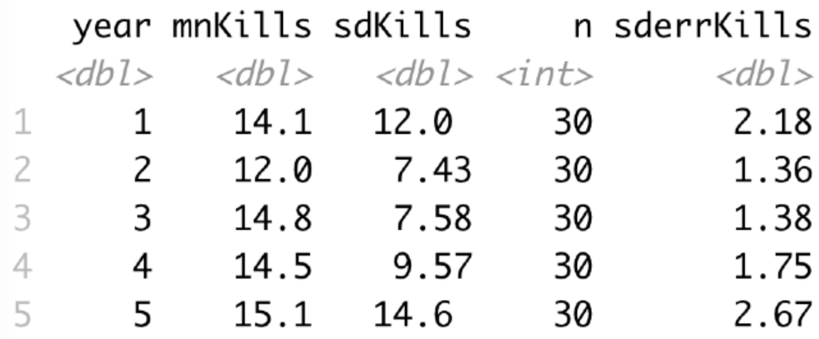


	name	year	zombie	kills
1	hI3NTxaE	1	imp	6
2	DleBHkKd	1	imp	7
3	kIeNVLTK	1	imp	8
4	bnkvZwrW	1	imp	7
5	qbRuQwrs	1	imp	3
6	m3mad4dm	1	imp	7
7	2m82rLXH	1	imp	3
8	9h2uaChd	1	imp	6
9	gFJ9R5d7	1	imp	5
10	7320R5lf	1	imp	6

Technical: R

Tidyverse allows you to easily manipulate data

```
d_sum <- d %>%
  group_by(year) %>%
  summarise(mnKills = mean(kills),
            sdKills = sd(kills),
            n = n(),
            sderrKills = sdKills/sqrt(n)) %>%
  ungroup()
```



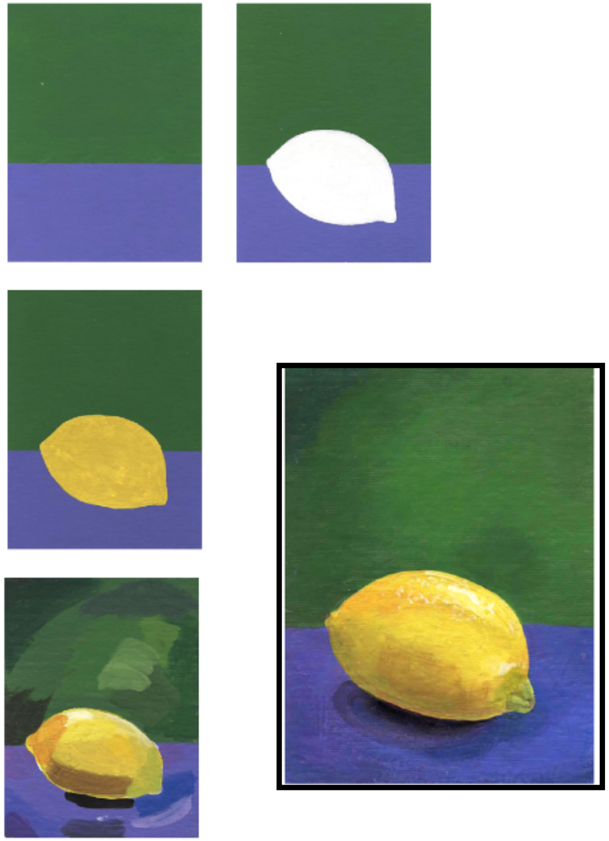
```
   year mnKills sdKills     n sderrKills
  <dbl>   <dbl>   <dbl> <int>      <dbl>
1     1    14.1   12.0     30       2.18
2     2    12.0    7.43    30       1.36
3     3    14.8    7.58    30       1.38
4     4    14.5    9.57    30       1.75
5     5    15.1   14.6     30       2.67
```

R for Data Science: https://r4ds.had.co.nz/

Technical: R

ggplot is a package that lets you draw figures

- A grammar
  - Combine & reuse smaller parts in a structured way
- Of graphics
  - Like a painter
  - Figure is built by layering

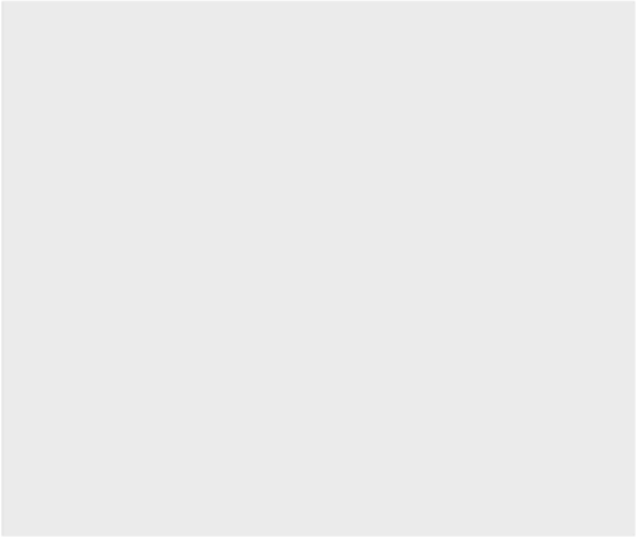


R for Data Science: https://r4ds.had.co.nz/

https://nancyreyner.com/2017/12/25/what-is-layering-for-painting/

Technical: R

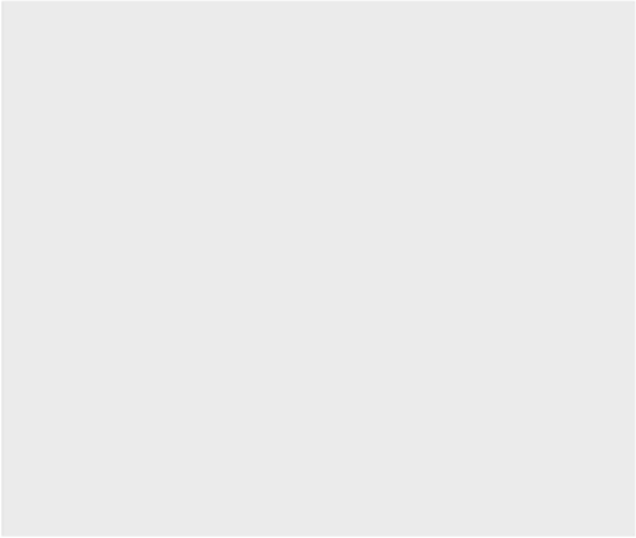
```
ggplot()
```



Sets a blank canvas

Technical: R

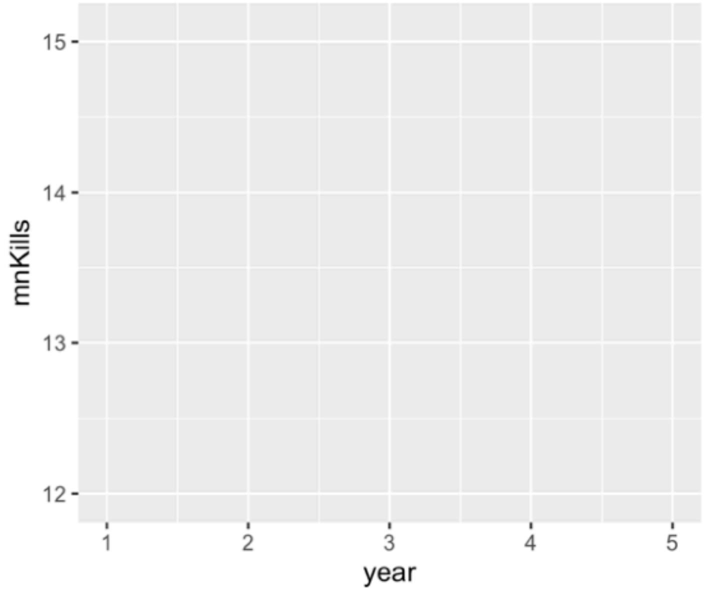
```
d_sum %>%
  ggplot()
```



Specifies the data (but don’t know what to do with it yet)

Technical: R

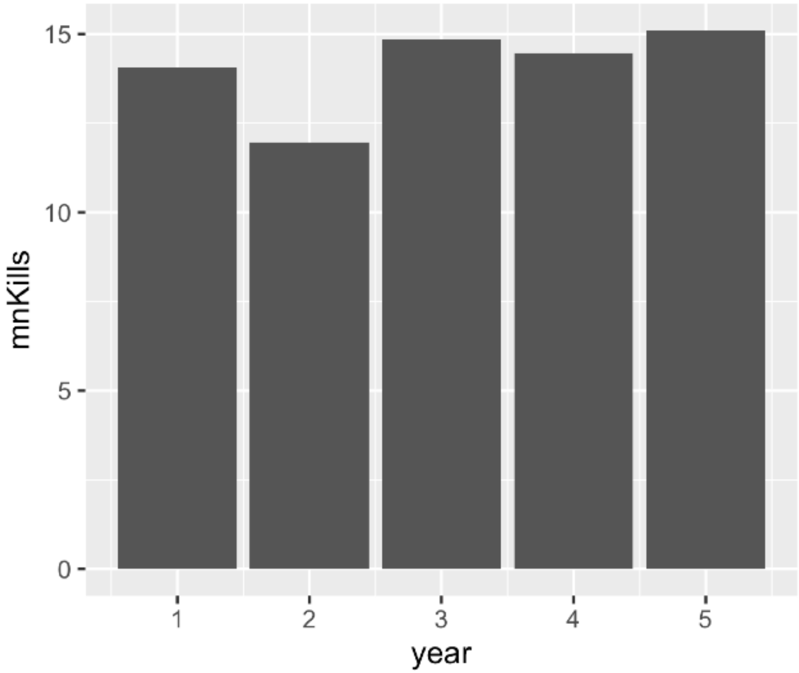
```
d_sum %>%
  ggplot(
    mapping = aes(x = year,
      y = mnKills))
```



Specifies a **mapping** to the plot **aes**thetics (in this case, the x and y axis)

Technical: R

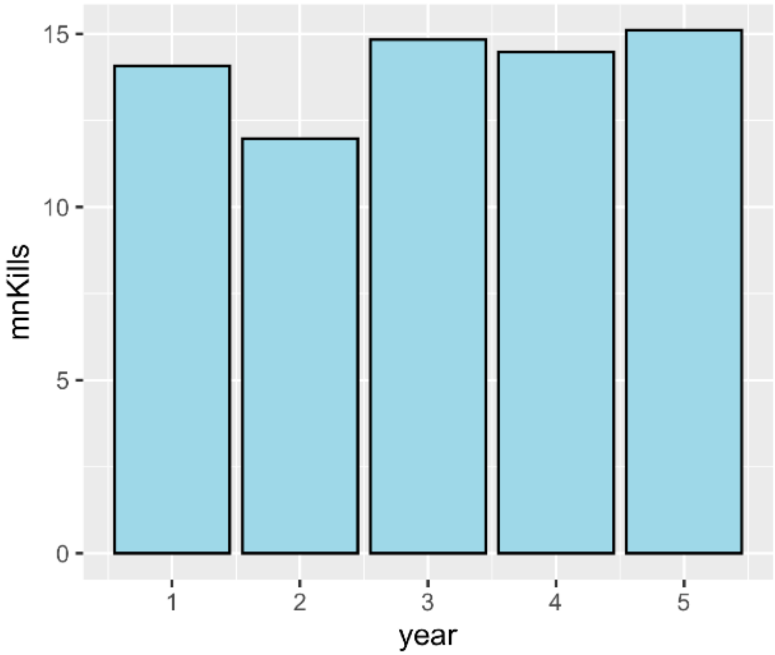
```
d_sum %>%
  ggplot(
    mapping = aes(x = year,
      y = mnKills)) +
  geom_col()
```



Add a **plot** layer (the points, lines, bars, histograms, etc)

Technical: R

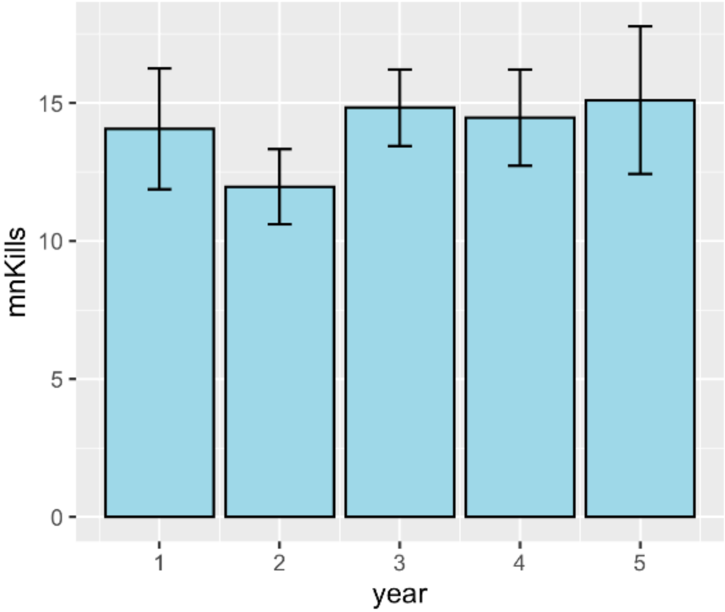
```
d_sum %>%
  ggplot(
    mapping = aes(x = year,
      y = mnKills)) +
  geom_col(colour=“black",
           fill="lightblue")
```



Add aesthetics to the **plot** layer

Technical: R

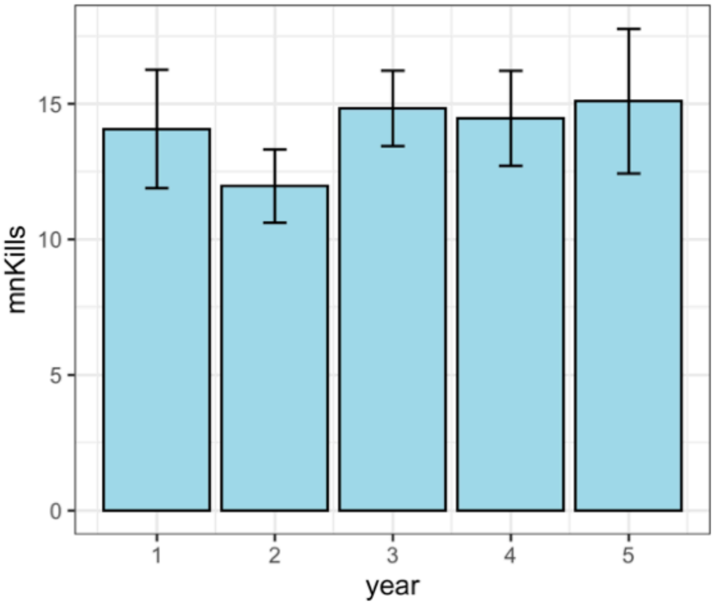
```
d_sum %>%
  ggplot(
    mapping = aes(x = year,
      y = mnKills)) +
  geom_col(colour=“black",
           fill="lightblue") +
  geom_errorbar(
    mapping = aes(ymin = mnKills-sderrKills,
                  ymax = mnKills+sderrKills),
    width=0.2)
```



Add another plot layer with its own mapping and aesthetics

Technical: R

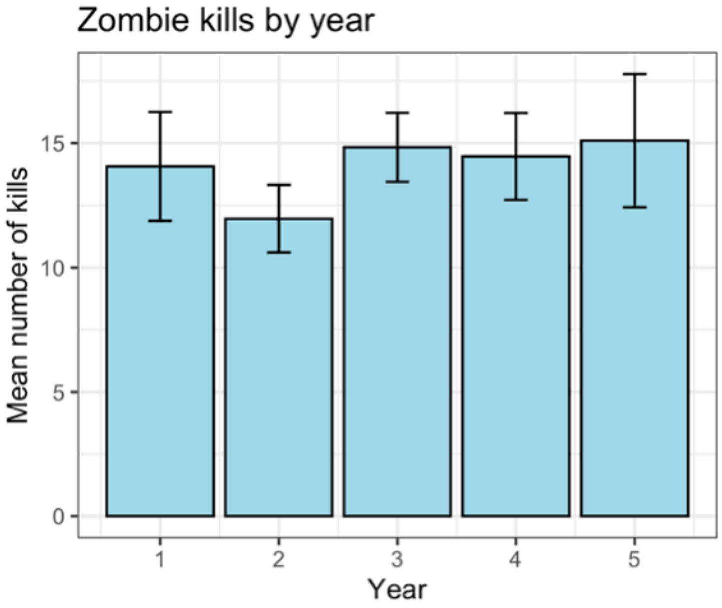
```
d_sum %>%
  ggplot(
    mapping = aes(x = year,
      y = mnKills)) +
  geom_col(colour=“black",
           fill="lightblue") +
  geom_errorbar(
    mapping = aes(ymin = mnKills-sderrKills,
                 ymax = mnKills+sderrKills),
    width=0.2) +
  theme_bw()
```



Modify the theme to make it look nicer…

Technical: R

```
d_sum %>%
  ggplot(
    mapping = aes(x = year,
      y = mnKills)) +
  geom_col(colour=“black",
           fill="lightblue") +
  geom_errorbar(
    mapping = aes(ymin = mnKills-sderrKills,
                  ymax = mnKills+sderrKills),
    width=0.2) +
  theme_bw() +
  labs(title = "Zombie kills by year",
    x = "Year",
    y = "Mean number of kills")
```



Add title and labels

Skills for good data visualisation

- Technical
  - Good tools combine ease & power
  - In R, ggplot & tidyverse are great
- Active, critical, aesthetic
  - Guided by scientific questions
  - Some tips to get you started

Critical & aesthetic tips

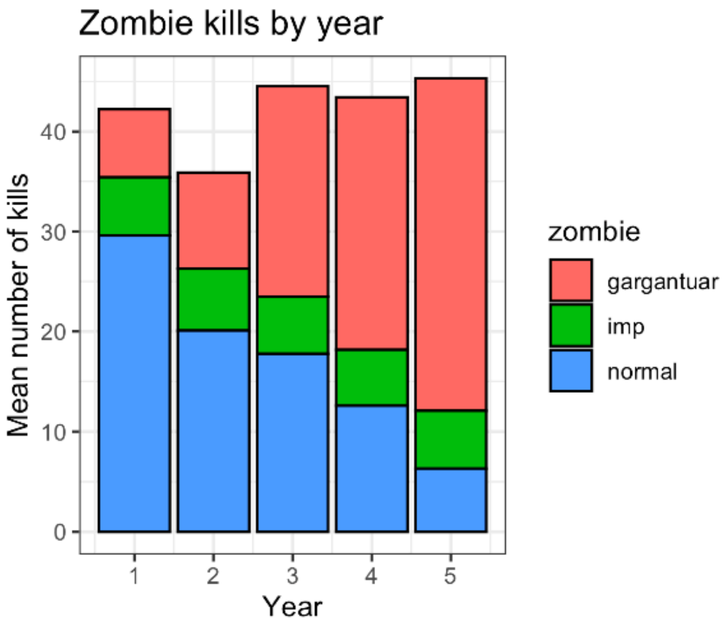
- Break down your data: don’t just summarise!

```
d_sum2 <- d %>%
  group_by(year,zombie) %>%
  summarise(mnKills = mean(kills),
            sdKills = sd(kills),
            n = n(),
            sderrKills = sdKills/sqrt(n)) %>%
  ungroup()
```

Critical & aesthetic tips

- Aesthetic choices should visualise important things

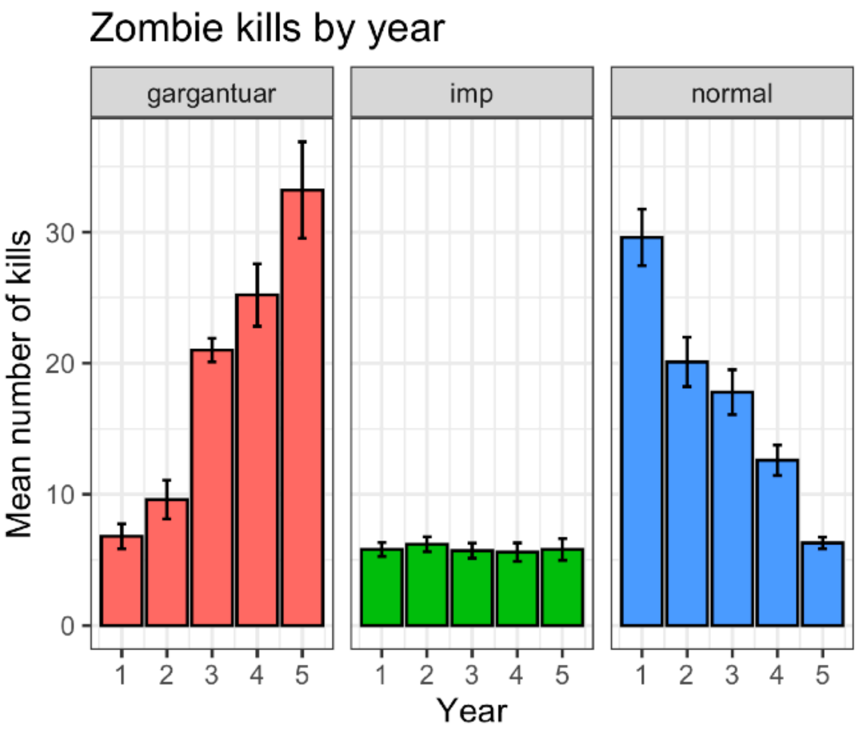
```
d_sum2 %>%
  ggplot(mapping = aes(x = year,
      y = mnKills,
      fill = zombie)) +
  geom_col(colour="black") +
  theme_bw() +
  labs(title = "Zombie kills by year",
    x = "Year",
    y = "Mean number of kills")
```



Critical & aesthetic tips

- Faceting (making multiple panels) is GREAT!

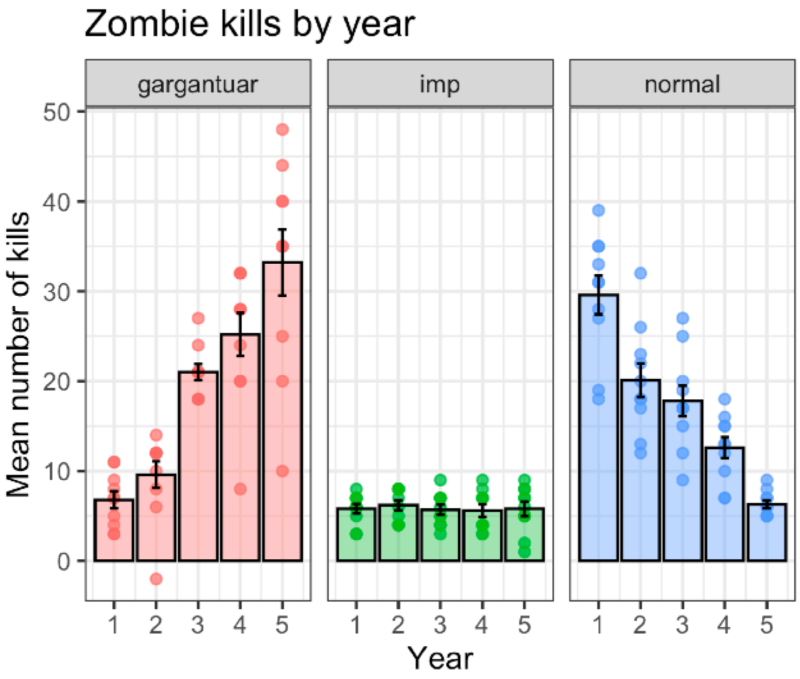
```
d_sum2 %>%
  ggplot(mapping = aes(x = year,
      y = mnKills,
      fill = zombie)) +
  geom_col(colour="black",
           show.legend=FALSE) +
  geom_errorbar(
    mapping = aes(ymin = mnKills-sderrKills,
                 ymax = mnKills+sderrKills),
    width=0.2) +
  facet_wrap(~zombie) +
  theme_bw() +
  labs(title = "Zombie kills by year",
    x = "Year",
    y = "Mean number of kills")
```



Critical & aesthetic tips

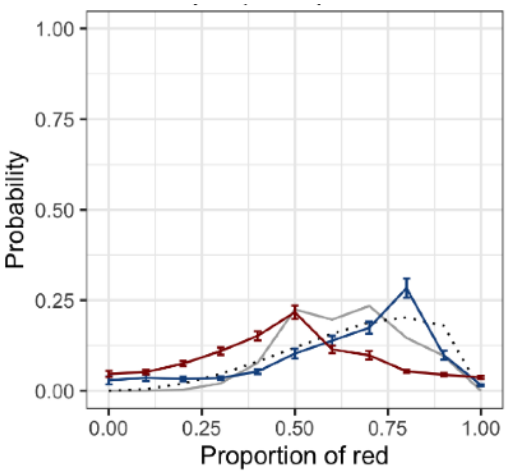
- Show distributions, not just summary stats

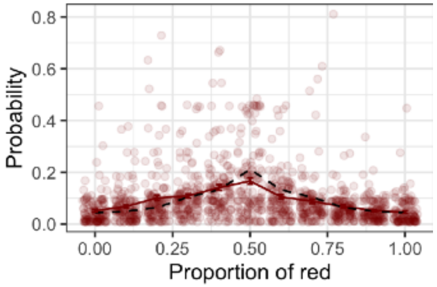
```
d_sum2 %>%
  ggplot(mapping = aes(x = year,
      y = mnKills,
      fill = zombie)) +
  geom_point(data=d,
             mapping=aes(x=year,y=kills,
                         colour=zombie),
             alpha=0.7,show.legend=FALSE) +
  geom_col(colour="black",
           alpha=0.4,
           show.legend=FALSE) +
  geom_errorbar(
    mapping = aes(ymin = mnKills-sderrKills,
                  ymax = mnKills+sderrKills),
    width=0.2) +
  facet_wrap(~zombie) +
  theme_bw() +
  labs(title = "Zombie kills by year",
    x = "Year",
    y = "Mean number of kills")
```

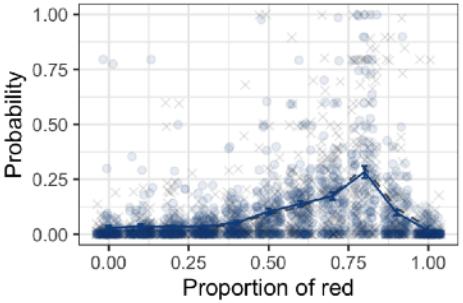


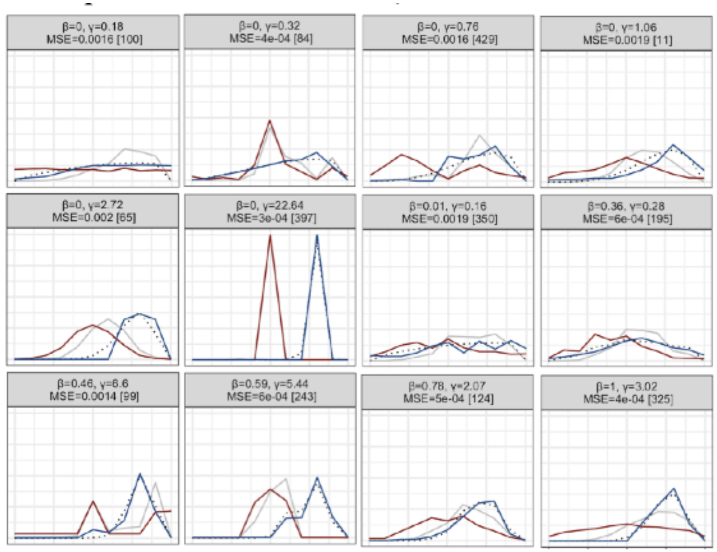
Critical & aesthetic tips

- Look at individuals, not just the aggregate



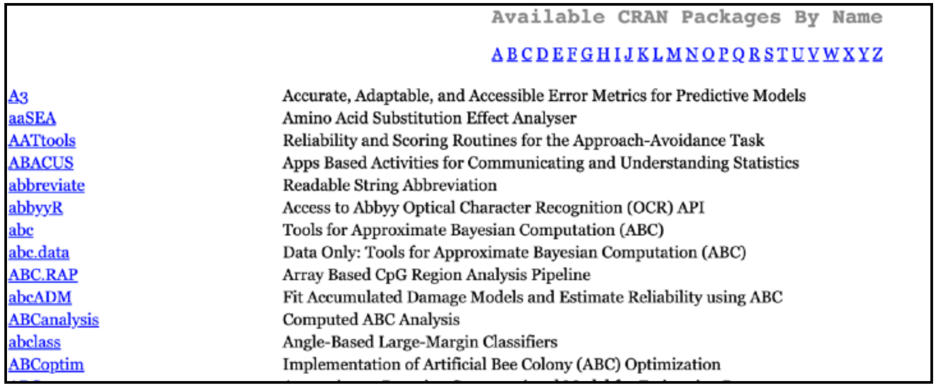






Critical & aesthetic tips

- Use the tools at your disposal to get ideas and look at things in multiple ways



Available CRAN Packages By Name

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

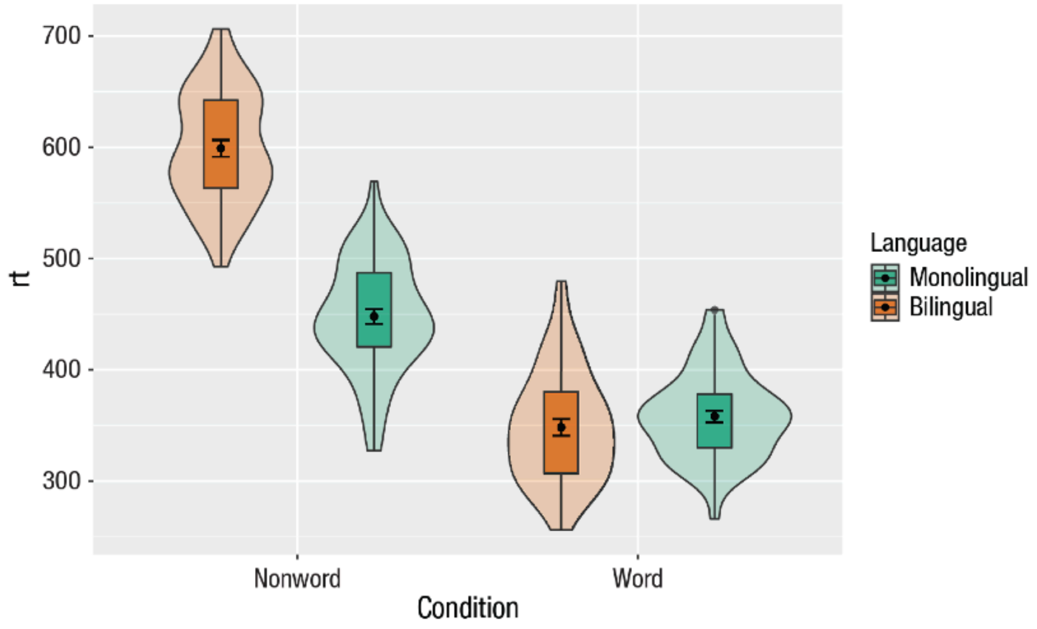
A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
aaSEA	Amino Acid Substitution Effect Analyser
AATtools	Reliability and Scoring Routines for the Approach-Avoidance Task
ABACUS	Apps Based Activities for Communicating and Understanding Statistics
abbreviate	Readable String Abbreviation
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
ABC.RAP	Array Based CpG Region Analysis Pipeline
abcADM	Fit Accumulated Damage Models and Estimate Reliability using ABC
ABCanalysis	Computed ABC Analysis
abclass	Angle-Based Large-Margin Classifiers
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization

R contains many packages (always being added) — you don't need to reinvent the wheel!

Nordmann E, McAleer P, Toivo W, Paterson H, DeBruine LM. Data Visualization Using R for Researchers Who Do Not Use R. *Advances in Methods and Practices in Psychological Science*. April 2022. doi:10.1177/25152459221074654

Critical & aesthetic tips

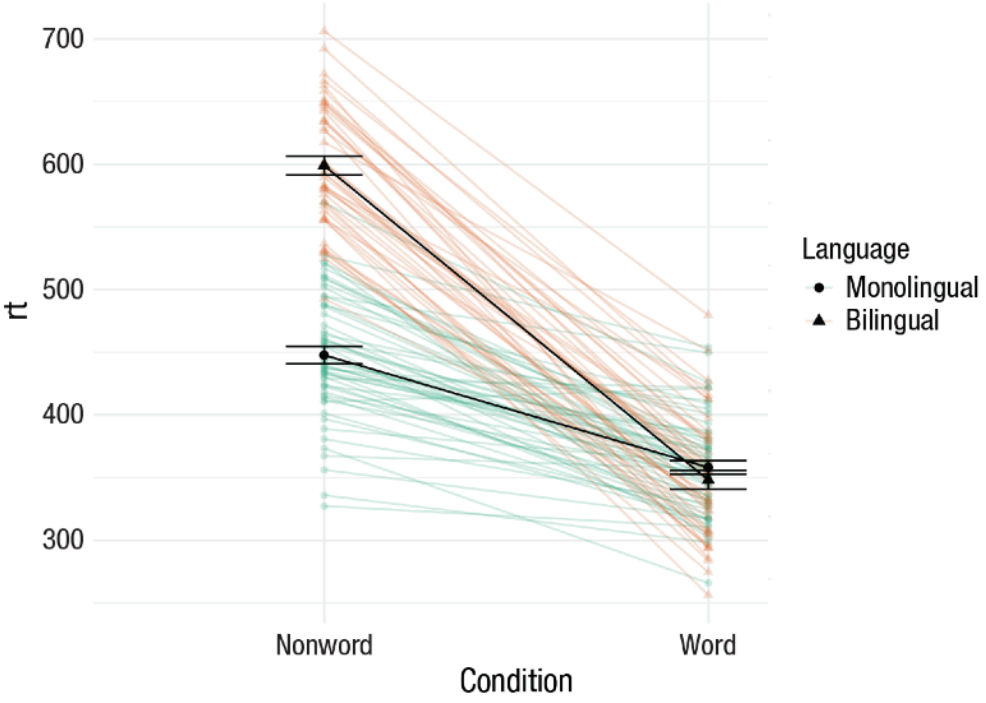
- Use the tools at your disposal to get ideas and look at things in multiple ways



Violin plot +
Bar plot +
Means with errors +
Sensible grouping

Critical & aesthetic tips

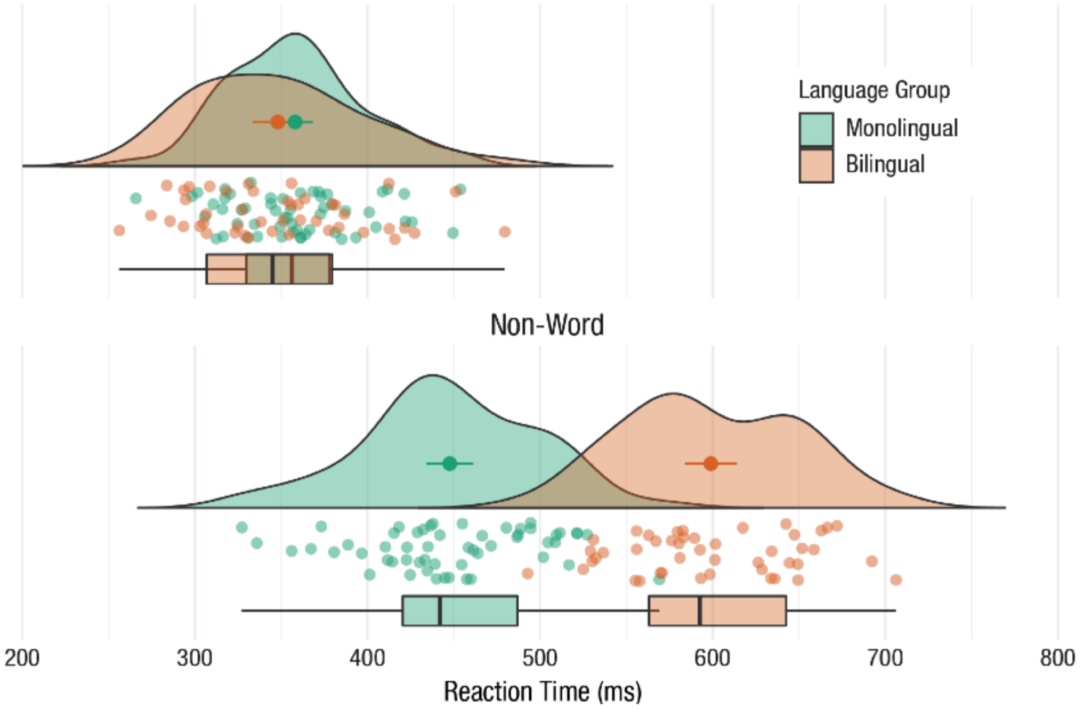
- Use the tools at your disposal to get ideas and look at things in multiple ways



Interaction plots are easy, including by participant

Critical & aesthetic tips

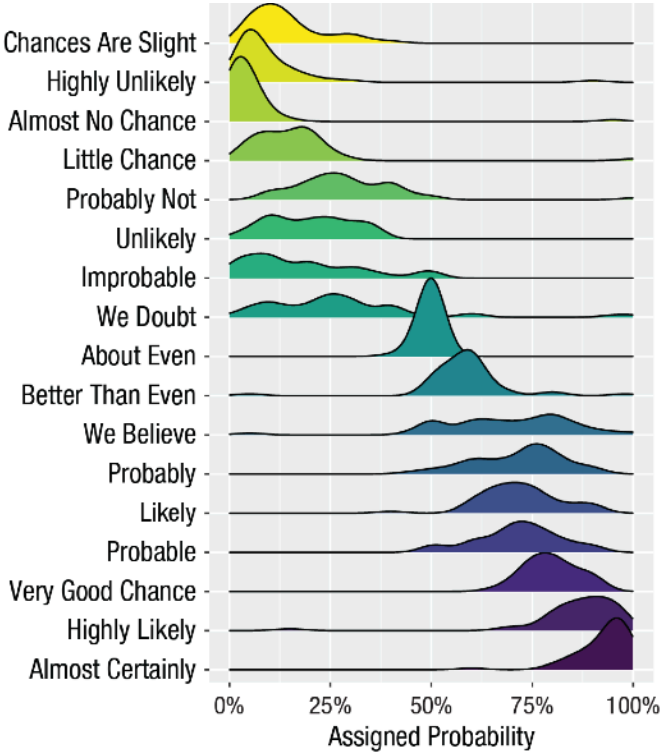
- Use the tools at your disposal to get ideas and look at things in multiple ways



Raincloud plots are a beautiful way of viewing distributions and overlap

Critical & aesthetic tips

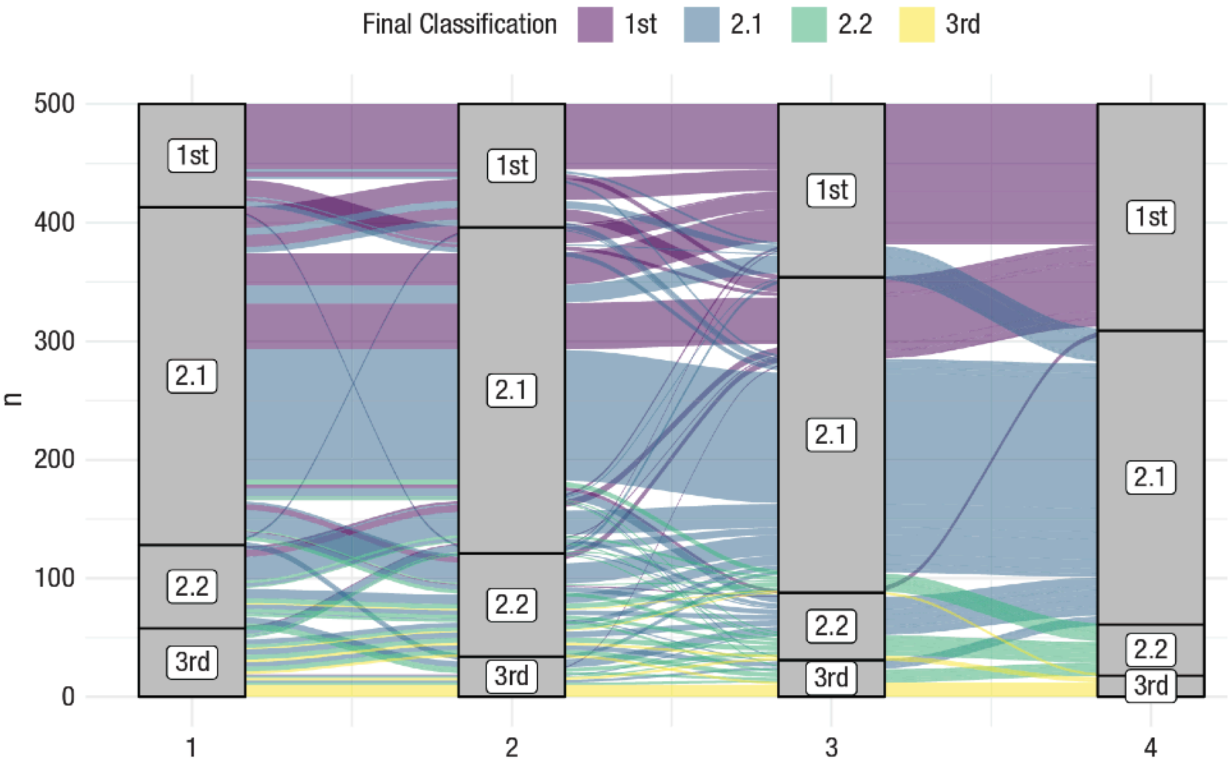
- Use the tools at your disposal to get ideas and look at things in multiple ways



Ridge plots are good when you have a lot of distributions you want to compare

Critical & aesthetic tips

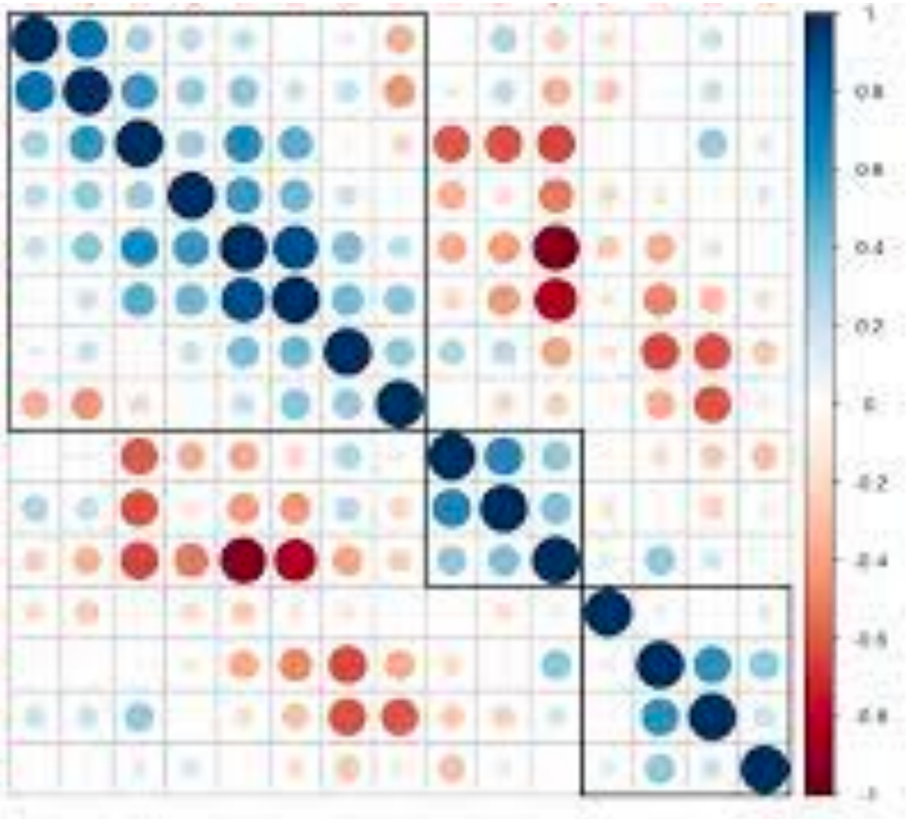
- Use the tools at your disposal to get ideas and look at things in multiple ways



Alluvial plots let you look at change over time

Critical & aesthetic tips

- Use the tools at your disposal to get ideas and look at things in multiple ways

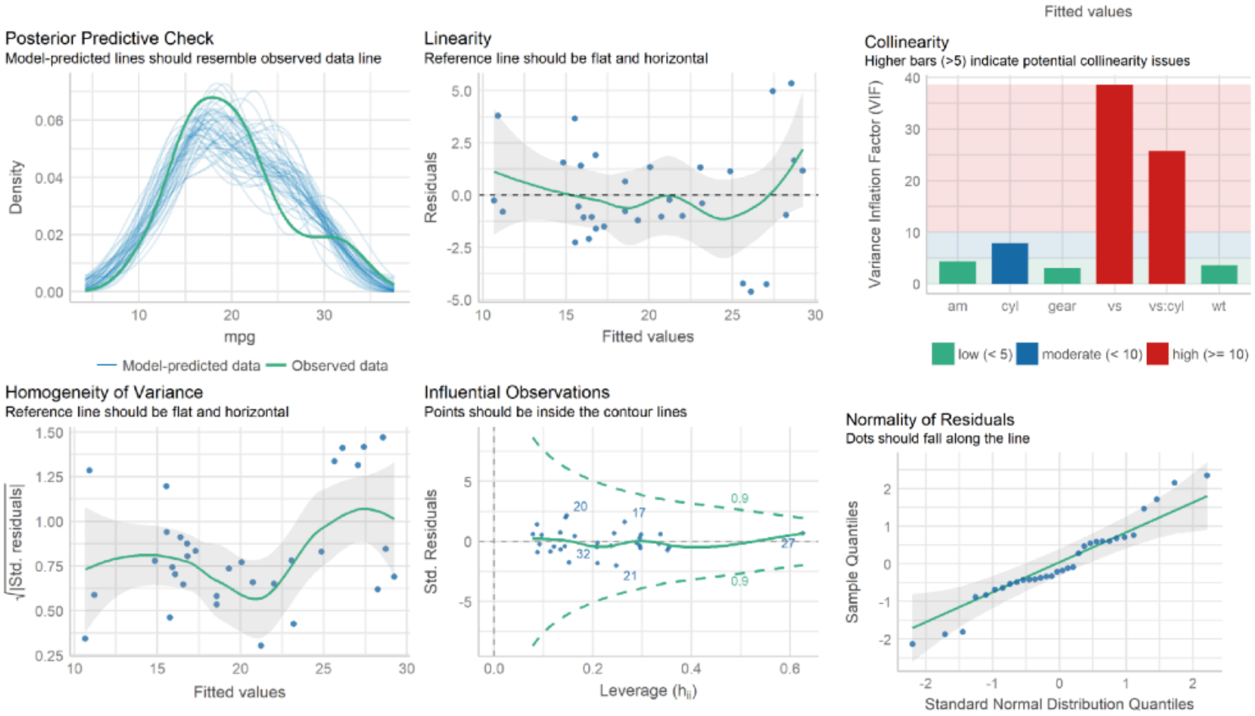


Heatmaps are often way better than correlation tables for identifying patterns

`corrplot`

Critical & aesthetic tips

- Figures are even useful when you're just doing assumption checks and diagnostics!



Posterior Predictive Check

Model-predicted lines should resemble observed data line

Linearity

Reference line should be flat and horizontal

Collinearity

Higher bars (>5) indicate potential collinearity issues

Homogeneity of Variance

Reference line should be flat and horizontal

Influential Observations

Points should be inside the contour lines

Normality of Residuals

Dots should fall along the line

```
performance::compare_performance()
```

Take-home points

- Data visualisation isn’t just for communication, it’s an essential part of the discovery process
- Do lots of things, lots of ways
- Have fun!