

More effective data visualisation

Visualisation for data exploration

Dr Nicola Rennie

Introduction

Data visualisation specialist, mainly working with R, Python, and D3.



Background in statistics, operational research, and data science consultancy.

Co-author of Royal Statistical Society's *Best Practices for Data Visualisation* guidance.

Outline

This session will include:

- Understand why data visualisation is necessary, and what it can be used for.
- Explore common types of visualisation and how they can be leveraged to uncover relationships in data.
- Be able to adjust default settings of charts to get a more truthful representation of your data.

Why visualise data?

Why visualise data?

Data visualisation has two main purposes:

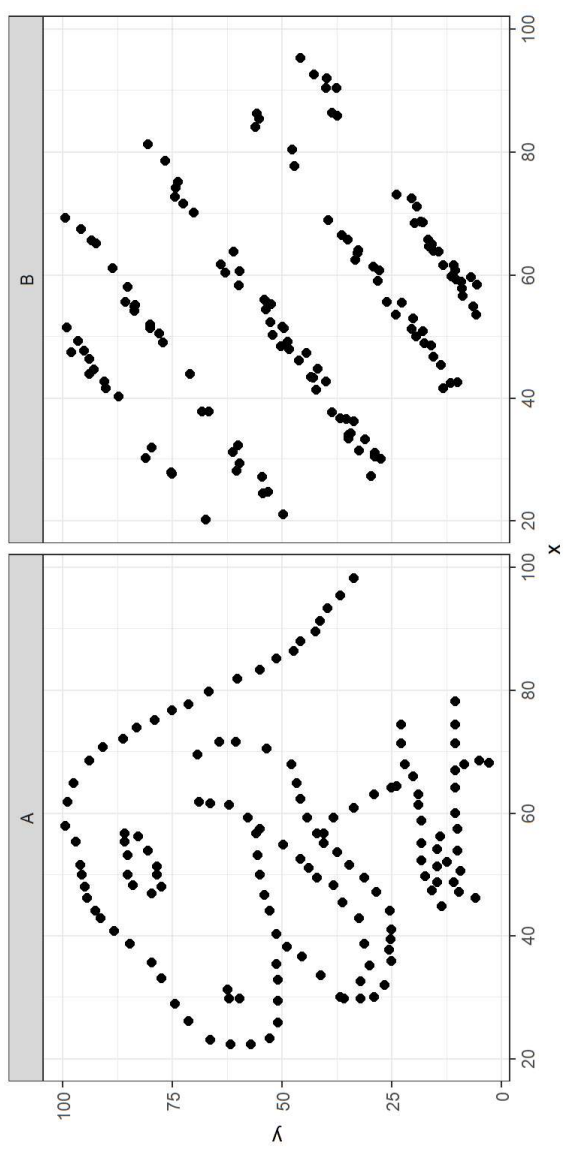
- Exploratory data analysis and identifying data issues
- Communicating insights and results



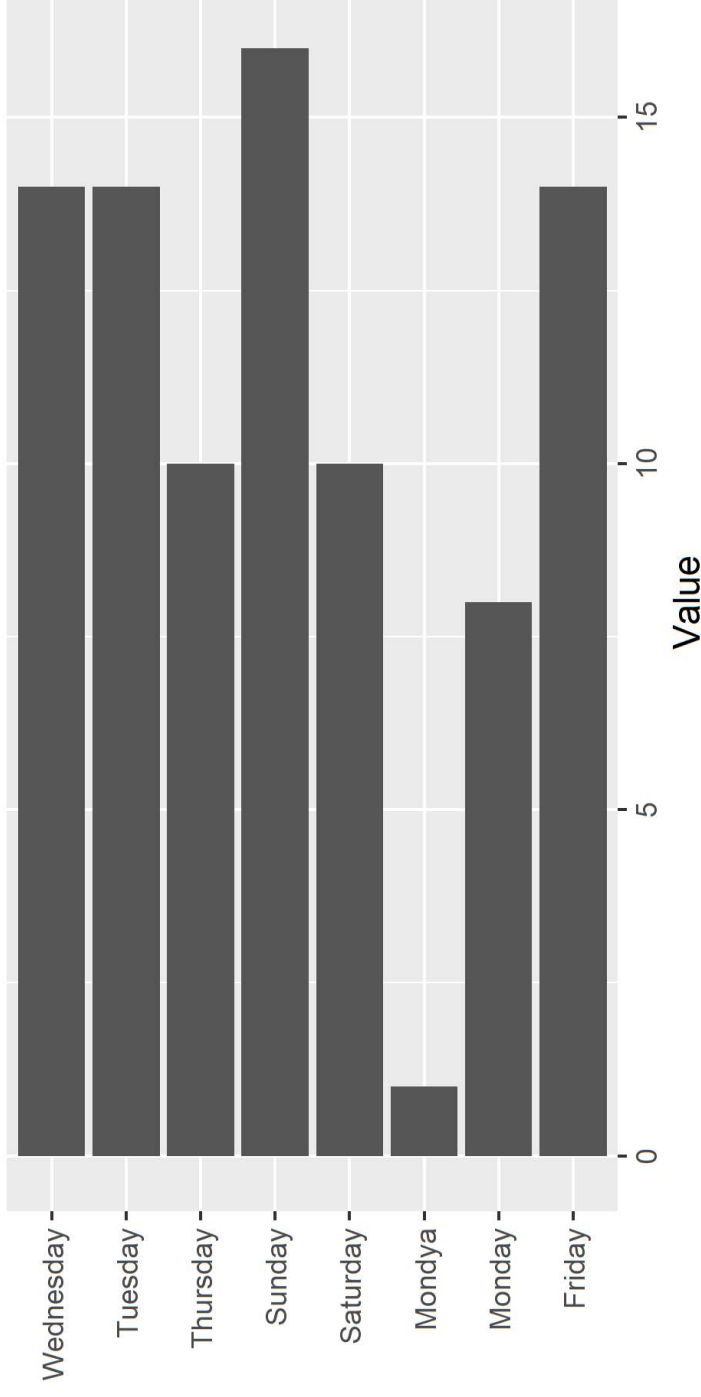
Exploratory data visualisation

Because summary statistics aren't enough...

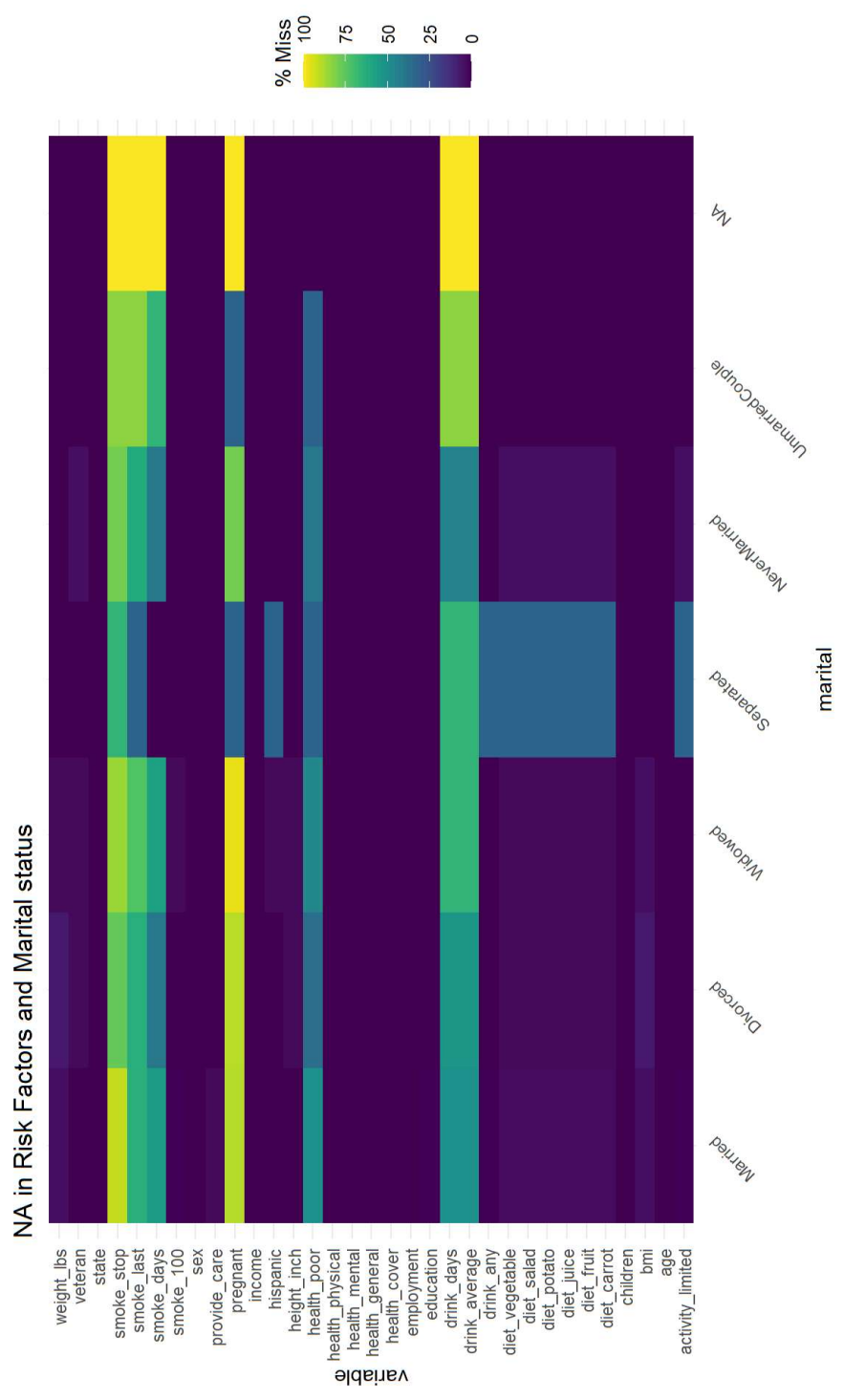
Dataset	A		B	
mean_x	54.2632732	54.2658818		
mean_y	47.8322528	47.8314957		
sd_x	16.7651420	16.7688527		
sd_y	26.9354035	26.9386081		
cor_xy	-0.0644719	-0.0686092		



Identifying issues with data



Understanding missing data

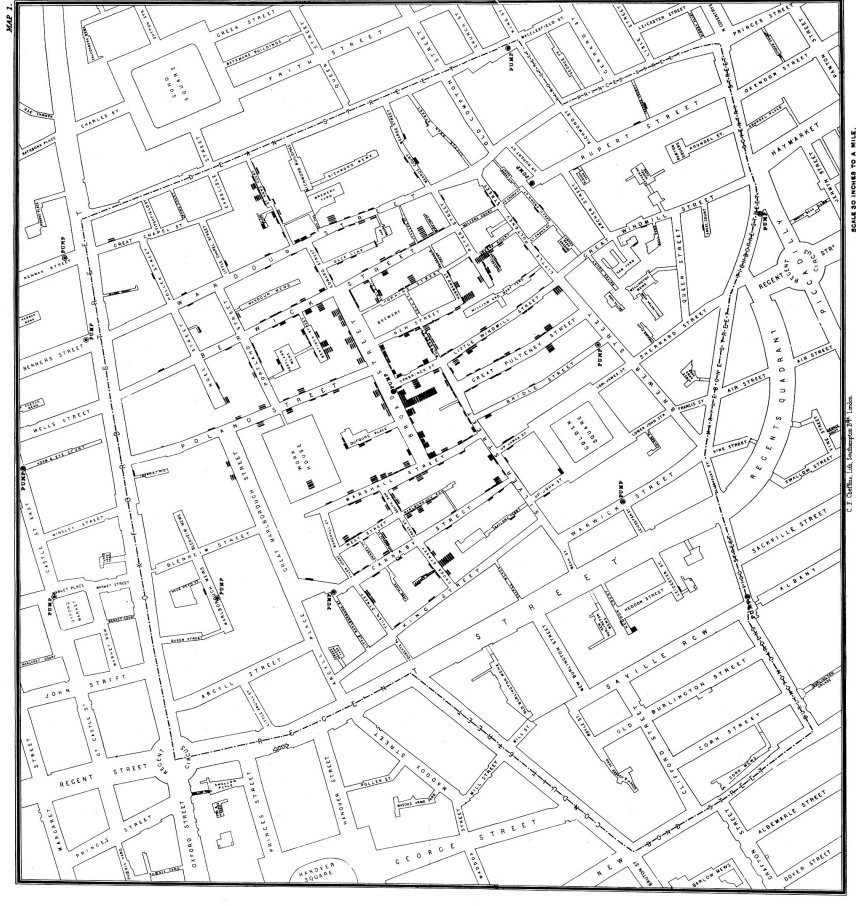


Exploring patterns in data

John Snow collected data on cholera deaths and created a visualisation where the number of deaths was represented by the height of a bar at the corresponding address in London.

This visualisation showed that the deaths clustered around Broad Street, which helped identify the cause of the cholera transmission, the Broad Street water pump.

Snow. 1854.



Choosing a visualisation type

A note about terminology



A note about terminology

Is it a **plot**? Is it a **chart**? Is it a **graph**? Or is it an **infographic**?

- **Plot:** Raw data e.g. scatter plot.
- **Chart:** Plot plus assumptions or interpretation e.g. scatter plot with line of best fit.
- **Graph:** Mathematical representation of a network.
- **Infographic:** Complex or combinations of chart that take longer to understand.

What are you trying to understand?

Data visualisations must serve a purpose.

Ask yourself:

- What is the purpose?
- Does the visualisation support the purpose?
- Is it quick, accurate, and intuitive?

A frequent aim: comparison.

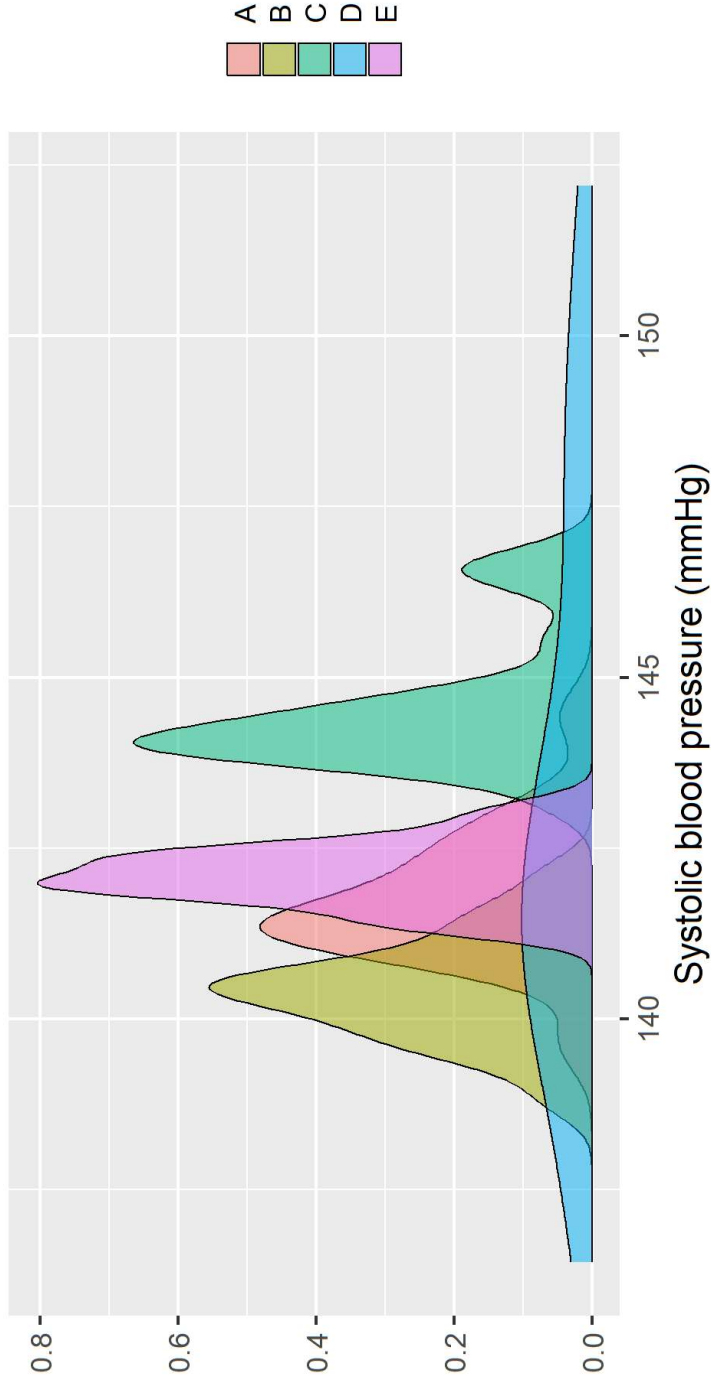
Choosing a chart type

What is the purpose of the chart?

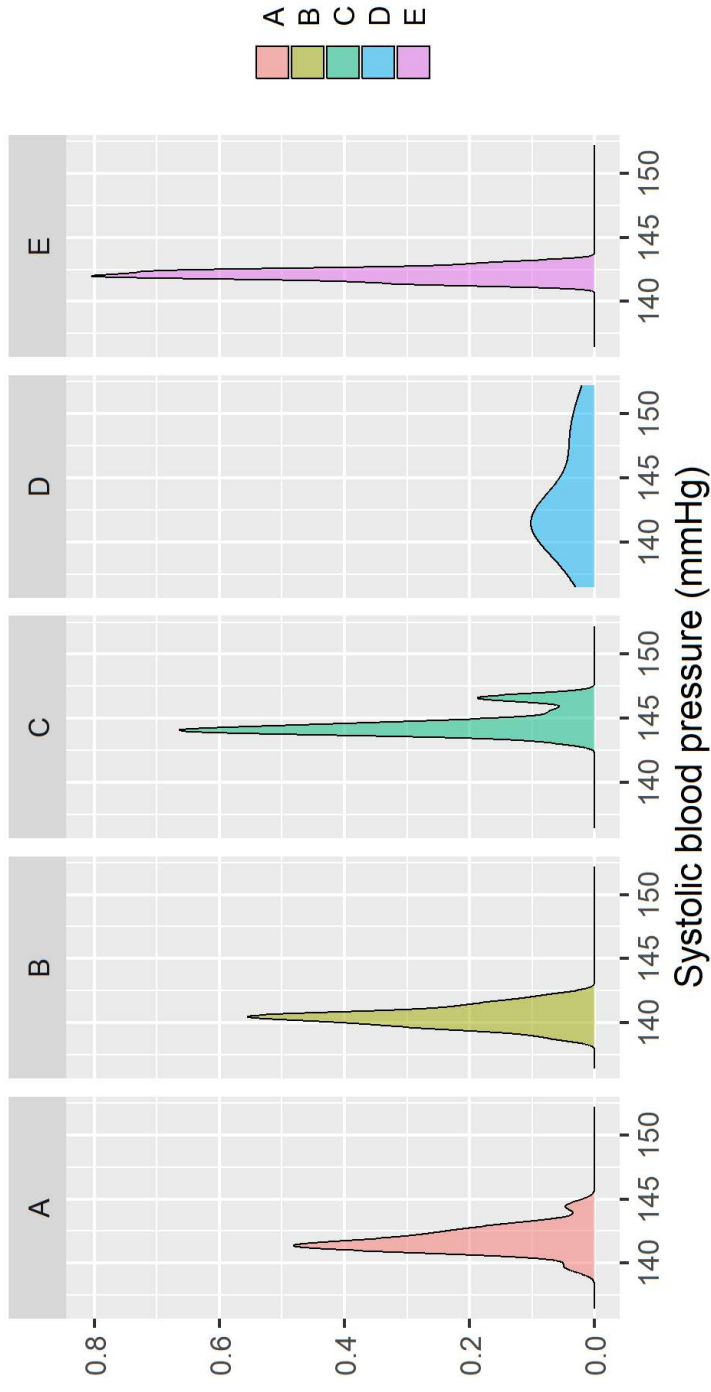
Try: ft-interactive.github.io/visual-vocabulary



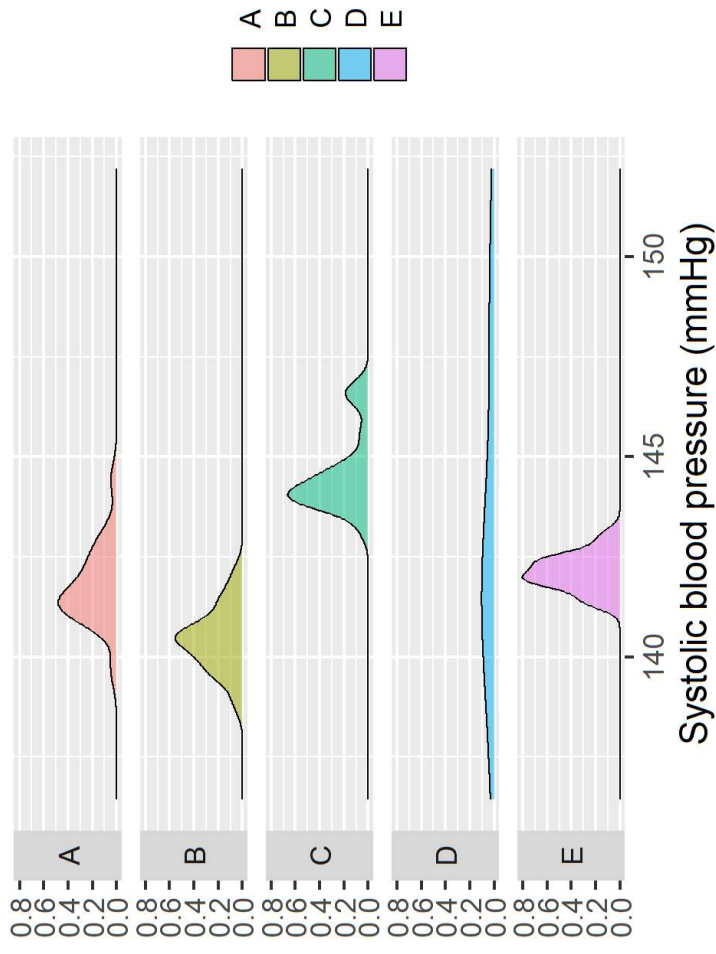
Comparing distributions for different groups



Comparisons with small multiples

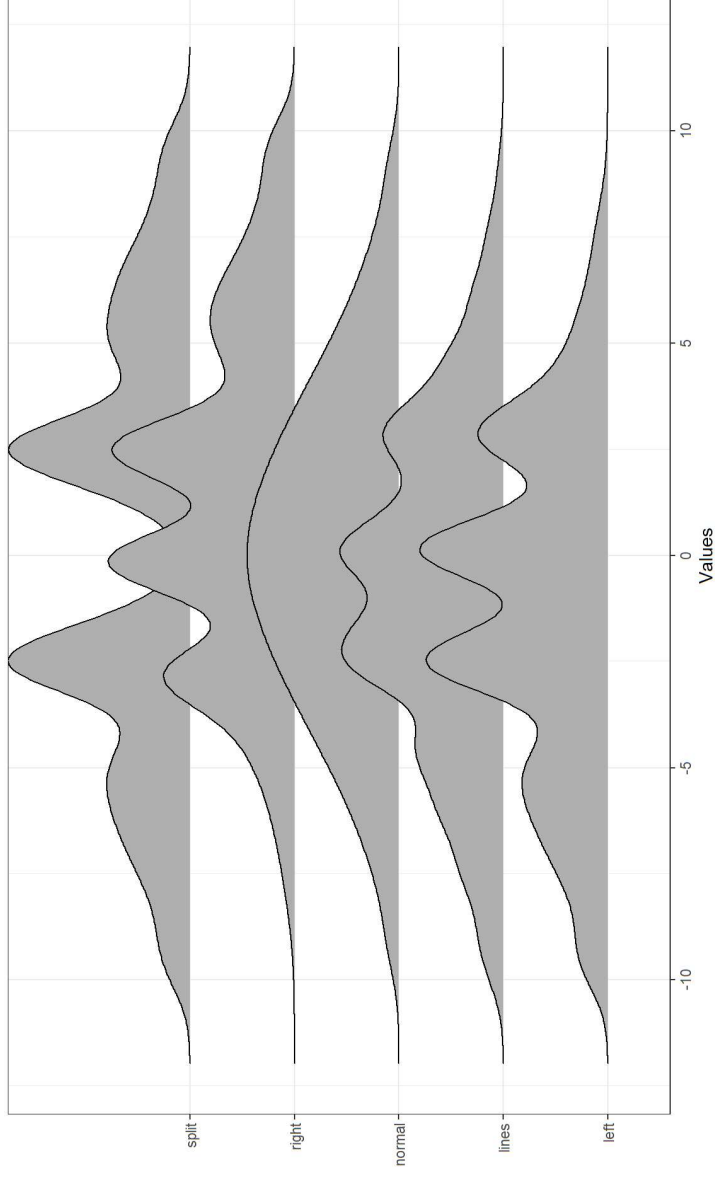
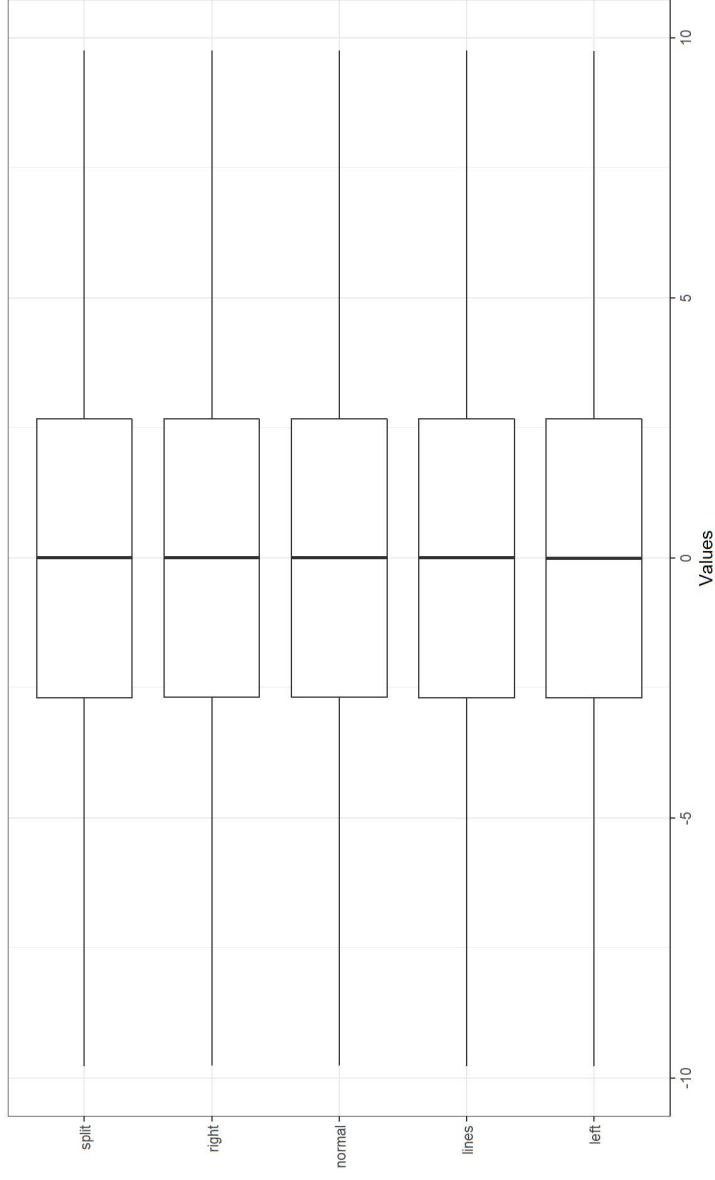


What are you comparing?



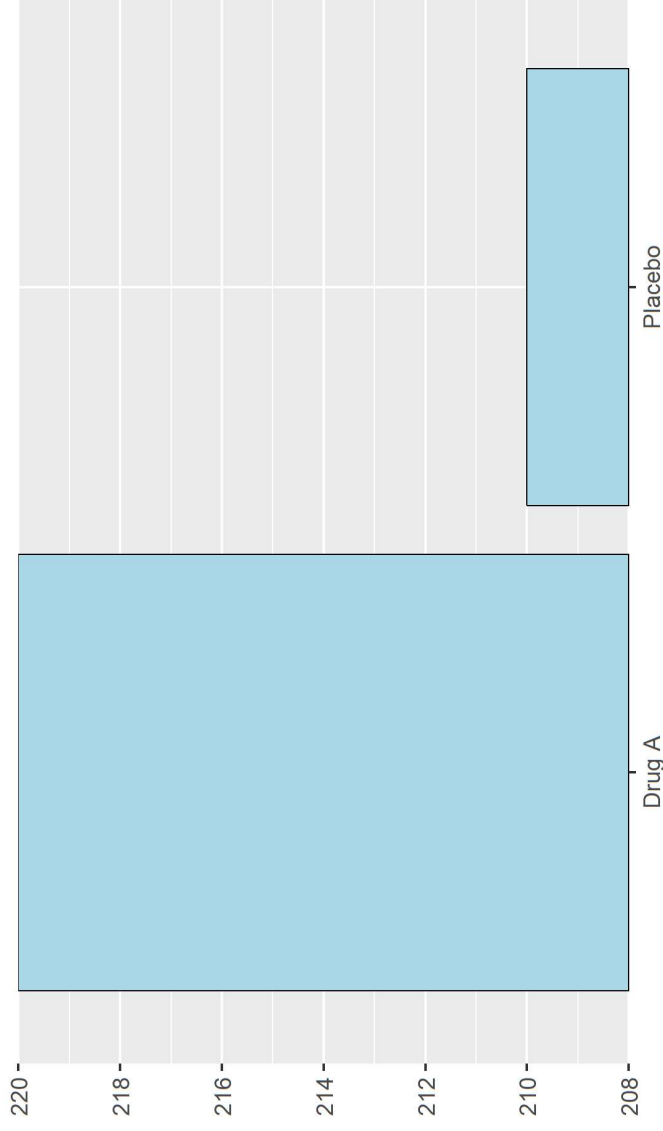
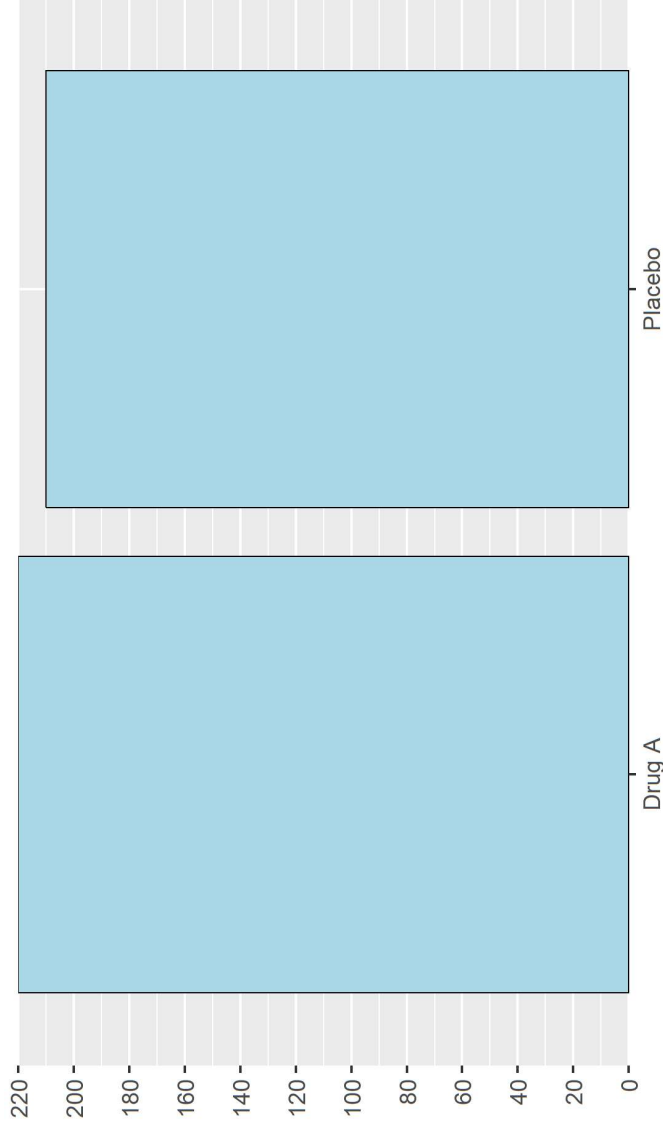
Box plots are just summary statistics in disguise...

Box plots hide information.



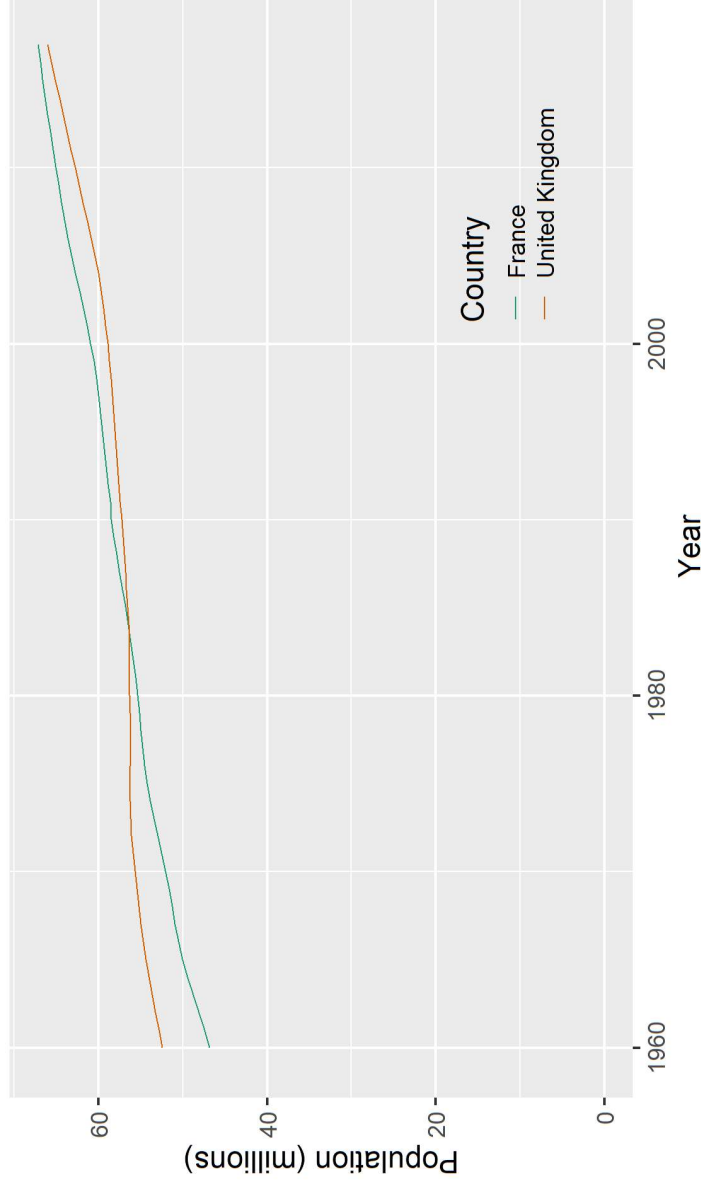
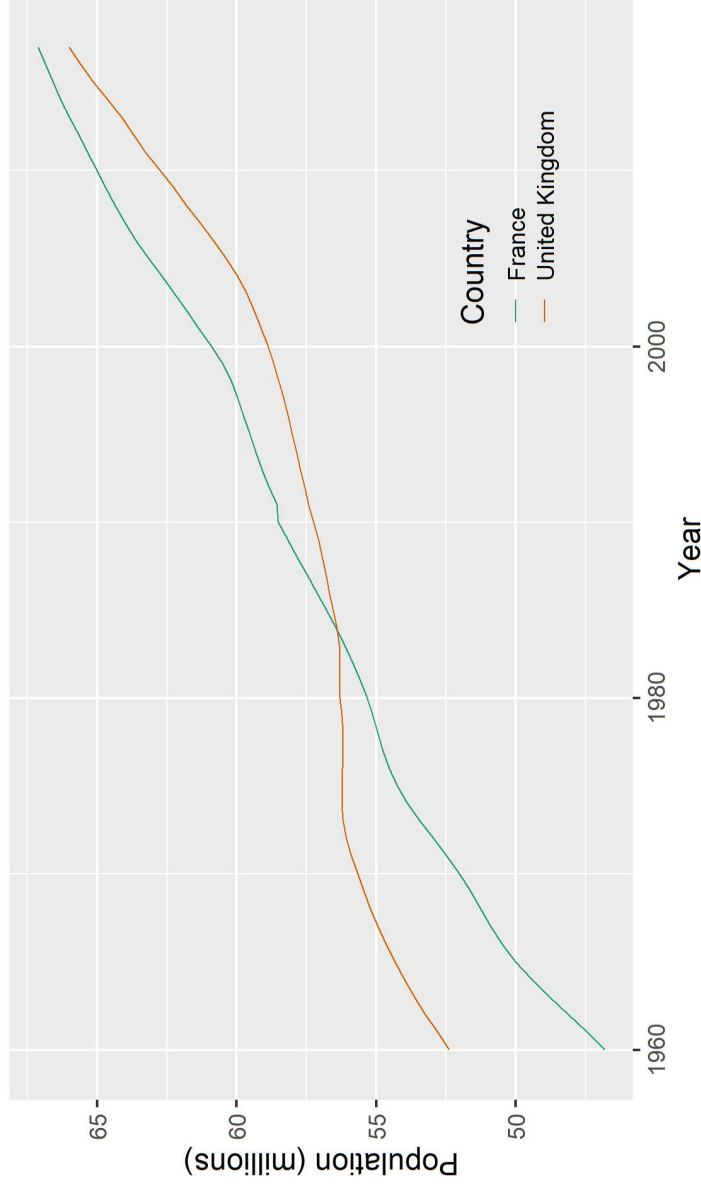
Layouts, aspect ratios, and axes

Should the axes start at 0?

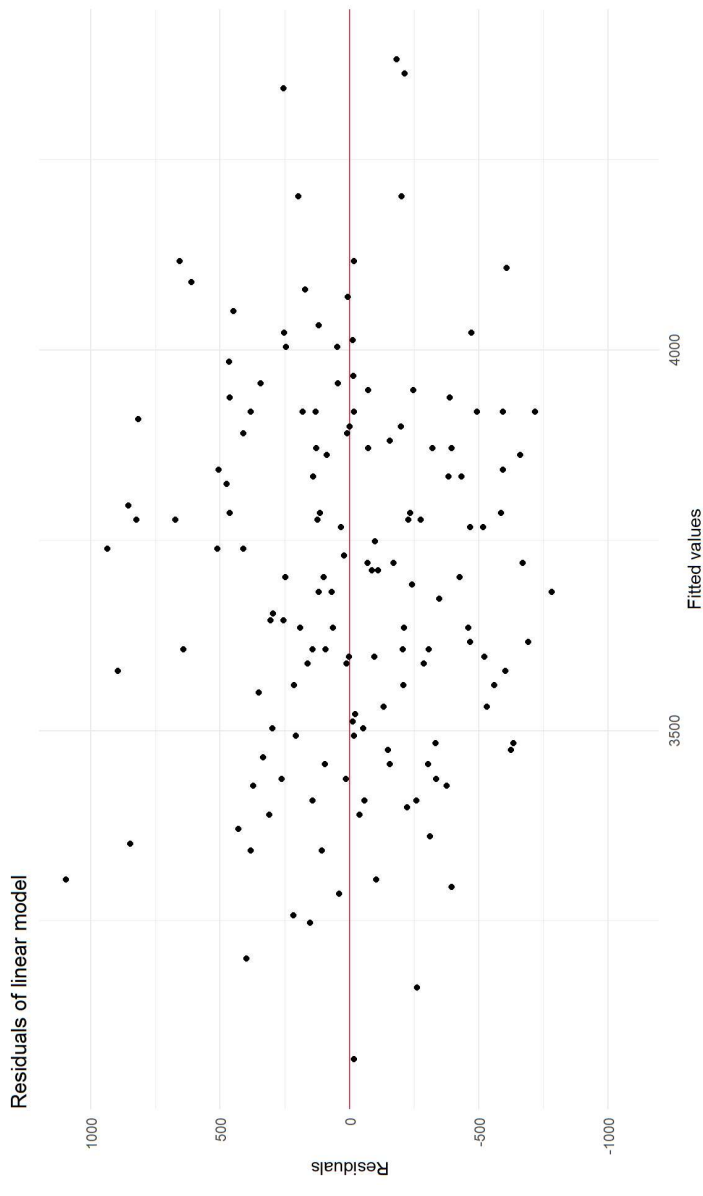
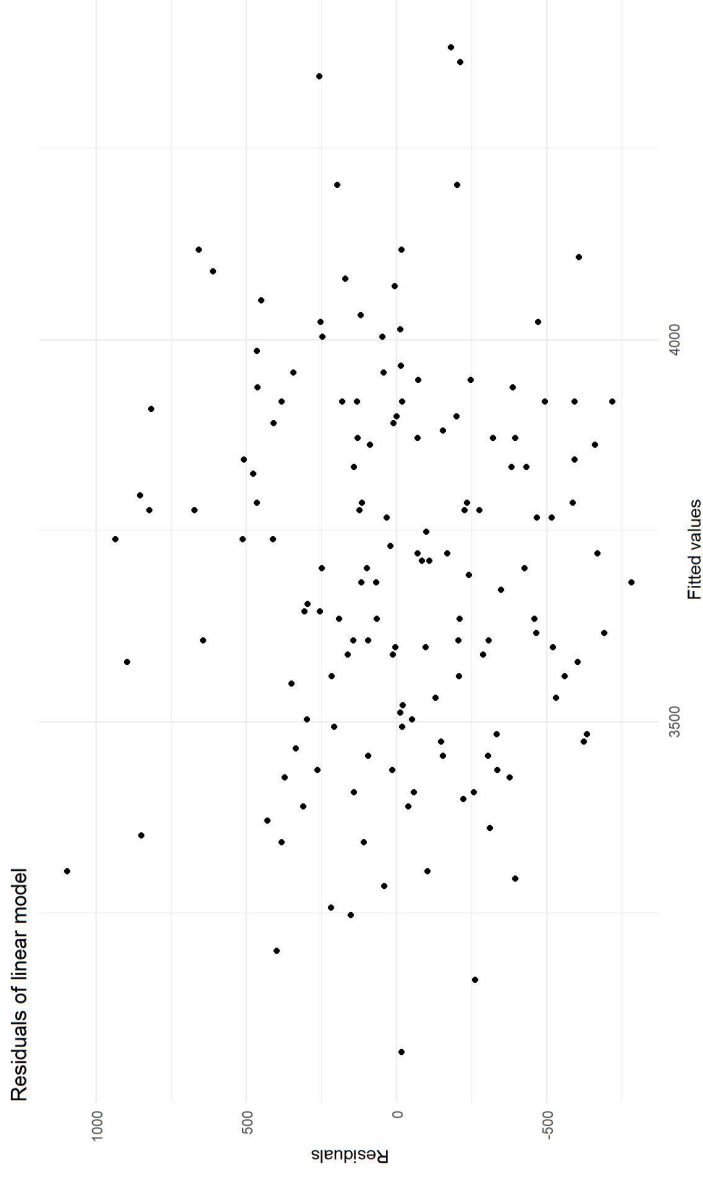


Layouts, aspect ratios, and axes

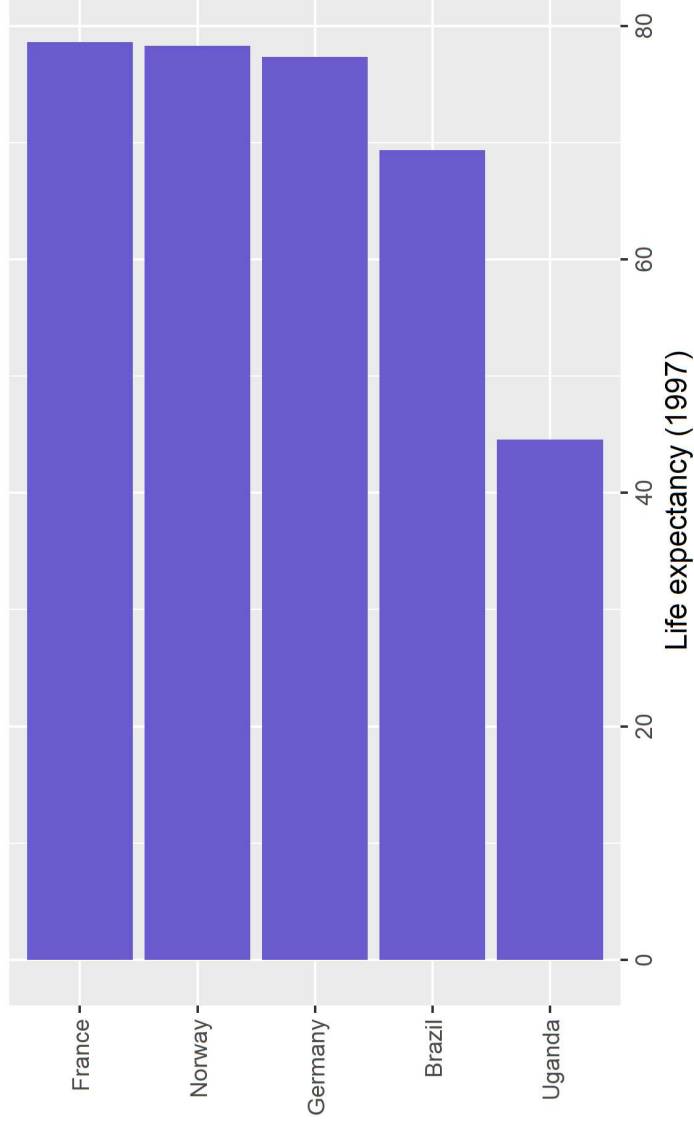
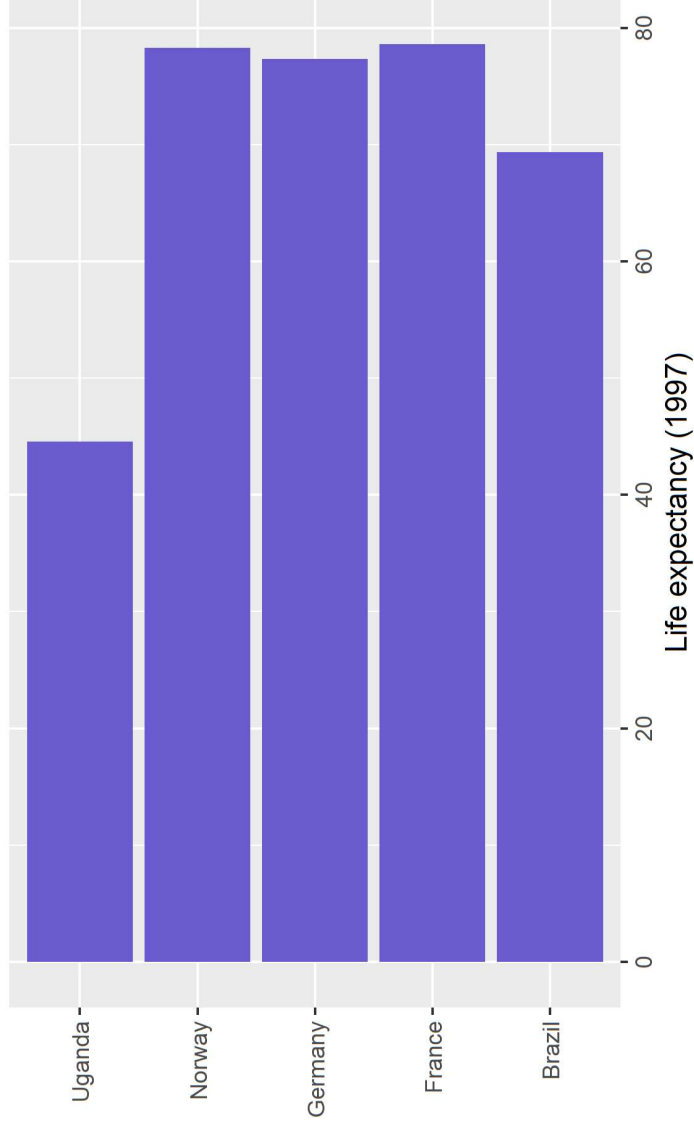
They don't *always* have to start at zero...



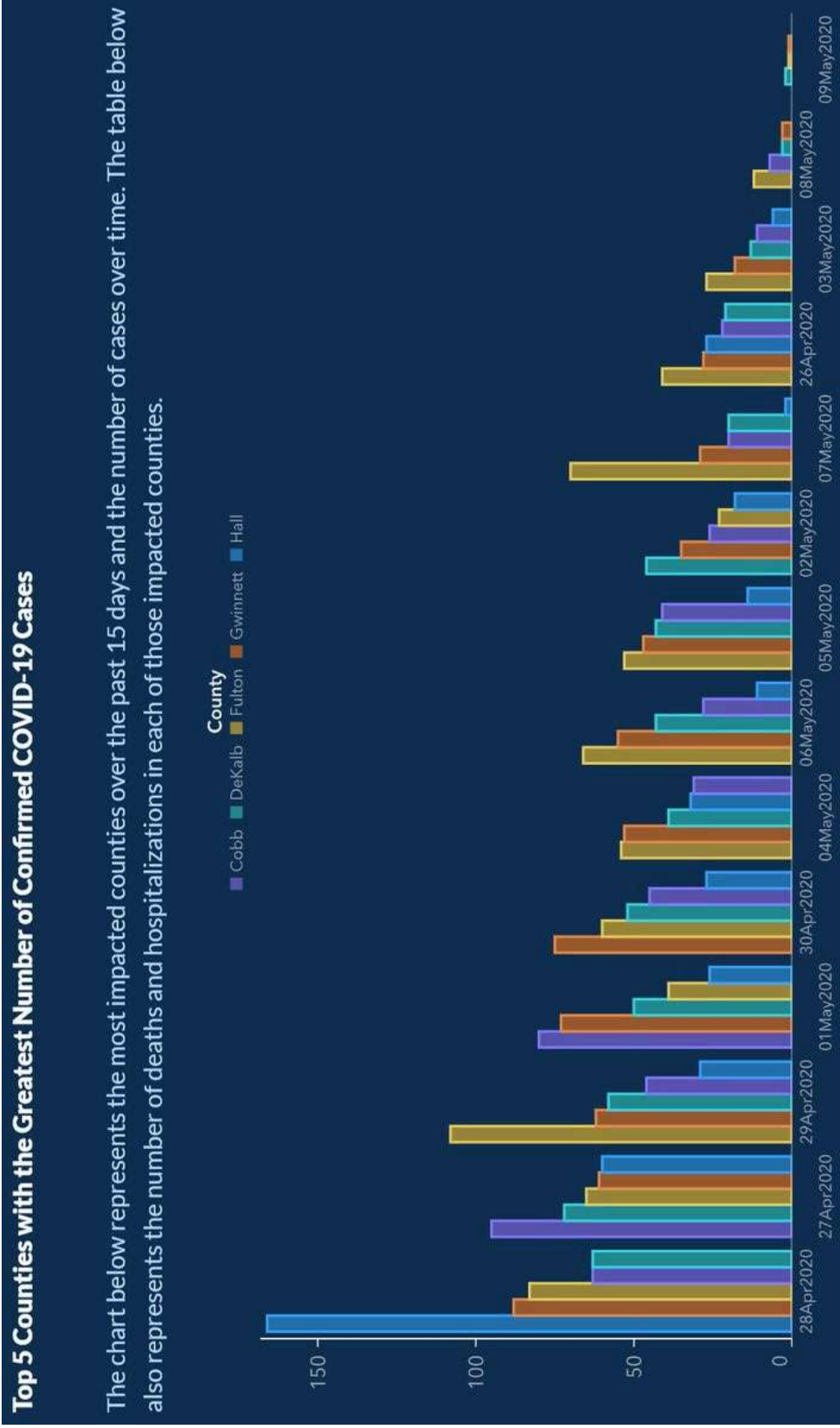
Axis ranges



Comparing categories



Use common sense!



Source: Georgia Department of Public Health

nrennie.rbind.io/training-data-visualisation

Order based on magnitude unless the category order has meaning...

Colours

Why use colours in data visualisation?

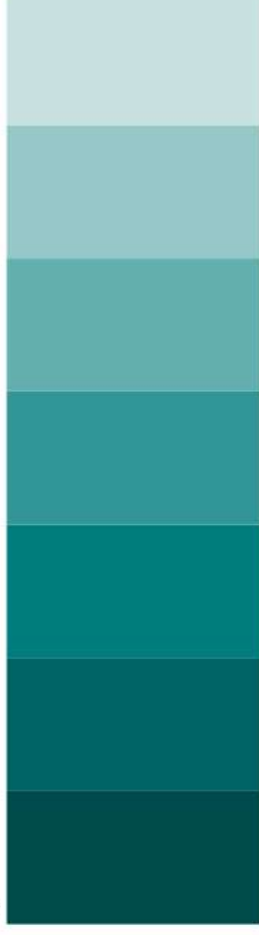
- Colours should serve a purpose, e.g. discerning groups of data
- Colours can highlight or emphasise parts of your data.
- Not always the most effective for, e.g. communicating differences between variables.

Colour palette types

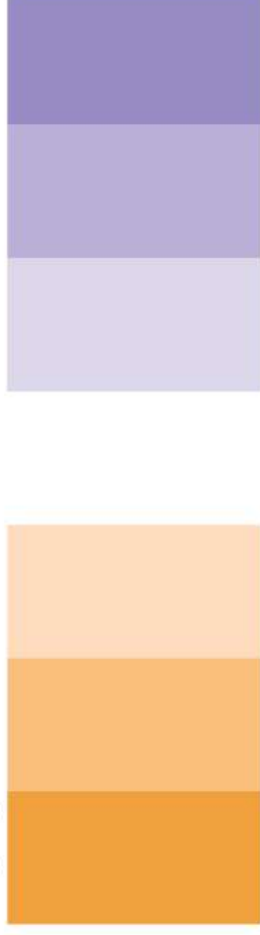
Different types of colour palettes...

... for different types of data.

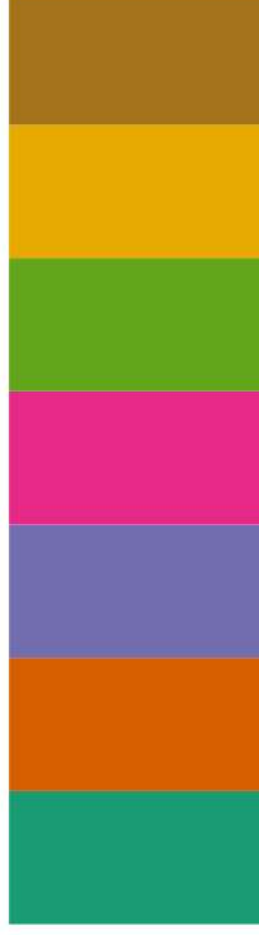
Sequential



Diverging



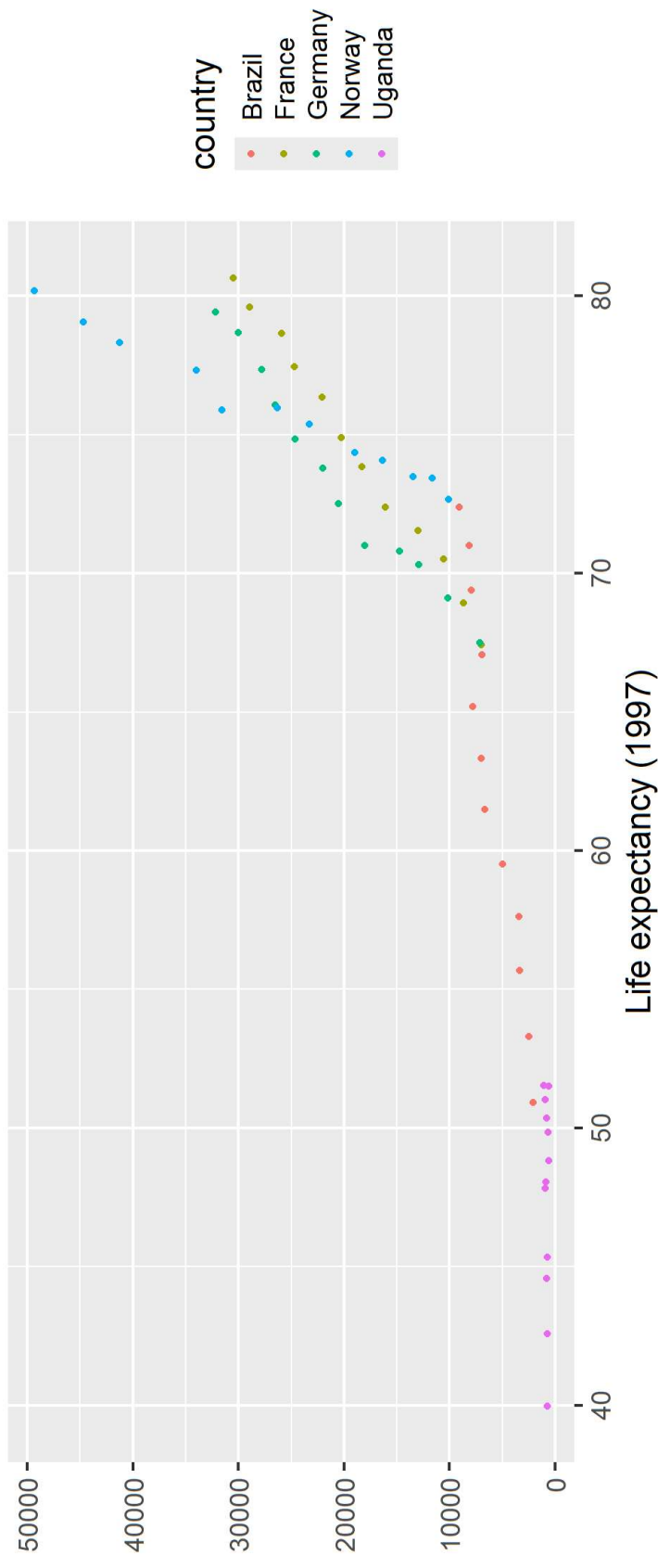
Qualitative





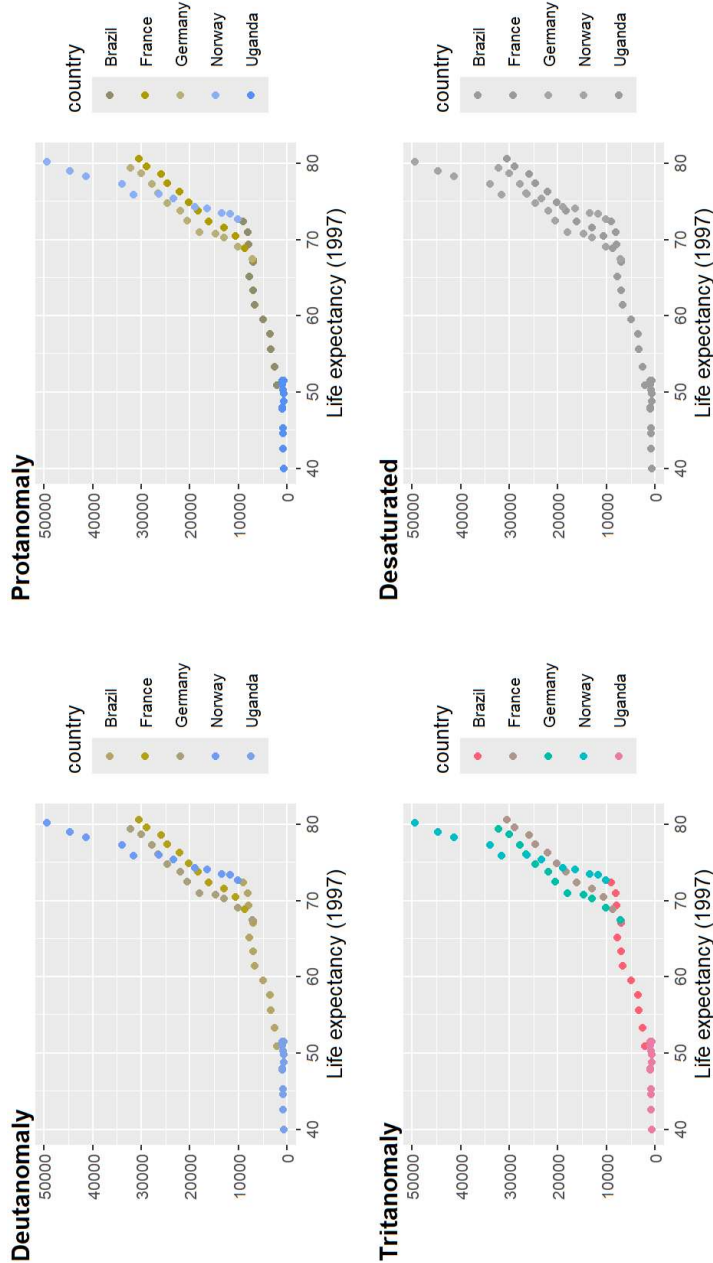
Choosing colours

Is this a good choice of colour?



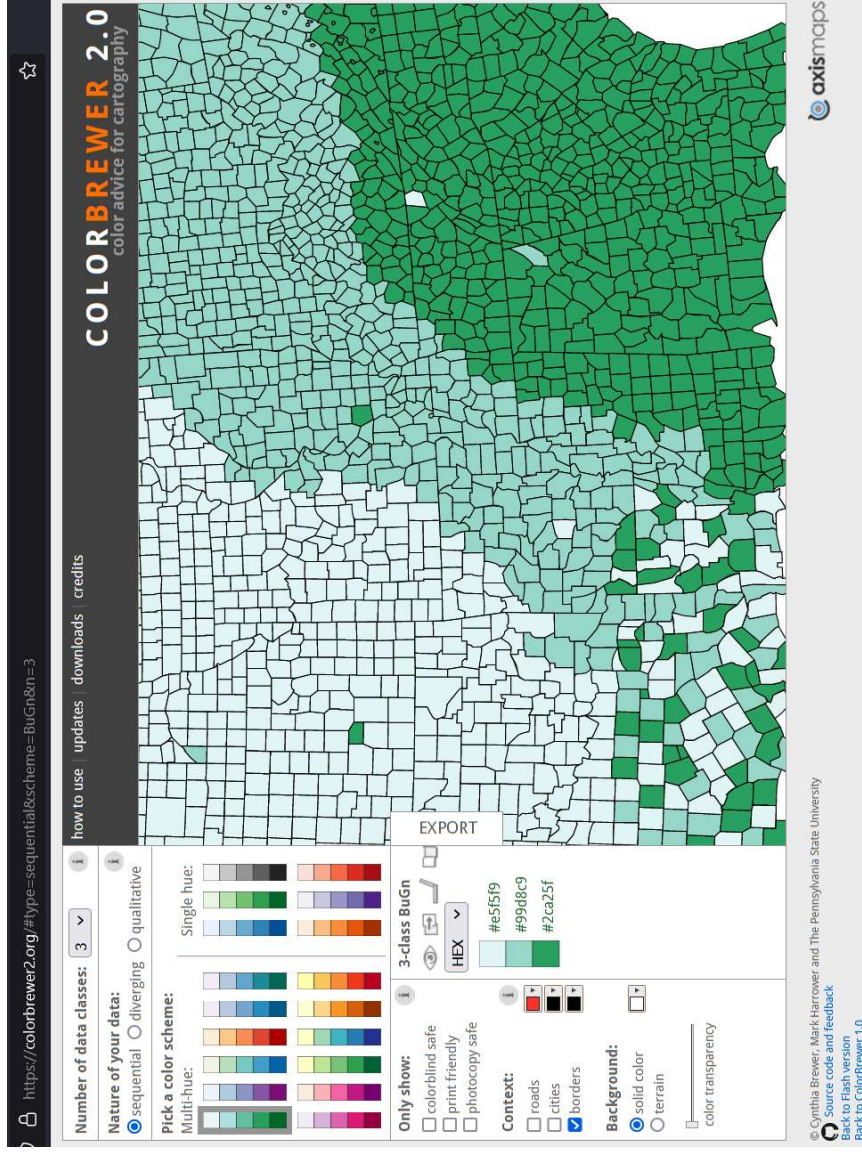
Accessible colours

Check for colourblind friendly plots with `colorblindr::cvd_grid(g)`.



Where do you find colours?

- In base R via `{RColorBrewer}`:
`brewer.pal(4, "Dark2")`
- In `{ggplot2}`, `scale_fill_brewer()`,
`scale_fill_distiller()`,
`scale_colour_brewer()`,
`scale_colour_distiller()`.
- In Python via `ColorBrewer`



colorbrewer2.org

Discussion

03:00

In groups, discuss how you might visualise the following data?

study_id	treatment	dosing_regimen_for_scurvy	gum_rot_d6	skin_sores_d6	weakness_of_the_knees_d6	lassitude_d6	fit_for_duty_d6
001	cider	1 quart per day	2_moderate	2_moderate	2_moderate	2_moderate	0_no
002	cider	1 quart per day	2_moderate	1_mild	2_moderate	3_severe	0_no
003	dilute_sulfuric_acid	25 drops of elixir of vitriol, three times a day	1_mild	3_severe	3_severe	3_severe	0_no
004	dilute_sulfuric_acid	25 drops of elixir of vitriol, three times a day	2_moderate	3_severe	3_severe	3_severe	0_no
005	vinegar	two spoonfuls, three times daily	3_severe	3_severe	3_severe	3_severe	0_no

A Treatise on the Scurvy in Three Parts. James Lind. 1757.

Initial exploration of data

Example data

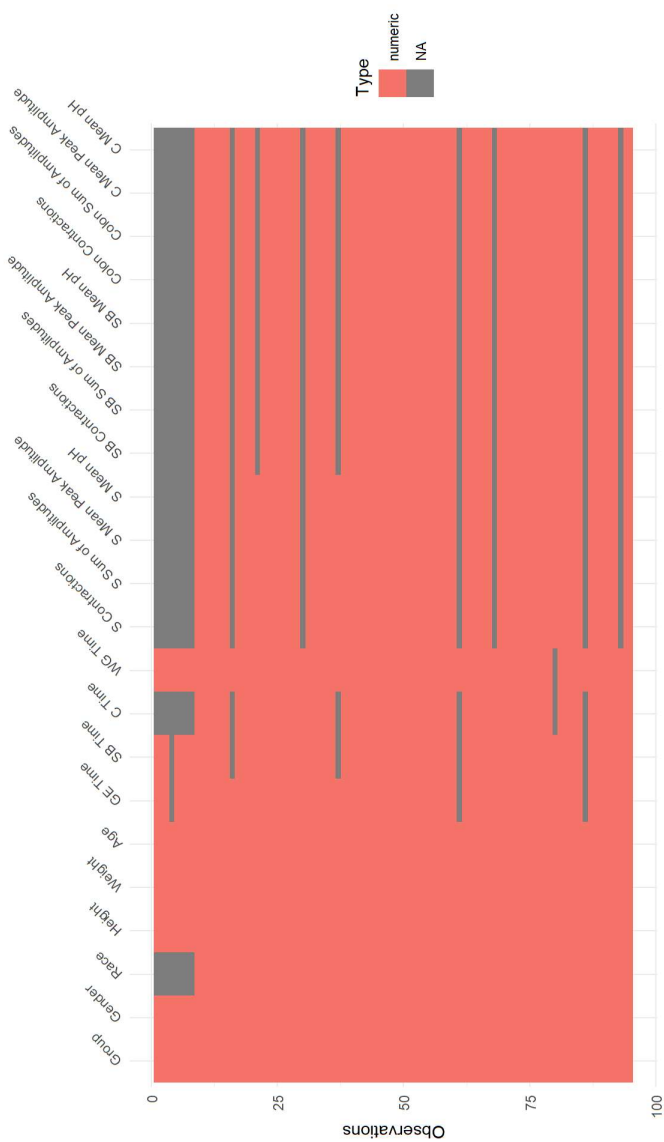
Smart Pill Data: causeweb.org/tshs/smart-pill

Group	Gender	Race	Height	Weight	Age	GE	SB	C	WG	S	S Sum of	S Mean	S	SB	SE
Time Time Time Time Time Time Time Time Time Time Amplitude pH															
0	1	NA	182.880	102.05820	25	74.30	8.40	NA	816.00	NA	NA	NA	NA	NA	NA
0	1	NA	180.340	102.05820	39	73.30	13.80	NA	168.00	NA	NA	NA	NA	NA	NA
0	1	NA	180.340	68.03880	44	4.30	6.70	NA	240.00	NA	NA	NA	NA	NA	NA
0	1	NA	175.260	69.85317	53	NA	NA	NA	216.00	NA	NA	NA	NA	NA	NA
0	0	NA	152.400	44.90561	57	13.90	5.10	NA	120.00	NA	NA	NA	NA	NA	NA
0	1	NA	185.120	61.80072	42	72.20	8.70	NA	284.00	NA	NA	NA	NA	NA	NA



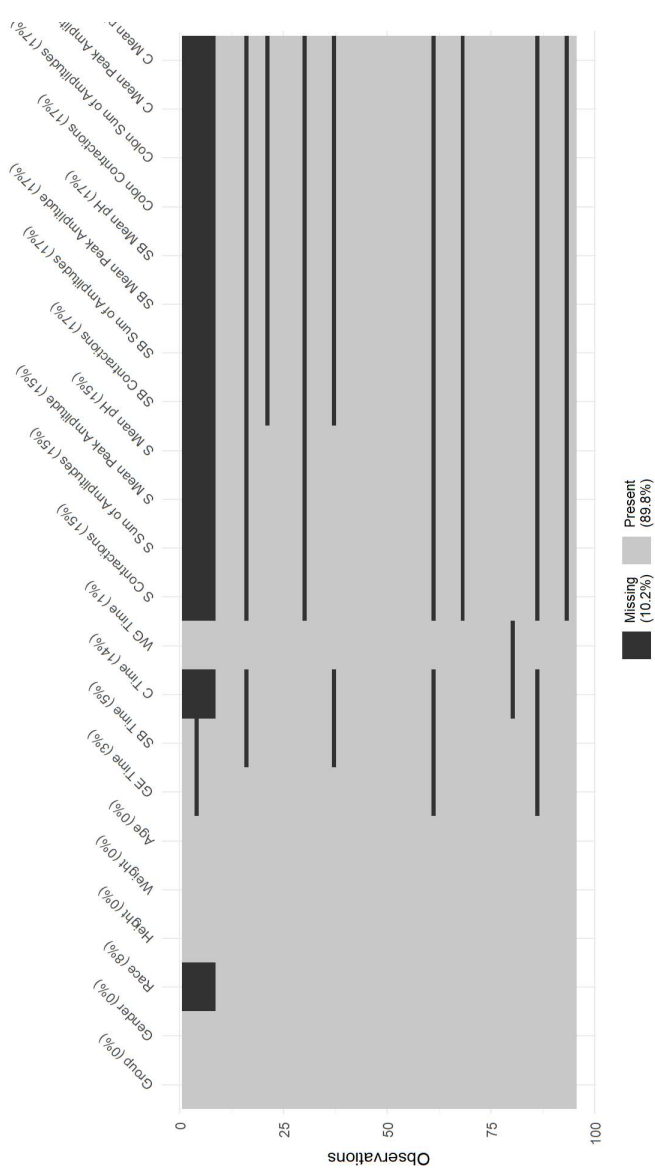
Look at your data

```
1 library(visdat)
2 vis_dat(smart_pill)
```



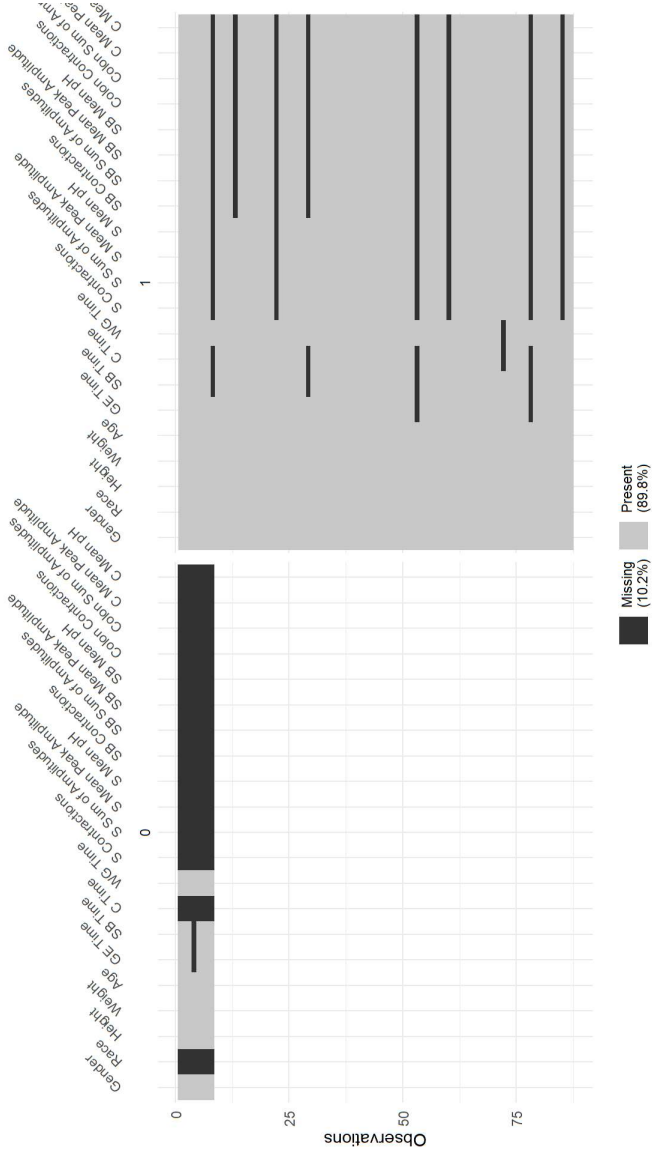
Identifying patterns of missingness

```
1 vis_miss(smart_pill)
```



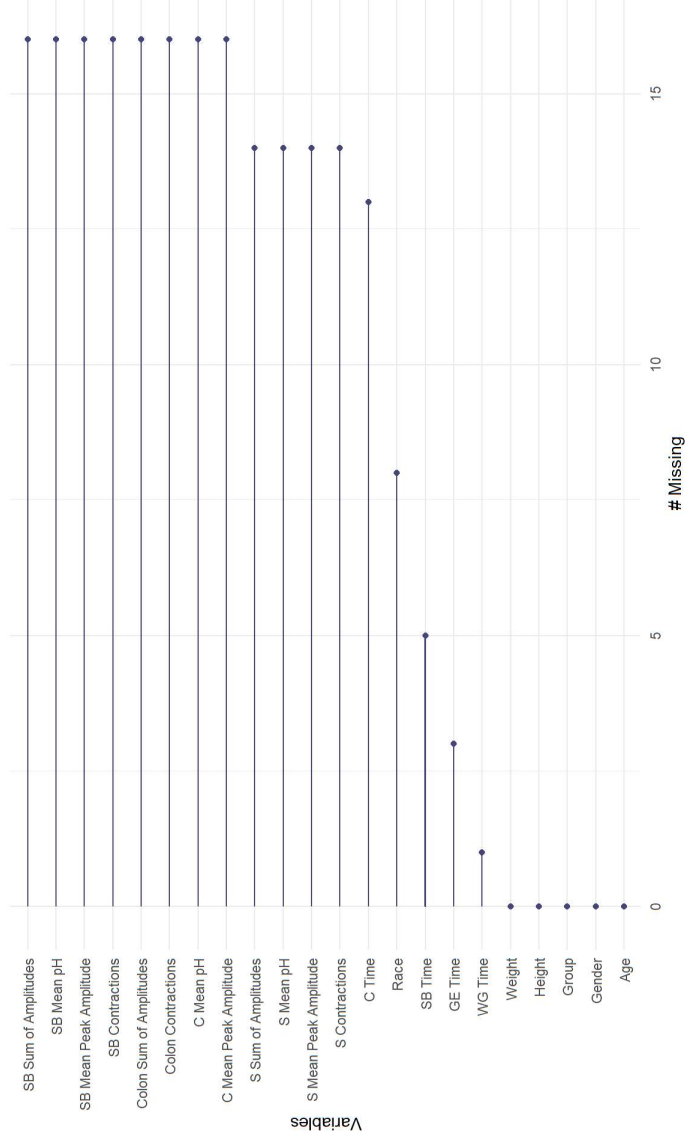
Identifying patterns of missingness

```
1 vis_miss(smart_pill, facet = Group)
```



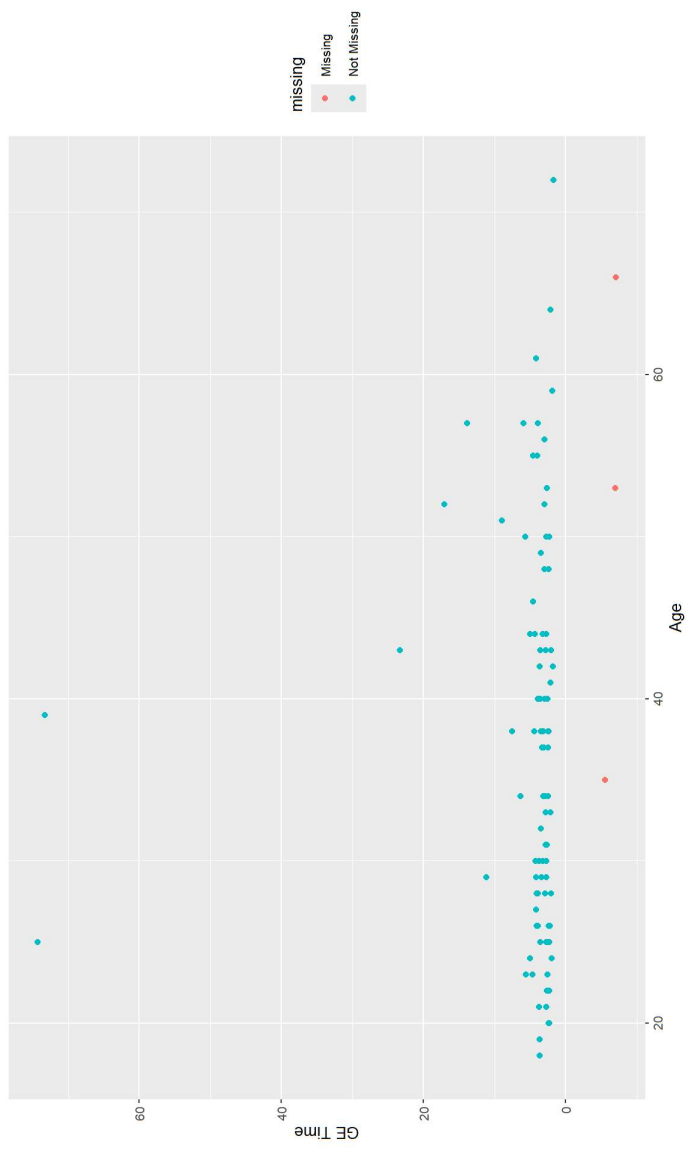
Identifying patterns of missingness

```
1 library(naniar)
2 gg_miss_var(smart_pill)
```



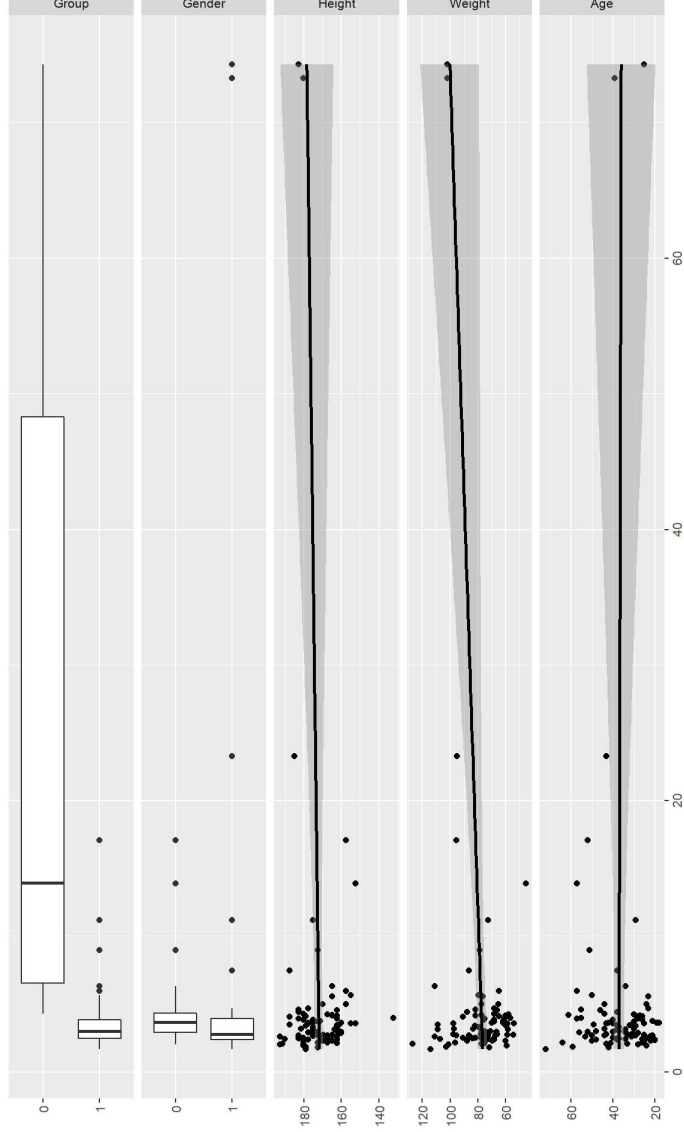
Identifying patterns of missingness

```
1 ggplot(smart_pill, aes(x = Age, y = `GE Time`)) +  
2   geom_miss_point()
```



Relationships with outcome variable

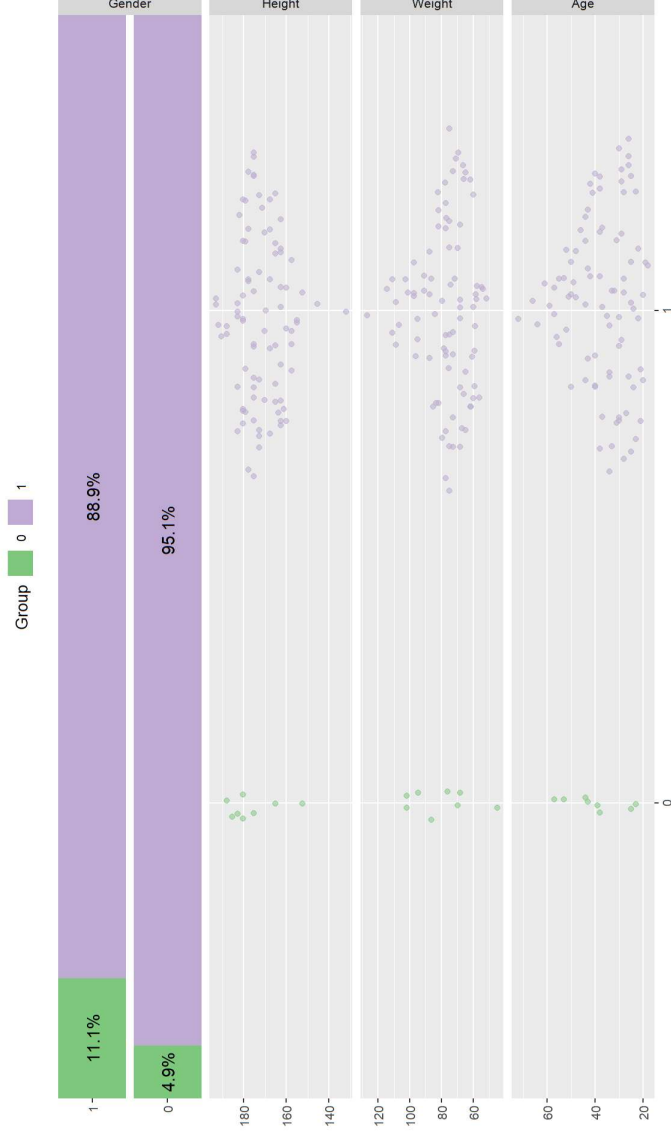
```
1 library(GGally)
2 smart_pill |>
3   mutate(across(c(Group, Gender), as.factor)) |>
4   ggbiivariate(
5     outcome = "GE Time",
6     explanatory = c("Group", "Gender", "Height", "Weight",
7     types = list(comboVertical = "autopoint")
8   )
```



How might you change this plot?

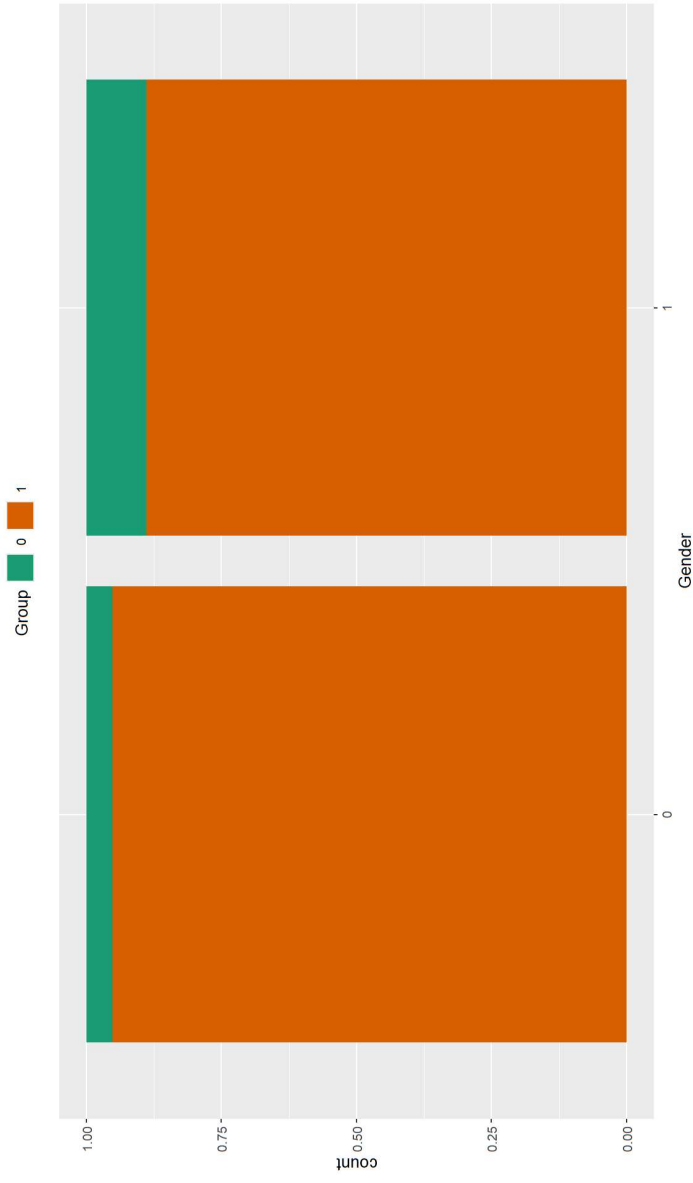
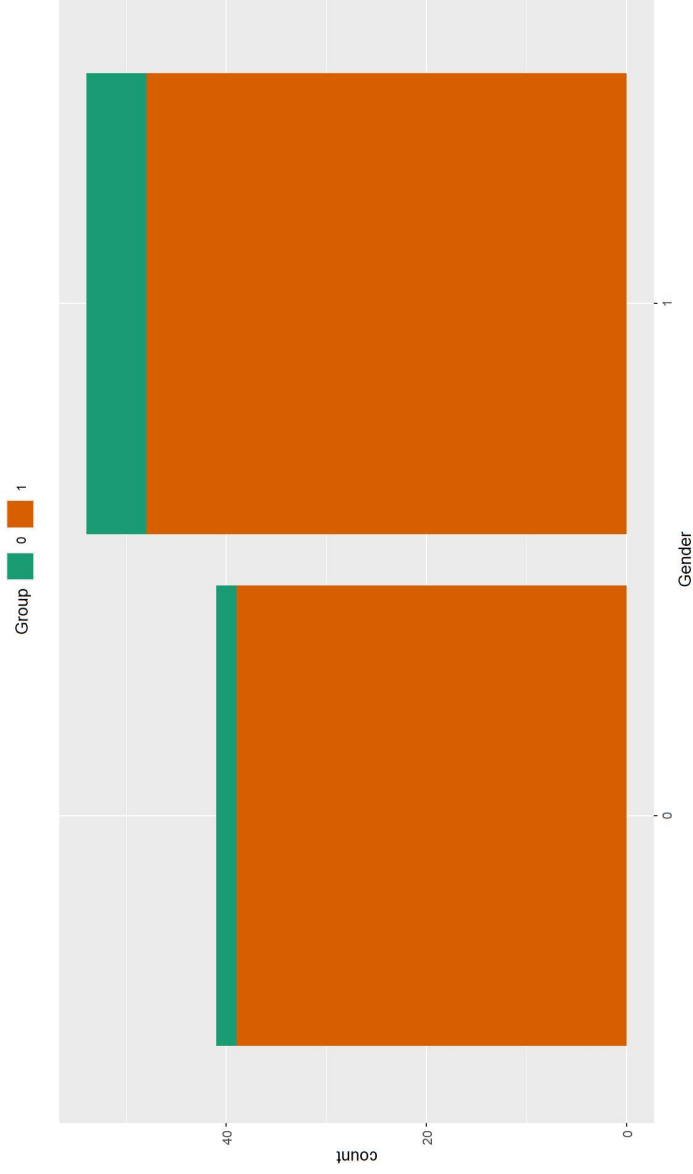
Relationships with outcome variable

```
1 smart_pill |>
2   mutate(across(c(Group, Gender), as.factor)) |>
3   ggbiplot(
4     outcome = "Group",
5     explanatory = c("Gender", "Height", "Weight", "Age"
6     types = list(comboVertical = "autopoint")
7   ) +
8     scale_fill_brewer(type = "qual") +
9     scale_colour_brewer(type = "qual")
```



How might you change this plot?

Absolute values or percentages?



Resources

nrennie.rbind.io/training-data-visualisation



Workshop website

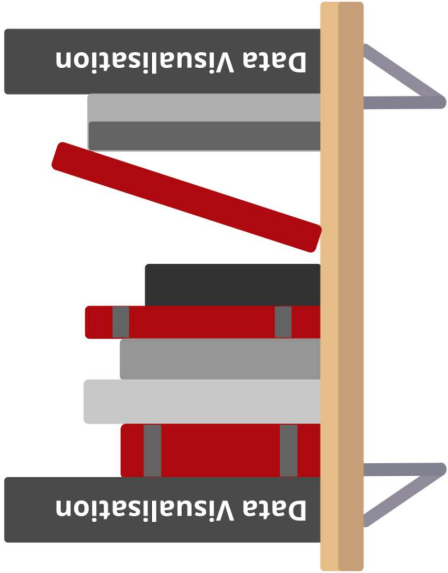
Workshop website: `nrennie.rbind.io/training-data-visualisation`

Welcome!

Welcome to the *More effective data visualisation* workshop! The workshop website can be found at nrennie.rbind.io/training-data-visualisation, where you will also find a copy of the slides.

Contents

You'll often hear me describe visualisation as the *book ends of analysis projects*. Essentially, data visualisation is one of the first things you'll do (to better understand the data you have) and also one of the last things you'll do (to communicate your results to others). Although there are many similarities between visualisations for these two purposes, there are also some differences.



Therefore, this workshop is split into two parts:

RSS Data Visualisation Guide

royal-statistical-society.github.io/datavisguide

Royal Statistical Society

Best Practices for Data Visualisation

Royal Statistical Society publications

Introduction

How to use this guide

The guide

Why we visualise data

Principles and elements of visualisations

Choosing a visualisation type

Styling for accessibility

Styling for RSS publications

References and resources

About the authors

Join us at RSS Conference

Terms and conditions

Coming to the RSS Conference this September? Find out more about [our conference session](#) and how to become a guide contributor!

Statistics is “the science of collecting, analyzing, presenting, and interpreting data” (Williams, Anderson, and Sweeney 2023). Presentation of data is a key means to support and guide interpretation and subsequent decision making. Techniques exist for effective display. This is what this guide is all about.

Good data visualisation requires appreciation and careful consideration of the technical aspects of data presentation. But it also involves a creative element. Authorial choices are made about the “story” we want to tell, and design decisions are driven by the need to convey that story most effectively to our audience. Software systems use default settings for most graphical elements. However, each visualisation has its own story to tell, and so we must actively consider and choose settings for the visualisation under construction.

This guide covers both aspects of data visualisation: the art and the science. It is written primarily for contributors to [Royal Statistical Society publications](#) – chiefly, [Significance magazine](#), the [Journal of the Royal Statistical Society Series A](#), and [Real World Data Science](#) – but we trust you will find the information and advice within to be of broad relevance and use to any data visualisation task.

42

nrennie.rbind.io/training-data-visualisation

Data visualisation resources

Fundamentals of data visualization - clauswilke.com/dataviz

ONS Guidance - service-manual.ons.gov.uk/data-visualisation

DataWrapper Blog - blog.datawrapper.de

R Graphics Cookbook - r-graphics.org

ggplot2: Elegant Graphics for Data Analysis - ggplot2-book.org

 nrennie.rbind.io

 nicola-rennie

 nrennie