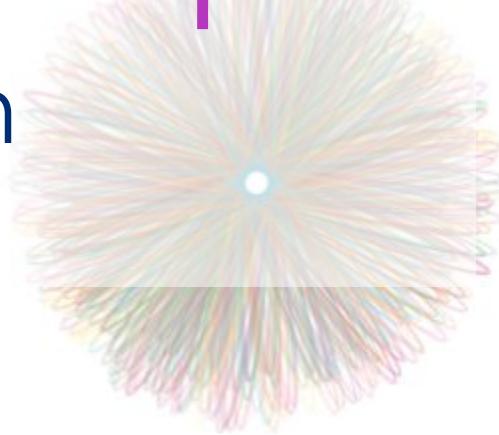
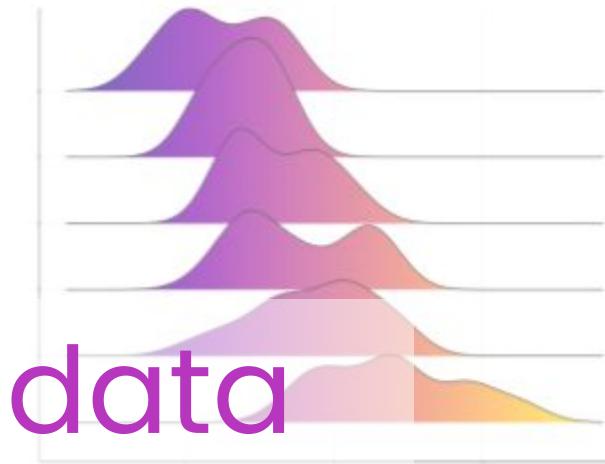
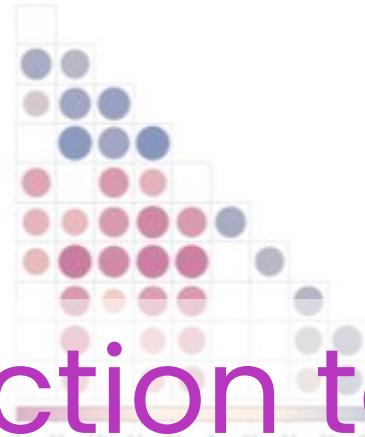


Introduction to data visualisation principles

August 19th

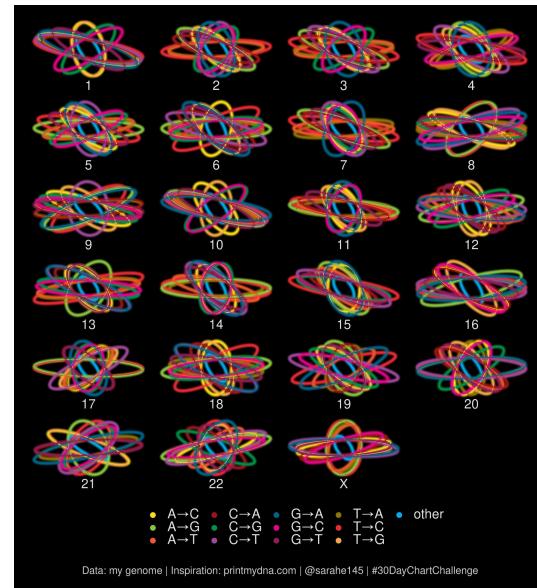
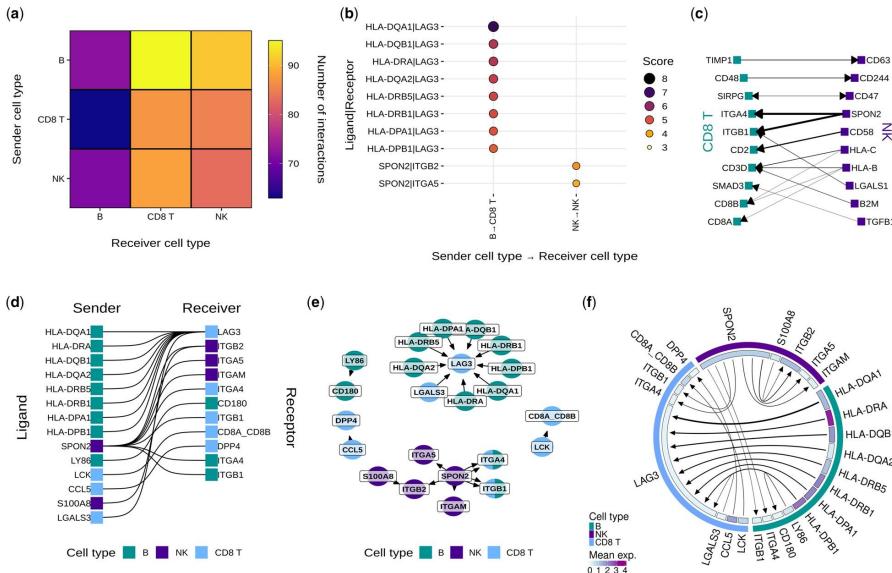


What to expect

- Introduction
- Why you should care about data visualisation
- Different plot types and how to choose the correct one
- Other considerations
- Case study

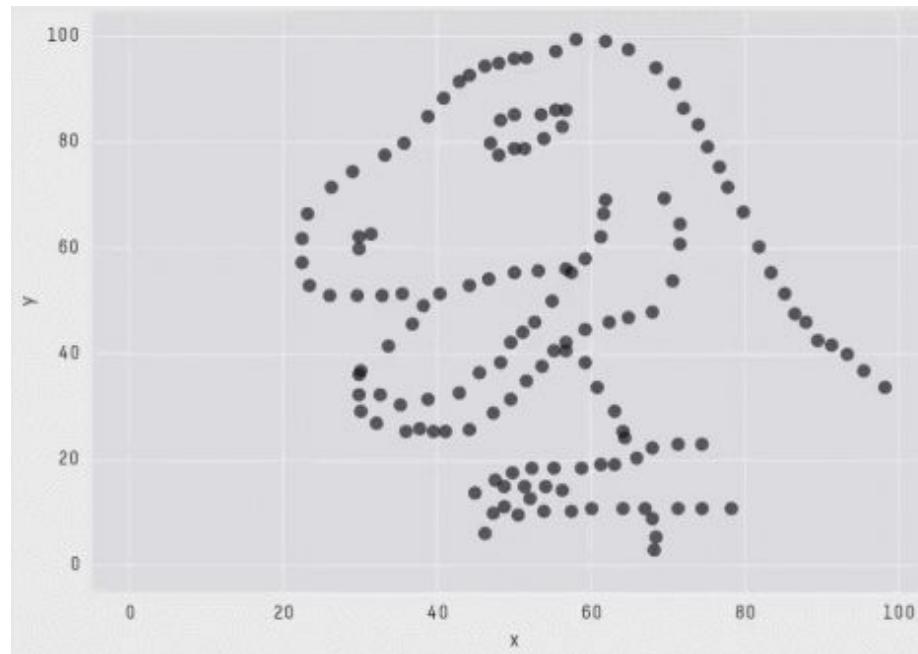
Introduction

- PhD in Genomics Data Science at University of Galway
 - Interested in data visualisation
 - Currently work for a drug discovery startup based in Toronto



Why you should care about data visualisation

Find hidden patterns in data



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Why you should care about data visualisation

Charts change minds

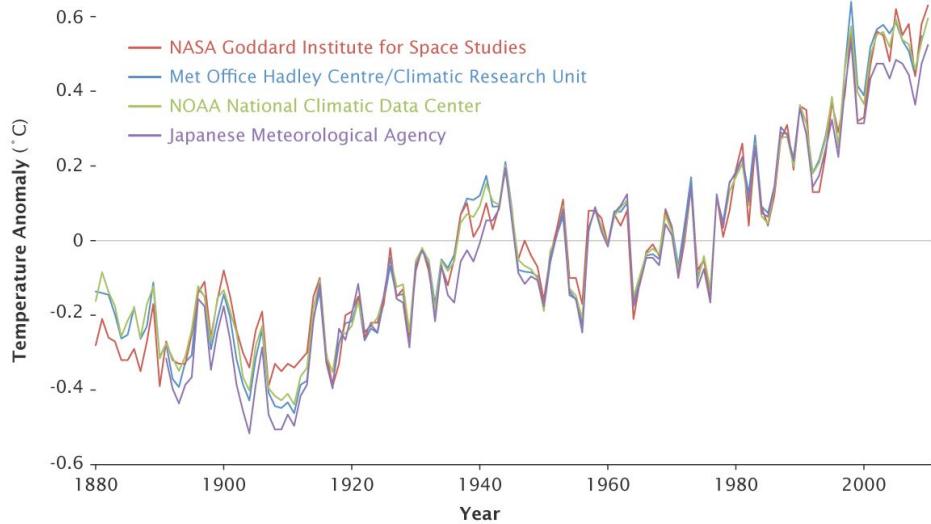
Text treatment

"Groups of scientists tally data collected by temperature monitoring stations spread around the world. They each show an increase in average global surface temperatures of approximately 0.5 degrees Celsius over the last three decades. Data from each source also indicate that the last decade is the warmest since 1940."

Chart treatment

Global Surface Temperatures

Four independent records show nearly identical long-term warming trends.



Why you should care about data visualisation

Charts change minds

"Would you say that the average global surface temperature has gone up, stayed about the same, or gone down in the last thirty years?"

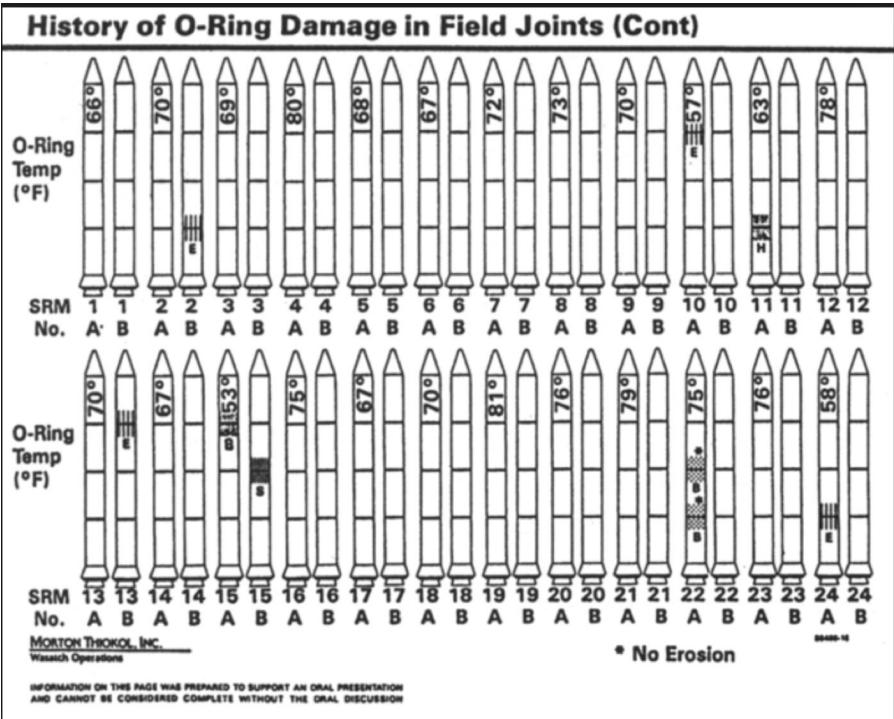
Share of US Republicans who do not believe global warming is happening



Source: Nyhan and Reifler, 2017

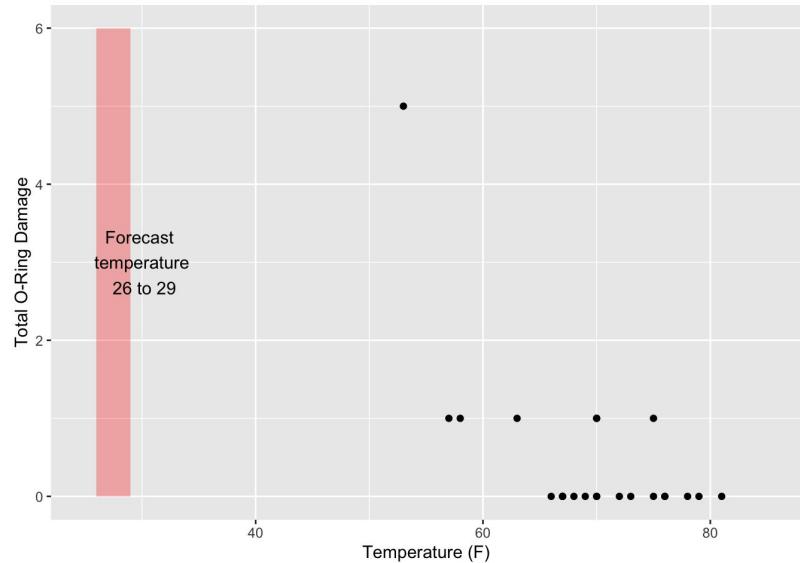
Why you should care about data visualisation

Good data viz can save lives



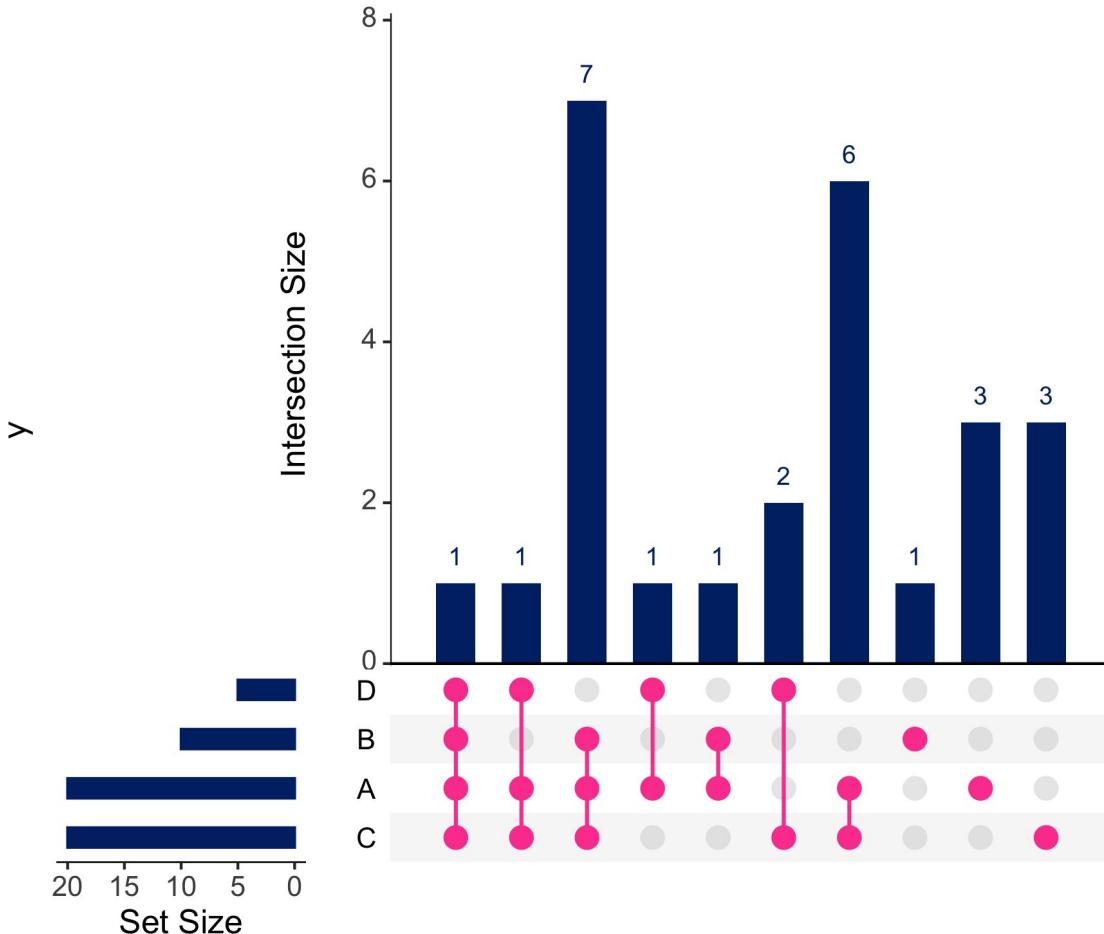
Why you should care about data visualisation

Good data viz can save lives



Plot types

- Scatter



Choosing the correct chart type

Considerations

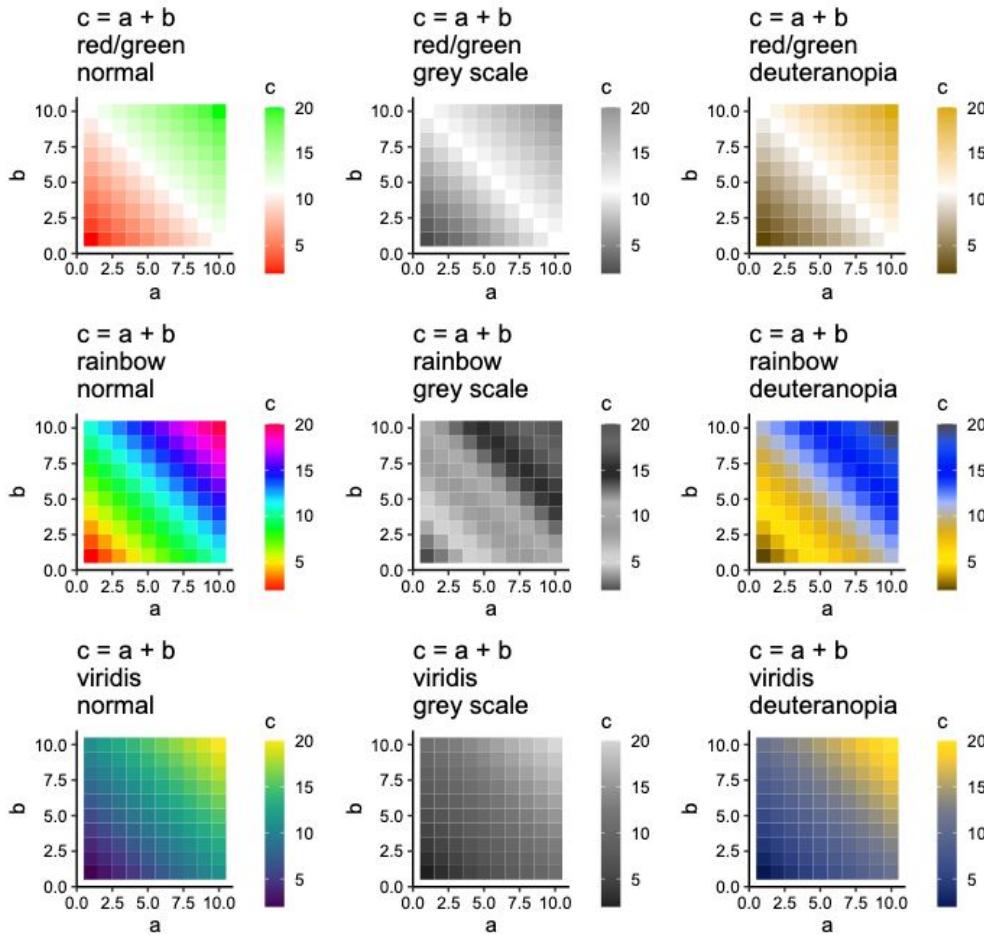
- ? What type of data are you working with – distribution or trend in a single variable or a relationship or comparison between multiple variables
- ? What is the audience for the visualisation – lab meeting, conference presentation, thesis/publication
- ? What question do you want the plot to answer

Choose a chart type that will visualise the message you want to convey in the most effective and obvious way

Colour

Colour-blindness

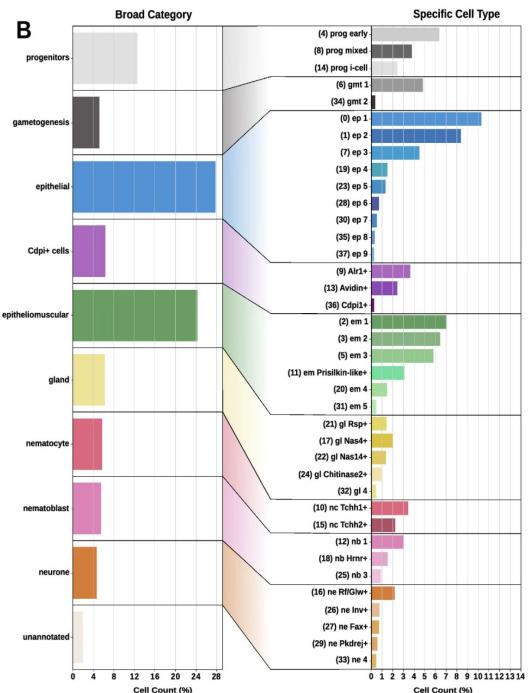
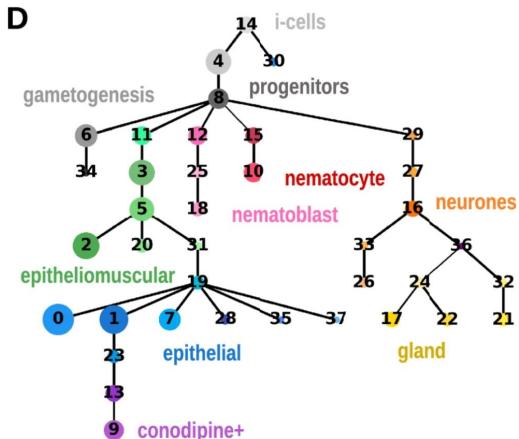
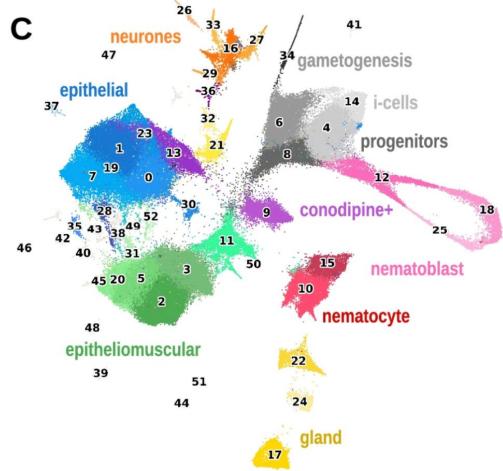
- 4.5% population has some form of colour-blindness (1 in 12 men)
- Making charts colour-blind-friendly also means they can be printed in black/white
- Viridis palettes are a good accessible alternative



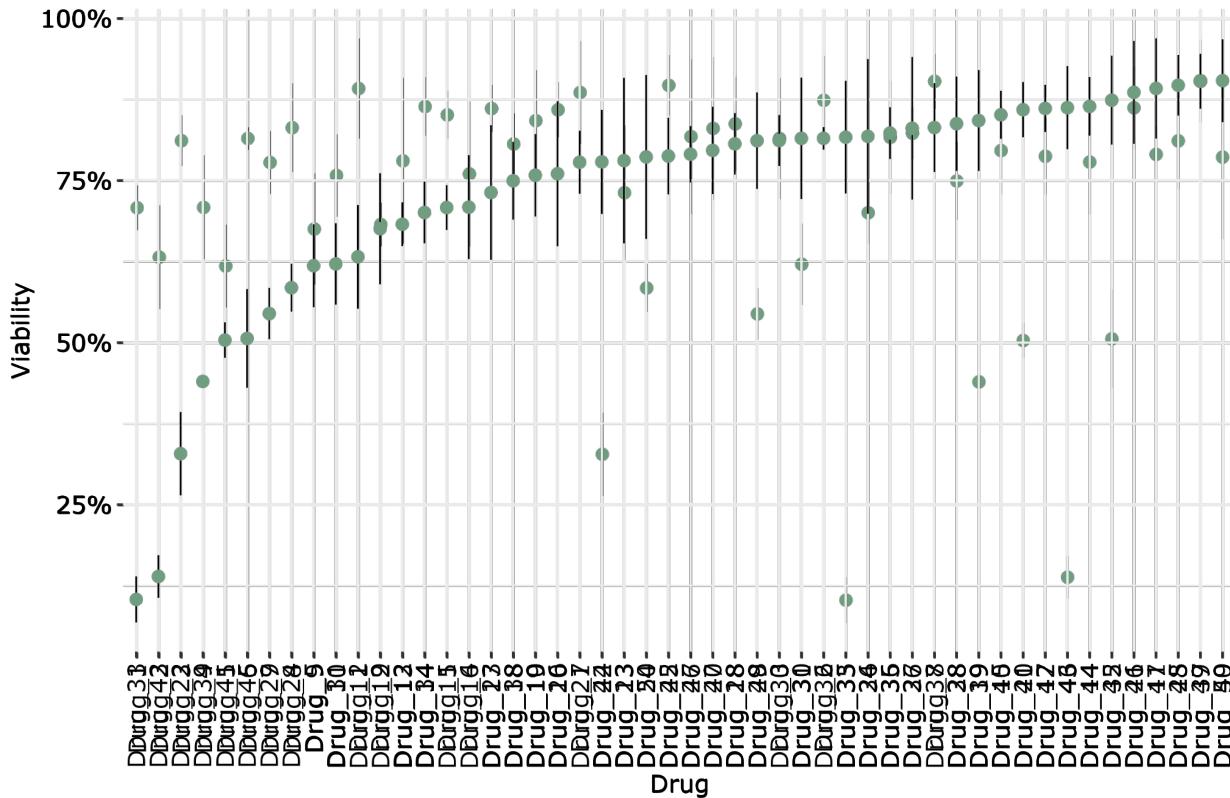
Colour



Be consistent



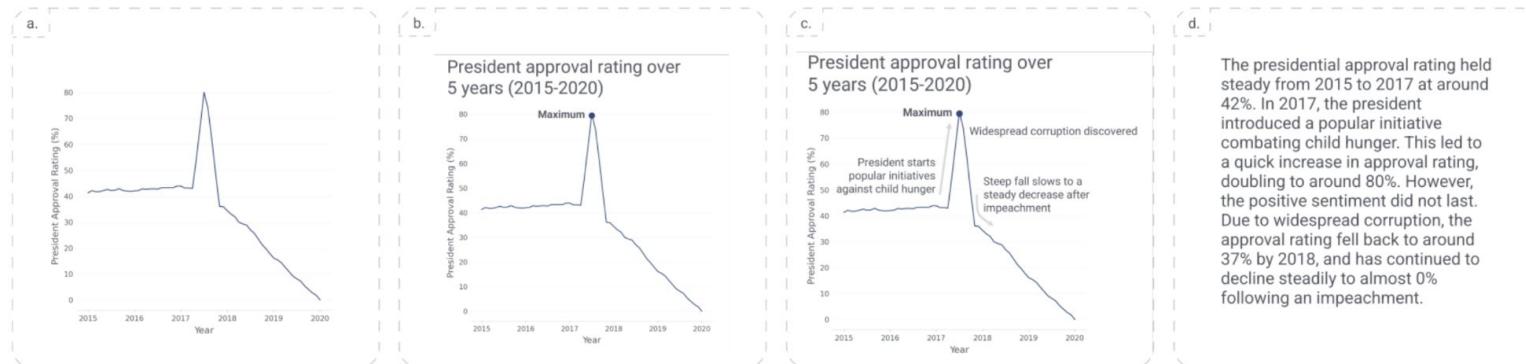
Ordering variables



Use annotations and labels

Striking a Balance: Reader Takeaways and Preferences when Integrating Text and Charts

Chase Stokes, Vidya Setlur (*Member*), Bridget Cogley, Arvind Satyanarayan, and Marti A. Hearst



4th



Resources

-  [Data to viz - choosing the correct chart type](#)
-  [Datawrapper blog](#)
-  [R graph gallery](#)
-  [Tidy Tuesday](#)

Case study

The screenshot shows a web browser displaying a dataset page from [data.cdc.gov](https://data.cdc.gov/Foodborne-Waterborne-and-Related-Diseases/NORS/5xkq-dg7x/about_data). The page title is "NORS | Data | Centers for Disease Control and Prevention". The main content area is titled "What's in this Dataset?" and displays the following information:

Rows: **66.7K** Columns: **19** Each row is a **Outbreak**

Columns (19)

Column Name	Description	API Field Name	Data Type
# Year	Year of earliest date of reported illness onset	year	Number
# Month	Month of earliest date of reported illness onset	month	Number
Tr State	The state where the exposure occurred. For a single state of exposure, the state will be listed. For multiple states of exposure, "Multistate" will be listed.	state	Text
Tr Primary Mode	Primary mode of transmission.	primary_mode	Text
Tr Etiology	Genus and species of identified etiology. Multiple	etiology	Text

https://data.cdc.gov/Foodborne-Waterborne-and-Related-Diseases/NORS/5xkq-dg7x/about_data

Case study



ChatGPT

Prompt : *"I have a dataframe that contains information on outbreaks from 1971 to 2023. Each row represents a recorded outbreak and there's a column for Year and Primary.Mode for each outbreak. Can you write me some R code that would create a stream graph from this dataframe where the x axis depicts year, the y axis depicts number of outbreaks and the streams are coloured by primary mode of transmission."*

Case study

R Code for Stream Graph:

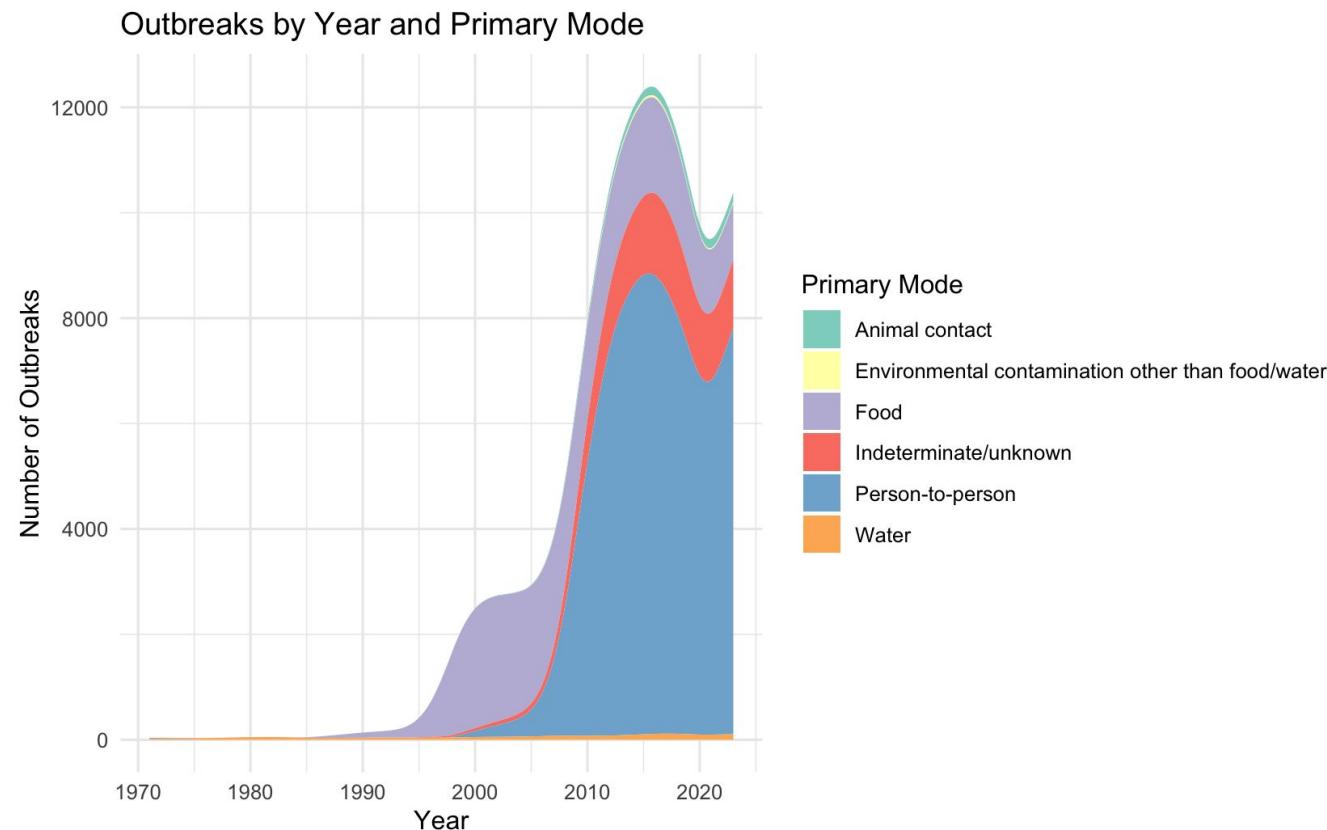
```
r

# Load required libraries
library(ggplot2)
library(ggstream)

# Summarize number of outbreaks per year per PrimaryMode
library(dplyr)
df_summary <- df %>%
  group_by(Year, PrimaryMode) %>%
  summarise(Outbreaks = n(), .groups = "drop")

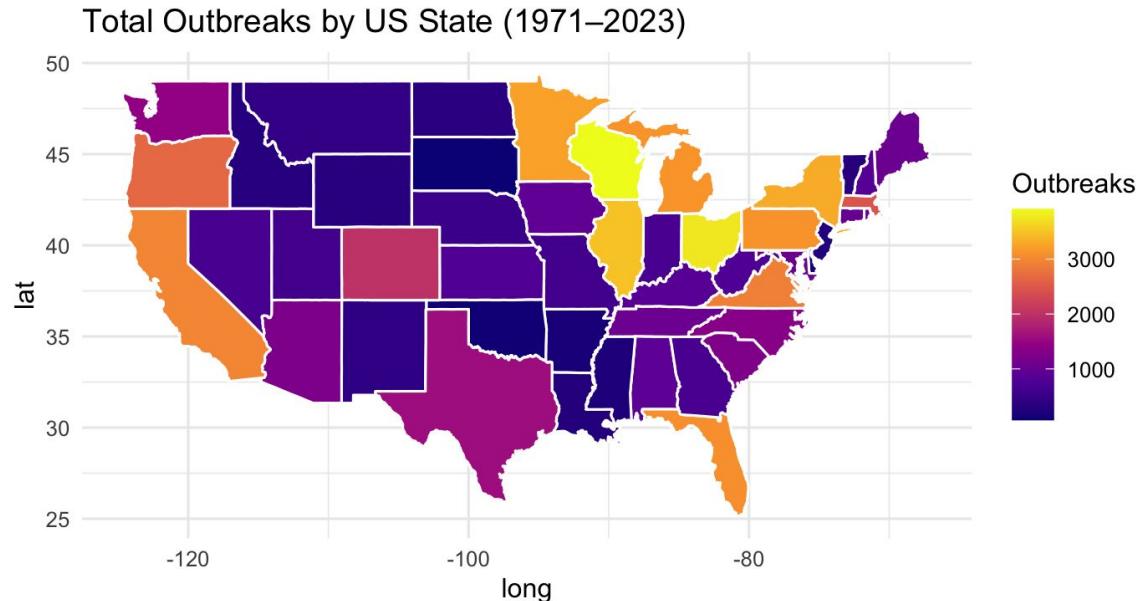
# Create the stream graph
ggplot(df_summary, aes(x = Year, y = Outbreaks, fill = PrimaryMode)) +
  geom_stream(type = "ridge") +
  scale_fill_brewer(palette = "Set3") + # Choose a color palette
  theme_minimal() +
  labs(title = "Outbreaks by Year and Primary Mode",
       x = "Year",
       y = "Number of Outbreaks",
       fill = "Primary Mode")
```

Case study



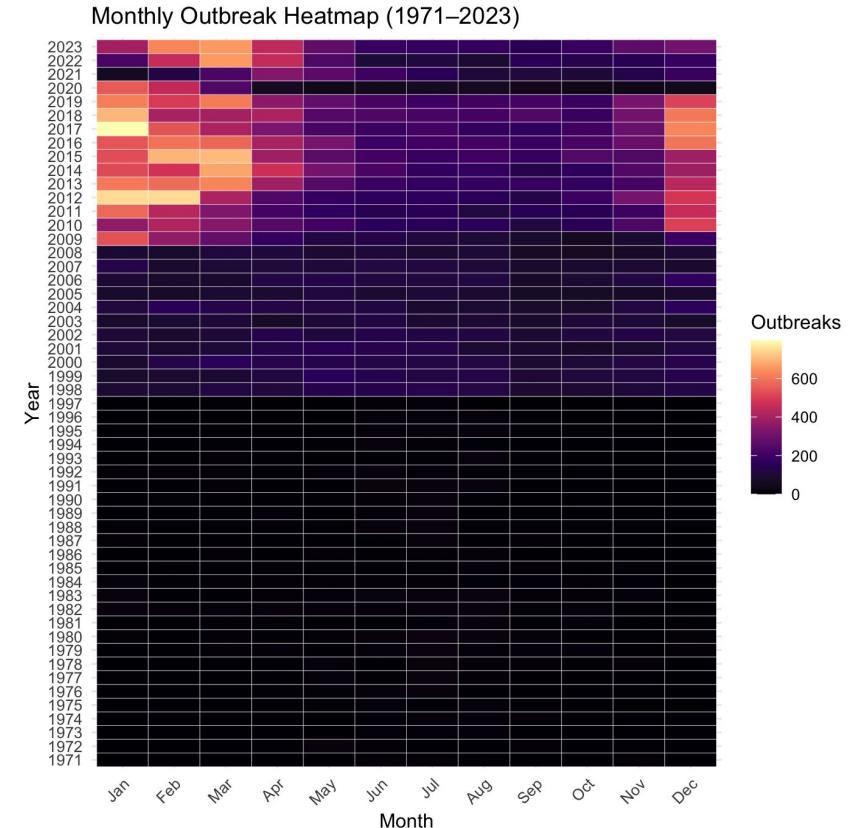
Case study

"The dataframe also has a column called State that represents which state (in the US) each outbreak was recorded in. Could you write some R code that would create a **choropleth map** of this data where the states are coloured by total number of outbreaks."



Case study

*"The dataframe also has a Month column. Could you write some R code that would produce a **heatmap** where each row represents a year, the column represents the month and the tiles are coloured by total number of outbreaks detected in that month/year."*



Summary

- The best plot is the plot that gets your message across in the most obvious and impactful way
- Things such as colour, annotations and the order of variables can make a huge difference to your plots - use them wisely
- The internet is full resources for data visualisation tips, tricks and inspiration
- If stuck, ask ChatGPT

Happy plotting! 😊

