Week Six

Open Science

Philip Leftwich

25.10.2021

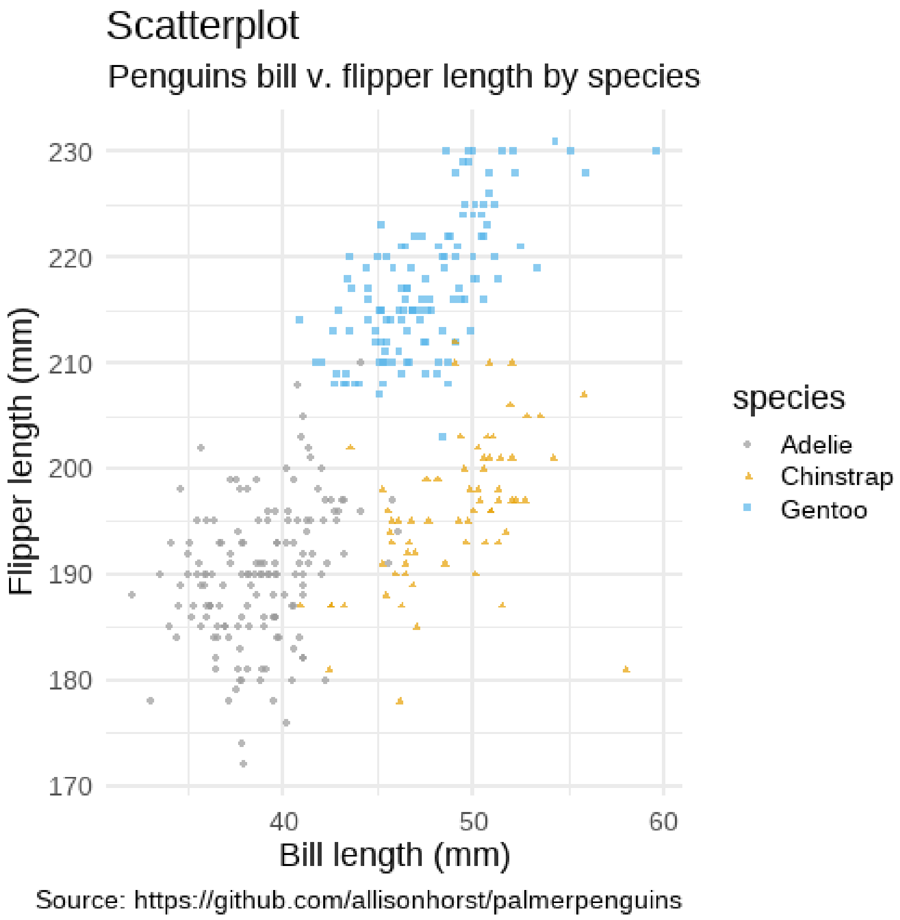
Hello! 👋

- Let me know how you are today in the Q&A on today's Slido!
- Go to Slido.com #944385

Good data visualisations

```
penguins %>%
  remove_missing() %>%
  ggplot(aes(x = bill_length_mm, y = f
        color = species, shape = specie
  geom_point(alpha = 0.7) +
  labs(x = "Bill length (mm)",
      y = "Flipper length (mm)",
     title = "Scatterplot",
     subtitle = "Penguins bill v. flipper le
     caption = "Source: https://github.
```

Slido.com #944385



Scatterplot

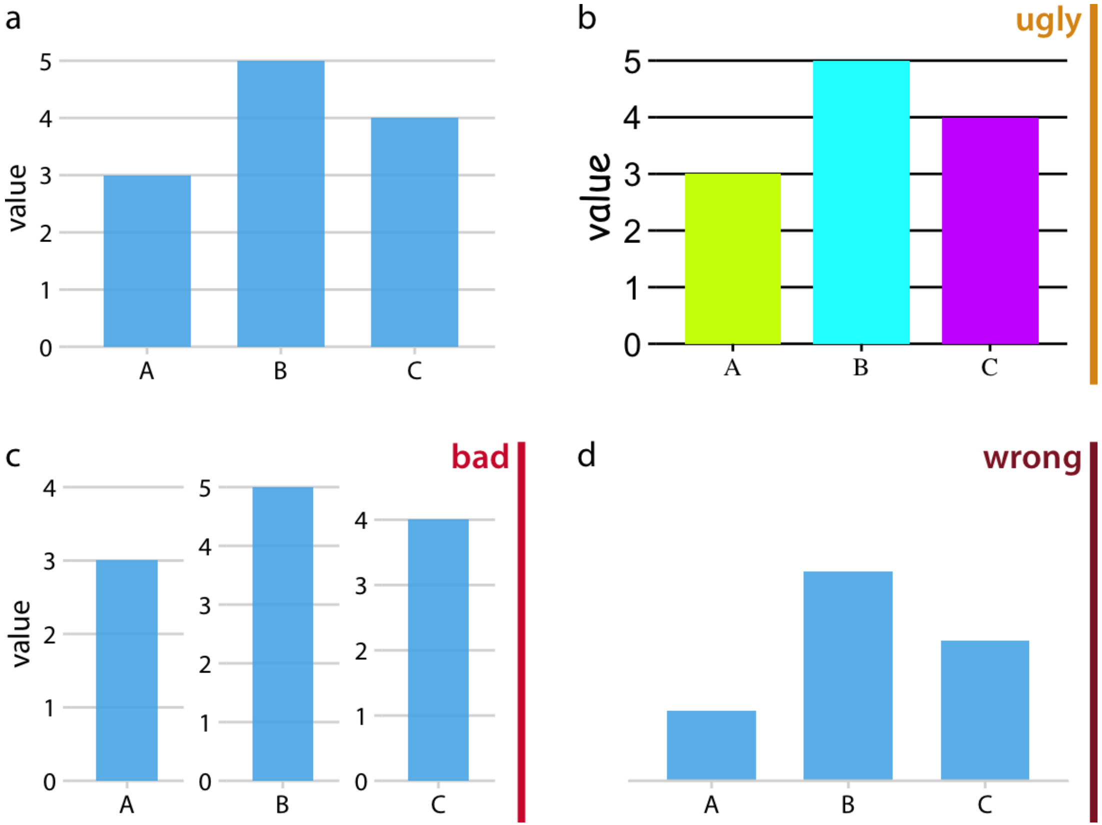
Penguins bill v. flipper length by species

Source: https://github.com/allisonhorst/palmerpenguins

Requirements of a good graph

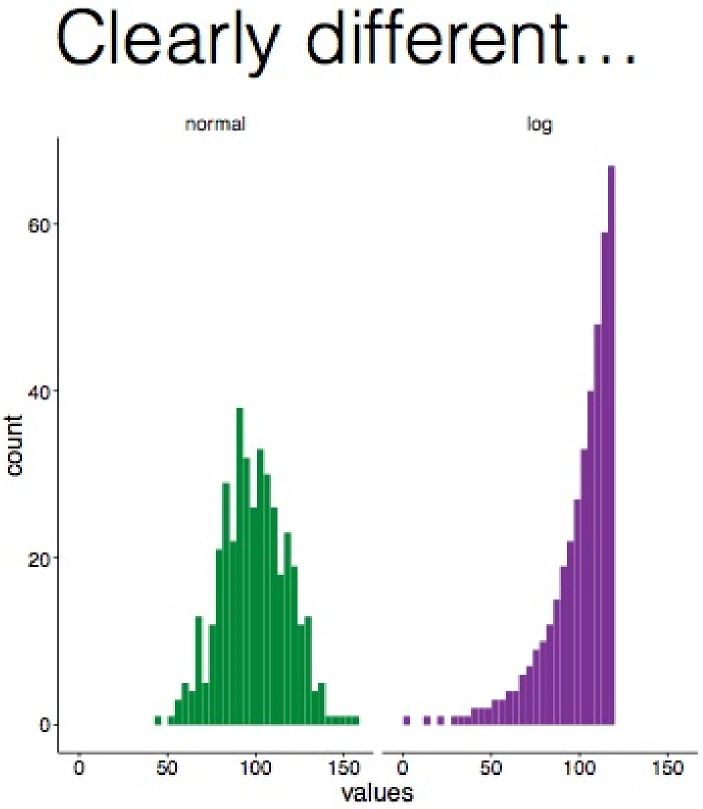
- Visualisations must accurately reflect the data
- Tell a story
- Look professional

The Good, the Bad & the Ugly



Choosing the right visual for your data

Choosing the right visual for your data

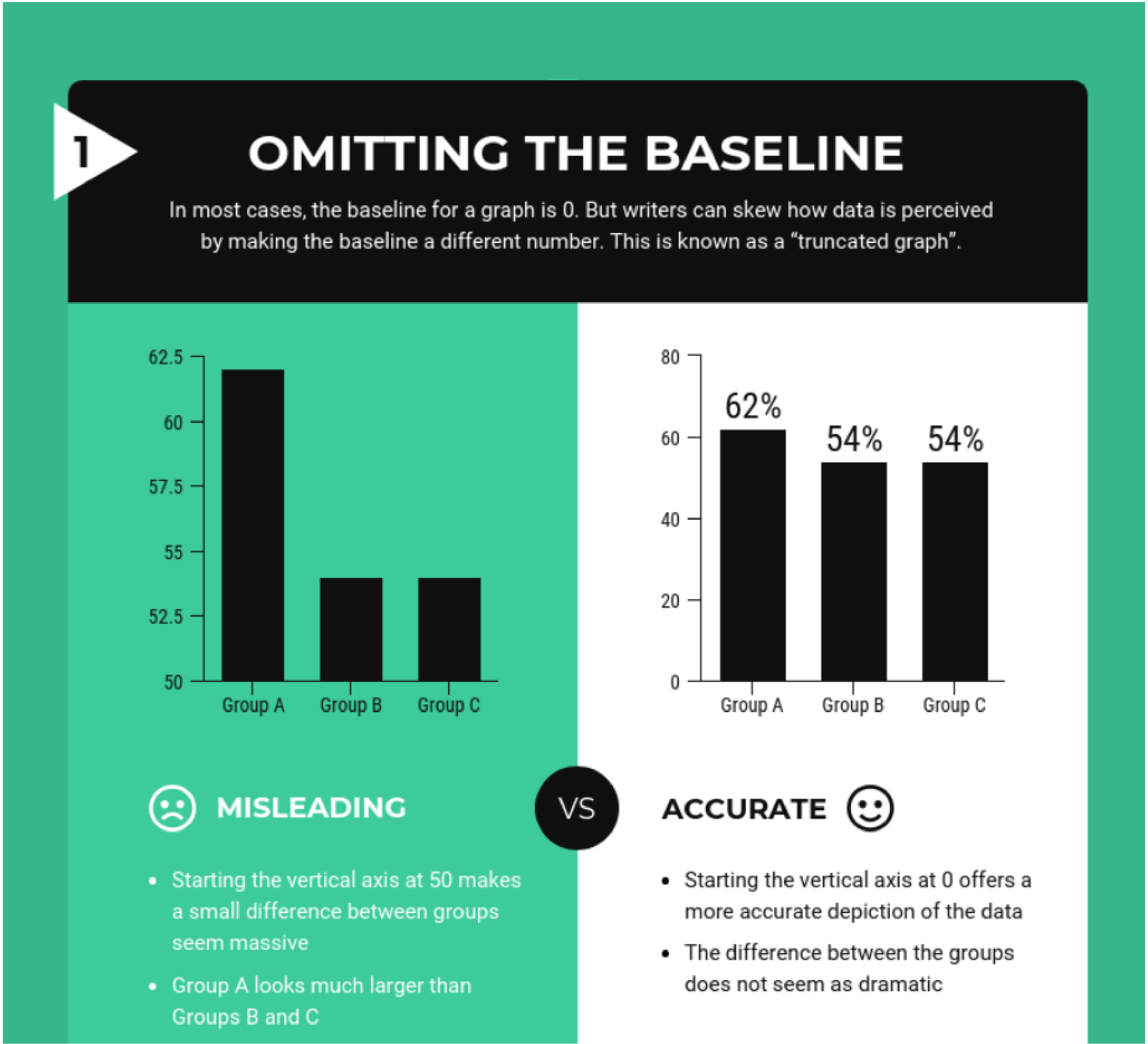


Choosing the right visual for your data

Choosing a data visual

- Choosing the right data visual *requires* understanding your data
- You **must** clearly explain any non-obvious features
- We will cover the different types of figures over the next few weeks

Five common ways graphs can mislead you



1 OMITTING THE BASELINE

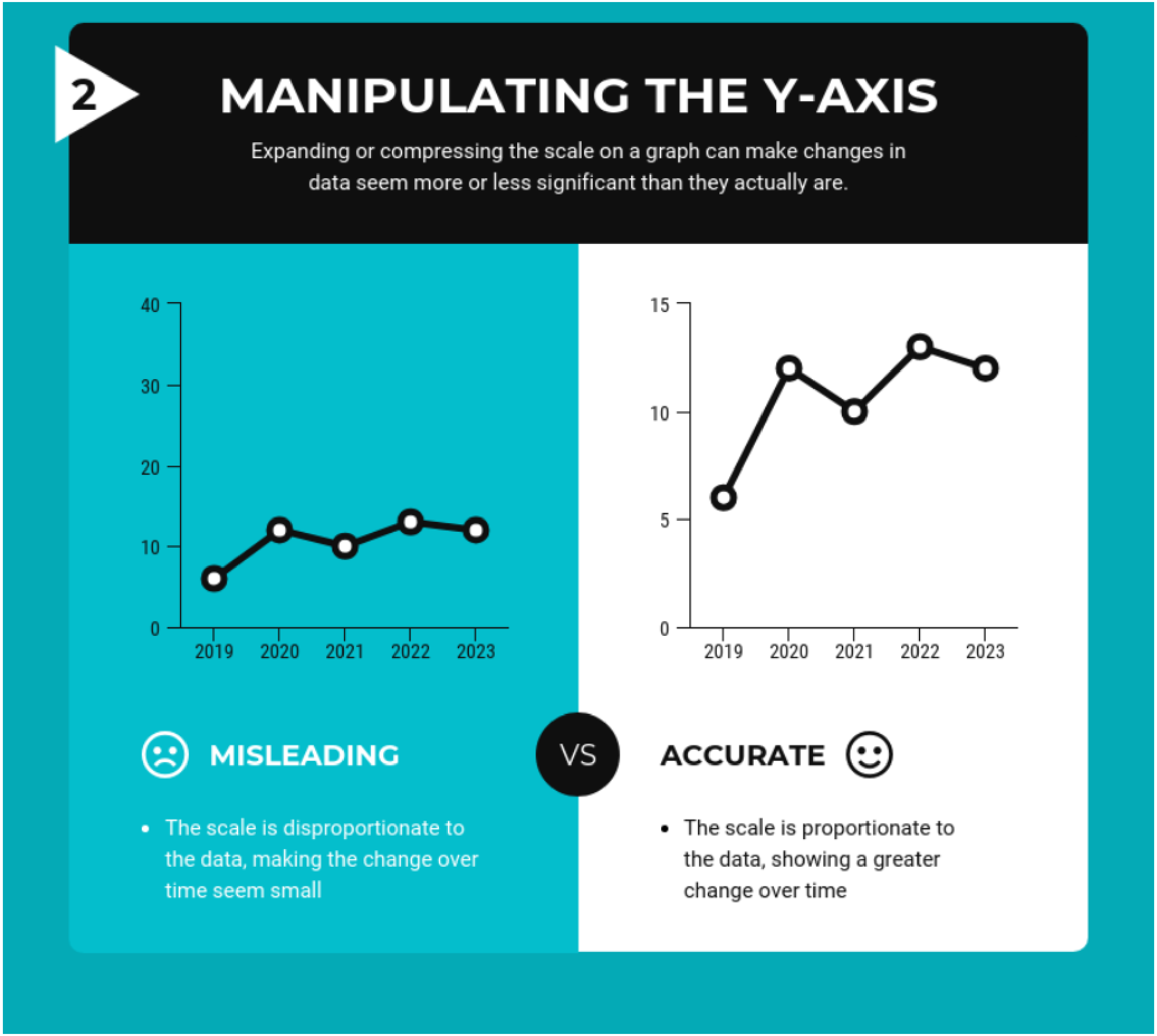
In most cases, the baseline for a graph is 0. But writers can skew how data is perceived by making the baseline a different number. This is known as a “truncated graph”.

MISLEADING

- Starting the vertical axis at 50 makes a small difference between groups seem massive
- Group A looks much larger than Groups B and C

ACCURATE

- Starting the vertical axis at 0 offers a more accurate depiction of the data
- The difference between the groups does not seem as dramatic



2

MANIPULATING THE Y-AXIS

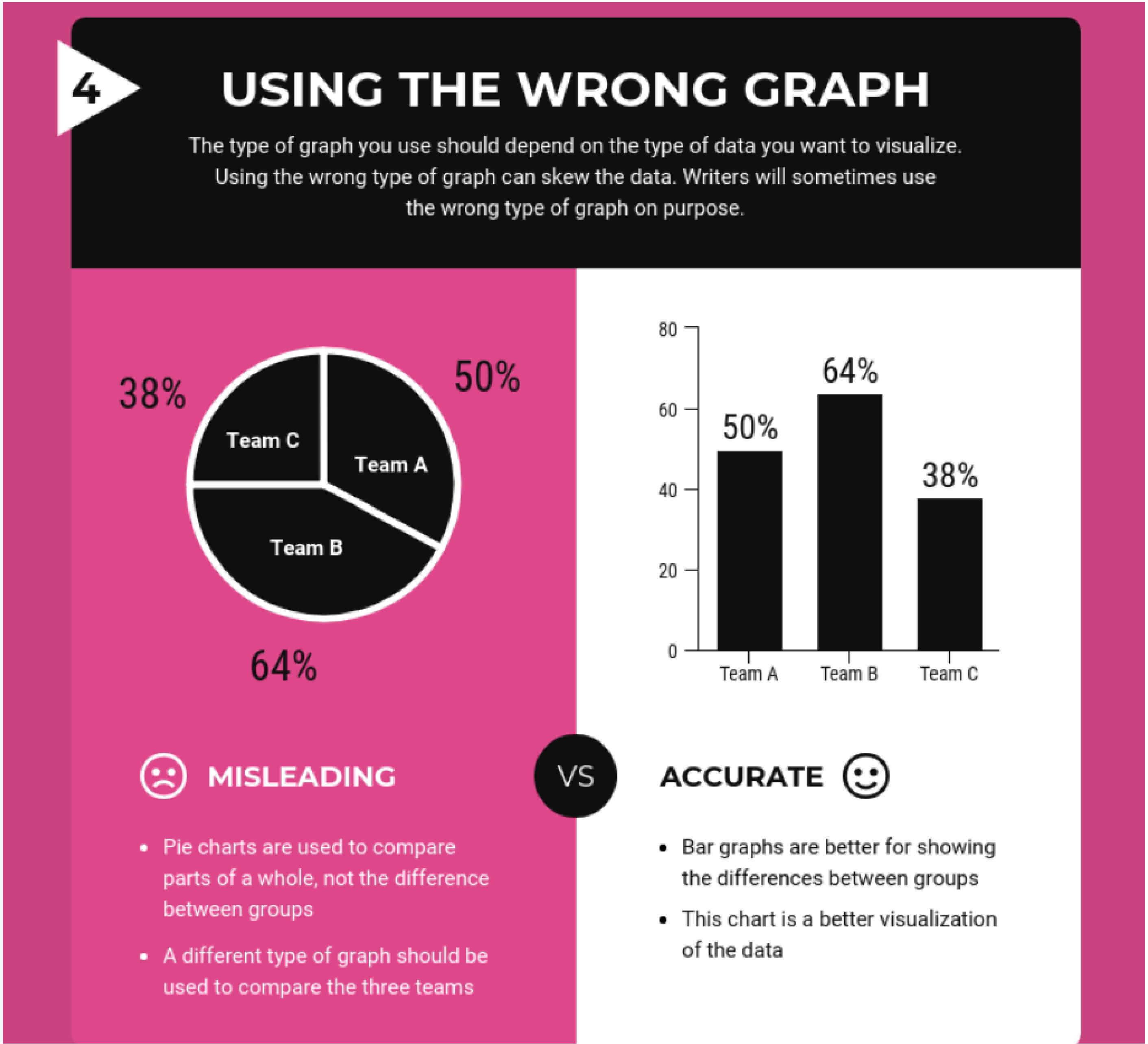
Expanding or compressing the scale on a graph can make changes in data seem more or less significant than they actually are.

MISLEADING

- The scale is disproportionate to the data, making the change over time seem small

ACCURATE

- The scale is proportionate to the data, showing a greater change over time



USING THE WRONG GRAPH

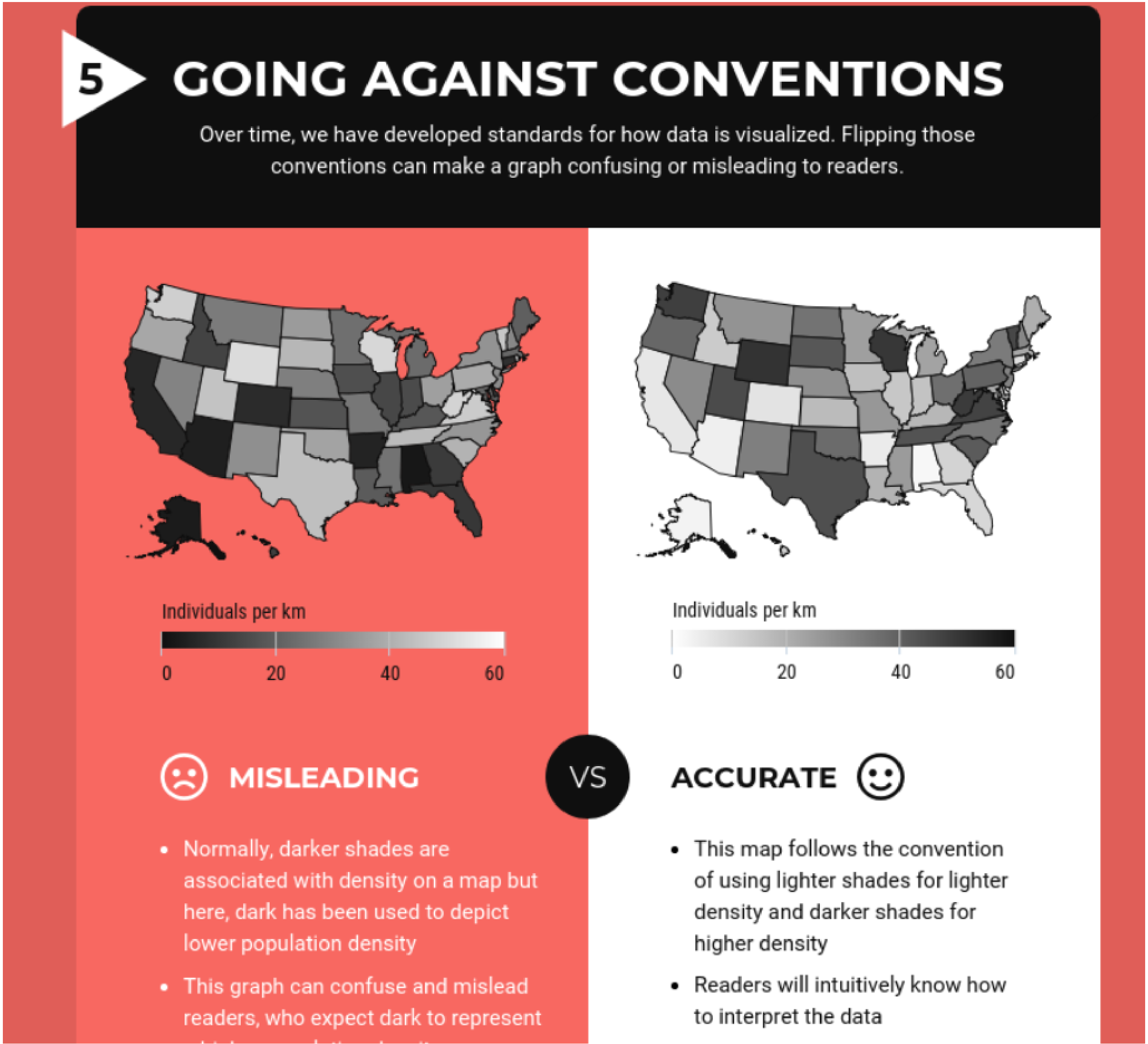
The type of graph you use should depend on the type of data you want to visualize. Using the wrong type of graph can skew the data. Writers will sometimes use the wrong type of graph on purpose.

MISLEADING

- Pie charts are used to compare parts of a whole, not the difference between groups
- A different type of graph should be used to compare the three teams

ACCURATE

- Bar graphs are better for showing the differences between groups
- This chart is a better visualization of the data



5 GOING AGAINST CONVENTIONS

Over time, we have developed standards for how data is visualized. Flipping those conventions can make a graph confusing or misleading to readers.

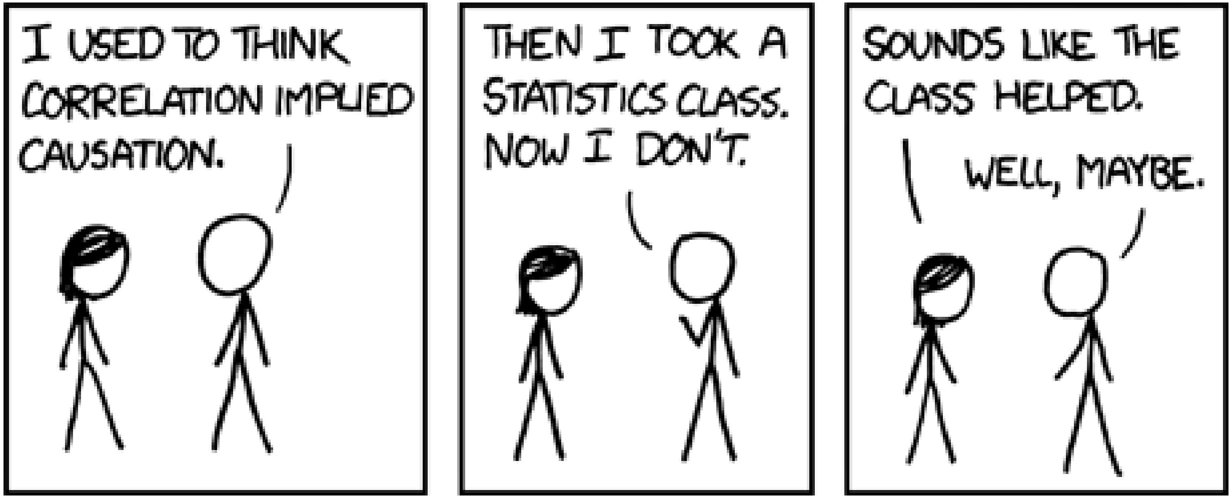
☹ MISLEADING

- Normally, darker shades are associated with density on a map but here, dark has been used to depict lower population density
- This graph can confuse and mislead readers, who expect dark to represent

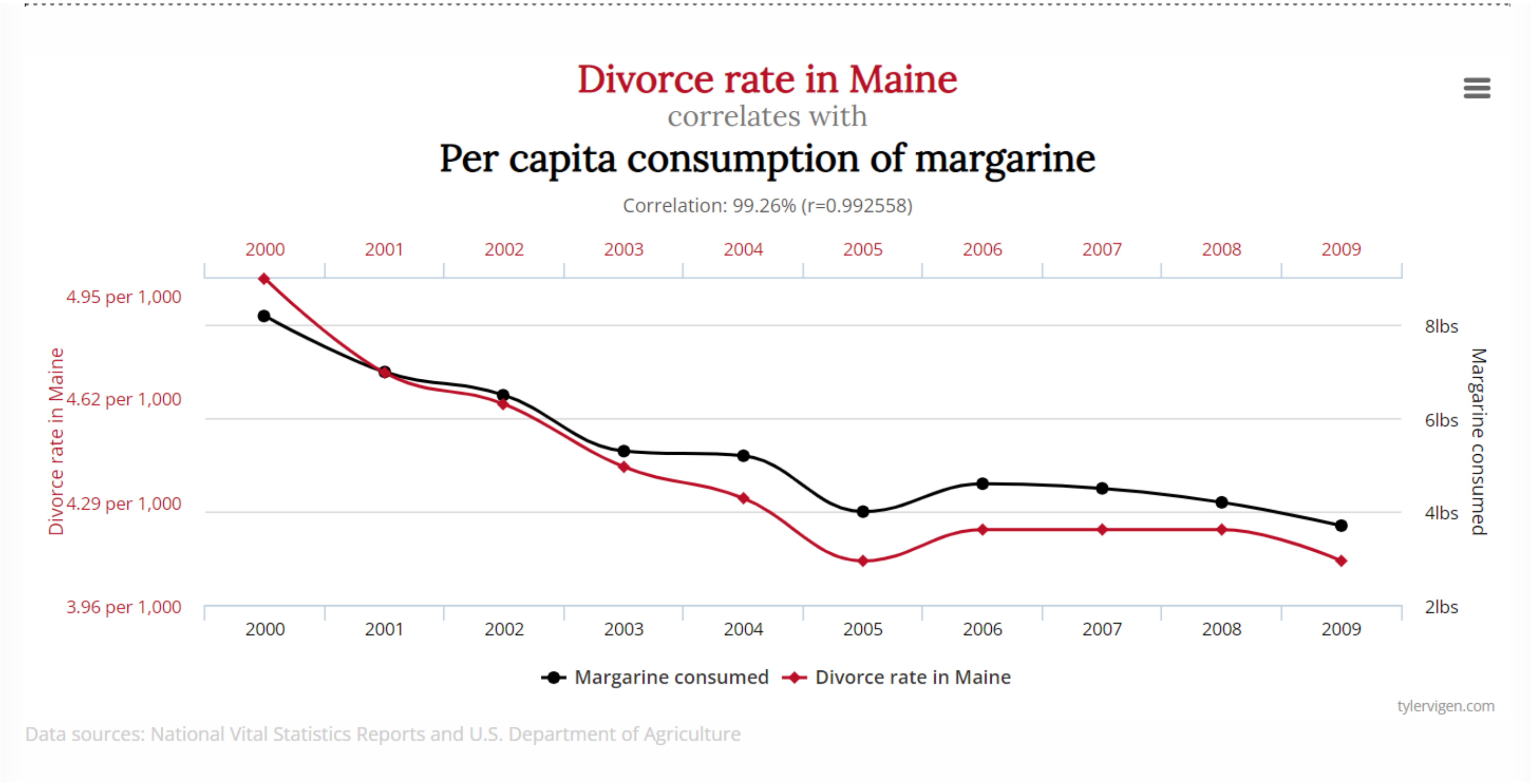
ACCURATE ☺

- This map follows the convention of using lighter shades for lighter density and darker shades for higher density
- Readers will intuitively know how to interpret the data

BONUS: Spurious correlation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

BONUS: Spurious correlation



Divorce rate in Maine

correlates with

Per capita consumption of margarine

Correlation: 99.26% (r=0.992558)

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

http://tylervigen.com/spurious-correlations

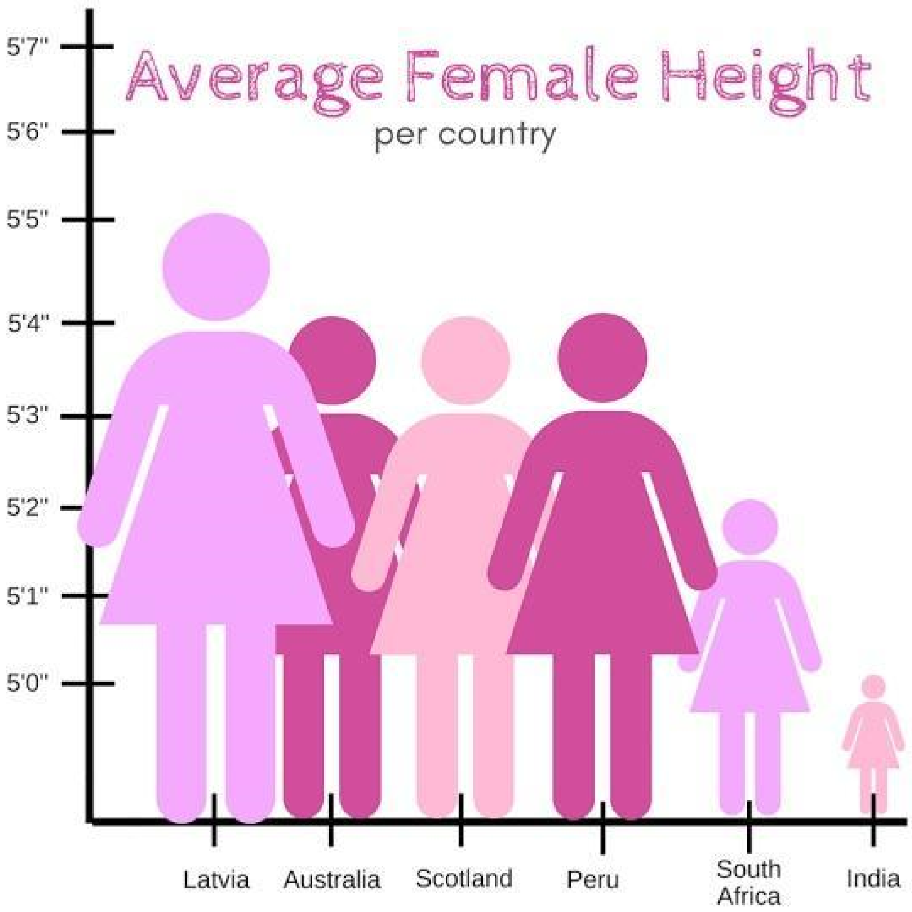
BONUS: Should have been a log scale?

BONUS: Should have been a log scale?

Caveat - does everyone undestand logs? More on this in future sessions

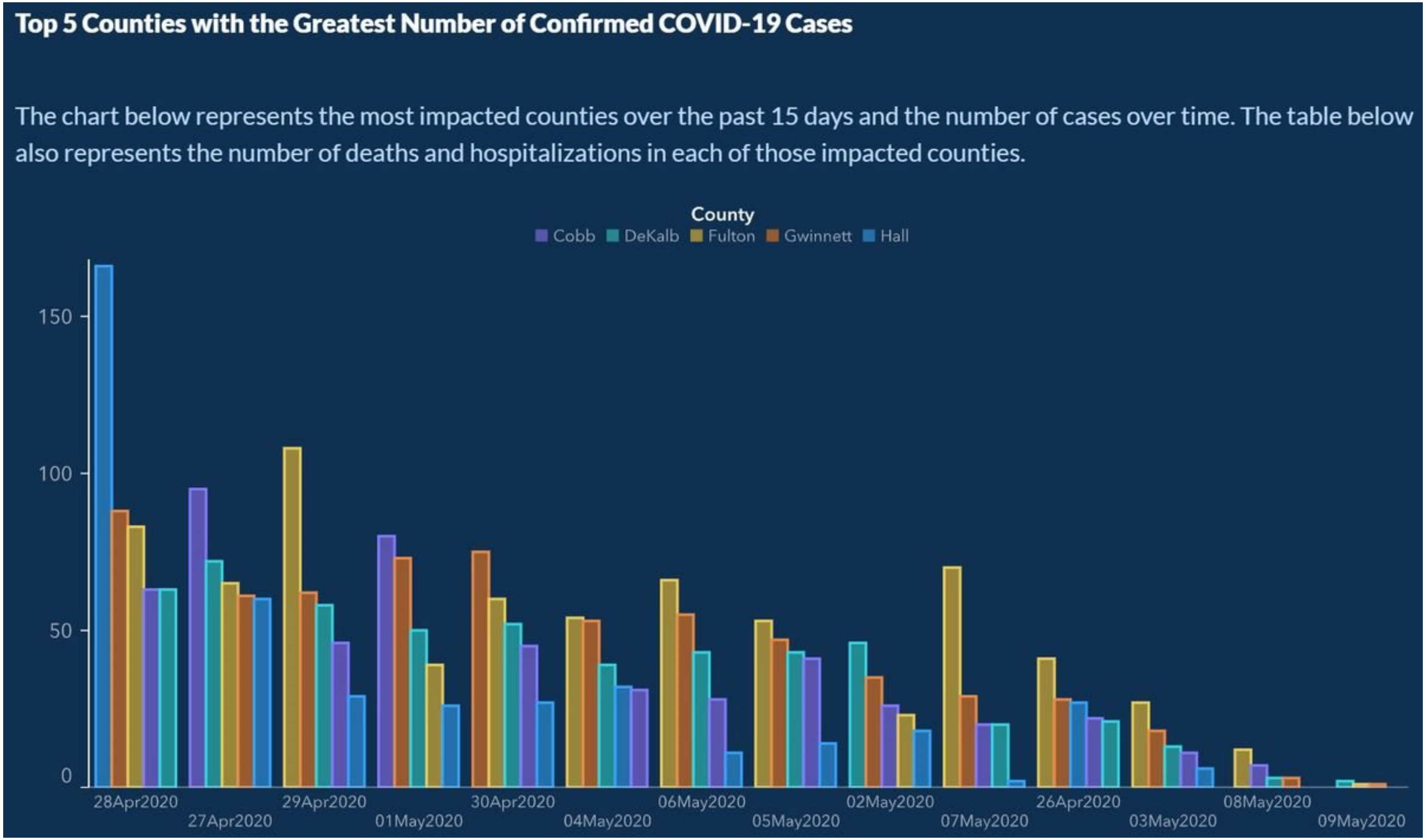
Graph Crimes

Graph Crimes



Graph Crimes

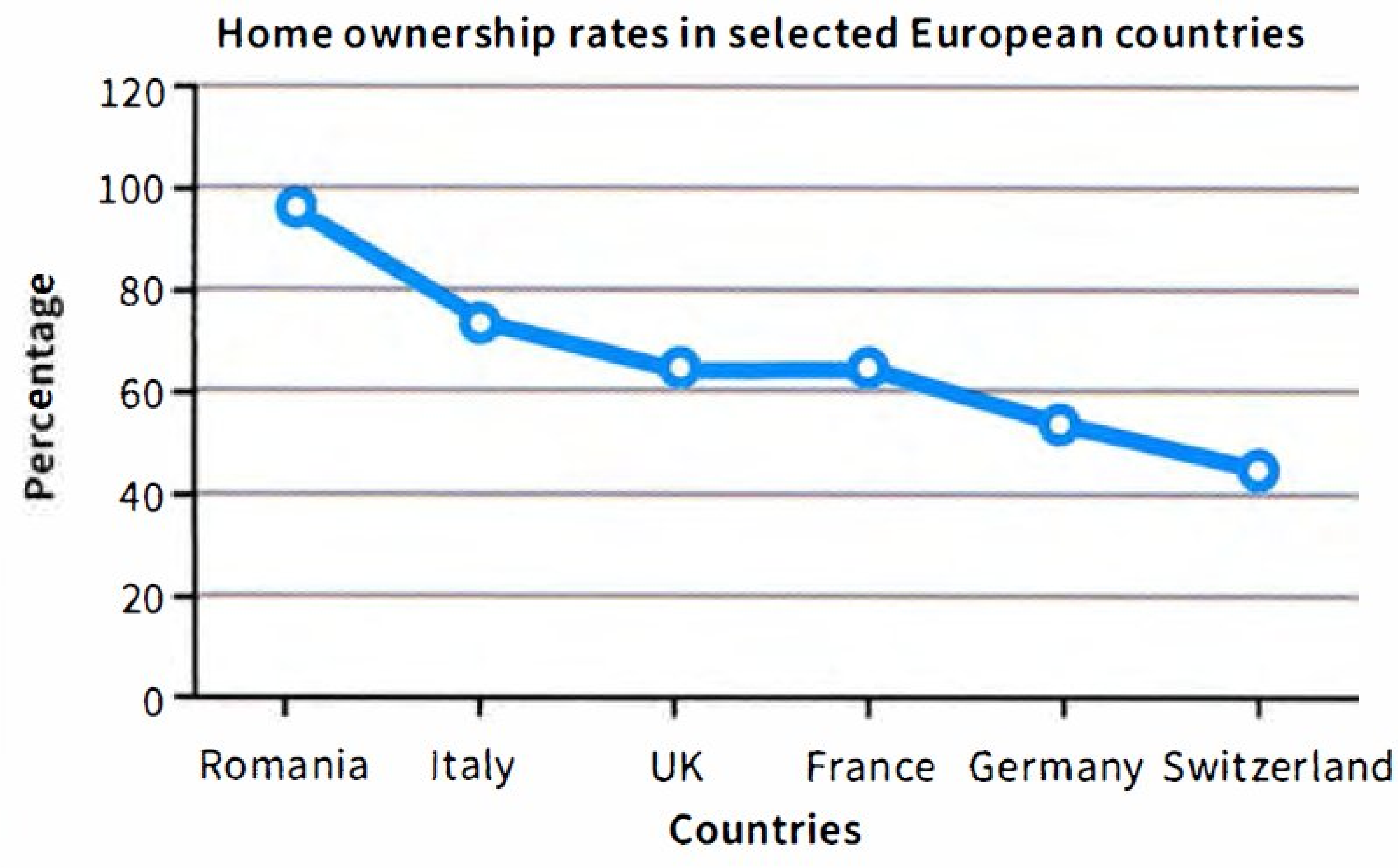
Graph Crimes



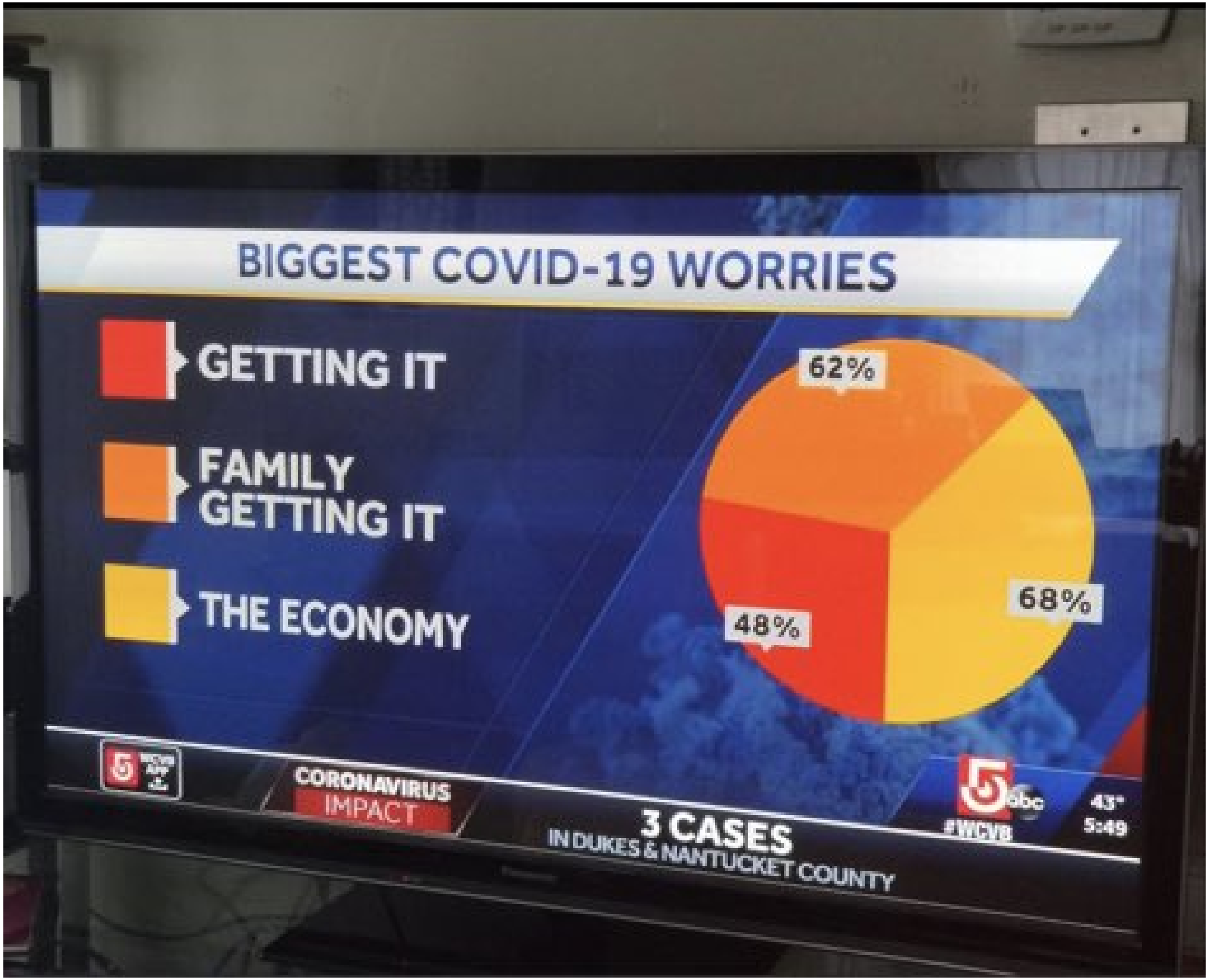
Top 5 Counties with the Greatest Number of Confirmed COVID-19 Cases

The chart below represents the most impacted counties over the past 15 days and the number of cases over time. The table below also represents the number of deaths and hospitalizations in each of those impacted counties.

Graph Crimes



Graph Crimes

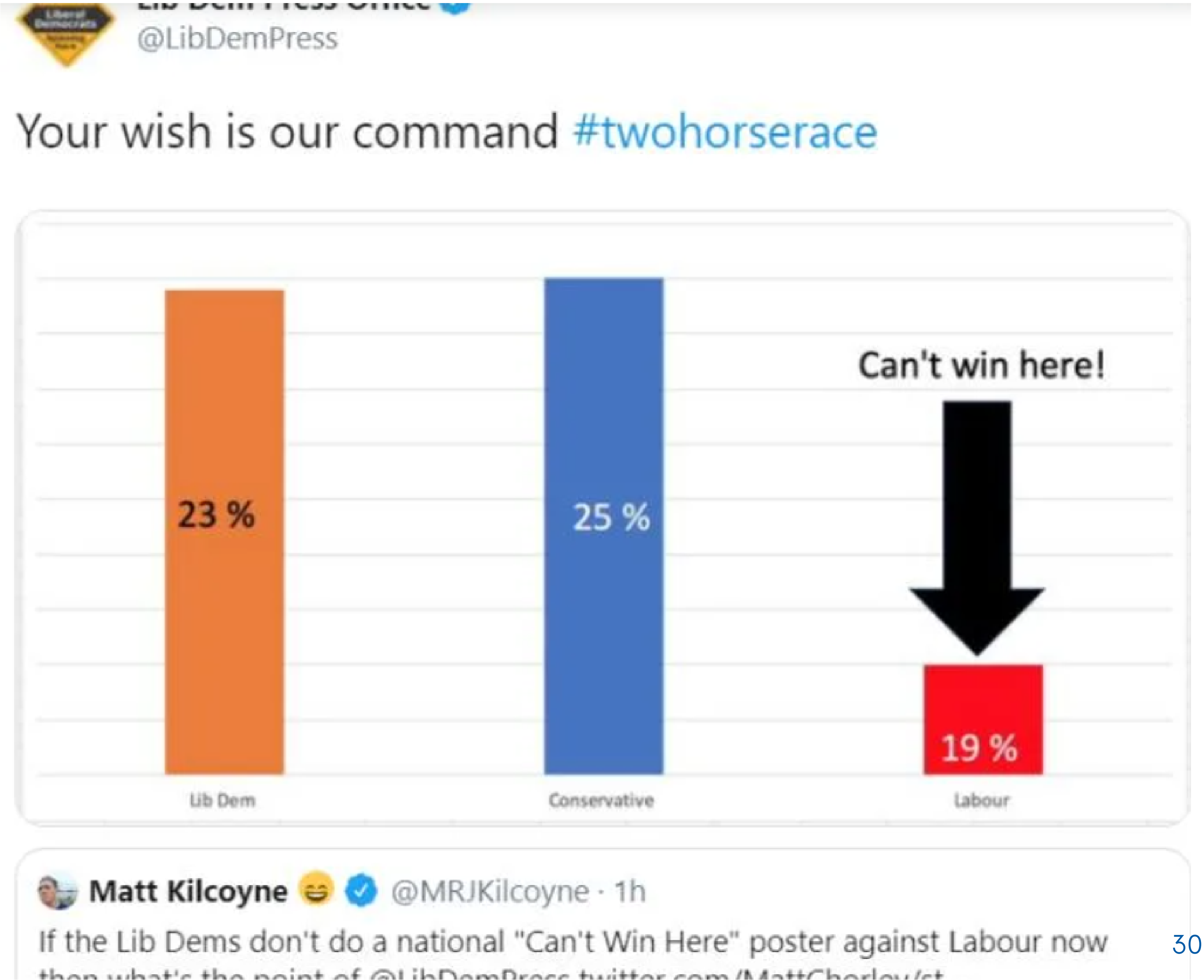


Graph Crimes

Graph Crimes

Graph Crimes

Lib Dem party *dislikes* accurate graphs

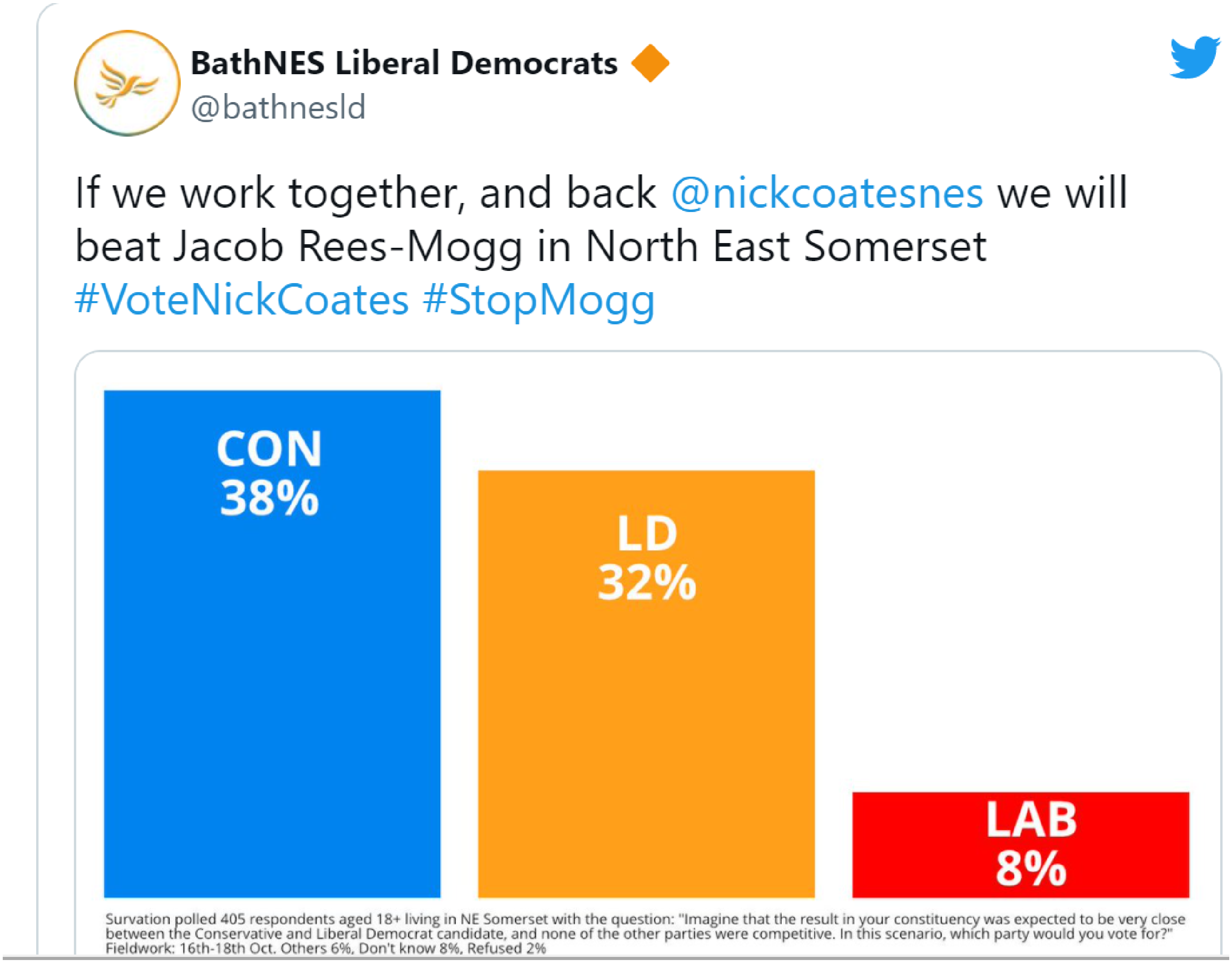


Lib Dem Press Office
@LibDemPress

Your wish is our command #twohorserace

Matt Kilcoyne 😄 @MRJKilcoyne · 1h

If the Lib Dems don't do a national "Can't Win Here" poster against Labour now then what's the point of @LibDemPress twitter.com/MattChorley/st



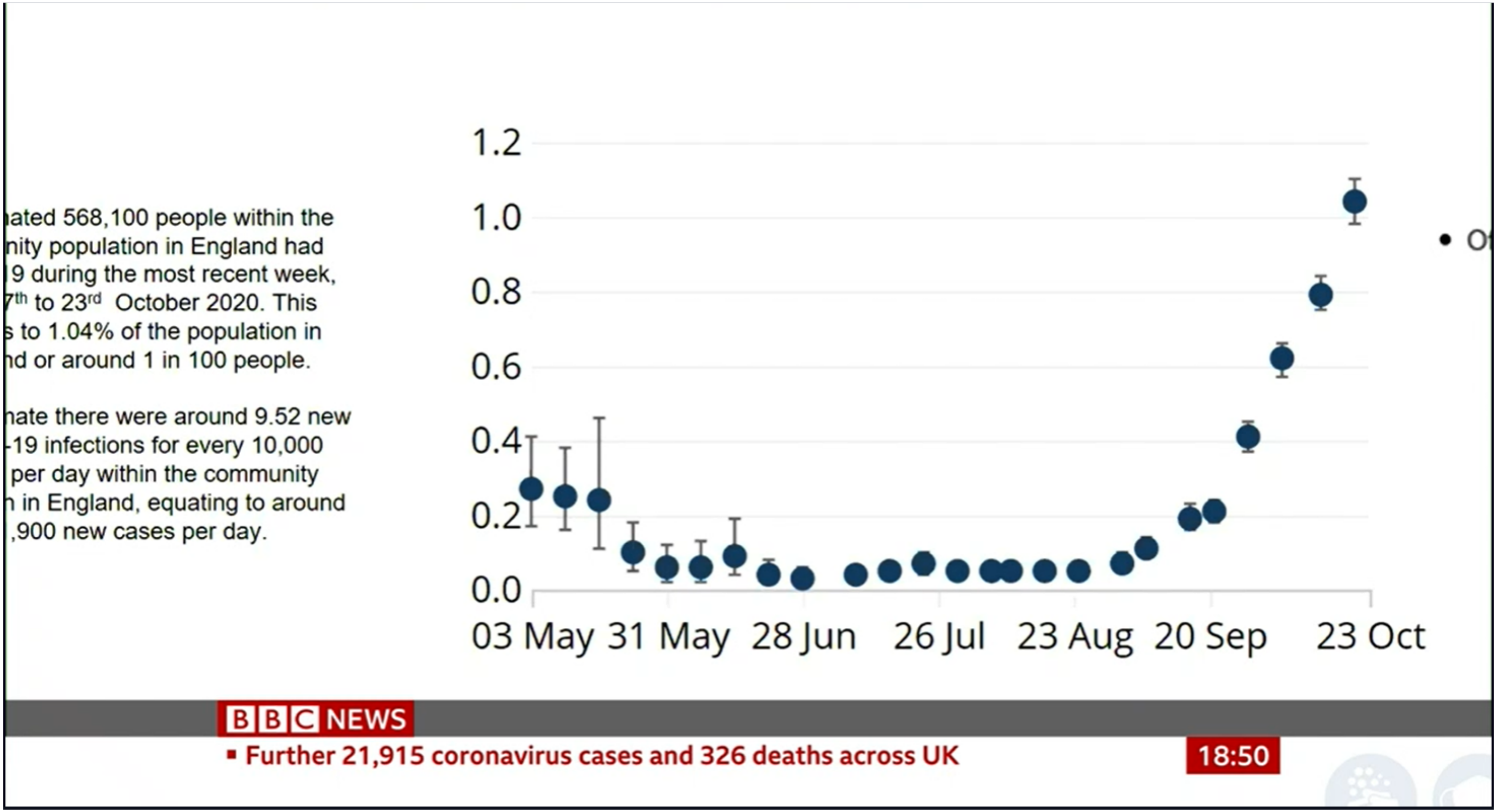
BathNES Liberal Democrats
@bathnesld

If we work together, and back @nickcoatesnes we will beat Jacob Rees-Mogg in North East Somerset #VoteNickCoates #StopMogg

Survation polled 405 respondents aged 18+ living in NE Somerset with the question: "Imagine that the result in your constituency was expected to be very close between the Conservative and Liberal Democrat candidate, and none of the other parties were competitive. In this scenario, which party would you vote for?" Fieldwork: 16th-18th Oct. Others 6%, Don't know 8%, Refused 2%

Government Daily Briefings 2020

Daily Briefings via the BBC



ıated 568,100 people within the
nity population in England had
9 during the most recent week,
7th to 23rd October 2020. This
s to 1.04% of the population in
nd or around 1 in 100 people.

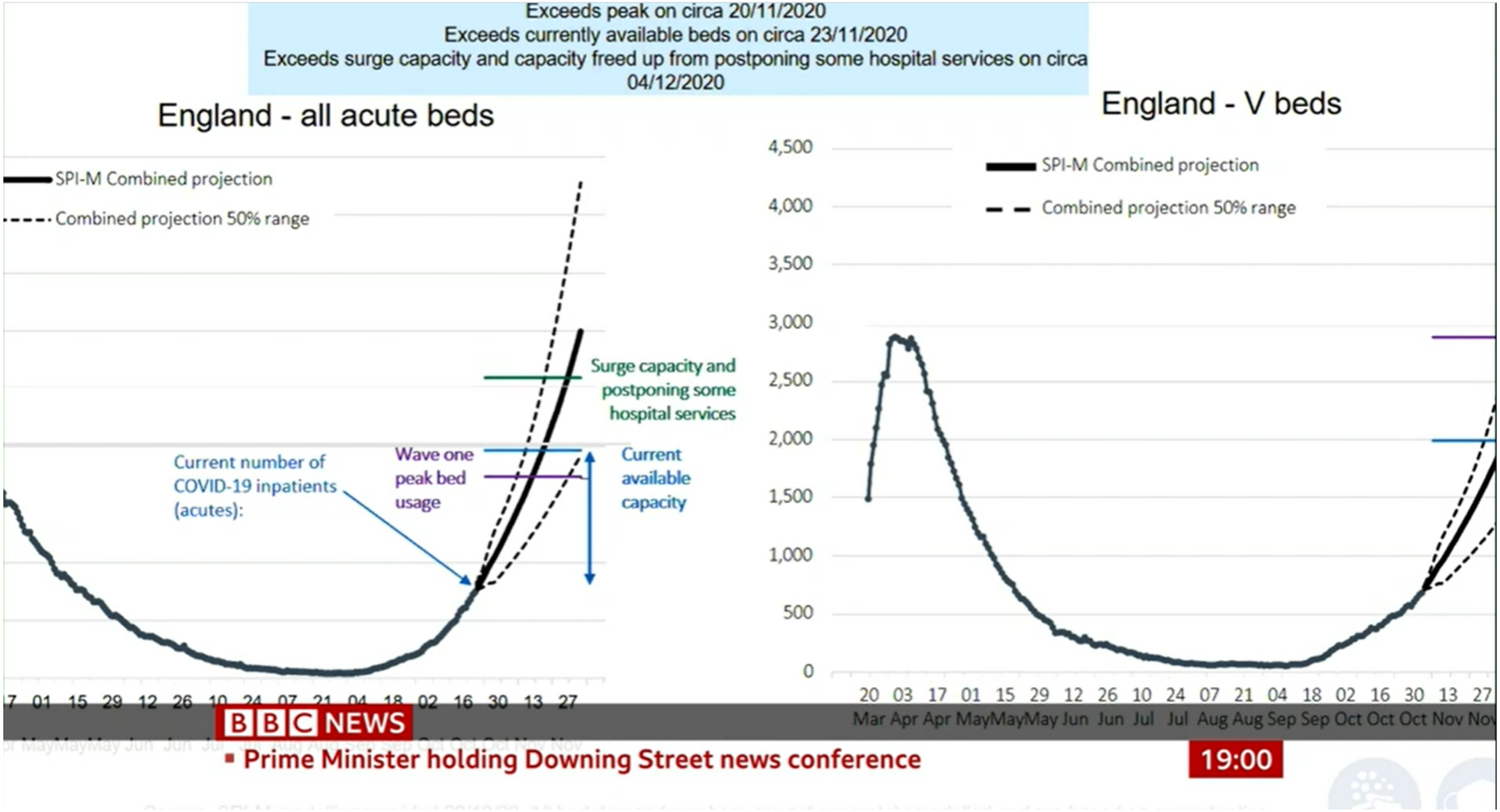
nate there were around 9.52 new
-19 infections for every 10,000
per day within the community
n in England, equating to around
,900 new cases per day.

BBC NEWS

▪ Further 21,915 coronavirus cases and 326 deaths across UK

18:50

Daily Briefings via the BBC



Exceeds peak on circa 20/11/2020
Exceeds currently available beds on circa 23/11/2020
Exceeds surge capacity and capacity freed up from postponing some hospital services on circa 04/12/2020

England - all acute beds

England - V beds

BBC NEWS

▪ Prime Minister holding Downing Street news conference

19:00

SPI-M Combined projection
Combined projection 50% range
Surge capacity and postponing some hospital services
Current number of COVID-19 inpatients (acutes):
Wave one peak bed usage
Current available capacity
17 01 15 29 12 26 10 24 07 21 04 18 02 16 30 13 27
Apr May May May Jun Jun Jul Jul Aug Aug Sep Sep Oct Oct Oct Nov Nov
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
SPI-M Combined projection
Combined projection 50% range
20 03 17 01 15 29 12 26 10 24 07 21 04 18 02 16 30 13 27
Mar Apr Apr May May May Jun Jun Jul Jul Aug Aug Sep Sep Oct Oct Oct Nov Nov

♻

Essential Reading:

Fundamentals of Data Visualisation - Claus O. Wilke

R Graphics Cookbook - Winston Chang

A ggplot tutorial for beautiful plotting in R - Cédric Scherer