Bad Graphs

And Good Principles

Miguel Salema

Católica-Lisbon SBE

October 1, 2023

Section 1

Good Plots Matter

Anscombe's Quartet

These four datasets have the same conventional statistics and lead to the same linear model. However, the data are quite different.

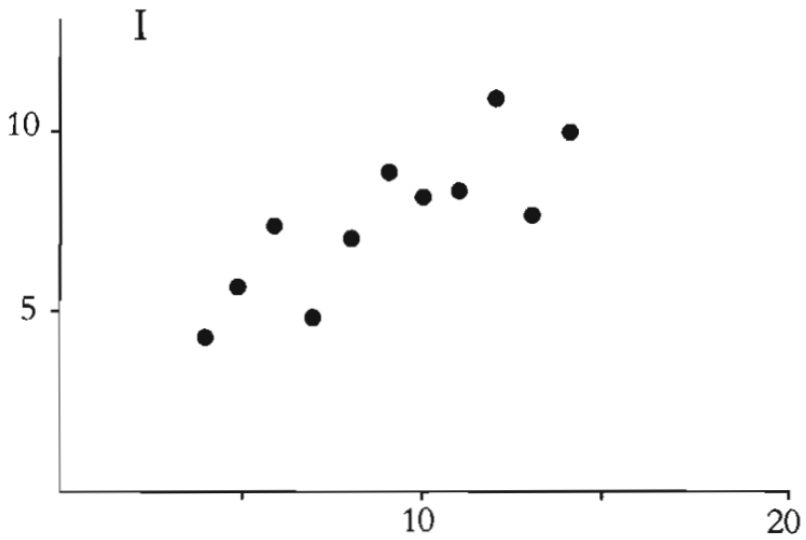
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

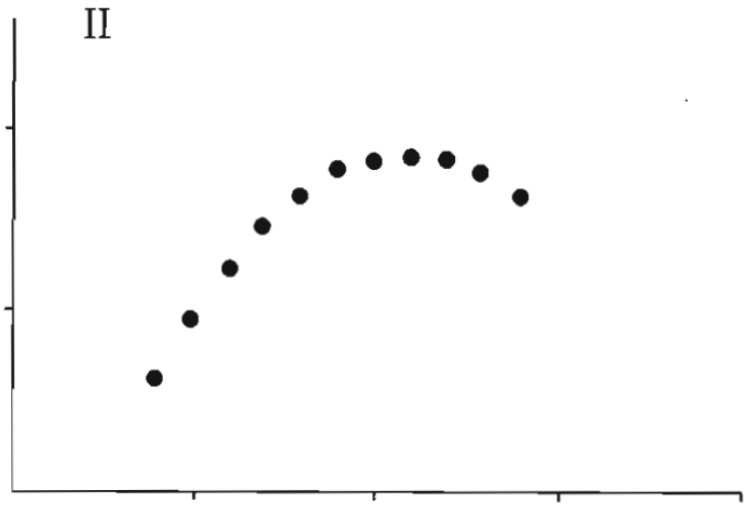
$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$
$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$

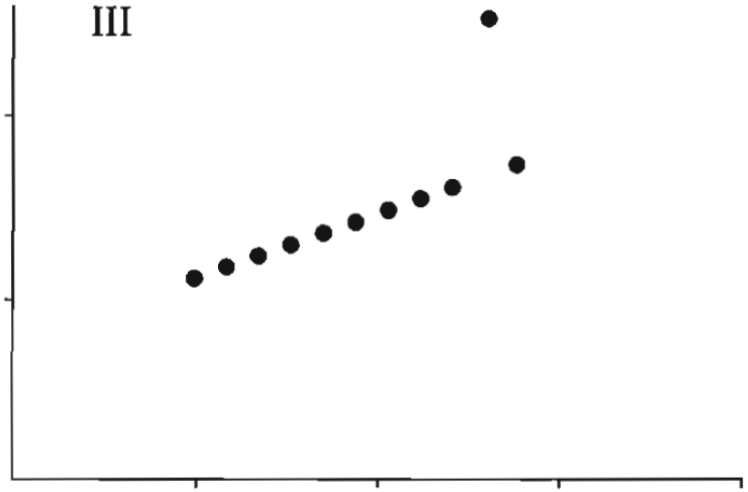
Solution: Plot your data

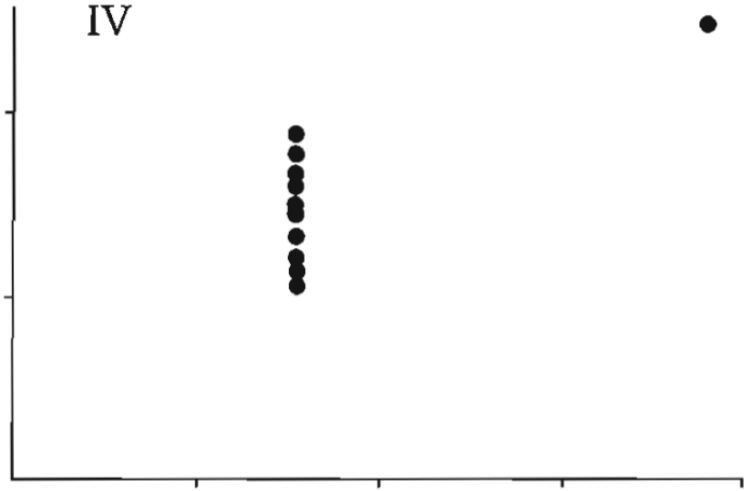
Graphics *can* be more precise than conventional statistics, contrary to popular belief.

Graphics Reveal Data



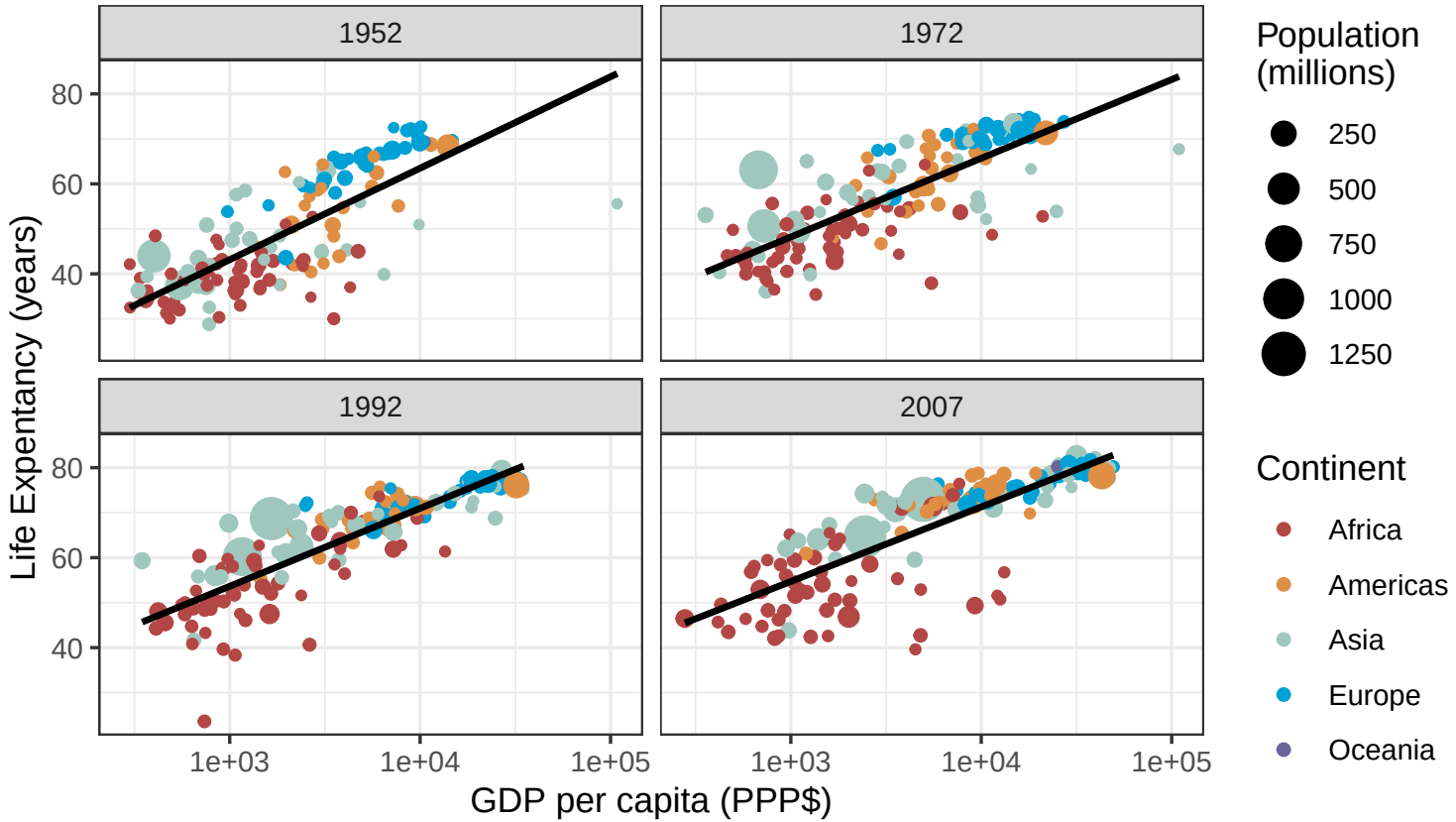






Previously Seen in Our Classes...

We already talked about how to do good graphs.

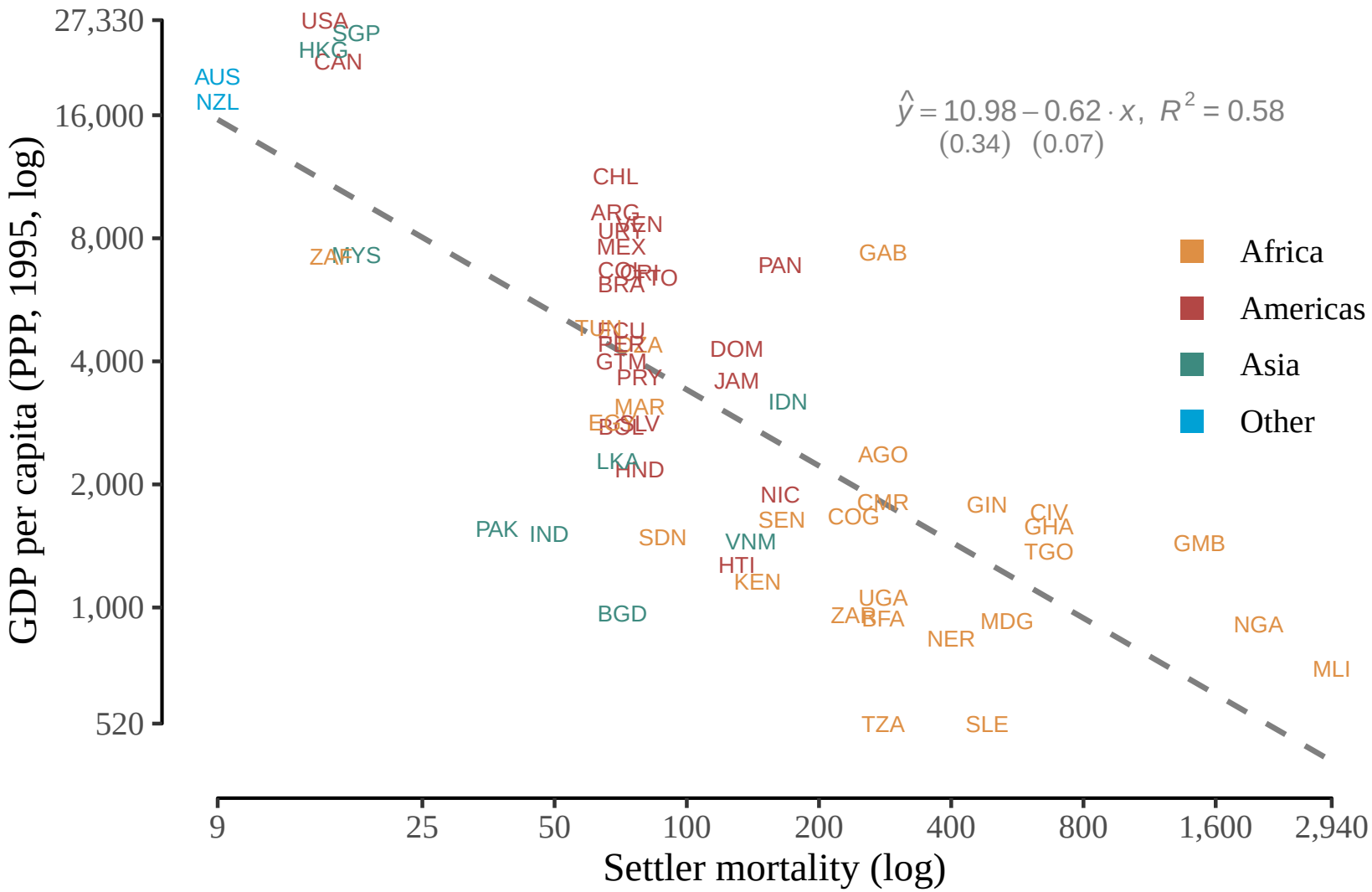


But can you detect a bad plot?

Code of previous plot

```
library(tidyverse)
library(gapminder)
gapminder %>%
  filter(year %in% c(1952, 1972, 1992, 2007)) %>%
  ggplot(aes(y = lifeExp,
             x = gdpPercap)) +
  geom_point(aes(size = pop/1000000,
                 color = continent)) +
  scale_x_log10() +
  labs(x = "GDP per capita (PPP$)",
       y = "Life Expentancy (years)",
       color = "Continent",
       size = "Population\n(millions)") +
  facet_wrap(vars(year)) +
  scale_color_manual(values = c("#b34745",
                                "#de8f44",
                                "#A0C7BE",
                                "#00a1d5",
                                "#696598")) +
  geom_smooth(method = "lm",
              se = FALSE,
              color = "black") +
  theme_bw()
```

My plot from AJR (2001)



Section 2

Graphical Integrity

Don't Be a Liar

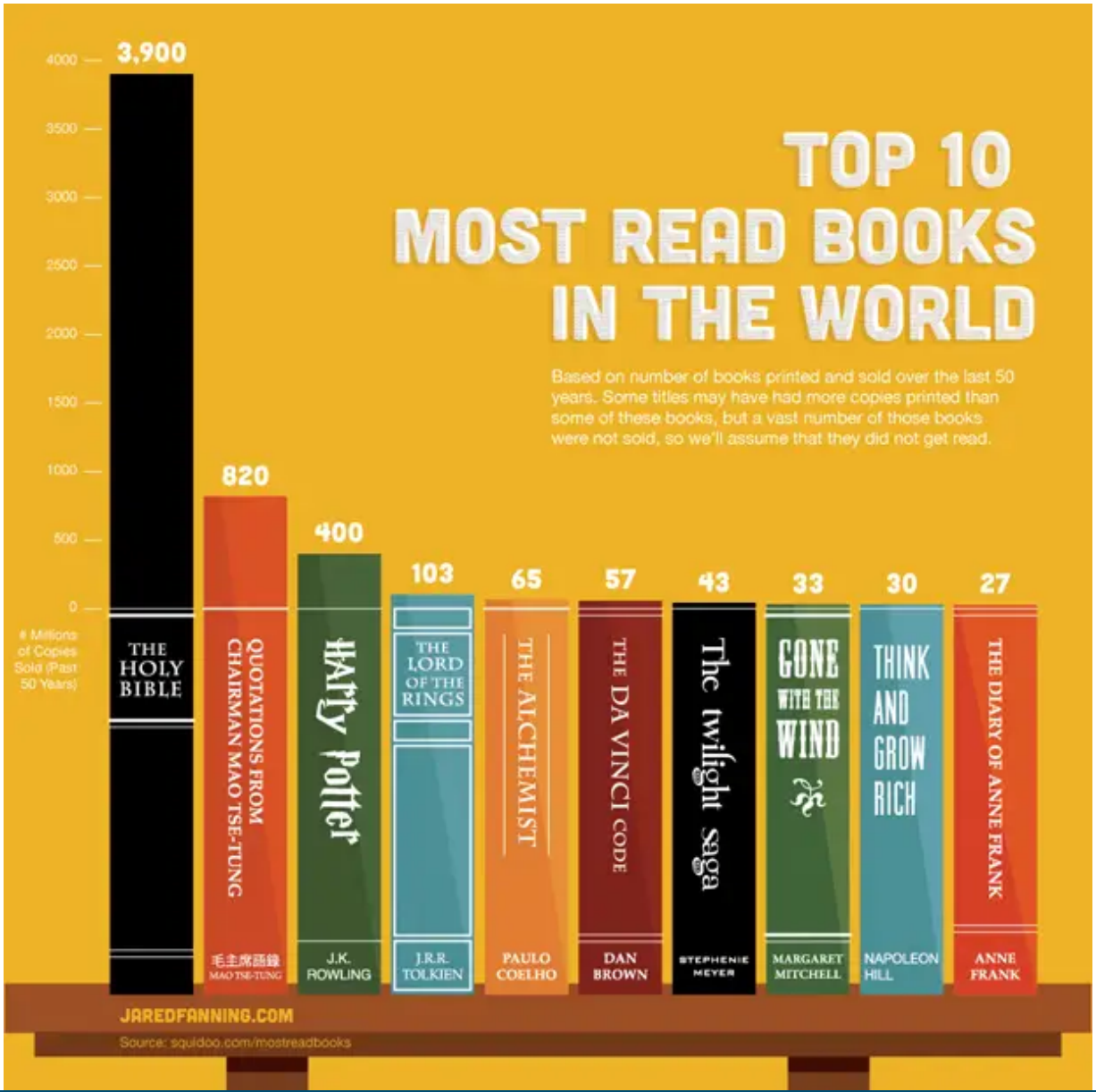
Many people think about statistical graphs as a tool to lie. However, graphs are no different than words in this regard. We just need to follow the principles of graphical integrity.

Follow the rule

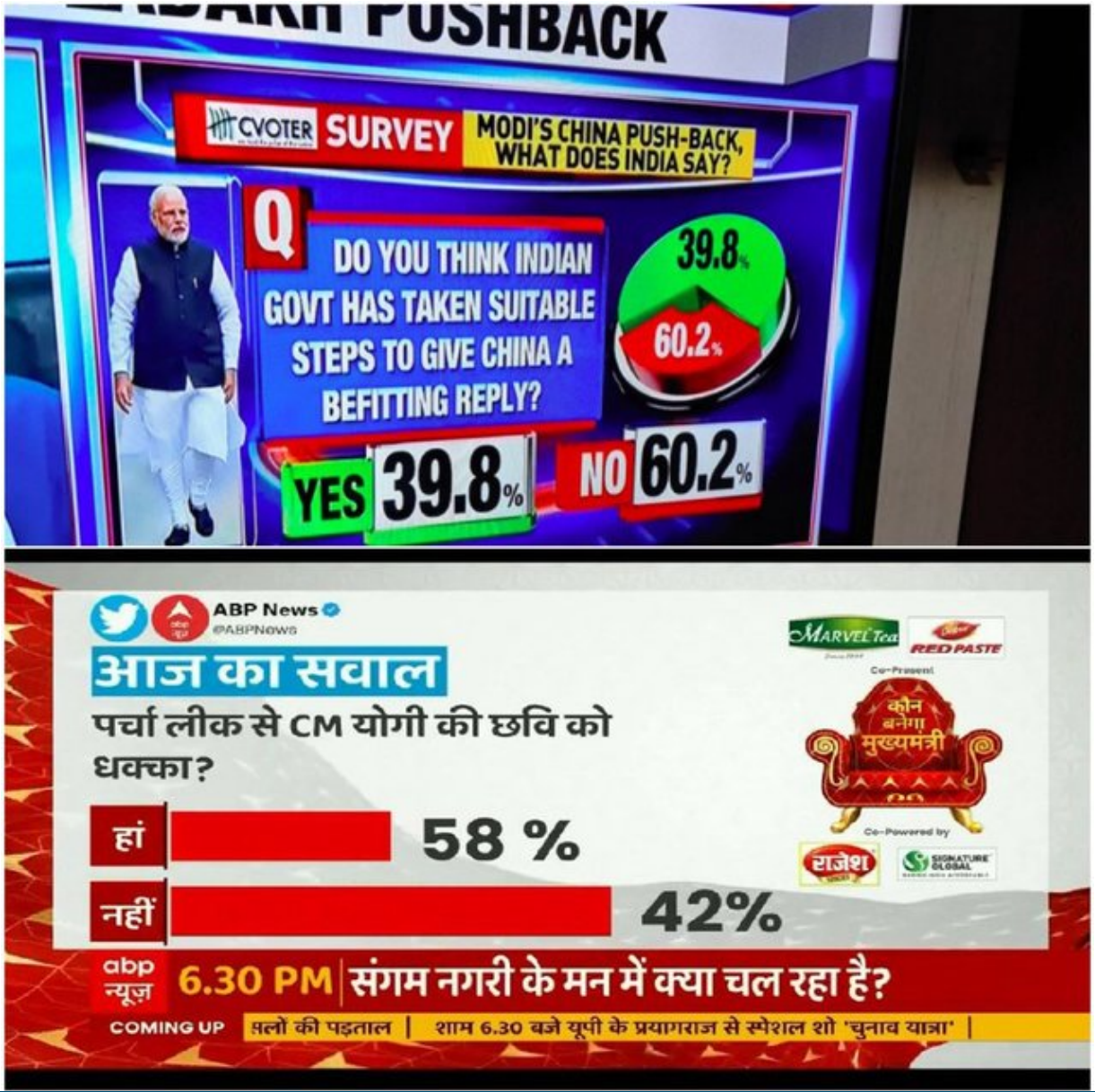
$$\text{Lie Factor} = \frac{\text{size of the effect shown in graphic}}{\text{size of the effect in data}}$$

The lie factor must be between 0.95 and 1.05 for the graphic to do a reasonable job at representing data.

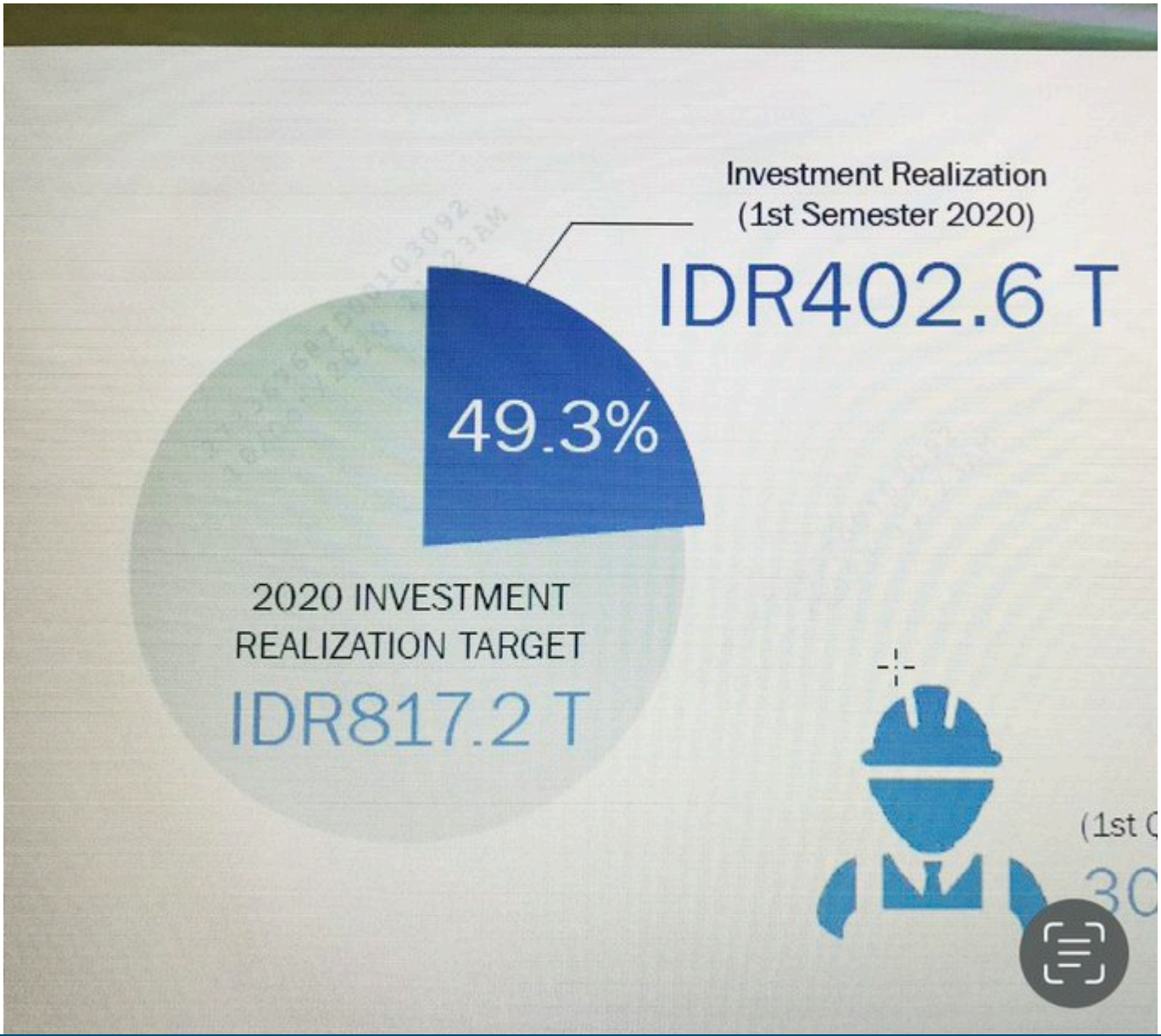
Present Correct Proportions I



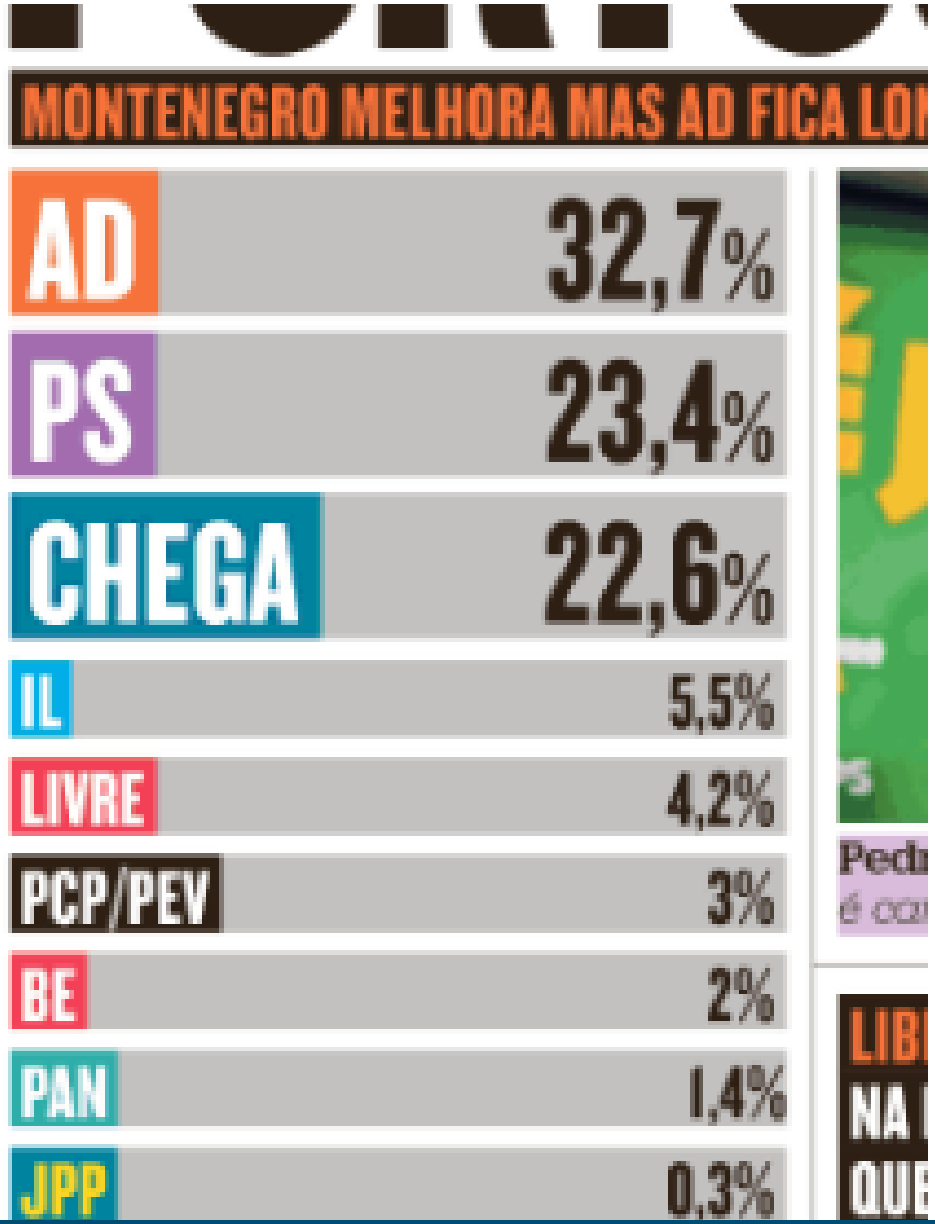
Present Correct Proportions II



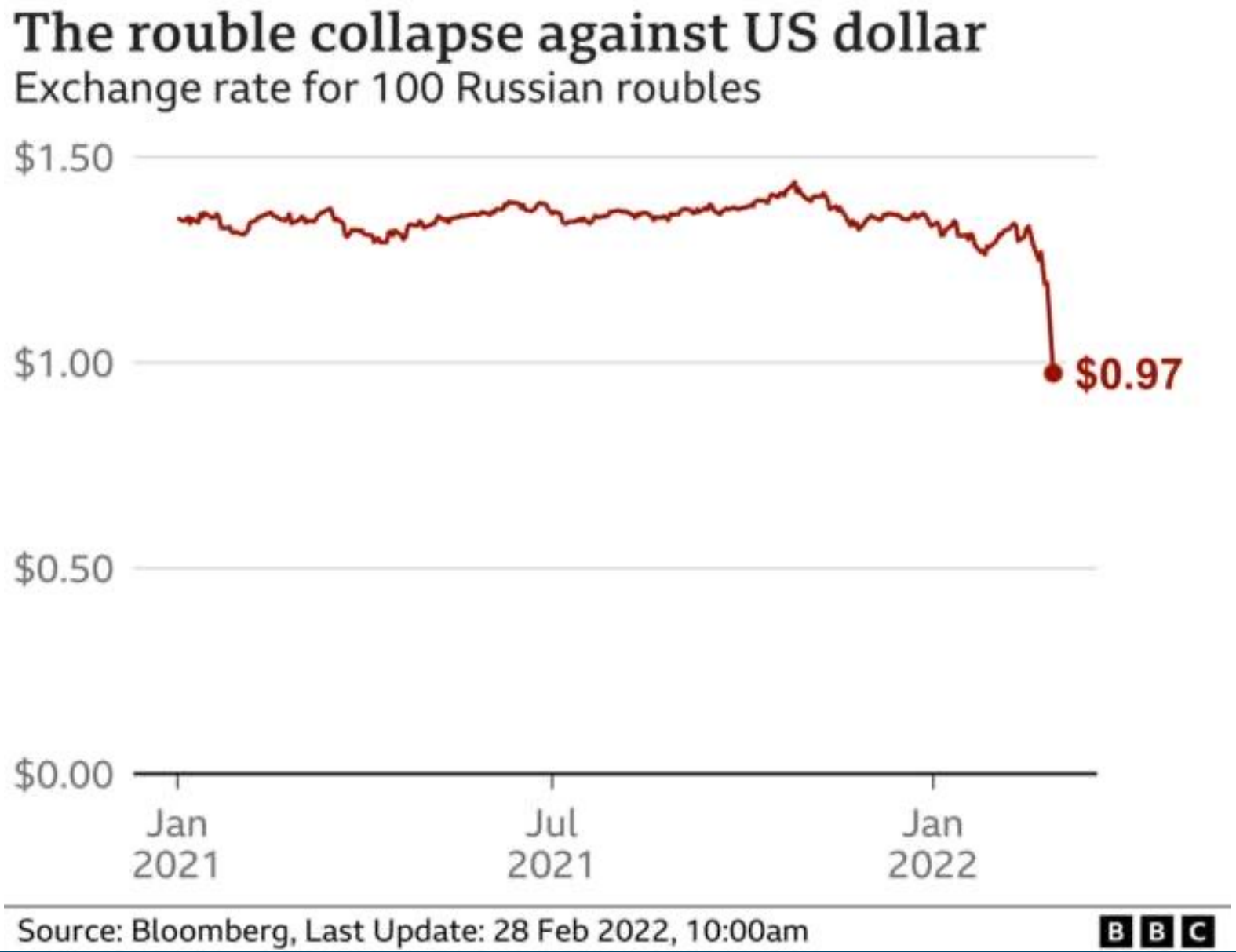
Present Correct Proportions III



Present Correct Proportions IV



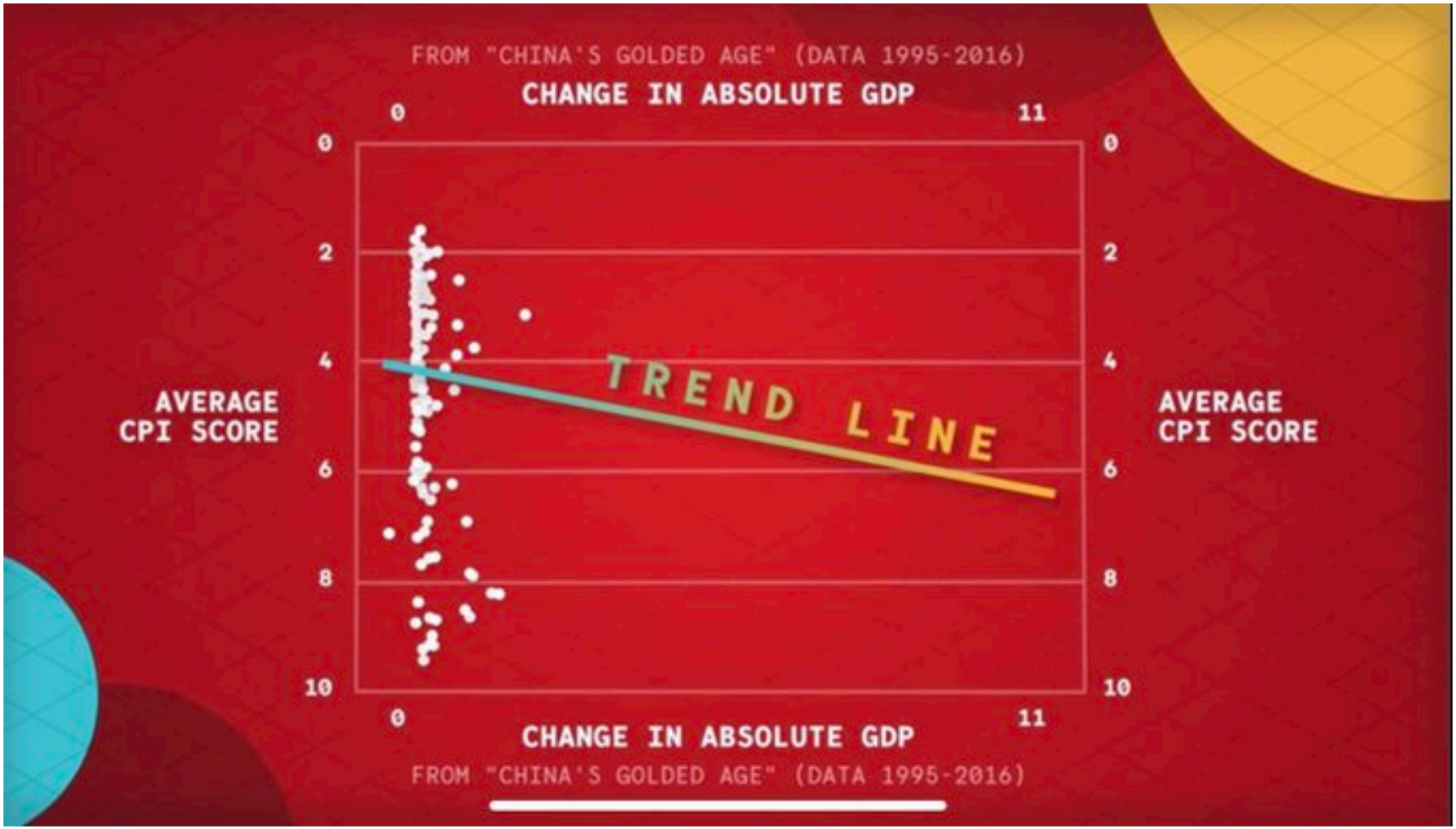
Line Plots Don't Have to Start at Zero...



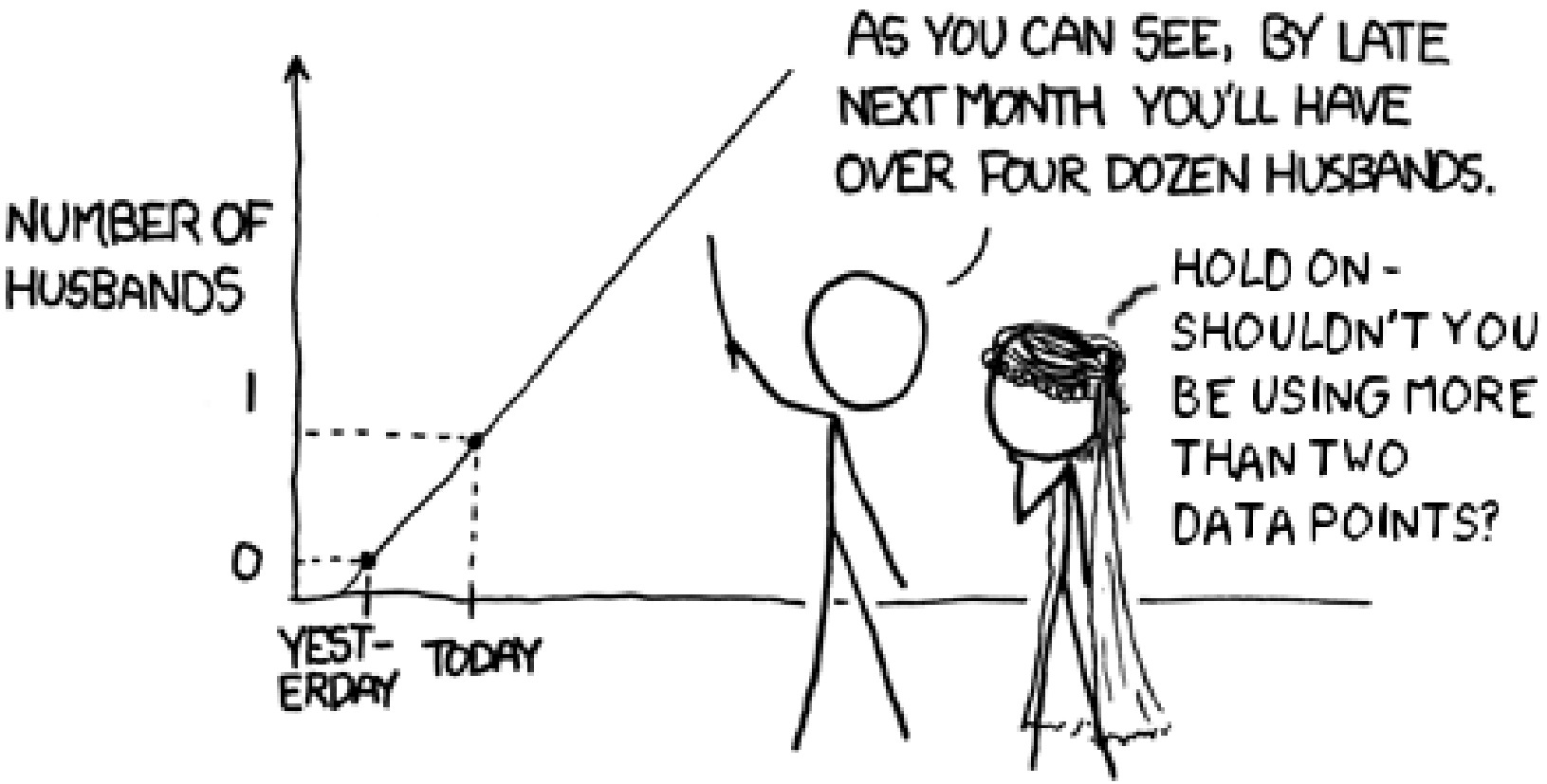
... But Bar Plots Generally Do

Same Scale for Side by Side Plots

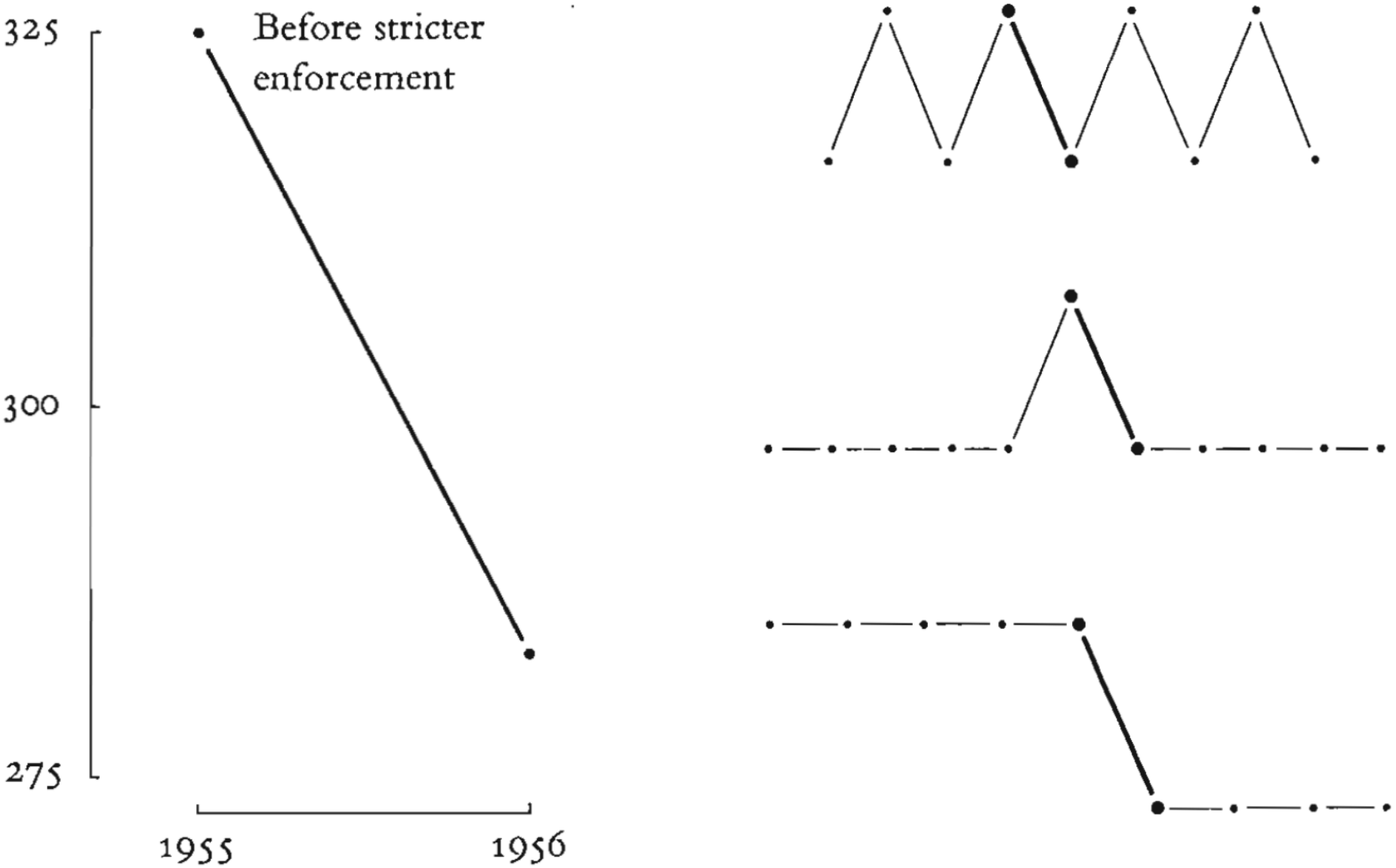
Don't Extrapolate I



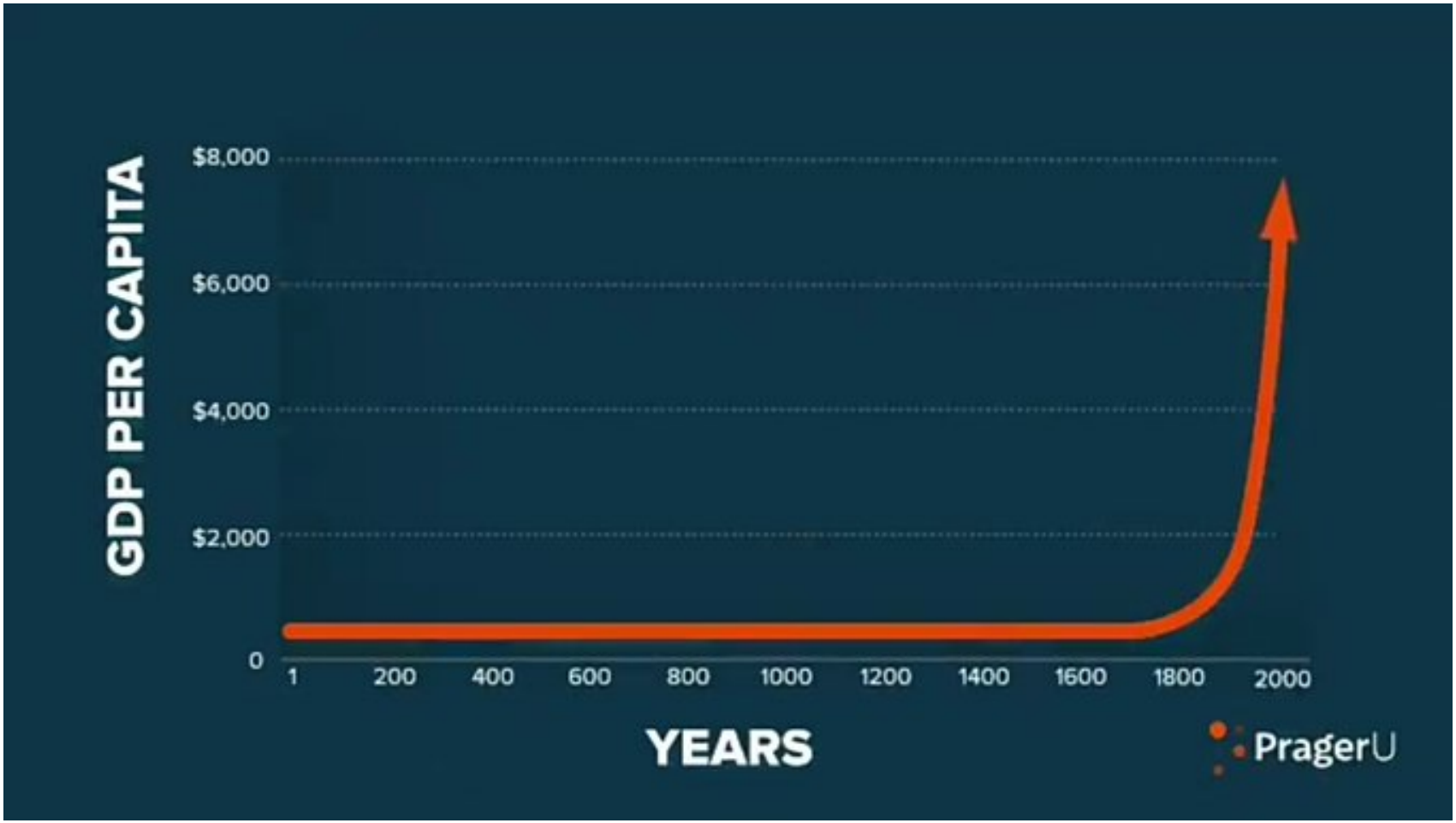
Don't Extrapolate II



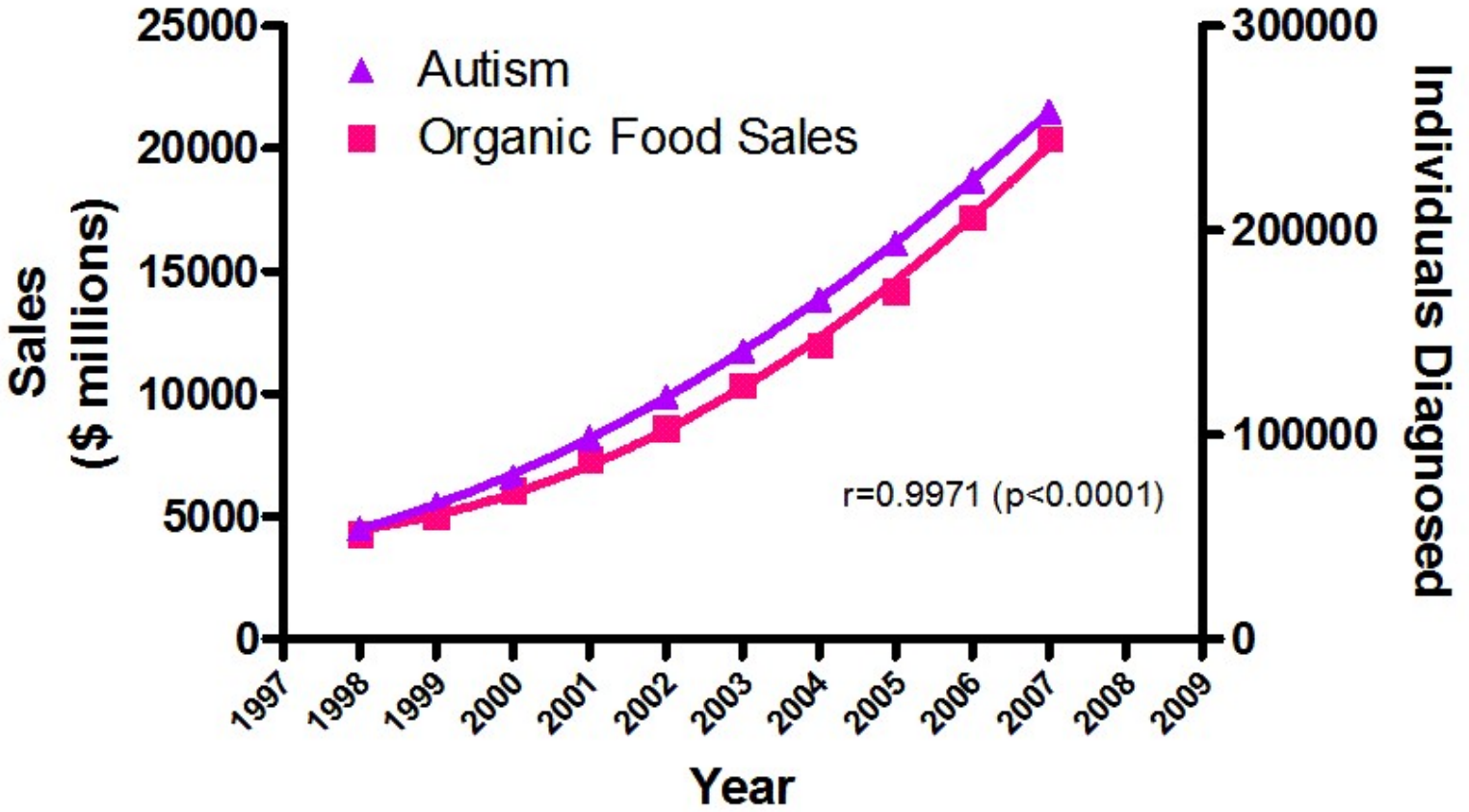
Context is Important



Be Precise: Don't Just Draw

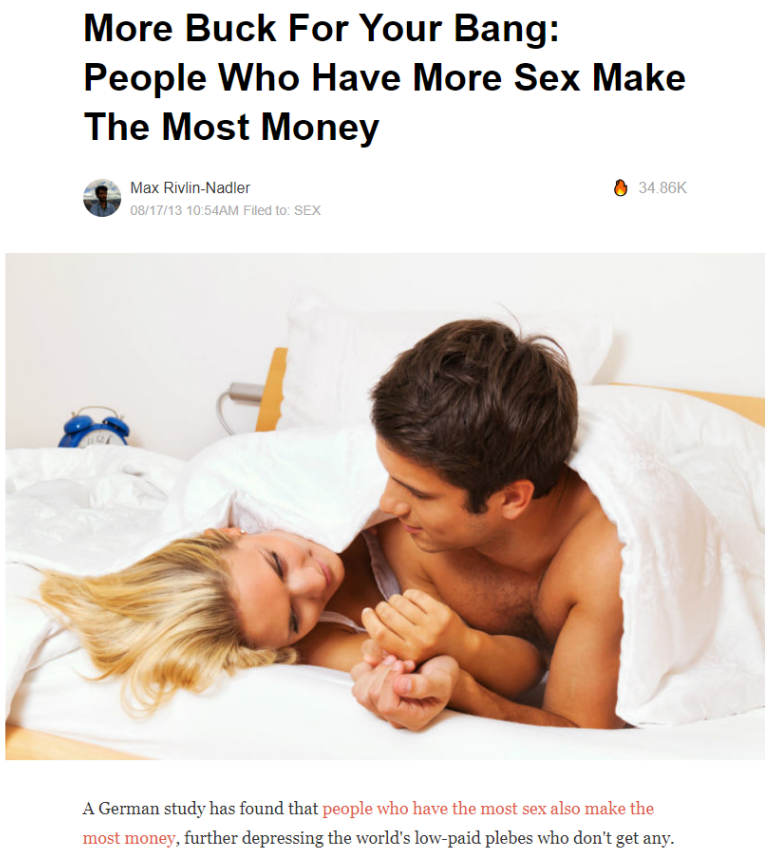


Correlation or Causation? I



Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals with Disabilities Education Act

Correlation or Causation? II



More Buck For Your Bang: People Who Have More Sex Make The Most Money

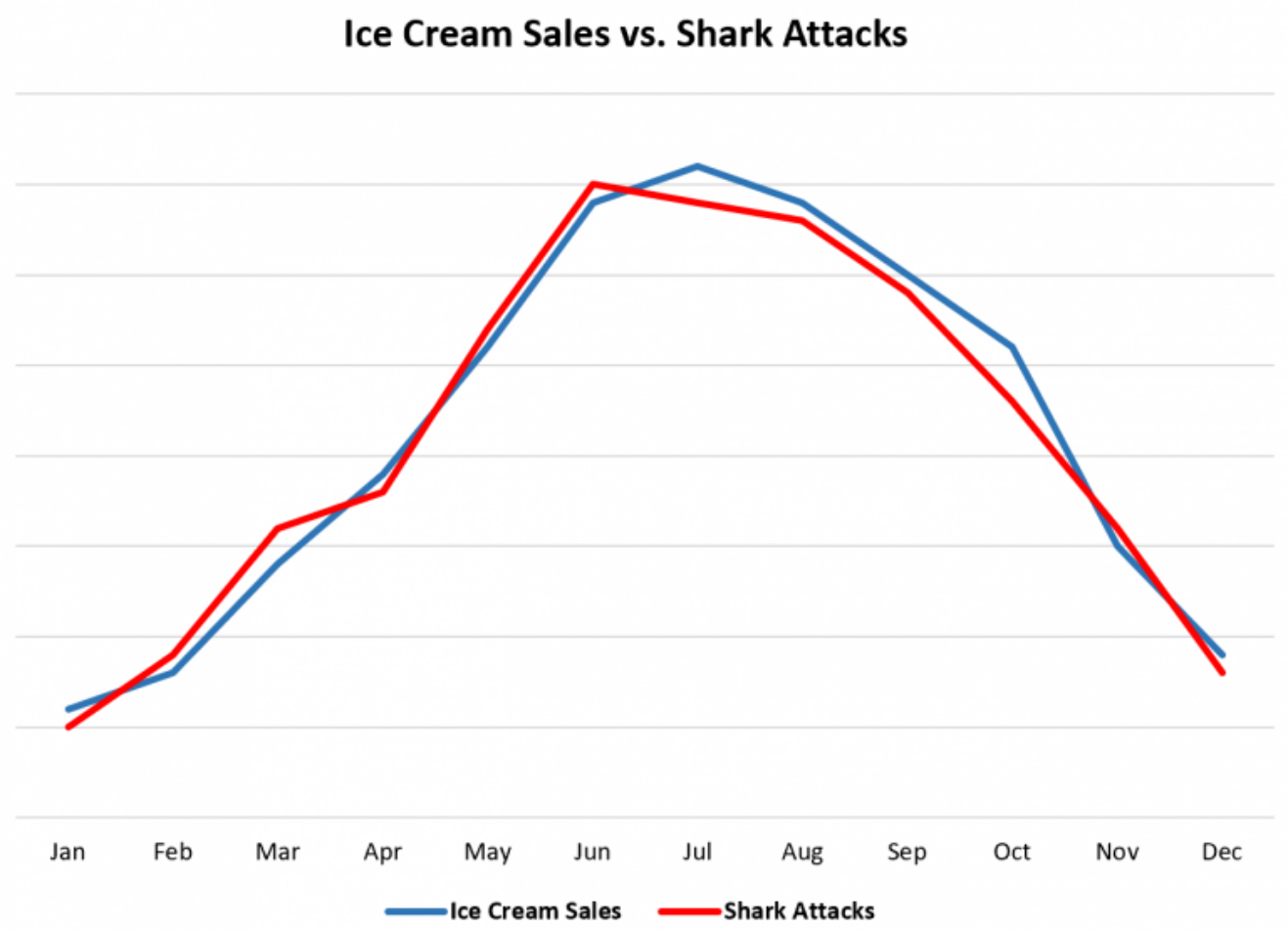
Max Rivlin-Nadler
08/17/13 10:54AM Filed to: SEX

34.86K

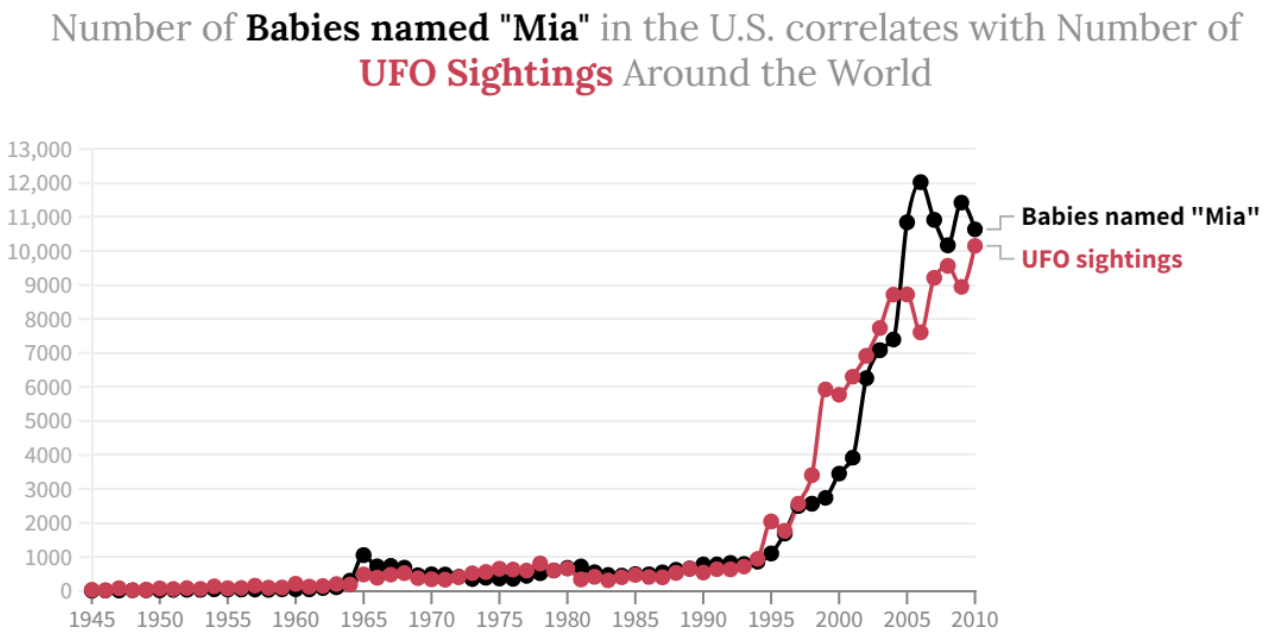
A German study has found that people who have the most sex also make the most money, further depressing the world's low-paid plebes who don't get any.

Causality only exists in a very specific occupation... Extroversion is the third omitted variable.

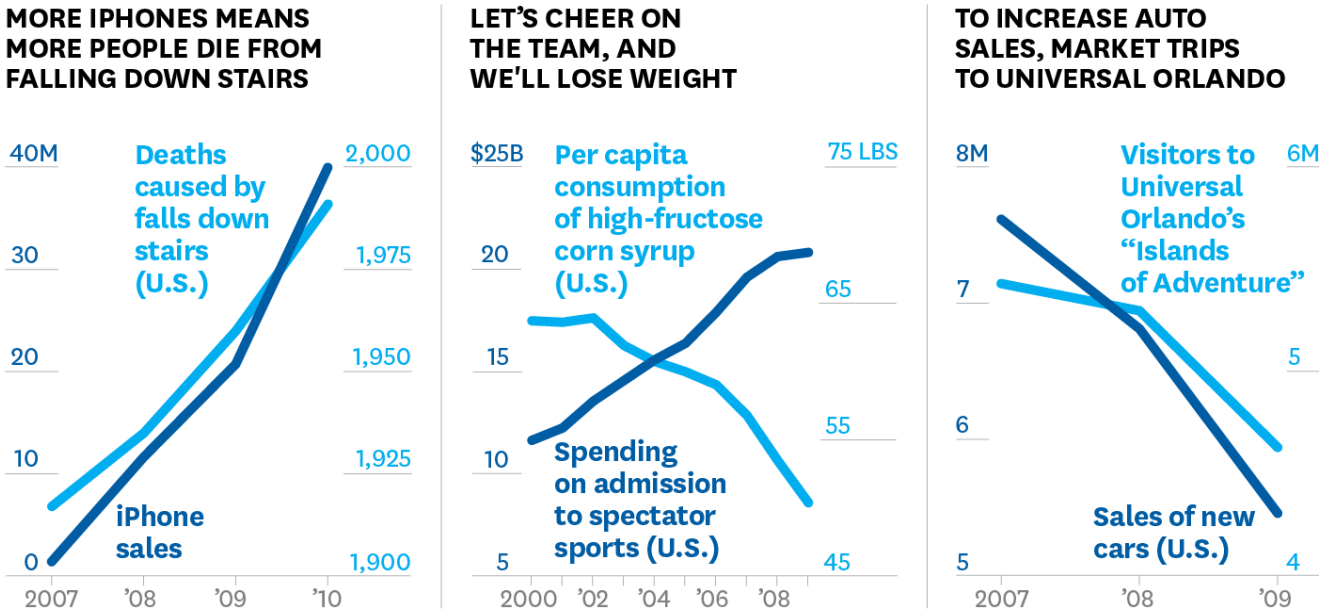
Correlation or Causation? III



Correlation or Causation? IV



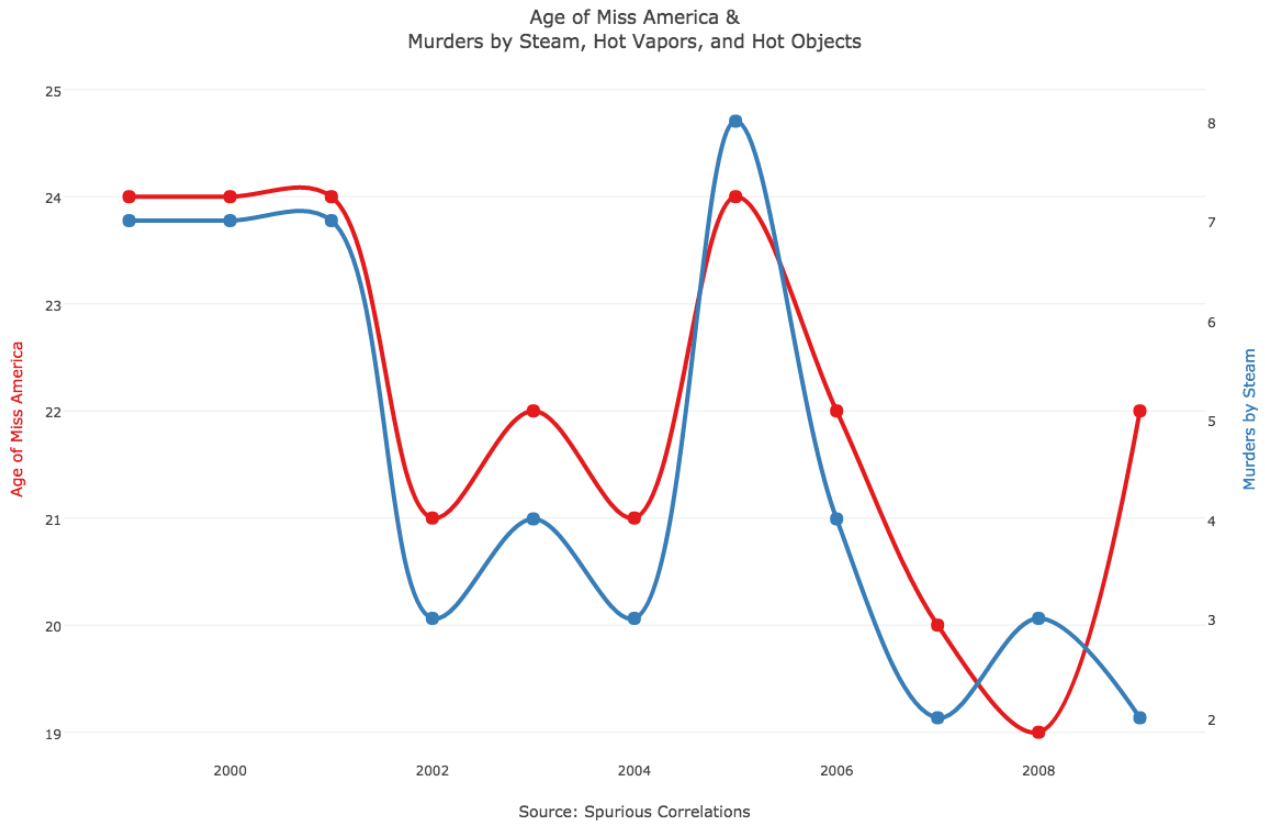
Correlation or Causation? V



SOURCE TYLERVIGEN.COM
FROM "BEWARE SPURIOUS CORRELATIONS," JUNE 2015

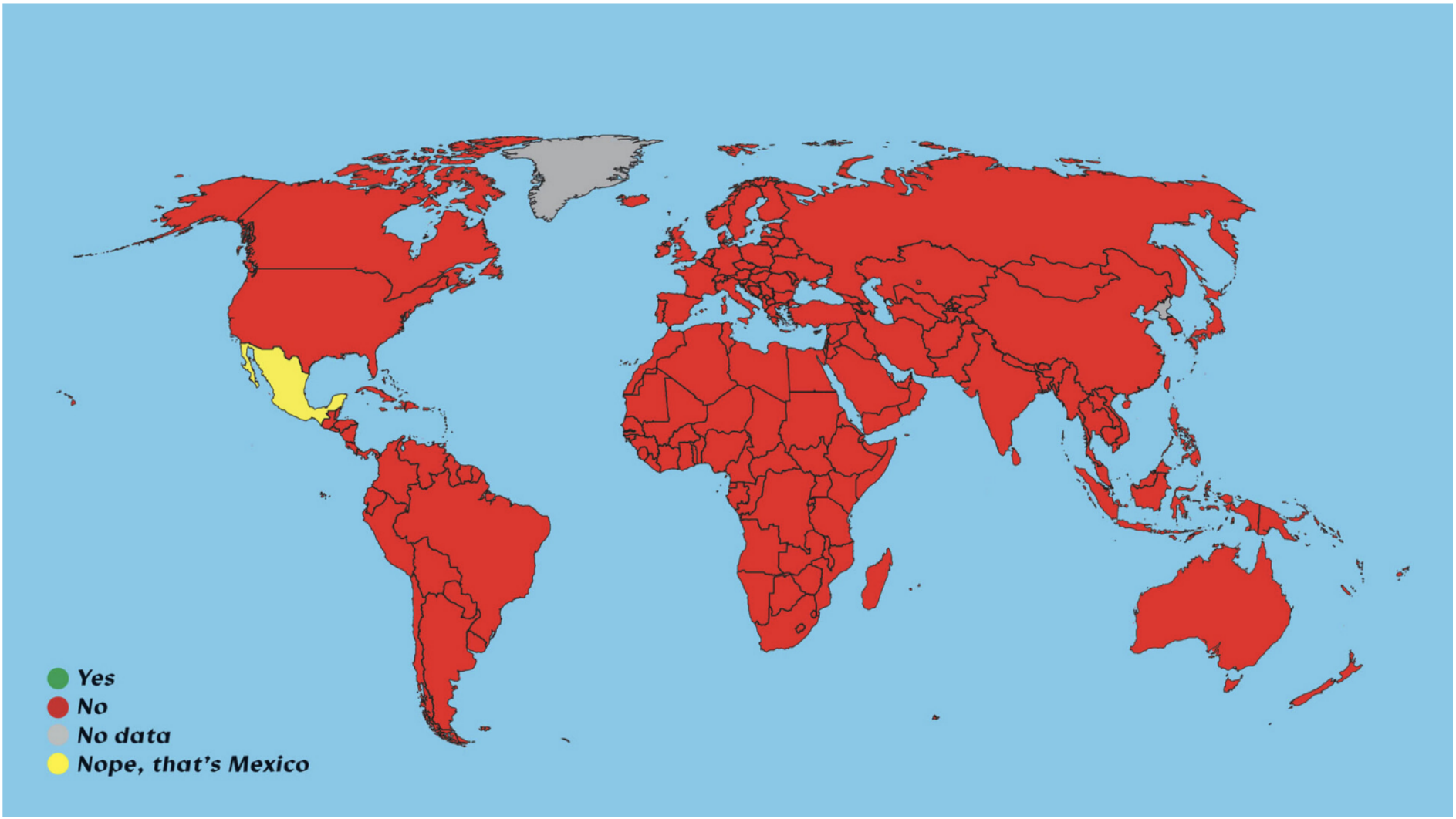
© HBR.ORG

Correlation or Causation? VI



Careful with Missing Data

Is this country Greenland?



Section 3

Graphical Sophistication

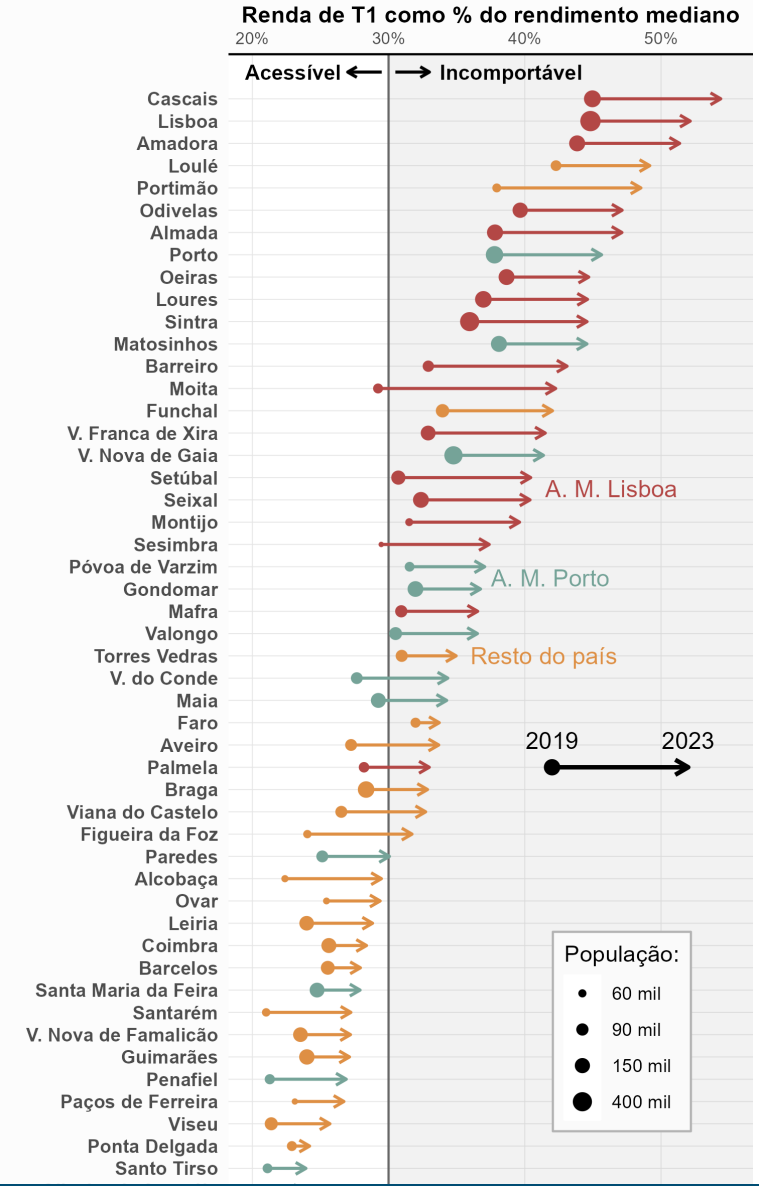
Don't Waste Ink

Don't make graphs that are:

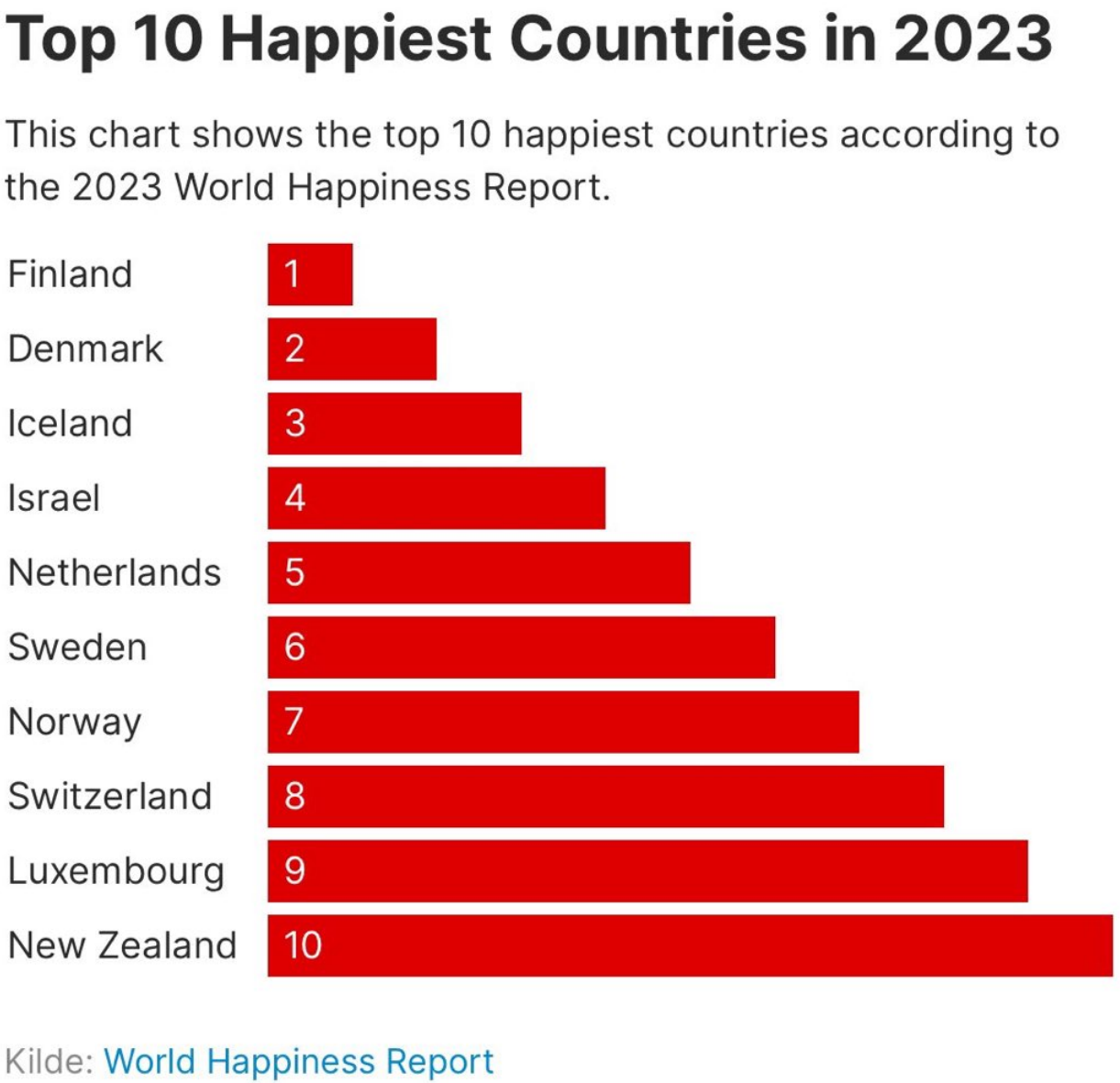
- Too simple. Show more than one variable. Use colors, size, arrows, to show more than one variable.
- Too obvious.

Short attention spans

Since you have people's attention, transmit the maximum information you can



Useless I

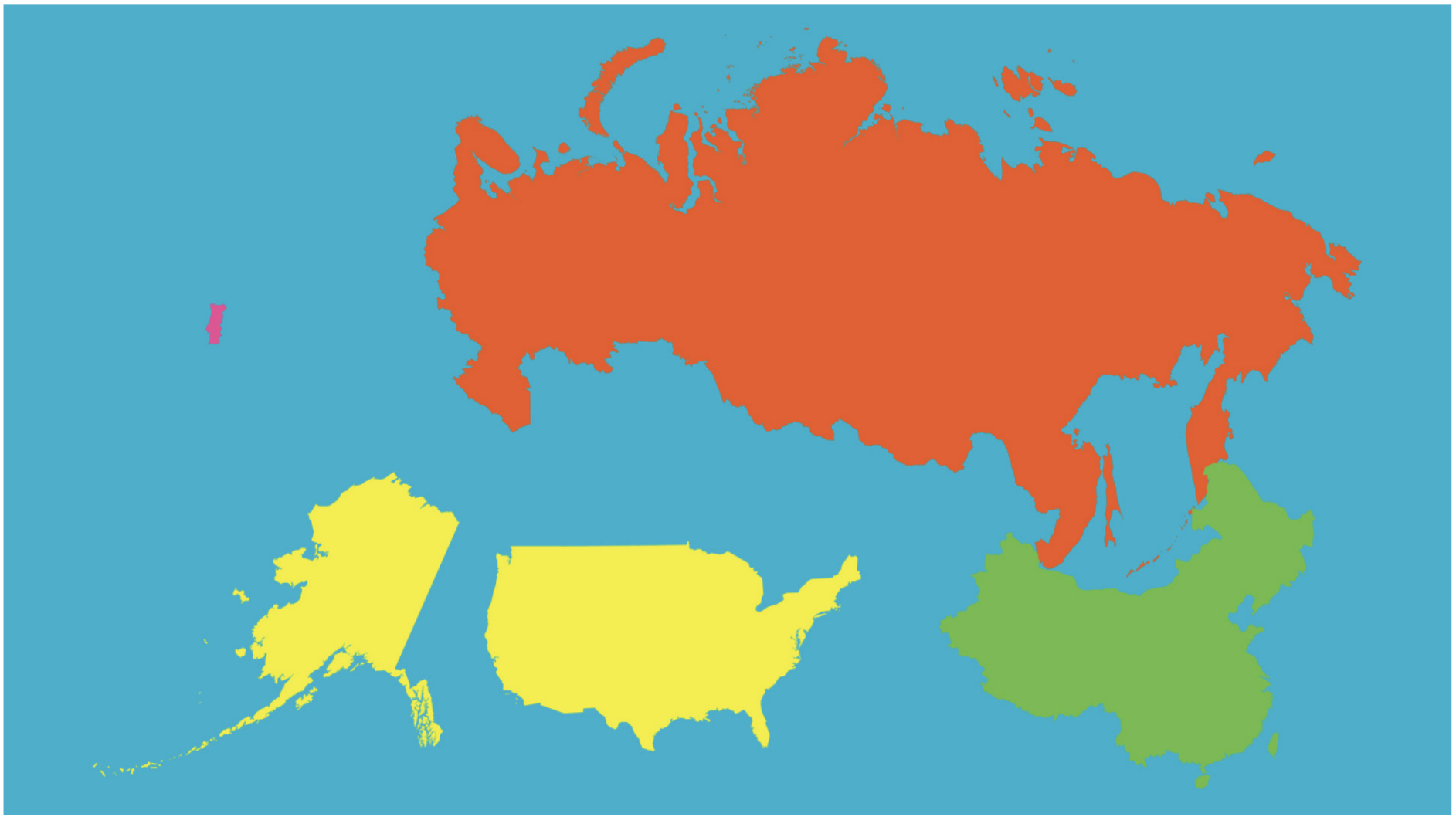

Top 10 Happiest Countries in 2023
This chart shows the top 10 happiest countries according to the 2023 World Happiness Report.
Finland 1
Denmark 2
Iceland 3
Israel 4
Netherlands 5
Sweden 6
Norway 7
Switzerland 8
Luxembourg 9
New Zealand 10
Kilde: World Happiness Report

Obvious I

Coldest Season of the Year

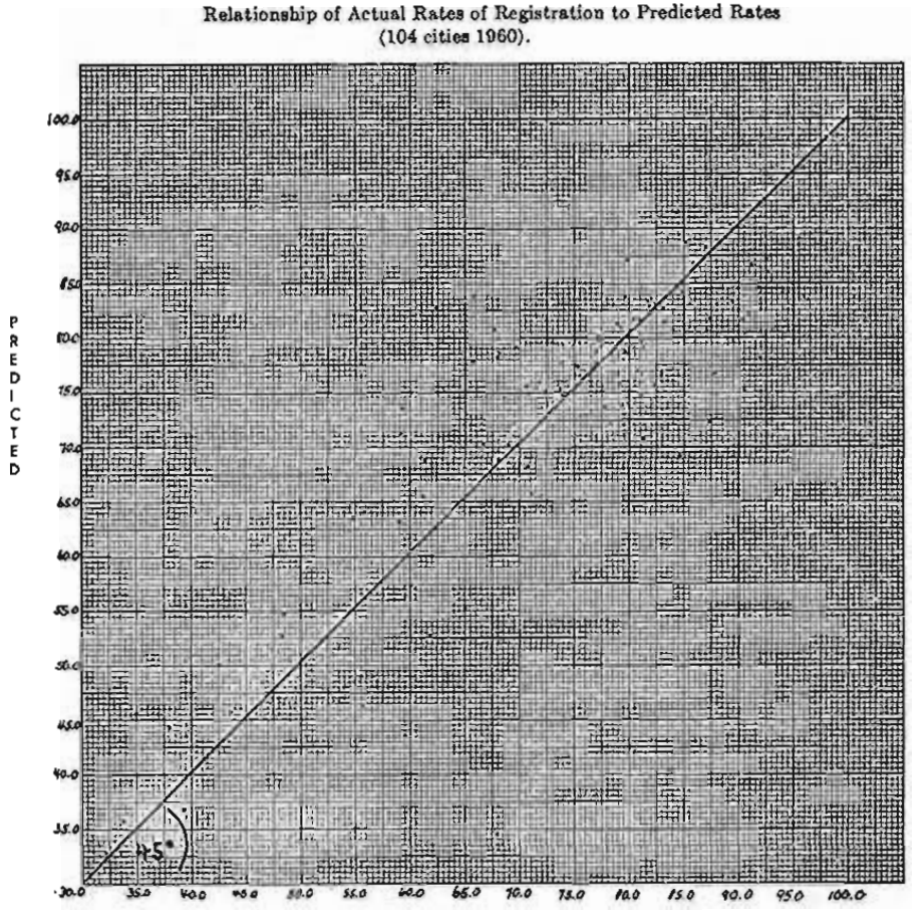
Obvious II

Portugal is smaller and the USA, Russia and China combined!



Use little Non-Data Ink

Non Data Ink Can be erased without loss of information. This graph has too much ink in the grid:



Keep It Simple I

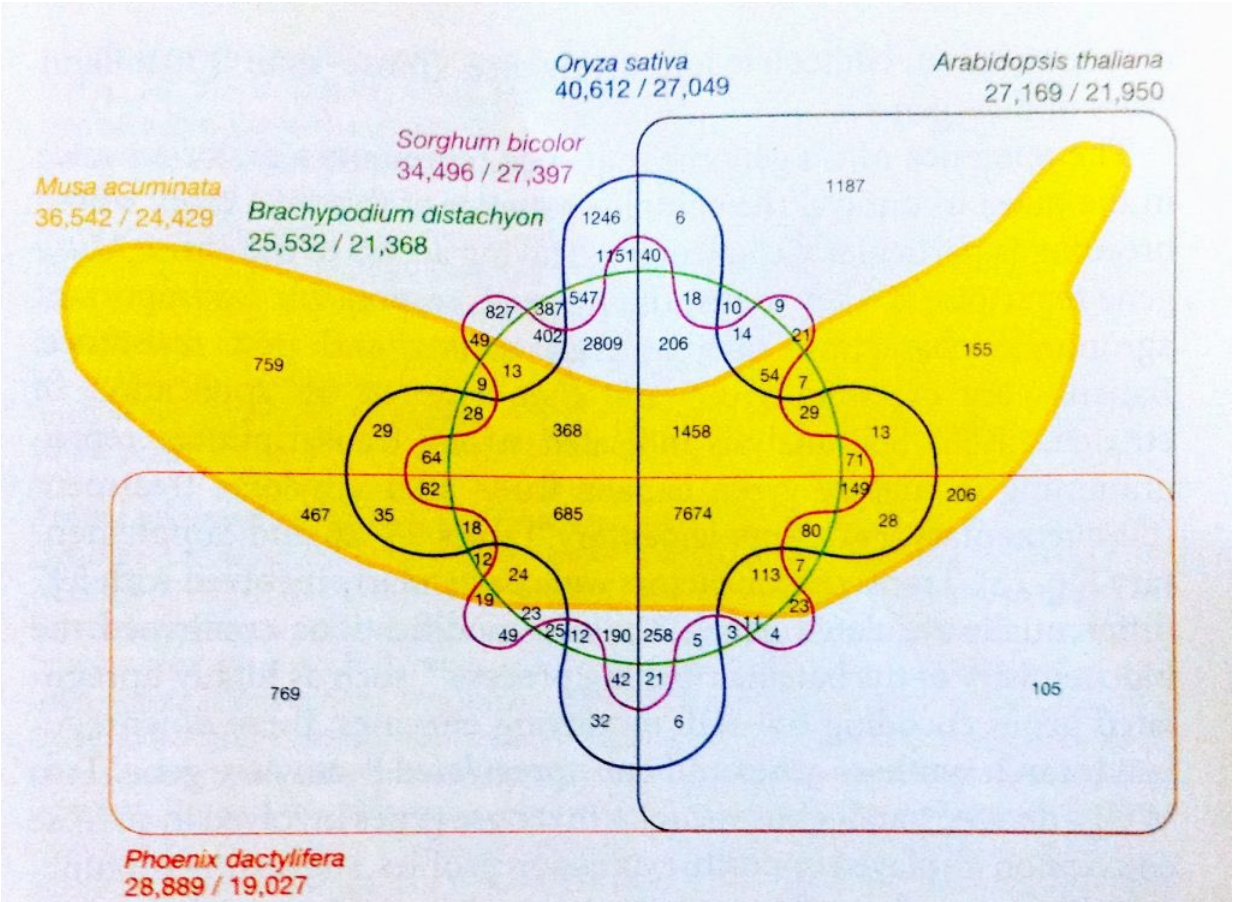
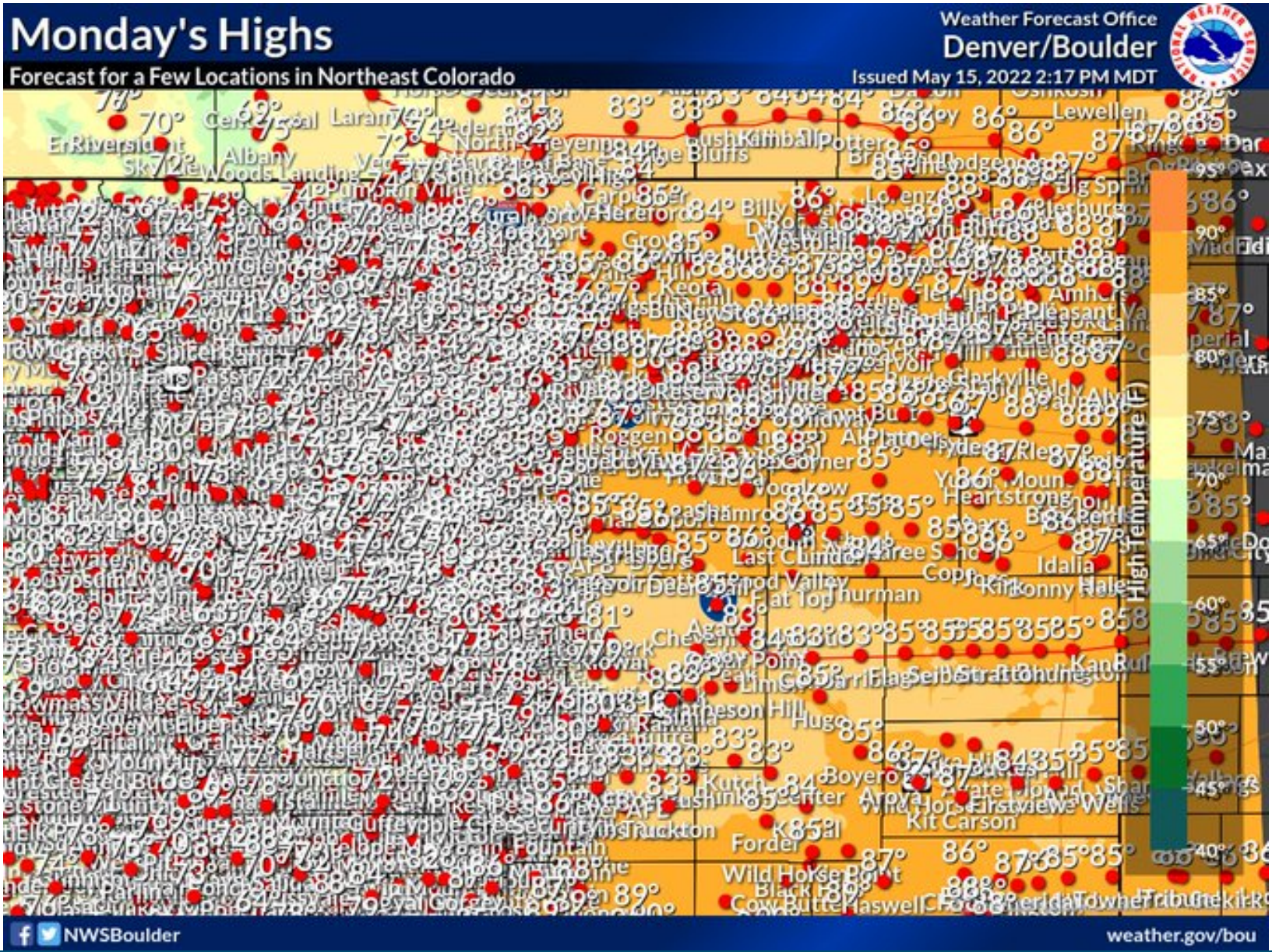


Figure 4 | Six-way Venn diagram showing the distribution of shared gene families (sequence clusters) among *M. acuminata*, *P. dactylifera*, *Arabidopsis thaliana*, *Oryza sativa*, *Sorghum bicolor* and *Brachypodium distachyon* genomes. Numbers of clusters are provided in the intersections. The total number of sequences for each species is provided under the species name (total number of sequences/total number of clustered sequences).

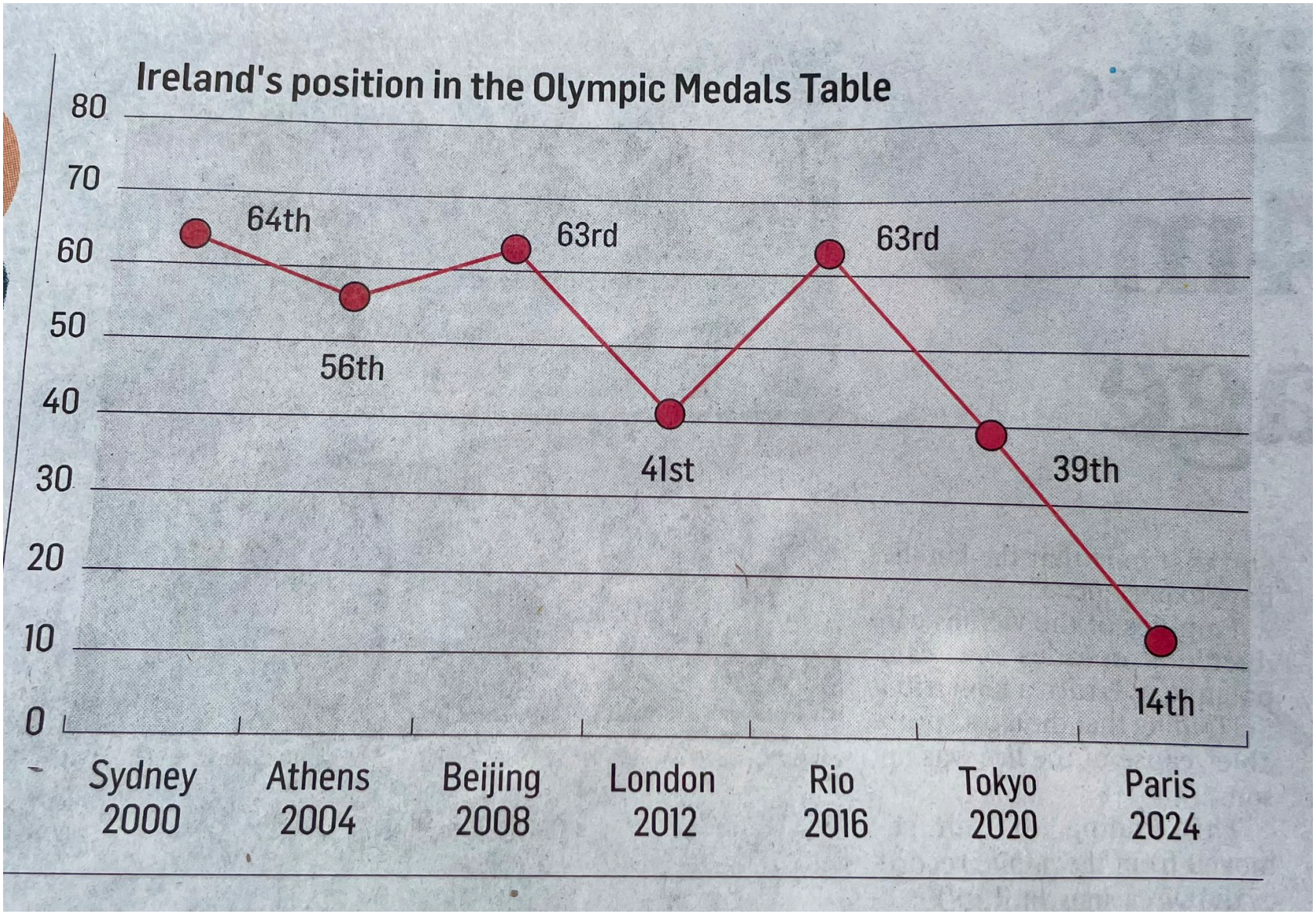
Keep It Simple II



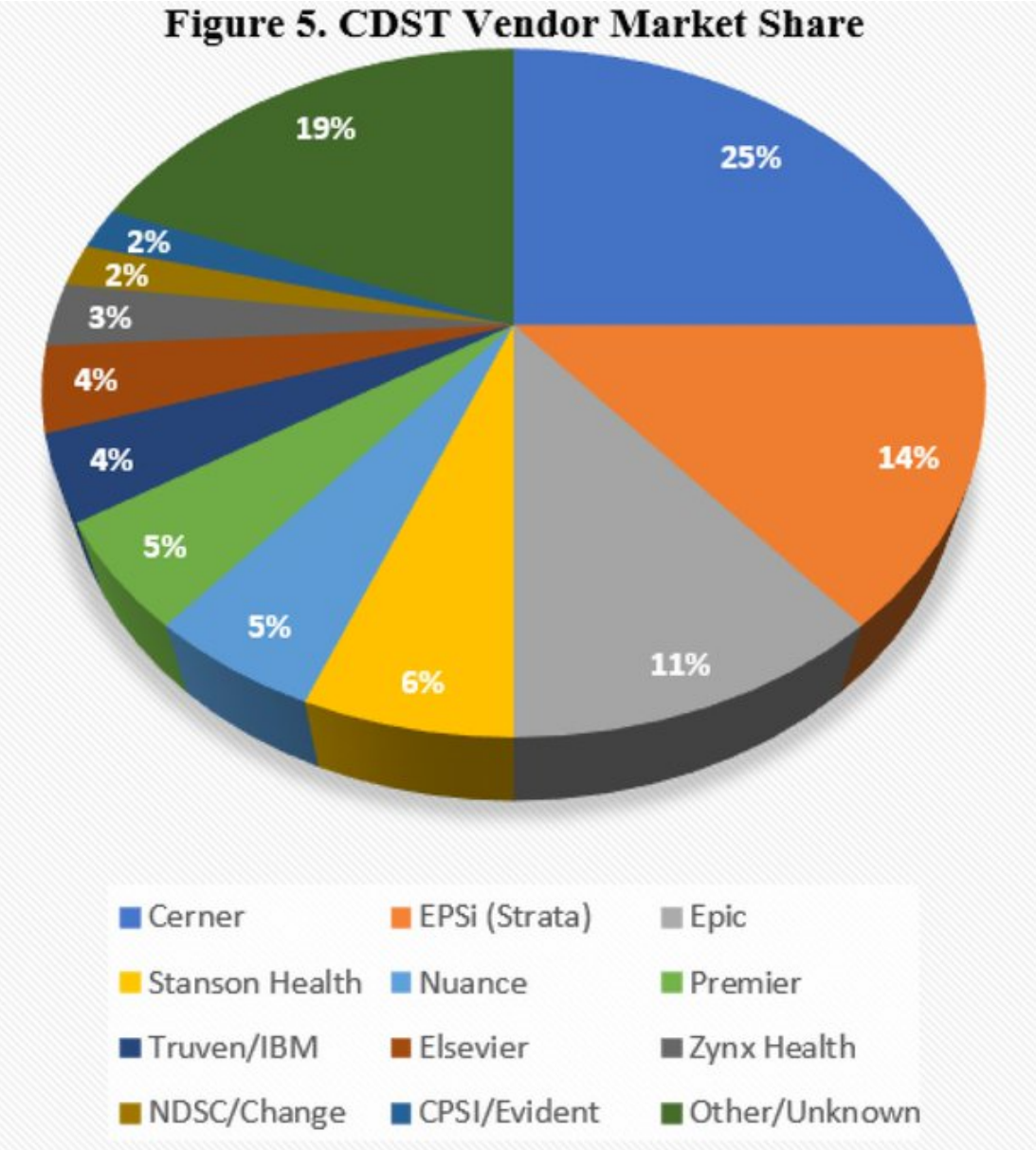
Section 4

Make it Beautiful and Effortless

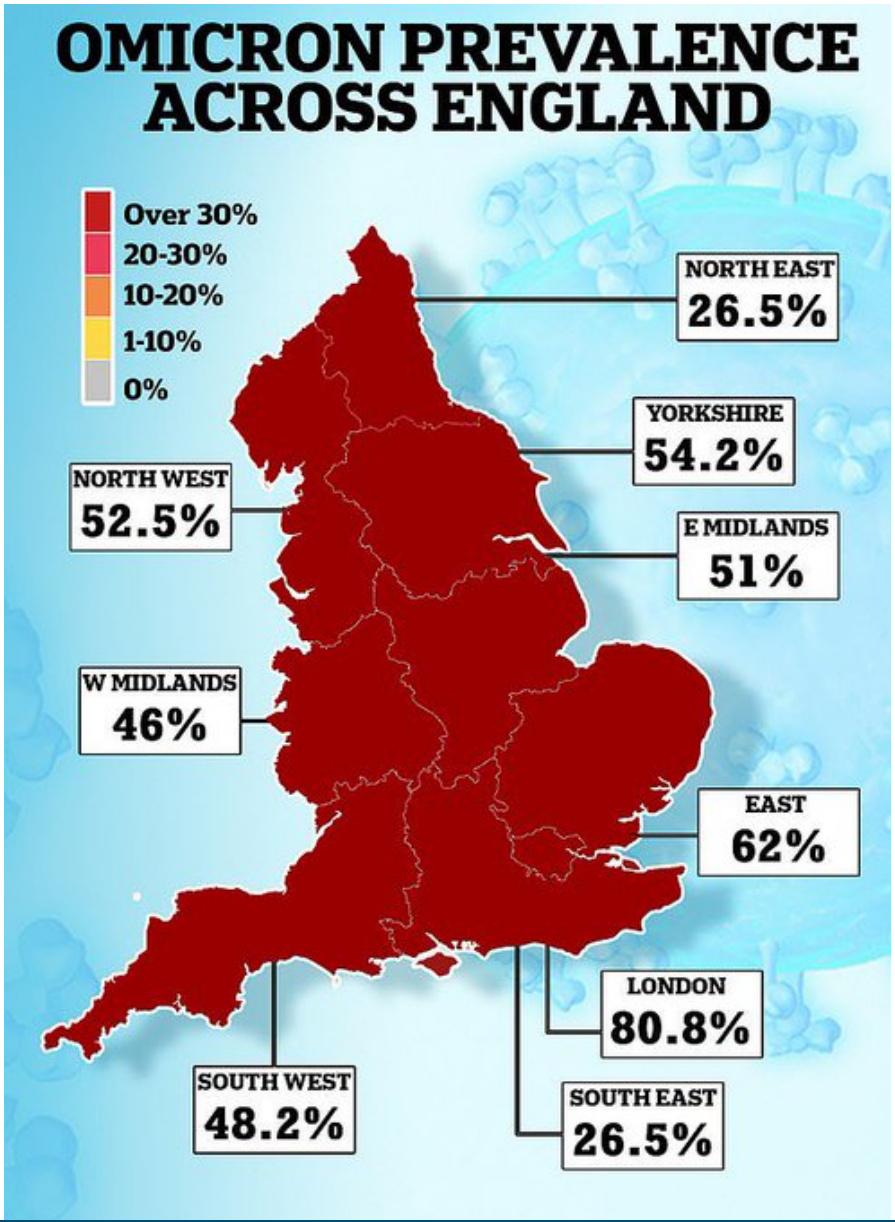
Make It Intuitive



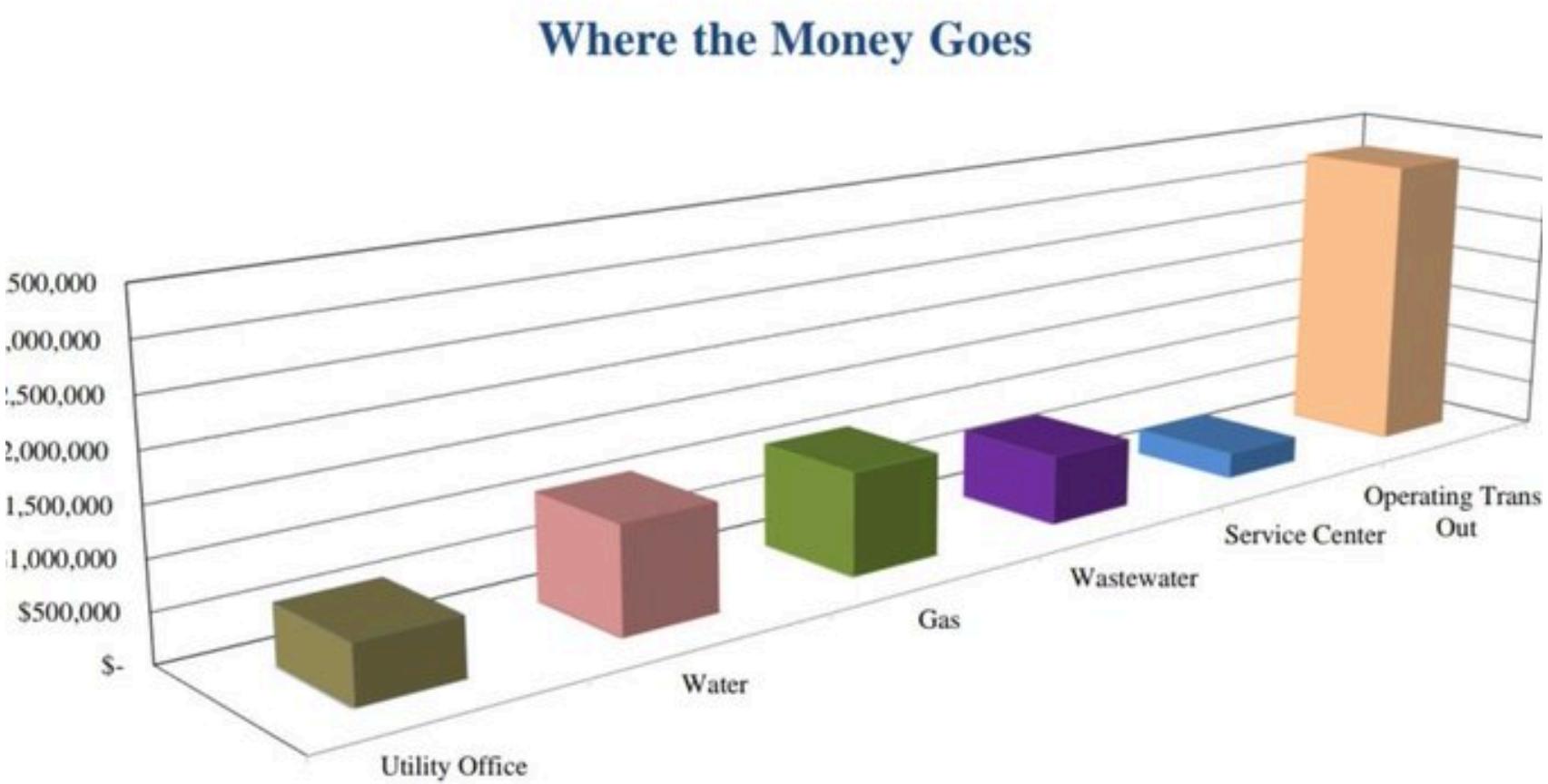
Your Palette Matters



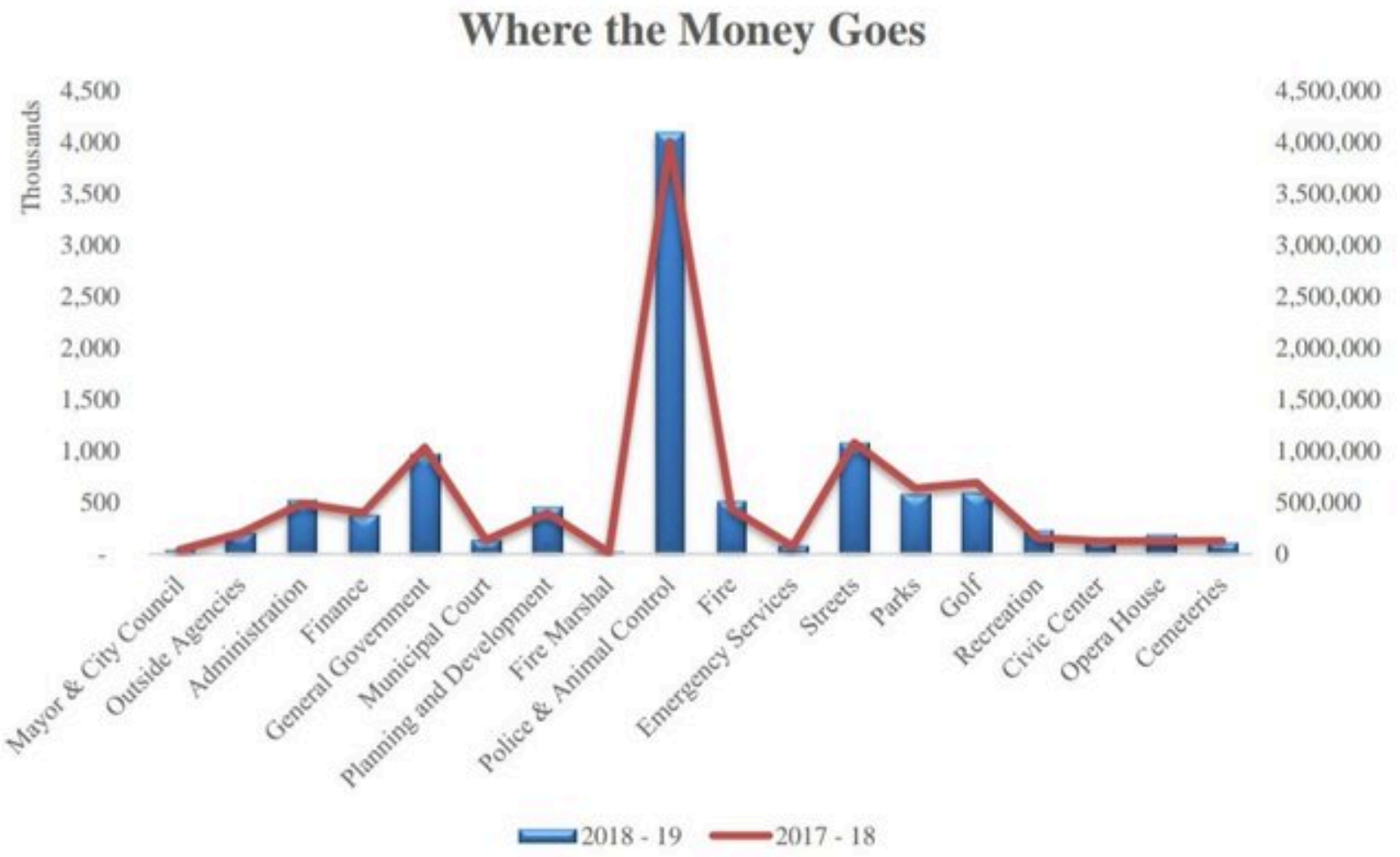
Colors Matter



3D Is Not "Cool"



Lines Should Show Trends I



Lines Should Show Trends II

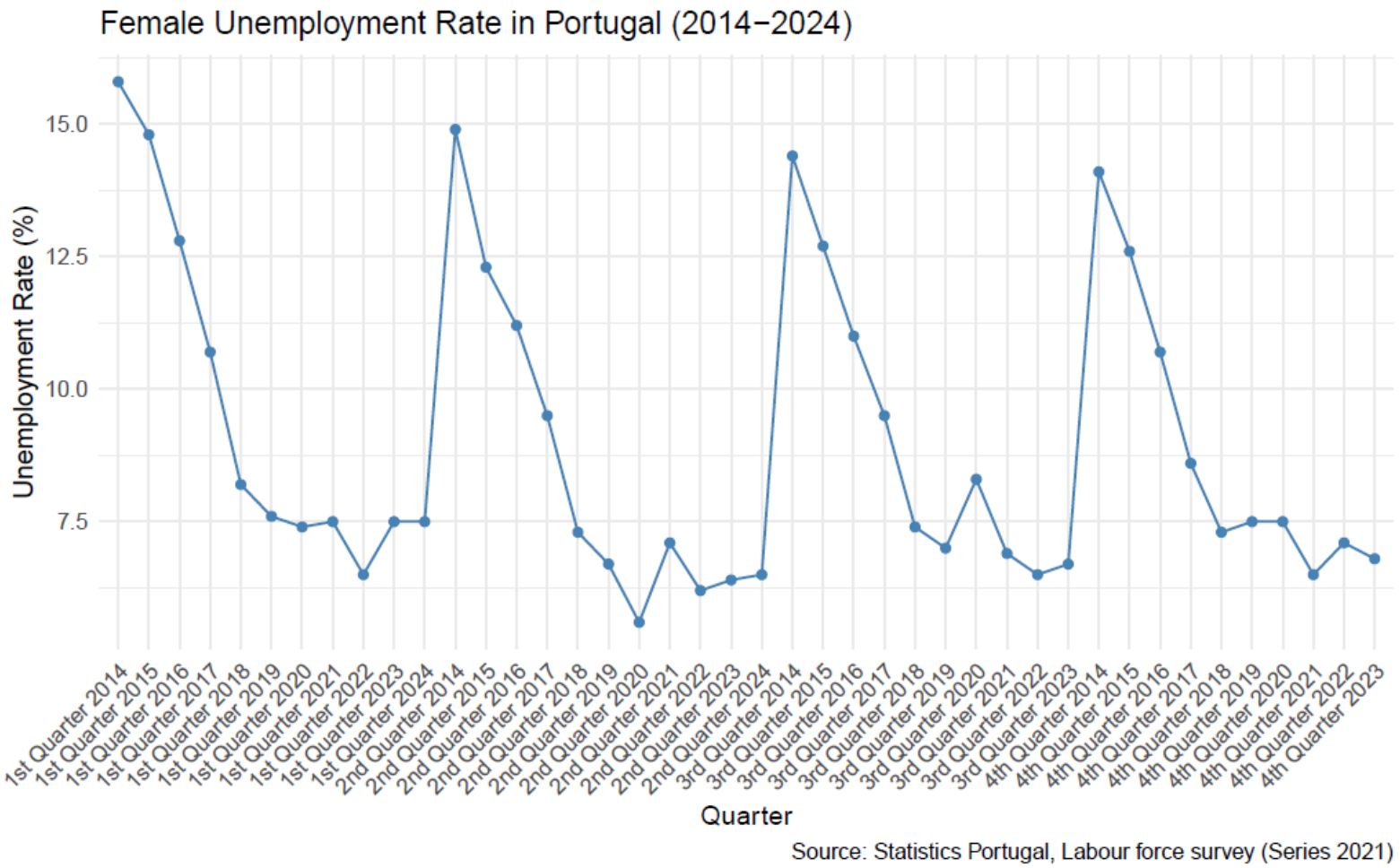


Figure 2: Female unemployment rate in Portugal with linear trend