

# Lecture 5 - Data Visualization

Topics in Econometrics

Vincent Bagilet

2025-09-30

# Housekeeping

- Questions on replication games?
- On your research proposal?

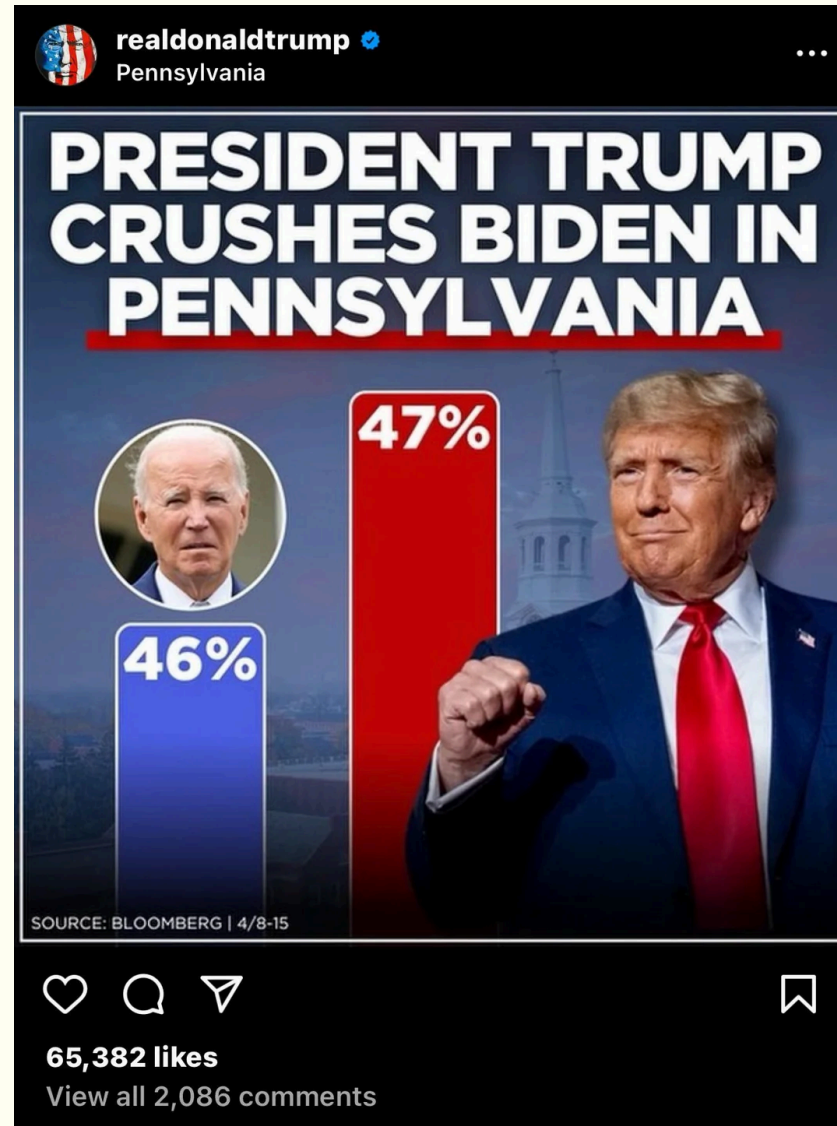
# Introduction

# Everywhere but not so simple

- Data viz is **everywhere**
- We work with data, we routinely (need to) visualize it
- **Seems pretty simple**, we all know how to make graphs
- Sure BUT **there are a few things we need to think about** when visualizing data
- Once you pay attention to data viz, it is fun, instructive and satisfying!

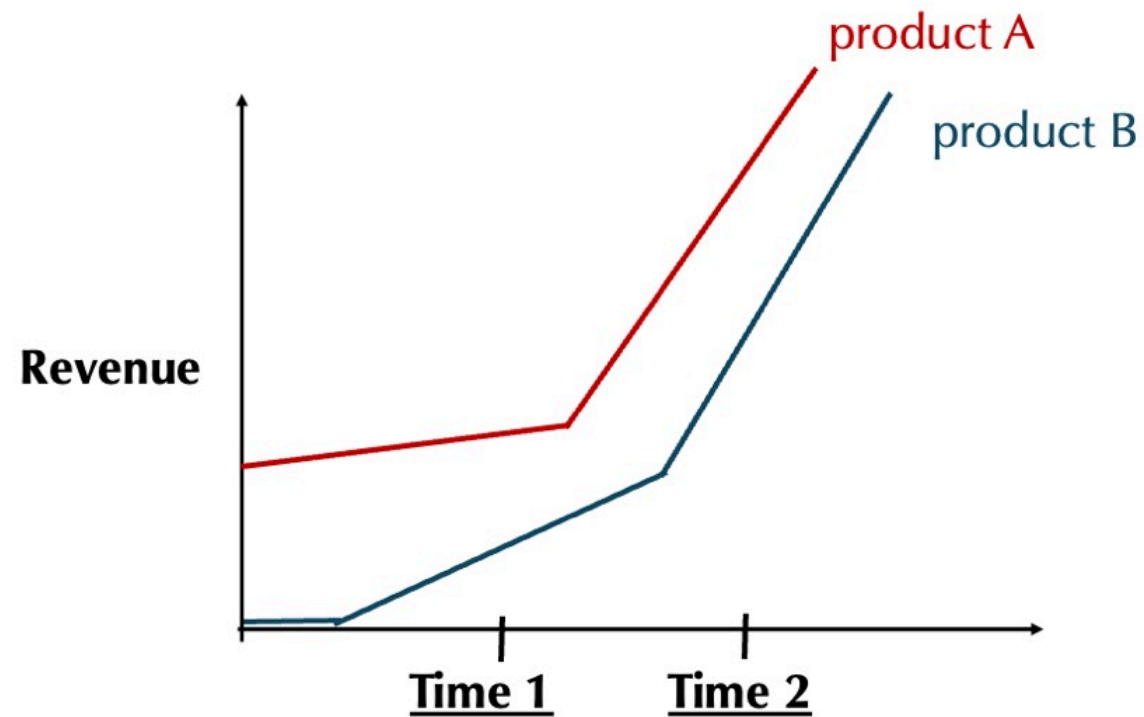


# Data viz can be obviously deceptive



# ...or difficult to interpret

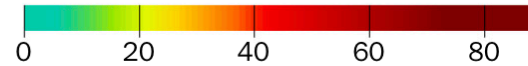
On which day was there a bigger difference in revenue between A and B?



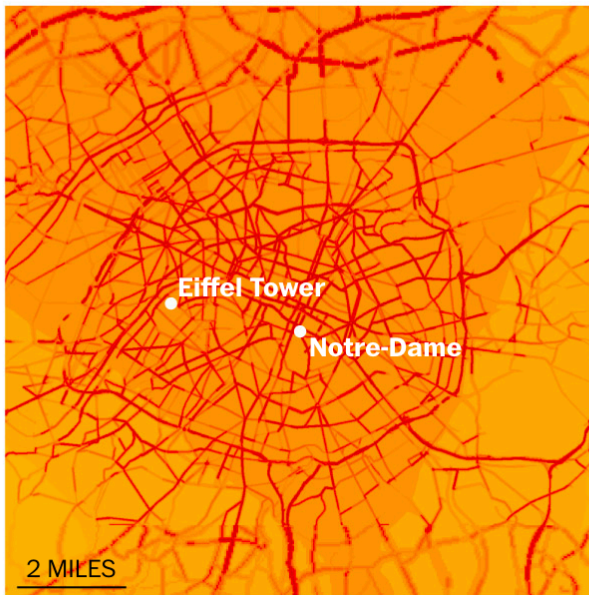
# They can be memorable and insightful

**Average PM 2.5 concentration in Paris**

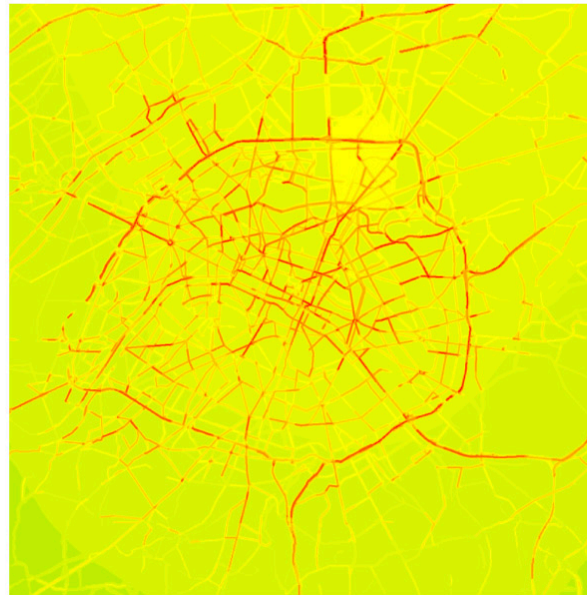
Micrograms per cubic meter



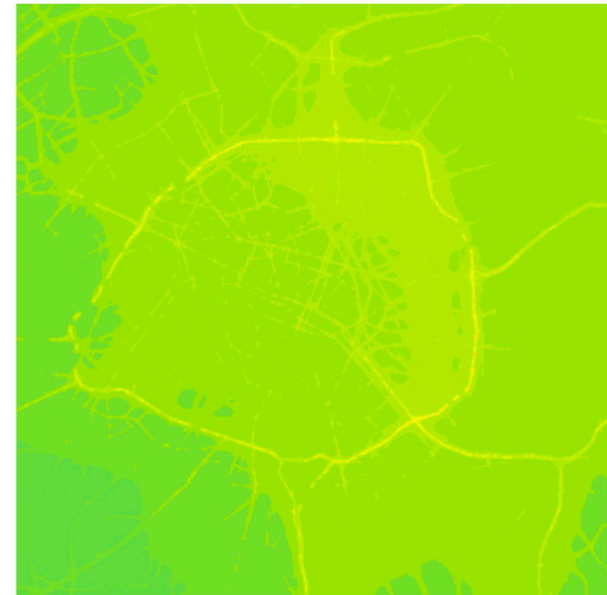
**2007**



**2014**



**2023**



Source: Airparif

**...and also beautiful**



# Outline

1. Why is data viz important?
2. Key data viz principles
3. Building a graph
4. Data viz for research in economics

**Why is data viz important?**

# When and why use data viz?

## Graphs to explore

- Analyze
- Confirm

## Graphs to explain

- Inform
- Convince

They have **different goals and audiences**

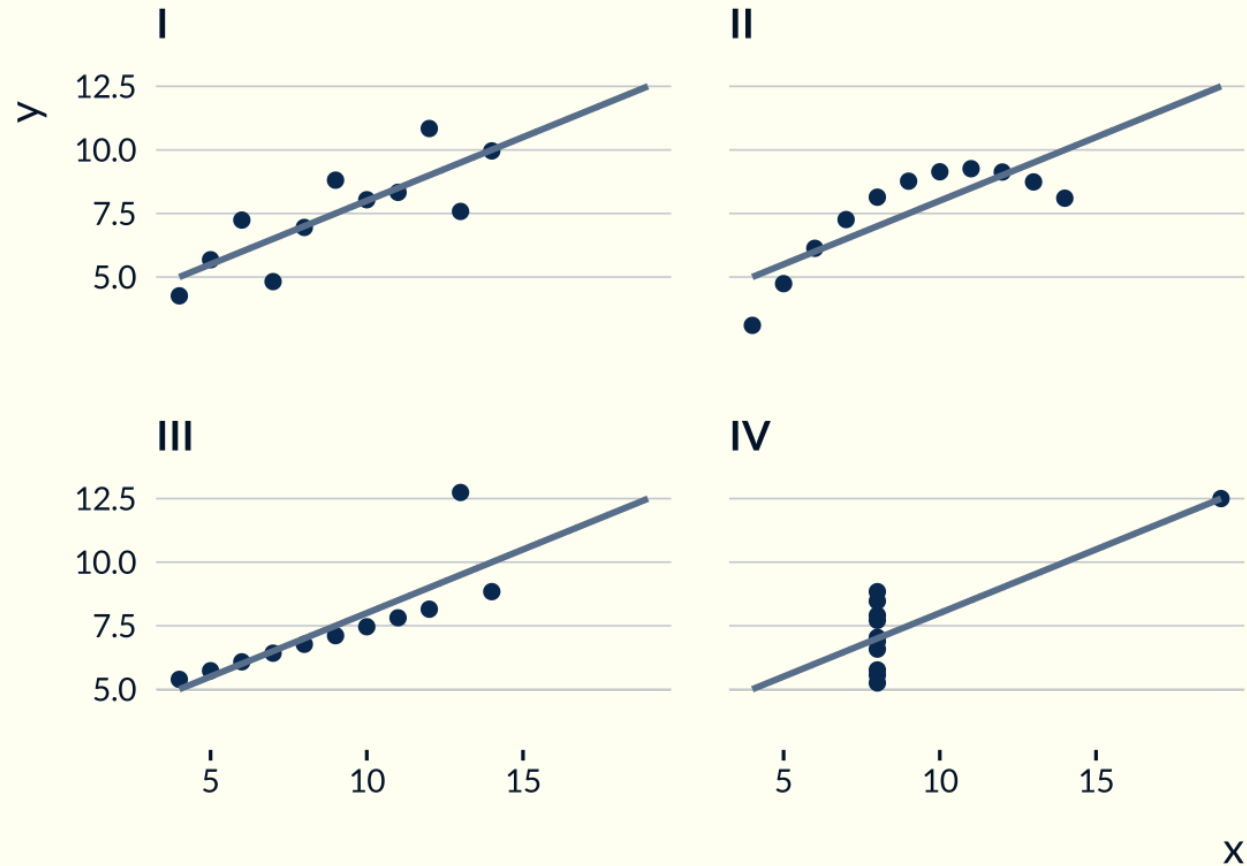
# Explore: make sense of your data

- Data often contain patterns; data viz can be tremendously helpful to identify them
- But need to look at your raw data
- That's the role of the exploratory data analysis (EDA)
- It may also help you
  - Formulating hypotheses (to test on other data sets)
  - Assess the relevance of some key assumptions



# Look at your raw data

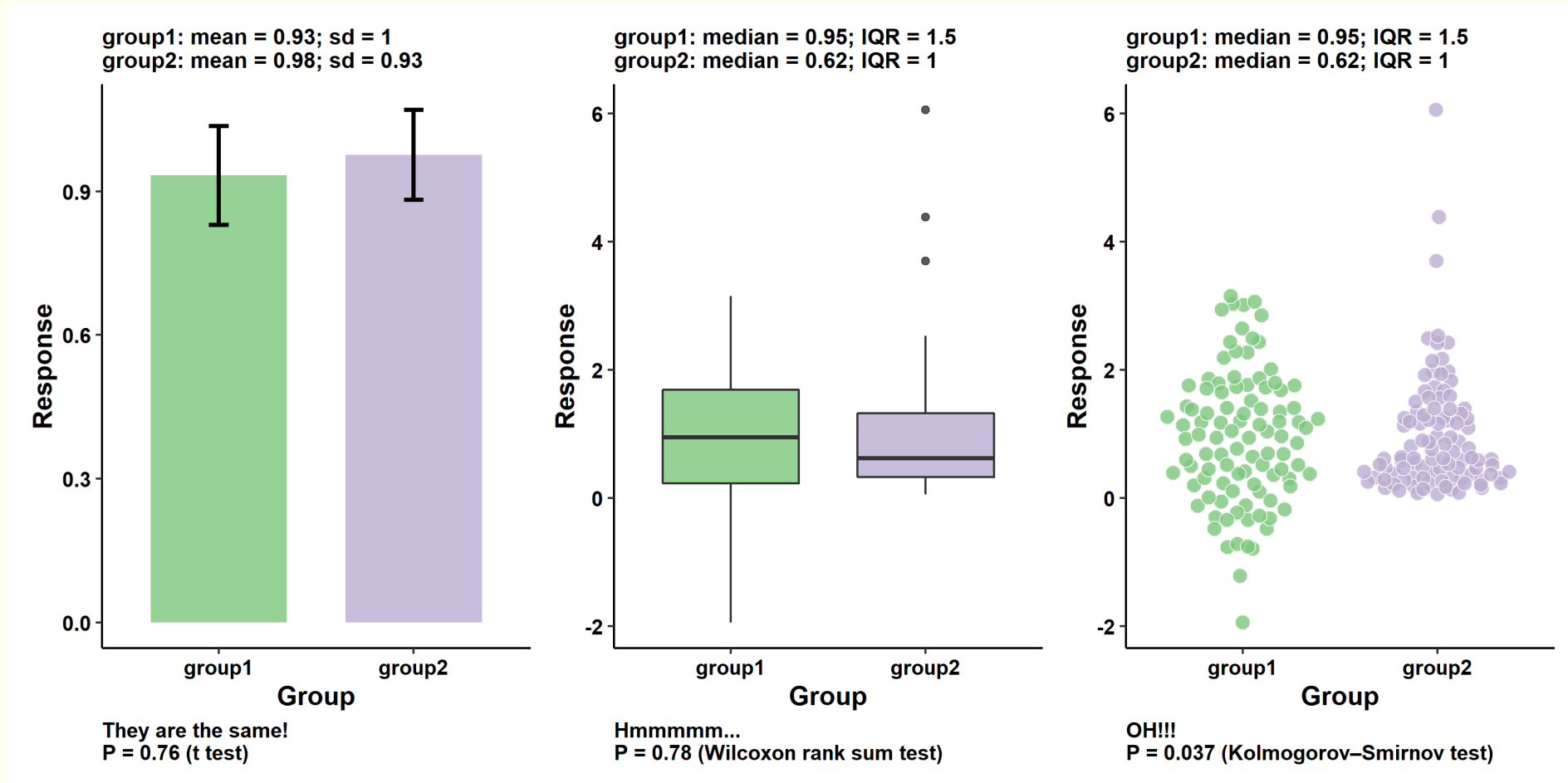
Anscombe's quartet



Same relation, different patterns

- eg might be helpful to **evaluate your modeling assumptions**

# Look at your raw data



- Some plots (or summary statistics) help summarize but can also **hide** information
- eg might be helpful to **explore balance**

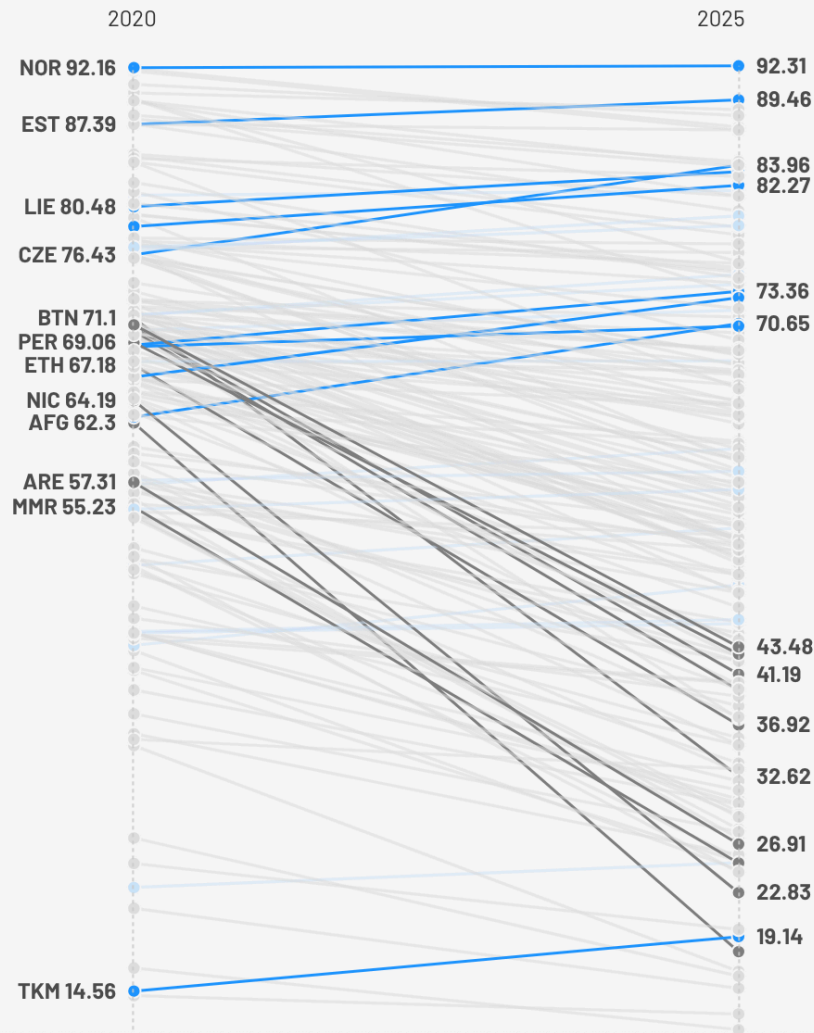
# Explain: tables might be arrid

Country	ISO	2020	2020 Ranking	2025	2025 Ranking	Region	Pattern
Afghanistan	AFG	62.30	122	17.88	175	Asia-Pacific	Lower
Albania	ALB	69.75	84	58.18	80	EU & Balkans	Lower
Algeria	DZA	54.48	146	44.64	126	MENA	Lower
Andorra	AND	76.77	37	63.30	65	EU & Balkans	Lower
Angola	AGO	66.08	106	52.67	100	Africa	Lower
Argentina	ARG	71.22	64	56.14	87	Americas	Lower
Armenia	ARM	71.40	61	73.96	34	EECA	Higher
Australia	AUS	79.79	26	75.15	29	Asia-Pacific	Lower
Austria	AUT	84.22	18	78.12	22	EU & Balkans	Lower

**Data viz can help getting your point across**

# Press freedom freefalls

Average scored plummeted 10.5 points globally. This chart highlights the 10 countries with the sharpest **increase** or **decrease**

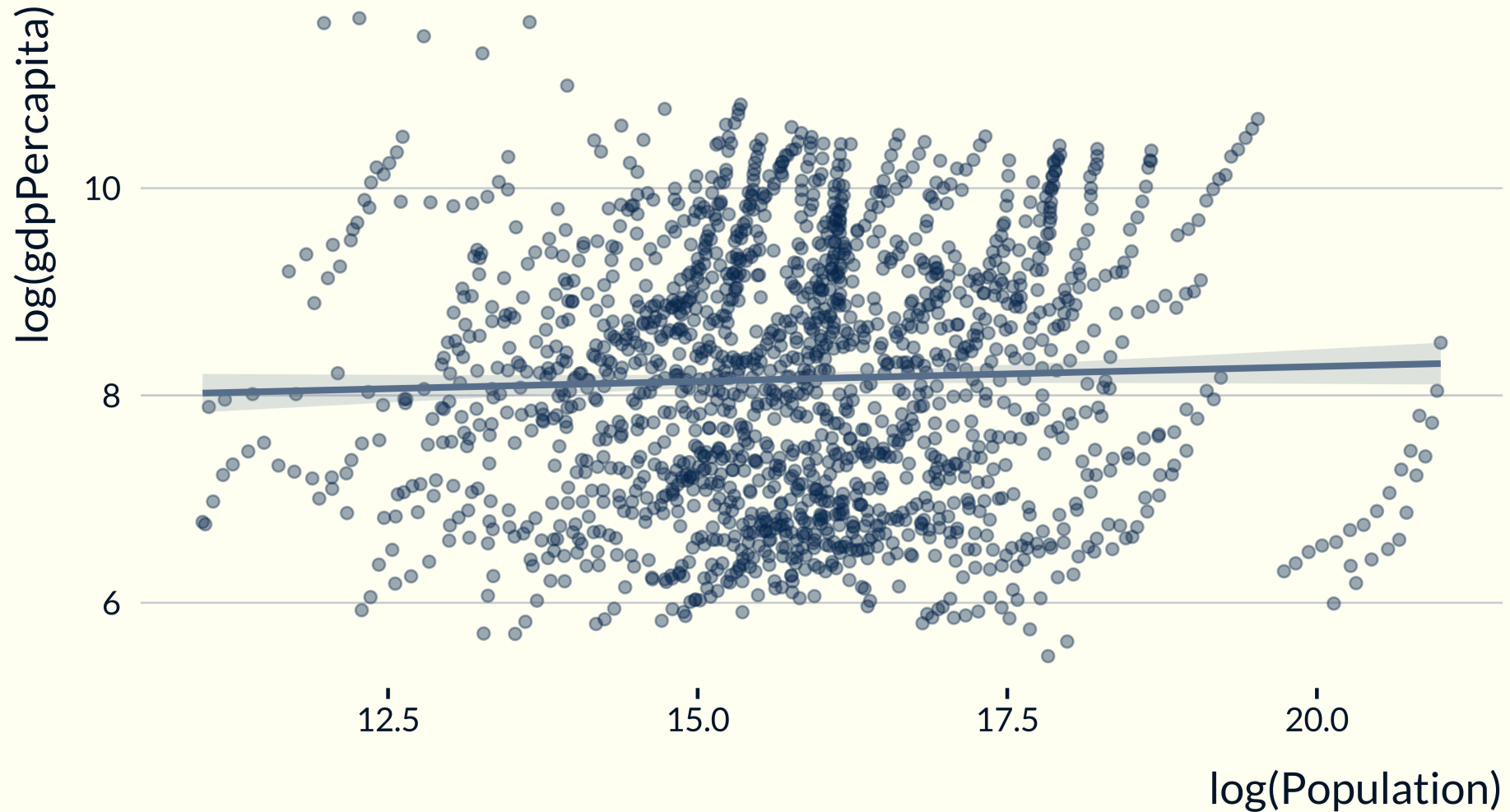


Source: [Reporters Without Borders](#) • Made with Flourish

- A data viz can be much clearer than a table (not always)
- Can help focus on one specific point

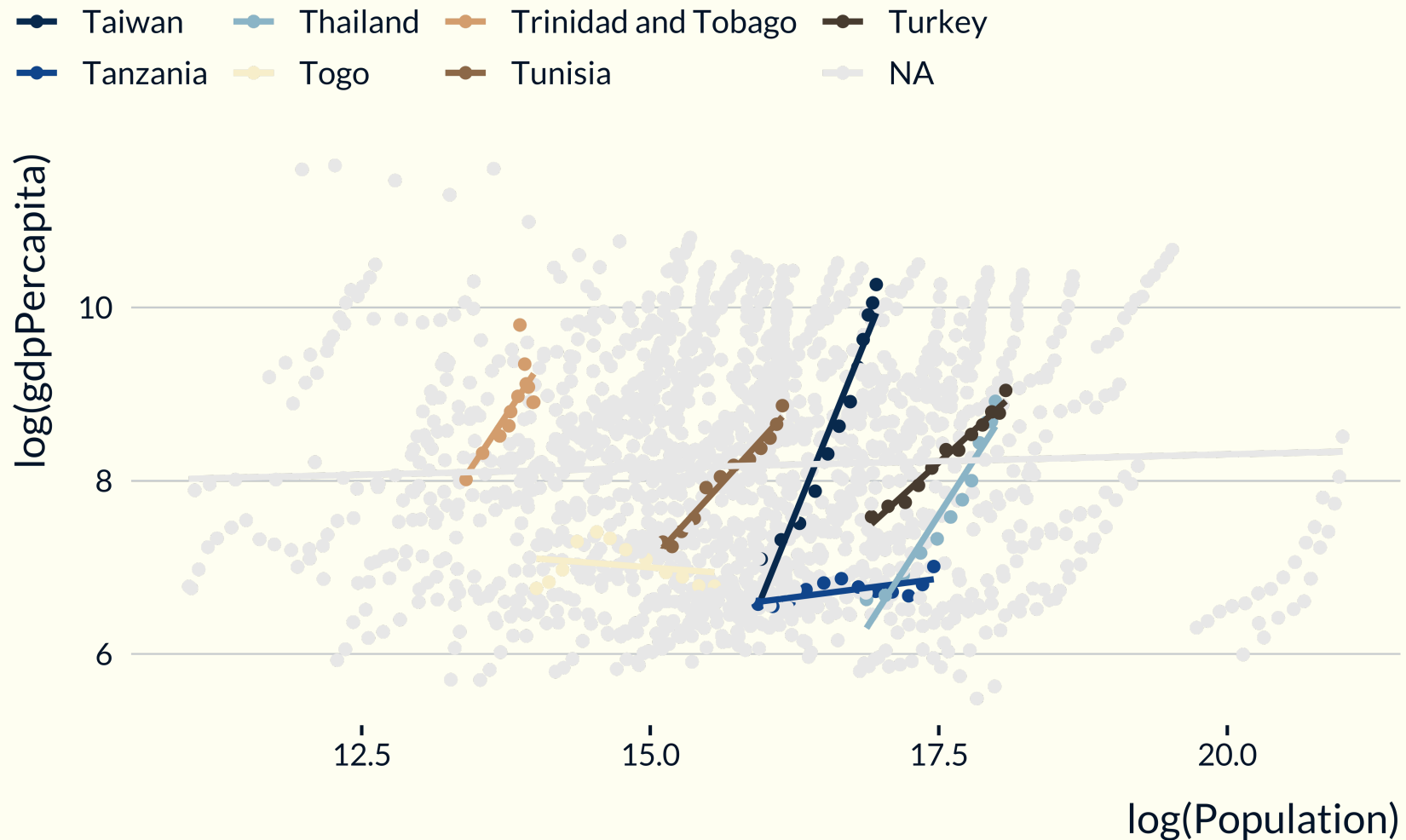
# Data viz can be a helpful rhetorical tool

Population Against GDP per Capita



# Data viz can be a helpful rhetorical tool

Within Country Population Against GDP per Capita  
*For a subset of countries*



# The power of data viz

*We can easily see patterns presented in certain ways, but if they are presented in other ways, they become invisible [..]*

*Following perception-based rules, we can present our data in such a way that the important and informative patterns stand out. If we disobey the rules, our data will be incomprehensible or misleading.*

Ware, C. (2012). Information Visualization, Third Edition: Perception for Design

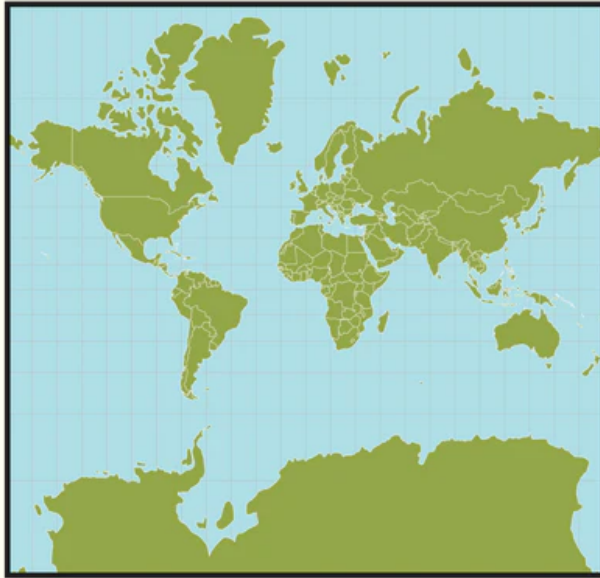


# Data viz can be misleading

- We have briefly discussed that before
- There is a breadth of ways in which they can be misleading
- Charts can be wrong. They can also be correct BUT misleading
- See *Defense Against Dishonest Charts* on Flowing Data

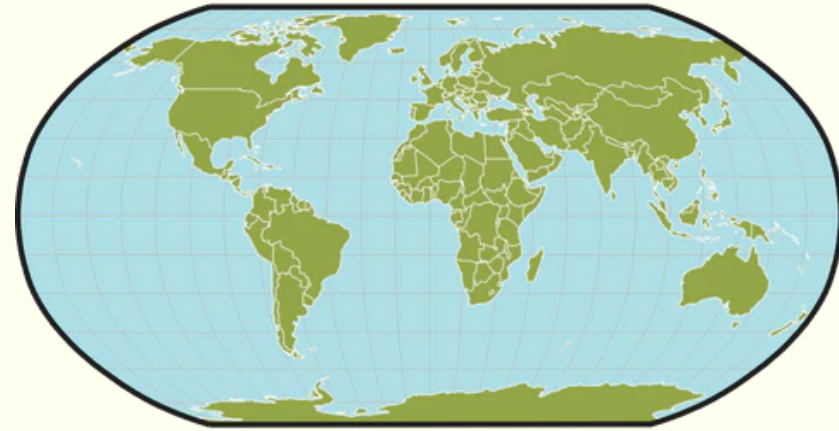
# Map projections distort the reality

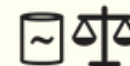
MERCATOR



 Gerardus Mercator - 1569

ROBINSON



 Arthur H. Robinson - 1963

# Cutting 0 on the y-axis

Shut up about the y-axis. It shouldn't always start at zero.

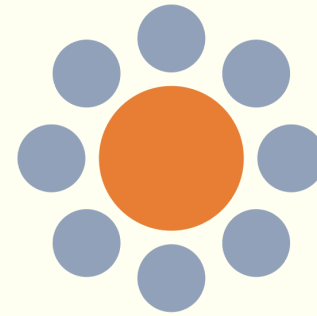
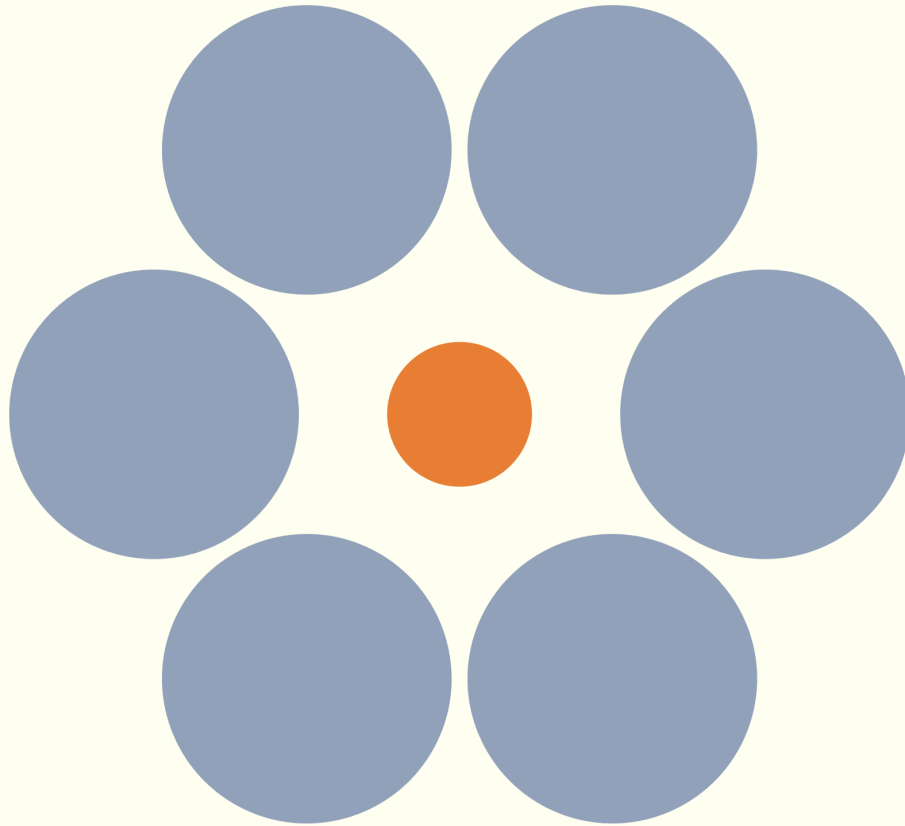


# Key data viz principles

# Theory

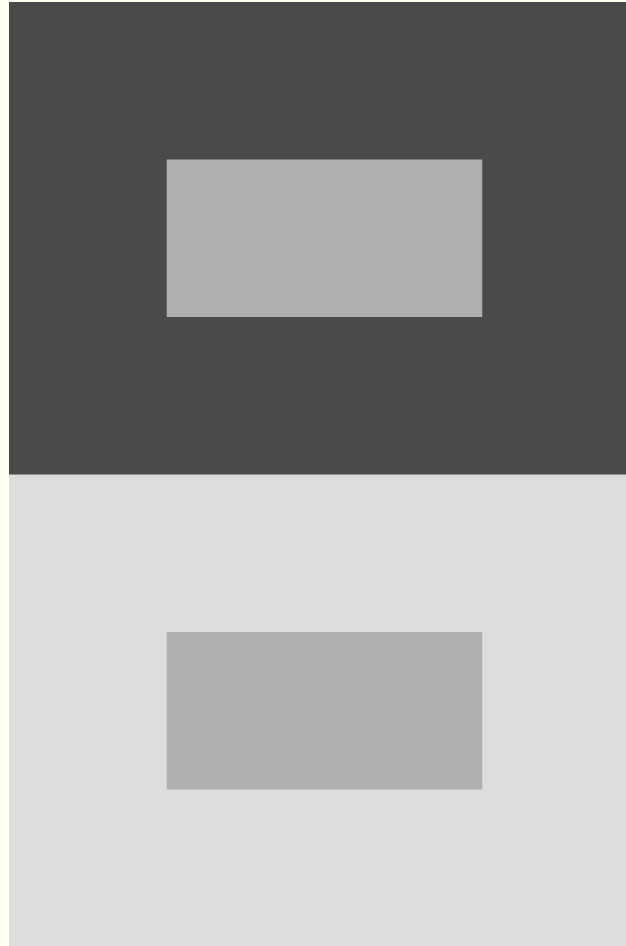
- There is actually a lot of theory behind data viz:
  - Perception,
  - Colors,
  - Design,
  - etc
- Worth learning about it and being aware of key principles
- Leverage it to make better data viz

# Perception



Ebbinghaus illusion

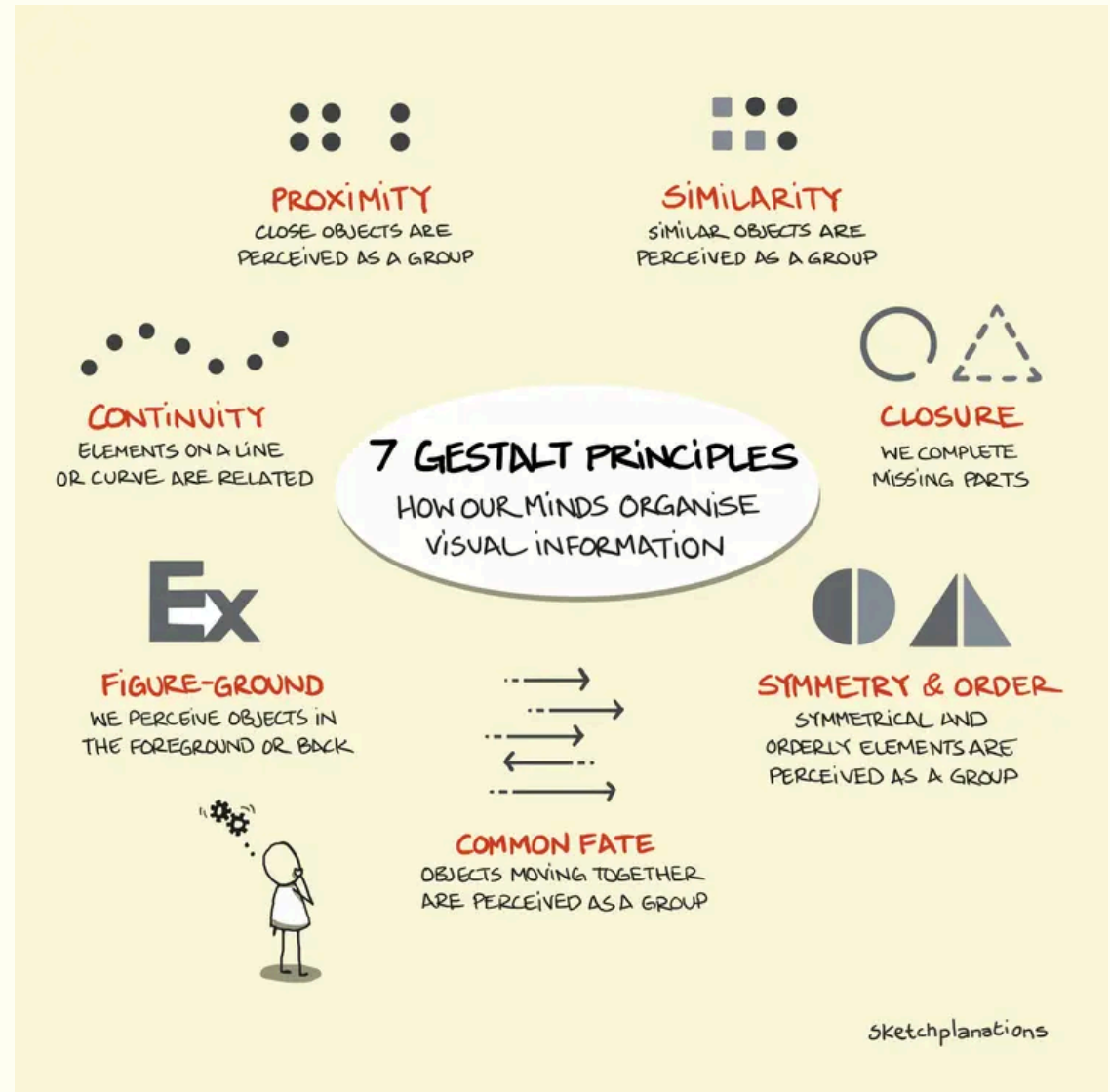
# Relative differences matter



Law of simultaneous contrast

# Gestalt principles

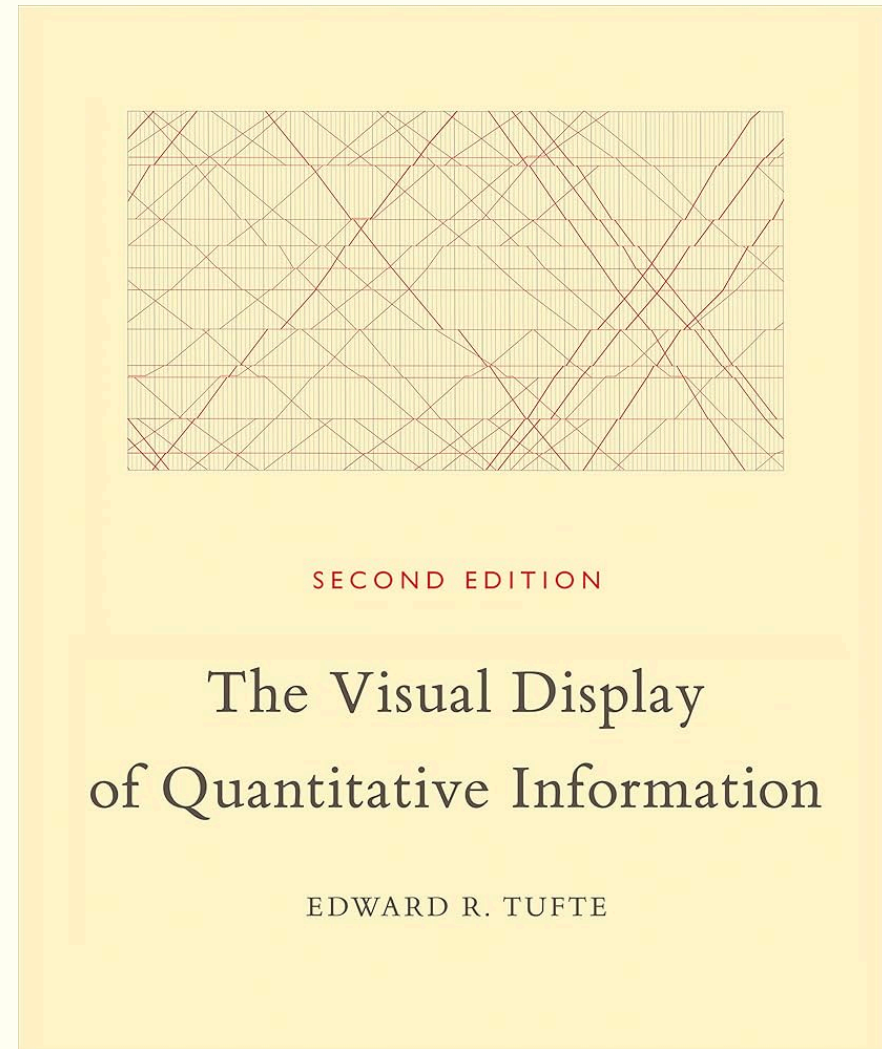
- How our brain interprets what we see
- How it organize visual information
- How we group elements together
- Use them to highlight some patterns and downplay others





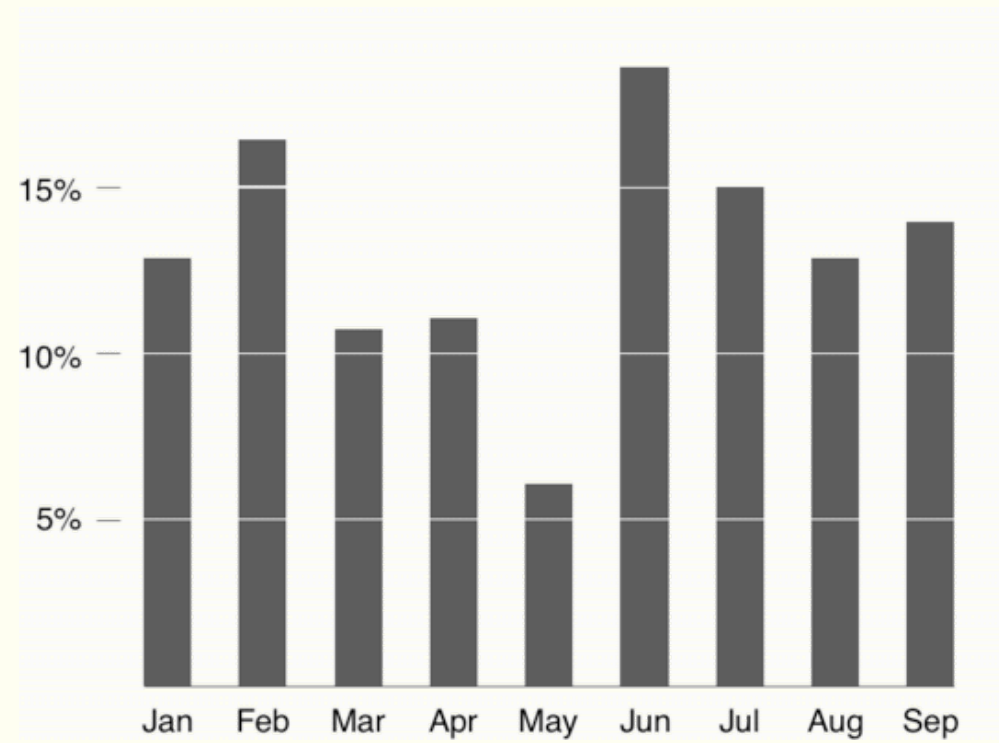
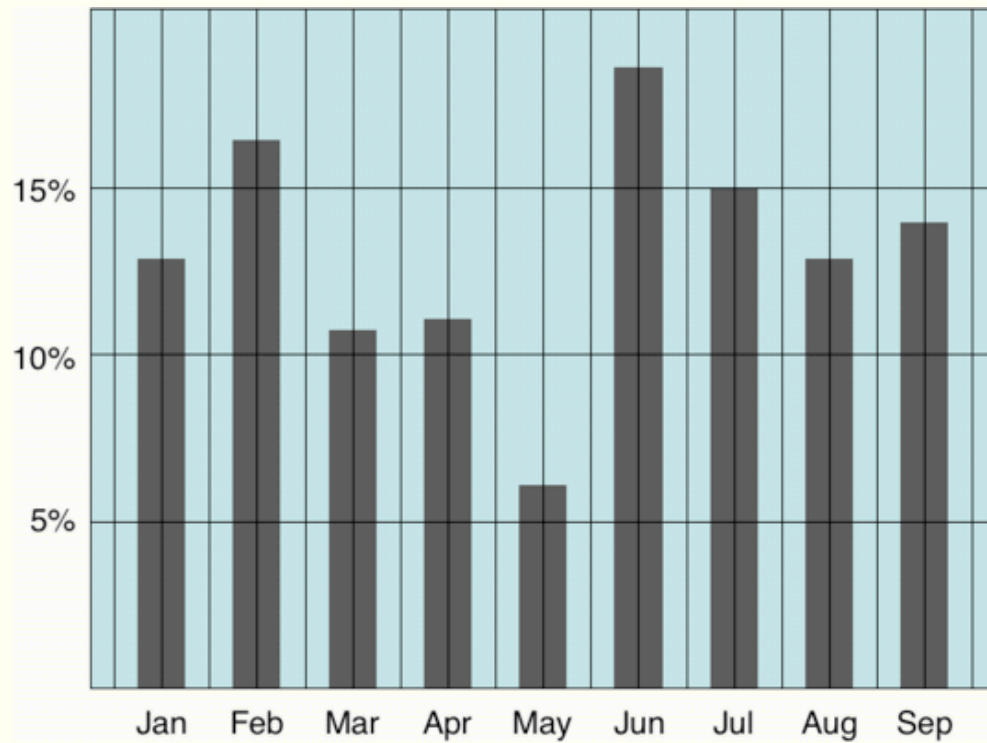
# Data-to-ink ratio

- Introduced by Edward Tufte
- In a nutshell: **avoid clutter**
- Erase non-essential and redundant information
- *“Above all else show the data”*
- Sometimes good to break these rules



# Data-to-ink ratio

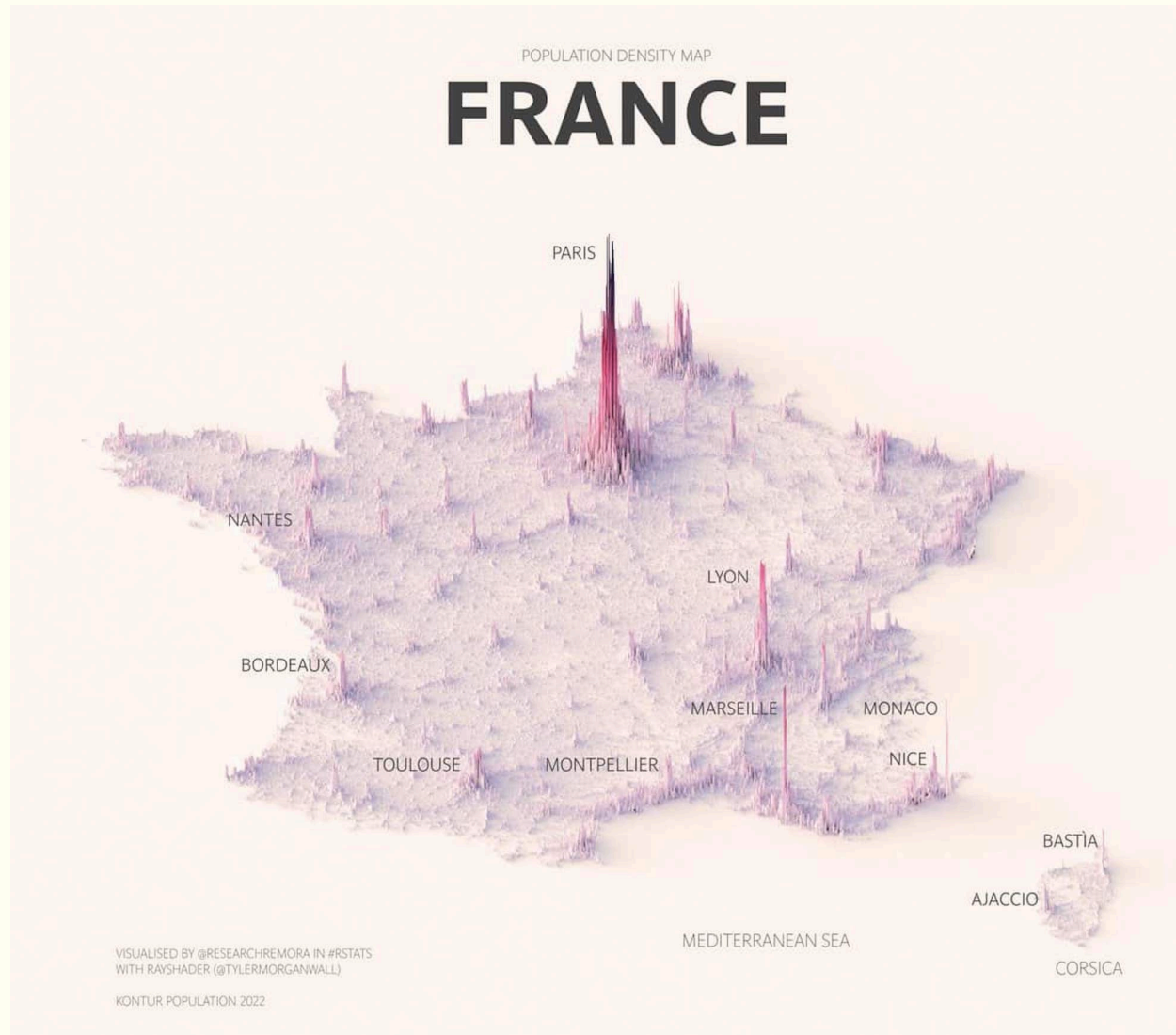
$$\text{data-ink ratio} = \frac{\text{data ink}}{\text{total ink on graph}}$$



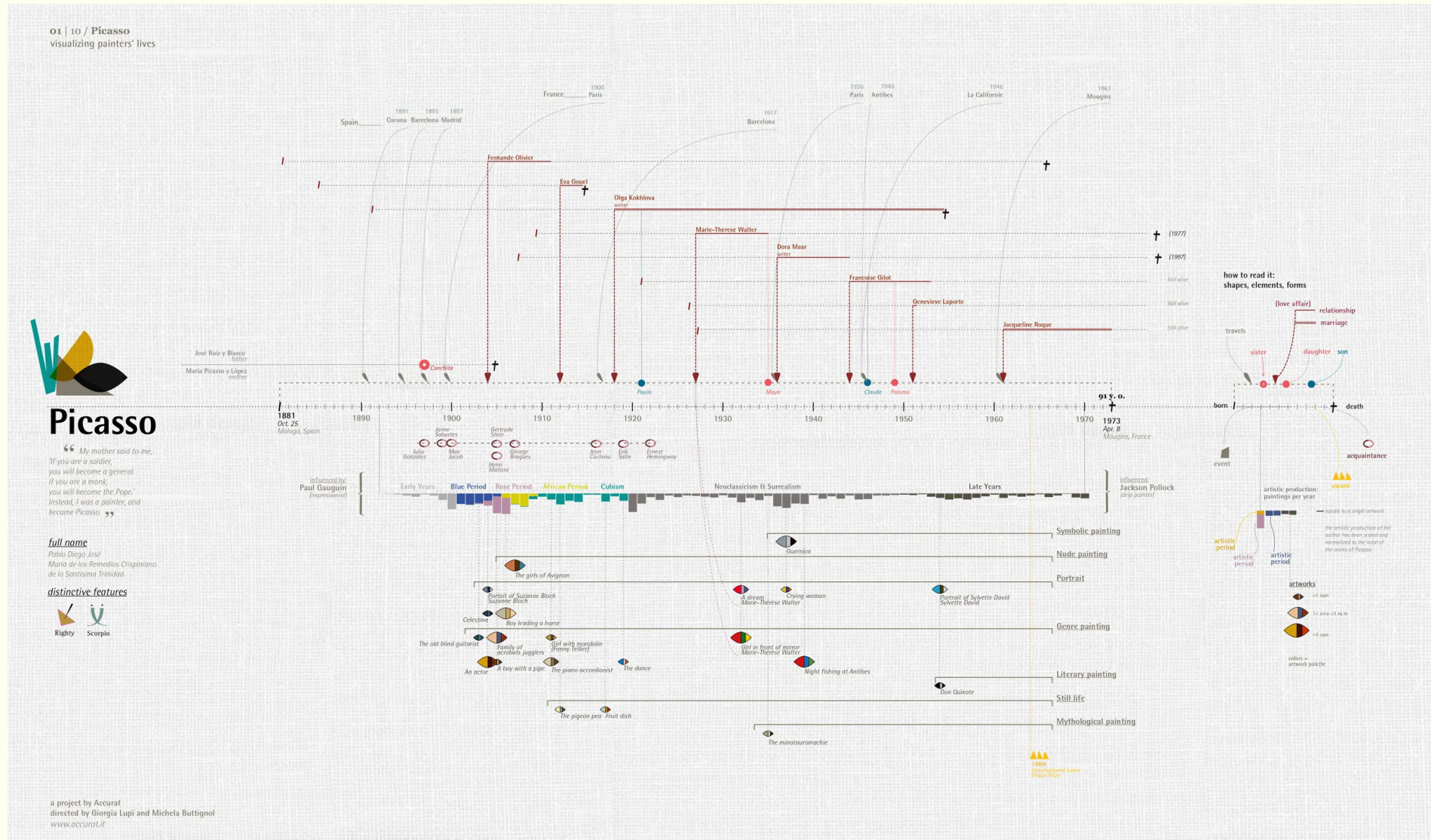
# Graph aesthetics

- Why make nice looking visualizations?
- To trigger interest, to intrigue, to **catch the eye**
- That affects how people perceive information
- Nice looking visuals may be more **memorable**
- In that sense, not everything is *chartjunk*

# Pretty and memorable



# Engaging the audience





# Graph aesthetics in academia

- Pretty graphs are useful in data journalism and so on, but what about academia?
- They are also only more pleasing to look at, they also make readers want to **engage more** with them
- Maybe better to keep the design rather minimalist
- Pretty does not always mean non-simple. Simple graphs have value.
- Opinion on **credibility** partly based on aesthetics

*“This paper did not receive the care it deserved” comment given to a now senior researcher when they submitted a paper with a sketchy graph*

# Credibility and aesthetics

## VIOLET S. MANGANESE

5419 HOLLYWOOD BLVD. STE. C731, LOS ANGELES CA 90027 (323) 555-1435 VIOLET@GMAIL.COM

### Education

**UCLA Anderson School of Management** Los Angeles, California  
**August 2011 to June 2013**  
❖ Cumulative GPA: 3.98  
❖ Academic interests: real-estate financing, corporations, money  
❖ Henry Murtaugh Award

**Hartford University** Cambridge, Massachusetts  
**September 2003 to June 2007**  
❖ B.A. *summa cum laude*, Economics  
❖ Extensive coursework in Astrophysics, Statistics  
❖ Van Damme Scholarship

### Business experience

**Boxer Bedley & Ball Capital Advisors** New York, New York  
**June 2008 to August 2011**  
Equity Analyst  
❖ Performed independent research on numerous American industries, including:  
❖ Steelmaking, croquet, semiotics, and butterscotch manufacturing  
❖ Led company in equities analyzed in two quarters

### Other work experience

**Proximate Cause** Los Angeles, California  
**June 2007 to May 2008**  
Assistant to the Director  
❖ Helped devise fundraising campaigns for this innovative nonprofit  
❖ Handled lunch orders and general errands

**Hot Topic** Boston, Massachusetts  
**February 2004 to March 2006**  
Retail sales associate  
❖ Inventory management  
❖ Training and recruiting

### Skills and interests

- ❖ Fluent in Mandarin, Esperanto; conversational knowledge of Gaelic
- ❖ Writer of U.S. Senate-themed fan fiction
- ❖ Ocean kayaking and free diving
- ❖ Travel, cooking, hiking, playing with my dog
- ❖ Ceramics
- ❖ Backgammon
- ❖ Making paper planes

## TRIXIE B. ARGON

5419 HOLLYWOOD BLVD STE C731, LOS ANGELES CA 90027  
(323) 555 1435 TRIXIEARGON@GMAIL.COM

### EDUCATION

**UCLA Anderson School of Management** 2011–13  
• Cumulative GPA: 3.98  
• Academic interests: real-estate financing, criminal procedure, corporations  
• Henry Murtaugh Award

**Hartford University** 2003–07  
• B.A. *summa cum laude*, Economics  
• Extensive coursework in Astrophysics, Statistics  
• Van Damme Scholarship

### BUSINESS EXPERIENCE

**Boxer Bedley & Ball Capital Advisors** 2008–11  
*Equity analyst*  
• Performed independent research on numerous American industries, including:  
• Steelmaking, croquet, semiotics, and butterscotch manufacturing  
• Led company in equities analyzed in two quarters

### OTHER WORK EXPERIENCE

**Proximate Cause** 2007–08  
*Assistant to the director*  
• Helped devise fundraising campaigns for this innovative nonprofit  
• Handled lunch orders and general errands

**Hot Topic** 2004–06  
*Retail-sales associate*  
• Top in-store sales associate in seven out of eight quarters  
• Inventory management  
• Training and recruiting

# “Originality” VS “familiarity”

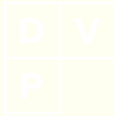
- **Original** graphs may trigger interest
- **Familiar** graphs may convey the point more easily
- My take:
  - Use the best type of graph, regardless of its originality/familiarity
  - If it is different from what people are used to, make it easy to read



# Building a graph

# Know what chart types exist

- Know what is possible to do to find what you need
- Know **graph names** (to search the internet regarding how to code them)
- Refer to existing graph (type) galleries:
  - Data Viz Project
  - Flowing Data - Chart type
  - Dataviz Inspiration
  - R graph gallery



ALL  
([HTTPS://DATAVIZPROJECT.COM/](https://datavizproject.com/))

FAMILY

INPUT

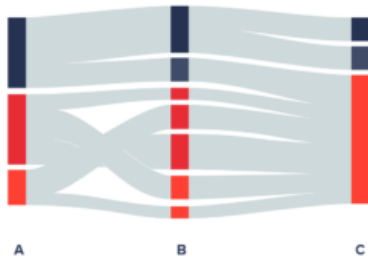
FUNCTION

SHAPE

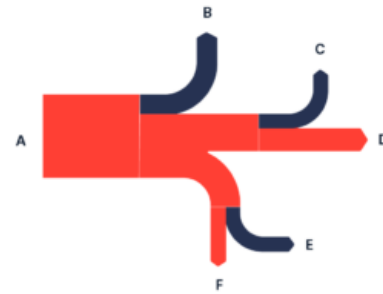
([HTTPS://DATAVIZPROJECT.COM/](https://datavizproject.com/))

(<https://datavizproject.com/>)

**Alluvial Diagram**



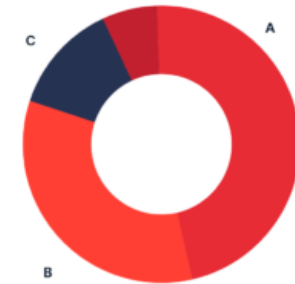
**Sankey Diagram**



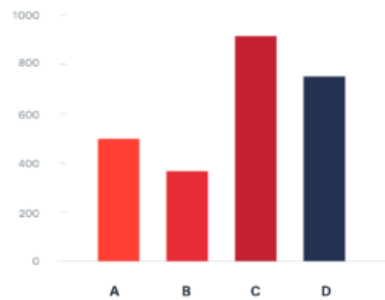
**Line Graph**



**Donut Chart**



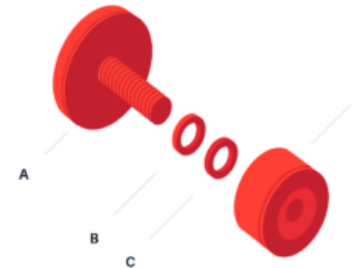
(<https://datavizproject.com/data-> (<https://datavizproject.com/data-> (<https://datavizproject.com/data-> ([\*\*Bar Chart \(Vertical\)\*\*](https://datavizproject.com/data-</a></p></div><div data-bbox=)



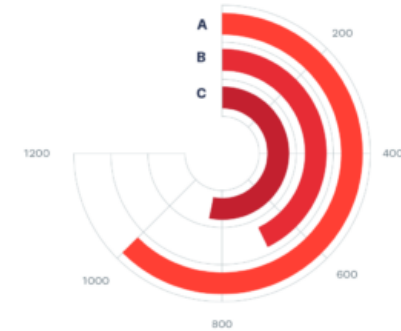
**Polar Area Chart**



**Exploded View Drawing**



**Radial Bar Chart**



# Steps to choose a chart type

- What do you want to show in your graph?
- **What is the main message you want to convey?**
- With the same data, you can:
  - Tell a lot of different stories
  - Emphasize different points
- Making a graph = choosing a lighting for your data

# 1 dataset, many graphs, many stories



1 dataset 100 visualizations

ALL  
([HTTPS://100.DATAVIZPROJECT.COM/](https://100.datavizproject.com/))

STORY

PROPERTY

SHAPE

(HT

(<https://100.datavizproject.com/>)

## 1 dataset 100 visualizations

Can we come up with 100 visualizations from one simple dataset?

As an information design agency working with data visualization every day, we challenged ourselves to accomplish this using insightful and visually appealing visualizations.

We wanted to show the diversity and complexity of data visualization and how we can tell different stories using limited visual properties and assets.

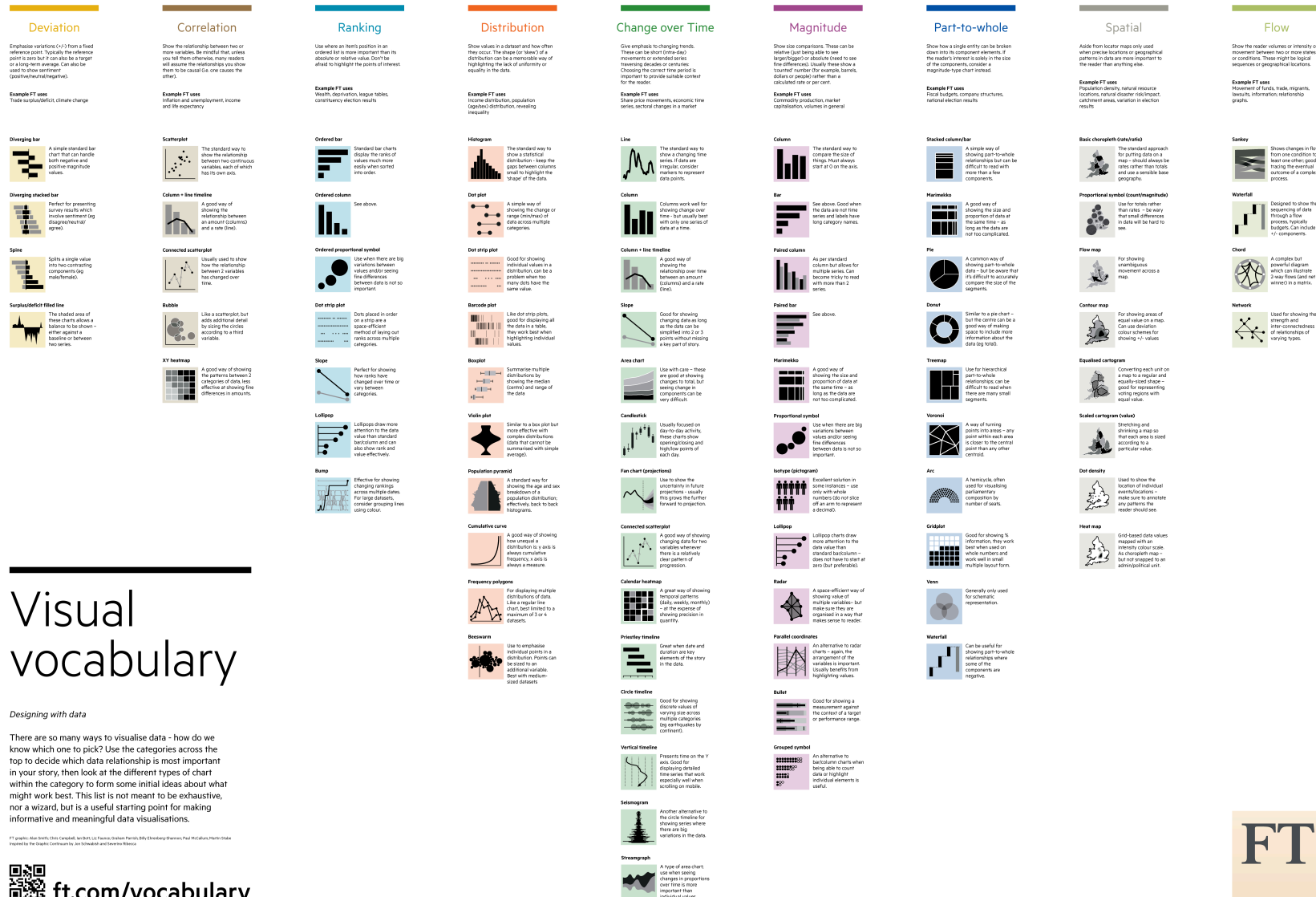
**Number of World Heritage Sites**

# Type of relationship to show

In your graph, you may want to show a:

- Distribution
- Evolution over time
- Magnitude
- Part of a whole
- Ranking
- Geographical patterns
- Flow
- Correlations
- Deviation

# Graph type decision tree

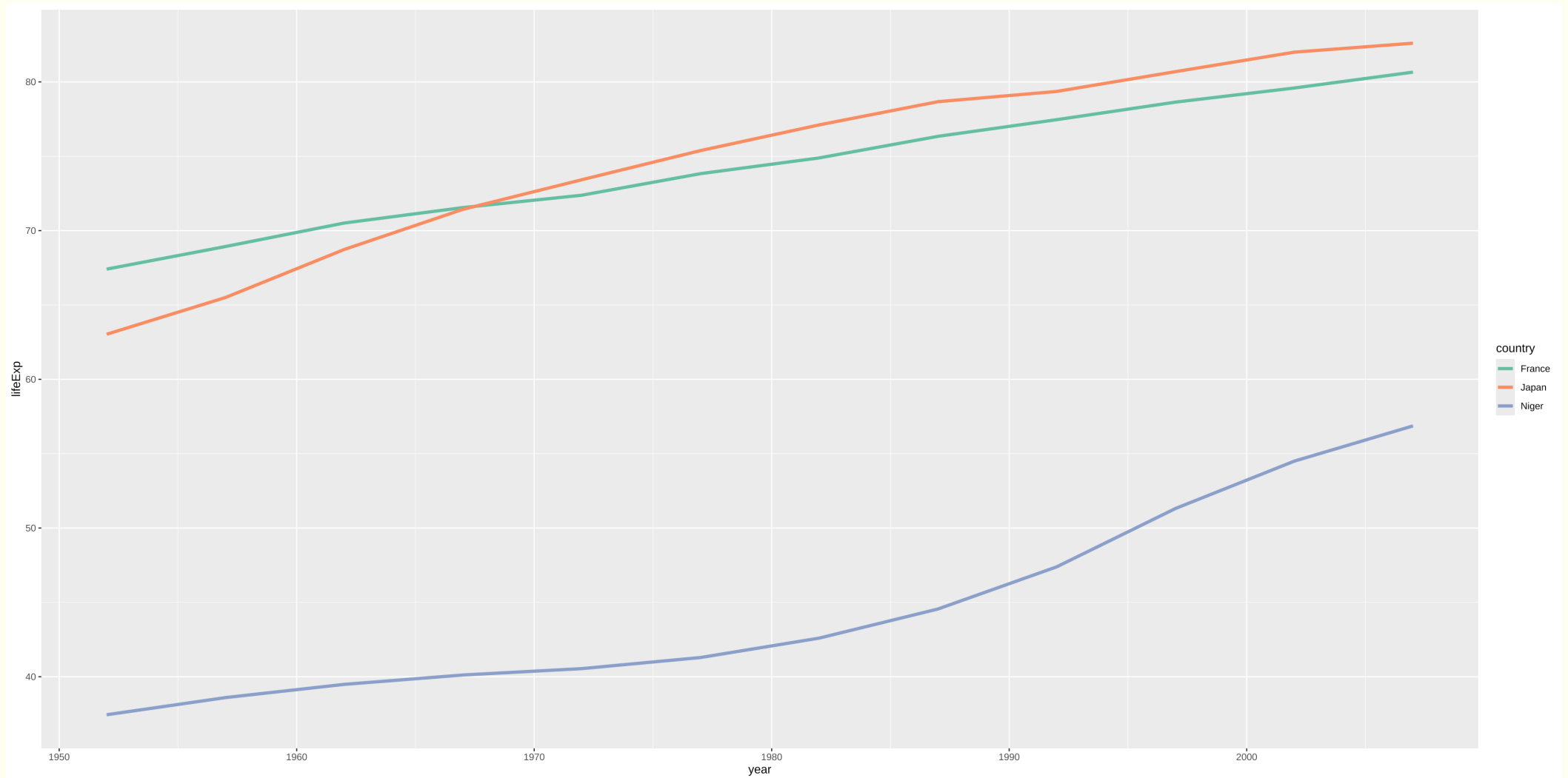


# Example data

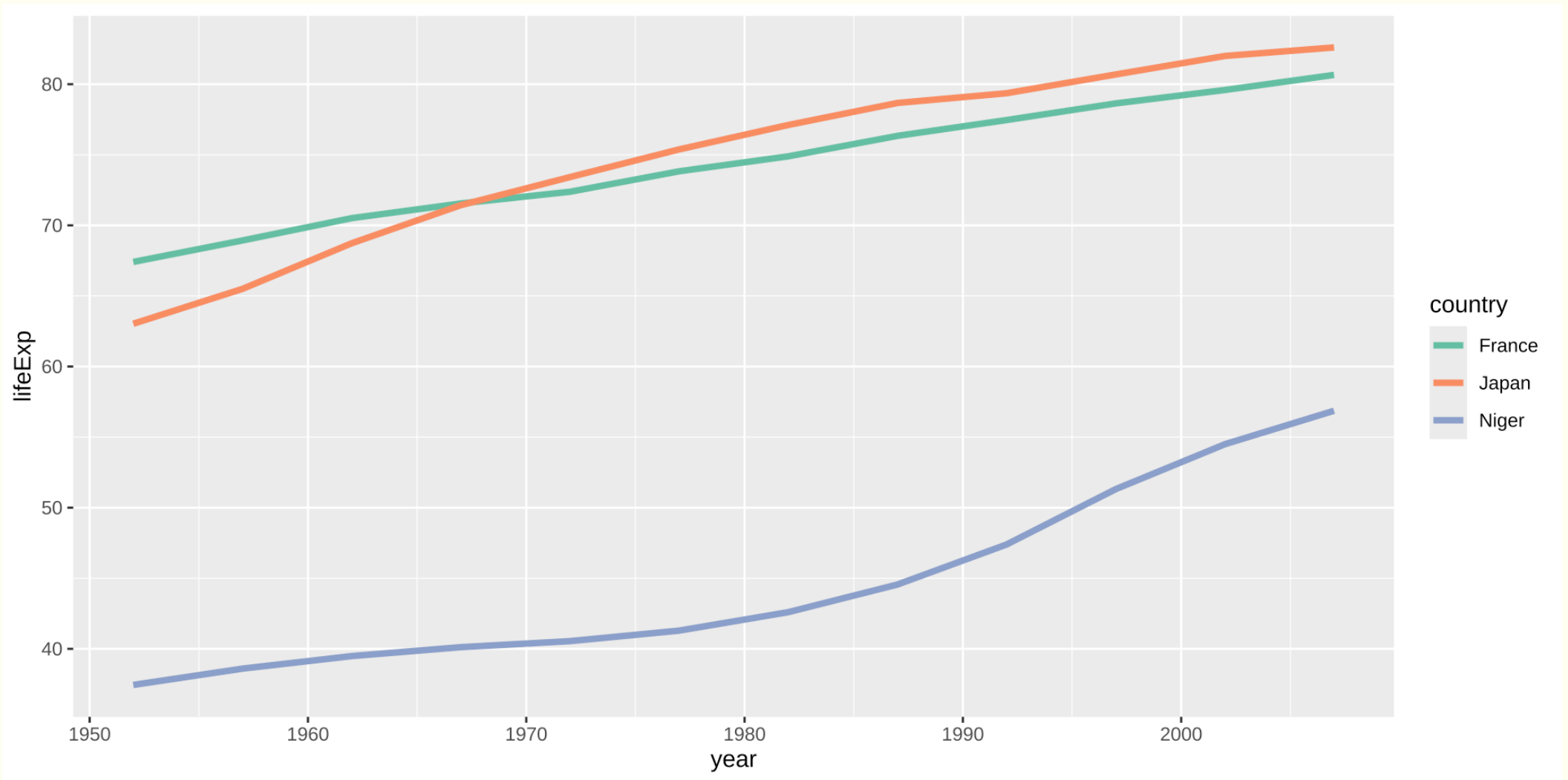
country	year	lifeExp
France	1952	67.410
France	1957	68.930
France	1962	70.510
France	1967	71.550
France	1972	72.380
France	1977	73.830
France	1982	74.890
France	1987	76.340
France	1992	77.460
France	1997	78.640
France	2002	79.590
France	2007	80.657
Japan	1952	63.030
Japan	1957	65.500



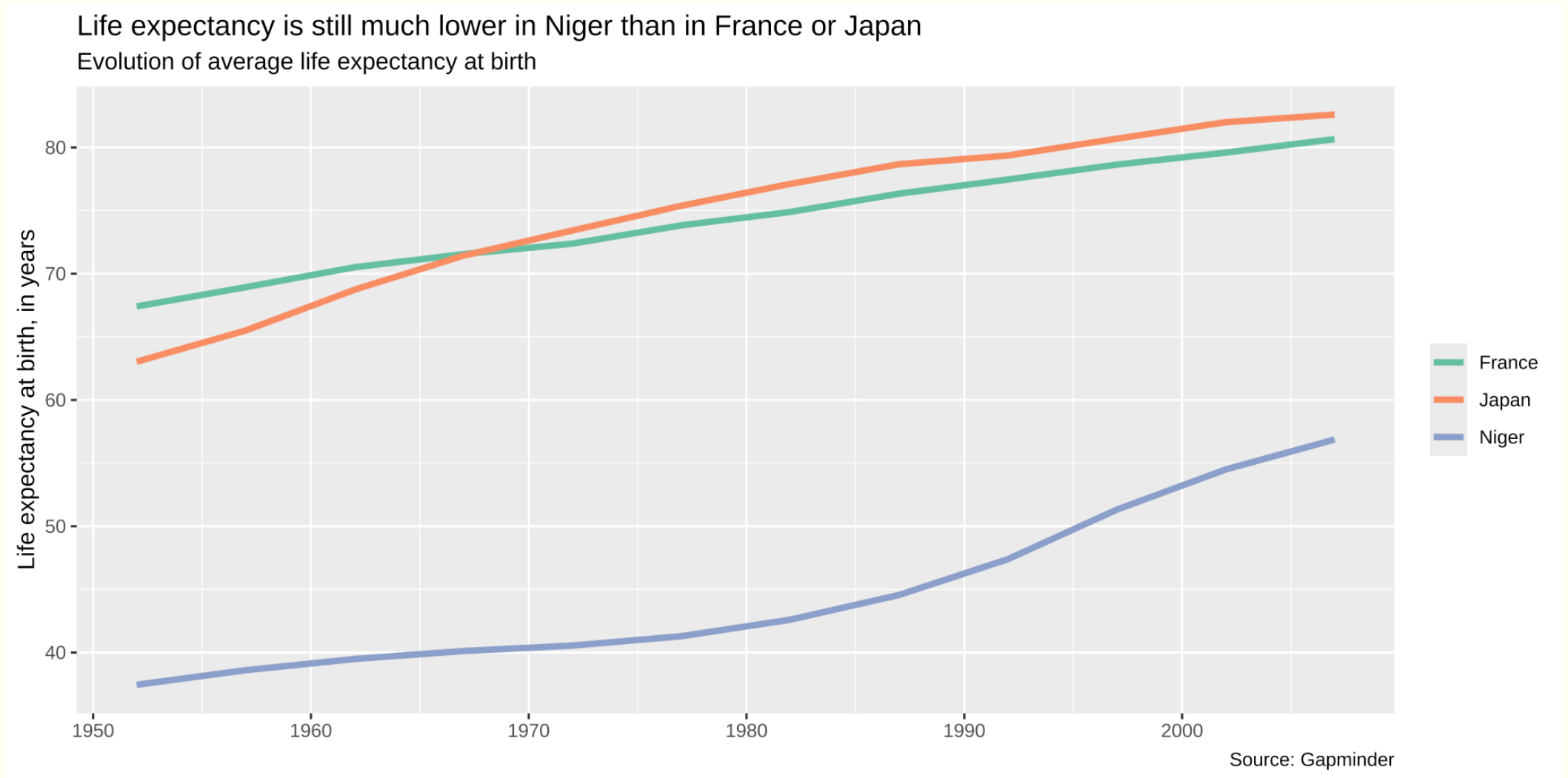
# A concrete example



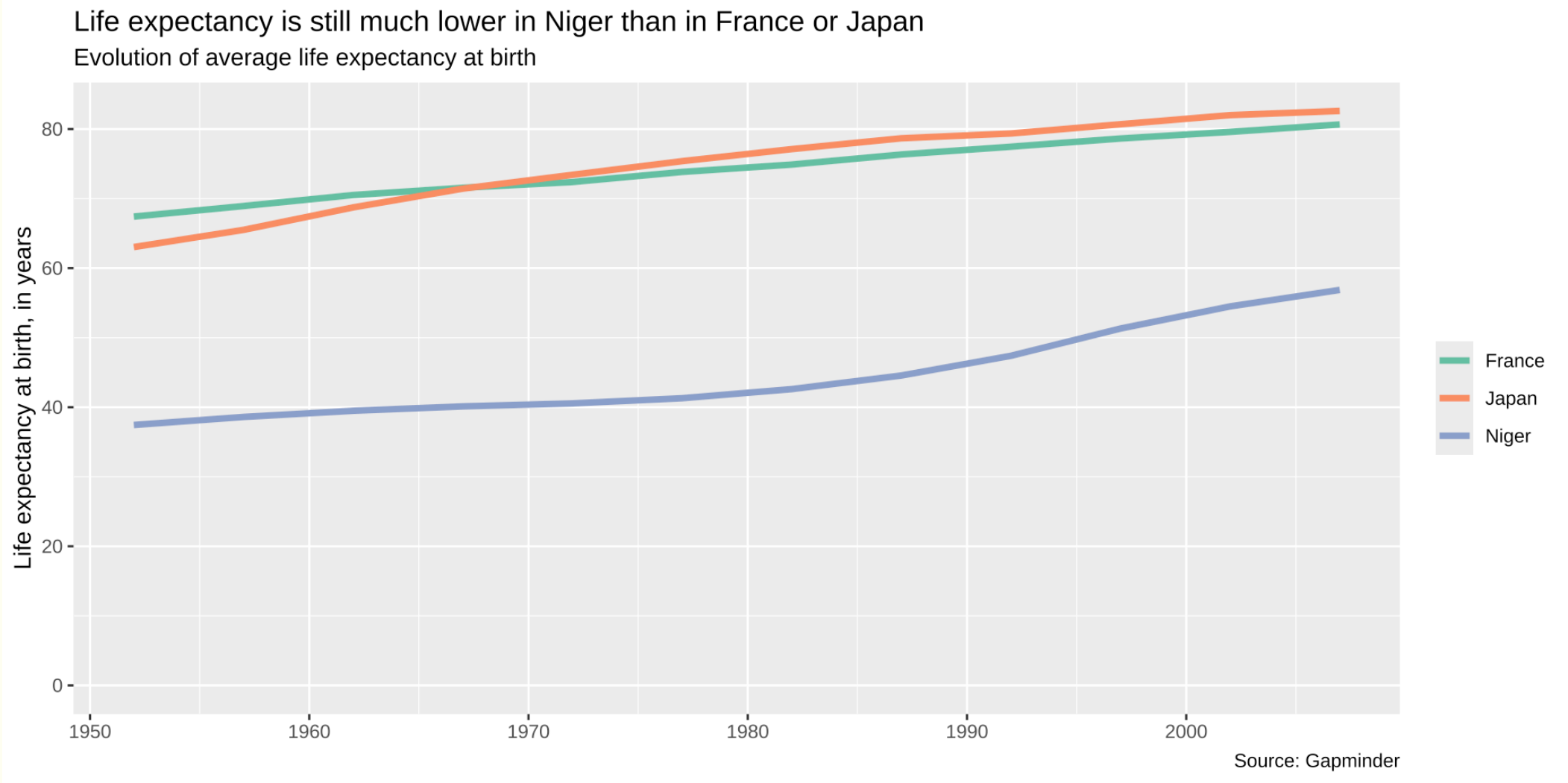
# Legible text



# Title, clear axis labels and source

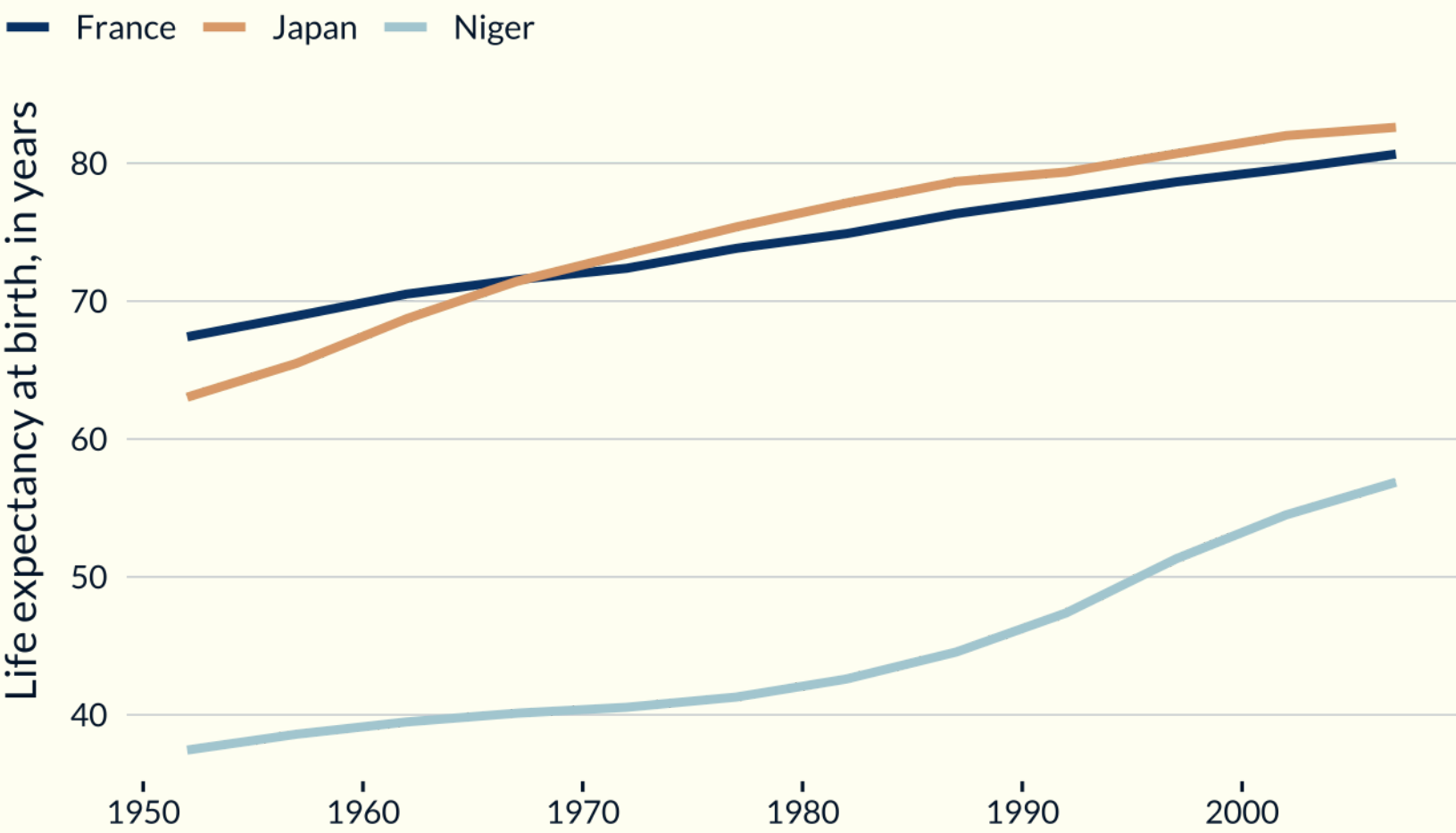


# Scale of the y-axis



# Stylize

Life expectancy is still much lower in Niger than in France or Japan  
*Evolution of average life expectancy at birth*

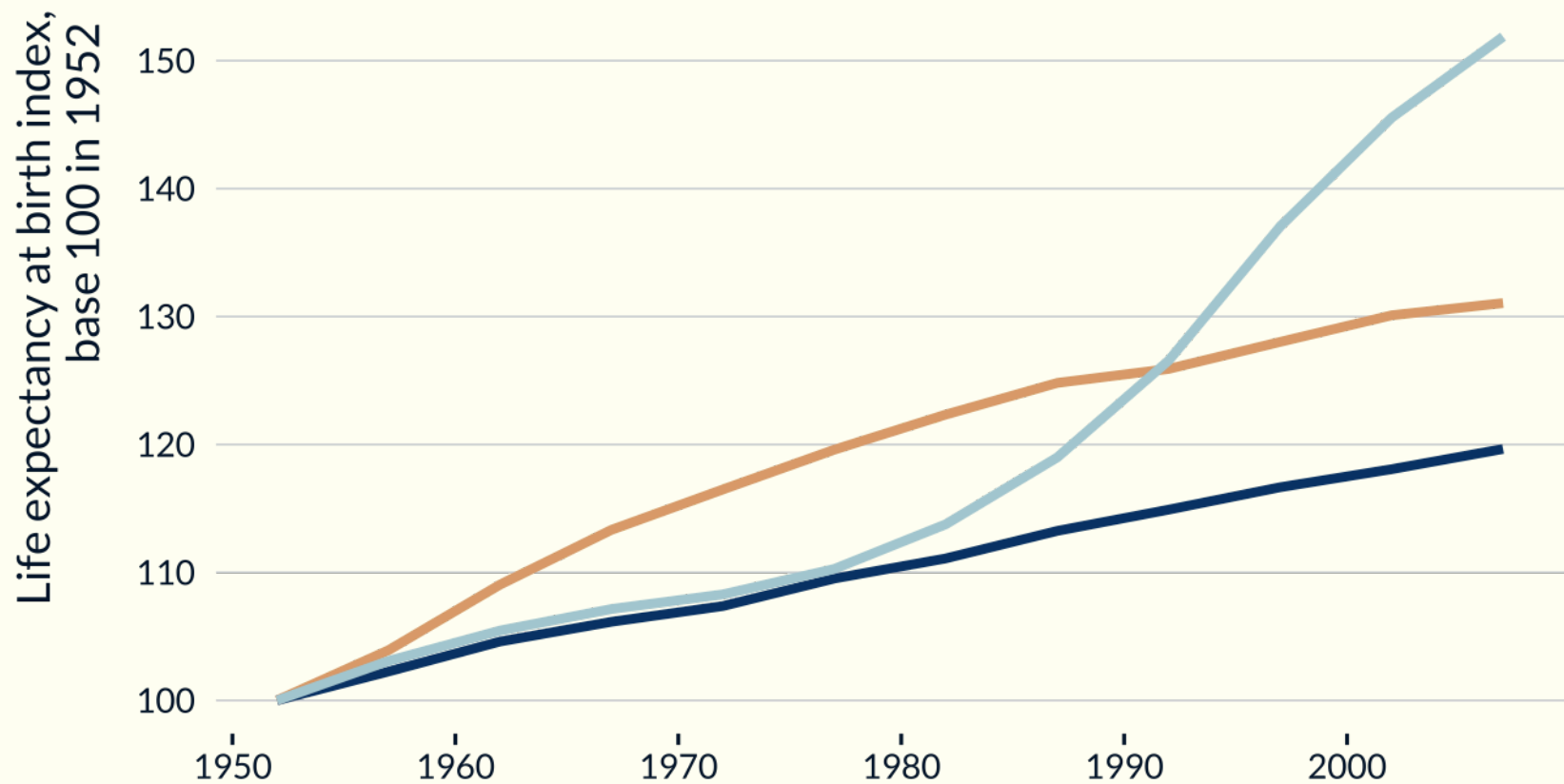


Source: Gapminder

# An alternative story

Life expectancy at birth increased sharply in Niger since 1980  
*Evolution of an average life expectancy at birth index*

■ France ■ Japan ■ Niger



Source: Gapminder

# Graphs in an oral presentation

- Explain orally what your graph represents!
  - What is on the x-axis?
  - What is on the y-axis?
  - What is the message you want to convey?

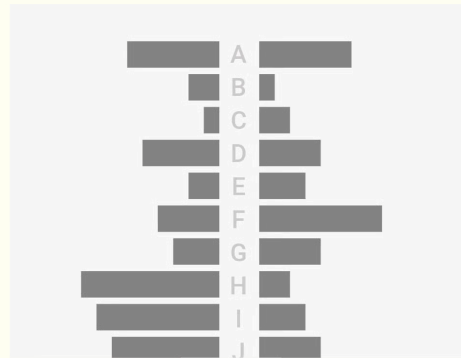
# Avoid using many different colors

If need more than 7 colors or so:

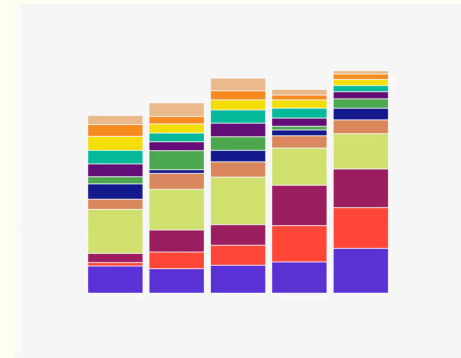
- Use another graph
- Group categories together



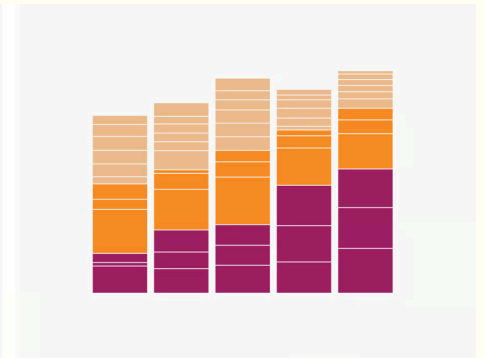
NOT IDEAL



BETTER



NOT IDEAL



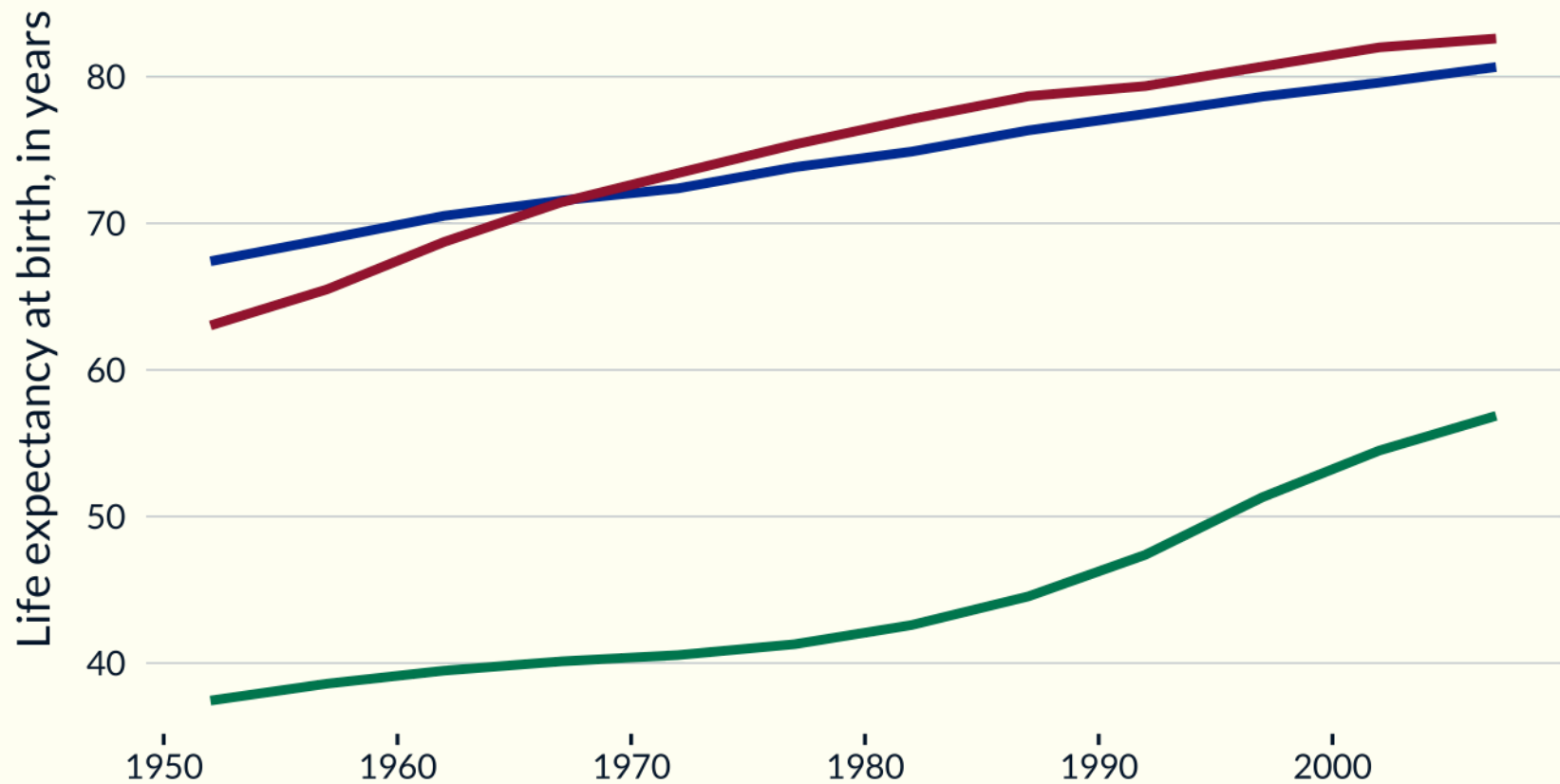
BETTER



# Use intuitive colors

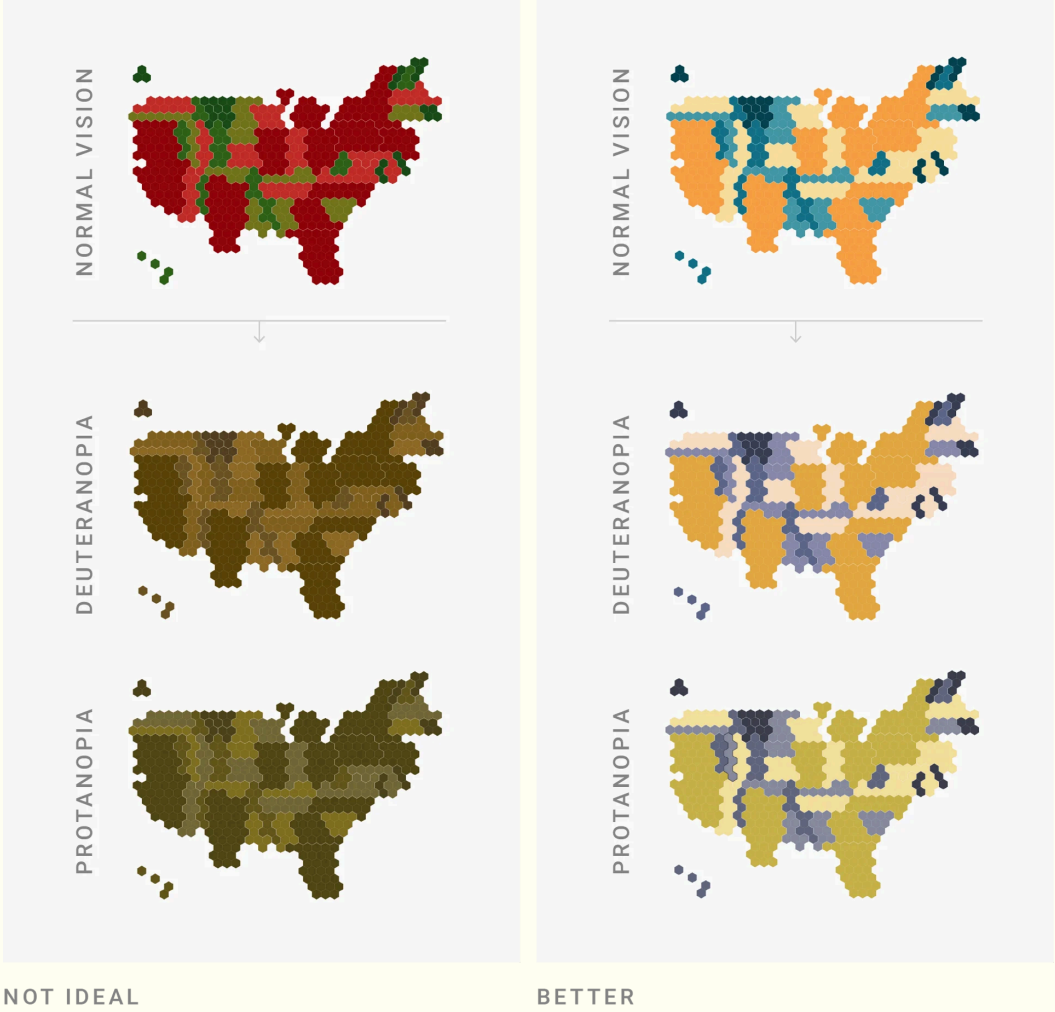
Life expectancy is still much lower in Niger than in France or Japan  
*Evolution of average life expectancy at birth*

■ France ■ Japan ■ Niger



Source: Gapminder

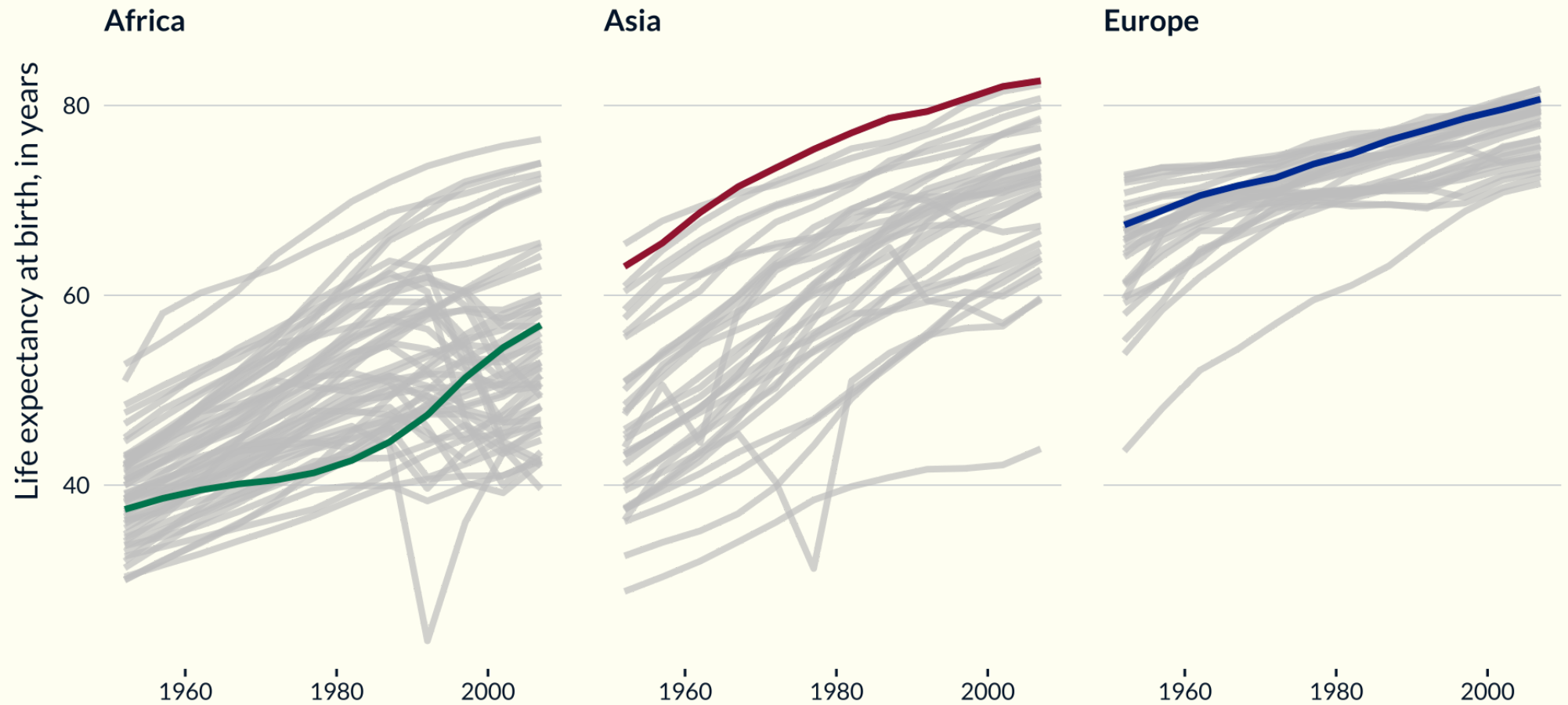
# Colorblind-friendly visualizations



# Use gray. Emphasize.

Life expectancy in France, Japan and Niger as compared to their own continent  
*Evolution of average life expectancy at birth, by continent*

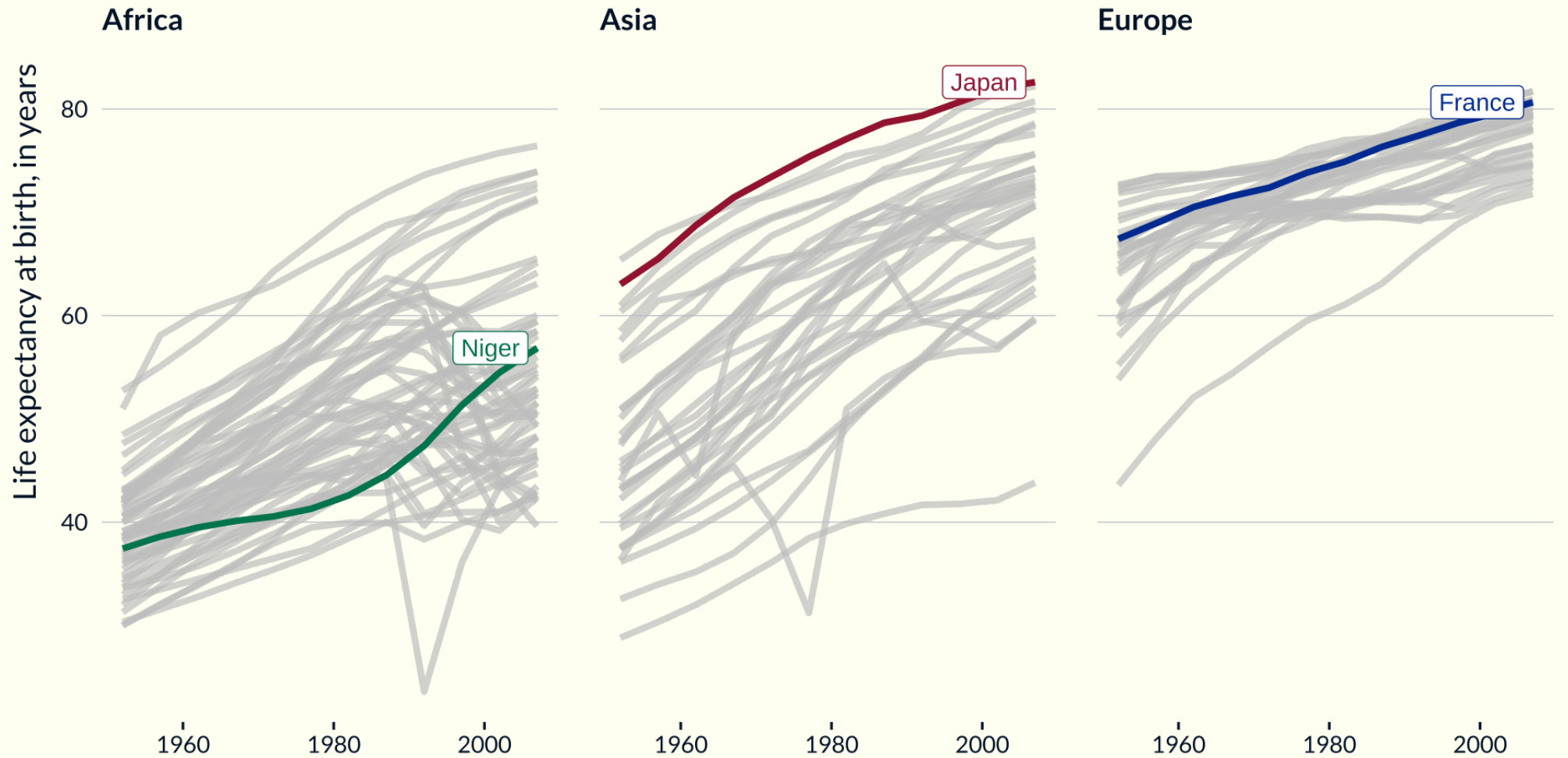
— France — Japan — Niger



Source: Gapminder

# Label directly

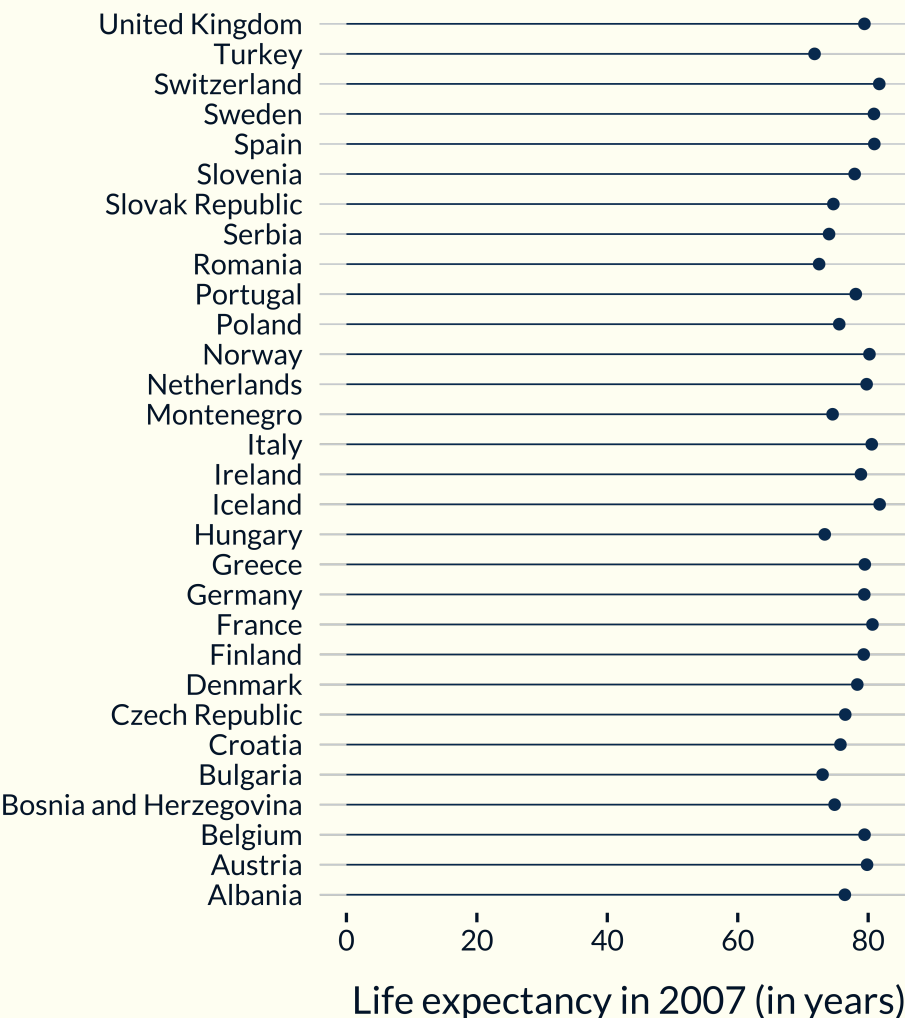
Life expectancy in France, Japan and Niger as compared to their own continent  
*Evolution of average life expectancy at birth, by continent*



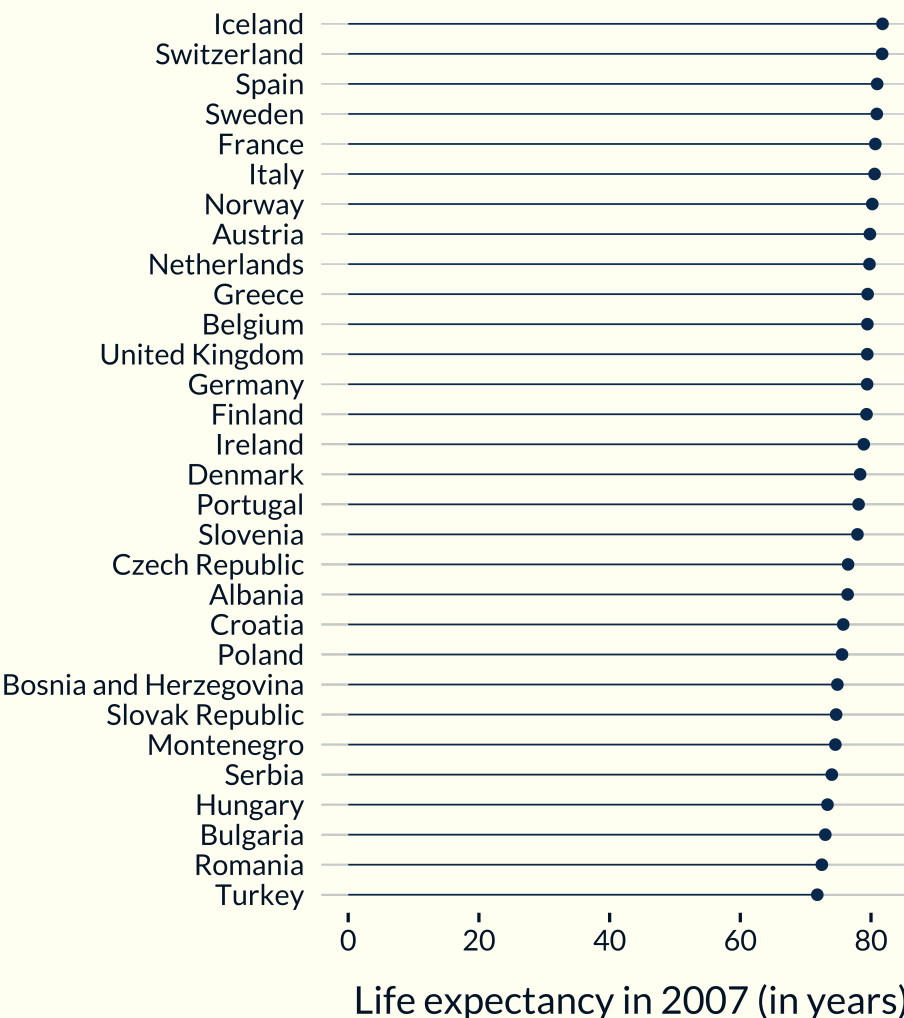
Source: Gapminder

# Order your data

Life Expectancy in Europe



Life Expectancy in Europe



# Data viz caveats

## CAVEATS

*A collection of dataviz caveats by [data-to-viz.com](https://data-to-viz.com)*

Show all

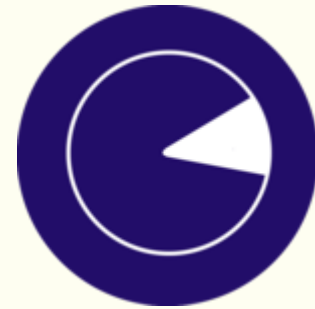
Top 10

Improvement

Misleading

Map

Bar



# Sometimes you may break the rules

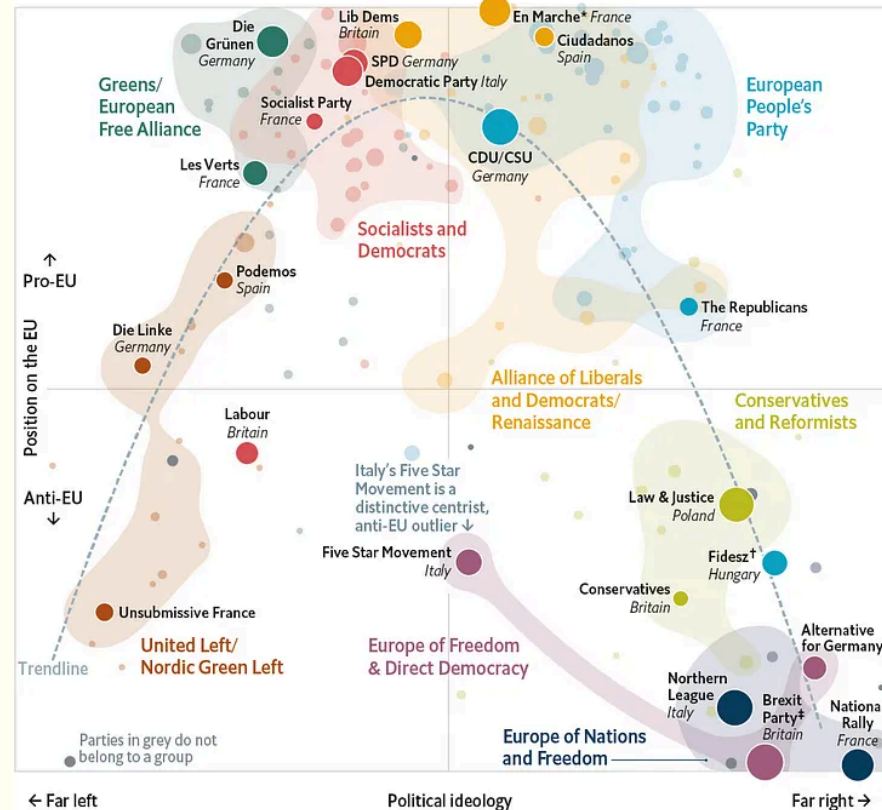
Anti-EU parties cluster at ideological extremes, whereas pro-EU ones are centrist

European Parliament political parties and groupings

By ideology and position on the EU

Party name — Parliamentary grouping

2019 election, provisional results, seats ○ 10 ○ 20



\*Includes MoDem and UDI †Currently suspended from the EPP group ‡Ideology and difference based on UKIP 2014  
Sources: Chapel Hill Expert Survey (2014/2017); ECFR; European Parliament

9 colors

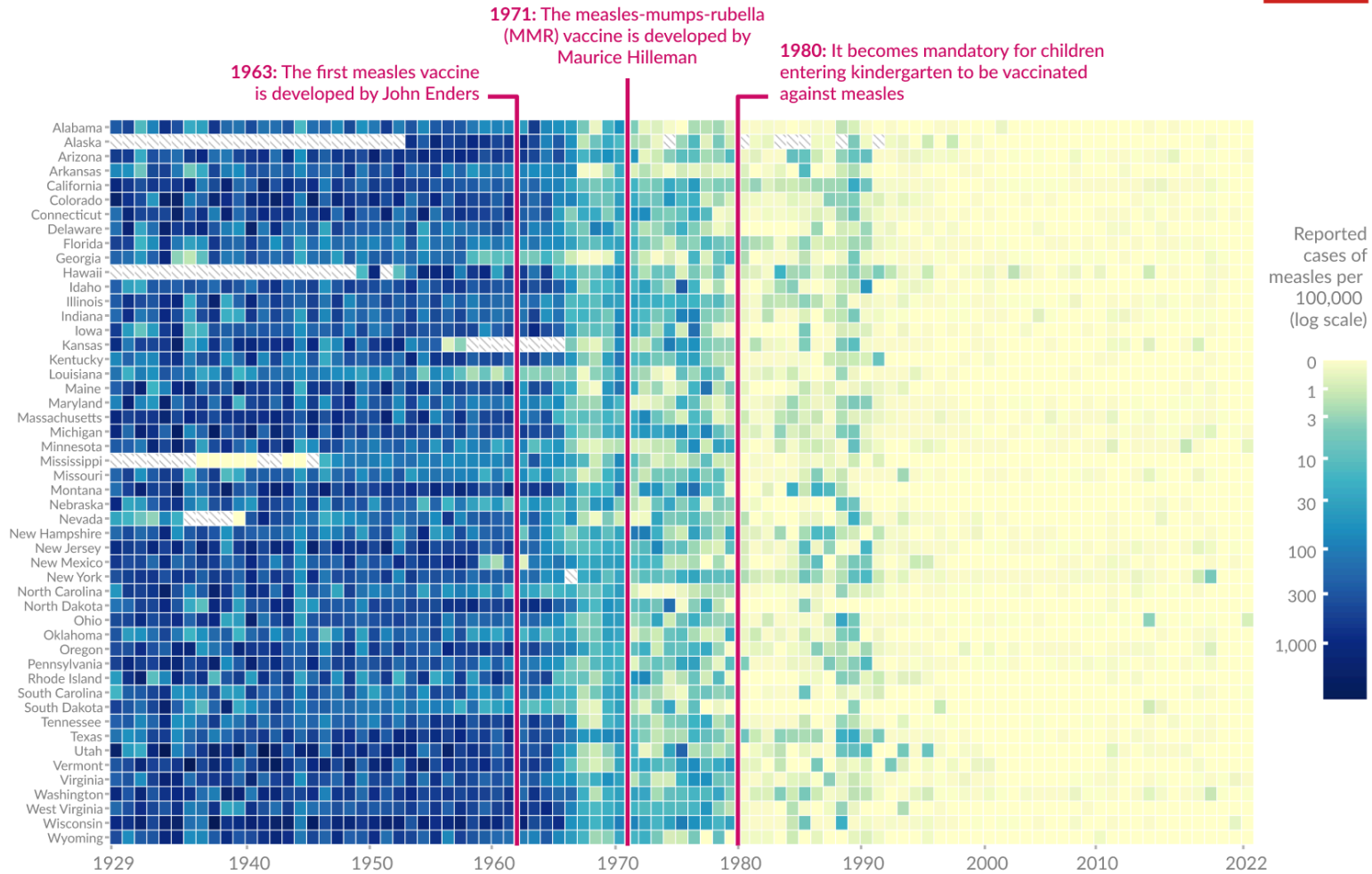
BUT

- Labeled directly
- Shade reinforces grouping

# Nice data viz

## Vaccines reduced measles cases across US states

Our World  
in Data



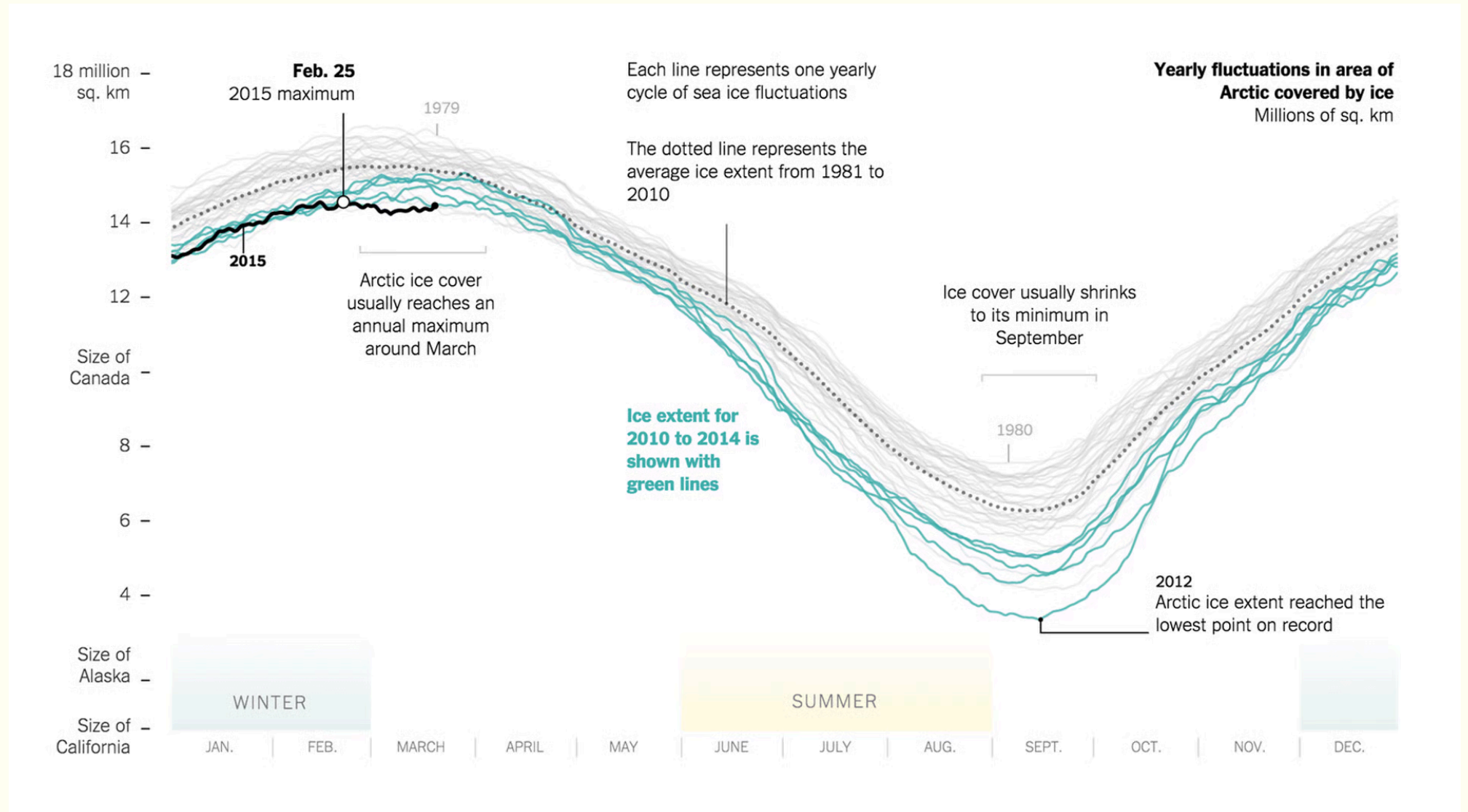
Data source: Project Tycho (2018); Centers for Disease Control and Prevention (1959–2022)

OurWorldinData.org — Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the author Fiona Spooner



# Nice data viz



# Pay attention to data viz

- If you start paying attention to data viz when you see them, you will see what works, what does not
- It is a process that takes place in the “background”: you will learn quickly and not realize it
- You will see nice looking graphs (and hopefully enjoy it)
- You will build better and more impactful graphs i

# Summary of concrete recommendations

## Take-away messages

- Build legible, understandable and nice looking graphs.
- Have a title and explicit axes; present them.
- Limit the number of colors you use. Use gray.
- Label your graphs directly, add annotations.
- Think twice before cutting the y-axis
- Overall, facilitate the retrieval of information.

# Data viz in economics

# Specificities of viz in economics

- We often have to deal with a **massive number of observations**
- We often want to display a specific type of graph: **estimation output**
- Our analysis are often based on **identifying assumptions** that we can sometimes check through graphs
- We often use very complex models. Visualization can help **understand what we are actually estimating**

# Graphs hierarchy

- We make graphs for different audiences
- They thus need to be more or less polished
  - For **you** (and your future self): can be quite rough on the edges, but you will want to be able to understand it in the future
  - For **presentations**: you have some leeway for explaining orally your graphs
  - For the **paper**: there is only a couple of graphs in a paper; make them perfect

# When do we use graphs in econ?

- As **rhetorical** visualization tools for models
- To **explore** our data
- To check the validity of our **models**
- As **diagnostics**
- To **communicate** results

# Look at your data

Data	What's in it?	Summary	Take-aways	Cleaning
------	---------------	---------	------------	----------

```
1 library(AER)
2 data("Fatalities")
3
4 Fatalities |> head(5) |> kable()
```

state	year	spirits	unemp	income	emppop	beertax	baptist	mormon	drin
al	1982	1.37	14.4	10544.15	50.69204	1.539379	30.3557	0.32829	1
al	1983	1.36	13.7	10732.80	52.14703	1.788991	30.3336	0.34341	1
al	1984	1.32	11.1	11108.79	54.16809	1.714286	30.3115	0.35924	1
al	1985	1.28	8.9	11332.63	55.27114	1.652542	30.2895	0.37579	1
al	1986	1.23	9.8	11661.51	56.51450	1.609907	30.2674	0.39311	2



# Explore and understand relationships

Code for graphs

Graph in levels

Graph in logs

```
1 fatal_jail <- fatalities |>
2   ggplot(aes(x = jail_name, y = fatal)) +
3   # geom_hline(aes(yintercept = mean(fatal, na.rm = TRUE))) +
4   geom_jitter(width = 0.25) +
5   labs(
6     title = "Jail penalty for drunk driving and fatal accidents",
7     y = "Number of fatal accidents per state",
8     x = "State law on drunk driving"
9   )
10
11 fatal_jail_log <- fatal_jail +
12   scale_y_log10() +
13   labs(y = "Number of fatal accidents\nper state (log scale)")
```

# Finding sources of variation

Code

Evolution

Map

```
1 # Evolution of law
2 law_evol <- fatalities |>
3   group_by(year) |>
4   # summarise(prop_jail = mean(jail, na.rm = TRUE))
5   summarise(prop_jail = mean(jail, na.rm = TRUE)) |>
6   ggplot(aes(year, prop_jail)) +
7   geom_line() +
8   labs(
9     title = "Adoption of jail sentence for drunk driving",
10    y = "Proportion of states with a jail law",
11    x = NULL
12  )
13
14 # Making the map
15 states_sf <- tigris::states(
16   cb = TRUE, resolution = "20m", year = 2024, progress_bar = FALSE) |>
17   tigris::shift_geometry() |>
18   rename(state = STUSPS)
19
20 fatalities_sf <- fatalities |>
21   filter(jail) |>
22   group_by(state) |>
23   mutate(first_year = min(year, na.rm = TRUE)) |>
24   ungroup() |>
25   filter(year == first_year) |>
26   dplyr::full_join(states_sf, by = join_by(state)) |>
27   sf::st_as_sf()
28
29 law_map <- fatalities_sf |>
```

# Balance plots (explore and communicate)

Code

Unemployment

Income

Spirits consumption

```
1 graph_balance <- function(balance_var) {  
2   fatalities |>  
3   ggplot(aes(x = {{ balance_var }}, fill = jail_name, color = jail_name)) +  
4   geom_density() +  
5   labs(  
6     title = paste("Balance plot for", substitute(balance_var)),  
7     fill = NULL,  
8     color = NULL,  
9     y = "Density"  
10  )  
11 }
```

# Balance table

Why?

Table

---

- **Formal statistical tests**
- Graphs only provide visual information that can hide informations or lack clarity

```
1 balance_table <- fatalities |>  
2   select(-state) |>  
3   modelsummary::datasummary_balance(formula = ~jail)
```

# Identification strategy

- The setting calls for a TWFE approach (staggered roll-out)
- Identifying assumptions?
- Threats to identification?
  - Trends in number of fatalities before adoption
  - Other shocks at the time of adoption
  - Anticipation: decrease in fatalities before the implementation
  - Change in composition of the states
  - Spillover effects
- How to explore these? Are graphs helpful?

# Parallel trends and event study graph

Code

Event study graph

```
1 first_year <- fatalities |>
2   filter(jail) |>
3   group_by(state) |>
4   mutate(treat_year = min(year, na.rm = TRUE)) |>
5   ungroup() |>
6   filter(year == treat_year) |>
7   select(treat_year, state)
8
9 dat <- fatalities |>
10  left_join(first_year, by = join_by(state)) |>
11  mutate(
12    first_year = replace_na(treat_year, 0)
13  )
14
15 event_study_reg <- feols(
16   log(fatal) ~ sunab(treat_year, year) | state + year,
17   data = dat
18 )
19
20 event_study_graph <- event_study_reg |>
21   tidy(conf.int = TRUE) |>
22   mutate(term = as.integer(str_remove(term, "year::"))) |>
23   rbind(c(-1, rep(0, 6))) |>
24   ggplot(aes(x = term, y = estimate)) +
25   geom_point() +
26   geom_pointrange(aes(ymin = conf.low, ymax = conf.high)) +
27   geom_vline(xintercept = -1) +
28   labs(
29     title = "Event study graph: law and impact on fatalities"
```

# Model specification

- We will discuss that in the last session
- Along with inference aspects
- Let's just consider a simple model for now, even though it might present obvious issues

# Estimation

```
1 reg <- feols(  
2   data = fatalities,  
3   log(fatal) ~ jail | state + year  
4 )  
5  
6 reg_table <- msummary(  
7   list(`log fatalities` = reg),  
8   gof_omit = "IC|Log|R",  
9   fmt = 4  
10 )
```

log fatalities	
jailTRUE	0.0247
	(0.0444)
Num.Obs.	335
Std.Errors	by: state
FE: state	X
FE: year	X



# Visualizing estimation output

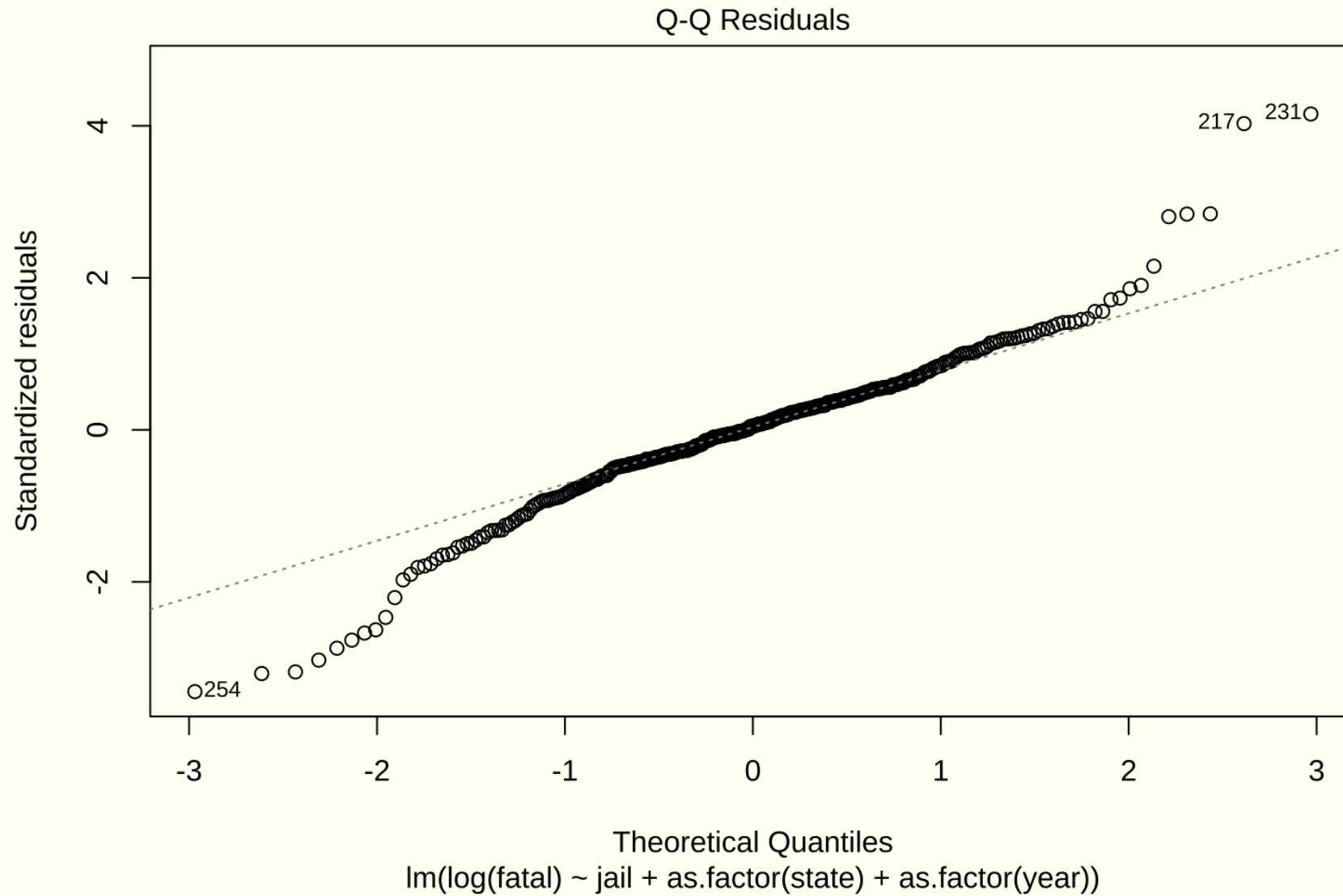
Code

Coef plot

Distribution

```
1 reg_ctrl <- feols(  
2   data = fatalities,  
3   log(fatal) ~ jail + log(unemp) + log(income) | state + year  
4 )  
5  
6 coef_plot <- modelsummary::modelplot(list(`No controls` = reg, `Controls` = reg_ctrl)) +  
7   geom_vline(xintercept = 0) +  
8   labs(title = "Coefficient plot", caption = "R package: modelsummary")  
9  
10 distrib_coef_plot <- reg_ctrl |>  
11   broom::tidy() |>  
12   ggplot(aes(y = term)) +  
13   ggdist::stat_halfeye(  
14     aes(xdist = distributional::dist_student_t(  
15       df = df.residual(reg), mu = estimate, sigma = std.error)),  
16     fill = colors_mediocre$complementary,  
17     color = colors_mediocre$base,  
18     # size = 10,  
19     alpha = 0.6  
20   ) +  
21   geom_vline(xintercept = 0) +  
22   labs(  
23     title = "Distribution of estimates",  
24     x = "Point estimate",  
25     y = NULL,  
26     caption = "R package: ggdist"  
27   )
```

# Diagnostic plots



# Handling large numbers of observations

Code

Raw

Opacity

Heatmap

Binscatter

```
1 ex_duration <- readRDS("~/Documents/Teaching/data_viz_summer/content/slides/data/ex_duration.RDS")
2 ex_duration <- readRDS("data/ex_duration.RDS")
3
4 raw <- ex_duration |>
5   ggplot(aes(x = date, y = duration)) +
6   geom_point() +
7   scale_y_log10() +
8   labs(
9     title = "Evolution of the duration of items in time",
10    x = NULL,
11    y = "Duration (in s)"
12  )
13
14 opacity <- ex_duration |>
15   ggplot(aes(x = date, y = duration)) +
16   geom_point(alpha = 0.01) +
17   scale_y_log10() +
18   labs(
19     title = "Evolution of the duration of items in time, by channel",
20    x = NULL,
21    y = "Duration (in s)"
22  ) +
23   facet_wrap(~ channel)
24
25 heat_map <- ex_duration |>
26   ggplot(aes(x = date, y = duration)) +
27   geom_bin2d(bins = 70) +
28   scale_y_log10() +
29   labs(
```

# Main take-away points

# Data viz at large

- Data viz is **powerful**, harness its power
- It can be super **insightful** or equally **deceptive**
- It can make your point **memorable**
- It can also be truly **beautiful**
- Leverage perception and data viz **principles**

# How to build a graph?

## Take-away messages

- There are **many rules** in data viz
- The main goal is to **facilitate the transmission of your message**
- What is the main point you want to convey?
- Choose (one of) the right graph types
- Explain your graphs, orally and by facilitating reading, on your graph
- BUT avoid clutter

# Data viz for economics

- Most data viz principles and ideas also apply to economics and academia in general
- There are however some specificities
- In particular, some graphs and **types of analyses** are specific to academic research
- Data viz can be extremely useful for research and communication

**Thanks**