DATA ANALYSIS AND MACHINE LEARNING · WEEK 3

Exploratory Data Analysis and Visualization

Javier Garcia-Bernardo

Department of Methodology & Statistics, Utrecht University

Builds on earlier versions by Daniel L. Oberski and Robert A. Bagheri

Reading: R for Data Science (2nd ed.), chapters 1, 9 and 10 · r4ds.hadley.nz

Student Assistant: Web Scraping

Skills:

- Python or Javascript without AI
- SQL
- Web technologies (HTTP, HTML, website logins)
- Basic Unix / SSH
- Ideal: Work with web crawlers / scrapers

Contact: j.m.b.haslbeck@uu.nl

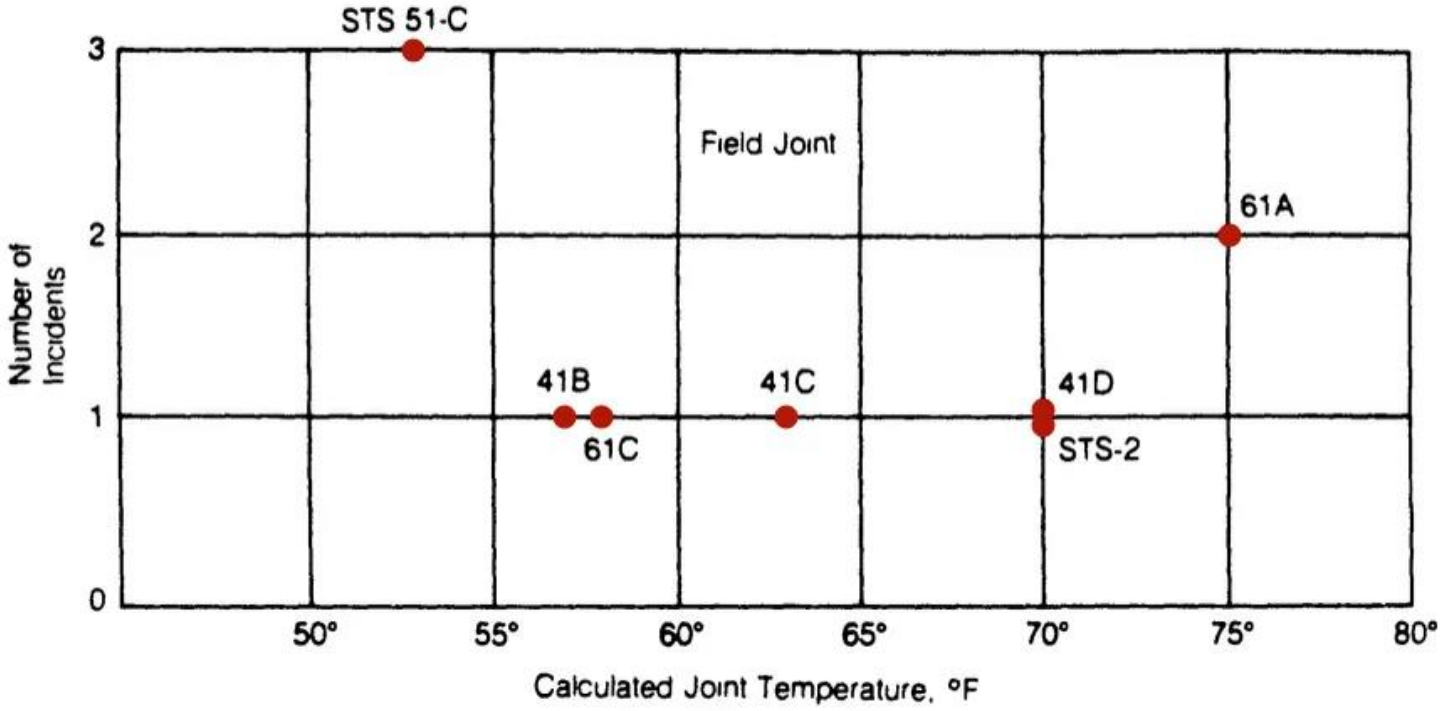
Project: Scraping news articles from news websites in 23 countries. Will be used for large-scale analysis of climate coverage in the new media using genAI.

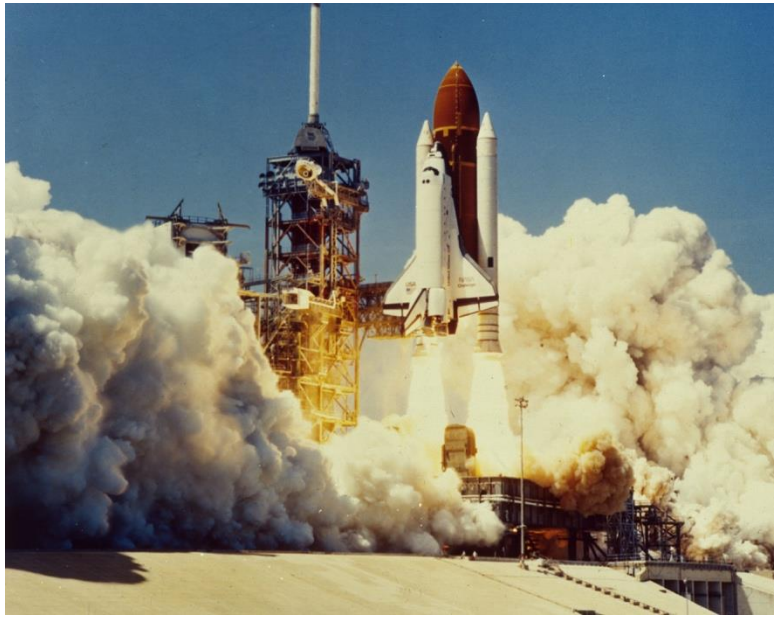
Employment:

- Student assistant, 0.2fte
- 6 Months, perhaps longer
- Start ~Nov 2026
- Supervised by CS expert

Challenger Space Shuttle, 1986

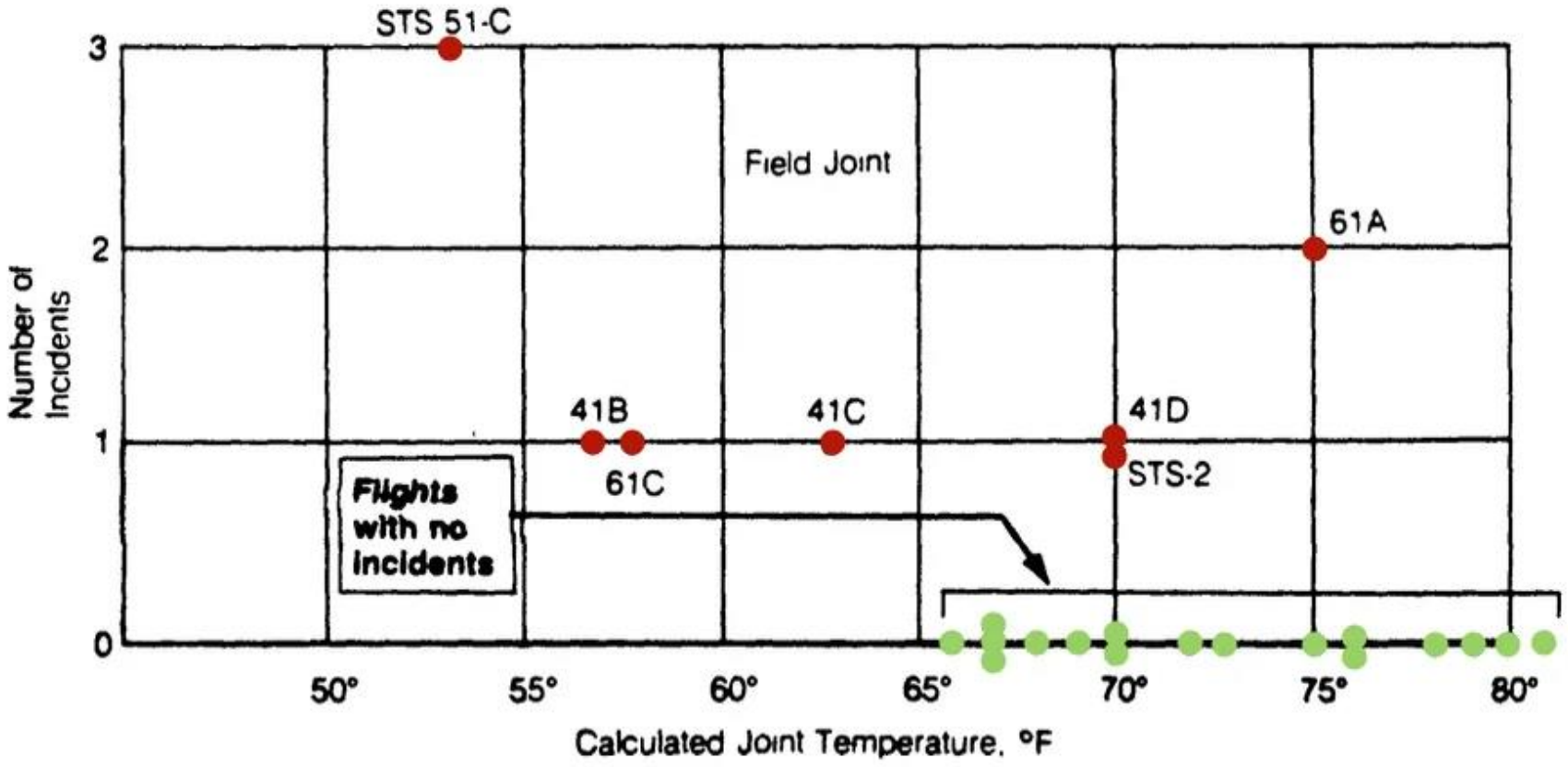
On 28 January 1986, Challenger was expected to launch into the coldest weather any Shuttle had flown in, 31F (-1C). This graph played a key role to understand if there was a relationship between temperature and failure in one component (O-ring).







The full data



The same chart with the 16 flights that had no incidents added back. Every launch below 65 °F showed O-ring distress, and the temperature the engineers were asked about was 31 °F.

Exploratory data analysis helps us understand the data and refine our questions.

Visualization makes patterns easier to see and findings easier to communicate.

Learning goals of today

Part 1: Exploratory Data Analysis
Use EDA to understand data and refine questions.

Part 2: Why we plot
Explain why we visualize data.

Part 3: Principles of visualization
Choose and justify plots using the grammar of graphics and visual channels.

Part 4: Making the argument (time allowing)
Communicate findings through design and storytelling

PART 01

Exploratory data analysis

Use EDA to understand data and refine questions.

John Tukey, 1915–2000



John Tukey at Bell Laboratories

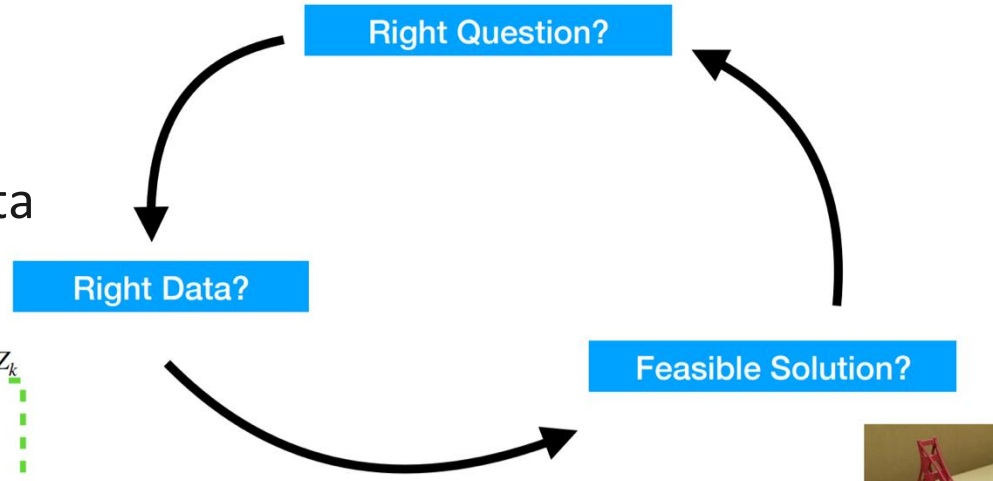
Huge influence in data science. Published Exploratory Data Analysis in 1977.

He developed the boxplot, coined the term *bit*, co-developed the fast Fourier transform, and so much more.

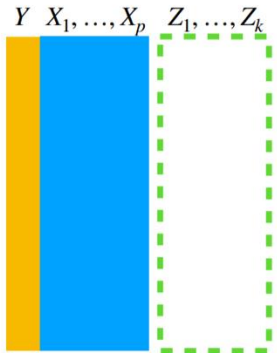
"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." (The future of data analysis, 1962)

The EDA cycle: ask, look, refine

The hardest job: what do we want to learn? 🤔



Can these data answer it?





Simple analyses to understand if getting an answer is possible

Source: R.D. Peng, https://rdpeng.org/ads2020/week-7.html#a-framework-for-exploratory-analysis

A systematic first pass

Right question

1. Write the question and sketch the expected pattern(s).

Right data

2. Read the codebook: define the rows, variables and unit of analysis
3. Count: rows, groups, duplicates and missing values.
4. Plot variation: does it look as expected? Zoom in rare or impossible values.
5. Check aggregates against an independent source.
6. Plot covariation between variables.

Feasible solution

7. Try a simple analysis that could answer the question.
8. Try to break the result with other scales, periods or outcomes.

Adapted from https://rdpeng.org/ads2020/week-7.html

Worked example: did ending compulsory school swimming increase child drownings?



A school swimming lesson in the Netherlands, mid twentieth century.

1950s
Drowning mortality was high, especially among children.

1960s and 1970s
School swimming expanded as municipalities built indoor pools and small teaching pools. Compulsory school swimming starts

Around 1985
The national obligation ended, while local programs and private lessons continued.

Data.

CBS (Statistics Netherlands): deaths, 1950 to 2024, by year, age group and cause.

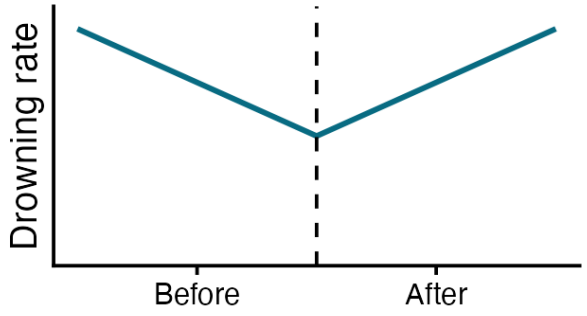
1. Write the question and sketch the expected pattern.

What type of claim? Association, prediction, causation?
What comparison? Across groups, time, conditions?
What evidence would support or weaken it?

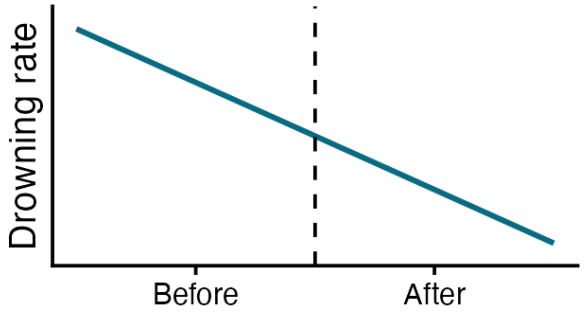
Did ending compulsory school swimming increase child drownings?

Causal claim; EDA examines associations.
Compare time: before and after.
Rates rising or declining more slowly afterward supports the hypothesis.

Supportive evidence

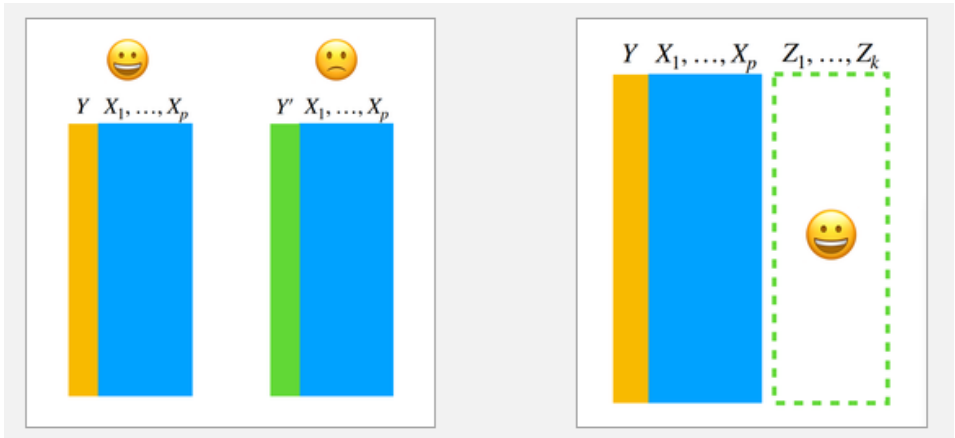


No evidence



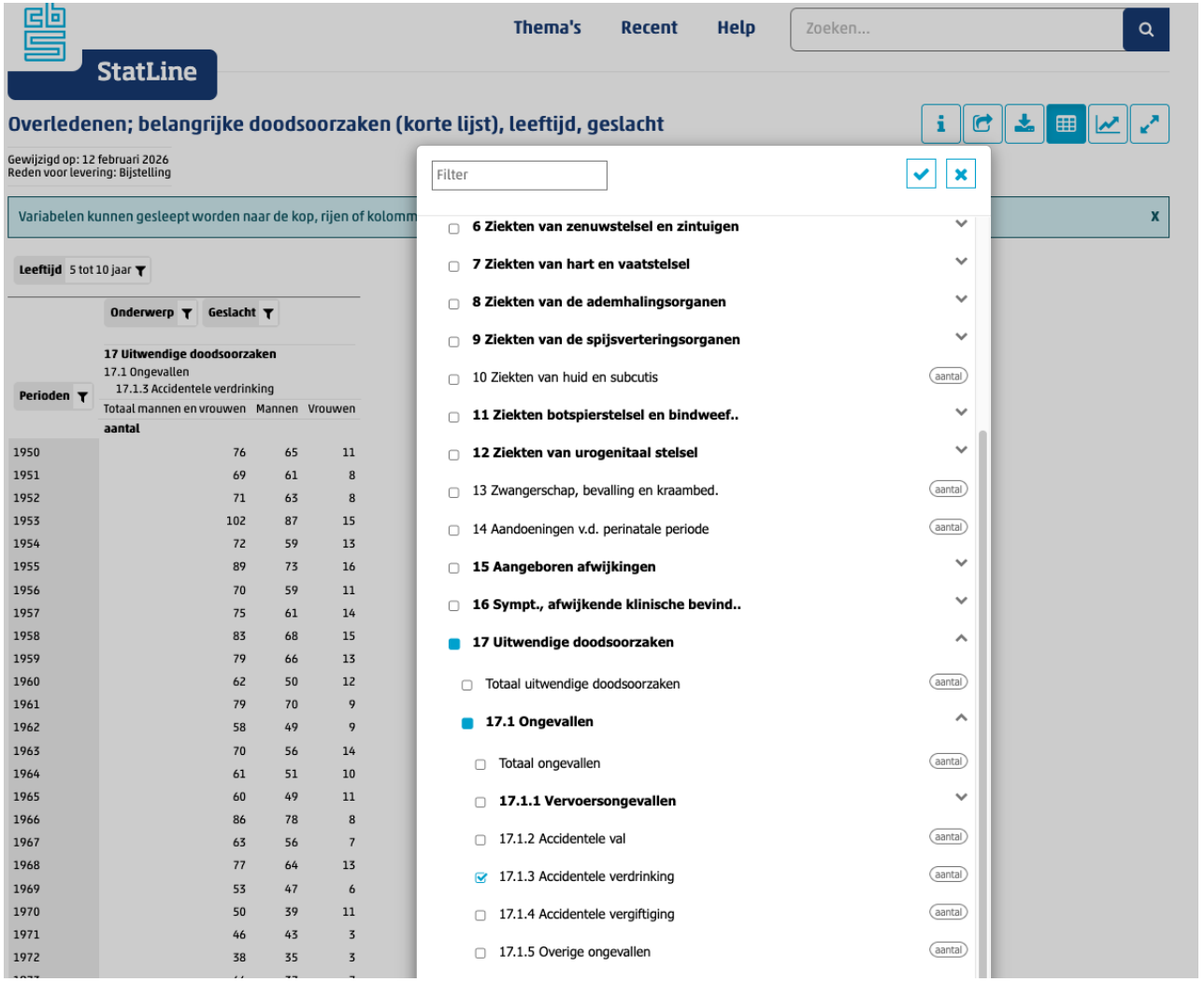
The right data are more than a clean file

What is the unit of analysis for your data? (person, person-year, etc). Does it match what your question requires?



Right outcome? Needed predictors? Missing data? Why? Unusual data? Why?

2. Read the codebook: define the rows, variables and units.



What does one row represent, and what exactly counts as a drowning death?

Accidental drowning among residents of the Netherlands (ICD-10: W65–W74)

Aggregated by age at death and gender

W65	Drowning and submersion while in bathtub
W66	Drowning and submersion following fall into bathtub
W67	Drowning and submersion while in swimming pool
W68	Drowning and submersion following fall into swimming pool
W69	Drowning and submersion while in natural water
W70	Drowning and submersion following fall into natural water
W73	Other specified drowning and submersion
W74	Unspecified drowning and submersion

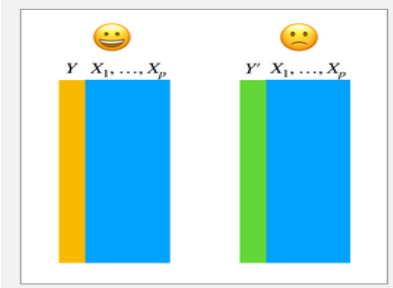
2. Read the codebook: define the rows, variables and units.

What's in your data?

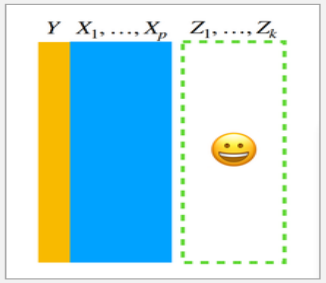
```
glimpse(child_rates)
Rows: 75
Columns: 6
$ year       <int> 1950, 1951, 1952, 1953, 1954,…
$ age_code   <chr> "70200", "70200", "70200", "7…
$ age_group  <chr> "5-9", "5-9", "5-9", "5-9", "…
$ deaths     <int> 76, 69, 71, 102, 72, 89, 70, …
$ population <dbl> 931820, 989357, 1072359, 1127…
$ rate       <dbl> 8.156082, 6.974227, 6.620917,…
```

One row: one year and age group.

year: year
age_code: CBS age-category code
age_group: age label
deaths: annual accidental drownings
population: the same age group
rate: deaths / population × 100,000



Right outcome?



Needed predictors?

3. Count rows, groups, gaps, duplicates and missing values.

Check your n()

```
drowning |> summarise(rows = n())  # 1575

drowning |> summarise(
  across(everything(), ~sum(is.na(.)))
)
# year age_code age_group deaths population rate
#    0        0         0      0          0    0

drowning |> count(age_group)
# n = 75 for all 21 age groups

drowning |> count(year)
# n = 21 for all 75 years

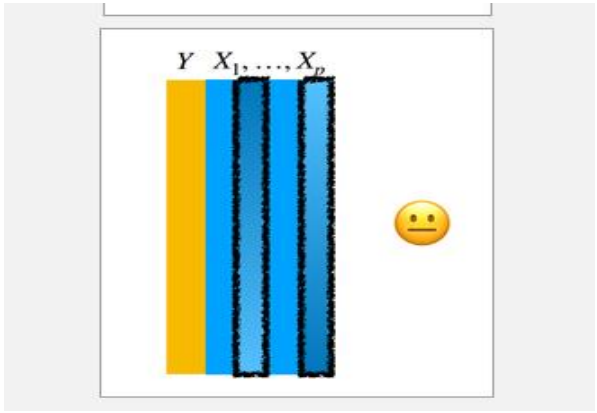
drowning |> count(year, age_group) |> filter(n > 1)
# 0 rows: no duplicate year-age combinations
```



Missing data?

CBS data is very clean; this is not usually the case.

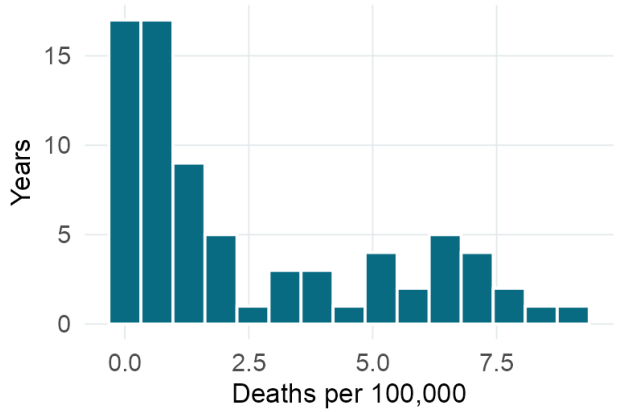
4. Plot variation within each variable; inspect rare or impossible values.



Unusual data? Why?

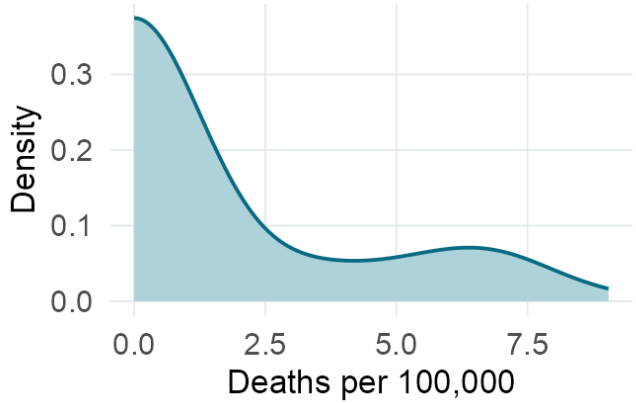
Histogram

`geom_histogram()`



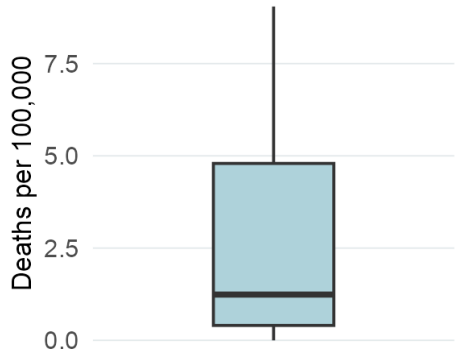
Density

`geom_density()`



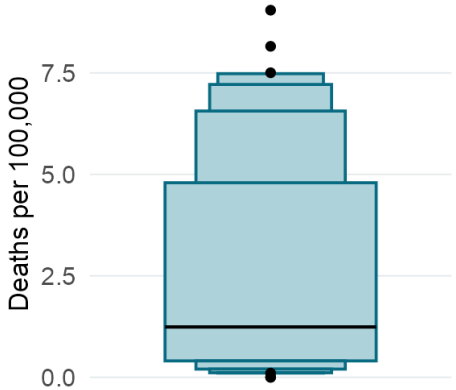
Boxplot

`geom_boxplot()`



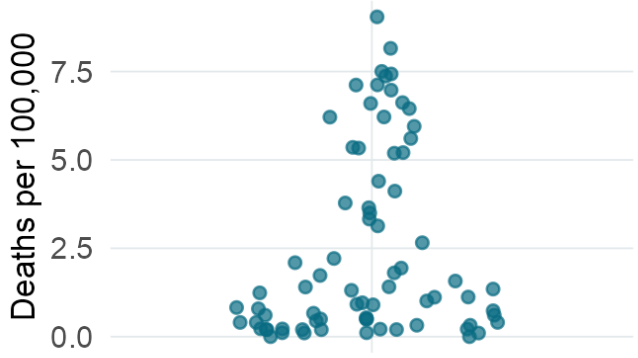
Boxen plot

`geom_rect()`



Sina plot

`ggforce::geom_sina()`



5. Check one value against an independent source.

Use a similar same year, age group and definition.
If no independent source is available, check a related quantity.

1950

All deaths: 603 (64.71 per 100,000)
Accidental deaths: 230 (24.68 per 100,000)
Of which drowning: 76 (8.16 per 100,000)

2024

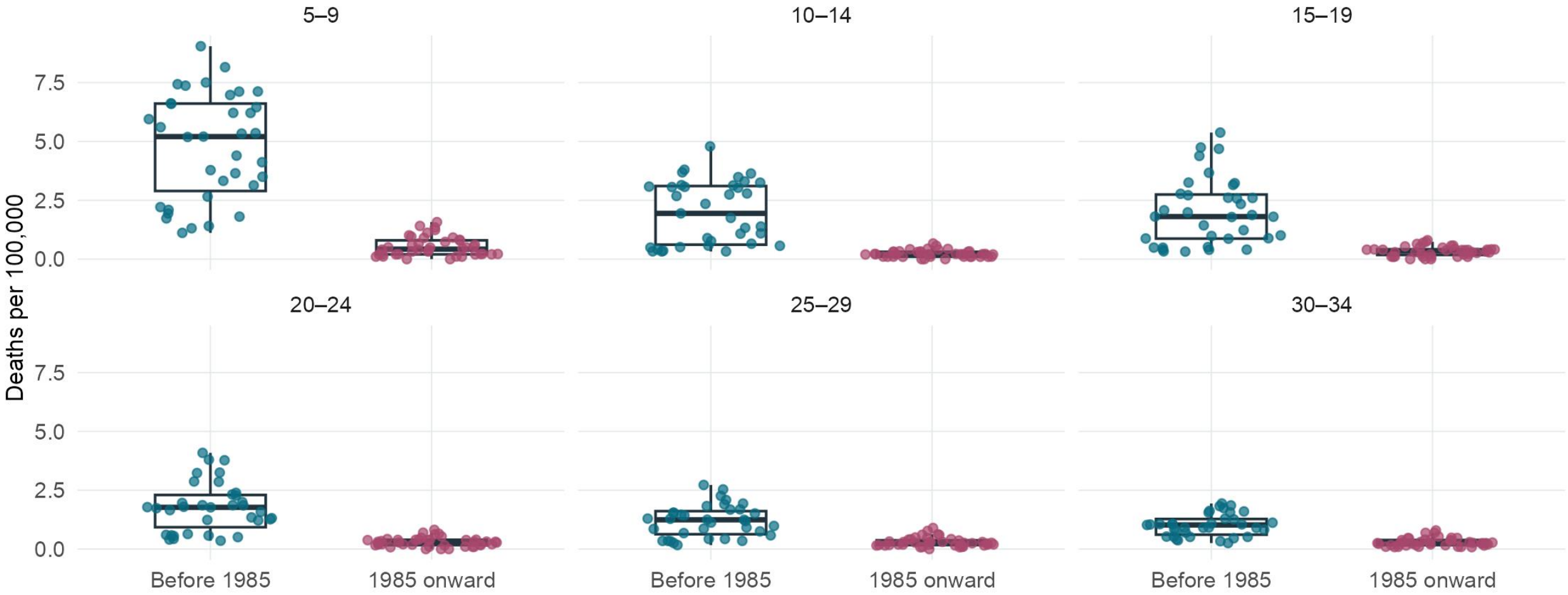
All deaths: 68 (7.59 per 100,000)
Accidental deaths: 9 (1.00 per 100,000)
Of which drowning: 2 (0.22 per 100,000)

Do these rates seem plausible for children aged 5–9?

CBS StatLine 7052_95 | Both sexes, ages 5–9 | Drownings are included in accidents

Right data

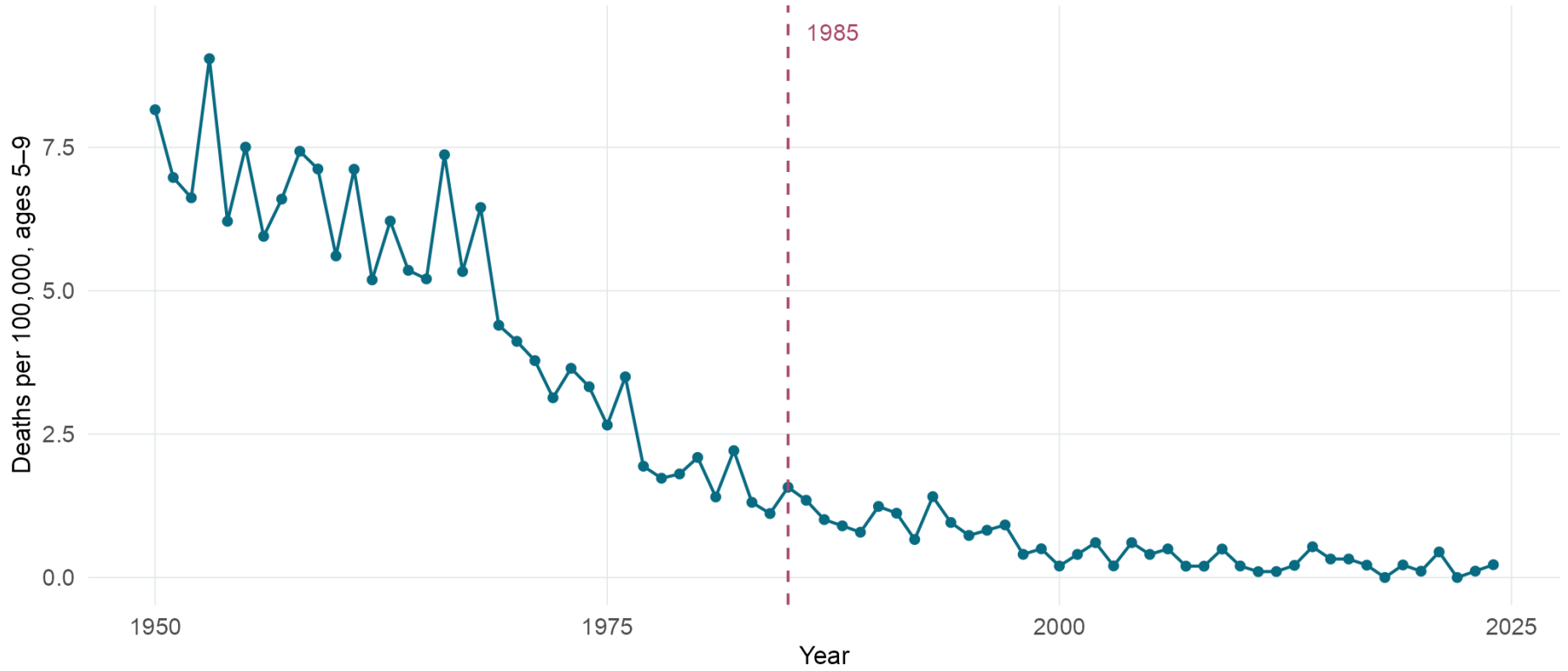
6. Plot covariation between variables and inspect relevant subgroups.



CBS StatLine 7052_95 | Netherlands, 1950–2024

Feasible solution

7. Try the simplest analysis that could answer the question.



There is not a visible increase in deaths after 1985. It seems to be decaying much earlier.

CBS StatLine 7052_95 | Netherlands, 1950–2024

8. Try to break the result with other scales, periods, groups or outcomes.

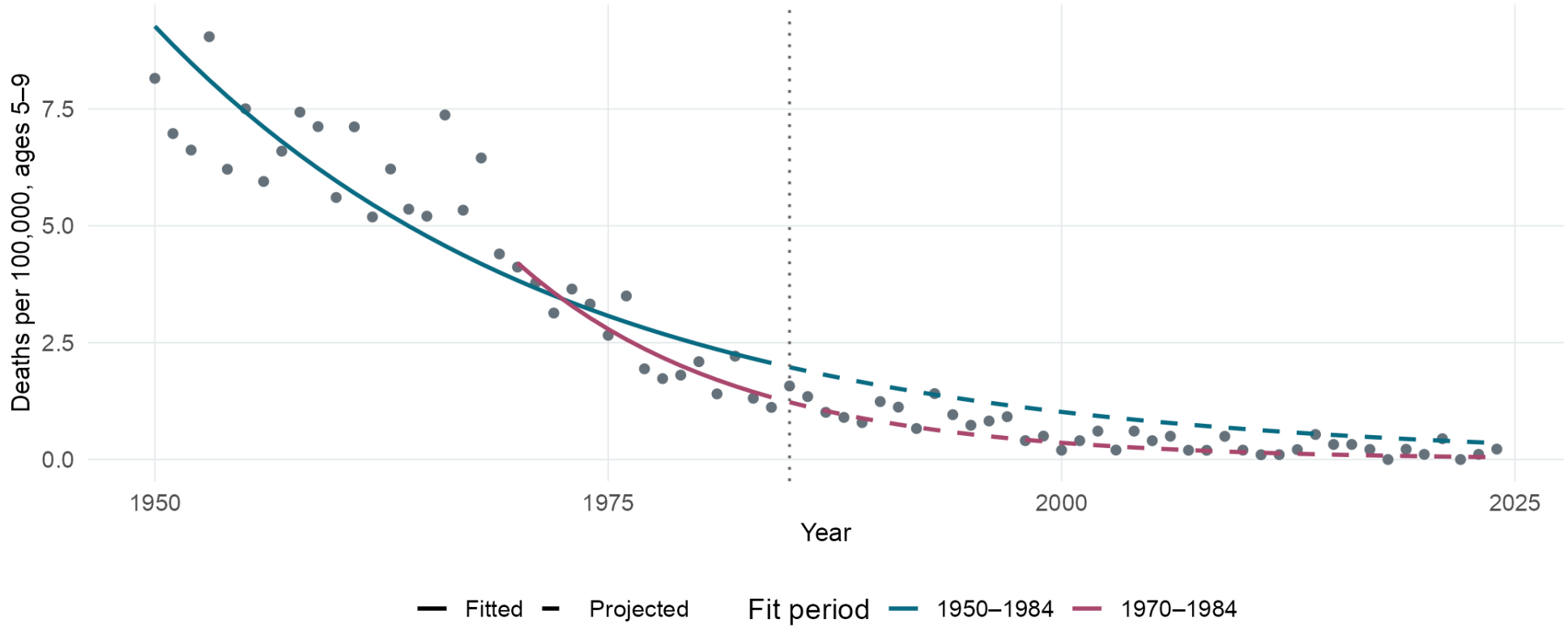
Does the pattern change with another scale or time period?

Does the pattern hold only for children, or also for other age groups?

Does the pattern hold for other accidental deaths?

A temporal baseline

What would we expect after 1985 if the earlier trend continued?

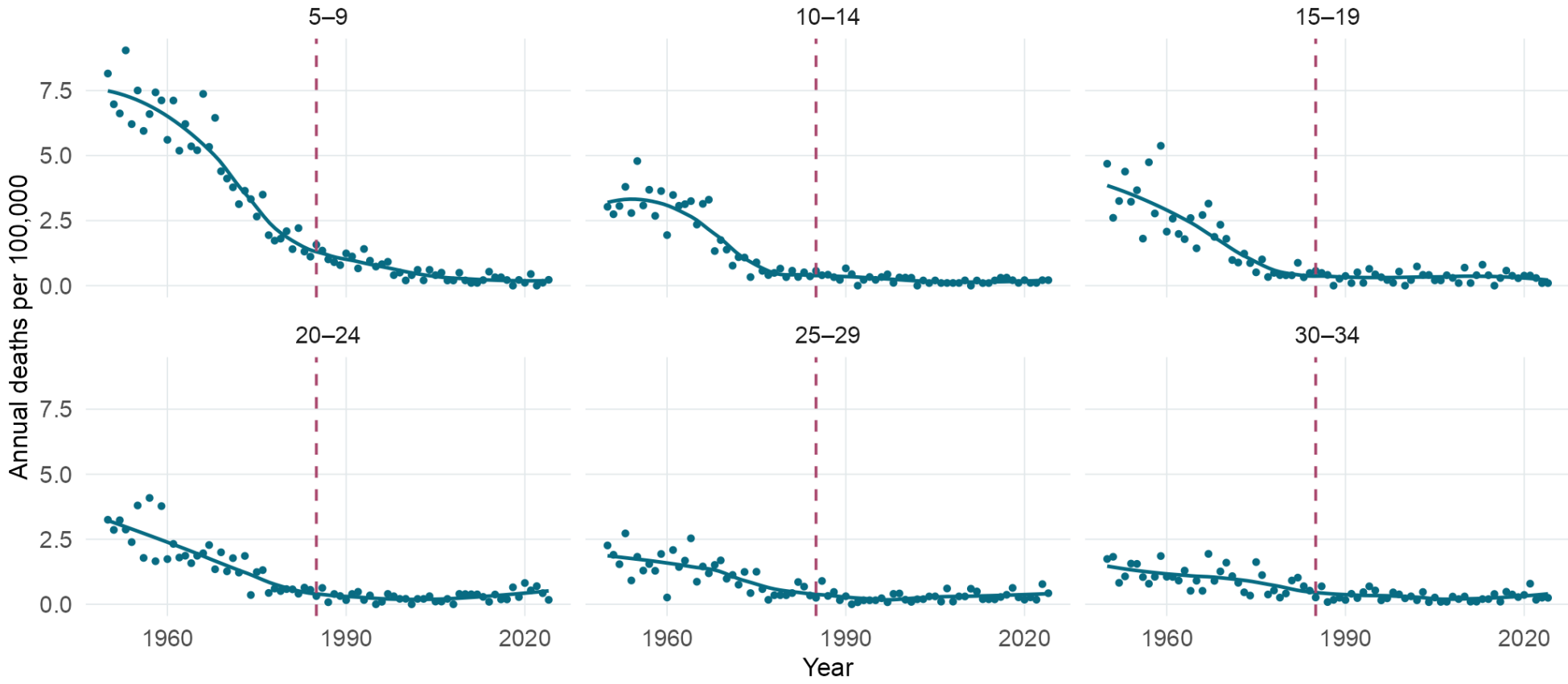


.

CBS StatLine 7052_95 | Netherlands, 1950–2024

An age-group baseline

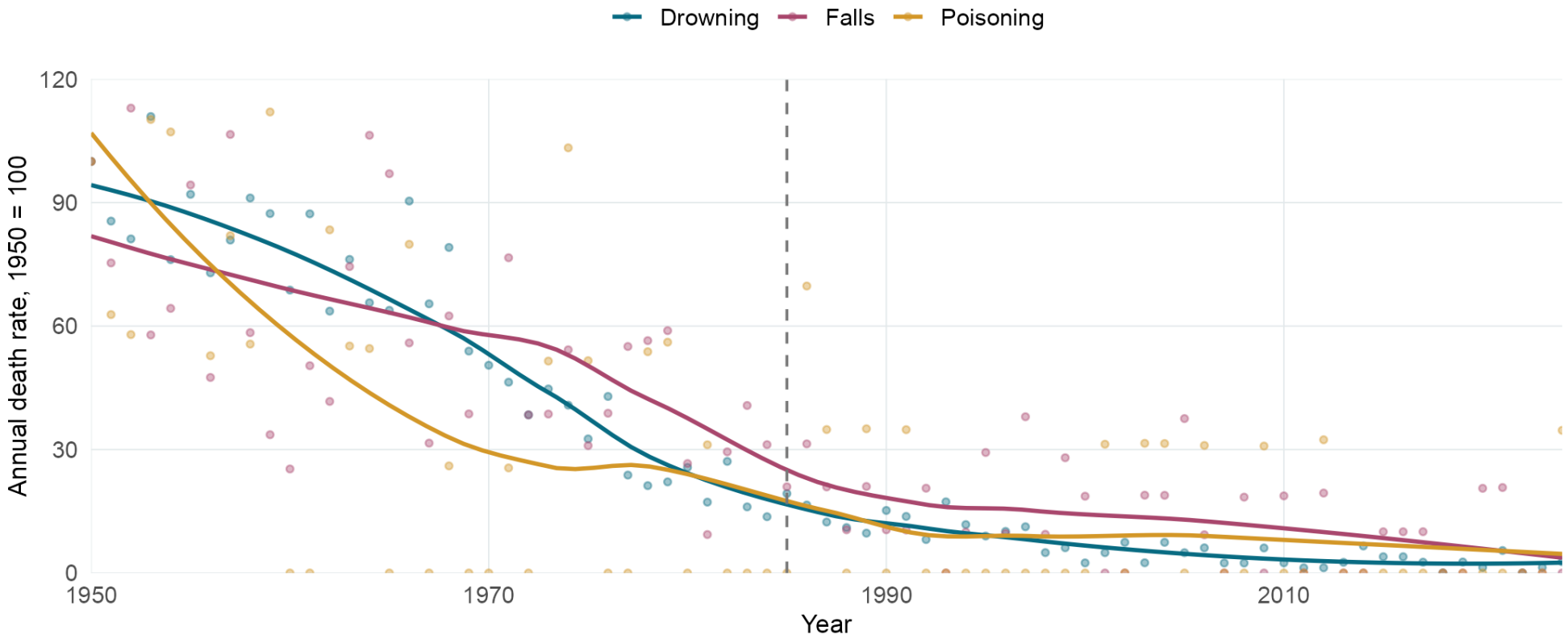
Could other age groups (not so affected by school swimming lessons) provide a useful comparison for school-age children?



CBS StatLine 7052_95 | Netherlands, 1950–2024

Other accidental deaths as a baseline

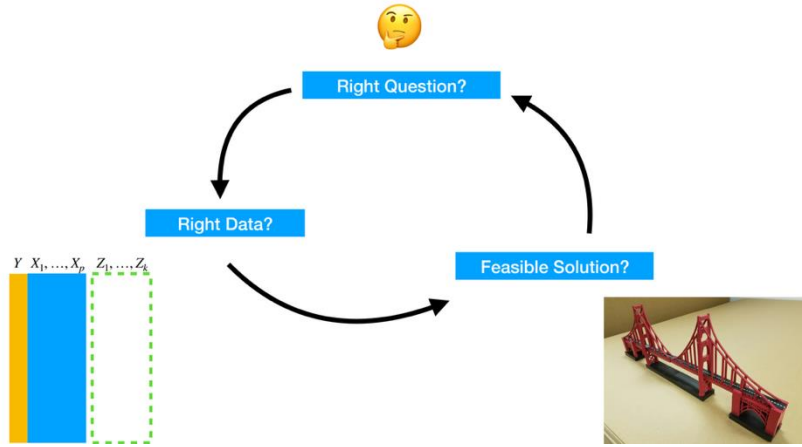
Did falls and poisoning decline too?



Each cause = 100 in 1950. Compare relative changes, not mortality levels.

CBS StatLine 7052_95 | View limited to 120; smooths use all observations.

EDA reshapes the question



Starting question

Did ending compulsory school swimming increase child drownings?

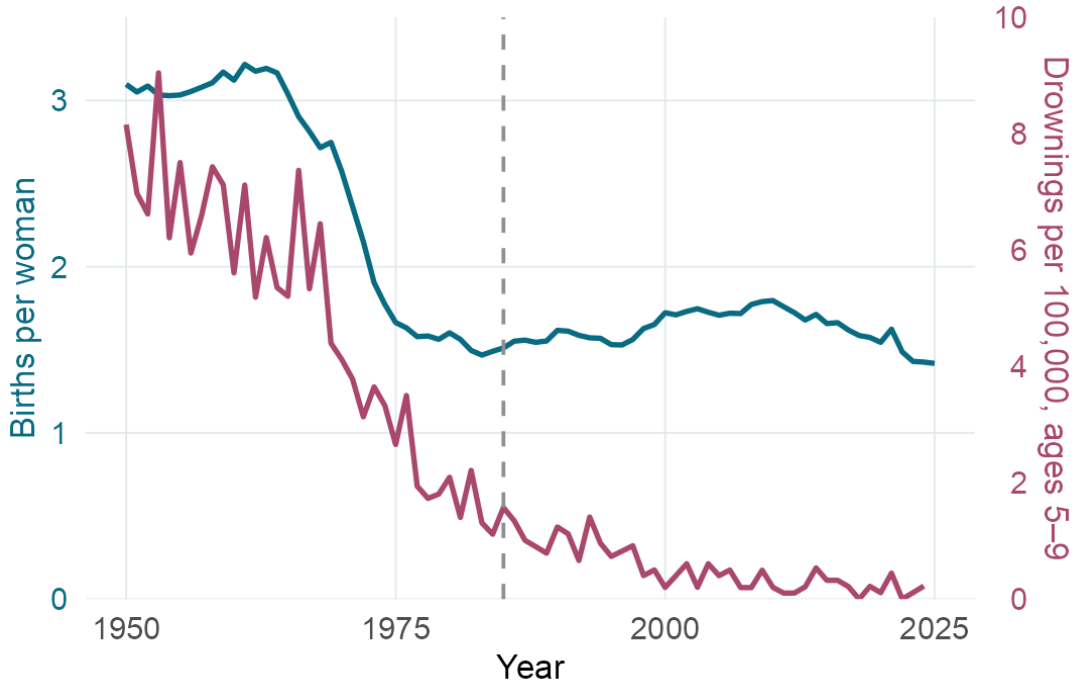
Doesn't seem to have a large impact → rates didn't change or barely, and for all age groups similarly
These comparisons help assess possible explanations, but do not establish the policy's effect

After EDA

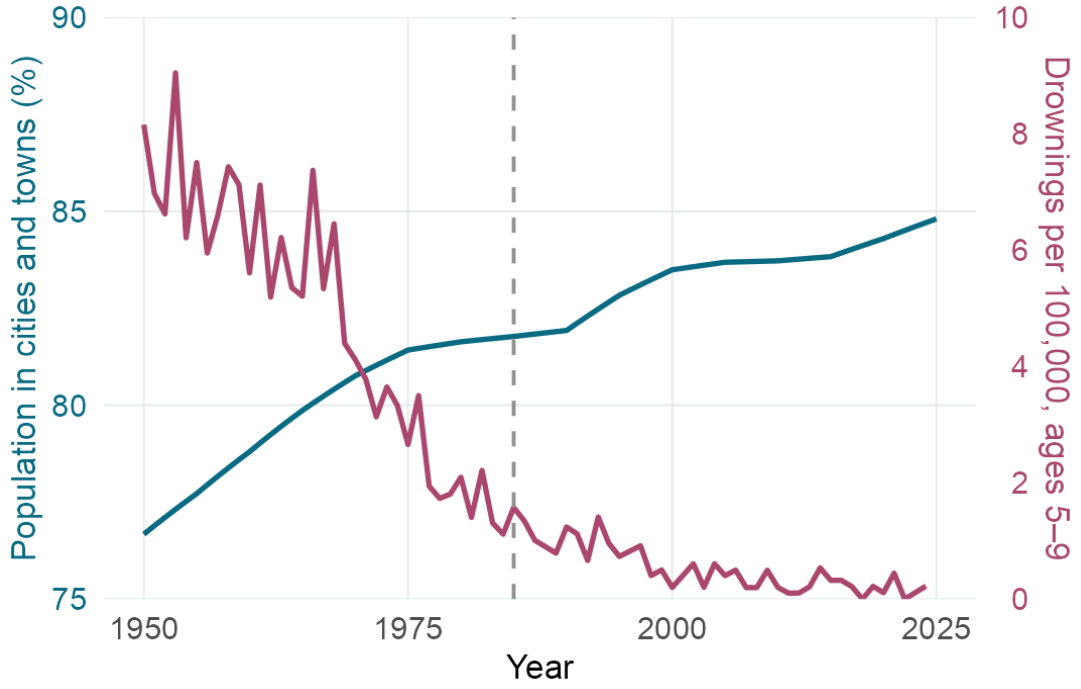
All accidental deaths decreased, are there other factors (e.g. cultural or demographic) impacting risky behavior in the Netherlands?

Other changes in the Netherlands

Fertility



Urbanization

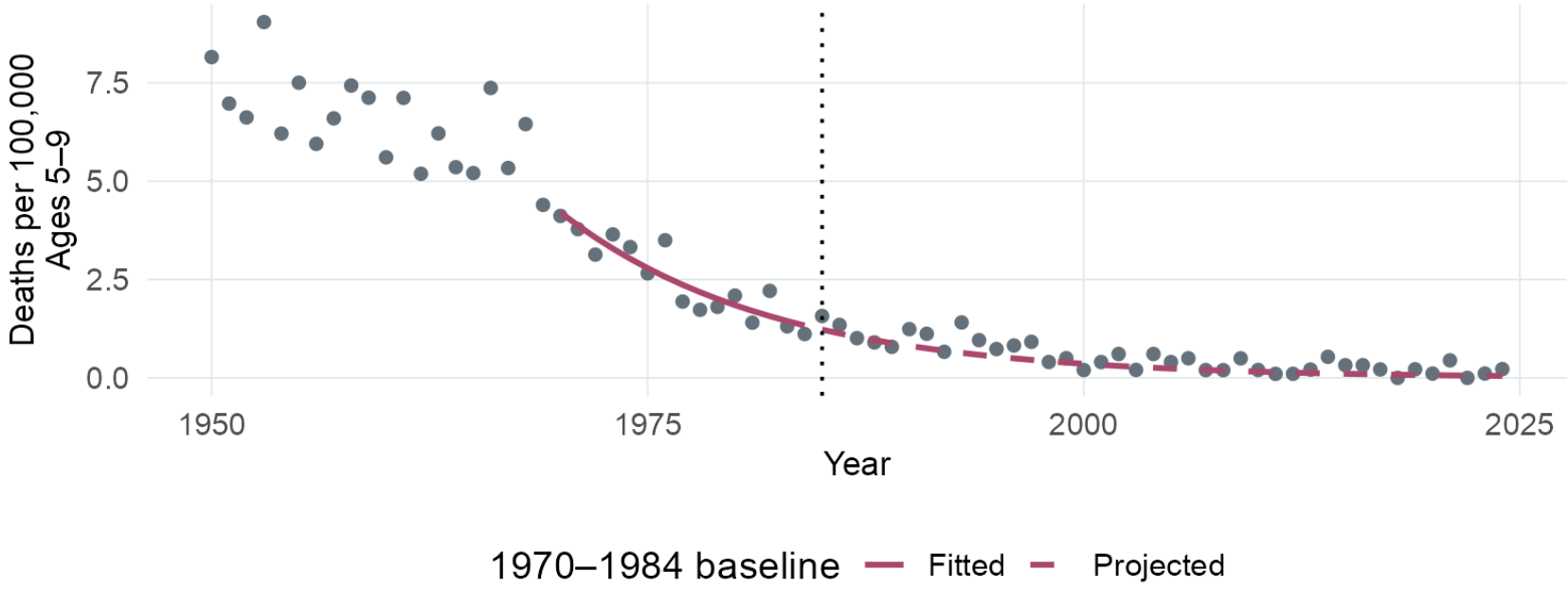


Could changes in family life or living environments help explain the decline?

CBS 7052_95 and 85722NED; UN WUP 2025 | Pink: drowning rates, right axis | Dashed line: 1985

Warning: EDA can produce chance findings

Try enough age groups, periods and outcomes, and something may look convincing by chance. You, as a model learning from the data, may be overfitting. This is also called p-hacking if you do statistical analyses.



Explore openly, keep track of what you tried, try to break your finding don’t stop at the first result, and check new findings with new data (you can set a subset aside for this check)

What we learned from the drowning data

Right data?
→ We have the data needed to examine whether child drowning rates increased.

Right question?
→ Partially, but interesting questions came up from the EDA

Approximate answer:
→ Deaths from the accident categories we examined had already declined before the 1980s. There is no clear change around the end of the requirement. Other factors may help explain these trends. Fertility rates seems like an interesting direction.

Part 1: what you should be able to do now

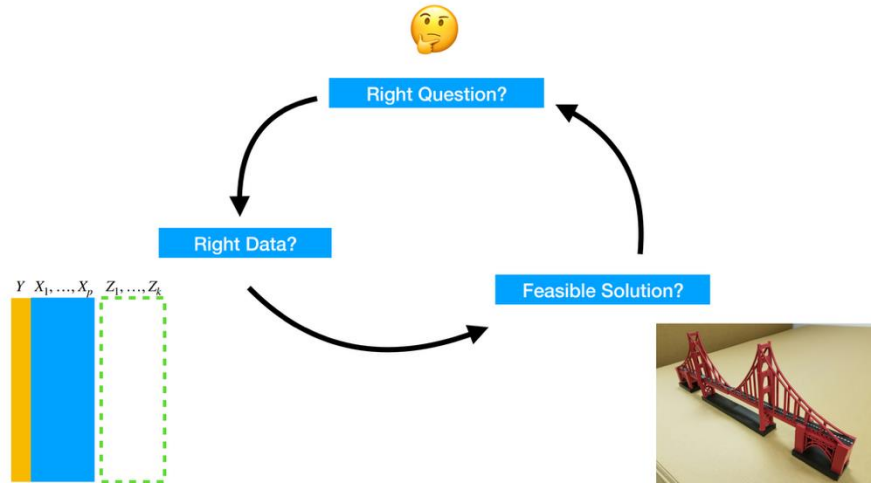
Use EDA to understand data and refine questions.

State a question and the expected pattern.

Check what the data measure and identify possible problems.

Examine variation and covariation.

Challenge the result and refine the question.



We just used many plots to explore one question.

Why not read the numbers?

PART 02

Why we plot

Explain why we visualize data.

A plot carries data to the brain through a picture

DATA



Values usually in rows and columns.

Encode

PICTURE



Map variables to position, colour or shape.

Decode

BRAIN



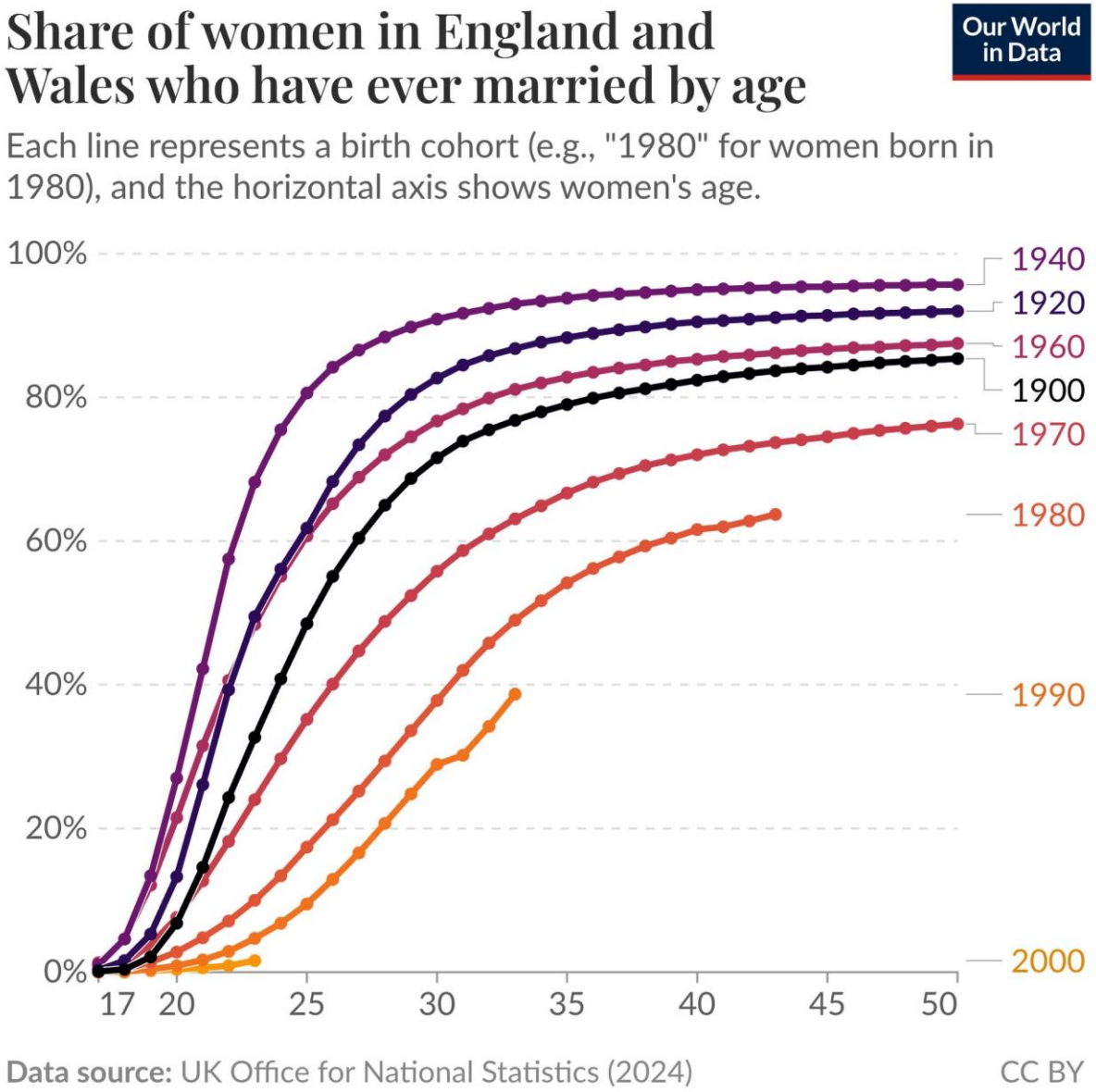
Interpret the patterns.

Done correctly: *Extremely efficient way to help people understand and remember information*

Our brains process visual information much faster than text or numbers:

- The human retina can transfer around 10^6 or 10^7 bits per second to the brain;
- Reading transfers about 3 words, so ~10^2 or 10^3 bits/s;

Marriage has gone from near-universal rite of passage to an increasingly optional milestone for women in England and Wales. A new chart from the UK's Office for National Statistics traces eight birth cohorts—women born in 1900, 1920, 1940, 1960, 1970, 1980, 1990 and 2000—and shows how the share who have "ever married" rises with age and then levels off. Women born at the turn of the twentieth century were almost certain to tie the knot in their twenties: by age 24, more than 80 percent of the 1900 cohort were married, and by age 30 that figure climbed into the mid-90s. The pattern holds—and even accelerates—for the 1920 and 1940 cohorts: purple-hued lines soaring above 90 percent before most reach 28. But from the 1960 cohort onward, marriage curves began to flatten and shift rightward. Women born in 1960 didn't quite hit 80 percent until their late twenties, and those born in 1970 paused near 75 percent even by age 30. By contrast, among women born in 1980, barely two-thirds had ever married by their mid-thirties. The most dramatic change appears among the youngest groups. For those born in 1990, the climb to 40 percent married stretches into their mid-thirties—and the 2000 cohort, now barely in their early twenties, has only just begun to steady above zero. Today's young women are marrying later, if at all, reshaping long-standing social norms around partnerships, family formation and adulthood.



Know your audience and goal

Exploration

Helps **you** understand the data and avoid mistakes and refine the question.

EDA: understand data and refine questions by visualizing variation (distributions) and covariation (e.g., comparisons, relationships).

Explanation

Helps **others** understand your findings.

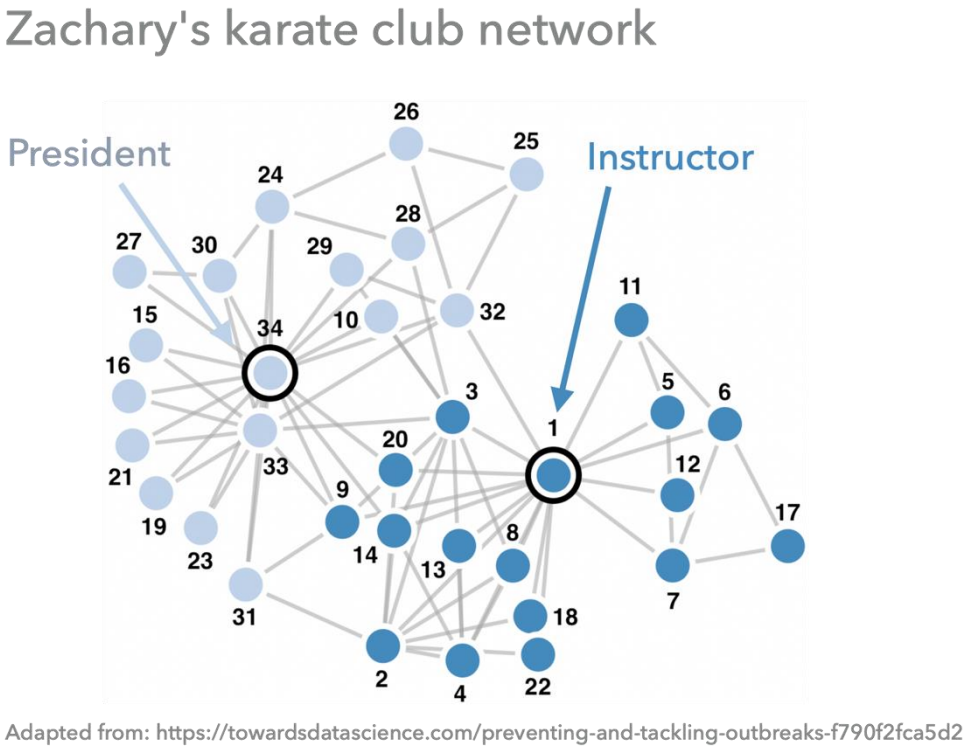
Communicate key insights clearly and effectively.

Example: Explanatory data visualization

Every explanatory plot should have one (and ideally only one) message, which should be written in advance.

Individual Number

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
2	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0
3	1	1	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0
4	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	
10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
11	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
14	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
17	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
22	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1
29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	1
31	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
32	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1
33	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	1	1	0	0	1
34	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0



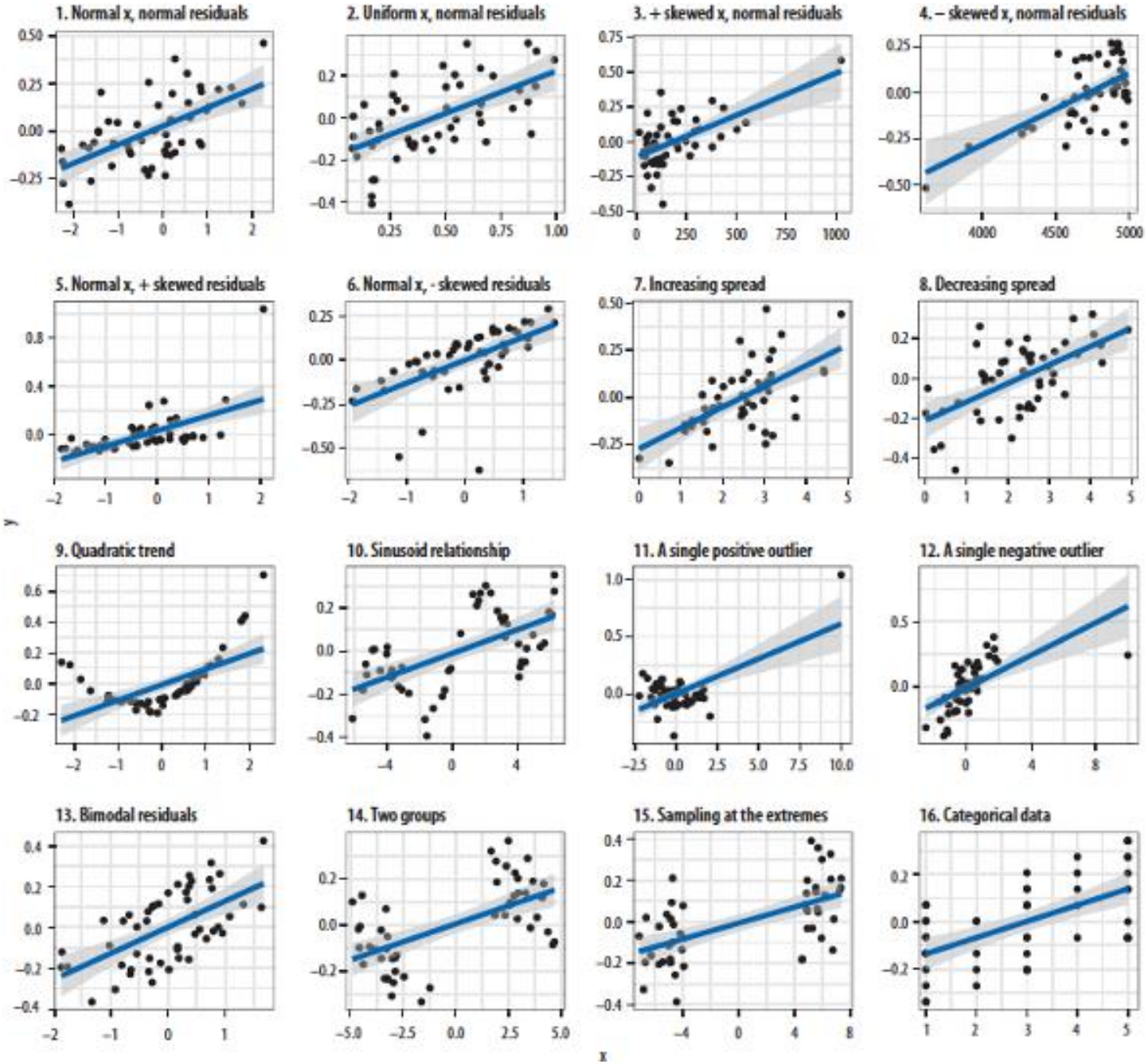
Adapted from: https://towardsdatascience.com/preventing-and-tackling-outbreaks-f790f2fca5d2

Two communities centered around the instructor and president.

Example: Exploratory data visualization

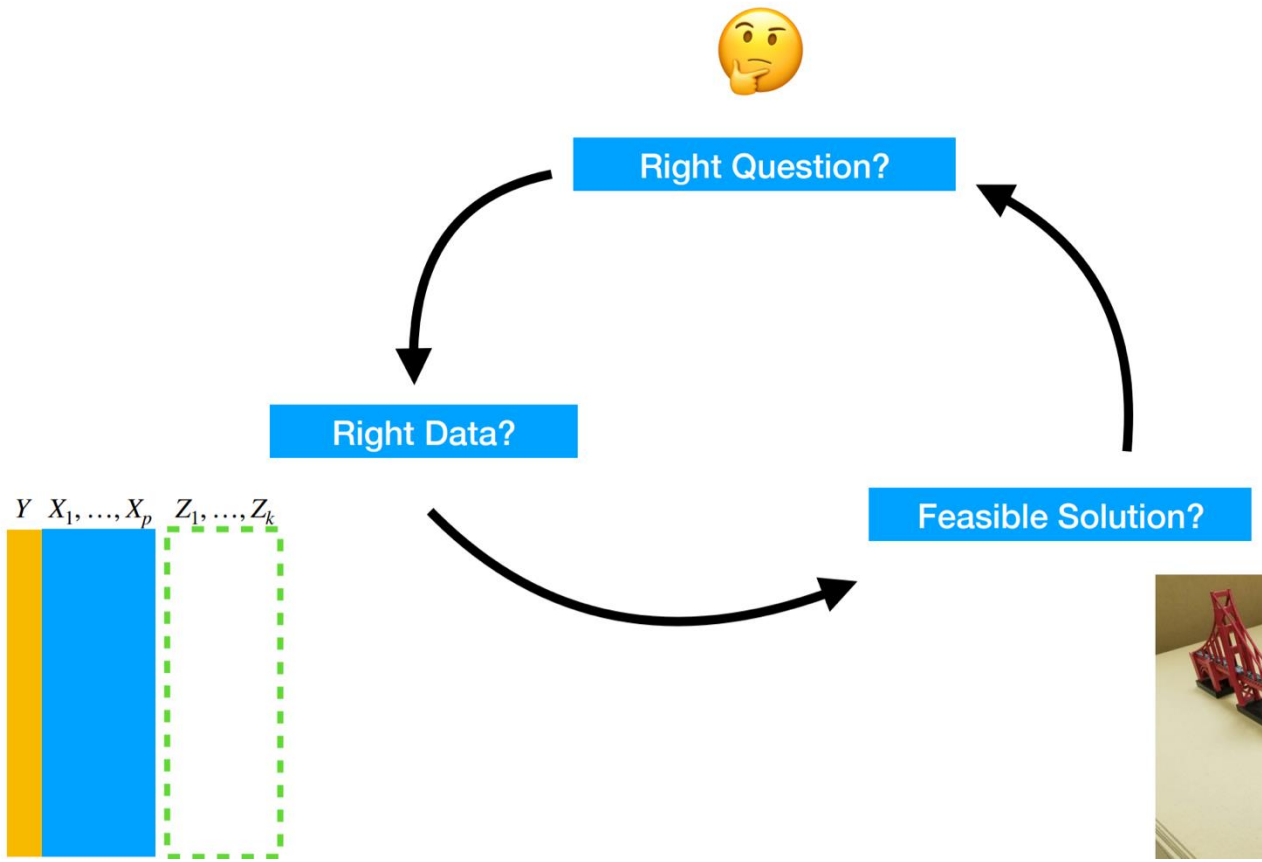
All 16 panels have a correlation of 0.6.

The plots reveal different distributions, trends and unusual observations.



Source: Healy (2019)

Visualization in the EDA cycle





Think about what you expect to see.

Plot the data and compare with that expectation.

Use the difference to validate the data or refine the question.

Part 2: what you should be able to do now

Explain why we visualize data.

Visualization is an extremely efficient way to interpret data.

Understand the difference between exploratory and explanatory visualization



Break

Back at …

The goal of visualization is to specify how map data to picture, so the correct type and largest amount possible is transferred.

But how can we leverage the human visual system to analyze data?

PART 03

Principles of visualization

Choose and justify plots using the grammar of graphics and visual channels.

The grammar of graphics

The *grammar of graphics* breaks a graph into parts.
In R, ggplot2 combines data, aesthetic mappings and geometric objects.

1. Data

```
ggplot(data = drownings)
```

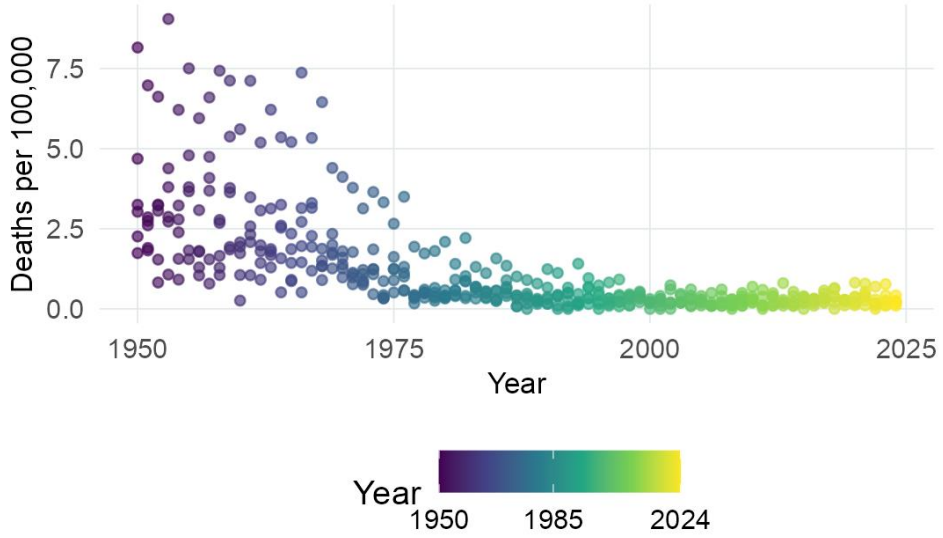
year	age	deaths	rate
1950	5-9	76	8.16
1951	5-9	69	6.97
1952	5-9	71	6.62

2. Aesthetic mappings

```
ggplot(data = drownings,
       aes(x = year,
           y = rate,
           color = year)
)
```

3. Geometric objects

```
+ geom_point()
```

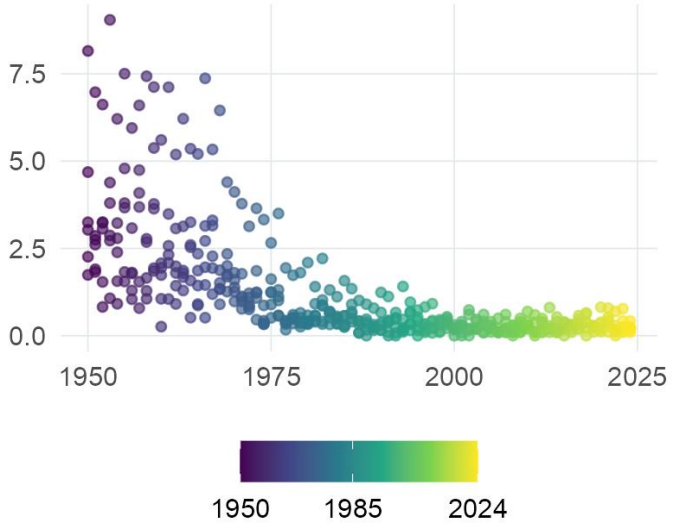


Note: this is illustrative, don't map the same variable to two channels!

Scales, facets and transformations

Without labels

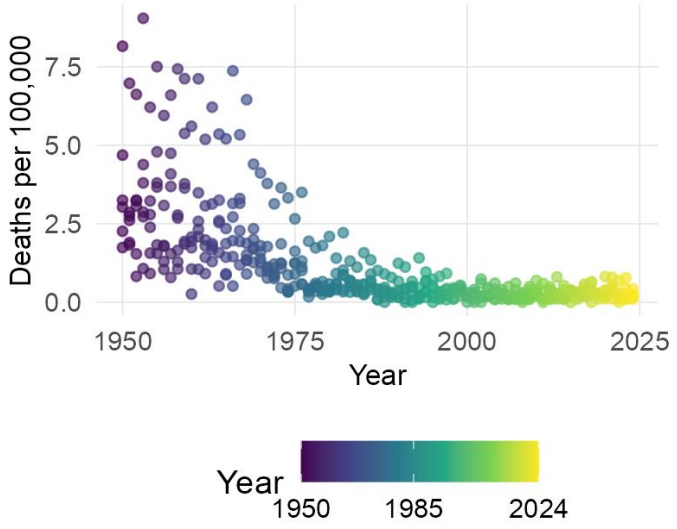
The plot from the previous slide.



Add labels

Name the variables and units.

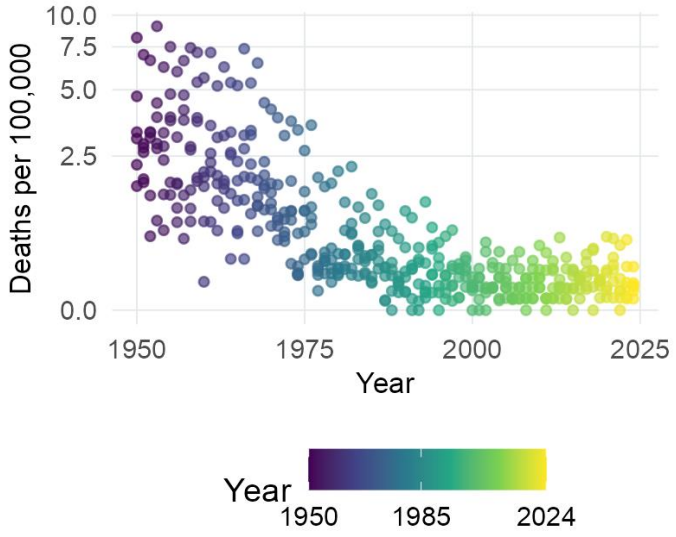
```
+ labs(x = "Year",
 y = "Deaths per 100,000",
 colour = "Year")
```



Change the scale

log1p retains zero-rate years.

```
+ scale_y_continuous(
   transform = "log1p"
)
```



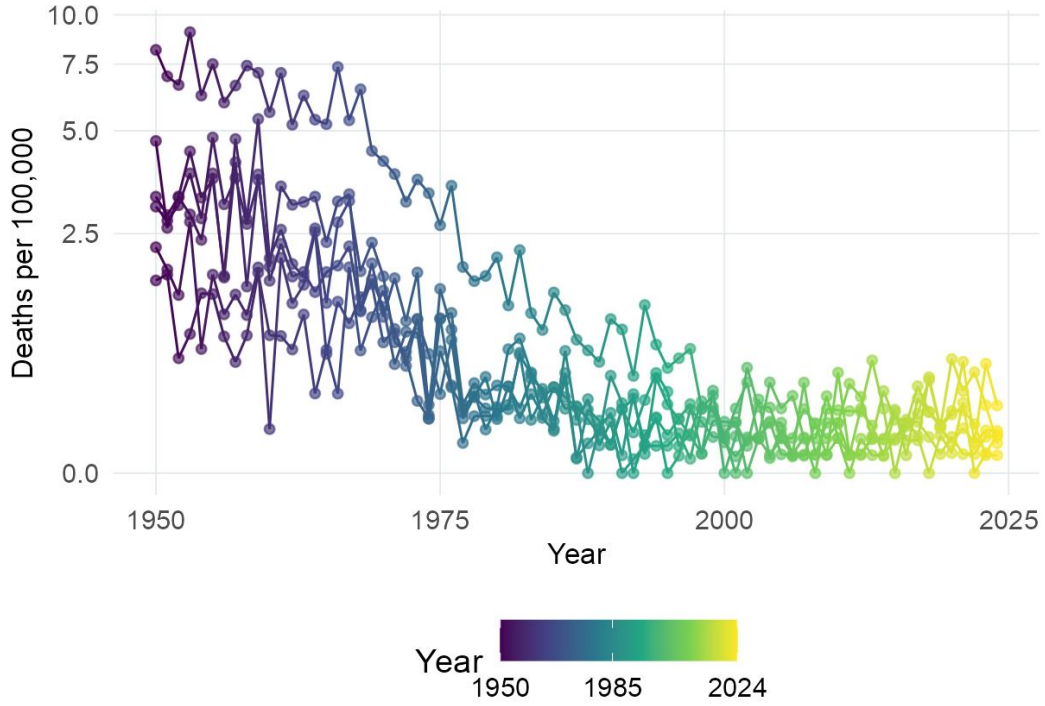
CBS StatLine 7052_95. Six age groups, 1950–2024. Each panel adds one change.

Lines and facets

Add lines

Connect years within each age group.

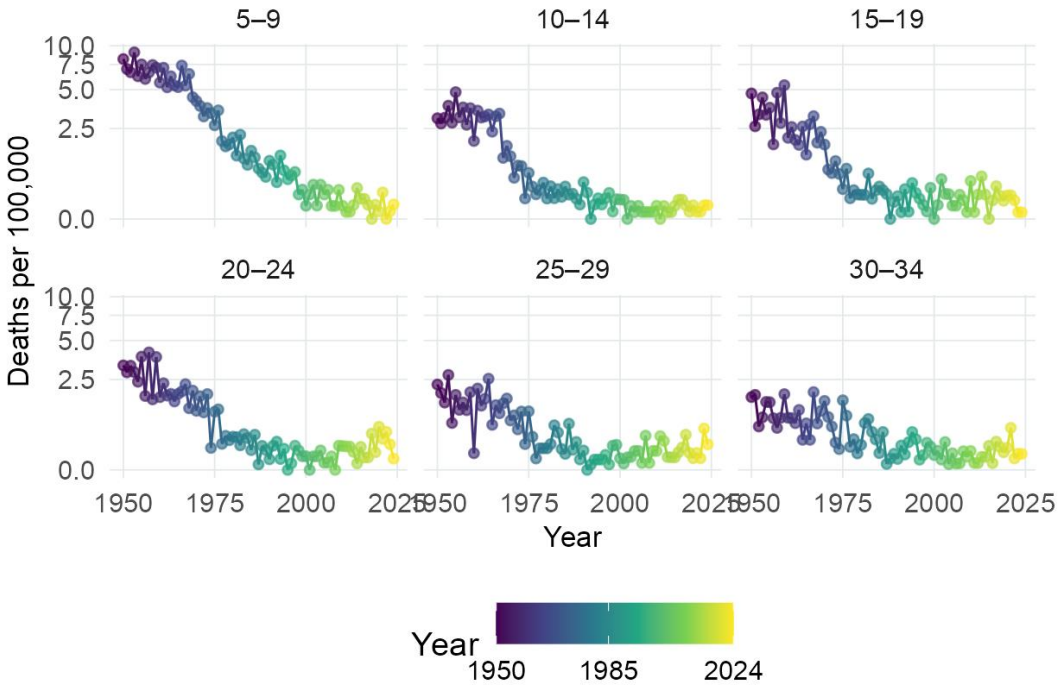
```
+ geom_line(aes(group = age_group))
```



Add facets

Separate the age groups into panels.

```
+ facet_wrap(~ age_group, ncol = 3)
```



Scales, facets and transformations

Statistical transformations

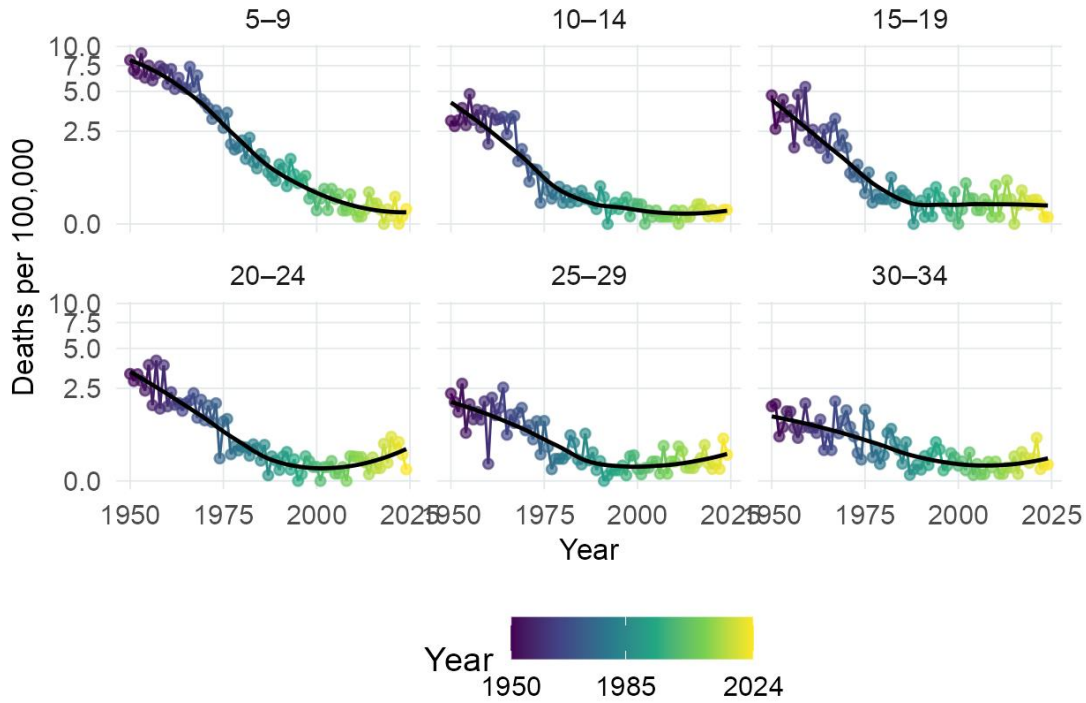
Count, bin or smooth observations.

```
geom_smooth(method = "loess",
            se = FALSE,
            colour = "black")
```

Position adjustments

Move marks to reduce overlap.

```
position_jitter()
```

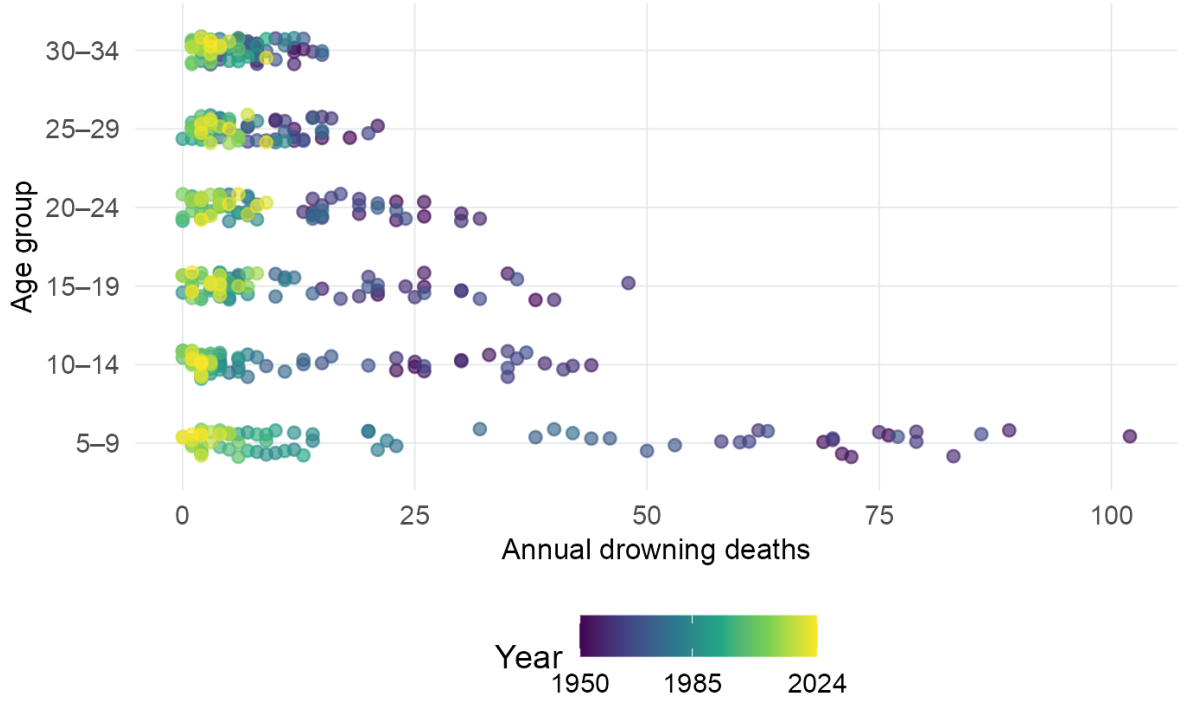


Add a smooth trend to each age group.

CBS StatLine 7052_95. Six age groups, 1950–2024. Rates per 100,000, on a log1p scale.

Example: a drowning plot with ggplot

```
ggplot(six,
       aes(x = deaths,
           y = age_group,
           colour = year
       )
)
+ geom_jitter(height = 0.18, alpha = 0.65)
+ scale_colour_viridis_c()
+ theme_minimal()
```



Which visual channels are used?

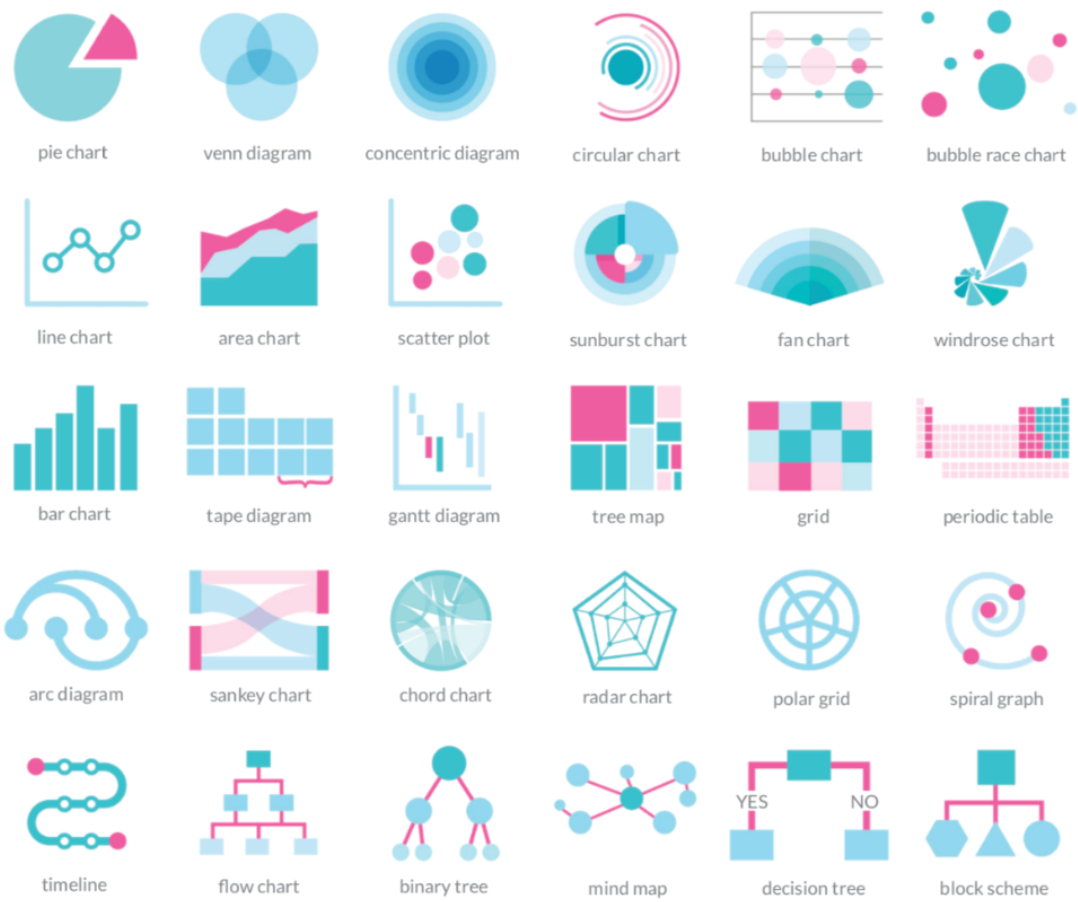
Which geometry and position adjustment?

Is this a good way to visualize these data?

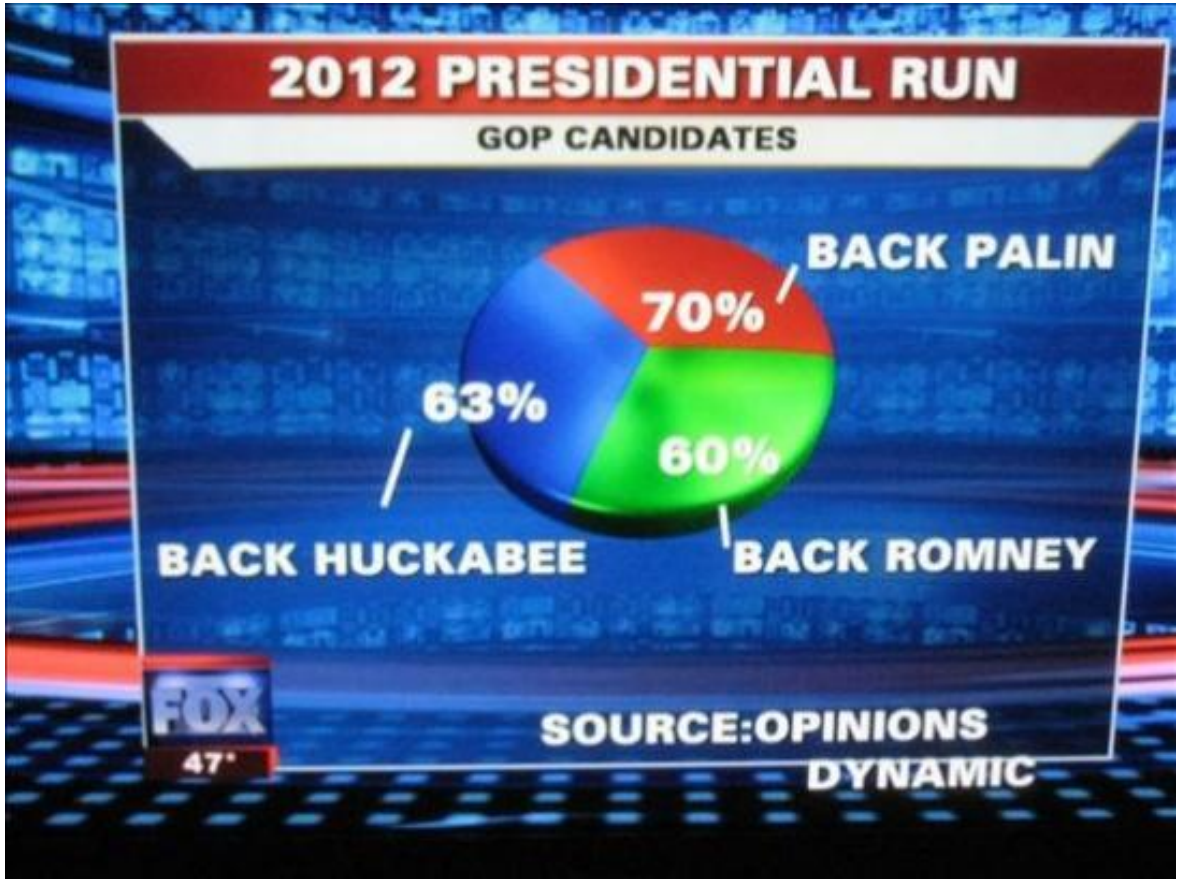
CBS StatLine 7052_95 | Netherlands, 1950–2024

Which plot should I use?

We know how to make a plot with ggplot(). But how do we know which visual channels (aesthetics) and types of plots (geometry) to use?



Separate data problems from encoding problems



Fox News graphic, November 2009.

Data problem

The percentages sum to 193 percent. No visual redesign can repair incorrect numbers.

Encoding problem

The chart uses angle, area and depth for a comparison. A bar chart on a common baseline would be easier to read.

Compare visual channels

tinyurl.com/uu-dataviz

How much more saturation?

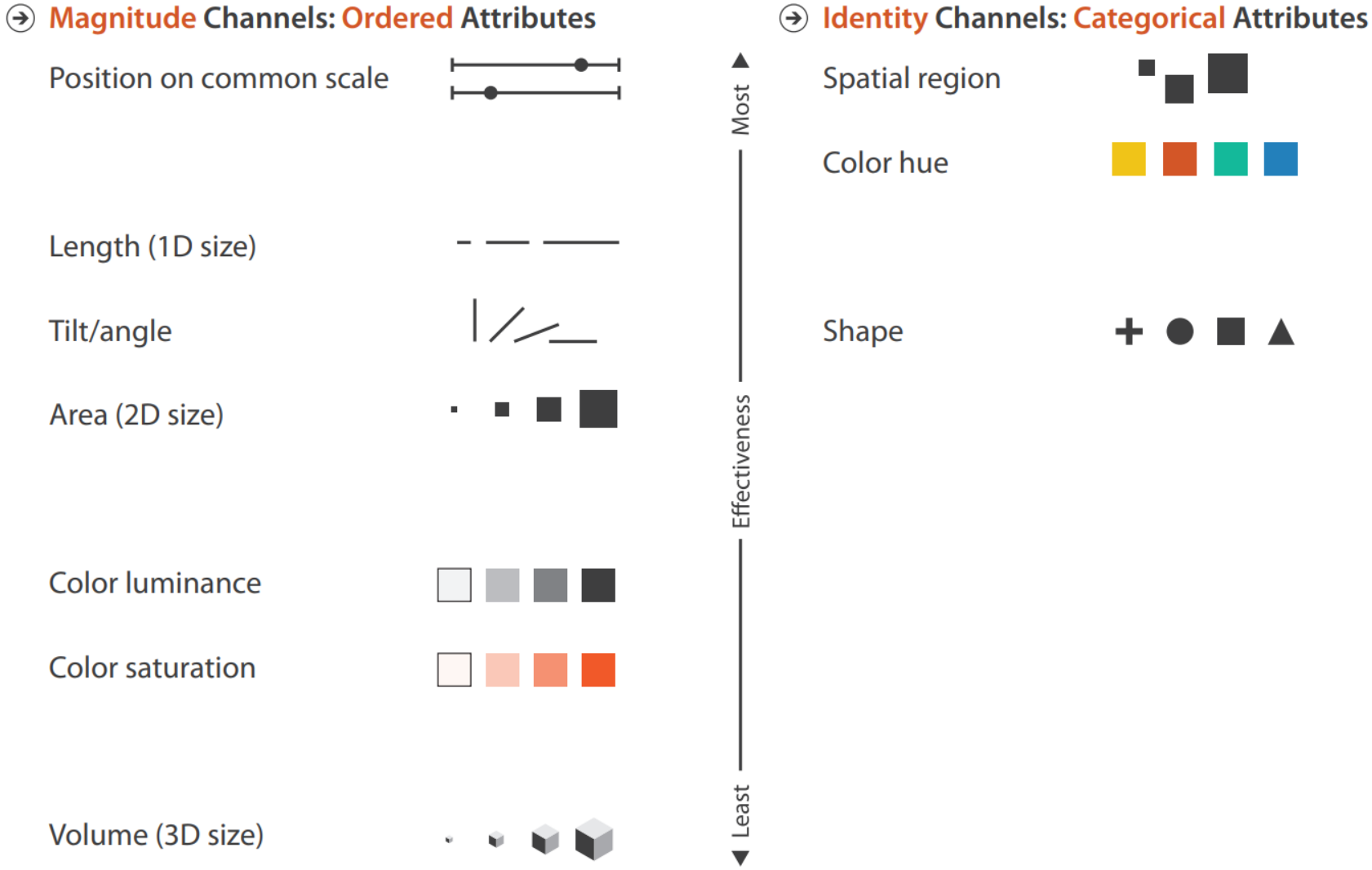


How much more vivid is the square on the left?
Saturation left / saturation right (1=equal saturation)

1 2 3 4 5 6 7 8 9 10

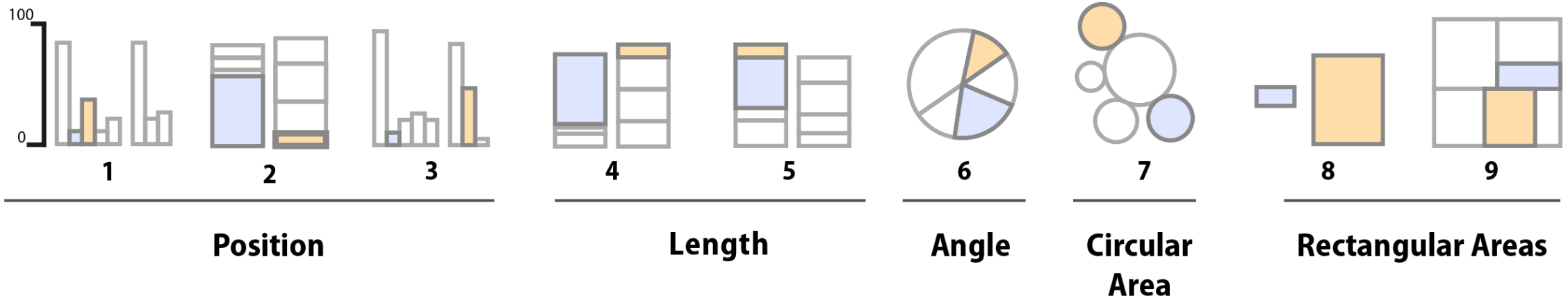
Visual channels differ in accuracy

Channels: Expressiveness Types and Effectiveness Ranks



Munzner, Visualization Analysis and Design, 2014.

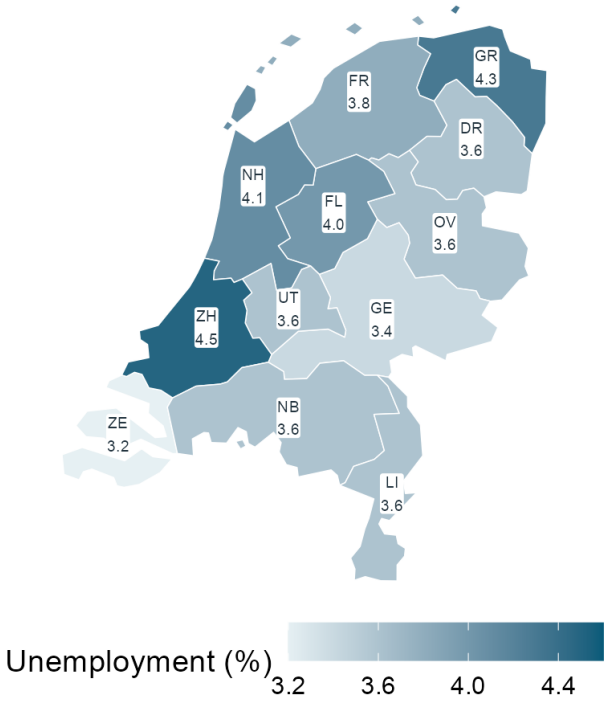
Position and length support precise comparisons



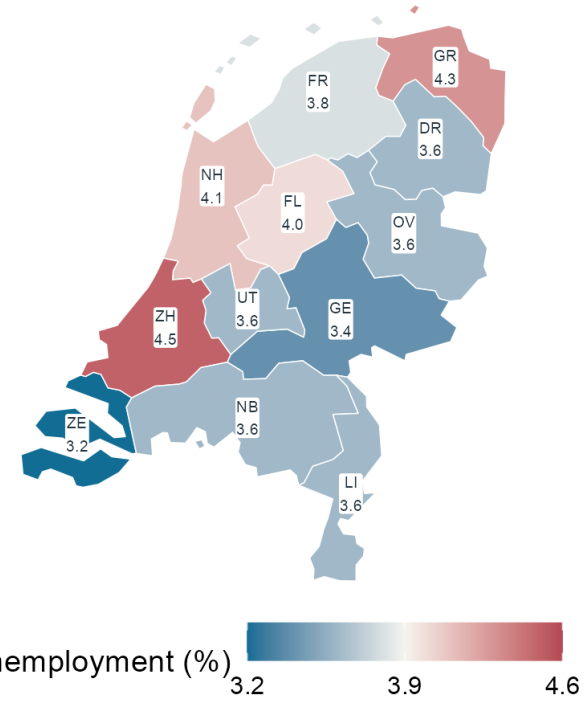
Effectiveness depends on the task.

Use a color palette that fits your message

Sequential: compare unemployment levels



Diverging: compare with the national rate

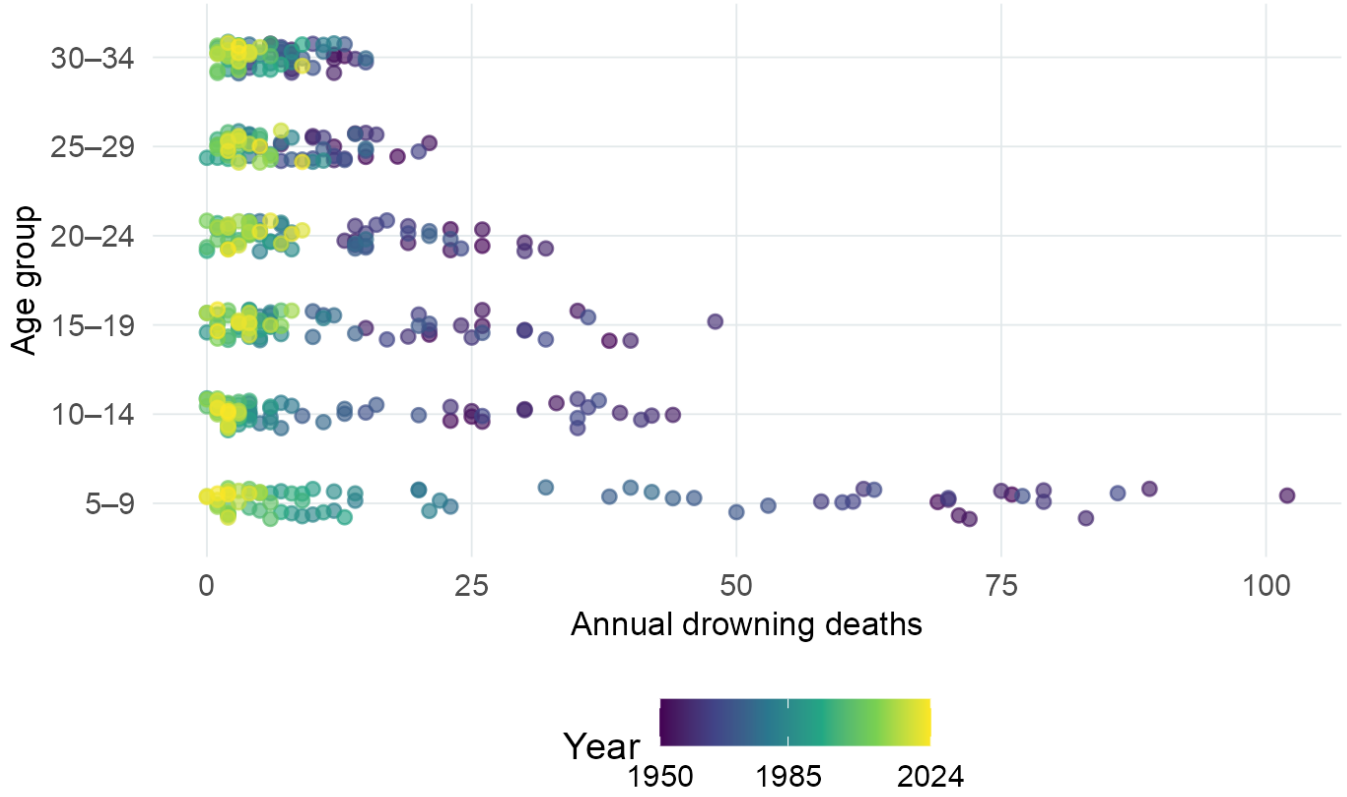


Netherlands, 2025. National unemployment: 3.9%.

Sequential for ordered values; diverging around a meaningful midpoint; distinct hues for categories.

CBS, De regionale economie 2025, table 4.5 | Boundaries: Eurostat GISCO

Improve the mapping of variables to visual channels



Which variables are visualized?

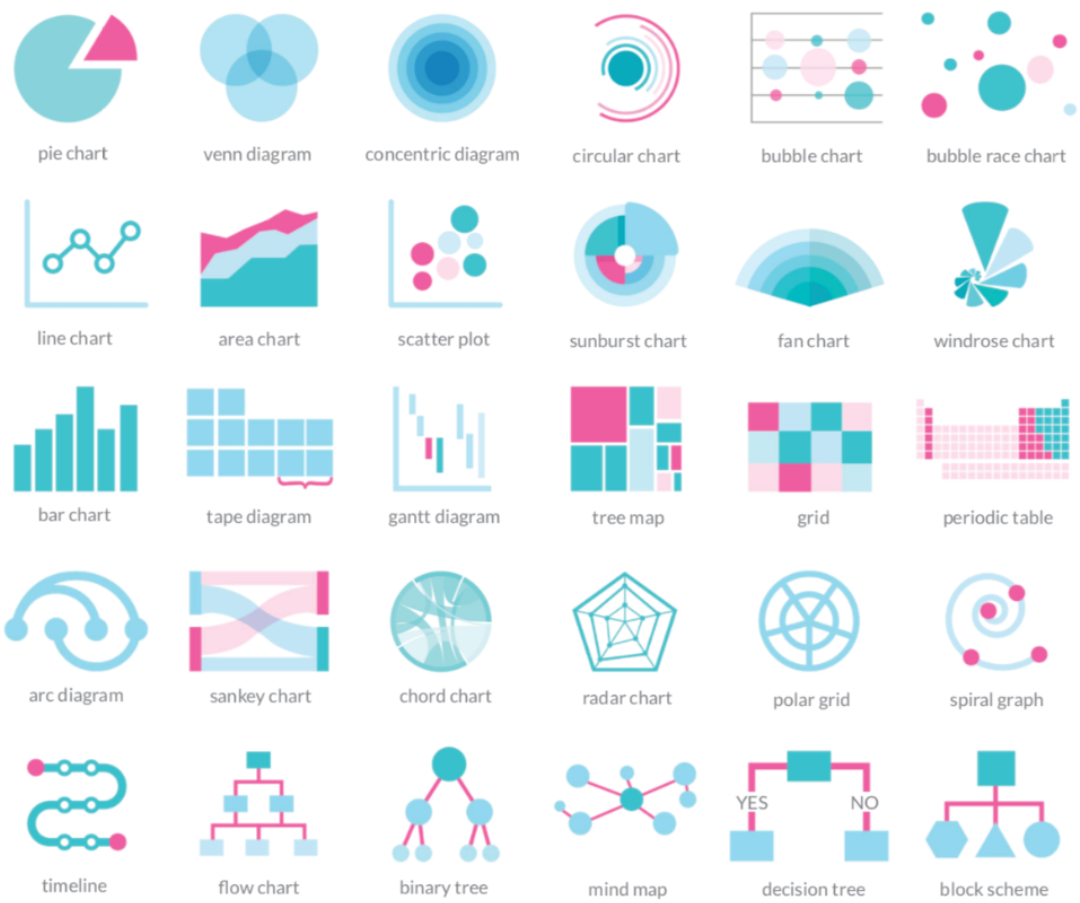
To which visual channels are they mapped?

Are the most important variables mapped to channels with low bias?

What plot should we use?

Which plot should I use?

Now that we know which visual channels (aesthetics) are less biased, which plot (geometries) shall we use?

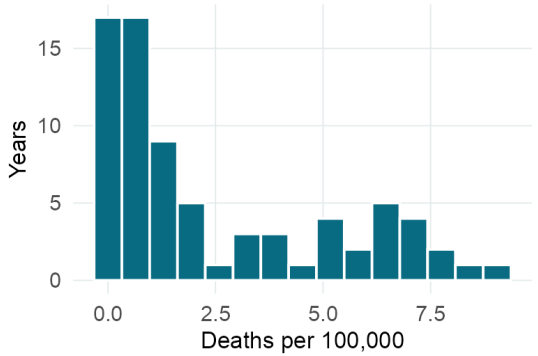


Variation and covariation

Variation

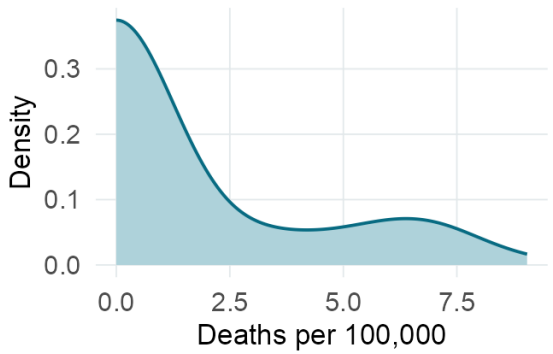
Histogram

`geom_histogram()`



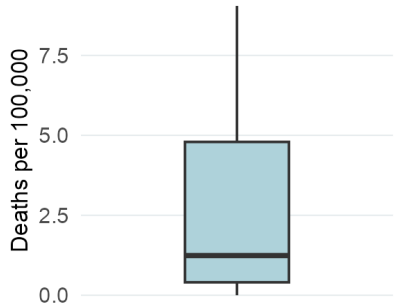
Density

`geom_density()`



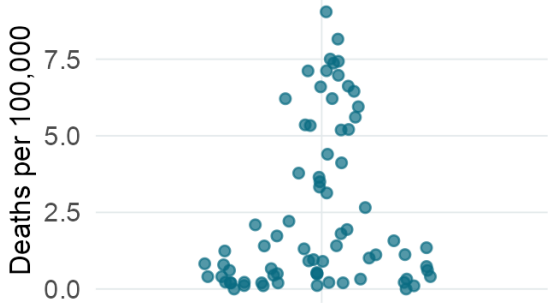
Boxplot

`geom_boxplot()`



Sina plot

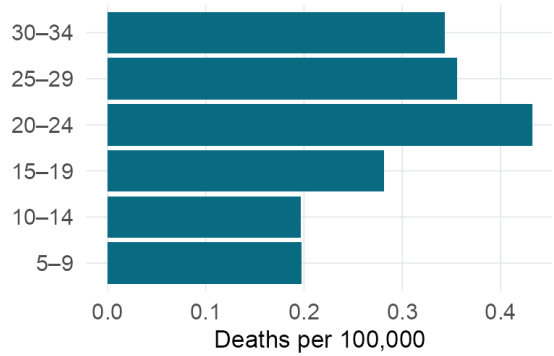
`ggforce::geom_sina()`



Covariation

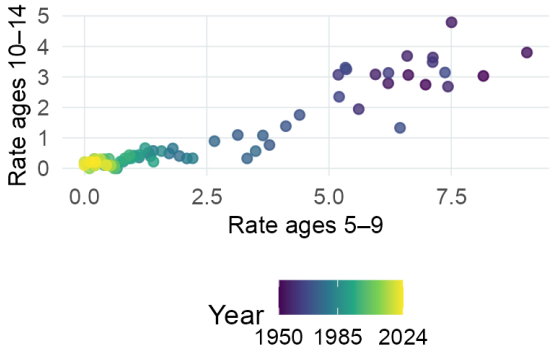
Comparisons

`geom_col()`



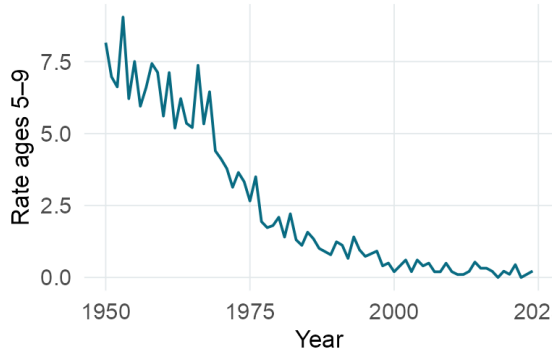
Correlations

`geom_point()`



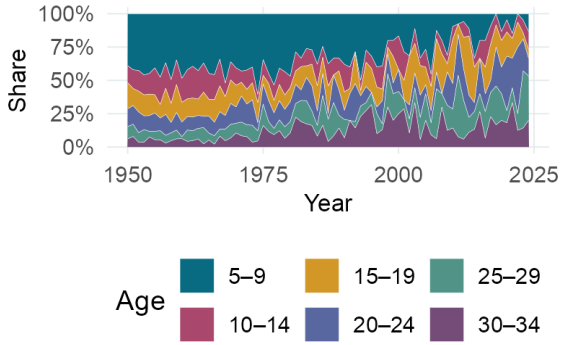
Time series

`geom_line()`



Compositions

`geom_area()`

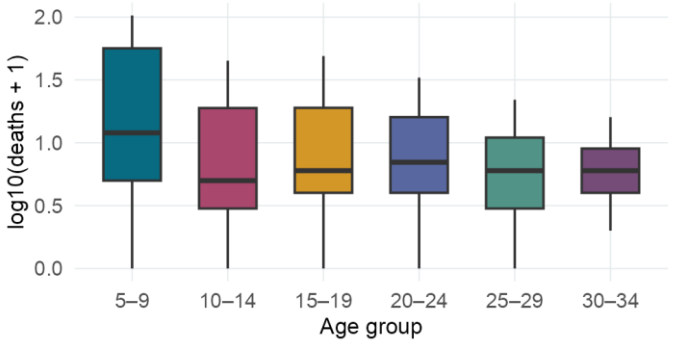


Variation and covariation: distributions

One distribution shows variation. Distributions per age group show covariation.

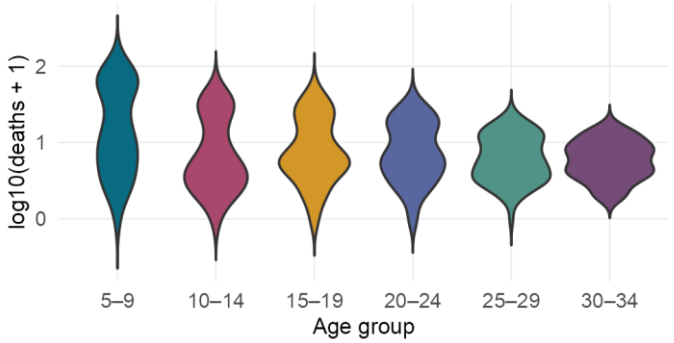
Boxplot

`geom_boxplot()`



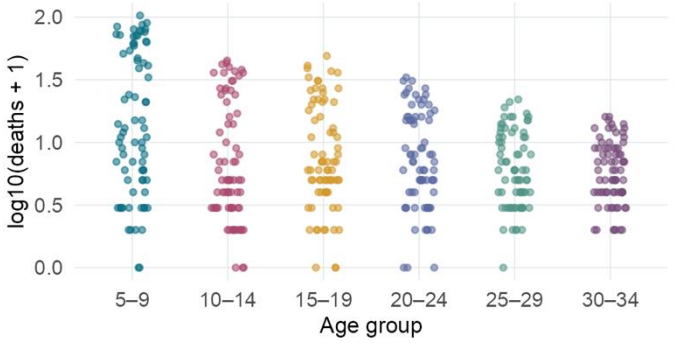
Violin plot

`geom_violin()`



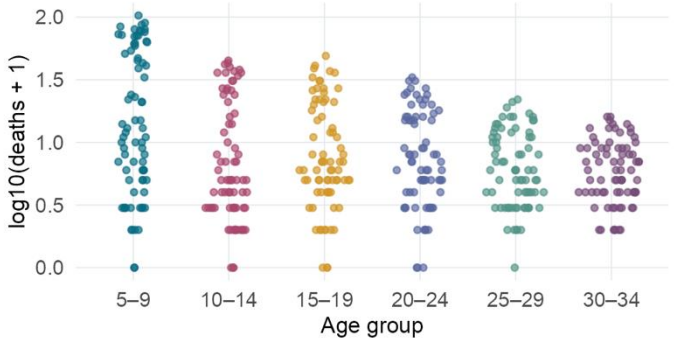
Strip plot

`geom_jitter()`



Sina plot

`ggforce::geom_sina()`

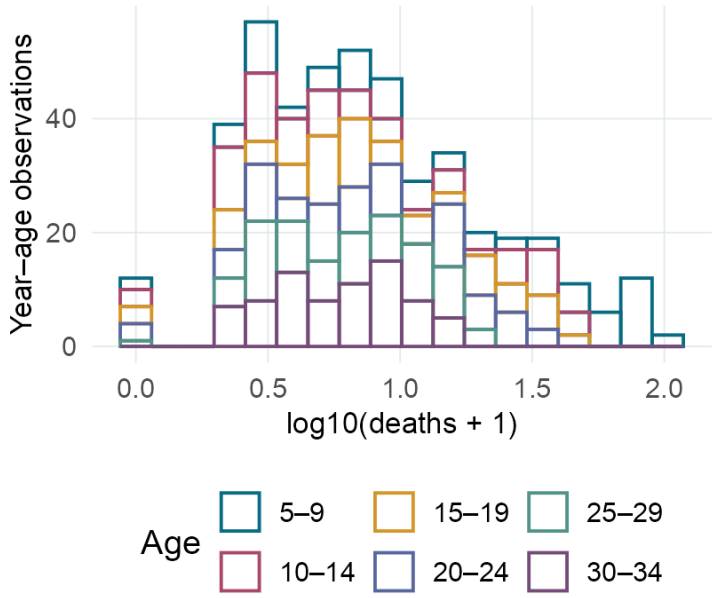


Annual drowning counts, six age groups, 1950–2024 | log10(deaths + 1) retains zero-death years

Variation and covariation: distribution shapes

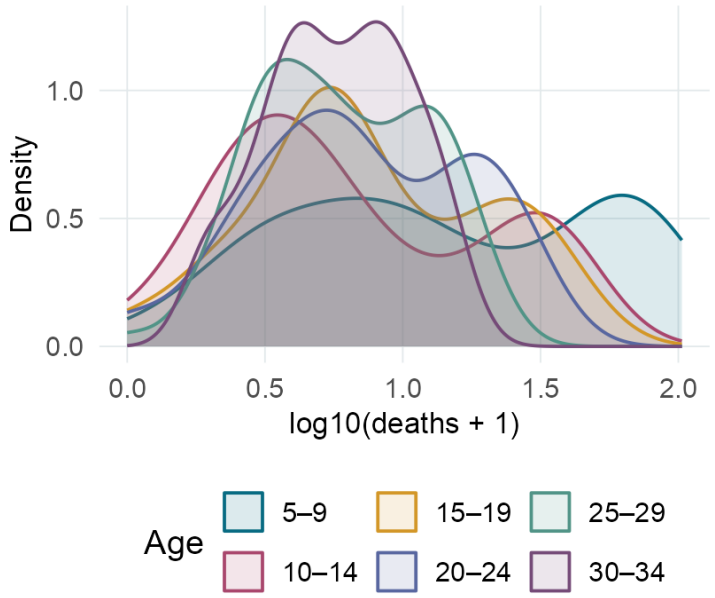
Stacked histograms

`geom_histogram()`



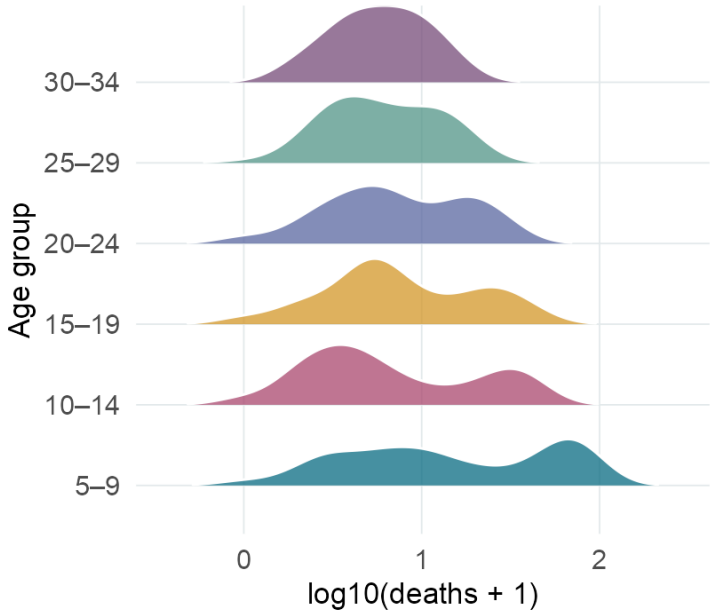
Overlapping densities

`geom_density()`



Ridge plot

`ggridges::geom_density_ridges()`

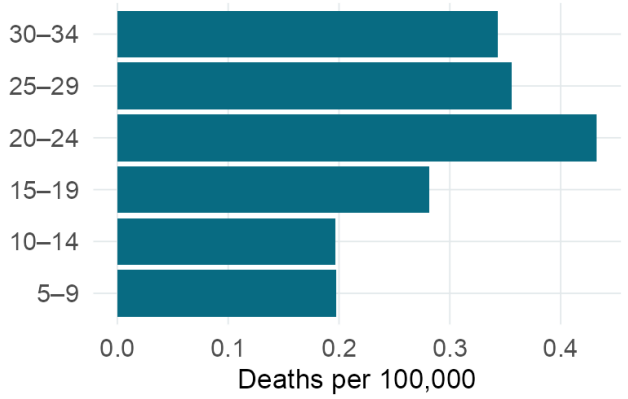


Check bin width or bandwidth before interpreting peaks. Zero-death years remain in log10(deaths + 1).

Comparisons and compositions

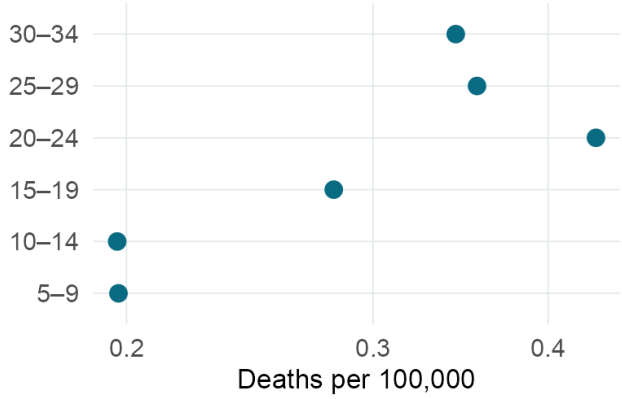
Bars Compare rates

`geom_col()`



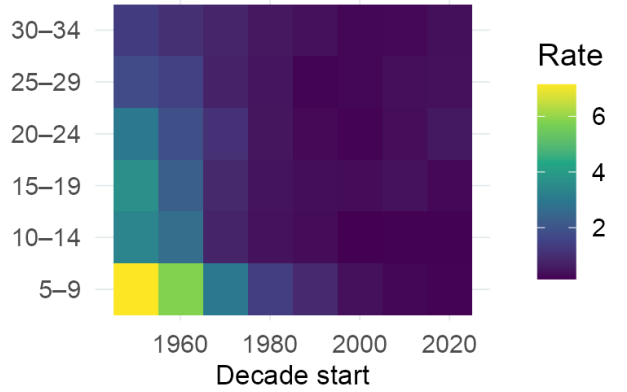
Dots Use dots with log scales

`geom_point()`



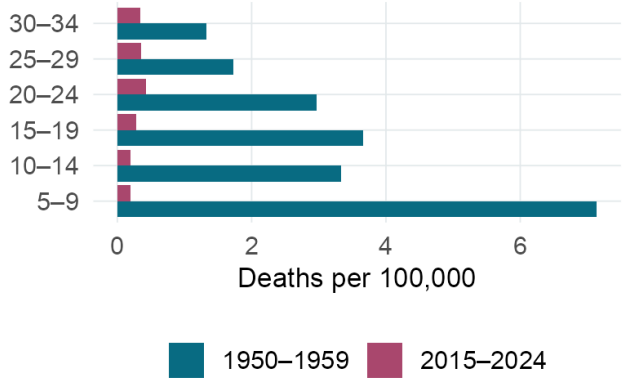
Heatmap Trends across age groups

`geom_tile()`



Grouped bars

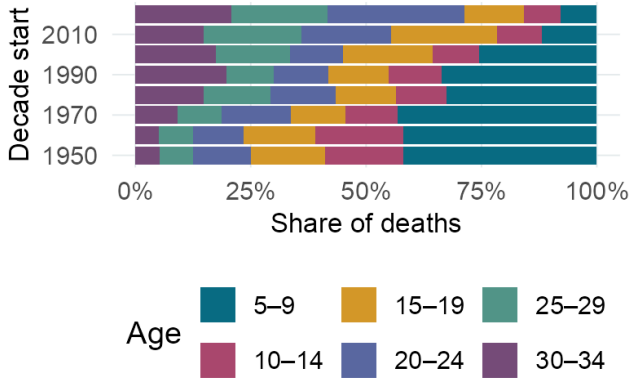
`geom_col(position = "dodge")`



Differences within each group

Stacked bars

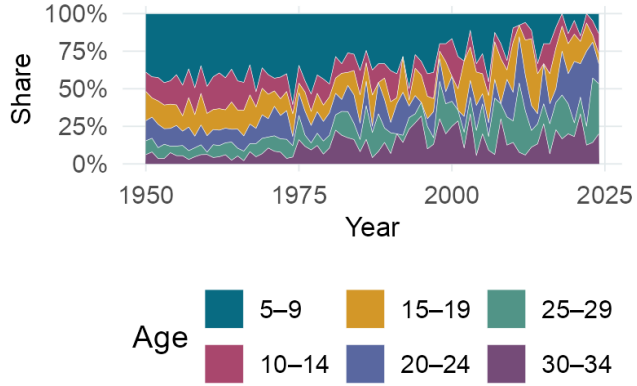
`geom_col(position = "fill")`



Proportions across decades

Stacked area

`geom_area()`

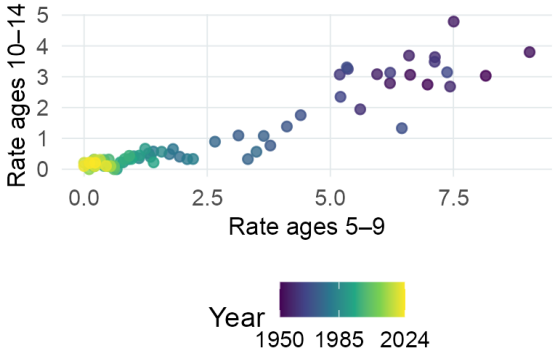


Proportions over time

Covariation: correlations and time series

Scatterplot

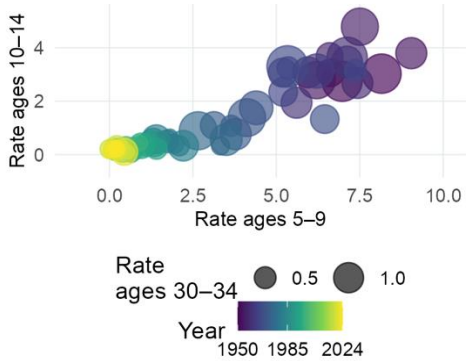
`geom_point()`



Each point is one year

Bubble plot

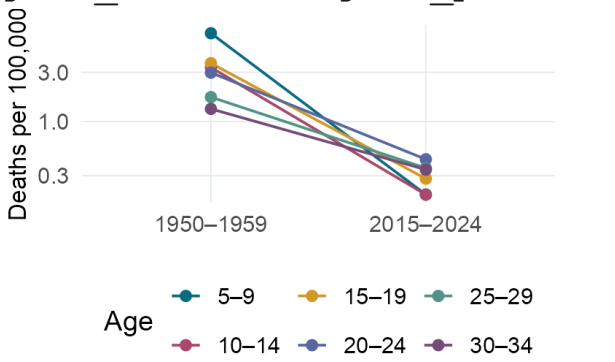
`geom_point()`



Size adds the rate at ages 30–34

Slopegraph

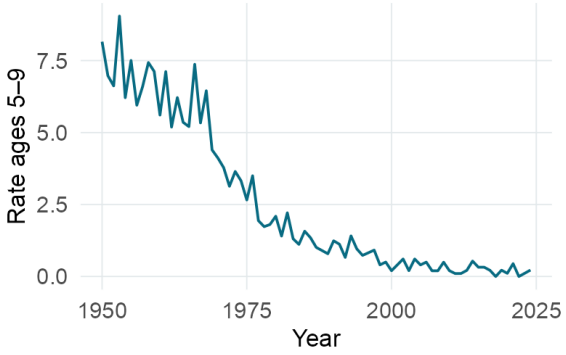
`geom_line() + geom_point()`



Compare two periods

Line graph

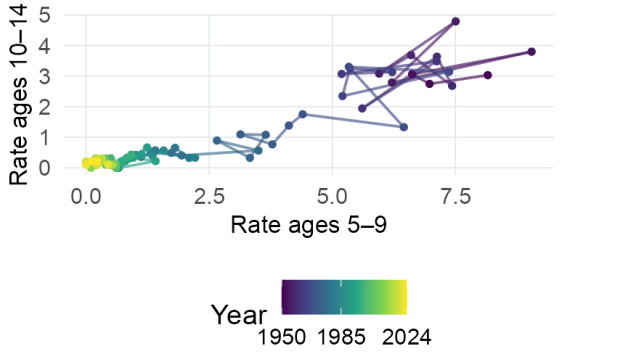
`geom_line()`



Follow change over time

Connected scatterplot

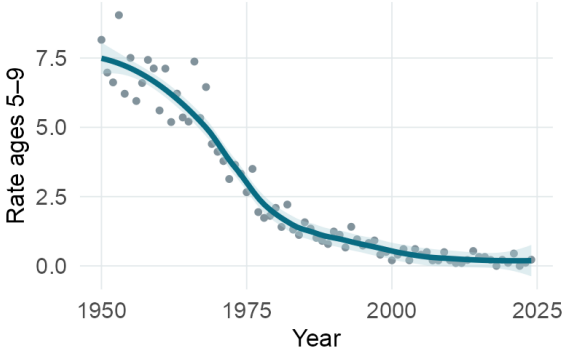
`geom_path() + geom_point()`



Connect years in order

Smoothed time series

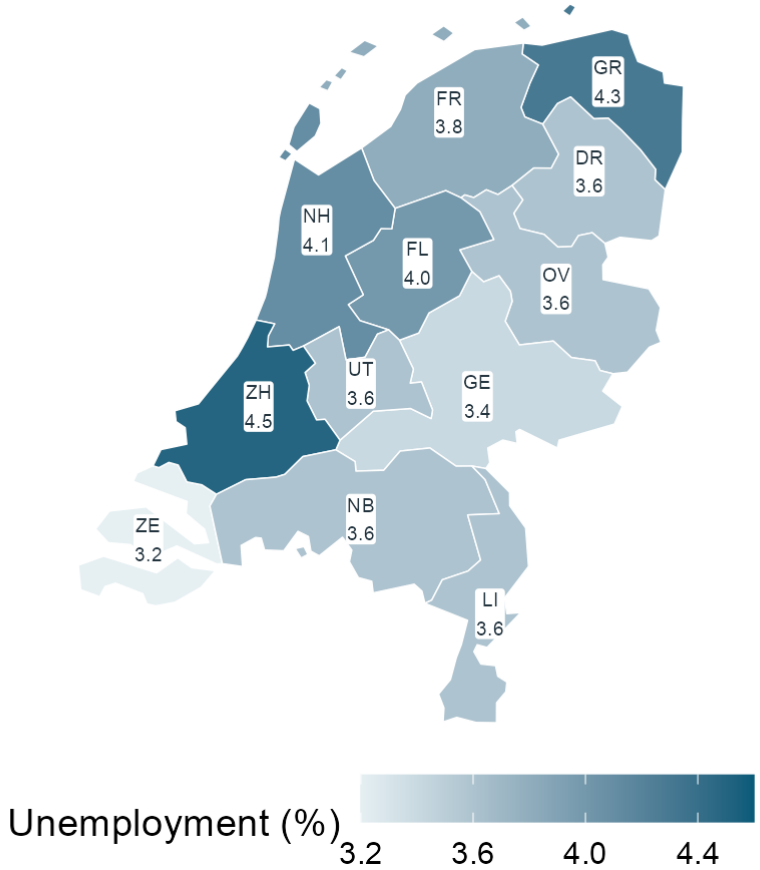
`geom_smooth()`



Inspect the trend

Color and continuous variables: geographical data

Where are similar rates located?



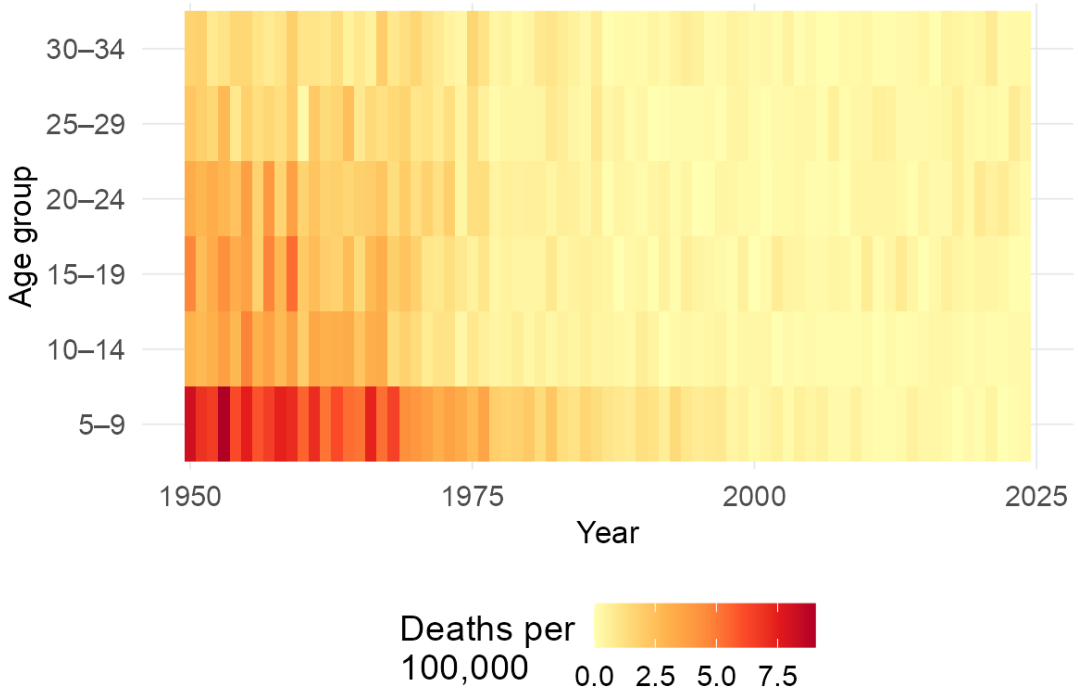
Which provinces have the highest rates?



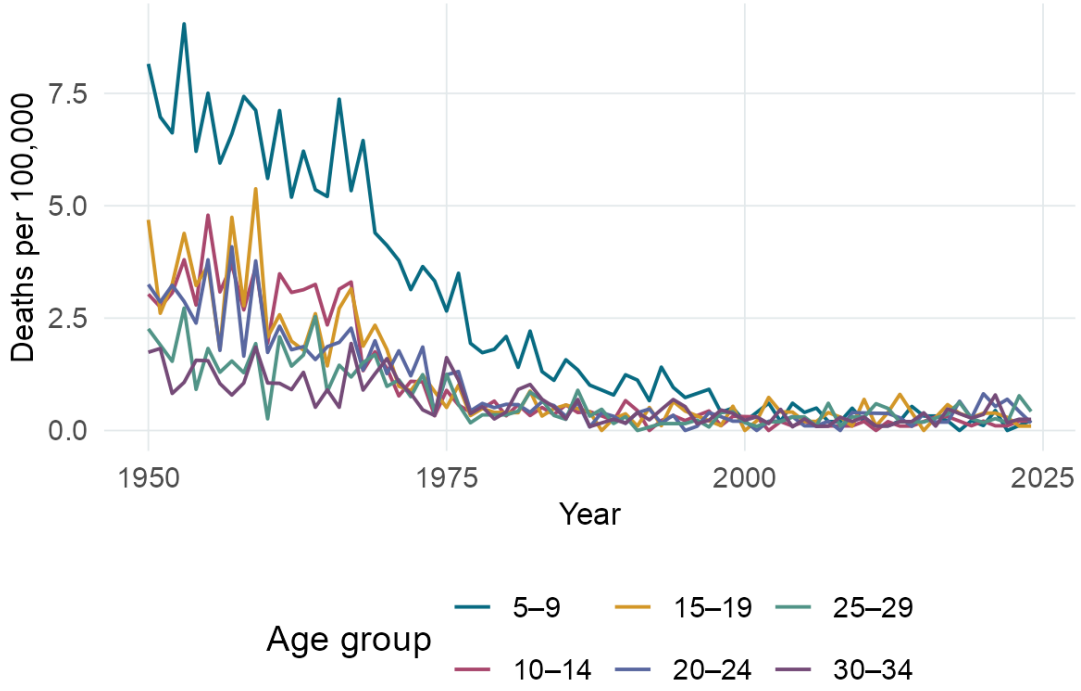
Unemployment, 2025 | CBS, De regionale economie 2025 | Boundaries: Eurostat GISCO

Color and continuous variables: map patterns

Rate mapped to colour



Rate mapped to vertical position



Which plot makes trends easier to compare?

Part 3: what you should be able to do now

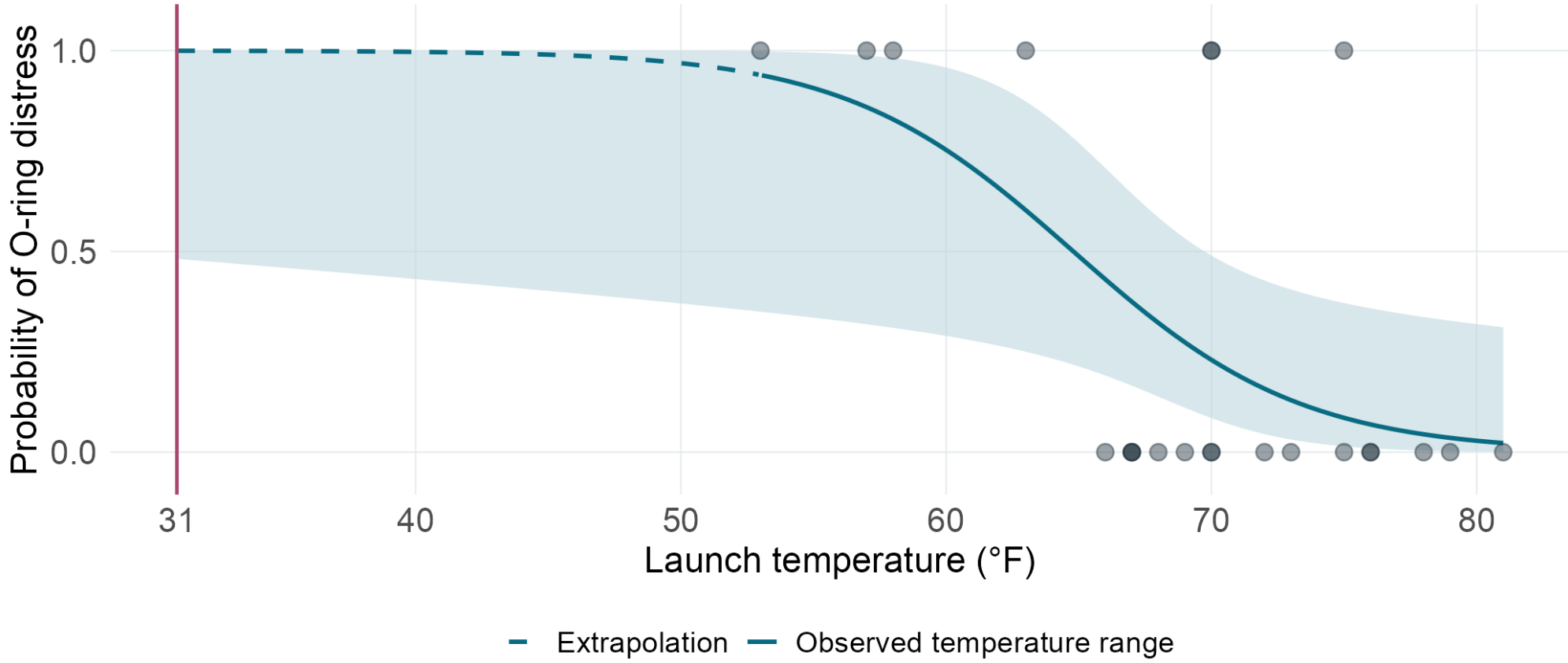
Choose and justify plots using the grammar of graphics and visual channels.

Identify data, mappings, geoms, scales, facets and statistical transformations.

Explain which visual channels are less biased (position, length)

Choose a plot for a variation or covariation (group comparison, composition, correlation or time series)

Example: Exploratory data visualization



23 earlier flights | Logistic fit with pointwise 95% confidence band | UCI DOI: 10.24432/C5PW2T

PART 04

Making the argument

Communicate findings through design and storytelling

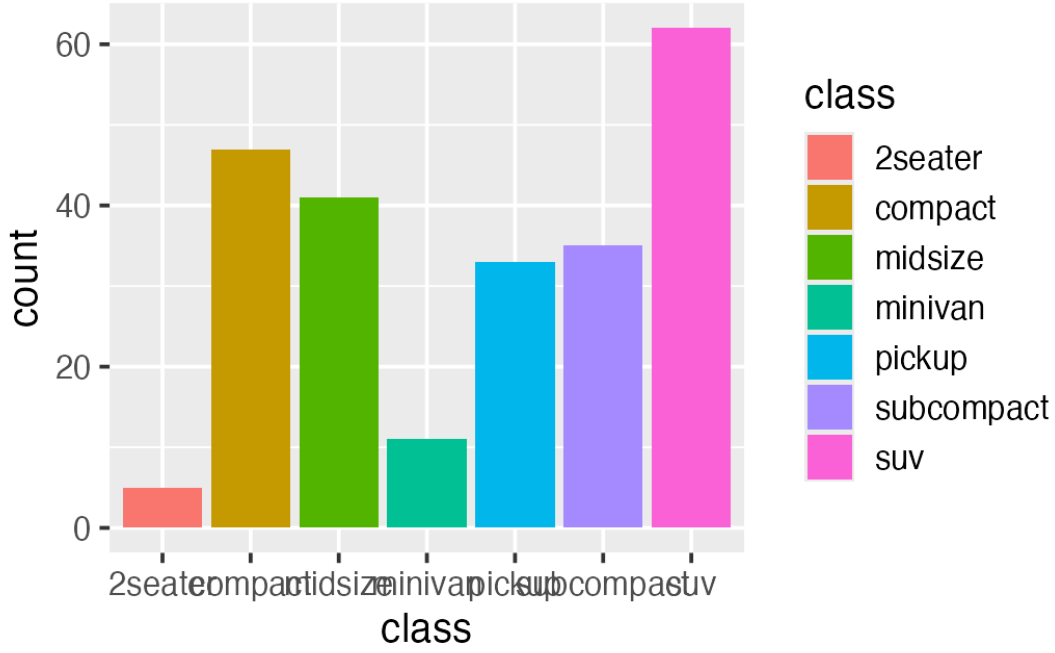
CRUCIAL: State what the reader should remember

Write the main message in one sentence before designing the figure.

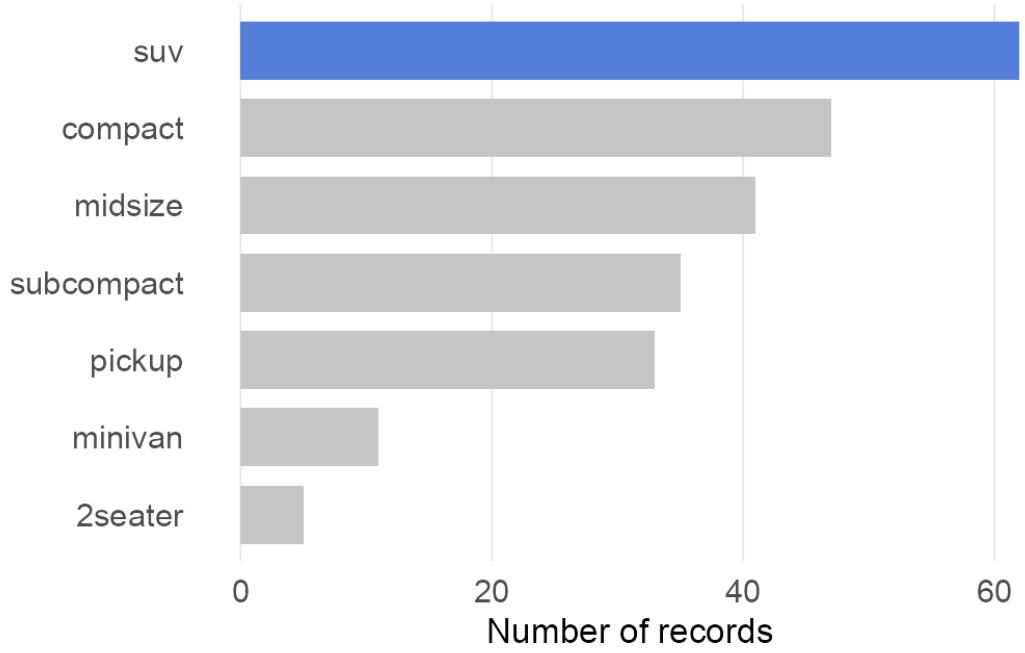
Choose the comparisons and emphasis that help explain that message.

Guide the reader through the plot

Design principles, guidance, storytelling



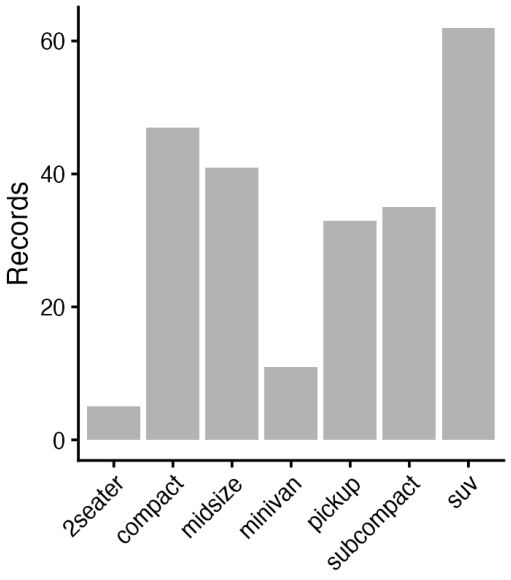
SUVs are the most popular car



Design: CRAP principles

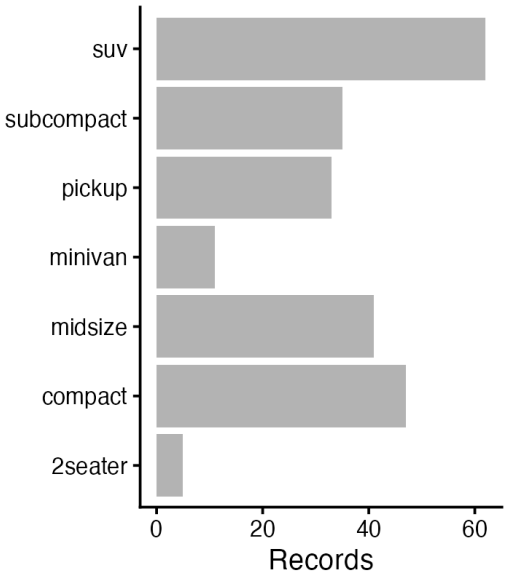
Contrast

Remove colour
and background.



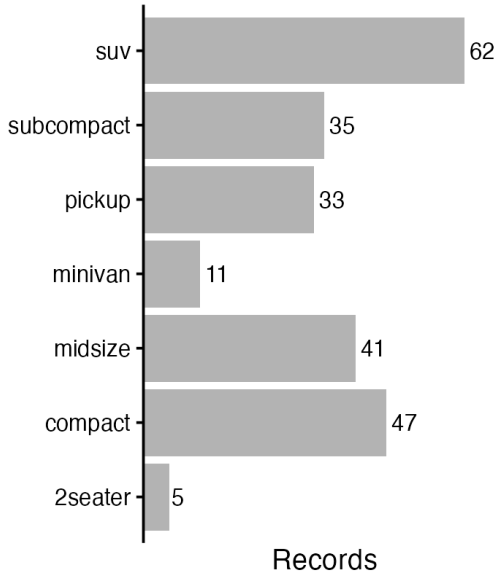
Alignment

Flip the chart
for readable labels.



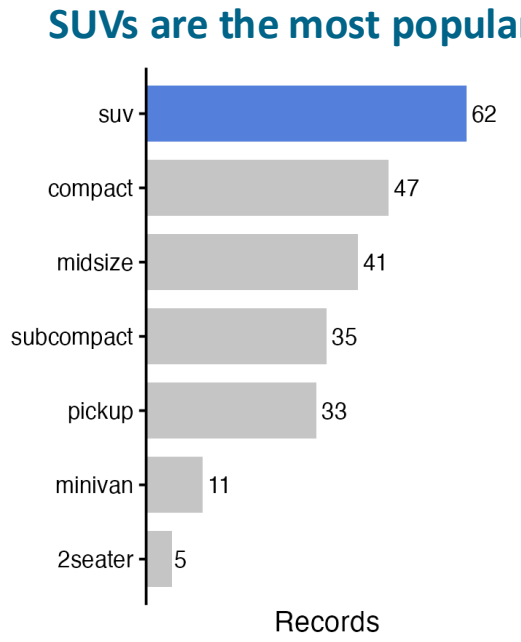
Proximity

Place values
next to the bars.



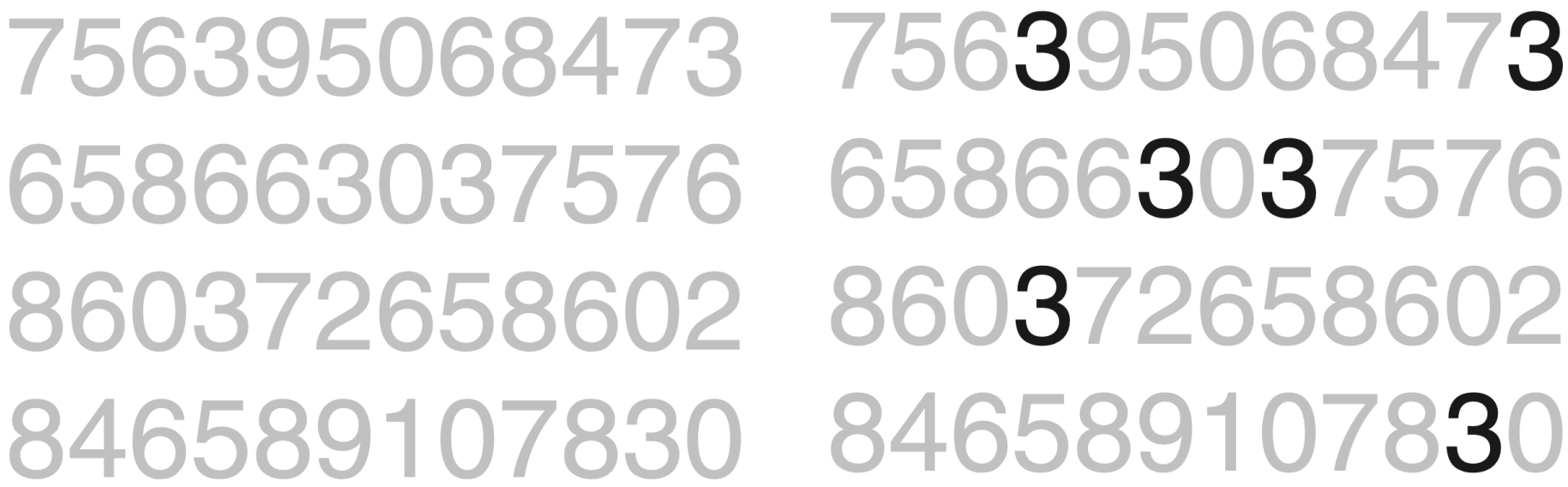
Guide attention / Contrast

Sort, highlight
and add a title.



Repetition: use colours, labels and styles consistently across plots.

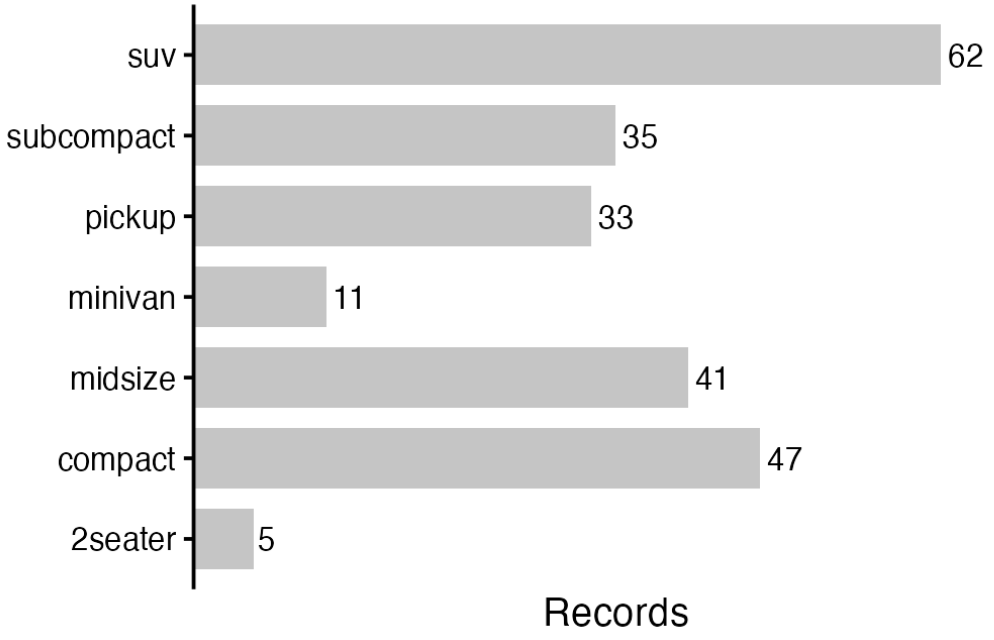
Guide the reader: pre-attentive attributes

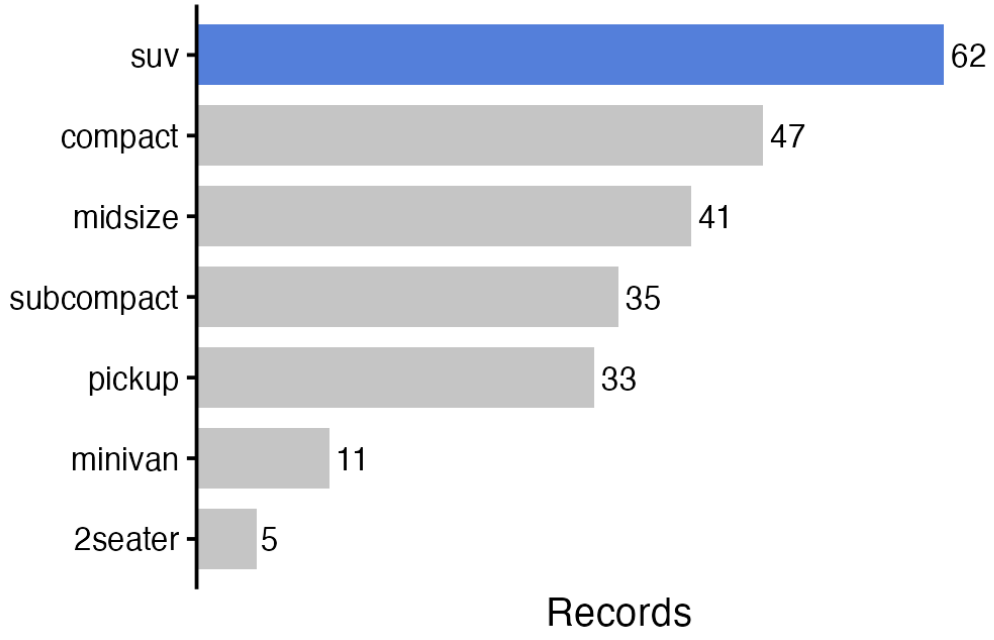


Color and contrast can guide attention before deliberate reading.

After Knaflic, Storytelling with Data, 2015.

Guide attention to the main comparison





Color and emotion

455 smokers compared dark and mustard-coloured packs (Pantone 448C).

Less willing to smoke from the pack

55% vs 19%

Dark colour vs mustard

Source: GfK Blue Moon (2011), Study 4, Table 32. Background: screen approximation.

Storytelling: humans get high on stories

Exposition

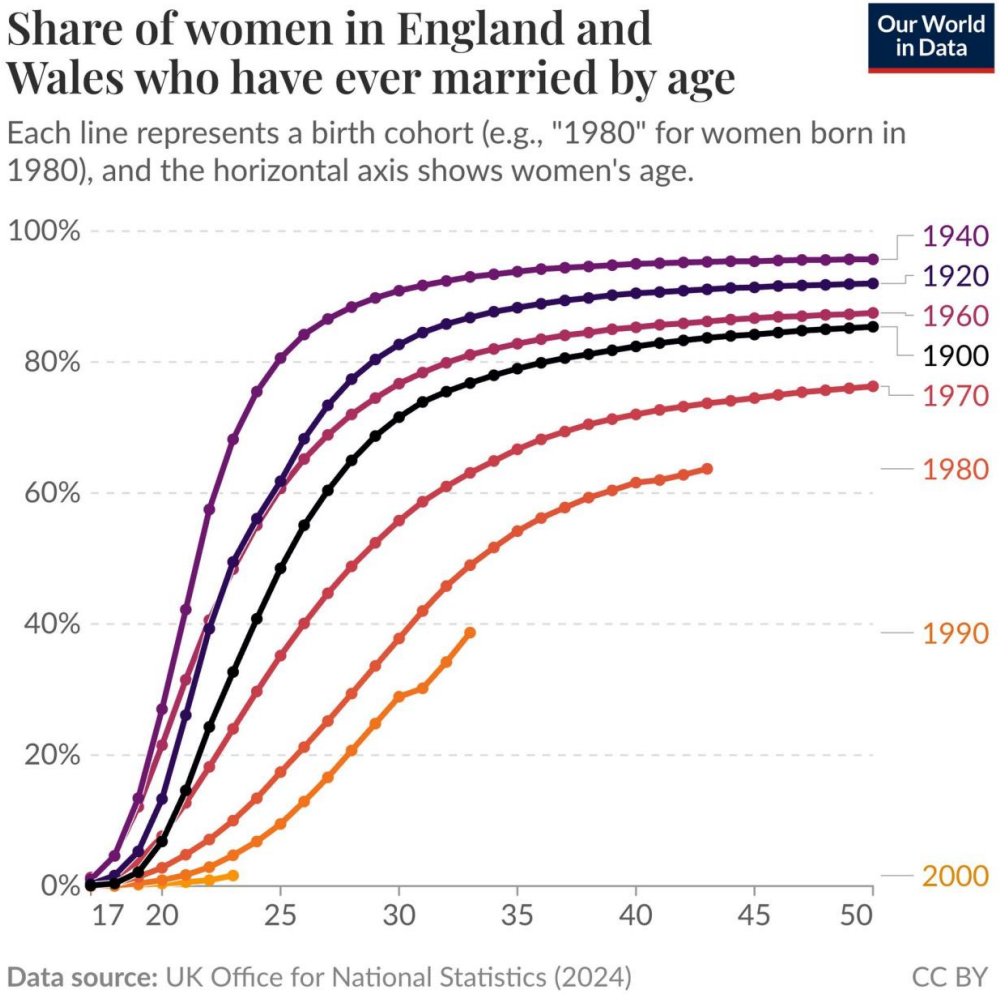
Compare the share ever married at the same ages across birth cohorts.

Conflict

Between cohorts, in different colors

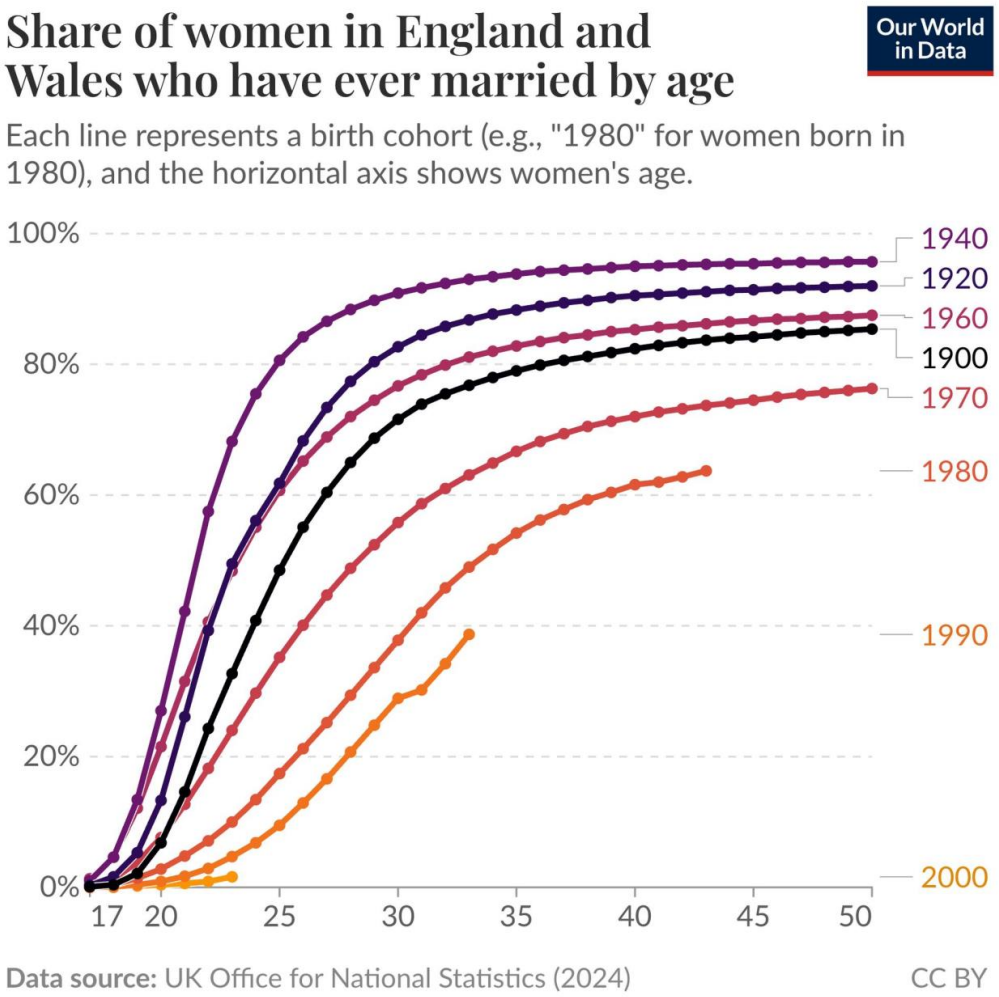
Resolution

Marriage happens later in the more recent cohorts. Their curves rise more slowly.



You will read this first.

You will read this second.

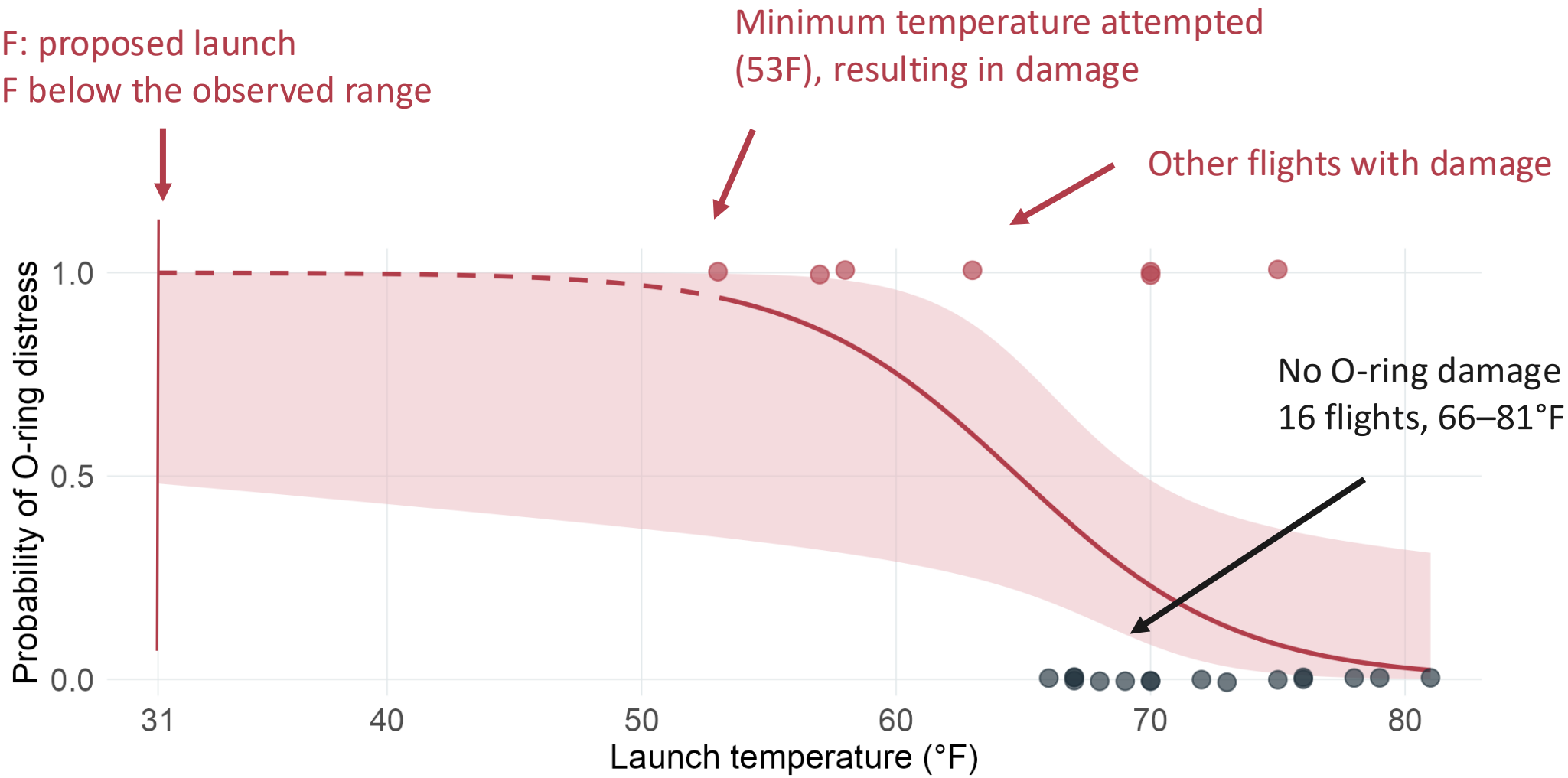


You will read this third.

You will read this fourth.

O-ring damage is very likely at 31F

All flights under 65F had damages



O-ring distress, not catastrophic failure | Dashed curve: extrapolation | Shaded band: 95% confidence interval

Learning goals of today

1. Use EDA to understand data and refine questions.

   Right question, right data, feasible solution.
   Try to break the result and refine the question.

2. Explain why we visualize data.

   Visual patterns help us understand and remember.
   Explore for yourself, explain to others.

3. Choose and justify plots using the grammar of graphics and visual channels.

   Choose geoms for the task and map variables to suitable channels.
   Position and length support accurate comparisons.

4. Communicate findings through design and storytelling.

   Write one message. Use design, guide attention, build a story.

Next week: missing data