Week 2: Graphics and Visualization

MATH-517 Statistical Computation and Visualization

Linda Mhalla

September 20th 2024

Graphics

"The simple graph has brought more information to the data analyst's mind than any other device." – *John W. Tukey*

"The greatest value of a picture is when it forces us to notice what we never expected to see." – *John W. Tukey*

One can think of graphics (and also models, for that matter) as a low-dimensional representation for data

Anscombe's Quartet

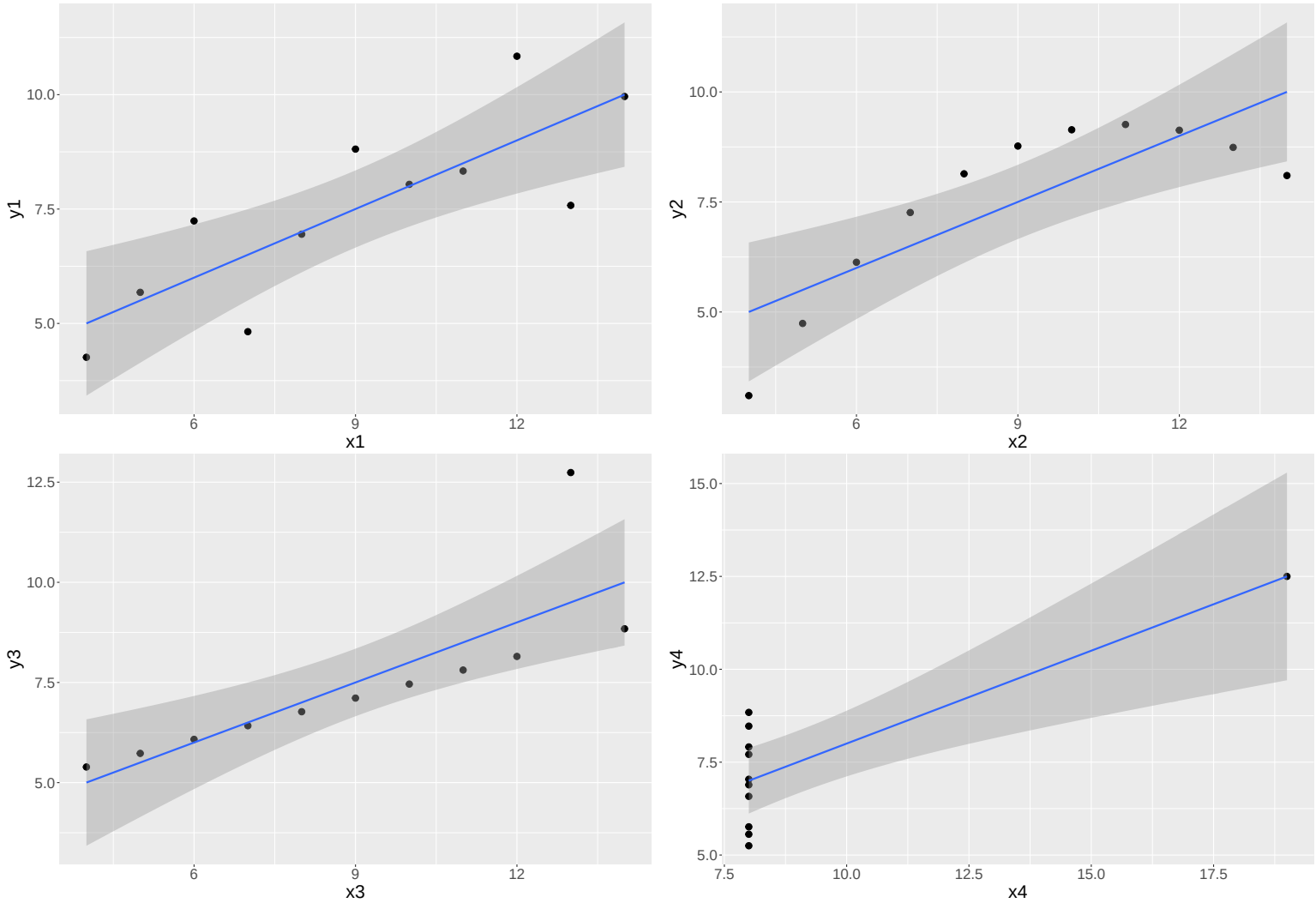
Four data sets with very similar descriptive statistics, each with

- one response variable y
- one regressor x

```
##     (Intercept)         x R-squared
## lm1    3.000091 0.5000909 0.6665425
## lm2    3.000909 0.5000000 0.6662420
## lm3    3.002455 0.4997273 0.6663240
## lm4    3.001727 0.4999091 0.6667073
```

$\Rightarrow$ how to fool the linear regression model ...

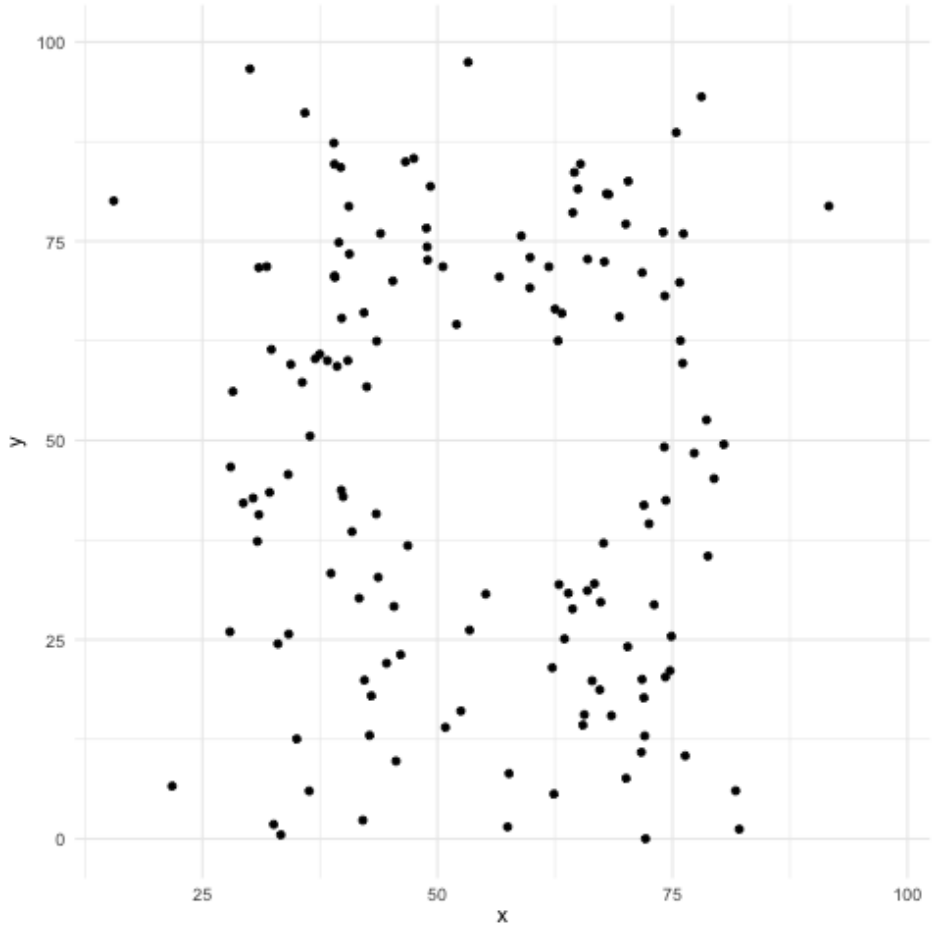
Anscombe's Quartet



$\Rightarrow$ qualitatively different structures

Datasaurus Dozen

A group of twelve datasets with almost identical summary statistics but are distinctively different when plotted!



Purposes of Visualization

The most common purposes of a visualization is

- data insight (exploratory graphs)
  - large data
  - detect important trends/patterns
  - find strange observations
- presentation (explanatory graphs)
  - result communication
  - decision making

$\Rightarrow$ a good choice of axes, axis limits, labels and symbols can facilitate substantially the extraction of information from the data

Datasets used for illustration are described in details in the Lecture Notes 2

What makes a good graph?

". . . that which gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space" – *Edward Tufte*

". . . graphical excellence requires telling the truth about the data." – *Edward Tufte*

Tufte's design principles for effective data visualization

- Principle 1: maximize data-ink ratio
- Principle 2: minimize chart junk
- Principle 3: minimize lie factor

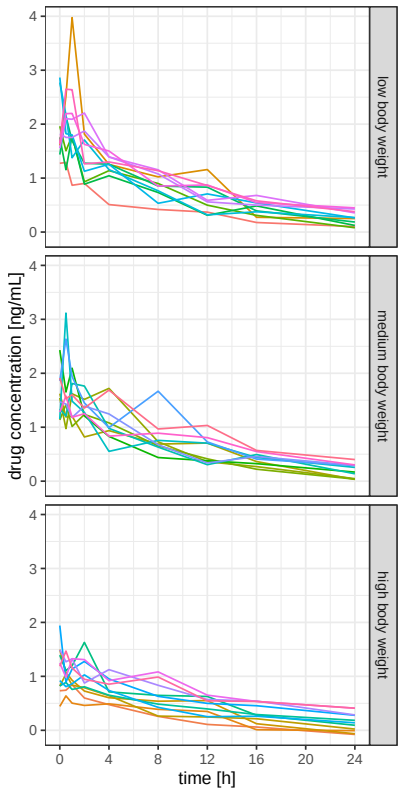
Section 1

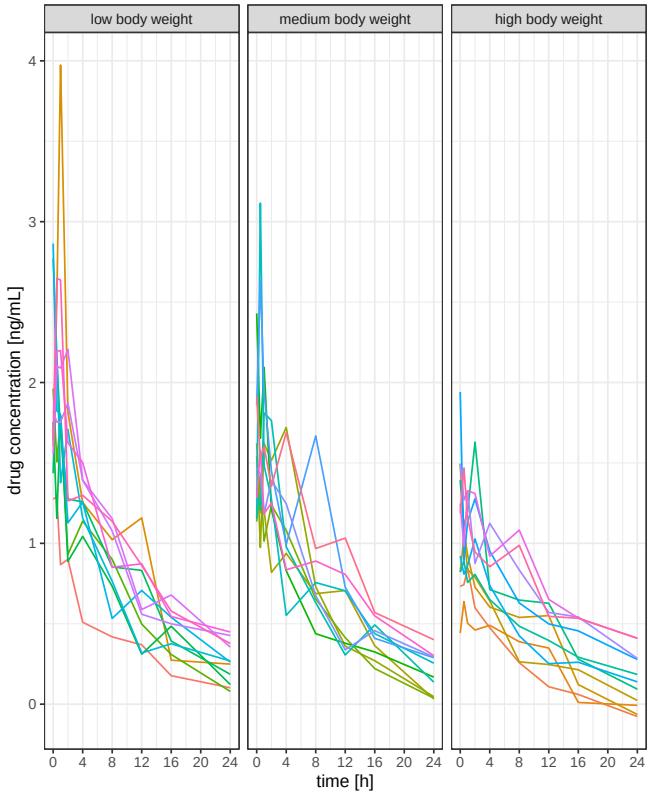
Elements of Visualization

Elements of Visualization: Layout

The arrangement of multiple panels is used for efficient comparison. To compare data on

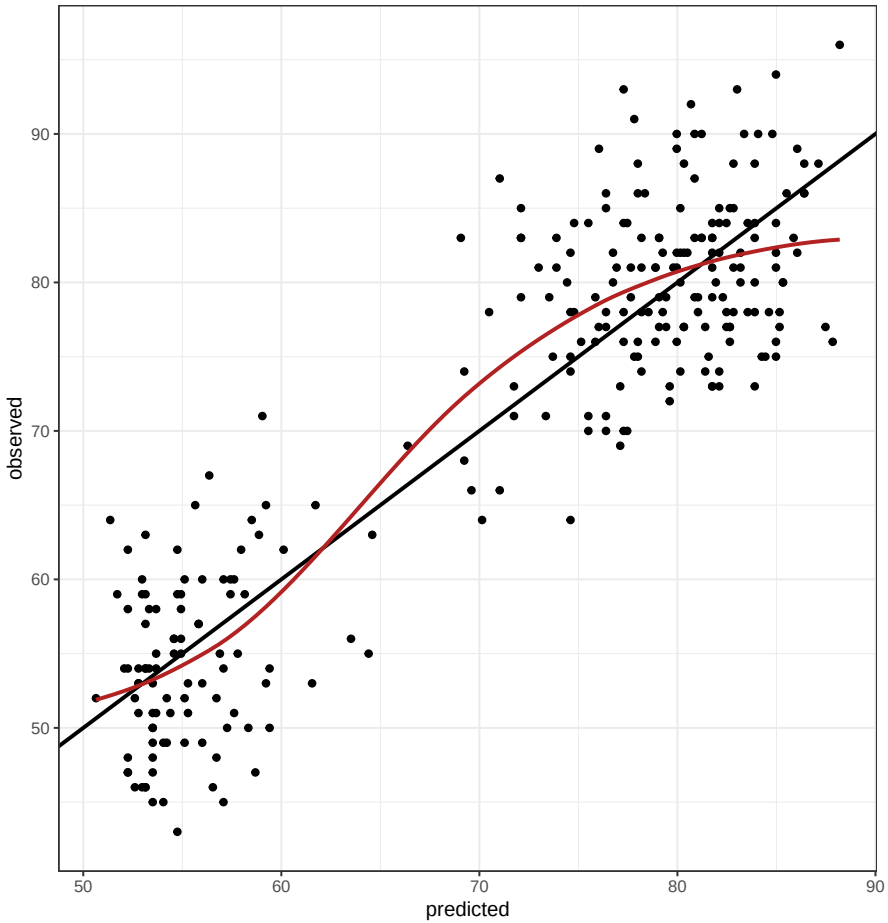
- the x-axis, stack panels sharing a single x-axis
- the y-axis, use a single y-axis with all panels aligned horizontally

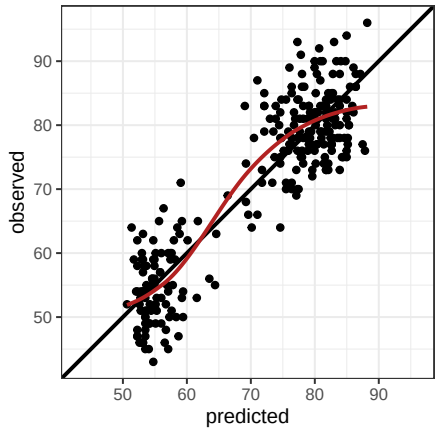




Elements of Visualization: Aspect Ratio

- If measurements on both axes reflect the same quantity (e.g., before vs after treatment, observed vs modelled), a square figure (1:1 aspect ratio) could avoid visual bias + make distances consistent and comparable

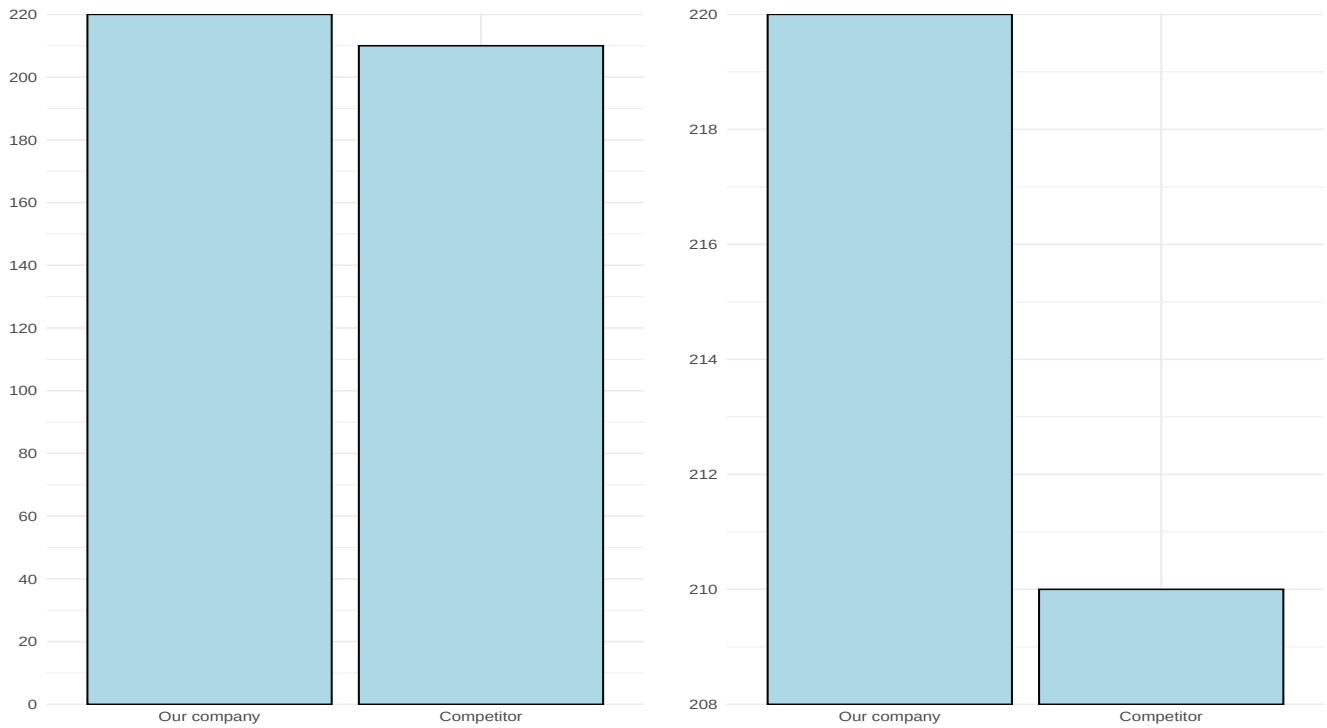




Elements of Visualization: Axes

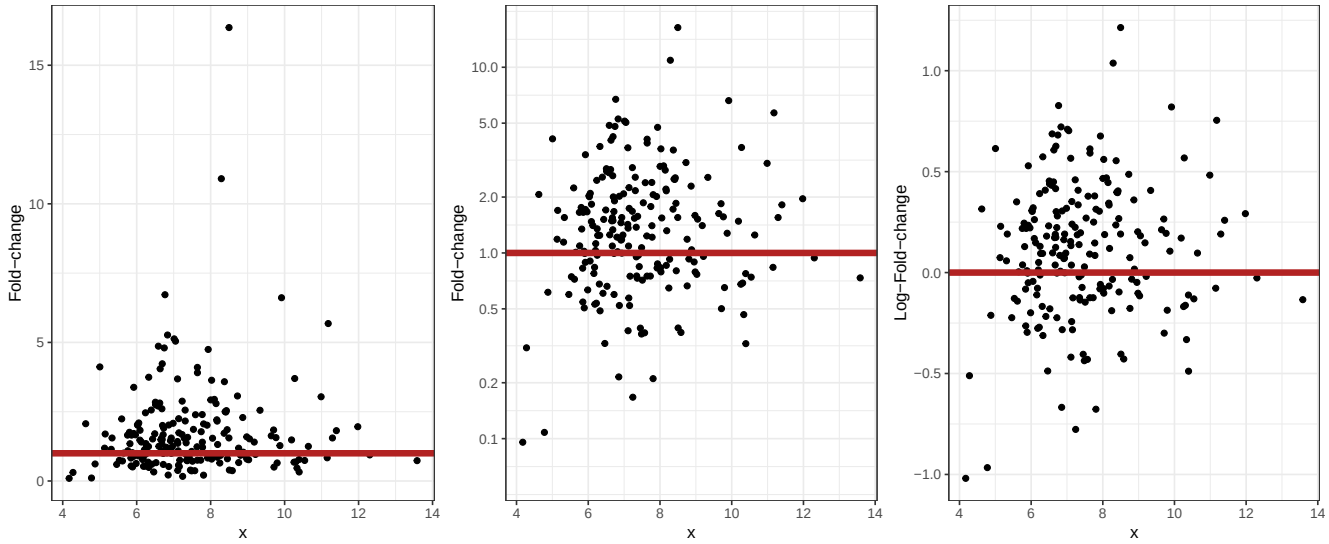
- If values are positive, axes should start at zero (unless there is a good reason for different choice) and not contain negative values

$\rightarrow$ this usually applies to barplots used to highlight relative differences



Elements of Visualization: Axes

- If values are ratios or relative changes, axes should be logarithmic and symmetric around the point of no change (e.g., 0)



Elements of Visualization: Colours and Shapes

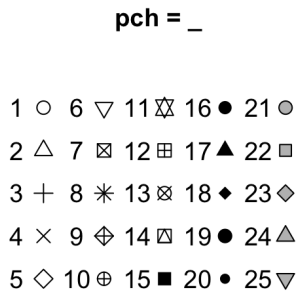
Chart features (`ggplot`'s arguments) for points (and similarly for lines)

- `colour`: help identify different groups (do not use for pure decoration)

Tufte pointed out that because "they do have a natural visual hierarchy, varying shades of gray show varying quantities better than color", and "the shades of gray provide an easily comprehended order to the data measures. This is the key".

Elements of Visualization: Colours and Shapes

- `colour`
- `shape`: if data are ordered, use ordered symbols (number of vertices)



- `size`
- `alpha` (opacity/transparency): can be used to clarify plots with many points

These features can be used

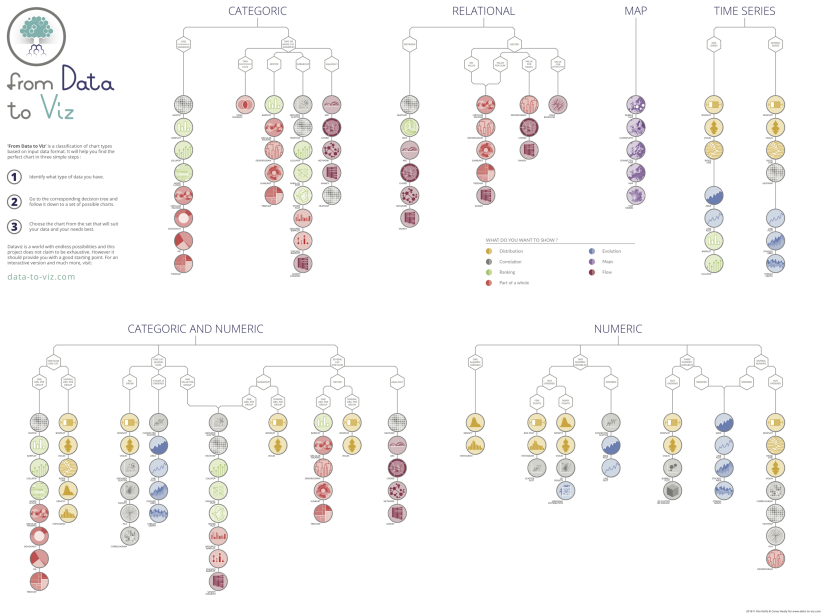
- to include additional information (or dimensions, i.e., to include additional variables) in a scatterplot
- to combat overplotting

Section 2

Visualization types

What type of chart?

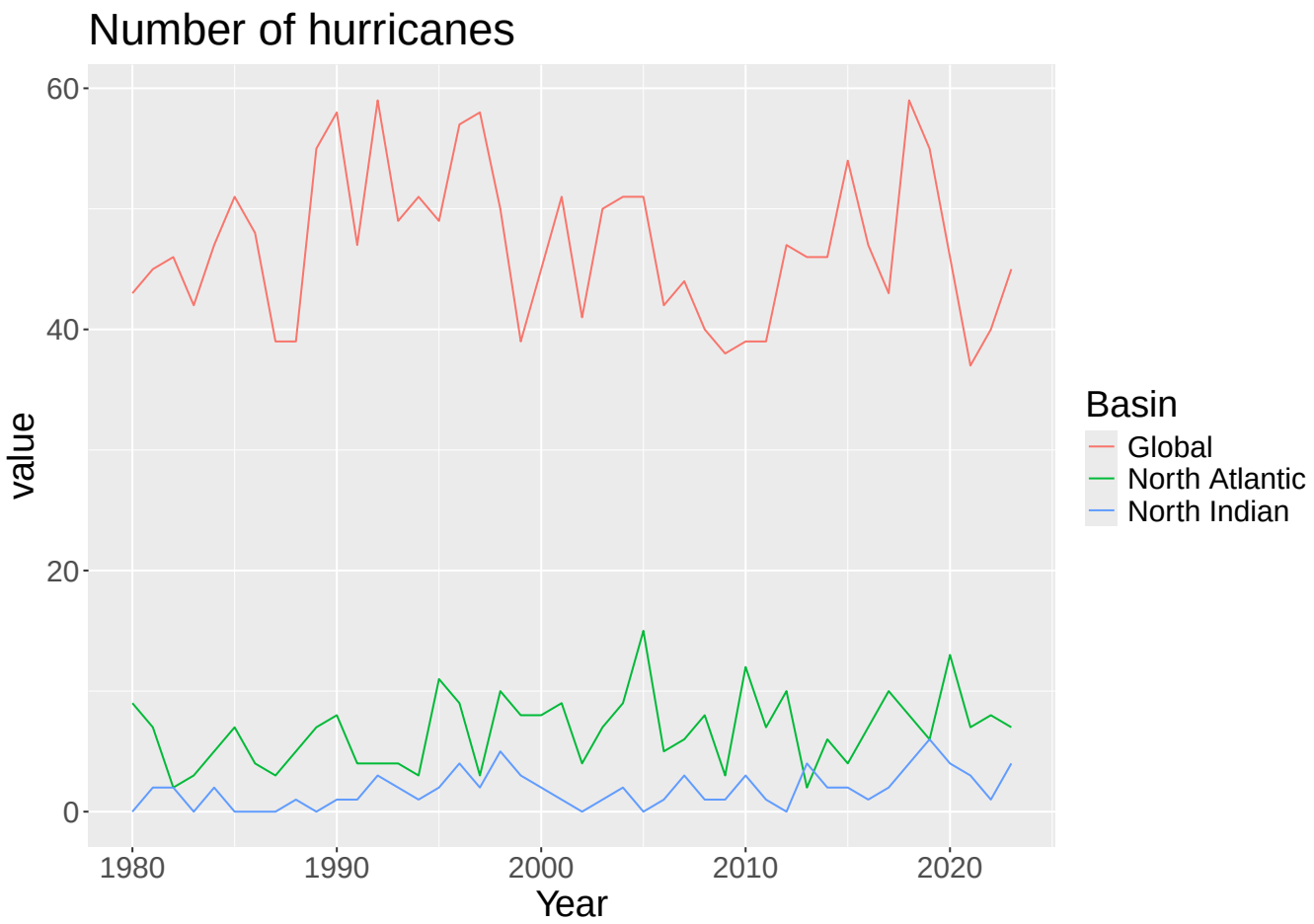
- Be clear about the intended purpose of a data visualization (compare groups, highlight correlation, show evolution over time, . . .)
- Choose the right chart according to the type of data and purpose of visualization: From Data to Viz



- Further general guidance can be found in Robbins' *Creating Better Graphs*

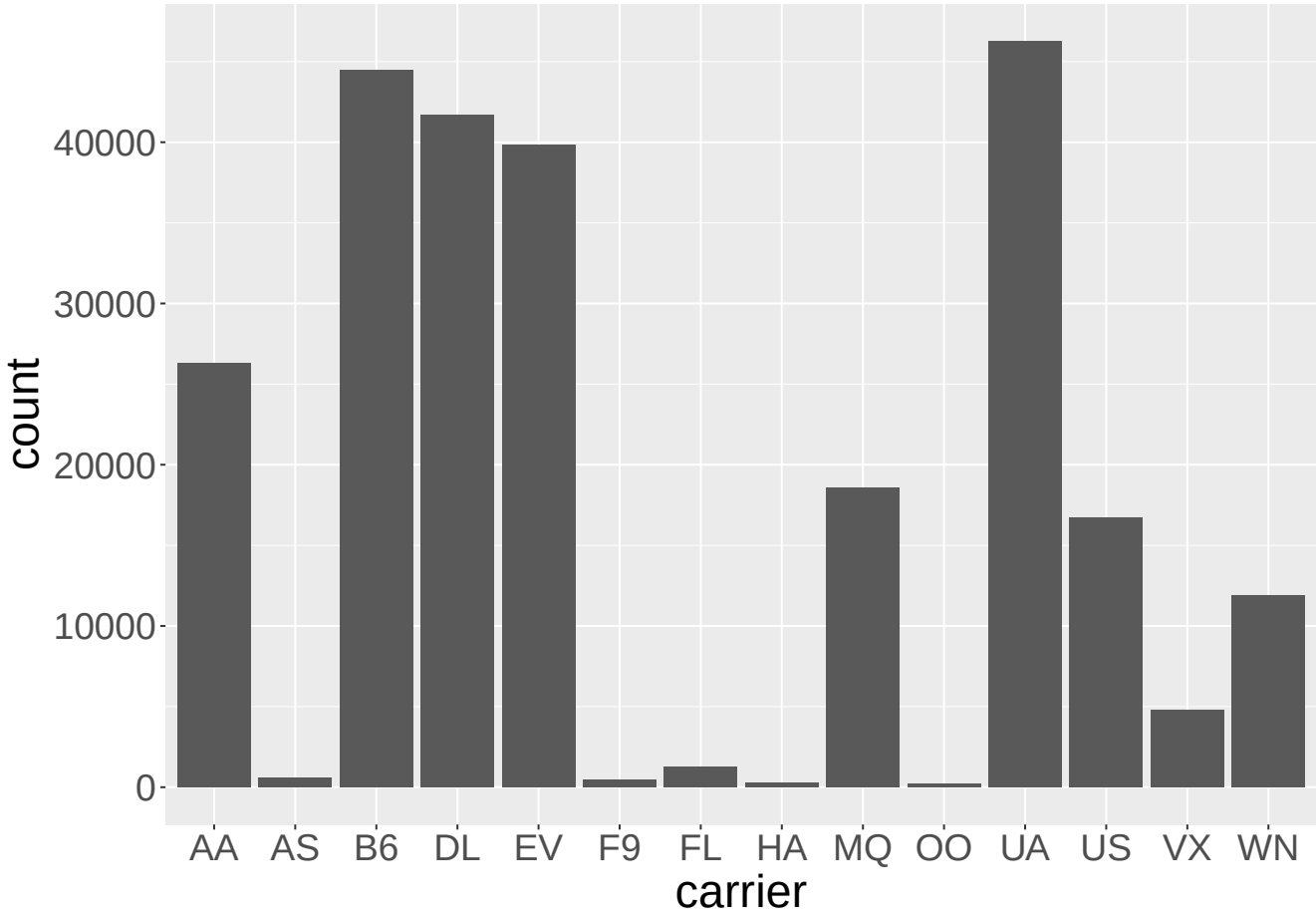
Line Plot

- there needs to be a linearly ordered variable, typically time
- if multiple groups with an inherent order, choose line types with an order (e.g., of thickness or dash density)



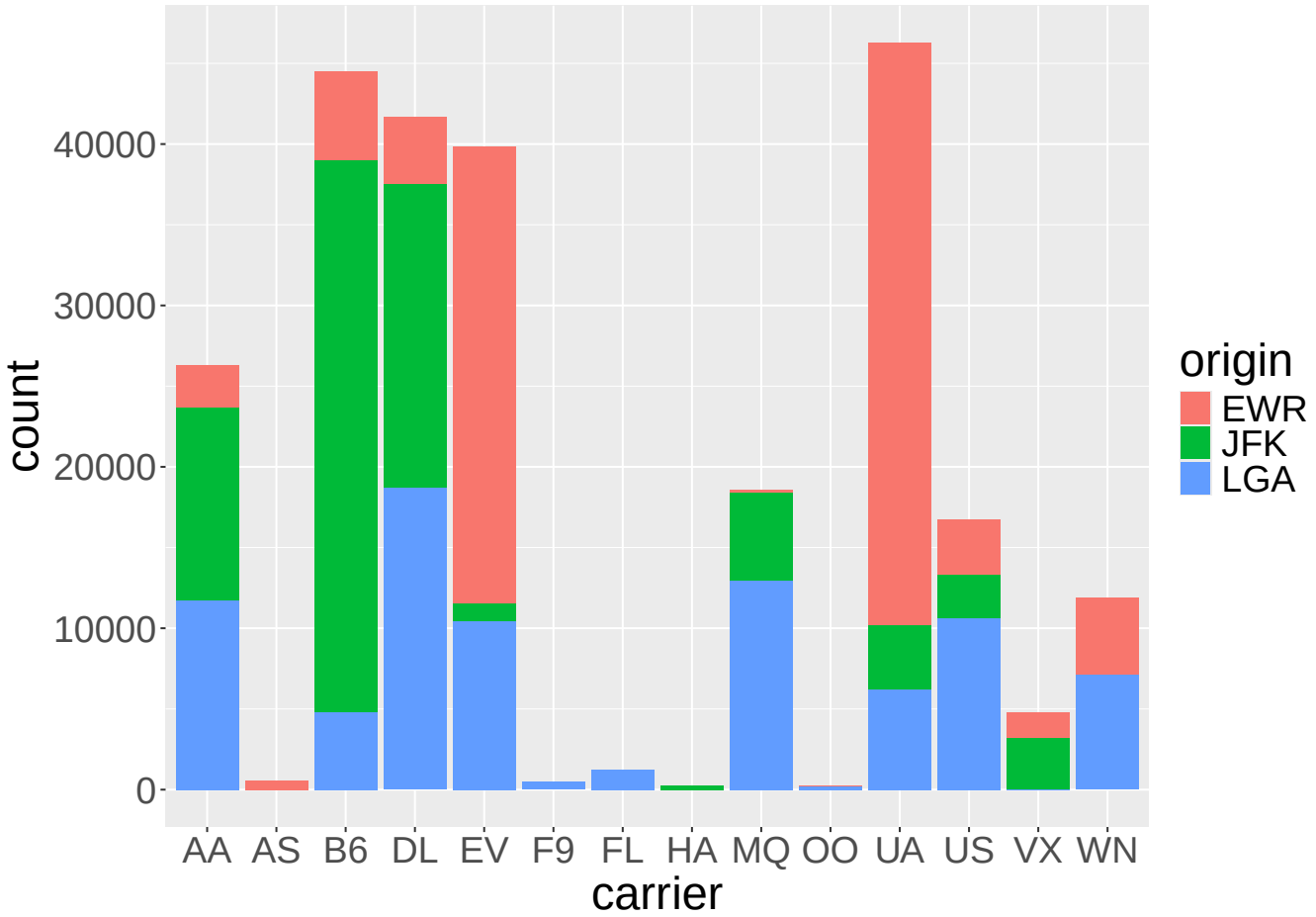
Bar Plot

- shows the relationship between a numeric and a categorical variable



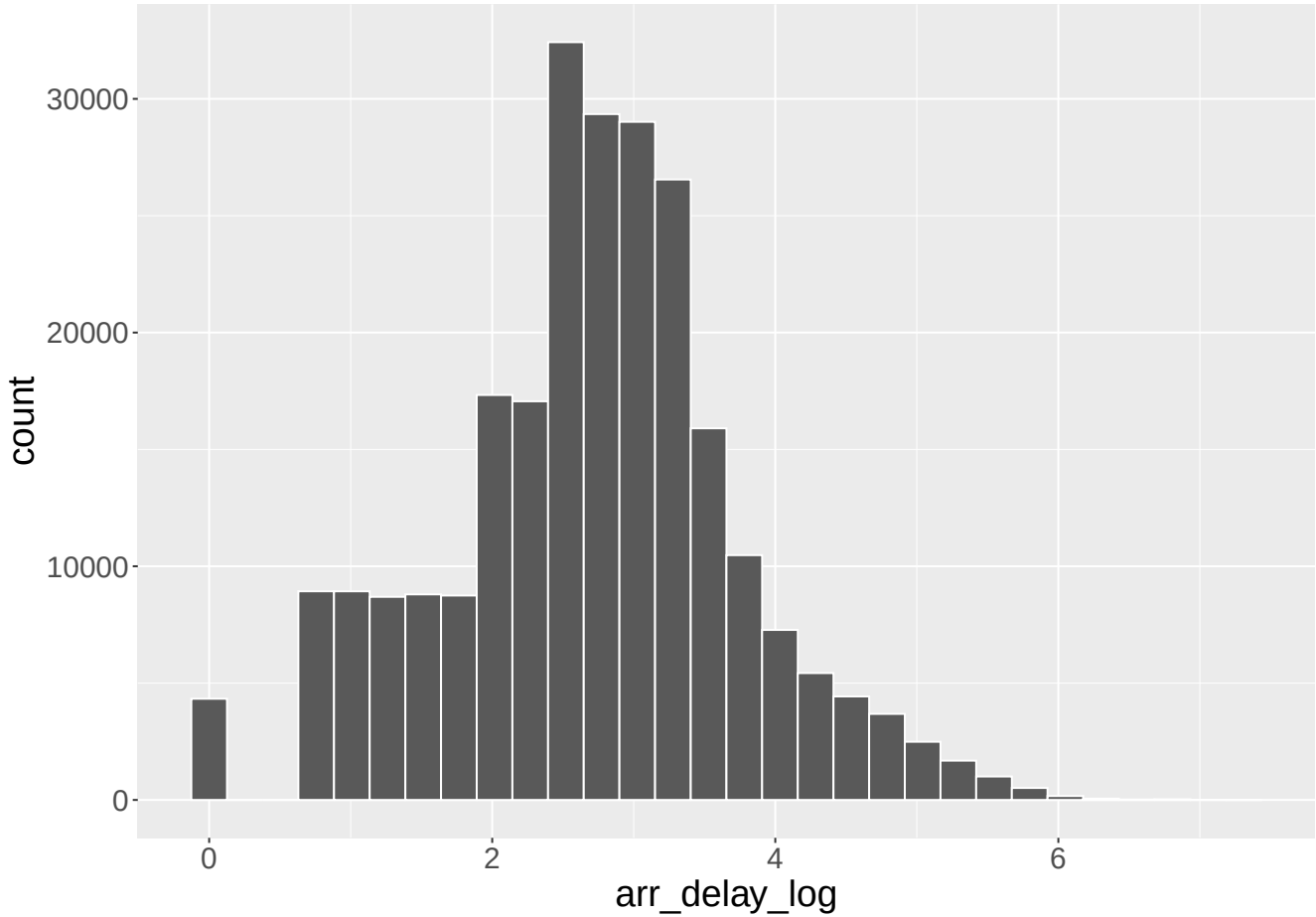
Bar Plot

- shows the relationship between a numeric and a categorical variable
- or displays values for several grouping levels



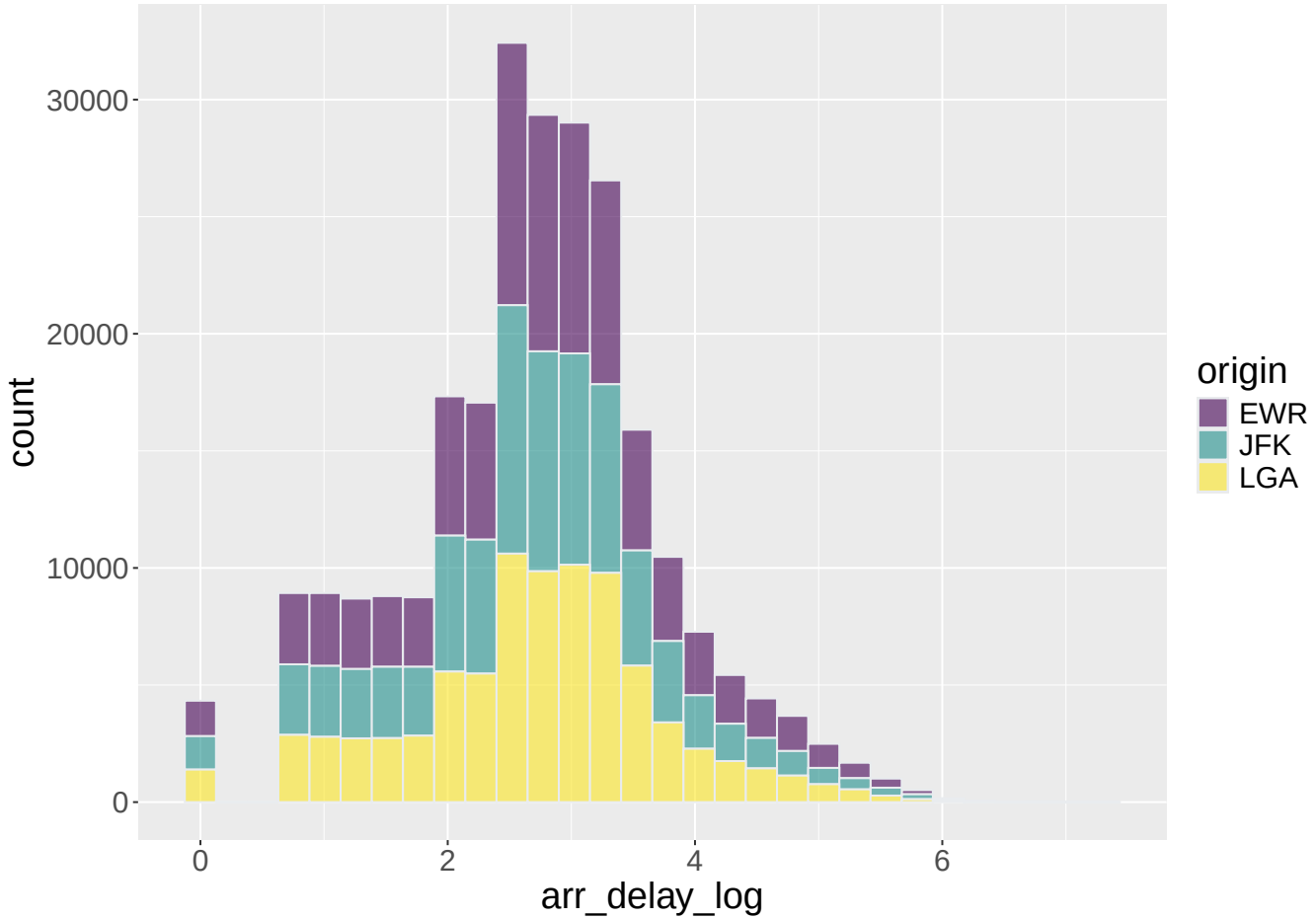
Histogram

- the variable is cut into several bins, and the number of observations per bin is represented by the height of the bar
- graphical representation of the distribution of a numeric variable
- allows to check the distribution for mistakes, outliers, or extremes



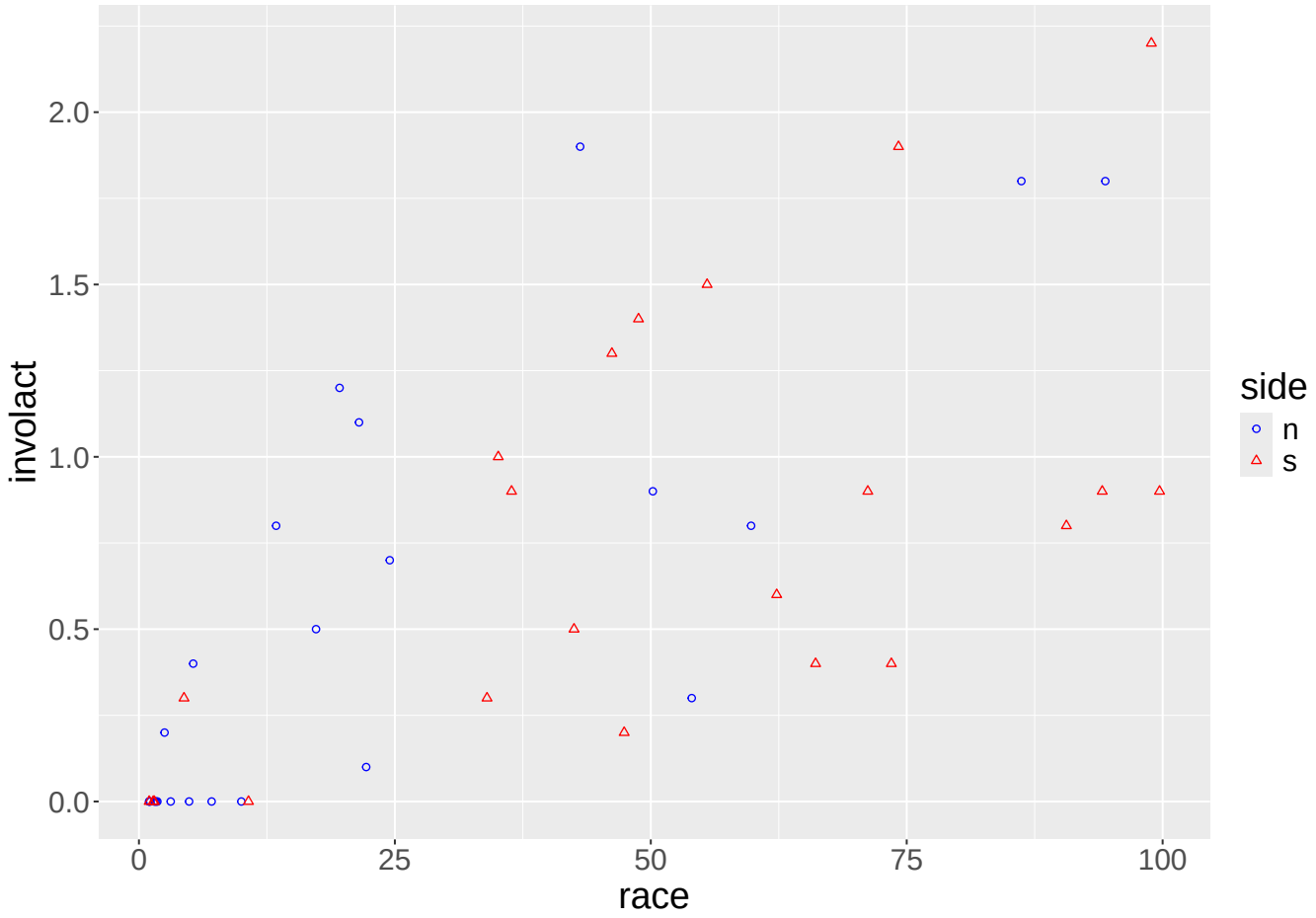
Histogram

- histogram allows to compare the distribution of several variables (not too many. . .)



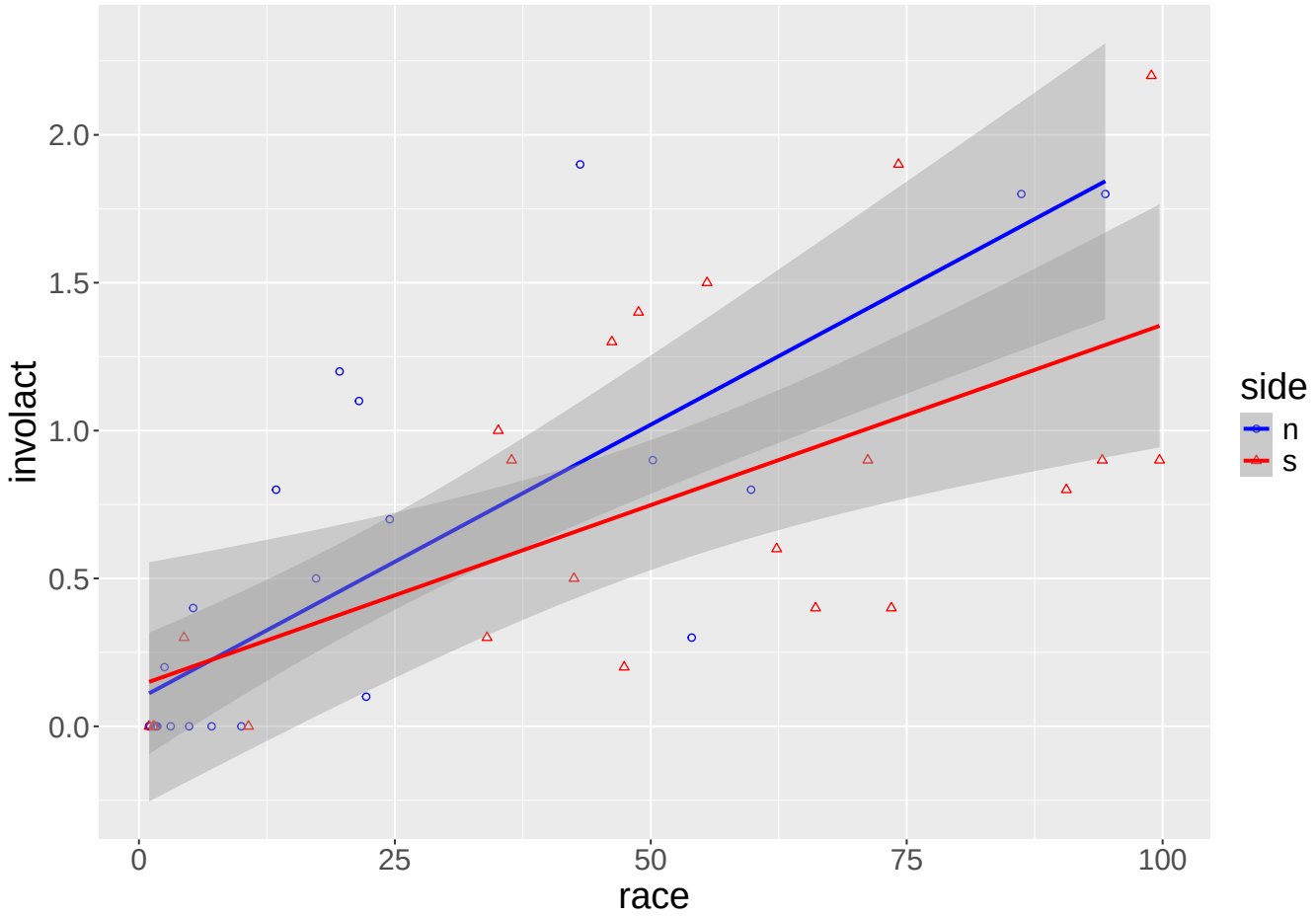
Scatterplot

- allows to study the relationship between two numeric variables
- subgroups can be added for more insights into hidden patterns

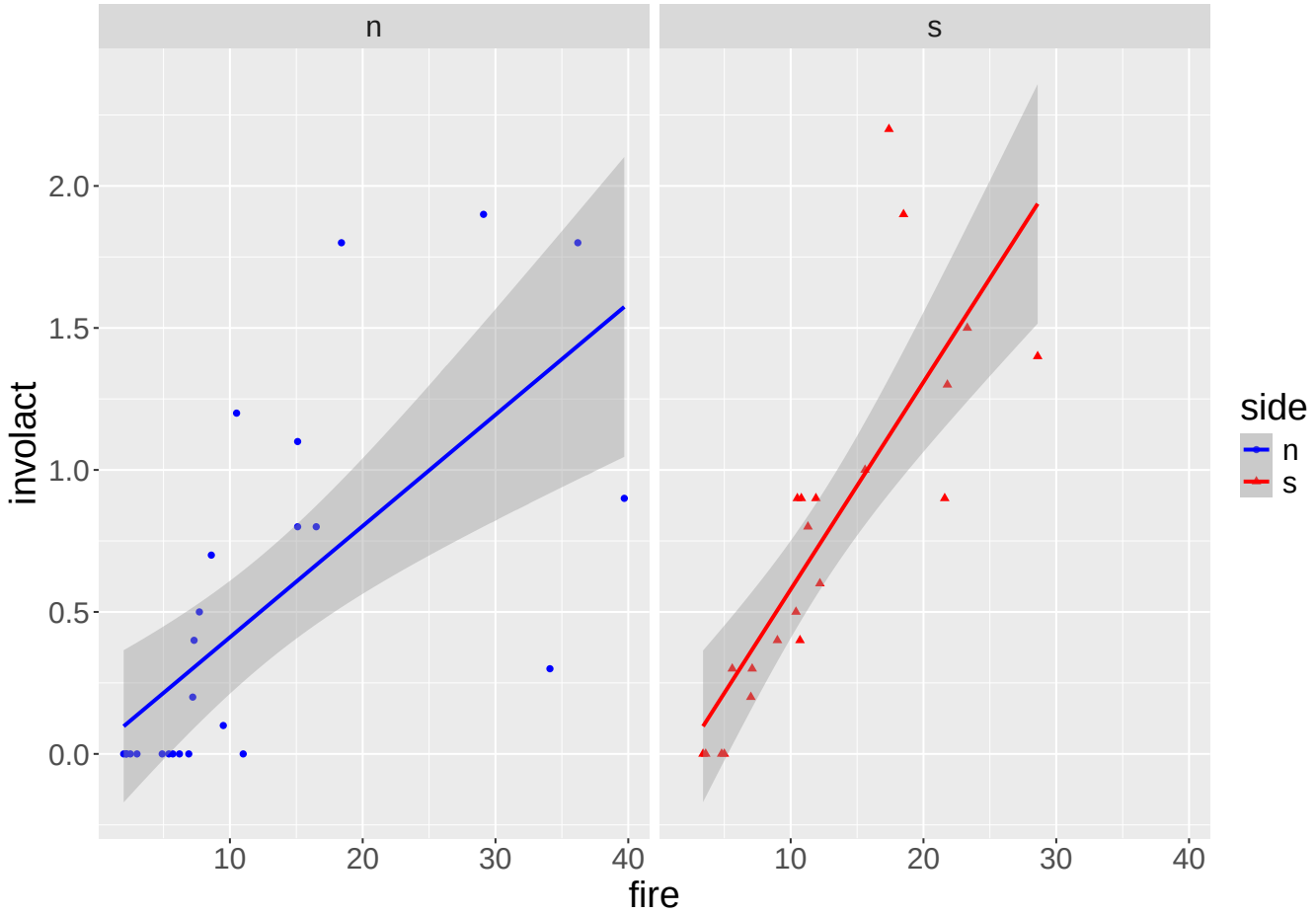


Scatterplot with Regression Lines

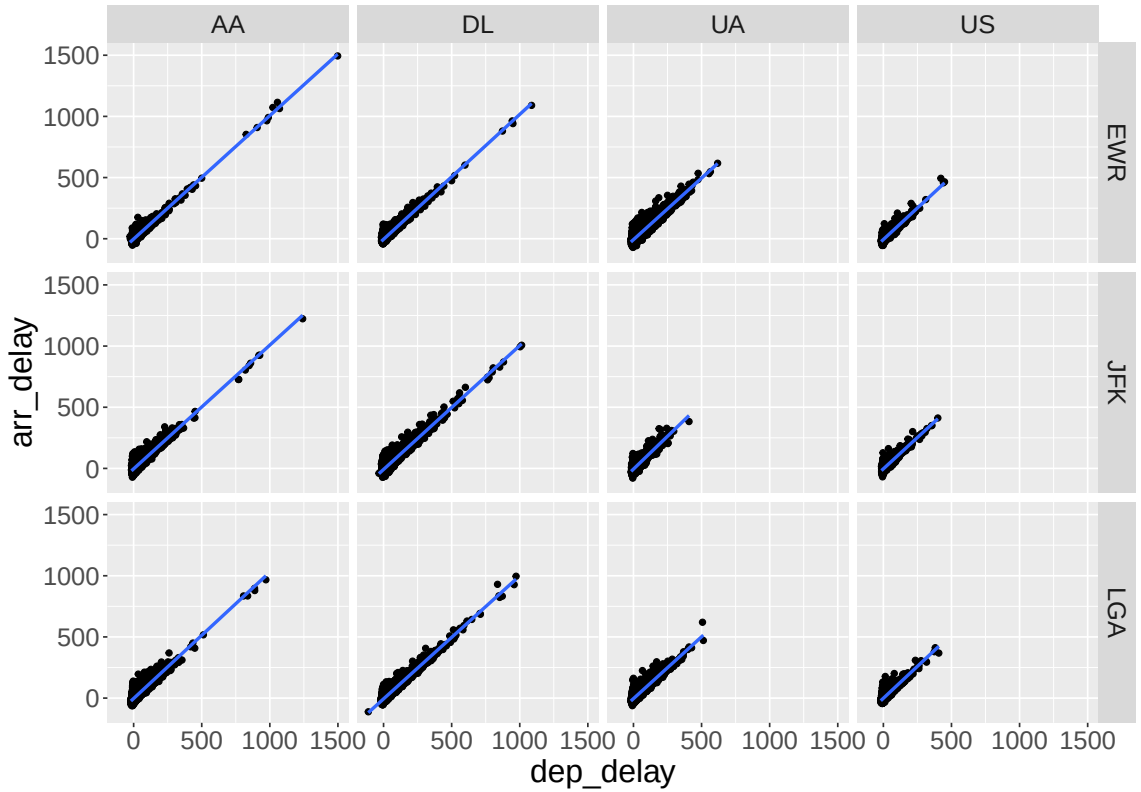
- different types of relationships can be detected by adding fitted regression curves



Plot by Factor

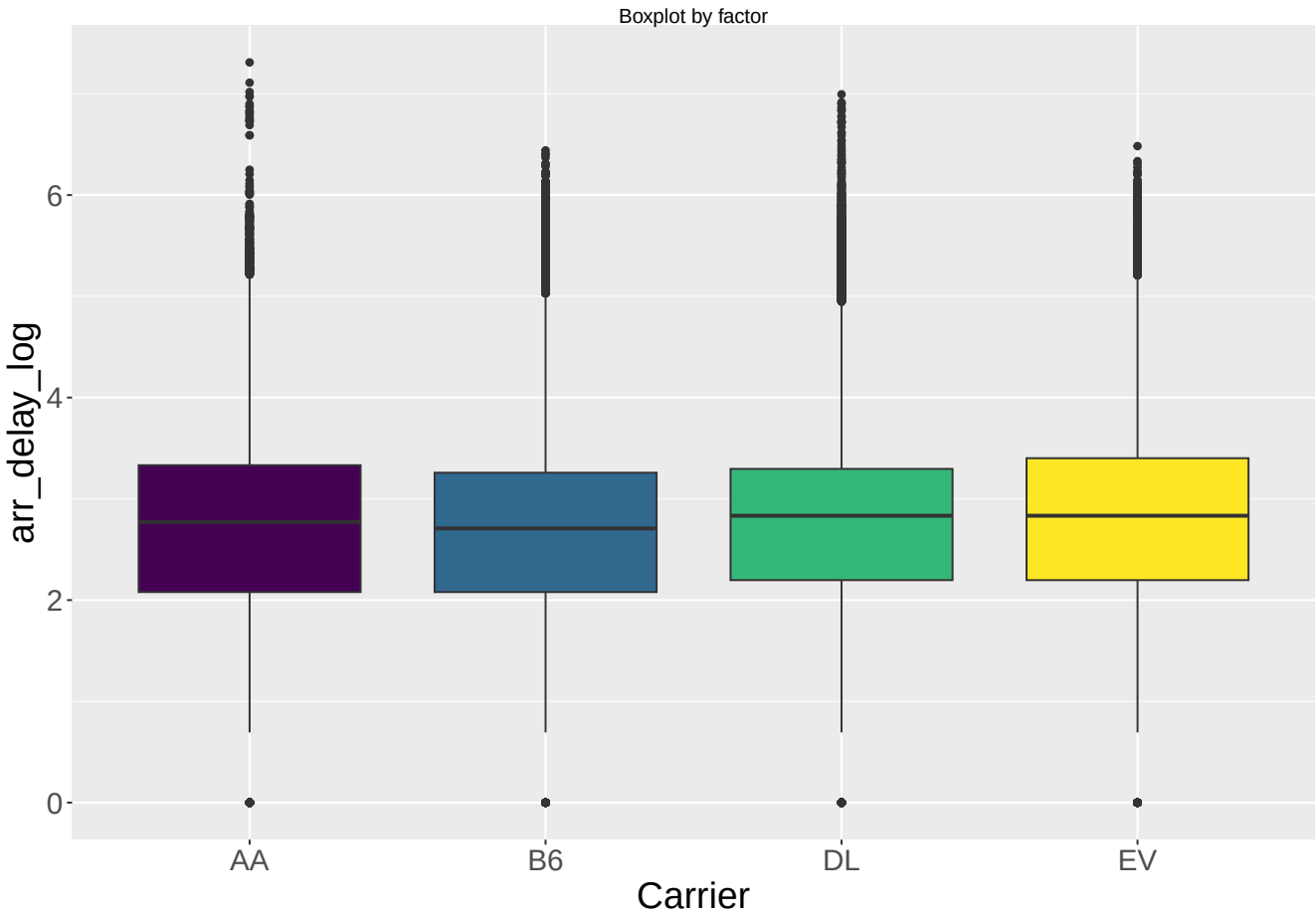


Plot by Two Factors

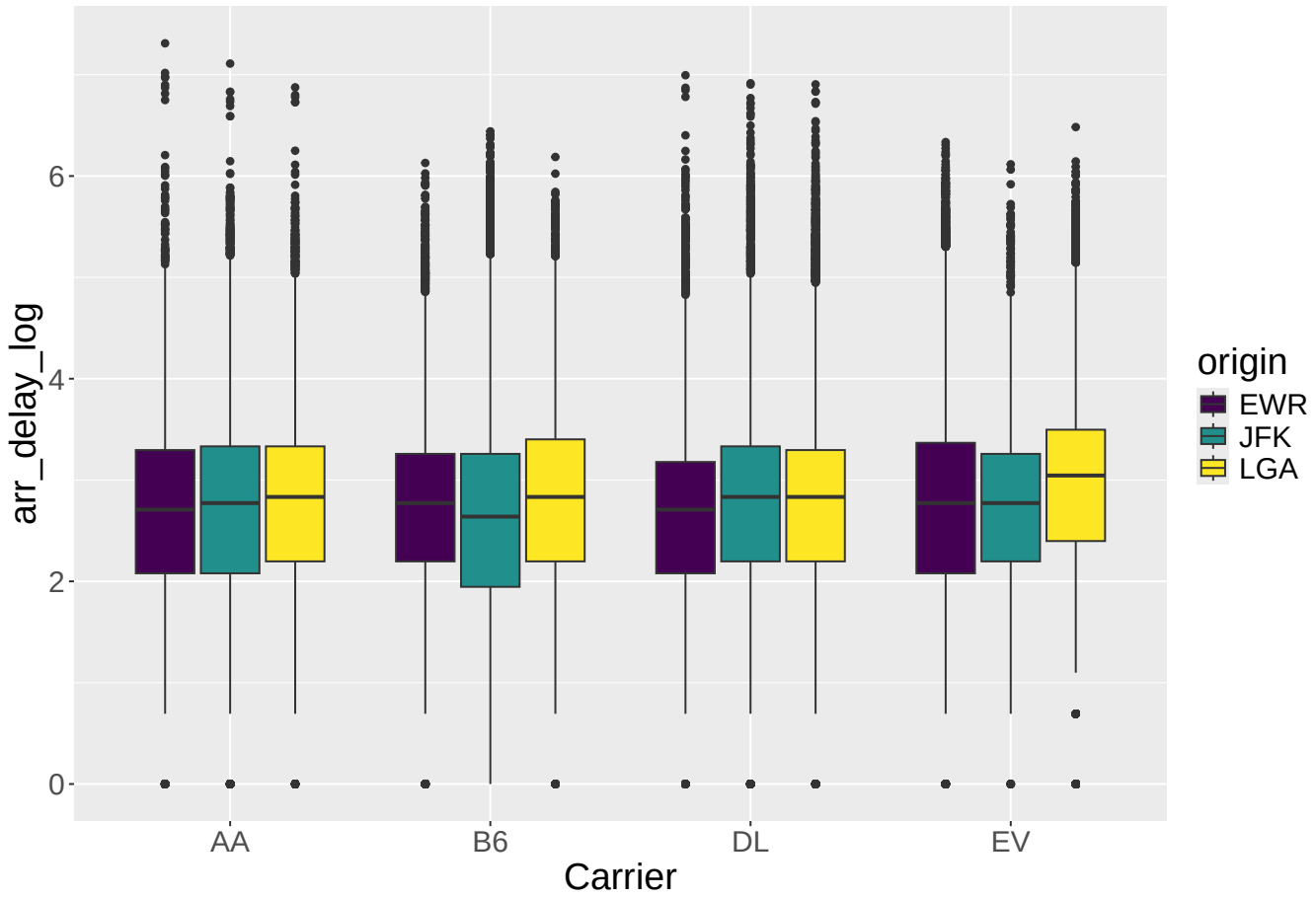


Boxplot

- summarizes the distribution of a numeric variable for several groups (presence of outliers, symmetry, dispersion, . . .)

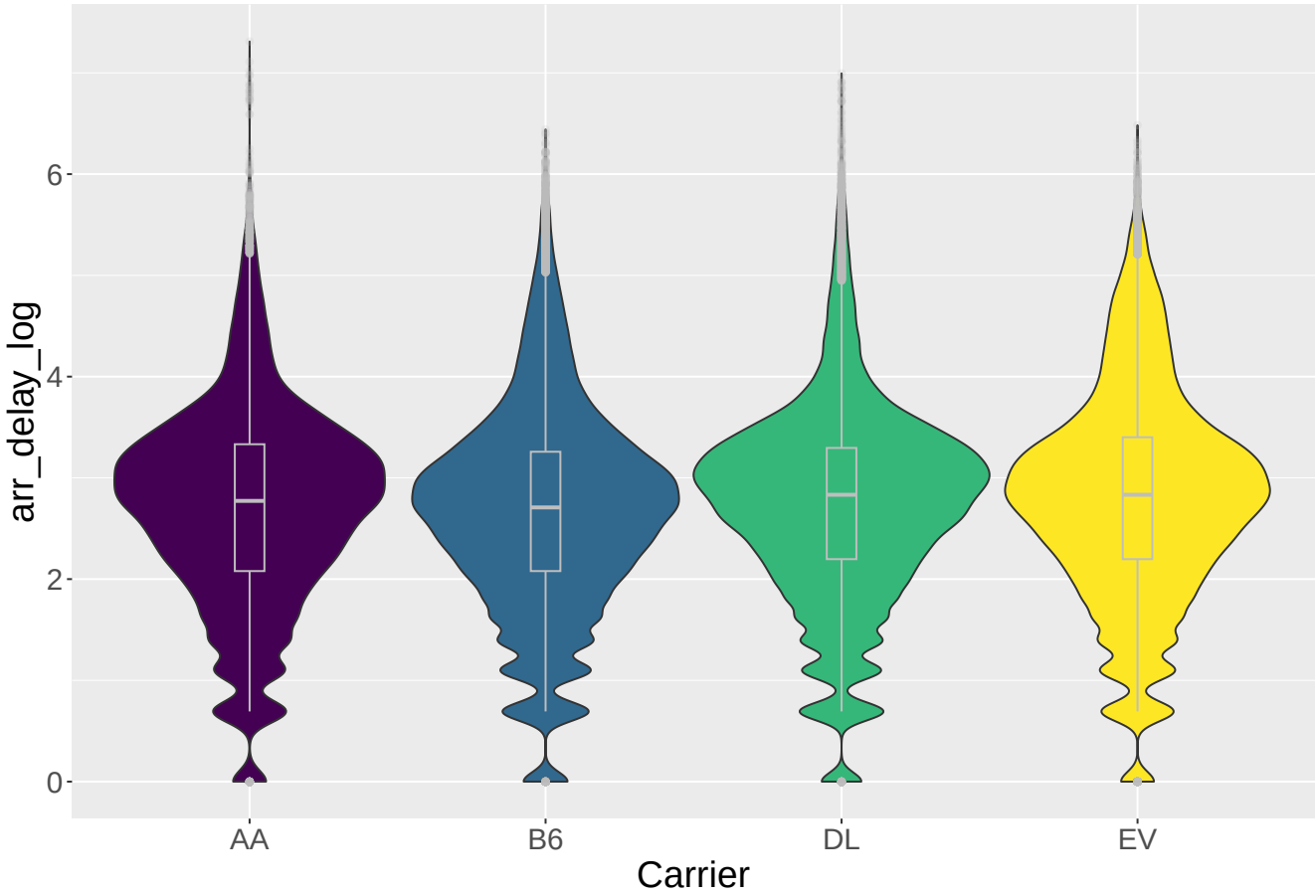


Boxplot by Two Factors



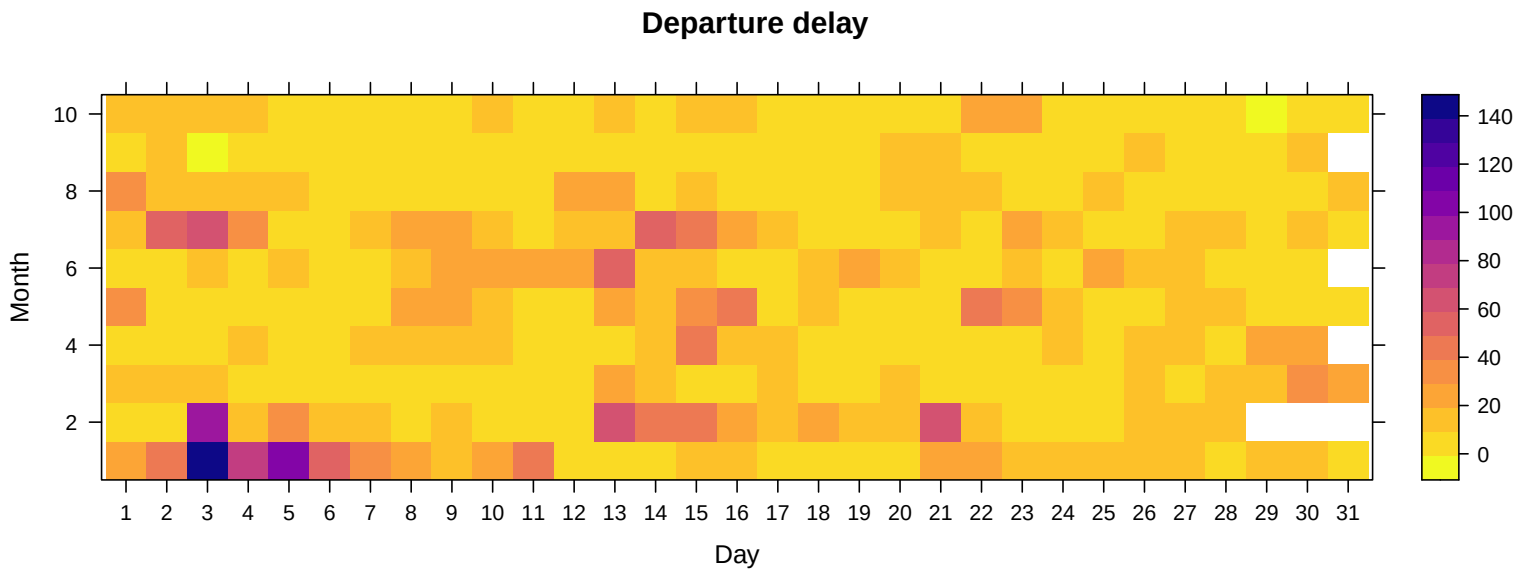
Improved version of a boxplot: The violin plot

- adds a representation of the distribution (a smooth version of the histogram)



Heatmaps

- represents a large matrix of data with colours reflecting their values (widely used for gene expression data)
- can be combined with dendrograms to display to show clustering, e.g., of genes



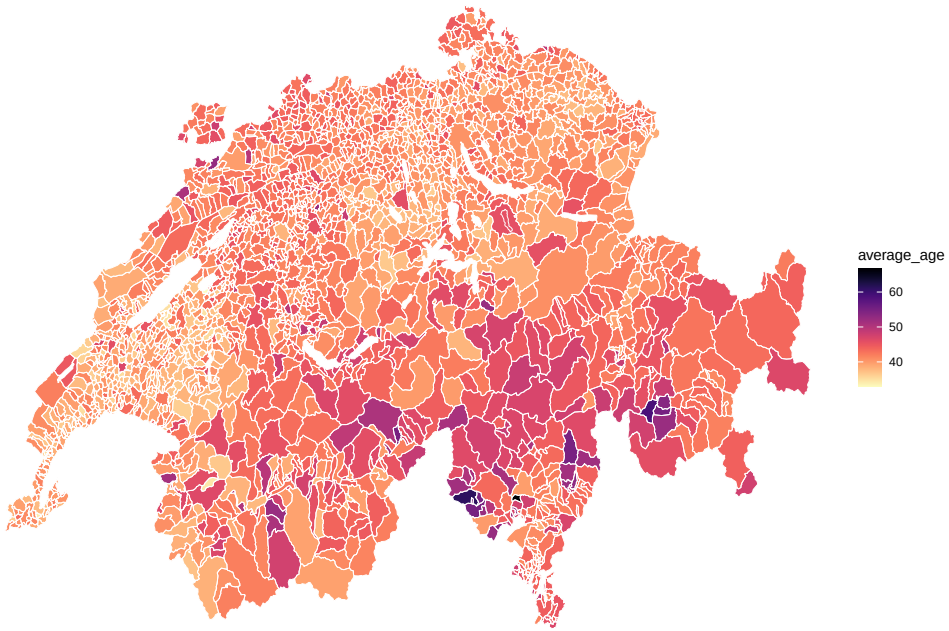
- could also be used for correlation matrices

Visualization of Spatial Data

Spatial data are complicated due to different

- data structures (vector or raster data)
- data sources
- data processing packages
- visualization packages

(Not so) short course about visualizing spatial data here (only if interested)

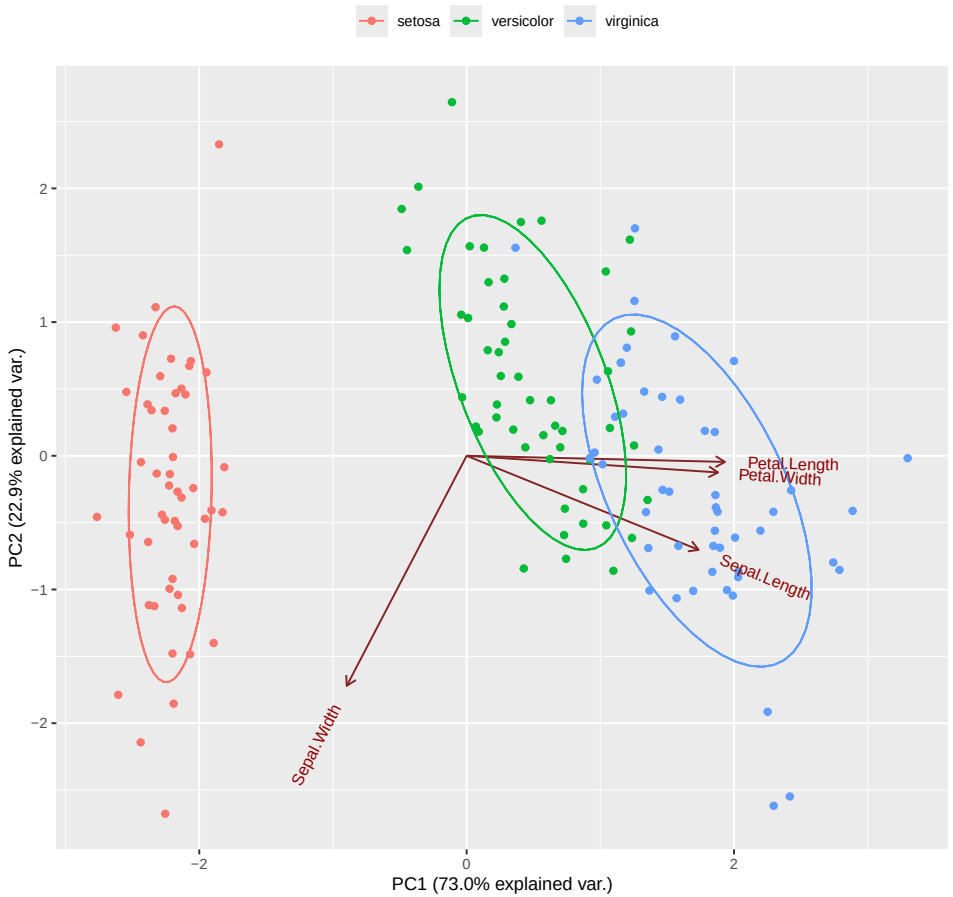


inspired by this blogpost

Visualization of High-dimensional Data

Exploration of high-dimensional data can be done using

- clustering methods: clusters are formed based on similarities between features (e.g., K-means, hierarchical clustering, . . .)
- PCA: dimension reduction technique that preserves most of the data variability



Good Visualization Practices

- context: exploratory vs. explanatory analysis/graphics
  - exploratory . . . helps you understand the patterns
  - explanatory . . . designed to communicate your understanding
- provide context (in text **and** in caption)
- take the audience into account: yourself? readers of your paper? listeners of your talk? what is their background?
- ask yourself if the graph
  - answers a/your question
  - achieves your aim (of convincing?)

Good Visualization Practices

- gray scale often preferable
  - color-blindness (friendly palettes, e.g. Coolors)
- axes (scale, gaps, etc.)
  - text of appropriate size
- publication-specific conditions
- seek simplicity and clarity, though you can be artistic!
  - sometimes bend the rules (responsibly and justifiably). As Nigel Holmes stated: "As long as the artist understands that the primary function is to convey statistics and respects that duty, then you can have fun (or be serious) with the image; that is, the form in which these statistics appear."
  - unusual complex graphs are more memorable
  - But, remember that visualization must be **accurate**, **easy to comprehend**, and **appropriate to the context**

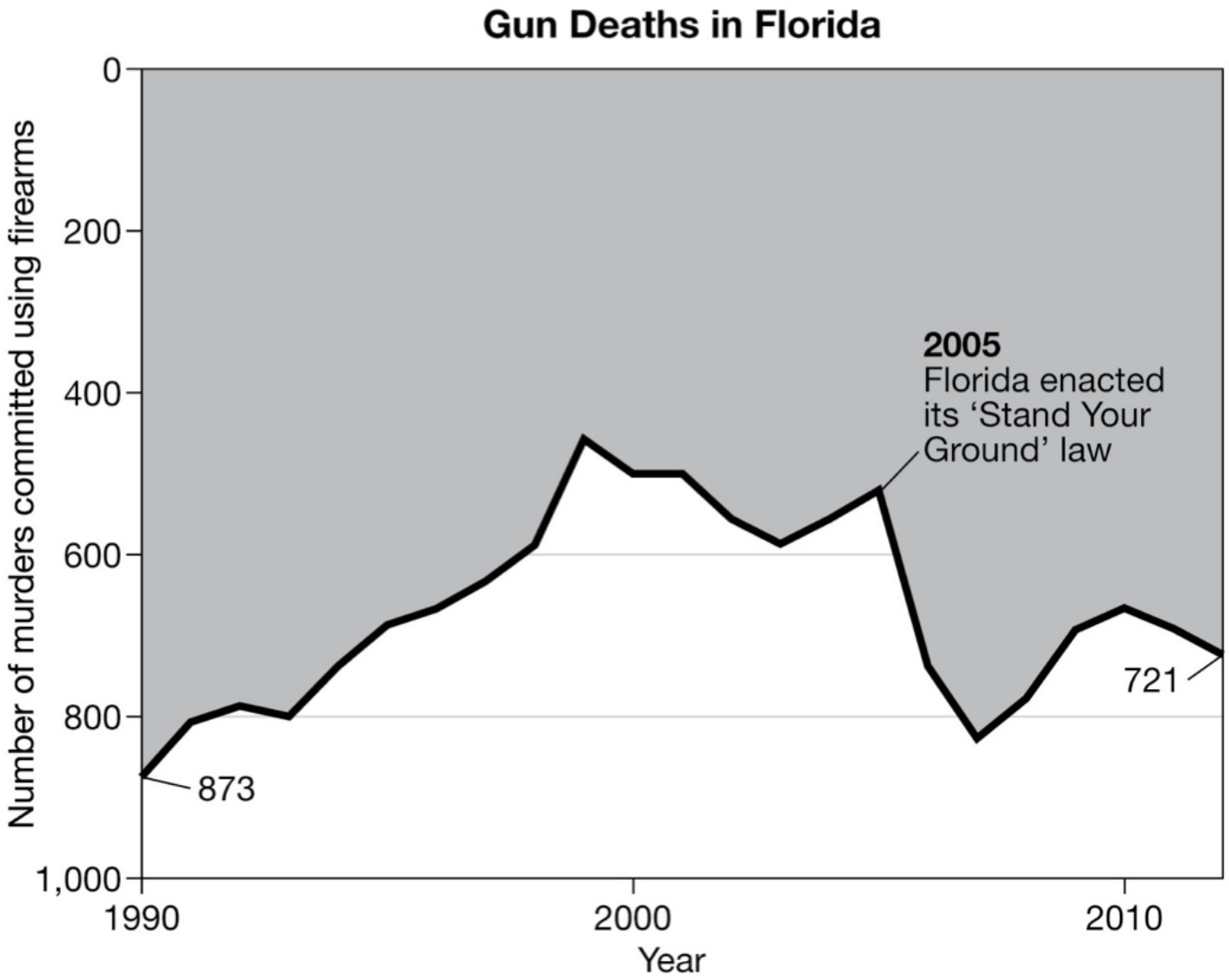
Find inspiration in The R Graph Gallery.

Beware when exporting graphics

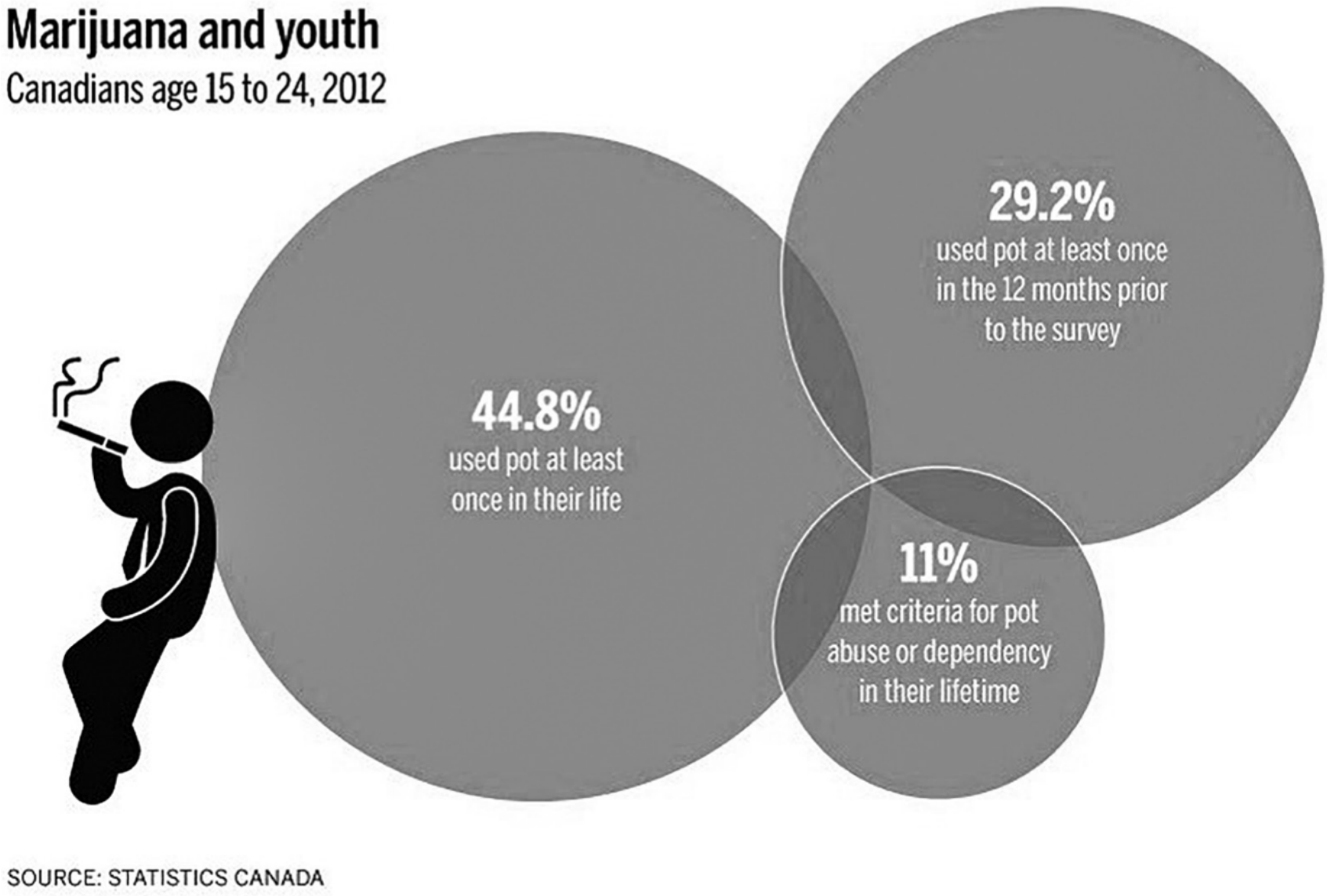
Section 3

Bad Visualization Practices

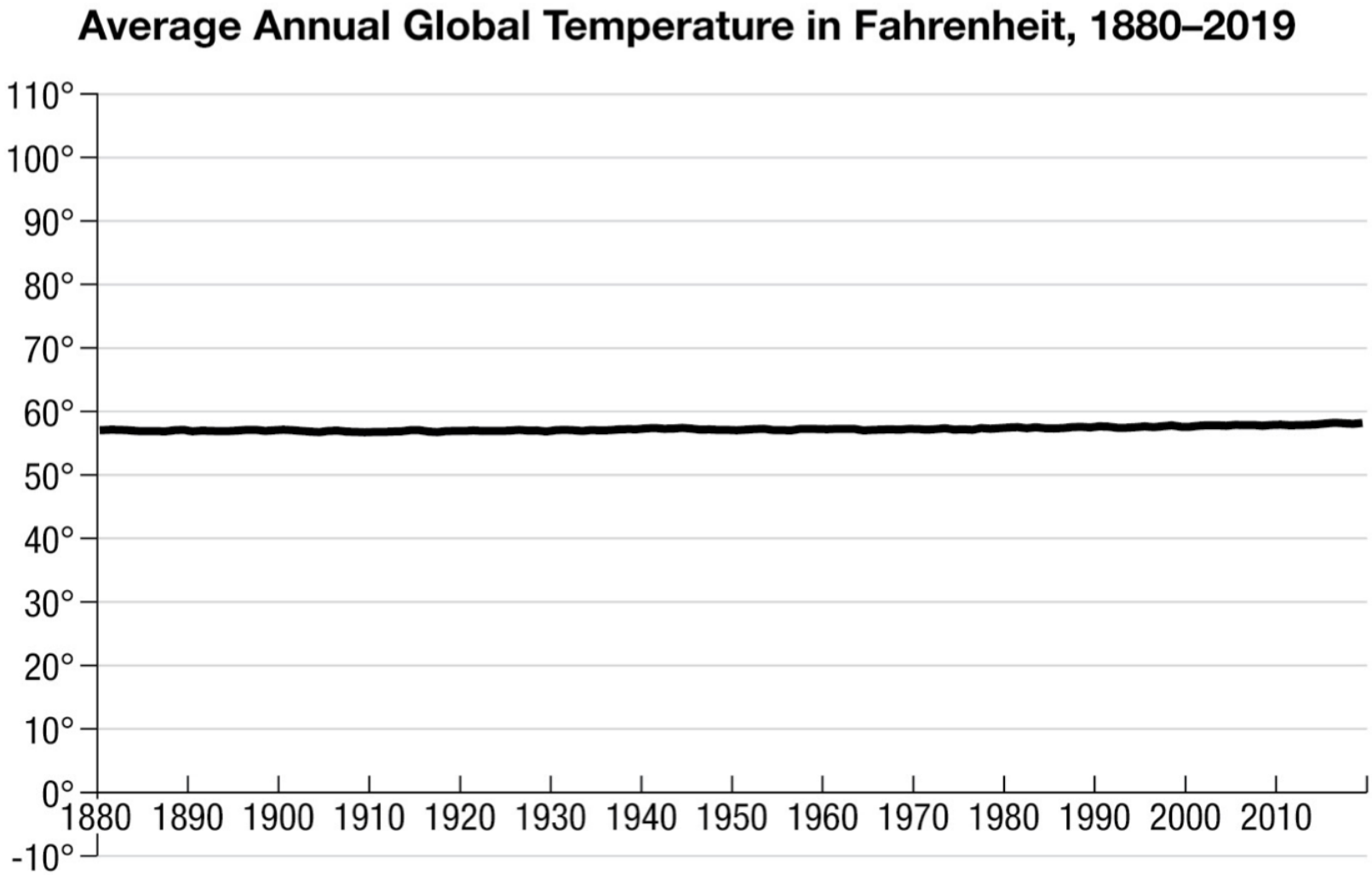
Reverted Axis



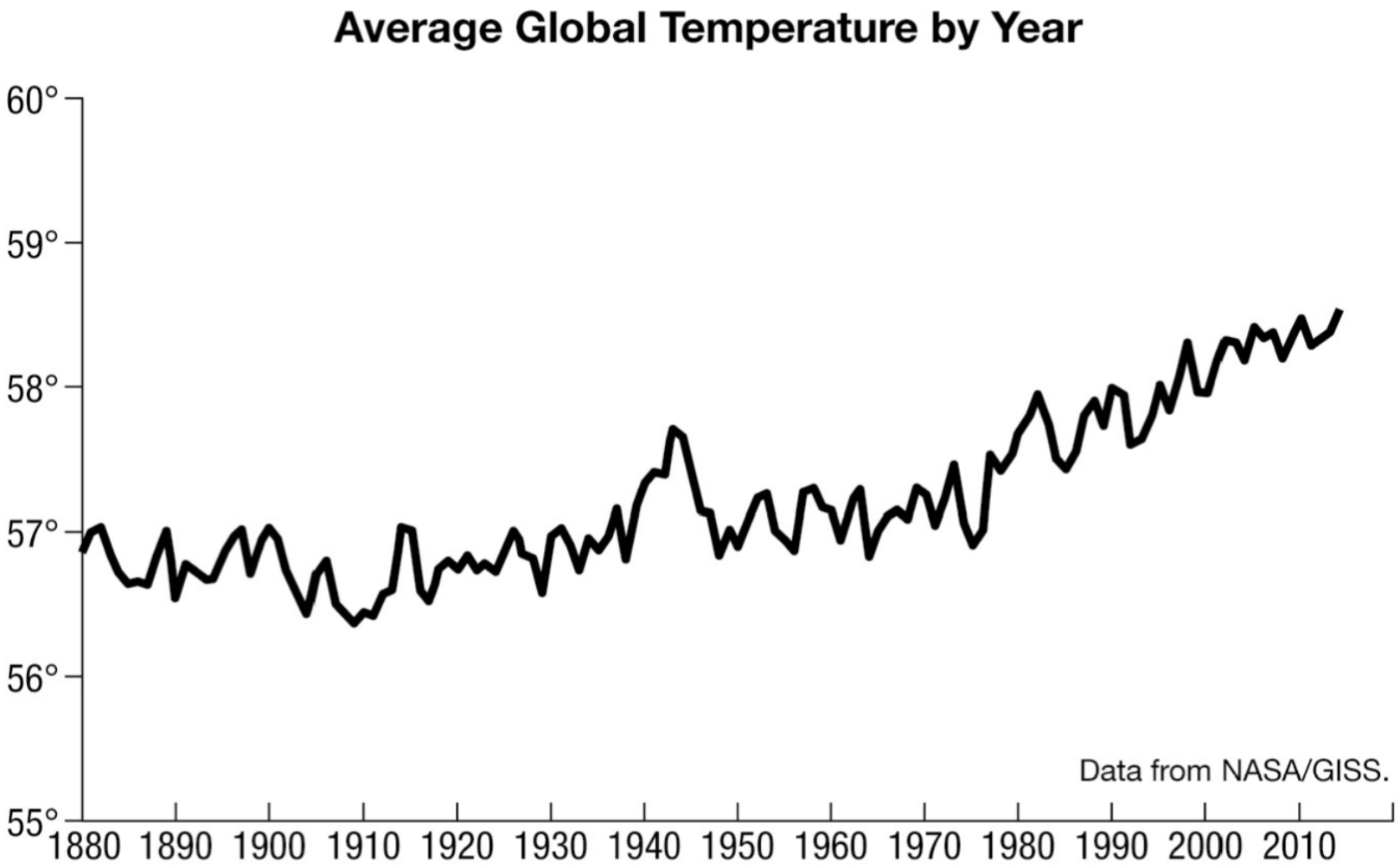
False Venn's Diagrams



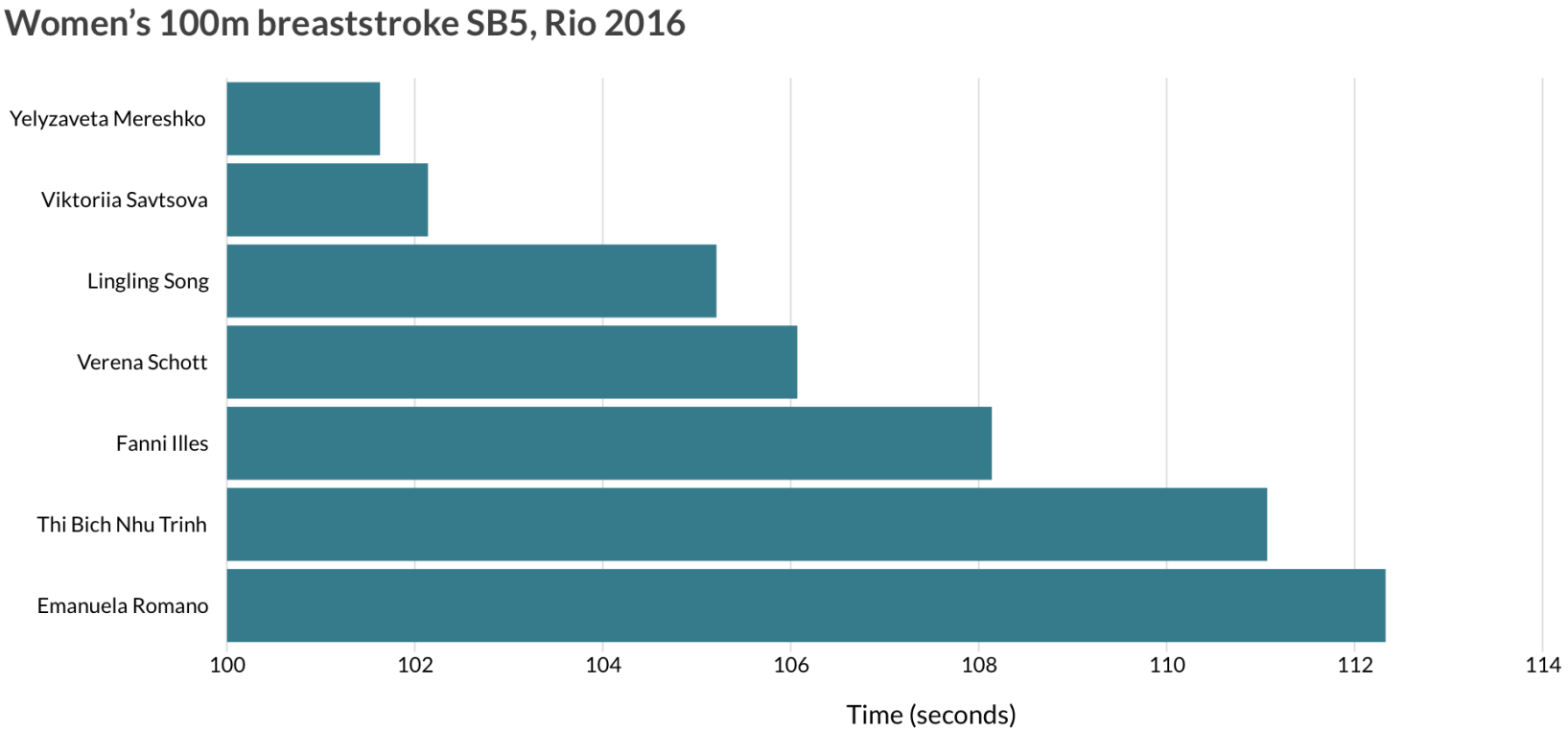
Axis Starting at Zero



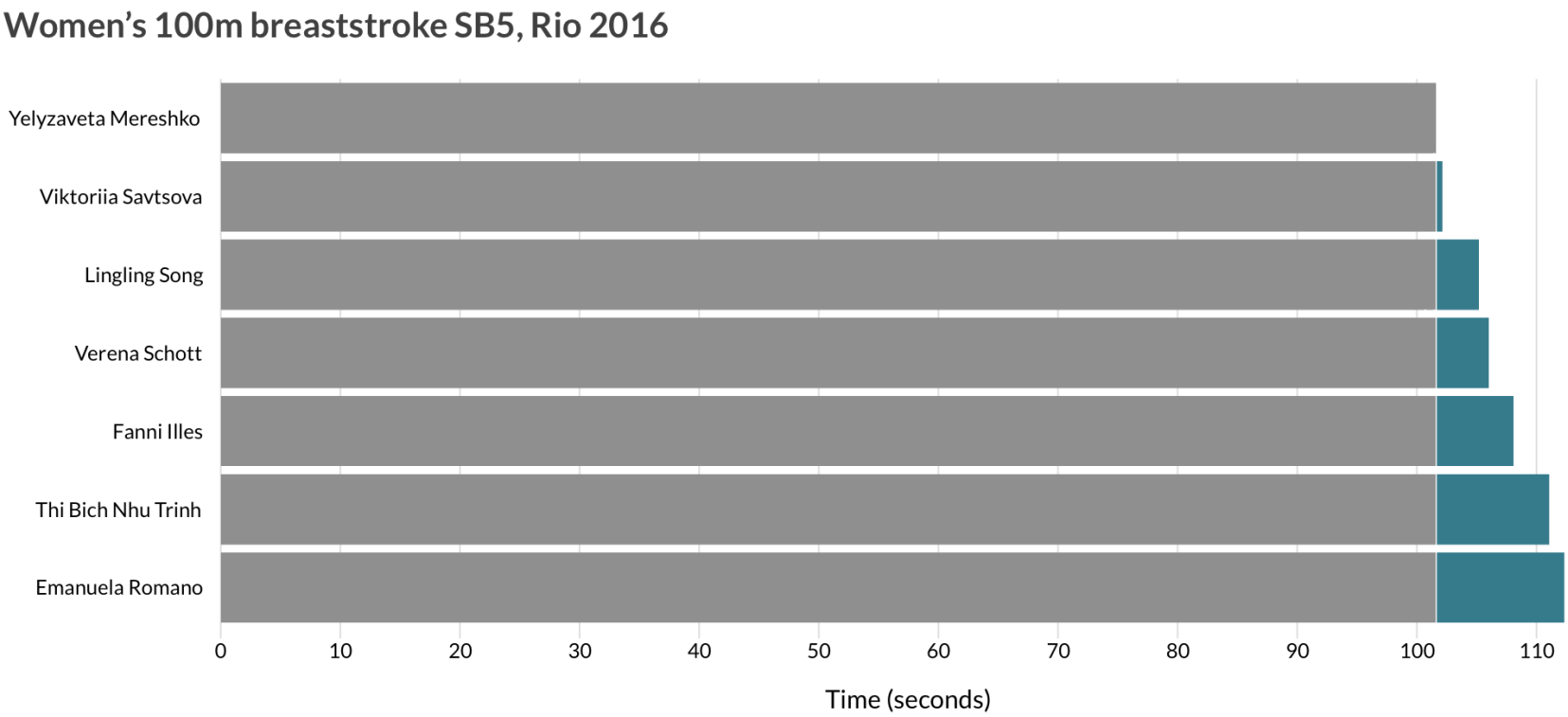
Better


Average Global Temperature by Year
60°
59°
58°
57°
56°
55°
1880 1890 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010
Data from NASA/GISS.

Over-exaggeration/ Missing Baseline



Better


Women's 100m breaststroke SB5, Rio 2016
Yelyzaveta Mereshko
Viktoriia Savtsova
Lingling Song
Verena Schott
Fanni Illes
Thi Bich Nhu Trinh
Emanuela Romano
0
10
20
30
40
50
60
70
80
90
100
110
Time (seconds)

To cut or not to cut?

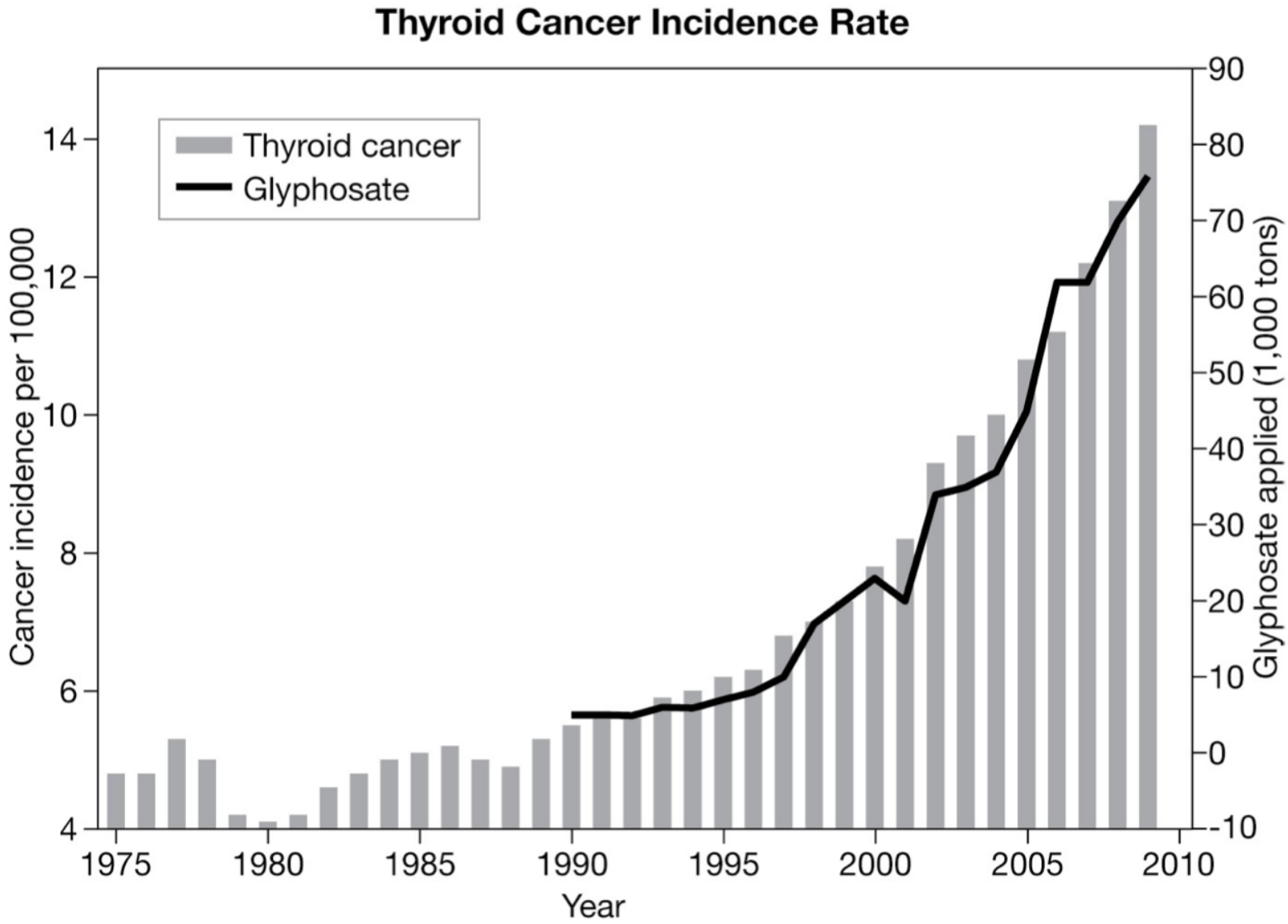
"In general, in a time-series, use a baseline that shows the data not the zero point" – *Edward Tufte*

"don't spend a lot of empty vertical space trying to reach down to the zero point at the cost of hiding what is going on in the data line itself" – *Edward Tufte*

So what?

- Barplot: With this kind of chart there is consensus: your Y-axis should start at 0
- Line plot: Here however there is no consensus, even if in general you don't have to start at 0

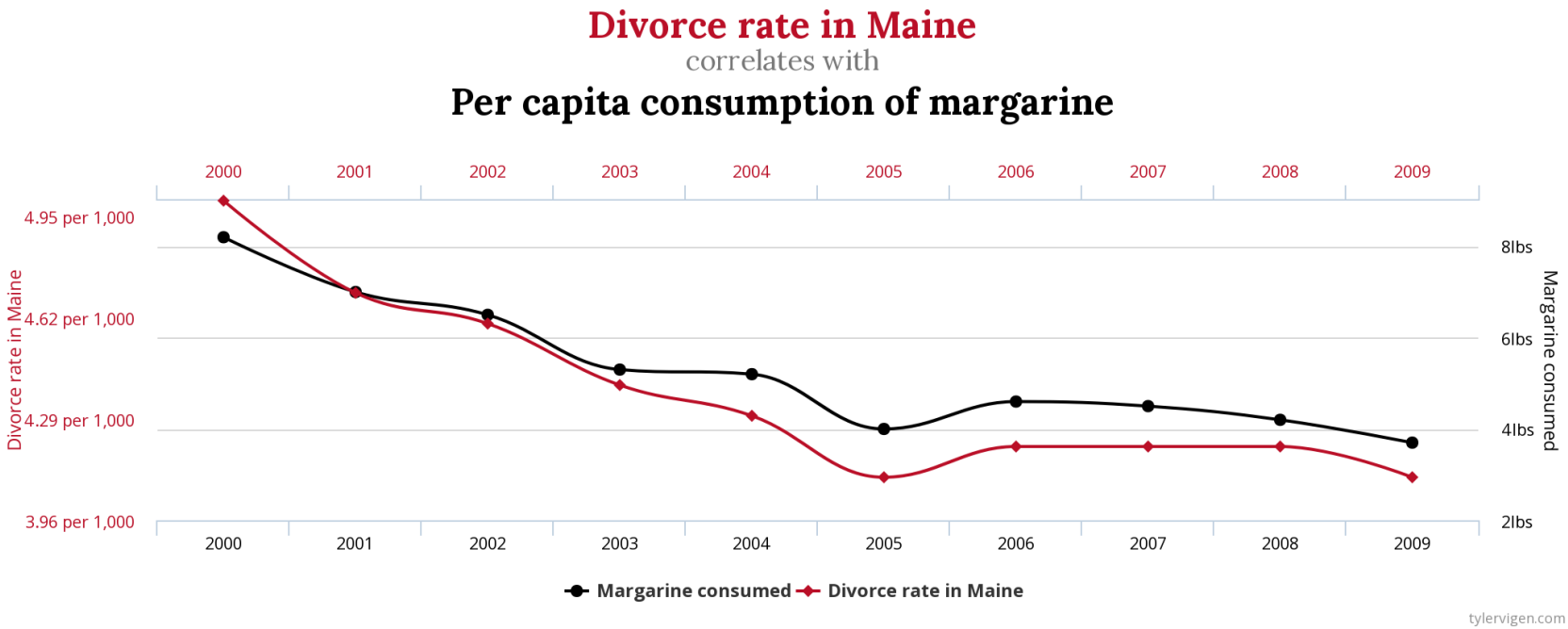
Double Axes



- This is actually quite good, but double axes are usually problematic

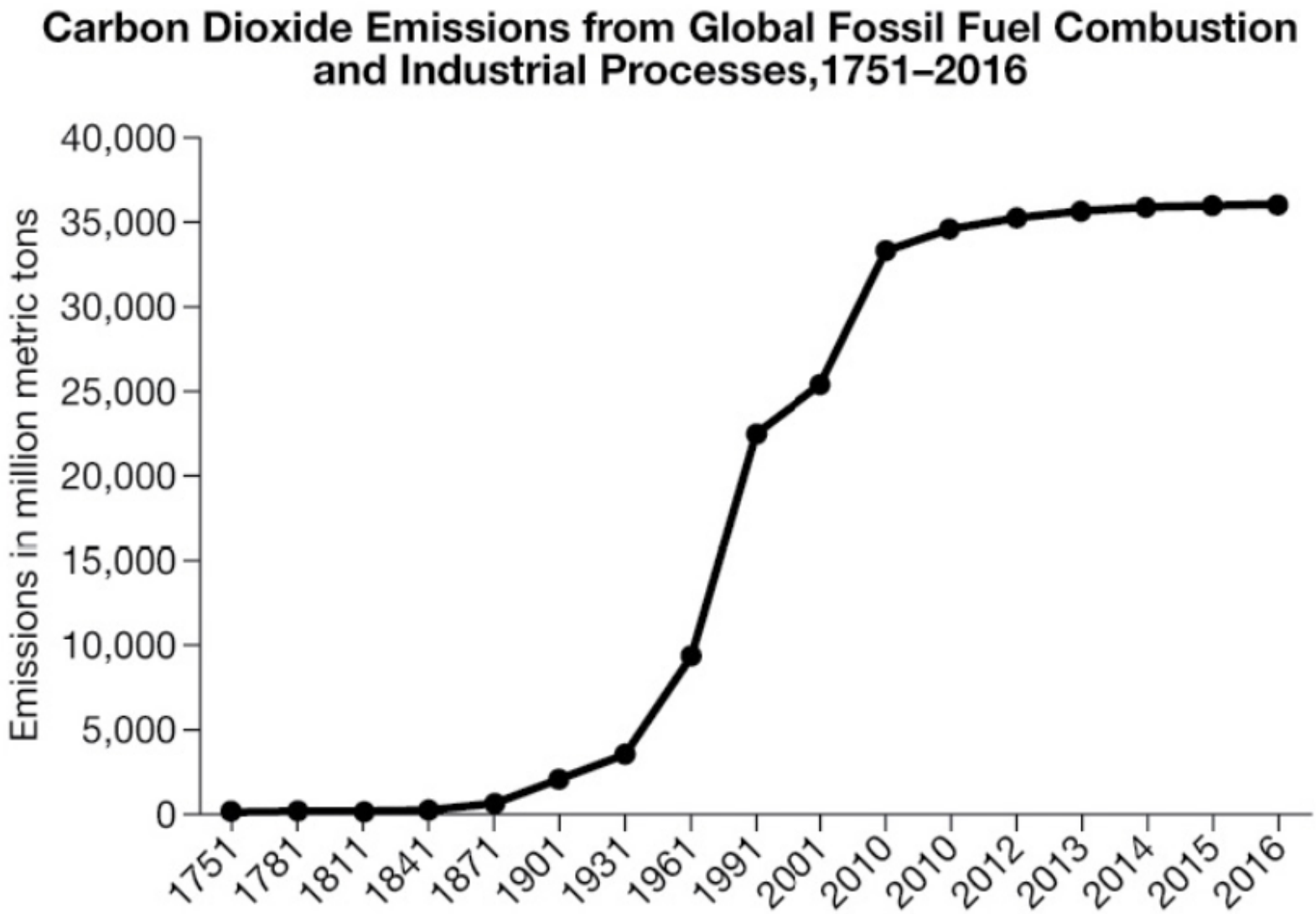
False Causation

Correlation does not imply causation

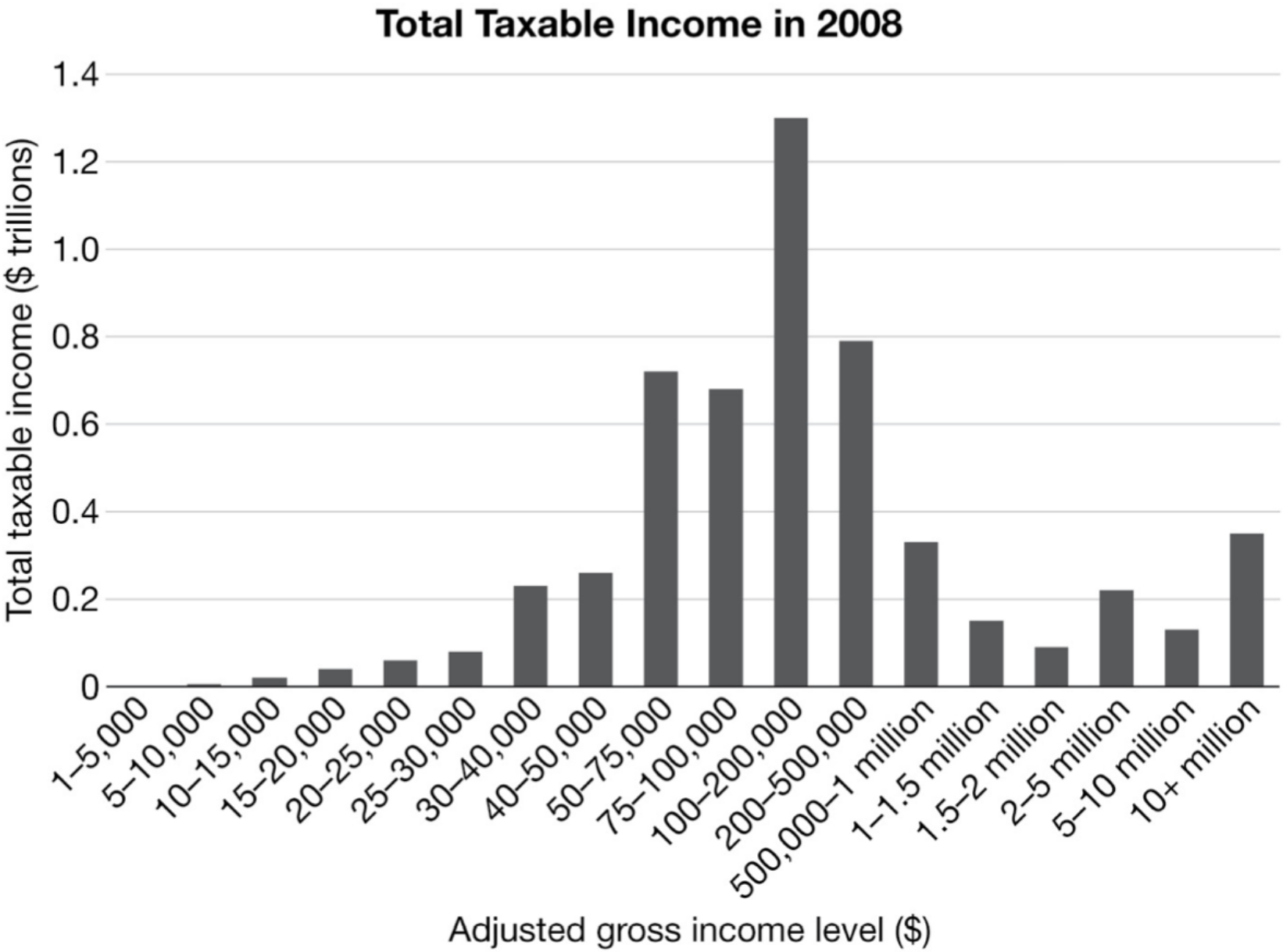


Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Tweaking Axis

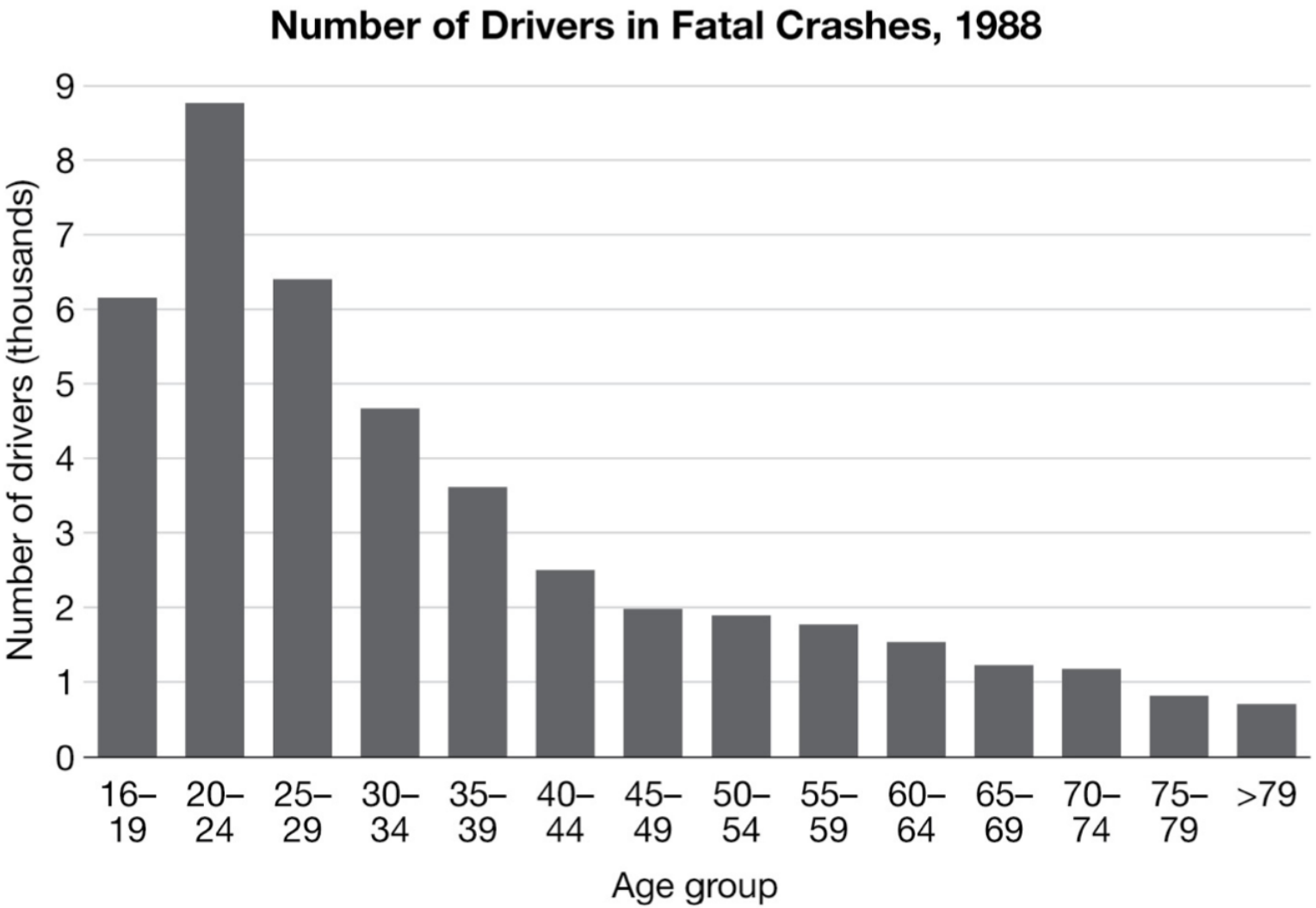


Binning

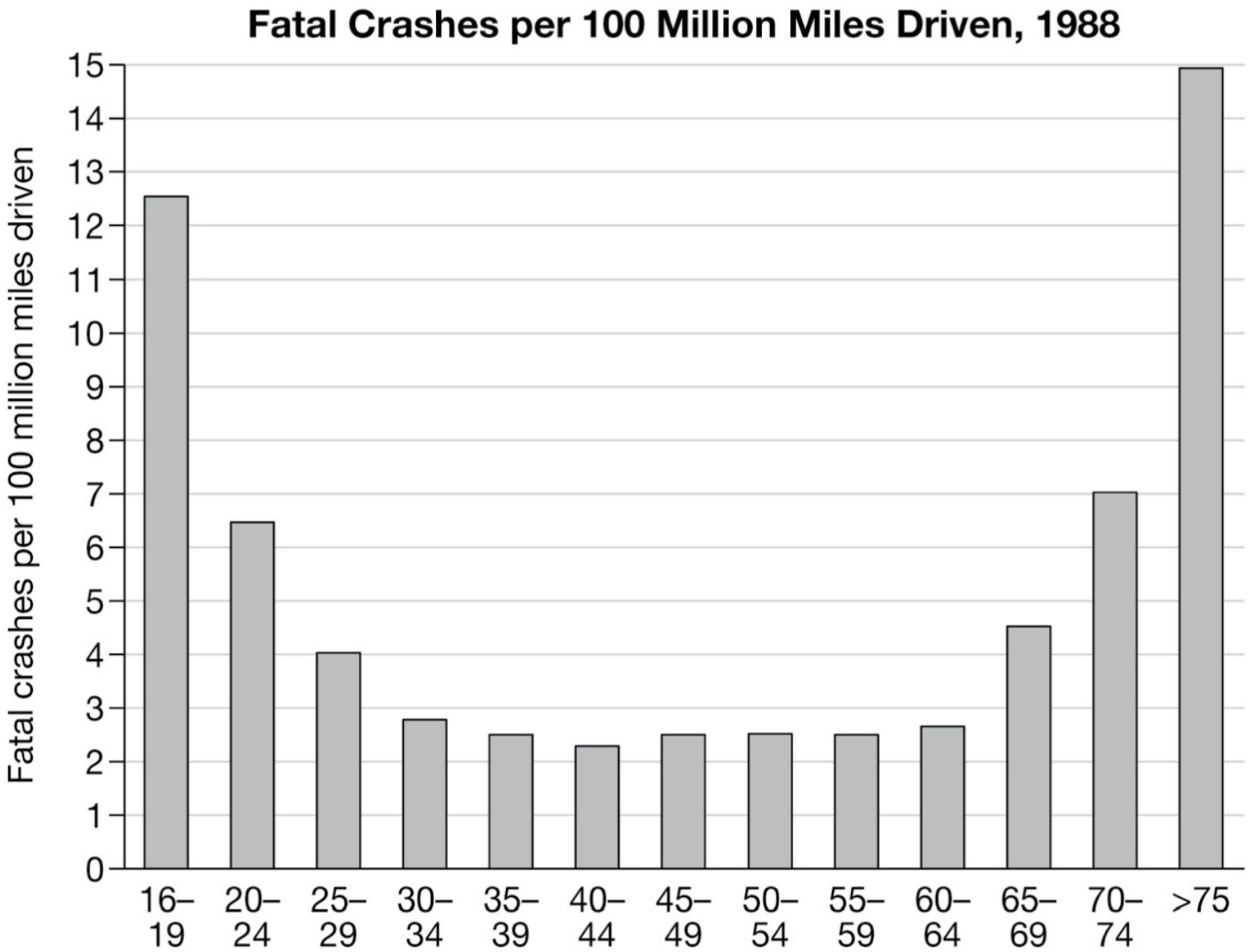

Total Taxable Income in 2008
Total taxable income ($ trillions)
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
1–5,000
5–10,000
10–15,000
15–20,000
20–25,000
25–30,000
30–40,000
40–50,000
50–75,000
75–100,000
100–200,000
200–500,000
500,000–1 million
1–1.5 million
1.5–2 million
2–5 million
5–10 million
10+ million
Adjusted gross income level ($)

Different Kinds of Binning

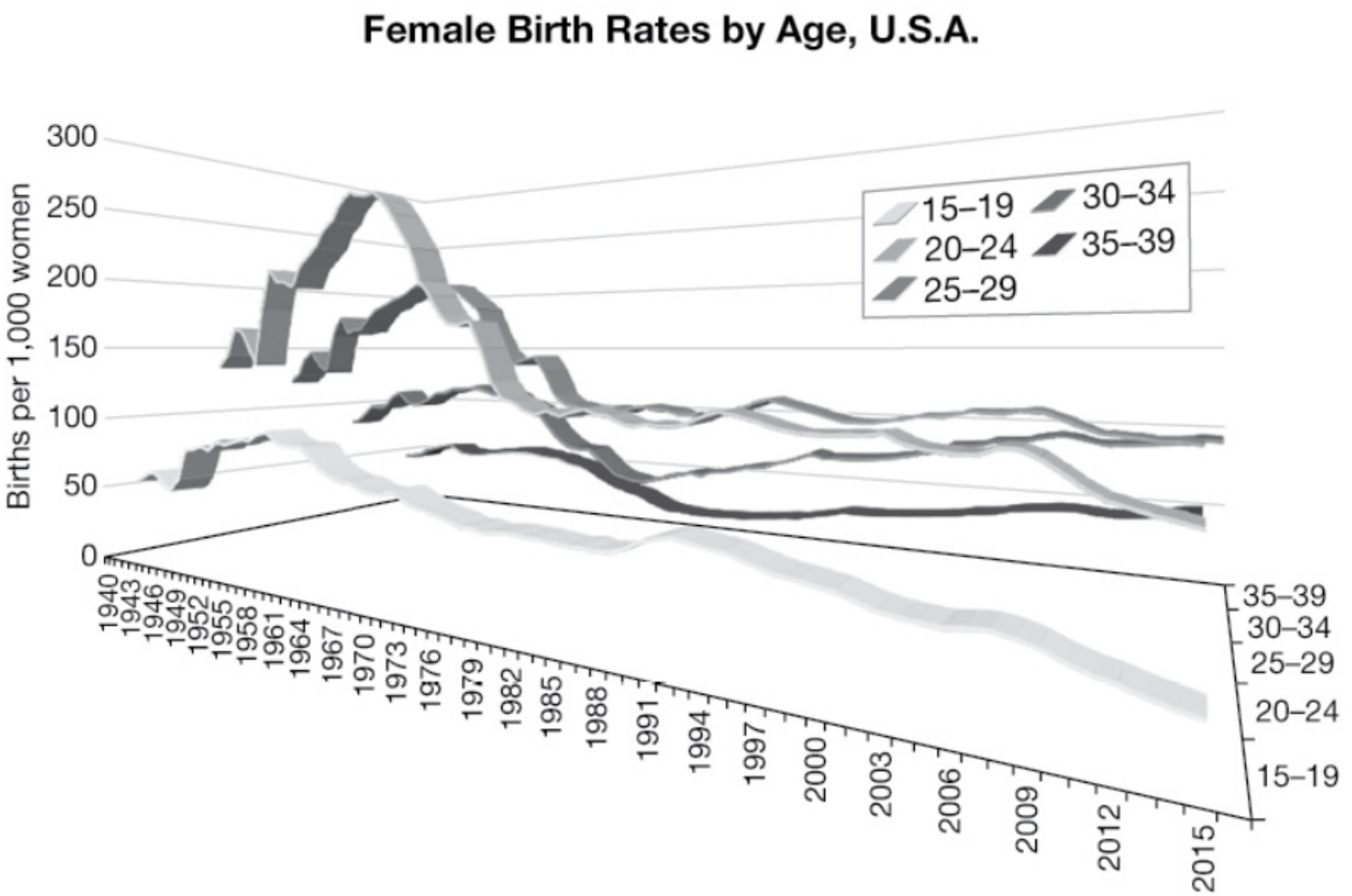
Total


Number of Drivers in Fatal Crashes, 1988
Number of drivers (thousands)
0
1
2
3
4
5
6
7
8
9
16–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
75–79
>79
Age group

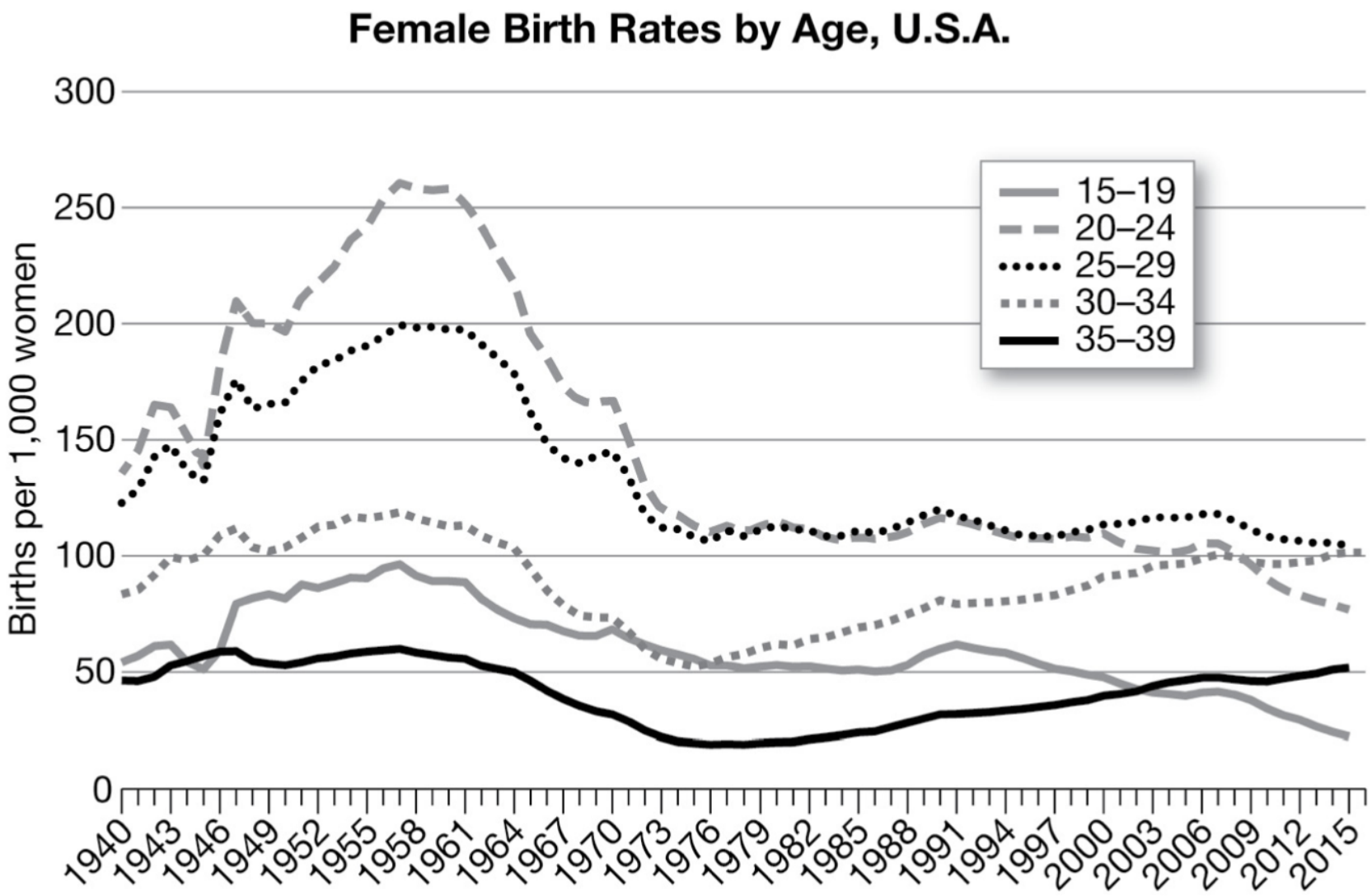
Relative


Fatal Crashes per 100 Million Miles Driven, 1988
Fatal crashes per 100 million miles driven
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16–19
20–24
25–29
30–34
35–39
40–44
45–49
50–54
55–59
60–64
65–69
70–74
>75

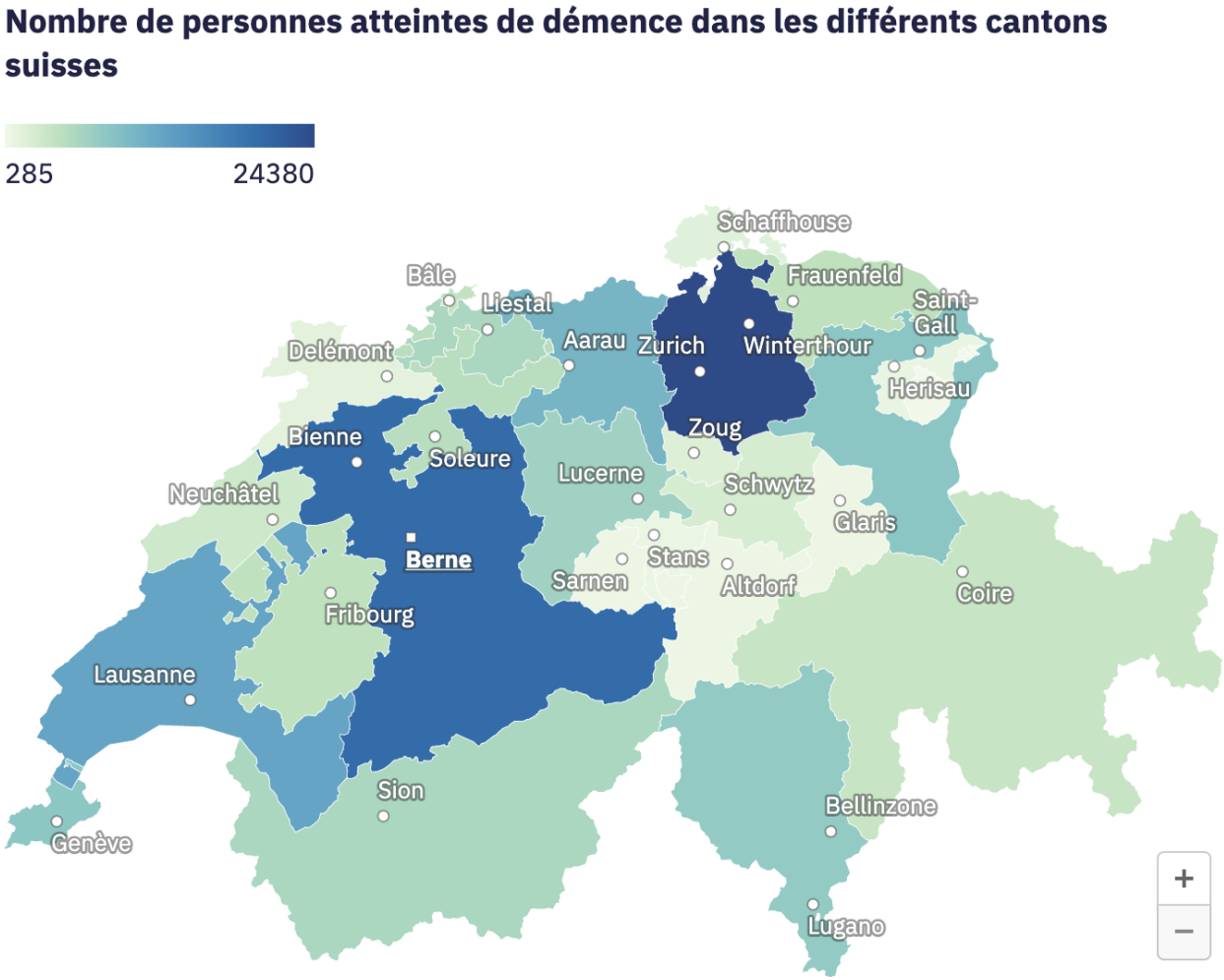
Useless 3D


Female Birth Rates by Age, U.S.A.
Births per 1,000 women
300
250
200
150
100
50
0
15–19
20–24
25–29
30–34
35–39
1940
1943
1946
1949
1952
1955
1958
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015
35–39
30–34
25–29
20–24
15–19

Better 2D

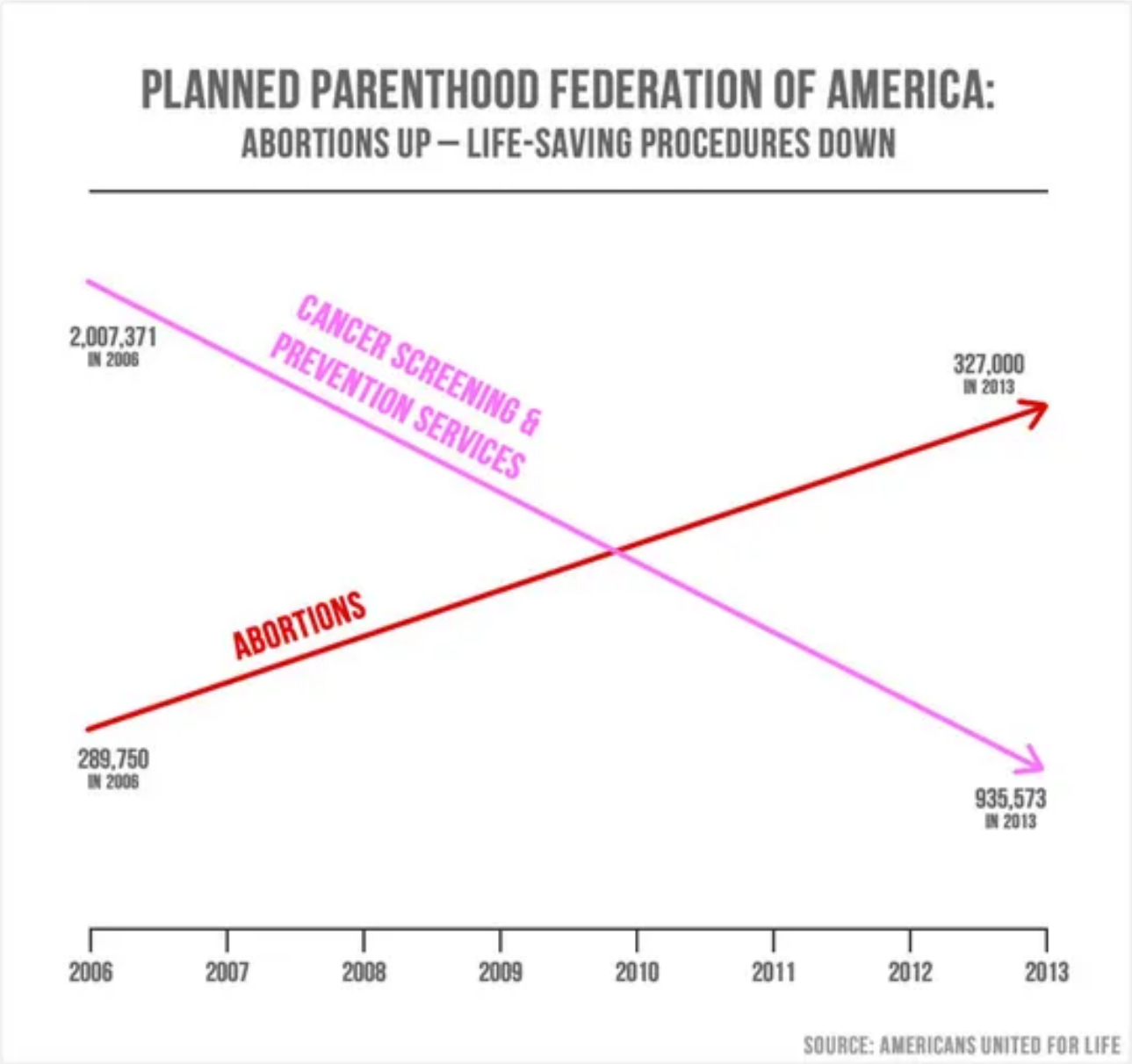

Female Birth Rates by Age, U.S.A.
Births per 1,000 women
300
250
200
150
100
50
0
15–19
20–24
25–29
30–34
35–39
1940
1943
1946
1949
1952
1955
1958
1961
1964
1967
1970
1973
1976
1979
1982
1985
1988
1991
1994
1997
2000
2003
2006
2009
2012
2015

Total vs. Relative Again

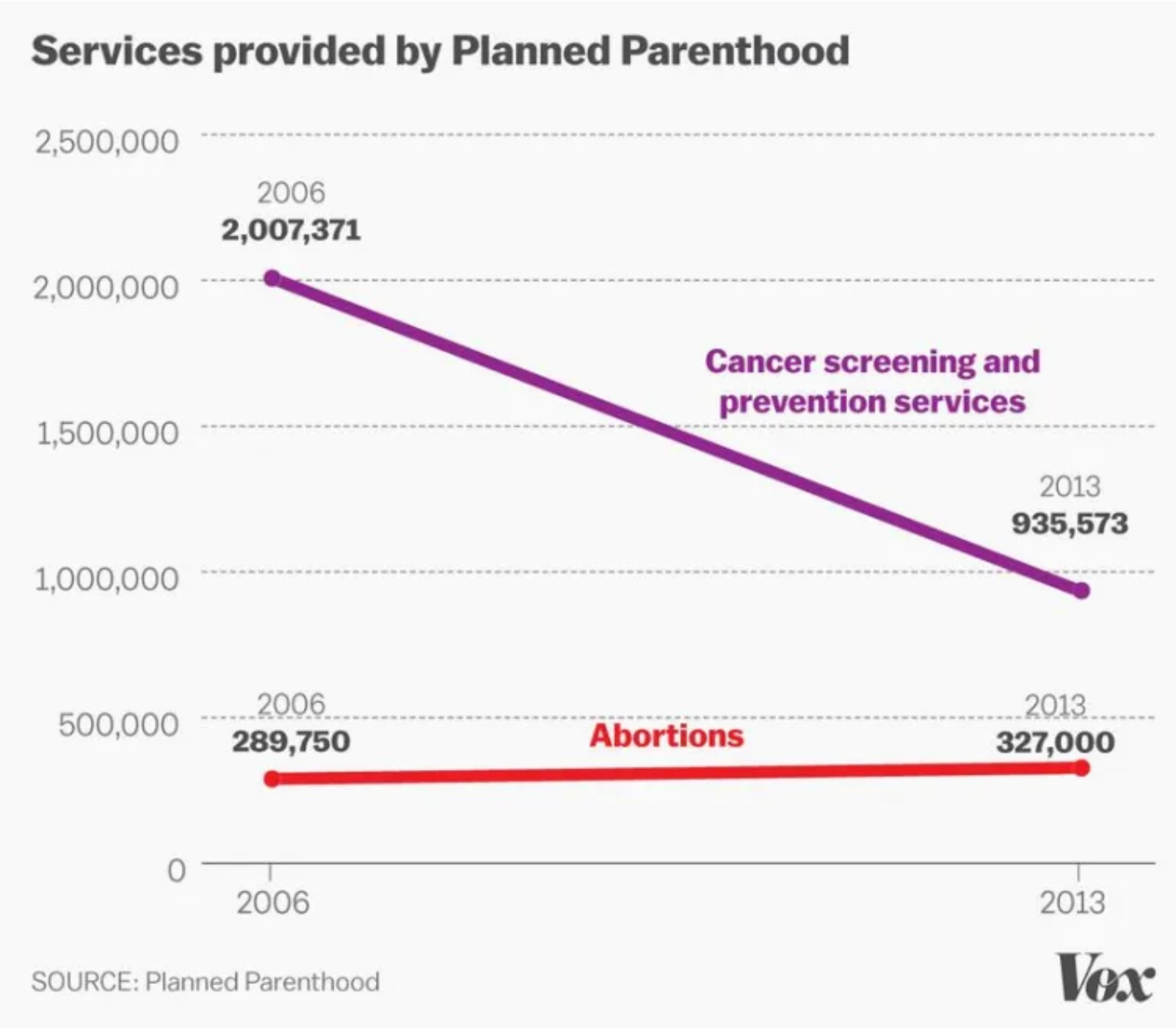


Carte: G. Laplace. Nombre exact par canton en se positionnant dessus avec la souris; Source: Faits et chiffres (Alzheimer Suisse, 2021); Récupérer les données

Missing Axis & Misguidance (1)



Missing Axis & Misguidance (2)



Data Visualization Caveats

A collection of caveats in From Data to Viz

Assignment

Small project [20%]. Deadline on Week 5.

The goal of this project is *data exploration*. Details can be found on the dedicated course page.

Some links to open data can be found here.

References

- Krause, Andreas, Nicola Rennie and Brian Tarran (2023) Best Practices for Data Visualisation
- Poldrack (2019) Statistical Thinking for the 21st Century
- JASA Ethical Guidelines for Statistical Practice
- Gelman (2018) Ethics in statistical practice and communication
- Wickham & Grolemund (2017) R for Data Science