

Data Wrangling and Data Analysis

Data Visualization

Daniel L. Oberski & Erik-Jan van Kesteren

Department of Methodology & Statistics

Utrecht University



Social & Behavioural Science team

Erik-Jan

e.vankesteren1@uu.nl



Ayoub

a.bagheri@uu.nl



Daniel

d.loberski@uu.nl



Timo

timo_kuil@hotmail.com



Hanne

h.i.oberman@uu.nl



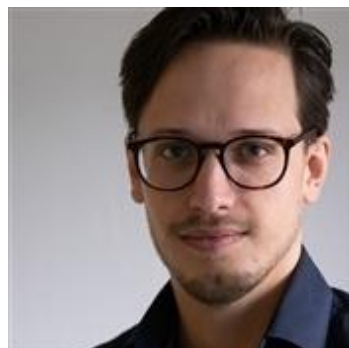
Elena

e.candellone@uu.nl



Jelle

j.j.teijema@uu.nl



Özgür

o.togay@uu.nl



Mahdi

m.shafieekamalabad@uu.nl



Thom

t.b.volker@uu.nl



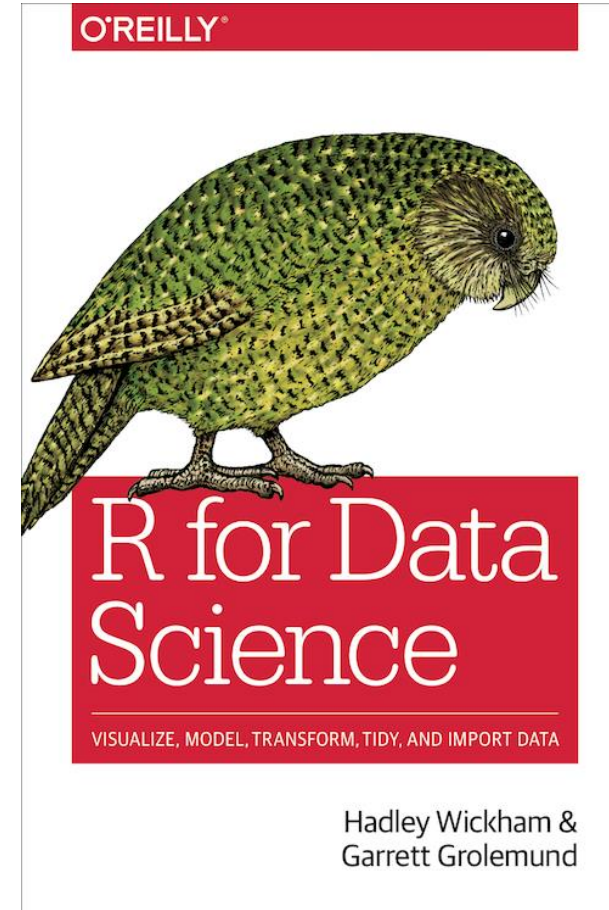
This week

1. Data preparation 1
2. Data preparation 2
- 3. Data visualization principles & Grammar of graphics**



Reading materials for this week

- Chapters from **R for Data Science (R4DS)**, open access book at:
- <https://r4ds.had.co.nz>
- Chapters: 1,3,9,10.



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

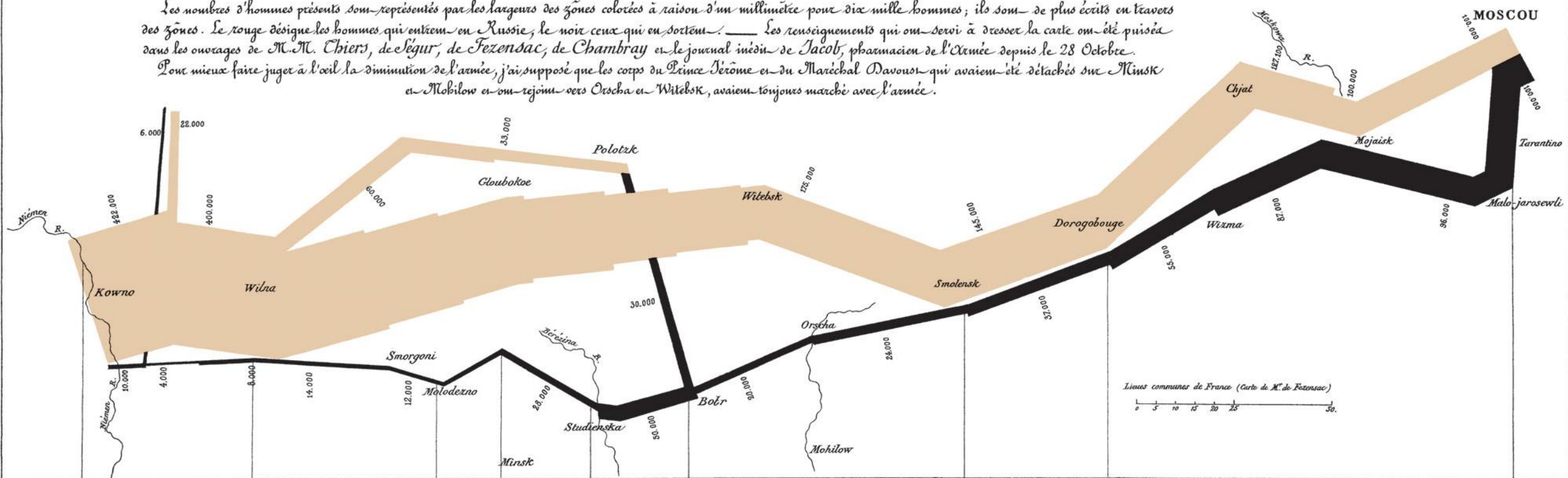
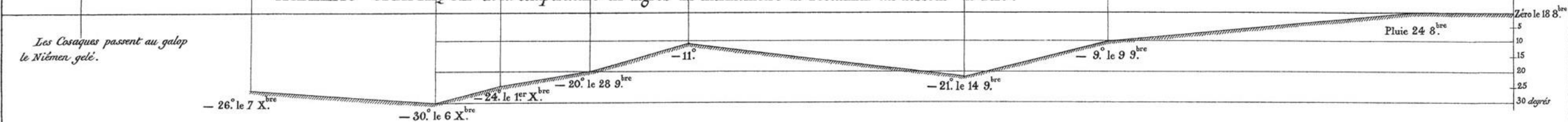
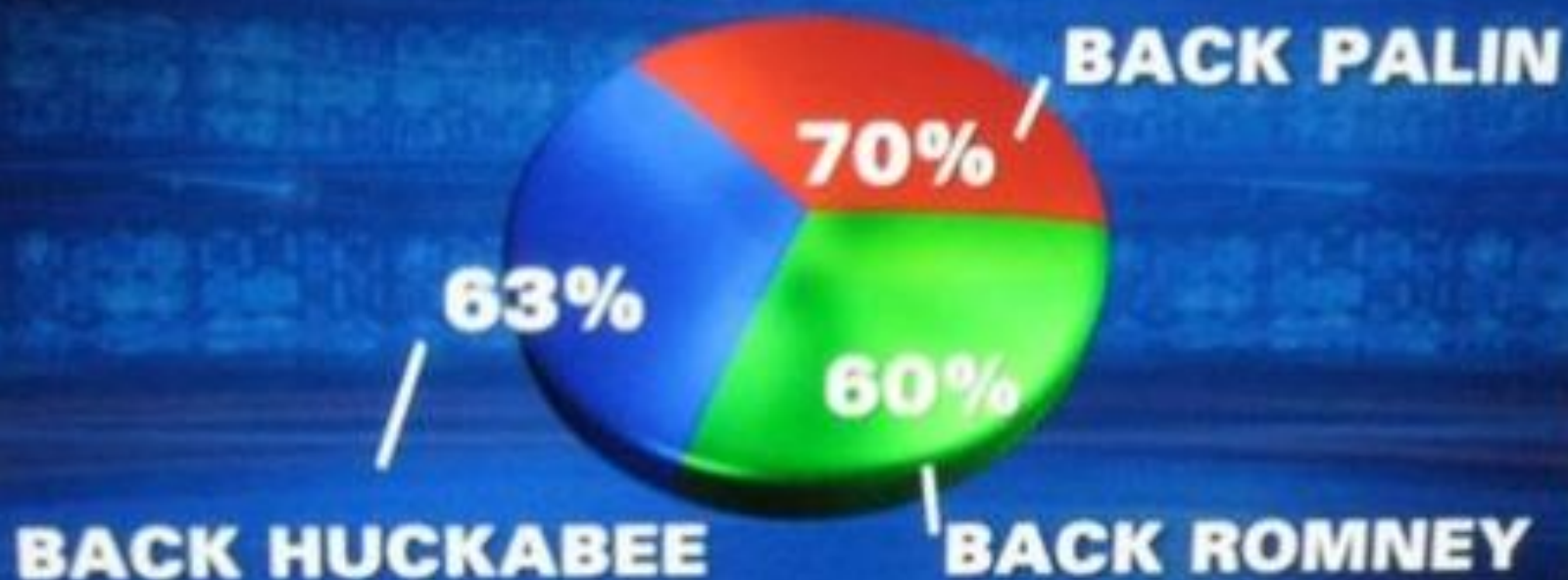


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



2012 PRESIDENTIAL RUN

GOP CANDIDATES



FOX

47°

SOURCE: OPINIONS

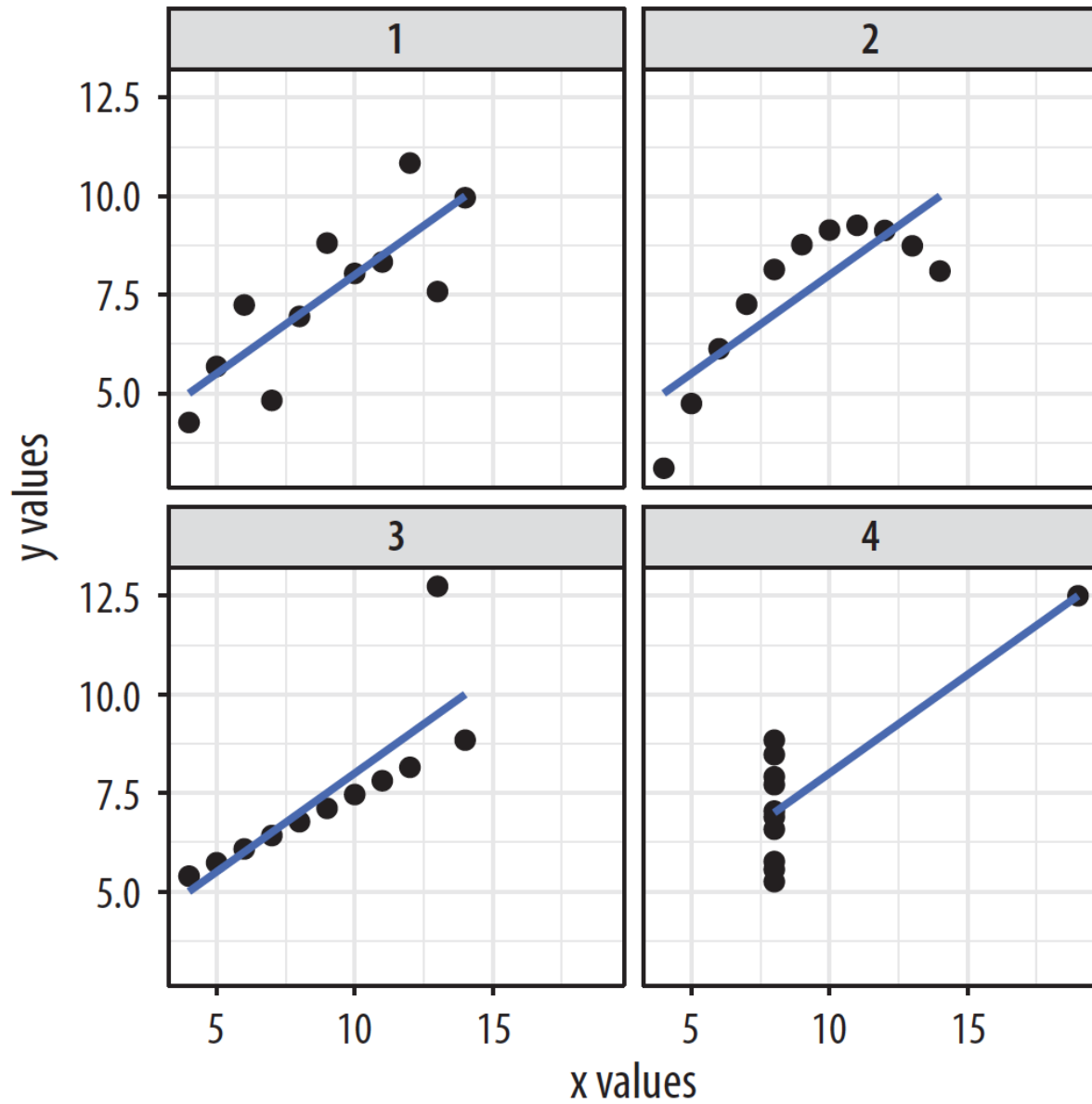
DYNAMIC

Today: visualization principles

Data visualization

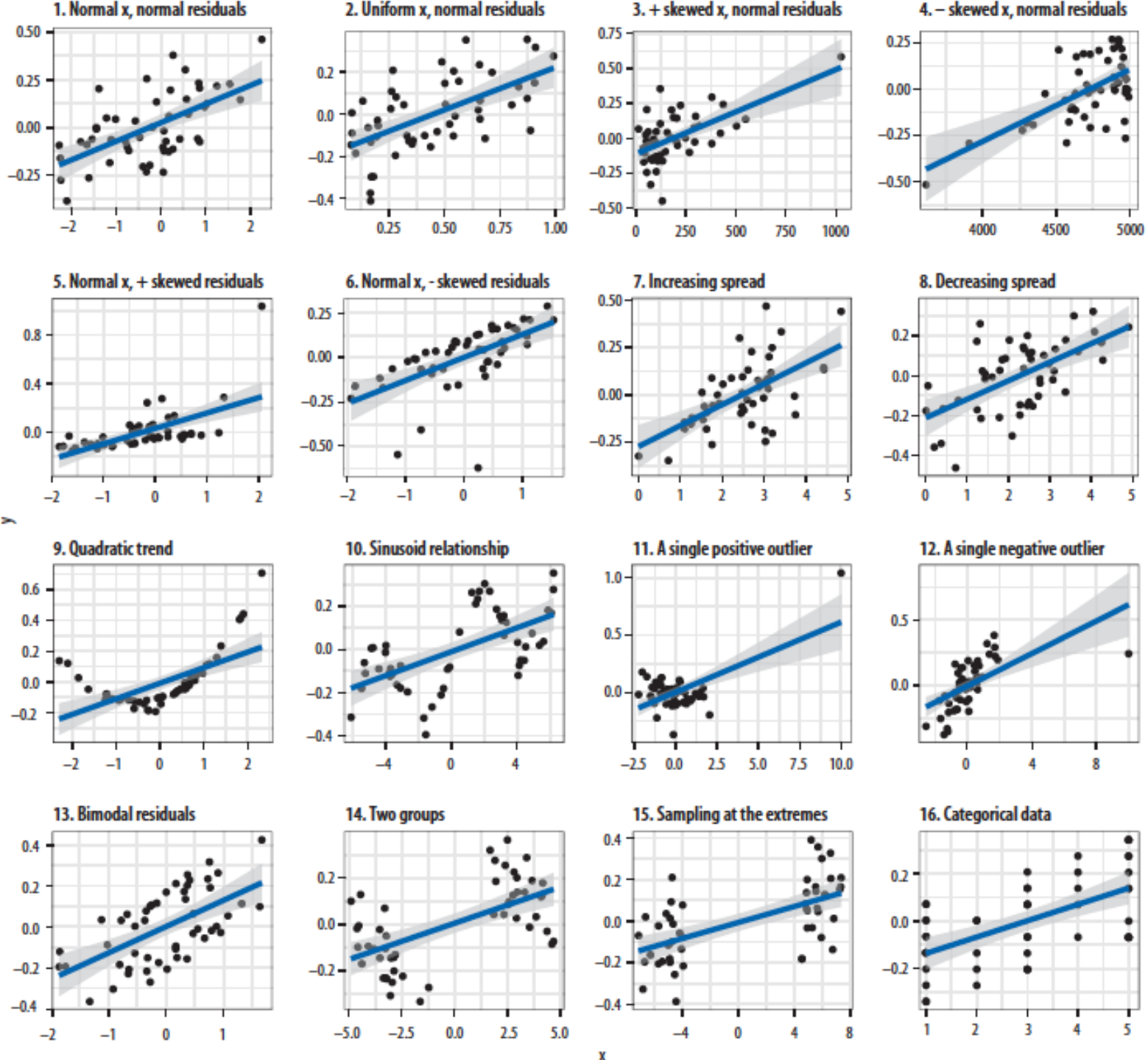
- **For exploration, data analysis** ←
- For communication
- For entertainment

Anscombe's quartet



Source: Healy (2019)





Graphics for data analysis

- The **human retina** can transfer around 10^6 or 10^7 bits per second to the brain;
- **Reading** transfers about 3 words, so $\sim 10^2$ or 10^3 bits/s;
- Potentially (!) visualization is about 4 orders of magnitude more powerful.

How can we leverage the human visual system to analyze data?



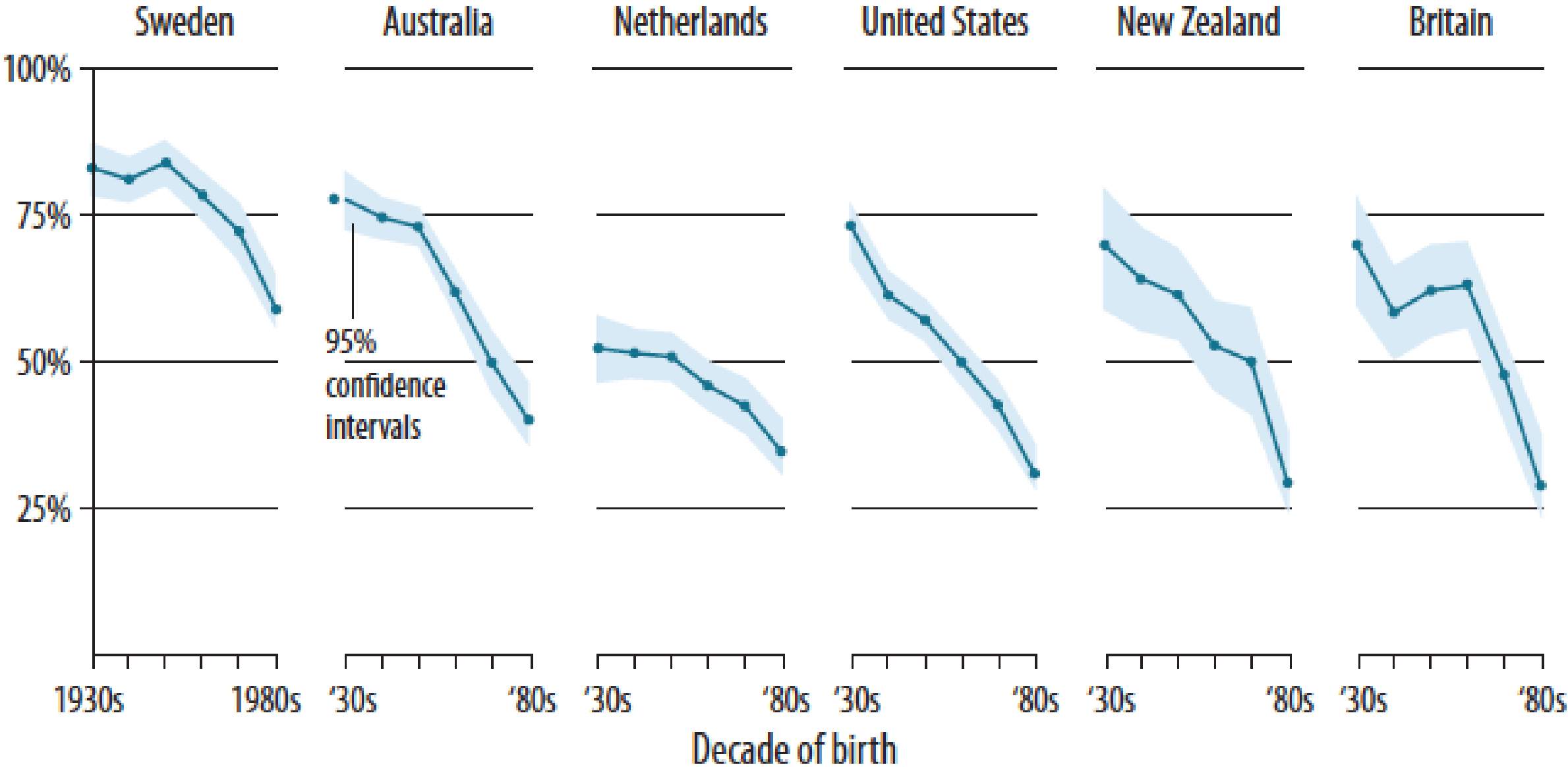
Plotting the right thing

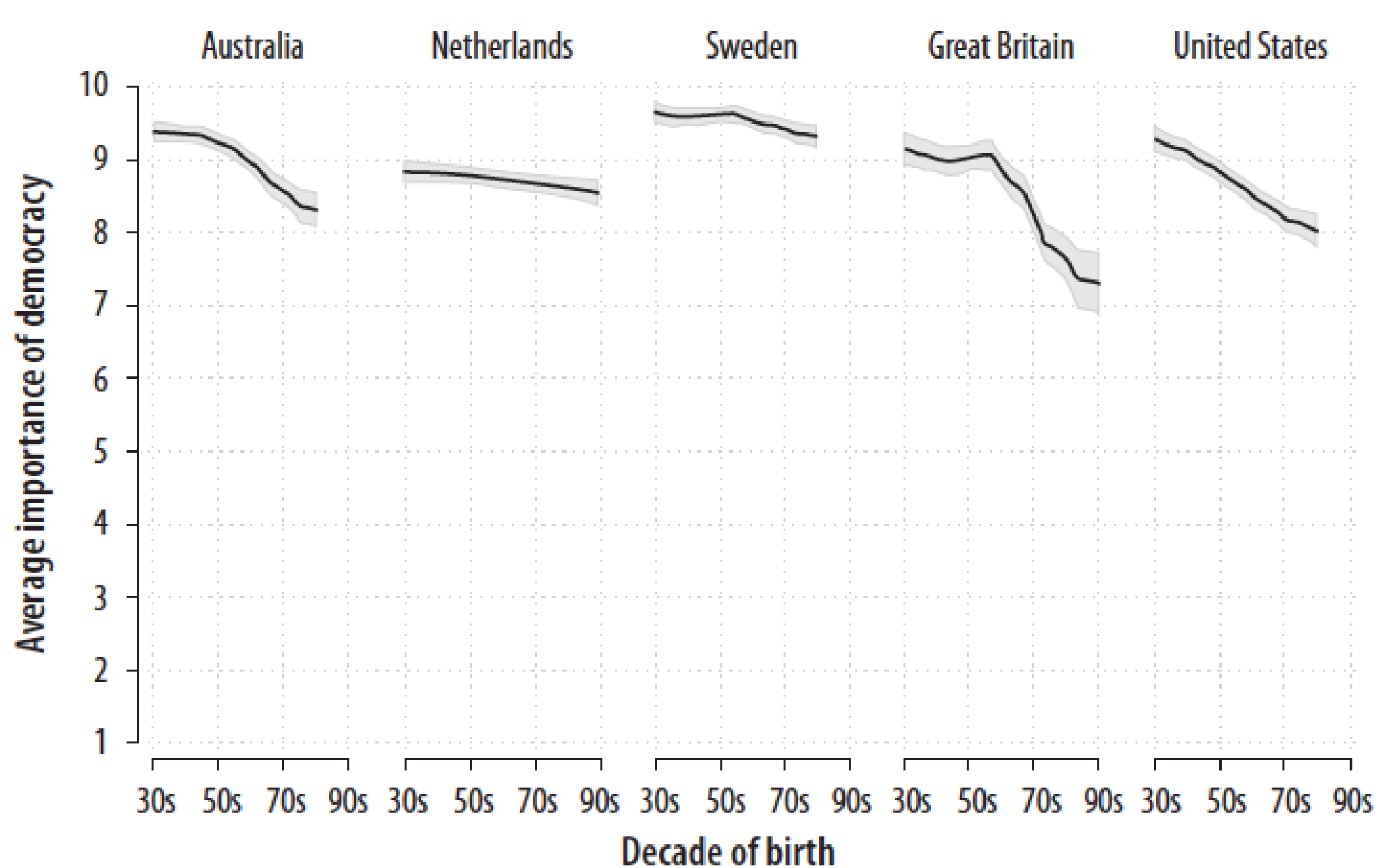
Most common problems:

- (Accidentally) misrepresenting what is being plotted
- Omitting baselines



Percentage of people who say it is “essential” to live in a democracy



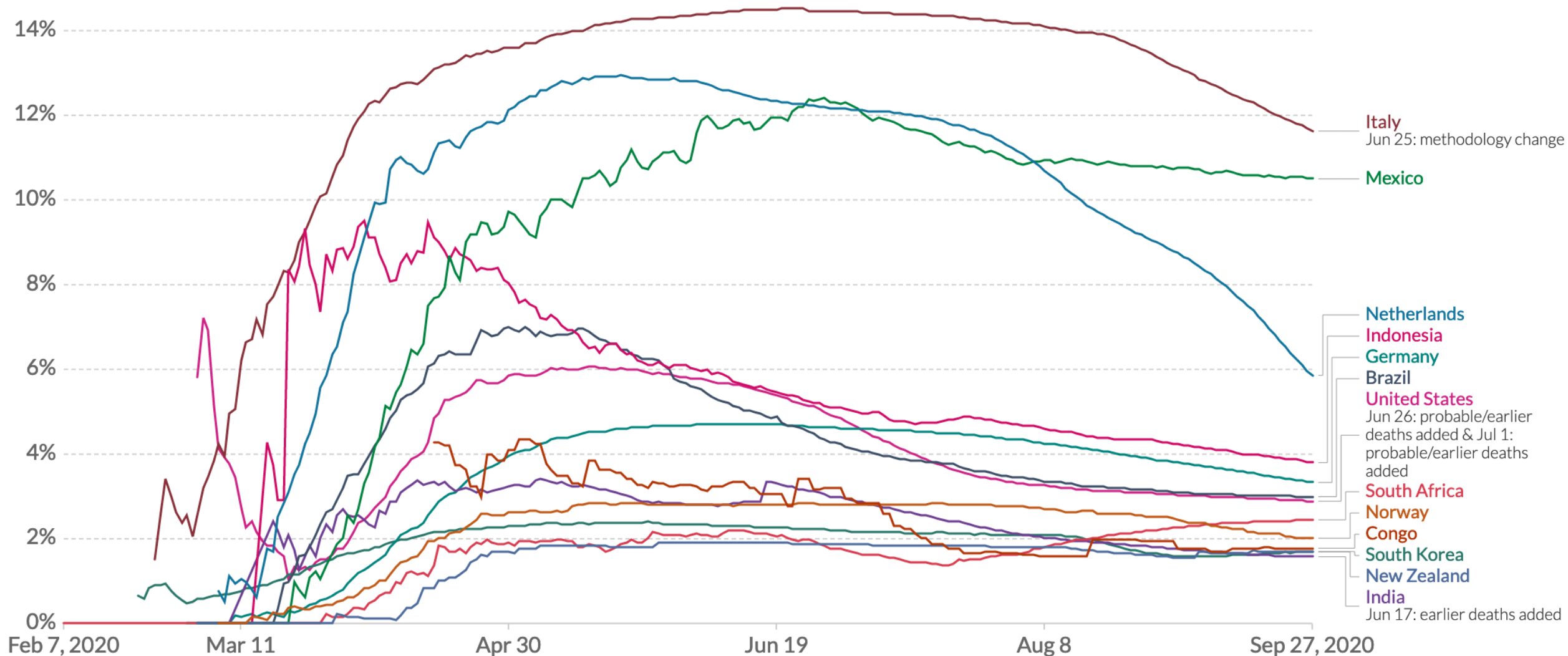


Graph by Erik Voeten, based on WVS 5

Source: Healy (2019)

Case fatality rate of the ongoing COVID-19 pandemic

The Case Fatality Rate (CFR) is the ratio between confirmed deaths and confirmed cases. During an outbreak of a pandemic the CFR is a poor measure of the mortality risk of the disease. We explain this in detail at OurWorldInData.org/Coronavirus

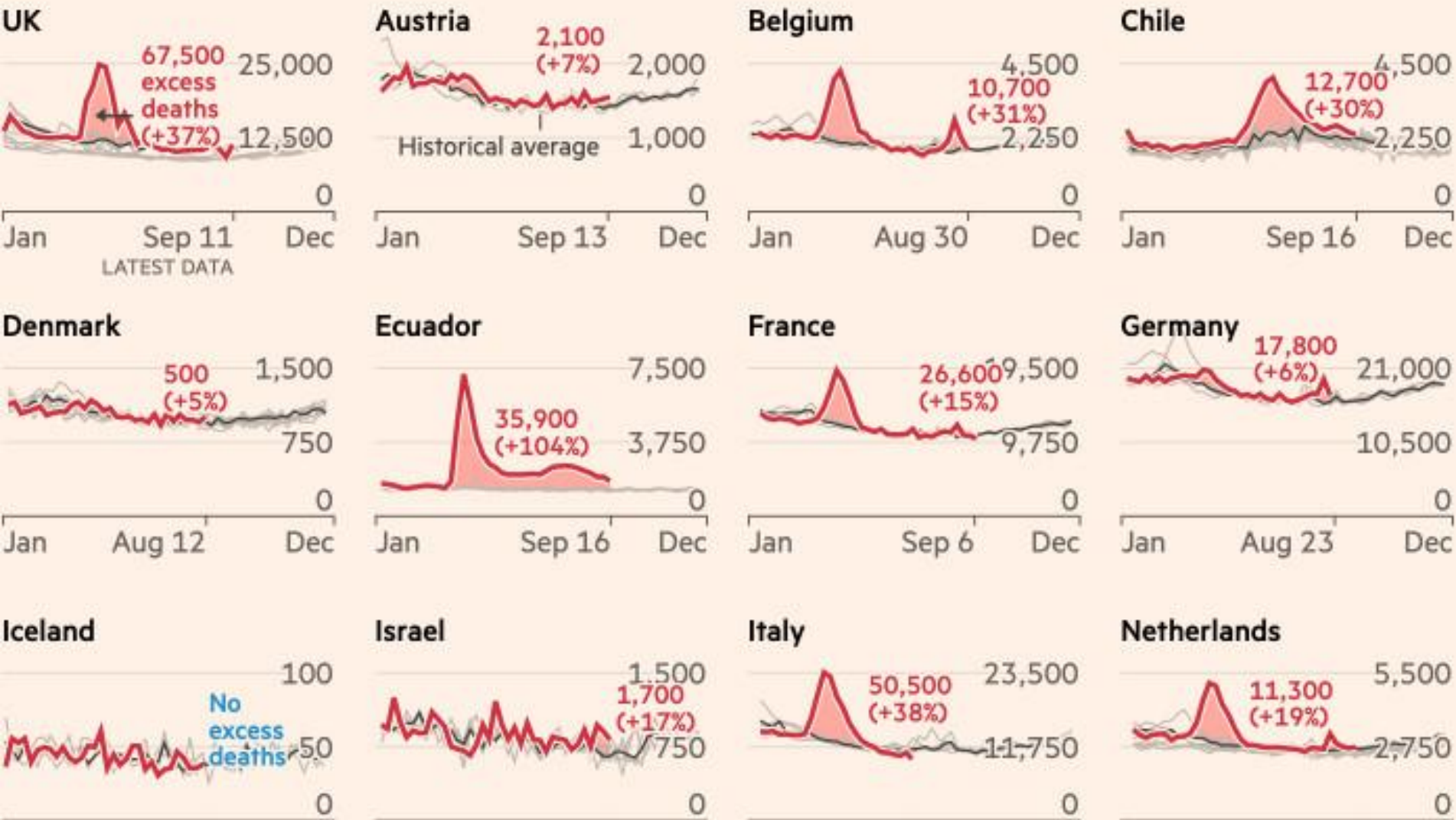


Death rates have climbed far above historical averages in many countries that have faced Covid-19 outbreaks

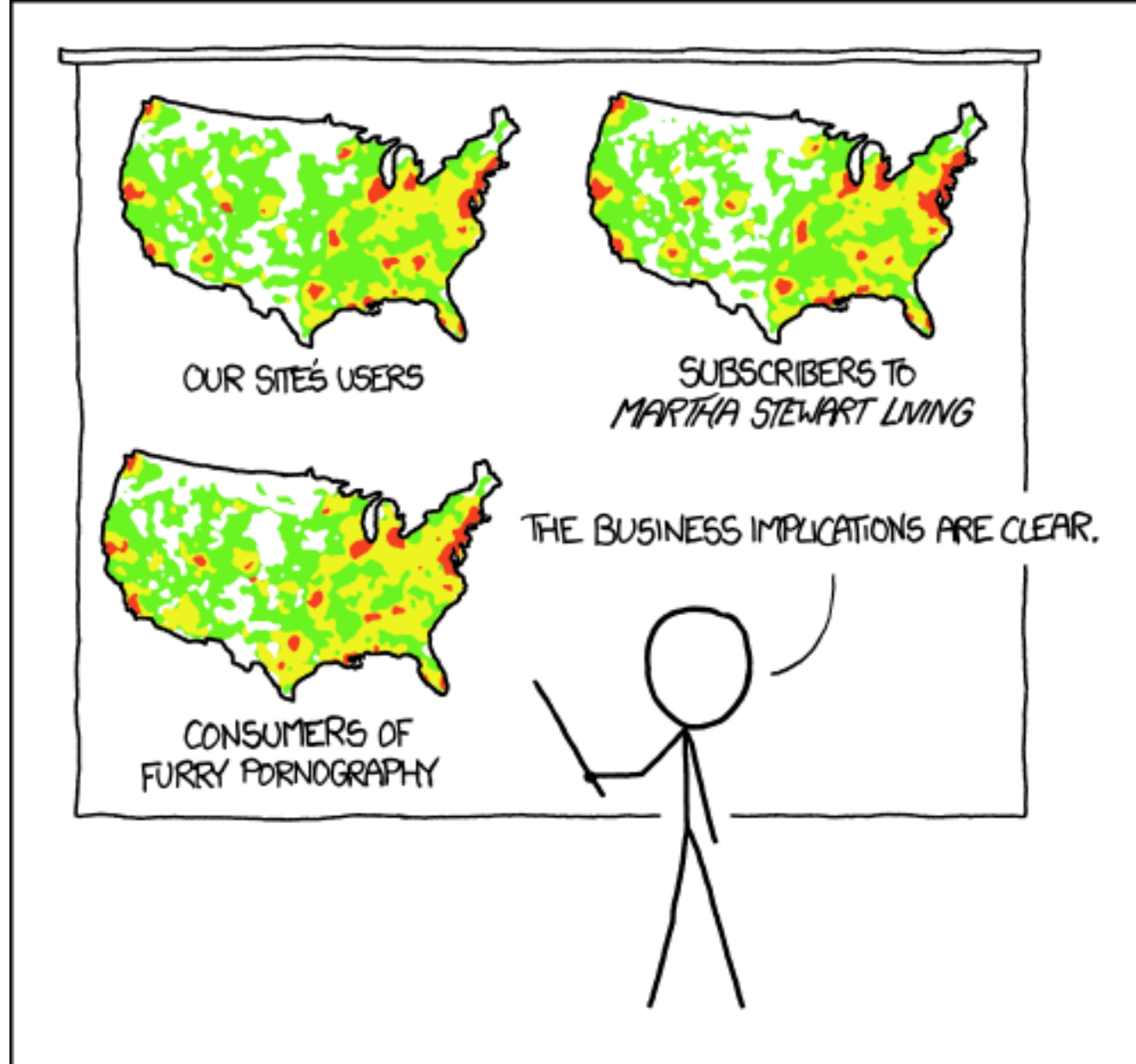
Good example (FT)

Number of deaths per week from all causes, 2020 vs recent years:

Shading indicates total excess deaths during outbreak



<https://www.ft.com/content/a298-5eb7-4633-b89c-cbdf5b38693>



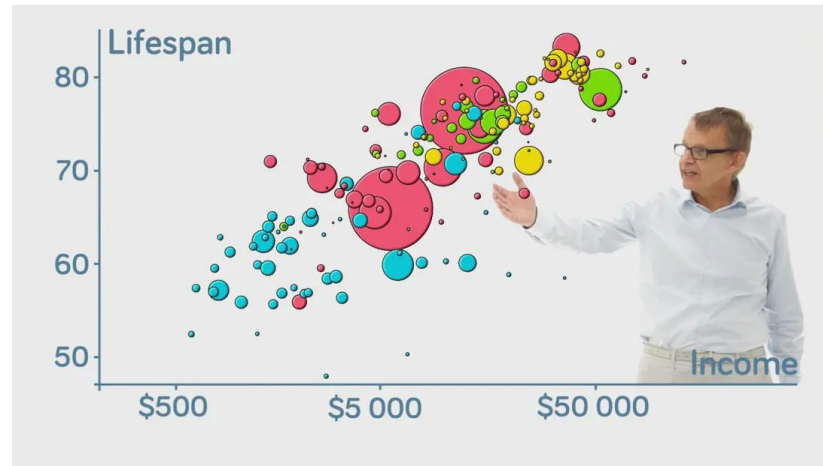
PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

Data

Category	0-Adolescent (ref.)	No. of trials	Mean lifespan (days ± SEM)	P-value	Residual CI
NE	0	7	15.2 ± 0.7	0.0124	4.0
	1	7	20.2 ± 0.3 ***	<0.0001	7.0
	2	7	20.9 ± 0.3 **	0.0062	7.5
NE+	0	3	35.7 ± 2.8	<0.0001	17.3
	1	3	40.2 ± 1.8 ***	<0.0001	21.0
	2	3	40.2 ± 3.2 ***	<0.0001	21.0
NE-2	0	4	21.9 ± 1.7	0.0005	2.3
	1	4	22.9 ± 2.2 NS	0.0735	1.8
	2	4	22.3 ± 2.8 NS	0.0735	1.8
NE+2	0	4	12.4 ± 0.3	<0.0001	4.0
	1	4	11.4 ± 0.3 ***	<0.0001	4.0
	2	4	12.8 ± 0.3 ***	<0.0001	4.0
NE+2+	0	2	18.5 ± 2.8	0.0180	7.6
	1	2	18.0 ± 1.7 *	0.0062	12.4
	2	2	26.8 ± 3.3 **	0.0062	12.4
NE+2+	0	4	11.5 ± 0.3	0.5441	8.6
	1	4	11.5 ± 0.3 NS	<0.0001	8.6
	2	4	14.0 ± 0.3 ***	<0.0001	8.6
NE+2+	0	3	18.1 ± 0.8	<0.0001	14.4
	1	3	20.7 ± 0.3 ***	<0.0001	14.4
	2	3	26.3 ± 0.4 ***	<0.0001	12.1



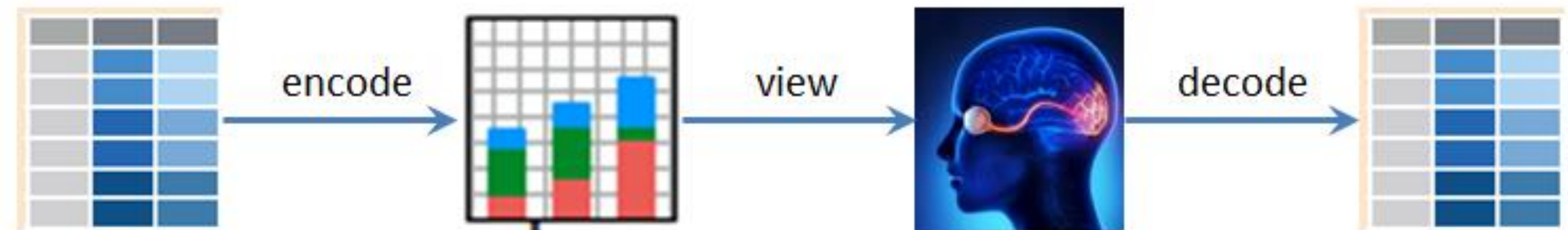
Picture



Brain



shutterstock.com · 1666544830



Making pictures that help analyze data

- We'd like to make, not just any kind of picture or graph, but one that transfers some part of the data to our brain
- How do we make sure that the graphs we make transfer:
 - The right part of the data, and;
 - As much of it as possible?

This is where the “**grammar of graphics**” comes in.

Goal is to **specify how data map to picture**, so the correct type and largest amount possible is transferred



Grammar of graphics (Wickham version)

- <http://r4ds.had.co.nz/visualize.html>
- Map raw data to following elements:
 - Aesthetics (position, shape, color, ...)
 - Geometric objects (points, lines, bars, ...)
 - Scales (continuous, discrete, ...)
 - Facets (small multiples)
- Additionally, can apply:
 - Statistical transformation (identity, binning, median, ...)
 - Coordinate system (Cartesian, polar, parallel, ...)



Grammar of graphics (Wickham version)

In R, grammar of graphics is implemented in `ggplot`, a function in the `ggplot2` package.



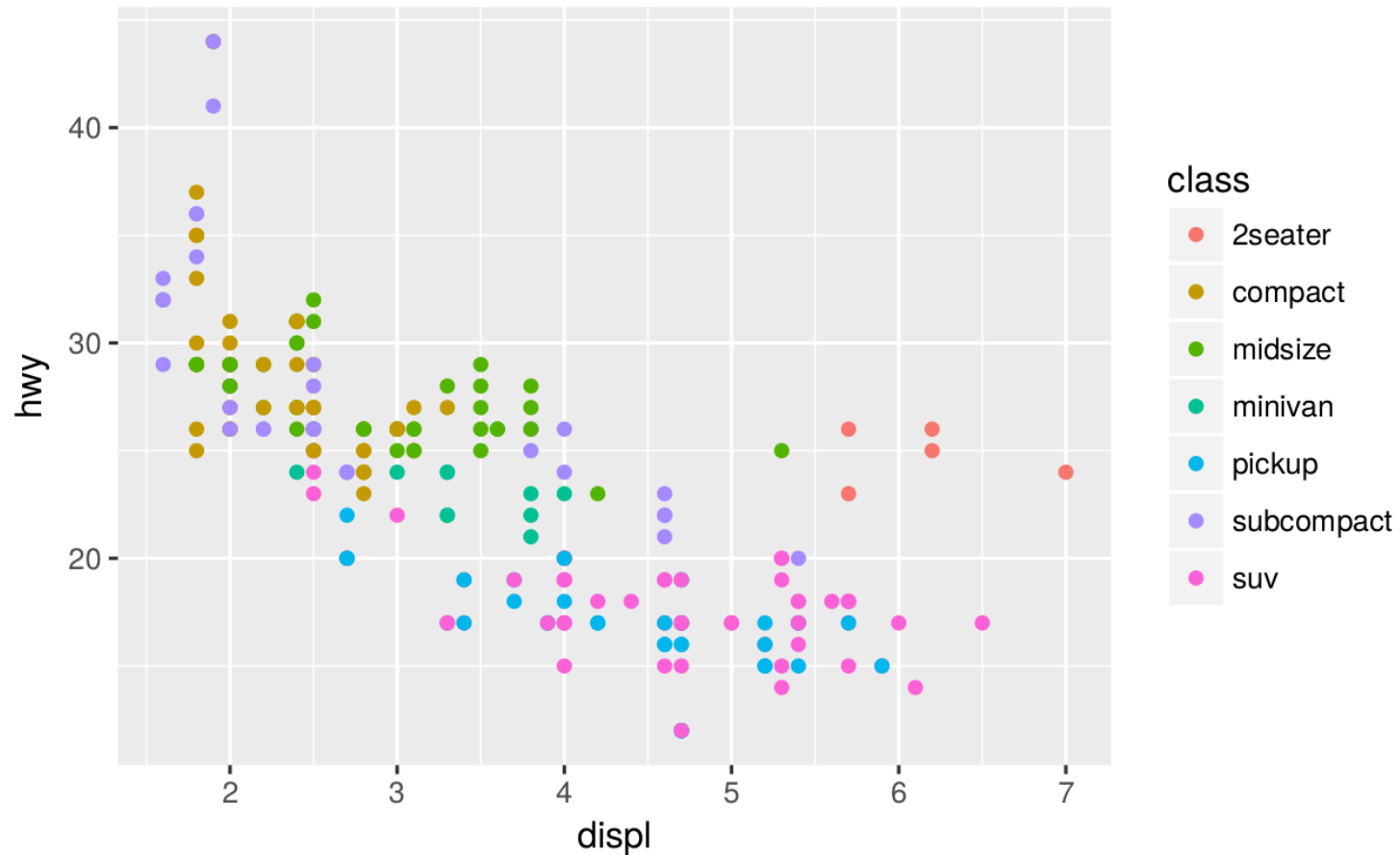
Example data set: cars

mpg

```
#> # A tibble: 234 × 11
#>   manufacturer model displ  year  cyl      trans  drv   cty   hwy fl
#>   <chr>    <chr>  <dbl> <int> <int>    <chr> <chr> <int> <int> <chr>
#> 1      audi    a4    1.8  1999     4  auto(l5)   f    18    29   p
#> 2      audi    a4    1.8  1999     4 manual(m5)   f    21    29   p
#> 3      audi    a4    2.0  2008     4 manual(m6)   f    20    31   p
#> 4      audi    a4    2.0  2008     4  auto(av)   f    21    30   p
#> 5      audi    a4    2.8  1999     6  auto(l5)   f    16    26   p
#> 6      audi    a4    2.8  1999     6 manual(m5)   f    18    26   p
#> # ... with 228 more rows, and 1 more variables: class <chr>
```

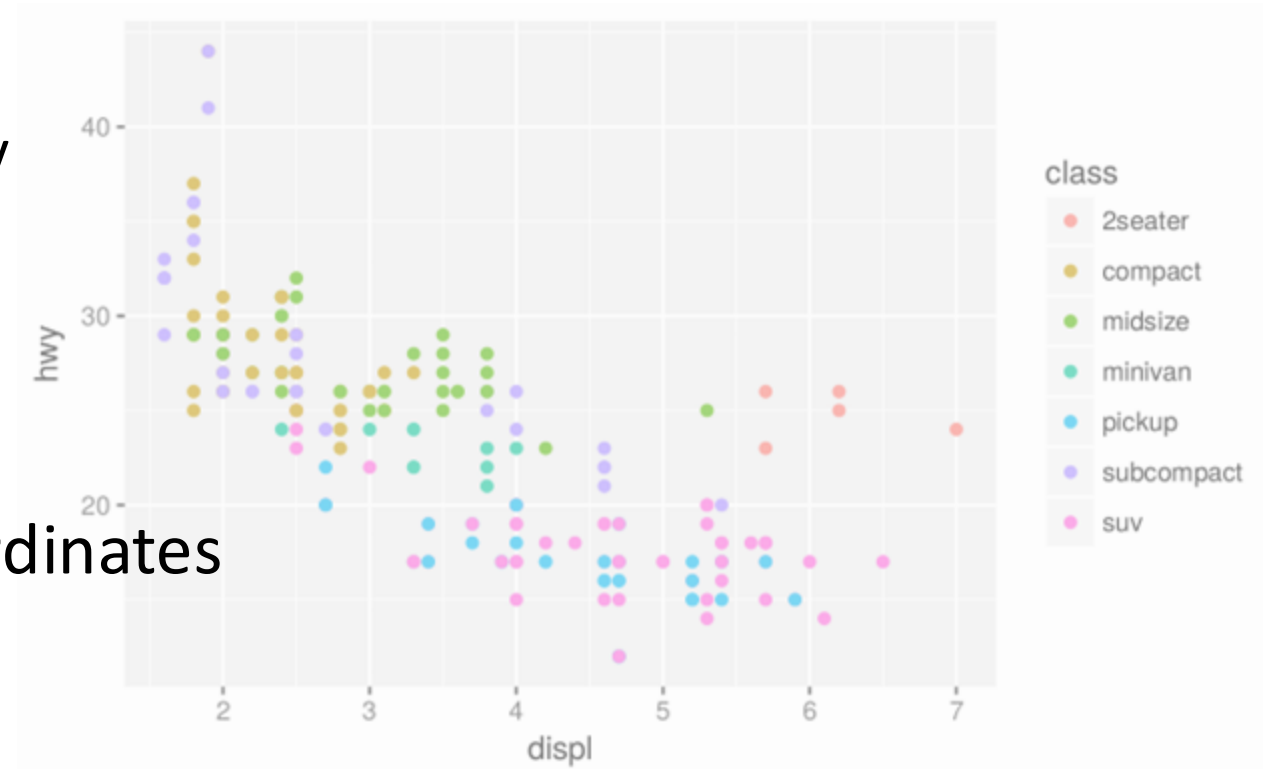


```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy,  
                           color = class))
```



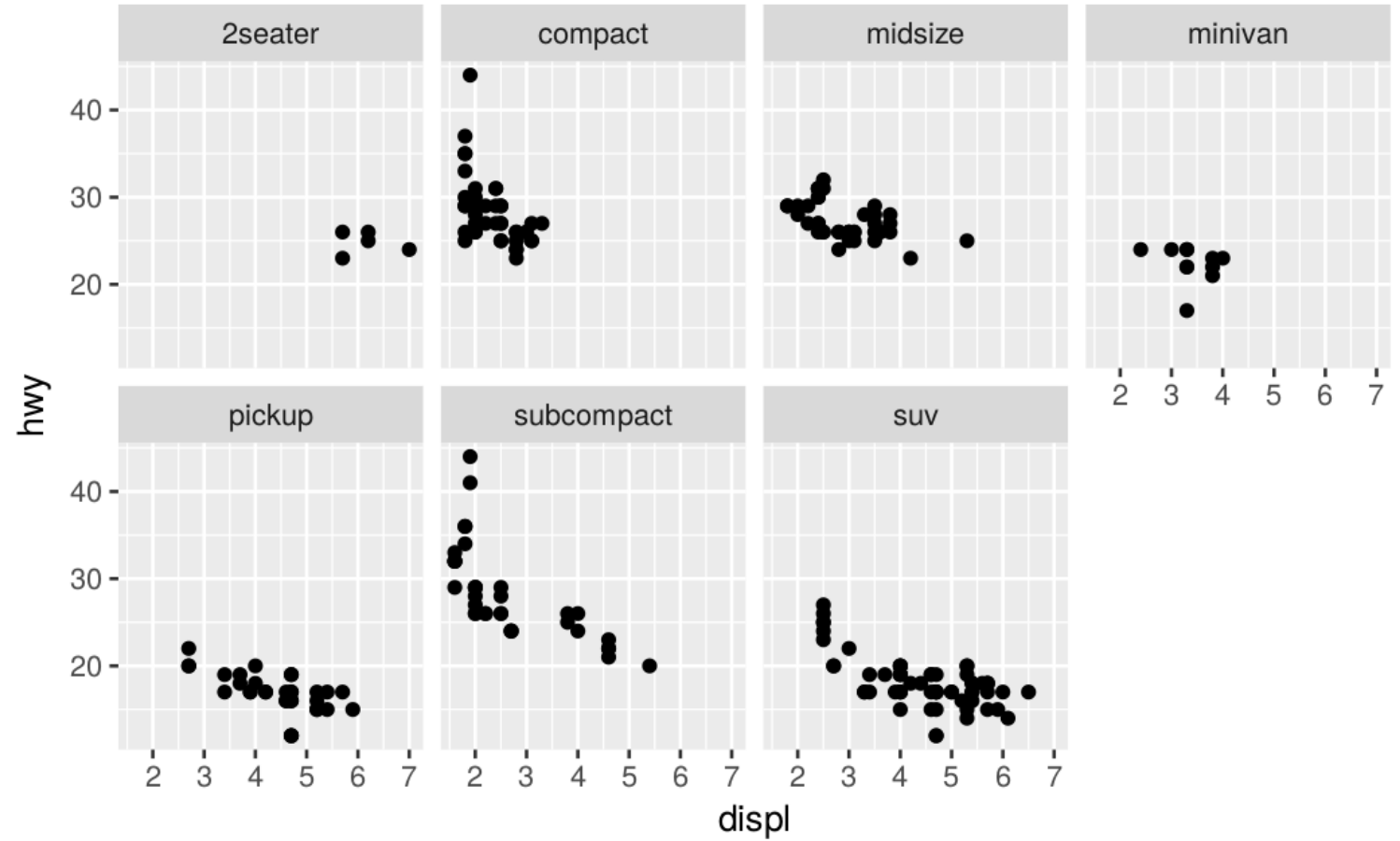
```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy,  
                           color = class))
```

- Aesthetics:
 - x-position mapped to engine size
 - y-position mapped to fuel efficiency
 - color mapped to car type
- Geometric objects: points
- Transformation: identity
- Scales: continuous, cartesian coordinates
- No facets



Facets

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy)) +  
  facet_wrap(~ class, nrow = 2)
```



Transformation (stats)

1. **geom_bar()** begins with the **diamonds** data set

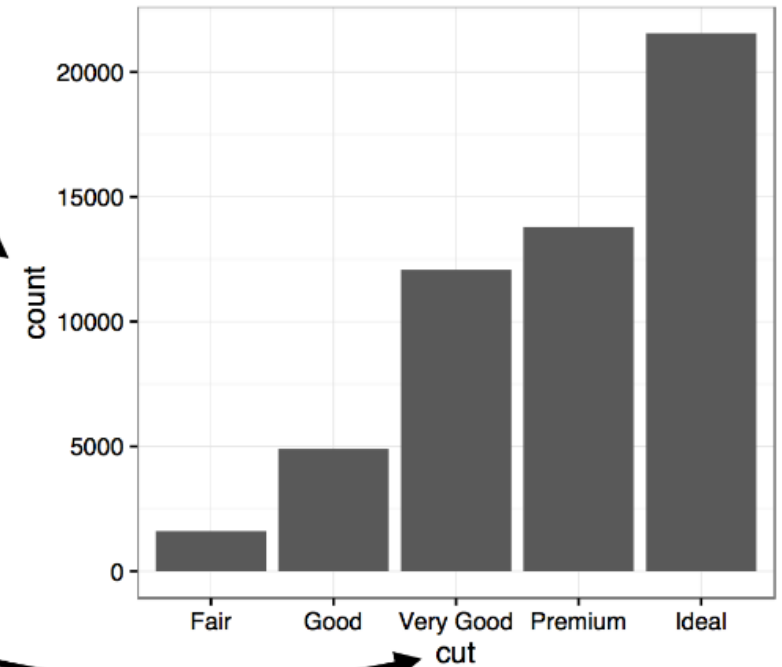
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
...	...	...	...	...	...	...	...	...	...

2. **geom_bar()** transforms the data with the "count" stat, which returns a data set of cut values and counts.

stat_count()

cut	count	prop
Fair	1610	1
Good	4906	1
Very Good	12082	1
Premium	13791	1
Ideal	21551	1

3. **geom_bar()** uses the transformed data to build the plot. cut is mapped to the x axis, count is mapped to the y axis.



What should I choose?

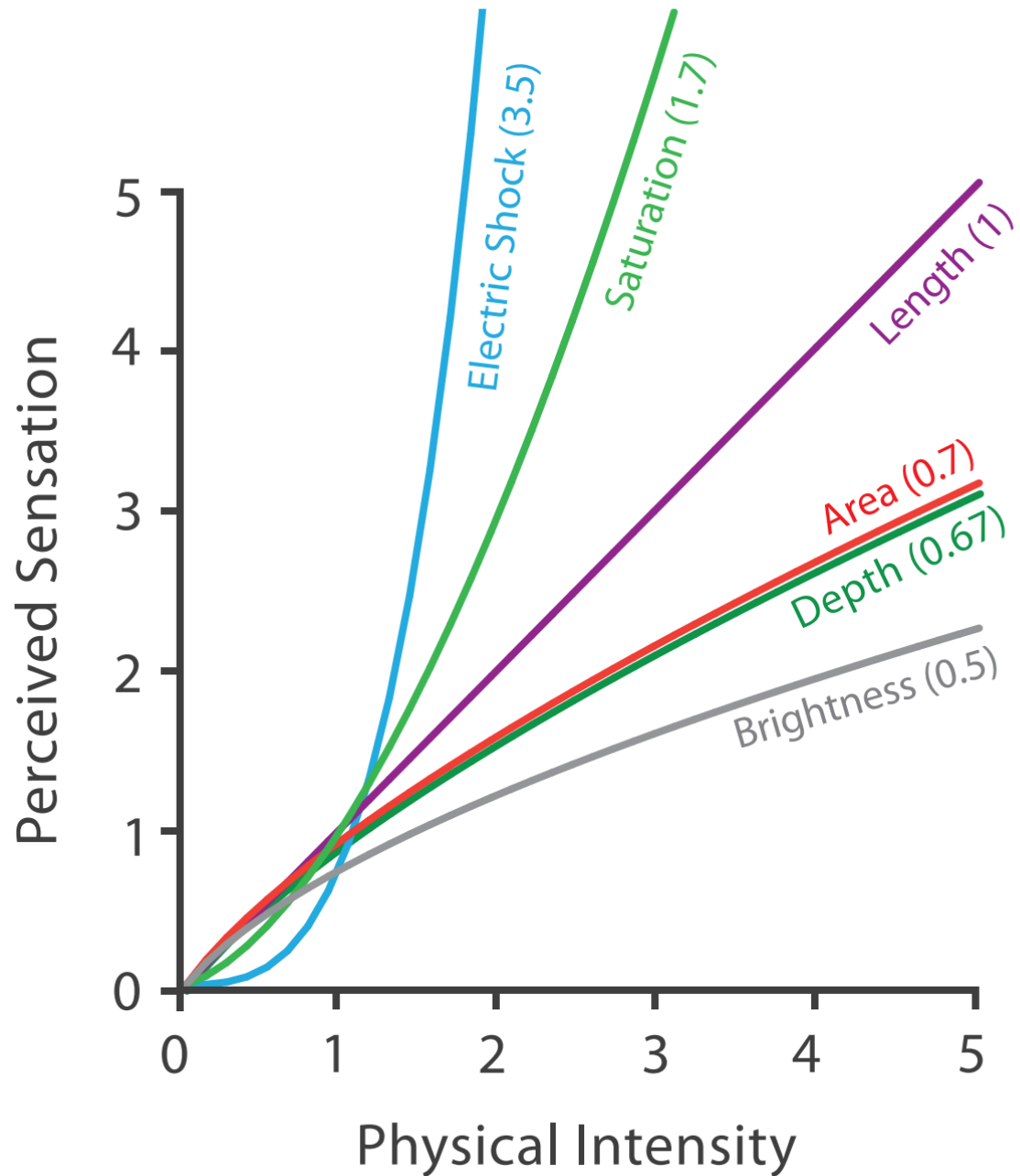
LES VARIABLES DE L'IMAGE

			POINTS			LIGNES			ZONES	
XY 2 DIMENSIONS DU PLAN										
	Z	TAILLE								
		VALEUR								

LES VARIABLES DE SÉPARATION DES IMAGES

	GRAIN				
	COULEUR				
	ORIENTATION				
	FORME				

Steven's Psychophysical Power Law: $S = I^N$



Source: **Tamara Munzer** (2014). Visualization Analysis and Design.

Channels: Expressiveness Types and Effectiveness Ranks

➔ **Magnitude** Channels: **Ordered** Attributes

Position on common scale



Position on unaligned scale



Length (1D size)



Tilt/angle



Area (2D size)



Depth (3D position)



Color luminance



Color saturation



Curvature



Volume (3D size)



Same
Same
Same

Most
Effectiveness
Least

➔ **Identity** Channels: **Categorical** Attributes

Spatial region



Color hue



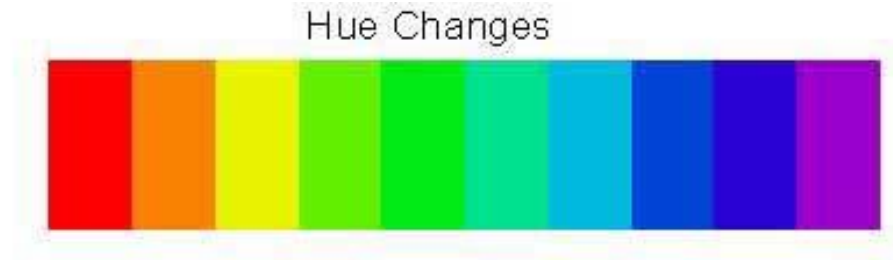
Motion

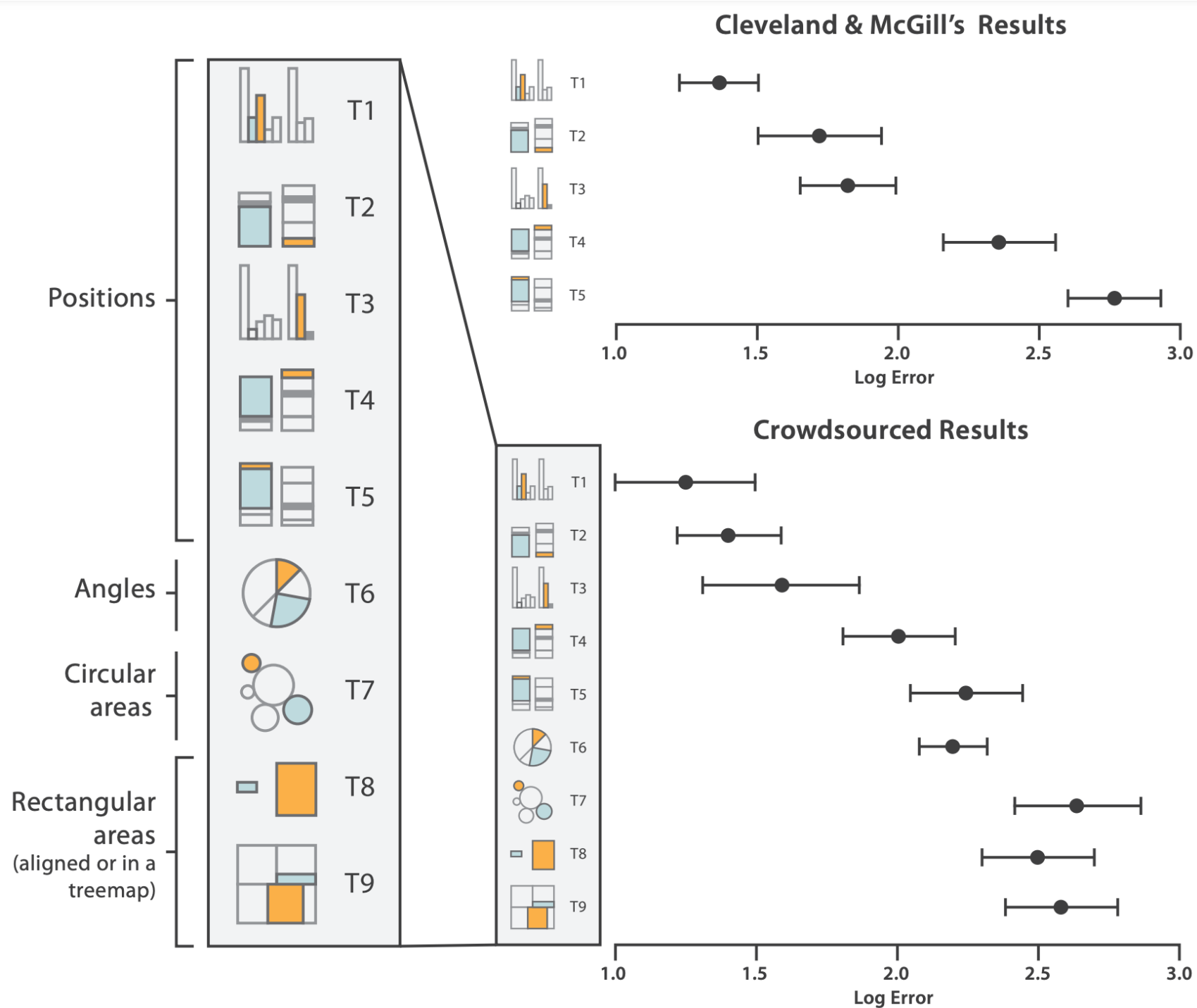


Shape



Color: hue-saturation-brightness (HSB)





Source: **Tamara Munzer** (2014).
Visualization Analysis and Design.

How many 5s in this display?

1561321203658413076510374627
4173127527327592732990709742
1703707774179527931749270973
4019743217909370945179279417

How many 5s in this display?

1561321203658413076510374627
4173127527|327592732990709742
1703707774179527931749270973
4019743217909370945179279417

Numerals differ only in **shape**, and are high-level symbols

You have to literally scan them **all** & count the 5s.

The distinction of **color** is immediate & **pre-attentive**

You only have to scan & count the 5s.

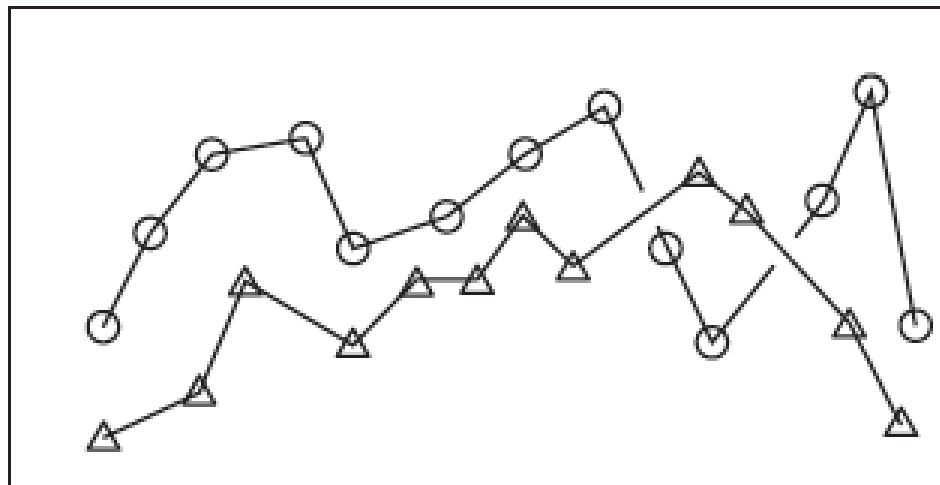
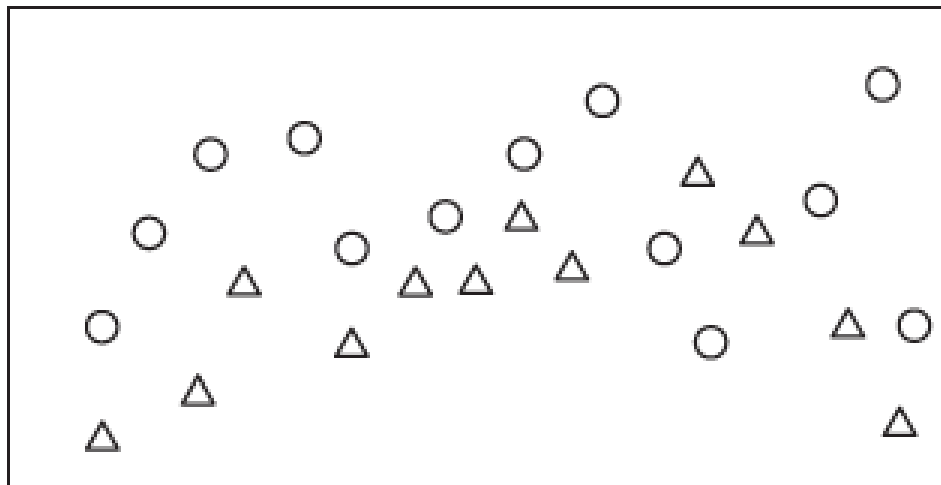
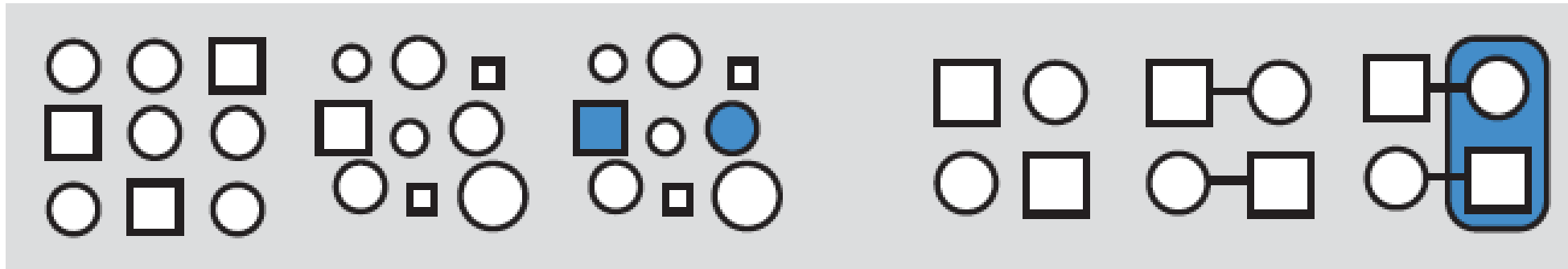
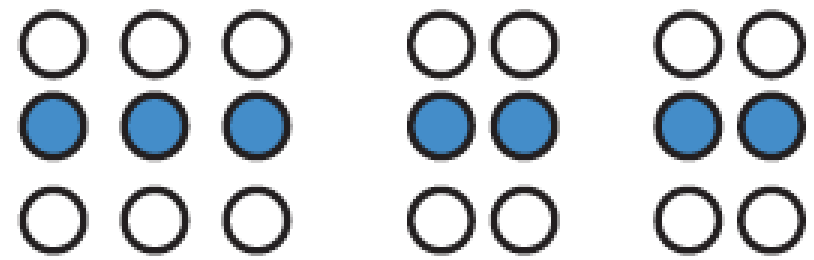
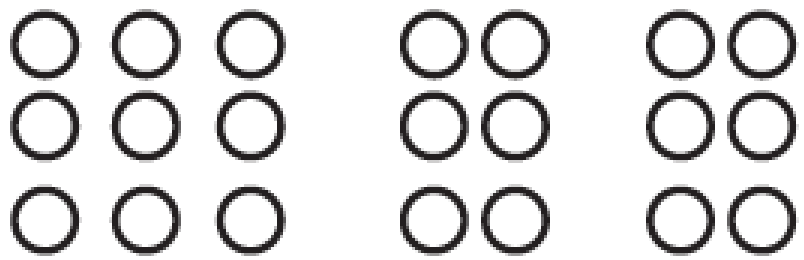
This is why **color** is an important visual attribute for a **categorical** variable in graphs



Gestalt principles of relatedness

- **Proximity:** Things that are spatially near to one another seem to be related.
- **Similarity:** Things that look alike seem to be related.
- **Connection:** Things that are visually tied to one another seem to be related.
- **Continuity:** Partially hidden objects are completed into familiar shapes.
- **Closure:** Incomplete shapes are perceived as complete.
- **Figure and ground:** Visual elements are taken to be either in the foreground or in the background.
- **Common fate:** Elements sharing a direction of movement are perceived as a unit.

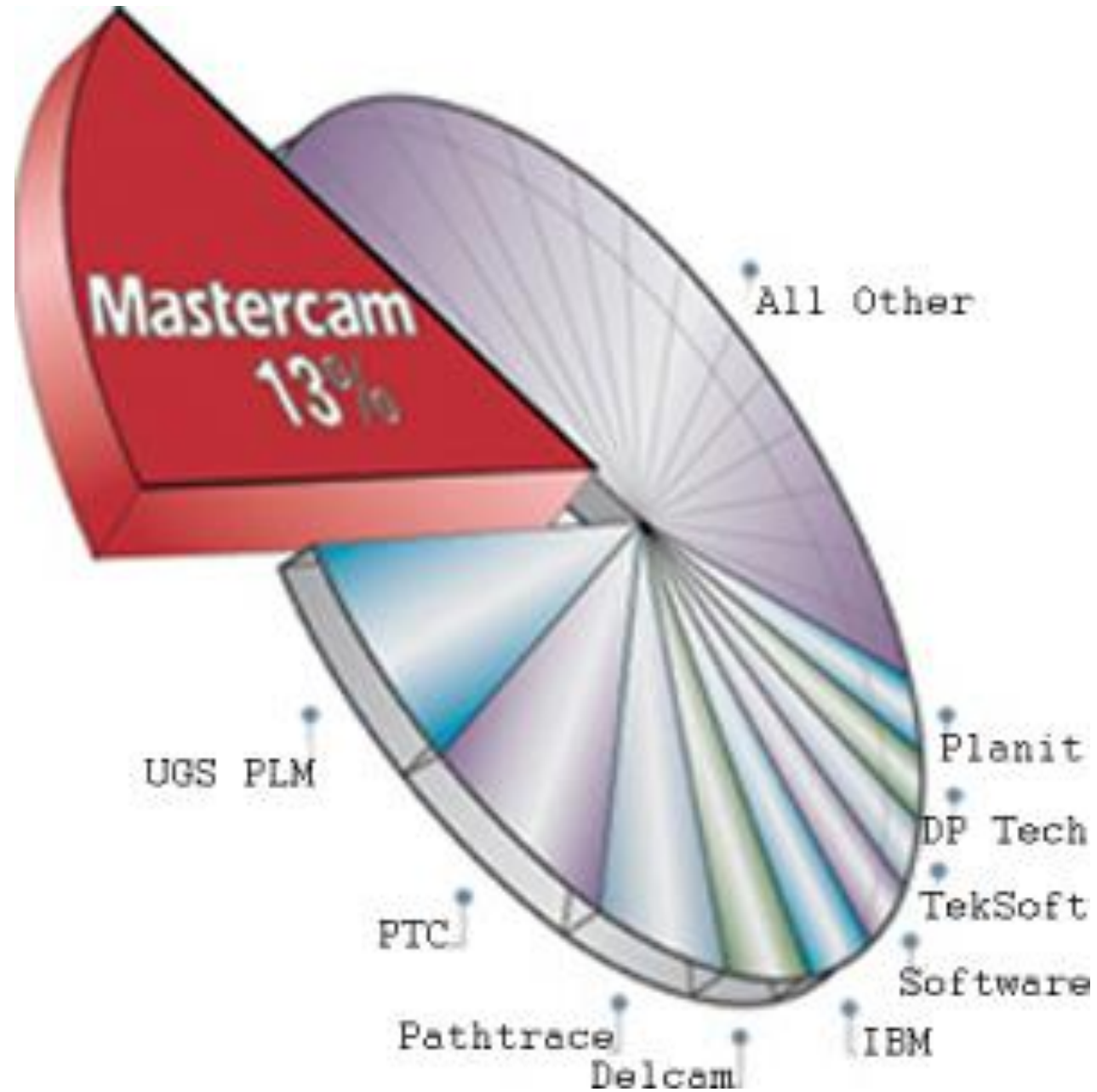




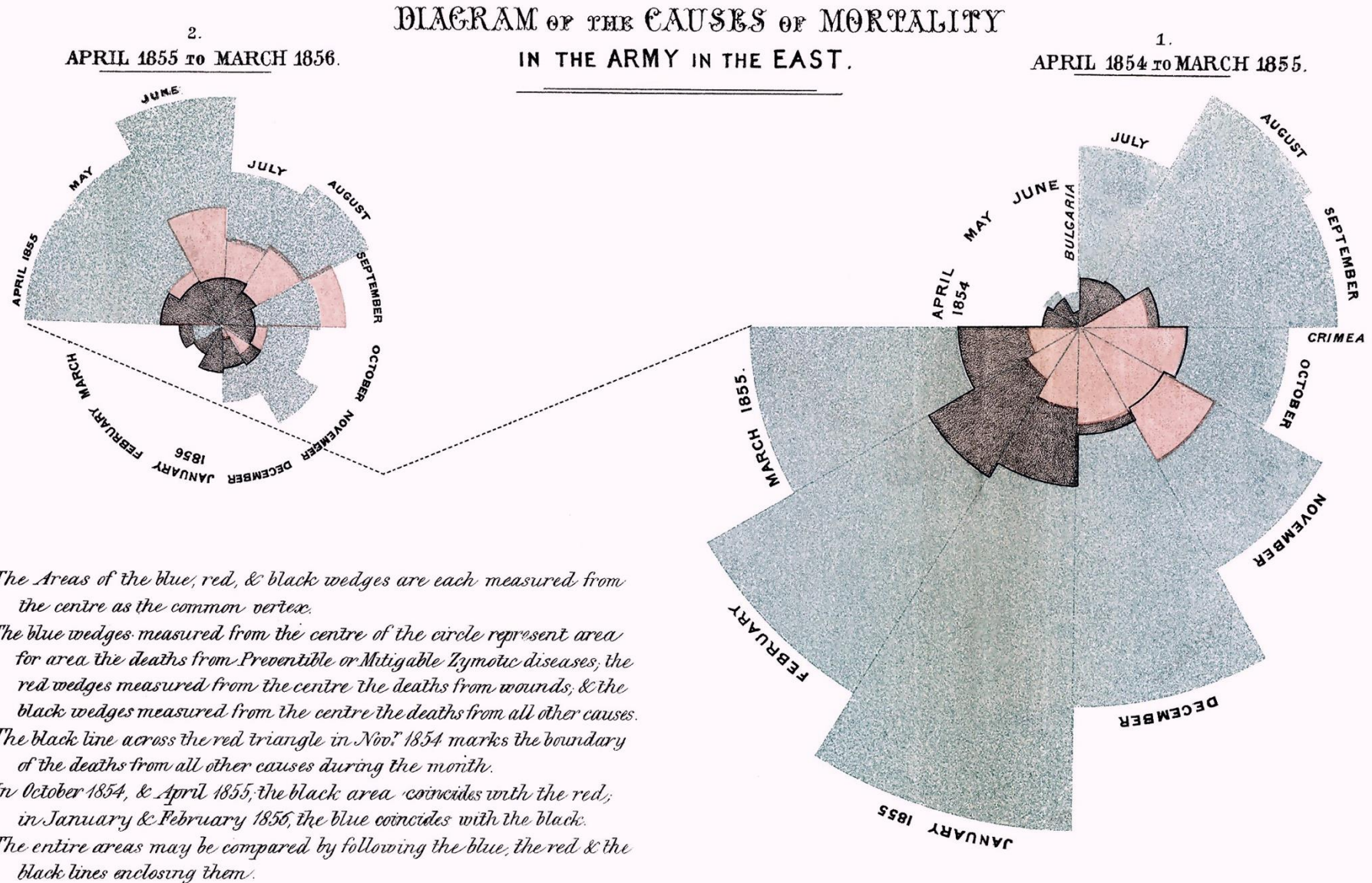
Some (distilled) principles from Tufte

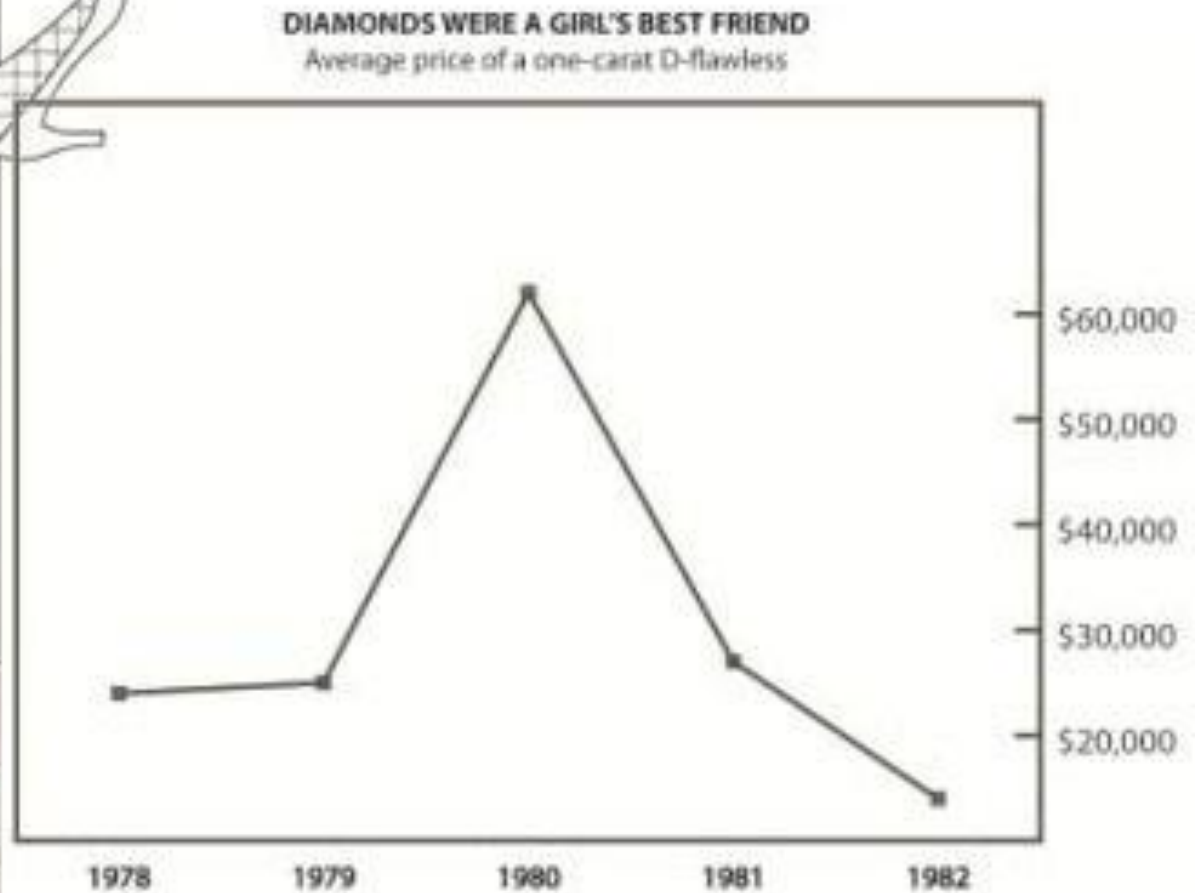
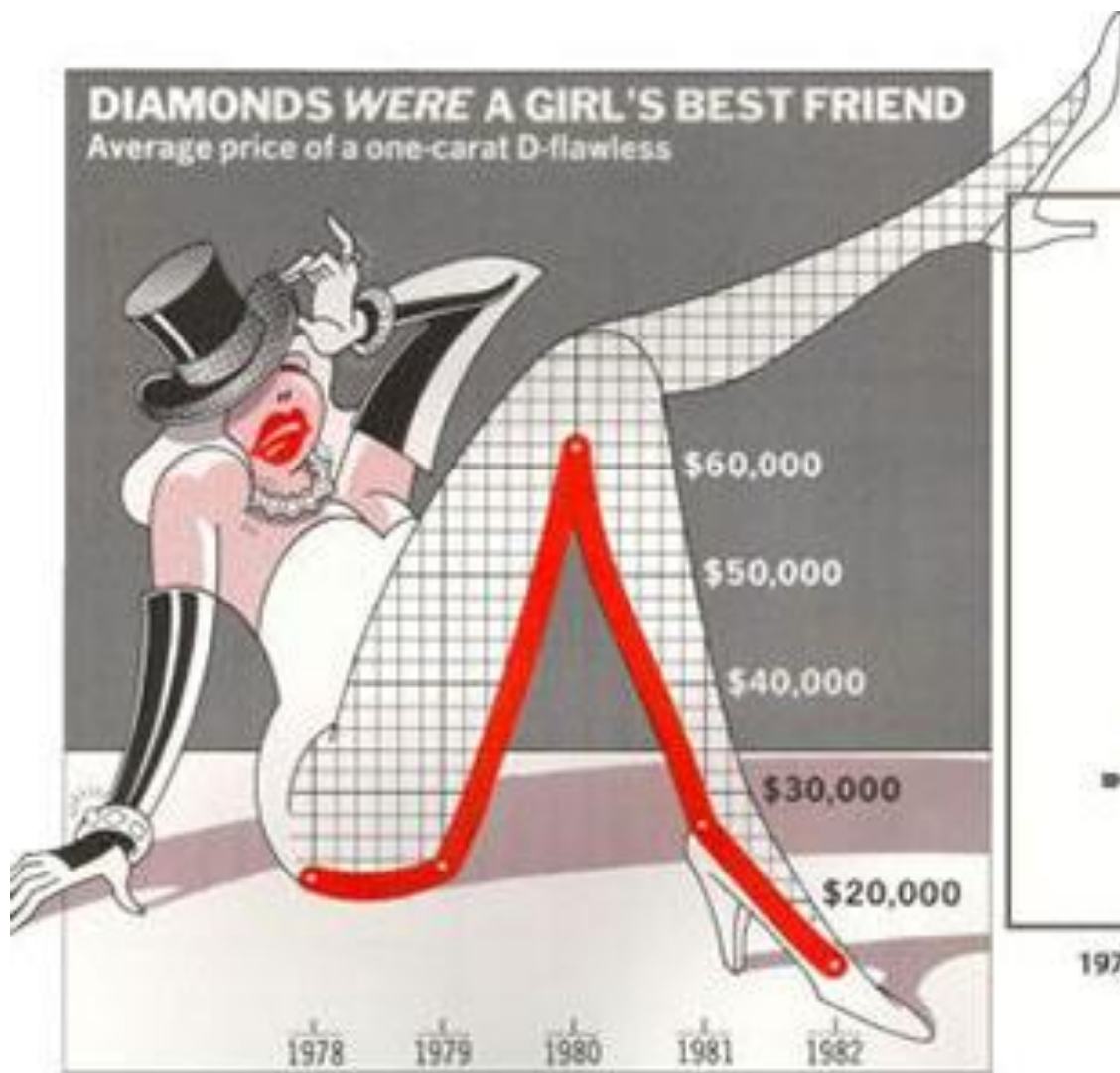


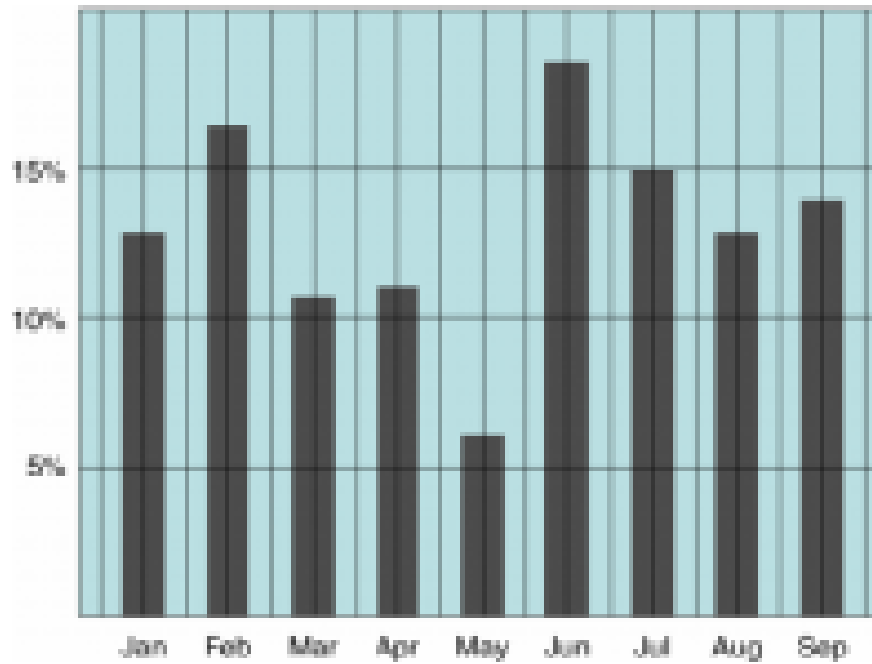
- Ask how data maps to perception
- Ask which comparisons you want, guide eye to those
- Maximize data-to-ink ratio
- Present more data (without losing interpretability)
- Use levels of detail
- (Remember narrative)



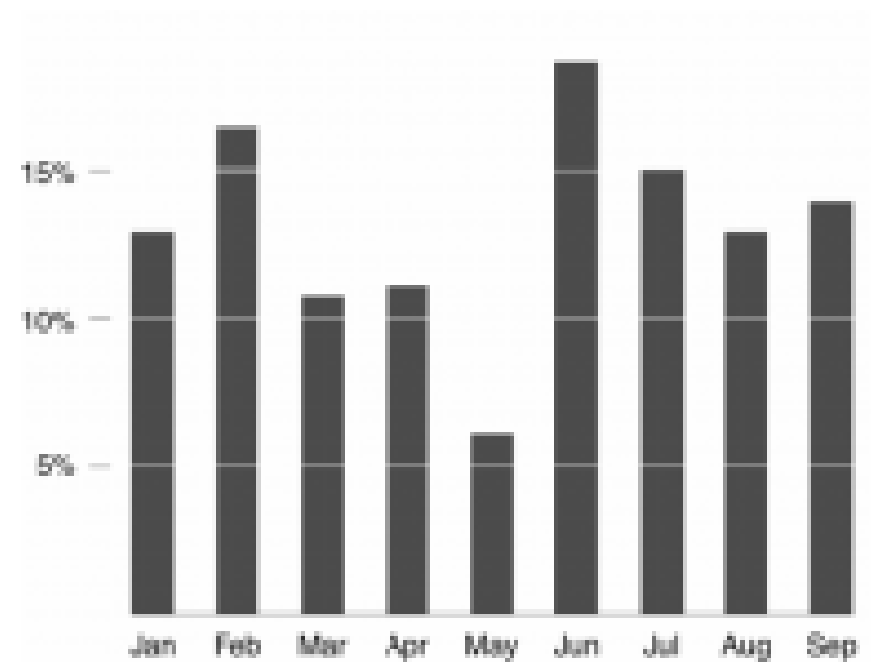
Nightingale Rose / Coxcomb chart



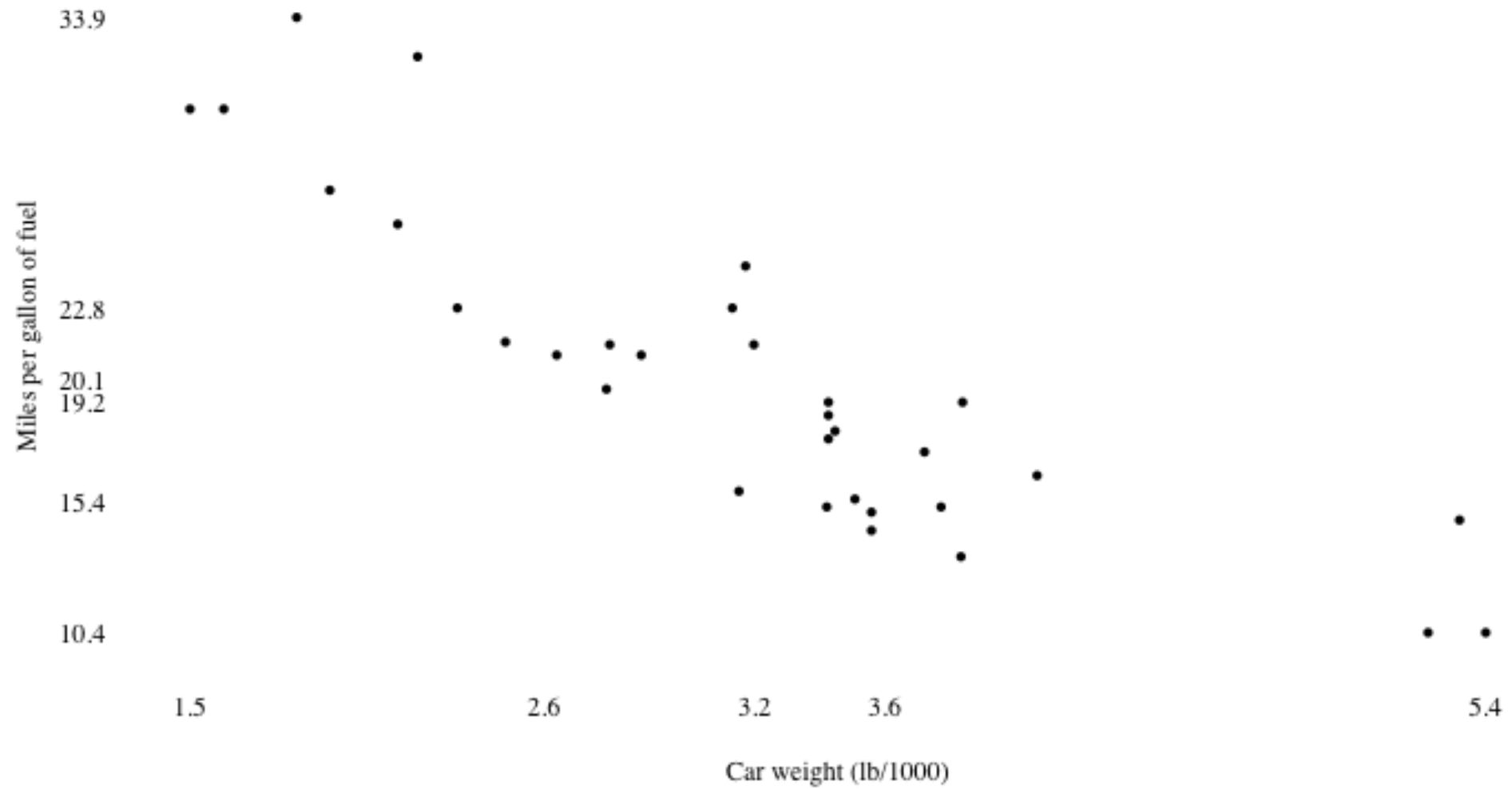


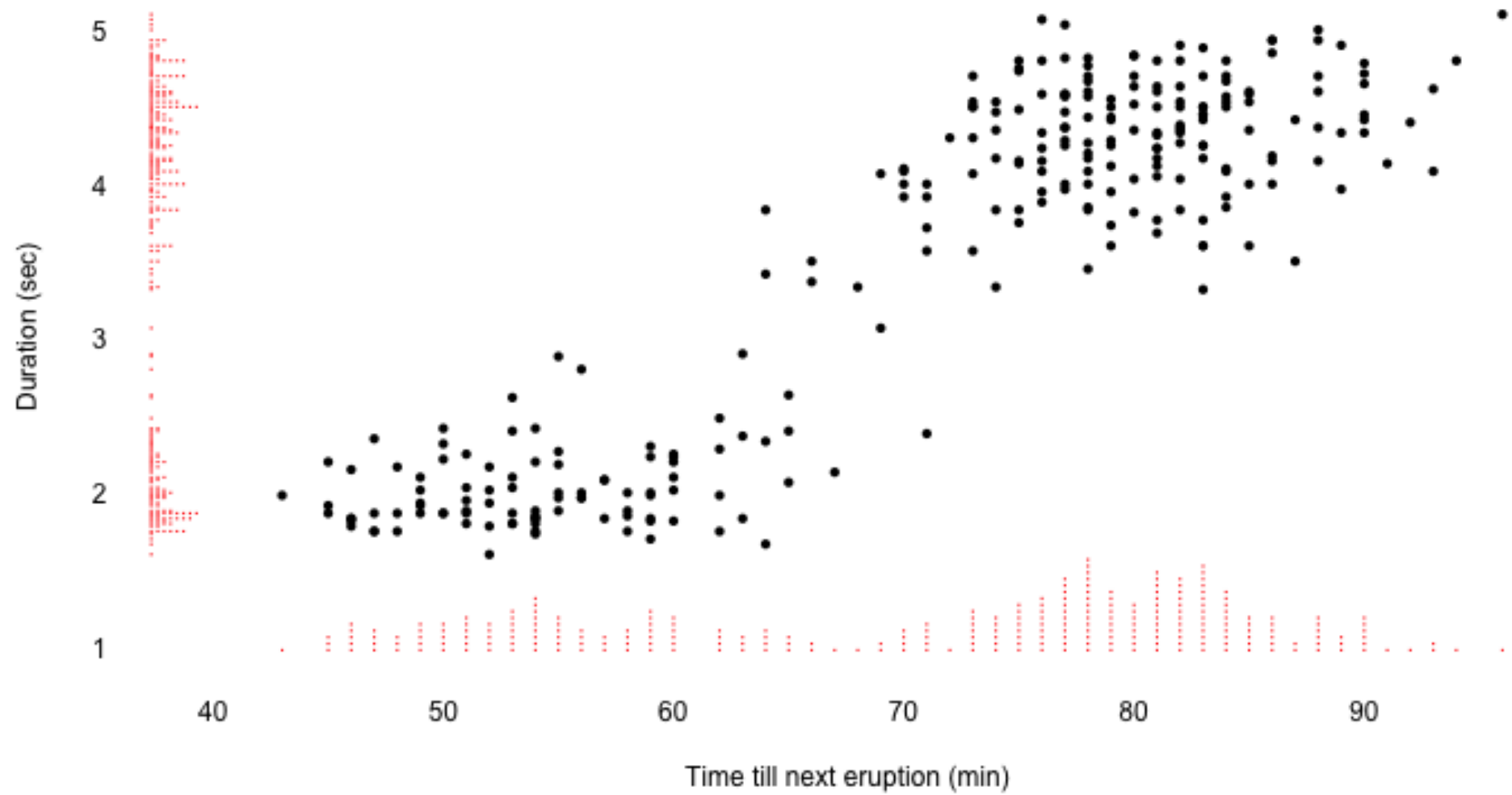


Low Data/Ink

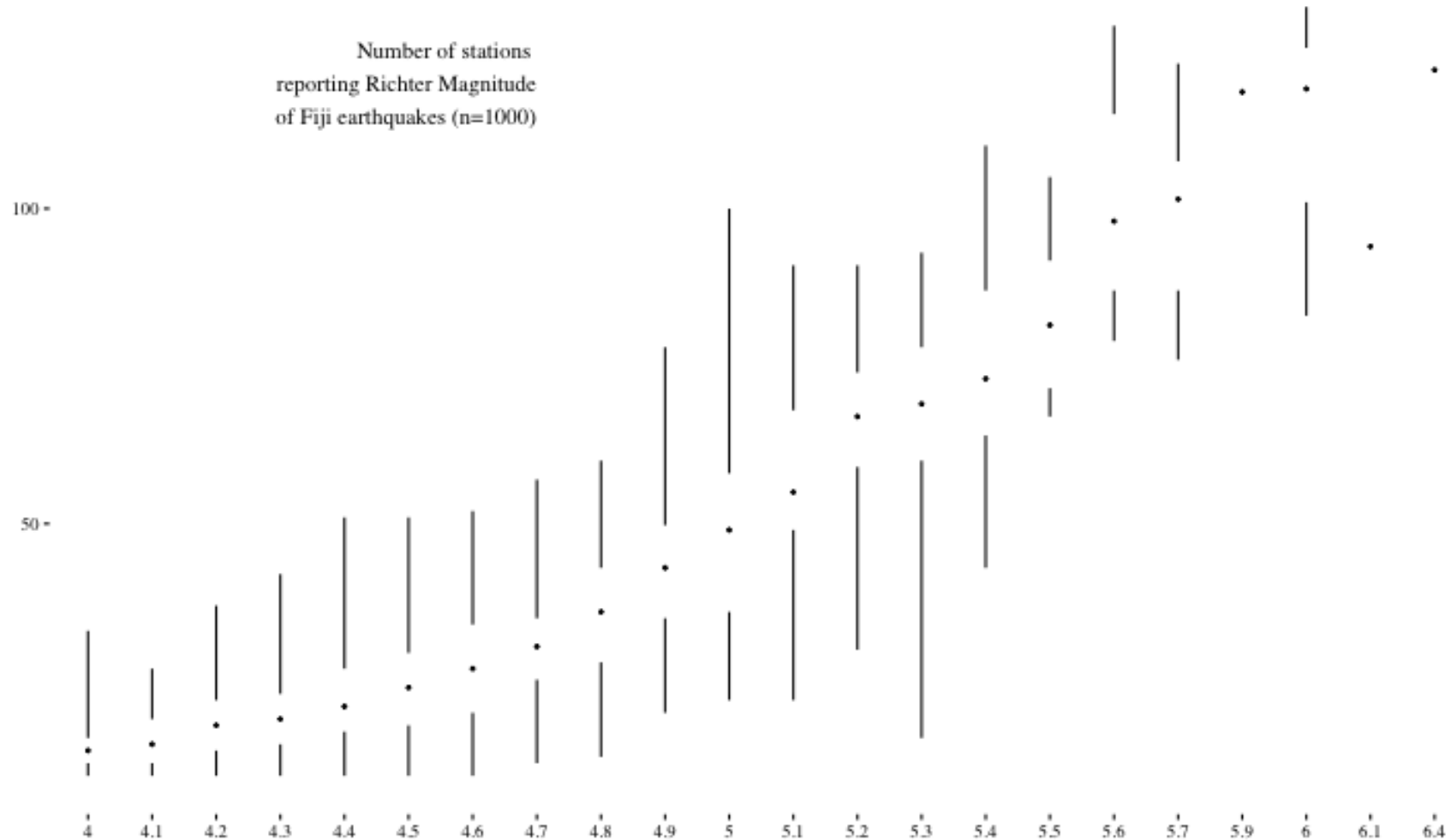


High Data/Ink





```
ggplot(quakes, aes(factor(mag), stations)) +
  theme_tufte() +
  geom_tufteboxplot(outlier.colour = "transparent") +
  theme(axis.title = element_blank())
```



Tufte wisdom

- Tufte's principles are more oriented to communication and can be taken too far
- Better data/ink → display more information without overload;
- Thinking about perception can help you choose better geoms, aesthetics.

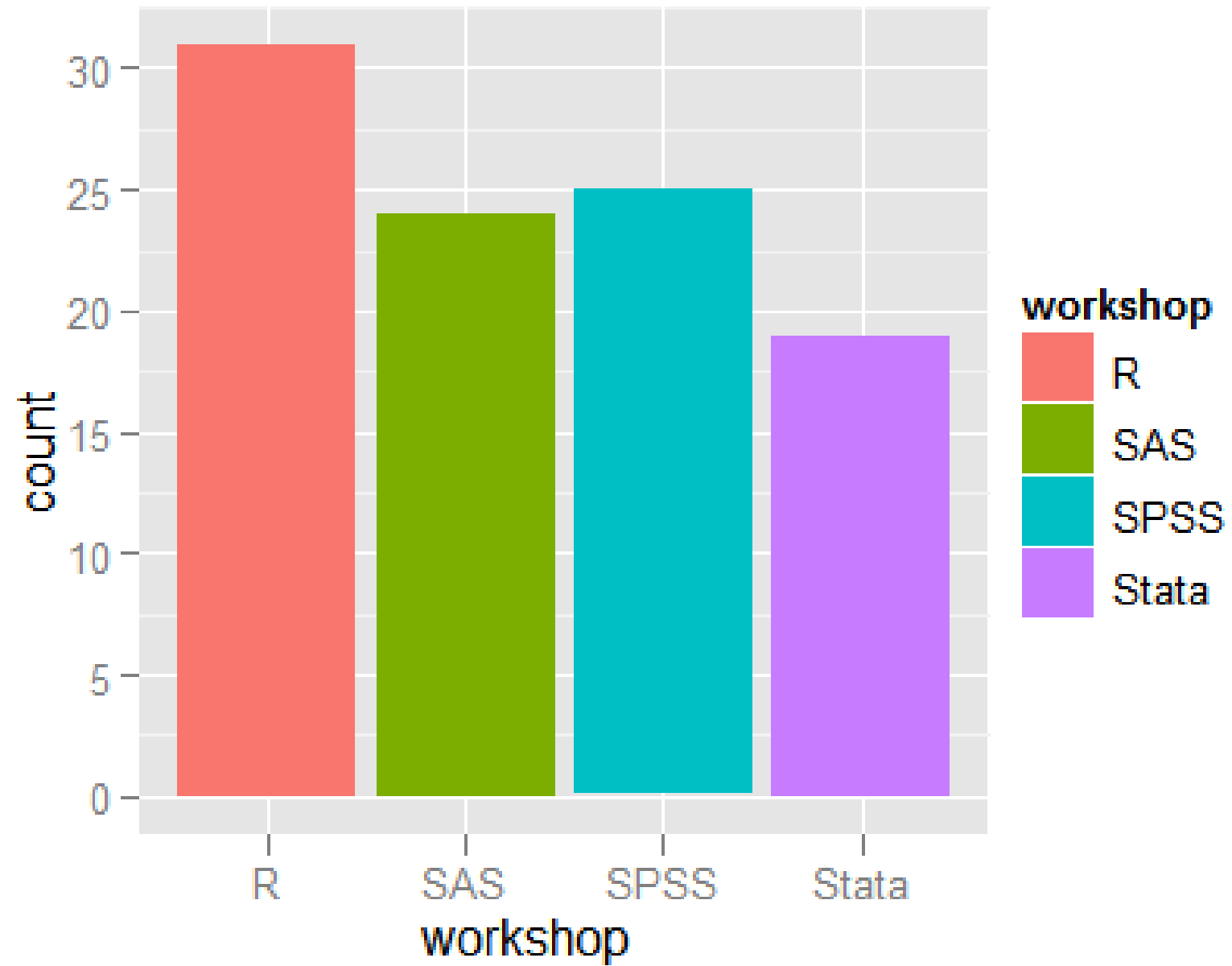


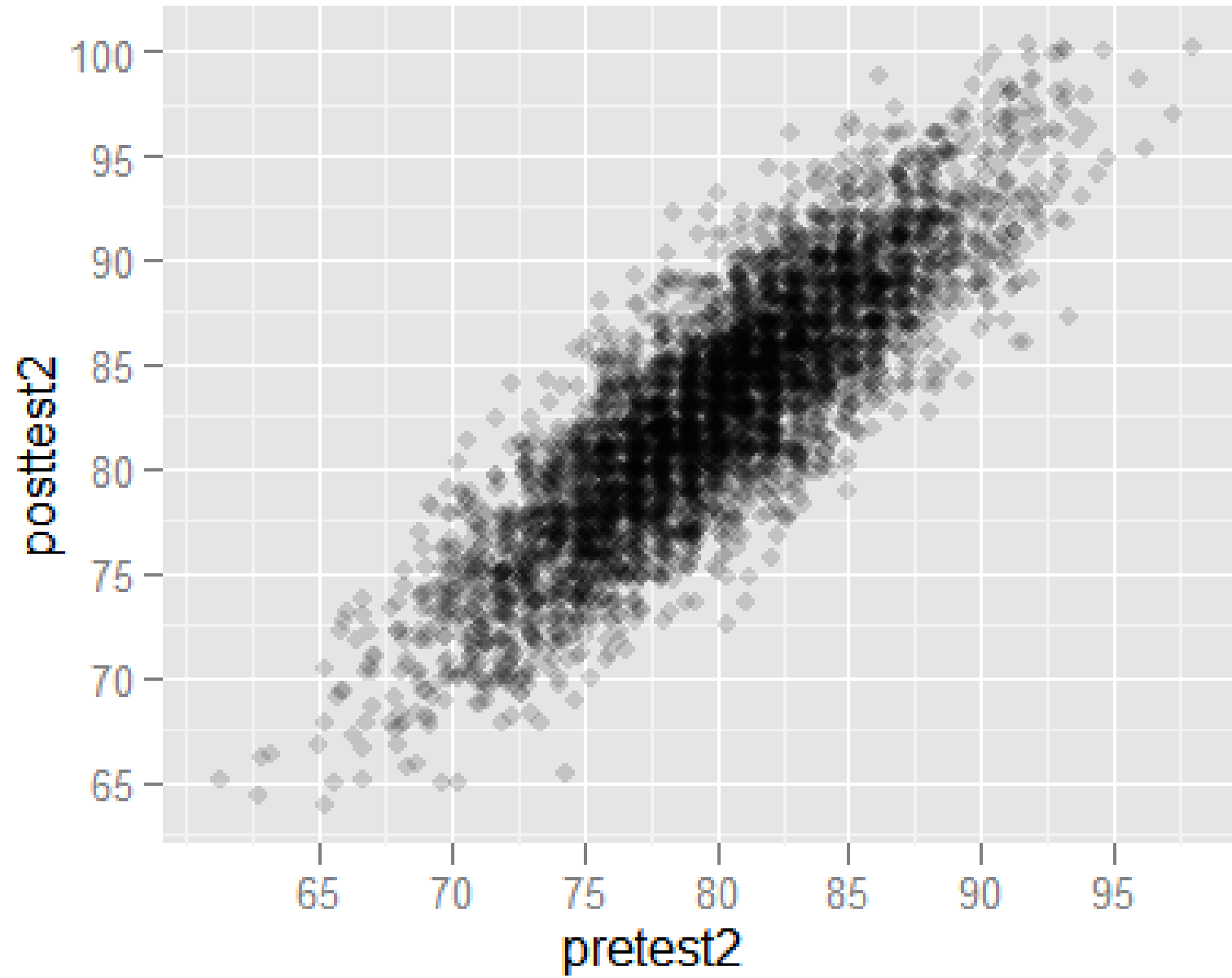
Some practice

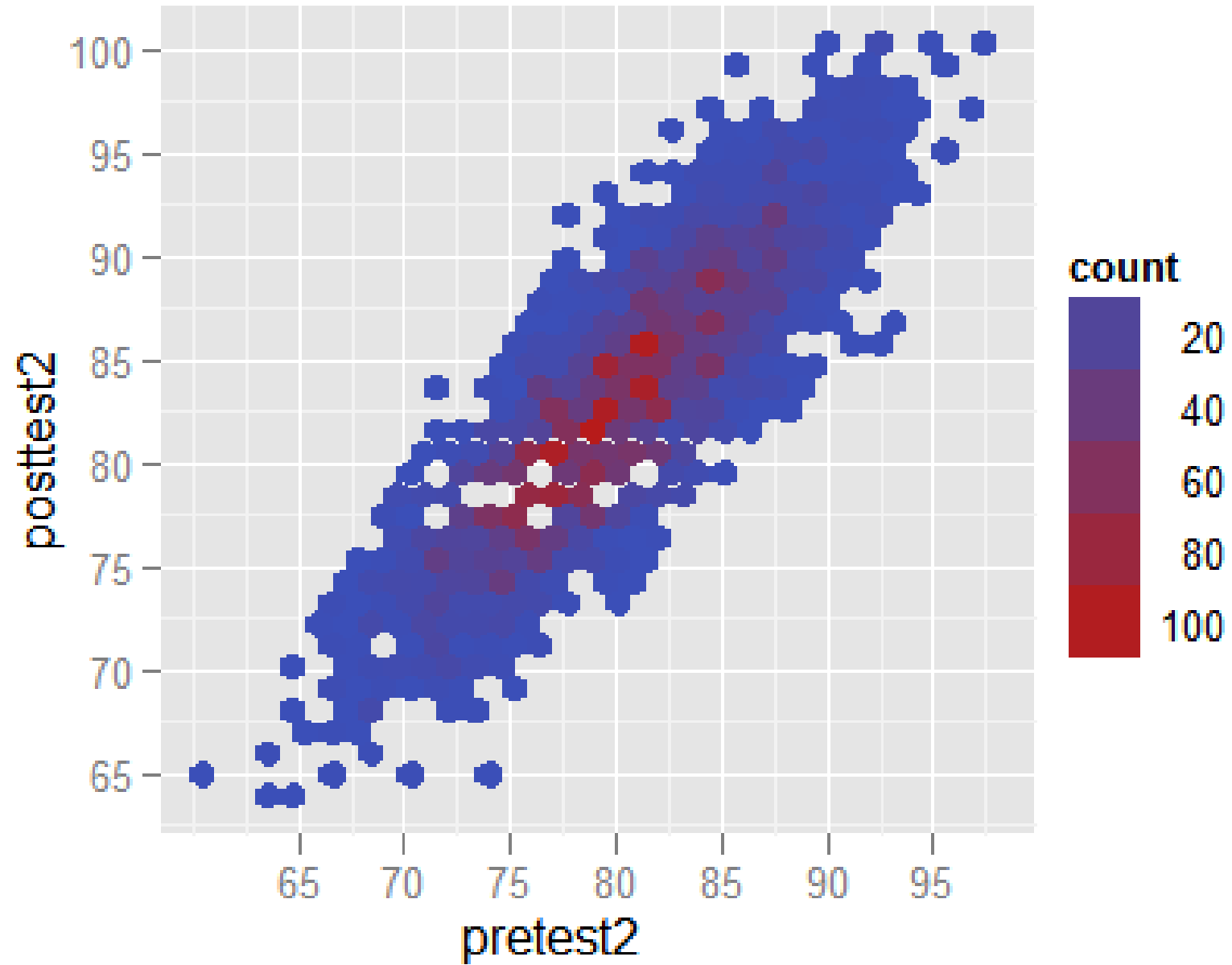
Answer these questions:

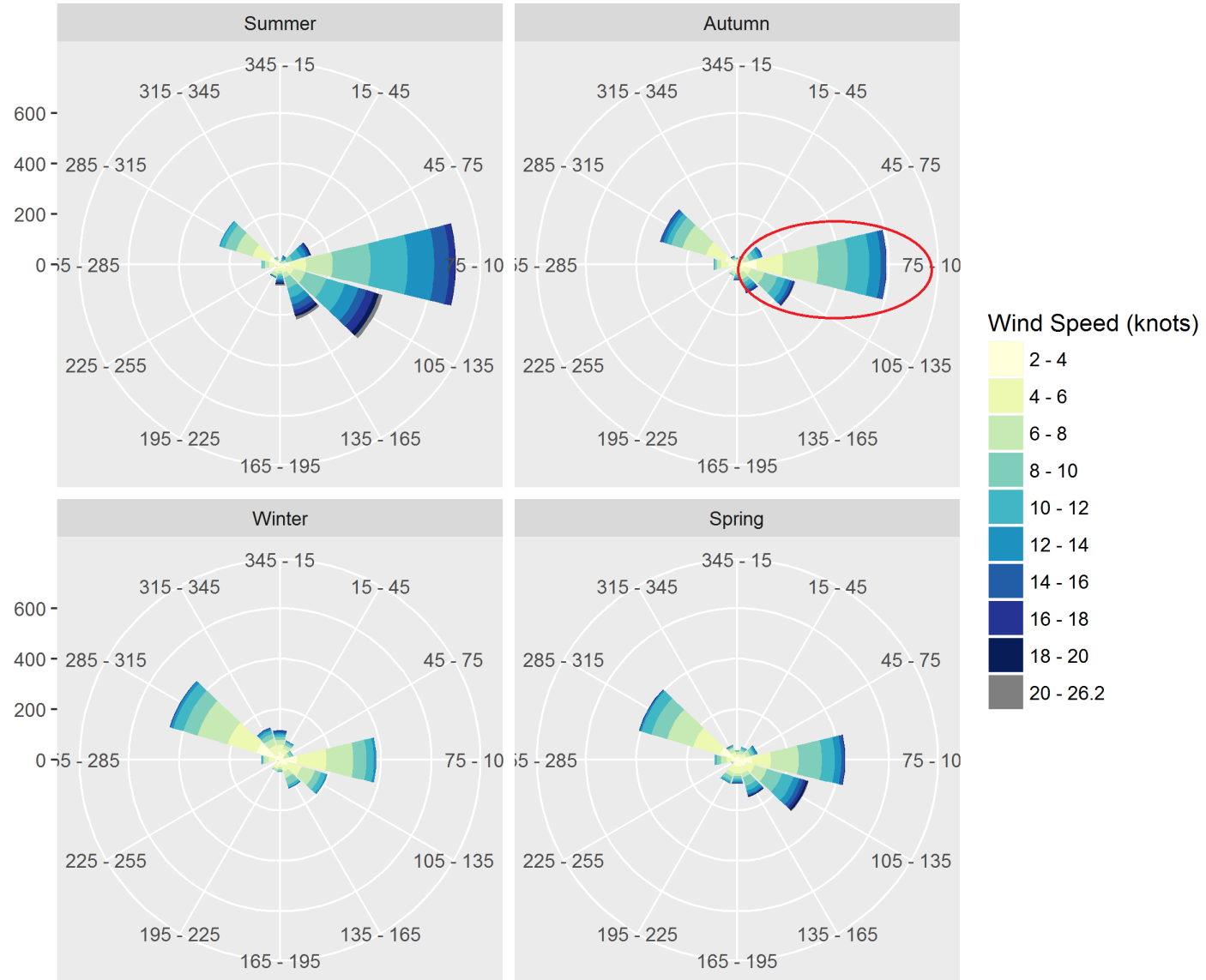
- Are we plotting the right thing?
- What are: aesthetics, geom, scale, facets, transformation, coordinate system
- How is data/ink?
- Is perception considered optimally?
- Can you think of questions you can't answer from this plot which are in the data?











Conclusion

Conclusion

- Data visualization is data analysis + psychology;
- Sticking to **basic principles** helps:
 - **Map data** to aesthetics, geoms, scales, facets;
 - Perception research guides choices;
 - **Which comparisons** do I want?
 - Maximize **data-ink** (within reason).
- Some standard recipes (e.g. “barplot”, “histogram”, “line graph”), but pros do not need the cookbook...
- Don’t believe everything you hear (“do’s and don’ts”)

