Data Wrangling and Data Analysis
Data Visualization

Daniel L. Oberski & Erik-Jan van Kesteren

Department of Methodology & Statistics

Utrecht University

This week

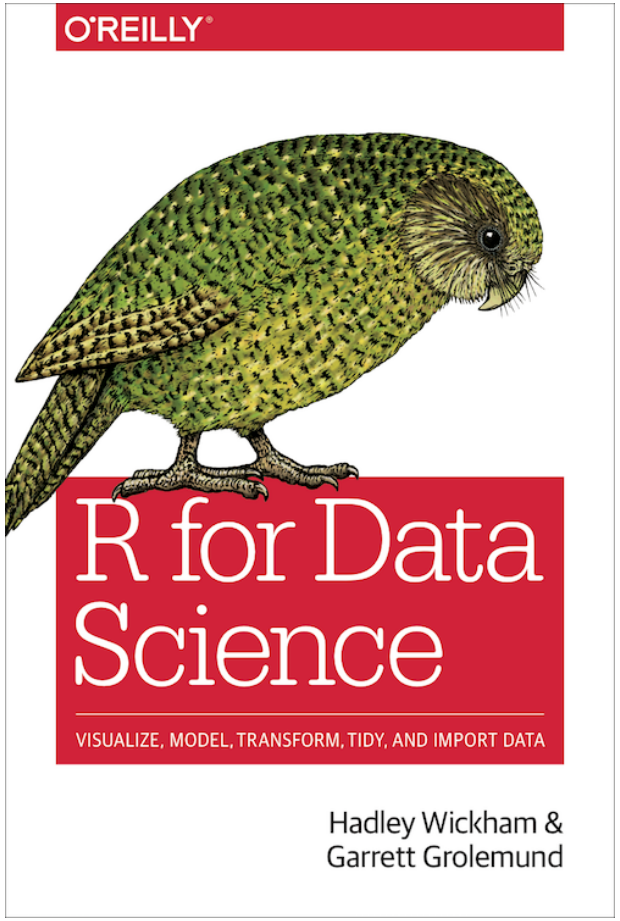
- **Data visualization principles**
- **The grammar of graphics**
- Exploratory data analysis (EDA)

- Goal of the week: know how to create and improve data visualizations, and know how to use them for exploration

Reading materials for this week

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

- Today: ch 3 visualization
- Tomorrow: chh 3, 5, 7



Assignments this week

- Monday: Tidy data visualization with dplyr and ggplot
- Tuesday: Exploratory data analysis
- Thursday: either (a) resit for the test, or (b) assignment on advanced data visualization

John Tukey (1915 – 2000)

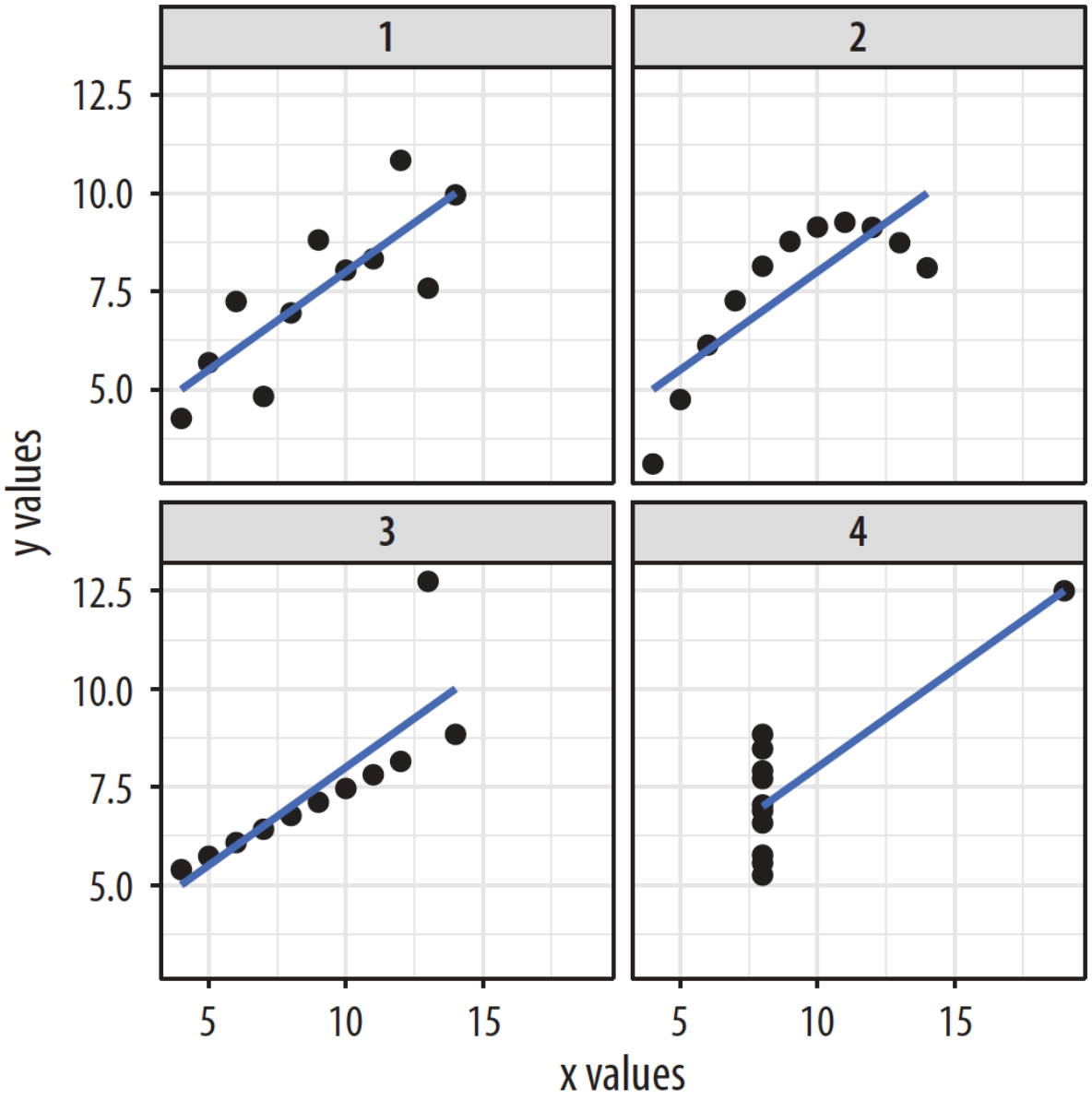
- Data Scientist patient zero
- Inventor of:
  - The boxplot
  - The term “exploratory data analysis”
  - The Fast Fourier Transform
  - “Tukey’s test”
  - The word “bit”
  - So, so much more (Wikipedia)

Today: visualization principles

Some data visualization principles

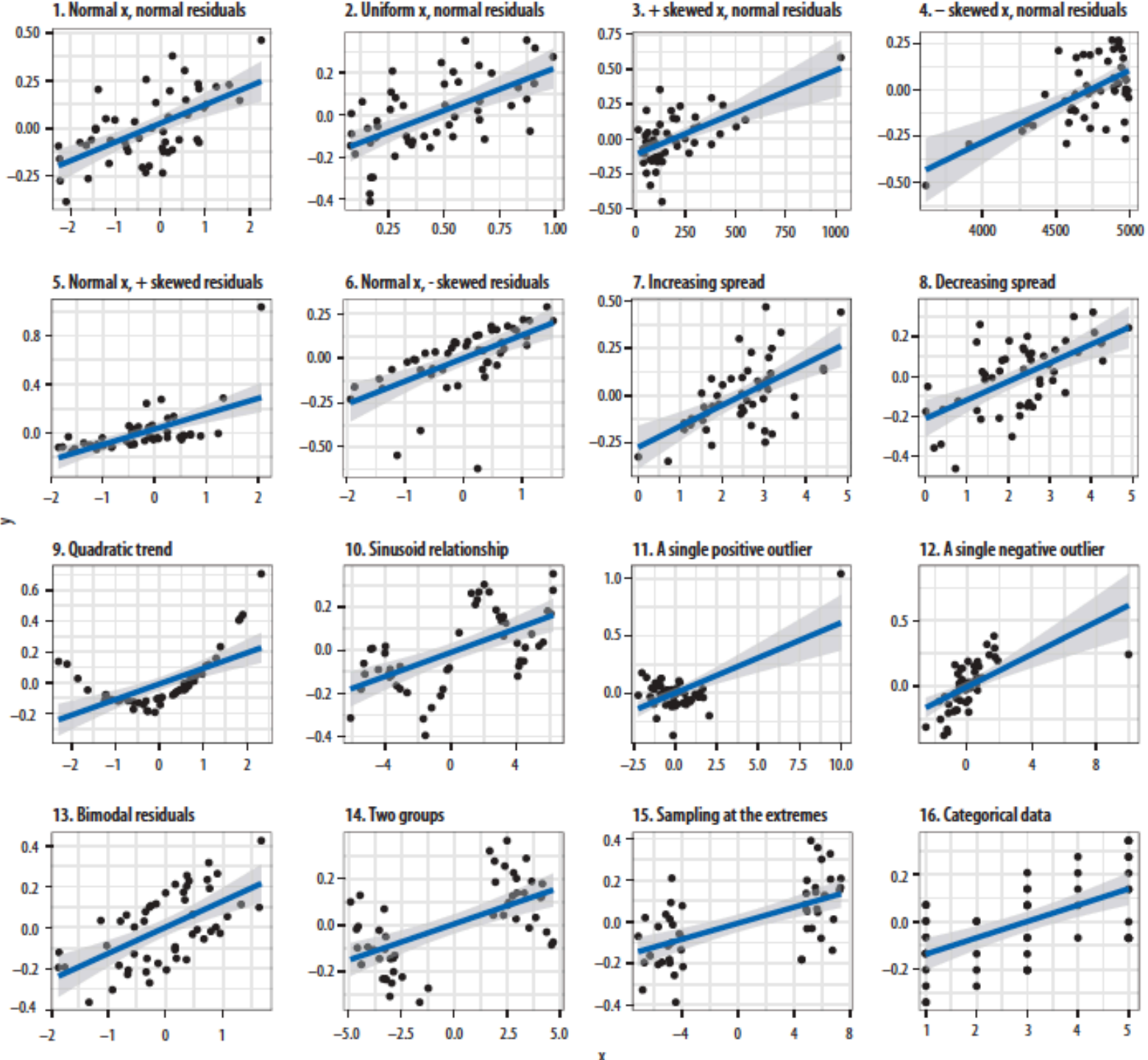
Data visualization

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



Anscombe's quartet

Source: Healy (2019)



Source: Healy (2019)

Graphics for data analysis

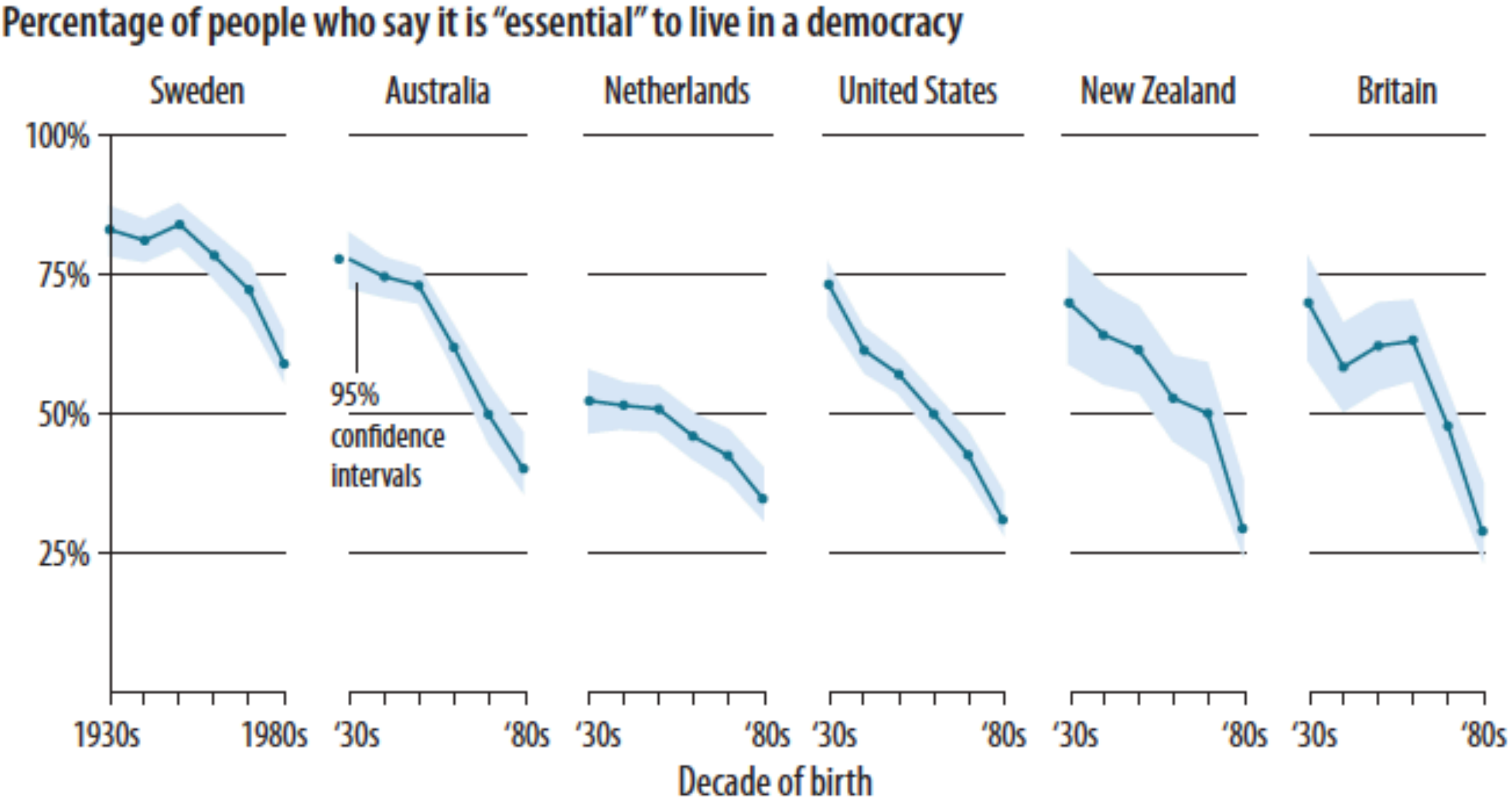
- The **human retina** can transfer around 10^6 or 10^7 bits per second to the brain;
- **Reading** transfers about 3 words, so $\sim 10^2$ or 10^3 bits/s;
- Potentially (!) visualization is about 4 orders of magnitude more powerful.

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

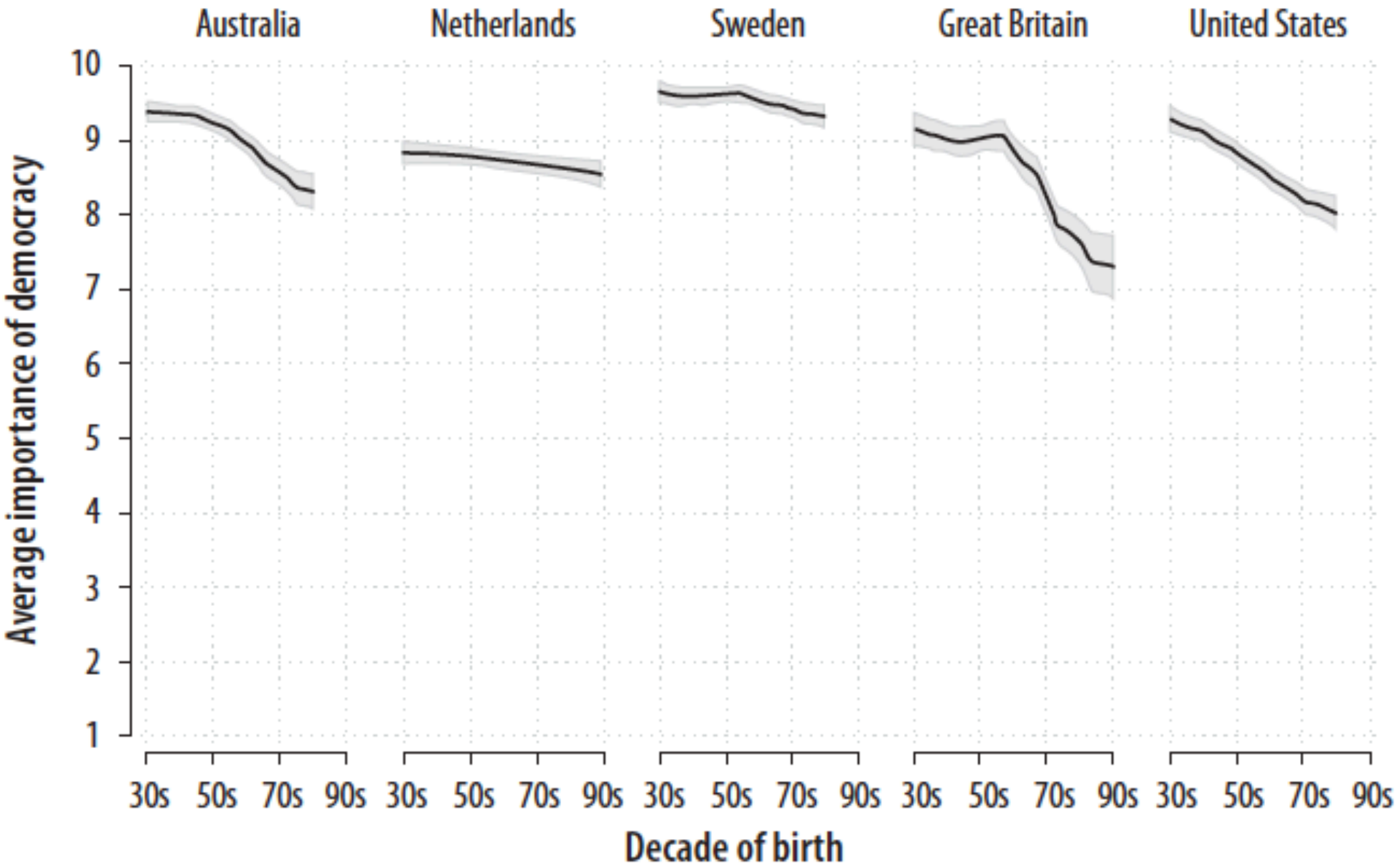
Plotting the right thing

Most common problems:

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



Source: Healy (2019)

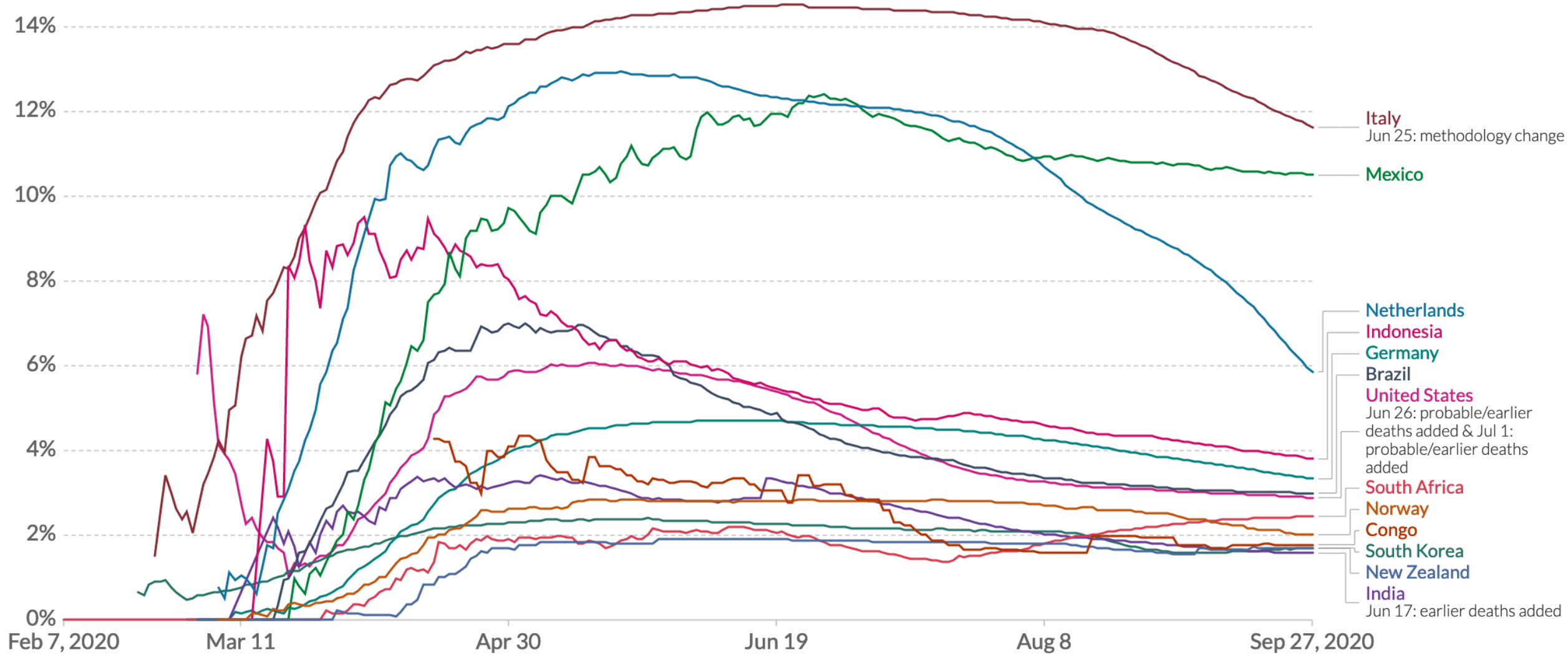


Source: Healy (2019)

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

Our World in Data

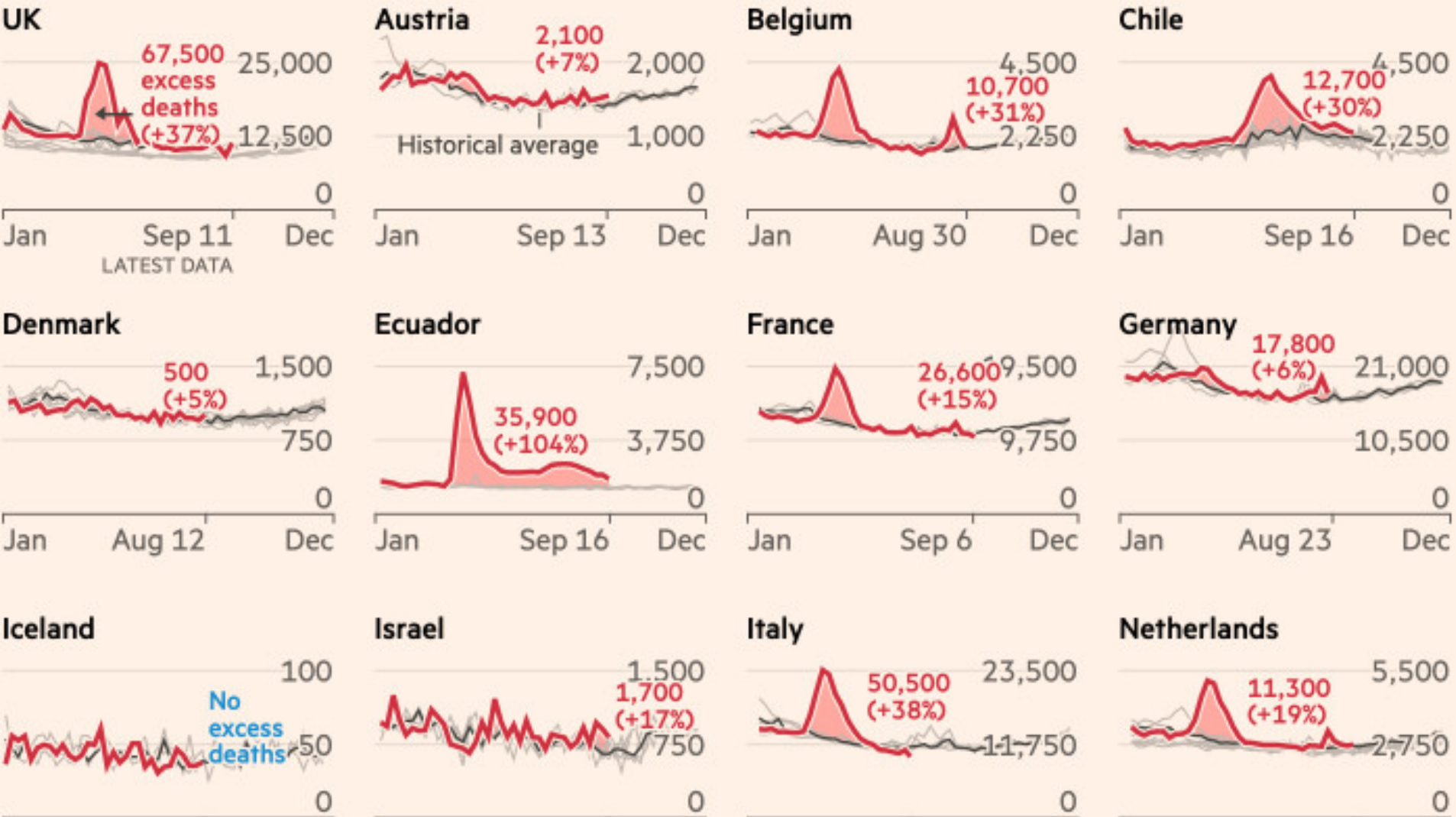


Source: European CDC – Situation Update Worldwide – Last updated 27 September, 10:05 (London time)

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

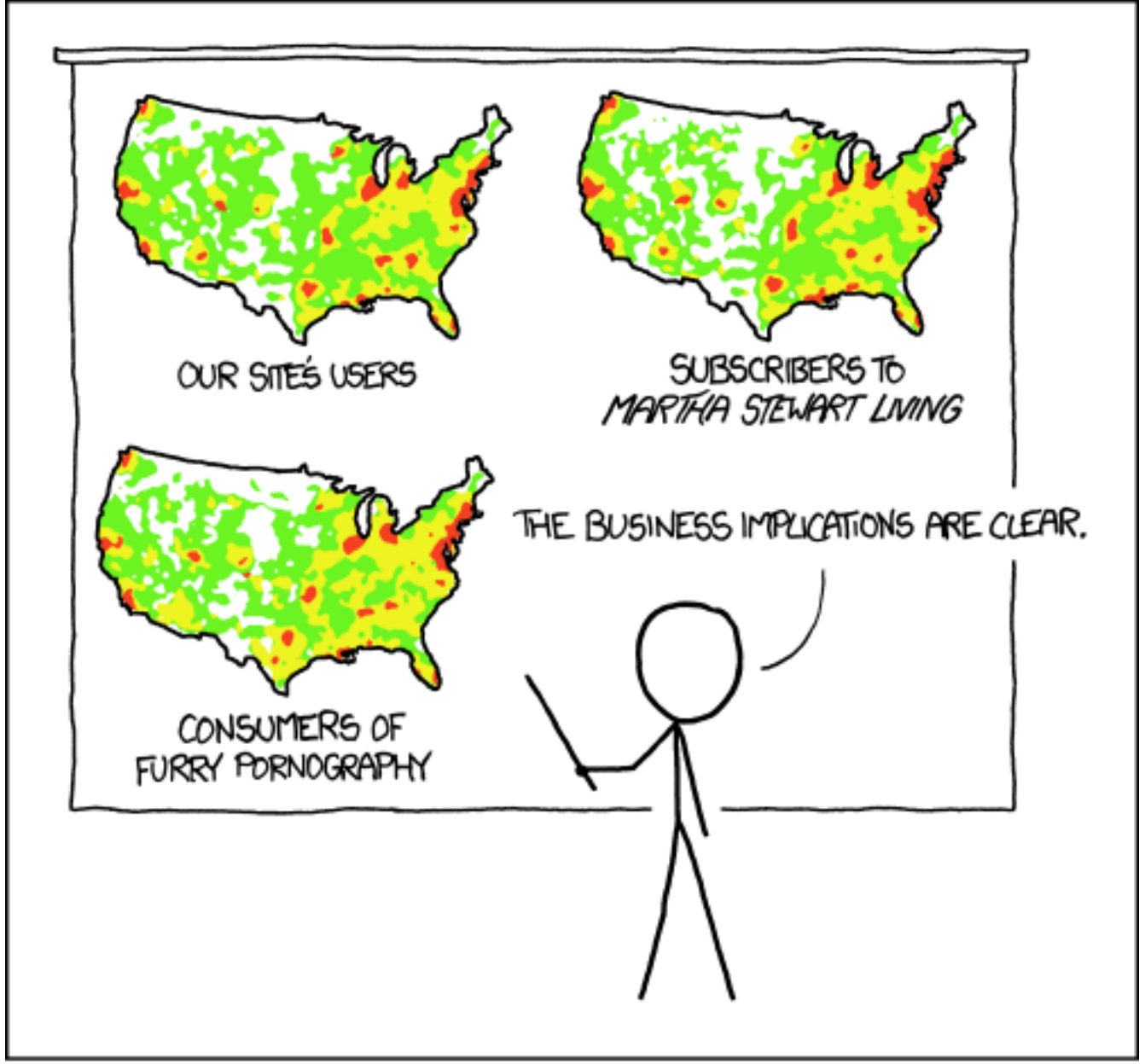
Number of deaths per week from all causes, 2020 vs recent years:

Shading indicates total excess deaths during outbreak



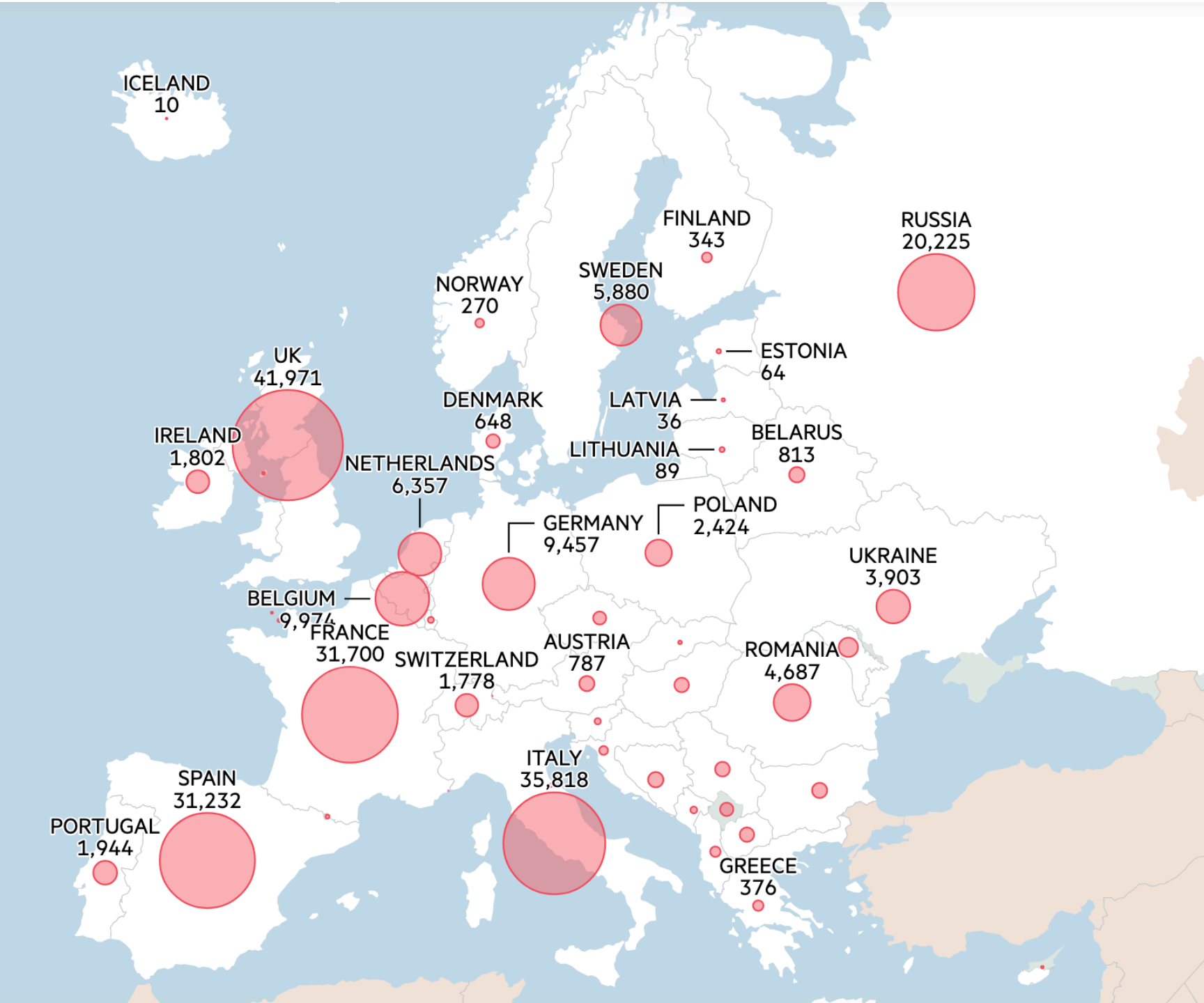
Good example (FT)

https://www.ft.com/content/a2
8-5eb7-4633-b89c-cbdf5b38693



PET PEEVE #208:

GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS



Missing a baseline…?

https://www.ft.com/content/a2901ce8-5eb7-4633-b89c-cbdf5b386938

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, ...

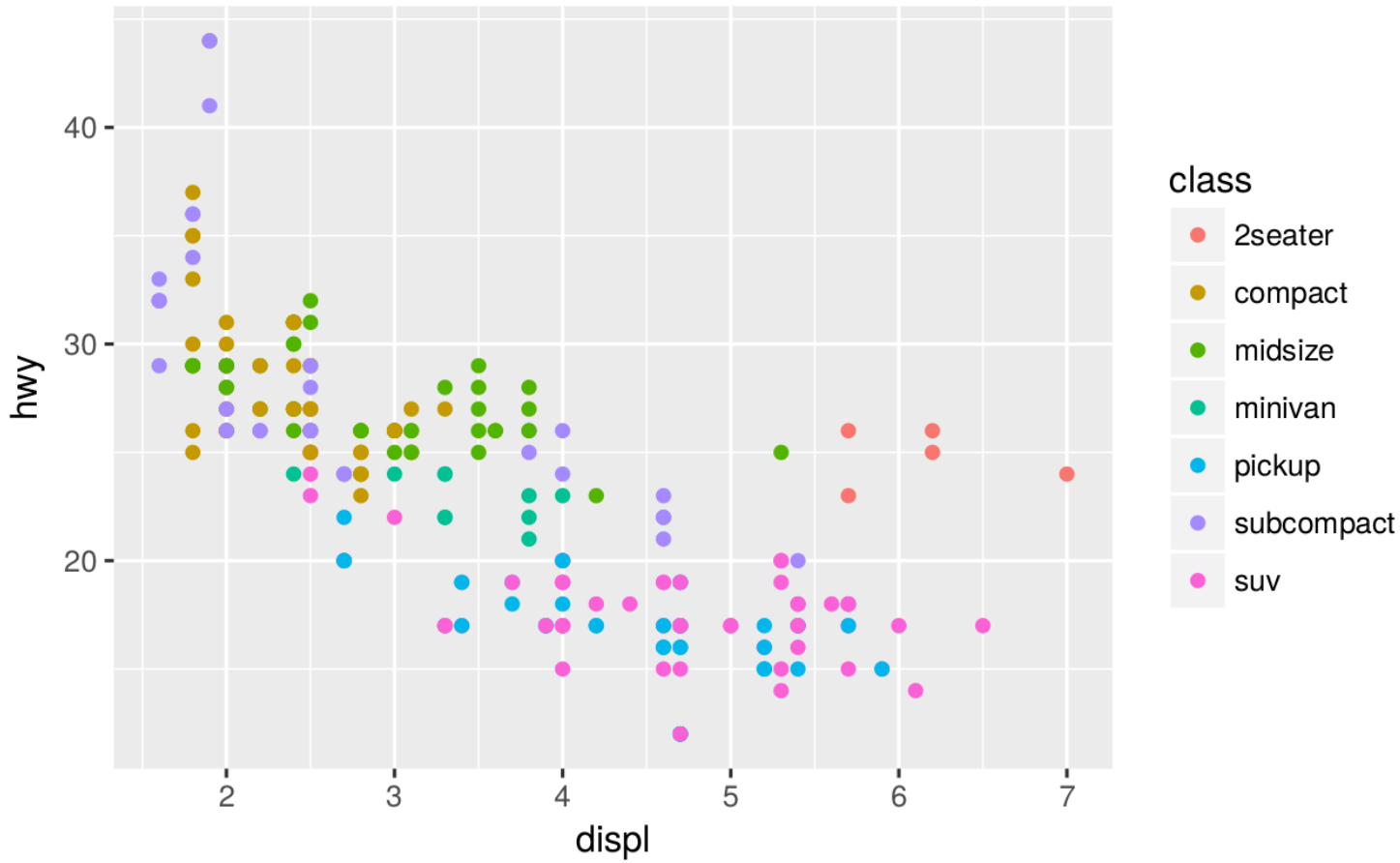
Grammar of graphics (Wickham version)

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

Example data set: cars

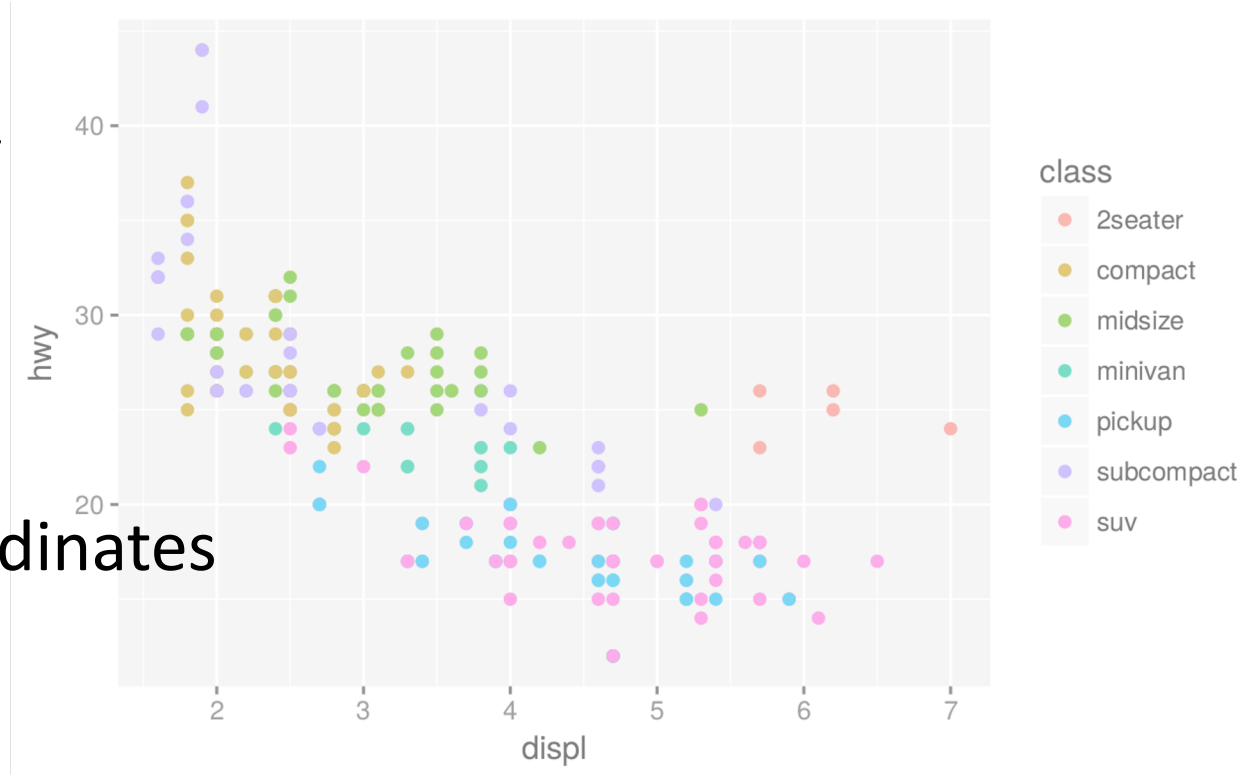
```
mpg
#> # A tibble: 234 × 11
#>    manufacturer model displ  year   cyl      trans   drv   cty   hwy    fl
#>           <chr> <chr> <dbl> <int> <int>      <chr> <chr> <int> <int> <chr>
#> 1          audi    a4   1.8  1999     4   auto(l5)     f    18    29     p
#> 2          audi    a4   1.8  1999     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))
```



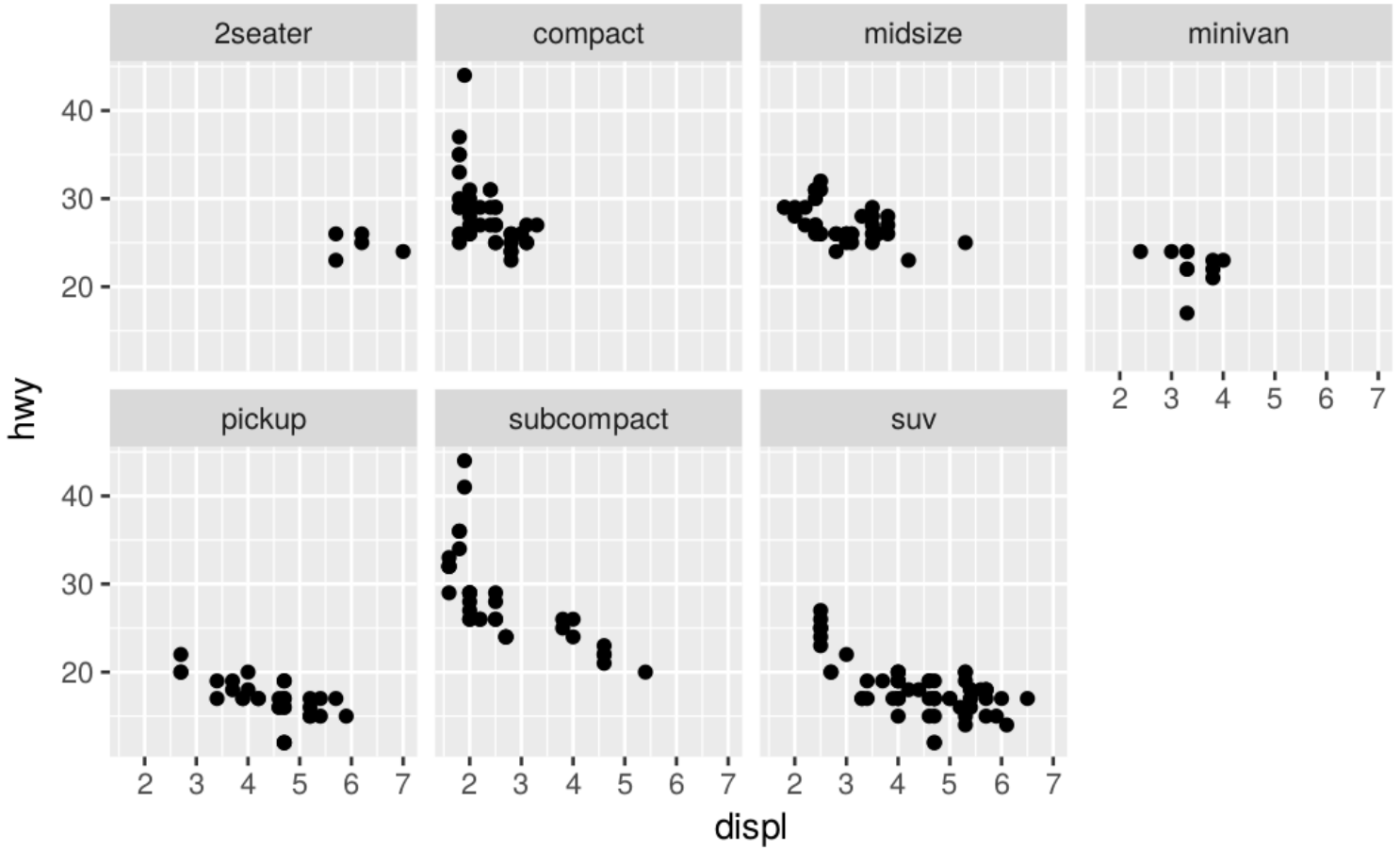
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy,
                           color = class))
```

- 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

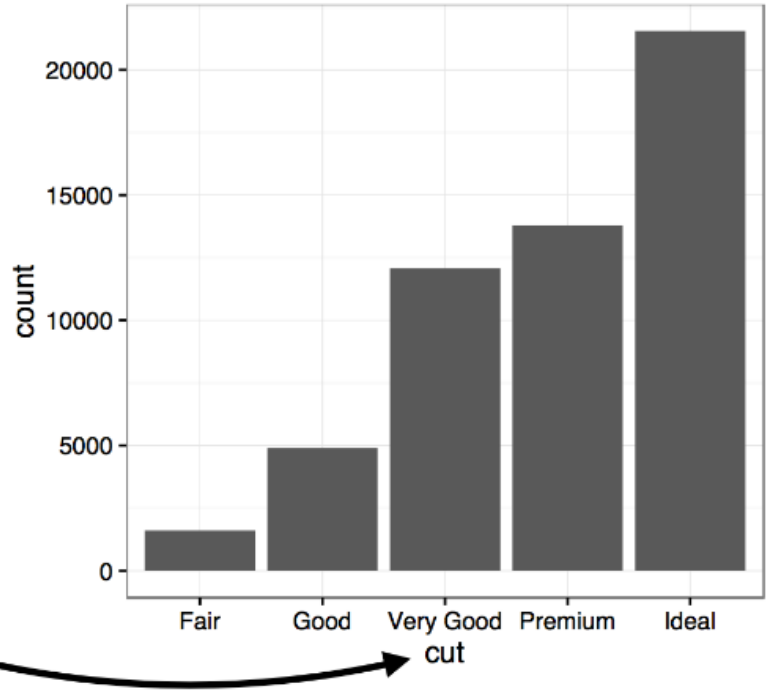
2. **`geom_bar()`** transforms the data with the "count" stat, which returns a data set of cut values and counts.

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.

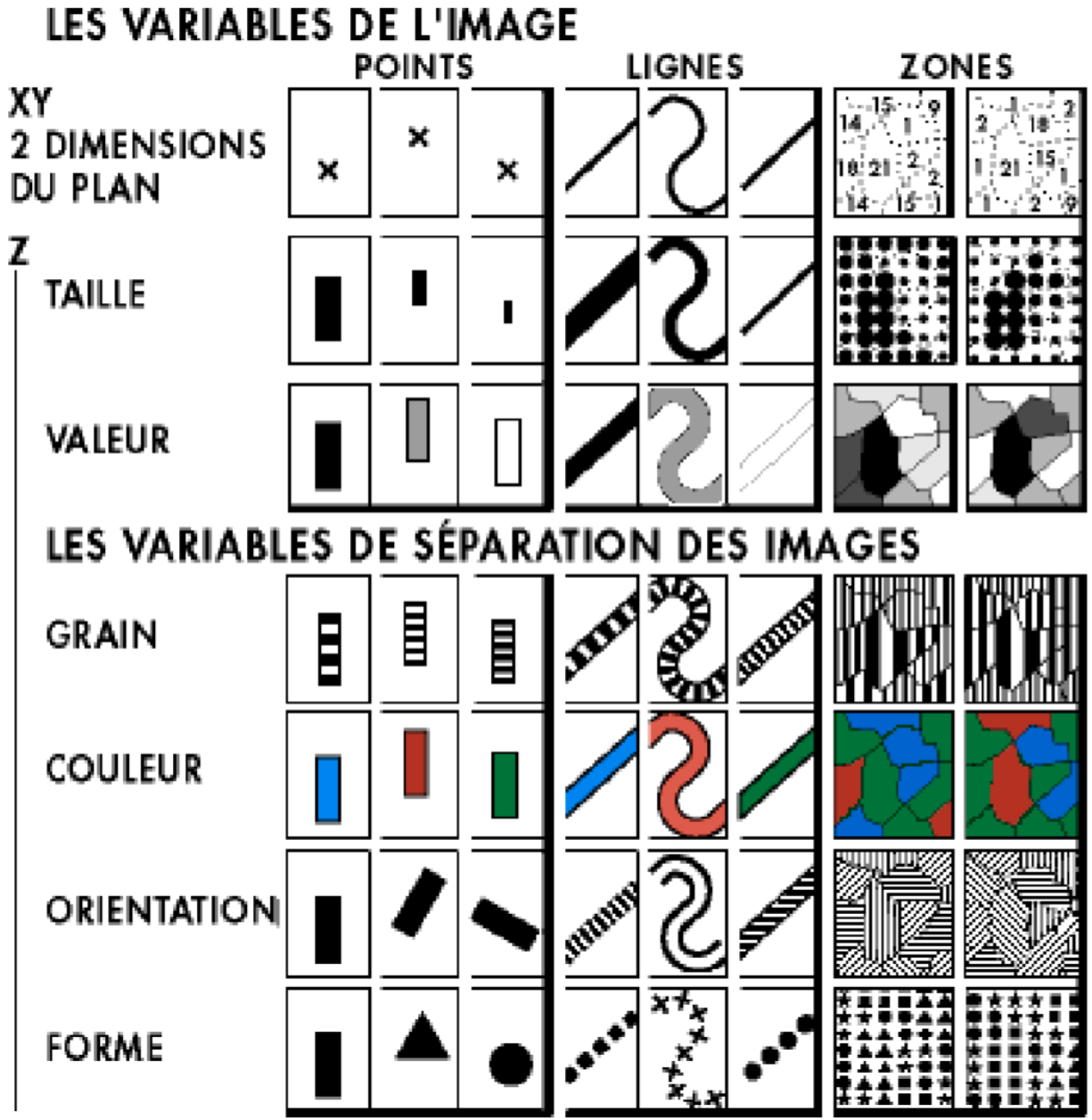
carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
…	…	…	…	…	…	…	…	…	…

stat_count()

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



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

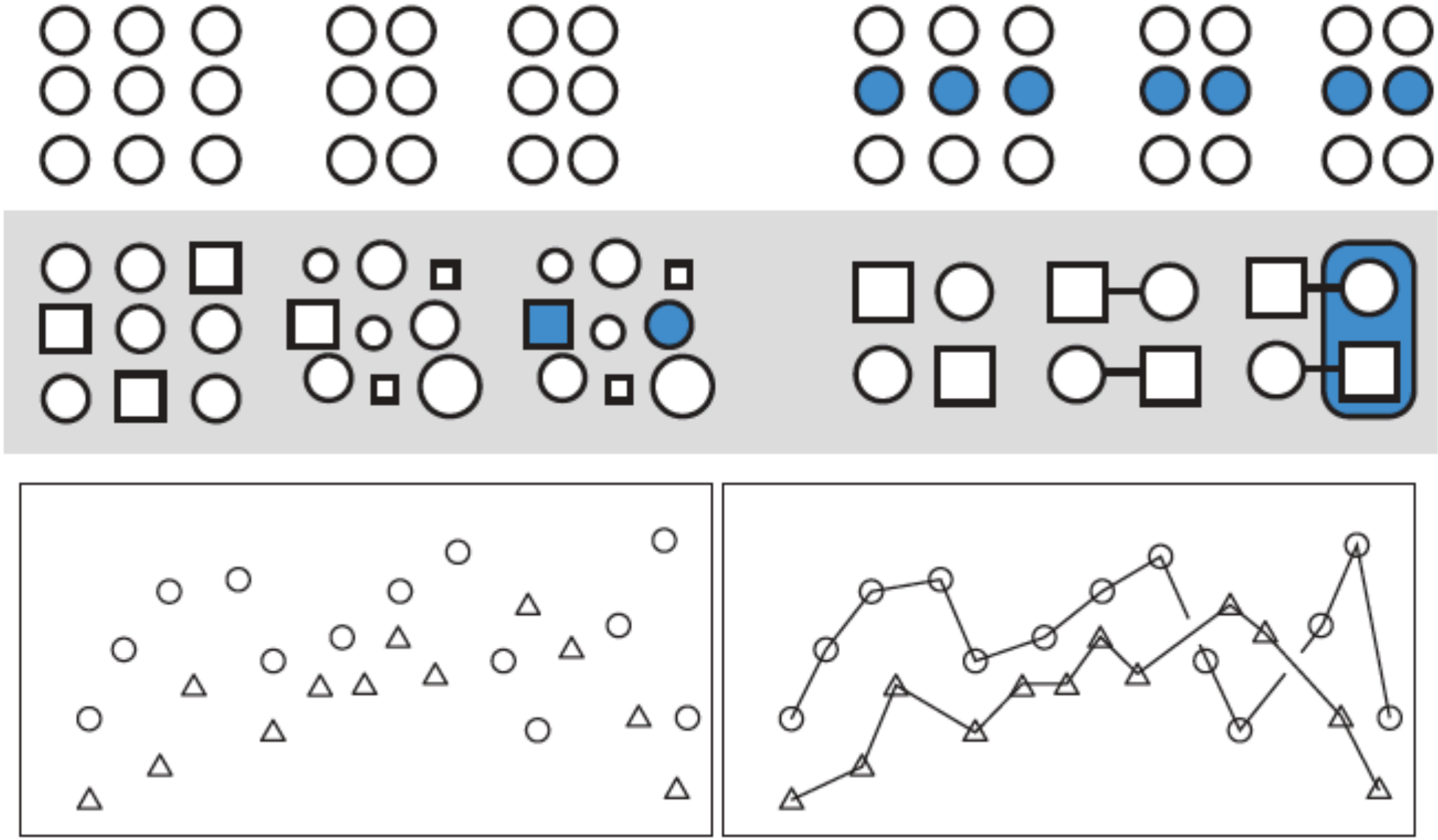
Color: hue-saturation-brightness (HSB)

Mackinlay's ranking of encodings

Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Color hue
Angle	Color saturation	Texture
Slope	Color Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Color saturation
Color saturation	Length	Shape
Color hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

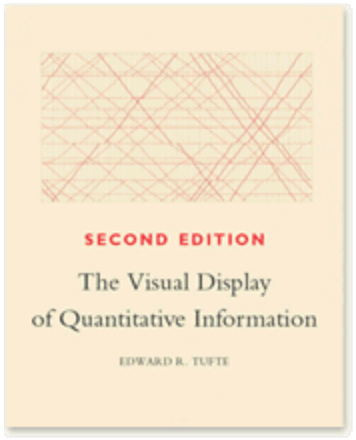
Gestalt principles of relatedness

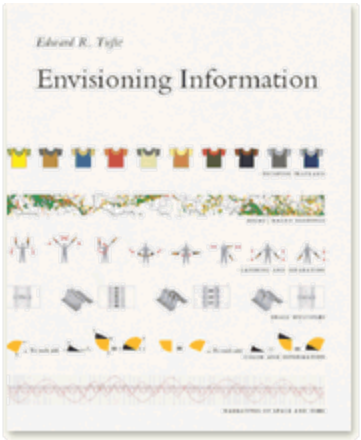
- **Proximity**: Things that are spatially near to one another seem to be related.
- **Similarity:** Things that look alike seem to be related.
- **Connection**: Things that are visually tied to one another seem to be related.
- **Continuity**: Partially hidden objects are completed into familiar shapes.
- **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.

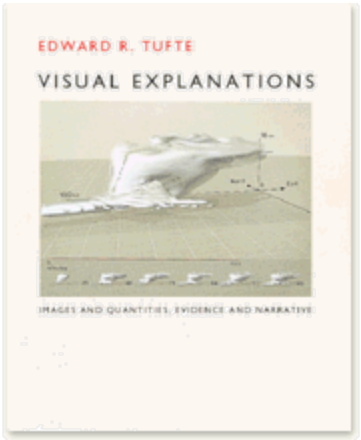


Source: Healy (2019)

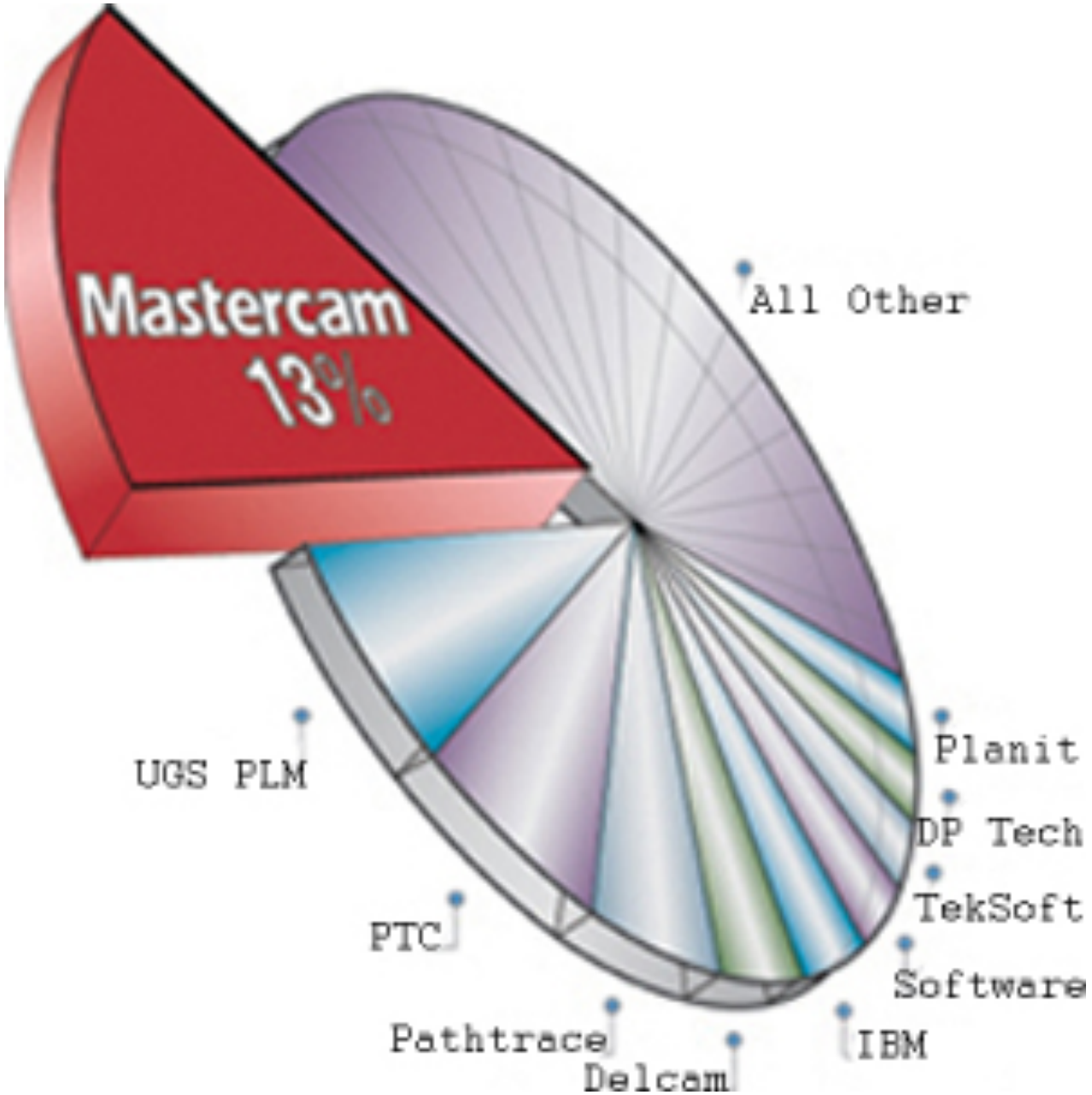
Some (distilled) principles from Tufte



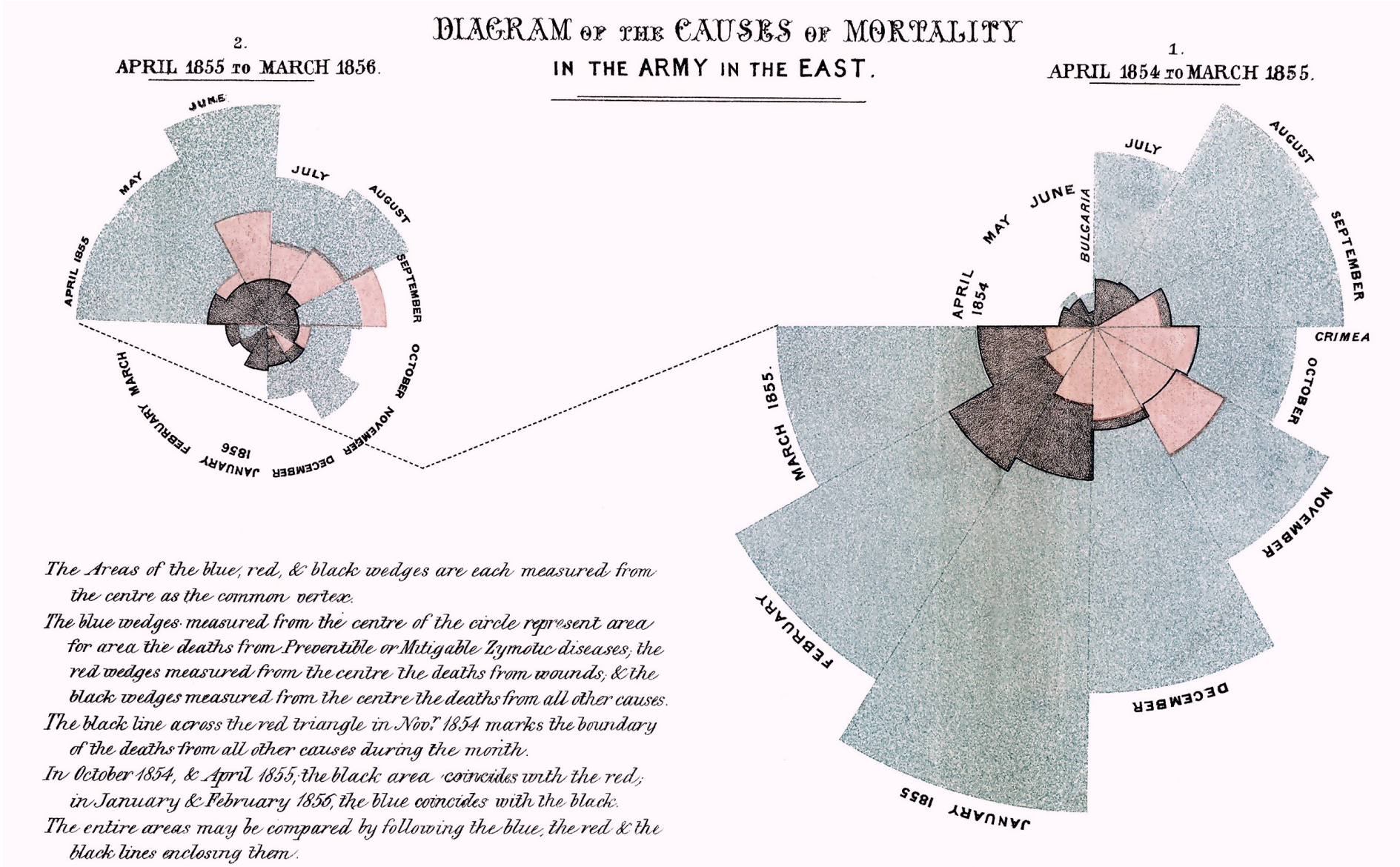


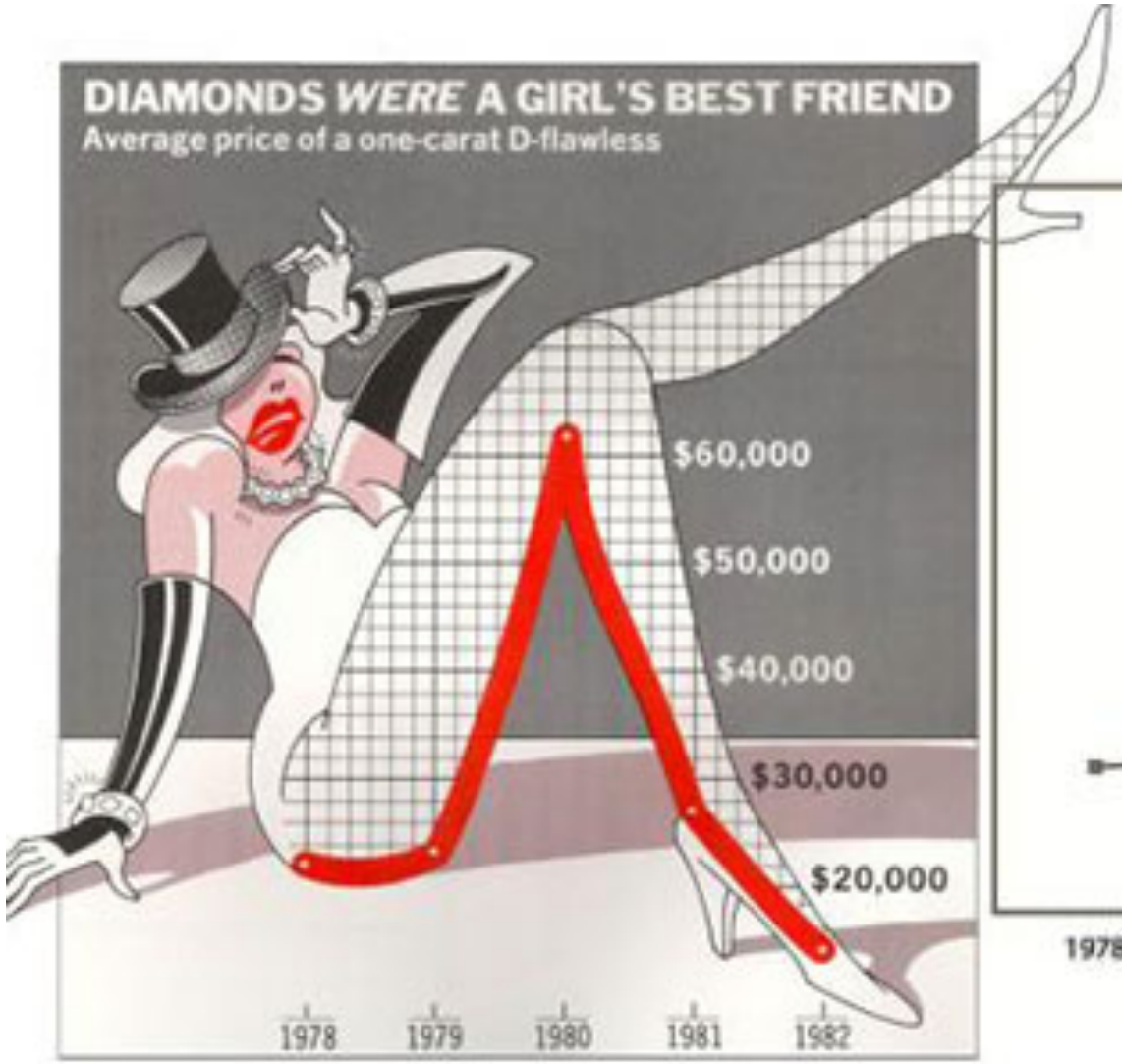


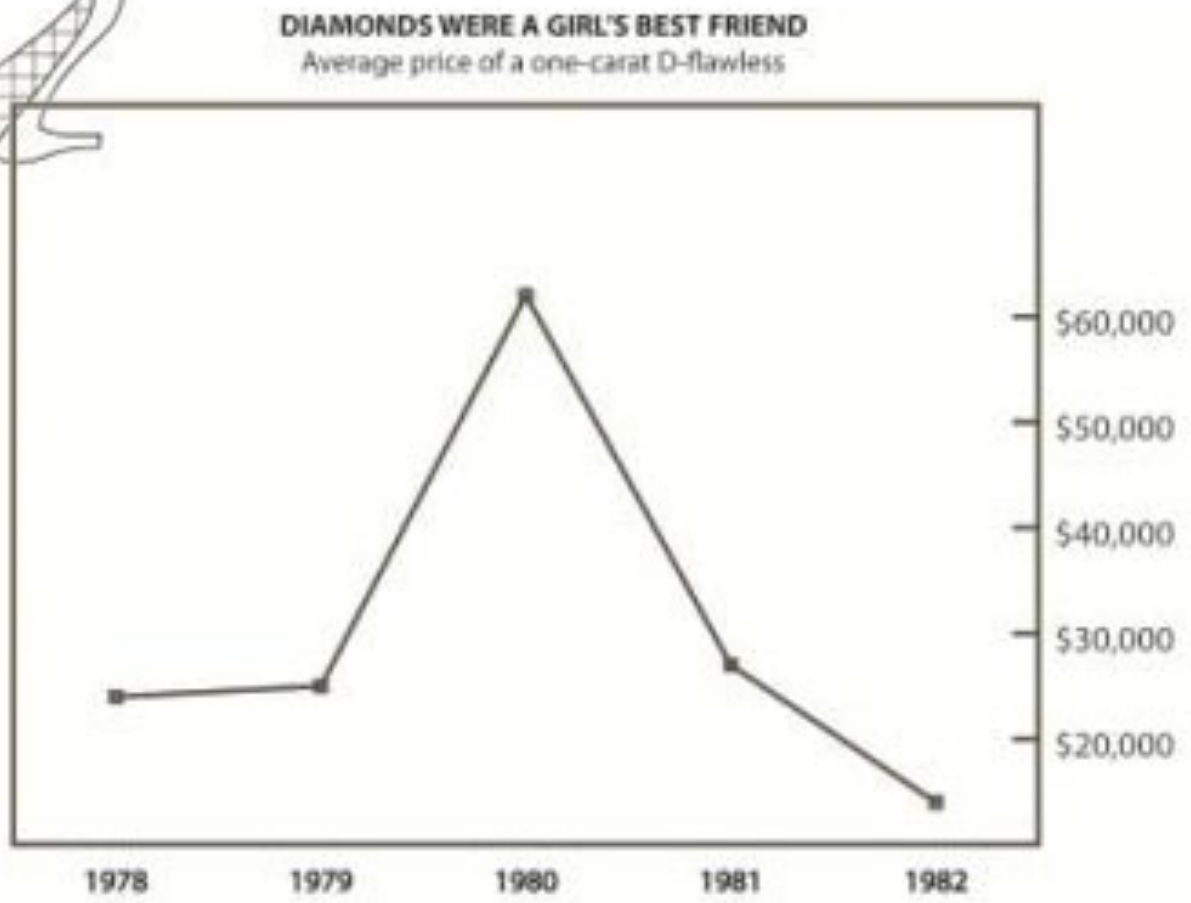
- Ask how data maps to perception
- Ask which comparisons you want, guide eye to those
- Maximize data-to-ink ratio
- Present more data (without losing interpretability)
- Use levels of detail
- (Remember narrative)

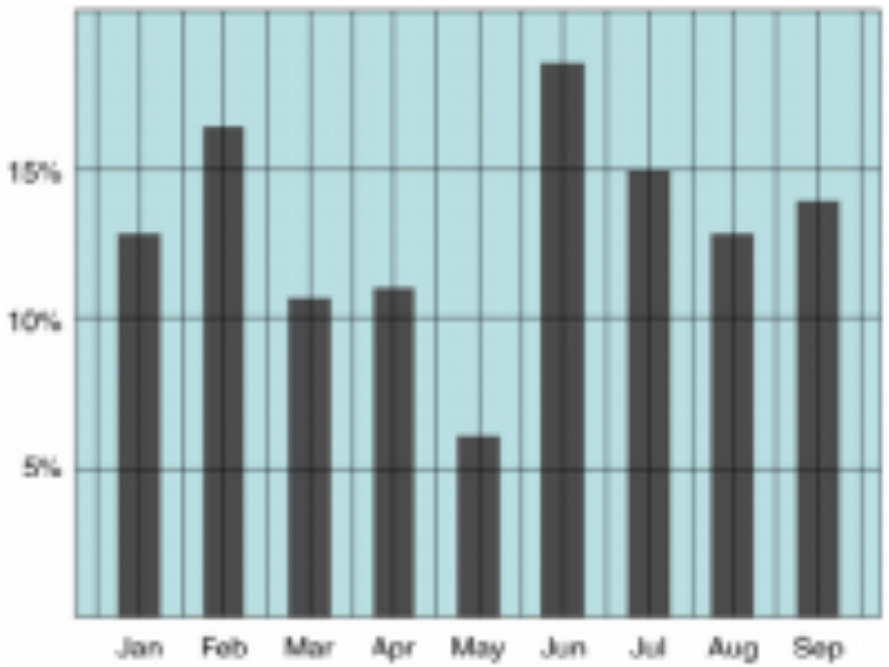

Mastercam
13%
All Other
UGS PLM
Planit
DP Tech
TekSoft
PTC
Software
Pathtrace
IBM
Delcam

Nightingale Rose / Coxcomb chart

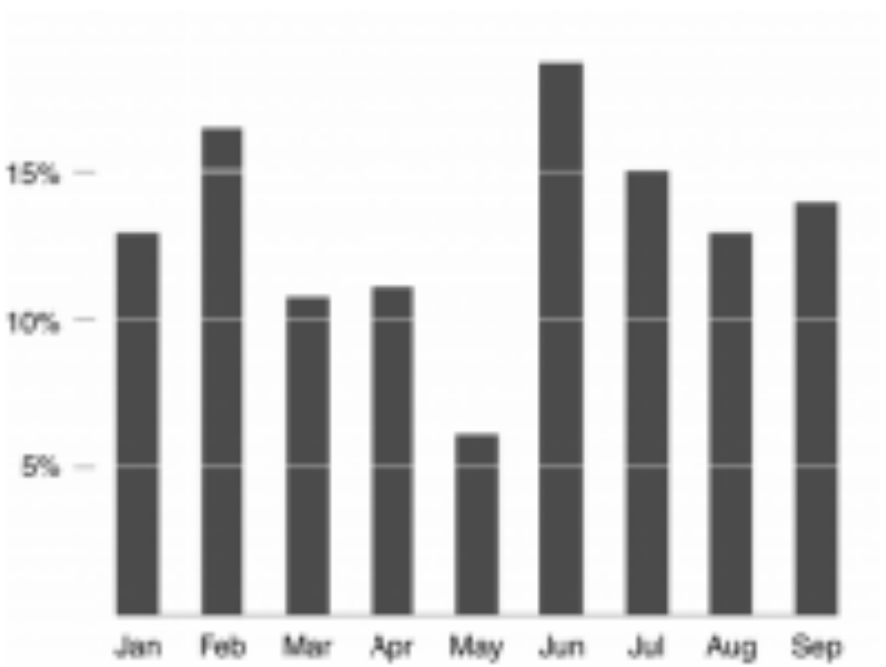



DIAMONDS *WERE* A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982

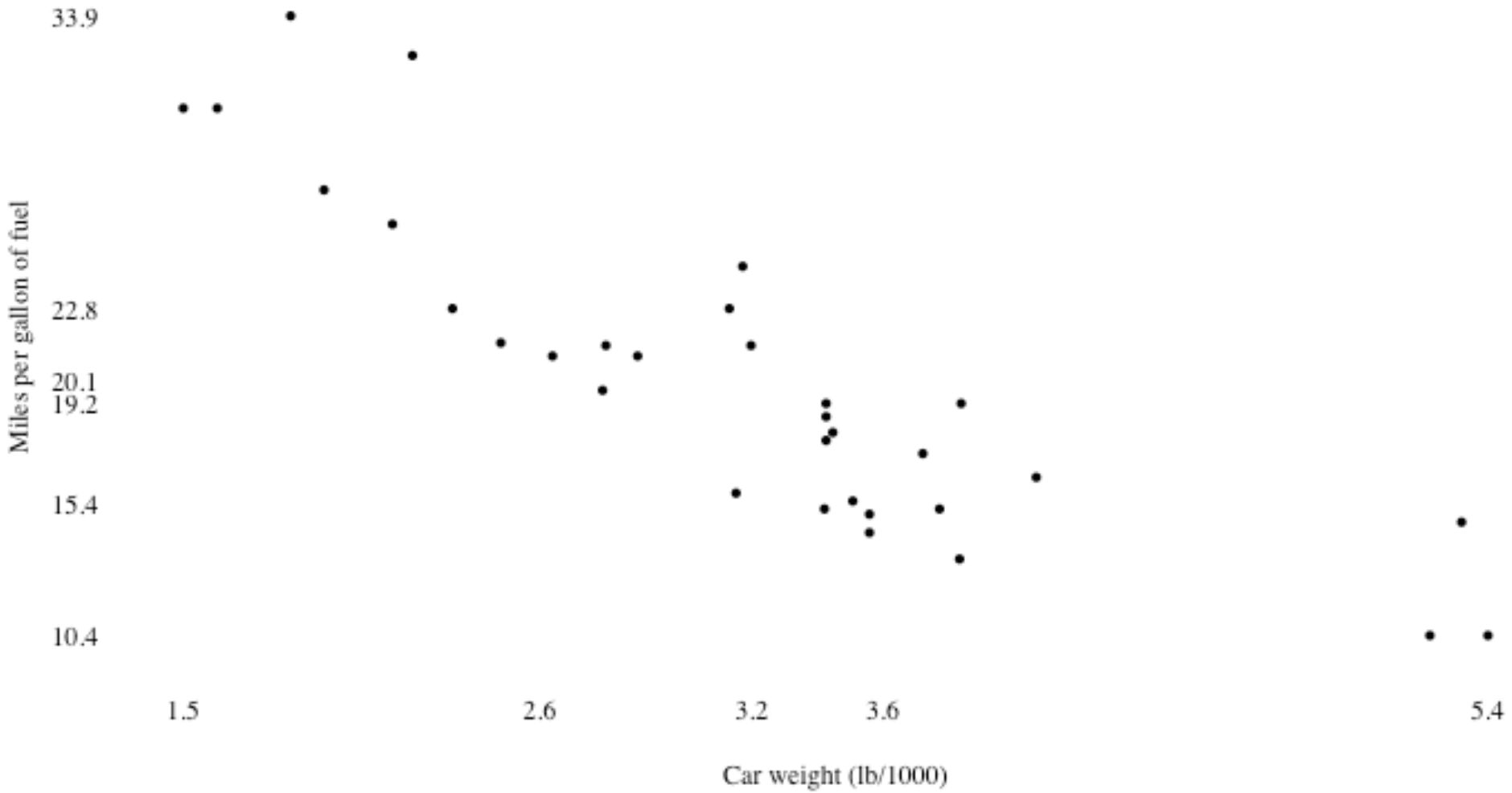

DIAMONDS WERE A GIRL'S BEST FRIEND
Average price of a one-carat D-flawless
$60,000
$50,000
$40,000
$30,000
$20,000
1978
1979
1980
1981
1982

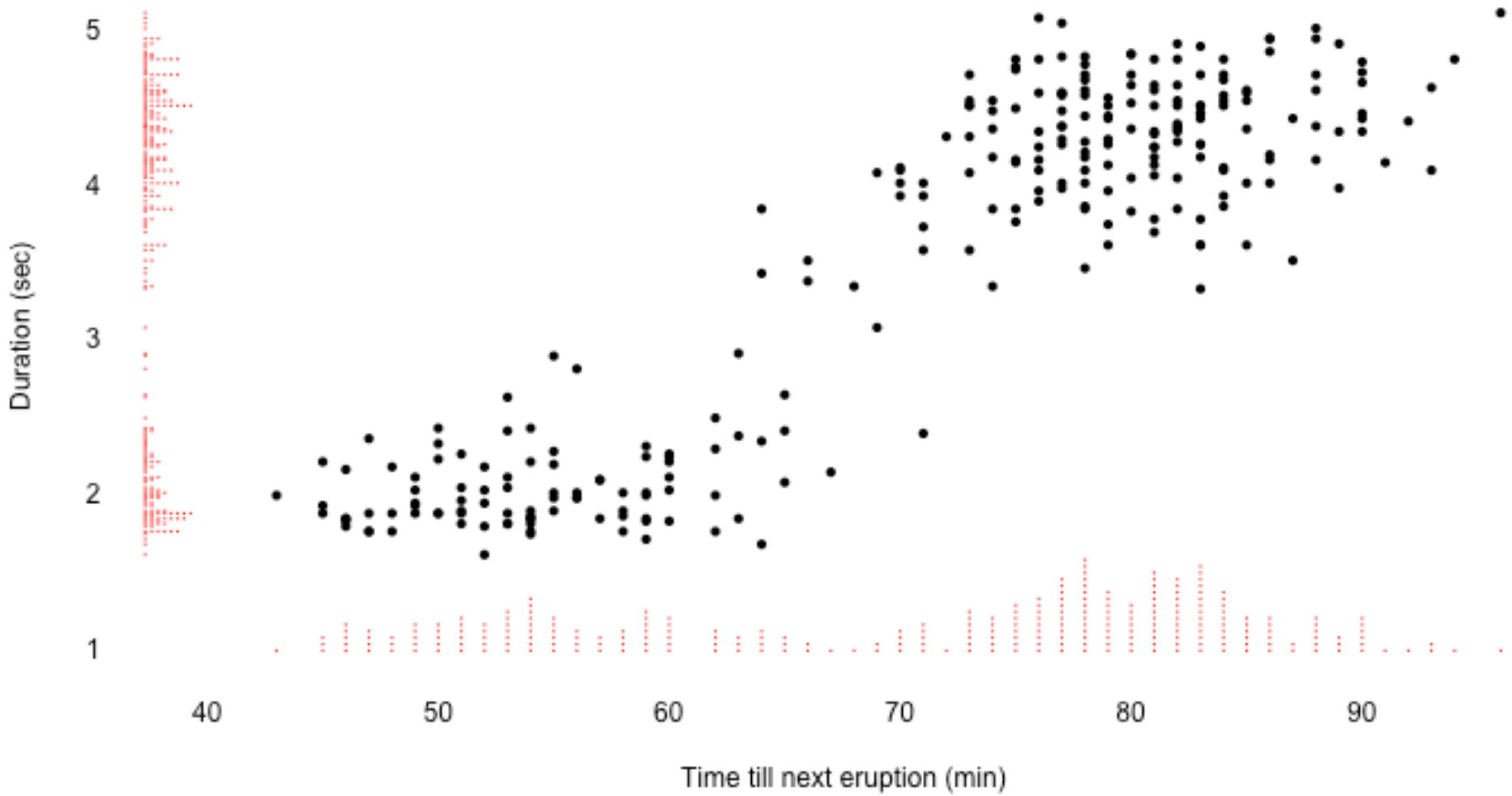


Low Data/Ink

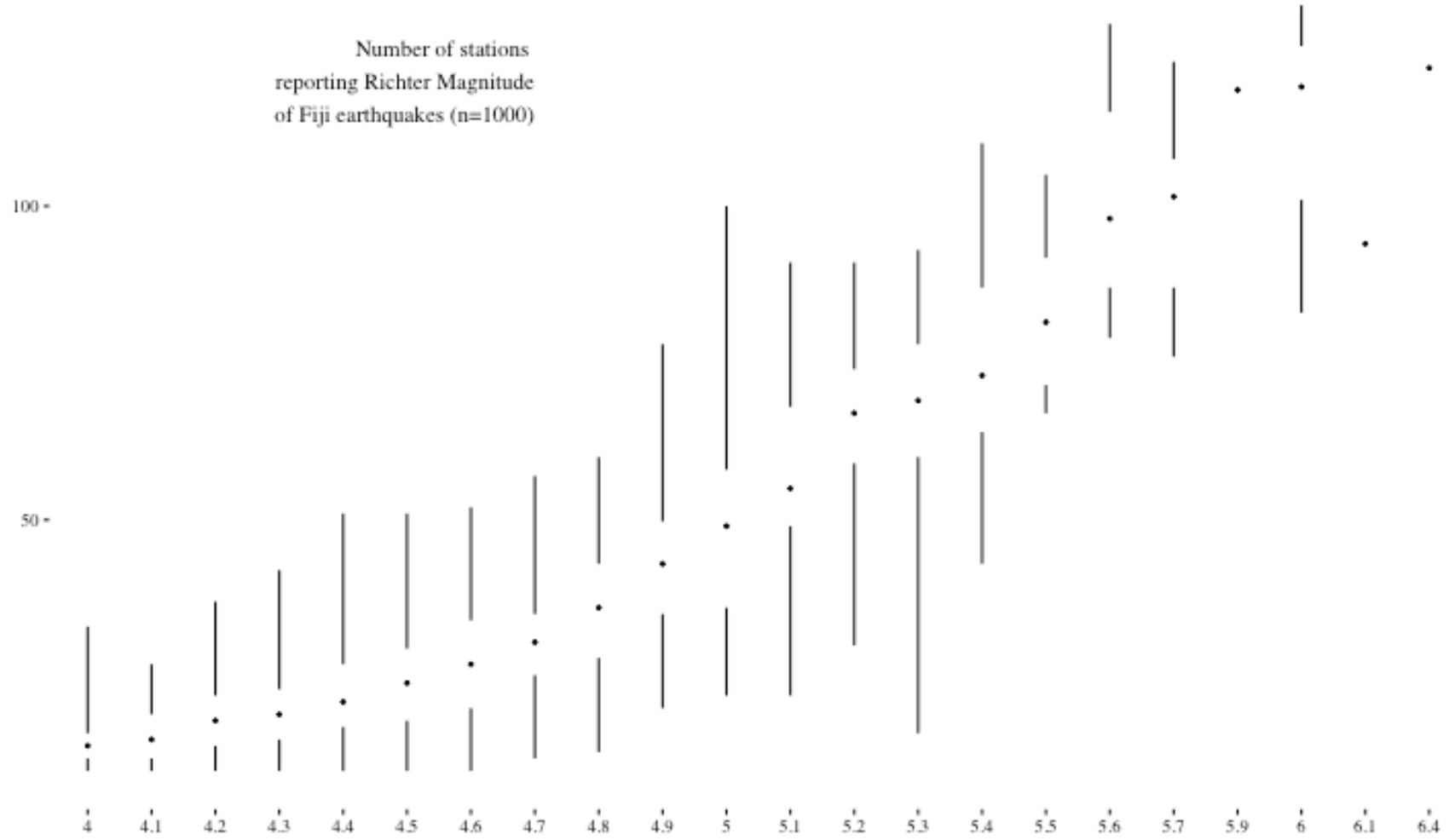


High Data/Ink


33.9
22.8
20.1
19.2
15.4
10.4
Miles per gallon of fuel
1.5
2.6
3.2
3.6
5.4
Car weight (lb/1000)


Duration (sec)
5
4
3
2
1
40
50
60
70
80
90
Time till next eruption (min)

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



Tufte wisdom

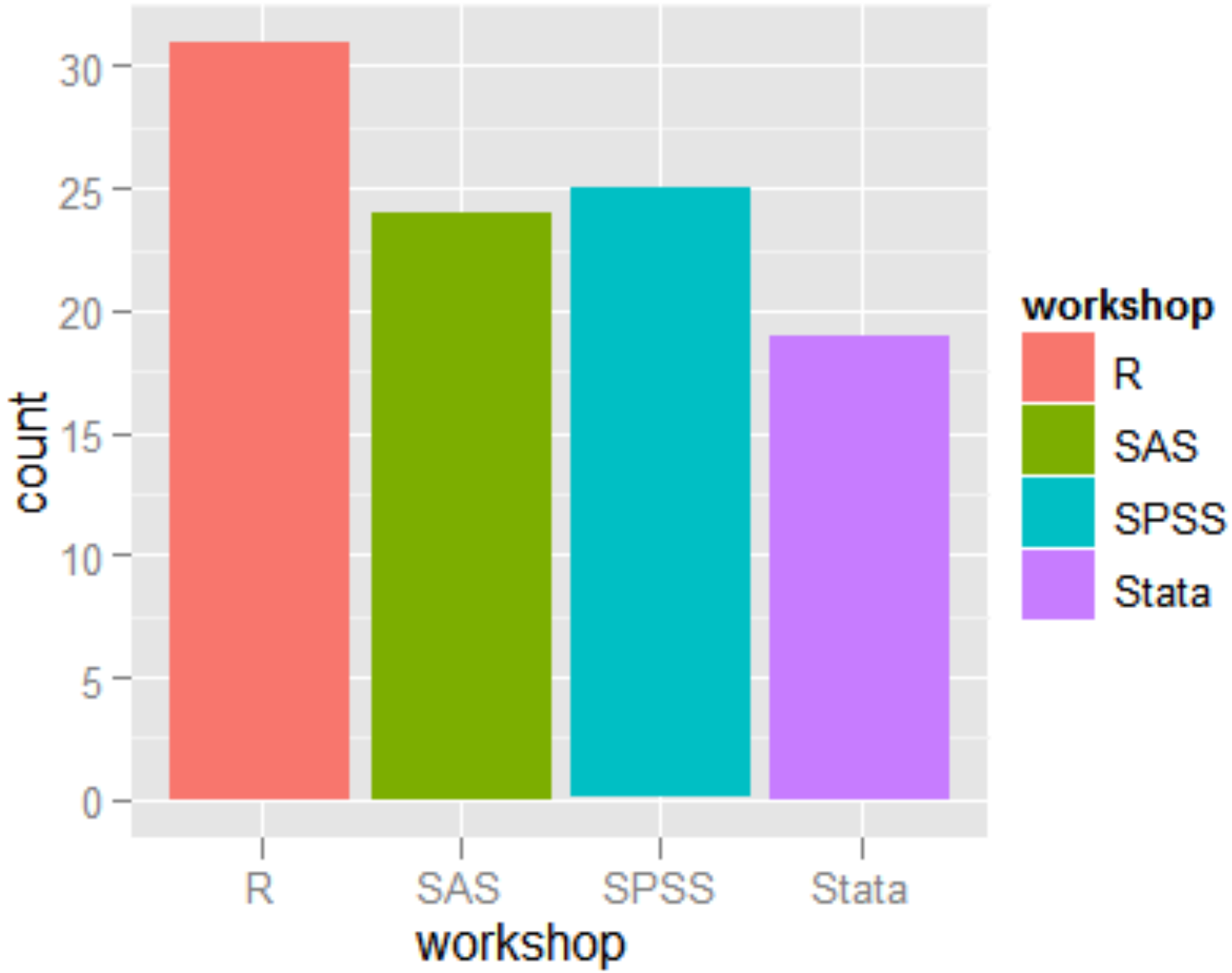
- Tufte's principles are more oriented to communication and can be taken too far

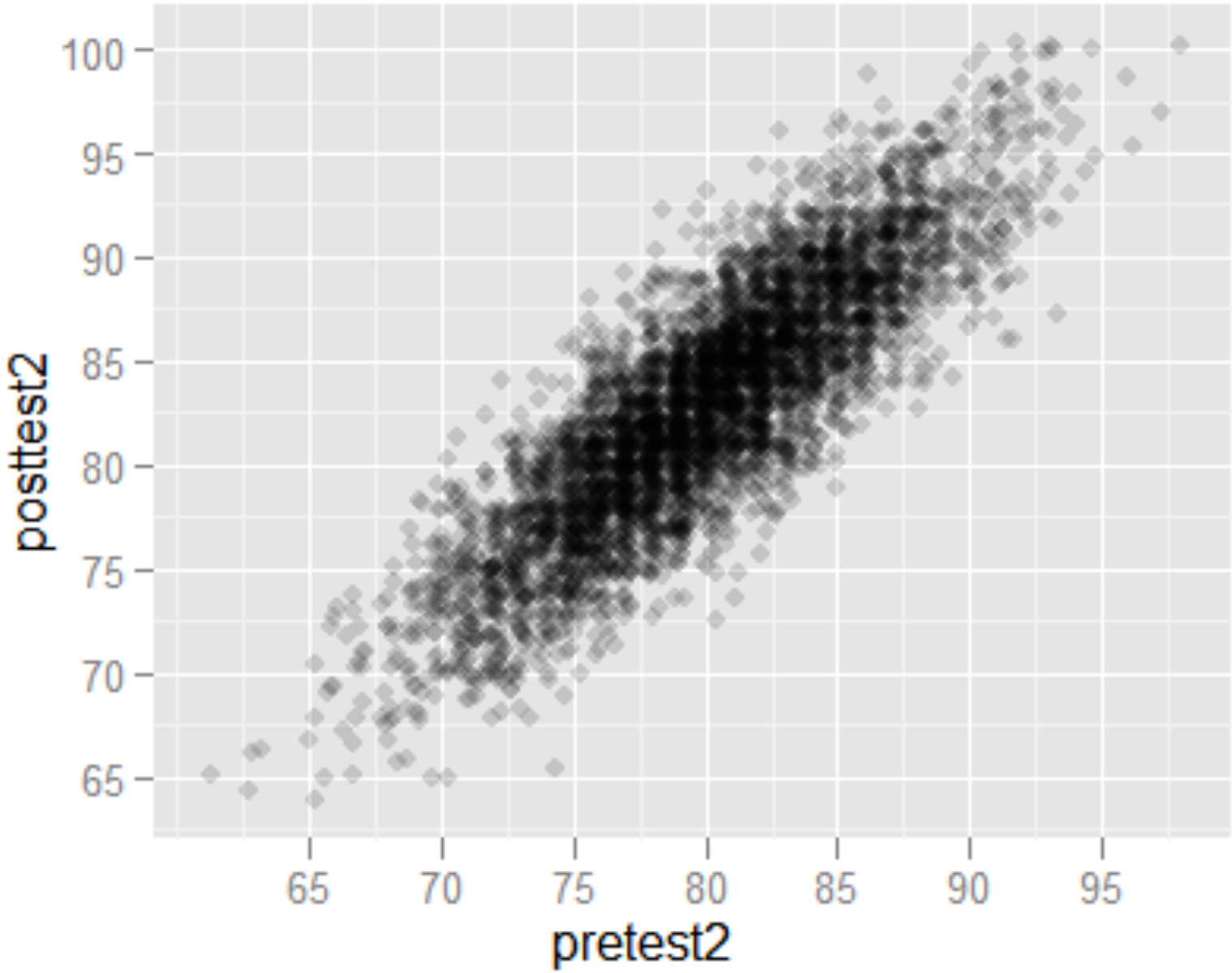
- Better data/ink → display more information without overload;
- Thinking about perception can help you choose better geoms, aesthetics.

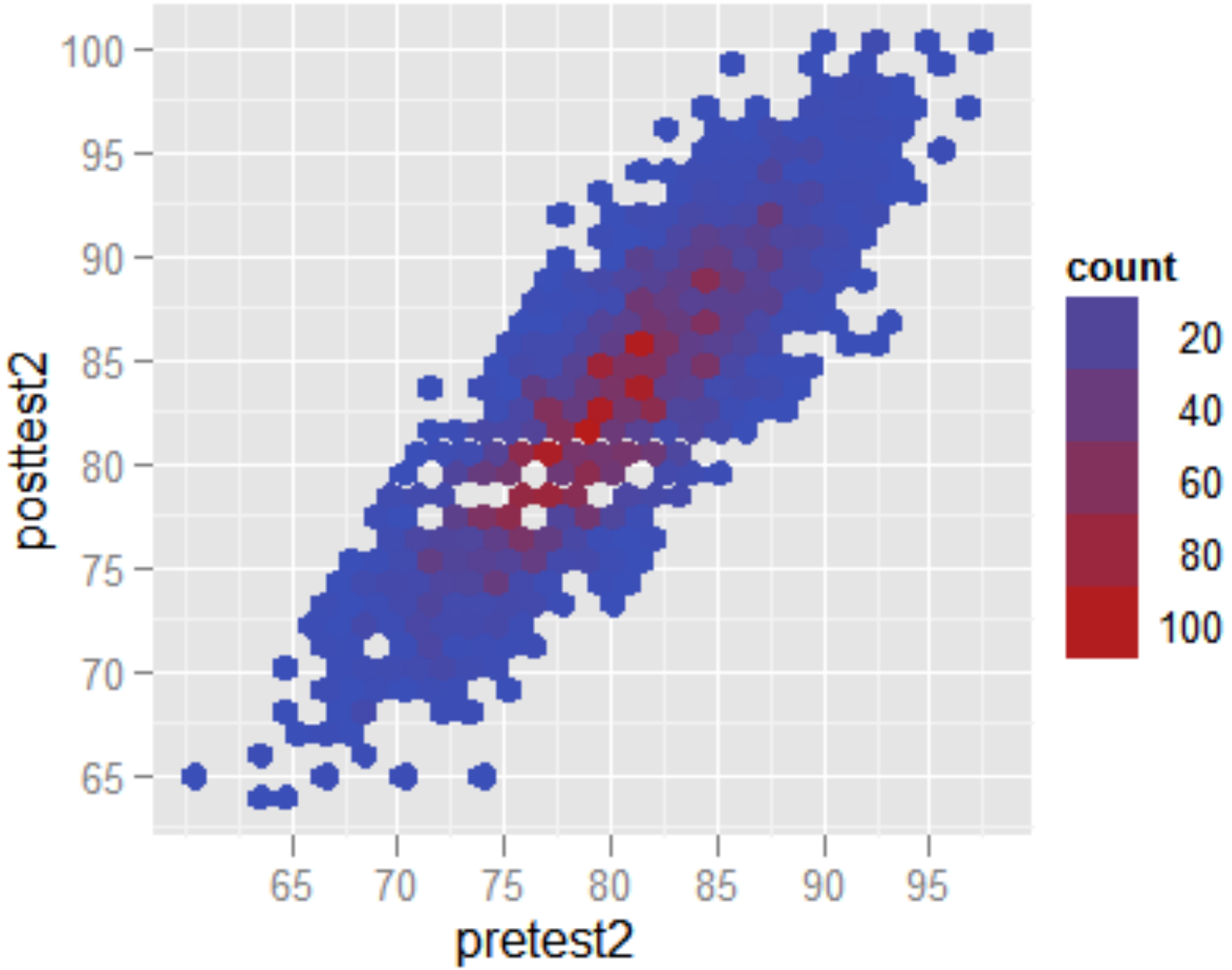
Some practice

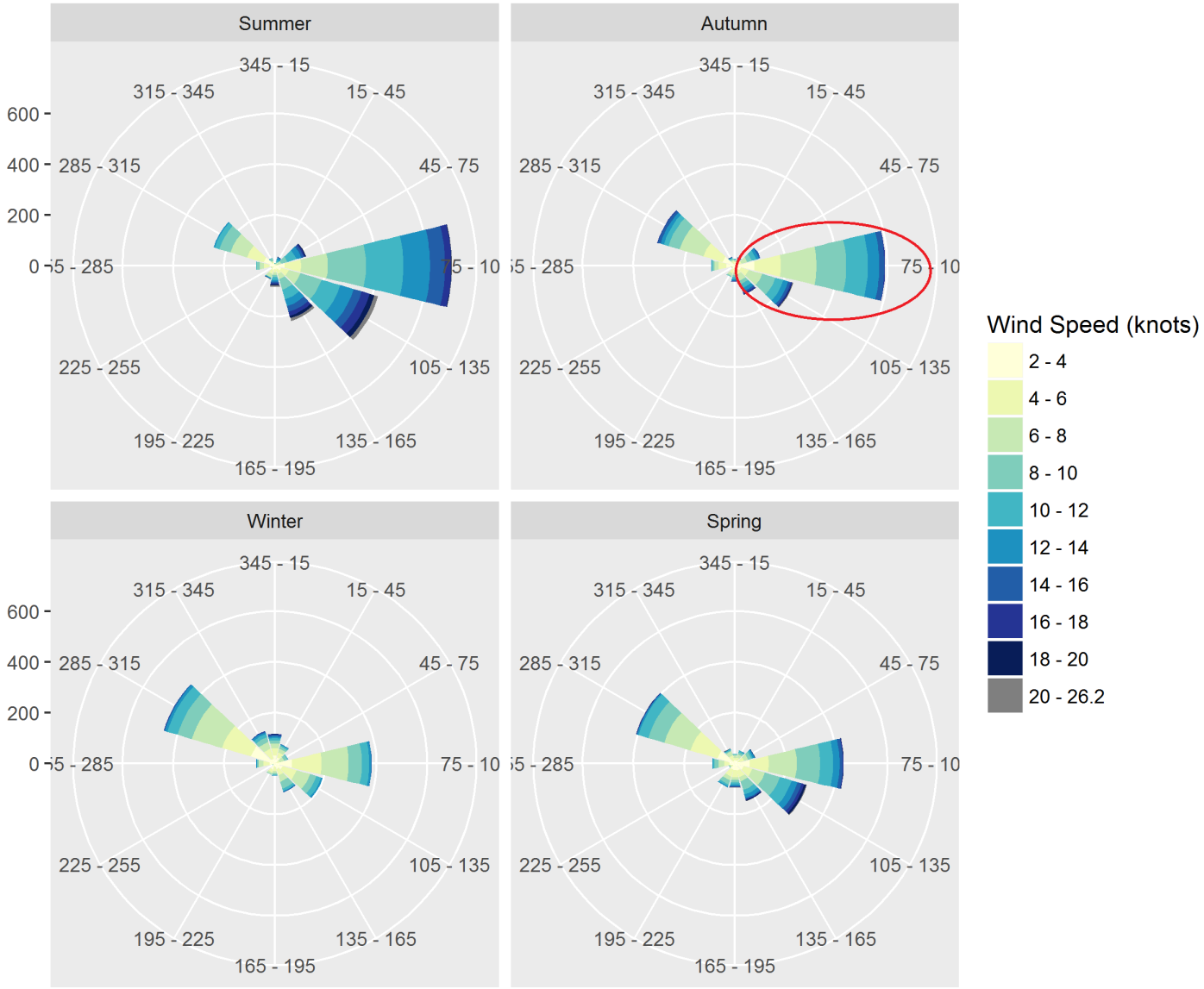
Answer these questions:

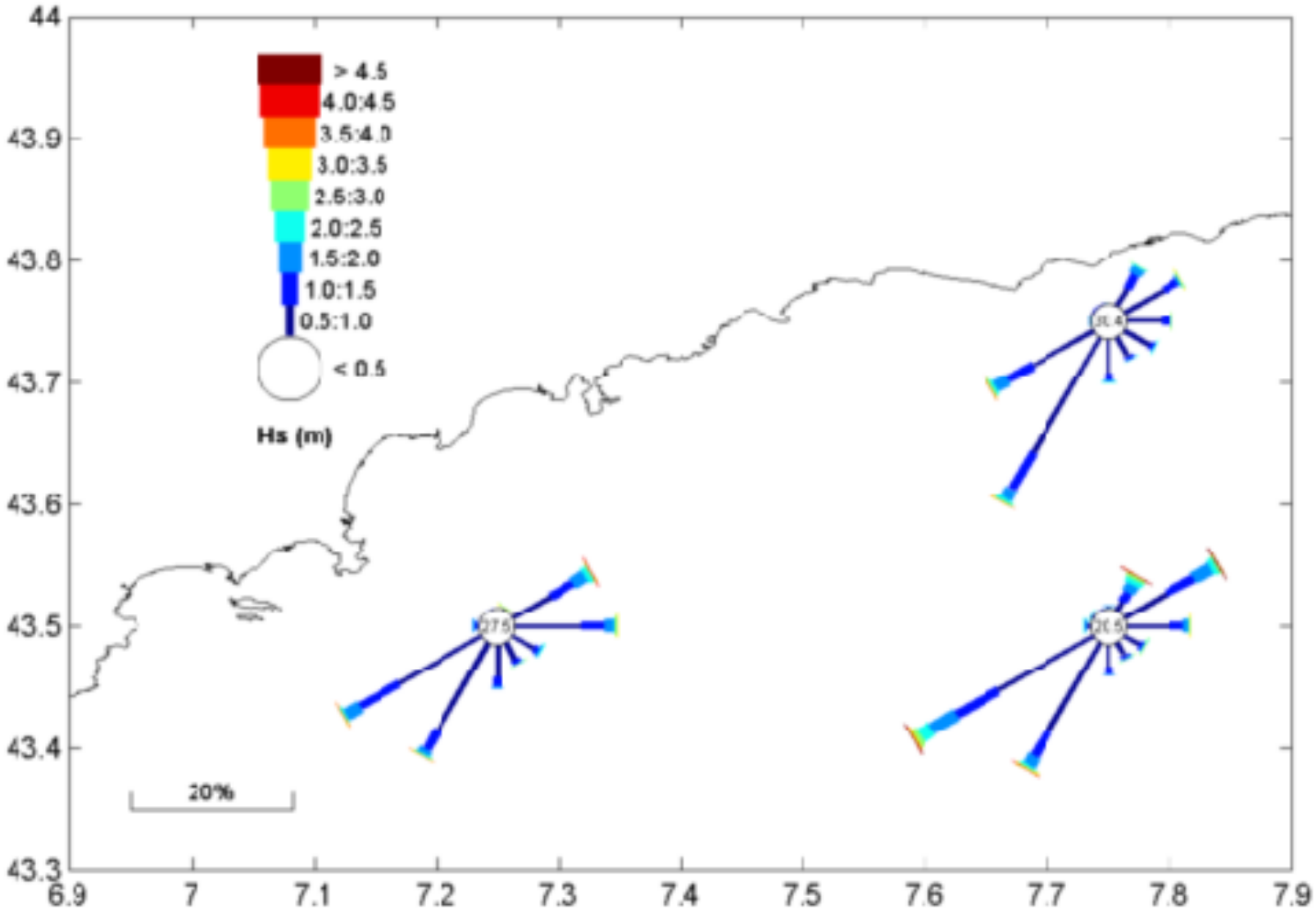
- Are we plotting the right thing?
- What are: aesthetics, geom, scale, facets, transformation, coordinate system
- How is data/ink?
- Is perception considered optimally?
- Can you think of questions you can’t answer from this plot which are in the data?


30
25
20
15
10
5
0
count
R
SAS
SPSS
Stata
workshop
workshop
R
SAS
SPSS
Stata

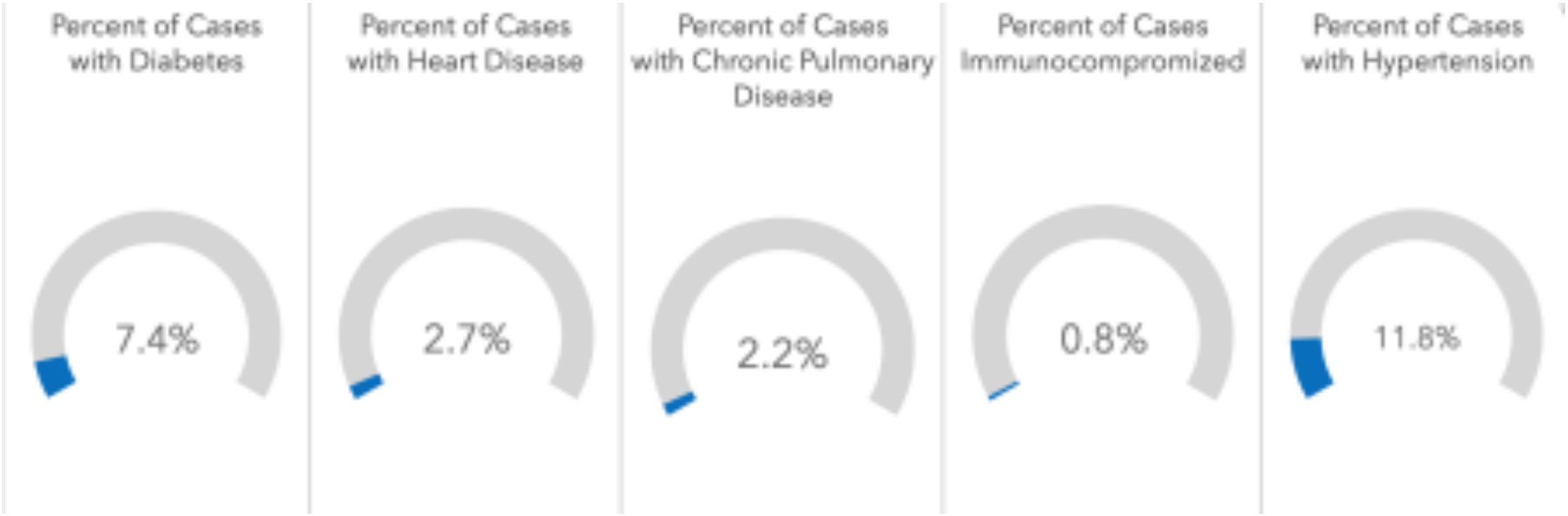

posttest2
100
95
90
85
80
75
70
65
65
70
75
80
85
90
95
pretest2


posttest2
100
95
90
85
80
75
70
65
65
70
75
80
85
90
95
pretest2
count
20
40
60
80
100


Summer
Autumn
Winter
Spring
345 - 15
15 - 45
45 - 75
75 - 10
105 - 135
135 - 165
165 - 195
195 - 225
225 - 255
285 - 315
315 - 345
600
400
200
0
Wind Speed (knots)
2 - 4
4 - 6
6 - 8
8 - 10
10 - 12
12 - 14
14 - 16
16 - 18
18 - 20
20 - 26.2


> 4.5
4.0:4.5
3.5:4.0
3.0:3.5
2.5:3.0
2.0:2.5
1.5:2.0
1.0:1.5
0.5:1.0
< 0.5
Hs (m)
20%
44
43.9
43.8
43.7
43.6
43.5
43.4
43.3
6.9
7
7.1
7.2
7.3
7.4
7.5
7.6
7.7
7.8
7.9

Arkansas COVID-19 dashboard



Conclusion

Conclusion

- Data visualization is a huge field;
- 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).