Social Science Inquiry III

Week 9: Data Visualization

Molly Offer-Westort

Department of Political Science,
University of Chicago

Spring 2024

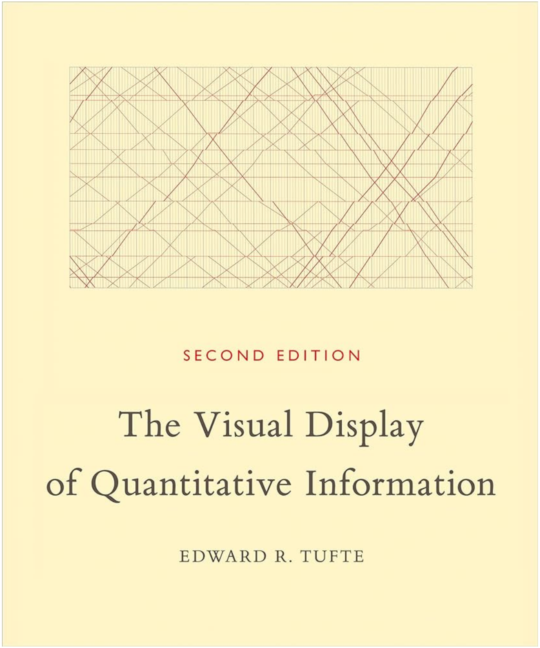
Loading packages for this class

```r
set.seed(60637)
# For plotting:
library(ggplot2)
# library(devtools)
# devtools::install_github("wilkelab/ungeviz")
library(ungeviz)
library(ggridges)
library(ggthemes)
devtools::source_url(
  'https://raw.githubusercontent.com/bearloga/Quartile-frame-
```

```
## i SHA-1 hash of file is
"fe88d63ea7111be1a61ea5d36df1bb9c196fba73"
```

```r
library(khroma)
```

Tufte (2001)



SECOND EDITION

The Visual Display of Quantitative Information

EDWARD R. TUFTE

Edward Tufte

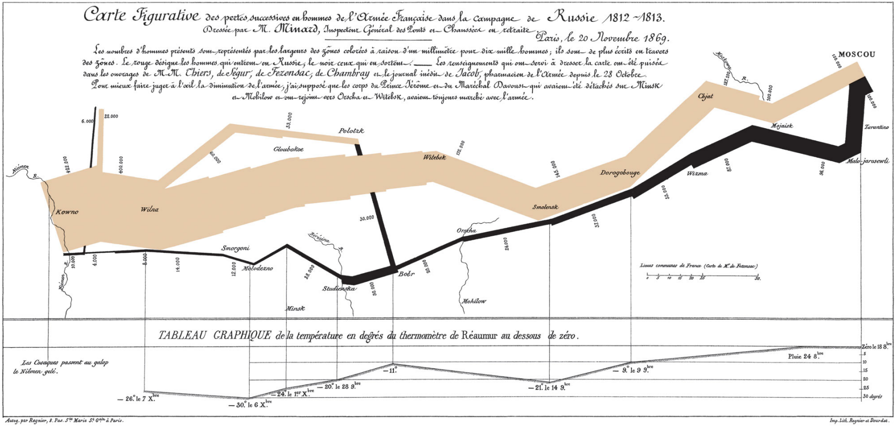
- Statistician and Professor Emeritus of Political Science, Statistics, and Computer Science at Yale University.

Rocket Science 3: Airstream Interplanetary Explorer 2011-2012 steel, aluminum, stainless steel, electronics
length 84' x height 31' Photo by Fred Orkin

Graphical excellence (in Tufte's words)

- Graphical displays should:
    - show the data
    - induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
    - avoid distorting what the data have to say
    - present many numbers in a small space
    - make large data sets coherent
    - encourage the eye to compare different pieces of data
    - reveal the data at several levels of detail, from a broad overview to the fine structure
    - serve a reasonably clear purpose: description, exploration, tabulation, or decoration
    - be closely integrated with the statistical and verbal descriptions of a data set

The Minard Map



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 zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. 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. Thiers, 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'oeil 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 Mobilow 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.

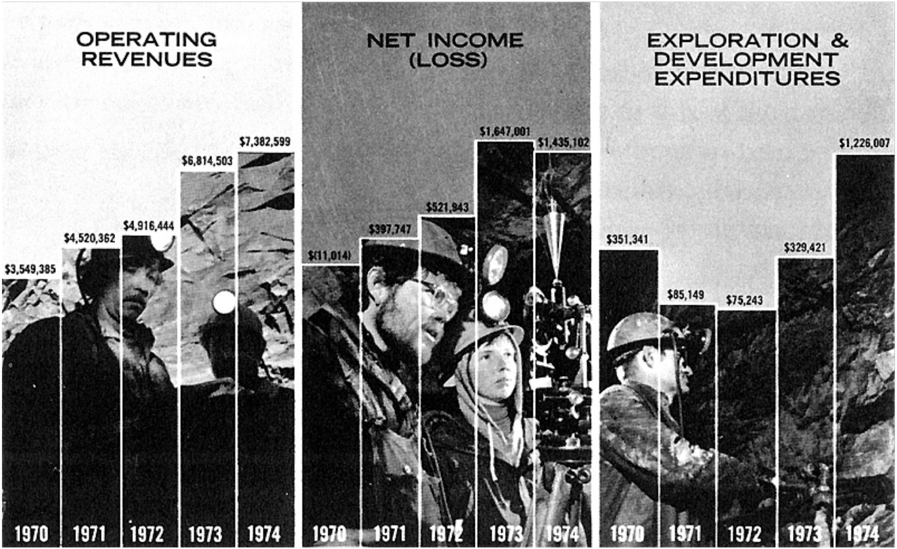
Les Cosaques passent au galop le Niémen gelé.

Graphical integrity

- Visual representations of data should accurately reflect the data itself. Representations of numbers on graphs should be proportional to the data they represent.

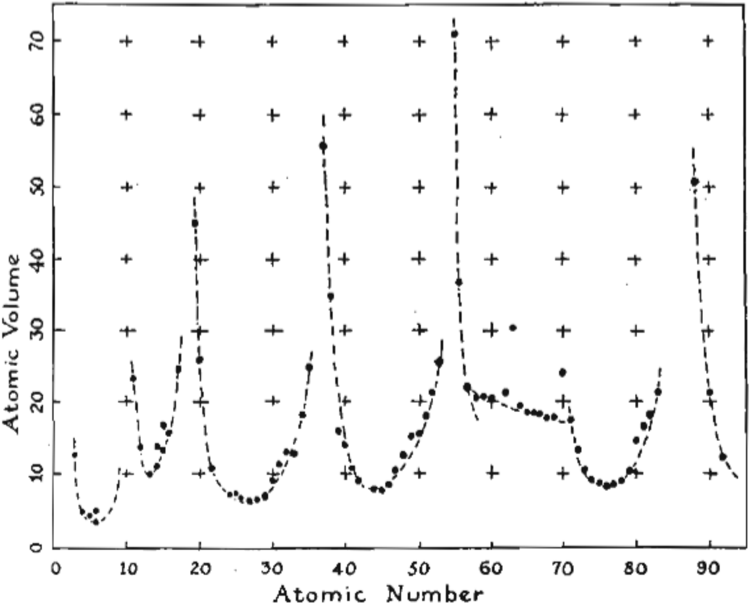
- Label with clarity and detail.

- Don’t include more “information-carrying dimensions” than exist in the data.

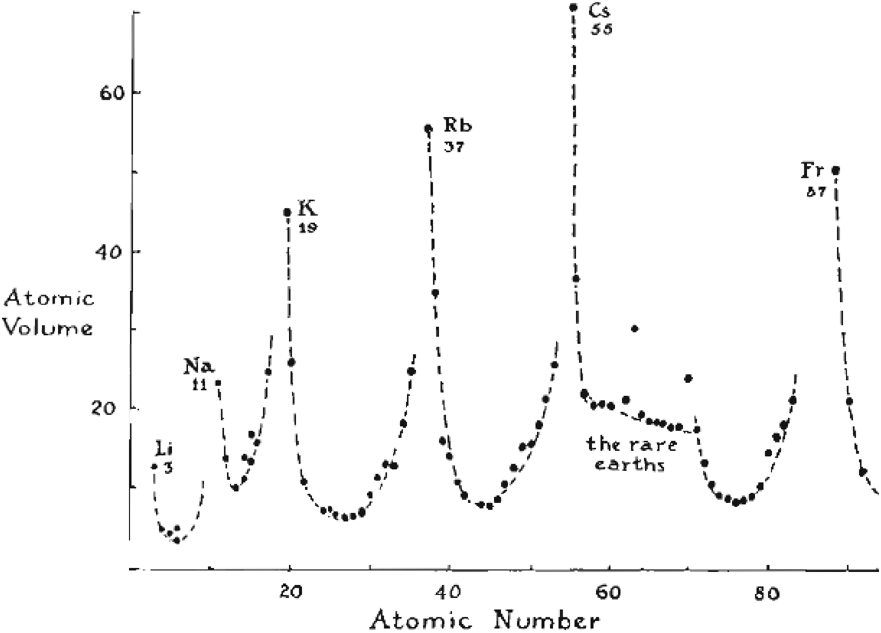

OPERATING REVENUES
$3,549,385
$4,520,362
$4,916,444
$6,814,503
$7,382,599
1970
1971
1972
1973
1974
NET INCOME (LOSS)
$(11,014)
$397,747
$521,943
$1,647,001
$1,435,102
1970
1971
1972
1973
1974
EXPLORATION & DEVELOPMENT EXPENDITURES
$351,341
$85,149
$75,243
$329,421
$1,226,007
1970
1971
1972
1973
1974

Five priniples in the theory of data graphics

- Above all else show the data.
- Maximize the data-ink ratio.
- Erase non-data ink.
- Erase redundant data-ink.
- Revise and edit.

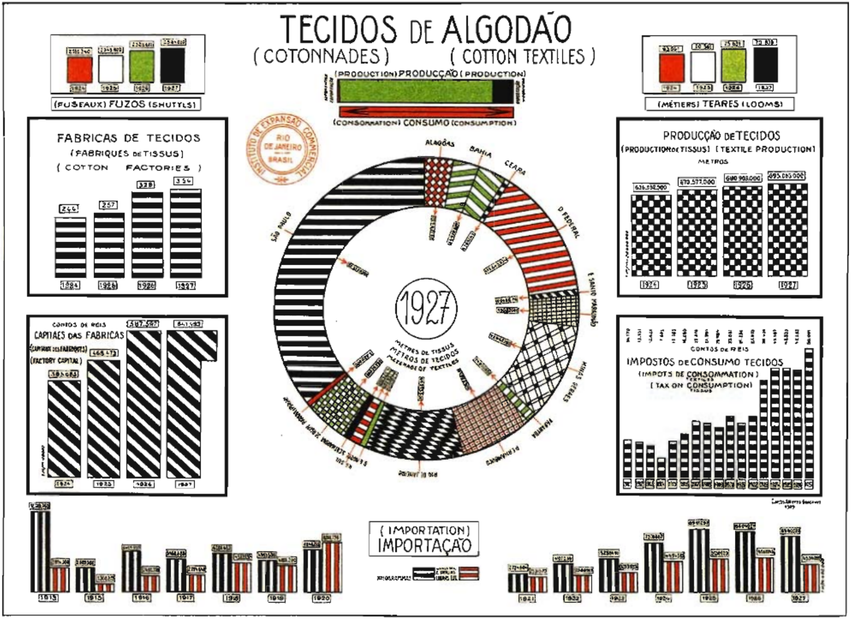


Linus Pauling, *General Chemistry* (San Francisco, 1947), p. 64.


Cs
55
Rb
37
Fr
87
K
19
Na
11
Li
3
the rare earths
Atomic Volume
60
40
20
20
40
60
80
Atomic Number

Avoid “chartjunk”

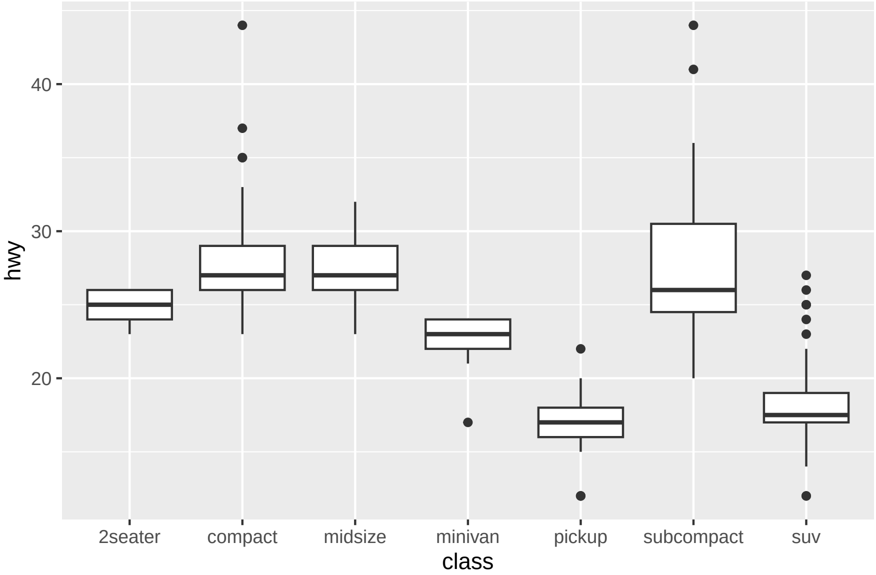
- Chartjunk: Non-essential or redundant information in graphics.
- Avoid distractions that do not enhance understanding.
- No meaningless patterns or dimensions, no grids, no chart-as-decoration.

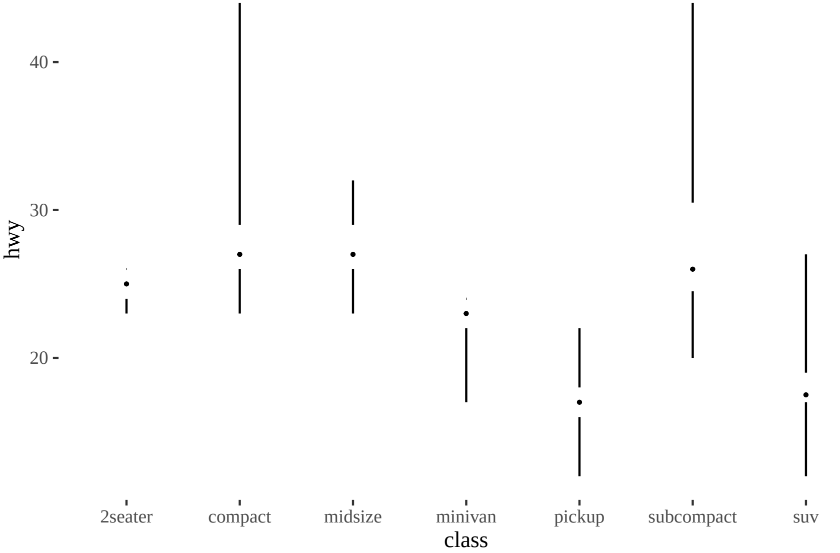


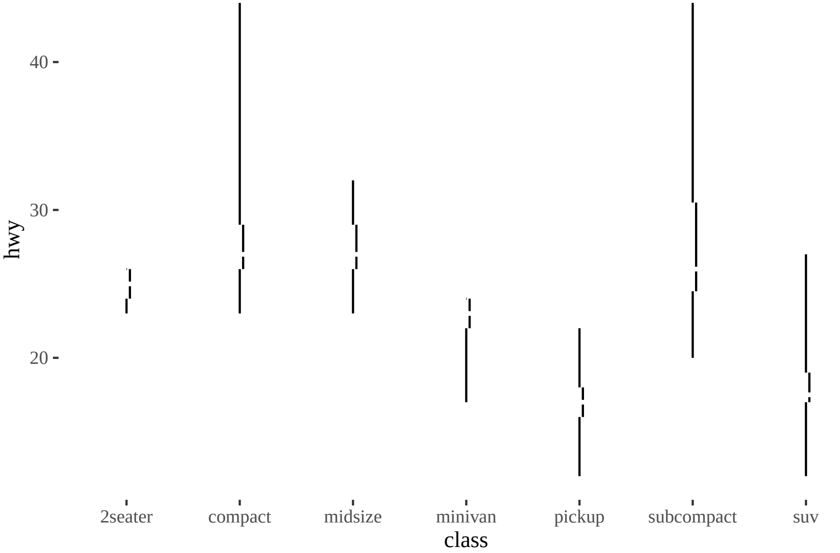
Institute de Expansão Commercial, *Brasil: Graphicos Economics-Estatisticas*, (Rio de Janeiro, 1929) p. 15.

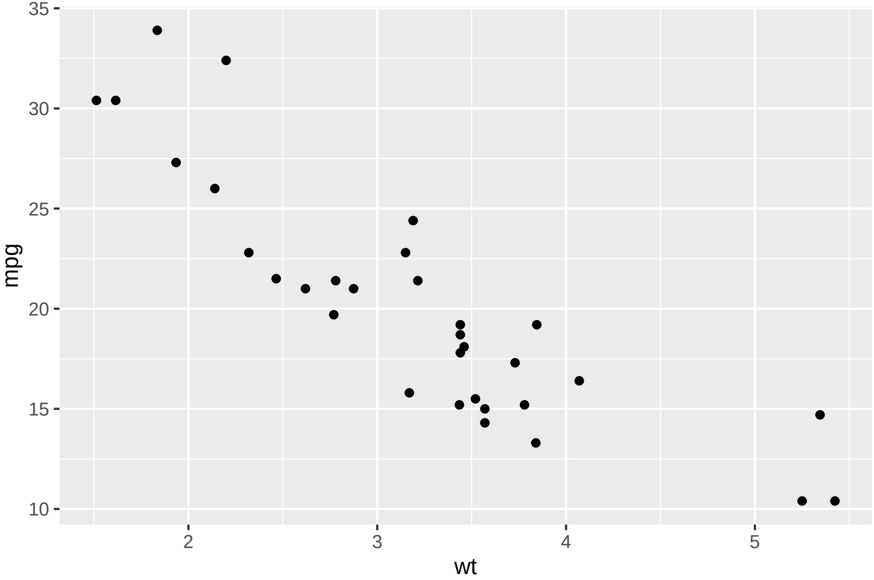
Data-ink maximization

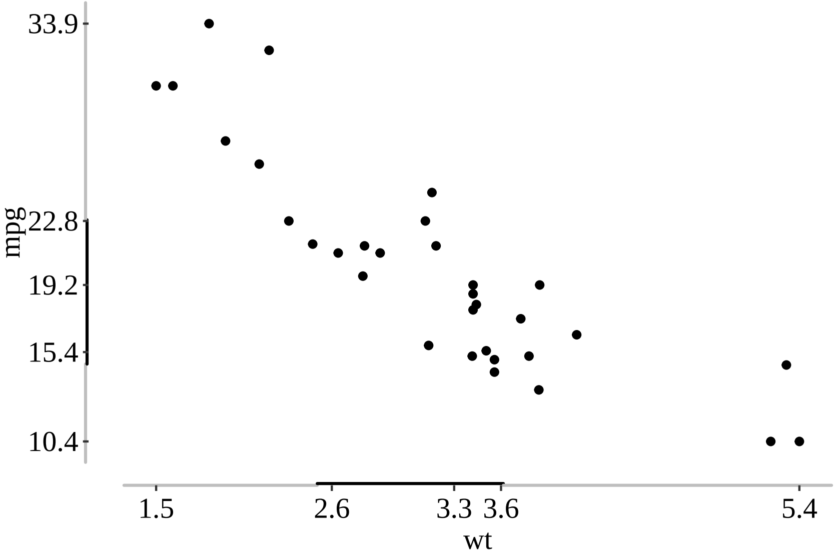
- New graphical forms.

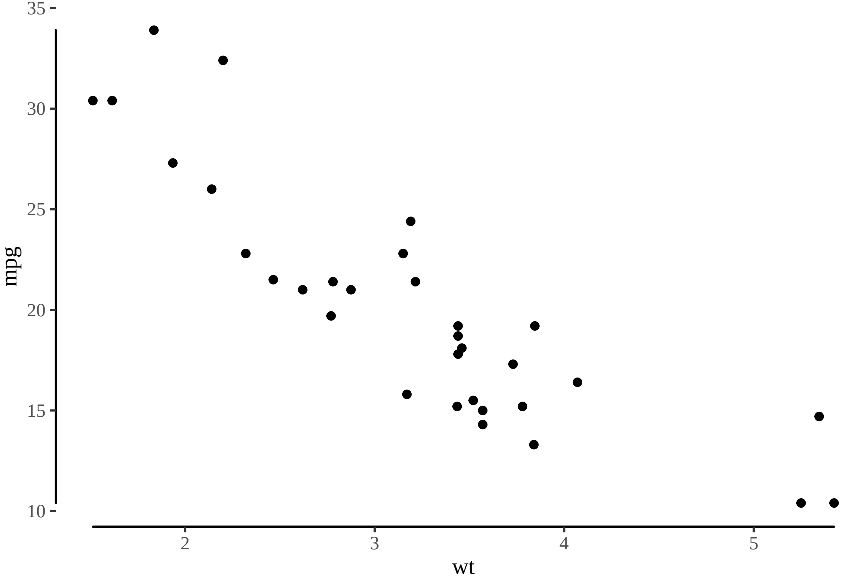

40
30
20
hwy
2seater
compact
midsize
minivan
pickup
subcompact
suv
class

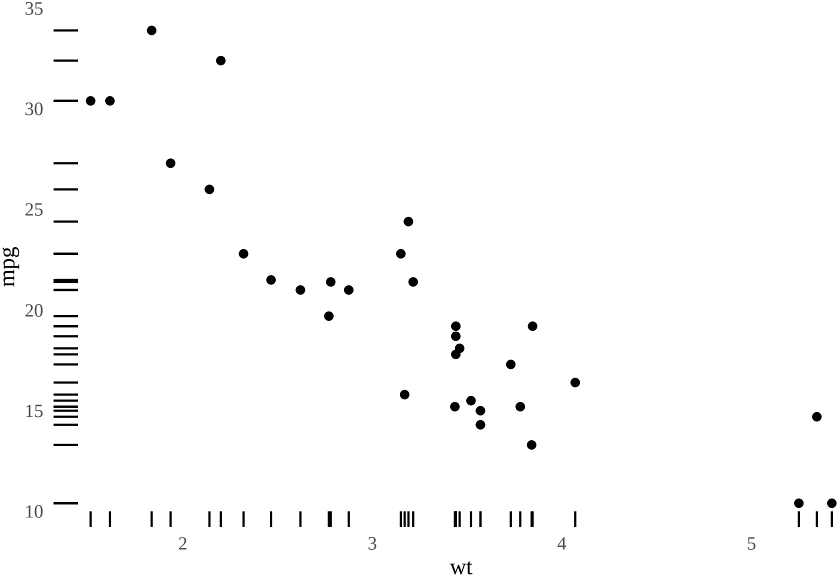

40
30
20
hwy
2seater
compact
midsize
minivan
pickup
subcompact
suv
class


40
30
20
hwy
2seater
compact
midsize
minivan
pickup
subcompact
suv
class


35
30
25
20
15
10
mpg
2
3
4
5
wt


33.9
22.8
19.2
15.4
10.4
mpg
1.5
2.6
3.3
3.6
5.4
wt

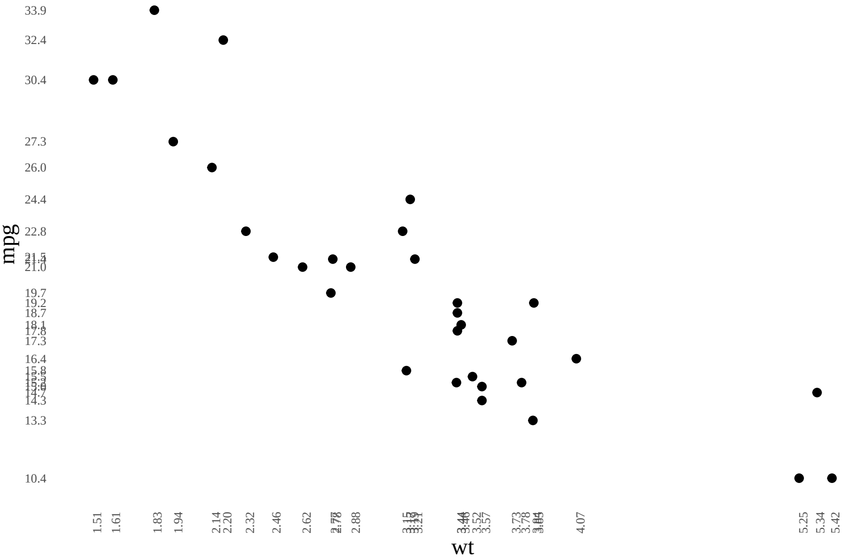

35
30
25
20
15
10
mpg
2
3
4
5
wt


35
30
25
20
15
10
mpg
2
3
4
5
wt

Multi-functioning graphical elements

- Use elements that serve more than one purpose.

- Combine text and images for efficient communication.


mpg
33.9
32.4
30.4
27.3
26.0
24.4
22.8
21.5
21.4
21.0
19.7
19.2
18.7
18.1
17.8
17.3
16.4
15.8
15.5
15.2
15.0
14.7
14.3
13.3
10.4
1.51
1.61
1.83
1.94
2.14
2.20
2.32
2.46
2.62
2.77
2.78
2.88
3.15
3.17
3.19
3.21
3.44
3.46
3.52
3.57
3.73
3.78
3.84
3.85
4.07
5.25
5.34
5.42
wt

Other principles

- Present a large amount of data in a small space.

- Use small multiples to make efficient comparisons, revealing trends.

Aesthetics

- Employ visual balance.
- Combine words, numbers, and pictures–all together in a graphic.
- Lines should be thin. Add weight to add meaning.
- Label series of words horizontally rather than stacked vertically.
- For causal or predictive graphs, plot the response on the Y-axis, the cause or predictor on the X-axis.
- On shape:
  - If the nature of the data suggests the shape of the plot, follow it.
  - "smoothly-changing curves can stand to be taller rather than wide, but a wiggly curve needs to be wider than tall…" - John W. Tukey, *Exploratory Data Analysis* (1977) p. 129.
  - Otherwise, opt for horizontally oriented plots with ratios 3:2 in width:height.

Using color

Colorblind palettes

- Use colorblind-friendly palettes to ensure effective communication.

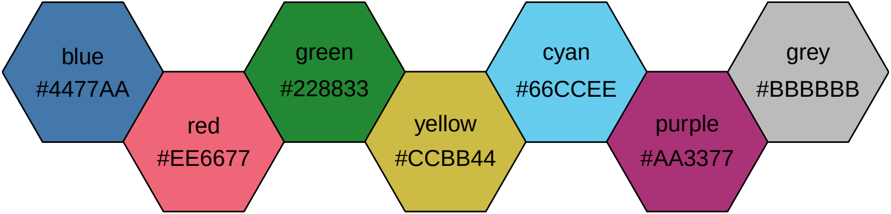
Principles for creating your own colorblind-friendly palettes

- Use high contrast between colors.
- Avoid using red and green together.
- Use shades to differentiate data points.
- You can test your visualizations with colorblindness simulation tools.

Paul Tol's color schemes

Tol (2021)

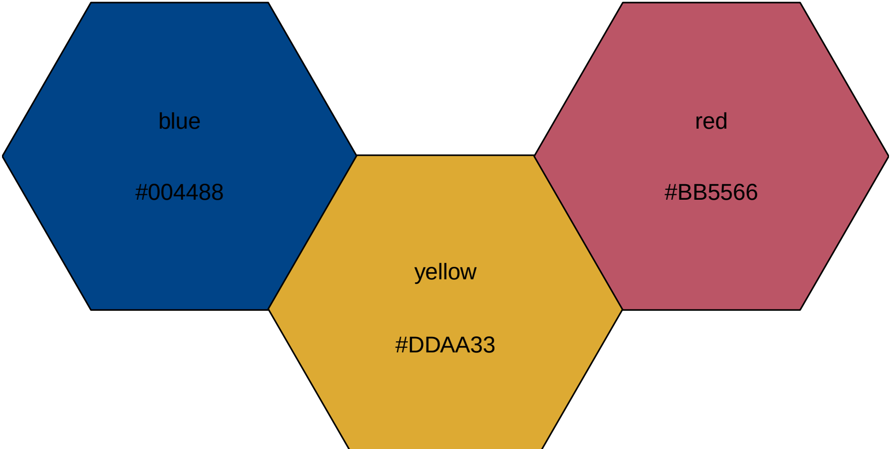
```
bright <- color("bright")
plot_scheme(bright(7), colours = TRUE, names = TRUE, size = 0
```



Paul Tol’s color schemes

Tol (2021)

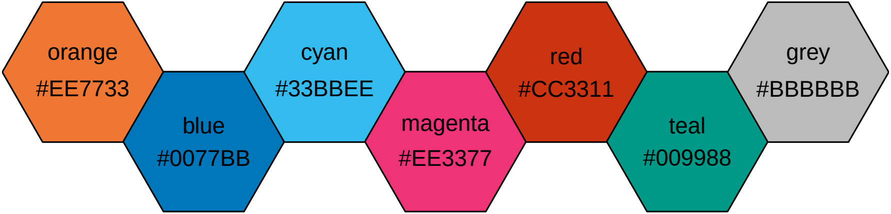
```
highcontrast <- color("high contrast")
plot_scheme(highcontrast(3), colours = TRUE, names = TRUE, si
```



Paul Tol’s color schemes

Tol (2021)

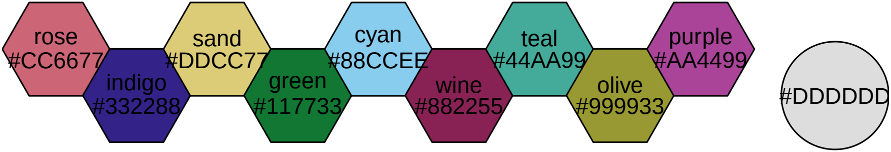
```
vibrant <- color("vibrant")
plot_scheme(vibrant(7), colours = TRUE, names = TRUE, size =
```



Paul Tol’s color schemes

Tol (2021)

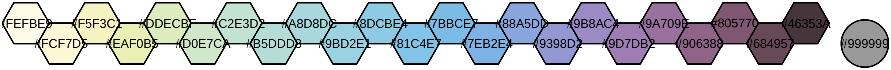
```
muted <- color("muted")
plot_scheme(muted(9), colours = TRUE, names = TRUE,
            size = 0.9)
```



Paul Tol’s color schemes

Tol (2021)

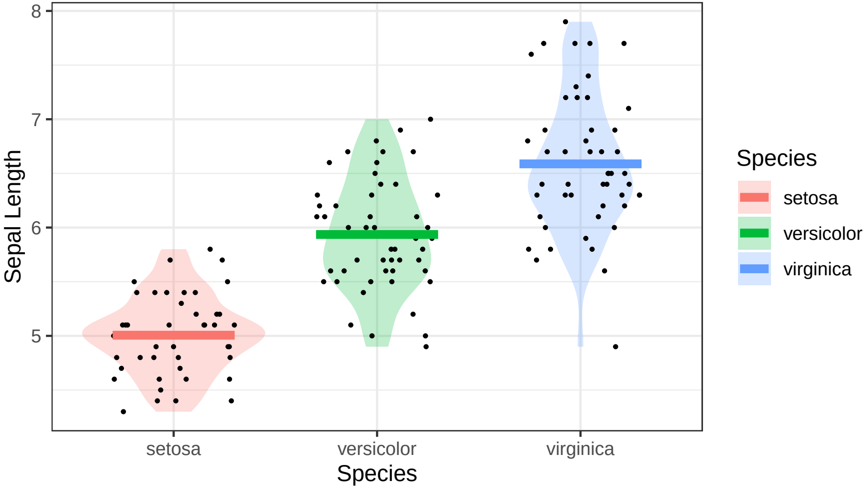
```r
iridescent <- color("iridescent")
plot_scheme(iridescent(23), colours = TRUE, size = 0.5)
```



Visualizing uncertainty

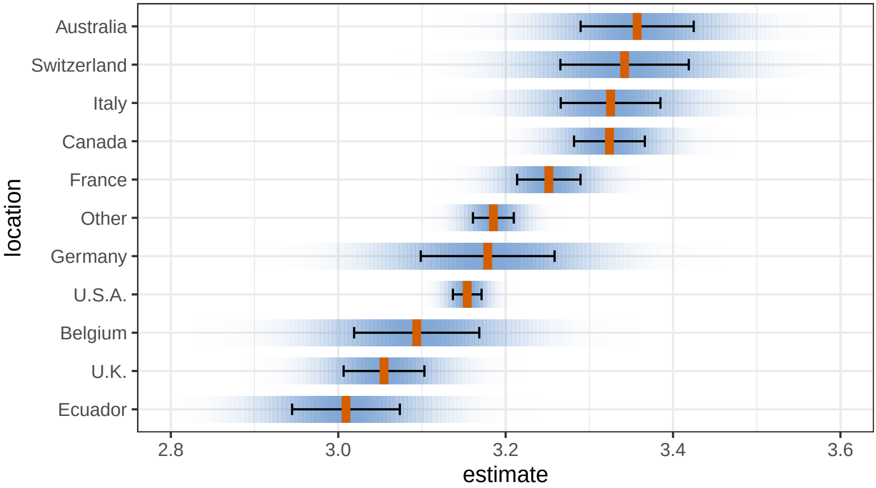
Show the underlying data.

```
ggplot(iris, aes(Species, Sepal.Length, fill = Species)) +
    geom_violin(alpha = 0.25, color = NA) +
    geom_point(position = position_jitter(width = 0.3, height = 0), size = 0.5) +
    geom_hpline(aes(colour = Species), stat = "summary", width = 0.6,
                fun = 'mean')
```



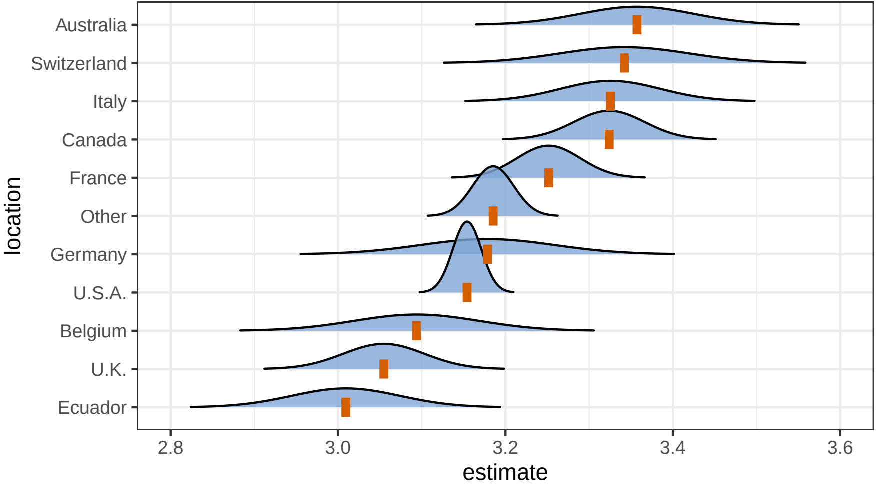
Shaded confidence strips.

```
ggplot(cacao_means, aes(x = estimate, y = location)) +
    stat_confidence_density(aes(moe = std.error), confidence = 0.68, fill = "#81A7D6", height = 0.7) +
    geom_errorbarh(aes(xmin = estimate - std.error, xmax = estimate + std.error), height = 0.3) +
    geom_vpline(aes(x = estimate), size = 1.5, height = 0.7, color = "#D55E00")
```



Confidence densities.

```
ggplot(cacao_means, aes(x = estimate, y = location)) +
    stat_confidence_density(
        aes(moe = std.error, height = after_stat(density)), geom = "ridgeline",
        confidence = 0.68, fill = "#81A7D6", alpha = 0.8, scale = 0.08, min_height = 0.1) +
    geom_vpline(aes(x = estimate), size = 1.5, height = 0.5, color = "#D55E00")
```



References I

- Paul Tol's color schemes: https://personal.sron.nl/~pault/; vignettes: https://cran.r-project.org/web/packages/khroma/vignettes/tol.html
- Clause Wilke: https://wilkelab.org/ungeviz/index.html

Tol, P. (2021). Introduction to colour schemes. *Paul Tol's Notes: Color Schemes and Templates*.

Tufte, E. R. (2001). *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.