Social Science Inquiry III

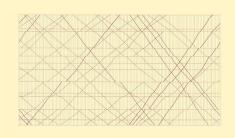## Week 1: Course introduction

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/bearl

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

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

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

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

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

Graphical excellence (in Tufte's words)
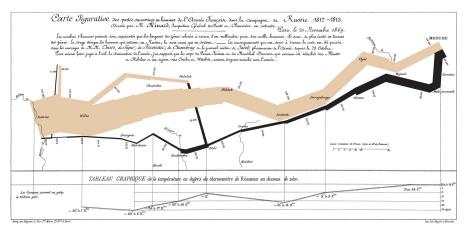
- Graphical displays should:
    - show the data
    - induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
    - avoid distorting what the data have to say
    - present many numbers in a small space
    - make large data sets coherent

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

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

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

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

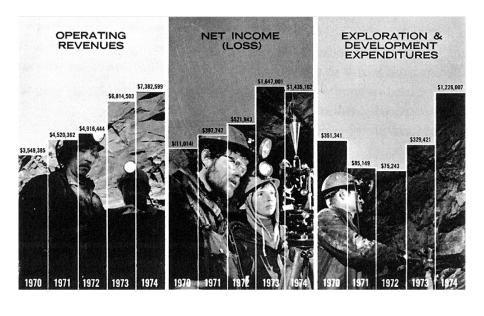Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.

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

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

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.

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.

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.
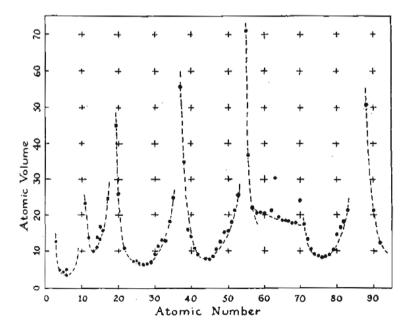
OPERATING REVENUES

Year	Amount
1970	$3,549,385
1971	$4,520,362
1972	$4,916,444
1973	$6,814,503
1974	$7,382,599

NET INCOME (LOSS)

Year	Amount
1970	$(11,014)
1971	$397,747
1972	$521,943
1973	$1,647,001
1974	$1,435,102

EXPLORATION & DEVELOPMENT EXPENDITURES

Year	Amount
1970	$351,341
1971	$85,149
1972	$75,243
1973	$329,421
1974	$1,226,007

9

Five priniples in the theory of data graphics

- Above all else show the data.

Five priniples in the theory of data graphics

- Above all else show the data.

- Maximize the data-ink ratio.

Five priniples in the theory of data graphics

- Above all else show the data.

- Maximize the data-ink ratio.

- Erase non-data ink.

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.

Five priniples in the theory of data graphics

- Above all else show the data.

- Maximize the data-ink ratio.

- Erase non-data ink.
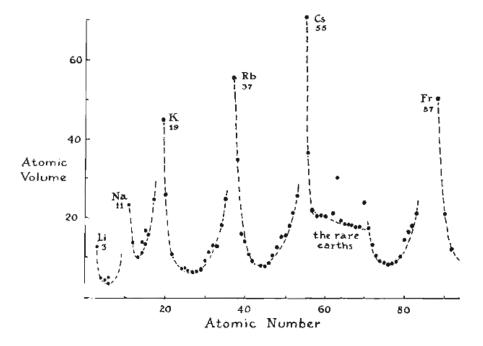
- Erase redundant data-ink.

- Revise and edit.
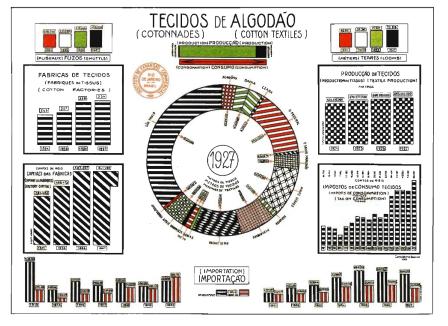
Linus Pauling, *General Chemistry* (San Francisco, 1947), p. 64.

12

Avoid "chartjunk"
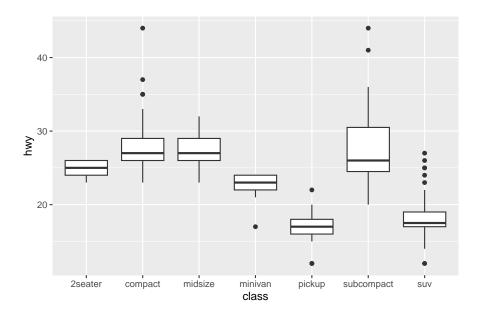
- Chartjunk: Non-essential or redundant information in graphics.

- Avoid distractions that do not enhance understanding.

- No meaningless patterns or dimensions, no grids, no chart-as-decoration.
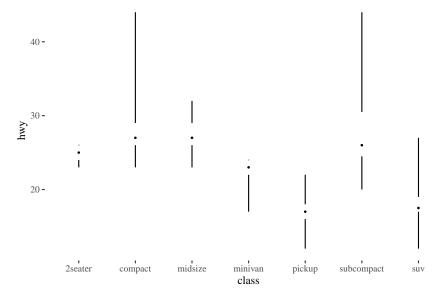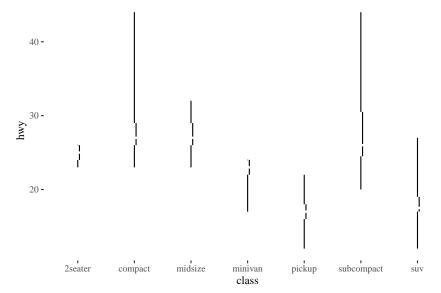
Institute de Expansão Commercial, *Brasil: Graphicos Economics-Estatisticas*, (Rio de Janeiro, 1929) p. 15.
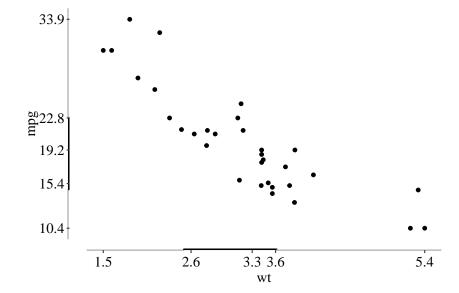
14
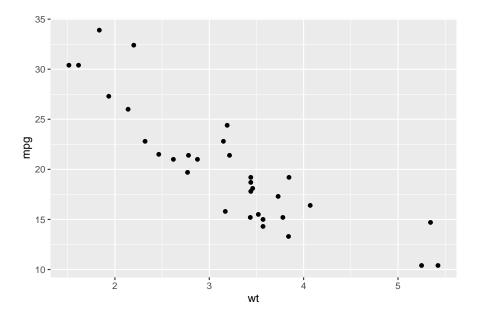
Data-ink maximization
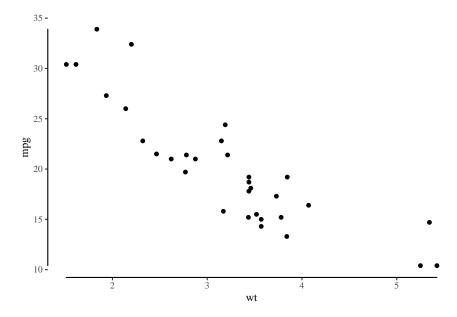
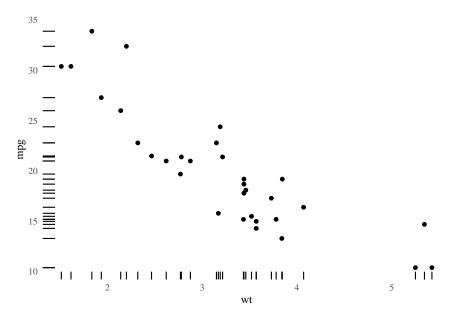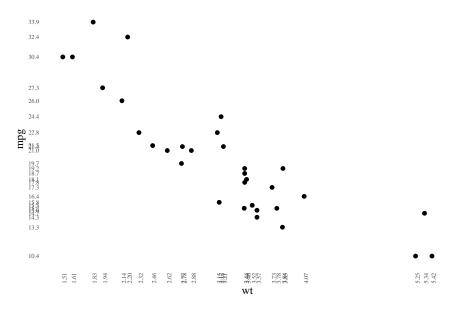- New graphical forms.

20

Multi-functioning graphical elements

- Use elements that serve more than one purpose.

Multi-functioning graphical elements

- Use elements that serve more than one purpose.

- Combine text and images for efficient communication.

Other principles

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

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.

Aesthetics

- Employ visual balance.

- Combine words, numbers, and pictures–all together in a graphic.

Aesthetics

- Employ visual balance.

- Combine words, numbers, and pictures–all together in a graphic.

- Lines should be thin.

Aesthetics

- Employ visual balance.

- Combine words, numbers, and pictures–all together in a graphic.

- Lines should be thin. Add weight to add meaning.

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.

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.

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.

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.

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

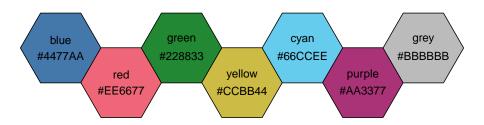- Use colorblind-friendly palettes to ensure effective
  communication.

Principles for creating your ownd colorblind-friendly palettes

- Use high contrast between colors.

Principles for creating your ownd colorblind-friendly palettes

- Use high contrast between colors.
- Avoid using red and green together.
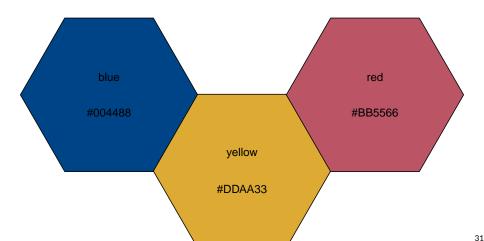
Principles for creating your ownd colorblind-friendly palettes

- Use high contrast between colors.
- Avoid using red and green together.
- Use shades to differentiate data points.

Principles for creating your ownd colorblind-friendly palettes

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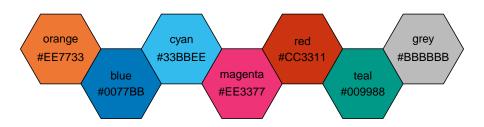- You can test your visualizations with colorblindness simulation tools.

Paul Tol's color schemes

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


blue
#4477AA

red
#EE6677

green
#228833

yellow
#CCBB44

cyan
#66CCEE

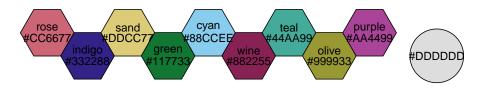purple
#AA3377

grey
#BBBBBB

30

Paul Tol's color schemes
Tol (2021)

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

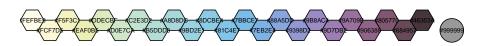# Paul Tol's color schemes

```
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)
```
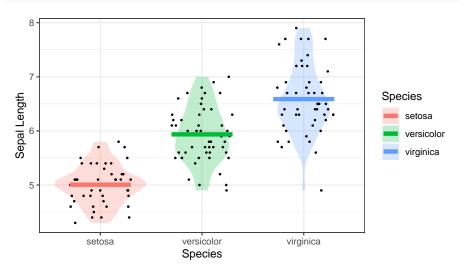
Paul Tol's color schemes
Tol (2021)

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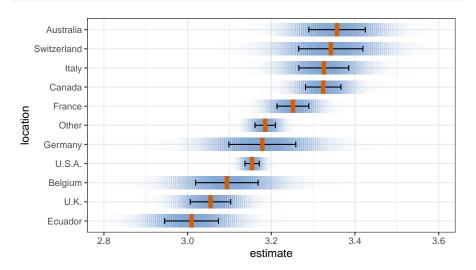
Visualizing uncertainty

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')
```
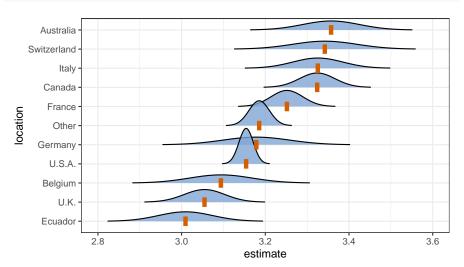
Shaded confidence strips.

```
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.