Data Visualization with R

E. Le Pennec

Fall 2015

Outline

Focus of this lecture

- Standard data visualization techniques,
- Review of various graphical techniques,
- Principle of good data presentation,
- Example of implementation with R.

Not the focus of this lecture

- *Infographics*
- Cognitive aspect of data perception...

Goal

- Exposure to various plotting techniques.
- Proof of concept with R.
- *Visualize* the power of appropriate data graphics techniques

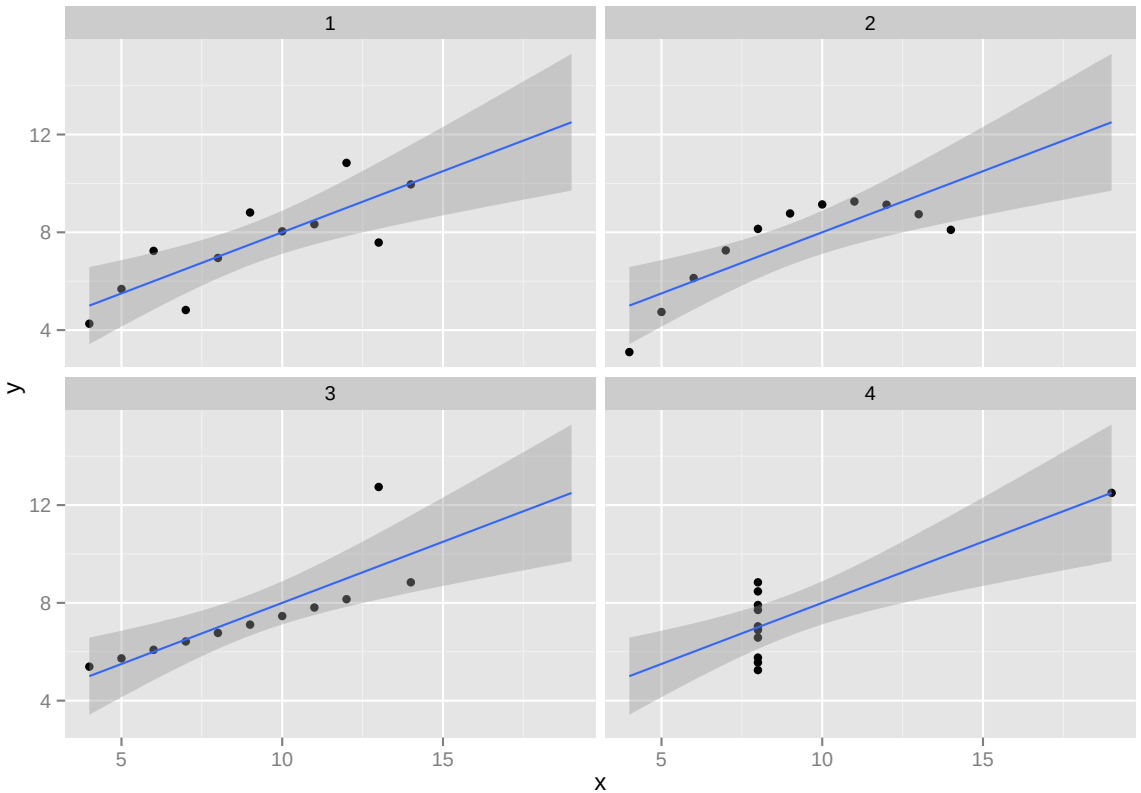
- Credit: this lecture is inspired from one workshop of R. Womack.

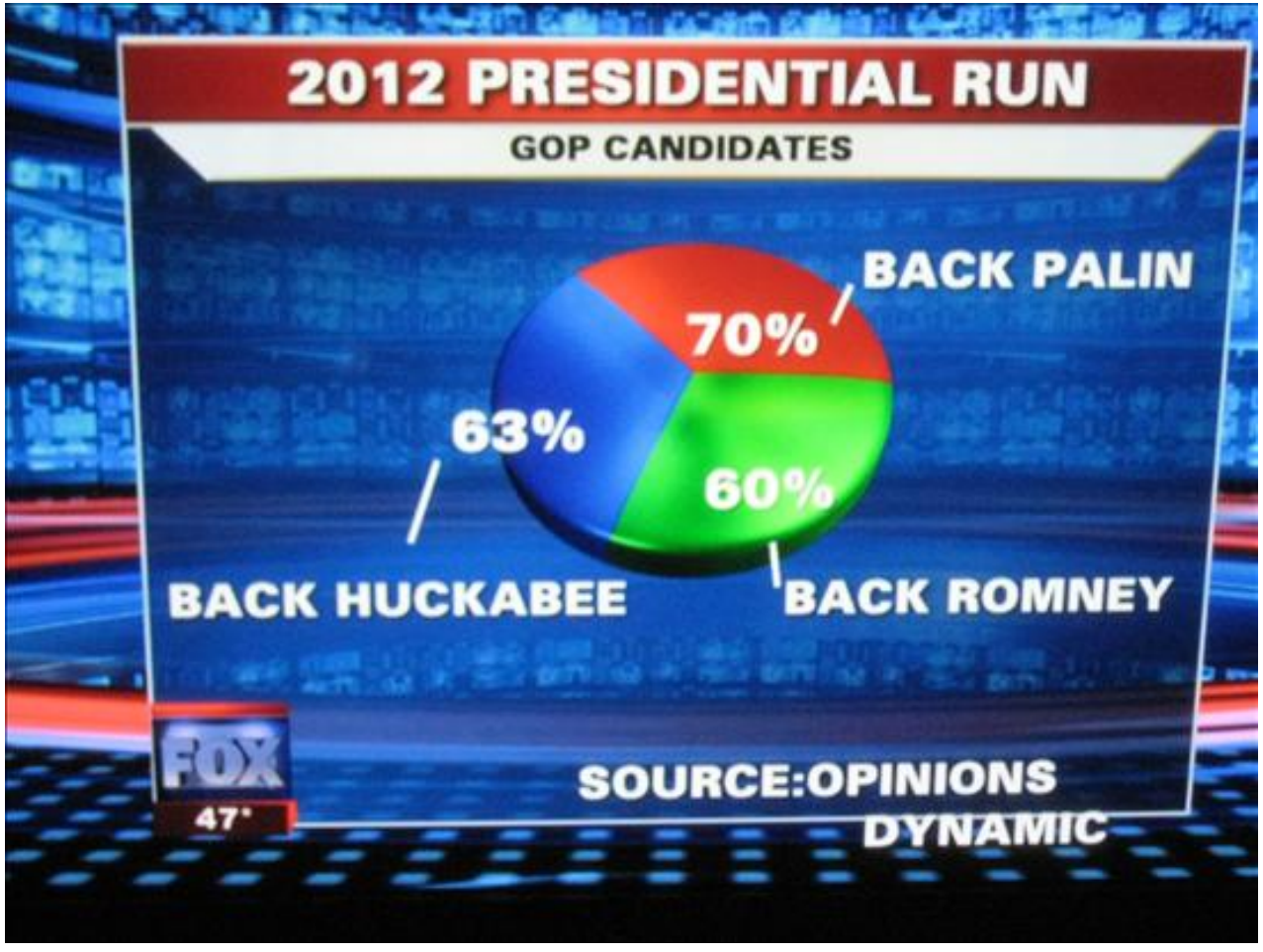
- Data visualization can:
  - provide a clear understanding of patters in data
  - detect hidden structures in data
  - condense information

Introduction

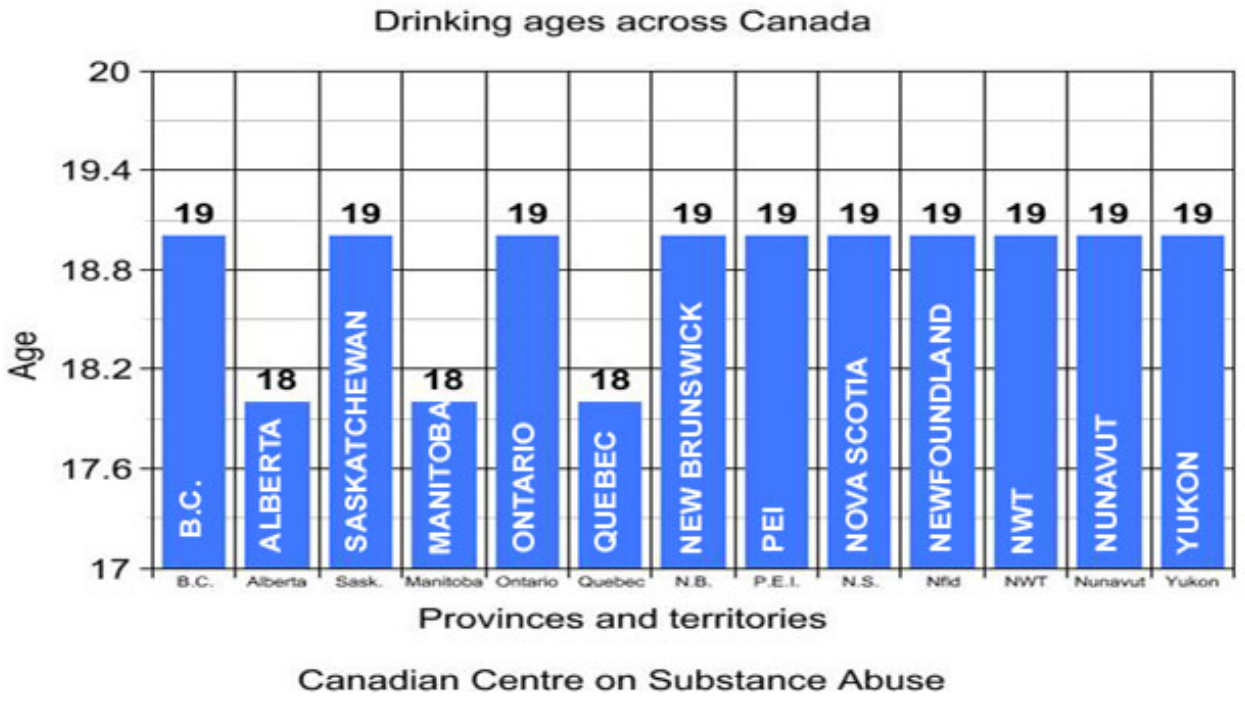
Why Data Visualization?

- Data visualization can:
  - provide a clear understanding of patters in data
  - detect hidden structures in data
  - condense information
- Anscombe's quartet example:

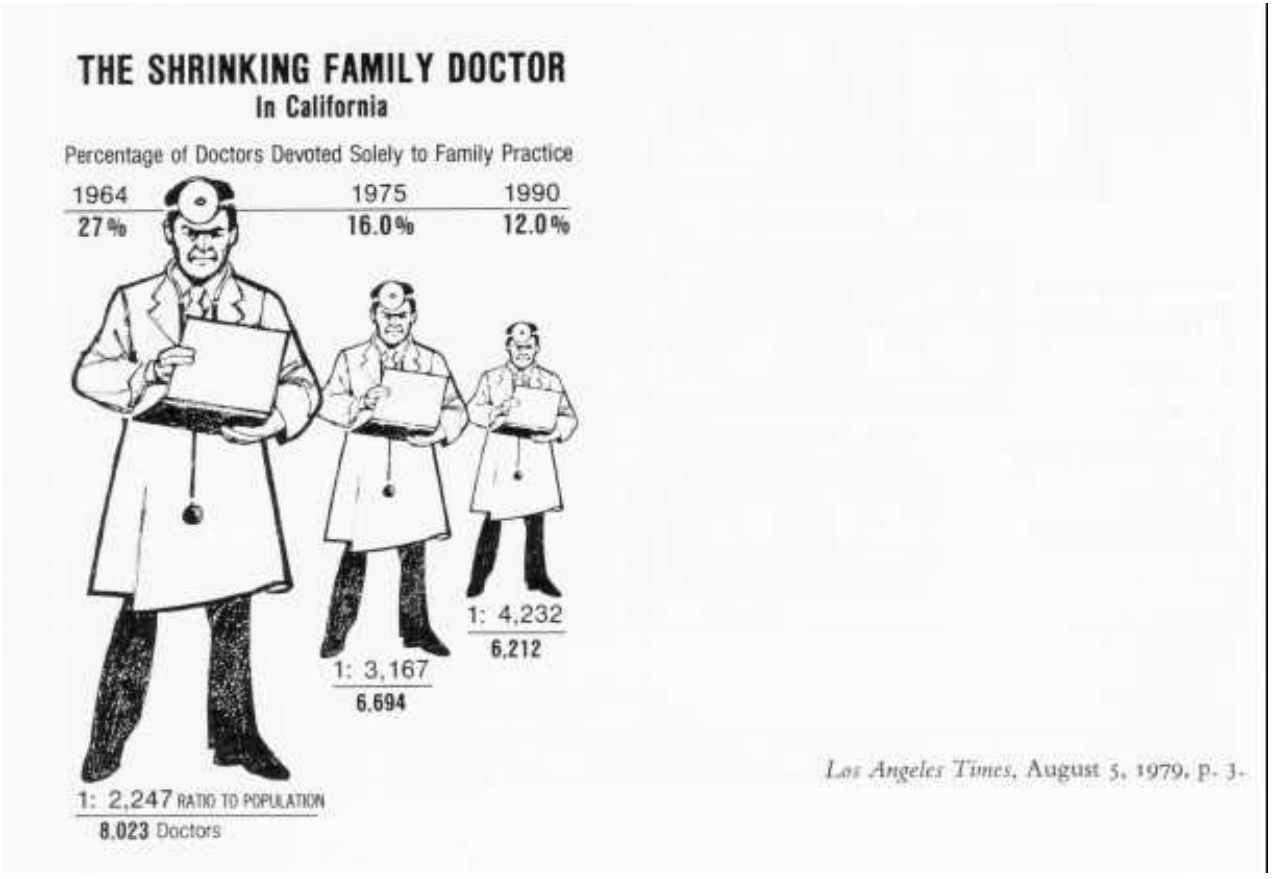



2012 PRESIDENTIAL RUN
GOP CANDIDATES
BACK PALIN
70%
63%
60%
BACK HUCKABEE
BACK ROMNEY
FOX
47'
SOURCE:OPINIONS
DYNAMIC

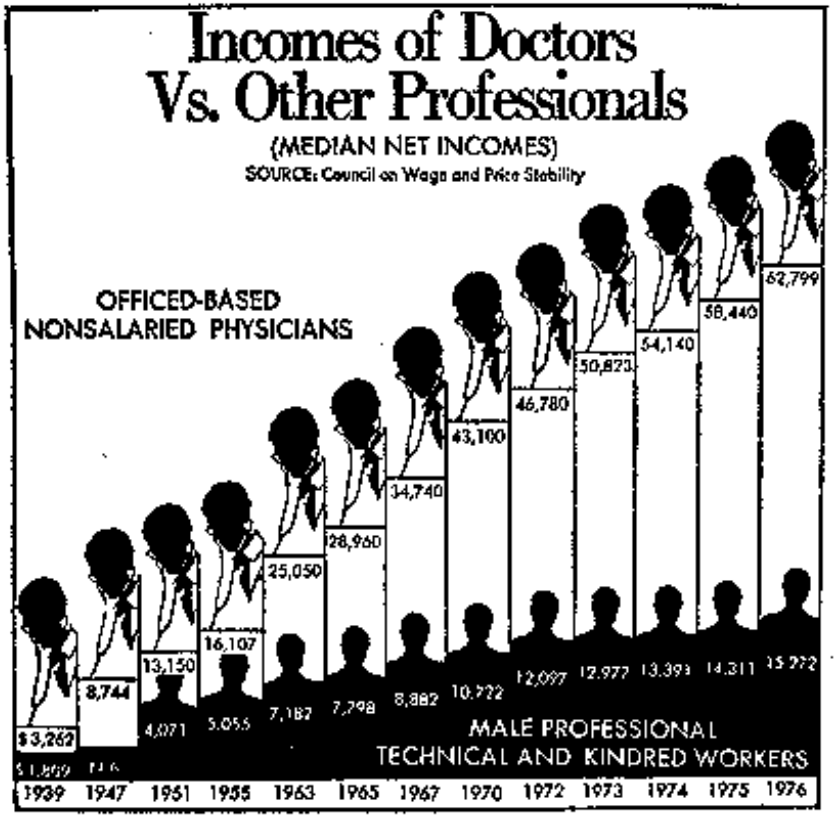
- No comment!



- Clutter Issue


THE SHRINKING FAMILY DOCTOR
In California
Percentage of Doctors Devoted Solely to Family Practice
1964
27%
1975
16.0%
1990
12.0%
1: 2,247 RATIO TO POPULATION
8,023 Doctors
1: 3,167
6,694
1: 4,232
6,212
Los Angeles Times, August 5, 1979, p. 3.

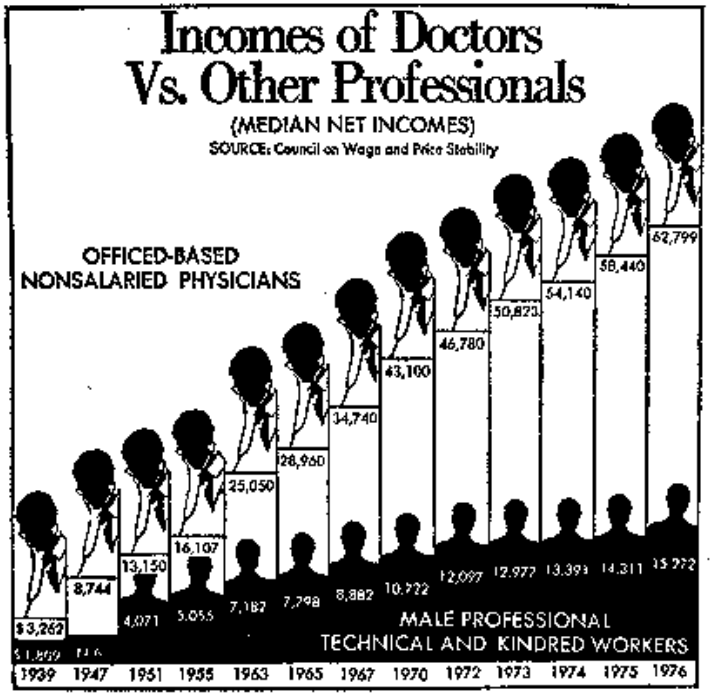
- Lie factor!

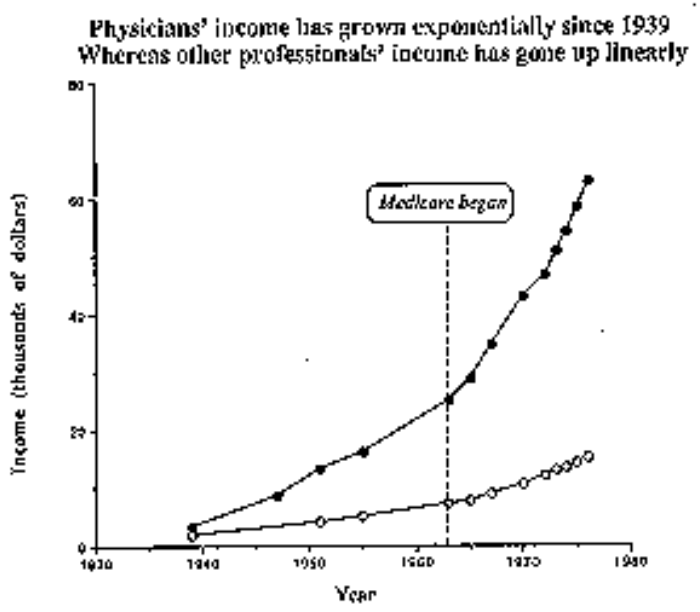

Incomes of Doctors
Vs. Other Professionals
(MEDIAN NET INCOMES)
SOURCE: Council on Wage and Price Stability
OFFICED-BASED
NONSALARIED PHYSICIANS
MALE PROFESSIONAL
TECHNICAL AND KINDRED WORKERS
$3,262
8,744
13,150
16,107
25,050
28,960
34,740
43,100
46,780
50,823
54,140
58,440
62,799
4,071
5,055
7,187
7,298
8,882
10,772
12,097
12,977
13,393
14,311
15,272
1939
1947
1951
1955
1963
1965
1967
1970
1972
1973
1974
1975
1976

- Scale issue...

Introduction

Bad Data Visualization





Introduction
Bad Data Visualization

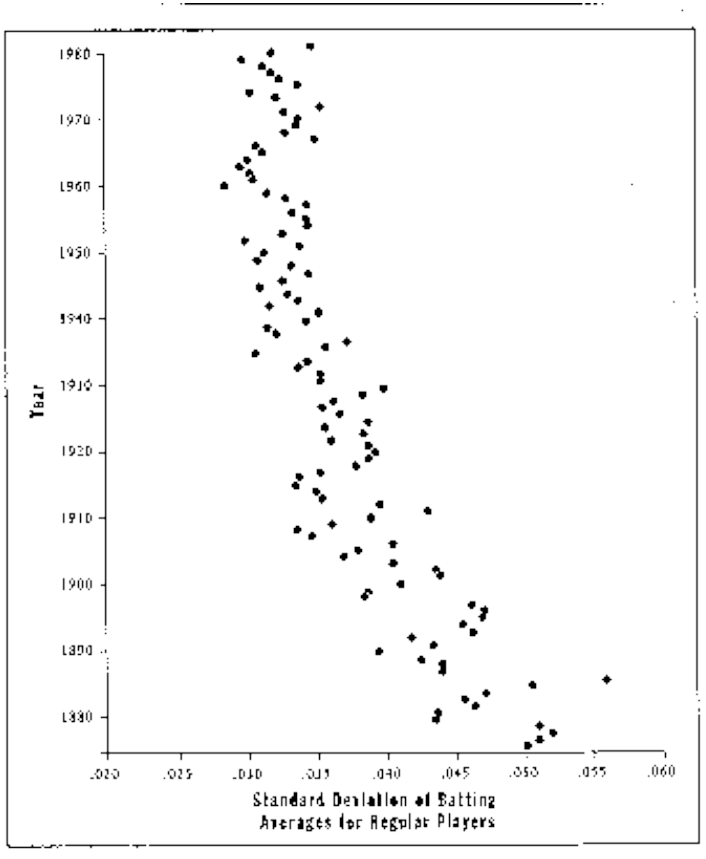


FIGURE 16
Standard deviation of batting averages for all full-time players by year for the first 100 years of professional baseball. Note the regular decline.

- Goosed up...

Introduction
Bad Data Visualization

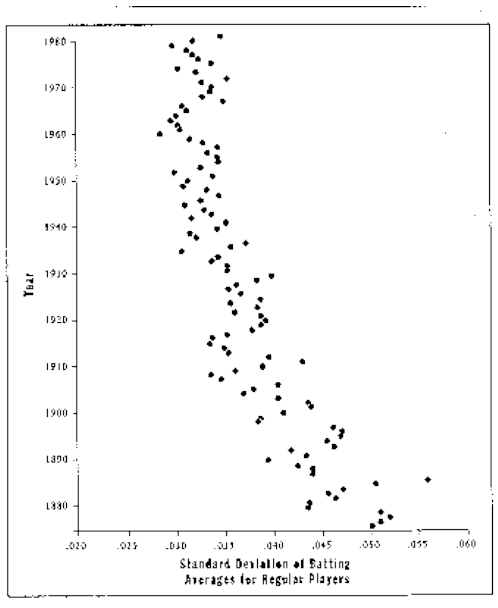
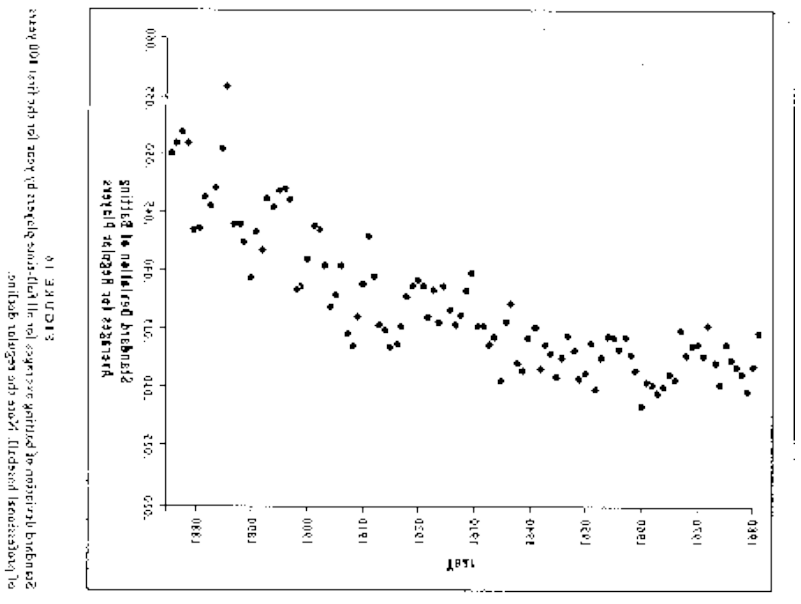
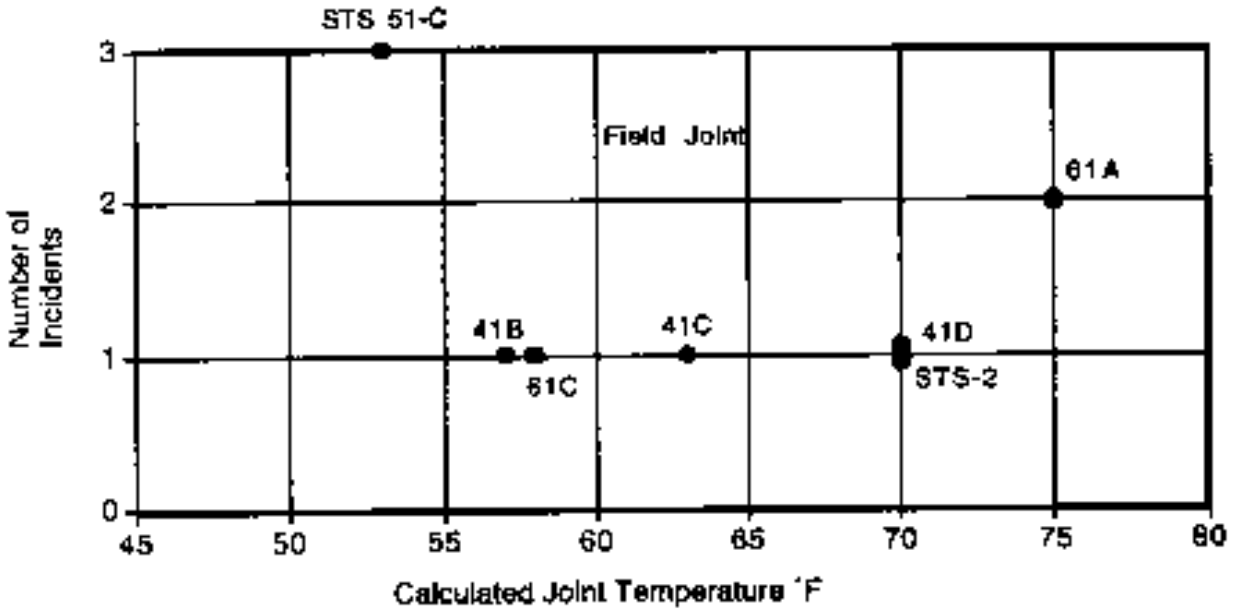


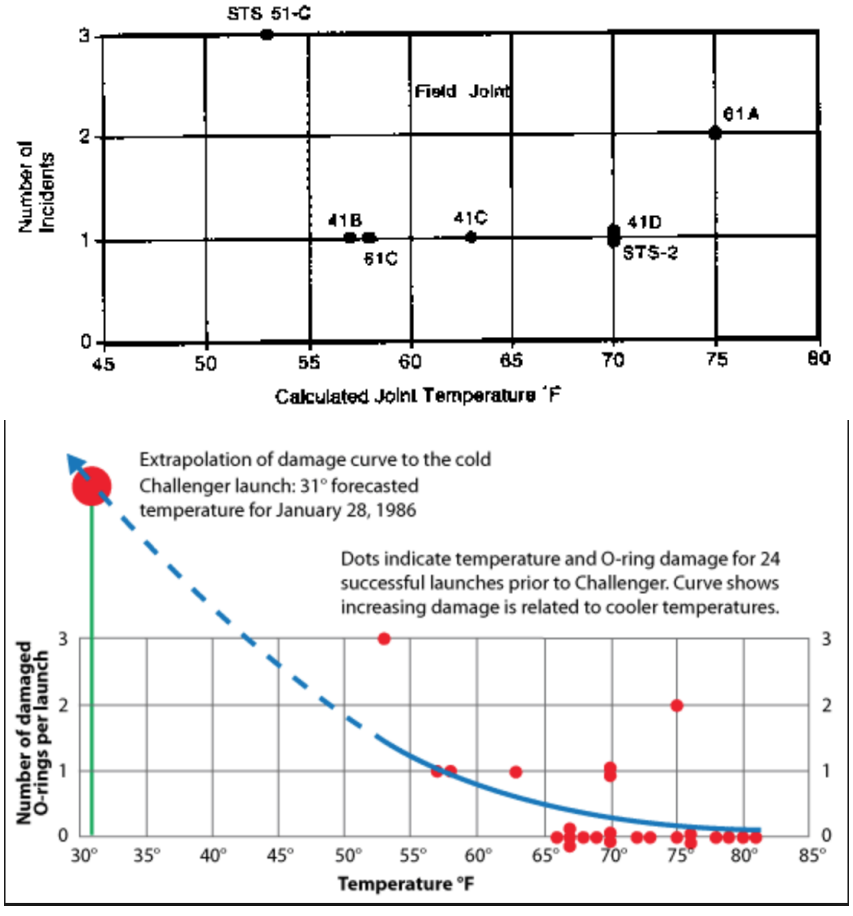
FIGURE 16
Standard deviation of batting averages for all full-time players by year for the first 100 years of professional baseball. Note the regular decline.



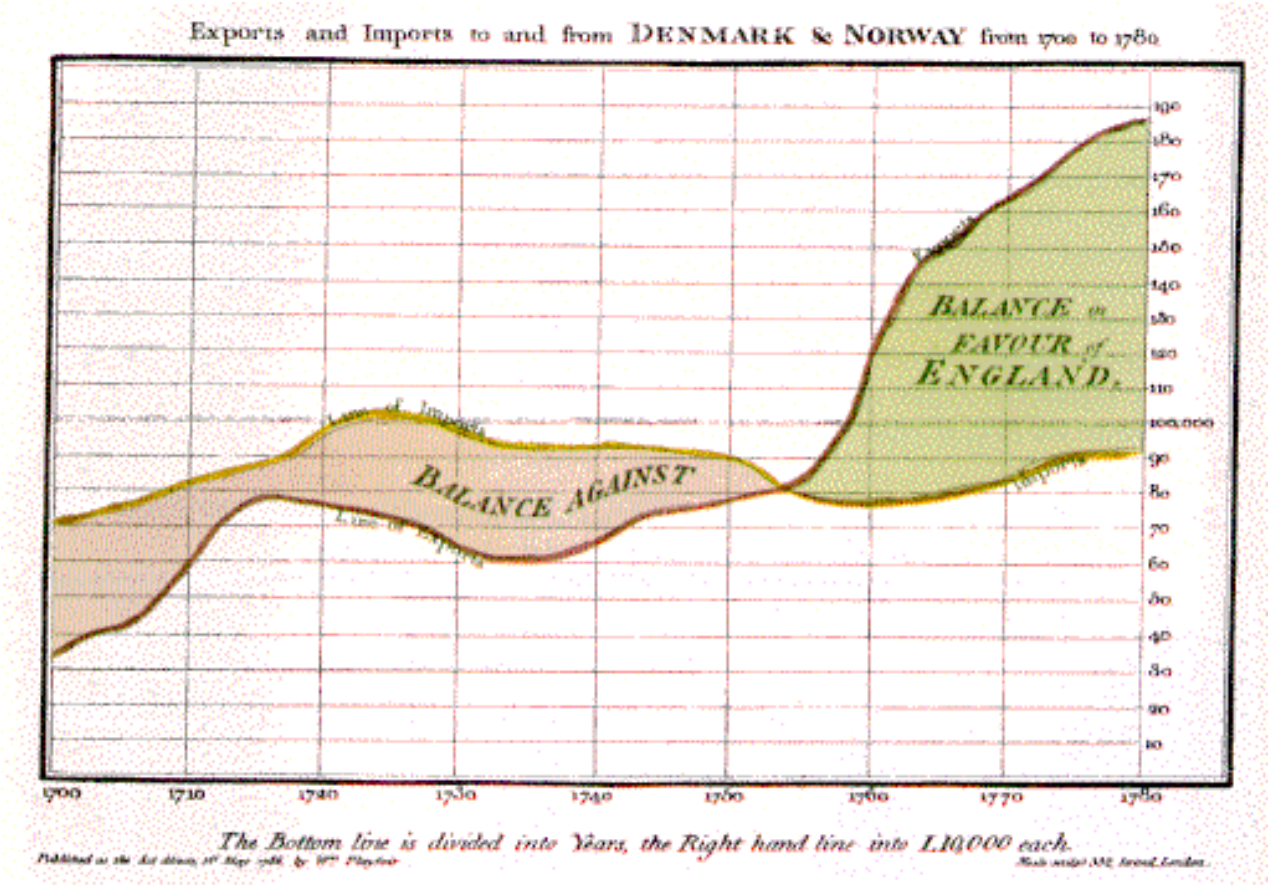


- Catastrophic!

Introduction
Bad Data Visualization

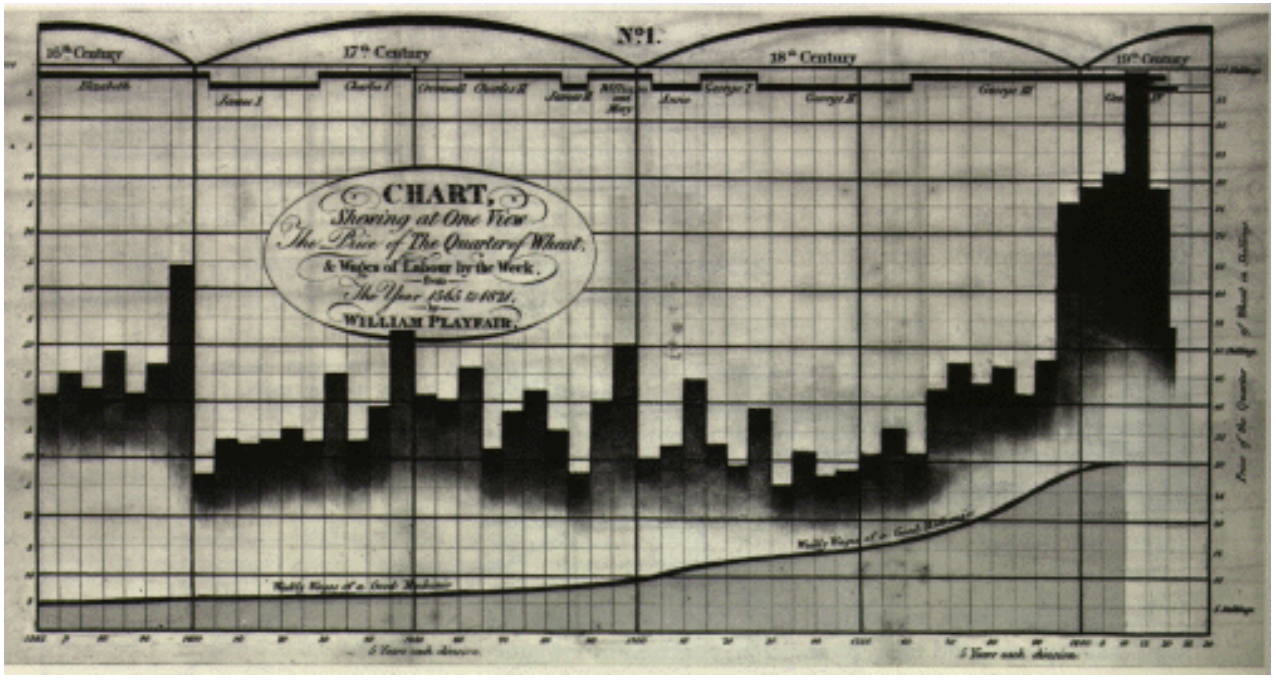


- William Playfair (1759-1823) generally viewed as the inventor of most of the common graphical forms used to display data: line plots, bar chart and pie chart



Historical Milestones

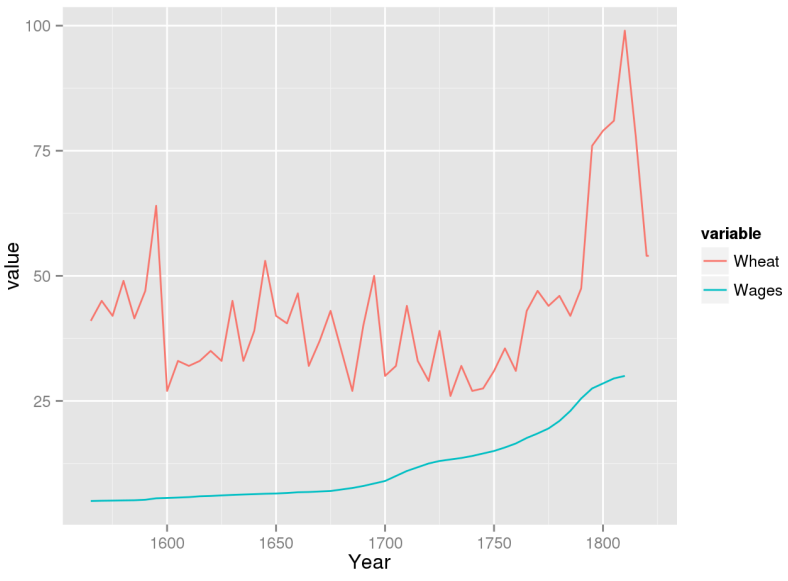
Playfair

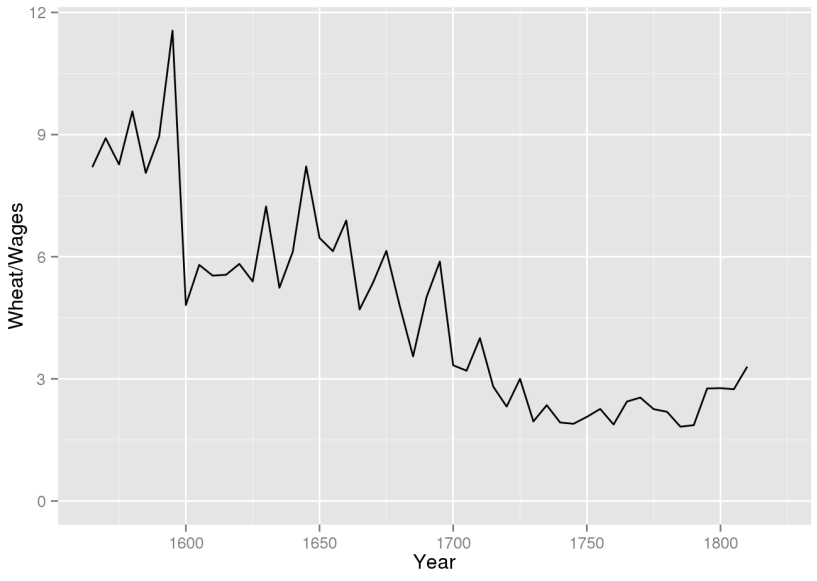


- Unfortunately often flawed...

Historical Milestones

Playfair

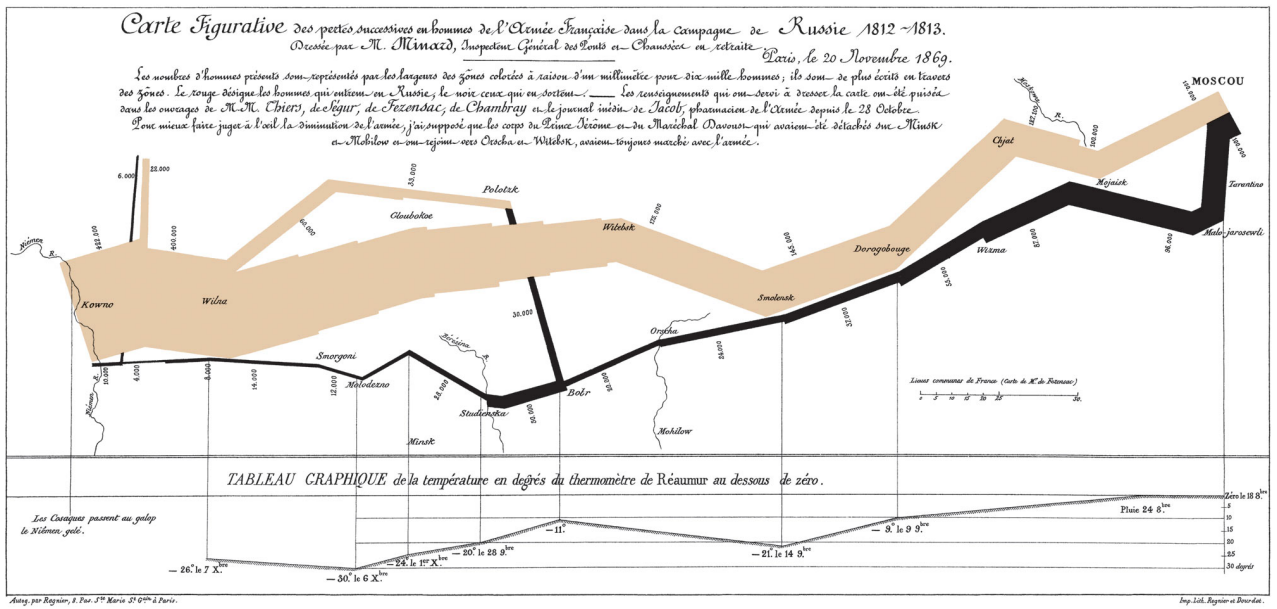




Historical Milestones

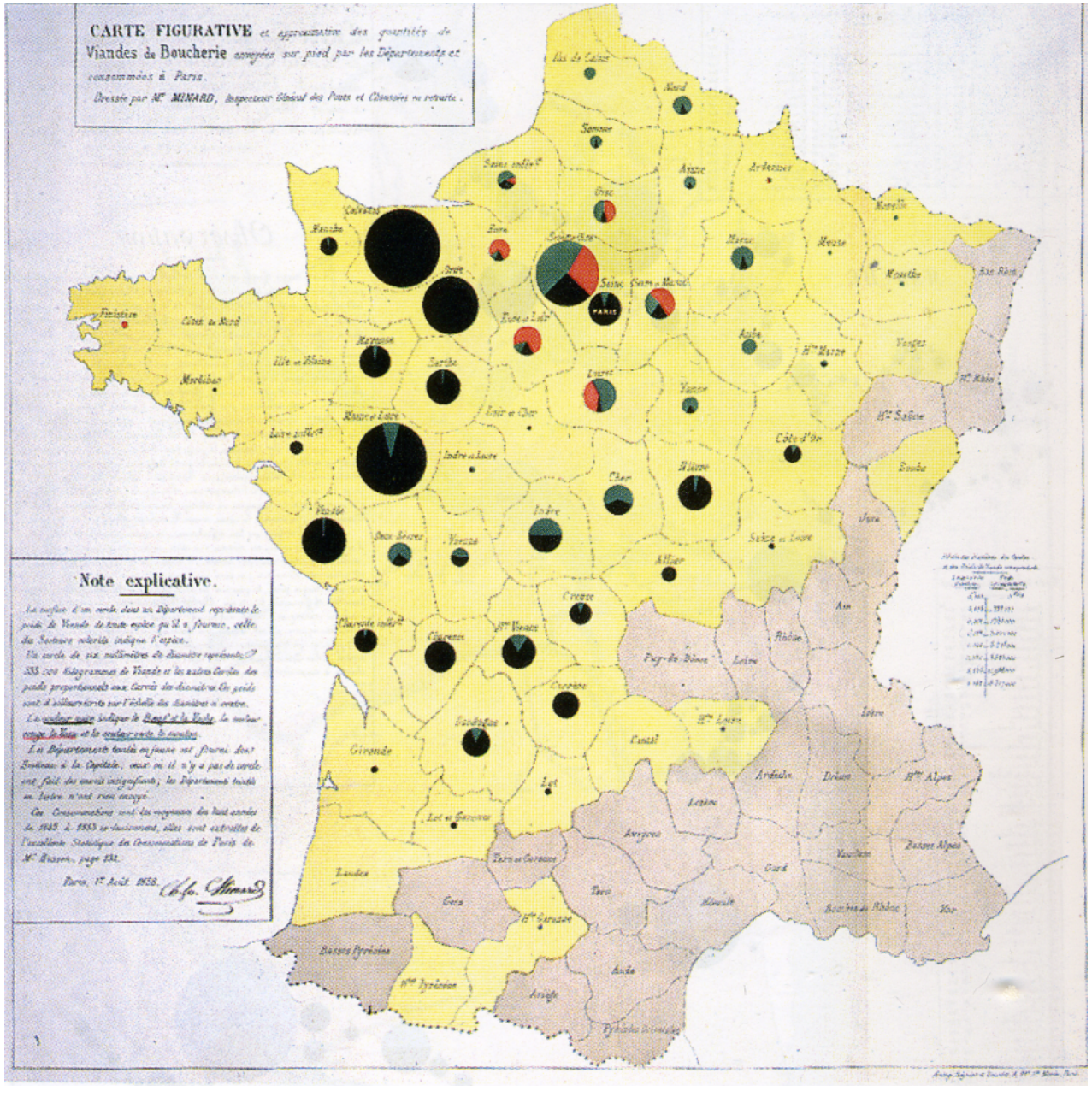
Minard

- Charles Minard (1781-1870) contributes significantly in the field of information graphics in civil engineering and statistics and in particular in geographic maps.

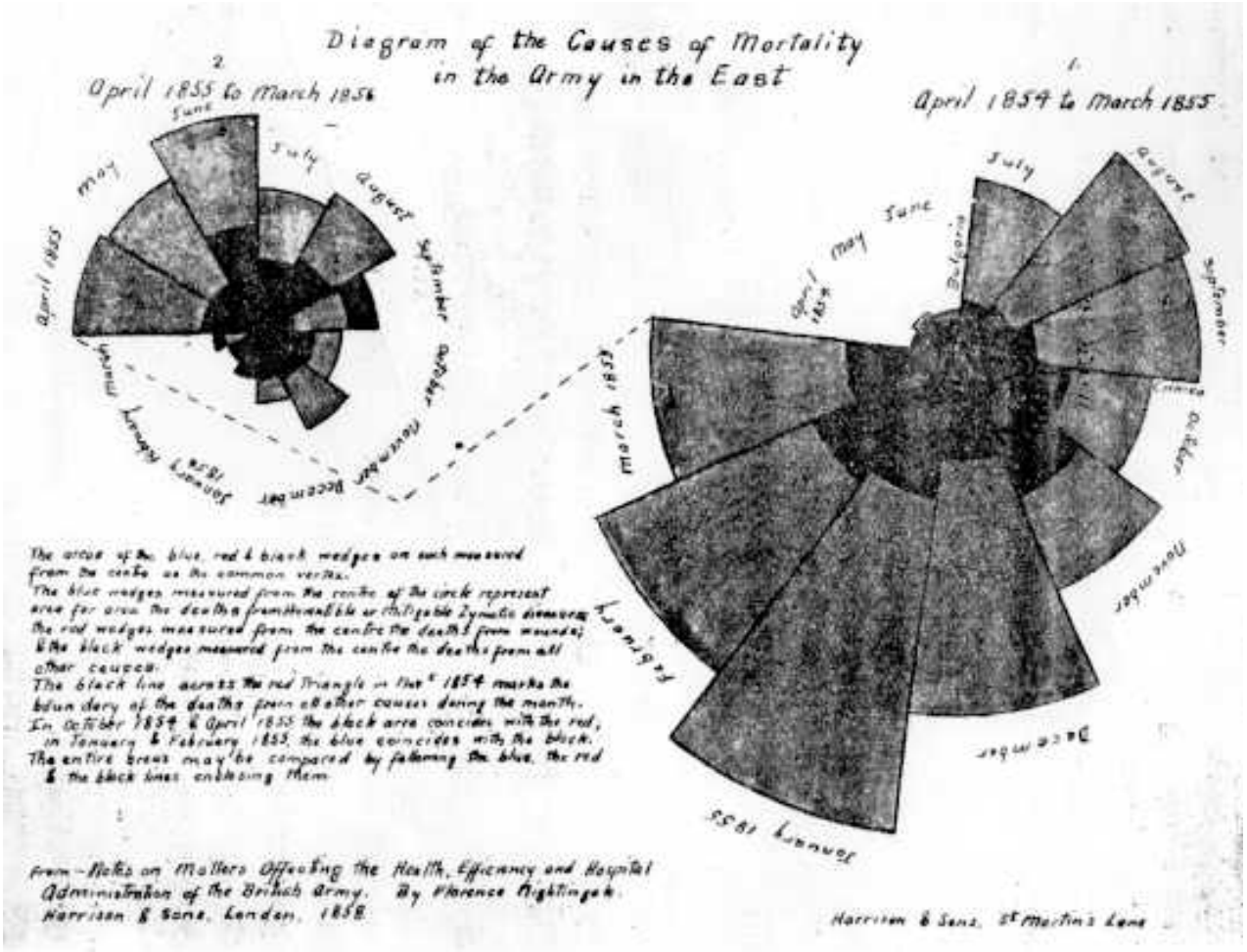


Historical Milestones

Minard



- Florence Nightingale (1820-1910) is mostly famous as a the mother of modern nursing. She also contributed to the use of graphical representation.

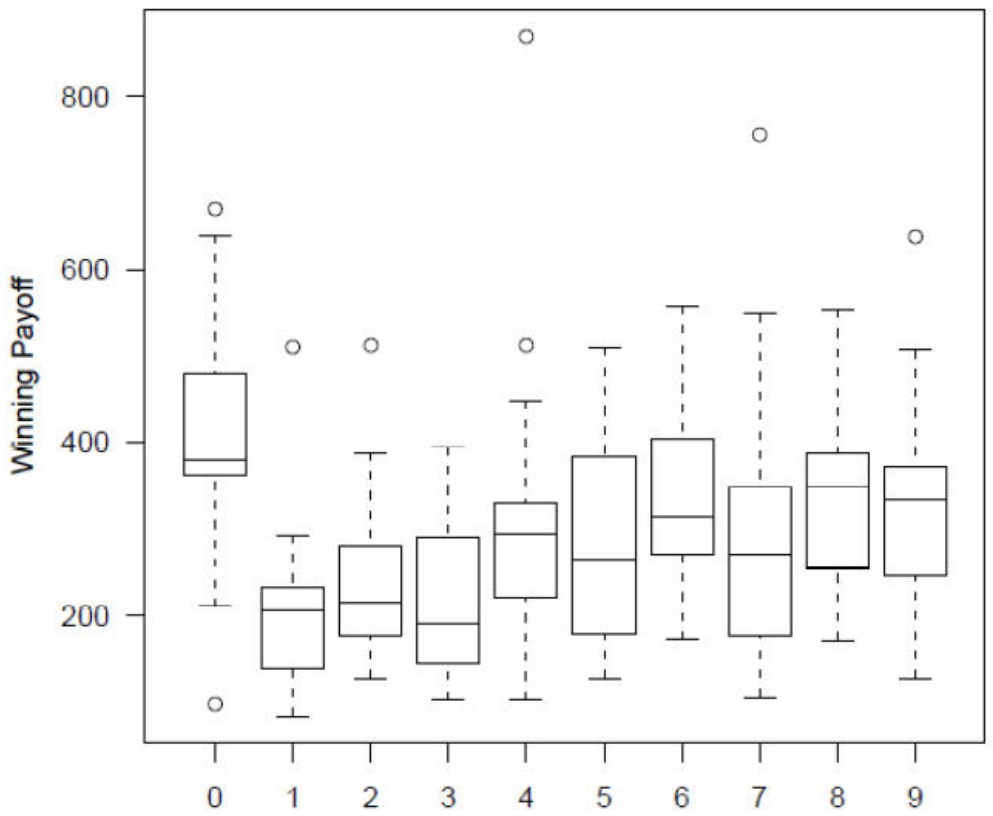


- John Snow (1813–1858) was an English physician. He is famous for tracing the source of a cholera outbreak in London.

Historical Milestones

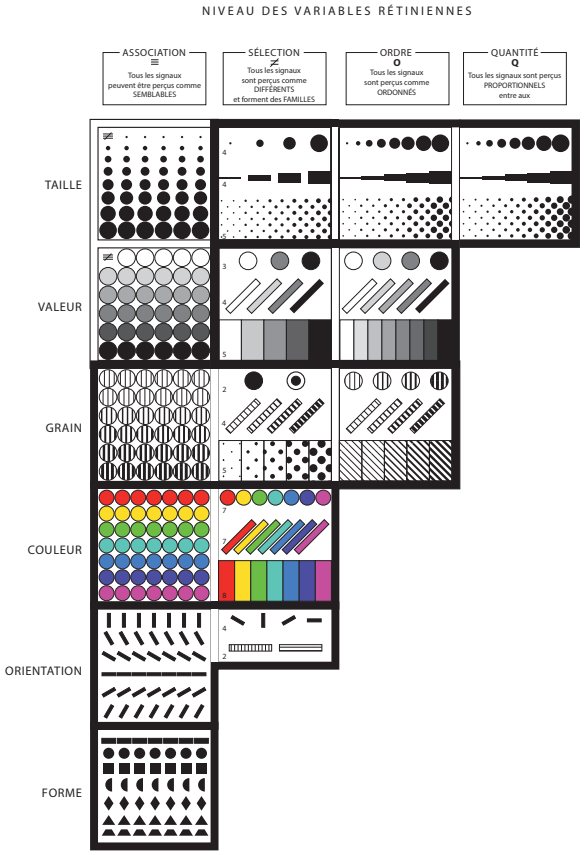
Fisher and Tuckey

- Ronald Fisher (1890-1962) and John Tukey (1915-2000): advance graphical methods for the analysis of data.
- Fisher: plot the data to understand relationships.
- Tukey promoted Exploratory Data Analysis!
- Tukey created the box plot and the stem and leaf plot.



Historical Milestones

Bertin



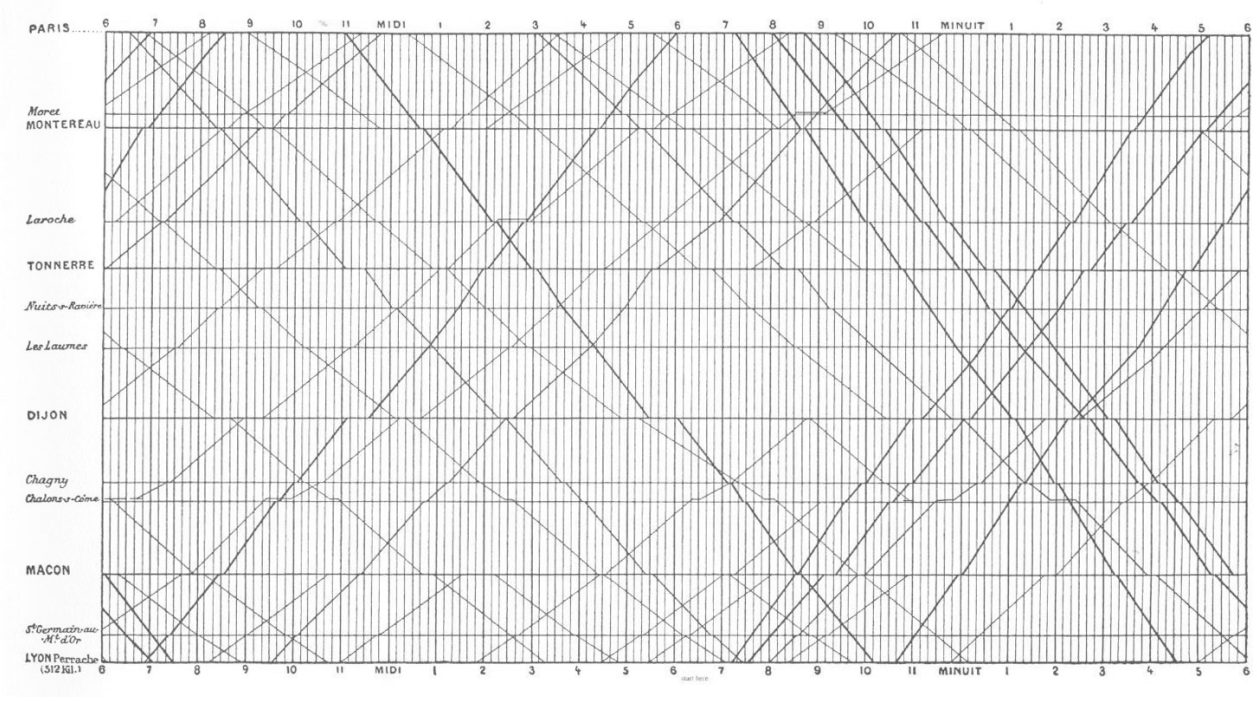
Jacques Bertin, "Sémiologie Graphique", 1973.

- Jacques Bertin (1918-2010): *sémiologie graphique*!
- Systematic system of sign for information transmission.

Historical Milestones

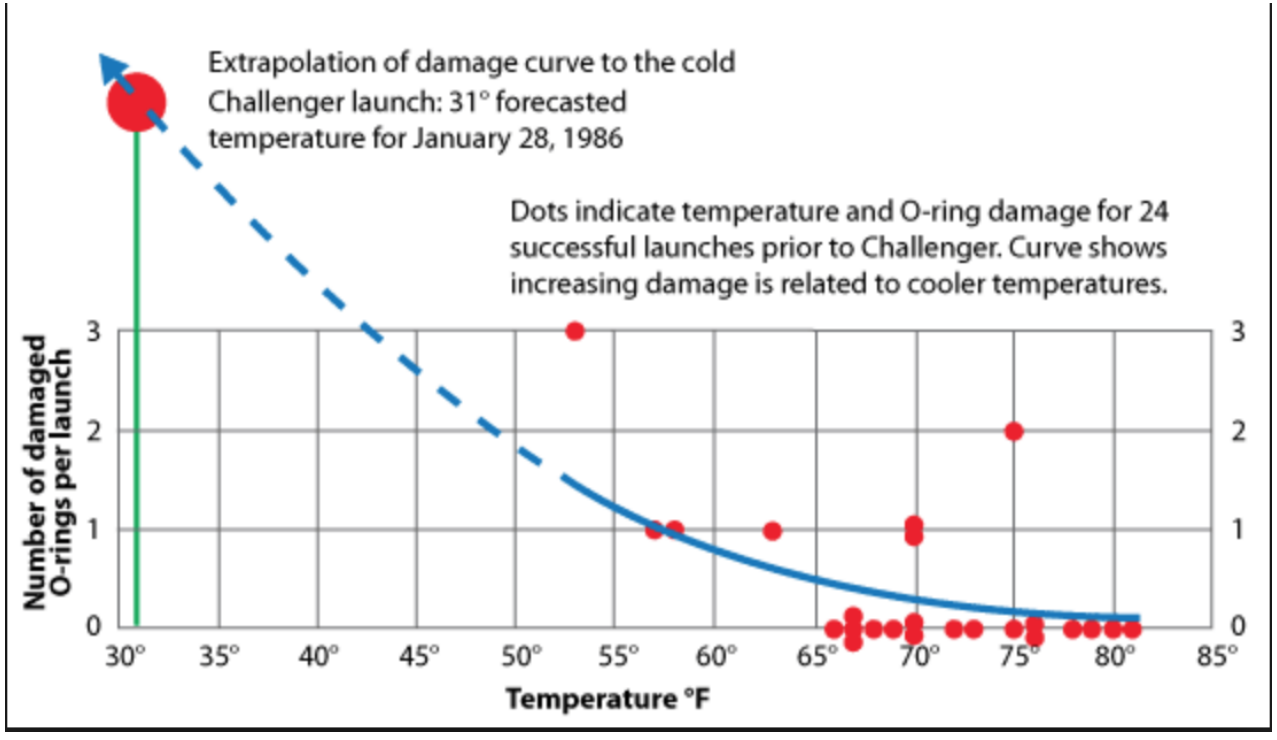
Tufte

- Edward Tufte (1942-): the most widely know works on data visualization.
- Highly compressed, elegant, and informative data, as expressed in dense printed graphics.
- Importance of *beauty* aspect...



Historical Milestones

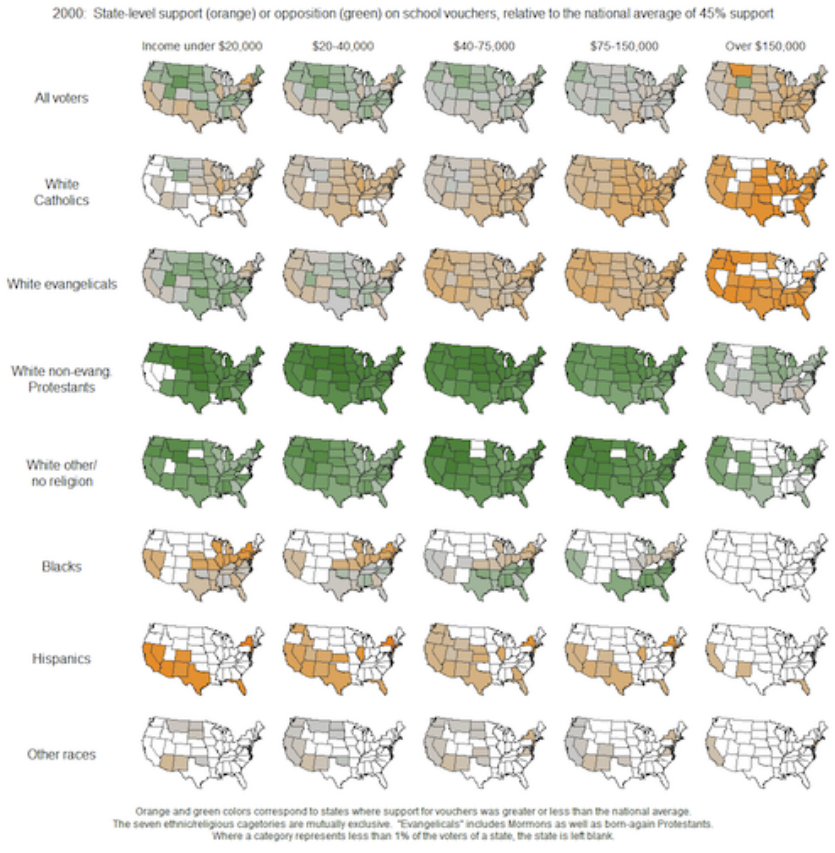
Tufte



- Challenger corrected!

Historical Milestones

Tufte



- Small Multiple

Example sparklines in small multiple

Index	Day	Value	Change
Dow Jones		10765.45	–32.82 (–0.30%)
S&P 500		1256.92	–8.10 (–0.64%)

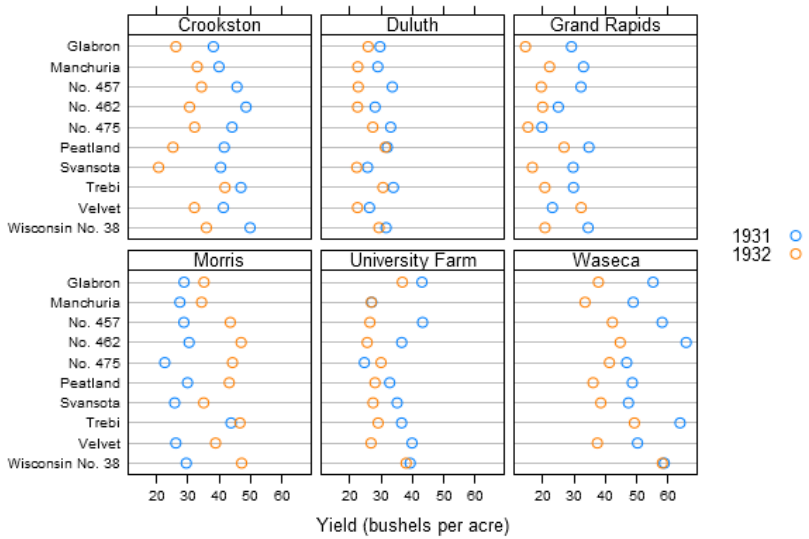
Example sparklines of the Dow Jones Industrial Average and S&P 500 on February 7, 2006

- Sparklines

- Tufte has developed and popularized numerous principles and terminology:
  - Graphics reveal data - show the data without distorting it - "above all else show the data"
  - Small multiple - understanding one slice makes understanding others easier
  - Lie factor - effect shown/effect in reality
  - Graphical Integrity - no lies, let data vary, not design
  - Data density - maximize data/ink ratio
  - Sparklines - seems they haven't caught on
  - chartjunk - self-explanatory
  - Powerpoint is responsible for most of the world's sorrows [The Cognitive Style of Powerpoint]

Historical Milestones

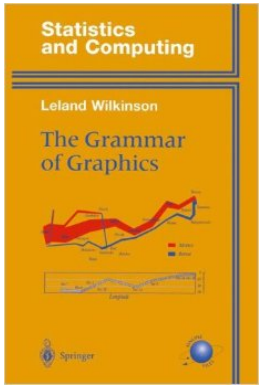
Cleveland



- William Cleveland's Elements of Graphing Data and Visualizing Data pioneered systematic considerations of legibility
- Cleveland is particularly known for promoting the dot plot as a alternative to bars and pies.
- The dot plot provides clarity and easy comparison of data.
- Cleveland also pioneered Trellis graphics that emphasizes comparison of multiple panels of data.

Historical Milestones

Wilkinson



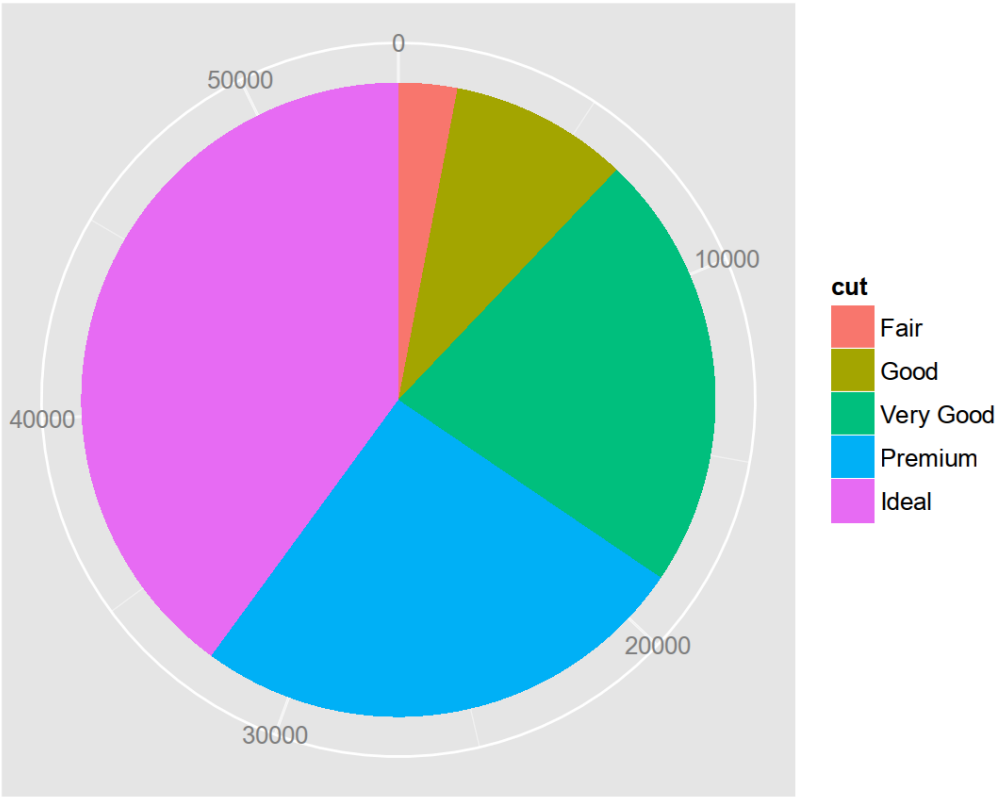
- The Grammar of Graphics, by Leland Wilkinson (1945-), was extremely influential in thinking about graphics
  - Grammar means "rules for art and science"
  - Specifies rules both mathematical and aesthetic
  - Earlier graph producers focused on aesthetics of static content
  - Dynamic graphics and scientific visualization, by contrast, require sophisticated designs to enable brushing, drill-down, zooming, linking
  - The Grammar of Graphics is easily adapted to this approach
- **ggplot2** (Bradley Wickham) is inspired by this formalism!

Historical Milestones

Wilkinson

- DATA - weighting, reshaping, counting, bootstrapping
- VARIABLES - transform, sort, log, ranking, residuals, quantiles
- ALGEBRA - nesting or blending data
- SCALES - nominal, ordinal, interval, ratio must be specified
- STATISTICS - static methods available to all graph types e.g, mean, sd, smoothing
- GEOMETRY - line, area, etc., along with modifiers like jitter and dodge
- COORDINATES - refers to the coordinate system of the graph (cartesian, polar, etc.)
- AESTHETICS - color, texture, size, position, etc. of the data points. Includes using color to classify.
- FACETS - subgroups, multiway tables
- GUIDES - legends, axes, color scales, keys

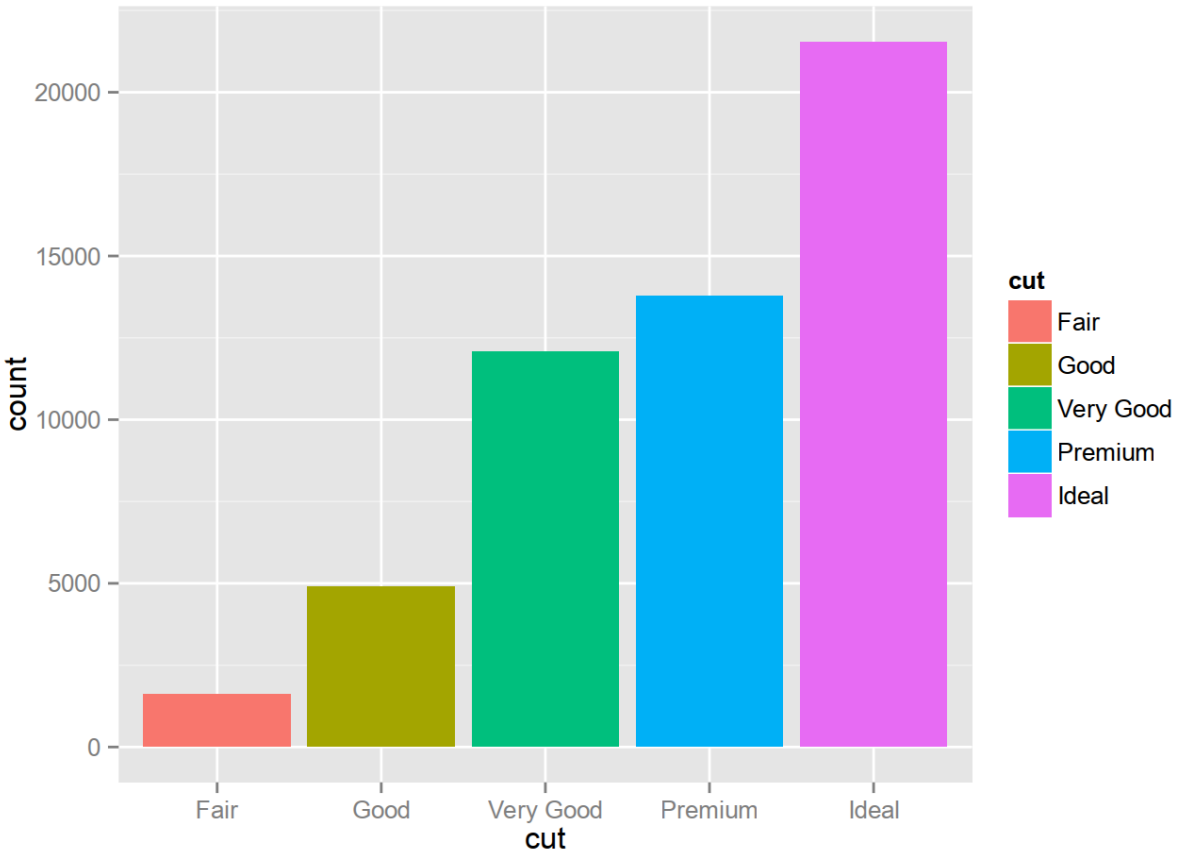
Classical graph
Pie



- Should not be used...

Classical graph

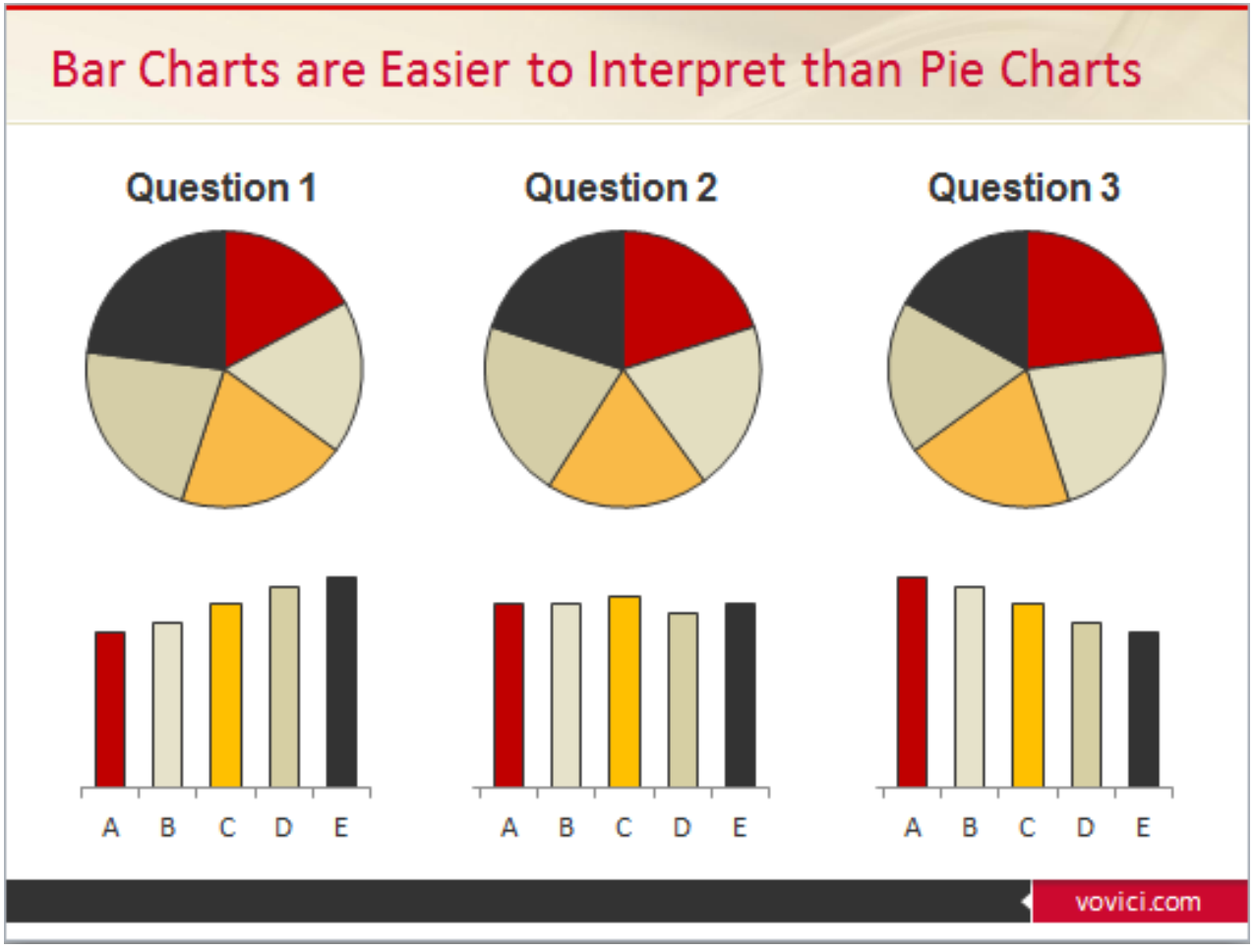
Bar



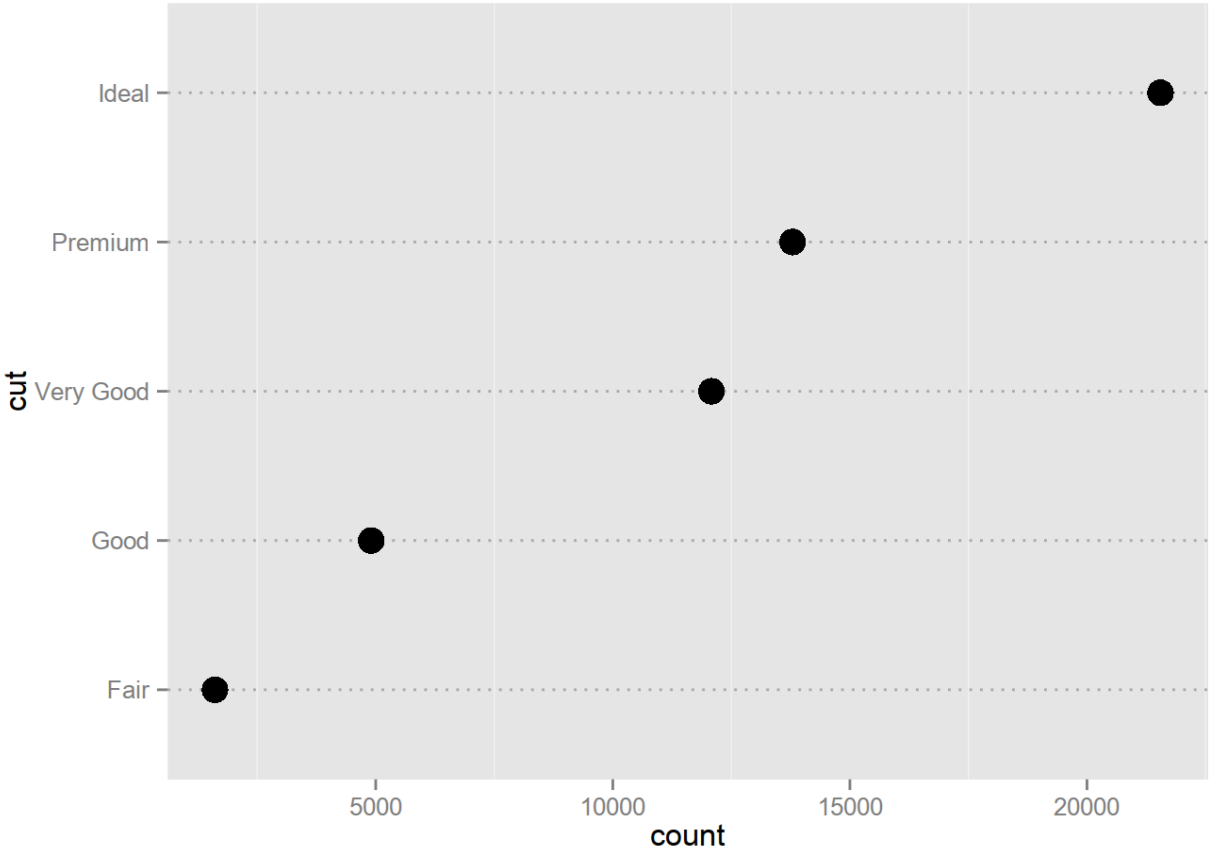
- Allow better comparison.
- Adpated to counts and quantities.

Classical graph

Bar



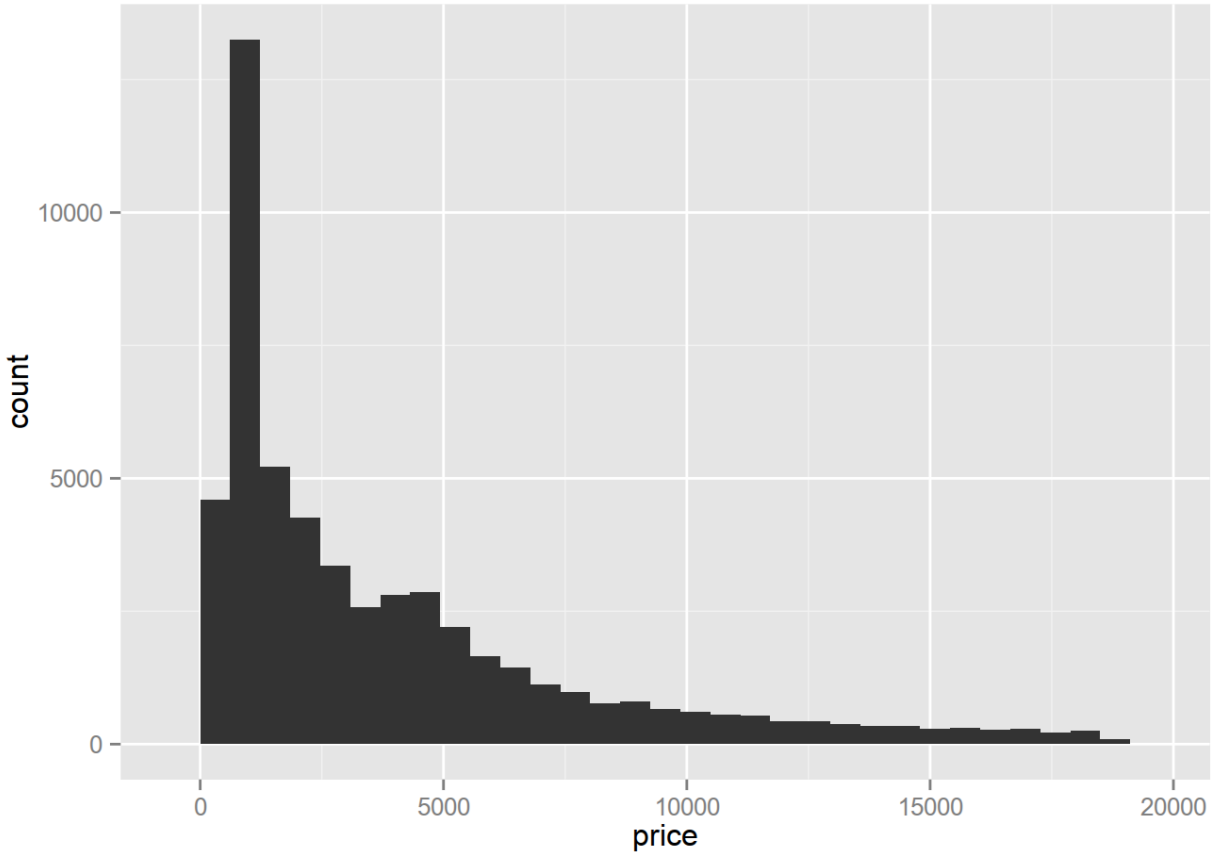
Classical graph
Cleveland Dot plots



- Less *ink*, more pleasant...

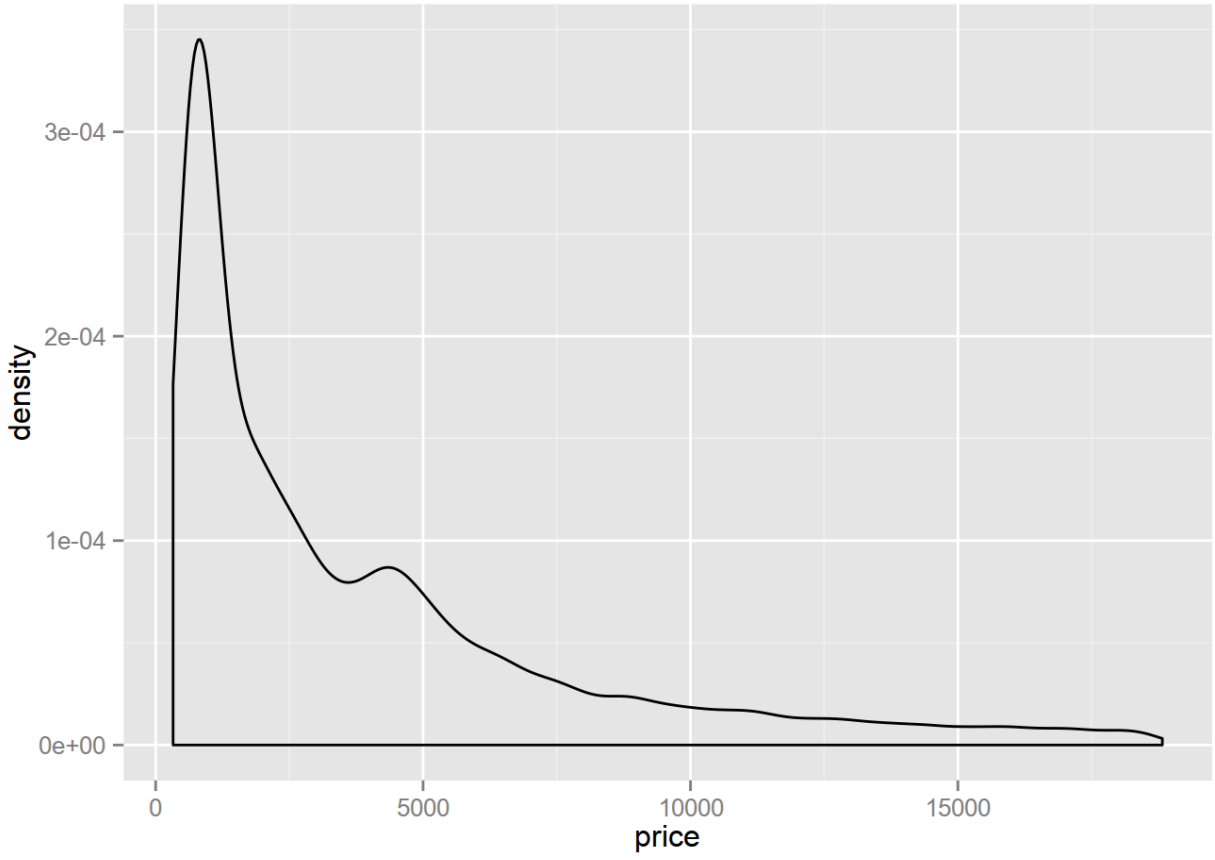
Classical graph

Histogram and density



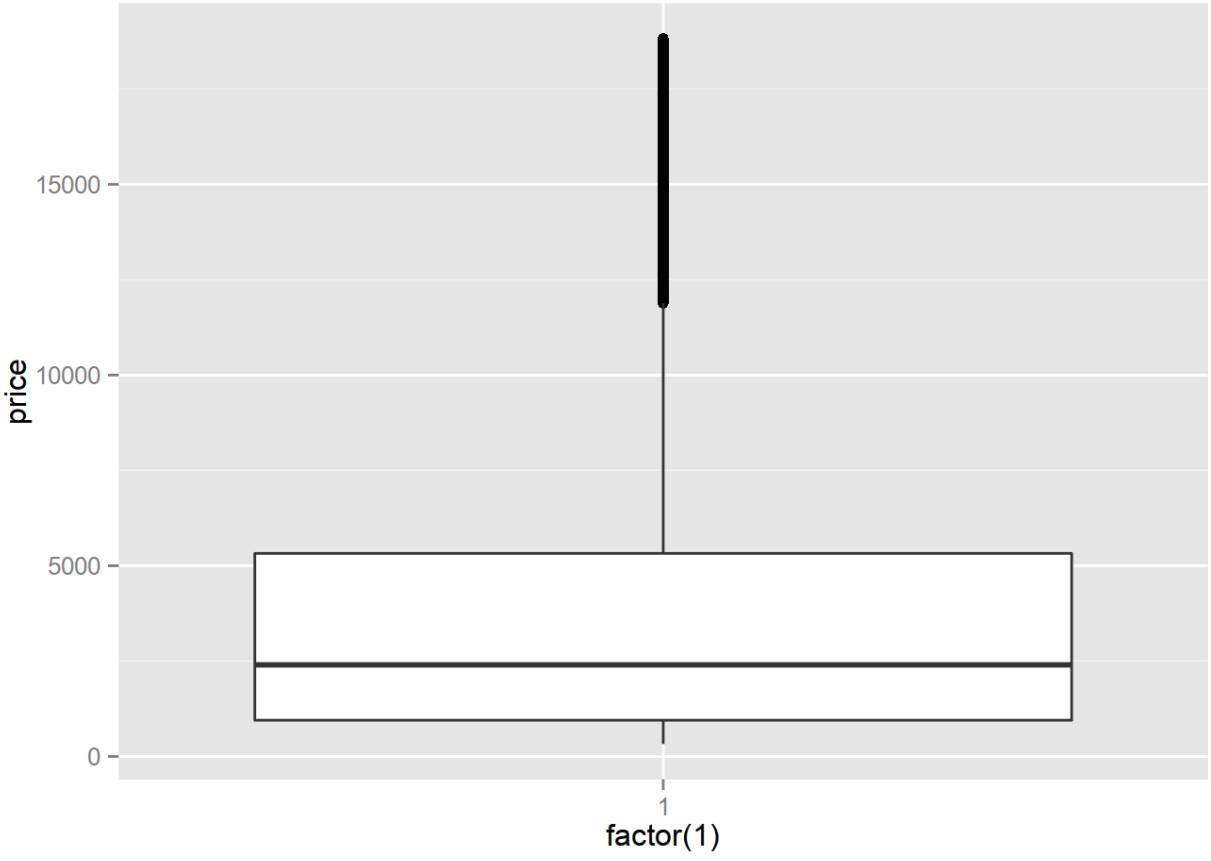
- Easily interpretable
- Adapted to continuous variable.

Classical graph
Histogram and density



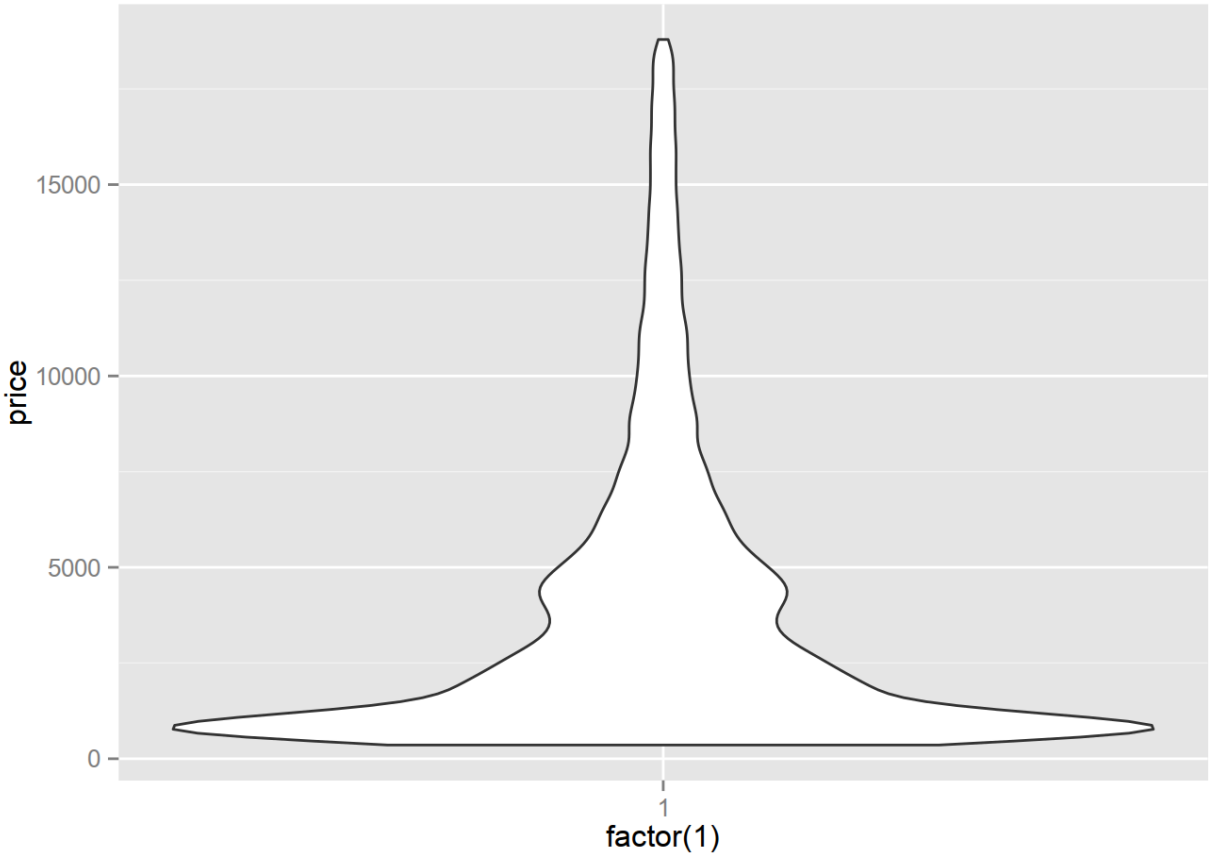
- Regularized view...

Classical graph
Box and Whiskers



- Most classical representation after pie...

Classical graph
Violin plot



- Combined box plot and density estimation.

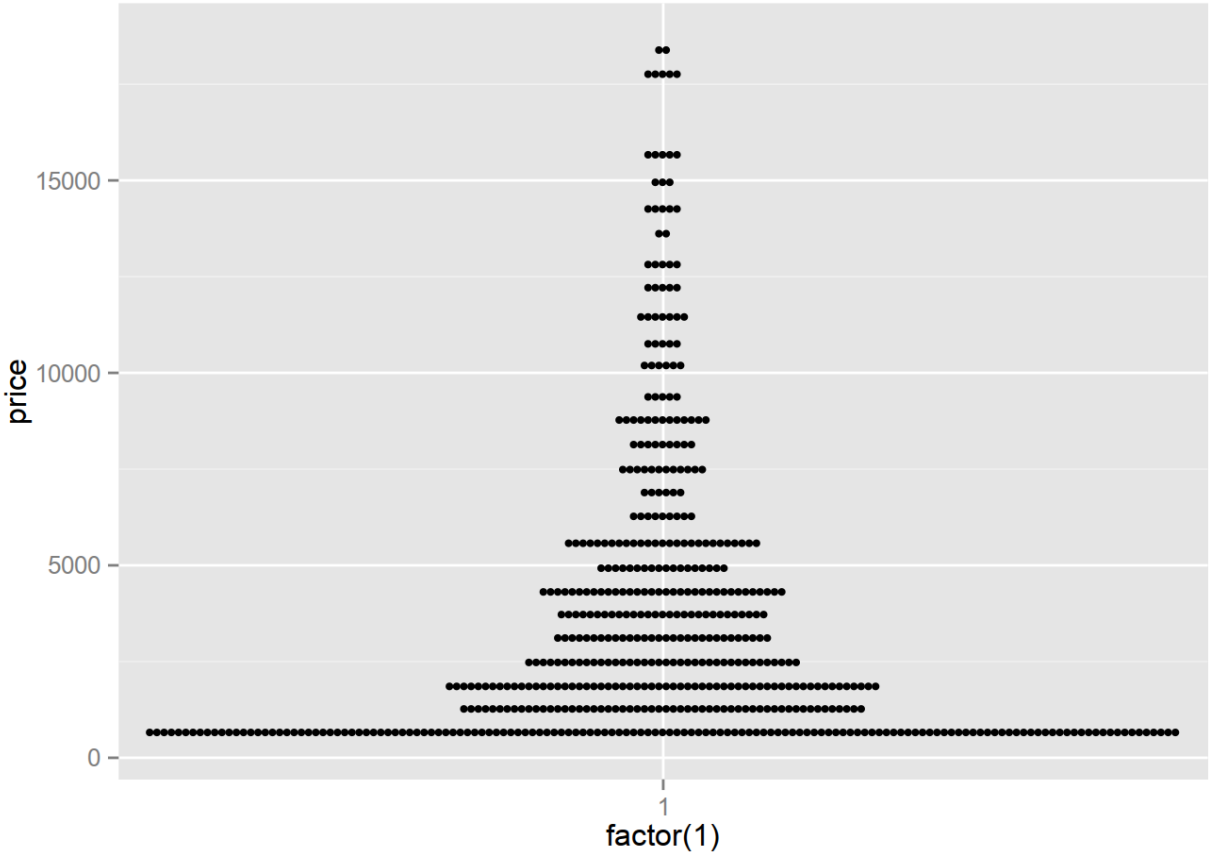
Classical graph

Stem and Leaf

- Textual analog of histogram...

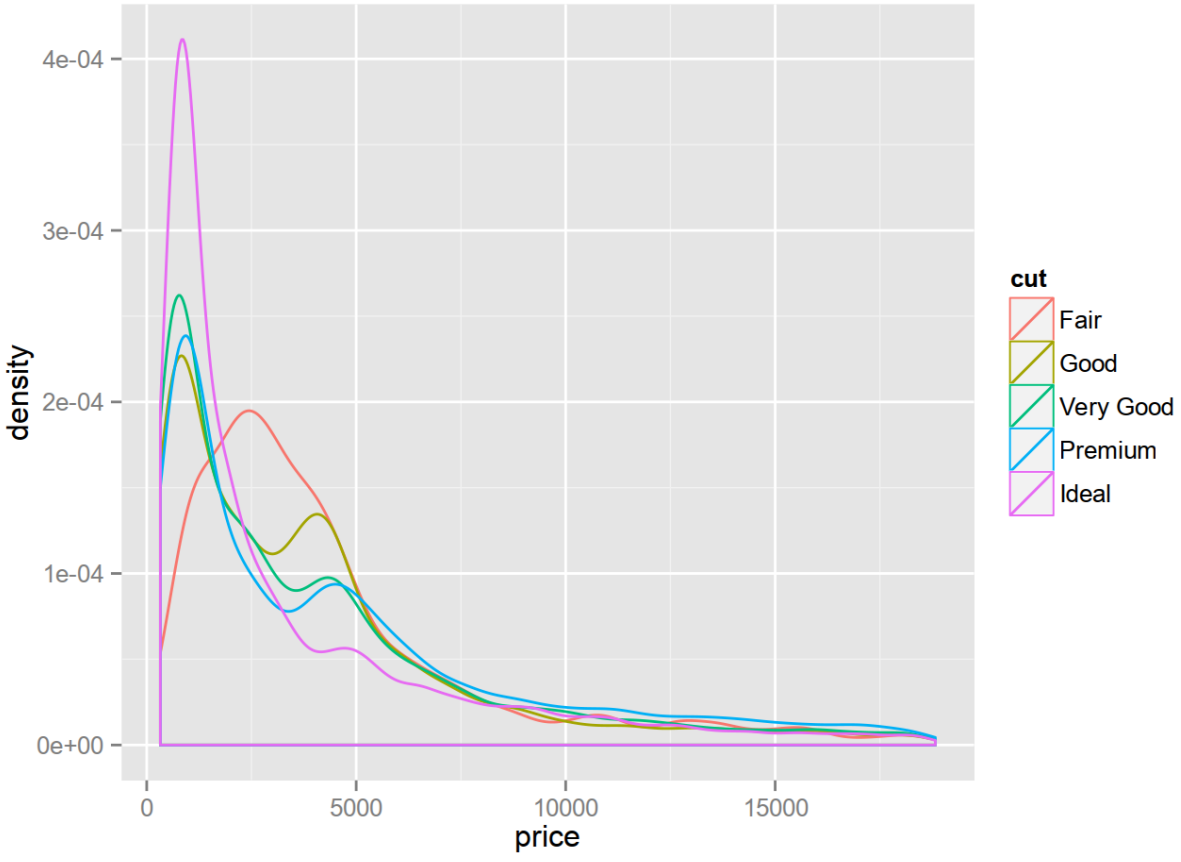
```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 44444444
##    0 | 55555555555555555566666666666666666666666666677777777777777777777777+54
##    1 | 00000000001111111111111112222222222222223333344444444
##    1 | 66666666666677777777778888888888999999999999
##    2 | 00000000000011122222333333444444444
##    2 | 55555555556666677788888999
##    3 | 00001111111222233333444
##    3 | 55555666777777888888999999
##    4 | 00011111222234444444444444444
##    4 | 5555666677777788999
##    5 | 00111123334
##    5 | 55555556666666777788889
##    6 | 0012344
##    6 | 556788
##    7 | 0224
##    7 | 555566677899
##    8 | 0233444
##    8 | 55666666799
##    9 | 01133
##    9 | 679
##   10 | 0012
##   10 | 566679
##   11 | 23
##   11 | 566679
##   12 | 024
##   12 | 5666
##   13 | 113
```

Classical graph
Dot plots



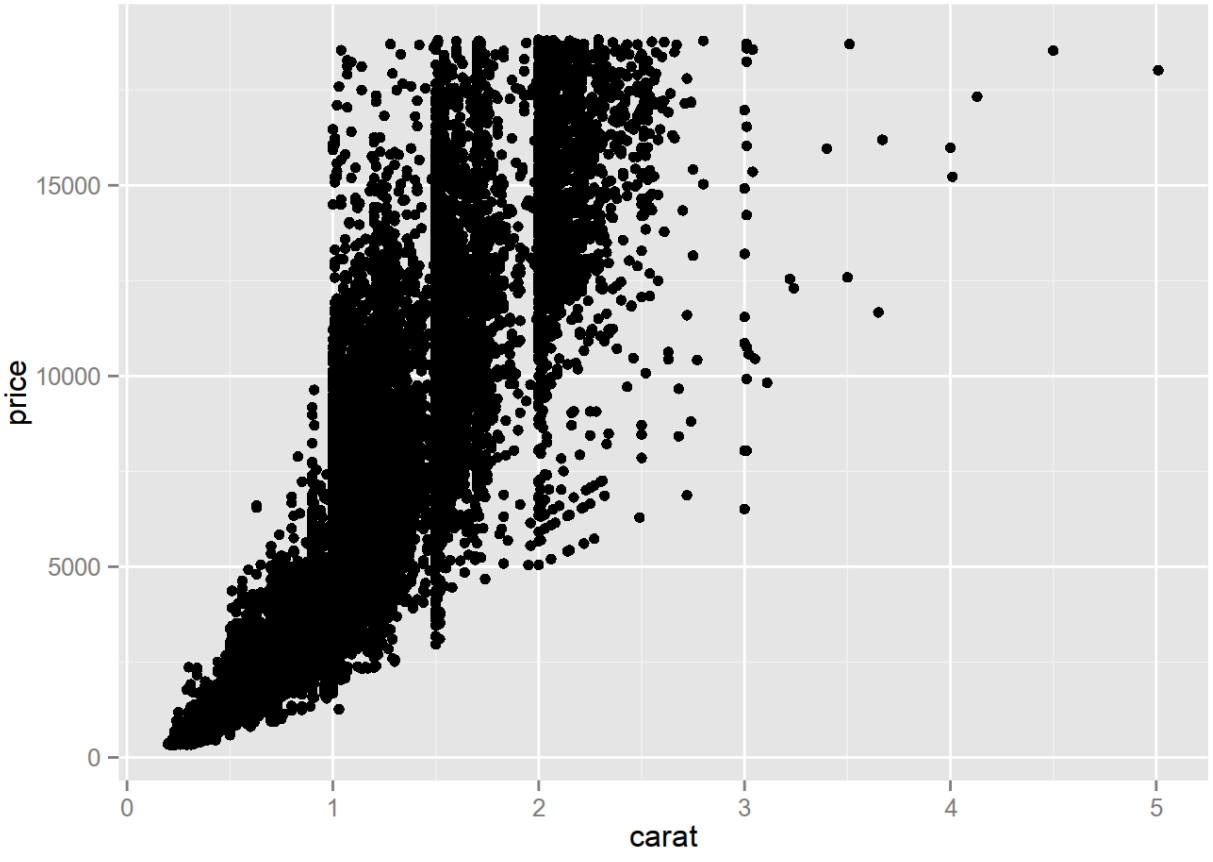
- Combined binning and individuals...

Classical graph
Grouping



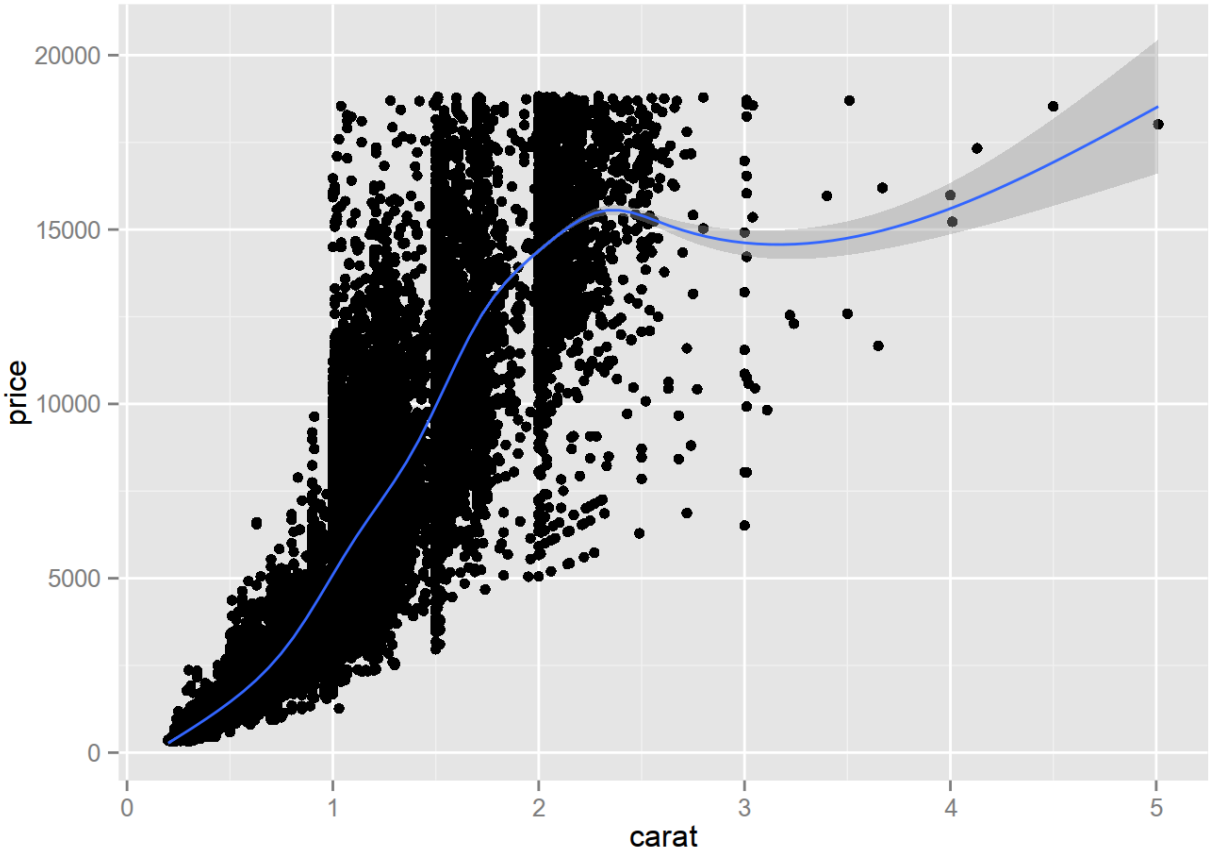
- Key to construct complex representation.

Classical graph
Scatter Plot



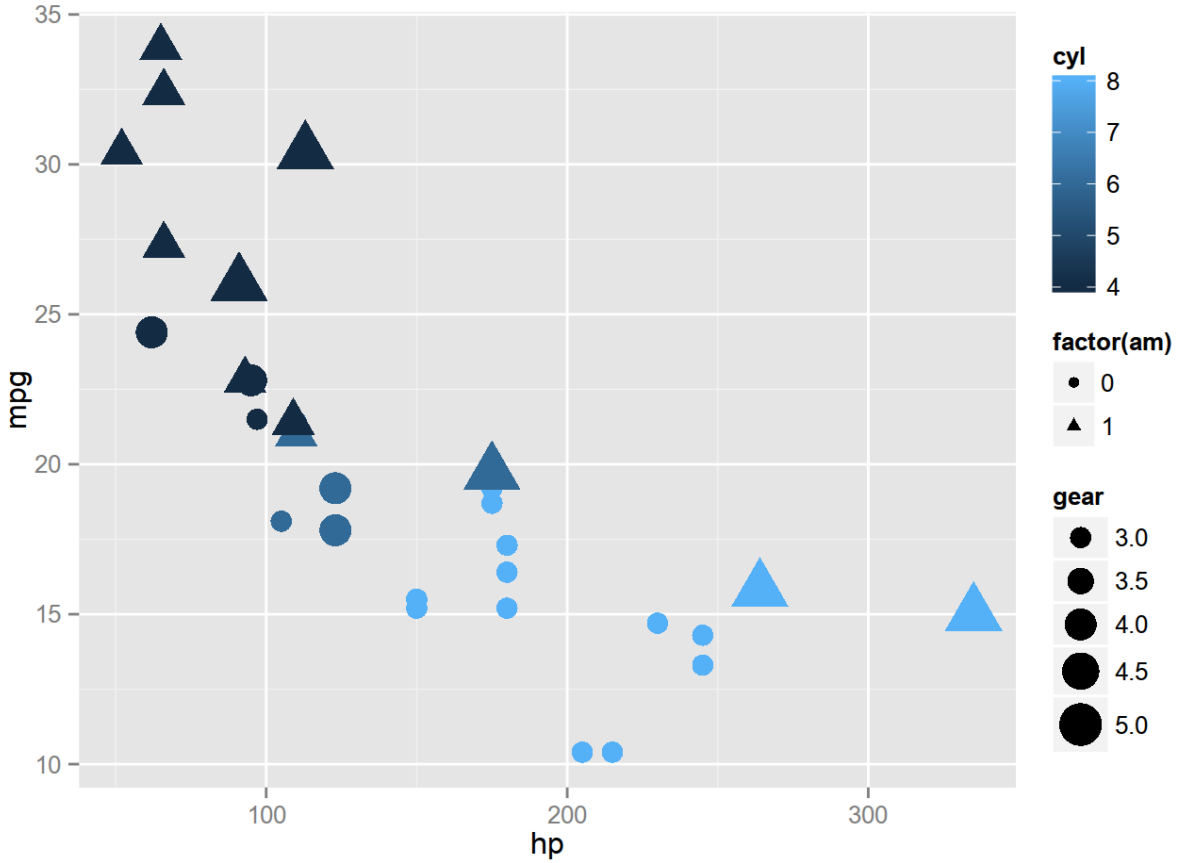
- Very useful to visualize the dependencies.

Classical graph
Smoothing



- Strong visual help.

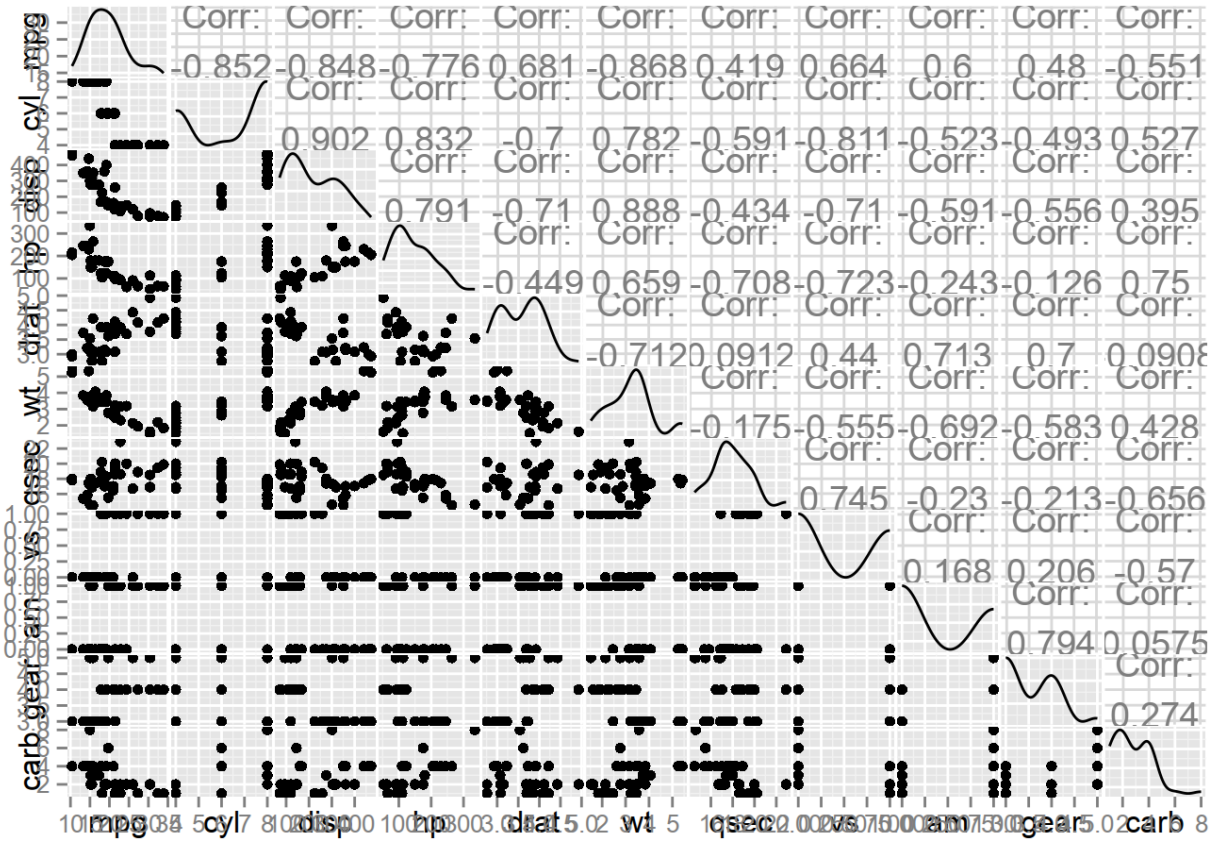
Classical graph
Decoration



- Good idea to augment the information density...
- but can lead to too much complexity.

Classical graph

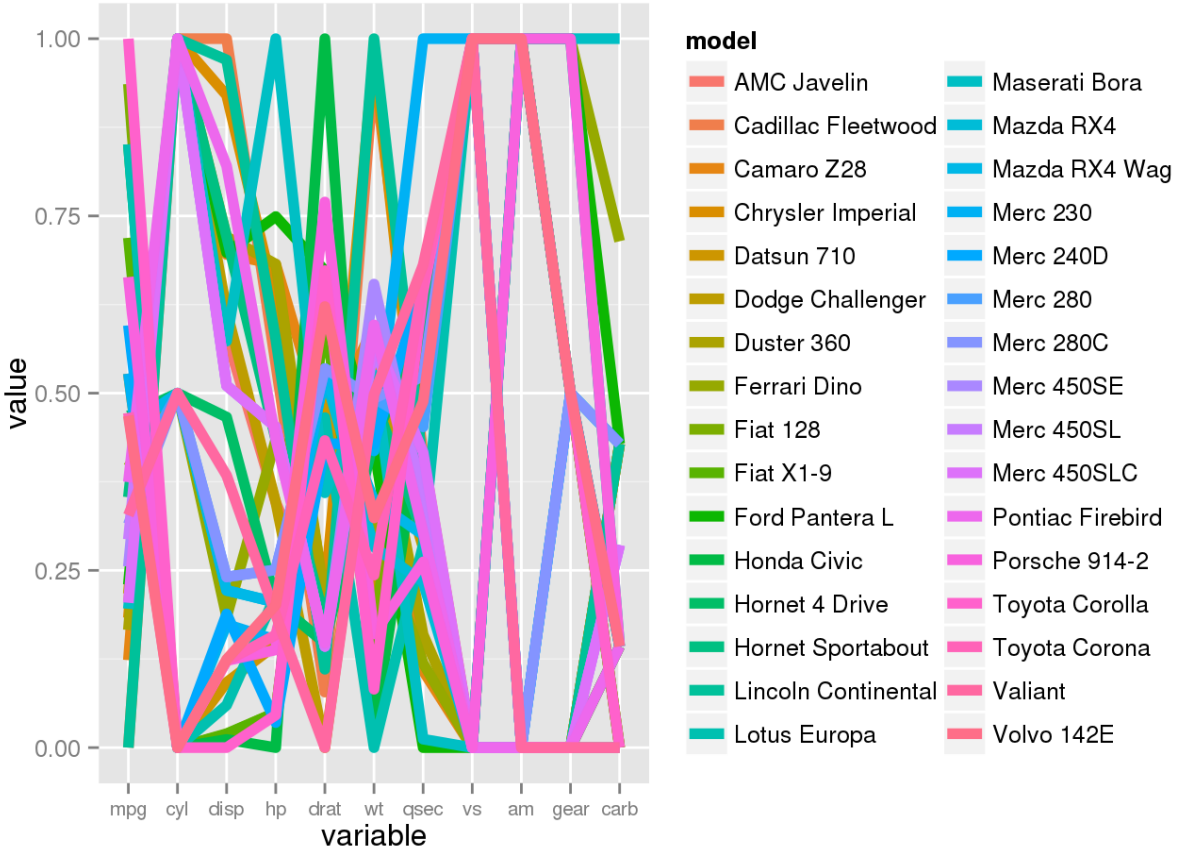
Scatter Plot Matrix



- Gather all the dependecies...

Classical graph

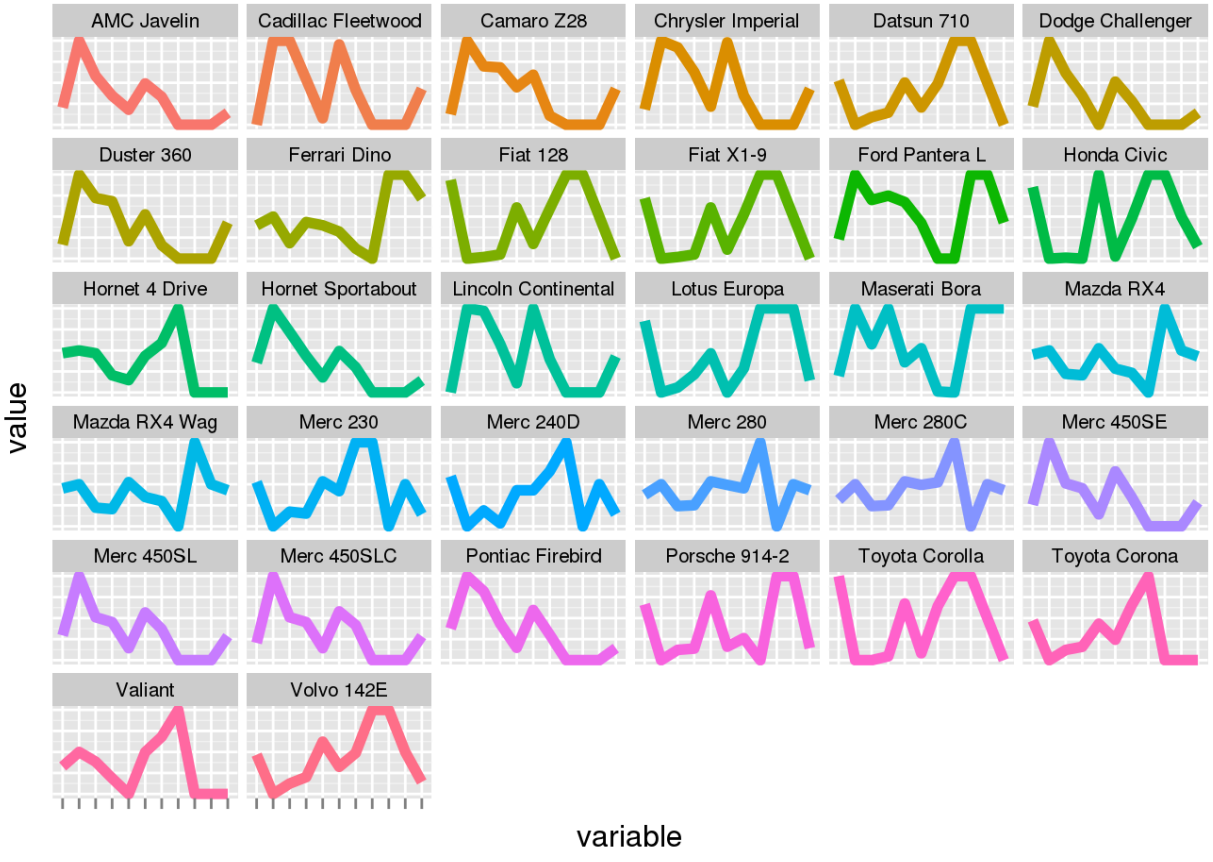
Parallel Coordinates and Radar Plot



- Clever ideas to visualize groups.

Classical graph

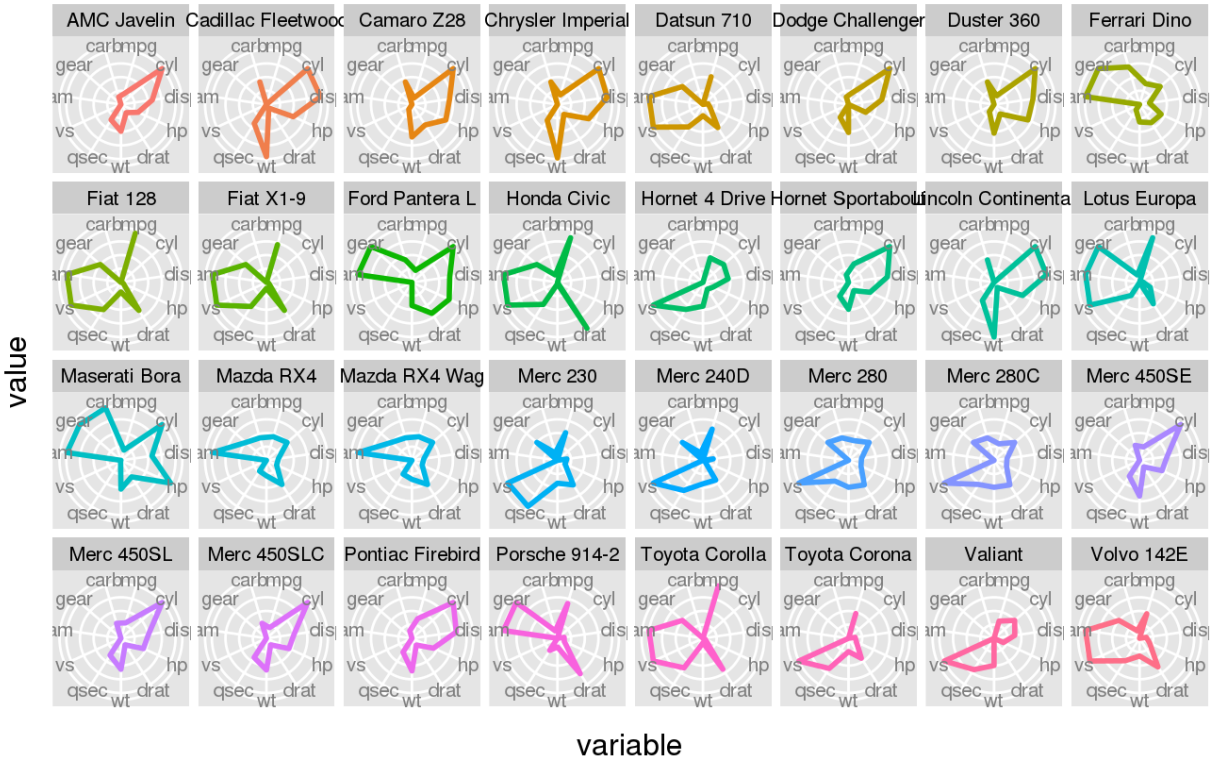
Parallel Coordinates and Radar Plot



- Clever ideas to visualize groups.

Classical graph

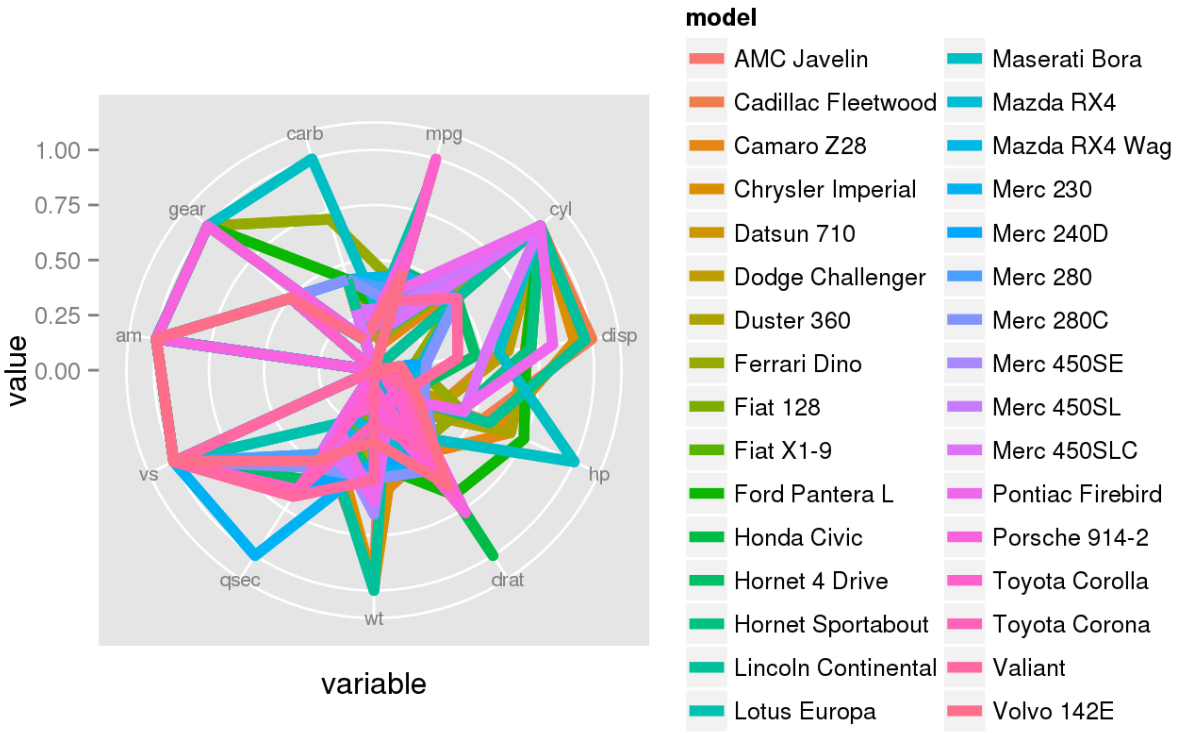
Parallel Coordinates and Radar Plot



- Clever ideas to visualize groups.

Classical graph

Parallel Coordinates and Radar Plot



- Clever ideas to visualize groups.

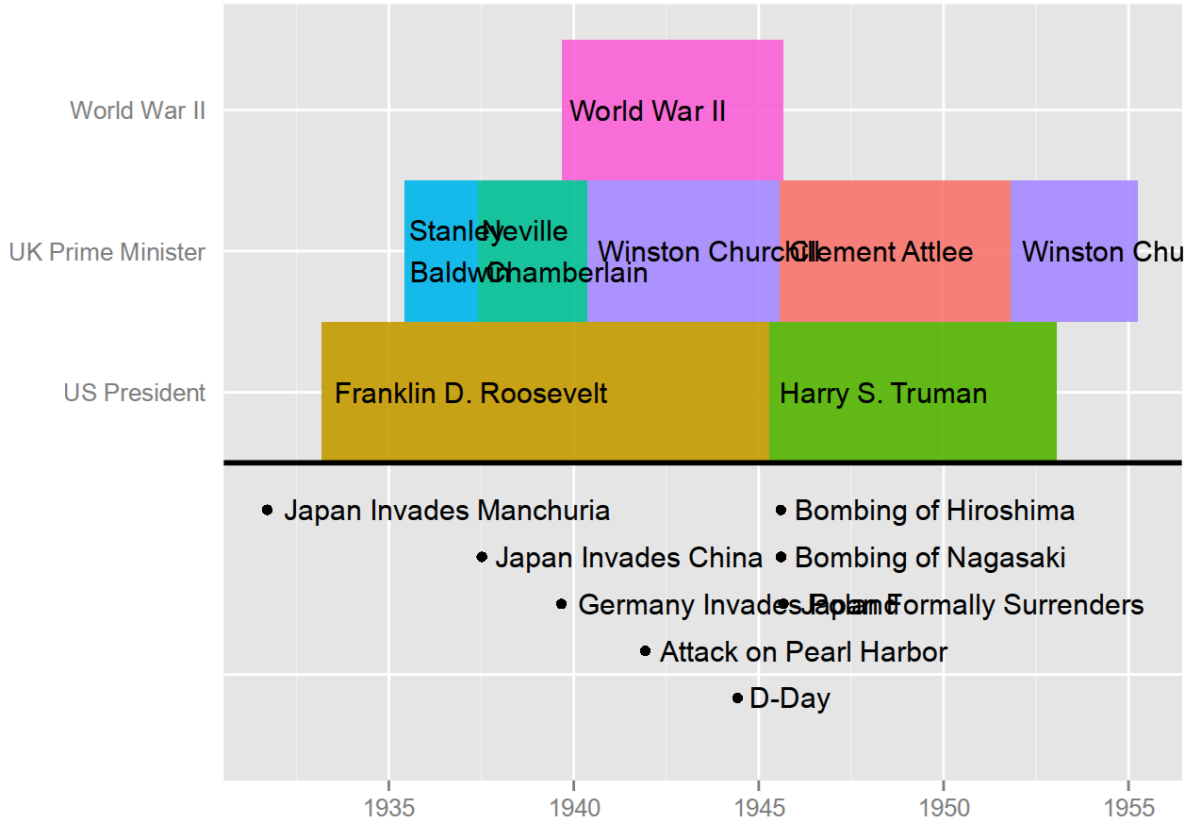
Classical graph

Time series

- Order makes lines pertinent...

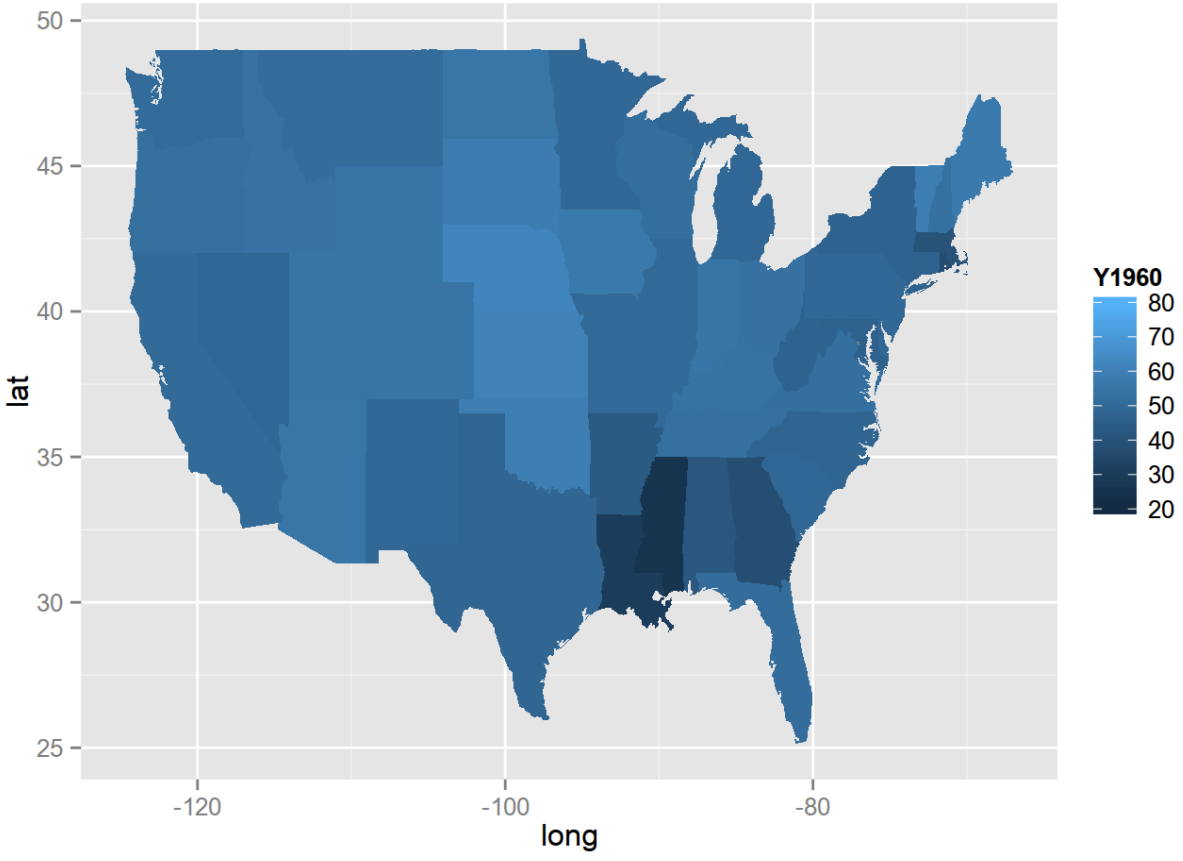
Classical graph

Timeline



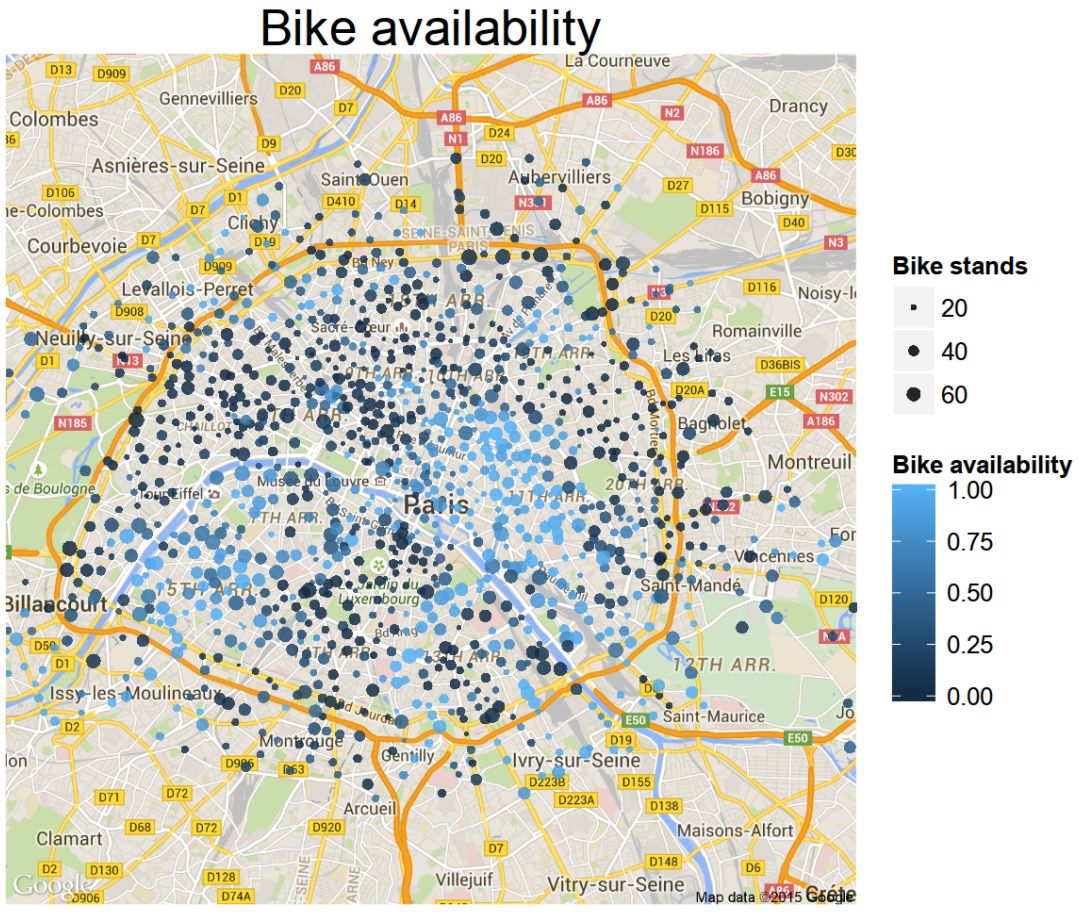
- Is this really a plot?

Classical graph
Choroplets



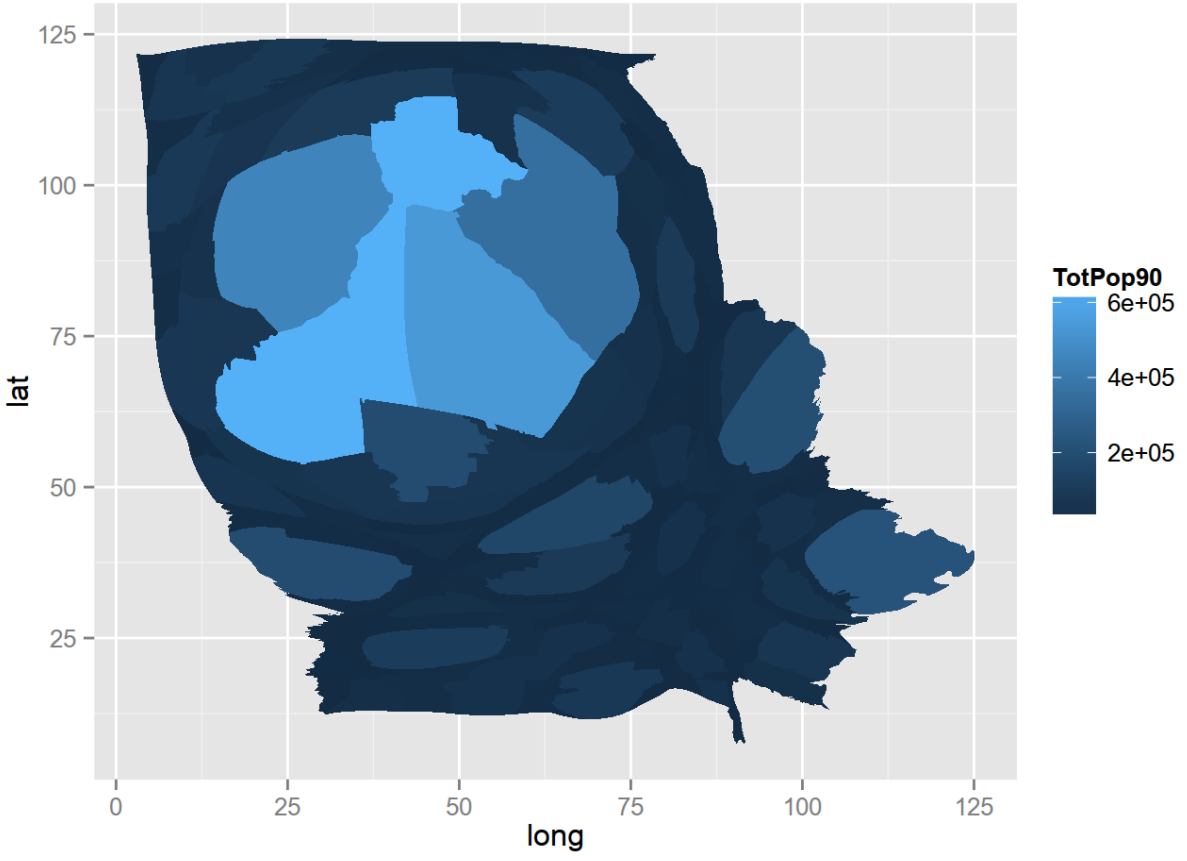
- Strong visual impact!

Classical graph
Symbols Maps



- Same ideas than decoration
- Could be extended to quite complex decorations...

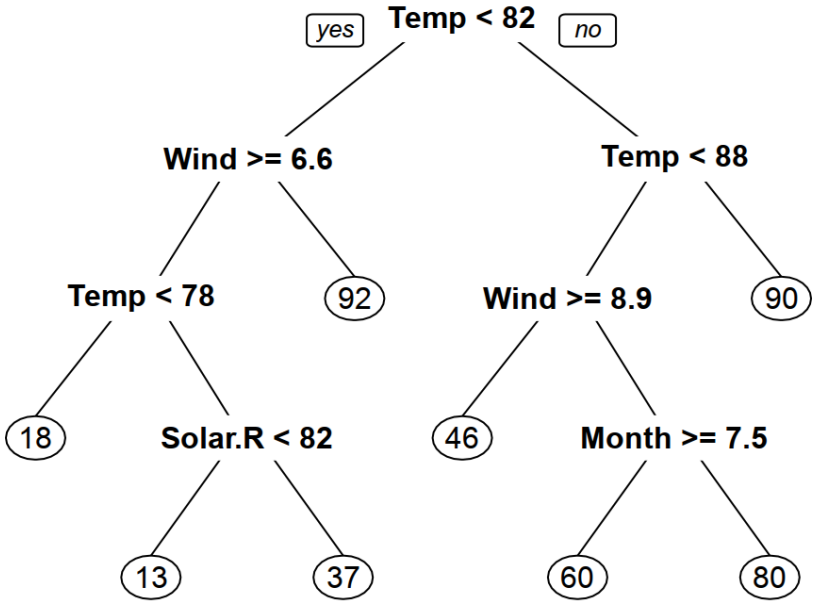
Classical graph
Cartograms



- Mainly useful when the reference is known.

Classical graph

Trees



- Often use in classification...

Classical graph

Tree Graph

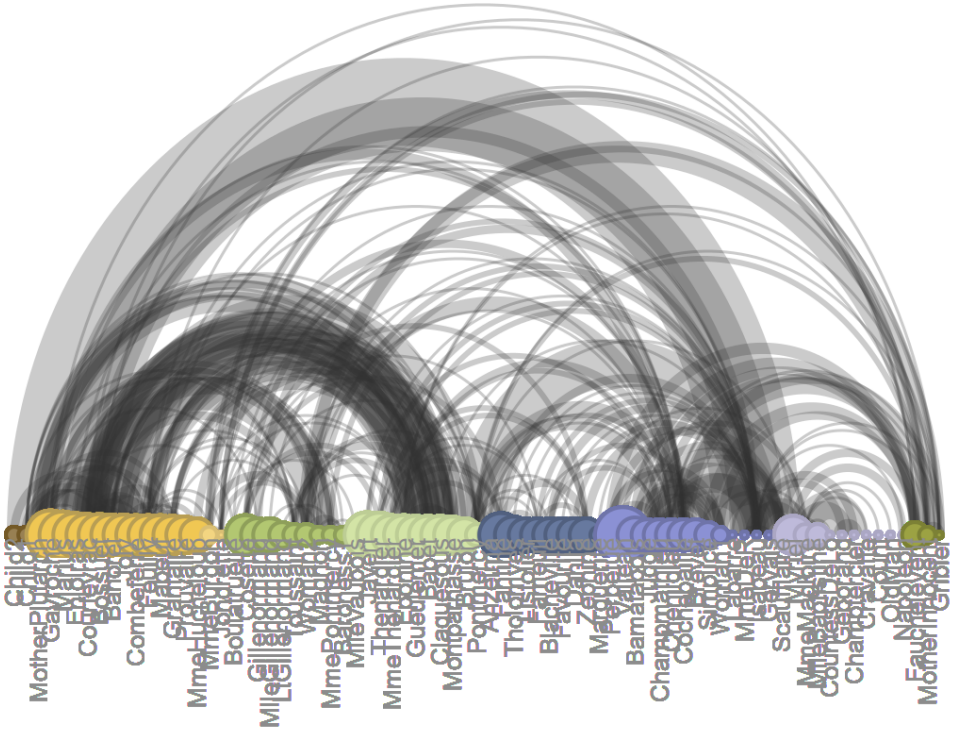
- Polar variant.

Classical graph

Planar Layout

- Many possible layouts.

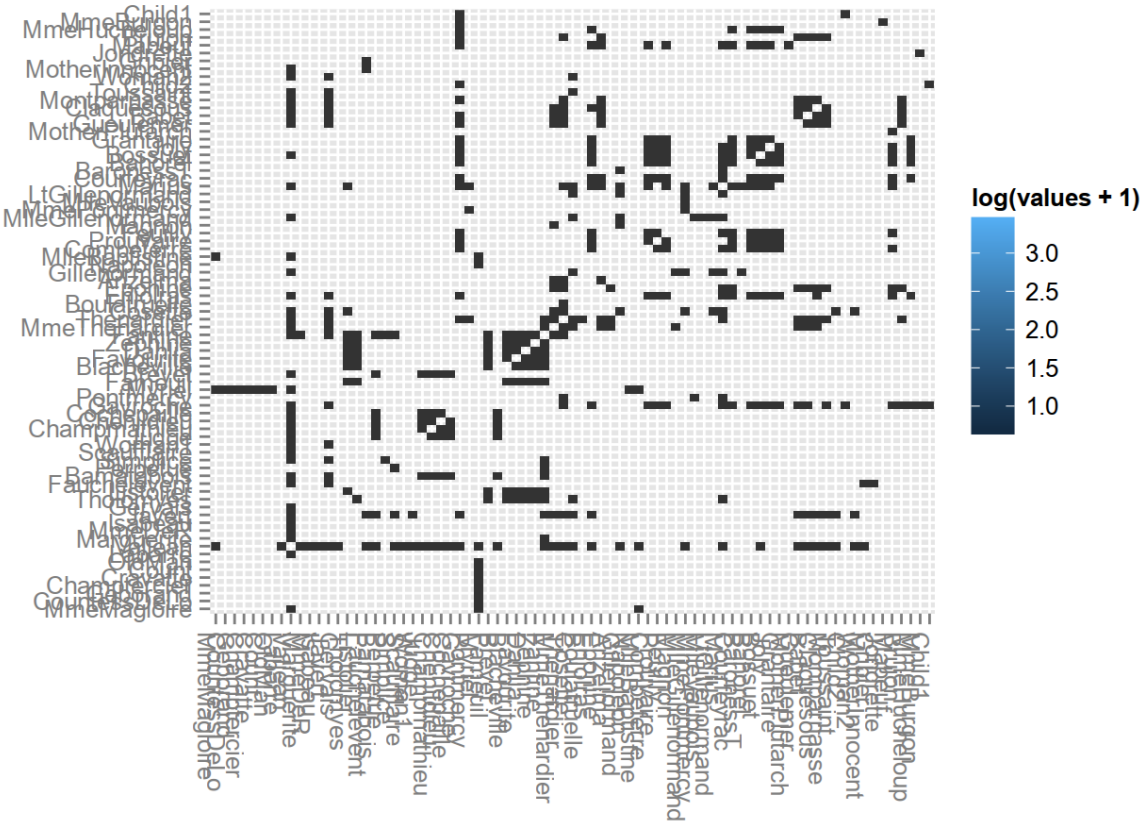
Classical graph
Arc Diagram



- Very different layout...

Classical graph

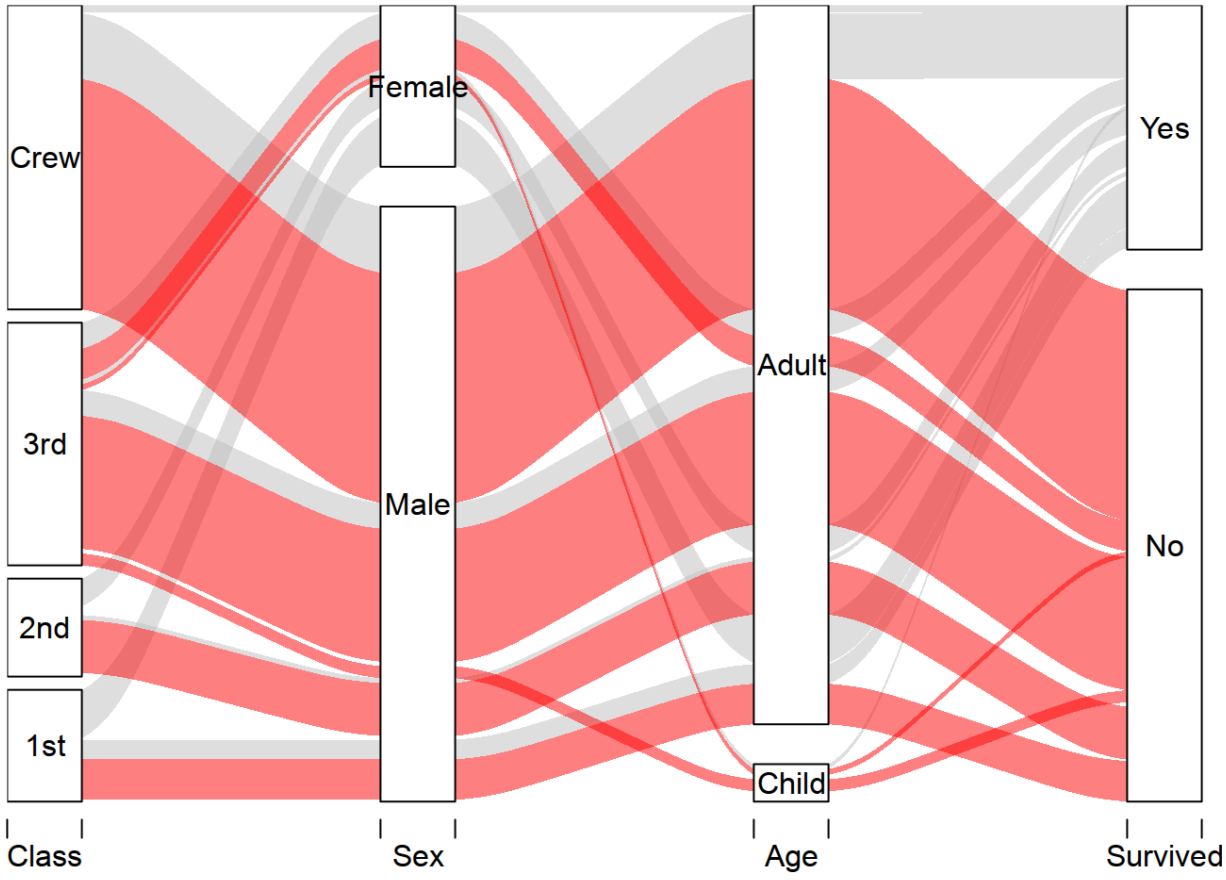
Matrix View



- Adjacency matrix visualization.

Classical graph

Flow

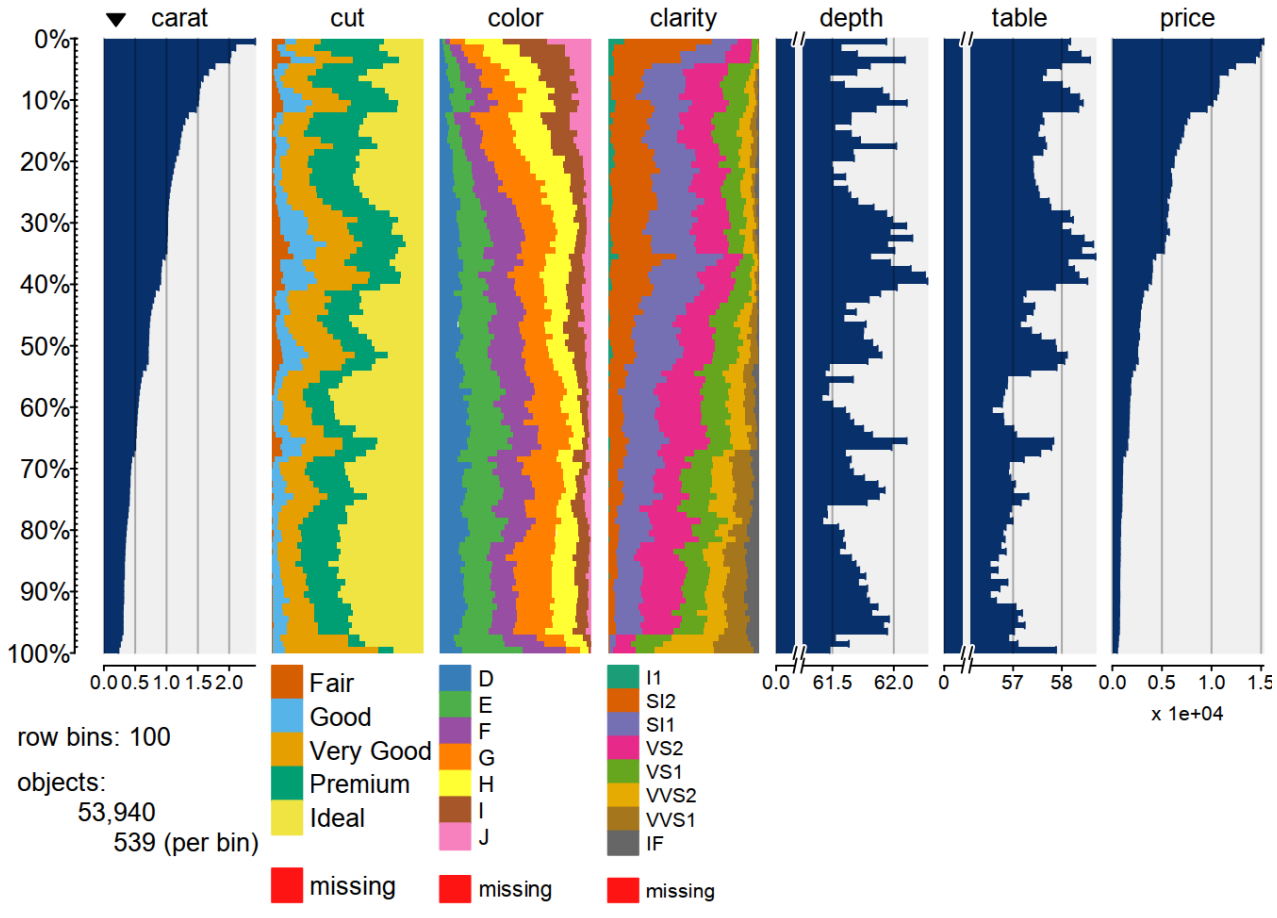


- Vertice oriented visualization.

Classical graph
3D

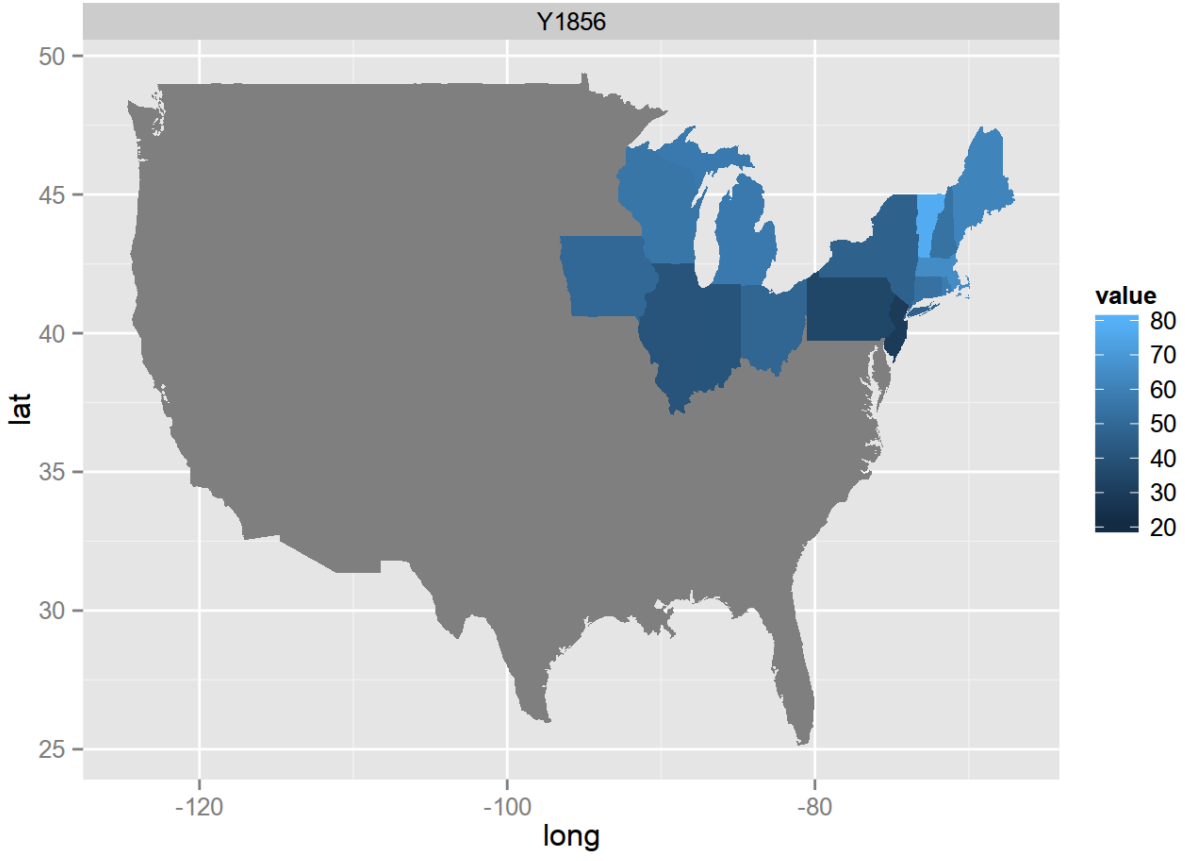
- Interaction is required

Classical graph
Linked Data Panels



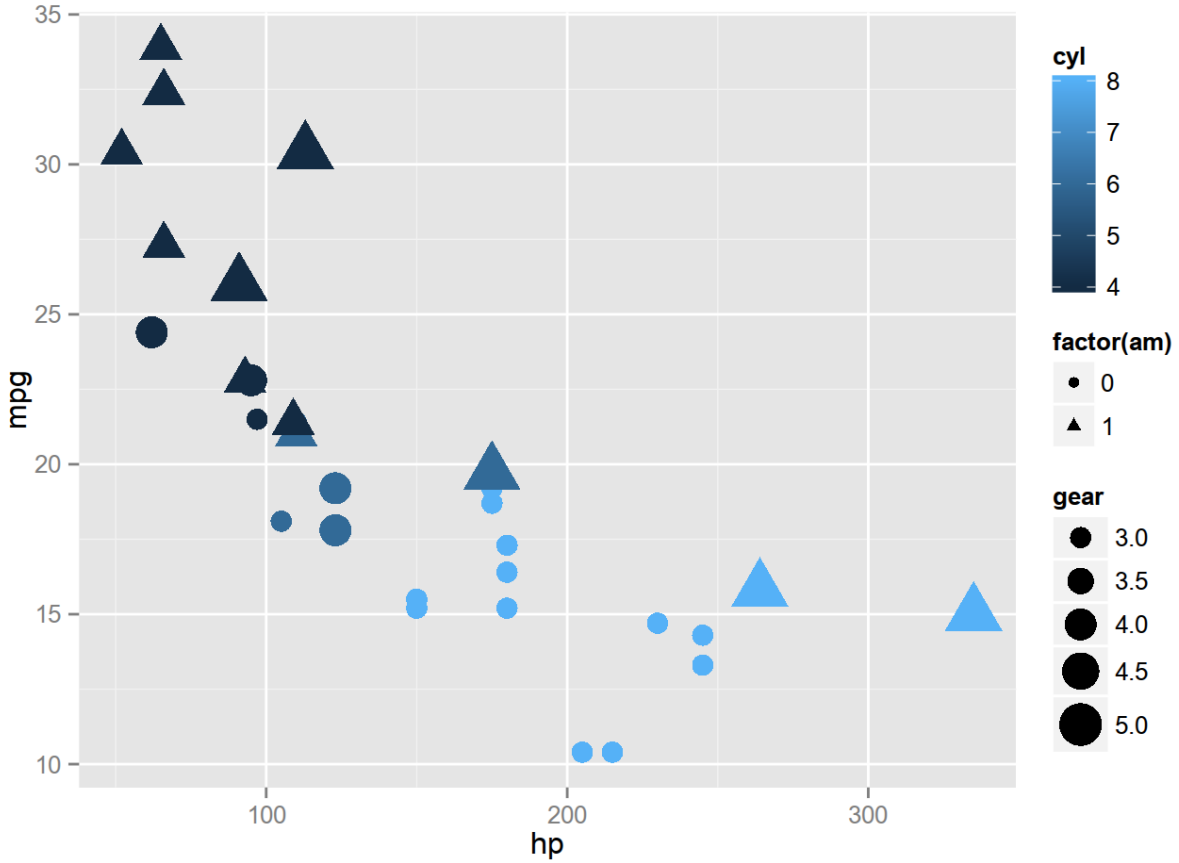
- Several linked views.

Classical graph
Animation



- Adapted to 1D axis...

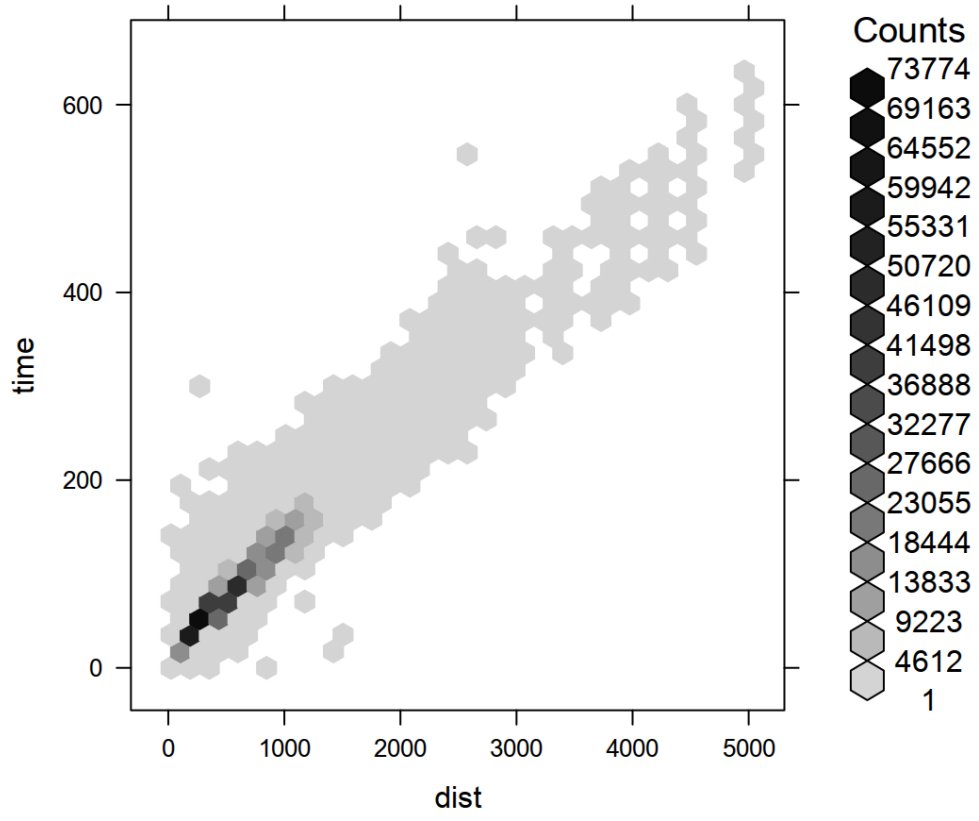
Classical graph
Interactivity



- Javascript based frameworks...
- Shiny

- More data point than pixels!
- Even if the processing possible, it is almost impossible to visualize faithfully the data!
- Summarization required...
- Grouping by categories or binning

Classical graph
Binning



- Binning with ggvis...