Data Visualisation avec R

E. Matzner-Løber

ce cours est basé sur des supports de R. Womack,

de nombreuses (très) discussions avec E. Le Pennec, P-A. Cornillon, B. Thieurmel, J. Petit...

Aussois 2015

Outline

Objectifs de cours

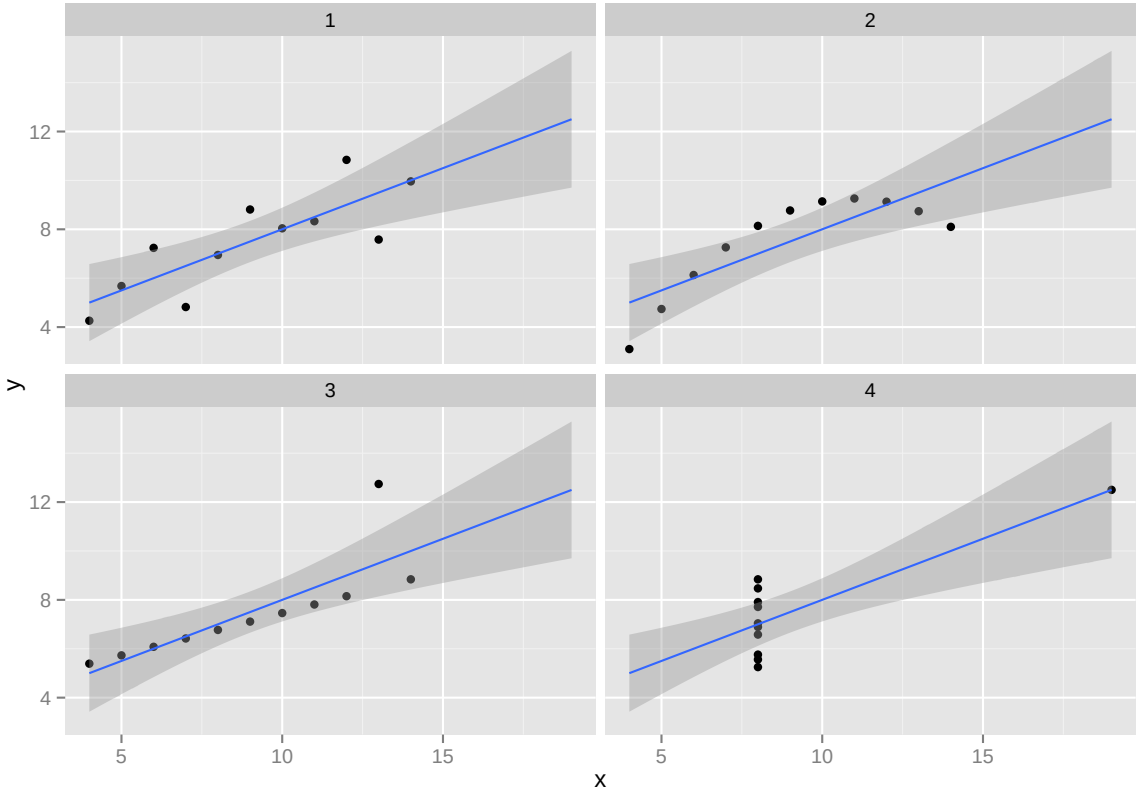
- Représentations standards,
- Principe de “bonnes” représentation,
- Exemples d’implémentations avec R

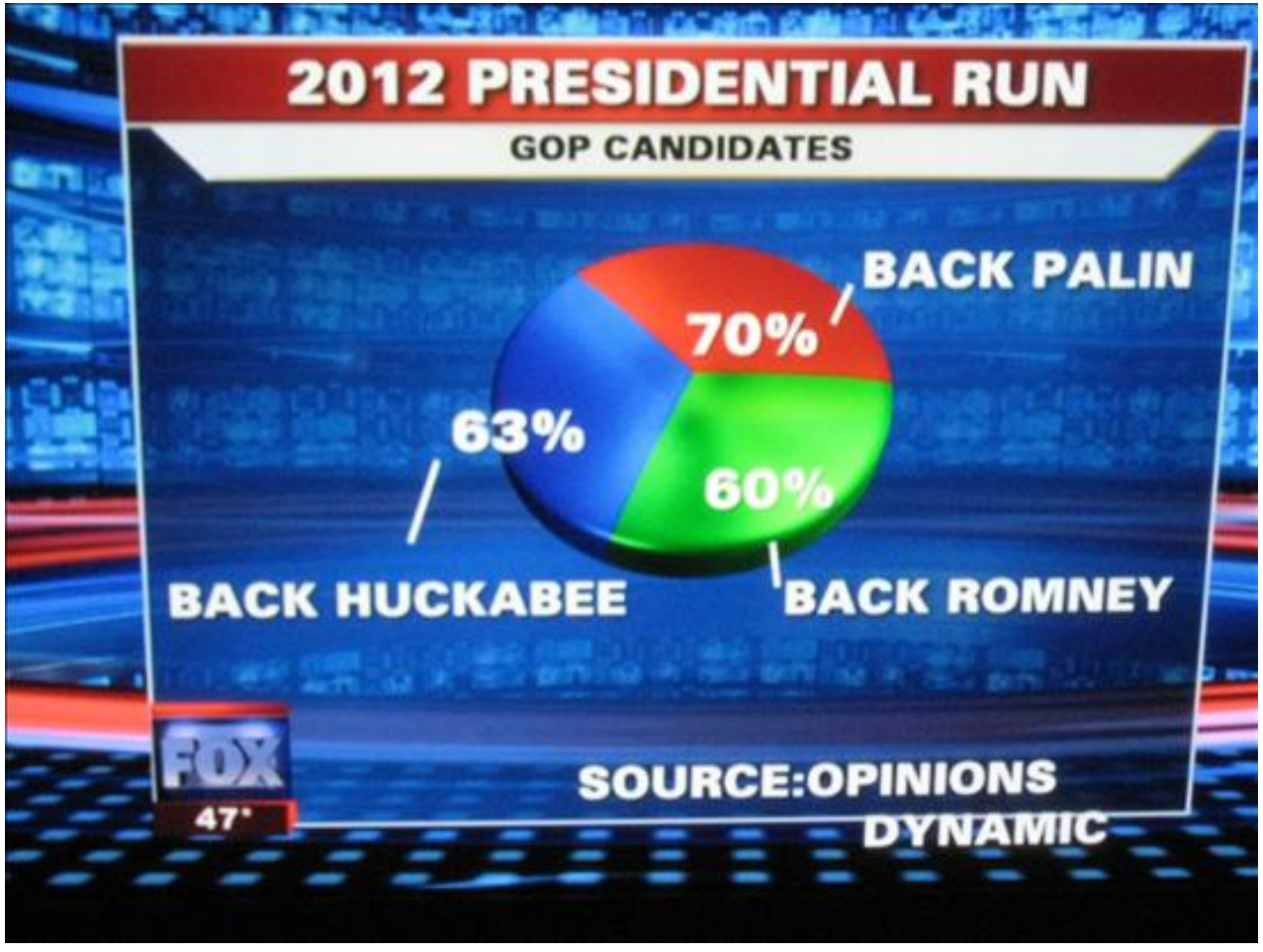
Ce n’est pas

- *Infographics*
- aspect cognitif de la perception...

- La dataVis peut :
  - proposer une réelle compréhension des “pattern” des données
  - détecter des structures cachées dans les données
  - un résumé simple

- La dataVis peut :
  - proposer une réelle compréhension des “pattern” des données
  - détecter des structures cachées dans les données
  - un résumé simple
- Anscombe’s quartet example:

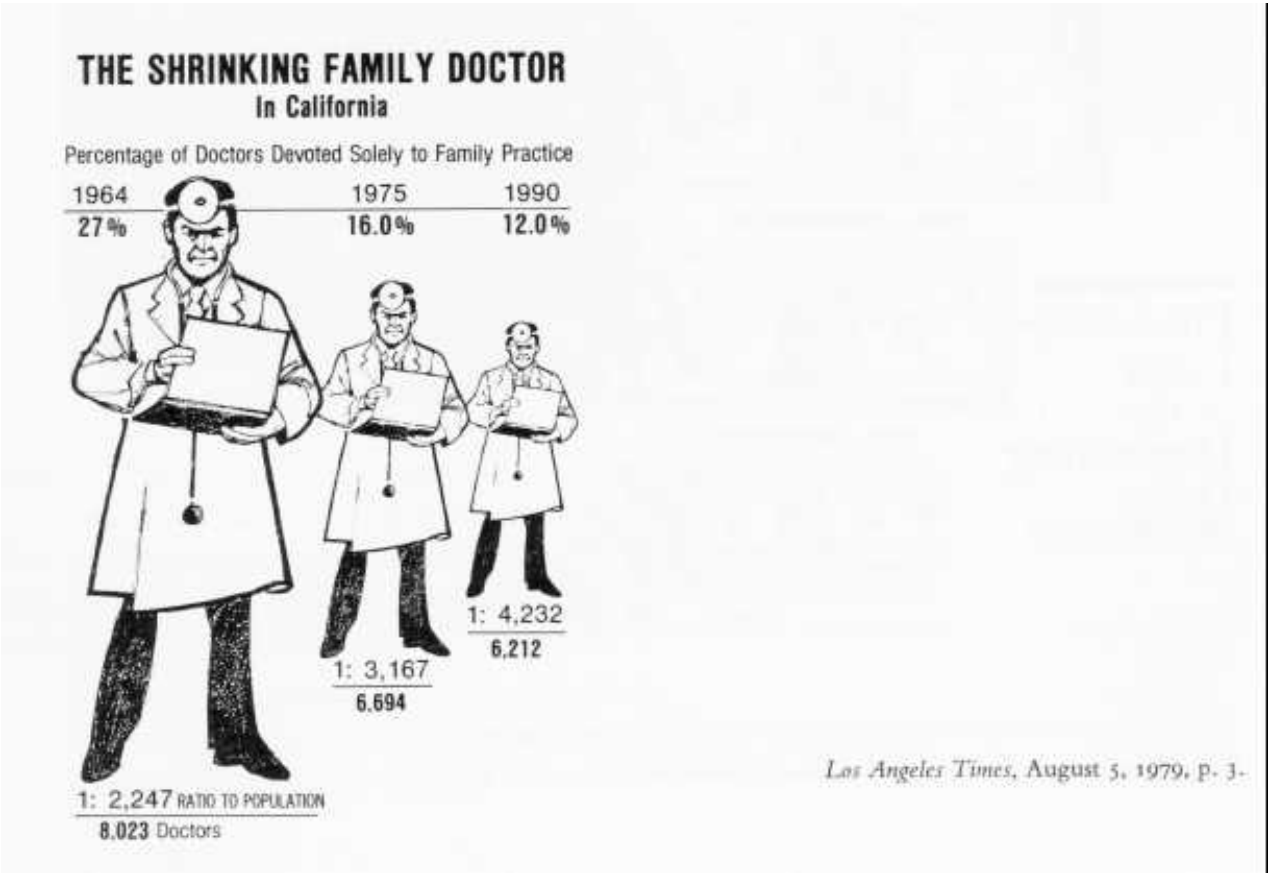




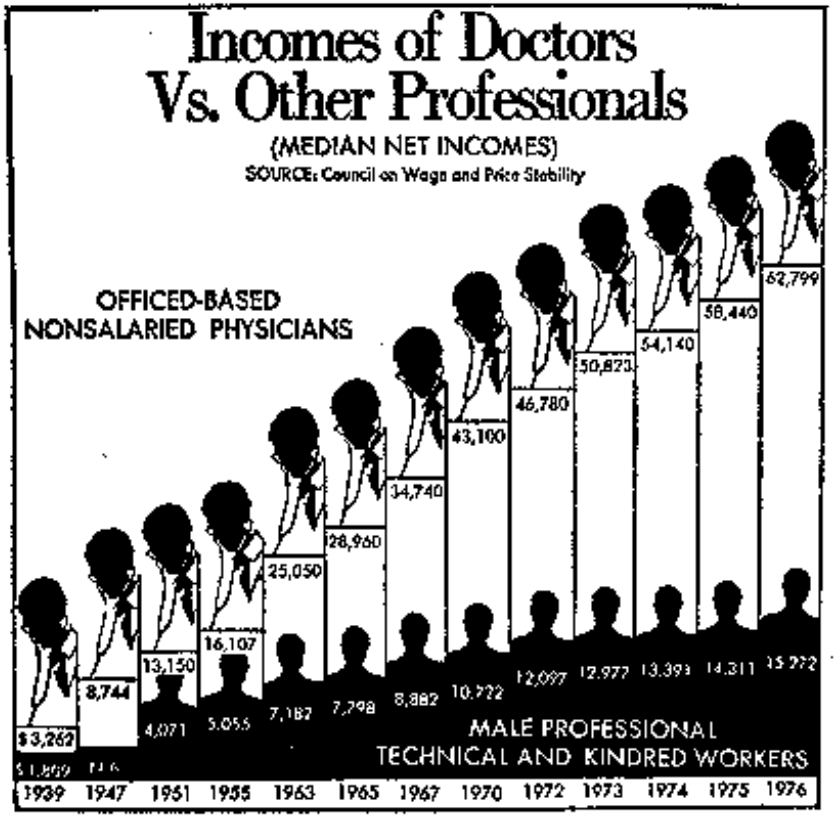
- Sans commentaire !!!

Introduction
Mauvaise DataVis

- Pas informatif

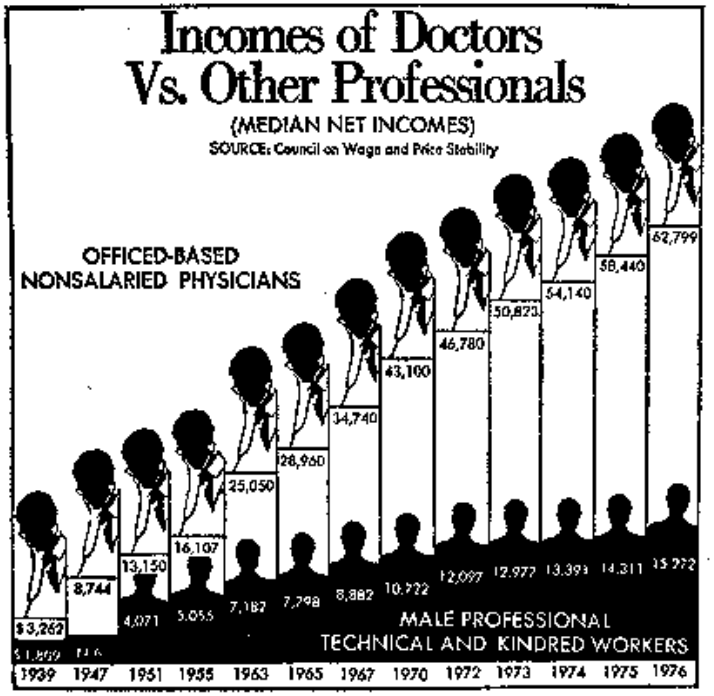


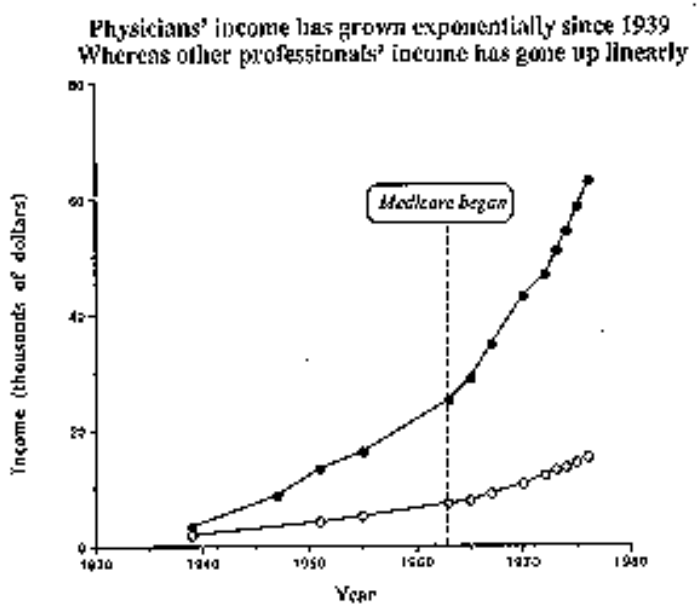
- facteur d'échelle faux !



- problème de facteur d'échelle

Introduction
Mauvaise DataVis





Introduction
Mauvaise DataVis

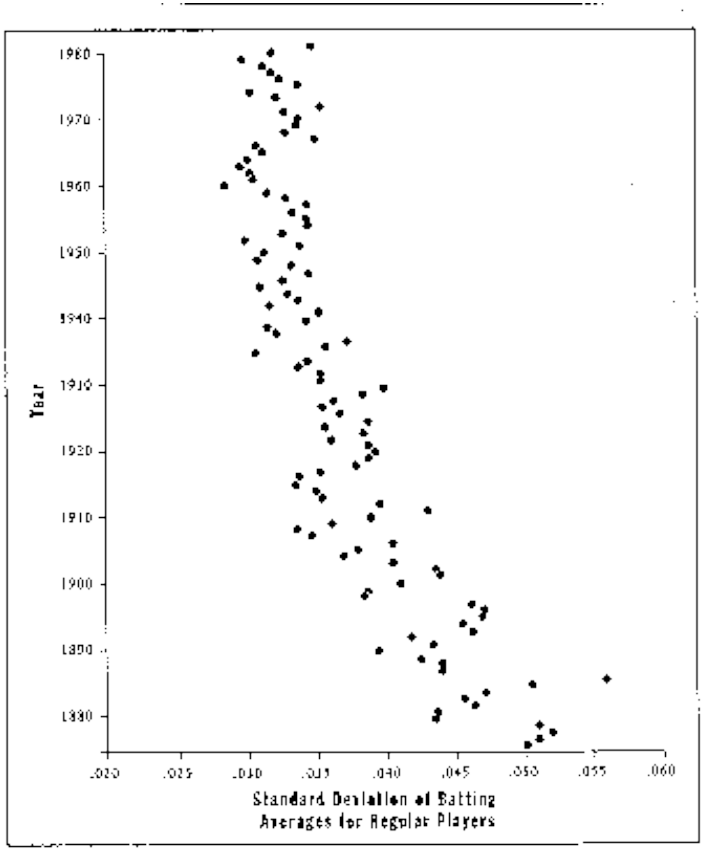


FIGURE 16
Standard deviation of batting averages for all full-time players by year for the first 100 years of professional baseball. Note the regular decline.

- inversion !

Introduction
Mauvaise DataVis

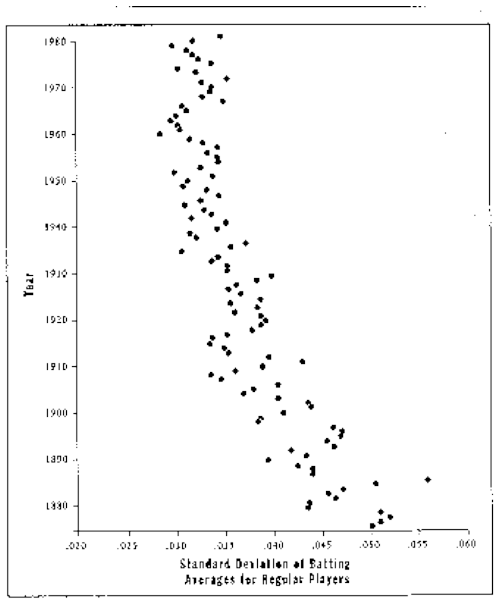
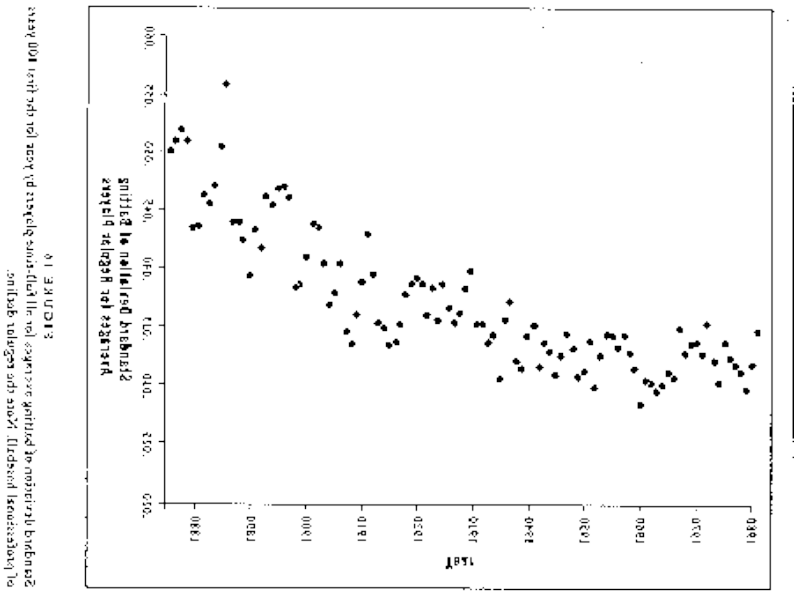
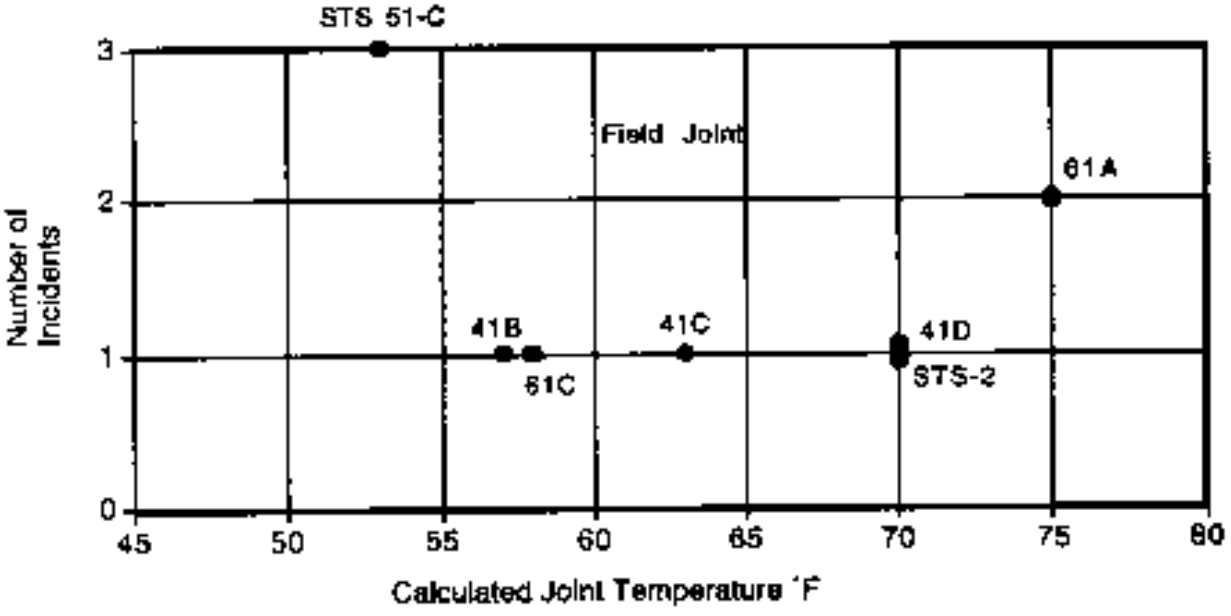


FIGURE 16
Standard deviation of batting averages for all full-time players by year for the first 100 years of professional baseball. Note the regular decline.



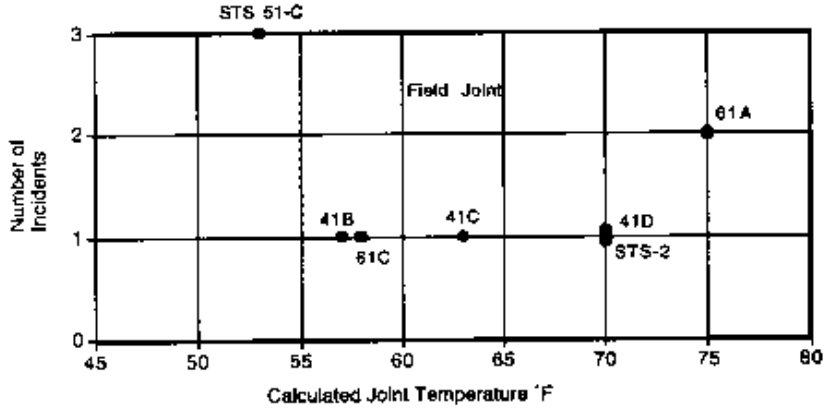
Incidents en fonction de la température

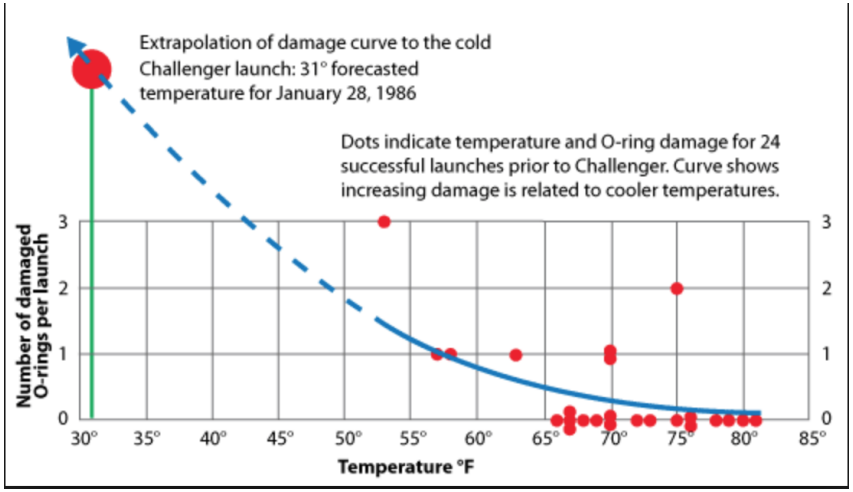


- Catastrophe, Challenger 1986 !

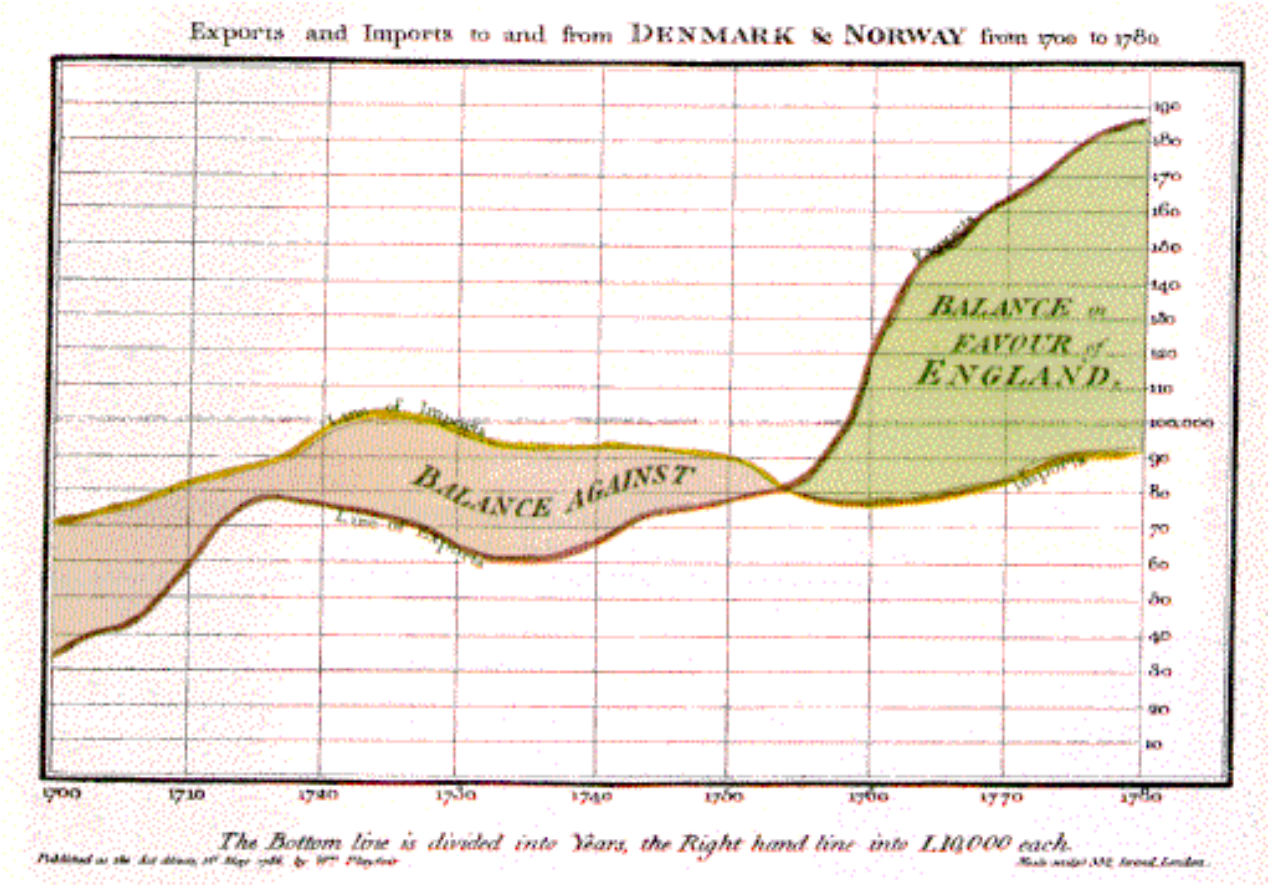
Introduction

Mauvaise DataVis



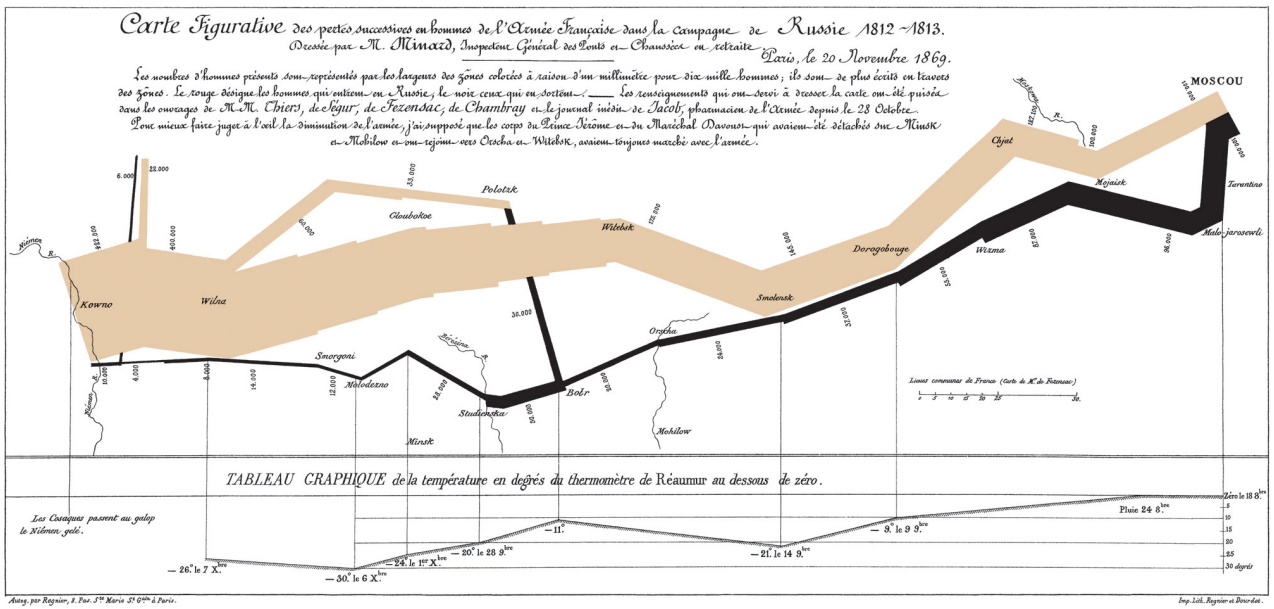


- William Playfair (1759-1823) considéré comme l'inventeur des formes communes de graphique pour représenter les données : line plots, bar chart and pie chart

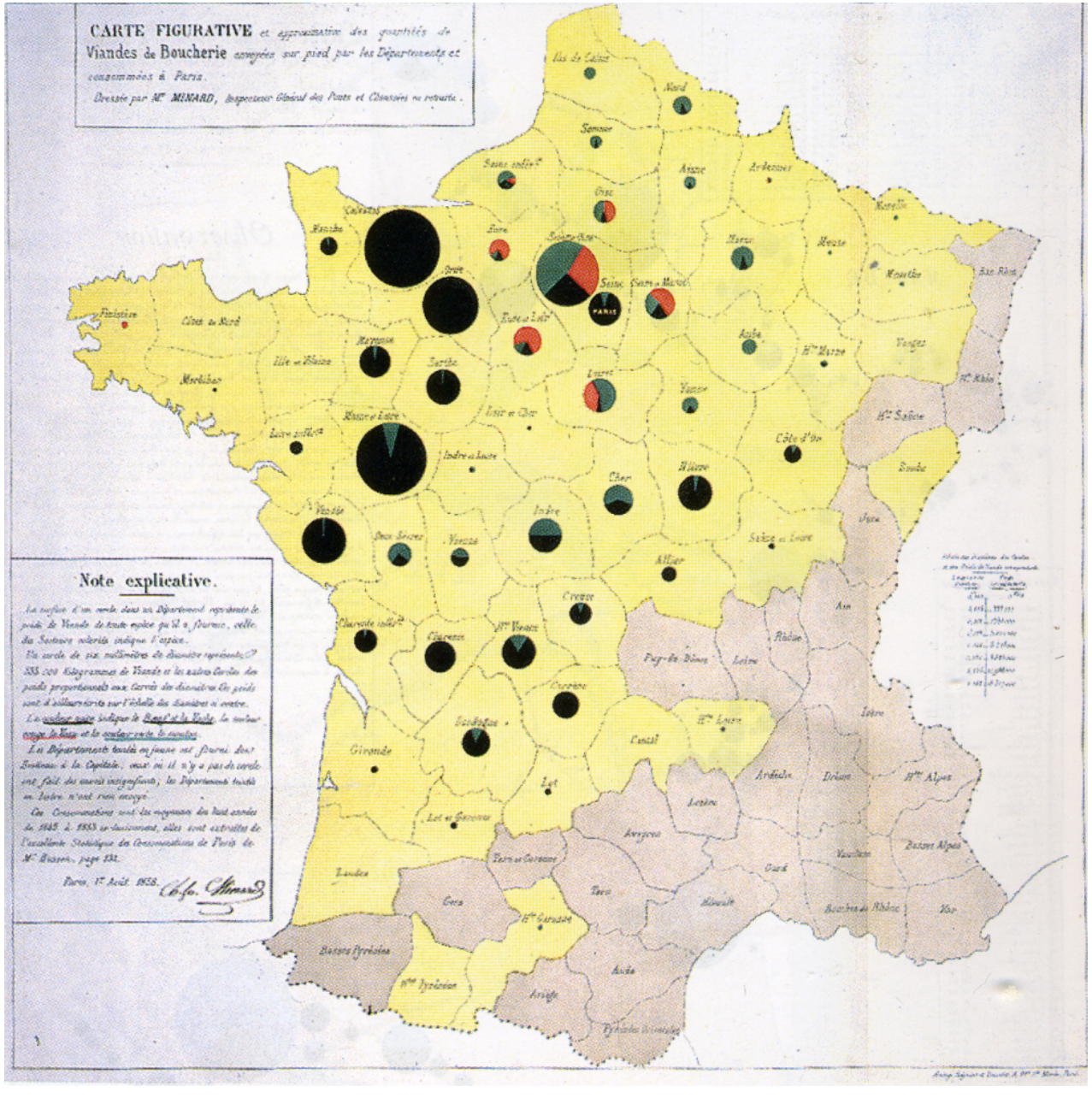


Historique
Minard

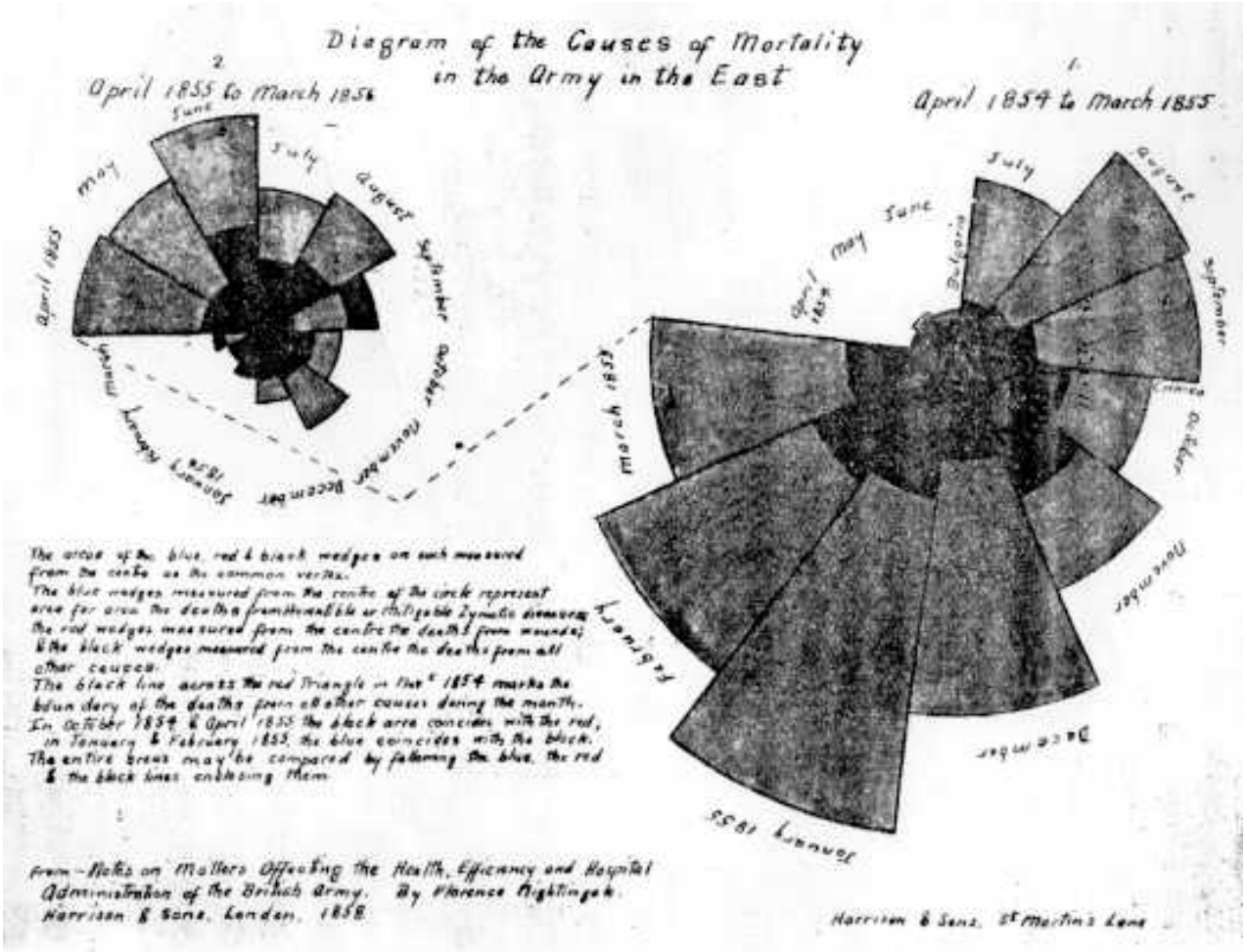
- Charles Minard (1781-1870) a fortement contribué au domaine de l'information graphique et statistique en particulier avec l'utilisation de cartes.



Historique
Minard

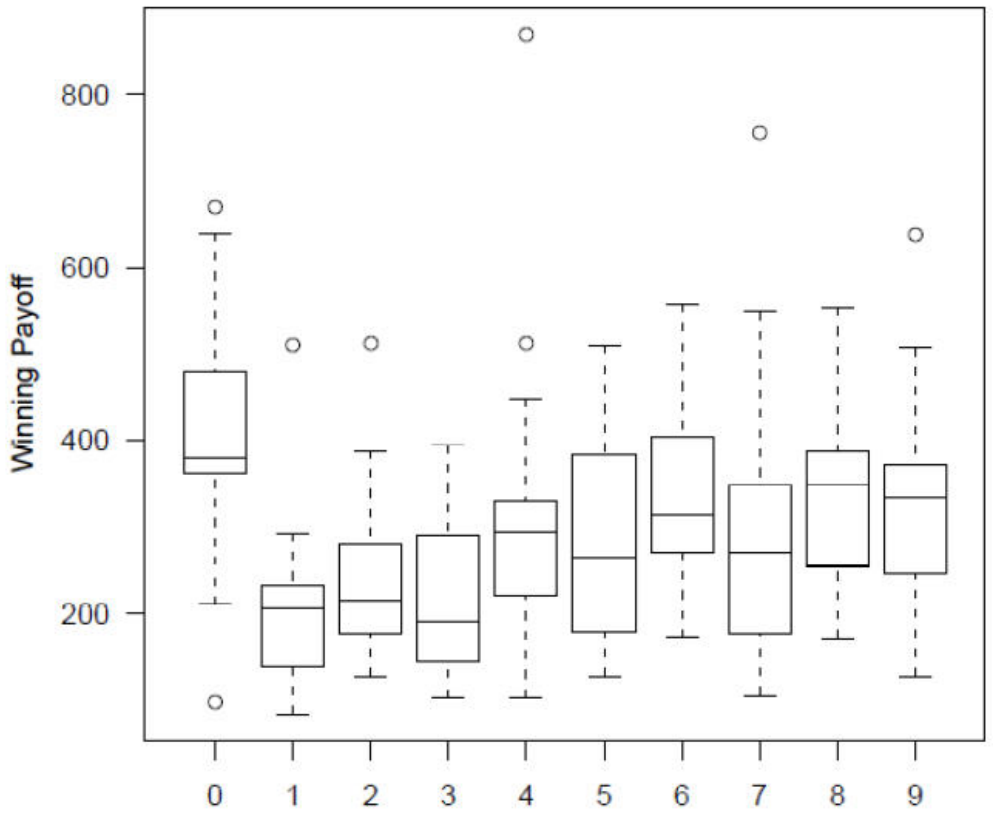


- Florence Nightingale (1820-1910) est connue comme étant la "mère" de la profession d'infirmière moderne. Elle a aussi contribué aux représentations graphiques.

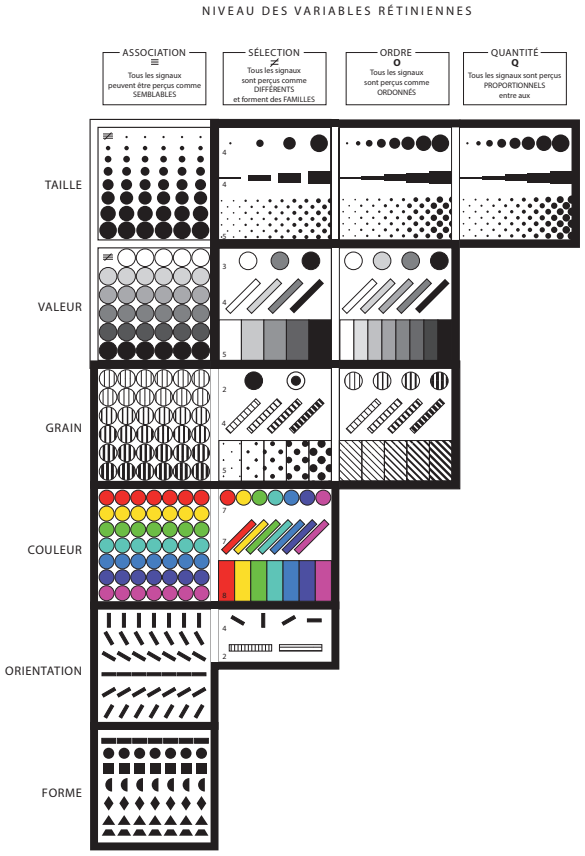


- John Snow (1813–1858) connu pour avoir tracé (détecté) les sources de Choléra à Londres.

- Ronald Fisher (1890-1962) et John Tukey (1915-2000) : méthodes graphiques avancées pour l'analyse des données.
- Fisher : dessin des données pour comprendre les relations
- Tukey : promotion de l'analyse de données exploratoires, il a créé en particulier le box plot, le stem plot et le leaf plot.



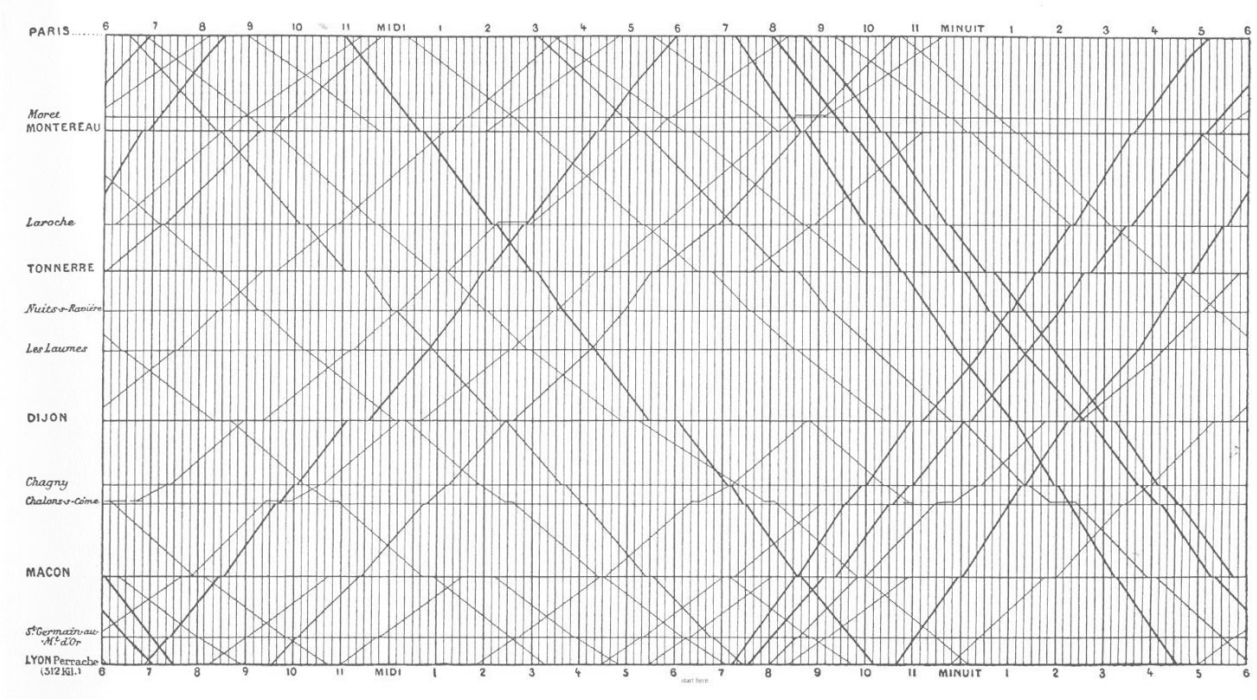
Historique
Bertin

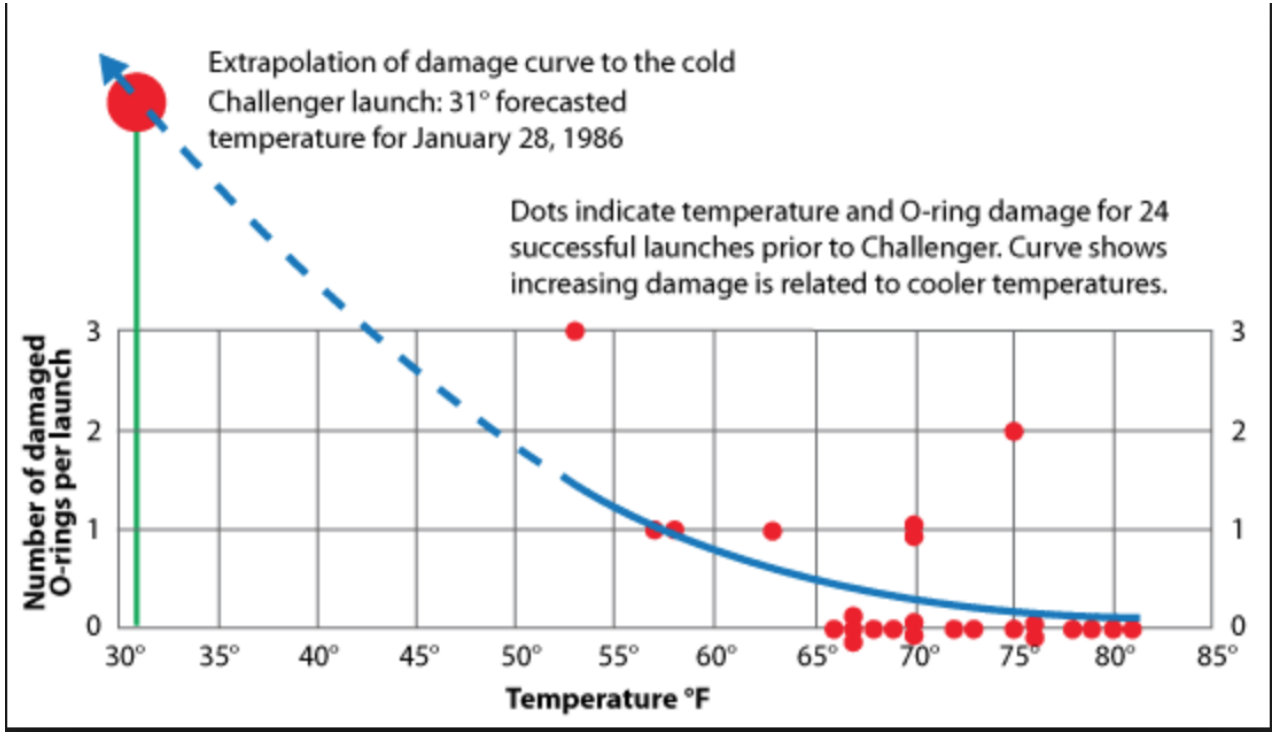


Jacques Bertin, "Sémiologie Graphique", 1973.

- Jacques Bertin (1918-2010): *sémiologie graphique*!
- Système de signes pour la transmission de l'information.

- Edward Tufte (1942-) a rédigé *The Visual Display of Quantitative Information*
- Importance de l'aspect mais pas à n'importe quel prix !

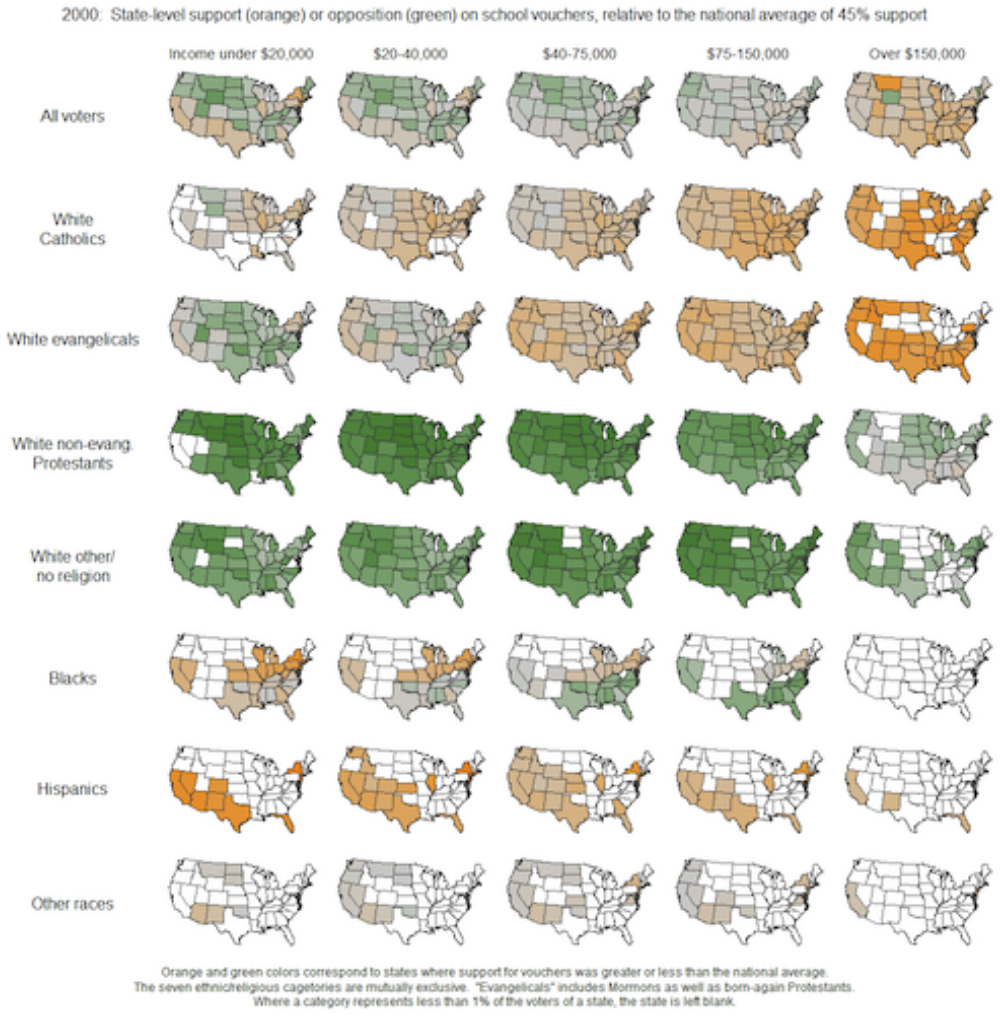




- Challenger revisité !

Historique

Tufte

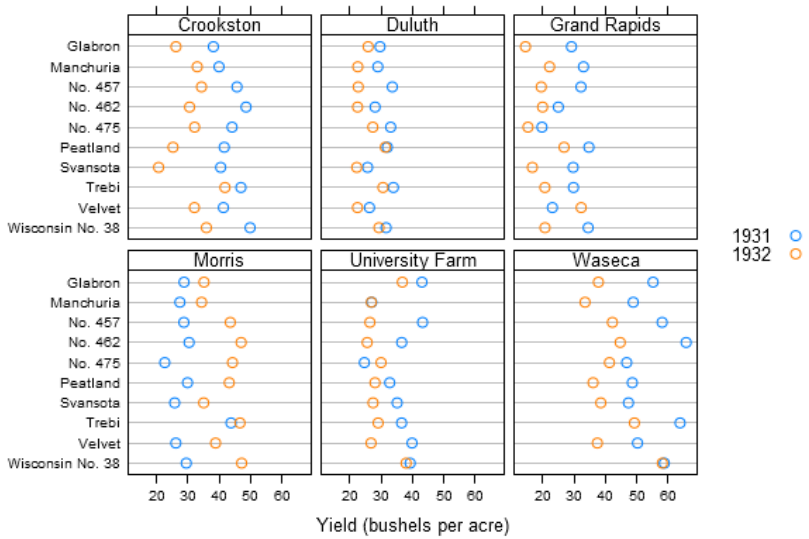


Example sparklines in small multiple

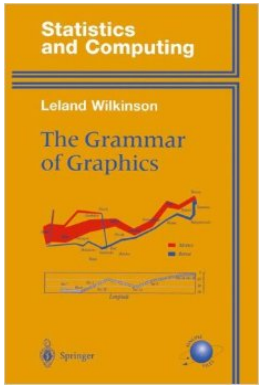
Index	Day	Value	Change
Dow Jones		10765.45	–32.82 (–0.30%)
S&P 500		1256.92	–8.10 (–0.64%)

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

- Tufte a développé et popularisé de nombreux principes :
  - 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]

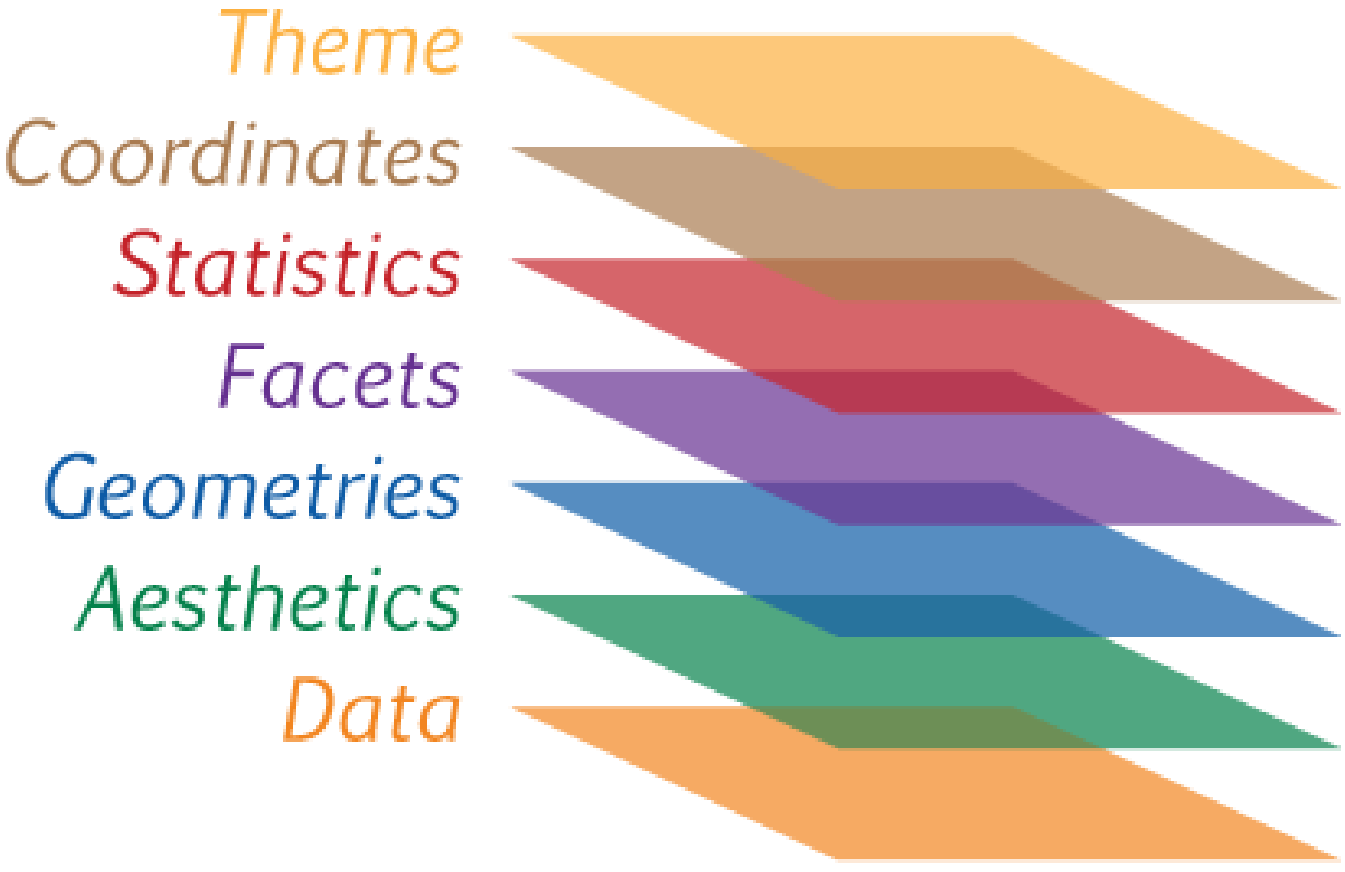


- William Cleveland's Elements of Graphing Data and Visualizing Data
- Cleveland est connu pour promouvoir le dot plot comme alternative aux barres, camembert...
- Le dot plot permet une certaine clarté et une facilité de comparaison des données.
- Cleveland est aussi un pionnier dans les treillis et les comparaisons de panels.



- The Grammar of Graphics de Leland Wilkinson (1945-), a eu une influence importante sur la facon de penser les graphes.
  - Grammaire signifie règles mathématiques et esthétiques
  - "Avant" on se focalisait surtout sur le coté esthétique d'un contenu statique
  - Par opposition, les graphiques dynamiques demandent une réflexion plus importante pour pouvoir zoomer, flouter, lier...
  - La Grammar... s'adapte facilement à cette nouvelle approche
- **ggplot2** (B. Wickham) inspiré par ce formalisme !

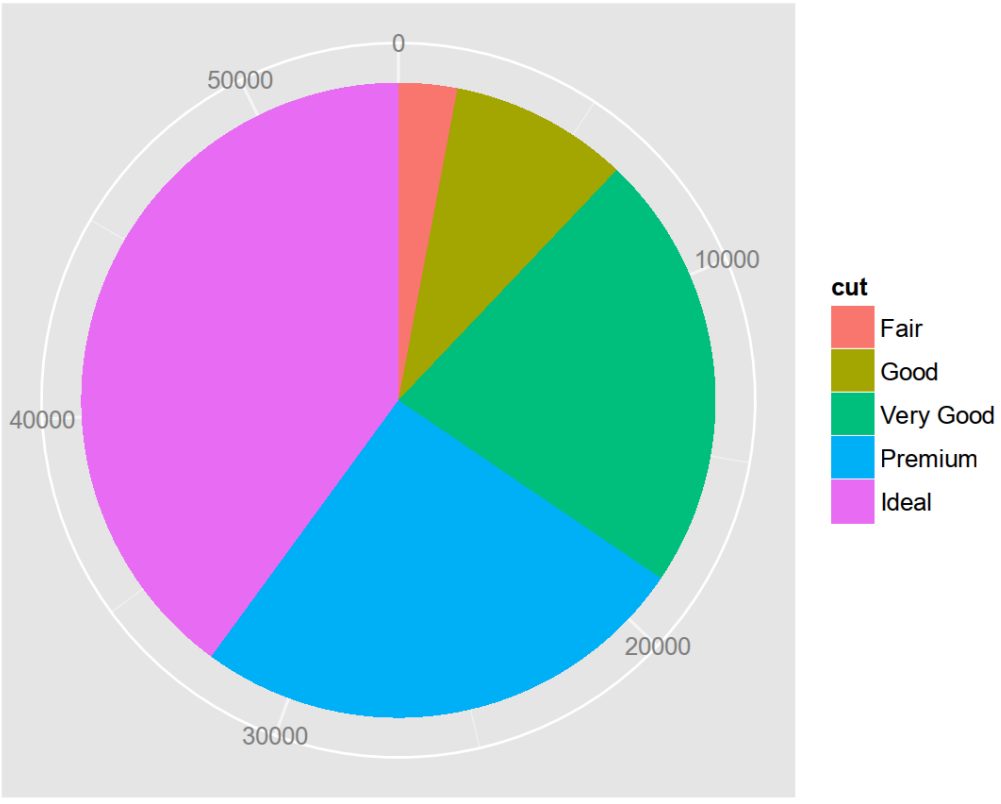
- DATA - weighting, reshaping, counting, bootstrapping
- VARIABLES - transform, sort, log, ranking, residuals, quantiles
- ALGEBRA - nesting or blending data
- SCALES - nominal, ordinal, interval, ratio must be specified
- STATISTICS - static methods available to all graph types e.g, mean, sd, smoothing
- 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


Theme
Coordinates
Statistics
Facets
Geometries
Aesthetics
Data

- Data : Les variables à afficher
- Aesthetics mapping : Les dimensions selon lesquelles les données sont représentées
- Geometries : Formes utilisées pour représenter les données
- Facets : Tableau (lignes et colonnes) de graphes
- Statistics : Modèles ou transformations statistiques des données
- Coordinates : L'espace de représentation (horizontal, vertical, cartésien, polaire)
- Scales : L'échelle des axes (linéaire, logarithmique, à l'envers), les couleurs de remplissage
- Thèmes : Description de l'arrière plan

Les classiques

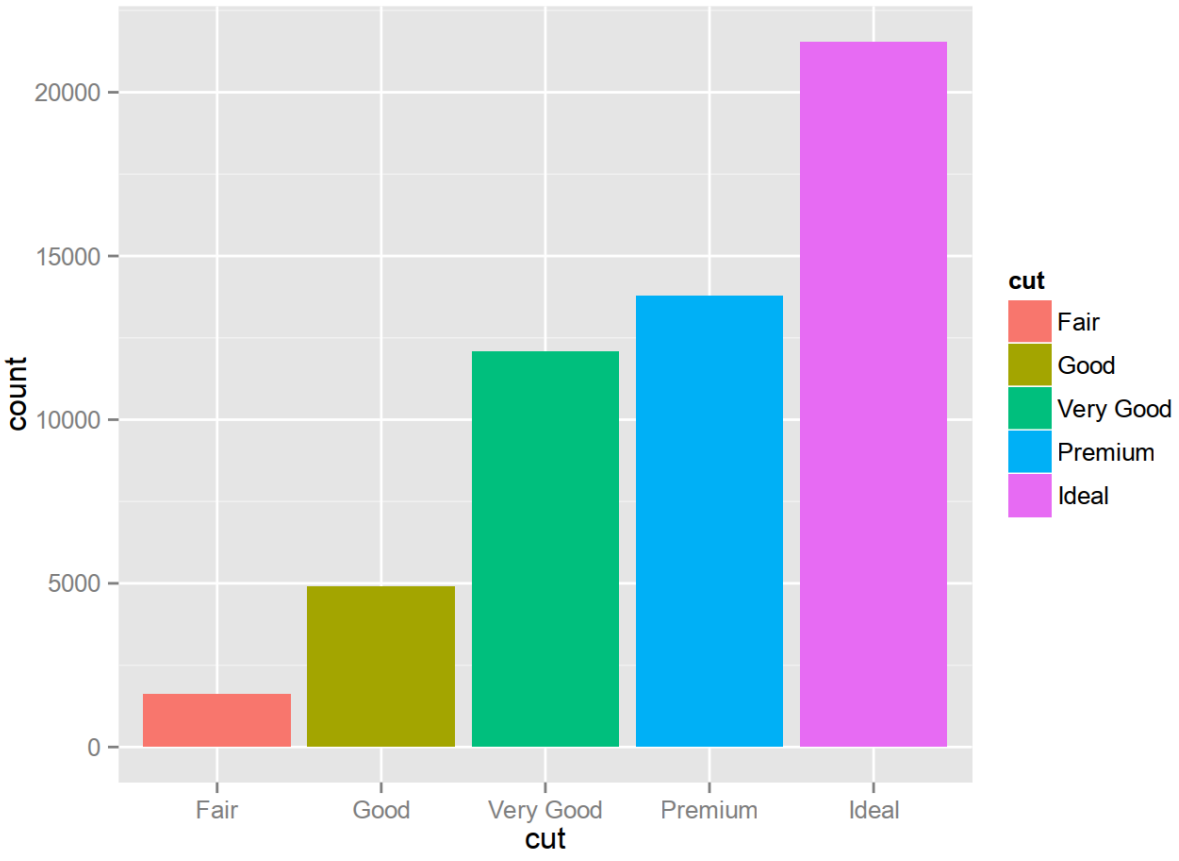
Camembert ou la tarte



- à ne pas utiliser !

Les classiques

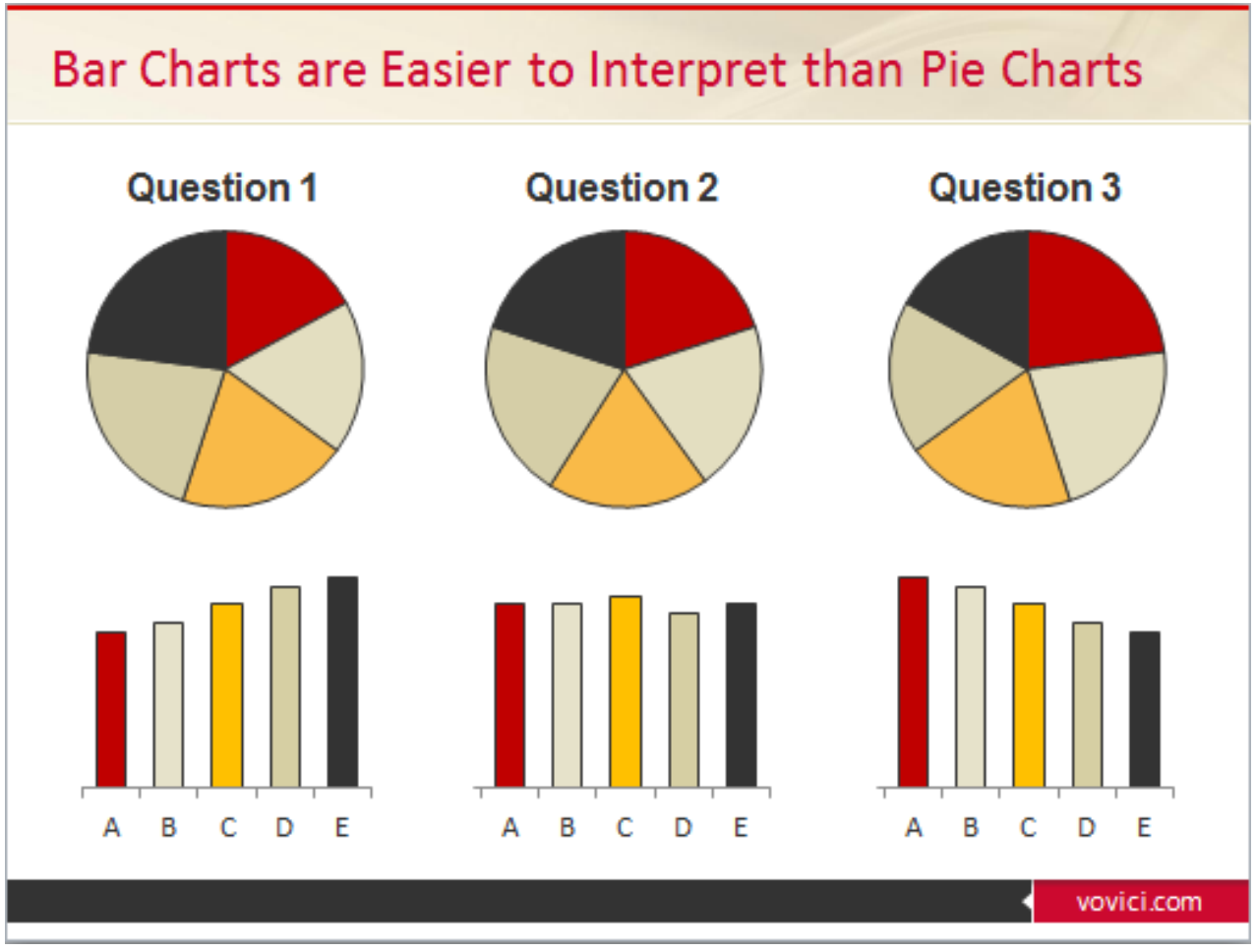
Barres **barplot** ou **plot**



- permet de meilleures comparaison
- adapté aux variables qualitatives

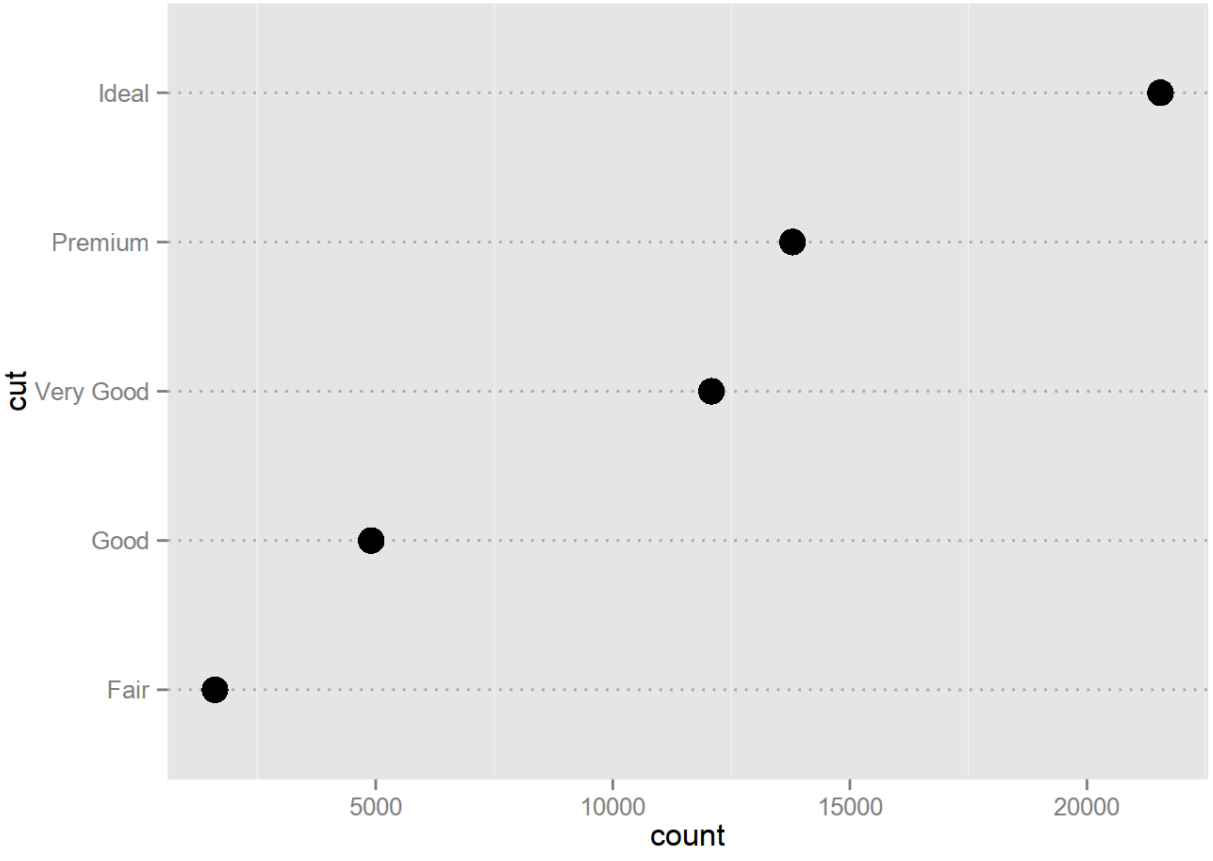
Les classiques

Barres **barplot** ou **plot**



Les classiques

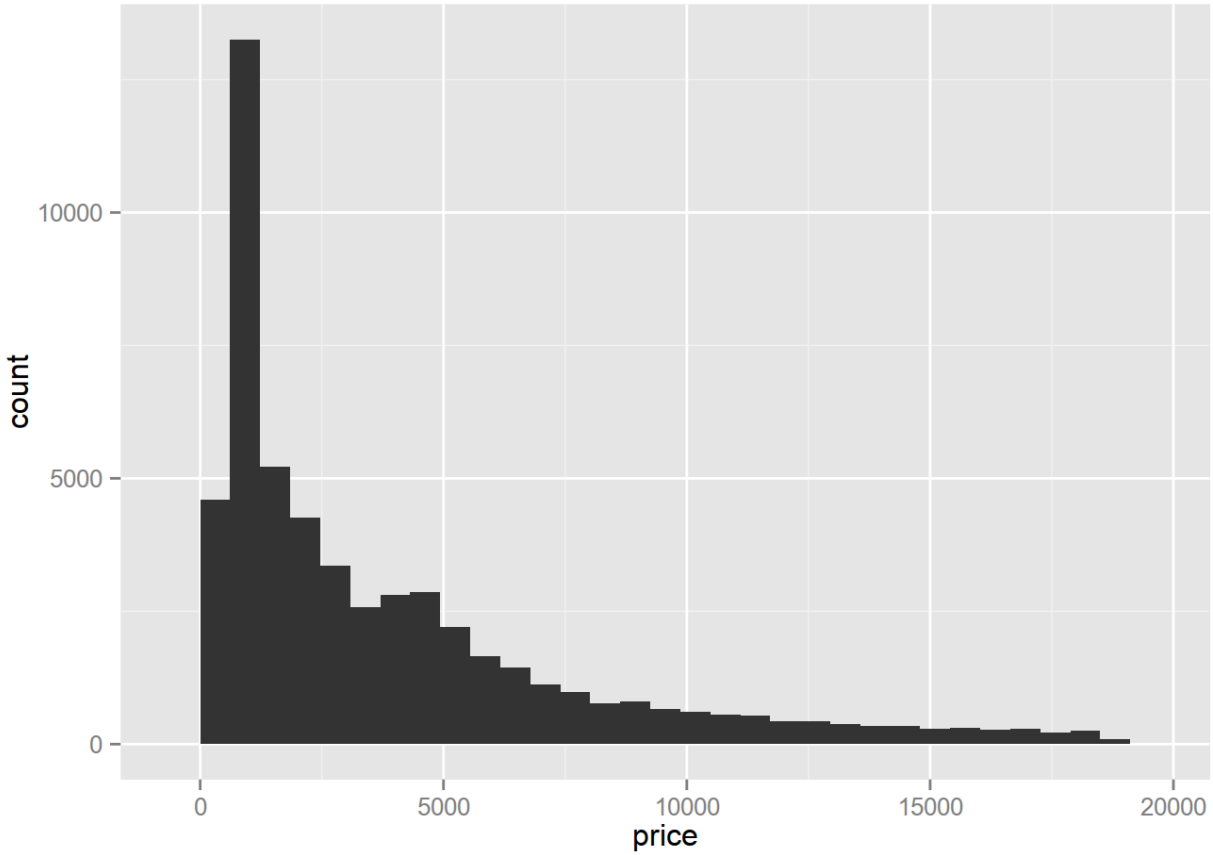
Les dot plot de Cleveland



- Less *ink*, more pleasant...

Les classiques

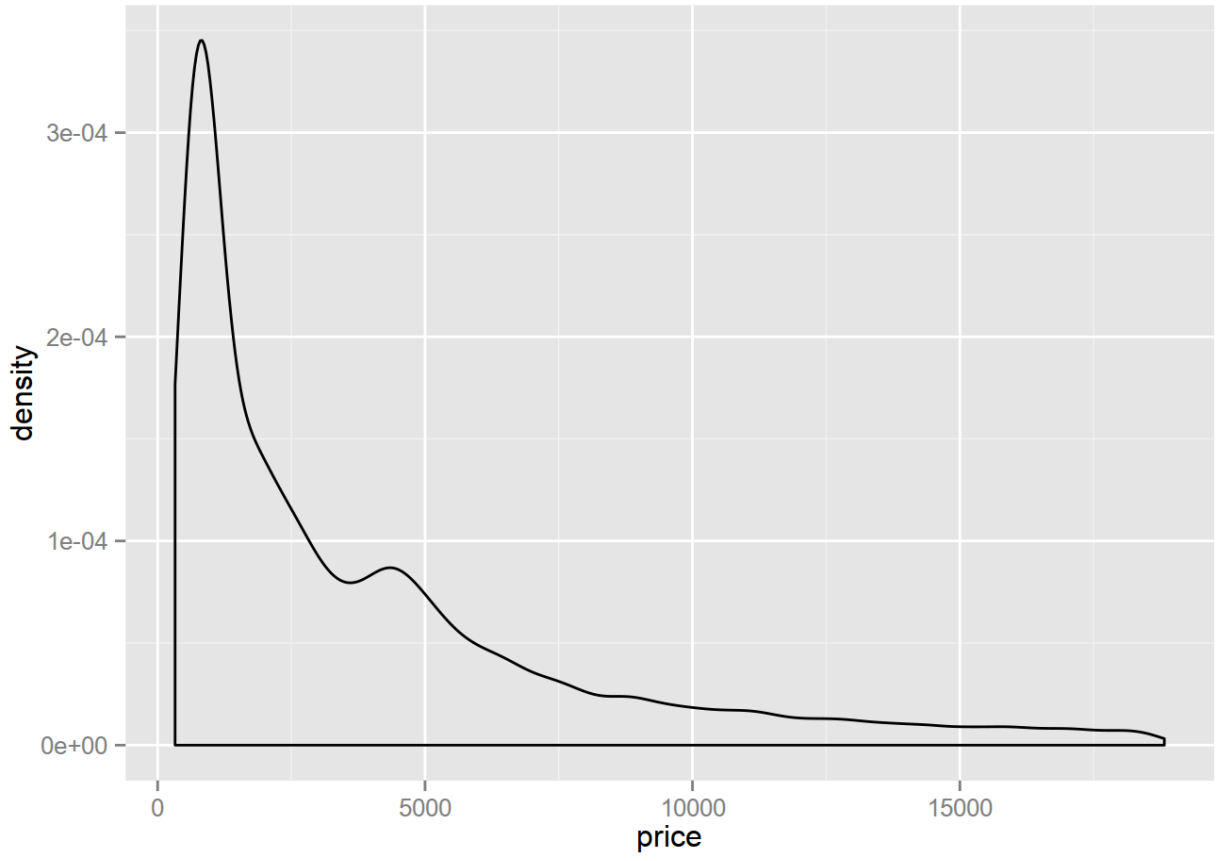
Histogramme et estimateur de la densité



- facilement interprétable
- adapté aux variables quantitatives

Les classiques

Histogramme et estimateur de la densité



- lissée