Types of data: Measurement scales

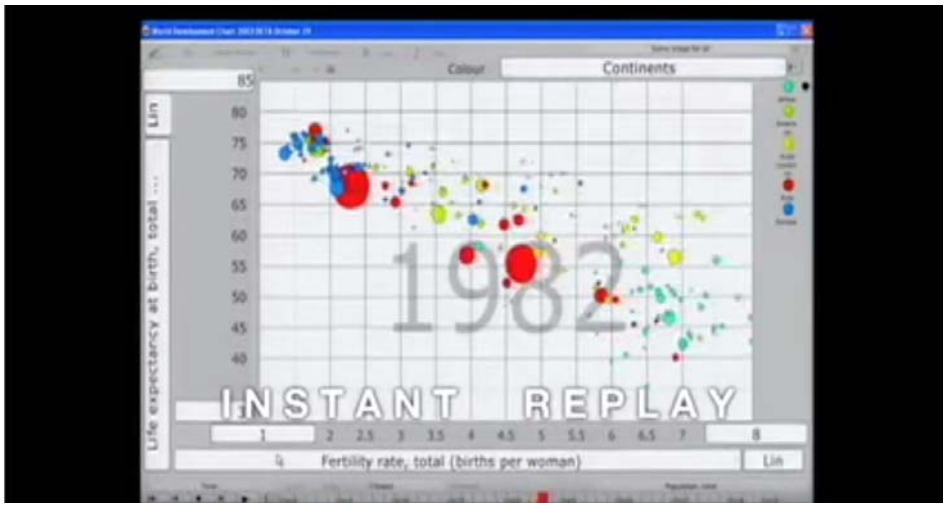
- Nominal
  - Example: gender
- Ordinal
  - Example: educational level
- Interval
  - Example: annual income
- Why does it matter?
  - Different measurement, scales or variables are analyzed differently (for example: it wouldn't be very informative to compute the average race in a population)

If we collected the data, now what?...

Exploratory Data analysis (EDA)

- This analysis should precede the data analysis
  - Detect data errors
  - Outliers
  - Check statistical assumptions
  - Understand your data! (variables relationships, trends, heterogeneous differences, etc.)
- This could include, for example:
  - Frequency tables
  - Distribution: Central tendency and spread of data
- Some graphical tools (we assume that you are familiar with these and their interpretation):
  - Boxplots (IQRs and outliers) and scatter plots
  - Quantile-normal plots
  - Cross tabulation
- You will do an exercise about this in the review session

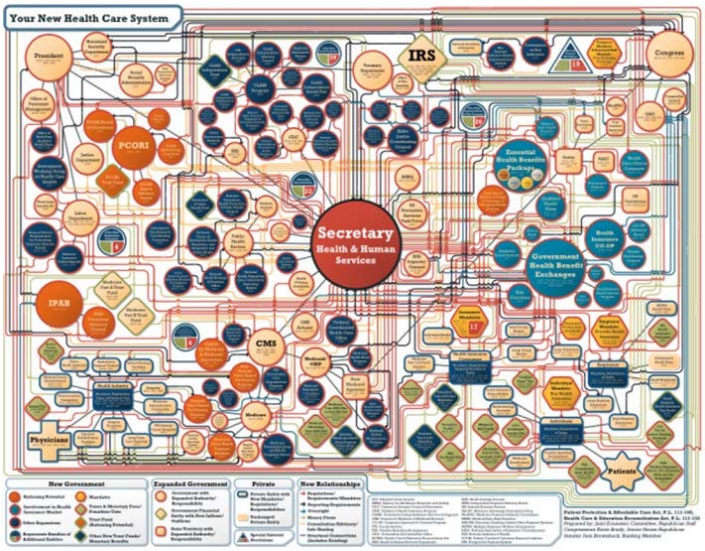
Data visualization: Motivation



- You can build your own examples in gapminder.org

Data visualization

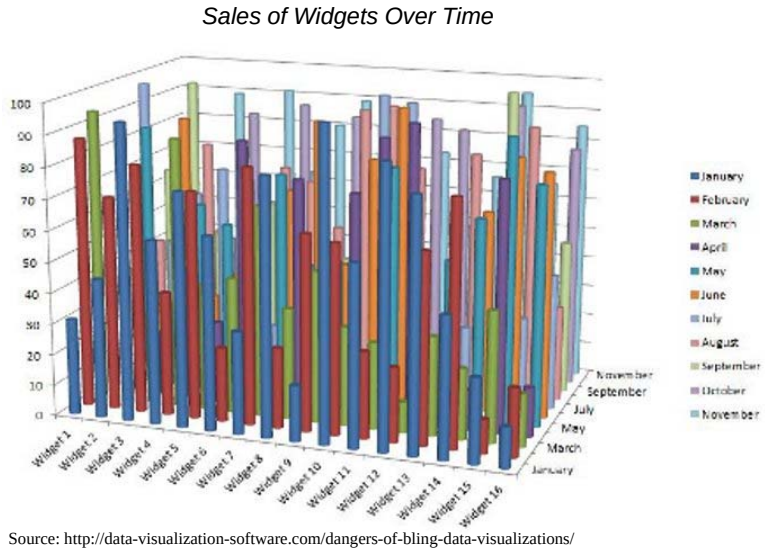
- Let's start with bad data visualization…



Source: https://www.seo.com/blog/infographics-vs-infocrapics-the-good-the-bad-the-ugly/

Data visualization

- Let's start with bad data visualization (cont'd)…



Source: http://data-visualization-software.com/dangers-of-bling-data-visualizations/

Data visualization

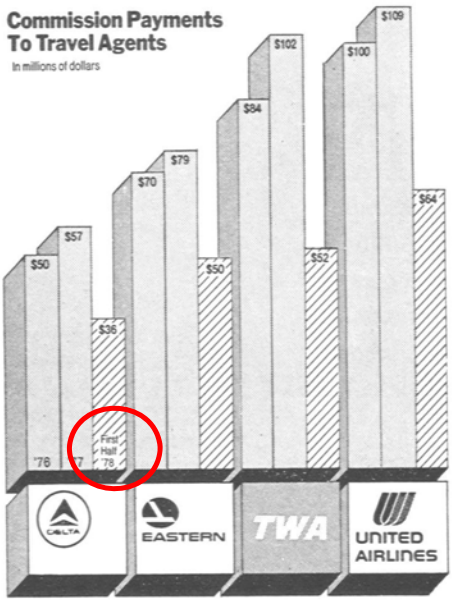
- Let's start with bad data visualization (cont'd)…

Source: http://data-visualization-software.com/dangers-of-bling-data-visualizations/

Data visualization: Graphical excellence

- Let's see some principles (based on Tufte's principles)
- Communicate complex ideas with:
  - Clarity
  - Precision
  - Efficiency
- Provide the greatest number of ideas in the shortest time with the least ink in the smallest space
- Tell the truth about the data

Data visualization: Graphical integrity

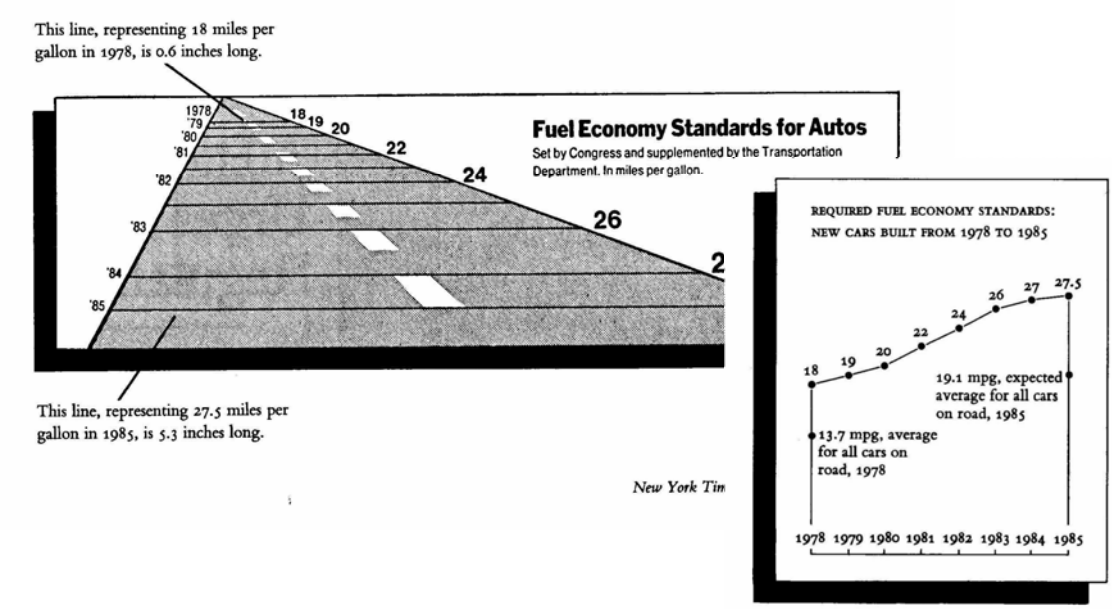


- Distortion: Visual representation of data is not consistent with the numerical representation.
- How can we test this?

$$Lie\ Factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data}$$

  - Small distortion: 0,95 < LF < 1,05
  - Example…

Data visualization: Graphical integrity



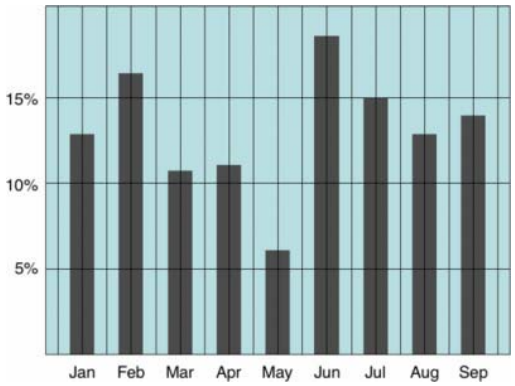
Data visualization: "Above all else show the data"

- When you see a graph you should be drawn to essence of the data (the focus shouldn't be the visualization technique)
- Data-ink (the part that cannot be erased to display the core of the data)
  - Most of the graph should show data information

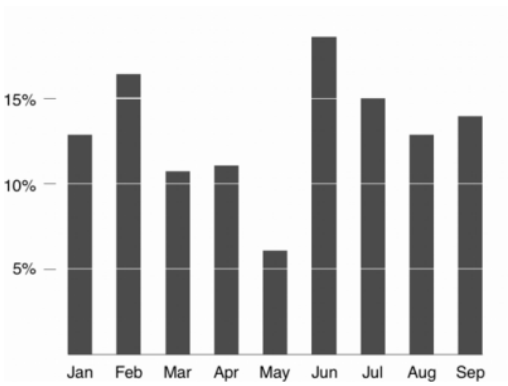
$$Data-ink\ ratio = \frac{data-ink}{total\ ink\ used\ to\ print\ graphic} = 1 - proportion\ of\ a\ graphic\ that\ can\ be\ erased\ without\ loss\ of\ data-information$$

  - Maximize the numerator or erase non-data ink, with reasons (e.g. avoid redundancy, unless facilitate comparison)

Data visualization: “Above all else show the data”



Low Data-Ink Ratio

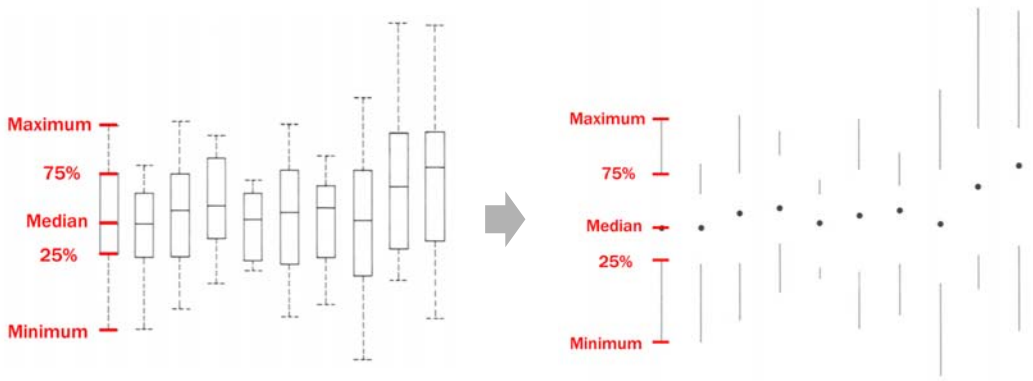


High Data-Ink Ratio

Source: http://www.infovis-wiki.net/index.php/Data-Ink_Ratio

Applications

- Boxplots (this is just a suggestion! – based on Tufte’s principles) – what are we missing???



- Criticism (Stryjewski, 2010)

Ethical principles

- Experimental research
  - Tuskegee Syphilis study

The New York Times

Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

  - Role of International Review Boards
    - Respect for persons (role of informed consent / coercion)
    - Beneficence (role of Zimbardo's prison study)
    - Justice

Ethical principles

- Confidentiality
- Anonymity
  (e.g., in surveys)

- How about when knowing about the study may affect behavior

Ethical principles: Plagiarism

Los "copy paste" más escandalosos de los políticos en los últimos años

Desde Wikipedia hasta el guión de una película hollywoodense han utilizado algunos honorables para salir del paso a la hora de elaborar importantes documentos. Revise acá los "copy paste" más célebres en Chile y el extranjero.

Emol domingo, 14 de abril de 2013 8:22

CNN fires news editor Marie-Louise Gumuchian for plagiarism [Updated]

By Erik Wemple May 16, 2014

Confesiones de un plagiador

El chileno Rodrigo Núñez Arancibia había construido una exitosa carrera de historiador en México. El problema es que estaba sustentada en los trabajos de otros investigadores, y el mes pasado su caso se destapó. "Yo sabía que iba a chocar como un tren contra una pared".

Ethical principles: Data fabrication

The Scientist » News & Opinion

Parkinson's Researcher Fabricated Data

Neuroscientist Mona Thiruchelvam agrees to retract two studies linking neurodegeneration to pesticides.

By Hayley Dunning | June 29, 2012

Journal Retracts Faked Study About Gay People Changing Voters' Minds

Michael LaCour, a graduate student, apparently made up the results of the much-publicized study in the journal *Science*. The study was formally retracted Thursday.

Originally posted on May 20, 2015, at 9:34 a.m.
Updated on May 28, 2015, at 3:42 p.m.

Virginia Hughes
BuzzFeed News Science Editor

Ethical principles: Data manipulation and "*p*-hacking"

NEUROBONKERS

The statistical significance scandal: The standard error of science?

by SIMON OXENHAM

P<0.05 is the figure you will often find printed on an academic paper, that is commonly (mis)understood as indicating that the findings have a one in twenty chance of being incorrect. The phenomenon has become a somewhat universal barrier which scientists must cross but in many cases has also inadvertently become a barrier to readers of scientific research accessing the very important numbers often hidden underneath this indication of statistical significance. The US Supreme Court for example has investigated cases where statistical significance of findings in medical trials has been used in place of undisclosed adverse events:

"In a nutshell, an ethical dilemma exists when the entity conducting the significance test has a vested interest in the outcome of the test."

Table 1. Likelihood of Obtaining a False-Positive Result

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science.*

About the cases on ethics we have discussed, what are the main motivations that lead people to do it?