

# DATA VISUALIZATION WITH R

## Principles and Practice

---

Constantin Manuel Bosancianu

May 5, 2021

Wissenschaftszentrum Berlin

*Institutions and Political Inequality*

[bosancianu@icloud.com](mailto:bosancianu@icloud.com)

# Preamble

---

# Plan for today

Things to speak about:

- ✓ The importance of data visualization;
- ✓ Basics of *good* data visualization;
- ✓ “The *good*, the *bad*, and the *ugly*” when it comes to data visualization—examples.

The rest of the course is hands-on, using *R* and *ggplot2*.

# Introduction

---

# Importance

---

There is more data than ever waiting to be analyzed, mined for patterns, summarized, or linked to other data.

We also observe a phenomenal level of growth in individual-level data: Internet, GPS-enabled smartphones, e-scooters, smartwatches, fitness trackers, automated sensors, smart speakers etc.

# Importance

---

Presenting such information in an accurate and intuitive way for the purpose of highlighting causal connections will be crucial for our ability to make adequate choices in a democracy.

# Principles

---

# Data visualization (DV)

At the confluence between statistics and design, dealing with the search for the most effective and graphically intuitive way of making an argument on the basis of data.

In 2000, an estimated 900 billion ( $9 * 10^{11}$ ) to 2 trillion ( $2 * 10^{12}$ ) graphs were generated every year (Tufte, 2001).

# Goals of DV

---

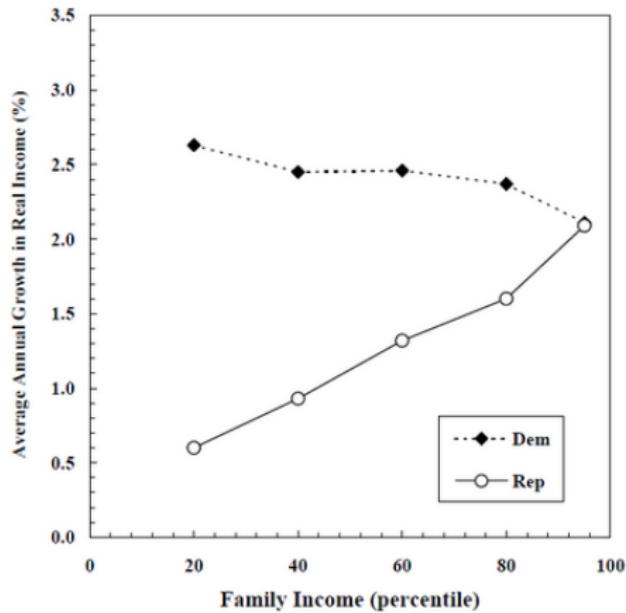
Multiple:

- ✓ Making an argument;
- ✓ Minimizing any distractions from the central argument;
- ✓ Ensuring the integrity of the argument;

“Making a presentation is a moral act as well as an intellectual activity.” (Tufte, 2006, p. 141)

# Simple, yet effective

**Figure 1: Income Growth by Income Level in Democratic and Republican Administrations, 1948-2001**



Economic performance under US presidents (Bartels, 2008)

# Goals of DV

---

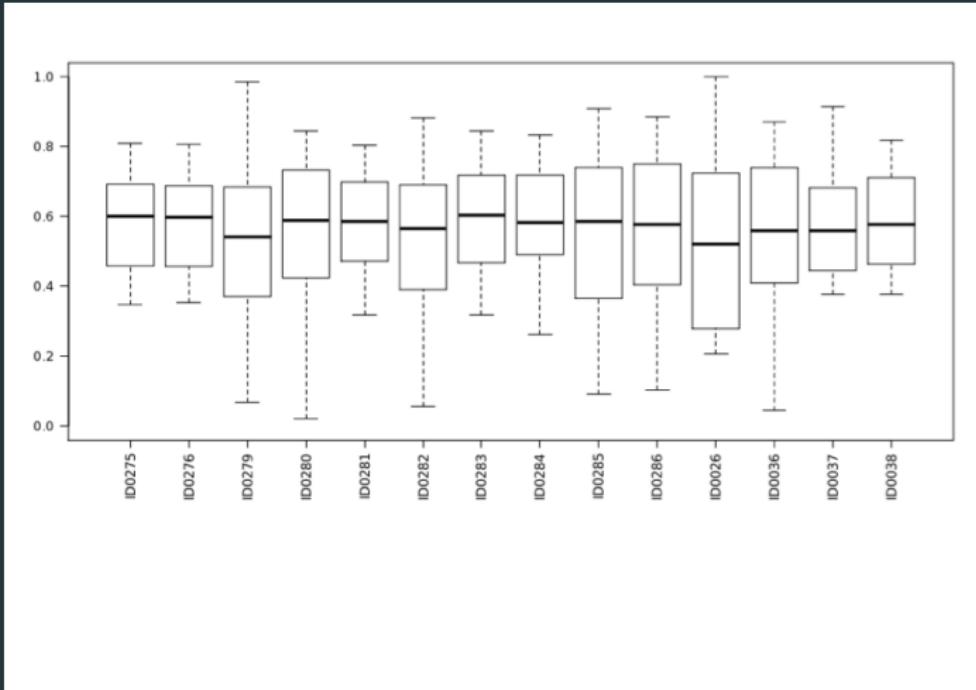
Two more:

- ✓ Summarizing a lot of information in a reduced space;
- ✓ Encouraging comparison.

## Principles of DV (adapted from Tufte, 2001)

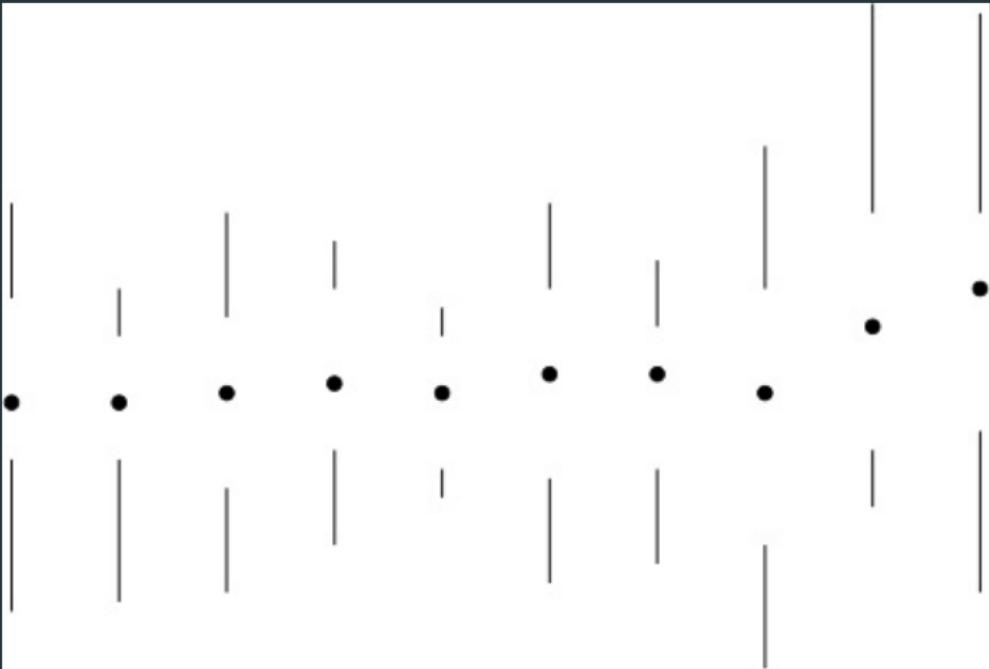
- ✓ The overarching purpose is to show the data;
- ✓ Minimize the data-ink ratio, as much as possible;
- ✓ Erase non-data-ink, as much as possible;
- ✓ Minimize redundant data-ink, as much as possible;
- ✓ Revise and edit;
- ✓ Mobilize every graphical element needed.

# Keeping it simple



Should be as simple as it needs to be, but not more

# Keeping it simple



## ACCENT principles (from Burn, 1993)

---

**A**pprehension: ability to correctly perceive relations among variables.

**C**larity: Ability to visually distinguish all the elements of a graph.

**C**onsistency: Ability to interpret a graph based on similarity to previous graphs.

# ACCENT principles

---

**E**fficiency: Ability to portray a possibly complex relation in as simple a way as possible.

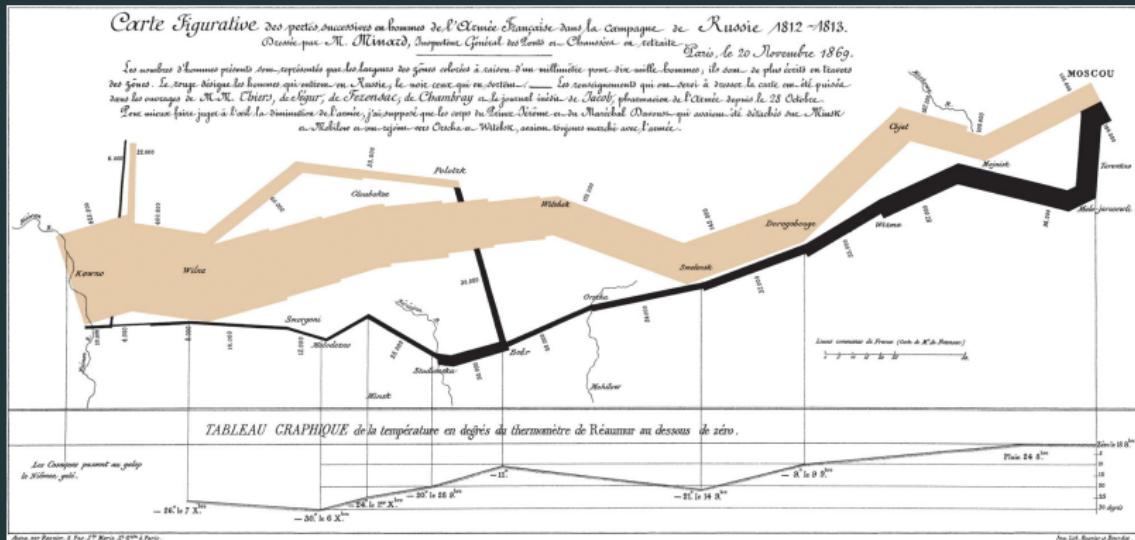
**N**ecessity: The need for the graph, and the graphical elements.

**T**ruthfulness: Ability to determine the true value represented by any graphical element by its magnitude relative to the implicit or explicit scale.

# Good examples

---

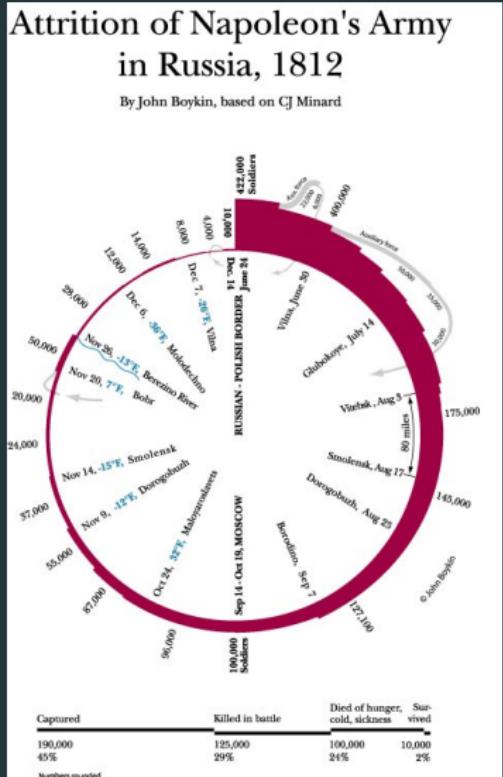
# Napoleon's 1812–1813 campaign



Multiple dimensions on a 2-dimensional map

Temperature, deaths, movement, advance/retreat.

# Multiple ways of presenting

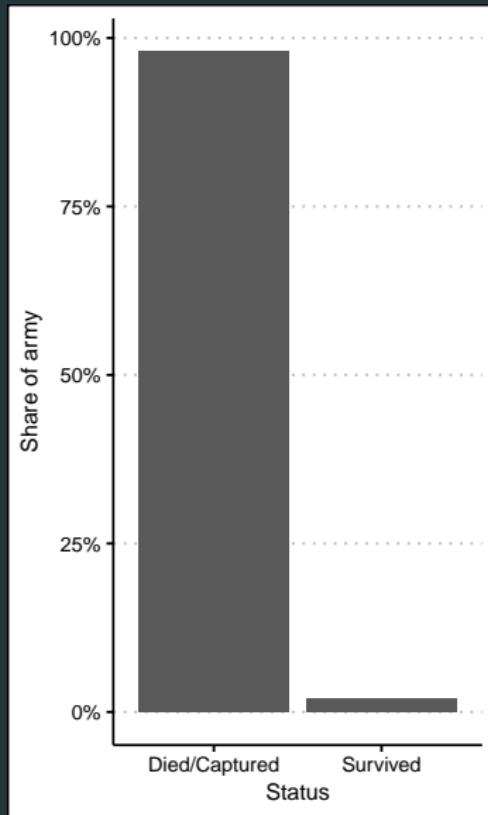


Map matters less than number of deaths, advance/retreat, and temperature.

Emphasis on dismal rate of survival.

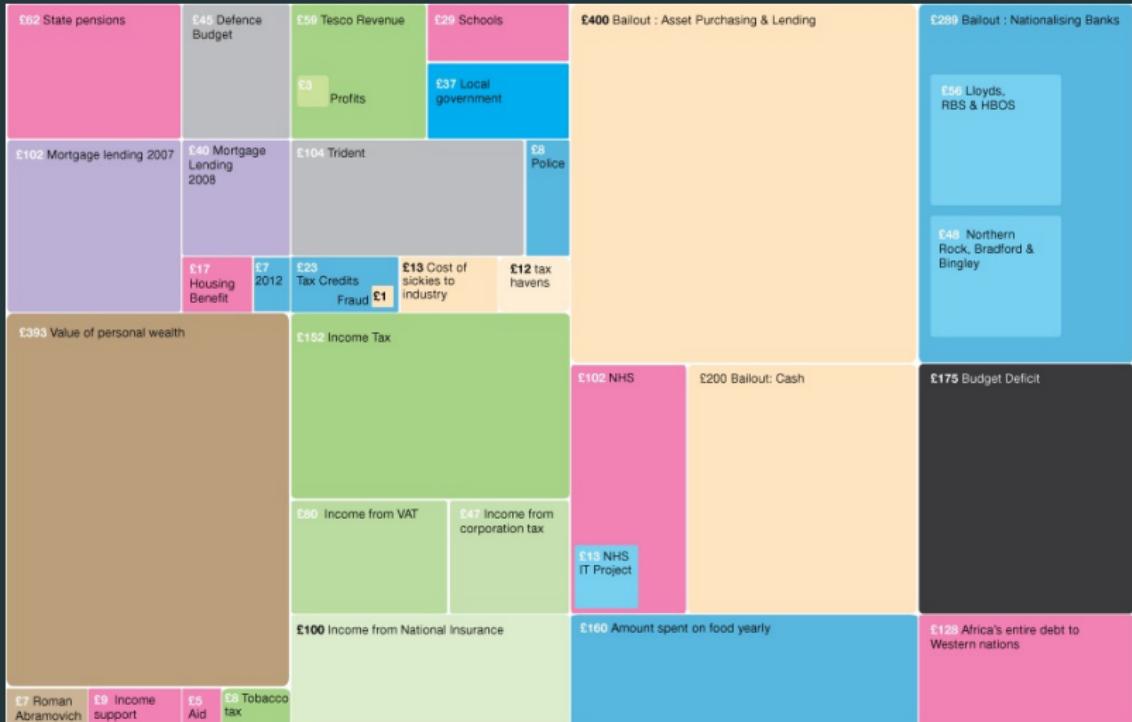
Temperature matters more in winter.

# Focus on a single aspect



A simple point shouldn't require a complex graph.

# Getting a sense of large numbers

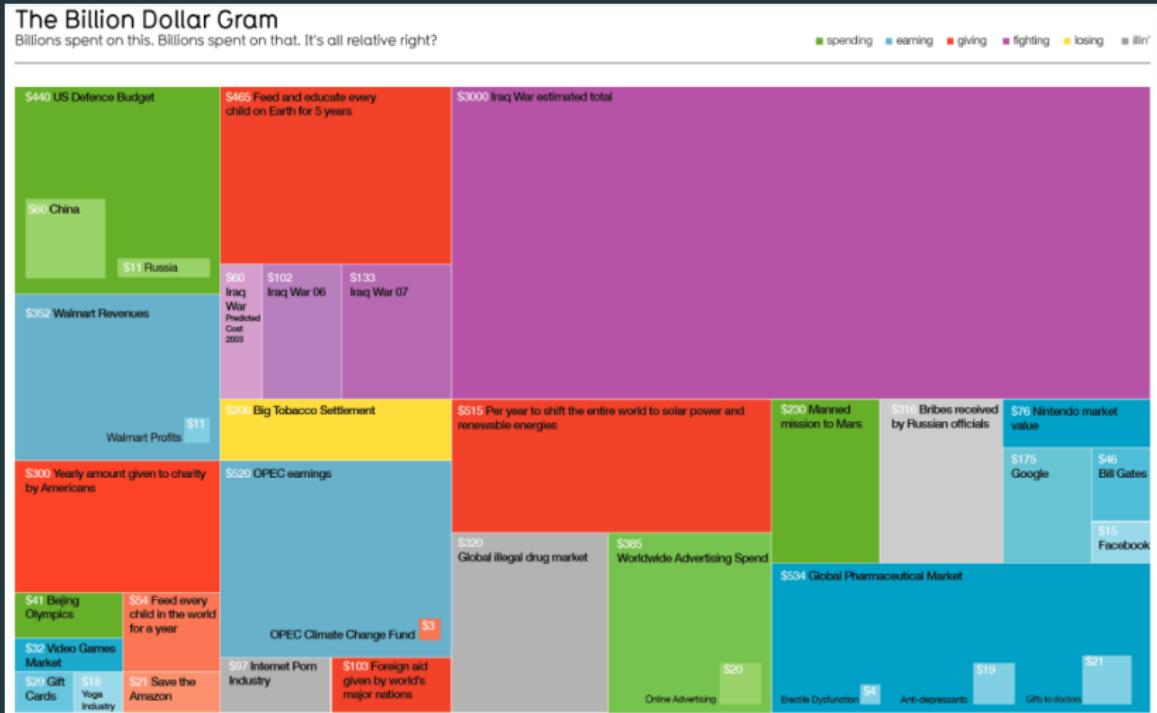


## The Billion Pound-O-Gram

David McCandless / [InformationIsBeautiful.net](http://InformationIsBeautiful.net)

Source: UK Treasury, Guardian

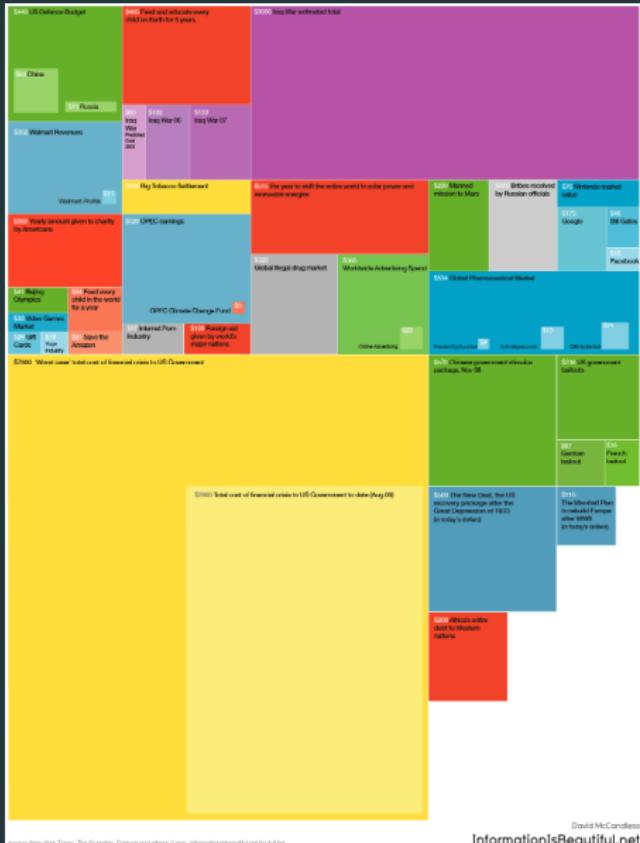
# (not really intended for accurate comparisons...



# ...but rather for relative dimensions)



# Putting it into perspective



Cost of crisis over 2008–2016 estimated at 4.6 trillion USD.

<https://hbr.org/2018/09/the-social-and-political-costs-of-the-financial-crisis-10-years-later>

# Ontario welfare benefits

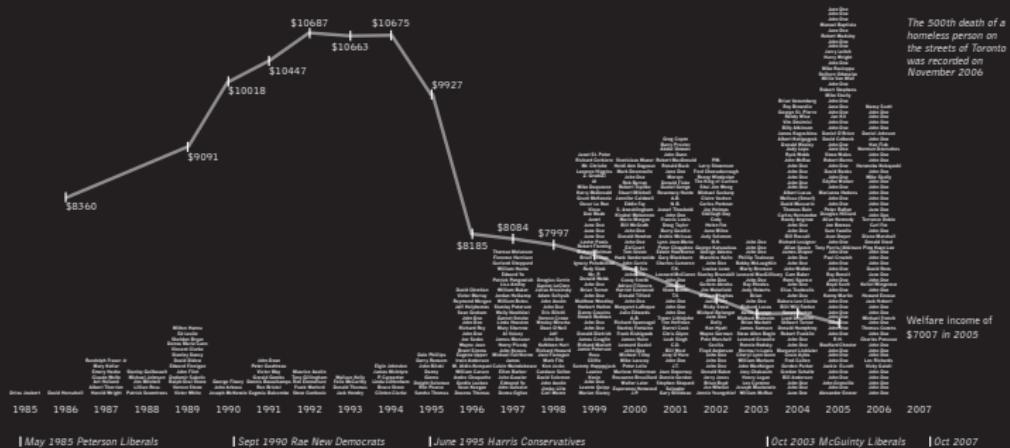
## Common Sense Revolution

Ontario Welfare Income for a Single Person in 2005 Constant Dollars &

Homeless Persons Who Have Died on the Streets of Toronto 1985–2006

(National Council of Welfare & the Toronto Disaster Relief Committee)

© Scott Sparling



| May 1985 Peterson Liberals

| Sept 1990 Rae New Democrats

| June 1995 Harris Conservatives

| Oct 2003 McGuinty Liberals | Oct 2007

# Bad examples

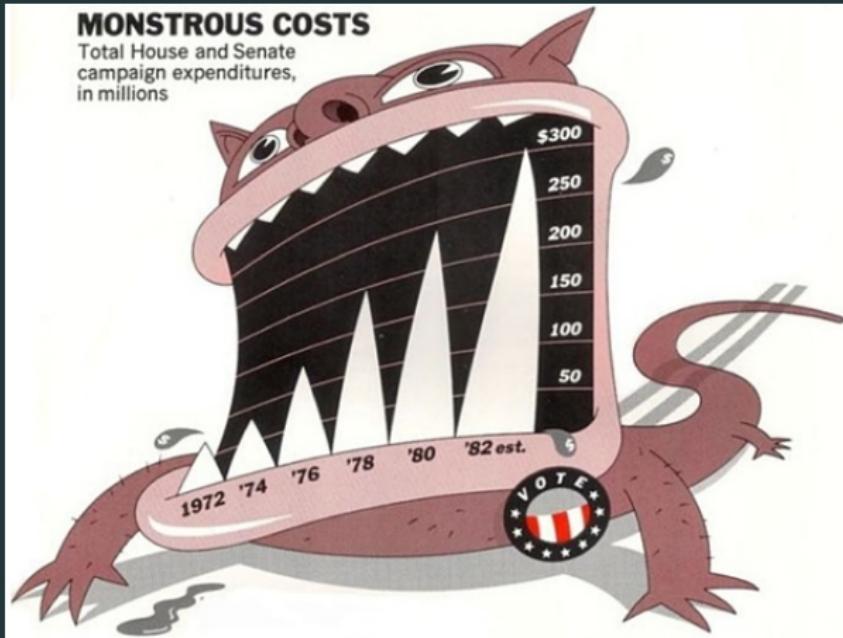
---

## Violations of basic rules (Tufte, 2001)

---

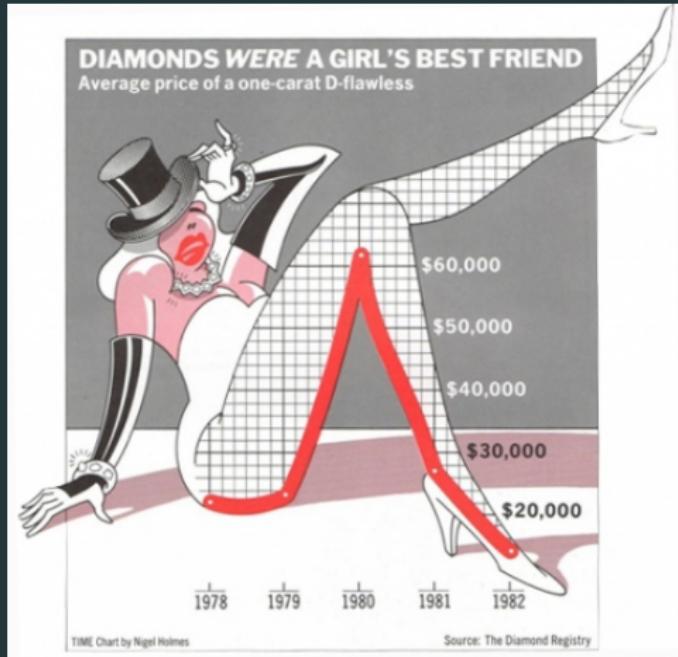
1. Showing more dimensions than there are in the data
2. Showing surfaces which are disconnected from the underlying quantities
3. Varying the design even though the data isn't varying
4. Failing to adjust for inflation, population change etc. in time-series data
5. Failing to provide adequate context for the data
6. Failing to put proper labels on axes

# Chartjunk



Unnecessary features galore (keep in mind, though: they might be easier to recall)

# Chartjunk



Unnecessary features (and sexism) galore (these are aesthetic violations, following Healy, 2018)

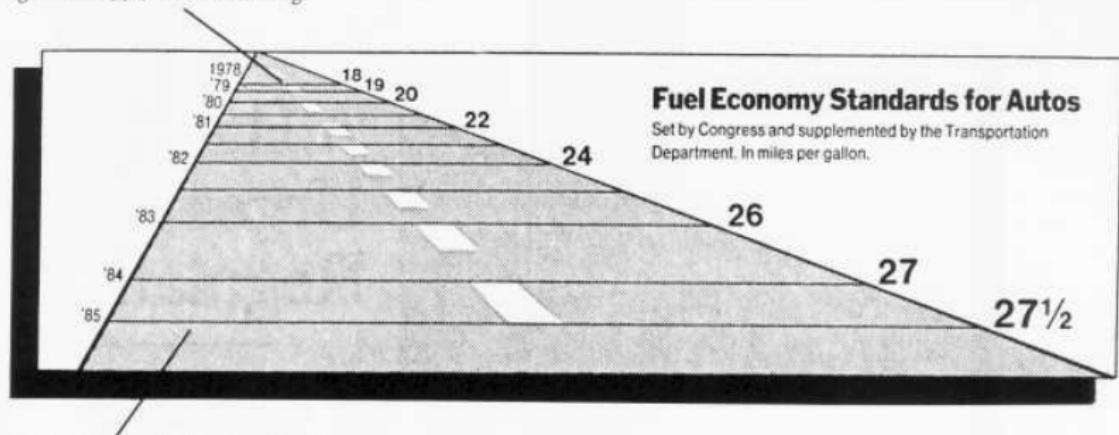
# Chartjunk



Unnecessary features, and poor choice of display

# Toying with perception

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

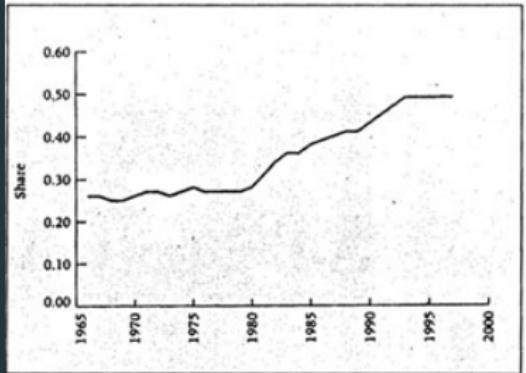


This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

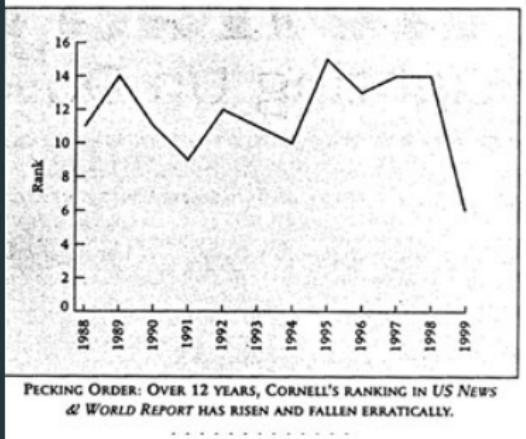
*New York Times*, August 9, 1978, p. D-2.

Supports the point of radical government overreach

# Toying with perception



By the Numbers: Over 35 years, Cornell's tuition has taken an increasingly larger share of its median student family income.



Pecking Order: Over 12 years, Cornell's ranking in US News & World Report has risen and fallen erratically.

3 things seem wrong here to me.

# Toying with perception



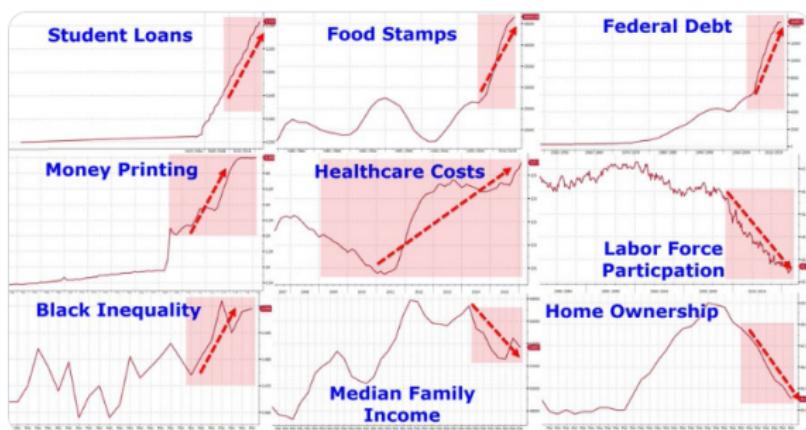
# Toying with perception



# Toying with perception

Donald J. Trump  @realDonaldTrump

"@TaylorEdwards99: THIS IS @POTUS'S LEGACY! AN ABSOLUTE DISASTER!!! WE NEED @realDonaldTrump NOW!! #MAGA #TRUMP2016

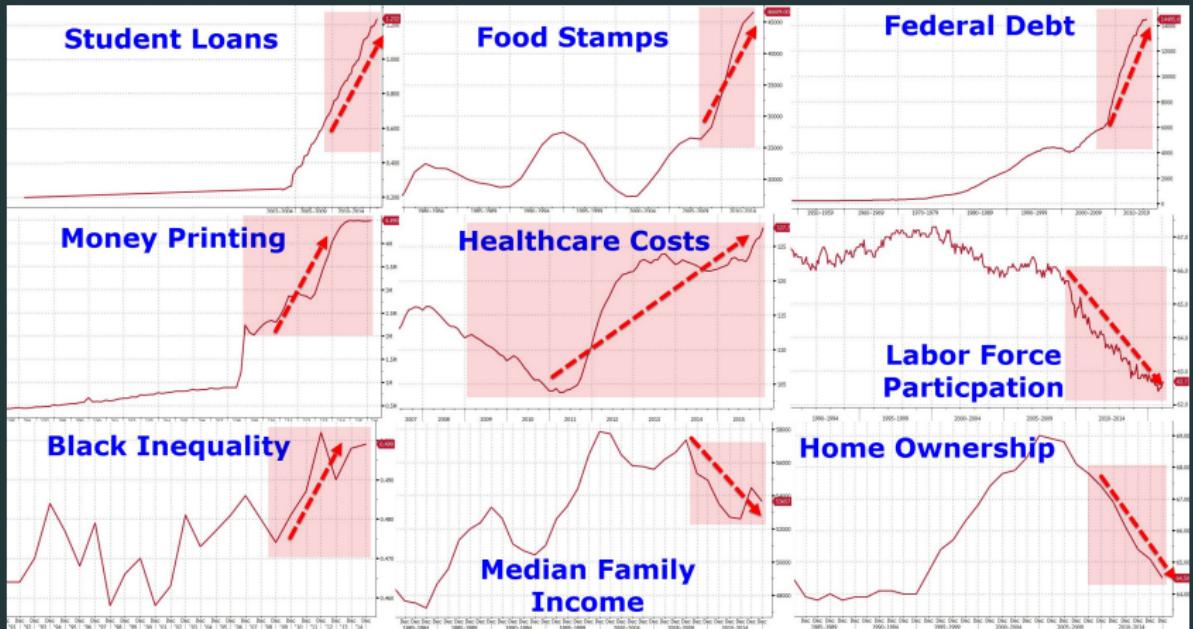


Student Loans      Food Stamps      Federal Debt  
Money Printing      Healthcare Costs      Labor Force Participation  
Black Inequality      Median Family Income      Home Ownership

12:20 AM · Jun 3, 2016 · Twitter for Android

5.6K Retweets 11.3K Likes

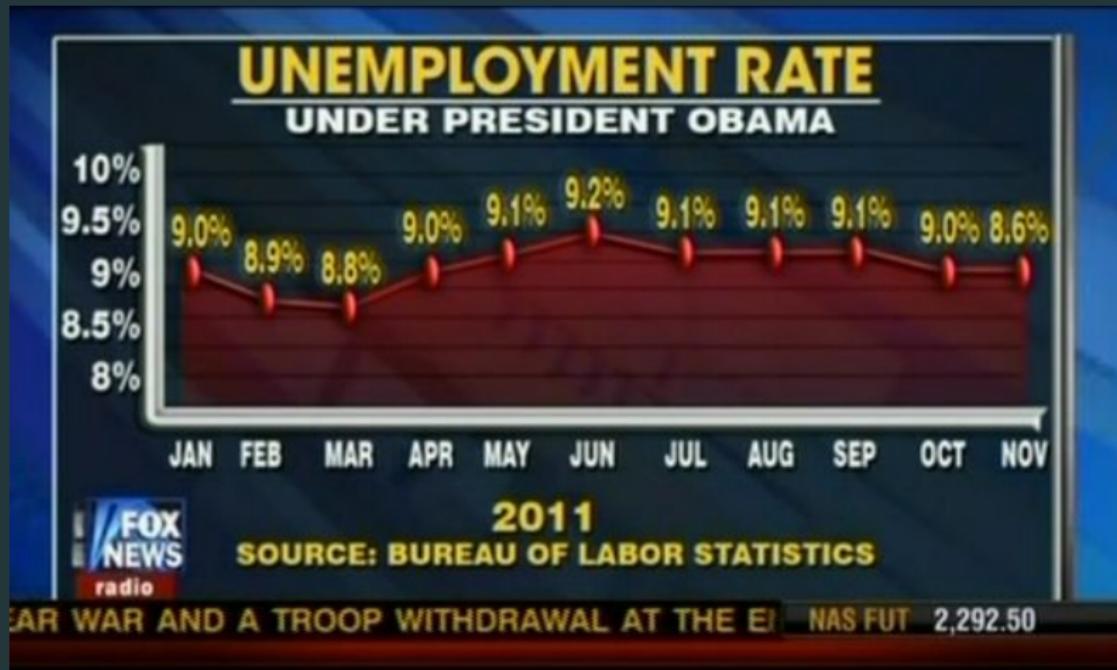
# Toying with perception



# Toying with perception



# "Perfect storm" of a bad plot

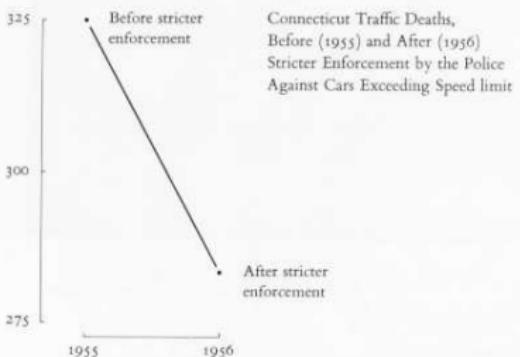


At least 3 things are wrong here

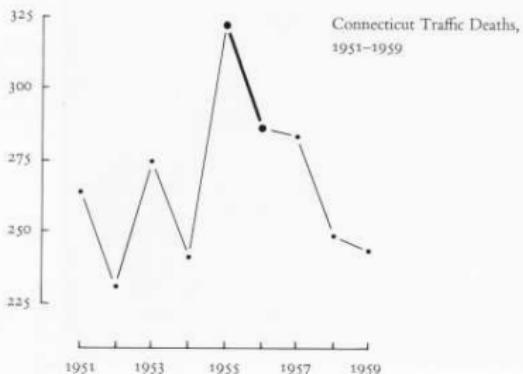
# Trimming a trend

Graphics must not quote data out of context.

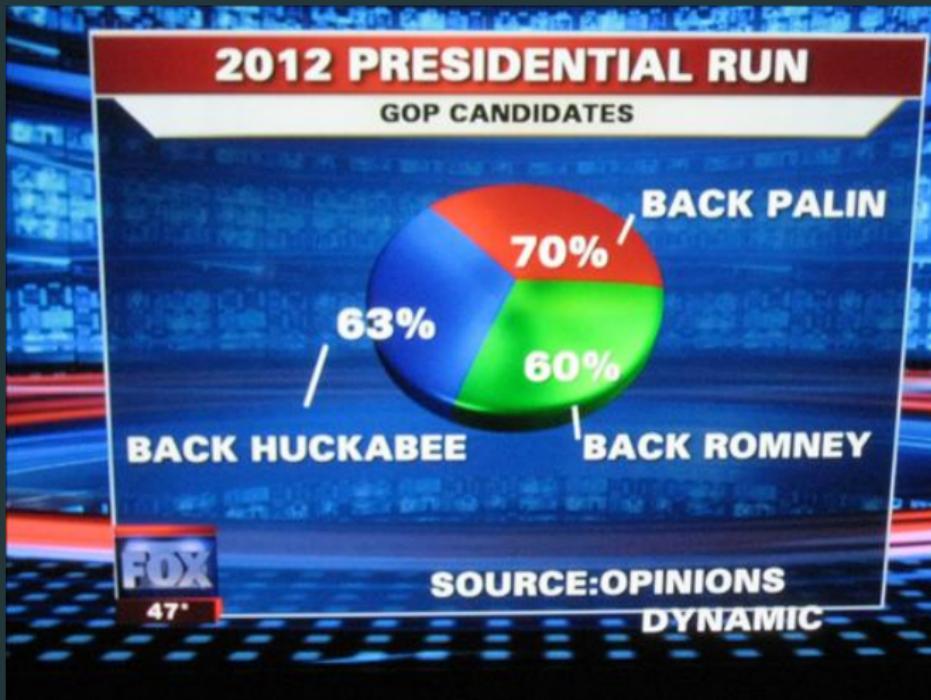
Nearly all the important questions are left unanswered by this display:



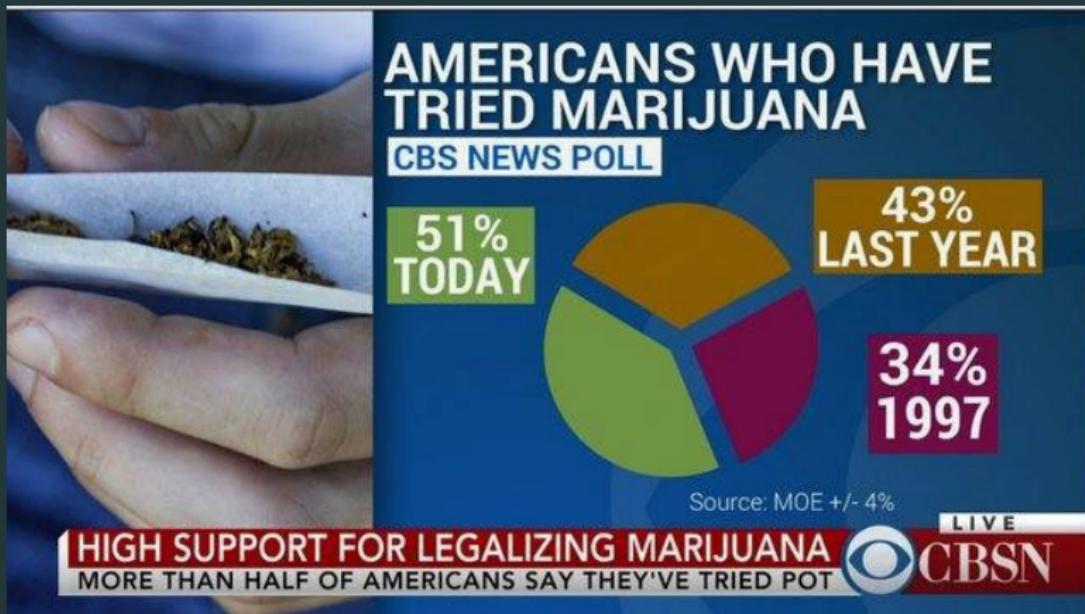
A few more data points add immensely to the account:



# Statistics misreported



# Statistics misreported

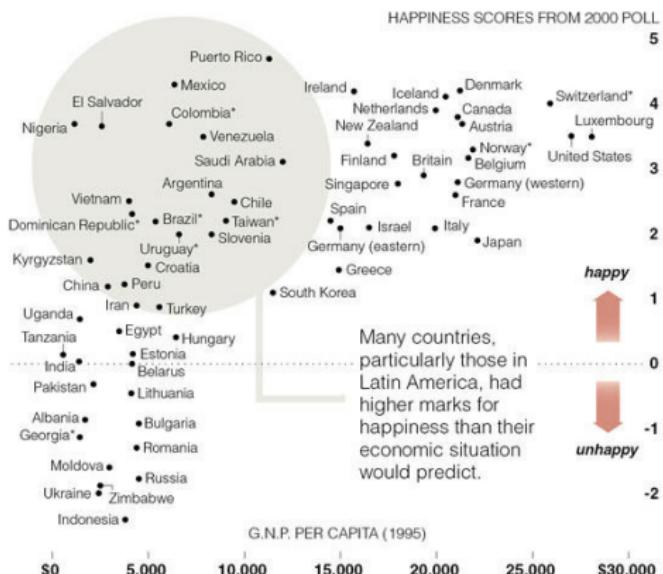


# Statistics misreported

## A Plateau of Happiness

A country's wealth may not always dictate the happiness of its people.

As part of the World Values Survey project, inhabitants of different countries and territories were asked how happy or satisfied they were. Below is a sampling of happiness rankings, along with economic status.



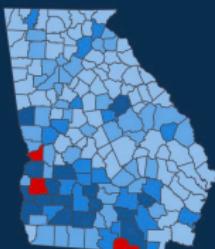
\*Poll results for these countries were from 1995.

Source: Ronald Inglehart, "Human Beliefs and Values : A Cross-Cultural Sourcebook Based on the 1999-2002 Values Surveys"

The puzzle is based on the assumption that the relationship should be linear.

# Statistics misreported

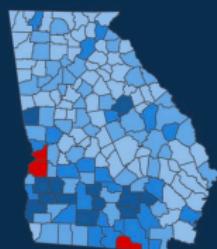
Cases per 100K▼



Cases per 100K

None	1,071 - 1,622
1 - 620	1,623 - 2,960
621 - 1,070	2,961 - 4,661

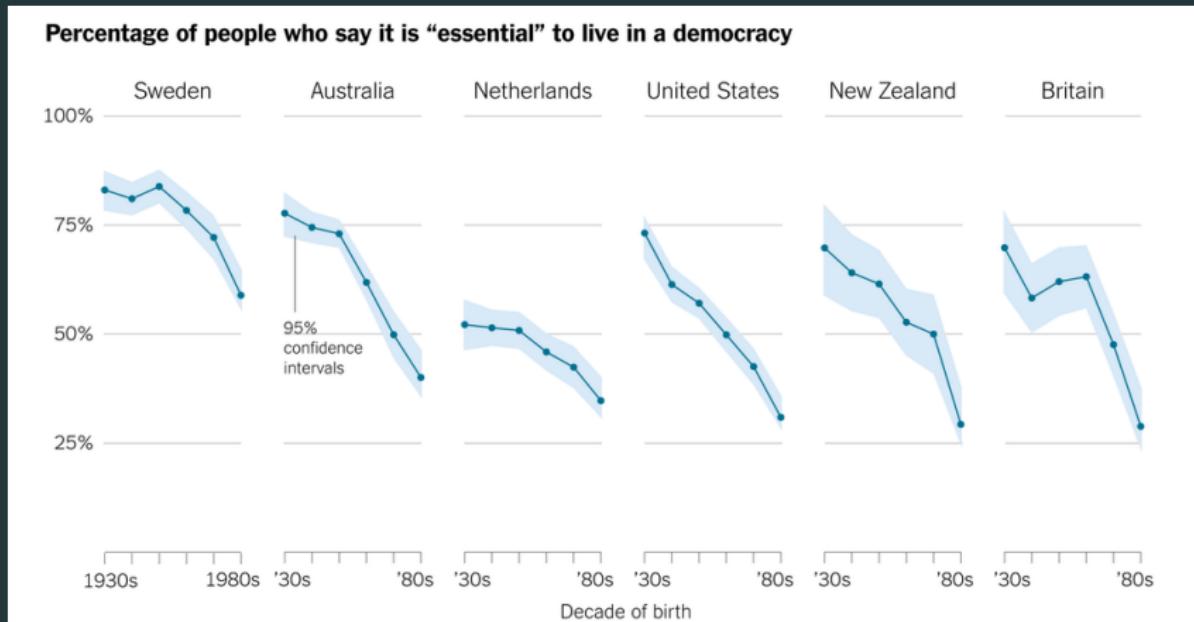
Cases per 100K▼



Cases per 100K

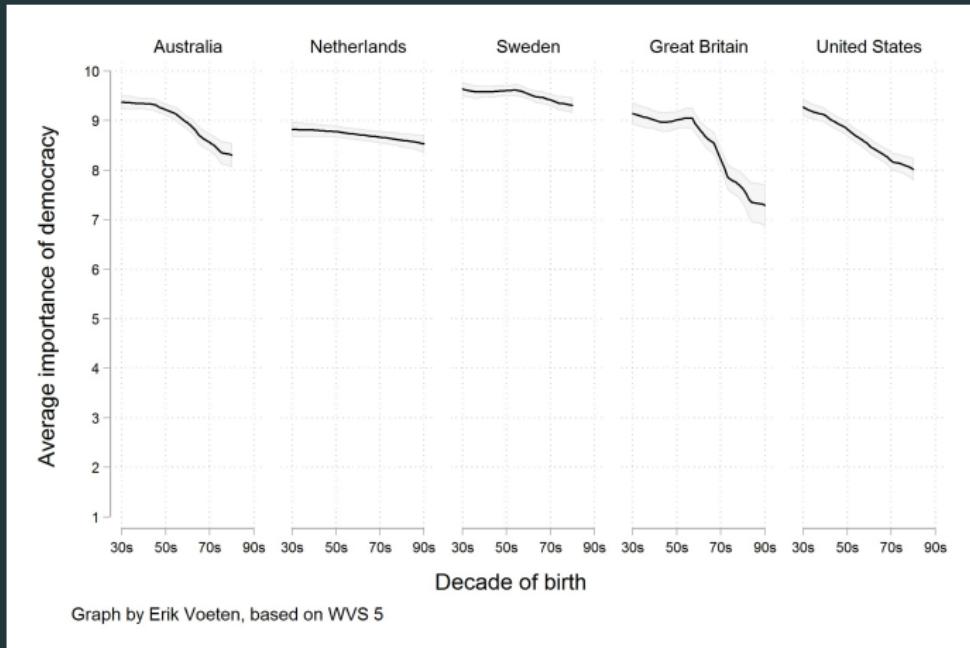
None	1,556 - 2,336
1 - 949	2,337 - 3,768
950 - 1,555	3,769 - 5,165

# Bad data



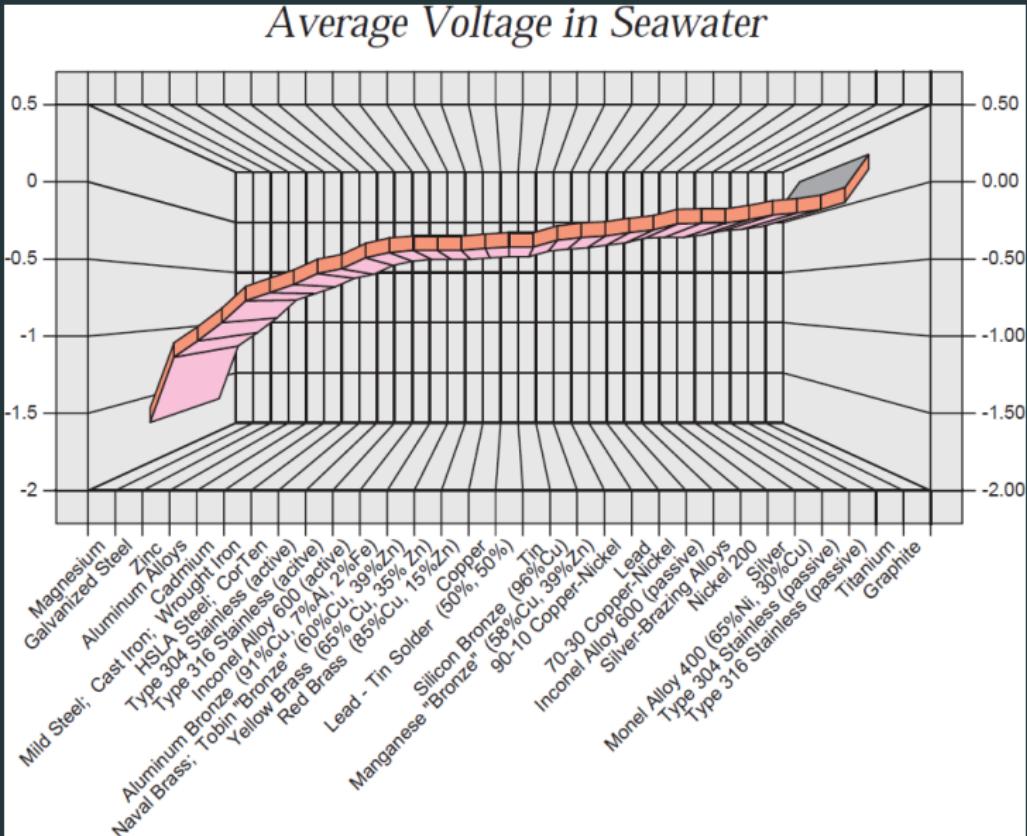
Difference between those who respond “10” and the rest.

# Better data

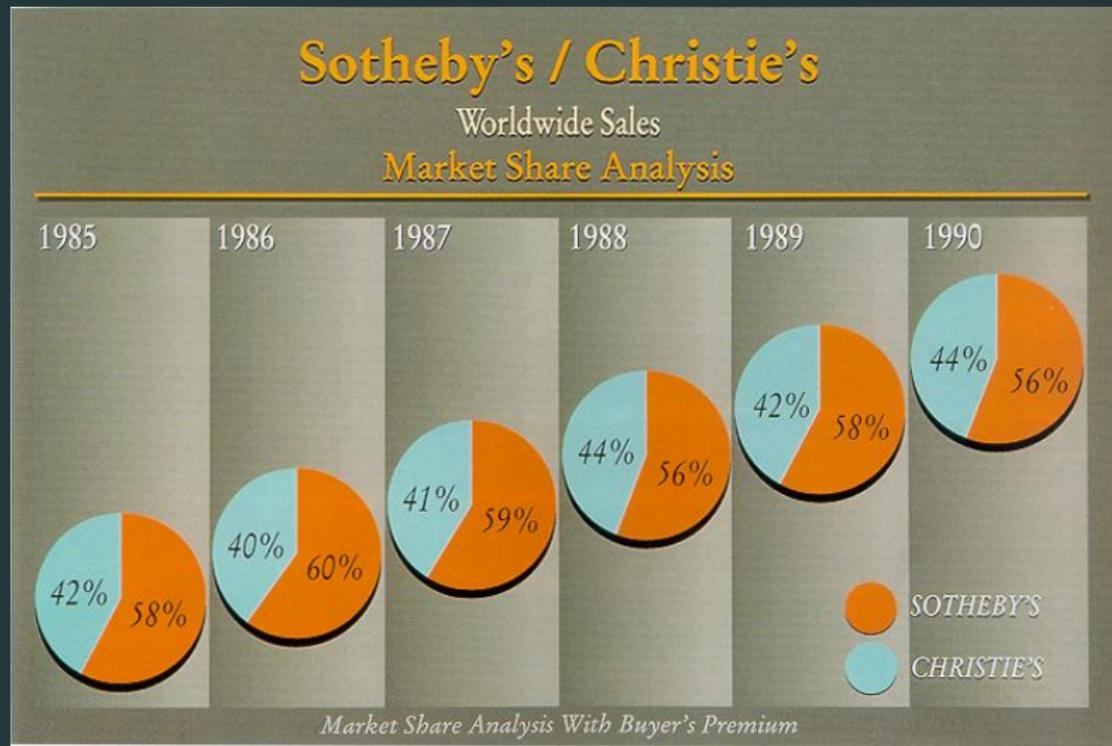


Only 2 countries out of 5 keep their pattern; less concerning than before.

# Poor choice of graphs

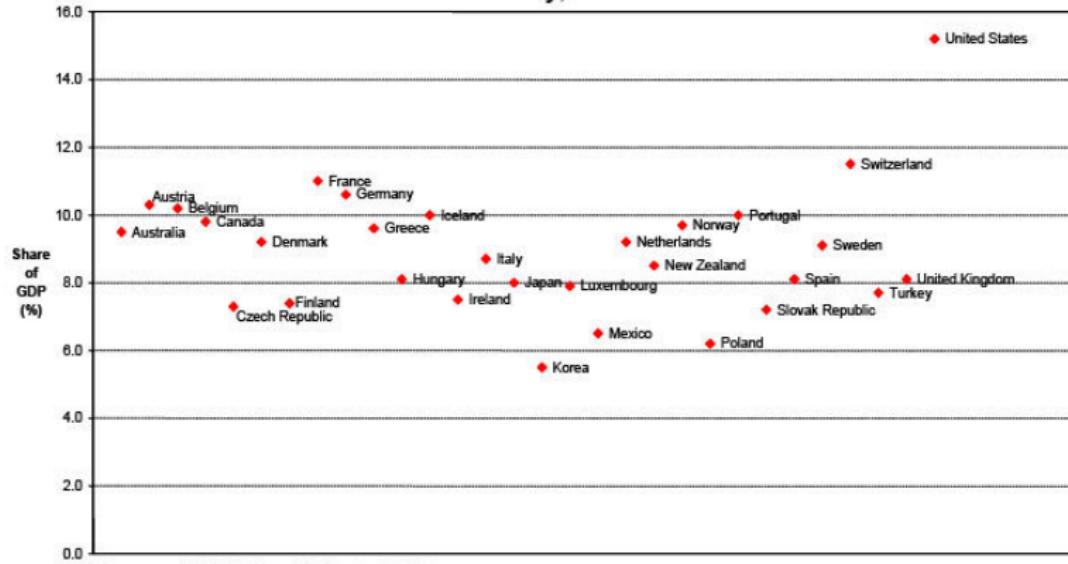


# Poor choice of graphs



# Poor choice of graphs

**Chart 2 - Total Expenditures on Health as a Percentage Share of GDP, by OECD Country, 2004**



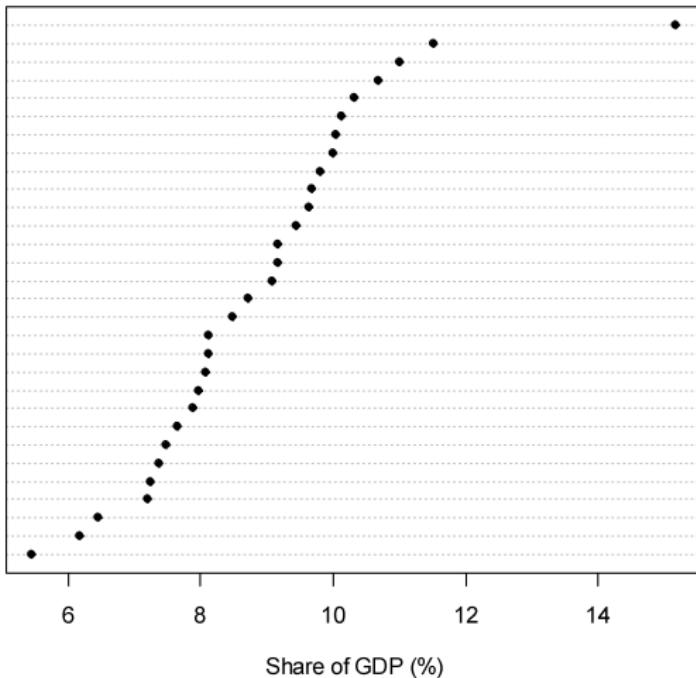
Source: OECD Health Data 2007.

Note: For the United States the 2004 data reported here do not match the 2004 data point for the United States in Chart 1 since the OECD uses a slightly different definition of "total expenditures on health" than that used in the National Health Expenditure Accounts.

# Improved display

Expenditures on Health as Percentage of GDP  
for OECD Countries, 2004

United States  
Switzerland  
France  
Germany  
Austria  
Belgium  
Iceland  
Portugal  
Canada  
Norway  
Greece  
Australia  
Netherlands  
Denmark  
Sweden  
Italy  
New Zealand  
United Kingdom  
Spain  
Hungary  
Japan  
Luxembourg  
Turkey  
Ireland  
Finland  
Czech Republic  
Slovak Republic  
Mexico  
Poland  
Korea



# Conclusion

---

# Conclusion

---

Good data visualization involves thinking about the argument to be made, making choices among alternatives, and taking into consideration issues such as audience, parsimony, integrity.

It will rarely result from canned routines and default options found in statistical packages.

# Conclusion

---

With *ggplot2*, we wouldn't be able to fall prey to some bad practices, like creating chartjunk.

However, there is still potential for other design issues: redundant information, large ink/data ratio, etc.

Thank **you** for the kind  
attention!

# References

- Bartels, L. M. (2008). *Unequal Democracy: The Political Economy of the New Gilded Age*. New York: Russell Sage Foundation.
- Burn, D. A. (1993). Designing Effective Statistical Graphs. In C. R. Rao (Ed.), *Handbook of statistics* (pp. 745–773). Amsterdam: Elsevier.
- Healy, K. (2018). *Data Visualization: A Practical Introduction*. Princeton, NJ: Princeton University Press.
- Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
- Tufte, E. R. (2006). *Beautiful Evidence*. Cheshire, CT: Graphics Press.