Data Visualisation ©

Dr Diarmuid McDonnell & Prof. Vernon Gayle
AQMEN
University of Edinburgh

March 2019

Welcome

Why are we here today?

U.K. Industrial Strategy aims to utilise big data to improve economic performance and increase productivity.

Major barrier is the lack of a suitably trained workforce.

U.K. social science stakeholders (e.g. ESRC, Nuffield) believe this discipline can make a major contribution to Industrial Strategy.

Why this type of training?

"Many organizations can barely find a way to use their R/Python programmers on reasonable datasets."

"The piece missing from the data science movement right now is really simple: intelligent application of data science tools."

https://www.linkedin.com/pulse/data-science-dead-5-years-less-justin-b-dickerson-phd-mba-pstat- accessed 16.07.2018.


Administrative Data Research Centre Scotland
An ESRC Data Investment
Better Knowledge Better Society
Supporting researchers to use administrative data for economic and social research
Enabling studies that benefit the public
Helping inform and improve policy-making
W. adrn.ac.uk
E. adrc-s@ed.ac.uk
AQMeN
Applied Quantitative Methods Network
In partnership with the Fair Institute of Scotland and the Scottish Government

What is the social science contribution?

Data Science ≠ Computer Science

Big Data ≈ Small Data

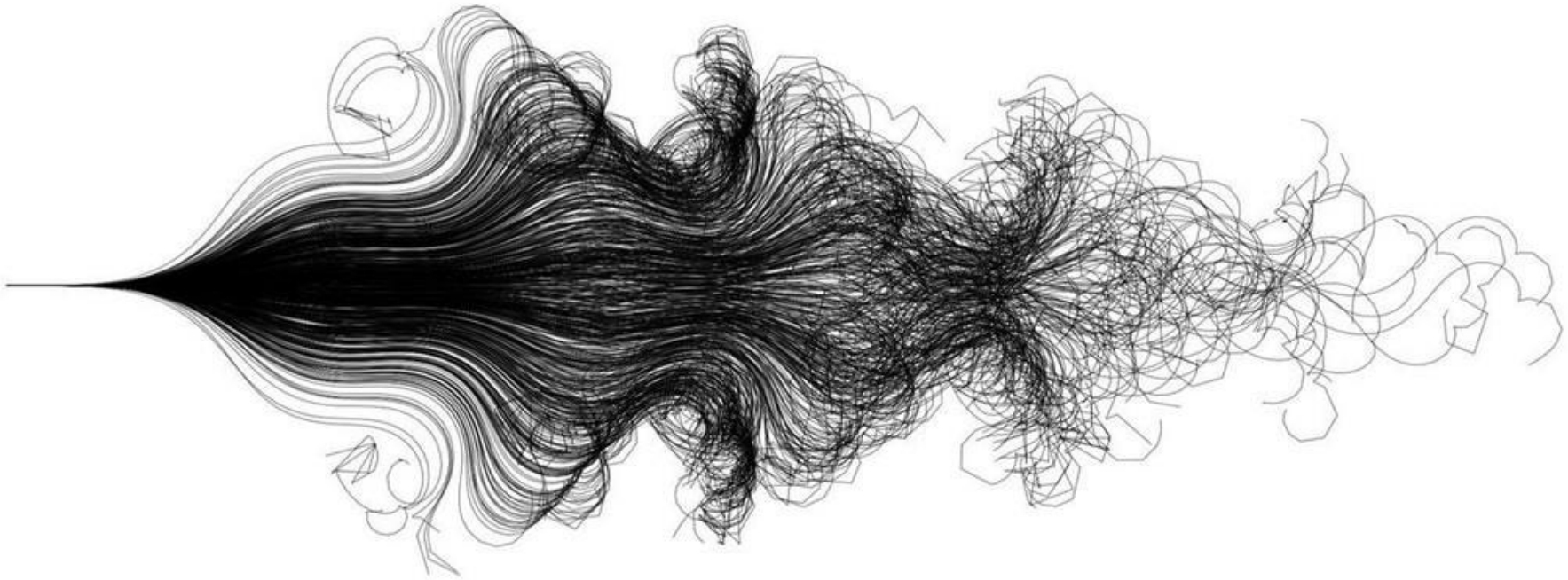
QM Social Scientists = Data Literate

Data Visualisation

In essence

"Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design … [It] consists of complex ideas communicated with clarity, precision, and efficiency. … [It] is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space … [It] is nearly always multivariate … And graphical excellence requires telling the truth about the data. (Tufte, 1983, p. 51)."

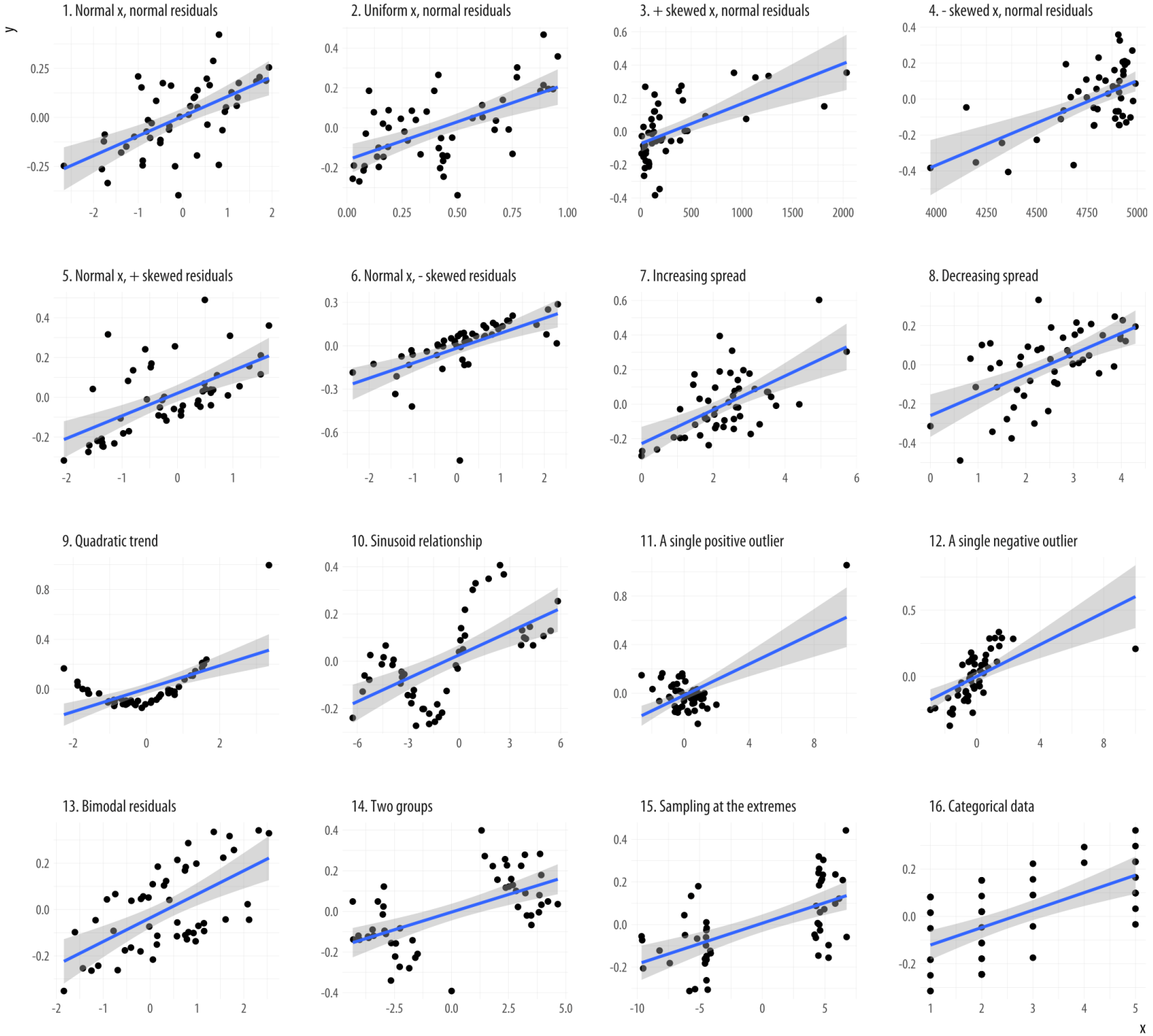
An example



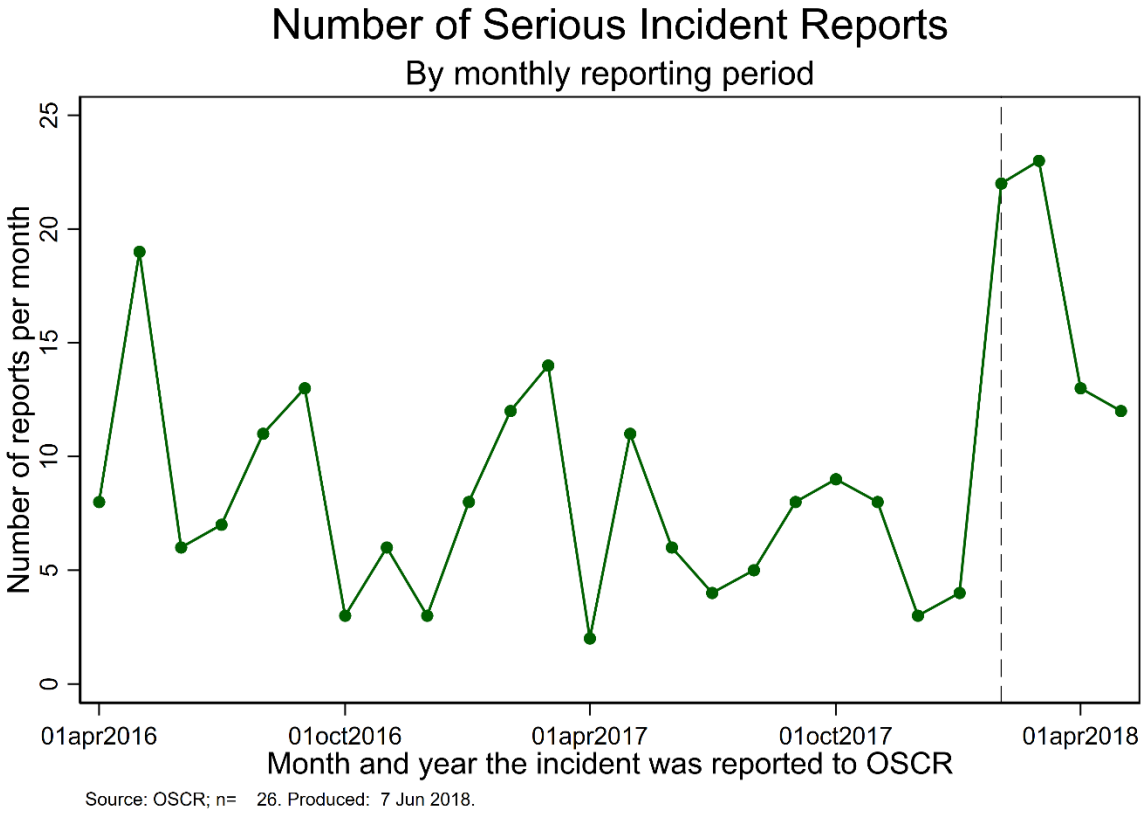
(Cook, 2004)

Why visualise?

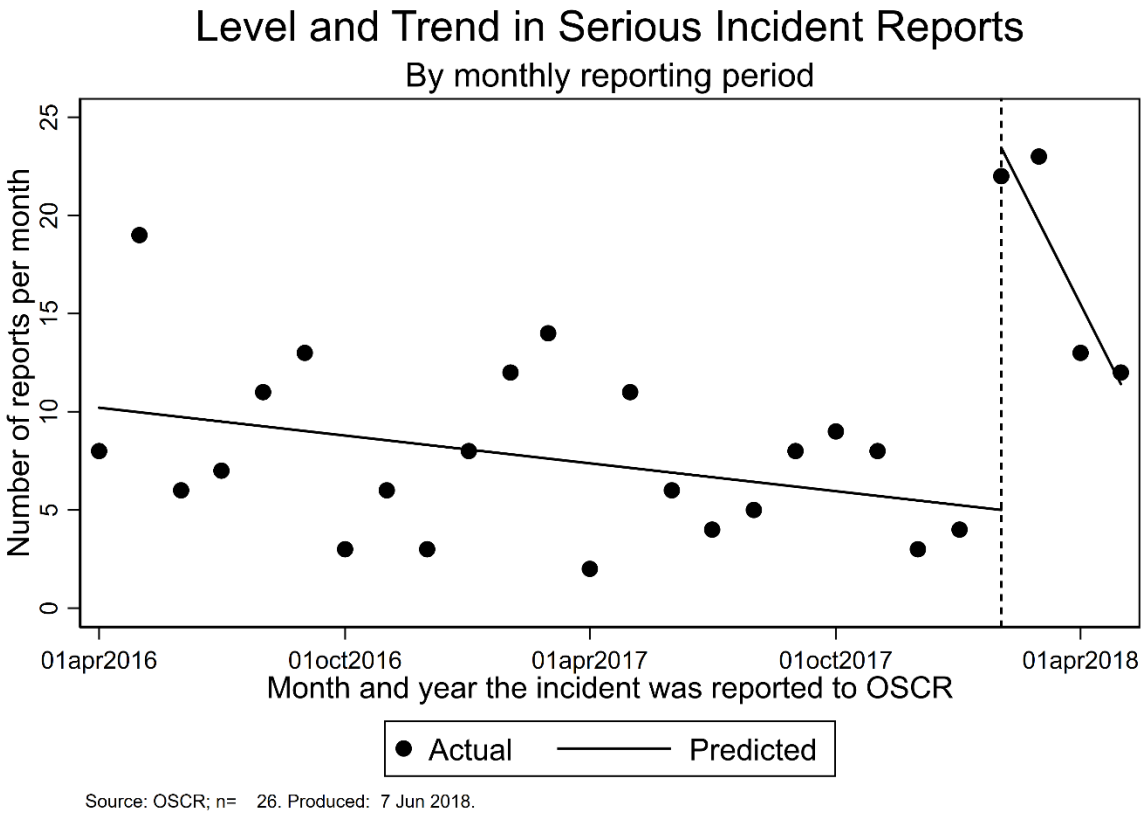
- Summary statistics can hide important differences between distributions.
- Essential aspect of data exploration, and the estimation, reporting and checking of statistical models/analyses.
- It’s the zeitgeist.


1. Normal x, normal residuals
2. Uniform x, normal residuals
3. + skewed x, normal residuals
4. - skewed x, normal residuals
5. Normal x, + skewed residuals
6. Normal x, - skewed residuals
7. Increasing spread
8. Decreasing spread
9. Quadratic trend
10. Sinusoid relationship
11. A single positive outlier
12. A single negative outlier
13. Bimodal residuals
14. Two groups
15. Sampling at the extremes
16. Categorical data
y
x

Exploratory Data Analysis



Model Estimation



Model Estimation

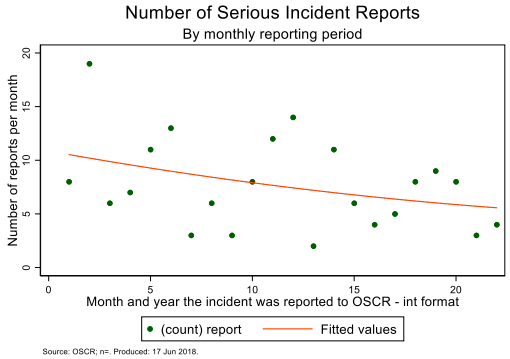
Table 2. Results of regression with Newey-West standard errors

	Coefficient	SE (Newey-West)	95% CI	
			Lower	Upper
Reporting period	-0.01	0.00	-1.69	0.11
Regulatory action	18.45***	2.34	7.89	0.00
Regulatory action*reporting period	-0.13***	0.02	-5.39	0.00
Constant	10.21***	2.06	4.96	0.00
n		26		

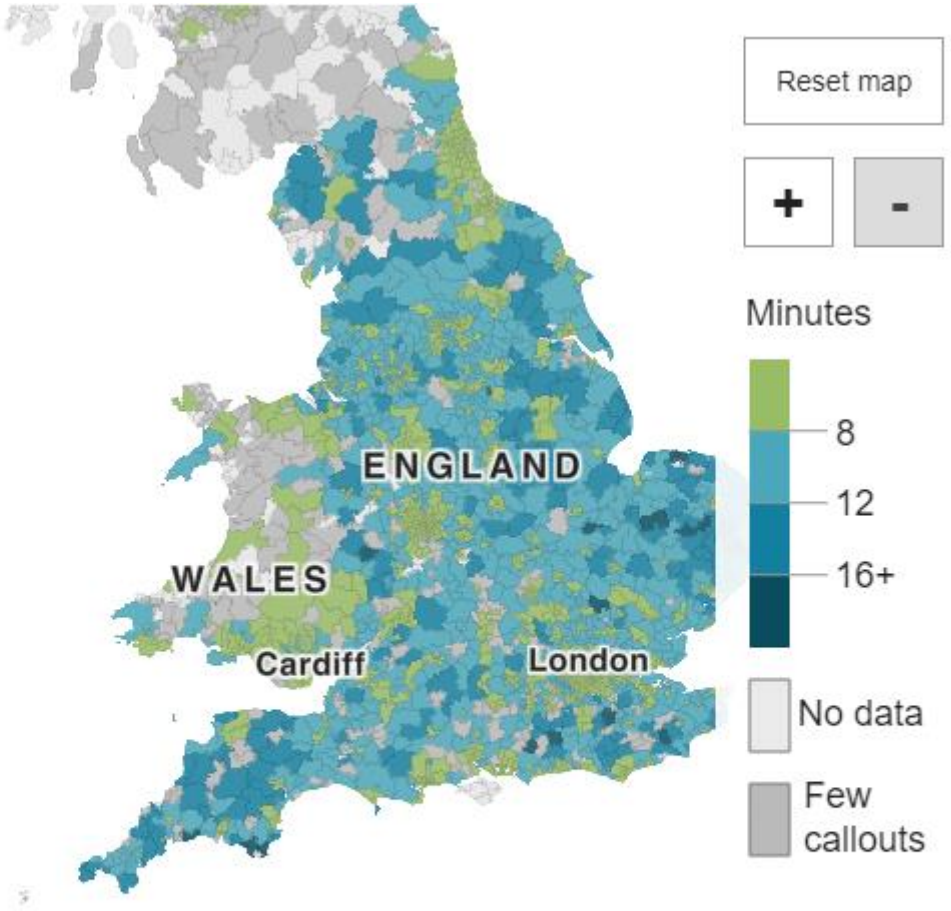
Note: Figures rounded to two decimal places. SE: standard error; CI: confidence interval. *p < .05.

**p < .01. ***p < .001.

Model Checking



Technique du jour



Source: Ambulance trusts. Data is shown for postcode districts with more than nine highest category callouts in January-October 2018. Districts with 10-49 callouts are labelled "low numbers". Northern Ireland does not use comparable categorisation.

Definition

The visual representation and presentation of data to facilitate understanding.

Perceive → Interpret → Comprehend

(Kirk, 2016)

Elements of a visualisation

1. Aesthetic
2. Substantive
3. Perceptual

(Healy, 2019)

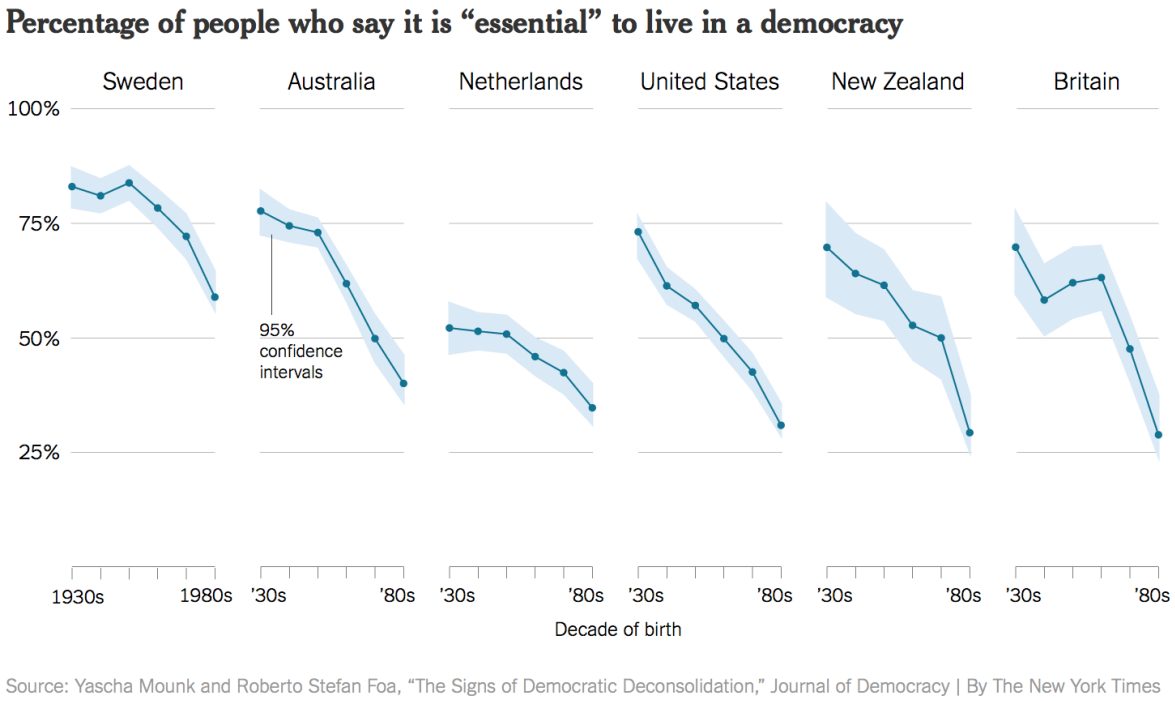
These elements can be corrupted or disregarded in poor quality data visualisations.

Aesthetic

That is, the arrangement of the visualisation, it's layout, titles, design, clarity, parsimony, tastefulness etc.

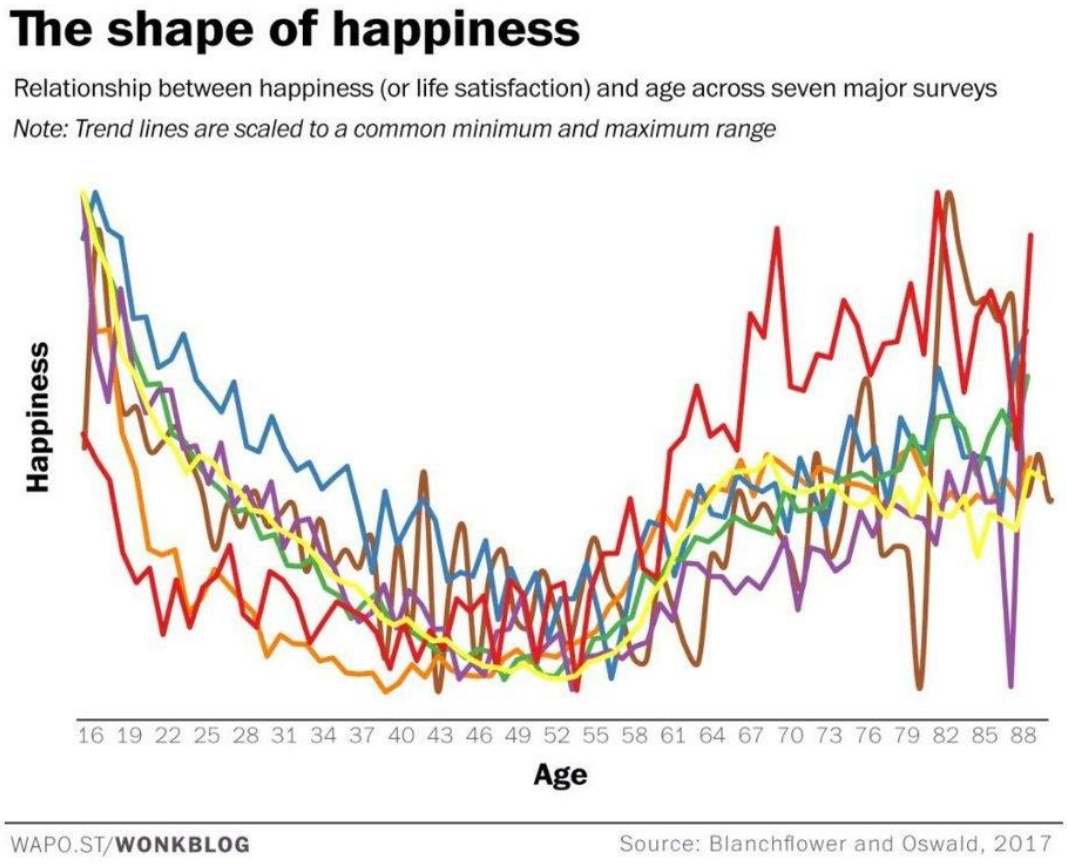
Substantive

That is, the quality of the raw data, the way it was marshalled and analysed, what observations were included and why etc.



Perceptual

That is, the use of colour, shapes, text size, transparency, lines etc.



What makes a good visualisation?

Principles

- Depends on audience and objectives.
- Nothing too fancy or clever.
- Tells a story but not too affecting.
- Doesn’t try to please everyone.

(Speigelhalter, 2016)

What makes a good visualisation?

Practices

- No pie charts.
- A good table counts as visualisation.
- Bar charts are good for conveying the gist of a comparison.
- Use words and numbers.

(Speigelhalter, 2016)

The role of trust

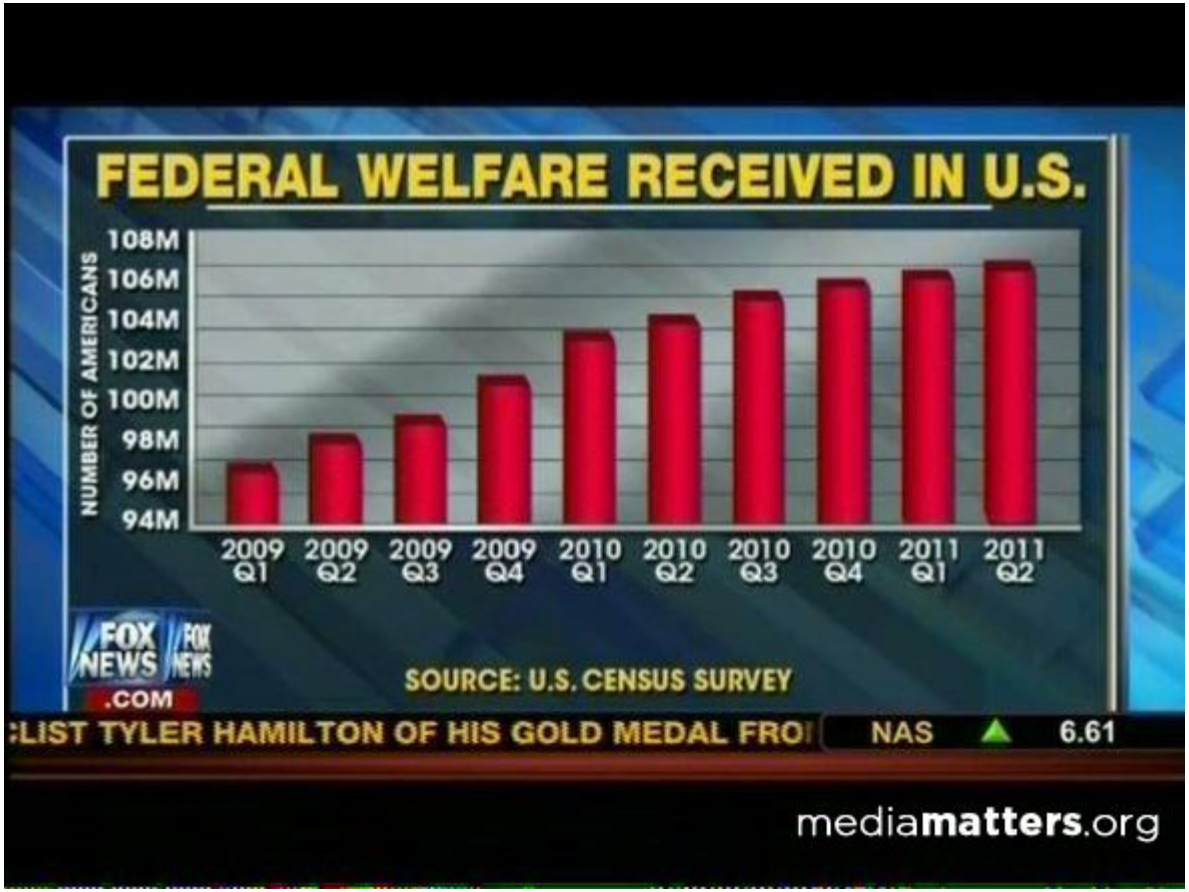
Individuals need to be able to intelligently place their trust in your findings.

Visualisations need to demonstrate trustworthiness with regards to their (and your):

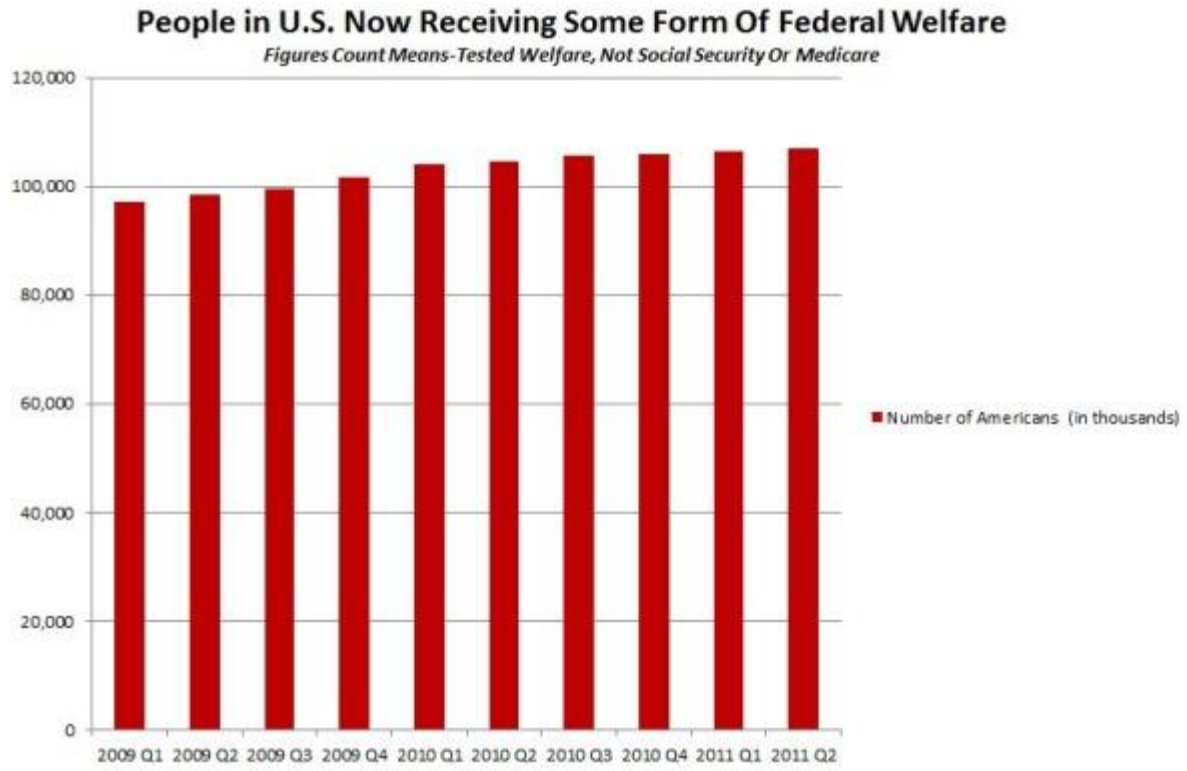
- Honesty
- Reliability
- Competence

(O'Neill, 2012)

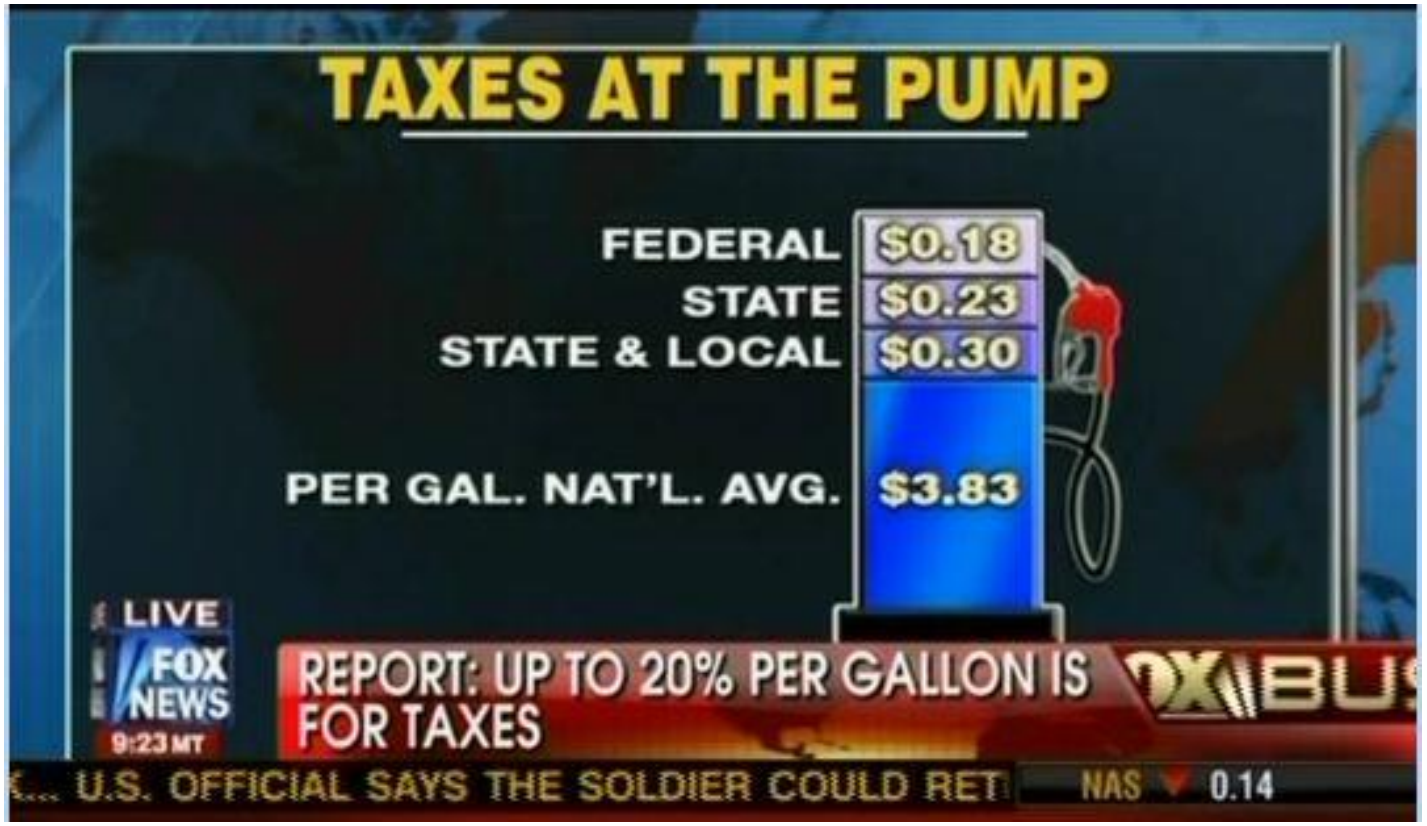
How to be untrustworthy



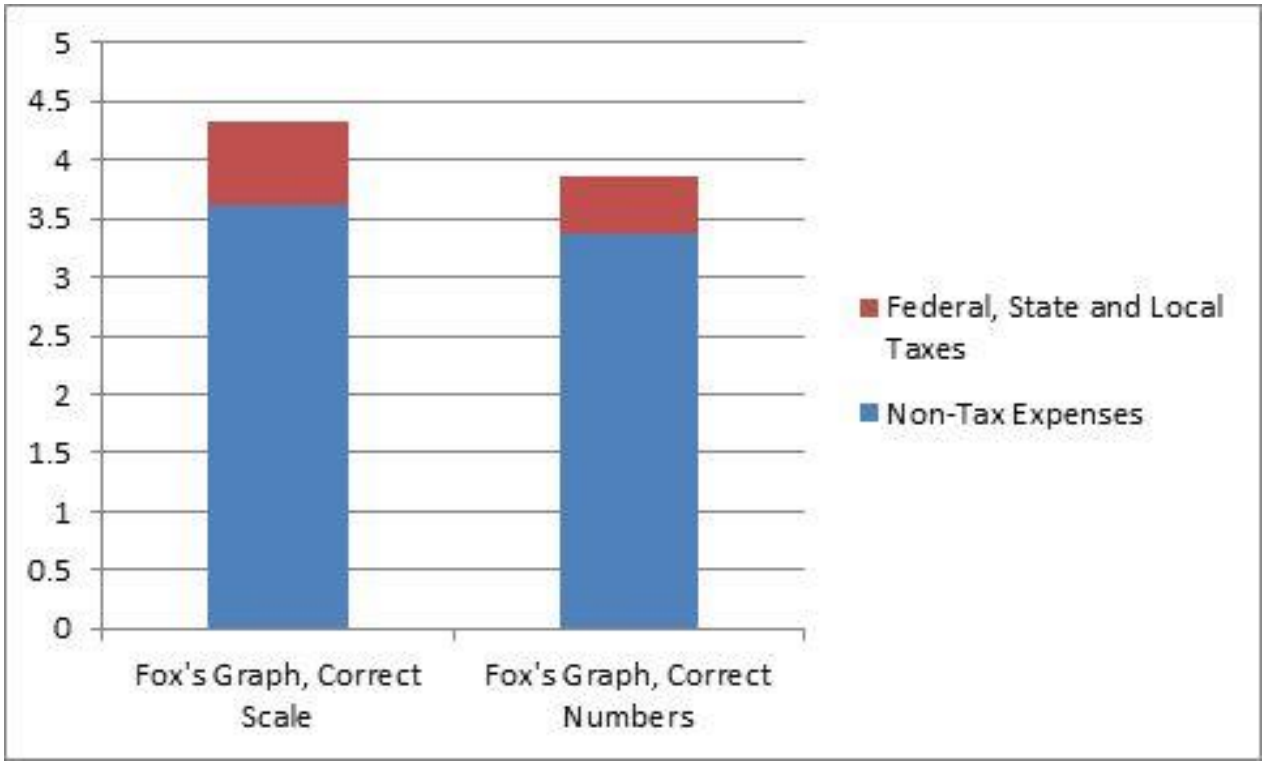
How to be untrustworthy



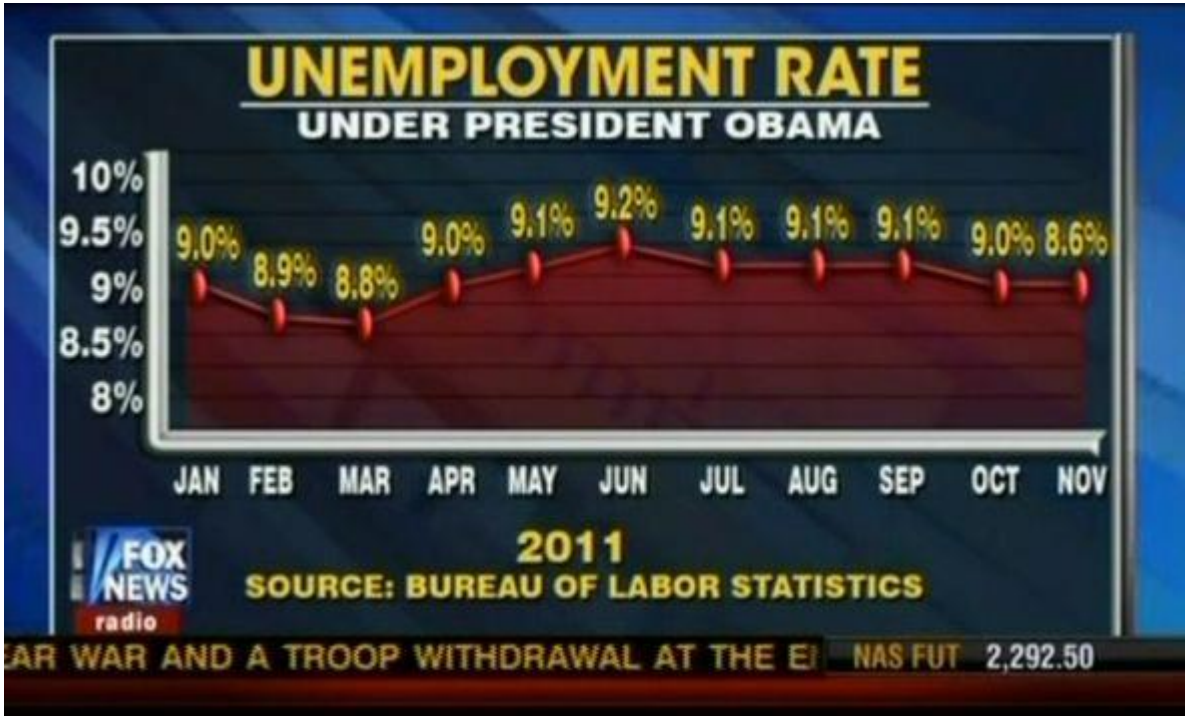
How to be untrustworthy



How to be untrustworthy

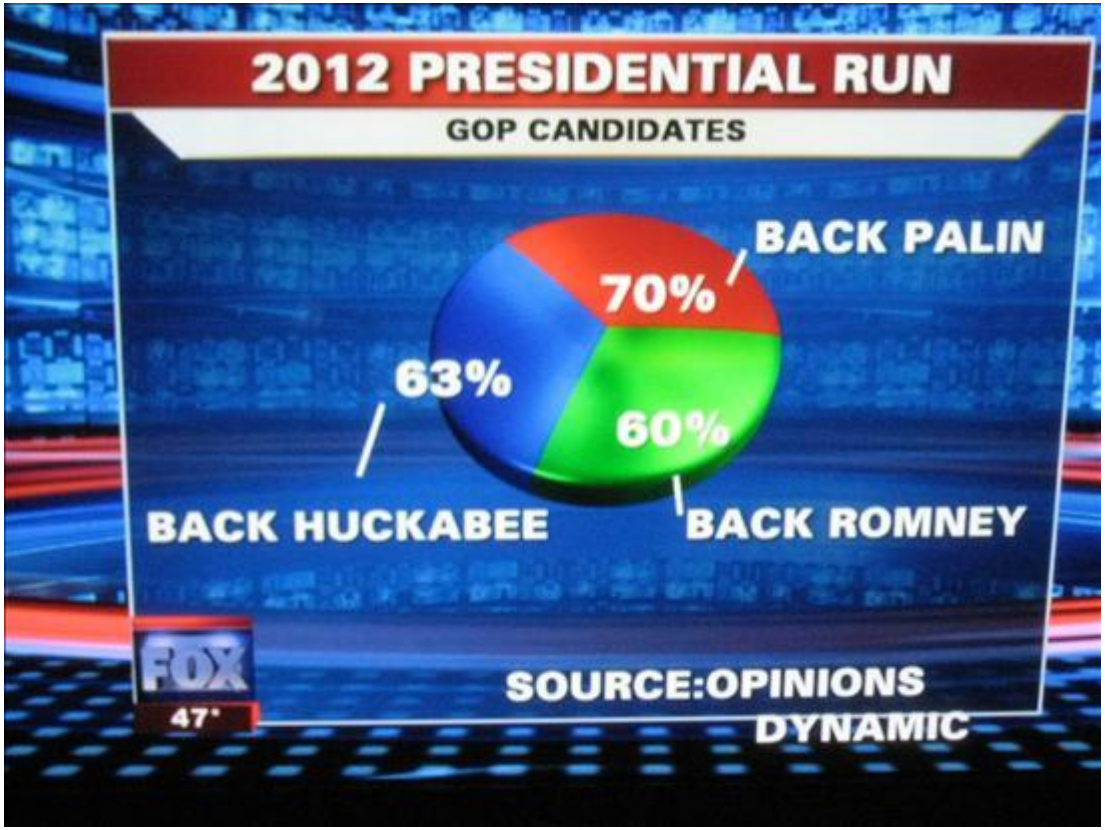


How to be untrustworthy



How to be untrustworthy

How to be untrustworthy



Visualisation done right

https://flowingdata.com/2013/09/25/the-most-unisex-names-in-us-history/

Workshop

R

- Growing in popularity (e.g. data science, statistics, science etc.)
- Popular with statisticians
- Free (open source)
- Difficult to learn
- Development and support is not commercial
- Help resources are under-developed

Programme

Mix of talks and self-directed practical activities.

Data Visualisation challenge (“Hackathon”).

Tutor and peer support.

Use of a variety of data sets, especially messy administrative records.

Top tips

1. Ask plenty of questions.
2. Take your time.
3. Complete as many of the tasks and exercises, and answer as many of the questions as you can.
4. Annotate your work.
5. Be positive.

Estimating Work Time…

Good Luck

Our aim is to equip you, as rapidly and painlessly as possible, with a proficiency in data visualisation using R.

We think it is an ambitious yet achievable goal.

Them: “Are you any good at data visualisation?”

You: “Yes, yes I am.”