Topic 4: Descriptive statistics & visualization

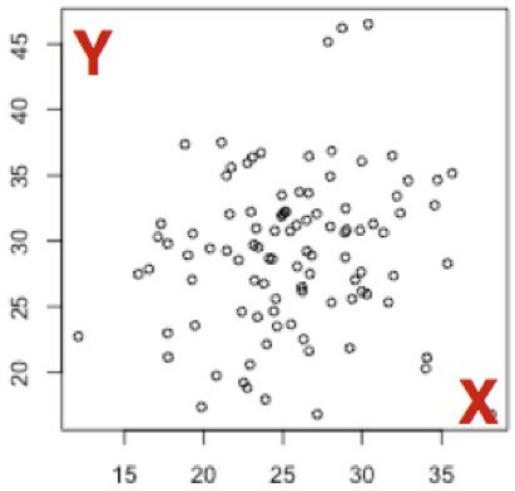
Lectures 8 & 9

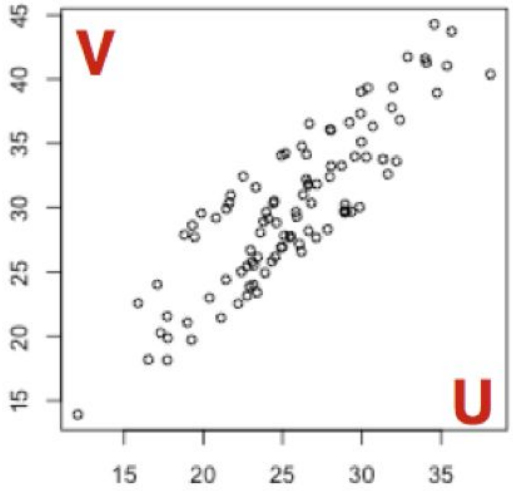
- Descriptive statistics
- Spurious correlations
- Visualization challenges

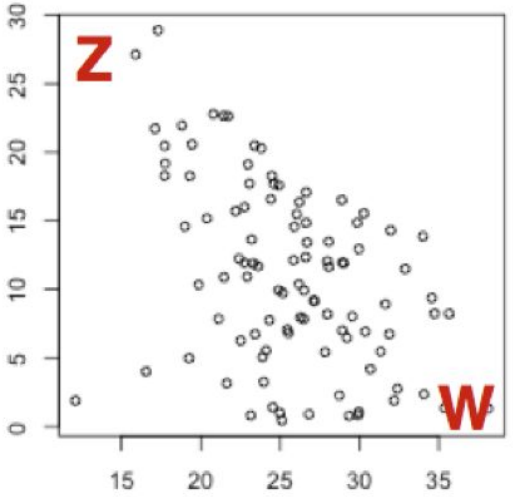
Calculating correlation

Variables → Attributes / Features

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68







Correlation coefficient

Pearson Correlation Coefficient

- Measures 'linear' relationship between variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is the sample size
- x_i, y_i are the single samples indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Wikipedia

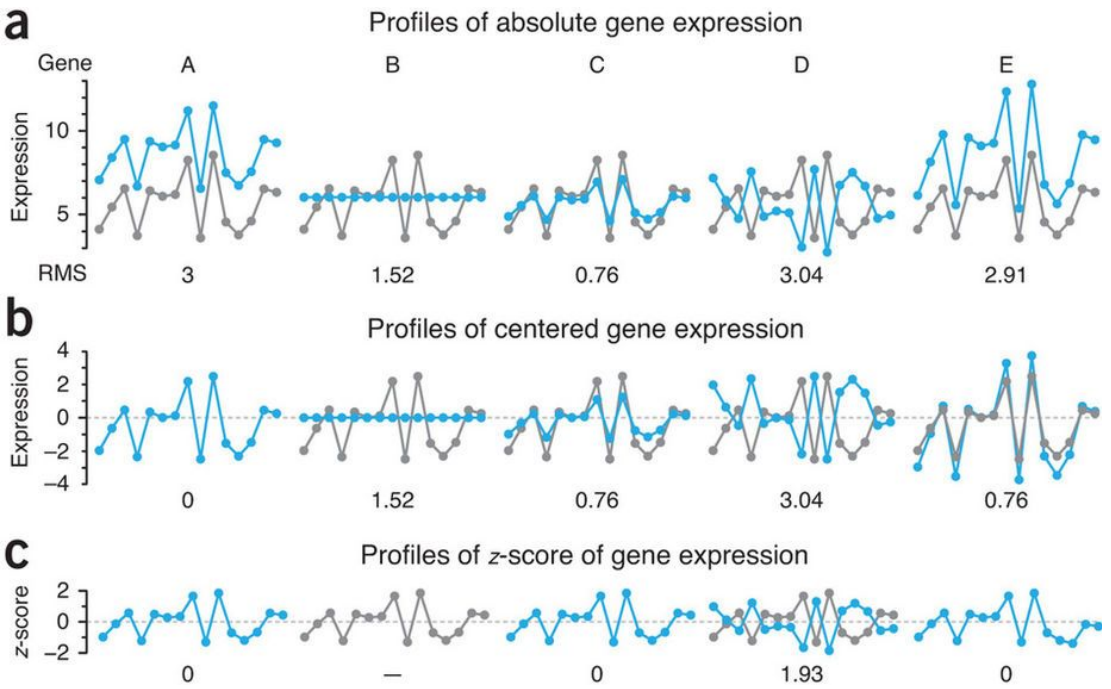
Correlation coefficient

Pearson Correlation Coefficient

- Captures the relationship between 2 vectors after centering each vector by its mean and scaling by its standard deviation.

- The final quantities for each vector are called z-scores.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$



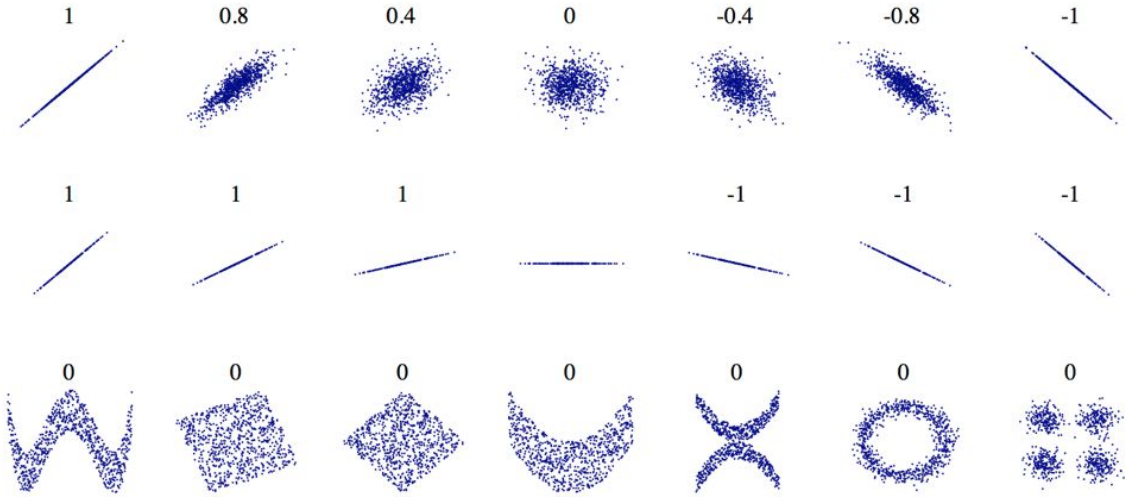
Altman & Krzywinski (2017) Nat. Meth.

Correlation coefficient

Pearson Correlation Coefficient

- Measures ‘linear’ relationship between variables.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$



$-1 \leq r \leq +1$

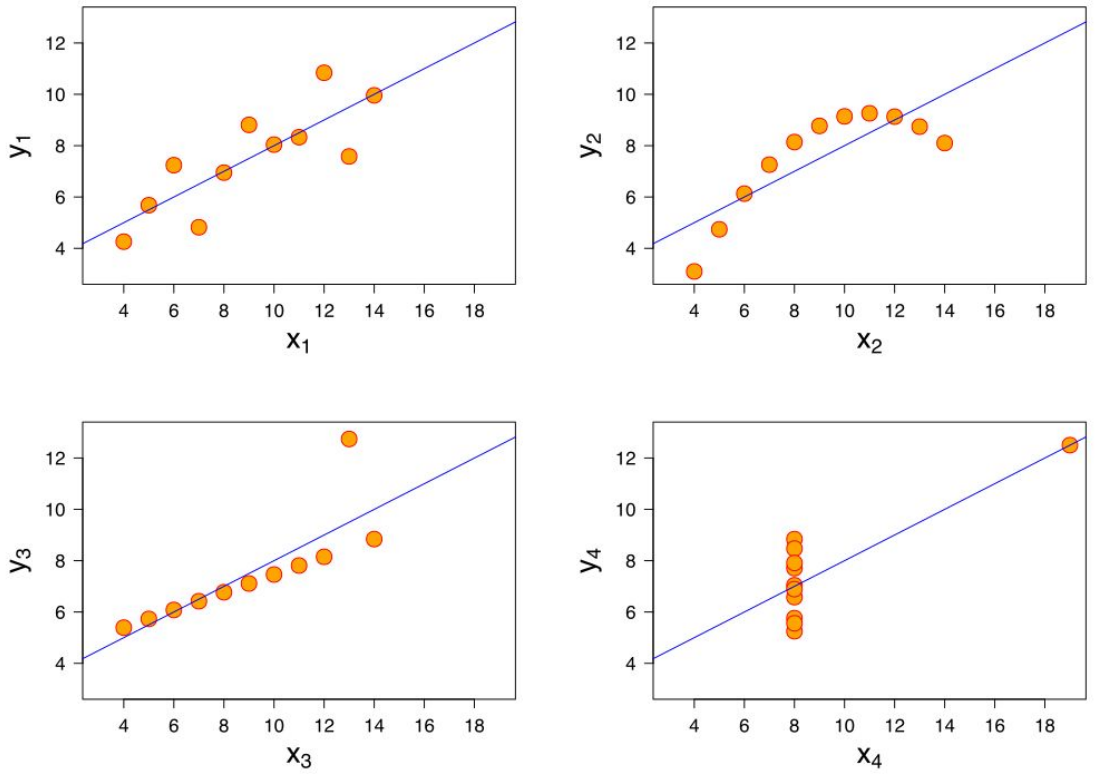
−1 is total −ve correlation | 0 is no correlation | +1 is total +ve correlation

Wikipedia

Anscombe's quartet: "calculation are exact; graphs are rough!"

11 datapoints

- Mean (x) = 9
- Var (x) = 11
- Mean (y) = 7.50
- Var (y) ~ 4.12
- Cor (x, y) = 0.816
- Linear regression line:
  - y = 3.00 + 0.500x

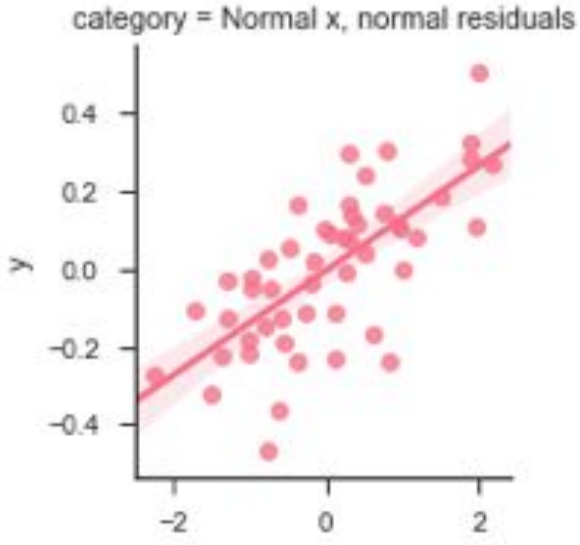


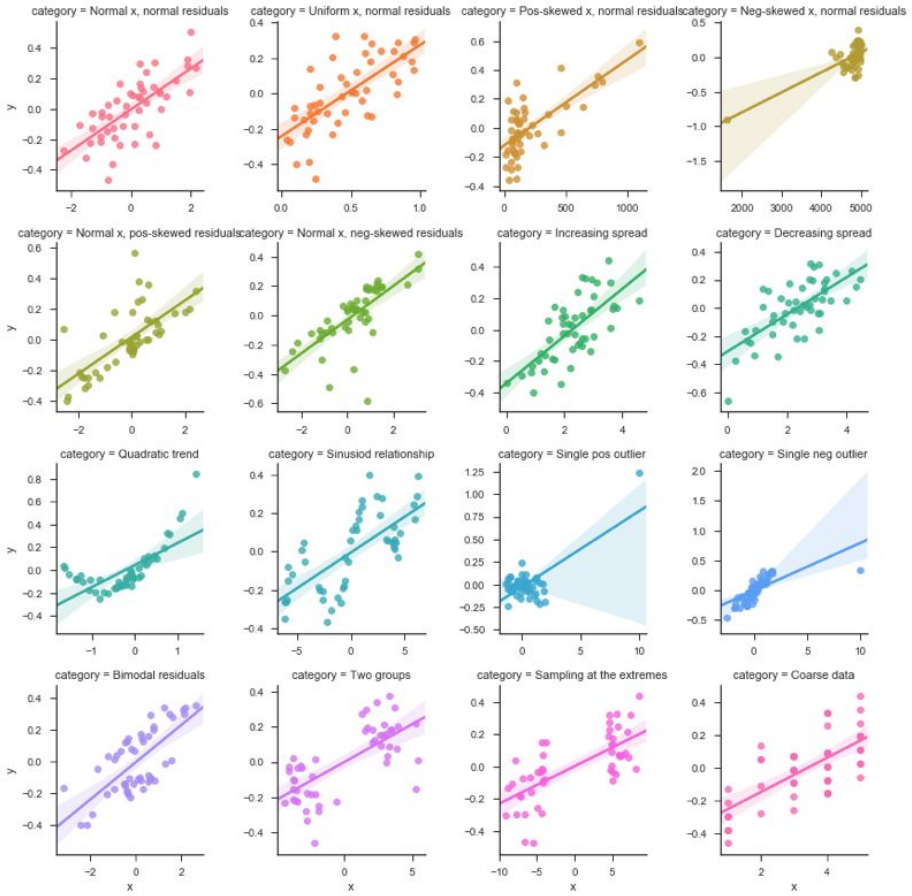
Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.

Wikipedia

What does a correlation coefficient tell you about the data?

Correlation = 0.7

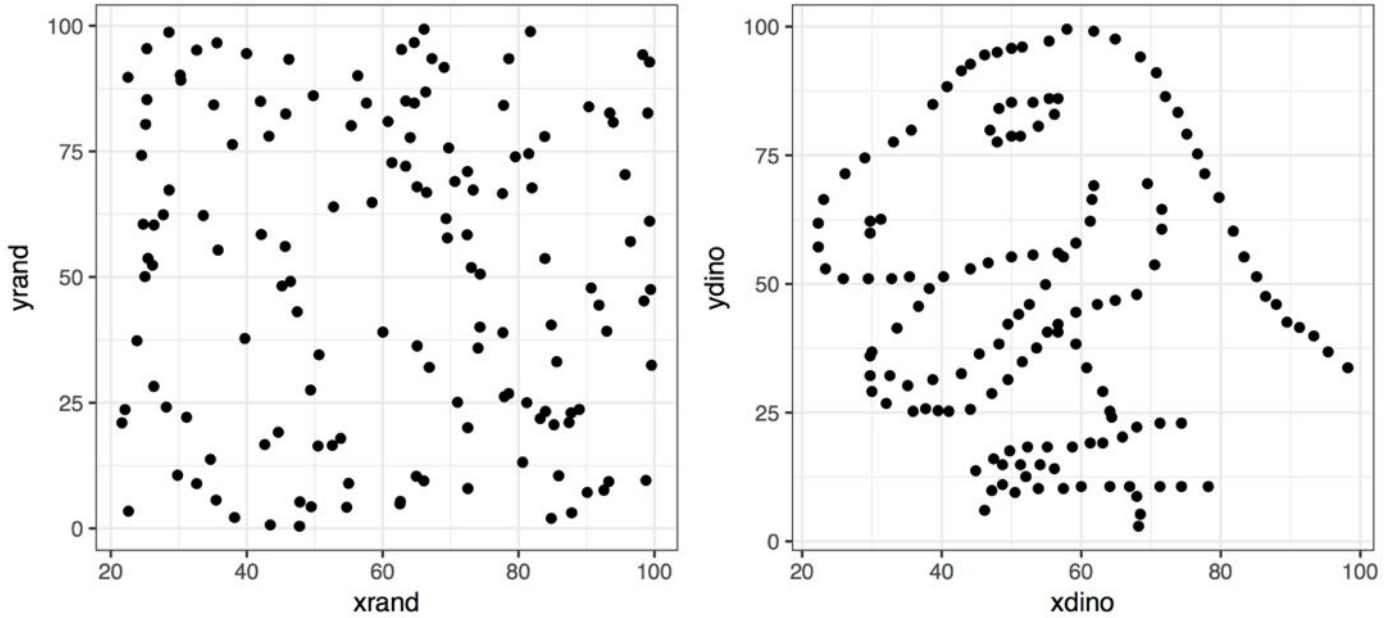




Wikipedia

What does a correlation coefficient tell you about the data?

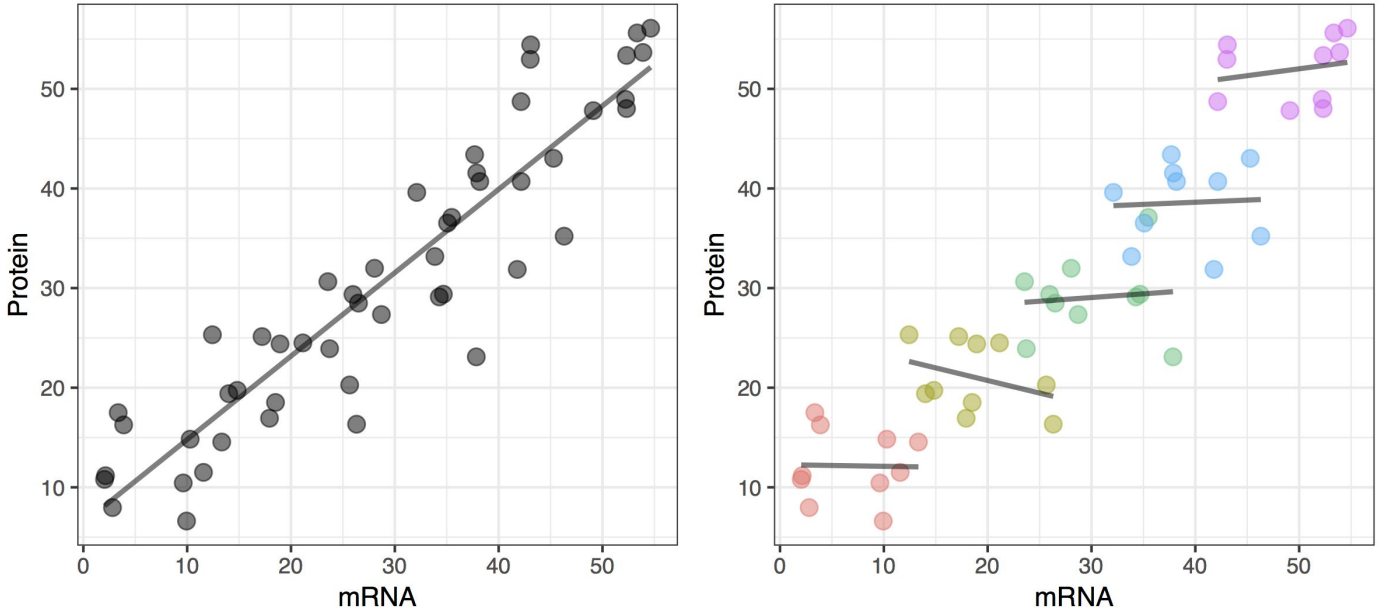
Correlation = -0.06



https://www.autodeskresearch.com/publications/samestats

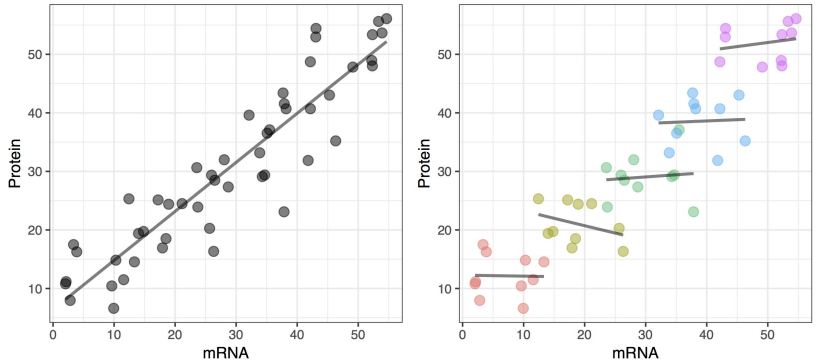
What does a correlation coefficient tell you about the data?

Simpson's Paradox



What does a correlation coefficient tell you about the data?

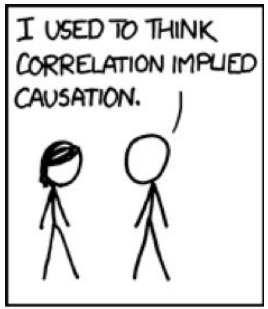
Simpson's Paradox

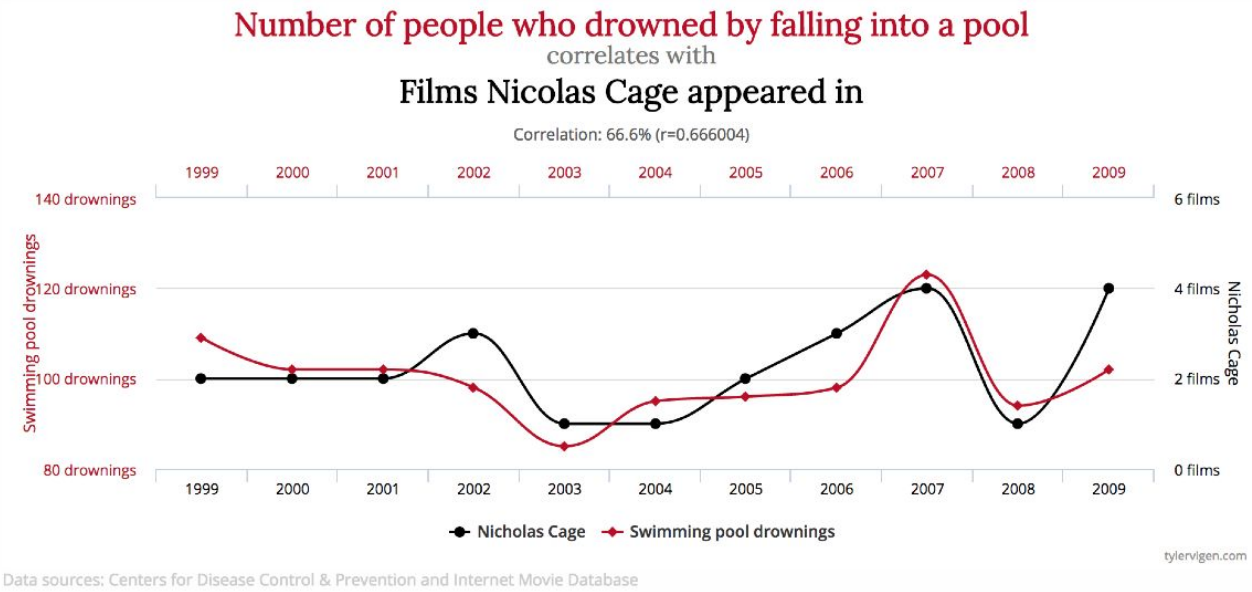


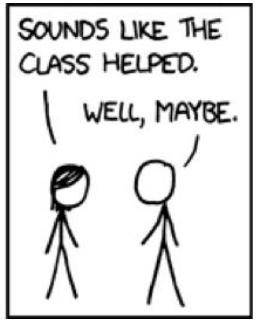
Success rates of kidney stone removal surgeries			
Treatment	**Diameter < 2 cm**	**Dia. ≥ 2 cm**	**Overall**
Open surgery	93%	73%	78%
Percutaneous nephrolithotomy	87%	69%	83%

Spurious correlations

What does Nicholas Cage have to do with people drowning in swimming pools?







Checkout https://www.google.com/trends/correlate

xkcd.com/552

Spurious correlations

Simulate fluctuations in correlation coefficients

- Repeat 10,000: Calculate correlation coefficients of $n = 10$ samples of two independent normally distributed variables ($\mu = 0$, $\sigma = 1$). Plot a histogram.
- Mark statistically significant coefficients ($\alpha = 0.05$).
- Plot the samples with the three largest and smallest correlation coefficients (statistically significant).
- Vary $\sigma = \{0.1, 0.5, 1.0\}$ and vary sample size $n = \{5, 10, 50\}$.

Many correlation/distance measures

Pearson Correlation Coefficient

Spearman Rank Correlation

Euclidean Distance

Mutual Information

...

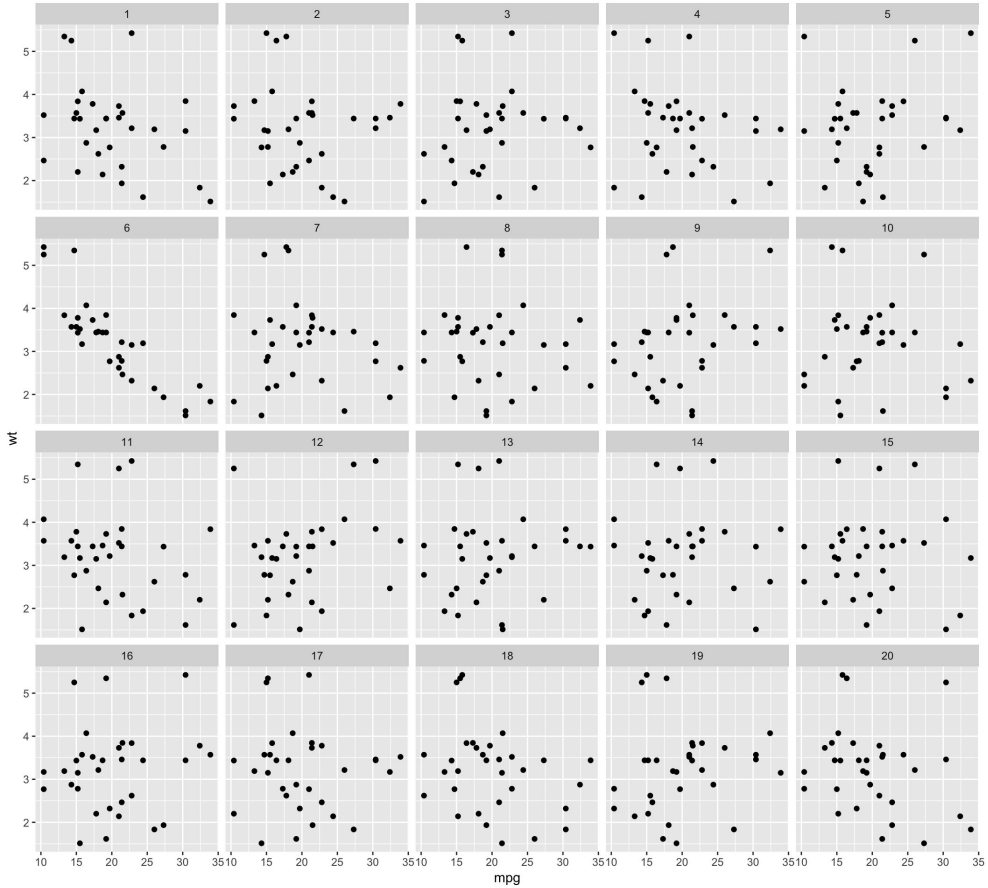
$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} [rank(x_i) - rank(y_i)]}{n(n^2 - 1)}$$

Wikipedia

Spurious correlations – But it *looks* associated!

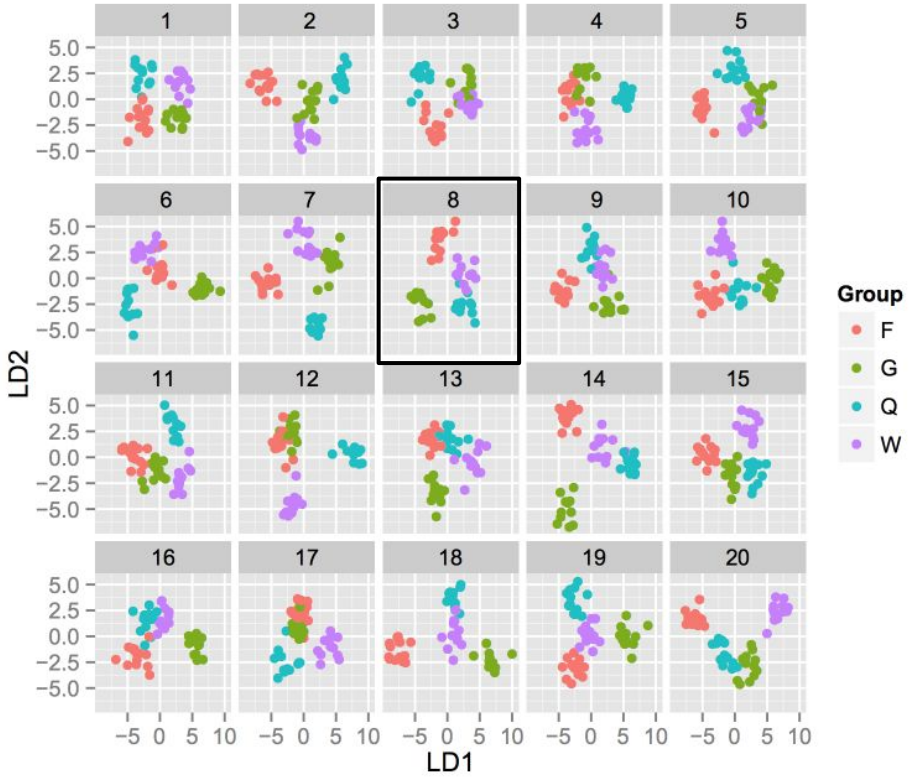


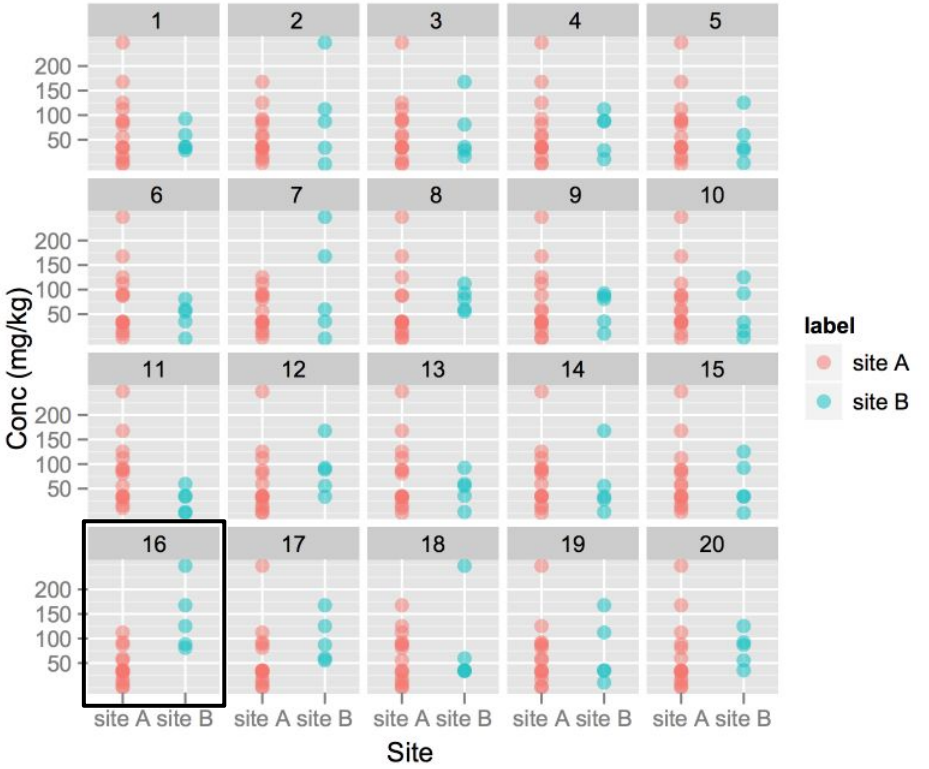
Create a lineup for visual inference

- Place the plot of the real data amongst a set of null plots to create a lineup; Null plots are generated in a way consistent with the null hypothesis.
- If the observer can pick the real data as different from the others, this puts weight on the statistical significance of the structure in the plot.

https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=4995&context=etd

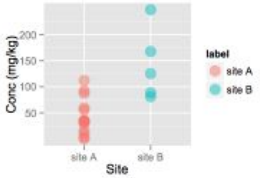
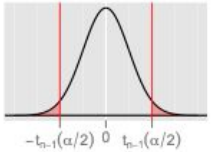
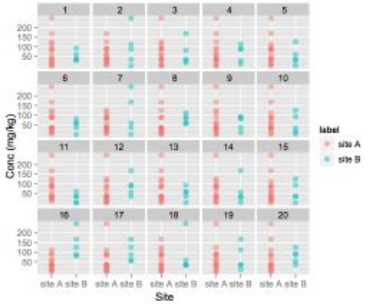
Spurious correlations – But it *looks* associated!



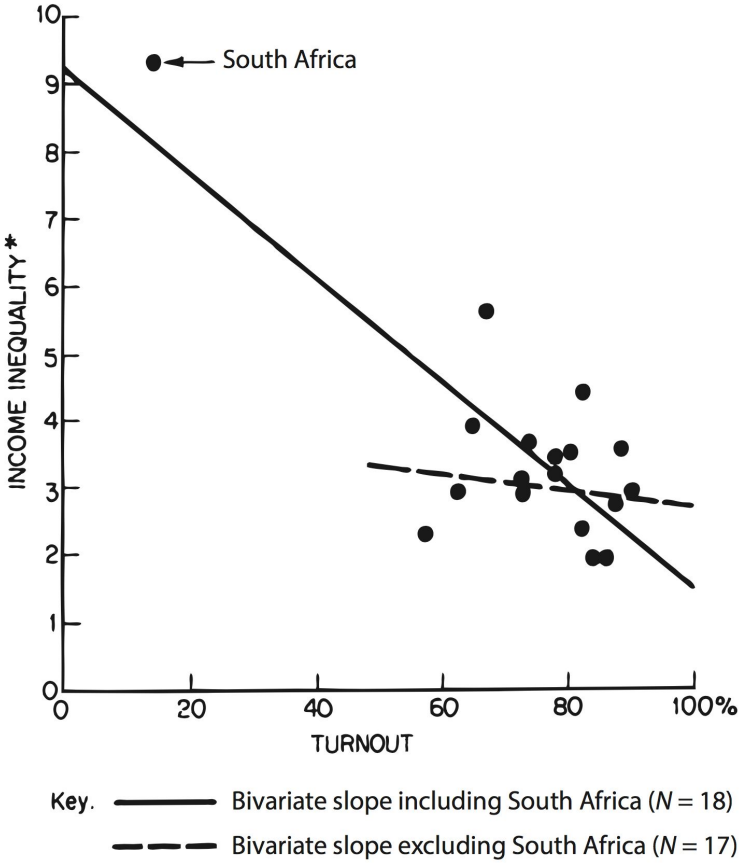


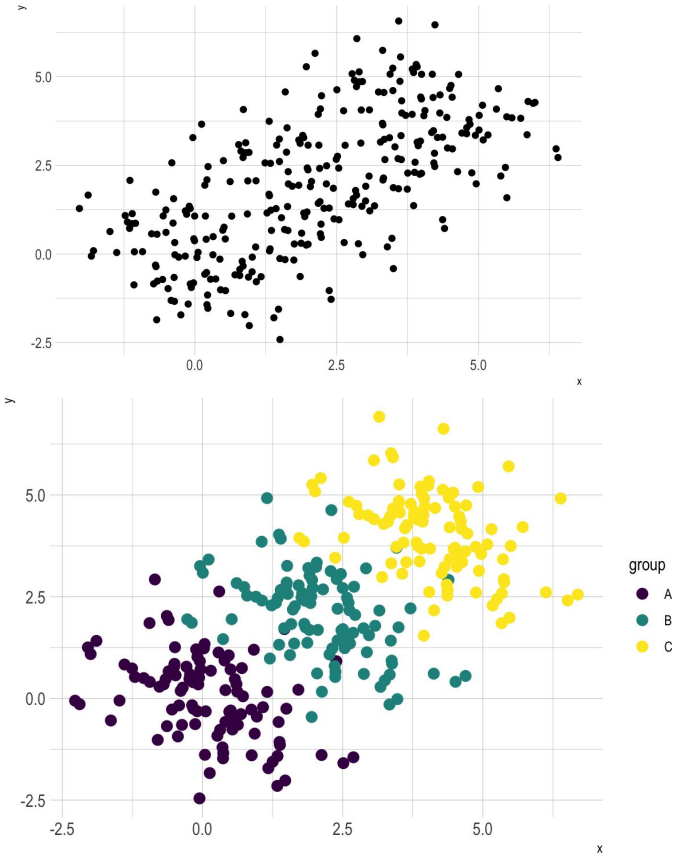
https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=4995&context=etd

Spurious correlations – But it *looks* associated!

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T(y) =$ 
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
Reject H_0 if	observed T is extreme	observed plot is identifiable

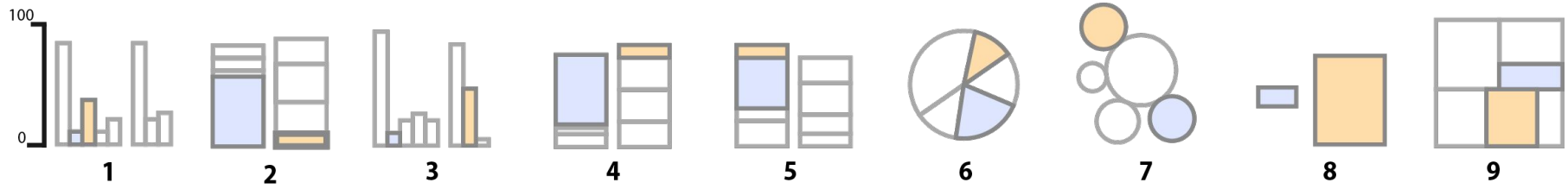
Visualization challenges – Don't do statistics without visualization

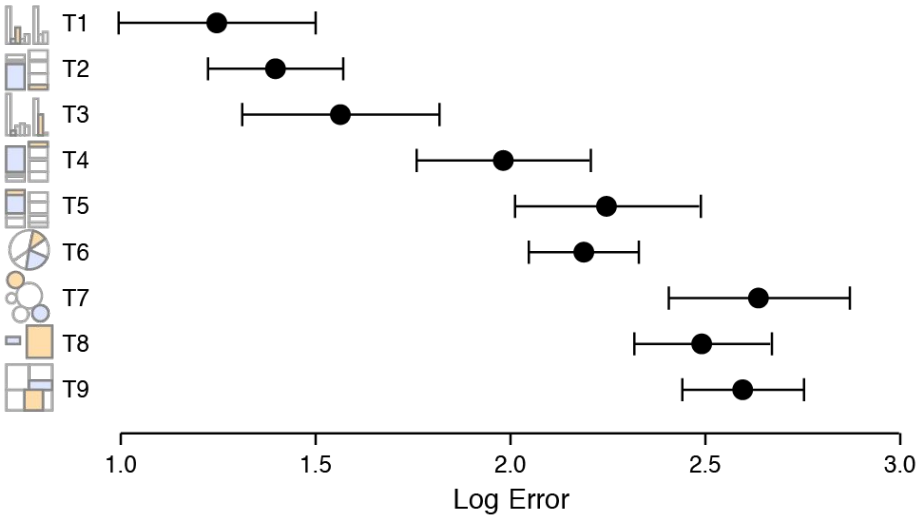




http://socviz.co/lookatdata.html; https://www.data-to-viz.com/caveat

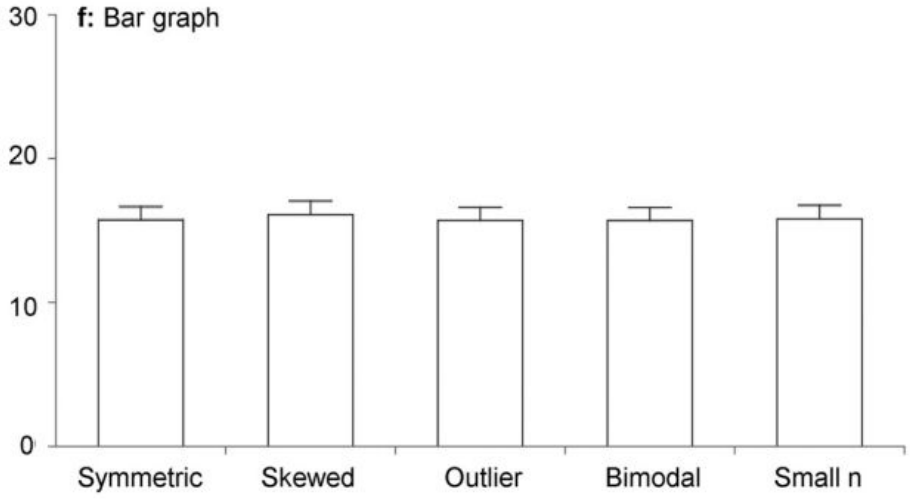
Visualization challenges

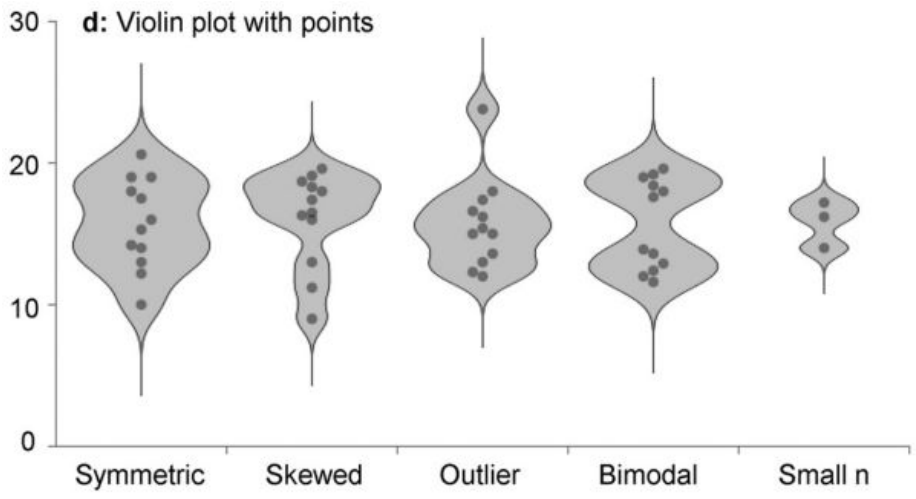




http://socviz.co/lookatdata.html

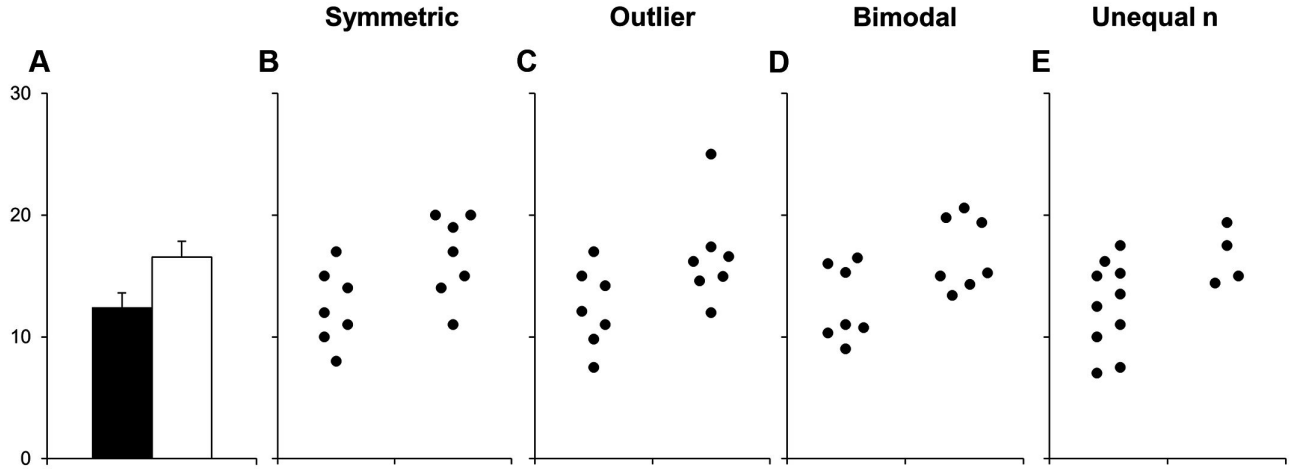
Visualization challenges – Bar plots, error bars





http://www.jbc.org/content/292/50/20592.full.pdf

Visualization challenges – Bar plots, error bars



Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

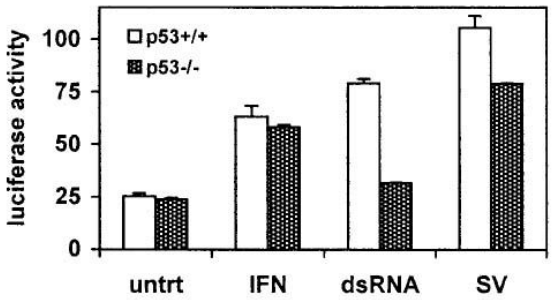
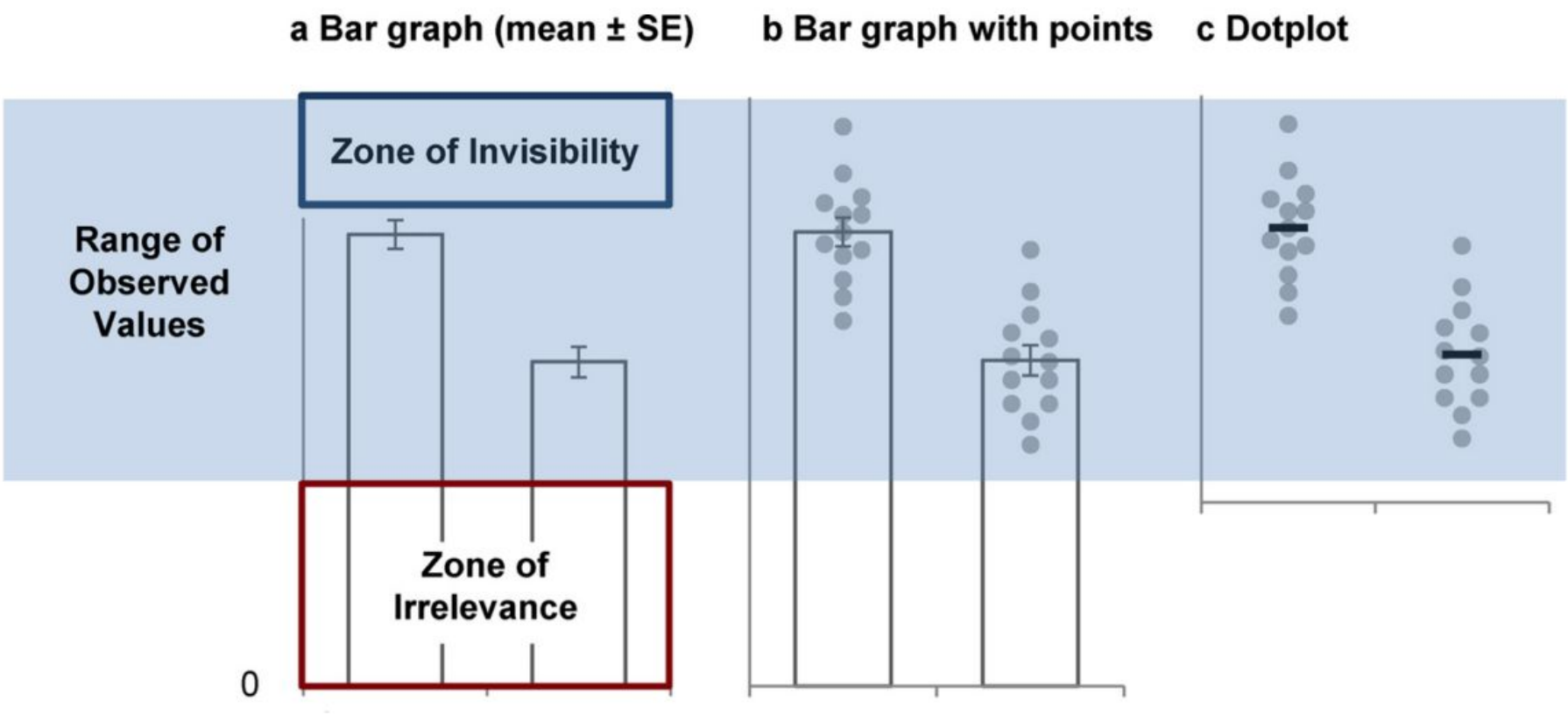


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were

differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are repre-

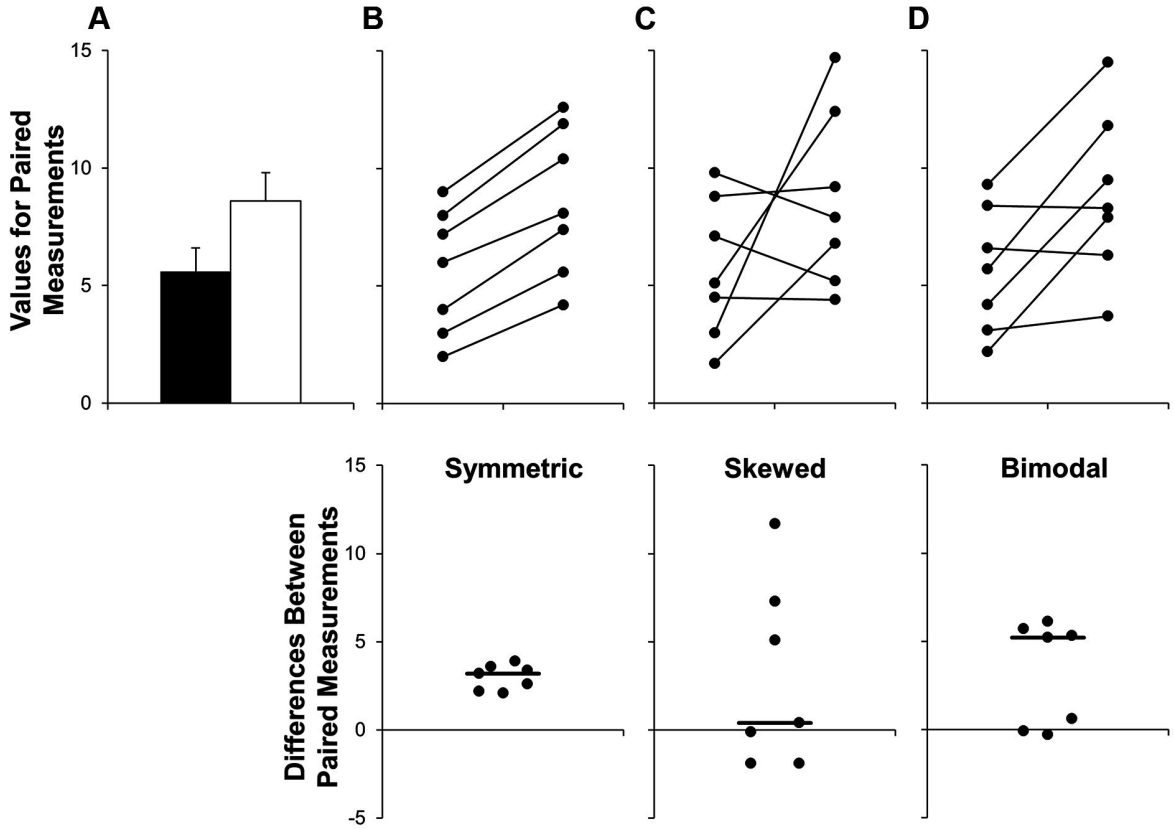
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

Visualization challenges – Bar plots, error bars



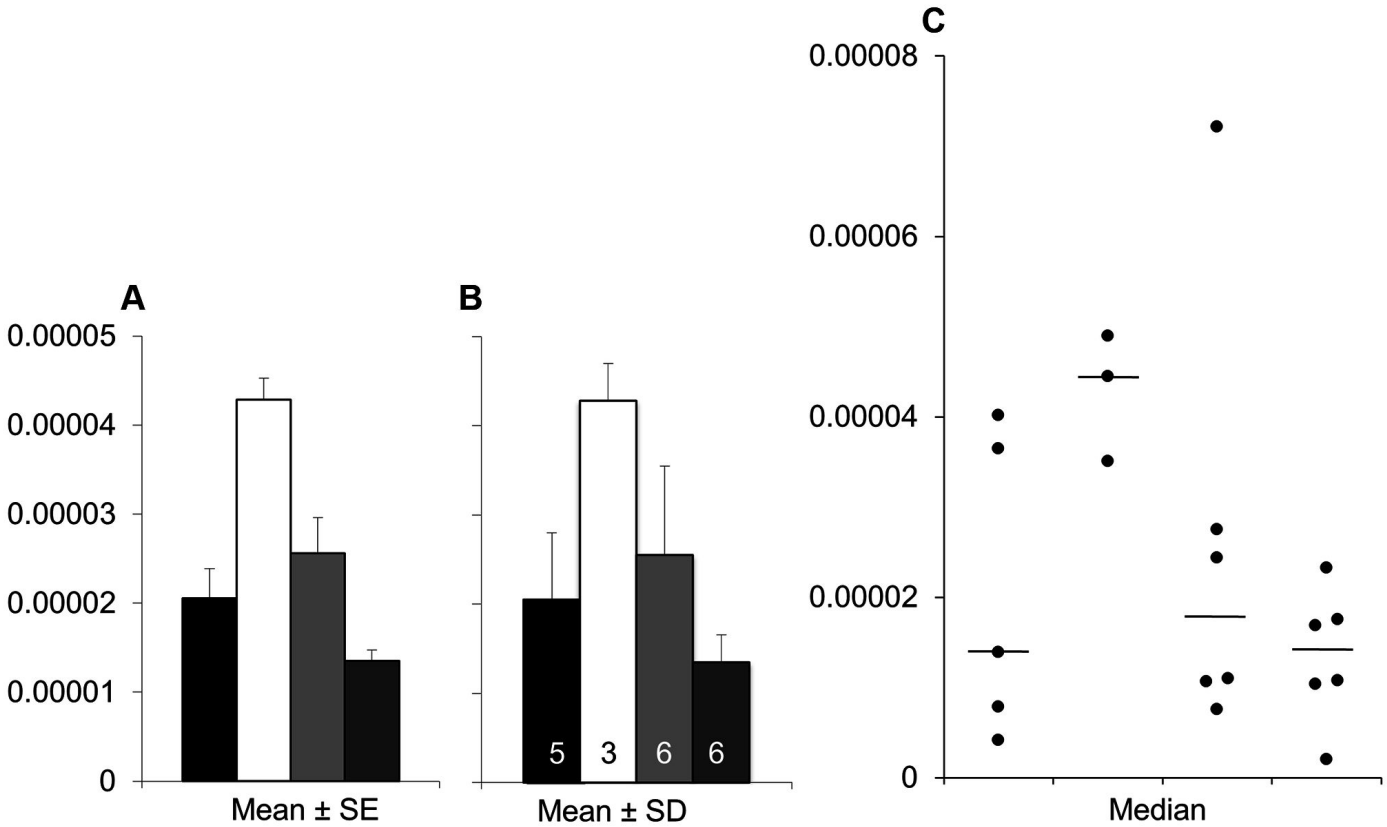
http://www.jbc.org/content/292/50/20592.full.pdf

Visualization challenges – Bar plots, error bars



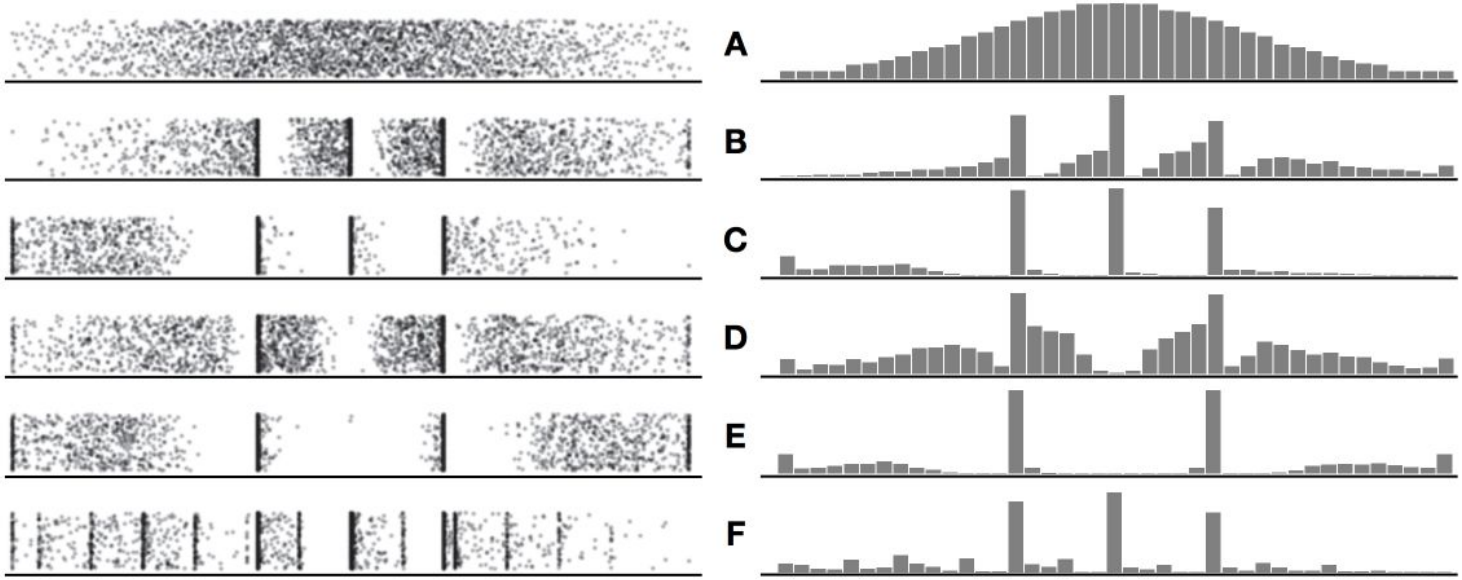
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

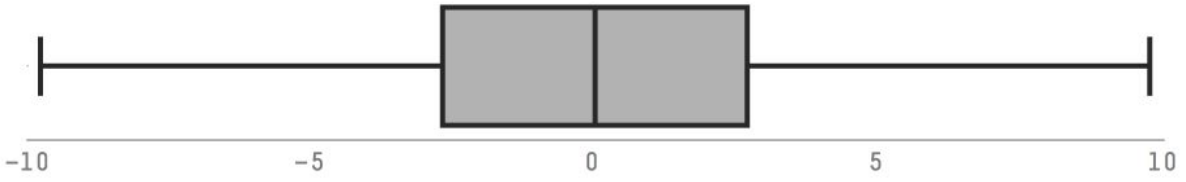
Visualization challenges – Bar plots, error bars



https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

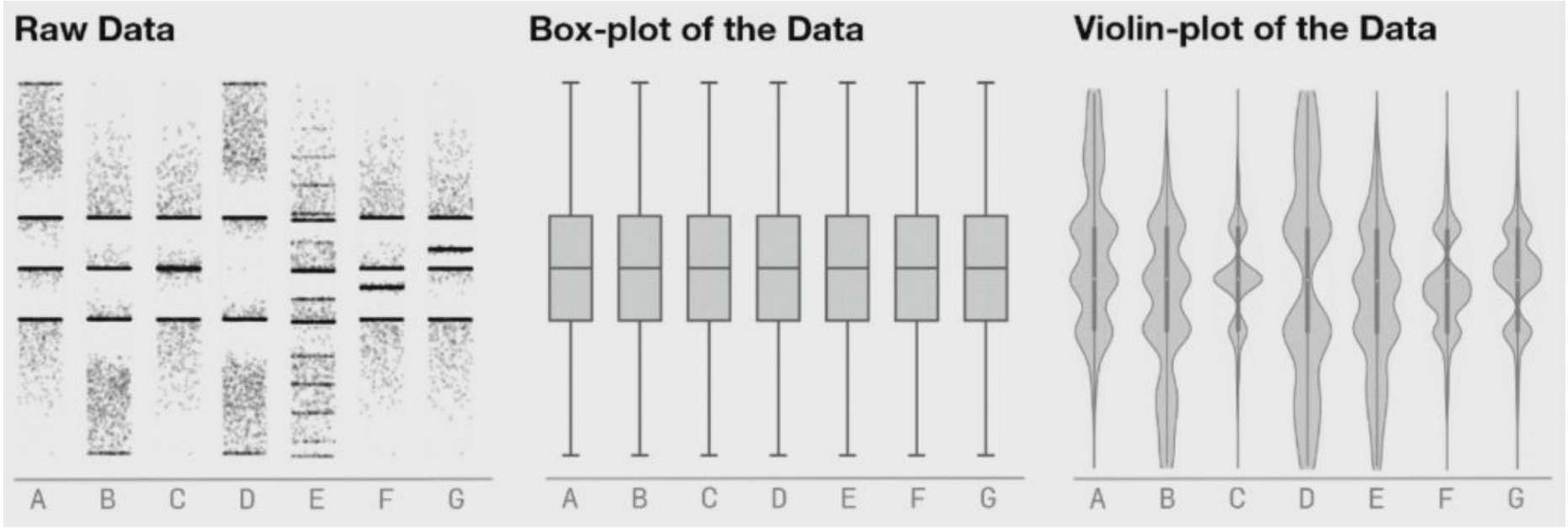
Visualization challenges – Different distributions





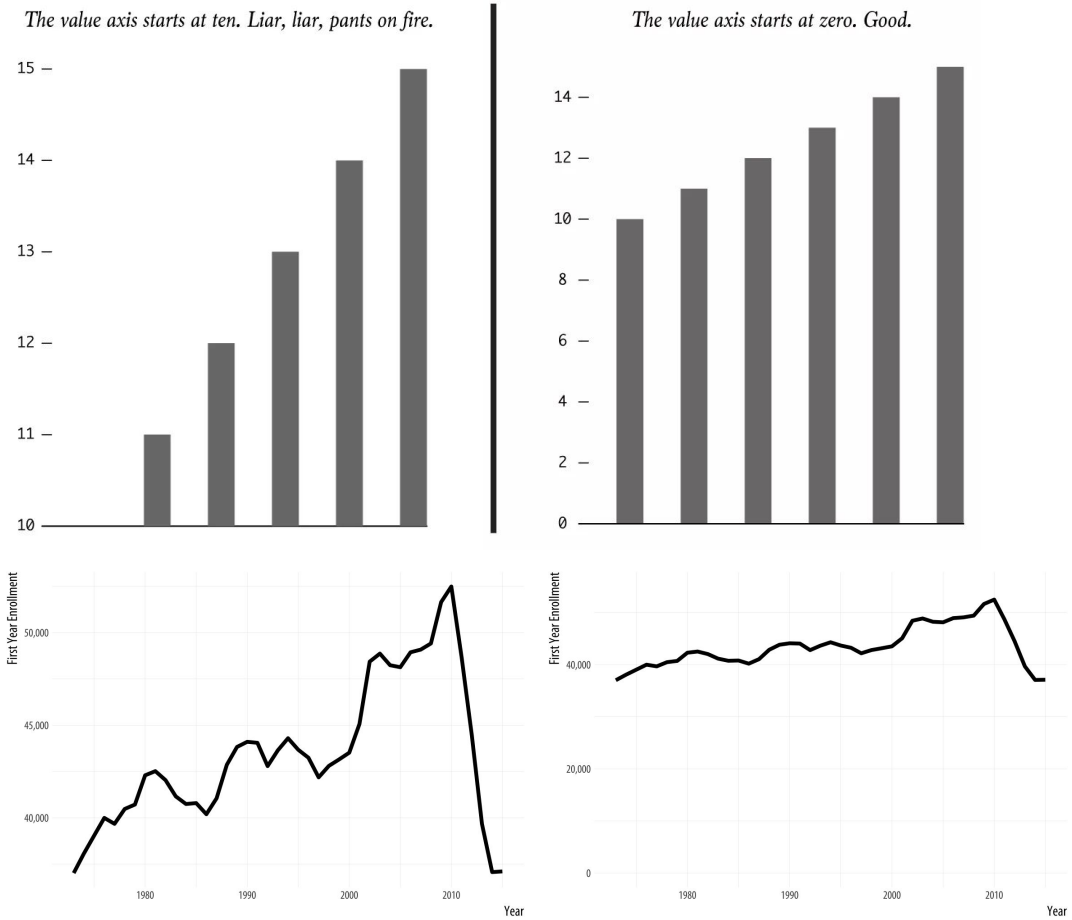
https://www.autodeskresearch.com/sites/default/files/SameStats-DifferentGraphs.pdf

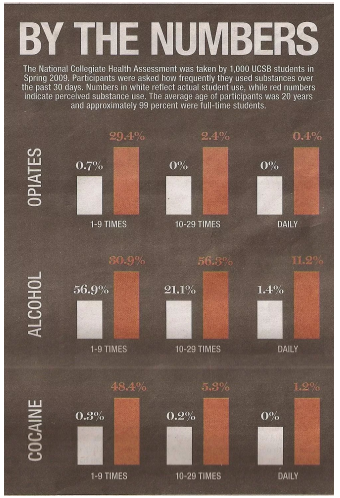
Visualization challenges – Different distributions



https://www.autodeskresearch.com/sites/default/files/SameStats-DifferentGraphs.pdf

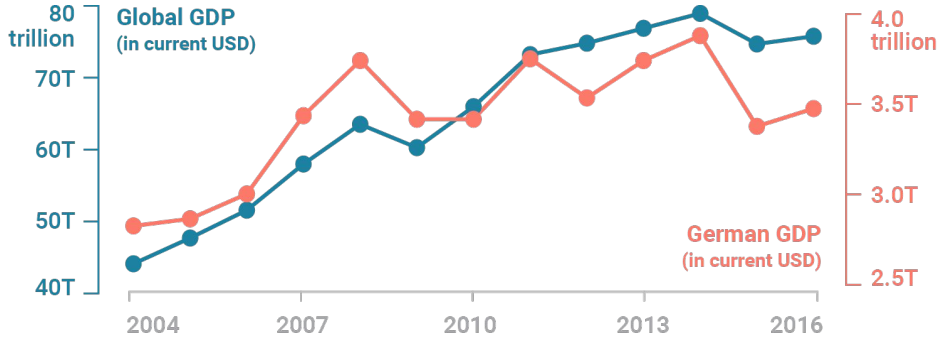
Visualization challenges – Truncated axis

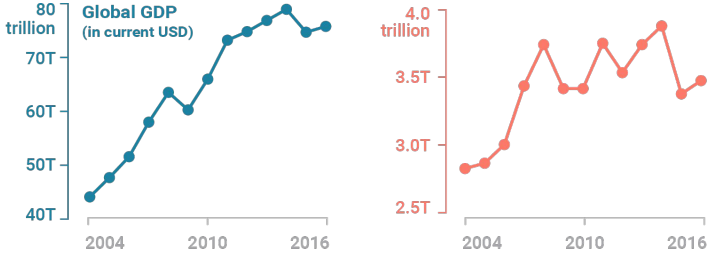


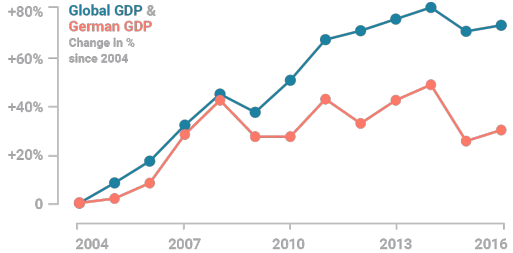


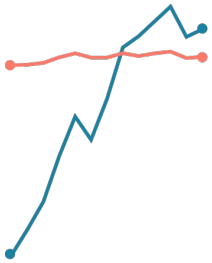
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/
http://socviz.co/lookatdata.html

Visualization challenges – Dual axes

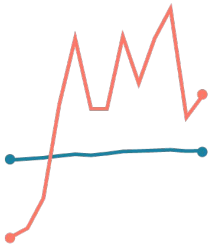




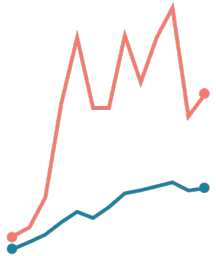




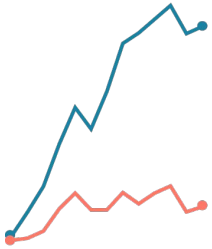
Orange steady,
Blue massively increasing.



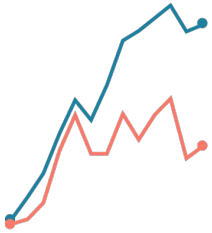
Blue steady,
Orange increasing.



Both started at the same level, but Orange increased far more than Blue.



Both started at the same level, but Blue increased far more than Orange.



Both started with the same increase, then Blue raced to the top.

Both steady.

https://blog.datawrapper.de/dualaxis/

Visualization challenges – Inappropriate axes

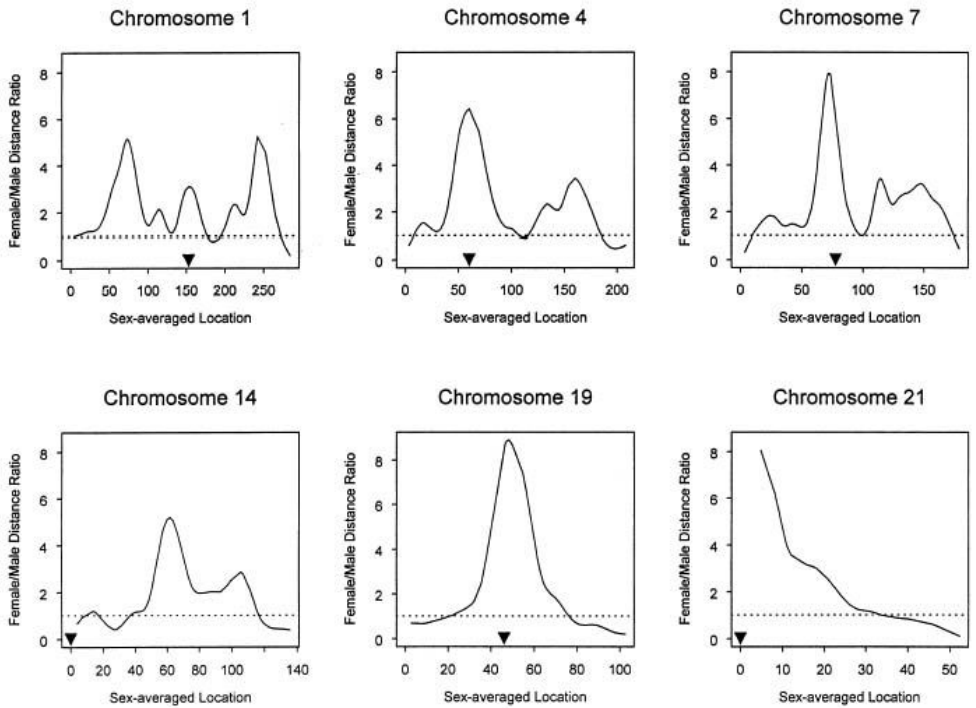
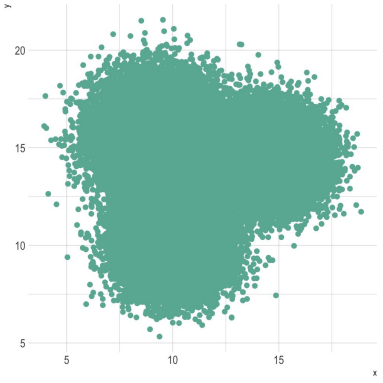
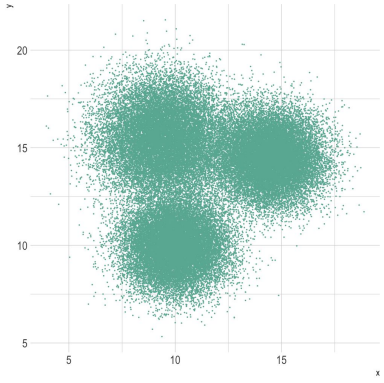


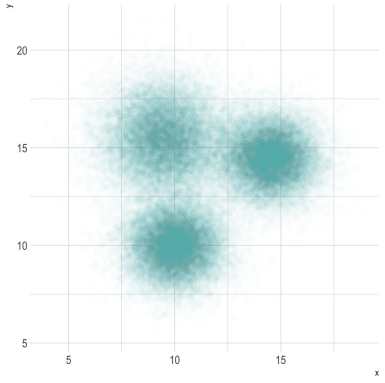
Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

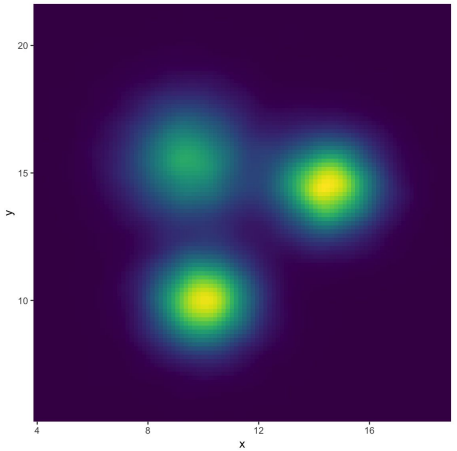
http://www.ncbi.nlm.nih.gov/pubmed/9718341

Visualization challenges – Overplotting



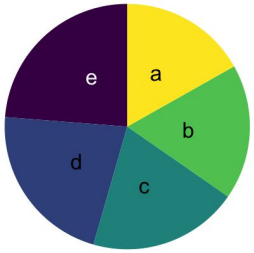


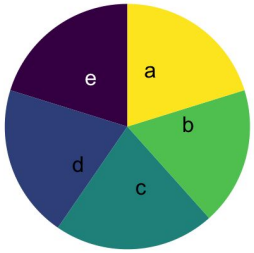


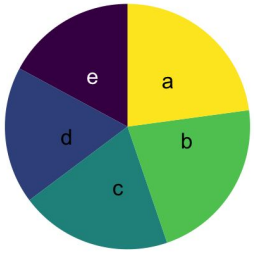


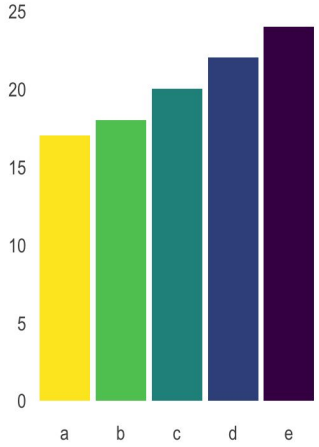
https://www.data-to-viz.com/caveat/

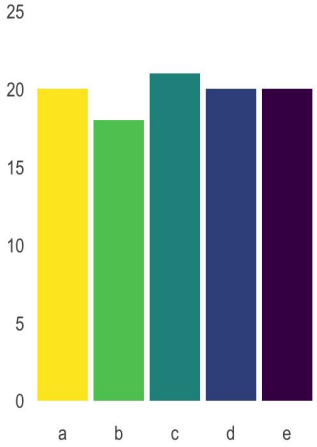
Visualization challenges – Pie charts - almost never a good idea

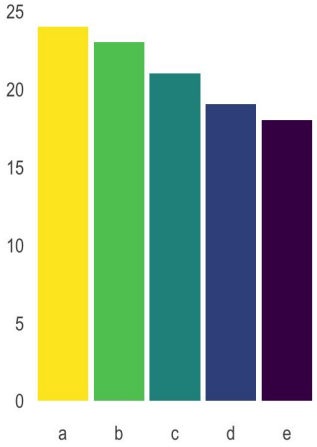






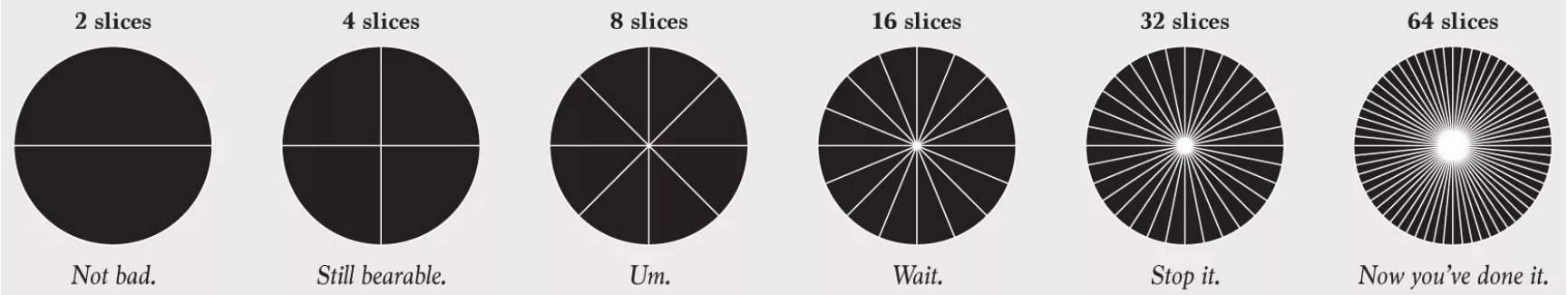


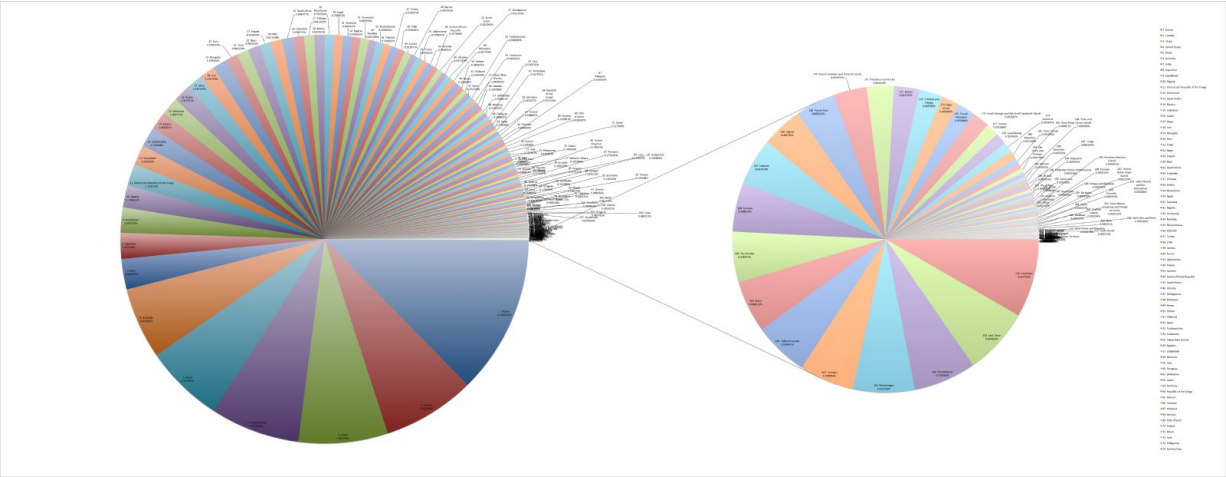




https://www.data-to-viz.com/caveat/

Visualization challenges – Pie charts - almost never a good idea

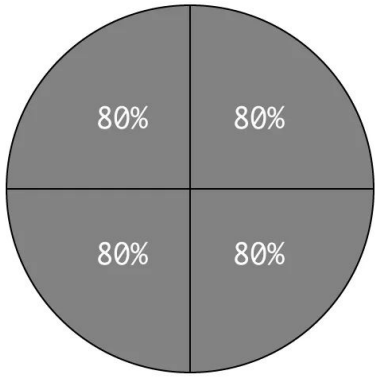


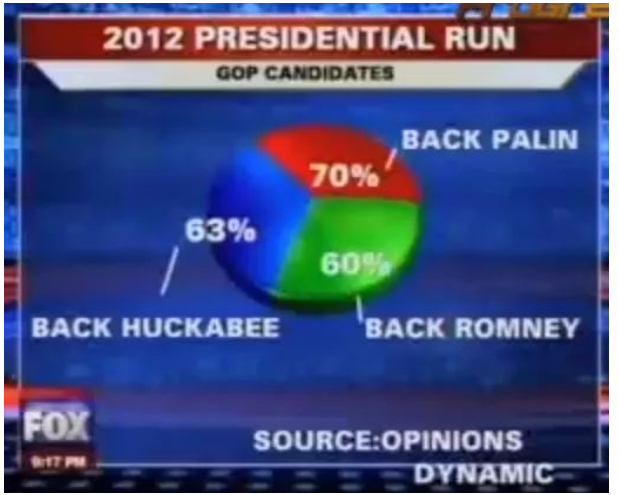


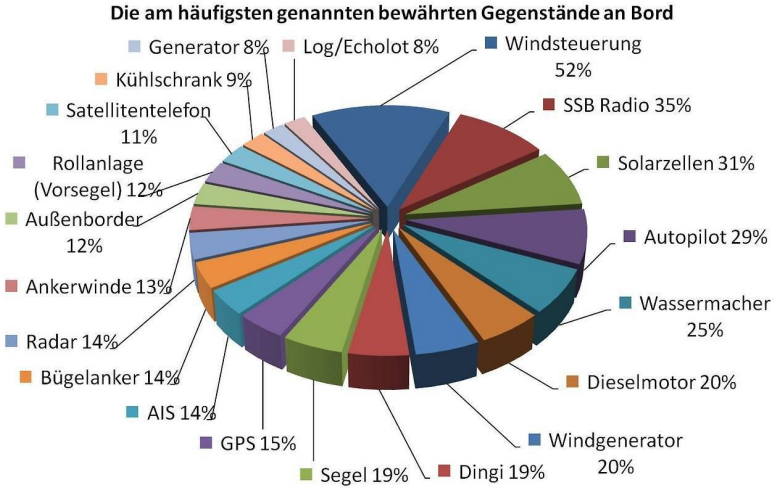
https://flowingdata.com/2015/08/11/real-chart-rules-to-follow/

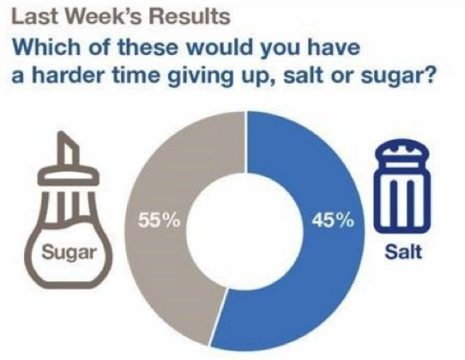
https://commons.wikimedia.org/wiki/File:Pie_chart_of_countries_by_area.png

Visualization challenges – When things don't add up



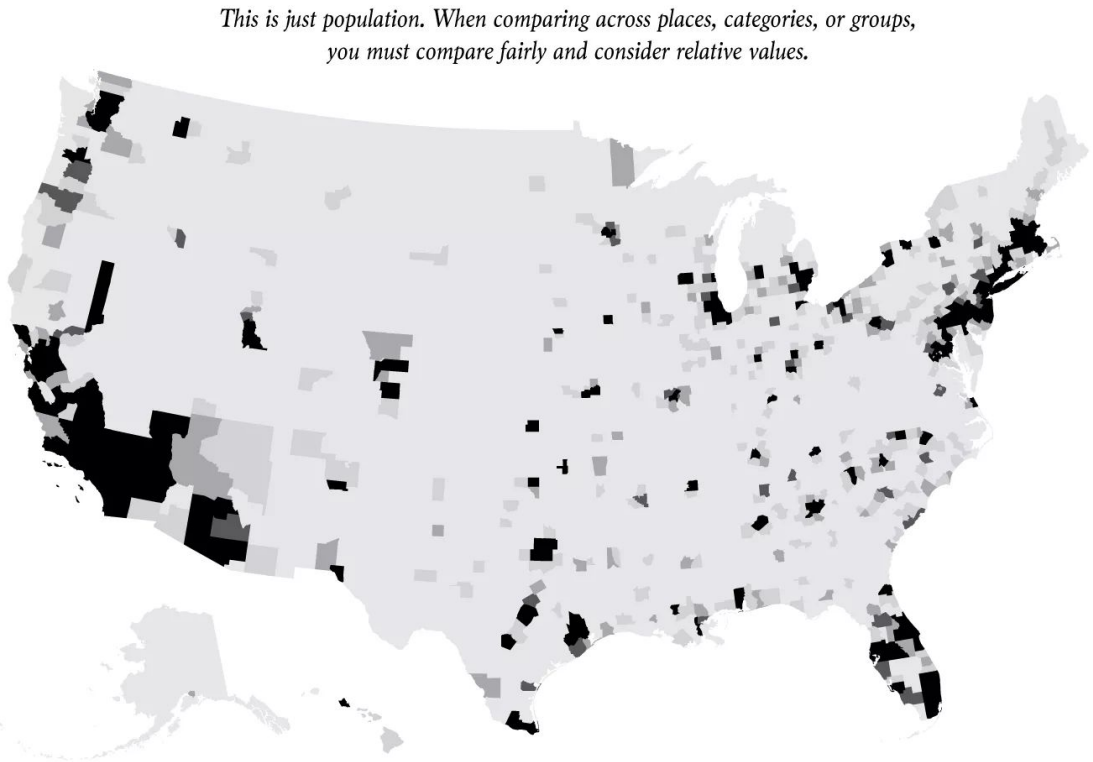


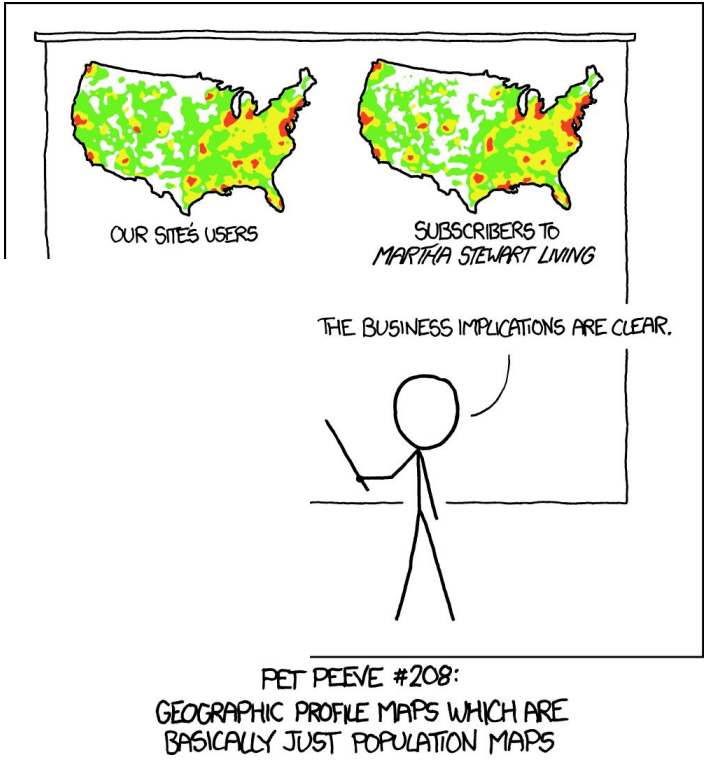




https://www.data-to-viz.com/caveat/
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

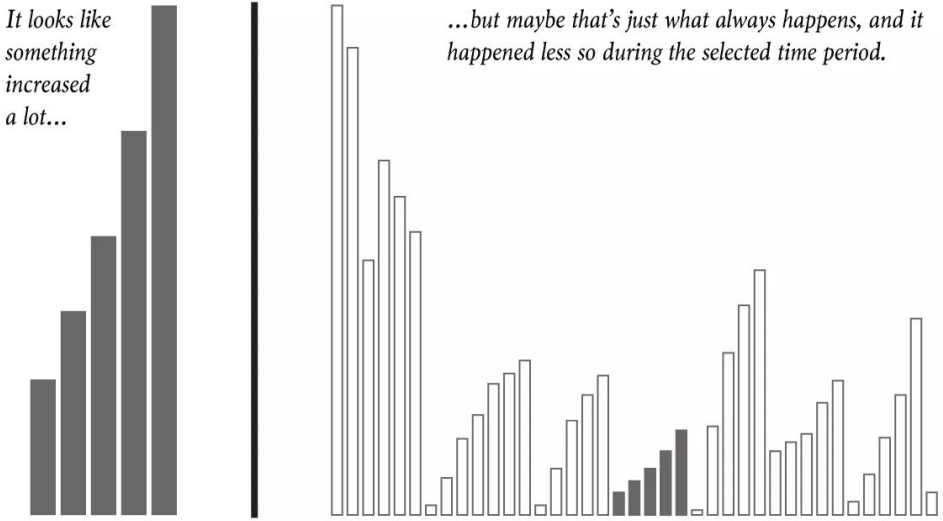
Visualization challenges – Absolutes vs. relative values

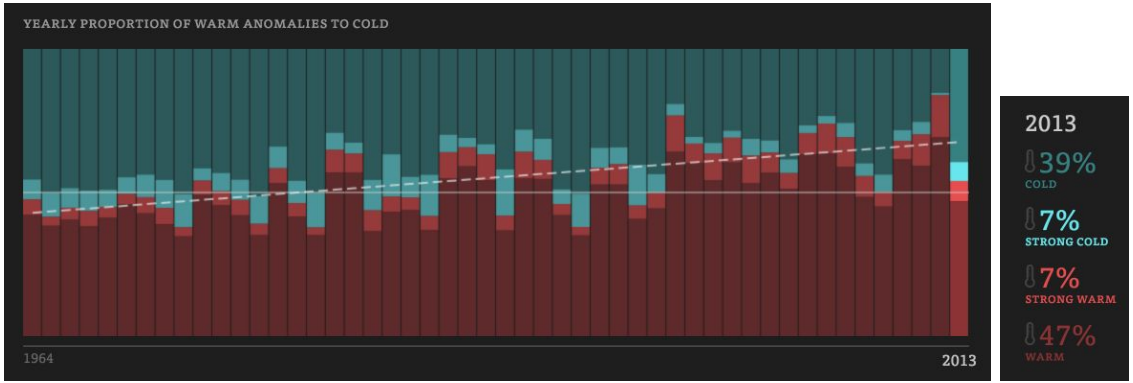




https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Visualization challenges – Limited scope



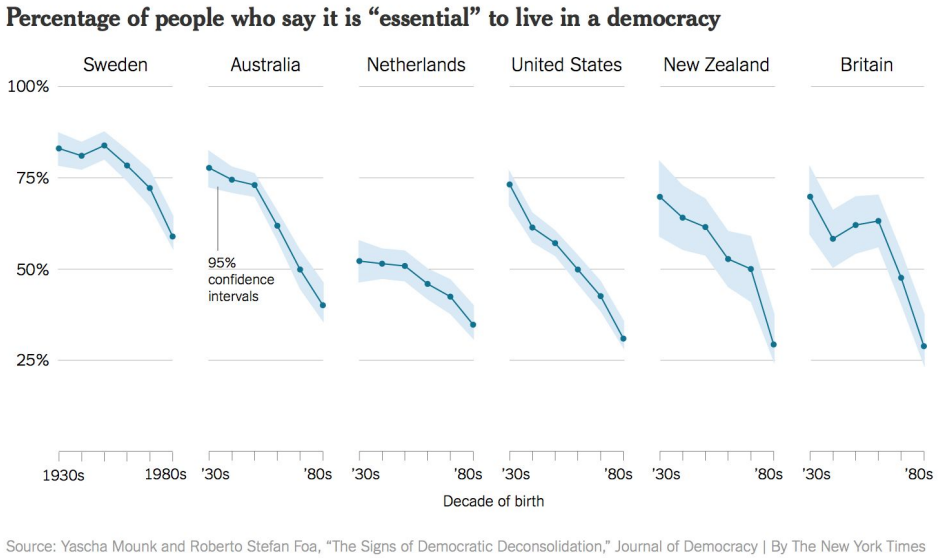


https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

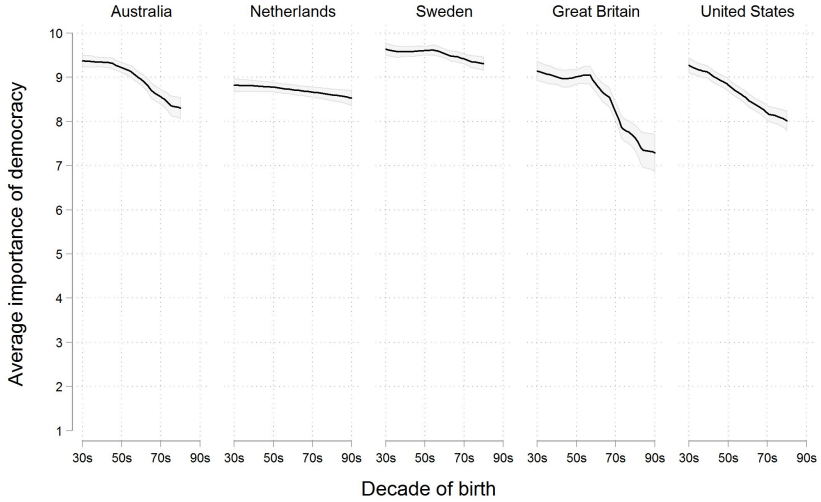
Visualization challenges – Choice of plot & data binning

Two bins. What's really in the 1+ category? Might be hiding something.

That's better. It can show more variation.



Source: Yascha Mounk and Roberto Stefan Foa, "The Signs of Democratic Deconsolidation," Journal of Democracy | By The New York Times



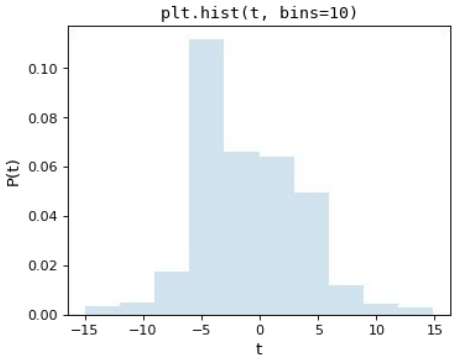
Graph by Erik Voeten, based on WVS 5

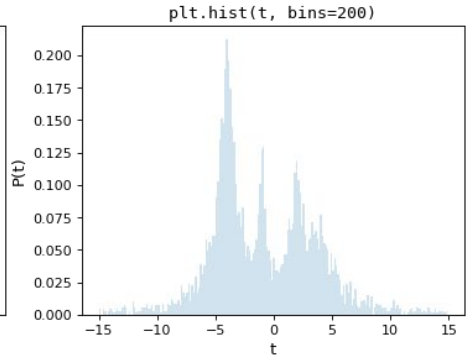
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

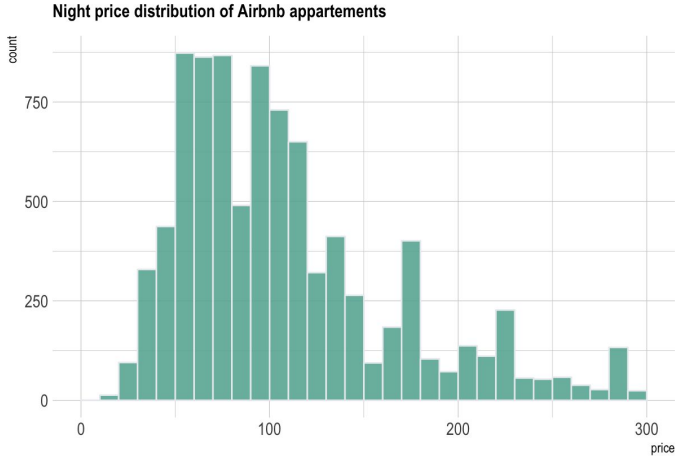
Visualization challenges – Choice of plot & data binning

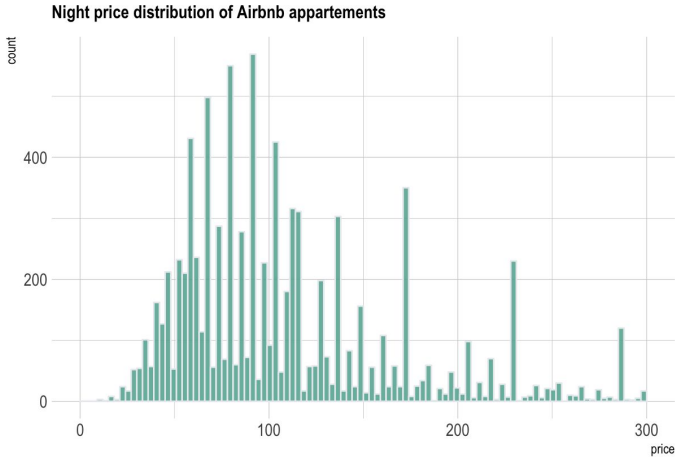
Two bins. What's really in the 1+ category? Might be hiding something.

That's better. It can show more variation.



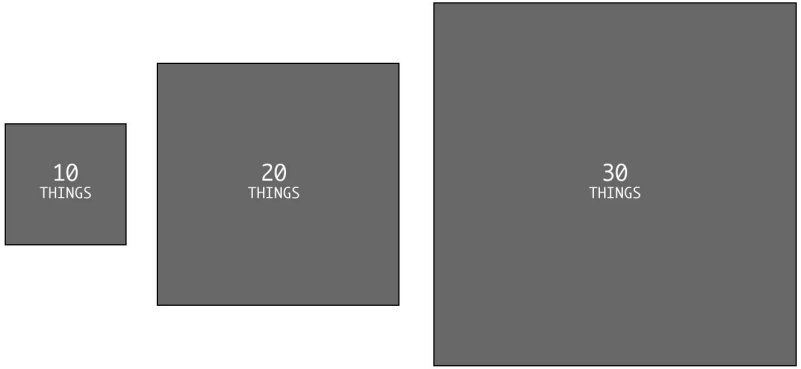


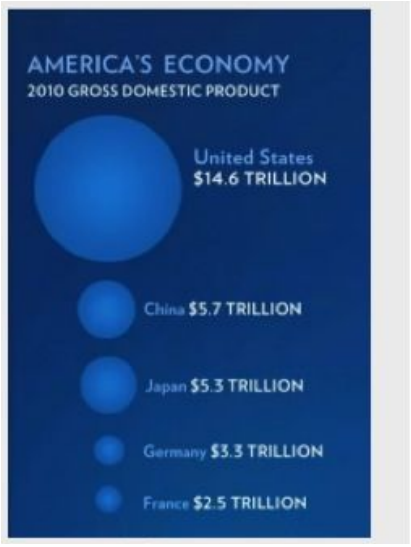


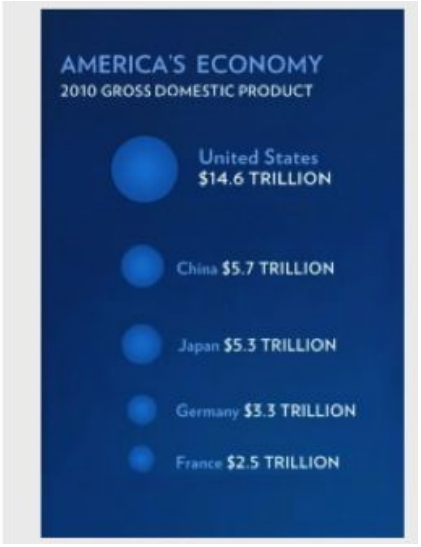


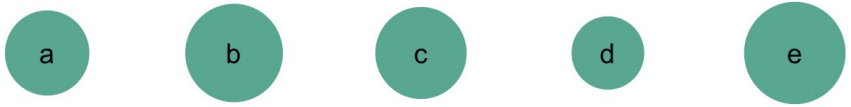
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

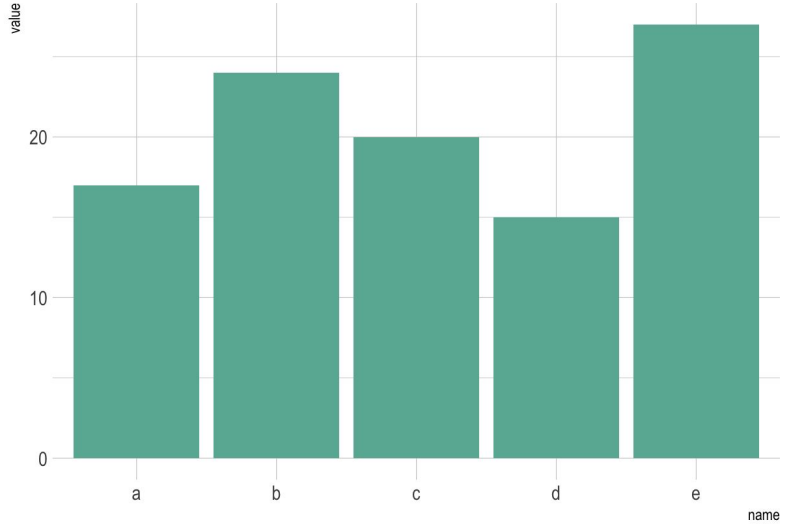
Visualization challenges – Area sized by dimension





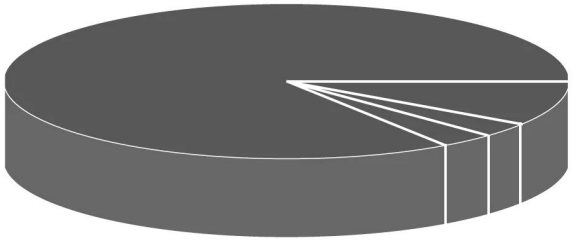






https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Visualization challenges – Unwarranted extra dimensions



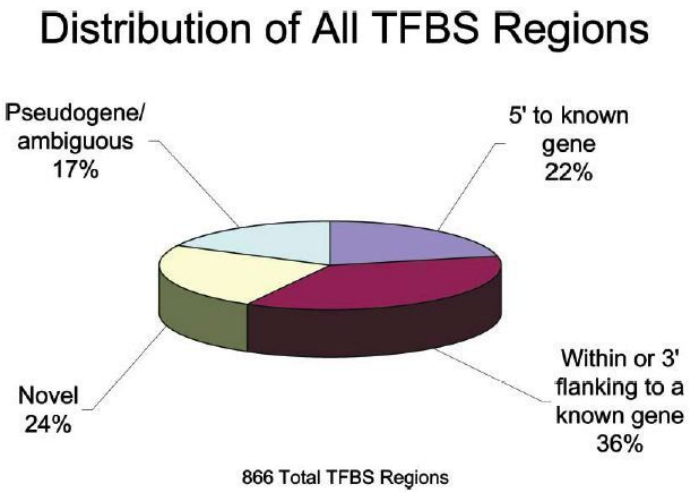
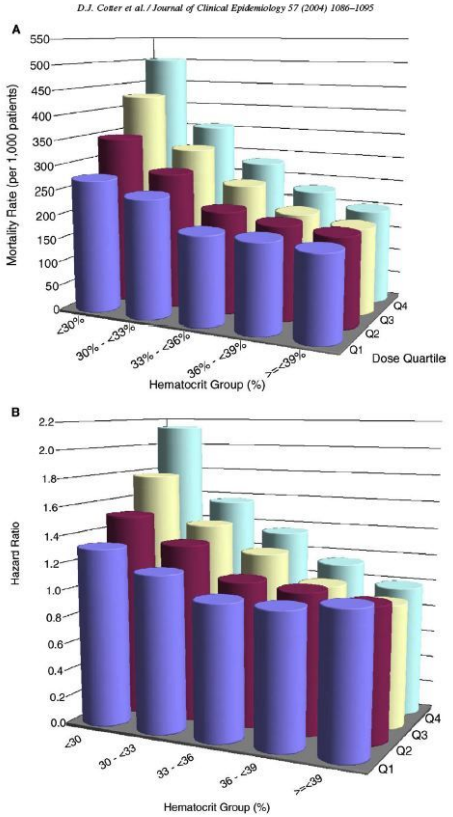


Figure 1. Classification of TFBS Regions

TFBS regions for Sp1, cMyc, and p53 were classified based upon proximity to annotations (RefSeq, Sanger hand-curated annotations, GenBank full-length mRNAs, and Ensembl predicted genes). The proximity was calculated from the center of each TFBS region. TFBS regions were classified as follows: within 5 kb of the 5′ most exon of a gene, within 5 kb of the 3′ terminal exon, or within a gene, novel or outside of any annotation, and pseudogene/ambiguous (TFBS overlapping or flanking pseudogene annotations, limited to chromosome 22, or TFBS regions falling into more than one of the above categories).



https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/