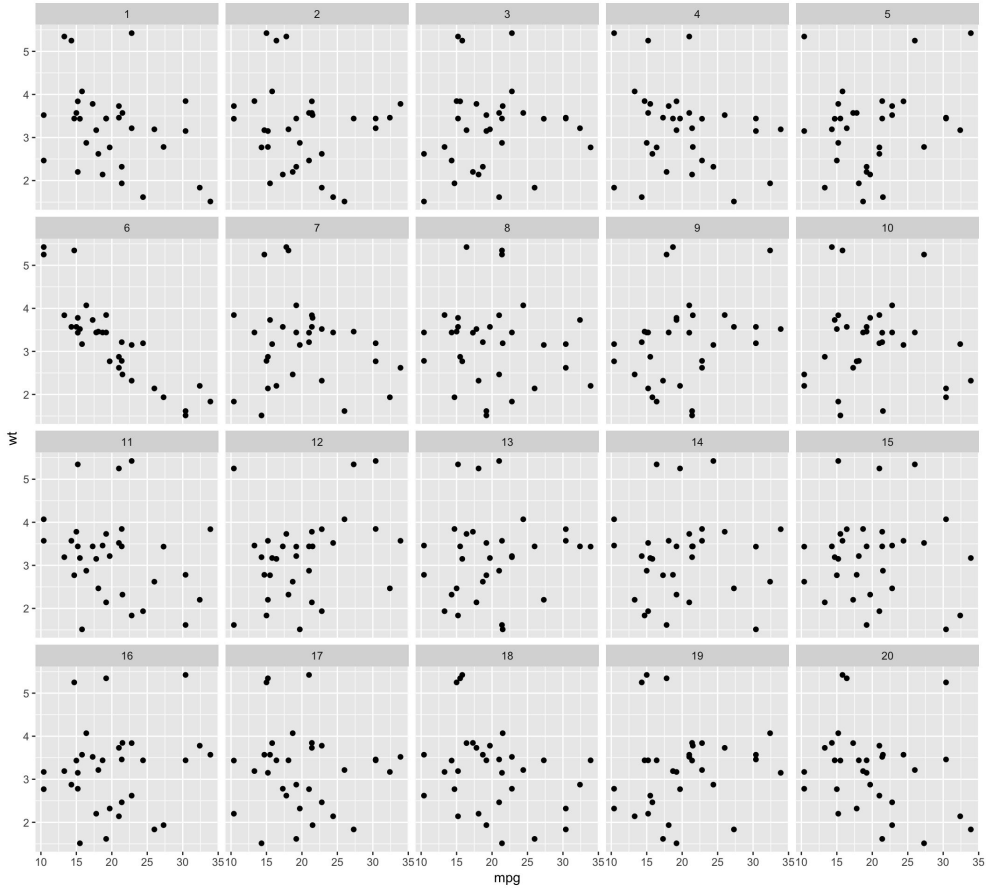
Day 08

Visualization challenges

- A survey of plots
- Issues with each type of plot

Spurious correlations – But it *looks* associated!

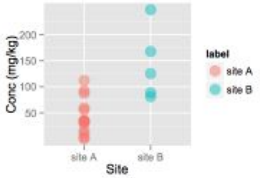
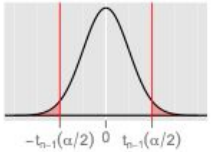
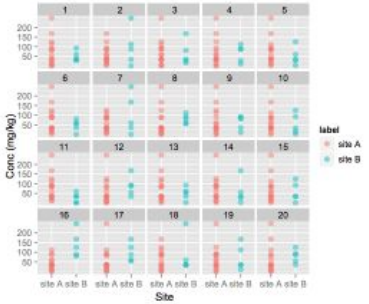


Create a lineup for visual inference

- 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

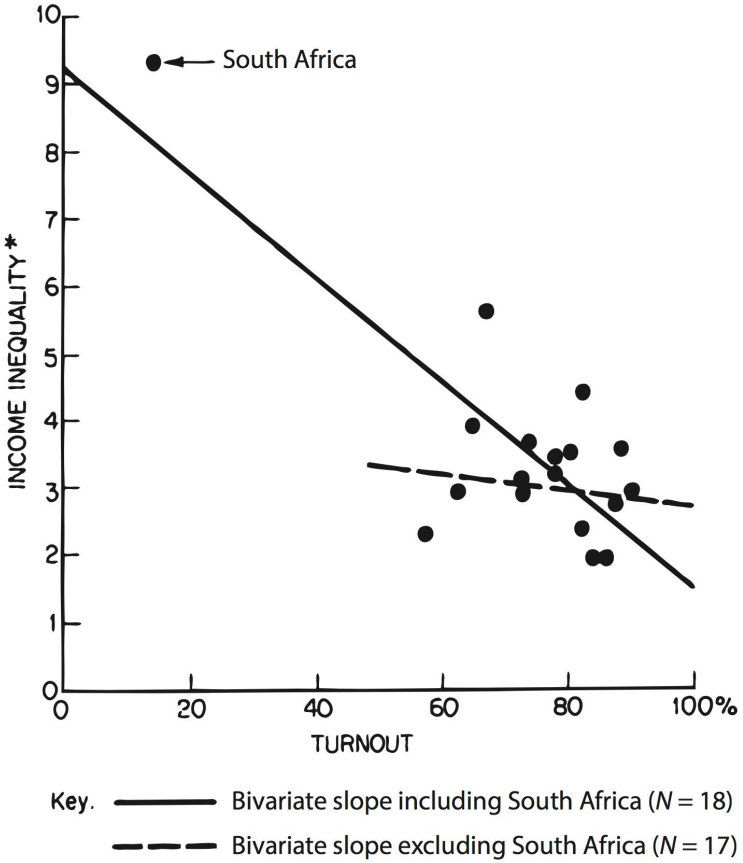
Spurious correlations – But it *looks* associated!

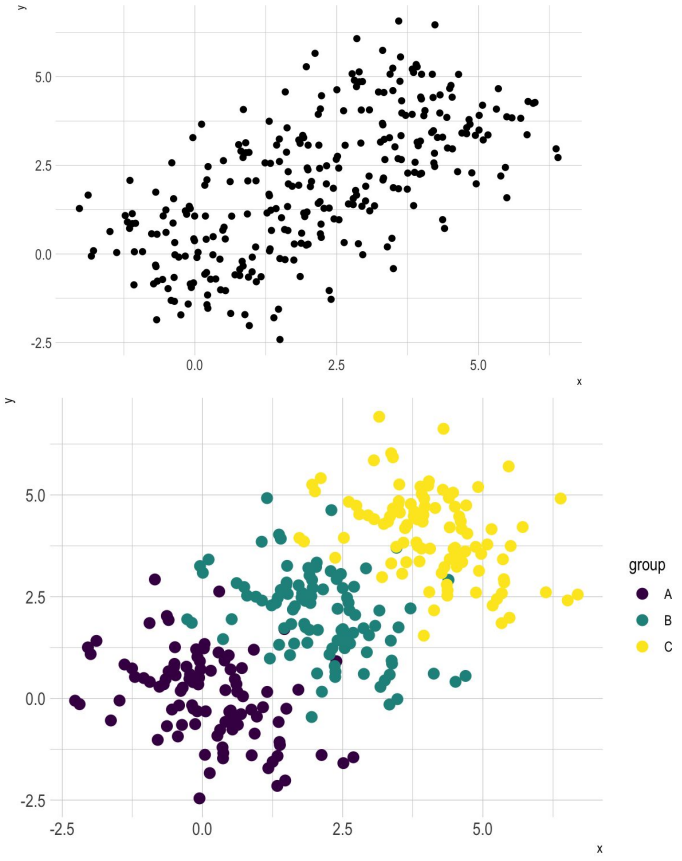
	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T(y) =$ 
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$;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

Purpose of data visualization

- Immediately convey information about study design.
- Illustrate important findings.
- Allow the reader to confirm that the statistical analysis is appropriate for the study design.
- Allow the reader to critically evaluate the data.

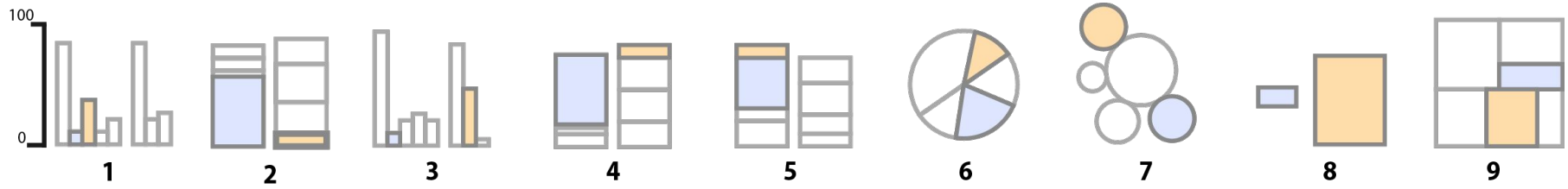
Visualization challenges – Don't do statistics without visualization

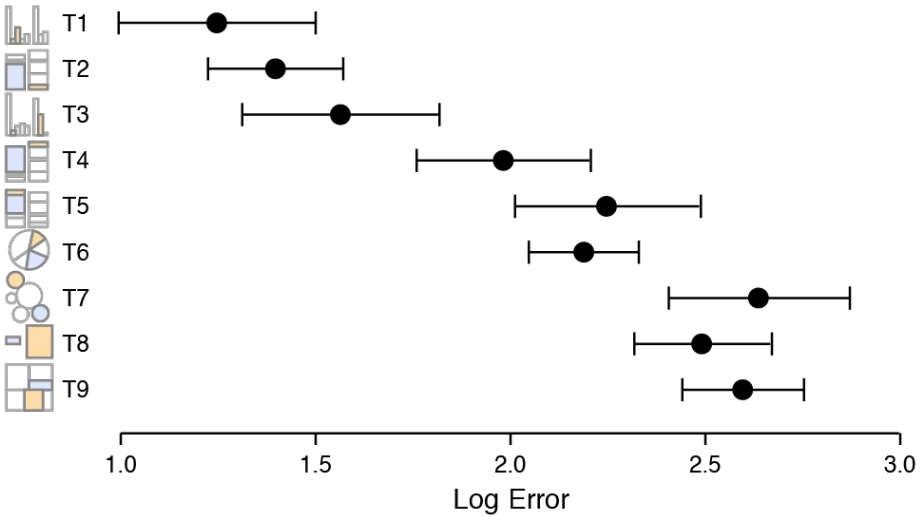




http://socviz.co/lookatdata.html; https://www.data-to-viz.com/caveat

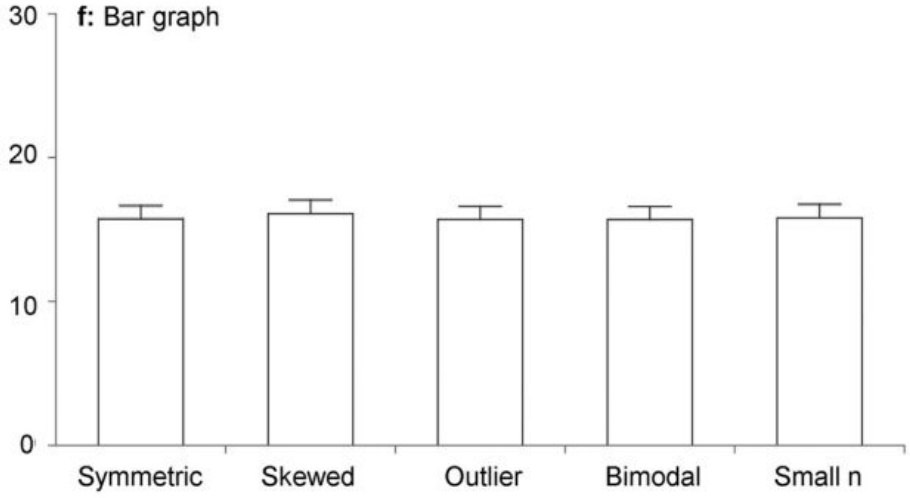
Visualization challenges

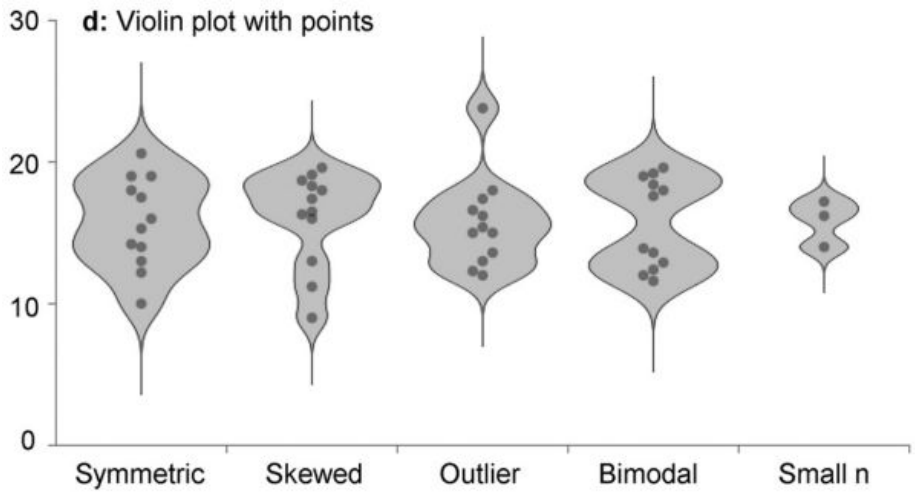




http://socviz.co/lookatdata.html

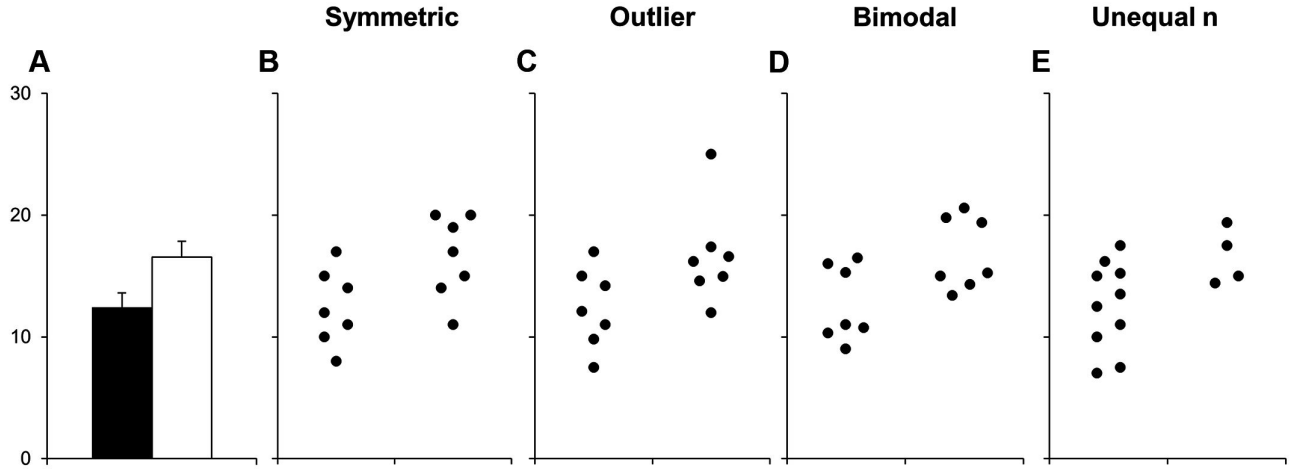
Visualization challenges – Bar plots, error bars





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

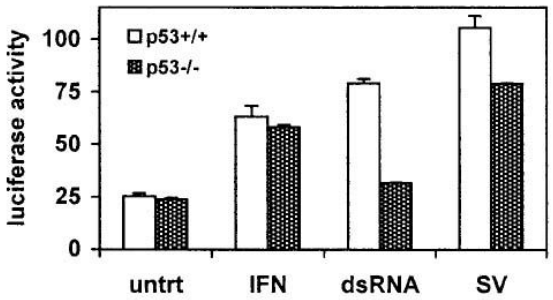


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were

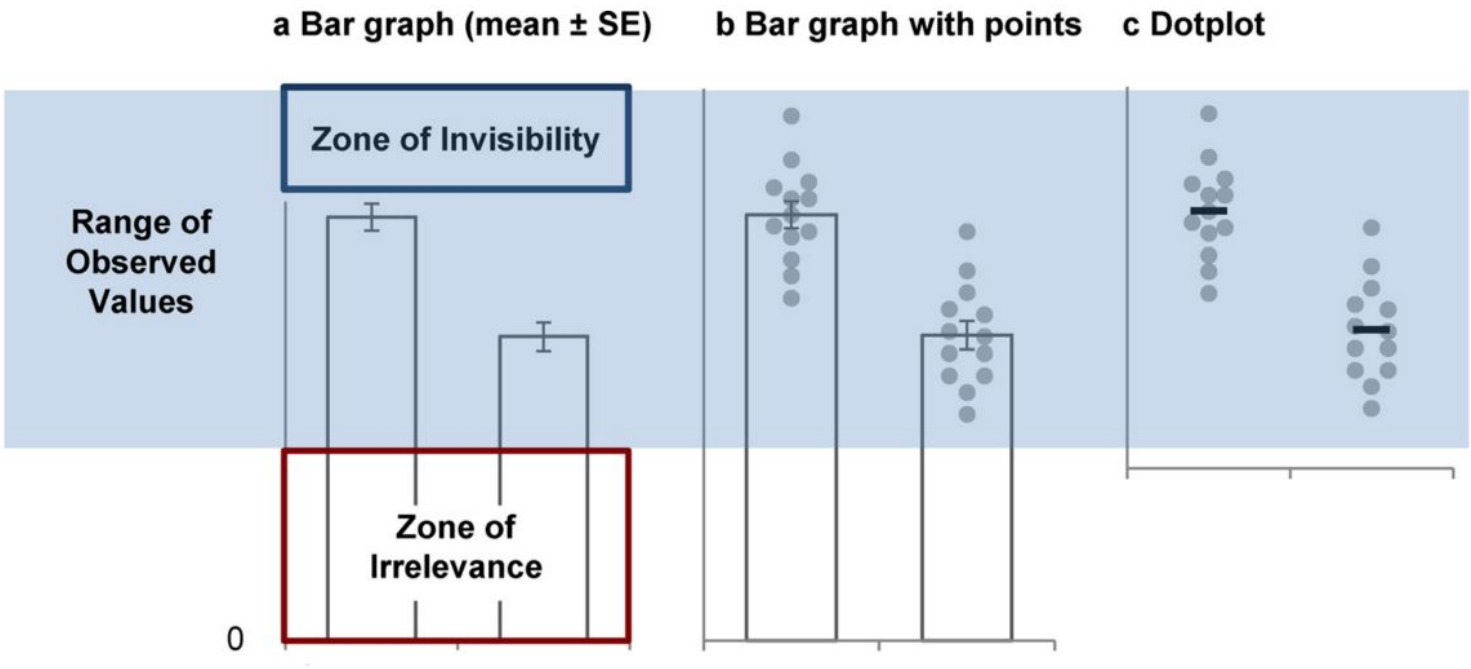
differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are repre-

Parametric vs.

Non-parametric test

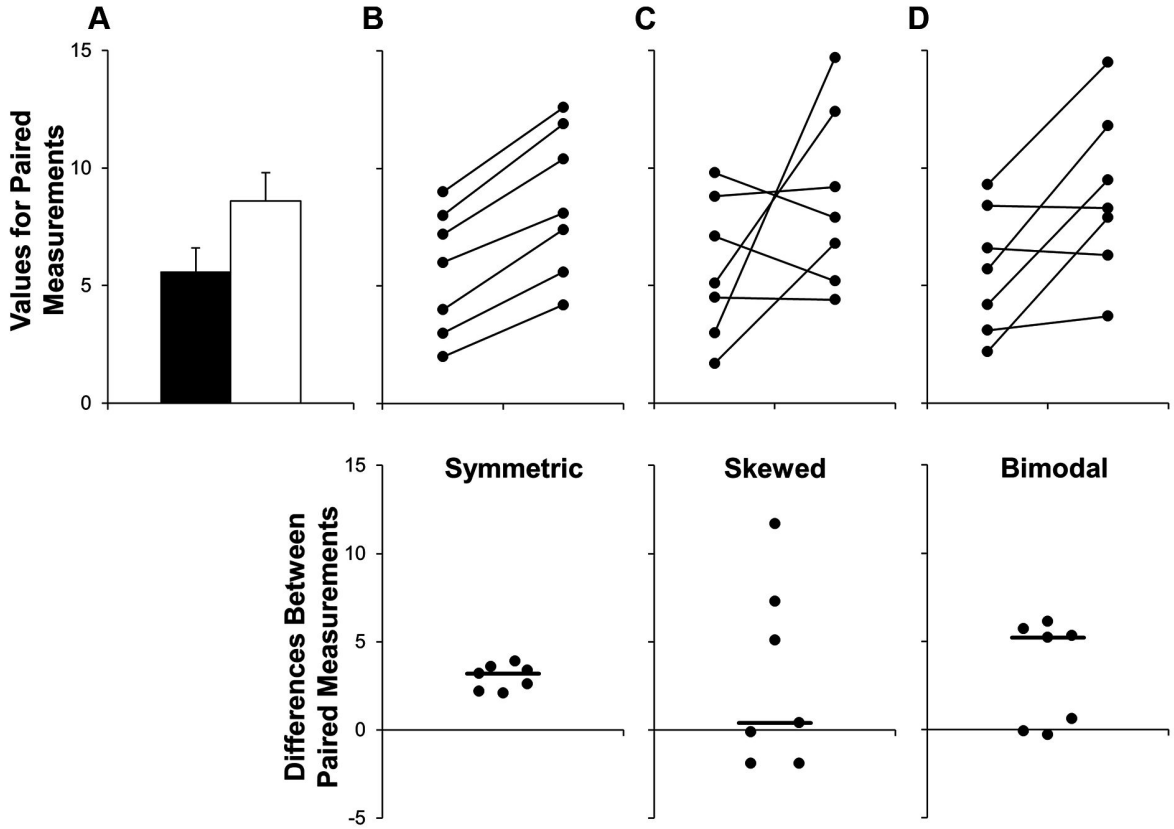
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

Visualization challenges – Bar plots, error bars



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

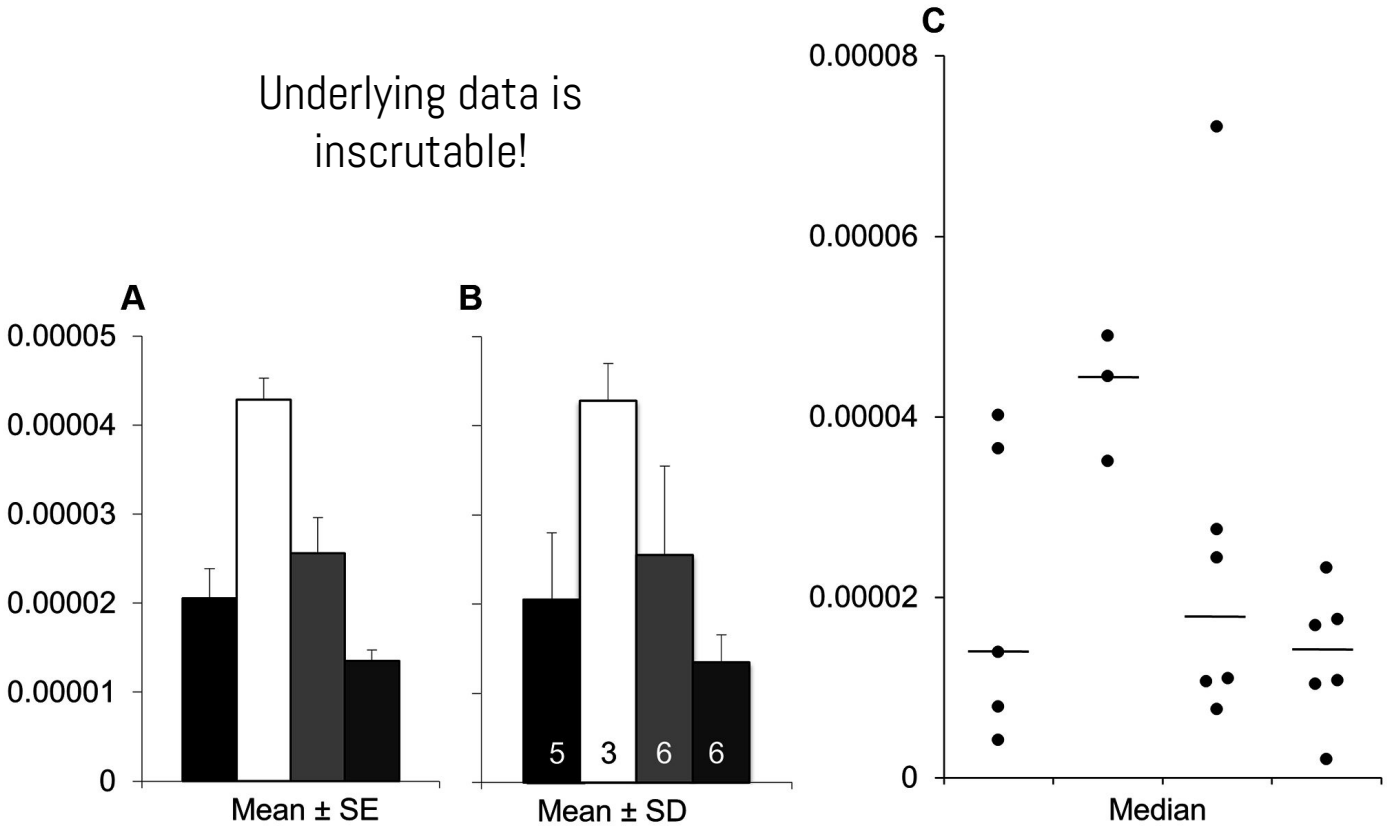
Visualization challenges – Bar plots, error bars



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

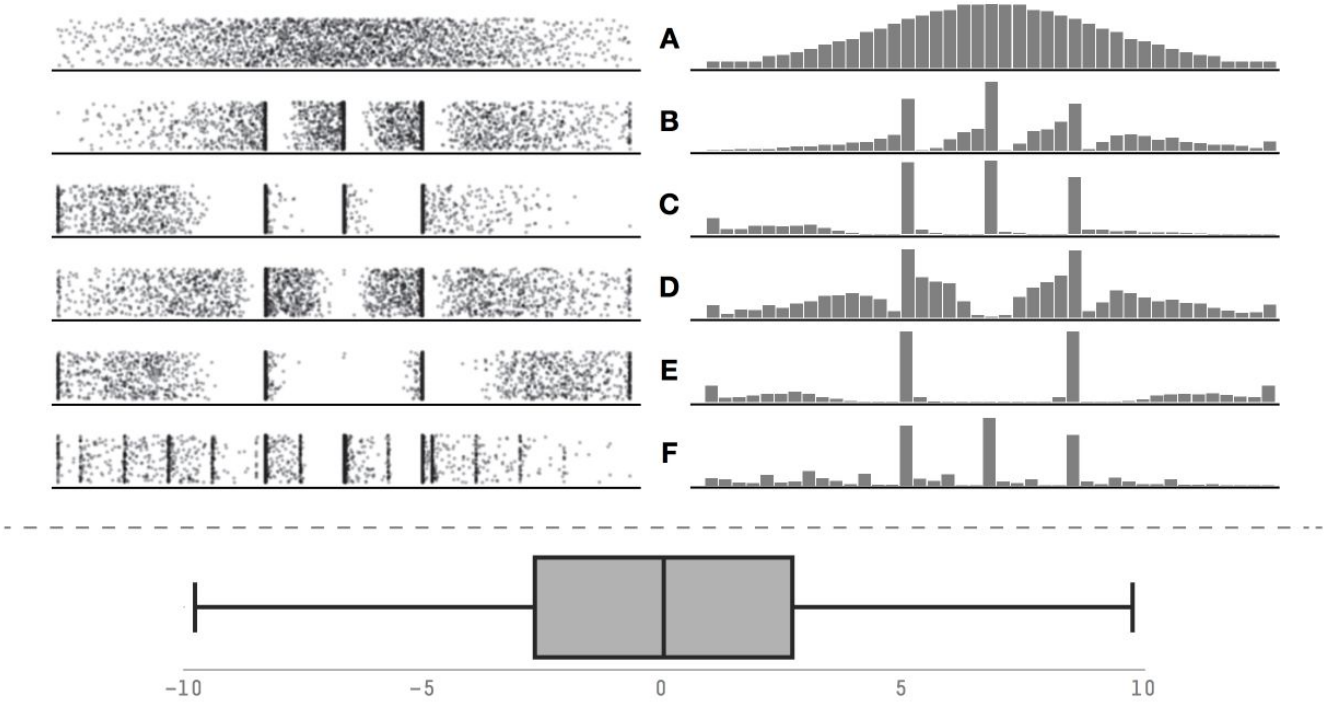
Visualization challenges – Bar plots, error bars

Underlying data is inscrutable!



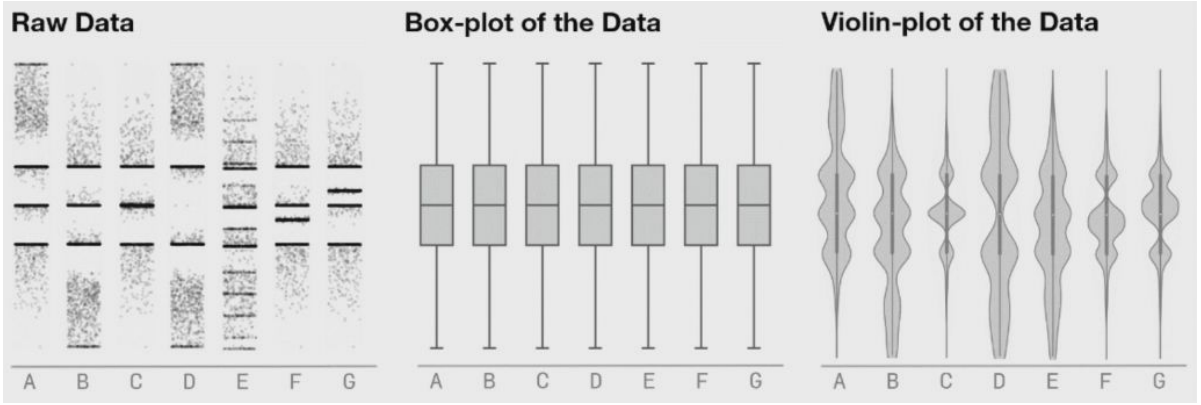
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

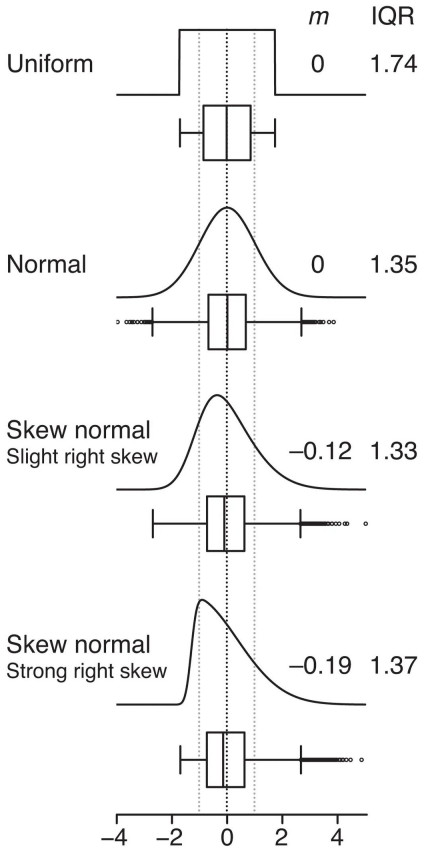
Visualization challenges – Different distributions



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

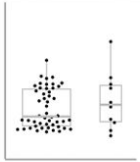
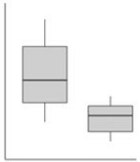
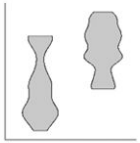
Visualization challenges – Different distributions





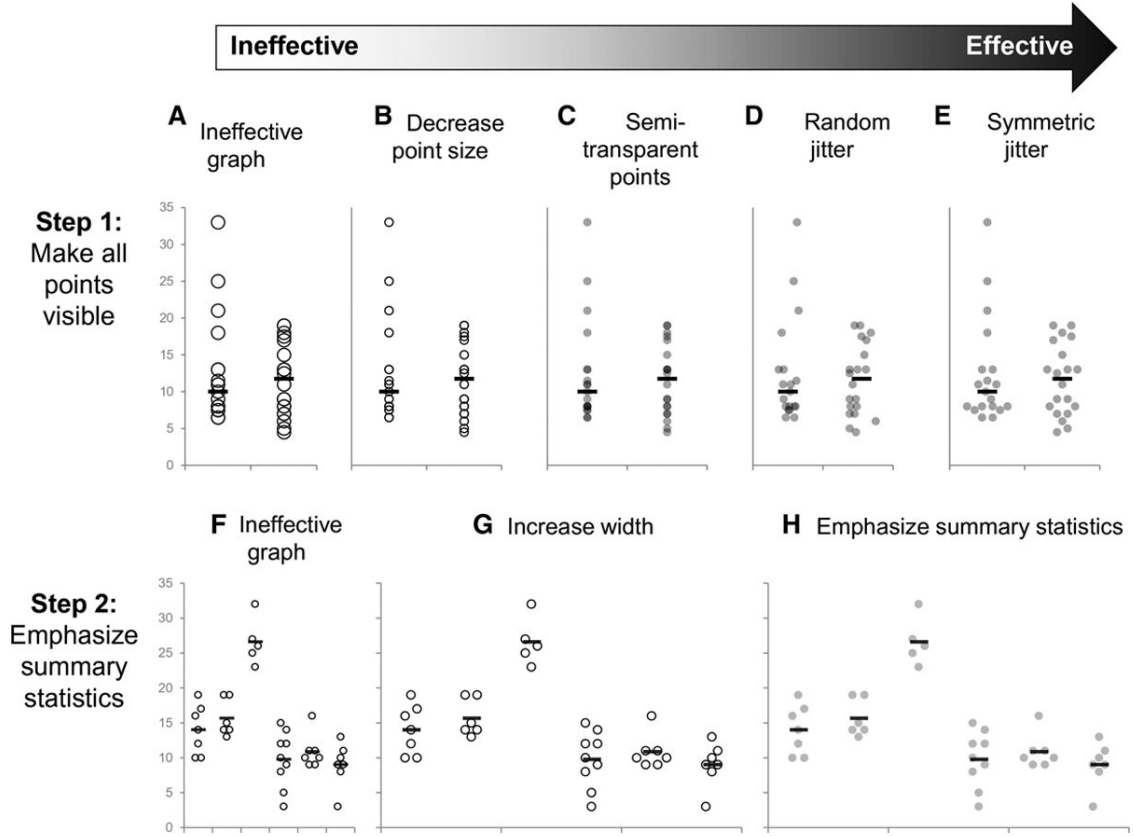
https://www.autodeskresearch.com/sites/default/files/SameStats-DifferentGraphs.pdf

Visualization challenges – Different distributions

Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**

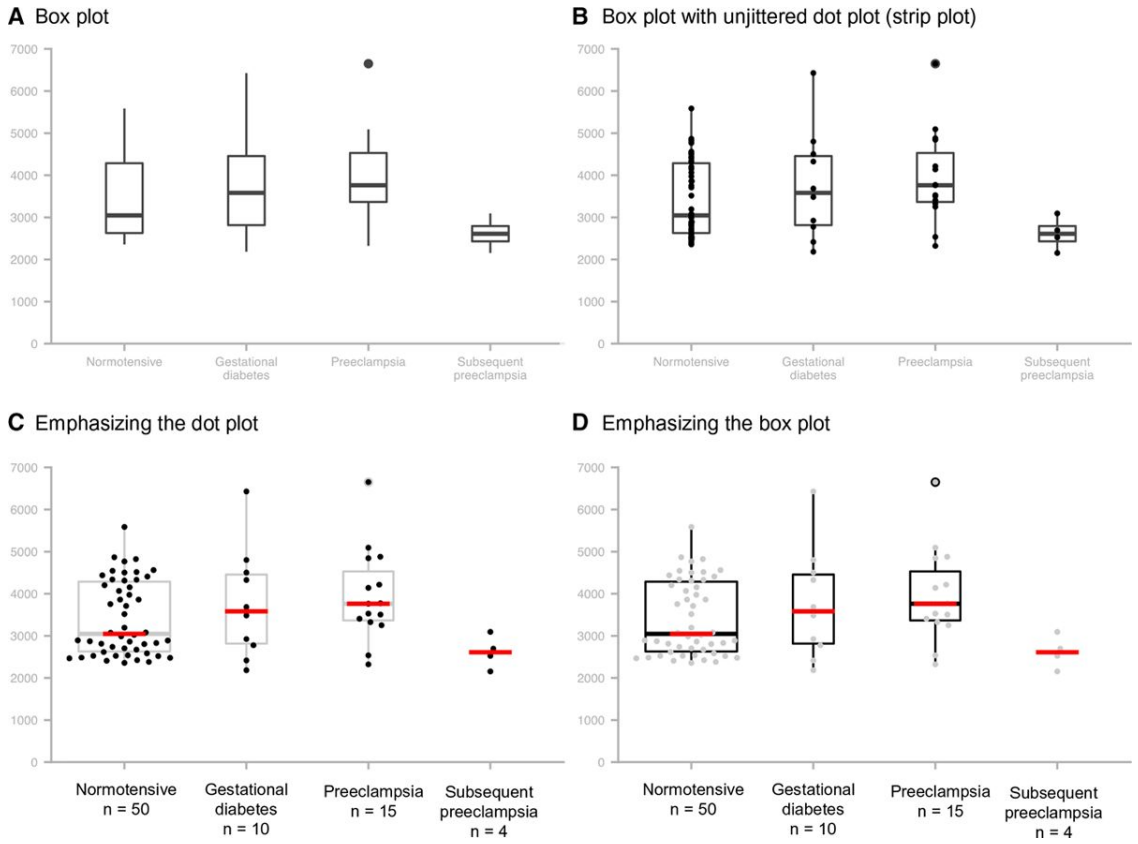
https://ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.037777

Visualization challenges – Different distributions



https://ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.037777

Visualization challenges – Different distributions



https://ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.037777

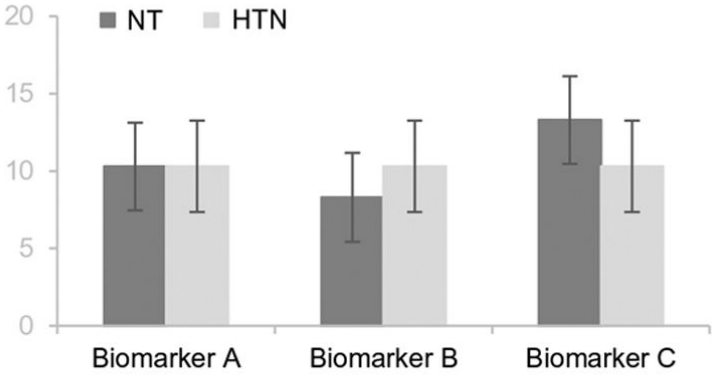
Visualization challenges – Reflecting study design

Experimental goal: Compare normotensive (NT) vs. hypertensive (HTN) patients
Statistical analysis: t-tests were used to compare values for each dependent variable (biomarker A, B and C)

A

Sending mixed messages

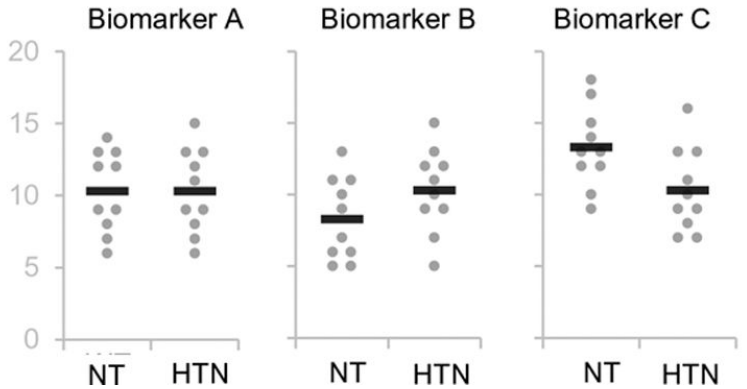
Figure structure erroneously suggests that authors also intended to compare biomarkers A, B and C



B

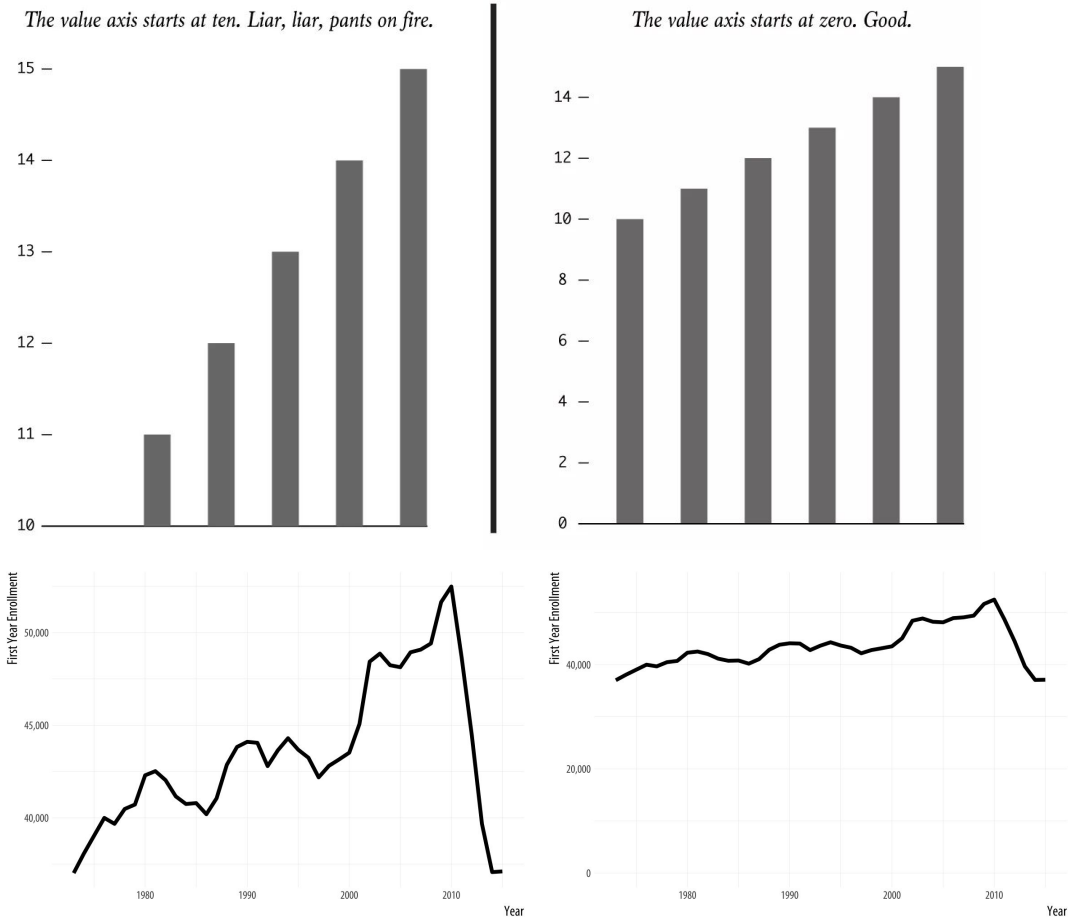
Clear communication

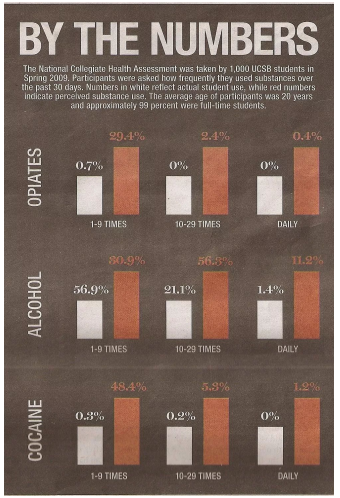
Figure structure matches study design & analysis, shows that the authors did not intend to compare biomarkers



https://ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.037777

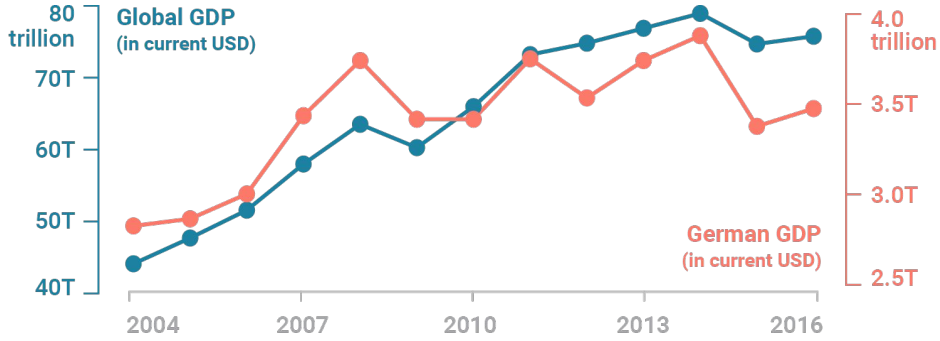
Visualization challenges – Truncated axis

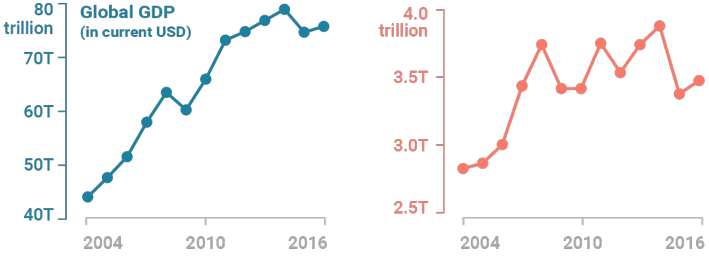


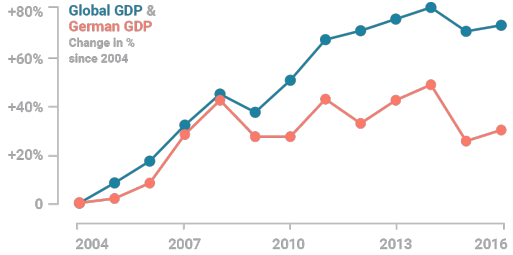


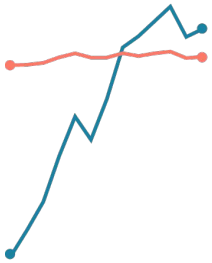
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/
http://socviz.co/lookatdata.html

Visualization challenges – Dual axes

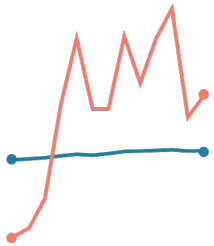




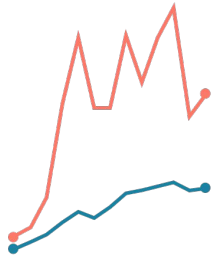




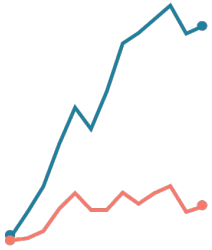
Orange steady,
Blue massively increasing.



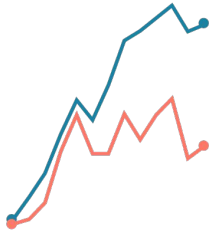
Blue steady,
Orange increasing.



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

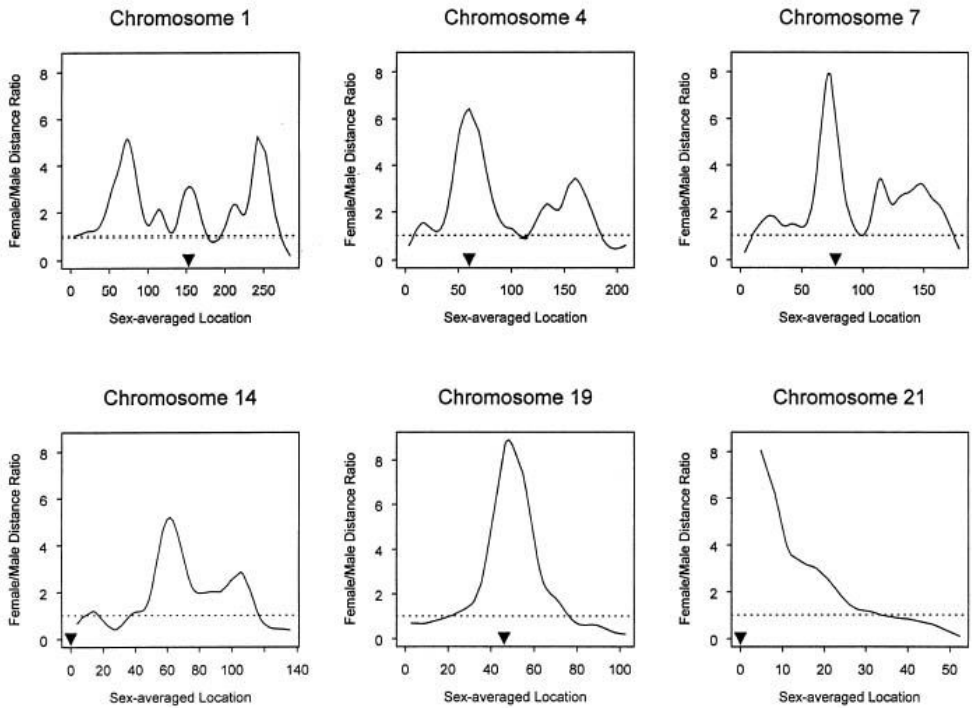
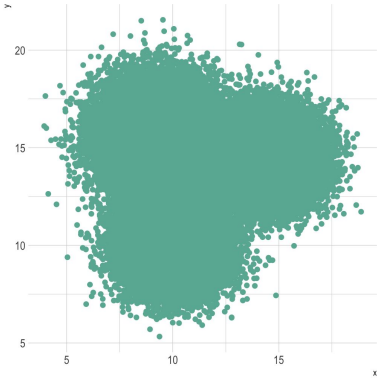
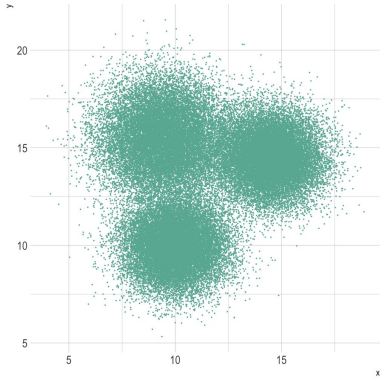


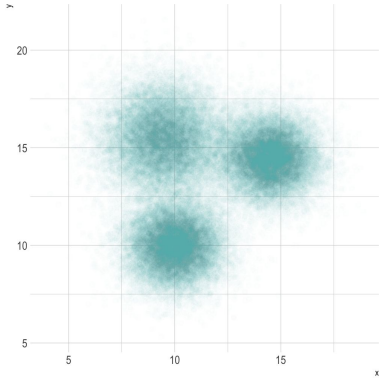
Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

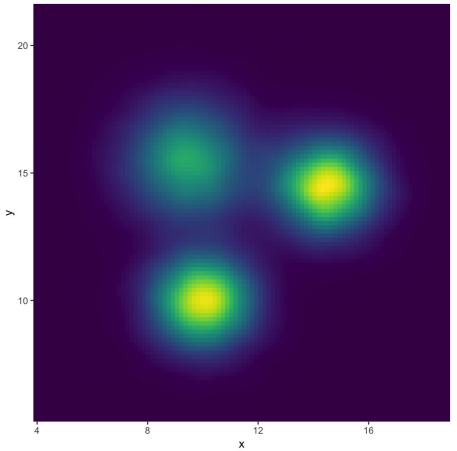
http://www.ncbi.nlm.nih.gov/pubmed/9718341

Visualization challenges – Overplotting



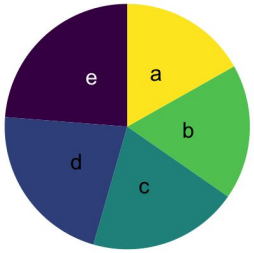


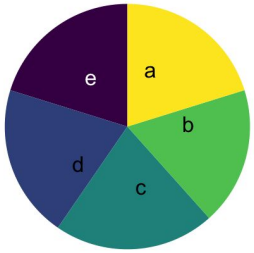


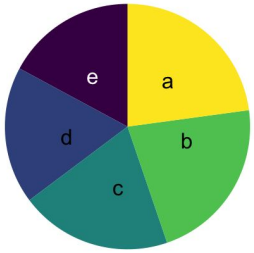


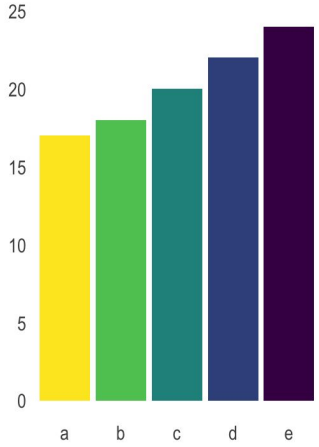
https://www.data-to-viz.com/caveat/

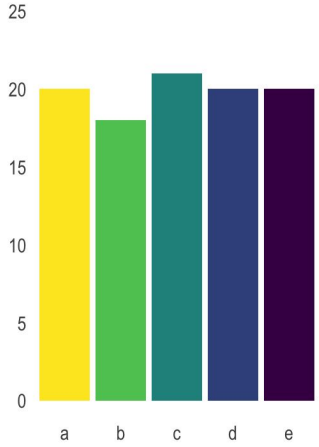
Visualization challenges – Pie charts - almost never a good idea

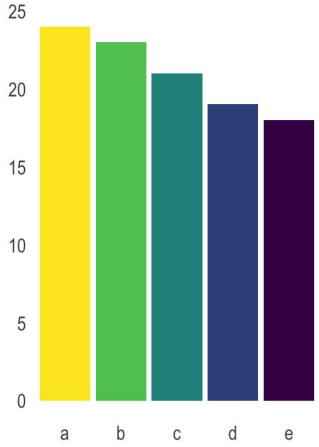






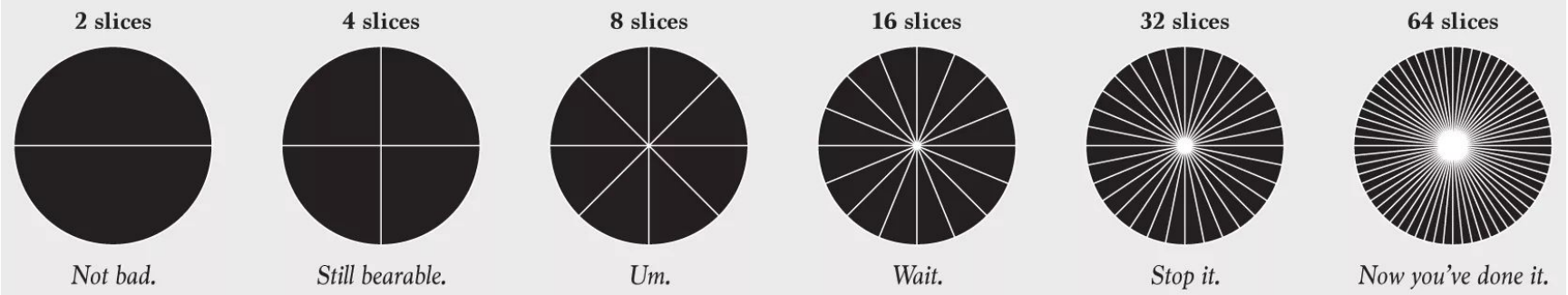


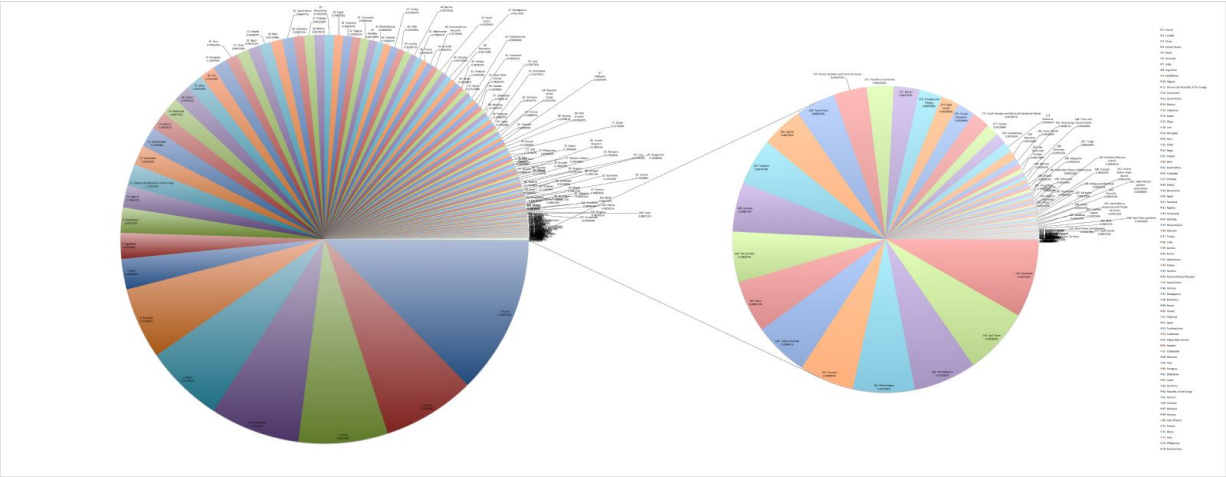




https://www.data-to-viz.com/caveat/

Visualization challenges – Pie charts - almost never a good idea

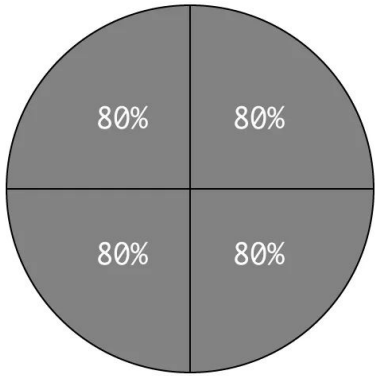


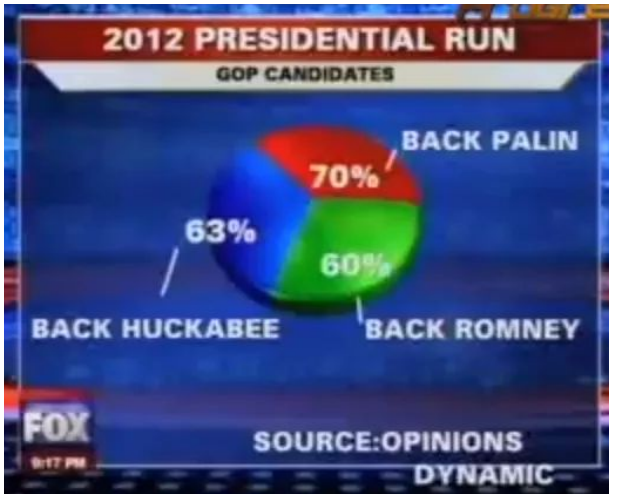


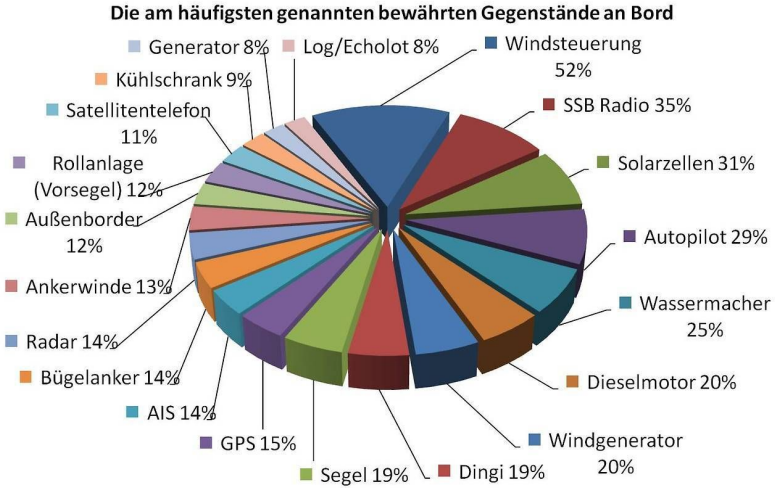
https://flowingdata.com/2015/08/11/real-chart-rules-to-follow/

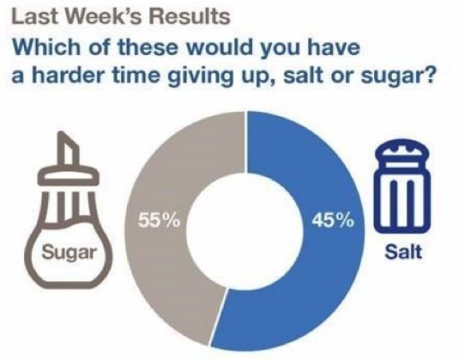
https://commons.wikimedia.org/wiki/File:Pie_chart_of_countries_by_area.png

Visualization challenges – When things don't add up



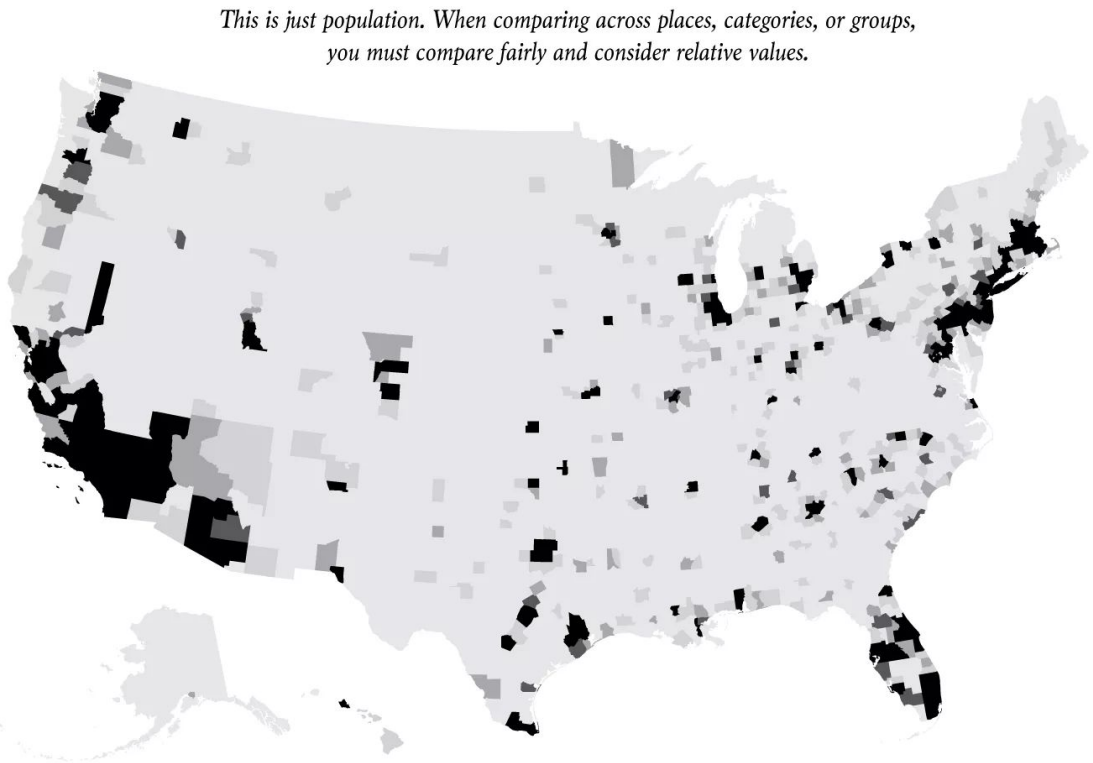


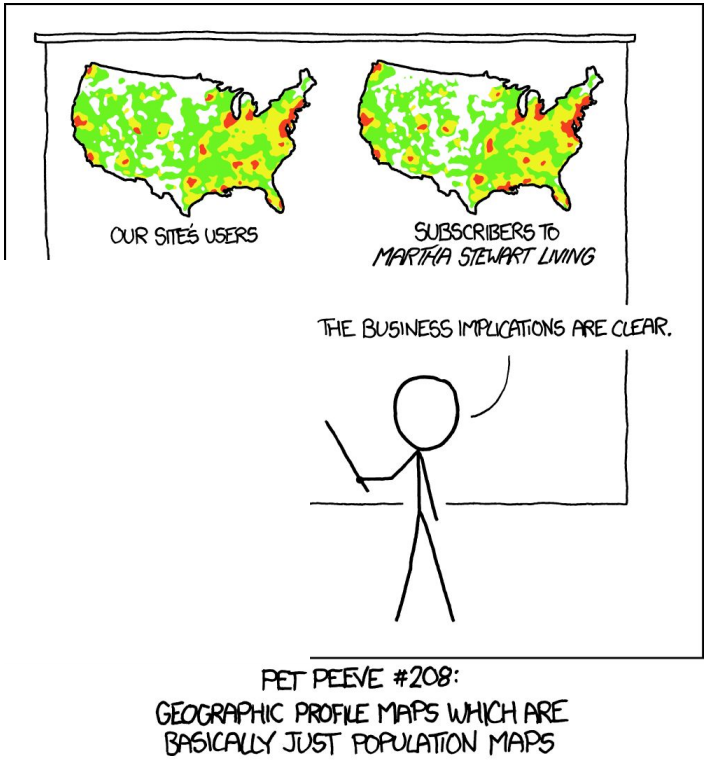




https://www.data-to-viz.com/caveat/
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

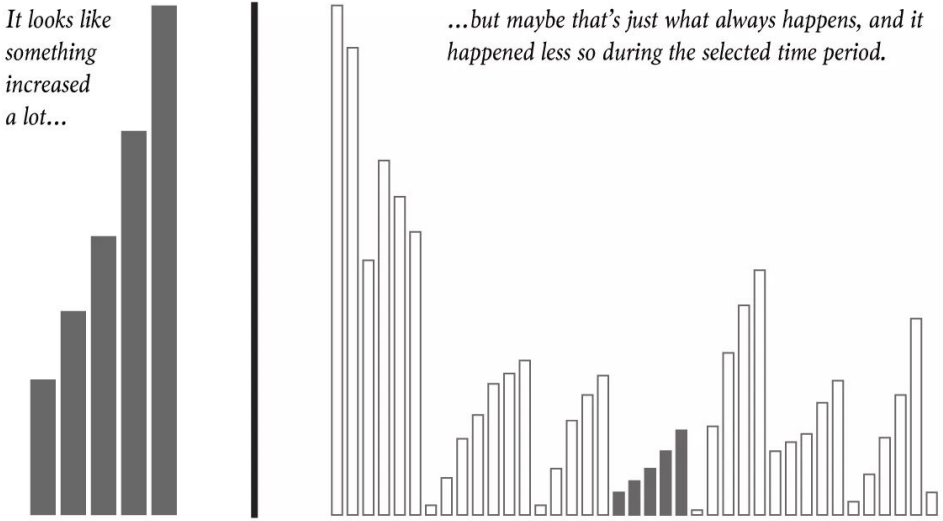
Visualization challenges – Absolutes vs. relative values

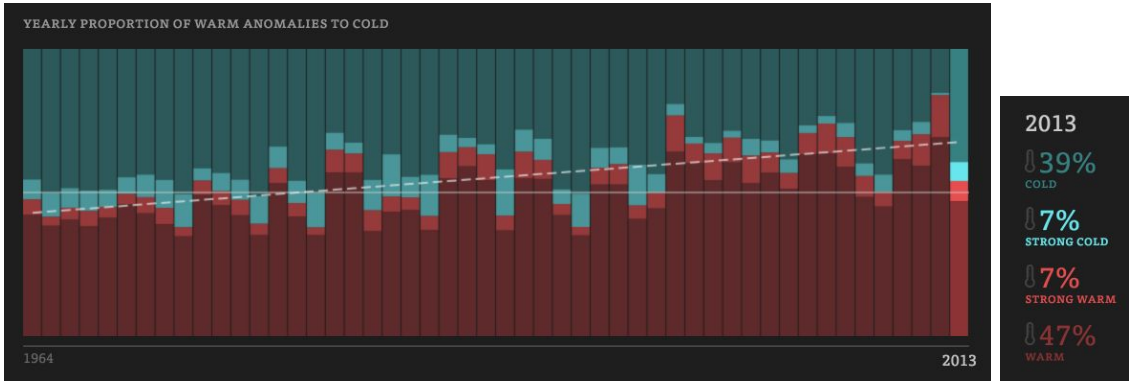




https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Visualization challenges – Limited scope



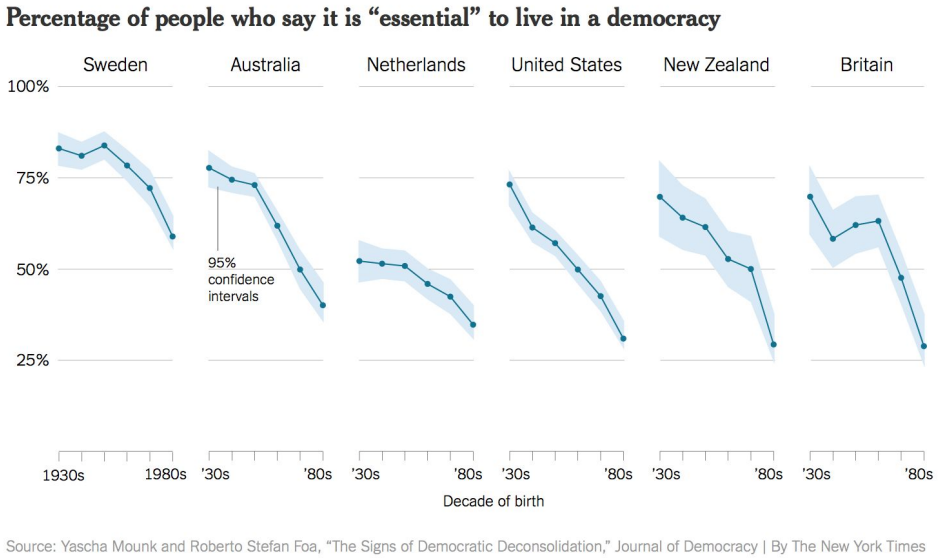


https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

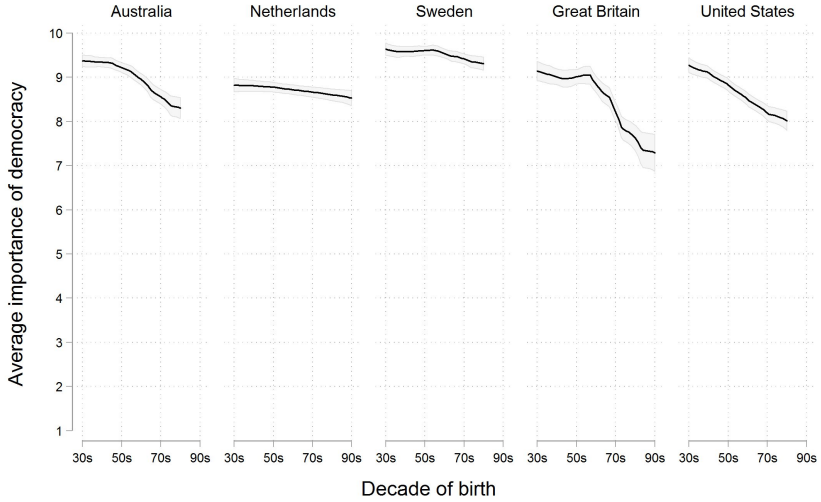
Visualization challenges – Choice of plot & data binning

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



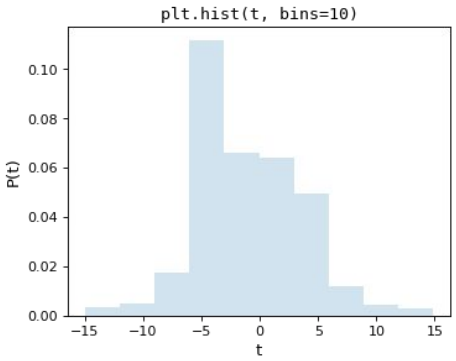
Graph by Erik Voeten, based on WVS 5

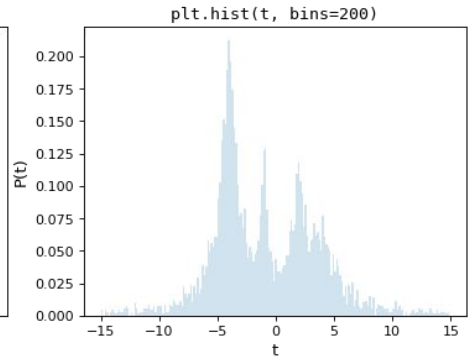
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

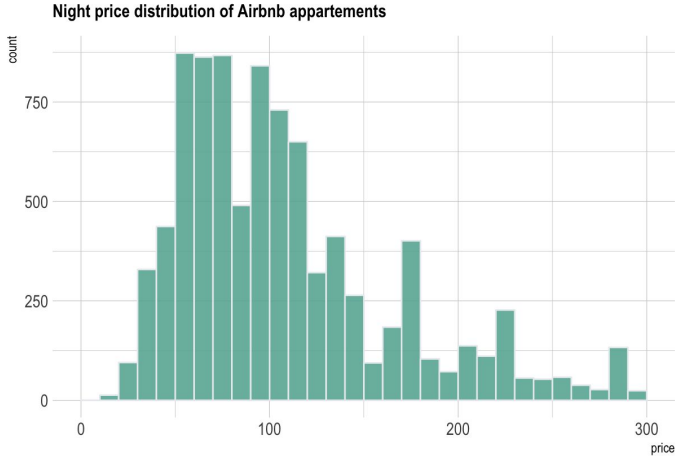
Visualization challenges – Choice of plot & data binning

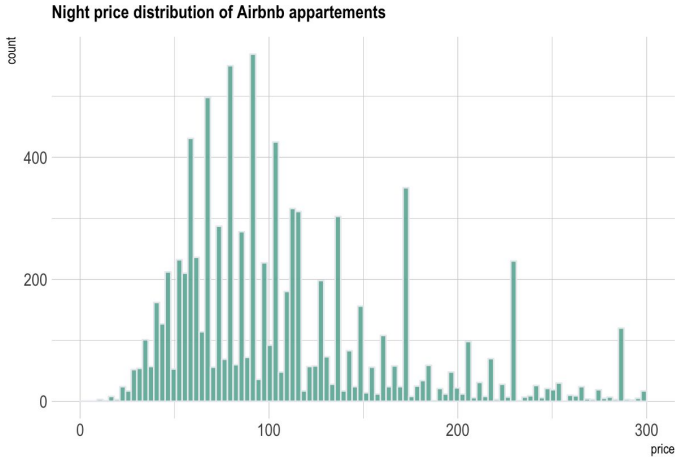
Two bins. What's really in the 1+ category? Might be hiding something.

That's better. It can show more variation.



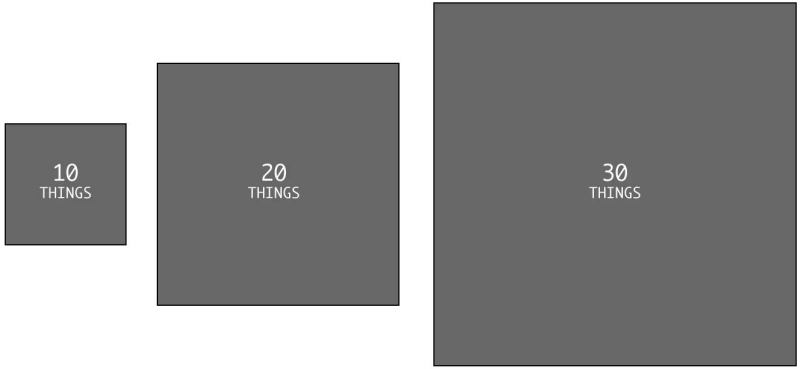


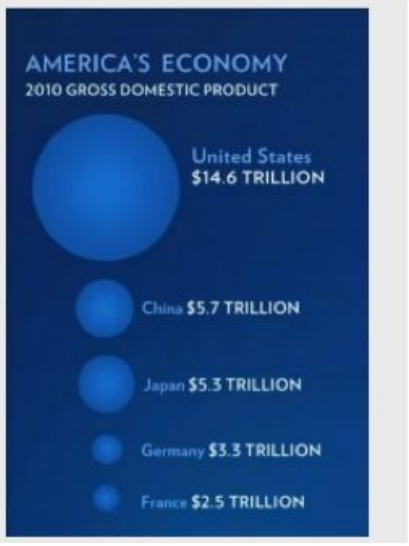


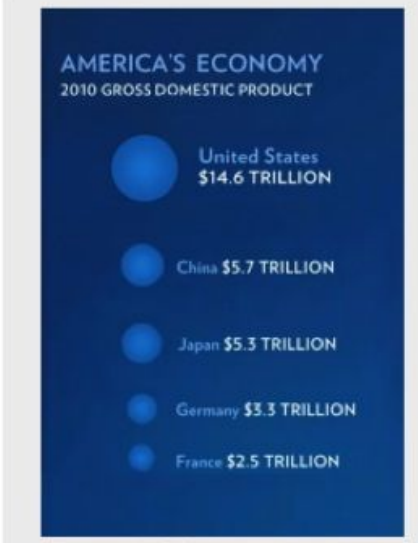


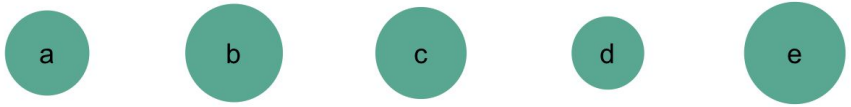
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

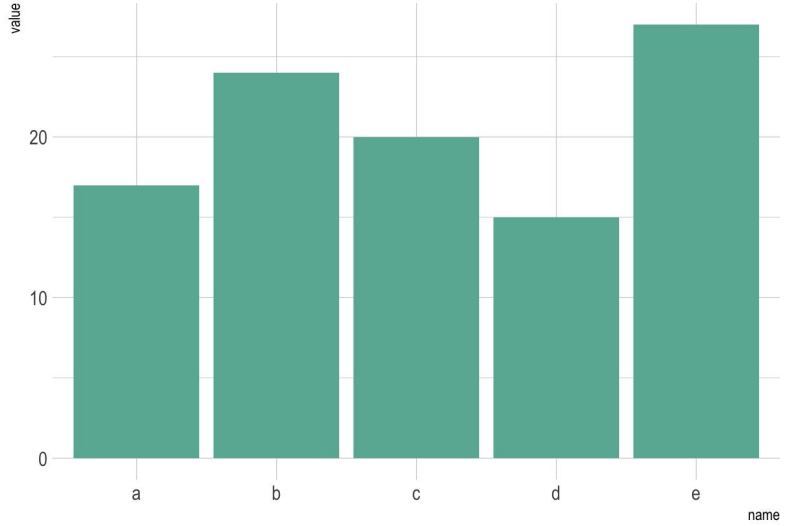
Visualization challenges – Area sized by dimension





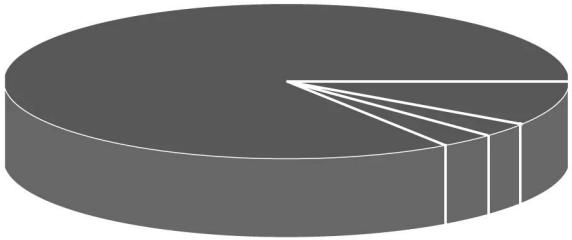






https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Visualization challenges – Unwarranted extra dimensions



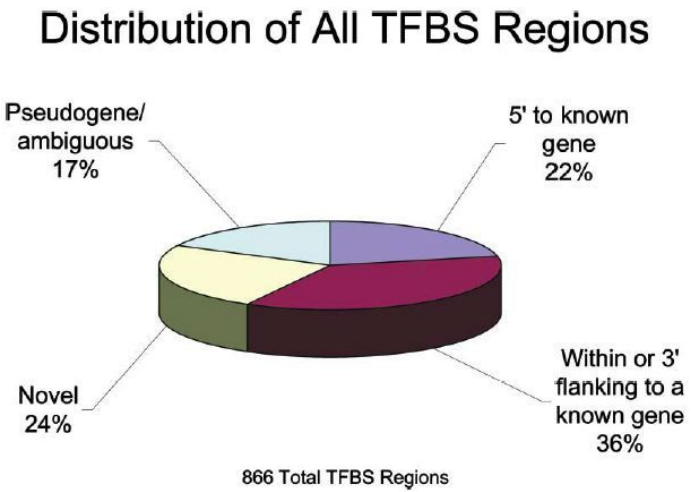
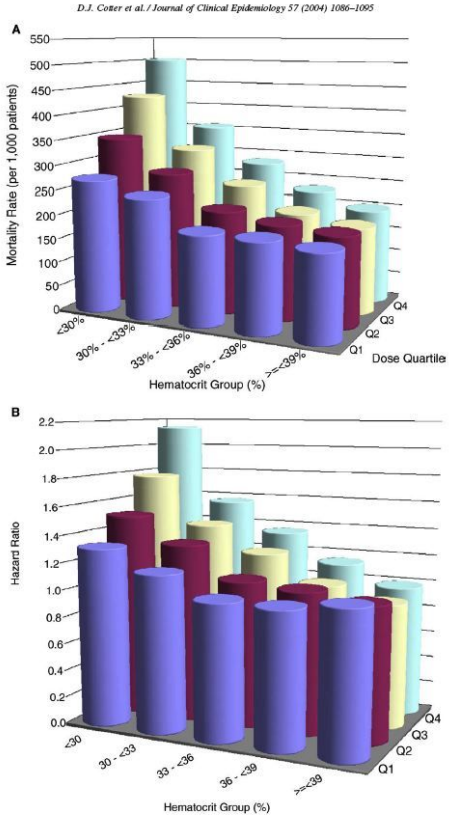


Figure 1. Classification of TFBS Regions

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/

Final exam

Select a primary research article published by you/your-group in the past 3–5 years that:

- Has a specific question (exploratory or hypothesis-driven).
- Contains experimental data (newly generated or previously published).
- Contains a “Methods" section.
- Involves at least **3 different results** – figures/tables in the main paper – obtained based on statistical data analysis. [This is critical.]

[Check with me to make sure the paper has enough statistical stuff to dig in for a final exam!]

Final exam

Carefully read the paper to dissect the statistical analyses and visualization choices that went into each of these results.

Remember that details about a single analysis/result might (unfortunately) be strewn across the entire paper:

- Results section
- Methods section
- Figure/table legend
- Supplementary materials/files, and
- Even older papers cited here!

For each result, take stock of all this information somewhere at least as rough notes, and bring them to the Final exam.

Complete online evaluation

https://sirsonline.msu.edu

I will also send you a survey.