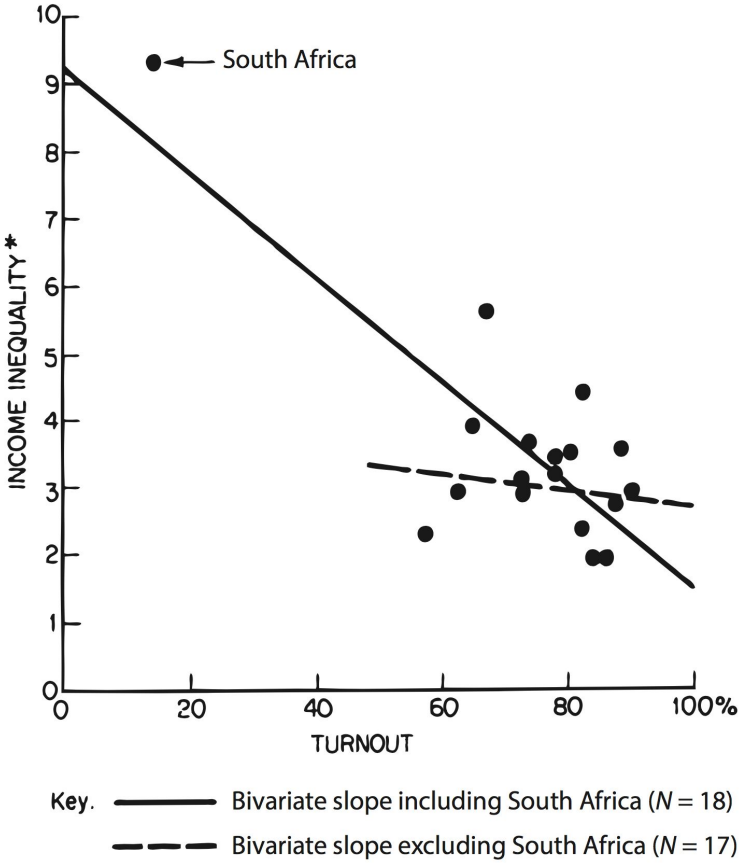
Primer 3: Presenting data & results in a computational biology project

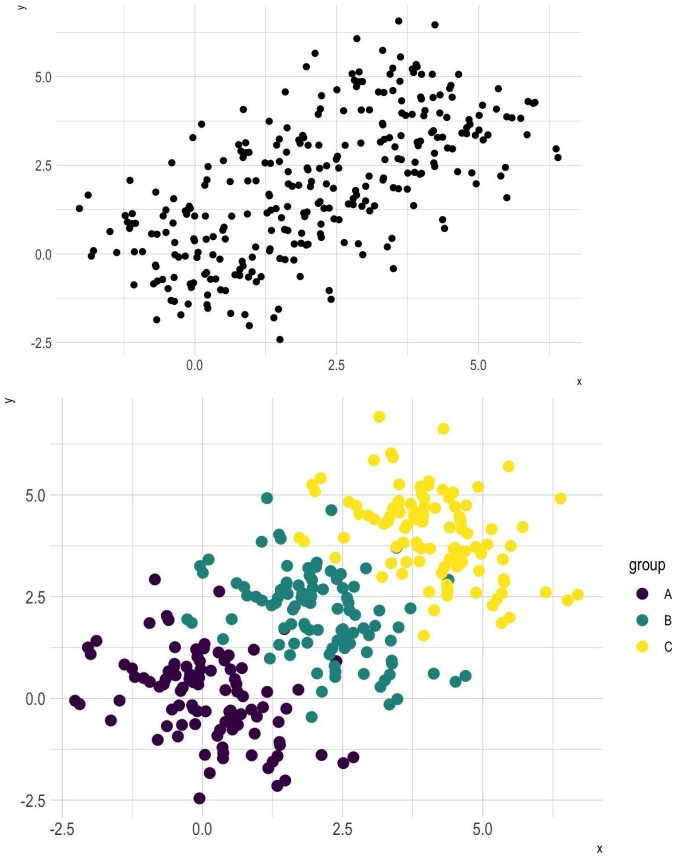
- The purpose of data visualization
- Choice of visualization
- Typical issues in aesthetics & data handling
- Improving visualization for clarity
- How to write a paper

Purpose of data visualization

- 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.
- A critical component of your analysis and research!

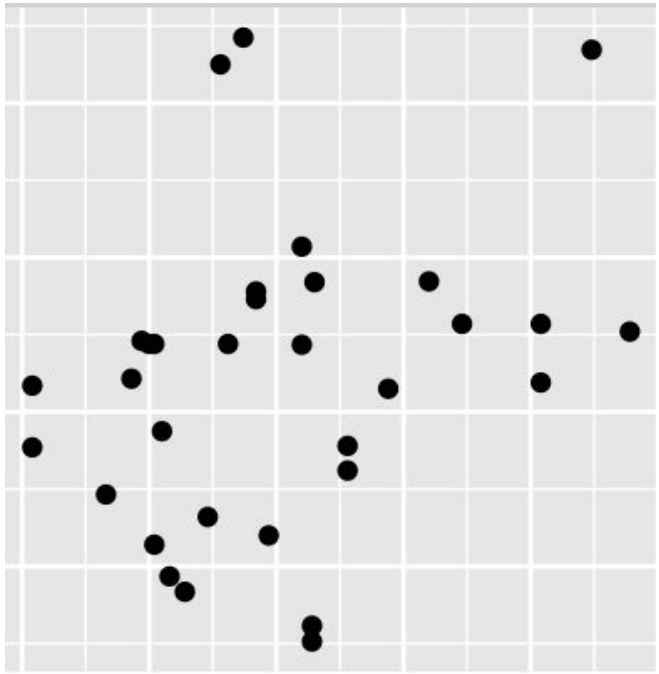
Don’t do statistics without visualization





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

Visualization is a rigorous tool for inference

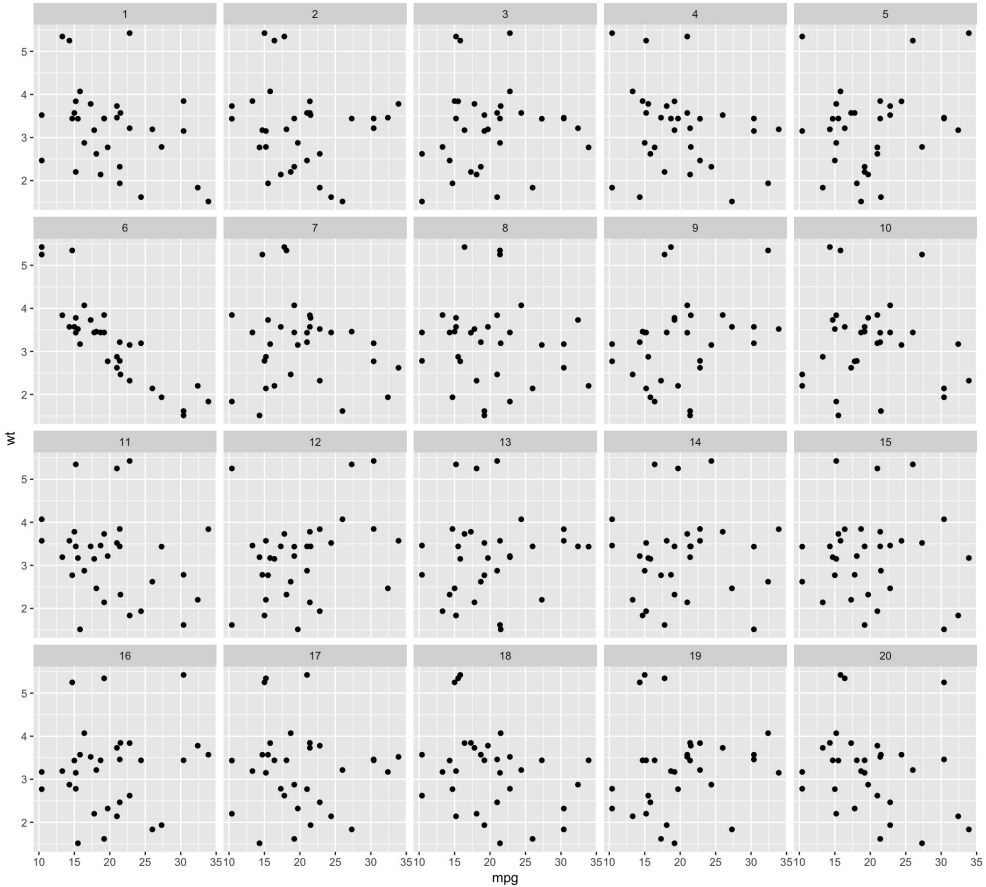


Create a lineup for visual inference

- 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 you 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

Visualization is a rigorous tool for inference

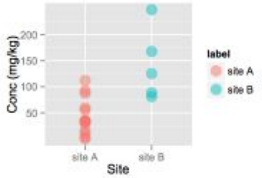
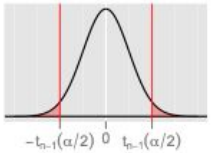
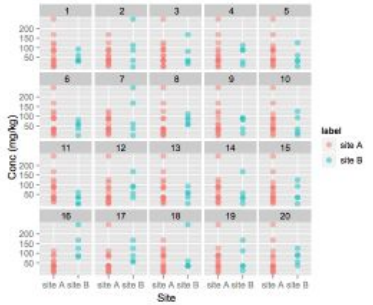


Create a lineup for visual inference

- 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 you 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

Visualization is a rigorous tool for inference

	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

Choice of visualization

- Choosing among many options
- Barplots
- Boxplots

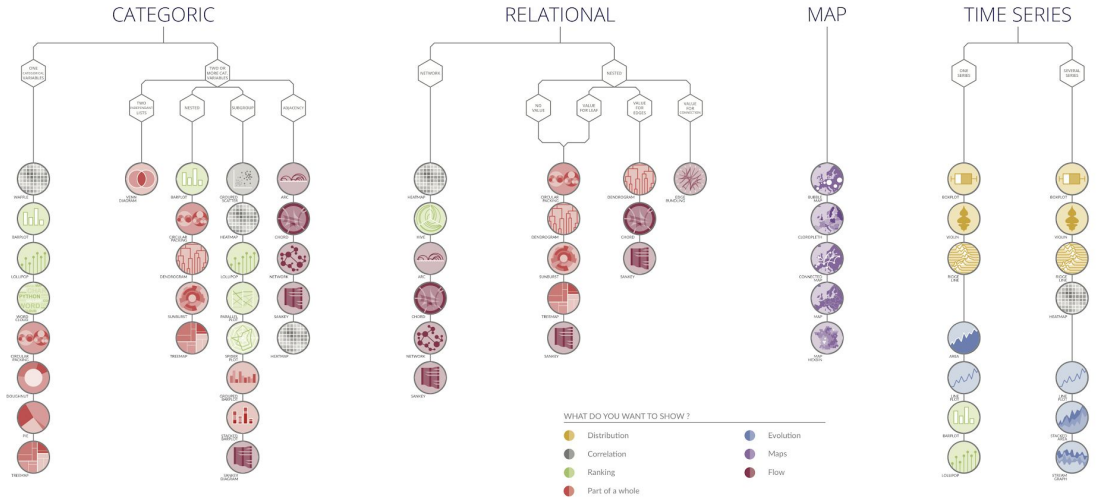
from Data to Viz

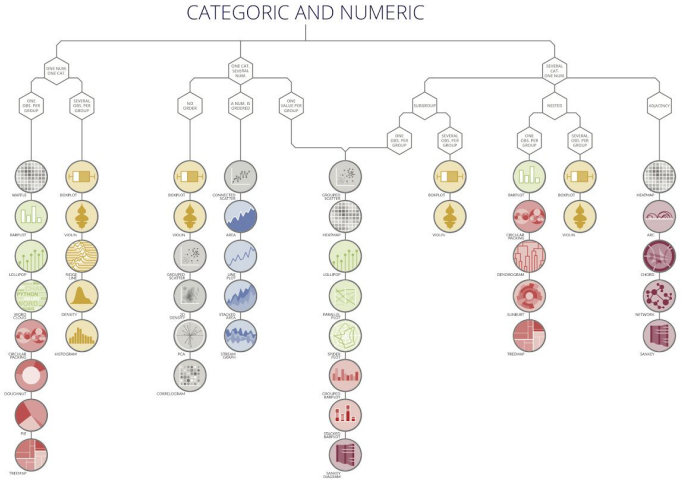
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

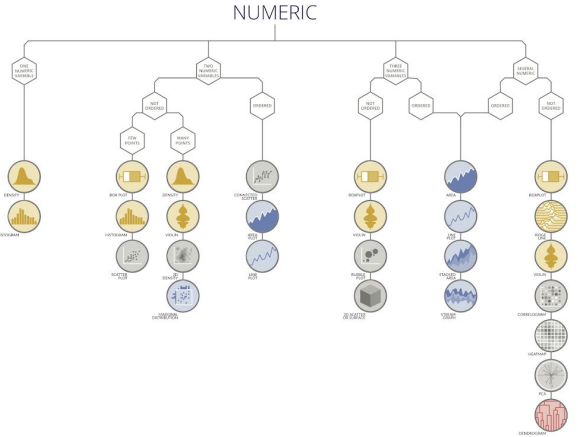
1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

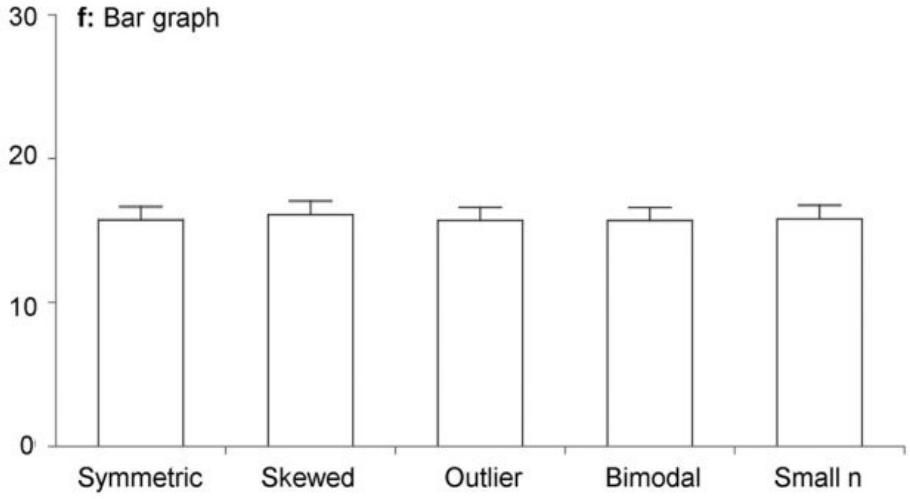
data-to-viz.com

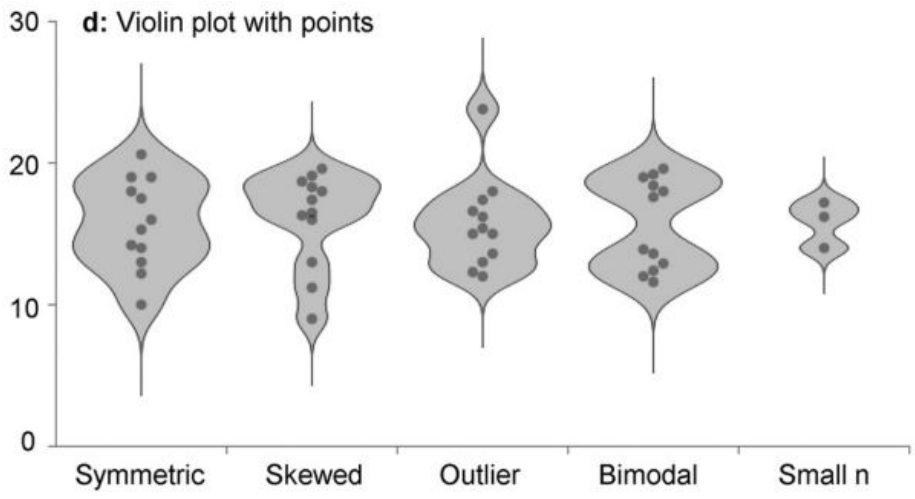






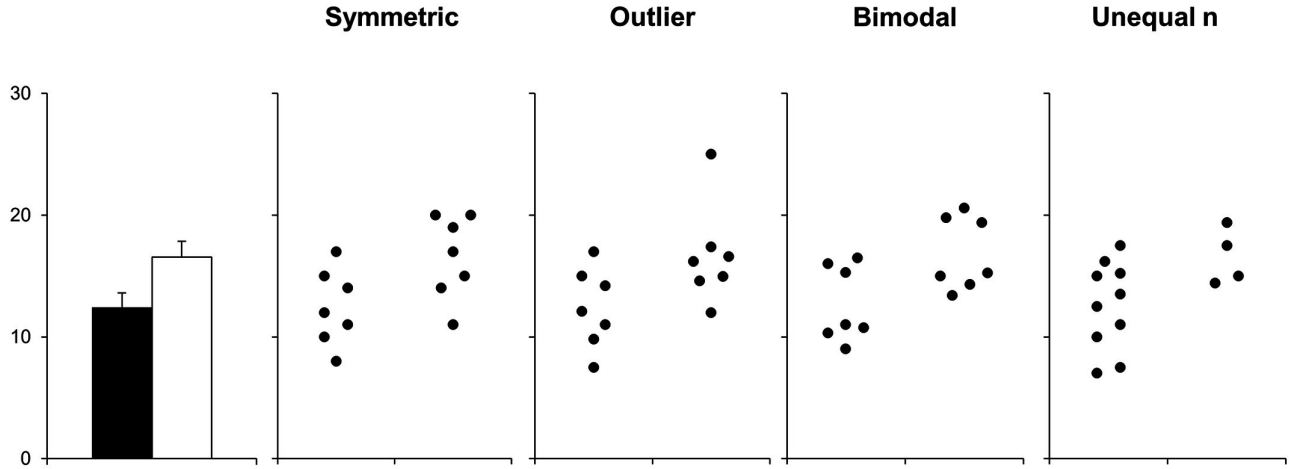
Bar plots w/ error bars – Almost never a good idea





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

Bar plots w/ error bars – Almost never a good idea



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

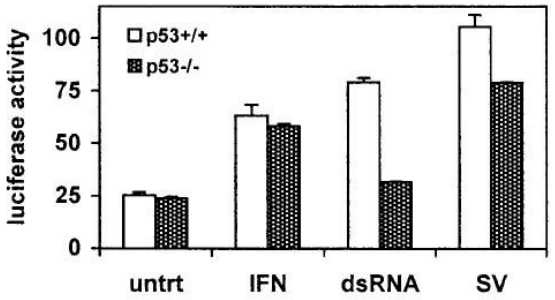


FIG. 4. ISG15 promoter activity mimics endogenous ISG15 mRNA regulation by p53, dsRNA, and virus. Cells (6×10^5 HCT 116) were

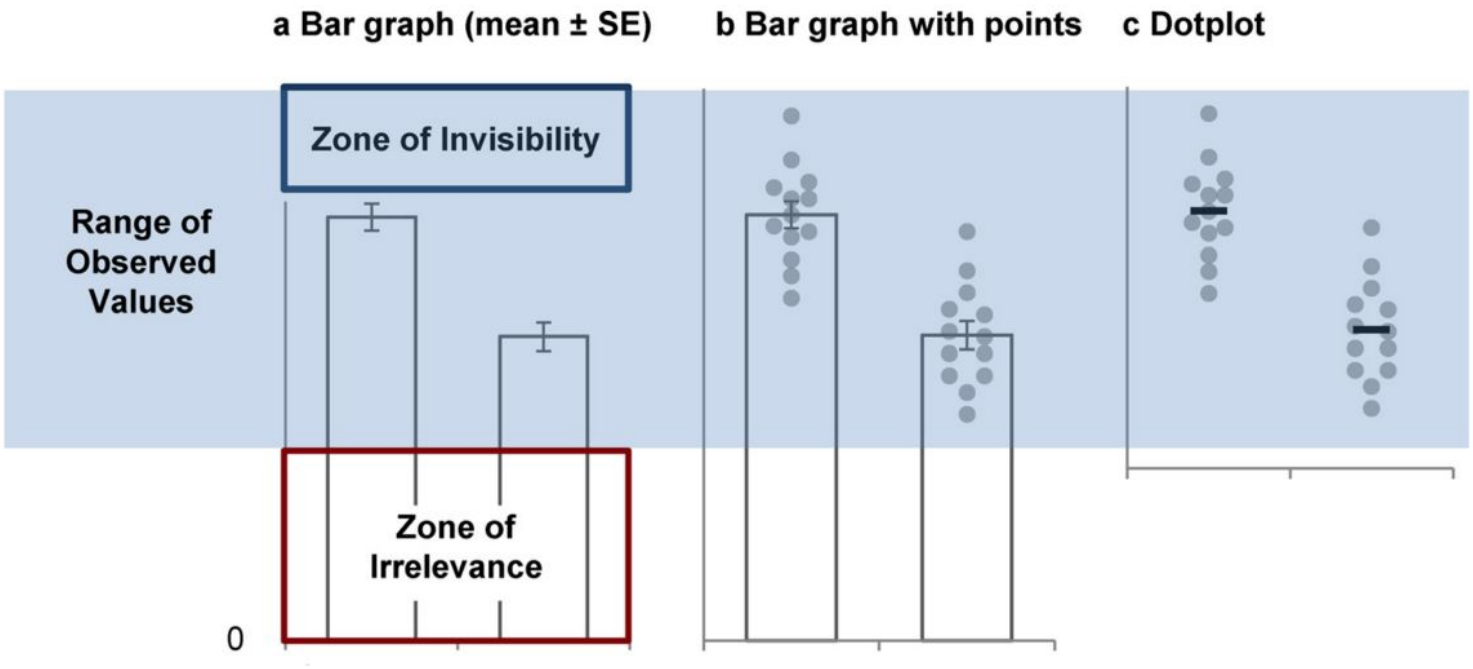
differences in transfection efficiency. Each data point is the mean of triplicate samples ± the standard error; the data presented are repre-

Parametric vs.

Non-parametric test

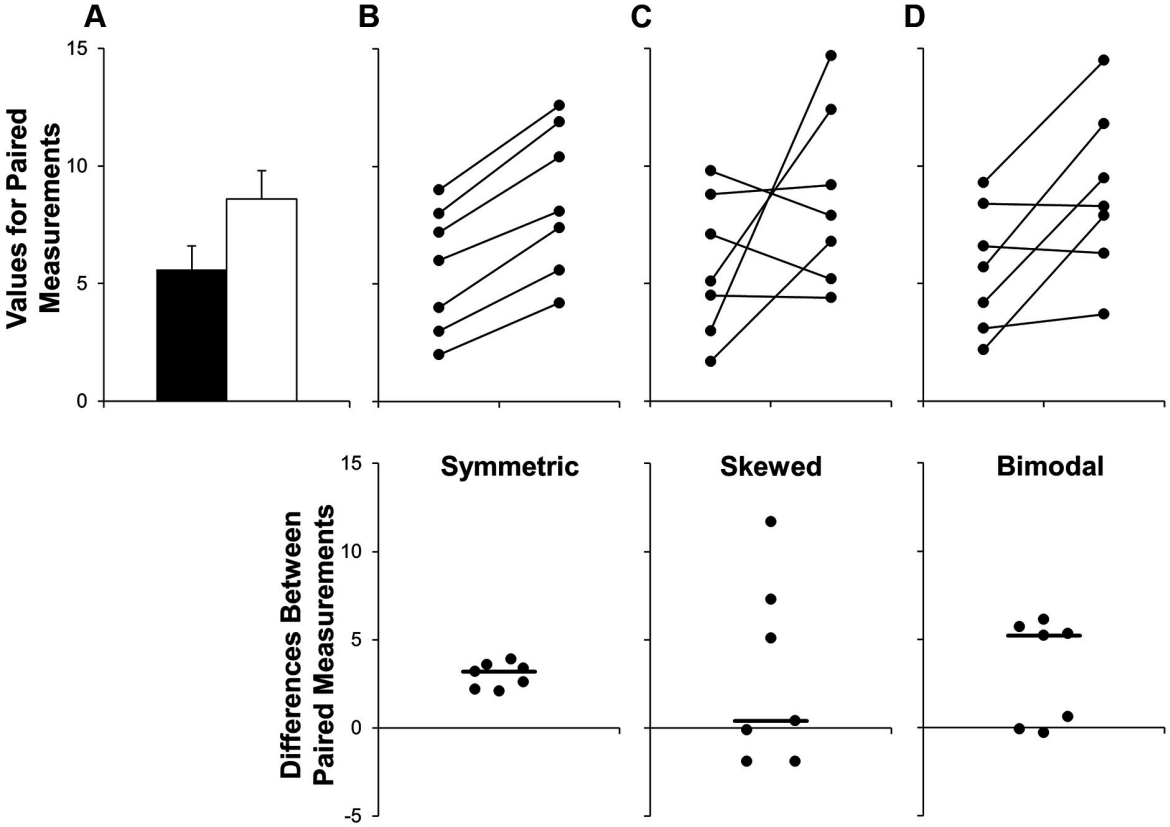
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

Bar plots w/ error bars – Almost never a good idea



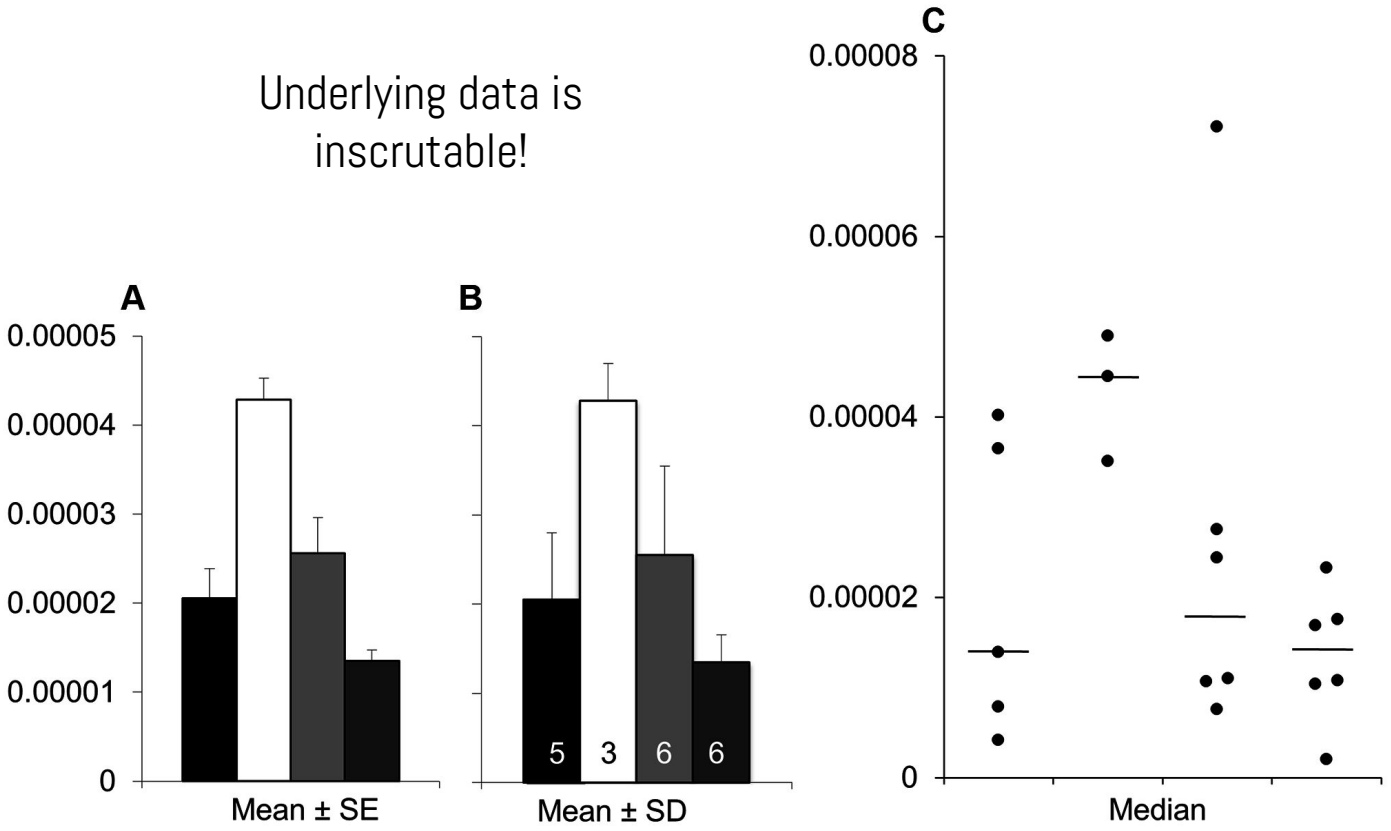
http://www.jbc.org/content/292/50/20592.full

Bar plots w/ error bars – Almost never a good idea



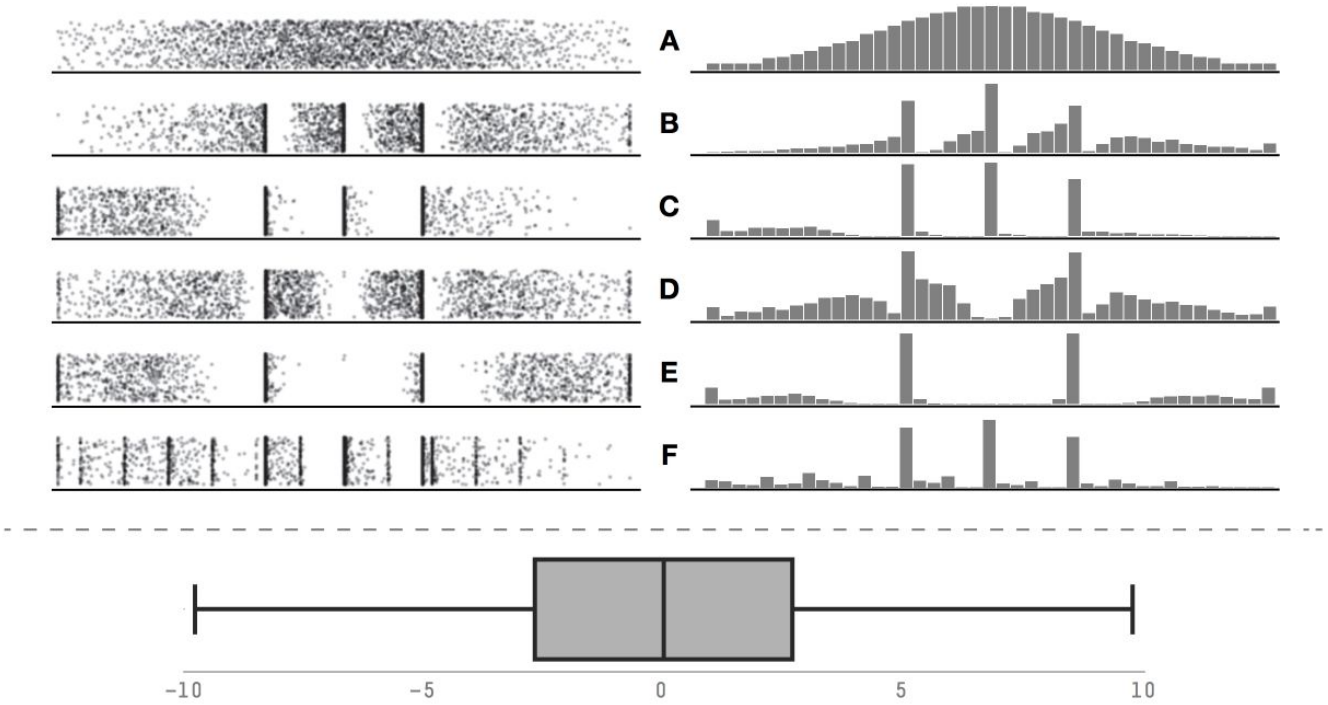
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

Bar plots w/ error bars – Almost never a good idea



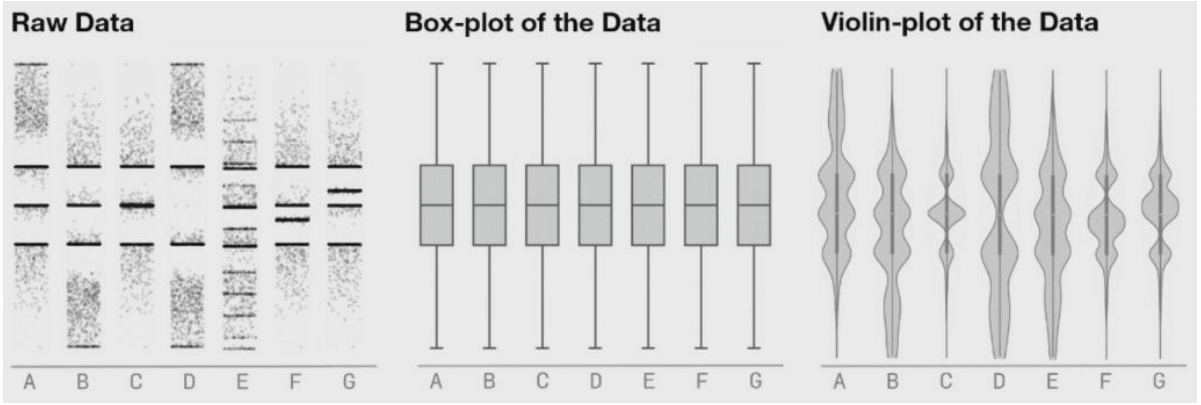
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002128

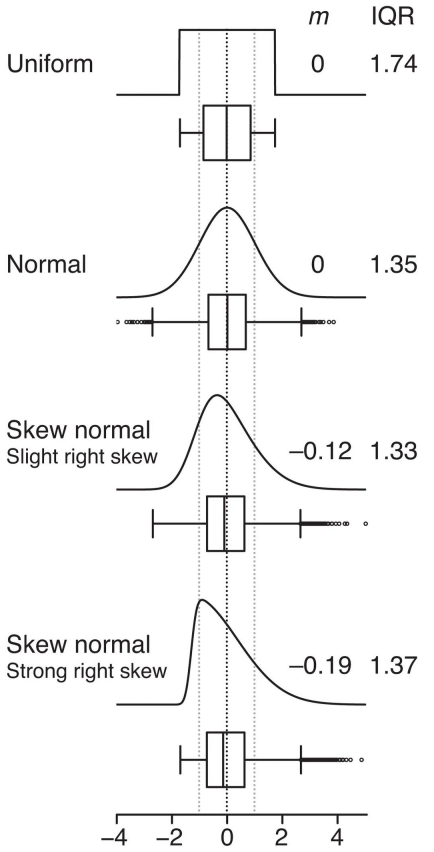
Boxplots can confound different distributions



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

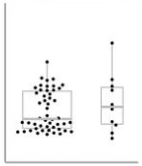
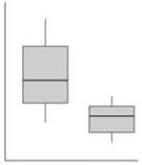
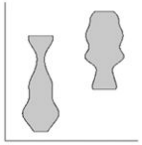
Boxplots can confound different distributions





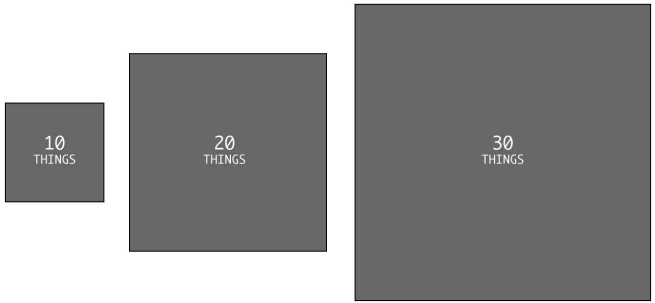
https://www.autodeskresearch.com/sites/default/files/SameStats-DifferentGraphs.pdf

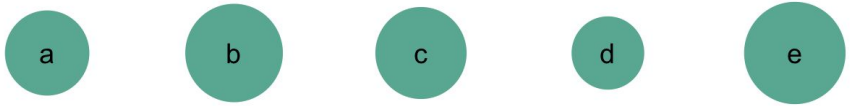
Pick the plot to show different distributions

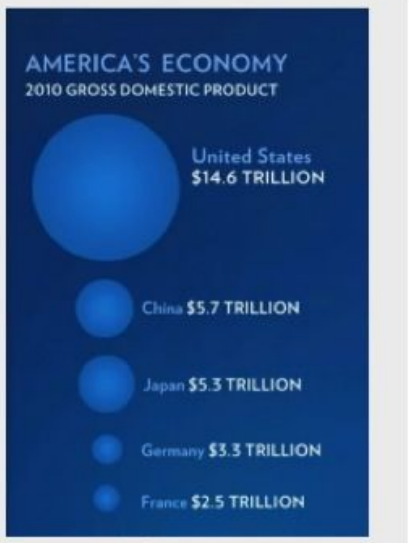
Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	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**

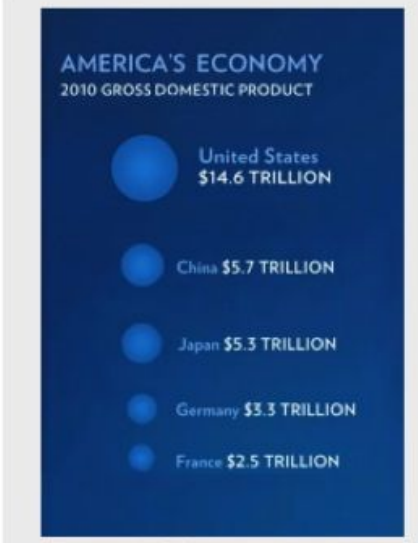
https://ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.037777

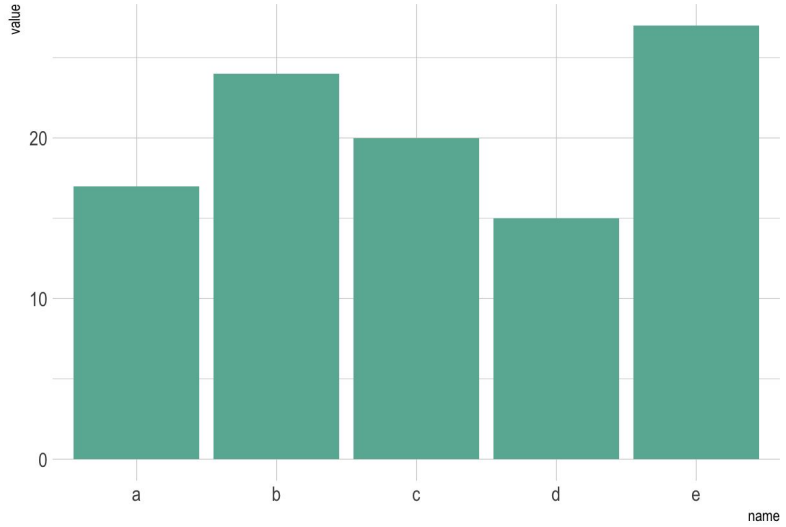
Area is a poor choice for dimension





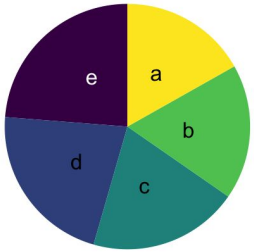


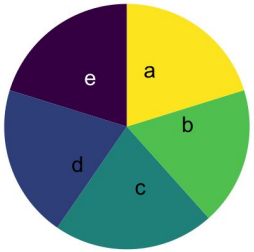


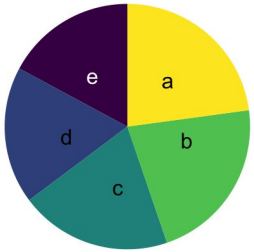


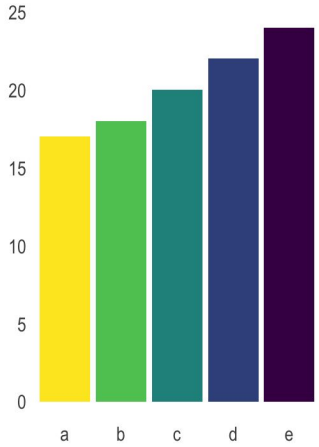
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

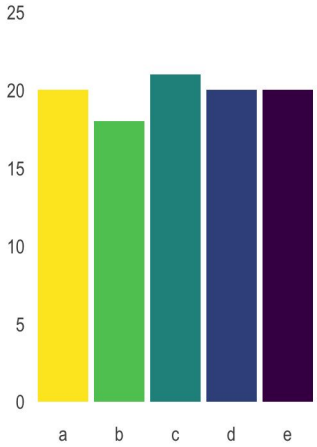
Pie charts – almost never a good idea

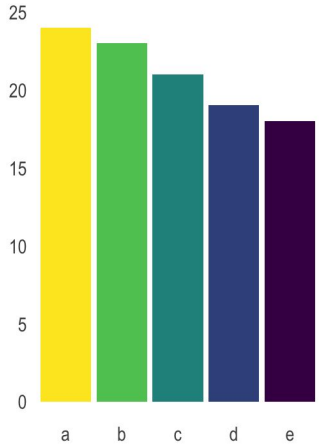






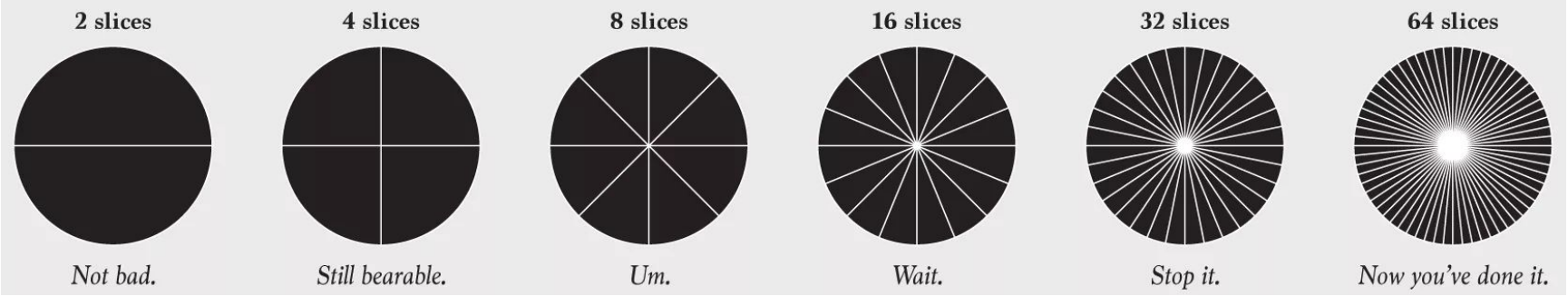


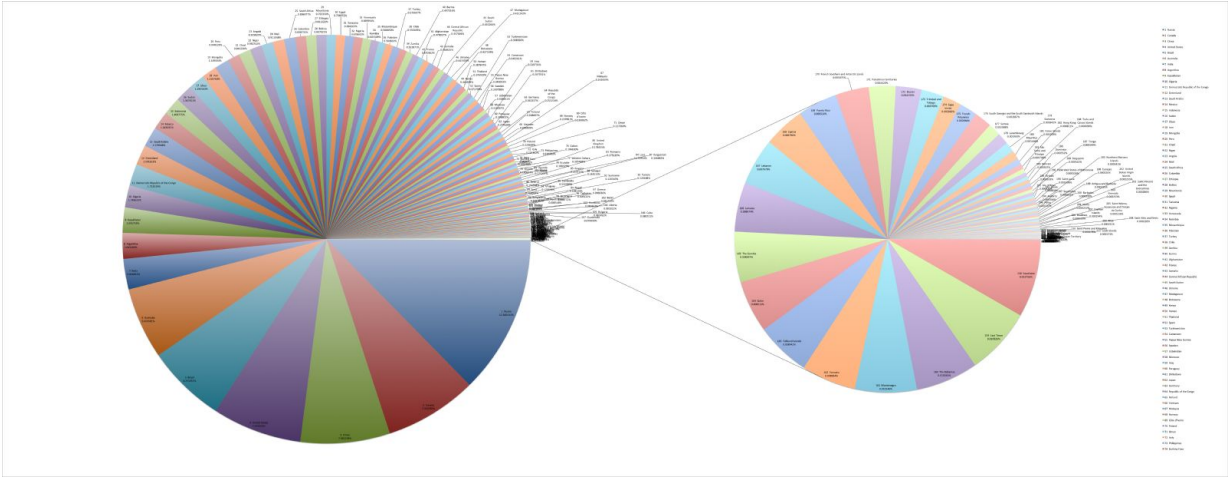




https://www.data-to-viz.com/caveat/

Pie charts – almost never a good idea

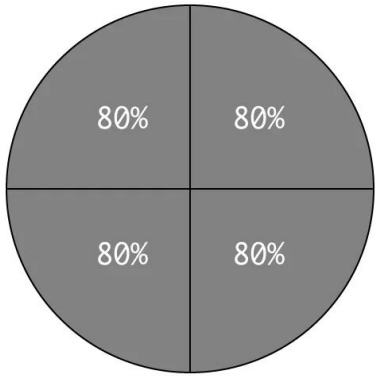


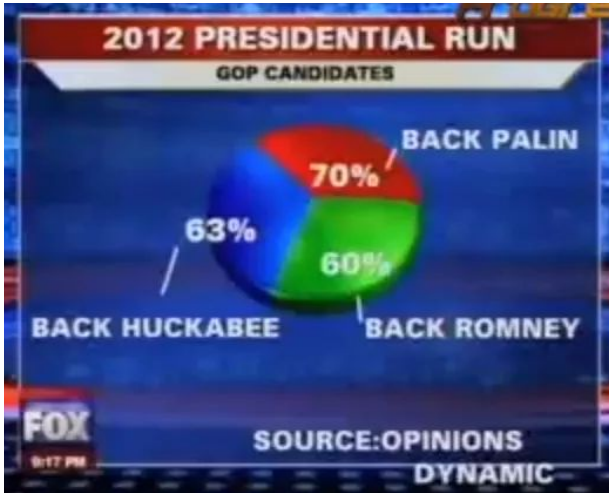


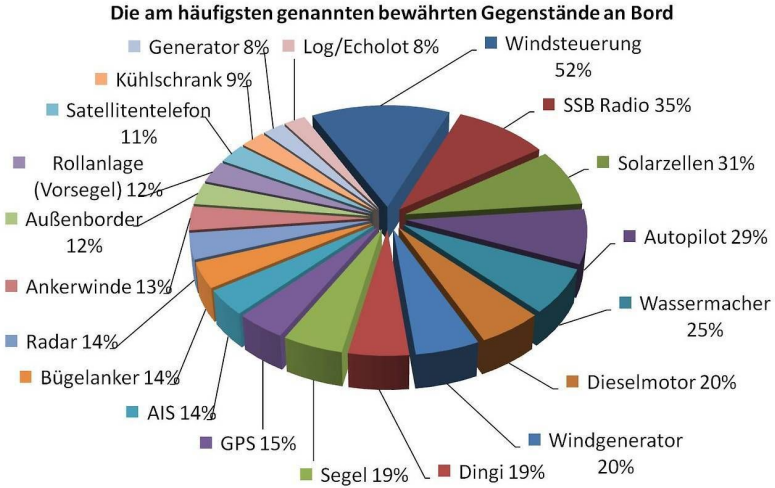
https://flowingdata.com/2015/08/11/real-chart-rules-to-follow/

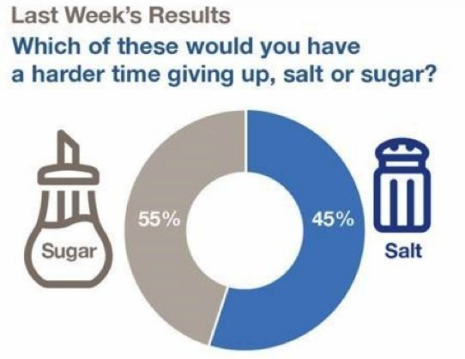
https://commons.wikimedia.org/wiki/File:Pie_chart_of_countries_by_area.png

… especially when things don't add up!



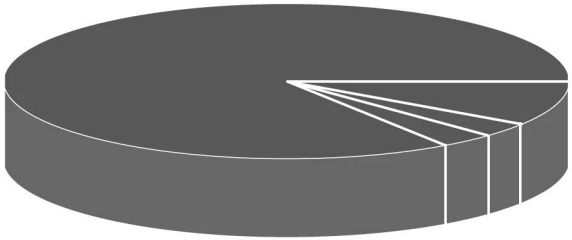






https://www.data-to-viz.com/caveat/
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

No 3D!



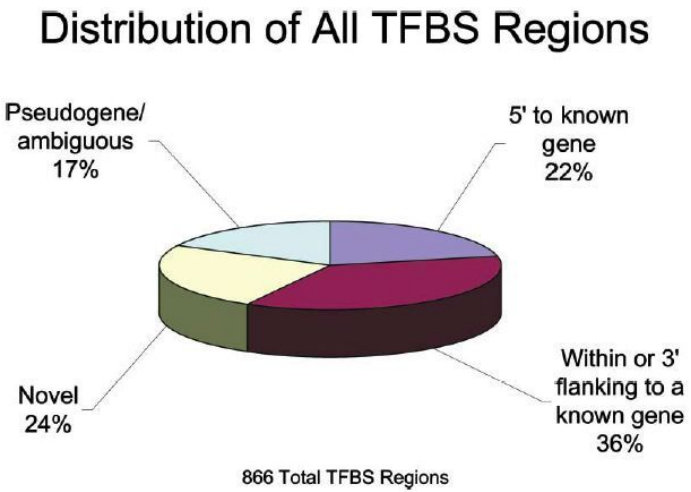
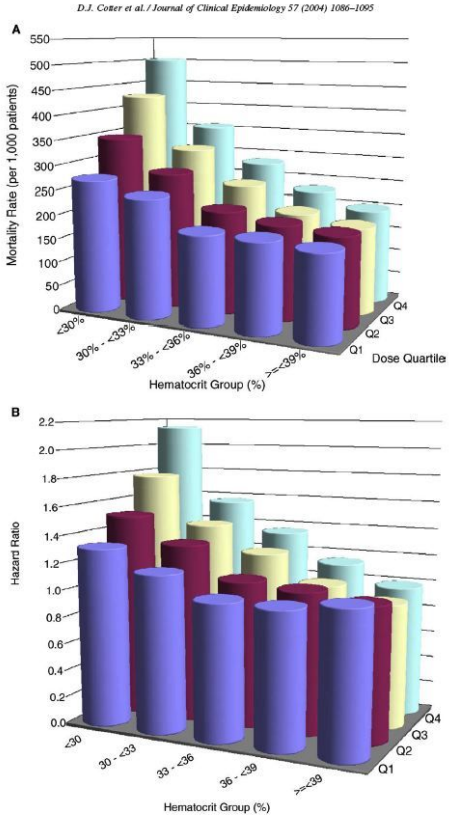


Figure 1. Classification of TFBS Regions

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).

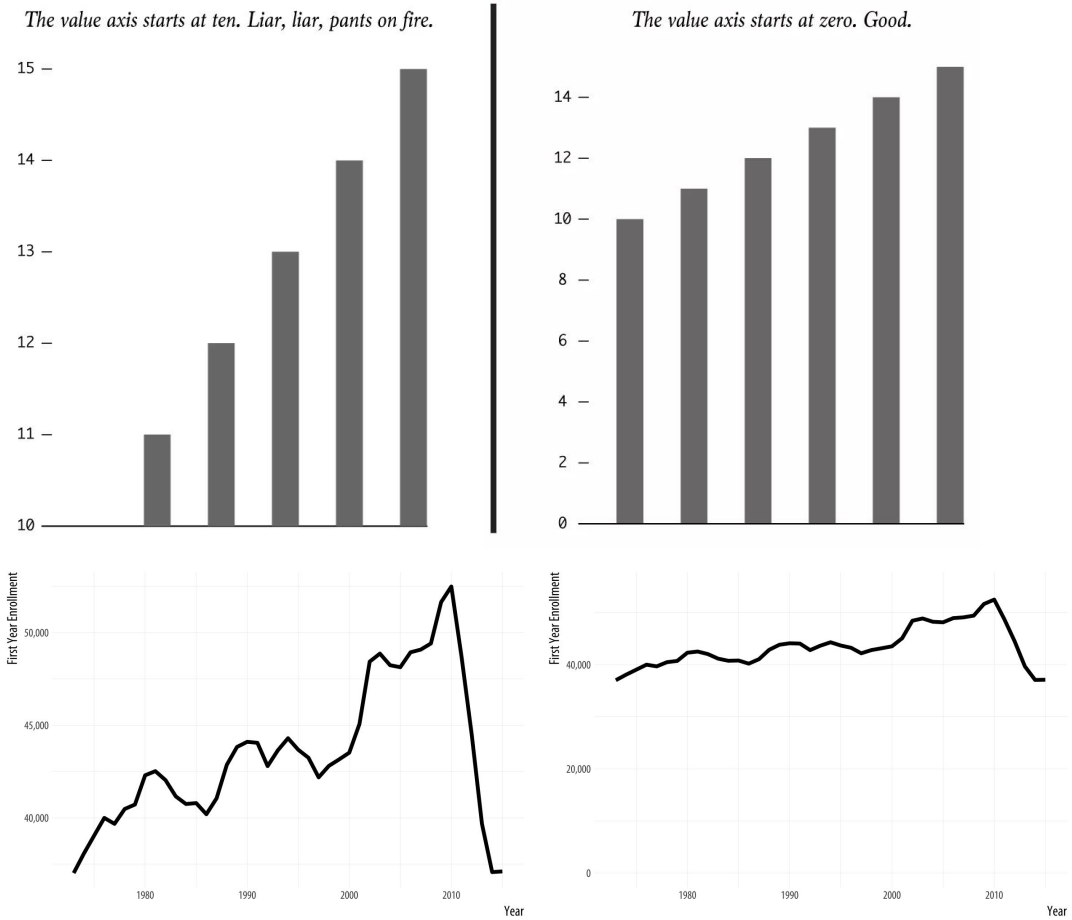


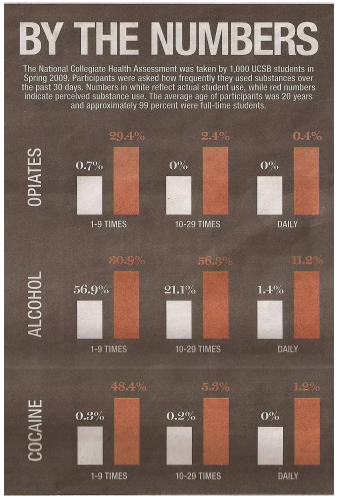
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Typical issues in aesthetics & data handling

- Plot axes
- Data scales
- Data scope
- Data binning

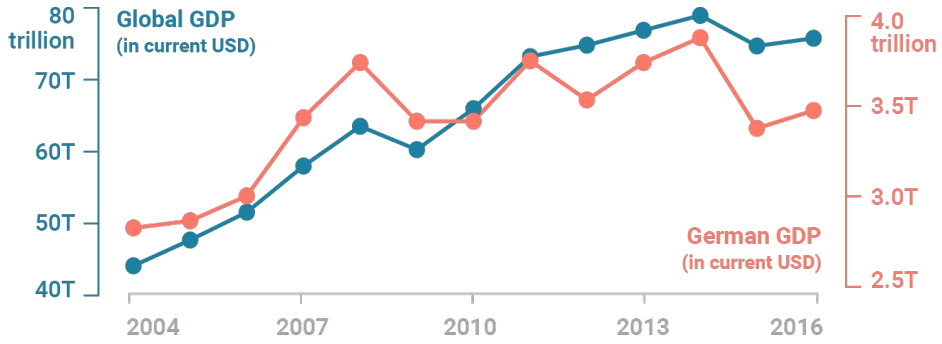
Avoid truncated axis

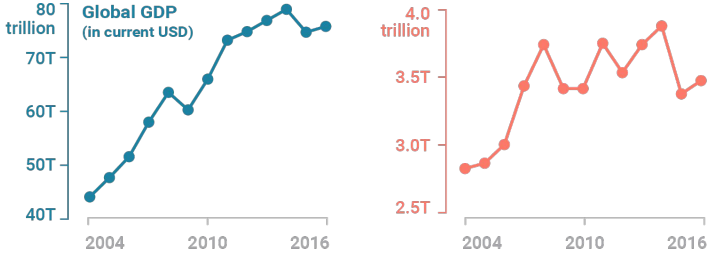


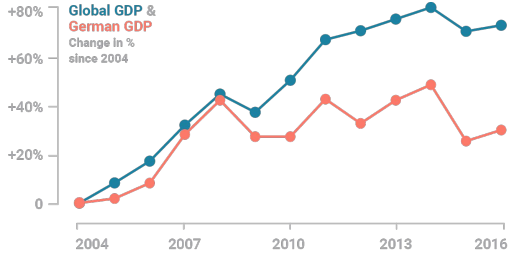


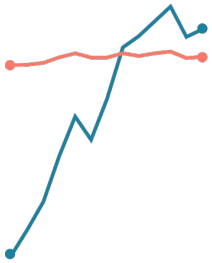
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/
http://socviz.co/lookatdata.html

No dual axes

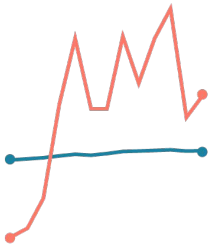




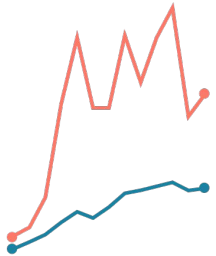




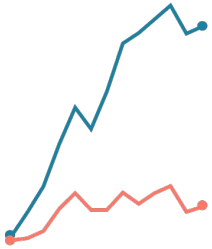
Orange steady,
Blue massively increasing.



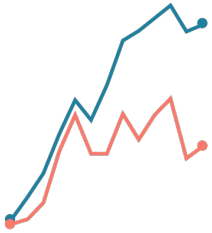
Blue steady,
Orange increasing.



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/

Pay attention to inappropriate axes

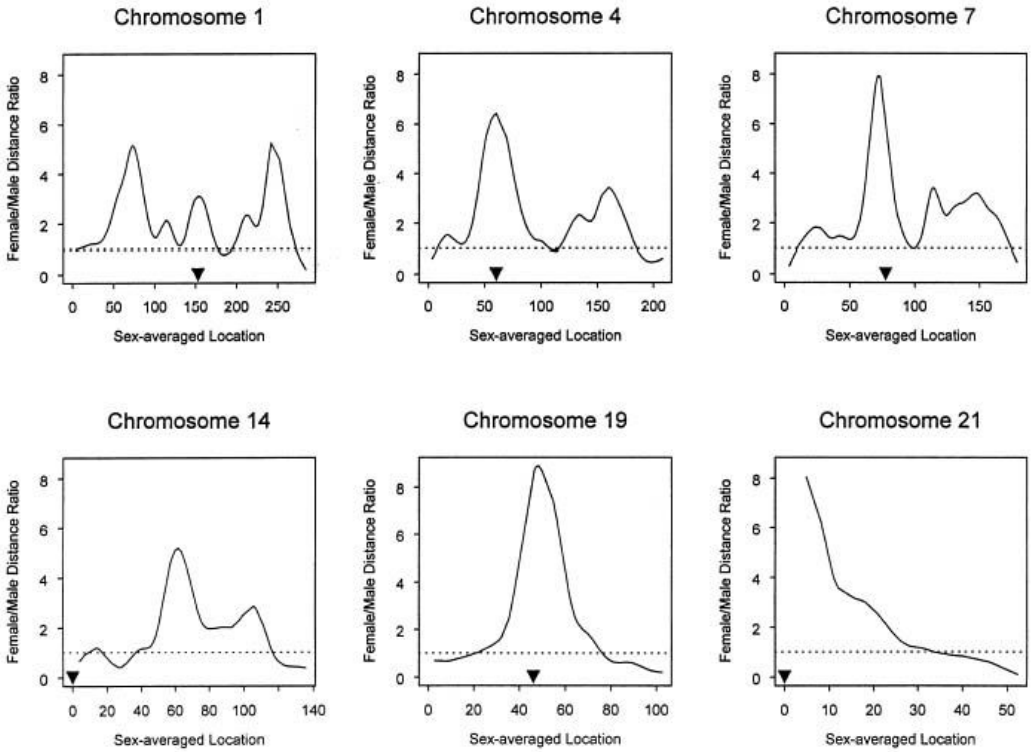
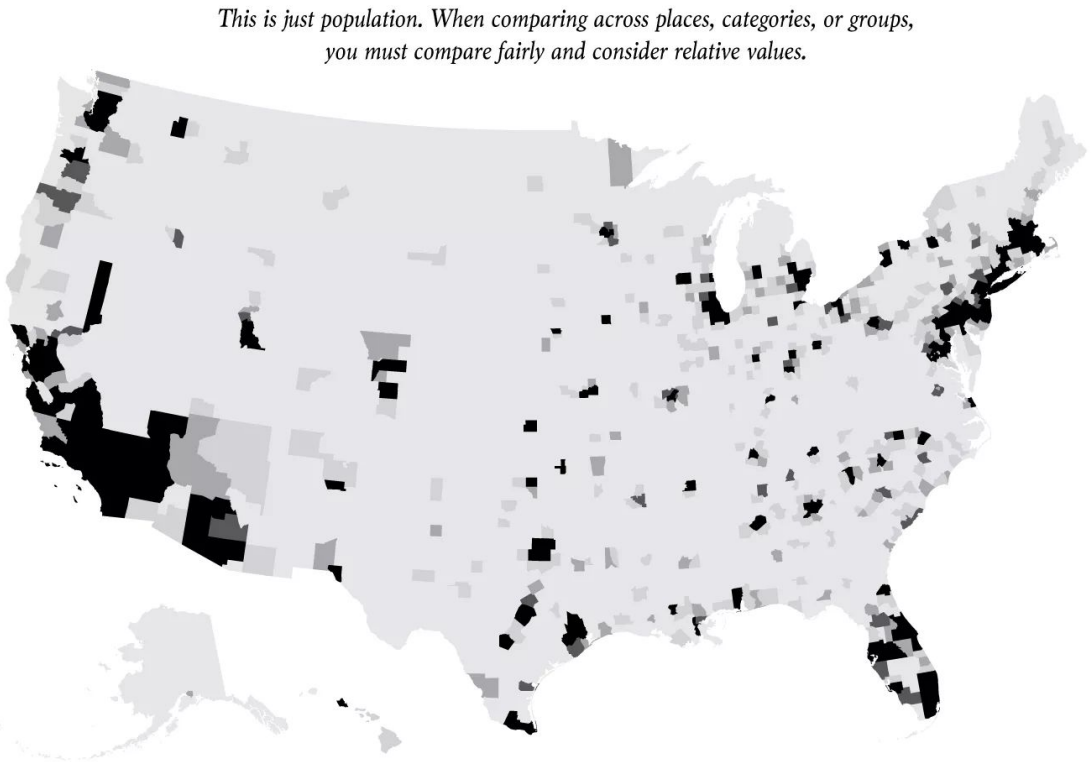
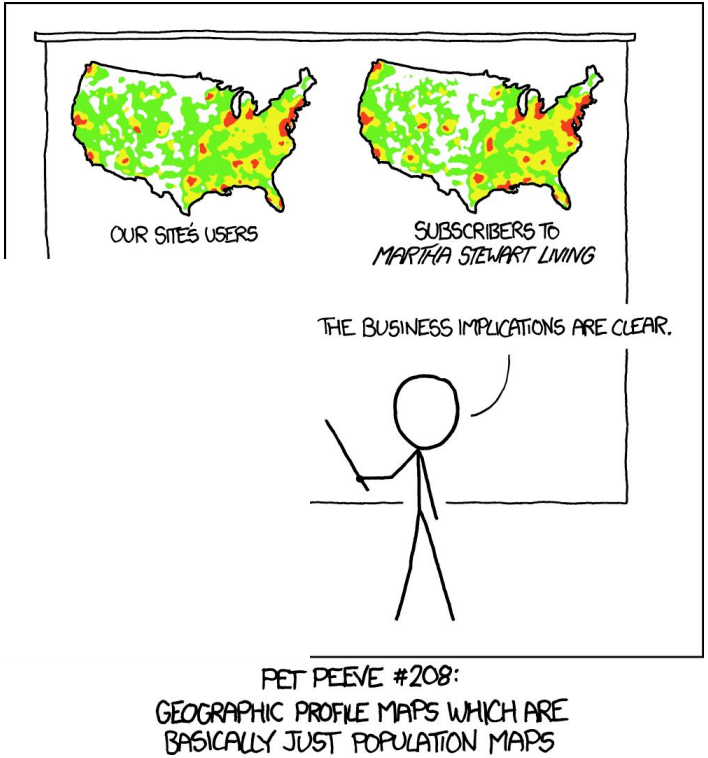


Figure 1 Plots of the female:male genetic-distance ratio against sex-averaged genetic location (in cM) along six selected chromosomes. Approximate locations of the centromeres are indicated by the triangles. The dashed lines correspond to equal female and male distances.

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

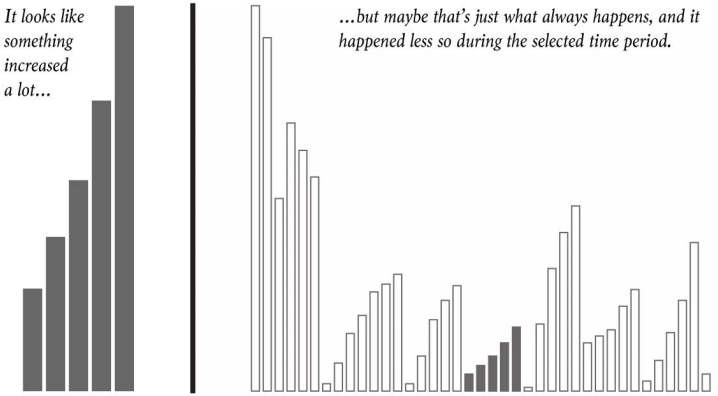
Beware of absolutes vs. relative values

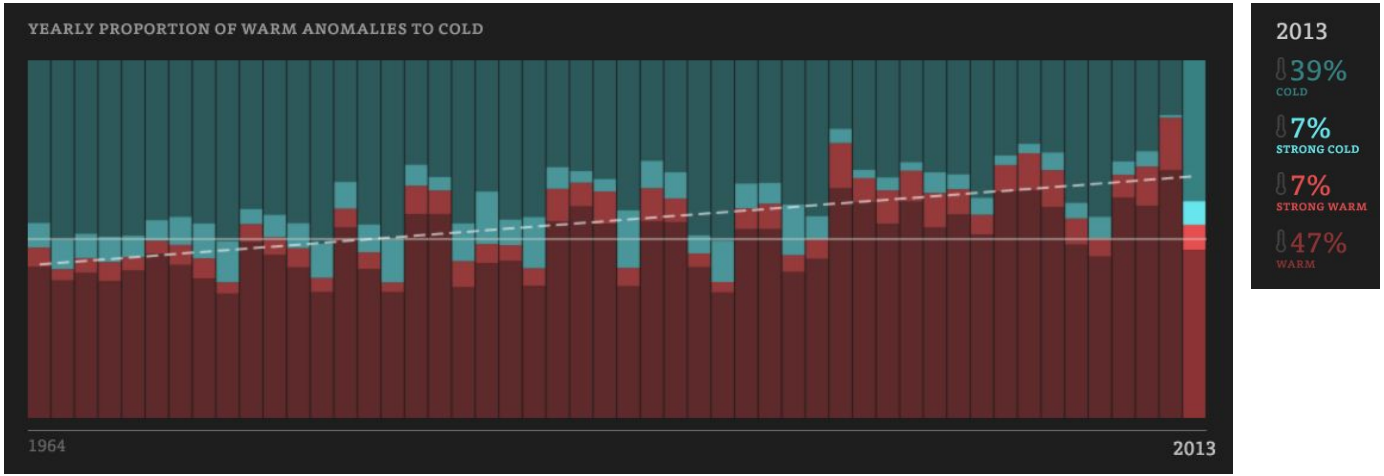




https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Showing full scope of the data is important



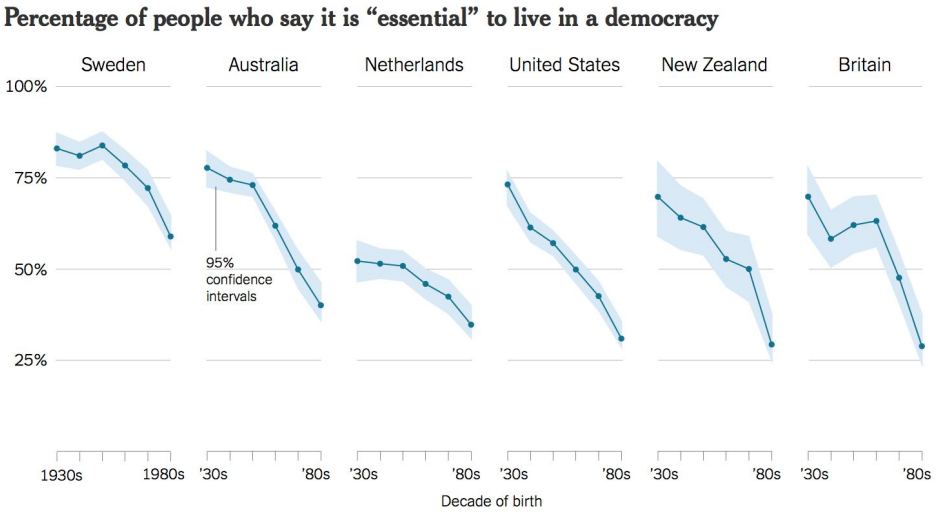


https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

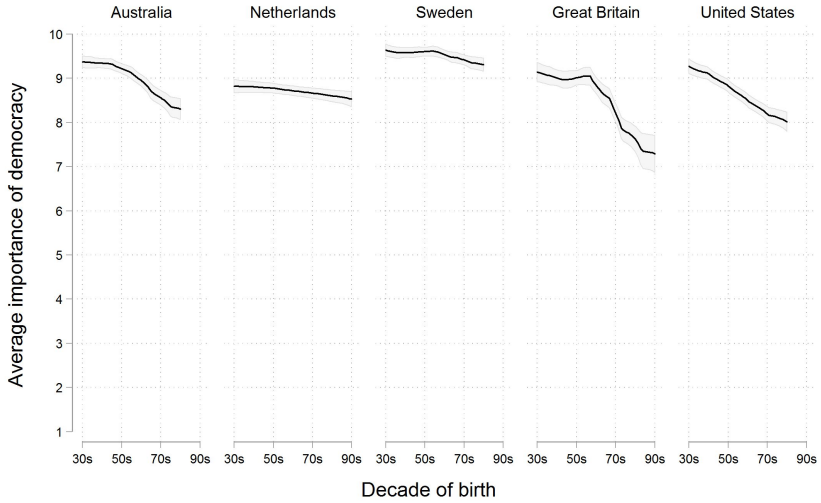
Data binning matters

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



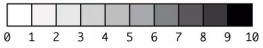
Graph by Erik Voeten, based on WVS 5

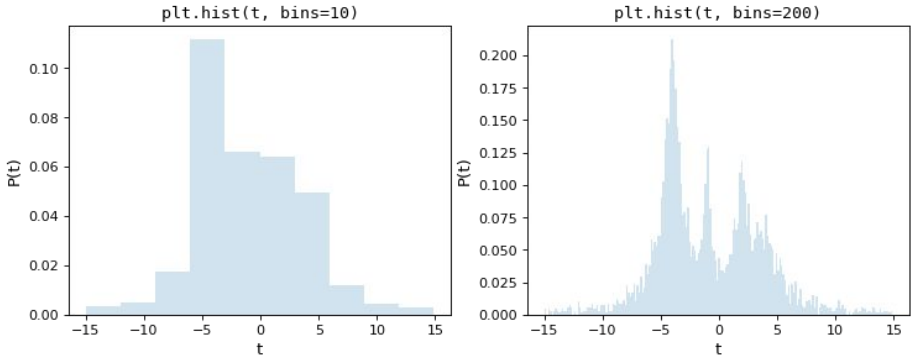
https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

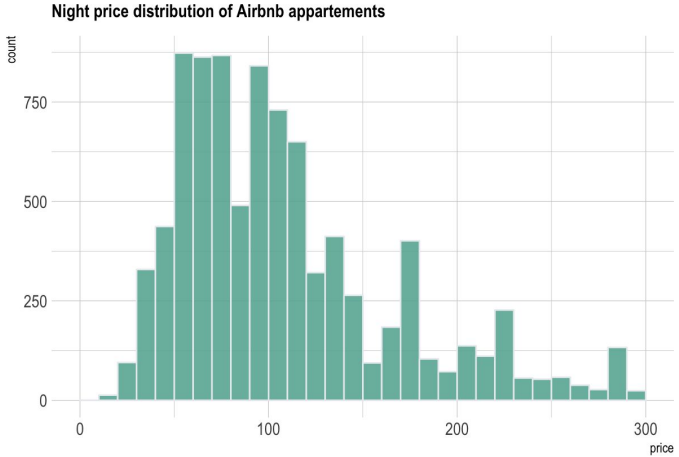
Data binning matters

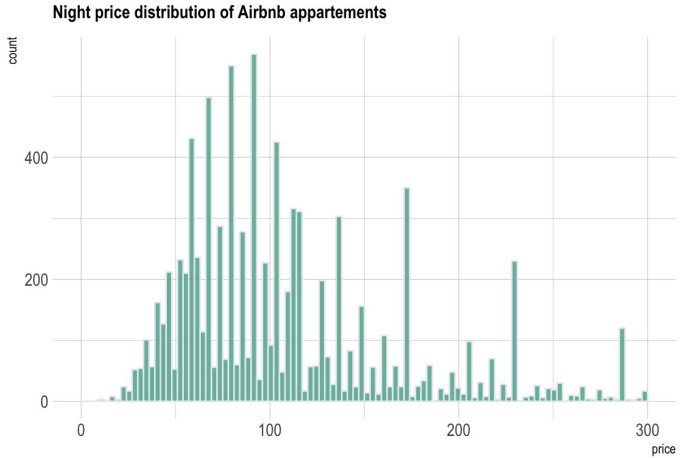
Two bins. What's really in the 1+ category? Might be hiding something.

That's better. It can show more variation.







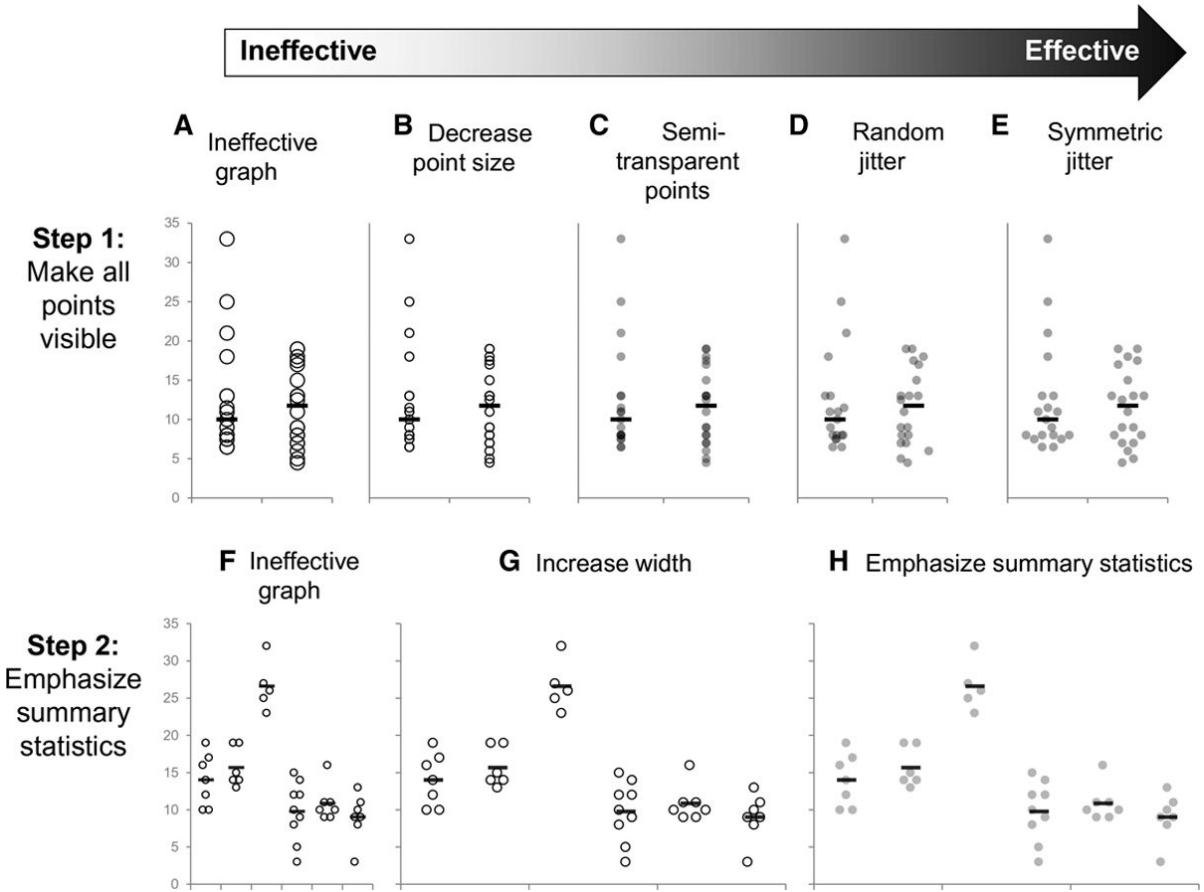


https://flowingdata.com/2017/02/09/how-to-spot-visualization-lies/

Improving visualization for clarity

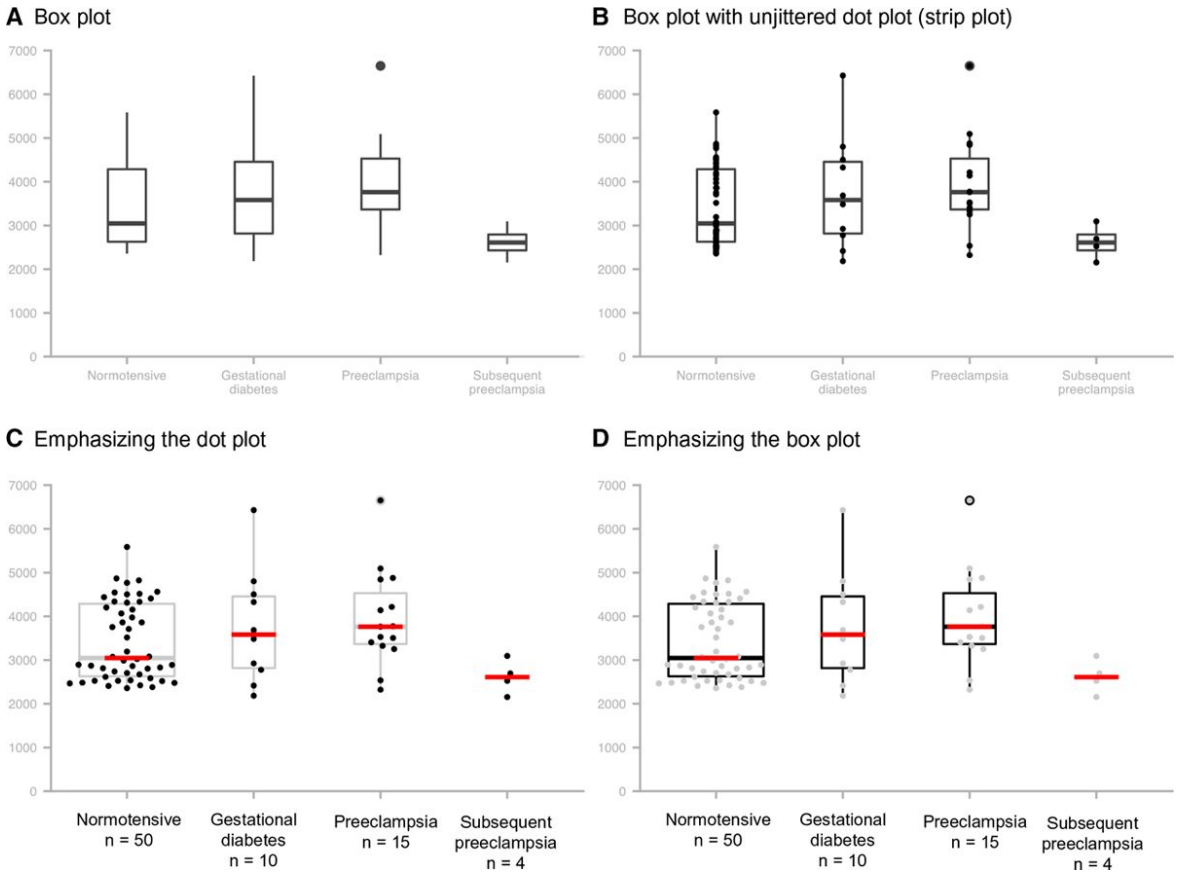
- Revealing underlying data as much as possible
- Changing emphasis based on data
- Reflecting study design / analysis
- Organizing & decluttering plots
- Choosing colors

Simple changes can make plots effective



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

Change emphasis based on data



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

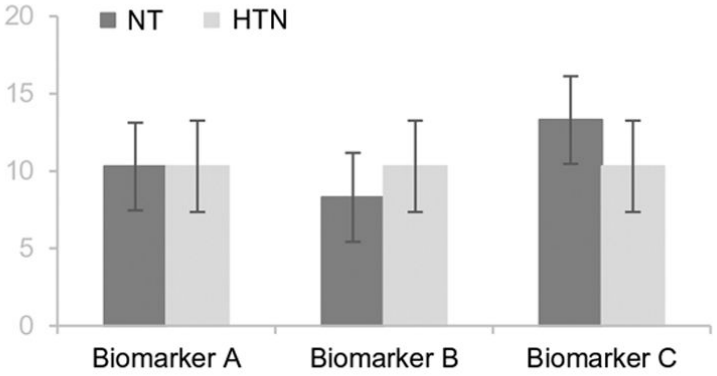
Make the plot reflect the study-design / analysis

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

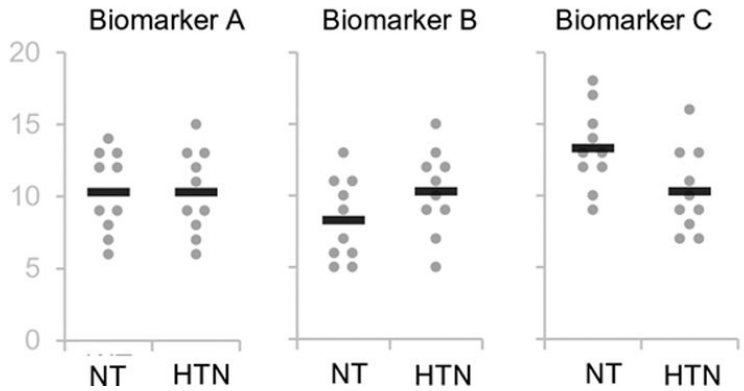
Figure structure erroneously suggests that authors also intended to compare biomarkers A, B and C



B

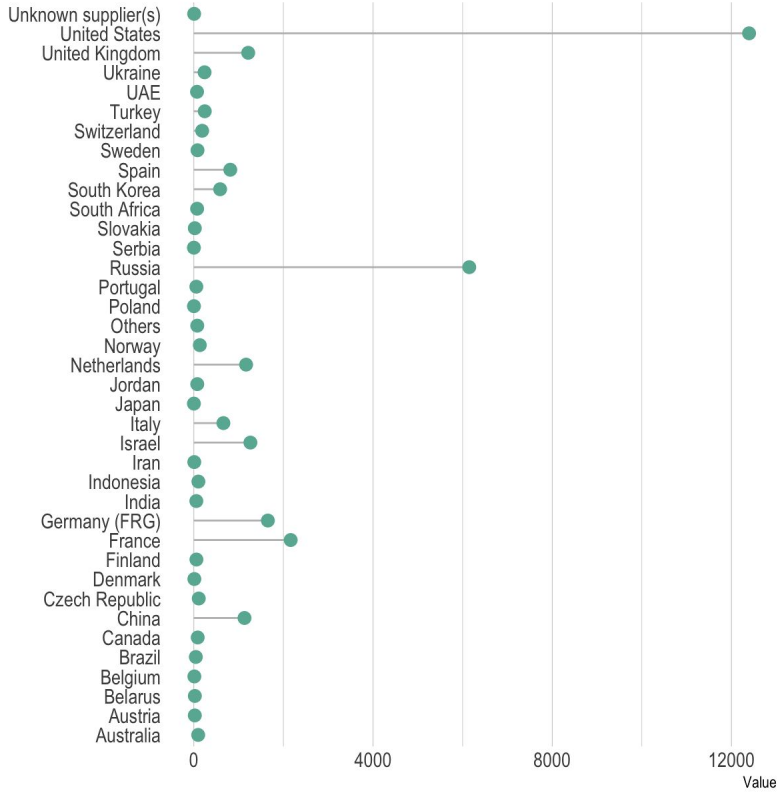
Clear communication

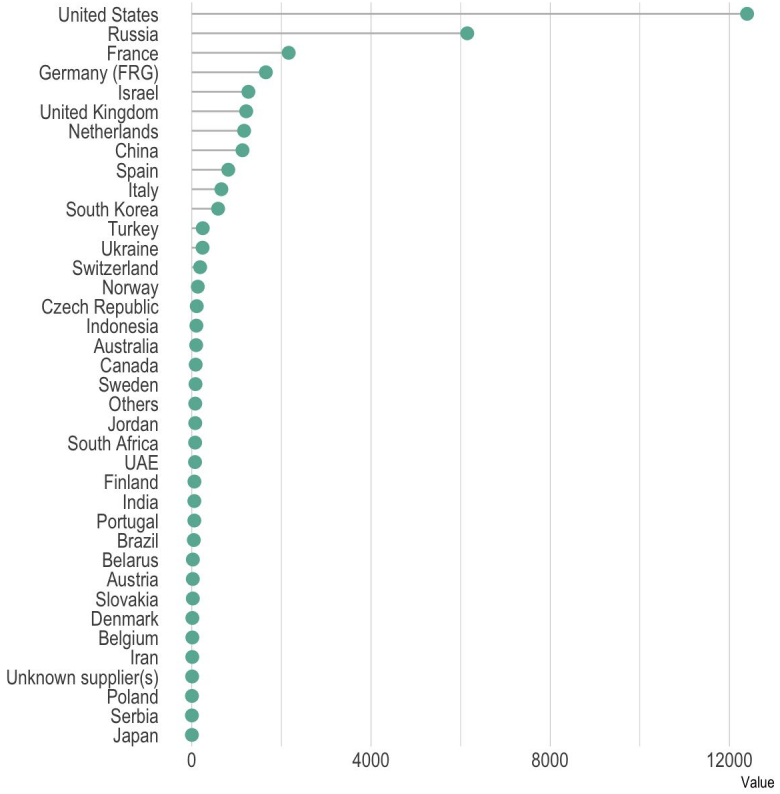
Figure structure matches study design & analysis, shows that the authors did not intend to compare biomarkers



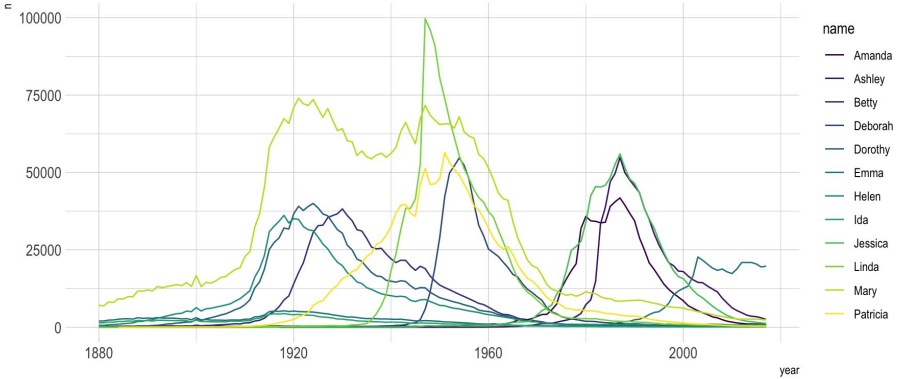
https://ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.118.037777

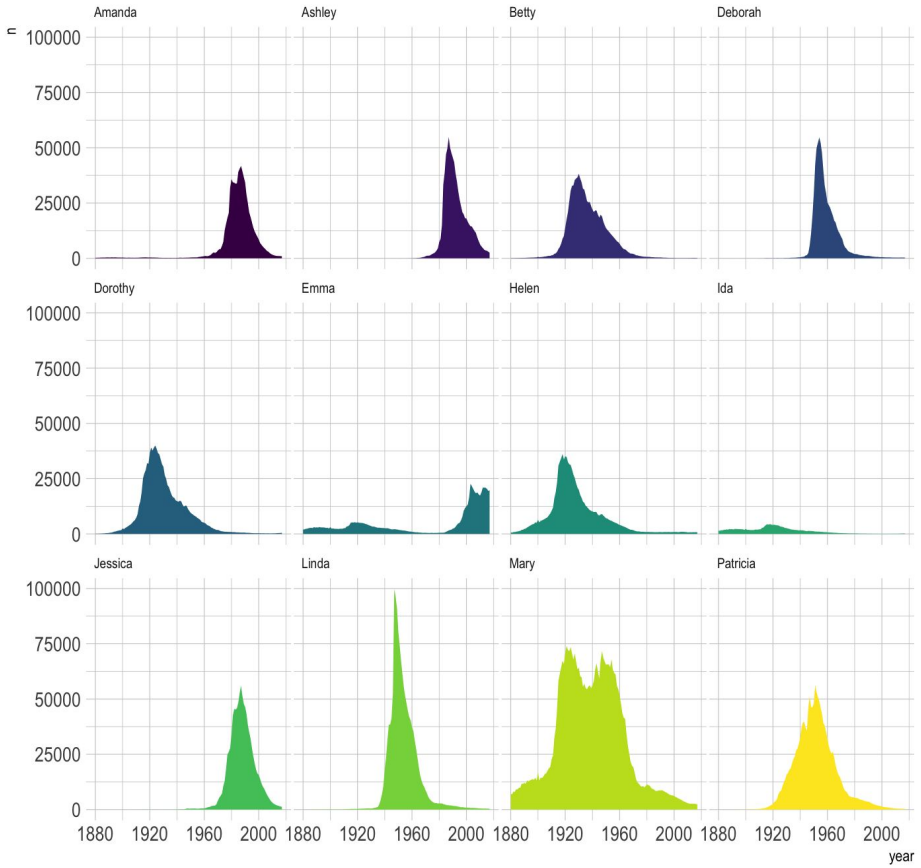
Reorder to make comparison clearer



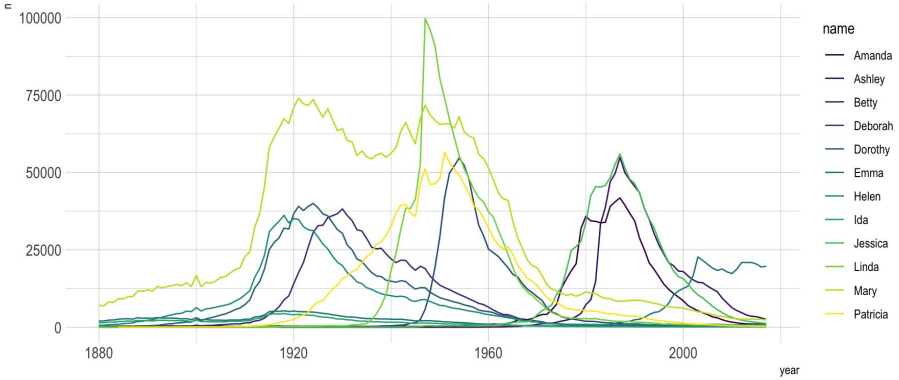


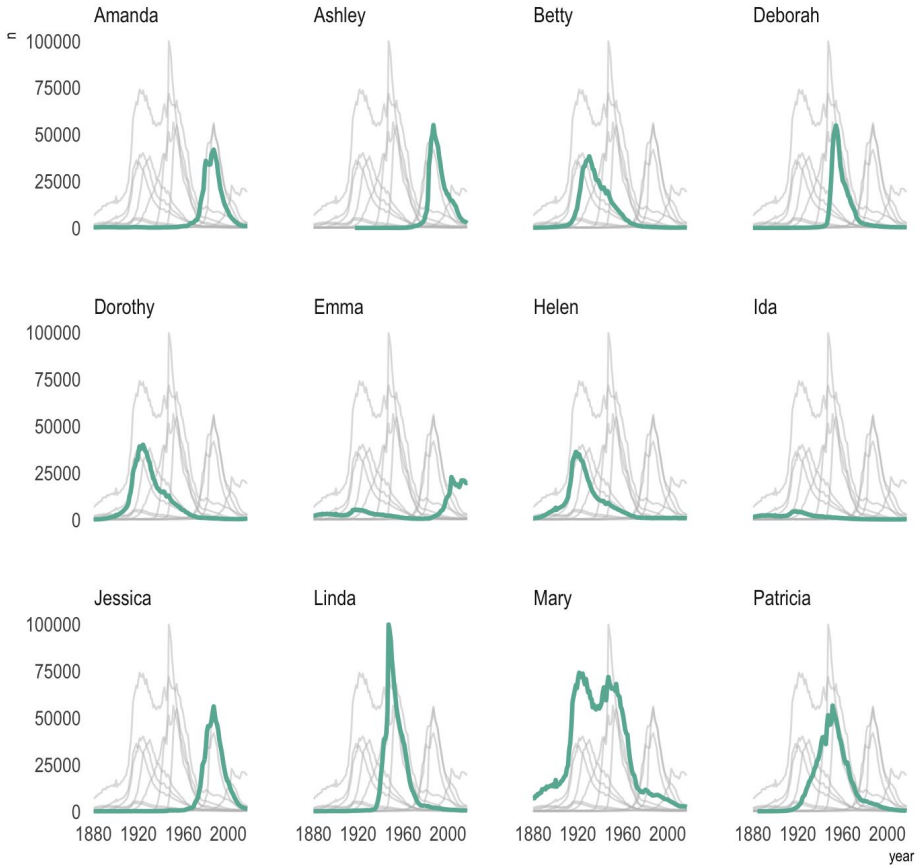
Split plots if complicated





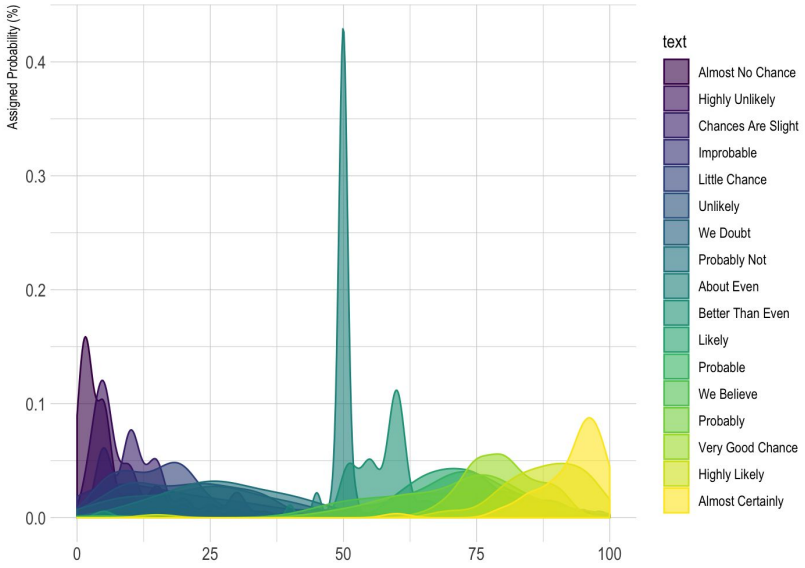
Split plots if complicated

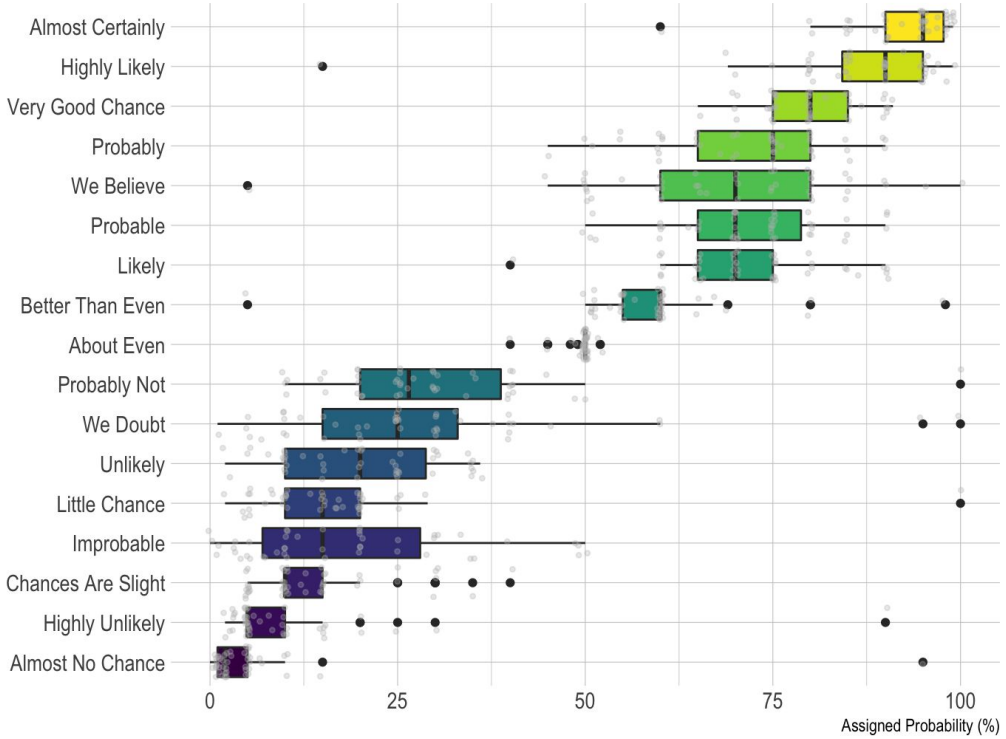




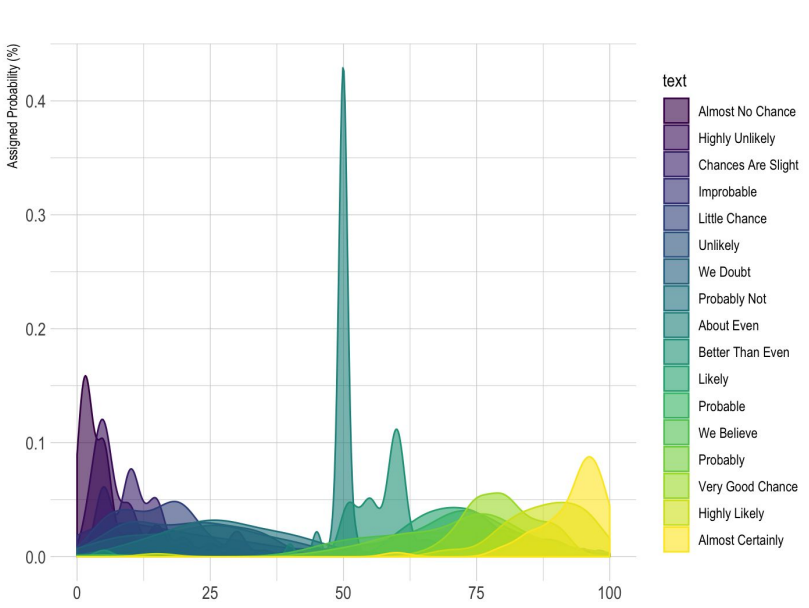
www.data-to-viz.com

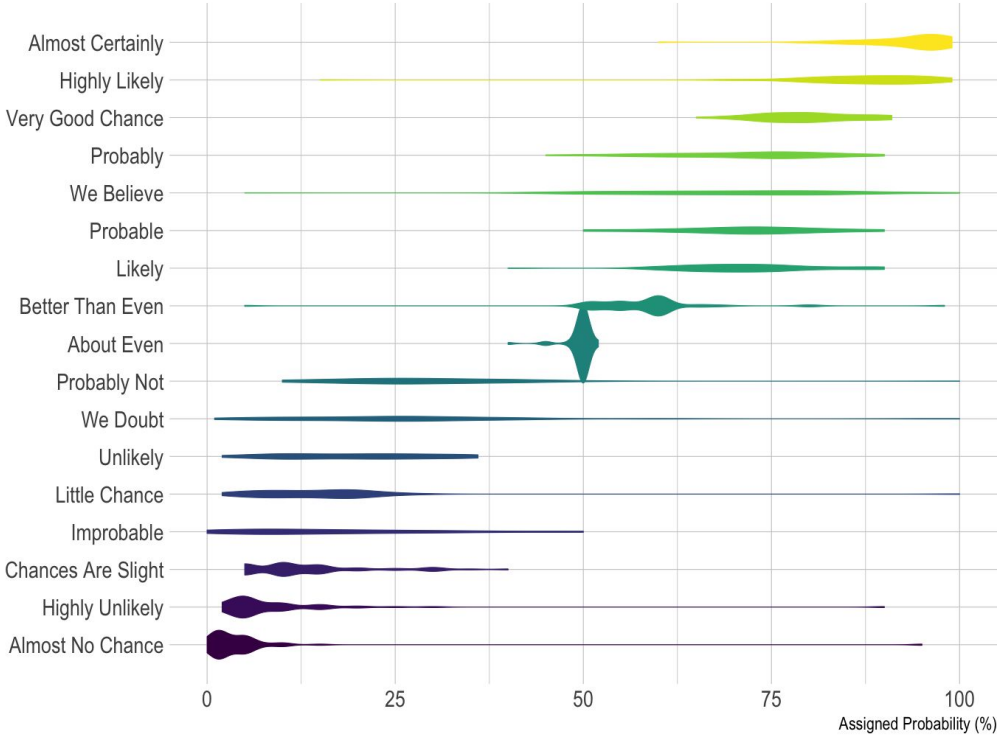
Split plots if complicated





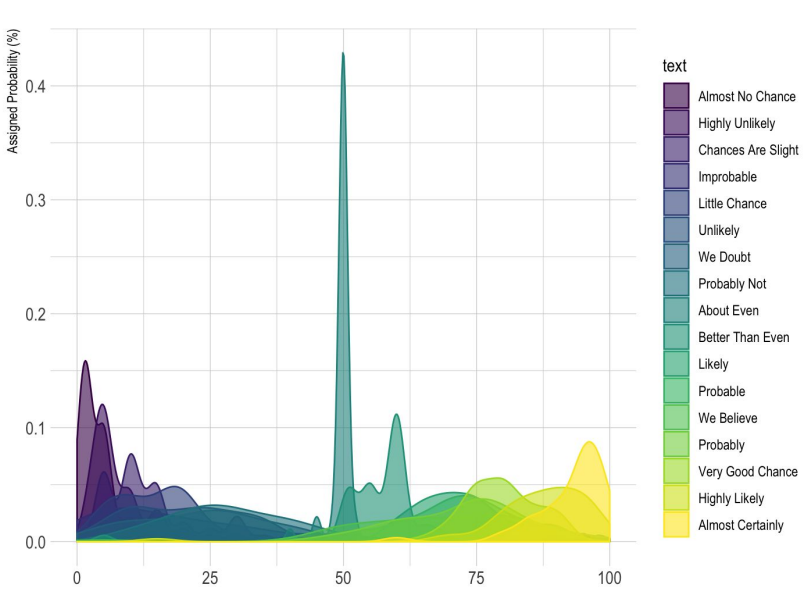
Split plots if complicated

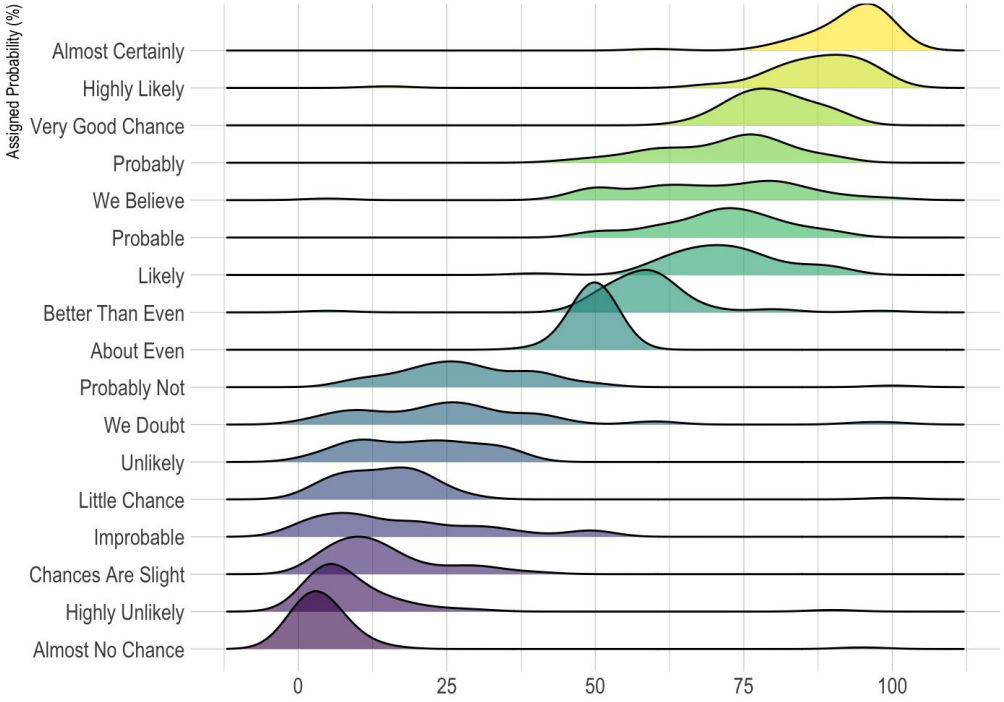




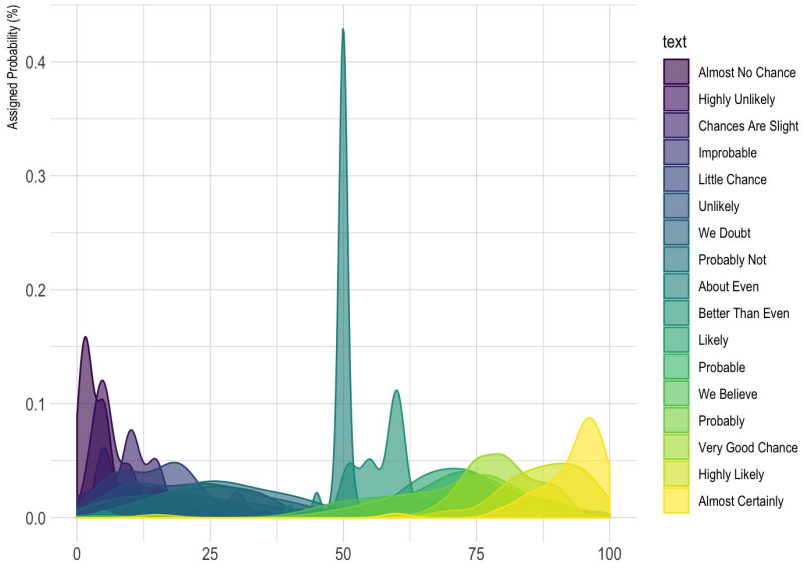
www.data-to-viz.com

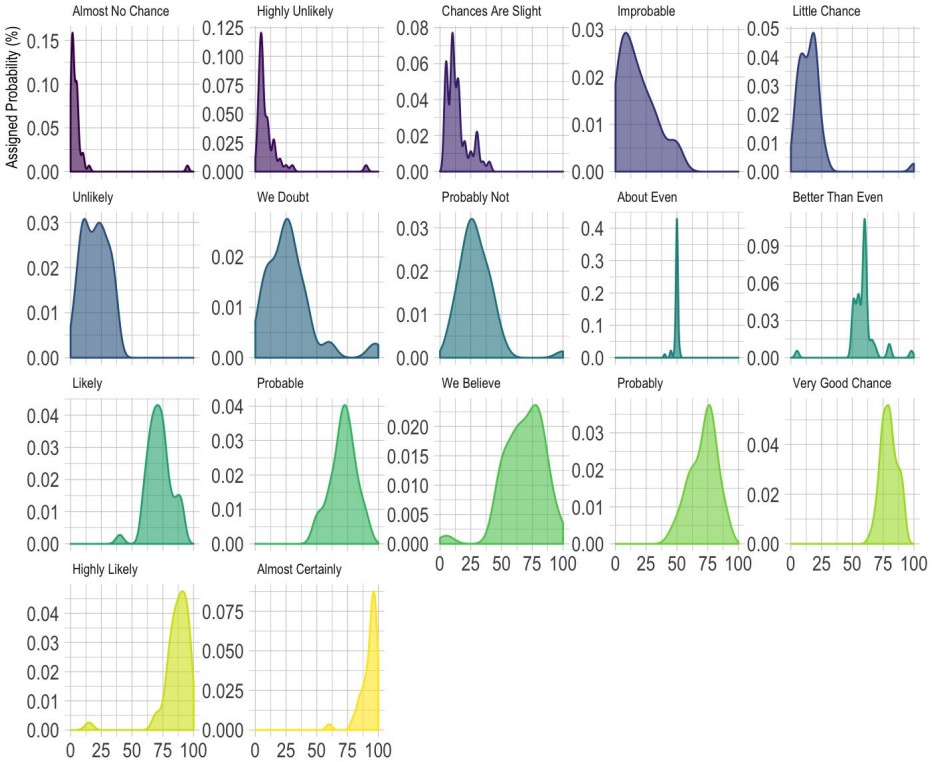
Split plots if complicated



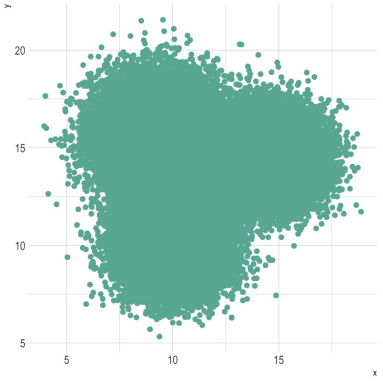


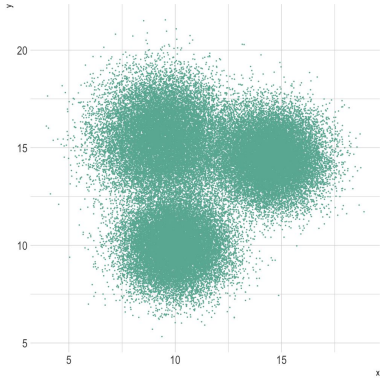
Split plots if complicated

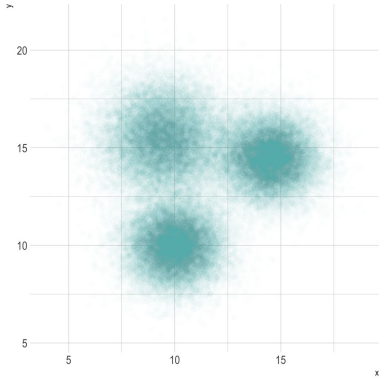


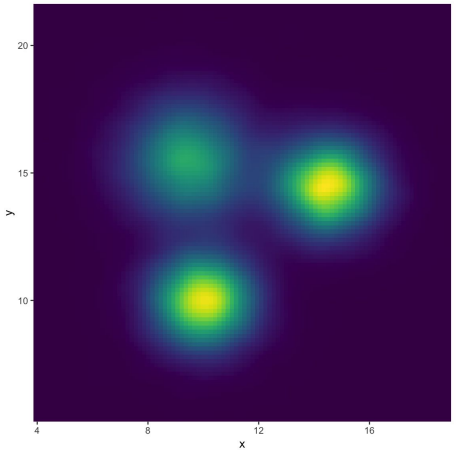


Avoid overplotting







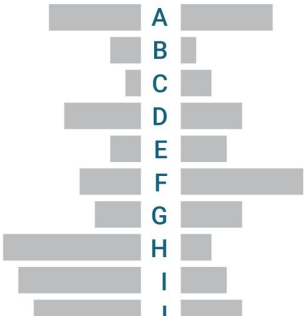


https://www.data-to-viz.com/caveat/

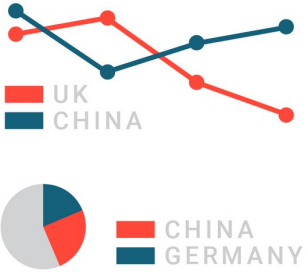
Choose colors carefully

NOT IDEAL

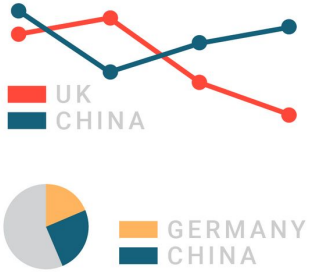
BETTER



NOT IDEAL



BETTER



COLOR KEY

CONTRAST

CONTRAST RATIOS

1.0				
1.1	Choose if you dislike readers.	That's bad.	That's bad.	Horrible.
1.5	Ok in 1% of the cases.	Not ideal.	That's bad.	My eyes!
2.5	Can be a good choice.	Ok.	Not ideal.	That's bad.
4.5	Safe choice.	Great.	Ok.	Not ideal.

NOT IDEAL

BETTER

https://blog.datawrapper.de/colors/

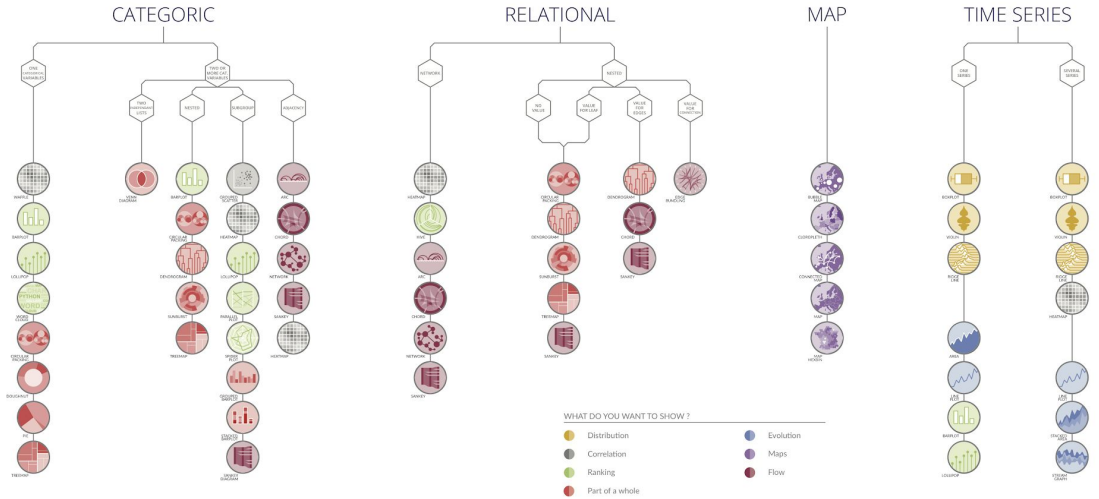
from Data to Viz

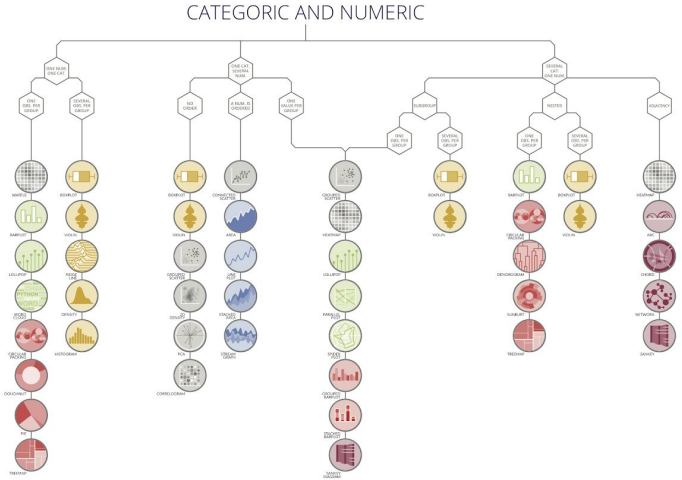
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

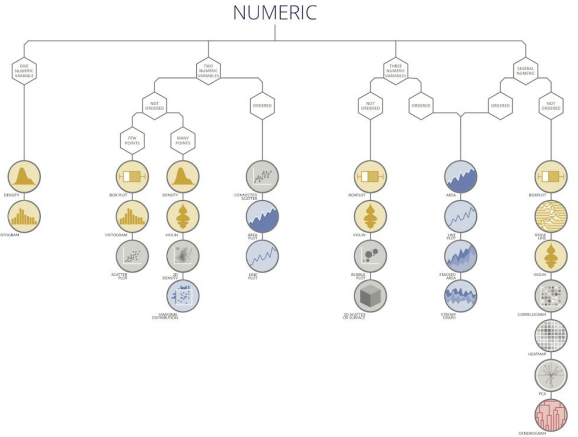
1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com







How to write a paper

1. Create a draft of your ideal set of Figures & Tables (in addition to supplements) that tells a coherent story.

- Write detailed figure/table legends to help understand what piece of the story each figure/table will convey.

2. Read widely and deeply. This exercise is going to entirely feed your Introduction and Discussion sections.

- Record all the papers in a Zotero library. Create a single Google Doc and make notes about each paper along with its title & link. The purpose of the notes is to make the points you will write in your manuscript to cite each paper.

3. Make all main and supplemental Figures and Tables along with declarative titles & detailed legends.

How to write a paper

4. Write a very detailed Methods section and prepare Code & Data to be released. As you write each subsection of Methods, organize and document the pertinent code, data, and results.

5. Write the Results section.

6. Then, write the Introduction and Discussion sections, and add References.

- The Introduction section should lead up to the main questions and results of the manuscript.
- The Discussion should put the new results in the context of existing work, describe novelty & potential impact, and conclude with opportunities for future work.

7. Write Title and Abstract