CSE 591: Data Science

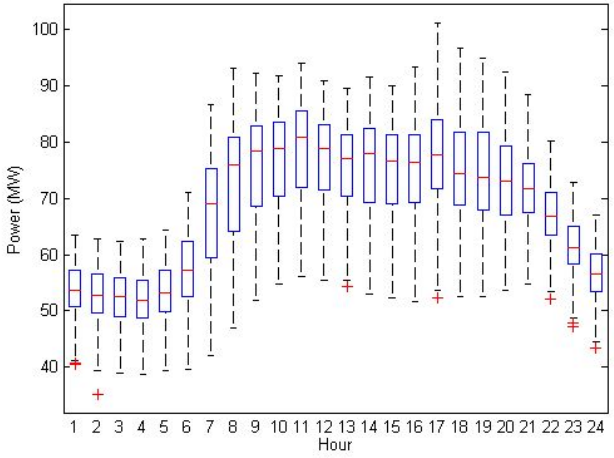
Steven Skiena
Stony Brook University

Lecture 12: Practice of Data Visualization

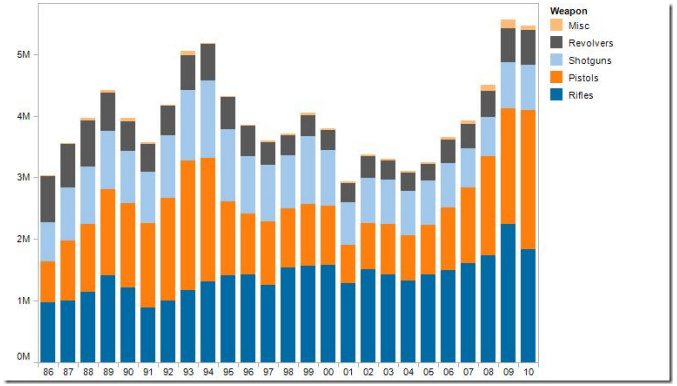
Box and Whisker Plots

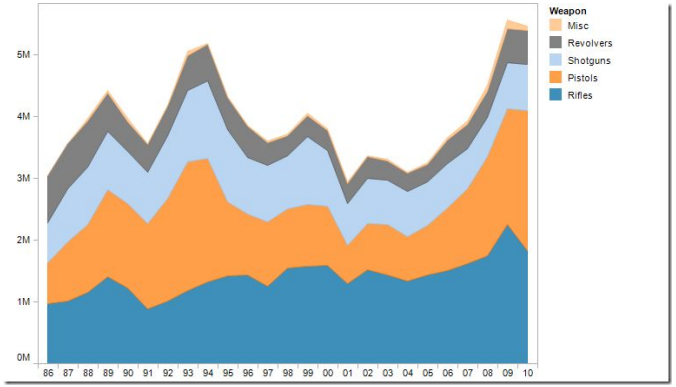
Box plots concisely show the range / quartiles (i.e. median and variance) of a distribution.

I personally prefer contour lines without the boxes.

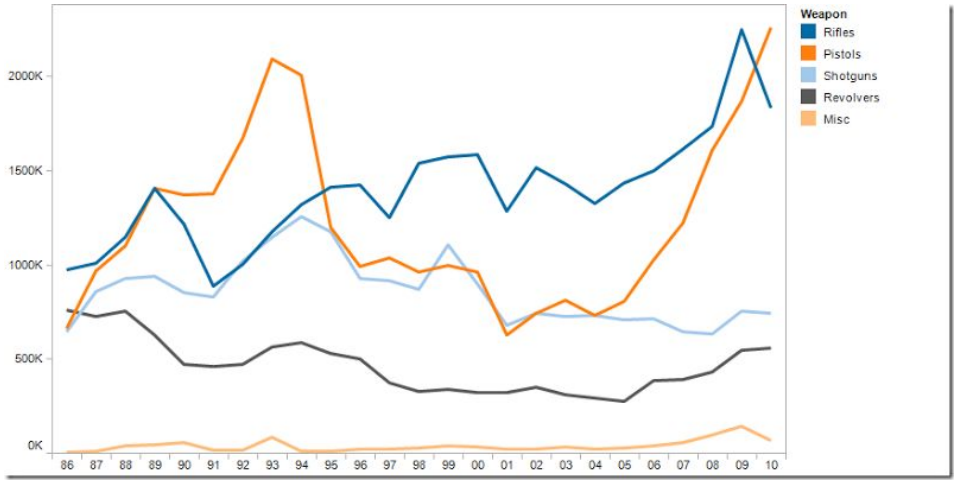


Stacked Area vs. Line Plots

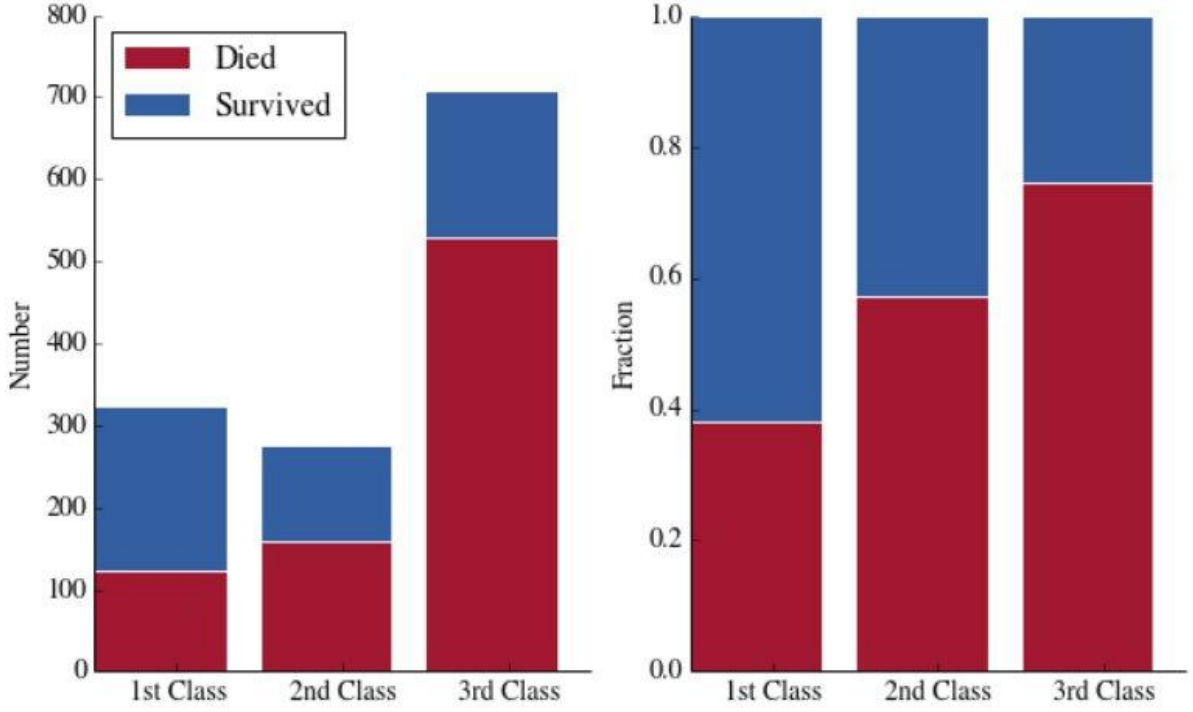




Hard to see trends in middle areas of stack:

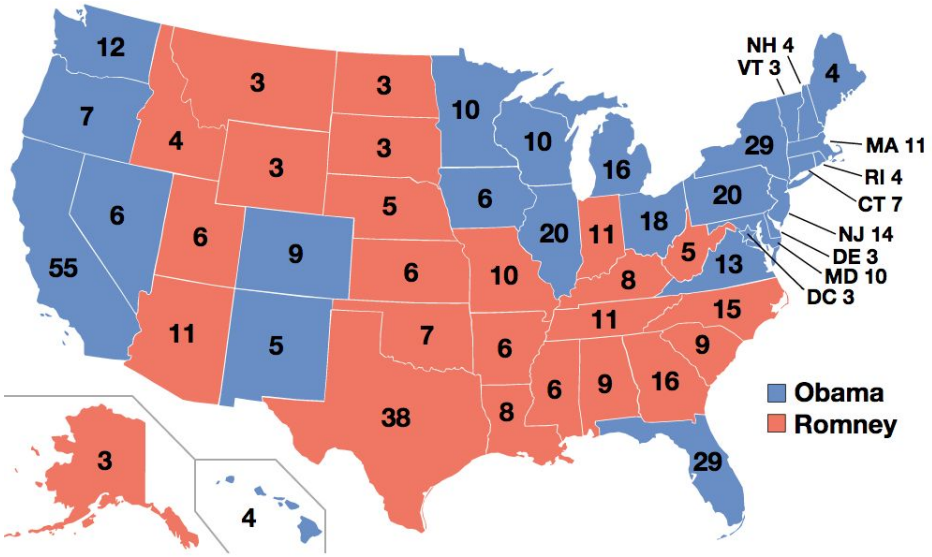


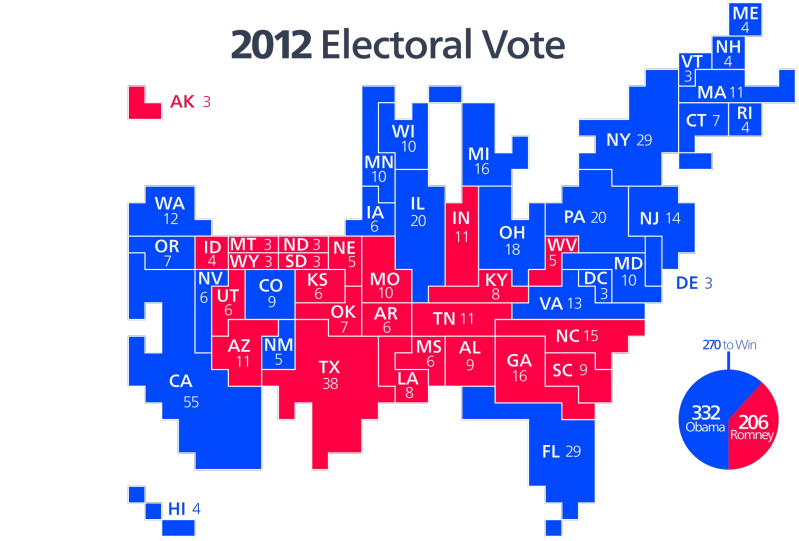
Stacked Bar Charts: Titanic



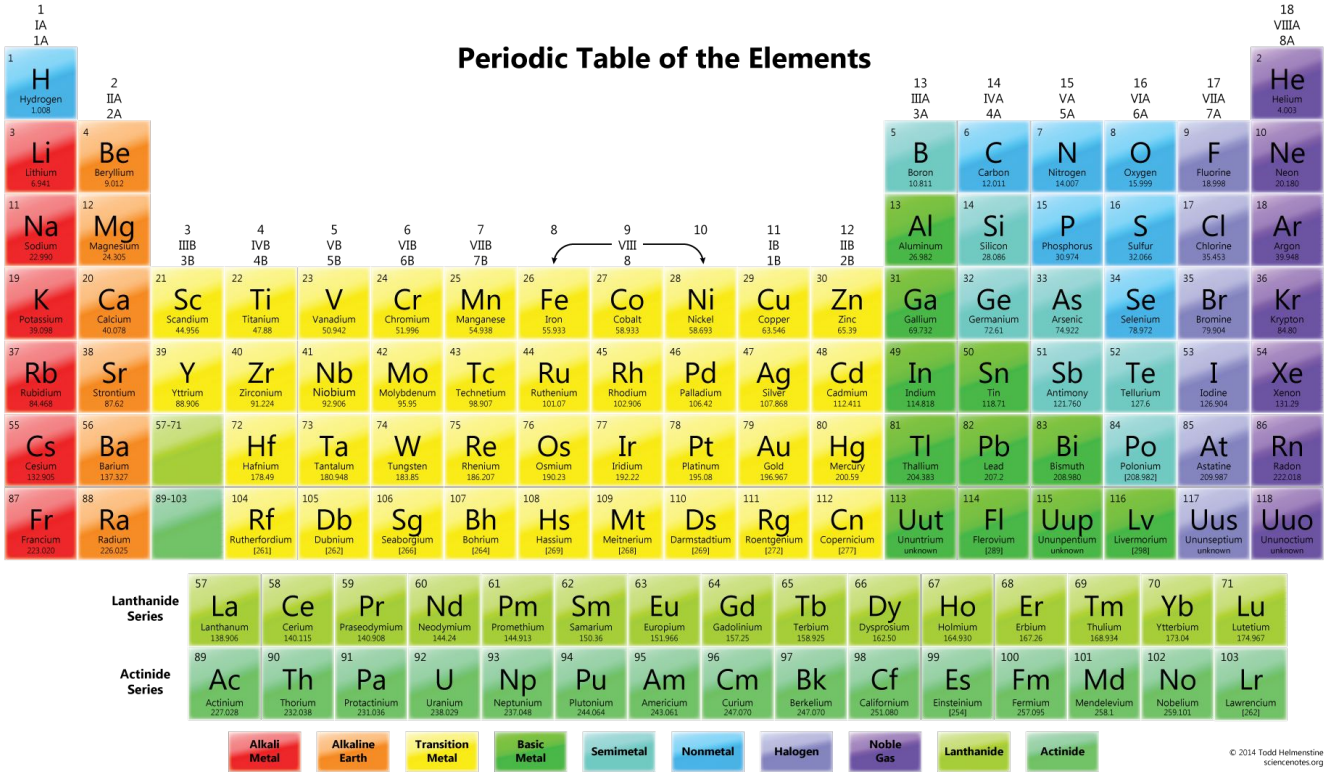
Data Maps and Cartograms

Cartograms distort regions to reflect an underlying variable.





Non-Geographic Data Maps



Tools for Data Visualization

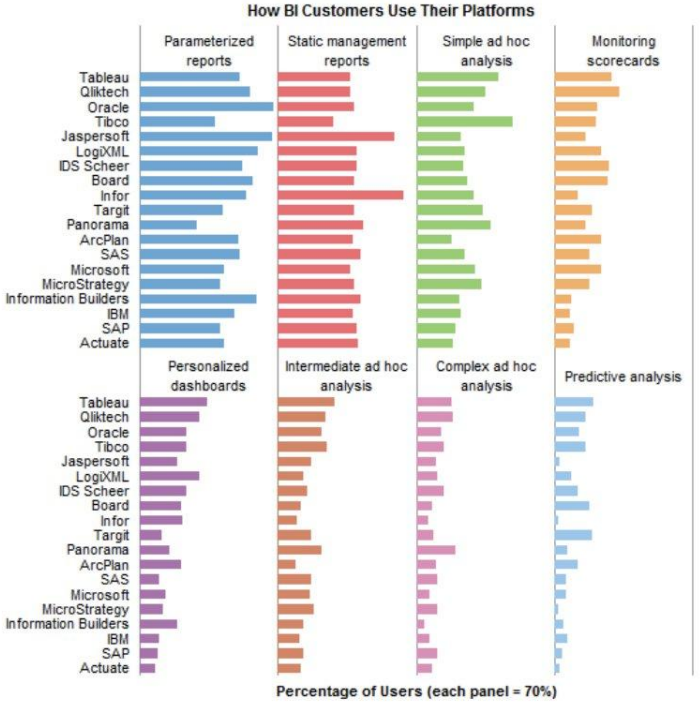
Just because Excel is very popular does not mean it produces good graphs/plots.

The statistical language R has a very extensive library of data visualizations.

MatPlotLib is your key to producing good graphs/plots in Python.

Repetitions for Multivariate Data

Small multiple plots / tables are good ways to represent multivariate data.

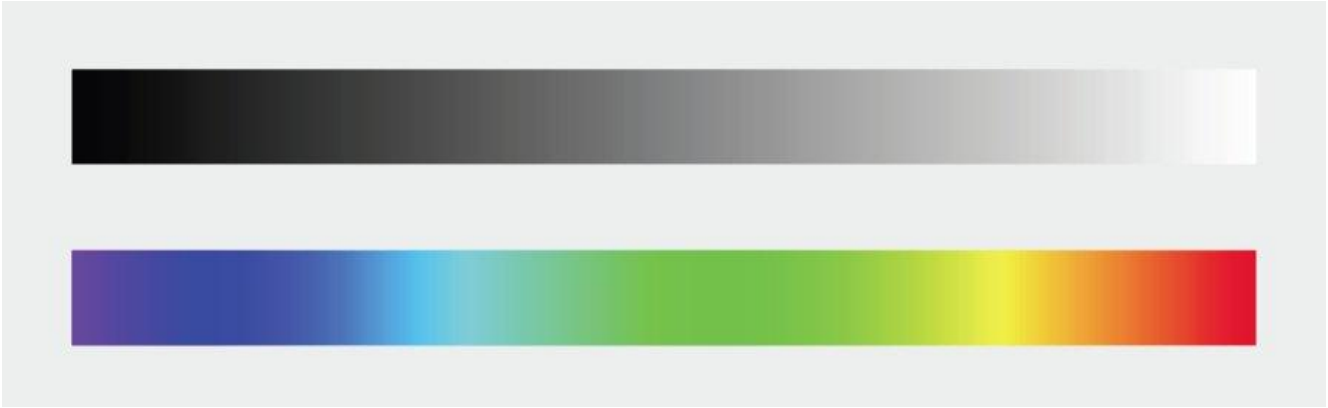


Understanding Color Scales

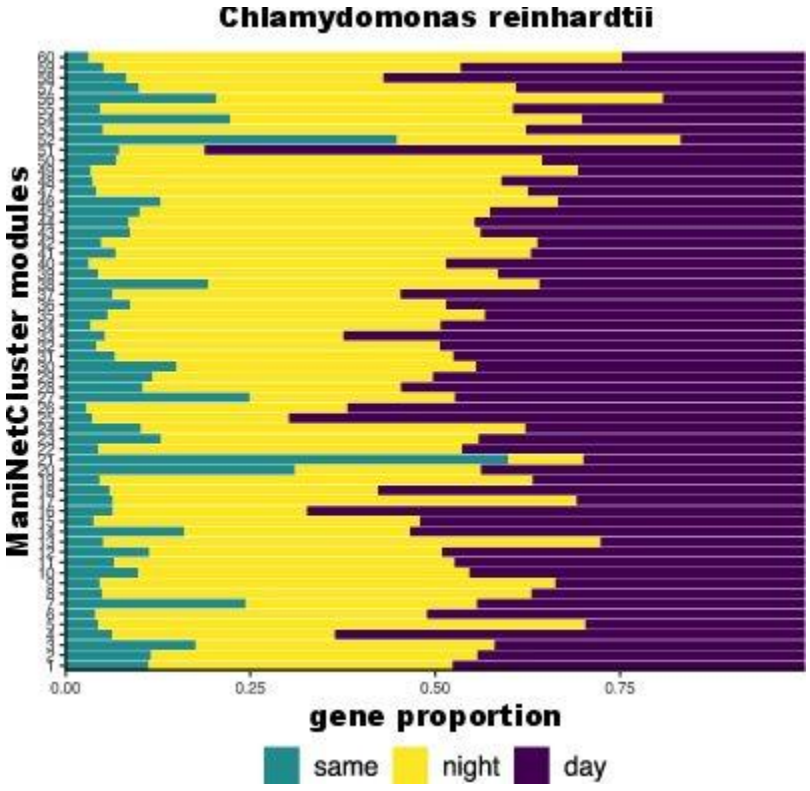
Rainbow Color Maps

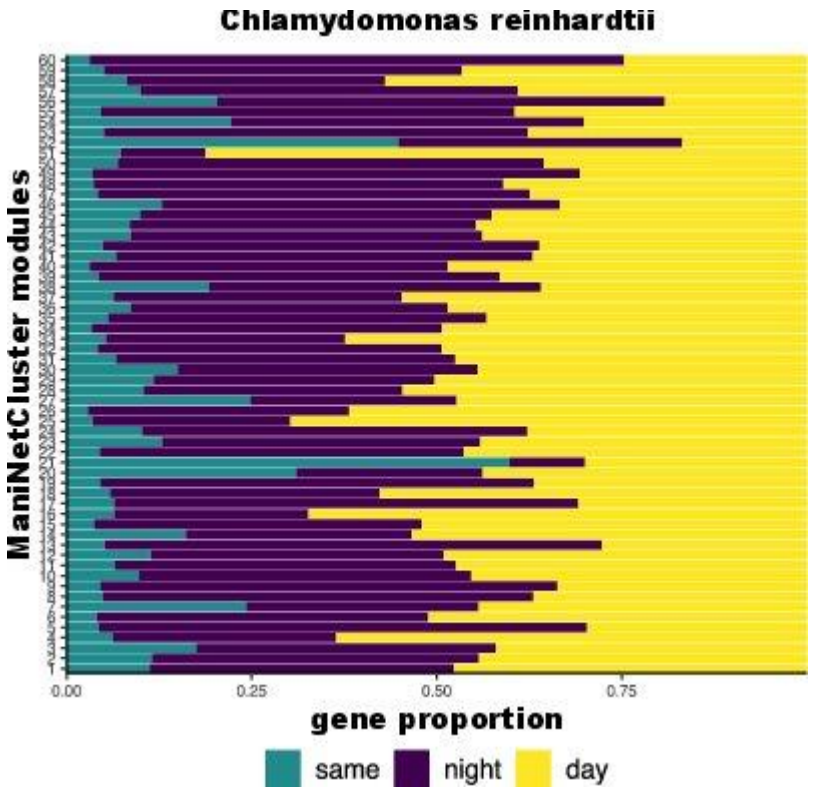
Rainbows are perceptually non-linear.

Distinct positive/negative colors reflected about a center make good scales.



Which Coloring is Better?





Appreciating Art: Which is Better?

Sensible appreciation of art requires developing a particular visual aesthetic.

Tufte's Visualization Aesthetic

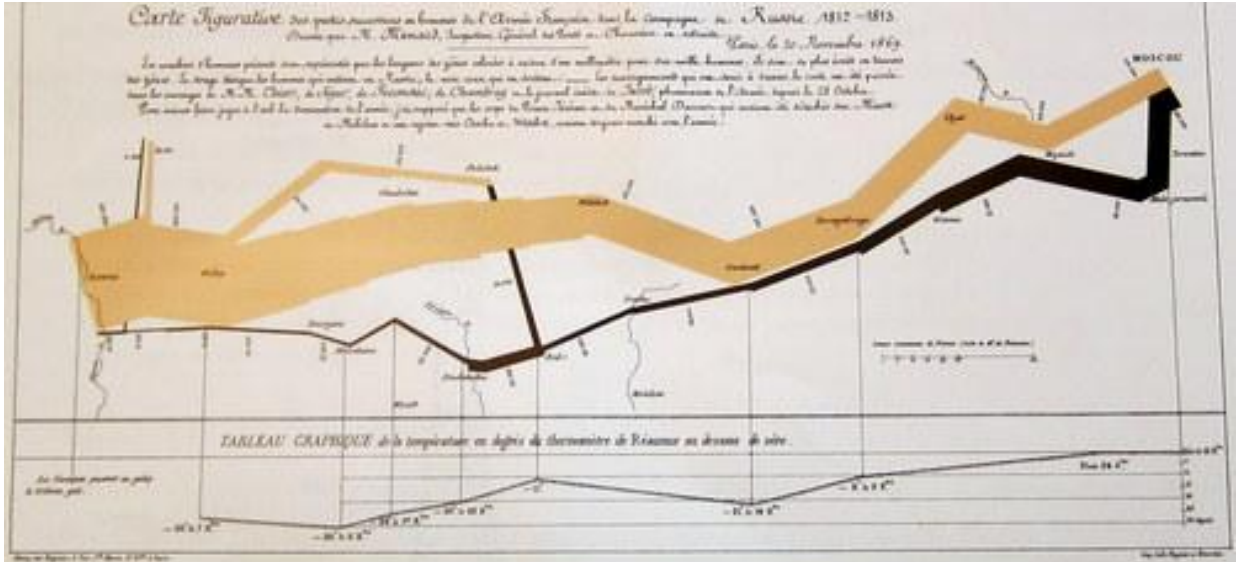
Distinguishing good/bad visualizations requires a design aesthetic, and a vocabulary to talk about data representations:

- Maximize data ink-ratio
- Minimize lie factor
- Minimize chartjunk
- Use proper scales and clear labeling

Great Data Visualizations

- Display data accurately and clearly.
- Tell a story that the data reveals.
- Are rich enough to make you want to look carefully and study the data.

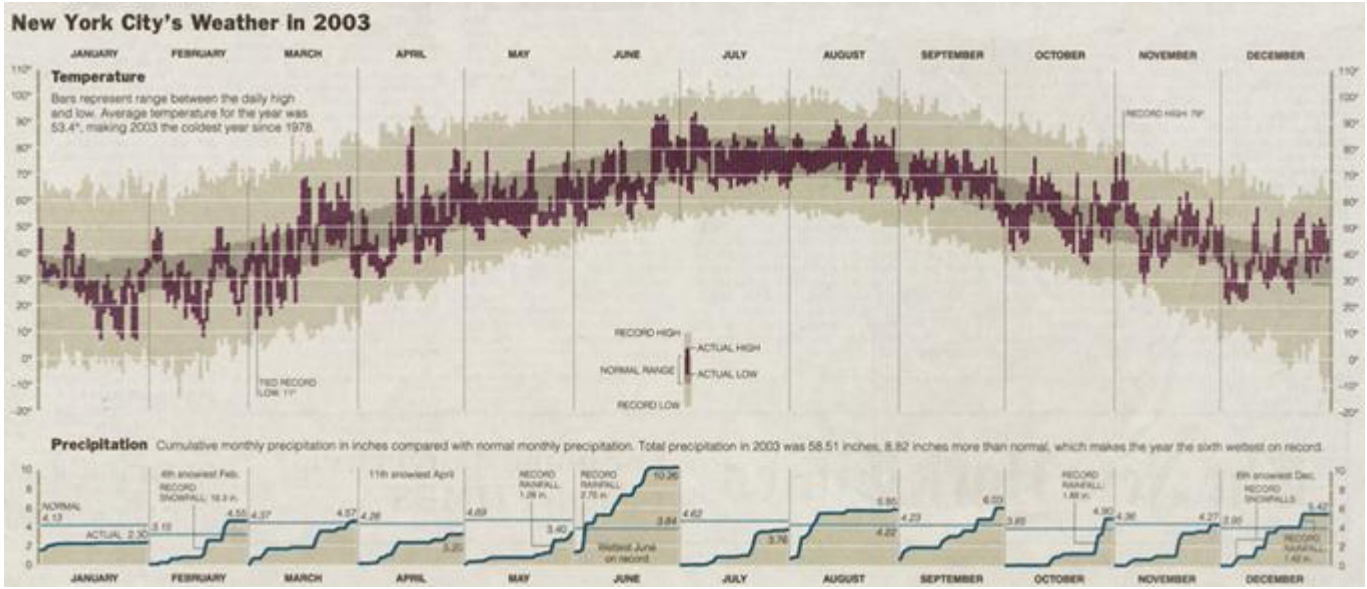
Napoleon's Advance and Retreat



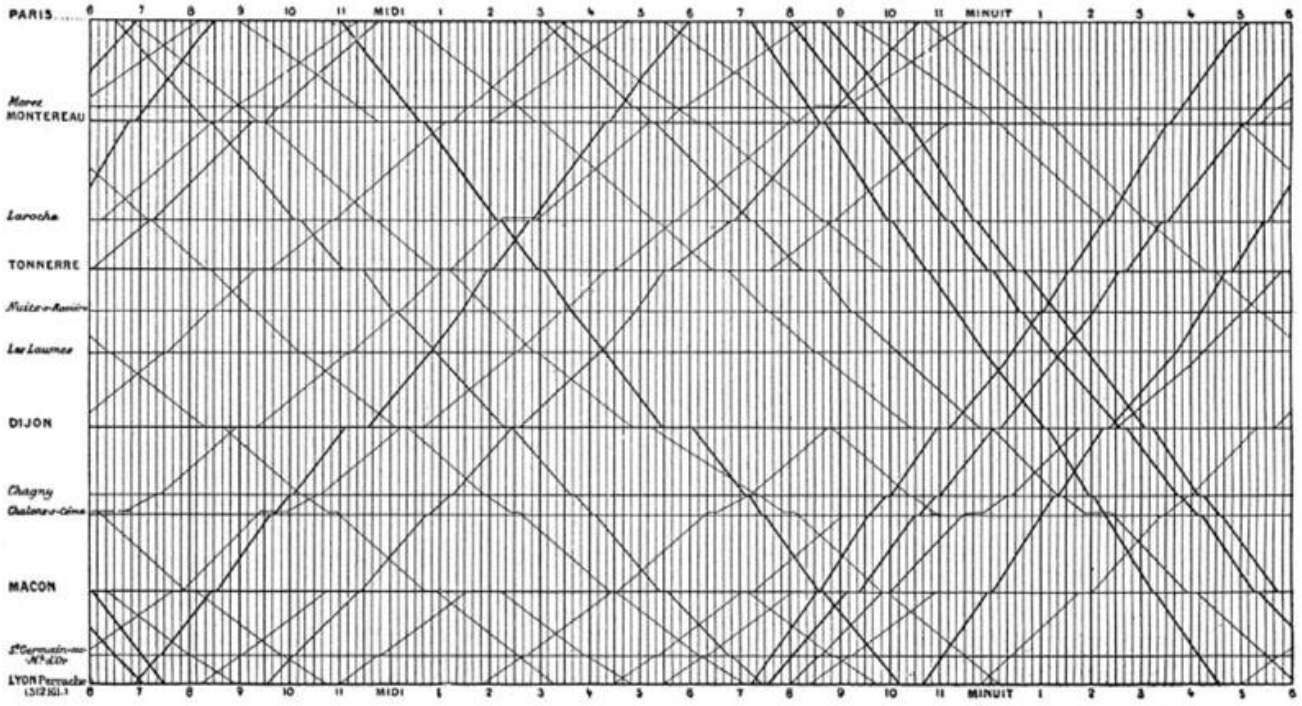
This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (422,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn. Napoleon's March poster $14 postpaid; English/French version $18 postpaid.

New York's Weather Year in Review

A clear story displaying over 3,000 numbers.



Marey's Train Schedule



What can you see here you cannot with normal train schedules?

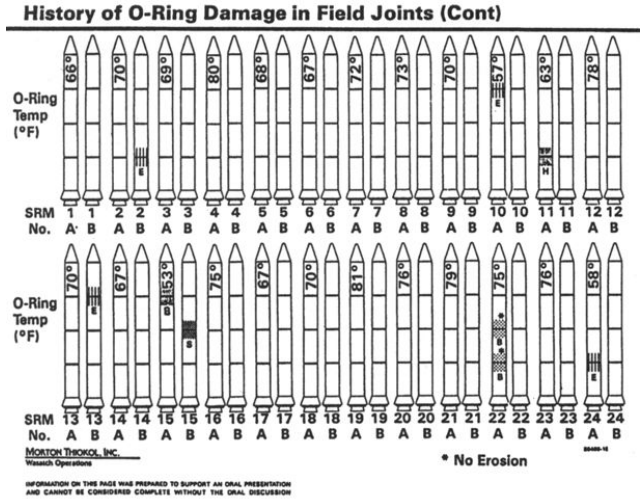
It would be even better with a lighter datagrid. Never imprison your data!

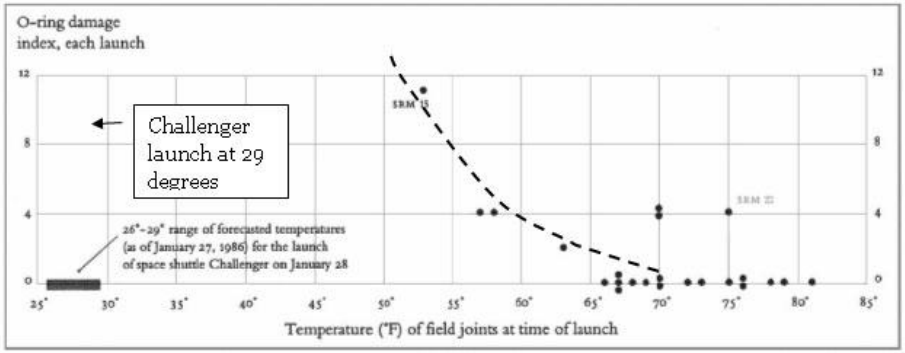
Discovering the Source of Cholera

John Snow used this data map to identify the source of an 1854 Cholera epidemic as a single contaminated water pump.

Trying to Stop the Challenge Launch

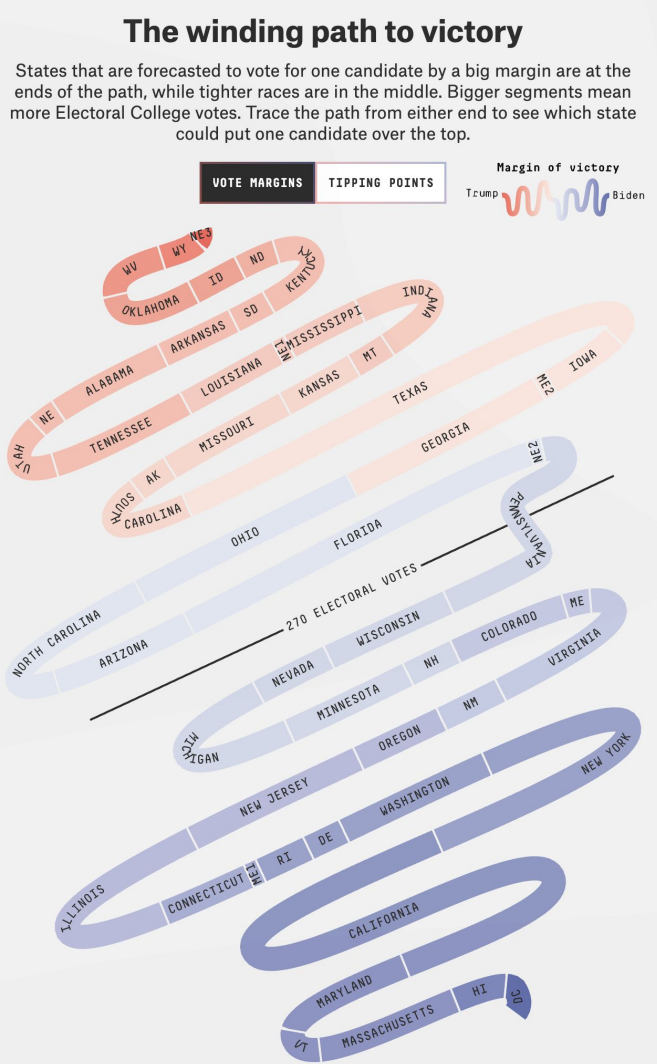
Engineers failed to convince management to call off the launch using a poor data visualization.





State of the Race

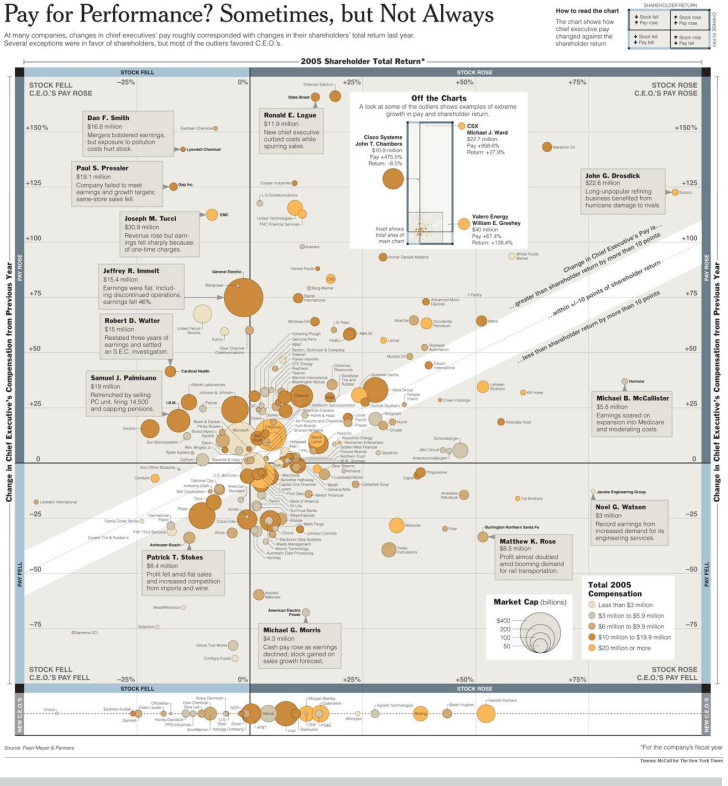
Effective use of what could have been chartjunk and color to capture the state of the 2020 election in one view



Which Executives Earn their Pay?

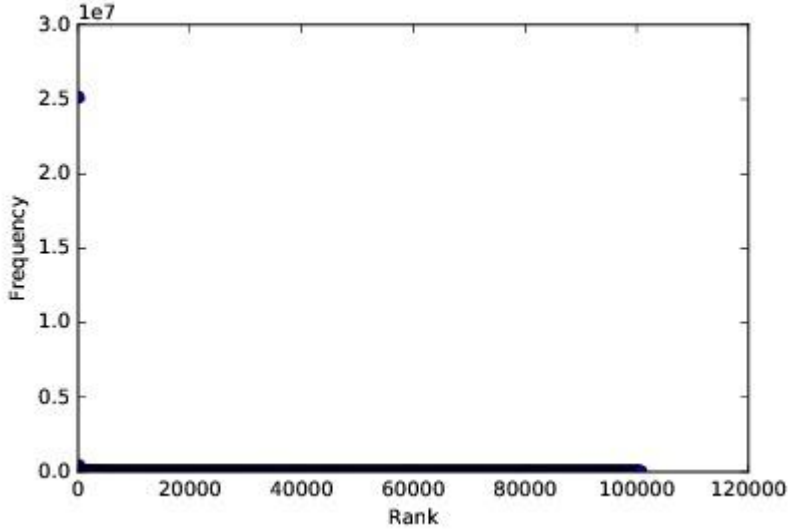
Chart plots points in 4D, shown by x, y, color, size.

Using attributes like point size/color is better than plotting points in 3D.



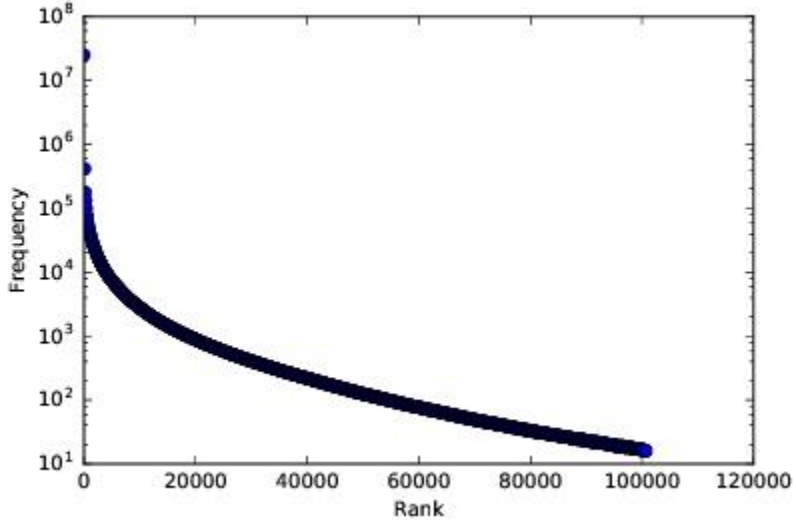
Terrible Student Visualizations...

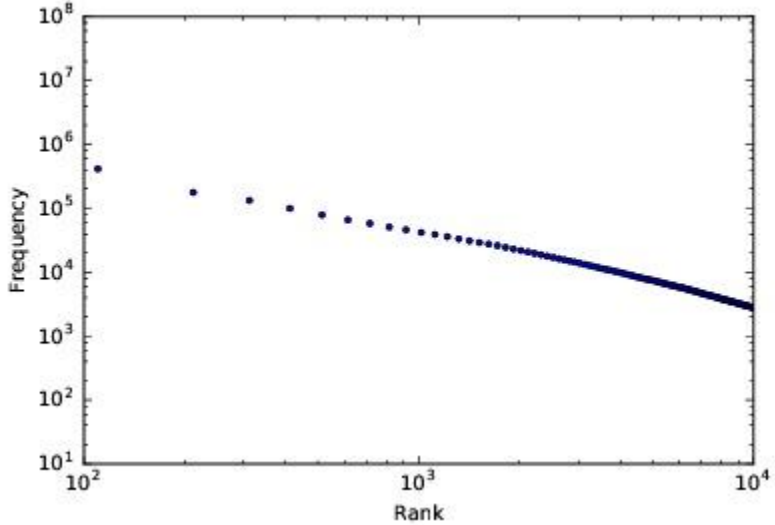
How do we this plot of word frequency?



Power Laws Need Log Plots!

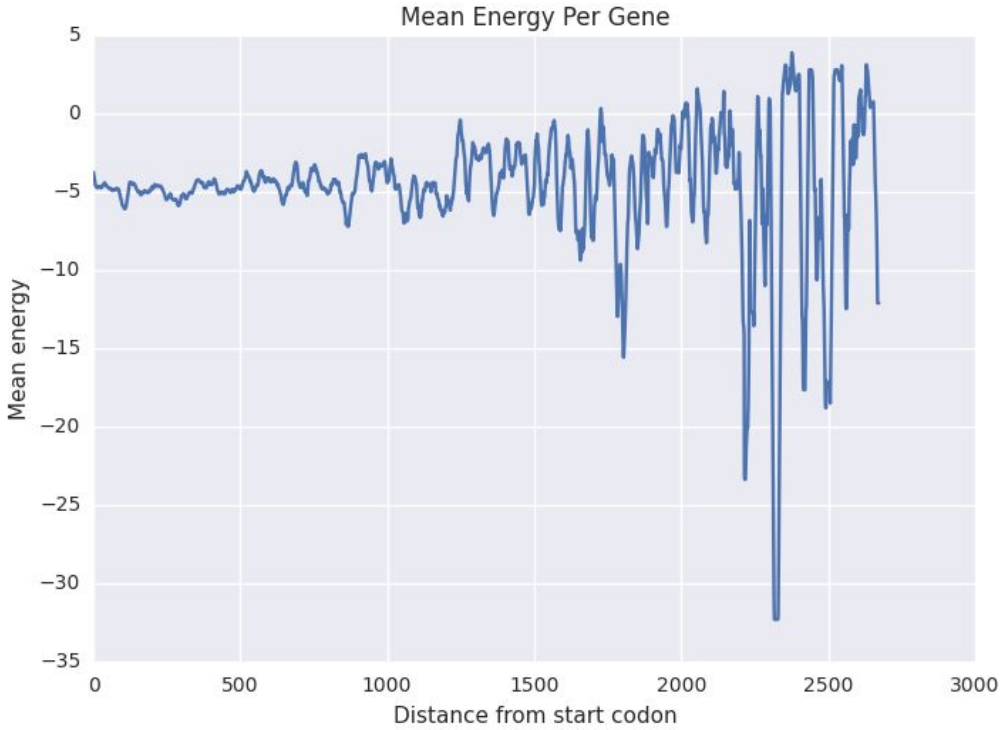
Log-Log plots can be even more revealing.





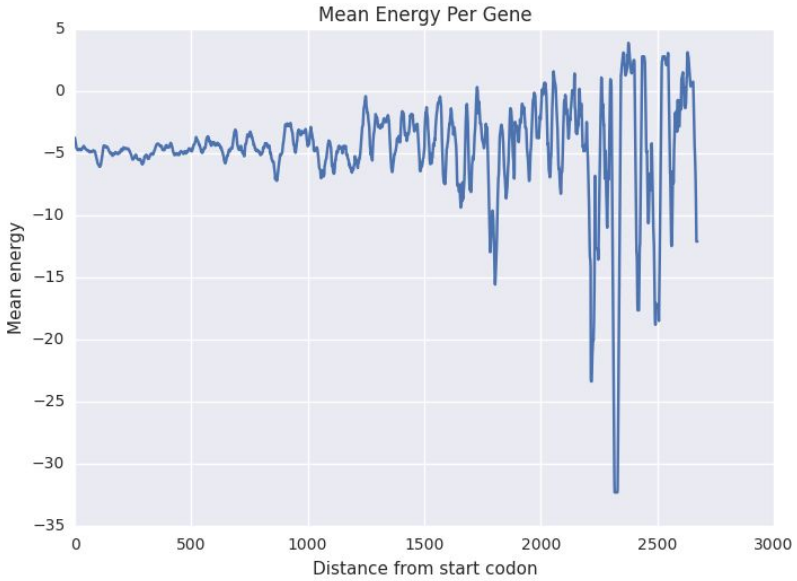
What Does this Graph Say?

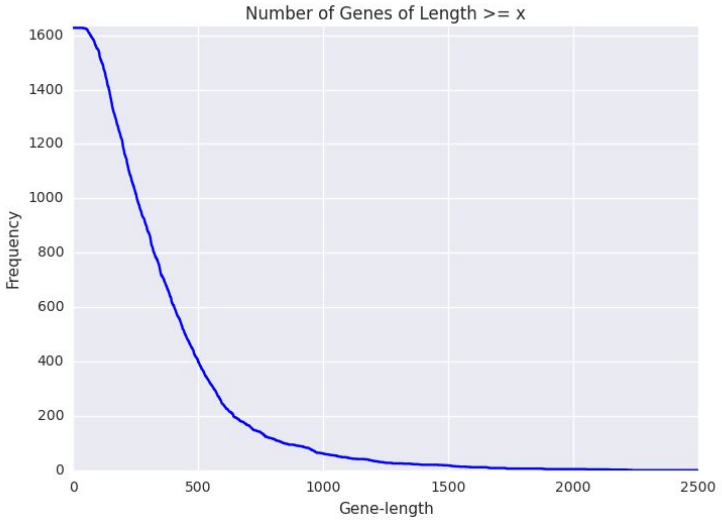
What is the trend that you see in this plot as a function of distance from starting position?



Overinterpreting Variance

With so few long genes the tail should be cut.



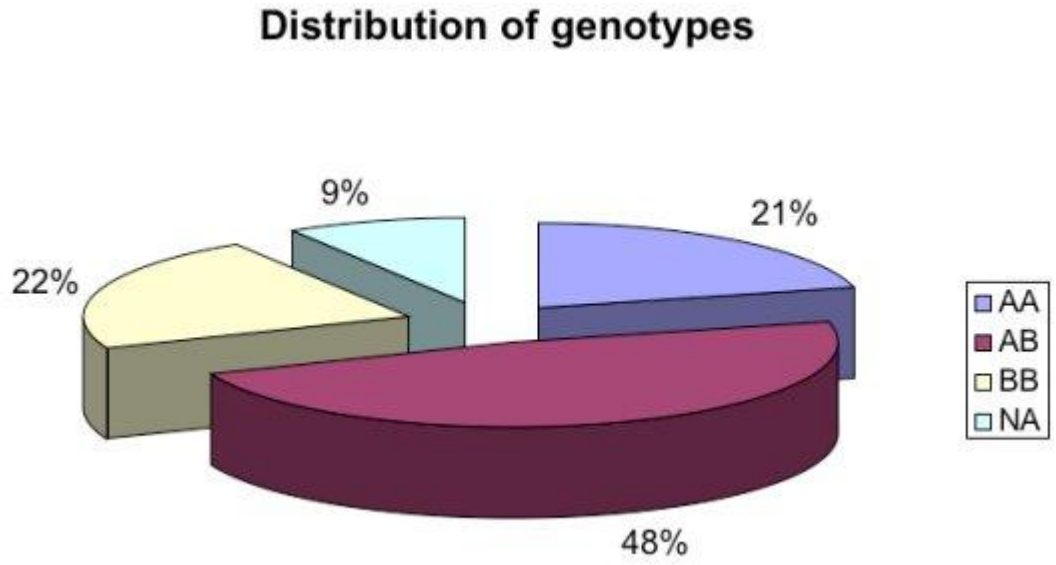


Terrible Professional Visualizations..

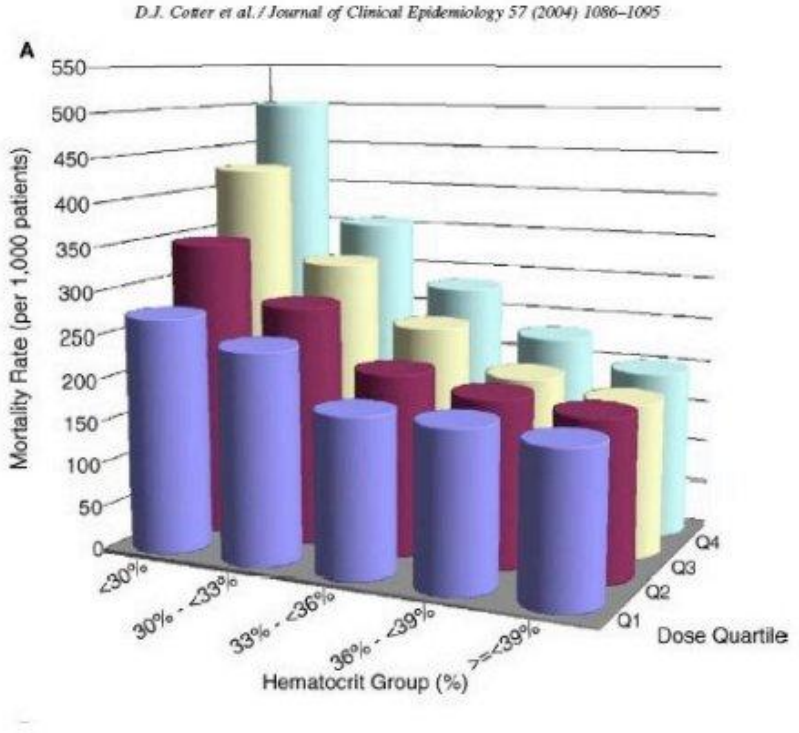
- Display as little information as possible.
- Obscure your data with chart junk, like pseudo-3D and excess color.
- Use poorly chosen scales.

Examples taken from http://wtfviz.net/

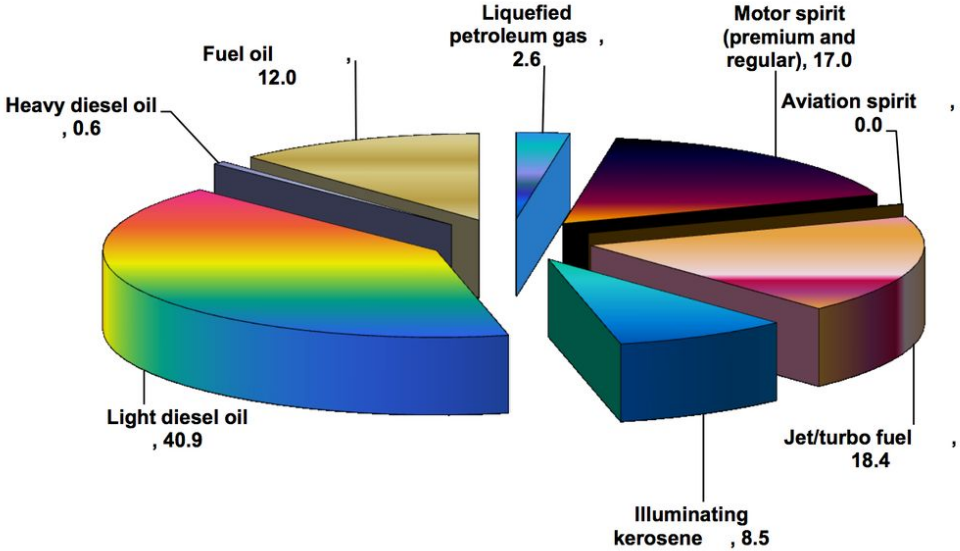
What is Wrong with this Pie Chart?



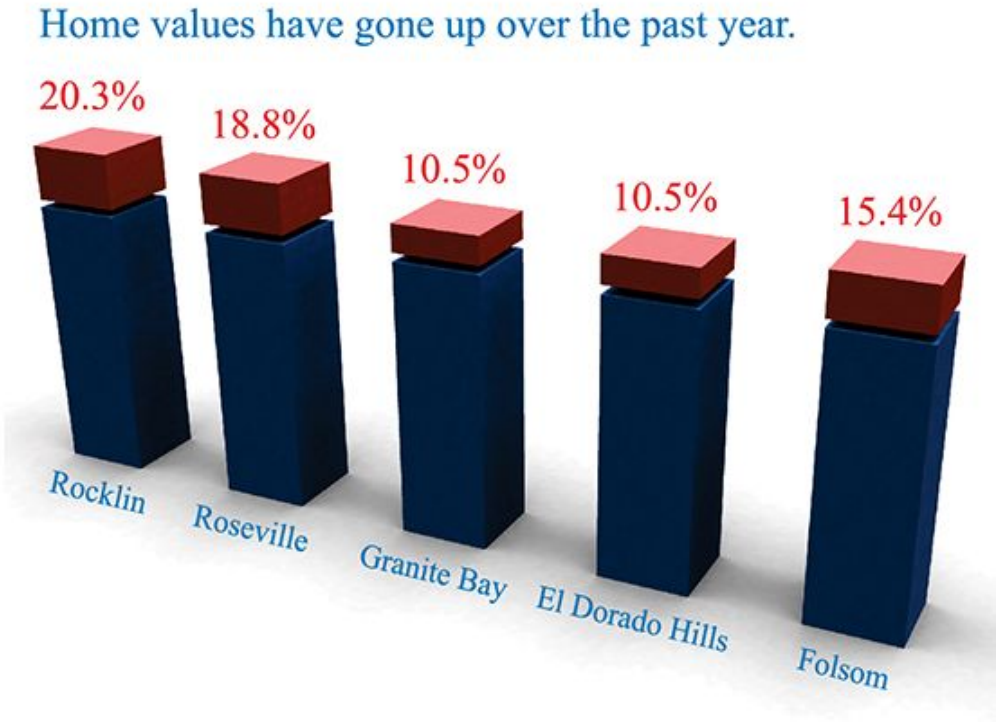
What's Wrong with this Bar Chart?



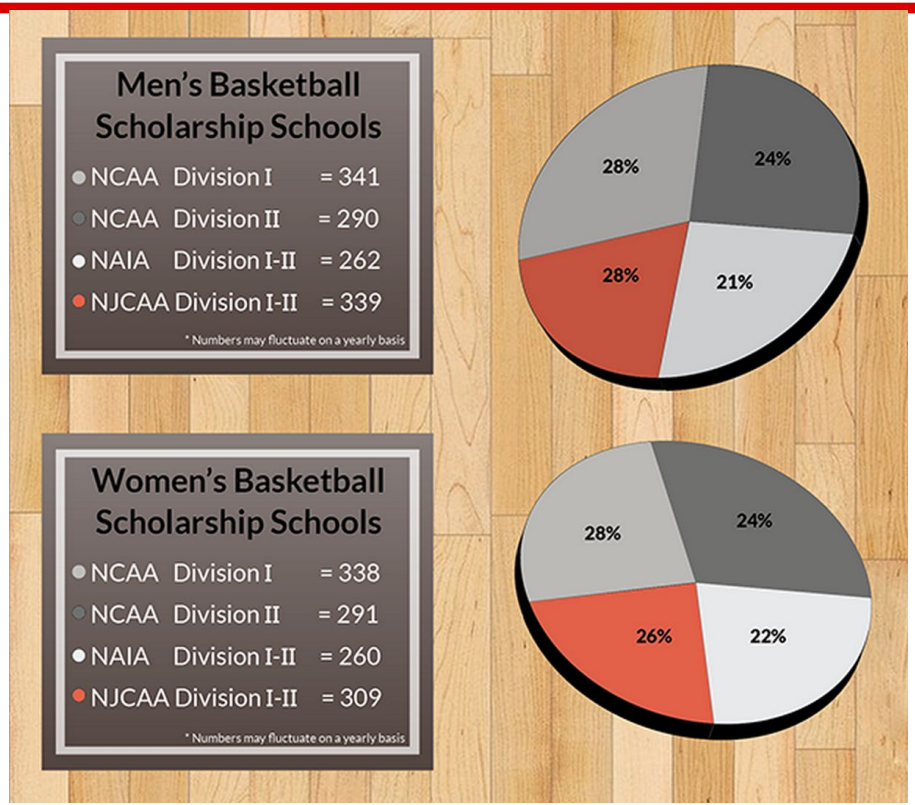
Color and Dimensionality



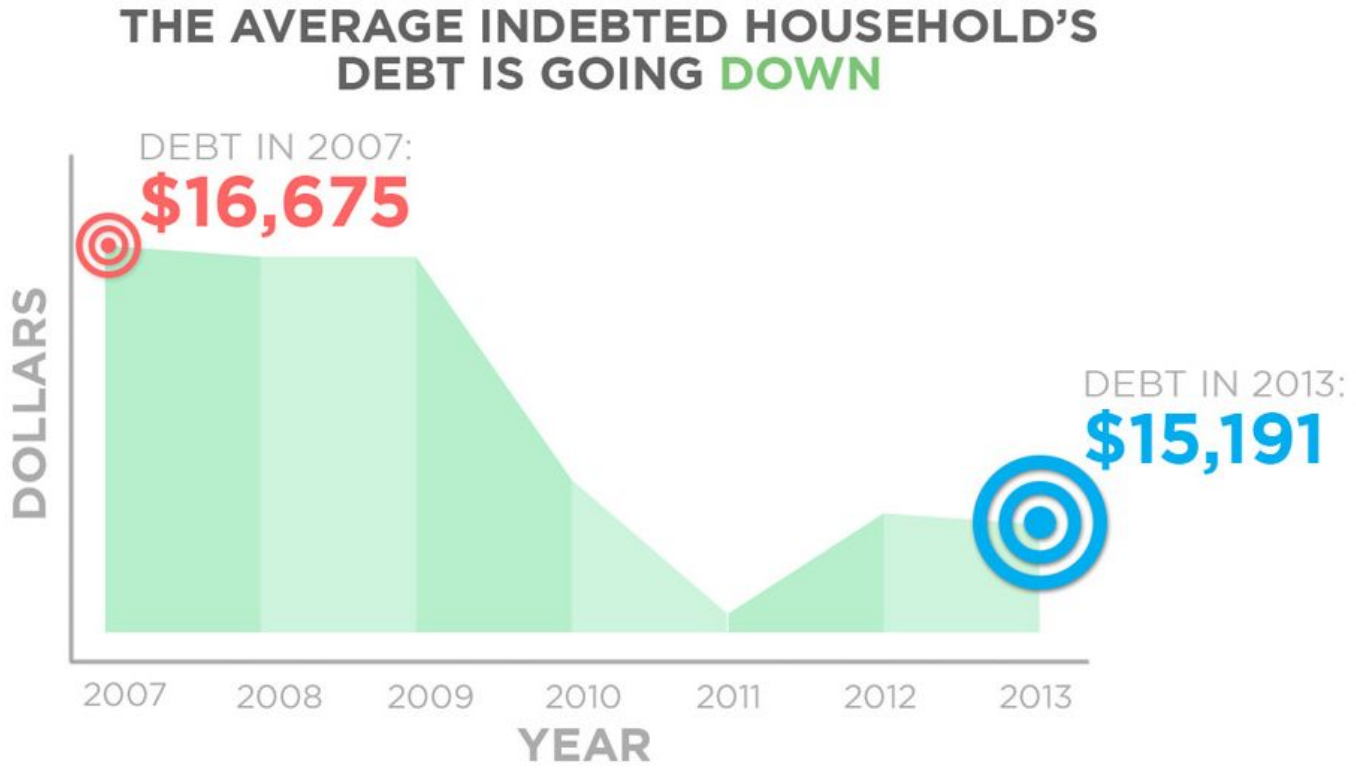
Volume/Value Comparisons



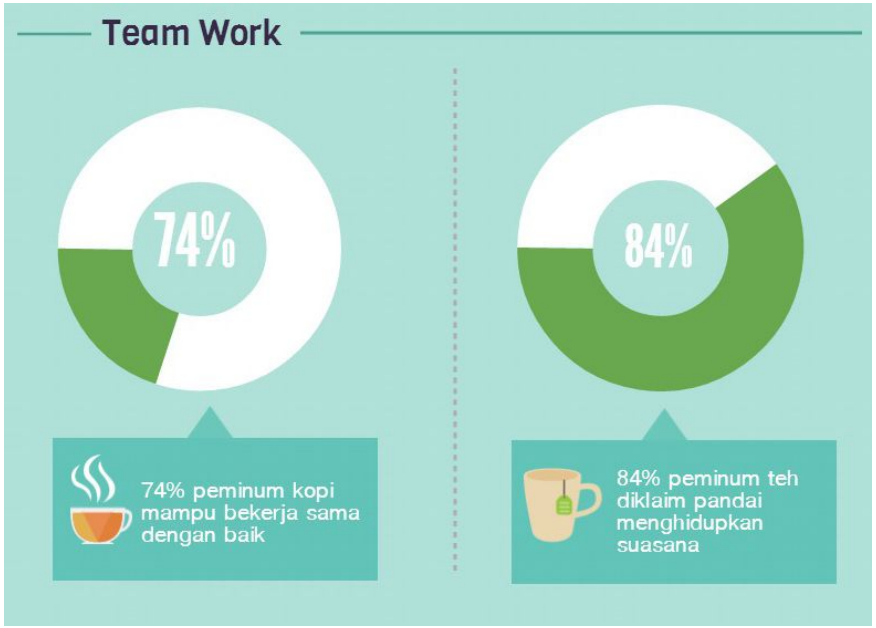
Oval Pie Charts?

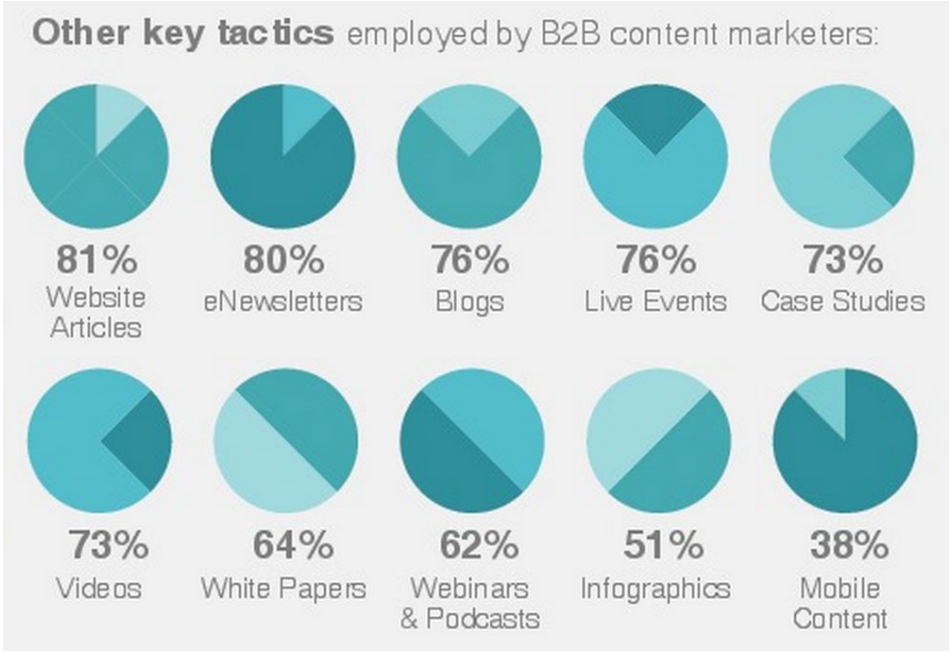


Range, Caption, and Symbol Sins

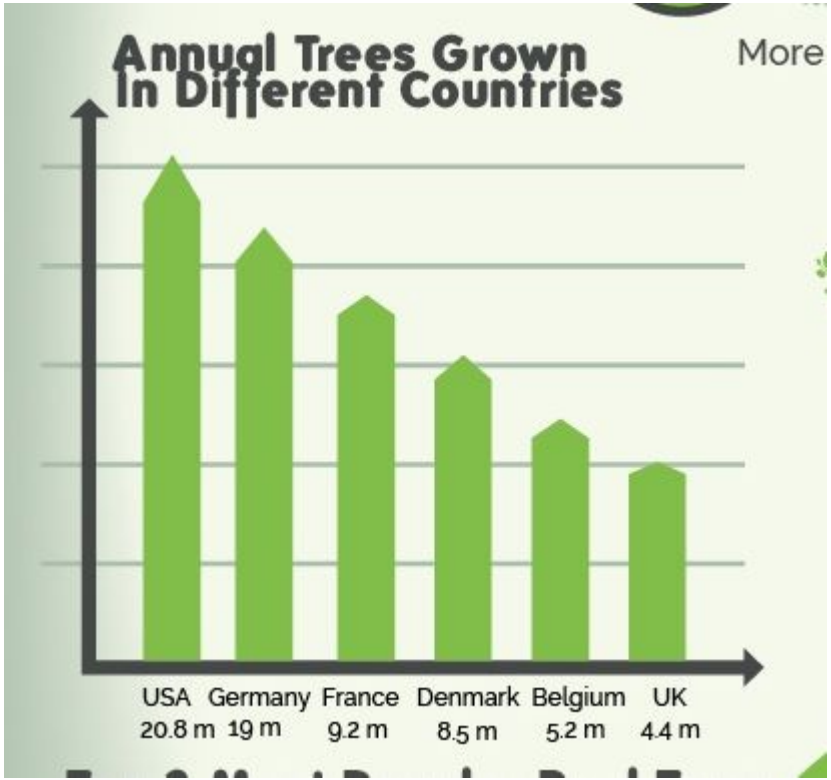


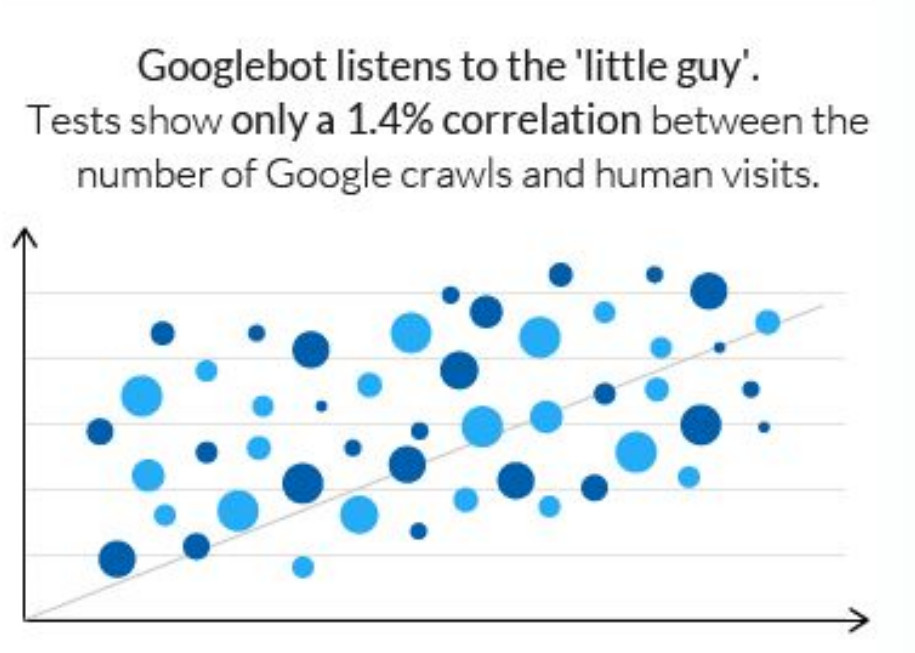
The Virtues of Consistency



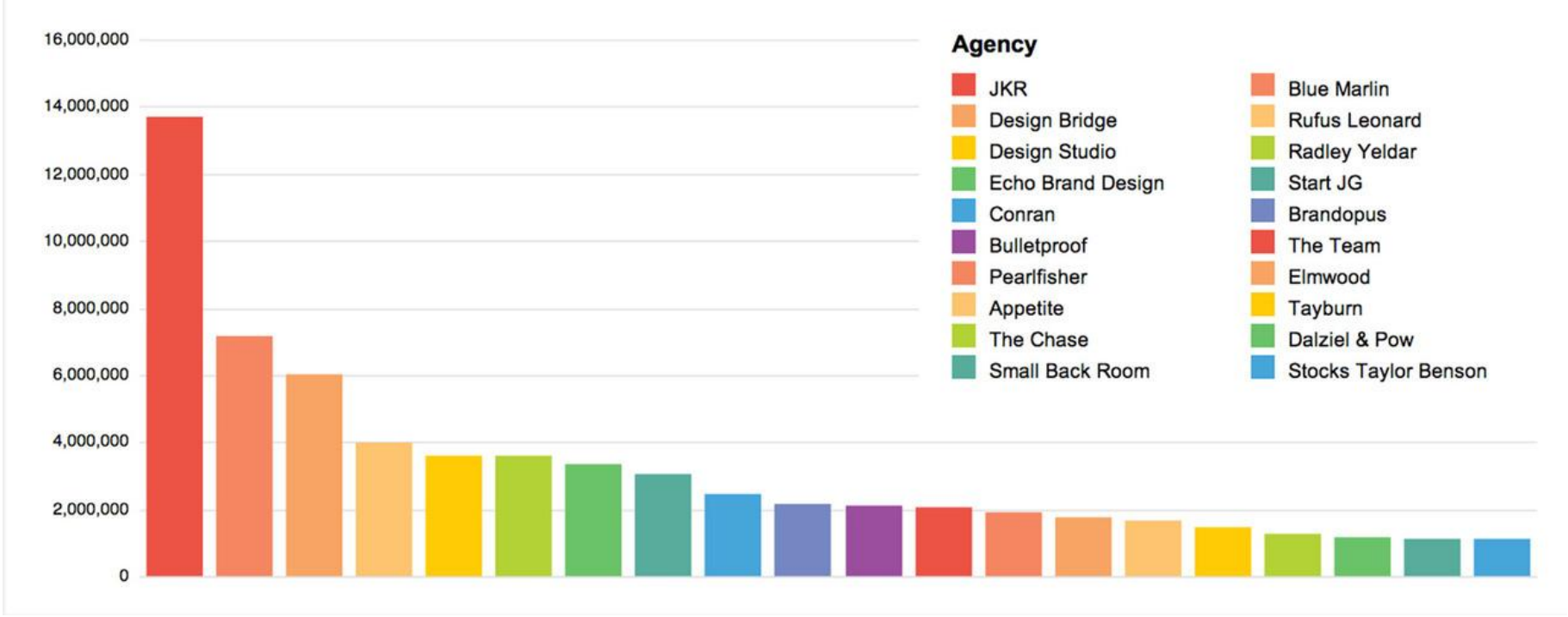


Provably Meaningless Charts

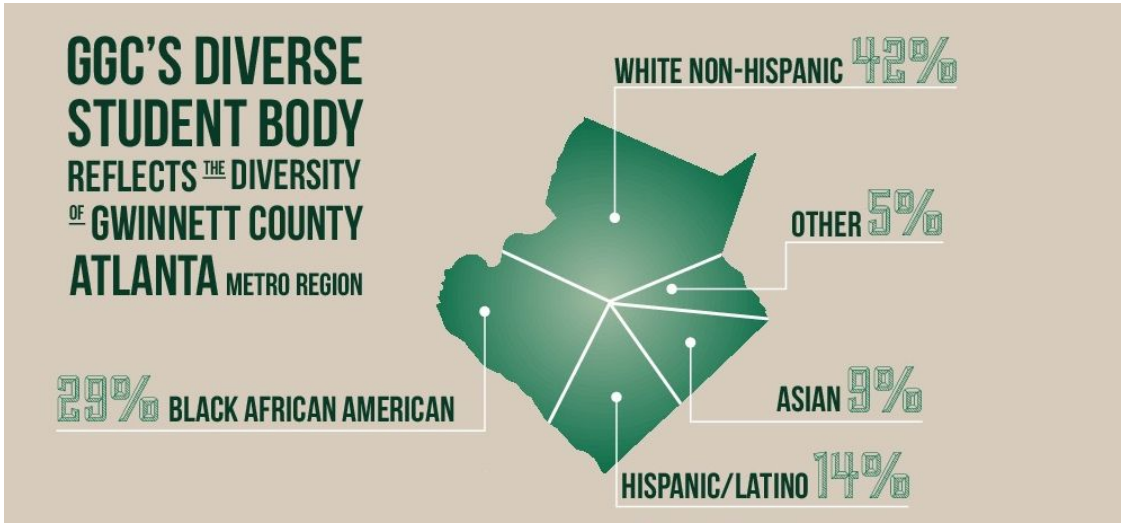


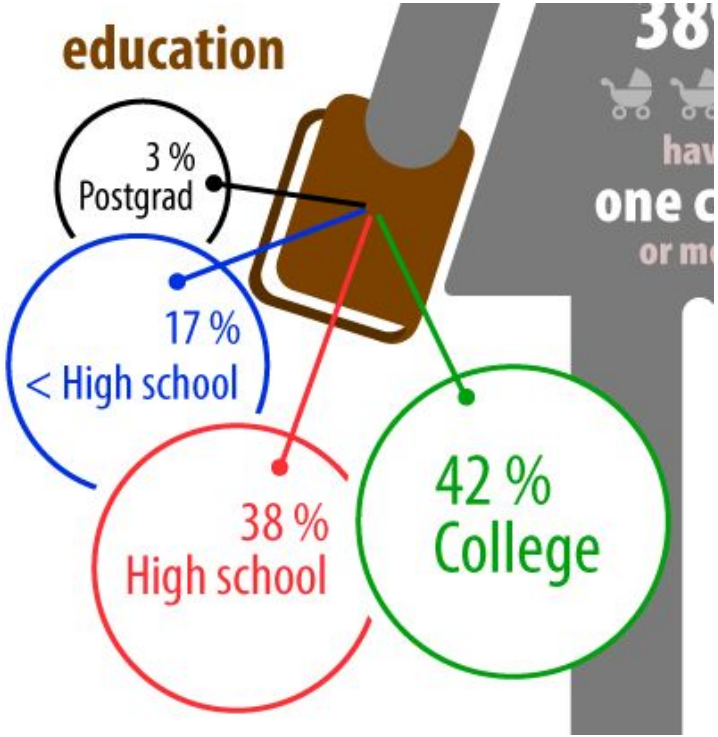


Dramatic Misuse/Reuse of Color

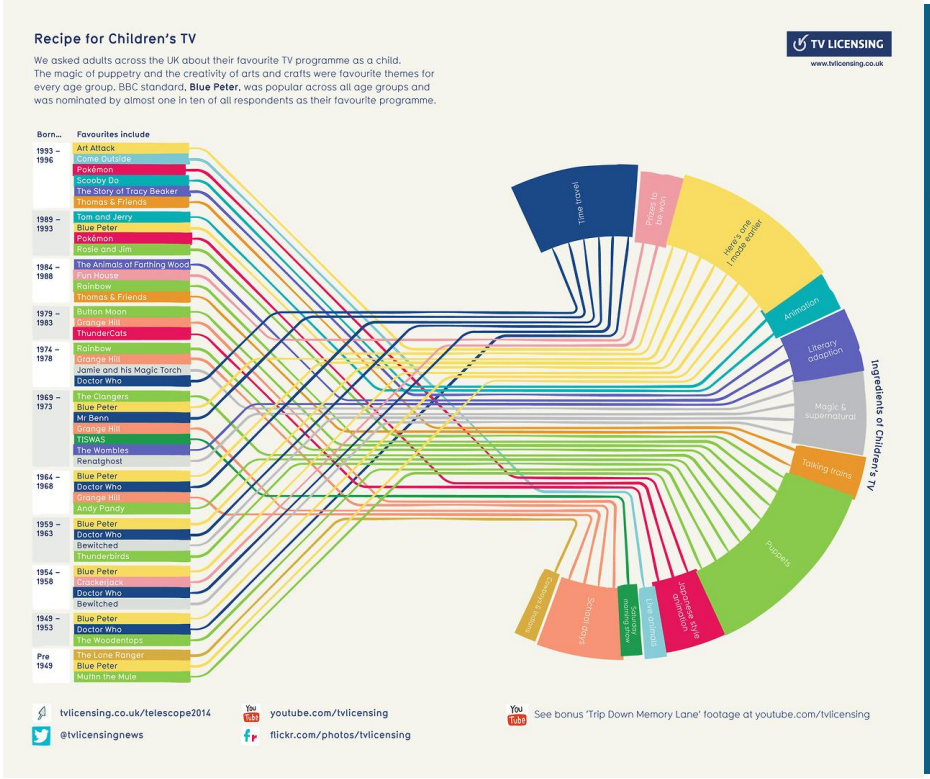


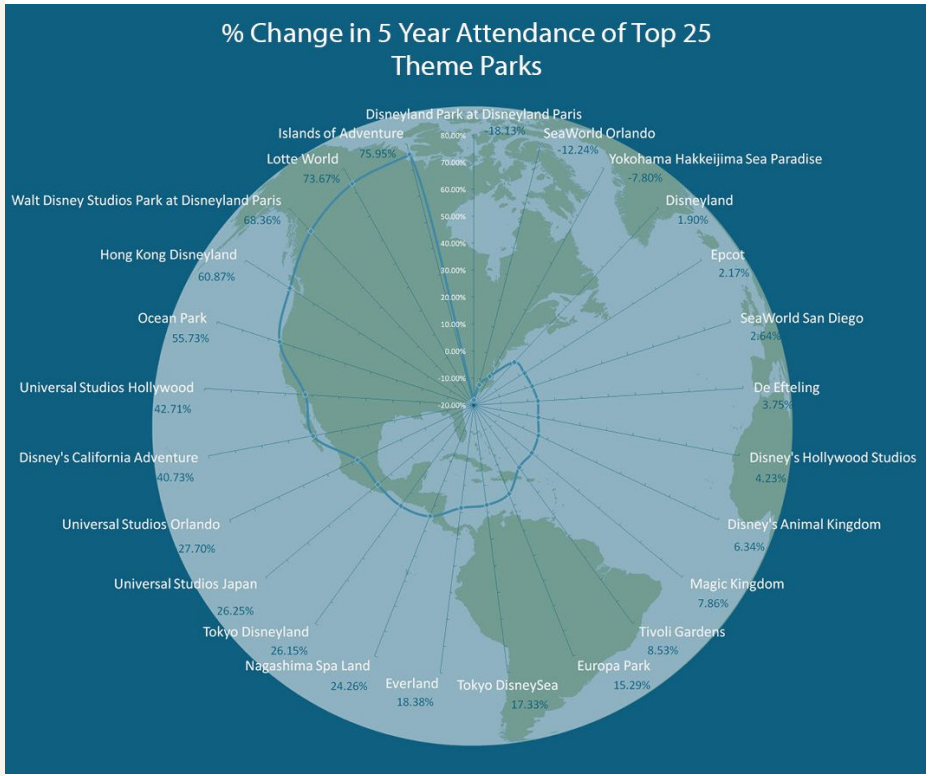
Graphics Size Matters



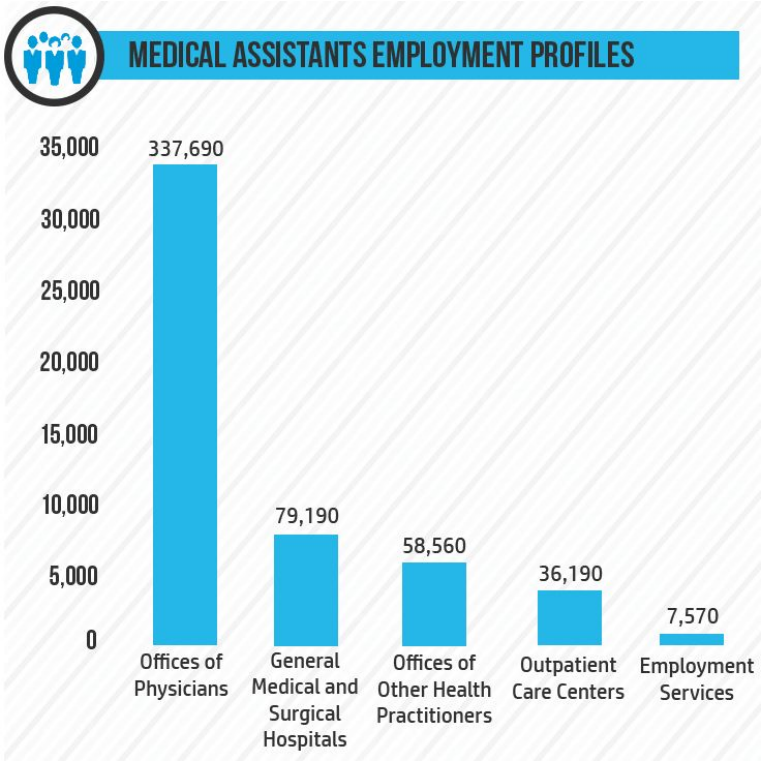


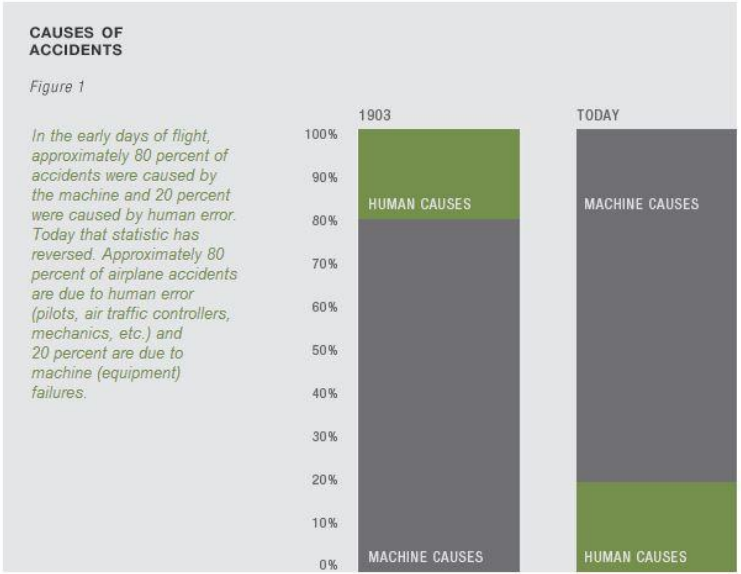
Impressive Chart Junk



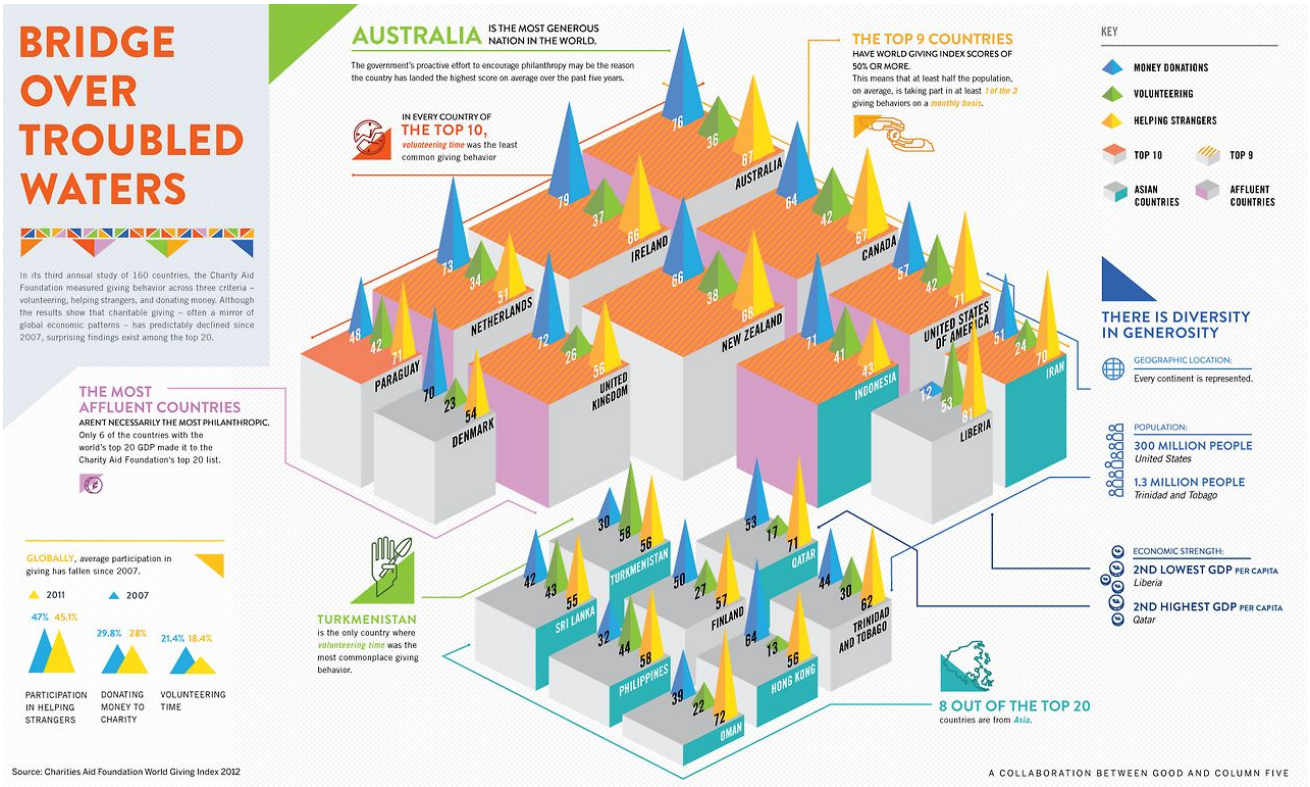


Labels Matter



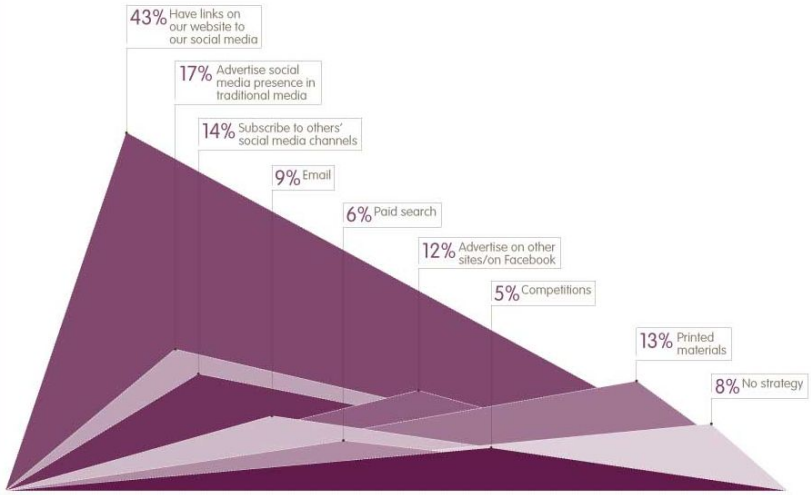


Way Too Many D

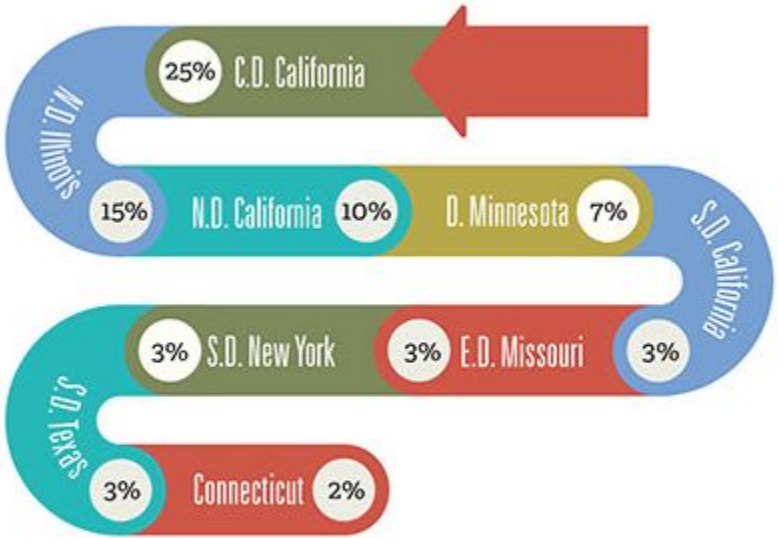


Size and Ordering Implications

How large businesses drive traffic to their social media



Courts in Which Complaints Were Filed



The largest number of complaints filed this quarter were filed in the Central District of California (25%) and the Northern District of Illinois (15% of complaints). The following chart shows the courts in which complaints were filed.

Keep a Critical Eye

Remember Tufte's principles whenever designing or interpreting data visualizations:

- Maximize data-ink ratio
- Minimize lie factor
- Minimize chartjunk
- Use proper scales and clear labeling

Beautiful data deserves beautiful visualization.