

---

# **CSE 591: Data Science**

## **Steven Skiena**

### **Stony Brook University**

---

Lecture 12: Practice of Data  
Visualization

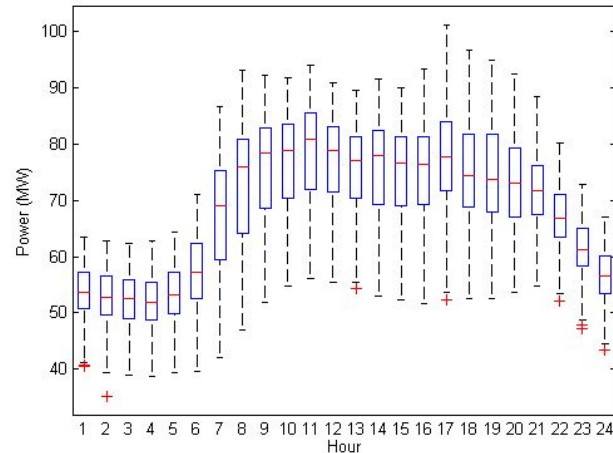
---

# Box and Whisker Plots

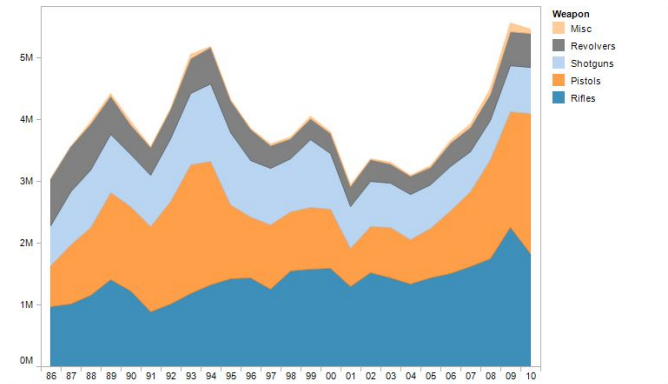
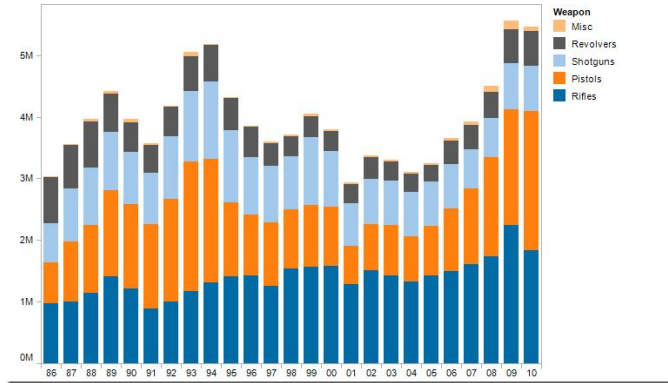
---

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

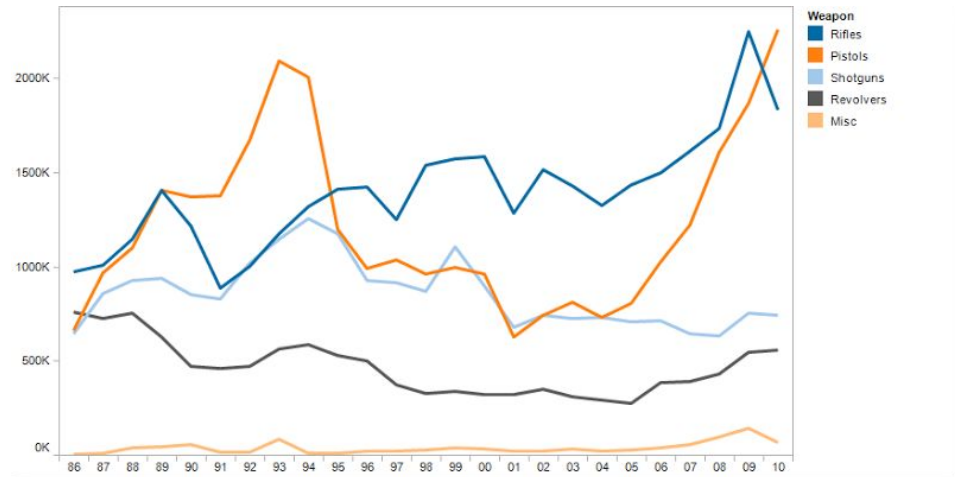
I personally prefer contour lines without the boxes.



# Stacked Area vs. Line Plots

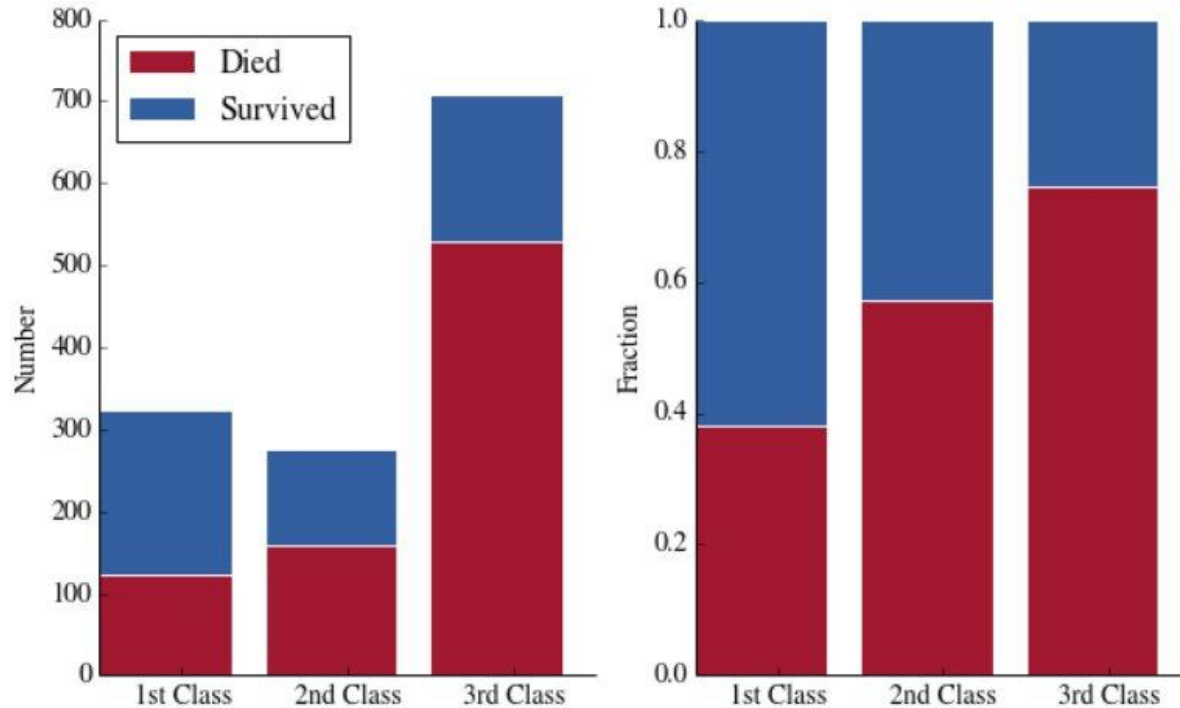


Hard to see trends in middle areas of stack:



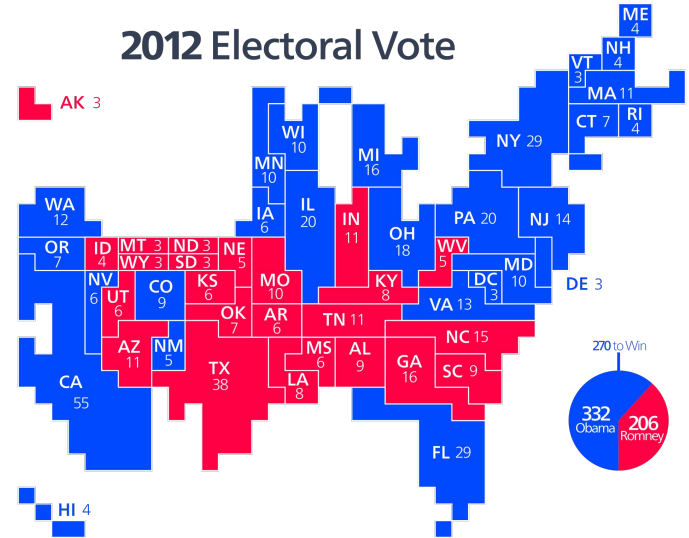
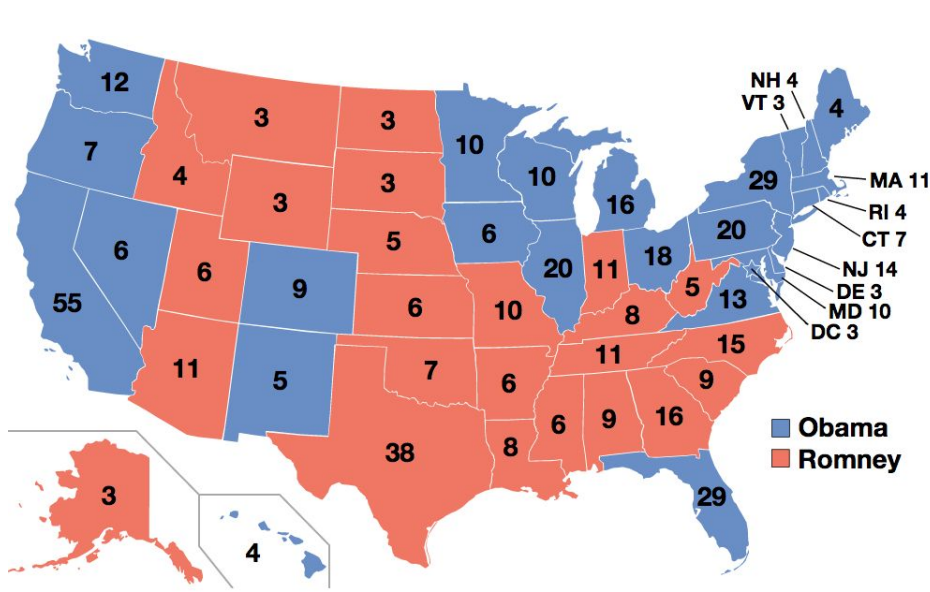
# Stacked Bar Charts: Titanic

---



# Data Maps and Cartograms

Cartograms distort regions to reflect an underlying variable.



# Non-Geographic Data Maps

Periodic Table of the Elements																		18 VIII 8A																	
1 IA 1A		2 IIA 2A												13 IIIA 3A		14 IVA 4A		15 VA 5A		16 VIA 6A		17 VIIA 7A		2 He Helium 4.003											
3 Li Lithium 6.941		4 Be Beryllium 9.012												5 B Boron 10.811		6 C Carbon 12.011		7 N Nitrogen 14.007		8 O Oxygen 15.999		9 F Fluorine 18.998		10 Ne Neon 20.180											
11 Na Sodium 22.990		12 Mg Magnesium 24.305												13 Al Aluminum 26.982		14 Si Silicon 28.086		15 P Phosphorus 30.974		16 S Sulfur 32.066		17 Cl Chlorine 35.453		18 Ar Argon 39.948											
19 K Potassium 39.098		20 Ca Calcium 40.078		3 IIIB 3B		4 IVB 4B		5 VB 5B		6 VIB 6B		7 VIIB 7B		8 VIII 8		9 VIII 8		10 VIII 8		11 IB 1B		12 IIB 2B		31 Ga Gallium 69.722		32 Ge Germanium 72.61		33 As Arsenic 74.922		34 Se Selenium 78.972		35 Br Bromine 79.904		36 Kr Krypton 84.80	
37 Rb Rubidium 84.468		38 Sr Strontium 87.62		39 Y Yttrium 88.906		40 Zr Zirconium 91.224		41 Nb Niobium 92.906		42 Mo Molybdenum 95.95		43 Tc Technetium 98.907		44 Ru Ruthenium 101.07		45 Rh Rhodium 102.906		46 Pd Palladium 106.42		47 Ag Silver 107.868		48 Cd Cadmium 112.411		49 In Indium 114.818		50 Sn Tin 118.71		51 Sb Antimony 121.760		52 Te Tellurium 127.6		53 I Iodine 126.904		54 Xe Xenon 131.29	
55 Cs Cesium 132.905		56 Ba Barium 137.327		57-71		72 Hf Hafnium 178.49		73 Ta Tantalum 180.948		74 W Tungsten 183.85		75 Re Rhenium 186.207		76 Os Osmium 190.23		77 Ir Iridium 192.22		78 Pt Platinum 195.08		79 Au Gold 196.967		80 Hg Mercury 200.59		81 Tl Thallium 204.383		82 Pb Lead 207.2		83 Bi Bismuth 208.980		84 Po Polonium [209]		85 At Astatine [210]		86 Rn Radon [222]	
87 Fr Francium 223.020		88 Ra Radium 226.025		89-103		104 Rf Rutherfordium [261]		105 Db Dubnium [262]		106 Sg Seaborgium [266]		107 Bh Bohrium [264]		108 Hs Hassium [269]		109 Mt Meitnerium [268]		110 Ds Darmstadtium [269]		111 Rg Roentgenium [272]		112 Cn Copernicium [277]		113 Uut Ununtrium [288]		114 Fl Flerovium [289]		115 Uup Ununpentium [289]		116 Lv Livermorium [293]		117 Uus Ununseptium [294]		118 Uuo Ununoctium [294]	
Lanthanide Series				57 La Lanthanum 138.906		58 Ce Cerium 140.115		59 Pr Praseodymium 140.908		60 Nd Neodymium 144.24		61 Pm Promethium 144.913		62 Sm Samarium 150.36		63 Eu Europium 151.966		64 Gd Gadolinium 157.25		65 Tb Terbium 158.925		66 Dy Dysprosium 162.50		67 Ho Holmium 164.930		68 Er Erbium 167.26		69 Tm Thulium 168.934		70 Yb Ytterbium 173.04		71 Lu Lutetium 174.967			
Actinide Series				89 Ac Actinium 227.028		90 Th Thorium 232.038		91 Pa Protactinium 231.036		92 U Uranium 238.029		93 Np Neptunium 237.048		94 Pu Plutonium 244.064		95 Am Americium 243.061		96 Cm Curium 247.070		97 Bk Berkelium 247.070		98 Cf Californium 251.080		99 Es Einsteinium [254]		100 Fm Fermium 257.095		101 Md Mendelevium 258.1		102 No Nobelium 259.101		103 Lr Lawrencium [262]			
Alkali Metal		Alkaline Earth		Transition Metal		Basic Metal		Semimetal		Nonmetal		Halogen		Noble Gas		Lanthanide		Actinide																	
																		© 2014 Todd Helmenstine sciencemaps.org																	

# Tools for Data Visualization

---

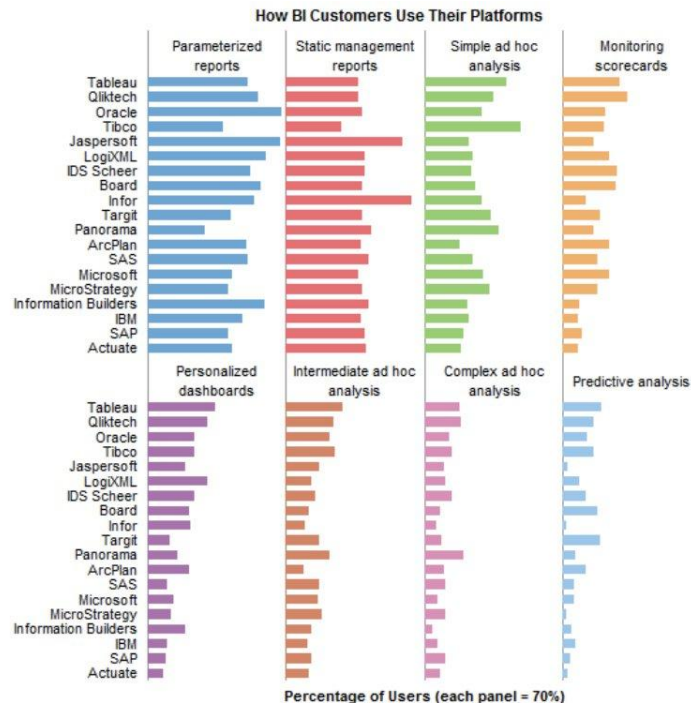
Just because Excel is very popular does not mean it produces good graphs/plots.

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

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

---

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



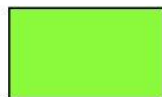


# Understanding Color Scales

---

Perceived as Ordered

Brightness



Saturation



Hue: not as much

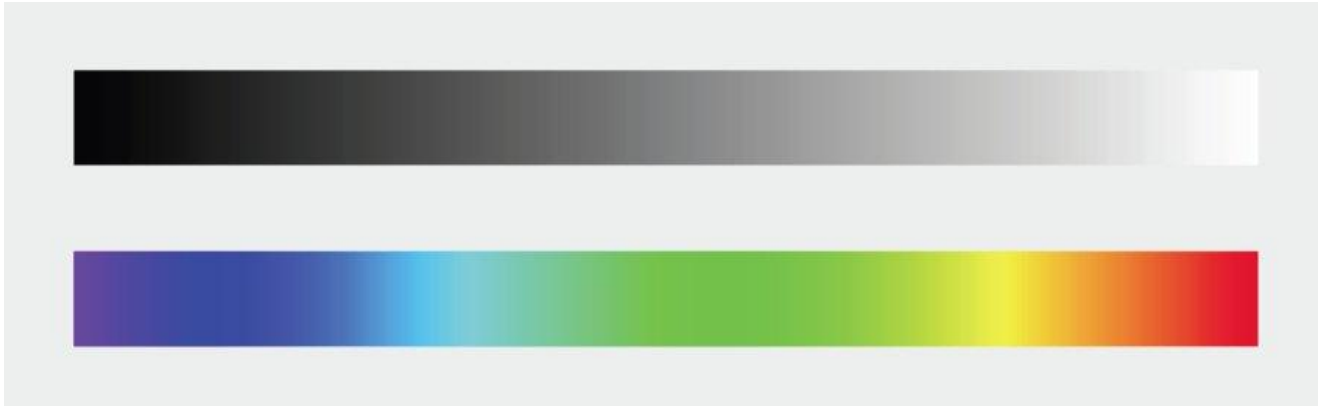


# Rainbow Color Maps

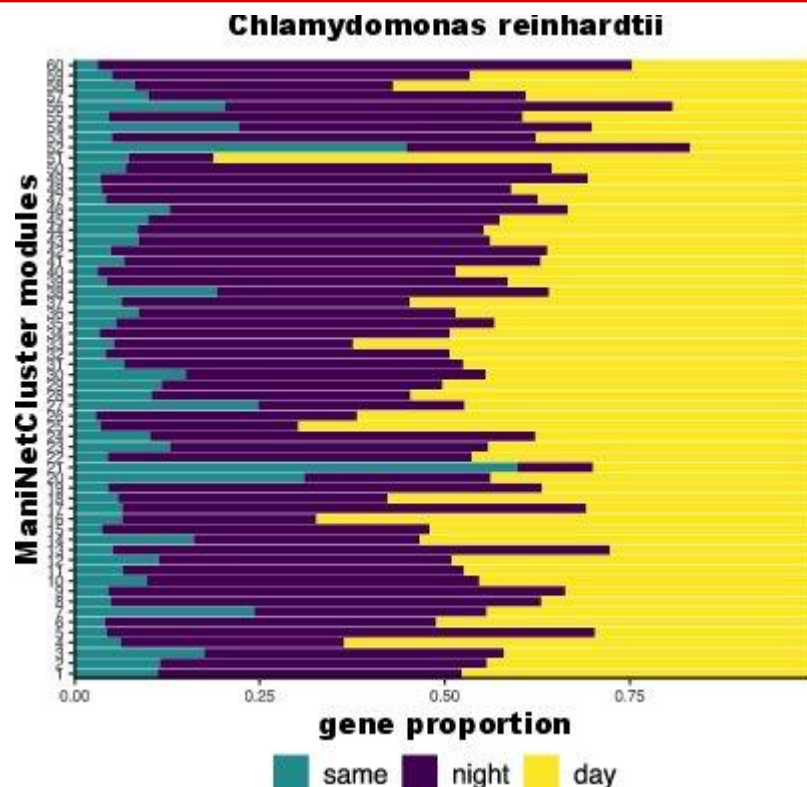
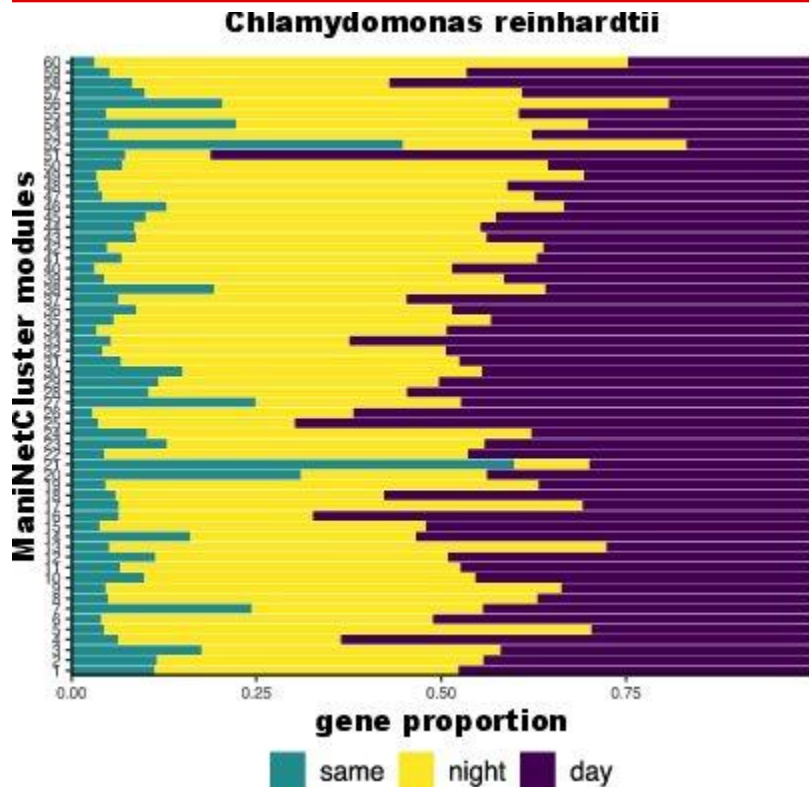
---

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

---

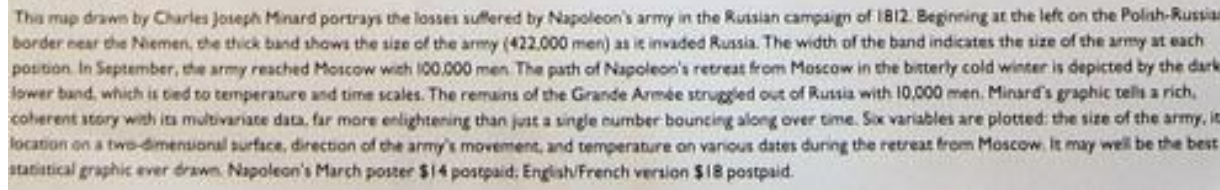
Distinguishing good/bad visualizations requires a design aesthetic, and a vocabulary to talk about data representations:

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

# Great Data Visualizations

---

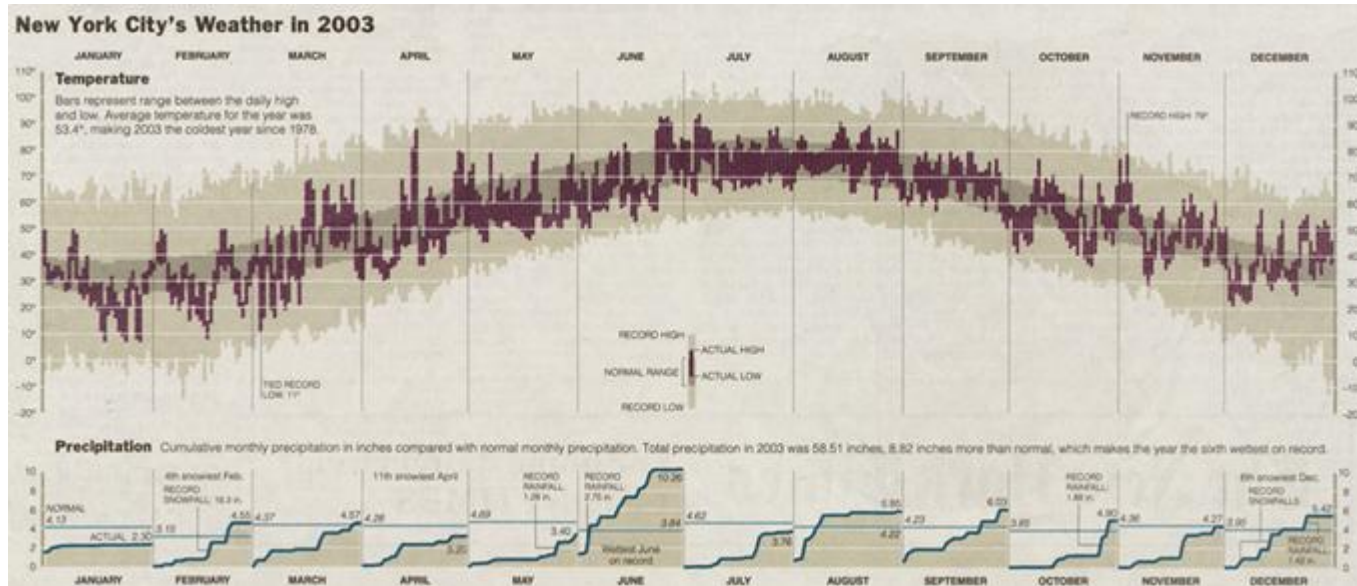
- Display data accurately and clearly.
  - Tell a story that the data reveals.
  - Are rich enough to make you want to look carefully and study the data.
-





# New York's Weather Year in Review

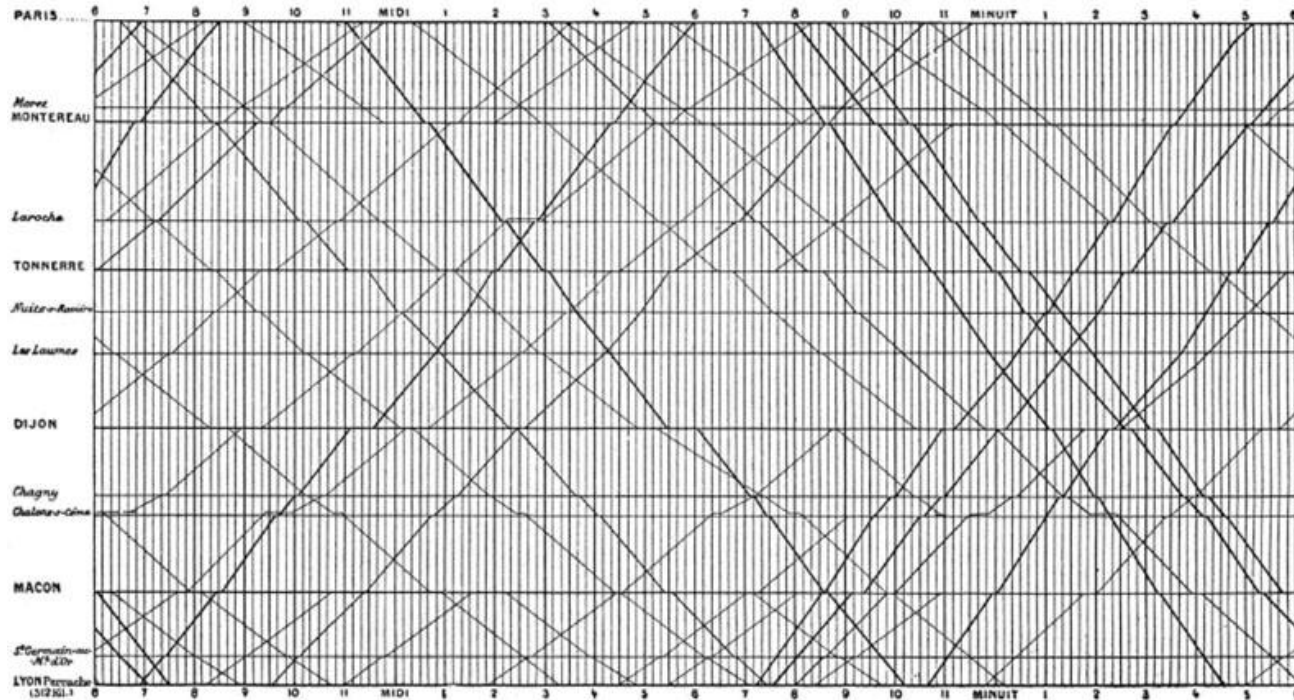
A clear story displaying over 3,000 numbers.





# 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!

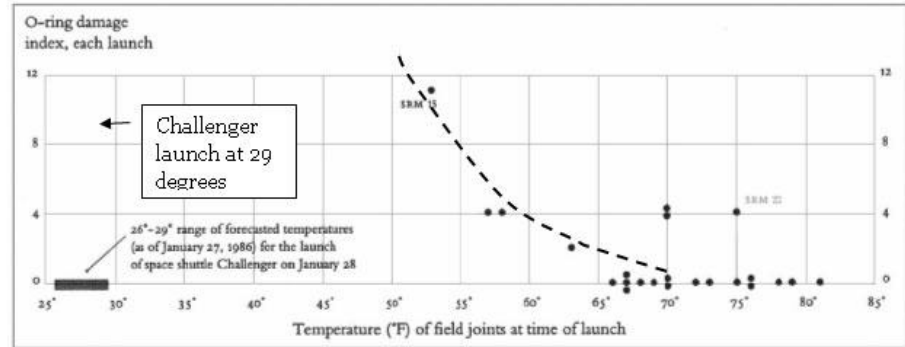
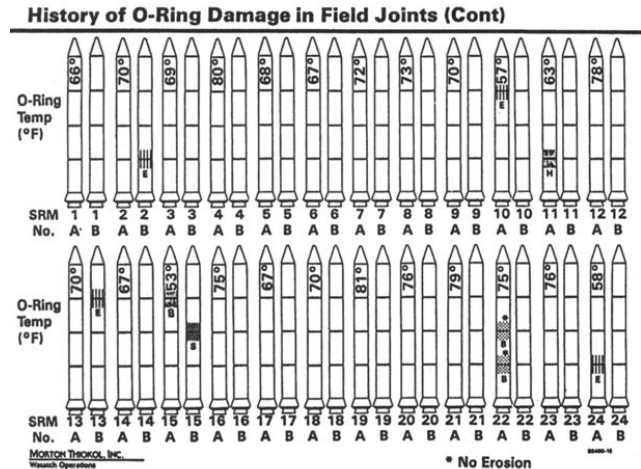
# Discovering the Source of Cholera

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.

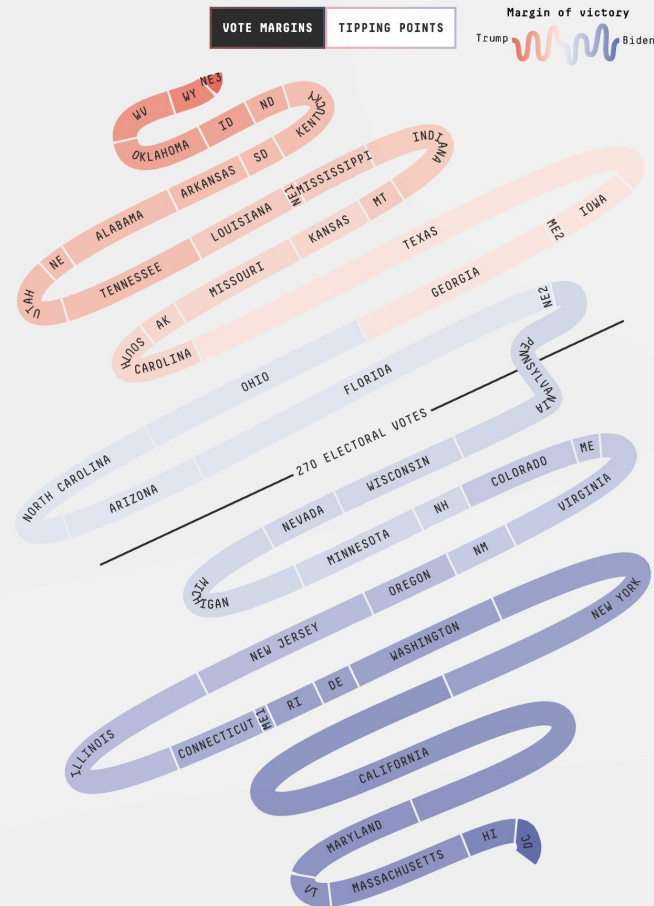


# State of the Race

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

## The winding path to victory

States that are forecasted to vote for one candidate by a big margin are at the ends of the path, while tighter races are in the middle. Bigger segments mean more Electoral College votes. Trace the path from either end to see which state could put one candidate over the top.

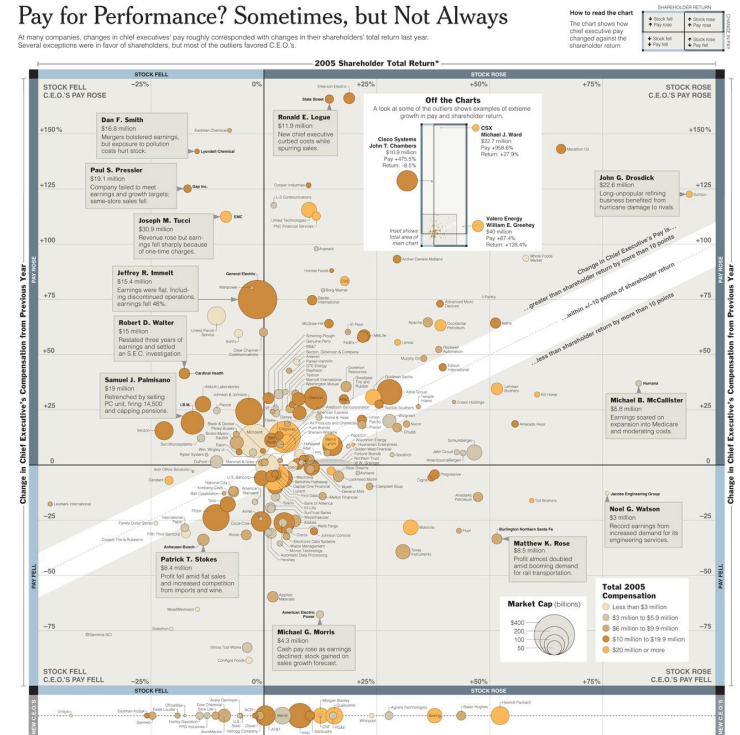


# Which Executives Earn their Pay?

Chart plots points in 4D, shown by x, y, color, size. Using attributes like point size/color is better than plotting points in 3D.

## Pay for Performance? Sometimes, but Not Always

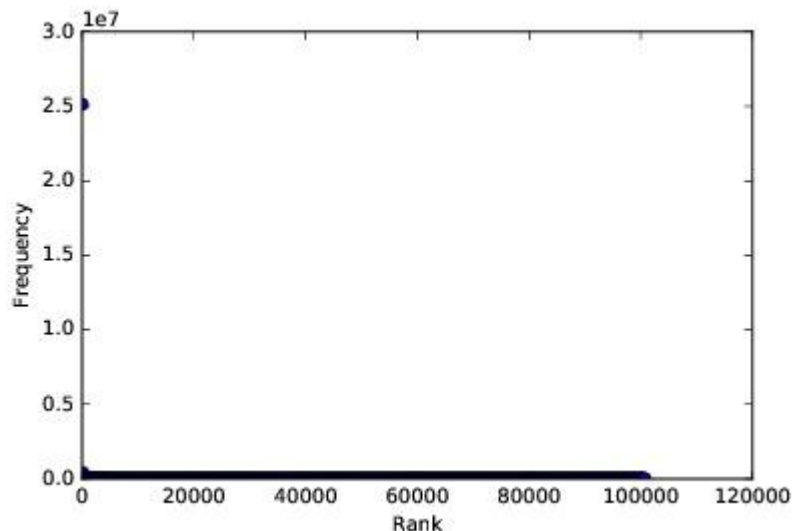
At many companies, changes in chief executive pay roughly corresponded with changes in their shareholders' total return last year. Several exceptions were in favor of shareholders, but most of the outliers favored C.E.O.'s.



# Terrible Student Visualizations...

---

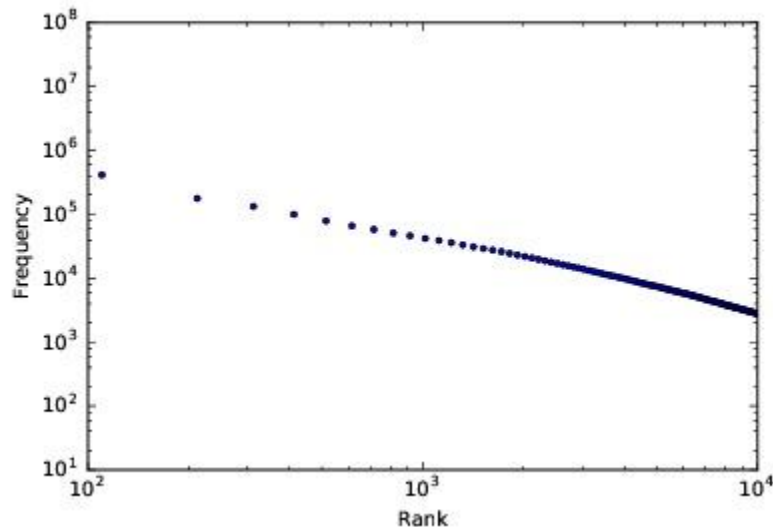
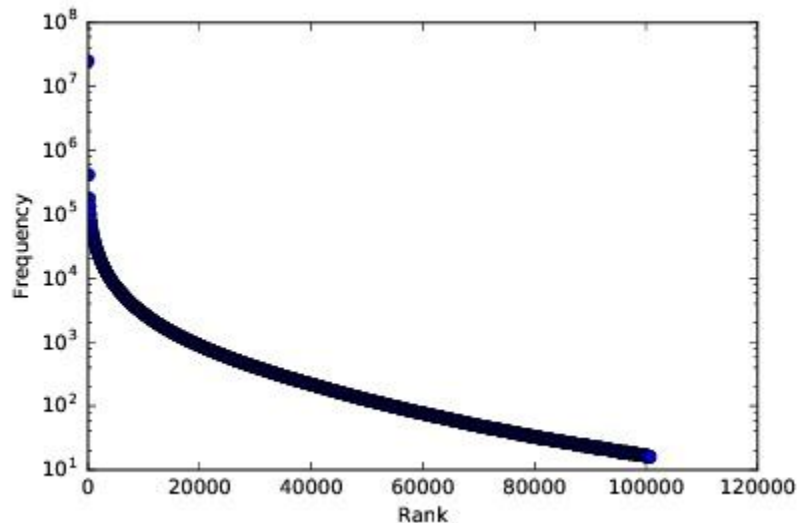
How do we this plot of word frequency?



# Power Laws Need Log Plots!

---

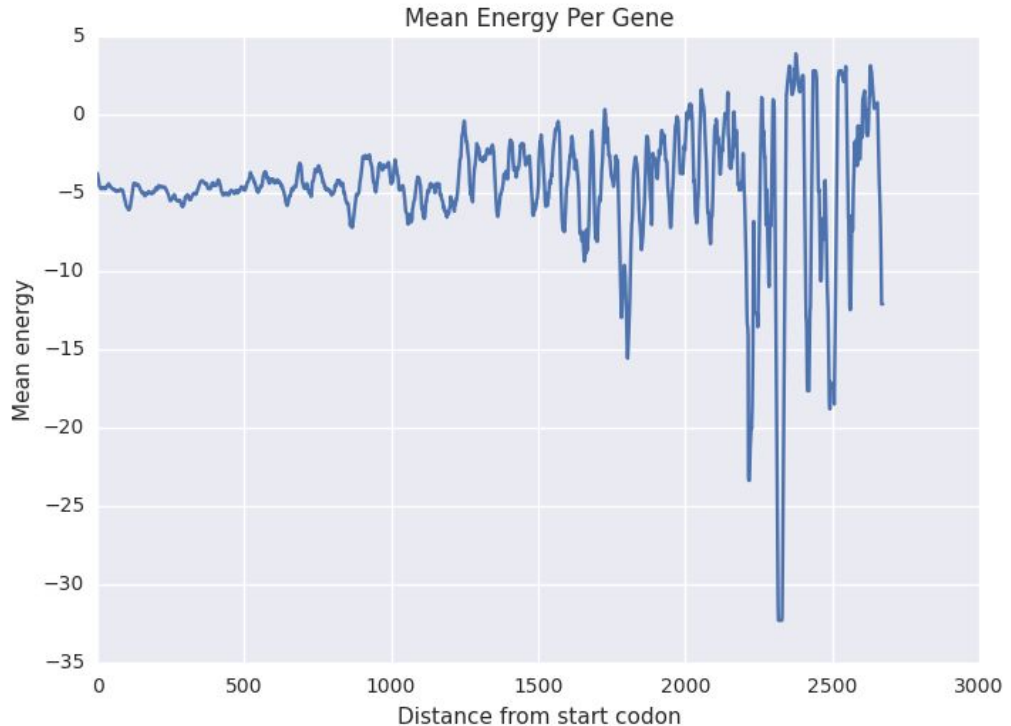
Log-Log plots can be even more revealing.



# What Does this Graph Say?

---

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

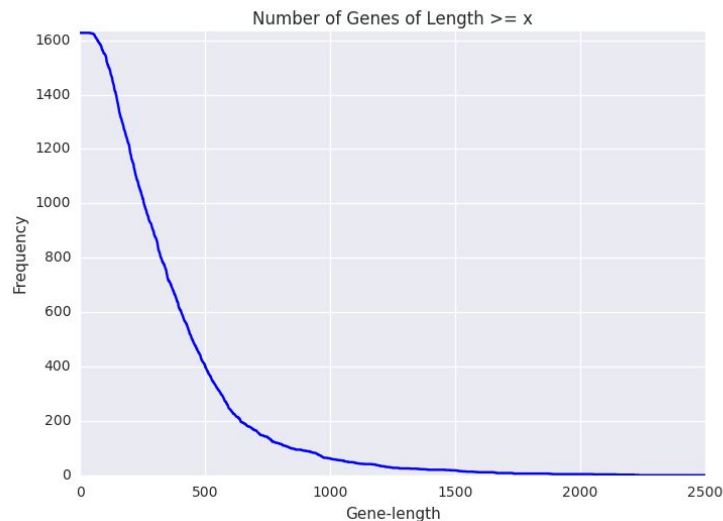
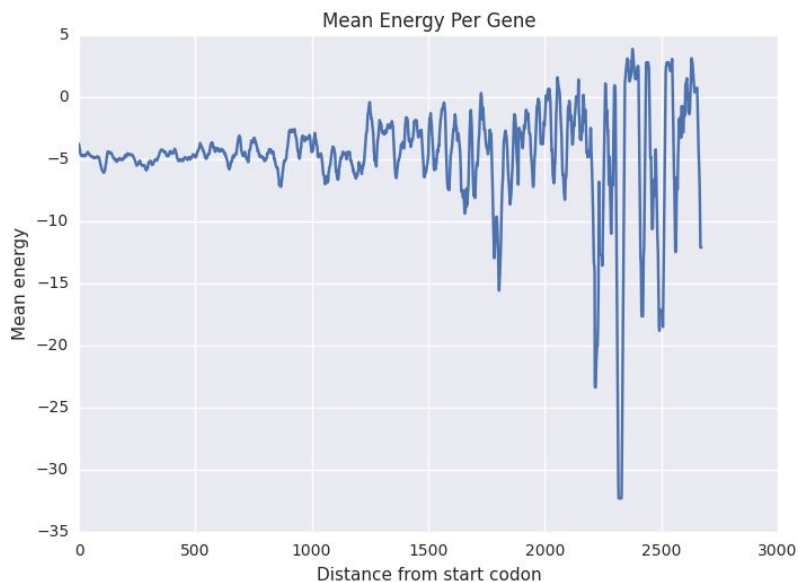




# Overinterpreting Variance

---

With so few long genes the tail should be cut.



# Terrible Professional Visualizations..

---

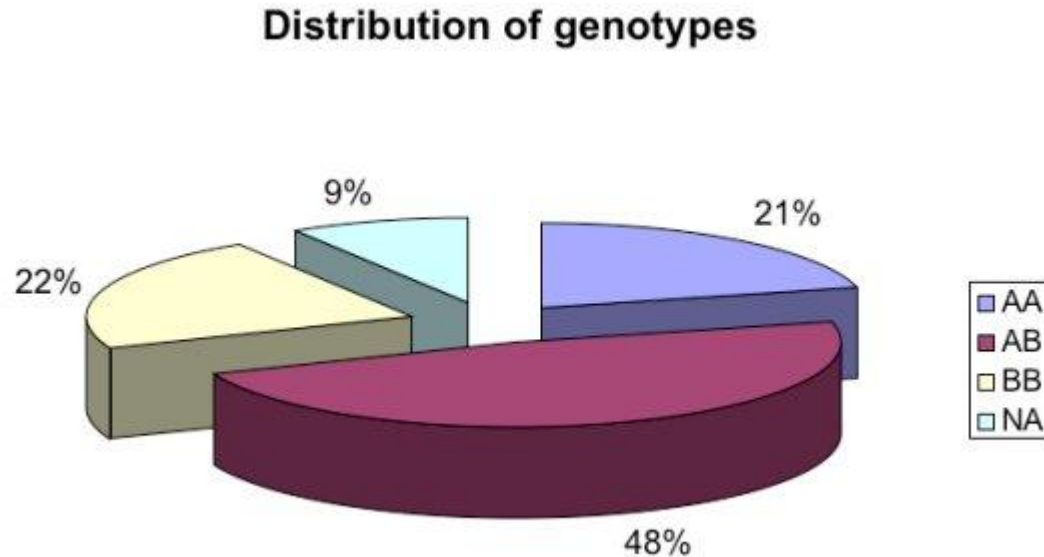
- Display as little information as possible.
- Obscure your data with chart junk, like pseudo-3D and excess color.
- Use poorly chosen scales.

Examples taken from <http://wtfviz.net/>

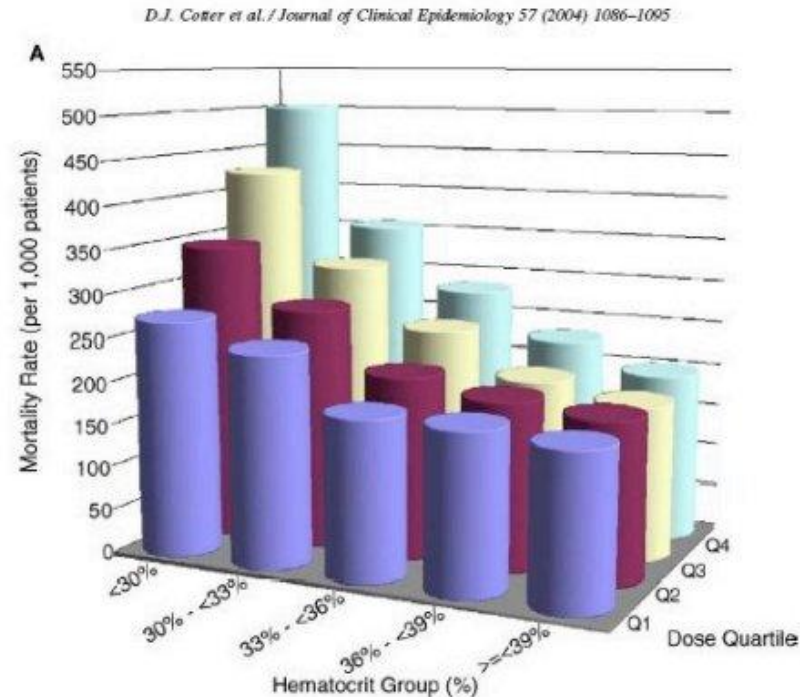
---

# What is Wrong with this Pie Chart?

---

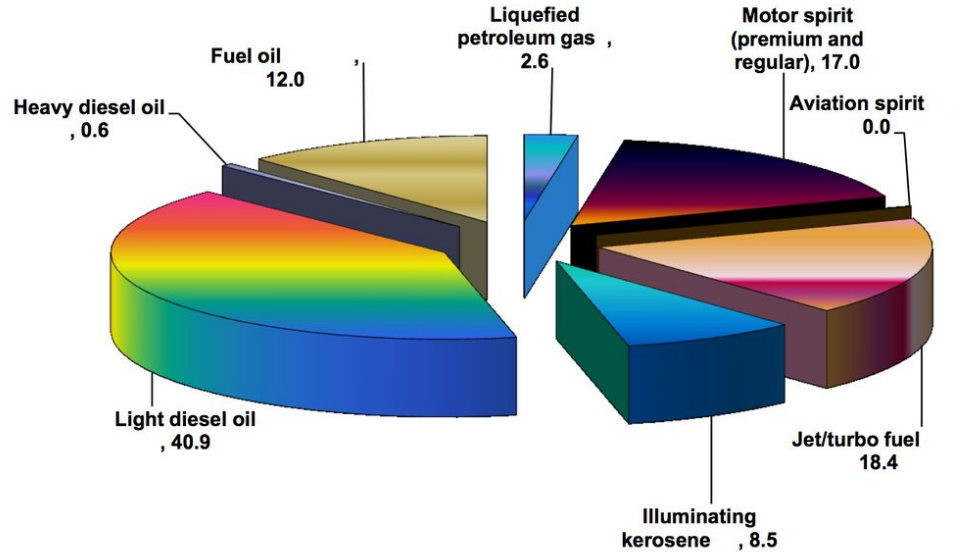


# What's Wrong with this Bar Chart?



# Color and Dimensionality

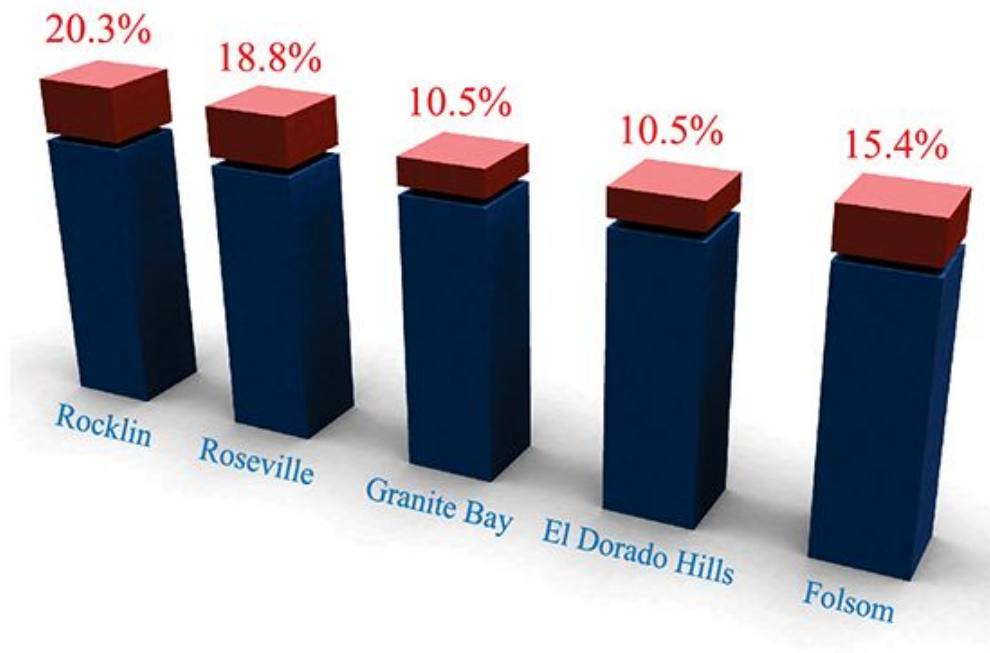
---



# Volume/Value Comparisons

---

Home values have gone up over the past year.

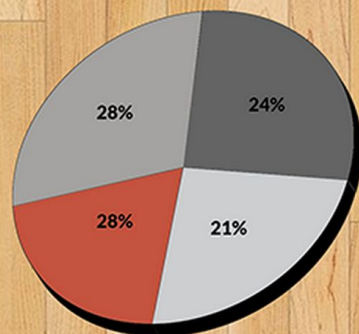


# Oval Pie Charts?

## Men's Basketball Scholarship Schools

- NCAA Division I = 341
- NCAA Division II = 290
- NAIA Division I-II = 262
- NJCAA Division I-II = 339

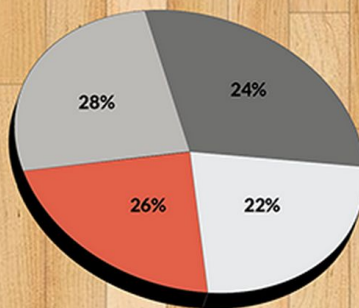
\* Numbers may fluctuate on a yearly basis



## Women's Basketball Scholarship Schools

- NCAA Division I = 338
- NCAA Division II = 291
- NAIA Division I-II = 260
- NJCAA Division I-II = 309

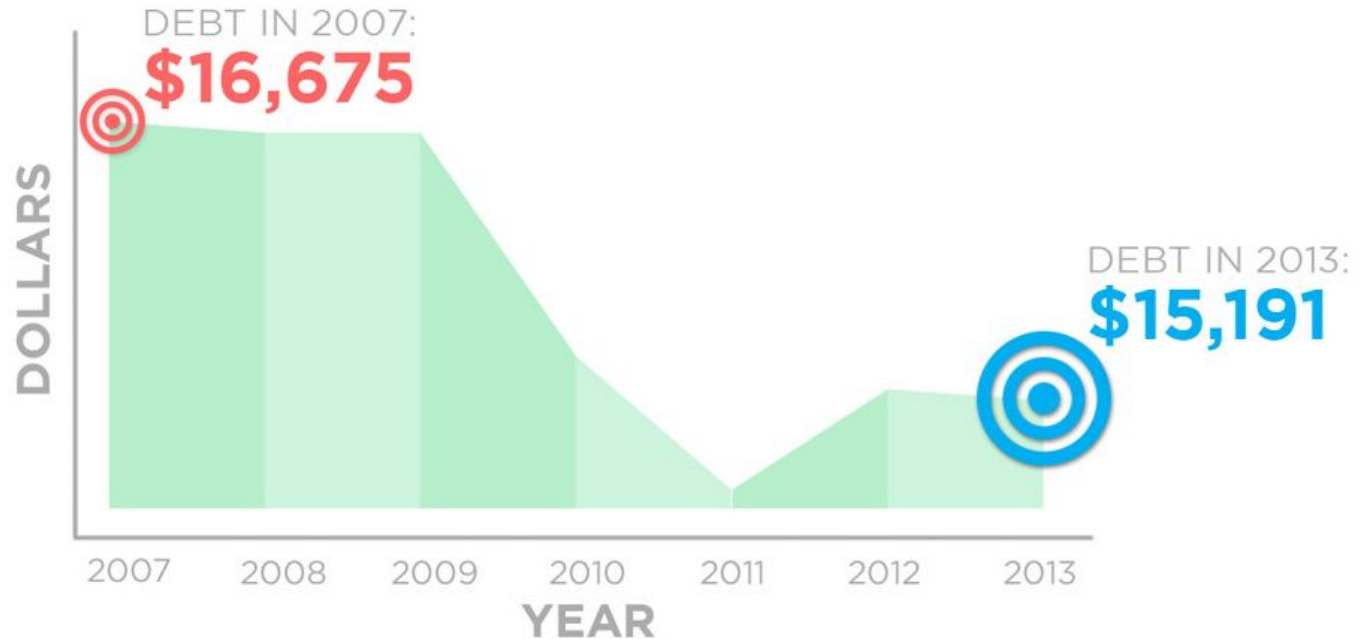
\* Numbers may fluctuate on a yearly basis



# Range, Caption, and Symbol Sins

---

THE AVERAGE INDEBTED HOUSEHOLD'S  
DEBT IS GOING **DOWN**



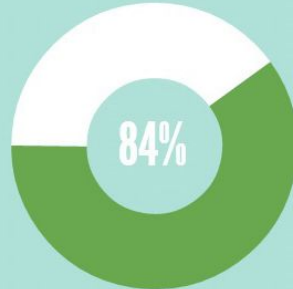


# The Virtues of Consistency

## Team Work



74% peminum kopi mampu bekerja sama dengan baik



84% peminum teh diklaim pandai menghidupkan suasana

## Other key tactics employed by B2B content marketers:



81%  
Website Articles



80%  
eNewsletters



76%  
Blogs



76%  
Live Events



73%  
Case Studies



73%  
Videos



64%  
White Papers



62%  
Webinars & Podcasts

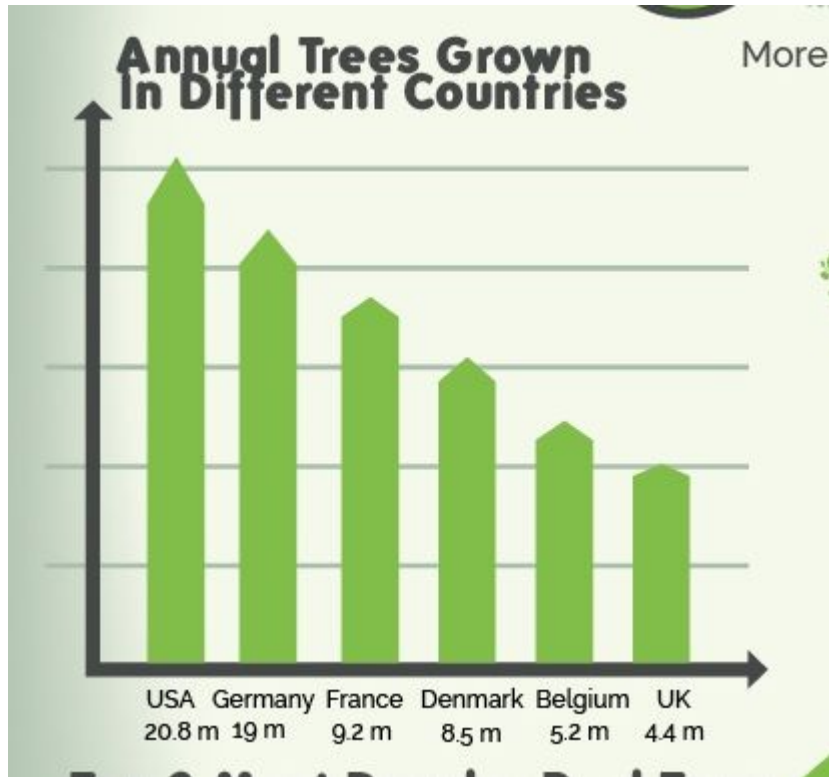


51%  
Infographics

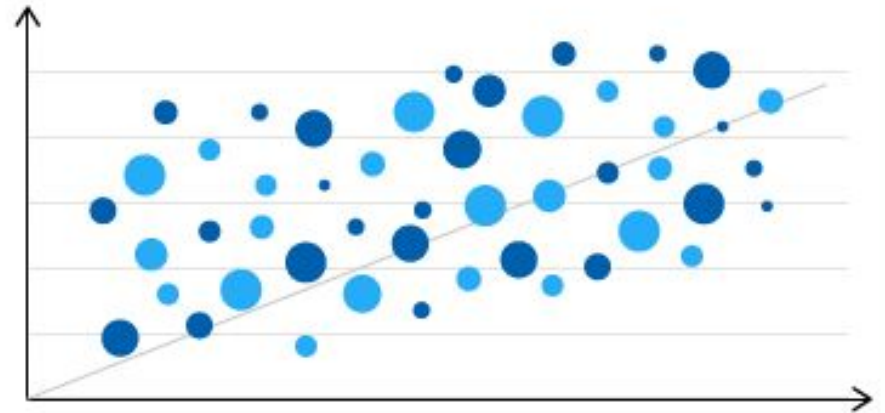


38%  
Mobile Content

# Provably Meaningless Charts

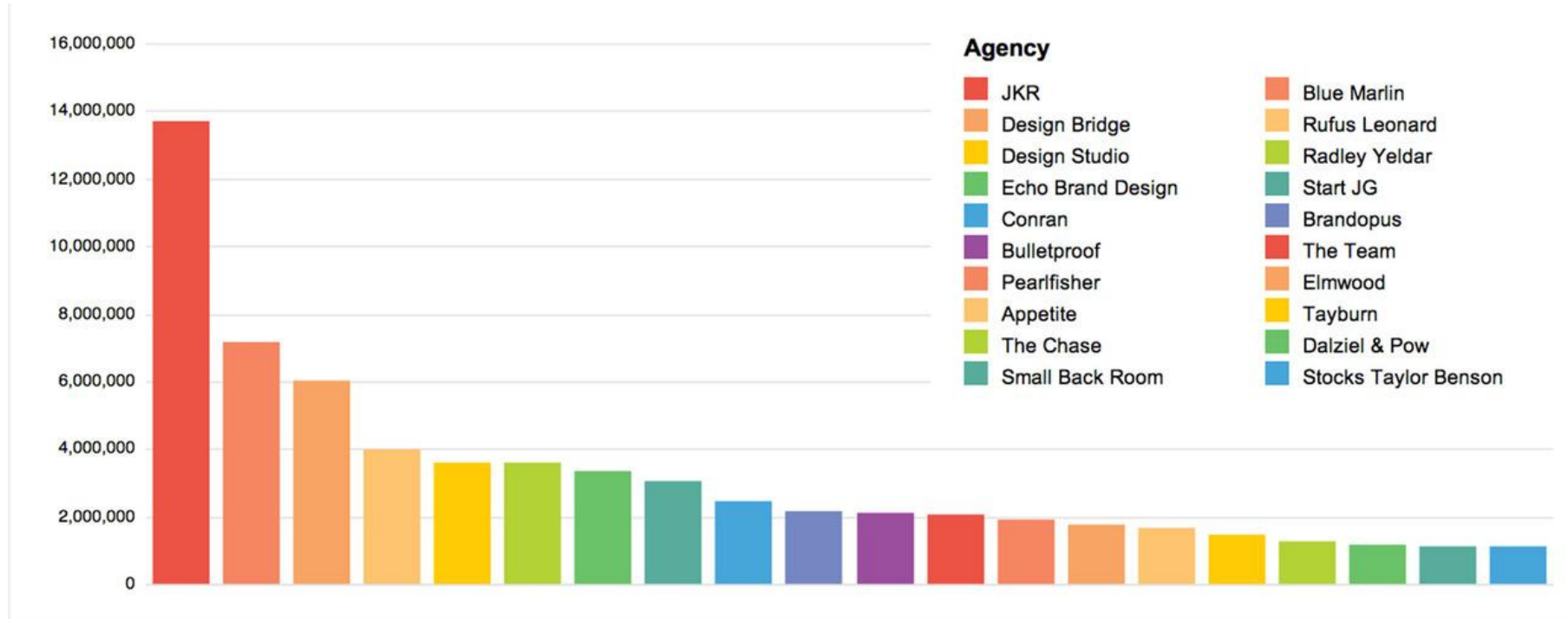


Googlebot listens to the 'little guy'.  
Tests show only a 1.4% correlation between the  
number of Google crawls and human visits.



# Dramatic Misuse/Reuse of Color

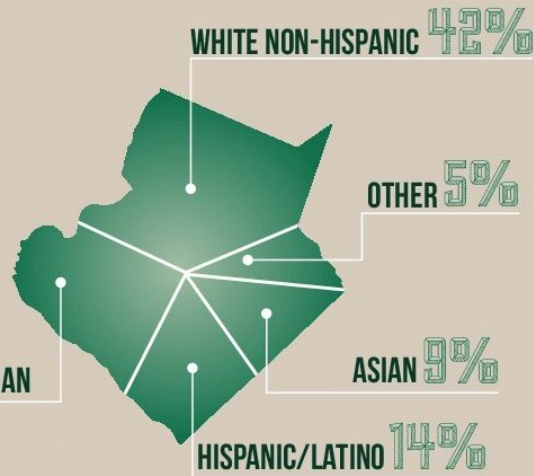
---



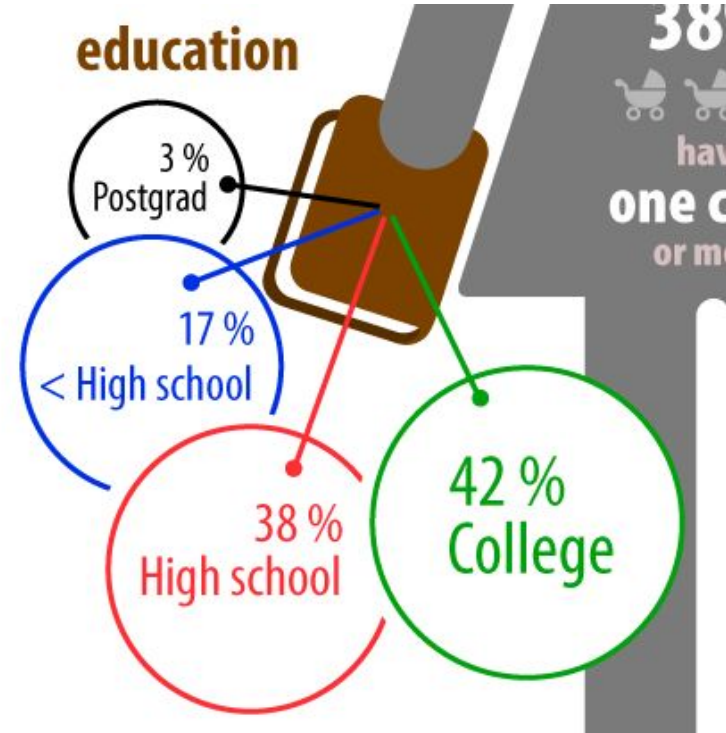
# Graphics Size Matters

GGC'S DIVERSE  
STUDENT BODY  
REFLECTS THE DIVERSITY  
OF GWINNETT COUNTY  
ATLANTA METRO REGION

29% BLACK AFRICAN AMERICAN



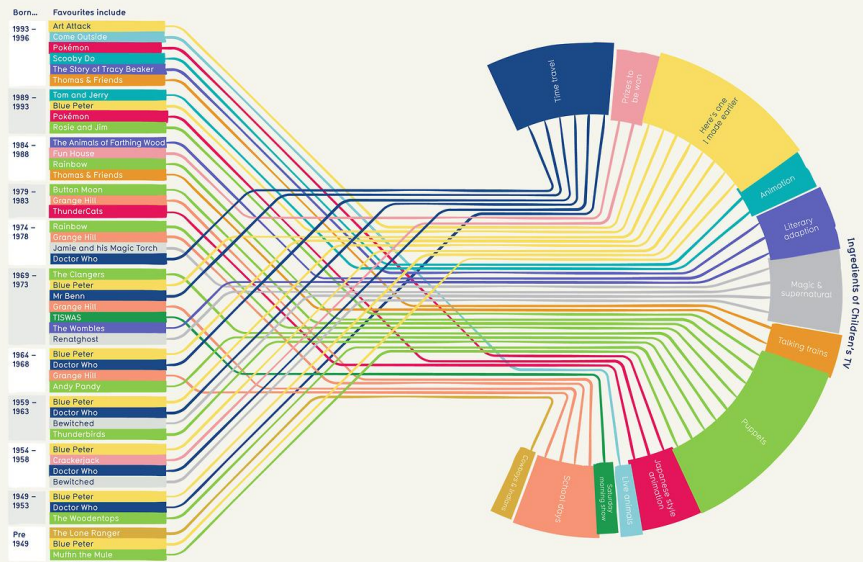
education



# Impressive Chart Junk

## Recipe for Children's TV

We asked adults across the UK about their favourite TV programme as a child. The magic of puppetry and the creativity of arts and crafts were favourite themes for every age group. BBC standard, **Blue Peter**, was popular across all age groups and was nominated by almost one in ten of all respondents as their favourite programme.



[twilicensing.co.uk/telescope2014](http://twilicensing.co.uk/telescope2014)

[@twilicensingnews](https://twitter.com/twilicensingnews)

[youtube.com/twilicensing](https://www.youtube.com/twilicensing)

[flickr.com/photos/twilicensing](https://www.flickr.com/photos/twilicensing)

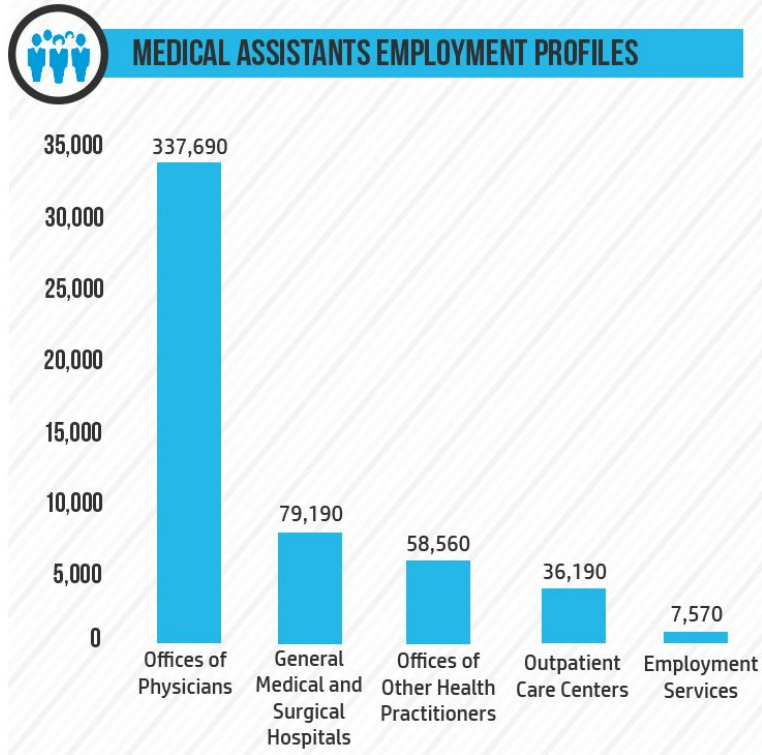
See bonus 'Trip Down Memory Lane' footage at [youtube.com/twilicensing](https://www.youtube.com/twilicensing)

## % Change in 5 Year Attendance of Top 25 Theme Parks





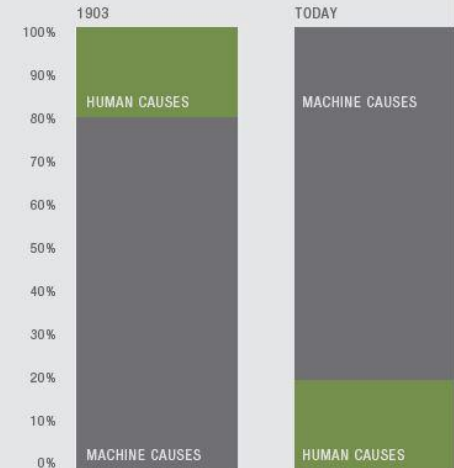
# Labels Matter



## CAUSES OF ACCIDENTS

Figure 1

*In the early days of flight, approximately 80 percent of accidents were caused by the machine and 20 percent were caused by human error. Today that statistic has reversed. Approximately 80 percent of airplane accidents are due to human error (pilots, air traffic controllers, mechanics, etc.) and 20 percent are due to machine (equipment) failures.*



# Way Too Many D

## BRIDGE OVER TROUBLED WATERS

In its third annual study of 160 countries, the Charity Aid Foundation measured giving behavior across three criteria – volunteering, helping strangers, and donating money. Although the results show that charitable giving – often a mirror of global economic patterns – has predictably declined since 2007, surprising findings exist among the top 20.

**THE MOST AFFLUENT COUNTRIES**  
AREN'T NECESSARILY THE MOST PHILANTHROPIC.  
Only 6 of the countries with the world's top 20 GDP made it to the Charity Aid Foundation's top 20 list.



**GLOBALLY**, average participation in giving has fallen since 2007.



PARTICIPATION IN HELPING STRANGERS  
DONATING MONEY TO CHARITY  
VOLUNTEERING TIME

Source: Charities Aid Foundation World Giving Index 2012

**AUSTRALIA** IS THE MOST GENEROUS NATION IN THE WORLD.

The government's proactive effort to encourage philanthropy may be the reason the country has landed the highest score on average over the past five years.



IN EVERY COUNTRY OF THE TOP 10, volunteering time was the least common giving behavior

**THE TOP 9 COUNTRIES**

HAVE WORLD GIVING INDEX SCORES OF 50% OR MORE. This means that at least half the population, on average, is taking part in at least 1 of the 3 giving behaviors on a monthly basis.



KEY

- MONEY DONATIONS
- VOLUNTEERING
- HELPING STRANGERS
- TOP 10
- TOP 9
- ASIAN COUNTRIES
- AFFLUENT COUNTRIES

**THERE IS DIVERSITY IN GENEROSITY**

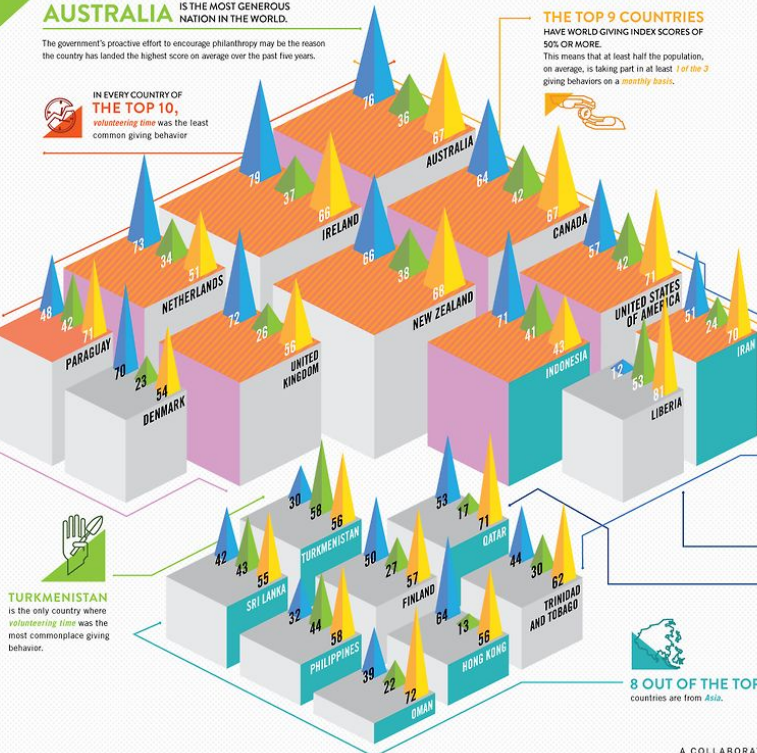
**GEOGRAPHIC LOCATION:**  
Every continent is represented.

**POPULATION:**  
300 MILLION PEOPLE  
United States  
1.3 MILLION PEOPLE  
Trinidad and Tobago

**ECONOMIC STRENGTH:**  
2ND LOWEST GDP PER CAPITA  
Liberia  
2ND HIGHEST GDP PER CAPITA  
Qatar



**8 OUT OF THE TOP 20**  
countries are from Asia.

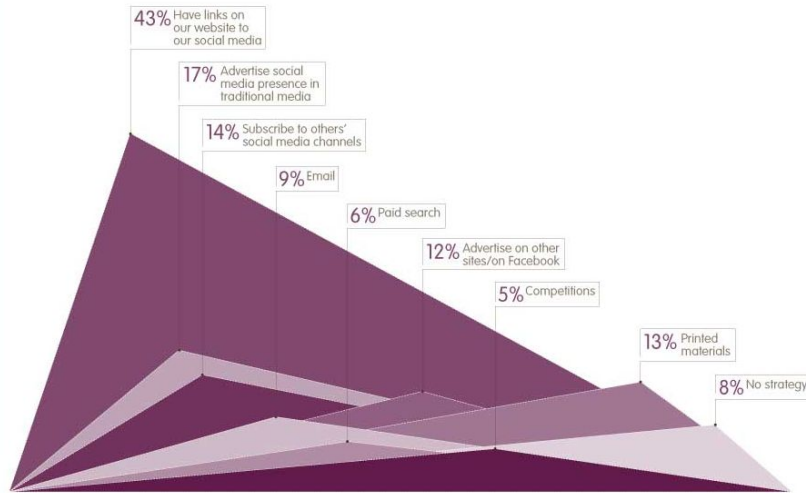


**TURKMENISTAN** is the only country where volunteering time was the most commonplace giving behavior.

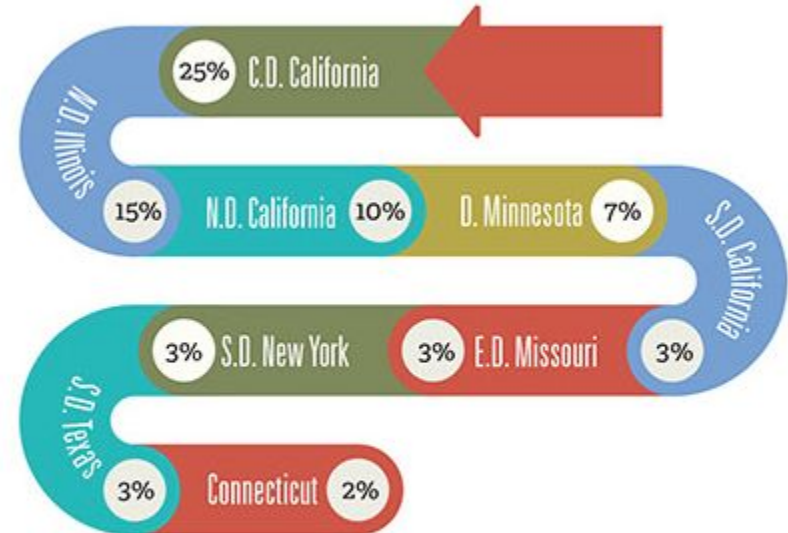
A COLLABORATION BETWEEN GOOD AND COLUMN FIVE

# 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.

---