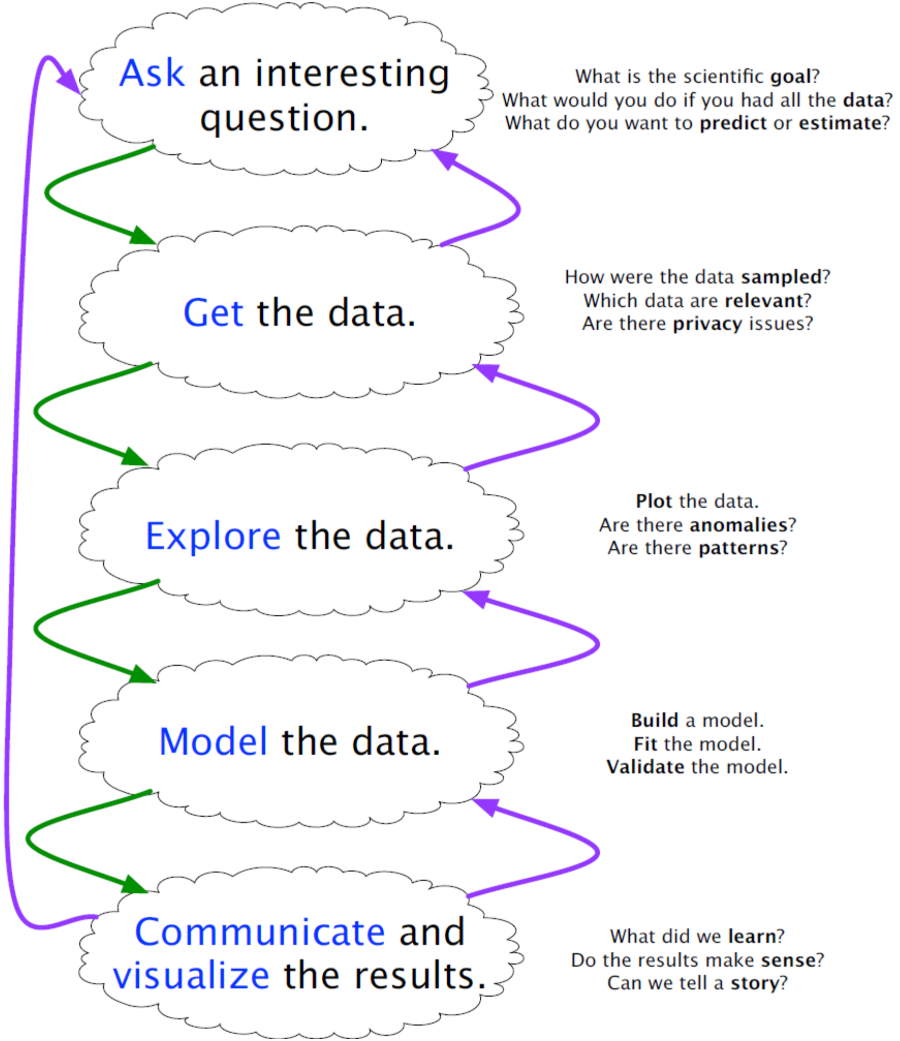
Data Science

Lecture 2: Statistical Analysis / Visualizing Data

Statistical Analysis

Typical Data Science Pipeline



Statistics and Data Science

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

- Josh Blumenstock (Univ. of Washington)

Statistical Data Distributions

Every observed random variable has a particular frequency/probability distribution.

Some distributions occur often in practice/theory:

- The Binomial Distribution
- The Normal Distribution
- The Power Law Distribution

Binomial Distributions

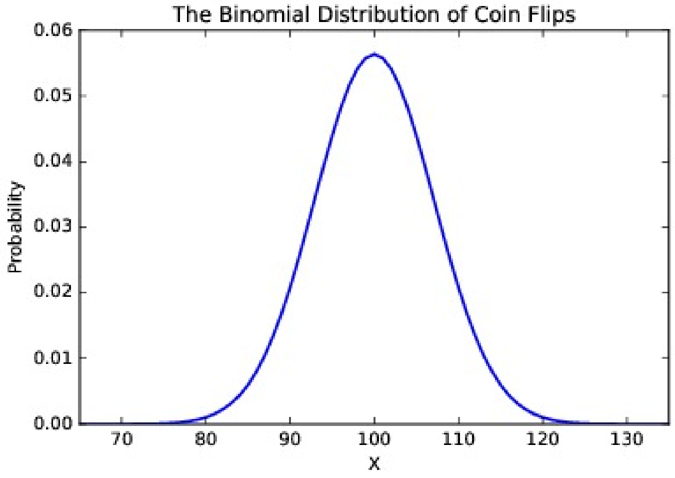
Experiments consist of *n identical, independent* trials which have two possible outcomes, with probabilities *p* and *(1-p)* like heads or tails.

$$P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x}$$

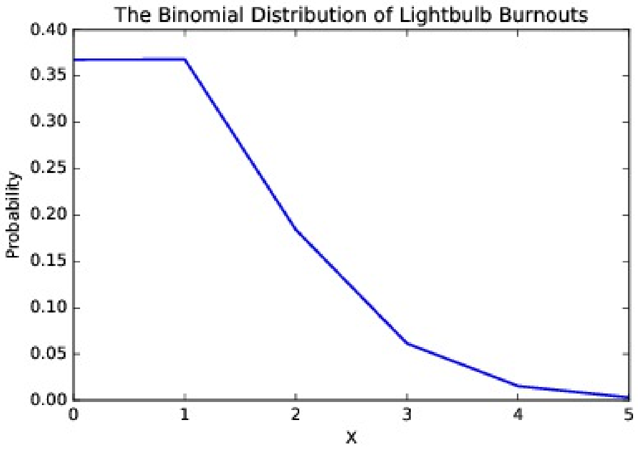
Properties of Binomial Distributions

Discrete, but bell (or half-bell) shaped

Coin flips: p=0.5 n=200



Lightbulb burnouts: p=0.001 n=1000



The distribution is a function of n and p.

The Normal Distribution

The bell-shaped distribution of height, IQ, etc.

Completely parameterized by mean and standard deviation:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma^2}$$

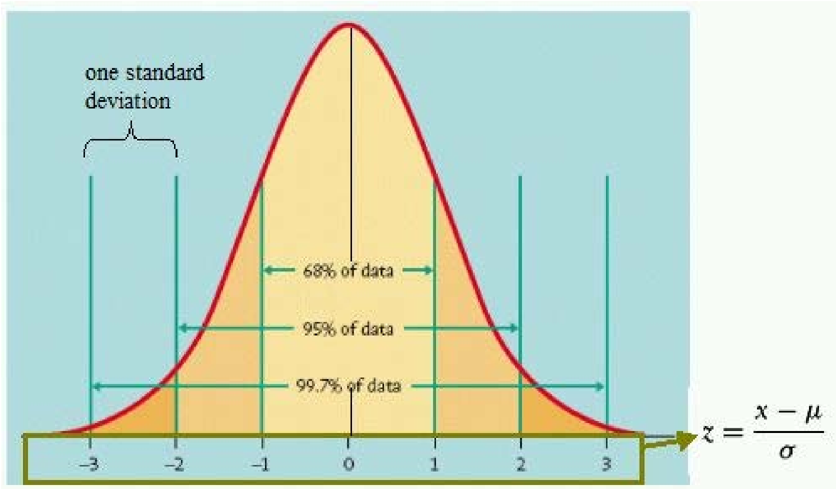
Not all bell-shaped distributions are normal but it is generally a reasonable start.

Interpreting the Normal Distribution

Tight bounds on probability follow for Z-scores from normally distributed random variables:

IQ is normally distributed, with mean 100 and standard deviation 15.

Thus about 2.5% of people have IQs above 130.



Power Law Distributions

Power laws are defined $P(X = x) = cx^{-\alpha}$ for exponent *α* and normalization constant *c*.

They do not cluster around a mean like a normal distribution, instead having very large values rarely but consistently.

They define 80-20 rules: 20% of the *X* get 80% of the *Y*.

City Population Yield Power Laws

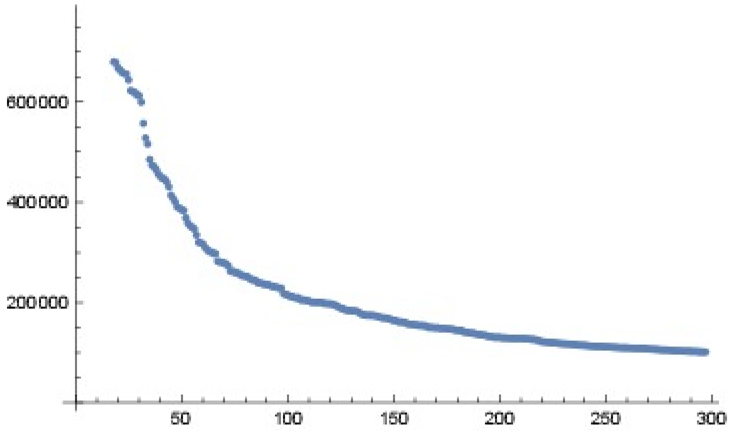
The average big US city has population 165,719. Even with a huge standard deviation of 410,730, New York city with 8,008,278 people is too many sigma away from the mean.

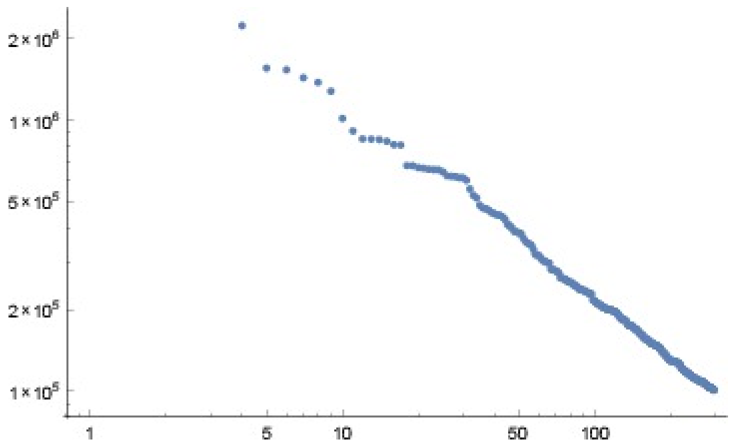
Power laws arise when the rich get richer.

Linear and Log-Log Plots for City Pop

Straight lines on log-log plots say power law.

The biggest values are out of scale on linear plots.

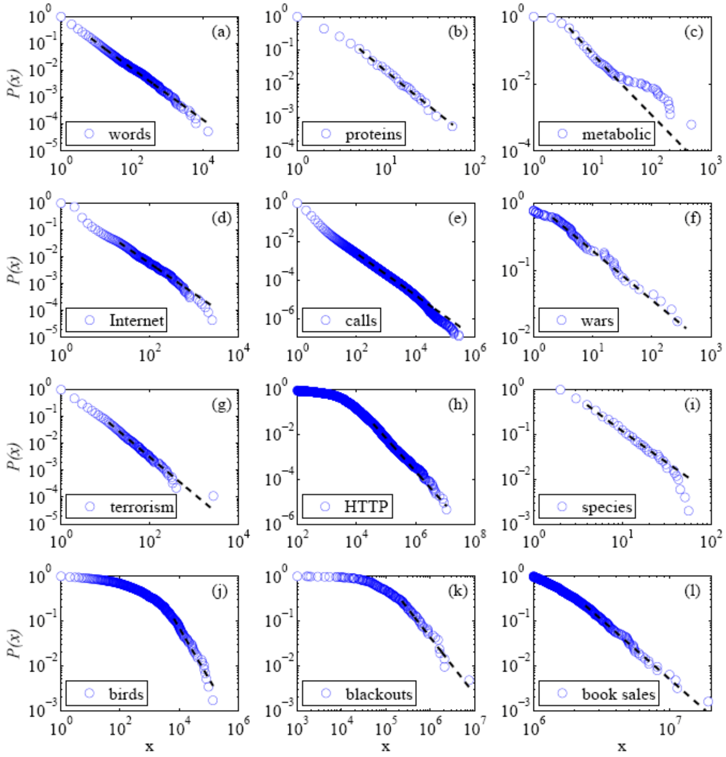




Many Distributions are Power Laws

- Internet sites with x inlinks.
- Frequency of earthquakes at x on the Richter scale
- Words used with a relative frequency of x
- Wars which kill x people

Power laws show as straight lines on log value, log frequency plots.



When is an Observation Meaningful?

Computational analysis readily finds patterns and correlations in large data sets.

But when is a pattern significant?

Sufficiently strong correlations on large data sets may seem ``obviously’’ significant, but the issues are often quite subtle.

Comparing Population Means

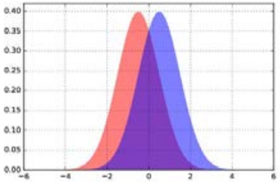
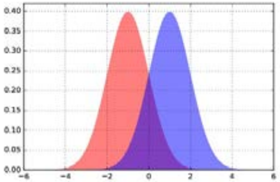
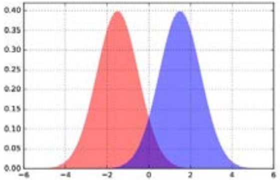
The T-test evaluates whether the population means of two samples are different.

Sample the IQs of 20 men and 20 women. Is one group smarter on average?

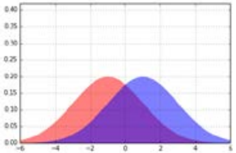
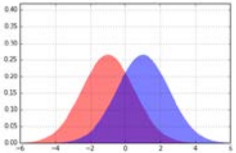
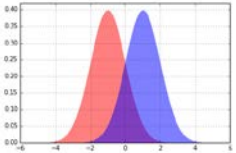
Certainly the sample means will differ, but is this difference significant?

Differences in Distributions

It becomes easier to distinguish two distributions as the means move apart...

... or the variance decreases:

The T-Test

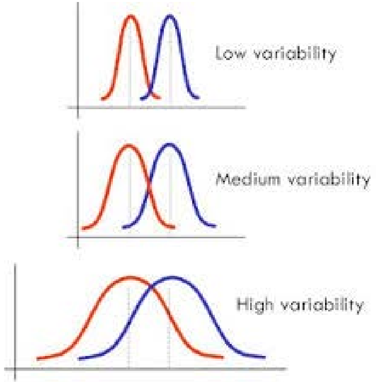
Two means differ significantly if:

- The mean difference is relatively large
- The standard deviations are small enough
- The samples are large enough

Welch's t-statistic is: $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where s^2 is the sample variance.

Significance is looked up in a table.



The Kolmogorov-Smirnov Test

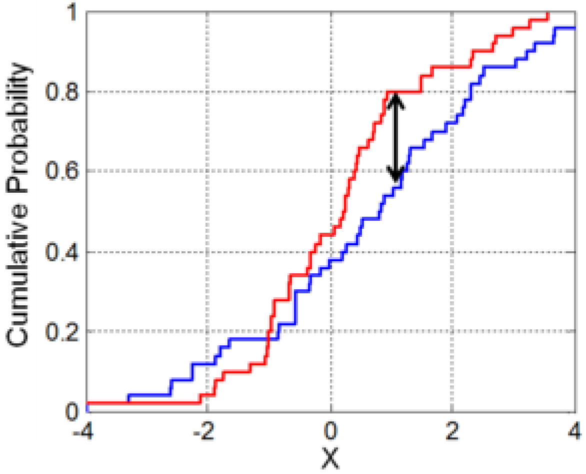
The KS-test quantifies the difference between two probability distributions by the maximum y-distance gap between the two cdfs.

The max distance between two cdfs is

$$D(C_1, C_2) = \max_{-\infty \leq x \leq \infty} |C_1(x) - C_2(x)|$$

two distributions differ at the significance level of α when:

$$D(C_1, C_2) > c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$



Permutation Tests and P-values

Traditional statistical tests evaluate whether two samples came from the same distribution.

Many have subtleties (e.g. one- vs. two-sided tests, distributional assumptions, etc.)

Permutation tests allow a more general, more computationally idiot-proof way to establish significance.

Permutation Tests and P-values

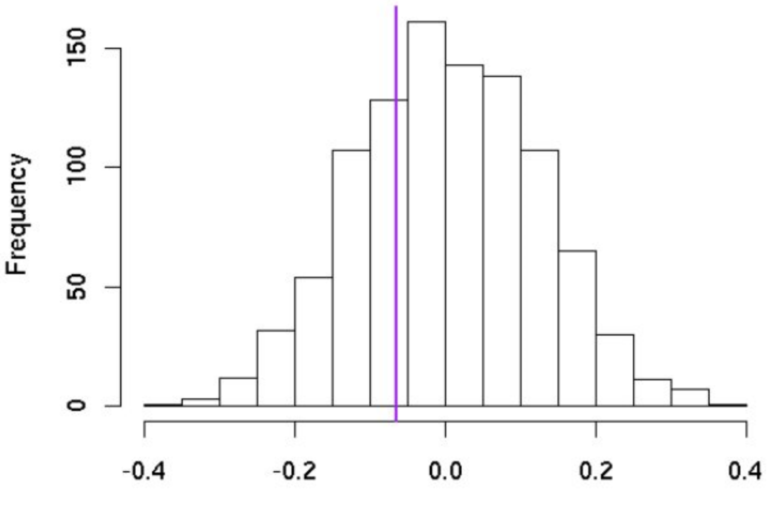
If your hypothesis is supported by the data, then randomly shuffled data sets should be less likely to support it.

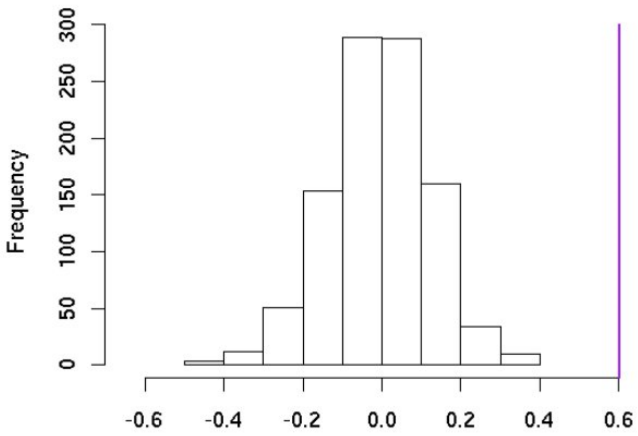
The ranking of the real test statistic among the shuffled test statistics gives a p-value.

You need statistic on your model you believe is interesting, e.g. correlation, std. error, etc.

Significance of a Permutation Test

The rank of the real data among the random permutations determines significance:





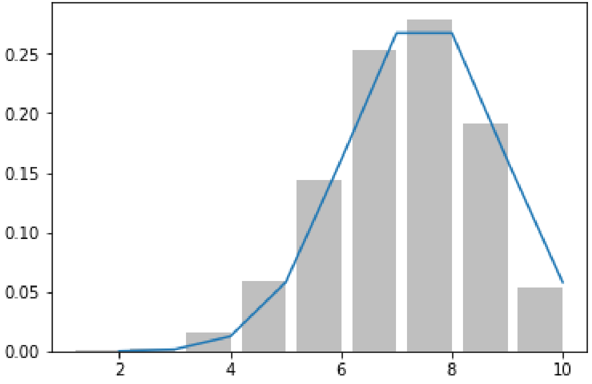
The Central Limit Theorem

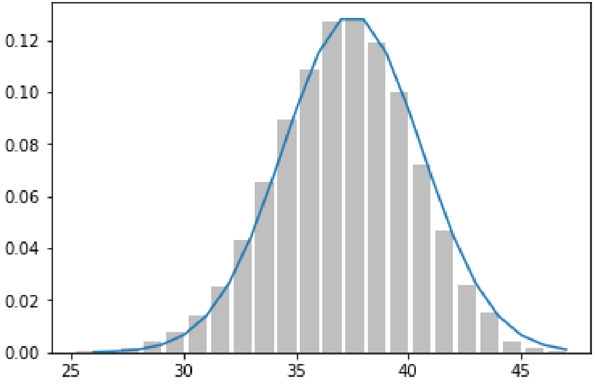
- A random variable defined as the average of a large #of independent and identically distributed(i.i.d.) random variables is itself approximately normally distributed.
- If $x_1, \ldots, x_n$ are r.v. with μ and σ^2, and if n is large:

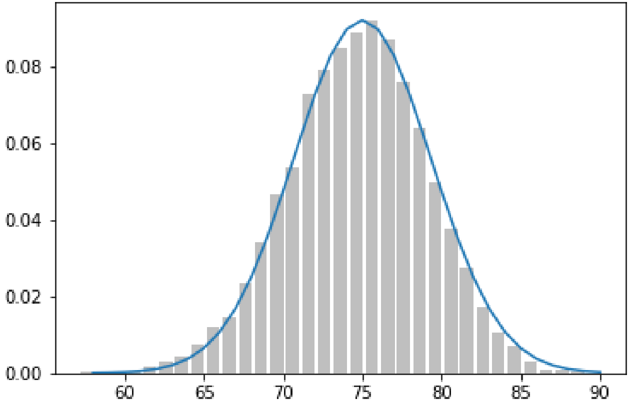
  $Z = 1/n\ (x_1 + \ldots + x_n)$ is approx. normally distributed

Significance of central limit theorem

- If n gets large, Binomial(n, p) ~ Normal(np, np(1-p))







Significance of central limit theorem

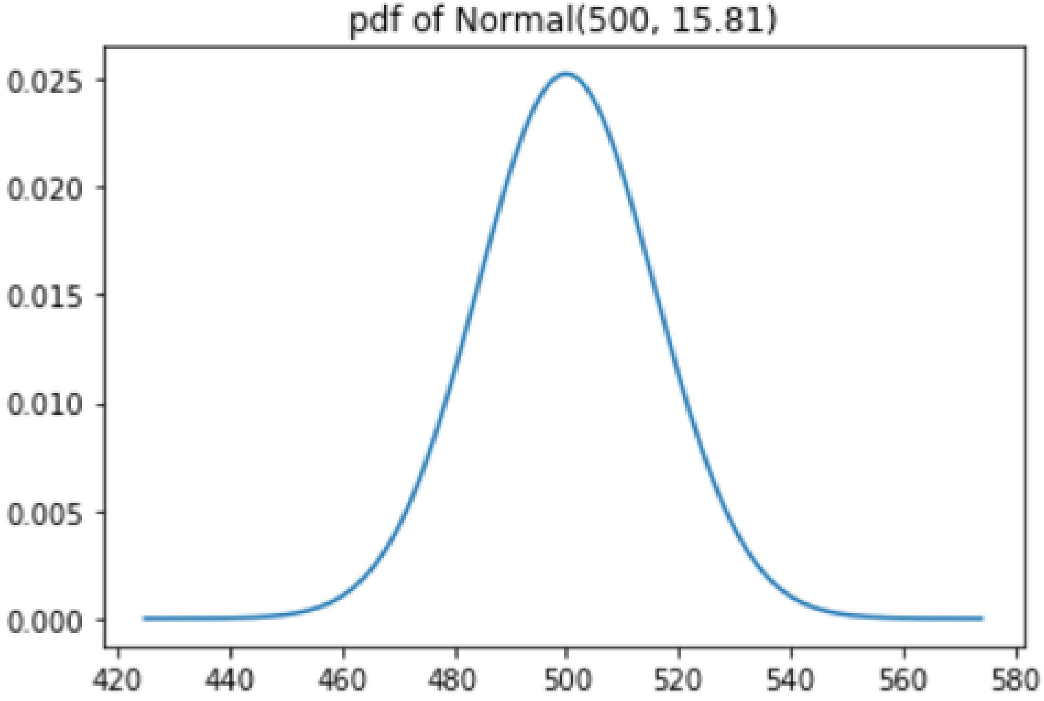
- It implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

Statistical Hypothesis Testing

- Example: Flipping a Coin – when speculating the coin is not fair.
  - **null hypothesis (H_0):** coin is fair, i.e., p = 0.5
  - **alternative hypothesis (H_1):** coin is not fair.

- We use statistics to decide whether we can reject H_0 as false or not.
  - In particular, flipping the coin *n* times and counting the #of heads *X*.
  - Each coin flip is a **Bernoulli** trial, meaning X is a **Binomial**(n,p).
  - Due to CLT, X can be approximated by **Normal**(np, np(1-p)).
  - Choose *significance* level– how willing to make a *type I error* (FP)
  - *Typical choices: 5% or 1%*

```
normal_two_sided_bounds(0.95, 500, 15.81) (469.01026640487555, 530.9897335951244)
normal_two_sided_bounds(0.99, 500, 15.81) (459.27260472187146, 540.7273952781286)
```



Types of errors

Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Reject	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)

Type I error is detecting an effect that is not present, while a type II error is failing to detect an effect that is present.

Statistical Hypothesis Testing w/ p-value

- P-value: probability (assuming H_0 is true) of seeing a value at least as extreme as the one we actually observed.
- Typical choices of significance level: 0.05 or 0.01
- If X = 530, p-value = 0.062, if X = 532, p-value = 0.0463

Example: Running an A/B Test

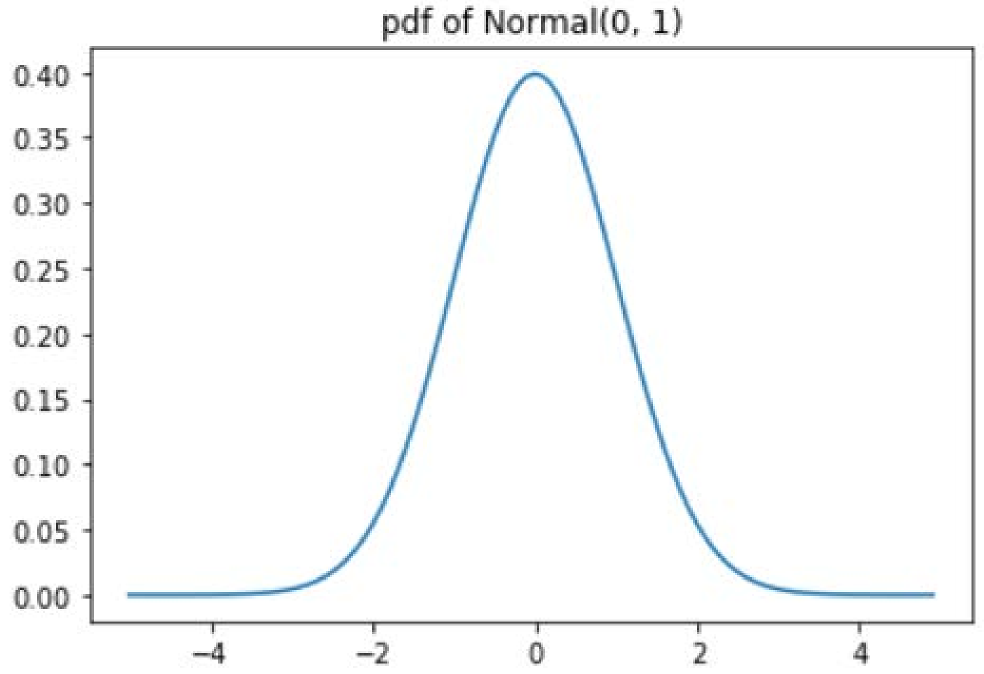
- You are trying to get people to click on advertisements.
- VP of ads. wants your help choosing between two ads A and B.
- You run an experiment by randomly showing site visitors one of the two ads and tracking how many people click on each ad.
- Let N_A=#of people seen A, n_A=#of people click on A, p_A=probability of someone clicking A.
- Then, n_A/N_A is approx. *Normal*$(p_A, p_A(1-p_A)/N_A)$.
- Similarly, n_B/N_B is approx. *Normal*$(p_B, p_B(1-p_B)/N_B)$.

Example: Running an A/B Test (cont'd)

- Then, n_A/N_A is approx. *Normal*(p_A, $p_A(1-p_A)/N_A$).
- Similarly, n_B/N_B is approx. *Normal*(p_B, $p_B(1-p_B)/N_B$).

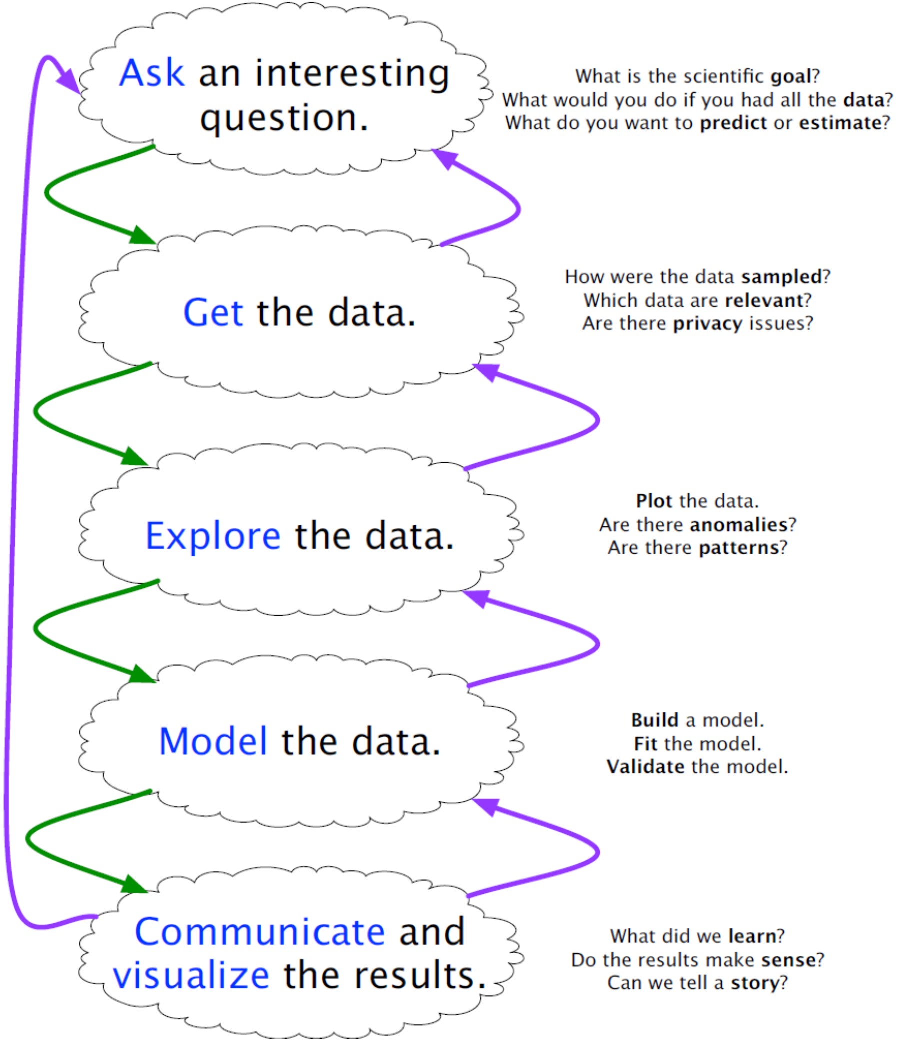
- The two normals are independent, thus their difference should also be normal
- Perform hypothesis test w/ H_0 – p_A and p_B are the same
- Suppose you have 1000, 200, 1000, 150 for N_A, n_A, N_B, n_B, respectively. And you set significance level for p-value as 0.05.
- Can you reject the null hypothesis?

```
a_b_test_statistic(1000, 200, 1000, 180) -1.1403464899
a_b_test_statistic(1000, 200, 1000, 150) -2.9488391231
```



Visualizing Data

Typical Data Science Pipeline



Exploratory Data Analysis

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey

Exploratory Data Analysis

Looking carefully at your data is important:

- to identify mistakes in collection/processing
- to find violations of statistical assumptions
- to observe patterns in the data
- to make hypothesis.

Feeding unvisualized data to a machine learning algorithm is asking for trouble.

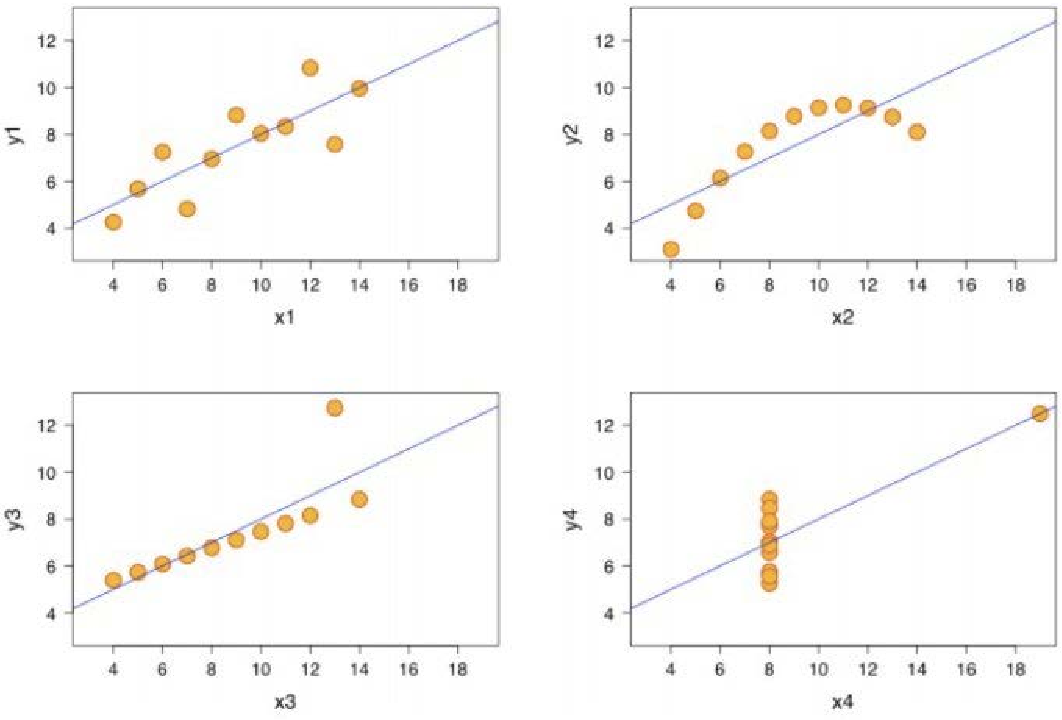
Anscombe's Quartet

All four data sets have exactly the same mean, variance, correlation, and regression line:

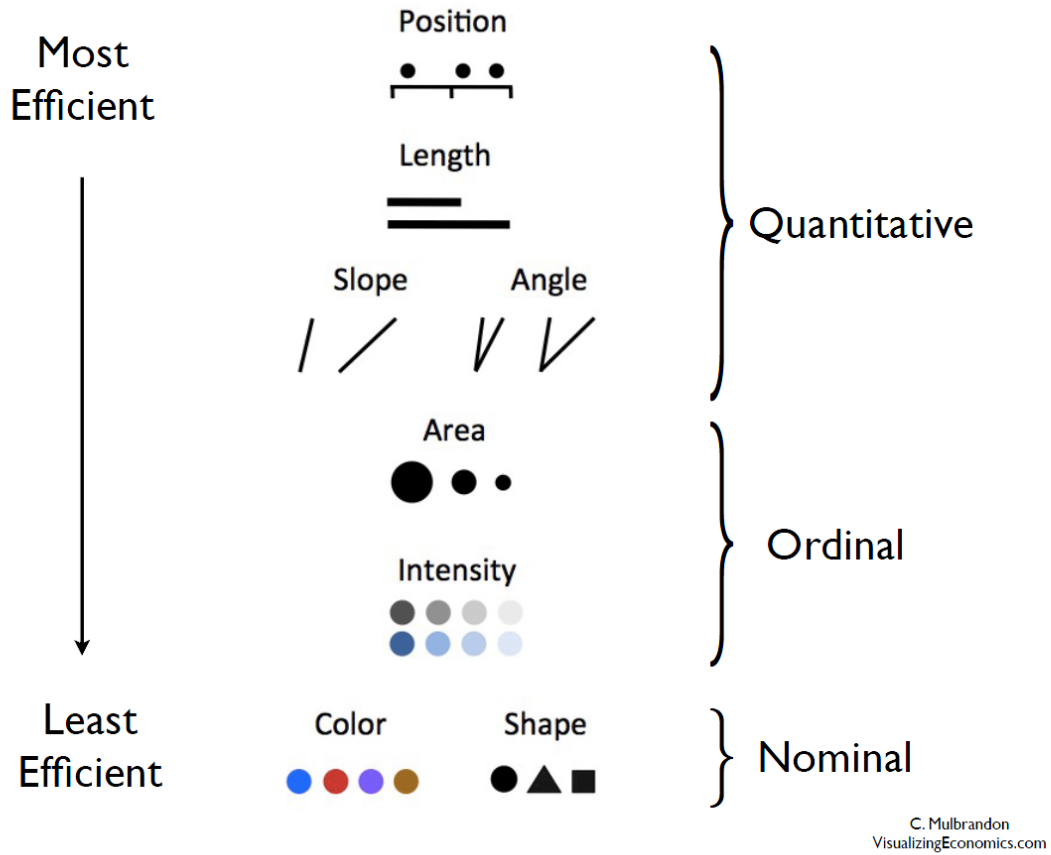
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
var.	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
corr.	0.816		0.816		0.816		0.816	

Plotting Anscombe's Quartet

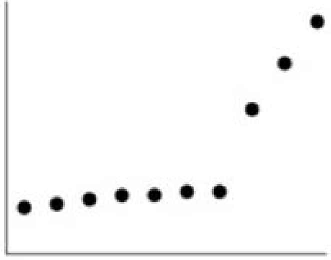
All four data sets have exactly the same mean, variance, correlation, and regression line:

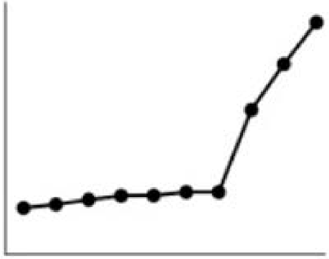


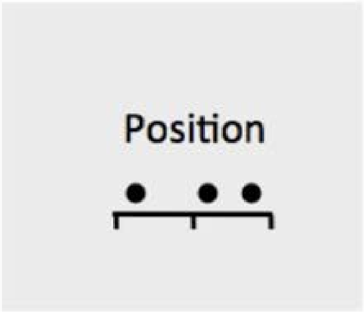
Mapping Data to Image

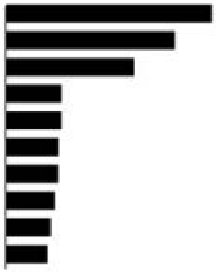


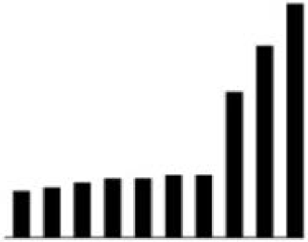
Most Effective

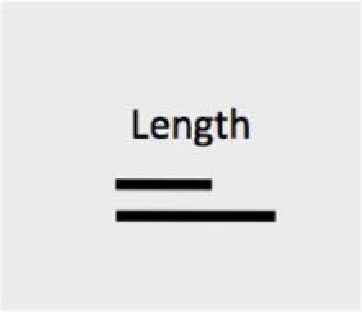




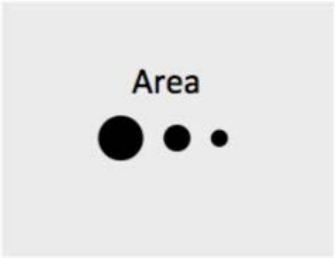




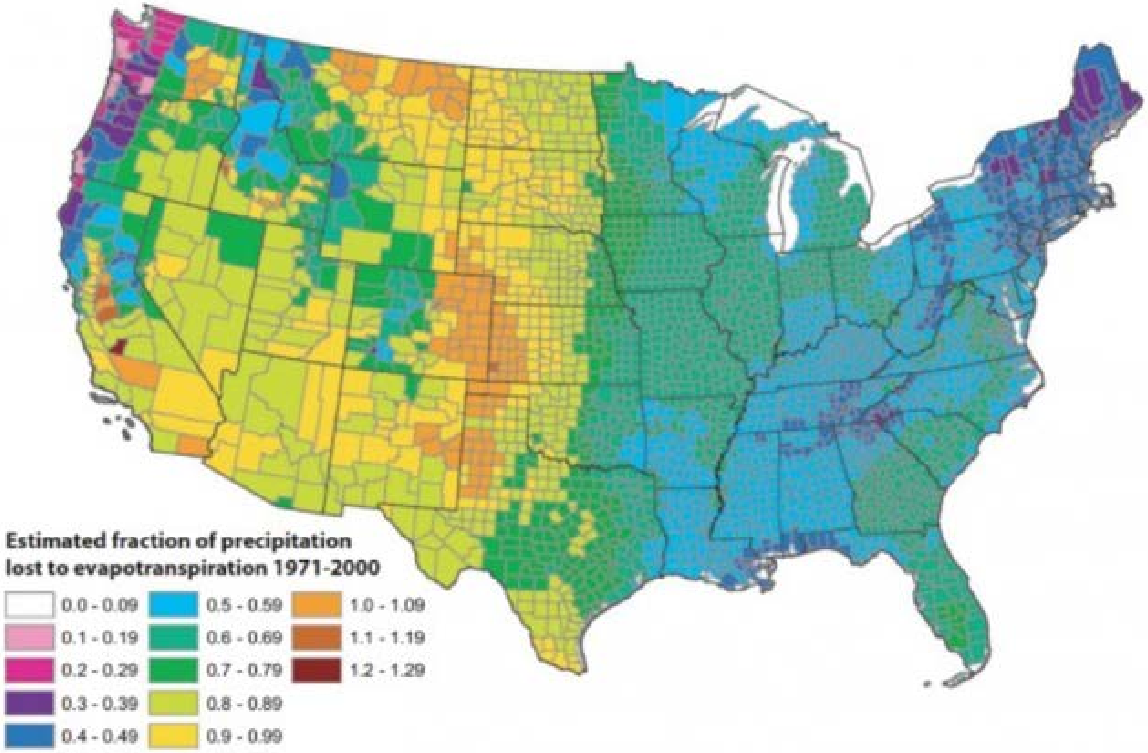




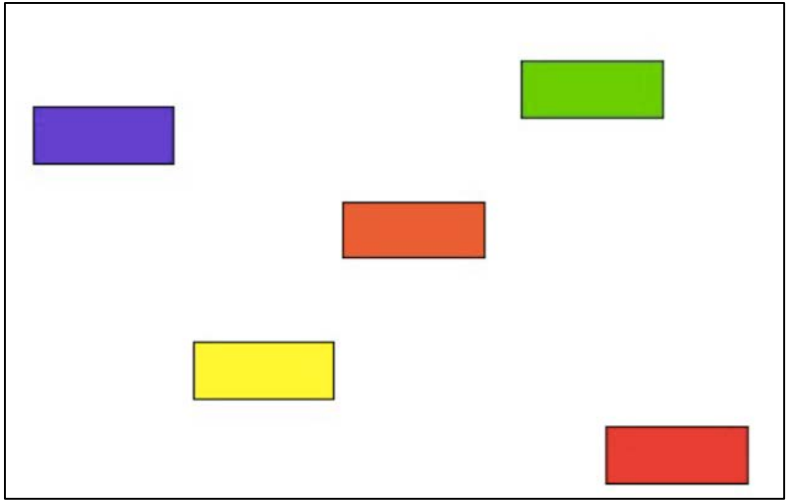
Less Effective

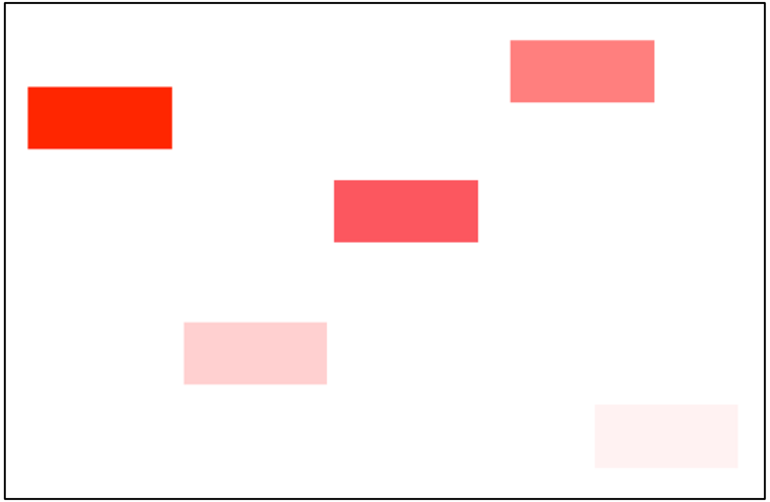


Least Effective



Order These Values





Perceived as Ordered

Examples: Not Effective

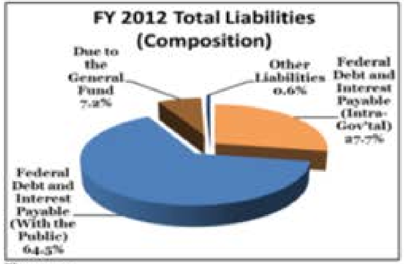
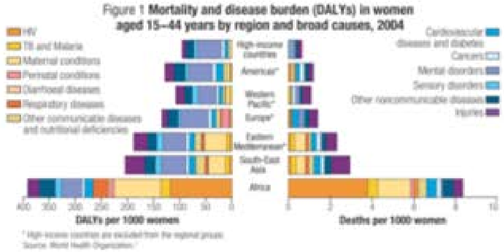
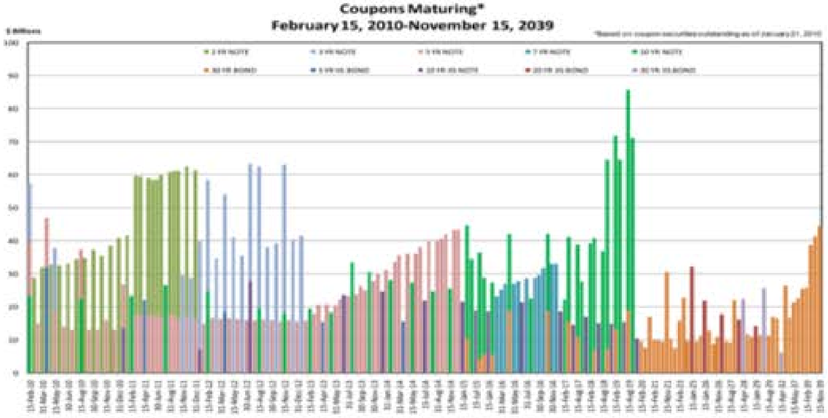


Figure 10





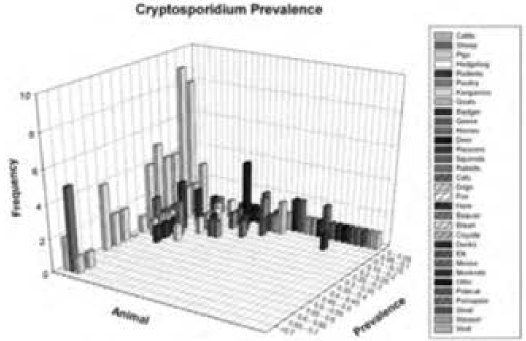
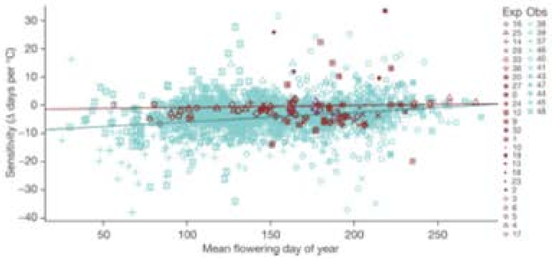
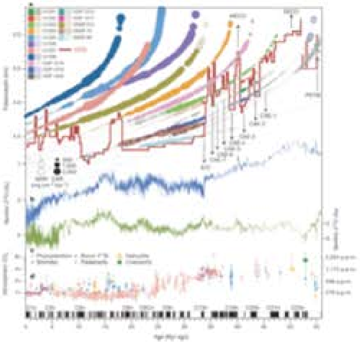


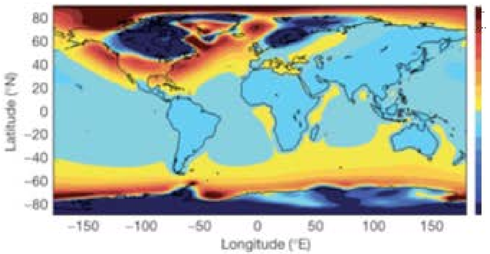
Figure 5.2 Mean prevalence rates of *Cryptosporidium* oocysts by animal species.

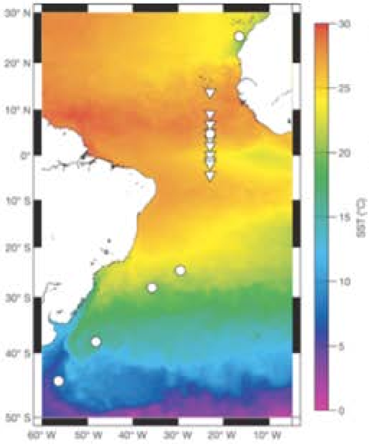
Sources: US Treasury and WHO reports

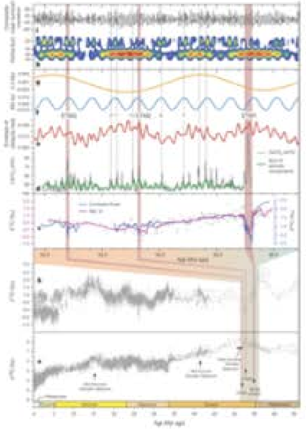
Examples: Not Effective

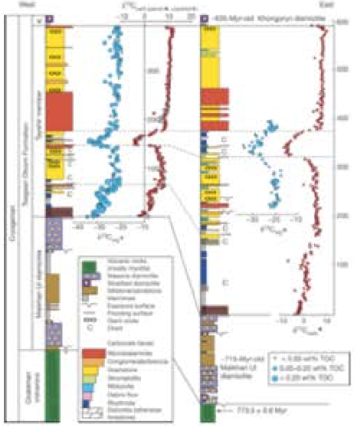


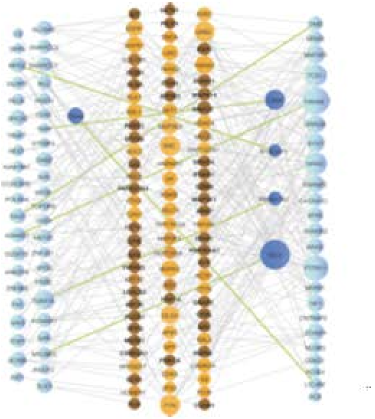






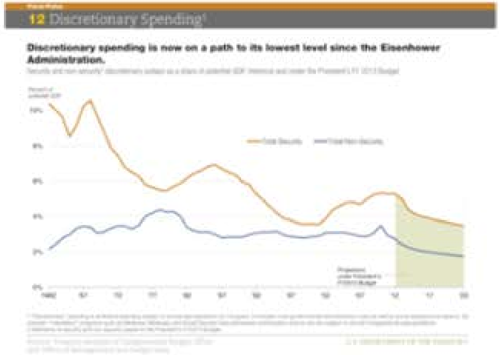


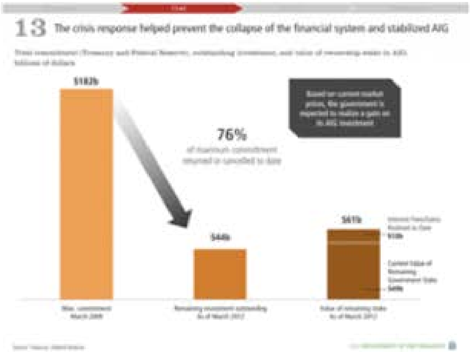


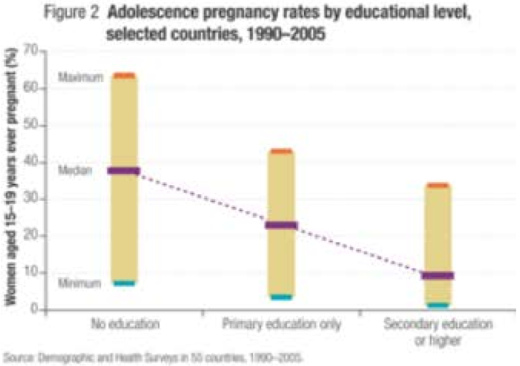


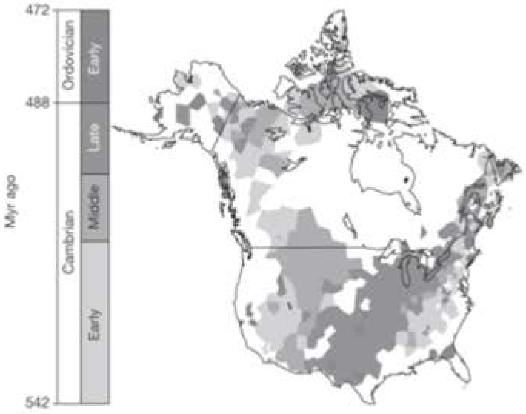
Source: Nature

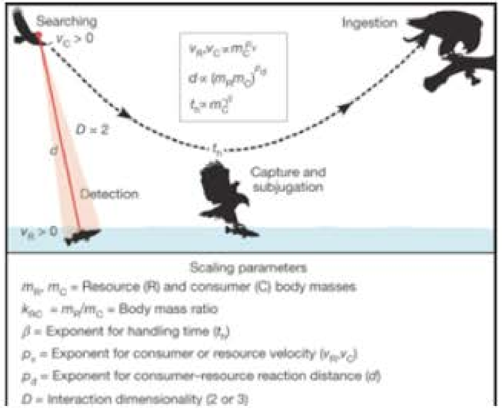
Examples: Much Better

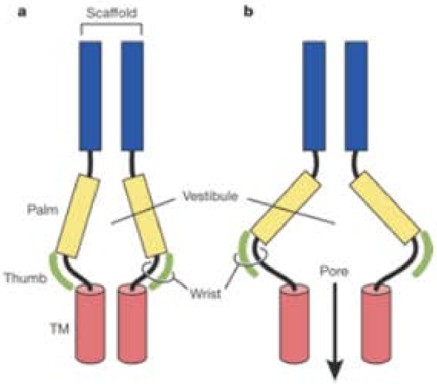












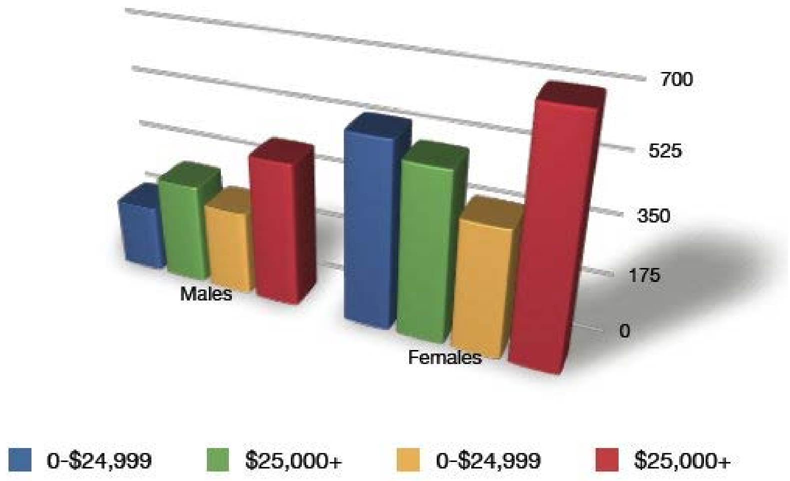
Tufte's Design Principle

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

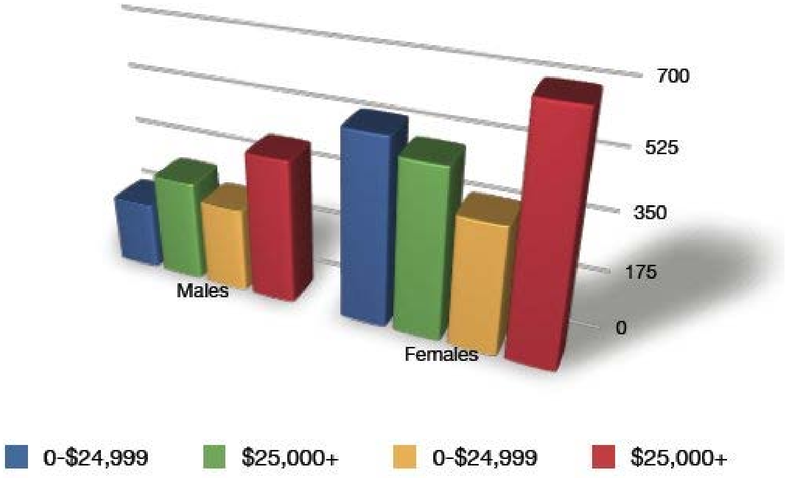
Maximize Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$



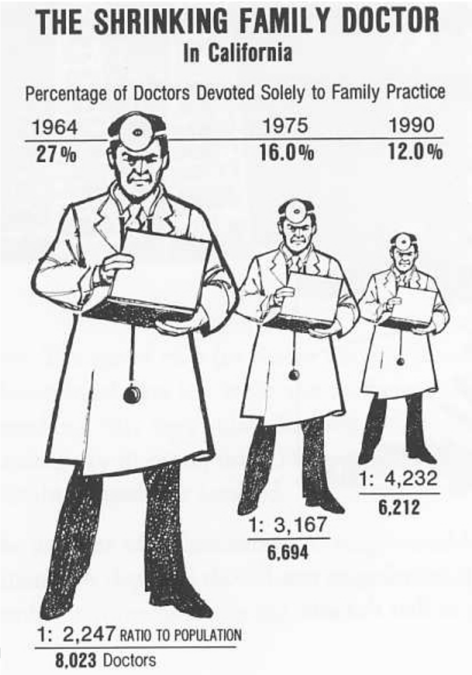
Maximize Data-Ink Ratio

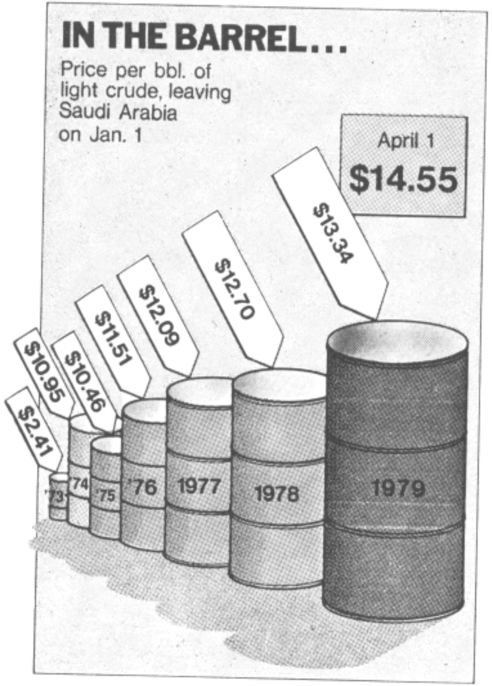
$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

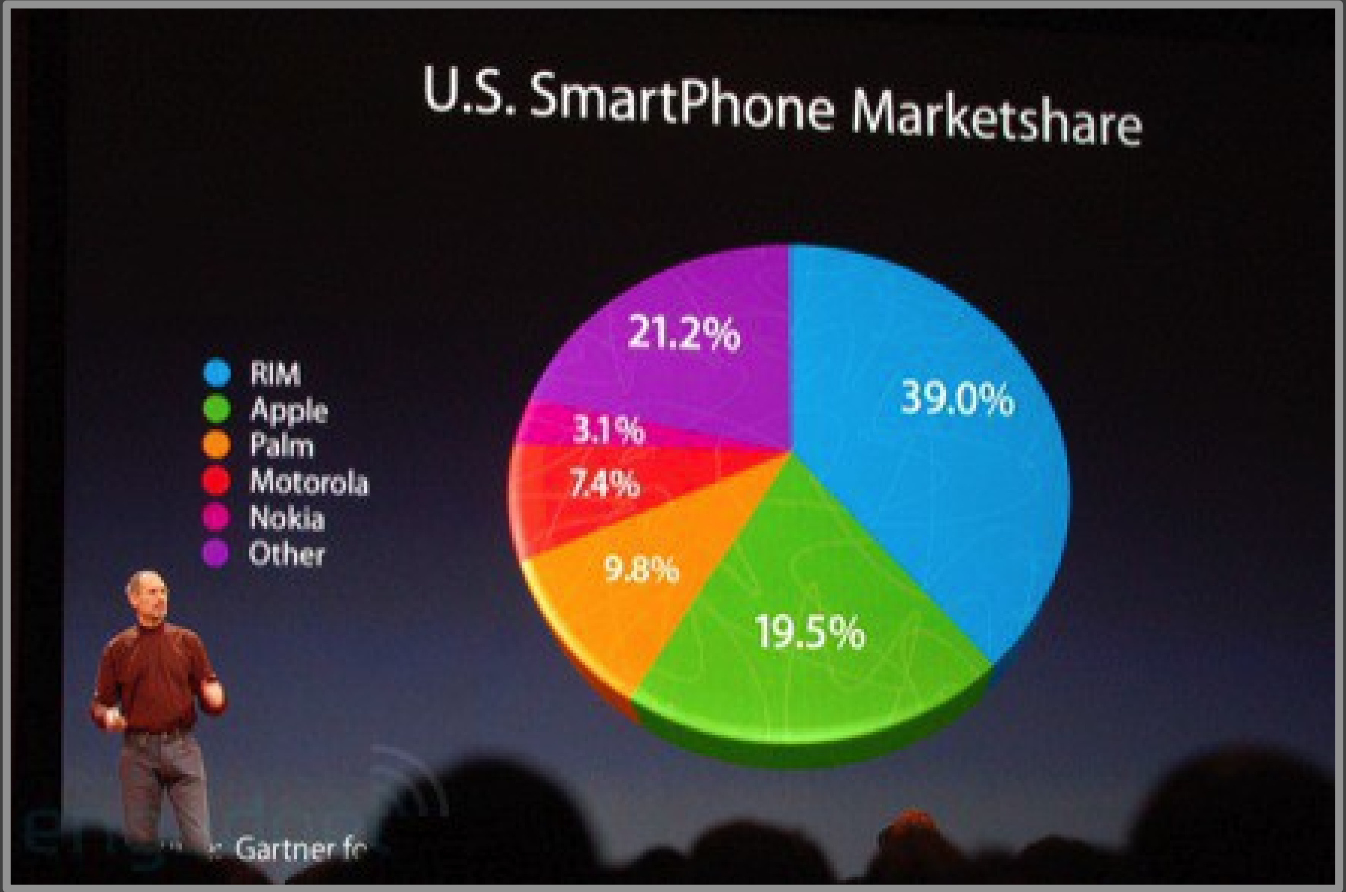


The Lie Factor

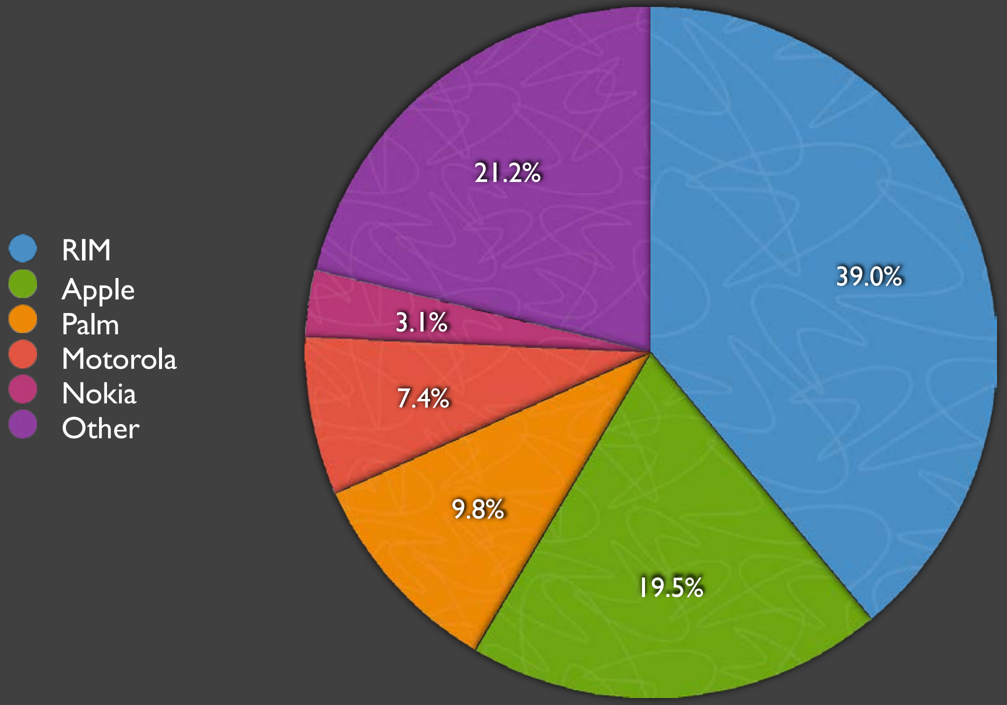
$$\frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$



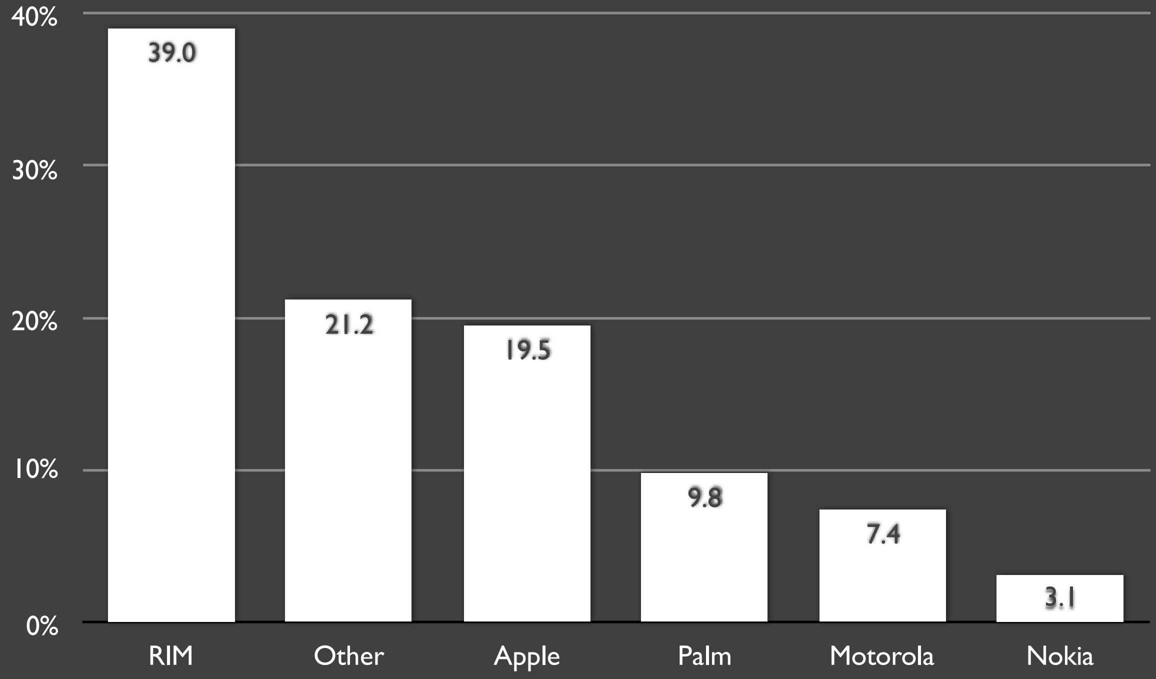



U.S. SmartPhone Marketshare
RIM
Apple
Palm
Motorola
Nokia
Other
21.2%
39.0%
3.1%
7.4%
9.8%
19.5%

U.S. SmartPhone Marketshare



U.S. SmartPhone Marketshare



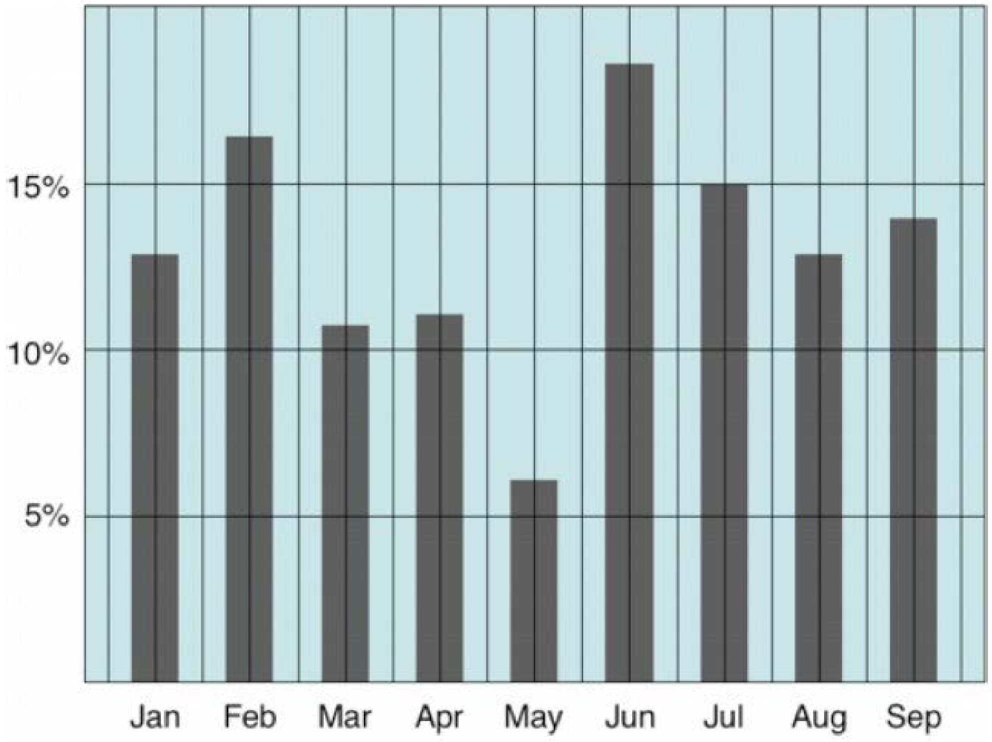
Reduce Chartjunk

Extraneous visual elements distract from the message the data is trying to tell.

- Extra dimensionality
- Uninformative coloring
- Excessive grids and figurative decoration

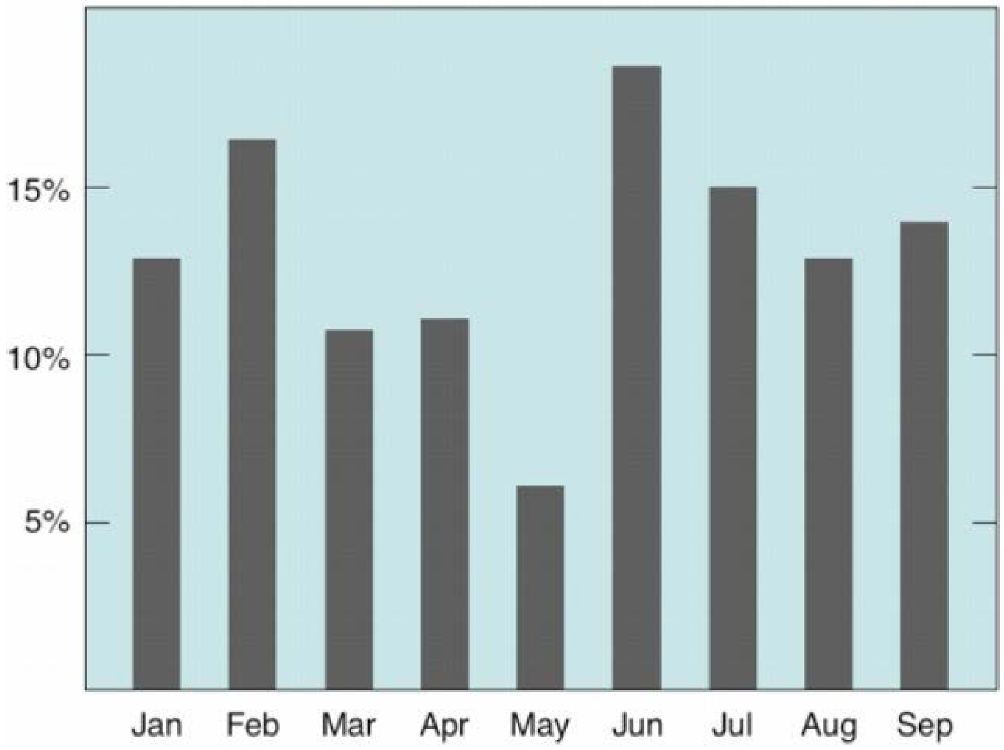
In an exciting graphic, the data tells the story, not the chartjunk.

Can you Simplify this Plot?

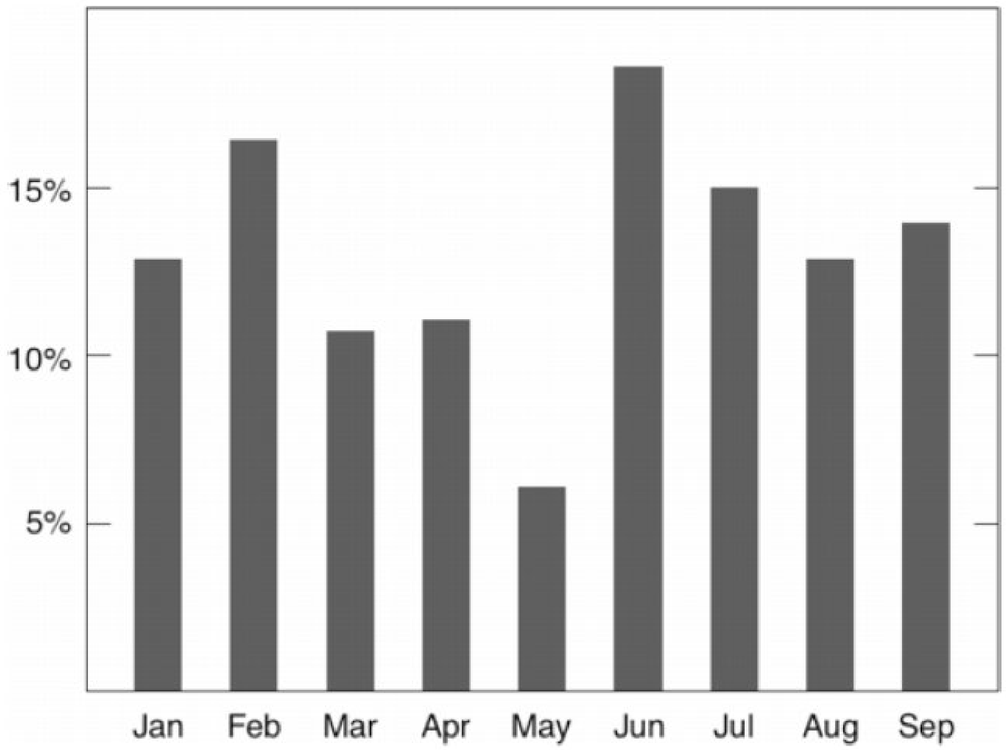


Tim Brey

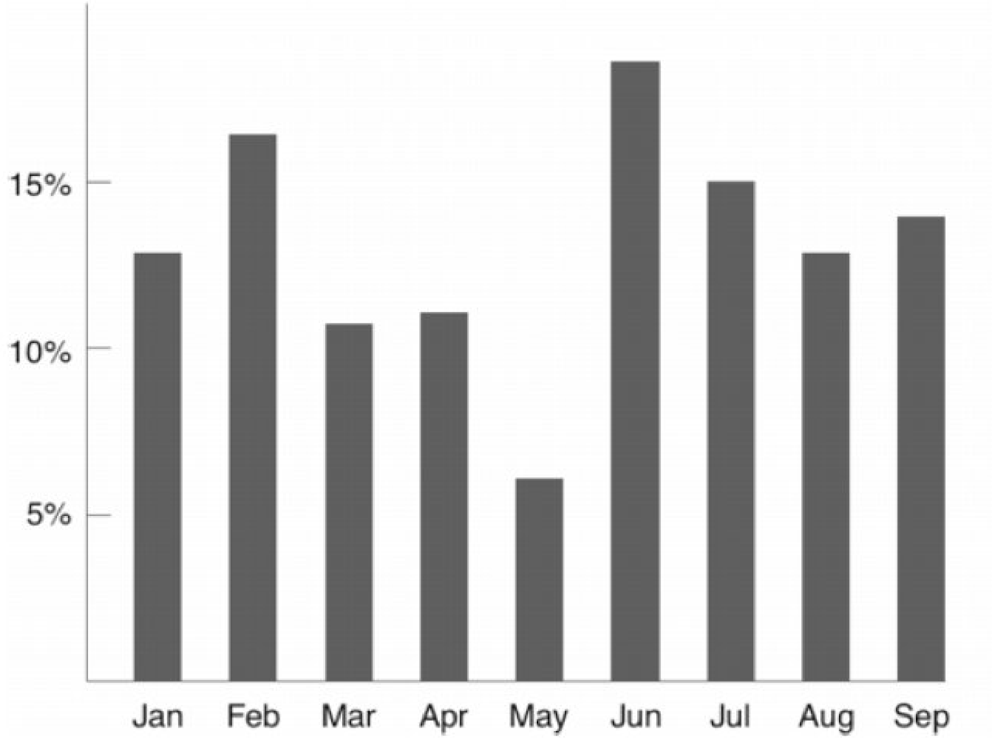
Can You Further Simplify?



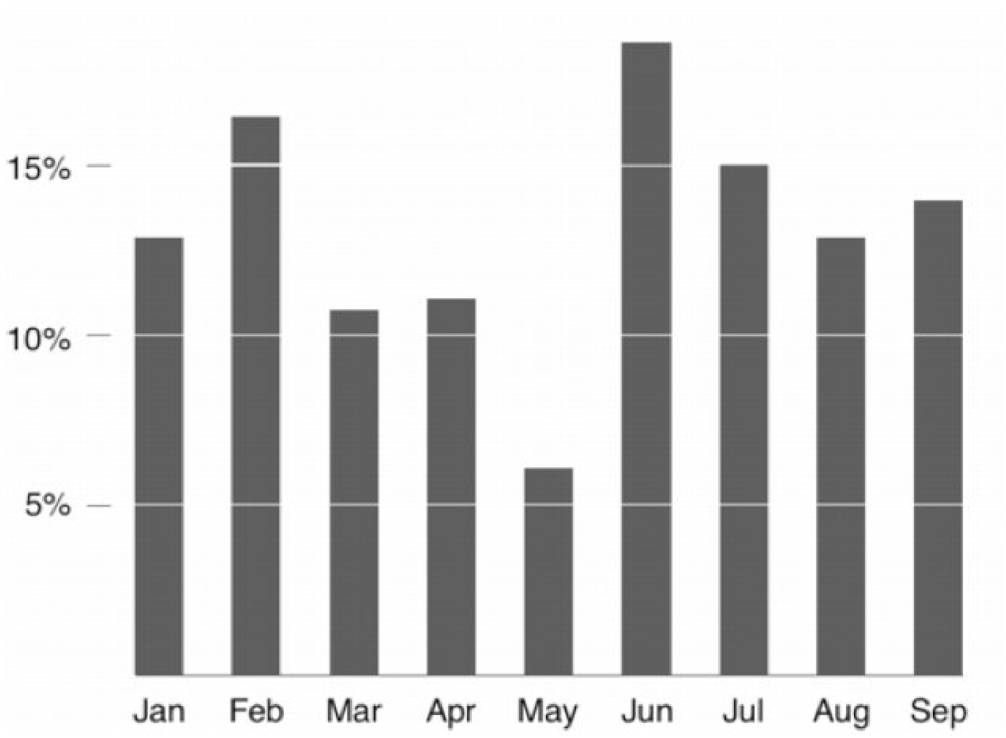
Better, but can you Further Simplify?



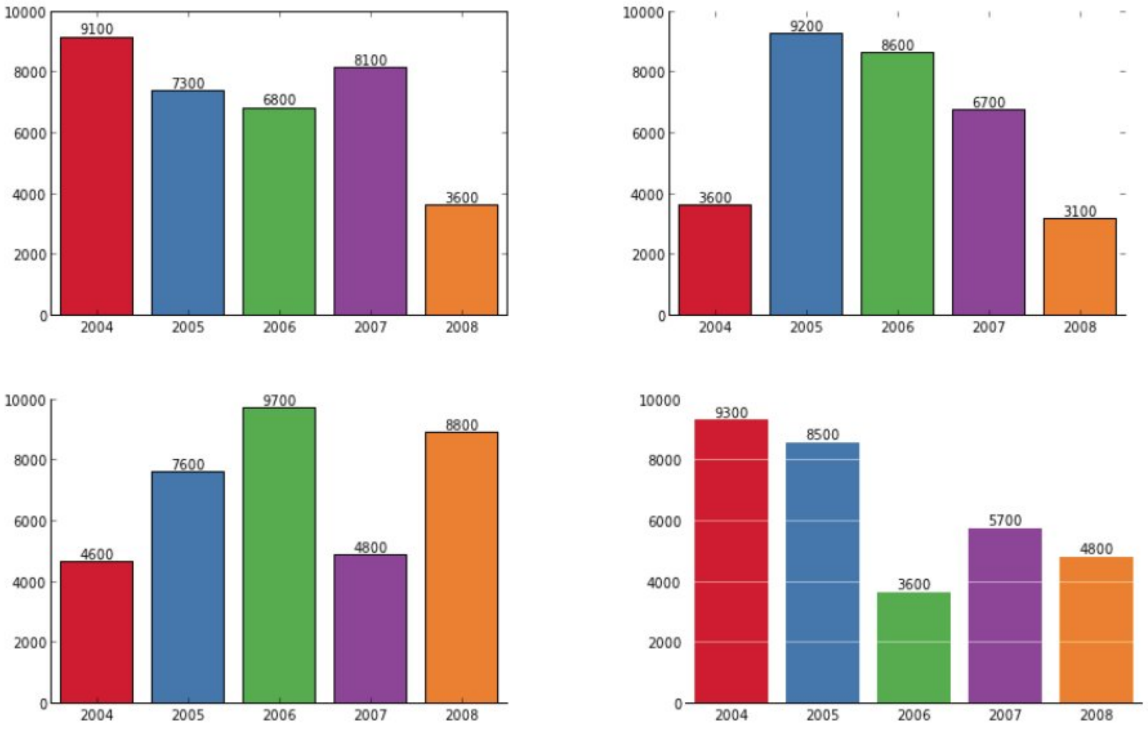
Anything Else that Can Go?



``Less is More''

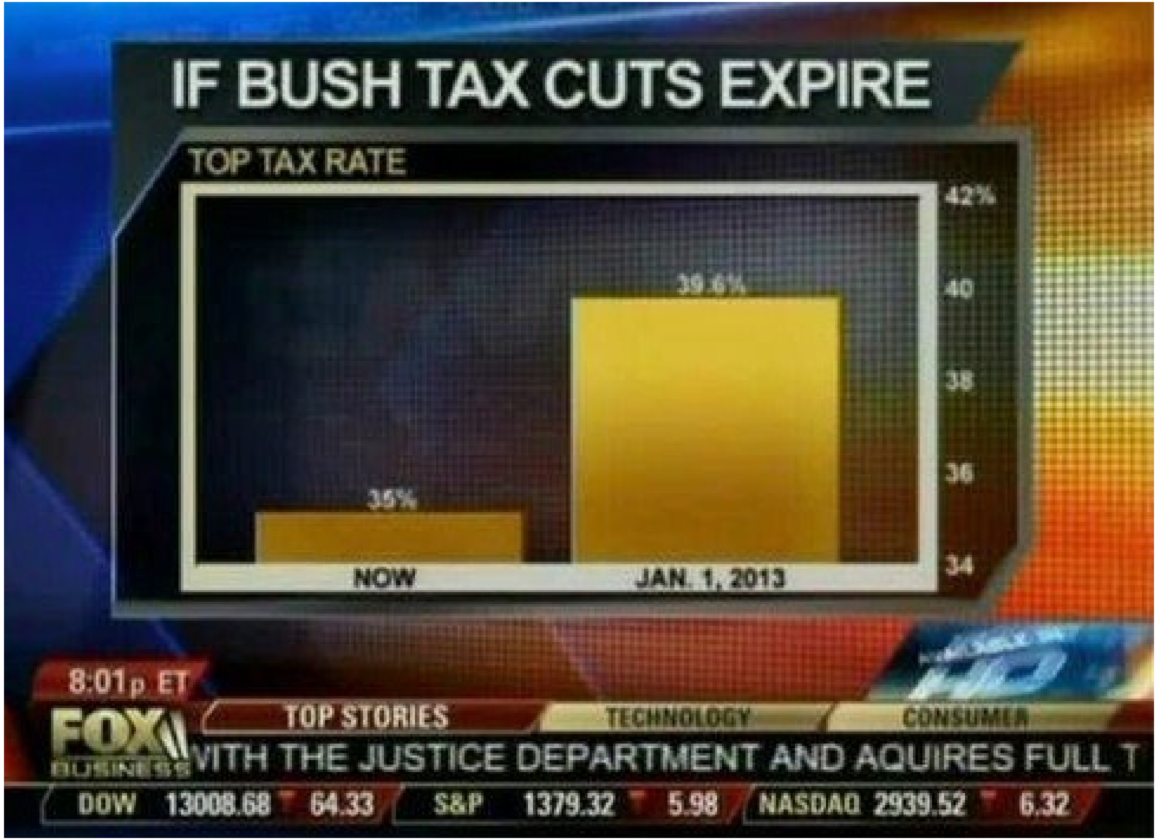


MatPlotLib Supports Nice Plots



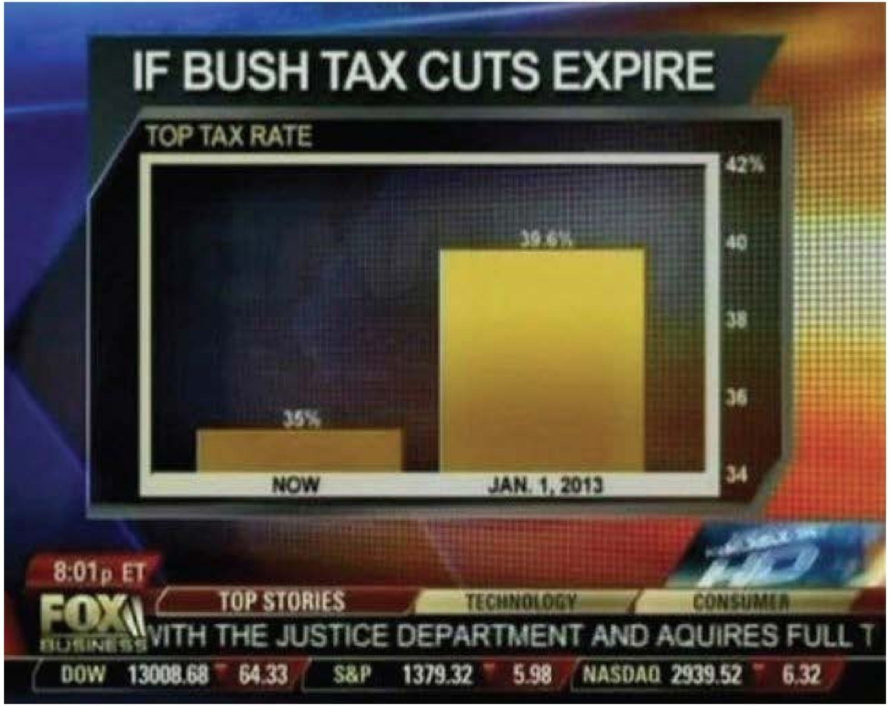
http://nbviewer.ipython.org/5357268

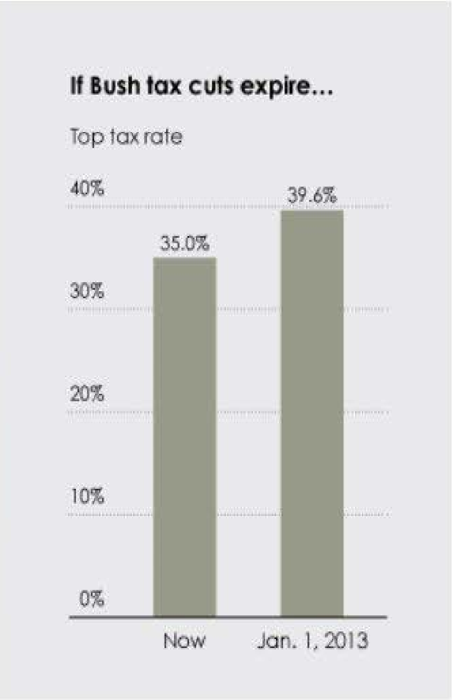
Graphical Integrity: Scale Distortion



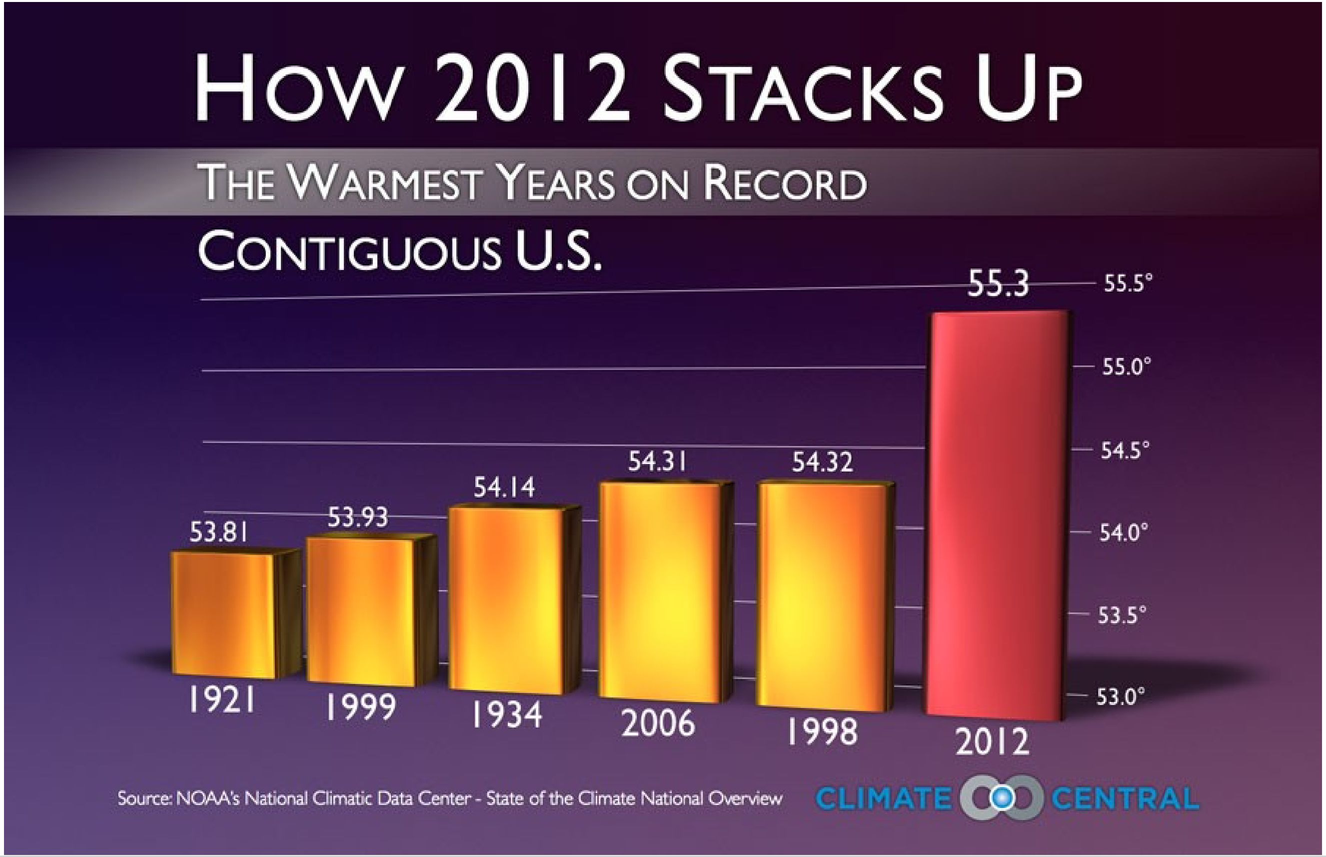
Graphical Integrity: Scale Distortion

Always start bar graphs at zero.



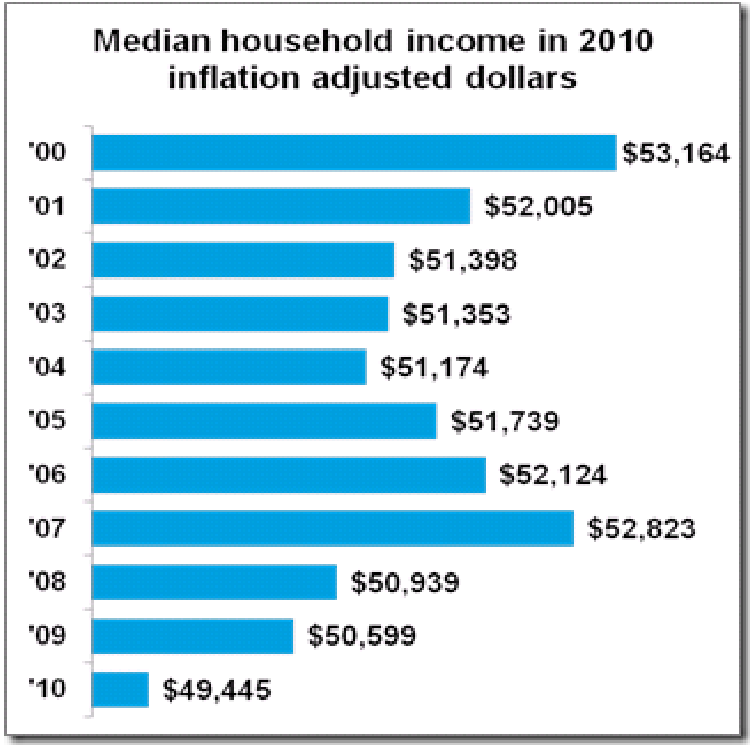


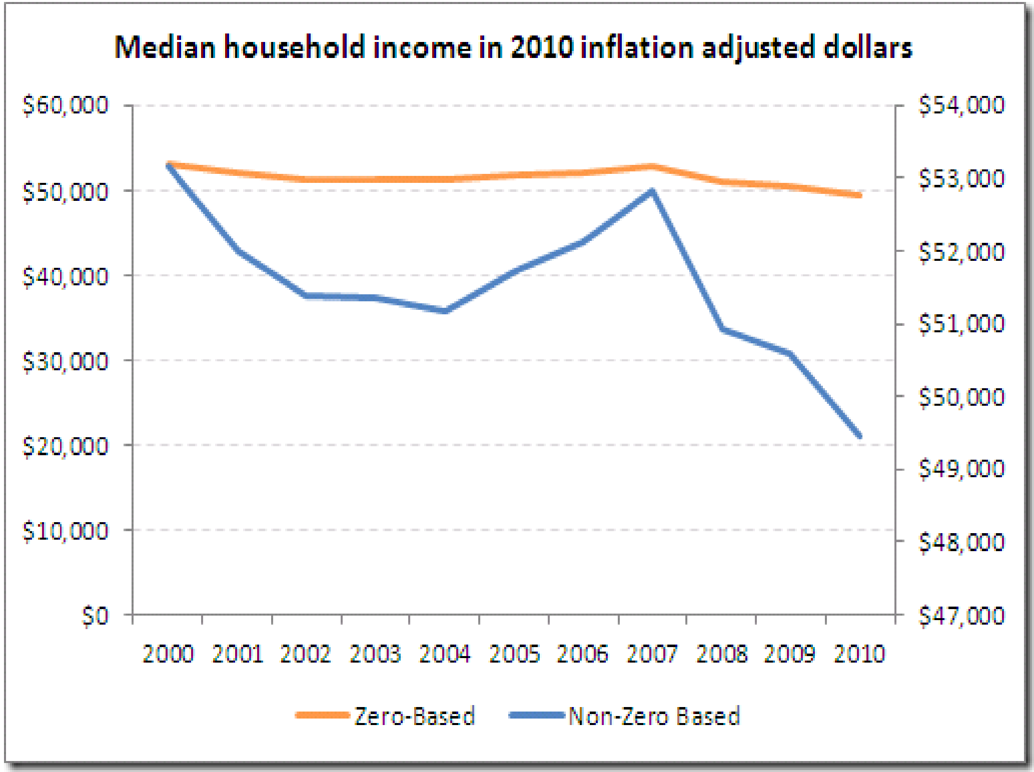
Scale Distortions



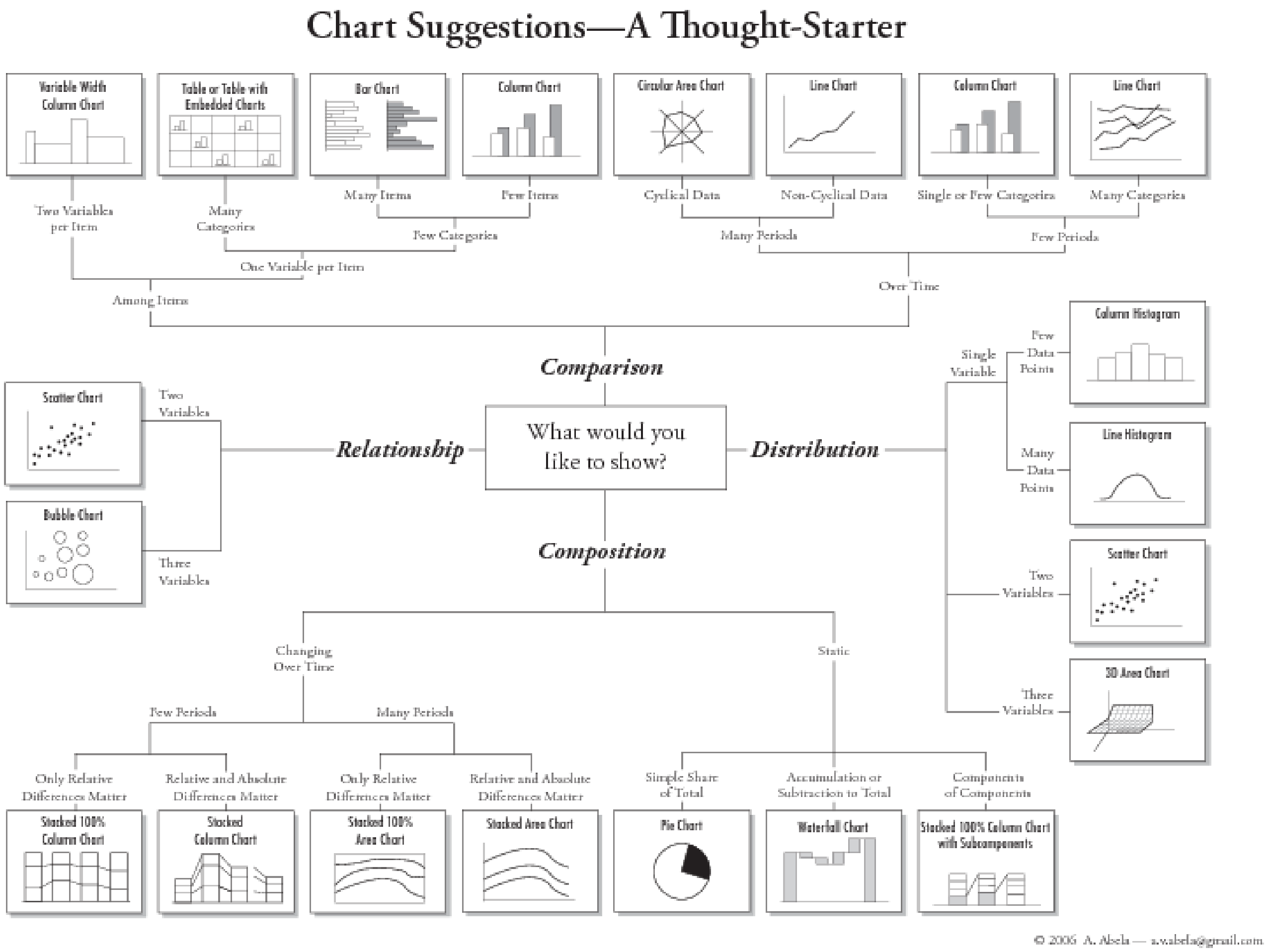
Scale Distortions

Always start your bar graphs at zero!

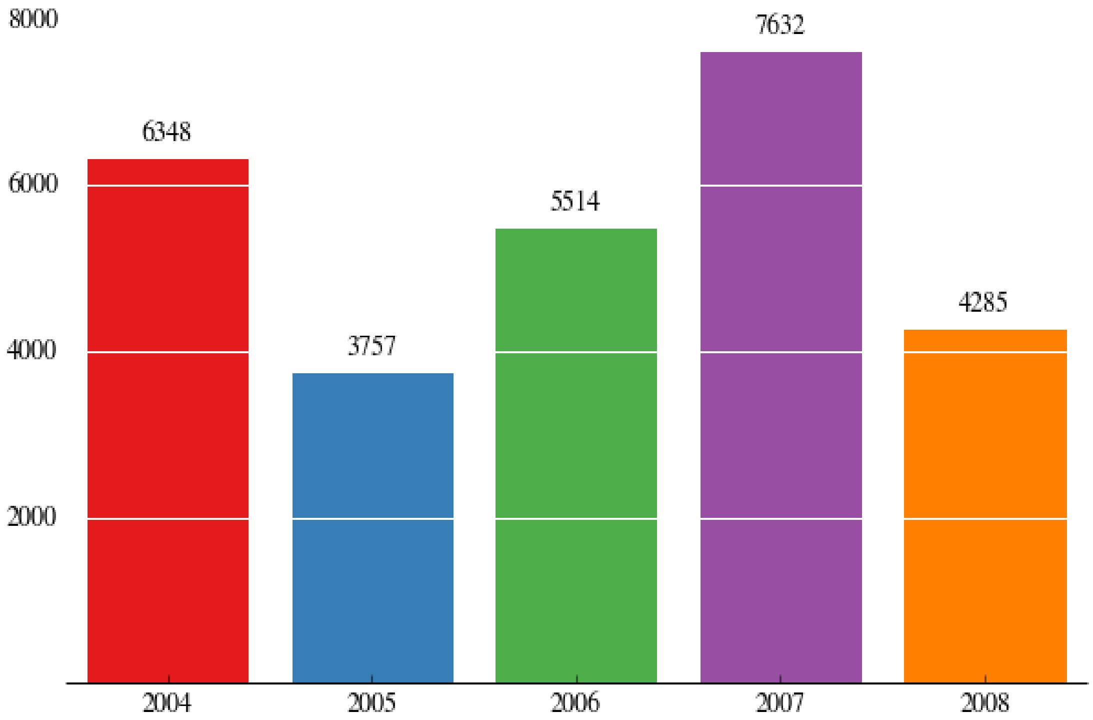




Which Chart to Use?

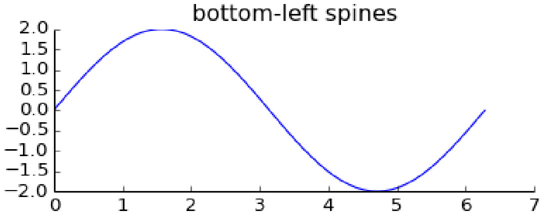


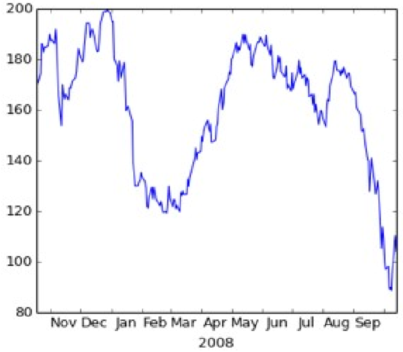
Bar Chart

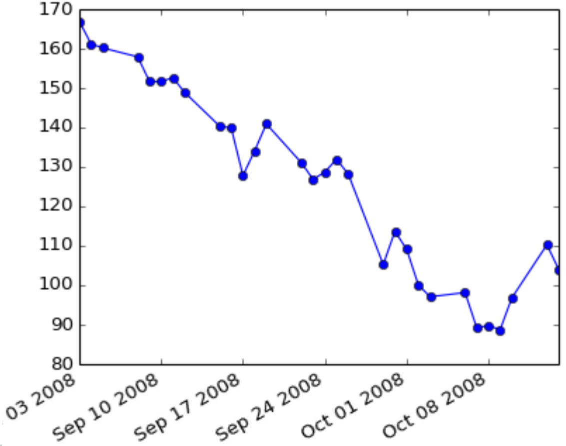


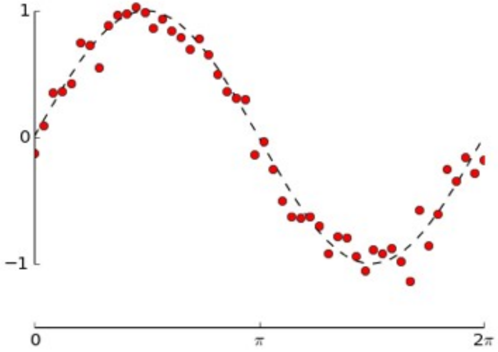
8000
6000
4000
2000
6348
3757
5514
7632
4285
2004
2005
2006
2007
2008

Line Charts - Trends Over Time





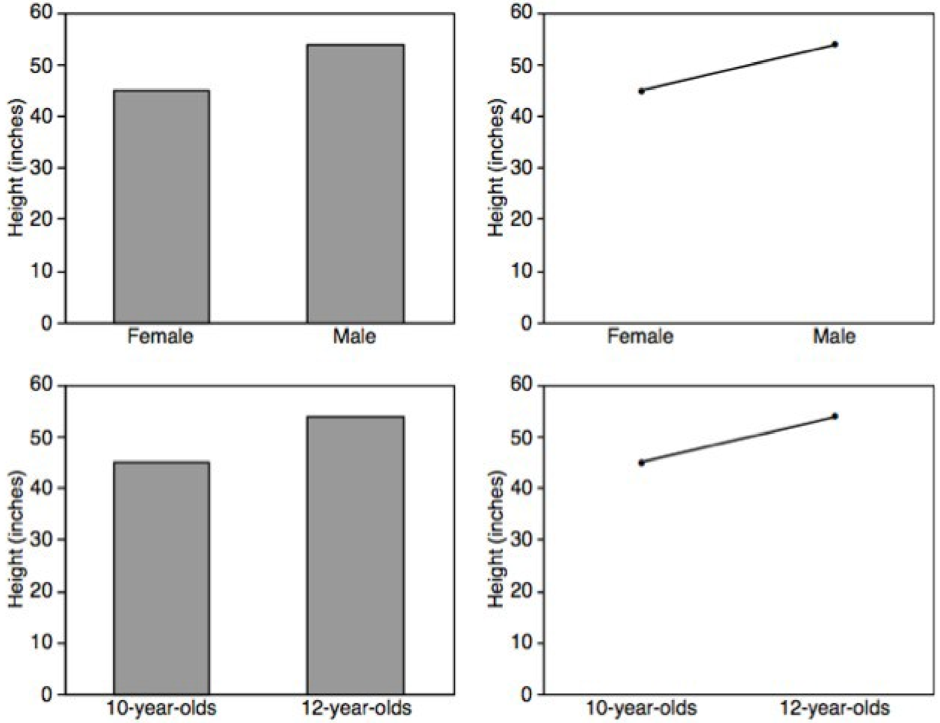




matplotlib gallery

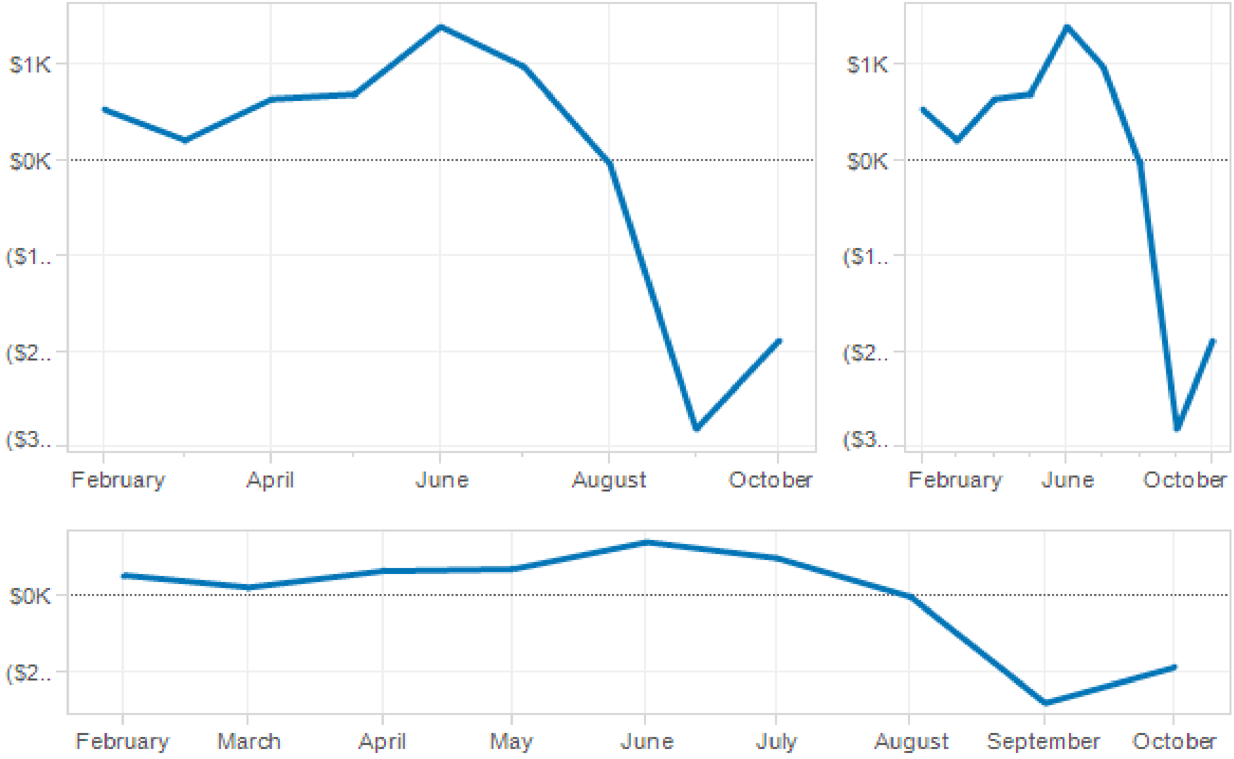
Bars vs. Lines

Lines imply connections - do not use for categorical data



Zacks 1999

Aspect Ratios

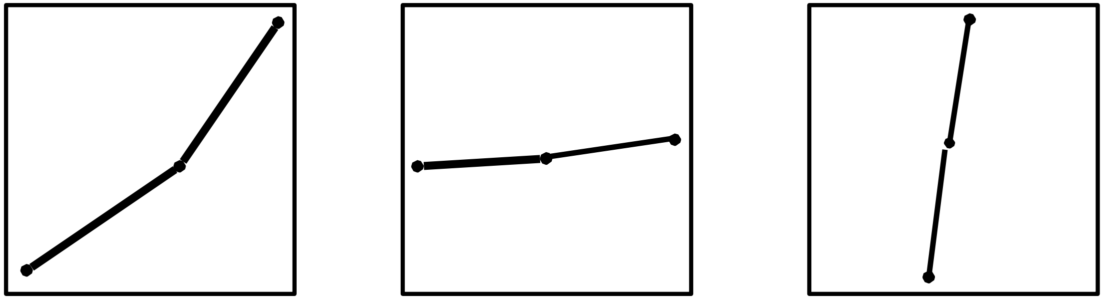


eagereyes.org

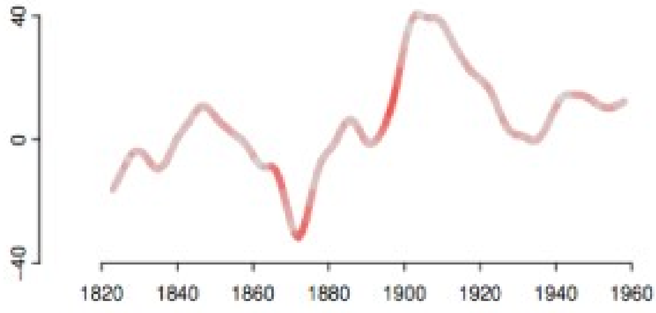
Banking to 45°

Two line segments are maximally discriminable when their average absolute angle is 45°

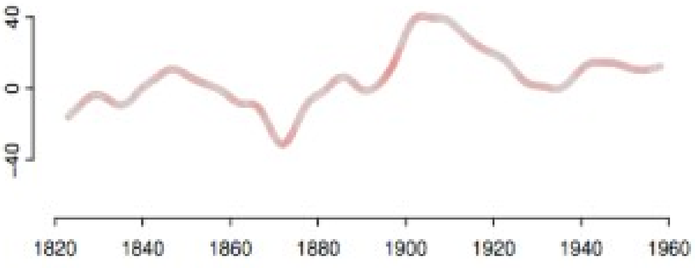
W. Cleveland



Banking to 45°

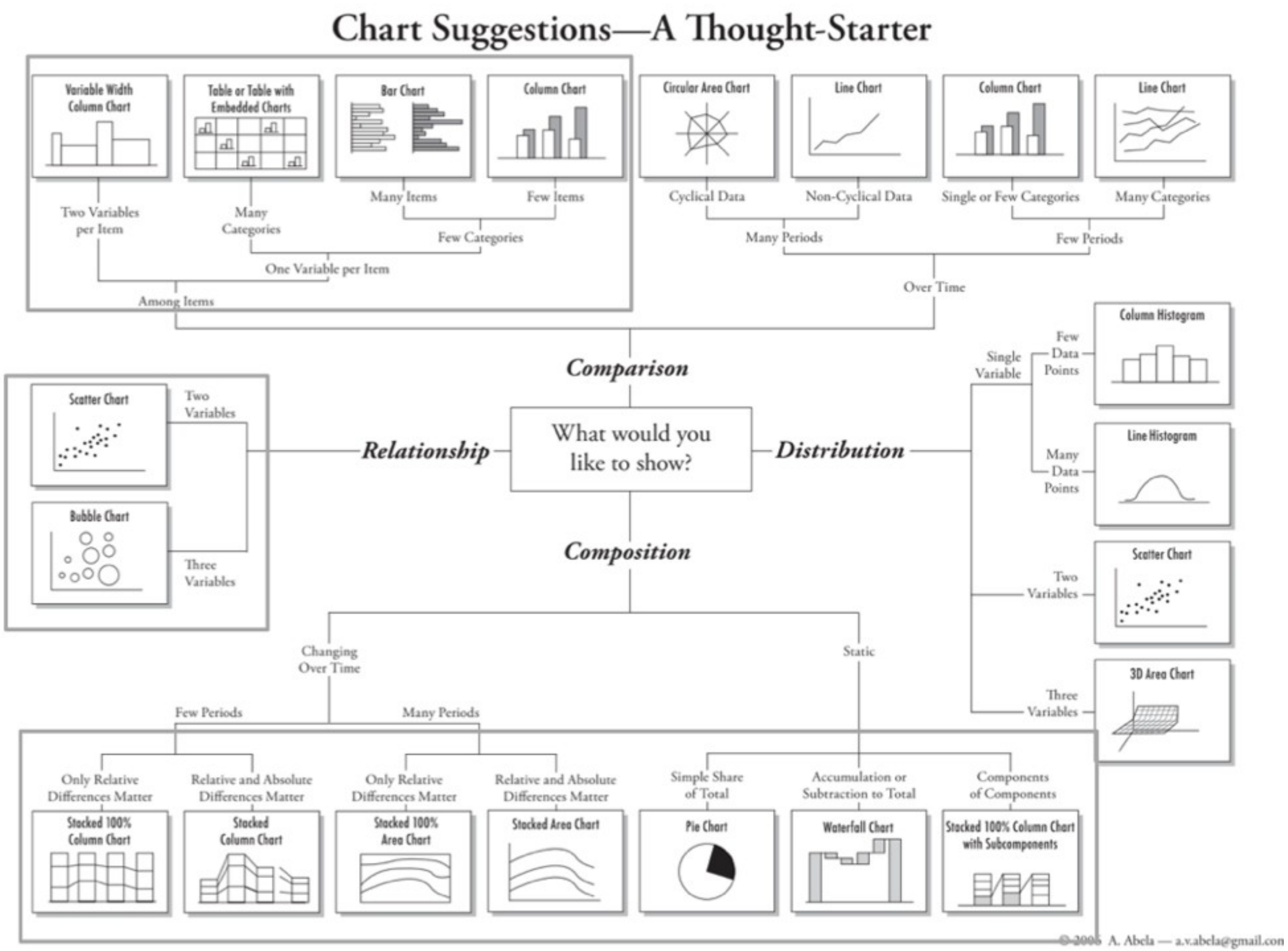


Error Prone



Optimal Aspect Ratio

An Empirical Model of Slope Ratio Comparisons,
Justin Talbot, John Gerth, Pat Hanrahan

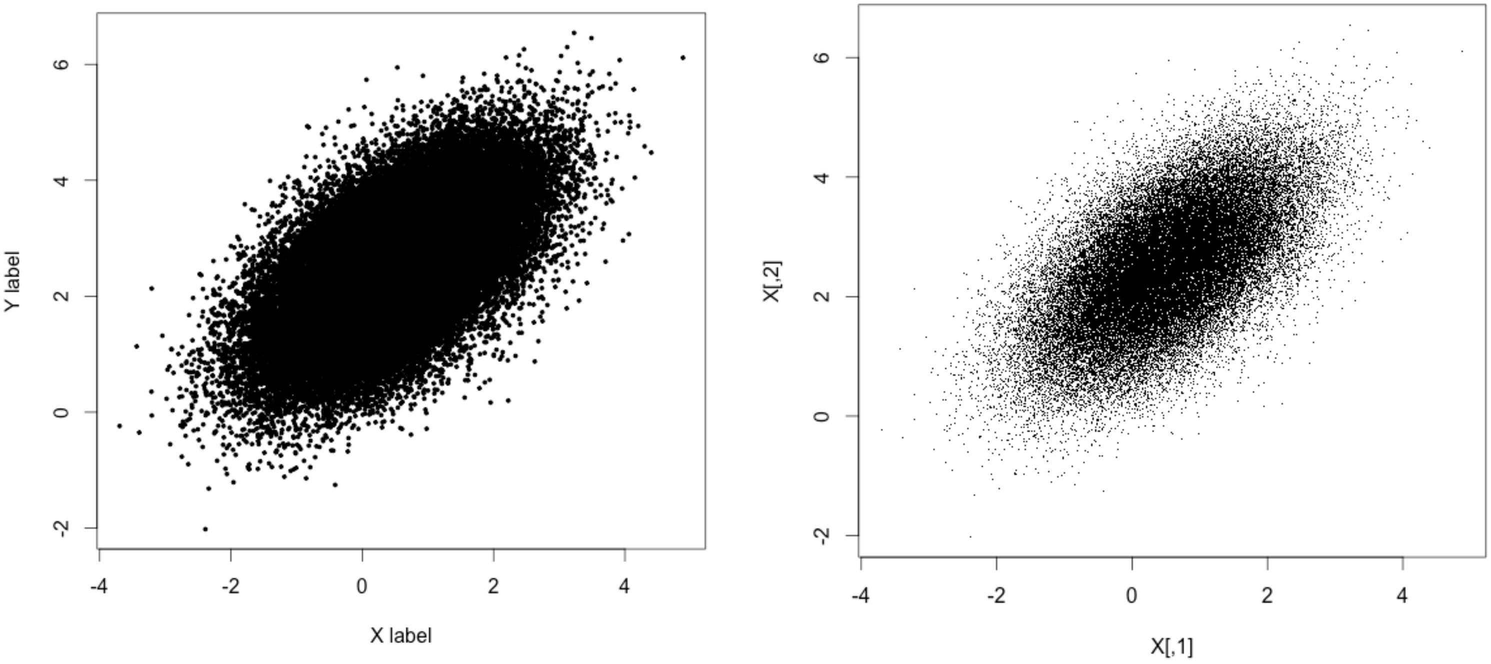


http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

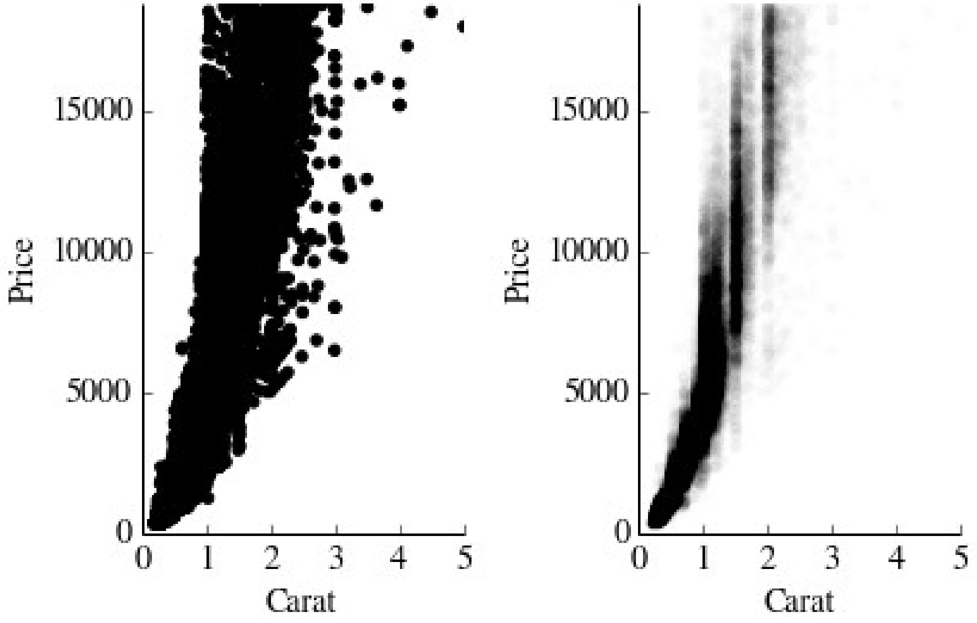
Scatter Plots / Bubble Charts

- Scatter plots show the values of each point, and are a great way to present 2D data sets.
- For data sets with three or four variables, use bubble charts.
- Higher dimensional datasets can be projected to 2D through principle component analysis.

Reduce Overplotting by Small Points



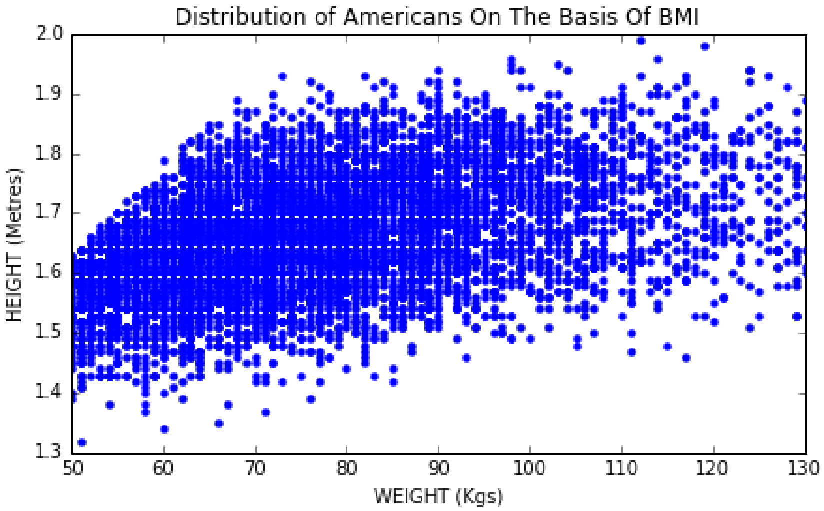
Reduce Overplotting by Opacity

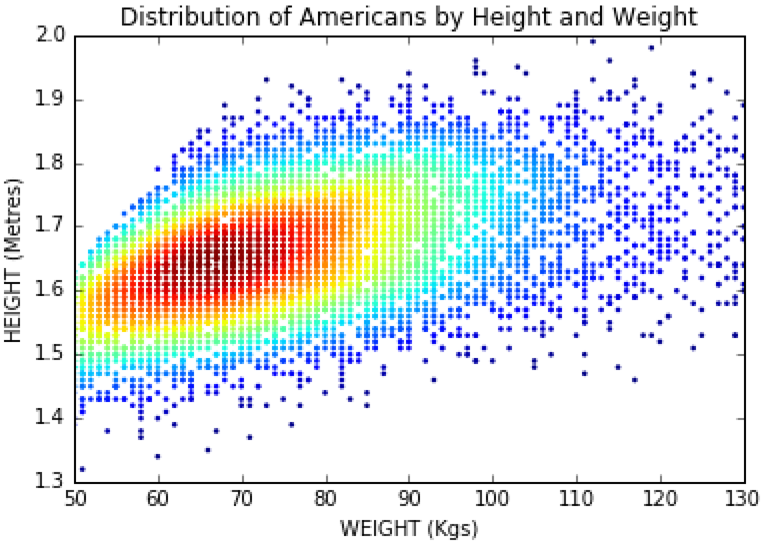


alpha = 1/100

Heatmaps Reveal Finer Structure

Color points on the basis of frequency

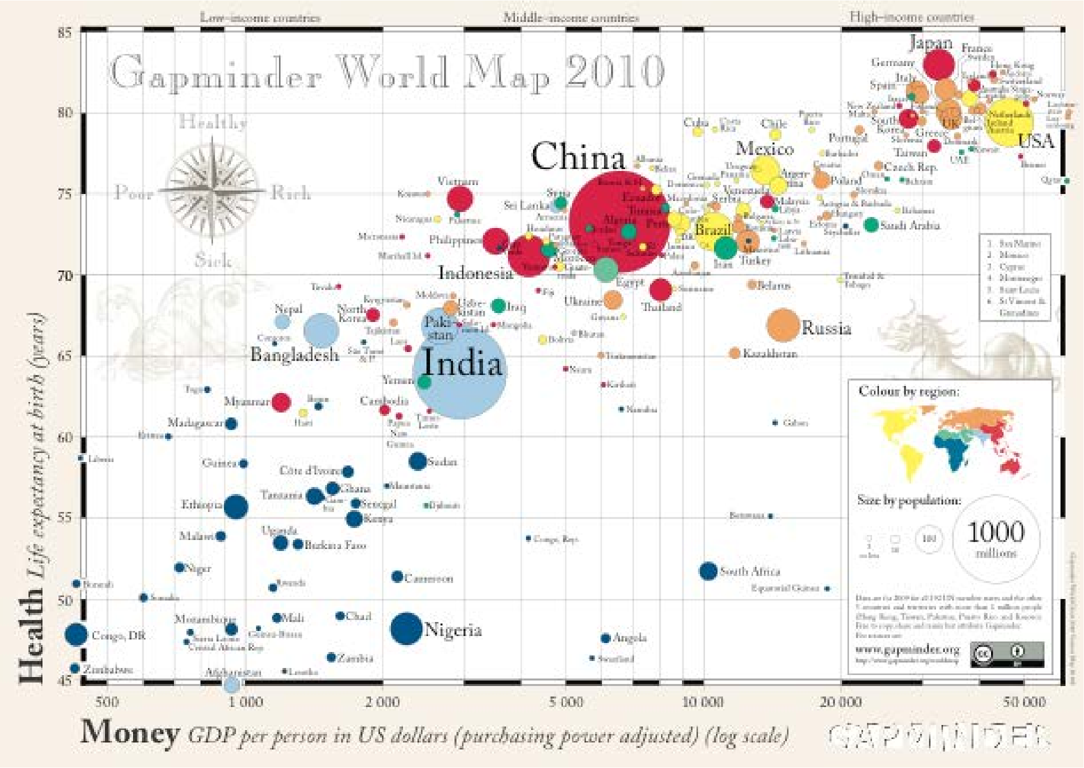




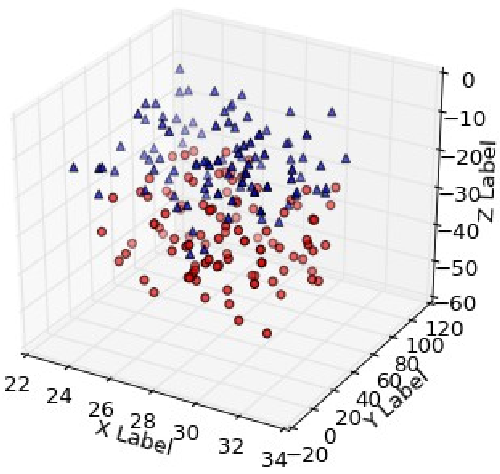
Bubble Charts for Extra Dimensions

Using color, shape, and size of “dots” enables dot plots to represent additional dimensions.

http://www.gapminder.org/videos/200-years-that-changed-the-world-bbc/



Don't



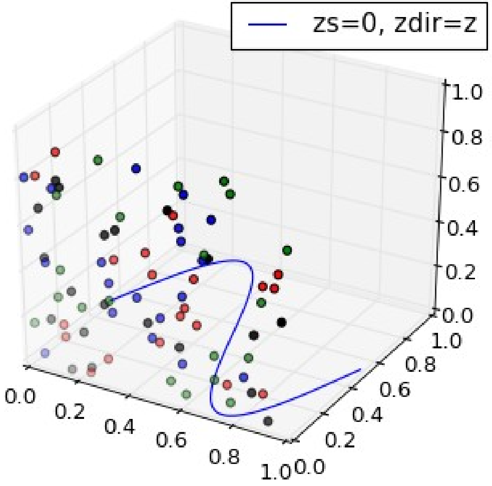
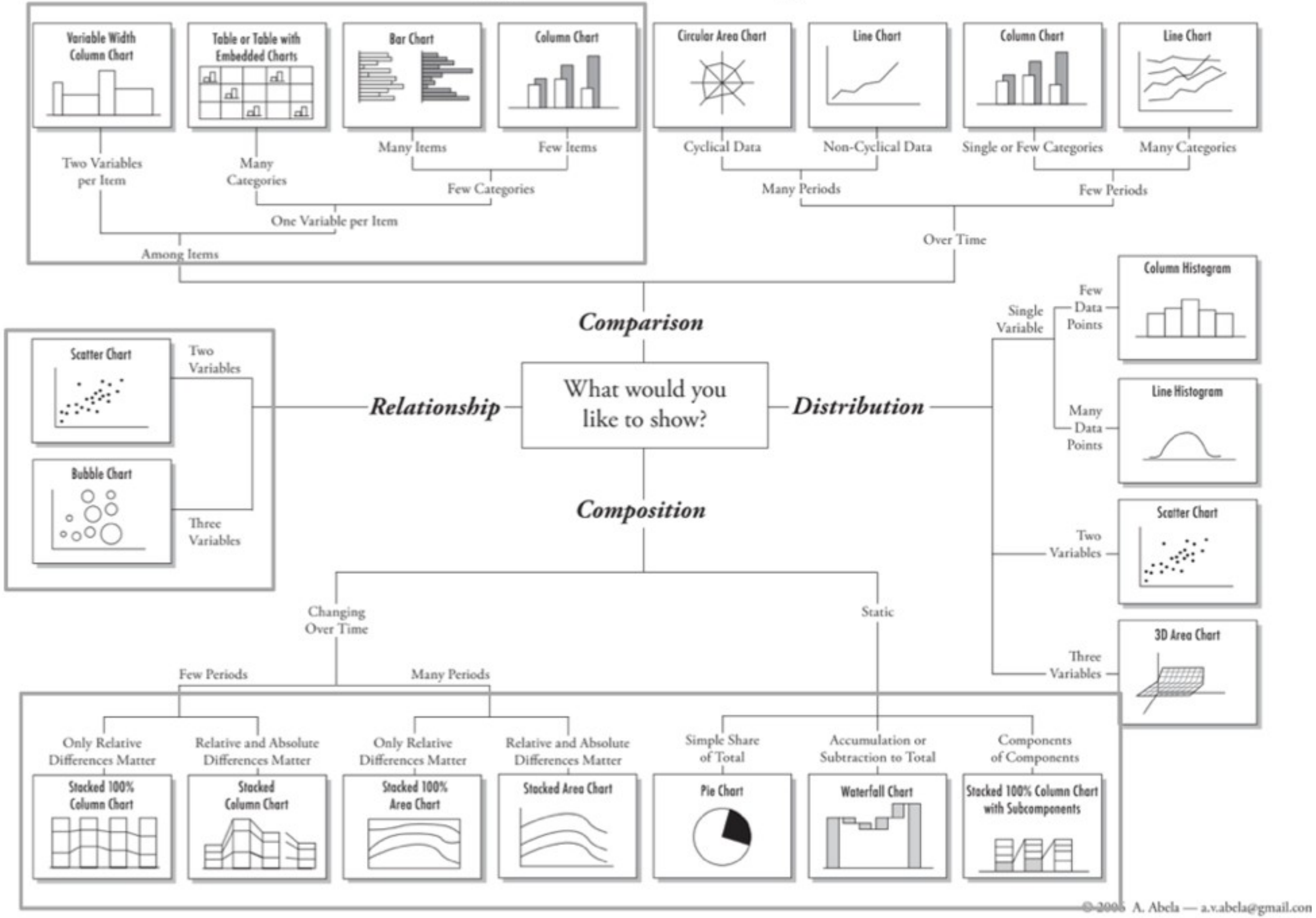
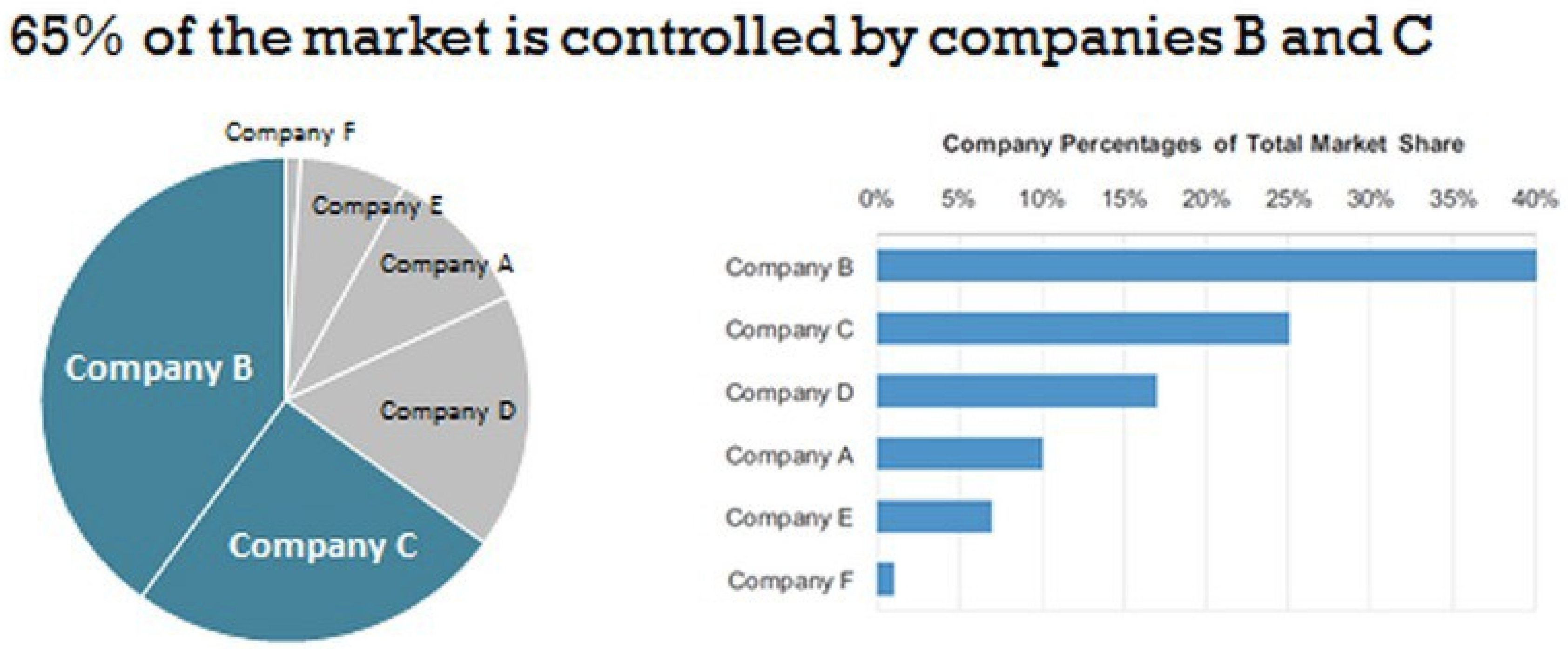


Chart Suggestions—A Thought-Starter

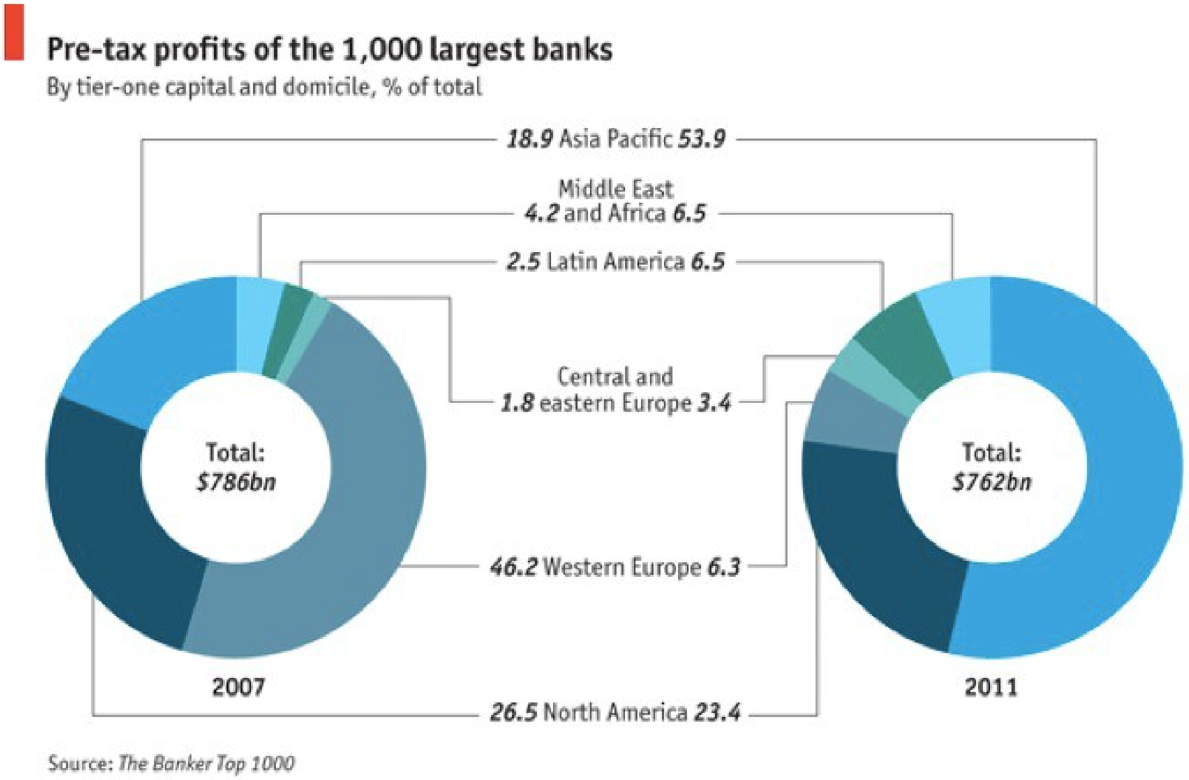


http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

Pie vs. Bar Charts

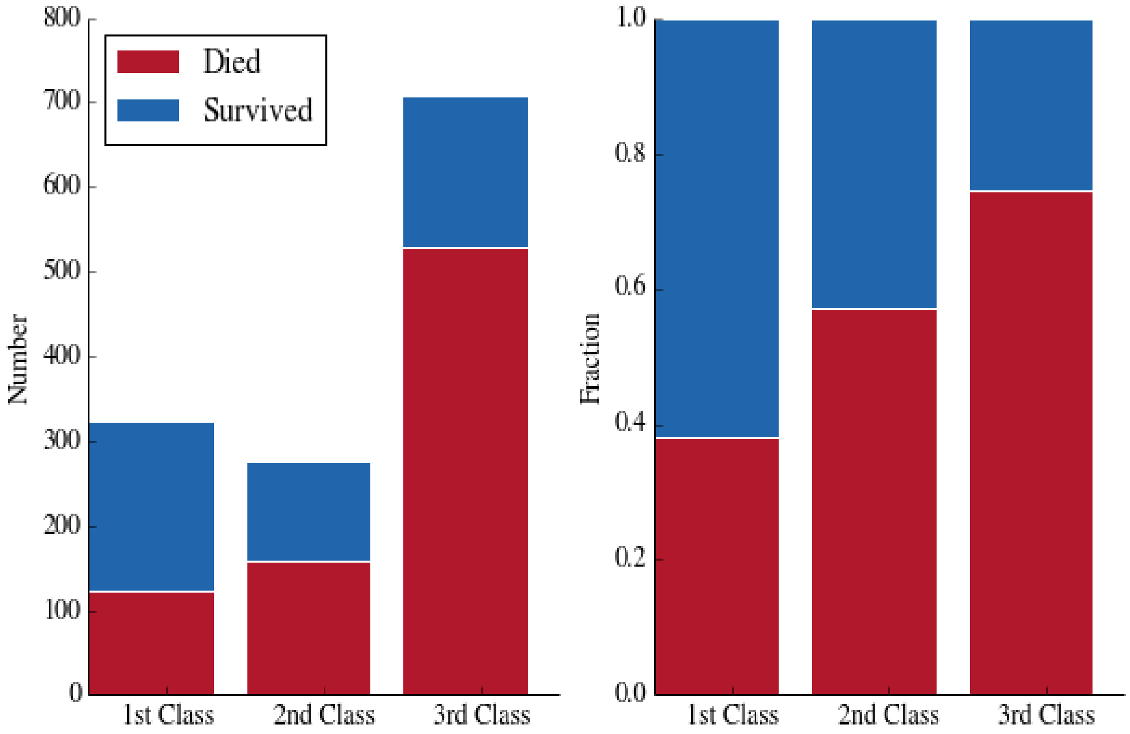


Donut Chart

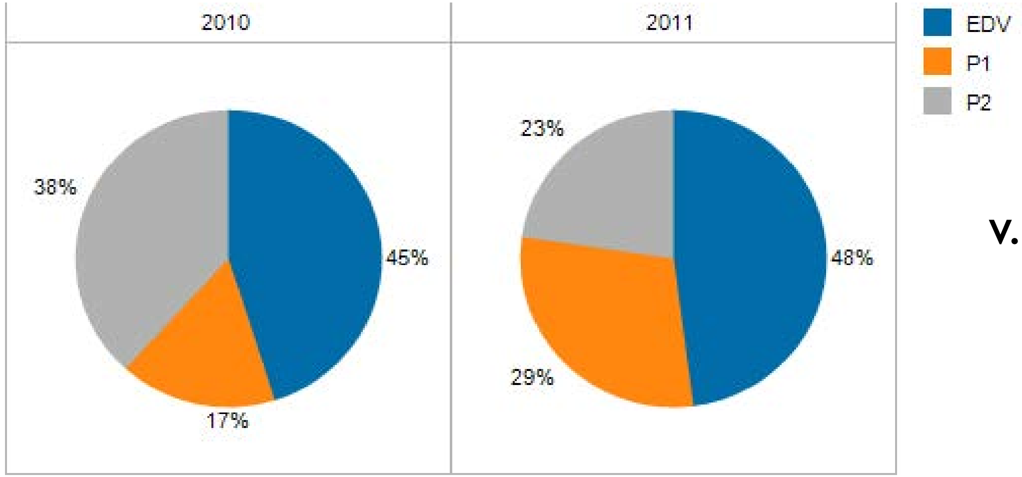


The Economist Daily Chart

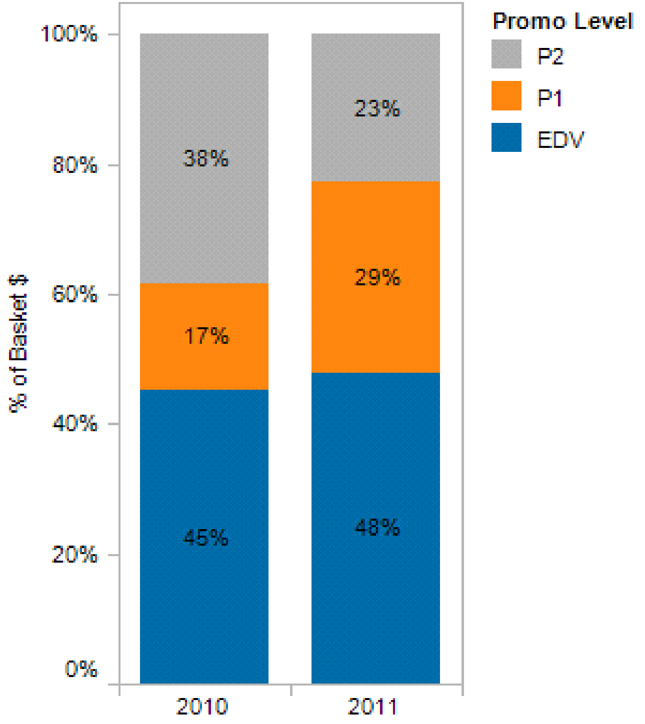
Stacked Bar Chart



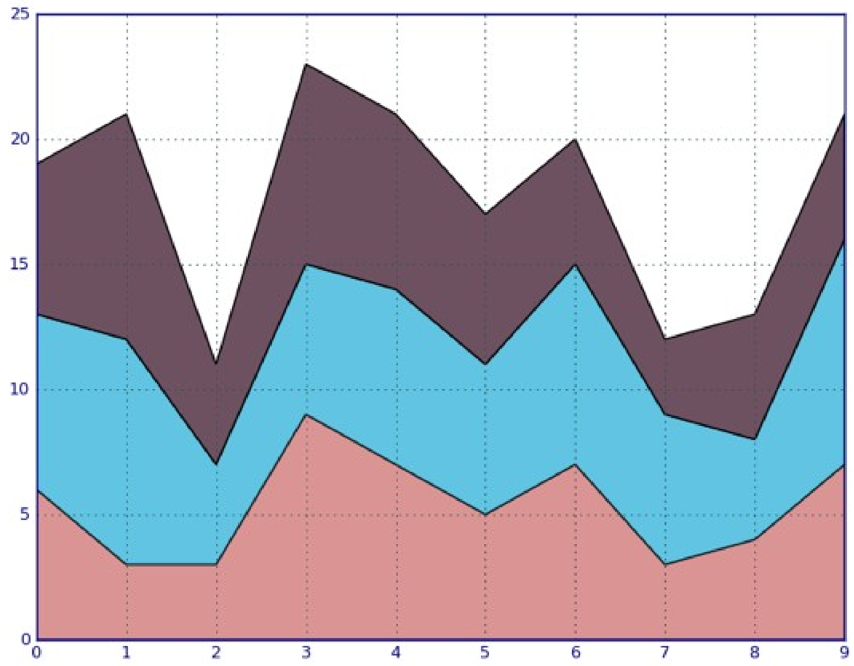
Stacked Bar Chart



v.s.

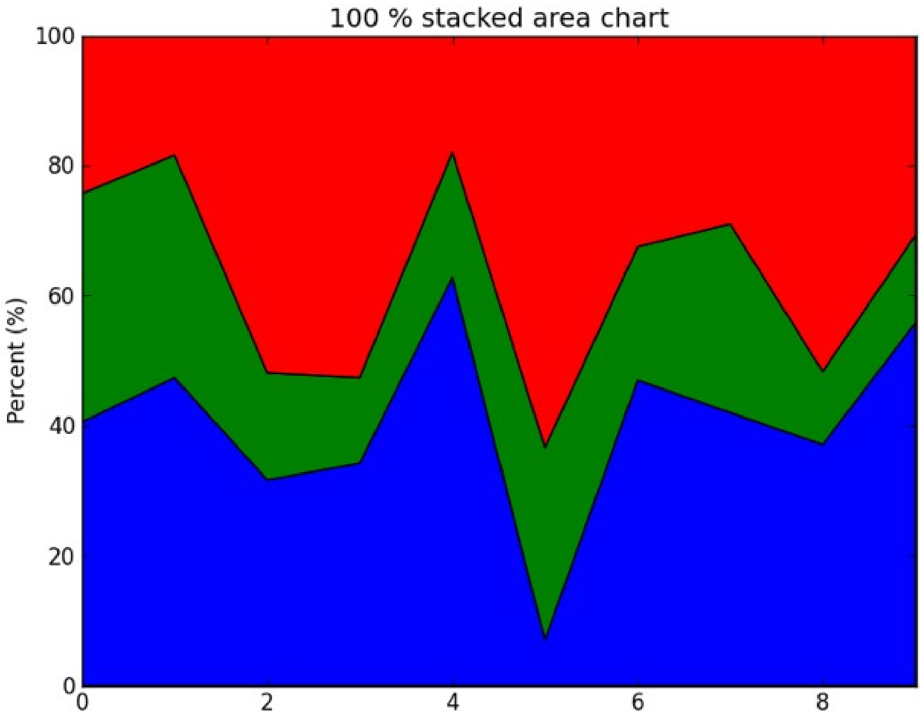


Stacked Area Chart



http://stackoverflow.com/questions/2225995/how-can-i-create-stacked-line-graph-with-matplotlib

100% Stacked Area Chart

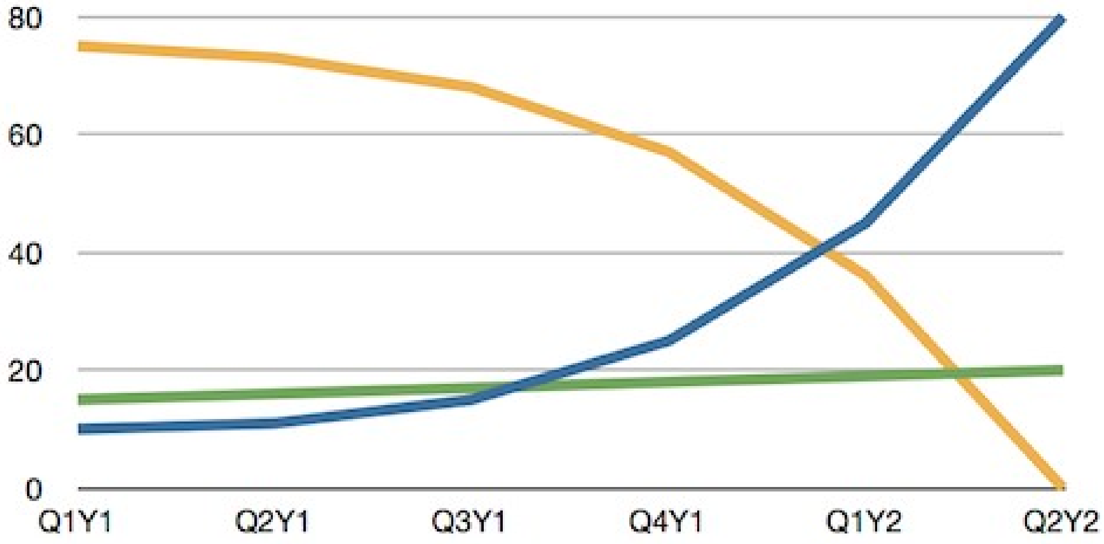


http://stackoverflow.com/questions/16875546/create-a-100-stacked-area-chart-with-matplotlib

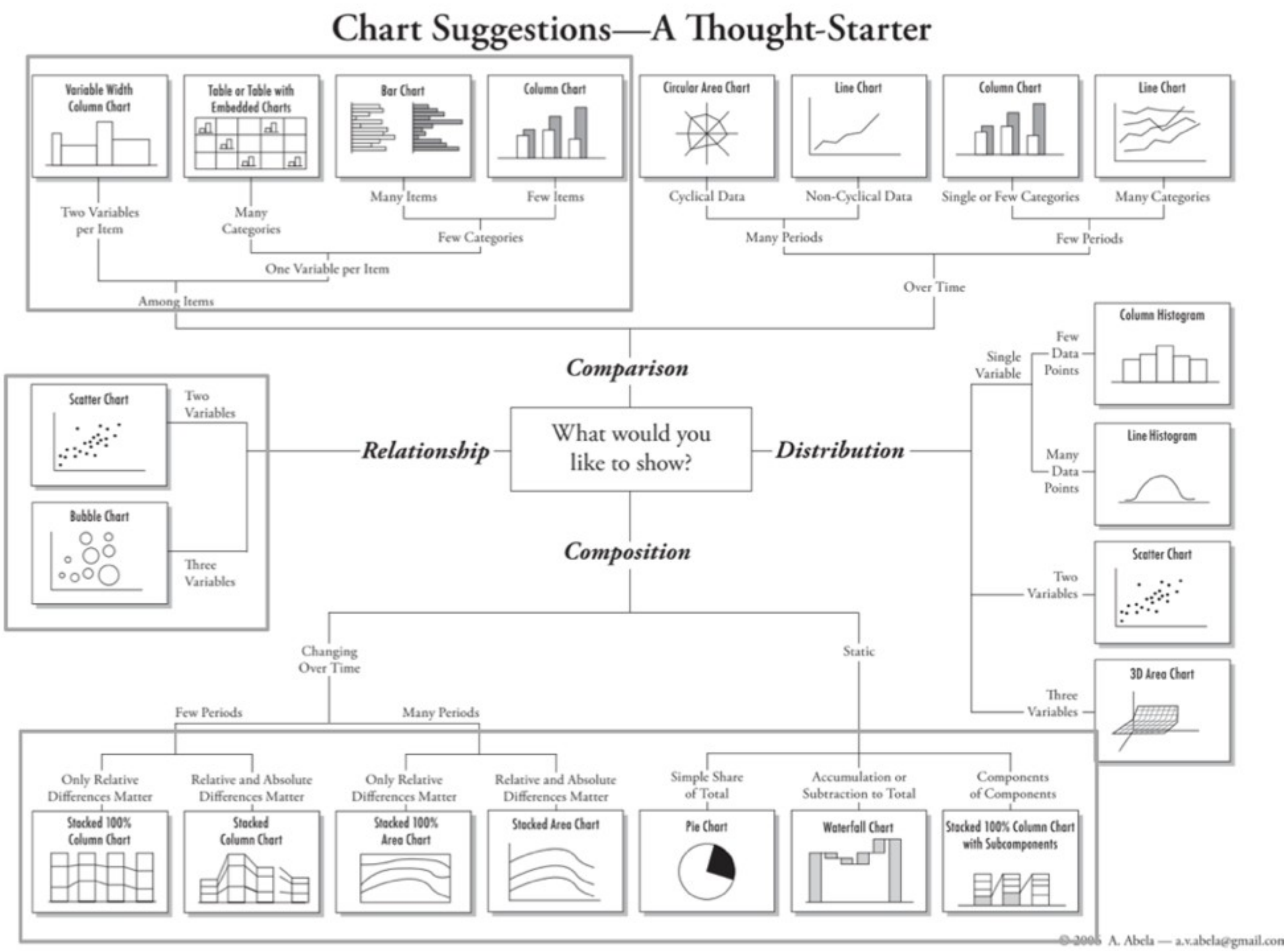
Stacked Area vs. Line Graphs

leancrew.com &
Practically Efficient

Stacked Area vs. Line Graphs



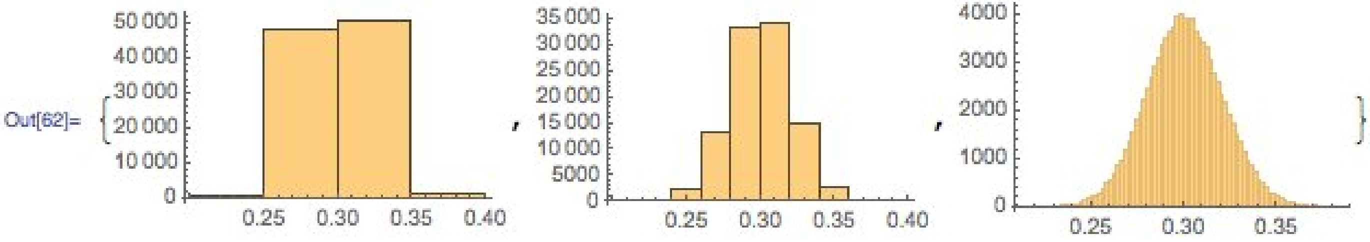
leancrew.com &
Practically Efficient



http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

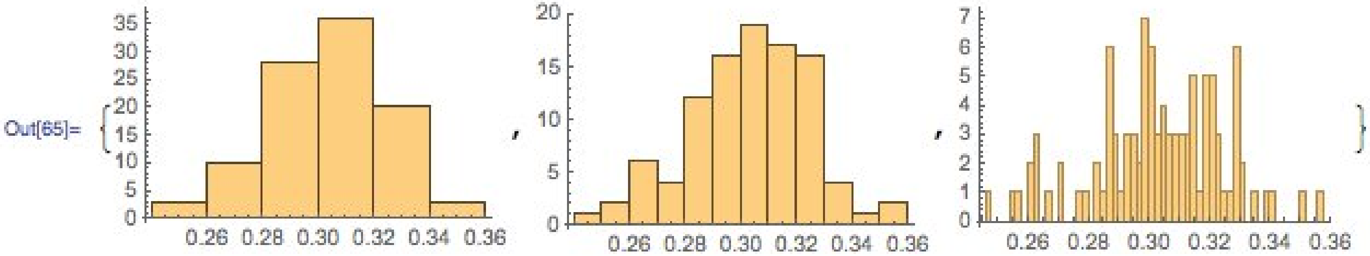
Histograms: Bin Size / Count Matters

```
In[62]:= {Histogram[d, 5], Histogram[d, 10], Histogram[d, 100]}
```



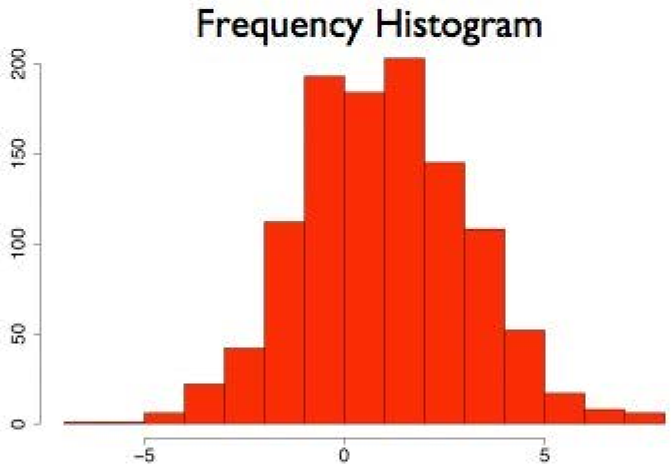

```
d100 = Take[d, 100];
```

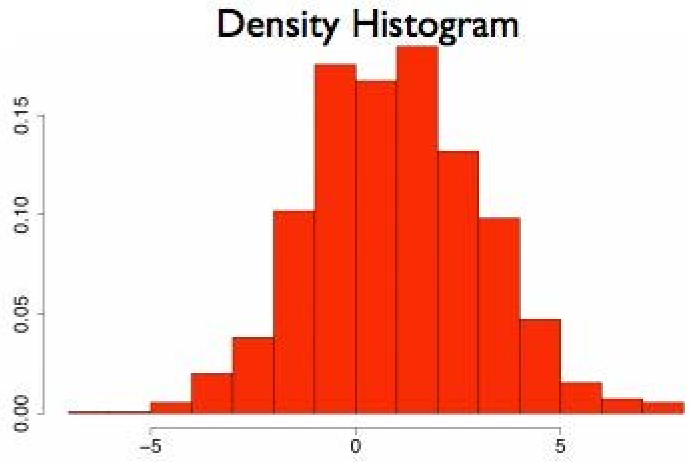
```
In[65]:= {Histogram[d100, 5], Histogram[d100, 10], Histogram[d100, 100]}
```



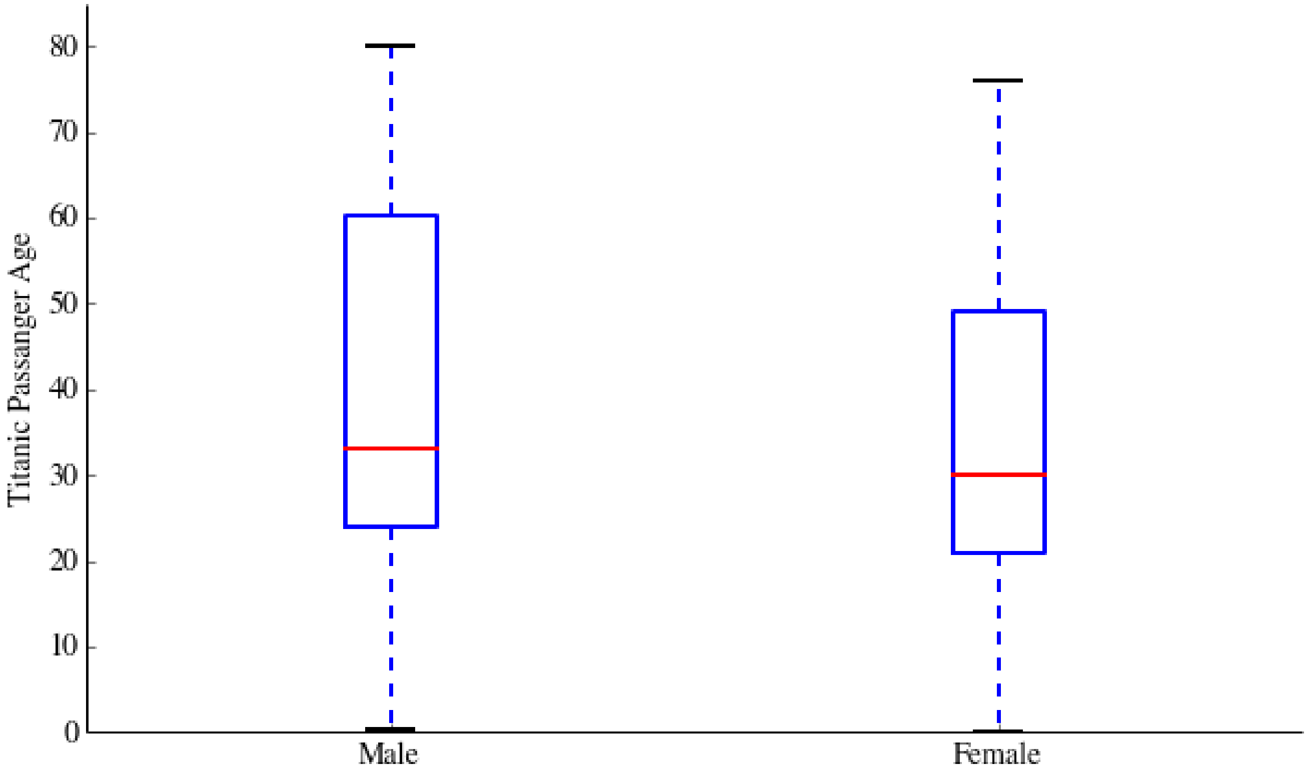
Frequency vs. Density Histograms

Dividing counts by the total yields a probability density plot, which is more interpretable:

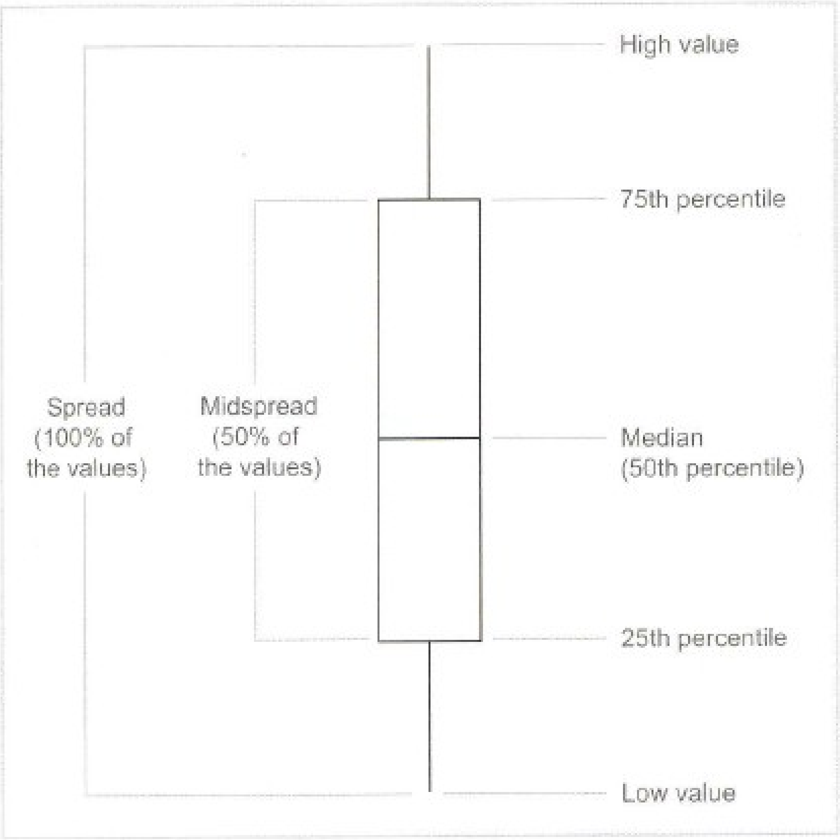


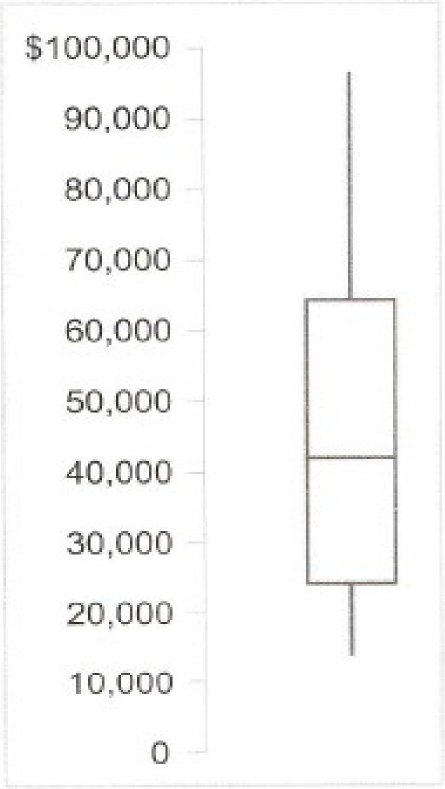


Box & Whisker Plots



Box & Whisker Plots





Few, "Show me the Numbers"

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.