CSE 591: Data Science

Steven Skiena
Stony Brook University

Lecture 11: Principles of Visualizing Data

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.

Why Data Visualization?

- Exploratory data analysis: what does your data really look like?
- Error detection: did you do something stupid?
- Presenting what you have learned to others.

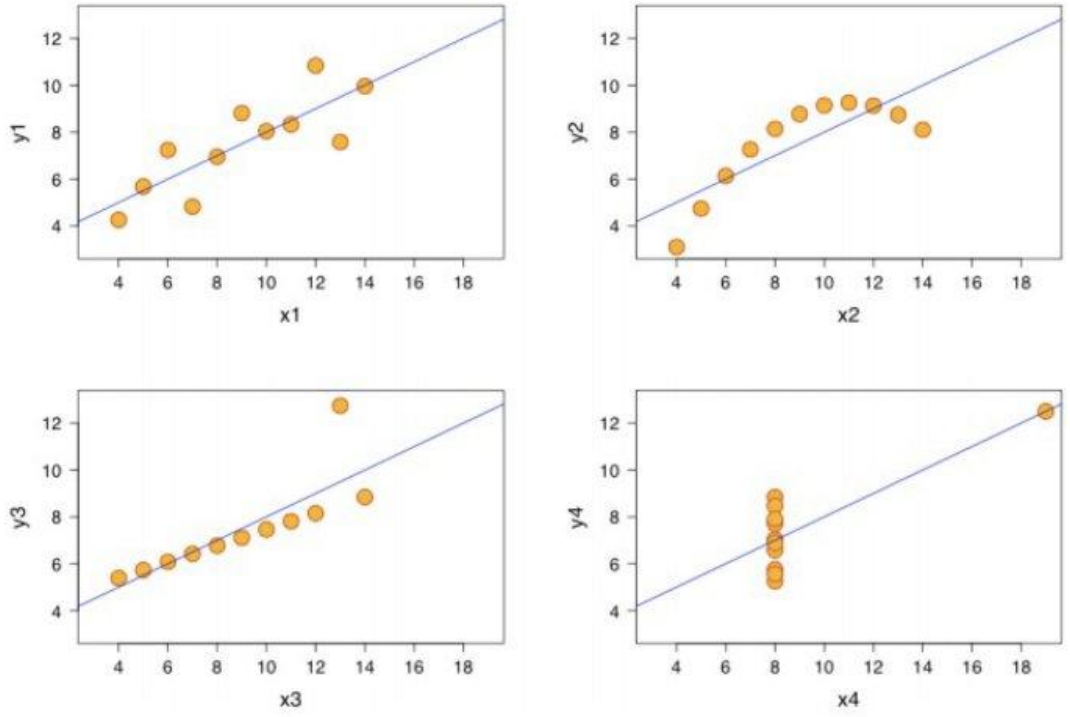
A large fraction of the graphs and charts I see are terrible: visualization is harder than it looks.

Ascombe'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 Ascombe's Quartet



Appreciating Art: Which is Better?

Sensible appreciation of art requires developing a particular visual aesthetic.

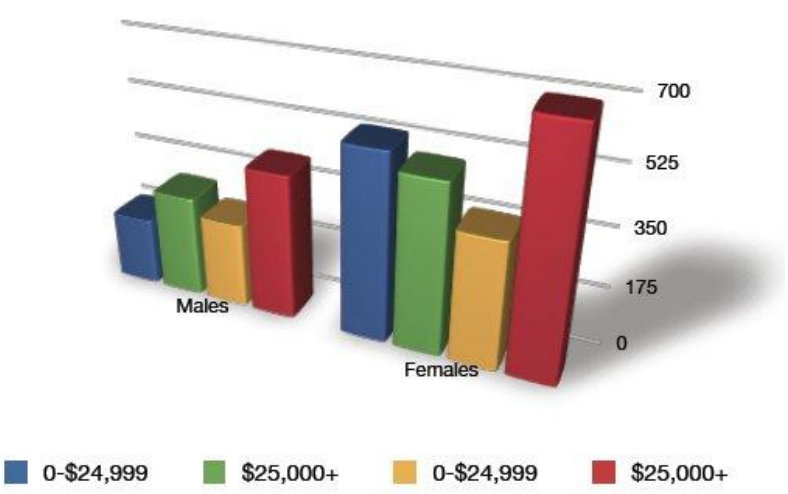
Tufte's Visualization Aesthetic

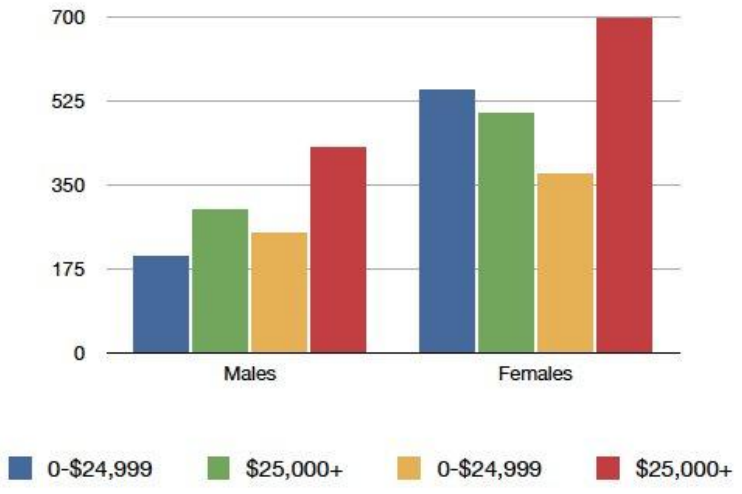
Distinguishing good/bad visualizations requires a design aesthetic, and a vocabulary to talk about data representations:

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

Maximize Data-Ink Ratio

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

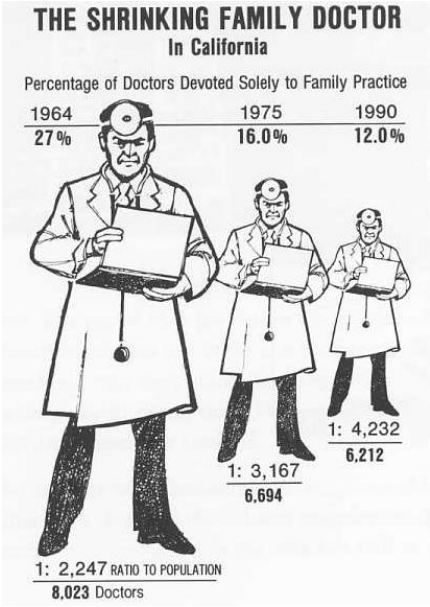


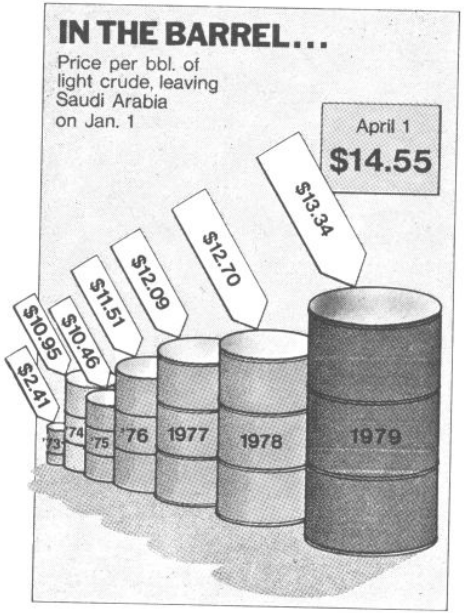


The Lie Factor: Dimensionality

(size of effect in graphic) / (size of effect in data)

The fixing a two- or three-dimensional representation by a single parameter yields a lie, because area or volume increase non-proportionally to length.



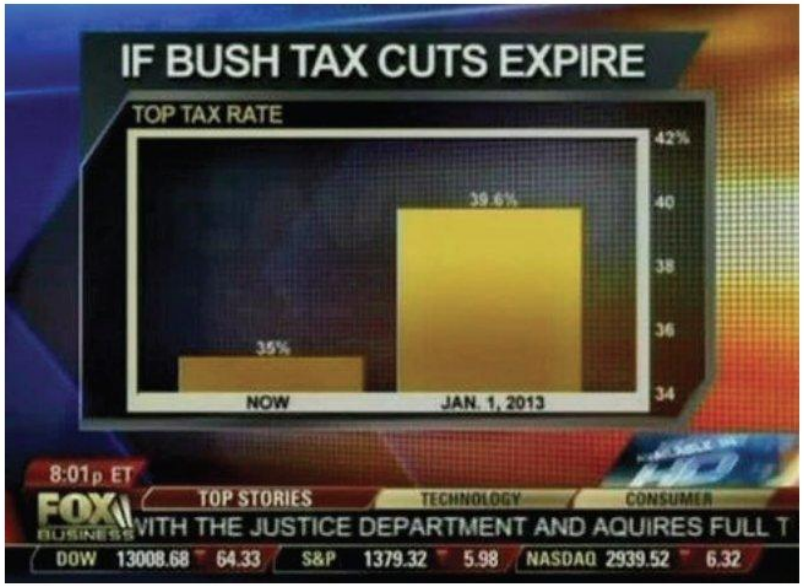


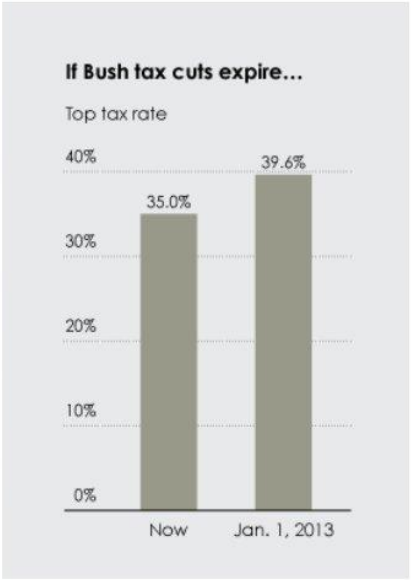
Graphical Integrity: Scale Distortion

Always start bar graphs at zero.

Always properly label your axes.

Use continuous scales: linear or labelled!



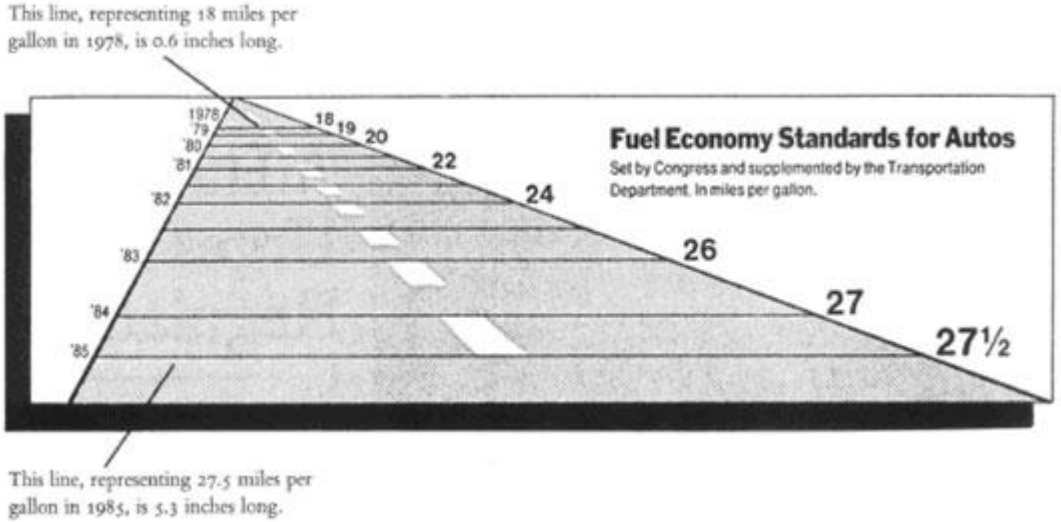


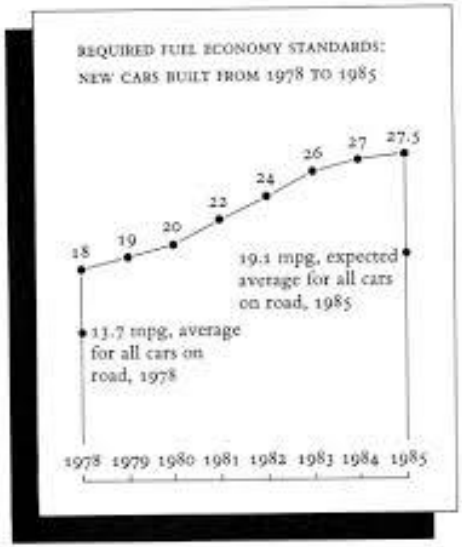
Aspect Ratios and Lie Factors

The steepness of apparent cliffs is a function of aspect ratio.

Aim for 45° lines or Golden ratio as most interpretable.

Can this be the Same Data?





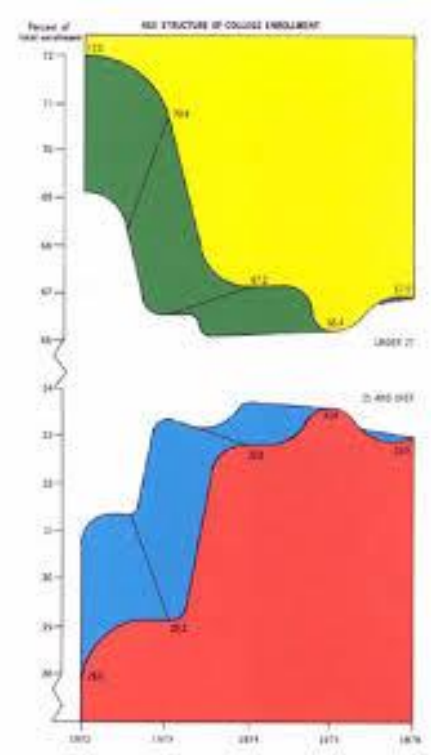
Reduce Chartjunk

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

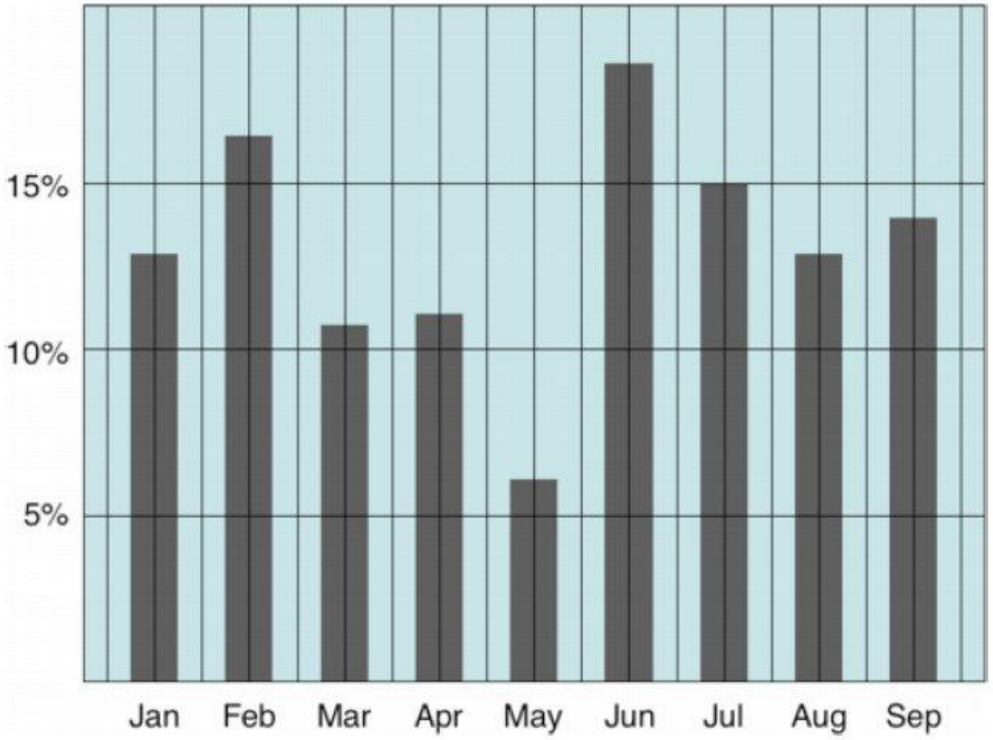
- Extra dimensionality
- Uninformative coloring
- Excessive grids and figurative decoration

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

Chartjunk / Graphical Ducks

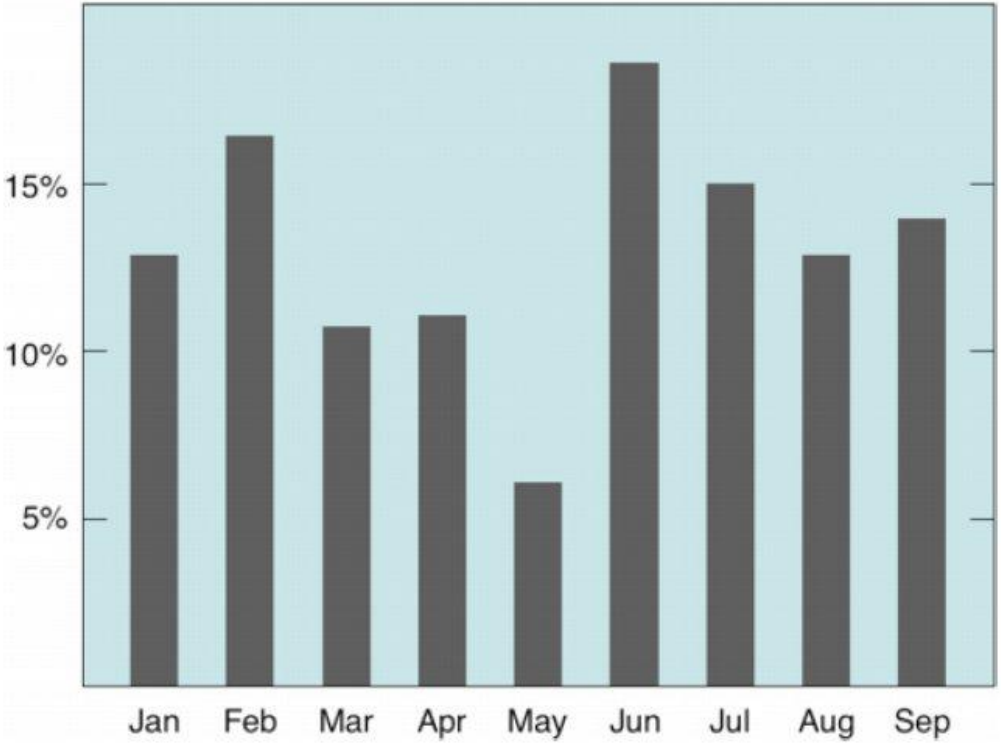


Can you Simplify this Plot?

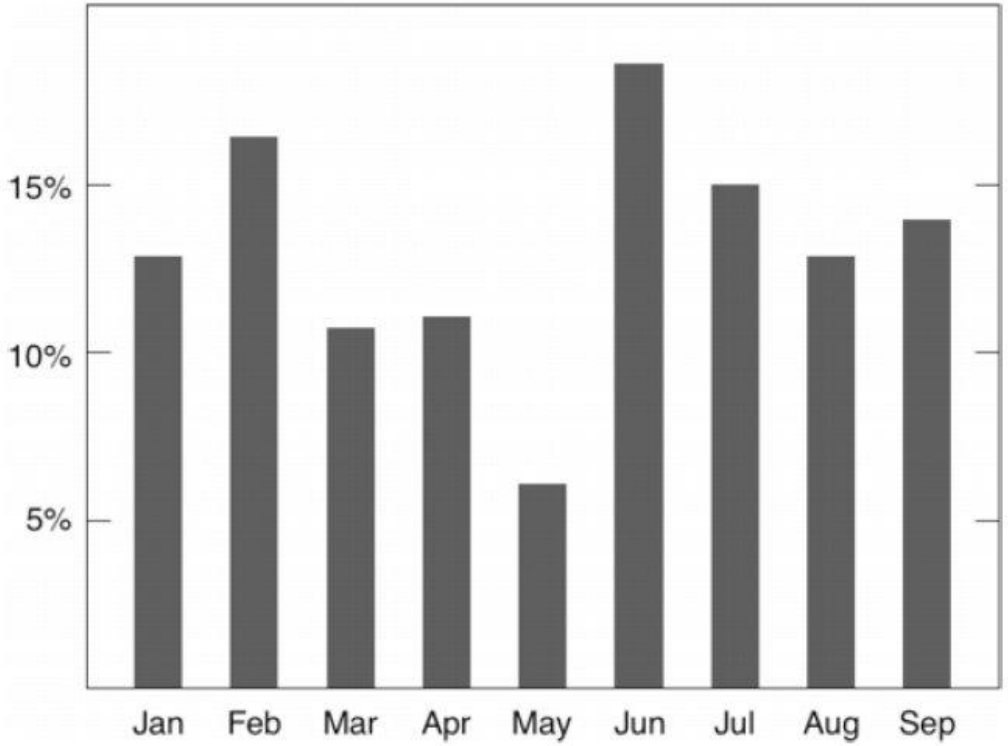


Tim Brey

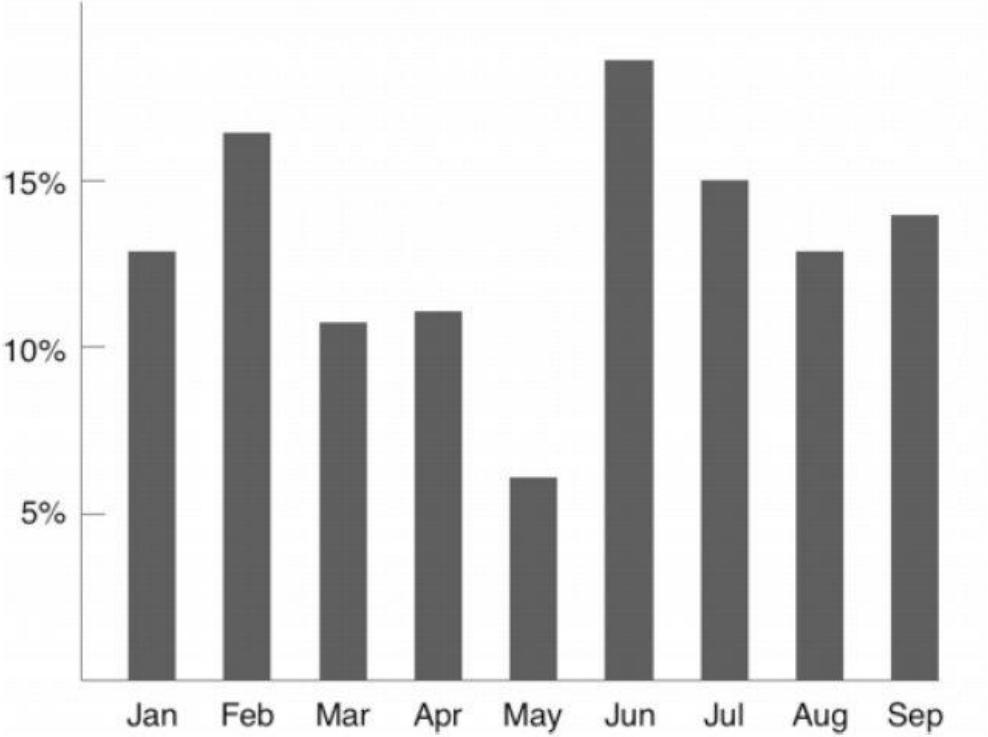
Can You Further Simplify?



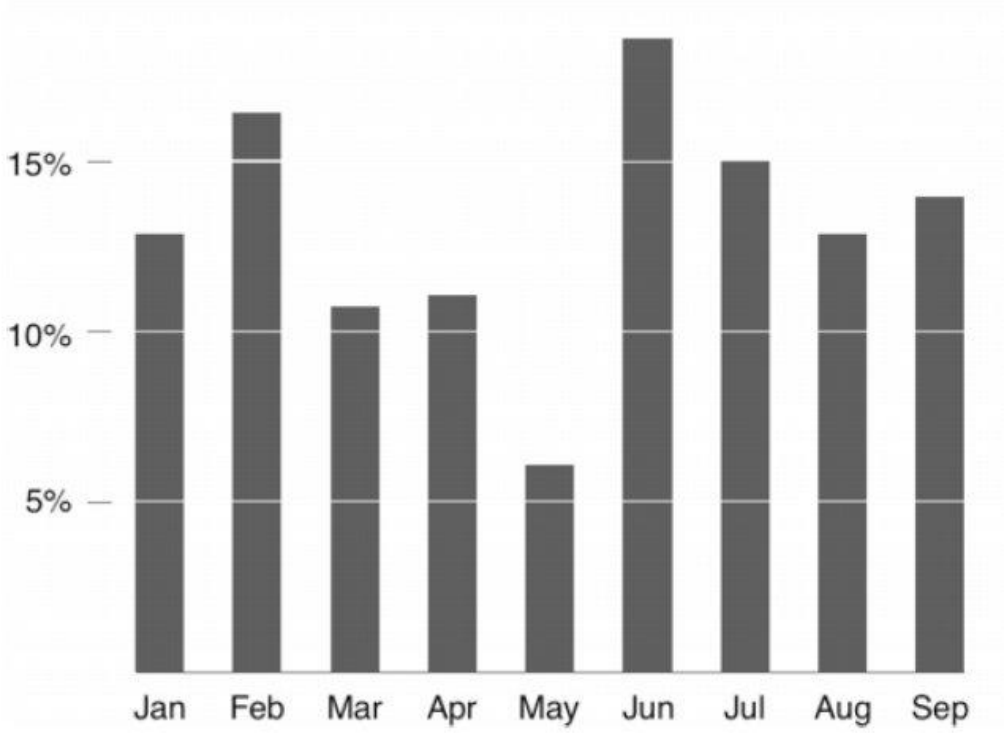
Better, but can you Further Simplify?



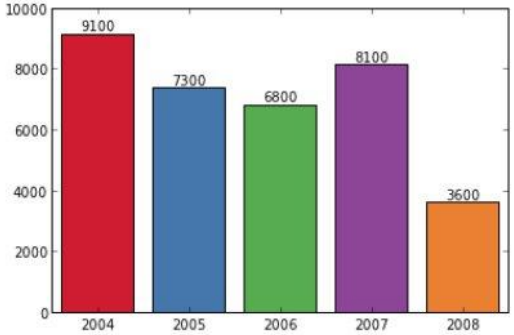
Anything Else that Can Go?

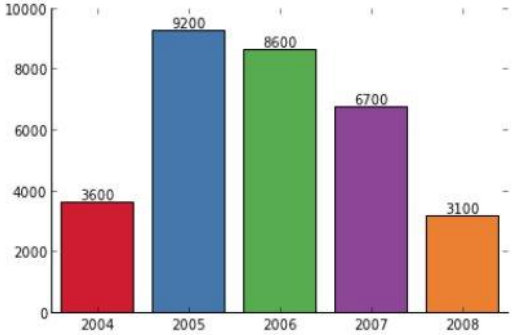


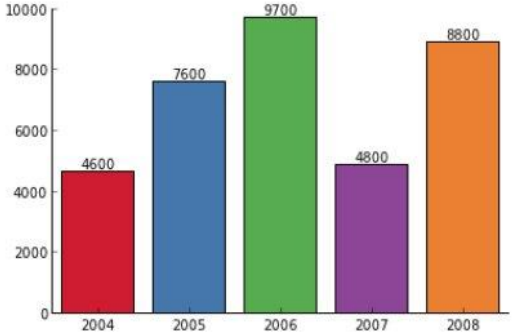
``Less is More”

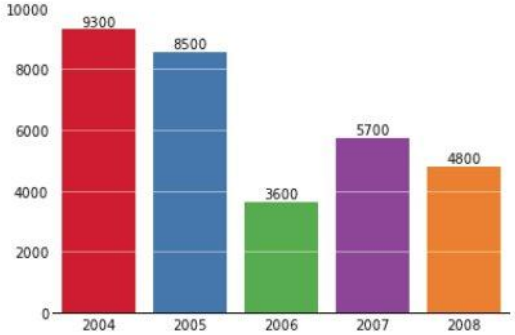


MatPlotLib Supports Nice Plots



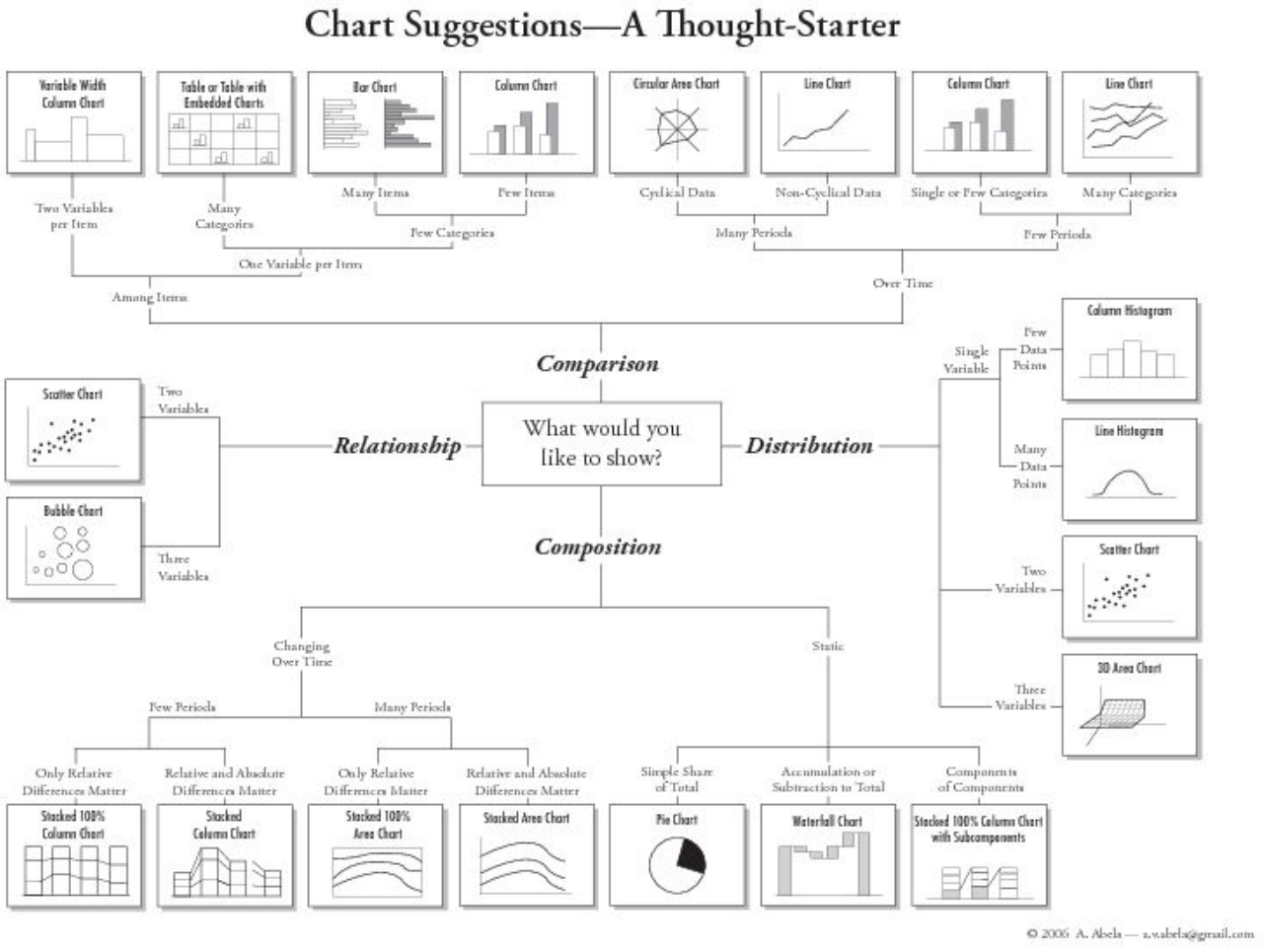






http://nbviewer.ipython.org/5357268

Which Chart to Use?



Tabular Data

Tables can have advantages over plots:

- Representation of numerical precision
- Understandable multivariate visualization: each column is a different dimension.
- Representation of heterogeneous data
- Compactness for small numbers of points.

Can this Table be Improved?

Country	Area	Density	Birthrate	Population	Mortality	GDP
Russia	17075200	8.37	99.6	142893540	15.39	8900.0
Mexico	1972550	54.47	92.2	107449525	20.91	9000.0
Japan	377835	337.35	99.0	127463611	3.26	28200.0
United Kingdom	244820	247.57	99.0	60609153	5.16	27700.0
New Zealand	268680	15.17	99.0	4076140	5.85	21600.0
Afghanistan	647500	47.96	36.0	31056997	163.07	700.0
Israel	20770	305.83	95.4	6352117	7.03	19800.0
United States	9631420	30.99	97.0	298444215	6.5	37800.0
China	9596960	136.92	90.9	1313973713	24.18	5000.0
Tajikistan	143100	51.16	99.4	7320815	110.76	1000.0
Burma	678500	69.83	85.3	47382633	67.24	1800.0
Tanzania	945087	39.62	78.2	37445392	98.54	600.0
Tonga	748	153.33	98.5	114689	12.62	2200.0
Germany	357021	230.86	99.0	82422299	4.16	27600.0
Australia	7686850	2.64	100.0	20264082	4.69	29000.0

Dimensions for Improvement

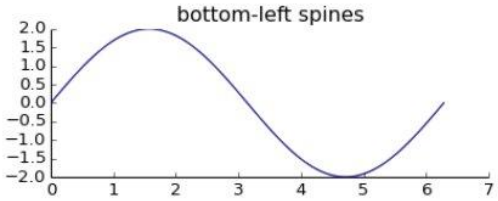
- Order rows to invite comparisons.
- Order columns to highlight importance or pairwise relationships.
- Right justify uniform-precision numbers
- Use **emphasis**, *font,* or color to highlight important entries.
- Avoid excessive-length column descriptors.

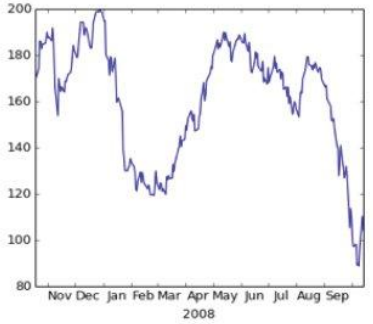
Improved Tabular Presentation

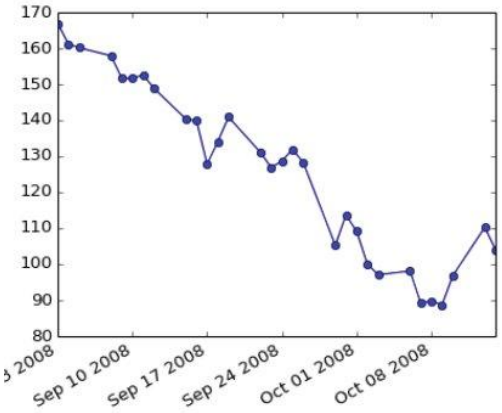
Country	Population	Area	Density	Mortality	GDP	Birth Rate
Afghanistan	31,056,997	647,500	47.96	**163.07**	700	36.0
Australia	20,264,082	7,686,850	2.64	4.69	29,000	**100.0**
Burma	47,382,633	678,500	69.83	67.24	1,800	85.3
China	**1,313,973,713**	9,596,960	136.92	24.18	5,000	90.9
Germany	82,422,299	357,021	230.86	4.16	27,600	99.0
Israel	6,352,117	20,770	305.83	7.03	19,800	95.4
Japan	127,463,611	377,835	**337.35**	3.26	28,200	99.0
Mexico	107,449,525	1,972,550	54.47	20.91	9,000	92.2
New Zealand	4,076,140	268,680	15.17	5.85	21,600	99.0
Russia	142,893,540	**17,075,200**	8.37	15.39	8,900	99.6
Tajikistan	7,320,815	143,100	51.16	110.76	1,000	99.4
Tanzania	37,445,392	945,087	39.62	98.54	600	78.2
Tonga	114,689	748	153.33	12.62	2,200	98.5
United Kingdom	60,609,153	244,820	247.57	5.16	27,700	99.0
United States	298,444,215	9,631,420	30.99	6.50	**37,800**	97.0

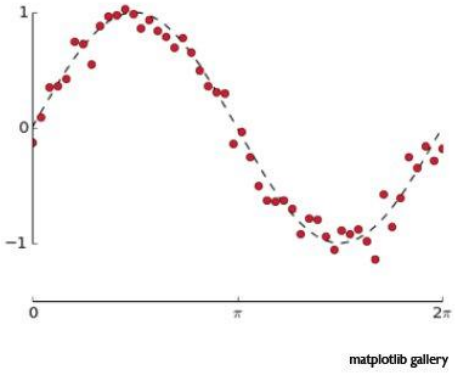
Line Charts

- Show data points, not just fits.
- Line segments show connections, so do not use in categorical data.
- Connecting points by lines is often chartjunk. Better is usually a trend line or fit with the data points.







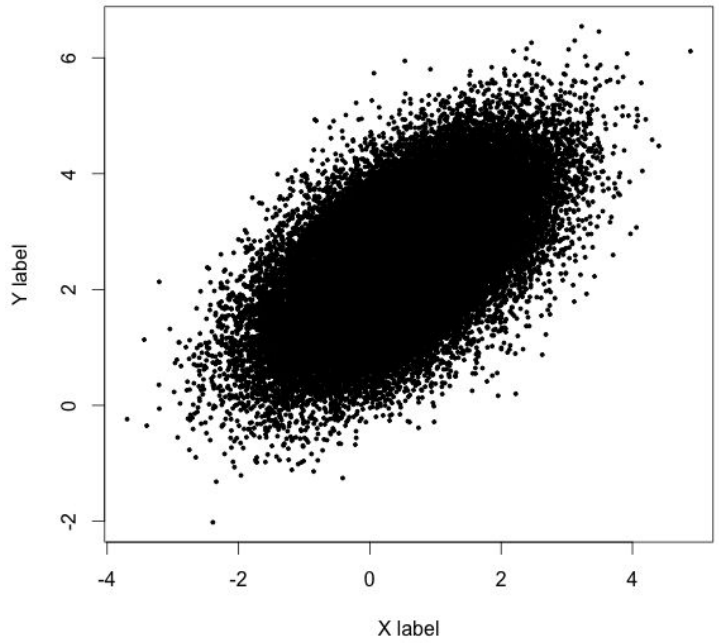


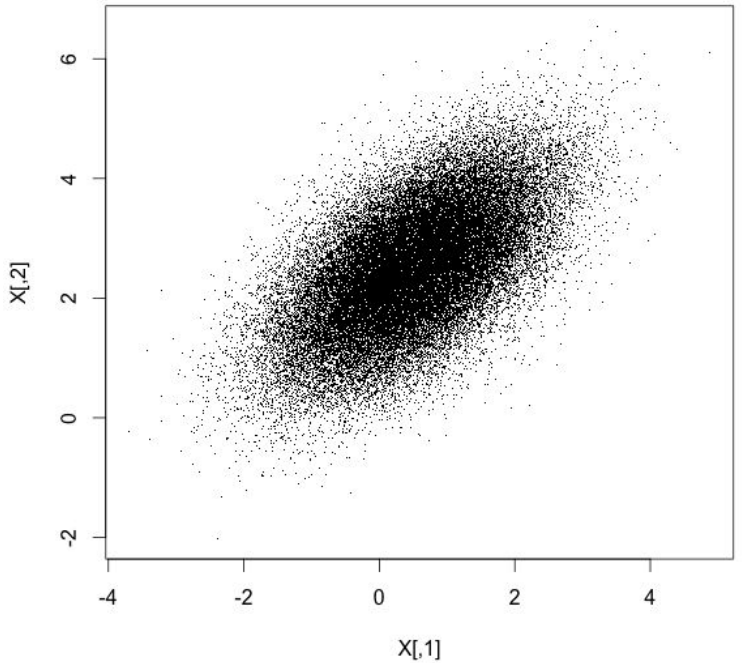
Scatter Plots / Multivariate Data

Scatter plots show the values of each point, and are a great way to present 2D data sets.

Higher dimensional datasets are often best projected to 2D, through self-organizing maps or principle component analysis, although can be represented through bubble plots.

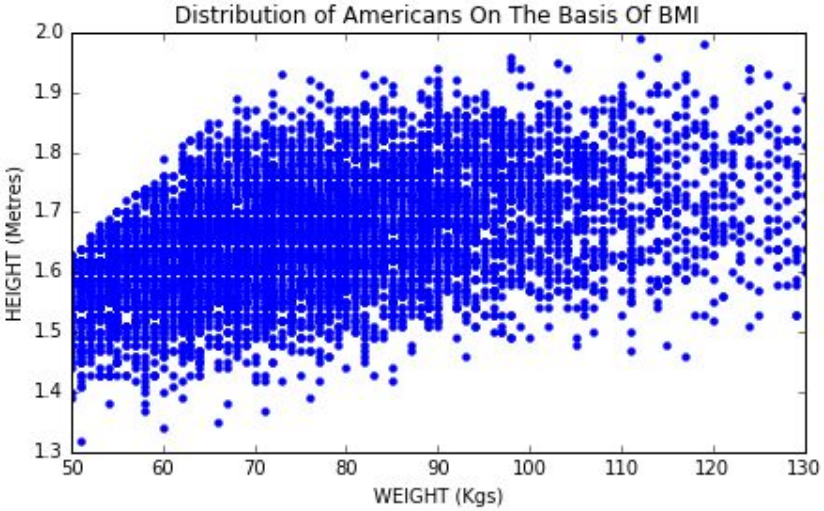
Reduce Overplotting by Small Points

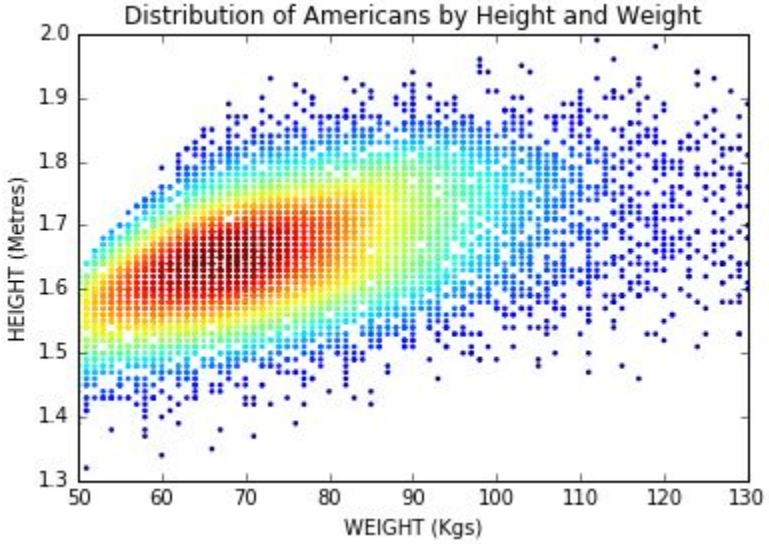




Heatmaps Reveal Finer Structure

Color points on the basis of frequency

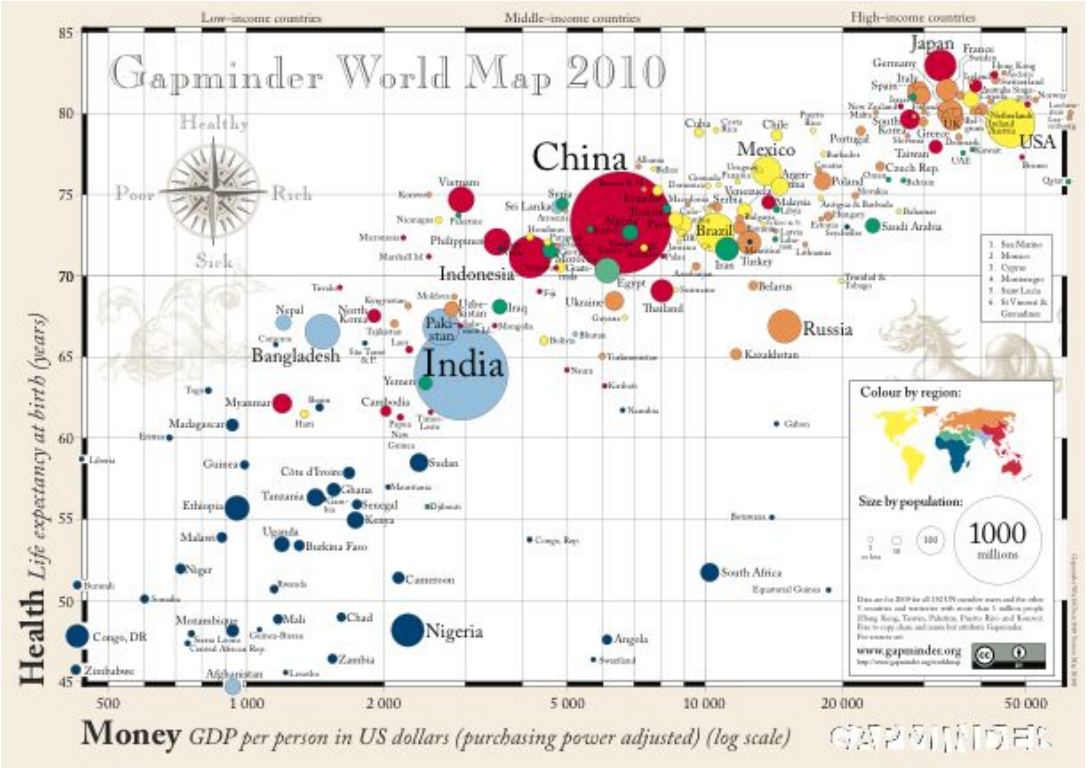




Bubble Charts for Extra Dimensions

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

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

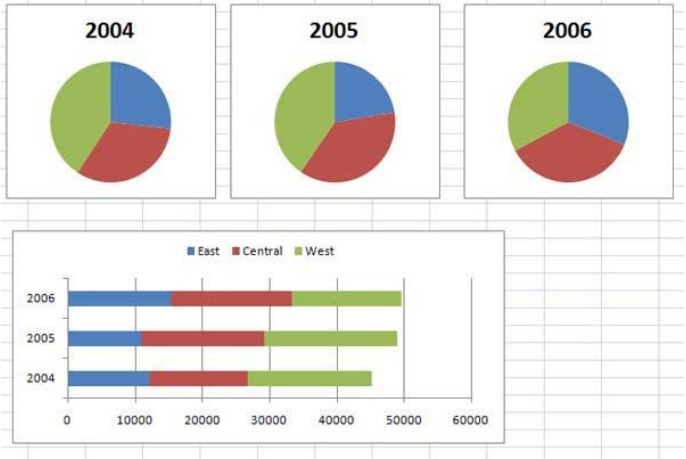


Bar Plots vs. Pie Charts

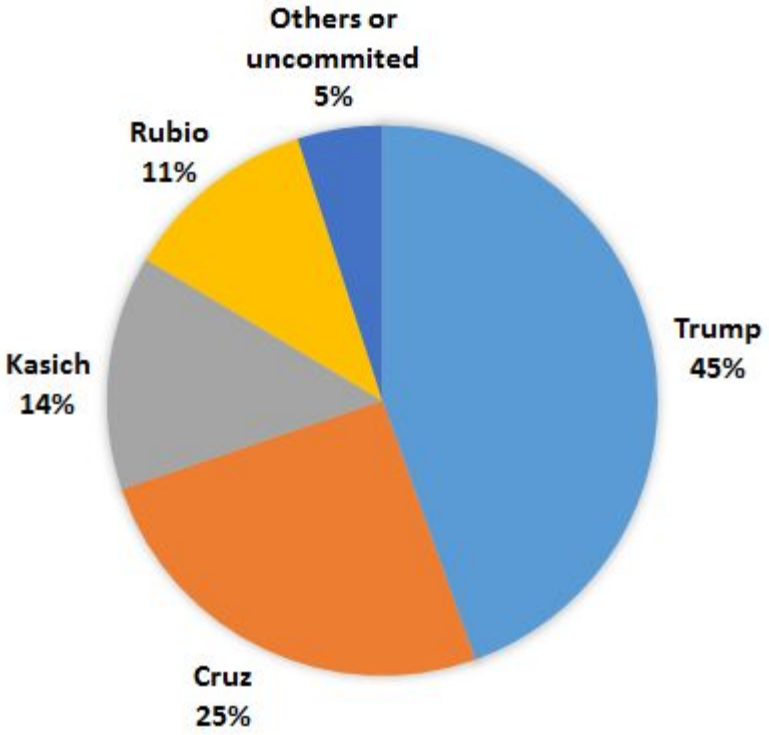
Bar plots show the frequency of proportion of categorical variables. Pie charts use more space and are harder to read and compare.

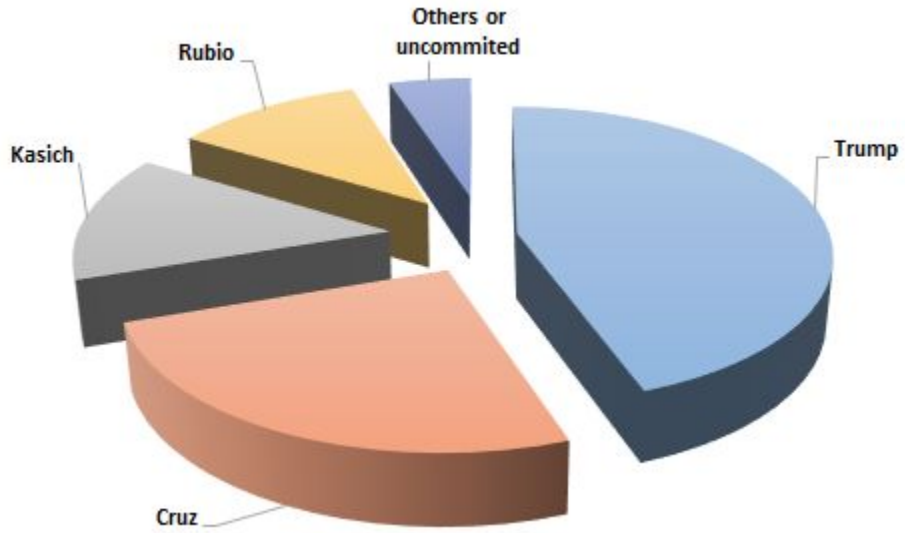
Partitioning each bar into pieces yields the stacked bar chart.

Pie charts are arguably better for showing percentages of totality, and people do seem to like them, so they may be harmless in small amounts.



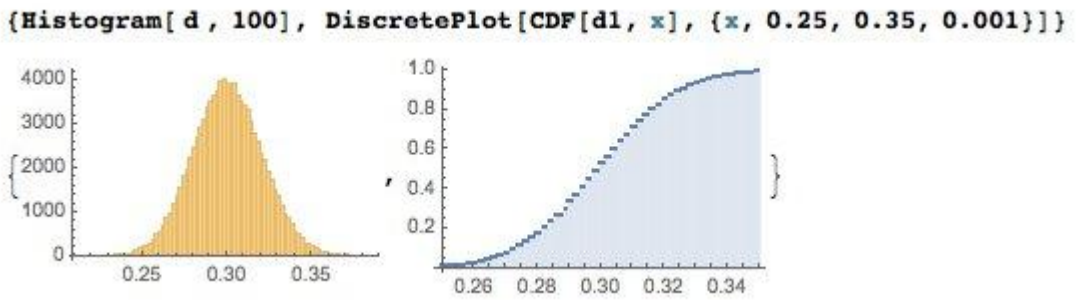
Which Pie Chart is Better?





Histograms

Histograms (and CDFs) visualize distributions over continuous variables:

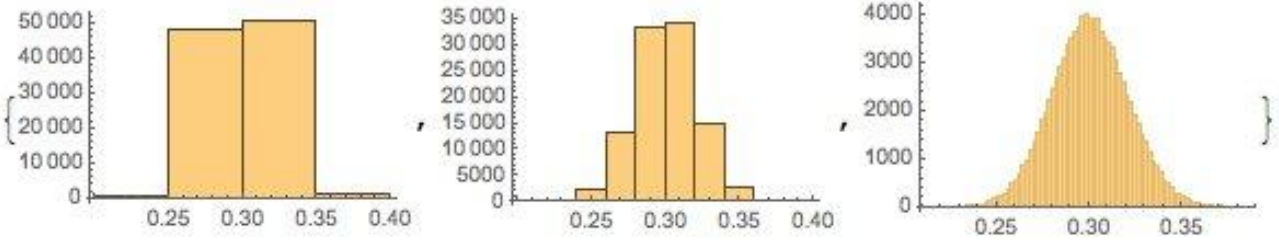


Histograms are better for displaying peaks, CDFs for showing tails.

Histograms: Bin Size / Count Matters

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

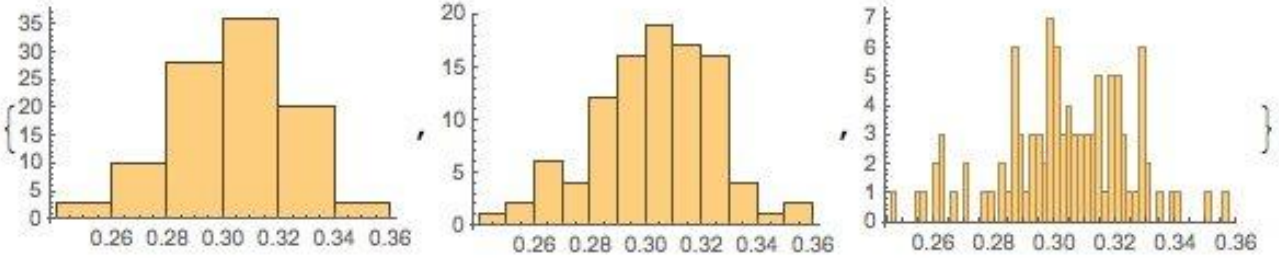
Out[62]=




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

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

Out[65]=



Frequency vs. Density Histograms

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