Data Analysis, Statistics, Machine Learning

Leland Wilkinson

Adjunct Professor
UIC Computer Science
Chief Scientist
H2O.ai

leland.wilkinson@gmail.com

Visualizing

- Visualizations represent data
  - Tallies, stem-and-leaf plots, pie charts, bar charts, …
- Statistical visualizations represent variables
  - Histograms, probability plots, density plots, …
- Statistical visualizations aid diagnosis of models
  - Does a variable derive from a given distribution?
  - Are there outliers and other anomalies?
  - Are there trends (or periodicity, etc.) across time?
  - Are there relationships between variables?
  - Are there clusters of points (cases)?

Visualizing

- Visualizations are diverse
  - Diagrams, charts, maps, graphs, images, scenes, …
- Every visualization rests on a model
  - How does a human (animal, robot) process a visualization?
  - Processing begins with perception and ends with understanding
  - A visualization model specifies the rules for parsing a visualization
    - Semiotic (Bertin)
    - Ecological (Gibson)
    - Algebraic (Wilkinson)
    - Connectionist (Pavlov)
  - Analytic visualizations
    - Visualizations for statistics and ML
    - These visualizations rest on graphs of functions

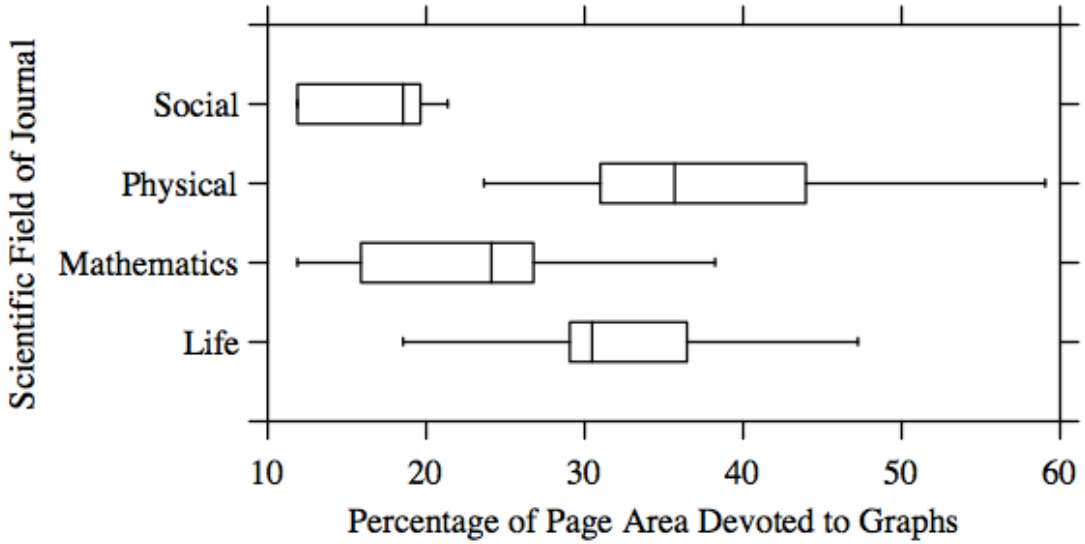
Visualizing

Sad, isn't this?

Cleveland, W.S. (1984). Graphs in scientific publications. *The American Statistician*, *38*, 261-269.

The hard sciences use more graphics than the soft!

Maybe social scientists think publishing numbers is more "scientific" than publishing graphs



Visualizing

- Data visualizations
  - Counts and densities
    - Tallies
    - Dot plots
    - Stem and leaf plots
    - Bar charts
    - Line/area charts
  - Proportions
    - Divided bar charts
    - Pie charts
  - Batches
    - Jittered 1D scatterplots
    - Stripe plots
    - Box plots

Visualizing

Data visualizations

Tallies (from Latin *talea*, a stick)

- Tally sticks have ancient origins, used for counting
- Scale may be categorical or continuous

28.0 to 28.9	II	2
29.0 to 29.9	𝍸	5
30.0 to 30.9	𝍸 II	7
31.0 to 31.9	𝍸 𝍸 𝍸	15
32.0 to 32.9	𝍸 𝍸 𝍸 𝍸	20
33.0 to 33.9	𝍸 𝍸 III	13
34.0 to 34.9	𝍸 III	8
35.0 to 35.9	III	3
36.0 to 36.9	II	2

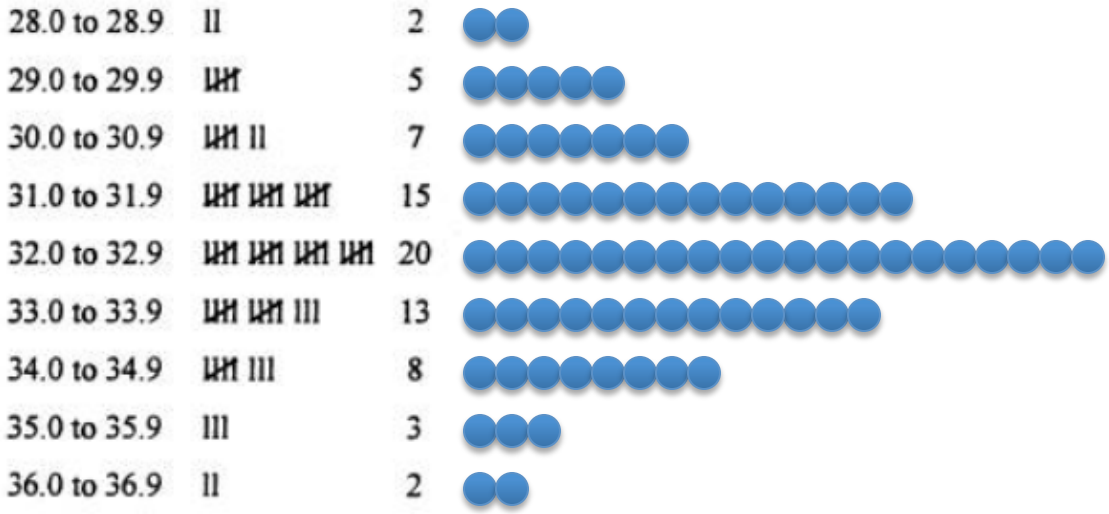
Transportation Research Board, National Academy of Sciences

Visualizing

Data visualizations

Dot plots

- Scale may be categorical or continuous
- One dot per observation
- A count scale can help
- For large batches, dot sizes can be reduced or each dot can represent many cases

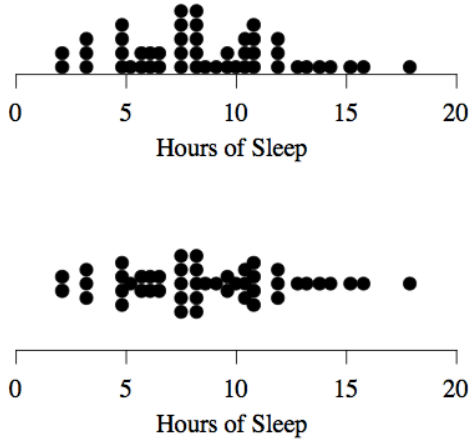


Visualizing

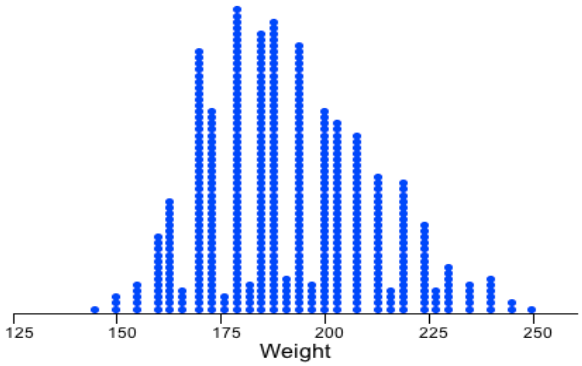
Data visualizations

Dot plots

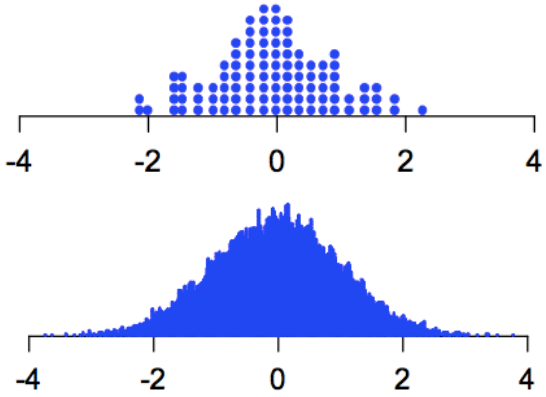
Symmetric or asymmetric versions



Reveal granularity (unlike histograms)



Small (100) or large (10,000)



Visualizing

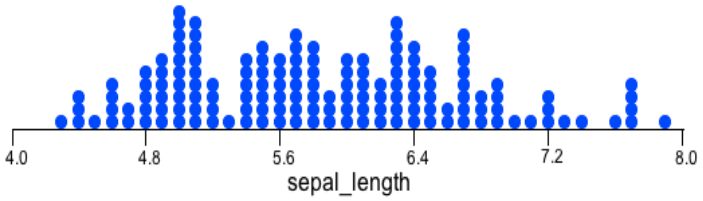
Data visualizations

How to make a dot plot

Wilkinson, L. (1999). Dot plots. *The American Statistician, 53,* 276–281.

A dot plot is not a histogram of dots

1) Start with the smallest data value, $x_j = x_1$. The first stack of dots always begins here.
2) Count the number of values, n_j, within one dot's width (h) to the right of x_j.
3) Place n_j dots above x_j, or offset to the right of x_j by the average of the n_j values if the n_j data values differ.
4) Move right to the next largest data value not included in the current stack of dots.
5) Repeat steps 2-4 until there are no data values left to plot.



Visualizing

Data visualizations

Dot plots can sometimes be appropriate in 3D

Dang, T.N., Wilkinson, L., and Anand , A. (2010). Stacking Graphic Elements to Avoid Over-plotting. *IEEE Transactions on Visualization and Computer Graphics 16*, 1044-1052.

3D dot plot showing enormous range in density not as well represented in heatmap

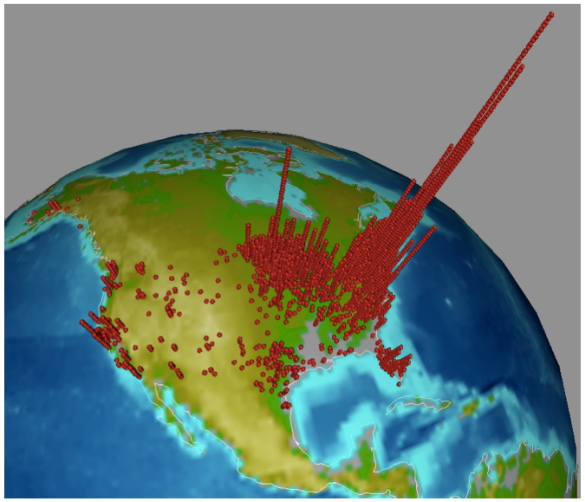
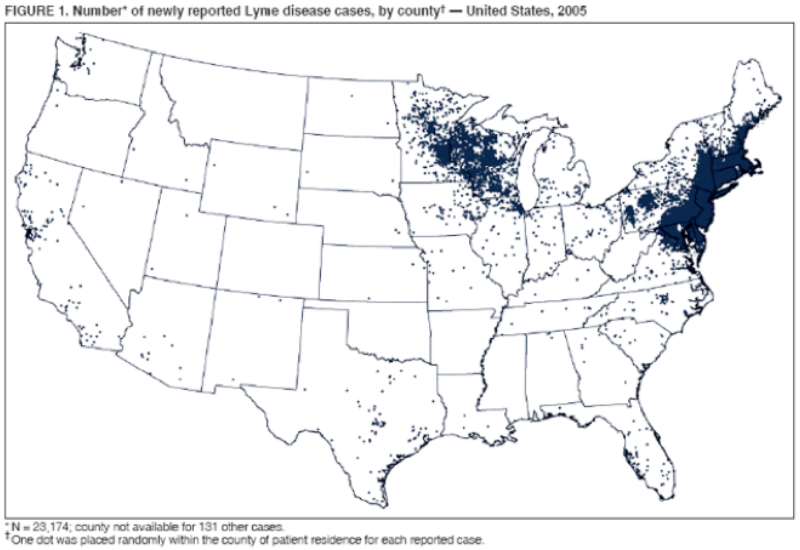


Fig. 8. Dot plot presenting Lyme disease from 2003 to 2005.

Same data from CDC map



Visualizing

Data visualizations (counts)

Stem and leaf plots

- Scale may be categorical or continuous
- Tukey invented these
- Instead of a dot or tally stick, we use next digit for placeholder

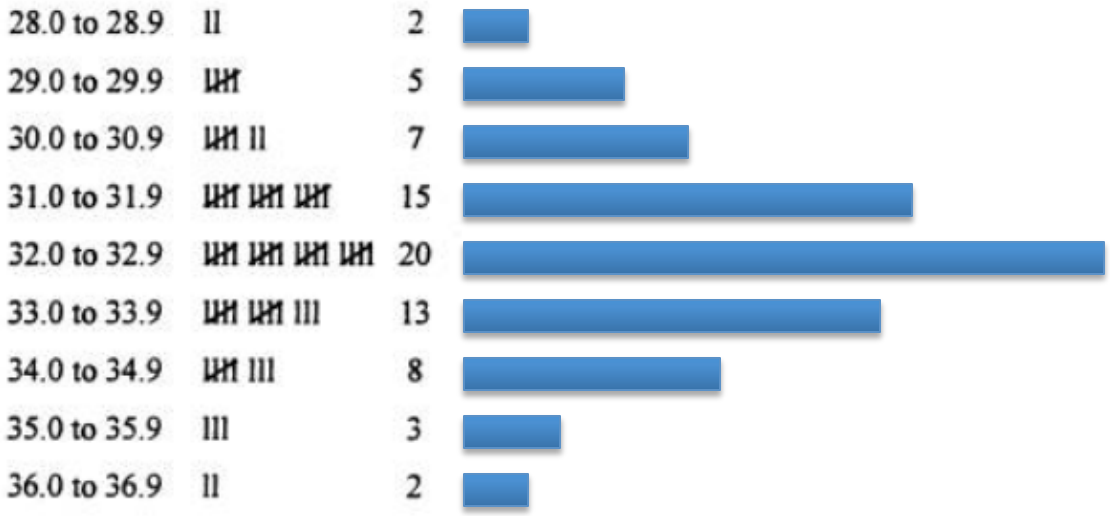
28.0 to 28.9	II	2	28 36
29.0 to 29.9	𝍸	5	29 13378
30.0 to 30.9	𝍸 II	7	30 2224589
31.0 to 31.9	𝍸 𝍸 𝍸	15	31 111224666677889
32.0 to 32.9	𝍸 𝍸 𝍸 𝍸	20	32 00112223335667888999
33.0 to 33.9	𝍸 𝍸 III	13	33 0222344466889
34.0 to 34.9	𝍸 III	8	34 11344668
35.0 to 35.9	III	3	35 333
36.0 to 36.9	II	2	36 48

Visualizing

Data visualizations (counts)

Bar charts

Scale may be categorical or continuous

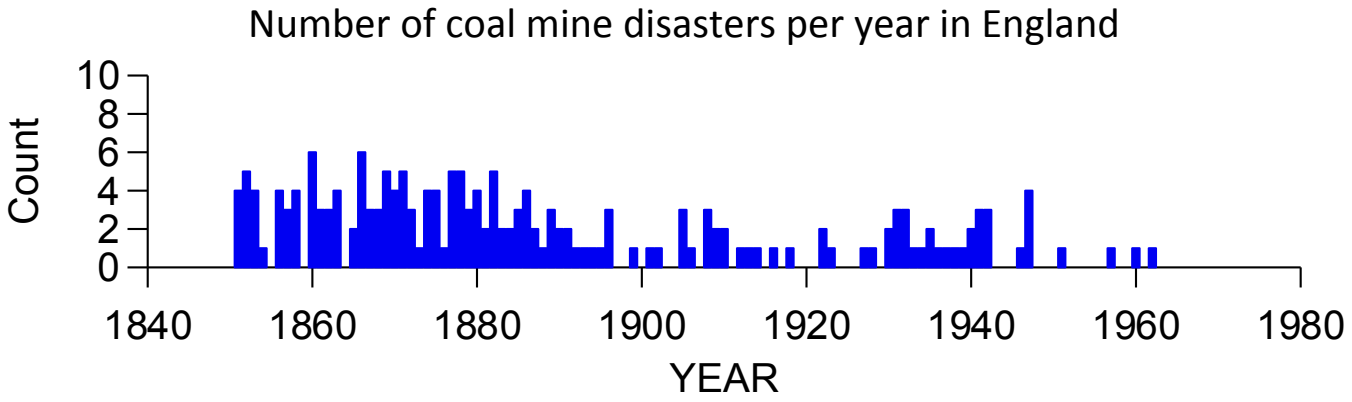


Visualizing

Data visualizations

Bar chart on continuous scale

- There is no justification for restricting bar graphs to categorical scales
- Excel and other software enforces this restriction unnecessarily
- If bars overlap, that's a fair representation of the data
- Bar graphics on time scales should be continuous, just like dot plots
- They should allow for unevenly spaced time points



Visualizing

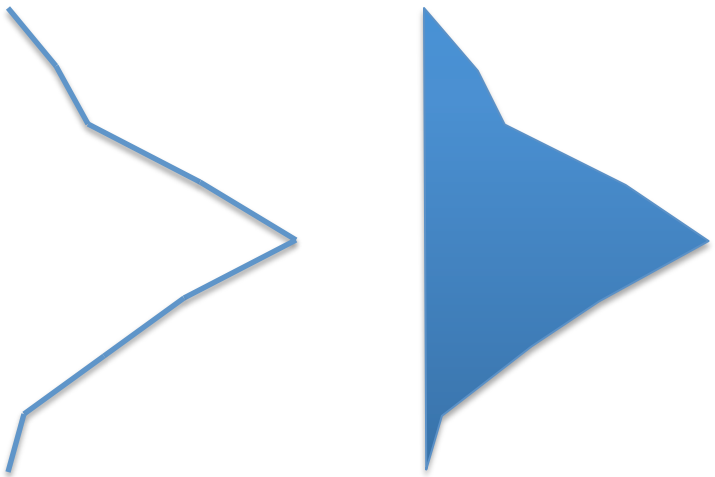
Data visualizations (counts)

Line/area charts

Nonsensical if scale is categorical

- Because slopes of lines are a function of category spacing rather than interval scale
- Unfortunately, line charts on categorical scales are very popular
- Better to use bars

28.0 to 28.9	II	2
29.0 to 29.9	𝍸	5
30.0 to 30.9	𝍸 II	7
31.0 to 31.9	𝍸 𝍸 𝍸	15
32.0 to 32.9	𝍸 𝍸 𝍸 𝍸	20
33.0 to 33.9	𝍸 𝍸 III	13
34.0 to 34.9	𝍸 III	8
35.0 to 35.9	III	3
36.0 to 36.9	II	2

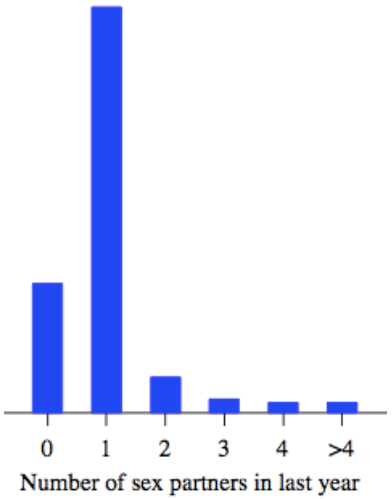


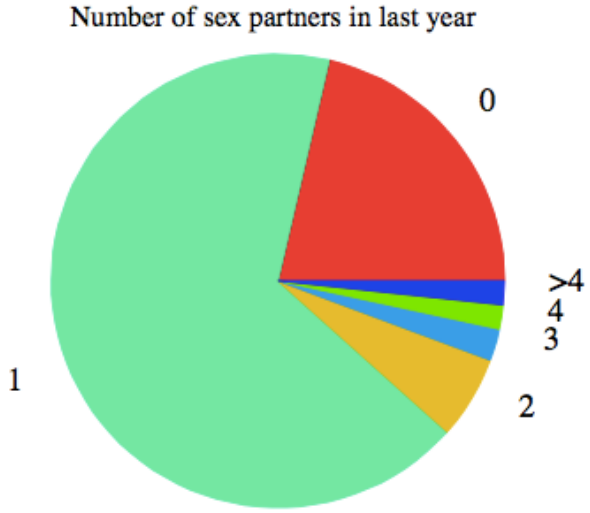
Visualizing

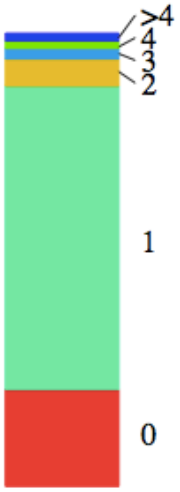
Data visualizations (proportions)

Bar chart, pie chart and divided bar chart

- Pie better than ordinary bars for proportion of whole data
- Divided bars usually less effective than pies or bars
- Too many slices bad news for pies







Visualizing

Pie charts

Viz gurus who claim pie charts are always bad are ignorant

The research says otherwise

Simkin, D., & Hastie, R. (1987). An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association, 82*(398), 454

Hollands, J.G., and Spence, I. (1998). Judging proportion with graphs: The summation model. *Applied Cognitive Psychology, 12*, 173-190.

Lewandowsky and Spence (1998). The perception of statistical graphs. *Sociological Methods and Research, 18*. 200-242.

Spence, Ian (1990). Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 683-692.

Spence, Ian, and Lewandowsky, Stephan (1991). Displaying proportions and percentages. *Applied Cognitive Psychology, 5,* 61–77.

Spence, Ian (2005). No Humble Pie: The Origins and Usage of a Statistical Chart. *Journal of Educational and Behavioral Statistics,* 30, No. 4, pp. 353–368.

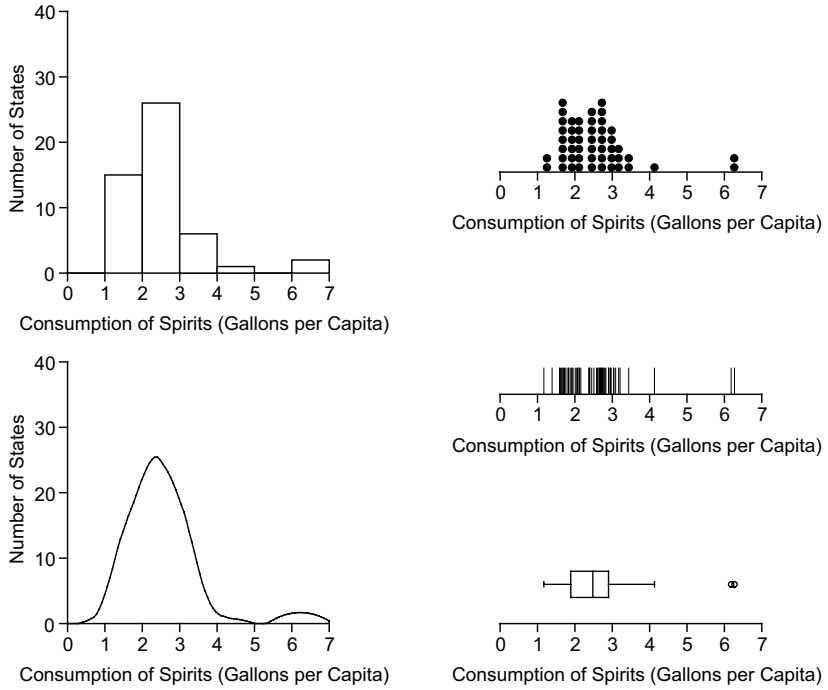
Visualizing

- Statistical visualizations
  - Density representations
    - Histograms
    - Density polygons
    - Dot plots
    - Kernel density estimates
  - Distribution assessments
    - Hanging histograms
    - Quantile plots
    - Probability plots
  - 2D statistical plots
    - Scatterplots
    - Enhancing scatterplots
    - Multigroup densities
  - 3D statistical plots

Visualizing

Statistical visualizations

Density representations



Visualizing

Statistical visualizations (density representations)

How to make a histogram (forget what your stat book told you)

n = sample size

k = number of bins

h = bin width

$k = ceil([max_x - min_x] / h)$

Or, $k = 3 + \log_2(n)\ \log_{10}(n)$ works pretty well, but there's lots of research on better formulas

h (or k) depends on n and distribution shape

$bin_j = [lower_j, upper_j)$ (this is a half-open interval)

bin_j contains all x_i such that $\{lowerj <= x_i < upper_j\}$

It's nice to have the bin limits coincide with tick marks

This makes it easier to describe the contents of a given bin

- Move optimal bin width to nearest scale (sub)interval

Bin widths do not have to be all the same

- It's area that counts, not the height of the bars
- The area must sum to 1 (or n)

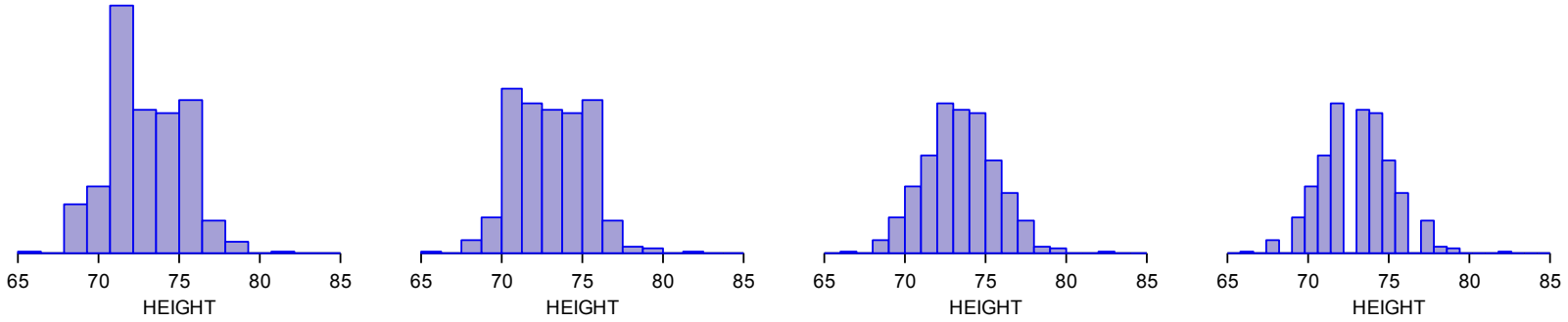
Simple, huh?

- Intro stat books make it look simple, but their algorithm is usually wrong or insufficient

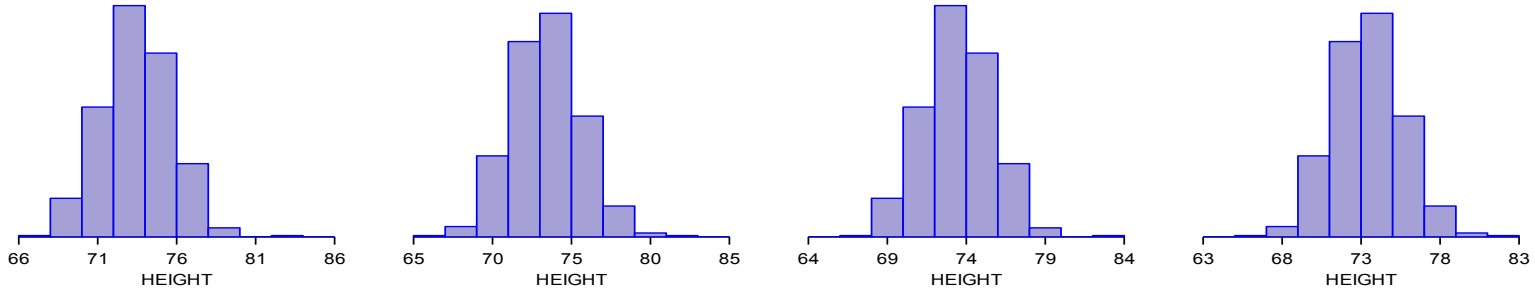
Visualizing

Statistical visualizations (density representations)

Histogram shape is sensitive to number of bars (different bar widths)



And to scale values (same bar width). That's scary.

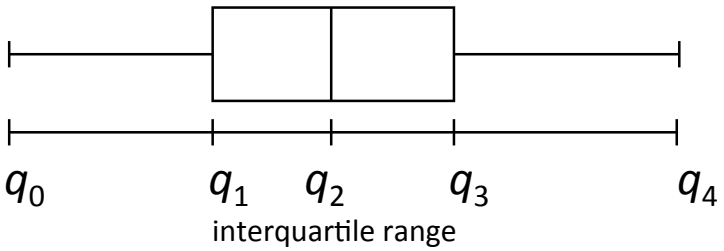


Visualizing

Statistical visualizations (density representations)

How to make a box plot (based on Tukey's letter values)

- Split the sorted batch to get the median
- Split each half batch and get its median (a quartile)
- This makes a 5 number summary
  - q_0 : minimum value (lower whisker limit)
  - q_1 : lower quartile (lower box edge)
  - q_2 : median (middle of box)
  - q_3 : upper quartile (upper box edge)
  - q_4 : maximum value (upper whisker limit)
- We'll discuss outliers in the Exploring lecture
  - They are based on an approximation to the normal distribution

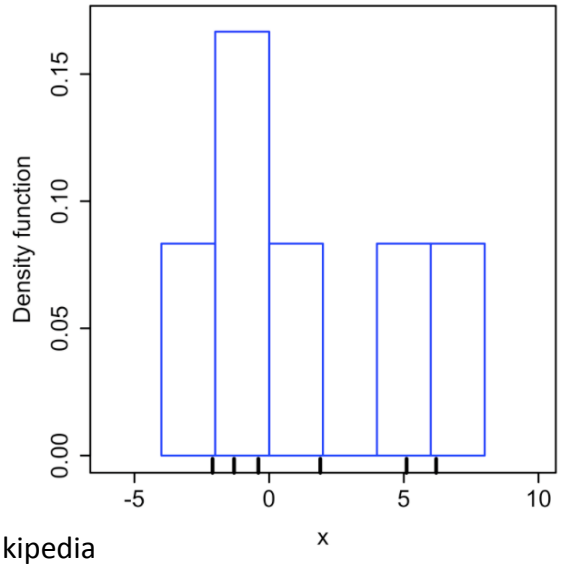


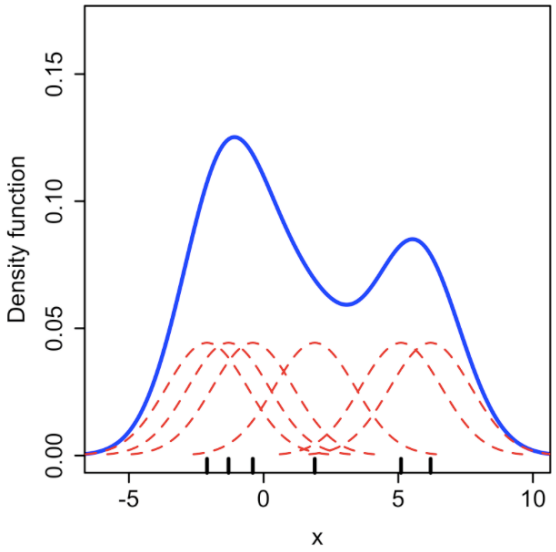
Visualizing

Statistical visualizations (density representations)

How to make a kernel density plot

- Choose a kernel function (red below)
- Choose the bandwidth (spread) for the kernel
- Evaluate it at each data point
- Sum kernels across domain





Wikipedia

Visualizing

Statistical visualizations

Hanging rootogram

Normalizes the counts to have equal variances (Freeman-Tukey variates)

Then the discrepancies can be compared between the tails and the center

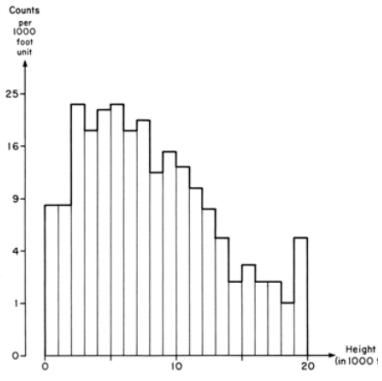


Fig. 18.16 Rootogram for heights of 218 volcanoes.

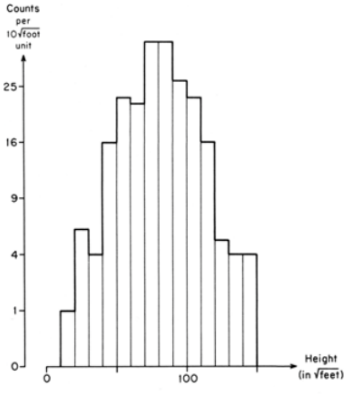


Fig. 18.17 Rootogram for $\sqrt{\text{height}}$ of 218 volcanoes.

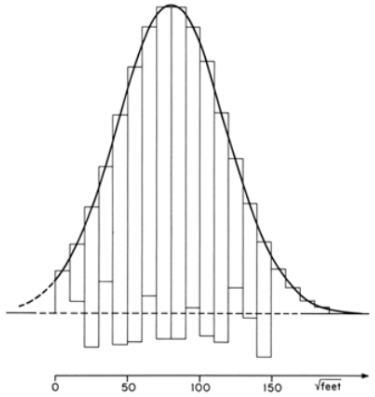


Fig. 18.18 A hanging rootogram for 218 volcanoes.

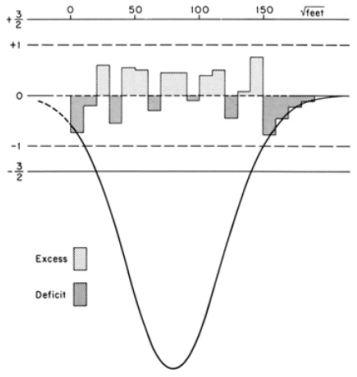


Fig. 18.19 Suspended histogram of the data in Figure 18.18.

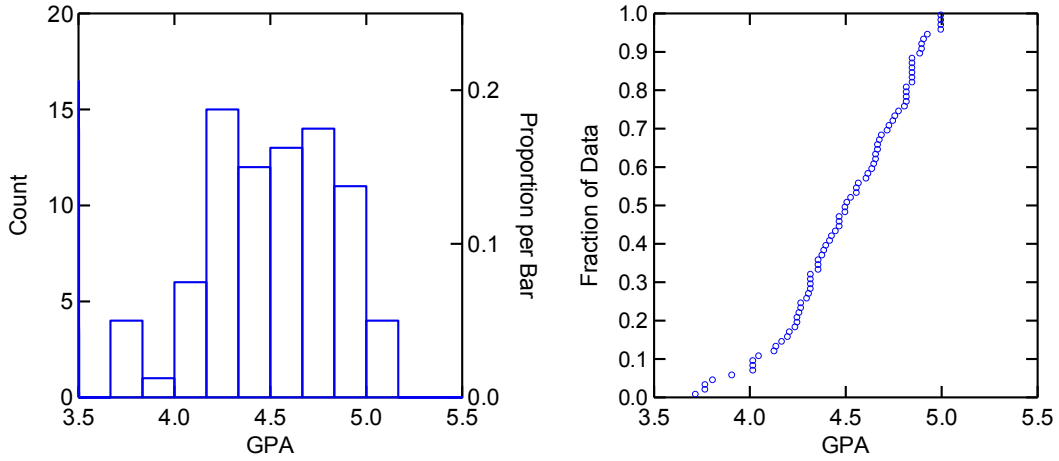
J.W. Tukey, Some Graphic and Semigraphic Displays, 1972

Visualizing

Statistical visualizations

Quantile Plot (QPLOT)

- Sort values and plot the fraction against the variable
- This is a sample cumulative distribution function (CDF)
- Best done alongside a histogram
  - Because it's hard to remember which is positive or negative skewness in the QPLOT

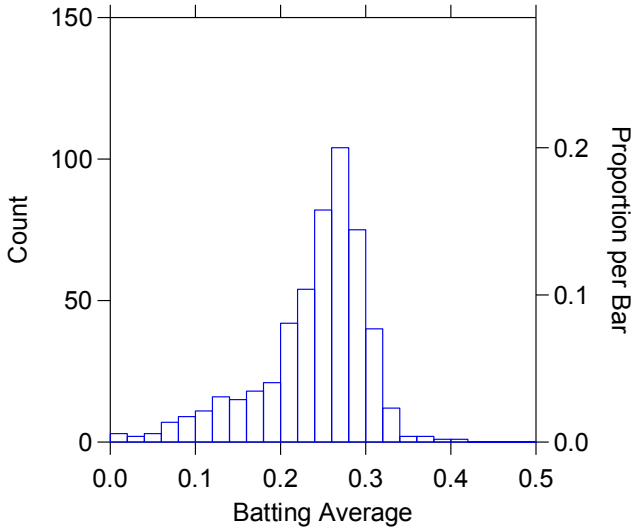


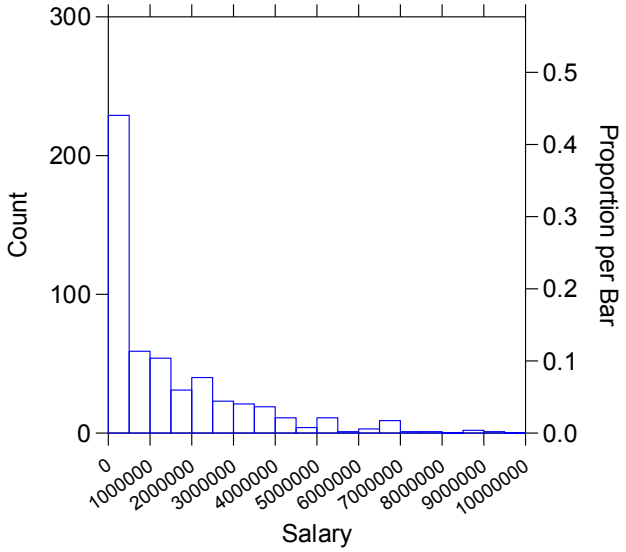
Visualizing

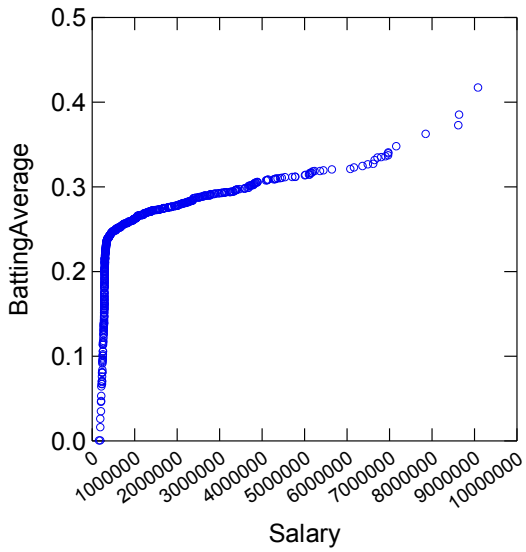
Statistical visualizations

Quantile-Quantile Plot (QQPLOT)

- Sort values on each variable
- Plot values against each other
- The more they depart from a diagonal line, the more different their shapes are







Visualizing

Statistical visualizations

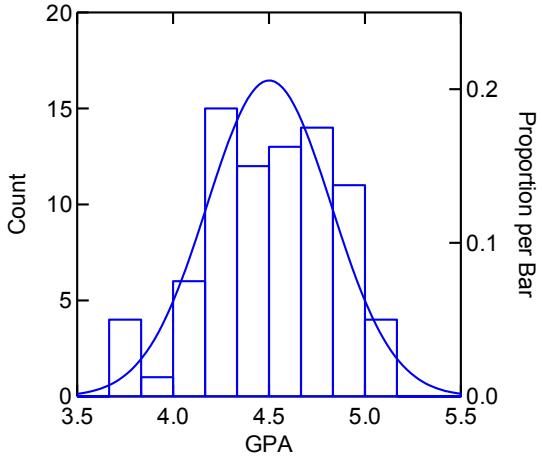
Probability Plot (PPLOT)

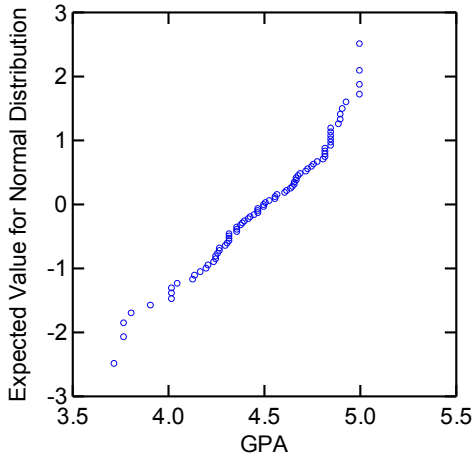
Same as QPLOT, but use a probability distribution as a reference

The more it departs from a diagonal line, the worse it fits the distribution

The example below is a normal probability plot

Again, best to do this alongside a histogram with the superimposed distribution

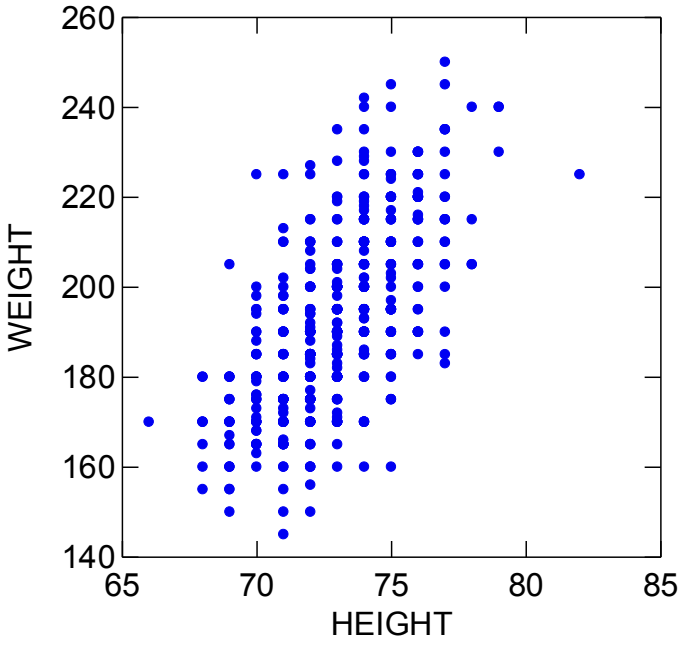


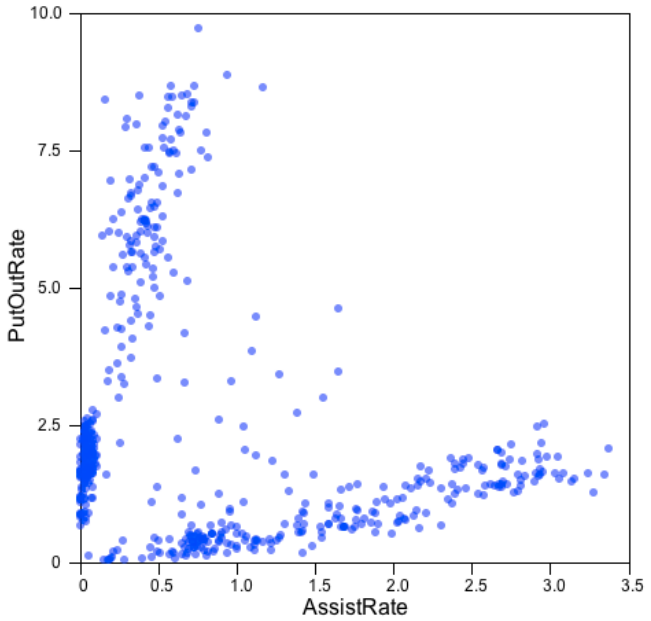


Visualizing

Scatterplot

The best way to show the relationship between two variables





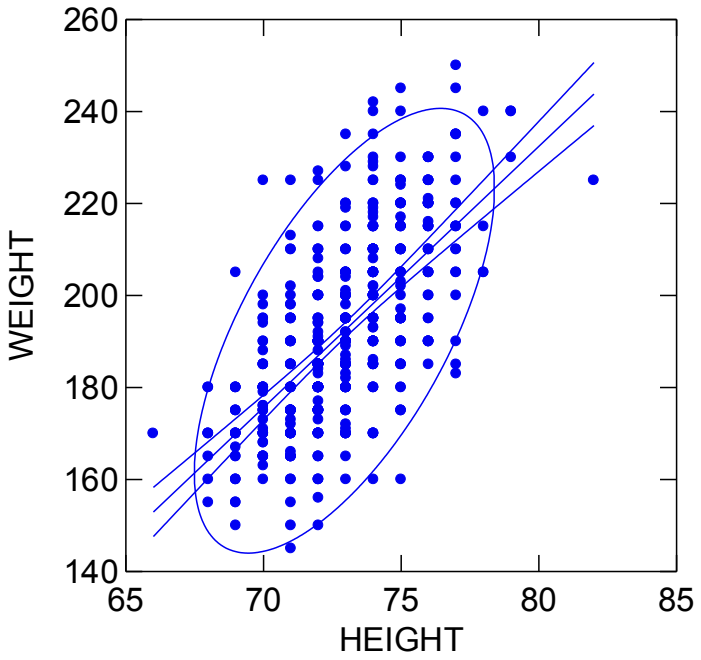
Visualizing

Enhancing Scatterplots

We can do all sorts of things, depending on the joint distribution

Here, we assume the distribution is bivariate normal

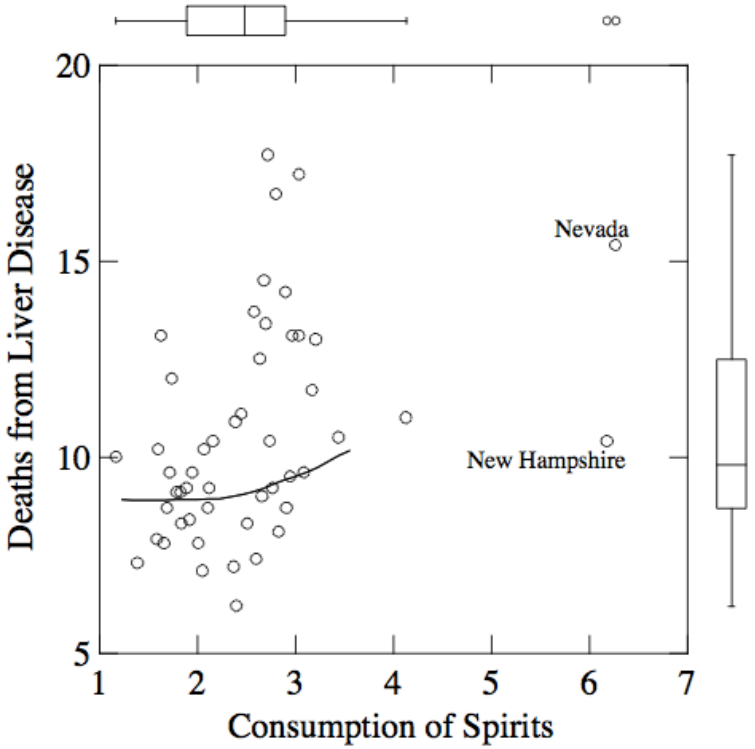
So we fit a joint confidence ellipse, a regression line, and confidence intervals on the line



Visualizing

Enhancing Scatterplots (bordering and smoothing)

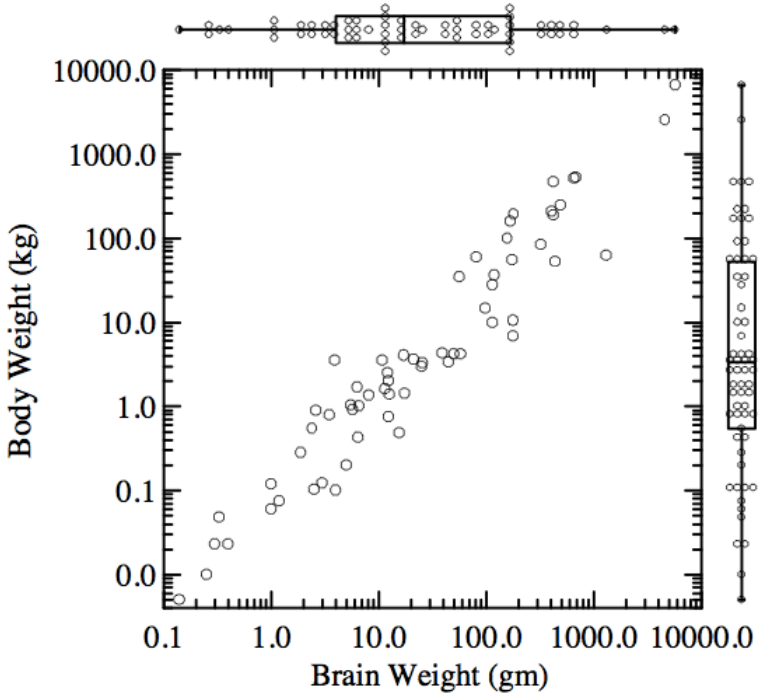
This conveys more information in the same space



Visualizing

Another bordered scatterplot

Dot-box plots are great for bordering

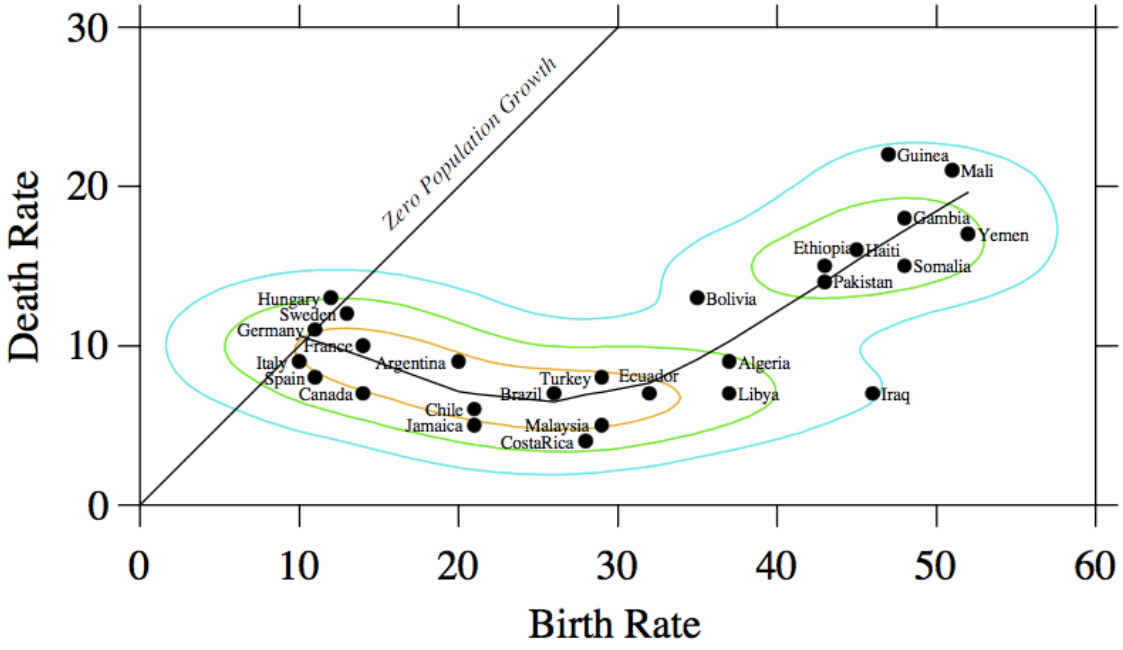


Visualizing

Enhancing Scatterplots (contouring density)

This graphic tells it all

Sometimes you don't need a statistical analysis

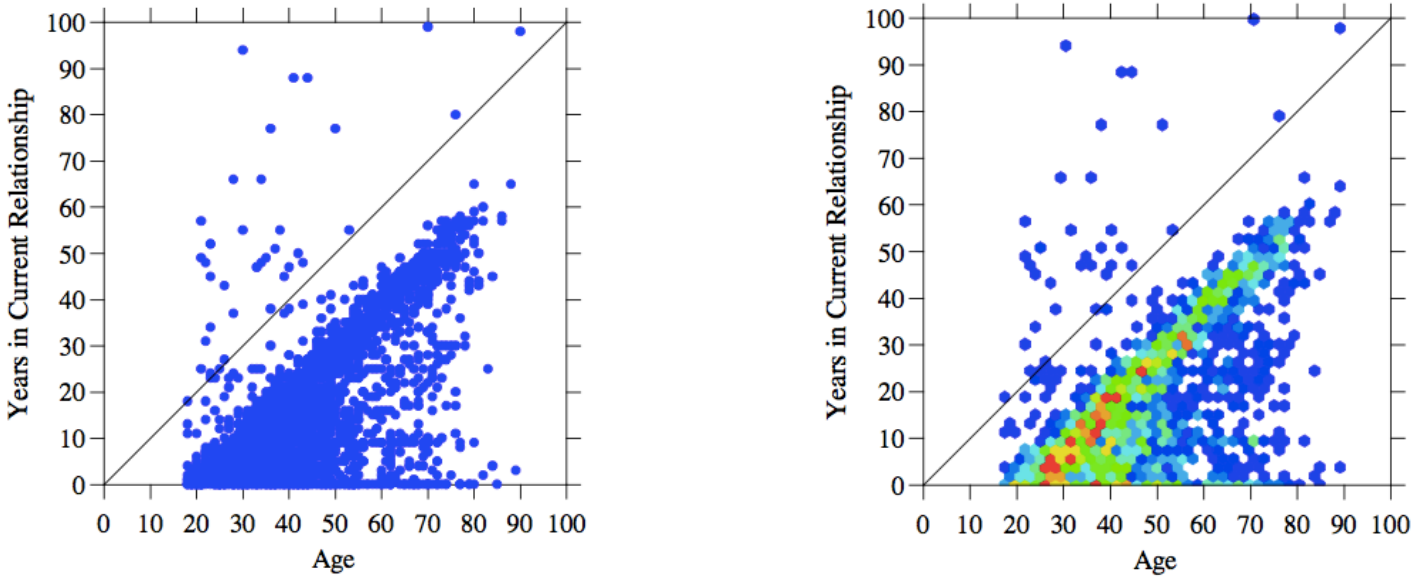


Visualizing

Enhancing Scatterplots by Hex Binning (Dan Carr)

Hex binning can handle huge datasets

- It helps reveal a density when points overlap in a conventional scatterplot (left)
- Color the hexagons using the number of cases in a bin
- Better than rectangular bins because it reduces Moiré patterns (anisotropy)



Visualizing

Enhancing Scatterplots by Hex Binning (Dan Carr)

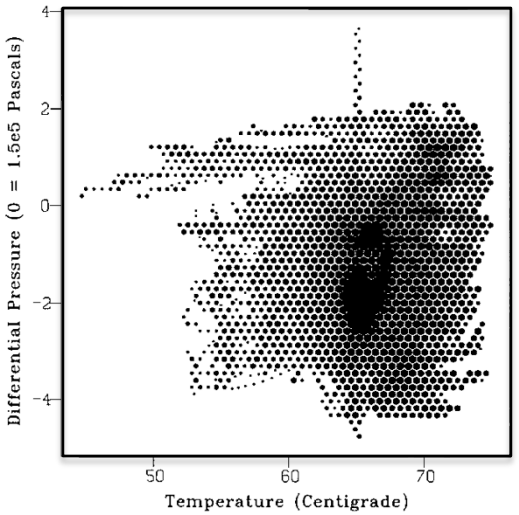
Sizing hex bins as alternative to coloring

Use circles instead of hexagons and place them at centroids of hex bins

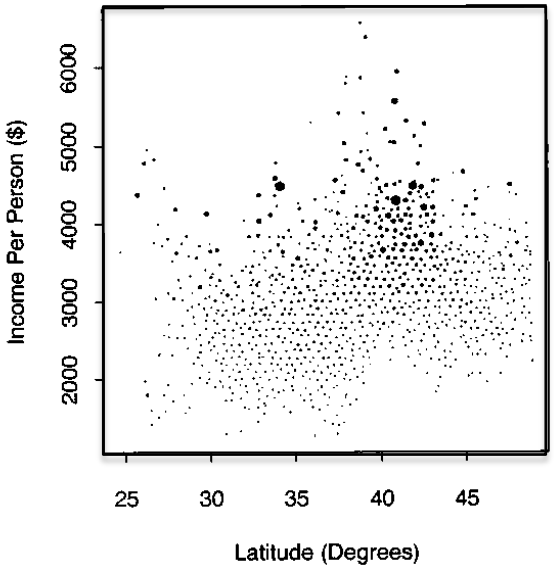
Scale circles so the largest size is less than the size of a hexagon and smallest is a pixel

Centering circles on bin centroids removes artifactual regularity

Don't do this!



Do this

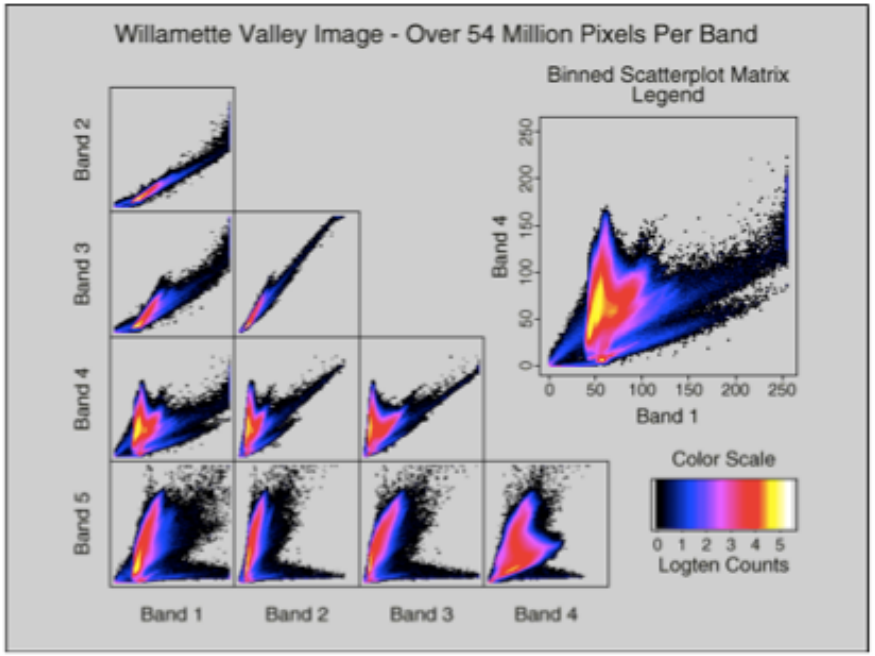


Visualizing

Enhancing Scatterplots by Hex Binning (Dan Carr)

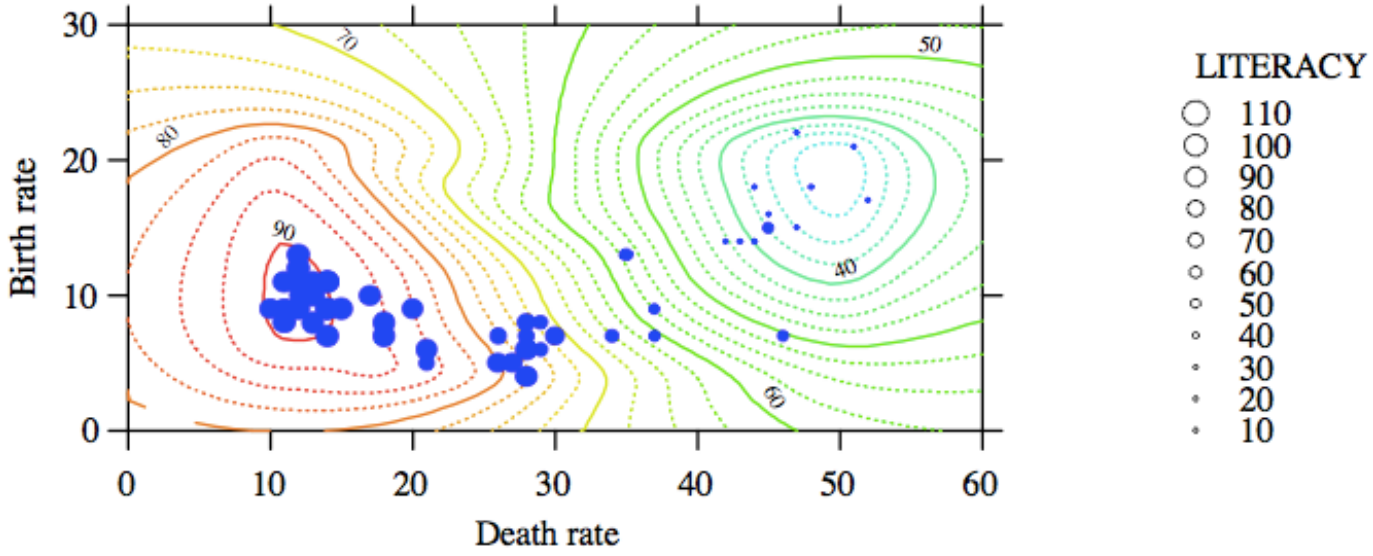
A very big dataset

Dan Carr, *Intensity of Landsat Pixels within Selected Bandwidths for Willamette Region of Oregon*, 1996.



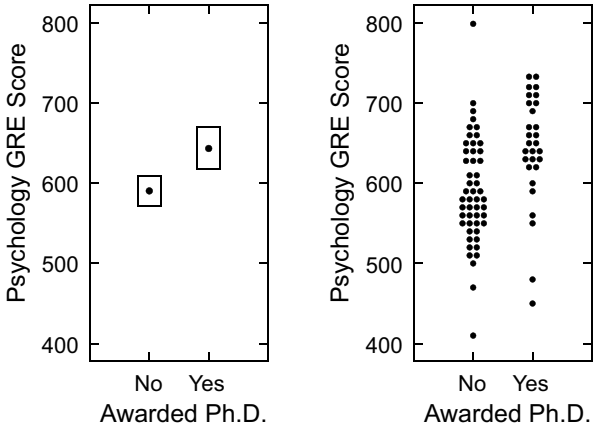
Visualizing

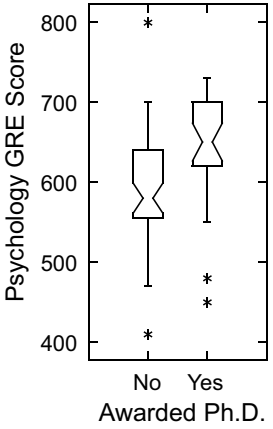
Enhancing Scatterplots (contouring third variable & bubbles)

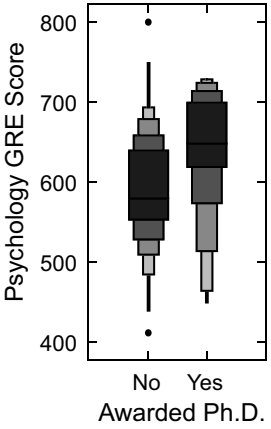


Visualizing

Multigroup densities







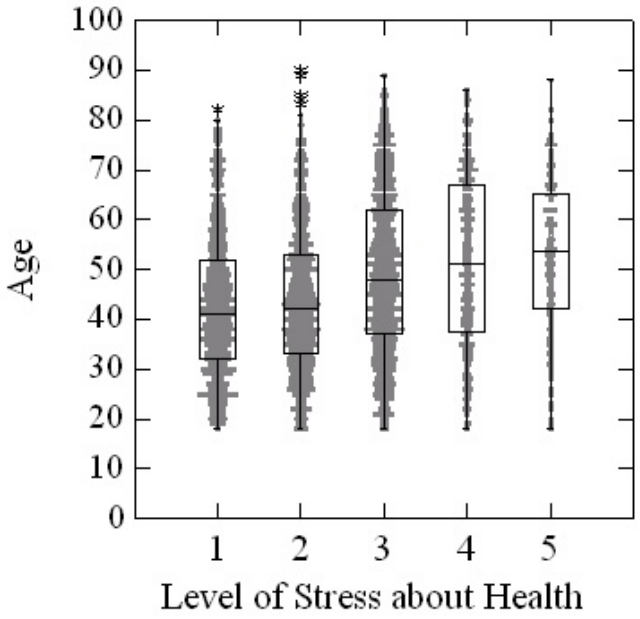
McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of Box Plots. *The American Statistician, 32*, 22-16.

Hofmann, H., Kafadar, K., and Wickham, H. (2006). Letter-value box plots. Technical Report 10, Department of Statistics, Iowa State University.

Visualizing

Multigroup densities

Dot-box plots

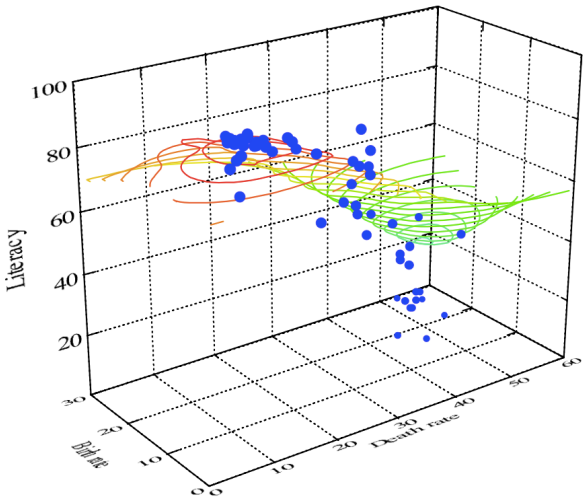


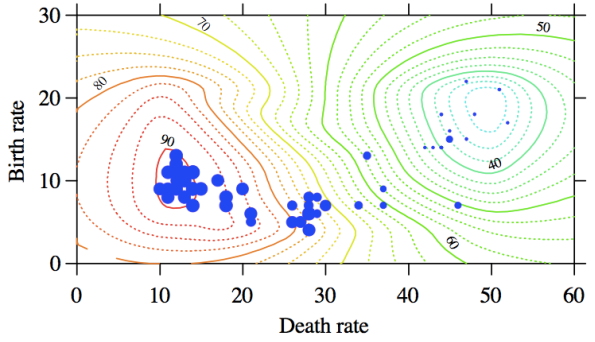
Visualizing

3D

Use it sparingly

If there is coherence/smoothness in the third dimension, it can be useful



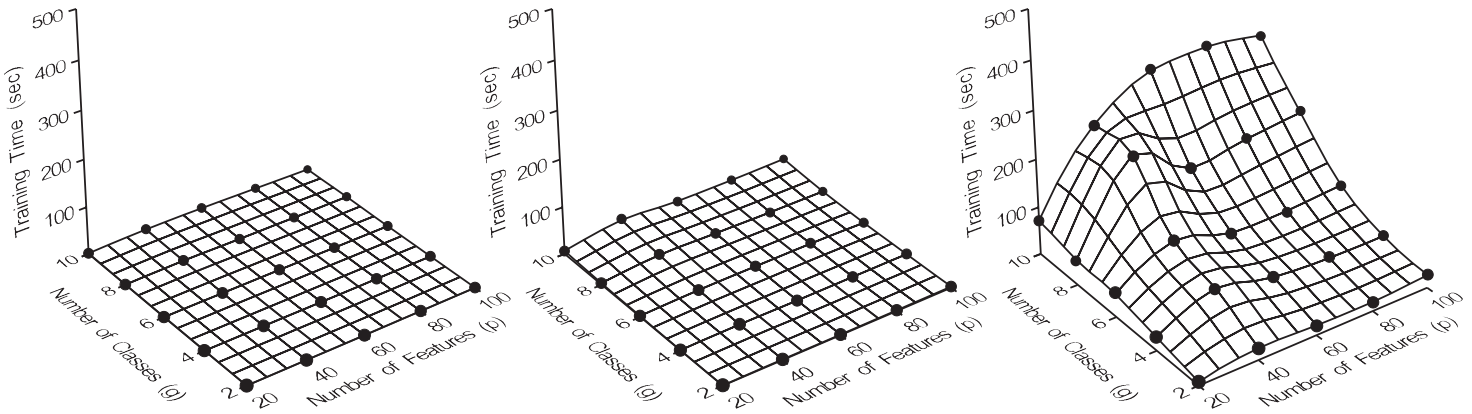


Visualizing

3D

Works well for mathematical functions and nonparametric smooths

n = 500 | n = 5,000 | n = 50,000

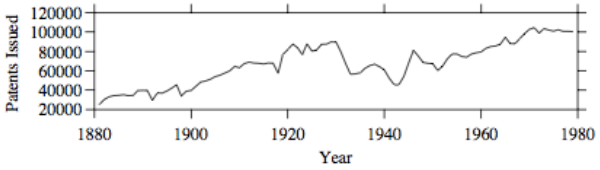


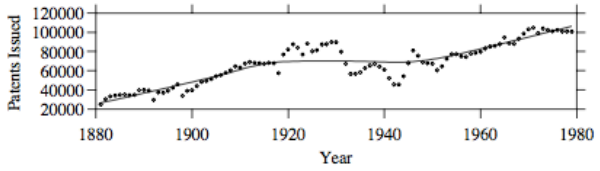
Wilkinson, Anand, Dang (2012)

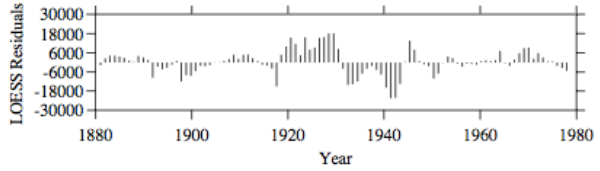
Visualizing

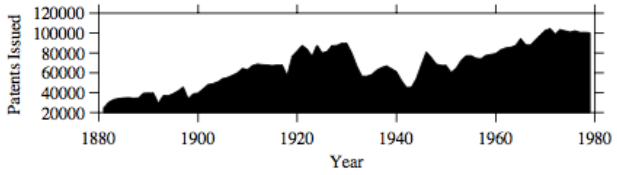
Time Series

- There are several good methods
- The second shows Loess smoother
  - This is a good trend estimator
- The third shows residuals from trend
- The fourth can highlight slope changes







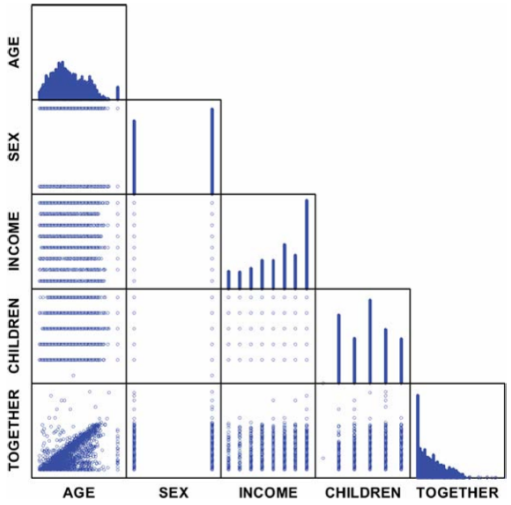


Visualizing

Scatterplot matrices

Perhaps the best single multivariate display method

- Works for up to about 25 variables
- Lensing (or megapixel displays) can improve this limitation
- These data are from a survey
  - The lower left plot reveals that a number of respondents reported they had been together with their partners longer than they were alive!

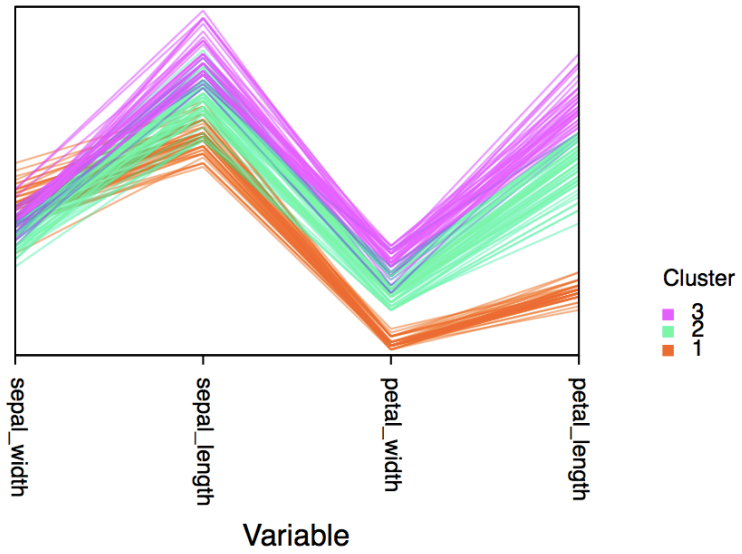


Visualizing

Parallel coordinate plot (PC)

Gannet & Hughes (1883), Hartigan (1975), Inselberg (1985), Wegman (1990)

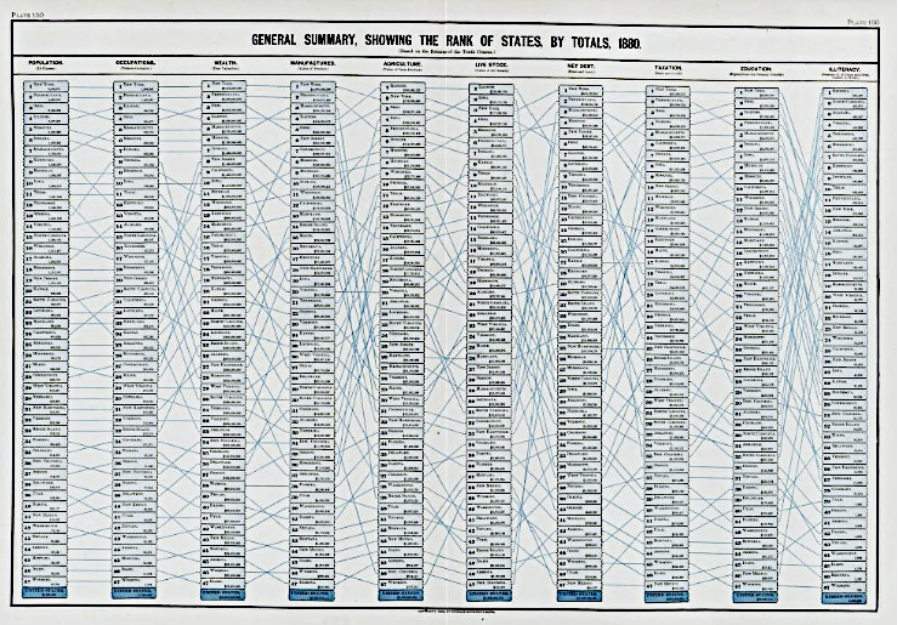
Reading the Wikipedia Talk page on PCs is a hoot



Visualizing

Parallel coordinate plot (PC)

Gannet and Hughes (1883)

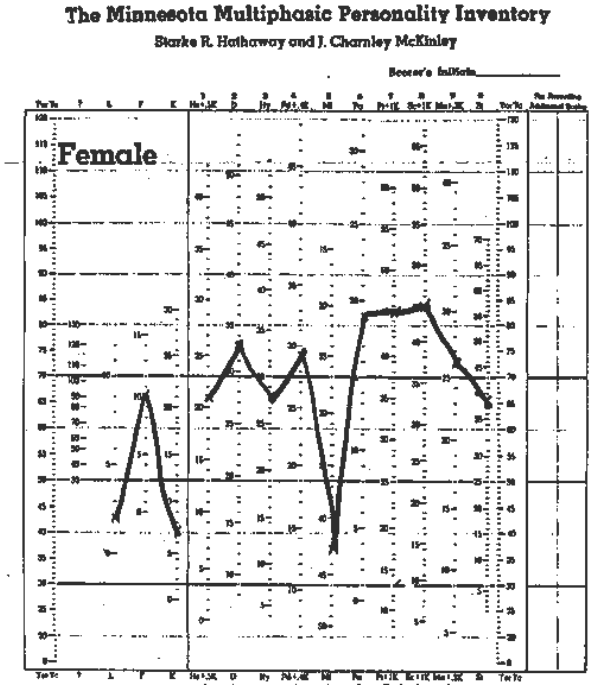


Gannett, H., and Hughes, F.W. (1883). General Summary. Scribner's statistical atlas of the United States. New York: Charles Scribner's Sons.

Visualizing

Parallel coordinate plot (PC)

Hathaway & McKinley (1943)



Hathaway, S. R., & McKinley, J. C. (1943). *Manual for the Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation.

Visualizing

Parallel coordinate plot (PC)

Hartigan (1975)

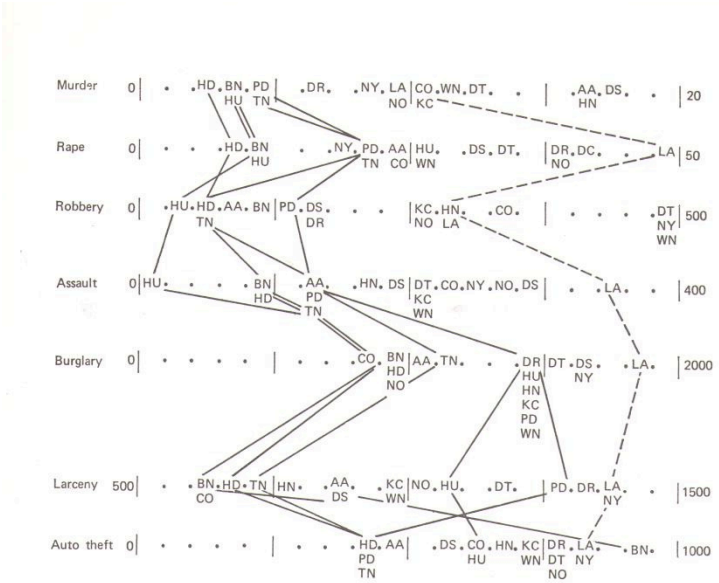
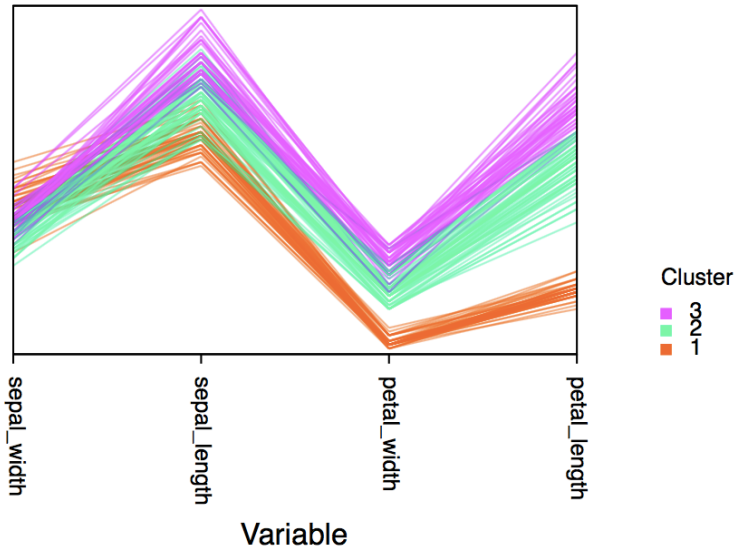


Figure 1.1 Initial profiles of city crime.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: John Wiley & Sons Inc.

Visualizing

- Parallel coordinate plot (PC)
  - Hartigan (1975), Inselberg (1985), Wegman (1990)
    - Reading the Wikipedia Talk page on PCs is a hoot
      - Some people claim to have invented a method that is more than 100 years old
  - Extremely popular
    - Except among statisticians (see weaknesses below)
  - Strengths
    - Many variables can be accommodated
    - Reveals clusters when they are compact
    - Many variations (publication mill)
      - Density representations (Wegman)
      - Stacking (Dang & Wilkinson)
      - Pargnostics (Kosara)
  - Weaknesses
    - Depends on ordering of variables
      - Wrong ordering creates crossings
    - Can conceal correlations
      - Unless variables are adjacent
    - No good for outliers (see next)

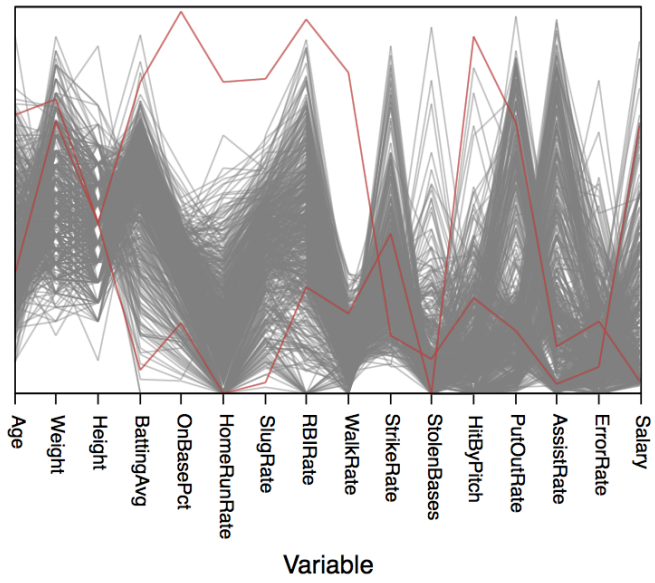


Visualizing

Parallel coordinate plot

Cannot be used for outlier detection

- The upper red profile is a multivariate outlier
- The lower red profile is also a multivariate outlier, but it doesn't stand out



Visualizing

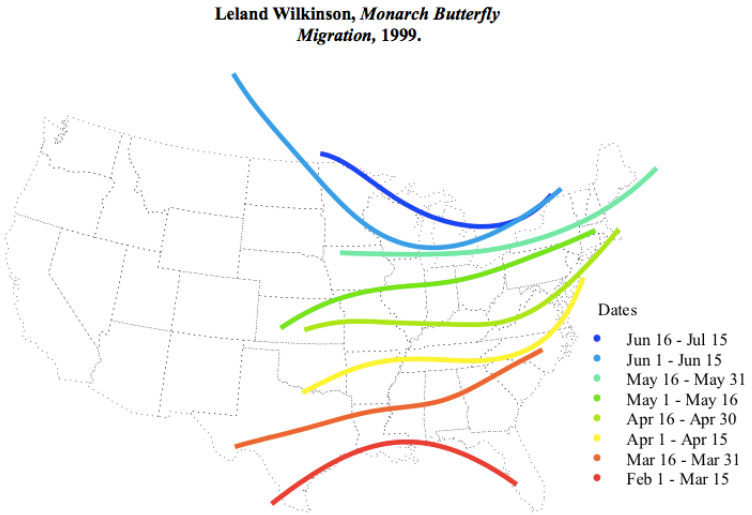
Statistics on maps

Weather maps show contours of temperature, pressures, etc.

These are usually smoothed with kernels

Other statistical methods can be useful

These contours are distance-least-squares estimates of migration waves



Visualizing

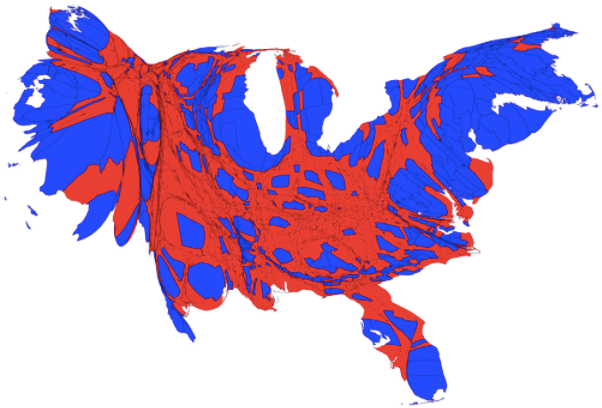
Statistics on maps

Distortion maps and cartograms

- The left map is distorted by price of airline tickets
- The right map is a cartogram that distorts counties to fit voter data
  - It uses a warping algorithm designed by a physicist
    - It's cool, but one might argue about whether it improves perception of the data patterns

Leland Wilkinson, *US Map Warped to fit Cost of Airline Travel*, 1999.

Mark Newman, county-level elections, 2012
http://www-personal.umich.edu/~mejn/election/2012/



Mark Newman, county-level elections, 2012
http://www-personal.umich.edu/~mejn/election/2012/

Visualizing

Statistics on maps

Scatterplots on maps

The map on the left was made by hand at the Census Bureau

The map on the right is a nighttime satellite photograph

US Bureau of the Census

Population Distribution in the US, 1970.

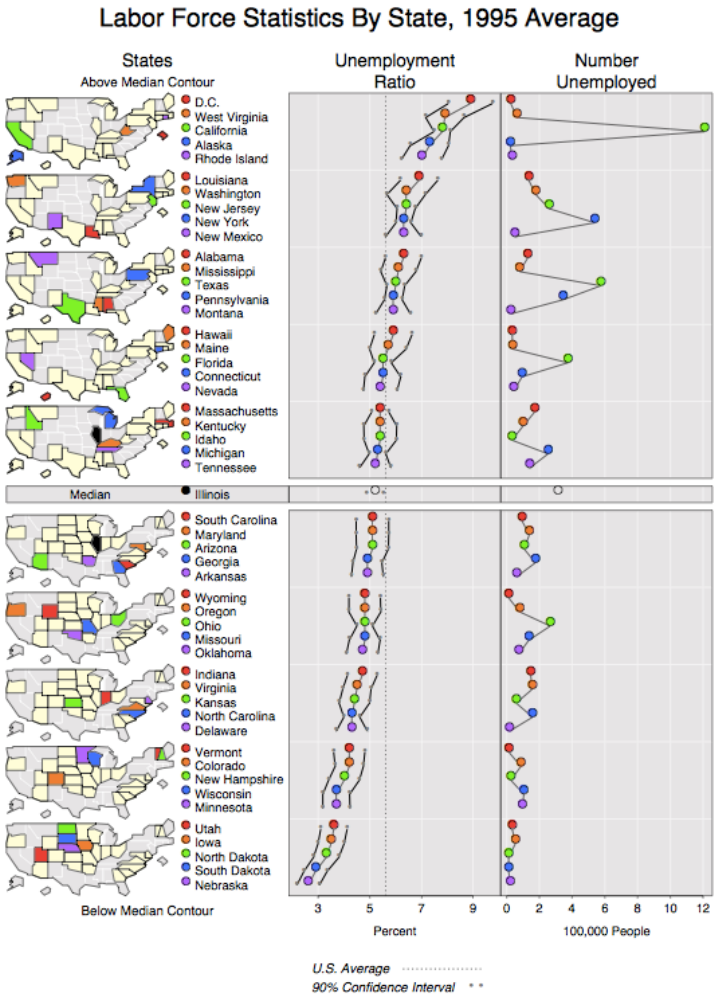
US NOAA

Nighttime Lights of the Continental USA, 2000.

Visualizing

Statistics on maps

Linked Micromaps (Dan Carr)



Carr D.B., Pickle L.W. (2010). *Visualizing Data Patterns with Micromaps*. NY: Chapman & Hall.

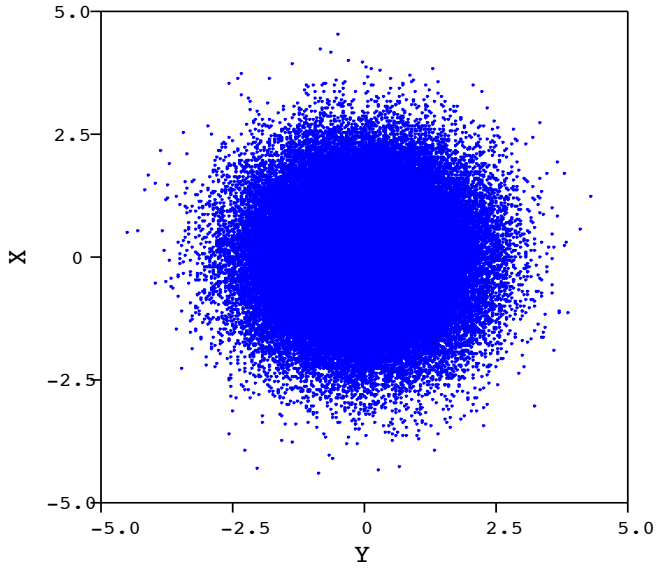
Visualizing

Big Data

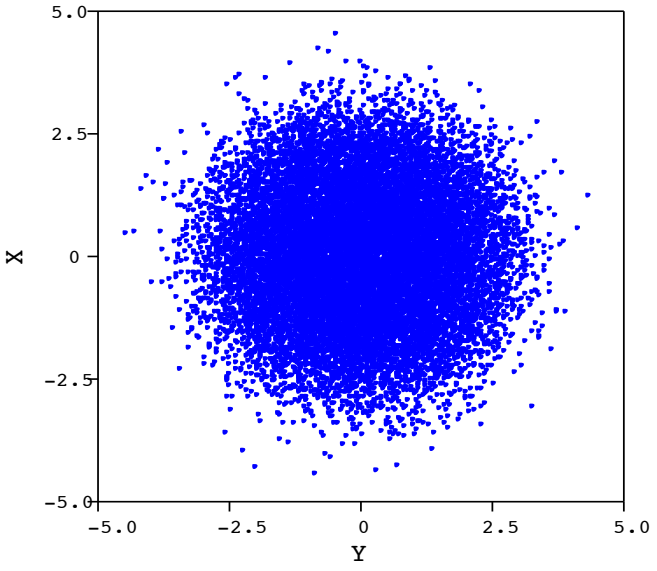
Clustering algorithms can be used to reduce plotted points

- An alternative to hex binning
- Unlike random sampling, outliers survive binning
- Can be employed on higher-dimensional data

100,000 points



13,000 points (compressed)



Visualizing

References

Beniger, J. R. and Robyn, D. L. (1978). Quantitative graphics in statistics: A brief history. *The American Statistician*, *32*, 1-11.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Monterey, CA: Wadsworth.

Cleveland, W. S. (1994). *The Elements of Graphing Data* (Rev. Ed.). Summit, NJ: Hobart Press.

Cleveland, W. S. (1995). *Visualizing Data*. Summit, NJ: Hobart Press.

Fienberg, S. (1979). Graphical methods in statistics. *The American Statistician, 33*, 165-178.

Stigler, S. (1983). *The History of Statistics*. Cambridge, MA: Harvard University Press.

Wainer, H. (1997). *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. New York: Springer-Verlag.

Wainer, H., and Spence, I. (1997). Who was Playfair? *Chance, 10,* 35-37.

Wainer, H., Velleman, P. F. (2001). Statistical graphs: Mapping the pathways of science. *The Annual Review of Psychology, 52*.