IRDS: Visualization

Charles Sutton
University of Edinburgh

Why visualisation?

- Goal 1: Have a data set that I want to understand. This is called exploratory data analysis.
  - Today's lecture.
- Goal II: Want to display data (i.e., for publication)
  - Will save this for later lecture (if time)
- Find or display relationships in the data
- This is a prelude to model building (what is most important to model?)
- Major goal is inter-ocular impact

Visualisations that we won't be interested in

Graphics provide little additional information

For an interesting perspective on this difference, see:
Gelman and Unwin. Infovis and statistical graphics: Different goals, different looks (with discussion). *Journal of Computational and Graphical Statistics.* 2013

[source: Wikipedia]

Univariate data

```
52.6 47.5 18.8 29.8 16.4 46.2 22.1 18.6 23.8 43.7 24.7 33.5 29.3 42.9 29.6 28.9 33.8 23.1 37.8 31.3
18.8 28.8 32.7 34.2 32.0 32.1 21.7 22.7 24.3 23.8 30.7 39.9 34.6 25.7 33.6 29.5 33.6 25.0 12.0 22.8
 3.2 27.4 18.8 41.2 31.1 35.8 26.5 14.2 31.4 38.6 29.2 19.4 33.2 22.4 16.1 14.0 35.7 36.9 14.4 33.2
25.4  0.0 32.9 33.8 35.8 33.7 24.4 50.6 41.8 32.3 11.3 23.5 39.4 47.8 24.2 25.2 27.0 23.8 24.7 26.7
23.2 21.7 33.7 36.6 32.1 26.1 26.8 57.3 32.0  5.5 21.8  3.3 32.2 21.8 17.8 12.0 45.0 36.4 35.9 27.7
22.6 37.7 17.1 39.7 35.1 32.3 28.7 26.5 18.7 37.3 26.1 37.1 21.4 24.6 34.5 34.1 30.2 28.5 44.3 23.7
22.9 37.9 34.4 31.8 25.5 27.1 28.0 21.1 45.0 27.1 35.6 17.2 21.9 41.0 11.8 41.2 39.8 11.1 32.9 22.2
25.5 29.6 31.1 31.7 38.7 28.8 23.0 18.0 36.6 34.7 30.4 25.2 22.6  8.5 19.2 11.3 30.5 13.7 32.3 16.9
33.1 45.8 27.2 35.1 44.7 23.1 14.9 29.6 44.7 27.8 18.2 20.4 24.1 30.4 29.8 30.5 21.5 28.1 38.7 32.7
32.8 27.3 29.9 42.3 12.0 25.0 27.2 37.2 20.9 20.7 30.7 21.5 21.7 16.3 14.2  5.9 21.2 17.1 28.3 19.0
34.9 36.7 32.5 30.8 10.8 19.7 43.5 35.3 18.6 29.0 25.3 26.0 44.7 25.3 24.1 28.0 33.2 29.2 21.7 23.3
30.9 24.2 10.6  8.1 37.7 16.1 17.7 18.5 20.2 31.1 35.6 28.7 18.5 19.3 21.0 12.7 26.5 36.9 24.1 14.2
                    28.0 14.6 21.6 28.5 33.5 31.1 1.0 32.6 34.2 32.5
```

Summaries

Mean 27.7
Std Dev 9.5

Min 0.00
1Q 21.7
Median 28.0
3Q 33.6
Max 57.3

Median and quartiles

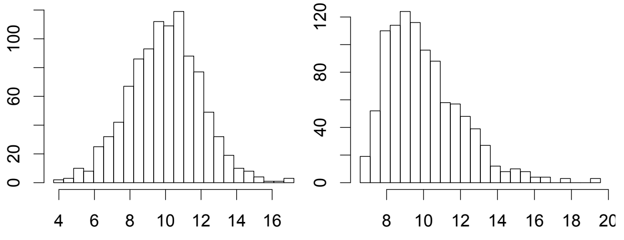
Sample mean

$$\bar{x} = \frac{1}{N}\sum_i x_i$$

Sample standard deviation

$$s_x = \sqrt{\frac{1}{N-1}\sum_i (x_i - \bar{x})}$$

Histograms

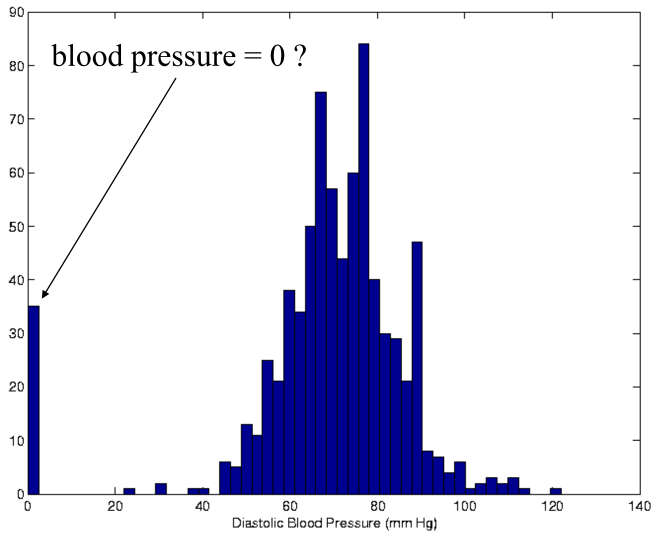


skew

multimodality

these three have same summary statistics!

Outliers in histograms

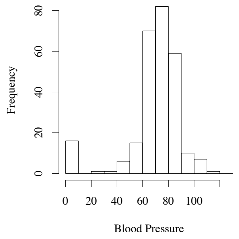


Blood pressure data set

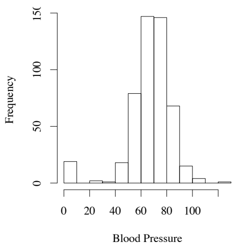
UCI ML repository says no missing data
(well, for 20 years it did)

[Source: Padhraic Smyth]

Class-Conditional Histograms

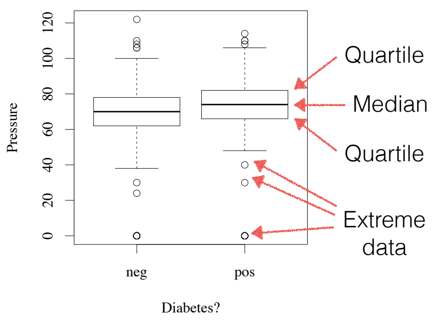


Positive (diabetes)



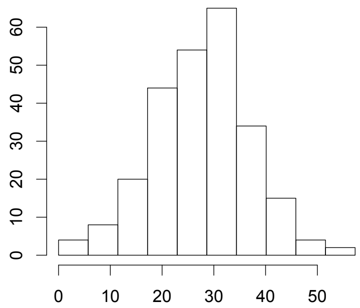
Negative

Alternative: Box plot

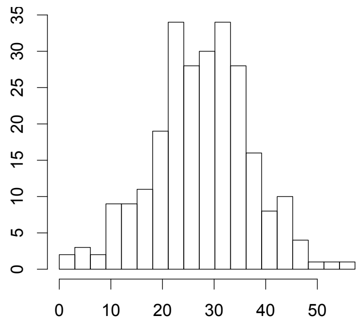


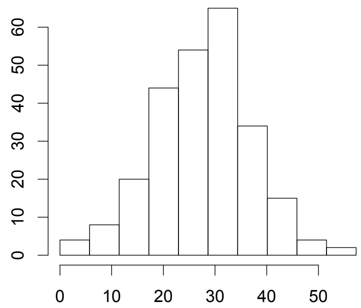
Maybe for only 2 groups, graphs not necessary.
For more visual comparisons, can be helpful.

Effect of bin size

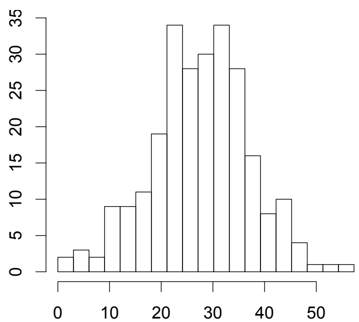


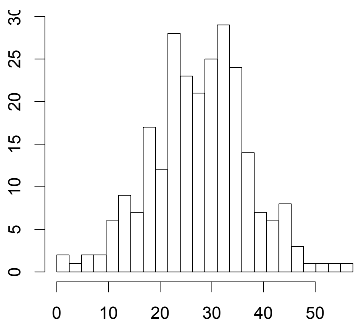
Effect of bin size



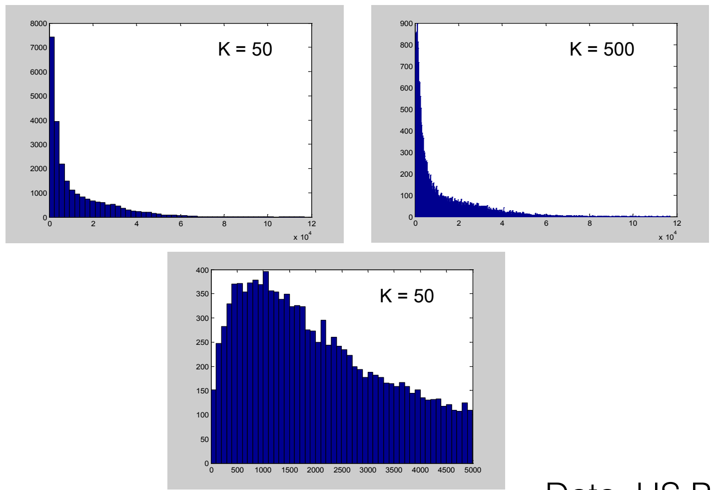


Effect of bin size





More misleading histograms



Data: US Post Codes

[Source: Padhraic Smyth]

Bivariate data

Numerical bivariate summaries

Data are $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$

Sample covariance:

$$s_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})$$

Sample correlation:

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

where as before

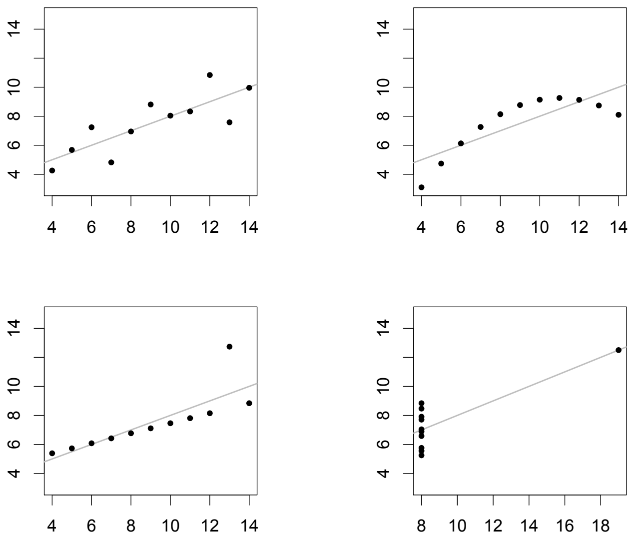
$$\bar{x} = \frac{1}{N} \sum_i x_i$$

$$\bar{y} = \frac{1}{N} \sum_i y_i$$

$$s_x = \sqrt{\frac{1}{N-1} \sum_i (x_i - \bar{x})}$$

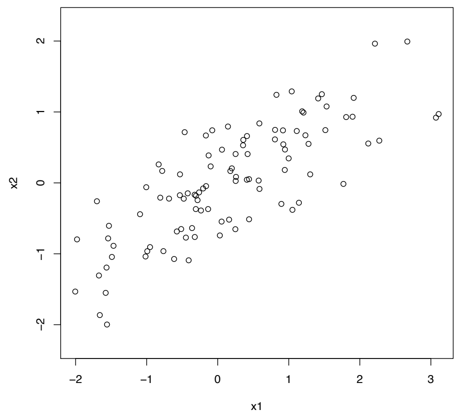
$$s_y = \sqrt{\frac{1}{N-1} \sum_i (y_i - \bar{y})}$$

Dangers of correlation

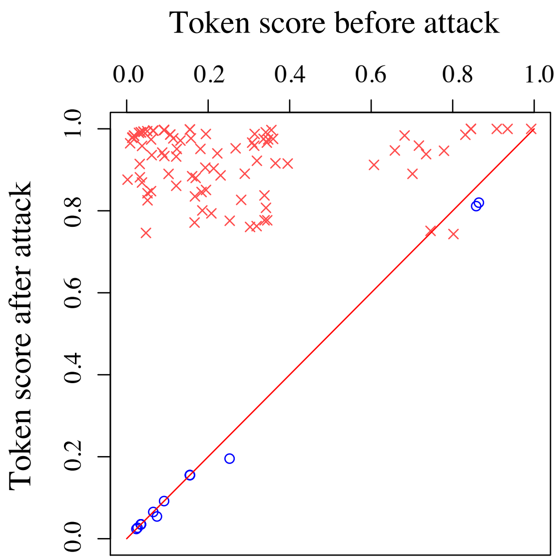


[Anscombe, 1973]

Scatterplots



Colour in Scatterplots

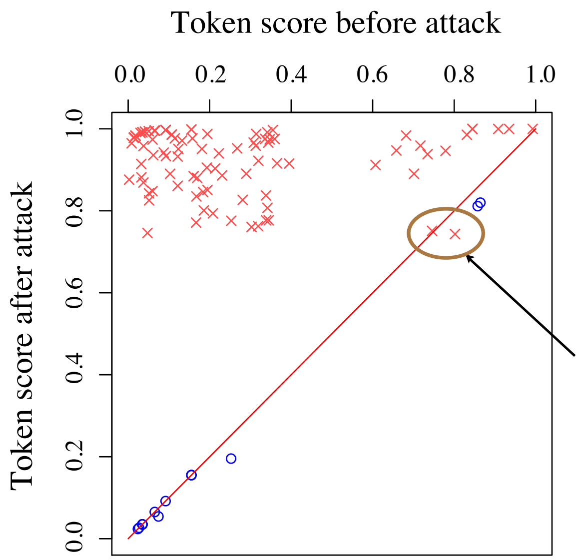


Each point is a word
Entire plot: one email
Axes: “Spam score”

Colour: Whether token was part
of an attack on the spam filter

[Nelson et al, 2008]

Colour in Scatterplots

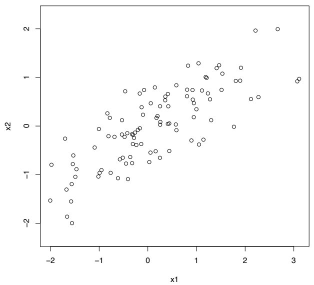


For our purposes, note:

- Use of colour to add a categorical variable
- Without this colour would not have seen these two outliers
- Use of y=x line to add the eye

[Nelson et al, 2008]

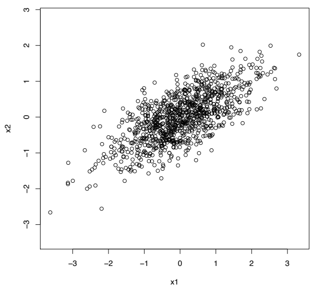
Overplotting



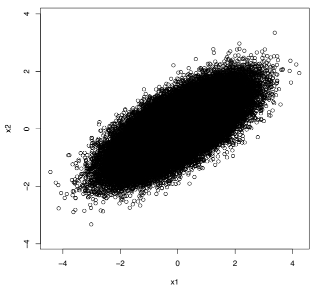
100 data points

samples from bivariate normal

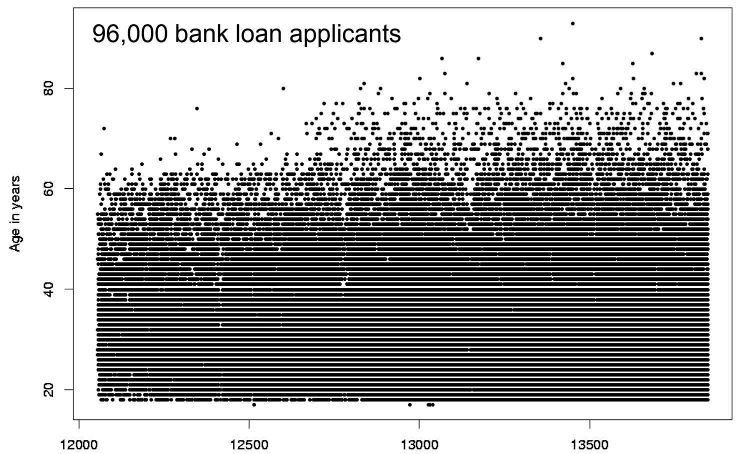
also: notice the axes!



1000 data points



100,000 data points

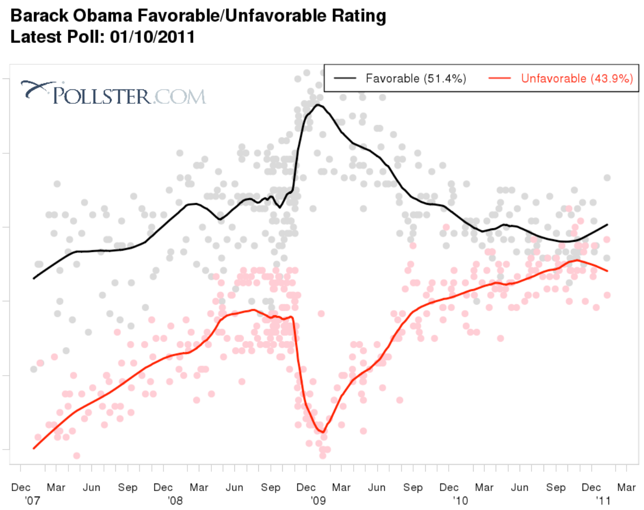


[Source: Hand, Manila, and Smyth]

To fix overplotting, could consider:

- Jittering points
- Subsampling points (i.e., plot only 10%)
- Averaging (if this makes sense)
- Add trend lines (e.g., quantile lines)

Fitted line



This fit is from loess (local linear regression).

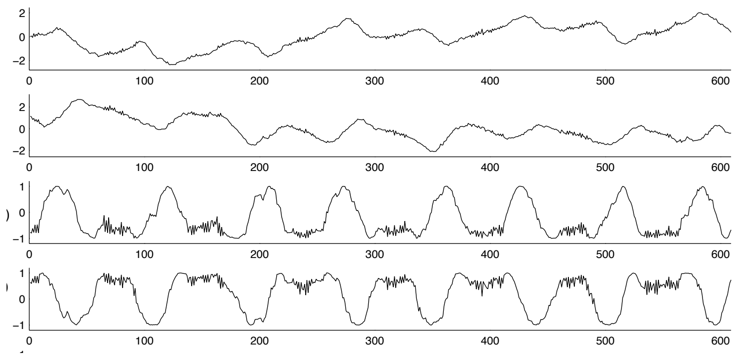
Time Series

Examples

- Financial data
- Network traffic
- Energy usage
- Human traffic
- Building occupancy

Visualization tricks include:

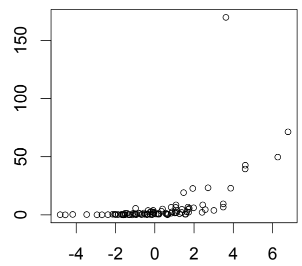
- Smoothing
  - (running mean, median)
- Repeated multiples



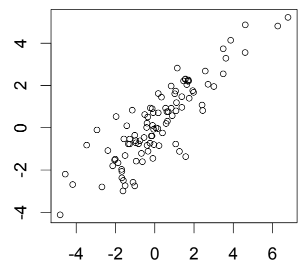
[Oh et al, 2006], figure from [Xuan and Murphy, 2007]

Transformations

Consider powers, logs.
Occasionally reciprocals (e.g., rates).
Also square root

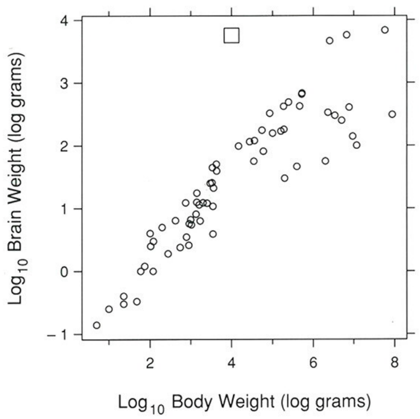


Before



After

Example Transformation



Why log log here? Hint: Imagine a spherical cow

[Source: William Cleveland, Visualizing Data]

Wait, what if you have categorical data?

Tools here include:

- Colour
- Contingency tables
- Multiple plots (e.g., class-conditional histograms)

Three-Dimensional Data

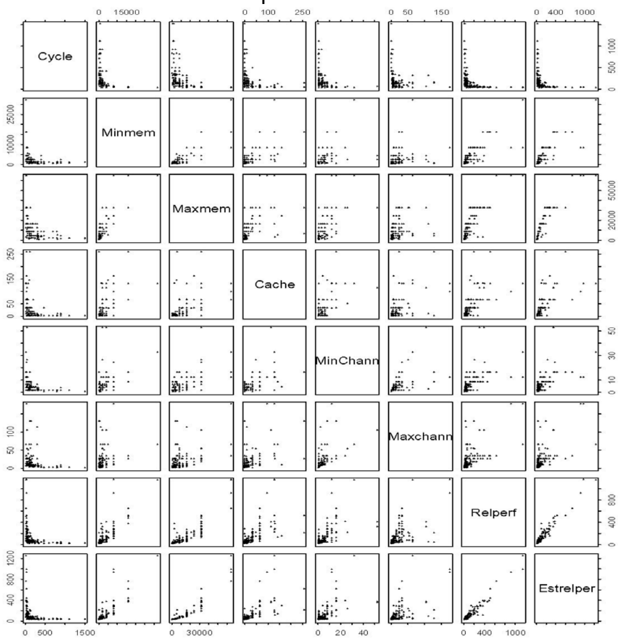
- Generally hard
- 3-D plots are not usually useful
- Usually better to use colour on a 2-D plot
- Or show multiple 2D plots for each value of third variable

High-Dimensional Data

Two main options:

- Project the data down to 2-D
  - Many techniques
    - Principal Components Analysis (IAML, MLPR)
    - Multidimensional scaling
    - Modern nonlinear methods: t-SNE, LLE, Isomap, Eigenmaps
  - Problem: Sometimes this will obscure high-D structure and nonlinear structure
- Another option: Scatterplot matrix (see next)

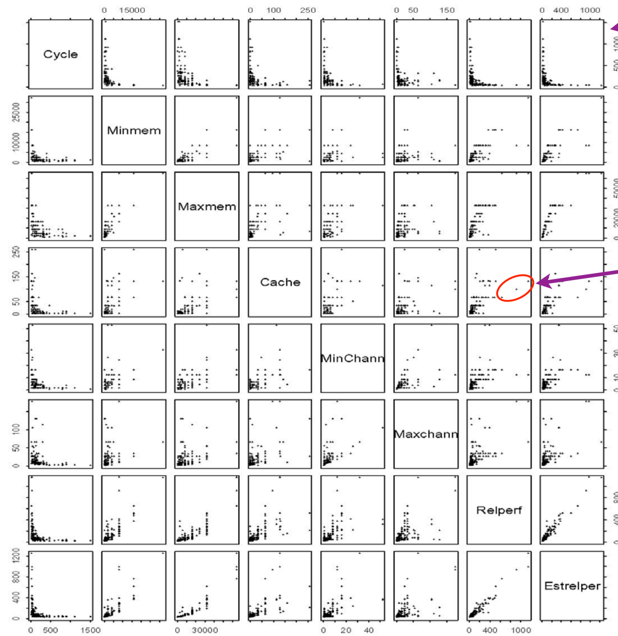
Scatterplot matrix



This is performance data for (very old) CPUs

Important:
Scales must be matched

Scatterplot matrix



Maybe want to use transformed variables up here

Might be worth understanding points like these

This row is the variable we want to predict

This is the prediction according to somebody's model (explains strong relationship)

What are you looking for?

- Anomalies. If something looks weird, figure out why. It could be an error in your data.
  - Learn from your data but do not trust it! (Not completely.)
- Relationships. Hypothesis-based visualization. What relationships do you *expect* to exist? Can you *see* them?
- Use visualization to inform models and vice versa
  - e.g., Can help with feature construction, e.g., whether a relationship is "really" nonlinear
- Fancy 3D graphs … meh
- These techniques also useful for the *outputs* of learning!

If you really like this stuff

- Tukey, Exploratory Data Analysis
- Bill Cleveland, Visualizing Data
- Edward Tufte, all books