
irds

Visualization

- Visualization means two things:
  - Visualization as part of exploratory data analysis.
  - Visualization as a means to present data/results in order to clearly illustrate a particular point or summary.
  - The big gap I put between these was deliberate.

Visualization for Exploratory Data Analysis

- Visualization Principles

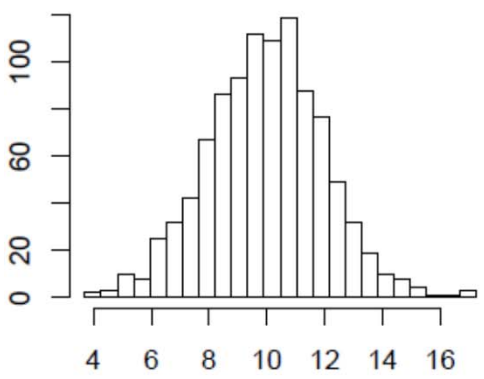
1. Maximize the useful information hitting your retina.
2. View the data as an adversary. It is out to get you! It wants you to fail in your data science endeavour.
3. Never (ever) let any anomaly in the data pass you by without having a potential explanation for it.
4. Search for the information that might help you do your job, and the information that is going to make your job hard.
5. Never forget what you learnt during your exploratory data visualization.

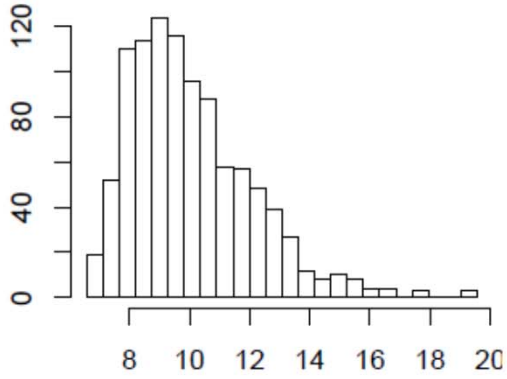
Maximize Useful Information

- Never plot 3D graphs. Colour is your third dimension.
- Avoid overplotting.
- Use less aggregation rather than more.

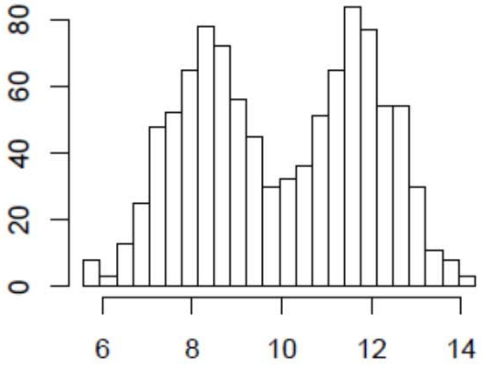
No brainers

- Things to do with every numeric item:
  - Compute summary statistics…
    - Mean, standard deviation, median, range, interquartile range, etc.
    - Correlation, Covariance.
  - Plot histograms (note histograms below all have same mean and variance).



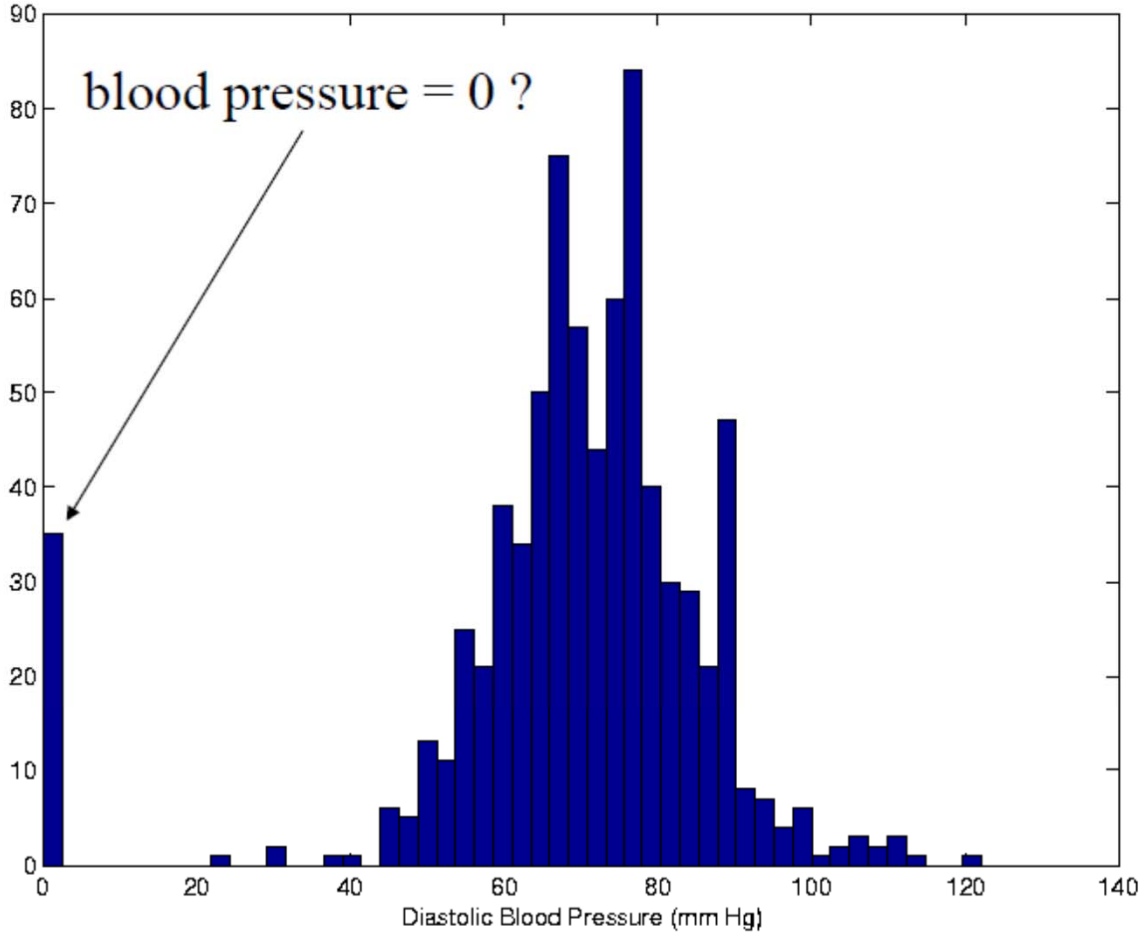


A skew distribution



A bimodal distribution

anomalies



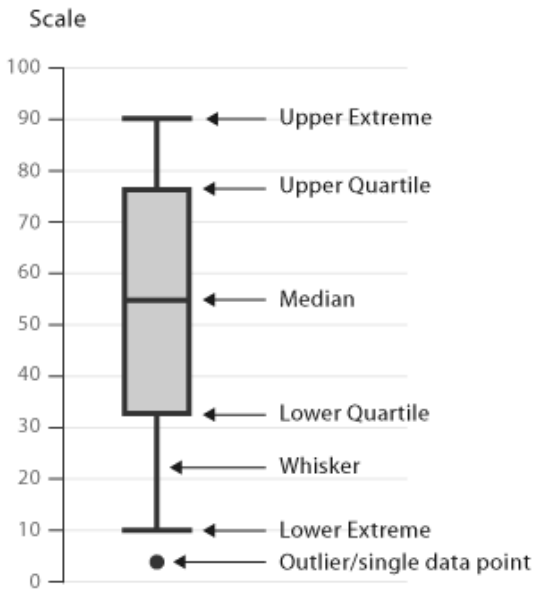
Blood pressure data set

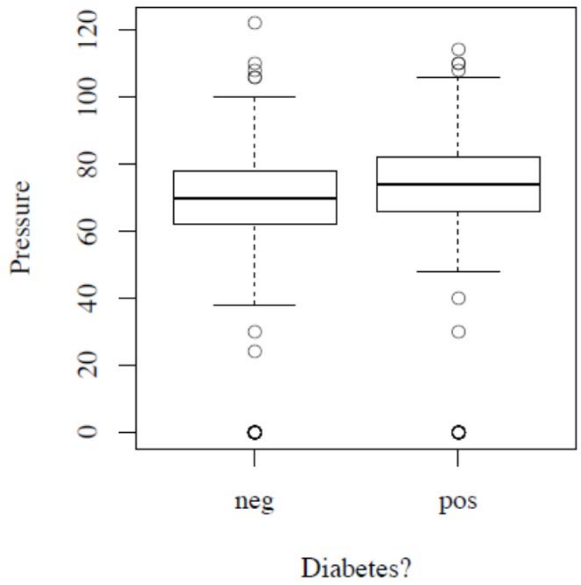
UCI ML repository says no missing data
(well, for 20 years it did)

[Source: Padhraic Smyth]

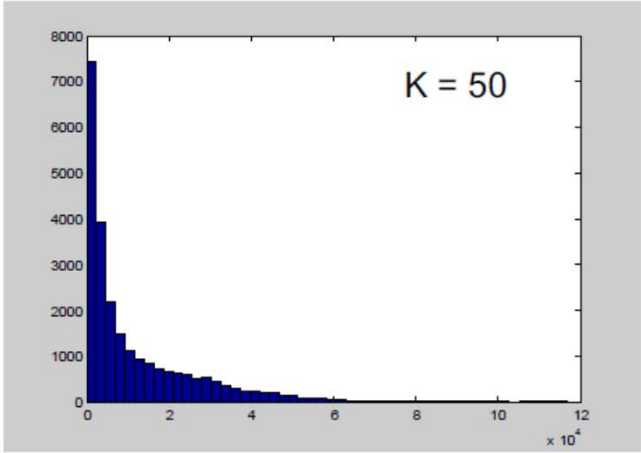
Class conditional data

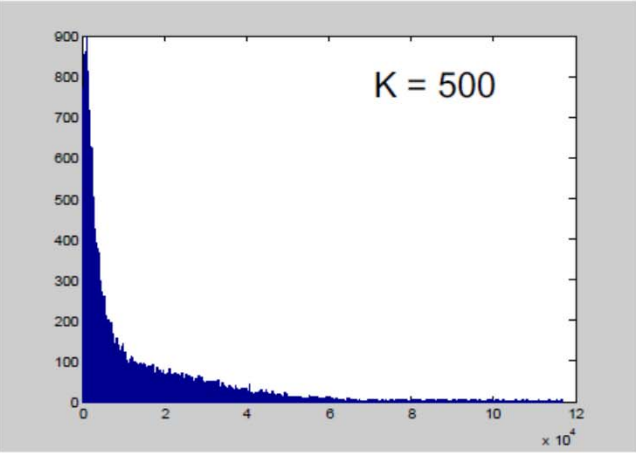
- Boxplots
- Make sure you know what type of boxplot
- Median, interquartile most common.

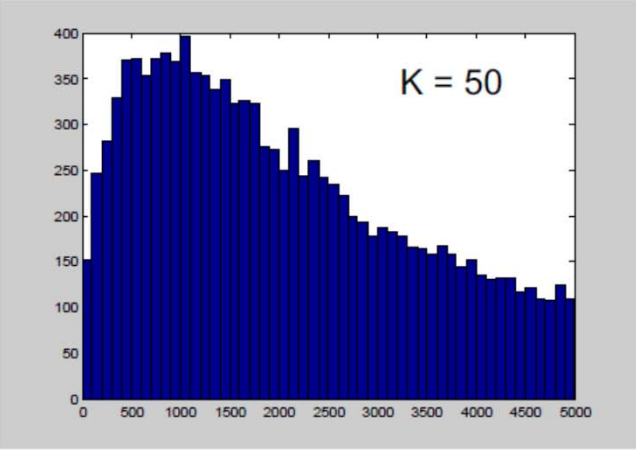




Check deeply







Data: US Post Codes

[Source: Padhraic Smyth]

Between two variables

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Let

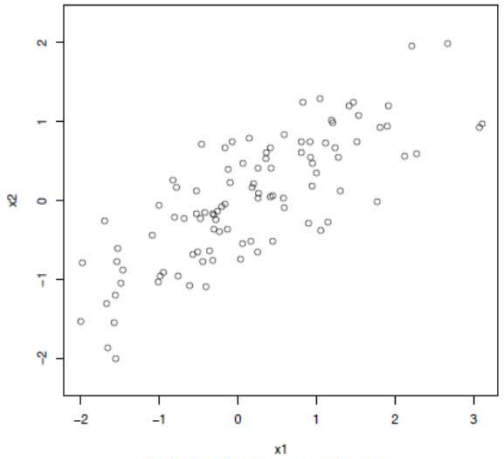
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad s_x = \sqrt{\frac{1}{N-1}\sum_{i}(x_i - \bar{x})^2}.$$

Likewise $\bar{y}$, s_y. Then the sample covariance is

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

and the sample correlation is $\rho_{xy} = \frac{s_{xy}}{s_x s_y}$.

Scatterplots and overplotting

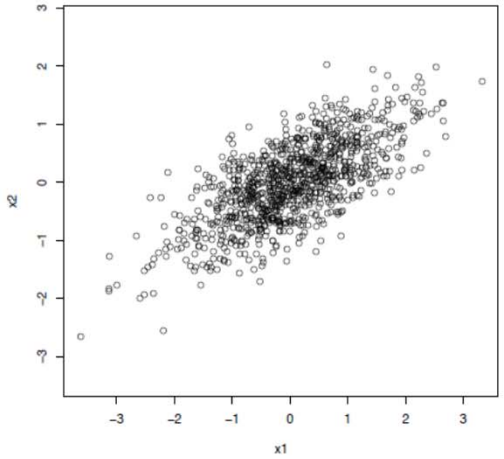


100 data points

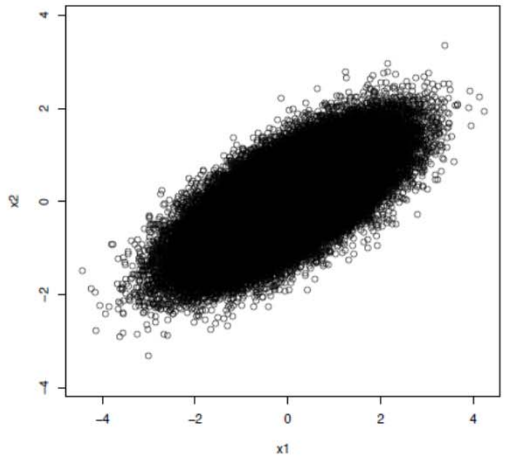
Overplotting

samples from bivariate normal

also: notice the axes!



1000 data points

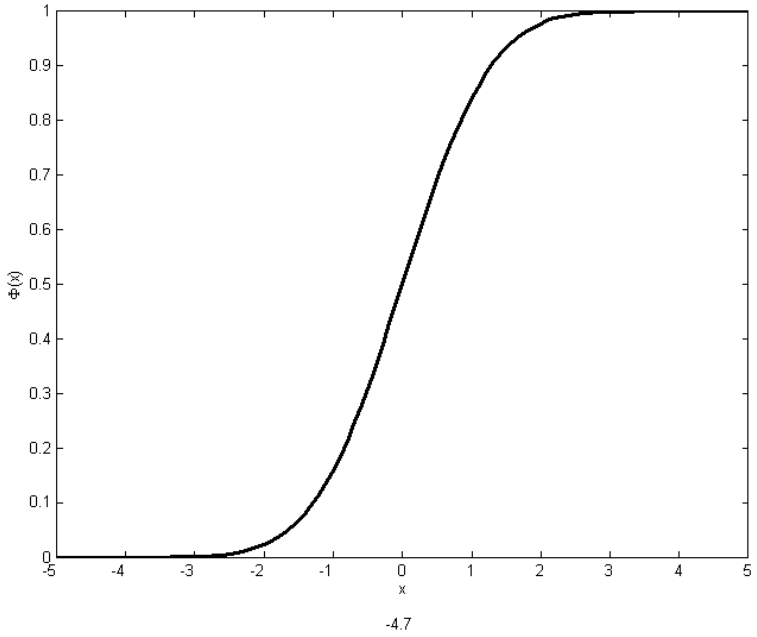


100,000 data points

Amos Storkey

Transformations

- Log? Exp? Sqroot? Square? Sin/Cos?
- Also Gaussianization?



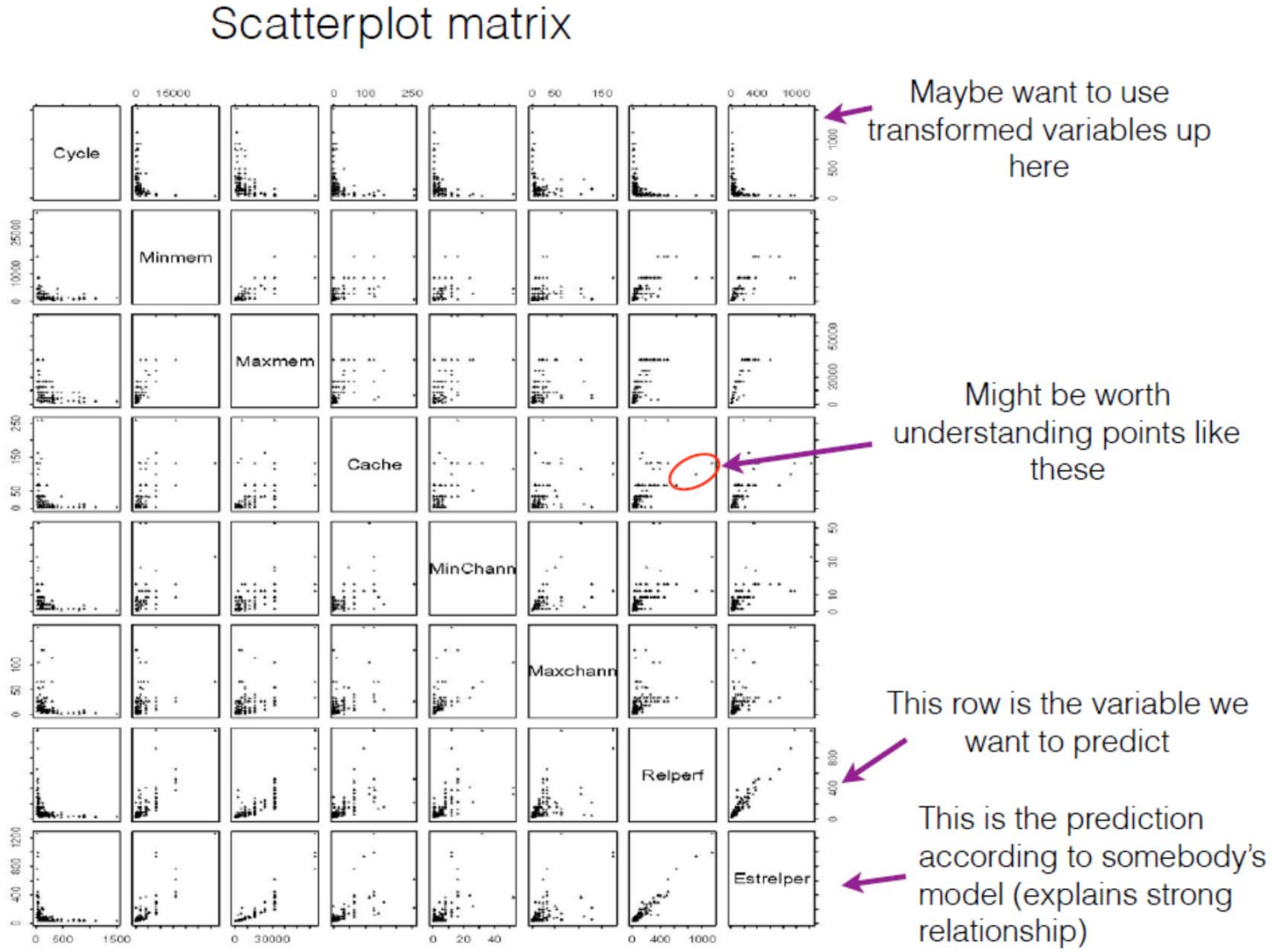
-4.7

What is the appropriate distribution?

- Consider the tails of distributions
- No infinite uniform distribution
- Power law distributions?
- Boundedness
- Truncated distributions
- Thresholded distributions

- This constrains appropriate models.

Scatterplot matrices



Irrelevant variables?

- Test the informativeness of variables.

- How?

Summary

- Visualization is essential but not scalable (in dimension or size).
- For exploration simple is good. Histograms. Scatterplots.
- Do lots fast.
- Principle of small multiples
- Colour is the fourth dimension
- Understand the right axes and transformations.
- Understand all oddities. Find oddities you don't understand.
- How do you expect things to relate? Do you have evidence for this?
- Visualization informs models and vice versa.
- The data is out to get you. Defend yourself with knowledge.

Thanks

- Charles Sutton for material from previous IRDS slides.