Data Exploration and Graphical Myths

What do users seek? (and what do Graphics offer?)

Hard facts	**General information**
Answers to known questions	Clarification of ideas
Specific precise details	Summaries in context
Standard database queries	Exploration is necessary
Facts	Insight

Caveats

- Data always need cleaning
- Data often have to be transformed or restructured
- There is never all the data you want
- There is never all the background information needed
- Datasets are rarely independent random samples (as is assumed in Statistics), so generalise with care
- Large numbers of variables are hard to manage

Recipe dataset

- 4646 recipes from a magazine website
  - tried and tested
- 3398 variables
  - ID and name, date added
  - ingredients (binary variables)
  - properties (calories, carbohydrates, fat, protein)
  - time, standard, vegetarian, alcohol

Recipe problems

- Multiple recipes?
- Variations of ingredient names
  - alternatives
  - singular/plural
  - misspellings
- Errors in quantitative variables
  - large errors are easy to spot, other are not

Case study: Shipman dataset

In 2000 the British doctor, Harold Shipman, was convicted of murdering 15 of his patients. The official report (www.the-shipman-inquiry.org.uk/), which examined the deaths of all patients under his care over twenty years, concluded that he had probably murdered over 200. Details of the deaths of 508 of his patients where there was doubt about the cause of death have been taken from Appendix F of the report.

Variable	Description
ID	patient number
Day	day of death
Month	month of death
Year	year of death
Weekday	day of week of death
Date	days since 1/1/1904
Name	full name of patient
Surname	surname of patient
Sex	gender of patient
Age	age at death
Location	place of death
Decision	official view on Dr.'s guilt

From the Appendix

APPENDIX F **Chronological List of Decided Cases**

Date of Death	Name of Deceased	Age of Deceased	Place of Death	Decision
1974				
10/5/74	Ruth Highley	72	Own home	Natural death
22/6/74	Edith Annie Bill	67	Own home	Natural death
23/7/74	Colin Whitham	26	Own home	Natural death
2/8/74	Stanley Uttley	58	Surgery	Natural death
9/10/74	Hena Cheetham	77	Ambulance	Natural death
10/11/74	Harold Edward Jackman	78	Hospital	Natural death
9/12/74	Sean Stuart Callaghan	18	Hospital	Natural death
16/12/74	Moira Kelly	26	Hospital	Natural death
29/12/74	Sarah Ann Thomas	86	Own home	Insufficient evidence for decision
There is also a decision in respect of Frances Elaine Oswald, relating to an incident which took place on 21/08/74				
1975				
21/1/75	Lily Crossley	73	Own home	Suspicion of unlawful killing
21/1/75	Robert Henry Lingard	62	Own home	Suspicion of unlawful killing

Graphical Myths (1)

- General
  - One graphics window is enough
  - A caption should be as short as possible
  - Graphics should be easy to understand
  - Graphics are for reading off values
  - Missing values can be ignored

Graphical Myths (2)

- General
  - Always use the default scales provide by software
  - The more gridlines and tickmarks the better
  - The more detailed the scales the better
  - The more colour/decoration the better
  - Fake 3-d improves the look of plots
  - Vaguely relevant images in the background help

Graphical Myths (3)

- Histograms
  - Histograms are for density estimation
  - Determine histogram bins by number not by binwidth
  - Binwidth should be calculated by a formula
  - You can find an optimal histogram
- Boxplots
  - There is no standard definition of a boxplot
  - Boxplots should be as wide as the plot frame allows

Graphical Myths (4)

- Scatterplots
  - The ideal point display glyph is an empty circle
  - Sunflowers are good for displaying overlapping points
  - Jittering is a good for coping with overlapping points
  - Include case ID numbers in scatterplots
  - Plots of data v case index are useful

Graphical Myths (5)

- Categorical data
  - Nominal variable categories should be numbers not text
  - Hatching bars is OK
  - Always add numbers to columns in barcharts
  - Point plots with jittering can represent categorical data
  - You can't draw displays for multivariate categorical data
  - Sieveplots are useful

Interactive Graphical Myths

- Interactive graphics
  - Dynamic graphics are more important than interactive
  - Brushing comes before static linking
  - 3-d rotating plots are very useful
  - Animation is an important exploratory tool
  - Graphics without labels are no good
  - You have to be able to log what you have done