



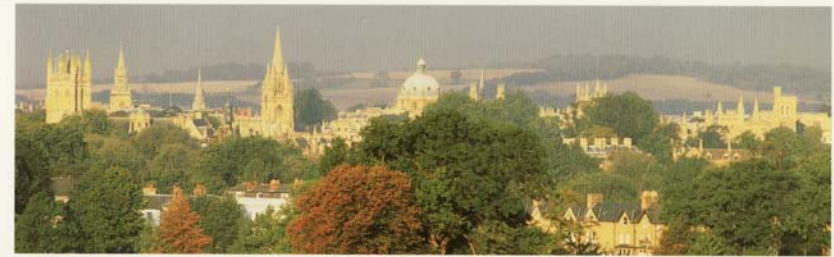
What You Get Is What You See – Graphics for Data Analysis

Antony Unwin
Augsburg University

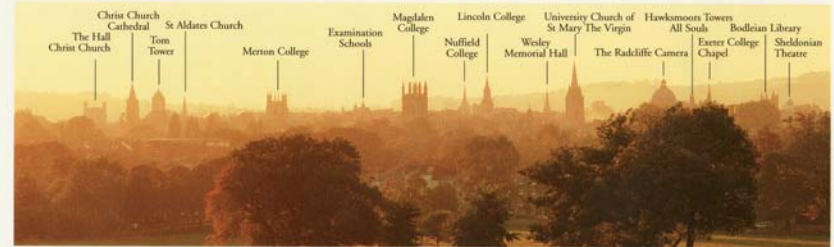
WYGIWYS

Antony Unwin

Oxford, 18th /25th February 2010



UNIVERSITY OF OXFORD



What do you think of when you
hear the term “graphics”?

WYGIWYS

Antony Unwin

Oxford, 18th /25th February 2010

Some questions to ask



- Which graphics might help?
- Why?
- What stories do the graphics tell?
- Could the stories be told better?
 - with reformatting
 - with other graphics

WYGIWYS

Antony Unwin

Oxford, 18th /25th February 2010

Examples from R



- Aspirin
- Anorexia
- Divorce
- Lanza
- Sexual Fun

Plus some
media examples in between

Plus some
Interactive Graphics

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010

Aspirin (HSAUR2)



Efficacy of Aspirin in preventing death after a myocardial infarct.

A data frame with 7 observations on the following 4 variables.

Description

- dp** number of deaths after placebo
- tp** total number subjects treated with placebo
- da** number of deaths after Aspirin
- ta** total number of subjects treated with Aspirin

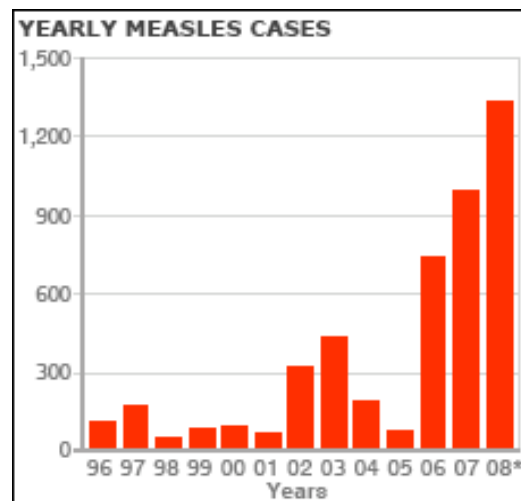
Suggested code

```
data("aspirin", package = "HSAUR2")
aspirin
```

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010



*provisional SOURCE: Health Protection Agency

BBC 6th February, 2009

Anorexia (MASS + ...)



The anorexia data frame has 72 rows and 3 columns. Weight change data for young female anorexia patients.

This data frame contains the following columns:

Treat

Factor of three levels: "Cont" (control), "CBT" (Cognitive Behavioural treatment) and "FT" (family treatment).

Prewt

Weight of patient before study period, in lbs.

Postwt

Weight of patient after study period, in lbs.

WYGIWYS

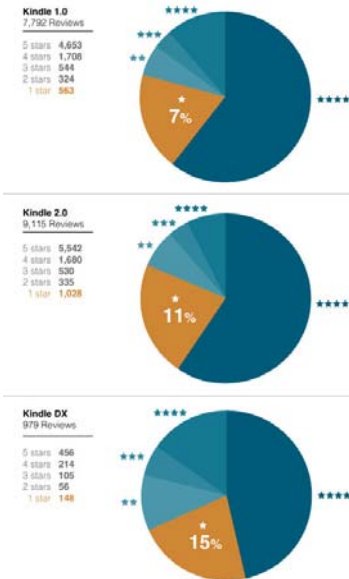
Antony Unwin

Oxford, 18th/25th February 2010



Anorexia is also in

- *granova* (with graphic)
- *glmperm* (permutation likelihood tests)
- *multcompView* (with graphic)
- *Deducer* (testing normality of data)
- *pgirmess* (an ecology package)
- *rrcov* (robust methods)
- *accuracy* (testing accuracy of results)



Times Customer feedback ratings for three versions of the Amazon Kindle.



Interactive Graphics: what?

Exploring data through graphical displays

Direct manipulation of statistical objects

Fast, flexible and forgiving



Interactive Graphics: why?

I hear and I forget

I see and I remember (=> graphics)

I do and I understand (=> interaction)

Chinese proverb

IG and the movies



- Movies data downloaded from the web
- Just over 87,000
- Information on
 - Year, Length, Budget
 - Type (7 binary variables)
 - Classification
 - Viewers' ratings

Some movie questions



- What missing value patterns are there?
- What is the distribution of ratings?
- Do modern films get more votes and higher ratings?
- What sort of ratings do action films get?
- Are short films less often rated than non-shorts?
- What kind of films are long films?

More movie questions



- What kinds of film get high ratings based on few votes?
- What combinations of film types are there?
- What are individual ratings distributions like for the most highly rated films?
- Which film titles occur most often and are these films all from different years?

Interactive Graphics: how?



- Multiple views
- Querying
- Linking
- Sorting and reformatting
- Zooming

Software



- Mondrian (Martin Theus)
 - interactive graphics
 - crossplatform, links to R via Rserve
- JGR (Markus Helbig/Simon Urbanek)
 - GUI for R (available as R package)



stats.math.uni-augsburg.de

What have we learnt?





What You Get Is What You See – Graphics for Data Analysis (2)

Antony Unwin
Augsburg University

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010



What did you learn?

- Graphics can be informative
- You have to look at graphics carefully
- Graphics should tell a story
- Interactive graphics are powerful
 - multiple views, querying, linking
 - sorting and reformatting, zooming

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010

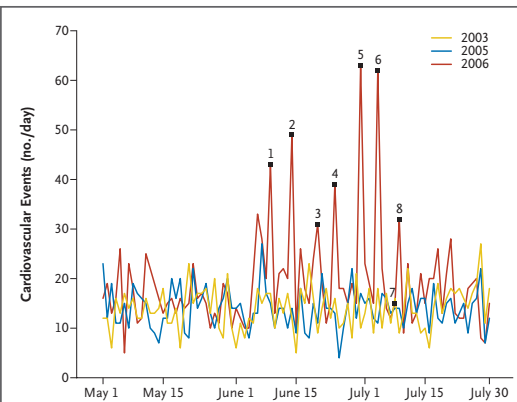


Figure 1. Daily Cardiovascular Events in the Study Population from May 1 to July 31 in 2003, 2005, and 2006.

The FIFA World Cup 2006 in Germany started on June 9, 2006, and ended on July 9, 2006. The 2006 World Cup matches with German participation are indicated by numbers 1 through 7: match 1, Germany versus Costa Rica; match 2, Germany versus Poland; match 3, Germany versus Ecuador; match 4, Germany versus Sweden; match 5, Germany versus Argentina; match 6, Germany versus Italy; and match 7, Germany versus Portugal (for third-place standing). Match 8 was the final match, Italy versus France.

Study population is patients in the greater Munich area



Graphic displays have

- Graphics
- Titles
- Captions
- (Scales, legends, annotations)
- Accompanying text
 - they should all tell the same story

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010



Divorce (vcd)

A 4-dimensional array resulting from cross-tabulating 1036 observations on 4 variables. The variables and their levels are as follows:

Name Levels

MaritalStatus Divorced, Married

ExtramaritalSex Yes, No

PremaritalSex Yes, No

Gender Women, Men

A sample of about 500 people who had petitioned for divorce, and a similar number of married people were asked two questions regarding their pre- and extra-marital sexual experience:
 (1) "Before you married your (former) husband/wife, had you ever made love with anyone else?",
 (2) "During your (former) marriage (did you) have you had any affairs or brief sexual encounters with another man/woman?"
 (Thornes and Collard 1979)

, , PremaritalSex = Yes, Gender = Women

ExtramaritalSex
 MaritalStatus Yes No
 Divorced 17 54
 Married 4 25

, , PremaritalSex = No, Gender = Women

ExtramaritalSex
 MaritalStatus Yes No
 Divorced 36 214
 Married 4 322

in R

, , PremaritalSex = Yes, Gender = Men

ExtramaritalSex
 MaritalStatus Yes No
 Divorced 28 60
 Married 11 42

, , PremaritalSex = No, Gender = Men

ExtramaritalSex
 MaritalStatus Yes No
 Divorced 17 68
 Married 4 130

For Mondrian

Gender	Prem	Extram	Divorced	Count
m	n	n	y	68
m	n	n	n	130
m	n	y	y	17
m	n	y	n	4
m	y	n	y	60
m	y	n	n	42
m	y	y	y	28
m	y	y	n	11
w	n	n	y	214
w	n	n	n	322
w	n	y	y	36
w	n	y	n	4
w	y	n	y	54
w	y	n	n	25
w	y	y	y	17
w	y	y	n	4



Divorce (Michael Friendly)

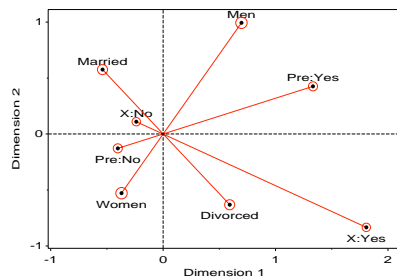
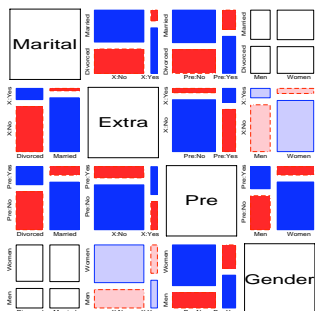
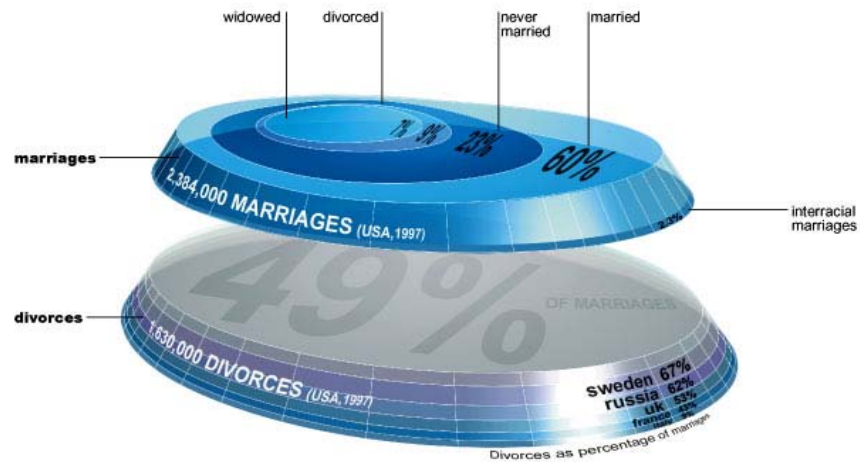


Figure 15: 2D multiple correspondence analysis display for marital status data

Figure 14: Mosaic matrix for marital status data. Each panel shows the bivariate marginal association.

marital status



Questions for media graphics



- Where do the data come from? How were they collected, by whom, when?
- How good are the data likely to be?
- Does the display do justice to the data?
- Do the conclusions make sense?
- Is there corroborating information?

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010

Lanza (HSAUR2)



Data from four randomised clinical trials on the prevention of gastrointestinal damages by Misoprostol reported by Lanza et al. (1987, 1988a,b, 1989).

A data frame with 198 observations on the following 3 variables.

study

a factor with levels I, II, III, and IV describing the study number.

treatment

a factor with levels Misoprostol Placebo

classification

an ordered factor with levels 1 < 2 < 3 < 4 < 5 describing an ordered response variable.

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010

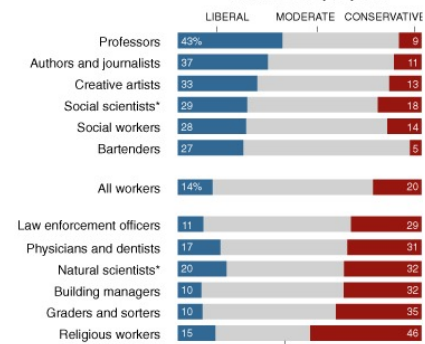
The New York Times

January 18, 2010

Ideology at Work

Professors are more likely to identify themselves as liberals than those in any other occupation, according to an analysis of General Social Survey data from 1996 to 2008.

Percent who say they are:



*Non-academic Moderate includes "slightly liberal" and "slightly conservative."

Source: Neil Gross, University of British Columbia

The New York Times
TWITTER

SIGN IN TO RECOMMEND

Close Window

Sexual Fun (vcd)



A 2-dimensional array resulting from cross-tabulating the ratings of 91 married couples. The variables and their levels are as follows:

Name Levels

Husband Never Fun, Fairly Often, Very Often, Always Fun

Wife Never Fun, Fairly Often, Very Often, Always Fun

Data from Hout et al. (1987) given by Agresti (1990) summarizing the responses of married couples to the questionnaire item: Sex is fun for me and my partner: (a) never or occasionally, (b) fairly often, (c) very often, (d) almost always.

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010

Why visualize?



- Looking for global trends
 - overall structure
- Looking for local features
 - data quality
 - groups or clusters
 - outliers, tail distributions and extremes
 - patterns of all kinds

WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010

Graphics examples



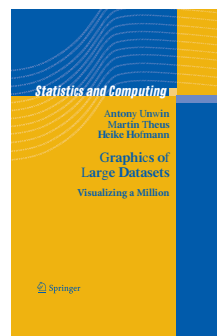
- in R
 - not always good
 - usually complex
 - often without a storyline
- in the media
 - not always good
 - sometimes with a different storyline

WYGIWYS

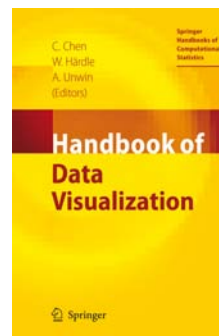
Antony Unwin

Oxford, 18th/25th February 2010

Recent graphics references



and



WYGIWYS

Antony Unwin

Oxford, 18th/25th February 2010