Mosaic Mind Games
— Visualising Categorical Data

Antony Unwin

Augsburg University, Germany

Summary 140

Graphics for categorical data

Many alternative mosaicplots

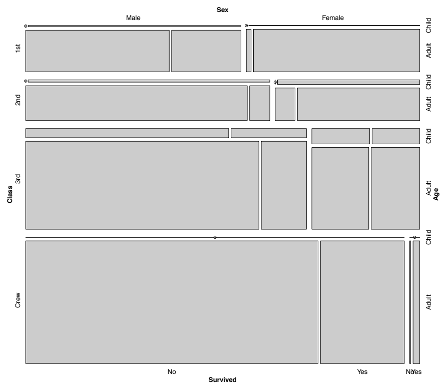
Hard to choose

Need explaining

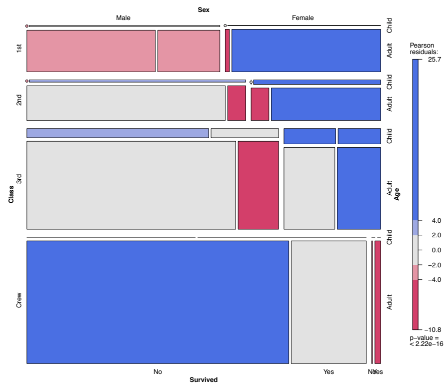
Principles / advice / no solution

Mosaicplots
(and other displays)
of the Titanic dataset

Mosaicplot examples



Titanic



Titanic — independence model

Mosaicplots

- Variable category combinations are represented by rectangles. There are gaps between rectangles (ideally smaller by level).
- Rectangle area is (almost always) proportional to frequency. Rectangles may have equal width (height), so that height (width) is proportional to frequency.
- Rectangles may be aligned in various ways, rotated, coloured.
- Mosaicplots need to be interactive.

Titanic dataset

2201 passengers and crew by gender, age (child/adult), ship's class (1st, 2nd, 3rd, crew), survived or died.

(R. J. MacG. Dawson, J. Statistics Education 3 no 3, 1995)

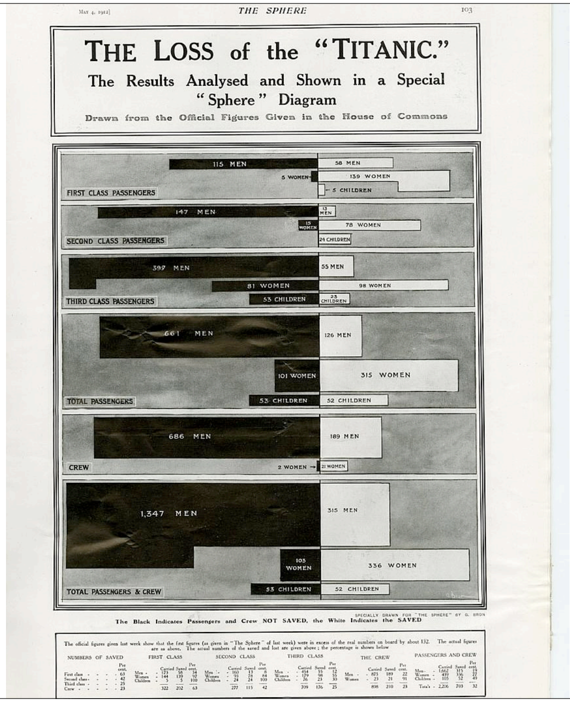
Aside: Titanics in R

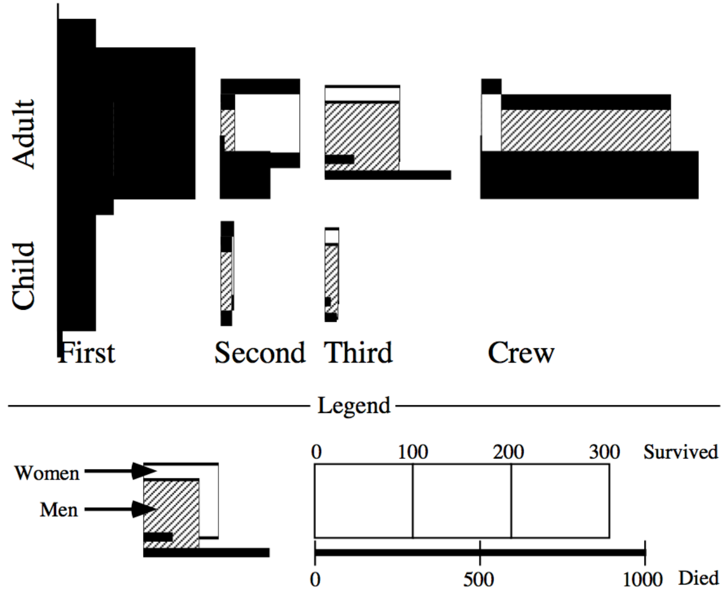
- Titanic {*datasets*}, Titanic {*effects*}
- TitanicMat {*RelativeRisk*}
- titanic {*alr4*}, titanic {*prLogistic*}, titanic {*msme*}
- titanic.dat {*exaxtLoglinTest*}
- titan.Dat {*elrm*}
- etitanic {*earth*}
- ptitanic {*rpart.plot*}
- Lifeboats {*vcd*}

So take care!

and please treat datasets with respect

and maybe there are more I have not found ....





Bertin “Semiologie Graphique”

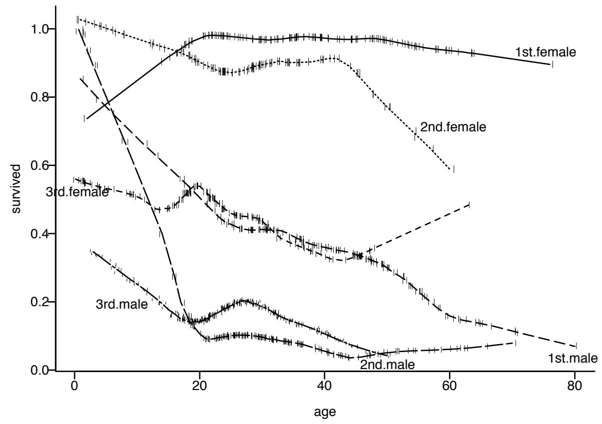
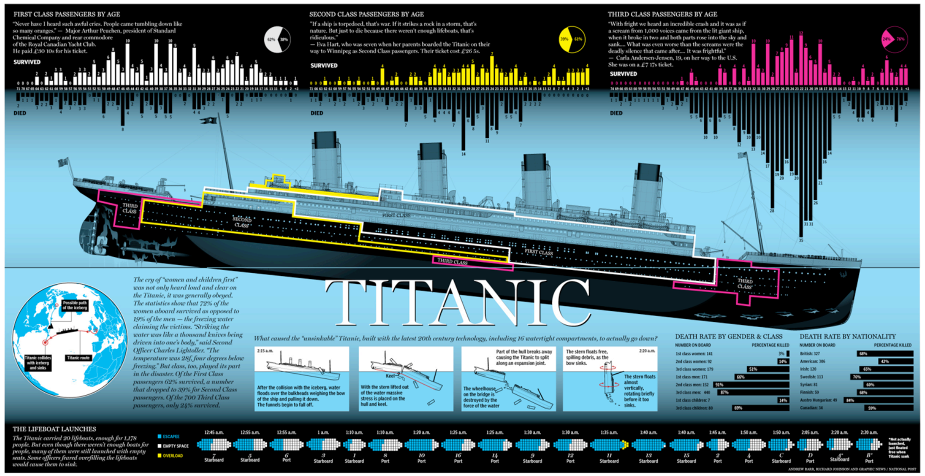
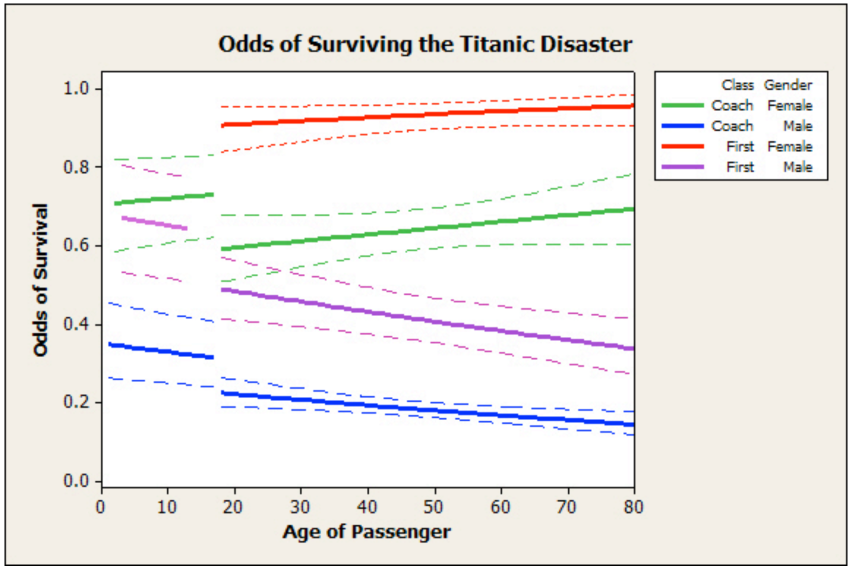


Figure 5.1: `loess` *smoothed estimates of the probability of surviving the* Titanic *as a function of passenger age, sex, and ticket class*

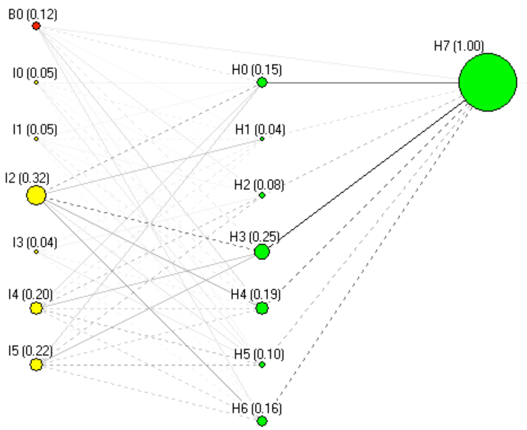
Harrell “Statistical Graphics Course”





The Minitab Blog

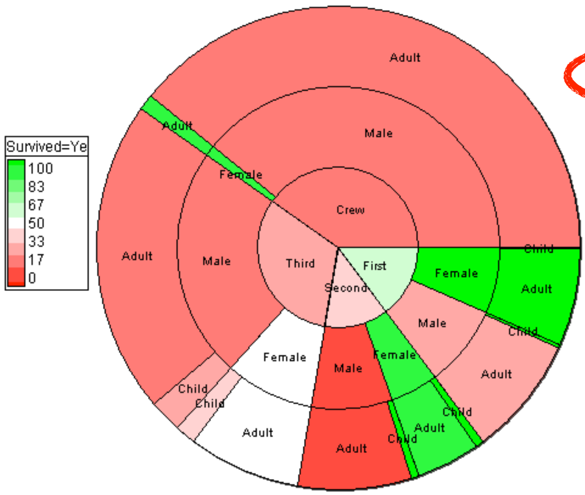
The *really* interesting thing about the neural network's solution to the problem isn't that it reached any kind of useful error rate (which it didn't), but that its weights encode the relative importances of the various inputs.



We can see that the most useful inputs for survival prediction were "3rd Class", "Age", and "Sex".

Logical Genetics blog

Using Colour to Show Survival Rates



A number of things can be seen at a glance:

Female survival rate was higher than males within every passenger type

Survival rates for passenger class was in the order First, Second, Third, and finally Crew

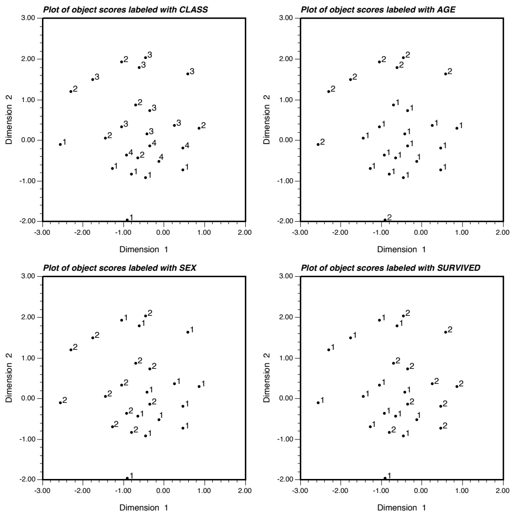
First class, female, adult passenger survival rate was close to 100% (actually 97.22%)

Male children in third class berths had a survival rate comparable to crew (around 25%)

Within almost all Class/Gender combinations, children had higher surivial rates than adults

The worst survival rate was for second class adult males (around 8%)

Neoformix blog



The configuration clearly distinguishes two groups, the outer points representing children and the people scattered around the origin representing adults (upper right panel of Figure 2.4). One reason for this result is that there were not many children on board of the ship, and objects of low frequency tend to be located in the periphery of the plot.

The lower right panel of Figure 2.4 shows that most survivors are located on the left-hand side of the first axis. Since all women (lower left panel of Figure 2.4) as well as most children (upper right panel of Figure 2.4) are located in this very same area, the PIONEER analysis shows that the rule 'women and children first' seems to have been applied in the rescue operation of the sinking Titanic.

Finally, there is a slight indication that first and second class passengers were rescued more often then the others (upper left panel of Figure 2.4).

Figure 2.4. Labeled profile scores plots of PIONEER analysis of Titanic survival data.

Patrick J.F. Groenen, Jacques J.F. Commandeur
Department of Data Theory, Leiden University

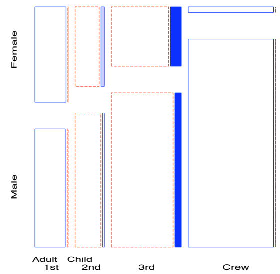


Figure 9: Titanic data, background variables

Figure 9 shows the frequencies of the background variables, Class, Gender and Age by the sizes of the boxes. It also shows the association between Age and Class–Gender combinations by shading. There were no children among the crew, and the overall proportion of children was quite small (about 5 %). But among the passengers, the proportion of children increases from first class to third class. The large positive residuals for children among the 3rd class passengers likely represents families traveling or emmigrating together.

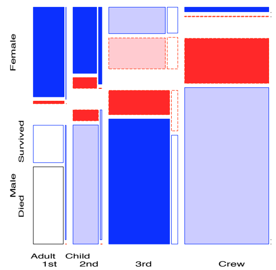


Figure 10: Titanic data, Joint independence: Survival ⊥ {Class, Gender, Age}

Figure 10 shows an initial four-way mosaic for the full table, and fits the model which asserts that survival is independent of Class, Gender, and Age jointly. This is the minimal null model when the first three variables are explanatory. It is clear that greater proportions of women survived than men in all classes, but with greater proportions of women surviving in the up- per two classes. Among males, the proportion who survived also increases with economic class. However, this model fits very poorly ($G^2(15) = 671.96$), and we may try to fit a more adequate model by adding associations between survival and the explanatory vari- ables.

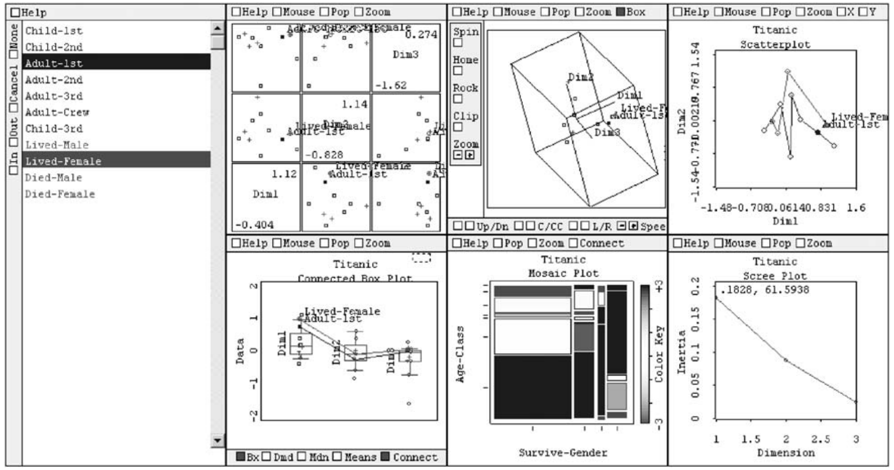
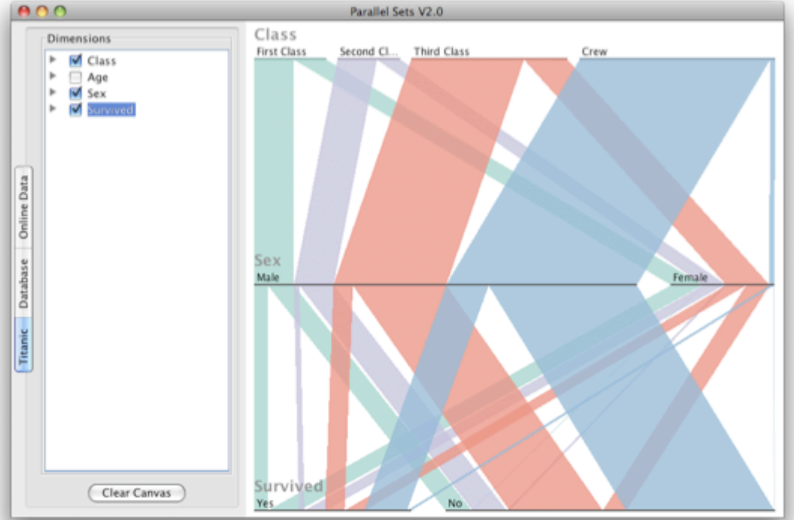


Fig. 2. Spreadplot for correspondence analysis.

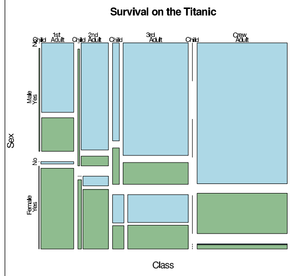
Visualizing categorical data in ViSta

Pedro Valero-Mora, Forrest W. Young, Michael Friendly

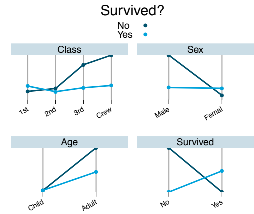


eagereyes blog (Kosara)

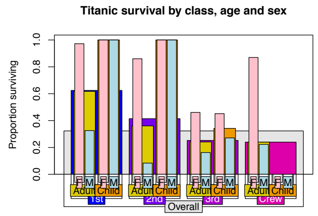
and plots for Titanic in R ...



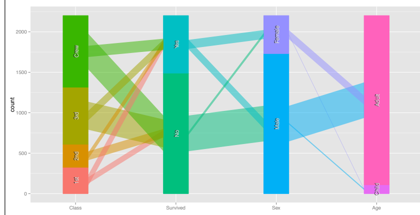
Ihaka "Statistics 120 Mosaic Plots"



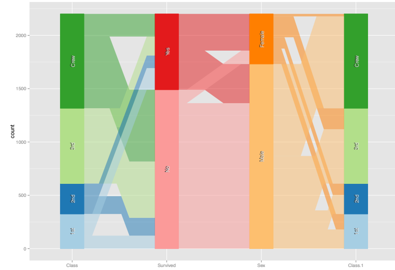
Marginal plot (*latticeExtra*)



barnest (*plotrix*)



Hammock plot (*ggparallel*)



Angle plot (*ggparallel*)

Titanic models (Simonoff)

Table of models

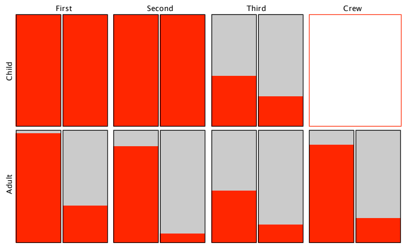
Model	G	Predictors	D	H–L p	AIC_C
G	434.5	1	.40		−430.5
E	180.9	3	.28		−172.9
A	19.6	1	.05		−15.6
E, G	540.5	4	.49	.000	−530.6
A, G	440.4	2	.42	.000	−434.9
E, A	206.5	4	.30	.001	−196.5
E, A, G	559.4	5	.52	.000	−547.4
E, G, EG	605.7	7	.50		−589.7
A, G, AG	456.7	3	.43		−448.7
E, A, EA	235.7	6	.30		−221.7
E, A, G, EG	626.1	8	.53	.941	−608.0
E, A, G, EA	595.1	7	.52	.000	−579.0
E, A, G, AG	577.4	6	.52	.000	−563.4
E, A, G, EA, EG	670.3	10	.53	.944	−648.2
E, A, G, EG, AG	634.7	9	.53	.992	−614.6
E, A, G, EA, AG	606.9	8	.53	.000	−588.9
E, A, G, EA, EG, AG	672.0	11	.54	1.000	−647.9

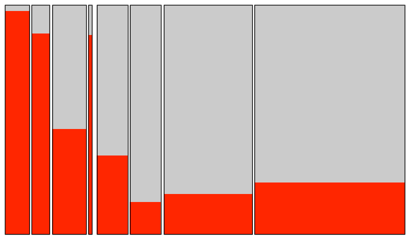
Gender*Class + Age model

```
Logistic Regression Table
                                                       Odds       95% CI
Predictor              Coef       StDev         Z       P   Ratio    Lower    Upper
Constant             1.8971      0.6191      3.06 0.002
Economic
 First class         1.6608      0.8003      2.08 0.038    5.26     1.10    25.26
 Second class       -0.0199      0.6869     -0.03 0.977    0.98     0.26     3.77
 Third class        -2.2247      0.6370     -3.49 0.000    0.11     0.03     0.38
Age grou
 Child               1.0537      0.2304      4.57 0.000    2.87     1.83     4.51
Gender
 Male               -3.1469      0.6245     -5.04 0.000    0.04     0.01     0.15
Economic*Gender
 First class*Male   -1.0862      0.8197     -1.33 0.185    0.34     0.07     1.68
 Second class*Male  -0.6379      0.7250     -0.88 0.379    0.53     0.13     2.19
 Third class*Male    1.7763      0.6522      2.72 0.006    5.91     1.65    21.21
```


Mosaicplots for the Titanic in Mondrian

Two of the alternatives





Equal binsize plot
Titanic survival rates by class gender, age
Women to the left, men to the right

Doubledecker plot
Titanic survival rates by class and gender
Women to the left, men to the right

Categorical data and visualizations thereof

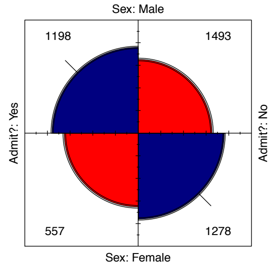
Categorical data

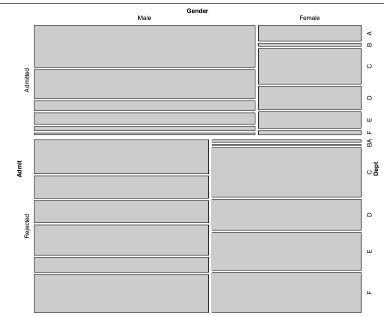
- Nominal data (perhaps grouped, e.g. geographically)
  – Occupation, Experimental treatments, Cities, …
- Binary or logical data
  – Gender, Yes/No, True/False, …
- Ordinal data
  – Survey responses, Income group, Fitness, …
- Discrete data (and discretised continuous data)
- Examples: Titanic, Rochdale, Divorce, Bowling Alone, …

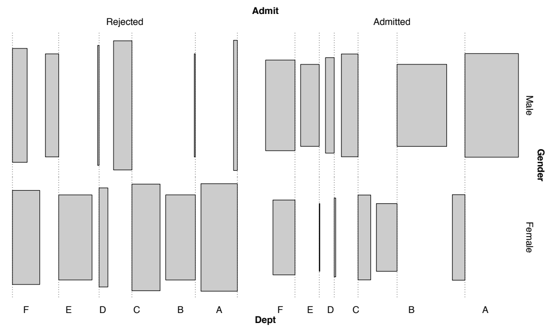
Plots for categorical data

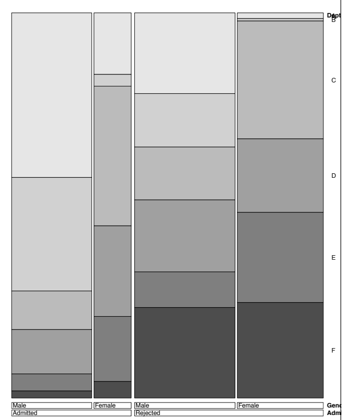
- barcharts, stacked barcharts, dodged barcharts
- piecharts, agreement plots (Bangdiwala)
- fourfold displays, sieve diagrams, association plots, cpcp
- 3-d and trellis versions
- mosaicplots
- familes of plots in R
  – strucplot (Meyer, Zeileis, Hornik) *vcd* package
  – productplots (Wickham, Hofmann) *productplots* package

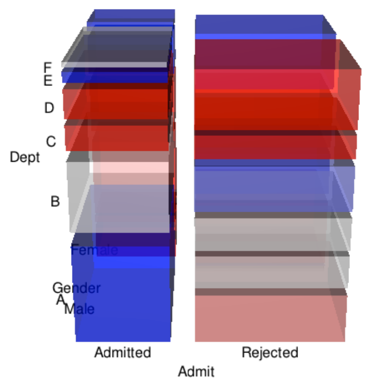
Plots from R



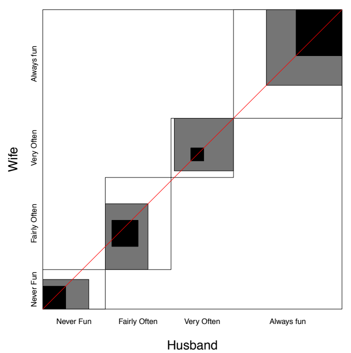








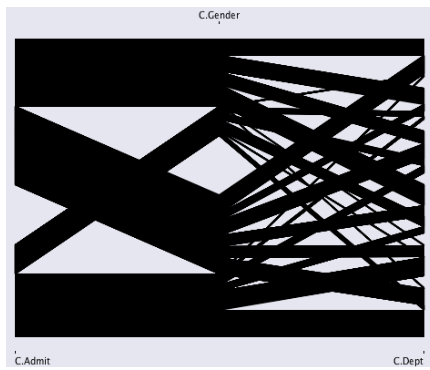
3D mosaicplot (*vcdExtra*)



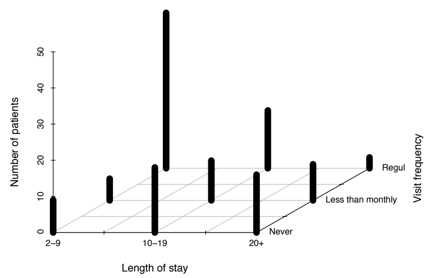
Bangdiwala agreementplot (*vcd*)

For those of a sensitive disposition I would suggest avoiding Figure 13.8, which apparently shows a three-dimensional mosaicplot. Up to this point in the book I had agreed with Paul Murrell's statement in his Preface that "no plot type is all bad".

@Antony: Since this is a Fig, I take it that you are prepared to argue that fluctuation diagrams are an alternative (maybe better) display for agreements. I caution you not to say that in print, because you would be wrong... Michael Friendly



cpcp plot (*extracat*)



3D barchart (Meyer)

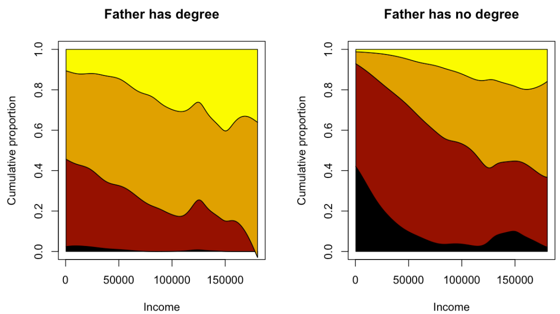
Mosaicplot variants

Another dataset?

- Possibilities
  - *wong.df* in James Curran’s *dafs*
  - *wffc* in Thomas Yee’s *VGAM*
  - *nhanes* in Thomas Lumley’s *survey*
  - *diabetes, murder, NORC, Auckland* .... in Stats 330

- Decided to use *Intergenerational inequality* from StatsChat
  - GSS (USA) with education, parents’ education, age, sex, family income, survey weighting, ...

Intergenerational inequality



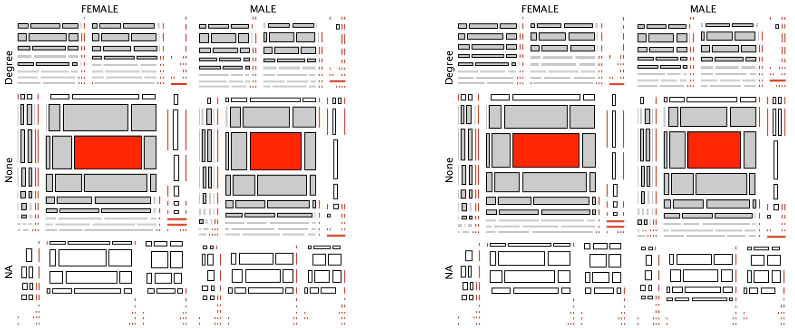
In each panel, black is less than high school, dark red is high school, light brown is university or junior college and yellow is postgraduate. These are plotted by family income (in inflation-adjusted US dollars).

www.statschat.org.nz/2013/03/25/intergenerational-inequality/

Inequality topics

- What other factors might be relevant?
  - mother’s degree, age, sex, interview year ....
- What about missing values and data quality?
  - "face-to-face interview over 90 minutes"
- How much data are there in the various groupings?
  - are the irregularities in the smooths real?
  - what would you expect?
- ...

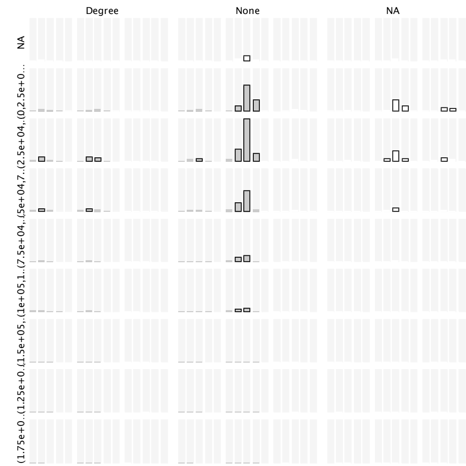
Classical mosaicplots



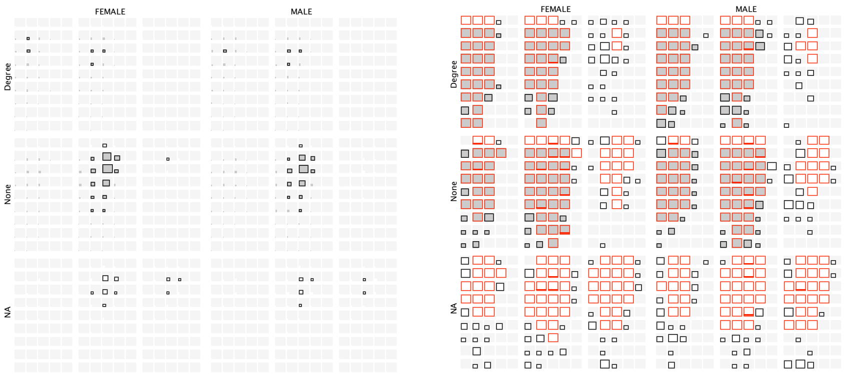
Sex, father's ed, mother's ed, family income, education
left: unweighted right: weighted

Multiple barcharts

Father's ed, family income, mother's ed, education

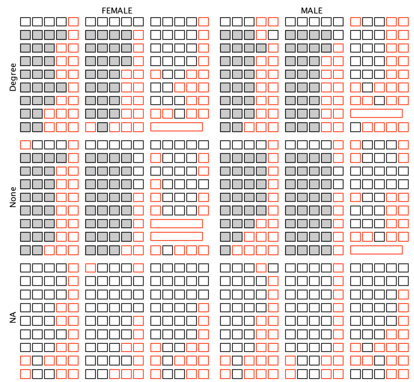


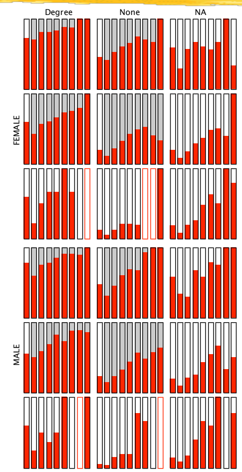
Fluctuation diagrams



Sex, father's ed, mother's ed, family income, education
right: ceiling censored zoomed, high weight cases selected

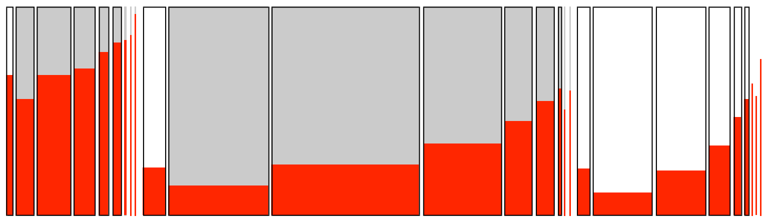
Same binsize





More interesting with highlighting here

Doubledecker plots



Education university or graduate
by income group within father's education
(All splits are horizontal)

Mosaicplots also include

- Residual plots (by expected / observed, association plots)
- Relative multiple barcharts
- Multiple spineplots
- Treemaps

though no single software package implements all (yet)

Mosaicplot options

- Choice of variables
- Order of variables
- Whether each variable is horizontal or vertical
- Form of mosaicplot
- Orders of categories within nominal variables, ordering direction for ordinal variables
- Grouping categories
- Display options: spacing between levels and between categories, plot size, aspect ratio, colour

Numbers of mosaicplots

		Titanic
≥ 8	Variants	8
$2^m - 1$	Choice from *m* variables	7
r!	Orderings of *r* variables	24
2^r	Directions of variables (horizontal / vertical)	16
$\prod c_j !$	Orderings of categories within variables (or 2^r for direction of ordinal variables)	24*2*2*2
?	Aggregations of categories	...
?	New derived variables	...

Choices and principles

Variant choices

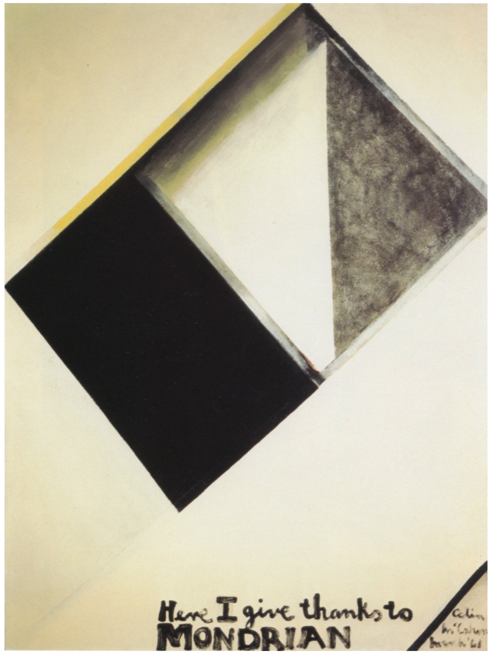
- Classical mosaicplots: for cumulative rates
- Residual plots: for supporting model building
- Fluctuation diagrams: for sparse structures
- Multiple barcharts: for non-binary target variables
- Same binsize: for rates across all groups and missing values
- Doubledecker plots: for rates across all groups with cell sizes
- Relative multiple barcharts: for distribution shape
- Treemaps: for splitting by different variables

Design principles (1)

- Variable ordering
  - Target variable should usually be last
  - Binary target variables are best displayed using linking
  - A grouping variable should be first, possibly rotated
  - Comparisons and context determine the order of conditioning variables (+ overall height/width)
  - Variables with unequal distributions are better early (?)

Design principles (2)

- Category ordering
  - can be determined by context
  - by what you want to compare
  - can sort by count, absolute proportions, relative proportions
- Vary aspect ratio of cells
  - square (fluctuation diagrams), otherwise height > width
- One graphic is usually not enough



Mondrian

- Mondrian for interactive graphical analysis
  - one of the Augsburg Impressionists
  - stats.math.uni-augsburg.de/Mondrian/
  - for Windows, Unix, MacOS
  - by Martin Theus

Summary

- Categorical data are difficult to visualise
- Several related plots are more effective than one single plot
- Mosaicplots are a general, flexible family of displays for categorical data though
  - they are often puzzling to interpret
  - and they need thoughtful design
  - so sometimes they seem more like mind games