Pie Charts

- In the period 1910 – 1920 there was a great deal of discussion or the relative merits of pie charts and divided bar charts in the *Journal of the American Statistical Society*.
- Eventually consensus was reached that divided bar charts were a superior way of presenting proportions.
- Since 1980, the study of graphical of graphical perception has revealed why bar charts are preferable.

  Humans are much better at decoding numbers presented in the form of lengths or positions than they are at decoding numbers presented as angles or areas.

Comments on Bar Charts I

Becker R., and Cleveland W. S. (1996).
The Splus Trellis Graphics User Manual.
Page 50.

> *Pie charts have severe perceptual problems. Experiments in graphical perception have shown that compared with dot charts, they convey information much less reliably. But if you want to display some data, and perceiving the information is not so important, then a pie chart is fine.*

Bill Cleveland is one of the world's foremost authorities on how information is extracted from graphs.

Comments on Pie Charts II

Tufte, E. (1983).
The Visual Display of Quantitative Information.
Page 178.

> *A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies . . . Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.*

Ed Tufte was Professor of Statistics, Political Science and Graphic Design at Yale University. He has written some of the best-selling books on information display.

Comments on Pie Charts III

Bertin, J. (1981).
Graphics and Graphic Information Processing.
Page 111.

> Bertin describes multiple pie charts as "*completely useless.*"

Jarques Bertin is one of the major names in *semiotics* (the study of signs). He has written a number of very influential books on graphical presentation.

Comments on Pie Charts IV

The Energy Information Agency (EIA) is part of the U.S. Department of Energy and is charged with compiling and disseminating information about energy to the government and private sectors.

EIA maintains a large standards manual for graphical presentation.

`http://www.eia.doe.gov/neic/graphs/preface.htm`

Comments on Pie Charts IV

- William Eddy of Carnegie-Mellon University, formerly vice chair of the American Statistical Association (ASA) Committee on Energy Statistics, said of pie charts at the April 1988 ASA committee meetings in a session on the EIA Standards Manual, "*death to pie charts.*"
- Howard Wainer of the Educational Testing Service stated in a 1987 *Independent Expert Review of EIA Statistical Graphs Policies* that "*the use of pie charts is almost never justified*" and that they "*ought not to be used.*" Wainer recommended to EIA that dot charts be used instead of pie charts in EIA products.

Comments on Pie Charts V

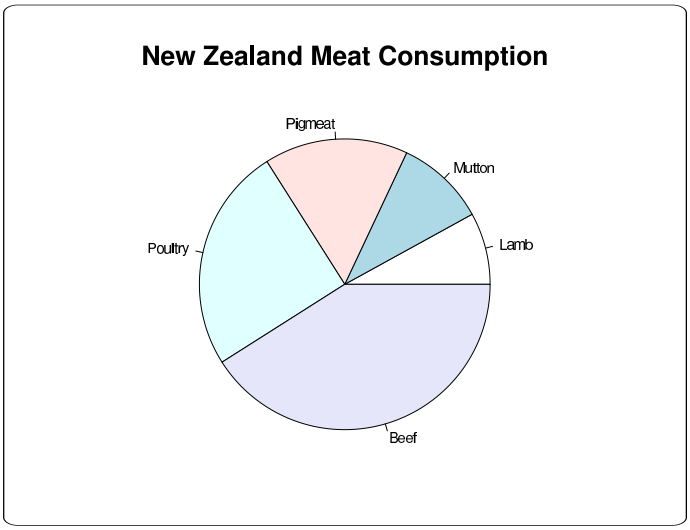
- During revision for the STAT 120 (Information Visualisation) exam in 2002, Ross Ihaka said:

  If you want to fail this course, just show me a pie chart.

Drawing Pie Charts with R

A basic pie chart is produced from a vector of named values. such a vector can be created as follows:

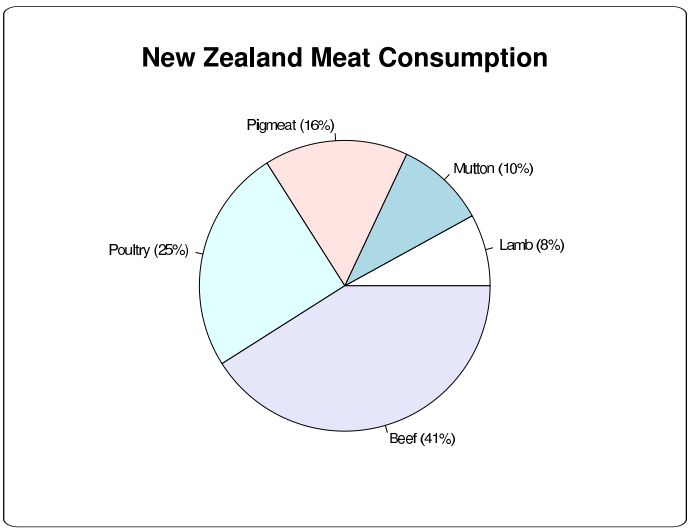
```
> meat = c(8, 10, 16, 25, 41)
> names(meat) = c("Lamb",
                  "Mutton",
                  "Pigmeat",
                  "Poultry",
                  "Beef")
> pie(meat,
      main = "New Zealand Meat Consumption",
      cex.main = 2)
```



Annotating Pie Charts

Because it is so hard to decode values from pie charts, it is common to include the values as text in the plot.

```
> meat = c(8, 10, 16, 25, 41)
> names(meat) = c("Lamb",
                  "Mutton",
                  "Pigmeat",
                  "Poultry",
                  "Beef")
> pie(meat,
      labels = paste(names(meat),
                     " (", meat, "%)",
                     sep = ""),
      main = "New Zealand Meat Consumption",
      cex.main = 2)
```



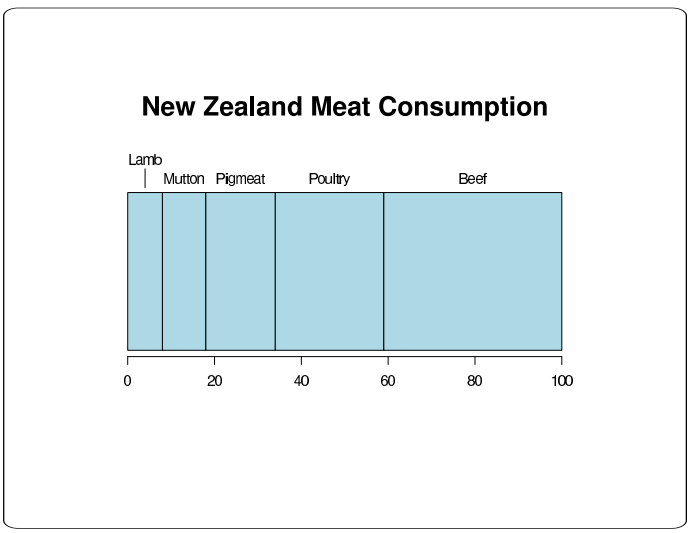
A Tabular Representation

In this case, the information is easier to extract from a table than from a pie chart.

New Zealand Meat Consumption

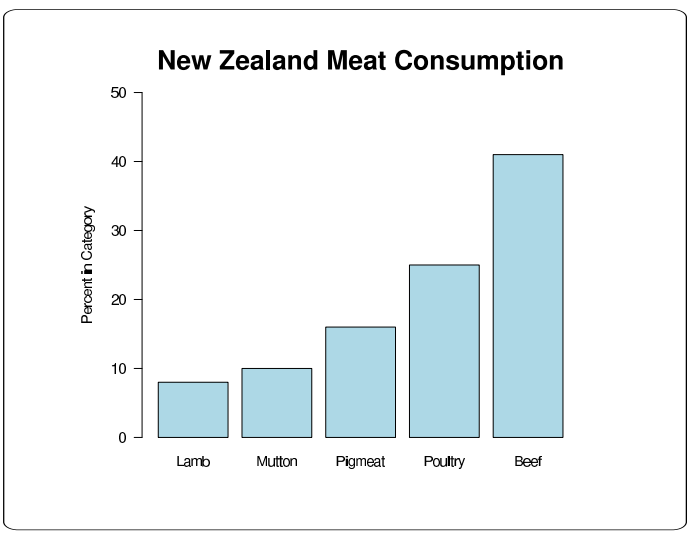
Lamb	8%
Mutton	10%
Pigmeat	16%
Poultry	25%
Beef	41%

(This table has deliberately been kept simple. No boxes or lines have been used.)



A Simple Bar Chart

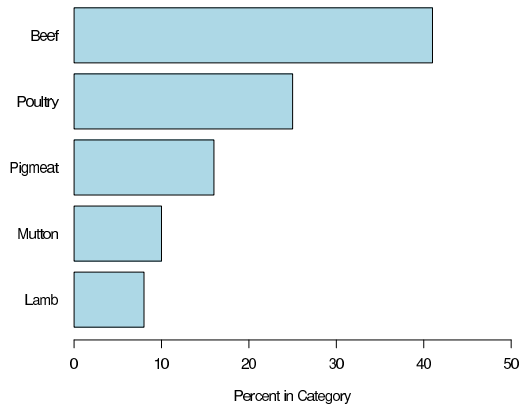
```
> barplot(meat, ylim = c(0, 50),
          col = "lightblue",
          main = "New Zealand Meat Consumption",
          ylab = "Percent in Category",
          cex.main = 2, las = 1)
```



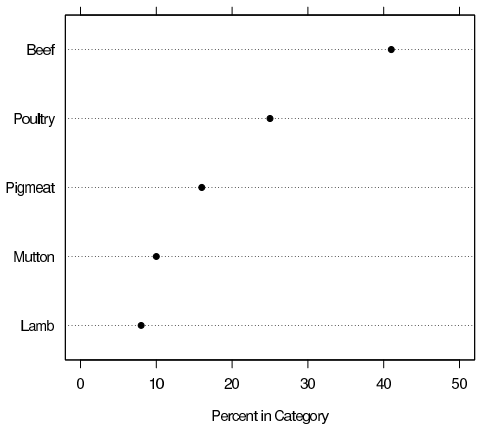
A Horizontal Bar Chart

```
> barplot(meat, xlim = c(0, 50),
          col = "lightblue",
          main = "New Zealand Meat Consumption",
          xlab = "Percent in Category",
          cex.main = 2, las = 1,
          horiz = TRUE)
```

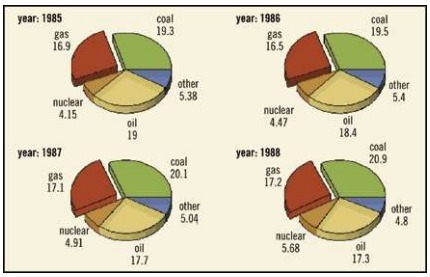
New Zealand Meat Consumption



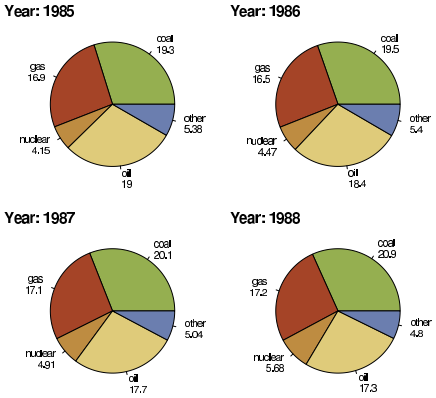
New Zealand Meat Consumption



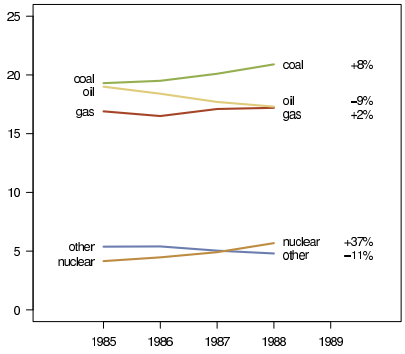
United States Energy Production I



United States Energy Production II



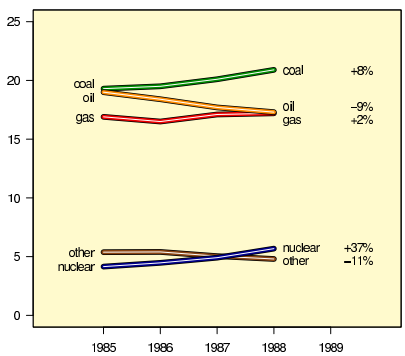
United States Energy Production III



Decorating Plots

- One common complaint about R is that the plots it produces are "plain" or "boring."
- In fact, if you are prepared to put a little effort in, you can produce a wide variety of "interesting" effects.
- Of course, there is no substitude for having a graph which shows that something interesting is going on.

United States Energy Production IV



Filling a Plot's Background

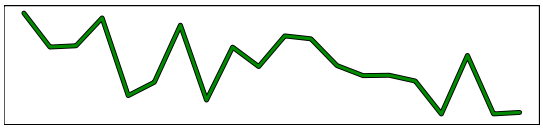
Colouring the background of the plot region is simple. After setting the up the axis scales, determine the coordinates of the edges of the plotting region and draw a filled rectangle which fills the area completely.

```
plot.new()
plot.window(xlim = xlimits, ylim = ylimits)
usr = par("usr")
rect(usr[1], usr[3], usr[2], usr[4],
     col = "lemonchiffon")
```

Thick Lines

Thick lines can be drawn by first drawing the lines n units wide in black and then drawing them $n - 3$ units wide in the fill colour. This works for all colours.

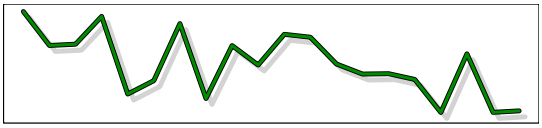
```
> lines(x, y, lwd = 8, col = "black")
> lines(x, y, lwd = 5, col = "green4")
```



Drop Shadows

A drop shadow effect can be obtained by first drawing the line in gray, offset down and to the left, and then drawing the line itself.

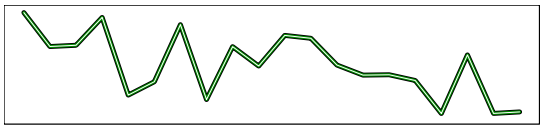
```
> lines(x + xinch(.1), y - yinch(.1),
        lwd = 8, col = "lightgray",
        border = NA)
> lines(x, y, lwd = 8, col = "black")
> lines(x, y, lwd = 5, col = "green4")
```



Specular Reflections

It is also possible to create a three dimensional look by adding what appears to be a specular highlight.

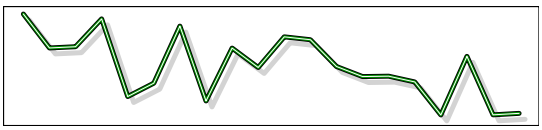
```
> lines(x, y, lwd = 8, col = "black")
> lines(x, y, lwd =  5, col = "green4")
> lines(x, y, lwd =  1, col = "white")
```



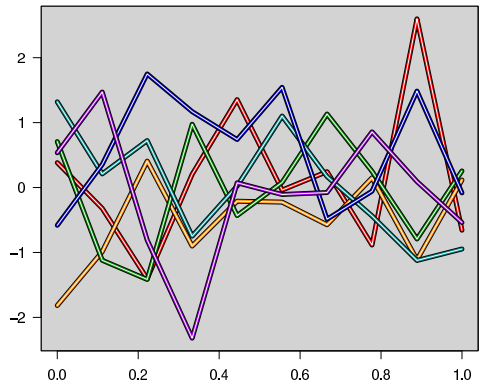
Combined Effects

It is of course possible to include all three of these effects in in a single graph.

```
> lines(x + xinch(.1), y - yinch(.1),
        lwd = 8, col = "lightgray",
        border = NA)
> lines(x, y, lwd = 8, col = "black")
> lines(x, y, lwd =  5, col = "green4")
> lines(x, y, lwd =  1, col = "white")
```

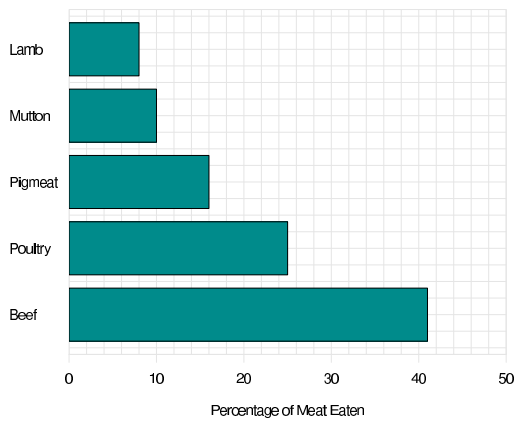


Spaggetti Anyone?

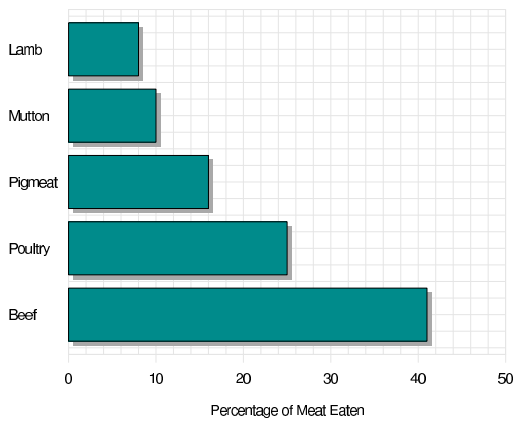


A Useful Set of Line Colours

New Zealand Meat Consumption



New Zealand Meat Consumption



Mosaic Plots

- Hartigan, J. A., and Kleiner, B. (1981), "Mosaics for contingency tables," In W. F. Eddy (Ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*. New York: Springer-Verlag.
- Hartigan, J. A., and Kleiner, B. (1984) "A mosaic of television ratings." *The American Statistician*, **38**, 32–35.
- Friendly, M. (1994) "Mosaic displays for multi-way contingency tables." *Journal of the American Statistical Association*, **89**, 190–200.

Who Listens To Classical Music?

The following table of values shows a sample of 2300 music listeners classified by age, education and whether they listen to classical music.

	Education			
	High		*Low*	
	Classical Music			
Age	*Yes*	*No*	*Yes*	*No*
Old	210	190	170	730
Young	194	406	110	290

This is a $2 \times 2 \times 2$ contingency table.

Old Versus Young

The effect of age and education on muscial taste can be investigated by breaking the observations down into more homogenous groups. The most obvious split is by age. There are 1300 older people and 1000 younger people.

Old	*Young*
56.5%	43.5%

This is almost certainly a result of the way in which the sample was taken.

Education Level

Within the old and young groups we can now find the proportions falling into each of the high and low education categories.

Old		*Young*	
High Ed.	*Low Ed.*	*High Ed.*	*Low Ed.*
30.8%	69.2%	60.0%	40.0%

The *young* group is clearly more highly educated than the *old* group.

Music Listening

Finally, we can compute the proportion of people who listen to classical music in each of the age/education groups.

Old		*Young*	
High Ed.	*Low Ed.*	*High Ed.*	*Low Ed.*
52.5%	18.9%	32.3%	27.5%

The music-listening habits of younger people seem to be fairly independent of education level. This is not true for older people.

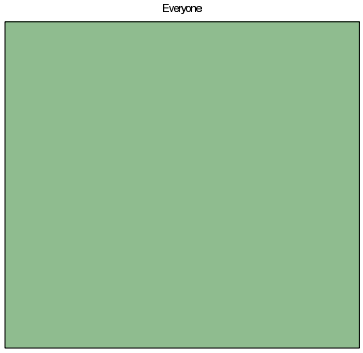
Summary

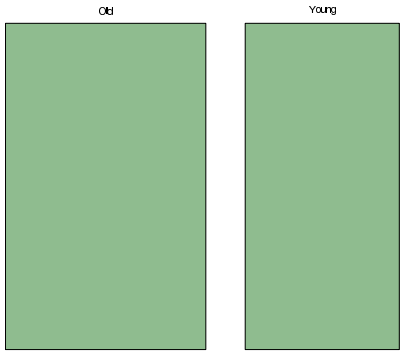
The result of our "analysis" is a series of tables. From these tables we can see:

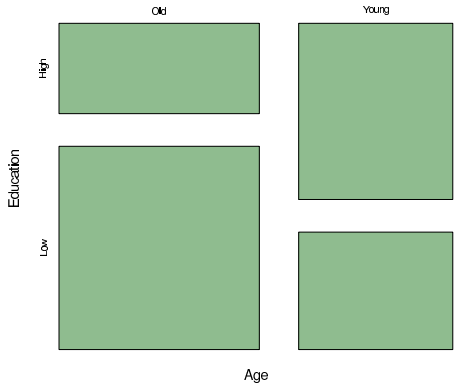
- There are slightly more old people than young people in the sampled group.
- The younger people are more highly educated than the older ones.
- The likelihood of listening to classical music depends on both age and education level.

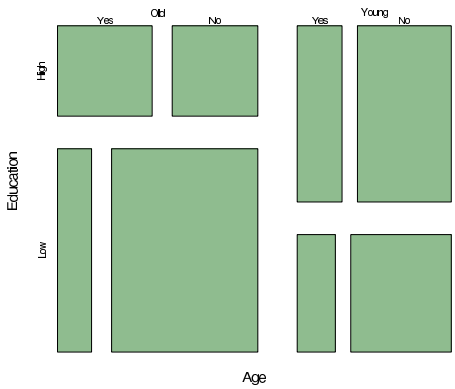
Mosaic Plots

- Mosaic plots give a graphical representation of these successive decompositions.
- Counts are represented by rectangles.
- At each stage of plot creation, the rectangles are split parallel to one of the two axes.









The Perceptual Basis for Mosaic Plots

- It is tempting to dismiss mosaic plots because they represent counts as rectangular areas, and so provide a distorted encoding.
- In fact, the important encoding is length.
- At each stage the comparison of interest is of the lengths of the sides of pieces of the most recently split rectangle.

Creating Mosaic Plots

- In order to produce a mosaic plot it is neccessary to have:
  - A contingency table containing the data.
  - A preferred ordering of the variables, with the "response" variable last.

Data Entry

```
> music = c(210, 194, 170, 110,
            190, 406, 730, 290)

> dim(music) = c(2, 2, 2)

> dimnames(music) =
      list(Age = c("Old", "Young"),
           Education = c("High", "Low"),
           Listen = c("Yes", "No"))
```

Data Inspection

```
> music
, , Listen = Yes

       Education
Age     High Low
  Old    210 170
  Young  194 110

, , Listen = No

       Education
Age     High Low
  Old    190 730
  Young  406 290
```

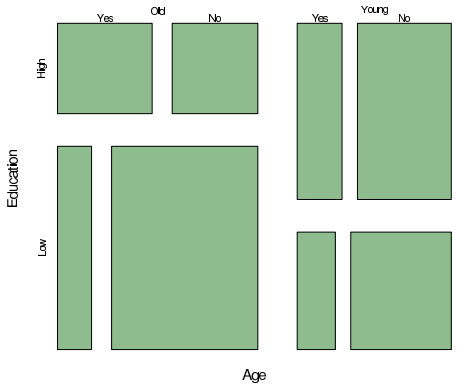
Producing A Mosaic Plot

The R function which produces mosaic plots is called `mosaicplot`. The simplest way to produce a mosaic plot is:

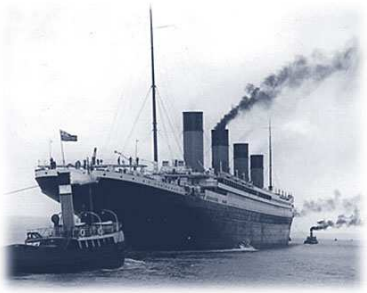
```
> mosaicplot(~ Age + Education + Listen,
             data = music)
```

It is also easy to colour the plot and to add a title.

```
> mosaicplot(~ Age + Education + Listen,
      data = music,
      col = "darkseagreen",
      main = "Classical Music Listening")
```



Example: Survival on the Titanic



On Sunday, April 14th, 1912 at 11:40pm, the RMS Titanic struck an iceberg in the North Atlantic. Within two hours the ship had sunk. At best reckoning 705 survived the sinking, 1,523 did not.

The Data

- There is very good documentation on who survived and who did not survive the sinking of the Titanic.
- R has a data set called "Titanic" which gives data on the passengers on the Titanic, cross-classified by:
  - Class: 1st, 2nd, 3rd, Crew.
  - Sex: Male, Female.
  - Age: Child, Adult.
  - Survived: No, Yes.

Adults	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	57	140	118	4
2nd Class	14	80	154	13
3rd Class	75	76	387	89
Crew	192	20	670	3

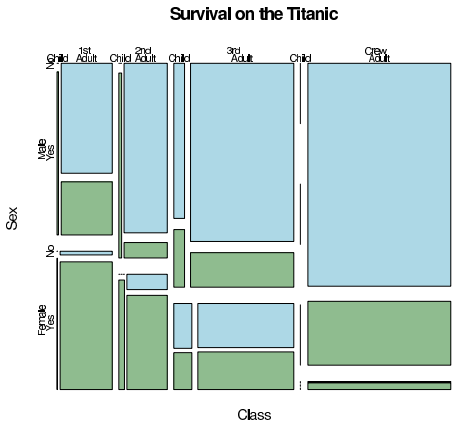
Children	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	5	1	0	0
2nd Class	11	13	0	0
3rd Class	13	14	35	17
Crew	0	0	0	0

Producing a Mosaic Plot

The following command produces the mosaic.

```
> mosaicplot(~ Class + Sex + Age + Survived,
             data = Titanic,
             main = "Survival on the Titanic",
             col = c("lightblue", "darkseagreen"),
             off = c(5, 5, 5, 5))
```

Note the use of `col=` to produce alternating coloured rectangles — green for survivors and blue for non-survivors. Also note that the `off=` argument is used to squeeze out a little of the space between the blocks.



Example: Sexual Discrimination at Berkeley

- In the 1980s, a court case brought against the University of California at Berkeley by women seeking admission to graduate programs there.
- The women claimed that the proportion of women admitted to Berkeley was much lower than that for men, and that this was the result of discimination.

Gender	*Admitted*	*Rejected*	*%Admitted*
Male	1198	1493	44.5
Female	557	1278	30.4

- It is clear that a higher proportion of males is being admitted.

The University Case

The Dean of Letters and Science at Berkeley was a famous statistician (called Peter Bickel) and he was able to argue that the difference in admissions rates was not caused by sexual discrimination in the Berkeley admissions policy, but was caused by the fact that males and females generally sought admission to different departments.

The Dean broke the admissions data down by department and showed that within each program there was no admission discrimination against women. Indeed, there seemed to be some admissions bias in favour of women.

		Admitted	*Rejected*	% Admitted
Department A	*Male*	512	313	62
	Female	89	19	82
Department B	*Male*	353	207	63
	Female	17	8	68
Department C	*Male*	120	205	37
	Female	202	391	34
Department D	*Male*	138	279	33
	Female	131	244	35
Department E	*Male*	53	138	28
	Female	94	299	24
Department F	*Male*	22	351	6
	Female	24	317	7

Producing The Berkeley Mosaic

We relabel the Admit/Reject levels so that the labels will fit across the plot.

```
> x = UCBAdmissions
> dimnames(x)[[1]] = c("Ad", "Rej")
> mosaicplot(~ Dept + Gender + Admit,
      data = x,
      col = c("darkseagreen", "pink"),
      main = "Student Admissions at UC Berkeley")
```