lab 7: Simpson's paradox and mosaic plots

The aim of this lab session is to look at multi-way **contingency tables (count data)**, and to practice using **mosaic plots** to visualise it.

Very often one needs to understand a table of count data, in various categories. For example,

- thousands of students may apply to a university each year;
- each applicant is accepted or rejected;
- each applicant applies to a particular department; -and each applicant is male or female.]

We then have a 3-dimensional table of counts. We can print this as a table, **but is there a way to visualise it?**

Tables of counts of this type are called **contingency tables**, and they crop up all the time. How to visualise them?

One way to visualize such contingency tables is to use **mosaic plot**.

A mosaic plot

- is a special type of stacked bar chart that shows percentages of data in groups.
- The plot is a graphical representation of a contingency table.
- The mosaic plot resides in the category of visualizations that feature part-to-whole relationships.

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.
- At each stage the comparison of interest is of the lengths of the sides of pieces of the most recently split rectangle.

We will look at two celebrated data-sets of counts:

- applicants to the University of California at Berkeley in 1973,
- and the passengers and crew of the Titanic.

The UC Berkeley (UCB) admissions data

In 1973 the admissions counts for UCB were as follows (see the dataset UCBAdmissions)

```
head(UCBAdmissions)
```

```
## , , Dept = A
##
##           Gender
## Admit      Male Female
##   Admitted  512     89
##   Rejected  313     19
##
## , , Dept = B
##
##           Gender
## Admit      Male Female
##   Admitted  353     17
##   Rejected  207      8
##
## , , Dept = C
##
##           Gender
## Admit      Male Female
##   Admitted  120    202
##   Rejected  205    391
##
## , , Dept = D
##
##           Gender
## Admit      Male Female
##   Admitted  138    131
##   Rejected  279    244
##
## , , Dept = E
##
##           Gender
## Admit      Male Female
##   Admitted   53     94
##   Rejected  138    299
##
## , , Dept = F
##
##           Gender
## Admit      Male Female
##   Admitted   22     24
##   Rejected  351    317
```

What are the comparative fractions of men and women who are admitted? Do you think you need to do a significance test?

The university was taken to court for discriminating against women in the admissions process in one of the first cases of its kind. Do you think they were guilty? Can you think of other explanations for these figures?

```
# The UC Berkeley (UCB) admissions data
#install.packages("vcd")
#install.packages("vcdExtra")
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 4.0.5
```

```
## Loading required package: grid
```

```
library(vcdExtra)
```

```
## Loading required package: gnm
```

```
library(gcookbook)
library(datasets)
```

`mosaic()`

- provides a wide range of options for the directions of splitting,
- the specification of shading,
- labeling,
- spacing,
- legend and many other details.

`mosaic(formula, data, highlighting = NULL, highlighting_fill =  rev(gray.colors(tail(dim(x), 1))), direction = NULL, ….)`

`formula` : a formula specifying the variables used to create a contingency table from data. For convenience, conditioning formulas can be specified; the conditioning variables will then be used first for splitting. If any, a specified response variable will be highlighted in the cells.

`highlighting` : character vector or integer specifying a variable to be highlighted in the cells.

`highlighting_fill` : color vector or palette function used for a highlighted variable, if any.

`direction` : character vector of length k, where k is the number of margins of x (values are recycled as needed). For each component, a value of `"h"` indicates that the tile(s) of the corresponding dimension should be split horizontally, whereas `"v"` indicates vertical split(s).

For example,

```
summary(UCBAdmissions)
```

```
## Number of cases in table: 4526
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 2000.3, df = 16, p-value = 0
```

```
dim(UCBAdmissions)
```

```
## [1] 2 2 6
```

```
names(UCBAdmissions)
```

```
## NULL
```

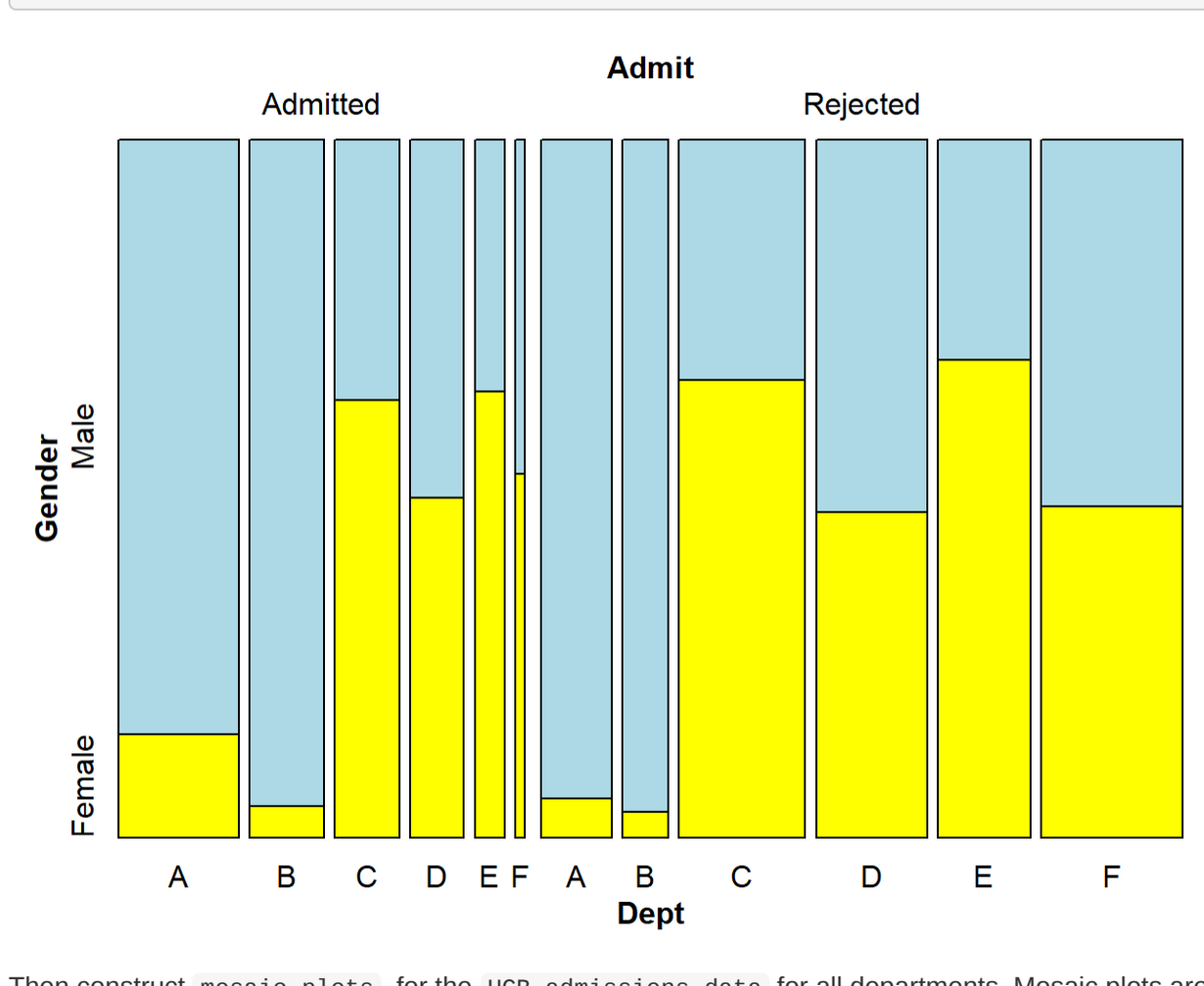
produces the following `mosaic plot` with a variable splitting order:

- first admissions status,
- then gender,
- then department.

The direction specifies how each variable will be split:

- the first variable, Admit, is split vertically;
- the second variable, Gender, is split horizontally;
- and the third variable, Dept, is split vertically.

```
mosaic(~Admit+Gender+Dept, data=UCBAdmissions, highlighting="Gender",
highlighting_fill=c("lightblue", "yellow"), direction=c("v", "h", "v"))
```



Then construct `mosaic plots`, for the `UCB admissions data` for all departments. Mosaic plots are that very rare phenomenon – a clever visualisation method that is actually readable and useful!

1. How many different mosaic plots can be constructed using this data?
2. Which ones enable you to best compare the proportions of men and women admitted to each department?
3. Why do different mosaic plots allow different comparisons?
4. In a mosaic plot, which comparisons can be made accurately, and which can be made approximately? Why? What is the suggested explanation for the apparent bias in UCB admissions?

Titanic passenger survival data

The sinking of the Titanic (aboard were 2,435 passengers and 892 crew members) in 1912 is a famous event, and it is often remembered as a symbol of class inequality: the first class passengers survived, whereas the third class passengers perished.

A 4-way **contingency table** of the survival data for Titanic passengers is in the dataset `Titanic`.

1. How many mosaic plots can you construct for this data?
2. Which mosaic plots allow the best comparison of survival rates for different classes of passenger?
3. Is it true that the first class passengers survived whereas the third class passengers perished?
4. Are there any more striking findings in this data? Do you think that the same pattern of survival would happen today?

A little more information about the Titanic data is in: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/Titanic.html

Exercise: Make up a dataset that shows Simpson's paradox

Simpson's paradox is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

A real life example of this phenomenon is when the University of California, Berkeley was sued for bias against women who had applied for admission to graduate schools in 1973. Admission figures showed that men applying were more likely than women to be admitted, and the difference was so substantial that one would conclude that discrimination existed. However, when examining individual academic departments, it appeared that no department was significantly biased against women.

Suppose that a drug company wants to market an expensive new drug ("Redpills", perhaps).

They perform clinical trials on Redpills, testing them on men and women. There is a 3-way contingency table:

1. Redpills / no Redpills;
2. Man / Woman;
3. Felt Better / Felt Worse.

More women than men signed up to participate in the clinical trials.

They want to market Redpills for both football injuries and knitting injuries, and they perform clinical trials on both men and women. You may assume that some women play football, and some men knit.

Unfortunately, when properly analysed, the results of the trials show that Redpills have little effect. However, an executive suggests a summary of the data that appears to show that Redpills make most people feel better. How might this have been done? Would it always be possible?

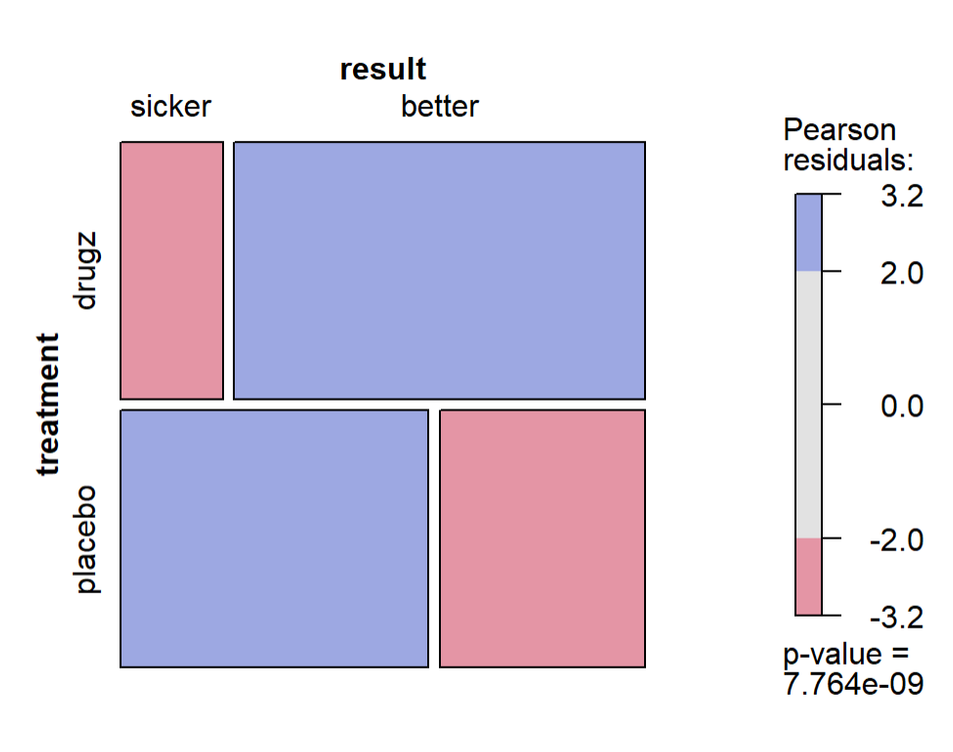
```
drugz <- data.frame(expand.grid(treatment=c("drugz","placebo"),
                                result=c("sicker","better")),count=c(20,60,80,40))
drugz
```

```
##   treatment result count
## 1     drugz sicker    20
## 2   placebo sicker    60
## 3     drugz better    80
## 4   placebo better    40
```

```
drugztable <- xtabs(formula = count ~ treatment + result, data = drugz)
drugztable
```

```
##          result
## treatment sicker better
##   drugz       20     80
##   placebo     60     40
```

```
mosaic(~treatment+result, drugztable,shade=TRUE,split_vertical=c(FALSE,TRUE))
```




```
GSS <- data.frame(expand.grid(sex=c("female", "male"),party=c("dem", "indep", "rep")),
                 count=c(279,165,73,47,225,191))
GSS
```

```
##      sex party count
## 1 female   dem   279
## 2   male   dem   165
## 3 female indep    73
## 4   male indep    47
## 5 female   rep   225
## 6   male   rep   191
```

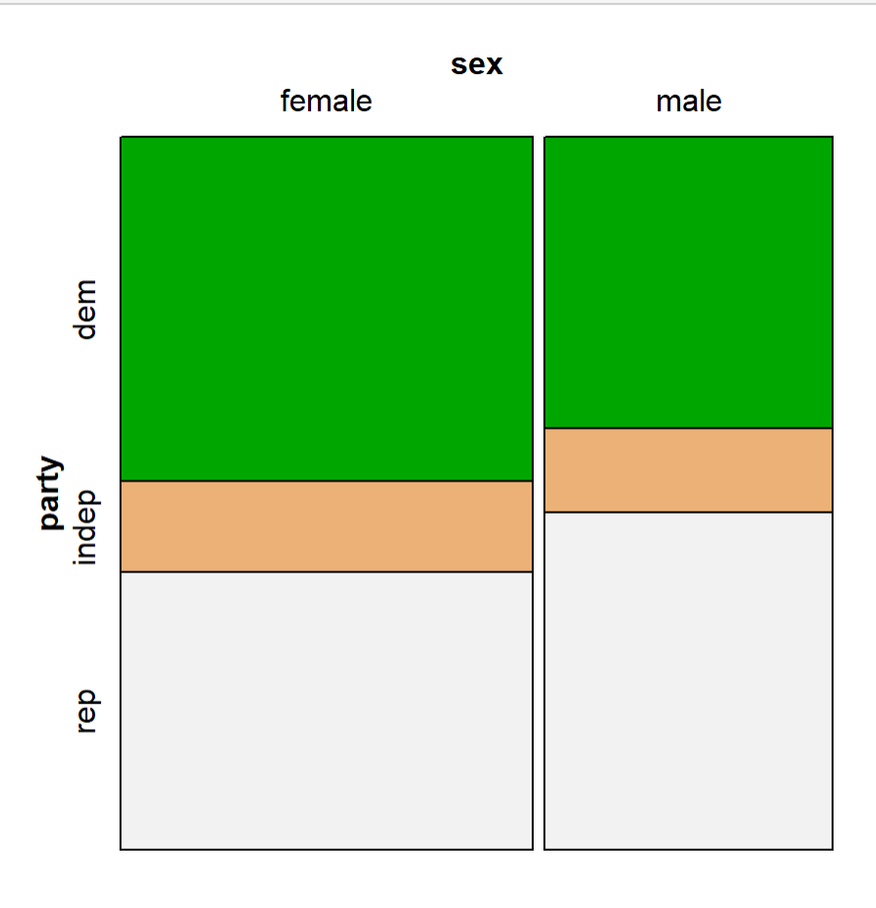
```
gsstable <- xtabs( formula = count ~ sex + party, data = GSS)
dimnames(gsstable)
```

```
## $sex
## [1] "female" "male"
##
## $party
## [1] "dem"   "indep" "rep"
```

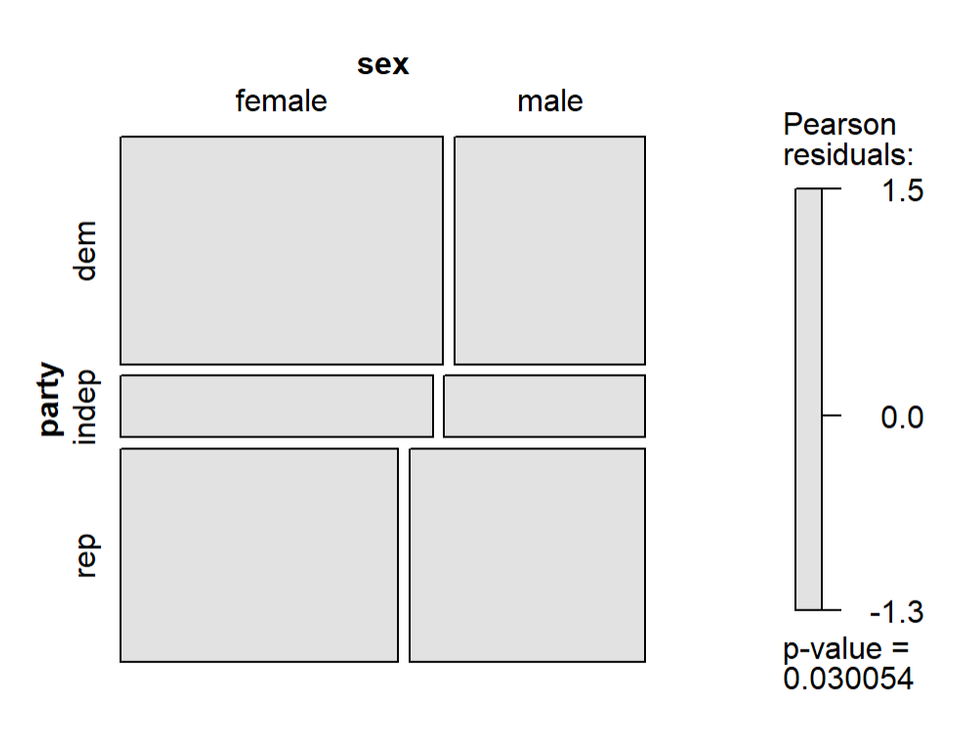
```
gsstable
```

```
##         party
## sex      dem indep rep
##   female 279    73 225
##   male   165    47 191
```

```
mosaic( ~party + sex, data=gsstable, highlighting=TRUE,highlighting_fill=terrain.colors(3))
```

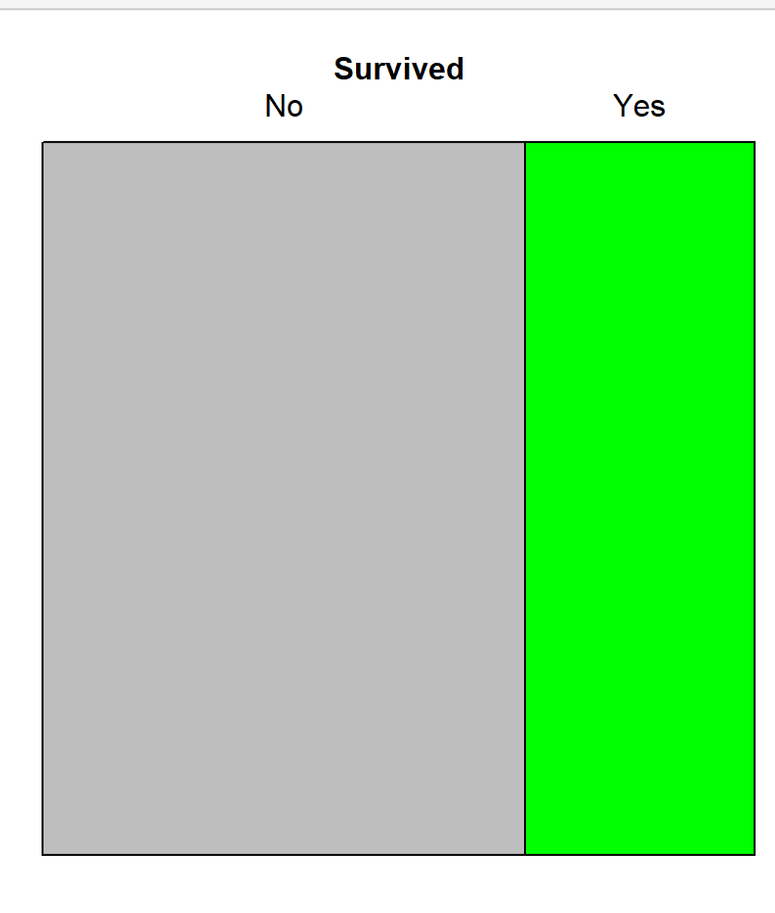



```
mosaic( ~party + sex, data=gsstable, shade=TRUE)
```

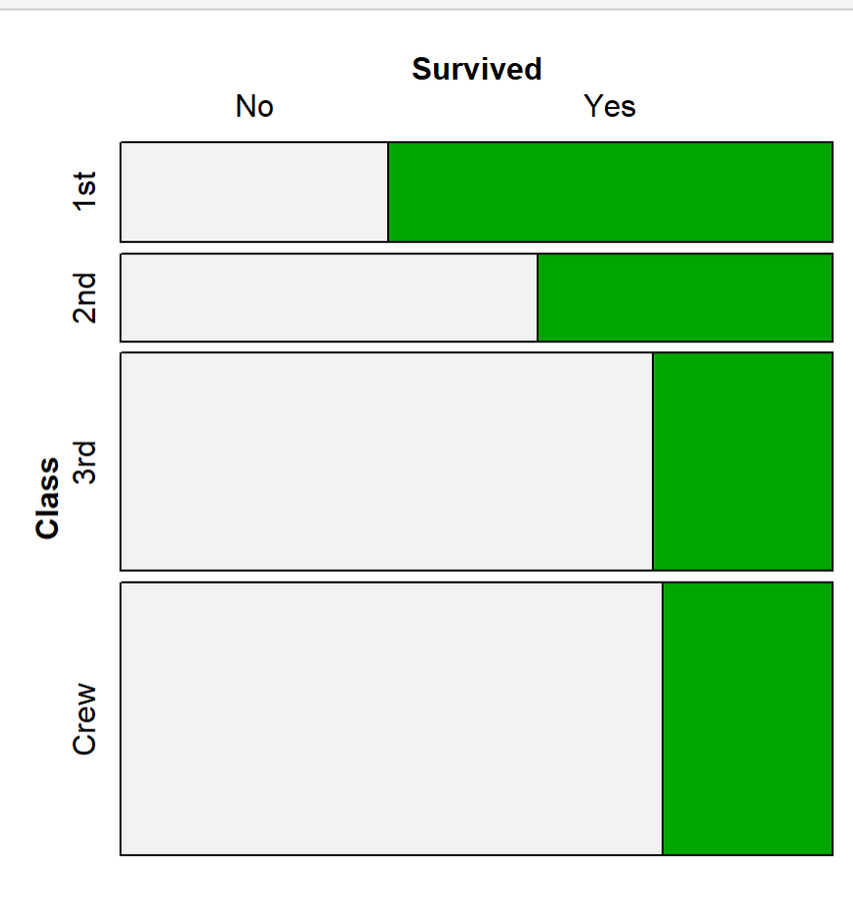


Titanic

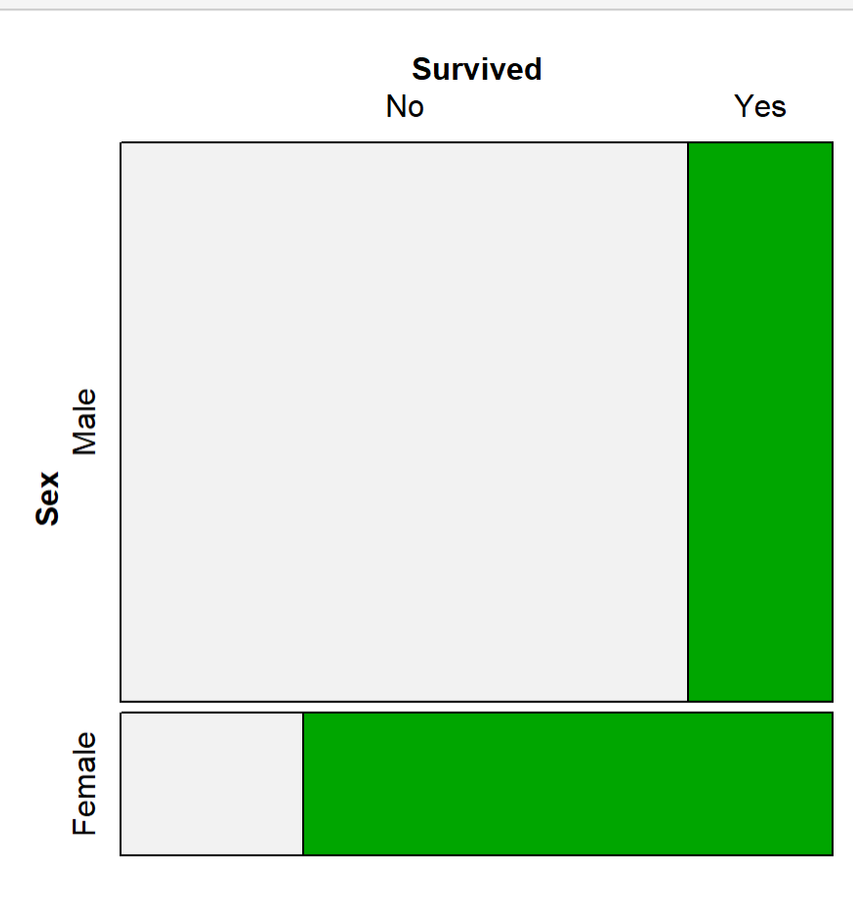
```
mosaic(~Survived, data=Titanic, split_vertical=c(TRUE),
      highlighting_fill=c("grey","green"), highlighting=TRUE)
```



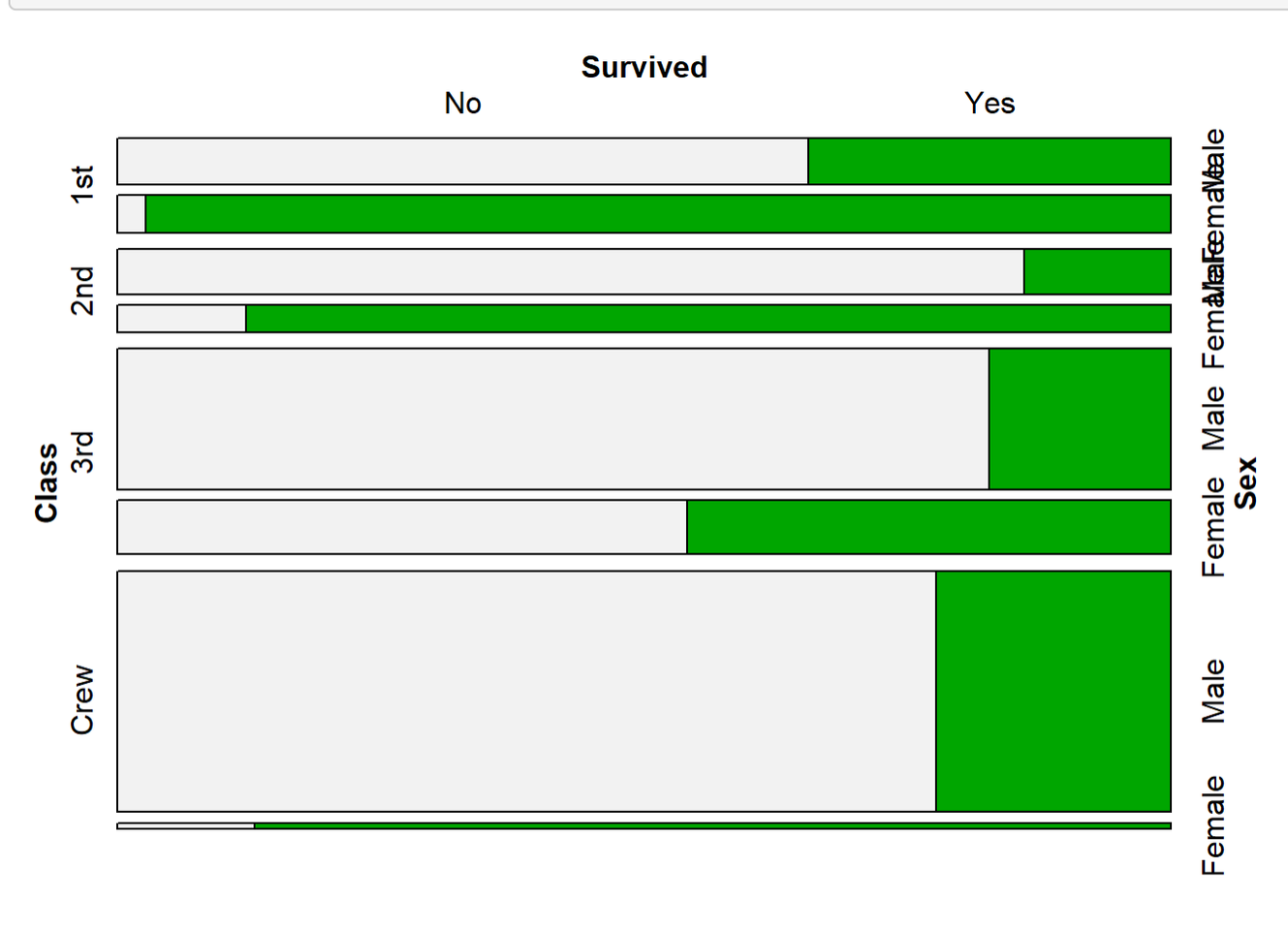

```
mosaic(~Survived + Class, data=Titanic, split_vertical=c(TRUE,FALSE,FALSE),
      highlighting_fill=rev(terrain.colors(2,alpha=1)),highlighting=TRUE)
```




```
mosaic(~Survived + Sex, data=Titanic[,,"Adult",], split_vertical=c(TRUE,FALSE),
      highlighting_fill=rev(terrain.colors(2)), highlighting=TRUE)
```




```
mosaic(~Survived + Class + Sex, data=Titanic, split_vertical=c(TRUE,FALSE,FALSE),
      highlighting_fill=rev(terrain.colors(2,alpha=1)),highlighting=TRUE)
```



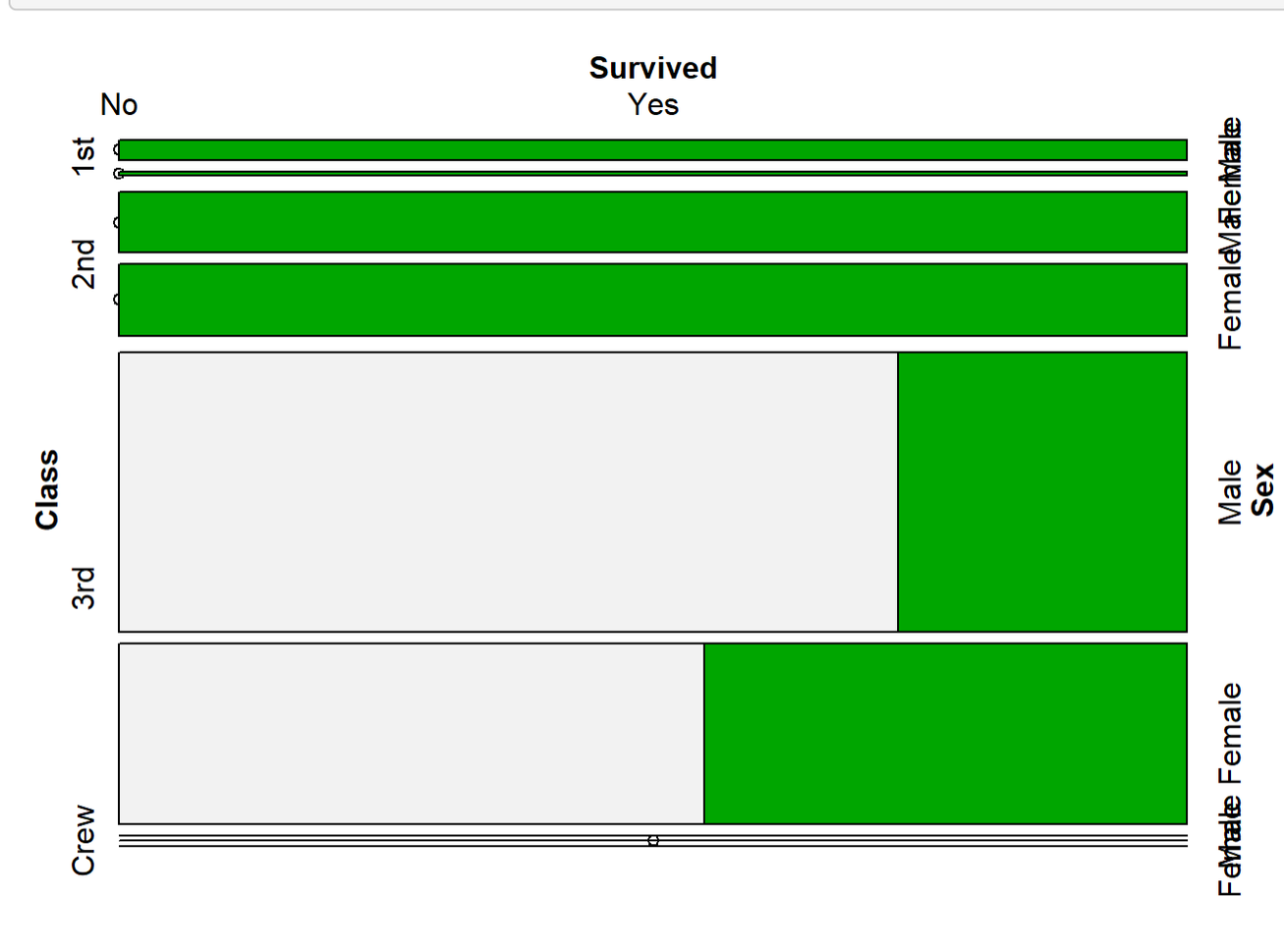

```
dimnames(Titanic)
```

```
## $Class
## [1] "1st"  "2nd"  "3rd"  "Crew"
##
## $Sex
## [1] "Male"   "Female"
##
## $Age
## [1] "Child" "Adult"
##
## $Survived
## [1] "No"  "Yes"
```

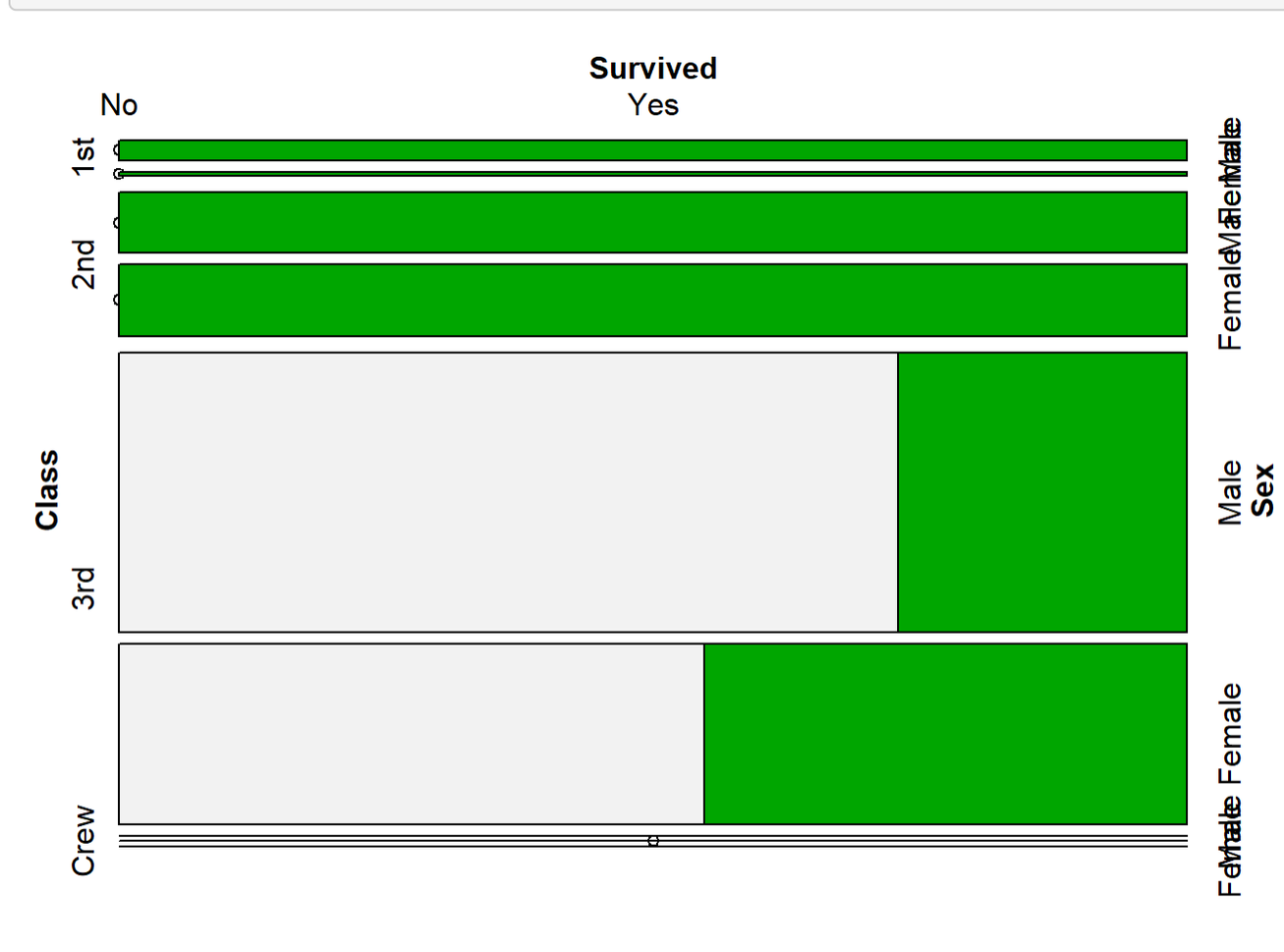
```
Titanic[ , , "Child", ]
```

```
## , , Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
```

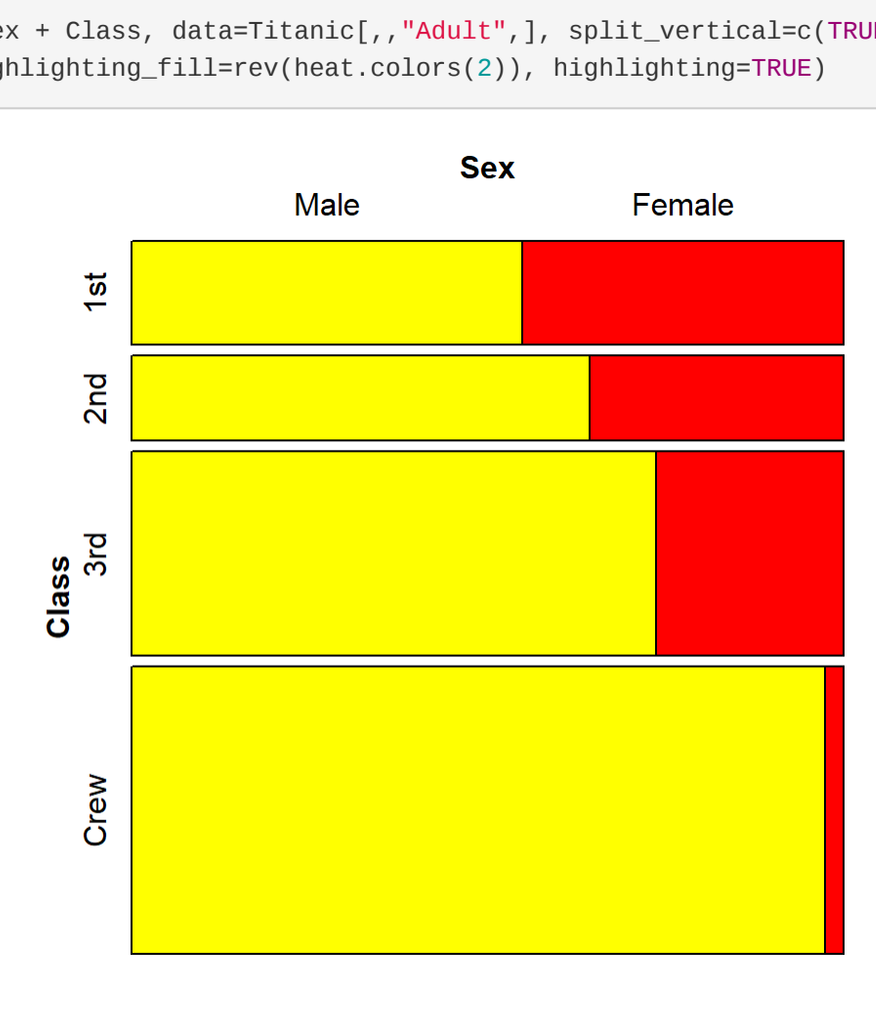
```
mosaic(~Survived + Class + Sex, data=Titanic[,,"Child",],
      split_vertical=c(TRUE,FALSE,FALSE), highlighting_fill=rev(terrain.colors(2)),
      highlighting=TRUE)
```



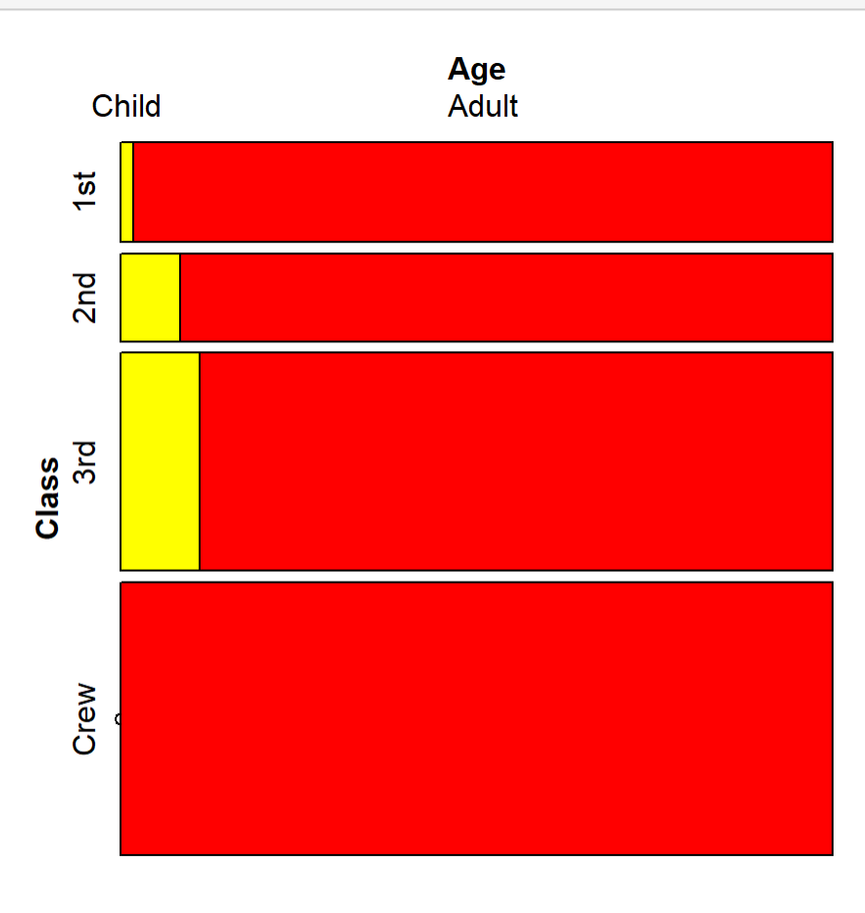

```
mosaic(~Survived + Class + Sex, data=Titanic[,,"Child",],
      split_vertical=c(TRUE,FALSE,FALSE, FALSE), highlighting_fill=rev(terrain.colors(2)),
      highlighting=TRUE)
```




```
mosaic(~Sex + Class, data=Titanic[,,"Adult",], split_vertical=c(TRUE,FALSE),
      highlighting_fill=rev(heat.colors(2)), highlighting=TRUE)
```




```
mosaic(~Age + Class, data=Titanic, split_vertical=c(TRUE,FALSE),
      highlighting_fill=rev(heat.colors(2)), highlighting=TRUE)
```



UCBAdmissions

```
dimnames(UCBAdmissions)
```

```
## $Admit
## [1] "Admitted" "Rejected"
##
## $Gender
## [1] "Male"   "Female"
##
## $Dept
## [1] "A" "B" "C" "D" "E" "F"
```

```
mosaic( ~Admit + Gender + Dept, data=UCBAdmissions, highlighting=TRUE,
       highlighting_fill=terrain.colors(2))
```


```
mosaic( ~Admit + Dept + Gender, data=UCBAdmissions, highlighting=TRUE,
       highlighting_fill=terrain.colors(2),split_vertical=c(TRUE,FALSE,FALSE))
```