Titanic

coop711

2015년 4월 13일

Titanic Data Analysis

Main Question : Are women and children saved first?

Data

`datasets` 패키지에 들어있으므로 불러들이기만 하면 됨. 자료의 구조 파악.

```
data(Titanic)
str(Titanic)
```

```
##  table [1:4, 1:2, 1:2, 1:2] 0 0 35 0 0 0 17 0 118 154 ...
##  - attr(*, "dimnames")=List of 4
##   ..$ Class   : chr [1:4] "1st" "2nd" "3rd" "Crew"
##   ..$ Sex     : chr [1:2] "Male" "Female"
##   ..$ Age     : chr [1:2] "Child" "Adult"
##   ..$ Survived: chr [1:2] "No" "Yes"
```

```
Titanic
```

```
## , , Age = Child, Survived = No
##
##       Sex
## Class  Male Female
##   1st     0      0
##   2nd     0      0
##   3rd    35     17
##   Crew    0      0
##
## , , Age = Adult, Survived = No
##
##       Sex
## Class  Male Female
##   1st   118      4
##   2nd   154     13
##   3rd   387     89
##   Crew  670      3
##
## , , Age = Child, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st     5      1
##   2nd    11     13
##   3rd    13     14
##   Crew    0      0
##
## , , Age = Adult, Survived = Yes
##
##       Sex
## Class  Male Female
##   1st    57    140
##   2nd    14     80
##   3rd    75     76
##   Crew  192     20
```

4-dimensional array table이므로 보기 쉽게 `ftable` (flat table) 적용.

```
ftable(Titanic)
```

```
##                     Survived  No Yes
## Class Sex    Age
## 1st   Male   Child            0   5
##              Adult          118  57
##       Female Child            0   1
##              Adult            4 140
## 2nd   Male   Child            0  11
##              Adult          154  14
##       Female Child            0  13
##              Adult           13  80
## 3rd   Male   Child           35  13
##              Adult          387  75
##       Female Child           17  14
##              Adult           89  76
## Crew  Male   Child            0   0
##              Adult          670 192
##       Female Child            0   0
##              Adult            3  20
```

4-dimensional array 인 점을 감안하여 각 변수의 주변합을 구해보면

```
apply(Titanic, 1, sum)
```

```
##  1st  2nd  3rd Crew
##  325  285  706  885
```

```
apply(Titanic, 2, sum)
```

```
##   Male Female
##   1731    470
```

```
apply(Titanic, 3, sum)
```

```
## Child Adult
##   109  2092
```

```
apply(Titanic, 4, sum)
```

```
##   No  Yes
## 1490  711
```

Crosstable 을 구하되 상황 파악이 편하게 열과 행을 조정.

```
apply(Titanic, 1:2, sum)
```

```
##       Sex
## Class  Male Female
##   1st   180    145
##   2nd   179    106
##   3rd   510    196
##   Crew  862     23
```

```
apply(Titanic, 2:1, sum)
```

```
##         Class
## Sex      1st 2nd 3rd Crew
##   Male   180 179 510  862
##   Female 145 106 196   23
```

```
apply(Titanic, c(3,1), sum)
```

```
##        Class
## Age     1st 2nd 3rd Crew
##   Child   6  24  79    0
##   Adult 319 261 627  885
```

```
apply(Titanic, c(4,1), sum)
```

```
##         Class
## Survived 1st 2nd 3rd Crew
##      No  122 167 528  673
##      Yes 203 118 178  212
```

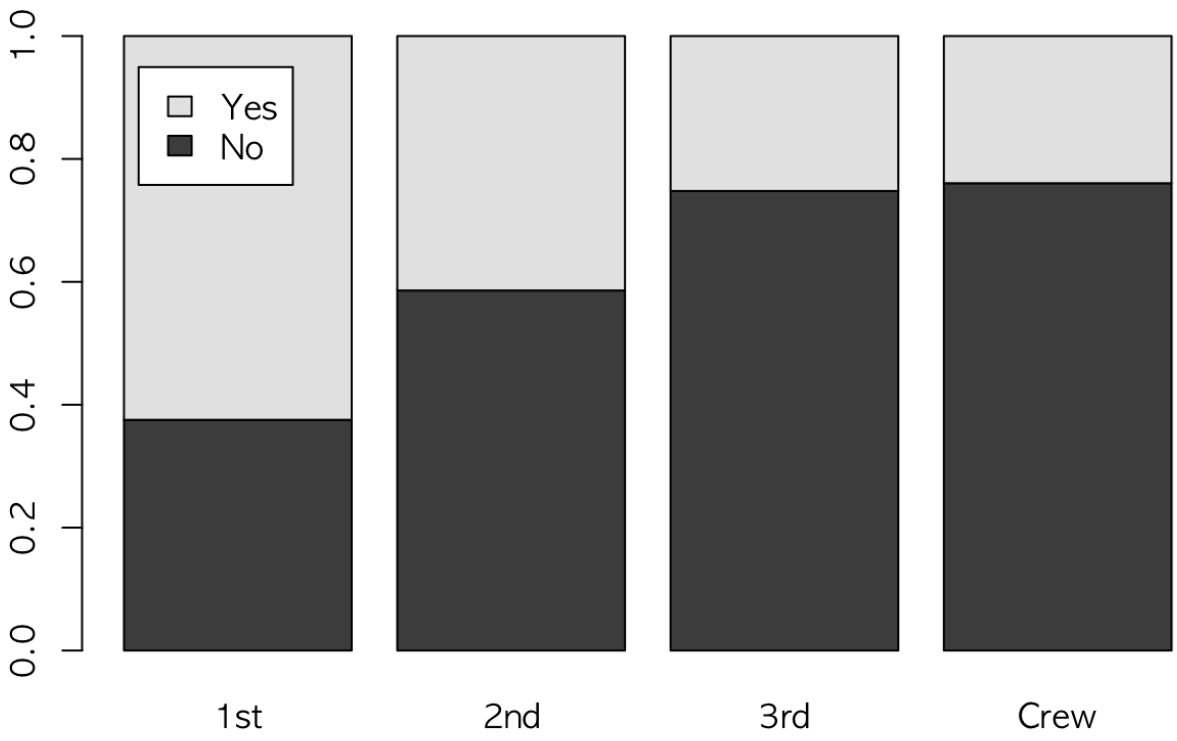
등급별 생존률을 비교하려면. (우선, 자릿수를 정해 놓고)

```
options(digits=2)
prop.table(apply(Titanic, c(4,1), sum), margin=2)
```

```
##         Class
## Survived  1st  2nd  3rd Crew
##      No  0.38 0.59 0.75 0.76
##      Yes 0.62 0.41 0.25 0.24
```

이를 barplot으로 나타내는 데 있어서 각 `argument` 가 왜 필요한지 시행착오를 겪어 볼 것.

```
barplot(prop.table(apply(Titanic, c(4,1), sum), margin=2), legend.text=T, arg
s.legend=list(x="topleft", inset=0.05))
```



Cross-table 을 계속 작성해 가자면

```
apply(Titanic, 2:3, sum)
```

```
##         Age
## Sex      Child Adult
##   Male      64  1667
##   Female    45   425
```

```
apply(Titanic, c(2,4), sum)
```

```
##         Survived
## Sex        No Yes
##   Male   1364 367
##   Female  126 344
```

```
apply(Titanic, c(4,2), sum)
```

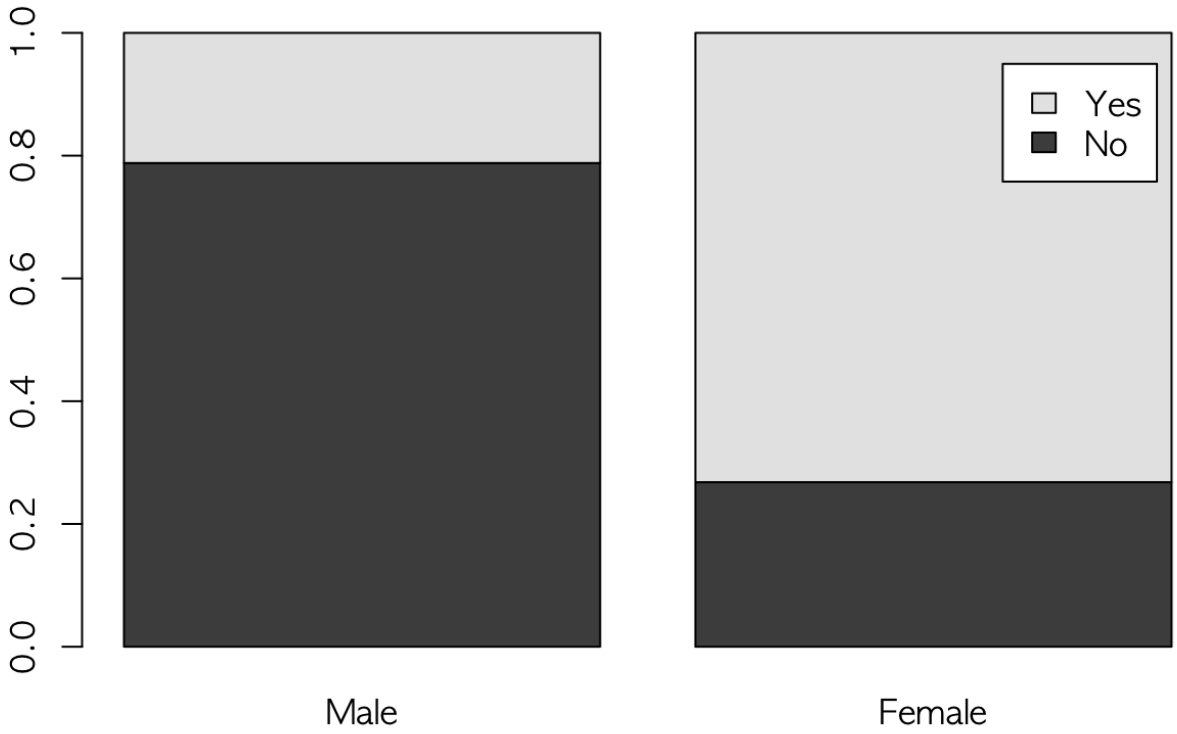
```
##         Sex
## Survived Male Female
##      No  1364    126
##      Yes  367    344
```

남녀 생존률을 비교하려면,

```
prop.table(apply(Titanic, c(4,2), sum), margin=2)
```

```
##          Sex
## Survived Male Female
##      No  0.79   0.27
##      Yes 0.21   0.73
```

```
barplot(prop.table(apply(Titanic, c(4,2), sum), margin=2), legend.text=T, arg
s.legend=list(x="topright", inset=0.05))
```



남은 cross-table 은

```
apply(Titanic, 4:3, sum)
```

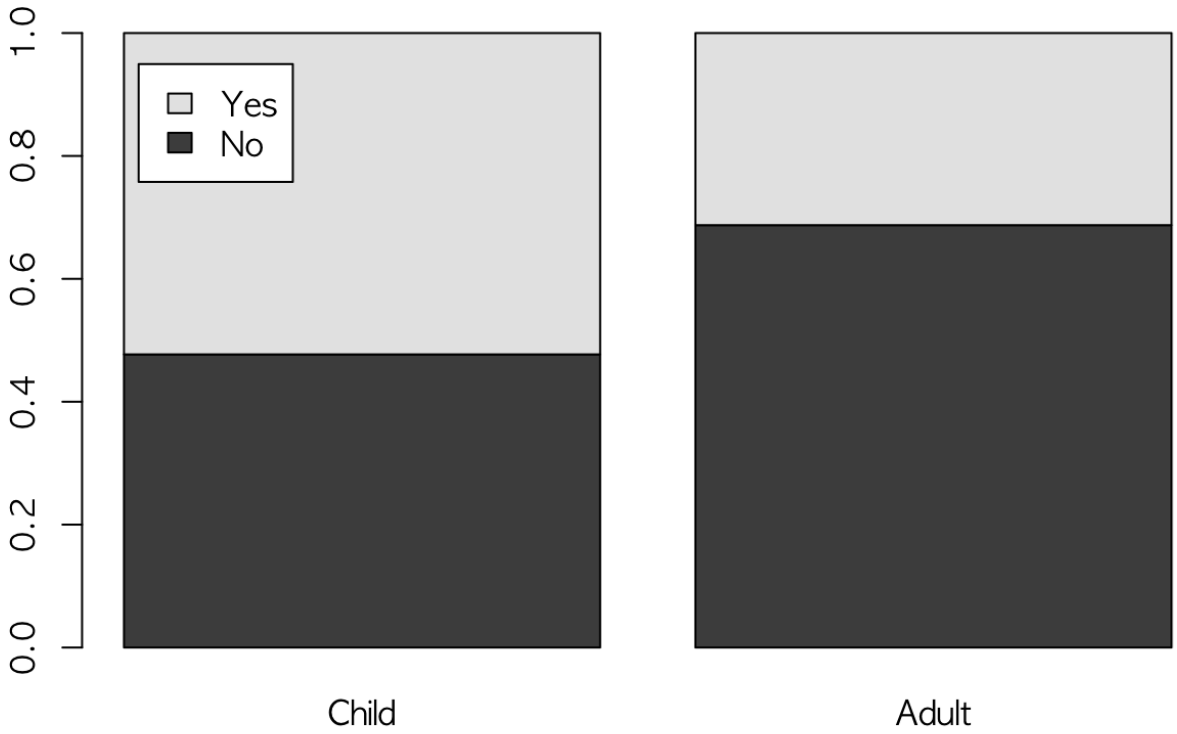
```
##          Age
## Survived Child Adult
##      No     52  1438
##      Yes    57   654
```

성인과 어린이의 생존률을 비교하려면

```
prop.table(apply(Titanic, 4:3, sum), margin=2)
```

```
##          Age
## Survived Child Adult
##       No   0.48  0.69
##       Yes  0.52  0.31
```

```
barplot(prop.table(apply(Titanic, c(4,3), sum), margin=2), legend.text=T, arg
s.legend=list(x="topleft", inset=0.05))
```



객실 등급별로 어린이들과 어른들의 생존률을 비교하려면

```
apply(Titanic, c(3, 4, 1), sum)
```

```
## , , Class = 1st
##
##        Survived
## Age      No Yes
##   Child   0   6
##   Adult 122 197
##
## , , Class = 2nd
##
##        Survived
## Age      No Yes
##   Child   0  24
##   Adult 167  94
##
## , , Class = 3rd
##
##        Survived
## Age      No Yes
##   Child  52  27
##   Adult 476 151
##
## , , Class = Crew
##
##        Survived
## Age      No Yes
##   Child   0   0
##   Adult 673 212
```

```
ftable(apply(Titanic, c(3, 4, 1), sum))
```

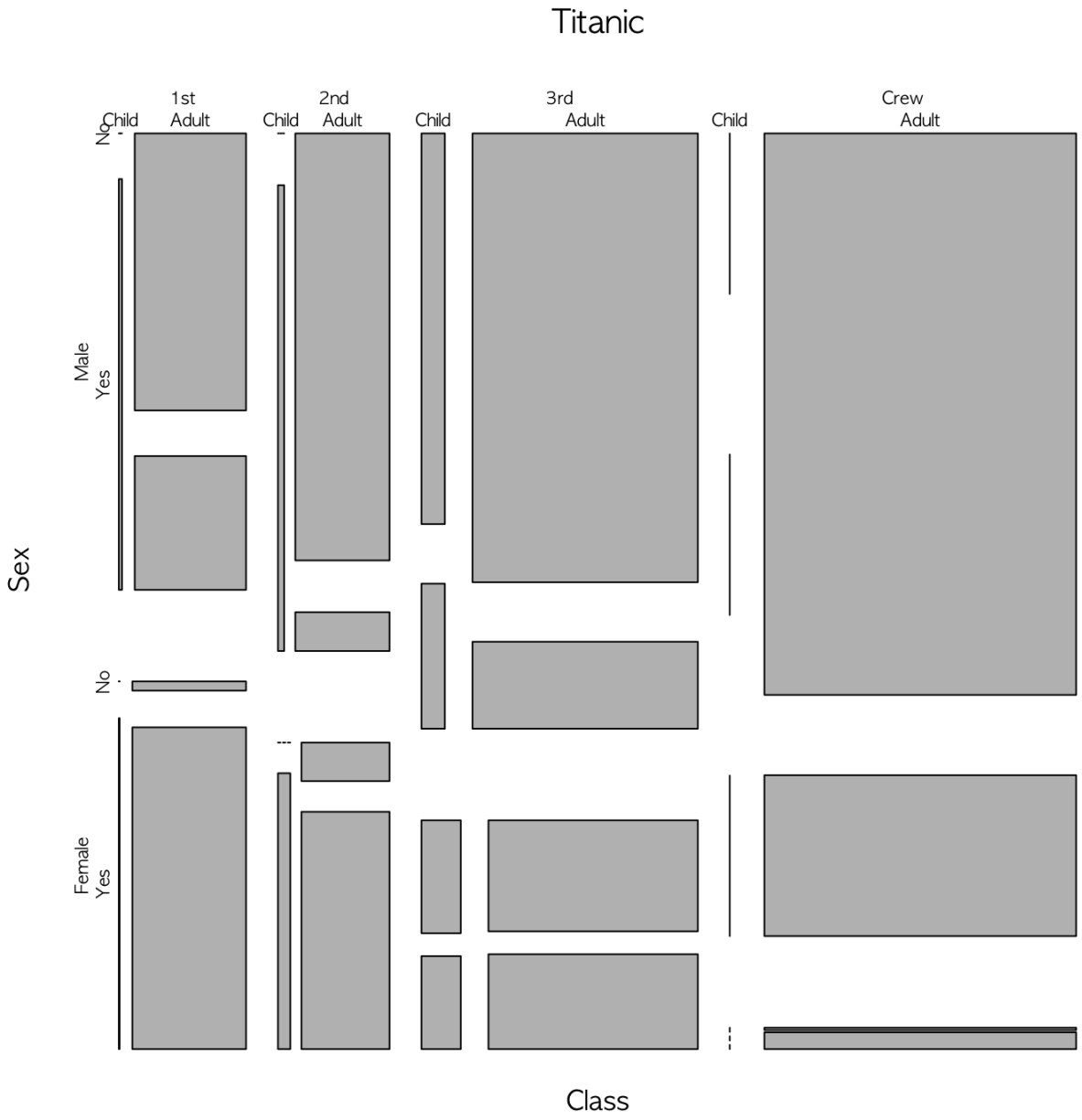
```
##                Class 1st 2nd 3rd Crew
## Age   Survived
## Child No               0   0  52    0
##       Yes              6  24  27    0
## Adult No             122 167 476  673
##       Yes            197  94 151  212
```

```
child.by.class<-prop.table(ftable(apply(Titanic, c(3, 4, 1), sum))[1:2,], margi
n=2)
adult.by.class<-prop.table(ftable(apply(Titanic, c(3, 4, 1), sum))[3:4,], margi
n=2)
child.adult.by.class<-rbind(child.by.class, adult.by.class)
dimnames(child.adult.by.class)<-list(c("child.no", "child.yes", "adult.no", "ad
ult.yes"), dimnames(Titanic)[[1]])
child.adult.by.class
```

```
##            1st  2nd  3rd Crew
## child.no  0.00 0.00 0.66  NaN
## child.yes 1.00 1.00 0.34  NaN
## adult.no  0.38 0.64 0.76 0.76
## adult.yes 0.62 0.36 0.24 0.24
```

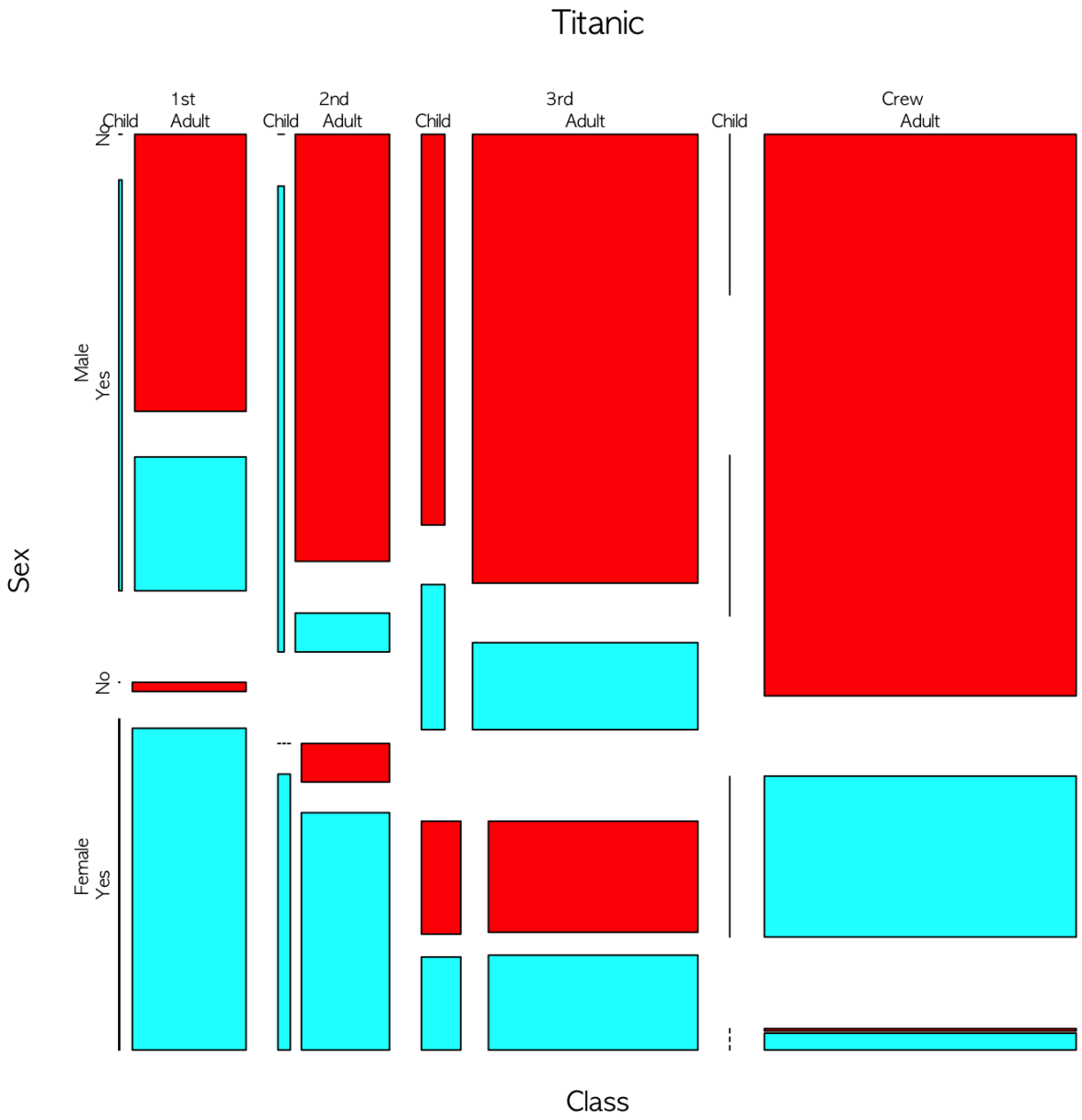
한꺼번에 살펴보려면

```
mosaicplot(Titanic)
```



컬러로 구분하려면

```
mosaicplot(Titanic, col=rainbow(2))
```



이 자료를 보다 익숙한 data frame 으로 작업하려면 `as.data.frame()` 을 이용하여

```
Titanic.df<-as.data.frame(Titanic)
str(Titanic.df)
```

```
## 'data.frame':    32 obs. of  5 variables:
##  $ Class   : Factor w/ 4 levels "1st","2nd","3rd",..: 1 2 3 4 1 2 3 4 1 2
...
##  $ Sex     : Factor w/ 2 levels "Male","Female": 1 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : Factor w/ 2 levels "Child","Adult": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Survived: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Freq    : num  0 0 35 0 0 0 17 0 118 154 ...
```

Survived factor의 `"Yes"`, `"No"` level을 바꾸려면

```
Titanic.df$Survived<-factor(Titanic.df$Survived, levels=c("Yes", "No"))
str(Titanic.df)
```

```
## 'data.frame':    32 obs. of  5 variables:
##  $ Class   : Factor w/ 4 levels "1st","2nd","3rd",..: 1 2 3 4 1 2 3 4 1 2
...
##  $ Sex     : Factor w/ 2 levels "Male","Female": 1 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : Factor w/ 2 levels "Child","Adult": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Survived: Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Freq    : num  0 0 35 0 0 0 17 0 118 154 ...
```

이제는 `xtabs()` 등의 익숙한 함수를 이용할 수 있음.

```
xtabs(Freq~Survived, data=Titanic.df)
```

```
## Survived
##  Yes   No
##  711 1490
```

```
addmargins(xtabs(Freq~Survived, data=Titanic.df))
```

```
## Survived
##  Yes   No  Sum
##  711 1490 2201
```

```
xtabs(Freq~Survived+Class, data=Titanic.df)
```

```
##          Class
## Survived 1st 2nd 3rd Crew
##      Yes 203 118 178  212
##      No  122 167 528  673
```

```
addmargins(xtabs(Freq~Survived+Class, data=Titanic.df))
```

```
##          Class
## Survived  1st  2nd  3rd Crew  Sum
##      Yes  203  118  178  212  711
##      No   122  167  528  673 1490
##      Sum  325  285  706  885 2201
```

```
xtabs(Freq~Survived+Sex, data=Titanic.df)
```

```
##         Sex
## Survived Male Female
##      Yes  367    344
##      No  1364    126
```

```
addmargins(xtabs(Freq~Survived+Sex, data=Titanic.df))
```

```
##         Sex
## Survived Male Female  Sum
##      Yes  367    344  711
##      No  1364    126 1490
##      Sum 1731    470 2201
```

```
xtabs(Freq~Survived+Age, data=Titanic.df)
```

```
##         Age
## Survived Child Adult
##      Yes    57   654
##      No     52  1438
```

```
addmargins(xtabs(Freq~Survived+Age, data=Titanic.df))
```

```
##         Age
## Survived Child Adult  Sum
##      Yes    57   654  711
##      No     52  1438 1490
##      Sum   109  2092 2201
```

```
ftable(xtabs(Freq~Age+Survived+Class, data=Titanic.df))
```

```
##                Class 1st 2nd 3rd Crew
## Age   Survived
## Child Yes              6  24  27    0
##       No               0   0  52    0
## Adult Yes            197  94 151  212
##       No             122 167 476  673
```