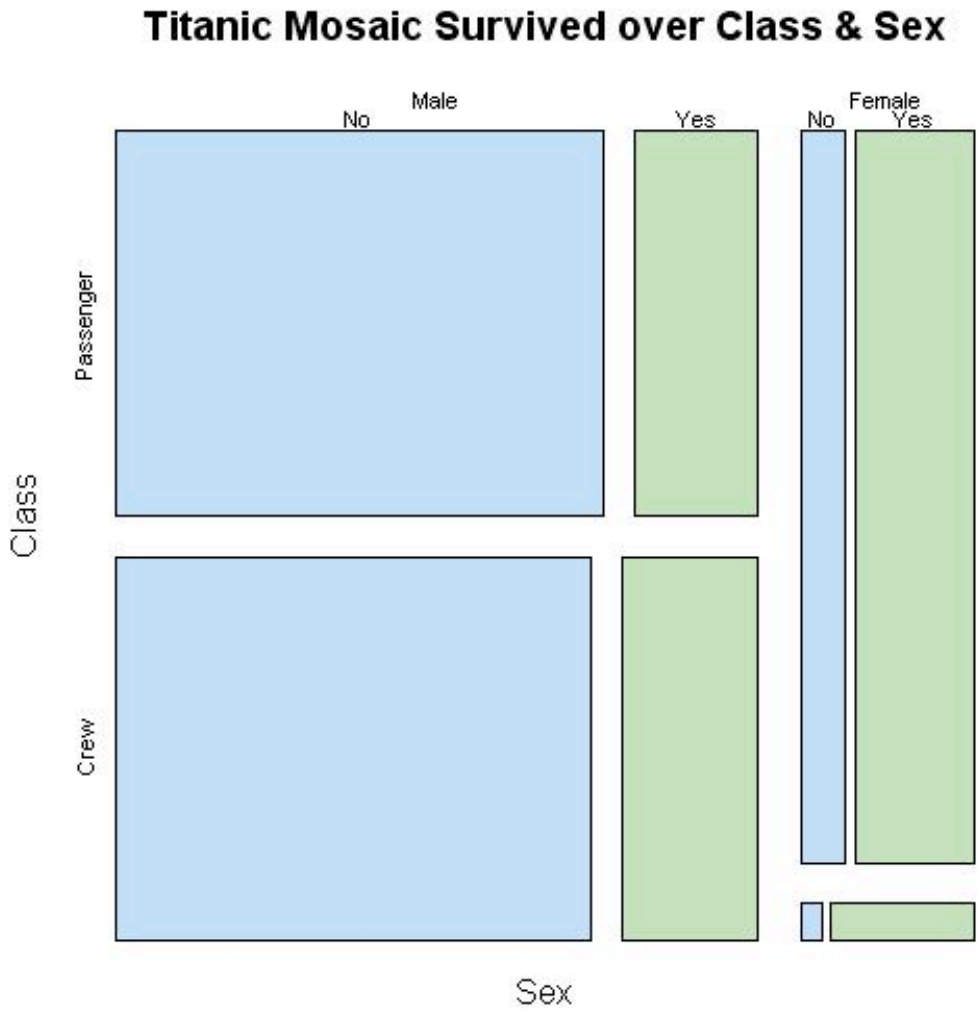
# Mosaic Plot

A graphical display of cross-classified data

유 충현
블로그 모음 1탄(http://blog.naver.com/bdboys) • (주)오픈베이스 • 2012년 10월 3일



## Mosaic Plot

단변량의 양적 데이터의 분포를 설명하는 도구에 Histogram이라는 훌륭한 Chart가 있다. 단, 계급의 수를 적절하게 산정해야만 데이터의 분포를 이해할 수 있는 Chart가 된다. 물론 통계분석 프로그램이나 Chart 생성 프로그램에서는 적정 수의 계급을 산정해서 Chart를 그려주므로 크게 신경쓸 일은 아니다.

단변량의 범주형 데이터의 분포를 설명하는 도구에는 Pie Chart라는 훌륭한 Chart가 있다. 전체 도수 대비 해당 도수의 비율 만큼 파이의 크기(각도)가 달라지는 Chart로 Class의 수가 많은면 가독성이 떨어지는 단점이 있기는 하다.

위 두가지 Chart는 일상적으로 많이 접하는 Chart들이다. Bar Chart, 산포도 등등 대중적인 Chart들은 교육 과정이나 인문, 사회과학이나 경제분야에서 독자들에게 쉽게 지표를 설명해주는 Information Visualisation Tool이라 할 수 있다.

그러면 본격적으로 Mosaic Plot에 대해서 이야기를 해 보자.

초등학교 시절에 미술시간에 색종이와 신문지 등을 오려서 모자이크 그림을 그렸던 추억이 있다. 아니, 그리기 보다는 만들었다는 표현이 더 가까울 수도 있겠다. 가위로 오려서 도화지에 풀로 붙혀서 그림을 그리고, 오리는 작업이 번거로워 밑그림이 그려진 도화지에 풀을 발라 놓고, 색종이를 그 위에 올려 연필로 눌러서 찢어 붙힌 기억도 있다.

그러면 생소한 **Mosaic Plot**은 어떤 것이고, 어떨 때 사용하는 Chart인가?

Mosaic Plot은 1981년 Hartigan와 Kleiner가 분할표의 수를 표현하기 위해서 처음으로 시도한 Chart로 Spine Plot에서 기본 아이디어를 얻었다고 한다. 그리고 1984년 이들을 다음과 같이 Mosaic Plot을 정의하였다.

"a graphical display of **cross-classified data** in which each count is represented by a **rectangle** of area **proportional to the count**"

Mosaic Plot은 다변량의 범주형 데이터의 분포를 설명하는 도구이다. 단변량의 범주형 데이터의 분포에도 사용할 수는 있지만 여타 Plot에 비해 실효성을 떨어지기 때문에 사용하지 않는 게 일반적이다. 그러나 다변량 분석에 앞서 자료의 분포를 조망하는 EDA 단계에서 유요할 도구로 사용될 수 있다. 위력을 발휘하는 분야로는 2차원 이상의 교차 분류 자료 (cross-classified data)의 표현이다.

앞서 예를 든 Histogram, Pie Chart 등은 유클리드 공간에서의 면적의 크기가 도수의 양과 비례하는 중요한 특성이 있다. 이 특성으로 면적의 대소를 보면서 전체 집합으로부터 해당 Class들의 분포를 이해하는 것이다.

즉, 비율이라는 수치가 2차원 공간에서 면적으로 시각화되어 독자의 이해를 쉽게 도와주는 것이고, 이것이 바로 수치 Table을 Chart로 변환하는 배경이기도 하다.

이 Chart들에는 다음과 같은 중요한 법칙이 있다.

1. 각 조각의 면적 합은 1이 된다. (Total mass 1)

$$\sum X_i = 1$$

2. 각 조각들은 중첩하지 않는다.

$$\sum (X_i \bigcap X_{i+1}) = \phi$$

Mosaic Plot은 해당 변수에서 각각의 Class 도수의 비율 만큼의 직사각형의 면적을 색종이에서 오려서 도화지에 붙혀 놓은 Chart라고 볼 수 있다. 이때 해당 변수내의 각 조각의 색종이의 합은 1이 되게 하고, 중첩하지 않고 구분할 수 있게 도화지에 붙혀넣는게 전부이다.

변수 차원이 1일 경우에는 가로축으로, 변수 차원이 2차원일 경우에는 가로축, 세로축으로 구분해서 붙혀 넣는다. 그리고 3차원, 4차원으로 확장할 수도 있다. 색종이의 색깔은 모든 변수의 Class의 수가 된다. 또한 오린 색종이의 개수는 각각 변수의 Class의 곱이 된다. 색종이의 크기는 모든 Class의 돗수의 합 대비 해당 Class의 돗수의 비율과 비례하게 크기가 정해지며, 차원과 Class의 수가 많을 수록 그 조각들은 작아지게 된다.

예를 들면,

2*2*2 분할표를 표현하기 위한 색종이의 색깔 수 (굳이 색으로 구분하자면)는 2+2+2 = 6이 되고, 색종이의 조각은 2*2*2 = 8이 된다.

다음의 예로 들면서 Mosaic Plot에 대해서 알아 보자.

어떤 부류에서 클래식 음악을 많이 듣는지 가상으로 조사한 자료가 있다. 2300명으로 구성된 표본을 나이와 교육수준, 클래식을 듣는 여부의 세가지 변수로 분류를 한 자료가 다음과 같다.

2*2*2의 분할표이다.

| | 교육수준 | | | |
|---|---|---|---|---|
| 나이 | 고학력 | | 저학력 | |
| | 클래식음악 듣기 | | | |
| | 예 | 아니오 | 예 | 아니오 |
| 고연령 | 210 | 190 | 170 | 730 |
| 저연령 | 194 | 406 | 110 | 290 |

고연령 대 저연령의 나이로 비교해 보면 1300 대 1000으로 고연령의 표본이 더 많다.

| 고연령 | 저연령 |
|---|---|
| 56.5% | 43.5% |

여기에 교육수준을 추가해서 살펴보자. 저연령의 집단이 고연령보다 학력이 높은 것을 알 수 있다.

| 고연령 | | 저연령 | |
|---|---|---|---|
| 고학력 | 저학력 | 고학력 | 저학력 |
| 30.8% | 69.2% | 60.0% | 40.0% |

마지막으로 각각의 그룹에서 클래식을 듣는 사람들의 비율을 비교해 보자. 저연령대에서는 교육수준과 클래식을 듣는 것과는 차이가 없는 반면에 고연령대에서는 교육수준이 높을 수록 클래식을 듣는 비율이 높음을 알수 있다.

| 고연령 | | 저연령 | |
|---|---|---|---|
| 고학력 | 저학력 | 고학력 | 저학력 |
| 52.5% | 18.9% | 32.3% | 27.5% |

정리하면

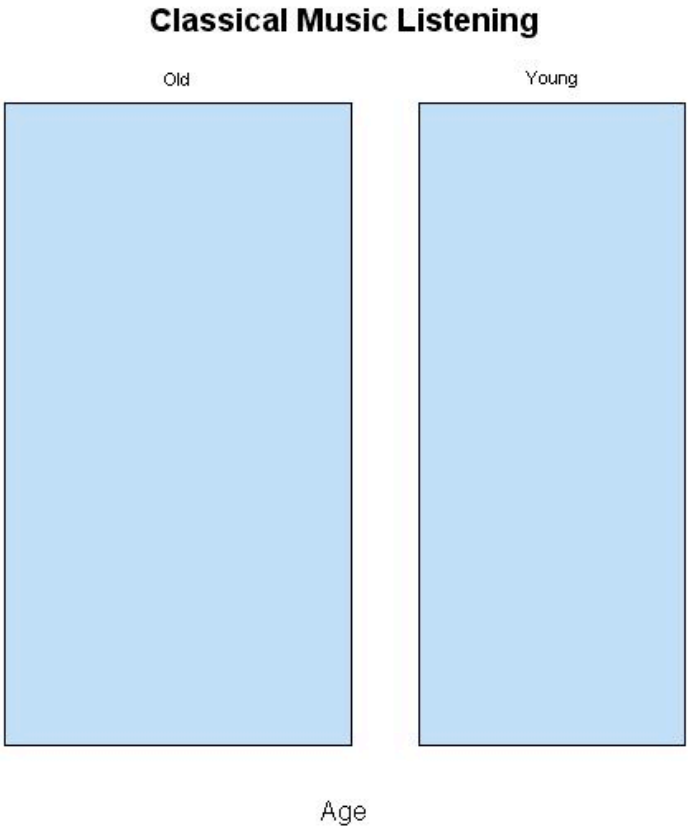
1. 표본은 고연령대보다 저연령대가 더 많다.

2. 저연령대는 고연령대보다 교육수준이 높은 사람이 많다.

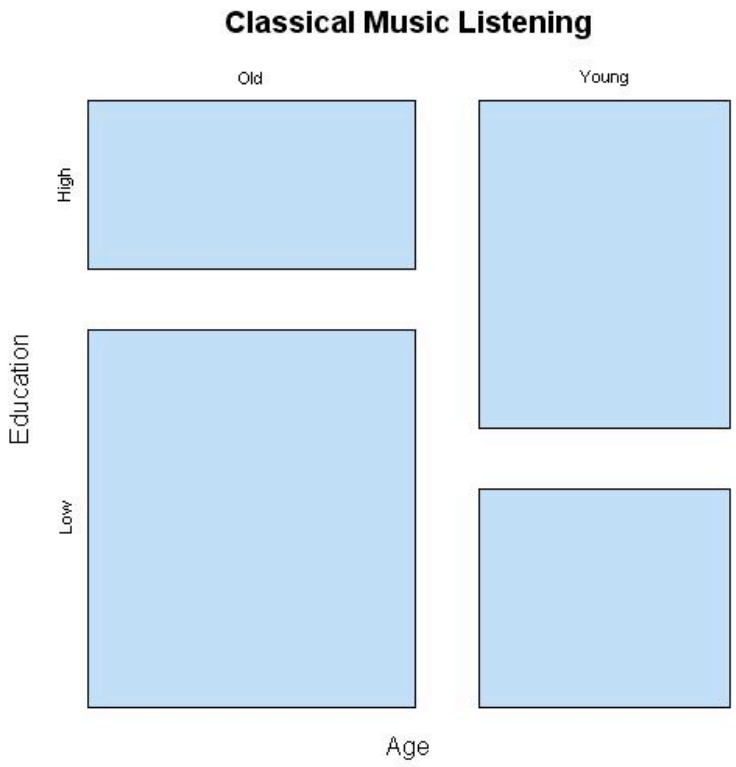
3. 클래식 음악을 즐겨 듣는 것은 나이와 교육수준의 두 level에 종속적이다.

자, 그러면 Mosaic Plot을 만들어 보자.

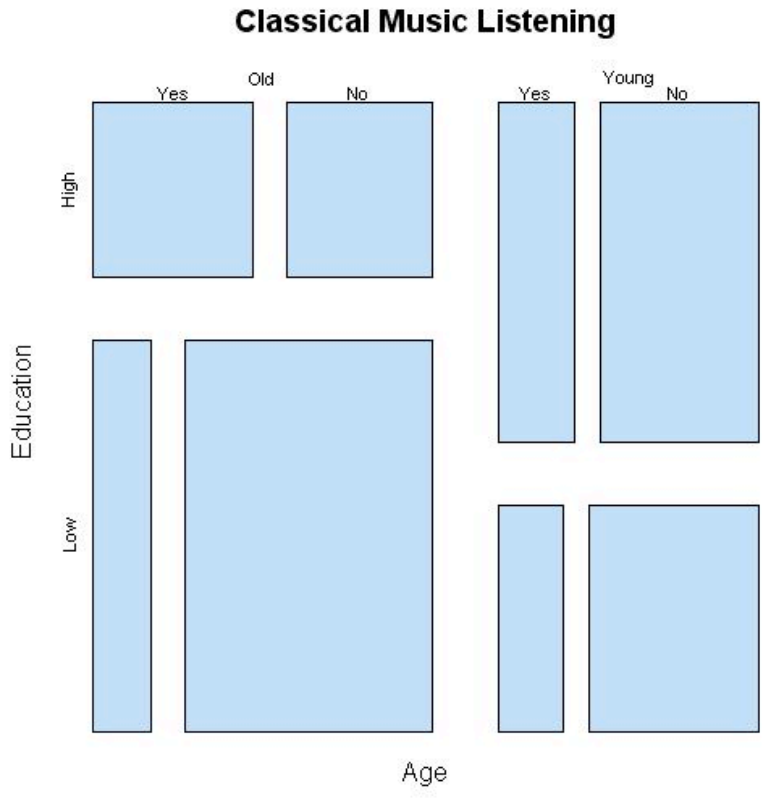
우선 연령 기준의 1차원 Mosaic Plot을 만들어 보자.




다음으로 연령과 학력수준의 2차원 Mosaic Plot을 만들어 보자.




마지막으로 연령과 학력수준과 클래식 선호의 3차원 Mosaic Plot을 만들어 보자.



이 예제의 R Program Script는 다음과 같다.

```
> 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"))

> mosaicplot(apply(music,1,sum), col = hcl(240),
    main = "Classical Music Listening", sub="Age")

> mosaicplot(apply(music,1:2,sum), col = hcl(240),
    main = "Classical Music Listening")

> mosaicplot(music, col = hcl(240), main = "Classical Music Listening")
```

다음으로 타이타닉 승선자들의 생존에 대한 자료를 예로 Mosaic Plot에 대해 알아보자.

타이타닉 승선자 자료는 R의 datasets 패키지에 Titanic이라는 데이터 객체 이름으로 만들어져 있다. 4차원 array이며 2201의  도수를 가지고 있으며 자료의 구조는 다음과 같다.

- 객실 Class : 1등실, 2등실, 3등실, 승무원
- 성별 : 남자, 여자
- 나이 : 아이, 어른
- 생존여부 : 사망, 생존

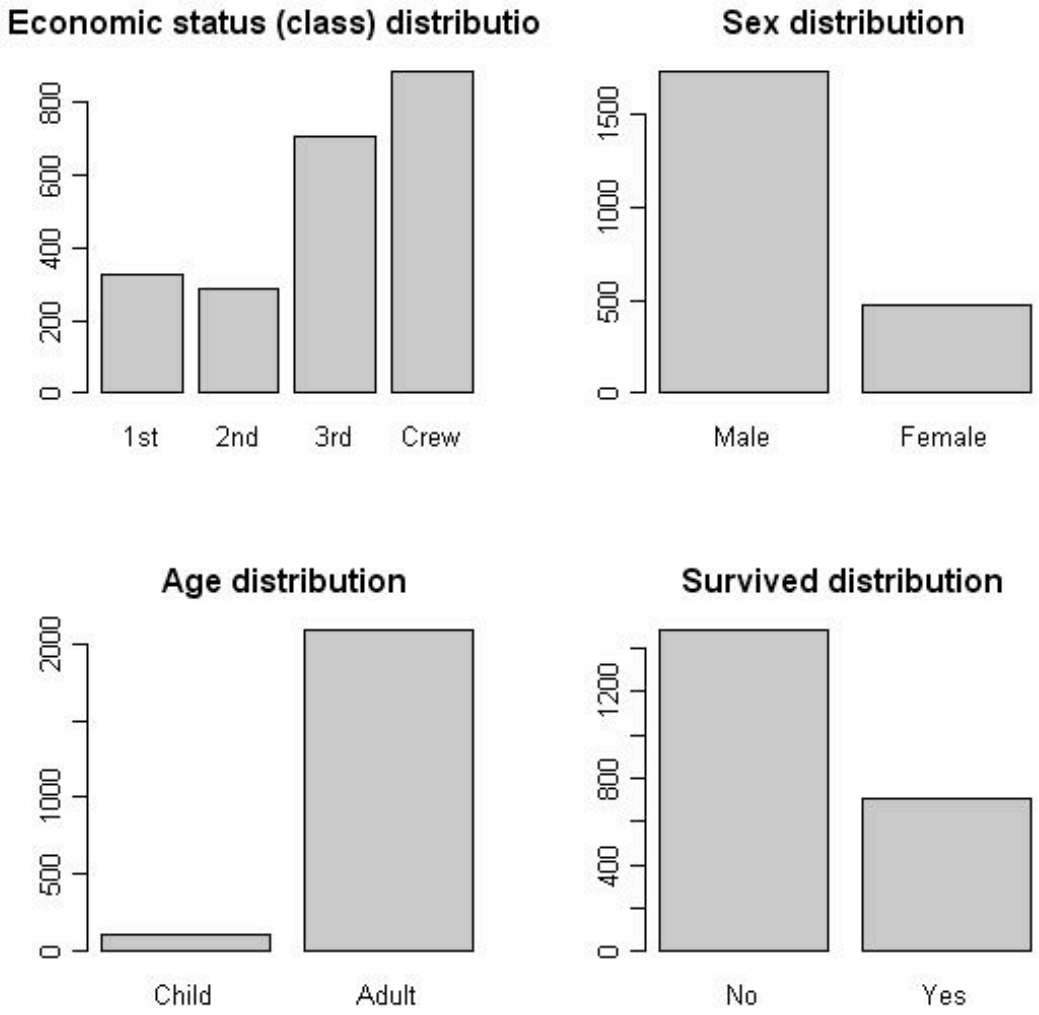
다음에 4차원 array를 4*2*2*2 분할표로 만들었다.

| 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 |

각각의 변수들의 분포를 barplot으로 나타내면 다음과 같다.

이 barplot은 단순히 개별 변수들에서 class의 도수분포 비율을 나타낸 것이다.




이 barplot은 단순히 개별 변수들에서 class의 도수분포 비율을 나타낸 것이다.

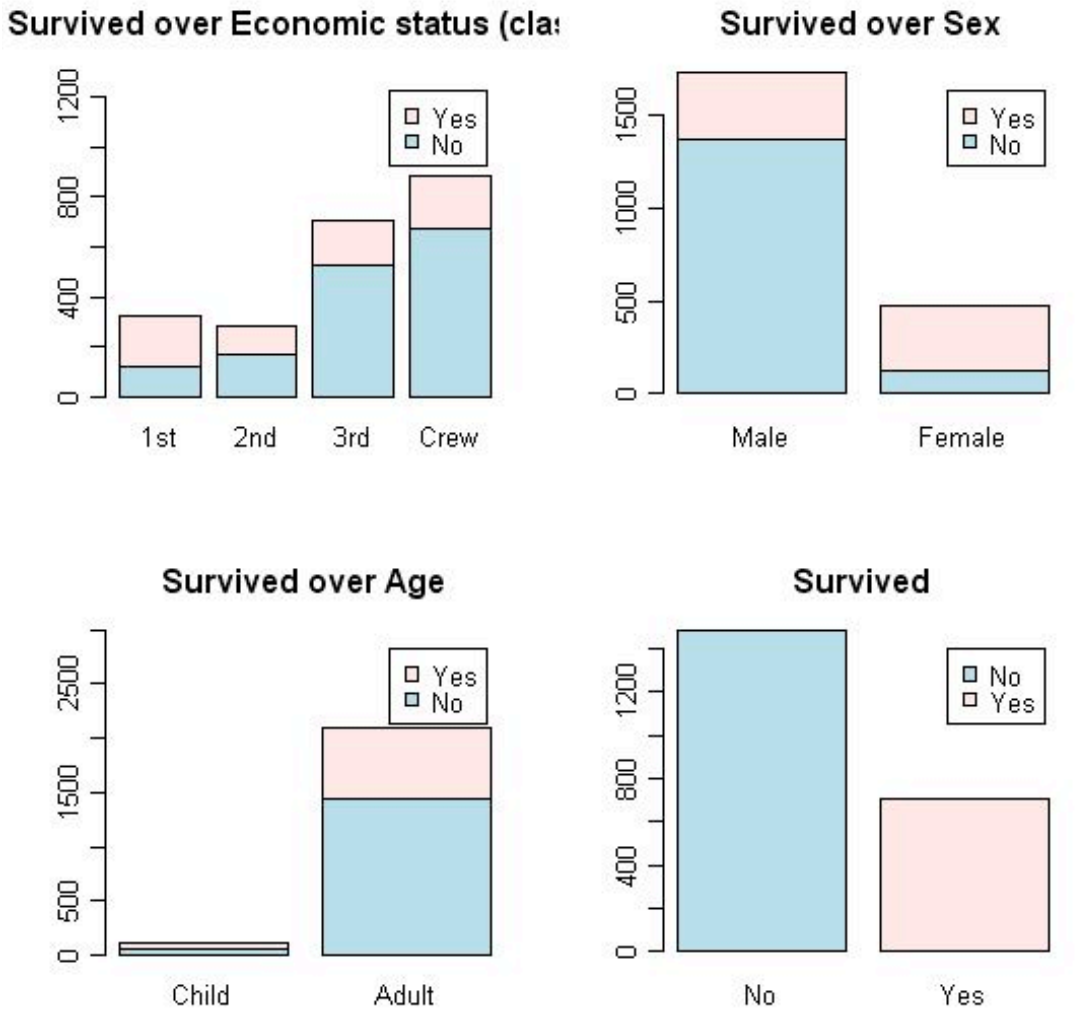
탑승자 중에서 승무원의 수가 가장 많고, 2등석 탑승자가 가장 적었다. 성별로는 남자가 여성보다 3배 가량 많았다. 나이별로는 어른이 아이들보다 월등히 많았으며, 생존자보다 사망자가 두배 정도 많음을 알 수 있다.

R script는 다음과 같다.

```
> par(mfrow=c(2,2))
> barplot(apply(Titanic, 1, sum), main="Economic status (class) distribution")
> barplot(apply(Titanic, 2, sum), main="Sex distribution")
```

```
> barplot(apply(Titanic, 3, sum), main="Age distribution")
> barplot(apply(Titanic, 4, sum), main="Survived distribution")
```

이번에는 barplot으로 객실 Class, 성별, 나이별로 생존자 및 사망자를 비교해보자.




탑승자 중에서 1등석 탑승자가 생존율이 가장 높고, 승무원의 생존율이 가장 낮고 성별로는 여성의 생존율이 높고 나이별로는 아이의 생존율이 높았으며,전체적으로 생존자보다 사망자가 두배 정도 많음을 알 수 있다.

아래의 분할표로도 이 사실을 알 수 있으나 chart보다는 덜 직관적이다.

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

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

      Age
Survived Child Adult
    No     52  1438
    Yes    57   654

 No  Yes
1490  711
```

R script는 다음과 같다.

```
> par(mfrow=c(2,2))

> apply(Titanic,c(4,1),sum)
> barplot(apply(Titanic,c(4,1),sum),col=c("lightblue", "mistyrose"),
    main="Survived over Economic status (class)",ylim=c(0,1300),
    legend=rownames(apply(Titanic,c(4,1),sum)))

> apply(Titanic,c(4,2),sum)
> barplot(apply(Titanic,c(4,2),sum),col=c("lightblue", "mistyrose"),
    main="Survived over Sex",
    legend=rownames(apply(Titanic,c(4,2),sum)))

> apply(Titanic,c(4,3),sum)
> barplot(apply(Titanic,c(4,3),sum),col=c("lightblue", "mistyrose"),
    main="Survived over Age",ylim=c(0,3000),
    legend=rownames(apply(Titanic,c(4,3),sum)))

> apply(Titanic,4,sum)
> barplot(apply(Titanic,4,sum),col=c("lightblue", "mistyrose"),
    main="Survived",
    legend=dimnames(Titanic)$Survived)
```

다시 객실 Class를 1등실, 2등실, 3등실을 묶어 승객이라는 class로 만들어 승무원과 비교해 보자.

- 객실 Class : 승객, 승무원
- 성별 : 남자, 여자

- 나이 : 아이, 어른
- 생존여부 : 사망, 생존

다음에 4차원 array를 2*2*2*2 분할표로 만들었다.

| Adults | Survivors | | Non-Survivors | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| Passenger | 146 | 296 | 659 | 106 |
| Crew | 192 | 20 | 670 | 3 |

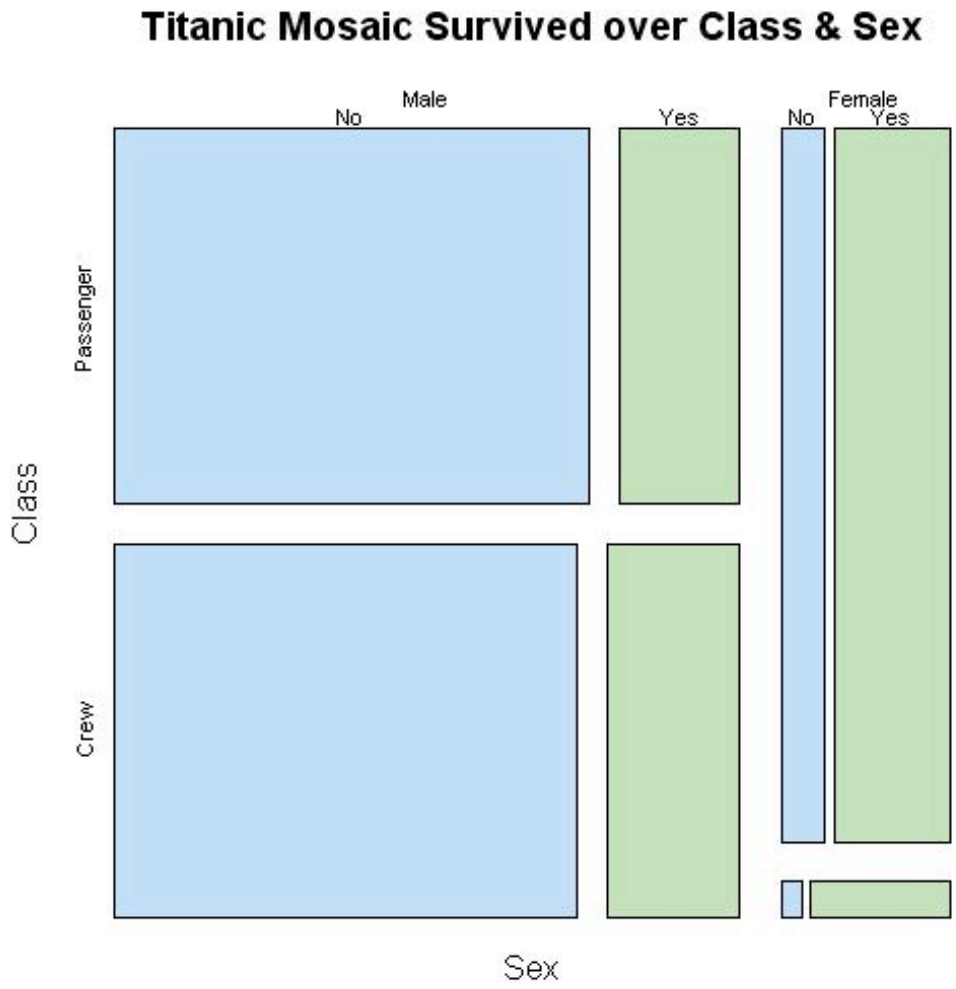
| Children | Survivors | | Non-Survivors | |
|---|---|---|---|---|
| | Male | Female | Male | Female |
| Passenger | 29 | 28 | 35 | 17 |
| Crew | 0 | 0 | 0 | 0 |

```
> temp=c(apply(Titanic[dimnames(Titanic)[[1]]!="Crew",,,],2:4,sum),
                Titanic[dimnames(Titanic)[[1]]=="Crew",,,])
> dim(temp)=c(2,2,2,2)
> dimnames(temp)=list(Sex = c("Male", "Female"),
    Age = c("Child", "Adult"),
    Survived = c("No", "Yes"),
    Class = c("Passenger", "Crew"))
```

승객과 승무원의 생존관계를 알아보기 위해 2*2 분할표를 만들어 보았다.

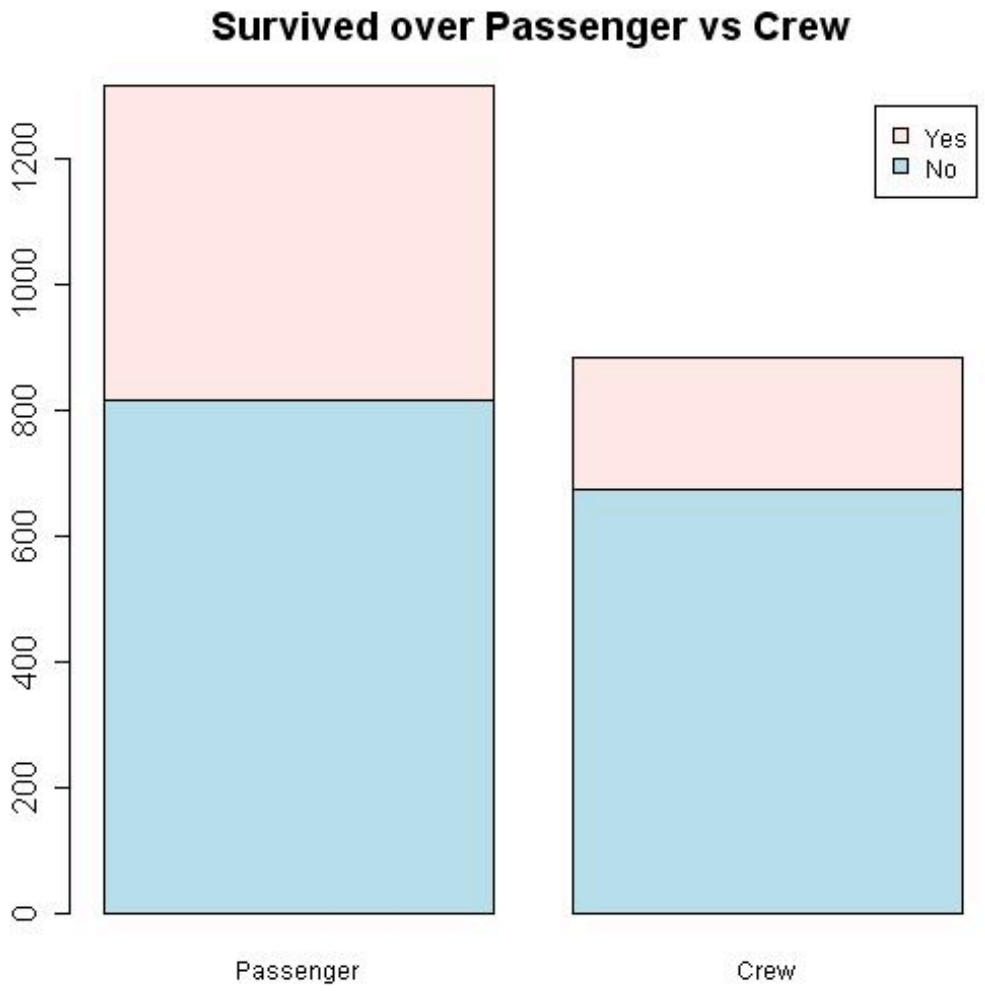
```
> apply(temp,3:4,sum)
       Class
Survived Passenger Crew
    No       817  673
    Yes      499  212
```

승객과 승무원으로 재분류한 타이타닉 자료에서 나이를 제외한 변수들을 이용한 Mosaic Plot은 다음과 같다.




이 Mosaic Plot을 보면, 승무원의 경우 남자의 경우와 여자의 경우 모두 생존율이 승객의 것보다 높게 보인다. 세로축으로 비교해 보면 쉽게 파악된다.

그러면 다음의 Barchart를 보자. 이 Chart는 위 Chart에서 성별 변수를 제거한 후 그린 Chart이다.




이 Barchart를 보면 승무원의 사망율이 승객의 사망율 보다 훨씬 높게 나타나 있다. 그런데 앞서 그린 Mosaic Plot에서는 승무원의 경우 남자의 경우와 여자의 경우 모두 생존율이 승객의 것보다 높게 나타나 있다. 이러한 현상을 심프슨의 파라독스(Simpson-paradox)라 한다.

두 Chart의 R Script는 다음과 같다.

```
> mosaicplot(apply(temp,c(1,4,3),sum),
      main="Titanic Mosaic Survived over Class & Sex",
      col = hcl(c(240, 120)),
      off = c(5, 5, 5))
> barplot(apply(temp,3:4,sum),col=c("lightblue", "mistyrose"),
      main="Survived over Passenger vs Crew",
      legend=rownames(apply(temp,3:4,sum)))
```