다변량 데이터의 시각화

그래프 방법의 장점

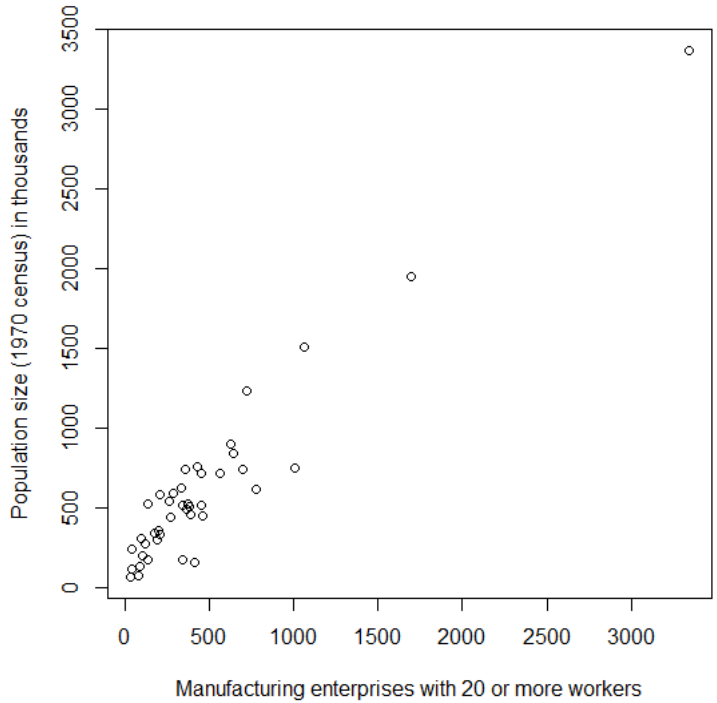
- 독자의 관심과 주의를 끄는데 효율적
- 시각적 관계는 더 쉽게 파악, 기억
- 시간절약
- 종합적 그림 제공

"인간은 실제로 그곳에 존재하는 미묘한 패턴을 잘 구별한다. 그러나 마찬가지로 패턴이 존재하지 않을 때도 상상을 잘 한다."

- 단변량 변수
  - ✓양적변수: 히스토그램, 상자그림
  - ✓범주형변수: 막대그래프, 파이차트
- 둘 혹은 세 변수 사이의 관계
  - ✓양적 vs. 양적: 산점도, 산점도행렬, 버블차트, 분포 시각화
  - ✓범주형 vs. 범주형: mosaic plot
  - ✓양적 vs. 범주형: 상자그림'들'
- 다변량 변수 시각화
  - ✓별그림, 나이팅게일차트, heatmap,

산점도

- 미국의 대기오염 데이터 USairpollution

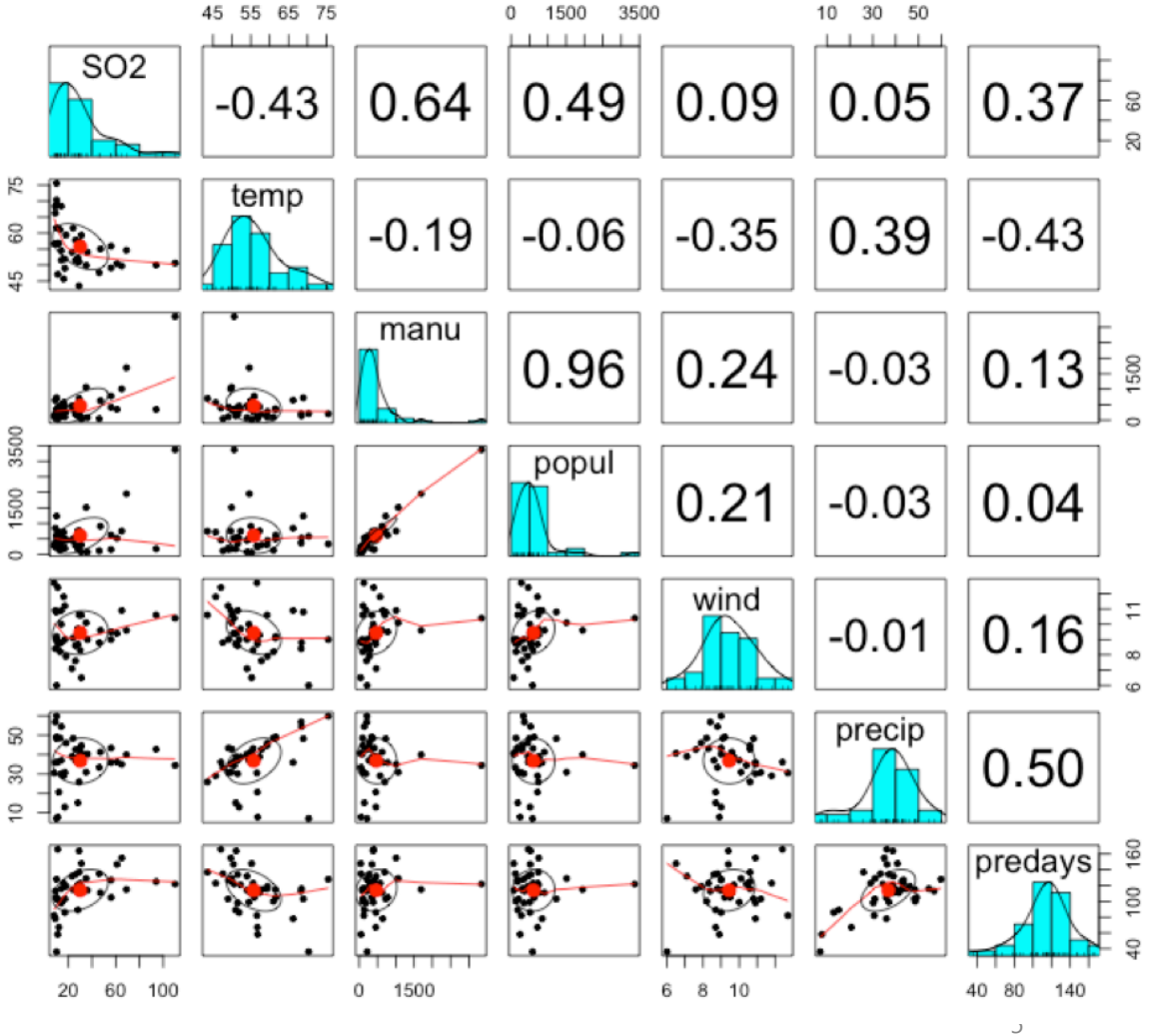



```
> mlab <- "Manufacturing enterprises with 20 or more workers"

> plab <- "Population size (1970 census) in thousands"

> plot(popul ~ manu, data = USairpollution,
+      xlab = mlab, ylab = plab)
```

산점도 행렬

- 이변량 산점도와 단변량 분포를 함께 그리는 것 좋음
  - ✓이상값 식별 용이
    - 이상값에 해당하는 점을 다른 색이나 모양으로 표현
    - Identify 함수 사용

```
> library(psych)
Warning message:
package 'psych' was built under R version 3.1.3
> pairs.panels(USairpollution)
```



상관계수

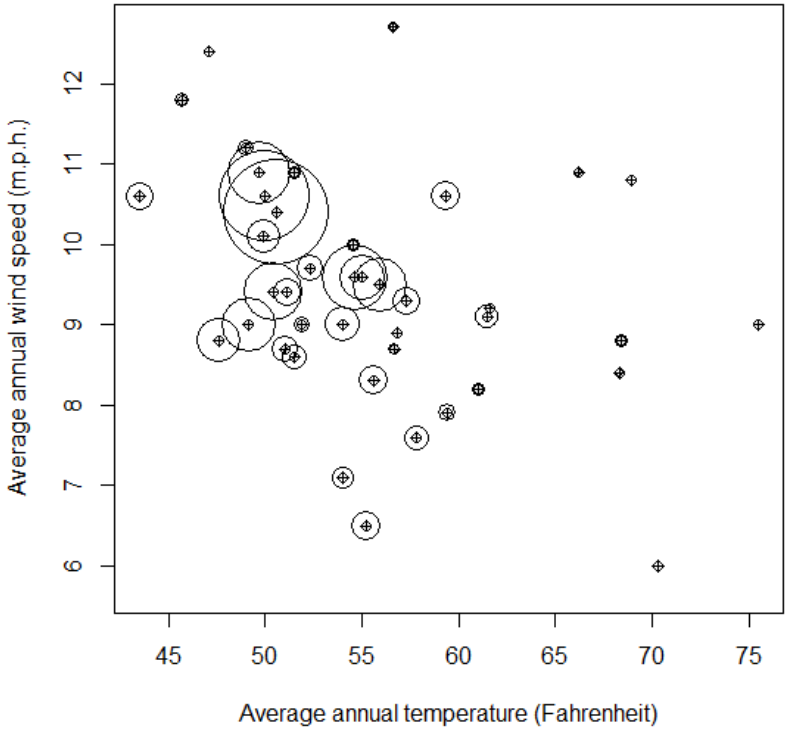
- 이상치 포함 전 후의 상관계수의 차이 관찰
- 네 개의 이상치를 제거한 후 상관계수가 0.96➔ 0.80으로 감소

```
> with(USairpollution, cor(manu, popul))
[1] 0.9552693
> outcity <- match(c("Chicago", "Detroit", "Cleveland", "Philadelphia"),
+                  rownames(USairpollution))
> with(USairpollution, cor(manu[-outcity], popul[-outcity]))
[1] 0.7955549
```

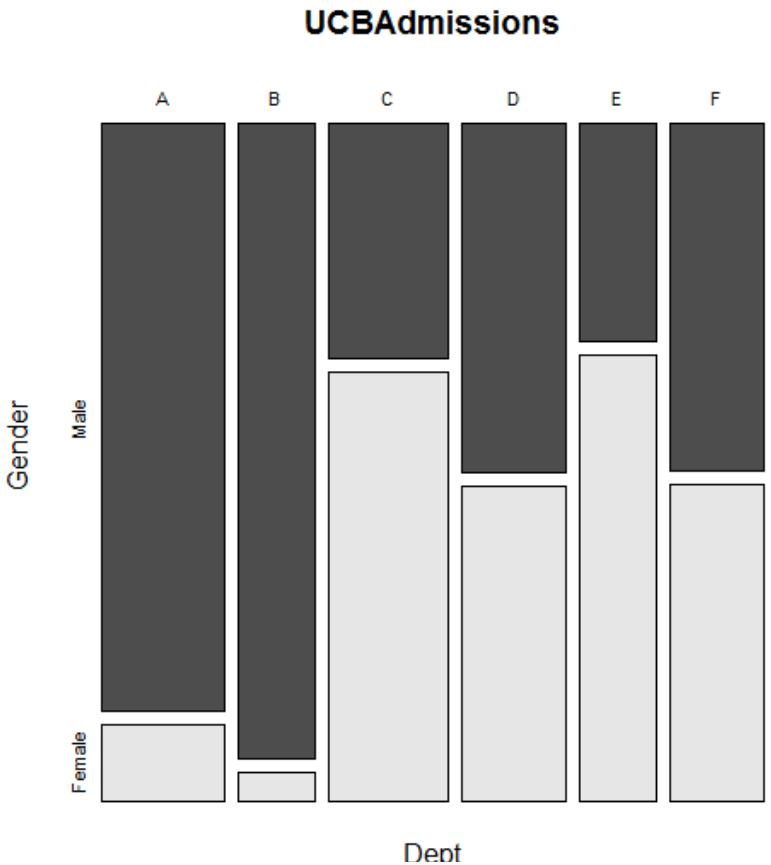
버블차트

- 두 개의 변수를 사용한 산점도에 세번째 변수의 값을 함께 표현
- temp, wind, SO2 의 관계
  - ✓적절한 연간 온도와 풍속을 갖는 도시들의 대기 오염이 심함

```
ylim <- with(USairpollution, range(wind)) * c(0.95, 1)

plot(wind ~ temp, data = USairpollution,
       xlab = "Average annual temperature (Fahrenheit)",
       ylab = "Average annual wind speed (m.p.h.)", pch = 10,
       ylim = ylim)

with(USairpollution, symbols(temp, wind, circles = SO2,
                             inches = 0.5, add = TRUE))
```



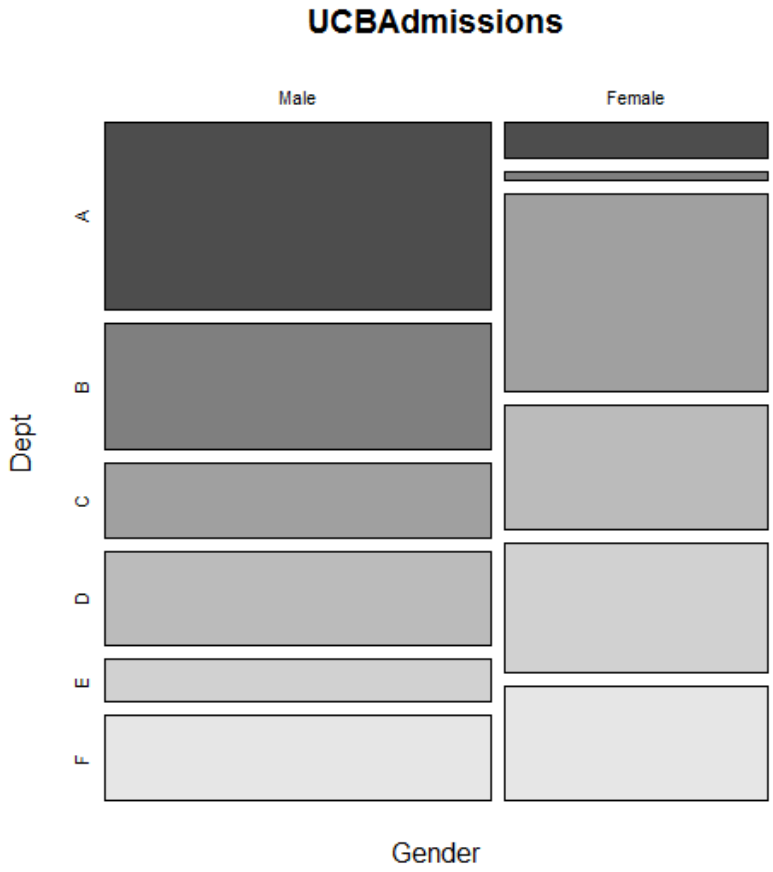
Mosaic plot



- **<u>지원학과별 남녀 빈도비교</u>**
- 두 범주형 변수의 관계를 표현
- 행과 열에 어느 변수를 사용하느냐에 따라 다른 비교 가능
- 막대폭=지원학과의 빈도에 비례
- 막대길이=학과 내 성별 빈도에 비례

```
mosaicplot(~Dept+Gender,data=UCBAdmissions,color=T)
```

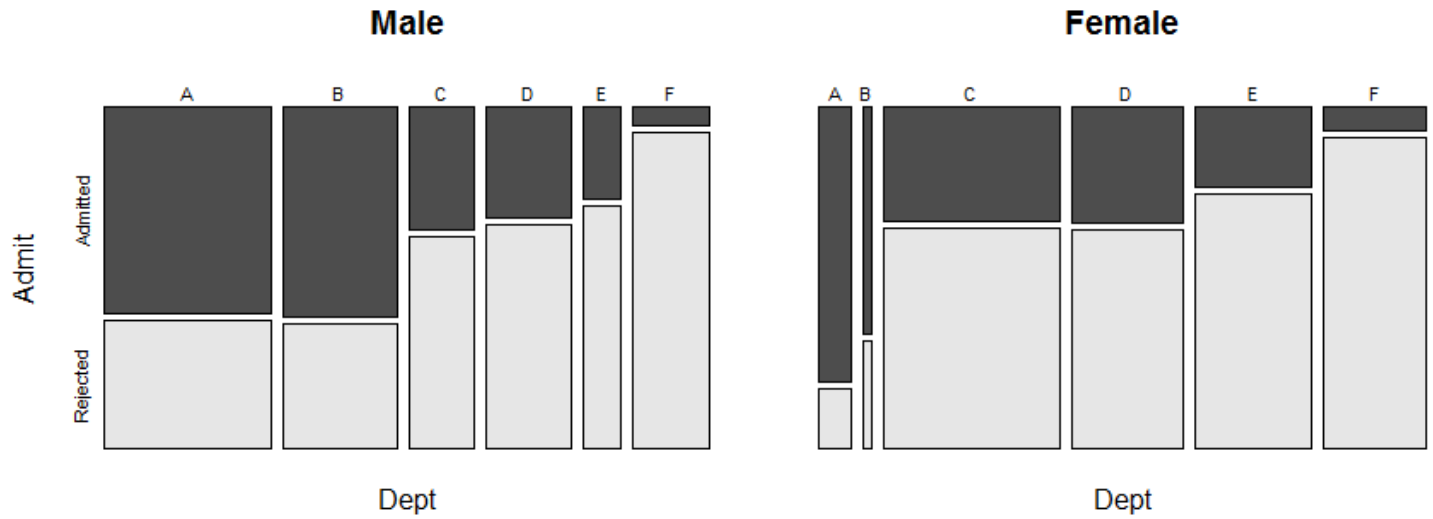
Mosaic plot



- **<u>남여 별 지원학과 빈도비교</u>**
- 막대폭=남녀의 빈도에 비례
- 막대길이=학과 내 성별 빈도에 비례

Mosaic plot

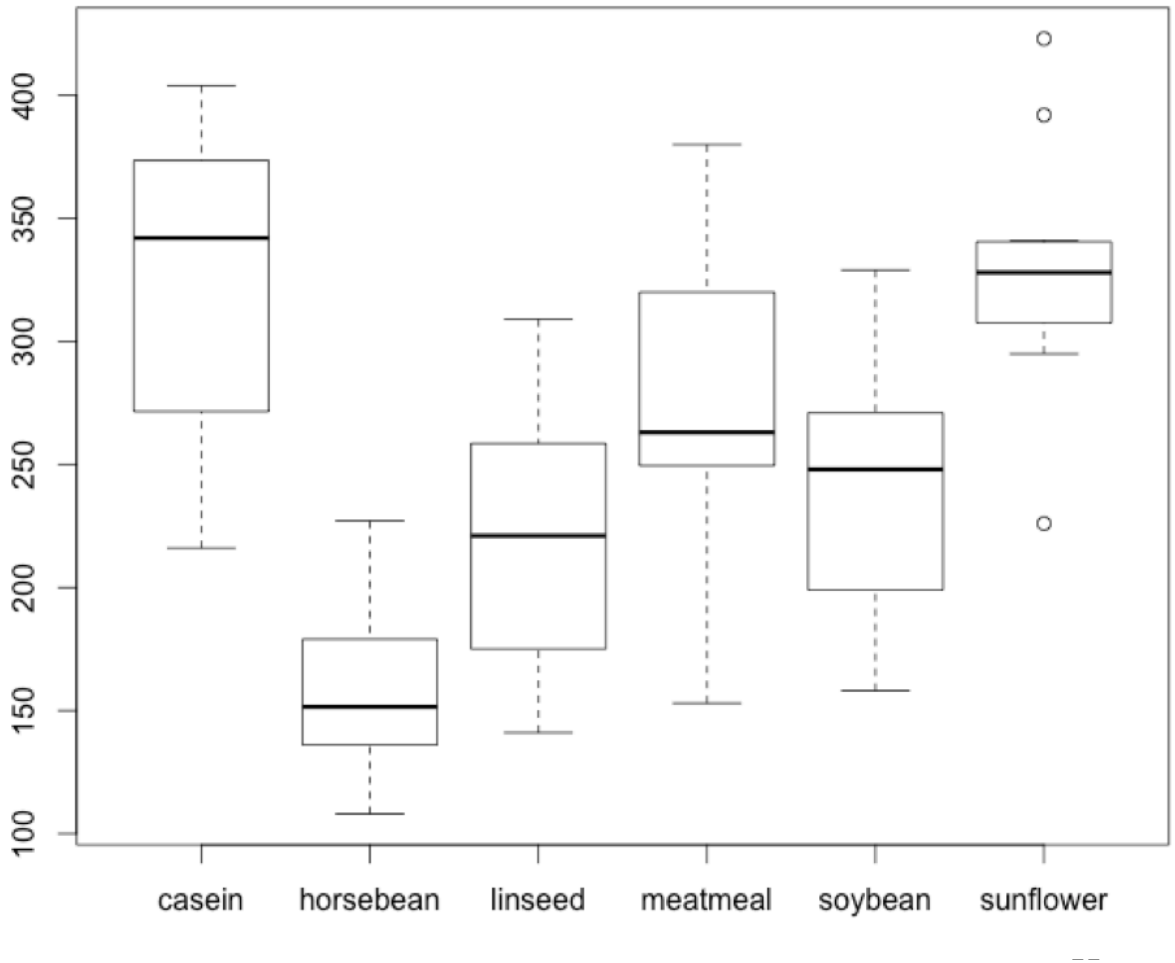
- **지원학과별 남녀의 합격률 비교**




```
tab.m=as.table(tab.m)
tab.f=as.table(tab.f)
mosaicplot(~Dept+Admit,data=tab.m,color=T,main="Male")
mosaicplot(~Dept+Admit,data=tab.f,color=T,main="Female")
```

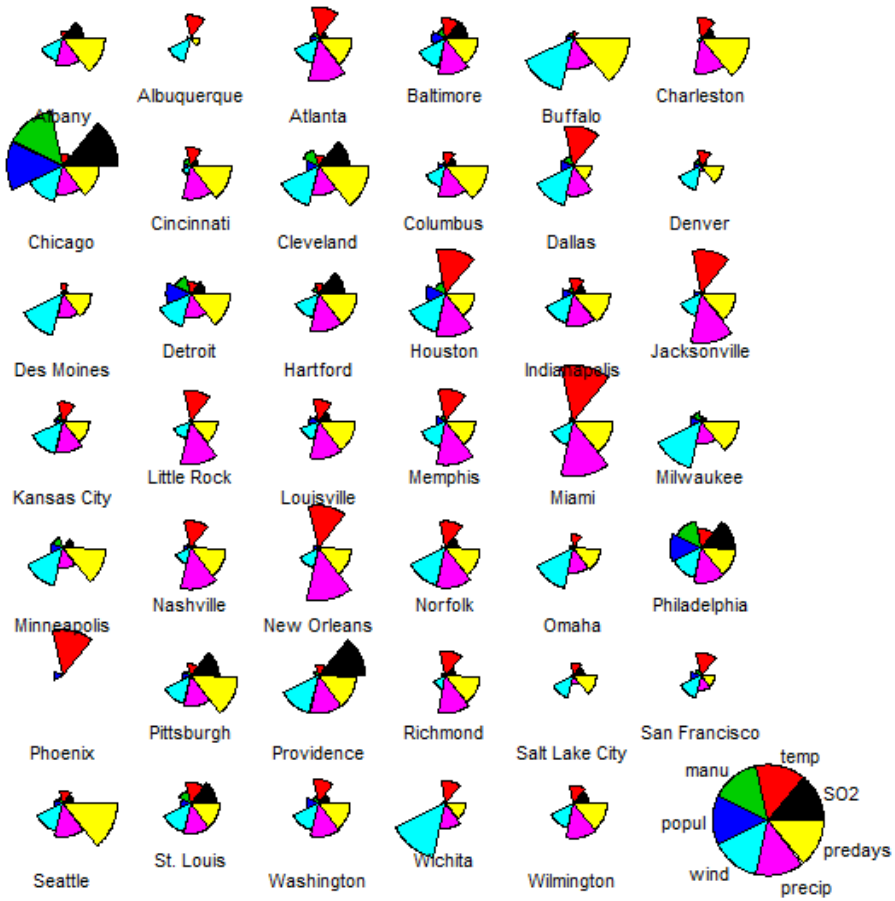
상자그림

```
> boxplot(weight~feed,chickwts)
```



Nightingale's Chart

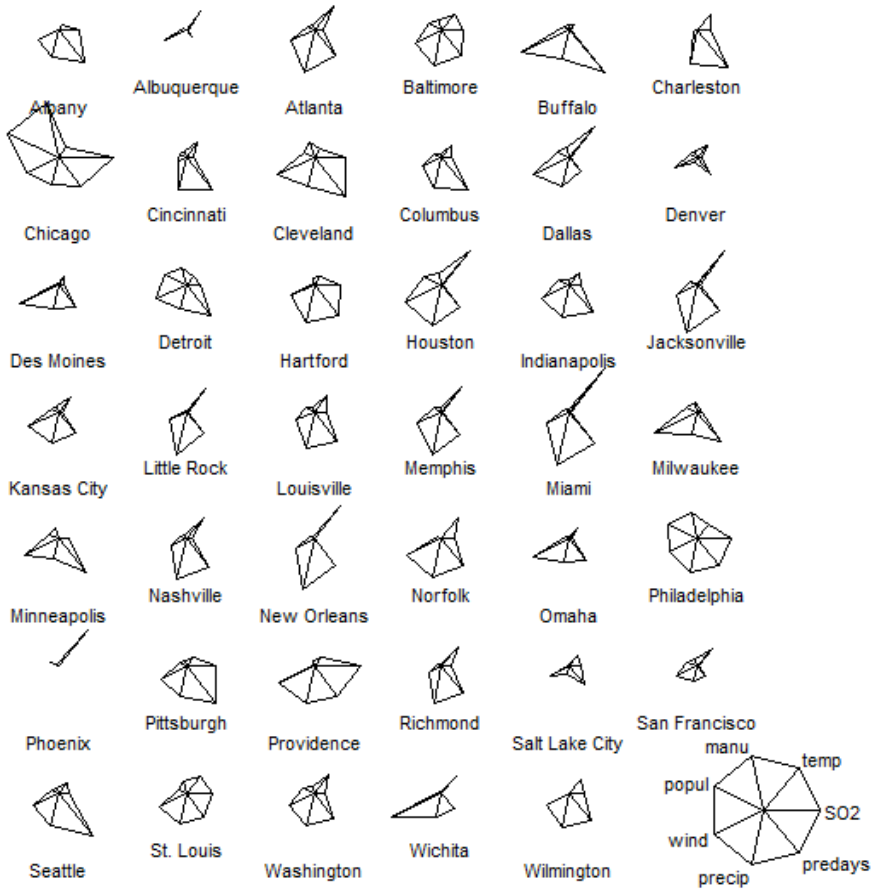
```
stars(USairpollution, cex=0.7,key.loc = c(15, 2), draw.segment=TRUE)
```



Star Plot

- 각 도시를 위한 7개의 변수를 7개의 변을 갖는 별로 표현
  - ✓뉴올리언즈, 마이애미, 잭슨빌, 아틀란타: 연간온도가 높은 특이한 모양

```
stars(USairpollution, cex=0.7,key.loc = c(15, 2))
```

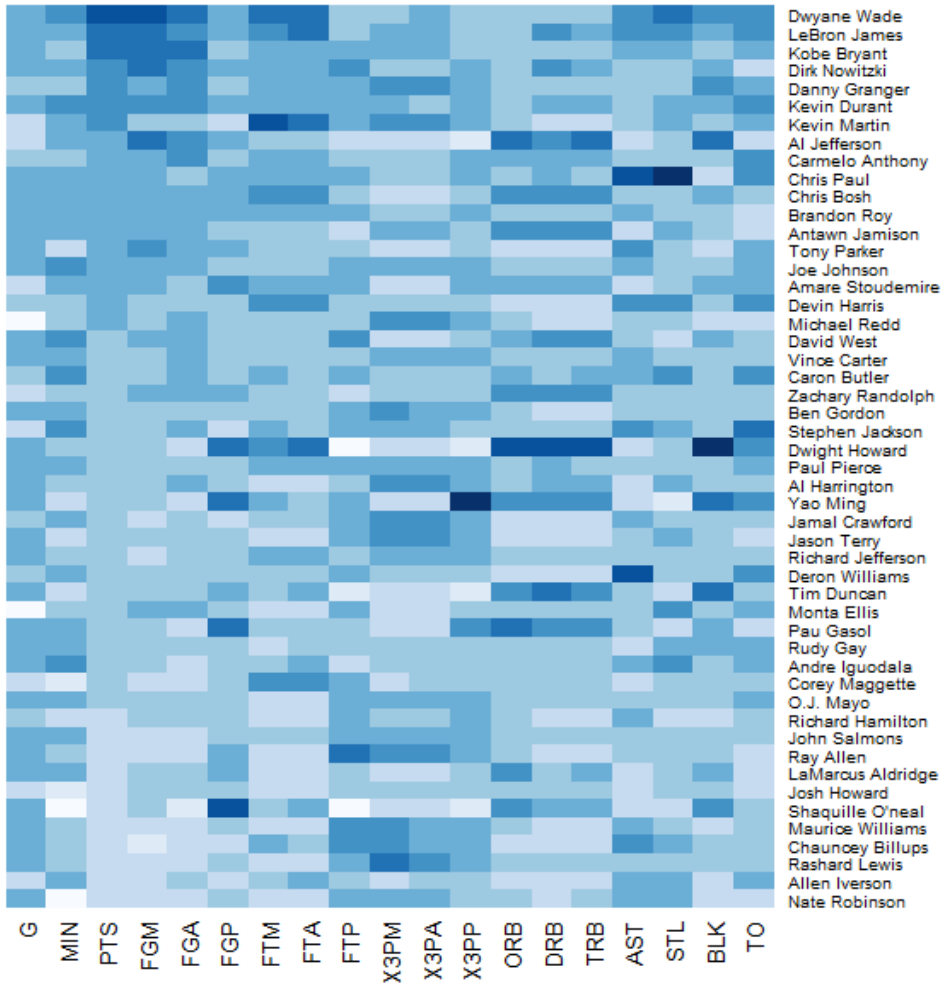


히트맵 (Heatmap)

- 색상으로 값들의 높고 낮은 관계를 표시
- 데이터가 지나치게 많은 경우 여전히 혼란스러움
  - ✓색상 선택
  - ✓정렬 과정

```r
bball=read.csv("http://datasets.flowingdata.com/ppg2008.csv")
rownames(bball)=bball[,1]
bball=bball[,2:20]
bball=as.matrix(bball)

bball=bball[order(bball[,3],decreasing=FALSE),]
library(RColorBrewer)
heatmap(bball,Rowv=NA,Colv=NA,scale="column",margins=c(5,10),col=brewer.pal(9,"Blues"))
```

- heatmap 함수는 input이 matrix 형태이어야 함
- 득점(PTS)을 기준으로 내림차순으로 정렬



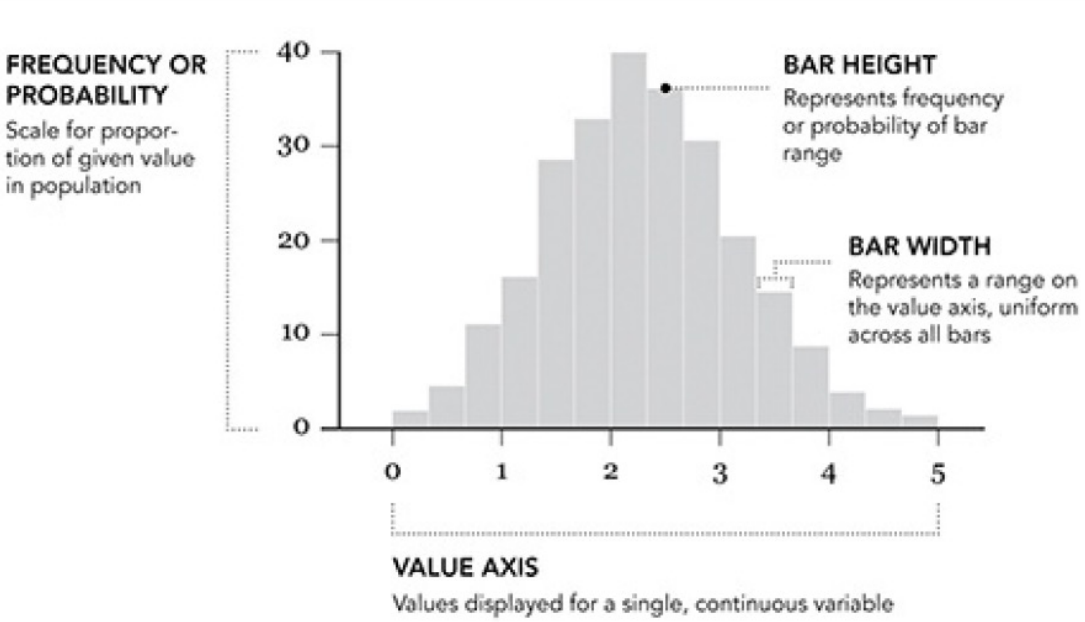
분포 시각화

- 분포에 대한 요약통계량
  - ✓평균, 중앙값, 최빈값, Q1, Q3
  - ✓표준편차, 분산, IQR

- 분포의 시각화
  - ✓히스토그램
  - ✓Boxplot
  - ✓연속밀도함수

단변량 자료의 분포: 히스토그램 (Histogram)

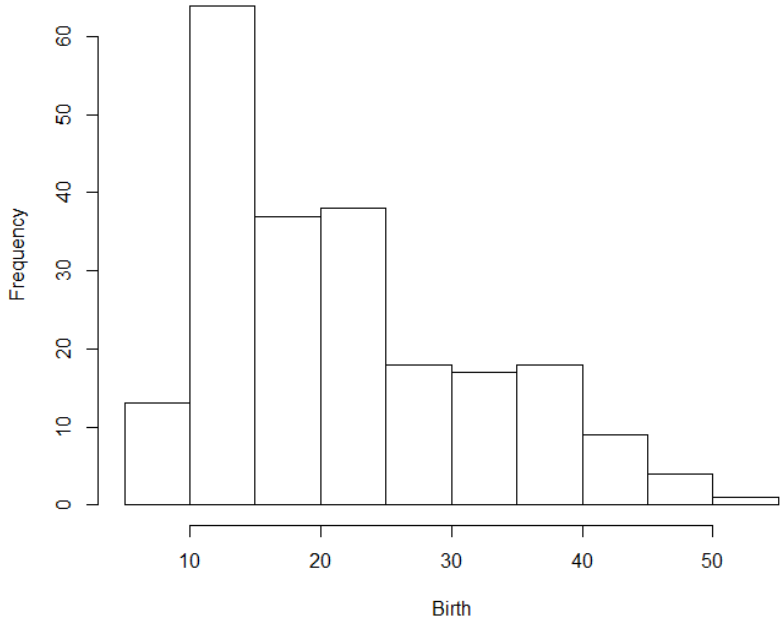
- 어느 영역에 데이터가 몰려있는지 눈으로 확인
- 가로축의 변수가 연속적이므로 막대 사이의 간격이 없음
- 그래픽에 익숙치 않은 사람들은 가로축을 시간으로 생각하는 오류를 범할 수 있으므로 설명

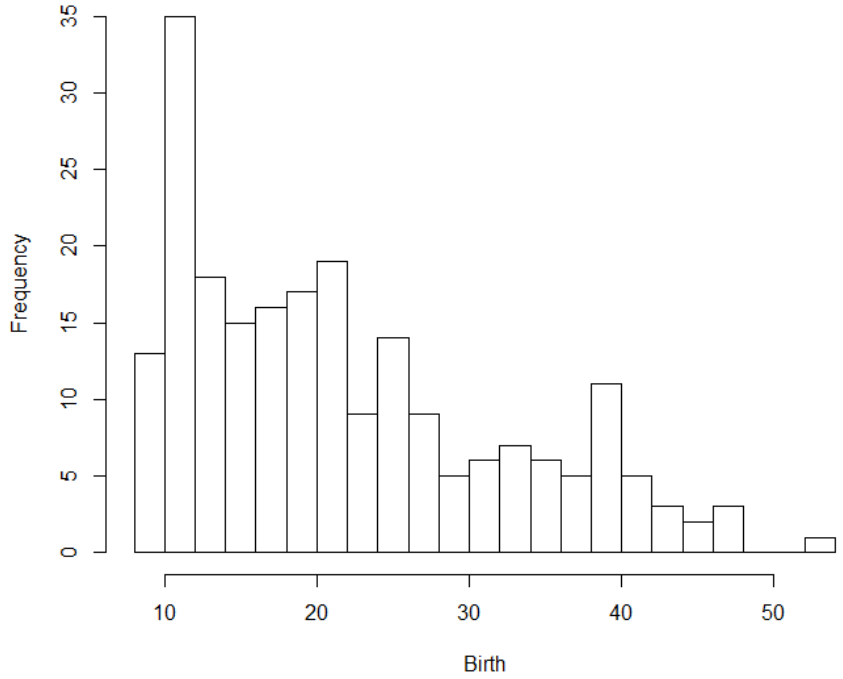


- 세계의 출산율 분포

```
> birth=read.csv("http://datasets.flowingdata.com/birth-rate.csv")
> head(birth)
                Country    X1960  X1961  X1962  X1963  X1964    X1965  X1966
1                 Aruba 36.40000 35.179 33.863 32.459 30.994 29.51300 28.069
2           Afghanistan 52.20100 52.206 52.208 52.204 52.192 52.16800 52.130
3                Angola 54.43200 54.394 54.317 54.199 54.040 53.83600 53.585
4               Albania 40.88600 40.312 39.604 38.792 37.913 37.00800 36.112
5 Netherlands Antilles 32.32100 30.987 29.618 28.229 26.849 25.51800 24.280
```

```
hist(birth$X2008,xlab="Birth",main=NA)
hist(birth$X2008,breaks=20,xlab="Birth",main=NA)
```

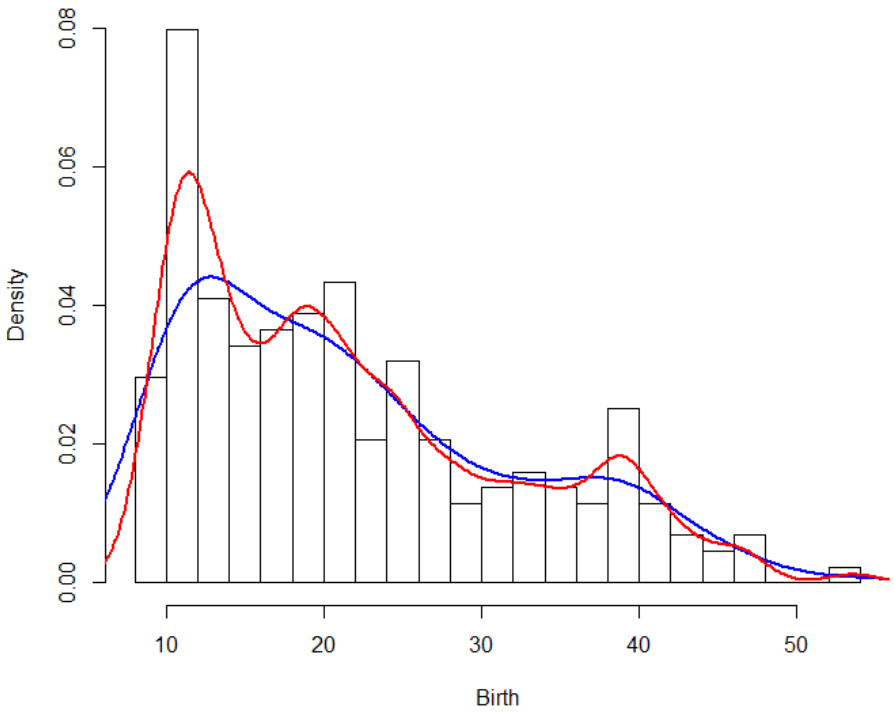




- 히스토그램의 구간 수의 기본값이 언제나 최상은 아님
- 다양한 구간 수의 비교를 통해 데이터를 가장 잘 설명해주는 값을 찾는 것이 바람직

단변량 자료의 분포: 연속밀도함수

- 히스토그램의 막대 대신 연속된 선으로 분포 표현



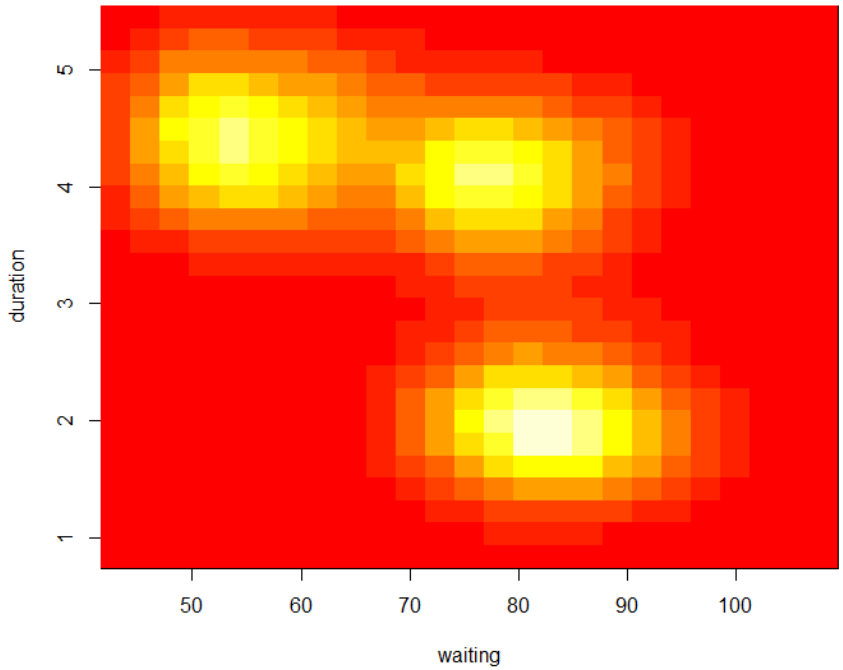
- 히스토그램의 구간 수를 조정하는 것과 유사하게 density 함수의 adjust 옵션을 통해 연속밀도함수의 smoothness 를 조정 가능

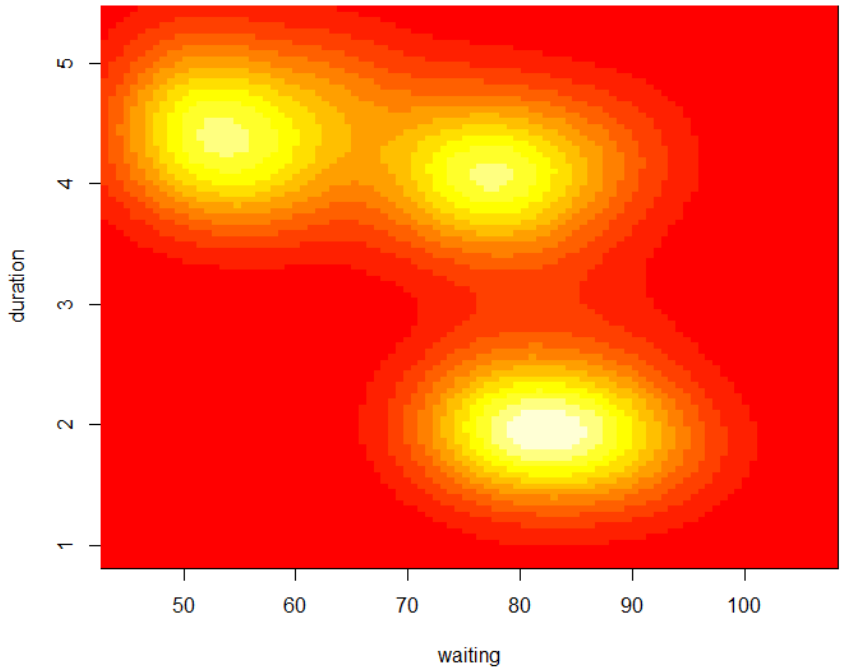
```
birth2008=na.omit(birth$X2008)
hist(birth2008,breaks=20,xlab="Birth",main=NA,freq=FALSE)
lines(density(birth2008),col="blue",lwd=2)
lines(density(birth2008,adjust=0.5),col="red",lwd=2)
```

이변량 자료의 분포

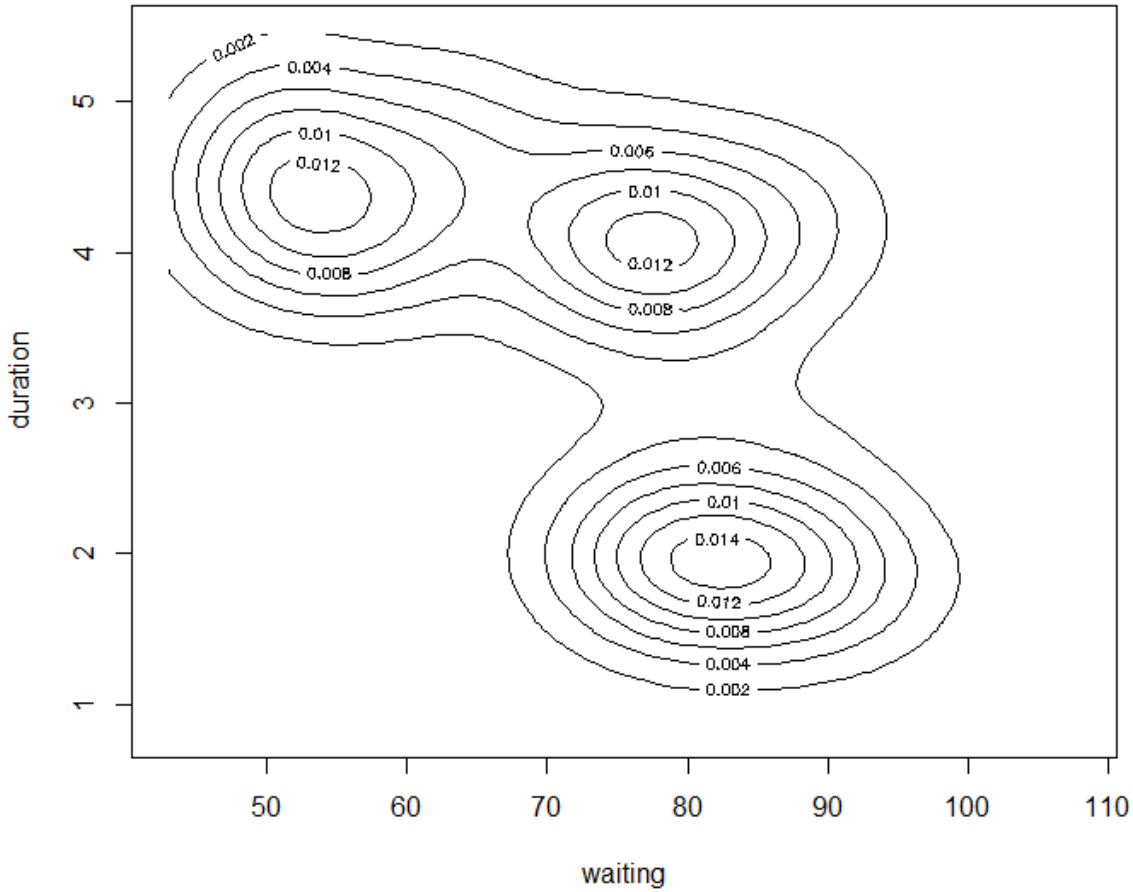
- 미국의 Yellowstone 국립공원의 간헐 온천 분출의 지속시간(duration)과 해당 분출까지 waiting time의 자료 (in minutes)

```
library(MASS)
attach(geyser)
density1=kde2d(waiting,duration, n=25)
image(density1,xlab="waiting",ylab="duration")

density2=kde2d(waiting,duration, n=100)
image(density2,xlab="waiting",ylab="duration")
```



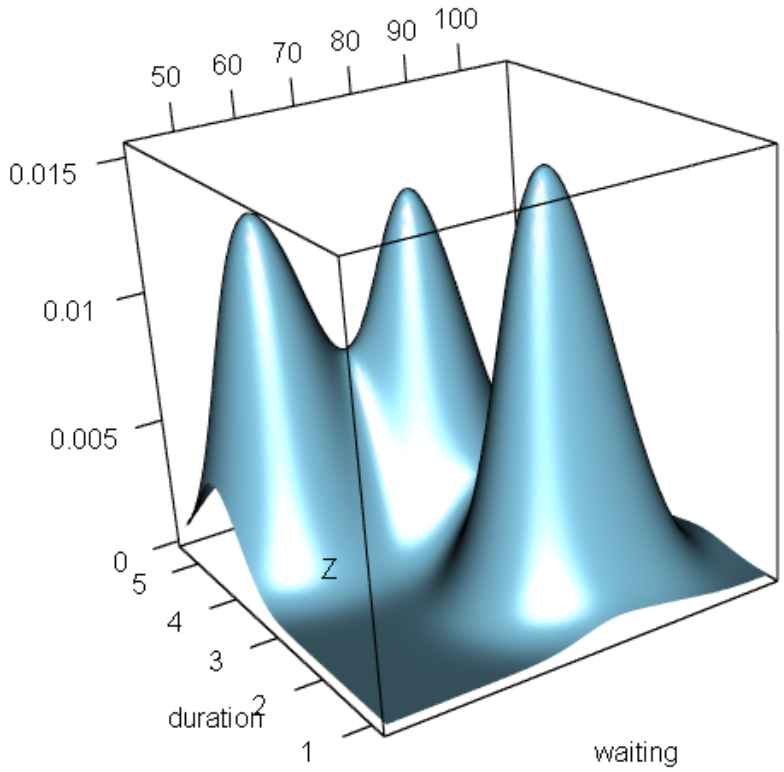


- 등고선 (Contour plot)




```
contour(density2)
```

- 3차원 분포그래프




```
persp3d(density2,back="lines",col="skyblue",xlab="waiting",ylab="duration")
```