# 12강 다변량 자료의 시각화 이해 I

정보통계학과 이태림 교수

1. 다변량 자료의 시각화를 이해한다
2. 별그림을 이용하여 다변량 자료를 표현할 수 있다
3. 다변량 자료의 상관성 검토를 위한 산점도 행렬을 작성하고 해석할 수 있다
4. 다변량 시각화의 R에 의한 표현

# 학습개요(다변량 자료의 시각화 I)

다변량자료의 시각화

별그림(star plot)

산점도 행렬
(scatter plot matrix)

- 변수가 3개 이상인 다변량 자료
- 다변량 자료는 다차원의 자료

- 별그림(star plot)
- 파이조각그림
  (pie segments plot)

- 산점도 행렬
- 조건부 플롯

# 1. 다변량 자료의 정의

# 1 다변량 자료의 정의

## 다변량 자료 (Multivariate Data Analysis) :

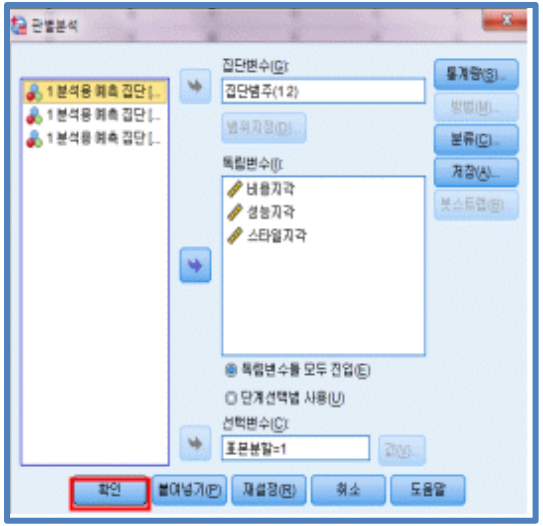
MDA

변수가 3개 이상인 자료의 분석

- 3차원을 넘는 시각화 정보는 처리하기가 어렵다
- 평면상에서의 삼차원 공간표현을 위한 조처가 필요

# 다변량 자료의 예제(판별분석)

- Y= 집단범주 (1=구매자) (2=비구매자)
- X1=비용지각 X2=성능지각 X3=스타일지각 X4=표본분할상태 (1=분석표본) (2=검증표본)
- 비용지각, 성능지각, 스타일지각에서 0=매우 불량 10=매우 훌륭임
- 집단범주 : (1=구매자) (2=비구매자)
- 비용지각, 성능지각, 스타일지각에서 0=매우 불량 10=매우 훌륭임
- 표본분할 : (1=분석표본) (2=검증표본)

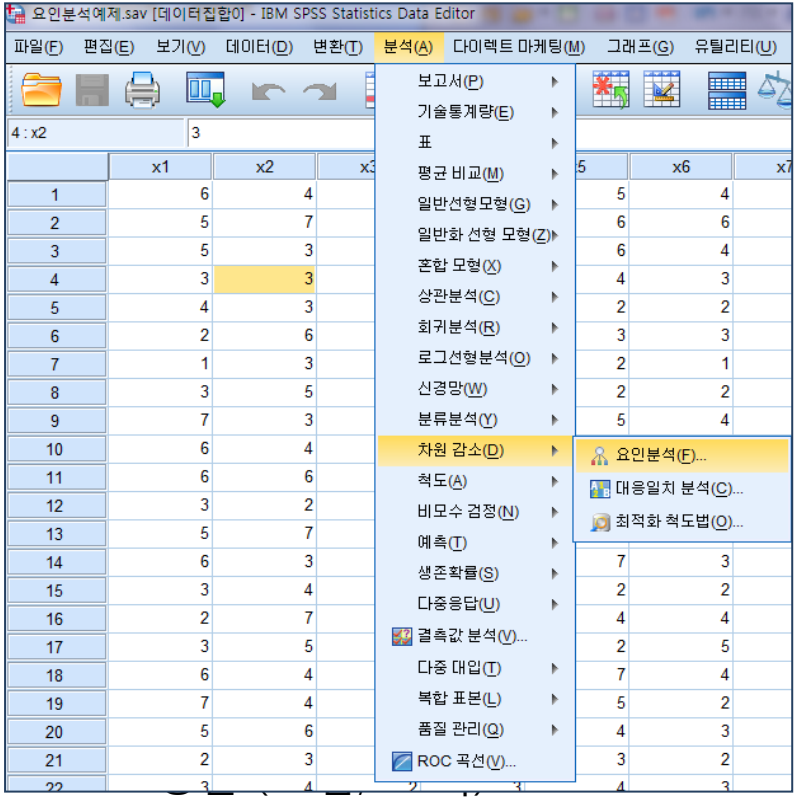


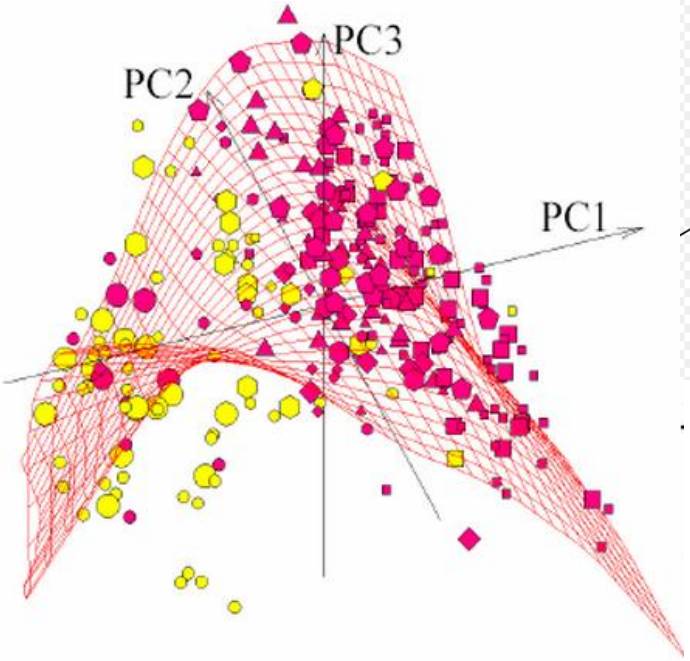
# 다변량 자료의 예제(요인분석)

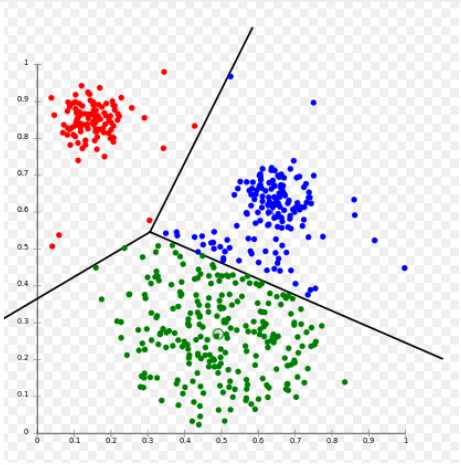
♣ 데이터는 구내식당에 대한 학생들의 반응을 조사한 설문결과.

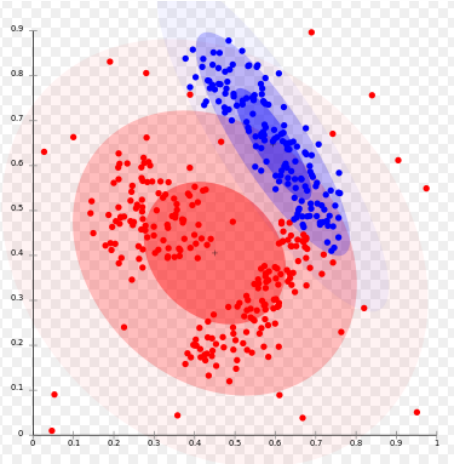
- 설문1. 귀하는 구내식당을 이용할 때 다음의 항목들을 어느 정도 중요하게 생각하십니까?
- x1=청결상태, x2=음식량, x3=대기시간, x4=음식 맛, x5=친절
- 설문2. 귀하는 구내식당을 평균적으로 일주일에 몇 번 정도 이용하십니까?
- x6=이용횟수
- 설문3. 귀하의 성별
- x7=성별 (1: 남, 2: 여)

# 다변량 자료의 예제(요인분석)

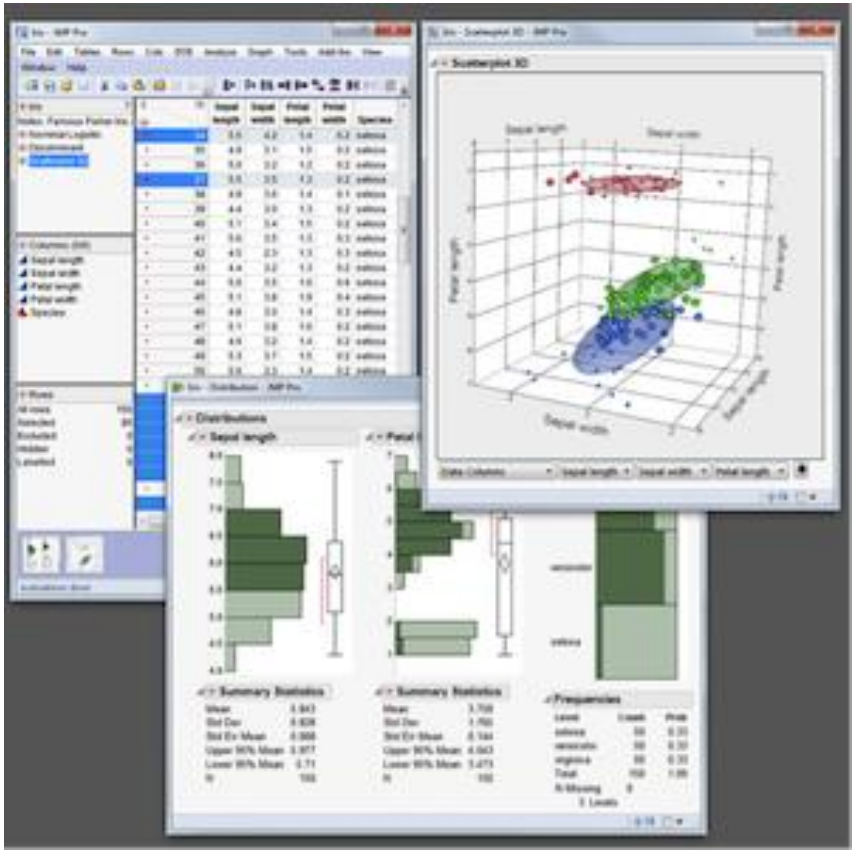



PC3
PC2
PC1

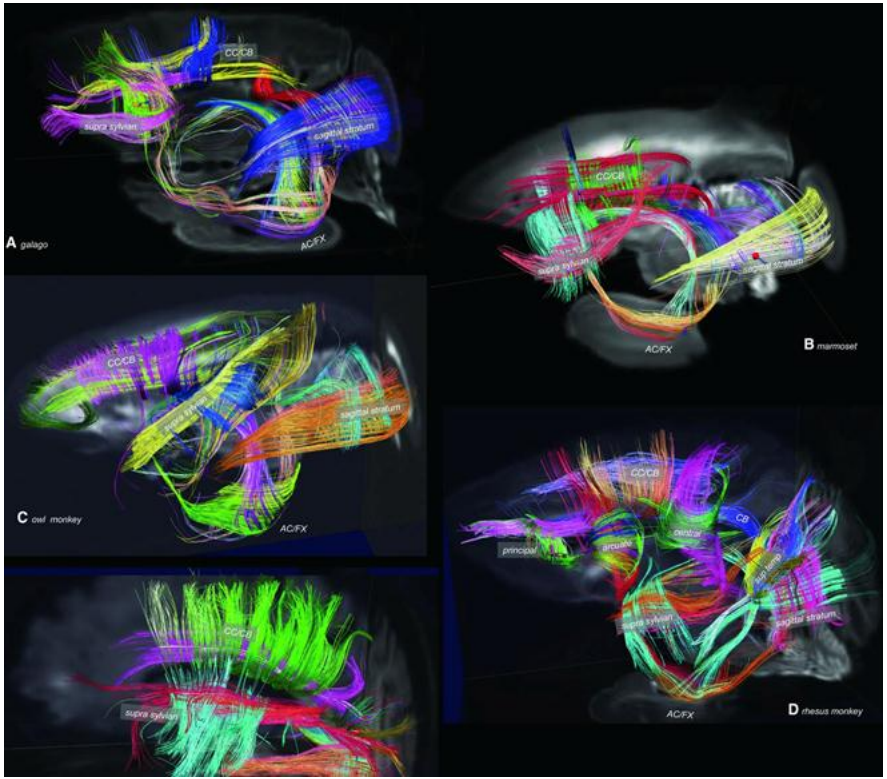


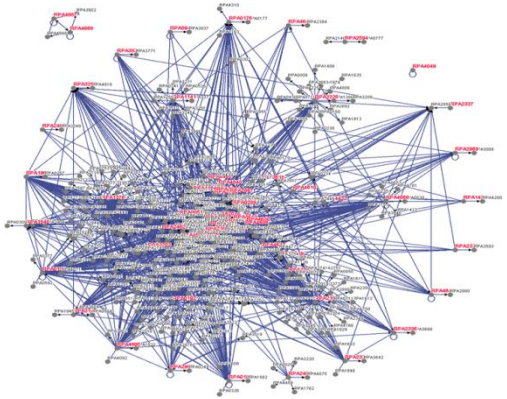


# 다변량 자료의 예제

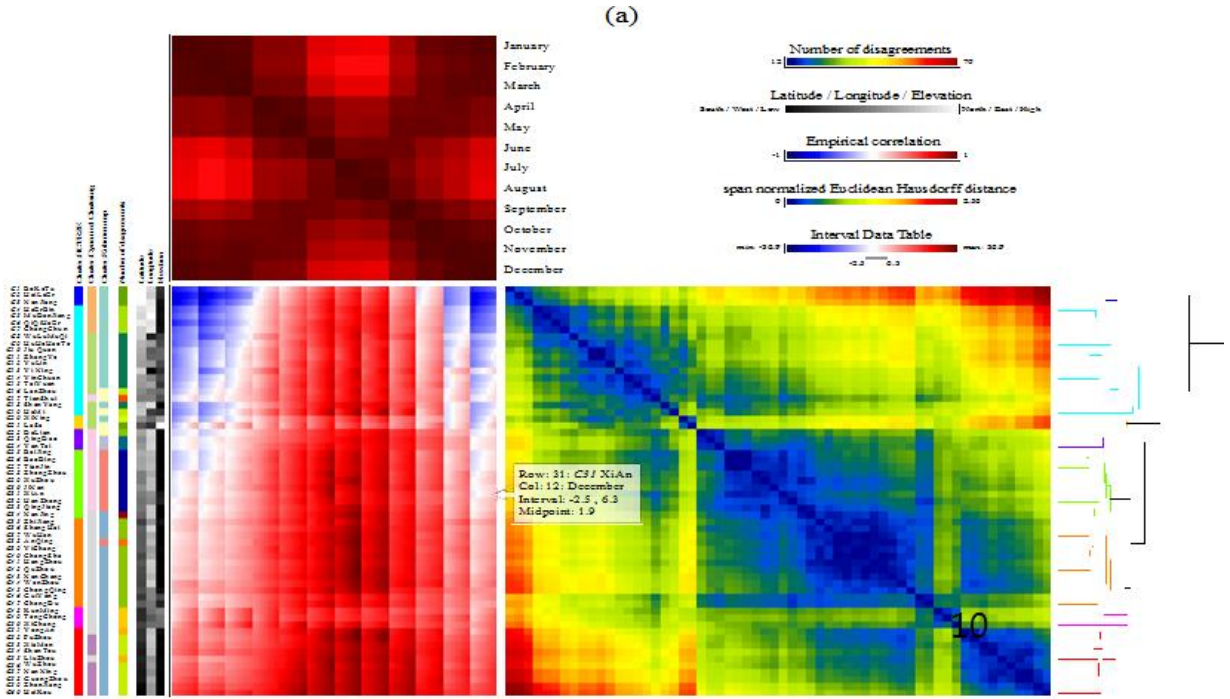

Scatterplot 3D
Distributions
Sepal length
Summary Statistics
Frequencies

# GAP에 의한 뇌기전 그래프



# 3.1 Meteorological Stations in China



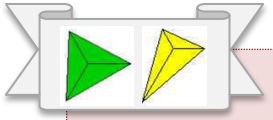
# 2. 별그림

# 2 별그림

## 별그림(Star plot) :

**Star plot**

다변량을 동시에 별의 형태로 나타낸 그림



- x, y, z 세 축으로 나타내지는 3차원의 경우 시각화하고자 하는 화면이 평면인 2차원이므로 시각화를 위해서는 평면상에서의 삼차원 공간표현을 위한 조처가 필요

# 2 별그림

## 별그림 예제

우리나라 8개 프로야구 구단의 2006년도 전력 자료

| 구단 | 팀타율 | 출루율 | 팀방어율 |
|---|---|---|---|
| 현대 | 0.264 | 0.347 | 3.26 |
| 한화 | 0.255 | 0.329 | 3.34 |
| 삼성 | 0.249 | 0.332 | 3.05 |
| SK | 0.265 | 0.343 | 3.63 |
| KIA | 0.253 | 0.323 | 3.29 |
| 두산 | 0.235 | 0.303 | 3.07 |
| LG | 0.254 | 0.317 | 4.47 |
| 롯데프로 | 0.218 | 0.293 | 3.63 |

# 2 별그림

## 별그림 예제

우리나라 8개 프로야구 구단의 2006년도 전력 자료

- 타율과 출루율은 클수록 좋고 방어율은 작을수록 좋기 때문에 3개 변수를 비교할 수 있는 공통의 스케일로 표시되도록 변수 값들이 0.2과 1사이의 공통 범위로 변환하여 0과 1 사이의 범위값을 취하도록 다음 식으로 변환

# 2 별그림

## 별그림 예제

우리나라 8개 프로야구 구단의 2006년도 전력 자료

- 팀 타율 〉〉 0.2+0.8 (팀 타율 – 최소값)/(최대값–최소값)
- 팀 방어율
  〉〉 0.2+0.8×(최대값– 팀 방어율)/(최대값–최소값)
- 출루율 〉〉 0.2+0.8×(출루율 – 최소값)/(최대값–최소값)
- 변수변환 후 각 변수는 최대값 1, 최소값 0을 취한다
- 팀 타율이 크고 방어율이 작고 출루율이 클수록 큰 삼각형 형성

| 구단 | 팀타율 | 출루율 | 팀방어율 |
|---|---|---|---|
| 현대 | 0.264 | 0.347 | 3.26 |
| 한화 | 0.255 | 0.329 | 3.34 |
| 삼성 | 0.249 | 0.332 | 3.05 |
| SK | 0.265 | 0.343 | 3.63 |
| KIA | 0.253 | 0.323 | 3.29 |
| 두산 | 0.235 | 0.303 | 3.07 |
| LG | 0.254 | 0.317 | 4.47 |
| 롯데프로 | 0.218 | 0.293 | 3.63 |

# 2 별그림

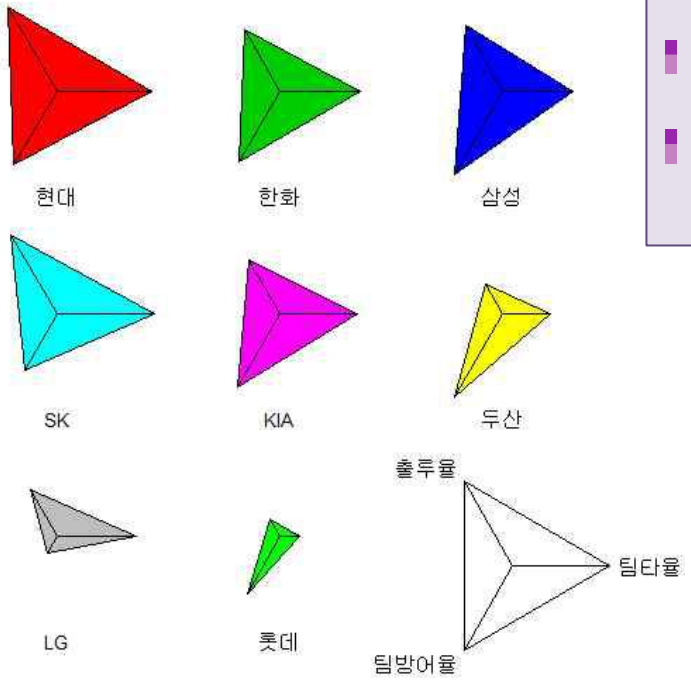
## 별그림 작성을 위한 R 프로그램

```
baseball <- read.csv("프로야구 20060602.csv", header=T)
x <- baseball[,2:4]
x[,1] <- 0.2+0.8*(x[,1]-min(x[,1]))/(max(x[,1])-min(x[,1]))
x[,2] <- 0.2+0.8*(x[,2]-min(x[,2]))/(max(x[,2])-min(x[,2]))
x[,3] <- 0.2+0.8*(max(x[,3])-x[,3])/(max(x[,3])-min(x[,3]))
rownames(x) <- baseball[,1]
stars(x, scale=F, key.loc = c(7,2), col.stars=2:9)
```

```
x11(); stars(x, scale=F, draw.segments=T, full=F,
key.loc = c(7,2))
```

# 2 별그림

## 별그림 작성



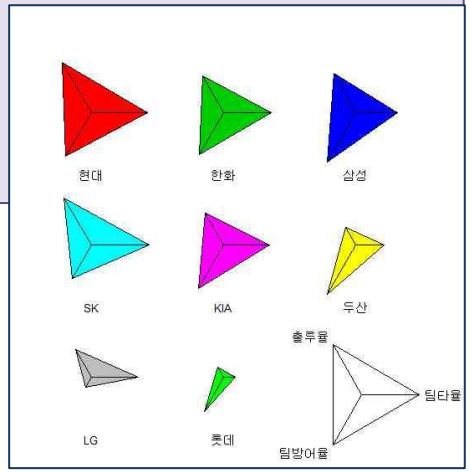

- 롯데와 LG는 작은 삼각형을 형성
- 삼성은 방어율은 좋지만 타율과 출루율에서 저조
- SK는 반대로 방어율이 좋지 않지만 타율과 출루율이 좋음

# 2 별그림

## 별그림 작성

우리나라 8개 프로야구 구단의 2006년도 전력 자료

- 롯데와 LG는 작은 삼각형을 형성
- 삼성은 방어율은 좋지만 타율과 출루율에서 저조
- SK는 반대로 방어율이 좋지 않지만 타율과 출루율이 좋음



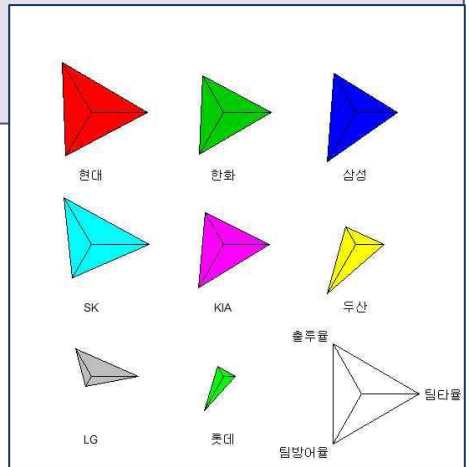

| 구단 | 팀타율 | 출루율 | 팀방어율 |
|---|---|---|---|
| 현대 | 0.264 | 0.347 | 3.26 |
| 한화 | 0.255 | 0.329 | 3.34 |
| 삼성 | 0.249 | 0.332 | 3.05 |
| SK | 0.265 | 0.343 | 3.63 |
| KIA | 0.253 | 0.323 | 3.29 |
| 두산 | 0.235 | 0.303 | 3.07 |
| LG | 0.254 | 0.317 | 4.47 |
| 롯데프로 | 0.218 | 0.293 | 3.63 |

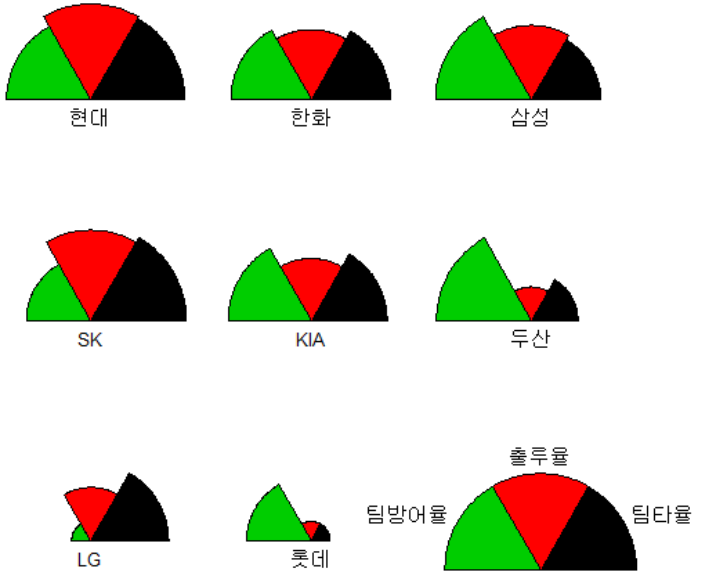
# 3. 파이조각그림

# 3 파이조각그림 (pie segments plot)

## 파이조각그림 (pie segments plot)

- 각 변수는 파이조각의 면적으로 나타내진다.
- 단순한 그림은 다변수 특성을 갖는 개체들의 특성을 표현해낼 수 있지만 깊이 있는 분석을 하기는 어렵다.



# 4. 산점도 행렬

# 4 산점도 행렬

## 산점도 행렬 (scatterplot matrix)

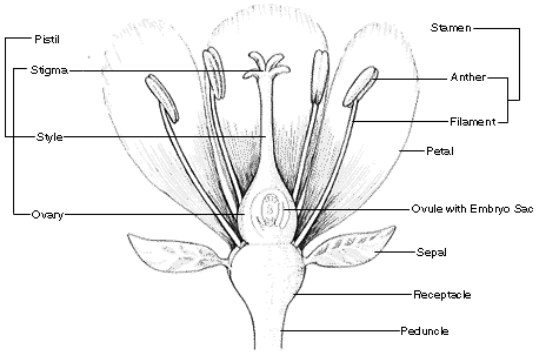
- 2개의 변수 쌍으로 작성된 산점도를 변수의 개수 만큼 P×P행렬에 체계적으로 배열해놓은 그래프
- 다변량 자료의 변수간 관련성을 눈으로 확인할 수 있는 그림으로 표현하는 것이 가능
- p*(p–1)/2개의 개별 산점도를 포함

- 자료의 구조 및 특징을 파악을 위하여 효과적이고 신뢰성 있는 자료의 요약과 그래프 기법의 활용

# 4 산점도 행렬

## 산점도 예제 Fisher의 붓꽃자료 (iris data)

- 4개 변수; 꽃받침 길이(Sepal.Length), 꽃받침의 폭(Sepal.Width), 꽃잎의 길이(Petal Length), 꽃잎 폭(Petal.Width)로 ahems 2개 변수간의 관계를 보여준다.

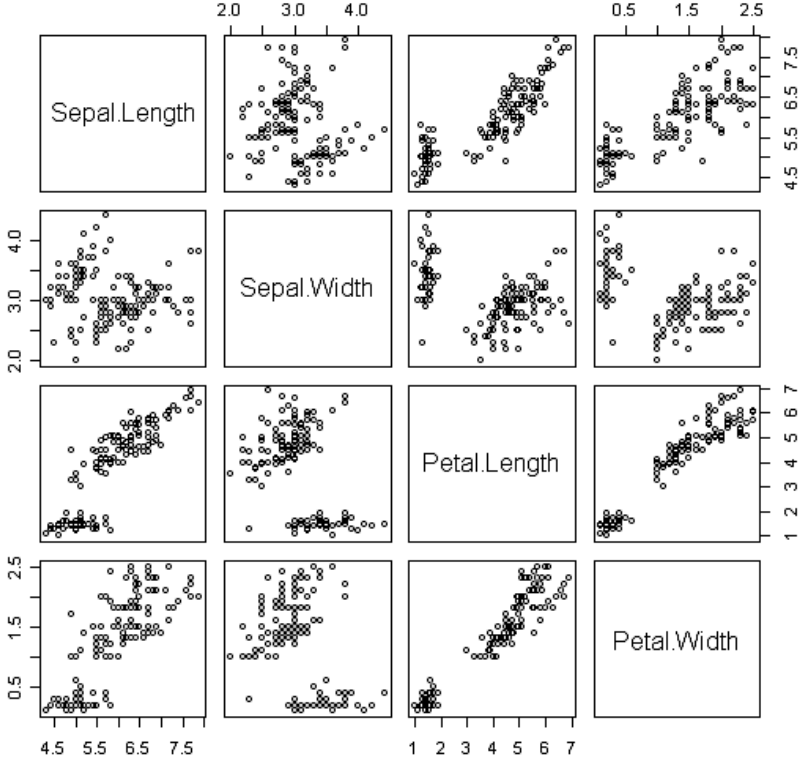



```
data(iris) attach(iris) pairs(iris[1:4],
main="Iris Data")
```

# 4 산점도 행렬

## 산점도 예제 Fisher의 붓꽃자료 (iris data)

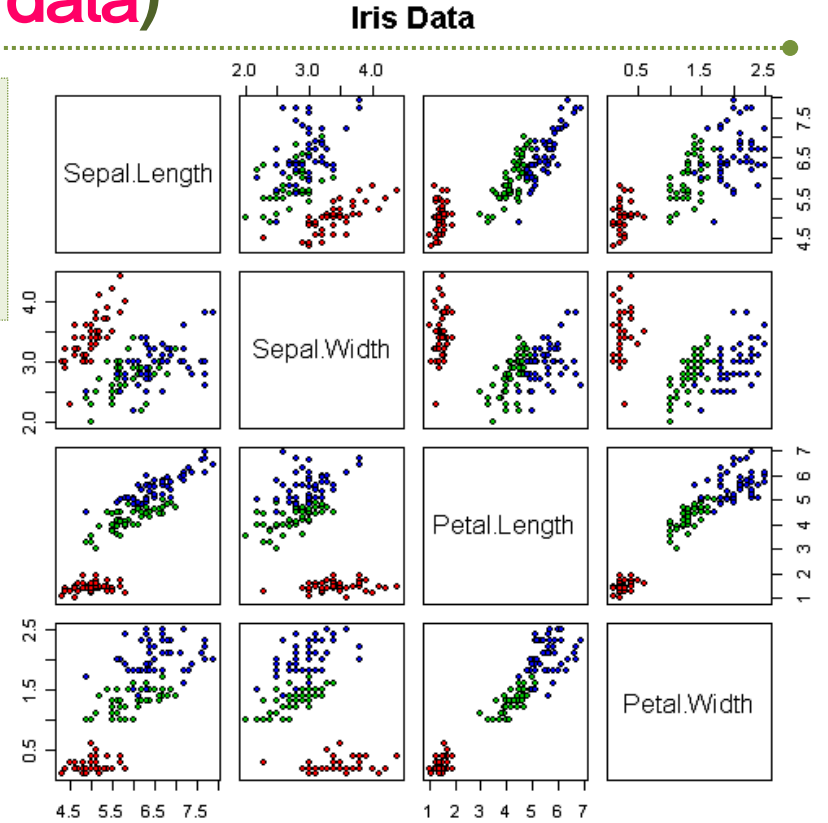
- Sepal.Length와 Sepal.Width의 산점도에서 2개 정도의 군집이 있음
- Petal.Length와 Petal.Width에서도 2개 정도의 군집



# 4 산점도 행렬

## 품종별 색구분 산점도 행렬 (iris data)

```
pairs(iris[1:4], main = "Iris Data", pch
= 21, bg = c("red", "green3",
"blue")[unclass(iris$Species)])
```



# 5. 조건부 플롯

# 5 조건부 플롯 (conditioning plot)

- 3개 변수 X, Y와 Z의 관계를 탐색하는 과정에서 변수 X와 Y의 관계를 Z에 조건화하여 보는 그래프

- 3개 변수 X, Y와 Z의 관계를 탐색하는 과정에서 변수 X와 Y의 관계를 Z에 조건화하여 볼 수 있다.

- 예 (Z=0이면 고교 재학생, Z=1이면 재수생, Z=2이면 삼수생 이상). 이런 경우 수능 언어와 수능 수리의 관계를 탐색하는 과정에서 재수기간(Z) 별로 X와 Y의 산점도

# 5 조건부 플롯 (conditioning plot)

## Z가 연속형인 경우

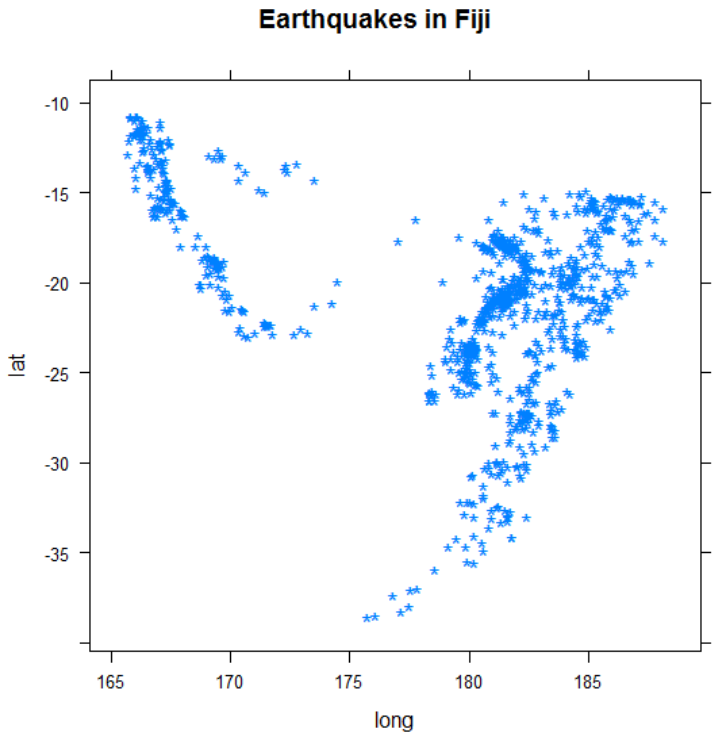
Z를 몇 개의 구간으로 나누어 범주화

equal.count(z, number=*k1*, overlap=*k2*)

- z는 조건화 변수이고 *k1*은 구간 수이고 *k2*는 0과 1사이의 수로 구간이 겹치는 정도를 제어하는 파라미터이다.
- 이에 따른 결과로는 z 값에 따라 관측개체를 *k1*개의 구간 중 하나(또는 2개 이상)에 배속

# 5 조건부 플롯 (conditioning plot)

## 예제 피지 섬 지진의 경도(long)와 위도(lat)

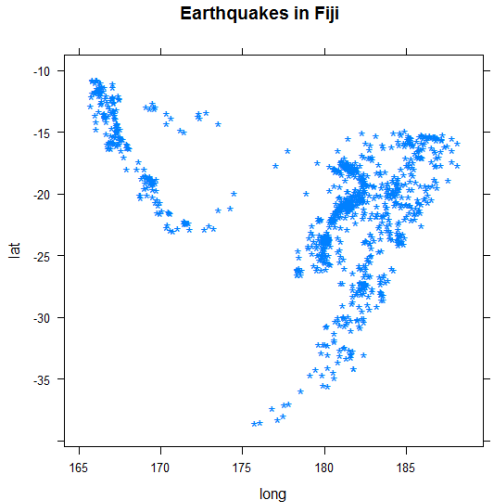


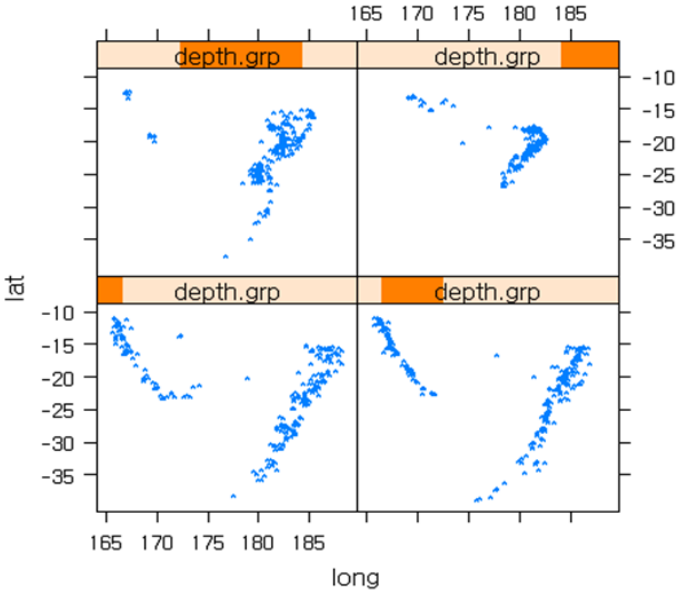

```
library(lattice) data(quakes) str(quakes)
attach(quakes) xyplot(lat ~ long,
main="Earthquakes in Fiji", pch="*",
cex=2)
```

# 5 조건부 플롯 (conditioning plot)

## 예제 피지 섬 지진의 경도(long)와 위도(lat)

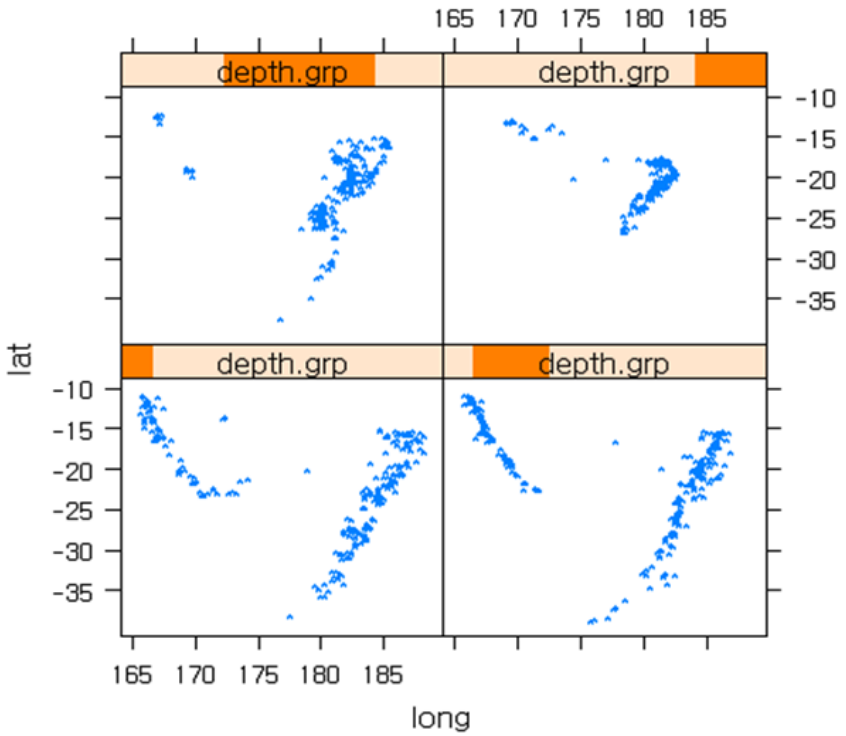
```
library(lattice) data(quakes) str(quakes) attach(quakes) xyplot(lat ~ long, main="Earthquakes in Fiji", pch="*", cex=2)
```



# 5 조건부 플롯 (conditioning plot)

## 예제 피지 섬 지진의 경도(long)와 위도(lat)




- depth.grp=1,2,3,4에 해당하는 long과 lat의 산점도가 왼쪽 아래, 오른쪽 아래, 왼쪽 위, 오른쪽 위의 순서로 놓여 있다.
- 4개의 산점도 중 아래 2개와 위 2개는 서로 달라 보인다. 즉 피지 섬 바깥쪽 낮은 곳과 안쪽 깊은 곳에 단층이 있다는 추론이 가능

다음시간안내

# 다변량 자료의 시각화 이해 2