7강 R 통계 그래픽스(1)

이화여대 통계학과 이은경교수

목 차

1 lattice의 소개

lattice의 소개

- R에서 Trellis 그림을 그릴 수 있도록 제공하고 있는 패키지
- Trellis
  - 빌 글리브랜드(Bill Cleveland)가 제안한 그래픽의 '디자인 원칙'을 구현
  - 자료의 정보들을 좀 더 정확하고 충실하게 전달할 수 있도록 함
  - R의 기본 그래픽과는 달리 멀티패널조건(multipanel conditioning)을 제공하고 있어 자료를 쉽게 범주형 변수의 범주에 따라 나누어 그림을 그릴 수 있음

2 자료 소개

autompg 자료

- 398대의 자동차 연비에 관한 자료
- UCI machine learning repository에 저장
- 변수 설명
  - mpg : 연비
  - cylinder : 실린더수
  - displacement : 배기량
  - horsepower : 마력
  - weight : 무게
  - acceleration : 가속능력
  - year : 연도
  - origin : 만들어진 곳
  - name : 차종

autompg 자료

```
> setwd("자료가 있는 디렉토리")
> autompg<-read.csv("auto-mpg.csv",header=TRUE,na.string=".")
> dim(autompg)
[1] 398   9
> head(autompg)
  mpg cylinder displacement horsepower weight acceleration year origin                  name
1  18        8          307        130   3504         12.0   70      1 chevrolet chevelle malibu
2  15        8          350        165   3693         11.5   70      1 buick skylark 320
3  18        8          318        150   3436         11.0   70      1 plymouth satellite
4  16        8          304        150   3433         12.0   70      1 amc rebel sst
5  17        8          302        140   3449         10.5   70      1 ford torino
6  15        8          429        198   4341         10.0   70      1 ford galaxie 500
```

tipping 자료

- 레스토랑 고객들의 팁에 대한 습성을 알아보기 위하여 미국 뉴욕 근교에서 수집된 자료
- 변수 설명
  - obs : 관측번호
  - tip : 팁
  - smoker : 흡연석/금연석
  - time : 점심/저녁
  - totbill : 전체 가격
  - sex : 계산한 사람의 성별
  - day : 요일
  - size : 일행 수
  - tiprate = tip/totbill * 100 : 전체가격에 대한 팁의 비율

tipping 자료

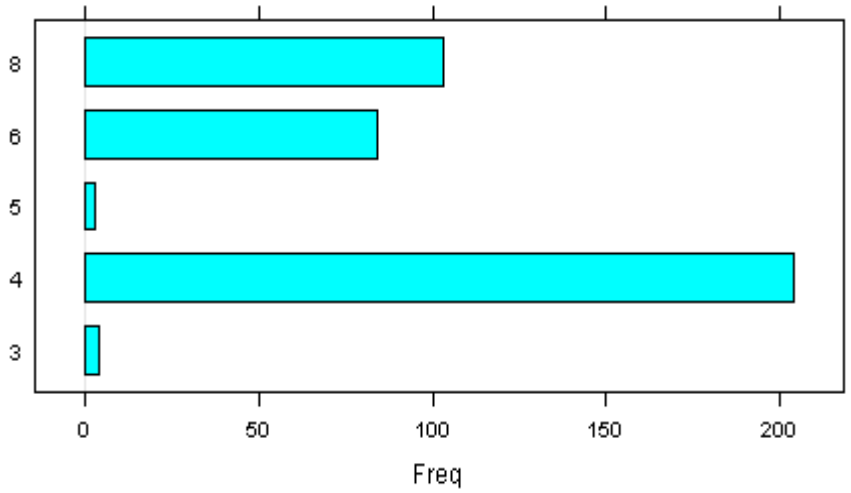
```
> tipping <- read.csv("tips.csv", header=TRUE)
> tipping$tiprate <- tipping$tip/tipping$totbill * 100
> dim(tipping)
[1] 244   9
> head(tipping)
  obs totbill  tip    sex        smoker        day    time    size   tiprate
1   1   16.99 1.01 Female   Non-smoker    Sun   Dinner     2      5.94
2   2   10.34 1.66   Male   Non-smoker    Sun   Dinner     3     16.05
3   3   21.01 3.50   Male   Non-smoker    Sun   Dinner     3     16.66
4   4   23.68 3.31   Male   Non-smoker    Sun   Dinner     2     13.98
5   5   24.59 3.61 Female   Non-smoker    Sun   Dinner     4     14.68
6   6   25.29 4.71   Male   Non-smoker    Sun   Dinner     4     18.62
```

3 lattice를 이용한 단변량 그래프

막대그래프 (1)

- 범주형 변수의 각 범주에 대한 도수를 막대그림으로 나타낸 것
- barchart 함수를 이용
- 문자형 또는 factor 변수를 이용
- 기본형은 가로형 막대그림

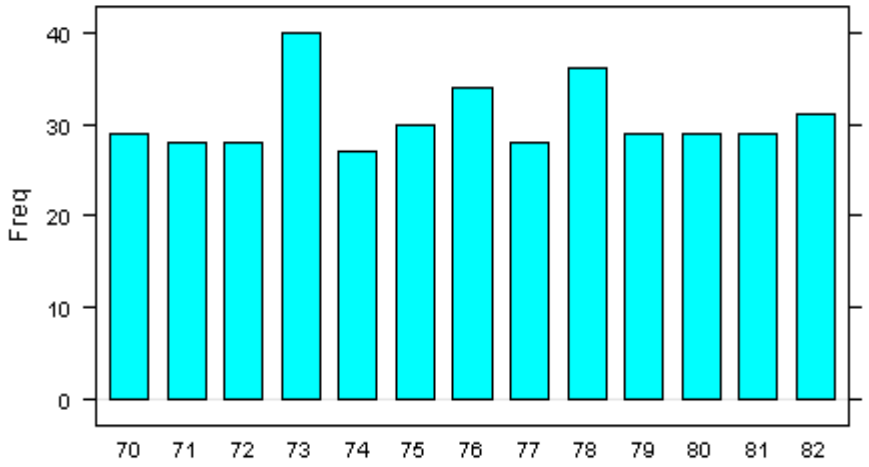
```
> library(lattice) # call lattice library
> barchart(as.factor(autompg$cylinder))
```



막대그래프 (2)

- 세로형 막대그림을 위해서는 horizontal = FALSE 옵션을 이용

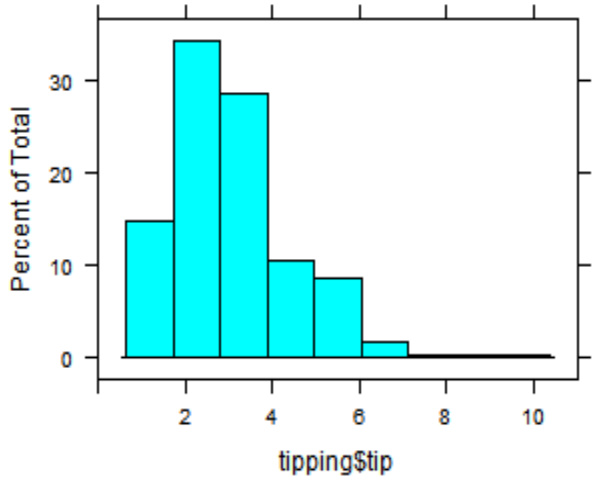
```
> barchart(as.factor(autompg$year),horizontal=FALSE)
```



히스토그램 (1)

- 연속자료의 분포를 쉽게 파악하기 위하여 그리는 그림
- histogram 함수를 이용
- 숫자변수를 이용

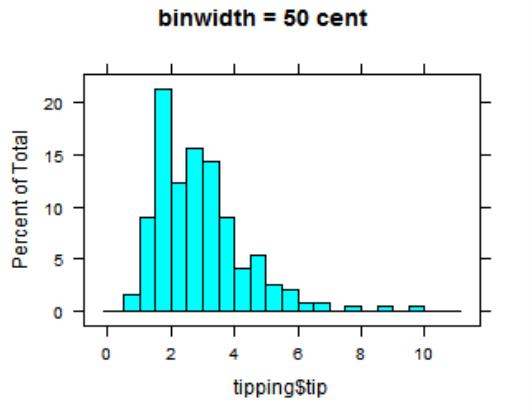
```
> histogram(tipping$tip)
```

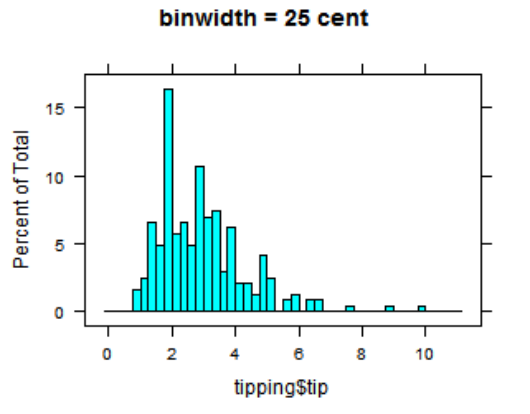


히스토그램 (2)

- breaks 옵션을 이용하여 binwidth 설정
- main 옵션을 이용하여 그림의 title 지정

```
> histogram(tipping$tip, breaks = seq(0,11,0.5), main = "binwidth = 50 cent")
> histogram(tipping$tip, breaks = seq(0,11,0.25), main = "binwidth = 25 cent")
```

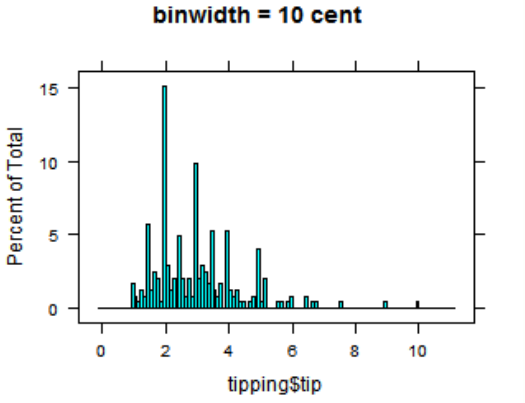


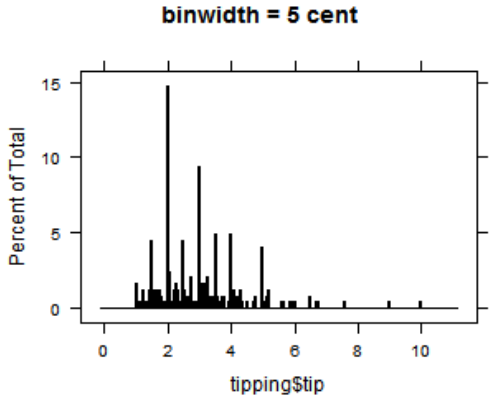


히스토그램 (3)

- binwidth를 달리하면 다른 패턴을 발견할 수 있음

```
> histogram(tipping$tip, breaks = seq(0,11,0.1), main = "binwidth = 10 cent")
> histogram(tipping$tip, breaks = seq(0,11,0.05), main = "binwidth = 5 cent")
```

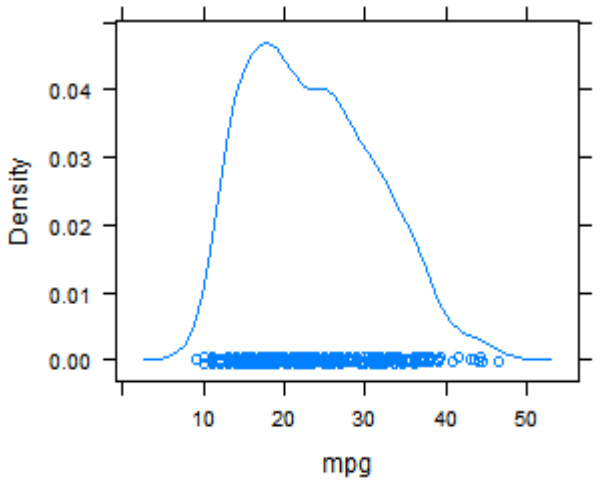




밀도그림

- 연속자료의 분포를 파악하기 위한 그림
- density 함수 이용
- xlab 옵션을 이용하여 X축 레이블 지정

```
> densityplot(autompg$mpg, xlab="mpg")
```

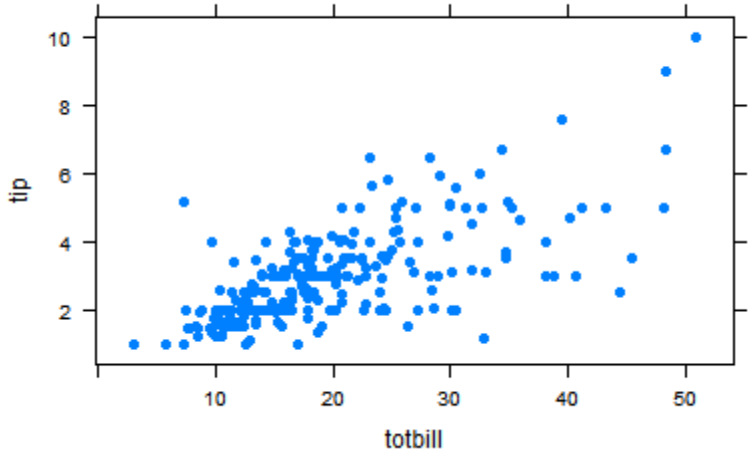


4 lattice를 이용한 이변량 그래프

연속변수 vs. 연속변수

- 산점도(scatter plot) : 두 연속변수의 관계를 알아보기 위한 그림
- xyplot 함수를 이용
- ‘Y축 변수 ~ X축 변수’의 식 형태로 나타냄

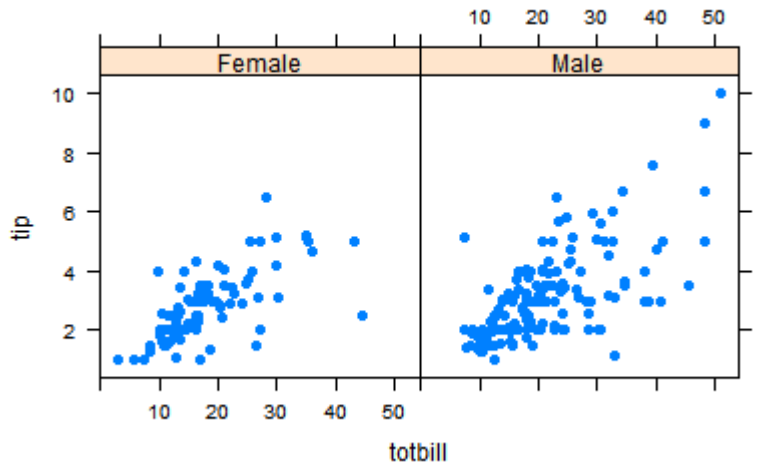
```
> xyplot(tip ~ totbill, pch=16, data = tipping)
```



범주별 산점도 (1)

- ‘Y축 변수 ~ X축 변수 | 범주형 변수’의 식으로 범주형 변수별 산점도를 그릴 수 있다

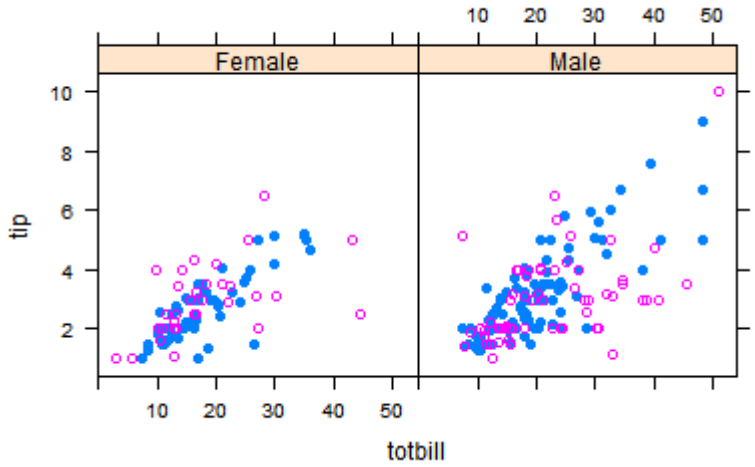
```
> xyplot(tip~totbill | sex, pch=16, data = tipping)
```



범주별 산점도 (2)

- group 옵션을 이용하여 제 3의 범주형 변수에 따라 점의 모양/색을 달리할 수 있다.

```
> xyplot(tip~totbill | sex, group = smoker, pch = c(16,1), data = tipping)
```



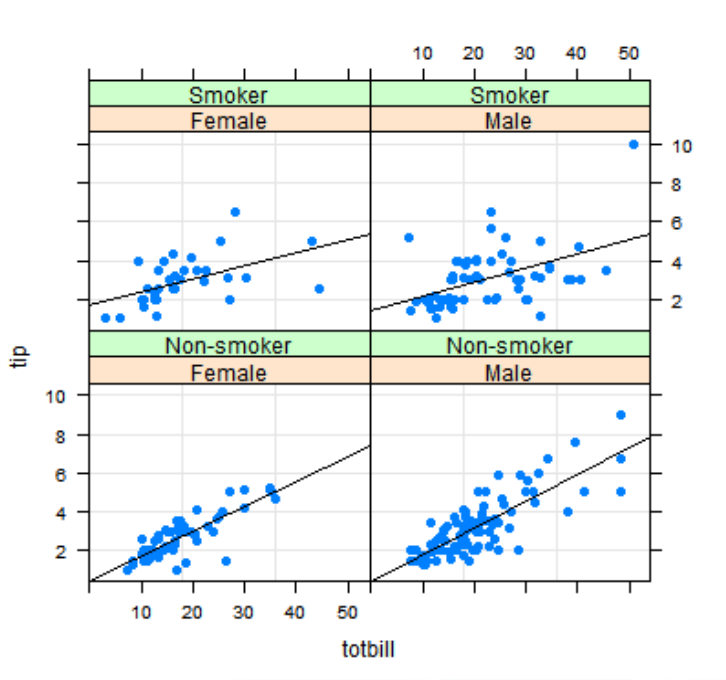
panel을 이용한 그림 (1)

- panel 함수 내에 사용자가 각 패널의 그림을 다양한 형태로 정의 가능
- panel.grid 함수를 이용하여 각 패널에 눈금선을 삽입
  - ‘h = –1’ : 수평으로 각 눈금마다 눈금선을 넣기
  - ‘v = 2’ : 수직으로는 2개의 눈금선을 넣기
- panel.xyplot 함수를 이용하여 각 패널에 xyplot을 지정
  - ‘pch = 16’ : 16번 점모양(●)을 이용하여 산점도 그리기
- panel.lmline 함수를 이용하여 각 패널에 회귀직선 삽입

```
> xyplot(tip ~ totbill | sex + smoker,
+        data = tipping,
+        panel = function(x, y) {
+            panel.grid(h = -1, v = 2)
+            panel.xyplot(x, y, pch=16)
+            panel.lmline(x, y)
+        })
```

panel을 이용한 그림 (2)

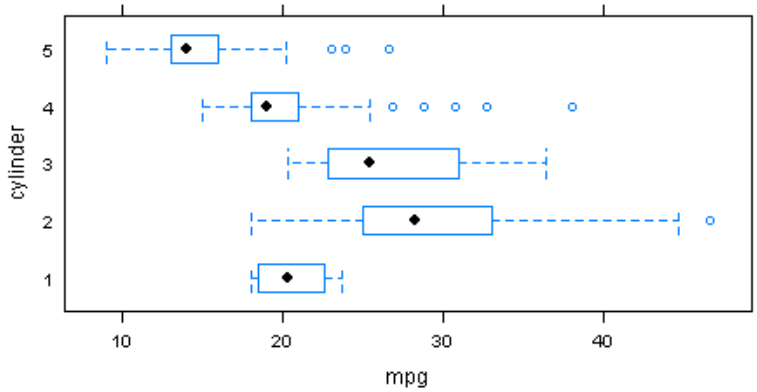
```
> xyplot(tip ~ totbill | sex + smoker,
+        data = tipping,
+        panel = function(x, y) {
+            panel.grid(h = -1, v = 2)
+            panel.xyplot(x, y, pch=16)
+            panel.lmline(x, y)
+        })
```



연속변수 vs. 범주형 변수 (1)

- 평행상자그림 : 각 범주별로 연속변수의 분포를 비교하기 위한 그림
  - 예 : 일원분산분석(연속변수의 그룹간 차이를 보기 위한 것)
- 범주형 변수의 각 범주별로 연속변수의 상자그림을 그려 비교
- bwplot 함수를 이용
- 기본형은 가로형태의 상자그림

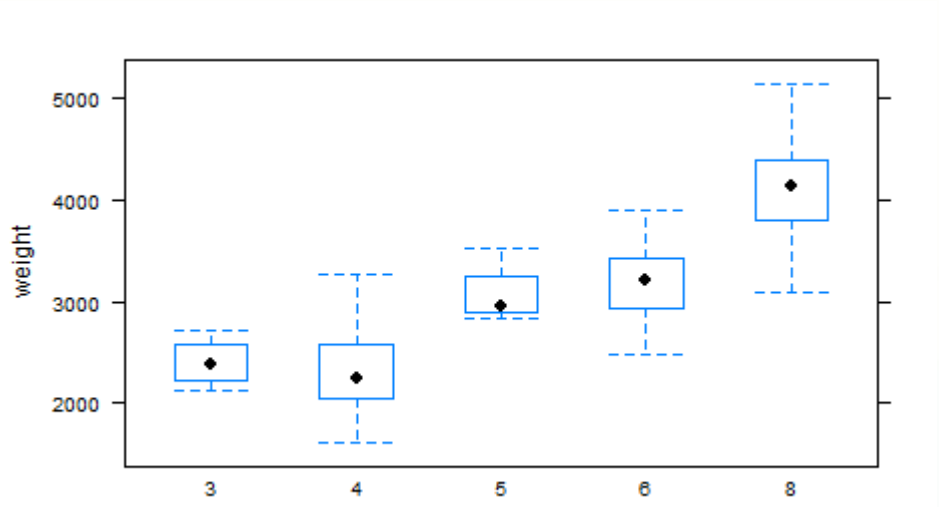
```
> bwplot(cylinder ~ mpg, data = autompg)
```



연속변수 vs. 범주형 변수 (2)

- 세로형태의 상자그림을 위하여 horizontal = FALSE 옵션을 이용

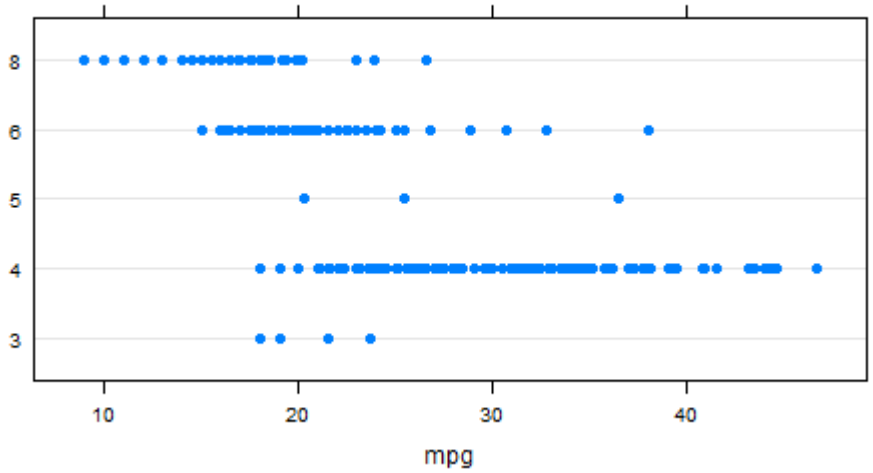
```
> bwplot(weight ~ as.factor(cylinder), data = autompg, horizontal  = FALSE)
```



연속변수 vs. 범주형 변수 (3)

- 평행점그림 : 점그림(dotplot)을 평행하게 그려 비교
- dotplot 함수를 이용

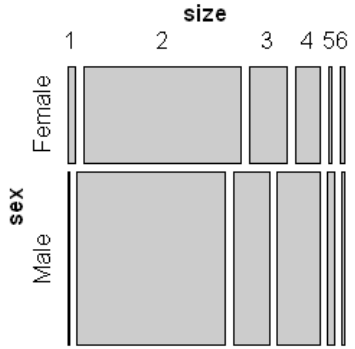
```
> dotplot(as.factor(cylinder) ~ mpg, data = autompg)
```



범주형 변수 vs. 범주형 변수 (1)

- mosaic 그림 : 이차원 분할표로 정리된 자료에서 두 범주형 변수 중 하나의 변수를 조건으로 할 때 나머지 한 변수의 분포를 나타내는 그림
- Trellis 형태의 mosaic 그림은 vcd 패키지에서 mosaic 함수로 제공
- ‘ ~ A + B’ 의 식에서는 A 변수를 조건으로 하여 A 변수의 각 범주 내에서 B 변수의 범주비율을 그림으로 나타냄

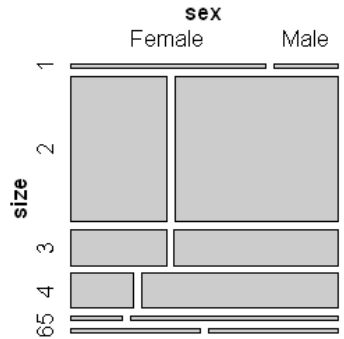
```
> library(vcd)
> mosaic(~ sex + size, data = tipping)
```



범주형 변수 vs. 범주형 변수 (2)

- 변수의 순서를 달리하면 다른 그림이 됨

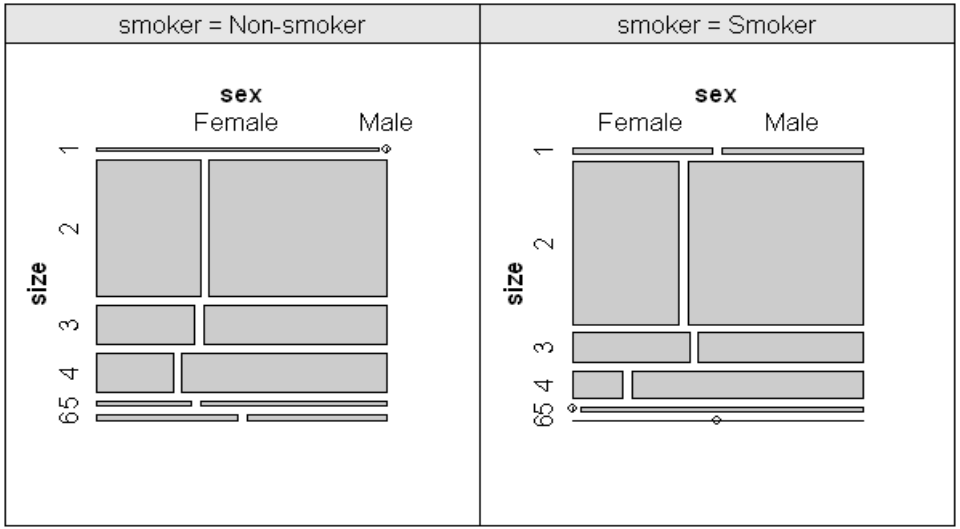
```
> mosaic(~ size + sex, data = tipping)
```



범주형 변수 vs. 범주형 변수 (3)

- 제 3의 범주형 변수 별 mosaic 그림을 위해서는 cotabplot 함수를 이용

```
> cotabplot(~ size+sex | smoker, data = tipping, panel = cotab_mosaic)
```

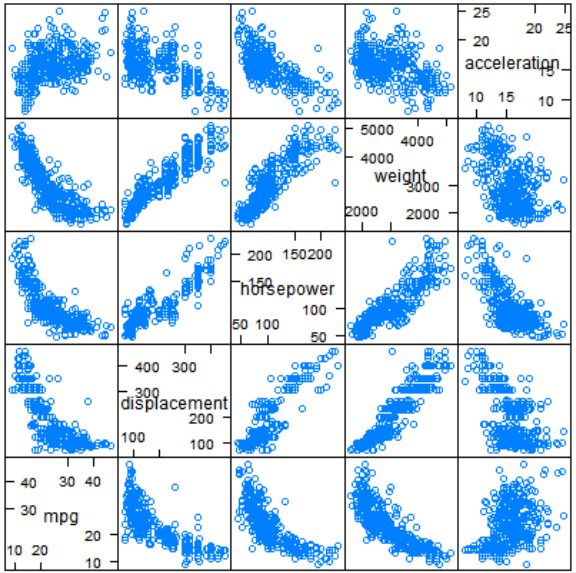


5 lattice를 이용한 다변량 그래프

산점도 행렬 (1)

- 여러 개의 연속변수를 동시에 살펴보기 위한 그림
- 두 연속변수씩 짝지어 그린 산점도를 행렬 형태로 나타낸 것
- splom 함수를 이용하여 그림

```
> splom(~autompg[,c(1,3:6)], data = autompg)
```

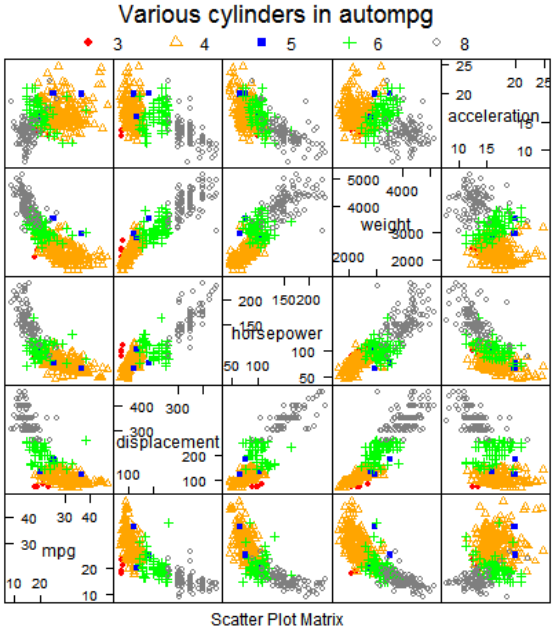


Scatter Plot Matrix

산점도 행렬 (2)

- group, pch, col 등의 옵션을 이용하여 점의 모양, 색깔을 그룹별로 지정
- key 옵션을 이용하여 범례의 내용을 명시

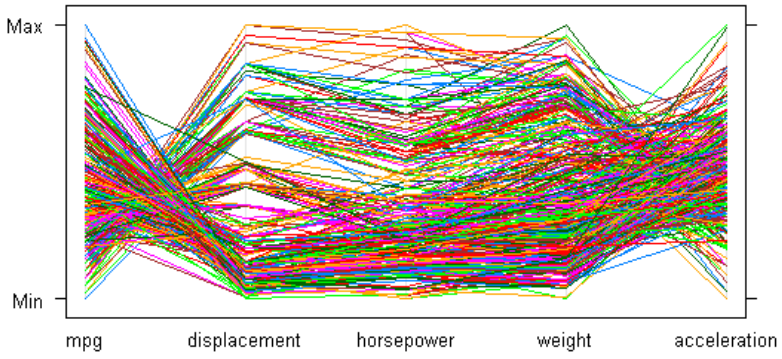
```
> splom(~autompg[,c(1,3:6)], groups = cylinder, data = autompg,
+       col=c("red","orange","blue","green","grey50"),
+       pch=c(16,2,15,3,1),cex=0.7,
+       key = list(title = "Various cylinders in autompg",
+                  columns = 5,
+                  points = list(pch =c(16,2,15,3,1),
+                  col = c("red","orange","blue","green","grey50")),
+                  text = list(c("3","4","5","6","8"))))
```



평행좌표그림 (1)

- 연속변수의 수가 많아 산점도 행렬로는 파악이 힘든 경우 이용
- parallelplot 함수를 이용
- 기본은 수평그림, 수직그림으로 바꾸기 위해서는 horizontal = FALSE 옵션을 이용

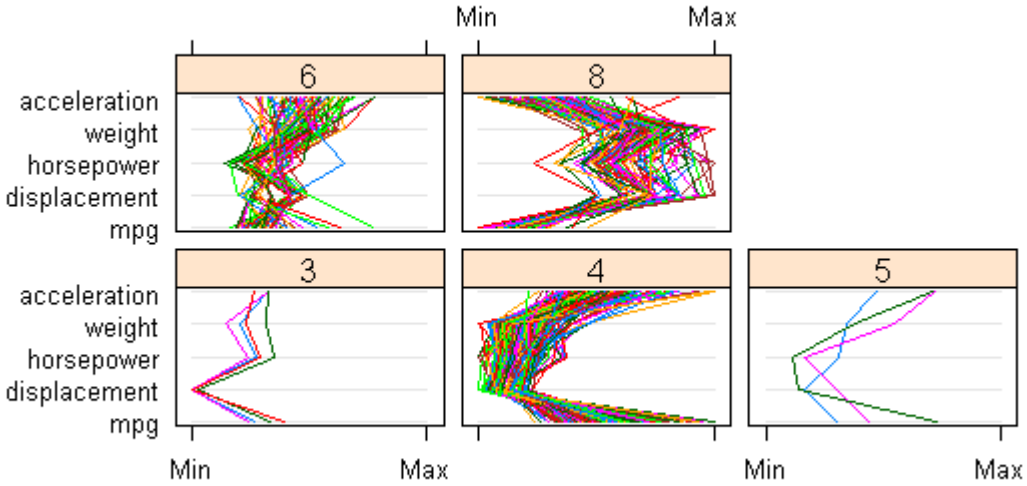
```
> parallelplot(~ autompg[,c(1,3:6)] ,data = autompg, horizontal=FALSE)
```



평행좌표그림 (2)

- 조건식을 이용하여 여러 그룹별로 평행좌표그림을 따로 그려 비교

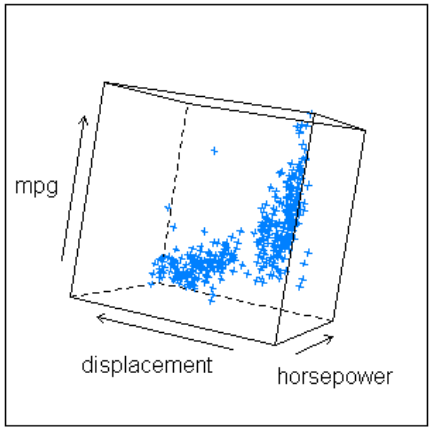
```
> parallelplot(~ autompg[,c(1,3:6)] | as.factor(cylinder), data = autompg)
```



3차원 산점도 그림

- 연속변수가 3개인 경우 3차원 상에서의 자료분포를 알아보기 위한 그림
- cloud 함수를 이용, ‘Z ~ X * Y’ 형태의 식으로 이용

```
> cloud(mpg ~ horsepower*displacement, data = autompg,
+                                    screen=list(x=-80,y=70))
```



다음시간 안내

R 통계 그래픽스 (2)