R 데이터 분석 입문

7주 차

데이터 시각화

오 세 종

DANKOOK UNIVERSITY

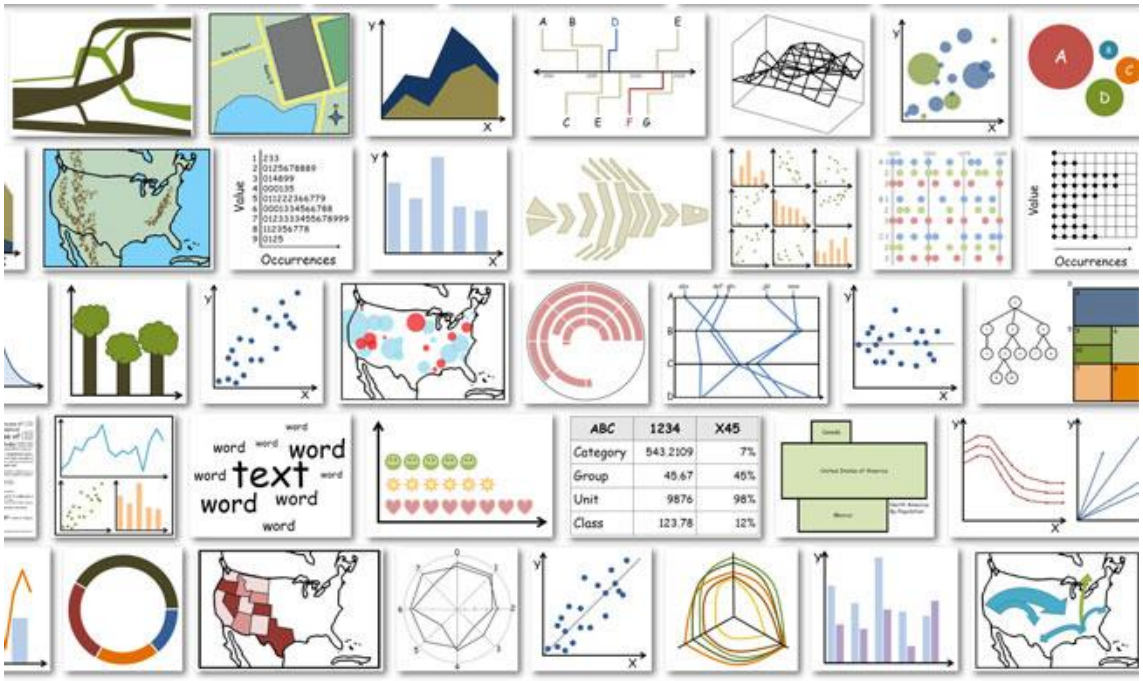
Contents

- 1. 나무지도
- 2. 버블차트
- 3. 다중상자그림
- 4. 모자이크 플롯
- 5. ggplot

본 강의 자료는 다음 블로그 '데이터과학의 둘레' **(http://blog.daum.net/huh420/19)** 및 '데이터 시각화 (허명회著, 자유아카데미)' 의 자료를 참고로 작성되었음

개요

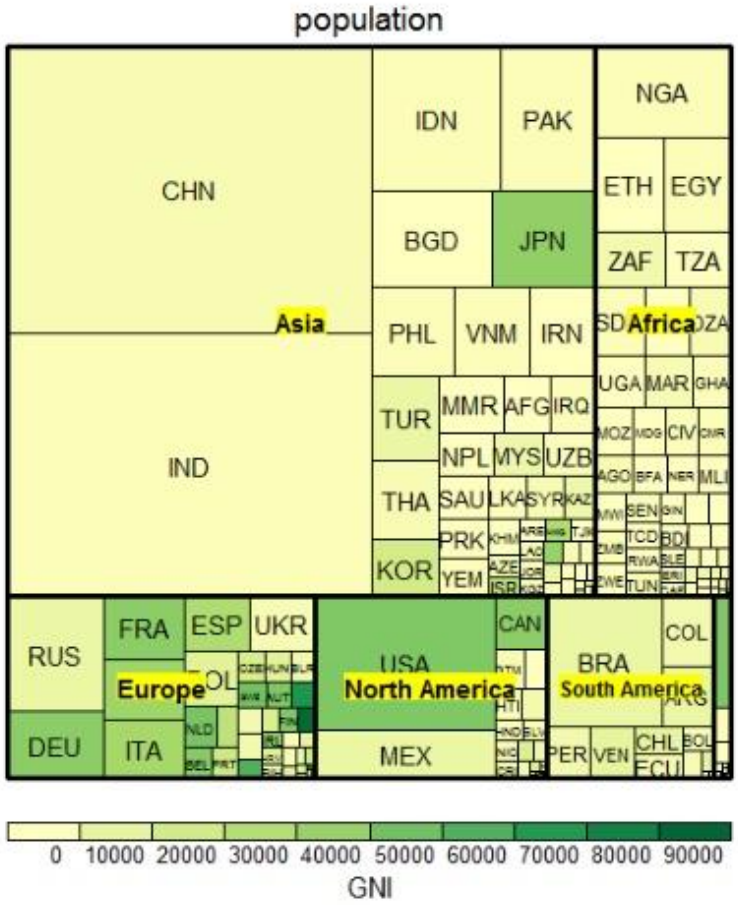
- 데이터 시각화의 중요성
  - 분석 결과를 숫자, 문자, 표로만 제공하면 이해가 어렵고, 데이터가 담고 있는 의미를 발견하기 어렵다.
  - 인간은 시각을 통한 정보 습득에 익숙함
  - 분석, 결과, 데이터를 그래프, 그림과 같은 시각적 도구를 사용하여 표현하면 정보 전달이 용이해지고, 영감을 얻게 한다.



https://www.seas.harvard.edu/news/2013/10/what-makes-data-visualization-memorable

1. 나무지도(tree map)

- 나무지도는 데이터가 갖는 계층구조를 타일 모양으로 표현한 것
- 타일은 계층적 속성을 가지며, 계층은 컬러로 표현된다



참조 : http://blog.daum.net/huh420/19

1. 나무지도(tree map)

- 설치가 필요한 패키지
  - treemap

- 실습에 사용할 데이터셋
  - GNI2014 (treemap)
  - 208개 국가의 1인당 총소득(gross national income) 데이터
  - 국가는 대륙(continent)으로 그룹핑되고 국가명은 국제표준(iso3)으로 지칭된다.
  - 국가정보는 population(인구)과 GNI(1인당 국민소득)이다

```
> head(GNI2014)
  iso3           country     continent population    GNI
3  BMU           Bermuda North America      67837 106140
4  NOR            Norway        Europe    4676305 103630
5  QAT             Qatar          Asia     833285  92200
6  CHE       Switzerland        Europe    7604467  88120
7  MAC Macao SAR, China          Asia     559846  76270
8  LUX        Luxembourg        Europe     491775  75990
```

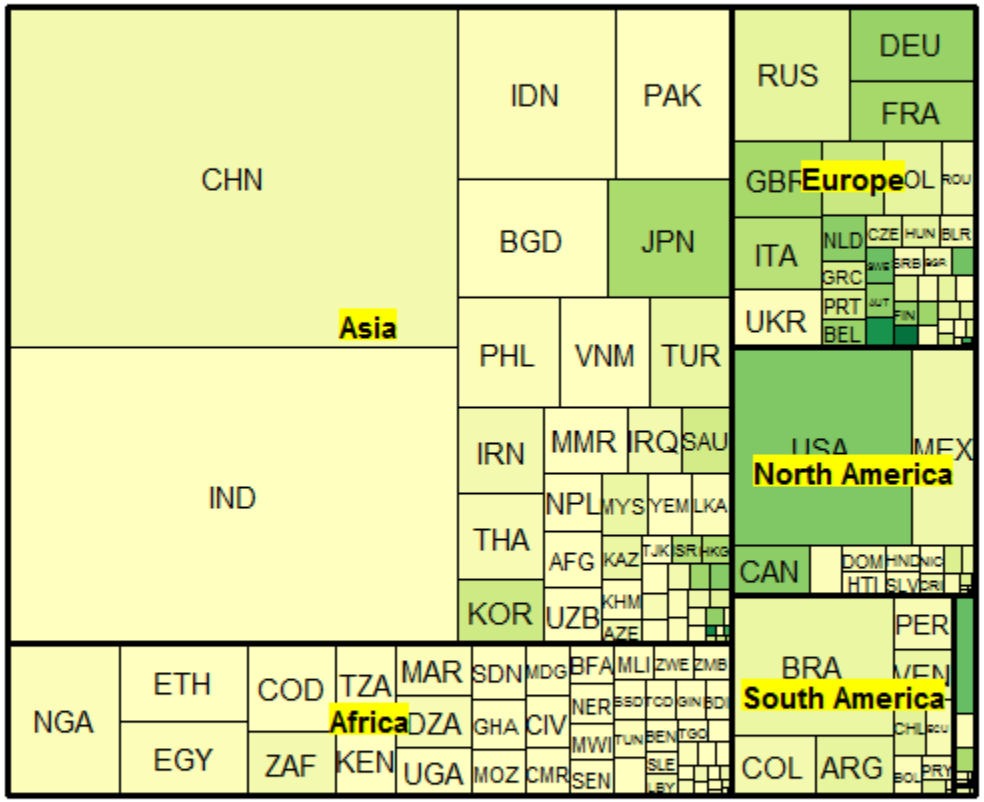
1. 나무지도(tree map)

```
library(treemap)
data(GNI2014)                          # 데이터 불러오기
str(GNI2014)                           # 데이터 내용보기
treemap(GNI2014,
        index=c("continent","iso3"),
        vSize="population",   # 타일의 크기
        vColor="GNI",         # 타일의 컬러
        type="value",         # 타일 컬러링 방법
        bg.labels="yellow")   # 레이블의 배경색
```

index=c("continent","iso3")
: 개체의 단위를 지정하는데 계층적 구조를 갖는 경우 상위 층을 먼저 넣는다. 대륙을 먼저 표현하고 그 안에 국가를 넣으라는 의미

type="value"
: **vColor** 에서 지정한 값에 의해서 타일의 컬러가 결정됨

population

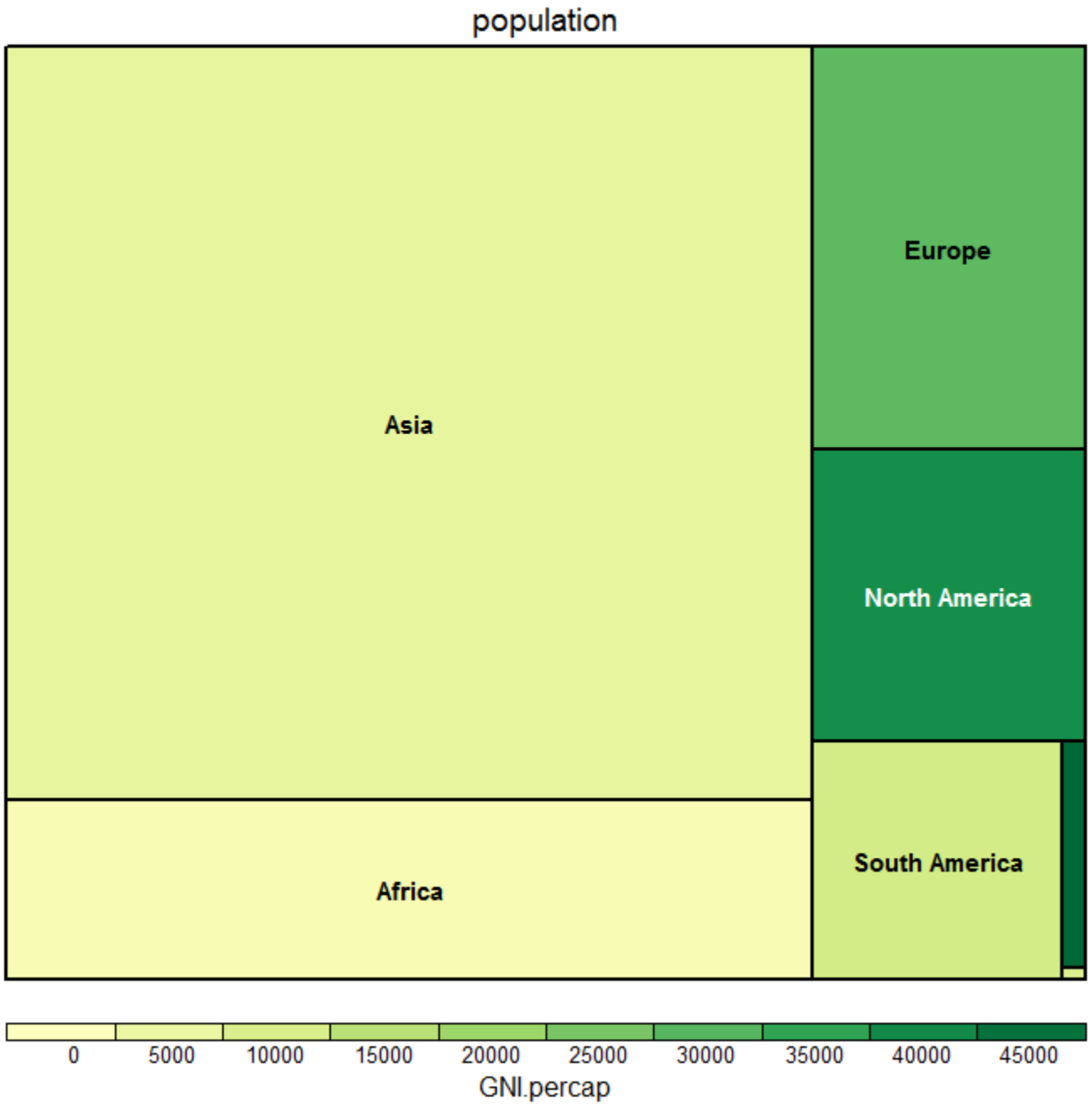


타일의 크기 : 국토면적
타일의 색 : 국민소득

1. 나무지도(tree map)

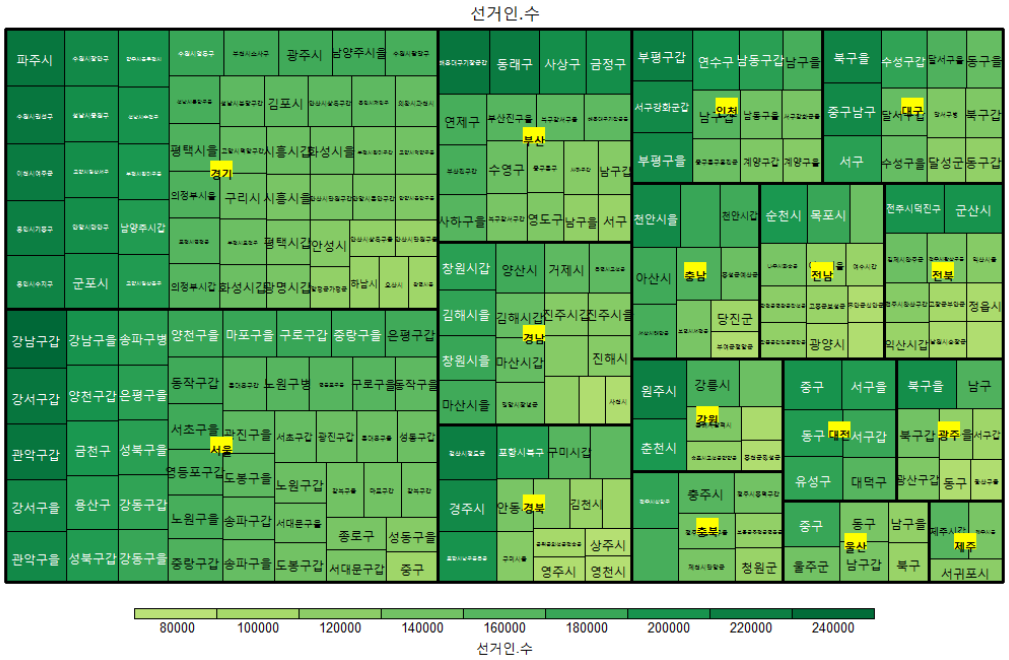
```
# 대륙별 인구, 소득
# 국가별 국민 총소득을 계산해서 GNI.total 컬럼에 저장
GNI2014$GNI.total <-
    GNI2014$population*GNI2014$GNI
head(GNI2014)
# 국가별 국민 총소득을 대륙별로 합계내서 GNI2014.a 에 저장
GNI2014.a <- aggregate(GNI2014[,4:6],
                by=list(GNI2014$continent),sum)
# 대륙별 합계를 대륙 인구수로 나누어 GNI.percap 컬럼에 저장
GNI2014.a$GNI.percap <-
    GNI2014.a$GNI.total/GNI2014.a$population

treemap(GNI2014.a,
        index=c("Group.1"),
        vSize="population",
        vColor="GNI.percap",
        type="value",
        bg.labels="yellow")
```


aggregate(): 배열에 저장된 값을 기준 컬럼에 의해 집계값를 계산한다


population
Asia
Europe
North America
South America
Africa
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
GNI.percap

[연습문제 1]

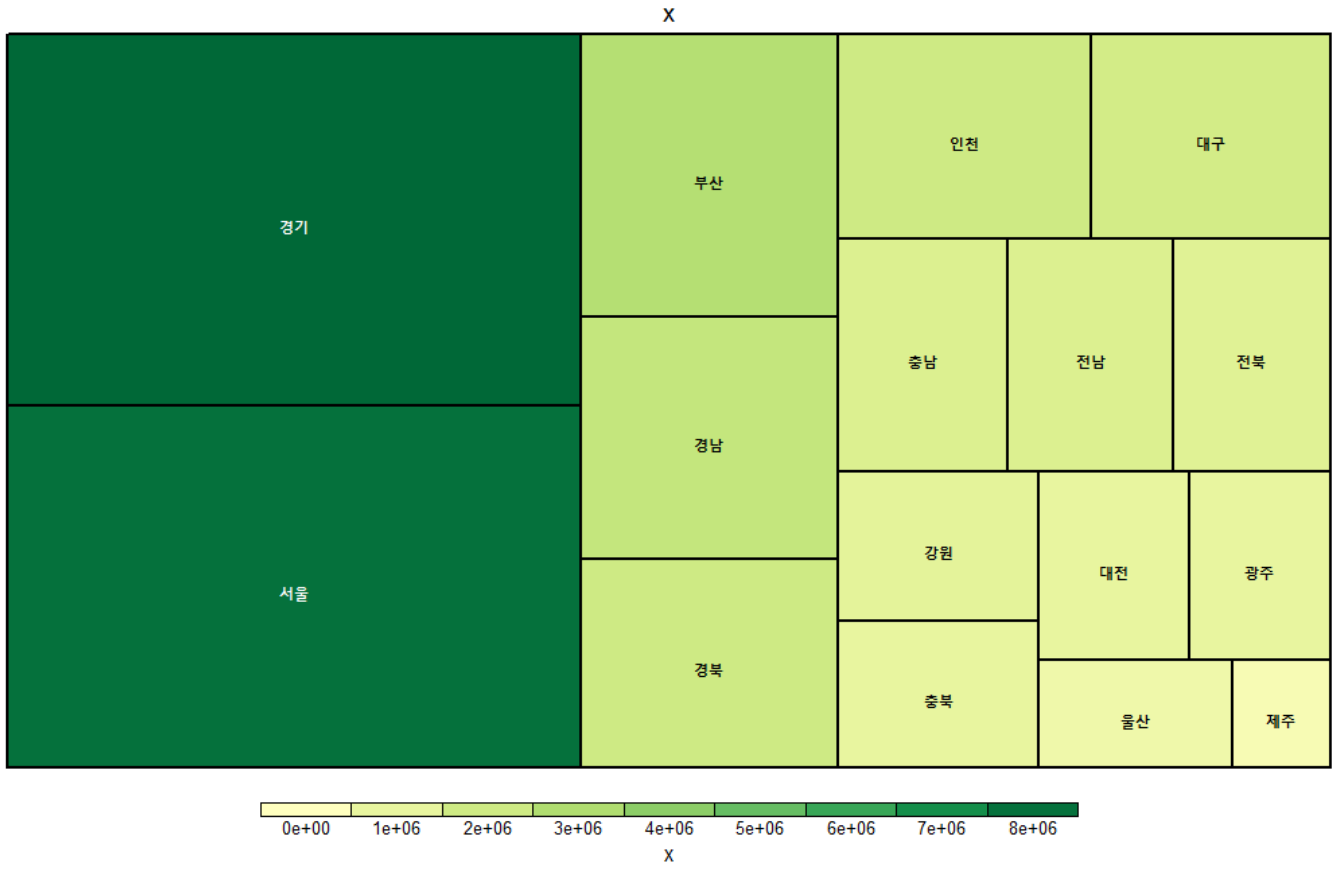
- 1. “국회의원_선거구_유권자수.csv” 파일의 내용을 가지고 다음과 같은 treemap 을 작성하시오



- 타일 하나는 각 선거구를 의미
- 굵은 검은띠 블록은 선거구가 속한 시도를 의미
- 타일의 면적, 색깔은 선거인수를 의미

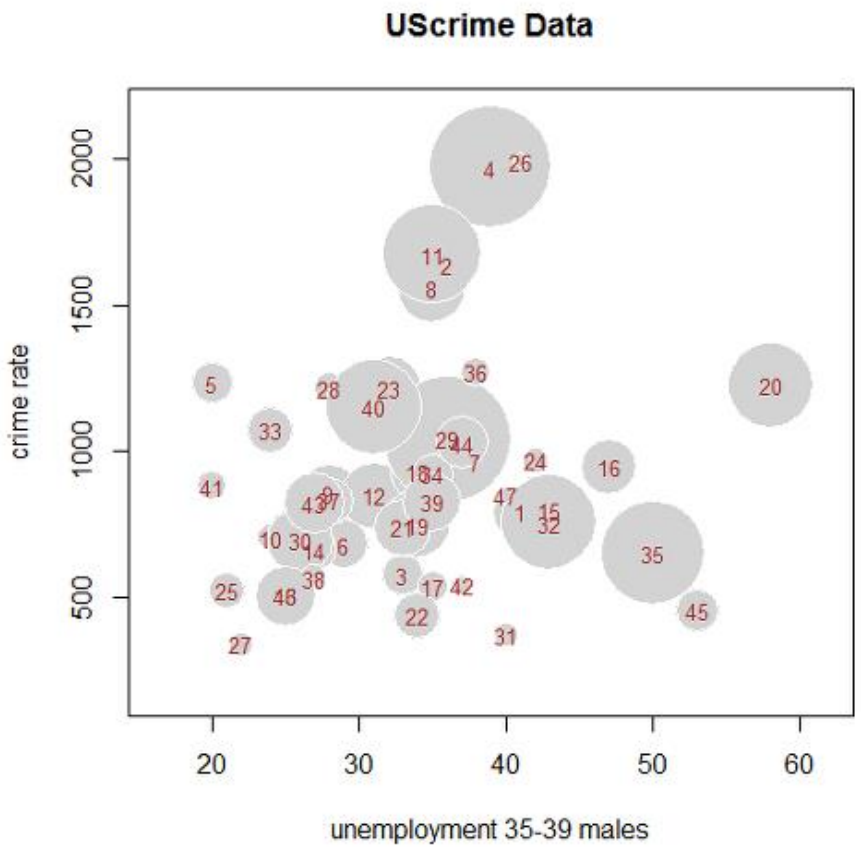
[연습문제 1]

- 2. “국회의원_선거구_유권자수.csv” 파일의 내용으로 부터 시도별 선거인수를 집계(합계계산)하여 다음과 같은 treemap 을 작성하시오



2. 버블 차트 (bubble chart)

- 산점도는 두개의 변수간 상관 관계를 표시한다.
- 버블 차트는 산점도에 제3의 변수를 크기에 비례하는 버블(원)으로 표현한 그림이다.



실업률(남자 35-39세) x와 범죄율 y 간 관계를 보여주는 버블차트 (원의 넓이는 인구수) http://blog.daum.net/huh420/19

2. 버블 차트 (bubble chart)

- 설치가 필요한 패키지
  - MASS

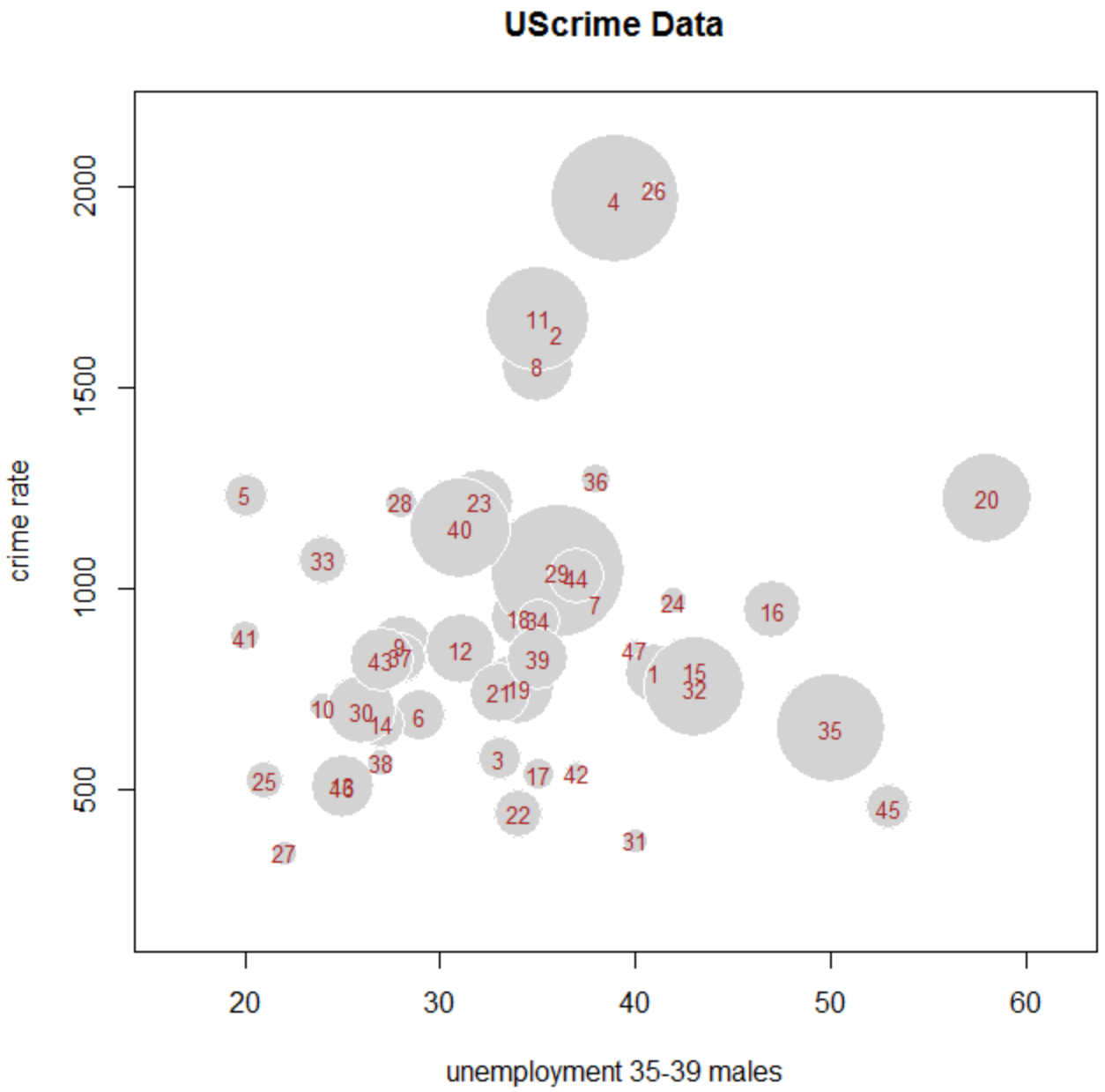
- 실습에 사용할 데이터셋
  - UScrime (MASS)

```
> head(UScrime)
    M So  Ed Po1 Po2  LF  M.F Pop  NW  U1 U2 GDP Ineq     Prob     Time    y
1 151  1  91  58  56 510  950  33 301 108 41 394  261 0.084602 26.2011  791
2 143  0 113 103  95 583 1012  13 102  96 36 557  194 0.029599 25.2999 1635
3 142  1  89  45  44 533  969  18 219  94 33 318  250 0.083401 24.3006  578
4 136  0 121 149 141 577  994 157  80 102 39 673  167 0.015801 29.9012 1969
5 141  0 121 109 101 591  985  18  30  91 20 578  174 0.041399 21.2998 1234
6 121  0 110 118 115 547  964  25  44  84 29 689  126 0.034201 20.9995  682
```

- Pop : 인구수
- U2 : 실업률(35~39세)
- y : 범죄율

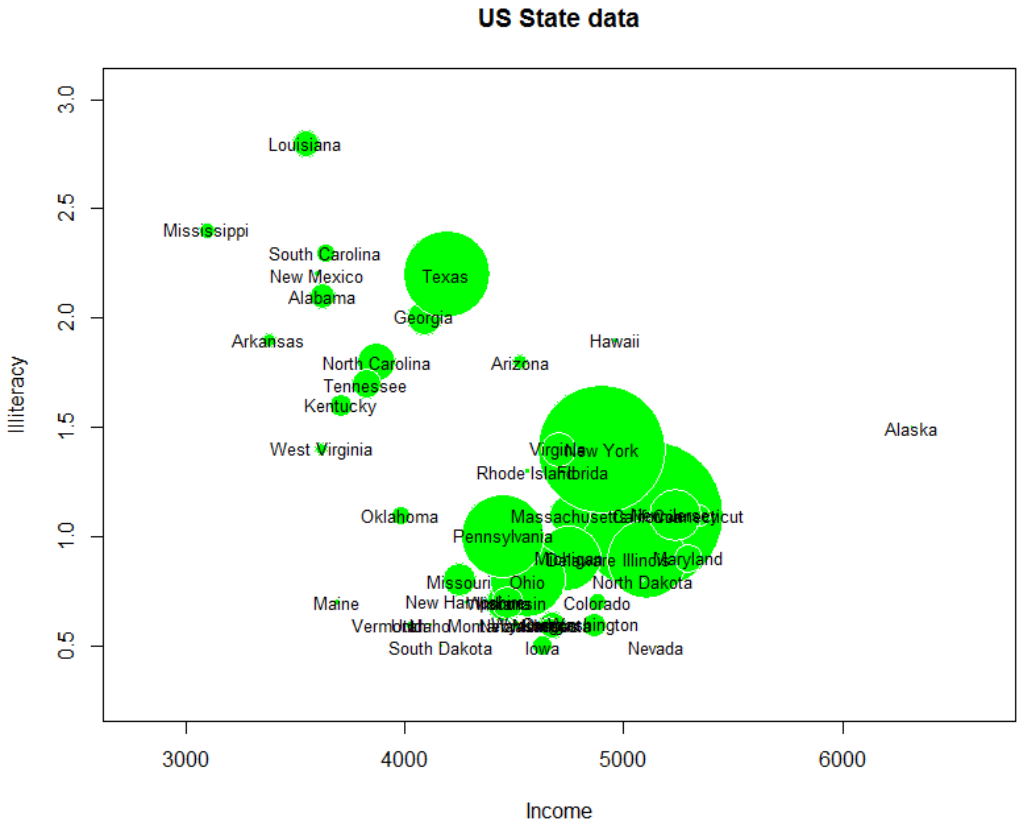
2. 버블 차트 (bubble chart)

```
library(MASS)
head(UScrime)
radius <- sqrt(UScrime$Pop) # 원의 반지름(값이커서 줄임)
symbols(UScrime$U2, UScrime$y, # 원의 x,y 좌표값
    circles=radius,            # 원의 반지름값
    inches=0.4,                # 원의 크기 조절값
    fg="white",                # 원의 테두리 색
    bg="lightgray",            # 원의 바탕색
    lwd=1.5,                   # 원의 테두리선 두께
    xlab="unemployment 35-39 males",
    ylab="crime rate",
    main="UScrime Data")
text(UScrime$U2, UScrime$y,   # 텍스트가 출력될 x,y좌표
    1:nrow(UScrime),          # 출력할 텍스트
    cex=0.8,                  # 폰트 크기
    col="brown")              # 폰트 color
```


UScrime Data
crime rate
500
1000
1500
2000
20
30
40
50
60
unemployment 35-39 males

[연습문제 2]

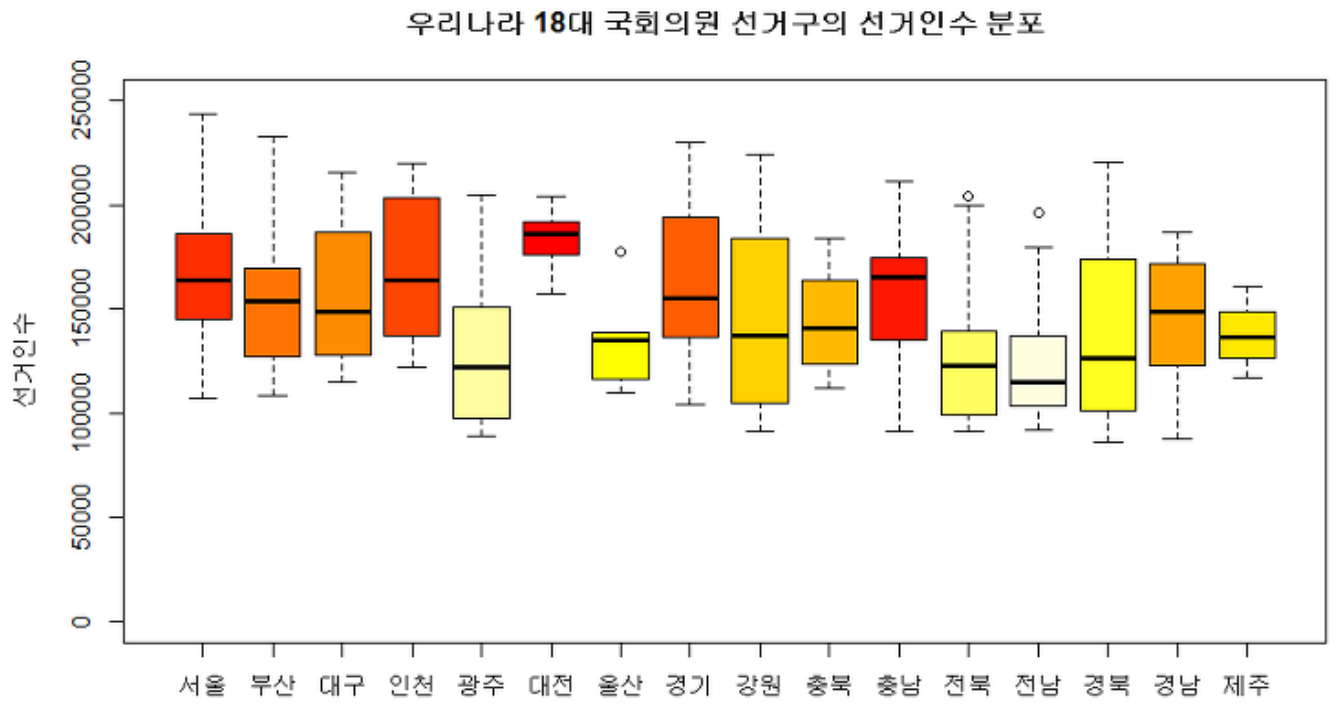
- state.x77 데이터로 부터 다음과 같은 버블차트를 작성하시오
  - st <- data.frame(state.x77) 과 같이 matrix를 data frame 으로 변환하여 사용
  - 원의 크기는 인구(Population) 수를 의미



  - 이 그래프로부터 관찰할 수 있는 것은 무엇인가

3. 다중 상자그림(Boxplot)

- 상자그림(box plot)은 일변량 연속형 자료를 상자와 선, 그리고 점으로 표현한 그림
- 다중 상자 그림은 총 자료가 여러 개의 자료 묶음(data batch)으로 구성되어 있는 경우 묶음 간 비교에 있어 시각적 효과가 탁월하다



http://blog.daum.net/huh420/19

3. 다중 상자그림(Boxplot)

- 설치가 필요한 패키지
  - 없음

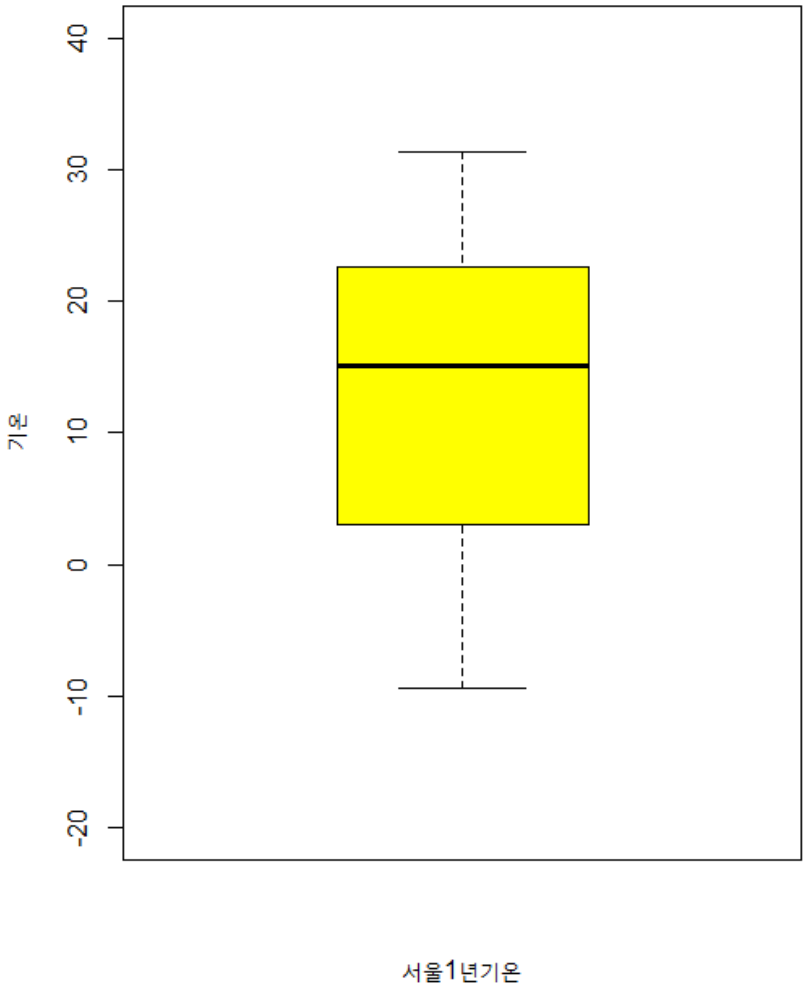
- 실습에 사용할 데이터셋 (2017년 서울 일별 평균기온)
  - Seoul_temp_2017.csv

```
         date month avg_temp
1  2017-01-01     1      2.7
2  2017-01-02     1      5.0
3  2017-01-03     1      2.0
4  2017-01-04     1      3.9
5  2017-01-05     1      3.8
6  2017-01-06     1      5.4
>
```

3. 다중 상자그림(Boxplot)

```
setwd("c:/Rworks") # 읽어올 데이터 파일이 있는 폴더지정
ds <- read.csv("seoul_temp_2017.csv")
head(ds)
summary(ds$avg_temp)

# 서울 1년 기온 분포
boxplot(ds$avg_temp,
        col="yellow",
        ylim=c(-20,40),
        xlab="서울1년기온",
        ylab="기온")
```

3. 다중 상자그림(Boxplot)



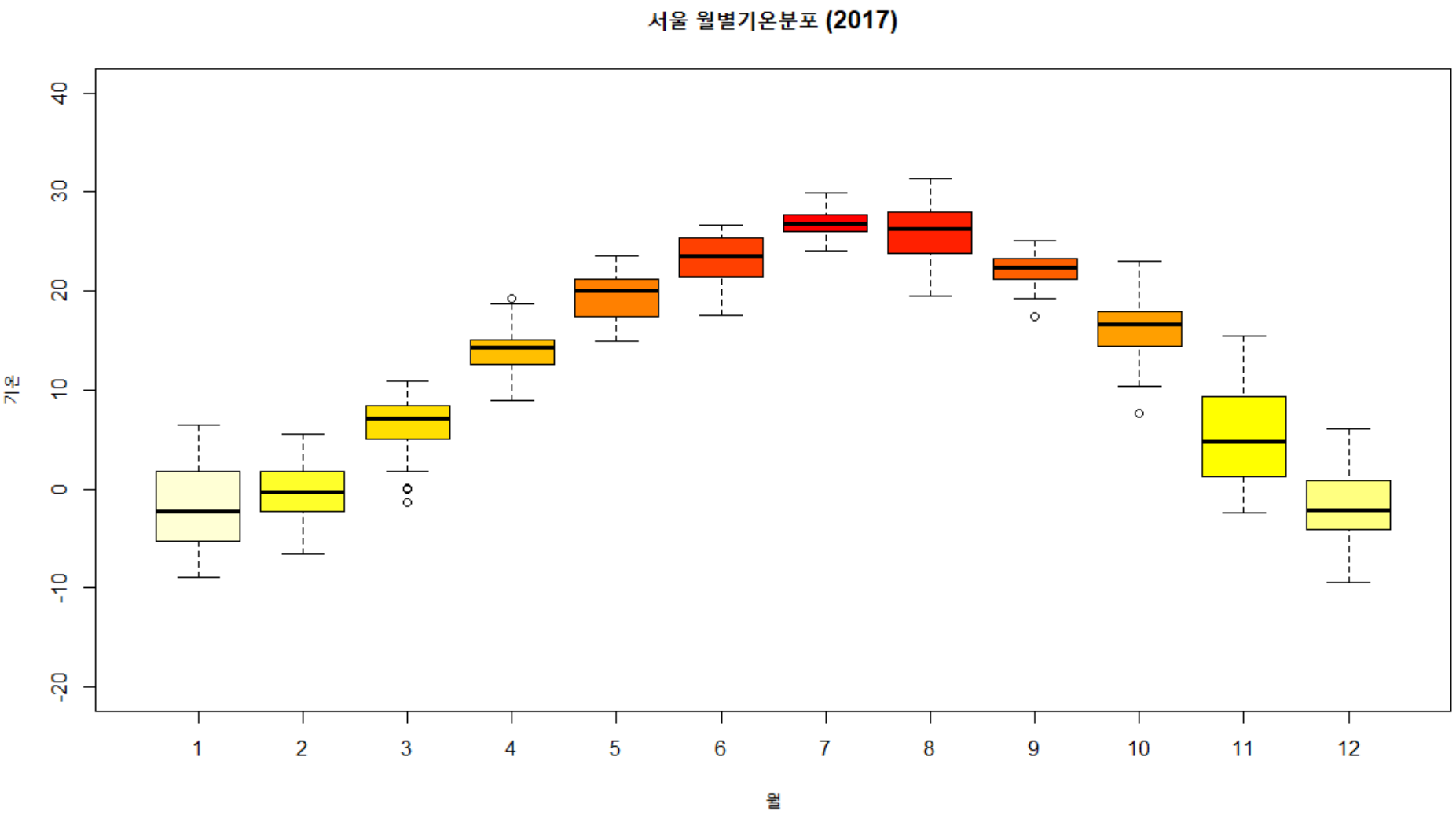
3. 다중 상자그림(Boxplot)

```
# 월별 평균기온계산
month.avg <- aggregate(ds$avg_temp,
                       by=list(ds$month),median)[2]

# 평균기온 순위 계산 (내림차순)
odr <- rank(-month.avg)

# 월별 기온분포
boxplot(avg_temp~month, data=ds,
            col=heat.colors(12)[odr], # 상자의 색을 지정
            ylim=c(-20,40),
            ylab="기온",
            xlab="월",
            main="서울 월별기온분포 (2017)")
```

col=heat.colors(12)[odr]

: 각 box 의 색을 heat.colors 에서 12개의 색을 취하여 그린다.
어느값을 취할지는 odr 에 따른다.

3. 다중 상자그림(Boxplot)

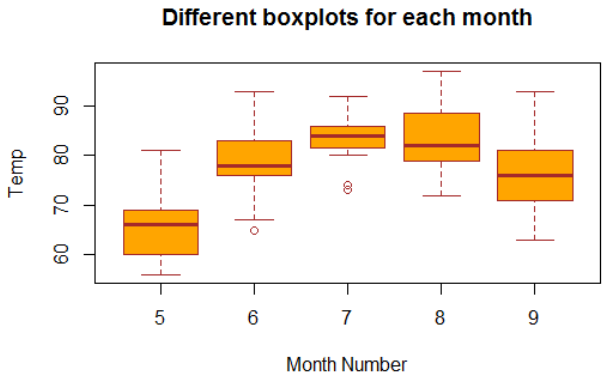


온도가 높을수록 붉은색, 낮을수록 연한 노랑색

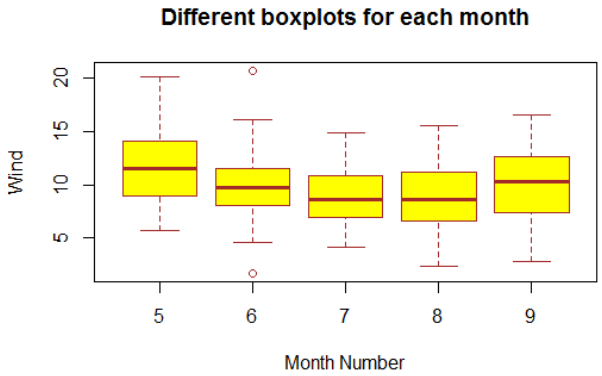
[연습문제 3]

- R 에서 제공하는 airquality 데이터셋을 이용하여 다음 문제를 해결하시오

1. 월별(Month) 기온(Temp)을 boxplot 으로 작성하시오

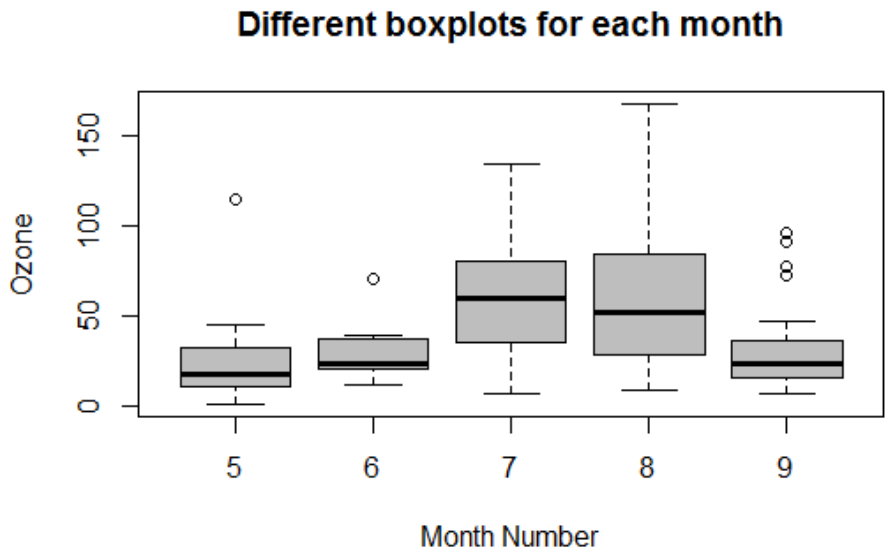


2. 월별(Month) 풍속(Wind)을 boxplot 으로 작성하시오



[연습문제 3]

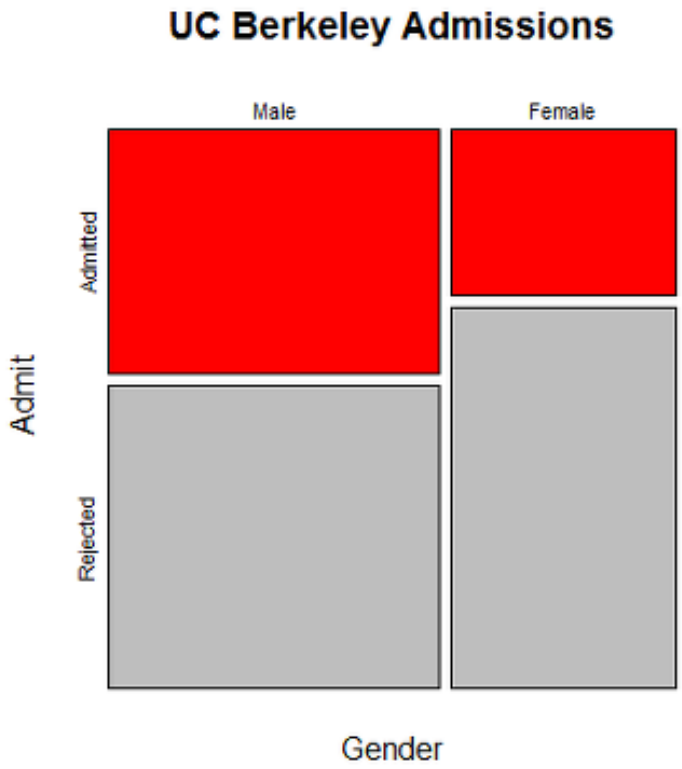
3. 월별(Month) 오존농도(Ozone)을 boxplot 으로 작성하시오



4. 각각의 boxplot 으로 부터 관찰할 수 있는 정보는 무엇인가

4. 모자이크 플롯 (mosaic plot)

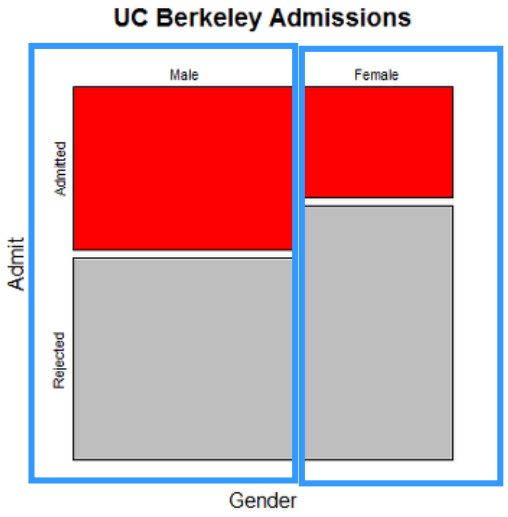
- 모자이크 플롯(mosaic plot)은 2원 3원 교차표의 시각화이다. 전체 정사각 도형을 교차표의 행 빈도에 비례하는 직사각 도형으로 나누고 다시 각 도형을 행 내 열의 빈도에 해당하는 직사각 도형으로 나눈다.



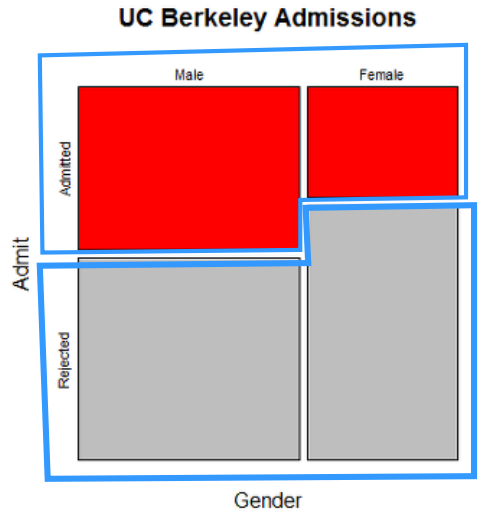
UC Berkeley 대학원 입시 통계

참조 : http://blog.daum.net/huh420/20

4. 모자이크 플롯 (mosaic plot)

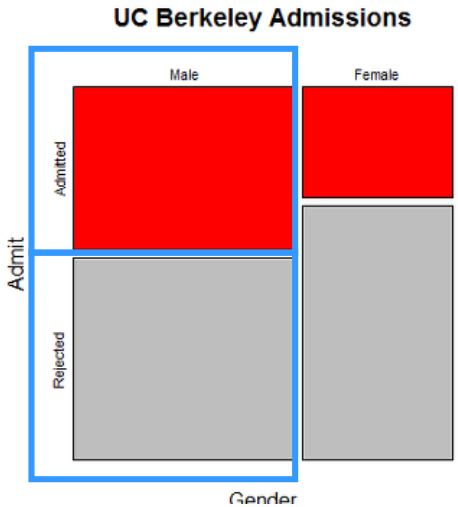


버클리 지원자 중 남성, 여성의 비율
(면적의 크기가 비율을 나타낸다)

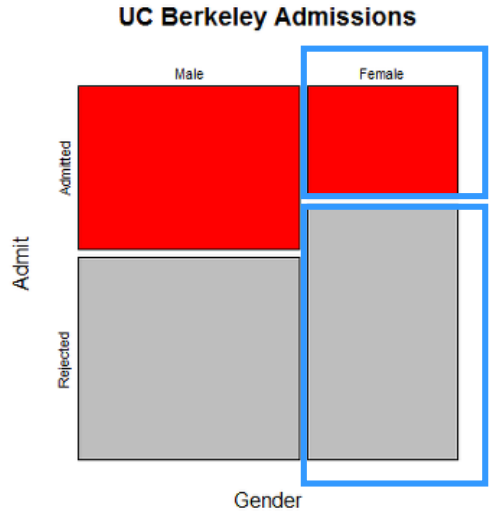


버클리 지원자 중 합격자와, 불합격자의 비율
(면적의 크기가 비율을 나타낸다)

4. 모자이크 플롯 (mosaic plot)

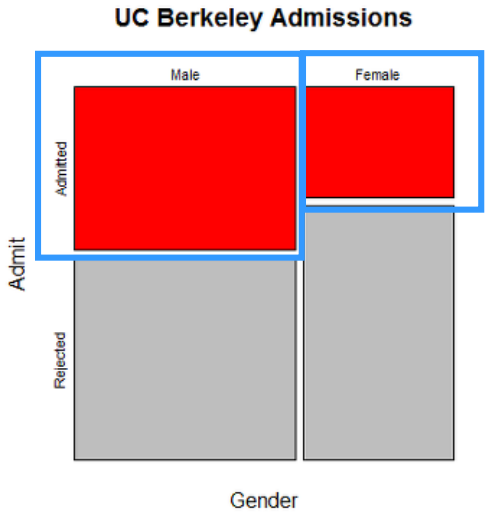


버클리 남성 지원자 중 합격자, 불합격자의 비율
(면적의 크기가 비율을 나타낸다)



버클리 여성 지원자 중 합격자, 불합격자의 비율
(면적의 크기가 비율을 나타낸다)

4. 모자이크 플롯 (mosaic plot)



버클리 남성 합격자와, 여성 합격자의 비율
(면적의 크기가 비율을 나타낸다)

(전체적으로는 남성이 합격자 수, 합격률에 있어서 여성보다 앞서는 것을 알 수 있다. -> 남녀차별 문제 제기)

이와 같이 모자이크 플롯은 여러가지 정보를 한눈에 표현할 수 있다

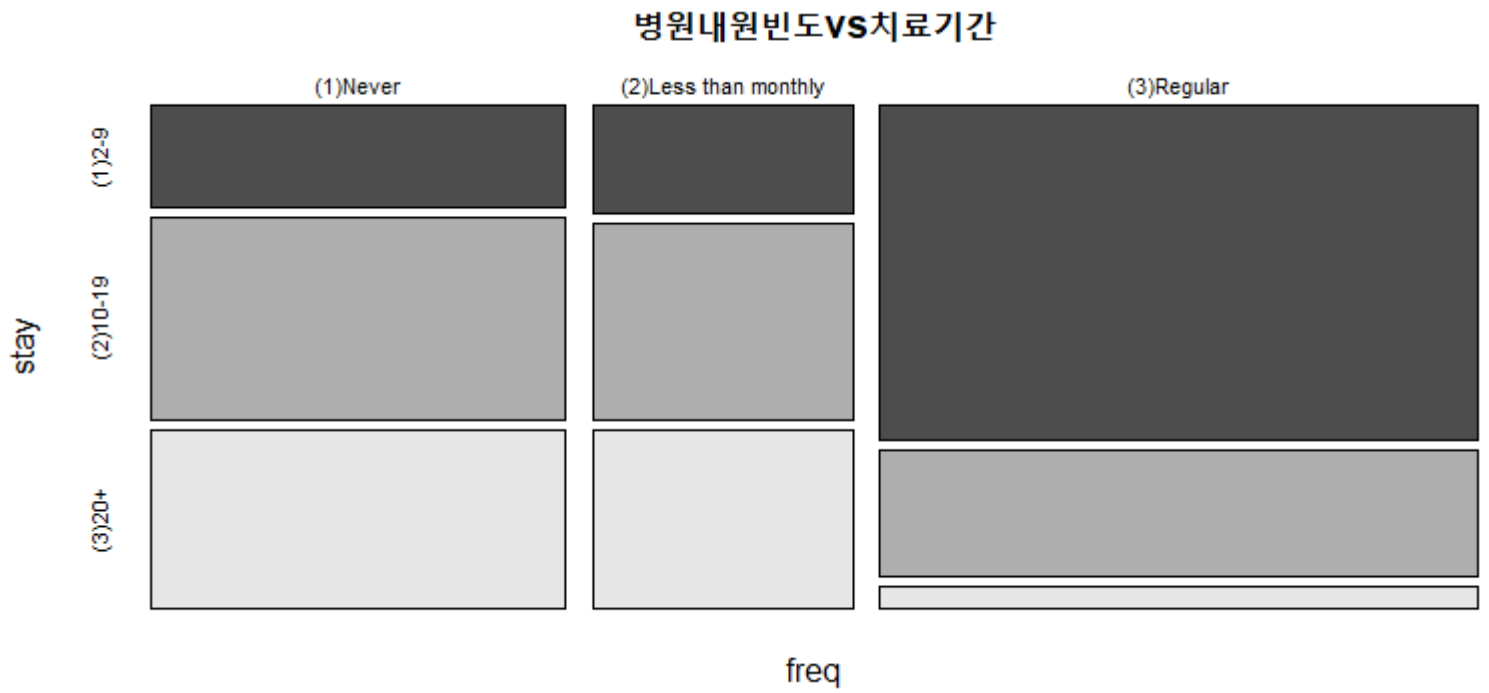
4. 모자이크 플롯 (mosaic plot)

- 설치 필요 패키지
  - 없음
- 실습용 데이터셋
  - mtcars
  - Titnic

4. 모자이크 플롯 (mosaic plot)

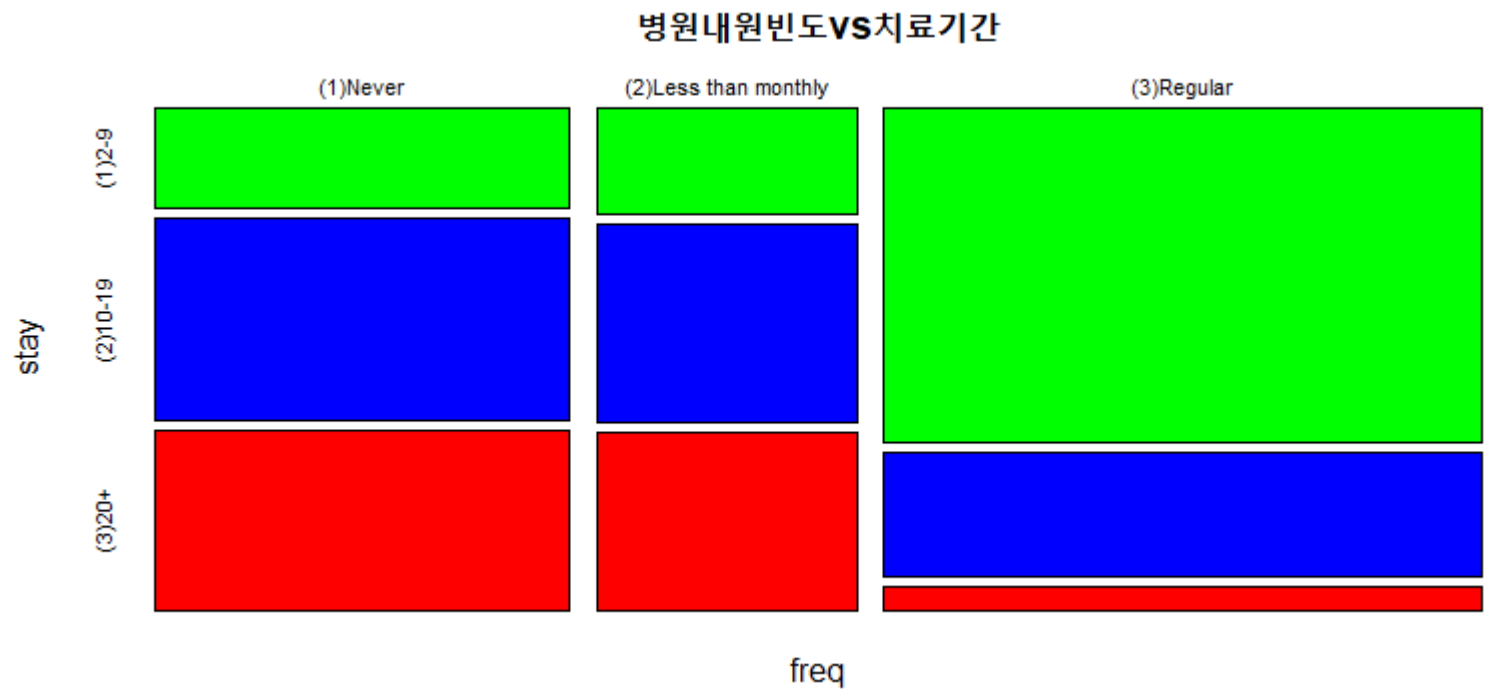
```
# matrix 형태로 데이터가 존재하는 경우
hospital <- read.csv("c:/Rworks/hospital.csv")
head(hospital)
table(hospital)
mosaicplot(~freq+stay, data = hospital, color=TRUE,
           main ="병원내원빈도vs치료기간")
```

132명의 조현병환자



4. 모자이크 플롯 (mosaic plot)

```
mosaicplot (~freq+stay, data = hospital,
            color =  c("green","blue","red"),
            main ="병원내원빈도vs치료기간")
```



4. 모자이크 플롯 (mosaic plot)

- Note.

```
mosaicplot(~freq+stay, data = hospital, color=TRUE,
main ="병원내원빈도vs치료기간")
```

```
# 교차표 형태의 데이터
tbl <- table(hospital)
mosaicplot(tbl, color=TRUE,
           main ="병원내원빈도vs치료기간")
```

```
> head(hospital)
        freq   stay
1 (3)Regular (1)2-9
2 (3)Regular (1)2-9
3 (3)Regular (1)2-9
4 (3)Regular (1)2-9
5 (3)Regular (1)2-9
6 (3)Regular (1)2-9
```

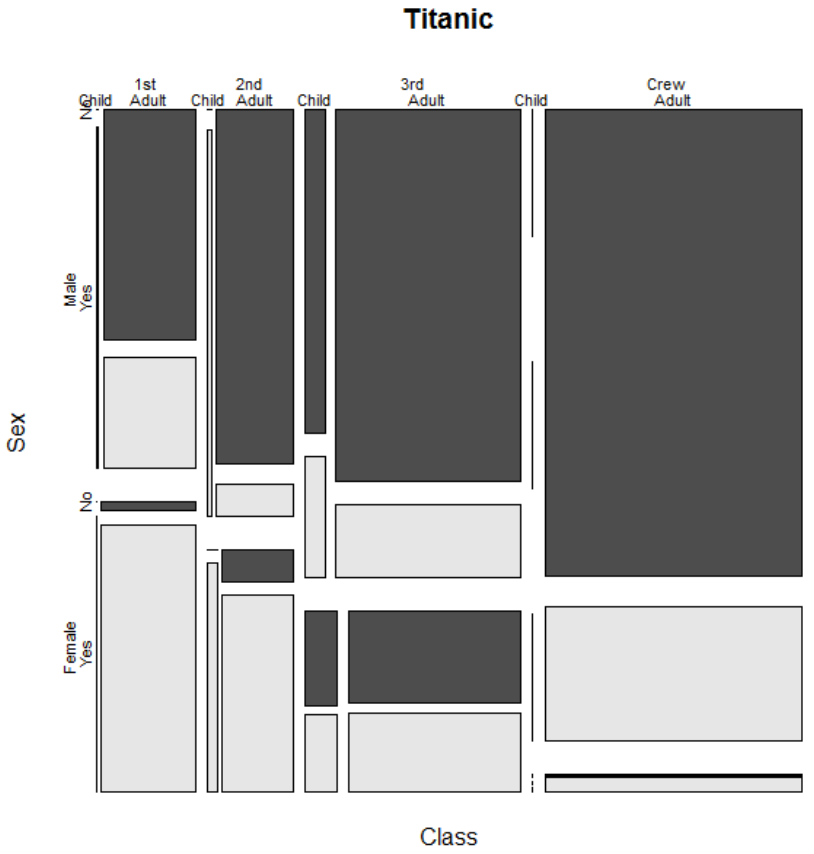
```
> table(hospital)
                     stay
freq                  (1)2-9 (2)10-19 (3)20+
  (1)Never                 9        18     16
  (2)Less than monthly     6        11     10
  (3)Regular              43        16      3
> .
```

4. 모자이크 플롯 (mosaic plot)

```
# 3차원 교차표 형태로 데이터가 존재하는 경우
Titanic
mosaicplot(Titanic, color = TRUE, off=5)
```

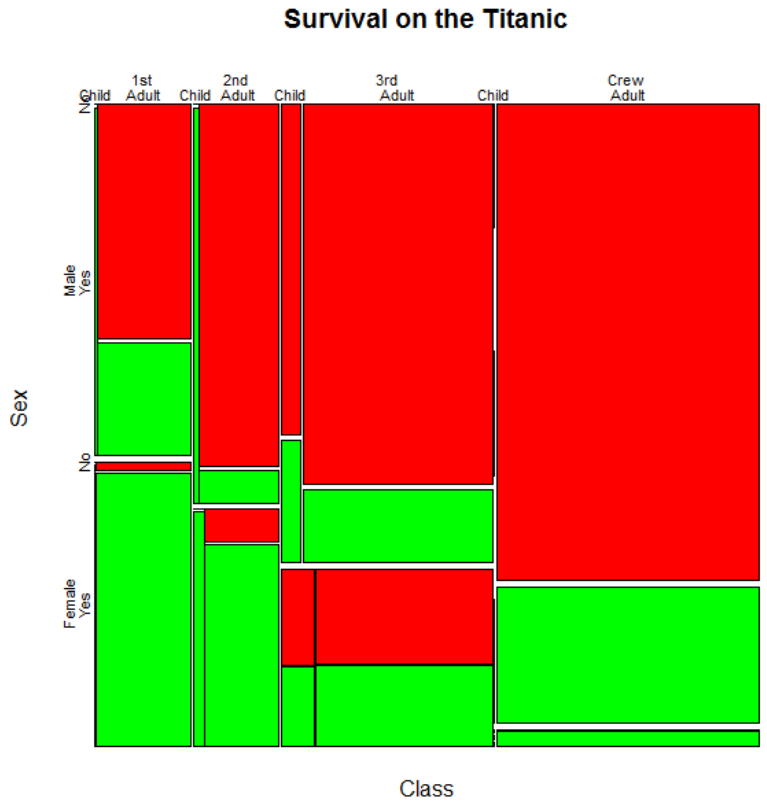
Block 간격 조절

```
> 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
```



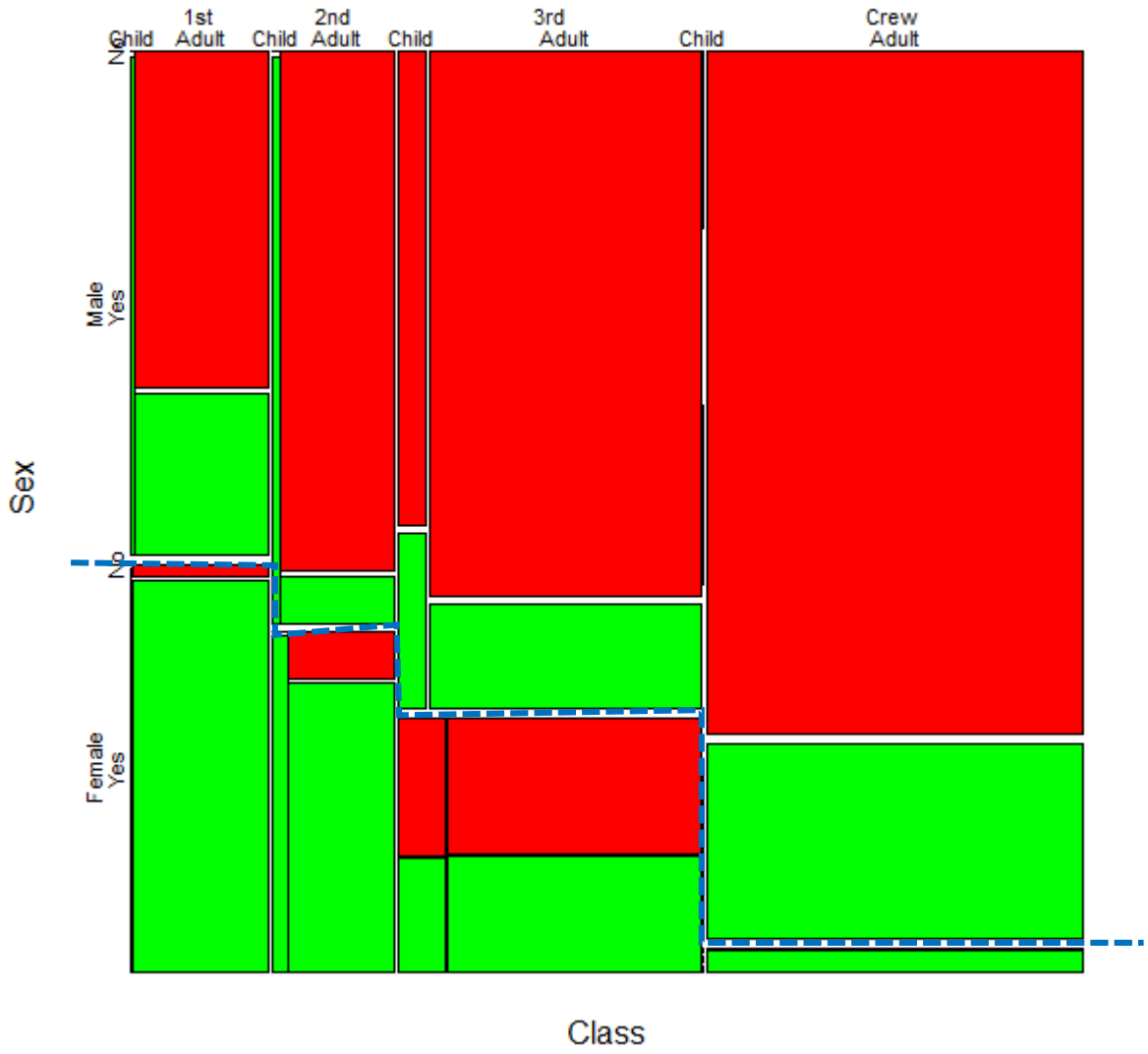
4. 모자이크 플롯 (mosaic plot)

```
mosaicplot(Titanic,
           main = "Survival on the Titanic",
           color = c("red","green"),
           off=1) # 블럭들 사이의 간격 지정
```



붉은색 : 사망
연두색 생존

Survival on the Titanic



붉은색 : 사망
연두색 생존

[연습문제 4]

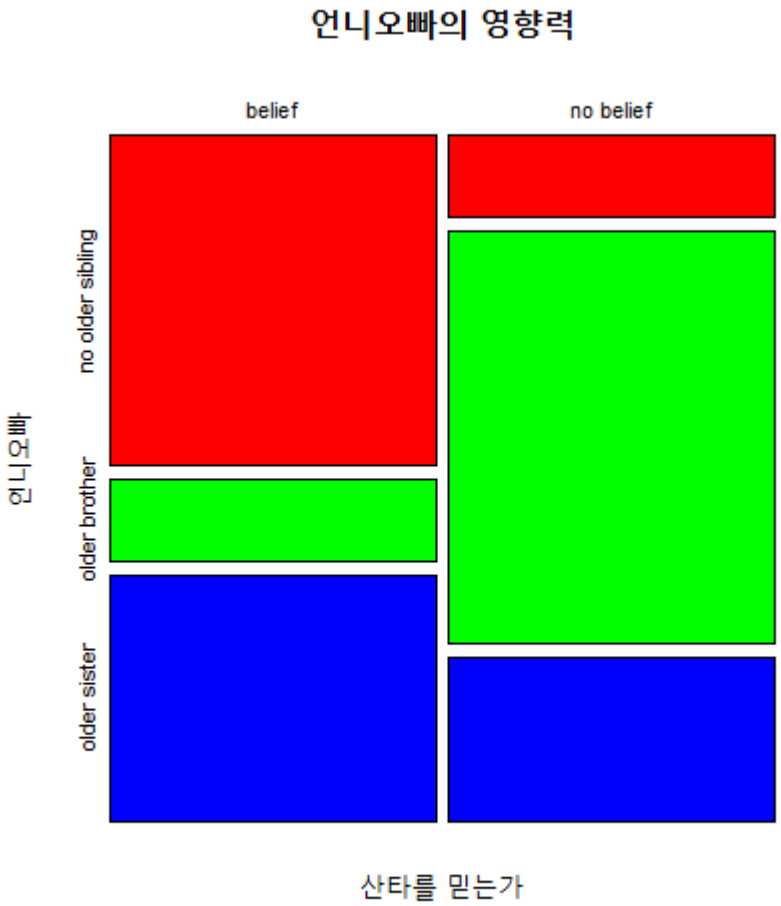
1. HairEyeColor 데이터셋에 대해 모자이크 플롯을 작성하시오. 여기서 관찰할 수 있는 정보는 무엇인가

2. 다음의 santa data 에 대해 모자이크 플롯을 작성하시오 (다음 slide 참조). 여기서 관찰할 수 있는 정보는 무엇인가

```
santa <- data.frame(belief=c('no belief','no belief','no belief','no belief',
                             'belief','belief','belief','belief',
                             'belief','belief','no belief','no belief',
                             'belief','belief','no belief','no belief'),
                    sibling=c('older brother','older brother','older brother','older sister',
                              'no older sibling','no older sibling','no older sibling','older sister',
                              'older brother','older sister','older brother','older sister',
                              'no older sibling','older sister','older brother','no older sibling')
                    )
```

* belief : 산타를 믿는지여부, sibling: 손위 형제가 있는지 여부

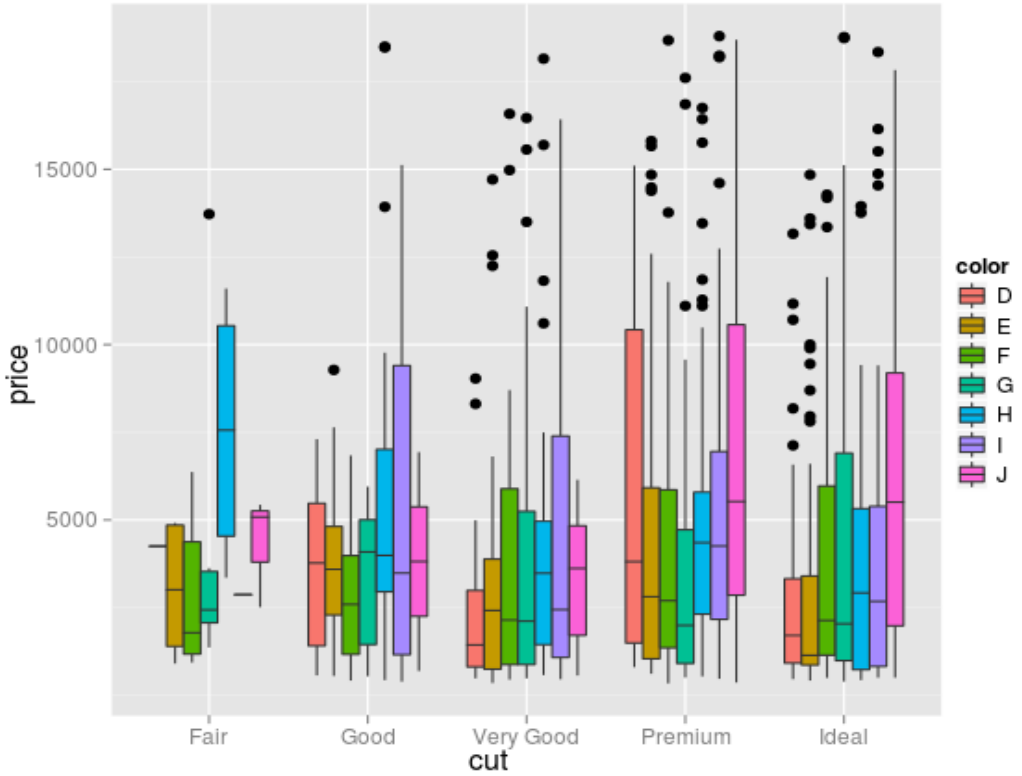
[연습문제 4]



http://www.cyclismo.org/tutorial/R/intermediatePlotting.html

5. ggplot

- R 에서 제공하는 기본 그래프 도구들로도 시각화가 가능하지만 보다 미적인 시각화를 위해서는 ggplot 을 사용한다.
- 설치 패키지 : ggplot2



http://chrisladroue.com/extra/ggplot2Intro/introductionGGplot2/

5. ggplot

- 기본 문법

Style 1

```
ggplot(data=xx, aes(x=x1,y=x2))+geom_xx()+ ..
```

그래프를 그릴 대상 dataset

X축 데이터

Y축 데이터

그래프의 형태 지정

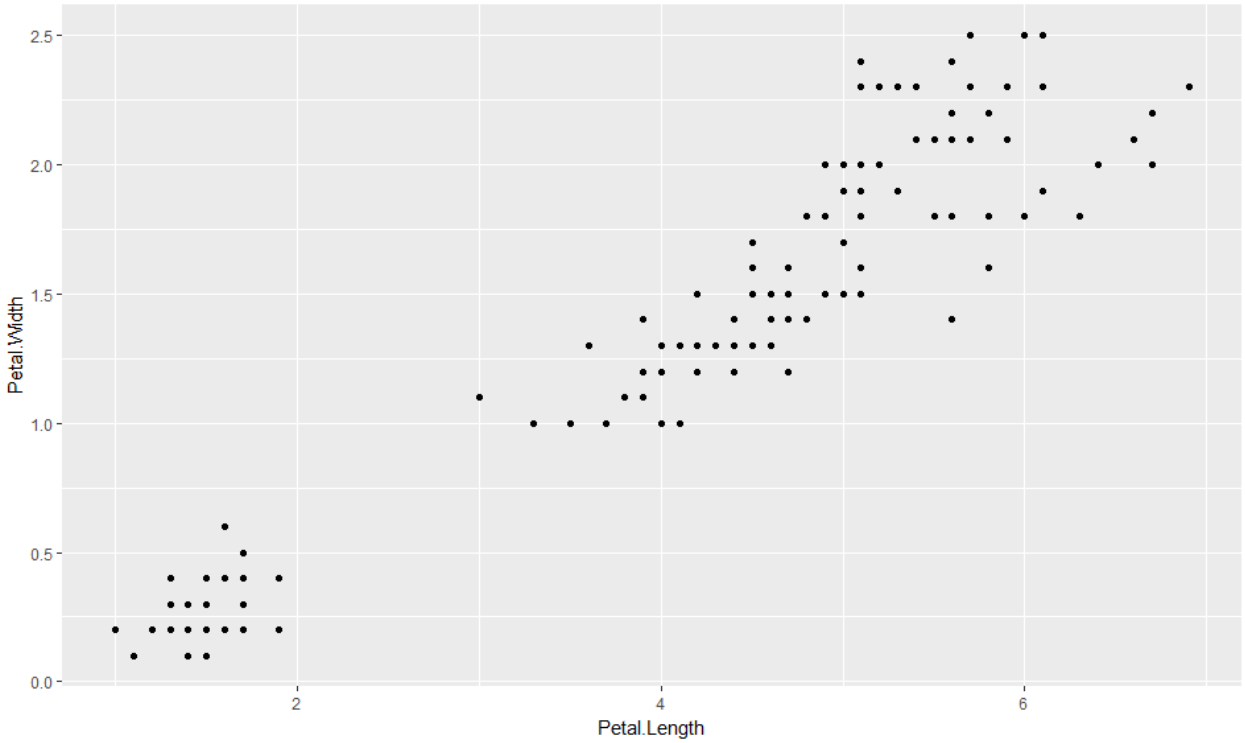
Style 2

```
ggplot()+geom_xx(data=xx, aes(x=x1,y=x2))+ ..
```

Style 1 은 단일 그래프를 그릴 때, style2 는 여러 그래프를 하나로 겹쳐 그릴 때 편리하다

5. ggplot

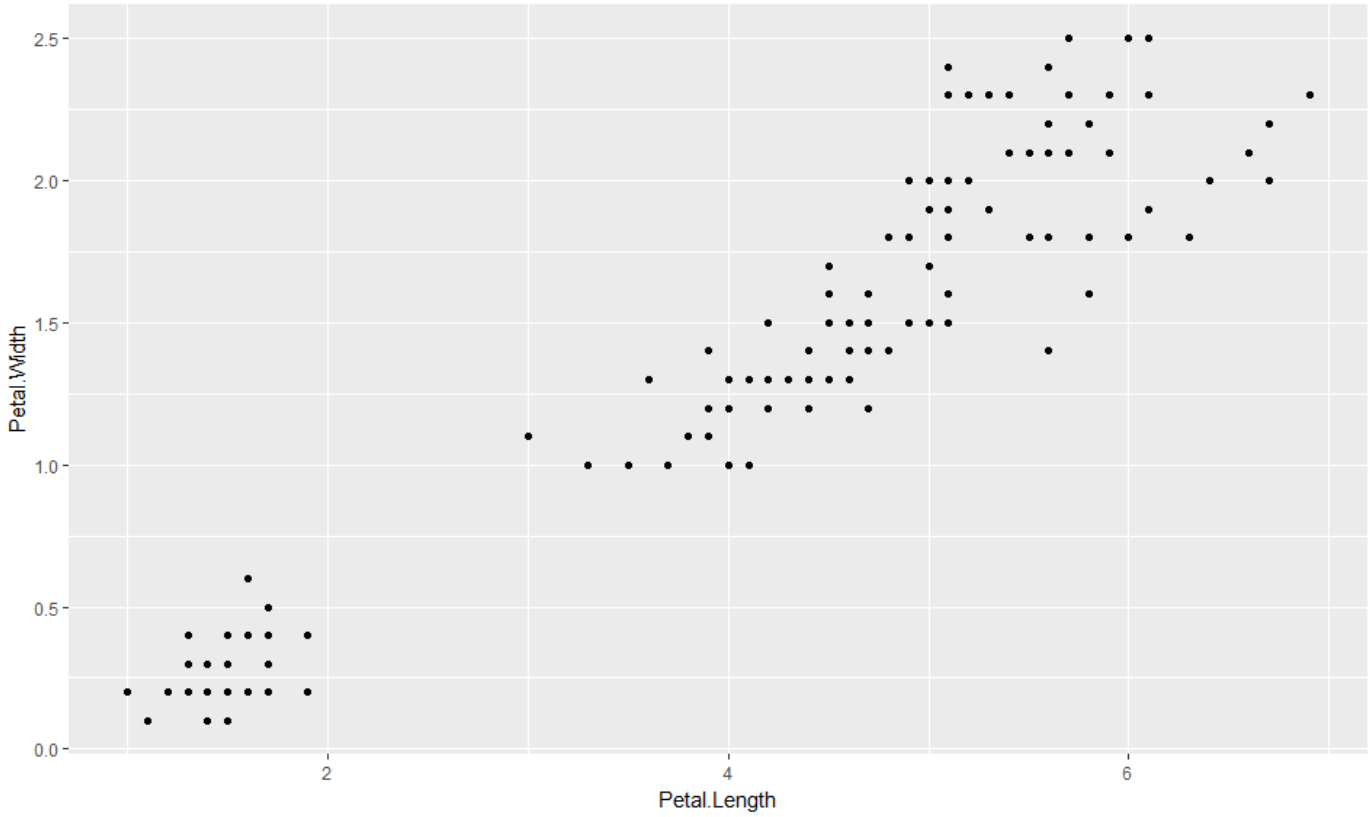
```
library(ggplot2)

ggplot(data=iris, aes(x=Petal.Length,
       y=Petal.Width))+geom_point()
```



 geom_point(): 산점도

5. ggplot

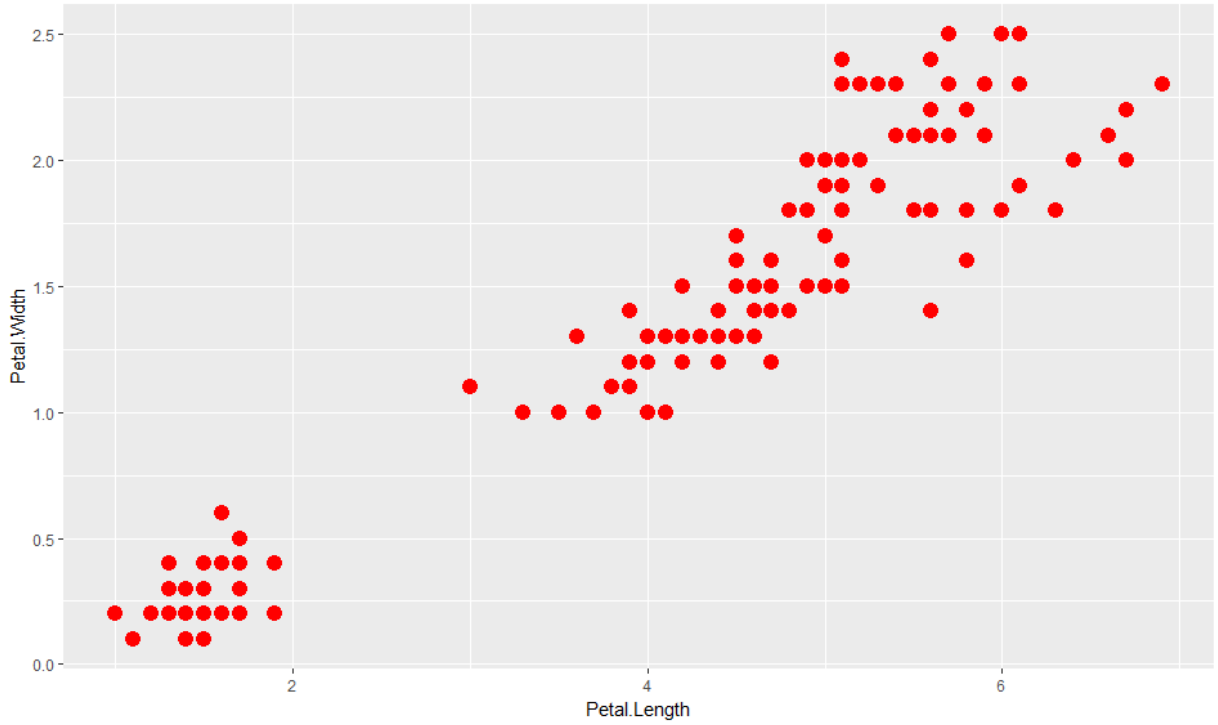
```
ggplot()+geom_point(data=iris,aes(x=Petal.Length,
                                  y=Petal.Width))
```



5. ggplot

- 옵션의 추가

```
ggplot()+geom_point(
    data=iris,
    aes(x=Petal.Length, y=Petal.Width),
    color="red",
    size=4)
```

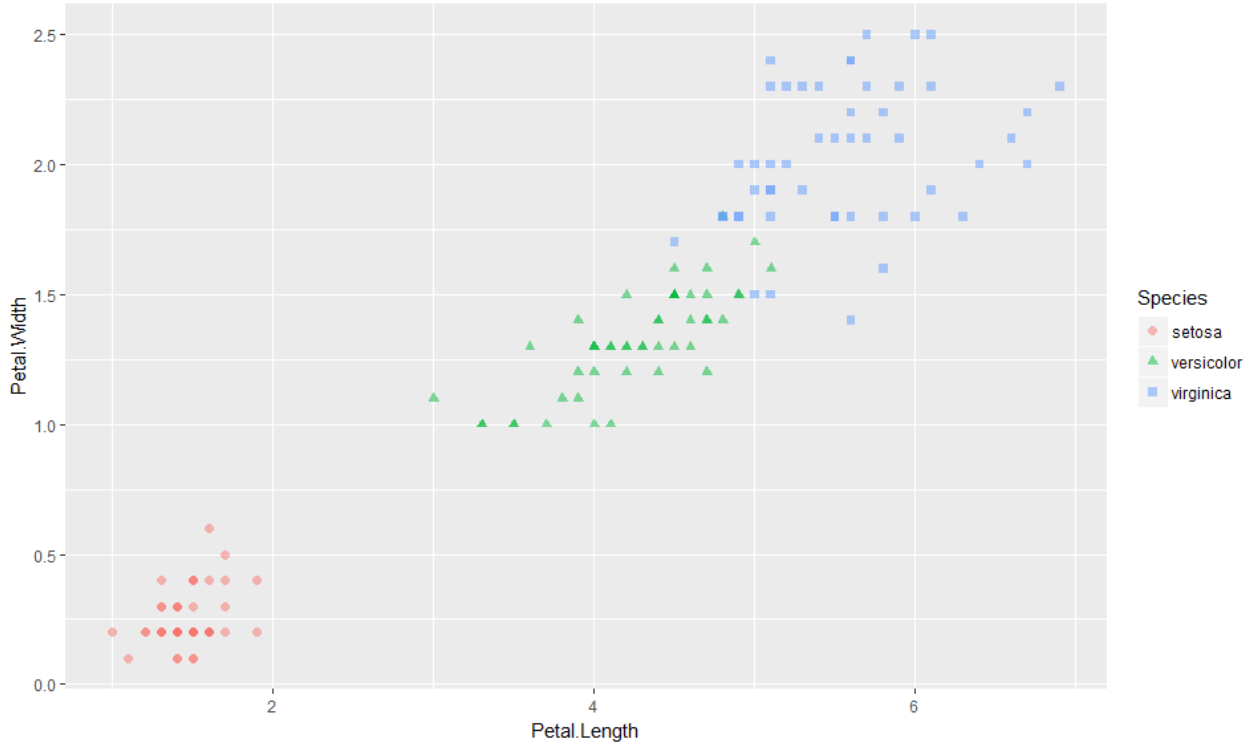


5. ggplot

- geom_point() 옵션
  - x : x축 데이터
  - y : y축 데이터
  - alpha : 점의 투명도
  - colour(color) : 점의 색깔
  - fill : 점안을 채울 색깔 (shape=21 과 같은 경우 ○)
  - group : 데이터의 그룹정보 (그룹에 따라 점의 모양, 색깔을 달리할때)
  - shape : 점의 모양
  - size : 점의 크기
  - Stroke : 테두리 굵기 ○

5. ggplot

```
ggplot(data=iris,
       aes(x=Petal.Length, y=Petal.Width))+
  geom_point(
    aes(color=Species, shape=Species),
    alpha=0.5,
    size=2)
```



5. ggplot

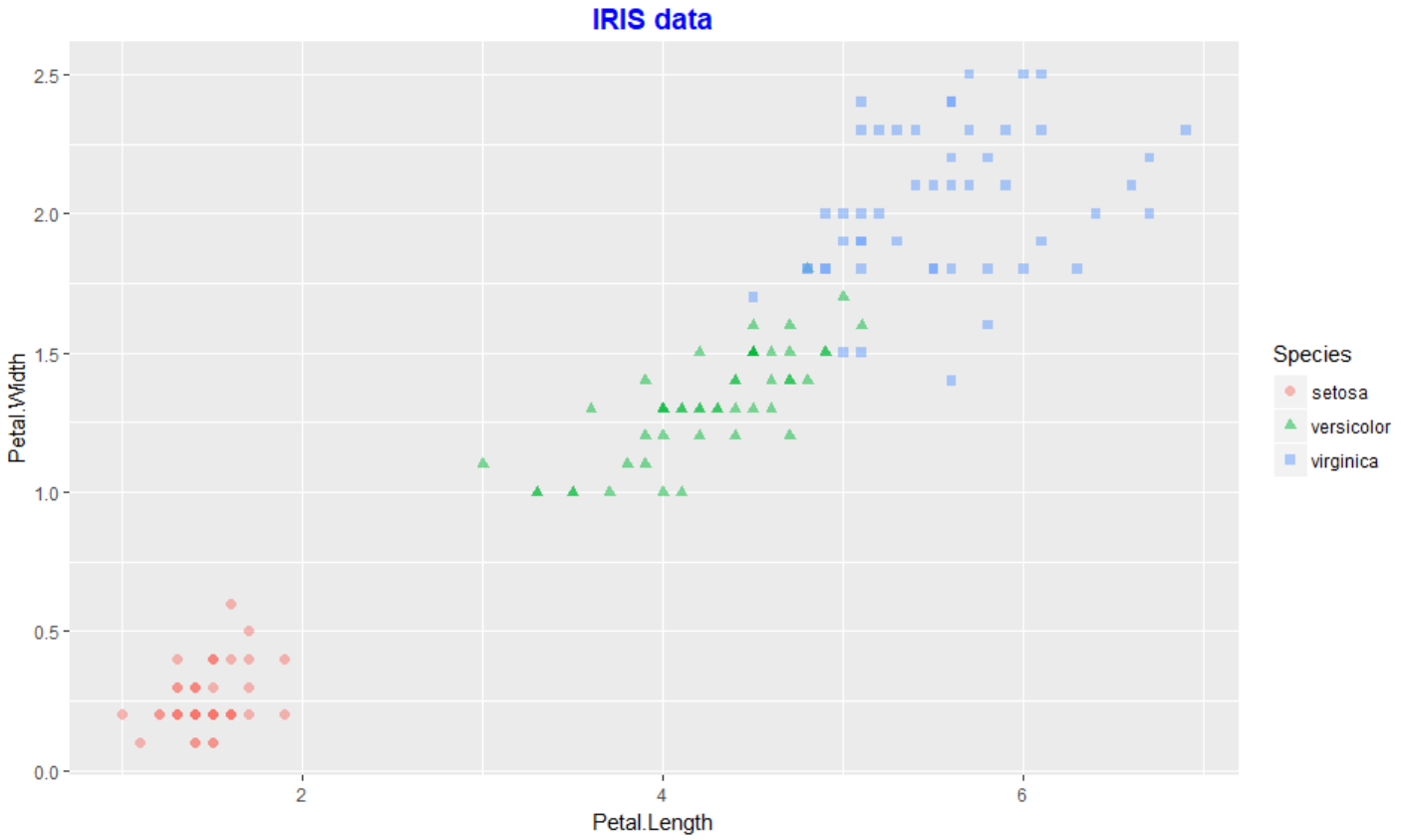
- 그래프 title

```
gp <- ggplot(data=iris,
         aes(x=Petal.Length, y=Petal.Width))+
      geom_point(
         aes(color=Species, shape=Species),
         alpha=0.5,
         size=2)
gp
```

5. ggplot

- 그래프 title

```
gp+ggtitle("IRIS data")+
  theme(plot.title = element_text(size=14,
  face="bold",color="blue", hjust=0.5))
```



5. ggplot

- 그래프 title

```
gp2 <- gp+ggtitle("IRIS data")+
        theme(plot.title = element_text(size=14,
        face="bold",color="blue", hjust=0.5))
gp2
```

hjust=0.5: title을 가운데 정렬
(0:왼쪽정렬, 1:오른쪽 정렬)

Note

- 여러줄에 걸쳐 명령어를 작성할 때, 다음줄에 명령이 이어진다는 것을 표현해 주어야 한다

```
ggplot(data=my.iris, aes(x=Petal.Length,
       y=Petal.Width))
+geom_point()
```

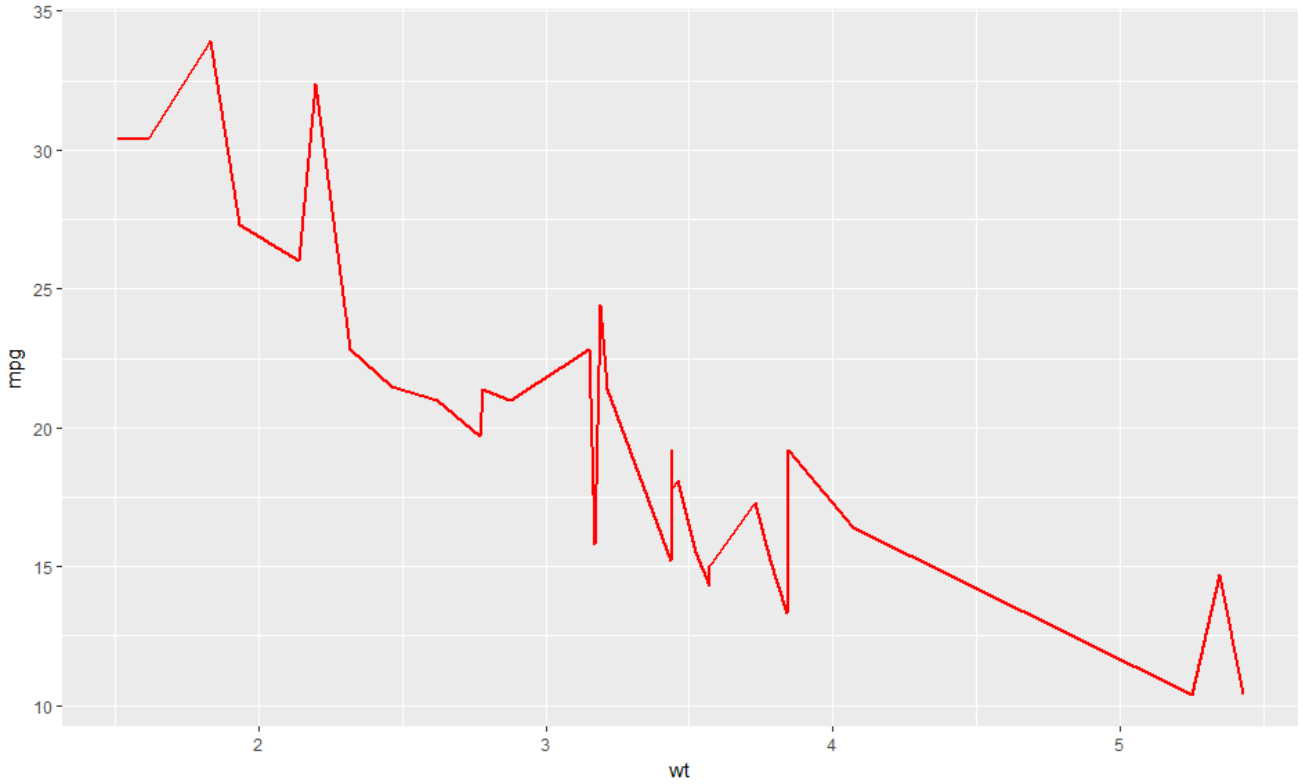
```
ggplot(data=my.iris, aes(x=Petal.Length,
       y=Petal.Width))+
geom_point()
```

실습x

5. ggplot

- 꺽은선 그래프 예제

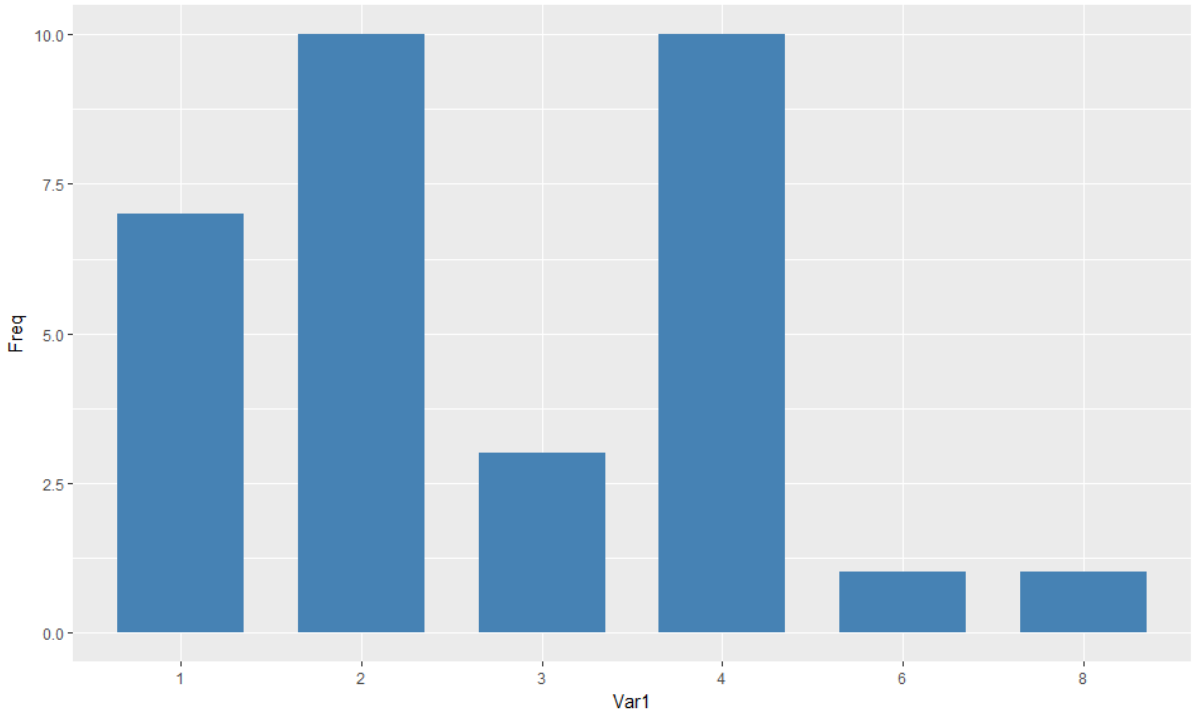
```
ggplot(mtcars, aes(x=wt, y=mpg)) +
  geom_line(color='red', size=1)
```



5. ggplot

- 막대 그래프 예제

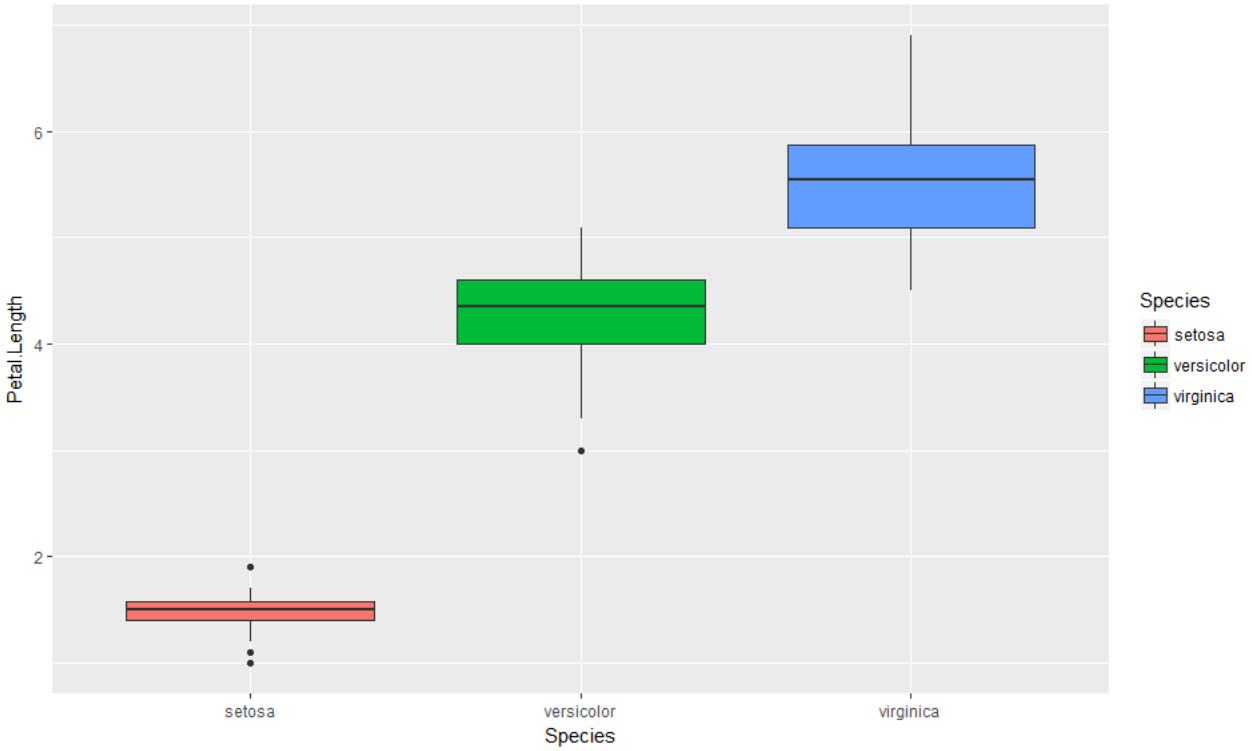
```
df = data.frame(table(mtcars$carb))
df
ggplot(df, aes(x=Var1,y=Freq))+
  geom_bar(stat="identity",  width=0.7,
  fill="steelblue")
```



5. ggplot

- 상자그림 예제

```
ggplot()+
  geom_boxplot(data=iris,
               aes(x=Species,y=Petal.Length,
               fill=Species))
```



5. ggplot

- Reference list

- Introduction to R graphics with ggplot2 (http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html)

- Top 50 ggplot2 Visualizations (http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html)

실습x

[연습문제 5]

1. R에서 제공하는 state.x77 데이터셋의 Income, Illiteracy 데이터를 가지고 ggplot 으로 산점도를 작성하시오

2. R에서 제공하는 mtcars 데이터셋의 gear 데이터를 가지고 기어수별 빈도에 대해 ggplot 으로 막대 그래프를 작성하시오

3. R에서 제공하는 airmiles 데이터셋은 1937년~1960년까지 비행기 탑승객의 여행거리가 저장되어 있다. ggplot 으로 선그래프를 작성하시오 (x축:년도, y축:여행거리. airmiles 는 벡터가 아니기 때문에 다음과 같이 벡터로 바꾼 다음 실행한다. am <- as.numeric(airmiles)

4. R에서 제공하는 iris 데이터셋의 Petal.Width 에 대해 품종(Species)별 상자그림(boxplot) 을 ggplot 으로 작성하시오