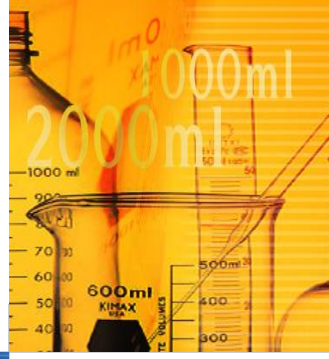
R 프로그래밍



Chapter 5

다변량 자료 탐색

Sejong Oh

Bio Information technology Lab.

Content

- 개요
- 산점도
- 선 그래프
- 상관 분석
- 데이터분석의 실제: iris

개요

- 다변량 자료 :
  - 키와 몸무게의 관계와 같이 두개 이상의 변수를 동시에 다루어야 하는 자료
  - 두개인 경우를 특히 이변량 자료라고 한다
  - 일변량 자료는 vector 에 저장하여 분석할 수 있고, 다변량 자료는 matrix 또는 data frame 에 저장하여 분석한다

  - 변수(variable)는 데이터셋에서 열로 표현된다 (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width)

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

개요

- 용어

variable,
feature,
attribute

observation,
instance

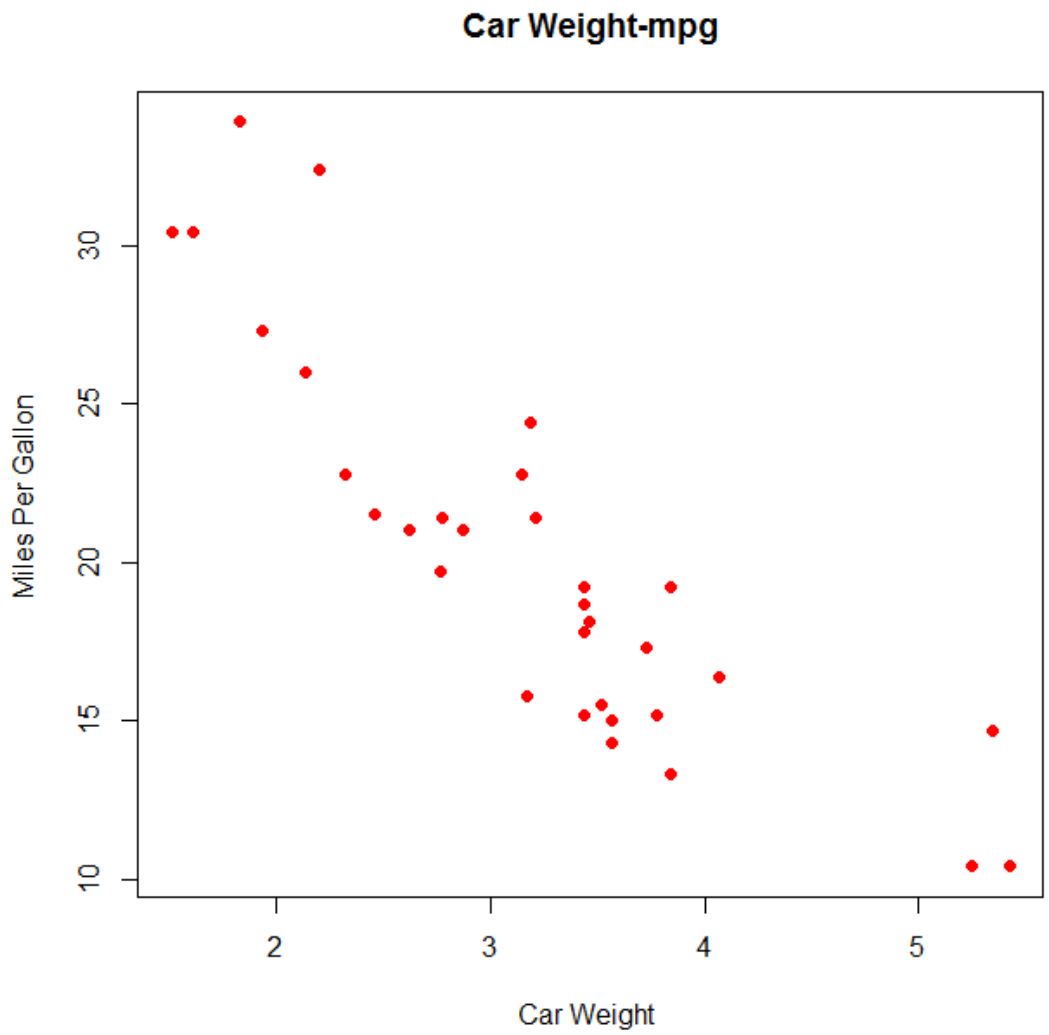
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

산점도(Scatter plot)

- 2변량 자료의 분포 및 상관관계를 시각적으로 확인
  - mtcars 데이터셋에서 자동차 중량(wt) 와 연비(mpg) 의 상관관계를 산점도를 통해 확인해 보자

```
wt <-mtcars$wt
mpg <- mtcars$mpg
plot(wt, mpg,                          # 2개 변수
     main="Car Weight-mpg",            # 제목
     xlab="Car Weight ",               # x축 레이블
     ylab="Miles Per Gallon ",         # y축 레이블
     col="red",                        # point 의 color
     pch=19)                           # point 의 종류
```

산점도(Scatter plot)



중량이 큰 차일수록 연비가 떨어지는 것을 관찰할 수 있다

산점도(Scatter plot)

- plot (x축 데이터, y축 데이터, 옵션)
- 옵션↓

인수	설명
main="메인제목"	제목설정
sub="서브제목"	서브제목설정
xlab="문자", ylab="문자"	x,y 축에 사용할 문자열을 지정
ann=F	x,y 축 제목을 지정하지 않음
tmag=2	제목 등에 사용되는 문자의 확대률 지정
axes =F	x,y 축을 표시하지 않음
axis	x,y 축을 사용자의 지정값으로 표시

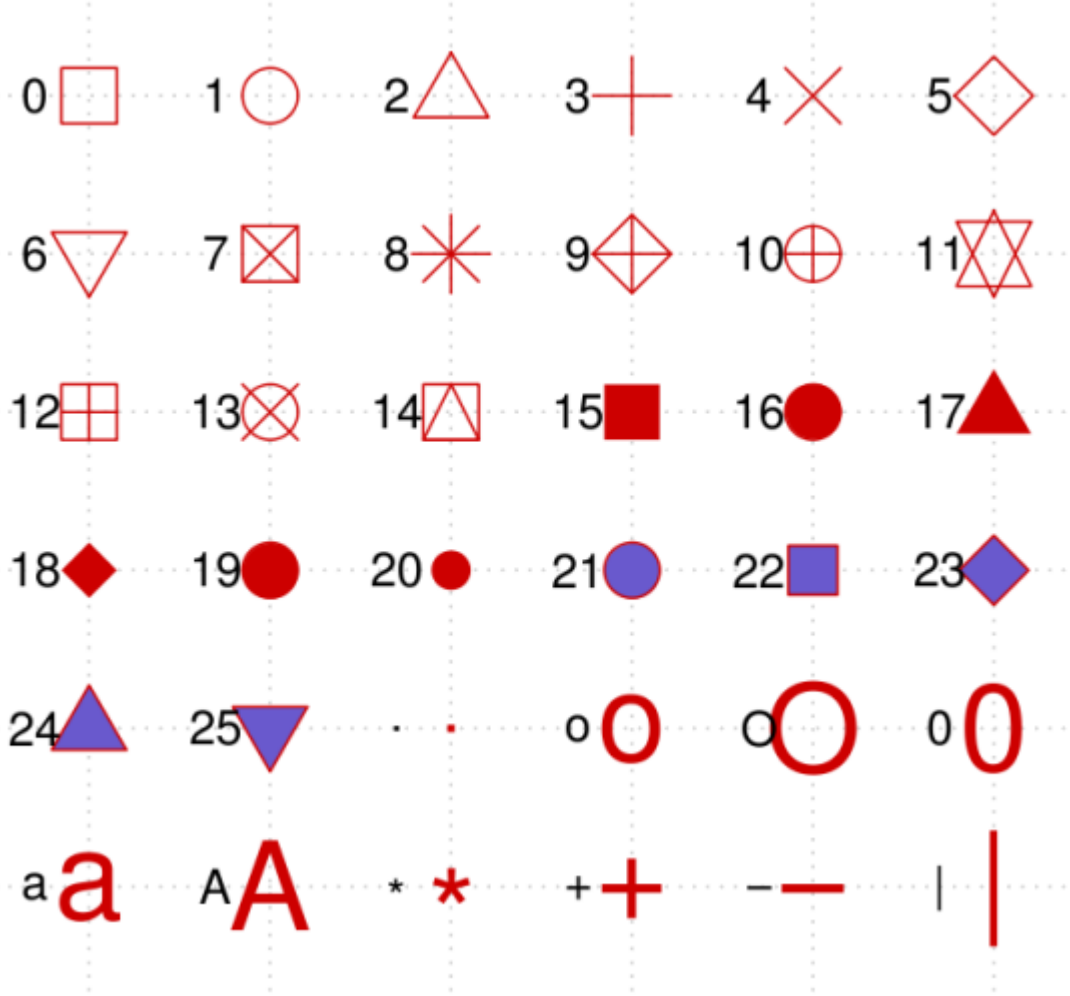
그래프 타입 선택	
type="p"	점 모양 그래프 (기본값)
type="l"	선 모양 그래프 (꺾은선 그래프)
type="b"	점과 선 모양 그래프
type="c"	"b"에서 점을 생략한 모양
type="o"	점과 선을 중첩해서 그린 그래프
type="h"	각 점에서 x축 까지의 수직선 그래프
type="s"	왼쪽값을 기초로 계단 모양으로 연결한 그래프
type="S"	오른쪽 값을 기초로 계단모양으로 연결한 그래프
type="n"	축만 그리고 그래프는 그리지 않음

산점도(Scatter plot)

선의 모양 선택	
lty=0, lty="blank"	투명선
lty=1, lty="solid"	실선
lty=2, lty="dashed"	대쉬선
lty=3, lty="dotted"	점선
lty=4, lty="dotdash"	점선과 대쉬선
lty=5, lty="longdash"	긴 대쉬선
lty=6, lty="twodash"	2개의 대쉬선
색, 기호 등	
col=1, col="blue"	기호의 색지정 1:검정, 2:빨강, 3:초록, 4:파랑, 5:연파랑, 6:보라, 7:노랑, 8:회색
pch=0, pch="문자"	점의 모양 지정
bg ="blue"	그래프의 배경색 지정
lwd="숫자"	선을 그릴때 선의 굵기 지정
cex="숫자"	점이나 문자를 그릴때 점이나 문자의 굵기를 지정

산점도(Scatter plot)

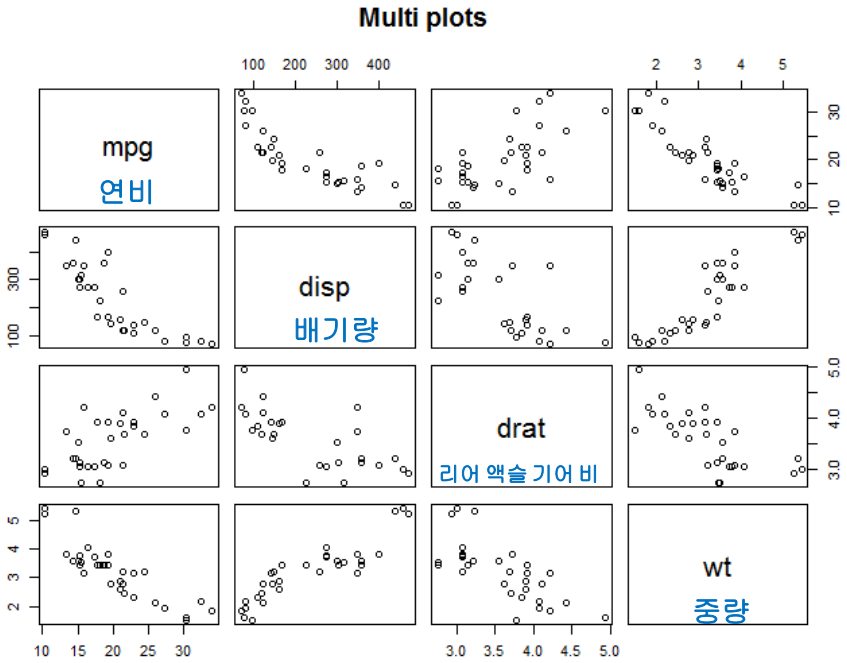
- 포인트의 종류 (pch)



산점도(Scatter plot)

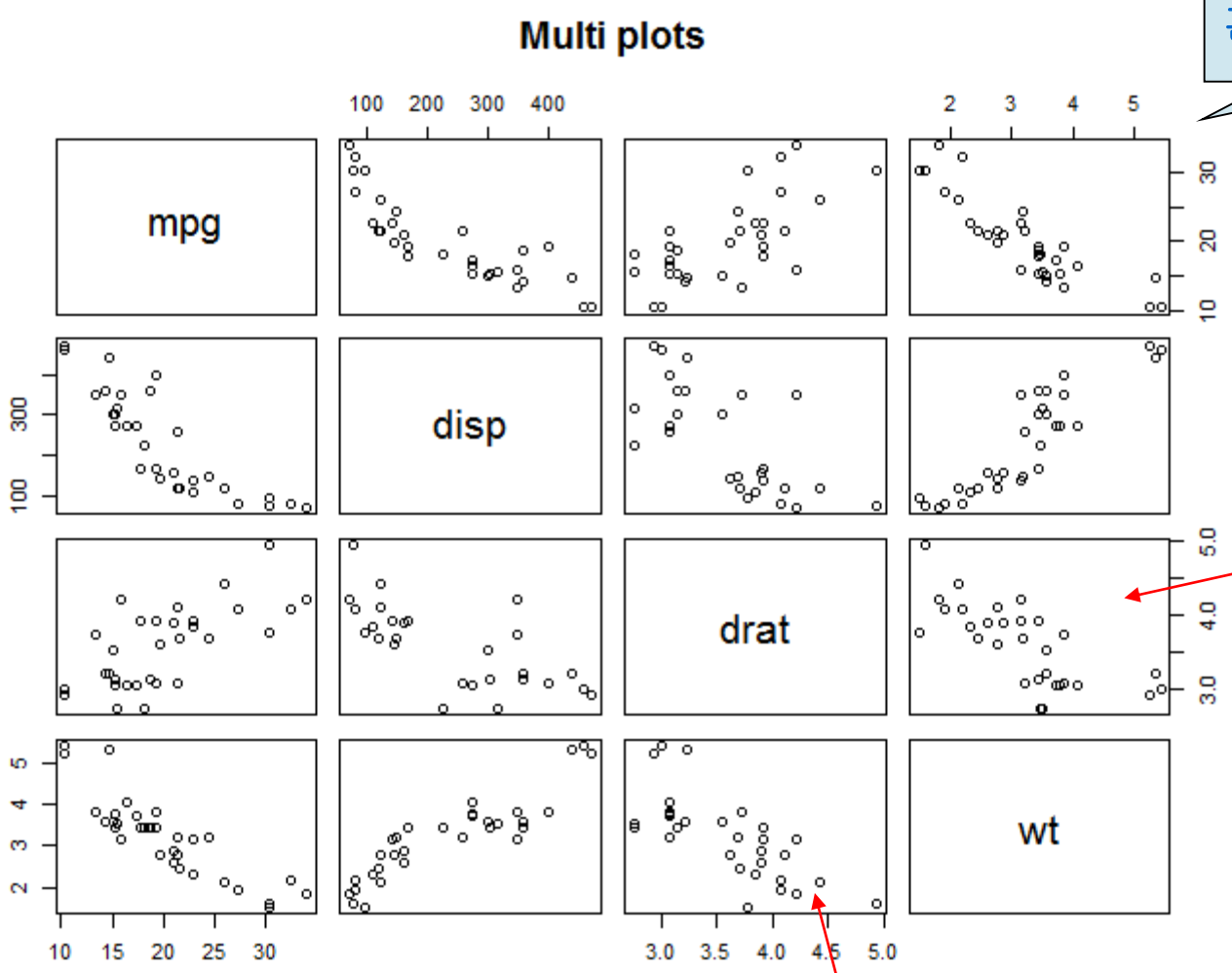
- pairs() : 여러 변수들 사이의 상관관계를 한번에 확인

```
vars <- c("mpg","disp","drat","wt") # 대상 변수
target <- mtcars[,vars]
pairs(target,                       # 대상 데이터
      main="Multi plots")
```



산점도(Scatter plot)

이 그래프를 보고
관찰할 수 있는 것은?



x축: wt
y축: drat

x축: drat
y축: wt

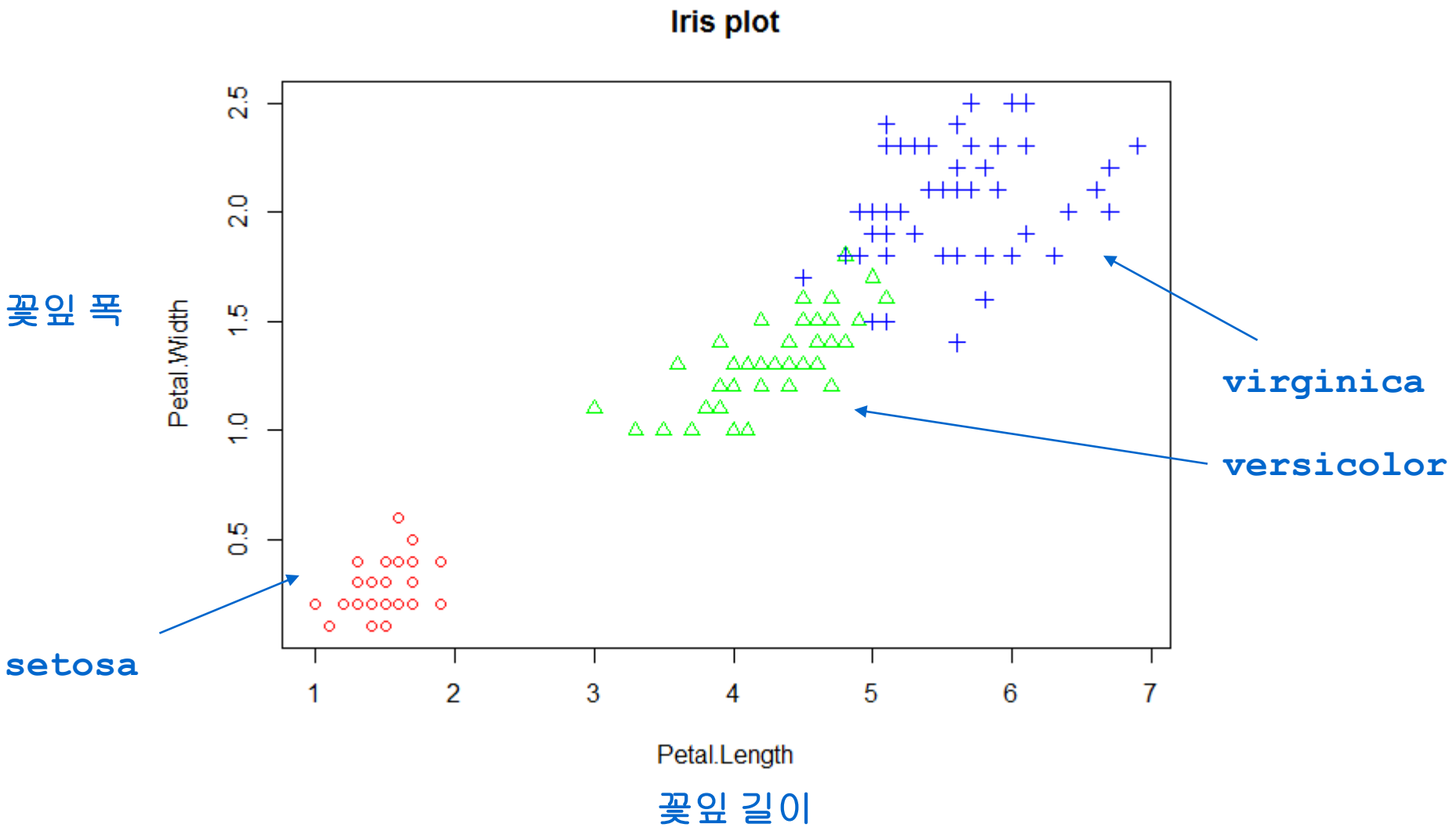
산점도(Scatter plot)

- 그룹 정보가 있는 2변량 데이터의 분포 보기
  - iris 데이터셋에서 Species 정보에 따른 Petal.Length, Petal.Width 의 분포를 알아 보자

```
iris.2 <- iris[,3:4]                    # 데이터
point <- as.numeric(iris$Species)       # 포인트 모양
color <- c("red","green","blue")        # 포인트 컬러
plot(iris.2,
      main="Iris plot",
      pch=c(point),
      col=color[iris[,5]]
)
```

```
color[iris[,5]]
  iris[,5] 는 species 정보.
  setosa :1, versicolor:2, virginica:3
  이렇게 값이 변환되어 사용됨
```

산점도(Scatter plot)



산점도(Scatter plot)

- Iris plot 을 보고 알아낼 수 있는 정보
  - 붓꽃(iris)은 꽃잎의 폭과 넓이 정보만 있으면 품종을 구별할 수 있다
  - Setosa 품종은 꽃잎의 폭과 넓이가 다른 두 종에 비해 매우 작다
  - virginica 품종은 꽃잎의 폭과 넓이가 가장 큰 품종이다.
  - virginica 품종과 versicolor 품종은 데이터가 겹치는 영역이 있어서 품종 구분이 정확히 안될수도 있다.

주어진 수치나 그래프로 부터 유용한 정보를 얻어내는 것이 데이터 분석의 목적임을 잊지 말자

[연습 1]

1. R에서 제공하는 cars 데이터셋을 이용해서 speed 와 dist 에 대한 산점도를 그리시오 (x축이 speed). speed 와 dist (제동거리)에 대한 상관 관계를 설명해 보시오

2. R에서 제공하는 pressure 데이터셋을 이용해서 temperature 와 pressure 에 대한 산점도를 그리시오 (x축이 temperature). 두 변수간 상관 관계를 설명해 보시오

3. R에서 제공하는 state.x77 데이터셋에서 Population, Income, Illiteracy, Area 변수간 산점도를 그려 상관관계를 관찰하시오 (pairs() 함수 이용)

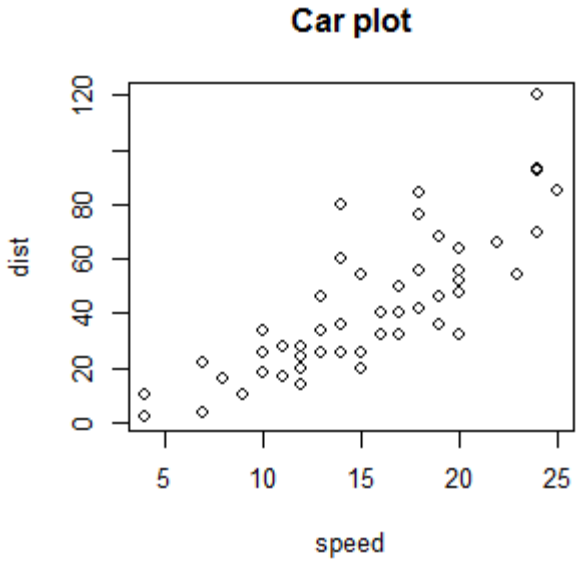
4. iris 데이터셋에서 Species 정보에 따른 Sepal.Length, Sepal.Width (꽃받침의 길이, 폭)의 분포를 알아 보시오

R에서 제공하는 데이터셋에 대한 설명을 보고 싶으면 **`help()`** 함수 이용
예) **`help(cars)`** 또는 Rstudio 의 Help 탭에서 cars 검색

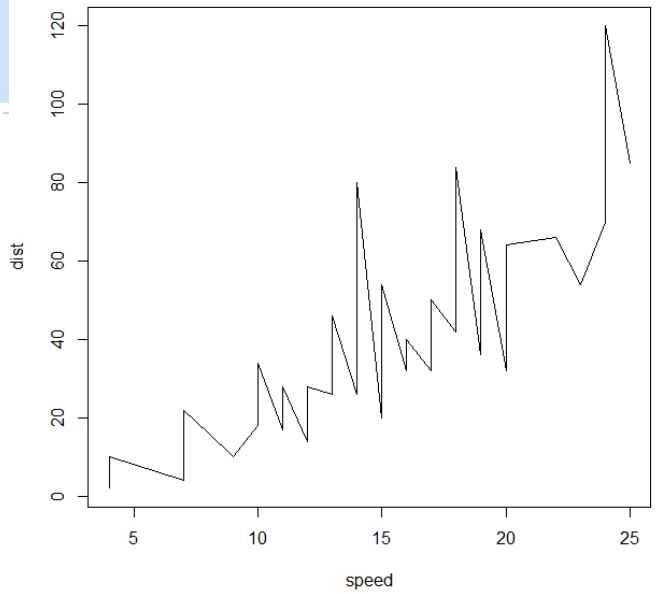
선 그래프

- 산점도를 작성하는 plot() 함수를 이용하여 선 그래프를 그릴 수 있다.

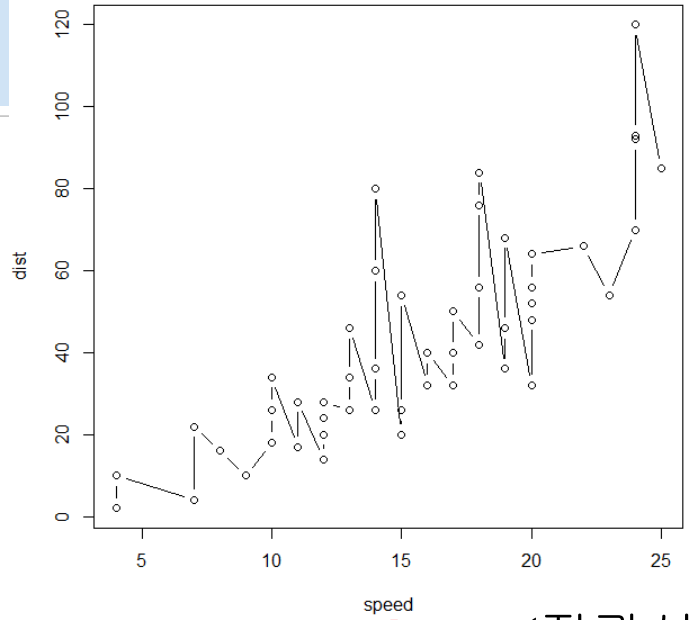
```
plot(cars,
       main="Car plot",
       pch=c(point),
       type="p"                  # 그래프의 종류 선택.
)
```



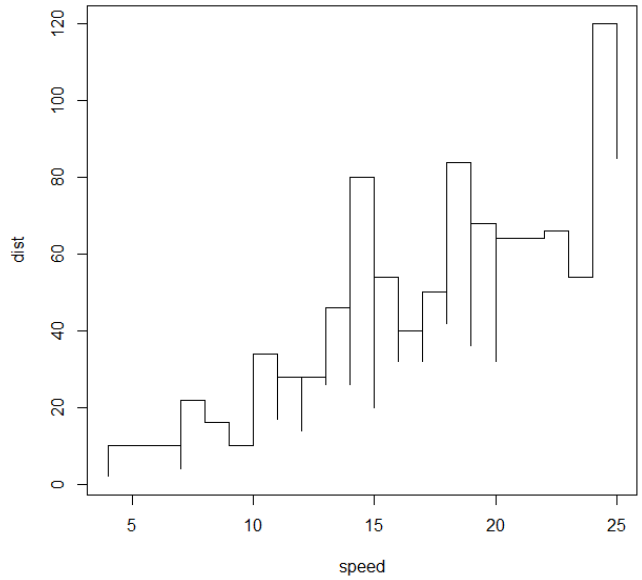
type="p" (점)



type= "l" (선)



type= "b" (점과선)



type= "s" (계단형)

type= "o" (점 위의선)

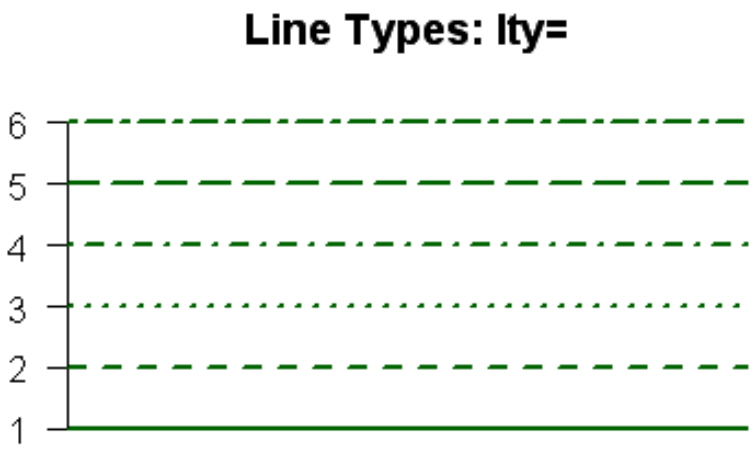
type= "h" (값에 해당하는 수직선)

type= "S" (계단형 2)

선 그래프

- 선그래프에서 선의 종류 지정

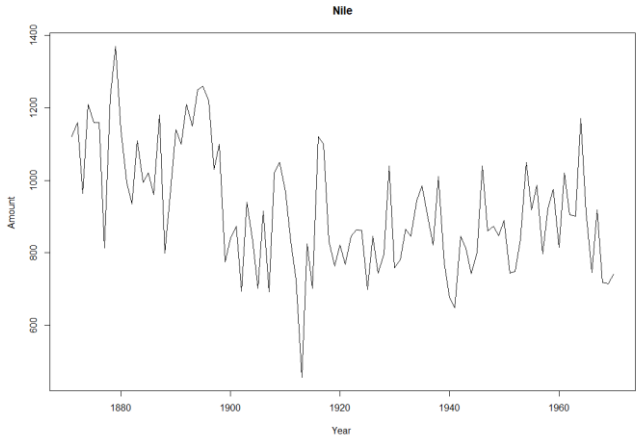
```
plot(cars,
      main="Car plot",
      pch=c(point),
      type= "l",                # 그래프의 종류 선택.
      lty=1,                    # 선의 종류(line type) 선택
      lwd=1                     # 선의 굵기 선택
)
```



선 그래프

- 1871~1970 나일강 범람 데이터

```
plot(1871:1970,                  # x data
       as.numeric(Nile),         # y data
       main="Nile",
       type= "l",                # 그래프의 종류 선택.
       lty=1,                    # 선의 종류(line type) 선택
       lwd=1,                    # 선의 굵기 선택
       xlab="Year ",             # x축 레이블
       ylab="Amount "            # y축 레이블
)
```



[연습 2]

1. R에서 제공하는 pressure 데이터셋을 이용해서 temperature 와 pressure에 대한 선 그래프를 그리시오 (x축이 temperature). 선의 종류와 색을 다양하게 바꾸어 작성해본다

2. R에서 제공하는 state.x77 데이터셋에서 Income, Illiteracy 변수를 가지고 선 그래프를 그려 보시오

3. R에서 제공하는 lynx 데이터셋을 이용하여 1821~1934의 lynx 선 그래프를 그려 보시오

lynx : 살쾡이류의 동물

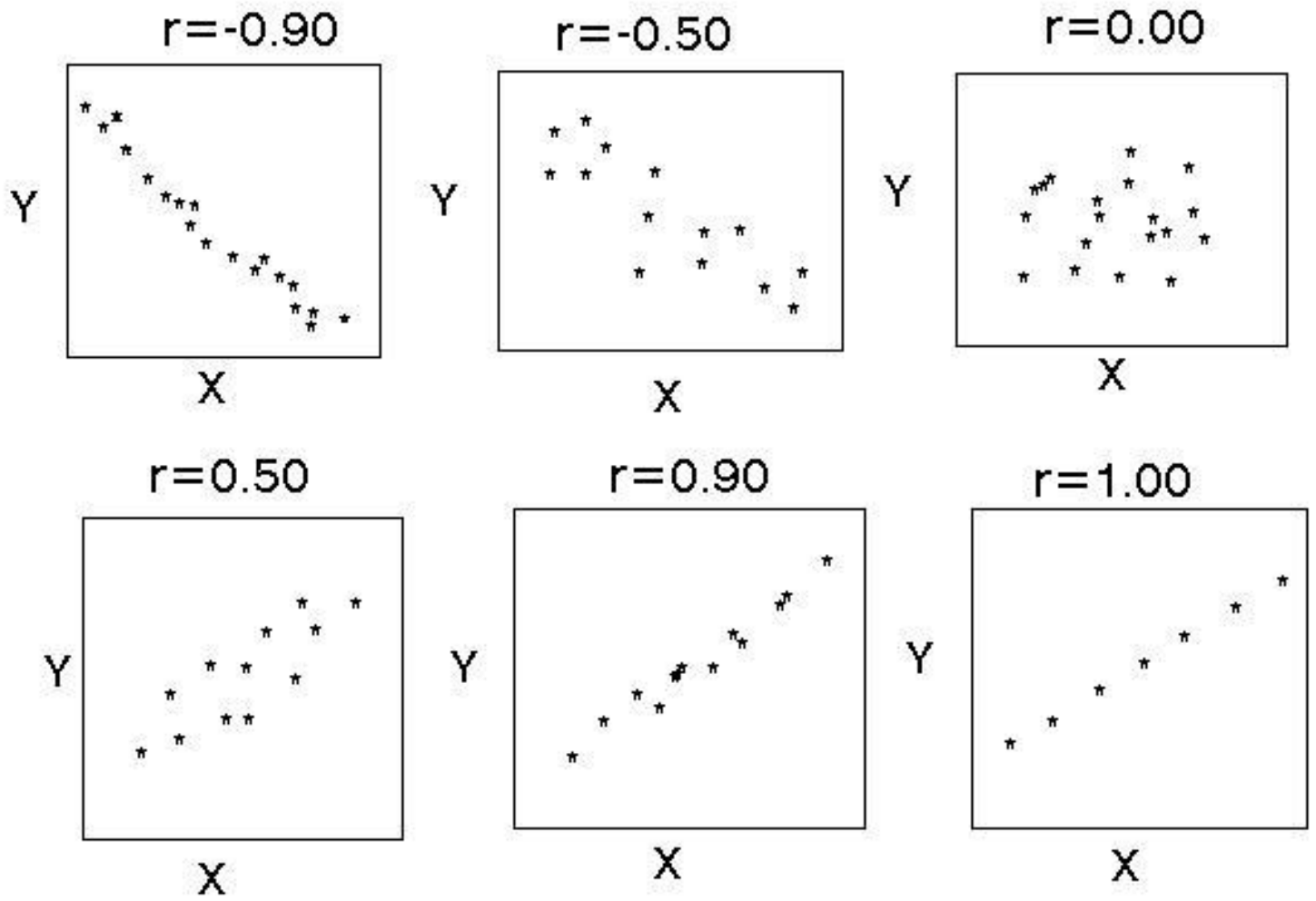
상관분석(Correlation Analysis)

- 두 변수 X와 Y 간의 선형성의 정도를 측정하는 통계량으로 다음과 같이 정의 됨

$$r = \frac{\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

- 일반적으로
  - $-1 \le r \le 1$
  - $r > 0$: 양의 상관 관계
  - $r < 0$: 음의 상관관계
  - 1 이나 -1 에 가까울수록 상관성이 높다

상관분석(Correlation Analysis)



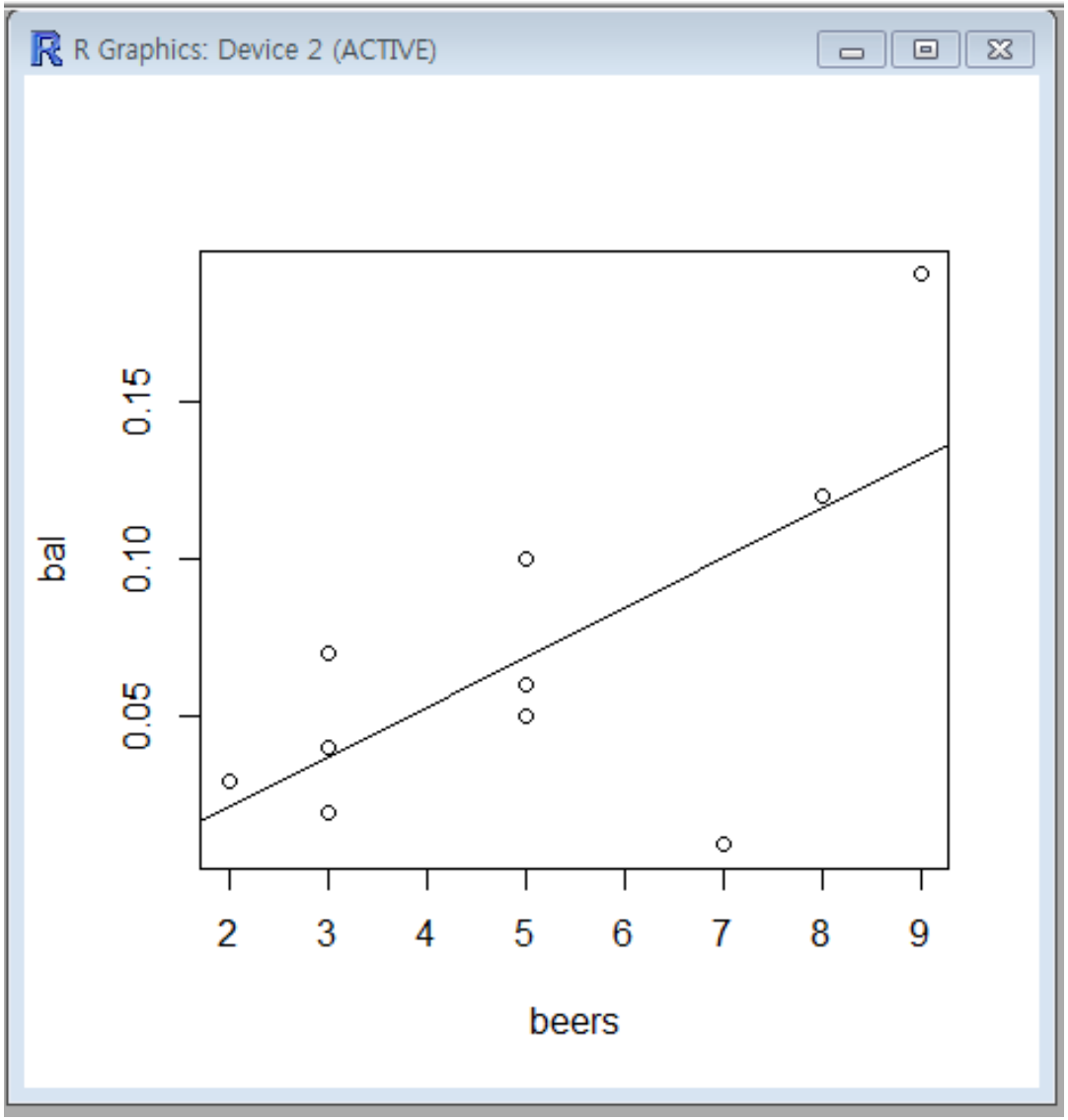
R 실습 : cor()

- 음주 정도와 혈중 알코올 농도의 상관도 분석

Beers	5	2	9	8	3	7	3	5	3	5
BAL	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06	0.02	0.05

```
beers = c(5,2,9,8,3,7,3,5,3,5)
bal = c(0.1,0.03,0.19,0.12,0.04,0.0095,0.07,
        0.06,0.02,0.05)
tbl = data.frame(cbind(beers,bal))
tbl; class(tbl)
plot(bal~beers,data=tbl)     # 산점도
res=lm(bal~beers,data=tbl)   # 회귀식 도출
abline(res)                  # 회귀선그리기
cor(beers,bal)               # 상관성 분석 시행
```

R 실습 : cor()




```
> cor(beers,bal)
[1] 0.6797025
```

R 실습 : cor()

```
tbl = data.frame(cbind(beers,bal))
```

- data.frame : 데이터를 테이블 형태로 관리
- cbind() : 두 벡터를 컬럼(열) 방향으로 합친다
  (cf. rbind() : 두 벡터를 행 방향으로 합친다)

```
plot(bal~beers,data=tbl)      # 산점도 (beers 이 x축)
```

- 두 벡터 데이터를 가지고 산점도를 그린다.
- **plot(tbl), plot(tbl[,1],tbl[,2])** 도 동일한 결과 도출

```
res=lm(bal~beers,data=tbl)
```

- 산점도를 가장 잘 표현할 수 있는 선형 모델(회귀식)을 구한다.
  (회귀모델에 대해서는 나중에 자세히 배우기로 한다)

R 실습 : cor()

```
abline(res)                    # 선그리기
```

- 구한 선형모델을 가지고 산점도 위에 선을 그린다

```
cor(beers,bal)                 # 상관성 분석 시행
```

- 두 벡터자료로 부터 상관계수를 계산한다.

[연습 3]

1. 다음은 10명의 수입과 교육받은 기간을 조사한 표이다. 수입과 교육기간 사이에 어느정도 상관관계가 있는지 조사하시오 (산점도, 상관계수 구하기)

Income	Years of Education
125,000	19
100,000	20
40,000	16
35,000	16
41,000	18
29,000	12
35,000	14
24,000	12
50,000	16
60,000	17

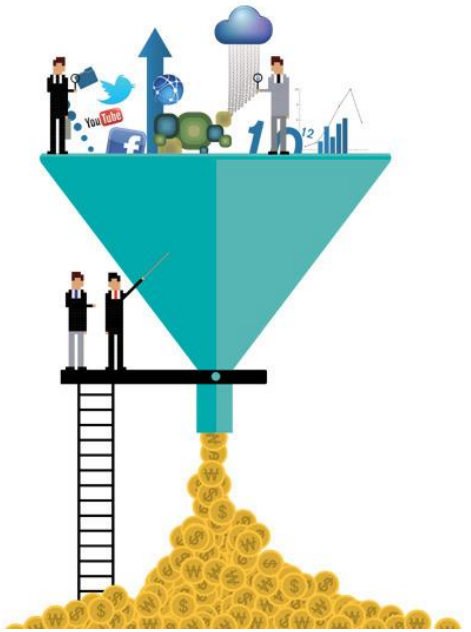
[연습 3]

2. 다음은 학생 10명의 성적과 TV 시청시간을 조사한 표이다. 성적과 TV시청시간 사이의 상관관계를 조사하시오. (산점도, 상관계수 구하기)

GPA	TV in hours per week
3.1	14
2.4	10
2.0	20
3.8	7
2.2	25
3.4	9
2.9	15
3.2	13
3.7	4
3.5	21

[연습 3]

3. R에서 제공하는 mtcars 데이터셋에서 mpg 와 다른 변수들 간의 상관 계수를 구하시오. 어느 변수가 mpg 와 가장 상관성이 높은지 산점도와 함께 제시하시오.



http://news.mk.co.kr/newsRead.php?year=2012&no=614917

데이터분석의 실제: iris

- Step.1 데이터셋 일반 정보

```
class(iris)
head(iris)
dim(iris)
table(iris$Species)
```

```
> class(iris)
[1] "data.frame"
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
> dim(iris)
[1] 150   5
> table(iris$Species)

    setosa versicolor  virginica
        50         50         50
```

데이터분석의 실제: iris

<해석>

- 이 데이터셋의 자료구조는 data frame (열 선택시 $ 사용 가능)
- 총 5개의 열(변수) 을 포함. 앞의 4개는 수치 데이터. 5번째는 각 행에 대한 그룹 정보 포함
- 이 데이터셋은 총 150 개의 행과 5개의 열로 구성
- 각 행들은 3개의 그룹중 하나 (setosa, versicolor, virginica)
- 각 그룹에 속한 행의 개수는 각각 50개씩 균등하다.

Note 1. 단순히 어떤 그룹이 있는지만 알아 보려면

```
> unique(iris$Species)
[1] setosa     versicolor virginica
Levels: setosa versicolor virginica
```

위와 같이 Level 정보가 표시되면 Specis 열은 타입이 factor 임

Note 2. 자료구조가 data frame 이면 열 데이터를 추출할때 iris$Species 가 가능. 만일 iris 가 matrix 이면 iris[,5], iris[,"Species"] 처럼 해야 한다

데이터분석의 실제: iris

- Step.2 4개 열 데이터에 대한 데이터 분포 확인

```
summary(iris[,1])
summary(iris[,2])
summary(iris[,"Petal.Length"])
summary(iris$Petal.Width)
```

```
> summary(iris[,1])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
> summary(iris[,2])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.800   3.000   3.057   3.300   4.400
> summary(iris[,"Petal.Length"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.600   4.350   3.758   5.100   6.900
> summary(iris$Petal.Width)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   0.300   1.300   1.199   1.800   2.500
```

데이터분석의 실제: iris

```
sd(iris[,1])   # Sepal.Length
sd(iris[,2])   # Sepal.Width
sd(iris[,3])   # Petal.Length
sd(iris[,4])   # Petal.Length
```

```
> sd(iris[,1])
[1] 0.8280661
> sd(iris[,2])
[1] 0.4358663
> sd(iris[,3])
[1] 1.765298
> sd(iris[,4])
[1] 0.7622377
```

<해석>

- Sepal.Width 는 데이터의 편차가 작고, Petal.Length 는 편차가 크다

데이터분석의 실제: iris

- Step 3. 각 열 데이터에 대해 그룹별 분포를 확인

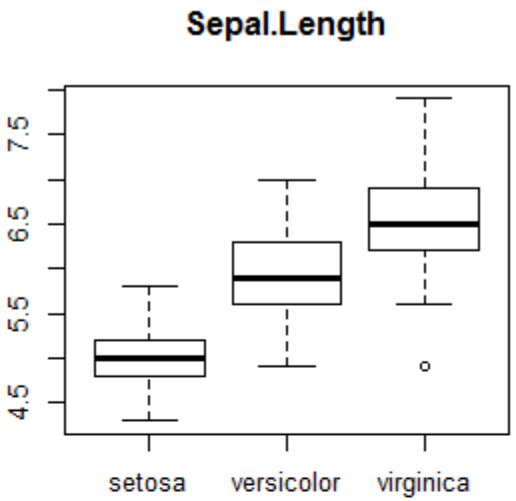
```
par(mfrow = c(2, 2))
boxplot(Sepal.Length~Species, data = iris,
        main = "Sepal.Length")
boxplot(Sepal.Width~Species, data = iris,
        main = "Sepal.Width")
boxplot(Petal.Length~Species, data = iris,
        main = "Petal.Length")
boxplot(Petal.Length~Species, data = iris,
        main = "Petal.Width")
```

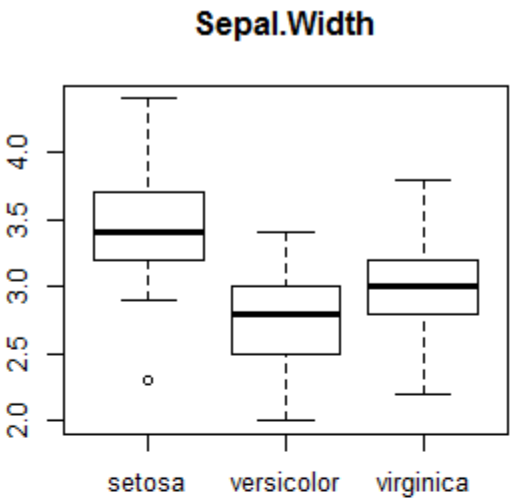
Sepal.Length~Species, data = iris

Iris 데이터셋의 **Sepal.Length**에 대해 boxplot 을 그리되 **Species** 에 따라 그룹을 구분하여 그리시오

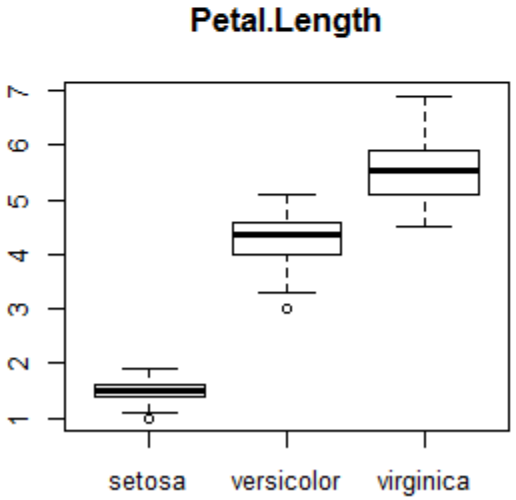
데이터분석의 실제: iris

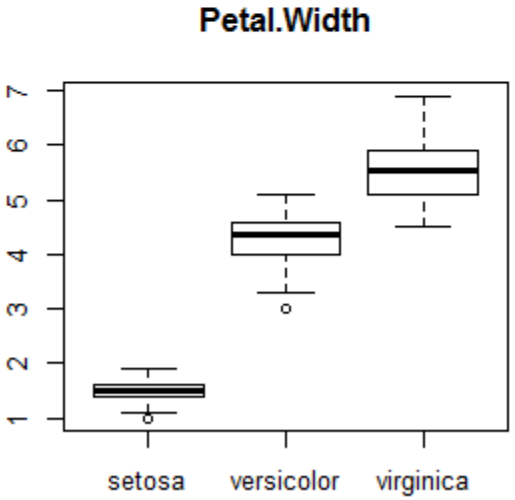
꽃받침





꽃잎





데이터분석의 실제: iris

<해석>

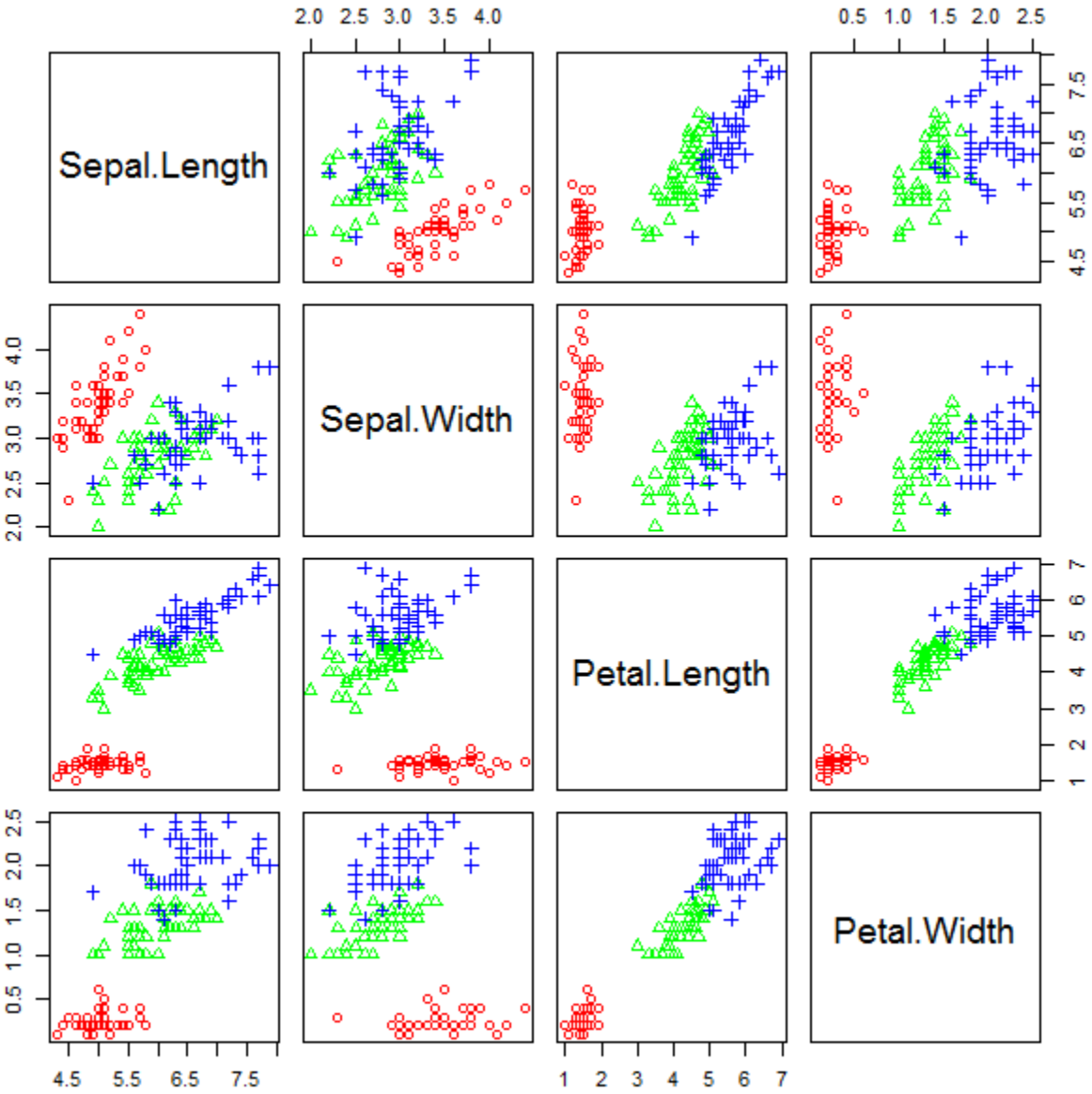
- 4개 변수에서 각 그룹간 데이터의 크기는 차이가 존재한다
- Sepal.width 와 Sepal.length 에서는 그룹간 데이터가 겹치는 부분이 넓다
- Setosa 품종의 경우는 petal.width 와 petal.length 에서 데이터의 편차가 매우 적다
- 이상치에 속하는 데이터가 일부 있다

데이터분석의 실제: iris

- Step 4. 각 열 데이터에 대해 그룹별 분포를 산점도를 통해 확인

```
point <- as.numeric(iris$Species) # 포인트 모양
color <- c("red","green","blue")  # 포인트 컬러
pairs(iris[,-5],
      pch=c(point),
      col=color[iris[,5]]
)
```

데이터분석의 실제: iris



데이터분석의 실제: iris

<해석>

- 4개 변수에서 각 그룹간 데이터의 크기는 차이가 존재한다
- Sepal.length 와 petal.length, 그리고 petal.length 와 petal.width 는 강한 양의 상관 관계를 보인다.

```
> cor(iris$Sepal.Length, iris$Petal.Length)
[1] 0.8717538
> cor(iris$Petal.Width, iris$Petal.Length)
[1] 0.9628654
```

[연습 4]

- 공공데이터 포털 사이트에서 2016년도 경륜 매출액 데이터를 다운 받아 분석해 보시오 (https://www.data.go.kr/dataset/15004119/fileData.do)