# 제5강
# 단일변수 자료의 탐색

# Section 01

# 자료의 종류

# Section 02

# 단일변수 범주형 자료의 탐색

# Section 03

# 단일변수 연속형 자료의 탐색

# 3. 단일변수 연속형 자료의 탐색

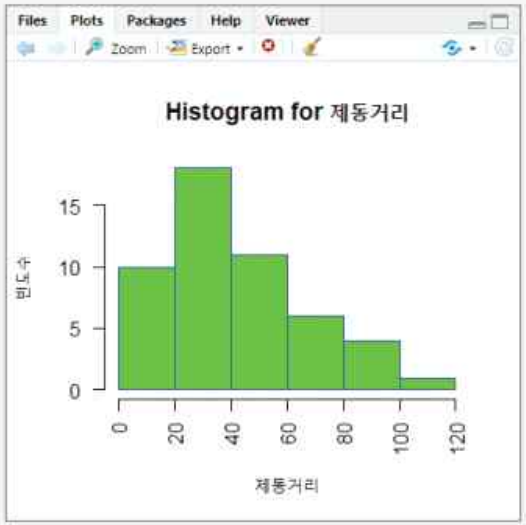
## 4. 히스토그램

- 히스토그램(histogram)은 외관상 막대그래프와 비슷한 그래프로, 연속형 자료의 분포를 시각화 할 때 사용
- 막대그래프를 그리려면 값의 종류별로 개수를 셀 수 있어야 하는데, 키와 몸무게 등의 자료는 값의 종류라는 개념이 없어서 종류별로 개수를 셀 수 없음
- 대신에 연속형 자료에서는 구간을 나누고 구간에 속하는 값들의 개수를 세는 방법을 사용

코드 5-8

```
dist <- cars[,2]                                  # 자동차 제동거리
hist(       dist,                                 # 자료(data)
            main="Histogram for 제동거리",        # 제목
            xlab ="제동거리",                     # x축 레이블
            ylab="빈도수",                        # y축 레이블
            border="blue",                        # 막대 테두리색
            col="green",                          # 막대 색
            las=2,                                # x축 글씨 방향(0~3)
            breaks=5)                             # 막대 개수 조절
```

# 3. 단일변수 연속형 자료의 탐색




**hist()함수**
**las : x축에 표시되는 값들(0,20,40…120)의 출력방향을 조절하는 매개변수이다. 2일때는 글씨가 세로방향으로 출력됨**
**breaks : 값이 커지면 구간의 개수도 늘어나고, 값이 작아지면 구간의 개수도 줄어든다.**

**위와 같이 히스토그램은 관측값들이 어느 구간에 분포하는지를 쉽게 파악해 줄 수 있도록 해준다.**

## 여기서 잠깐! 막대그래프와 히스토그램 비교

히스토그램은 외관상 막대그래프와 유사

일반적으로 막대 사이에 간격 있으면 막대그래프, 간격 없이 막대들이 붙어 있으면 히스토그램

막대그래프에서는 막대의 면적이 의미가 없지만 히스토그램에서는 막대의 면적도 의미가 있음

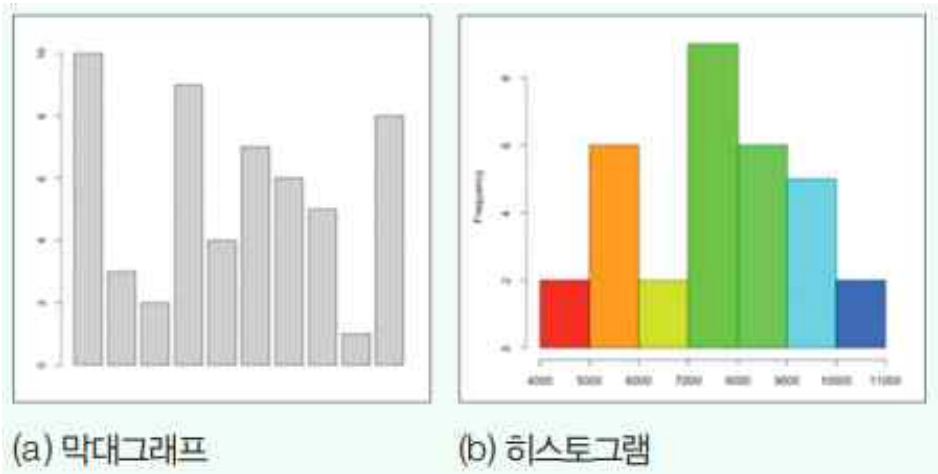
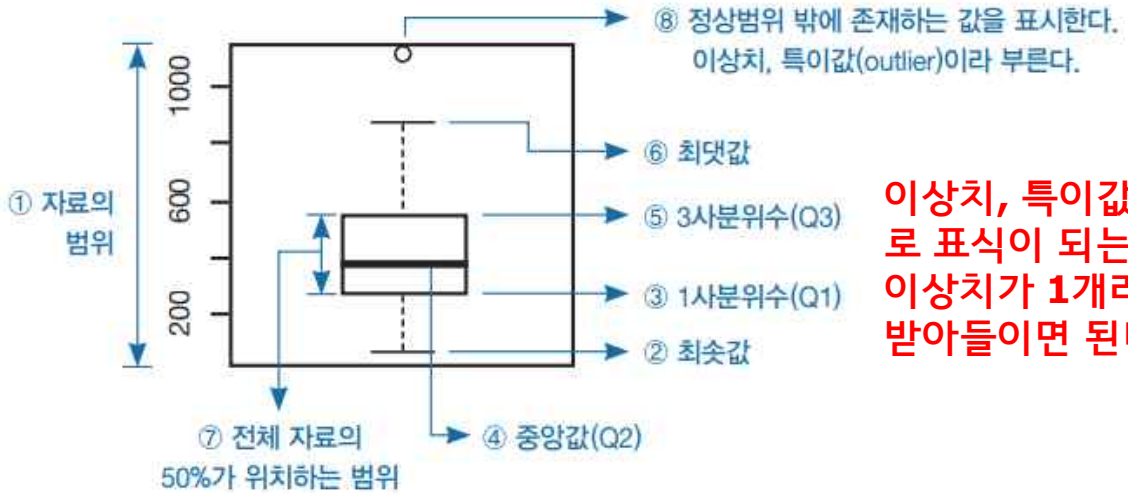



그림 5-8 막대그래프와 히스토그램

# 3. 단일변수 연속형 자료의 탐색

## 5. 상자그림

- **상자그림(box plot)은 상자 수염 그림(box and whisker plot)으로도 부르며, 사분위수를 시각화하여 그래프 형태로 나타낸 것**
- 하나의 그래프로 데이터의 분포 형태를 포함한 다양한 정보를 전달하기 때문에 단일변수 수치형 자료를 파악하는 데 자주 사용




**이상치, 특이값은 동그라미 하나로 표식이 되는데 1개가 있으면, 이상치가 1개라는 의미로 받아들이면 된다.**

그림 5-9 상자그림의 구성 요소

# 3. 단일변수 연속형 자료의 탐색

코드 5-9

```
dist <- cars[,2]                                    # 자동차 제동거리(단위: 피트)
boxplot(dist, main="자동차 제동거리")
```

```
> dist <- cars[,2]                    # 자동차 제동거리(단위: 피트)
> boxplot(dist, main="자동차 제동거리")
```

# 3. 단일변수 연속형 자료의 탐색

코드 5-10

```
boxplot.stats(dist)
```

```
> boxplot.stats(dist)
$stats
[1]  2 26 36 56 93

$n
[1] 50

$conf
[1] 29.29663 42.70337

$out
[1] 120
```

**boxplot()함수를 통해서 그린 박스상자 그림을 구체적으로 수치를 알고 싶다면, boxplot.stats()함수를 이용하면 된다.**

**좌측 $stats는 최솟값, 1사분위수, 중앙값, 3사분위수, 최댓값을 의미한다.**

**$n은 관측값의 개수를 의미한다.**

**$conf는 중앙값에 대한 신뢰구간을 의미한다.**

**$out는 특이값(이상치)에 대한 목록을 나타낸다.**

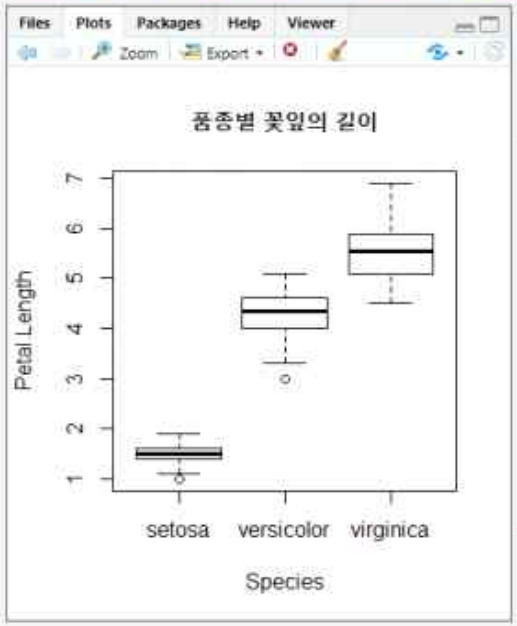
# 3. 단일변수 연속형 자료의 탐색

## 6. 그룹이 있는 자료의 상자그림

코드 5-11

```
boxplot(Petal.Length~Species, data=iris, main="품종별 꽃잎의 길이")
```

```
> boxplot(Petal.Length~Species, data=iris, main="품종별 꽃잎의 길이")
```




**boxplot()함수의 매개변수 Petal.length~Species는 꽃잎의 길이(Petal.length) 자료를 품종(Species)별로 나누어 상자그림을 그리라는 의미이며, 반드시 Petal.length와 Species가 포함되어 있는 변수가 따라와야 하는데, 여기서는 data=iris가 위의 변수를 포함한 데이터셋이라고 알려주는 것이다.**

**좌측 그림을 보면 관찰할 수 있는 사실은 setosa품종의 꽃잎의 길이가 가장 작고, virginica품종에 대한 꽃잎의 길이가 전반적으로 가장 크다는 사실을 알 수 있으며, setosa품종은 꽃잎의 길이가 비슷하다는 사실을 모여있으니깐 알수 있고 나머지 2품종은 꽃잎의 길이가 퍼져 있다는 것을 알 수가 있다.**

## 여기서 잠깐! 한 화면에 그래프 여러 개 출력하기

```
par(mfrow=c(1,3))            # 1x3 가상화면 분할
barplot(table(mtcars$carb),
        main="Barplot of Carburetors",
        xlab="#of carburetors",
        ylab="frequency",
        col="blue")
barplot(table(mtcars$cyl),
        main="Barplot of Cylender",
        xlab="#of cylender",
        ylab="frequency",
        col="red")
barplot(table(mtcars$gear),
        main="Barplot of Grar",
        xlab="#of gears",
        ylab="frequency",
        col="green")
par(mfrow=c(1,1))            # 가상화면 분할 해제
```

**par()함수로 화면을 분할을 할 수가 있는데, 매개변수로 mfrow의 값이 1,3인 것은 1행 3열을 의미하여 1개의 행으로 3개의 막대그래프를 나열하라는 의미가 된다.**

## 여기서 잠깐! 한 화면에 그래프 여러 개 출력하기

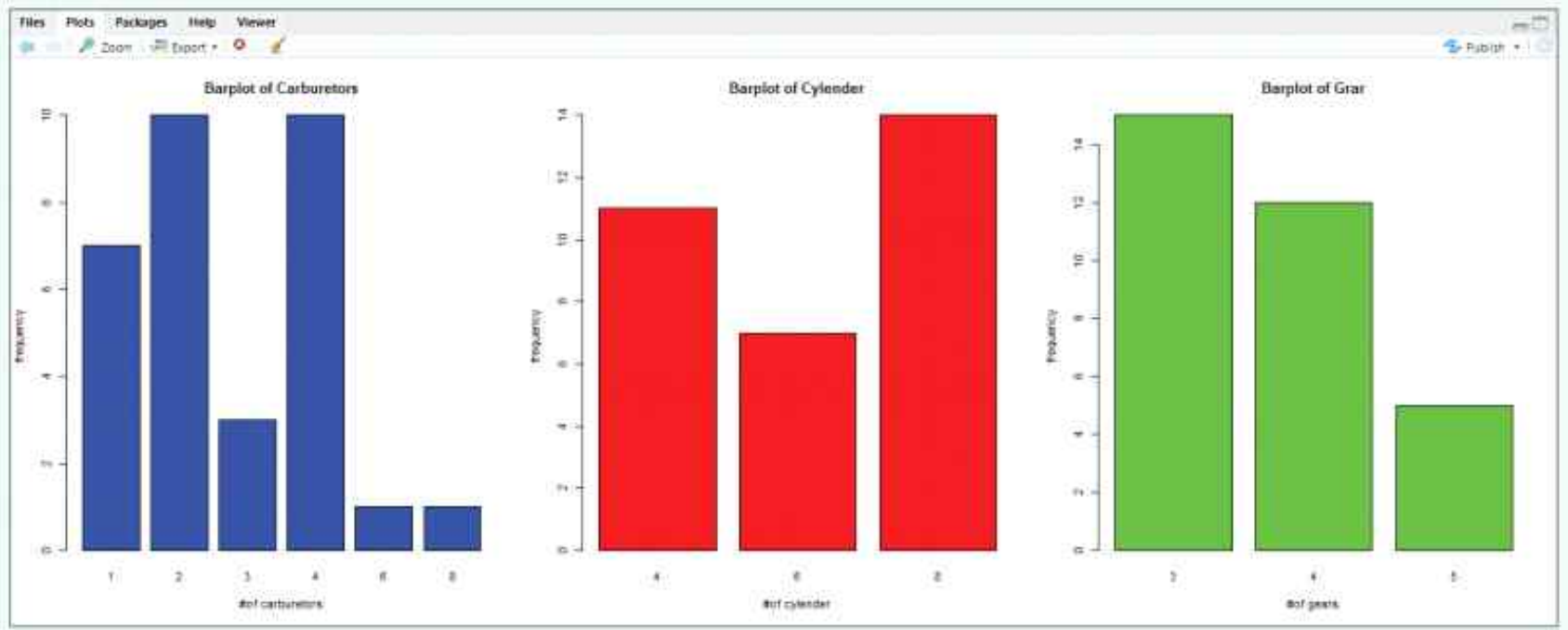



그림 5-11 한 화면에 여러 개의 그래프 출력

# 감사합니다.