Part 1. R 프로그래밍 (데이터 분석 전문가 양성과정)

02

내장 데이터셋

경북대학교 배준현 교수
(joonion@knu.ac.kr)

02. 내장 데이터셋

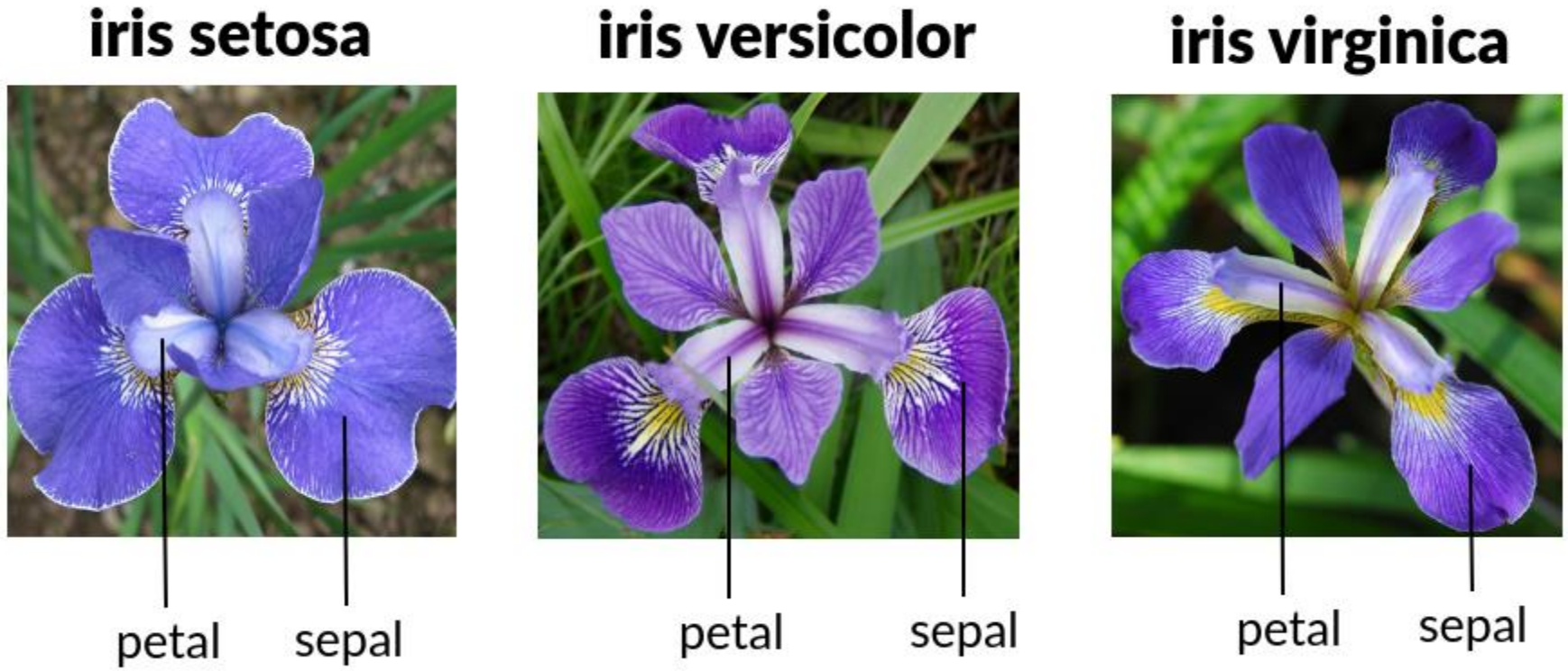
- 데이터셋: *dataset*
  - 통계분석에 필요한 데이터를 2차원 형태(*data frame*)로 만들어 놓은 데이터
  - R의 내장 데이터셋:
    - R에서 별도로 불러오지 않아도 기본적으로 사용할 수 있는 데이터셋
    - 외부 데이터셋: 데이터셋을 사용하기 위해 외부 패키지에서 불러와야 함
  - 두 개의 내장 데이터셋:
    - *iris*: 붓꽃의 품종에 관한 데이터셋
    - *mtcars*: 자동차 로드 테스트에 관한 데이터셋

02. 내장 데이터셋

- *iris*: Edgar Anderson's Iris Data
  - 1935년, 붓꽃의 품종 연구를 위해 수집한 꽃잎과 꽃받침의 길이와 너비 정보
  - 5개의 **변수**에 대한 150개의 **관측값**을 포함하고 있음

  - **변수** or **변량**: *variable* or *variate*
    - 관찰, 조사, 분석, 연구의 대상이 가지는 특성
  - **관측값**: *observation*
    - 분석 대상의 특성을 관찰하여 측정한 값

02. 내장 데이터셋



https://machinelearninghd.com/iris-dataset-uci-machine-learning-repository-project/

02. 내장 데이터셋

```
> ?iris
```

iris {datasets} R Documentation

Edgar Anderson's Iris Data

Description

This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

Usage

```
iris
iris3
```

Format

`iris` is a data frame with 150 cases (rows) and 5 variables (columns) named `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`.

`iris3` gives the same data arranged as a 3-dimensional array of size 50 by 4 by 3, as represented by S-PLUS. The first dimension gives the case number within the species subsample, the second the measurements with names `Sepal L.`, `Sepal W.`, `Petal L.`, and `Petal W.`, and the third the species.

Source

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, Part II, 179–188.

The data were collected by Anderson, Edgar (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5.

References

02. 내장 데이터셋

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

- Sepal.Length: **꽃받침의 길이**
- Sepal.Width: **꽃받침의 너비**
- Petal.Length: **꽃잎의 길이**
- Petal.Width: **꽃잎의 너비**
- Species: **붓꽃의 품종** (setosa, versicolor, virginica)

02. 내장 데이터셋

```
> 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
```

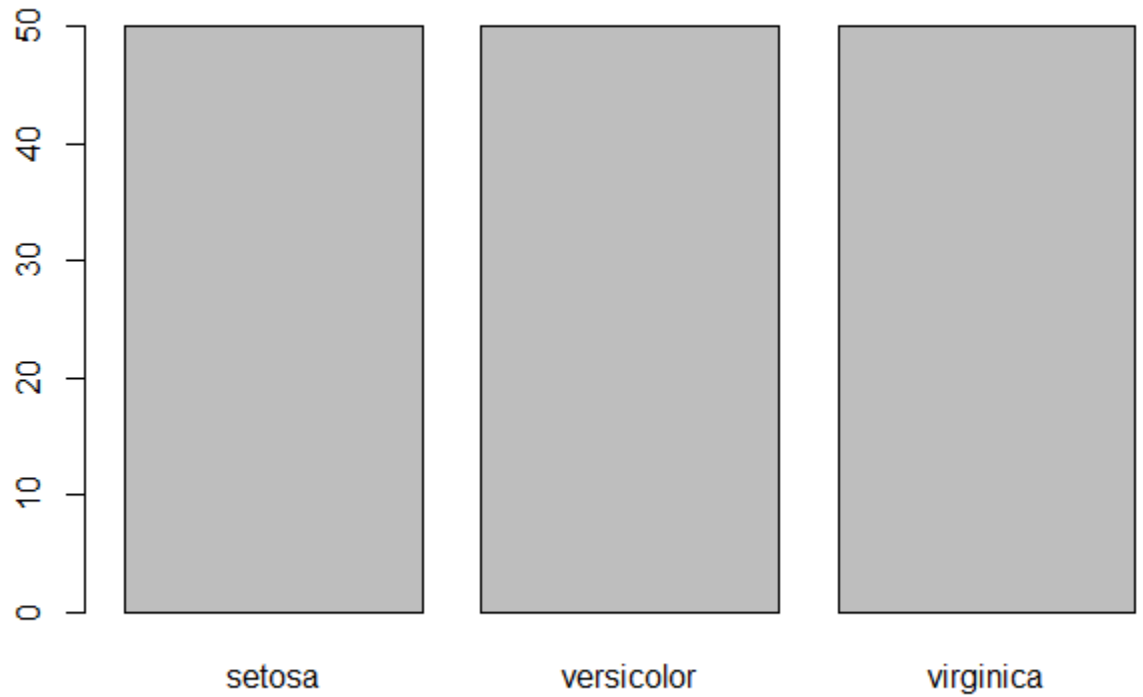
```
> tail(iris)
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
145          6.7         3.3          5.7         2.5 virginica
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

변수의 특성에 따른 분류

- **수치형** 자료: *numerical data*
  - 관측값이 크기를 나타내는 수치형으로 나타남
    - 예) **키**, **몸무게**
  - 대소비교와 산술연산이 가능하고, 평균, 표준편차 등의 특성을 가짐
  - **양적**(*quantitative*) 자료, **연속형**(*continuous*) 자료라고도 함
- **범주형** 자료: *categorical data*
  - 관측값이 그룹으로 구분할 수 있는 범주형으로 나타남
    - 예) **성별**, **혈액형**
  - 대소비교와 산술연산이 불가능하고, 빈도가 중요함
  - **질적**(*qualitative*) 자료, **명목형**(*nominal*) 자료라고도 함

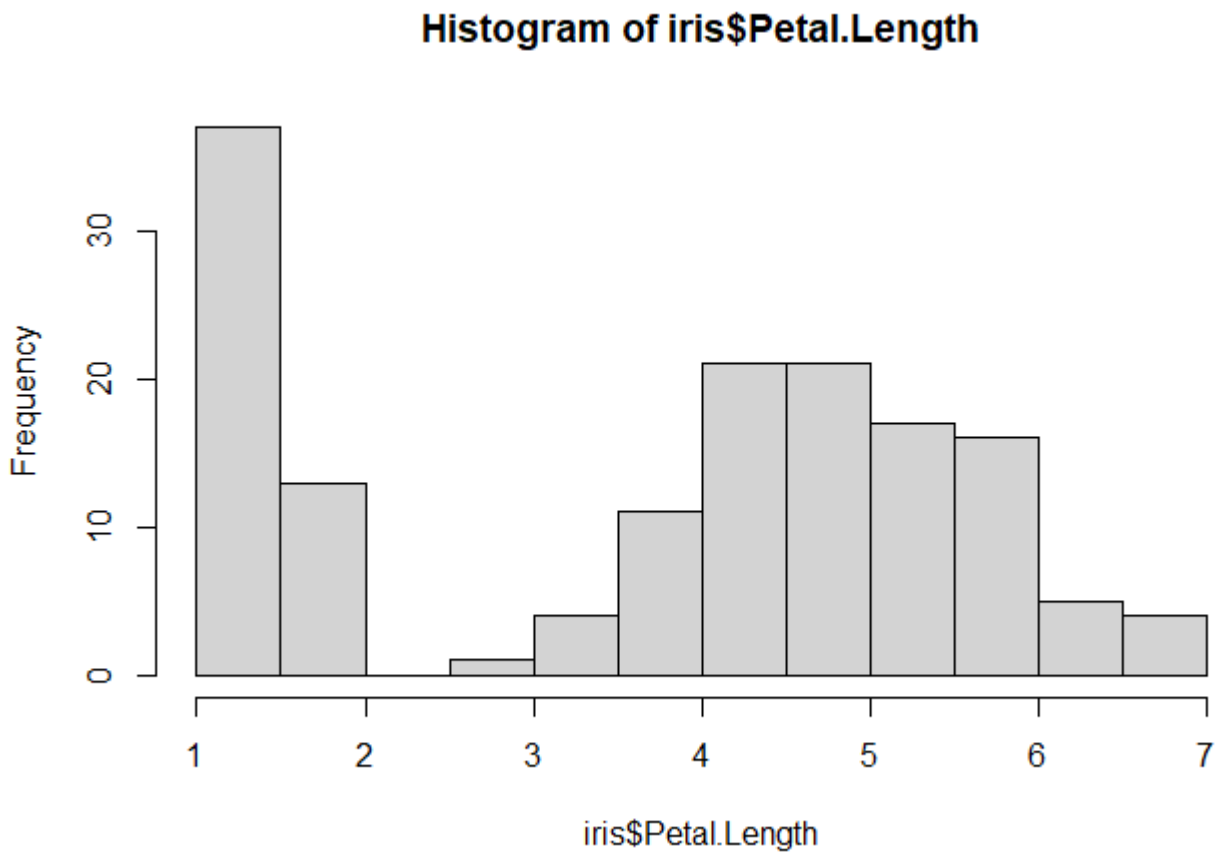
02. 내장 데이터셋

```
> class(iris$Species)
[1] "factor"

> levels(iris$Species)
[1] "setosa"     "versicolor" "virginica"

> table(iris$Species)

    setosa versicolor  virginica
        50         50         50

> barplot(table(iris$Species))
```



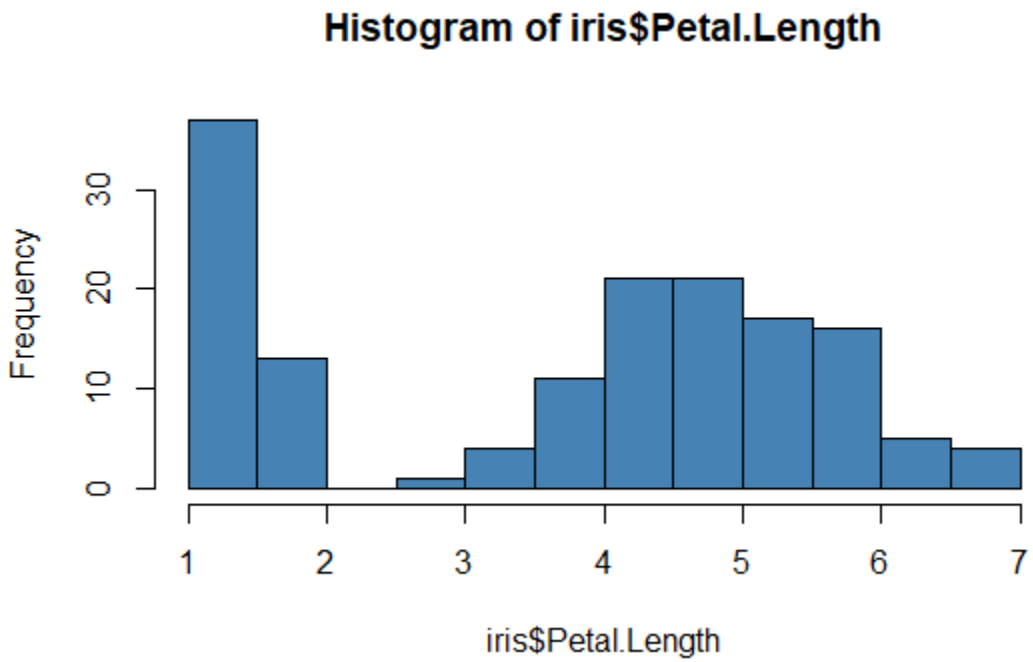
02. 내장 데이터셋

```
> class(iris$Petal.Length)
[1] "numeric"

> mean(iris$Petal.Length)
[1] 3.758

> var(iris$Petal.Length)
[1] 3.116278

> sd(iris$Petal.Length)
[1] 1.765298

> hist(iris$Petal.Length)
```

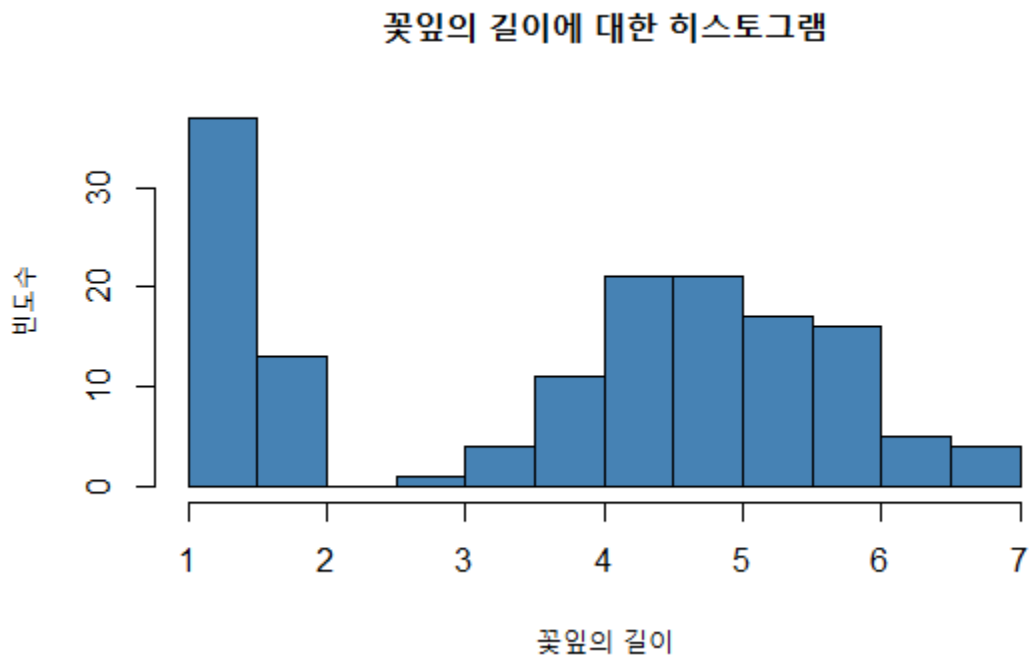


02. 내장 데이터셋

```
> hist(iris$Petal.Length, col = 'steelblue')
```



```r
hist(iris$Petal.Length, col = 'steelblue',
     main = '꽃잎의 길이에 대한 히스토그램',
     xlab = '꽃잎의 길이',
     ylab = '빈도수')
```



02. 내장 데이터셋

- *mtcars*: Motor Trend Car Road Tests
  - 1974년에 실시한 32개 자동차 모델의 연비에 대한 조사 결과
  - 11개의 변수에 대한 32개의 관측값을 포함하고 있음

02. 내장 데이터셋

```
> ?mtcars
```

mtcars {datasets} R Documentation

Motor Trend Car Road Tests

Description

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Usage

```
mtcars
```

Format

A data frame with 32 observations on 11 (numeric) variables.

[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)
[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

02. 내장 데이터셋

```
> str(mtcars)
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

[, 1]	mpg	Miles/(US) gallon (연비)
[, 2]	cyl	Number of cylinders (실린더 개수)
[, 3]	disp	Displacement (cu.in.) (배기량)
[, 4]	hp	Gross horsepower (마력)
[, 5]	drat	Rear axle ratio (후방 차축 비율)
[, 6]	wt	Weight (1000 lbs) (중량)
[, 7]	qsec	¼ mile time (1/4 마일 가는데 걸리는 시간)
[, 8]	vs	Engine (0 = V-shaped, 1 = straight) (엔진 모양)
[, 9]	am	Transmission (0 = automatic, 1 = manual) (변속기)
[,10]	gear	Number of forward gears (전진기어 개수)
[,11]	carb	Number of carburetors (기화기 개수)

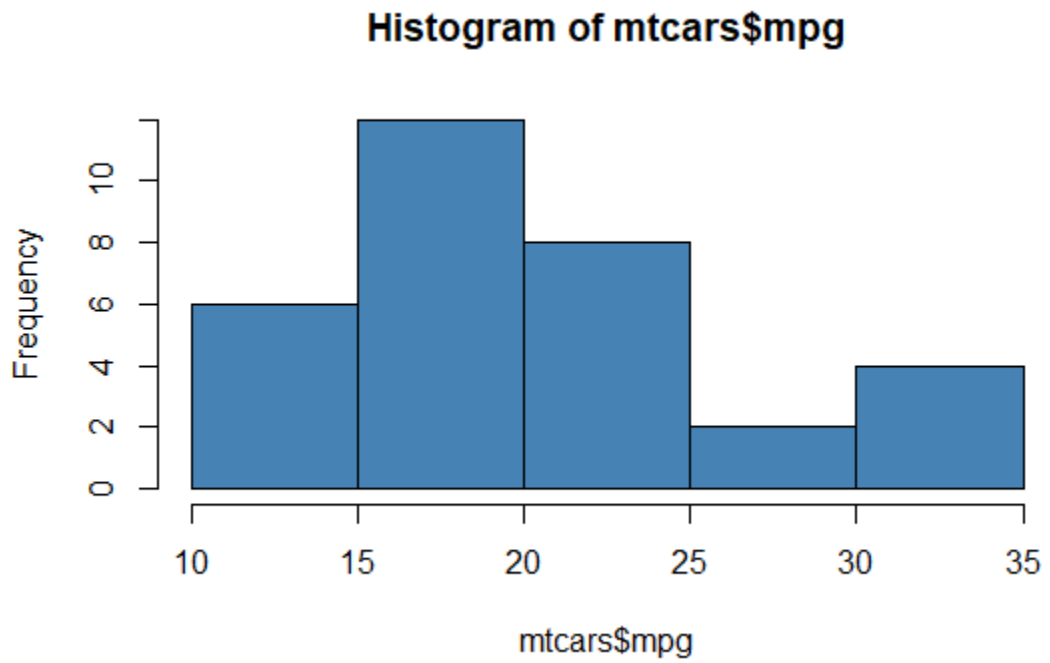
```
> mtcars$mpg
 [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
[16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
[31] 15.0 21.4

> mtcars$wt
 [1] 2.620 2.875 2.320 3.215 3.440 3.460 3.570 3.190 3.150 3.440 3.440 4.070
[13] 3.730 3.780 5.250 5.424 5.345 2.200 1.615 1.835 2.465 3.520 3.435 3.840
[25] 3.845 1.935 2.140 1.513 3.170 2.770 3.570 2.780
```

02. 내장 데이터셋

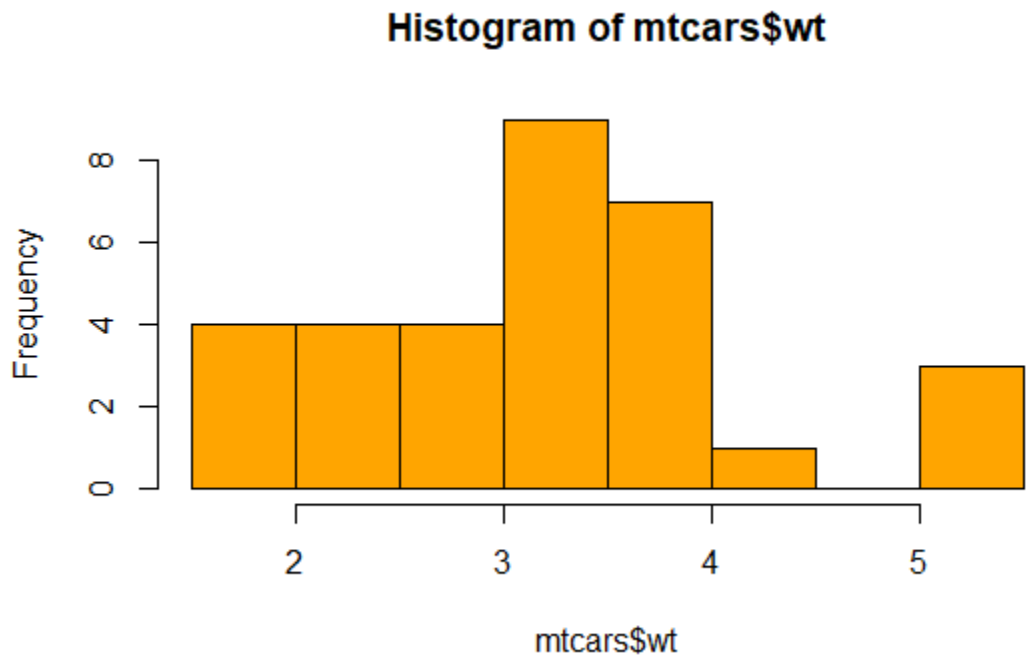
```
> summary(mtcars)
      mpg             cyl             disp             hp             drat       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930  
       wt             qsec             vs              am              gear      
 Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000  
 1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000  
 Median :3.325   Median :17.71   Median :0.0000   Median :0.0000   Median :4.000  
 Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688  
 3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000  
 Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000  
      carb      
 Min.   :1.000  
 1st Qu.:2.000  
 Median :2.000  
 Mean   :2.812  
 3rd Qu.:4.000
```

02. 내장 데이터셋

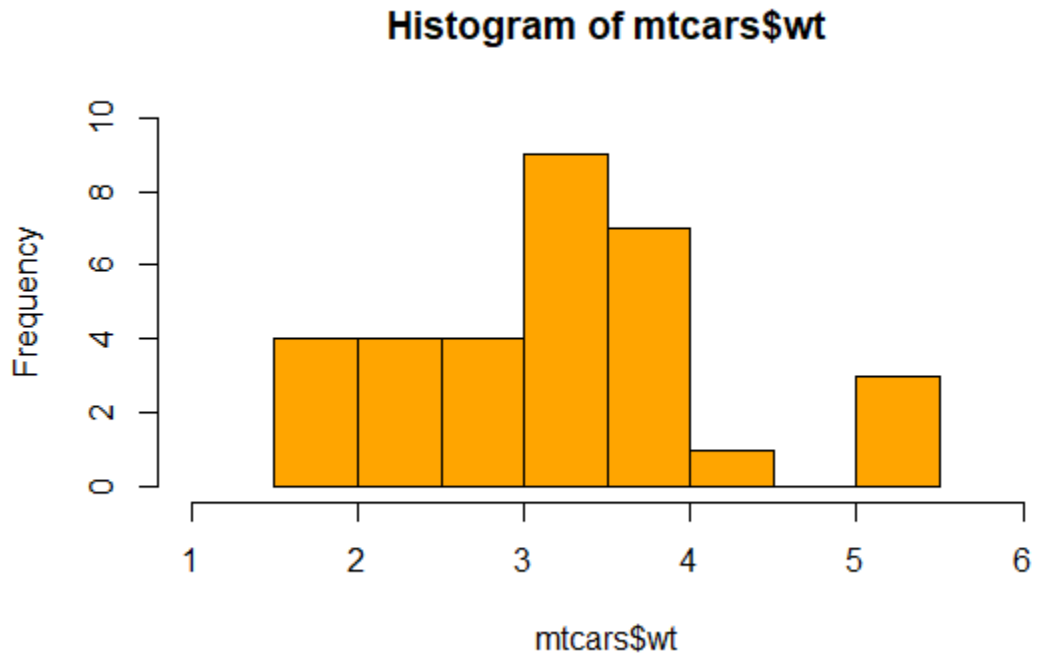
```
> hist(mtcars$mpg, col = 'steelblue')
```



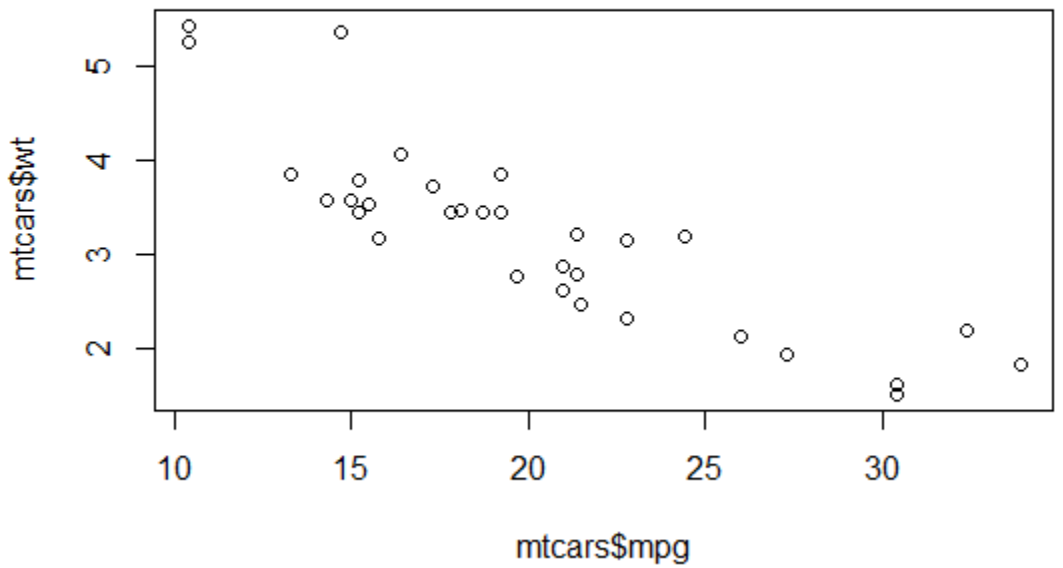
```
> hist(mtcars$wt, col = 'orange')
```



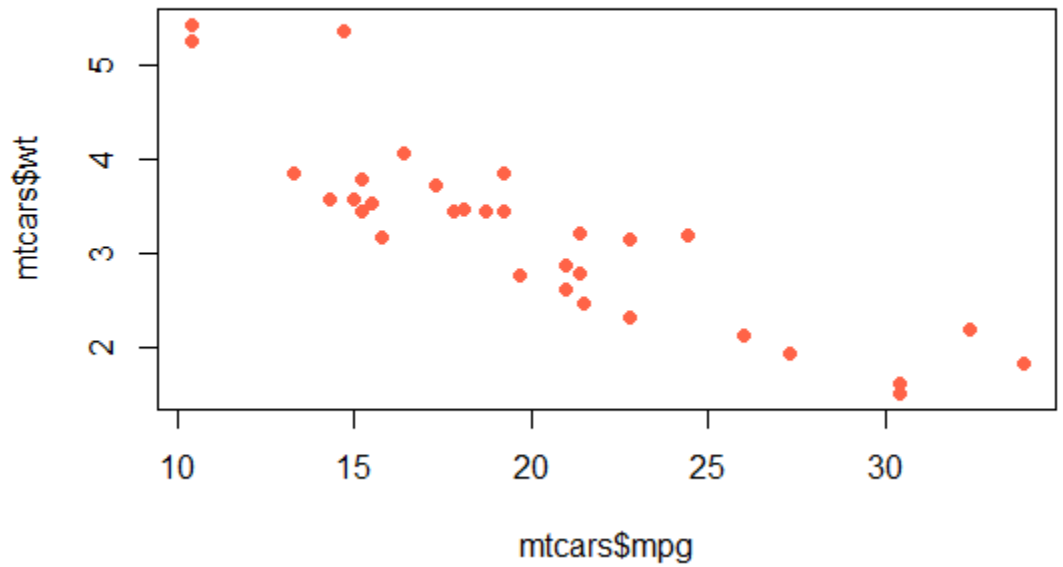
```
> hist(mtcars$wt, col = 'orange',
       xlim = c(1, 6),
       ylim = c(0, 10))
```



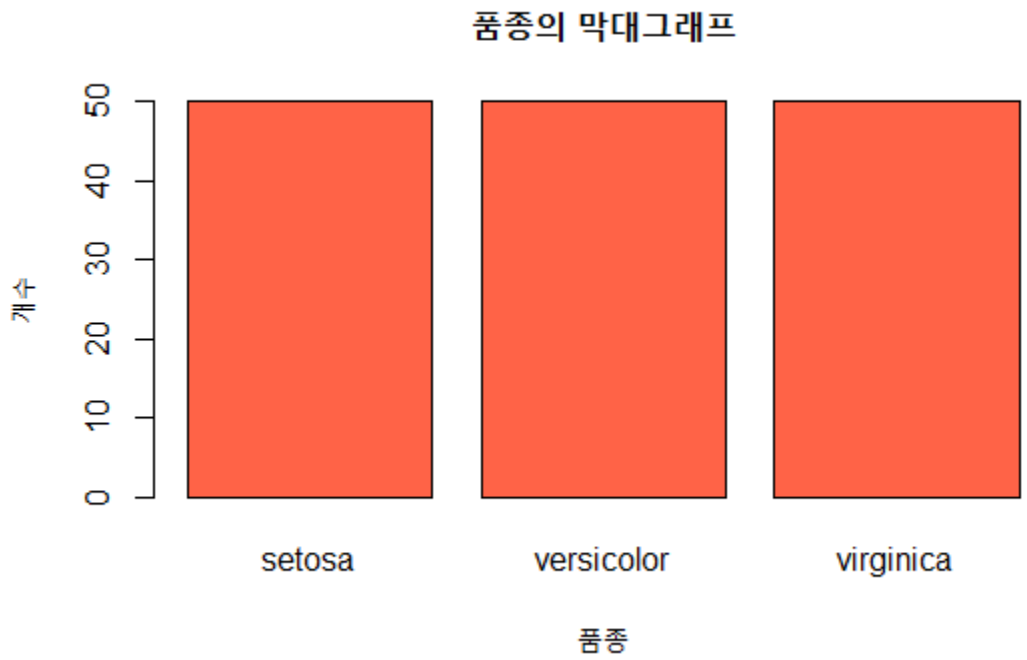
```
> plot(mtcars$mpg, mtcars$wt)
```



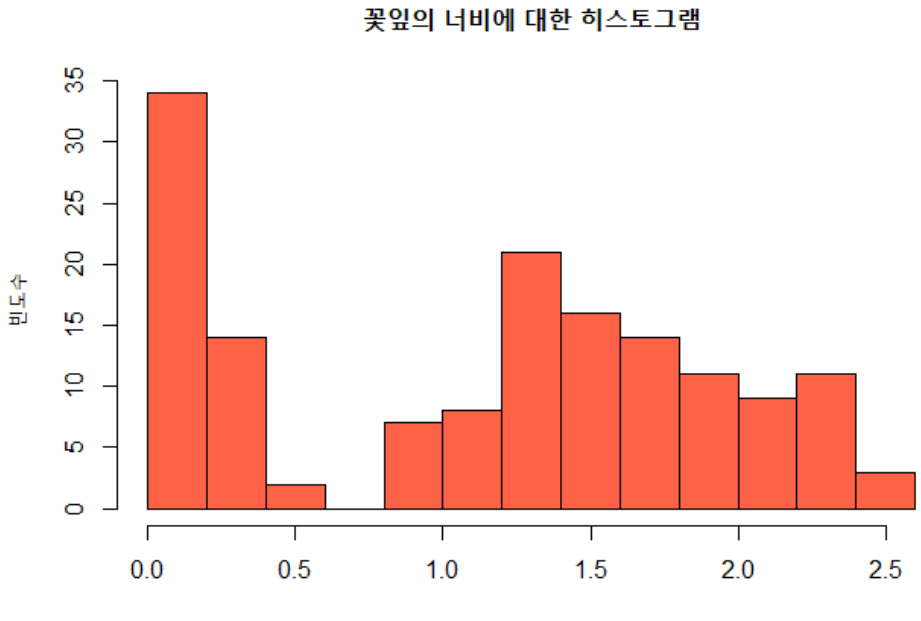
```
> plot(mtcars$mpg, mtcars$wt, col = 'tomato', pch = 19)
```



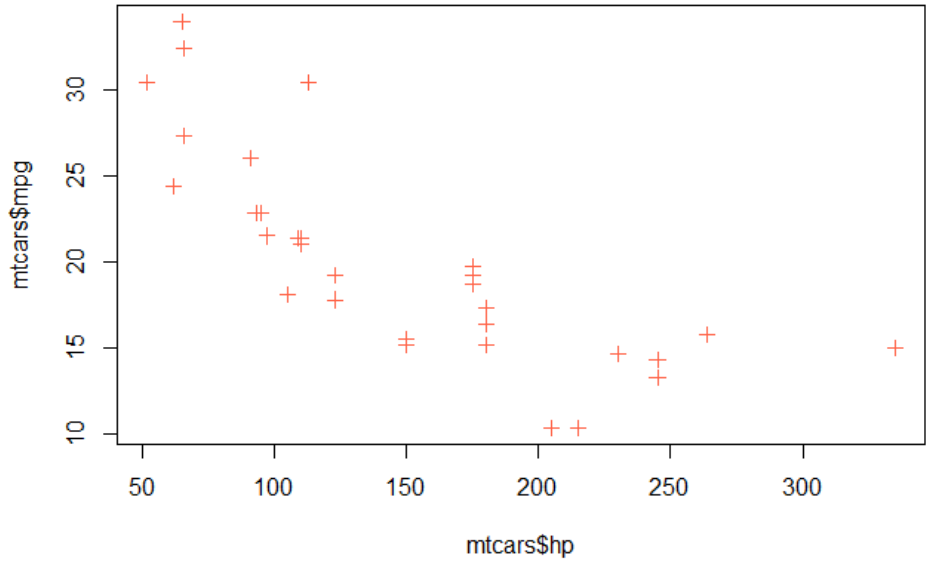
- **연습문제 2.1:**
  - 아래와 같이 iris 데이터셋의 Species 변수에 대해 막대그래프를 그리시오.
    - 막대의 색을 "tomato"색으로 칠해보시오.
    - 막대그래프의 제목과 축의 라벨을 다음과 같이 변경하시오.
      - 제목: "품종의 막대그래프"
      - 가로축: "품종"
      - 세로축: "개수"



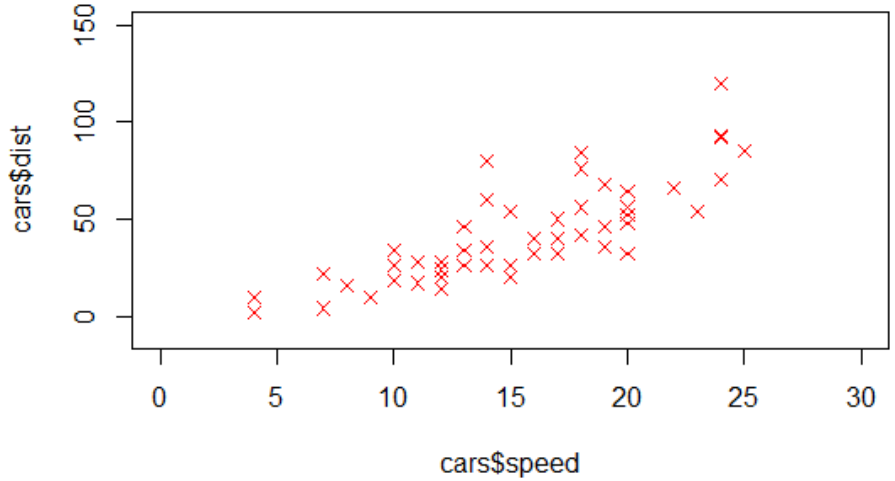
- **연습문제 2.2:**
  - 꽃잎의 너비(Petal.Width)에 대해서 다음 통계량을 구하시오.
    - 평균: *mean*
    - 분산: *var*
    - 표준편차: *sd*
  - 꽃잎의 너비(Petal.Width)에 대해서 히스토그램을 그려보시오.
    - 히스토그램의 색을 변경해보시오.
    - 제목과 축의 라벨을 변경해보시오.



- **연습문제 2.3:**
  - mtcars 데이터셋에서
    - 마력(hp)의 히스토그램을 그려보시오.
    - 히스토그램에서 축의 범위를 바꿔보시오. x축:c(0, 400), y축: c(0, 12)
  - mtcars 데이터셋에서
    - 마력(hp)과 연비(mpg)의 관계를 나타내는 산점도를 그려보시오.
    - 산점도에서 점의 색과 모양을 여러 가지로 바꿔보시오.



- **연습문제 2.4:**
  - R의 내장 데이터셋인 *cars* 데이터셋에 대하여
    - 변수와 관측값의 개수는 각각 얼마인가?
    - speed, dist 변수에 대해서 다음 통계량을 각각 구해보시오.
      - 평균, 중앙값, 최대값, 최소값, 1 사분위 값, 3 사분위 값
    - speed와 dist 변수의 관계를 나타내는 산점도를 그려보시오.
      - 산점도에서 점의 색과 모양을 바꿔보시오.
      - 산점도의 세로축과 가로축의 범위를 바꿔보시오.



Any Questions?