기상청

[STEP2 데이터 탐색]

시각화를 활용한 데이터 탐색 방법

기상청

기상기후
빅데이터 분석 플랫폼

[분석교육] 다양한 데이터 탐색 시각화 방법

1. 그래프 함수

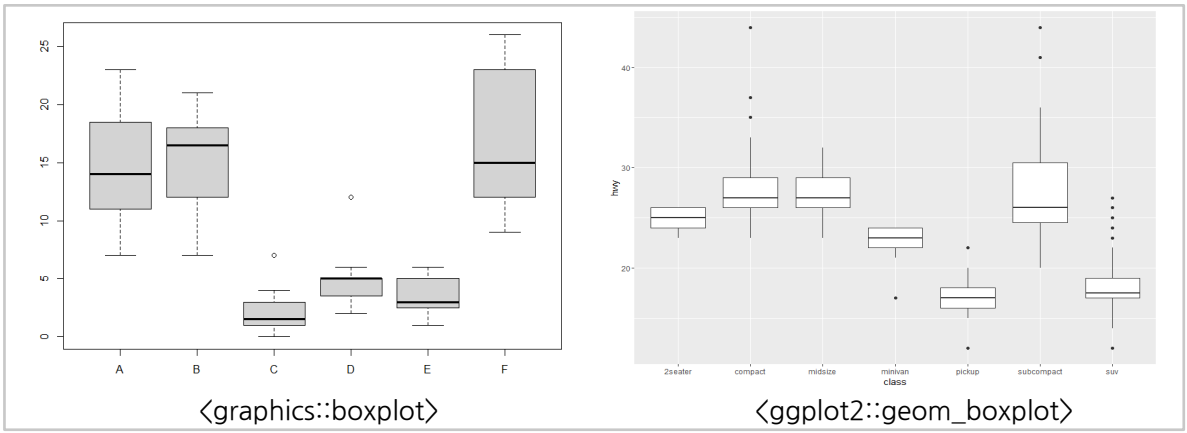
2. 그래프 내보내기 함수

[분석 교육] 다양한 데이터 탐색 시각화 방법(1/4)

데이터 시각화(Data Visualization)은 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하는 방법을 말한다. 데이터 시각화는 대부분 도표(Graph)라는 방법을 통해 표현된다. 분석에서 데이터 탐색을 위해 기본적으로 활용되는 그래프 종류에는 상자그림(Box Plot), 히스토그램(Histogram), 산점도(Scatter Plot), 밀도그림(Density Plot), 막대그래프(Bar Chart), 파이 그래프(Pie Chart) 등이 있다. R에서는 graphics, lattice, ggplot, ggplot2 등의 다양한 시각화 패키지를 지원한다.

● 상자그림(Box Plot)

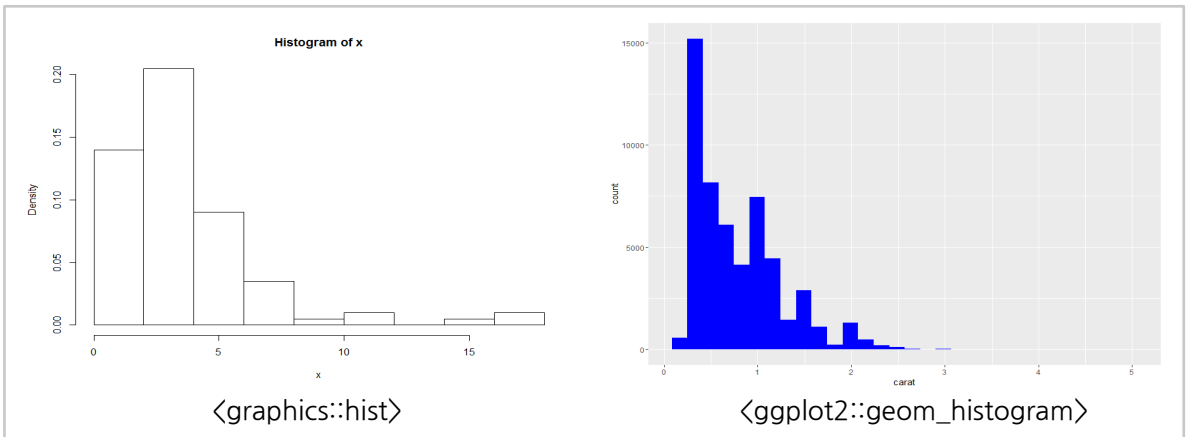
- 상자그림(Box Plot)은 데이터의 분포를 보여주는 그래프로, 가운데 상자는 제 1사분위수, 중앙값, 제 3사분위수를 보여준다. 상자의 좌우 또는 상하로 뻗어나간 선(whisker)은 이상치에 해당한다.



〈graphics::boxplot〉 〈ggplot2::geom_boxplot〉

● 히스토그램(Histogram)

- 히스토그램(Histogram)은 도수분포표를 그래프로 나타난 것으로, 데이터의 분포를 파악할 때 유용하다.보통 히스토그램은 가로축이 계급, 세로축이 도수를 뜻한다.

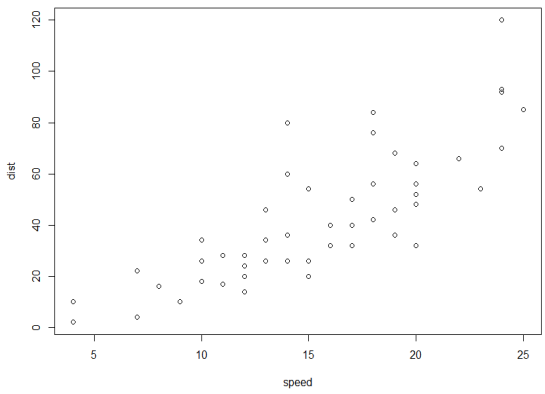


〈graphics::hist〉 〈ggplot2::geom_histogram〉

[분석 교육] 다양한 데이터 탐색 시각화 방법(2/4)

● 산점도(Scatter Plot)

- 산점도(Scatter Plot)은 주어진 데이터를 점으로 표시해 흩뿌리듯 시각화한 그래프이다. 데이터의 실제 값들의 분포를 파악하는 데 유용하다.

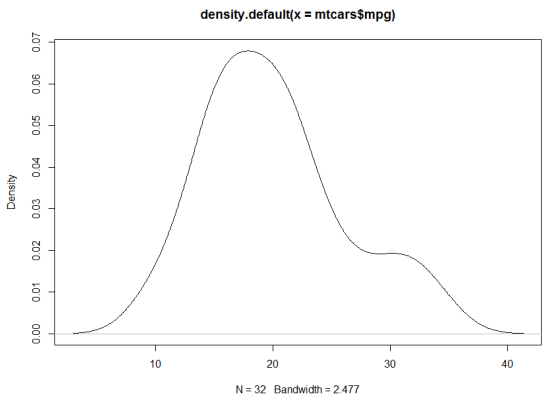


〈graphics::plot〉

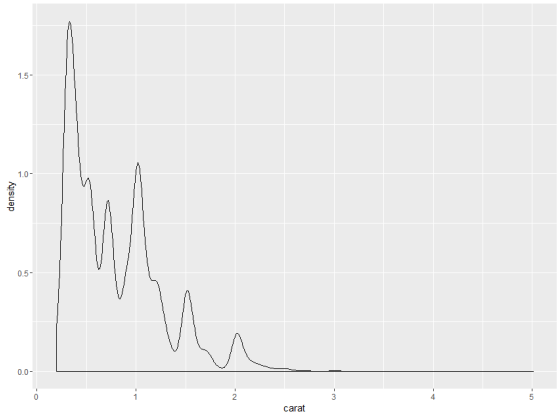
〈ggplot2::geom_point〉

● 밀도그림 (Density Plot)

- 히스토그램에서 막대 구간의 폭이 분포의 모양을 결정하는 중요한 요소인것에 비해, 밀도그림(Density Plot)은 데이터의 밀도를 추정하는 커널 밀도 추정(Kernel Density Estimation)에 의한 분포를 파악하기 위한 그래프이다.



〈stats::density〉

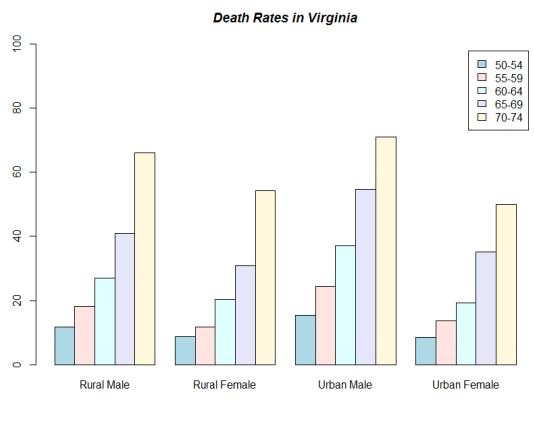


〈ggplot2::geom_density〉

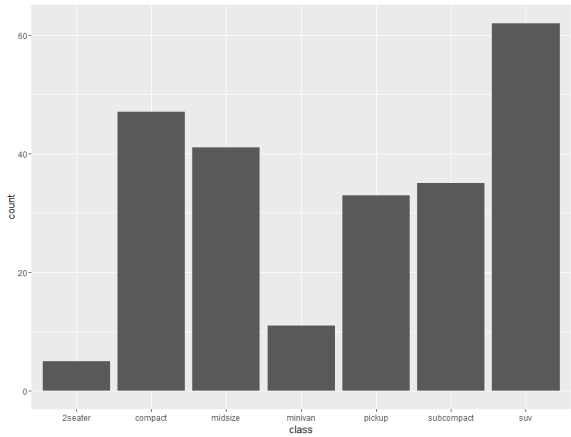
[분석 교육] 다양한 데이터 탐색 시각화 방법(3/4)

● 막대 그래프(Bar Chart)

- 막대 그래프(Bar Chart)는 하나의 축을 기준으로 데이터를 막대로 표현한 그래프이다. 가장 기본적인 차트이며 직관적인 정보를 전달 할 수 있다.



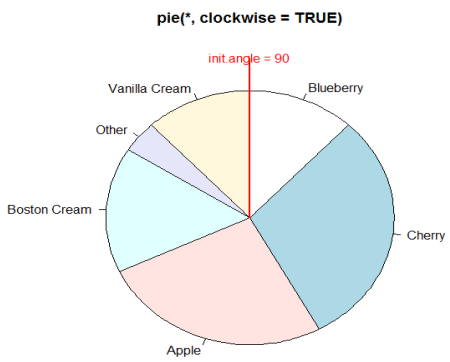
〈graphics::barplot〉



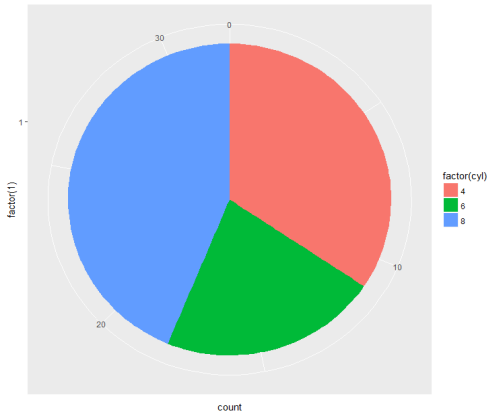
〈ggplot2::geom_bar〉

● 파이 그래프(Pie Chart)

- 파이차트(Pie Chart)는 원형 차트로 데이터에서 각 범주의 상대적인 크기나 비율을 파악하기 위해 사용된다.



〈graphics:pie〉



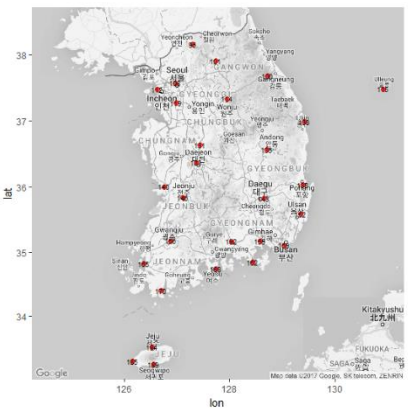
〈ggplot2::coord_polar〉

참고 자료 : R Documentation

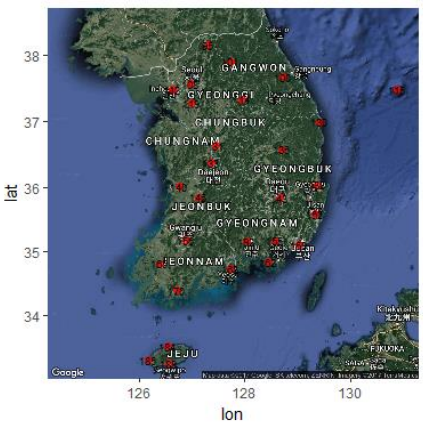
[분석 교육] 다양한 데이터 탐색 시각화 방법(4/4)

● 구글지도(Map)

- 위도와 경도로 표시된 위치 정보를 2차원 구글지도(Map) 위에 점으로 표시하는 공간시각화 방법이다. 분석 대상이 되는 지점의 위치를 파악할 수 있으며, 직관적으로 정보를 전달할 수 있다.



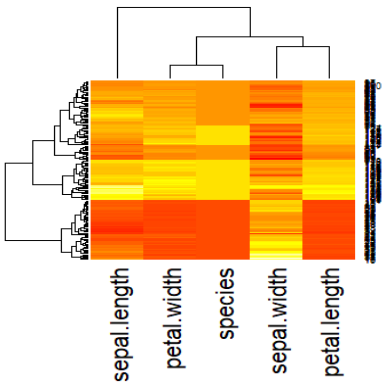
〈ggmap::roadmap〉



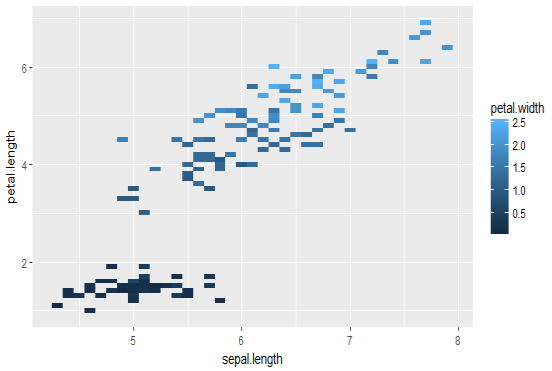
〈ggmap::hybrid〉

● 히트맵(Heat Map)

- 히트맵(Heat Map)은 다양한 정보를 2차원 평면상에 숫자 대신 색상으로 표현하는 그래프이다. 데이터의 빈도를 파악할 때 유용하며, 직관적으로 정보를 전달할 수 있다.



〈graphics::heatmap〉



〈ggplot2::geom_tile〉

참고 자료 : R Documentation

1. 그래프 함수(1/5)

- **boxplot**

  - 상자그림(Box Plot)을 그리는 함수

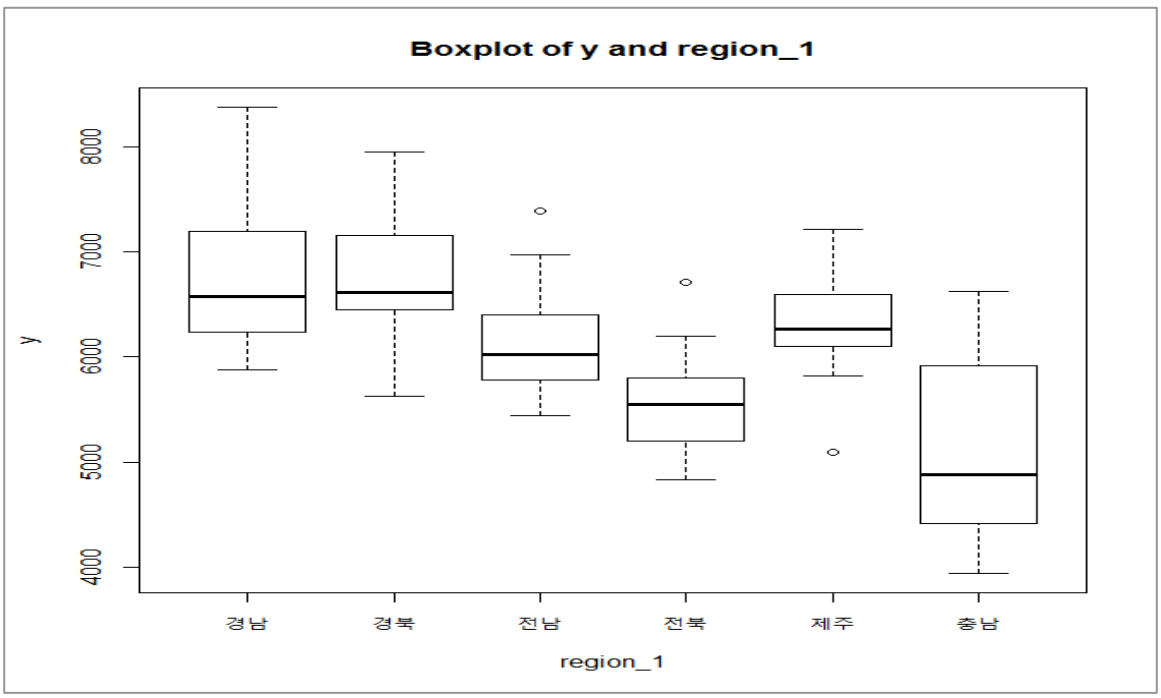
▪ **Usage**

boxplot(data = data, x ~group, …)

- data : 데이터 프레임 또는 리스트 타입의 데이터
- x : 상자그림으로 생성할 숫자 데이터 타입의 특정 변수
  특정 그룹으로 묶어 상자그림을 그려야 할 때는 formula 옵션 사용
- formula : 특정 숫자 데이터 타입의 변수를 그룹으로 묶어 상자그림 생성, y ~ group
- xlab : x축 라벨
- ylab : y축 라벨
- main : 그래프 제목

▪ **Examples**

```
boxplot(data=onion, y~region_1, xlab="region_1", ylab="y",
        main="Boxplot of y and region_1" )
```



1. 그래프 함수(2/5)

- **par**

  - 그래픽 파라미터를 지정

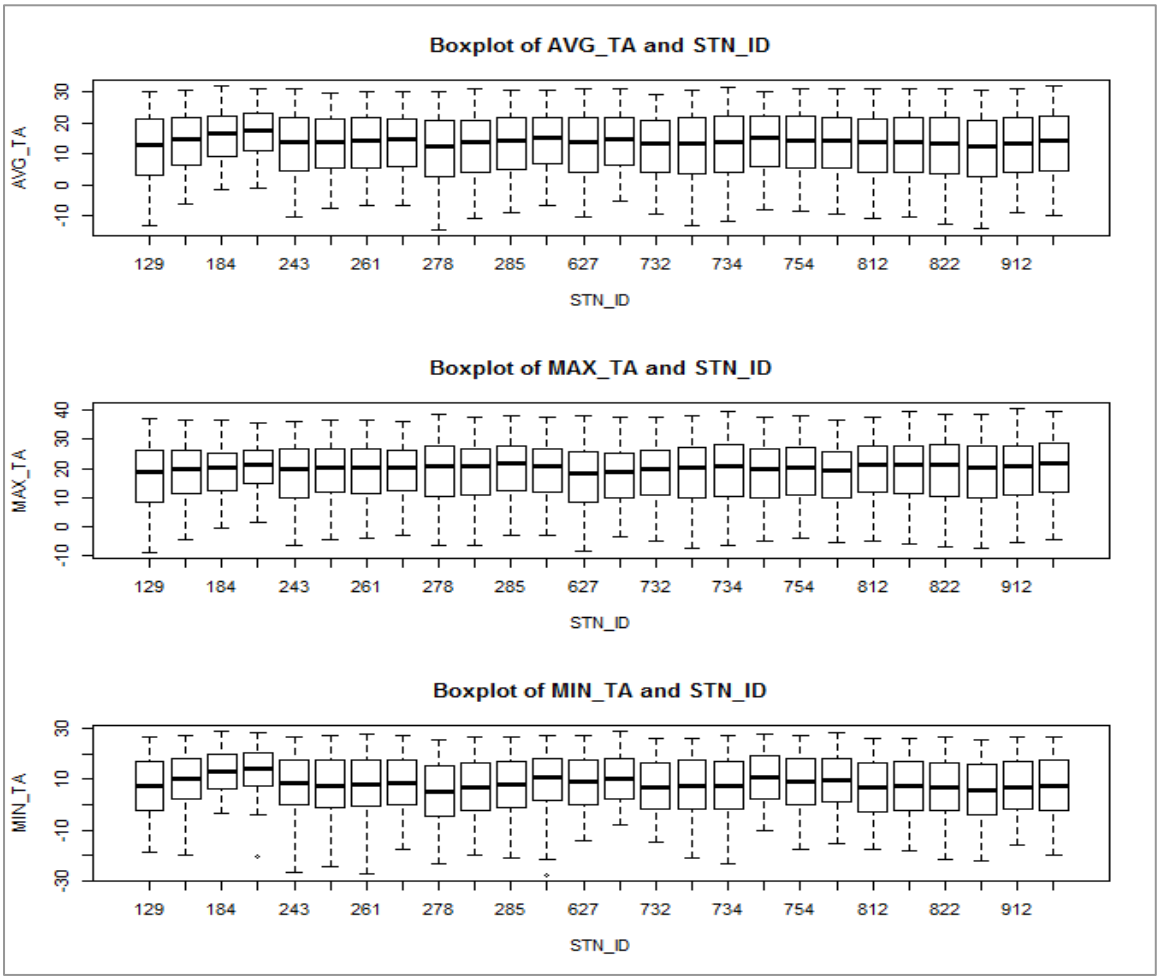
▪ **Usage**

par(mfrow = c(행,열))

- mfrow : 그래프 배열을 지정, c(행 수, 컬럼 수)로 표현

▪ **Examples**

```
par(mfrow=c(3,1)) # 3행 1열의 그래프 배열 지정
boxplot(data=TA_RN_SS, AVG_TA~STN_ID, xlab="STN_ID", ylab="AVG_TA",
        main="Boxplot of AVG_TA and STN_ID")
boxplot(data=TA_RN_SS, MAX_TA~STN_ID, xlab="STN_ID", ylab="MAX_TA",
        main="Boxplot of MAX_TA and STN_ID")
boxplot(data=TA_RN_SS, MIN_TA~STN_ID, xlab="STN_ID", ylab="MIN_TA",
        main="Boxplot of MIN_TA and STN_ID")
```



1. 그래프 함수(3/5)

- **hist**

  - 히스토그램(Histogram)을 그리는 함수

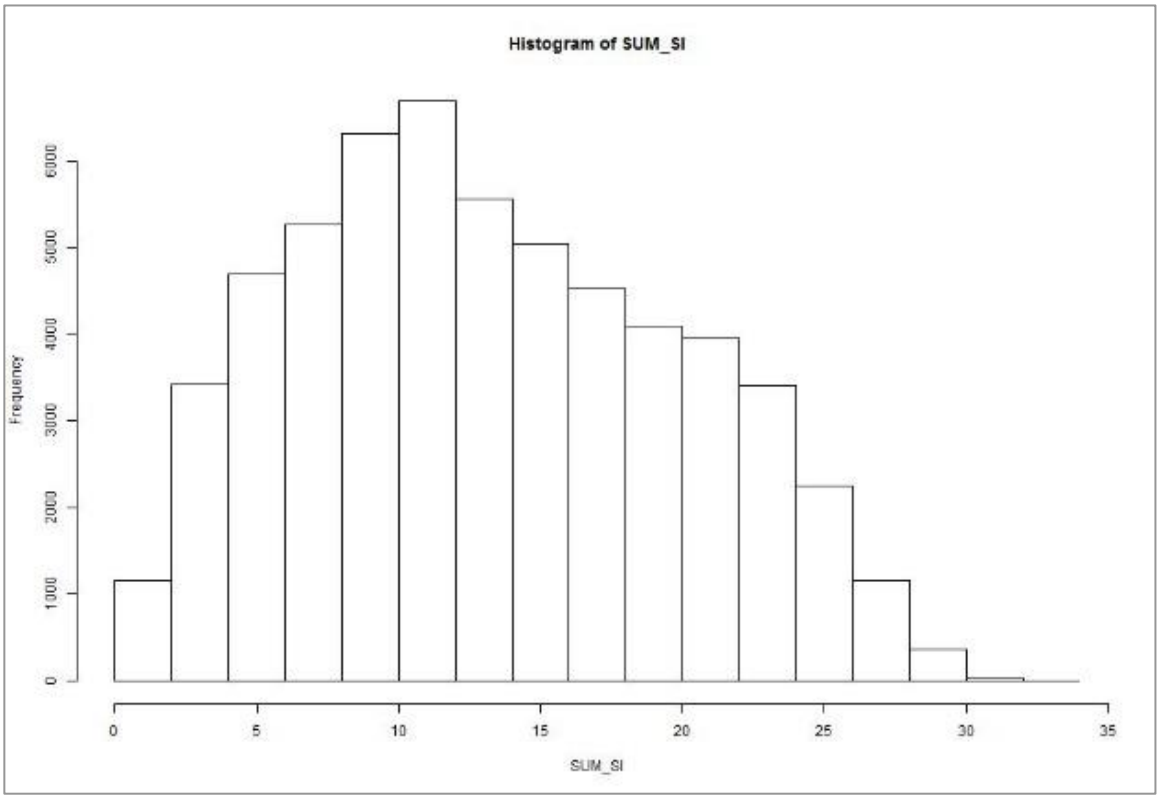
▪ **Usage**

hist(x, breaks = …, freq = TRUE, …)

- x : 히스토그램을 그리려는 데이터
- breaks : 히스토그램의 구간 갯수 설정
- freq : TRUE면 각 구간의 빈도를 히스토그램으로 표현, FALSE면 각 구간의 확률 밀도를 히스토그램으로 표현, 기본값은 TRUE임
- xlab : x축 라벨
- ylab : y축 라벨
- main : 그래프 제목
- xlim : x축의 범위
- ylim : y축의 범위

▪ **Examples**

```
hist(DATA[,colnames(DATA) == y] , breaks = 17, main = "Boxplot of SUM_SI",
     xlab = y, ylab = "Frequency")
```



1. 그래프 함수(4/5)

- **ggplot2::qplot**

  - ggplot2 패키지에서 그래프를 그리는 기본 함수, graphics::plot()과 같은 역할

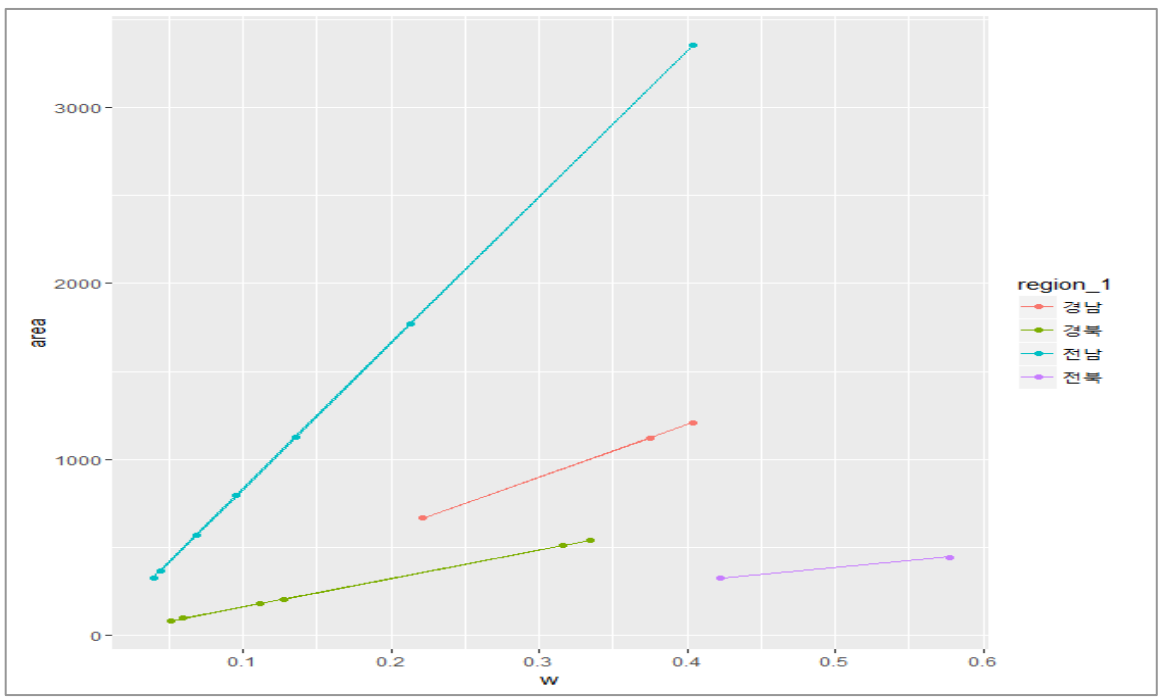
▪ **Usage**

qplot(x, y, data = data, …)

- x : x축으로 나타낼 변수
- y : y축으로 나타낼 변수
- data : 그래프를 생성할 데이터
- color : 각 카테고리를 색으로 구분, 구분할 변수명 입력
- shape : 각 카테고리를 도형 모양으로 구분, 구분할 변수명 입력
- size : 각 카테고리를 크기로 구분, 구분할 변수명 입력
- geom : 생성할 그래프 타입 설정
  : " point " , " smooth " , " boxplot " , " line " , " histogram " , " density " , " bar " , “path”, and " jitter " .
- xlab : x축 라벨
- ylab : y축 라벨
- main : 그래프 제목

▪ **Examples**

```
qplot(x = w, y = area, data = subset(onion.area, !is.na(w) & !is.na(area)),
      color = region_1, geom = c("path", "jitter"))
```



1. 그래프 함수(5/5)

- **heatmap**

  - 히트맵(Heatmap)을 그리는 함수

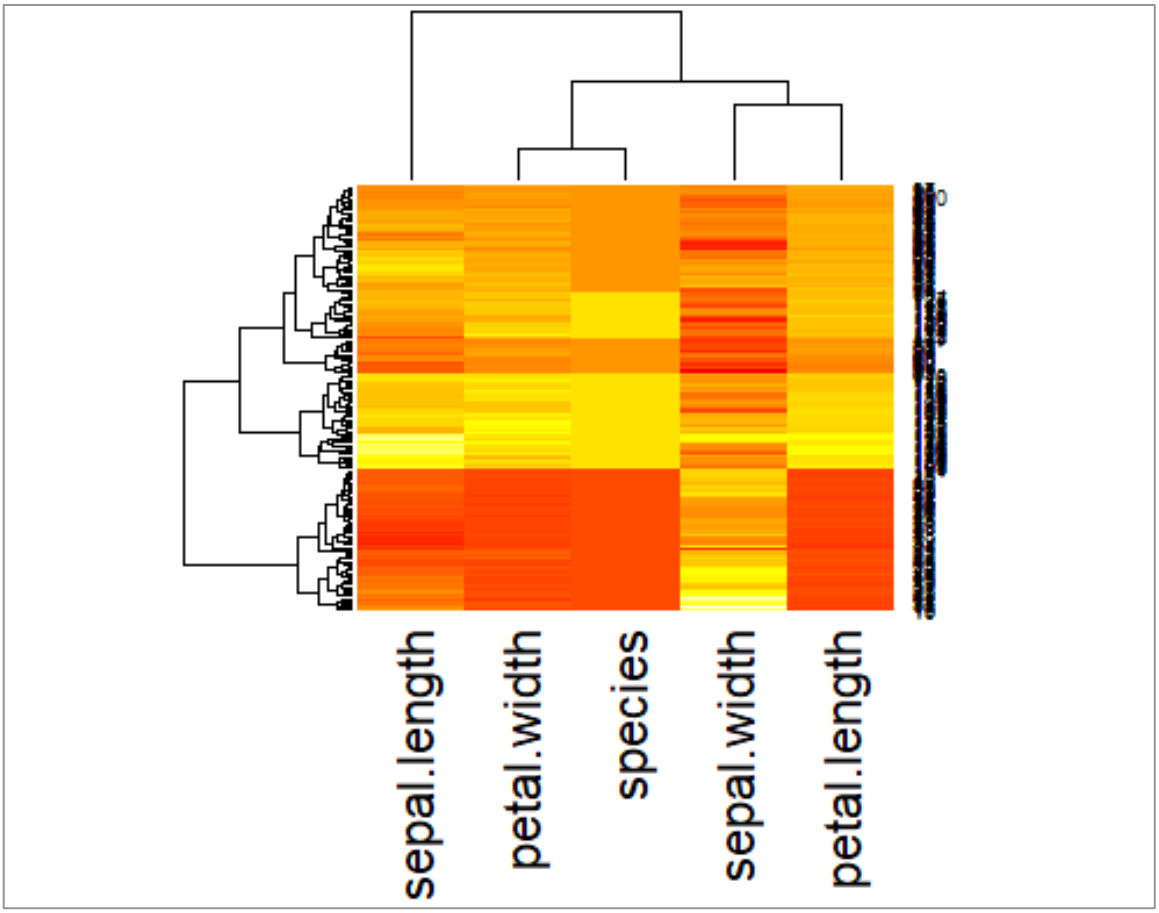
▪ **Usage**

heatmap(x, col = color, scale = …, margins = …)

- x : 시각화할 숫자형 매트릭스
- col : 색상
- scale : 값이 맞춰지는 방향
- margins : 행과 열 각각의 여백을 포함한 길이

▪ **Examples**

```
iris2 <- data.matrix(iris)
iris_heatmap <- heatmap(iris2, col = heat.colors(256), scale = "column“,
                        margins = c(8, 5)
```



2. 그래프 내보내기 함수

- **jpeg / pdf / png / tiff / bmp / postscript**
  - R에 생성된 그래프를 외부 파일로 내보냄

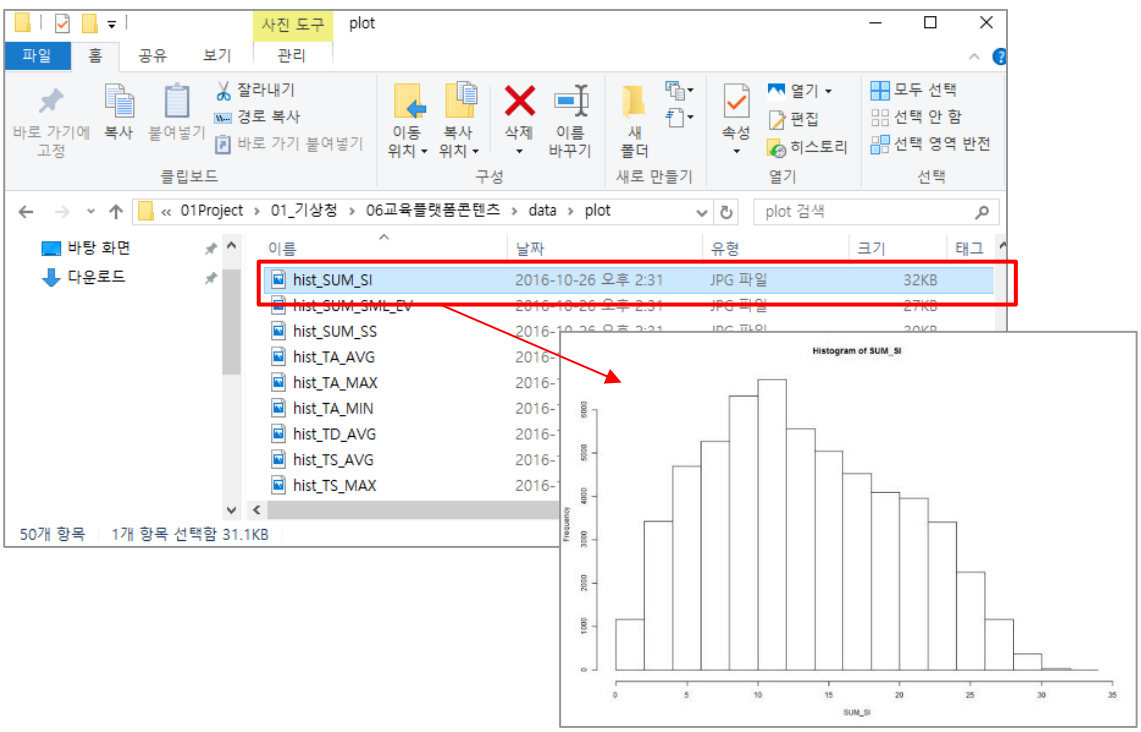
  ▪ **Usage**
  jpeg(filename, width = …, height = …, …)
  - filename : 외부에 저장할 파일명
  - width : 그래프 넓이
  - height : 그래프 높이
  - units : 그래프 넓이와 높이의 단위 “px”,”in”,”cm”,”mm”

  ▪ **Examples**
  ```
  jpeg("./data/plot/hist_SUM_SI.jpg", width = 950, height = 650, units = "px")
  hist(DATA[,colnames(DATA) == y] , main = mainname, xlab = y)
  ```

- **dev.off**
  - 현재 R에 생성된 그래프 장치를 종료함

  ▪ **Usage**
  dev.off()

  ▪ **Examples**
  ```
  hist(DATA[,colnames(DATA) == y] , main = mainname, xlab = y)
  dev.off()
  ```



본 문서의 내용은 기상청의 날씨마루(http://big.kma.go.kr) 내
R 프로그래밍 교육 자료입니다.