문공 A0015

R 프로그래밍

김 태 완

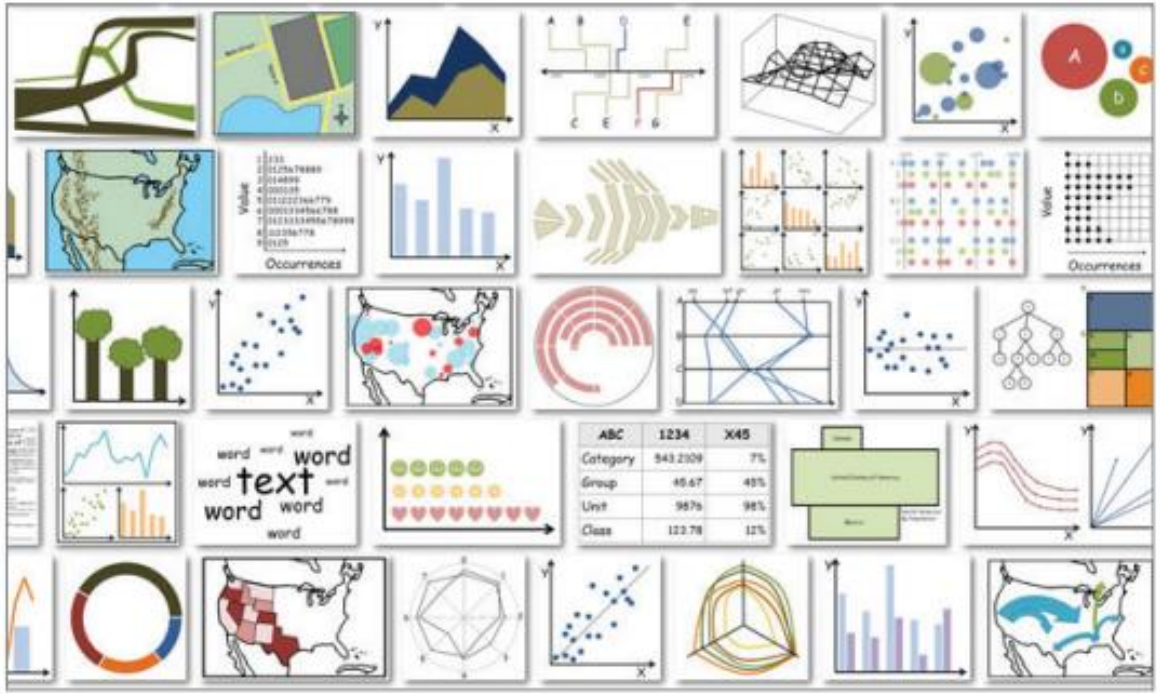
kimtwan21@dongduk.ac.kr

데이터 시각화

- 데이터 시각화 기법
  - 트리맵
  - 버블차트
  - 모자이크 플롯 (mosaic plot)
- ggplot 패키지
  - ggplot 명령문 구조
  - 막대그래프 (bar plot)
  - 히스토그램 (histogram)
  - 산점도 (scatter plot)
  - 상자그림 (box plot)
  - 선그래프 (line plot)

데이터 시각화

- 데이터 시각화의 중요성
  - 데이터 시각화(data visualization) : 숫자 형태의 데이터를 그래프나 그림 등의 형태로 표현하는 과정
  - 데이터 분석 과정에서 중요한 기술 중의 하나
  - 데이터를 시각화 하면 데이터가 담고 있는 정보나 의미를 보다 쉽게 파악
  - 경우에 따라서는 시각화 결과로부터 중요한 영감을 얻기도 함



데이터 시각화

- 데이터 시각화 기법
  - 트리맵

트리맵(tree map)

사각 타일 형태로 구성

각 타일의 크기와 색으로 데이터에 담긴 정보를 표현함
각각의 타일이 계층 구조이기 때문에 데이터의 계층 구조를 표현할 수 있음

treemap 패키지 설치 필요

```
install.packages("treemap")
```

데이터 시각화

- 데이터 시각화 기법
  - 트리맵

    - GNI2014 데이터셋 : 2014년도의 전 세계 국가별 인구, 국민총소득(GNI), 소속 대륙의 정보

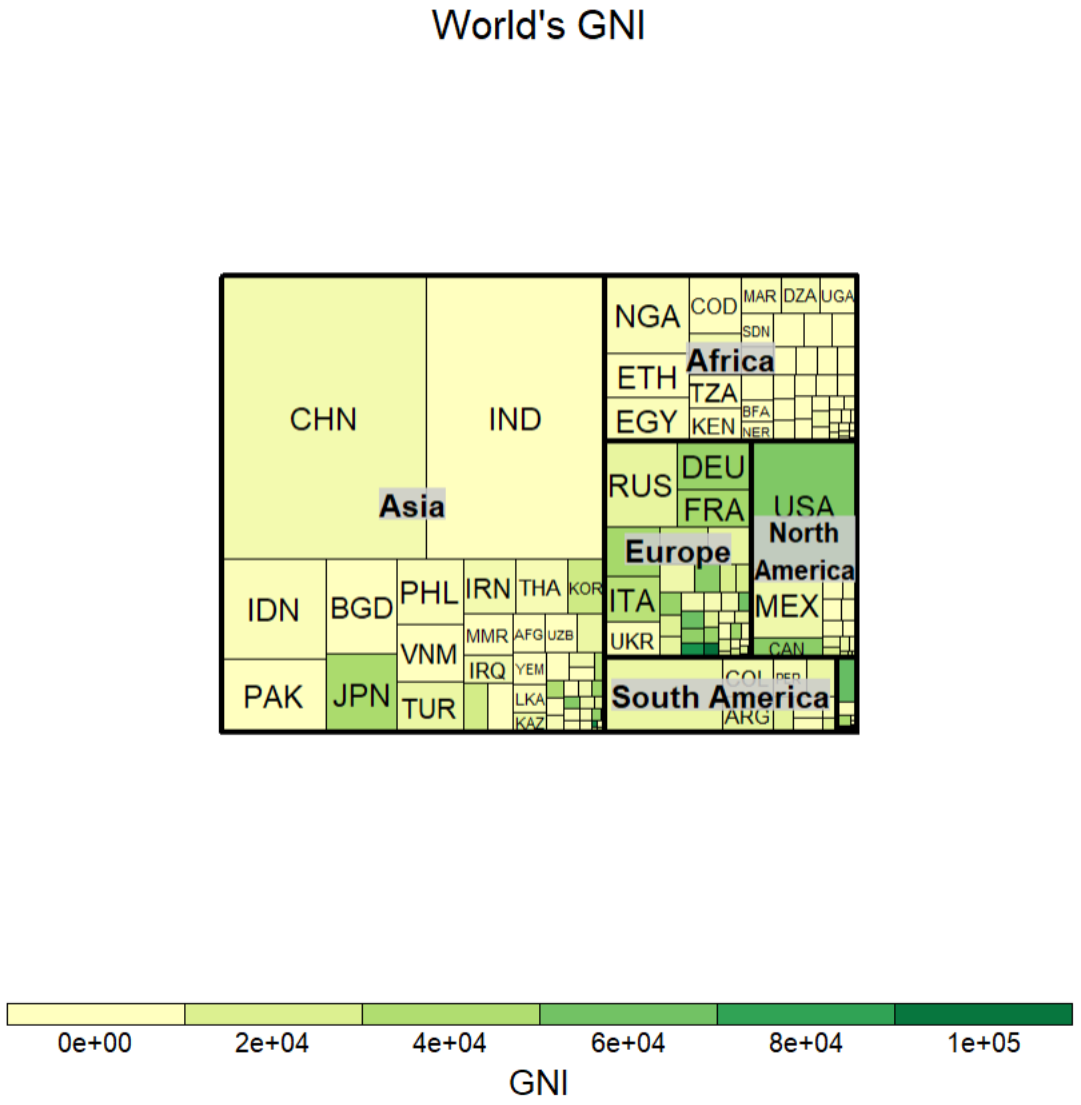
```
>   install.packages("treemap")          # treemap 패키지를 설치

>   library(treemap)                     # treemap 패키지 불러오기
>   data(GNI2014)                        # 데이터 불러오기
                                            ->GNI2014 : treemap 패키지 안에 포함된 데이터셋
>   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
...
```

데이터 시각화

- 데이터 시각화 기법
  - 트리맵

    - GNI2014 데이터셋으로 트리맵 작성하기

```
treemap(GNI2014,
        index=c('continent','iso3'),
        vSize='population',
        vColor='GNI',
        type='value',
        title = 'GNI')
```

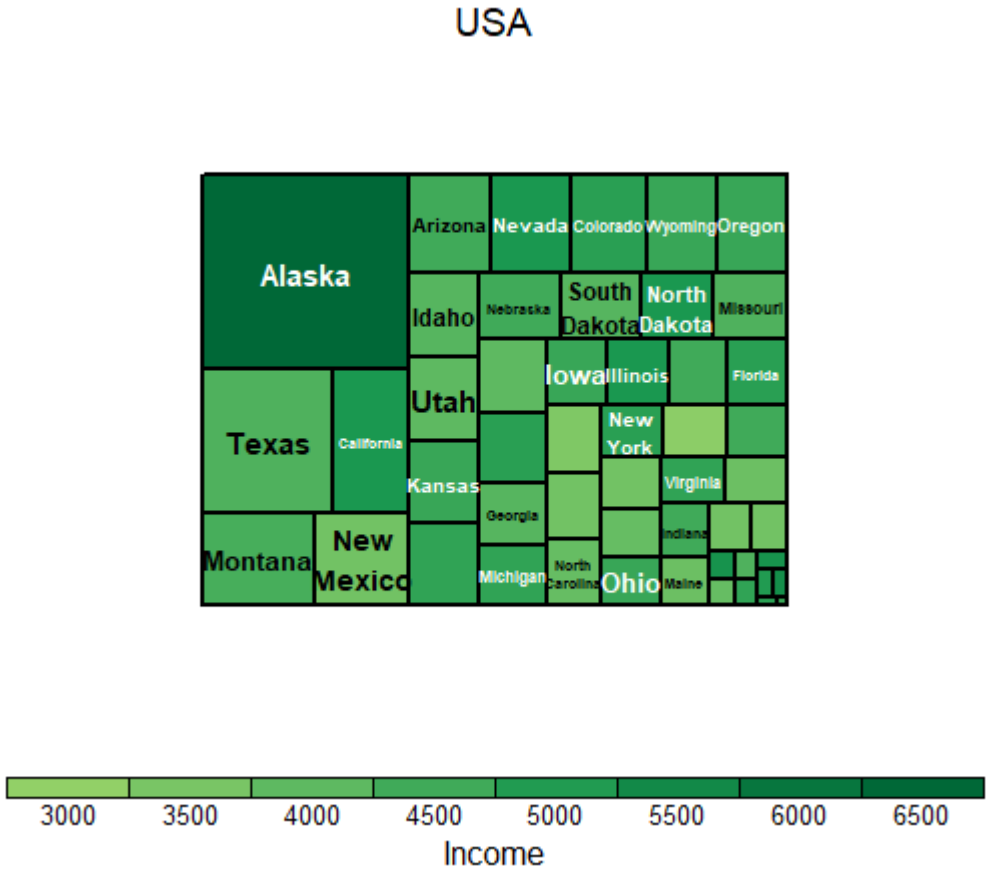


데이터 시각화

- 데이터 시각화 기법
  - 트리맵

    - state.x77 데이터셋으로 트리맵 작성하기

```
st <- data.frame(state.x77)
st <- data.frame(st,
stname=rownames(st))
treemap(st,
        index = c("stname"),
        vSize = 'Area',
        vColor = 'Income',
        type = 'value',
        title = 'USA')
```



데이터 시각화

- 데이터 시각화 기법
  - 버블차트

버블차트(bubble chart)

산점도 위에 버블의 크기로 정보를 표시하는 시각화 방법

산점도	→	2개의 변수에 의한 위치 정보를 표시
버블차트	→	3개의 변수 정보를 하나의 그래프에 표시

데이터 시각화

- 데이터 시각화 기법
  - 버블차트 : symbols( ) 함수와 text( ) 함수의 조합으로 생성

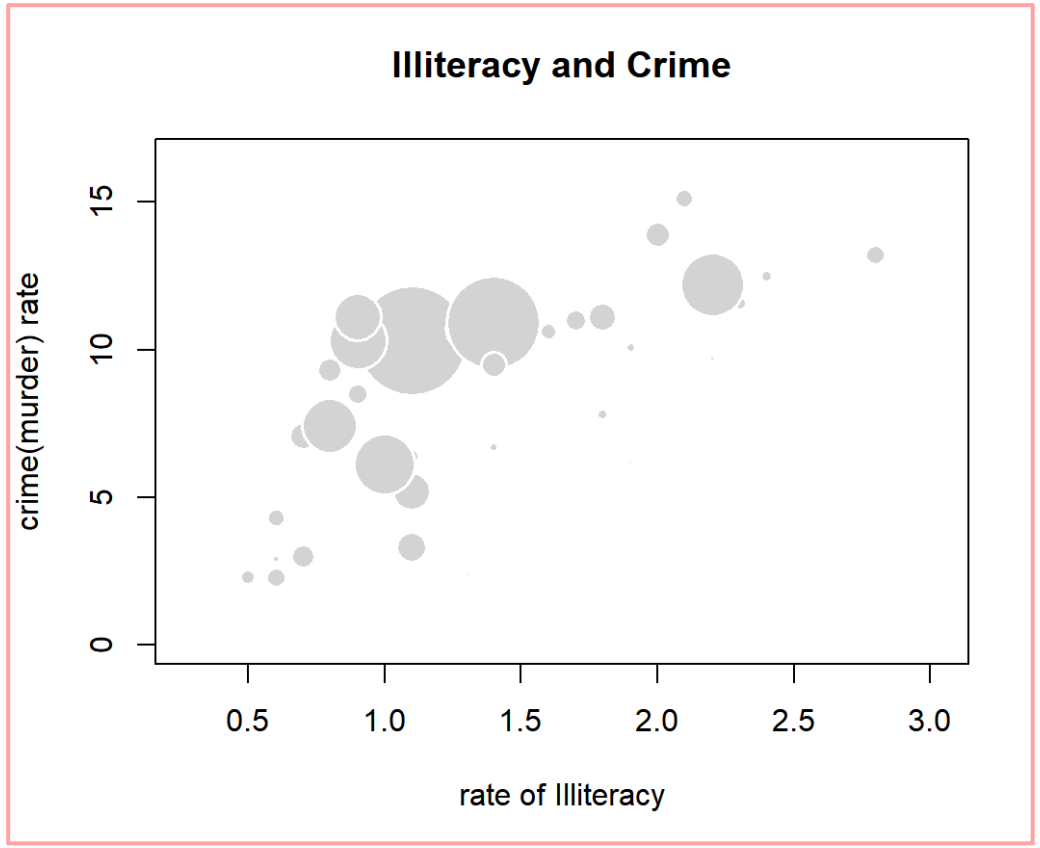
    - state.x77 데이터셋을 이용한 버블차트

```
>   st <- data.frame(state.x77)                  # state.x77(매트릭스)를 데이터프레임으로 변환
> symbols(st$Illiteracy, st$Murder,              # 2차원 좌표의 x축과 y축을 나타낼 열을 지정
+         circles = st$Population,               # 원의 크기(반지름)을 결정할 열을 지정
+                                                   ex)Population값이 커지면 원의 크기가 커짐
+         inches = 0.3,                          # 원의 크기를 조절하는 매개변수
+                                                   매개변수값이 클수록 원이 크게 그려짐
+         fg = 'white',                          # 원의 테두리선 색
+         bg = 'lightgray',                      # 원의 바탕색
+         lwd = 1.5,                             # 원의 테두리선 두께
+         xlab = 'rate of Illiteracy',           # x축의 레이블
+         ylab = 'crime(murder) rate',           # y축의 레이블
+        main = 'Illiteracy and Crime')          # 그래프의 제목
```

데이터 시각화

- 데이터 시각화 기법
  - 버블차트 : symbols( ) 함수와 text( ) 함수의 조합으로 생성

    - state.x77 데이터셋을 이용한 버블차트



symbols( ) 함수 실행 결과
-> 각 원이 무엇을 의미하는지 알 수 없음
-> text( ) 함수 필요

데이터 시각화

- 데이터 시각화 기법
  - 버블차트 : symbols( ) 함수와 text( ) 함수의 조합으로 생성

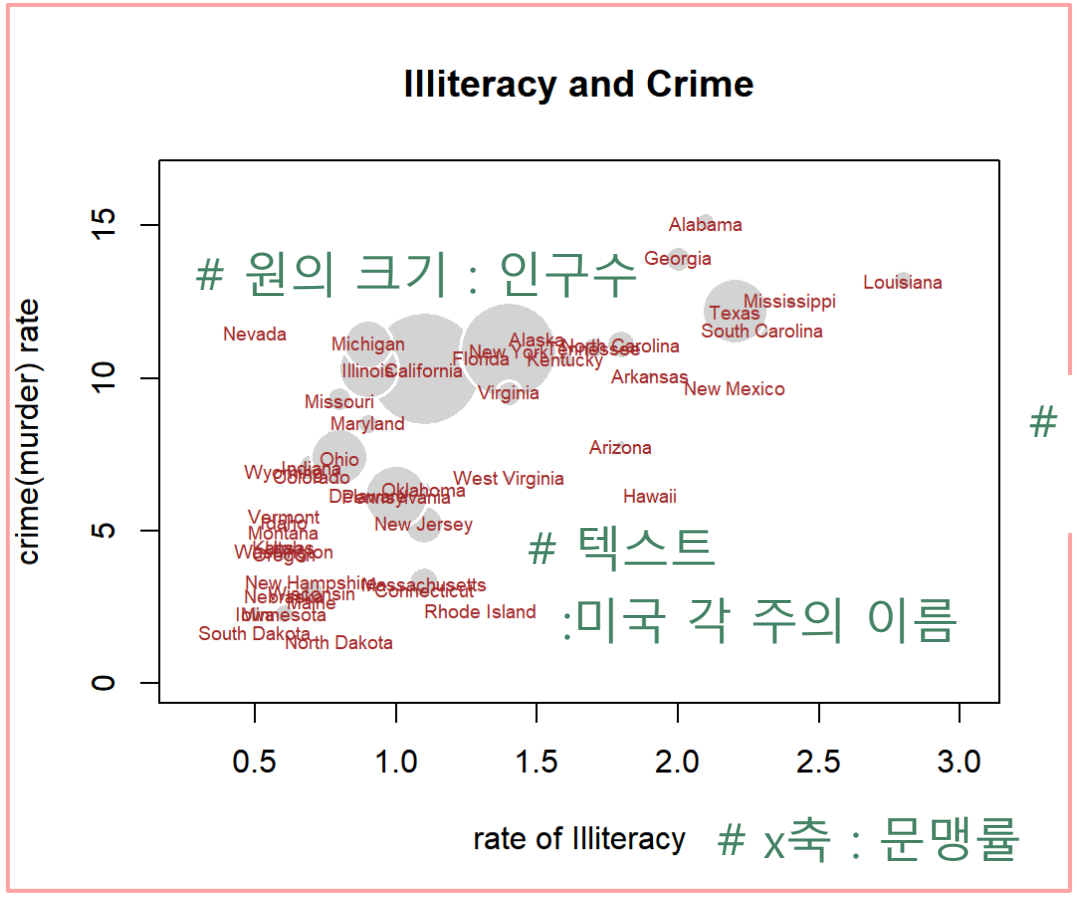
    - state.x77 데이터셋을 이용한 버블차트

```
> text(st$Illiteracy, st$Murder,      # 텍스트를 표시할 위치에 대한 x축과 y축 좌표값
+                                     symbols( ) 함수에 있는 원의 x축과 y축 좌표값과 일치시킴
+       rownames(st),                 # 표시할 텍스트를 지정
+       cex = 0.6,                    # 텍스트의 크기
+       col = 'brown')                # 텍스트의 색
```

데이터 시각화

- 데이터 시각화 기법
  - 버블차트 : symbols( ) 함수와 text( ) 함수의 조합으로 생성
    - state.x77 데이터셋을 이용한 버블차트

y축 : 범죄율

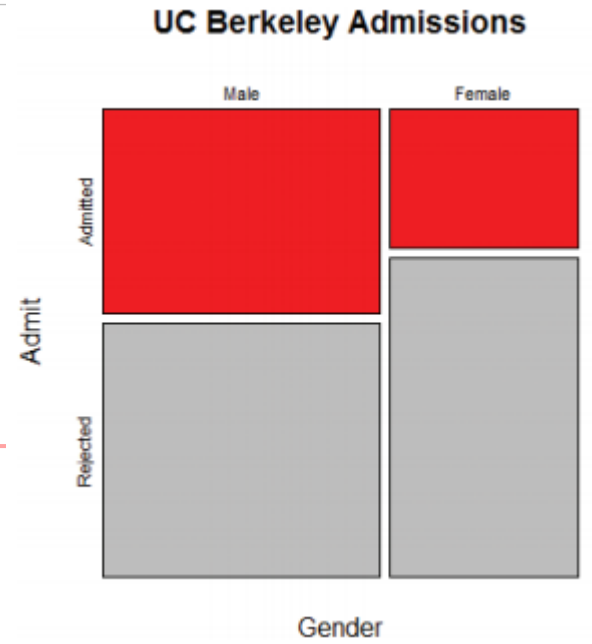


변수 : 원의 위치(문맹률, 범죄율), 원의 크기(인구 수)->3개

x축 : 문맹률

데이터 시각화

- 데이터 시각화 기법
  - 모자이크 플롯



모자이크 플롯(mosaic plot)

범주형 데이터에 대해 각 변수의 그룹별 비율을 면적으로 표시하여 정보를 전달

데이터의 형태가 범주형 자료 또는 개수를 셀 수 있는 정수형 자료여야 함

데이터 시각화

- 데이터 시각화 기법
  - 모자이크 플롯 : mosaicplot( ) 함수 이용

    - mtcars 데이터셋을 이용한 모자이크 플롯

```
>   head(mtcars)
                   mpg cyl  disp hp  drat    wt    qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
...
```

정수형 자료

모자이크 플롯을 작성할 대상 변수 — y축 변수의 그룹별로 음영을 달리하여 표시(색상 지정X)

```
>   mosaicplot(~gear+vs, data = mtcars, color = TRUE,        # 대상 변수
        main = 'Gear and Vs')                                  : ~ 다음 변수(gear)가 x축 방향,
                                                                 + 다음 변수(vs)가 y축 방향
```

데이터셋

모자이크 플롯의 제목

데이터 시각화

- 데이터 시각화 기법
  - 모자이크 플롯 : mosaicplot( ) 함수 이용

y축 : 엔진의 형태

기어가 3개인 것의 비율

기어가 5개인 것 중 엔진의 형태가 1타입인 것의 비율

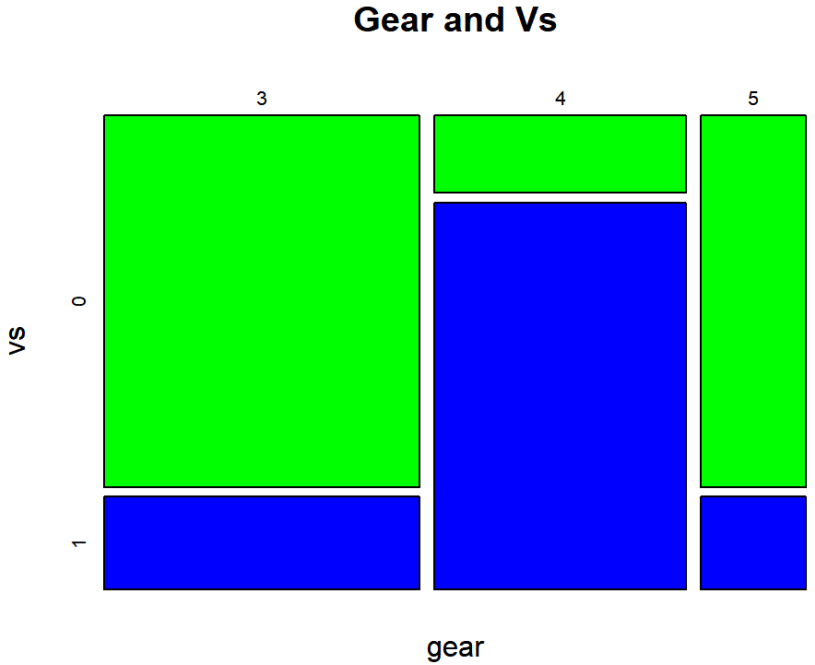
x축 : 기어의 개수

데이터 시각화

- 데이터 시각화 기법
  - 모자이크 플롯 : mosaicplot( ) 함수 이용

    - mtcars 데이터셋을 이용한 모자이크 플롯

```
> mosaicplot(~gear+vs, data = mtcars, color = c('green', 'blue'),
+            main = 'Gear and Vs')    # y축 변수에 색을 지정
                                      (vs가 0, 1 두개의 그룹이므로 순서대로 2 개의 색을 지정)
```



데이터 시각화

- ggplot 패키지

ggplot 패키지

R에서 제공하는 기본적인 함수들보다 미적인 그래프의 작성에 이용

복잡하고 화려한 그래프를 작성할 수 있음

데이터 시각화

- ggplot 패키지

ggplot 명령문의 기본 구조

사용할 데이터셋

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

x축, y축으로 사용할 열 이름

ggplot은 보통 하나의 ggplot( ) 함수와 여러 개의 geom_xx( ) 함수들이 +로 연결되어 하나의 그래프를 완성함

ggeom_xx( ) 함수
: 어떤 형태의 그래프를 그릴지 지정

```
ggplot(data=mtcars,aes(x=wt,y=mpg))+geom_point() +labs(x='weight'...)
```

ggplot 패키지 설치 필요

```
install.packages("ggplot2")
```

- ggplot 패키지
  - 막대그래프의 작성 : geom_bar( ) 함수 이용

    - 기본적인 막대그래프 작성하기

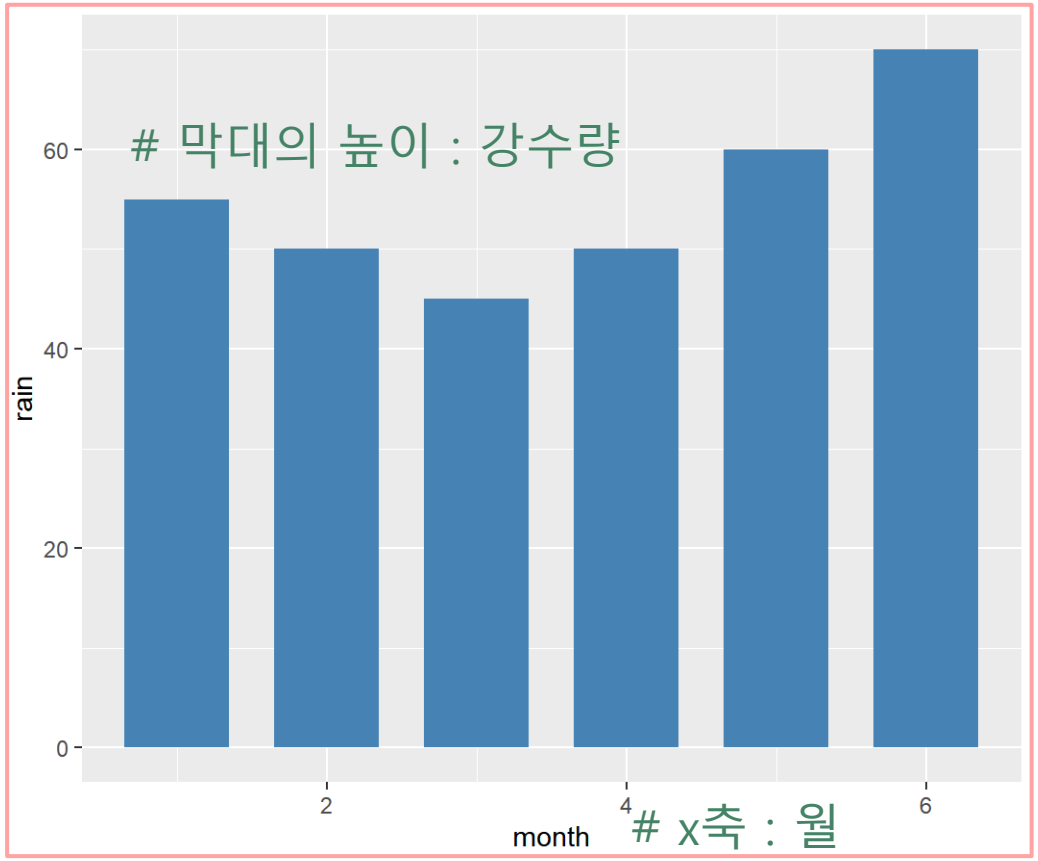
```
>   library(ggplot2)                          # ggplot2 패키지 불러오기
>   month <- c(1, 2, 3, 4, 5, 6)              # 1월부터 6월까지 데이터 저장
>   rain <- c(55, 50, 45, 50, 60, 70)         # 강우량 데이터 저장
>   df <- data.frame(month, rain)             # 데이터프레임에 저장하기
>   df
 month rain
1     1   55
2     2   50
...                                           # 그래프를 그릴 데이터 지정
> ggplot(df, aes(x=month, y=rain))+           # geom_bar( ) 함수 : 막대그래프 생성
+   geom_bar(stat='identity',                 # 막대의 높이는 ggplot( ) 함수에서 y축에 해당하는
+                                                열(rain)에 의해 결정되도록 지정
+            width = 0.7,                     # 막대의 폭 지정
+            fill = 'steelblue')              # 막대의 내부 색 지정
```

데이터 시각화

- ggplot 패키지
  - 막대그래프의 작성 : geom_bar( ) 함수 이용

    - 기본적인 막대그래프 작성하기

y축 : 강수량



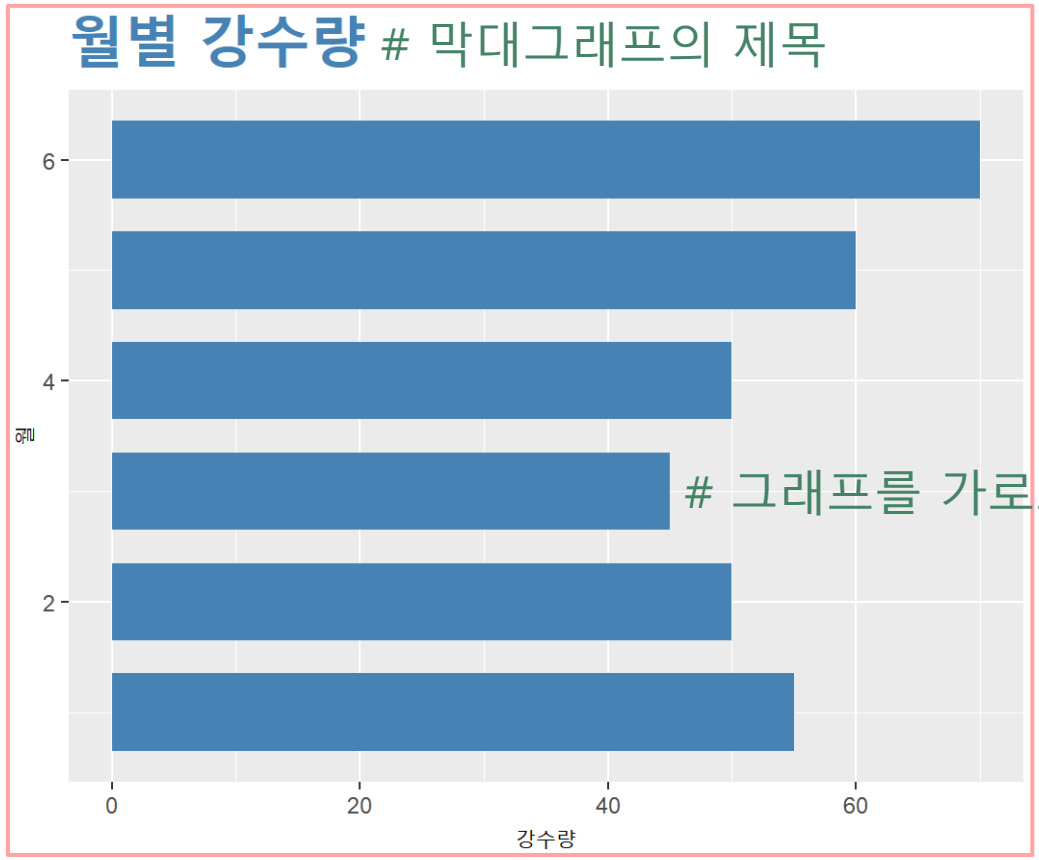
- ggplot 패키지
  - 막대그래프의 작성 : geom_bar( ) 함수 이용

    - 막대그래프 꾸미기

```
> ggplot(df, aes(x=month, y=rain))+
+   geom_bar(stat='identity',
+            width = 0.7,
+            fill = 'steelblue')+
+   ggtitle('월별 강수량')+                                    # 그래프의 제목 지정
+   theme(plot.title = element_text(size=25, face='bold', colour = 'steelblue'))+
+                                                              # 제목의 폰트 크기, 굵기, 색상 지정
+   labs(x='월', y='강수량')+                                  # x축 레이블과 y축 레이블 지정
+   coord_flip( )                                              # 막대를 가로로 표시
```

- ggplot 패키지
  - 막대그래프의 작성 : geom_bar( ) 함수 이용

    - 막대그래프 꾸미기

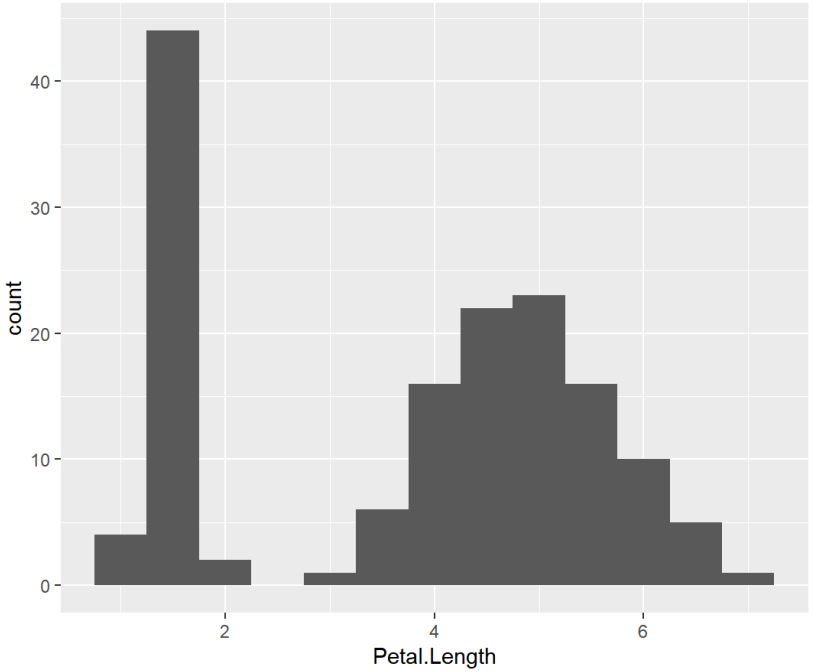


데이터 시각화

- ggplot 패키지
  - 히스토그램의 작성 : geom_histogram( ) 함수 이용

    - 기본적인 히스토그램 작성하기

```
> ggplot(iris, aes(x = Petal.Length)) +
+   geom_histogram(binwidth = 0.5)
```




```
# 그래프를 그릴 데이터 지정
# 히스토그램 작성
# binwidth : 구간의 길이를 지정하는 매개변수
   -> 꽃잎의 길이를 0.5 간격으로 나누라는 의미
```

데이터 시각화

- ggplot 패키지
  - 히스토그램의 작성 : geom_histogram( ) 함수 이용

    - 그룹별 히스토그램 작성하기

x축 : 꽃받침의 폭

막대의 윤곽선 색

```
> ggplot(iris, aes(x = Sepal.Width, fill = Species, color = Species)) +
```

막대 내부를 채울 색

```
+   geom_histogram(binwidth = 0.5, position = 'dodge') +
```

'dodge' : 3개의 막대가 겹치지 않고 병렬로 그려짐

```
+   theme(legend.position = 'top')
```

범례의 위치를 설정

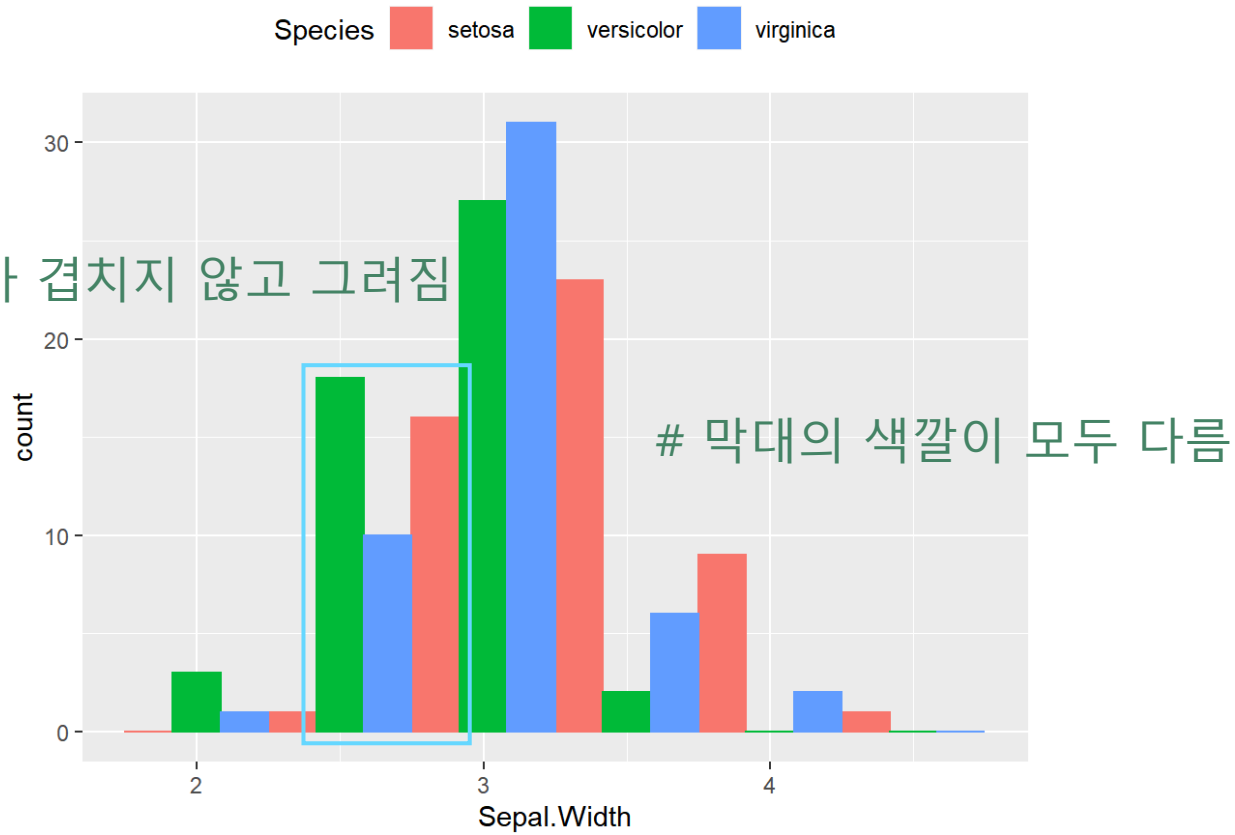
```
# fill = Species
  :  Species는 팩터(문자형 벡터)
-> 숫자 1, 2, 3으로 변환 가능
-> 품종별로 막대의 색이
   다르게 채워짐
```

데이터 시각화

- ggplot 패키지
  - 히스토그램의 작성 : geom_histogram( ) 함수 이용

    - 그룹별 히스토그램 작성하기

범례가 위에 위치



3개의 그래프가 겹치지 않고 그려짐

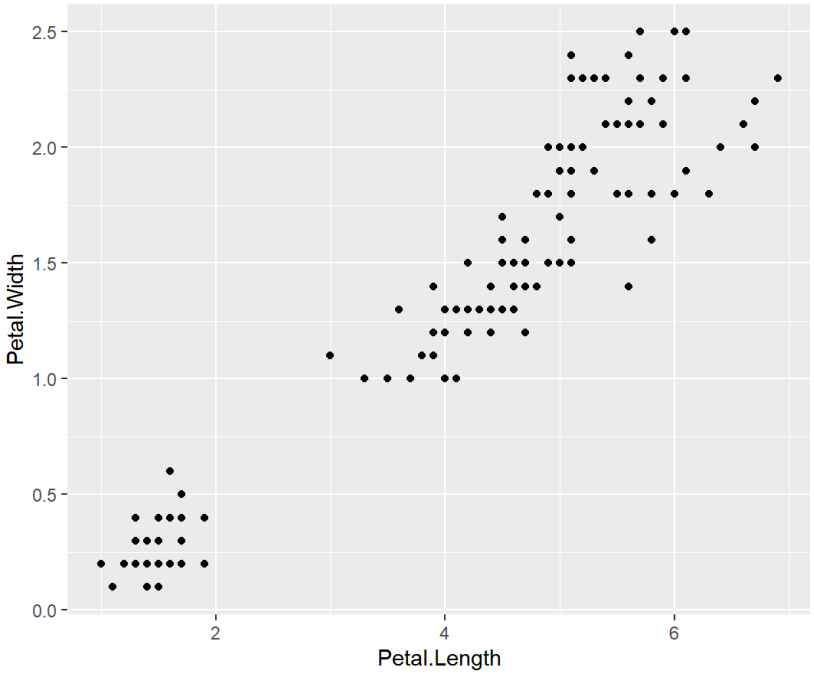
막대의 색깔이 모두 다름

데이터 시각화

- ggplot 패키지
  - 산점도의 작성 : goem_point( ) 함수 이용

    - 기본적인 산점도 작성하기

```
> ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width)) +        # 그래프를 그릴 데이터 지정
+    geom_point( )                                               # 산점도 작성
```



데이터 시각화

- ggplot 패키지
  - 산점도의 작성 : goem_point( ) 함수 이용

    - 그룹이 구분되는 산점도 작성하기

```
> ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width,
                        color=Species)) +
+   geom_point(size=3) +
+   ggtitle('꽃잎의 길이와 폭') +
+   theme(plot.title = element_text(size=25, face='bold', colour='steelblue'))
```

점의 색을 품종에 따라 다르게 함

산점도의 점의 크기를 설정

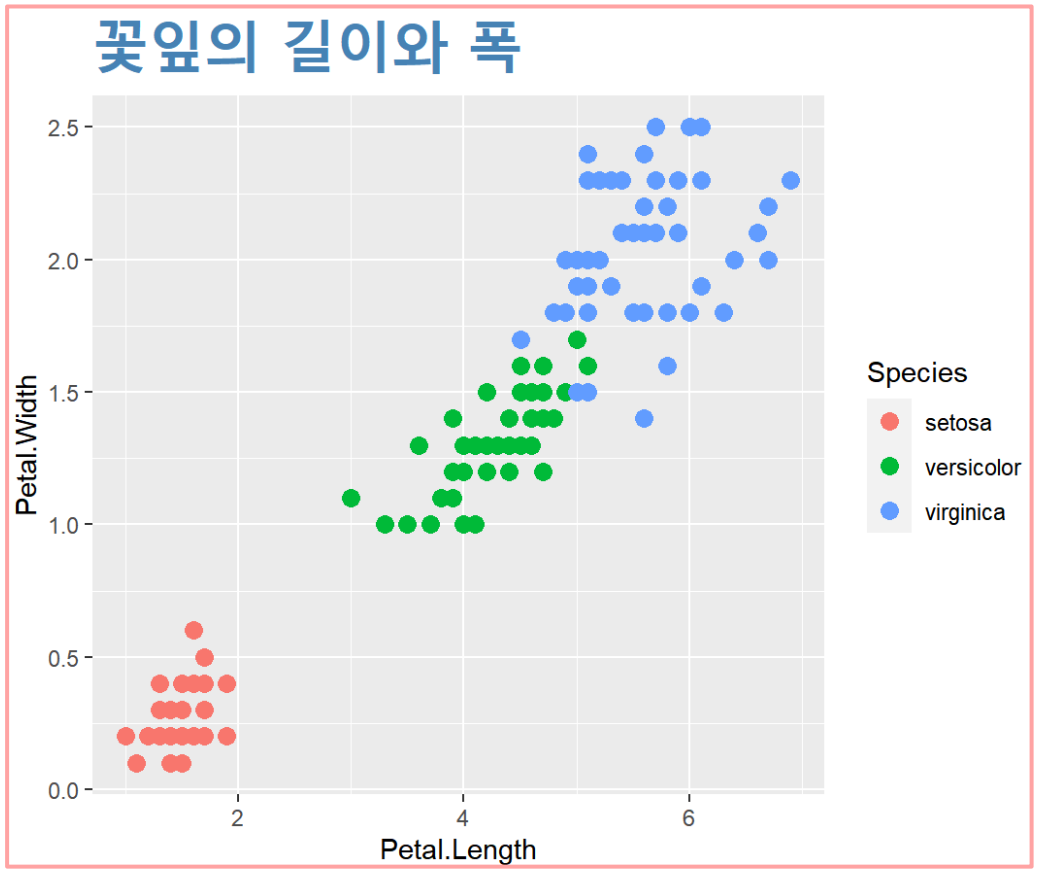
산점도의 제목

제목의 크기, 굵기, 색상

데이터 시각화

- ggplot 패키지
  - 산점도의 작성 : goem_point( ) 함수 이용

    - 그룹이 구분되는 산점도 작성하기



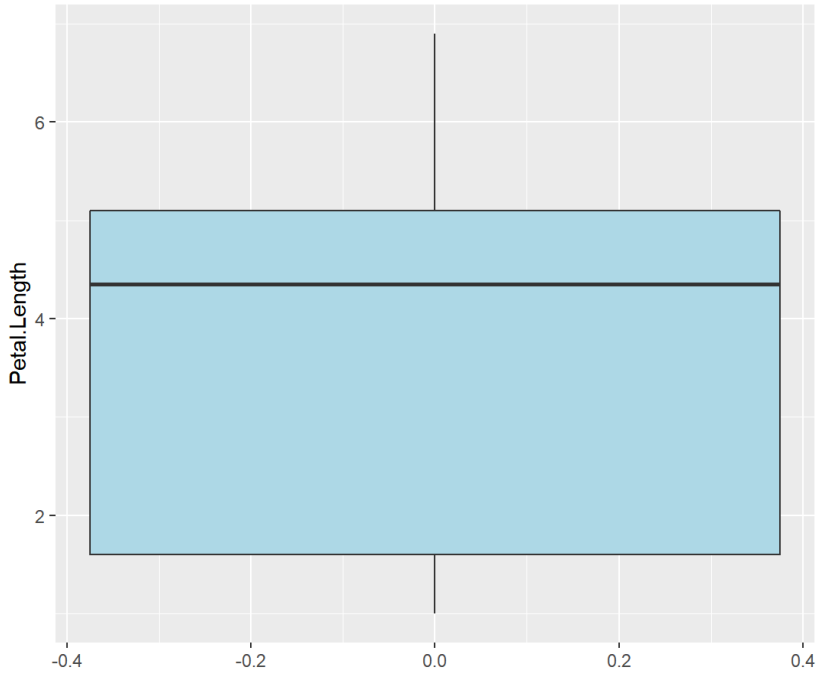
데이터 시각화

- ggplot 패키지
  - 상자그림의 작성 : geom_boxplot( ) 함수 이용

    - 기본적인 상자그림 작성하기

```
> ggplot(data=iris, aes(y=Petal.Length)) +
+    geom_boxplot(fill='lightblue')
```

```
# 그래프를 그릴 데이터 지정
# 상자그림 작성
# 배경색 : 연청색
```

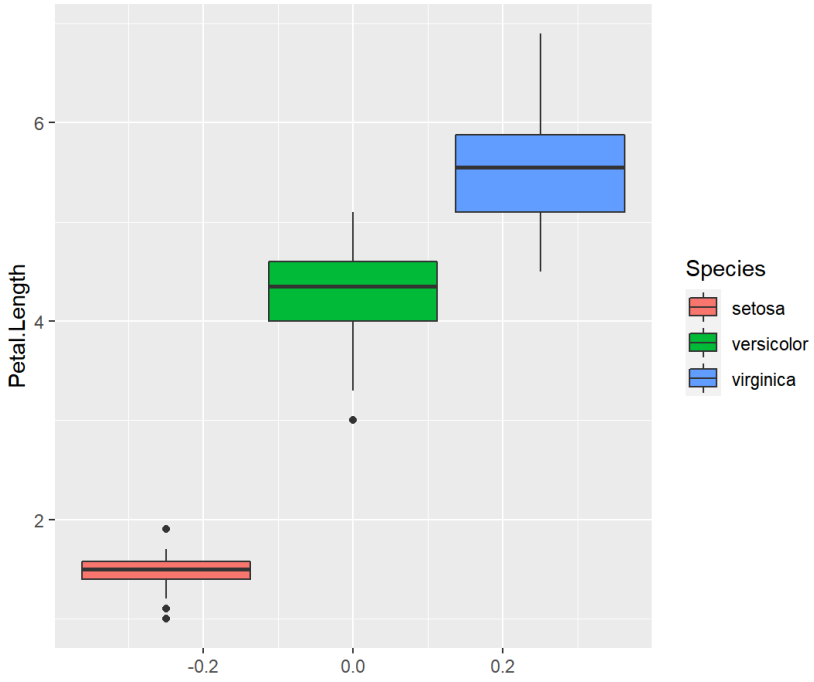


데이터 시각화

- ggplot 패키지
  - 상자그림의 작성 : geom_boxplot( ) 함수 이용

    - 그룹별 상자그림 작성하기

```
> ggplot(data=iris, aes(y=Petal.Length, fill = Species)) +# aes 안에 fill=Species추가
+   geom_boxplot()                                          -> Species의 그룹별로 상자그림 작성
```



데이터 시각화

- ggplot 패키지
  - 선그래프의 작성 : geom_line( ) 함수 이용

    - 1937년~1960년 사이 항공기에 의한 승객들의 이동거리 통계

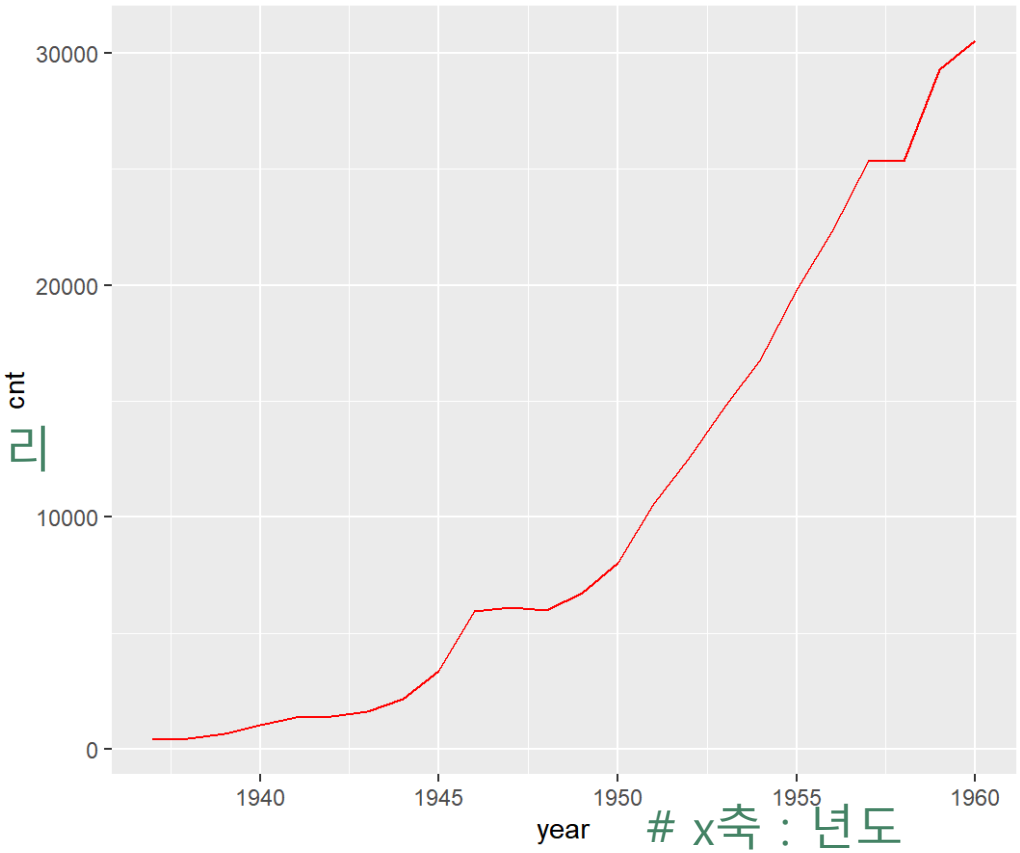
```
>   year <- 1937:1960                          # 1937년부터 1960년까지 데이터 저장
>   cnt <- as.vector(airmiles)                 # 항공기에 의한 승객들의 이동거리 데이터 저장
>   df <- data.frame(year, cnt)                # 데이터프레임에 저장하기
>   head(df)
 year  cnt
1 1937  412
2 1938  480
3 1939  683
...
> ggplot(data = df, aes(x = year, y = cnt)) + # 선그래프 작성
+   geom_line(col = 'red')
```

데이터 시각화

- ggplot 패키지
  - 선그래프의 작성 : geom_line( ) 함수 이용

    - 1937년~1960년 사이 항공기에 의한 승객들의 이동거리 통계

y축 : 항공기에 의한 승객들의 이동거리



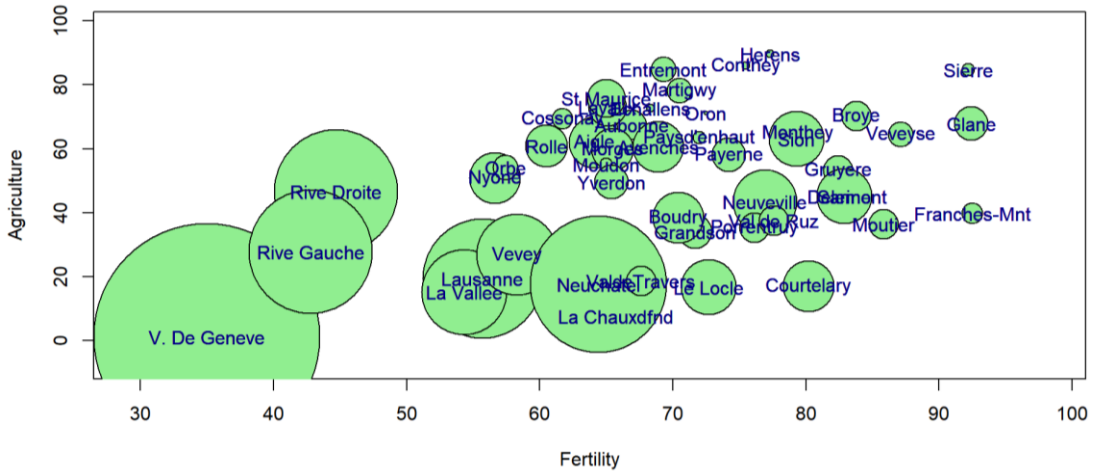
x축 : 년도

예시

- 예시 1 : swiss 데이터셋을 이용해 트리맵 작성하기

  - 실행 결과

```
> symbols(                                    ,   # x축 : Fertility, y축 : Agriculture
+           ,                                     # 원의 크기 : Education
+           bg = 'lightgreen',                    # 원의 색 : lightgreen
+           xlab = 'Fertility',                   # x축과 y축의 레이블이 각각 표시됨
+           ylab = 'Agriculture',
+           )
+  text(swiss$Fertility, swiss$Agriculture,       # 원 위에 주의 이름 표시(행의 이름)
+       rownames(swiss),
+      col = 'darkblue')                          # 폰트 색 :  darkblue
```

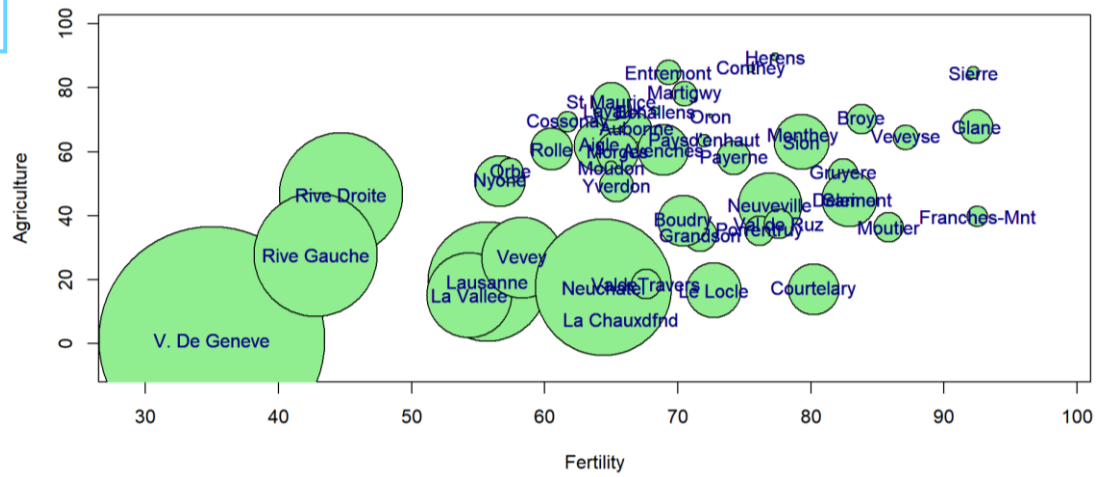


예시

- 예시 2 : swiss 데이터셋을 이용해 버블차트 작성하기
  - 실행 결과

```
> symbols(
+         circles =
+         bg =
+         xlab = 'Fertility',
+         ylab = 'Agriculture',
+         )
> text
+
+       col =
```

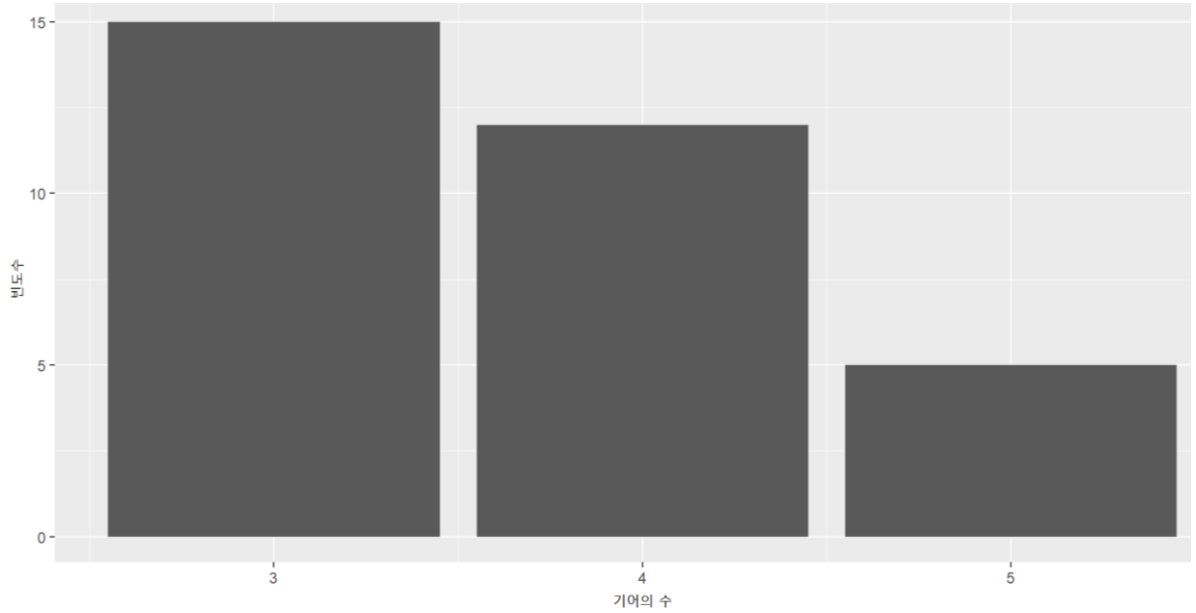
```
# x축 : Fertility, y축 : Agriculture
# 원의 크기 : Education
# 원의 색 : lightgreen
# x축과 y축의 레이블이 각각 표시됨

# 원 위에 주의 이름 표시(행의 이름)
# 폰트 색 :  darkblue
```



예시

- 예시 3 : ggplot을 이용해 mtcars 데이터셋 막대그래프 작성하기
  - 실행 결과

```
>    library(ggplot2)                          # ggplot2 패키지 실행
> ggplot(                          ) +         # x축 = gear
+    geom_bar(                    ) +          # 막대그래프 작성
+    labs(x = '기어의 수', y = '빈도수')        # x축 = '기어의 수', y축 = '빈도수' 레이블 표시
```

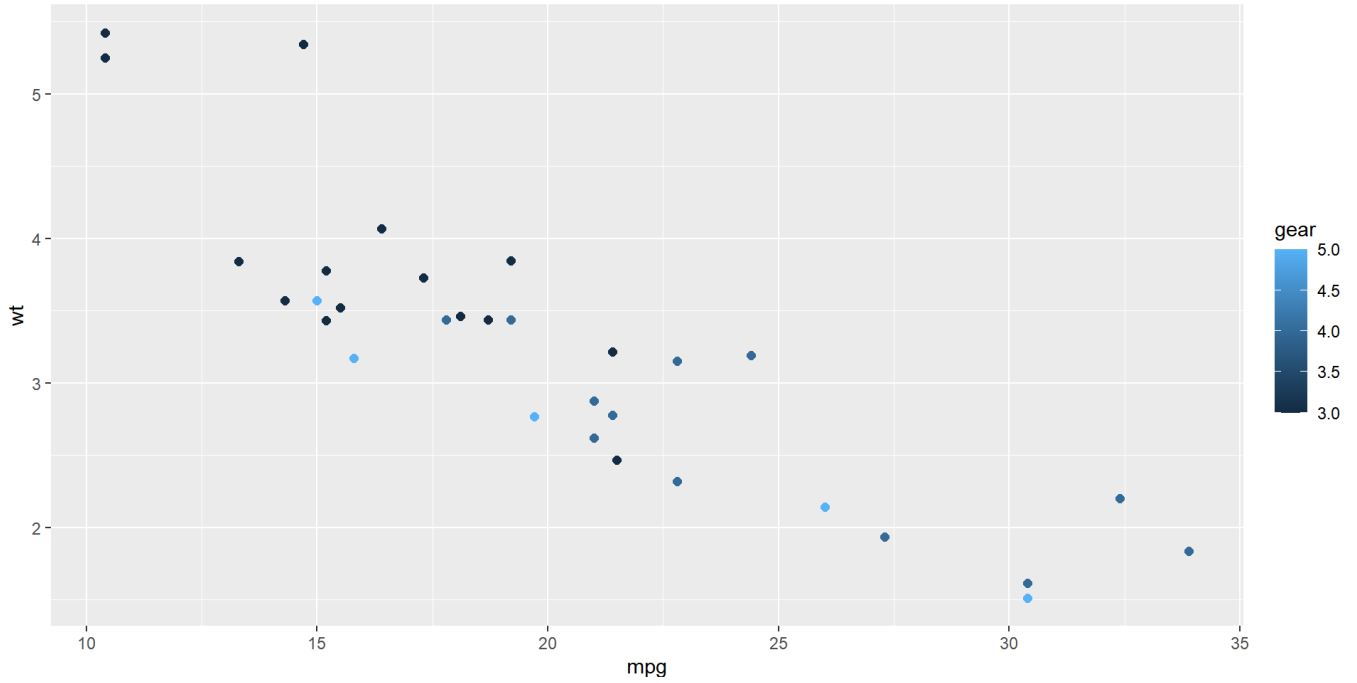


예시

- 예시 4 : mtcars를 이용해 산점도 작성하기
  - 실행 결과

```
> ggplot(                                    )
+                    +
+
```

```
# x축 = mpg, y축 = wt
# gear에 따라 점의 색을 다르게 함
# 산점도 작성
# 점의 크기 = 2
```



감사합니다

kimtwan21@dongduk.ac.kr

김 태 완