



7주차: 데이터 시각화

ChulSoo Park

School of Computer Engineering & Information Technology

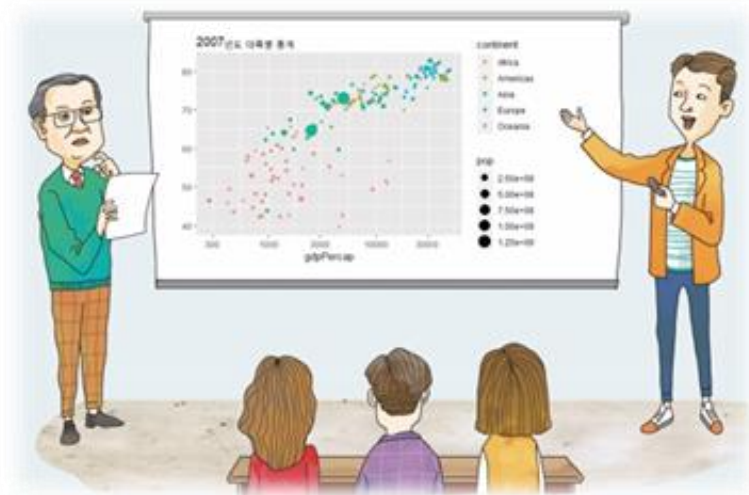
Korea National University of Transportation



06

CHAPTER

데이터 시각화



CONTENTS

- 6.1 데이터 시각화란?
- 6.2 시각화의 기본 기능
- 6.3 시각화 도구
- 6.4 시각화를 이용한 데이터 탐색

요약



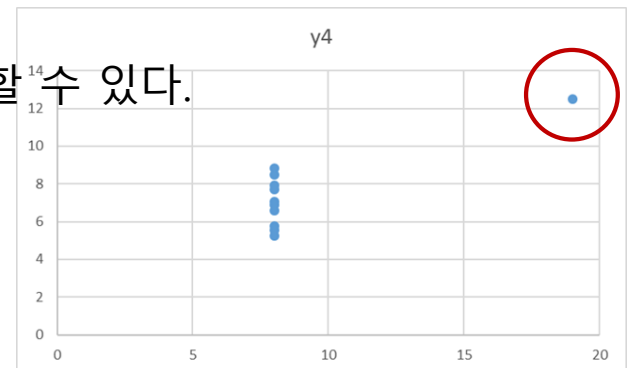
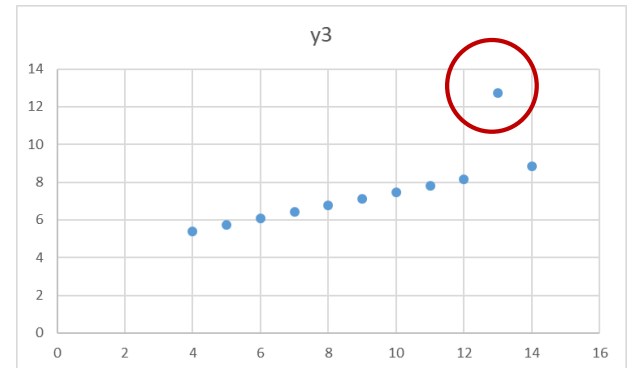
6.2 시각화의 기본 기능

■ 많은 량의 데이터를 효과적으로 관찰

- 시각화는 데이터를 올바르게 해석할 수 있게 해주는 동시에, 많은 양의 데이터를 효과적으로 관찰할 수 있게 해주는 역할을 한다.
- 최근의 데이터 과학은 신뢰도를 높이기 위해 점점 더 많은 데이터를 다루면서 복잡도도 더욱 높아졌다.

■ 시각화의 효과

- ✓ 직관(insight)을 얻을 수 있다.
- ✓ 핵심을 명확하게 이해할 수 있다.
- ✓ 평균적인 경향과 더불어 이상값(outlier)도 발견할 수 있다.
- ✓ 데이터에서 문제를 빨리 찾아낼 수 있다.






6.2 시각화의 기본 기능

■ 많은 량의 데이터를 효과적으로 관찰

■ gapminder 데이터의 직관적 이해(1)

- ✓ glimpse, str 같은 데이터 프레임 요약 함수를 이용해 데이터의 규모와 속성을 어느 정도 파악.
- ✓ 시각화를 할 수 있다면 요약 통계를 추출하는 과정 없이도 데이터를 직관적으로 이해할 수 있다.

```
Console C:/RSources/     
> str(gapminder)  
tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)  
$ country   : Factor w/ 142 levels "Afghanistan",...: 1 1 1  
1 1 1 1 1 1 1 ...  
$ continent: Factor w/ 5 levels "Africa","Americas",...: 3  
3 3 3 3 3 3 3 3 ...  
$ year      : int [1:1704] 1952 1957 1962 1967 1972 1977 19  
82 1987 1992 1997 ...  
$ lifeExp   : num [1:1704] 28.8 30.3 32 34 36.1 ...  
$ pop       : int [1:1704] 8425333 9240934 10267083 1153796  
6 13079460 14880372 12881816 13867957 16317921 22227415 ...  
$ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```



6.2 시각화의 기본 기능

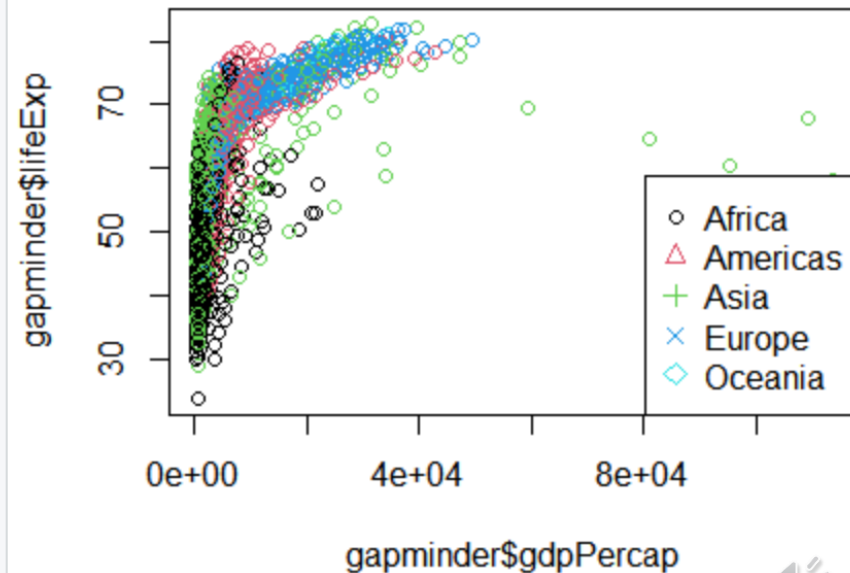
■ 많은 량의 데이터를 효과적으로 관찰

■ gapminder 데이터의 직관적 이해(2)

- ✓ 원 데이터의 속성들을 가능한 그대로 사용하여 모든 샘플들을 그래프에 표시하되 대륙 혹은 국가에 따라 구별된 마커를 사용해 gdpPercap, lifeExp, pop 항목의 범위와 특징, 상대적인 차이, 대략의 상관관계도 확인 가능

Console C:/RSources/

```
> plot(gapminder$gdpPercap, gapminder$lifeExp, col = gapminder$continent)
> legend("bottomright", legend = levels(gapminder$continent), pch = c(1:length(levels(gapminder$continent))), col = c(1:length(levels(gapminder$continent))))
> |
```



6.2 시각화의 기본 기능

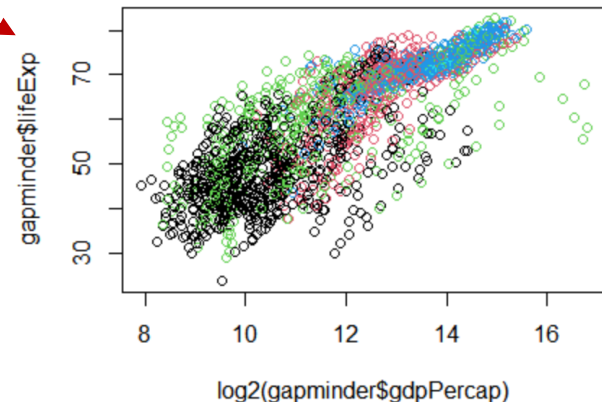
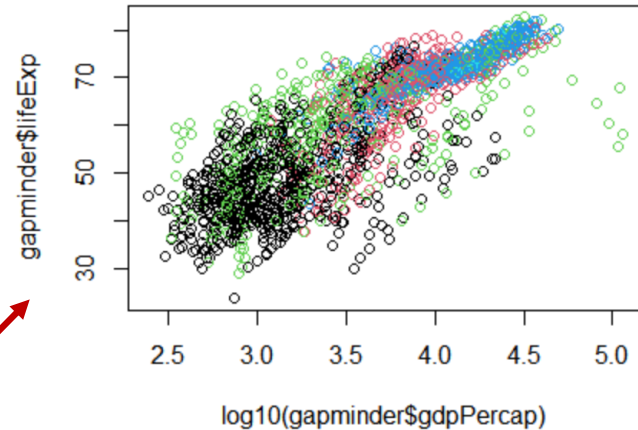
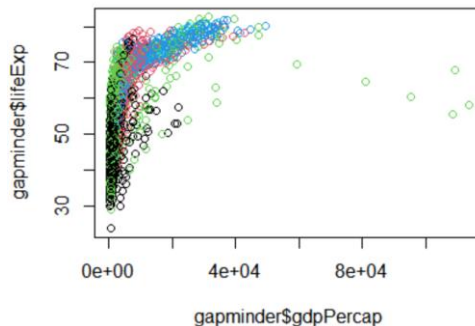
■ 많은 량의 데이터를 효과적으로 관찰

■ gapminder 데이터의 직관적 이해(3)

- ✓ gdpPercap 값의 전체 범위에 비해 낮은 범위에 샘플들이 많이 몰려 있어 관찰이 쉽지 않은 경우에는 로그 스케일(log scale)을 이용해 샘플들을 고르게 관찰 가능

```
Console C:/RSources/
> plot(log10(gapminder$gdpPercap), gapminder$lifeExp, col = gapminder$continent)
> plot(log2(gapminder$gdpPercap), gapminder$lifeExp, col = gapminder$continent)
> |
```

num	log2	log10
1	0.0	0.00
2	1.0	0.30
3	1.6	0.48
4	2.0	0.60
5	2.3	0.70
6	2.6	0.78
7	2.8	0.85
8	3.0	0.90
9	3.2	0.95
10	3.3	1.00
11	3.5	1.04
12	3.6	1.08
13	3.7	1.11
14	3.8	1.15
15	3.9	1.18
16	4.0	1.20
17	4.1	1.23
18	4.2	1.26
19	4.2	1.28
20	4.3	1.30



6.2 시각화의 기본 기능

■ 많은 량의 데이터를 효과적으로 관찰 (논문에서 log 사용 사례)

2) 유동인구 : 서울시의 10,000여 개 포인트에서 수집한 시간대별, 요일별 인구수의 일평균 값을 해당 지역의 대리점과 행정코드로 연결하여 각 수집 포인트의 유동인구 합을 유동인구 투입변수의 값으로 산출하였다. 대리점별로 유동인구의 정규성을 검사한 결과 Kolmogorov-Smirnov 검정에서 유의확률이 0.00이고 왜도(skewness)가 2.44로 높아 정적편포(positively skewed distribution)로 나타나 정규분포를 가지도록 유동인구 수에 자연로그를 취하여 사용했다.

- 자료의 대칭성을 알아보는 측도

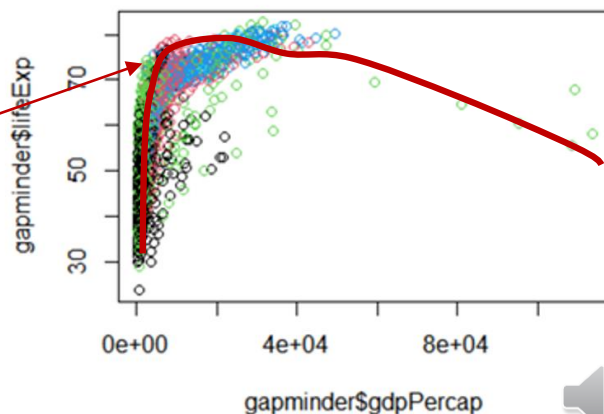
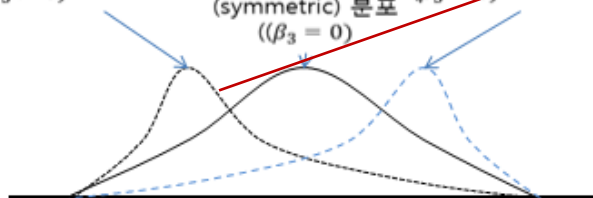
$$\text{Skewness} = \beta_3 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3$$

왜도
(skewness)

오른쪽으로 꼬리가 긴
(right-skewed) 분포
($\beta_3 > 0$)

좌우 대칭
(symmetric) 분포
($\beta_3 = 0$)

왼쪽으로 꼬리가 긴
(left-skewed) 분포
($\beta_3 < 0$)

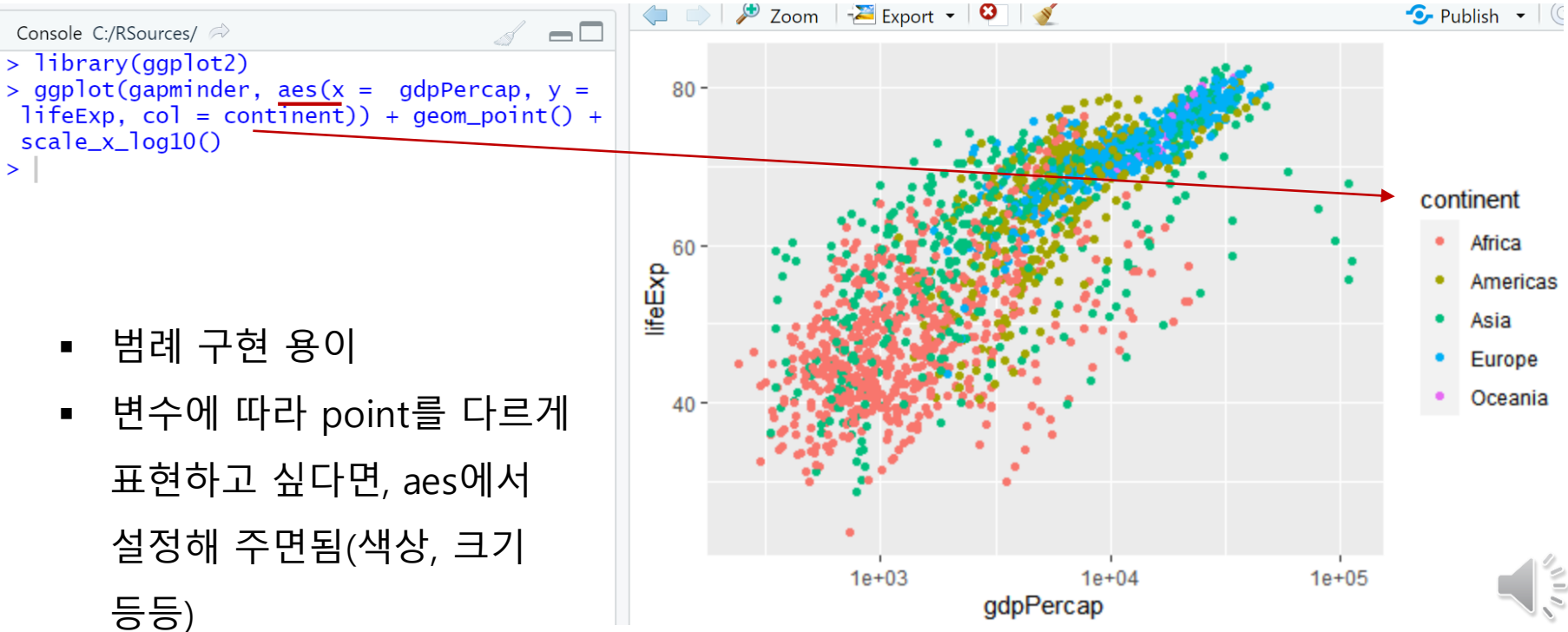


6.2 시각화의 기본 기능

■ 많은 량의 데이터를 효과적으로 관찰

■ gapminder 데이터의 직관적 이해(4)

- 베이스 R의 plot 함수를 이용해 기본적인 시각화가 가능하지만,
- 시각화 전용 라이브러리인 ggplot2를 사용하면 그래프의 추가 옵션을 간단히 지정할 수 있을 뿐 아니라 완성도 높은 시각화 결과를 훨씬 쉽게 얻을 수 있음.

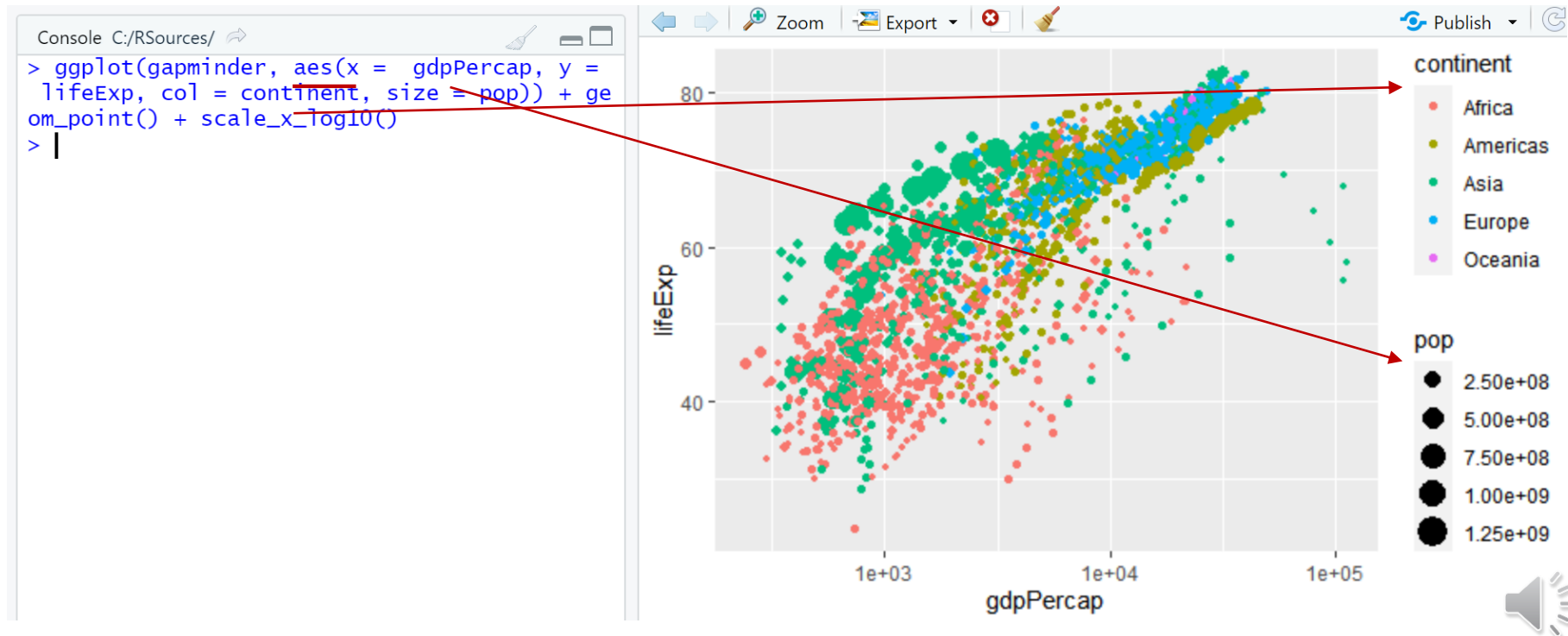


- 범례 구현 용이
- 변수에 따라 point를 다르게 표현하고 싶다면, aes에서 설정해 주면됨(색상, 크기 등등)

6.2 시각화의 기본 기능

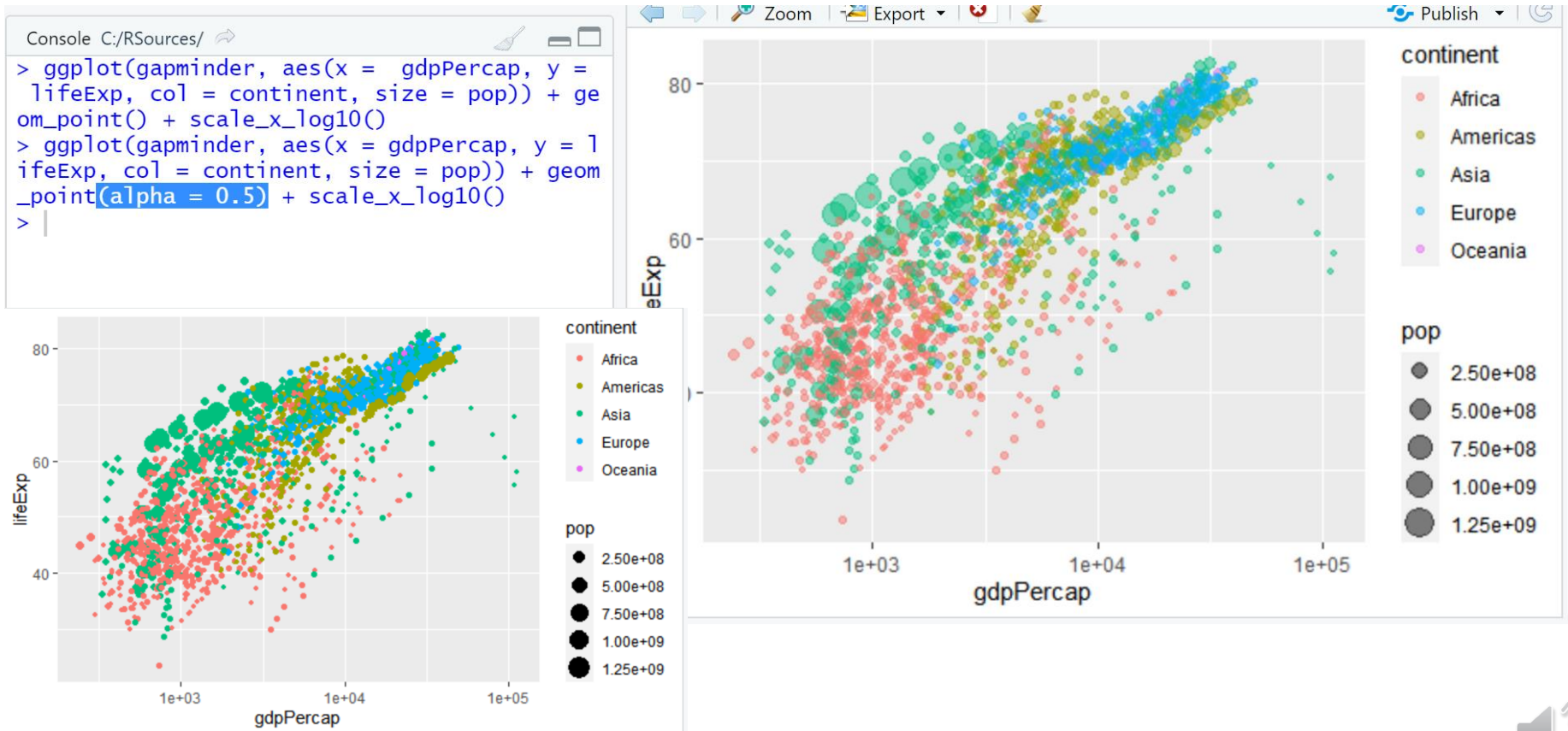
■ gapminder 데이터의 직관적 이해(5)

- ✓ ggplot 함수는 size = pop을 추가함으로써 플롯 마커의 크기가 각 국가의 인구에 비례하도록 지정 가능
- ✓ gplot2에서 제공하는 size 옵션을 활용하여 pop 항목도 하나의 그래프에 표시할 수 있게 됨으로써 다양한 속성들의 상호 관계를 쉽게 파악 가능



6.2 시각화의 기본 기능

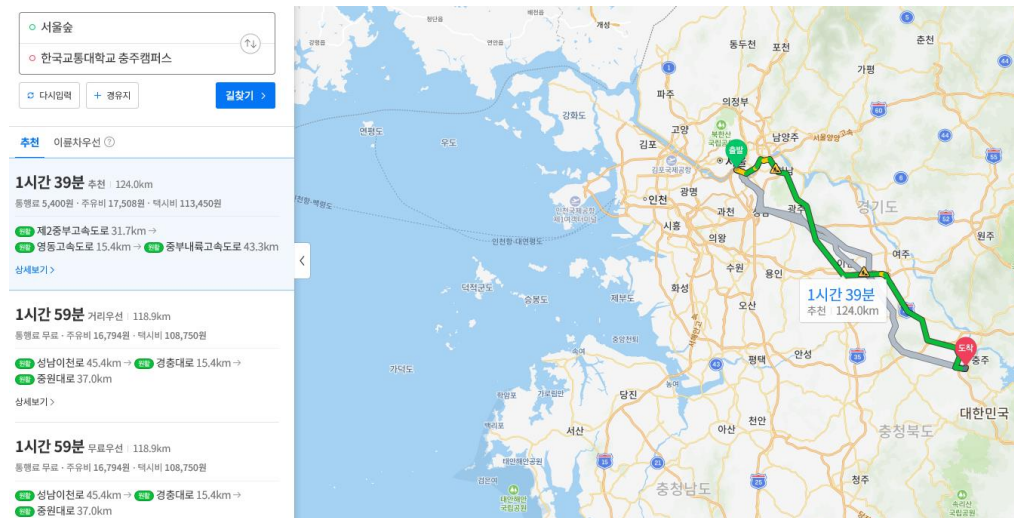
- gapminder 데이터의 직관적 이해(6) : 마커의 중첩 해결을 위한 투명도 설정 사용 예제



6.2 시각화의 기본 기능

■ gapminder 데이터의 직관적 이해(7)

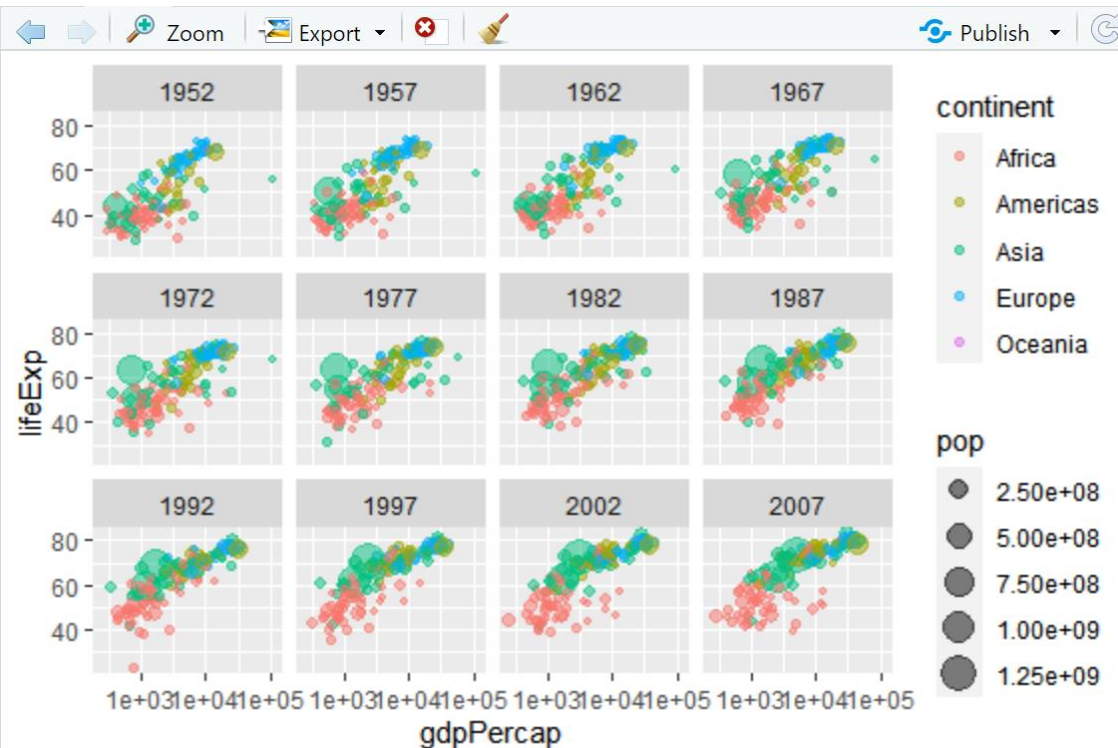
- 데이터를 정교하게 시각화하려면 관측 연도를 구분하여 표시하는 것이 더 효과적이다.
 - ✓ dplyr 라이브러리의 filter 함수를 이용해 각 연도의 데이터를 차례로 추출한 후 그래프를 반복하여 그리는 방법도 있으나,...
- ggplot2에서 제공하는 facet_wrap 함수를 이용하면 데이터 가공과 반복을 위한 프로그래밍을 간단히 대신할 수 있다.



6.2 시각화의 기본 기능

- gapminder 데이터의 직관적 이해(7)
 - ggplot2에서 제공하는 **facet_wrap** 함수를 이용

```
52:1 (top Level) R Script  
Console C:/Rsources/  
> ggplot(gapminder, aes(x=gdpPercap, y=lifeExp, col=continent, size=pop)) + geom_point(alpha=0.5) + scale_x_log10() + facet_wrap(~year)  
>  
> |
```



6.2 시각화의 기본 기능

■ 데이터를 여러 관점에서 시각화

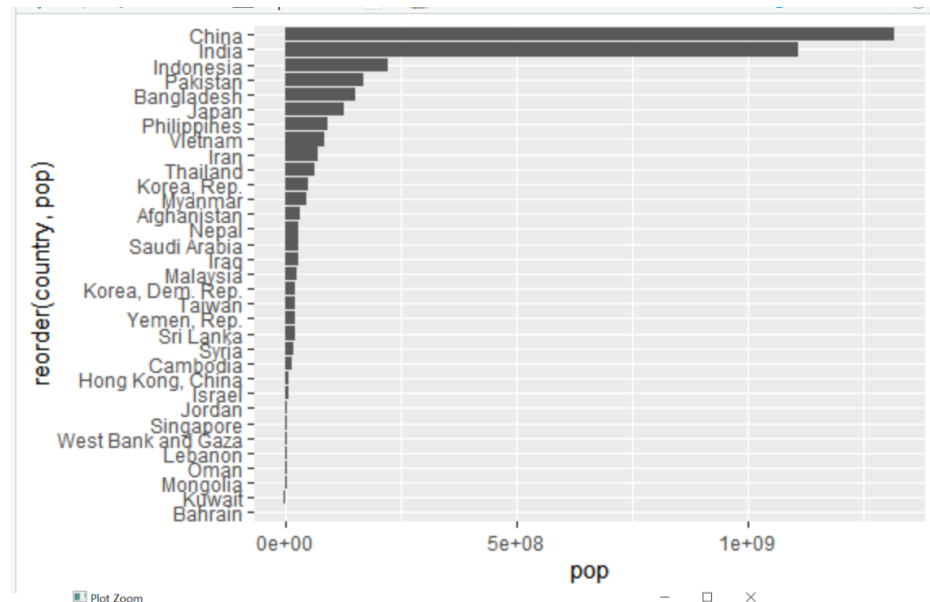
- 시각화의 공통 목적은 데이터에 내재된 의미, 즉 변화 · 구성 · 분포 · 상관관계 등을 명확히 드러내는 것이다.
- 인간의 인지 능력에는 한계가 있기 때문에 데이터에 포함된 '모든 변수의 모든 변화'를 한 번에 확인하는 것은 불가능하다.
- 따라서 데이터의 시각화는 반복적으로, 또 여러 관점에서 시도되어야 한다. 데이터를 바라보는 시각화 관점에 변화를 주면 데이터에 포함된 여러 가지 의미에 대한 통찰이 생긴다.
- 다양한 시각화 방법은 데이터 과학의 핵심 기술이다



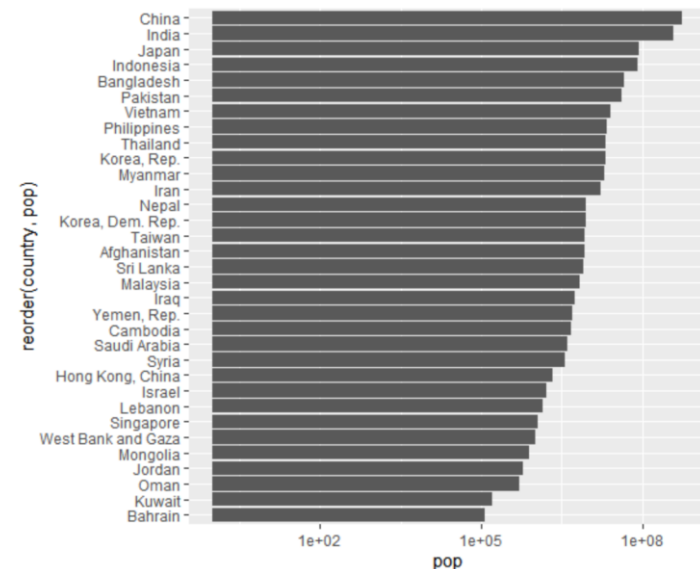
6.2 시각화의 기본 기능

■ 비교/순위(1)

- 1952년 아시아 대륙의 인구 분포에서 각 국가의 순위를 매겨보자.
- 겹치는 문제 해결을 위해 가로,세로 축 위치 바꾸기: `coord_flip`
- 상대적인 차이 해소를 위한 로그스케일 축 사용: `scale_y_log10`



Plot Zoom



Console C:/RSources/

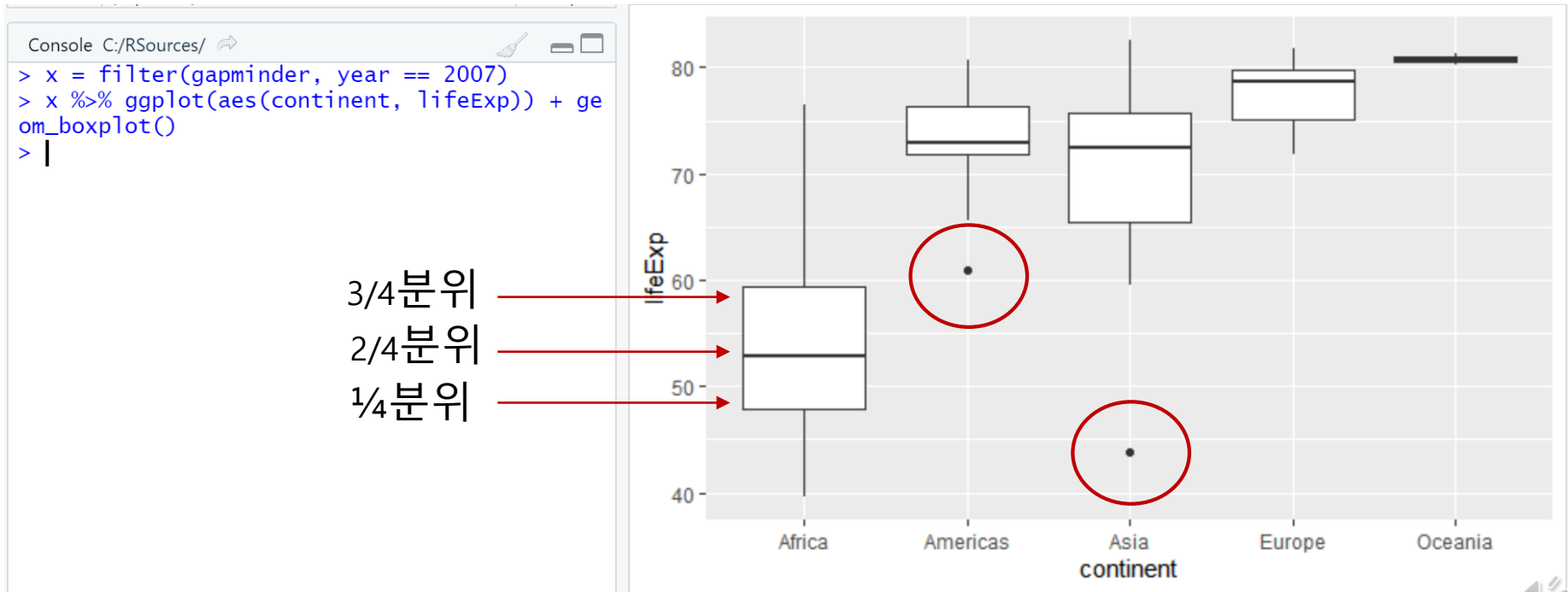
```
> gapminder %>% filter(year == 2007 & continent == "Asia") %>% ggplot(aes(reorder(country, pop), pop)) + geom_bar(stat = "identity") + coord_flip()
> gapminder %>% filter(year==1952 & continent== "Asia") %>% ggplot(aes(reorder(country, pop), pop)) + geom_bar(stat = "identity") + scale_y_log10() + coord_flip()
```



6.2 시각화의 기본 기능

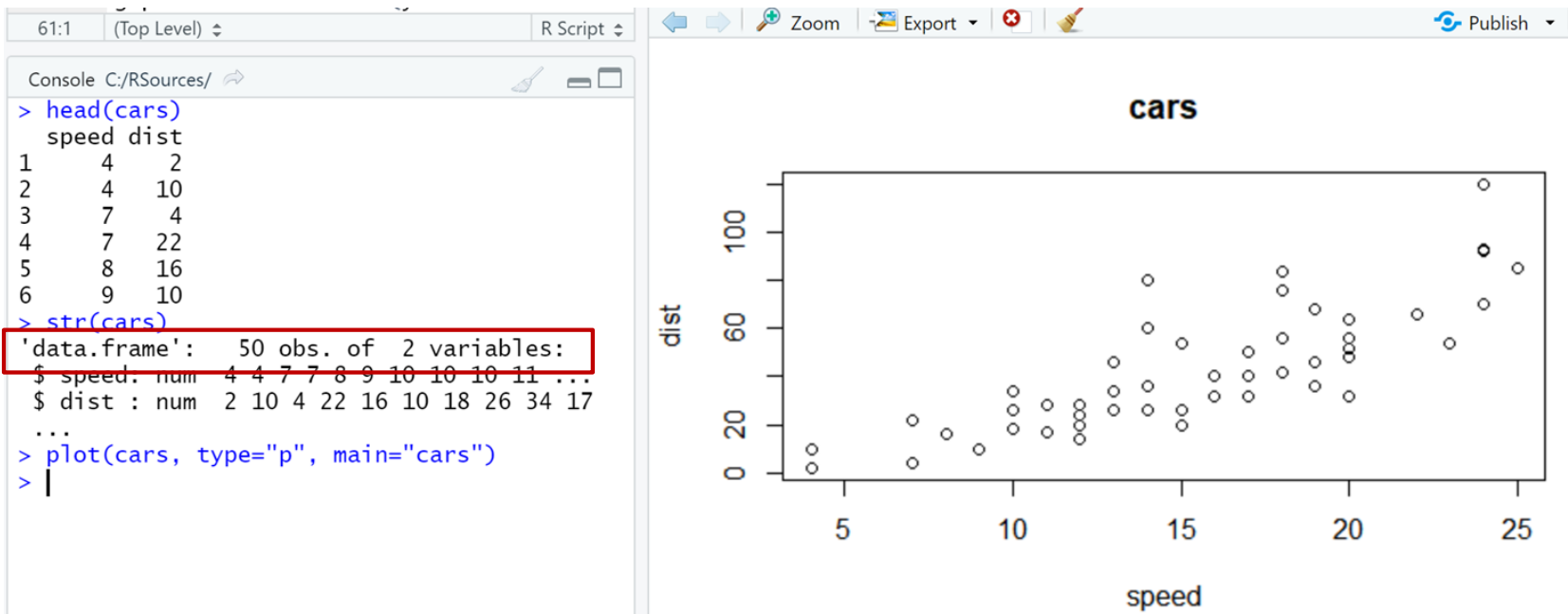
■ 분포 혹은 구성 비율 : boxplot 사용

- 앞의 두 그래프 모두 국가나 대륙의 구분 없이 해당 연도 모든 국가의 기대 수명을 종합한 분포를 보여준다.
- boxplot 함수를 이용하면 대륙별로 세분화된 분포 특성을 동시에 살펴볼 수 있다.



■ 베이스 R : 시각화 관련 함수의 체계적 정리

- R에 내장된 기본적인 시각화 함수로 시각화를 쉽게 할 수 있음
- 추가 옵션을 활용하여 그래프의 세부적인 설정 가능
- 데이터의 type과 필요에 따라 기본과 추가 Option을 병행 사용
- **Plot 함수** : 가장 일반적인 그래프 시각화 함수(점, 선 등의 형태 그래프)
- 베이스 R에 내장되어 있는 cars 데이터를 이용



6.3 시각화 도구

■ 베이스 R : 시각화 관련 함수의 체계적 정리

- Plot의 type
- `plot(cars, type="l", main = "cars")` # type="l"은 line

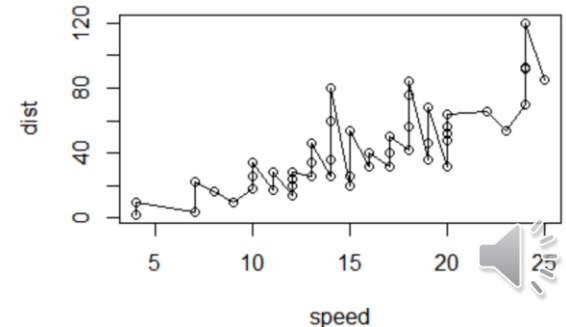
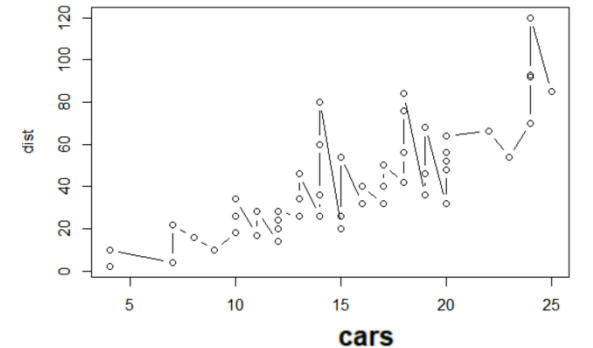
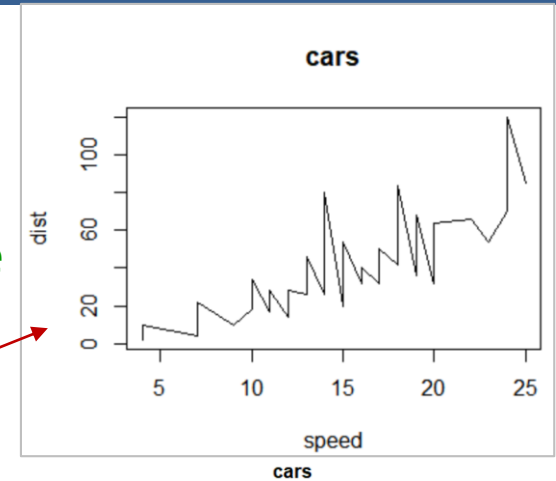
R: Generic X-Y Plotting ▾

Find in Topic

type

what type of plot should be drawn. Possible types are

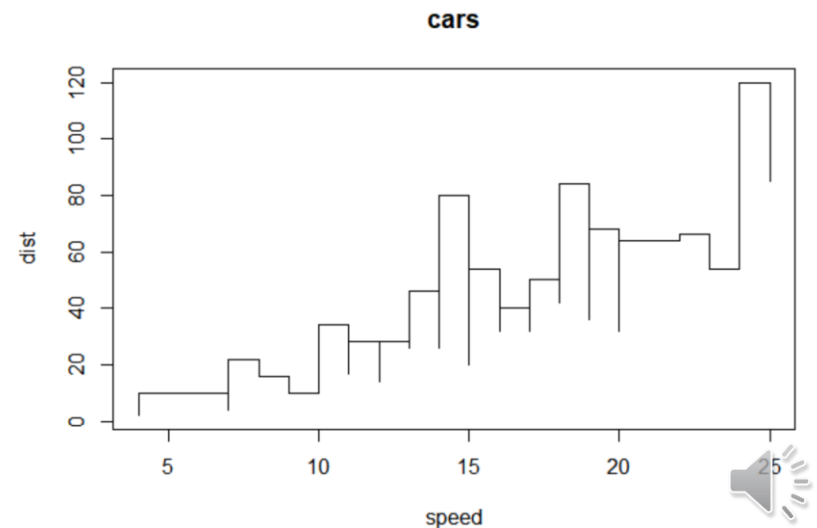
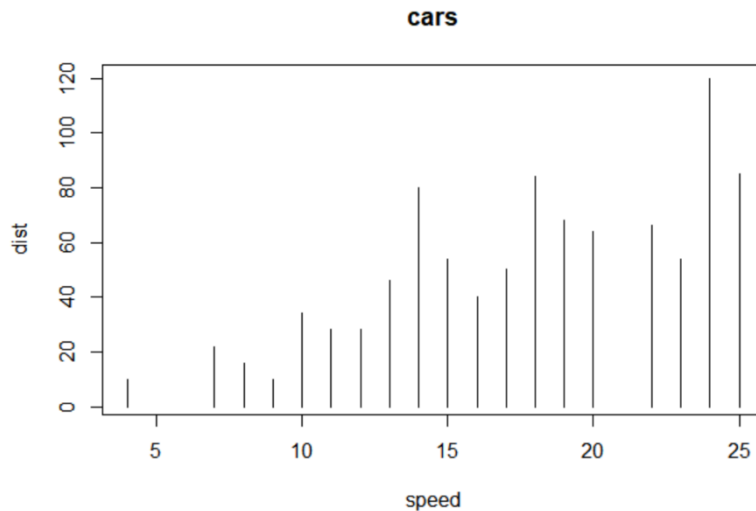
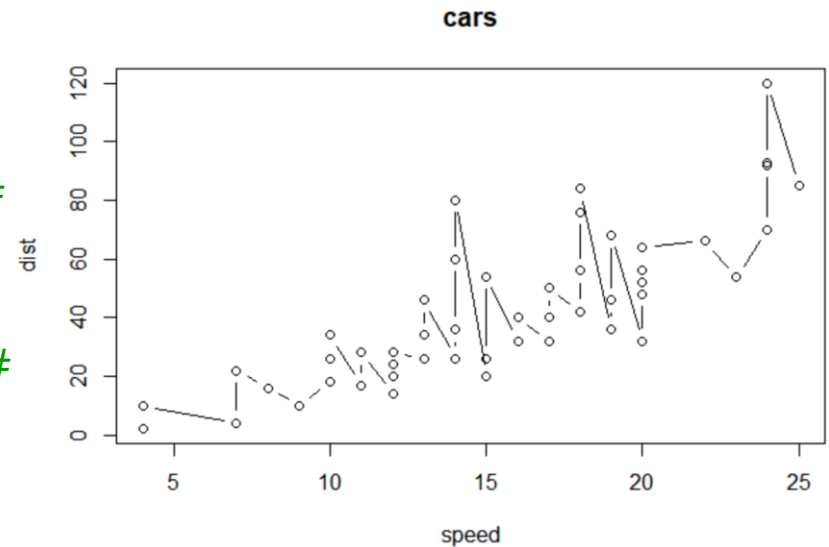
- "p" for **p**oints,
- "l" for **l**ines,
- "b" for **b**oth,
- "c" for the lines part alone of "b",
- "o" for both **o**verplotted,
- "h" for **h**istogram like (or 'high-density') vertical lines,
- "s" for stair **s**teps,
- "S" for other **s**teps, see 'Details' below,
- "n" for no plotting.



6.3 시각화 도구

■ 시각화 관련 함수의 체계적 정리

- Plot의 type
- `plot(cars, type="b", main = "cars")` #
type="b"는 점과 선을 모두 사용
- `plot(cars, type="h", main = "cars")` #
type="h"는 히스토그램과 같은 막대그래프
- `plot(cars, type="s", main = "cars")` #

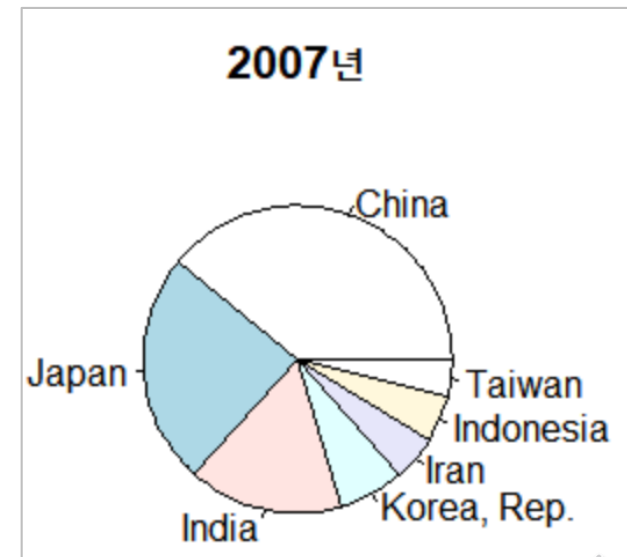
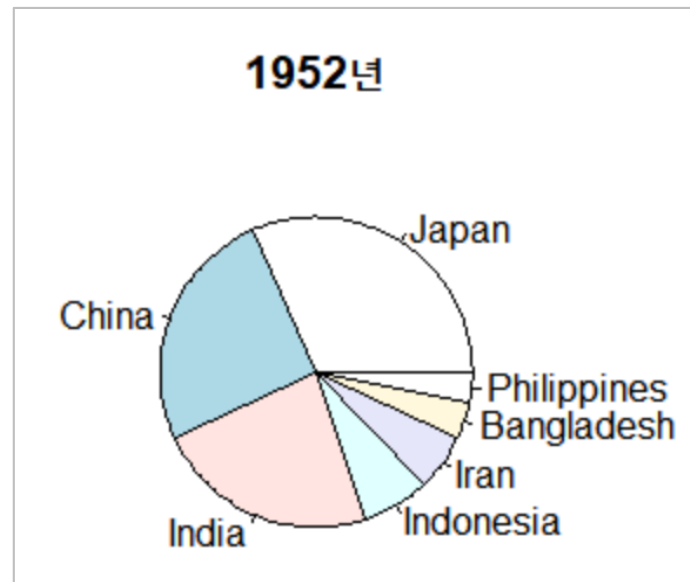


6.3 시각화 도구

■ 베이스 R : 시각화 관련 함수의 체계적 정리

- pie·barplot 함수 : 구성 비율, 순위 등을 시각적으로 확인할 때 유용
- `x = gapminder %>% filter(year == 1952 & continent == "Asia") %>% mutate(gdp = gdpPercap*pop) %>% select(country, gdp) %>% arrange(desc(gdp)) %>% head(7)`
- `pie(xgdp, xcountry, main="1952년")`

Pie 함수를 이용해 시각화한 1952, 2007년 아시아 국가들의 구성 순위

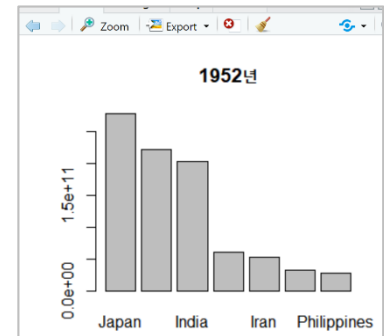


6.3 시각화 도구

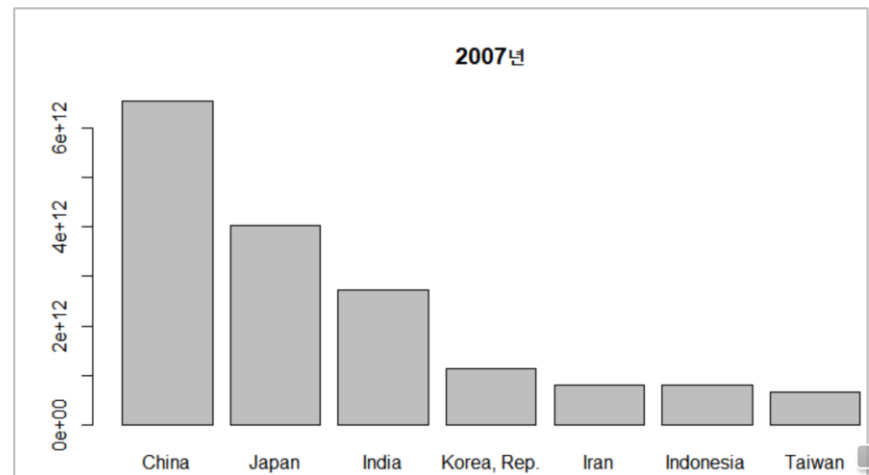
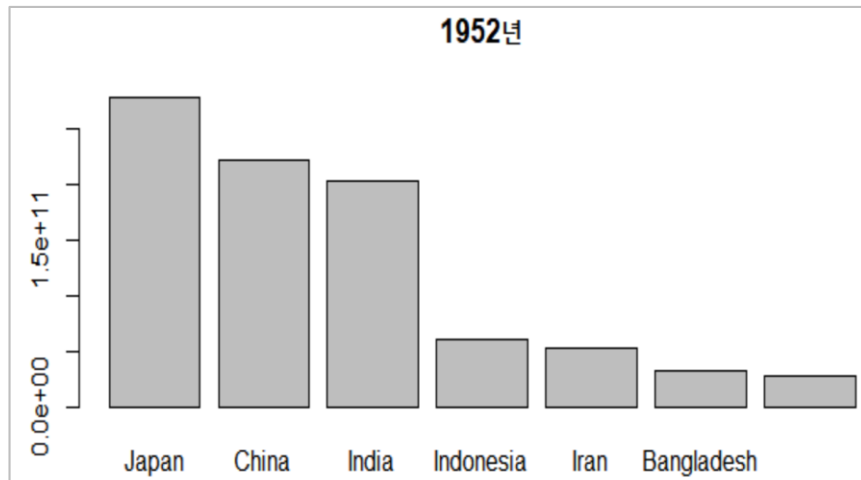
■ 베이스 R : 시각화 관련 함수의 체계적 정리

- pie.barplot 함수 : 구성 비율, 순위 등을 시각적으로 확인할 때 유용

```
Console C:/RSources/
> x = gapminder %>% filter(year == 2007 & continent == "Asia") %>% mutate(gdp = gdpPercap*pop) %>% select(country, gdp) %>% arrange(desc(gdp)) %>% head()
> pie(x$gdp, x$country)
> barplot(x$gdp, names.arg = x$country)
```



barplot 함수를 이용해 시각화한 1952,2007년 아시아 국가들의 구성 순위



■ 베이스 R : 시각화 관련 함수의 체계적 정리

- 대상 data set : iris : str, head 함수로 data 정보 파악 : `data()`
- `str(iris)`, `head(iris,7)`

```
Console C:/RSources/ ↗
> 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 ...
```

```
> head(iris,7)
  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
7          4.6          3.4          1.4          0.3  setosa
```

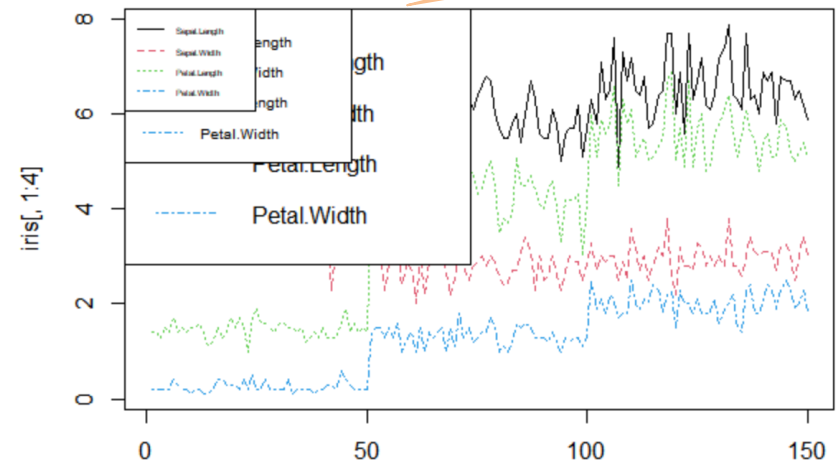
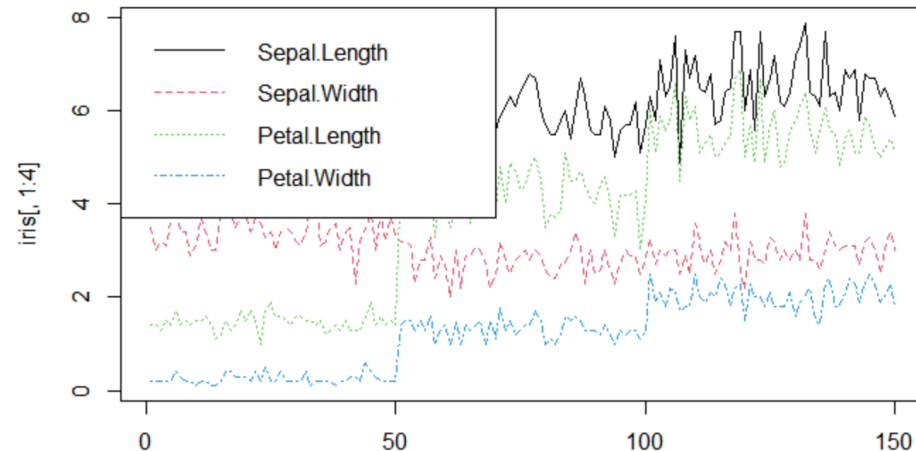


6.3 시각화 도구

■ 베이스 R : 시각화 관련 함수의 체계적 정리

- `matplot` 함수 : 벡터나 행렬 데이터를 이용한 다중 플롯을 빠르게 구현함
- `Legend` 함수를 이용한 범례 지정
- `matplot(iris[, 1:4], type='l')`
- `legend("topleft", names(iris)[1:4], lty = c(1, 2, 3, 4), col = c(1, 2, 3, 4), cex=0.6)`

cex = 0.4, 0.6, 1



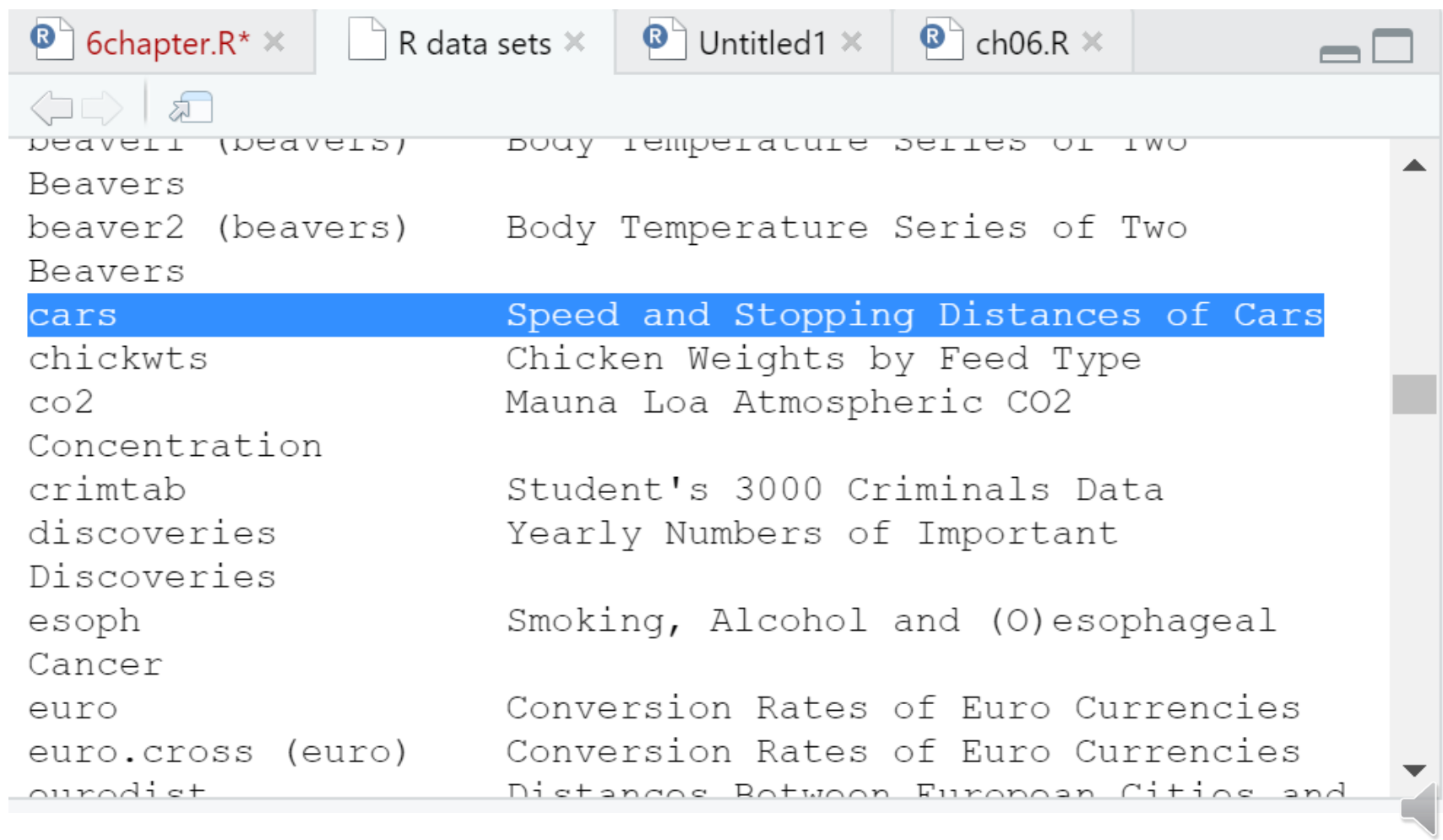
■ 베이스 R : 시각화 관련 함수의 체계적 정리

- 대상 data set : cars : str, head 함수로 data 정보 파악 : `data()`
- `str(cars), head(cars,10)`

```
Console C:/RSources/ ↗  
> str(cars)  
'data.frame':   50 obs. of  2 variables:  
 $ speed: num  4 4 7 7 8 9 10 10 10 11 ...  
 $ dist : num  2 10 4 22 16 10 18 26 34 17 ..  
> head(cars,10)  
  speed dist  
1     4    2  
2     4   10  
3     7    4  
4     7   22  
5     8   16  
6     9   10  
7    10   18  
8    10   26  
9    10   34  
10   11   17
```



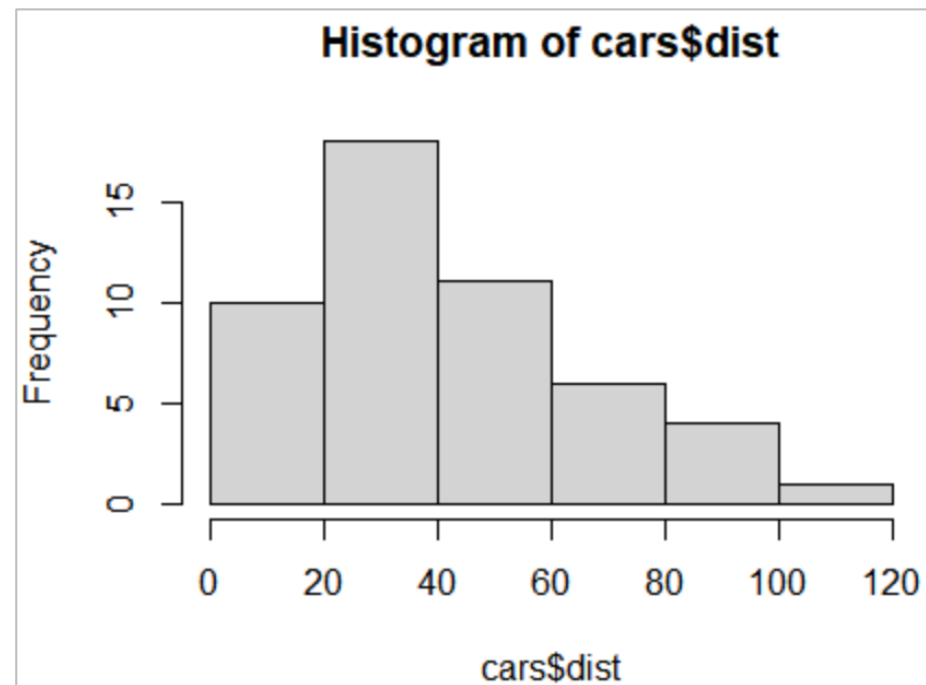
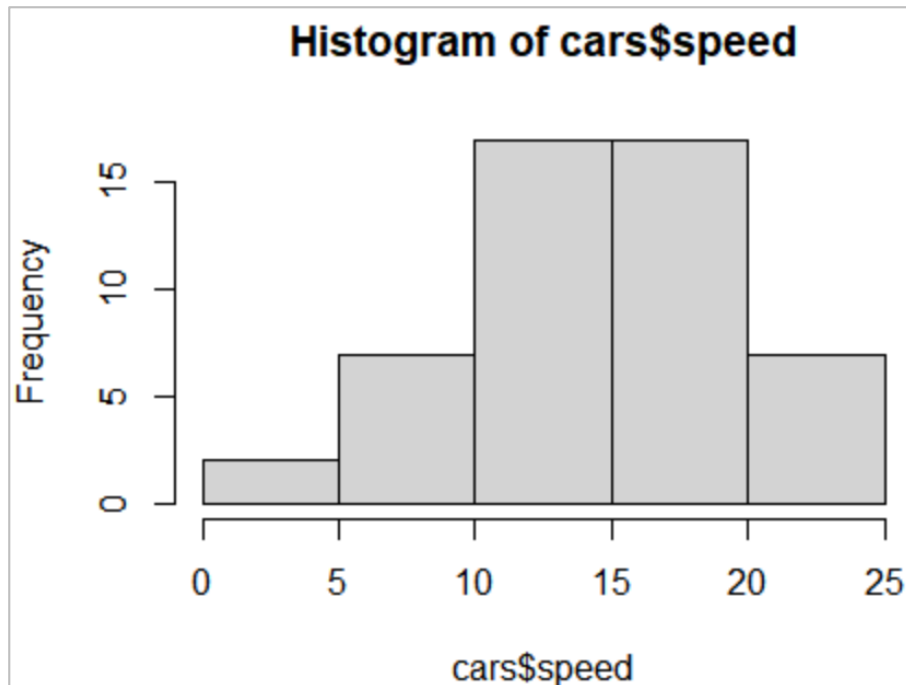
```
>data()
```



6.3 시각화 도구

■ 베이스 R : 시각화 관련 함수의 체계적 정리

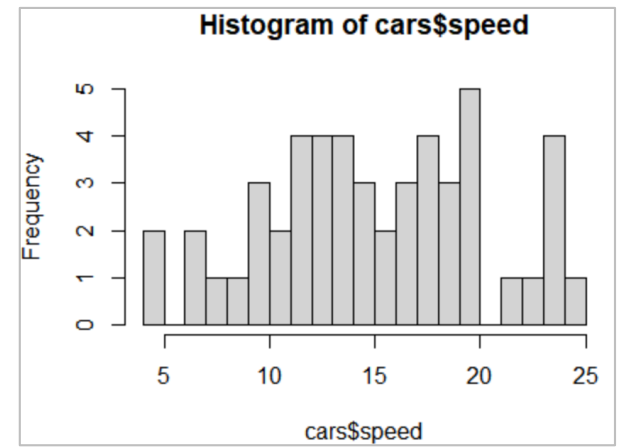
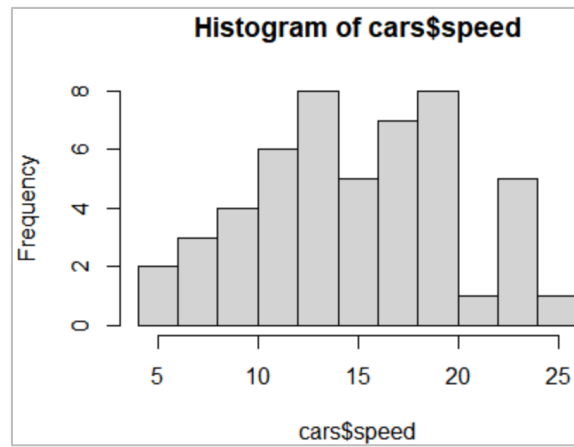
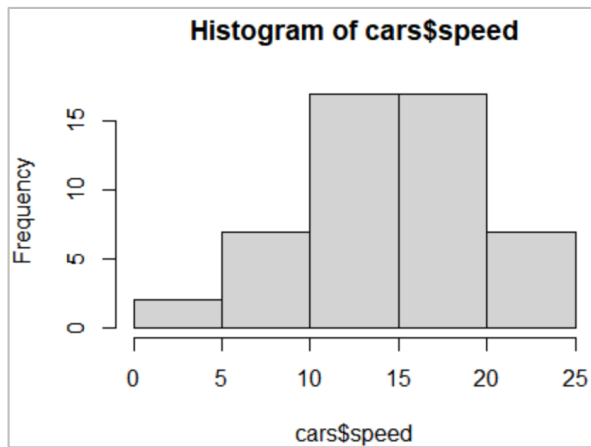
- 대상 data set : cars : str, head 함수로 data 정보 파악 : `data()`
- `hist(cars$speed), hist(cars$dist)`



6.3 시각화 도구

■ 베이스 R : 시각화 관련 함수의 체계적 정리

- 대상 data set : cars : str, head 함수로 data 정보 파악 : `data()`
- `hist(cars$speed), hist(cars$dist)`
- `hist(cars$speed, breaks=10)`
- `hist(cars$speed, breaks=15)`



Thank you

