7주차:
데이터 시각화

ChulSoo Park

School of Computer Engineering & Information Technology

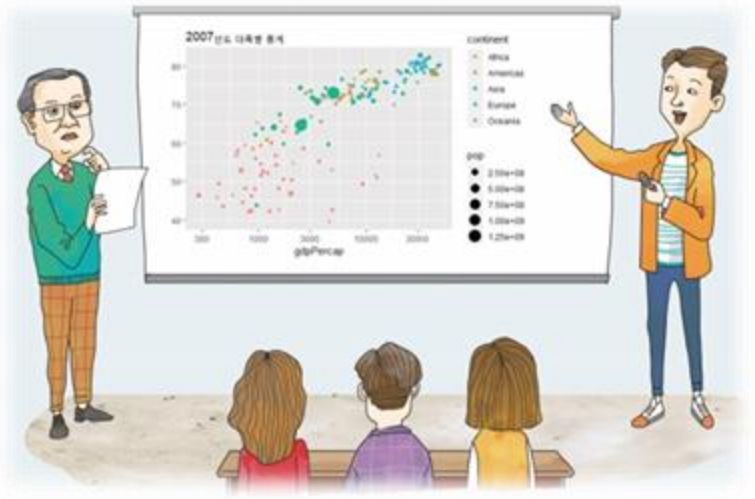
Korea National University of Transportation

학습목표 (7주차)

- 데이터 시각화 Library 사용법 이해
- 데이터 시각화 기본 기능 숙지
- 시각도 도구 숙지
- 시각화를 이용한 데이터 해석

06 CHAPTER

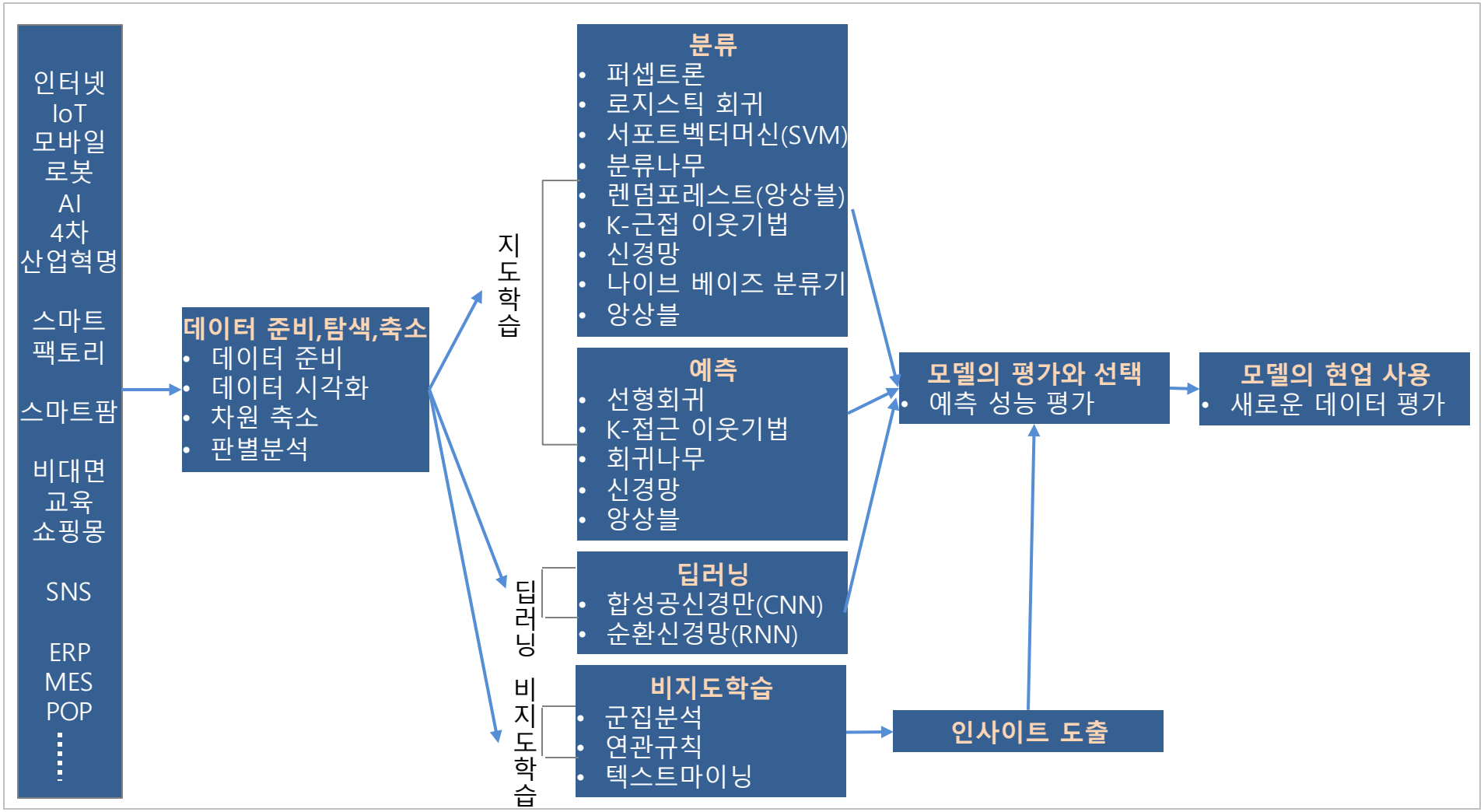
데이터 시각화



CONTENTS

Review

개론 7장 데이터 사이언스 방법 및 분석 기법



Review

관련 산업 트랜드

- **(텍스트코딩)** 파이썬을 활용한 온라인 코딩파티 콘텐츠 구현 및 내용·기능 보완

① 파이썬 데이터 분석과 시각화	
• **(권장연령)** 중학생 이상 • **(학습 프로그래밍 언어)** 파이썬 • **(내용)** 데이터 분석에 필요한 파이썬 기본 문법을 학습하고 파이썬을 활용해 데이터를 읽고 분석하는 기초적인 학습 진행 • **(인증서)** 각 스테이지 완료 시 발급(2회) • **(사용기기)** PC, 태블릿(웹브라우저) • **(교사용 수업지원도구)** 제공 필요* * 교사가 본 콘텐츠를 활용하여 수업을 진행할 수 있도록 사용법, 관련 배경 지식 등을 담은 별도 참고자료 제작	

6.3 시각화 도구

시각화에 특화된 ggplot2 라이브러리

- ggplot2는 R에서 가장 많이 사용되는 시각화 library
- gg는 grammar of graphics를 뜻함
- 사용할 data set : gapminder : str(gapminder), head(gapminder)

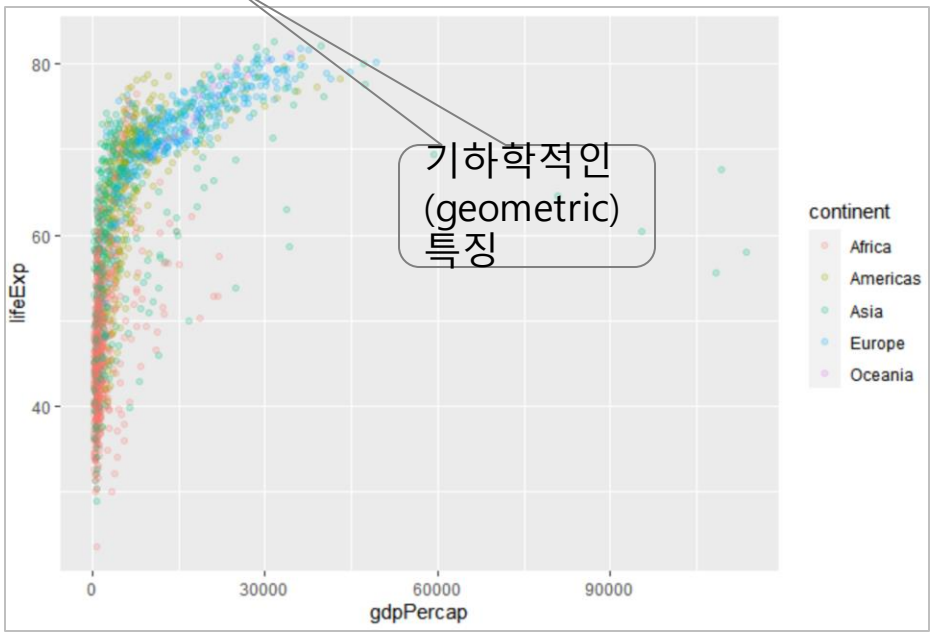
```
> str(gapminder)
tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
 $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1
 1 1 1 1 1 1 1 1 ...
 $ continent: Factor w/ 5 levels "Africa","Americas",..:
 3 3 3 3 3 3 3 3 3 3 ...
 $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977
 1982 1987 1992 1997 ...
 $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
 $ pop      : int [1:1704] 8425333 9240934 10267083 1153
7966 13079460 14880372 12881816 13867957 16317921 222274
15 ...
 $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

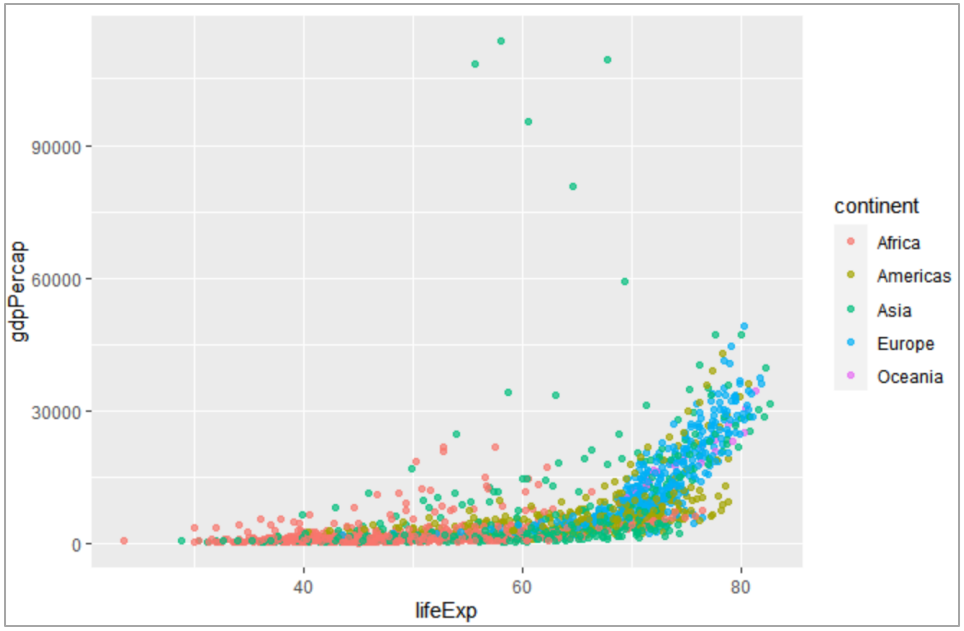
```
> head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanist~ Asia       1952    28.8  8.43e6       779.
2 Afghanist~ Asia       1957    30.3  9.24e6       821.
3 Afghanist~ Asia       1962    32.0  1.03e7       853.
4 Afghanist~ Asia       1967    34.0  1.15e7       836.
```

6.3 시각화 도구

시각화에 특화된 ggplot2 라이브러리

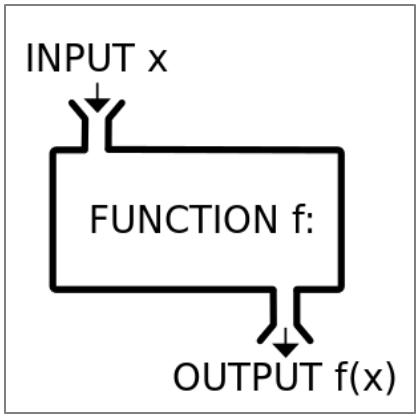
- ggplot2 라이브러리의 함수는 세 가지 요소로 구성된, 다음의 기본 표현식 사용
- 세 가지 기본 요소(①사용 data, ② x,y 축, ③ 그래프 type)
- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, col = continent)) + geom_point(alpha = 0.2)
- ggplot(gapminder, aes(x = lifeExp, y = gdpPercap, col = continent)) + geom_point(alpha = 0.7)



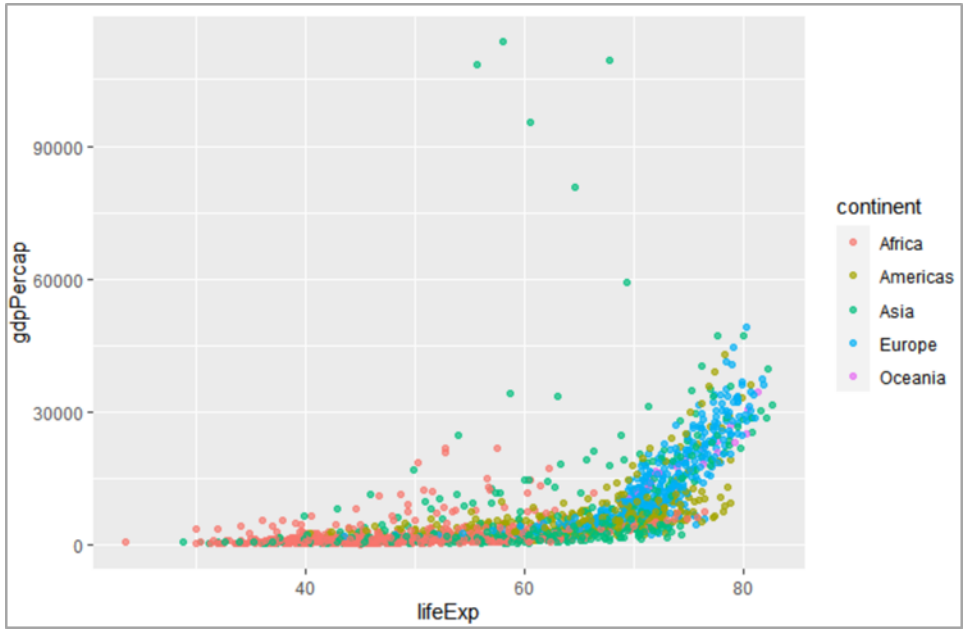


6.3 시각화 도구

수학에서 **함수(function)** 는 어떤 집합의 각 원소를 다른 집합의 유일한 원소에 대응시키는 이항 관계다



- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, col = continent)) + geom_point(alpha = 0.2)



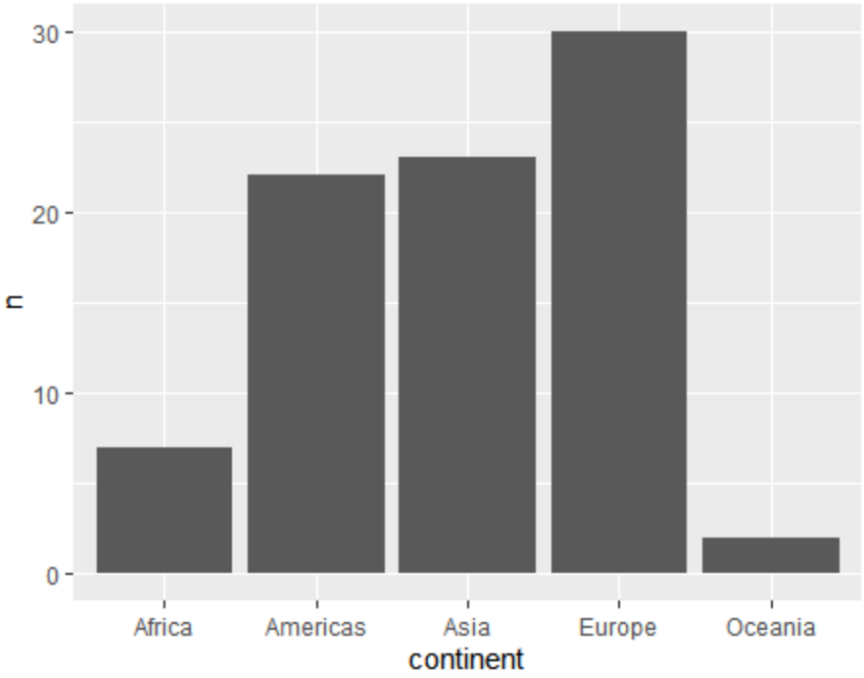
6.3 시각화 도구

ggplot2 함수

- 시각화 객체를 생성하는 역할을 한다.
  - 초기화 과정에서 입력 데이터와 가로축과 세로축에 대응될 항목을 지정해야 한다.
  - 내부에 aes(aesthetic)를 이용해 지정한다.
  - gapminder %>% filter(lifeExp>70) %>% group_by(continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x=continent, y=n)) + geom_bar(stat="identity")

데이터 프레임의 값을 그대로 사용해서 그래프를 그리라는 뜻

세로축에는 빈도수가 표시되므로 세로축에 대응될 변수를 별도로 지정하지 않아야함.



Ggplot2 라이브러리의 geom_bar 함수를 이용한 막대그래프

6.3 시각화 도구

geom_point 함수(1)

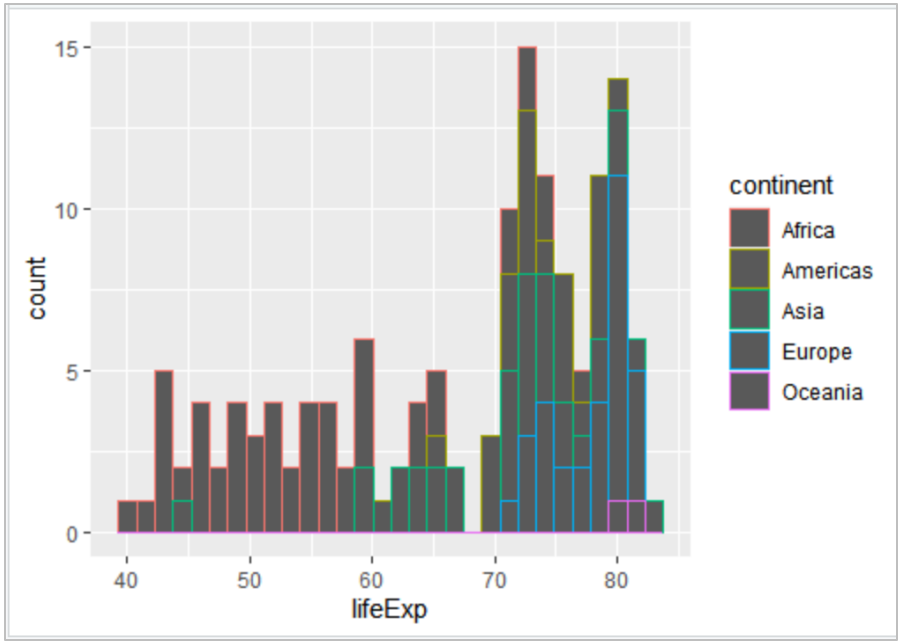
- 데이터를 점으로 표시하는 플롯을 그린다.
  - ✓ 내부에서 alpha 옵션을 통해 점의 불투명도(투명 0.0 ~ 불투명 1.0)를 설정할 수 있어, 마커들이 겹쳐 표시되더라도 데이터의 분포와 빈도를 확인 가능
- 위치에 사용할 수 있는 함수들
  - ✓ geom_line : 데이터를 선으로 표시한다.
  - ✓ geom_bar : 데이터를 막대그래프로 표시한다. 별도의 설정이 없으면 분포를 자동으로 계산해 geom_histogram과 동일하게 히스토그램을 그리기 때문에 히스토그램이 아닌 막대그래프, 즉 aes 함수에 x, y가 모두 지정된 그래프를 그리려면 stat="identity" 옵션을 지정한다.
  - ✓ geom_histogram : 히스토그램 전용의 플롯 함수다. 데이터가 그룹으로 구분되어 있는 경우 기본 옵션은 position ="stack"으로 막대를 위로 쌓아 올리도록 되어 있다. 막대를 나란히 옆으로 표시하려면 position ="dodge" 옵션을 지정한다.

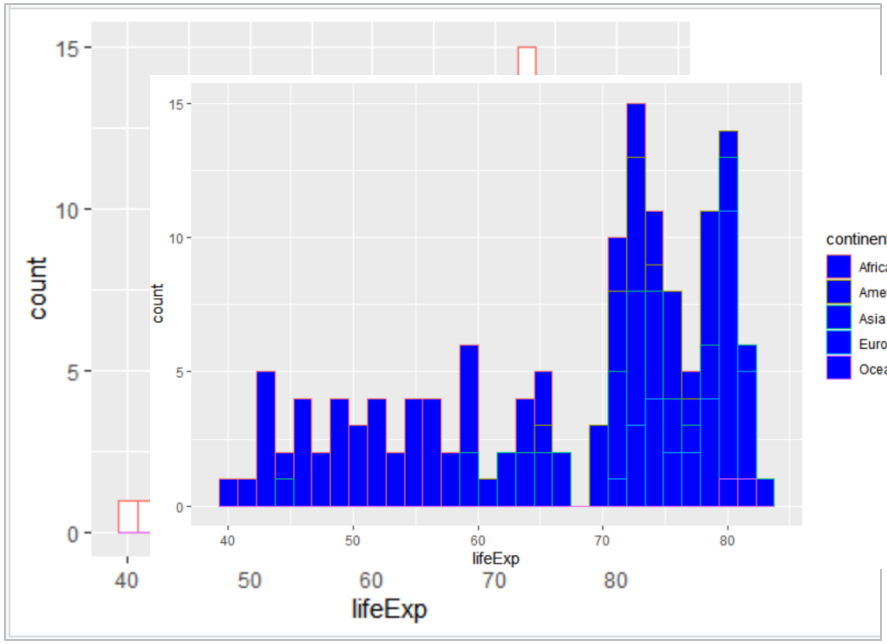
- 산점도 / 산포도 geom_point( )
- 선 그래프 geom_line( )
- 박스플롯 geom_boxplot( )
- 히스토그램 geom_histogram( )
- 막대 그래프 geom_bar( )

6.3 시각화 도구

geom_point 함수(2)

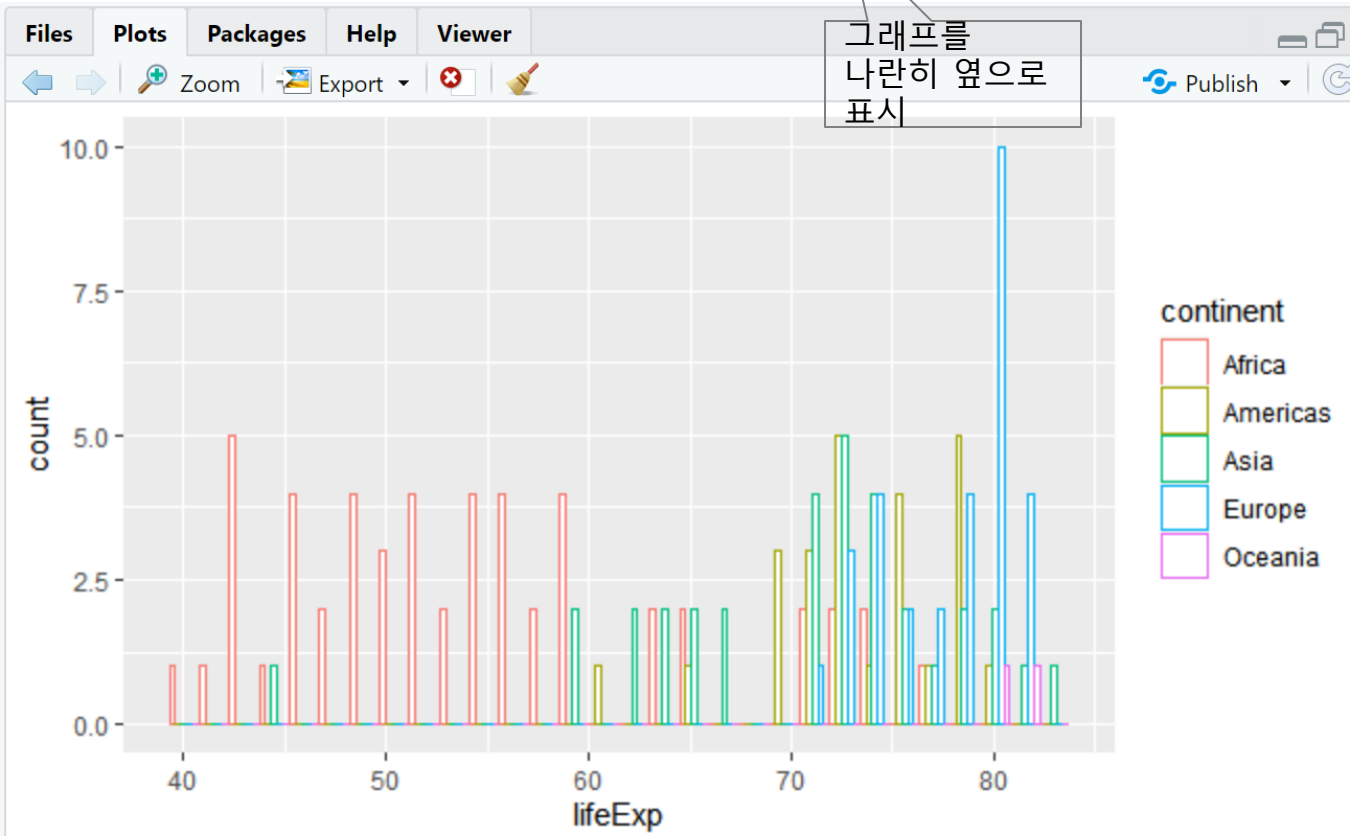
- gapminder %>% filter(year == 2007) %>% ggplot(aes(lifeExp, col = continent)) + geom_histogram()
- gapminder %>% filter(year == 2007) %>% ggplot(aes(lifeExp, col = continent)) + geom_histogram(fill = "white")





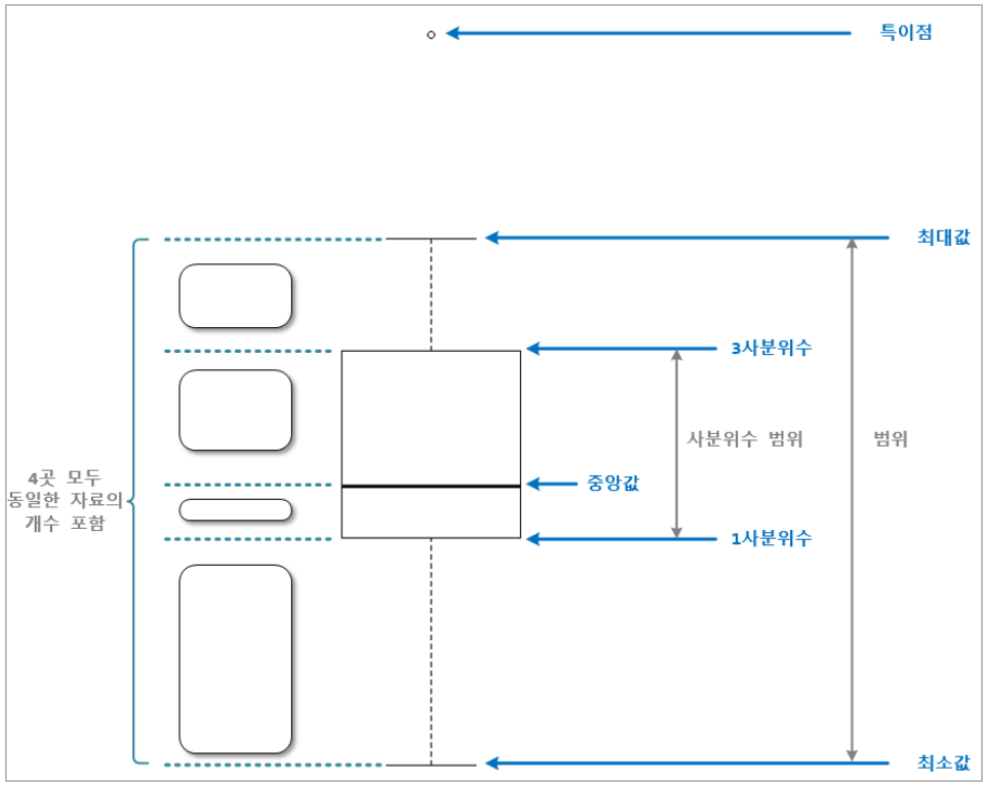
6.3 시각화 도구

- geom_point 함수(3)
  - gapminder %>% filter(year == 2007) %>% ggplot(aes(lifeExp, col = continent)) + geom_histogram(fill = "white",position="dodge")



6.3 시각화 도구

- geom_boxplot 함수
  - 여러 항목의 분포를 한꺼번에 관찰하는 함수이며, 이상값을 파악하는 데 유용
- 박스 플롯(box plot)의 해석



박스 플롯은 박스와 박스 바깥의 선(whisker)으로 이루어져 있습니다.

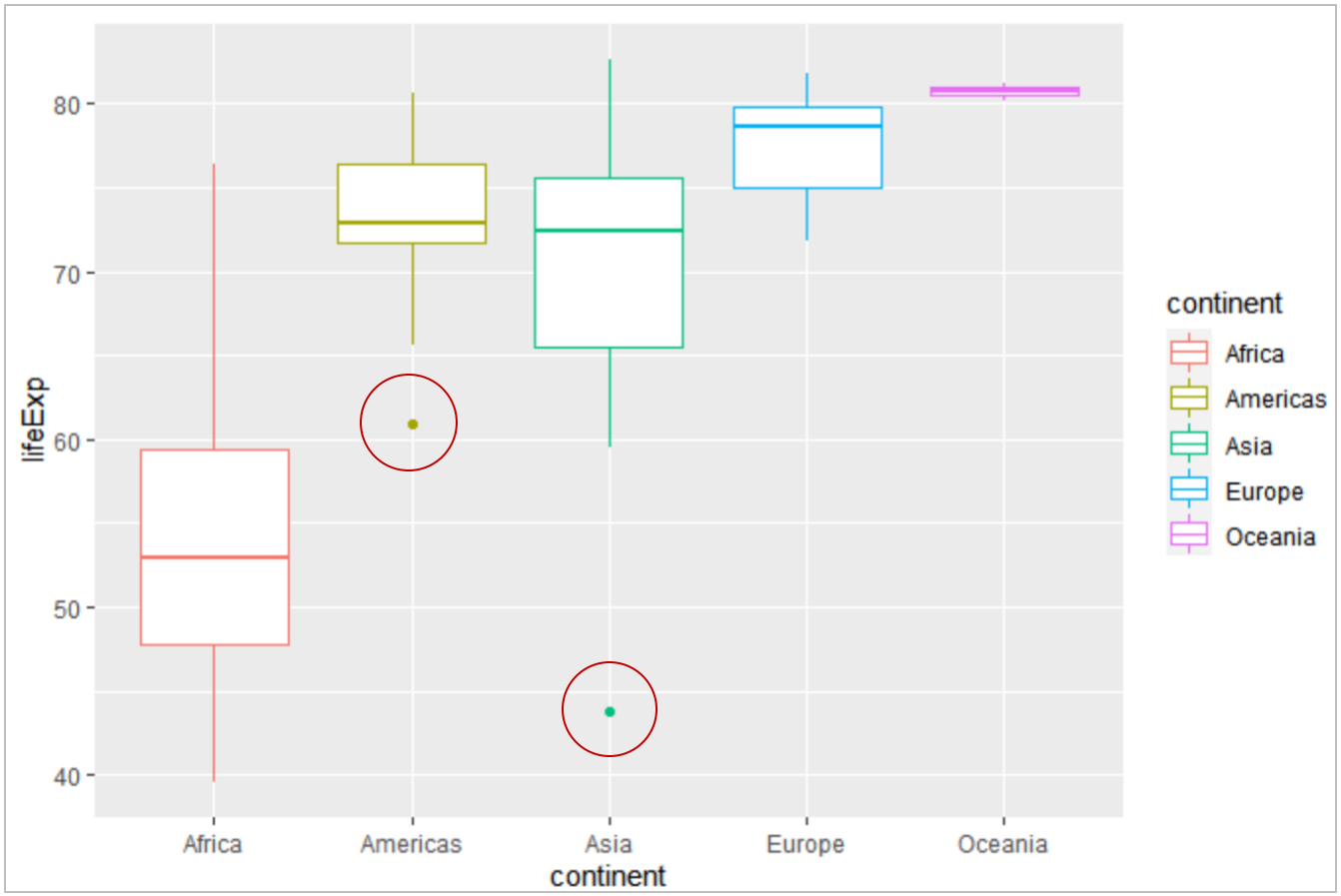
구분	설명
whisker	• 상자의 좌우 또는 상하로 뻗어나간 선
박스 내부의 가로선	• 중앙값을 나타냅니다.
lower whisker	• 최소값 • '중앙값 - 1.5 x IQR'보다 큰 데이터 중 가장 작은 값
upper whisker	• 최대값 • '중앙값 + 1.5 x IQR'보다 작은 데이터 중 가장 큰 값
IQR	• Inter Quartile Range • 제3사분위수 - 제1사분위수 • 실수 값 분포에서 1사분위수(Q1)와 3사분위수(Q3)를 뜻하고 이 3사분위수와 1사분수의 차이(Q3 - Q1)를 IQR(interquartile range)라고 합니다.
점	• 이상치(outlier; 아웃라이어) ≒ 특이점 • lower whisker보다 작은 데이터 또는 upper whisker보다 큰 데이터가 여기에 해당됩니다.

https://codedragon.tistory.com/7012

6.3 시각화 도구

geom_boxplot 함수

- gapminder %>% filter(year == 2007) %>% ggplot(aes(continent, lifeExp, col = continent)) + geom_boxplot()



6.3 시각화 도구

scale_x_log10 · scale_y_log10 함수

- scale_x_log10과 scale_y_log10 함수를 사용하면 데이터에 직접 로그를 취하지 않고도 축의 스케일을 바꾸어 동일한 효과를 얻을 수 있다.
- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, col = continent)) + geom_point(alpha = 0.7)
- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, col = continent)) + geom_point(alpha = 0.7) + scale_x_log10()

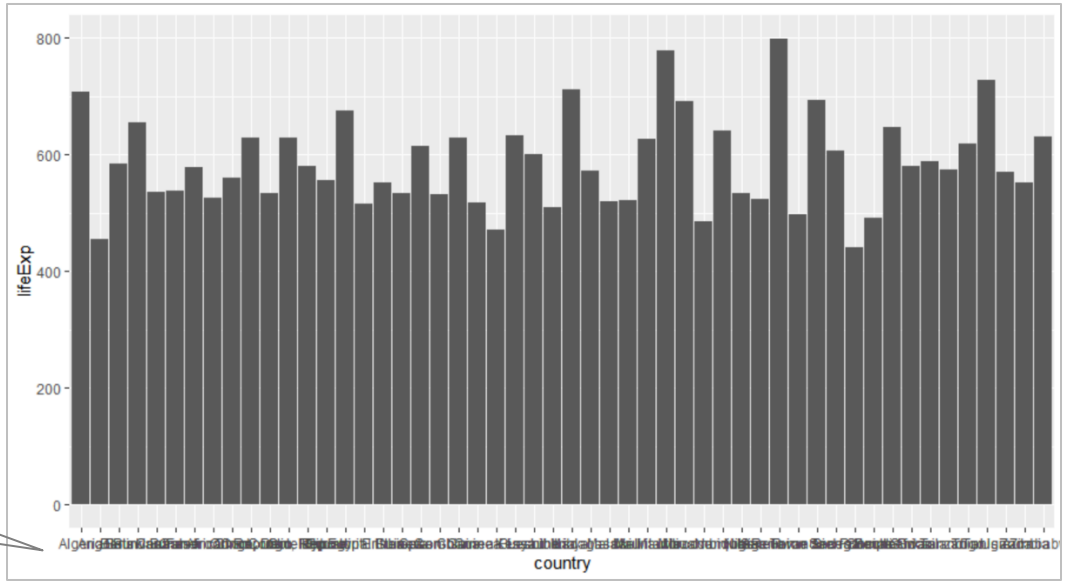
가로축에 log10 적용

6.3 시각화 도구

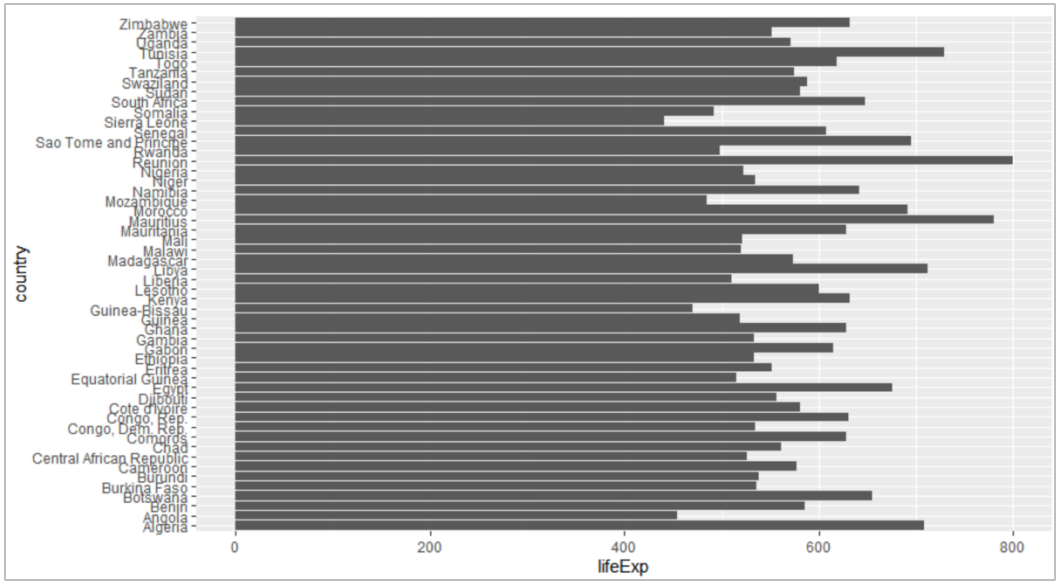
coord_flip 함수

- gapminder %>% filter(continent == "Africa") %>% ggplot(aes(country, lifeExp)) + geom_bar(stat = "identity")

나라 이름이 겹쳐서 보이지 않을 경우 사용

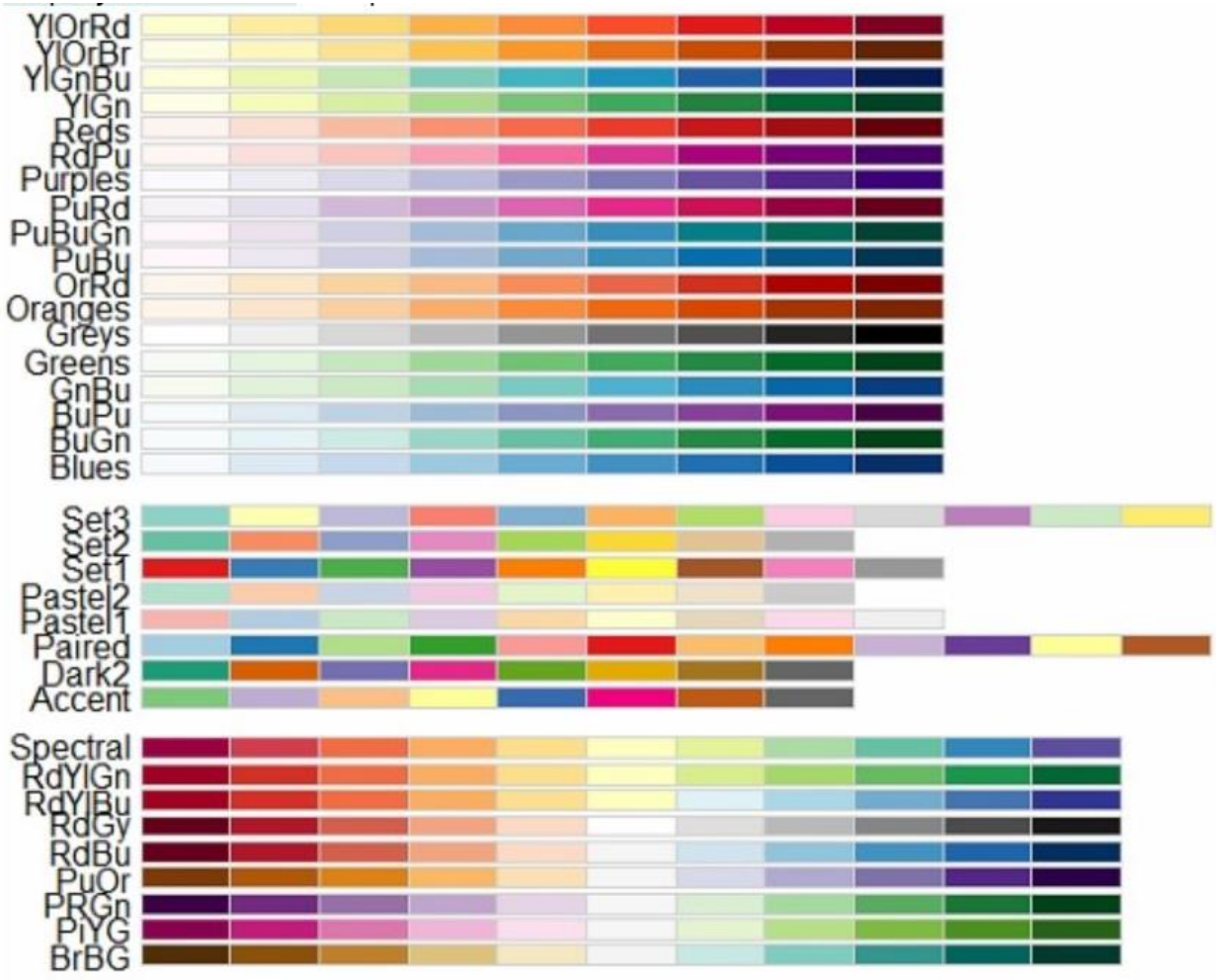


- gapminder %>% filter(continent == "Africa") %>% ggplot(aes(country, lifeExp)) + geom_bar(stat = "identity") + coord_flip()



6.3 시각화 도구

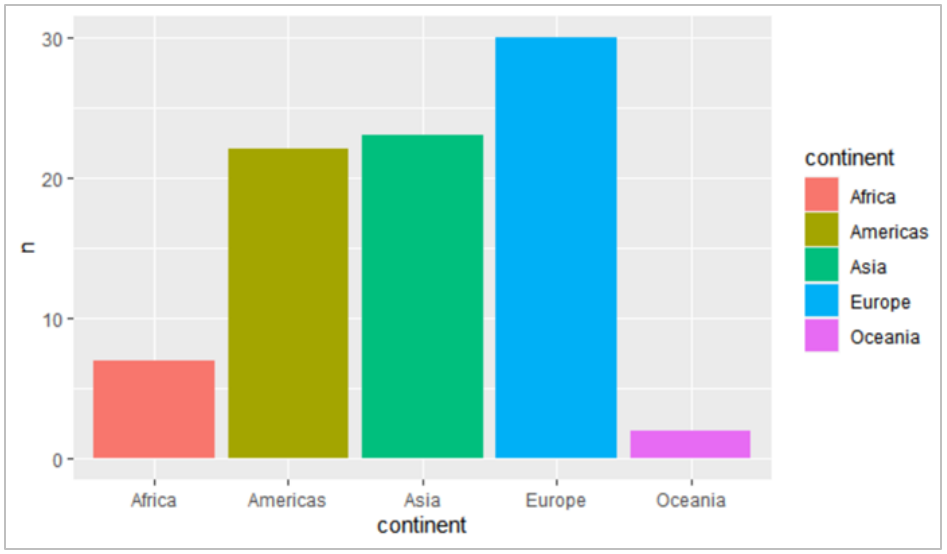
- scale_fill_brewer 함수(2)
  - install.packages("RColorBrewer")
  - library(RColorBrewer)
  - display.brewer.all()



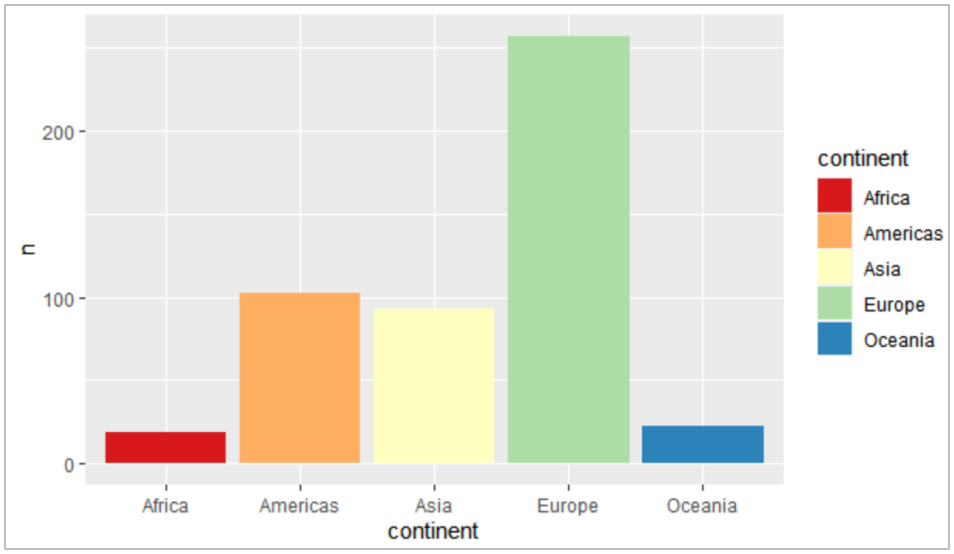
6.3 시각화 도구

scale_fill_brewer 함수(3)

- gapminder %>% filter(lifeExp>70) %>% group_by(continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x = continent, y = n)) + geom_bar(stat = "identity", aes(fill = continent)) # 기본 팔레트
- gapminder %>% filter(lifeExp>70) %>% group_by(year, continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x = continent, y = n)) + geom_bar(stat = "identity", aes(fill = continent)) + scale_fill_brewer(palette = "Spectral") # Spectral 적용



기본 팔레트 적용

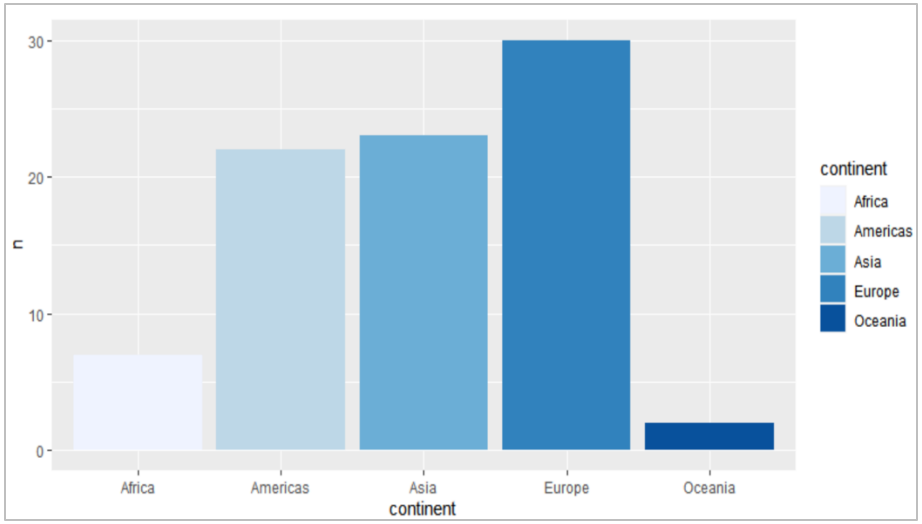


Spectral 팔레트 적용

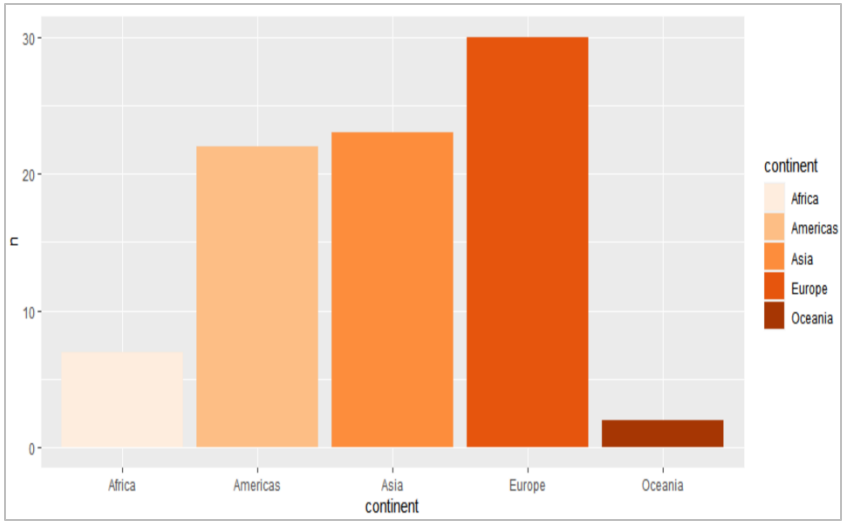
6.3 시각화 도구

scale_fill_brewer 함수(3)

- gapminder %>% filter(lifeExp>70) %>% group_by(continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x = continent, y = n)) + geom_bar(stat = "identity", aes(fill = continent)) + scale_fill_brewer(palette = "Blues")
- gapminder %>% filter(lifeExp>70) %>% group_by(continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x = continent, y = n)) + geom_bar(stat = "identity", aes(fill = continent)) + scale_fill_brewer(palette = "Oranges")



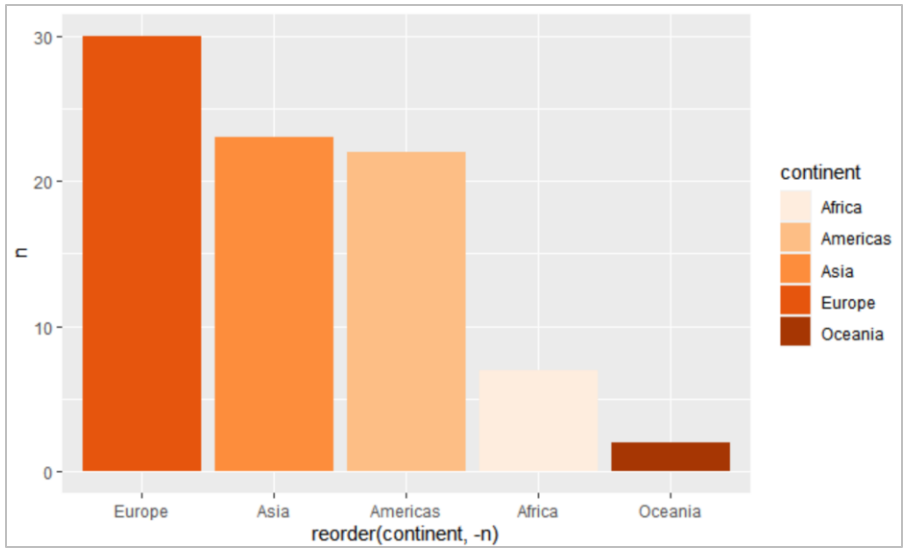
Blues 팔레트 적용

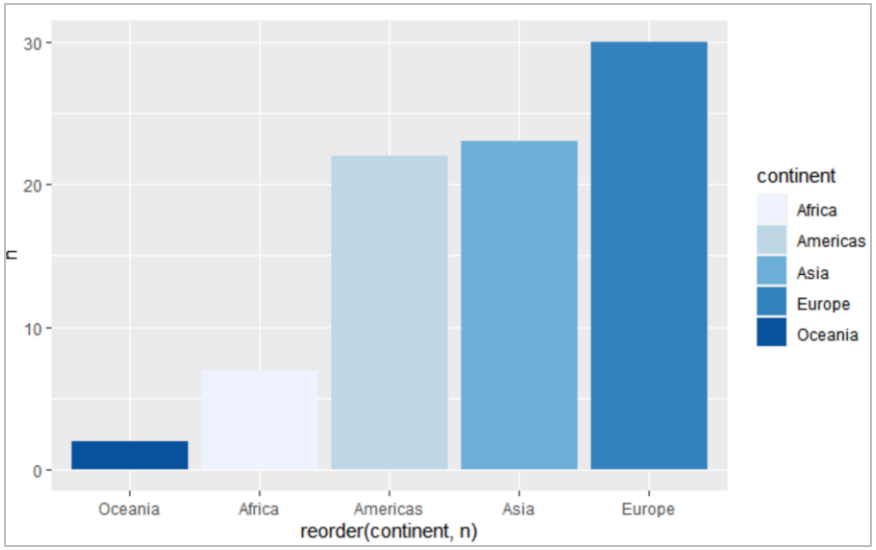


Oranges 팔레트 적용

6.3 시각화 도구

- scale_fill_brewer 함수(4) :
  - 그래프에 표시되는 데이터의 순서를 조정하기 위해서는 reorder 함수를 활용한다.
  - gapminder %>% filter(lifeExp >70) %>% group_by(continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x = reorder(continent, -n), y = n)) + geom_bar(stat = "identity", aes(fill = continent)) + scale_fill_brewer(palette = "Oranges")
  - gapminder %>% filter(lifeExp >70) %>% group_by(continent) %>% summarize(n = n_distinct(country)) %>% ggplot(aes(x = reorder(continent, n), y = n)) + geom_bar(stat = "identity", aes(fill = continent)) + scale_fill_brewer(palette = "Blues")





6.4 시각화를 이용한 데이터 탐색

- 시각화를 통해 종종 다른 방법으로 얻을 수 없는 **통찰을 얻을 수 있다**.
- 특히 R과 같은 효과적인 데이터 분석 도구들 덕분에 시각적 탐구는 과거에 비해 **훨씬 쉽고 흥미로운 작업이 됨**
- 시각적 탐구를 할 때는 찾고자 하는 아이디어가 무엇인지 아직 확실하지 않기 때문에 가능하면 데이터를 폭넓게 시각화하는 경향이 있으며, 때로는 여러 개의 데이터를 결합하여 활용하는 경우도 있다.
- 효율적인 도구들에 힘입어 시각화는 때론 보기 좋은 차트를 만들어내는 일에 힘을 쏟게 됨
- 그러나 시각화의 본질은 오히려 데이터의 의미를 통찰하고 분석해내는 시각적 탐구에 있다는 본래에 취지에 전념해야 함

6.4 시각화를 이용한 데이터 탐색

gapminder 데이터의 시각적 탐구(1)

- 세계 여러 나라의 경제 및 복지 수준, 국가나 지역에 따라 나타나는 공통점과 차이점, 아시아 일부 국가들이 지난 수십 년 동안 이룬 눈에 띄는 경제 성장 등 확인할 수 있음
- 또한 country, continent, year , lifeExp, pop, gdpPercap등 속성들 간의 상호 관계를 분석함으로써 이러한 차이와 변화가 나타난 원인에 대해서 생각해볼 수 있음

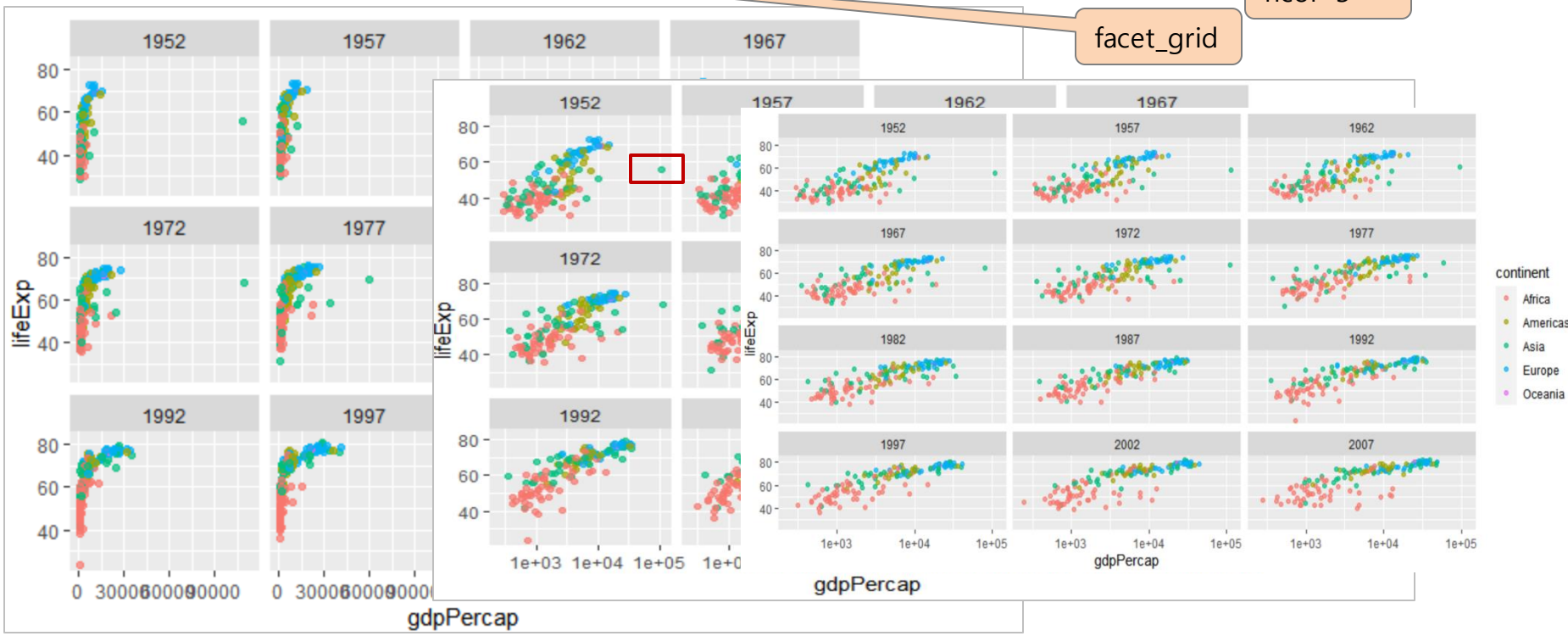
```
> str(gapminder)
tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
 $ country  : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1
 1 ...
 $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3
 3 3 3 ...
 $ year     : int [1:1704] 1952 1957 1962 1967 1972 1977 1982 1987 199
2 1997 ...
 $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
 $ pop      : int [1:1704] 8425333 9240934 10267083 11537966 13079460
 14880372 12881816 13867957 16317921 22227415 ...
 $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
> head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp      pop gdpPercap
  <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
1 Afghanistan Asia       1952    28.8  8425333      779.
2 Afghanistan Asia       1957    30.3  9240934      821.
3 Afghanistan Asia       1962    32.0 10267083      853.
4 Afghanistan Asia       1967    34.0 11537966      836.
5 Afghanistan Asia       1972    36.1 13079460      740.
6 Afghanistan Asia       1977    38.4 14880372      786.
```

6.4 시각화를 이용한 데이터 탐색

- gapminder 데이터의 시각적 탐구(2)
  - R을 이용해 대륙별의 경제 및 기대 수명과 이의 변화를 시각화해보기로 하자.
  - gapminder %>% ggplot(aes(gdpPercap, lifeExp, col = continent)) + geom_point(alpha = 0.7) + facet_wrap(~year) + scale_x_log10()
  - gapminder %>% ggplot(aes(gdpPercap, lifeExp, col = continent)) + geom_point(alpha = 0.7) + facet_wrap(Origin~year) + scale_x_log10()

ncol=3

facet_grid



6.4 시각화를 이용한 데이터 탐색

gapminder 데이터의 시각적 탐구(3)

- 그래프를 통해 알 수 있는 사항들을 정리 !
  - 유럽에서는 상당수의 국가들이 일찍부터 높은 gdpPercap과 lifeExp을 기록하였다.
  - 70년대 이후 아시아와 아메리카 대륙에 있는 많은 국가들의 gdpPercap과 lifeExp이 빠르게 증가하고 있다.
  - 아프리카 국가들의 상당수는 낮은 gdpPercap과 lifeExp에 머물고 있다.
  - gdpPercap과 lifeExp이 증가할수록 두 변수 사이의 관계가 점차로 선형화되는 경향이 뚜렷하게 나타난다.
  - 관측 기간 동안 거의 모든 국가의 gdpPercap과 lifeExp가 전체적으로 상승하였다(아프리카 일부 국가 제외).
  - gdpPercap의 최솟값과 최댓값의 차이가 증가하였다. 즉, 격차가 커졌다.

6.4 시각화를 이용한 데이터 탐색

쿠웨이트의 경제지표 변화와 특성(1)

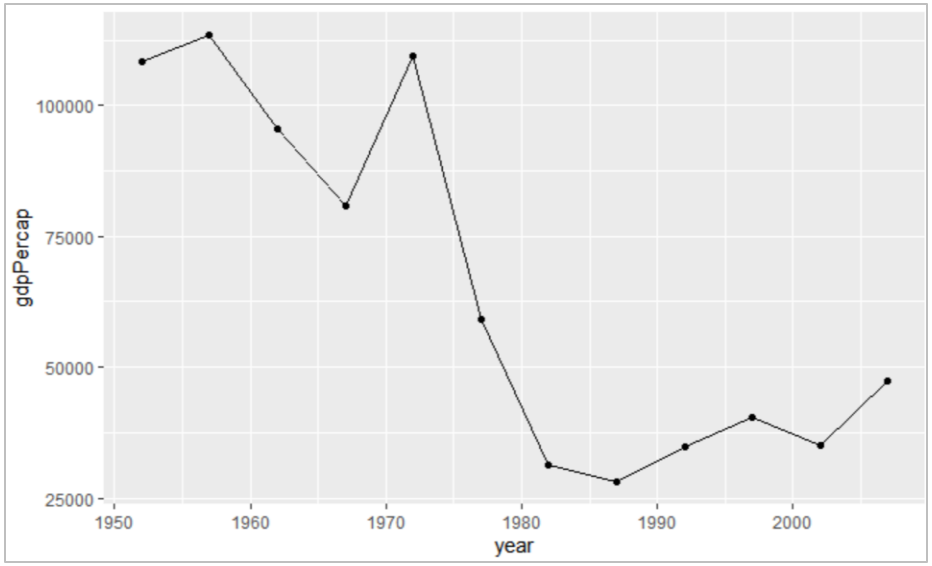
- 1952년에 1인당 gdpPercap이 매우 높은 **아시아 국가 하나**를 발견할 수 있다

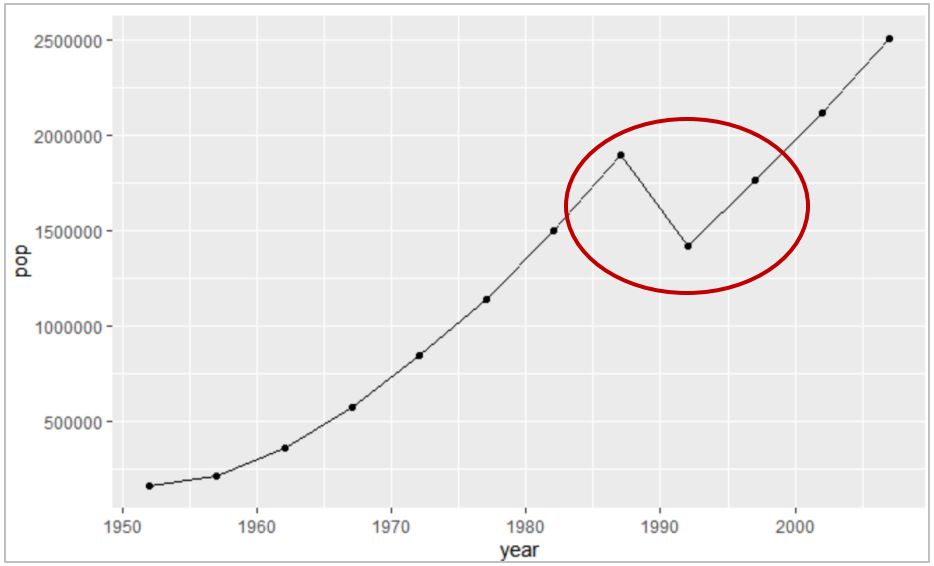
```
> gapminder %>% filter(year == 1952 & gdpPercap > 10000 & continent
 == "Asia")
# A tibble: 1 x 6
  country continent  year lifeExp    pop gdpPercap
  <fct>   <fct>     <int>   <dbl>  <int>     <dbl>
1 Kuwait  Asia       1952    55.6 160000   108382.
> gapminder %>% filter(country == "Kuwait")
# A tibble: 12 x 6
   country continent  year lifeExp     pop gdpPercap
   <fct>   <fct>     <int>   <dbl>   <int>     <dbl>
 1 Kuwait  Asia       1952    55.6  160000   108382.
 2 Kuwait  Asia       1957    58.0  212846   113523.
 3 Kuwait  Asia       1962    60.5  358266    95458.
 4 Kuwait  Asia       1967    64.6  575003    80895.
 5 Kuwait  Asia       1972    67.7  841934   109348.
 6 Kuwait  Asia       1977    69.3 1140357    59265.
 7 Kuwait  Asia       1982    71.3 1497494    31354.
 8 Kuwait  Asia       1987    74.2 1891487    28118.
 9 Kuwait  Asia       1992    75.2 1418095    34933.
10 Kuwait  Asia       1997    76.2 1765345    40301.
11 Kuwait  Asia       2002    76.9 2111561    35110.
12 Kuwait  Asia       2007    77.6 2505559    47307.
```

6.4 시각화를 이용한 데이터 탐색

쿠웨이트의 경제지표 변화와 특성(1)

- 1952년에 1인당 gdpPercap이 매우 높은 국가 쿠웨이트(Kuwait) 분석
- gapminder %>% filter(country == "Kuwait") %>% ggplot(aes(year, gdpPercap)) + geom_point() + geom_line()
- gapminder %>% filter(country == "Kuwait") %>% ggplot(aes(year, pop)) + geom_point() + geom_line()





6.4 시각화를 이용한 데이터 탐색

대한민국 경제지표 변화와 특성(2)

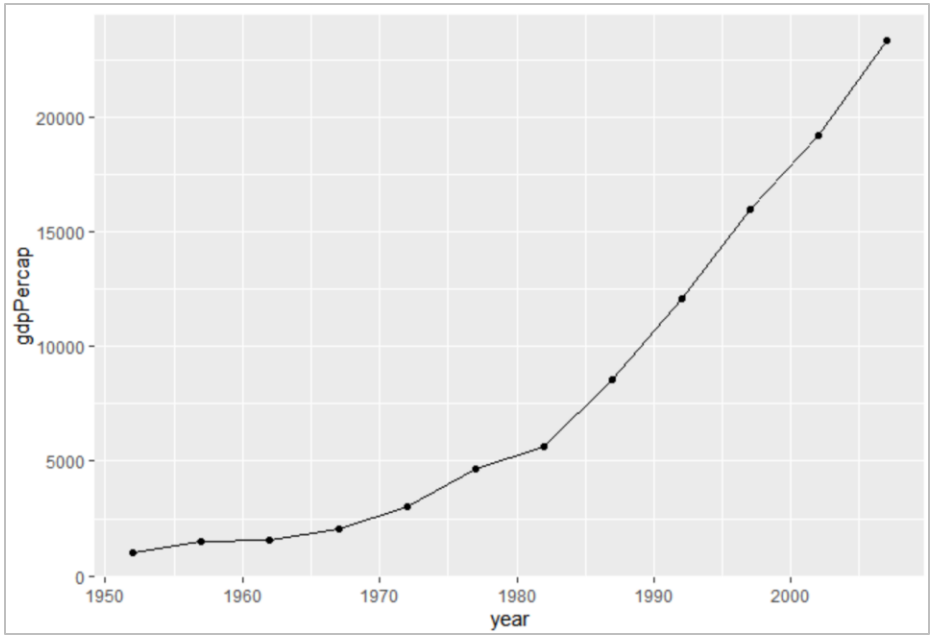
- gapminder$country[grep("Korea",gapminder$country)]
- gapminder %>% filter(country == "Korea, Rep.")

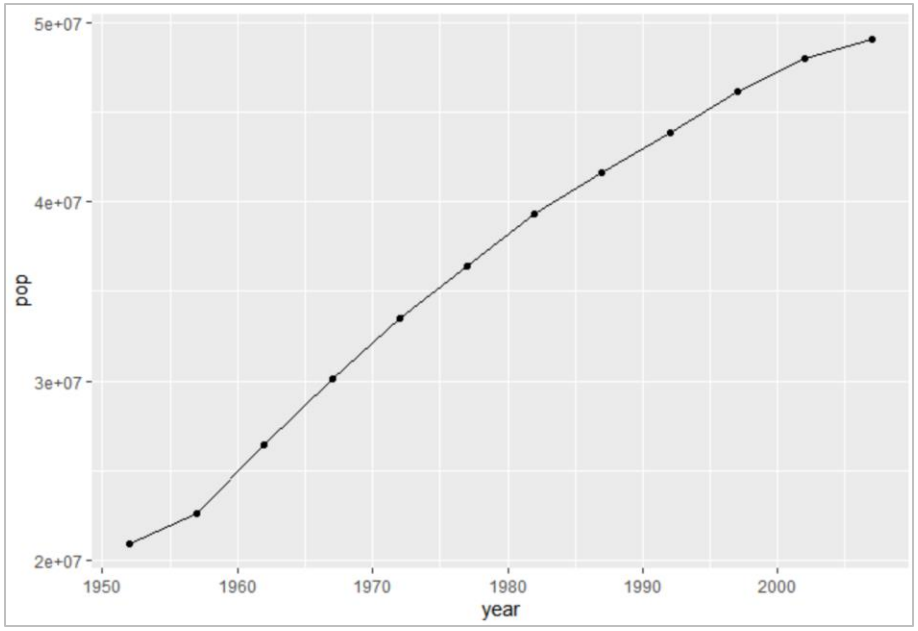
```
> gapminder$country[grep("Korea",gapminder$country)]
 [1] Korea, Dem. Rep. Korea, Dem. Rep. Korea, Dem. Rep.
 [4] Korea, Dem. Rep. Korea, Dem. Rep. Korea, Dem. Rep.
 [7] Korea, Dem. Rep. Korea, Dem. Rep. Korea, Dem. Rep.
[10] Korea, Dem. Rep. Korea, Dem. Rep. Korea, Dem. Rep.
[13] Korea, Rep.      Korea, Rep.      Korea, Rep.
[16] Korea, Rep.      Korea, Rep.      Korea, Rep.
[19] Korea, Rep.      Korea, Rep.      Korea, Rep.
[22] Korea, Rep.      Korea, Rep.      Korea, Rep.
142 Levels: Afghanistan Albania Algeria Angola ... Zimbabwe
> gapminder %>% filter(country == "Korea, Rep.")
# A tibble: 12 x 6
   country     continent  year lifeExp      pop gdpPercap
   <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
 1 Korea, Rep. Asia       1952    47.5 20947571     1031.
 2 Korea, Rep. Asia       1957    52.7 22611552     1488.
 3 Korea, Rep. Asia       1962    55.3 26420307     1536.
 4 Korea, Rep. Asia       1967    57.7 30131000     2029.
 5 Korea, Rep. Asia       1972    62.6 33505000     3031.
 6 Korea, Rep. Asia       1977    64.8 36436000     4657.
 7 Korea, Rep. Asia       1982    67.1 39326000     5623.
 8 Korea, Rep. Asia       1987    69.8 41622000     8533.
 9 Korea, Rep. Asia       1992    72.2 43805450    12104.
10 Korea, Rep. Asia       1997    74.6 46173816    15994.
11 Korea, Rep. Asia       2002    77.0 47969150    19234.
12 Korea, Rep. Asia       2007    78.6 49044790    23348.
```

6.4 시각화를 이용한 데이터 탐색

대한민국 경제지표 변화와 특성(2)

- gapminder %>% filter(country == "Korea, Rep.") %>% ggplot(aes(year, gdpPercap)) + geom_point() + geom_line()
- gapminder %>% filter(country == "Korea, Rep.") %>% ggplot(aes(year, pop)) + geom_point() + geom_line()

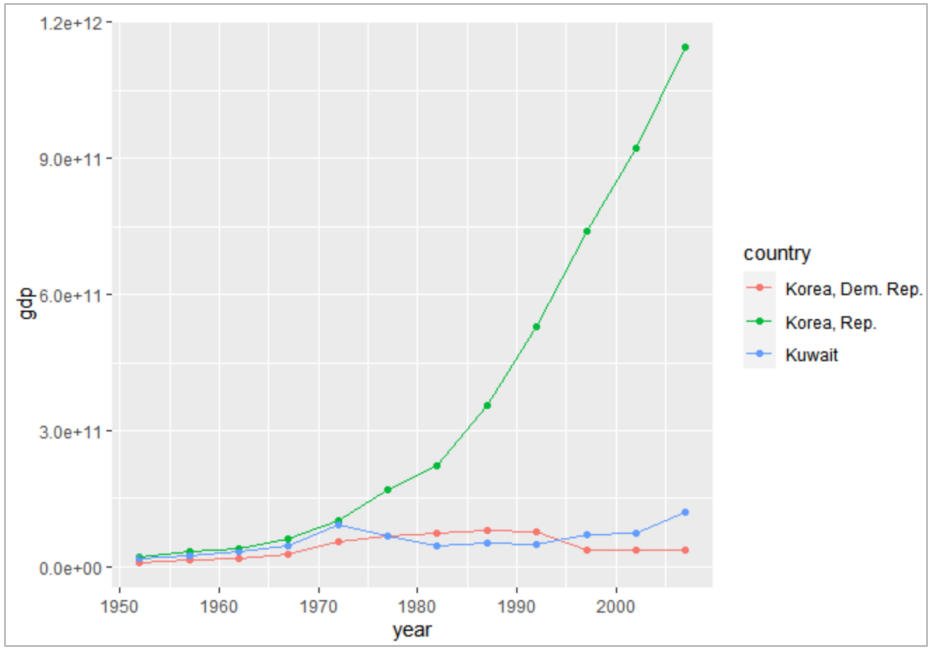


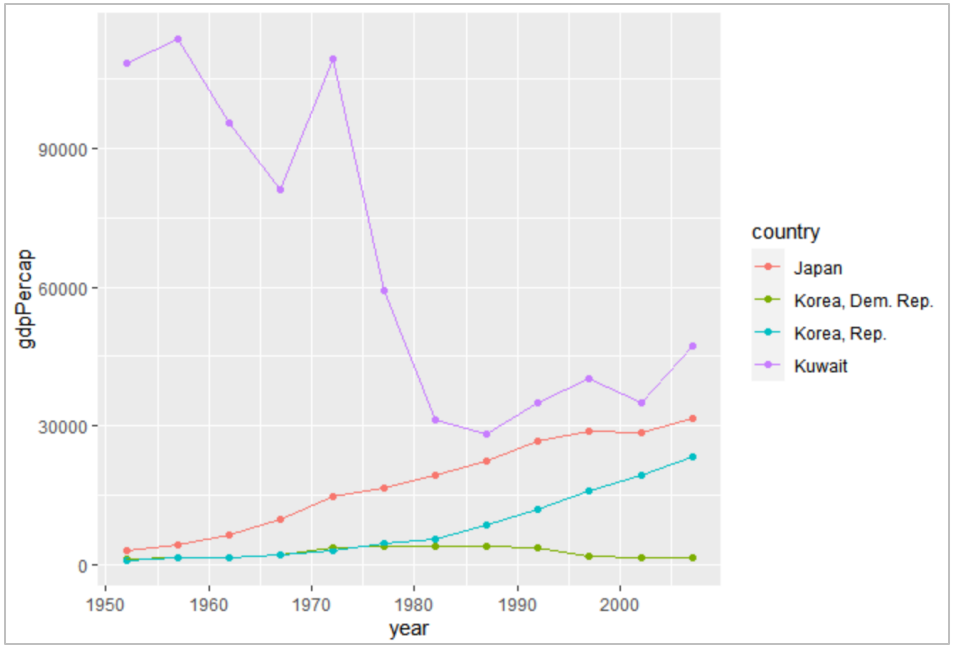


6.4 시각화를 이용한 데이터 탐색

특정 국가의 경제지표 변화 비교

- 동시에 변화하는 gdpPercap과 pop을 효과적으로 관찰하기 위해 gdp(국내총생산) 활용
- dplyr 라이브러리의 mutate 함수를 이용해 gdpPercap과 pop으로부터 국내총생산을 산출
- gapminder %>% filter(country == "Kuwait" | country == "Korea, Rep." | country == "Korea, Dem. Rep.") %>% mutate(gdp = gdpPercap*pop) %>% ggplot(aes(year, gdp, col = country)) + geom_point() + geom_line()





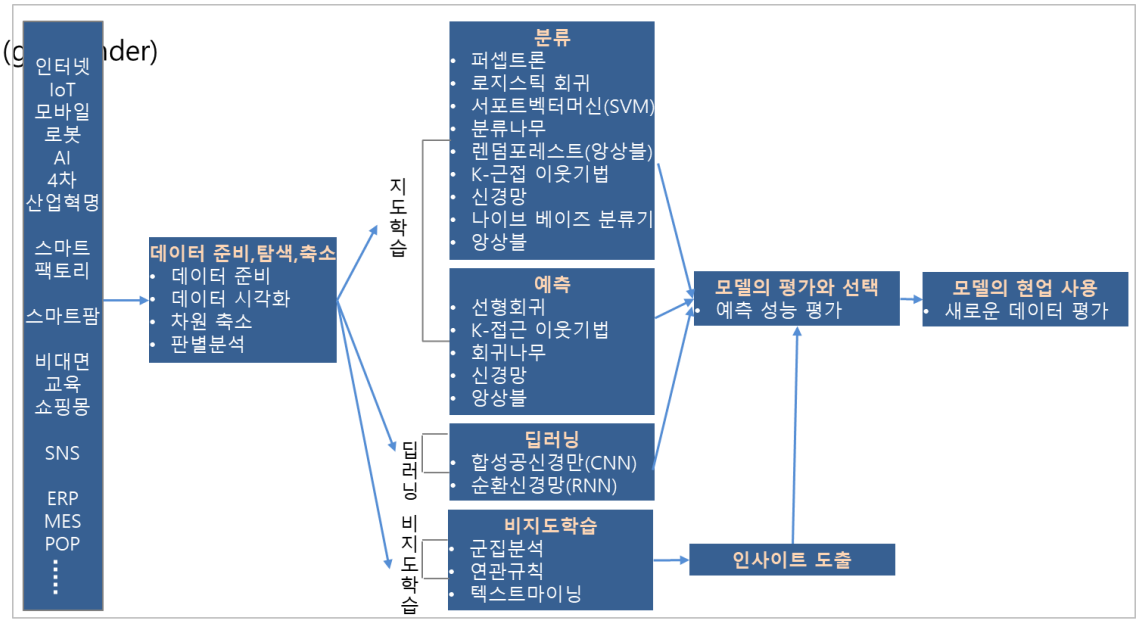
요약

- 시각화 사용 library
  - library(gapminder)
  - library(dplyr)
  - library(ggplot2)
  - library(RColorBrewer)

- **시각화에 특화된 ggplot2 라이브러리**
  - ggplot2는 R에서 가장 많이 사용되는 시각화 library
  - gg는 grammar of graphics를 뜻함
  - 사용할 data set : gapminder : str(gapminder), head(gapminder)

```
> 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
 1982 1987 1992 1997 ...
 $ lifeExp  : num [1:1704] 28.8 30.3 32 34 36.1 ...
 $ pop      : int [1:1704] 8425333 9240934 10267083 1153
7966 13079460 14880372 12881816 13867957 16317921 222274
15 ...
 $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
```

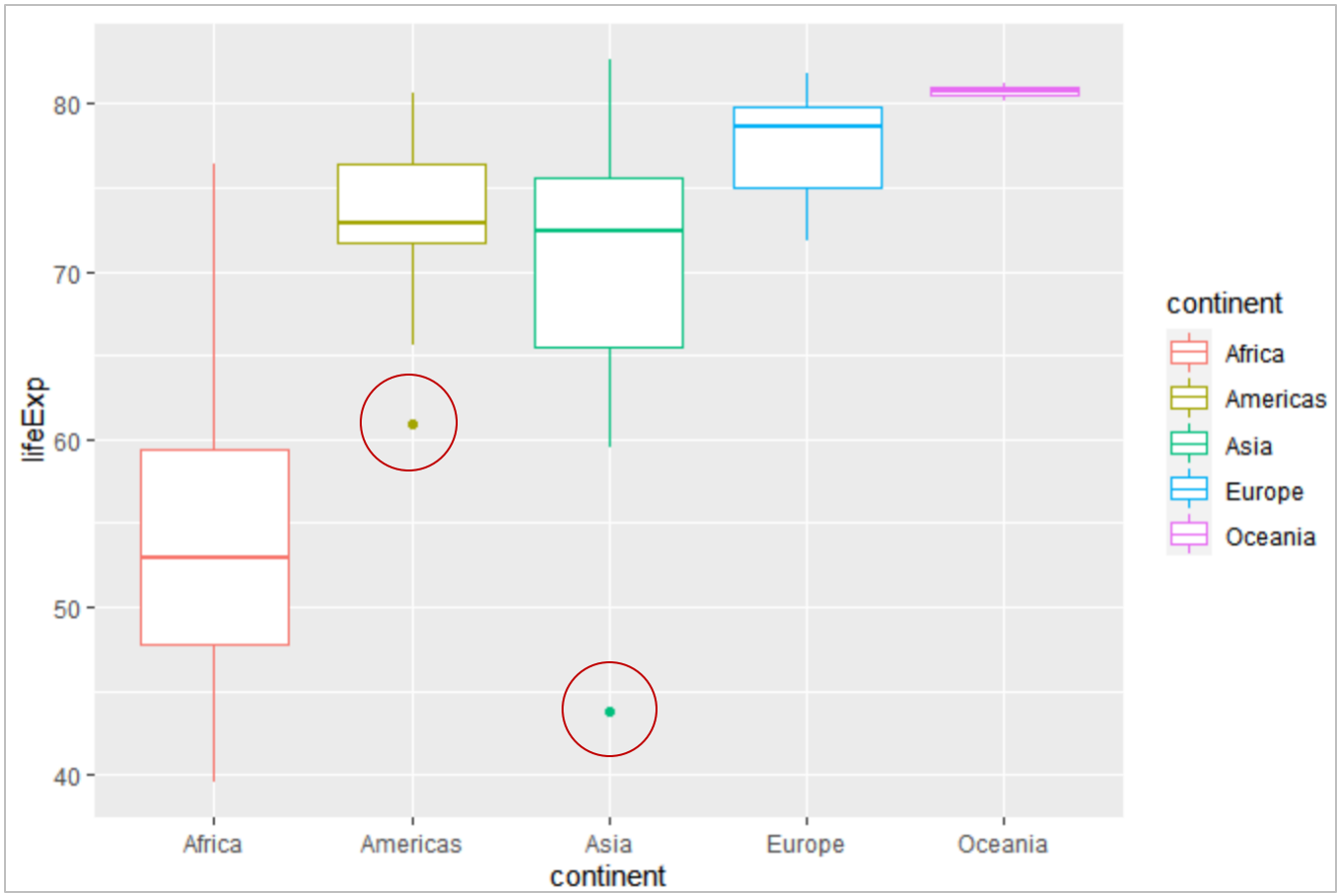
```
> head(gapminder)
# A tibble: 6 x 6
  country     continent  year lifeExp     pop gdpPercap
  <fct>       <fct>     <int>   <dbl>   <int>     <dbl>
1 Afghanist~ Asia       1952    28.8  8.43e6      779.
2 Afghanist~ Asia       1957    30.3  9.24e6      821.
3 Afghanist~ Asia       1962    32.0  1.03e7      853.
4 Afghanist~ Asia       1967    34.0  1.15e7      836.
```



요약

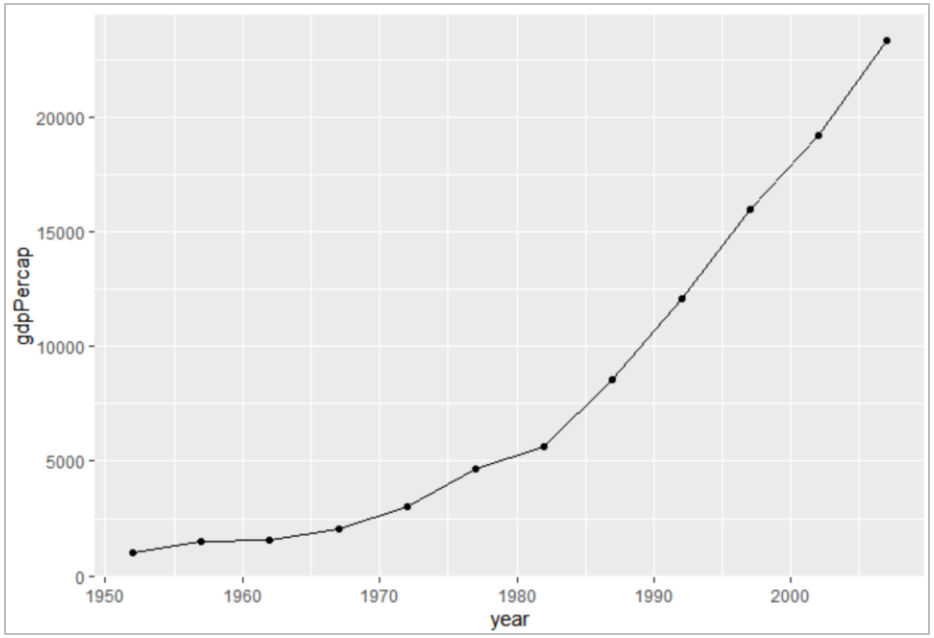
geom_boxplot 함수

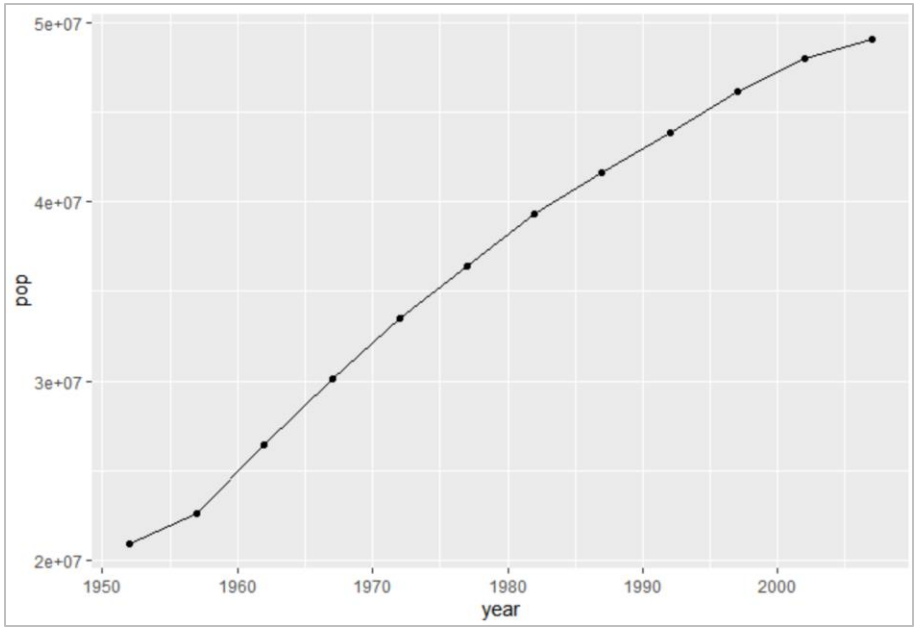
- gapminder %>% filter(year == 2007) %>% ggplot(aes(continent, lifeExp, col = continent)) + geom_boxplot()



요약

- 대한민국 경제지표 변화와 특성 그래프
  - gapminder %>% filter(country == "Korea, Rep.") %>% ggplot(aes(year, gdpPercap)) + geom_point() + geom_line()
  - gapminder %>% filter(country == "Korea, Rep.") %>% ggplot(aes(year, pop)) + geom_point() + geom_line()





Thank you


continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09
90
70
60
50
40
300
1000
3000
10000
30000
gdpPercap