R Notebook: PCA

Code

Hide

```
suppressPackageStartupMessages({
    library(dplyr)
    library(ggplot2)
    library(GGally)
    library(ggfortify)
})
```

Hide

```
# 국가 정보가 담긴 파일을 읽는다.
nation <- (
    read.delim('nation.txt', header=T, sep='\t', as.is=T,
               fileEncoding='UTF-8', col.names=c('name', 'population', 'area', 'gdp'))
    %>% tbl_df()
    %>% mutate(gdp_per_capita = round(gdp * 100000000 / population, digits=1)) # 1인당 GDP (gdp 필드의 단위는 "억 달러")
)
nation %>% sample_n(5) # 읽은 데이터 샘플 확인
```

name <chr>	population <int>	area <int>	gdp <dbl>	gdp_per_capita <dbl>
베네수엘라	29270000	910000	3337.15	11401.3
부탄	740000	38000	24.75	3344.6
루마니아	21660000	230000	1819.44	8400.0
이라크	37050000	430000	1484.11	4005.7
리히텐슈타인	37000	160	32.00	86486.5

5 rows

Hide

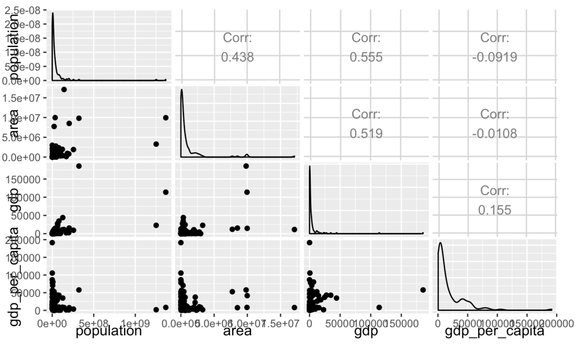
```
nation %>% filter(name == '대한민국')
```

name <chr>	population <int>	area <int>	gdp <dbl>	gdp_per_capita <dbl>
대한민국	49110000	9900	13211	26900.8

1 row

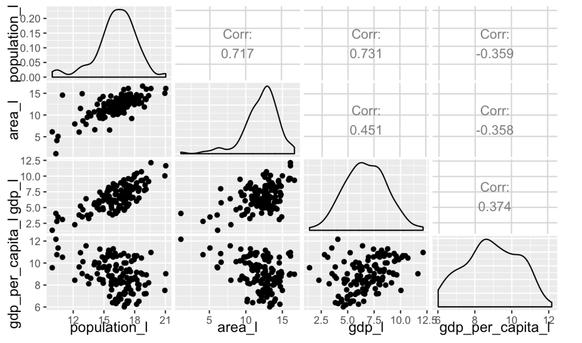
Hide

```
# 우선, 값의 분포와 요소 간의 상관관계를 보자.
ggpairs(nation %>% dplyr::select(population, area, gdp, gdp_per_capita))
```



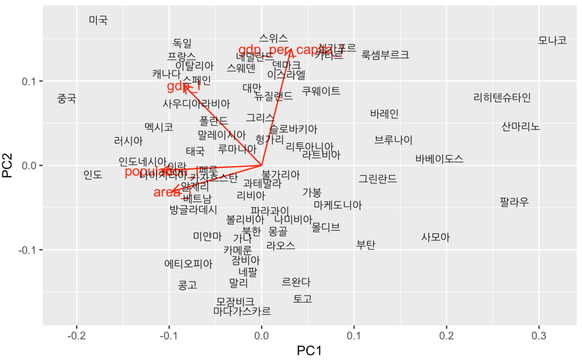
Hide

```
# 값이 아주 큰 몇몇 국가 때문에 분포 모양이 늘어지니까, 요소값에 로그를 취해준다.
nation <- (
    nation
    %>% mutate(population_l     = log(1+population),
               area_l           = log(1+area),
               gdp_l            = log(1+gdp),
               gdp_per_capita_l = log(1+gdp_per_capita)
               )
)
ggpairs(nation2 %>% dplyr::select(population_l, area_l, gdp_l, gdp_per_capita_l))
```



Hide

```
# 이 데이터로 주성분분석(Principal Component Analysis, PCA)을 해보자.
# base에 기본으로 제공하는 prcomp() 함수를 호출하면 결과가 나오고, ggfortify 패키지의 autoplot()에 인자로 넘기면 그림까지 예쁘게 그려준다.
m <- prcomp(~ population_l + area_l + gdp_l + gdp_per_capita_l, data=nation, scale=T)
autoplot(m, data=nation2, size=-1.0, label=F, loadings=T, loadings.label=T, loadings.label.size=4) +
    geom_text(aes(label=name, family='NanumGothic'), size=3, check_overlap=T)
```



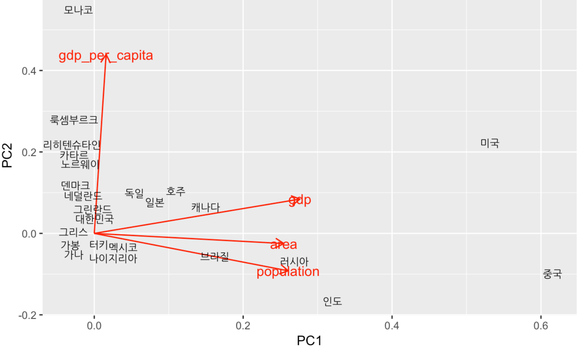
가장 의미있는 2개의 주성분을 X-Y축에 놓았을 때, 원래 데이터의 요소의 방향(벡간선)과 국가의 위치가 표현되었다. 이 경우 각 성분의 의미를 해석하는 것은 어렵지 않다. PC1은 인구나 영토같은 국가의 "크기"다. 상관관계가 높은 population_l과 area_l이 PC1과 비슷한 방향을 가리키고 있다. 실제로 왼쪽을 보면 미국, 중국, 인도, 러시아가 있고, 오른쪽에는 리히텐슈타인, 모나코같은 나라들이 있다. (크기순 랭킹을 하고 싶은데 어떤 요소를 기준으로 상아야 할지 모르겠다면 PC1값을 기준으로 해보면 어떨까?)

Y축은 gdp_per_capita(1인당 GDP)와 비슷한 방향인 걸로 보아 경제적인 부유함이라는 걸 알 수 있다. 그 결과 세계의 나라를 크기와 부유함을 기준으로 바라봤을 때 비슷한 나라들이 2차원 상에서 가깝게 찍혔다. 제일 위의 표나 요소의 분포/상관관계 그래프만으로는 보이지 않던 어떤 통찰을 주는 것 같기도 하다.

비교 삼아서, 로그를 취하지 않은 데이터로 주성분분석을 돌린 결과는 아래와 같다.

Hide

```
m <- prcomp(~ population + area + gdp + gdp_per_capita, data=nation, scale=T)
autoplot(m, data=nation, size=-1.0, label=F, loadings=T, loadings.label=T, loadings.label.size=4) +
    geom_text(aes(label=name, family='NanumGothic'), size=3, check_overlap=T)
```



R 외에 파이썬 코드도 실행할 수 있다.

```
for i in xrange(5):
    print i,
```

```
0 1 2 3 4
```

마스터 알고리즘 책에서 *"세상에서 가장 중요한 곡선"*이라고 했던 시그모이드 곡선의 수식은 Latex로 쓰면 된다.

$$\frac{1}{1+e^{-x}}$$