2024-2학기

다변량 자료분석 2

DATA : Global Country Information Dataset 2023

응용통계학과
2 0 1 8 3 3 2 9 4
한경찬

목차

01. 데이터 선택

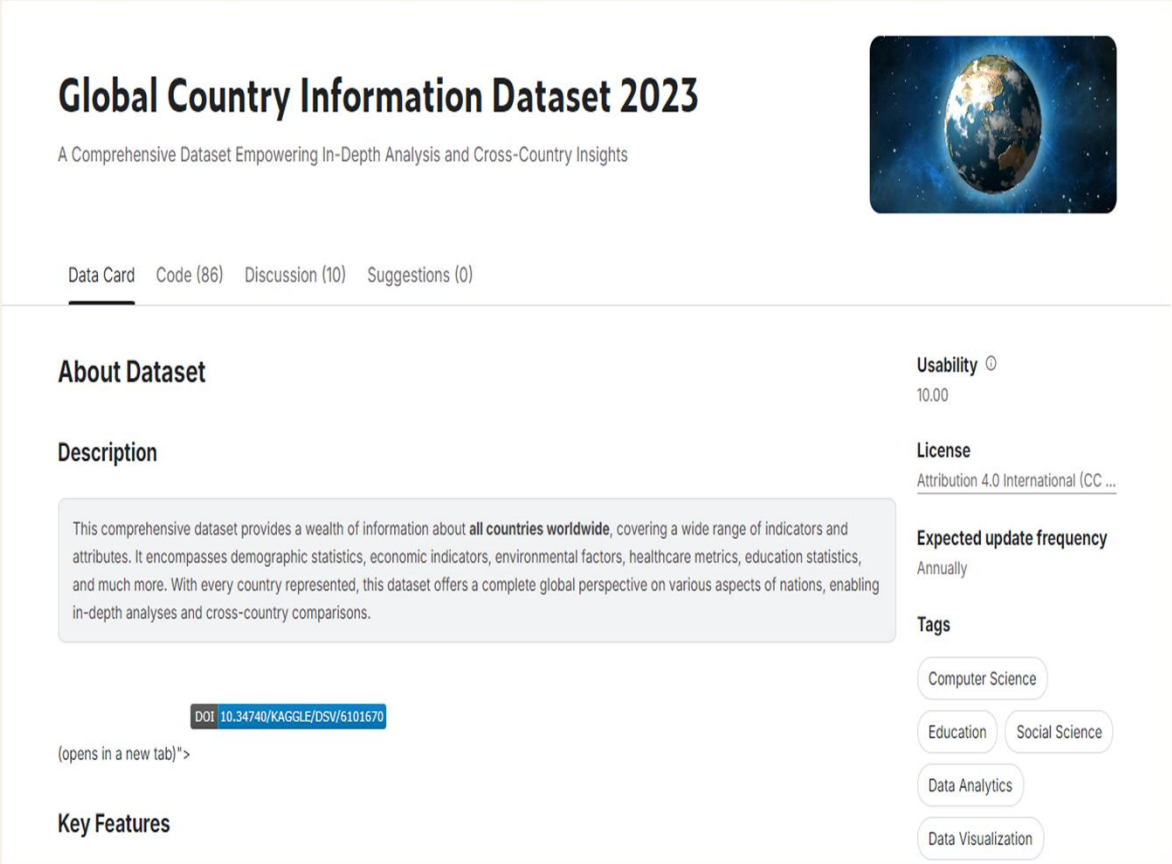
데이터 선택&선정 이유, 데이터 전처리

01. 데이터 선택

Kaggle
-> Global Country information Dataset 2023

데이터 선정 이유

각 국가별 인구, GDP, CPI 등 다양한 주요 지표들이 자세하게 나와 있어서 국가 간에 주요 변수들에 대하여 상관성을 분석해볼 수 있고, 더하여 각 변수들 간에 어떤 관계성이 있는지 파악할 수 있을 것을 기대하며 데이터를 선정함.



01. 변수 설명

총 35개의 변수, 국가별 주요 경제 지표들이 자세하게 나와있음.

country : 국가 이름
density : 단위 면적 당 인구 밀도
abbreviation : 국가 약자 or 코드
Agricultural Land (%) : 농업 목적용 토지 비율
Land Area : 국가 전체 토지 면적
Armed Forces Size : 군사력 규모
Birth Rate: 연간 1,000명당 태어나는 출생아 수
Calling Code : 국가 국제 번호
Capital/Major City : 수도/주요도시
CO2 Emissions : 이산화탄소 방출량(ton)
CPI: 소비자 물가 지수 ; 인프레이션과 구매력 측정
CPI Change (%) : 지난 해 대비 CPI 증감률
Currency_Code : 국가 통화 기호
Fertility Rate : 출산율
Forested Area (%) : 토지내 산림 면적 비율
Gasoline_Price: 리터당 휘발유 가격
GDP : 국내총생산
Gross Primary Education Enrollment (%) : GER ; 초등 교육 접근성
Gross Tertiary Education Enrollment (%) : GER ; 고등 교육 접근성
Infant Mortality : 신생아(1세 미만) 사망률
Largest City: 국가 내 가장 큰 도시
Life Expectancy : 출생 평균 기대수명
Maternal Mortality Ratio : MMR ; 100,000명당 모성 사망률
Minimum Wage: 최저 임금
Official Language: 공식 사용 언어
Out of Pocket Health Expenditure (%) : 의료 서비스 자가 부담률
Physicians per Thousand: 1000명 당 의사 수
Population : 인구
Population: Labor Force Participation (%) : 경제활동인구
Tax Revenue (%) : GDP 대비 세금 비율
Total Tax Rate : 총 기업이익 대비 세금 부담률
Unemployment Rate: 실업률
Urban Population: 도시 인구 비율
Latitude : 위도
Longitude : 경도

01. 데이터 전처리

데이터 로드 & 확인

비어있는 데이터가 있고,
결측값(NA)있는 데이터가 있음.

	Country	Density..P.Km2.	Abbreviation	Agricultural.Land....	Land.Area.Km2.	Armed.Forces.size	Birth.Rate	Calling.Code
56	Estonia	31	EE	23.10%	45,228	6,000	10.90	372
57	Eswatini	67			17,364		NA	268
58	Ethiopia	115	ET	36.30%	1,104,300	138,000	32.34	251
59	Fiji	49	FJ	23.30%	18,274	4,000	21.28	679
60	Finland	18	FI	7.50%	338,145	25,000	8.60	358
61	France	119	FR	52.40%	643,801	307,000	11.30	33
62	Gabon	9	GA	20.00%	267,667	7,000	31.61	241
63	The Gambia	239	GM	59.80%	11,300	1,000	38.54	220
64	Georgia	57	GE	34.50%	69,700	26,000	13.47	995
65	Germany	240	DE	47.70%	357,022	180,000	9.50	49
66	Ghana	137	GH	69.00%	238,533	16,000	29.41	233
67	Greece	81	GR	47.60%	131,957	146,000	8.10	30
68	Grenada	331	GD	23.50%	349		16.47	1
69	Guatemala	167	GT	36.00%	108,889	43,000	24.56	502
70	Guinea	53	GN	59.00%	245,857	13,000	36.36	224
71	Guinea-Bissau	70	GW	58.00%	36,125	4,000	35.13	245
72	Guyana	4	GY	8.60%	214,969	3,000	19.97	592
73	Haiti	414	HT	66.80%	27,750	0	24.35	509
74	Vatican City	2,003			0		NA	379
75	Honduras	89	HN	28.90%	112,090	23,000	21.60	504
76	Hungary	107	HU	58.40%	93,028	40,000	9.60	36
77	Iceland	3	IS	18.70%	103,000	0	12.00	354
78	India	464	IN	60.40%	3,287,263	3,031,000	17.86	91
79	Indonesia	151	ID	31.50%	1,904,569	676,000	18.07	62
80	Iran	52	IR	28.20%	1,648,195	563,000	18.78	98

01. 데이터 전처리

데이터 별 결측값 확인

```
> ##load data
> w ← read.csv("C:/mva/mva2/world_data_2023.csv")
> #NA check
> sapply(w, function(x) sum(is.na(x)))
                                    Country                                Density..P.Km2.
                                          0                                              0
                                Abbreviation                             Agricultural.Land....
                                          0                                              0
                             Land.Area.Km2.                                 Armed.Forces.size
                                          0                                              0
                                 Birth.Rate                                   Calling.Code
                                          6                                              1
                           Capital.Major.City                                Co2.Emissions
                                          0                                              0
                                         CPI                                  CPI.Change....
                                          0                                              0
                               Currency.Code                                 Fertility.Rate
                                          0                                              7
                            Forested.Area....                                 Gasoline.Price
                                          0                                              0
                                         GDP           Gross.primary.education.enrollment....
                                          0                                              0
  Gross.tertiary.education.enrollment....                                 Infant.mortality
                                          0                                              6
                                Largest.city                                Life.expectancy
                                          0                                              8
                  Maternal.mortality.ratio                                   Minimum.wage
                                         14                                              0
                         Official.language          Out.of.pocket.health.expenditure
                                          0                                              0
                   Physicians.per.thousand                                     Population
                                          7                                              0
Population..Labor.force.participation....                                   Tax.revenue....
                                          0                                              0
                               Total.tax.rate                              Unemployment.rate
                                          0                                              0
                           Urban_population                                       Latitude
                                          0                                              1
                                  Longitude
                                          1
```

01. 데이터 전처리

데이터 별 결측값 -> 제거
비어있는 관측값 -> 0으로 대체

data set 재확인

```
> #remove NA
> cw ← w[complete.cases(w),]
> #empty data → 0
> cw ← cw %>%
+   mutate(across(everything(), ~ ifelse(. == "" | is.na(.), 0, .)))
> #re-check
> sapply(cw, function(x) sum(is.na(x)))
                                Country                          Density..P.Km2.                             Abbreviation
                                      0                                        0                                        0
                    Agricultural.Land....                           Land.Area.Km2.                       Armed.Forces.size
                                      0                                        0                                        0
                              Birth.Rate                             Calling.Code                     Capital.Major.City
                                      0                                        0                                        0
                           Co2.Emissions                                      CPI                          CPI.Change....
                                      0                                        0                                        0
                           Currency.Code                          Fertility.Rate                       Forested.Area....
                                      0                                        0                                        0
                         Gasoline.Price                                      GDP       Gross.primary.education.enrollment....
                                      0                                        0                                        0
  Gross.tertiary.education.enrollment....                         Infant.mortality                             Largest.city
                                      0                                        0                                        0
                         Life.expectancy                    Maternal.mortality.ratio                          Minimum.wage
                                      0                                        0                                        0
                       Official.language             Out.of.pocket.health.expenditure                 Physicians.per.thousand
                                      0                                        0                                        0
                              Population      Population..Labor.force.participation....                     Tax.revenue....
                                      0                                        0                                        0
                        Total.tax.rate                        Unemployment.rate                          Urban_population
                                      0                                        0                                        0
                              Latitude                               Longitude
                                      0                                        0
```

01. 데이터 전처리

데이터 구조 확인

총 179개 관측치
35개 변수

(세계 각국의 여러가지 지표)

```
> str(cw)
'data.frame':   179 obs. of  35 variables:
 $ Country                                 : chr  "Afghanistan" "Albania" "Algeria" "Angola"  ...
 $ Density..P.Km2.                         : chr  "60" "105" "18" "26"  ...
 $ Abbreviation                            : chr  "AF" "AL" "DZ" "AO"  ...
 $ Agricultural.Land....                   : chr  "58.10%" "43.10%" "17.40%" "47.50%"  ...
 $ Land.Area.Km2.                          : chr  "652,230" "28,748" "2,381,741" "1,246,700"  ...
 $ Armed.Forces.size                       : chr  "323,000" "9,000" "317,000" "117,000"  ...
 $ Birth.Rate                              : num  32.5 11.8 24.3 40.7 15.3  ...
 $ Calling.Code                            : int  93 355 213 244 1 54 374 61 43 994  ...
 $ Capital.Major.City                      : chr  "Kabul" "Tirana" "Algiers" "Luanda"  ...
 $ Co2.Emissions                           : chr  "8,672" "4,536" "150,006" "34,693"  ...
 $ CPI                                     : chr  "149.9" "119.05" "151.36" "261.73"  ...
 $ CPI.Change....                          : chr  "2.30%" "1.40%" "2.00%" "17.10%"  ...
 $ Currency.Code                           : chr  "AFN" "ALL" "DZD" "AOA"  ...
 $ Fertility.Rate                          : num  4.47 1.62 3.02 5.52 1.99 2.26 1.76 1.74 1.47 1.73  ...
 $ Forested.Area....                       : chr  "2.10%" "28.10%" "0.80%" "46.30%"  ...
 $ Gasoline.Price                          : chr  "$0.70 " "$1.36 " "$0.28 " "$0.97 "  ...
 $ GDP                                     : chr  "$19,101,353,833 " "$15,278,077,447 " "$169,988,236,398 " "$94,635,415,870 "  ...
 $ Gross.primary.education.enrollment....  : chr  "104.00%" "107.00%" "109.90%" "113.50%"  ...
 $ Gross.tertiary.education.enrollment....  : chr  "9.70%" "55.00%" "51.40%" "9.30%"  ...
 $ Infant.mortality                        : num  47.9 7.8 20.1 51.6 5 8.8 11 3.1 2.9 19.2  ...
 $ Largest.city                            : chr  "Kabul" "Tirana" "Algiers" "Luanda"  ...
 $ Life.expectancy                         : num  64.5 78.5 76.7 60.8 76.9 76.5 74.9 82.7 81.6 72.9  ...
 $ Maternal.mortality.ratio                : int  638 15 112 241 42 39 26 6 5 26  ...
 $ Minimum.wage                            : chr  "$0.43 " "$1.12 " "$0.95 " "$0.71 "  ...
 $ Official.language                       : chr  "Pashto" "Albanian" "Arabic" "Portuguese"  ...
 $ Out.of.pocket.health.expenditure        : chr  "78.40%" "56.90%" "28.10%" "33.40%"  ...
 $ Physicians.per.thousand                 : num  0.28 1.2 1.72 0.21 2.76 3.96 4.4 3.68 5.17 3.45  ...
 $ Population                              : chr  "38,041,754" "2,854,191" "43,053,054" "31,825,295"  ...
 $ Population..Labor.force.participation....: chr  "48.90%" "55.70%" "41.20%" "77.50%"  ...
 $ Tax.revenue....                         : chr  "9.30%" "18.60%" "37.20%" "9.20%"  ...
 $ Total.tax.rate                          : chr  "71.40%" "36.60%" "66.10%" "49.10%"  ...
 $ Unemployment.rate                       : chr  "11.12%" "12.33%" "11.70%" "6.89%"  ...
 $ Urban_population                        : chr  "9,797,273" "1,747,593" "31,510,100" "21,061,025"  ...
 $ Latitude                                : num  33.9 41.2 28 -11.2 17.1  ...
 $ Longitude                               : num  67.71 20.17 1.66 17.87 -61.8  ...
```

01. 데이터 전처리

필요한 패키지 로딩

변수들 중에서,
'%,' , '$' 등의 기호들을
제거하여 숫자형으로 변환

```
> library(ggplot2)
> library(dplyr)
> library(ggrepel)
>
> #변수 선택
> convert ← c("Density..P.Km2.", "Agricultural.Land....", "Land.Area.Km2.","Armed.Force
s.size","Co2.Emissions","CPI","CPI.Change....","Forested.Area....","Gasoline.Price","GD
P","Gross.primary.education.enrollment....","Gross.tertiary.education.enrollment....","Li
fe.expectancy","Minimum.wage","Out.of.pocket.health.expenditure","Tax.revenue....","Tota
l.tax.rate","Unemployment.rate","Urban_population","Population", "Population..Labor.forc
e.participation....")
>
> #선택한 열들만 %와 , 제거하고 숫자형으로 변환
> cw[convert] ← lapply(cw[convert], function(x) {
+     as.numeric(gsub("[%|,|$]", "", x))
+ })
```

01. 데이터 전처리

변수 명들이 너무 길어서
재조정.

그후 summary() 함수로
대략적인 요약 통계량 파악

```
> ##column re-name
> g ← g %>%
+   rename(pd = Density..P.Km2., Agric_Land = Agricultural.Land...., Land = Land.Area.Km2.,CPI_R = CPI.
Change....,forest_R = Forested.Area...., GER_P = Gross.primary.education.enrollment...., GER_T = Gross.
tertiary.education.enrollment...., OOP = Out.of.pocket.health.expenditure, LFP = Population..Labor.forc
e.participation....,  TR = Tax.revenue....)
> global ← g
> summary(global)
   Country                pd          Abbreviation         Agric_Land
 Length:179         Min.   :   2   Length:179         Min.   : 0.60
 Class :character   1st Qu.:  32   Class :character   1st Qu.:21.60
 Mode  :character   Median :  83   Mode  :character   Median :39.70
                    Mean   : 209                      Mean   :39.24
                    3rd Qu.: 185                      3rd Qu.:55.65
                    Max.   :8358                      Max.   :82.60
      Land          Armed.Forces.size   Birth.Rate      Calling.Code
 Min.   :     298   Min.   :      0   Min.   : 6.40   Min.   :   1.0
 1st Qu.:   32190   1st Qu.:   9000   1st Qu.:11.45   1st Qu.:  73.5
 Median :  163610   Median :  27000   Median :18.18   Median : 252.0
 Mean   :  743555   Mean   : 151117   Mean   :20.43   Mean   : 358.5
 3rd Qu.:  584386   3rd Qu.: 131500   3rd Qu.:28.70   3rd Qu.: 506.5
 Max.   :17098240   Max.   :3031000   Max.   :46.08   Max.   :1876.0
 Capital.Major.City Co2.Emissions          CPI             CPI_R
 Length:179         Min.   :     66   Min.   :   0.0   Min.   : -4.300
 Class :character   1st Qu.:   3230   1st Qu.: 113.3   1st Qu.:  0.800
 Mode  :character   Median :  16670   Median : 124.3   Median :  2.300
                    Mean   : 186721   Mean   : 160.3   Mean   :  5.616
                    3rd Qu.:  69944   3rd Qu.: 155.9   3rd Qu.:  4.000
                    Max.   :9893038   Max.   :2740.3   Max.   :254.900
 Currency.Code      Fertility.Rate     forest_R     Gasoline.Price
 Length:179         Min.   :0.980   Min.   : 0.00   Min.   :0.0000
 Class :character   1st Qu.:1.710   1st Qu.:10.95   1st Qu.:0.7150
 Mode  :character   Median :2.260   Median :31.30   Median :0.9700
                    Mean   :2.704   Mean   :31.26   Mean   :0.9411
                    3rd Qu.:3.580   3rd Qu.:47.45   3rd Qu.:1.1850
                    Max.   :6.910   Max.   :98.30   Max.   :2.0000
```

01. 데이터 전처리

변수 명들이 너무 길어서
재조정.

그후 summary() 함수로
대략적인 요약 통계량 파악

```
      GDP                  GER_P           GER_T        Infant.mortality
 Min.   :     194647202   Min.   :  0.0   Min.   :  0.00   Min.   : 1.40
 1st Qu.:   12673957646   1st Qu.: 98.9   1st Qu.: 11.60   1st Qu.: 6.05
 Median :   40761142857   Median :102.3   Median : 28.50   Median :13.90
 Mean   :  514362054777   Mean   :102.0   Mean   : 37.34   Mean   :21.39
 3rd Qu.:  255999344430   3rd Qu.:107.5   3rd Qu.: 63.15   3rd Qu.:33.25
 Max.   :21427700000000   Max.   :142.5   Max.   :136.60   Max.   :84.50
 Largest.city       Life.expectancy Maternal.mortality.ratio  Minimum.wage
 Length:179         Min.   :52.80   Min.   :   2.0           Min.   : 0.000
 Class :character   1st Qu.:67.00   1st Qu.:  12.5           1st Qu.: 0.175
 Mode  :character   Median :73.40   Median :  53.0           Median : 0.600
                    Mean   :72.28   Mean   : 155.0           Mean   : 1.681
                    3rd Qu.:77.50   3rd Qu.: 185.5           3rd Qu.: 1.600
                    Max.   :84.20   Max.   :1140.0           Max.   :13.590
 Official.language       OOP        Physicians.per.thousand   Population
 Length:179         Min.   : 0.00   Min.   :0.010           Min.   :     97118
 Class :character   1st Qu.:17.30   1st Qu.:0.310           1st Qu.:   2901235
 Mode  :character   Median :30.90   Median :1.510           Median :   9770529
                    Mean   :32.61   Mean   :1.808           Mean   :  42599589
                    3rd Qu.:44.05   3rd Qu.:2.905           3rd Qu.:  31308952
                    Max.   :81.60   Max.   :8.420           Max.   :1397715000
      LFP              TR         Total.tax.rate   Unemployment.rate
 Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   : 0.000
 1st Qu.:55.50   1st Qu.:10.15   1st Qu.: 29.80   1st Qu.: 3.315
 Median :62.10   Median :14.80   Median : 37.10   Median : 5.110
 Mean   :60.96   Mean   :14.86   Mean   : 39.77   Mean   : 6.628
 3rd Qu.:68.80   3rd Qu.:20.55   3rd Qu.: 47.60   3rd Qu.: 9.040
 Max.   :86.80   Max.   :37.20   Max.   :219.60   Max.   :28.180
 Urban_population      Latitude        Longitude
 Min.   :    23800   Min.   :-40.901   Min.   :-175.198
 1st Qu.:  1610847   1st Qu.:  4.373   1st Qu.:  -8.837
 Median :  5256027   Median : 17.190   Median :  20.939
 Mean   : 23660469   Mean   : 18.944   Mean   :  18.077
 3rd Qu.: 16278774   3rd Qu.: 39.237   3rd Qu.:  48.047
 Max.   :842933962   Max.   : 64.963   Max.   : 178.065
```

01. 데이터 전처리

변경한 열 이름

```
> print(colnames(global))
 [1] "Country"                    "pd"
 [3] "Abbreviation"               "Agric_Land"
 [5] "Land"                       "Armed.Forces.size"
 [7] "Birth.Rate"                 "Calling.Code"
 [9] "Capital.Major.City"         "Co2.Emissions"
[11] "CPI"                        "CPI_R"
[13] "Currency.Code"              "Fertility.Rate"
[15] "forest_R"                   "Gasoline.Price"
[17] "GDP"                        "GER_P"
[19] "GER_T"                      "Infant.mortality"
[21] "Largest.city"               "Life.expectancy"
[23] "Maternal.mortality.ratio"   "Minimum.wage"
[25] "Official.language"          "OOP"
[27] "Physicians.per.thousand"    "Population"
[29] "LFP"                        "TR"
[31] "Total.tax.rate"             "Unemployment.rate"
[33] "Urban_population"           "Latitude"
[35] "Longitude"
```

02. 주성분 분석

주성분 분석을 통한 변수별 특징과 패턴 파악

02. 주성분 분석

주성분 분석에
사용할 데이터 선택
-> 총 26개 변수

어떤 데이터가
서로 관련성이 있는지
쉽게 판단하기 어려운 관계로,
데이터 중 수치형 변수들을
선택해서 주성분 분석 진행

변수들이 서로 다른 단위를
가지고 있기 때문에
주성분 분석을 하기 전
표준화를 후 분석.

```
> #PCA data choice
>
> pca_data ← global[,c("pd", "Agric_Land", "Land","Armed.Forces.size","Co2.E
missions","CPI","CPI_R","forest_R","Gasoline.Price","GDP","GER_P","GER_T","Li
fe.expectancy","Minimum.wage","OOP","TR","Total.tax.rate","Unemployment.rat
e","Urban_population","Fertility.Rate","Infant.mortality","Physicians.per.tho
usand", "Population", "LFP", "Birth.Rate","Maternal.mortality.ratio")]
> #standardized
> scaled_data ← scale(pca_data)
> pca_result ← prcomp(scaled_data, center =T, scale=T)
```

02. 주성분 분석

상관행렬을 계산해보아도,
변수들이 많아서
한 눈에 보기 어려움.

```
> #cor
> cor(global[,c("pd", "Agric_Land", "Land","Armed.Forces.size","Co2.Emissions","CPI","CPI_R","forest_R","Gasoline.Price","GDP","GER_P","GER_T","Life.expectancy","Minimum.wage","OOP","TR","Total.tax.rate","Unemployment.rate","Urban_population","Fertility.Rate","Infant.mortality","Physicians.per.thousand", "Population", "LFP", "Birth.Rate","Maternal.mortality.ratio")])
                                pd   Agric_Land         Land Armed.Forces.size Co2.Emissions           CPI
pd                       1.000000000 -0.112039317 -0.082828400      -0.001050169  -0.014582272 -0.0388551400
Agric_Land              -0.112039317  1.000000000 -0.033833128       0.042117184   0.062087184 -0.0369234067
Land                    -0.082828400 -0.033833128  1.000000000       0.555336161   0.589660029  0.0471876634
Armed.Forces.size       -0.001050169  0.042117184  0.555336161       1.000000000   0.741587367  0.0552346890
Co2.Emissions           -0.014582272  0.062087184  0.589660029       0.741587367   1.000000000 -0.0028900490
CPI                     -0.038855140 -0.036923407  0.047187663       0.055234689  -0.002890049  1.0000000000
CPI_R                   -0.046575669 -0.024705951  0.043659526       0.067621405   0.006605181  0.9319488849
forest_R                -0.097674026 -0.457243564 -0.002765995      -0.032038857  -0.021094969  0.0186567806
Gasoline.Price           0.073517276  0.075068848 -0.107880818      -0.084705203  -0.029900228 -0.1793540697
GDP                     -0.008433727  0.051520341  0.548435553       0.607110430   0.916662015 -0.0179800197
GER_P                    0.017218937 -0.026214495  0.019547614       0.059798821   0.007498681 -0.0004265569
GER_T                    0.114511696 -0.084109901  0.224860340       0.120044752   0.159540686  0.0413502515
Life.expectancy          0.167249298 -0.235022615  0.060322886       0.080835289   0.122514410 -0.0908555250
Minimum.wage            -0.027559508  0.021285793  0.169197585       0.003843914   0.107201426 -0.0914769124
OOP                      0.022377145  0.136591136 -0.020370883       0.116836593  -0.029835001  0.1243859563
TR                       0.004038936 -0.042825736 -0.099835307      -0.147127311  -0.079309798 -0.1442516208
Total.tax.rate          -0.072142924  0.133135511  0.099855174       0.065981342   0.069428747  0.1448885595
Unemployment.rate       -0.107047722  0.082230846  0.059179422      -0.010498237   0.017377900  0.1116099120
Urban_population        -0.005875229  0.101903301  0.544223854       0.852175376   0.926191297  0.0181195876
Fertility.Rate          -0.150135187  0.162623017 -0.071163935      -0.150851339  -0.144544926  0.0660147219
Infant.mortality        -0.119287974  0.200919872 -0.073163919      -0.079230725  -0.124621260  0.0847045354
Physicians.per.thousand  0.031214790 -0.013570020  0.083121420       0.030668220   0.060919970 -0.0936974547
Population               0.007269724  0.118111125  0.442656598       0.876524115   0.809473066  0.0181812393
LFP                      0.023833734  0.004884238  0.039316497      -0.008189825   0.016730841 -0.0260067188
Birth.Rate              -0.147876955  0.179358061 -0.079698222      -0.148760002  -0.161263184  0.0783977823
Maternal.mortality.ratio -0.090353052  0.213667649 -0.056198582      -0.105047907  -0.109038094  0.0557894602
```

02. 주성분 분석

표준화를 진행하고,
주성분 분석 진행 후 결과 확인

개별 분산과 주성분 점수 확인

```
> #standardized
> scaled_data ← scale(pca_data)
> pca_result ← prcomp(scaled_data, center =T, scale=T)
>
>
> #주성분의 표준편차와 주성분 점수
> print(pca_result)
Standard deviations (1, .., p=26):
 [1] 2.55654241 2.14533057 1.45410226 1.29717692 1.14849229 1.12843715 1.05156428 1.03474492 0.97856613 0.93528225
[11] 0.89407868 0.75947793 0.73401002 0.71642322 0.66964414 0.63899678 0.57454504 0.49238444 0.46199941 0.35955239
[21] 0.31571958 0.24517725 0.22487340 0.17759461 0.10905404 0.09653264

Rotation (n x k) = (26 x 26):
                                PC1          PC2          PC3          PC4         PC5          PC6          PC7
pd                         0.05138183 -0.032490010  0.031370159 -0.015398456  0.36665499 -0.4342745208  0.0429609276
Agric_Land                -0.06841364  0.076257439  0.078557087  0.537850876 -0.17352134 -0.0888131434 -0.2022792789
Land                       0.09480613  0.294483938 -0.019693010 -0.050163312 -0.13588879  0.0829072319  0.0398232723
Armed.Forces.size          0.10645427  0.393380948 -0.004758342 -0.039543255  0.12231001 -0.0335251718 -0.0107266000
Co2.Emissions              0.13280202  0.411119631  0.066778078 -0.033151722 -0.04172676  0.0297509231  0.0654808517
CPI                       -0.04555339  0.053572604 -0.623654006 -0.052700509 -0.18587924 -0.1316377520  0.0666910112
CPI_R                     -0.03454727  0.054193367 -0.625964626 -0.042533557 -0.20390338 -0.1444456448  0.0696371352
forest_R                   0.02633325 -0.036630924 -0.022605571 -0.572528866 -0.10680998  0.2516747996 -0.0185202238
Gasoline.Price             0.10406611 -0.076788345  0.204530429  0.204646189 -0.36472816 -0.1890468661  0.1205588724
GDP                        0.14620568  0.357670660  0.071698698 -0.024973758 -0.14009787  0.0445870901  0.0842021276
GER_P                      0.01846355  0.007129071  0.046267227 -0.281612581 -0.10579190 -0.2148694645 -0.5915770492
GER_T                      0.31554862 -0.039164413 -0.133394338  0.134625962 -0.08240535 -0.0671409988 -0.0719245879
Life.expectancy            0.36178312 -0.094827007 -0.031476914 -0.002153644  0.04914579 -0.0625657431 -0.0281647543
Minimum.wage               0.20426370 -0.030276556  0.063307851  0.115070366 -0.39711584  0.0471695573 -0.0809312412
OOP                       -0.13028504  0.091316532 -0.158638333  0.271962417  0.41075577 -0.0849035163 -0.2076259428
TR                         0.15579335 -0.138975418  0.122933526 -0.024364486 -0.19569131  0.2501788594 -0.2285652895
Total.tax.rate            -0.07321321  0.098394813 -0.175937172  0.132633561 -0.06236293  0.1514515509 -0.6027009028
Unemployment.rate          0.01544215  0.005124778 -0.141177297  0.267495548 -0.04288713  0.3623987920  0.2876620077
Urban_population           0.10948274  0.431263410  0.055848164 -0.016451407  0.02541984  0.0001067631 -0.0005756752
Fertility.Rate            -0.35151538  0.070639738  0.077239197 -0.006851614 -0.15298966  0.0241619636  0.0140478977
Infant.mortality          -0.35826161  0.098148479  0.040366428  0.038166165 -0.08215704  0.0224980653  0.0248356747
Physicians.per.thousand    0.29898318 -0.094983433 -0.035069712  0.169606526 -0.05813472 -0.0814989383  0.0182359225
Population                 0.07820182  0.412964720  0.058906972 -0.004971668  0.09112207 -0.0301140146 -0.0290833242
LFP                       -0.04529334  0.020601430  0.136416744 -0.146156400 -0.27746739 -0.6014770590  0.1034431718
Birth.Rate                -0.36307009  0.070814237  0.062246206 -0.015667234 -0.11231038  0.0185145958 -0.0203817763
Maternal.mortality.ratio  -0.32543632  0.084857366  0.075624174  0.062208726 -0.19279161 -0.0611074119  0.0552033769
```

02. 주성분 분석

주성분 분석에 대한 결과 요약

PC1이 약 25%,
PC2가 약 17%,
전체 분산에 대한 설명 비율을 가짐.

일반적으로 80%-90%
정도 설명 분산을 가지면
설명력이 있다고 판단을 하므로,

현재 결과로는 PC13은
되어야 전체 데이터에 대하여
적어도 90%는 설명한다고
볼 수 있다.

```
> summary(pca_result)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12
Standard deviation     2.5565 2.1453 1.45410 1.29718 1.14849 1.12844 1.05156 1.03474 0.97857 0.93528 0.89408 0.75948
Proportion of Variance 0.2514 0.1770 0.08132 0.06472 0.05073 0.04898 0.04253 0.04118 0.03683 0.03364 0.03075 0.02218
Cumulative Proportion  0.2514 0.4284 0.50972 0.57444 0.62517 0.67415 0.71668 0.75786 0.79469 0.82833 0.85908 0.88126
                          PC13    PC14    PC15   PC16   PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24
Standard deviation     0.73401 0.71642 0.66964 0.6390 0.5745 0.49238 0.46200 0.35955 0.31572 0.24518 0.22487 0.17759
Proportion of Variance 0.02072 0.01974 0.01725 0.0157 0.0127 0.00932 0.00821 0.00497 0.00383 0.00231 0.00194 0.00121
Cumulative Proportion  0.90199 0.92173 0.93897 0.9547 0.9674 0.97670 0.98491 0.98988 0.99371 0.99603 0.99797 0.99918
                          PC25    PC26
Standard deviation     0.10905 0.09653
Proportion of Variance 0.00046 0.00036
Cumulative Proportion  0.99964 1.00000
> #pc12 정도 되어야 설명 분산 비율이 88%되며, pc13은 설명분산 비율이 90%정도 된다.
```

02. 주성분 분석

각각 고유값 & 설명분산

```
> #고유값
> eigenvalues ← pca_result$sdev^2
> eigenvalues
 [1] 6.535909107 4.602443265 2.114413392 1.682667971 1.319034542 1.273370399 1.105787425 1.070697050 0.957591670
[10] 0.874752888 0.799376689 0.576806727 0.538770706 0.513262228 0.448423272 0.408316888 0.330102008 0.242442439
[19] 0.213443452 0.129277924 0.099678853 0.060111882 0.050568044 0.031539846 0.011892783 0.009318551
> #각각의 설명분산
> explained_variance ← eigenvalues / sum(eigenvalues)
> print(explained_variance)
 [1] 0.2513811195 0.1770170486 0.0813235920 0.0647179989 0.0507320978 0.0489757846 0.0425302856 0.0411806558
 [9] 0.0368304488 0.0336443418 0.0307452573 0.0221848741 0.0207219502 0.0197408549 0.0172470489 0.0157044957
[17] 0.0126962311 0.0093247092 0.0082093635 0.0049722278 0.0038338021 0.0023119955 0.0019449248 0.0012130710
[25] 0.0004574147 0.0003584058
> pca_data ← as.data.frame(pca_result$x[, 1:26])
> pca_data$Country ← global$Country
```

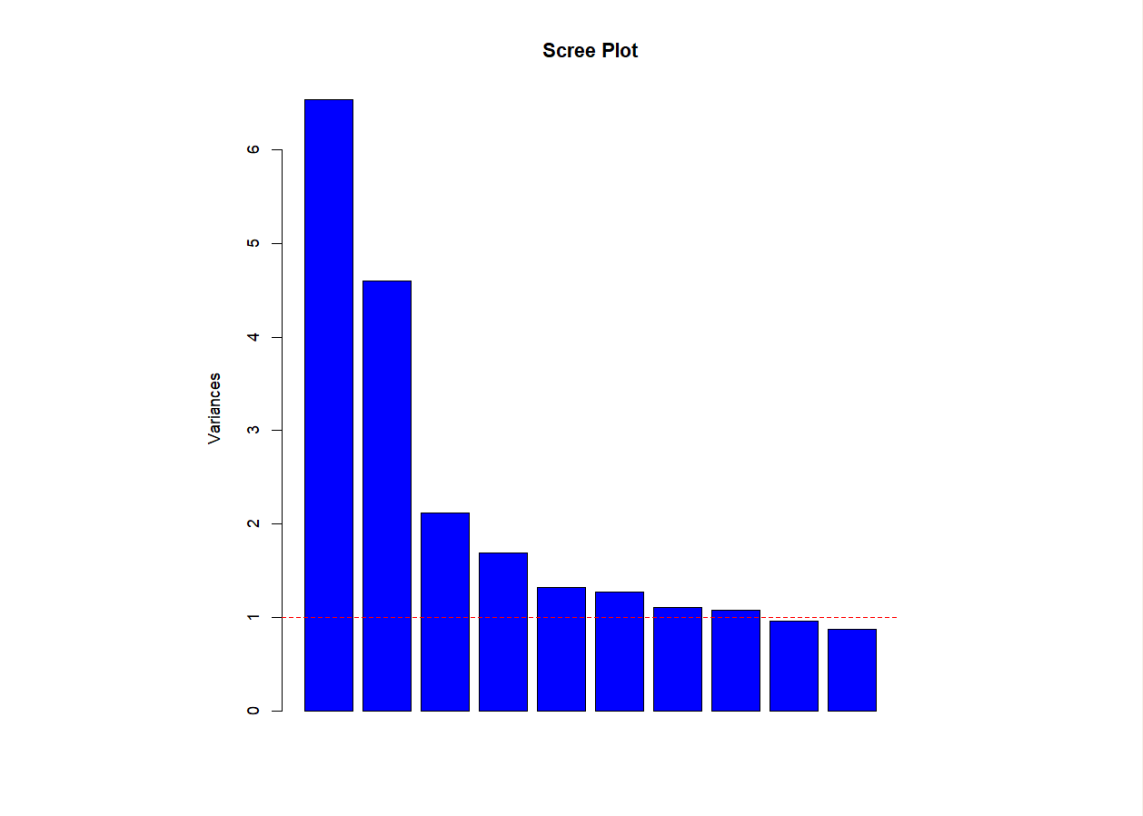
02. 주성분 분석

scree 도표

주성분 표준편차의 제곱이 고유값을 의미하며,
kaiser 기준에 따라
고유값이 1 이상인 주성분 선택

대략적으로 8번째까지
1근처에 있는 것을 알 수 있음

```
> #scree plot
> screeplot(pca_result, main = "Scree Plot", col = "blue", pch = 16)
> abline(h = 1, col = "red", lty = 2)
```



02. 주성분 분석

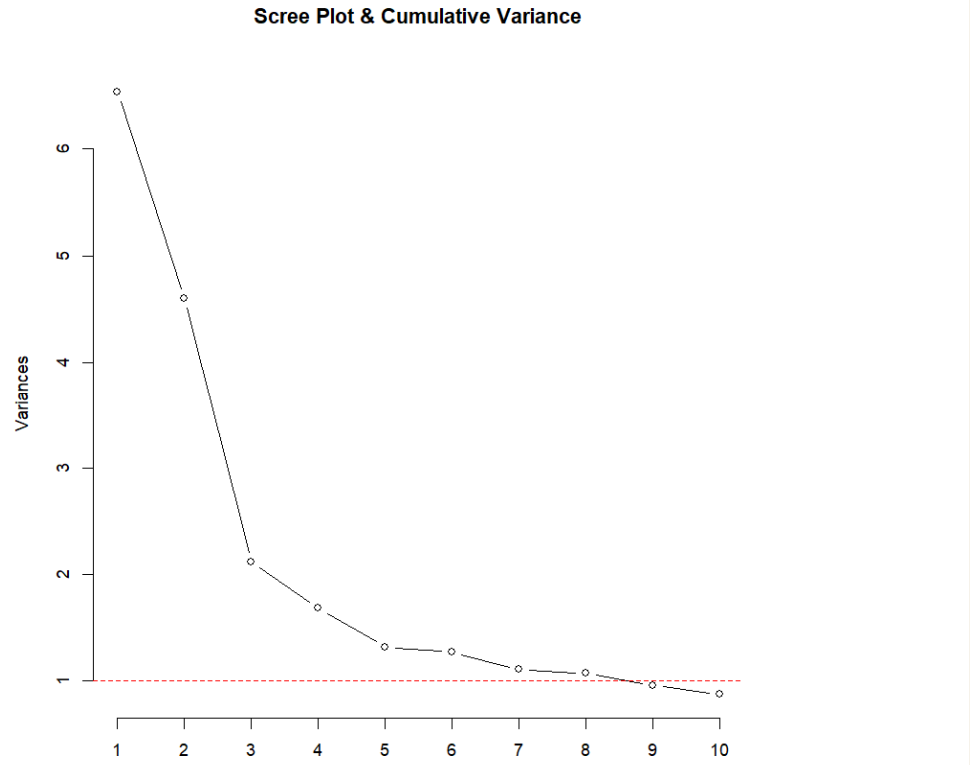
누적분산

누적 분산 비율에서 볼 때,
csree 도표와 비슷한 경향을 보이고,

그 개수가 급격히 감소하다가
완만해지는 포인트를 찾으면,
그 개수가 5개 또는 6개 정도이다.

결론적으로, 유의미한 주성분의 개수는
6개에서 8개 사이라고 추측할 수 있다.

```
> #누적분산 비율
> plot(pca_result, type = "l", main="Scree Plot & Cumulative Variance")
> abline(h = 1, col = "red", lty = 2)
```



02. 주성분 분석

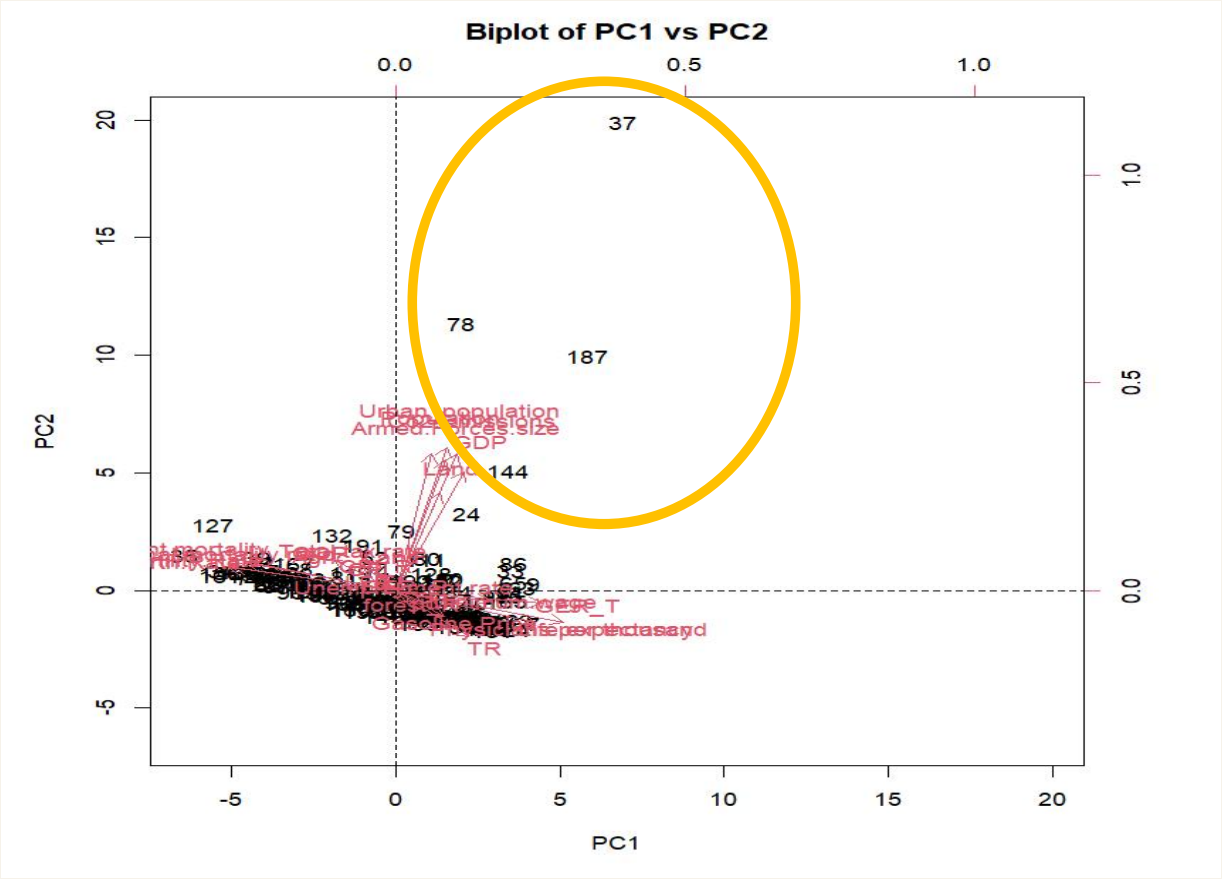
biplot(행렬도)

biplot으로 알 수 있는 것은,
특정 변수들의 관계성과 경향성,
PC1, PC2에 기여하는 정도가
높은 관측치를 알 수 있다.

그래프를 볼 때,
37, 78, 187,144 정도가
눈에 띄게 나타나는데,

이를 실제 데이터에서 해당 국가를 살펴보면,
37은 중국, 78은 인도,
144는 러시아, 187은 미국을 나타낸다.

```
> biplot(pca_result, scale = 0, main="Biplot of PC1 vs PC2")
> abline(h=0,v=0, lty=2)
> #시각화를 했을 때, 37, 187이 확연히 다른 국가들의 중심 포인트와 달랐고,
> #PC1, PC2에 기여하는 정도가 다른 관측치들에 비해 높다는 것을 알 수 있다.
> #추가로, PC2에 기여하는 정도가 높은 관측치는 144,24 정도이다.
```

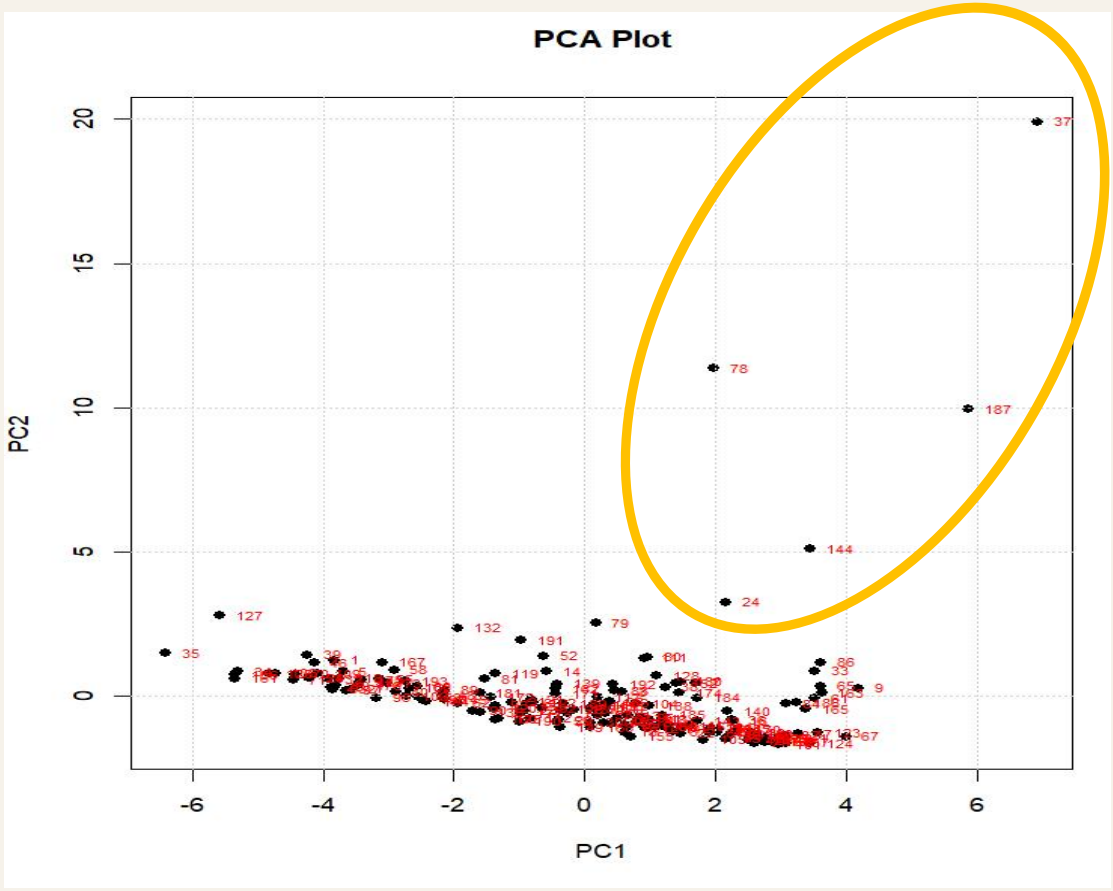


02. 주성분 분석

좀 더 자세히 PCA 그래프를 보면,

37,187,78,144,24번 등 앞에서 보았던
관측치들을 다시 확인 할 수 있다.

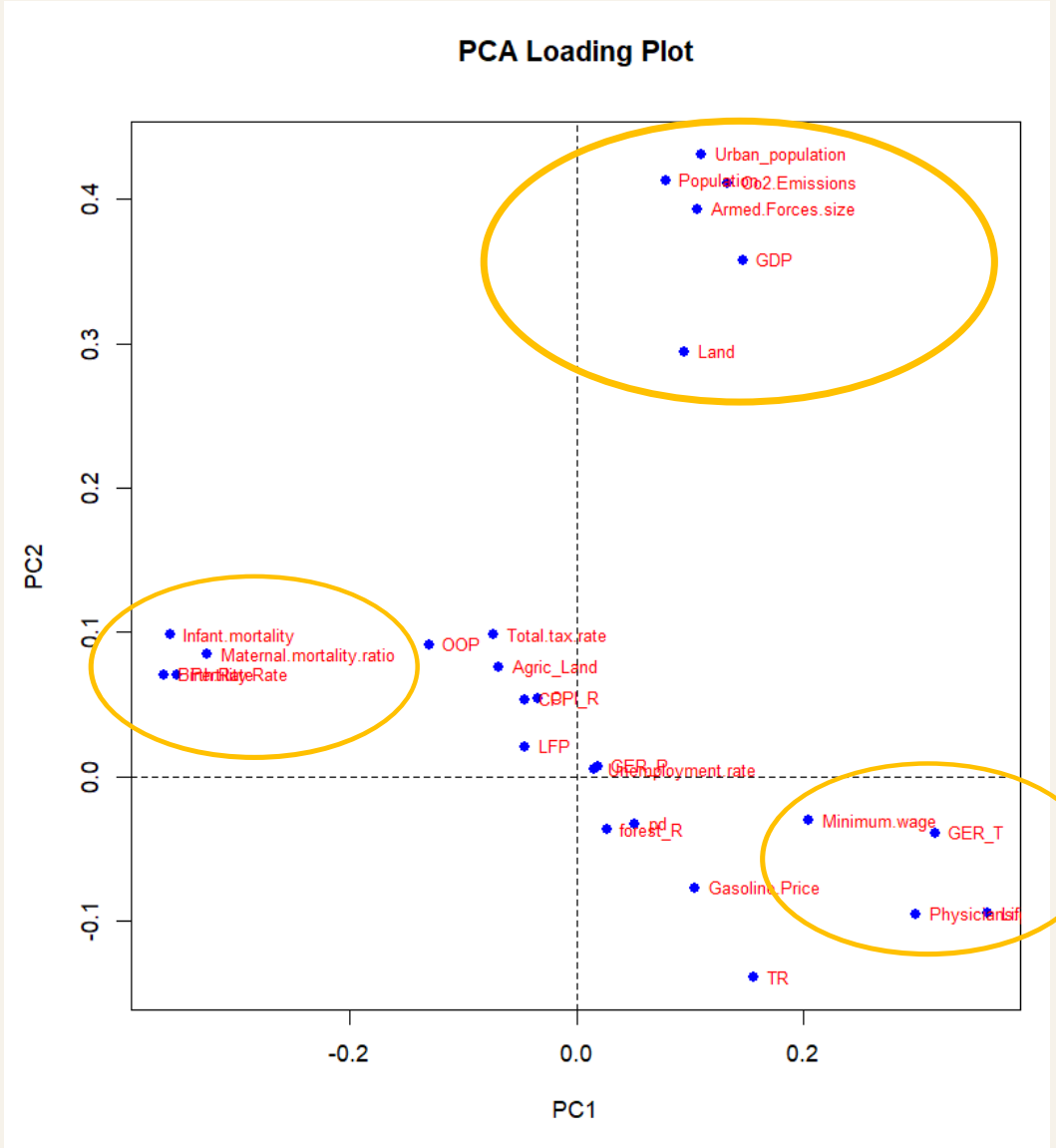
```
> #PCA visualization
> plot(pca_result$x[, 1], pca_result$x[, 2],
+      xlab = "PC1", ylab = "PC2",
+      main = "PCA Plot",
+      pch = 19, col = "black")
> grid()
> text(pca_result$x[, 1], pca_result$x[, 2],
+      labels = rownames(pca_result$x),
+      cex = 0.6, pos = 4, col = "red")
```



02. 주성분 분석

각 변수별로 비슷한 패턴을
가지고 있는 변수들의 분포를 확인할 수 있다.

```
> #loading visualization
> plot(loadings[, 1], loadings[, 2],
+      xlab = "PC1", ylab = "PC2",
+      main = "PCA Loading Plot",
+      pch = 19, col = "blue")
> text(loadings[, 1], loadings[, 2],
+      labels = rownames(loadings),
+      cex = 0.7, pos = 4, col = "red")
> abline(h=0,v=0, lty=2)
```



03. 결론

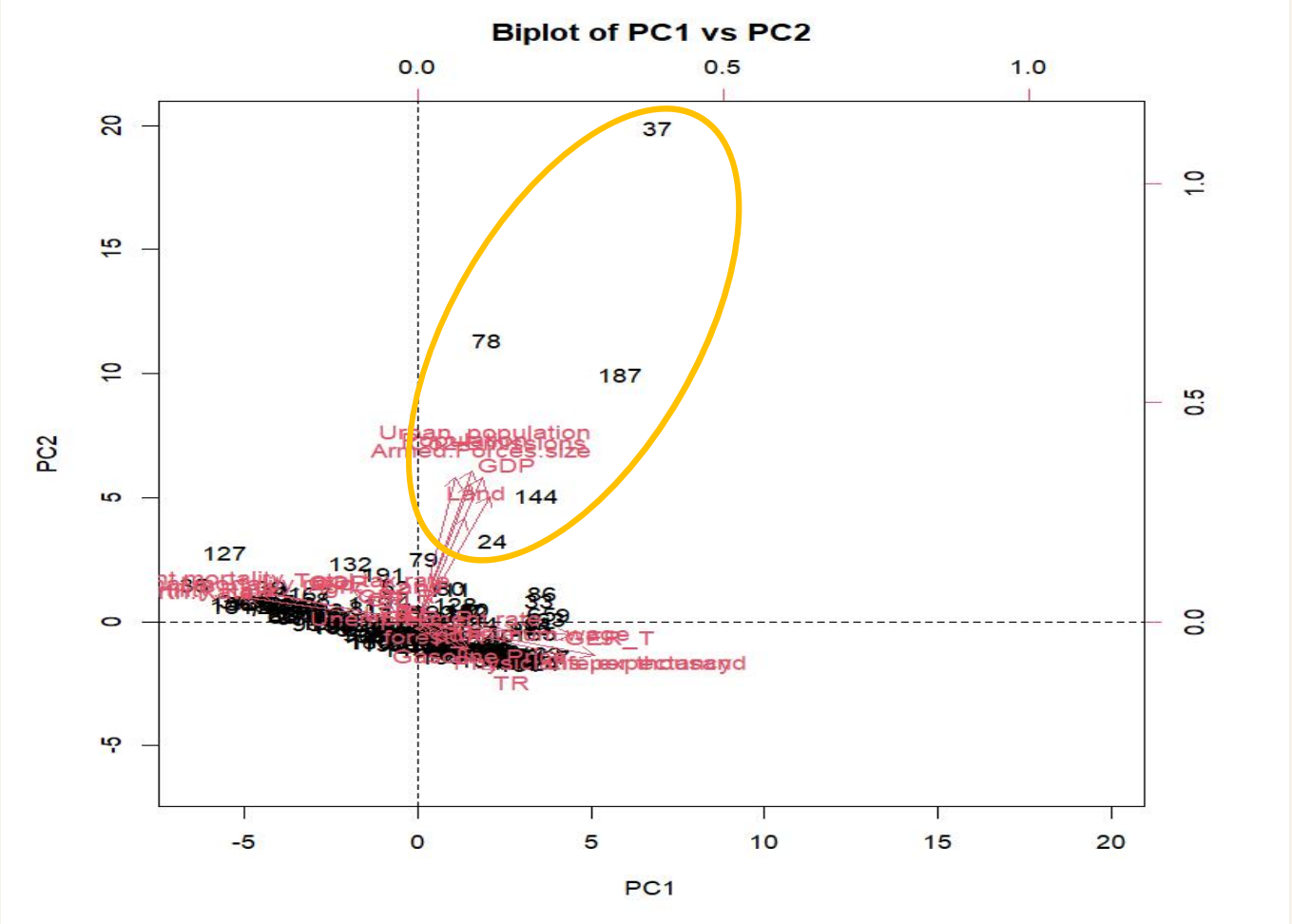
주성분 분석에 대한 결과 해석

03. 결론

각 국가별 관측치들을 볼 때,
PC1과 PC2에 영향을 많이 주는
관측치들은 37,78,187,144였고,
GDP, Land 등 특정 변수 들에서
크게 영향을 준다는 것을 알 수 있었고,

실제로 loading plot에서 밀접한 관련이
있다고 보이는 총 인구, co2배출량,
도시 인구, 군사력, GDP, 토지면적이
오른쪽 그래프에서와 동일한 결과가
나타난다는 것을 알 수 있다.

즉, 해당 변수들에서,
중국, 인도, 러시아, 미국의 GDP와
군사력 등이 다른 나라와 비교할 때
큰 차이가 있는 것을 알 수 있다.



03. 결론

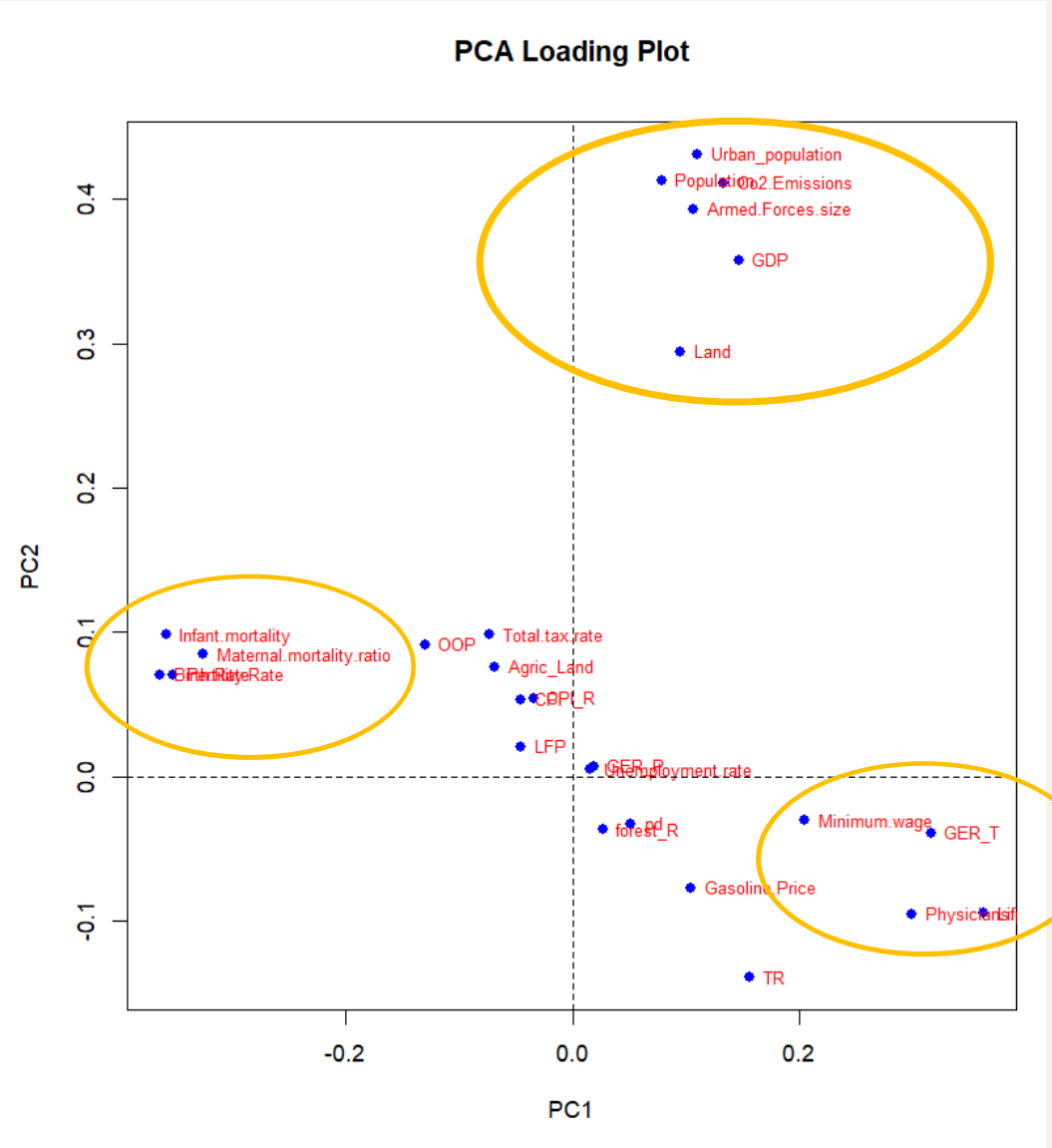
각 변수별 주성분 점수에 대한
loading plot을 보면,

PC1과 PC2 모두 중요한 영향을 미치는 변수들은
총 인구, co2배출량, 도시 인구, 군사력, GDP,
토지 면적이고 이들이 비슷한 패턴을 가진 다는 것을
알 수 있고, (인구 - 경제)

최저임금, 인구 1000명 당 의사 수, 고등 교육
보급률 변수들은 PC1에만 영향을 주는 것을
알 수 있고 이들 간에 관계성이 있다. (경제-의료)

신생아 사망률, 인구 100,000명 당 모성 사망률,
연간 1000명당 태어나는 출생아 수, 출산율이
서로 관계성이 있을 것이라고 생각할 수 있다. (복지)

또한, 가운데에 분포하는 변수들은
주성분 설명의 변동성과 관련이 적고,
상관관계가 적은 것을 알 수 있다.

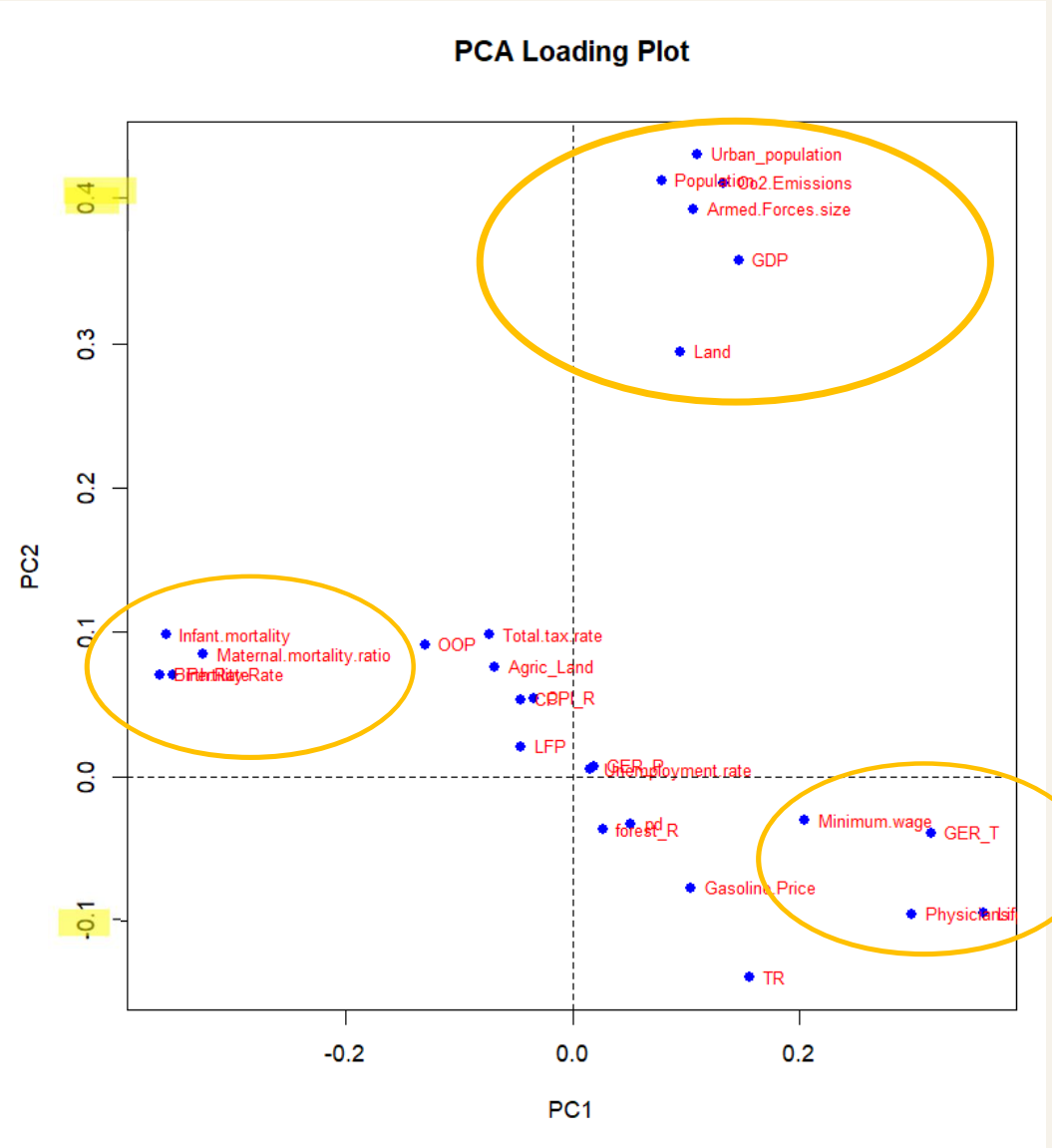


03. 결론

이에 대하여 알 수 있는 것은,

우선, PC1에 기여하고 있는 변수들이 많이 있지만,
그 중에서 최저임금, 고등교육 보급률, 인구
1000명당 의사 수가 높게 기여하는 것을 보았을 때,
이는 이 첫번째 주성분은 경제적-사회적 현상을
설명하는 변수로 생각할 수 있으며 이 값이 커질수록
교육과 의료 수준은 높고,
이에 따라 영아사망률, 모성사망률이 낮고,
출산율과 연간 신생아 수가 낮다는 것을 알 수 있다.

한편, PC2에 기여하는 변수들은 GDP, 군사력 등
6개의 변수가 있는데,
이는 이 두번째 주성분은 인구와 토지를 비롯한
국가의 경제력 등과 관련된 지표를 담고 있을
가능성이 높다.
그러나, PC2에서 기름값 또는 TR(GDP 대비 세금
비율)이 낮게 나오는 이유는 PC2가 이 변수들과는
크게 상관이 있지 않다는 의미이다.



다변량자료분석 프로젝트 기초보고서

주제(다변량기법)
주성분 분석을 통하여 각 변수들의 분포와 경향성을 파악.
데이터 제목
Global Country Information Dataset 2023
데이터내용 간략설명
세계에 다양한 경제,사회,교육,환경 등 여러 통계 지표들을 국가별로 정리해 놓은 데이터.
사이트링크
https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023
사이트내 로드맵
-토지면적과 인구 밀도 간 패턴 부석 -농지와 식량 안보 관계성 파악 -이산화탄소 배출량에 따른 기후변화 조사 -사회적 요인에 따른 경제 지표 사이의 상관성 등.