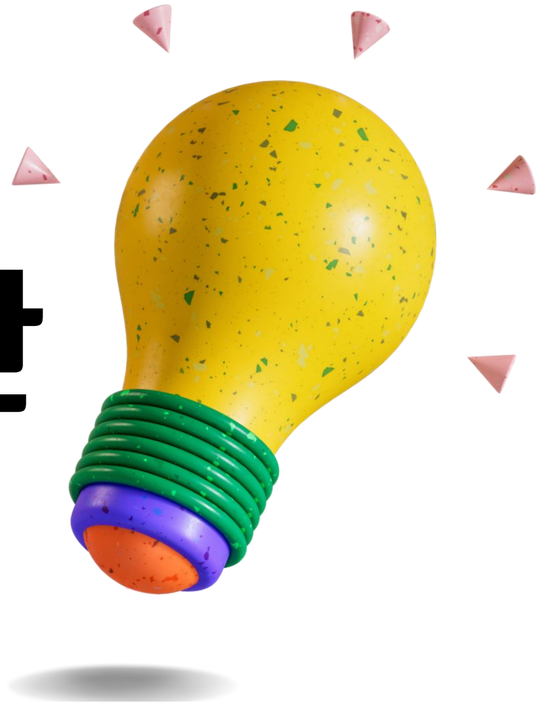
기대수명 데이터를 이용한 분류 학습 및 시각화



목차

1. 데이터 소개
   - 원자료
   - 데이터 범주화

2. 데이터 전처리 및 EDA
   - 분포
   - 이상치 및 상관관계
   - 전처리 후 데이터

3. 비지도 학습
   - PCA
   - k-means

4. 지도 학습
   - Lda
   - randomforest

데이터 살펴보기

데이터

🟡 data : Life_Expectancy_Data.csv 🟡

- 여러 국가를 기반으로 한 면역 요인, 사망률 요인, 경제적 요인, 사회적 요인 및 기타 건강 관련 요인으로 구성

- 총 22개의 변수로 구성, 1649개의 데이터

- 결측치 없음

데이터

변수명	정의	설명	유형
Country	국가	다양한 국가의 이름	범주형
Year	연도	데이터를 관찰한 연도 2000 ~ 2015	연속형
Status	국가의 상태	Developed : 선진국 Developing : 개발도상국	범주형
Life Expectancy	기대수명	기대 수명(나이)	연속형
Adult Mortality	성인 사망수	15세~60세 사이의 성인 1000명당 사망자 수	연속형
Infant Deaths	영아 사망수	유아 1000명당 사망자 수	연속형
Alcohol	알코올 소비량	1인당 알코올 소비량(리터)	연속형
Percentage Expenditure	지출 비율	GDP대비 보건 예산 지출 비율(%)	연속형
Hepatitis B	B형 간염	1세 아동의 B형 간염 예방 접종률(%)	연속형
Measles	홍역	인구 1000명 당 홍역 예방접종자 수	연속형
BMI	BMI	인구 평균 체질량 짖수	연속형
Under-Five Deaths	5세 미만 사망자 수	5세 이하 아동의 1000명당 사망자 수	연속형
Polio	소아마비	1세 아동의 소아마비 예방접종률(%)	연속형
Total Expenditure	총 지출	정부 총 예산 대비 보건 분야 예산(%)	연속형
Diphtheria	디프테리아	1세 아동의 디프테리아 예방접종률(%)	연속형
HIV/AIDS	HIV/AIDS	1000명당 HIV/AIDS으로 인한 사망률	연속형
GDP	국내총생산	1인당 GDP	연속형
Population	인구	국가 총 인구	연속형
Thinness 10-19 Years	저체중(10-19세)	10-19세 청소년의 저체중 비율	연속형
Thinness 5-9 Years	저체중(5-9세)	5-9세 어린이의 저체중 비율	연속형
Income Composition of Resources	ICOR	소득 분배 및 자원 접근성을 반영하는 종합 지수	연속형
Schooling	교육	평균 교육 기간	연속형

데이터 중 범주형 변수를 이용해 나머지 건강 변수들이 그룹 간 유의미한 차이가 존재하는지 살펴보고자 한다.

데이터 범주화 – data_new

변수명	정의	설명	유형
Country_g	국가	Africa, Asia, Europe, Middle_East, Oceania, South_America	범주형
Year_g	연도	y_0004, y_0509, y_1015	범주형
Status	국가의 상태	Developed : 선진국 Developing : 개발도상국	범주형
School	교육 기간	Elementary, high, middle, uni	범주형
Life Expectancy	기대수명	기대 수명(나이)	연속형
Adult Mortality	성인 사망수	15세~60세 사이의 성인 1000명당 사망자 수	연속형
⋮			
Population	인구	국가 총 인구	연속형
Thinness 1-19 Years	저체중(1-19세)	10-19세 청소년의 저체중 비율	연속형

범주형 변수 4개

연속형 변수 18개

최종 데이터

- **원데이터**

  - 범주형 변수 3개, 연속형 변수 19개
  - 총 22개 변수, 1649개 데이터

- **이상치 제거**

  - Measles(292) ,under5_death(15), GDP (116)

  - BMI 삭제

- **상관관계**

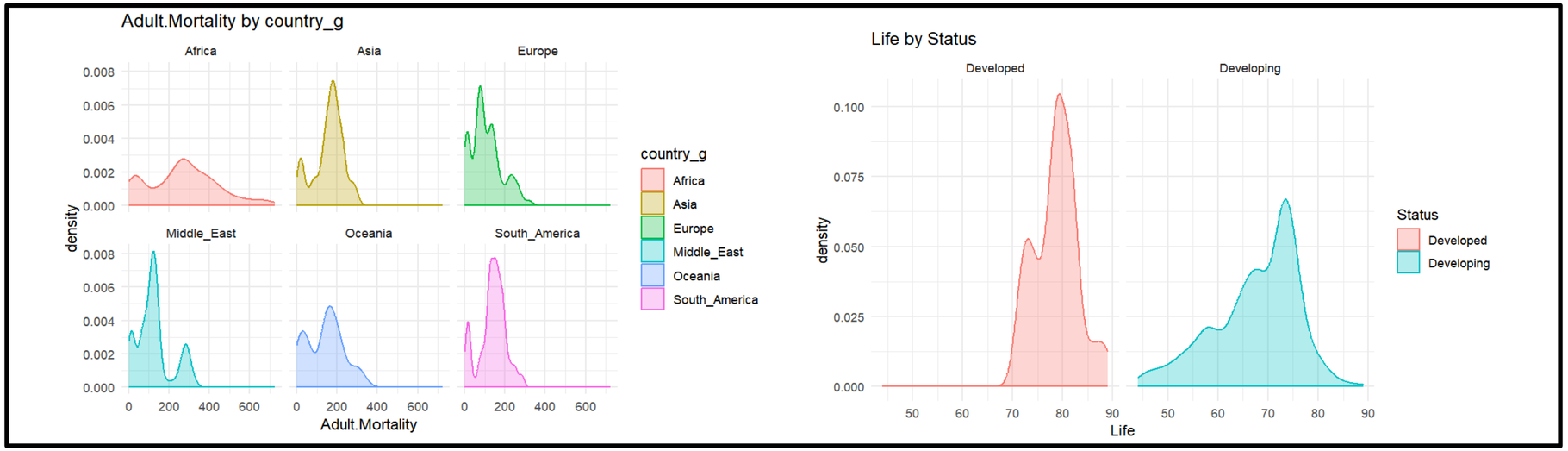
  - per_expenditure, infant.deaths 삭제

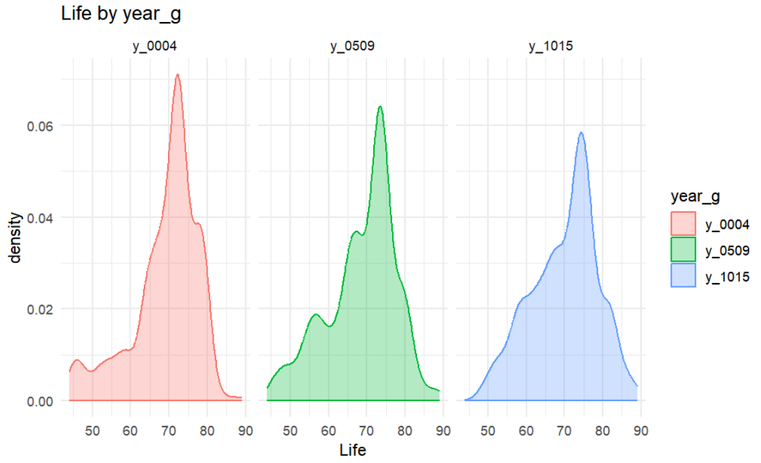
- **data_final**

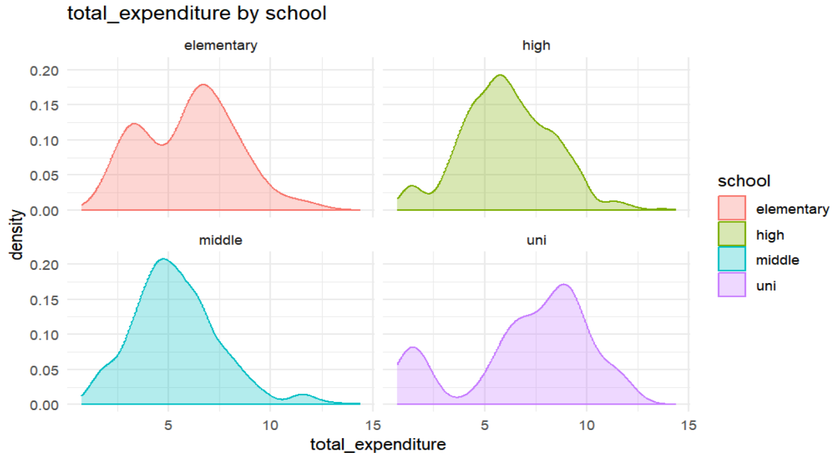
  - 범주형 변수 4개, 연속형 변수 15개
  - 총 19개 변수 , 1270개 데이터

데이터 분포

* country_g와 Status → 각 변수에 대해 레벨에 따른 분포 차이 존재

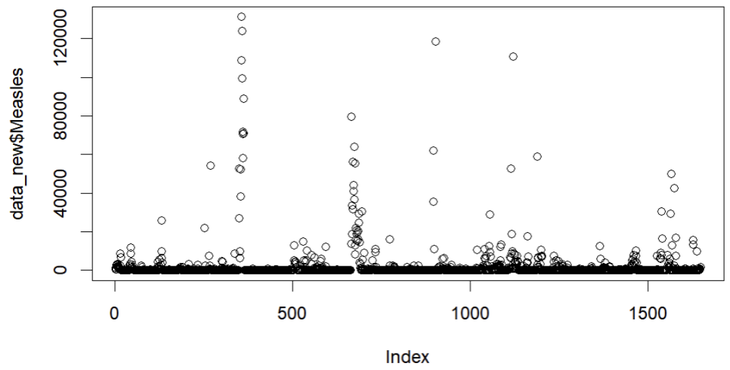


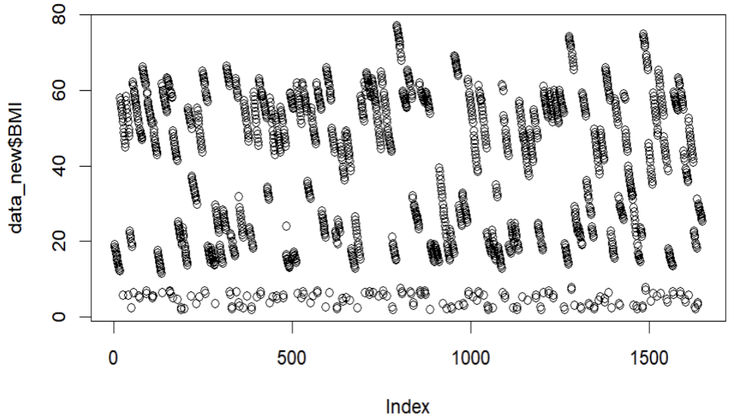


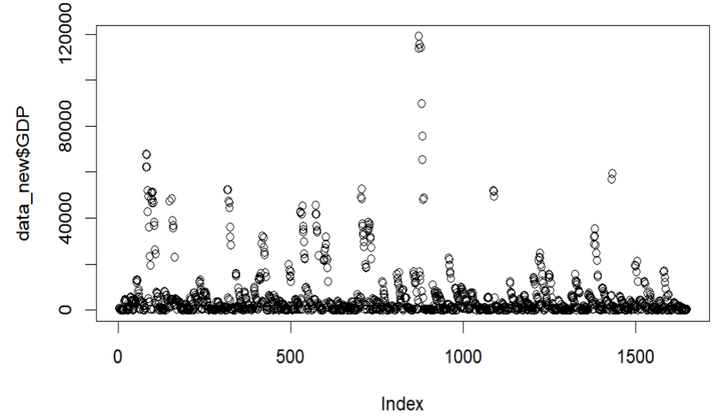


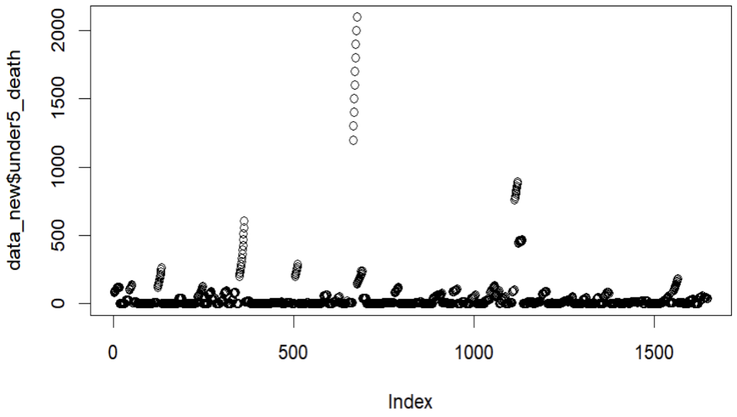
이상치 제거

* 변수 정의에 의한 이상치 판단









- Measles (292)

: 1000명 당 홍역 예방 접종자 수

1000 이상의 값을 이상치로 판단

- under5_death (15)

: 5세 이하 아동의 1000명당 사망자 수

1000 이상의 값을 이상치로 판단

- GDP (116)

: 국가 1인당 GDP(국내총생산)

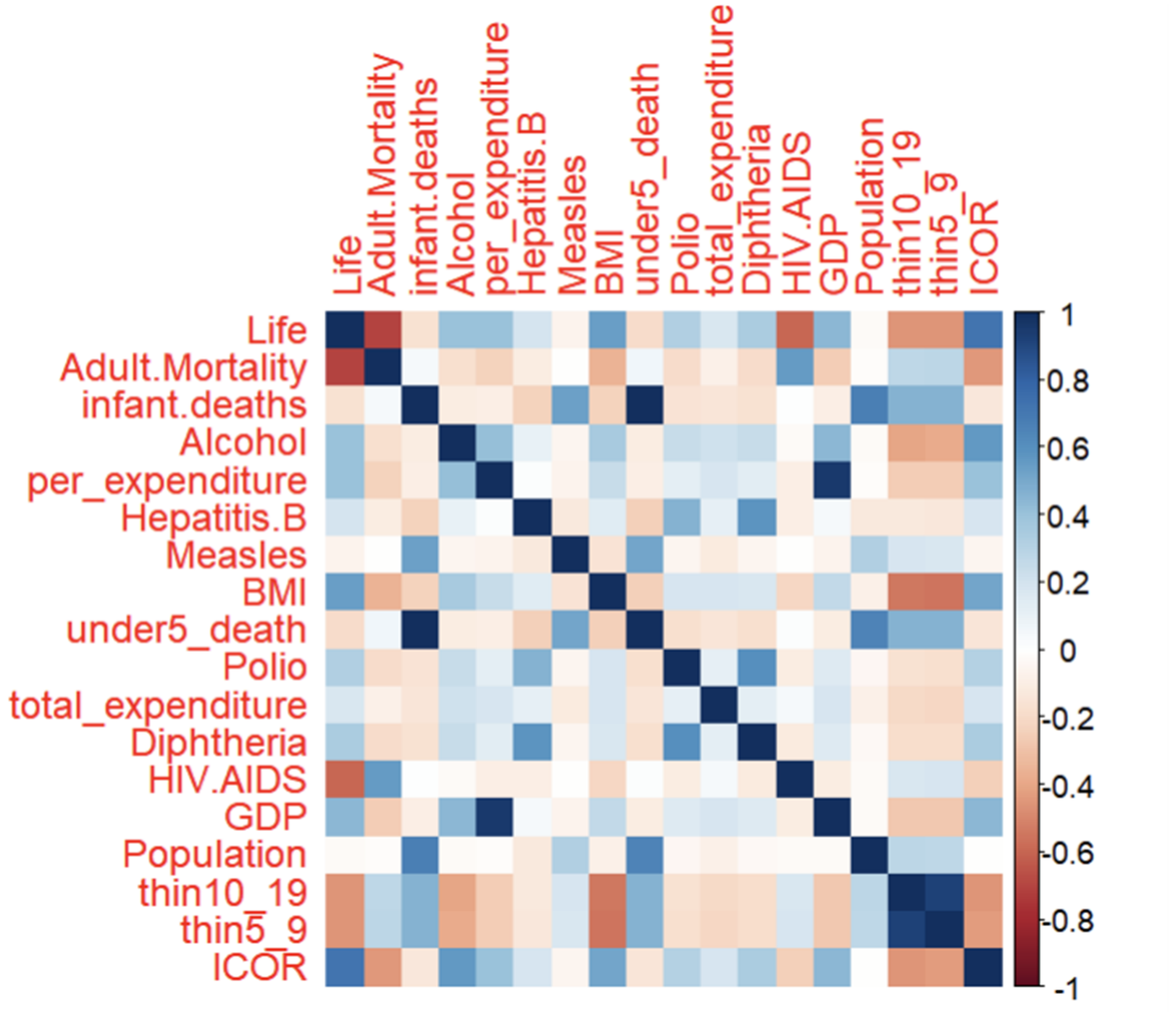
100 이하의 값을 이상치로 판단

- BMI

: 인구 평균 체질량 지수로, 50 이상인 경우 이상치로 판단.

단, 전체 데이터 중 이상치의 비율이 매우 높아 변수 제거

상관관계




```
[1] "under5_death"    "GDP"             "infant.deaths"   "per_expenditure"
```

- 양의 상관관계
: (infant.deaths,under5_death),
(GDP,per_expenditure),
(thin10_19,thin5_9)

- 음의 상관관계
: (Life,Adult.Mortality), (Life,HIV.AIDS)

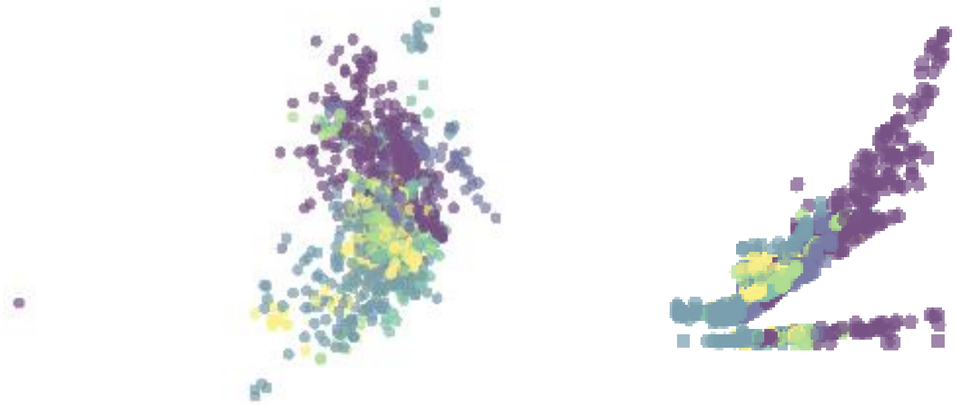
- 0.95 이상의 강한 상관관계
: (infant.deaths,under5_death), (GDP,per_expenditure)

→ per_expenditure, infant.deaths 제거

주성분 분석

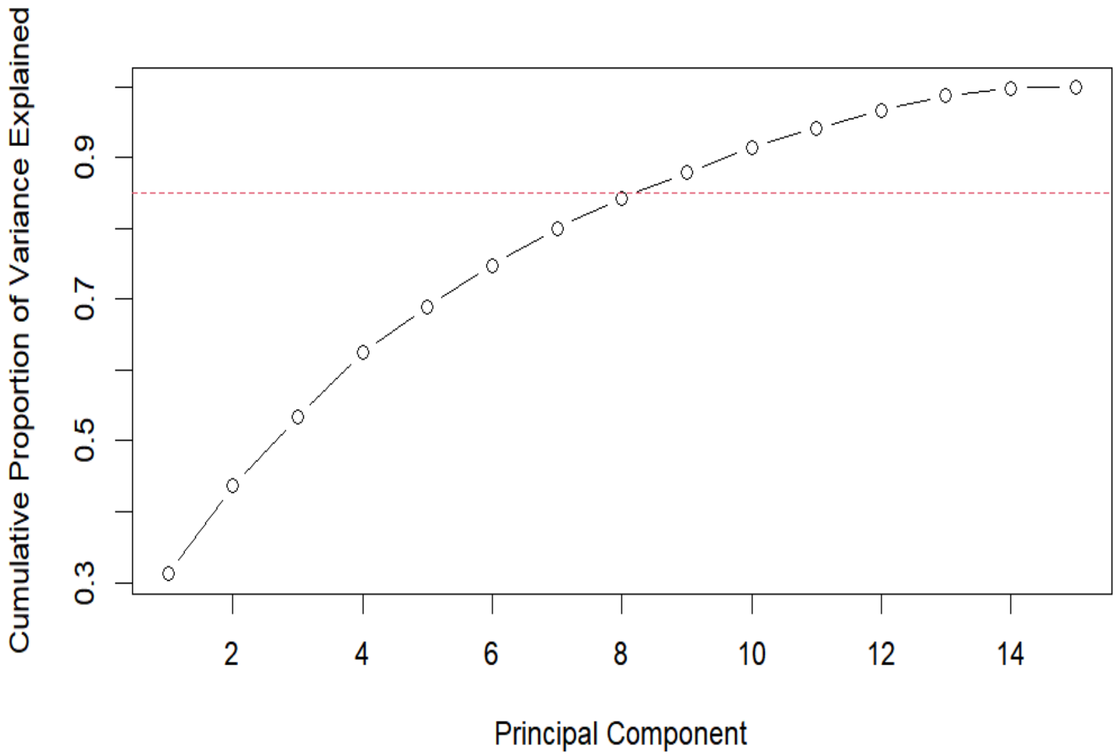
원자료

Country_g



status

주성분분석




```
Importance of components:
                           Comp.1    Comp.2     Comp.3
Standard deviation     2.1618776 1.3690937 1.20445463
Proportion of Variance 0.3118265 0.1250596 0.09679028
Cumulative Proportion  0.3118265 0.4368862 0.53367645
                           Comp.4     Comp.5     Comp.6
Standard deviation     1.16763614 0.98683582 0.93630825
Proportion of Variance 0.09096323 0.06497416 0.05849093
Cumulative Proportion  0.62463968 0.68961384 0.74810477
                           Comp.7     Comp.8     Comp.9
Standard deviation     0.88715531 0.79458951 0.73856811
Proportion of Variance 0.05251098 0.04212467 0.03639418
Cumulative Proportion  0.80061576 0.84274042 0.87913460
                          Comp.10    Comp.11    Comp.12
Standard deviation     0.72685688 0.64304691 0.60541402
Proportion of Variance 0.03524915 0.02758901 0.02445433
Cumulative Proportion  0.91438375 0.94197277 0.96642710
                          Comp.13    Comp.14     Comp.15
Standard deviation     0.57348661 0.38256296 0.167199568
Proportion of Variance 0.02194307 0.00976465 0.001865182
Cumulative Proportion  0.98837017 0.99813482 1.000000000
```

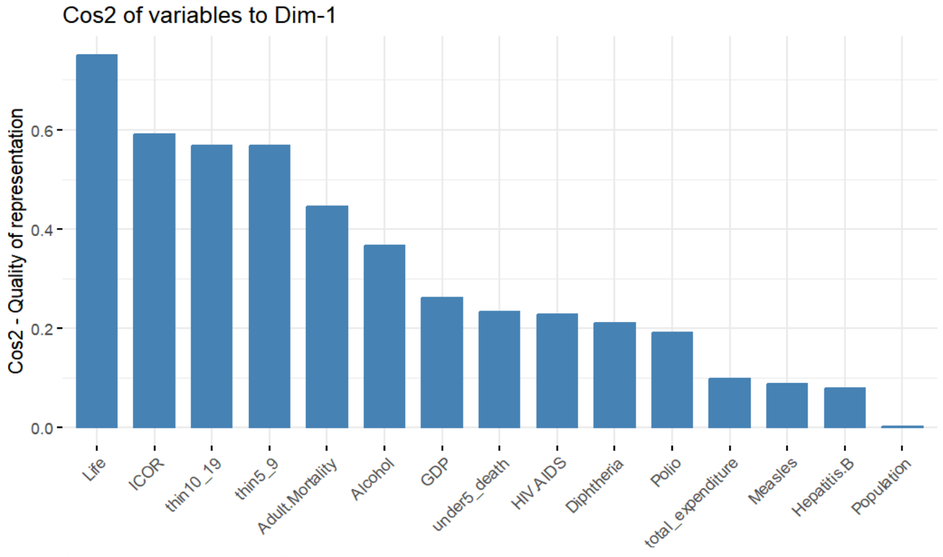
* 주성분 분석의 결과와 cumulative scree plot을 종합적으로 살펴봤을 때,
전체 데이터 분산의 약 85%를 설명할 수 있는 8개의 주성분을 사용

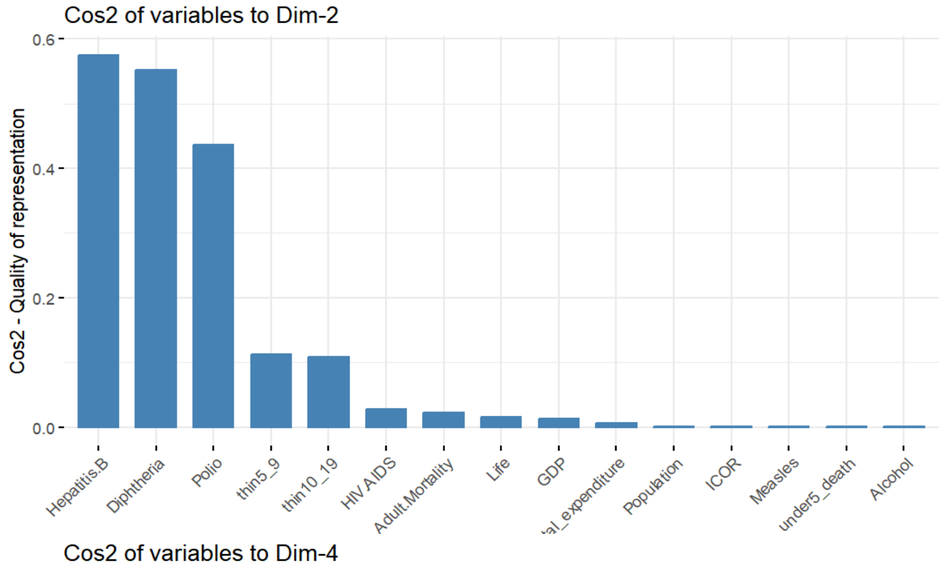
주성분분석

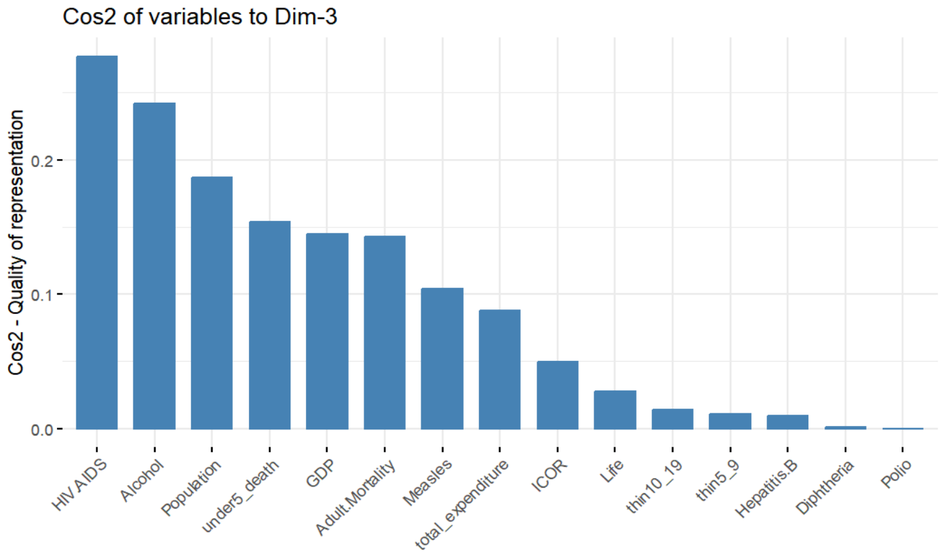
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Life	0.40066313	0.08989306	0.137509727	0.22422300	0.10446389	0.04084902	0.10419573	0.037476366
Adult.Mortality	-0.30850718	-0.10996242	-0.313882581	-0.31163244	-0.04423502	-0.15269768	-0.05748745	0.036608913
Alcohol	0.28013881	0.02019302	-0.408185388	-0.06777879	0.17320508	-0.23082068	0.12781177	-0.262630201
Hepatitis.B	0.12952056	0.55390286	0.080265784	0.01187229	-0.11871667	0.05089359	-0.14799788	-0.294397108
Measles	-0.13658545	0.02227158	-0.267563850	0.44287862	0.20243602	0.15405949	-0.70372006	0.322268951
under5_death	-0.22292443	0.02209065	-0.325352089	0.34170748	-0.23487083	0.12330395	-0.06048649	-0.711016975
Polio	0.20238432	0.48217671	-0.004237598	0.01105742	-0.02523550	-0.09225994	-0.06033891	0.161401435
total_expenditure	0.14466996	-0.05694961	-0.246061991	-0.19477310	-0.03336527	0.90171616	0.19302473	0.094305809
Diphtheria	0.21215367	0.54255210	-0.030678301	0.01251563	-0.04266236	-0.04260810	-0.11272465	0.045786021
HIV.AIDS	-0.22031058	-0.12196182	-0.436842775	0.42482412	0.09778889	-0.12054358	0.01208992	0.163878132
GDP	0.23642261	0.08474746	-0.315689369	0.10570145	0.52333885	-0.04199246	0.09158528	-0.027874195
Population	0.01496666	-0.02678657	-0.358698764	0.43283180	-0.53569748	-0.14834336	0.40560956	0.405887860
thin10_19	-0.34850327	-0.24071270	0.098221932	0.23604421	0.35273592	0.05434886	0.32670647	0.021148490
thin5_9	-0.34840548	-0.24523288	0.087404625	0.23414428	0.35084303	0.03655792	0.33119536	0.003997728
ICOR	0.35537707	0.02590021	-0.184709424	0.08497822	0.15934440	-0.11915834	0.07504585	-0.048239383

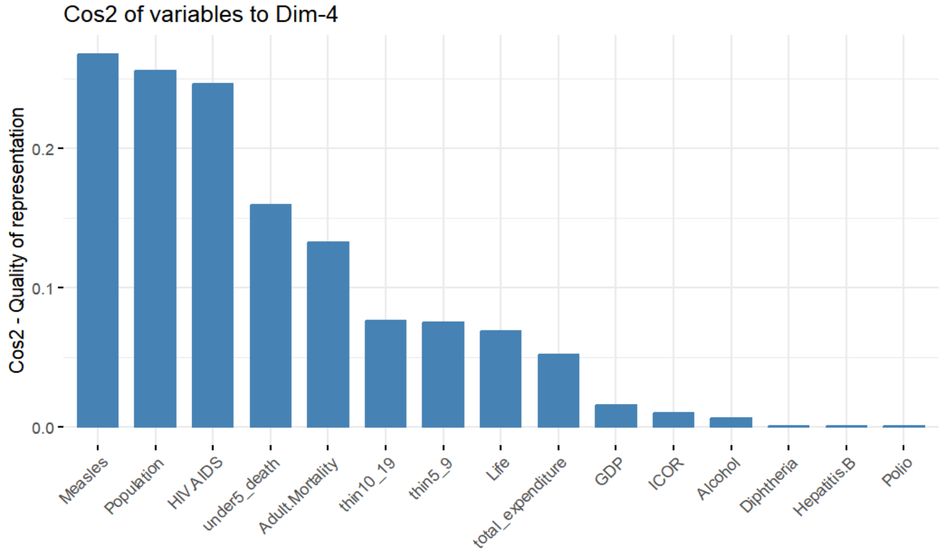
* comp.1 : Life, Thin10_19, Thin5_5 등과 관련
* comp.2 : Hepatitis B, Polio, Diphtheria 등과 관련 – 면역 요인
* comp.3 : HIV/AIDS. Alcohol, Under-Five Deaths, Adult Mortality 등과 관련 – 사망 요인

주성분분석

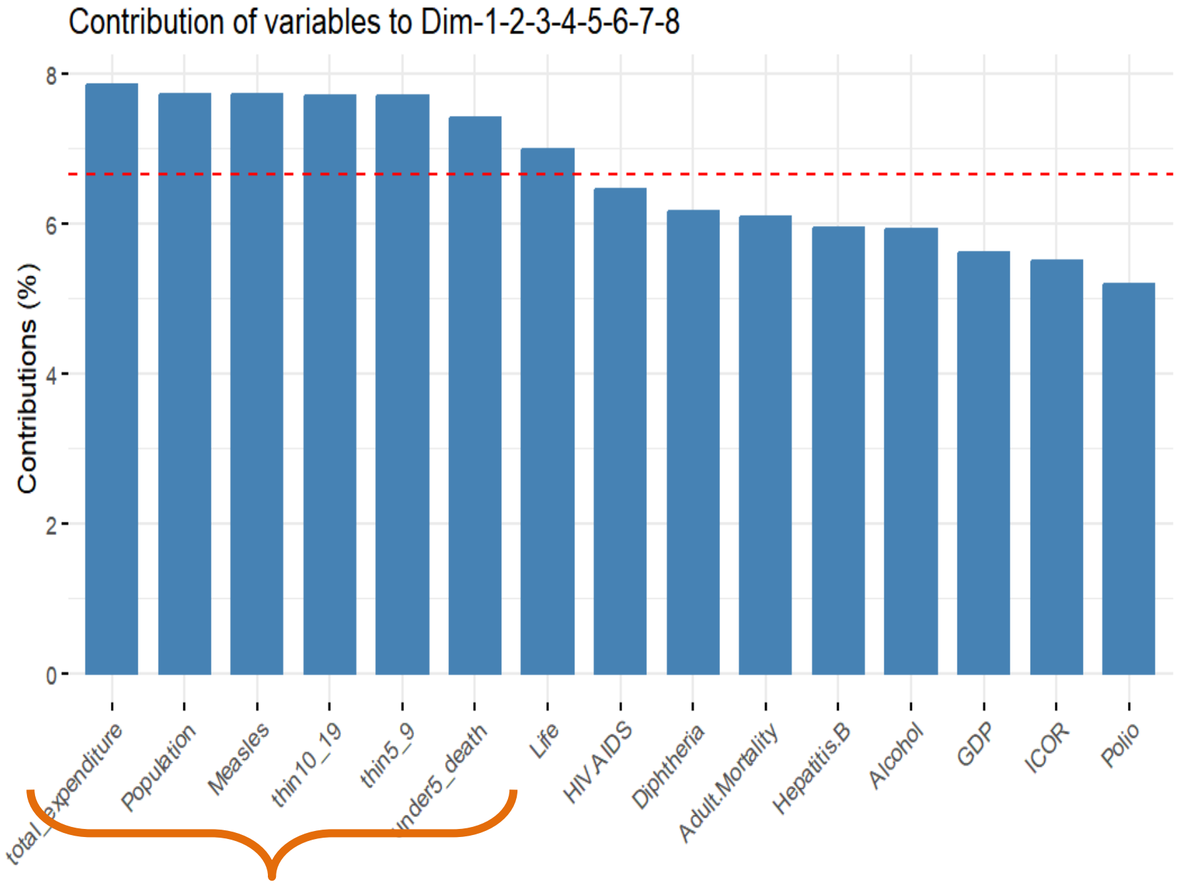

Cos2 of variables to Dim-1
Cos2 - Quality of representation
0.0
0.2
0.4
0.6
Life
ICOR
thin10_19
thin5_9
Adult.Mortality
Alcohol
GDP
under5_death
HIV.AIDS
Diphtheria
Polio
total_expenditure
Measles
Hepatitis.B
Population

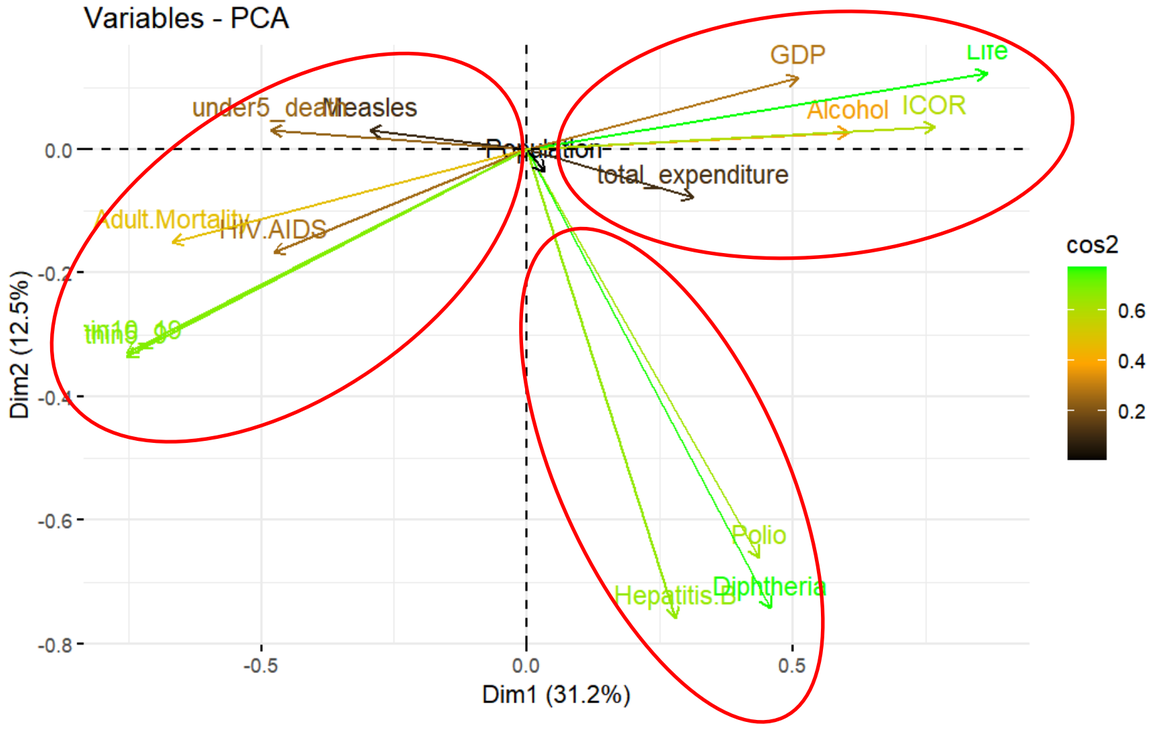

Cos2 of variables to Dim-2
Cos2 - Quality of representation
0.0
0.2
0.4
0.6
Hepatitis.B
Diphtheria
Polio
thin5_9
thin10_19
HIV.AIDS
Adult.Mortality
Life
GDP
Population
ICOR
Measles
under5_death
Alcohol


Cos2 of variables to Dim-3
Cos2 - Quality of representation
0.0
0.1
0.2
HIV.AIDS
Alcohol
Population
under5_death
GDP
Adult.Mortality
Measles
total_expenditure
ICOR
Life
thin10_19
thin5_9
Hepatitis.B
Diphtheria
Polio


Cos2 of variables to Dim-4
Cos2 - Quality of representation
0.0
0.1
0.2
Measles
Population
HIV.AIDS
under5_death
Adult.Mortality
thin10_19
thin5_9
Life
total_expenditure
GDP
ICOR
Alcohol
Diphtheria
Hepatitis.B
Polio

주성분분석

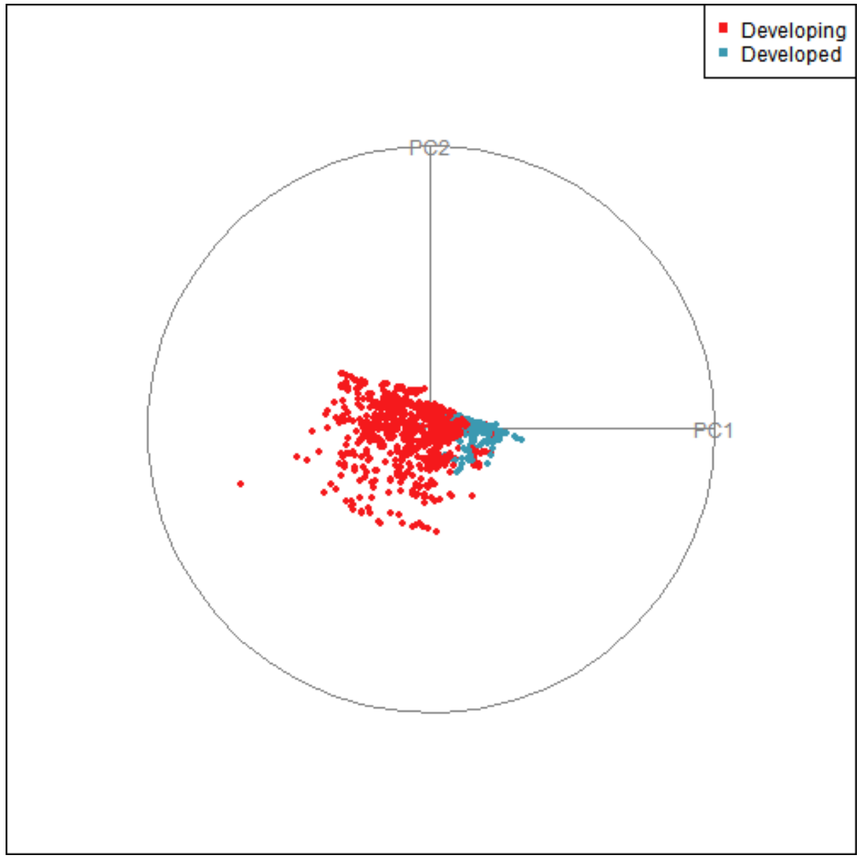


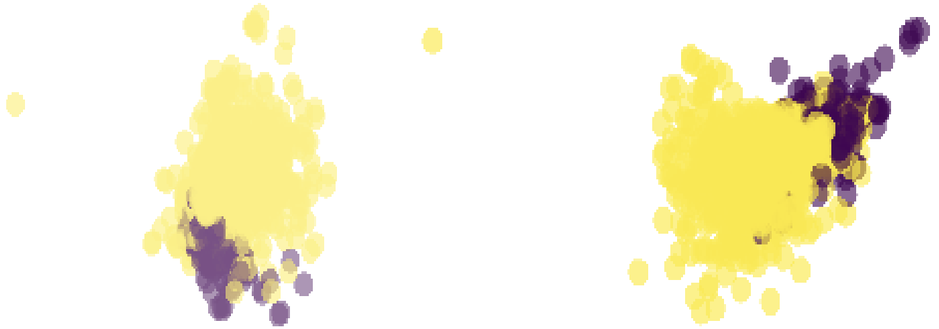


주성분 분석에 의해 발생한 주성분을 전체적으로 살펴봤을 때,

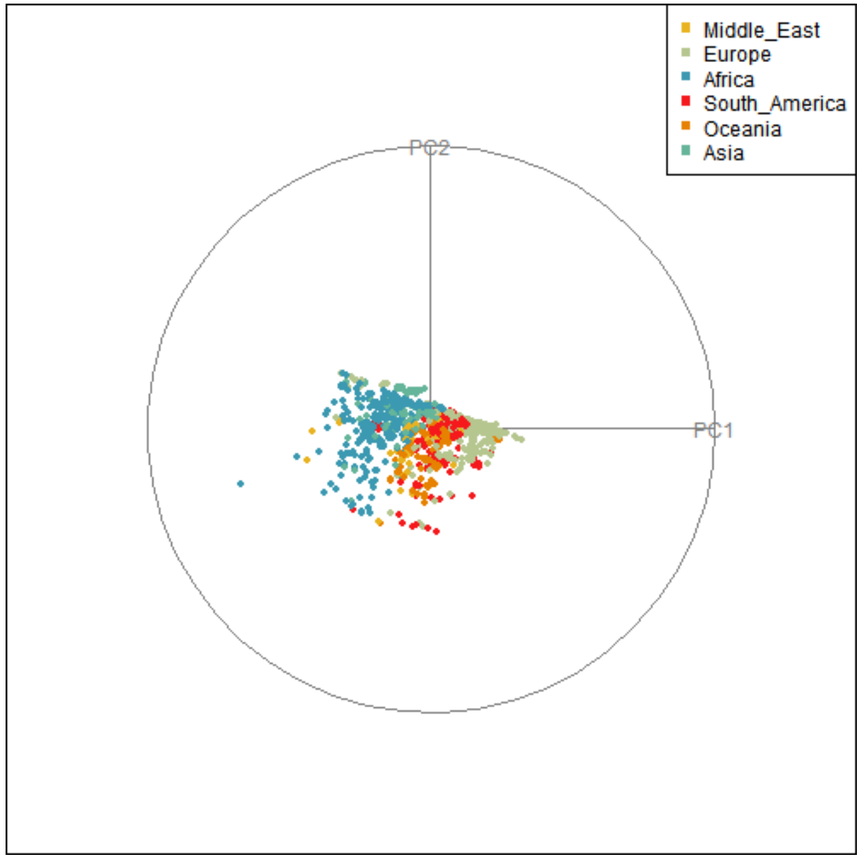
Total Expenditure, Population, Measles, Thin10_19, Thin5_9, Life 가 유의미하게 기여도가 높음

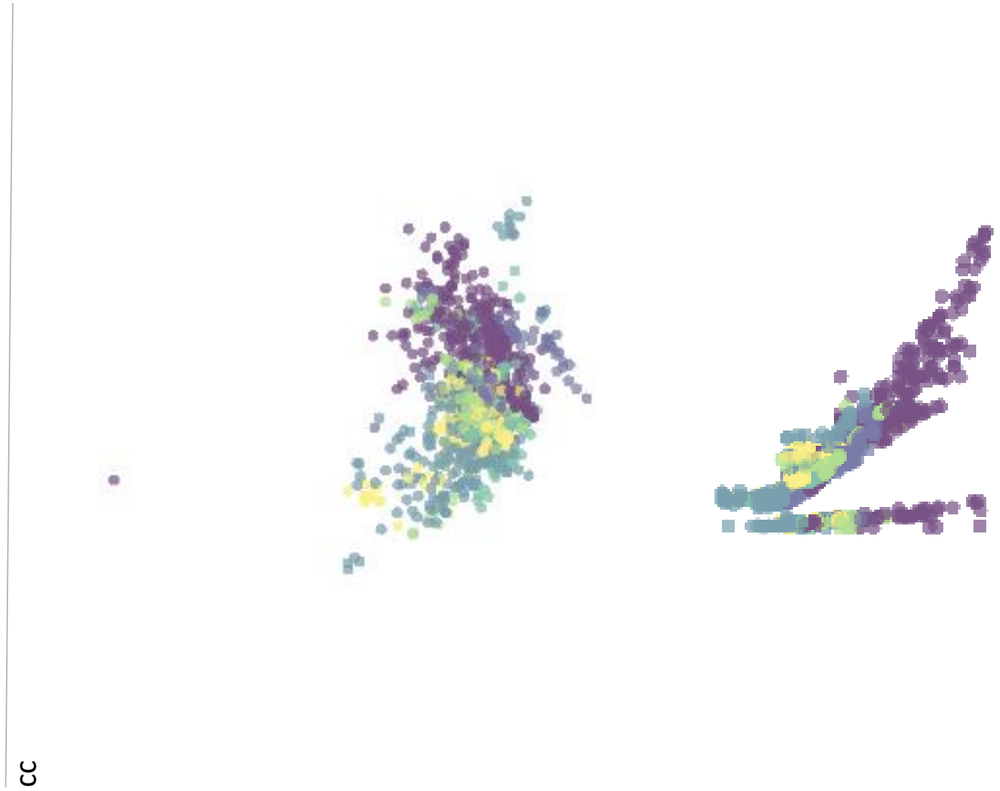
결과 비교 : Status





결과 비교 : country_g

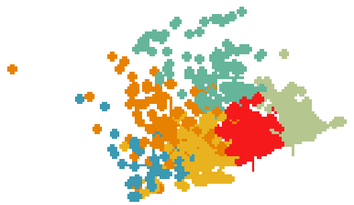


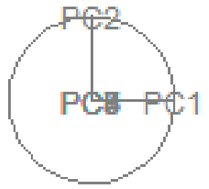


- 원자료의 모든 변수 22개를 이용했을 때 보다 주성분 분석을 통한 8개의 주성분을 이용했을 때 클러스터링 결과가 비교적 뚜렷함
  → 데이터의 정보 손실을 최소화 하며 고차원 데이터를 저차원 데이터로 차원축소 변환하여 시각화 용이

K-means

K-means : country_g





1	2	3	4	5	6
40	128	219	213	124	546

* 앞서 진행한 PCA 결과를 이용해, 주성분 8개로 구성된 데이터로 k-means 진행

* K-emans 실행 시 클러스터 개수는 6개로 지정

* 이상치 존재 → India(Asia) → 5번 클러스터

Population <dbl>	Country <chr>	Status <chr>
1290000000	India	Developing
1180000000	India	Developing
1160000000	India	Developing
1140000000	India	Developing
1130000000	India	Developing
255000000	Indonesia	Developing
249000000	Indonesia	Developing
243000000	Indonesia	Developing
236000000	Indonesia	Developing
233000000	Indonesia	Developing

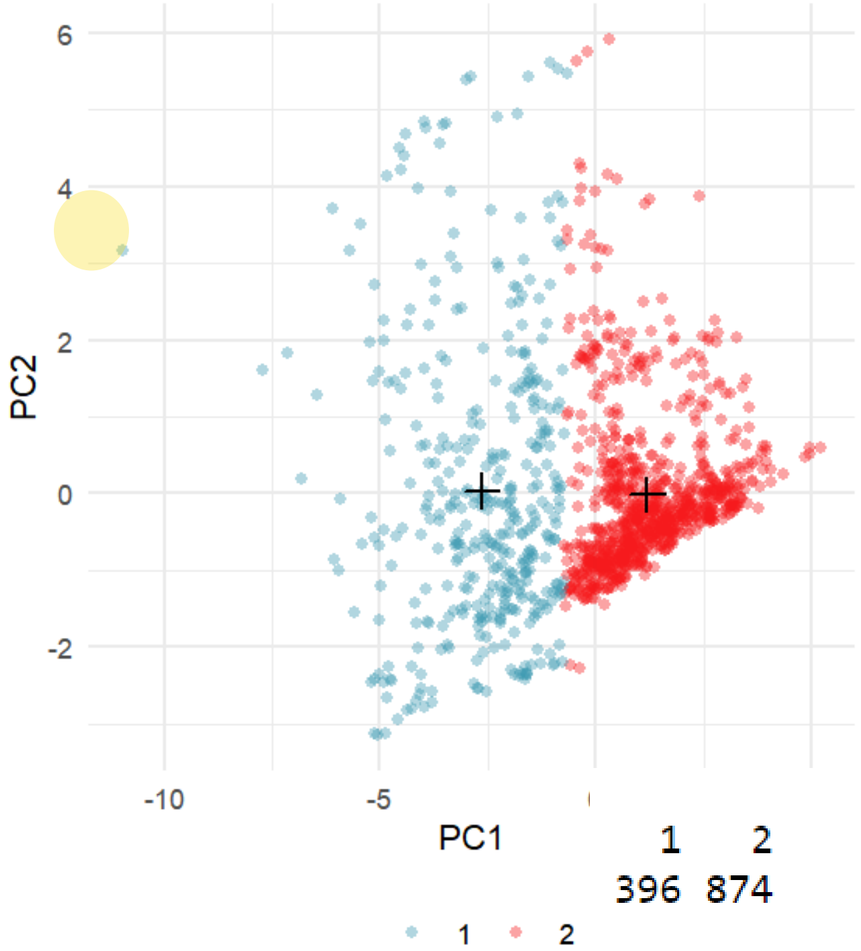
K-means : country_g

country_g <fctr>	Life <dbl>	Adult.Mortality <dbl>	Alcohol <dbl>	Hepatitis.B <dbl>	Measles <dbl>	under5_death <dbl>	Polio <dbl>
Africa	60.56981	263.3896	2.646169	76.68506	142.48052	33.961039	78.70779
Asia	68.69565	168.8783	2.054870	90.41739	150.54783	19.686957	91.03478
Europe	76.34420	105.6981	8.779596	83.61725	69.61995	1.469003	91.91644
Middle_East	74.02571	104.8714	1.287286	82.44286	211.27143	15.257143	80.11429
Oceania	70.82018	132.6789	2.344679	74.50459	23.87156	1.174312	79.33945
South_America	73.46094	133.8990	5.132997	81.15825	24.23906	11.760943	85.19529

k_cl <fctr>	Life <dbl>	Adult.Mortality <dbl>	Alcohol <dbl>	Hepatitis.B <dbl>	Measles <dbl>	under5_death <dbl>	Polio <dbl>
1	49.38500	522.90000	4.731750	76.40000	115.95000	26.750000	82.80000
2	69.97969	157.07031	2.914687	41.52344	36.39062	9.140625	50.42969
3	80.06895	72.94977	10.497032	81.16438	94.73516	1.141553	94.47945
4	66.16854	194.11737	1.985164	87.42723	64.30986	12.666667	89.86854
5	59.06210	253.58065	1.790403	69.31452	426.68548	69.016129	65.36290
6	72.73407	131.02930	4.895037	90.97436	27.41026	7.714286	92.79487

- Aisa로 예상되는 군집의 각 변수 평균 값을 비교해 본 결과
  - 통계량 값이 유사하지 않음
  - k-means를 이용하여 클러스터링을 한 결과, 군집이 잘 나눠졌지만, 원데이터와의 비교하기 어렵다.

K-means : Status



* 앞서 진행한 PCA 결과를 이용해, 주성분 8개로 구성된 데이터로 k-means 진행

* K-emans 실행 시 클러스터 개수는 2개로 지정

* 이상치 존재 → India (Developing)→ 1번 클러스터

K-means : Status

Status <fctr>	Life <dbl>	Adult.Mortality <dbl>	Alcohol <dbl>	Hepatitis.B <dbl>	Measles <dbl>	under5_death <dbl>	Polio <dbl>
Developed	78.73836	82.90868	10.546758	87.43379	76.85388	0.9315068	94.42466
Developing	68.84358	174.28449	3.681256	79.81637	90.18554	16.8924833	83.43673

k_cl <fctr>	Life <dbl>	Adult.Mortality <dbl>	Alcohol <dbl>	Hepatitis.B <dbl>	Measles <dbl>	under5_death <dbl>	Polio <dbl>	total_expenditure <dbl>
1	61.59066	254.6237	2.191010	70.72980	173.18434	33.093434	73.77020	5.45053
2	74.60915	114.9874	6.076773	85.84211	49.23913	5.552632	90.56979	6.43286

- Developing 으로 예상되는 군집의 각 변수 평균 값을 비교해 본 결과
  - 통계량 값이 유사하지 않음
  - k-means를 이용하여 클러스터링을 한 결과, 군집이 잘 나눠졌지만, 원데이터와의 비교하기 어렵다.

LDA

LDA : 데이터셋 분할

- lda 진행을 위해, 기존 data_final에서 범주형 변수를 제거하고, 스케일을 진행한 data_lda_s생성
- 지도 학습을 위해 train data(7)와 test data(3) 분할 생성

- train data : 889개

Africa	Asia	Europe	Middle_East	Oceania	South_America
217	85	250	49	81	207

- test data : 381개

Africa	Asia	Europe	Middle_East	Oceania	South_America
91	30	121	21	28	90

LDA : country_g

```
Coefficients of linear discriminants:
                                LD1         LD2         LD3         LD4          LD5
Life                        0.66363745  0.11092100 -1.08709528 -0.03333450  0.4885071353
Adult.Mortality            -0.06938632  0.11243123 -0.06207634  0.03605237  0.3343136600
Alcohol                     0.78137459  1.07830980  0.55492180 -0.14198530 -0.2444321160
Hepatitis.B                -0.14899431 -0.11389347 -0.24302136  0.12291429 -0.1059340768
Measles                    -0.10657479  0.04401924 -0.41531916 -0.11490825 -0.648513141.0
under5_death               -0.16830298  0.07934980  0.06139814 -0.45050834  0.737337783
Polio                      -0.04398315  0.06646687 -0.06414036  0.14117889 -0.2669777310
total_expenditure           0.08782725  0.00113375 -0.24138303 -0.67448155  0.0009371451
Diphtheria                 -0.04189410  0.40503667  0.33794720  0.21998205  0.4490421973
HIV.AIDS                   -0.02276926 -0.15922303  0.01397435 -0.51723286 -0.0577191967
GDP                        -0.15525406 -0.13310303  0.21224396  0.13662121 -0.590058798 2
Population                  0.01426769 -0.01198278 -0.07808703 -0.06182466  0.3353263969
thin10_19                   0.05479890  0.02745610  0.31057770 -1.84829029  0.0646652390
thin5_9                    -0.77532848  0.95014026 -0.66713689  1.76862146 -0.1743824324
ICOR                       -0.27778863  0.15589999 -0.44757193 -0.23920713  0.2246127176

Proportion of trace:
   LD1    LD2    LD3    LD4    LD5
0.4527 0.2496 0.1882 0.0826 0.0268
```

* Train data를 이용해 학습한 모델에 test data를 이용해 예측한 결과

* 판별함수 5개 생성 – LD1 : LD5

* 각 판별함수에 큰 영향을 주는 변수 – Life, Alcohol, Measles, Under-Five Deaths, Total_expenditure, HIV/AIDS, GDP, Thin5_9

LDA : country_g

Confusion Matrix and Statistics

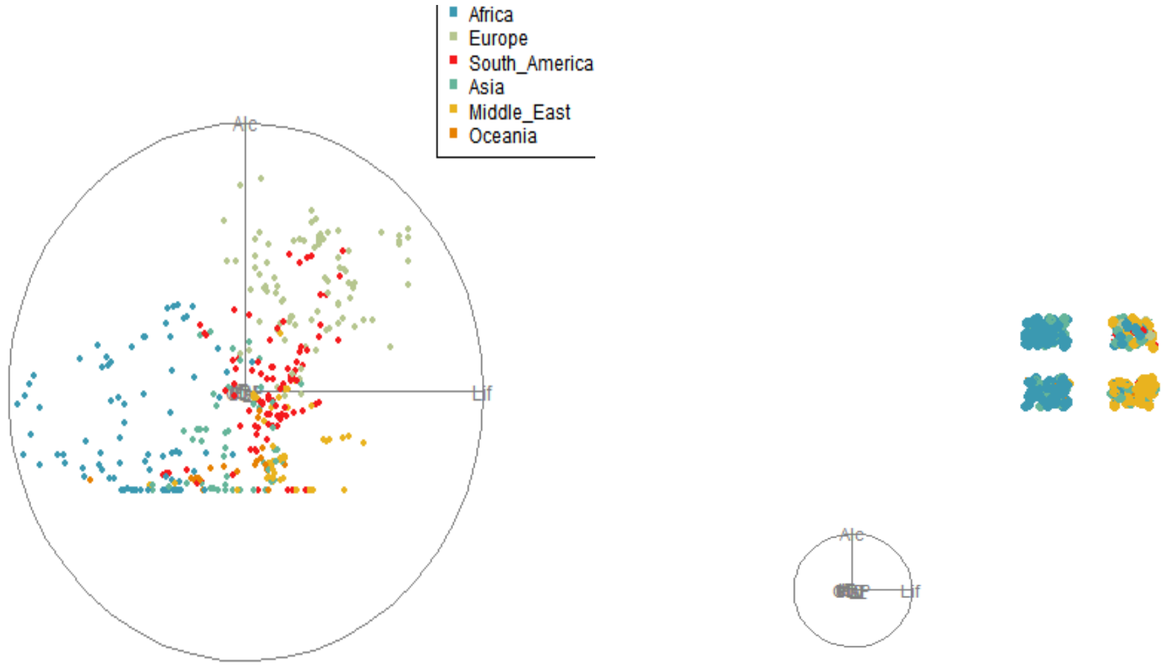
	test.pc					
	Africa	Asia	Europe	Middle_East	Oceania	South_America
Africa	66	19	8	8	4	2
Asia	0	18	0	3	7	7
Europe	0	7	74	4	2	13
Middle_East	2	0	0	18	2	3
Oceania	0	3	7	0	22	0
South_America	5	2	9	2	8	56

Overall Statistics

Accuracy : 0.6667
95% CI : (0.6169, 0.7139)
No Information Rate : 0.2572
P-Value [Acc > NIR] : < 0.00000000000000022

Kappa : 0.586

Mcnemar's Test P-Value : 0.00000001025



* LDA결과를 확인해 본 결과, 분류가 명확하지 않음 → LDA의 정확도가 약 67%%으로, 분류 정확도가 떨어짐
* 데이터를 6개의 대륙으로 분류할 때, 18개의 변수 중 Life, Alcohol, Measles, Under-Five Deaths, Total_expenditure 등에 의해 영향
  → 면역 요인, 사망 요인, 경제 요인 등 다양한 요인에 의해 분류

LDA : 데이터셋 분할

* lda 진행을 위해, 기존 data_final에서 Status를 제외한 범주형 변수를 제거하고, 스케일을 진행한 data_lda_s생성
* 지도 학습을 위해 train data(7)와 test data(3) 분할 생성

- train data : 889개

```
Developed Developing
      142         747
```

- test data : 381개

```
Developed Developing
       77         304
```

LDA : Status

```
Coefficients of linear discriminants:
                        LD1
Life              -0.371446228
Adult.Mortality    0.098060431
Alcohol           -1.000452367
Hepatitis.B       -0.114732328
Measles           -0.187044504
under5_death       0.024762911
Polio             -0.005123114
total_expenditure -0.020853887
Diphtheria         0.056541771
HIV.AIDS          -0.081653804
GDP               -0.360047079
Population         0.153360339
thin10_19         -0.063246462
thin5_9            0.064000051
ICOR               0.063832779
```

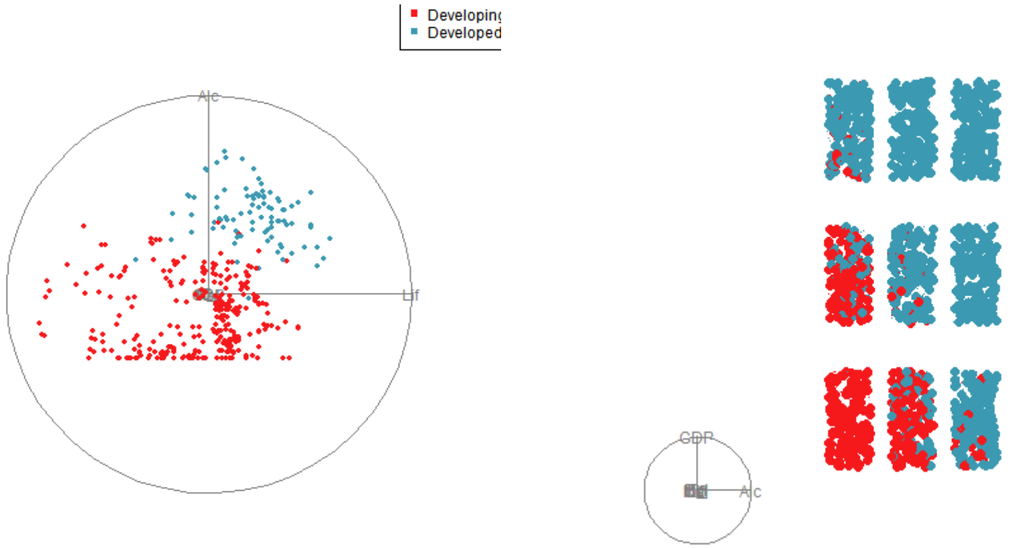
* Train data를 이용해 학습한 모델에 test data를 이용해 예측한 결과

* 판별함수 1개 생성 – LD1

판별함수에 큰 영향을 주는 변수 → Alcohol, Life, GDP , Measles ,Population

LDA : Status

```
Confusion Matrix and Statistics

            test.pc2
             Developed Developing
  Developed         74          3
  Developing        25        279

               Accuracy : 0.9265
                 95% CI : (0.8955, 0.9506)
    No Information Rate : 0.7402
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.7941
```



* accuracy= 92% 로, 판별분석에 의해 데이터가 비교적 잘 분류
→Status의 범주가 2개라 country_g보다 비교적 잘 된 것으로 판단

* 데이터를 2개의 Status로 분류할 때 Alcohol, Life, GDP , Measles ,Population 변수에 의해 영향
→ 건강, 사회, 면역 요인에 의해 Status가 분류

Random forest

LDA : 데이터셋 분할

* rondom forest진행을 위해, 기존 data_final에서 범주형 변수를 제거, 스케일을 진행한 data_rf생성

* 지도 학습을 위해 train data(7)와 test data(3) 분할 생성

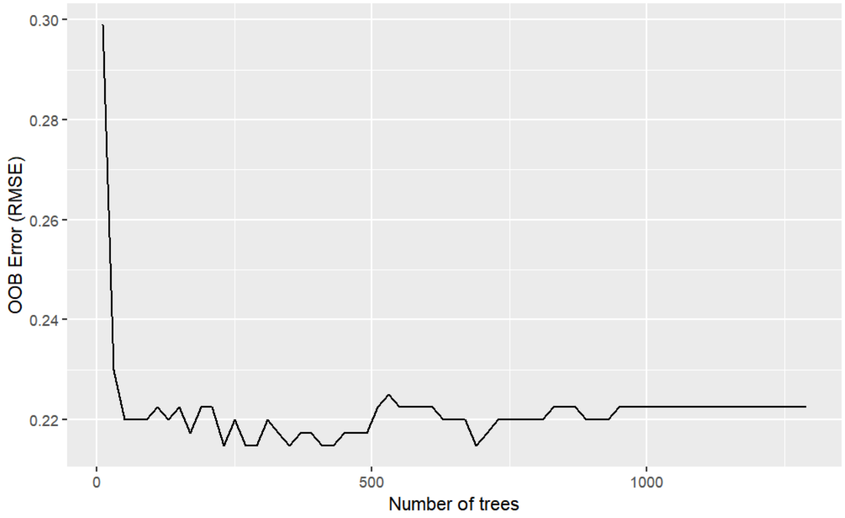
- train data : 889개

Africa	Asia	Europe	Middle_East	Oceania	South_America
217	85	250	49	81	207

- test data : 381개

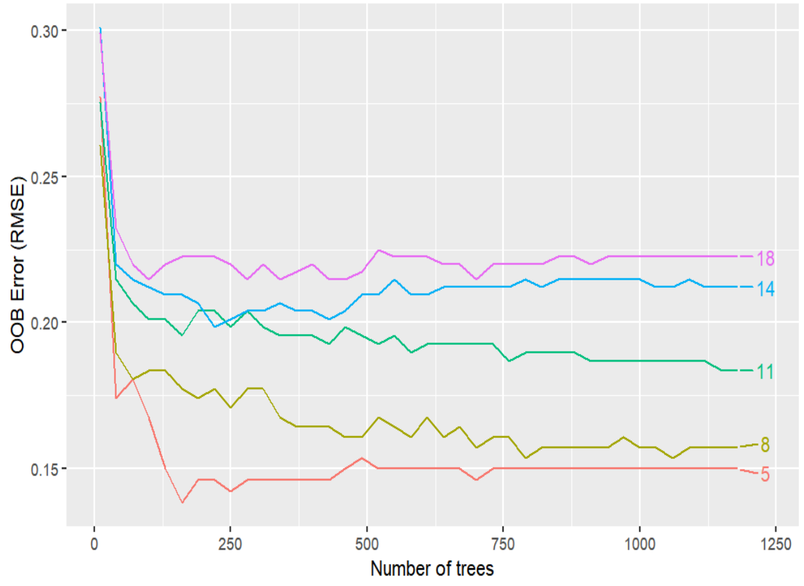
Africa	Asia	Europe	Middle_East	Oceania	South_America
91	30	121	21	28	90

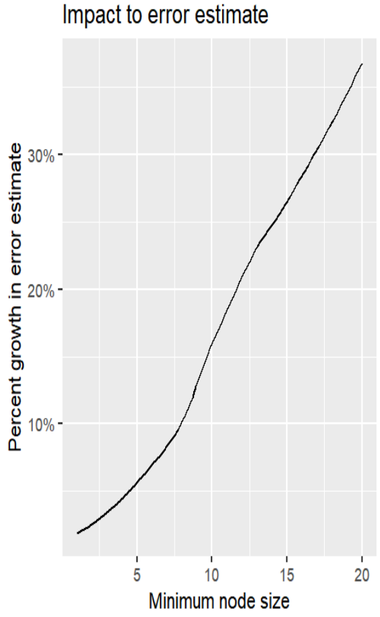
파라미터 탐색 (country_g): number of trees

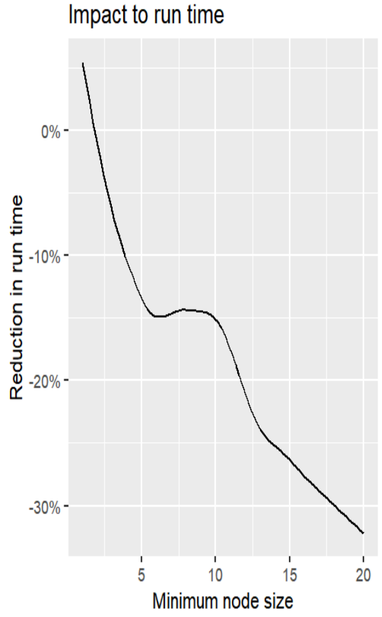


trees <dbl>	rmse <dbl>
230	0.2147539
270	0.2147539
290	0.2147539
350	0.2147539
410	0.2147539
430	0.2147539
690	0.2147539
170	0.2173571
330	0.2173571
370	0.2173571

* n tree = 230 ~ 690
* mtry =5
* Nodeseze = 5
* Sample fraction = 0.8







파라미터 결정 (country_g)

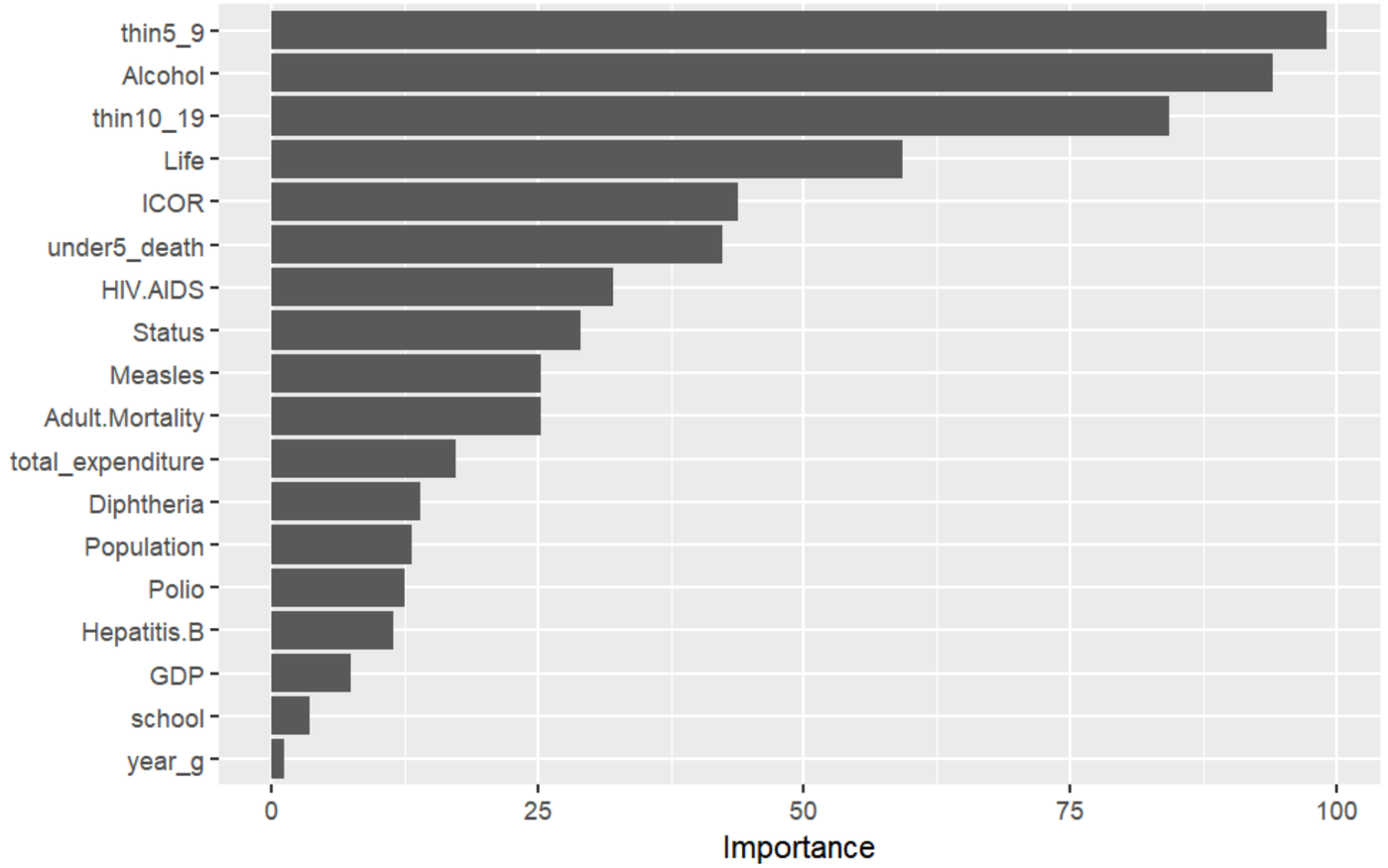
before & after

```
Confusion Matrix and Statistics


pred            Africa Asia Europe Middle_East Oceania South_America
  Africa            90    0      0           0       0             2
  Asia               0   33      1           0       0             0
  Europe             0    1    111           0       0             0
  Middle_East        0    0      0          21       0             0
  Oceania            0    0      0           0      39             0
  South_America      0    0      2           0       0            81

Overall Statistics

               Accuracy : 0.9843
                 95% CI : (0.966, 0.9942)
    No Information Rate : 0.2992
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.98

 Mcnemar's Test P-Value : NA
```

```
Confusion Matrix and Statistics

                rf_pred_im
                 Africa Asia Europe Middle_East Oceania South_America
  Africa             90    0      0           0       0             0
  Asia                0   34      0           0       0             0
  Europe              0    0    112           0       0             2
  Middle_East         0    0      0          21       0             0
  Oceania             0    0      0           0      39             0
  South_America       1    0      0           0       0            82

Overall Statistics

               Accuracy : 0.9921
                 95% CI : (0.9772, 0.9984)
    No Information Rate : 0.294
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.99
```

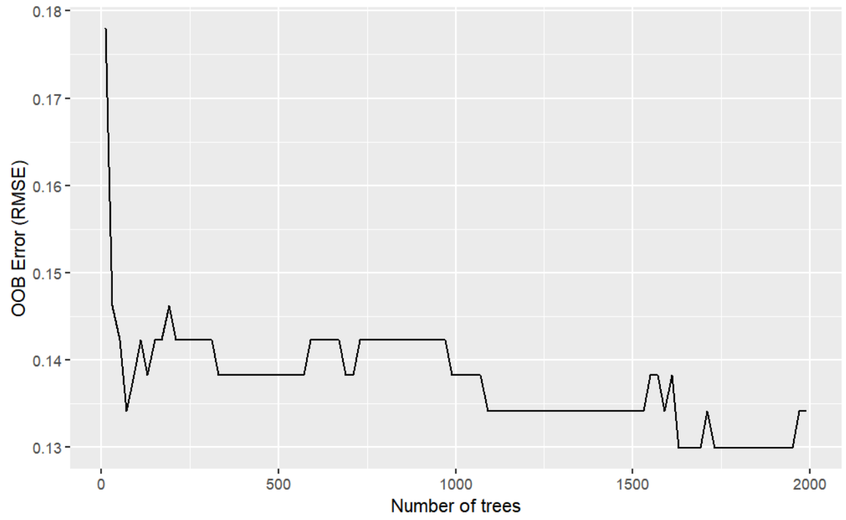
- ntree = 690, mtry=5, node.size=5, sample fraction=0.8을 기준으로 값을 조금씩 변경하여 최적의 파라미터 결정
  → ntree = 690, mtry=5, node.size=3, sample fraction=0.9 , replace = FALSE, importance = "impurity"
- 파라미터를 조정하지 않았을 때와 비교하여, accuracy가 미세하게 조정

Importance (country_g) - impurity



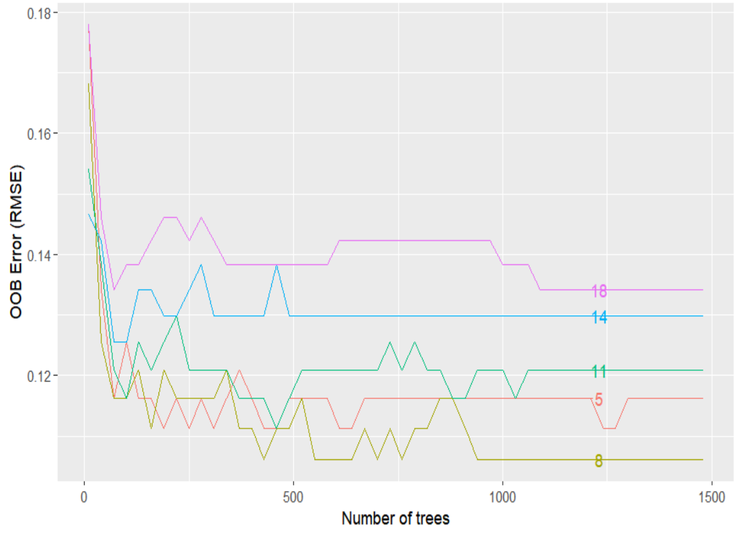
Variable <chr>	Importance <dbl>
thin5_9	99.137072
Alcohol	93.981948
thin10_19	84.275387
Life	59.311441
ICOR	43.854190
under5_death	42.328607
HIV.AIDS	32.140381
Status	28.986448
Measles	25.367704
Adult.Mortality	25.323017

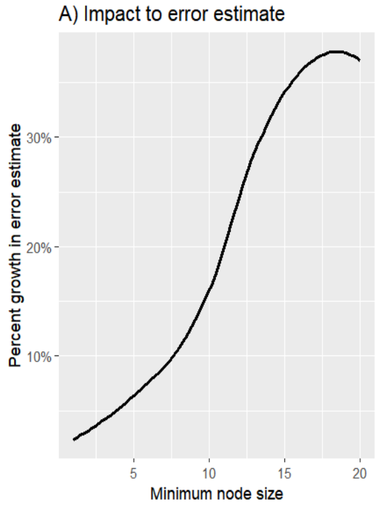
파라미터 탐색 (Status): number of trees

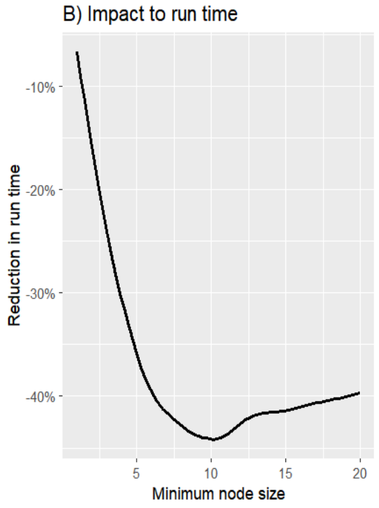


trees <dbl>	rmse <dbl>
710	0.1382845
990	0.1382845
1010	0.1382845
1030	0.1382845
1050	0.1382845
1070	0.1382845
1550	0.1382845
1570	0.1382845
1610	0.1382845
50	0.1422936

* n tree = 1610
* mtry = 8
* Nodesize = 5
* Sample fraction = 0.8







파라미터 탐색 (Status)

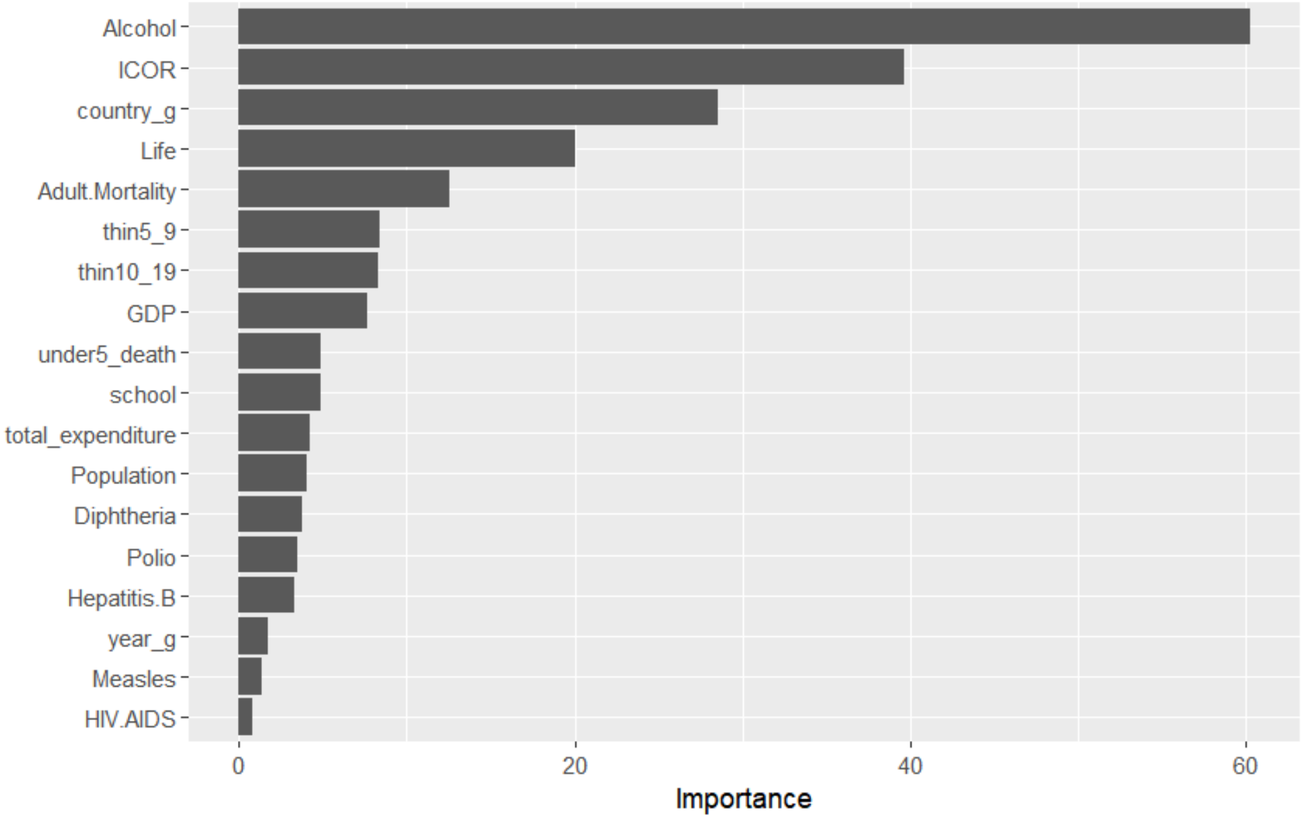
before & after

```
Confusion Matrix and Statistics

          rf_pred_none
           Developed Developing
Developed          66          3
Developing          2        310

               Accuracy : 0.9869
                 95% CI : (0.9696, 0.9957)
    No Information Rate : 0.8215
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.9555
```

```
Confusion Matrix and Statistics

          rf_pred_im
           Developed Developing
Developed          68          1
Developing          1        311

               Accuracy : 0.9948
                 95% CI : (0.9812, 0.9994)
    No Information Rate : 0.8189
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.9823
```

- ntree = 710, mtry=8 min.node.size =5, sample fraction=0.8을 기준으로 값을 조금씩 변경하여 최적의 파라미터 결정
  → ntrees = 900, mtry = 5, min.node.size = 5, sample.fraction = .90, replace = FALSE, importance = "impurity"
- 파라미터를 조정하지 않았을 때와 비교하여, accuracy가 미세하게 조정

Importance (Status) - impurity



Variable <chr>	Importance <dbl>
Alcohol	60.2345648
ICOR	39.6713886
country_g	28.5721123
Life	20.0198584
Adult.Mortality	12.5774338
thin5_9	8.3771058
thin10_19	8.2735085
GDP	7.6036349
under5_death	4.8764940
school	4.8445526

성능비교

LDA		Random forest	
Country_g(6)	Status(2)	Country_g(6)	Status(2)
0.6667	0.9265	0.9921	0.9948

성능비교

LDA		Random forest	
Country_g(6)	Status(2)	Country_g(6)	Status(2)
Life, Alcohol, Measles, Under-Five Deaths, Total_expenditure	Alcohol, Life, GDP , Measles ,Population	Thine 5-9, Alcohol, Thine10_19, Life	Thine 5-9, Alcohol, Thine10_19, Life