World Happiness Report

2019

Term Project

수강과목 : 다변량 통계학 (I)

담당교수 : 최용석 교수님

학　　과 : 통계학과

학　　번 : 201711505

이　　름 : 김도희

제출일자 : 2019.06.24.(월)

목차

1. Abstract

전쟁 방지와 평화 유지를 위해 설립된 국제기구인 유엔(UN:United Nations)에서는 에르네스토 일리 재단과 협력하여 2012년도부터 매년 세계행복보고서를 작성해오고 있다.

세계행복보고서는 세계 156개국을 대상으로 자국민이 스스로를 얼마나 행복하다고 생각하는지를 보여주는 조사이다. 올해 2019년도의 세계행복 보고서는 행복과 공동체에 초점을 맞추고 있다. 지난 12년도 동안 행복이 어떻게 진화해왔는지, 이러한 변화를 이끈 기술, 사회적 규범, 갈등과 정부 정책에 초점을 맞추고 있다.

행복지수를 산출하는 기준을 알지 못한다는 가정 하에 2019년도에서 나라가 갖는 행복지수를 포함한 1인당 GDP, 기대수명 등 특성의 차이를 파악해보고자 한다.

2. Introduce

※ 분석에 사용되는 데이터에 관한 설명은 3. Data Description 에서 볼 수 있다.

※ 자료에 대한 분석에 해당하는 4. Analysis and Interpretation 에서는,
PCA(주성분 분석), FA(인자 분석), CA(군집 분석) 총 3가지를 이용할 것이다.
주성분 분석을 통해, 존재하는 여러 변수들을 잘 설명해 내는 적은 수의 새로운 변수를 생성한다. 다음으로, PCFA와 MLFA를 이용한 인자 분석을 통해, 변수가 가진 공통 요인을 찾아 더 적은 차원으로 변수를 설명한다. 마지막으로, 군집 분석을 통해, 자료를 크게 계층 방법과 비계층 방법으로 군집화해 최종적으로 해석할 것이다.

※ 종합적인 해석 결과는 5. Conclusion 에 기입하였다.

※ 자료의 출처는 6. Reference, R code는 7. Rcode 를 통해 확인가능하다.

3. Data Description

1) 자료

	Country	Happiness score	Whisker-high	Whisker-low	Dystopia (1.88) + residual	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
1	Finland	7.7689	7.829888	7.707912	2.7136779	1.34024286	1.5872757	0.9861450	0.59589535	0.15270843	0.392912716
2	Denmark	7.6001	7.666658	7.533542	2.3928909	1.38343859	1.5725950	0.9960189	0.59235609	0.25231999	0.410473198
3	Norway	7.5539	7.615639	7.492160	2.2407641	1.48776698	1.5815483	1.0281229	0.60349983	0.27130419	0.340883523
4	Iceland	7.4936	7.613283	7.373917	2.4008756	1.38016319	1.6236511	1.0256525	0.59090537	0.35435641	0.117979728
5	Netherlands	7.4876	7.542098	7.433102	2.3928947	1.39602041	1.5219034	0.9993137	0.55707520	0.32243958	0.297978073
6	Switzerland	7.4802	7.552696	7.407703	2.2721138	1.45224464	1.5262786	1.0519891	0.57151413	0.26346397	0.342615664
7	Sweden	7.3433	7.416333	7.270267	2.2455273	1.38657725	1.4873067	1.0092030	0.57442039	0.26702431	0.373202175
8	New Zealand	7.3075	7.382892	7.232108	2.1267915	1.30258632	1.5572339	1.0256352	0.58514649	0.32984269	0.380280942
9	Canada	7.2781	7.356539	7.199661	2.1926885	1.36489606	1.5047410	1.0388116	0.58395189	0.28502035	0.308037907
10	Austria	7.2460	7.312841	7.179158	2.3775482	1.37554193	1.4752225	1.0157766	0.53207481	0.24356669	0.226221070
11	Australia	7.2280	7.314588	7.141413	2.0942869	1.37154543	1.5479574	1.0355320	0.55717164	0.33154917	0.289962173
12	Costa Rica	7.1674	7.254208	7.080592	2.9329858	1.03424954	1.4411207	0.9631057	0.55803537	0.14439800	0.093470611
13	Israel	7.1387	7.205976	7.071424	2.6650524	1.27583969	1.4546424	1.0289690	0.37057629	0.26147476	0.082120553
14	Luxembourg	7.0903	7.152490	7.028111	1.9540858	1.60876155	1.4788067	1.0124820	0.52616233	0.19430040	0.315696567
15	United Kingdom	7.0537	7.125649	6.981750	2.1117339	1.33295250	1.5375850	0.9960179	0.44952208	0.34824628	0.277595162

[그림 3.1] Raw Data

```
> dim(hap)
[1] 156  11
```

[그림 3.2] Raw Data 크기

그림 3.1은 'worldhappiness.report'에서 가져온 데이터로 확장명 .xls 파일을 불러온 상태이다. 자료명은 "Chapter2OnlineData.xls"에 해당하며 2019년도에서의 자료를 담은 두 번째 시트의 자료를 이용한다.

그림 3.2를 통해 156개에 해당하는 나라와 11개의 변수가 있을을 알 수 있다. 이 중, 나라이름에 해당하는 첫 번째 열을 행의 이름으로 지정하고, 행복지수를 나타내는 변수 'Happiness score','Whisker-high','Whisker-low' 중 'Happiness score'만을 이용하였다.

2) 변수

※ Happiness score : 행복 지수

※ Whisker-high : 행복 지수의 제 3사분위 수(Q_3)

※ Whisker-low : 행복 지수의 제 1사분위 수(Q_1)

※ Dystopia (1.88) + residual : 2016-2018 평균 수명 평가(=1.88) + 각 나라

의 자체 예측 오류

※ Explained by: GDP per capita : 1인당 GDP

※ Explained by: Social support : 사회적 지원

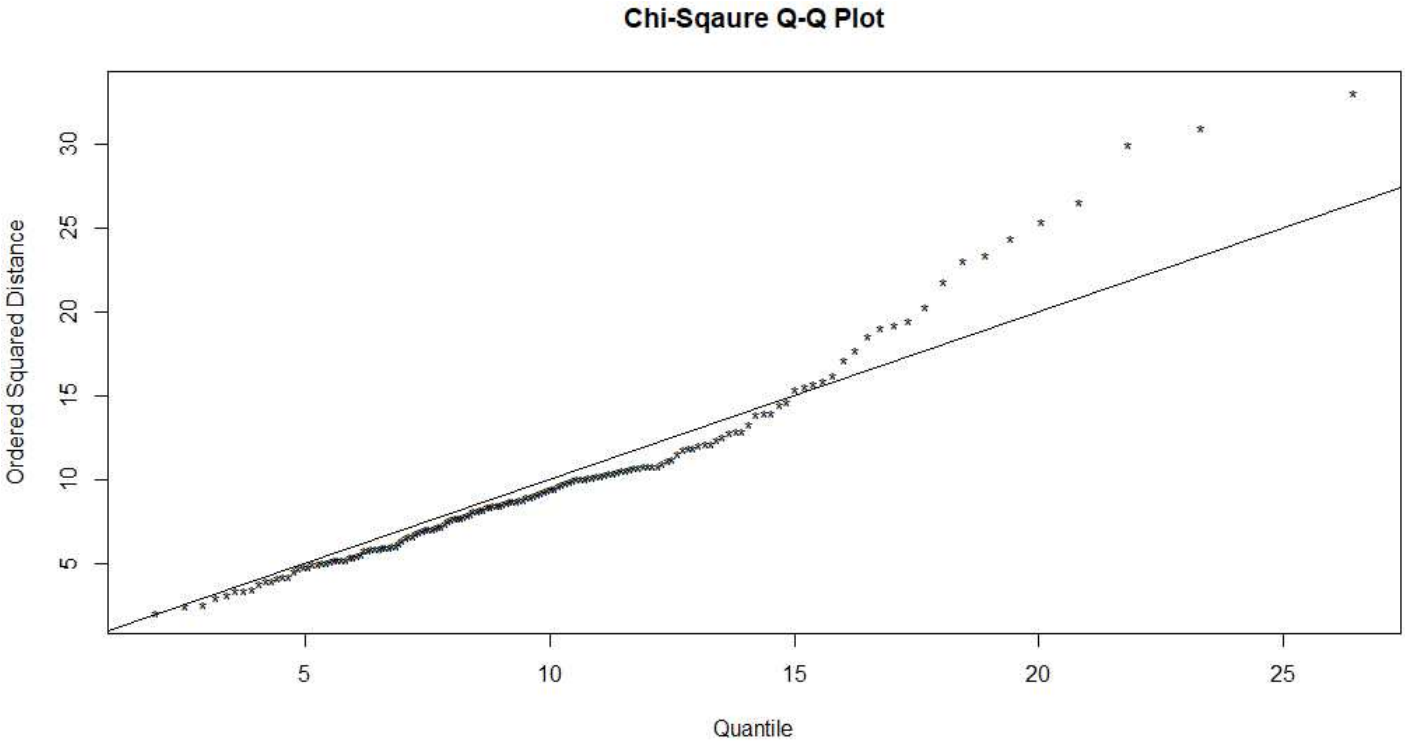
※ Explained by: Healthy life expectancy : 기대 수명(단위 100살)

※ Explained by: Freedom to make life choices : 직업 선택 자유의 정도

※ Explained by: Generosity : 관대하고 친사회적인 행동의 정도

※ Explained by: Perceptions of corruption : 부패에 대한 인식

3) 정규성



[그림 3.3] Chi-Square Q-Q Plot

```
> rq=cor(cbind(q, m))[1,2]
> rq
[1] 0.975556
```

[그림 3.4]

그림 3.2에서 대부분의 좌표점(*)이 직선상에 매우 가깝게 위치해 있는 것을 알 수 있다. 더군다나 그림3.4로 알 수 있듯이, 분위수와 마할라노비스 거리의 상관계수 0.975556의 값을 보면 거의 1이 되어 카이제곱그림의 직선성이 매우 인정되며 해당 데이터는 다변량 정규성을 만족한다고 볼 수 있다.

4. Analysis and Interpretation

1) PCA

① PCA에 이용할 공분산행렬과 상관행렬 중 선택

```
> round(gof, 2)
[1] 80.53 15.01  1.67  1.37  0.72  0.42  0.28  0.00
```

[그림 4.1]

그림 4.1은 공분산 행렬 결과이다. 첫 번째 eigenvalue의 설명비율의 합이 약 80.53%로 높은 설명력을 갖고 있어 하나의 주성분을 택한다.

```
> round(gof, 2)
[1] 47.92 17.87 14.53  7.65  6.81  3.27  1.97  0.00
```

[그림 4.2]

그림 4.2는 상관 행렬 결과이다. 첫 번째부터 세 번째 eigenvalue의 설명비율의 합이 약 80.32%로 70%보다 높아 3개의 주성분을 택한다.

공분산 행렬의 경우 1개의 성분만이 크게 영향을 받는 반면 상관행렬은 이를 해결한 것으로 볼 수 있다. 따라서 상관행렬을 사용하는 것이 좋다고 할 수 있다.

② 상관행렬을 이용한 PCA 결과

```
> V3
                                P1    P2    P3
Happiness score              -0.48  0.04  0.29
Dystopia (1.88) + residual   -0.09  0.07  0.91
GDP per capita               -0.45  0.20 -0.20
Social support               -0.43  0.20 -0.12
Healthy life expectancy      -0.45  0.17 -0.18
Freedom to make life choices -0.33 -0.36  0.05
Generosity                   -0.05 -0.70 -0.01
Perceptions of corruption    -0.24 -0.52 -0.06
```

[그림 4.3] 주성분 계수

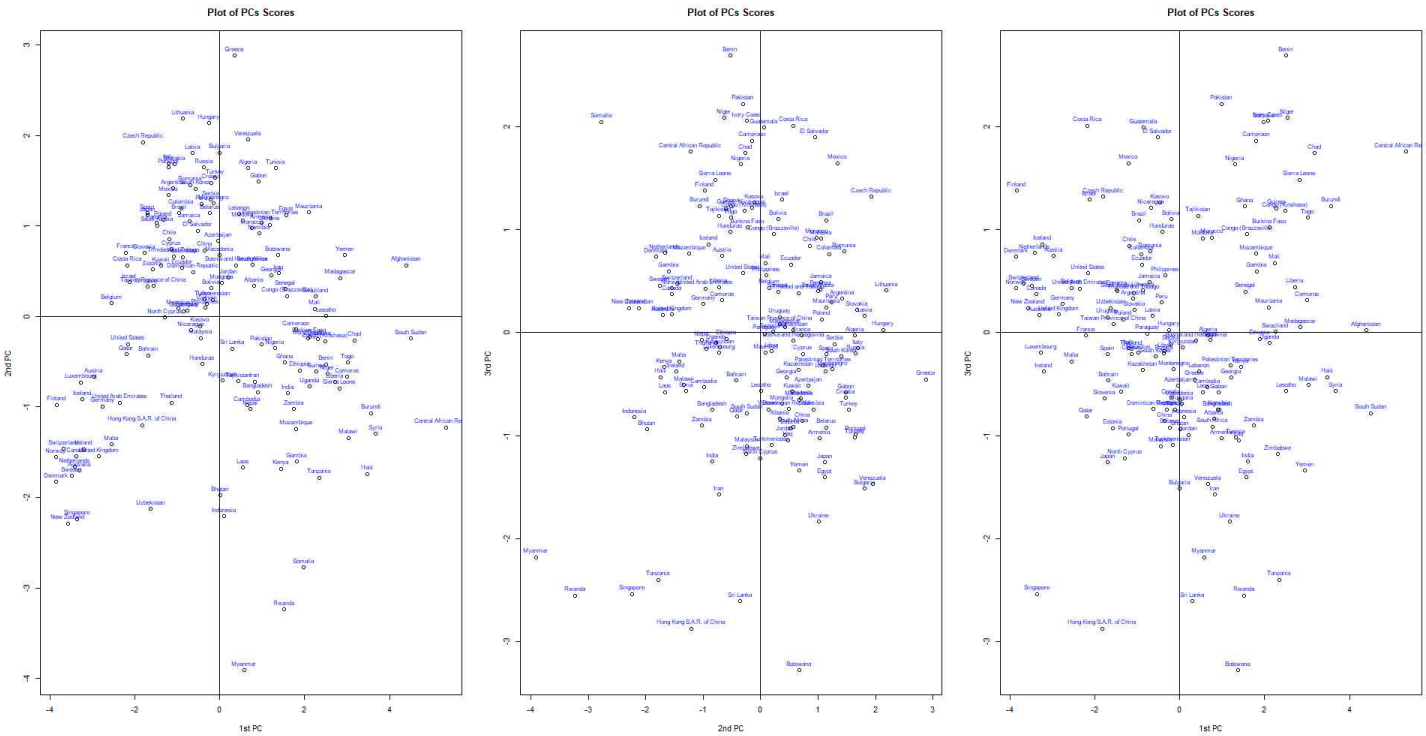
그림 4.3을 참고하여 주성분 식을 구하면 다음과 같이 표현될 수 있다.

$$P_1 = -0.48X_1 - 0.09X_2 - 0.45X_3 - 0.43X_4 - 0.45X_5 - 0.33X_6 - 0.05X_7 - 0.24X_8$$

$$P_2 = 0.04X_1 + 0.07X_2 + 0.20X_3 + 0.20X_4 + 0.17X_5 - 0.36X_6 - 0.70X_7 - 0.52X_8$$

$$P_3 = 0.29X_1 + 0.91X_2 - 0.20X_3 - 0.12X_4 - 0.18X_5 + 0.05X_6 - 0.01X_7 - 0.06X_8$$

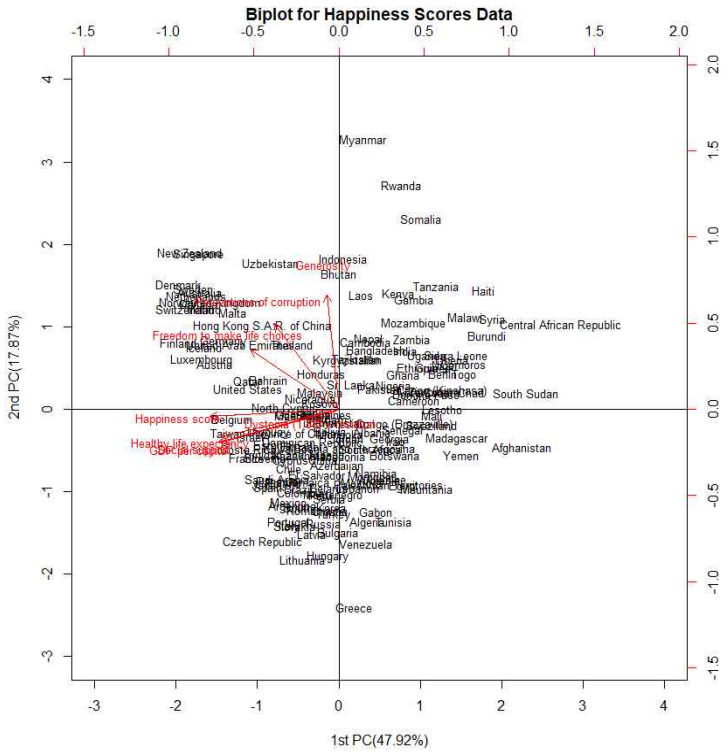
첫 번째 주성분의 경우 행복지수와 GDP, 사회적 지원, 기대수명에 가중치를 두고 있다. 두 번째 주성분의 경우 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식에 가중치를 두고 있다. 세 번째 주성분의 경우 Dystopia 변수에 가중치를 두고 있다.



[그림 4.4] Plot of PC's Scores

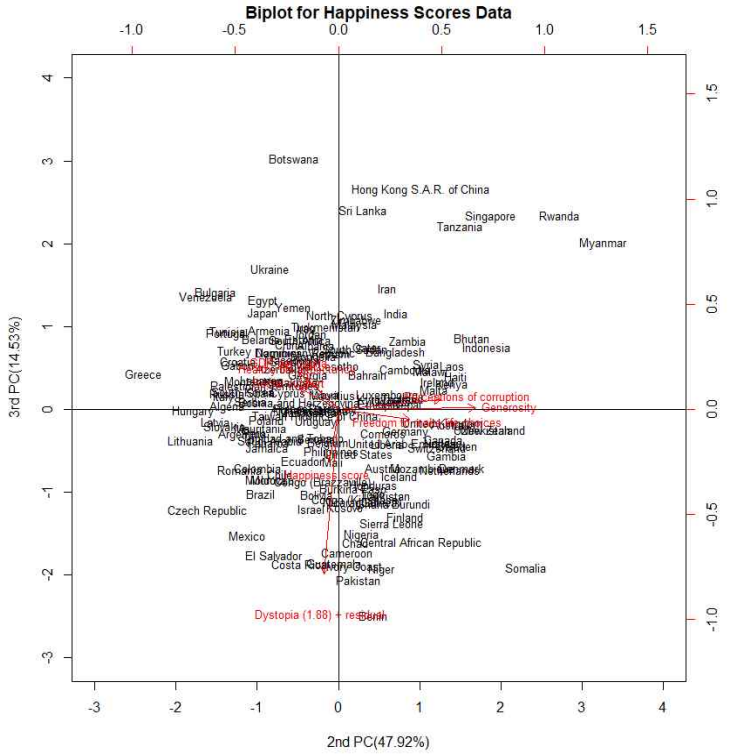
그림 4.4의 세 가지의 산점도 모두 타원형에 가까우므로 정규성을 만족한다.

③ Biplot



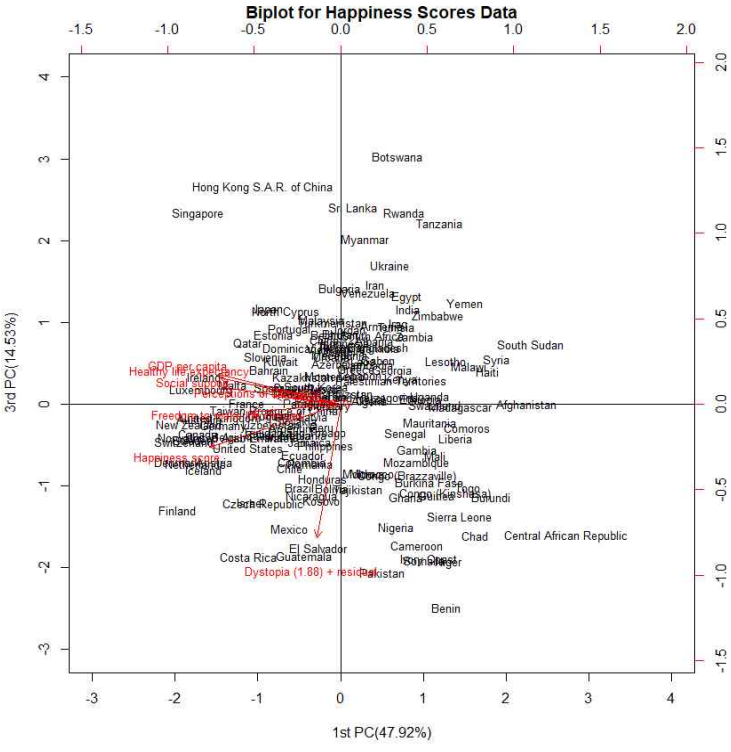
[그림 4.5] Biplot of 1ST PC & 2ND PC

그림 4.5의 Biplot을 통해 첫 번째 주성분에 대해 행복지수와 GDP, 사회적 지원, 기대수명이 음의 가중치를 주며, 두 번째 주성분에 대해 관대함의 정도가 가중치를 두고 있다.



[그림 4.6] Biplot of 2ND PC & 3RD PC

그림 4.6의 Biplot을 통해 두 번째 주성분에 대해 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식이 가중치를 주며, 세 번째 주성분에 대해 Dystopia가 가중치를 두고 있다.



[그림 4.7] Biplot of 1st PC & 3RD PC

그림 4.6의 Biplot에서도 각 주성분에 가중치를 두는 변수들이 동일하게 나타난다. 이는 주성분의 계수를 통해 해석한 결과와 같은 것을 알 수 있다.

2) FA

① PCFA와 MLFA의 결과 비교

	RC1	RC2	RC3
Happiness score	0.838	0.264	0.471
Dystopia (1.88) + residual			0.996
GDP per capita	0.935		
Social support	0.887		
Healthy life expectancy	0.909	0.103	
Freedom to make life choices	0.449	0.623	0.141
Generosity	-0.187	0.814	
Perceptions of corruption	0.249	0.746	

[그림 4.8]

	Factor1	Factor2	Factor3
Happiness score	0.846	0.471	0.241
Dystopia (1.88) + residual		0.997	
GDP per capita	0.992		
Social support	0.797		0.340
Healthy life expectancy	0.863		0.205
Freedom to make life choices	0.441		0.602
Generosity			0.489
Perceptions of corruption	0.335		0.399

[그림 4.9]

검은색 배경은 PCFA의 결과창이며, 하얀색 배경은 MLFA의 결과창으로 구별할 수 있다.

그림 4.8과 4.9를 통해, PCFA에서는 모든 변수들이 절대값이 0.6보다 큰 인자적재값을 갖고 세 가지의 공통인수에 대하여 해석이 가능하나, MLFA에서는 관대함의 정도와 부패에 대한 인식을 나타내는 변수의 인자적재 절대값이 최대 0.5가 안되는 값들로 다소 작은 값을 띄어 어느 인자 쪽으로도 해석되지 못한다. 따라서 이러한 관점에서는 PCFA가 선호된다.

	RC1	RC2	RC3
SS loadings	3.488	1.696	1.240

[그림 4.10]

	Factor1	Factor2	Factor3
SS loadings	3.387	1.229	0.985

[그림 4.11]

그림 4.10과 4.11을 통해 PCFA와 MLFA에서 총 기여율은 약 80.3%와 70%임을 알 수 있다. 이러한 관점에서 또한 PCFA가 선호된다.

```
> Psi
            Happiness score   Dystopia (1.88) + residual
                  0.007069639                 0.007636410
                 GDP per capita                Social support
                  0.118833206                 0.208385986
    Healthy life expectancy Freedom to make life choices
                  0.161851257                 0.390255404
                   Generosity     Perceptions of corruption
                  0.300976534                 0.380842676
```

[그림 4.12]

```
> Psi
            Happiness score   Dystopia (1.88) + residual
                  0.0050000                   0.0050000
                 GDP per capita                Social support
                  0.0050000                   0.2489364
    Healthy life expectancy Freedom to make life choices
                  0.2136678                   0.4346899
                   Generosity     Perceptions of corruption
                  0.7584466                   0.7287120
```

[그림 4.13]

그림 4.12와 4.13을 통해, PCFA와 MLFA에서 특정분산이 모두 0에 가까운 값으로 인자모형은 둘 다 자료를 잘 나타냈다고 보인다.

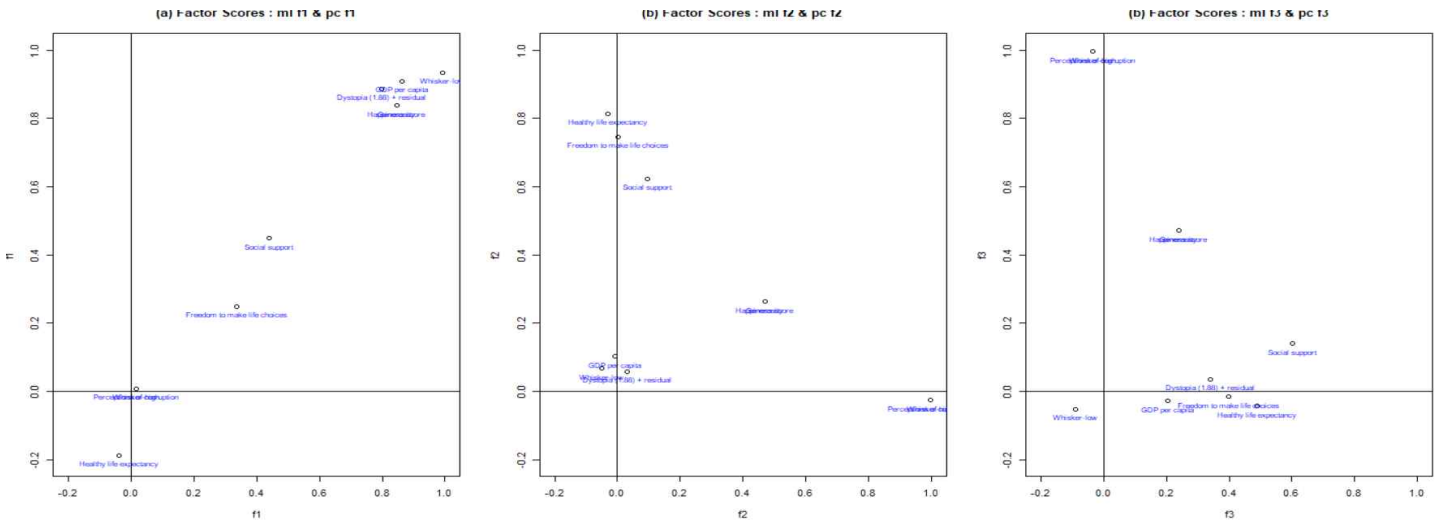
```
> round(Rm, 3)
                                X1     X2     X3     X4     X5     X6     X7     X8
Happiness score              0.000  0.005  0.018  0.002  0.005 -0.040  0.038 -0.012
Dystopia (1.88) + residual   0.005  0.000  0.012 -0.014  0.018 -0.051  0.013  0.025
GDP per capita               0.018  0.012  0.000 -0.076 -0.023 -0.075  0.038  0.014
Social support               0.002 -0.014 -0.076  0.000 -0.093  0.007  0.072 -0.082
Healthy life expectancy      0.005  0.018 -0.023 -0.093  0.000 -0.078  0.055 -0.009
Freedom to make life choices -0.040 -0.051 -0.075  0.007 -0.078  0.000 -0.147 -0.136
Generosity                   0.038  0.013  0.038  0.072  0.055 -0.147  0.000 -0.235
Perceptions of corruption   -0.012  0.025  0.014 -0.082 -0.009 -0.136 -0.235  0.000
```

[그림 4.14]

```
> round(Rm, 3)
                                X1     X2     X3     X4     X5     X6     X7     X8
Happiness score              0.000  0.000  0.000  0.007  0.005  0.005  0.005  0.004
Dystopia (1.88) + residual   0.000  0.000  0.000 -0.003 -0.002 -0.002 -0.002 -0.002
GDP per capita               0.000  0.000  0.000 -0.004 -0.003  0.000  0.000  0.001
Social support               0.007 -0.003 -0.004  0.000 -0.038 -0.111 -0.183 -0.221
Healthy life expectancy      0.005 -0.002 -0.003 -0.038  0.000 -0.112 -0.097 -0.076
Freedom to make life choices 0.005 -0.002  0.000 -0.111 -0.112  0.000 -0.005  0.051
Generosity                   0.005 -0.002  0.000 -0.183 -0.097 -0.005  0.000  0.145
Perceptions of corruption    0.004 -0.002  0.001 -0.221 -0.076  0.051  0.145  0.000
```

[그림 4.15]

그림 4.14와 4.15를 통해, PCFA와 MLFA에서 잔차행렬의 원소들이 모두 0에 가까운 값으로 모형이 매우 잘 적합되어 있다고 할 수 있다.



[그림 4.16]

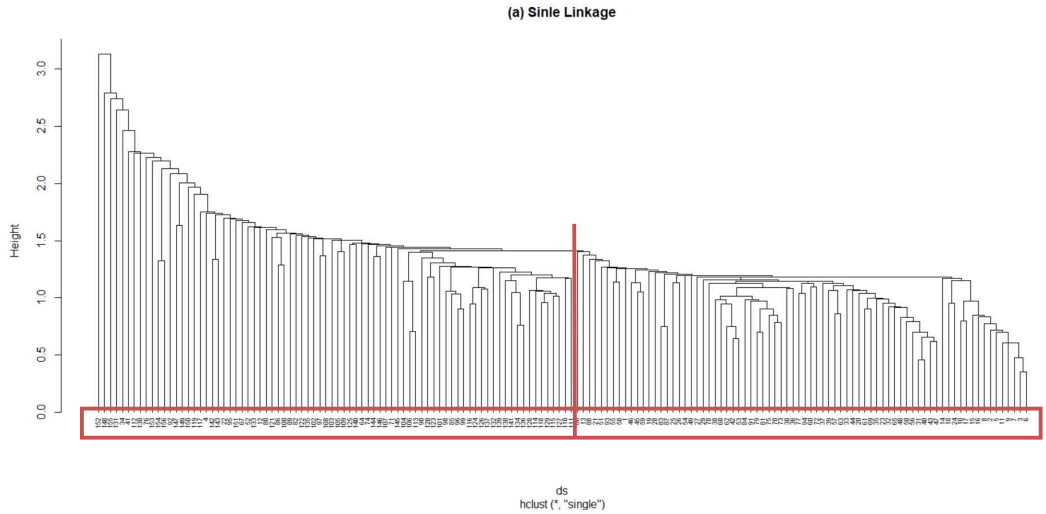
그림 4.16을 통해, 두 추정법 모두 (a)의 인자 f_1은 행복지수, 1인당 GDP, 사회적 지원, 기대수명으로 해석된다. 따라서 인자점수가 유사한 점을 원점에서 45^o직선을 형성하고 있는 것으로 확인할 수 있다. 직선을 벗어나는 개체는 없으므로 이상치 측정값은 찾을 수 없다고 본다. (b)의 인자 f_2는 PCFA에서는 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식으로 해석되며, MLFA에서는 Dystopia로 전혀 다르게 해석된다. (c)의 인자 f_3는 PCFA에서는 Dystopia로, MLFA에서는 직업 선택의 자

유로 전혀 다르게 해석된다.

비교 결과 상 해당 데이터에서는 MLFA보다 PCFA가 더 선호된다고 할 수 있다. PCFA로 차원축소 개념을 실현한 새로운 변수들은 PCA를 통해 얻어진 변수들의 해석 결과와 완벽히 일치하는 것을 볼 수 있다.

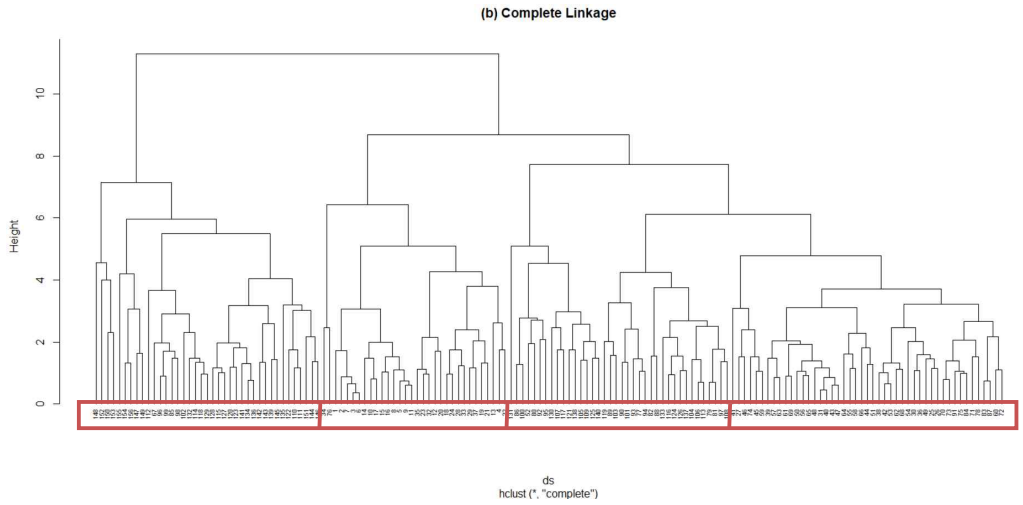
3) CA

① Hierarchical CA



[그림 4.17]

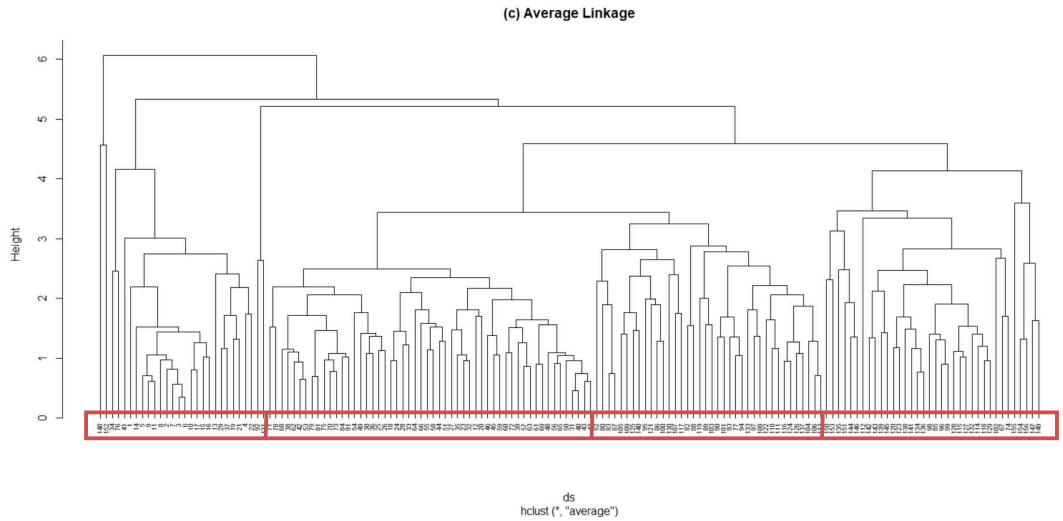
그림 4.17은 단일연결법에서의 덴드로그림으로 총 군집 수는 2개로 설정할 수 있다. 왼쪽 네모 칸부터 첫 번째 군집으로 네이밍하여 각 군집의 특성을 살펴보면, 첫 번째 군집은 행복지수 순위가 낮은 국가에 해당하는 나라들이 모여 있으며, 두 번째 군집은 행복지수가 상대적으로 높은 나라들로 구성되어 있다.



[그림 4.18]

그림 4.18은 완전 연결법에서의 덴드로그림으로 총 군집 수는 4개로 설정할 수 있다. 행복지수 최상위권 국가들로 구성된 군집부터 최하위권 국가들로 구성된 군집이 두 번

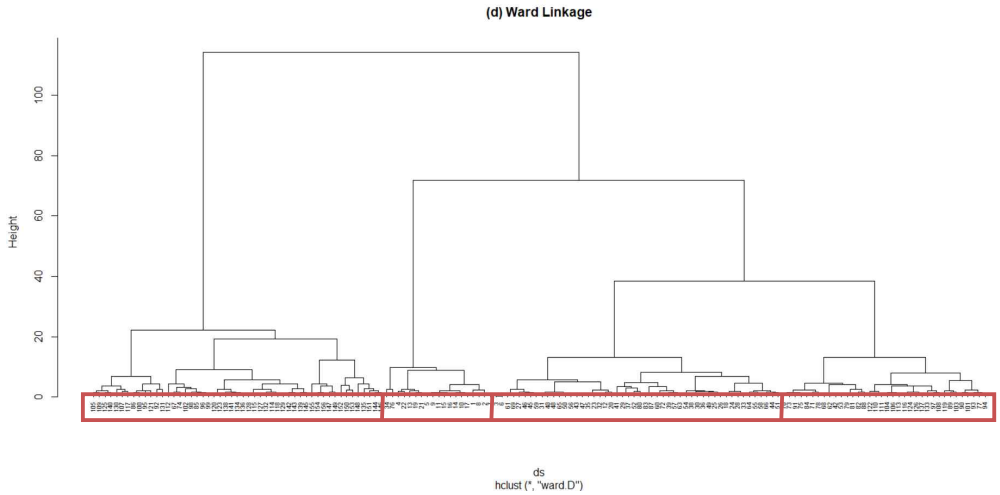
째 군집, 네 번째 군집, 세 번째 군집, 첫 번째 군집 순으로 나뉜 것을 확인할 수 있다.



[그림 4.19]

그림 4.19는 평균 연결법에서의 덴드로그림으로 총 군집 수는 4개로 설정할 수 있다.

첫 번째 군집부터, 두 번째, 세 번째, 네 번째 군집 순으로 행복지수 최상위권 국가에서 최하위권 국가들끼리 군집이 형성된 것을 알 수 있다.



[그림 4.20]

그림 4.20은 와드 연결법에서의 덴드로그림으로 총 군집 수는 4개로 설정할 수 있다.

이 또한 나라별 행복지수 순위를 기준으로 두 번째, 세 번째, 네 번째, 첫 번째 군집으로 나뉘어져 있는 것을 확인할 수 있다.

이때 말하는 행복지수 순위는 현재 데이터상에 들어있는 하나의 변수로서 2019년도에 한정되어 측정된 Happiness score 값의 순위를 말하는 것이 아닌 현재까지의 종합적인 나라별 행복지수 순위를 말하는 것이다.

② Non-Hierarchical CA

	rownames.X.	cluster
Finland	Finland	1
Denmark	Denmark	1
Norway	Norway	1
Iceland	Iceland	1
Netherlands	Netherlands	1
Switzerland	Switzerland	1
Sweden	Sweden	1
New Zealand	New Zealand	1
Canada	Canada	1
Austria	Austria	1
Australia	Australia	1
Luxembourg	Luxembourg	1
United Kingdom	United Kingdom	1
Ireland	Ireland	1
Germany	Germany	1
Belgium	Belgium	1
United States	United States	1
United Arab Emirates	United Arab Emirates	1
Malta	Malta	1
Qatar	Qatar	1
Singapore	Singapore	1
Uzbekistan	Uzbekistan	1

	rownames.X.	cluster
Finland	Finland	1
Denmark	Denmark	1
Norway	Norway	1
Iceland	Iceland	1
Netherlands	Netherlands	1
Switzerland	Switzerland	1
Sweden	Sweden	1
New Zealand	New Zealand	1
Canada	Canada	1
Austria	Austria	1
Australia	Australia	1
Luxembourg	Luxembourg	1
United Kingdom	United Kingdom	1
Ireland	Ireland	1
Germany	Germany	1
Belgium	Belgium	1
United States	United States	1
United Arab Emirates	United Arab Emirates	1
Malta	Malta	1
Singapore	Singapore	1
Uzbekistan	Uzbekistan	1

	rownames.X.	cluster
Costa Rica	Costa Rica	4
Israel	Israel	4
Czech Republic	Czech Republic	4
Mexico	Mexico	4
France	France	4
Taiwan Province of China	Taiwan Province of China	4
Chile	Chile	4
Guatemala	Guatemala	4
Saudi Arabia	Saudi Arabia	4
Spain	Spain	4
Panama	Panama	4
Brazil	Brazil	4
Uruguay	Uruguay	4
El Salvador	El Salvador	4
Italy	Italy	4
Bahrain	Bahrain	4
Slovakia	Slovakia	4
Trinidad and Tobago	Trinidad and Tobago	4
Poland	Poland	4
Lithuania	Lithuania	4
Colombia	Colombia	4
Slovenia	Slovenia	4
Nicaragua	Nicaragua	4
Kosovo	Kosovo	4
Argentina	Argentina	4
Romania	Romania	4
Cyprus	Cyprus	4
Ecuador	Ecuador	4
Kuwait	Kuwait	4
Thailand	Thailand	4
Latvia	Latvia	4
South Korea	South Korea	4
Estonia	Estonia	4
Jamaica	Jamaica	4
Mauritius	Mauritius	4
Japan	Japan	4
Honduras	Honduras	4
Kazakhstan	Kazakhstan	4
Bolivia	Bolivia	4
Hungary	Hungary	4
Paraguay	Paraguay	4
North Cyprus	North Cyprus	4
Peru	Peru	4
Portugal	Portugal	4
Russia	Russia	4
Philippines	Philippines	4
Serbia	Serbia	4
Moldova	Moldova	4

	rownames.X.	cluster
Costa Rica	Costa Rica	2
Israel	Israel	2
Czech Republic	Czech Republic	2
Mexico	Mexico	2
France	France	2
Taiwan Province of China	Taiwan Province of China	2
Chile	Chile	2
Guatemala	Guatemala	2
Saudi Arabia	Saudi Arabia	2
Qatar	Qatar	2
Spain	Spain	2
Panama	Panama	2
Brazil	Brazil	2
Uruguay	Uruguay	2
El Salvador	El Salvador	2
Italy	Italy	2
Bahrain	Bahrain	2
Slovakia	Slovakia	2
Trinidad and Tobago	Trinidad and Tobago	2
Poland	Poland	2
Lithuania	Lithuania	2
Colombia	Colombia	2
Slovenia	Slovenia	2
Nicaragua	Nicaragua	2
Kosovo	Kosovo	2
Argentina	Argentina	2
Romania	Romania	2
Cyprus	Cyprus	2
Ecuador	Ecuador	2
Kuwait	Kuwait	2
Latvia	Latvia	2
Estonia	Estonia	2
Jamaica	Jamaica	2
Mauritius	Mauritius	2
Japan	Japan	2
Honduras	Honduras	2
Kazakhstan	Kazakhstan	2
Bolivia	Bolivia	2
Paraguay	Paraguay	2
Peru	Peru	2
Portugal	Portugal	2
Philippines	Philippines	2

	rownames.X.	cluster
Libya	Libya	2
Montenegro	Montenegro	2
Tajikistan	Tajikistan	2
Croatia	Croatia	2
Hong Kong S.A.R. of China	Hong Kong S.A.R. of China	2
Dominican Republic	Dominican Republic	2
Bosnia and Herzegovina	Bosnia and Herzegovina	2
Turkey	Turkey	2
Malaysia	Malaysia	2
Belarus	Belarus	2
Greece	Greece	2
Mongolia	Mongolia	2
Macedonia	Macedonia	2
Kyrgyzstan	Kyrgyzstan	2
Turkmenistan	Turkmenistan	2
Algeria	Algeria	2
Morocco	Morocco	2
Azerbaijan	Azerbaijan	2
Lebanon	Lebanon	2
Indonesia	Indonesia	2
China	China	2
Vietnam	Vietnam	2
Bhutan	Bhutan	2
Bulgaria	Bulgaria	2
Nepal	Nepal	2
Jordan	Jordan	2
Gabon	Gabon	2
Laos	Laos	2
South Africa	South Africa	2
Albania	Albania	2
Venezuela	Venezuela	2
Cambodia	Cambodia	2
Palestinian Territories	Palestinian Territories	2
Namibia	Namibia	2
Armenia	Armenia	2
Iran	Iran	2
Georgia	Georgia	2
Tunisia	Tunisia	2
Bangladesh	Bangladesh	2
Iraq	Iraq	2
Sri Lanka	Sri Lanka	2
Myanmar	Myanmar	2
Ukraine	Ukraine	2
Egypt	Egypt	2
Botswana	Botswana	2

	rownames.X.	cluster
Thailand	Thailand	3
South Korea	South Korea	3
Hungary	Hungary	3
North Cyprus	North Cyprus	3
Russia	Russia	3
Serbia	Serbia	3
Moldova	Moldova	3
Libya	Libya	3
Montenegro	Montenegro	3
Croatia	Croatia	3
Hong Kong S.A.R. of China	Hong Kong S.A.R. of China	3
Dominican Republic	Dominican Republic	3
Bosnia and Herzegovina	Bosnia and Herzegovina	3
Turkey	Turkey	3
Malaysia	Malaysia	3
Belarus	Belarus	3
Greece	Greece	3
Mongolia	Mongolia	3
Macedonia	Macedonia	3
Kyrgyzstan	Kyrgyzstan	3
Turkmenistan	Turkmenistan	3
Algeria	Algeria	3
Morocco	Morocco	3
Azerbaijan	Azerbaijan	3
Lebanon	Lebanon	3
Indonesia	Indonesia	3
China	China	3
Vietnam	Vietnam	3
Bhutan	Bhutan	3
Bulgaria	Bulgaria	3
Nepal	Nepal	3
Jordan	Jordan	3
Gabon	Gabon	3
Laos	Laos	3
South Africa	South Africa	3
Albania	Albania	3
Venezuela	Venezuela	3
Cambodia	Cambodia	3
Palestinian Territories	Palestinian Territories	3
Namibia	Namibia	3
Armenia	Armenia	3
Iran	Iran	3
Tunisia	Tunisia	3
Bangladesh	Bangladesh	3
Iraq	Iraq	3
Sri Lanka	Sri Lanka	3
Myanmar	Myanmar	3
Ukraine	Ukraine	3
Egypt	Egypt	3
Botswana	Botswana	3

	rownames.X.	cluster
Pakistan	Pakistan	3
Nigeria	Nigeria	3
Cameroon	Cameroon	3
Ghana	Ghana	3
Ivory Coast	Ivory Coast	3
Benin	Benin	3
Congo (Brazzaville)	Congo (Brazzaville)	3
Senegal	Senegal	3
Somalia	Somalia	3
Niger	Niger	3
Burkina Faso	Burkina Faso	3
Guinea	Guinea	3
Gambia	Gambia	3
Kenya	Kenya	3
Mauritania	Mauritania	3
Mozambique	Mozambique	3
Congo (Kinshasa)	Congo (Kinshasa)	3
Mali	Mali	3
Sierra Leone	Sierra Leone	3
Chad	Chad	3
Ethiopia	Ethiopia	3
Swaziland	Swaziland	3
Uganda	Uganda	3
Zambia	Zambia	3
Togo	Togo	3
India	India	3
Liberia	Liberia	3
Comoros	Comoros	3
Madagascar	Madagascar	3
Lesotho	Lesotho	3
Burundi	Burundi	3
Zimbabwe	Zimbabwe	3
Haiti	Haiti	3
Syria	Syria	3
Malawi	Malawi	3
Yemen	Yemen	3
Rwanda	Rwanda	3
Tanzania	Tanzania	3
Afghanistan	Afghanistan	3
Central African Republic	Central African Republic	3
South Sudan	South Sudan	3

	rownames.X.	cluster
Pakistan	Pakistan	4
Tajikistan	Tajikistan	4
Nigeria	Nigeria	4
Cameroon	Cameroon	4
Ghana	Ghana	4
Ivory Coast	Ivory Coast	4
Benin	Benin	4
Congo (Brazzaville)	Congo (Brazzaville)	4
Senegal	Senegal	4
Somalia	Somalia	4
Niger	Niger	4
Burkina Faso	Burkina Faso	4
Guinea	Guinea	4
Georgia	Georgia	4
Gambia	Gambia	4
Kenya	Kenya	4
Mauritania	Mauritania	4
Mozambique	Mozambique	4
Congo (Kinshasa)	Congo (Kinshasa)	4
Mali	Mali	4
Sierra Leone	Sierra Leone	4
Chad	Chad	4
Ethiopia	Ethiopia	4
Swaziland	Swaziland	4
Uganda	Uganda	4
Zambia	Zambia	4
Togo	Togo	4
India	India	4
Liberia	Liberia	4
Comoros	Comoros	4
Madagascar	Madagascar	4
Lesotho	Lesotho	4
Burundi	Burundi	4
Zimbabwe	Zimbabwe	4
Haiti	Haiti	4
Syria	Syria	4
Malawi	Malawi	4
Yemen	Yemen	4
Rwanda	Rwanda	4
Tanzania	Tanzania	4
Afghanistan	Afghanistan	4
Central African Republic	Central African Republic	4
South Sudan	South Sudan	4

[그림 4.21]

그림 4.21의 왼쪽의 결과는 K-평균법, 오른쪽의 결과는 K-대표개체법이다.

결과를 보면 거의 모든 나라들을 4가지의 같은 군집 형태로 묶은 것을 확인할 수 있다.

첫 번째로 'Finland', 'Denmark', 'Norway', 'Sweden', 'Australia' 등의 나라로 구성된 군집의 특성을 보면, 행복지수가 가장 낮은 나라들이 모여 있으며, 사회적 지원, 기대수명, 직업 선택의 자유가 가장 낮고, 1인당 GDP, 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식 또한 낮은 것을 볼 수 있다.

두 번째로 'Turkey', 'Malaysia', 'Greece', 'China', 'Vietnam' 등의 나라로 구성된 군집의 특성을 보면, 행복지수가 가장 높은 나라들이 모여 있으며, 1인당 GDP, 사회적 지원, 기대수명, 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식 모두 가장 높은 것을 알 수 있다.

세 번째로 'Ghana', 'India', 'Kenya', 'Pakistan' 등의 나라로 구성된 군집의 특성을 보면, 행복 지수가 대체로 낮으며, 1인당 GDP, 사회적 지원, 기대수명 또한 낮은 축에 속하는 것을 알 수 있다.

네 번째로 'Israel', 'Mexico', 'Frnace', 'Chile' 등의 나라로 구성된 군집의 특성을 보면, 행복 지수가 대체로 높은 나라들이며, 1인당 GDP, 사회적 지원, 기대수명 또한 높은 축에 속하는 것을 알 수 있다.

```
> aggregate(X, by=list(kmeans$cluster),FUN=mean)
  Group.1 Happiness score Dystopia (1.88) + residual GDP per capita Social support
1       1        4.288419                    2.124509      0.3653088       0.840351
2       2        7.023639                    2.039350      1.3949717       1.494888
3       3        4.492070                    1.243882      0.7823275       1.101292
4       4        5.862921                    1.978556      1.0819626       1.356685
  Healthy life expectancy Freedom to make life choices Generosity Perceptions of corruption
1               0.3849968                    0.2817724  0.2008516                0.09062358
2               0.9904834                    0.5479258  0.2749234                0.28109788
3               0.6519419                    0.4054935  0.2129103                0.09422837
4               0.8498653                    0.3929380  0.1327447                0.07017306
```

[그림 4.22]

위의 각 군집 특성에 대한 해석을 그림 4.22를 통해 좀 더 쉽게 파악할 수 있다.

5. Conclusion

PCA, FA, CA를 통해 나온 결과를 종합해보면,

먼저 PCA의 경우 구한 3개의 주성분은 약 80%의 설명력을 갖고 주어져있던 변수들을 설명하는 새로운 변수들이다. 첫 번째 주성분의 경우 행복지수와 GDP, 사회적 지원, 기대수명에, 두 번째 주성분의 경우 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식에, 세 번째 주성분의 경우 Dystopia 변수에 가중치를 둔 주성분이다.

다음으로 FA의 경우, 인자 f_1은 행복지수, 1인당 GDP, 사회적 지원, 기대수명으로, f_2는 직업 선택의 자유, 관대함의 정도, 부패에 대한 인식으로, f_3는 Dystopia로 해석된다. 즉, PCA와 FA를 통해 차원 축소를 실현한 새로운 변수들이 동일한 특성을 갖는다고 할 수 있다.

마지막으로 CA의 경우, Hierachical CA(single, complete, average, ward)와 Non-Hierachical CA(K-means, K-medoids)의 결과로 묶인 나라들의 특성을 살펴보면, 현재까지의 행복지수 순위에 따라 군집화되어 나타나는 것을 확인할 수 있다.

이를 통해 행복지수가 높은 나라들의 특성은 1인당 GDP, 사회적 지원 정도, 기대수명, 직업선택의 자유, 관대함의 정도, 부패에 대한 인식이 높은 나라라는 것임을 알 수 있다. 따라서 행복지수는 위 변수들과의 상호 연관관계가 존재한다고 할 수 있다. 이에 자신의 나라가 행복지수가 높지 않다는 생각이 들어 이를 개선하고자 할 때에는 위의 수치들이 모두 높다고 판단될 수 있는지를 먼저 확인하는 과정이 필요하다고 할 수 있을 것이다.

6. References

※ 데이터 : 구글, "World happiness report"
, https://worldhappiness.report/ed/2019/, (2019.06.22.)

※ 참고 : John F. Helliwell, Richard Layard and Jeffrey D. Sachs, "World Happiness Report 2019"
R과 함께하는 다변량 자료분석(최용석 지음)
실습자료 R코드(https://stat.pusan.ac.kr/stat/49709/subview.do)
"UN", 위키피디아, https://en.wikipedia.org/wiki/United_Nations, (2019.06.23.)

7. R code

```
######################### Term Project_MVN
######################### 201711505_통계학과_김도희


setwd('C:\\Users\\kheed\\Desktop')

####install.packages

install.packages('readxl')
install.packages('dplyr')
install.packages('MVN')
install.packages('psych')
install.packages('biotools')
install.packages('cluster')


### 데이터 불러오기

library(readxl)
hap<-read_excel('Chapter2OnlineData.xls',sheet=2)
head(hap)
hap<-as.matrix(hap)
country<-hap[,1]
rownames(hap)<-country
hap<-hap[,-1]
head(hap)
View(hap)
dim(hap)
n=dim(hap)[1]
p=dim(hap)[2]
colnames(hap)
X<-matrix(NA, n,p )
for ( i in 1:p ) {
  X[,i]<-as.numeric(hap[,i])
}


colnames(X)<-colnames(hap)
rownames(X)<-rownames(hap)
X<-as.data.frame(X)

library(dplyr)
glimpse(X)
str(X)
```

```
plot(X)
colnames(X)[5:10]<-substr(colnames(X)[5:10],15,100)
X<-X[,-c(2,3)]
n=dim(X)[1]
p=dim(X)[2]

##### 3-3) 정규성 검토

xbar=colMeans(X)
S=cov(X)
m=mahalanobis(X, xbar, S)
m=sort(m)
id=seq(1, n)
pt=(id-0.5)/n
q=qchisq(pt, p)
plot(q, m, pch="*", xlab="Quantile", ylab="Ordered Squared Distance", main='Chi-Sqaure Q-Q Plot')
abline(0, 1)

rq=cor(cbind(q, m))[1,2]
rq


##### 4-1) PCA

# S 선택
S<-cov(X)
eigen.S=eigen(S)
round(eigen.S$values, 3) # Eigenvalus
V=round(eigen.S$vectors, 3) # Eigenvaectors
V
gof=eigen.S$values/sum(eigen.S$values)*100 # Goodness-of fit
round(gof, 2)

# R 선택
R=round(cor(X),3)
R
eigen.R=eigen(R)
round(eigen.R$values, 2) # Eigenvalues
V=round(eigen.R$vectors, 2) # Eigenvectors
gof=eigen.R$values/sum(eigen.R$values)*100 # Goodness-of fit
round(gof, 2)
plot(eigen.R$values, type="b", main="Scree Graph", xlab="Component Number", ylab="Eigenvalue")
V3=V[,1:3]
rownames(V3)<-colnames(X)
colnames(V3)<-c('P1','P2','P3')
V3
Z=scale(X, scale=T) # Standardized Data Matrix
Z
P=Z%*%V3              # PCs Scores
```

```
round(P, 3)
par(mfrow=c(1,3))
plot(P[,1], P[, 2], main="Plot of PCs Scores", xlab="1st PC", ylab="2nd PC")
text(P[,1], P[, 2], labels=rownames(X), cex=0.8, col="blue", pos=3)
abline(v=0, h=0)
plot(P[,2], P[, 3], main="Plot of PCs Scores", xlab="2nd PC", ylab="3rd PC")
text(P[,2], P[, 3], labels=rownames(X), cex=0.8, col="blue", pos=3)
abline(v=0, h=0)
plot(P[,1], P[, 3], main="Plot of PCs Scores", xlab="1st PC", ylab="3rd PC")
text(P[,1], P[, 3], labels=rownames(X), cex=0.8, col="blue", pos=3)
abline(v=0, h=0)


# Biplot
Y <- scale(X,scale=T)
svd.Y <- svd(Y)
U <- svd.Y$u
V <- svd.Y$v
D <- diag(svd.Y$d)
G <- (sqrt(n-1)*U)[,1:3]
H <- (sqrt(1/(n-1))*V%*%D)[,1:3]
rownames(G)<-rownames(X)
rownames(H)<-colnames(X)
par(mfrow=c(1,1))
lim<-range(pretty(G))
biplot(G[,1:2],H[,1:2], xlab="1st PC(47.92%)",ylab="2nd PC(17.87%)", main="Biplot for Happiness Scores
Data",
       xlim=lim,ylim=lim,cex=0.8,pch=16)
abline(v=0,h=0)
biplot(G[,2:3],H[,2:3], xlab="2nd PC(47.92%)",ylab="3rd PC(14.53%)", main="Biplot for Happiness
Scores Data",
       xlim=lim,ylim=lim,cex=0.8,pch=16)
abline(v=0,h=0)
biplot(G[,c(1,3)],H[,c(1,3)], xlab="1st PC(47.92%)",ylab="3rd PC(14.53%)", main="Biplot for Happiness
Scores Data",
       xlim=lim,ylim=lim,cex=0.8,pch=16)
abline(v=0,h=0)




##### 4-2) FA

###PCFA
library(psych)
R<-cor(X)
pcfa<-principal(R, nfactors=3, rotate="varimax")
pcfa$loadings
```

```
Psi=pcfa$uniquenesses
Rm = R-(L%*%t(L) + diag(Psi))
colnames(Rm)<-c('X1','X2','X3','X4','X5','X6','X7','X8')
round(Rm, 3)


###MLFA
library(psych)
mlfa<-factanal(covmat=R, factors = 3,  rotation="varimax") # rotation="none"
mlfa$loadings
Psi=mlfa$uniquenesses # specific variance
Rm = R-(L%*%t(L) + diag(Psi))
colnames(Rm)<-c('X1','X2','X3','X4','X5','X6','X7','X8')
round(Rm, 3)




## 결과비교
par(mfrow=c(1,3))
fpc<-pcfa$loadings[,1:3]
fml<-mlfa$loadings[,1:3]
lim<-range(pretty(fml))
plot(fml[,1], fpc[,1],main="(a) Factor Scores : ml f1 & pc f1",  xlab="f1", ylab="f1",
     xlim=lim, ylim=lim)
text(fml[,1], fpc[,1], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
plot(fml[,2], fpc[,2],main="(b) Factor Scores : ml f2 & pc f2",  xlab="f2", ylab="f2",
     xlim=lim, ylim=lim)
text(fml[,2], fpc[,2], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)
plot(fml[,3], fpc[,3],main="(b) Factor Scores : ml f3 & pc f3",  xlab="f3", ylab="f3",
     xlim=lim, ylim=lim)
text(fml[,3], fpc[,3], labels=rownames(L), cex=0.8, col="blue", pos=1)
abline(v=0, h=0)


##### 4-3) CA



#### Hierarchical CA
Z<-scale(X)
rownames(Z)<-seq(1:156)
ds <- dist(Z, method="euclidean")
round(ds, 3)
#단일연결법
single=hclust(ds, method="single")
```

```
plot(single, hang=-1, main="(a) Sinle Linkage", cex=0.6)

#완전연결법
complete=hclust(ds, method="complete")
plot(complete,hang=-1, main="(b) Complete Linkage", cex=0.6)

#평균연결법
average=hclust(ds, method="average")
plot(average, hang=-1, main="(c) Average Linkage", cex=0.6)

#와드연결법
ward=hclust(ds, method="ward.D")
plot(ward, hang=-1, main="(d) Ward Linkage", cex=0.6)



#### Non-Hierarchical CA

kmeans <- kmeans(Z, 4) # 4 cluster solution
cluster=data.frame(rownames(X),cluster=kmeans$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C1;C2;C3;C4

aggregate(X, by=list(kmeans$cluster),FUN=mean)



library(cluster)
kmedoids<-pam(Z, 4, metric='euclidean')
cluster<-data.frame(rownames(X), cluster=kmedoids$cluster)
C1=cluster[(cluster[,2]==1),]
C2=cluster[(cluster[,2]==2),]
C3=cluster[(cluster[,2]==3),]
C4=cluster[(cluster[,2]==4),]
C1;C2;C3;C4
```