다변량분석 Assignment 2

2014170824 황태민

[K-means clustering]

[Q1-1]

```
Clustering Methods:
 kmeans

Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                            2        3        4        5        6        7        8        9       10

kmeans APN             0.1287   0.0436   0.1344   0.1815   0.2321   0.1748   0.1632   0.2372   0.3118
       AD              5.0040   4.3162   4.2363   4.1467   4.1325   3.9275   3.8215   3.8121   3.8180
       ADM             0.7046   0.1872   0.6012   0.8891   1.1200   0.6406   0.5745   0.8429   1.1216
       FOM             0.9627   0.8144   0.7938   0.7653   0.7755   0.7569   0.7439   0.7389   0.7334
       Connectivity  100.0270 194.8433 279.2139 258.9845 300.7508 283.7972 398.0944 434.5683 458.6734
       Dunn            0.0842   0.0611   0.0439   0.0481   0.0481   0.0577   0.0577   0.0679   0.0947
       Silhouette      0.3201   0.2421   0.1966   0.1883   0.1840   0.1908   0.1682   0.1500   0.1343

Optimal Scores:

             Score    Method Clusters
APN            0.0436 kmeans 3
AD             3.8121 kmeans 9
ADM            0.1872 kmeans 3
FOM            0.7334 kmeans 10
Connectivity 100.0270 kmeans 2
Dunn           0.0947 kmeans 10
Silhouette     0.3201 kmeans 2

> end_time <- Sys.time()
> end_time - start_time
Time difference of 22.7762 secs
```

clValid() 함수를 사용하여 위와 같은 internal 및 stability 관련 타당성 지표 값을 산출하였다. 총 소요시간은 Sys.time() 함수를 사용하여 측정하였고 22.7762초 소요되었다. Dunn index 기준 최적의 군집 수는 10개이고, Silhouette index 기준 최적의 군집 수는 2개이다.

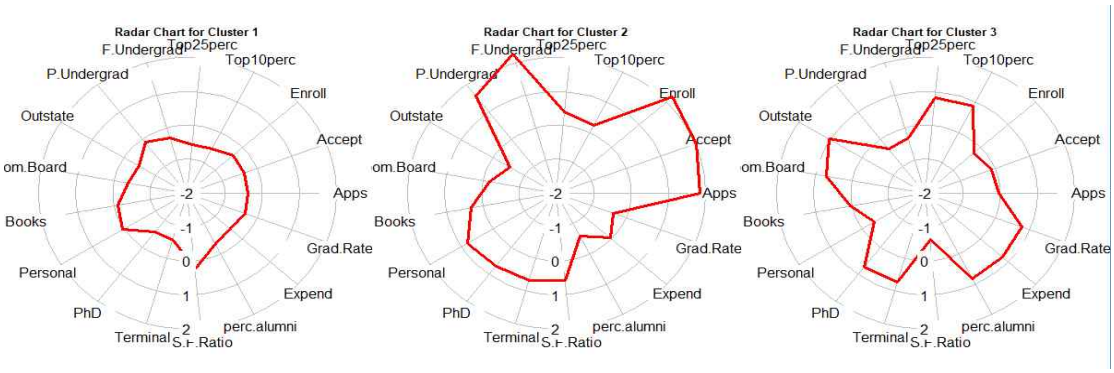
[Q1-2]

10회 반복 수행 시 군집의 순서만 달라졌을 뿐이지 모두 동일한 군집이 생성되었다.

[Q1-3]

K=10으로 군집화를 10회 반복 수행하였을 때, 10번 모두 다른 군집을 생성하였다.

[Q1-4]



K=3으로 군집화를 수행하고, Radar chart를 도시했을 때, 위와 같은 결과를 얻을 수 있다. 상대적으로 유사한 두 개의 군집은 cluster 1과 cluster 3이 유사한 것으로 보인다. Outstate, Room.board, Ph.D 등의 지표에서 차이를 보이긴 하지만 그나마 두 군집이 유사한 편이다. 가장 다른 두 개의 군집 쌍은 cluster 1과 cluster 2가 가장 확연한 차이를 보인다. 대부분의 지표에서 확연한 차이를 보인다.

[Q1-5]

#cluster 1 vs cluster 2

	v1	v2	v3
Apps	3.058227e-24	1.0000000000	1.529113e-24
Accept	8.506840e-27	1.0000000000	4.253420e-27
Enroll	1.835458e-35	1.0000000000	9.177291e-36
Top10perc	2.272173e-10	0.9999999999	1.136087e-10
Top25perc	4.337600e-16	1.0000000000	2.168800e-16
F.Undergrad	2.236426e-38	1.0000000000	1.118213e-38
P.Undergrad	1.399637e-12	1.0000000000	6.998186e-13
Outstate	4.480168e-01	0.2240083907	7.759916e-01
Room.Board	6.579076e-02	0.9671046214	3.289538e-02
Books	4.343236e-06	0.9999978284	2.171618e-06
Personal	4.085254e-10	0.9999999998	2.042627e-10
PhD	1.758412e-60	1.0000000000	8.792060e-61
Terminal	6.774873e-57	1.0000000000	3.387436e-57
S.F.Ratio	2.548733e-03	0.9987256337	1.274366e-03
perc.alumni	1.159043e-03	0.0005795213	9.994205e-01
Expend	2.541693e-07	0.9999998729	1.270846e-07
Grad.Rate	9.221614e-01	0.4610807226	5.389193e-01

유의수준은 0.01이라고 설정하자. v2가 0.01보다 작은 경우 cluster 1이 cluster 2보다 해당 지표에서 큰 수치를 가진다고 볼 수 있다. v3는 반대로 cluster 2가 cluster 1보다 큰 수치를 가진다고 볼 수 있다. perc.alumni의 지표에서는 cluster 1이 cluster 2 보다 큰 값을 가진다. 그리고 Outstate와 Room.Board의 지표는 두 cluster간의 차이가 있다고 보기 어렵다. 그 외의 모든 지표는 cluster 2가 cluster 1보다 크다.

#cluster 1 vs cluster 3

	v1	v2	v3
Apps	8.982960e-12	1.000000e+00	4.491480e-12
Accept	1.686494e-10	1.000000e+00	8.432471e-11
Enroll	2.082504e-03	9.989587e-01	1.041252e-03
Top10perc	4.705503e-57	1.000000e+00	2.352752e-57
Top25perc	7.271735e-82	1.000000e+00	3.635867e-82
F.Undergrad	7.440255e-01	6.279872e-01	3.720128e-01
P.Undergrad	8.973497e-15	4.486749e-15	1.000000e+00
Outstate	1.023578e-92	1.000000e+00	5.117892e-93
Room.Board	1.833853e-44	1.000000e+00	9.169264e-45
Books	5.283634e-02	9.735818e-01	2.641817e-02
Personal	2.231088e-08	1.115544e-08	1.000000e+00
PhD	3.644528e-86	1.000000e+00	1.822264e-86
Terminal	2.591698e-90	1.000000e+00	1.295849e-90
S.F.Ratio	5.700492e-36	2.850246e-36	1.000000e+00
perc.alumni	5.940226e-47	1.000000e+00	2.970113e-47
Expend	3.779045e-35	1.000000e+00	1.889523e-35
Grad.Rate	2.269348e-57	1.000000e+00	1.134674e-57

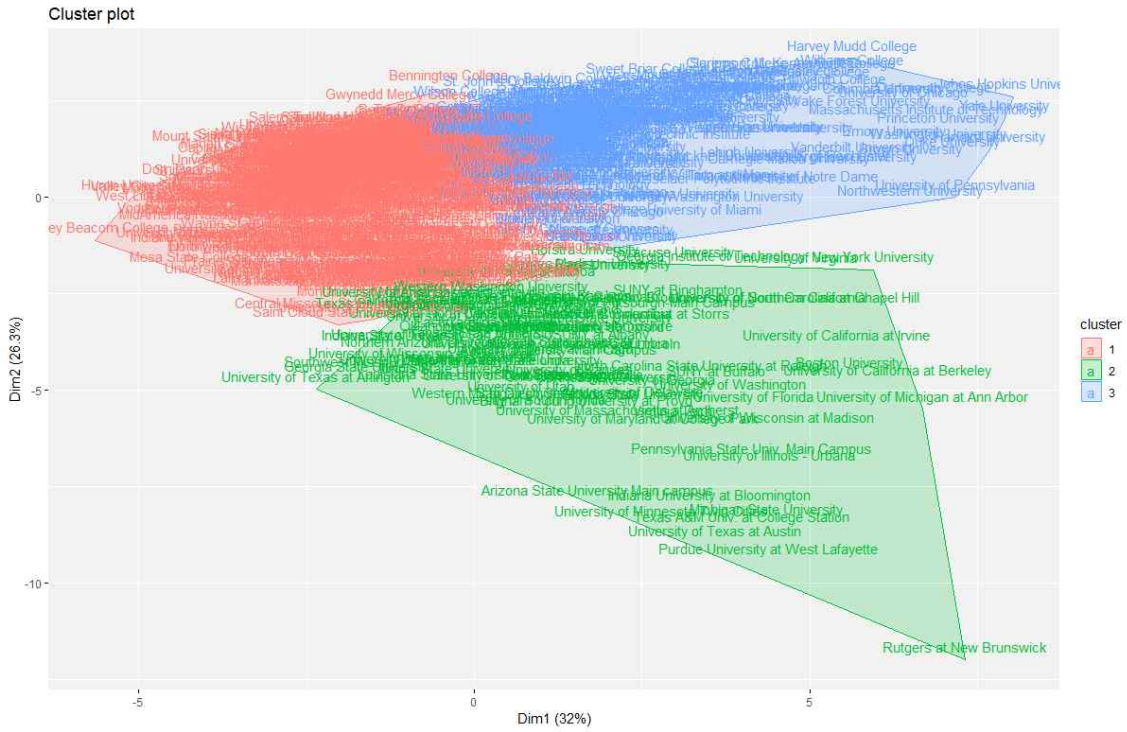
위에서와 동일한 방법으로 분석해보자. P.Undergrad, Personal, S.F.Ratio 지표에서는 cluster 1이 cluster 3보다 큰 값을 가진다. 그리고 F.undergrad와 Books 지표에서는 두 군집간의 차이가 있다고 보기 어렵다. 그 외의 모든 지표는 cluster 3이 cluster 2보다 큰 값을 가진다.

#cluster 2 vs cluster 3

	v1	v2	v3
Apps	6.304911e-20	3.152455e-20	1.000000e+00
Accept	8.170398e-24	4.085199e-24	1.000000e+00
Enroll	2.503837e-34	1.251919e-34	1.000000e+00
Top10perc	1.137471e-06	9.999994e-01	5.687356e-07
Top25perc	6.460040e-05	9.999677e-01	3.230020e-05
F.Undergrad	2.834424e-38	1.417212e-38	1.000000e+00
P.Undergrad	4.729789e-15	2.364895e-15	1.000000e+00
Outstate	1.769046e-40	1.000000e+00	8.845230e-41
Room.Board	8.032827e-13	1.000000e+00	4.016413e-13
Books	1.001898e-02	5.009490e-03	9.949905e-01
Personal	3.552026e-18	1.776013e-18	1.000000e+00
PhD	3.961200e-01	8.019400e-01	1.980600e-01
Terminal	2.860576e-01	8.569712e-01	1.430288e-01
S.F.Ratio	2.177290e-19	1.088645e-19	1.000000e+00
perc.alumni	2.701272e-37	1.000000e+00	1.350636e-37
Expend	2.225925e-14	1.000000e+00	1.112963e-14
Grad.Rate	7.964679e-21	1.000000e+00	3.982339e-21

위와 마찬가지로, cluster 2가 cluster 3보다 apps, accept, enroll, f.undergrad, p.undergrad, books, personal, s.f.ratio 지표에서 큰 값을 가진다. phD와 terminal은 두 군집간의 차이가 있다고 보기 어렵고, 그 외의 지표에서는 cluster 3이 cluster 2보다 큰 수치를 가진다.

[Q1-6]



다음과 같이 시각화를 할 수 있다.

[Hierarchical Clustering]

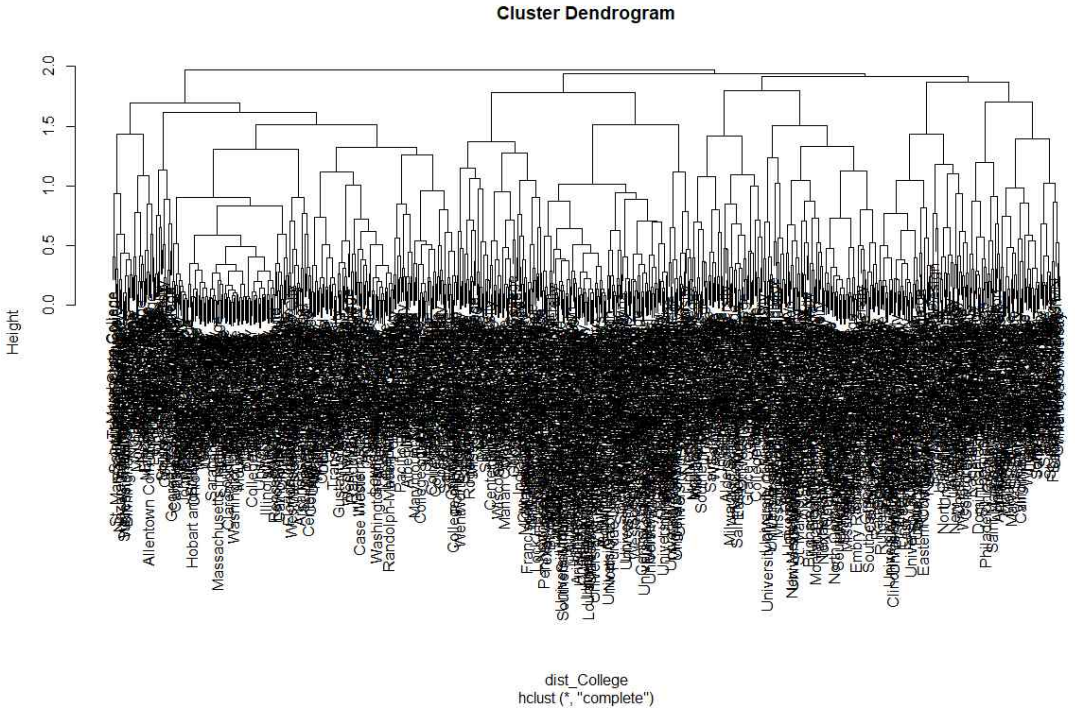
[Q2-1]

```
                                2      3      4      5      6      7      8      9     10
hierarchical APN           0.0003 0.0003 0.0042 0.0090 0.0176 0.0210 0.0251 0.0377 0.0729
             AD            5.3248 5.2914 5.2689 5.1870 5.1639 5.1406 5.1171 5.0968 5.0776
             ADM           0.0075 0.0074 0.0425 0.1195 0.1569 0.1938 0.2021 0.2913 0.4527
             FOM           0.9988 0.9944 0.9941 0.9812 0.9660 0.9658 0.9646 0.9575 0.9502
             Connectivity  2.9290 5.8579 8.7869 22.0956 25.0246 31.3833 34.3123 46.5571 46.5571
             Dunn          0.4033 0.4463 0.4393 0.1718 0.1718 0.1718 0.1718 0.1826 0.1826
             Silhouette    0.6777 0.6464 0.5802 0.4806 0.4291 0.3481 0.3015 0.2422 0.2125

Optimal Scores:

             Score  Method       Clusters
APN          0.0003 hierarchical 3
AD           5.0776 hierarchical 10
ADM          0.0074 hierarchical 3
FOM          0.9502 hierarchical 10
Connectivity 2.9290 hierarchical 2
Dunn         0.4463 hierarchical 3
Silhouette   0.6777 hierarchical 2

> end_time <- Sys.time()
> end_time - start_time
Time difference of 18.33196 secs
```

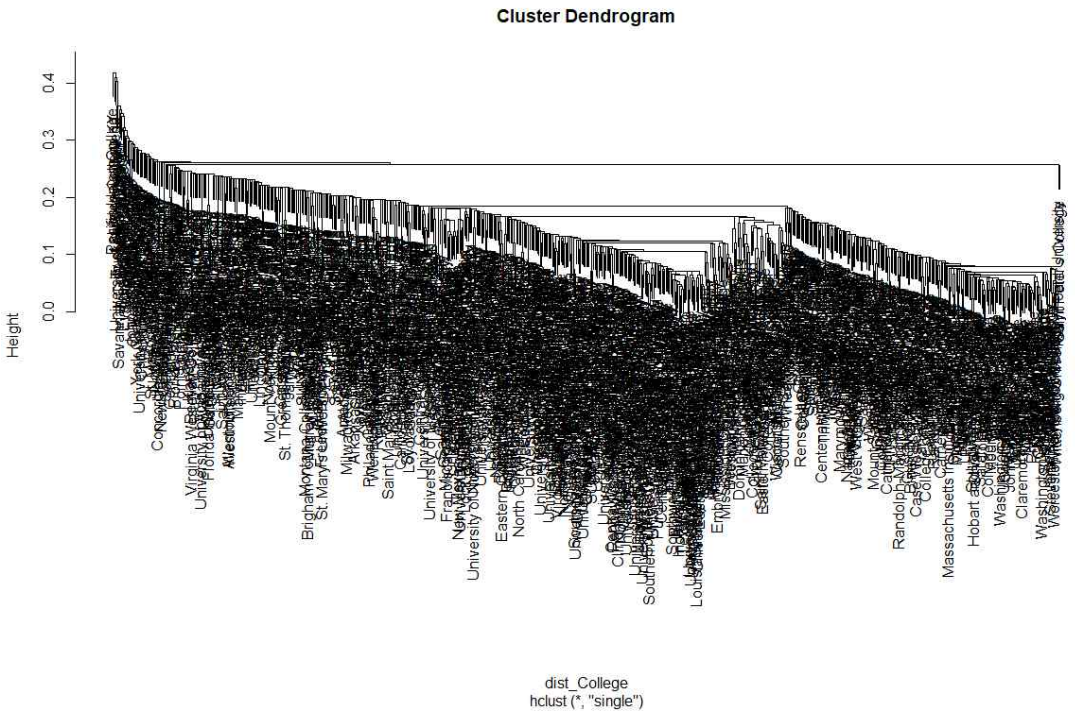
internal 및 stability 타당성 지표의 값들은 위와 같고, 소요시간은 18.33196초로 k-means clustering 보다 조금 덜 소요되었다. Dunn index 기준 최적의 군집수는 3개이고, silhouette index 기준으로는 2개가 최적이다.

[Q2-2]

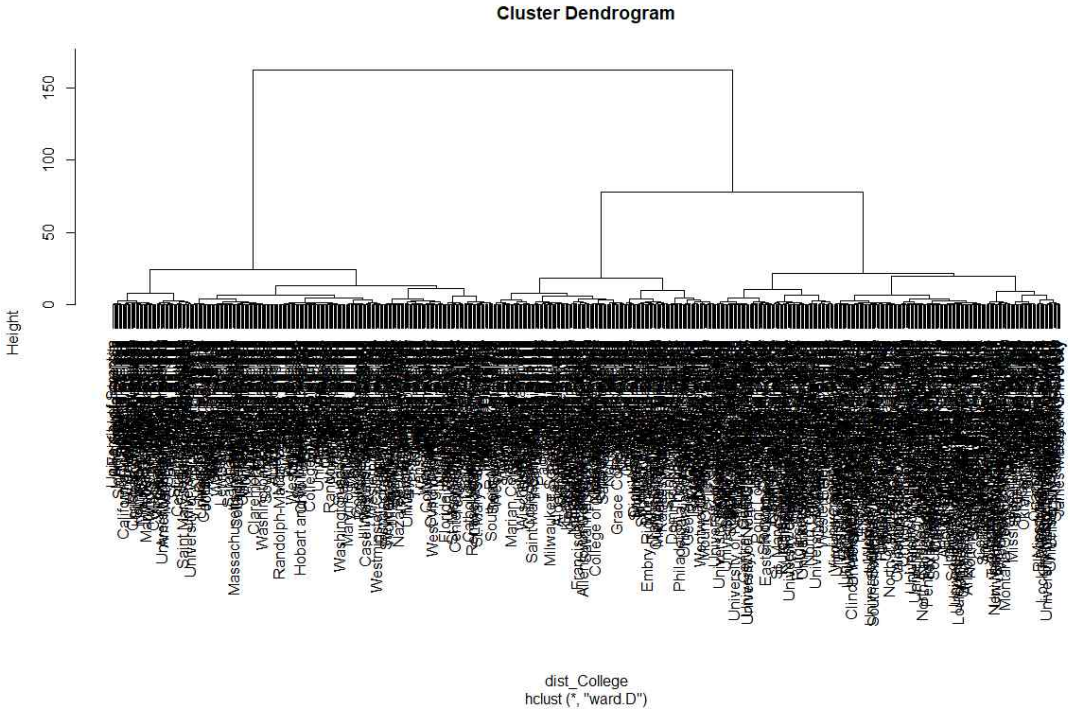
#complete linkage



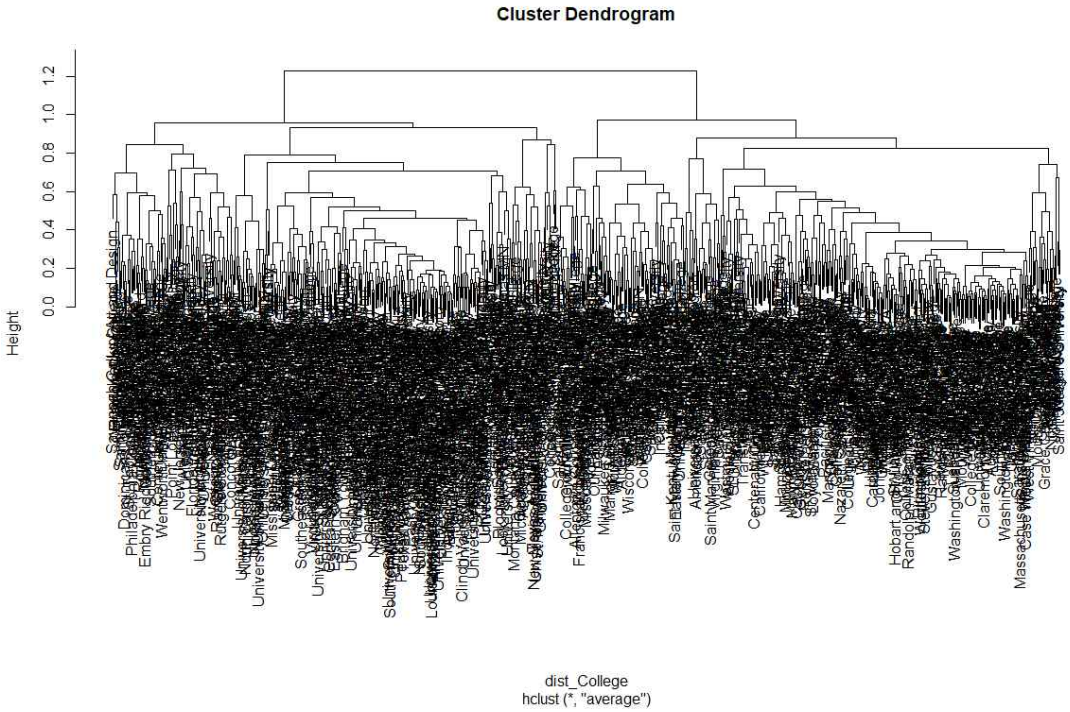
#single linkage



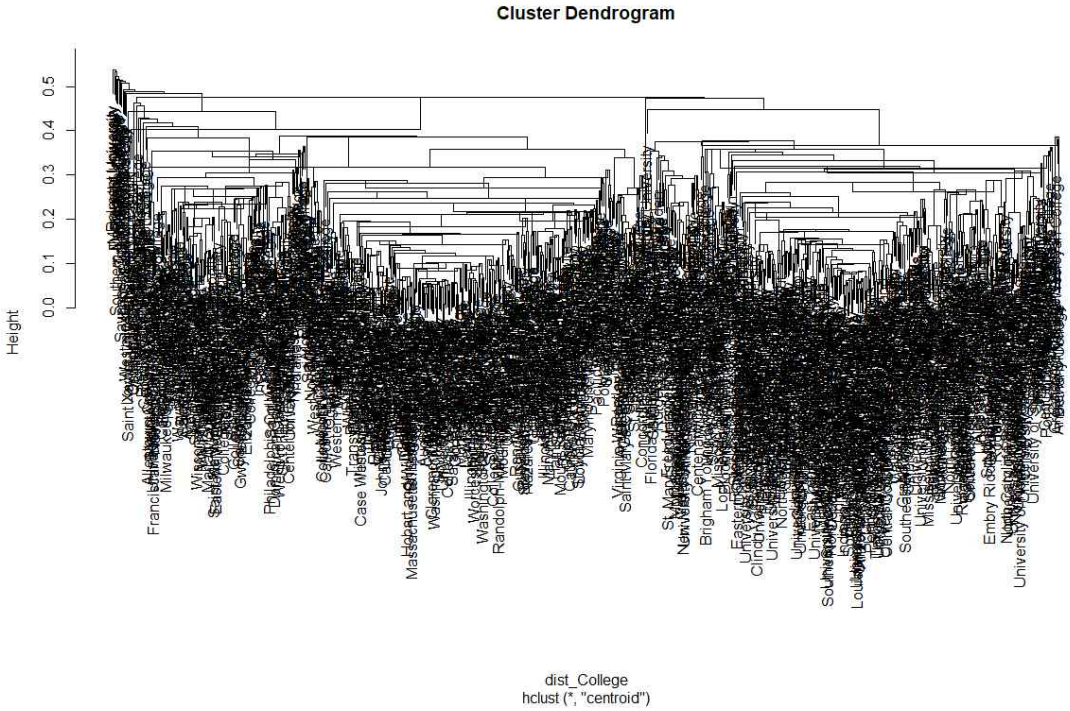
#ward method



#average linkage

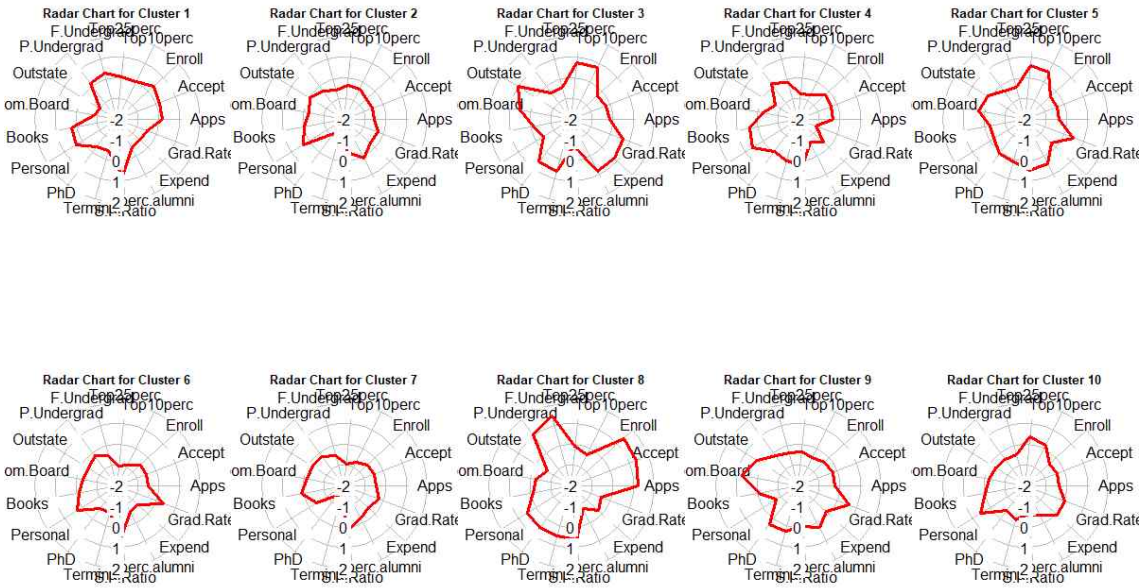


#centroid linkage



dunn index 기준에 따른 최적 군집의 개수를 2개라고 보았을 때, 각각에 방법론에 따른 dendrogram을 두 개의 군집으로 나누면 군집의 크기가 가장 극단적으로 차이가 날 것으로 예상되는 옵션은 single linkage와 centroid linkage가 가장 큰 차이를 보일 것이다.

[Q2-3]



가장 극명하게 차이가 나는 두 군집 조합은 cluster 3과 cluster 8로 보이고, 가장 유사한 두 군집 조합은 cluster 2와 cluster 7로 보인다.

#cluster 3 vs cluster 8

	V1	V2	V3
Apps	1.323863e-16	1.000000e+00	6.619313e-17
Accept	9.755304e-21	1.000000e+00	4.877652e-21
Enroll	4.162327e-26	1.000000e+00	2.081164e-26
Top10perc	7.636000e-27	3.818000e-27	1.000000e+00
Top25perc	7.983715e-16	3.991858e-16	1.000000e+00
F.Undergrad	6.102748e-30	1.000000e+00	3.051374e-30
P.Undergrad	5.681459e-16	1.000000e+00	2.840730e-16
Outstate	2.690637e-57	1.345319e-57	1.000000e+00
Room.Board	6.111057e-13	3.055528e-13	1.000000e+00
Books	7.662318e-01	6.168841e-01	3.831159e-01
Personal	7.868836e-19	1.000000e+00	3.934418e-19
PhD	3.445097e-01	1.722548e-01	8.277452e-01
Terminal	1.386780e-01	6.933899e-02	9.306610e-01
S.F.Ratio	6.504037e-29	1.000000e+00	3.252019e-29
perc.alumni	1.111755e-60	5.558776e-61	1.000000e+00
Expend	5.383138e-20	2.691569e-20	1.000000e+00
Grad.Rate	4.213407e-35	2.106703e-35	1.000000e+00

가장 극명하게 차이를 보이는 두 군집 조합에 대한 t test 결과는 다음과 같다. books와 phD 지표를 제외한 지표에서 극명한 차이를 보인다. top10perc, top25perc, outstate, room.board, terminal, perc.alumni, expend, grad.rate 지표에서는 cluster 3이 cluster 8보다 큰 수치를 가지고, 나머지 지표에서는 cluster 8이 더 큰 수치를 가진다.

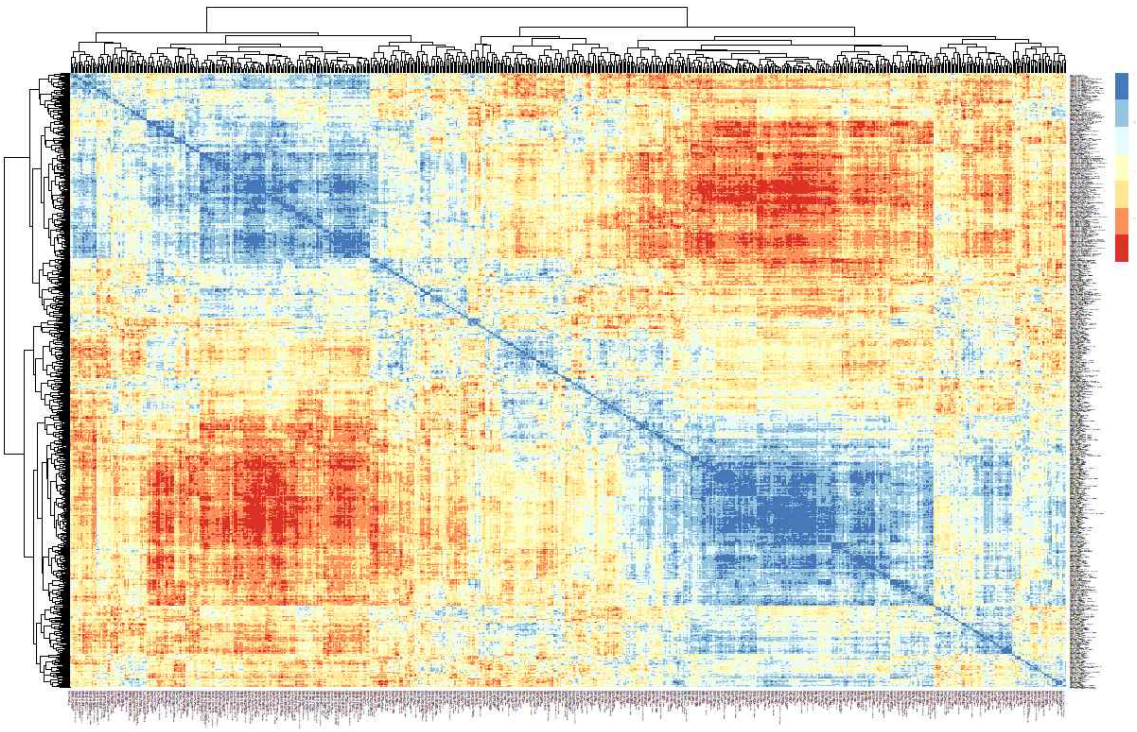
#cluster 2 vs cluster 7

	V1	V2	V3
Apps	2.856735e-02	9.857163e-01	0.014283673
Accept	3.479838e-02	9.826008e-01	0.017399190
Enroll	2.855483e-02	9.857226e-01	0.014277413
Top10perc	3.475639e-08	1.737819e-08	0.999999983
Top25perc	2.716152e-07	1.358076e-07	0.999999864
F.Undergrad	6.351390e-02	9.682430e-01	0.031756951
P.Undergrad	7.232383e-01	6.383809e-01	0.361619140
Outstate	8.996753e-02	4.498377e-02	0.955016234
Room.Board	5.675901e-01	7.162049e-01	0.283795052
Books	6.057908e-01	6.971046e-01	0.302895393
Personal	8.322624e-06	4.161312e-06	0.999995839
PhD	4.355805e-02	2.177903e-02	0.978220974
Terminal	7.156334e-01	3.578167e-01	0.642183324
S.F.Ratio	3.228996e-03	9.983855e-01	0.001614498
perc.alumni	1.630989e-02	8.154944e-03	0.991845056
Expend	1.363949e-01	6.819743e-02	0.931802571
Grad.Rate	7.203303e-01	6.398349e-01	0.360165127

가장 유사하게 보이는 두 군집 조합에 대한 t test 결과는 다음과 같다. 유의수준 0.01에서 top10perc와 top25perc 지표를 제외하고 두 군집간의 차이가 있다고 하기 어렵다.
top10perc와 top25perc에서는 cluster 2가 cluster7보다 더 큰 수치를 가진다.

[Q2-4]

pheatmap library를 사용하여 hierarchical clustering을 시각화 해본 결과는 다음과 같다.



크게 두 군집으로 나눌 수 있는 것을 볼 수 있다.

[DBSCAN]

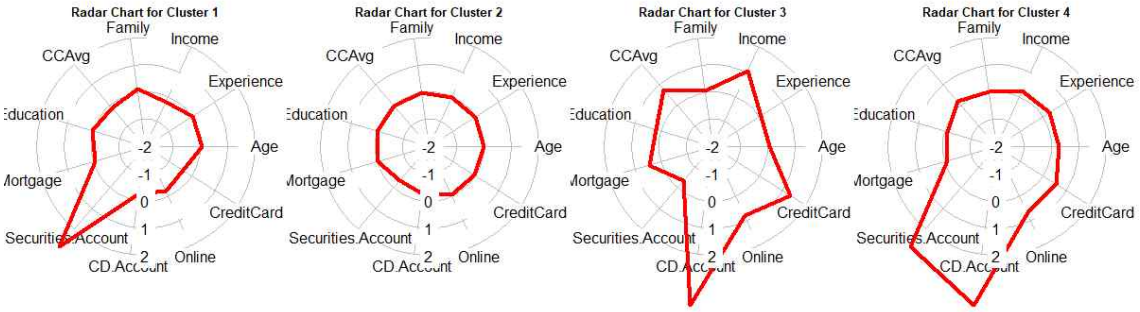
[Q3-1]

eps은 2.5부터 4.5까지, minPts는 4부터 8까지 바꾸면서 군집 수와 Noise 수를 구해보았다.

eps	minPts	군집 수	Noise 수
2.5	4	4	31
2.5	5	4	35
2.5	6	4	39
2.5	7	4	49
2.5	8	4	56
3	4	4	8
3	5	4	8
3	6	4	9
3	7	4	10
3	8	4	11
3.5	4	2	0
3.5	5	2	0
3.5	6	2	2
3.5	7	2	2
3.5	8	2	2
4	4	2	0
4	5	2	0
4	6	2	0
4	7	2	0
4	8	2	0
4.5	4	1	0
4.5	5	1	0
4.5	6	1	0
4.5	7	1	0
4.5	8	1	0

[Q3-2]

eps=3, minPts=5로 설정한 후 DBSCAN 방법을 사용하여 clustering을 진행한 결과를 radar chart를 통해 시각화하면 다음과 같은 결과를 얻는다.

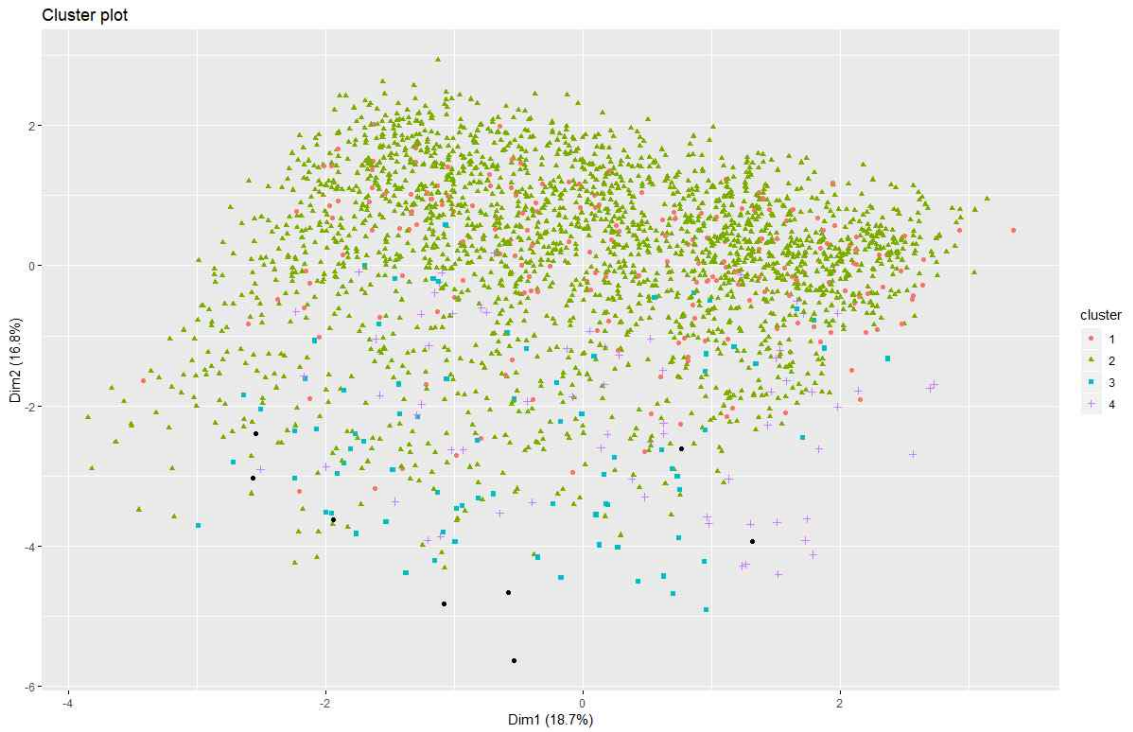


	noise	cluster 1	cluster 2	cluster 3	cluster 4
Age	0.06762434	0.05720942	-0.01542613	0.04700716	0.25592363
Experience	0.08714121	0.05013027	-0.01570137	0.06110819	0.26643729
Income	1.34025248	-0.20758320	-0.03242603	1.04484234	0.24343345
Family	-0.13622339	0.11614645	-0.01413241	0.06854278	0.04128289
CCAvg	1.16890859	-0.12215561	-0.02688594	0.76197700	0.17161597
Education	0.01123662	0.04052056	-0.00708426	0.19054441	-0.12060734
Mortgage	3.52483911	-0.05160341	-0.02107326	0.42402921	-0.09779151
Securities.Account	2.44964826	2.84970840	-0.35077273	-0.35077273	2.84970840
CD.Account	3.87551684	-0.25792689	-0.25792689	3.87551684	3.87551684
Online	0.31071614	-0.19386615	-0.03407728	0.79493584	0.67050787
CreditCard	0.46163560	-0.36290342	-0.03789669	1.37015032	0.59119207

cluster 1과 cluster 2는 대부분의 지표에서 비슷한 값을 보이지만 Securities.Account 지표에서 큰 차이를 보인다. Securities.Account와 CD.Account 지표는 0아니면 1의 값을 갖는 지표이다. 이에 따라 cluster 1에 속한 data들은 Securities.Account를 갖고 있는 군집이고 cluster2는 그렇지 않은 군집일 것이다. cluster 1,2 와 cluster 3,4 의 가장 큰 차이도 CD.Account를 갖고 있느냐 아니냐의 차이일 것이다. cluster 3,4는 CD.Account를 보유하고 있다. 또한 cluster 3은 다른 군집들과는 다르게 Income과 creditcard 지표의 값이 높은 것을 볼 수 있다. 마지막으로 cluster 4는 Securities.Account와 CD.Account를 모두 보유하고 있는 군집이다.

[Q3-3]

PCA를 이용하여 DBSCAN의 결과를 2차원에 시각화한 결과는 다음과 같다.



[Extra Question]

Extra Question에서는 수업시간에 다루지 않았던 fuzzy clustering에 대해서 알아보려고 한다. fuzzy clustering은 k-means clustering과 유사하지만 가장 큰 차이점은 k-mean clustering의 경우에는 data point가 하나의 cluster에만 속해야 하지만 fuzzy clustering의 경우에는 하나 이상의 cluster에 속하는 것이 가능하다.

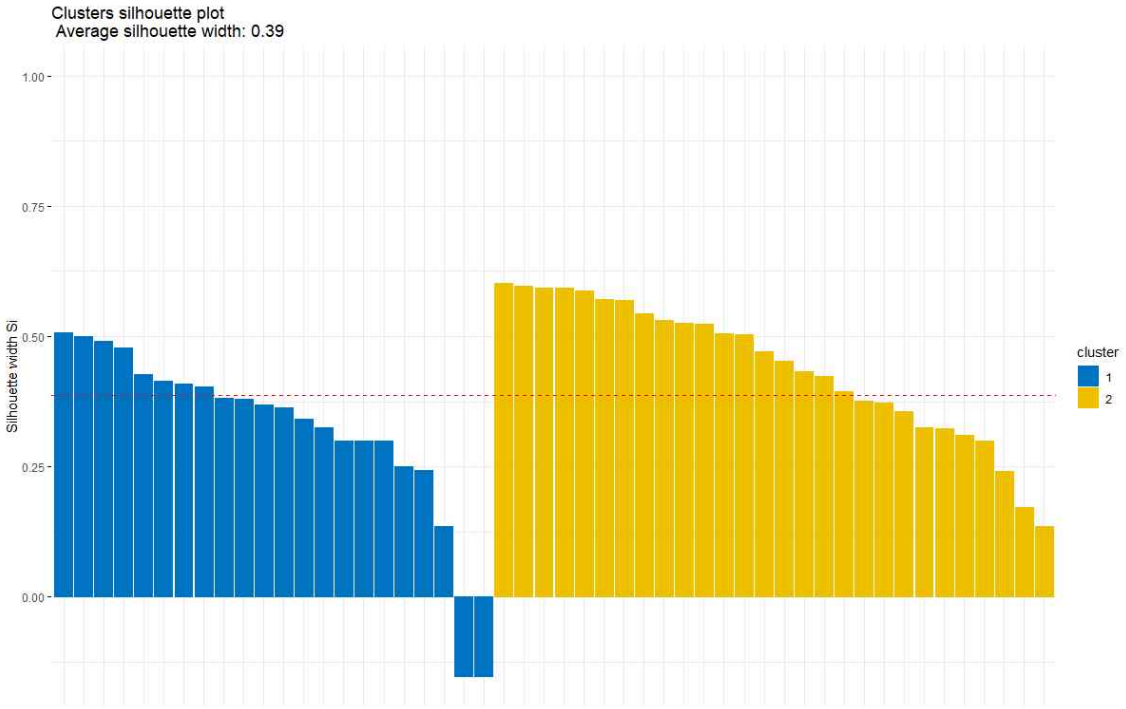
	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7
Connecticut	3.3	110	77	11.1
Delaware	5.9	238	72	15.8
Florida	15.4	335	80	31.9
Georgia	17.4	211	60	25.8
Hawaii	5.3	46	83	20.2
Idaho	2.6	120	54	14.2
Illinois	10.4	249	83	24.0
Indiana	7.2	113	65	21.0
Iowa	2.2	56	57	11.3
Kansas	6.0	115	66	18.0
Kentucky	9.7	109	52	16.3
Louisiana	15.4	249	66	22.2
Maine	2.1	83	51	7.8
Maryland	11.3	300	67	27.8
Massachusetts	4.4	149	85	16.3
Michigan	12.1	255	74	35.1
Minnesota	2.7	72	66	14.9
Mississippi	16.1	259	44	17.1

R에서 제공하는 USArrests의 데이터를 fuzzy clustering을 통해 분석하였다. 이 data set은 1973년 미국의 각 주에 따른 인구 100,000명당 assault, murder, rape로 체포된 사람의 수를 보여준다. 또한 도시에 사는 인구 비율도 보여준다.

다음과 같은 data를 scale을 통해 표준화 해준 뒤 fanny 함수를 통해 fuzzy clustering을 사용해보았다.

두 개의 군집으로 나누었을 때 다음과 같은 시각화가 가능하다.

k-means clustering과 다르게 몇 개의 data point가 두 군집에 모두 속해있는 것을 볼 수 있다.



cluster silhouette값이 1에 가까울수록 clustering이 잘 되었다고 볼 수 있는데, 이 데이터 set의 경우 군집의 개수를 2개로 했을 때 가장 높은 값을 보인다.