Multivariate Data Analysis

Assignment #2

Clustering: College Dataset

과목명	다변량분석
담당교수	강필성 교수님
제출일	2019-04-16
이름	박지원
학과명	산업경영공학부
학번	2014170856

목차

Dataset: College Dataset

해당 데이터셋은 미국 각 대학들의 신입생 선발에 대한 정보를 포함하는 데이터셋으로, 총 18개의 변수가 있습니다. 이 중에서 첫 번째 column인 Private을 제외한 총 17개의 수치형 변수에 대해서 군집화를 수행하고 그 결과를 해석했습니다.

1. K-Means Clustering

K-Means clustering은 각 변수의 값의 범위에 영향을 받을 수 있으므로 scale( )함수를 사용하여 모든 변수의 평균을 0, 표준편차를 1로 만드는 정규화를 수행했습니다.

```
#Standardization
College_x_scaled <- scale(College_x, center = TRUE, scale = TRUE)
```

[Q1-1] clValid( ) 함수를 사용하여 K-Means Clustering의 군집 수를 2개부터 10개까지 증가시켜 가면서 internal 및 stability 관련 타당성 지표 값들을 산출했습니다.

```
> summary(College_cvalid)

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
```

Dunn index와 Silhouette index 기준으로 가장 최적의 군집 수는 각각 10개와 2개로 판별되었습니다.

총 소요 시간은 22.2초였습니다.

```
> toc()
22.2 sec elapsed
```

[Q1-2] K=3으로 군집화를 10회 반복 수행하고 회차마다 각 군집의 Center와 Size를 확인해보았습니다. 다음은 첫번째와 두번째 반복 수행시의 center입니다. 순서와 값이 모두 같은 것을 알 수 있습니다. 같은 방식으로 나머지를 모두 비교했을 때, 순서는 다르더라도 값이 모두 같음을 발견했습니다.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
1	-0.03489665	-0.1165637	-0.2330693	0.8552988	0.8370369	-0.3018808	-0.3693397	1.0480076	0.7172261	0.0590233
2	-0.37097263	-0.3607812	-0.3367498	-0.5291736	-0.5548376	-0.3116495	-0.1277753	-0.4746087	-0.3668251	-0.1024233
3	1.83946866	2.0074927	2.2024888	0.2298334	0.3990083	2.2662920	1.5787435	-0.5368952	-0.1695508	0.3262545

	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
1	-0.37425038	0.7370637	0.7356034	-0.6350344	0.8164997	0.80516666	0.7674493
2	0.03786354	-0.5604491	-0.5571596	0.2325641	-0.3323420	-0.43979658	-0.3538001
3	0.81162753	0.6898820	0.6782520	0.5844663	-0.5945498	-0.05849565	-0.3637427

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
1	-0.03489665	-0.1165637	-0.2330693	0.8552988	0.8370369	-0.3018808	-0.3693397	1.0480076	0.7172261	0.0590233
2	-0.37097263	-0.3607812	-0.3367498	-0.5291736	-0.5548376	-0.3116495	-0.1277753	-0.4746087	-0.3668251	-0.1024233
3	1.83946866	2.0074927	2.2024888	0.2298334	0.3990083	2.2662920	1.5787435	-0.5368952	-0.1695508	0.3262545

	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
1	-0.37425038	0.7370637	0.7356034	-0.6350344	0.8164997	0.80516666	0.7674493
2	0.03786354	-0.5604491	-0.5571596	0.2325641	-0.3323420	-0.43979658	-0.3538001
3	0.81162753	0.6898820	0.6782520	0.5844663	-0.5945498	-0.05849565	-0.3637427

다음은 10번 반복 시 군집의 size입니다.

```
> cluster_size
   X93L X246L X438L
1    93   246   438
2   438    93   246
3   246    93   438
4   438    93   246
5    93   246   438
6   438    93   246
7   246   438    93
8    93   438   246
9   246   438    93
10   93   438   246
```

사이즈 역시 순서는 다르더라도 크기는 모두 동일했습니다. 이로부터 10번의 수행 모두 같은 군집이 생성되었음을 알 수 있습니다.

[Q1-3] K=10으로 군집화를 10회 반복 수행하고 회차마다 각 군집의 Center와 Size를 확인했습니다. 아래는 첫 번째 수행시의 center입니다.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
1	1.20715934	1.4823270	1.7983330	-0.24396642	-0.131310506	1.96615548	2.1587079	-0.6957375	-0.37620731	0.3290319	1.19862366	0.65486332
2	0.36729054	0.3814069	0.5689110	-0.52029973	-0.205216111	0.65822833	0.4308507	-1.0281841	-0.70082914	-0.1518491	0.46227261	0.17952535
3	3.47713512	3.5563767	3.4797343	1.05590940	1.188434428	3.42764096	1.0856640	-0.2799269	0.04595578	0.3153350	0.41391004	0.89183321
4	0.06066662	0.1105282	-0.0363445	0.31564108	0.450002880	-0.07379937	-0.0135076	0.3587180	0.67510480	0.5549893	0.54750434	0.64875261
5	1.34411723	0.4622458	0.4300183	2.87562970	1.954242910	0.21052368	-0.3342481	1.9402131	1.53152363	0.6055296	-0.12364078	1.33099091
6	-0.55500321	-0.5463893	-0.5668278	-0.53969354	-0.643056204	-0.53404177	-0.3110287	-0.2977119	-0.42328342	-0.2464263	0.02931203	-1.36582813
7	-0.15882950	-0.1859754	-0.3383168	1.06651889	1.068469106	-0.40196403	-0.4724319	1.4213377	0.71034922	-0.1574629	-0.75692643	0.91536071
8	-0.47535809	-0.4586877	-0.3983829	-0.79455194	-0.937115581	-0.39373498	-0.1363075	-0.8766652	-0.66152535	-0.1969442	-0.05622146	-0.44163466
9	-0.29704871	-0.3257244	-0.3316267	-0.61271742	-0.599854609	-0.33555317	-0.1917756	-0.2010728	0.28507459	4.2005645	0.58146170	-1.78078288
10	-0.46181395	-0.4414627	-0.4709932	-0.07844848	0.002967359	-0.49028617	-0.3375448	0.2688609	0.20966923	-0.3337131	-0.46957731	0.04589661

	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
1	0.62884500	0.6634034	-0.7072010	-0.2408122	-0.705005461
2	0.14264795	1.0227304	-0.7374435	-0.6238703	-0.589443666
3	0.86295308	0.4367514	-0.4127130	0.3035113	0.246853816
4	0.58765908	-0.2036606	-0.4012116	0.2016279	0.180833173
5	1.17975159	-1.7216026	1.3031980	3.4879823	1.528199359
6	-1.44198501	-0.3395108	-0.2269718	-0.3956769	-0.304271016
7	0.95378591	-0.6754701	1.2259947	0.8873245	0.826434611
8	-0.45642496	0.8097344	-0.6974121	-0.6282666	-0.845435310
9	-0.69866768	0.1541794	-0.5711212	-0.1856978	0.006986547
10	0.09957453	-0.2612913	0.5928496	-0.1230223	0.457334342

다음은 두 번째 수행시의 center입니다.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
1	-0.1449052	-0.07806567	-0.2281768	-0.3009325	-0.2657967	-0.2682544	-0.09995785	0.5322267	1.24934076	-0.02903598	-0.25056895	0.36830135
2	3.4771351	3.55637666	3.4797343	1.0559094	1.1884344	3.4276410	1.08566405	-0.2799269	0.04595578	0.31533496	0.41391004	0.89183321
3	1.4295074	0.47246911	0.4389555	2.9319939	1.9714105	0.2165199	-0.32191823	1.9706235	1.65485506	0.46575743	-0.14958866	1.33632788
4	-0.4355107	-0.42248386	-0.4400716	0.2533410	0.3995849	-0.4541072	-0.36969570	0.1127357	-0.17177812	-0.18221335	-0.20472673	0.08712111
5	0.3379908	0.34711720	0.5083785	-0.5033383	-0.1897619	0.5906657	0.39408702	-0.9952291	-0.69959591	-0.12856703	0.46413893	0.21396260
6	-0.3135257	-0.30904180	-0.3161319	-0.3325470	-0.3082259	-0.3256941	-0.19770401	-0.1364782	0.07990129	4.11068405	1.54666680	-1.15224765
7	-0.1046546	-0.13448822	-0.2800551	1.0757201	1.0491208	-0.3576084	-0.47026869	1.4416905	0.69076794	0.02137992	-0.65188819	0.94674386
8	1.1813493	1.43387533	1.7487479	-0.2284679	-0.1326368	1.9128468	2.16854111	-0.6950031	-0.33414070	0.35218329	1.17105231	0.67715855
9	-0.4933963	-0.48521298	-0.4329283	-0.7978538	-0.9739426	-0.4213991	-0.16216026	-0.8733799	-0.73422762	-0.13944919	0.07755249	-0.59794915
10	-0.5581275	-0.54465419	-0.5599744	-0.5408882	-0.6176566	-0.5251147	-0.28747828	-0.2251857	-0.33042192	-0.32248620	-0.14533606	-1.38361352

	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
1	0.4041113	-0.1827529	0.01540501	0.03263565	0.4279905
2	0.8629531	0.4367514	-0.41271296	0.30351133	0.2468538
3	1.2075034	-1.7850128	1.31184423	3.74605444	1.5727754
4	0.1120080	-0.2256376	0.42583017	-0.15791165	0.3684882
5	0.1717161	1.0218683	-0.68285757	-0.59635792	-0.5434435
6	-0.4134744	-0.1295441	-0.66216564	-0.12040167	-0.1613144
7	0.9543153	-0.7442849	1.21196983	0.96125340	0.7863083
8	0.6456061	0.6017645	-0.71018653	-0.20610408	-0.7178632
9	-0.6240053	0.6482125	-0.76069863	-0.58567789	-0.9755067
10	-1.4265274	-0.2910550	-0.10141441	-0.39519641	-0.1090702

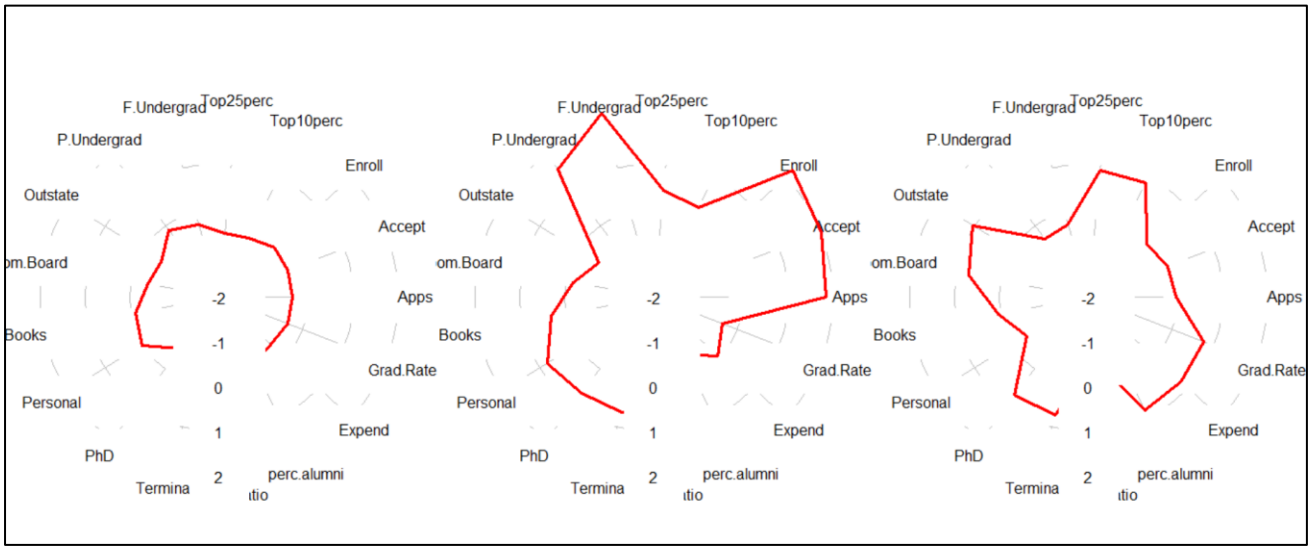
비교 결과, K=3일 때와 다르게 모두 다른 군집이 생성되었음을 파악했습니다.

다음은 각 수행시의 군집 size입니다.

```
> cluster_size
   X51L X71L X27L X79L X28L X131L X94L X118L X12L X166L
1    51   71   27   79   28   131   94   118   12   166
2    92   27   25  145   78    13  101    53  126   117
3   159  119   35  125   13    97   25    25   53   126
4    67  108  125   29   24    87   94    53   87   103
5   100   88   88   27   24    13  138   101  144    54
6    25   11  105  147   85    27   43    99  200    35
7    52  131   79  149   24    96   19    91   13   123
8   177  144   54  114   13    90   27    10   29   119
9    24  136  125   89   78    16   29    60   65   155
10   55   51  107   89   20    63  105   127   27   133
```

매 수행시마다 다른 군집이 생성되었습니다.

[Q1-4] K=3으로 군집화를 수행한 뒤, 각 변수들에 대해 정규화 이후의 값들을 이용하여 Radar chart를 도시했습니다. 결과는 다음과 같습니다.



첫번째와 두번째 군집은 P.Undergrad, F.Undergrad, Enroll, Apps, Personal 변수의 값이 다른 변수들의 값에 비해 상대적으로 크기 때문에 더 유사하다고 생각합니다. 반면, 두번째와 세번째 군집은 가장 다르다고 생각했습니다. 두번째 군집은 P.Undergrad, F.Undergrad, Enroll, Apps, Personal의 값이 높은 반면, 세번째

군집에서는 이들의 값이 모두 상대적으로 낮게 나와 서로 반대의 양상을 띄었기 때문입니다.

[Q1-5] 다음은 세 개의 두 군집 조합(Cluster 1 vs. Cluster 2, Cluster 1 vs. Cluster 3, Cluster 2 vs. Cluster 3)에 대해서 각 변수별 차이의 유의성에 대해서 t-test를 수행한 결과입니다.

(1) Cluster 1 vs. Cluster 2

```
> kmc_t_result
              V1        V2           V3
1  3.058227e-24 1.0000000 1.529113e-24
2  8.506840e-27 1.0000000 4.253420e-27
3  1.835458e-35 1.0000000 9.177291e-36
4  2.272173e-10 1.0000000 1.136087e-10
5  4.337600e-16 1.0000000 2.168800e-16
6  2.236426e-38 1.0000000 1.118213e-38
7  1.399637e-12 1.0000000 6.998186e-13
8  4.480168e-01 0.2240084 7.759916e-01
9  6.579076e-02 0.9671046 3.289538e-02
10 4.343236e-06 0.9999978 2.171618e-06
11 4.085254e-10 1.0000000 2.042627e-10
12 1.758412e-60 1.0000000 8.792060e-61
13 6.774873e-57 1.0000000 3.387436e-57
```

각각 cluster 1 = cluster 2, cluster 1>cluster2, cluster 1< cluster2에 대해 t-test를 수행했습니다. 1열을 보면, V1에 대해서는 3.058227e-24라는 매우 작은 값이 나왔습니다. 이는 cluster1과 cluster2에서 이 변수가 동일하지 않음을 나타냅니다. 1행 2열을 보면 1이라는 매우 큰 값이 도출되었습니다 p-value가 0.05보다 훨씬 크기 때문에 cluster 1>cluster2라는 귀무가설을 기각할 수 없습니다. 따라서 이 변수는 cluster1에서 큰 값을 가지고 cluster2에서 작은 값을 가져, 두 군집을 구별하는 유의미한 변수라는 것을 알 수 있습니다.

(2) Cluster 1 vs. Cluster 3

```
> kmc_t_result
              V1           V2           V3
1  8.982960e-12 1.000000e+00 4.491480e-12
2  1.686494e-10 1.000000e+00 8.432471e-11
3  2.082504e-03 9.989587e-01 1.041252e-03
4  4.705503e-57 1.000000e+00 2.352752e-57
5  7.271735e-82 1.000000e+00 3.635867e-82
6  7.440255e-01 6.279872e-01 3.720128e-01
7  8.973497e-15 4.486749e-15 1.000000e+00
8  1.023578e-92 1.000000e+00 5.117892e-93
9  1.833853e-44 1.000000e+00 9.169264e-45
10 5.283634e-02 9.735818e-01 2.641817e-02
11 2.231088e-08 1.115544e-08 1.000000e+00
12 3.644528e-86 1.000000e+00 1.822264e-86
13 2.591698e-90 1.000000e+00 1.295849e-90
```

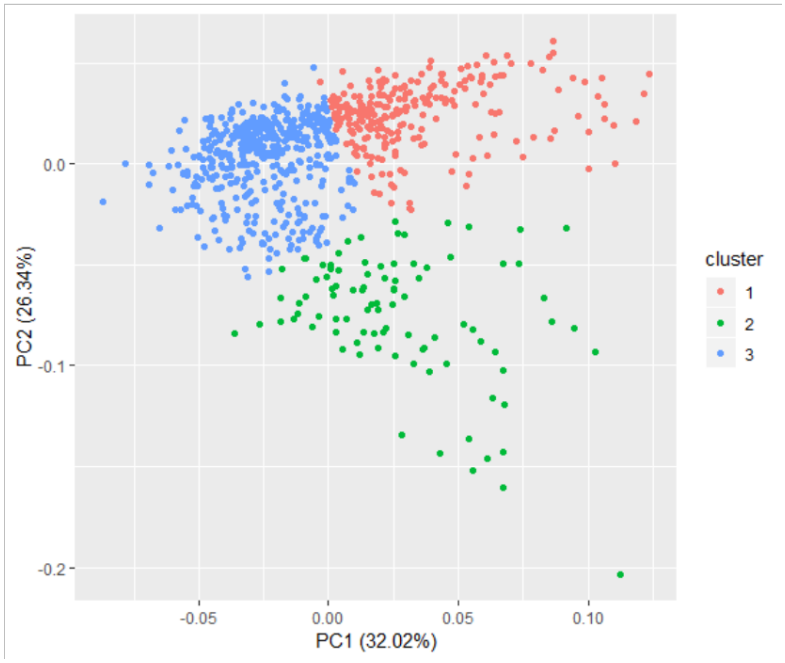
위와 마찬가지로 cluster 1 = cluster 3, cluster 1>cluster 3, cluster 1< cluster 3에 대해 t-test를 수행했습니다. 1열을 보면, V1에 대해서는 8.982960e-12라는 매우 작은 값이 나왔습니다. 이는 cluster1과 cluster3에서 이 변수가 동일하지 않음을 나타냅니다. 1행 2열을 보면 1이라는 매우 큰 값이 도출되었습니다 p-value가 0.05보다 훨씬 크기 때문에 cluster 1>cluster3라는 귀무가설을 기각할 수 없습니다. 따라서 이 변수는 cluster1에서 큰 값을 가지고 cluster3에서 작은 값을 가져, 두 군집을 구별하는 유의미한 변수라는 것을 알 수 있습니다.

(3) Cluster 2 vs. Cluster 3

```
> kmc_t_result
             V1           V2           V3
1  6.304911e-20 3.152455e-20 1.000000e+00
2  8.170398e-24 4.085199e-24 1.000000e+00
3  2.503837e-34 1.251919e-34 1.000000e+00
4  1.137471e-06 9.999994e-01 5.687356e-07
5  6.460040e-05 9.999677e-01 3.230020e-05
6  2.834424e-38 1.417212e-38 1.000000e+00
7  4.729789e-15 2.364895e-15 1.000000e+00
8  1.769046e-40 1.000000e+00 8.845230e-41
9  8.032827e-13 1.000000e+00 4.016413e-13
10 1.001898e-02 5.009490e-03 9.949905e-01
11 3.552026e-18 1.776013e-18 1.000000e+00
12 3.961200e-01 8.019400e-01 1.980600e-01
13 2.860576e-01 8.569712e-01 1.430288e-01
```

마지막으로 cluster 2와 cluster 3을 비교해보았습니다. 1열을 보면, cluster2 = cluster 3에 대해 6.304911e-20라는 매우 작은 p-value가 나왔습니다. 이는 cluster2과 cluster3에서 이 변수가 동일하지 않음을 나타냅니다. 반면, 3열의 cluster 2 < cluster 3을 보면 1이라는 매우 큰 값이 도출되었습니다 p-value가 0.05보다 훨씬 크기 때문에 cluster 2 < cluster 3이라는 귀무가설을 기각할 수 없습니다. 따라서 이 변수는 cluster3에서 큰 값을 가지고 cluster2에서 작은 값을 가져, 두 군집을 구별하는 유의미한 변수라는 것을 알 수 있습니다.

[Q1-6] K-Means Clustering의 결과물을 시각화하기 위해 ggfortify 패키지를 사용했습니다. K-means clustering을 수행한 후, PCA를 통해 가장 영향을 많이 끼치는 두 축을 찾아 이를 기준으로 다음과 같이 cluster를 시각화 했습니다. 1번 군집과 3번 군집은 PC1이라는 축에 의해 잘 구분이 되는 것으로 보이며, 2번 군집은 다른 군집들과 PC2라는 축에 의해 잘 구분이 되는 것으로 보입니다.



2. Hierarchical Clustering

[Q2-1] clValid( ) 함수를 사용하여 Hierarchical Clustering의 군집 수를 2개부터 10개까지 증가시켜 가면 서 internal 및 stability 관련 타당성 지표 값들을 산출했습니다. 그 결과는 다음과 같습니다.

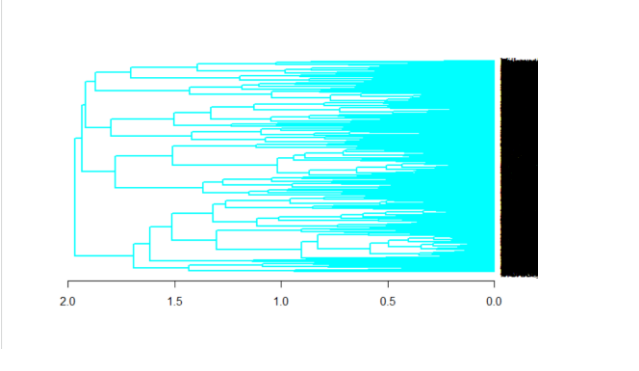
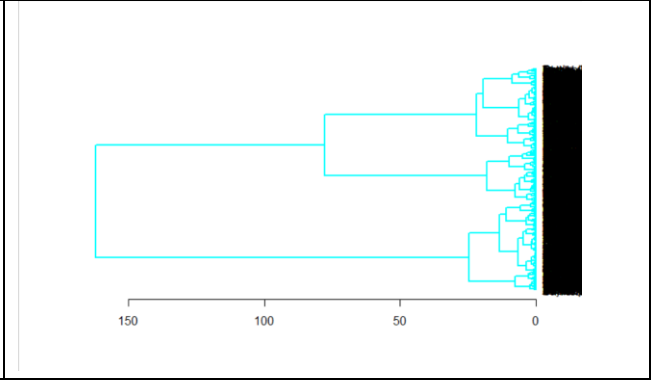
```
> summary(College_cvalid)

Clustering Methods:
 hierarchical

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

Validation Measures:
                                 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
```

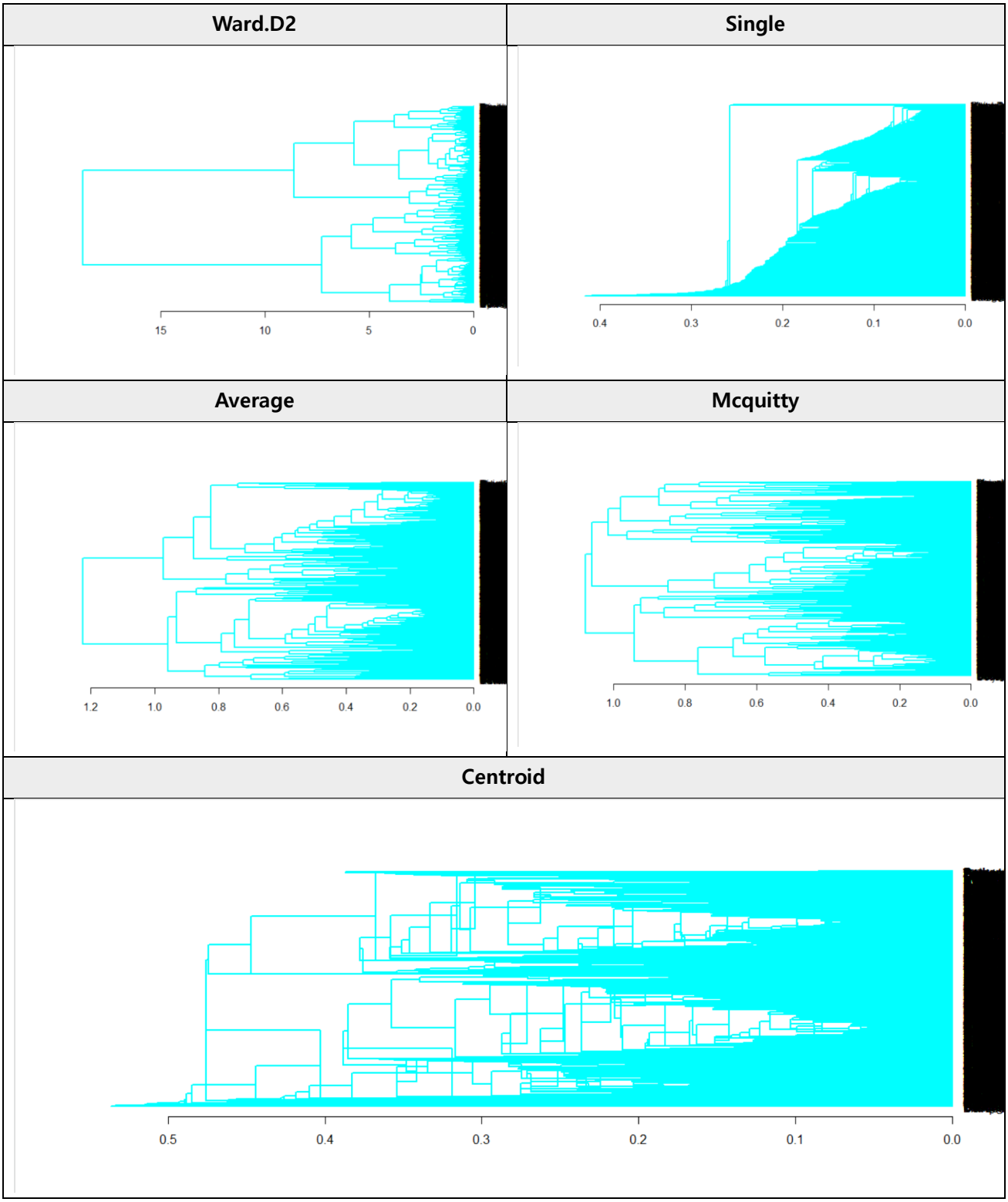
Dunn index와 Silhouette index 기준으로 가장 최적의 군집 수는 각각 3개와 2개로 판별되었습니다. 위에서 KMC를 수행할 때는 Dunn index 기준 10개가 나왔는데, 여기에서는 3개로 줄어들었습니다.

총 소요 시간은 25.36초입니다. K-Means Clustering과 비교할 경우 약간 더 오래 걸렸습니다.

```
> toc()
25.36 sec elapsed
```

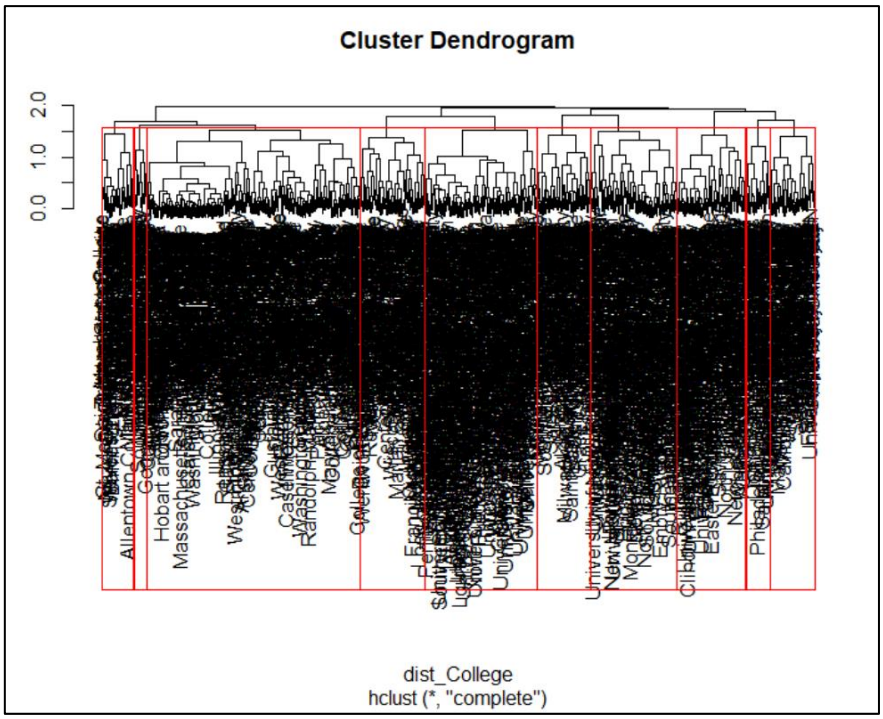
[Q2-2] Spearman correlation을 이용한 방식으로 distance matrix를 산출한 뒤, hclust( ) 함수의 method 옵션을 다양하게 조절해 가면서 dendrogram을 그려보았습니다. 사용한 method는 complete, ward.D, ward.D2, single, average, mcquitty, 그리고 centroid입니다. 각 method별 덴드로그램은 다음과 같습니다.

Complete	Ward.D
	

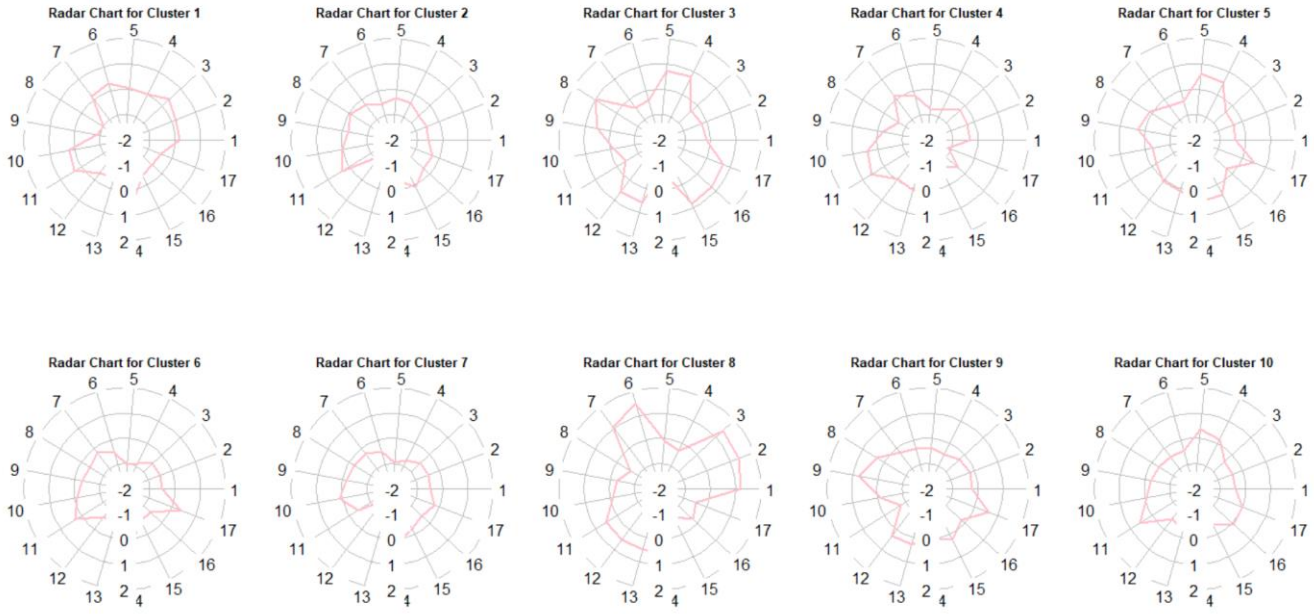


Single을 제외한 다른 method들의 경우 군집마다 비교적 고르게 개수가 할당된 반면, single의 경우 한쪽으로 지나치게 쏠려있는 모습이 발견되었습니다. 따라서 군집화를 수행했을 때, 군집의 크기가 가장 극단적으로 차이 날 것으로 예상합니다.

[Q2-3] Complete method를 사용해 생성된 Dendrogram으로부터 10개의 군집을 찾았습니다.



각 변수들에 대한 정규화 이후의 값들을 이용하여 Radar chart를 도시했습니다.



Cluster 3과 Cluster8은 1~10번 변수에서 반대의 양상을 보여 가장 차이가 극명하게 나타난다고 생각했습니다. 그리고, Cluster 1과 Cluster8은 12, 13번 변수를 제외한 나머지 변수들에 대해 비슷한 모습을 보여 가장 유사하다고 생각했습니다. 이렇게 두 가지 조합에 대해 변수별 차이의 통계적 유의성을 t-test를 통해 검증했습니다.

우선, Cluster 3과 Cluster 8을 비교한 결과입니다. 각 열은 cluster 3 = cluster 8, cluster 3 > cluster 8, cluster 3 < cluster 8에 대한 p-value를 나타냅니다.

```
> hc_t_result
              V1           V2           V3
1   1.323863e-16 1.000000e+00 6.619313e-17
2   9.755304e-21 1.000000e+00 4.877652e-21
3   4.162327e-26 1.000000e+00 2.081164e-26
4   7.636000e-27 3.818000e-27 1.000000e+00
5   7.983715e-16 3.991858e-16 1.000000e+00
6   6.102748e-30 1.000000e+00 3.051374e-30
7   5.681459e-16 1.000000e+00 2.840730e-16
8   2.690637e-57 1.345319e-57 1.000000e+00
9   6.111057e-13 3.055528e-13 1.000000e+00
10 7.662318e-01 6.168841e-01 3.831159e-01
11 7.868836e-19 1.000000e+00 3.934418e-19
```

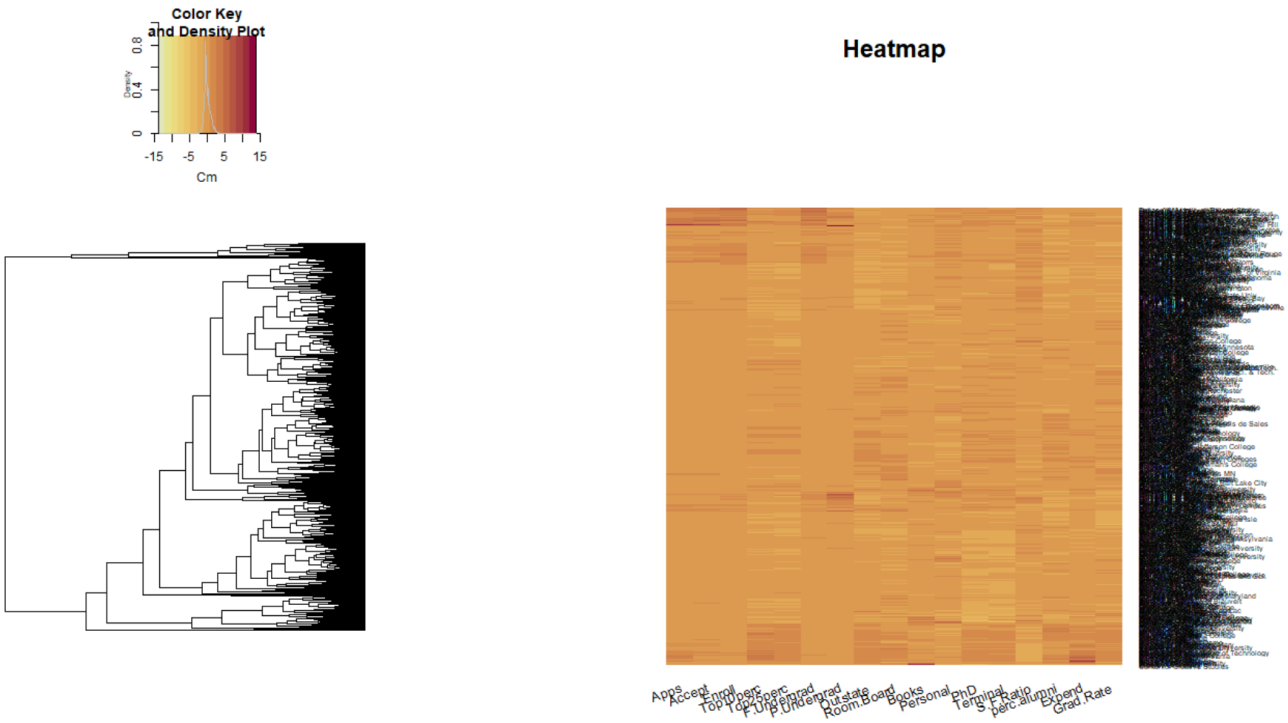
1행 1열을 살펴보면 1.323863e-16이라는 매우 작은 p-value가 계산되었습니다. 이는 cluster 3 = cluster 8이라는 귀무가설을 기각할 수 없음을 의미합니다. 2열을 살펴보면 1이라는 매우 큰 p-value가 계산되었습니다. 따라서 cluster 3 > cluster 8이라는 귀무가설을 기각할 수 있습니다. 위의 radar chart에서도 변수 1은 cluster 3에서는 매우 작은 값을 가지지만 cluster 8에서는 매우 큰 값을 가지는 것을 알 수 있습니다. 따라서 변수1은 두 군집을 구분하는 유의미한 변수입니다. 이와 같은 방법으로 살펴보았을 때, 10번 변수를 제외한 나머지 변수들의 차이가 통계적으로 유의미하다는 것을 알 수 있습니다.

다음으로, Cluster 1과 Cluster 8을 비교한 결과입니다. 각 열은 cluster 1 = cluster 8, cluster 1 > cluster 8, cluster 1 < cluster 8에 대한 p-value를 나타냅니다.

```
> hc_t_result
              V1           V2           V3
1   2.675866e-08 0.999999987 1.337933e-08
2   4.611044e-11 1.000000000 2.305522e-11
3   6.472266e-09 0.999999997 3.236133e-09
4   1.169724e-02 0.005848621 9.941514e-01
5   2.586032e-01 0.129301619 8.706984e-01
6   8.998310e-12 1.000000000 4.499155e-12
7   2.973351e-06 0.999998513 1.486675e-06
8   4.698747e-05 0.999976506 2.349373e-05
9   3.405370e-08 0.999999983 1.702685e-08
10 1.577810e-02 0.007889050 9.921110e-01
11 1.851574e-01 0.907421303 9.257870e-02
```

1행 1열을 살펴보면 2.675866e-08이라는 매우 작은 p-value가 계산되었습니다. 이는 cluster 1 = cluster 8이라는 귀무가설을 기각할 수 없음을 의미합니다. 2열을 살펴보면 거의 1에 가까운 p-value가 계산되었습니다. 따라서 유의수준 0.05 하에서 cluster 1 > cluster 8이라는 귀무가설을 기각할 수 있습니다. 이와 같은 방법으로 살펴보았을 때, 4번, 10번 변수를 제외한 나머지 변수들의 차이가 통계적으로 유의미하다는 것을 알 수 있습니다. Radar chart로 보았을 때에는 유사하다고 생각했으나 실제 통계적 수치를 비교해보니 크게 유사하지 않음을 알게되었습니다.

[Q2-4] Hierarchical Clustering의 결과물을 heatmap을 사용하여 도시하였습니다.



Heatmap으로부터 아래쪽 군집은 Top10perc, Top25perc, Outstate, Room.Board, Expend등의 값이 크다는 것을 알 수 있습니다. 위쪽 군집은 Apps, Accept, Enroll, S.F.Ration 등의 값이 비교적 컸습니다.

3. DBSCAN

실습에서 사용한 "Personal Loan.csv" 파일을 불러와 1번, 5번, 10번 column 제외한 후, 평균 = 0, 표준편차 = 1로 정규화 했습니다.

```
ploan <- read.csv("Personal Loan.csv")
ploan_x <- ploan[,-c(1,5,10)]

#Standardization
ploan_x_scaled <- scale(ploan_x, center = TRUE, scale = TRUE)
```

[Q3-1] dbscan( )함수의 eps 옵션과 minPts 옵션에 대해 각각 1부터 5의 값을 사용해 군집을 생성했습니다. 군집의 개수와 Noise points 수를 아래와 같이 정리했습니다.

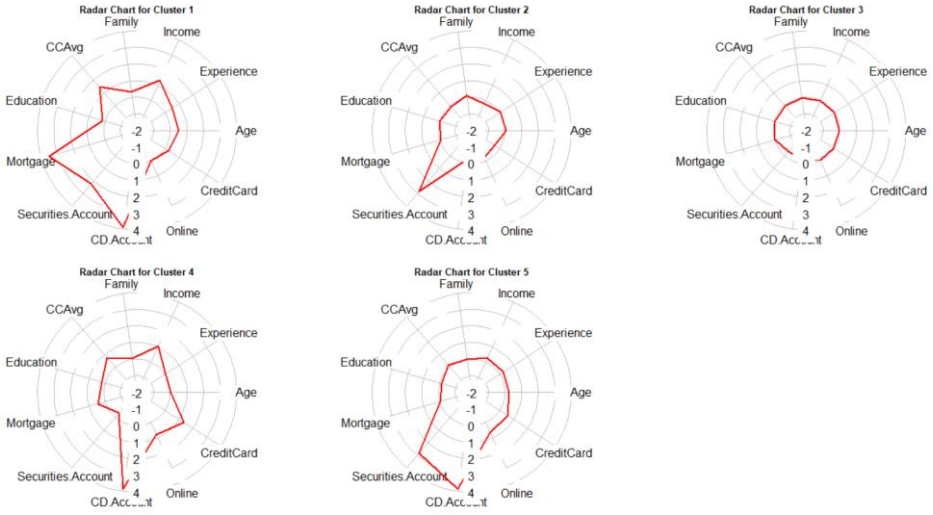
eps	minPts	군집 수	Noise 수
1	1	795	0
1	2	189	606
1	3	84	816
1	4	48	982
1	5	37	1122
2	1	86	0

2	2	28	58
2	3	16	82
2	4	14	104
2	5	11	131
3	1	10	0
3	2	5	5
3	3	4	7
3	4	4	8
3	5	4	8
4	1	2	0
4	2	2	0
4	3	2	0
4	4	2	0
4	5	2	0
5	1	1	0
5	2	1	0
5	3	1	0
5	4	1	0
5	5	1	0

같은 eps라면 minPts가 클수록 군집 수는 줄어들고 노이즈는 증가했습니다. Eps가 점점 커질수록 군집의 개수는 줄어들어 eps가 5일 때는 1개의 군집만 존재하고 노이즈가 0이었습니다. 모든 점이 하나의 군집에 할당 되었습니다.

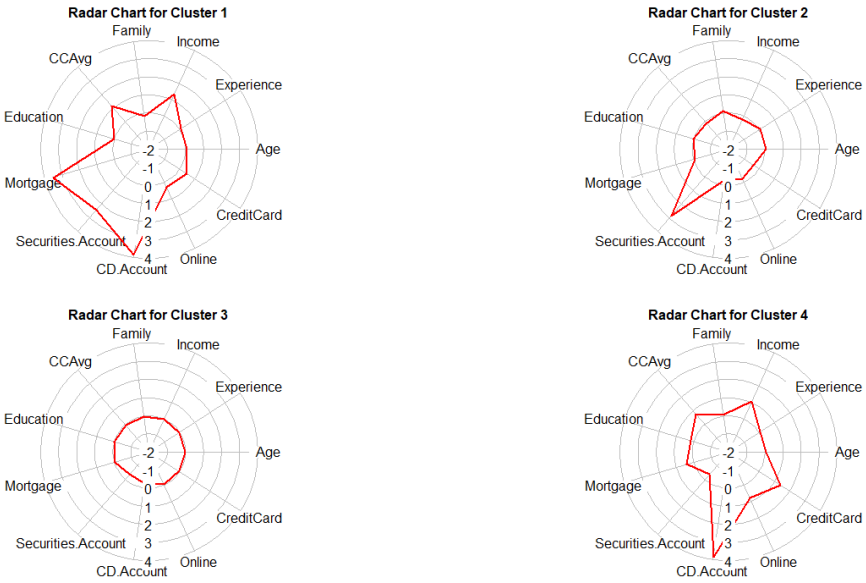
[Q3-2] 다음은 최소 3개 이상의 군집이 판별된 eps/minPts 조합에 대한 군집별 변수 Radar chart입니다.

(1) eps = 3, minPts = 2



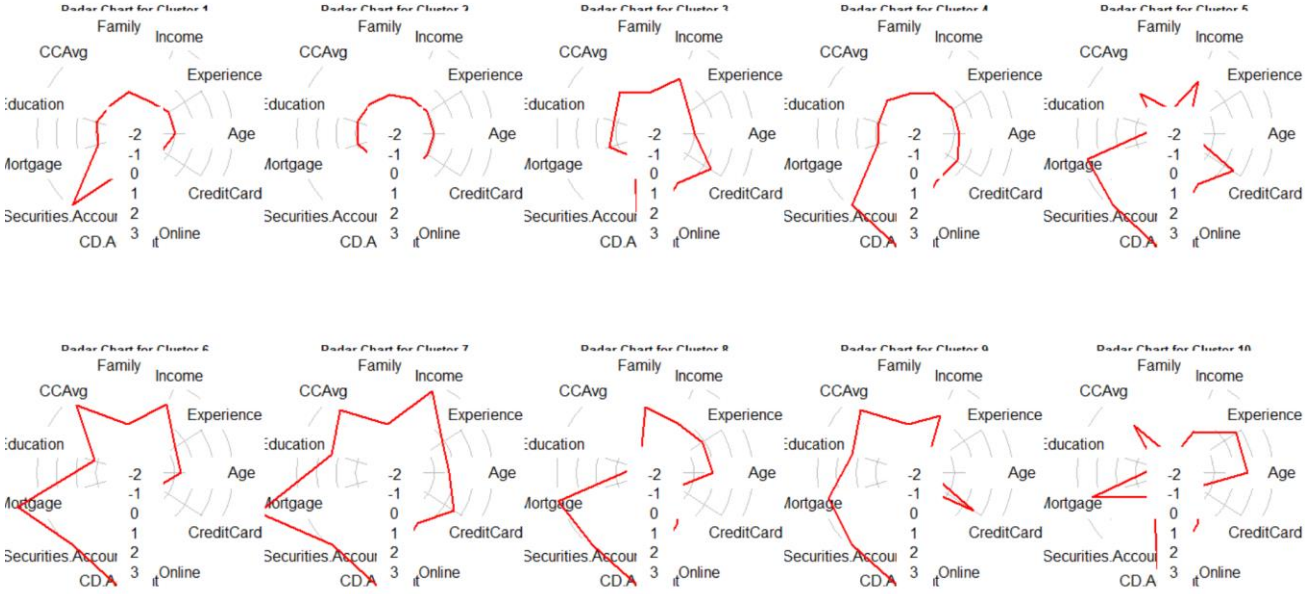
Cluster 1과 Cluster4, Cluster 5는 Cd. Account의 값이 매우 큰 특징이 있었습니다. Cluster 2는 Securities Account의 값을 제외하면 나머지 변수의 값은 대체로 작았습니다. Cluster 3은 대체로 모든 값이 작다는 특징이 있었습니다.

(2) eps = 3, minPts = 4



Cluster 1과 Cluster4는 CD.Account의 값이 매우 컸습니다. Cluster 1은 여기에 더해 Mortgage의 값 역시 컸습니다. Cluster 2는 Securities Account의 값을 제외하면 나머지 변수의 값은 눈에 띄게 크지 않았습니다. Cluster 3은 모든 값들이 작은 모습을 보였습니다.

(3) eps = 3, minPts = 1

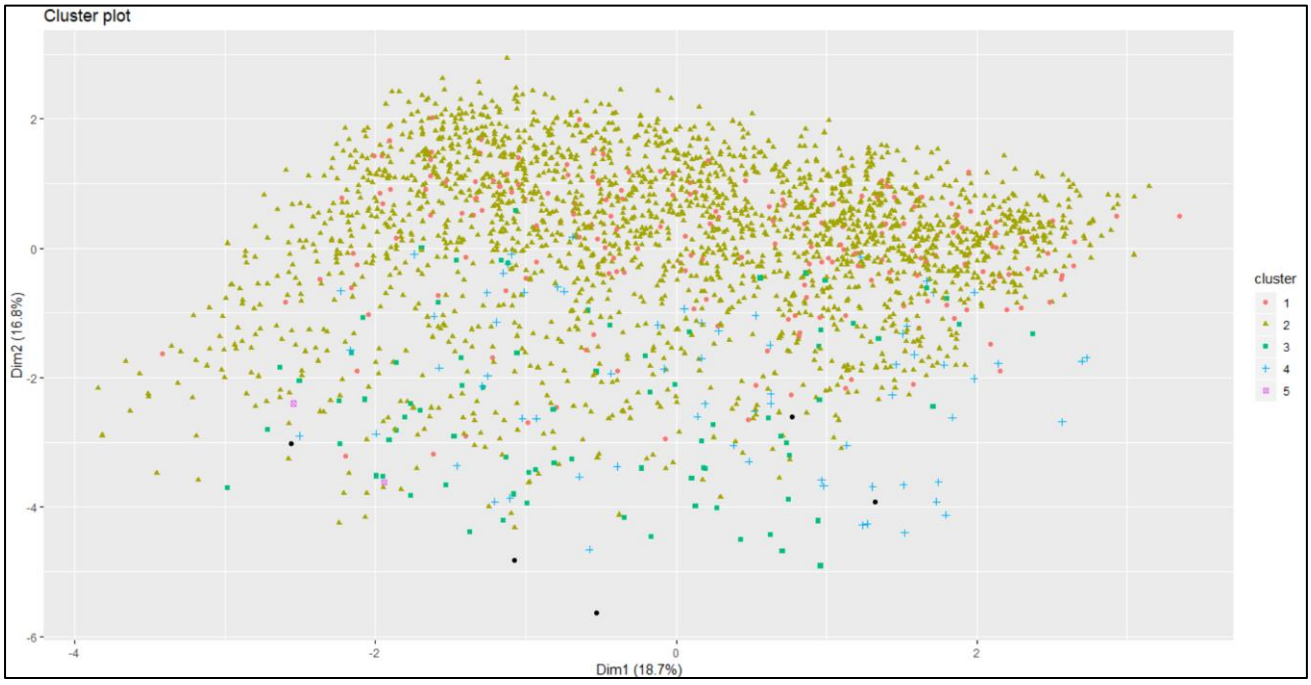


Cluster 6~10은 모두 mortgage의 값이 컸습니다. Cluster6과 7은 매우 비슷했는데, CreditCard의 값이

Cluster 6에서는 비교적 작고, Cluster 7에서는 크다는 차이가 있었습니다. Cluster 8은 Cluster 6에서 CCAvg의 값이 작아지고 Family의 값이 커졌습니다. Cluster 3, 5, 9는 CreditCard의 값이 컸습니다. Cluster 1, 4, 5, 6, 7, 8, 9는 모두 Securities Account의 값이 컸습니다.

[Q3-3] 원래 데이터를 PCA를 이용하여 2차원으로 축소시킨 후 [Q3-2]에서 선택한 군집의 수를 이용하여 군집들과 Noise points를 2차원 평면에 도시했습니다. 다음은 PCA로 축소시켰을 때의 축입니다. 여기서 상위 두 변수가 각각 18.65와 16.82만큼 설명하므로 이 둘을 축으로 하여 군집과 노이즈를 도시했습니다. 이때 사용한 eps와 minPts는 3과 2였습니다.

```
> res.pca$eig
         eigenvalue percentage of variance cumulative percentage of variance
comp 1  2.051519965             18.65018150                           18.65018
comp 2  1.850700824             16.82455294                           35.47473
comp 3  1.389886175             12.63532886                           48.11006
comp 4  1.049056203              9.53687457                           57.64694
comp 5  1.001297601              9.10270547                           66.74964
comp 6  0.956409490              8.69463173                           75.44428
comp 7  0.939572571              8.54156882                           83.98584
comp 8  0.898568022              8.16880020                           92.15464
comp 9  0.516871666              4.69883333                           96.85348
comp 10 0.340811064              3.09828240                           99.95176
comp 11 0.005306418              0.04824017                          100.00000
```



두 축이 합쳐서 전체의 35.47%만을 설명하기 때문에 군집이 제대로 분리되어 있지 않은 것처럼 보였습니다.