다변량데이터분석

과제 5

2021150456 이예지

목차

[다변량데이터분석][과제5]

[Part 1: Association Rule Mining]

[Q1] 원 데이터는 총 416,921건의 관측치와 22개의 변수가 존재하는 데이터프레임이다. 이 중에서 아래 그림과 같이 userid_DI (사용자 아이디)를 Transaction ID로 하고, institute (강좌 제공 기관), course_id (강좌코드), final_cc_cname_DI (접속 국가), LoE_DI (학위 과정)을 하나의 string으로 결합하여 Item Name으로 사용하는 연관규칙 분석용 데이터셋을 만드시오.

제공된 csv파일을 pandas 라이브러리를 통해 불러온 후, institute (강좌 제공 기관), course_id (강좌코드), final_cc_cname_DI (접속 국가), LoE_DI (학위 과정)을 "_"를 통해 연결한 "Item_Name" 컬럼을 새롭게 만들었다. 이후 "userid_DI", "Item_Name" 컬럼만을 이용하여 새로운 데이터프레임을 생성하고, "userid_DI"의 컬럼명을 "Transaction ID"로 변경함으로써 아래와 같은 데이터프레임이 만들어진 것을 확인할 수 있다. 만들어진 데이터프레임은 to_csv 함수를 통해 'association_dataset.csv'파일로 저장하였다.

	Transaction ID	Item_Name
0	MHxPC130313697	HarvardX_PH207x_India_Bachelor's
1	MHxPC130237753	HarvardX_PH207x_United States_Secondary
2	MHxPC130202970	HarvardX_CS50x_United States_Bachelor's
3	MHxPC130223941	HarvardX_CS50x_Other Middle East/Central Asia_...
4	MHxPC130317399	HarvardX_PH207x_Australia_Master's
5	MHxPC130191782	HarvardX_CS50x_Pakistan_Bachelor's
6	MHxPC130191782	HarvardX_ER22x_Pakistan_Bachelor's
7	MHxPC130267000	HarvardX_PH207x_Other South Asia_Master's
8	MHxPC130435800	HarvardX_CS50x_India_Bachelor's
9	MHxPC130284813	HarvardX_PH207x_United States_Bachelor's

[Q2-1] [Q1]에서 생성된 데이터를 읽어들이고 해당 데이터에 대한 탐색적 데이터 분석을 수행하여 데이터의 특징을 파악해보시오.

```
Data columns (total 2 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   Transaction ID  416921 non-null  object
 1   Item_Name       416921 non-null  object
```

info() 함수를 통해 416,921개의 데이터와 2개의 컬럼으로 이루어져 있는 것을 확인할 수 있다. 두 컬럼 모두 범주형 변수인 것 또한 확인할 수 있다.

[다변량데이터분석][과제5]

	Transaction ID	Item_Name
count	416921	416921
unique	335650	1405
top	MHxPC130386513	MITx_6.00x_United States_Bachelor's
freq	15	14412

총 데이터는 416,921개이나, 고유 Transaction ID는 335,650개, 고유 Item_Name은 1,405개인 것을 통해 Item_Name에 있어 중복되는 값이 많은 것을 알 수 있고, 이는 가장 많이 등장한 MITx_6.00x_United_States_Bachelor's의 빈도가 14,412인 것을 통해 확인할 수 있다. 반면, Transaction ID에서 가장 많이 등장한 MHxPC130386513의 빈도는 15인 것으로 보아, 동일한 강좌를 동일한 국가와 동일한 학위 과정 중에 들은 사람이 매우 많은 것을 파악할 수 있다. 한 명의 사용자가 여러 개의 강의를 수강한 경우 역시 존재하는 것을 파악할 수 있다.

```
Transaction ID    0
Item_Name         0
```

isnull() 함수를 통해 결측치를 확인해본 결과, 두 컬럼 모두 결측치가 존재하지 않는 것을 확인할 수 있었다.

```
Unique Transaction IDs: 335650
Most Frequent Transaction IDs:
Transaction ID
MHxPC130386513    15
MHxPC130027283    15
MHxPC130165828    15
MHxPC130308115    15
MHxPC130522464    14
MHxPC130358294    14
MHxPC130414295    14
MHxPC130464954    14
MHxPC130314286    14
MHxPC130156377    14
Name: count, dtype: int64
```

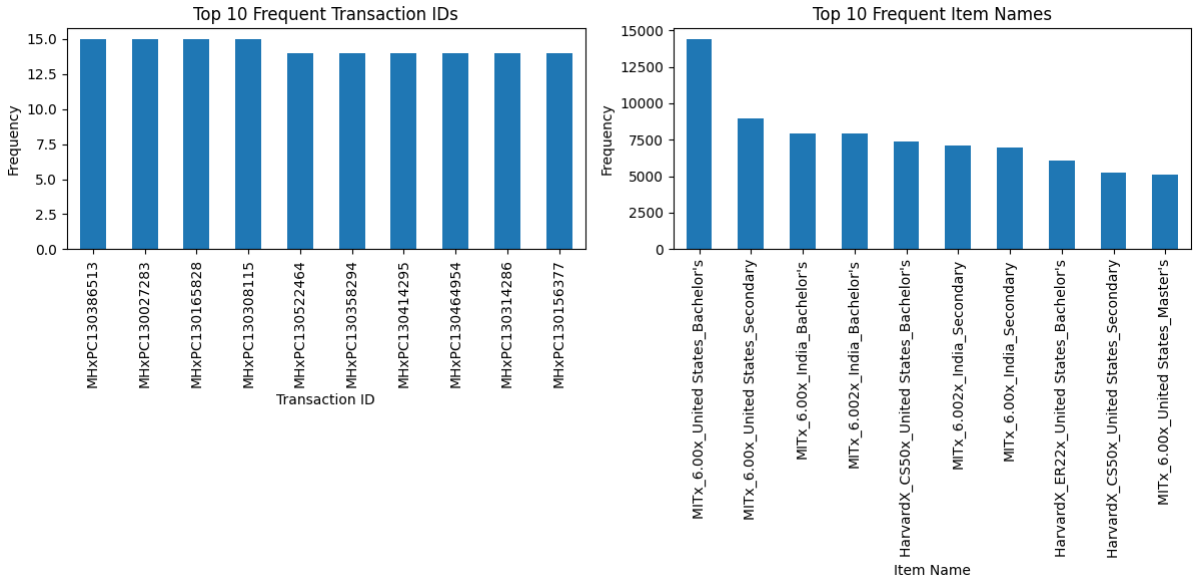
nunique() 함수를 통해 고유 Transaction ID의 수를, value_counts() 함수를 통해 각 사용자별 강좌 등록 횟수(상위 10명)를 구하였다. 고유 Transaction ID 수는 335,650으로 매우 많은 것을 알 수 있으며, 가장 많이 강좌를 등록한 사람은 MHxPC130386513, MHxPC130027283, MHxPC130165828, MHxPC130308115로, 총 15개의 강좌에 등록했으며, 그 다음은 14개의 강좌에 등록한 사람들인 것을 확인할 수 있다.

[다변량데이터분석][과제5]

```
Unique Item Names: 1405
Most Frequent Item Names:
Item_Name
MITx_6.00x_United States_Bachelor's        14412
MITx_6.00x_United States_Secondary          8944
MITx_6.00x_India_Bachelor's                 7963
MITx_6.002x_India_Bachelor's                7951
HarvardX_CS50x_United States_Bachelor's     7410
MITx_6.002x_India_Secondary                 7140
MITx_6.00x_India_Secondary                  7002
HarvardX_ER22x_United States_Bachelor's     6053
HarvardX_CS50x_United States_Secondary      5260
MITx_6.00x_United States_Master's           5093
Name: count, dtype: int64
```

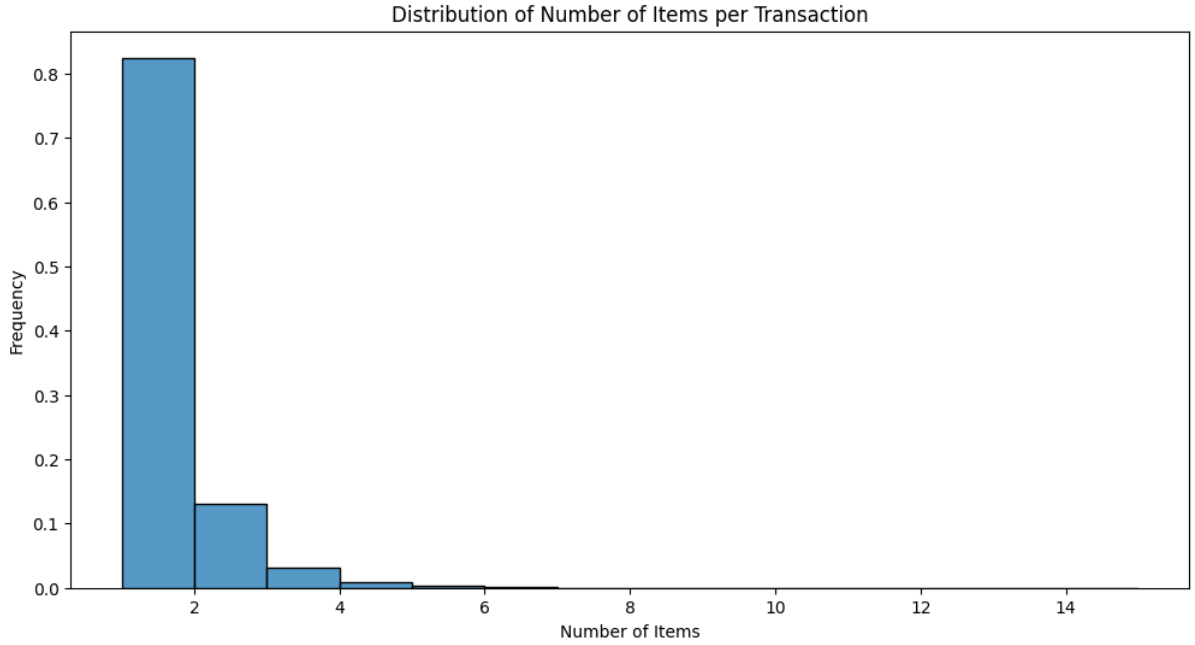
마찬가지로 고유 Item Names의 수와 각 Item 수(상위 10가지)를 구하였다. 고유 Item Names는 1,405로, 고유 Transaction ID보다 현저히 적은 것을 알 수 있다. 이를 통해 다른 사용자가 동일한 Item을 가질 수 있다는 것을 알 수 있다. MITx에서 제공하는 6.00x 강의를 United States에서 듣는 Bachelor가 14,412명으로 가장 많은 것을 알 수 있으며, MITx에서 제공하는 6.00x 강의를 United States에서 듣는 Secondary가 그 다음으로 많은 것을 알 수 있다.

이를 시각화하면 다음과 같다.



각 사용자별 강좌 등록 횟수 분포는 다음과 같다.

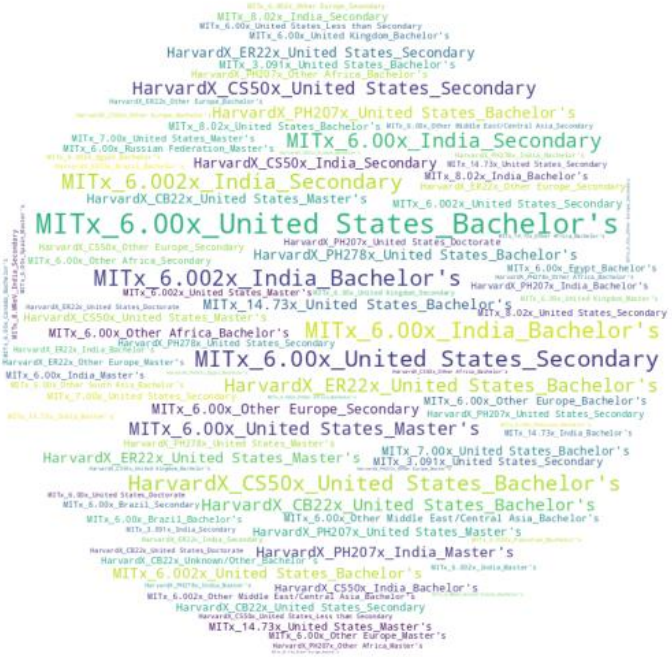
[다변량데이터분석][과제5]



1개의 강좌만을 등록한 사용자가 전체의 약 80% 정도를 차지하는 것을 알 수 있다. 2개의 강좌를 등록한 사용자는 10%를 초과하는 것을 볼 수 있다. 5개 이상의 강좌를 등록한 사용자는 거의 보이지 않을 정도로 매우 소수인 것을 확인할 수 있다. 대부분 1개~3개의 강좌를 등록하는 것을 알 수 있다.

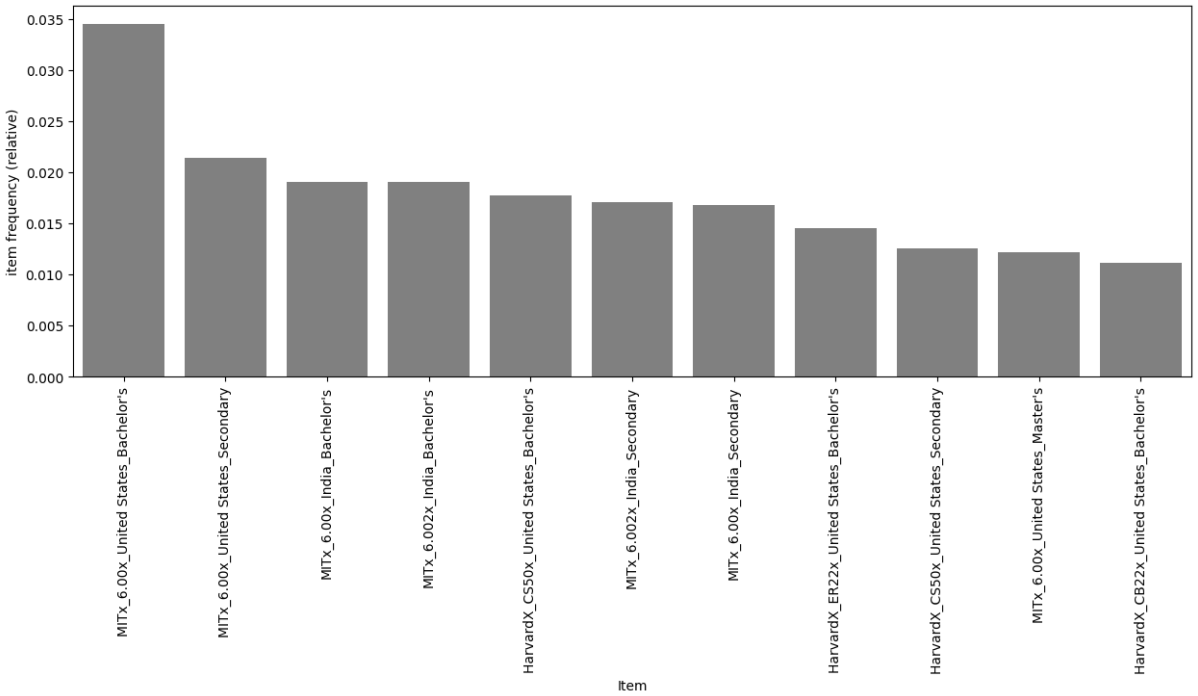
[Q2-2] 아이템 이름과 아이템 카운트를 이용하여 워드클라우드를 생성해 보시오.

[다변량데이터분석][과제5]



MITx_6.00x_United States_Bachelor's가 가장 큰 글씨로 보이는 것을 통해 해당 Item의 빈도가 가장 높았음을 알 수 있다. 다음으로는 MITx_6.00_United States_Secondary, MITx_6.002x_India_Bachelor's 등이 크게 보이는 것을 알 수 있으며, 이는 [Q2-1]에서 확인한 빈도수 상위 10개 Item 목록과 일치한다. MITx에서 제공하는 강좌의 등록인원이 많으며, United States에서 접속하는 경우, Bachelor인 경우가 많은 것도 확인할 수 있다.

[Q2-3] 최소 빈도 1% 이상 등장한 Items들의 Bar Chart를 도시하시오. 상위 5개의 Item에 대해 접속 국가는 각각 어느 국가인지 확인하시오.



상위 5개의 아이템은 MITx_6.00x_United States_Bachelor's, MITx_6.00x_United States_Secondary, MITx_6.00x_India_Bachelor's, MITx_6.002x_India_Bachelor's, HavardX_CS50x_United States_Bachelor's이므로, 각각의 접속 국가는 United States, United States, India, India, United States이다. 상위 10개의 Item으로 확장해보아도 United States가 가장 많고, 그 다음이 India인 것을 알 수 있다. 그 외 다른 국가는 상위 10개의 Item에서 확인되지 않는다.

[Q3-1] 최소 10개 이상의 규칙이 생성될 수 있도록 support와 confidence의 값을 조정해 가면서 각 support-confidence 조합에 대해 총 몇 가지의 규칙이 생성되는지 확인하고 그 결과를 아래 표와 같은 형태로 제시하시오. 최소한 3개 이상의 support, 3개 이상의 confidence, 총 9개 이상의 조합에 대한 규칙 생성을 수행하시오.

Number of rules	Confidence = 0.XXX	Confidence = 0.XXX	…
Support = 0.XXX			
Support = 0.XXX			
…			

총 데이터가 416,921개라 계속 런타임이 끊겨 학습이 제대로 진행되지 않았다. 따라서 랜덤으로 30,000개의 데이터만 추출하고, 해당 데이터들만으로 [Q3-1]과 [Q3-2]를 진행하였다. Item들은 "_"를 기준으로 split을 진행한 다음, 규칙을 생성하였다.

최소 10개 이상의 규칙이 생성되어야 하므로, 빈도 기준 상위 10번째 아이템이 전체 데이터셋에서 차지하는 비율인 약 0.01을 최소 Support 기준으로 설정하였고, 각 아이템은 4가지의 요소로 구성되어 있으므로, 더 많은 규칙이 생성될 수 있을 것이라 여겨 그보다

높은 0.05와 0.1까지 Support 범위로 설정하였다.

Confidence의 경우, Support보다는 커야 규칙으로서 의미가 있으므로, 최소 Confidence 기준을 0.1로 설정하였으며, Confidence가 높아질수록 생성되는 규칙의 수가 적어지므로, Confidence를 너무 크게 설정하면 10개 이상의 규칙이 생기지 않을 가능성이 있어 0.2와 0.3을 Confidence 후보로 설정하였다.

Number of rules	Confidence = 0.1	Confidence = 0.2	Confidence = 0.3
Support = 0.01	574	420	299
Support = 0.05	92	76	51
Support = 0.1	34	31	24

Support가 커질수록, Confidence가 커질수록 생성되는 규칙의 수가 줄어드는 것을 확인할 수 있다.

[Q3-2] support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들에 대해 다음 질문에 대한 답과 본인의 생각을 서술하시오.

- ✓ Support가 가장 높은 규칙은 무엇인가?
- ✓ Confidence가 가장 높은 규칙은 무엇인가?
- ✓ Lift가 가장 높은 규칙은 무엇인가?
- ✓ 만일 하나의 규칙에 대한 효용성 지표를 Support × Confidence × Lift로 정의한다면 효용성이 가장 높은 규칙 1위~3위는 어떤 것들인가?

다음은 support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들을 support를 기준으로 정렬한 데이터프레임이다.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
189	(Bachelor's)	(MITx)	0.437333	0.578800	0.257500	0.588796	1.017270	0.004371	1.024308	0.030172
190	(MITx)	(Bachelor's)	0.578800	0.437333	0.257500	0.444886	1.017270	0.004371	1.013606	0.040305
91	(MITx)	(6.00x)	0.578800	0.248567	0.248567	0.429452	1.727713	0.104696	1.317037	1.000000
92	(6.00x)	(MITx)	0.248567	0.578800	0.248567	1.000000	1.727713	0.104696	inf	0.560529
418	(MITx)	(Secondary)	0.578800	0.301733	0.193200	0.333794	1.106255	0.018557	1.048124	0.228038
...	...	...	...	...	...	...	...	...	...	...
1245	(Germany, 8.02x)	(MITx)	0.001000	0.578800	0.001000	1.000000	1.727713	0.000421	inf	0.421622
2178	(PH278x, Less than Secondary)	(HarvardX)	0.001000	0.421200	0.001000	1.000000	2.374169	0.000579	inf	0.579379
2177	(HarvardX, Less than Secondary)	(PH278x)	0.010200	0.058900	0.001000	0.098039	1.664503	0.000399	1.043393	0.403334
2880	(Russian Federation)	(6.002x, Secondary, MITx)	0.018000	0.041667	0.001000	0.055556	1.333333	0.000250	1.014706	0.254582
1866	(HarvardX, Mexico)	(CS50x)	0.003067	0.109300	0.001000	0.326087	2.983412	0.000665	1.321684	0.666858

4235 rows × 10 columns

총 4235개의 규칙이 생성된 것을 확인할 수 있다. Support가 가장 높은 규칙은 Bachelor's일 때 MITx인 것과 MITx일 때 Bachelor's인 것으로, support가 0.2575로 나타났다. 두 규칙의 support가 같게 나타날 수 있었던 것은, 현재 파이썬에서 support(A->B)가 P(A, B)로 계산되기 때문인 것 같

다. Bachelor's와 MITx가 함께 나타날 확률이 일정하므로 두 규칙의 support가 같게 나올 수밖에 없던 것으로 보인다. 이는 그 다음으로 support가 높은 규칙인 MITx일 때 6.00x인 것과 6.00x일 때 MITx인 것의 support 역시 동일한 것을 통해 확인할 수 있다.

[Q2]의 시각화들을 통해 빈도수 기준 상위권에서 MITx와 Bachelor가 함께 보이는 경우가 많았고, 워드클라우드에서도 큰 글씨로 두 단어가 함께 있는 경우가 많았던 것을 생각하면 이는 당연한 결과인 것 같다.

다음은 support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들을 confidence를 기준으로 정렬한 데이터프레임이다.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1014	(Morocco, 6.00x)	(MITx)	0.003733	0.578800	0.003733	1.000000	1.727713	0.001572	inf	0.422778
3973	(CS50x, Secondary, Other Europe)	(HarvardX)	0.004033	0.421200	0.004033	1.000000	2.374169	0.002334	inf	0.581144
3582	(Bachelor's, CS50x, Other Africa)	(HarvardX)	0.002467	0.421200	0.002467	1.000000	2.374169	0.001428	inf	0.580231
615	(Secondary, 2.01x)	(MITx)	0.003867	0.578800	0.003867	1.000000	1.727713	0.001629	inf	0.422835
1888	(CS50x, Other South Asia)	(HarvardX)	0.002567	0.421200	0.002567	1.000000	2.374169	0.001486	inf	0.580289
...	...	...	...	...	...	...	...	...	...	...
2889	(6.002x)	(Secondary, United States, MITx)	0.115133	0.046900	0.005767	0.050087	1.067950	0.000367	1.003355	0.071905
2887	(6.002x, MITx)	(Secondary, United States)	0.115133	0.083633	0.005767	0.050087	0.598886	-0.003862	0.964685	-0.430819
849	(6.002x)	(Secondary, United States)	0.115133	0.083633	0.005767	0.050087	0.598886	-0.003862	0.964685	-0.430819
3389	(Bachelor's, Other Europe)	(8.02x, MITx)	0.020633	0.051900	0.001033	0.050081	0.964948	-0.000038	0.998085	-0.035765
1237	(Bachelor's, Other Europe)	(8.02x)	0.020633	0.051900	0.001033	0.050081	0.964948	-0.000038	0.998085	-0.035765

4235 rows × 10 columns

Confidence가 가장 높은 규칙은 여러 개가 존재하는 것을 확인할 수 있는데, 그 중 가장 상위에 정렬된 규칙을 확인해보면, Morocco이고 6.00x일 때 MITx인 것을 확인할 수 있다. 현재 해당 규칙의 Confidence가 1인 것으로 보아, Morocco이고 6.00x일 때는 전부 Morocco, 6.00x, MITx에 해당하는 것을 알 수 있다.

현재 Confidence가 1인 다른 규칙들을 함께 보면, 한가지 공통점이 있는 것을 알 수 있다. antecedents에는 강좌 코드가 있고, consequent에는 강좌 제공 기관이 있다. 이는 강좌 코드가 강좌별로 다르며, 다른 강좌 제공 기관에서 같은 강좌 코드를 쓰는 경우가 없기 때문에 위와 같은 결과가 나온 것으로 보인다. 따라서 현재 Confidence를 그대로 받아들이면 안 될 것으로 보인다. 만약 Confidence를 규칙 평가 지표로 활용하고 싶다면, 강좌 제공 기관은 Item에서 제거하는 것이 더욱 유용한 규칙을 만드는 데 도움이 될 것 같다.

다음은 support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들을 lift를 기준으로 정렬한 데이터프레임이다.

[다변량데이터분석][과제5]

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
3551	(Bachelor's, CB22x)	(Unknown/Other, HarvardX)	0.029800	0.009333	0.003567	0.119687	12.823586	0.003289	1.125357	0.950339
3555	(Unknown/Other, HarvardX)	(Bachelor's, CB22x)	0.009333	0.029800	0.003567	0.382143	12.823586	0.003289	1.570266	0.930705
3904	(Unknown/Other, HarvardX, Secondary)	(CB22x)	0.002100	0.067933	0.001800	0.857143	12.617412	0.001657	6.524467	0.922682
3878	(Unknown/Other, HarvardX, Master's)	(CB22x)	0.002300	0.067933	0.001867	0.811594	11.946922	0.001710	4.947123	0.918409
1803	(CB22x)	(Unknown/Other, HarvardX)	0.067933	0.009333	0.007500	0.110402	11.828824	0.006866	1.113612	0.982184
...	...	...	...	...	...	...	...	...	...	...
3509	(Bachelor's, CB22x, HarvardX)	(India)	0.029800	0.167267	0.001600	0.053691	0.320992	-0.003385	0.879981	-0.685566
2804	(6.002x, India, MITx)	(Master's)	0.039133	0.213267	0.002433	0.062181	0.291563	-0.005913	0.838896	-0.716614
750	(6.002x, India)	(Master's)	0.039133	0.213267	0.002433	0.062181	0.291563	-0.005913	0.838896	-0.716614
2384	(PH207x, India)	(Secondary)	0.014800	0.301733	0.001000	0.067568	0.223931	-0.003466	0.748866	-0.778649
4132	(HarvardX, PH207x, India)	(Secondary)	0.014800	0.301733	0.001000	0.067568	0.223931	-0.003466	0.748866	-0.778649

4235 rows × 10 columns

향상도가 가장 높은 규칙은 총 두가지로, "Bachelor's이고 CB22x이면 Unknown/Other이고 HavardX이다"와 그 역이다. 이는 향상도의 식을 생각해보면 당연한 결과이다. 이 규칙 역시 antecedents에 강좌 코드가 존재하고, consequent에 강좌 제공 기관이 존재하는 점이 눈에 띈다. 다만, 이 경우에는 강좌 제공 기관 이외에도 접속 국가가 consequent에 존재해, 앞서 confidence가 높았던 규칙들보다는 조금 더 유용한 규칙일 것으로 보인다. 현재 향상도가 가장 높은 규칙은 향상도가 1을 초과하므로 조건절과 결과절이 서로 긍정적인 연관관계를 나타내고 있음을 알 수 있다.

다음은 support = 0.001, confidence = 0.05로 지정하여 생성된 연관규칙분석들을 performance(support*confidence*lift)를 기준으로 정렬한 데이터프레임이다.

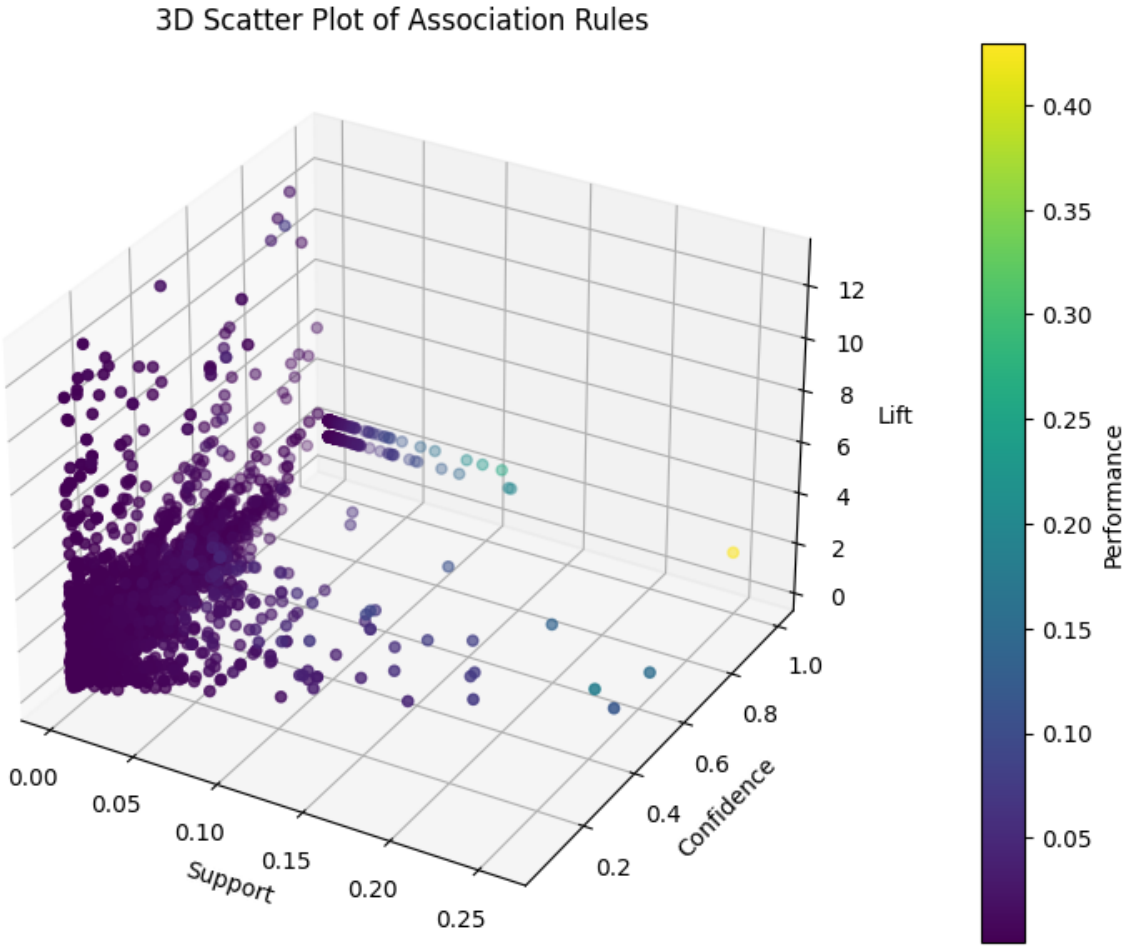
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	performance
92	(6.00x)	(MITx)	0.248567	0.578800	0.248567	1.000000	1.727713	0.104696	inf	0.560529	0.429452
255	(CS50x)	(HarvardX)	0.109300	0.421200	0.109300	1.000000	2.374169	0.063263	inf	0.649826	0.259497
310	(ER22x)	(HarvardX)	0.097433	0.421200	0.097433	1.000000	2.374169	0.056394	inf	0.641282	0.231323
366	(PH207x)	(HarvardX)	0.087633	0.421200	0.087633	1.000000	2.374169	0.050722	inf	0.634394	0.208056
52	(6.002x)	(MITx)	0.115133	0.578800	0.115133	1.000000	1.727713	0.048494	inf	0.476004	0.198917
...	...	...	...	...	...	...	...	...	...	...	...
3509	(Bachelor's, CB22x, HarvardX)	(India)	0.029800	0.167267	0.001600	0.053691	0.320992	-0.003385	0.879981	-0.685566	0.000028
1392	(Bachelor's, CB22x)	(India)	0.029800	0.167267	0.001600	0.053691	0.320992	-0.003385	0.879981	-0.685566	0.000028
1965	(Canada)	(Master's, MITx)	0.020700	0.104867	0.001067	0.051530	0.491384	-0.001104	0.943765	-0.513843	0.000027
4132	(HarvardX, PH207x, India)	(Secondary)	0.014800	0.301733	0.001000	0.067568	0.223931	-0.003466	0.748866	-0.778649	0.000015
2384	(PH207x, India)	(Secondary)	0.014800	0.301733	0.001000	0.067568	0.223931	-0.003466	0.748866	-0.778649	0.000015

4235 rows × 11 columns

performance가 가장 높은 규칙은 "6.00x이면 MITx이다"로, 강좌 제공 기관별로 강좌 코드를 설정하는 방식이 다르고, 각 강좌는 서로 다른 강좌 코드를 가지고 있기 때문에 Confidence가 1일 수밖에 없고, 6.00x는 MITx에서 제공하는 강좌이므로 둘이 함께 나타날 확률이 높아, Support와 Lift 역시 높은 값을 가질 수밖에 없다. 다만, MITx에서 제공하는 강좌는 여러가지가 존재하고, MITx에서 제공하는 6.002x 강좌도 높은 인기를 보였으므로 Lift가 매우 큰 값을 가질 수는 없다는 것을 예상할 수 있다. 현재 상위권에 위치한 규칙들은 모두 강좌 코드가 조건절로, 강좌 제공 기관이 결과절로 포함된 것을 알 수 있다. 따라서 앞서 언급했다시피, 해당 규칙들은 크게 유용한 규칙이 되지 못할 것으로 보이며, 유용한 규칙을 얻고 싶다면 강좌 제공 기관을 제거하는 것이 바람직해

보인다.

[Extra Question] 이 외 수업 및 실습 시간에 다루지 않은 연관규칙분석 시각화 및 해석을 시도해 보시오.



[Q3-2]에서 구한 규칙들을 3차원 산점도로 시각화해보았다. Support, Confidence, Lift를 세 축으로 사용하였으며, 앞서 구한 Performance(Support*Confidence*Lift)는 색을 통해 표현하였다. Support가 낮은 경우가 많은 것을 알 수 있는데, 이는 데이터 개수가 많지만 그 개별 값들이 동일한 경우가 많지 않았기 때문으로 보인다. 반면 상대적으로 Confidence는 높은 경우가 많으며, 이는 강좌 코드와 강좌 제공 기관 사이의 관계 때문인 것으로 보인다. Lift는 1을 넘어가는 규칙이 상당히 많은 것을 알 수 있는데, 이 역시도 강좌 코드와 강좌 제공 기관 사이의 매우 높은 관련성으로 인해 Lift의 값이 높아진 것으로 추측된다. Performance는 Support로 인해 거의 모든 규칙의 값이 매우 낮은 것을 확인할 수 있다. 대부분의 규칙이 크게 유용하지 않으며, Performance가 높더라도 각 요소별 연관성으로 인해 규칙의 내용을 면밀히 살펴볼 필요가 있다.

[다변량데이터분석][과제5]

[Part 2: Clustering]

[Q1] 데이터셋 선정하기

이 중에서 군집화 후 각 군집에 대한 속성 분석이 유의미할(또는 재미있을) 것으로 판단되는 데이터셋 하나를 선정하고 본인이 해당 데이터셋을 선정한 이유를 설명하시오.

선정한 데이터셋은 고객을 세분화하여 마케팅 전략을 수립하는데 도움을 주기 위해 만들어진 'Creadit Card Dataset for Clustering' 데이터셋이다.

편의점 아르바이트를 하면서, 어르신들은 주로 막걸리를, 아이들은 주로 튀김 같은 음식을 사가는 것을 많이 보았다. 편의점에 방문하는 손님들은 매주 비슷했으며, 구매 상품 역시 비슷한 것을 보았다. 그렇기에 방문하는 손님들을 몇 개의 그룹으로 나눈 뒤, 그에 맞는 서비스를 제공하면 매출이 조금 더 상승할 수 있지 않을까 하는 생각이 들었다.

따라서 고객 구매 데이터셋을 선정하여 군집화를 진행하고 각 군집에 대한 속성 분석을 진행하는 것이 흥미로울 것으로 판단했고, 본 데이터셋은 고객이 얼마나 자주 구매를 하는지, 구매 횟수는 얼마인지, 지불 금액은 얼마인지 등 고객과 관련된 다양한 attribute가 존재하기 때문에, 각 고객군의 특성을 파악해보는 연습을 하기에 적절할 것이라 생각했다.

실제로도 회사들은 제품이나 서비스를 보다 효과적으로 판매하기 위해 고객들의 군집을 이해하고, 이를 기반으로 각각의 군집에 대한 마케팅 전략을 수립해야 한다. 실제로 많은 기업들이 마케팅 및 고객 관리를 위해 고객들의 구매 데이터를 파악하고, 유사한 고객들을 하나의 군집으로 만들어 군집 분석을 수행하고 있다. 특정 군집의 고객들을 위한 맞춤형 프로모션을 제공하거나, 제품 타겟층을 명확히 하는 등의 전략 수립을 통해 각 기업은 이익을 극대화할 수 있기 때문이다.

데이터셋 링크: Credit Card Dataset for Clustering (kaggle.com)

[K-Means Clustering]

[Q2] K-Means Clustering의 군집 수를 2개부터 K(데이터의 규모에 따라서 본인이 적절히 선정할 것)개까지 증가시켜(증가 폭 역시 적절히 설정) 가면서 군집화 타당성 지표 값들을 산출하시오. 총 소요 시간은 얼마인가? Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

총 8,950개의 데이터가 존재하나, 일부 결측치가 존재하여 결측치를 제거한 8,636개의 데이터만으로 군집화를 수행하였다. 군집화에 있어 필요하지 않은 'CUST_ID'(고객 아이

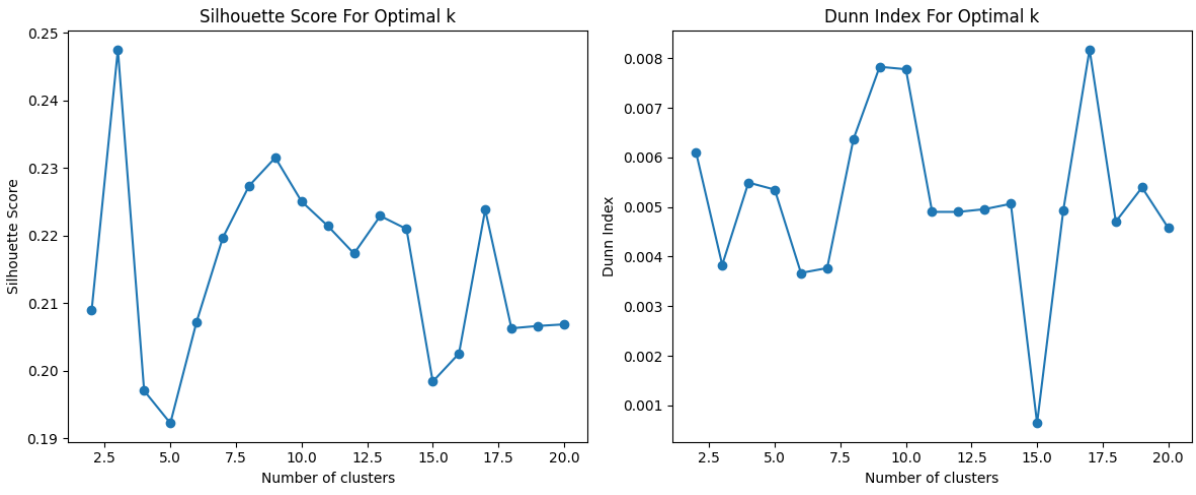
디) 컬럼은 제거하였으며, 나머지 변수들은 연속형이므로 단위에 영향을 받지 않도록 scaling을 수행하였다.

데이터가 약 8,500개로, 고객들을 분류하는데 있어 크게 많은 수치는 아니라고 생각했으며, 한 군집에 속하는 고객의 수가 너무 적다면 고객군을 만드는 비용이 그 효용에 비해 매우 클 것이라 생각했으므로 최대 군집 수 K를 20으로 설정하였다. 설정한 최대 군집 수가 크지 않으므로, 모든 경우의 수를 따져볼 필요가 있다고 생각하여 증가 폭은 1로 설정하였다.

군집 수에 따른 Dunn Index와 Silhouette Index 산출 결과는 다음과 같다.

```
Clusters: 2, Silhouette Score: 0.20892328220228673, Dunn Index: 0.006108152604600334
Clusters: 3, Silhouette Score: 0.24754638153191033, Dunn Index: 0.003831076255042093
Clusters: 4, Silhouette Score: 0.19703675772374923, Dunn Index: 0.005493746388396378
Clusters: 5, Silhouette Score: 0.19223817739937715, Dunn Index: 0.005351742352507408
Clusters: 6, Silhouette Score: 0.2072425705719404, Dunn Index: 0.0036708378639907944
Clusters: 7, Silhouette Score: 0.219694021309588, Dunn Index: 0.0037689339255379743
Clusters: 8, Silhouette Score: 0.2273670636332744, Dunn Index: 0.006366641861911749
Clusters: 9, Silhouette Score: 0.23154306508064237, Dunn Index: 0.007829713024410762
Clusters: 10, Silhouette Score: 0.22505725875570437, Dunn Index: 0.007780072766017558
Clusters: 11, Silhouette Score: 0.22143826685323337, Dunn Index: 0.004902237575505663
Clusters: 12, Silhouette Score: 0.21734016504333378, Dunn Index: 0.004902237575505663
Clusters: 13, Silhouette Score: 0.22291261721019756, Dunn Index: 0.004957348941106474
Clusters: 14, Silhouette Score: 0.22098363076626604, Dunn Index: 0.005065443812656851
Clusters: 15, Silhouette Score: 0.19838743330043473, Dunn Index: 0.0006399115159866519
Clusters: 16, Silhouette Score: 0.20254513121820442, Dunn Index: 0.004936700145446942
Clusters: 17, Silhouette Score: 0.22368826048347996, Dunn Index: 0.008177276382774084
Clusters: 18, Silhouette Score: 0.20627797834724196, Dunn Index: 0.004705241673745851
Clusters: 19, Silhouette Score: 0.20662418957260784, Dunn Index: 0.005395040063005294
Clusters: 20, Silhouette Score: 0.20686099537117467, Dunn Index: 0.0045780980001930078
Total computation time: 85.77254486083984 seconds
```

총 소요 시간은 약 86초 정도인 것을 확인할 수 있다. 군집화 타당성 지표값들을 시각화하면 다음과 같다.



Silhouette Index를 기준으로 보면, 군집 수가 3일 때 그 값이 가장 높은 것을 확인할 수

있으므로, 최적의 군집 수는 3으로 나타난다.

[Q3] [Q2]에서 선택된 군집의 수를 사용하여 K-Means Clustering을 10회 반복하고 회차마다 각 군집의 Centroid와 Size를 확인해보시오. 10회 반복 시 몇 가지 경우의 군집화 결과물이 도출되었으며 각 경우의 군집화는 몇번 반복되어 발생하는지 확인해보시오.

군집 수를 3으로 설정하고 K-Means Clustering을 10회 반복 실험한 결과는 다음과 같다.

회차	군집화 반복 수	Size	마지막 단계의 Centroid
1	17	[5863, 1211, 1562]	[-0.02519338, 0.29533849, -0.13433865]
2	17	[5862, 1211, 1563]	[-0.0253232, 0.29533849, -0.13371306]
3	17	[4236, 579, 3821]	[-0.08285707, 0.31038579, 0.04485655]
4	17	[5864, 1559, 1213]	[-0.02408763, -0.13919198, 0.29543716]
5	17	[1212, 1568, 5856]	[0.29523949, -0.13864385, -0.0237594]
6	17	[1544, 1219, 5873]	[-0.14068368, 0.2958286, -0.02454561]
7	17	[1211, 5862, 1563]	[0.29533849, -0.02525828, -0.13402565]
8	17	[5871, 1221, 1544]	[-0.02473971, 0.29597407, -0.14068368]
9	17	[5865, 1222, 1549]	[-0.0251285, 0.2960224, -0.13908205]
10	17	[5871, 1221, 1544]	[-0.02473971, 0.29597407, -0.14068368]

모든 경우에서 군집화 반복 수는 17로 확인되었으며, 군집의 Size는 매번 달라지는 것을 알 수 있었다. 그러나 5800, 1200, 1500 정도의 Size를 갖는 경우가 가장 많은 것을 알 수 있었다. 이는 초기 Centroid의 값에 따른 영향인 것으로 추측된다. Centroid의 경우, 모든 회차의 군집화 반복 과정에서 모든 Centroid를 출력하였으나, 지면 상 마지막 단계의 Centroid만을 적었다. 마지막 단계의 Centroid 값에 따라 군집의 Size가 달라지는 것을 확인할 수 있으며, Centroid가 비슷한 경우 군집의 Size 역시 비슷한 것을 확인할 수 있었다. 각 군집의 Centroid에 따라 경계면에 있는 점들이 어느 군집으로 속하게 되는지가 바뀌게 되어 약간의 Size 차이가 발생한 것으로 예상된다.

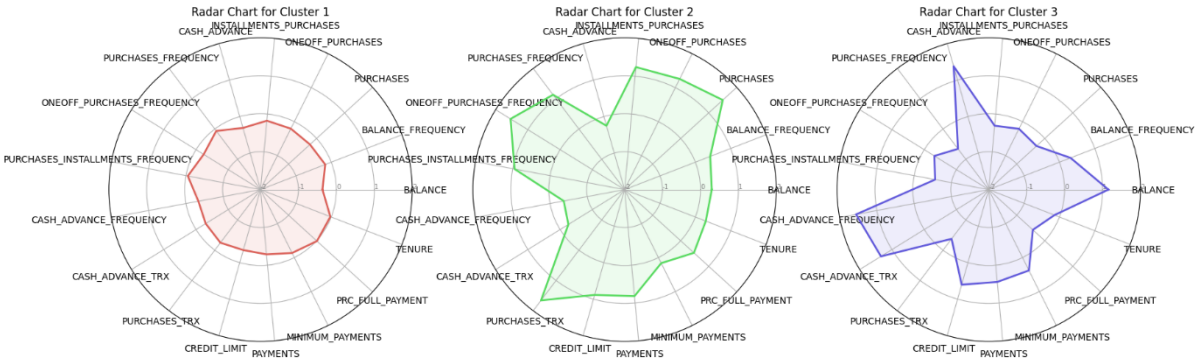
[Q4] [Q3]에서 가장 빈번하게 발생한 군집화 결과물에 대해서 각 군집별 변수들의 평균값을 이용한 Rader Chart를 도시해보시오. Rader Chart상으로 판단할 때, 군집의 속성이 가장 상이할 것으로 예상되는 두 군집(군집 A와 군집 B로 명명)과, 가장 유사할 것으로 예상되는 두 군집(군집 X와 군집 Y로 명명)을 각각 선택하고 선택 이유를 설명하시오.

[Q3]에서 가장 빈번하게 발생한 군집화 결과물에 대해서 각 변수들의 평균값을 구하면 다음과 같다.

[다변량데이터분석][과제5]

clusterID	0	1	2
BALANCE	-0.366003	0.298697	1.155502
BALANCE_FREQUENCY	-0.169929	0.420214	0.313842
PURCHASES	-0.235080	1.502889	-0.294606
ONEOFF_PURCHASES	-0.206072	1.257483	-0.210840
INSTALLMENTS_PURCHASES	-0.176978	1.241887	-0.309137
CASH_ADVANCE	-0.310455	-0.250593	1.378664
PURCHASES_FREQUENCY	-0.063122	1.131832	-0.655038
ONEOFF_PURCHASES_FREQUENCY	-0.237023	1.537543	-0.314625
PURCHASES_INSTALLMENTS_FREQUENCY	-0.049984	0.951181	-0.562133
CASH_ADVANCE_FREQUENCY	-0.333496	-0.366759	1.558141
CASH_ADVANCE_TRX	-0.299525	-0.256087	1.341447
PURCHASES_TRX	-0.246135	1.657173	-0.374580
CREDIT_LIMIT	-0.342962	0.884415	0.604703
PAYMENTS	-0.286834	0.818566	0.443350
MINIMUM_PAYMENTS	-0.134302	0.165362	0.379909
PRC_FULL_PAYMENT	0.013619	0.472774	-0.425656
TENURE	-0.024610	0.295829	-0.140363

이를 Rader Chart를 통해 시각화하면 다음과 같다.



군집의 속성이 유사한지, 상이한지에 대해서는 두 군집이 겹치는 부분을 기준으로 판별하였다.

군집의 속성이 가장 상이할 것으로 예상되는 두 군집은 Cluster 2와 Cluster 3이다. Cluster 2에서 높은 평균값을 가진 변수들이 Cluster 3에서 낮은 평균값을 보여주었고, 반대로 Cluster 3에서 높은 평균값을 보여준 변수들은 Cluster 2에서 낮은 평균값을 보여주었다. 따라서 두 군집은 겹치지 않는 부분이 Cluster 1과 다른 군집들을 비교하였을 때보다 많으므로 군집의 속성이 상이할 것으로 생각하였다.

군집의 속성이 가장 유사할 것으로 예상되는 두 군집은 Cluster 1과 Cluster 2이다. Cluster 1과 Cluster 2의 중복되는 면적이 Cluster 1과 Cluster 3의 중복되는 면적보다 넓기 때문에 두 군집의 속성이 가장 유사할 것으로 생각하였다. Cluster 2와 Cluster 3는 중복되는 면적도 어느 정도 있지만, 중복되지 않는 부분이 많았기에 유사하다고 판단하기 어려웠다.

[Q5] [Q4]에서 선택된 군집 A와 군집 B에 대해 각 변수별 평균값 차이에 대한 통계적 검정을 수

[다변량데이터분석][과제5]

행하시오. 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중은 얼마인가? 또한 [Q4]에서 선택된 군집 X와 군집 Y에 대해서도 각 변수별 평균값 차이에 대한 통계적 검정을 수행하시오. 이 경우, 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중은 얼마인가?

[Q4]에서 선택된 군집 A와 군집 B는 Cluster 2와 Cluster 3이다. 이 군집들에 대해 각 변수별 평균값 차이에 대해 t 검정을 진행한 결과는 다음과 같다.

t 검정의 귀무가설은 두 군집의 변수별 평균값에 차이가 없다는 것이고, 대립가설은 차이가 있다는 것이다. 이때 대립가설은 양측 검정을 기준으로 한 것이며, 단측 검정일 경우 한 군집의 평균값이 다른 군집의 평균값보다 높다는 것이 대립가설이 된다.

	two_sided	greater	less
BALANCE	3.366336e-67	1.000000e+00	1.683168e-67
BALANCE_FREQUENCY	1.184000e-10	5.920001e-11	1.000000e+00
PURCHASES	1.563931e-160	7.819657e-161	1.000000e+00
ONEOFF_PURCHASES	1.653874e-101	8.269371e-102	1.000000e+00
INSTALLMENTS_PURCHASES	1.080838e-121	5.404192e-122	1.000000e+00
CASH_ADVANCE	3.928341e-225	1.000000e+00	1.964170e-225
PURCHASES_FREQUENCY	0.000000e+00	0.000000e+00	1.000000e+00
ONEOFF_PURCHASES_FREQUENCY	0.000000e+00	0.000000e+00	1.000000e+00
PURCHASES_INSTALLMENTS_FREQUENCY	0.000000e+00	0.000000e+00	1.000000e+00
CASH_ADVANCE_FREQUENCY	0.000000e+00	1.000000e+00	0.000000e+00
CASH_ADVANCE_TRX	5.022229e-211	1.000000e+00	2.511115e-211
PURCHASES_TRX	1.085989e-234	5.429943e-235	1.000000e+00
CREDIT_LIMIT	9.208251e-11	4.604126e-11	1.000000e+00
PAYMENTS	1.405663e-10	7.028314e-11	1.000000e+00
MINIMUM_PAYMENTS	2.543862e-04	9.998728e-01	1.271931e-04
PRC_FULL_PAYMENT	1.573825e-103	7.869126e-104	1.000000e+00

먼저 양측검정부터 살펴보면, 모든 변수에서 유의확률이 0.05보다 작은 것을 알 수 있다. 따라서 모든 변수에서 귀무가설을 기각할 수 있다. 이는 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중이 100%라는 것을 의미한다.

단측검정을 살펴보면, greater의 경우, 13개의 변수에서 유의확률이 0.05보다 작은 것을 알 수 있다. 해당 변수들에서만 귀무가설을 기각할 수 있고, 이는 전체 변수 중 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중이 약 76%정도이고, cluster 2가 cluster 3보다 13개의 변수에서 평균값이 더 높다는 것을 의미한다.

less의 경우에는, 4개의 변수에서 유의확률이 0.05보다 작은 것을 알 수 있다. 마찬가지로 해당 변수들에서만 귀무가설을 기각할 수 있고, 이는 전체 변수 중 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중이 약 24%이며, cluster 3가 cluster 2보다 4개의 변수에서 평균값이 더 높다는 것을 의미한다.

양측검정의 결과가 두 단측검정의 결과를 합친 것과 동일하다는 점을 알 수 있다. 모든 변수의

[다변량데이터분석][과제5]

평균값에 차이가 있는 것으로 보아, 두 군집은 상이하다고 이야기할 수 있을 것으로 보인다.

[Q4]에서 선택된 군집 X와 군집 Y는 Cluster 1과 Cluster 2이다. 이 군집들에 대해 각 변수별 평균값 차이에 대해 t 검정을 진행한 결과는 다음과 같다.

	two_sided	greater	less
BALANCE	1.489650e-68	1.000000	7.448252e-69
BALANCE_FREQUENCY	9.903335e-227	1.000000	4.951668e-227
PURCHASES	1.885101e-153	1.000000	9.425504e-154
ONEOFF_PURCHASES	8.894561e-102	1.000000	4.447280e-102
INSTALLMENTS_PURCHASES	2.793691e-106	1.000000	1.396846e-106
CASH_ADVANCE	1.198994e-03	0.999401	5.994971e-04
PURCHASES_FREQUENCY	0.000000e+00	1.000000	0.000000e+00
ONEOFF_PURCHASES_FREQUENCY	0.000000e+00	1.000000	0.000000e+00
PURCHASES_INSTALLMENTS_FREQUENCY	9.244470e-237	1.000000	4.622235e-237
CASH_ADVANCE_FREQUENCY	1.025852e-01	0.051293	9.487074e-01
CASH_ADVANCE_TRX	2.190086e-02	0.989050	1.095043e-02
PURCHASES_TRX	2.507918e-215	1.000000	1.253959e-215
CREDIT_LIMIT	2.004439e-191	1.000000	1.002219e-191
PAYMENTS	1.595936e-101	1.000000	7.979682e-102
MINIMUM_PAYMENTS	5.536077e-11	1.000000	2.768038e-11
PRC_FULL_PAYMENT	3.123825e-30	1.000000	1.561913e-30

먼저 양측검정부터 살펴보면, 모든 변수에서 유의확률이 0.05보다 작은 것을 알 수 있다. 따라서 모든 변수에서 귀무가설을 기각할 수 있다. 이는 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중이 100%라는 것을 의미한다.

단측검정을 살펴보면, greater의 경우, 유의확률이 0.05보다 작은 변수가 존재하지 않는 것을 알 수 있다. 따라서 모든 경우에서 귀무가설을 기각할 수 없다. 이는 전체 변수 중 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중이 0%이고, cluster 1이 cluster 2보다 평균값이 높은 변수가 존재하지 않음을 의미한다.

less의 경우에는, 모든 변수에서 유의확률이 0.05보다 작은 것을 알 수 있다. 따라서 모든 변수들에서 귀무가설을 기각할 수 있다. 이는 전체 변수 중 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중이 100%이며, cluster 2가 cluster 1보다 모든 변수에서 평균값이 더 높다는 것을 의미한다.

마찬가지로 양측검정의 결과가 두 단측검정의 결과를 합친 것과 동일하다는 점을 알 수 있다. 모든 변수의 평균값에 차이가 있는 것으로 보아, 두 군집의 속성이 유사하다고 이야기하기는 어려울 수 있어 보인다.

[다변량데이터분석][과제5]

[Hierarchical Clustering]

[Q6] 두 객체 사이의 유사도를 측정하는 지표를 본인의 기준에 따라 정의하고(유클리드 거리, 상관계수 등) “single”과 “complete” 두 가지 linkage에 대해 군집 수를 2개부터 K(데이터의 규모에 따라서 본인이 적절히 선정할 것)개까지 증가시켜(증가 폭 역시 적절히 설정) 가면서 군집화 타당성 지표 값들을 산출하시오. Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

두 객체 사이의 유사도를 측정하는 지표로 상관계수를 설정하였다. 현재 데이터는 StandardScaler를 통해 scaling을 진행하여 단위의 영향을 따로 받지 않는다. 그러나 유사한 고객을 군집화하는 것이 목적이므로 단순히 두 점이 떨어진 거리보다는 두 점 사이의 상관관계를 가지고 유사도를 평가하는 것이 본래의 목적에 더 맞는다고 판단하였기 때문이다.

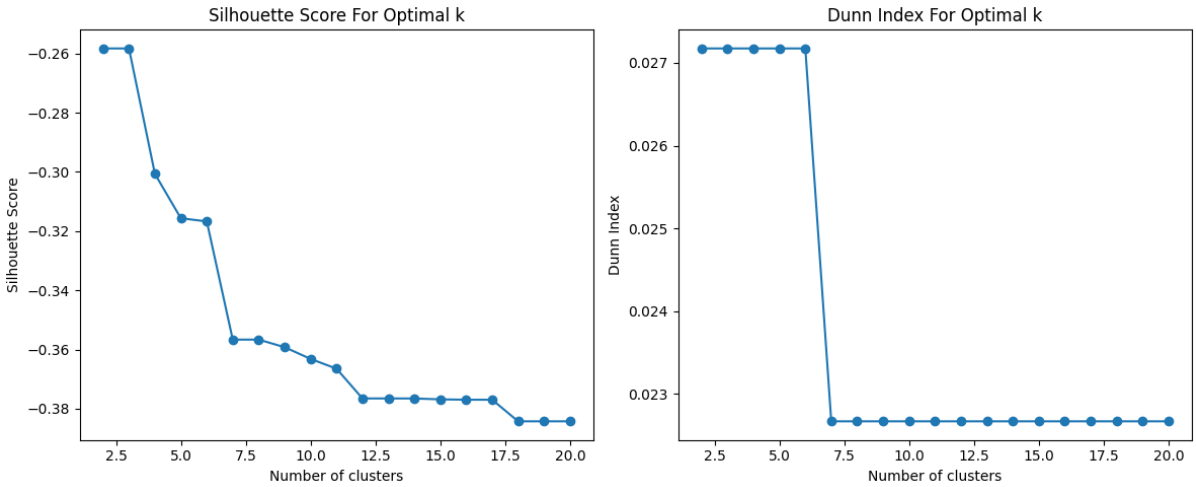
Hierarchical Clustering 역시 K-Means Clustering과 동일하게 최대 군집 수를 20으로 설정하였다. 이는 군집의 수를 너무 크게 하면 각 고객군을 관리하는 비용이 많이 드는 반면, 효용이 크지 않기 때문이며, 최대 군집의 수를 너무 작게 설정할 경우 최적 군집에 도달하지 못할 가능성이 있기 때문이다. 증가 폭 역시 이전과 동일하게 1로 설정하였다. 실험할 군집의 개수가 많지 않기 때문에, 모든 경우를 따져봐야 최적의 군집을 건너뛰지 않고 발견할 수 있기 때문이다.

먼저 single linkage를 사용한 경우에 대해 살펴보면, 군집화 타당성 지표 값들은 다음과 같이 산출되었다.

```
Clusters: 2, Silhouette Score: -0.26828468787950676, Dunn Index: 0.0271745010174048
Clusters: 3, Silhouette Score: -0.26828468787950676, Dunn Index: 0.0271745010174048
Clusters: 4, Silhouette Score: -0.30067420266808536, Dunn Index: 0.0271745010174048
Clusters: 5, Silhouette Score: -0.3166659738933481S, Dunn Index: 0.0271745010174048
Clusters: 6, Silhouette Score: -0.31671688400882875, Dunn Index: 0.0271745010174048
Clusters: 7, Silhouette Score: -0.35666161988968036, Dunn Index: 0.022668541206571436
Clusters: 8, Silhouette Score: -0.35666161988968036, Dunn Index: 0.022668541206571436
Clusters: 9, Silhouette Score: -0.35921596456980775, Dunn Index: 0.022668541206571436
Clusters: 10, Silhouette Score: -0.3631744572989749G, Dunn Index: 0.022668541206571436
Clusters: 11, Silhouette Score: -0.3664724407481991, Dunn Index: 0.022668541206571436
Clusters: 12, Silhouette Score: -0.37652668176171B2, Dunn Index: 0.022668541206571436
Clusters: 13, Silhouette Score: -0.37652668176171B2, Dunn Index: 0.022668541206571436
Clusters: 14, Silhouette Score: -0.37653764188185G5, Dunn Index: 0.022668541206571436
Clusters: 15, Silhouette Score: -0.37693221170367996, Dunn Index: 0.022668541206571436
Clusters: 16, Silhouette Score: -0.37696071172760937, Dunn Index: 0.022668541206571436
Clusters: 17, Silhouette Score: -0.37696071172760937, Dunn Index: 0.022668541206571436
Clusters: 18, Silhouette Score: -0.38425634470718T8, Dunn Index: 0.022668541206571436
Clusters: 19, Silhouette Score: -0.38425634470718T8, Dunn Index: 0.022668541206571436
Clusters: 20, Silhouette Score: -0.38425634470718T8, Dunn Index: 0.022668541206571436
```

이를 시각화하면 다음과 같다.

[다변량데이터분석][과제5]



군집 개수가 2 또는 3일 때 가장 높은 실루엣 계수를 가져, 군집 개수가 2 또는 3일 때가 최적인 것을 알 수 있다. 그러나 그 경우에도 실루엣 계수가 음수로, 성능이 좋지 않은 것을 알 수 있다.

실루엣 계수의 그래프를 보면, 군집 수가 증가할수록 실루엣 계수가 증가하는 경향이 나타난다. 특히 군집 수가 4개일 때와 군집 수가 7개일 때 급격히 감소하는 모습이 나타난다. 이는 군집의 수가 많아져 군집화가 잘 되지 않은 것을 의미한다.

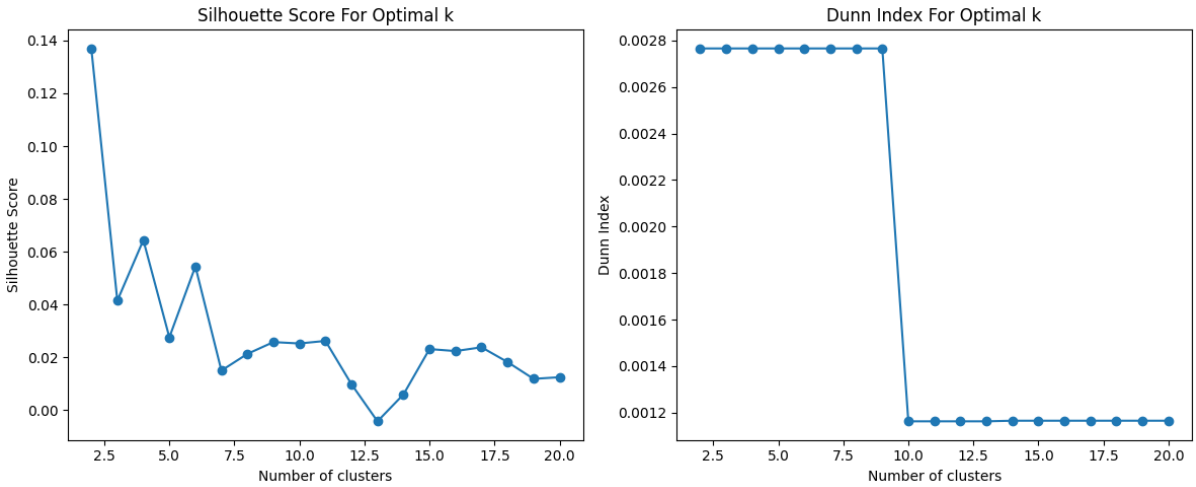
Dunn Index는 군집 개수가 2~6 사이일 때 가장 높게 나타나나, 군집 수가 7개가 되는 순간 점수가 급격히 하락하는 모습을 보인다. 이는 실루엣 계수와 동일하며, 군집의 개수를 7개 이상으로 설정하는 것은 바람직하지 않다고 해석할 수 있다.

complete linkage인 경우를 보면, 군집화 타당성 지표 값들은 다음과 같이 산출되었다.

```
Clusters: 2, Silhouette Score: 0.13700673899309725, Dunn Index: 0.0027669765788812974
Clusters: 3, Silhouette Score: 0.04144847812066619, Dunn Index: 0.0027669765788812974
Clusters: 4, Silhouette Score: 0.06440143993283122, Dunn Index: 0.0027669765788812974
Clusters: 5, Silhouette Score: 0.027546948292857518, Dunn Index: 0.0027669765788812974
Clusters: 6, Silhouette Score: 0.05442939931464243, Dunn Index: 0.0027669765788812974
Clusters: 7, Silhouette Score: 0.014974870314068486, Dunn Index: 0.0027669765788812974
Clusters: 8, Silhouette Score: 0.021291441521108234, Dunn Index: 0.0027669765788812974
Clusters: 9, Silhouette Score: 0.025769974309891098, Dunn Index: 0.0027669765788812974
Clusters: 10, Silhouette Score: 0.025234384277310347, Dunn Index: 0.0011626704289018072
Clusters: 11, Silhouette Score: 0.026209609899967983, Dunn Index: 0.001162670428901807
Clusters: 12, Silhouette Score: 0.009801450604816794, Dunn Index: 0.0011626704289018072
Clusters: 13, Silhouette Score: -0.004256249594869465, Dunn Index: 0.0011626704289018072
Clusters: 14, Silhouette Score: 0.005945896784059826, Dunn Index: 0.0011662489588934378
Clusters: 15, Silhouette Score: 0.023116669829644842, Dunn Index: 0.0011662489588934378
Clusters: 16, Silhouette Score: 0.022363723487698294, Dunn Index: 0.0011662489588934378
Clusters: 17, Silhouette Score: 0.023797475494919311, Dunn Index: 0.001166248958893437
Clusters: 18, Silhouette Score: 0.018199016399676813, Dunn Index: 0.0011662489588934378
Clusters: 19, Silhouette Score: 0.011831803720645991, Dunn Index: 0.0011662489588934378
Clusters: 20, Silhouette Score: 0.012447888200082764, Dunn Index: 0.0011662489588934378
```

이를 시각화하면 다음과 같다.

[다변량데이터분석][과제5]



군집 개수가 2일 때 실루엣 계수가 가장 높은 것을 알 수 있다. 따라서 complete linkage를 사용했을 때의 최적 군집 개수는 2이다.

single linkage와 동일하게 군집 수가 증가할수록 실루엣 계수가 낮아지는 경향이 나타난다. 다만, complete linkage는 실루엣 계수가 진동하는 모습이 나타난다.

Dunn Index는 군집의 개수가 9개일 때까지는 높은 점수를 유지하나, 군집의 개수가 10개를 넘으면 급격히 낮아지는 것을 확인할 수 있다. 이는 군집 내의 최대 거리와 군집 간의 최소 거리가 비슷해져 군집화가 잘 되지 않았음을 의미하며, 이는 실루엣 계수에서도 확인할 수 있다. 군집의 수가 10개 이상일 때 실루엣 계수가 0에 가까운 것을 알 수 있다.

따라서 군집의 수를 10개 이상으로 설정하는 것은 바람직하지 못하며, 최적의 군집 개수는 2개인 것을 알 수 있다.

complete linkage와의 보다 정확한 비교를 위해 single linkage 역시 군집 개수가 2일 때를 최적으로 판단하고 이후 문항들을 진행하려 한다.

[Q7] [Q6]에서 찾은 최적의 군집 수에 대해서 각 군집들의 변수값의 평균을 이용한 Rader Chart를 도시해보시오. Rader Chart를 바탕으로 판단할 때, K-Means Clustering과 보다 유사한 결과물이 나오는 방식은 어떤 Linkage인지 본인의 생각을 바탕으로 서술해보시오.

먼저 single linkage(K=2)에 대해 가장 빈번하게 발생한 군집화 결과물에 대해서 각 변수들의 평균값을 구하면 다음과 같다.

[다변량데이터분석][과제5]

cluster ID	1	2
BALANCE	-5.762348e-05	0.497579
BALANCE_FREQUENCY	4.285429e-05	-0.370047
PURCHASES	1.213634e-05	-0.104797
ONEOFF_PURCHASES	2.254679e-05	-0.194692
INSTALLMENTS_PURCHASES	-1.268902e-05	0.109570
CASH_ADVANCE	-1.207919e-04	1.043038
PURCHASES_FREQUENCY	-1.154472e-06	0.009969
ONEOFF_PURCHASES_FREQUENCY	-1.701831e-05	0.146953
PURCHASES_INSTALLMENTS_FREQUENCY	-1.391970e-05	0.120197
CASH_ADVANCE_FREQUENCY	-1.668005e-05	0.144032
CASH_ADVANCE_TRX	2.201383e-05	-0.190089
PURCHASES_TRX	-1.824466e-05	0.157543
CREDIT_LIMIT	6.991781e-07	-0.006037
PAYMENTS	4.204703e-05	-0.363076
MINIMUM_PAYMENTS	4.621578e-06	-0.039907
PRC_FULL_PAYMENT	6.227295e-05	-0.537727
TENURE	-4.113268e-05	0.355181

이를 Rader Chart를 통해 시각화하면 다음과 같다.

complete linkage(K=2)에 대해 가장 빈번하게 발생한 군집화 결과물에 대해서 각 변수들의 평균 값을 구하면 다음과 같다.

[다변량데이터분석][과제5]

cluster ID	1	2
BALANCE	-0.350190	0.254779
BALANCE_FREQUENCY	-0.237810	0.173018
PURCHASES	0.287959	-0.209503
ONEOFF_PURCHASES	0.155022	-0.112785
INSTALLMENTS_PURCHASES	0.395873	-0.288016
CASH_ADVANCE	-0.430232	0.313013
PURCHASES_FREQUENCY	0.625454	-0.455046
ONEOFF_PURCHASES_FREQUENCY	0.098742	-0.071839
PURCHASES_INSTALLMENTS_FREQUENCY	0.719952	-0.523798
CASH_ADVANCE_FREQUENCY	-0.617263	0.449087
CASH_ADVANCE_TRX	-0.445426	0.324068
PURCHASES_TRX	0.378310	-0.275238
CREDIT_LIMIT	-0.000587	0.000427
PAYMENTS	-0.024636	0.017924
MINIMUM_PAYMENTS	-0.067731	0.049277
PRC_FULL_PAYMENT	0.328277	-0.238837
TENURE	0.190114	-0.138317

이를 Rader Chart로 시각화하면 다음과 같다.

먼저, 군집화는 서로 다른 속성의 데이터를 군집으로 묶어, 군집 내부는 유사하게, 다른군집과는 상이하게 만드는 것이 중요하다 따라서 K-Means Clustering에서는 가장 상이한 두 군집인 Cluster 2와 Cluster 3가 Cluster 1에 비해 군집화의 목적을 잘 달성하였음을 알 수 있다. 그러므로 hierarchical clustering에서 어떤 linkage가 K-Means Clustering과 더 비슷한 결과를 보였는지는 K-Means Clustering의 Cluster 2와 Cluster 3를 기준으로 평가할 것이다.

Rader Chart의 대략적인 개형을 보면, single linkage의 Cluster 1은 별다른 특징을 보이지 않는 것을 확인할 수 있다. 이는 K-Means Clustering의 Cluster 2, 3와는 다른 모습이며, 오히려 Cluster 1과 더 비슷한 것을 알 수 있다.

complete linkage의 Rader Chart를 보면, Cluster 1은 K-Means의 Cluster 2와, Cluster 2는 K-Means의 Cluster 3와 유사하게, 특징적으로 높은 값을 갖는 변수가 있다는 것을 알 수 있다. 대부분 높

은 값을 갖는 변수와 낮은 값을 갖는 변수가 비슷하며, single linkage는 그렇지 못한 것을 알 수 있다. 따라서 single linkage보다는 complete linkage가 K-Means Clustering과 더 유사하다고 이야기할 수 있다.

이는 complete linkage가 군집을 원형으로 만드는 것을 선호하기 때문에 군집을 원형으로 만들 수밖에 없는 K-Means와 유사할 수밖에 없었을 것이라 해석할 수 있다.

[DBSCAN]

[Q8] DBSCAN 알고리즘의 eps 옵션과 minPts 옵션을 조정해가면서 [Q2]에서 선정한 최적 개수의 군집이 찾아지는 eps 값과 minPts 값을 찾아보시오.

eps 옵션은 0.1부터 2.1까지 0.1씩 증가하는 값, minPts는 2부터 11까지 1씩 증가하는 값을 갖도록 하고 모든 경우에 대해 최적 개수의 군집과 그 때의 noise point 개수, 실루엣 계수를 구할 수 있도록 하였다. 모든 경우에 대해 다음과 같이 계산되는 것을 알 수 있으며, 최적 클러스터 수가 매번 다양해지는 것을 확인할 수 있다.

[다변량데이터분석][과제5]

```
eps: 0.1, minPts: 2, Clusters: 97, Noise: 8274, Silhouette Score: -0.5069691256316601
eps: 0.1, minPts: 3, Clusters: 41, Noise: 8386, Silhouette Score: -0.4689633985711107
eps: 0.1, minPts: 4, Clusters: 22, Noise: 8453, Silhouette Score: -0.4195692899675594
eps: 0.1, minPts: 5, Clusters: 11, Noise: 8503, Silhouette Score: -0.3737304348974619
eps: 0.1, minPts: 6, Clusters: 5, Noise: 8543, Silhouette Score: -0.2750774407861322
eps: 0.1, minPts: 7, Clusters: 3, Noise: 8559, Silhouette Score: -0.25366918294691654
eps: 0.1, minPts: 8, Clusters: 2, Noise: 8569, Silhouette Score: -0.25090328181425725
eps: 0.1, minPts: 9, Clusters: 2, Noise: 8574, Silhouette Score: -0.251245624439190 85
eps: 0.1, minPts: 10, Clusters: 2, Noise: 8574, Silhouette Score: -0.25124562443919085
eps: 0.2, minPts: 2, Clusters: 209, Noise: 7706, Silhouette Score: -0.5070129606682875
eps: 0.2, minPts: 3, Clusters: 72, Noise: 7980, Silhouette Score: -0.46706738803126086
eps: 0.2, minPts: 4, Clusters: 37, Noise: 8112, Silhouette Score: -0.445887720475692
eps: 0.2, minPts: 5, Clusters: 31, Noise: 8164, Silhouette Score: -0.43628664231168524
eps: 0.2, minPts: 6, Clusters: 18, Noise: 8257, Silhouette Score: -0.4182056679674244
eps: 0.2, minPts: 7, Clusters: 13, Noise: 8297, Silhouette Score: -0.4132121908607301
eps: 0.2, minPts: 8, Clusters: 12, Noise: 8322, Silhouette Score: -0.4001150033265487
eps: 0.2, minPts: 9, Clusters: 8, Noise: 8365, Silhouette Score: -0.36424080550423216
eps: 0.2, minPts: 10, Clusters: 7, Noise: 8381, Silhouette Score: -0.3646960904158926
eps: 0.30000000000000004, minPts: 2, Clusters: 325, Noise: 7052, Silhouette Score: -0.47157549959842704
eps: 0.30000000000000004, minPts: 3, Clusters: 110, Noise: 7482, Silhouette Score: -0.47800400639962526
eps: 0.30000000000000004, minPts: 4, Clusters: 49, Noise: 7730, Silhouette Score: -0.42545050596702805
eps: 0.30000000000000004, minPts: 5, Clusters: 38, Noise: 7828, Silhouette Score: -0.4187151511195021
eps: 0.30000000000000004, minPts: 6, Clusters: 31, Noise: 7911, Silhouette Score: -0.4257845119771976
eps: 0.30000000000000004, minPts: 7, Clusters: 21, Noise: 7983, Silhouette Score: -0.412799048354344
eps: 0.30000000000000004, minPts: 8, Clusters: 16, Noise: 8039, Silhouette Score: -0.4041543763287258
eps: 0.30000000000000004, minPts: 9, Clusters: 13, Noise: 8087, Silhouette Score: -0.3967722299164506
eps: 0.30000000000000004, minPts: 10, Clusters: 11, Noise: 8124, Silhouette Score: -0.3939785064119622
eps: 0.4, minPts: 2, Clusters: 331, Noise: 6349, Silhouette Score: -0.4740117883202727
eps: 0.4, minPts: 3, Clusters: 126, Noise: 6759, Silhouette Score: -0.47524039591218253
eps: 0.4, minPts: 4, Clusters: 57, Noise: 7061, Silhouette Score: -0.4628104630159287
eps: 0.4, minPts: 5, Clusters: 38, Noise: 7227, Silhouette Score: -0.4060975624883885
eps: 0.4, minPts: 6, Clusters: 24, Noise: 7337, Silhouette Score: -0.3803019718022026
eps: 0.4, minPts: 7, Clusters: 15, Noise: 7434, Silhouette Score: -0.3331526641800845
eps: 0.4, minPts: 8, Clusters: 18, Noise: 7478, Silhouette Score: -0.37169835085049396
eps: 0.4, minPts: 9, Clusters: 19, Noise: 7517, Silhouette Score: -0.37855788346483277
eps: 0.4, minPts: 10, Clusters: 15, Noise: 7621, Silhouette Score: -0.3641765662492606
eps: 0.5, minPts: 2, Clusters: 336, Noise: 5529, Silhouette Score: -0.515994244752116
eps: 0.5, minPts: 3, Clusters: 128, Noise: 5945, Silhouette Score: -0.5244811903255837
eps: 0.5, minPts: 4, Clusters: 62, Noise: 6276, Silhouette Score: -0.5102458511113763
eps: 0.5, minPts: 5, Clusters: 36, Noise: 6488, Silhouette Score: -0.465117752115066
eps: 0.5, minPts: 6, Clusters: 20, Noise: 6652, Silhouette Score: -0.40228690465403466
eps: 0.5, minPts: 7, Clusters: 16, Noise: 6768, Silhouette Score: -0.3854519149634 75
eps: 0.5, minPts: 8, Clusters: 11, Noise: 6877, Silhouette Score: -0.3189461860956522
eps: 0.5, minPts: 9, Clusters: 12, Noise: 6938, Silhouette Score: -0.3259039450623435
eps: 0.5, minPts: 10, Clusters: 6, Noise: 7031, Silhouette Score: -0.26599173621699557
eps: 0.6, minPts: 2, Clusters: 306, Noise: 4785, Silhouette Score: -0.5396431150686052
eps: 0.6, minPts: 3, Clusters: 102, Noise: 5193, Silhouette Score: -0.5317048497633893
eps: 0.6, minPts: 4, Clusters: 45, Noise: 5475, Silhouette Score: -0.5081738124853403
eps: 0.6, minPts: 5, Clusters: 38, Noise: 5650, Silhouette Score: -0.5026395506643591
eps: 0.6, minPts: 6, Clusters: 17, Noise: 5863, Silhouette Score: -0.43789365194847857
eps: 0.6, minPts: 7, Clusters: 17, Noise: 5975, Silhouette Score: -0.4417041338873504
eps: 0.6, minPts: 8, Clusters: 15, Noise: 6102, Silhouette Score: -0.3636115713918533
eps: 0.6, minPts: 9, Clusters: 11, Noise: 6220, Silhouette Score: -0.39536675777230673
eps: 0.6, minPts: 10, Clusters: 8, Noise: 6331, Silhouette Score: -0.28446547852941273
eps: 0.7000000000000001, minPts: 2, Clusters: 311, Noise: 4113, Silhouette Score: -0.5186455320788802
eps: 0.7000000000000001, minPts: 3, Clusters: 93, Noise: 4549, Silhouette Score: -0.516547422399089
eps: 0.7000000000000001, minPts: 4, Clusters: 48, Noise: 4793, Silhouette Score: -0.4962803943320141
eps: 0.7000000000000001, minPts: 5, Clusters: 25, Noise: 4991, Silhouette Score: -0.43770103238013724
```

최적의 군집 수가 3인 조합이 여럿 존재하였지만, 그 중에서 실루엣 계수가 가장 높은 조합을 선정하였다.

	eps	minPts	n_clusters	n_noise	silhouette_score
120	1.4	6	3	1600	0.252571

그때의 eps는 1.4이며, minPts는 6인 것을 알 수 있다. eps 값이 비교적 작기 때문에, 현재 DBSCAN 알고리즘이 작은 반경 내에서만 이웃을 고려하여 군집을 형성하고 있다는 것을 알 수 있으며, 이는 데이터가 고밀도 구역에서 군집화되는 경향이 있기 때문으로 해석할 수 있다. minPts 역시 값이 크지 않기 때문에 minPts가 커지면 noise 수가 높아질 것이라 예측할 수 있으며, noise 데이터와 정상 데이터가 쉽게 구분되지 않는 위치에 존재할 가능성을 엿볼 수 있다.

[다변량데이터분석][과제5]

[Q9] [Q8]에서 찾은 군집화 결과물에서 Noise로 판별된 객체의 수가 몇 개인지 확인해 보시오.

최적의 군집 수가 3인 조합 중 실루엣 계수가 가장 높은 조합은 다음과 같다.

	eps	minPts	n_clusters	n_noise	silhouette_score
120	1.4	6	3	1600	0.252571

noise로 판별된 객체의 수가 1600개인 것을 확인할 수 있다. 상당히 많은 수의 점이 noise로 판별되었으며, eps와 minPts의 기준이 엄격한 편임을 알 수 있다. eps를 늘리거나 minPts를 줄인다면 상대적으로 noise가 줄어들 수 있을 것이다. 그러나, 만약 진짜 noise 데이터가 포함된다면 군집화 결과 활용에 있어 신뢰성이 떨어질 수 있다.

현재는 실루엣 계수가 약 0.25로 높지 않기 때문에, eps 값을 증가시키거나 minPts 값을 줄여 실루엣 계수를 높이는 것을 꾀해볼 수 있다.

[종합]

[Q10] 이 데이터셋에 가장 적합한 군집화 알고리즘은 무엇이라고 생각하는지 본인이 생각한 근거를 이용하여 서술하시오.

가장 먼저, 실루엣 지표를 기준으로 하면 hierarchical clustering이 가장 적절하지 않다고 볼 수 있다. 나머지 두 클러스터링 방법은 약 0.25의 점수를 보였으나, hierarchical clustering의 경우, 실루엣 지표 값이 음수, 혹은 약 0.14 정도로 계산된 것으로 보아, 가장 좋지 않은 성능을 보이고 있음을 알 수 있다. 또한, 이는 hierarchical clustering을 할 경우, 많은 데이터들이 잘못된 군집에 속할 수 있다는 것을 시사한다.

게다가, K-Means와 DBSCAN을 비교해보면, 최적의 군집 수가 다를 때 DBSCAN이 더 높은 실루엣 계수를 가질 수 있음을 확인할 수 있었다. 이는 K-Means가 구형의 군집을 찾는 데 최적화되어 있고, 노이즈나 이상치를 제거할 수 없기 때문에, 노이즈나 이상치가 함께 있는 실제 데이터에 대해서는 성능이 떨어지기 때문으로 해석할 수 있다. 그와는 달리, DBSCAN은 다양한 형태의 군집을 찾을 수 있으며, 하이퍼파라미터 조정을 통해 노이즈와 이상치에 더 강건한 성능을 보일 수 있다.

따라서 DBSCAN의 하이퍼파라미터를 잘 조정하면 noise point를 줄이면서도 실루엣 계수를 높일 수 있음을 알 수 있다. 이를 통해 DBSCAN이 K-Means보다 높은 수준의 성능에 도달할 수 있으면서, noise로 예상되는 포인트를 제거함으로써 일반화 성능 역시 더 높일 수 있을 것으로 기대된다. 그렇기 때문에 본 데이터셋에 대해서는 DBSCAN이 가장 적절한 군집화 알고리즘일 것이라 생각한다.