Multivariate Data Analysis Assignment #5

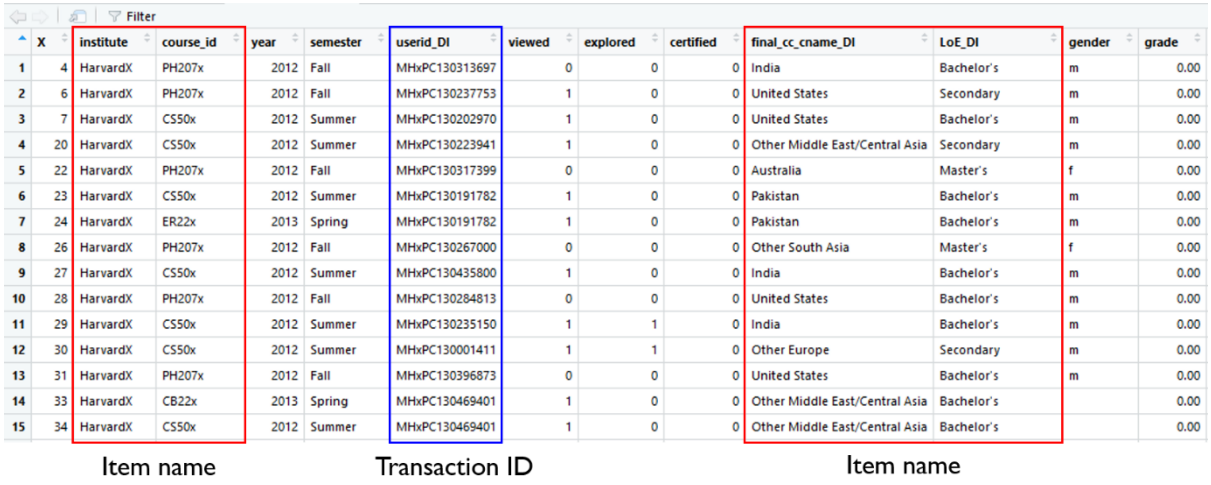
[Part 1: Association Rule Mining]

Dataset: MOOC Dataset (big_student_clear_third_version.csv)

해당 데이터셋은 MOOC 강좌를 수강한 수강생들에 대한 정보가 포함되어 있는 데이터 셋이다. 다음 각 Instruction에 따라 데이터를 변환하고 연관규칙분석을 수행하여 각 결과물을 제시하고 적절한 해석을 제공하시오.

[Step 1] 데이터 변환

[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.



Filter

	X	institute	course_id	year	semester	userid_DI	viewed	explored	certified	final_cc_cname_DI	LoE_DI	gender	grade
1	4	HarvardX	PH207x	2012	Fall	MHxPC130313697	0	0	0	India	Bachelor's	m	0.00
2	6	HarvardX	PH207x	2012	Fall	MHxPC130237753	1	0	0	United States	Secondary	m	0.00
3	7	HarvardX	CS50x	2012	Summer	MHxPC130202970	1	0	0	United States	Bachelor's	m	0.00
4	20	HarvardX	CS50x	2012	Summer	MHxPC130223941	1	0	0	Other Middle East/Central Asia	Secondary	m	0.00
5	22	HarvardX	PH207x	2012	Fall	MHxPC130317399	0	0	0	Australia	Master's	f	0.00
6	23	HarvardX	CS50x	2012	Summer	MHxPC130191782	1	0	0	Pakistan	Bachelor's	m	0.00
7	24	HarvardX	ER22x	2013	Spring	MHxPC130191782	1	0	0	Pakistan	Bachelor's	m	0.00
8	26	HarvardX	PH207x	2012	Fall	MHxPC130267000	0	0	0	Other South Asia	Master's	f	0.00
9	27	HarvardX	CS50x	2012	Summer	MHxPC130435800	1	0	0	India	Bachelor's	m	0.00
10	28	HarvardX	PH207x	2012	Fall	MHxPC130284813	0	0	0	United States	Bachelor's	m	0.00
11	29	HarvardX	CS50x	2012	Summer	MHxPC130235150	1	1	0	India	Bachelor's	m	0.00
12	30	HarvardX	CS50x	2012	Summer	MHxPC130001411	1	1	0	Other Europe	Secondary	m	0.00
13	31	HarvardX	PH207x	2012	Fall	MHxPC130396873	0	0	0	United States	Bachelor's	m	0.00
14	33	HarvardX	CB22x	2013	Spring	MHxPC130469401	1	0	0	Other Middle East/Central Asia	Bachelor's		0.00
15	34	HarvardX	CS50x	2012	Summer	MHxPC130469401	1	0	0	Other Middle East/Central Asia	Bachelor's		0.00

Item name　　Transaction ID　　Item name

[Step 2] 데이터 불러오기 및 기초 통계량 확인

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

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

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

[Step 3] 규칙 생성 및 결과 해석

[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			
…			

[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?

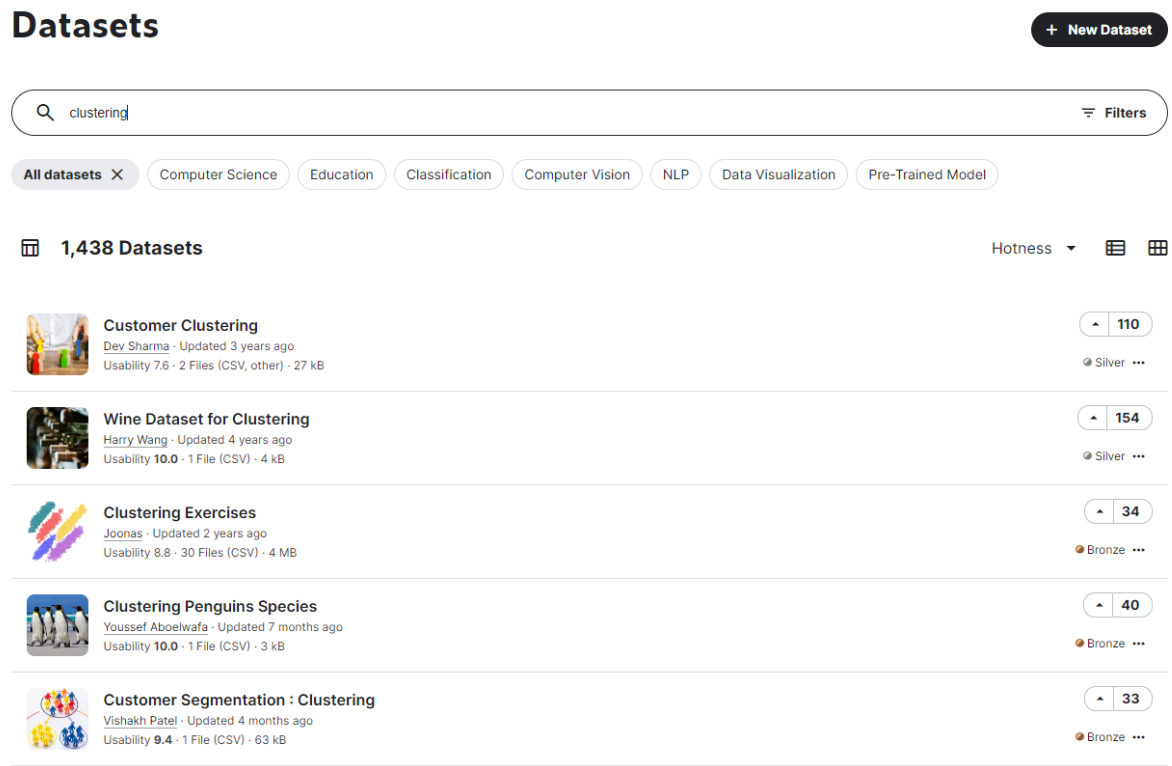
[Extra Question] 이 외 수업 및 실습 시간에 다루지 않은 연관규칙분석 시각화 및 해석을 시도해 보시오.

[Part 2: Clustering]

Dataset: Kaggle Clustering 데이터셋 중 1개 선택

Kaggle 사이트의 Datasets 항목에서 “clustering”을 키워드로 검색하면 총 1,438개의 데이터셋이 아래와 같이 검색됩니다(2024-05-26 기준).

https://www.kaggle.com/datasets?search=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.

[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?

[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.

[Q4] [Q3]에서 가장 빈번하게 발생한 군집화 결과물에 대해서 각 군집별 변수들의 평균값을 이용한 Radar Chart를 도시해보시오. Radar 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.

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

[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?

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

[DBSCAN]

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

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

[종합]

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