**데이터 전처리 과정**

**<코드>**

텍스트이(가) 표시된 사진

자동 생성된 설명

위 코드는 Association Rule Mining을 시작하기 전, 데이터를 전처리 하기 위해 사용된 코드이다. 본래 데이터는 MOOC 이라는 온라인 공개수업 켜뮤니티에서 강의를 수강한 학생들의 정보가 있는 데이터이다. 본래 데이터는 416921개의 데이터와 총 22개의 변수로 이루어져있다. 이때 본 데이터에서 연관분석을 위해 모든 변수들이 필요한 것이 아니기때문에 위 코드를 통해 주요 변수들만 추출을 해온다. 이를 통해 최종적으로는 연관분석을 하기 위한 single data를 만들기 위함에 목적이 있다.

<1단계>

Transaction\_ID, Institute, Course, Region, 그리고 Degree 변수들을 본래 데이터의 변수(열)들을 가져와 저장해주었다. 이때 Institute는 강좌 제공 기관, Course는 강좌 코드, Region은 접속국가, 그리고 Degree는 학생의 학위과정을 보여준다.

<2단계>

1단계에서 생성한 변수들 중, Region은 각 데이터 마다 공백이 존재하기 때문에 gsub함수를 통해 공백을 없애주었다. 아래와 같이 본래 띄어쓰기가 존재한 것을 없애주었다.



<3단계>

3단계에서는 1단계에서 정의한 Institute, Course, Region, Degree 변수들을 하나의 “Raw Transaction”이라는 이름의 변수로 저장함과 동시에, 각 변수들을 구분할 수 있도록 “\_”로 구분하여 paste 함수를 통해 저장해주었다.

<4단계>

3단계에서 정의한 RawTransaction과 Transaction\_ID 변수를 하나의 변수 “MOOC\_transaction”으로 저장하기 위해 paste함수와 sep= “ “를 통해 두 변수간의 공백을 만들어 저장해주었다.

<5단계>

마지막으로 앞서 생성한 MOOC\_transaction을 write.csv함수를 사용하여 csv 파일로 만들어 주었다. 이 모든 절차는 앞서 강조하였듯 연관분석을 하기 위한 새로운 데이터를 구축하는데 목적이 있으며, 최종 결과는 아래와 같다.

테이블이(가) 표시된 사진

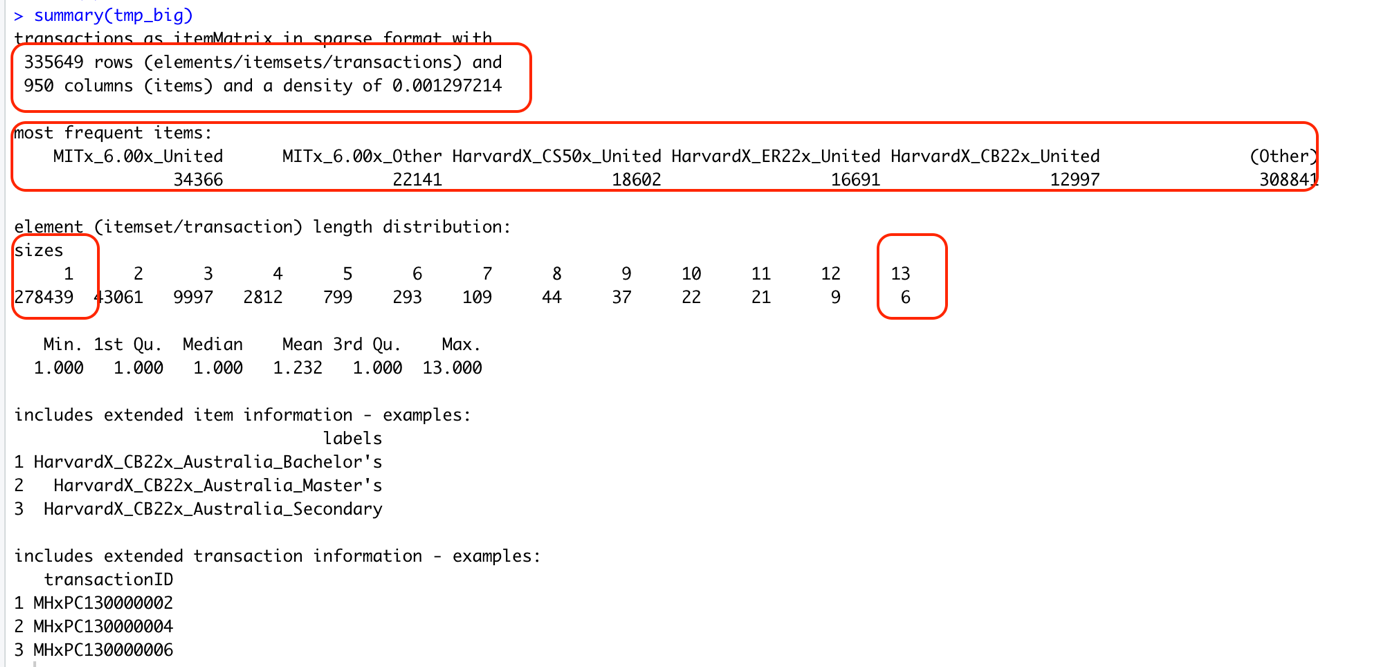
자동 생성된 설명좌측의 그림은 [Q1]에서 생성한 최종 데이터이다. 이때 첫번째 인덱스는 MHxPC1303131697의 결제 ID, 제공기관은 Harvard, 강좌코드는 PH207x, 접속 국가는 인도이며 학위는 Bachelor임을 나타낸다. 이와 같이 모든 데이터가 해당 강좌를 들은 학생의 정보를 포함 데이터임을 알 수 있다.

**<코드>**

**텍스트이(가) 표시된 사진

자동 생성된 설명**

Q1에서 최종적으로 구축한 데이터를 불러오기 위해 read.transactions함수를 통해 불러주었다. 이때 read.transactoin 함수에 format=”single”은 해당 데이터의 각 행들이 결제 id와, 나머지 필요 정보들을 가지고 있는 single 데이터 형태로 불러옴을 의미한다. 아래는 tmp\_big으로 저장한 transaction 데이터의 summary함수를 적용한 결과이다.

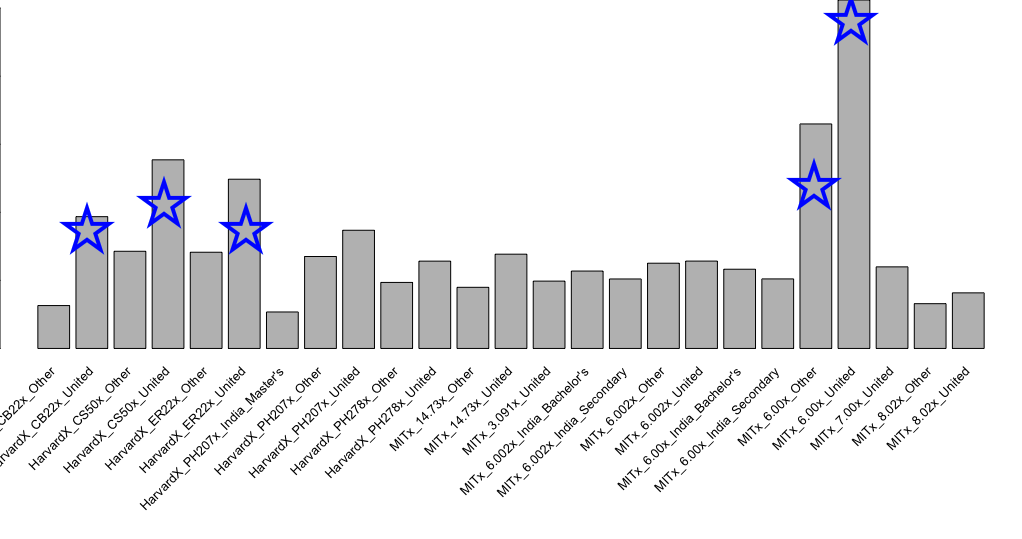


위 결과를 통해 해당 데이터는 총 335649개의 데이터를 transaction 데이터를 가지고 있으며 최소 1한번 이상 강좌를 시청한 학생의 데이터가 950개가 존재함을 알 수 있다. 또한 Density는 0이 아닌 1의 숫자를 가진 셀의 개수가 총 전체 데이터의 0.1%임을 나타낸다. Most frequent times 부분을 보면, MITx\_6.00x\_United는 미국에서 MIT에서 진행하는 6.00 강좌를 수강하는 결제 수가 34366로 가장 가장 많은 것을 알 수 있으며, 하나의 강의만 수강한 학생의 수가 278349 명, 13개의 강의를 들은 학생 수는 13명이 존재하는 것을 보여준다.

텍스트이(가) 표시된 사진

자동 생성된 설명위 그림은 wordcloud함수를 통해 구축한 transaction 데이터에 대한 wordcloud 출력결과이다. 이때 min frequency는 500으로 설정해주었는데, 즉 transaction 데이터 가 최소 500번이상 존재한 데이터들만 wordcloud에 나타난다는 것을 의미한다. 만약 min frequency가 적은 값을 가졌더라면, wordcloud생성이 불가할 뿐만 아니라 너무나도 많은 transaction 데이터들이 포함될 것이다. 또한 col함수를 정의하여 brewer palette에서 제공하는 11개의 spectral 색들을 사용하여 표시해 주었으며, scale함수는 가장 많이 등장하는 데이터의 사이즈를 2로 설정해주고 가장 적게 등장한 데이터를 0.2의 사이즈로 출력해주었음을 의미한다.

위 wordcloud를 보면 앞서 summary함수를 통해 봤듯이 MITx\_6.00x\_United에 대한 데이터가 가장 많이 존재함을 볼 수 있으며, 그 뒤로 MITx\_6.00X\_Other이 가장 큰 사이즈를 갖는데, 이를 통해 전반적으로 moot에서 제공하는 강의들 중 MITx\_6.00x를 제공하는 학생들이 가장 많이 수강한다는 정보를 준다. Wordcloud의 장점으로는 글자의 사이즈를 통해 데이터들의 분포를 상대적으로 비교는 가능하지만, 절대적으로 해당 데이터가 실질적으로 얼마나 존재하는 정보의 부재와 데이터가 많아질 수록 wordcloud의 글자 사이즈가 작아져 식별하기 힘든 것을 단점으로 꼽을 수 있겠다.



위 itemFrequencyPlot은 support=0.01을 설정하여 구축한 것이다. 이때 support=0.01은 전체 데이터들 중 빈도가 1%이상 인 데이터들만 출력하는 것으 의미한다. 앞서 wordcloud와 summary 함수와 동일하게 MITx\_6.00x\_United가 다른 transaction id 들에 비해 우월하게 많이 있음을 알 수 있다. 위 결과에서 별표 표시를 넣어준 막대 그래프들은 가장 접속이 많은 상위 5개 데이터들을 표시한 것이다. 가장 많이 접속한 transaction id는 ‘MITx\_6.00x\_United’ 로 미국이며, 그 외 데이터들을 보면 접속 국가가 Other과 United States이다. 이는 즉 moot를 수강하는 대부분의 학생들이 미국에 거주 중임과 동시에 나머지 국가들은 미미하기 때문에 others로 묶여 표시가 되었다고 할 수 있겠다.



위 코드는 support와 confidence를 변형해가면서 데이터에서 존재하는 규칙들의 생성 개수를 확인하기 위한 것이다. 이때 confidence와 support는 아래와 같은 값들로 설정해주었다

|  |  |
| --- | --- |
| Support | 0.0005, 0.001, 0.0015 |
| Confidence | 0.05, 0.01, 0.015 |

Support는 데이터의 상대적 지지도/빈도를 의미하는데, 이는 조건절과 결과절이 동시에 일어날 최소 확률을 의미한다. 만약 0.0005로 설정하였다면, 조건절과 결과절이 동시에 발생할 확률이 0.0005이상이 되어야 연관성으로 인정을 해준다는 것이다. Confidence는 향상도인데, 이는 조건절이 발생하였을 때 결과절이 발생할 확률을 의미한다. 위 표에서 나타나는 support와 confidence 값들을 이용하여 총 9개의 조합을 만들어 aprirori함수를 통해 연관분석을 실행해주었으며, 결과는 아래와 같다.

테이블이(가) 표시된 사진

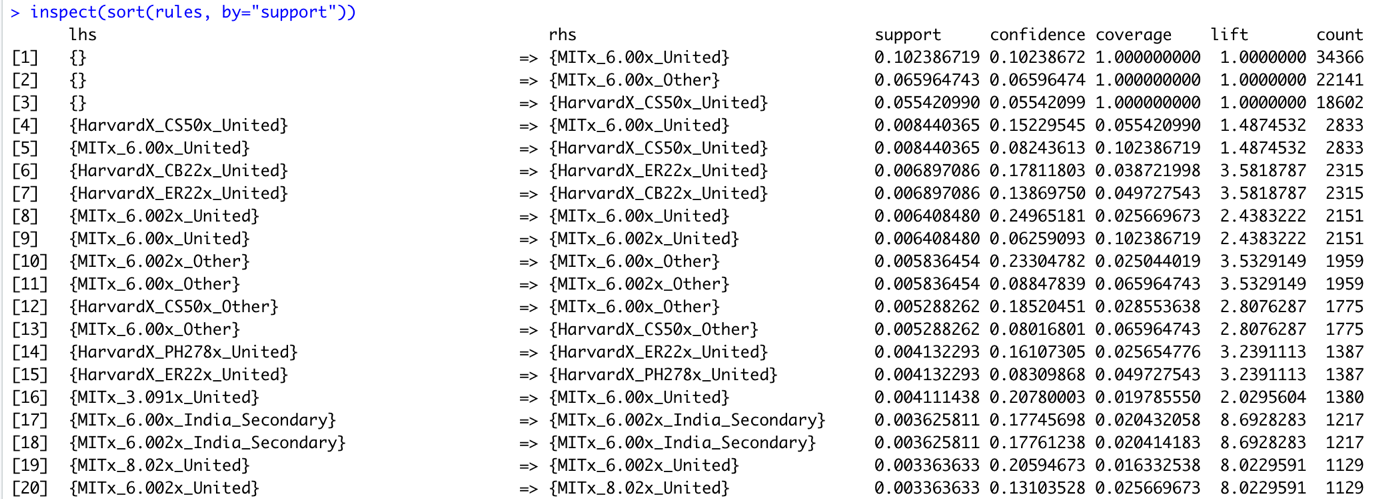
자동 생성된 설명

Classification과 regression과는 달리 연관분석에서는 성능을 평가할 기준이 존재하지 않는다. 그렇기때문에 어떤 confidence와 support 값의 조합이 최적이라고 할 수 없다. 위 표는 각 파라미터 조합으로 생성한 연관분석을 통해 나타나는 규칙의 개수이다. 규칙이 가장 많이 나타나는 조합은 confidence=0.01 & support=0.0005 이며 가장 규칙이 적은 것은 confidence=0.05 & support=0.0015 조합이다. 전반적으로 해당 표를 통해 support 값이 증가할 수록 규칙의 개수가 줄어드는 것을 확인 할 수 있었다. 이는 support 값이 증가함에 따라 규칙으로 인정하는 기준이 더 까다로워졌기 때문이라고 할 수 있겠다.

Support=0.001 그리고 confidence=0.05로 연관분석을 실행해준 결과를 각기 다른 기준을 통해 정렬을 해준다.

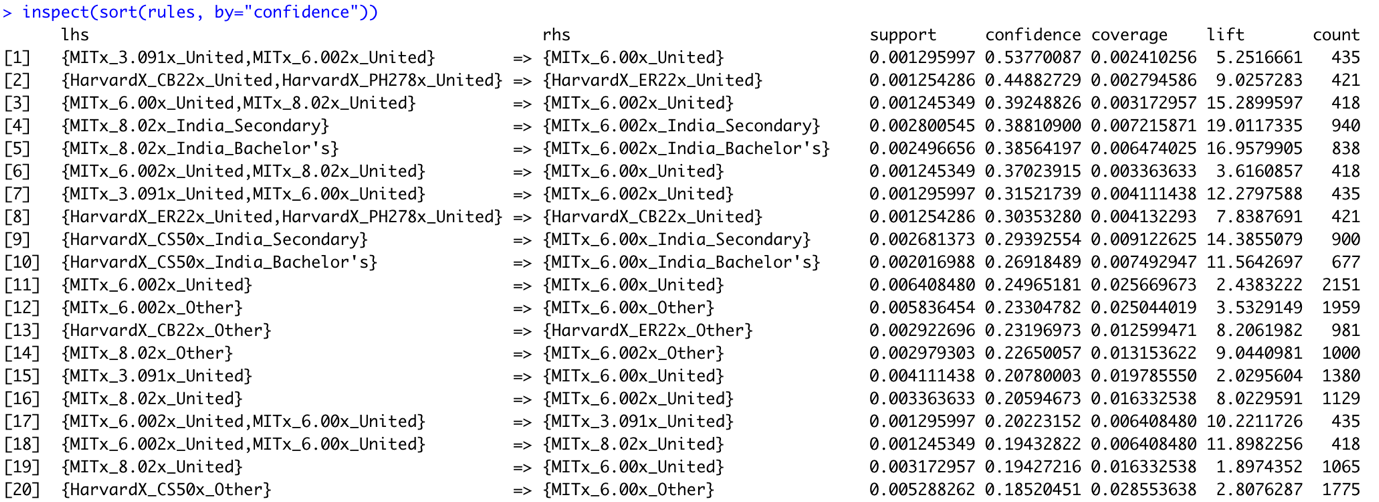
**<코드>**

**<Support가 가장 높은 규칙>**



위 결과는 연관분석을 support를 기준으로 정렬해준 결과이다. 이때 첫 세개의 인덱스는 조건절이 아무것도 없음으로, 조건절과 결과절의 연관성이 존재하지 않고 그저 빈도수가 가장 높은 item들을 나타낸다. 그럼으로 4번째 인덱스부터 실질적으로 조건절과 결과절의 연관성이 존재한다고 할 수 있다. Support가 가장 높게 나타난 규칙은 HarvardX\_CS50x\_united 과MITx\_.6.00x\_United의 규칙이다. 이때 support는 0.008으로 두 item을 동시에 수강할 확률이며, confidence는 0.15로 HarvardX\_CS50x\_united를 수강했을 때 MITx\_.6.00x\_United를 수강할 확률이다.

**<Confidence가 가장 높은 규칙은 무엇인가?>**



위 결과는 연관분석을 confidence를 기준으로 내림차순 정렬을 해준 결과이다. Confidence가 가장 높은 규칙은 MITx\_3.091x\_United, MITx\_6.002x\_United 과 MITx\_6.00x\_United이다. 이때 confidence는 0.53으로 MITx\_3.091x\_United, MITx\_6.002x\_United를 수강하는 학생이 MITx\_6.00x\_United를 동시에 수강할 확률을 나타내며, 해당 규칙은 총 435번 발생하였다. Lift는 5.25로, 이는 MITx\_3.091x\_United, MITx\_6.002x\_United , MITx\_6.00x\_United 각 과목을 따로 수강하는 확률보다 MITx\_3.091x\_United, MITx\_6.002x\_United 를 수강하는 학생이 MITx\_6.00x\_United를 같이 수강할 확률이 5.2배 많은 것을 알려준다.

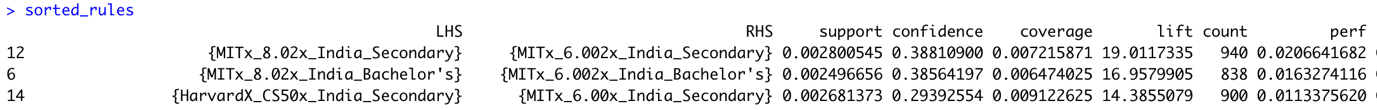
**<Lift가 가장 높은 규칙은 무엇인가?>**

텍스트이(가) 표시된 사진

자동 생성된 설명

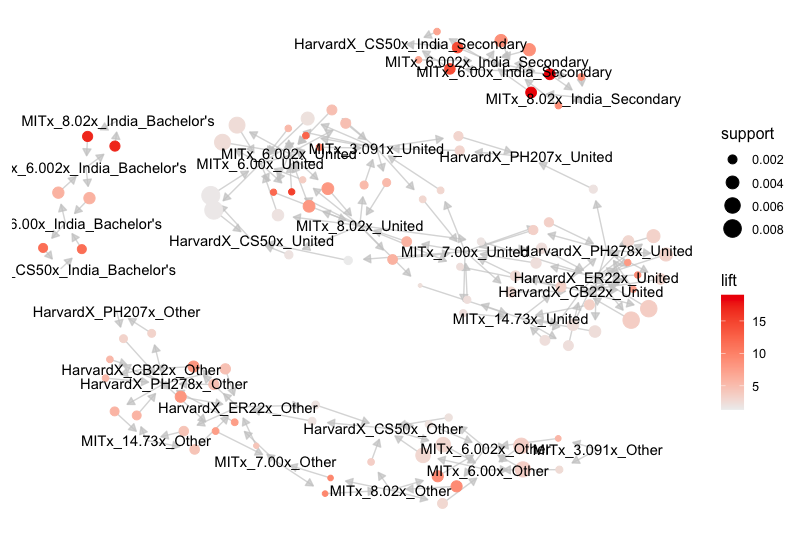
위 결과는 규칙들을 lift를 기준으로 내림차순 정렬해준 것이다. 이때 가장 높은 lif빈도를 기록한 규칙은 MITx\_6.002x\_India\_Secondary와 MITx\_8.02x\_India\_Secondary 간의 규칙이다. 해당 규칙은 lift가 19.01이고 총 940회 나타난 규칙이다. 이때 lift가 19.01는 MITx\_6.002x\_India\_Secondary와 MITx\_8.02x\_India\_Secondary 각 과목을 독립적으로 수강할 확률보다 둘을 같이 수강할 확률이 19.01배 많다는 것을 의미한다.

**<Support \* Confidence \* Lift가 가장 높은 규칙>**

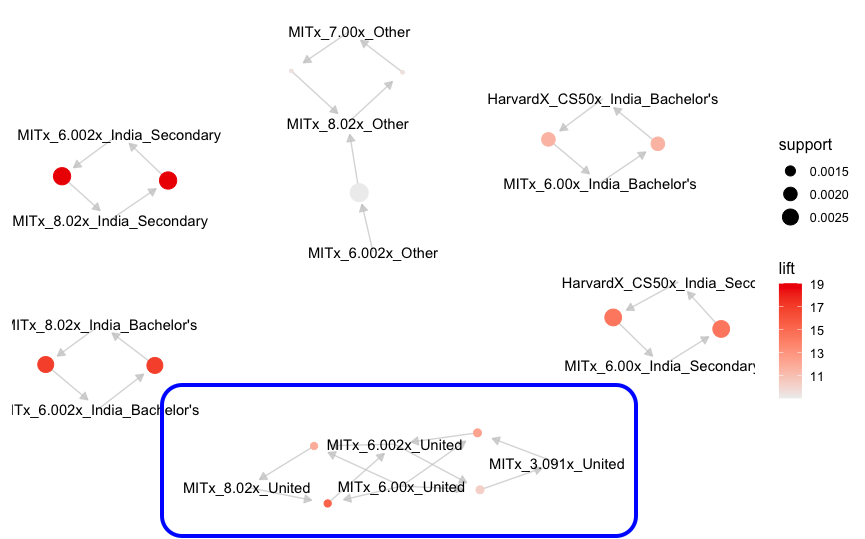


위 결과는 연관분석을 Support\*Confidence\*Lift “perf”를 기준으로 정렬해준 결과들 중 상위 1~3위들이다. 앞서 다른 결과들은 하나의 기준만으로 정렬을 해준 것이지만, 위 결과는 3가지 기준을 모두 한번에 평가를 해준 것이다. 이는 즉 support, confidence, lift가 모두 높은 것을 고려해 가장 연관성이 높은 연관규칙들을 찾는다는 것이다. 이때 가장 연관성이 높은 규칙은 MITx\_.8.02x\_India\_Secondary와 MITx\_6.002x\_India\_Secondary으로 perf=0.02이고, 2위는 MITx\_.8.02x\_India\_Bachelor’s와 MITx\_6.002x\_India\_Bachelor’s, perf=0.16, 마지막으로 3위는 HarvardX\_CS50x\_India\_Secondary와 MITx\_6.00x\_India\_Secondary 규칙이고 perf=0.011이다.

**<연관관계 visualization>**

****

위 그림은 plot함수에서 graph형태를 사용하여 데이터에 존재하는 연관관계를 시각화해준 결과이다. 이때 위 그림을 보면, 루프의 형태를 같는 연관관계를 볼 수 있는데, 이는 X -> Y 규칙과 Y->X이 동시에 존재한다는 것을 의미한다. 모든 규칙을 도출한 결과에선 이 관계를 완벽하게 식별하기 어렵기 때문에 15개의 결과만을 뽑아 새로 시각화를 해주었다.



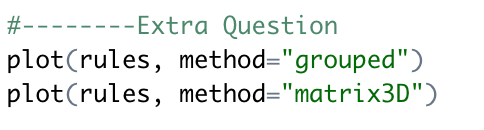
위 결과들 중 파란사각형으로 표시한 부분은 총 4가지 규칙들이 서로 얽혀있는 것을 볼 수 있다. 이때 아래의 규칙들에 대하여 조건절과 결과절을 바꿔가며 지표들에 변화를 측정해 보았다.

**<MITx\_6.002x\_United, MITx\_3.091x\_United, MITx\_6.00x\_United>**

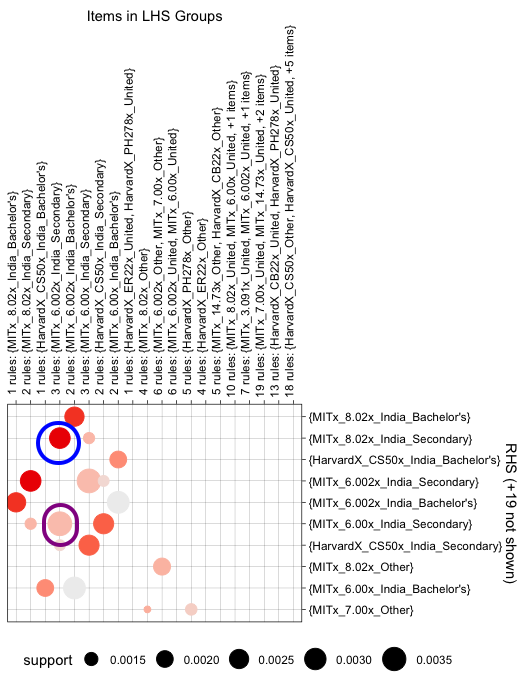
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **조건절** | **결과절** | **Support** | **Confidence** | **Lift** |
| MITx\_6.002x\_United | MITx\_3.091x\_United | 0.0024 | 0.093 | 4.75 |
| MITx\_6.002x\_United | MITx\_6.00x\_United | 0.006 | 0.249 | 2.438 |
| MITx\_3.091x\_United | MITx\_6.002x\_United | 0.0024 | 0.128 | 4.74 |
| MITx\_3.091x\_United | MITx\_6.00x\_United | 0.004 | 0.2078 | 2.029 |
| MITx\_6.00x\_United | MITx\_6.002x\_United | 0.006 | 0.06 | 2.438 |
| MITx\_6.00x\_United, MITx\_6.002x\_United | MITx\_3.091x\_United | 0.00129 | 0.202 | 10.22 |

위 결과는 MITx\_6.002x\_United, MITx\_3.091x\_United, MITx\_6.00x\_United를 조건절과 결과절으로 두고 바꿔가며 해당 구매내역들 간의 연관관계를 살핀 결과이다. 구매내역의 구분을 편리해주게 하기 위해 **MITx\_6.002x\_United=”1”, MITx\_3.091x\_United=”2”, MITx\_6.00x\_United=”3**” 과 같이 칭해주겠다. 동일한 구매내역들 간의 연관관계를 비교하였음에도 조건절과 결과절이 바뀜에 따라 support, confidence, 그리고 lift의 값들이 바뀌는 것을 볼 수 있었다. 이는 예를 들어 맥주를 사서 과자를 안주로 사는 경우는 있지만, 반대로 과자를 산다고 해서 맥주를 사는 결과가 나타나지 않는 것과 같은 현상이 존재해서이다. 따라서 조건절이 1이고 결과절이 3일 때 confidence는 0.249이지만, 반대로 조건절이 3이고 결과절이 1일 때 confidence=0.06으로 나타는 것이다. 이는 즉 1이 발생할 때 2가 발생할 확률은 높으나, 2가 발생한다해서 1이 발생할 확률은 미미하다는 것을 보여준다. 더하여 표의 마지막 행을 보면 2는 1과3이 동시에 발생해야만 발생하는 것을 알 수 있다. 즉 3만 발생한다해서 2가 발생하지는 않음을 보여준다.

**<코드>**



**[method=”grouped”]**



위 그림은 plot함수에서 grouped 메소드를 활용하여 규칙들의 관계를 보여준다. 이때 matrix안에 각각 색이 다른 원들이 보이는데, 원의 사이즈는 규칙들 간의 support수치를 보여주며, 원의 색은 lift를 나타낸다. 이때 원의 색이 진할수록(더 빨갈수록) 규칙간의 lift가 높음을 의미한다.

이때 색이 가장 진한 규칙을 파란원으로 표시하였는데, 해당 규칙은 조건절이 MITx\_6.002x\_India\_Secondary이고 결과절이 MITx\_8.02x\_India\_Secondary이다. 이는 앞서 Q3-2에서 lift로 정렬한 결과와 일치함을 알 수 있다. 또한 원이 큰 규칙을 보라색 원으로 표시하였는데, 이는 Q3-2에서 support로 정렬한 결과들 중 7번째 높은 support를 기록한 규칙이었다. 해당 시각화 방법의 단점은 만약 더 많은 거래내역이 들어간다면 이를 모두 matrix안에 표시하기는 힘들다는 점과, 색과 원의 크기는 상대적 비교만 해줄뿐 절대적인 수치를 알 수 없다는 점이다.

**[method=”3D”]**

**안테나이(가) 표시된 사진

자동 생성된 설명**

위 결과는 plot 함수에서 method=matrix3D로 설정해주어 도출한 결과이다. 해당 시각화방법은 규칙들 간의 lift 지표를 표현한다. 이때 matrix3D는 x축이 조건절, y축이 결과절 그리고 z축을 lift로 두는 3차원의 그래프를 구축해준다. 이때 3차원에 그려지는 규칙의 z축의 크기가 클 수록 규칙간의 lift가 높음을 의미한다. 해당 시각화모델의 단점은 조건절과 규칙절의 labeling이 존재하지 않아 어떤 조건절과 결과절의 규칙을 표시하는지를 알 수 없다는 점이다.