다변량분석 7번째 과제

Assignment 7

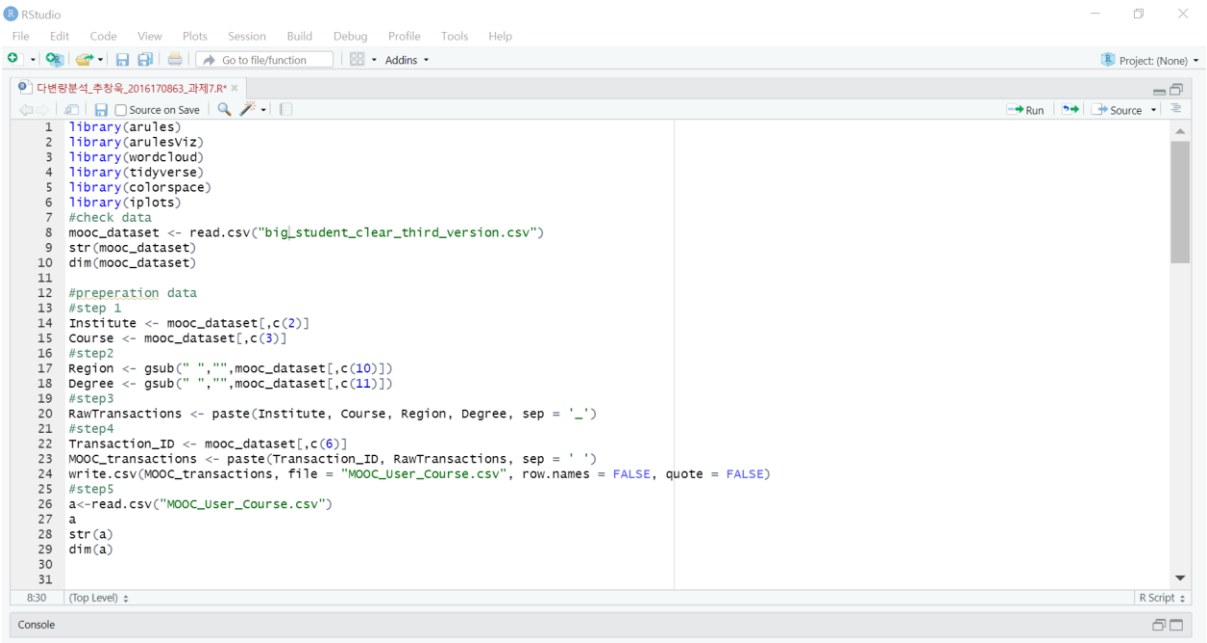
Association Rule Mining

학과 산업경영공학부

학번 2016170863

이름 추창욱

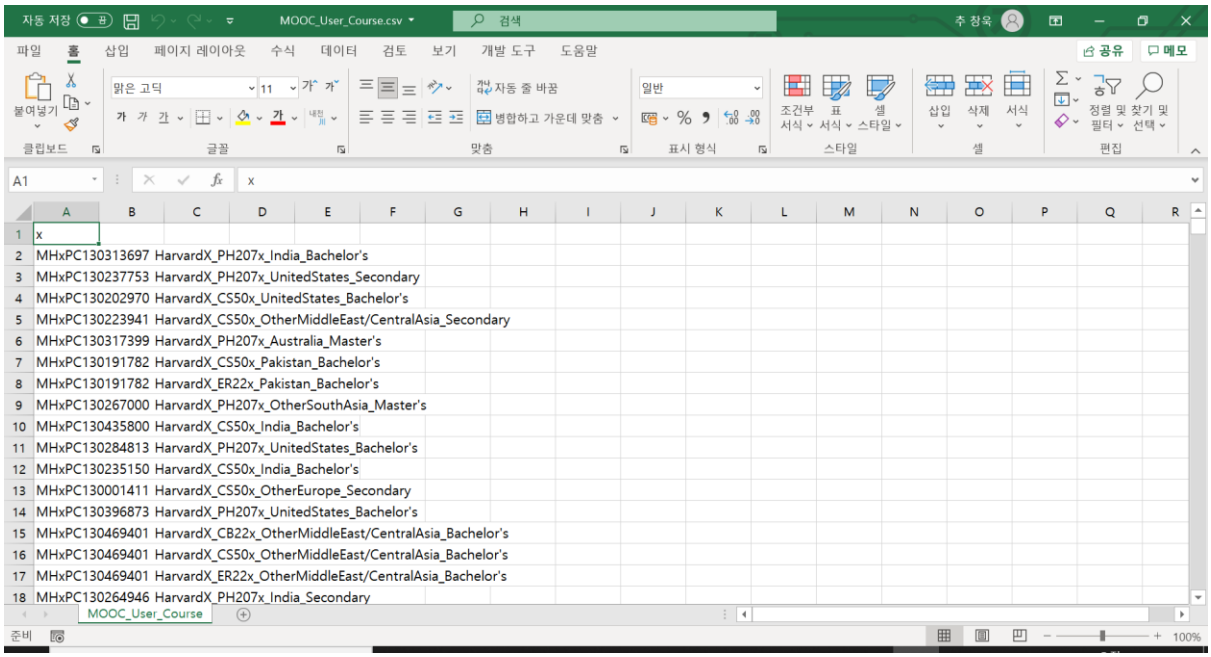
[Step 1] 데이터 변환



```
library(arules)
library(arulesViz)
library(wordcloud)
library(tidyverse)
library(colorspace)
library(iplots)
#check data
mooc_dataset <- read.csv("big_student_clear_third_version.csv")
str(mooc_dataset)
dim(mooc_dataset)

#preperation data
#step 1
Institute <- mooc_dataset[,c(2)]
Course <- mooc_dataset[,c(3)]
#step2
Region <- gsub(" ","",mooc_dataset[,c(10)])
Degree <- gsub(" ","",mooc_dataset[,c(11)])
#step3
RawTransactions <- paste(Institute, Course, Region, Degree, sep = '_')
#step4
Transaction_ID <- mooc_dataset[,c(6)]
MOOC_transactions <- paste(Transaction_ID, RawTransactions, sep = ' ')
write.csv(MOOC_transactions, file = "MOOC_User_Course.csv", row.names = FALSE, quote = FALSE)
#step5
a<-read.csv("MOOC_User_Course.csv")
a
str(a)
dim(a)
```

가장 먼저 분석을 진행할 데이터를 불러왔다. 이 데이터는 mooc_dataset에 csv 파일을 불러 저장 해주었다. 우선 가이드라인에 제시된 순차대로 진행을 하여 데이터를 준비해주는 과정을 거쳤는데, step 1(주석 기준)은 새로운 변수명으로 데이터의 칼럼인덱스를 이용하여 저장하였다. 이후 진행된 step2 에서는 Region과 Dgree라는 새로운 변수명에 각각에 해당하는 칼럼의 인덱스를 이용하여 저장해주었는데, 이 과정에서 공백을 제거하기 위해 gsub() 함수를 써주었다. 이렇게 총 네가지의 새로운 변수를 만들 수 있었다. Step 3에서는 위에서 진행한 네 가지 변수를 밑줄로 연결해주기 위해서 paste 함수를 써 생성한 순차적으로 RawTransactions에 저장해주는 과정을 거쳐 주었다. Step 4에서는 Transaction ID 에 해당하는 변수와 step 3에서 생성한 결과물을 공백으로써 연결을 해주기 위해 paste() 함수에 sep=" " 를 넣어주어 연결해주면서 이를 Mooc_transactions로 정의된 변수에 저장하였다. 마지막 단계로 이전 단계에서 생성한 새로운 변수를 하나의 csv팡ㄹ로 저장을 하였고 저장된 결과의 파일은 아래와 같았다.



	x
2	MHxPC130313697 HarvardX_PH207x_India_Bachelor's
3	MHxPC130237753 HarvardX_PH207x_UnitedStates_Secondary
4	MHxPC130202970 HarvardX_CS50x_UnitedStates_Bachelor's
5	MHxPC130223941 HarvardX_CS50x_OtherMiddleEast/CentralAsia_Secondary
6	MHxPC130317399 HarvardX_PH207x_Australia_Master's
7	MHxPC130191782 HarvardX_CS50x_Pakistan_Bachelor's
8	MHxPC130191782 HarvardX_ER22x_Pakistan_Bachelor's
9	MHxPC130267000 HarvardX_PH207x_OtherSouthAsia_Master's
10	MHxPC130435800 HarvardX_CS50x_India_Bachelor's
11	MHxPC130284813 HarvardX_PH207x_UnitedStates_Bachelor's
12	MHxPC130235150 HarvardX_CS50x_India_Bachelor's
13	MHxPC130001411 HarvardX_CS50x_OtherEurope_Secondary
14	MHxPC130396873 HarvardX_PH207x_UnitedStates_Bachelor's
15	MHxPC130469401 HarvardX_CB22x_OtherMiddleEast/CentralAsia_Bachelor's
16	MHxPC130469401 HarvardX_CS50x_OtherMiddleEast/CentralAsia_Bachelor's
17	MHxPC130469401 HarvardX_ER22x_OtherMiddleEast/CentralAsia_Bachelor's
18	MHxPC130264946 HarvardX_PH207x_India_Secondary

Mooc_transactions라는 변수를 생성할 때 Transaction_ID에 해당하는 변수를 먼저 넣고 이전에 '_'를 이용하여 순차적으로 연결해준 값들이 공백으로 연결된 데이터로서 하나의 변수 x 칼럼 아래에 저장되어 있다는 것을 확인해 볼 수 있었다.

[Step 2] 데이터 불러오기 및 기초 통계량 확인

[Q2-1]

위에서 저장한 csv 파일을 read.transactions( )함수를 이용하여 읽어 들였다. 이 csv 파일을 생성하는 과정에서 의도치 않은 column name이 생기게 되었는데, 이 칼럼이 들어간 부분은 transaction으로써 취급하지 말아야 하기 때문에 'skip=1'을 사용하여 원하는 데이터만을 불러올 수 있었다.

```
MOOC_single_format <- read.transactions("MOOC_User_Course.csv", format = "single",
                                        header = TRUE, cols = c(1,2), rm.duplicates = TRUE, skip = 1)
```

이후 이렇게 불러온 데이터를 summary 함수를 이용하여 살펴보면 다음과 같이 나온다.

```
> summary(MOOC_single_format)
transactions as itemMatrix in sparse format with
 335649 rows (elements/itemsets/transactions) and
 1405 columns (items) and a density of 0.0008771195

most frequent items:
    MITx_6.00x_UnitedStates_Bachelor's      MITx_6.00x_UnitedStates_Secondary            MITx_6.00x_India_Bachelor's
                                 14192                                   8841                                   7813
         MITx_6.002x_India_Bachelor's HarvardX_CS50x_UnitedStates_Bachelor's                                  (Other)
                                  7633                                   7410                                 367749

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13
278439  43061   9997   2812    799    293    109     44     37     22     21      9      6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.232   1.000  13.000

includes extended item information - examples:
                              labels
1 HarvardX_CB22x_Australia_Bachelor's
2   HarvardX_CB22x_Australia_Master's
3  HarvardX_CB22x_Australia_Secondary

includes extended transaction information - examples:
   transactionID
1 MHxPC130000002
2 MHxPC130000004
3 MHxPC130000006
```

우선 첫번째로 'transactions as itemMAtrix in saprse format with'로 시작하는 부분을 보자면 335649개의 행과 1405개의 items로 구성이 되어 있다는 것을 확인할 수 있다. 즉, 행은 거래라고 생각해볼 수 있을 것이다. 밀도(density)가 0.0008이라는 것은 총 데이터 셀(335649 X 1405) cell 중에서 약 0.087%가 1의 값을 가지고 있다는 것이다. 현재 이 데이터에서 밀도가 매우 작다는 것을 확인할 수 있다.

그 다음으로 'most frequent items'라는 부분은 매우 직관적인데 셀에서 가장 많이 나타난 데이터를 의미하는 것이다. 이 데이터셋 기준으로 말을 하자면 학생들이 가장 많이 수강한 강의가 될 것이다. 이때, summary를 다시 보자면 14192건으로 MITx_6.00x_UnitedStates_Bachelor's 가 가장 많이 수강 된 것을 확인할 수 있다. 이후에 있는 강의 들은 방금과 비슷한 형식으로 해석을 하면 될 것이다.

'element'로 시작하는 부분은 이 데이터셋에 맞춰서 해석을 하자면 1과 그 아래 있는

278439는 한 학생이 한 과목만 수강한 경우가 278439만큼 있었다는 것을 의미한다. 그 아래에 평균이 나와있는 것을 보면 이 학교의 한 학생은 평균적으로 1.232 번 과목을 수강했고 그에 따라 숫자가 13까지 나와있는 것을 보면 가장 많이 수강한 학생은 13 과목을 들었다는 것을 알 수 있다.

[Q2-2]

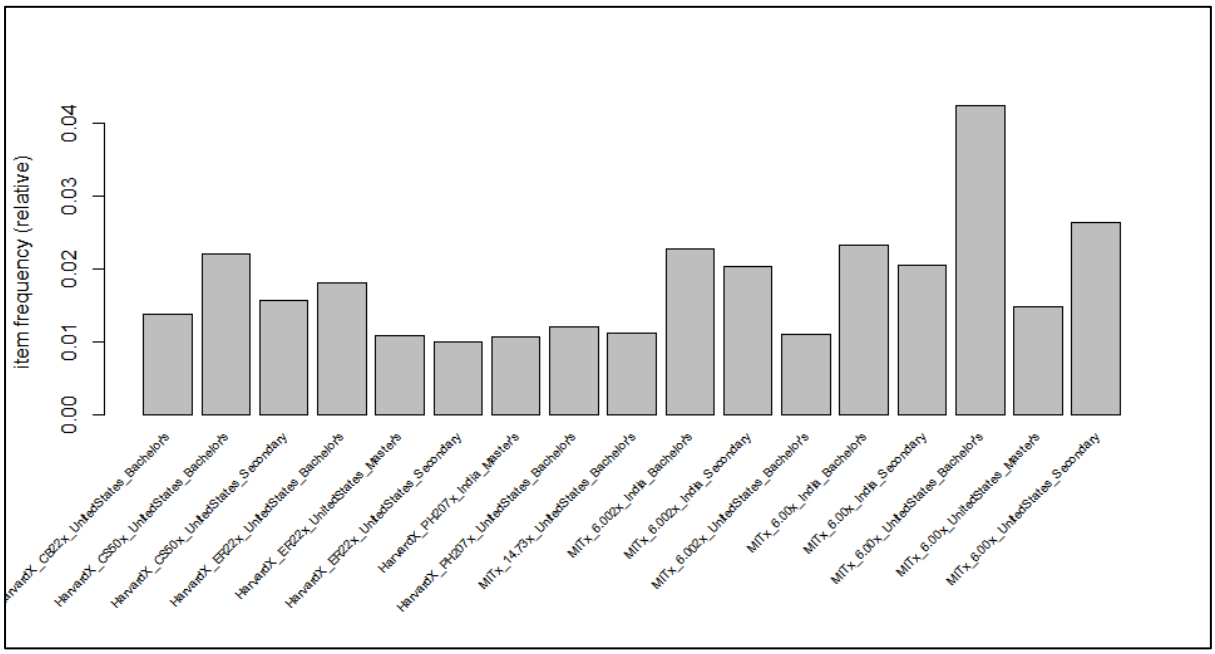
```
#Question2_2
item_name <- itemLabels(MOOC_single_format)
item_count <- itemFrequency(MOOC_single_format)*nrow(MOOC_single_format)
col <- brewer.pal(10,"Paired")
wordcloud(words = item_name, freq = item_count, min.freq = 100,
          scale = c(1,0.2), col = col,  random.order = FALSE)
```

위와 같은 코드를 이용하여 아래와 같은 워드 클라우드를 생성해낼 수 있었다. 워드클라우드를 통해서 알 수 있는 점은 비교적 다양한데 대부분 지지도를 간접적으로 알기 위해서 사용한다. Item의 이름과 item이 얼마나 빈번하게 나오는 지에 대한 count 및 색상이 필요하다. 이때 min.freq=100으로 지정해주었는데, 이는 그래프 표현되기 위해서 최소한 만족해야 하는 빈도수를 의미하는 것이고, scale=c(1, 0.2) 로 설정해주었는데, 이는 많이 일어나는 것과 적게 일어나는 것을 구분 할 수 있도록 scale해주는 것이다. 이 모든 것들이 종합적으로 합쳐져 아래의 워드클라우드가 생성되었다. 자주 나오는 것과 적게 나오는 것의 분간을 명확히 하기 위해 scale을 위와 같이 설정을 한 것이고 개인적으로 어떤 단어들이 많이 나오는지 더 많이 보고 싶은 마음에 min.freq은 100(문제에서 제시한 최소수치)으로 설정하여 주었다.

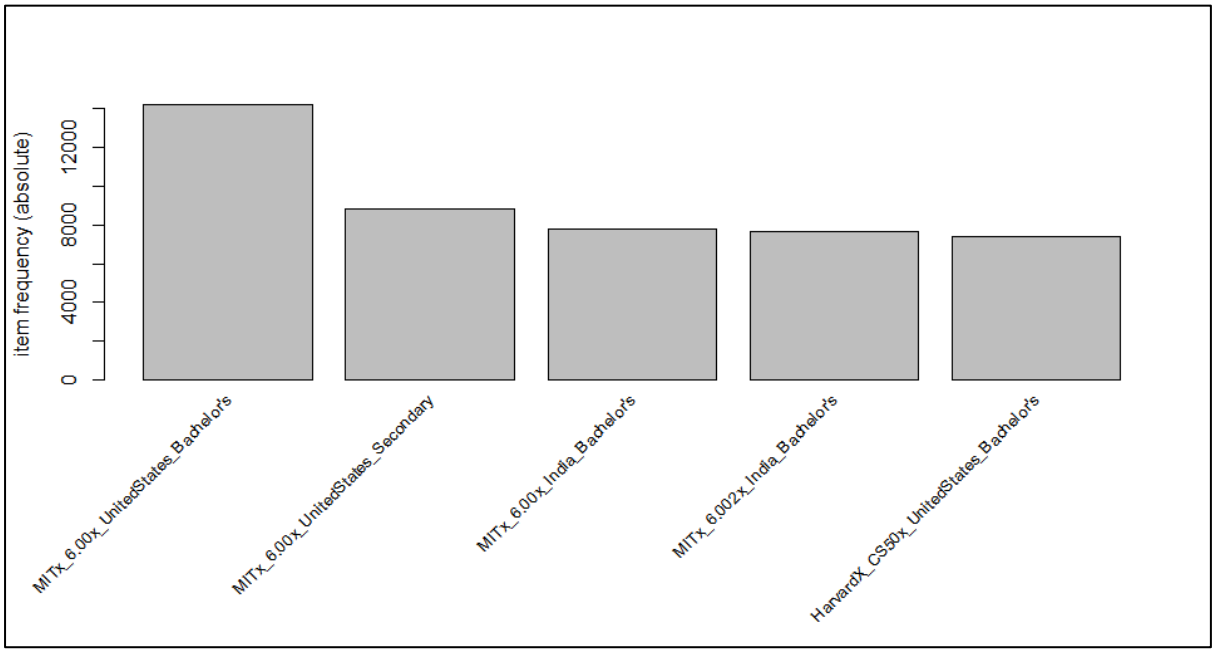
위의 word cloud를 보면 알 수 있다시피 MITx_6.00x_UnitedStates_Bachelor's 가 가장 많이 나왔다는 것을 알 수 있다. 보라색으로서 가장 크며 주변의 크기와 색깔로서 구별을 쉽게 할 수 있어 어떤 과목이 언급이 많이 되는지를 파악하기가 용이하였다.

[Q2-3]

itemFrequencyPlot( )함수를 사용하여 최소 빈도 1% 이상 등장한 Items들을 살펴보았다. itemFrequencyPlot( )함수에 대해 정확히 말하자면 Transaction대비 상대적인 빈도로 # %이상 등장한 변수들을 시각화 하는 함수인데, 1%이상 등장한 변수를 보고 싶으므로 support를 0.01로 설정하여 바플롯을 만들어 보았다.



아래에 있는 바 플롯은 당위 5개의 Item에 대하여 접속 국가가 어디인지를 확인하기 위해서 상위 5개의 아이템에 해당하는 것들만 모아온 것이다. 상위 다섯가지는 아래에 보이는 것과 같다는 것을 알 수 있다. 우리가 살펴보고자 하는 이 아이템들이 생성된 국가이므로 순차적으로 나열을 해보자면 United States, United States, India, India, United States라는 것을 확인할 수 있었다. 즉 United States와 India에서 생성된 item이 주로 쓰이고 있다는 것을 확인할 수 있었다.



[Step 3] 규칙 생성 및 결과 해석

[Q3-1]

```
Support <- c(0.0005,  0.001, 0.0025, 0.005)
Confidence <- c(0.0005, 0.001, 0.005)
```

Support와 Confidence 각각 네 가지, 3가지의 후보군을 설정하여 보았다. 이를 반복문을 규칙들을 생성하였고, Apriori를 이용하여 도출된 각 파라미터 조합에 따라 생성된 규칙의 수와 support, confidence 후보들을 결합하여 데이터 프레임을 생성하였다. 이에 따라 생성된 결과는 아래와 같았다.

Number of rules	Confidence = 5e-04	Confidence = 0.001	Confidence = 0.005
Support = 5e-04	654	448	240
Support = 0.001	307	307	99
Support = 0.0025	114	114	57
Support = 0.005	43	43	43

생성된 규칙의 수가 모두 10개 이상이라는 것을 확인할 수 있었다. 한가지 자명하게 보이는 결과는 Support가 가장 낮고 Confidence가 가장 낮은 부분에서 규칙이 가장 많이 생성되었다는 것이다. 조건절과 결과절이 함께 발생하는 확률을 낮춤으로 많은 규칙들이 통과하여 더 많이 생성되었다는 것을 알 수 있다.

[Q3-2]

Support와 confidence를 각각 0.01%와 5%로 지정을 하여 연관규칙을 분석해보았다. Inspect와 str함수 등 여러가지 방법을 통해 확인해본 결과 총 51개의 규칙이 생성되었다. 우선 support가 가장 높은 규칙을 보기 위해 support를 기준으로 내림차순을 해볼 결과 다음과 같은 결과물이 나왔다.

```
> inspect(sort(rules, by = "support"))
     lhs                                                rhs                                                support     confidence coverage    lift     count
[1]  {HarvardX_CS50x_UnitedStates_Bachelor's}  => {MITx_6.00x_UnitedStates_Bachelor's}      0.003643687 0.16504723 0.022076634  3.903462  1223
[2]  {MITx_6.00x_UnitedStates_Bachelor's}      => {HarvardX_CS50x_UnitedStates_Bachelor's} 0.003643687 0.08617531 0.042282265  3.903462  1223
[3]  {MITx_6.00x_India_Secondary}              => {MITx_6.002x_India_Secondary}            0.003625811 0.17745698 0.020432058  8.692828  1217
[4]  {MITx_6.002x_India_Secondary}             => {MITx_6.00x_India_Secondary}             0.003625811 0.17761238 0.020414183  8.692828  1217
[5]  {MITx_6.002x_India_Bachelor's}            => {MITx_6.00x_India_Bachelor's}            0.003092516 0.13598847 0.022741018  5.842109  1038
[6]  {MITx_6.00x_India_Bachelor's}             => {MITx_6.002x_India_Bachelor's}           0.003092516 0.13285550 0.023277293  5.842109  1038
[7]  {MITx_6.002x_UnitedStates_Bachelor's}     => {MITx_6.00x_UnitedStates_Bachelor's}     0.002818420 0.25484914 0.011059172  6.027329   946
[8]  {MITx_6.00x_UnitedStates_Bachelor's}      => {MITx_6.002x_UnitedStates_Bachelor's}    0.002818420 0.06665727 0.042282265  6.027329   946
[9]  {MITx_8.02x_India_Secondary}              => {MITx_6.002x_India_Secondary}            0.002800545 0.38810900 0.007215871 19.011734   940
[10] {MITx_6.002x_India_Secondary}             => {MITx_8.02x_India_Secondary}             0.002800545 0.13718622 0.020414183 19.011734   940
[11] {HarvardX_CS50x_India_Secondary}          => {MITx_6.00x_India_Secondary}             0.002681373 0.29392554 0.009122625 14.385508   900
[12] {MITx_6.00x_India_Secondary}              => {HarvardX_CS50x_India_Secondary}         0.002681373 0.13123360 0.020432058 14.385508   900
[13] {HarvardX_CB22x_UnitedStates_Bachelor's}  => {HarvardX_ER22x_UnitedStates_Bachelor's} 0.002589014 0.18728448 0.013823965 10.385239   869
```

{HarvardX_CS50x_UnitedStates_Bachelor's}의 조건절, {MITx_6.00x_UnitedStates_Bachelor's}의 결과절을 가진 규칙이라는 것을 확인할 수 있었다. Support의 수치를 보면 0.003643687이였고 즉 0.3%로 나왔다는 의미가 된다. Confidence는 0.165로 HarvardX_CS50x_United States_Bachelor's이 수강 될 때 조건절과 결과절이 동시에 수강 될 확률(16.5%)을 의미한다. Lift의 의미는 각각의 독립사건으로 일어날 때 두 사건이 함께 발생할 확률이다. 생각을 해보면 지금 이 규칙이 3.9배 더 같이 수강 되었다는 의미로 해석을 할 수 있을 것 같다. 그러므로 Lift가 크다는 것을 이 규칙이 효과적인다 라고 할 수 있을 것이다.

```
> inspect(sort(rules, by = "confidence"))
     lhs                                          rhs                                          support confidence    coverage      lift count
[1]  {MITx_8.02x_India_Secondary}              => {MITx_6.002x_India_Secondary}              0.002800545 0.38810900 0.007215871 19.011734   940
[2]  {MITx_8.02x_India_Bachelor's}             => {MITx_6.002x_India_Bachelor's}             0.002496656 0.38564197 0.006474025 16.957990   838
[3]  {HarvardX_CS50x_India_Secondary}          => {MITx_6.00x_India_Secondary}               0.002681373 0.29392554 0.009122625 14.385508   900
[4]  {MITx_6.002x_UnitedStates_Secondary}      => {MITx_6.00x_UnitedStates_Secondary}        0.001939526 0.28194023 0.006879210 10.703875   651
[5]  {HarvardX_CS50x_India_Bachelor's}         => {MITx_6.00x_India_Bachelor's}              0.002016988 0.26918489 0.007492947 11.564270   677
[6]  {MITx_6.002x_UnitedStates_Bachelor's}     => {MITx_6.00x_UnitedStates_Bachelor's}       0.002818420 0.25484914 0.011059172  6.027329   946
```

Confidence로 분석을 해보자면 가장 높은 confidence를 가지는 규칙은 맨 첫번째 줄과 같다는 것을 확인할 수 있다. {MITx_8.02x_India_Secondary}의 조건절, {MITx_6.002x_India_Secondary}의 결과절을 가지고 있다. Support 는 0.002800545 confidence는 0.38810900 Lift 는 19.011734 라는 것을 확인할 수 있는데, confidence의 수치를 기준으로 분석하자면 조건절의 수업이 수강 될 때, 조건절과 결과절의 강의가 같이 수강 될 확률이 0.39 정도 된다는 것이다. Lift를 보면 알 수 있듯이 독립적 사건일 때와 비교해보면 함께 수강 된 것이 19배 더 같이 수강된다 라고 분석할 수 있을 것이다. Confidence가 높다는 점에서 좋은 규칙이지는 않다고 생각해 볼 수 있을 것이다.

```
> inspect(sort(rules, by = "lift"))
     lhs                                          rhs                                          support confidence    coverage      lift count
[1]  {MITx_8.02x_UnitedStates_Bachelor's}      => {MITx_6.002x_UnitedStates_Bachelor's}      0.001391334 0.21620370 0.006435294 19.549719   467
[2]  {MITx_6.002x_UnitedStates_Bachelor's}     => {MITx_8.02x_UnitedStates_Bachelor's}       0.001391334 0.12580819 0.011059172 19.549719   467
[3]  {HarvardX_CB22x_UnitedStates_Secondary}   => {HarvardX_ER22x_UnitedStates_Secondary}    0.001540300 0.19240789 0.008005387 19.106957   517
[4]  {HarvardX_ER22x_UnitedStates_Secondary}   => {HarvardX_CB22x_UnitedStates_Secondary}    0.001540300 0.15295858 0.010070043 19.106957   517
[5]  {MITx_6.002x_India_Secondary}             => {MITx_8.02x_India_Secondary}               0.002800545 0.13718622 0.020414183 19.011734   940
[6]  {MITx_8.02x_India_Secondary}              => {MITx_6.002x_India_Secondary}              0.002800545 0.38810900 0.007215871 19.011734   940
```

Lift를 기준으로 보았을 때의 조건절과 결과절은 위에 있는 맨 첫번째 줄과 같다. Lift가 19.549719라는 것을 볼 수 있는데, 조건절과 결과절을 독립적으로 수강했을 때에 비해 동시에 수강한 것이 19.5배 정도 높았다는 것을 의미한다.

다음으로 효용성이 가장 높은 규칙 1~3위를 보도록 하겠다. 이때의 효용성 지표는 Support × Confidence × Lift라고 하였다. 이를 보기 위해서 각각의 Support, Confidence, Lift 값을 곱한 결과들을 저장하는 변수를 따로 하나 만들어 하나의 데이터 프레임으로 만들어 주어 결과를 확인해 주었다.

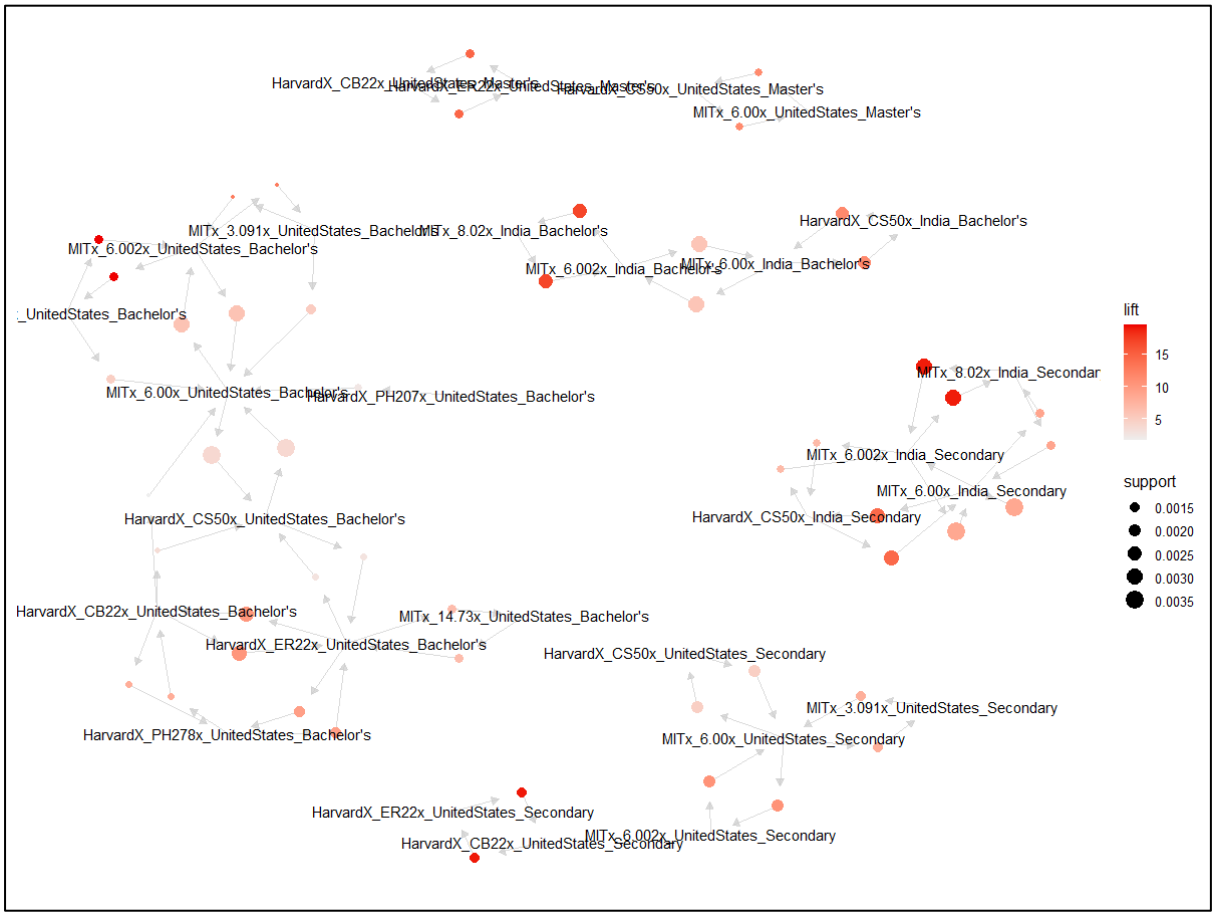
```
> rules_df
                                      LHS                                          RHS     support confidence    coverage      lift
23             {MITx_8.02x_India_Secondary}             {MITx_6.002x_India_Secondary} 0.002800545 0.38810900 0.007215871 19.011734
5             {MITx_8.02x_India_Bachelor's}            {MITx_6.002x_India_Bachelor's} 0.002496656 0.38564197 0.006474025 16.957990
25          {HarvardX_CS50x_India_Secondary}              {MITx_6.00x_India_Secondary} 0.002681373 0.29392554 0.009122625 14.385508
```

```
Perf_Mea_New
0.0206641682
0.0163274116
0.0113375620
```

우선 1~3위 까지에 해당하는 규칙과 Support, Confidence, Lift값은 위와 같다. 각각의 효용성 지표 값은 왼쪽과 같이 나왔다는 것을 확인할 수 있다.

마지막으로 생성된 규칙을 plot( )함수의 "graph" method를 이용하여 도시할 경우 두 아이템이 서로 조건절/결과절을 달리해서 생성되는 경우가 존재함을 확인할 수 있는데, 세가지 규칙들에 대해 Support/Confidence/Lift 값을 확인해보고 조건절과 결과절의 위치에 따라서 어떠한 지표 값들이 차이가 나는지에 대해서 살펴보도록 하겠다.

가장 먼저 graph 함수를 사용하여 어떤 형태의 그래프가 나오는지를 확인해보았다.



51가지의 규칙에 대한 그래프를 plotting 해본 결과는 위와 같았다. 위의 plot을 확인해보면 양방향을 가지고 있는 규칙들이 있는데 그중에서 %pin%함수를 이용하여 이에 해당하는 값을 도출하여 보았다. 그중에 상위에 나온 값들의 세개(양방향이므로 6가지)를 보았다.

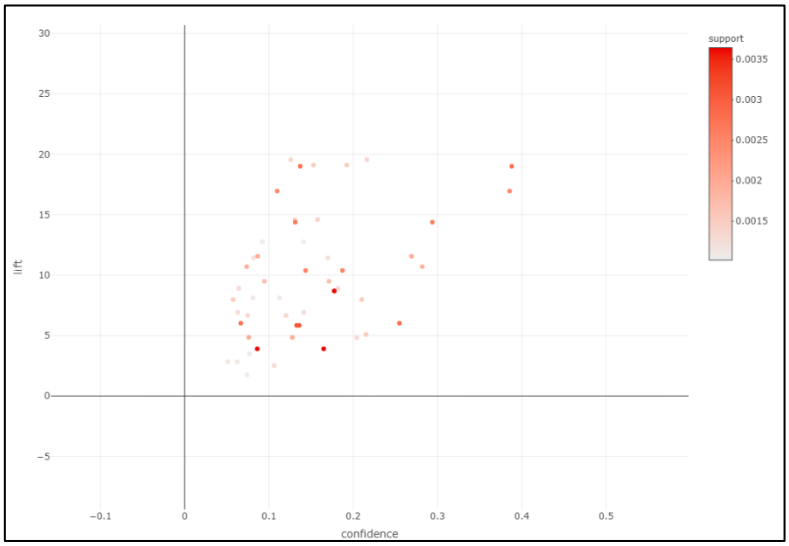
```
> inspect(rule_xy)
     lhs                                      rhs                                      support     confidence coverage    lift      count
[1]  {MITx_8.02x_India_Bachelor's}         => {MITx_6.002x_India_Bachelor's}         0.002496656 0.38564197 0.006474025 16.957990  838
[2]  {MITx_6.002x_India_Bachelor's}        => {MITx_8.02x_India_Bachelor's}          0.002496656 0.10978645 0.022741018 16.957990  838
[3]  {MITx_3.091x_UnitedStates_Secondary} => {MITx_6.00x_UnitedStates_Secondary}    0.001516465 0.21024370 0.007212892  7.981912  509
[4]  {MITx_6.00x_UnitedStates_Secondary}  => {MITx_3.091x_UnitedStates_Secondary}  0.001516465 0.05757267 0.026340016  7.981912  509
[5]  {MITx_6.002x_UnitedStates_Secondary} => {MITx_6.00x_UnitedStates_Secondary}    0.001939526 0.28194023 0.006879210 10.703875  651
[6]  {MITx_6.00x_UnitedStates_Secondary}  => {MITx_6.002x_UnitedStates_Secondary}  0.001939526 0.07363420 0.026340016 10.703875  651
```

각 규칙들을 보면 서로 crossover된 규칙들 끼리는 Support와 lift가 같다는 것을 확인할 수 있었고 confidence가 다르게 나오는 것을 확인할 수 있었다. Confidence는 P(A∩B)/P(A)인데 각 조건절을 구성하는 P(A)가 달라지기 때문이다. 이는 다른 말로 support의 값이 달라지기 때문이다. 조건절이 일어날 확률이 커지면 confidence는 작아지고, 조건절이 일어날 확률이 작아지면 confidence는 커질 거이다. 그러므로 조건절이 일어나는 확률을 대략적으로 가늠할 수 있게 된다. Confidence를 보면 처음 규칙이 0.38, 두번째 규칙은 0.109이다. 한마디로 두번째 규칙에서 조건절이 일어날 확률이 첫번째보다 더 높다는 것으로 생각해볼 수 있을 것이다. 이외 나머지의 규칙들도 비슷하게 해석을 하면 될 것이다.

[Extra Question]

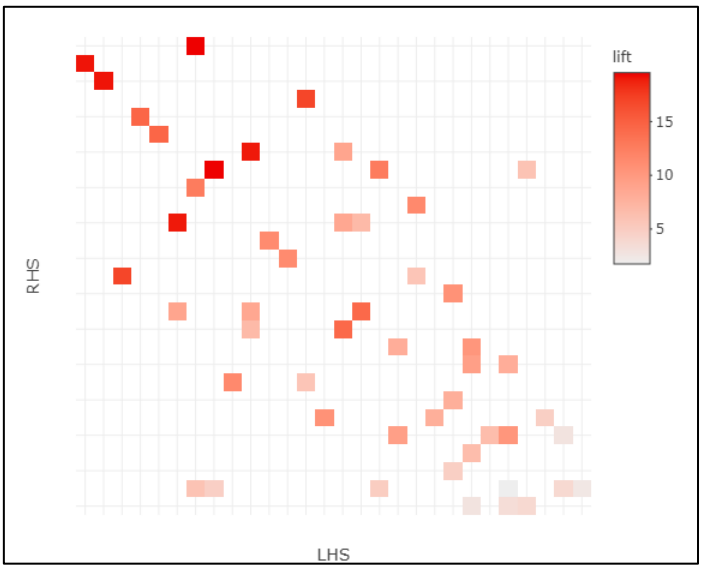
arulesViz에는 여러가지 plotting기법 들이 있음을 확인할 수 있다. 이를 확인해보자면 다음과 같다. methods : "scatterplot", "two-key plot", "matrix", "matrix3D", "mosaic", "doubledecker", "graph", "paracoord", "grouped", "iplots" 등의 방법들이 존재한다. 또한 이 외에도 여러가지 값들을 넣어줌으로써 분석을 더 용이 학 할 수 있는데, 예를 들어 measure 은 관심있는 시각화 지표를 넣어주는 란이고, interactive를 통해 생성한 plot을 조작할 수 있다. 이를 이용하여 몇가지 플롯팅을 해보고자 한다.

-Scatter plot



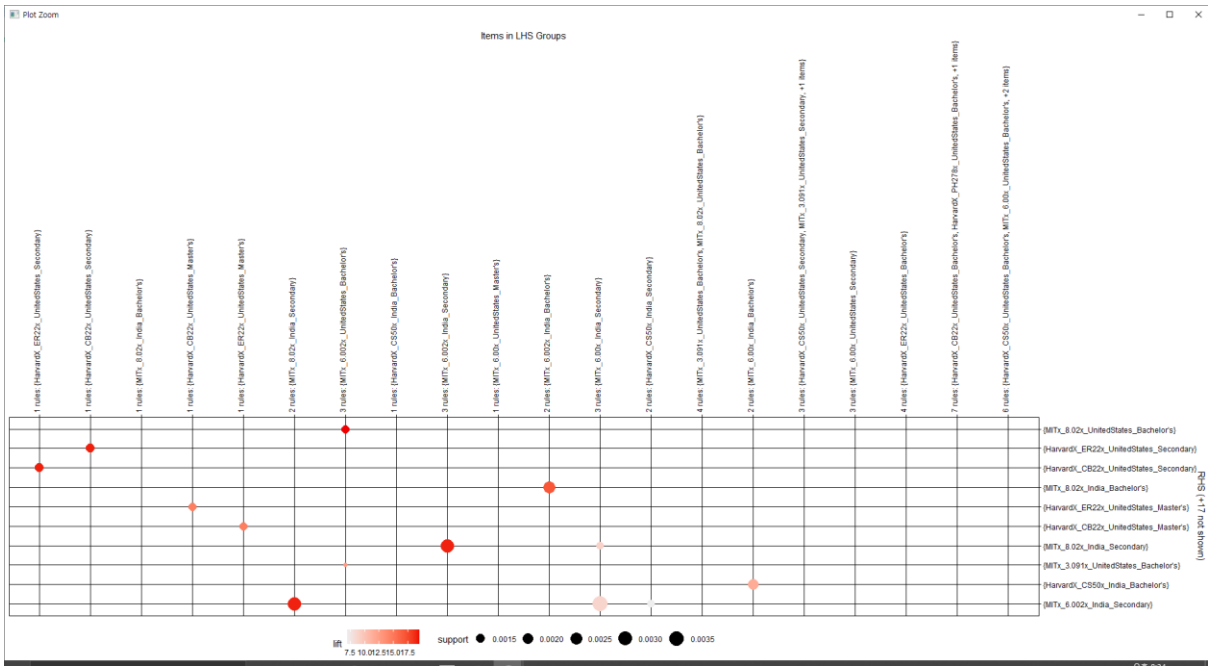
상당히 직관적으로 분석할 수 있는 표이다. x축은 confidence를 y축은 lift의 값을 넣었고, 각 점들의 색깔은 support를 표현하게 함으로써 테이블로 보아야 하는 수치를 한번에 체크할 수 있다. 또한 각 점들에 마우스를 가져다 대면 그 점에 해당하는 상세 정보를 확인할 수 있다.

-matrix plot



현재 위의 표는 매트릭스 plot으로써 scatterplot과는 x축은 LHS를 y축에는 RHS를 넣고 색깔을 Lift를 표현해 주었다. 각 선행 itemset과 후행 itemset들을 하나의 표에 표현을 함으로써 이 또한 함축적으로 많은 정보를 살펴볼 수 있는데, scatter plot과 다른 점에 대해서 생각을 해보자면 scatter plot의 값들은 연속적인 정보를 x와 y축에 담아서 보고, 매트릭스 플롯은 itemset과 같이 discrete한 정보를 표현하기에 용이하다고 생각하였다.

-grouped



이 또한 LHS와 RHS를 보여주는 plot이다. 각 원의 크기는 support의 크기를 의미하고 색깔은 lift를 의미한다는 것을 알 수 잇다. 총 51개의 규칙에 대하여 결과를 보여주고 51개의 규칙에 대한 상관관계를 보여주고 있다.