Multivariate-data-Analysis

Assignment1. Association rule

산업경영공학부
2014170849 배정모

Dataset: MOOC Dataset (big_student_clear_third_version.csv)

[Step 1] 데이터 변환

-1단계: 제공된 csv파일을 읽어온 뒤 Item Name에 해당하는 네 개의 변수를 각각 Institute, Course, Region, Degree로 정의한 변수명에 저장.

```
######<step1>######
mooc_dataset <- read.csv("big_student_clear_third_version.csv")
View(mooc_dataset)

#Q1-1
Institute<-mooc_dataset$institute
Course<-mooc_dataset$course_id
Region<-mooc_dataset$final_cc_cname_DI
Degree<-mooc_dataset$LoE_DI
```

-2단계: Region에 해당하는 변수의 공백을 제거하는 전처리.

```
#Q1-2
Region<-gsub(" ","",Region)
Region
```

```
 [919] "UnitedStates"                        "UnitedStates"                        "UnitedStates"
 [922] "Portugal"                            "UnitedStates"                        "UnitedStates"
 [925] "UnitedKingdom"                       "UnitedStates"                        "UnitedStates"
 [928] "UnitedKingdom"                       "UnitedStates"                        "Egypt"
 [931] "Mexico"                              "UnitedStates"                        "UnitedStates"
 [934] "UnitedStates"                        "UnitedStates"                        "OtherNorth&CentralAmer.,Caribbean"
 [937] "Australia"                           "UnitedKingdom"                       "UnitedStates"
 [940] "UnitedKingdom"                       "UnitedKingdom"                       "OtherEurope"
 [943] "UnitedStates"                        "UnitedStates"                        "UnitedStates"
 [946] "OtherNorth&CentralAmer.,Caribbean"   "Brazil"                            "UnitedKingdom"
 [949] "Brazil"                              "UnitedStates"                        "UnitedStates"
 [952] "UnitedStates"                        "Brazil"                              "Brazil"
 [955] "OtherNorth&CentralAmer.,Caribbean"   "UnitedStates"                       "Canada"
 [958] "Brazil"                              "UnitedStates"                        "UnitedStates"
 [961] "UnitedStates"                        "UnitedStates"                        "UnitedKingdom"
 [964] "UnitedStates"                        "UnitedStates"                        "India"
 [967] "UnitedStates"                        "OtherSouthAmerica"                   "OtherSouthAmerica"
 [970] "UnitedKingdom"                       "OtherAfrica"                         "UnitedStates"
 [973] "UnitedStates"                        "Mexico"                              "UnitedStates"
 [976] "UnitedStates"                        "Brazil"                              "UnitedStates"
 [979] "UnitedKingdom"                       "Brazil"                              "UnitedStates"
 [982] "UnitedStates"                        "UnitedStates"                        "UnitedStates"
 [985] "UnitedStates"                        "Brazil"                              "UnitedStates"
 [988] "UnitedStates"                        "OtherNorth&CentralAmer.,Caribbean"   "Brazil"
 [991] "UnitedStates"                        "UnitedStates"                        "Canada"
 [994] "UnitedStates"                        "Brazil"                              "Brazil"
 [997] "Australia"                           "Australia"                           "UnitedStates"
[1000] "UnitedStates"
 [ reached getOption("max.print") -- omitted 415921 entries ]
```

-3단계: 네 변수를 밑줄로 연결하여 RawTransactions로 정의된 하나의 변수에 저장.

```
#Q1-3
RawTransactions<-paste(Institute,Course,Region,Degree,sep='_')
RawTransactions
```

```
 [976] "HarvardX_PH207x_UnitedStates_Bachelor's"
 [977] "HarvardX_PH207x_Brazil_Secondary"
 [978] "HarvardX_PH207x_UnitedStates_Bachelor's"
 [979] "HarvardX_CS50x_UnitedKingdom_Bachelor's"
 [980] "HarvardX_CS50x_Brazil_Master's"
 [981] "HarvardX_PH207x_UnitedStates_Bachelor's"
 [982] "HarvardX_CS50x_UnitedStates_Bachelor's"
 [983] "HarvardX_CB22x_UnitedStates_Secondary"
 [984] "HarvardX_ER22x_UnitedStates_Secondary"
 [985] "HarvardX_PH207x_UnitedStates_Secondary"
 [986] "HarvardX_PH207x_Brazil_Bachelor's"
 [987] "HarvardX_CB22x_UnitedStates_Bachelor's"
 [988] "HarvardX_CS50x_UnitedStates_Bachelor's"
 [989] "HarvardX_ER22x_OtherNorth&CentralAmer.,Caribbean_Master's"
 [990] "HarvardX_PH207x_Brazil_Secondary"
 [991] "HarvardX_PH207x_UnitedStates_Bachelor's"
 [992] "HarvardX_ER22x_UnitedStates_Less than Secondary"
 [993] "HarvardX_PH207x_Canada_Doctorate"
 [994] "HarvardX_CS50x_UnitedStates_Master's"
 [995] "HarvardX_CS50x_Brazil_Secondary"
 [996] "HarvardX_PH207x_Brazil_Bachelor's"
 [997] "HarvardX_ER22x_Australia_Bachelor's"
 [998] "HarvardX_PH207x_Australia_Bachelor's"
 [999] "HarvardX_PH207x_UnitedStates_Master's"
[1000] "HarvardX_PH207x_UnitedStates_Bachelor's"
 [ reached getOption("max.print") -- omitted 415921 entries ]
```

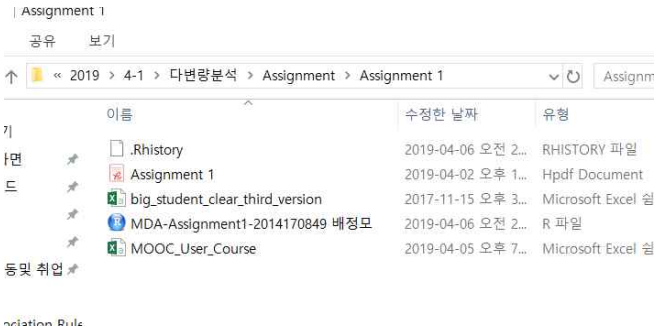
-4단계: Transaction ID에 해당하는 변수와 3단계의 결과물을 한 칸 공백으로 연결하여 MOOC_transactions로 정의된 변수에 저장.

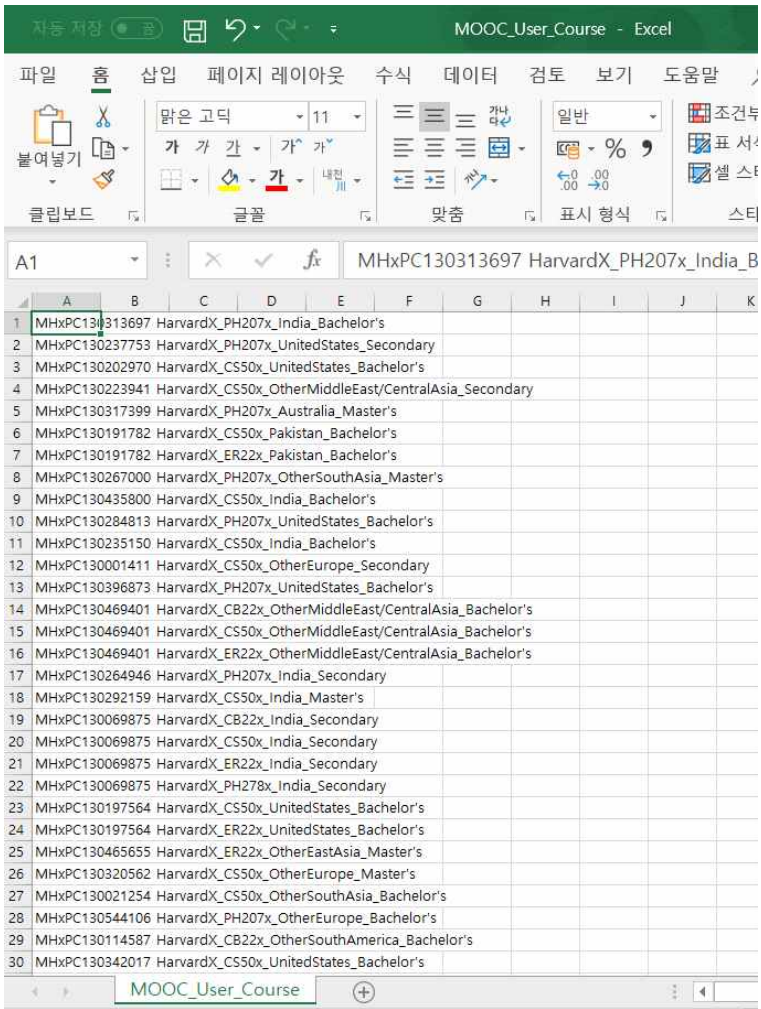
```
#Q1-4
MOOC_transactions<-paste(mooc_dataset$userid_DI,RawTransactions,sep=' ')
MOOC_transactions
```

```
 [974] "MHxPC130580906 HarvardX_CS50x_Mexico_Master's"
 [975] "MHxPC130143718 HarvardX_PH207x_UnitedStates_Secondary"
 [976] "MHxPC130020868 HarvardX_PH207x_UnitedStates_Bachelor's"
 [977] "MHxPC130355942 HarvardX_PH207x_Brazil_Secondary"
 [978] "MHxPC130343092 HarvardX_PH207x_UnitedStates_Bachelor's"
 [979] "MHxPC130153247 HarvardX_CS50x_UnitedKingdom_Bachelor's"
 [980] "MHxPC130250171 HarvardX_CS50x_Brazil_Master's"
 [981] "MHxPC130556560 HarvardX_PH207x_UnitedStates_Bachelor's"
 [982] "MHxPC130046677 HarvardX_CS50x_UnitedStates_Bachelor's"
 [983] "MHxPC130324535 HarvardX_CB22x_UnitedStates_Secondary"
 [984] "MHxPC130324535 HarvardX_ER22x_UnitedStates_Secondary"
 [985] "MHxPC130324535 HarvardX_PH207x_UnitedStates_Secondary"
 [986] "MHxPC130345784 HarvardX_PH207x_Brazil_Bachelor's"
 [987] "MHxPC130247585 HarvardX_CB22x_UnitedStates_Bachelor's"
 [988] "MHxPC130461065 HarvardX_CS50x_UnitedStates_Bachelor's"
 [989] "MHxPC130183647 HarvardX_ER22x_OtherNorth&CentralAmer.,Caribbean_Master's"
 [990] "MHxPC130258134 HarvardX_PH207x_Brazil_Secondary"
 [991] "MHxPC130565082 HarvardX_PH207x_UnitedStates_Bachelor's"
 [992] "MHxPC130583003 HarvardX_ER22x_UnitedStates_Less than Secondary"
 [993] "MHxPC130286913 HarvardX_PH207x_Canada_Doctorate"
 [994] "MHxPC130131242 HarvardX_CS50x_UnitedStates_Master's"
 [995] "MHxPC130125503 HarvardX_CS50x_Brazil_Secondary"
 [996] "MHxPC130017472 HarvardX_PH207x_Brazil_Bachelor's"
 [997] "MHxPC130002966 HarvardX_ER22x_Australia_Bachelor's"
 [998] "MHxPC130002966 HarvardX_PH207x_Australia_Bachelor's"
 [999] "MHxPC130350076 HarvardX_PH207x_UnitedStates_Master's"
[1000] "MHxPC130277620 HarvardX_PH207x_UnitedStates_Bachelor's"
 [ reached getOption("max.print") -- omitted 415921 entries ]
>
```

-5단계: 4단계의 MOOC_transactions 변수를 'MOOC_User_Course.csv'라는 파일명으로 저장&확인.

```
#Q1-5
write.table(MOOC_transactions,file='MOOC_User_Course.csv',col.names = FALSE,row.names = FALSE,quote = FALSE)
```





	A	B	C	D	E	F	G	H	I	J	K
1	MHxPC130313697	HarvardX_PH207x_India_Bachelor's									
2	MHxPC130237753	HarvardX_PH207x_UnitedStates_Secondary									
3	MHxPC130202970	HarvardX_CS50x_UnitedStates_Bachelor's									
4	MHxPC130223941	HarvardX_CS50x_OtherMiddleEast/CentralAsia_Secondary									
5	MHxPC130317399	HarvardX_PH207x_Australia_Master's									
6	MHxPC130191782	HarvardX_CS50x_Pakistan_Bachelor's									
7	MHxPC130191782	HarvardX_ER22x_Pakistan_Bachelor's									
8	MHxPC130267000	HarvardX_PH207x_OtherSouthAsia_Master's									
9	MHxPC130435800	HarvardX_CS50x_India_Bachelor's									
10	MHxPC130284813	HarvardX_PH207x_UnitedStates_Bachelor's									
11	MHxPC130235150	HarvardX_CS50x_India_Bachelor's									
12	MHxPC130001411	HarvardX_CS50x_OtherEurope_Secondary									
13	MHxPC130396873	HarvardX_PH207x_UnitedStates_Bachelor's									
14	MHxPC130469401	HarvardX_CB22x_OtherMiddleEast/CentralAsia_Bachelor's									
15	MHxPC130469401	HarvardX_CS50x_OtherMiddleEast/CentralAsia_Bachelor's									
16	MHxPC130469401	HarvardX_ER22x_OtherMiddleEast/CentralAsia_Bachelor's									
17	MHxPC130264946	HarvardX_PH207x_India_Secondary									
18	MHxPC130292159	HarvardX_CS50x_India_Master's									
19	MHxPC130069875	HarvardX_CB22x_India_Secondary									
20	MHxPC130069875	HarvardX_CS50x_India_Secondary									
21	MHxPC130069875	HarvardX_ER22x_India_Secondary									
22	MHxPC130069875	HarvardX_PH278x_India_Secondary									
23	MHxPC130197564	HarvardX_CS50x_UnitedStates_Bachelor's									
24	MHxPC130197564	HarvardX_ER22x_UnitedStates_Bachelor's									
25	MHxPC130465655	HarvardX_ER22x_OtherEastAsia_Master's									
26	MHxPC130320562	HarvardX_CS50x_OtherEurope_Master's									
27	MHxPC130021254	HarvardX_CS50x_OtherSouthAsia_Bachelor's									
28	MHxPC130544106	HarvardX_PH207x_OtherEurope_Bachelor's									
29	MHxPC130114587	HarvardX_CB22x_OtherSouthAmerica_Bachelor's									
30	MHxPC130342017	HarvardX_CS50x_UnitedStates_Bachelor's									

[Step 2] 데이터 불러오기 및 기초 통계량 확인

[Q2-1] 생성된 single format의 데이터를 read.transactions( )를 이용하여 읽어 들이고 summary( )를 사용하여 해당 데이터의 속성 파악.

```
39  ######<step2>######
40  #Q2-1
41  MOOC_single <- read.transactions("MOOC_User_Course.csv",
42                                   format = "single", cols = c(1,2), rm.duplicates=TRUE)
43
44  summary(MOOC_single)
45:1  <step2> :

Console  Terminal
C:/Users/Bae jeong mo/Desktop/2014170849/2019/4-1/다변량분석/multivariate-data-analysis-master/02 Association Rule Mining/
> summary(MOOC_single)
transactions as itemMatrix in sparse format with
 335650 rows (elements/itemsets/transactions) and
 1405 columns (items) and a density of 0.000877119

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                                 367750

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13
278440  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
>
```

->335,650개의 item_sets(transactions), 1,405개의 items이 존재함.

->“ MITx_6.00x_Unitedstates_Bachelor’s ” item이 14,192번으로 가장 빈번함.

->item set의 크기별 개수를 알 수 있음.
ex) item_set의 크기가 1인 경우(1개의 item이 포함된 set) 총 335,650개 중 278,440개.

->item name과 transaction ID를 예시로 몇 개 볼 수 있음.

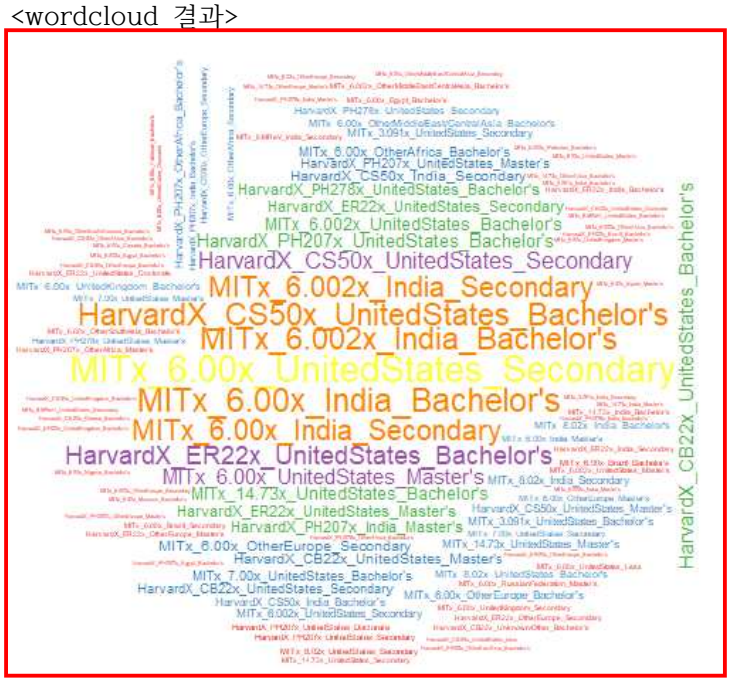
[Q2-2] color 변경, min.freq 최소 100 이상으로 wordcloud 생성.

```
#Q2-2
itemName <- itemLabels(MOOC_single)
itemCount <- itemFrequency(MOOC_single)*nrow(MOOC_single)
col <- brewer.pal(9, "Set1")
wordcloud(words = itemName, freq = itemCount, min.freq = 800, scale = c(2.5, 0.1), col = col , random.order = FALSE)
```

->color : 기존 'Dark2' 대신 'Set1'을 사용.

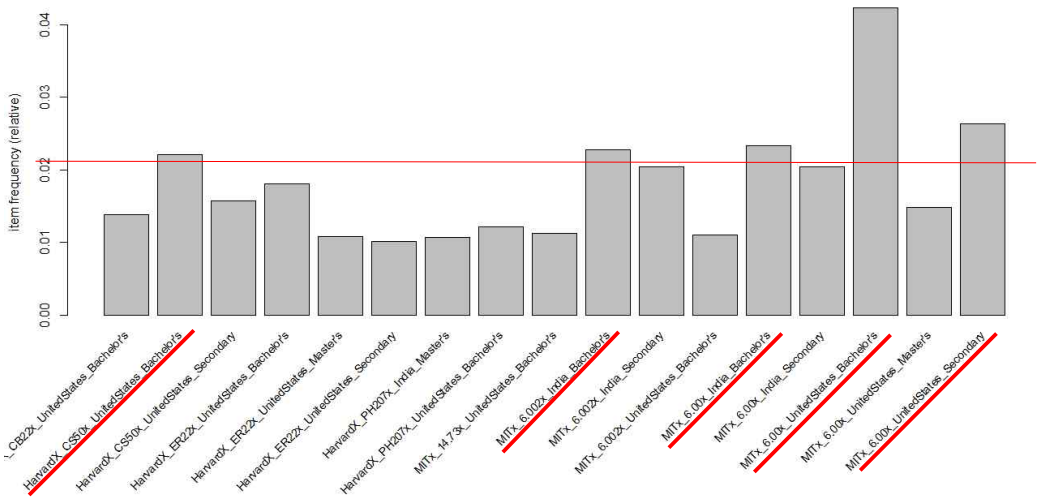
->min.freq=800, scale=c(2.5, 0.1)을 사용.

<wordcloud 결과>



[Q2-3] 최소 빈도 1% 이상 등장한 Items들의 Bar Chart 도시.

```
#Q2-3
itemFrequencyPlot(MOOC_single, support = 0.01, cex.names=0.7)
```



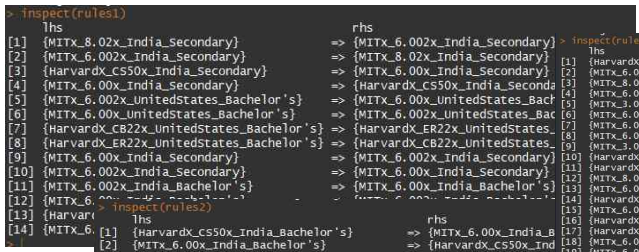
->item 빈도별 접속 상위국가 : 1, 2, 5위-United states, 3, 4위-India

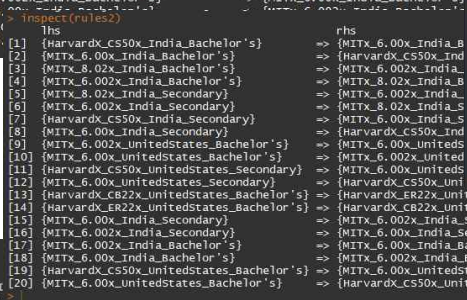
[Step 3] 규칙 생성 및 결과 해석

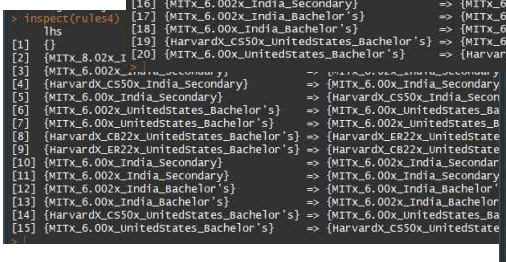
[Q3-1] 각 support, confidence마다 규칙 생성.

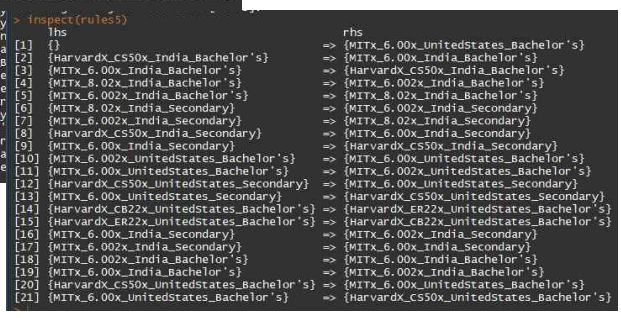
```
#Q3-1
rules1 <- apriori(MOOC_single, parameter=list(support=0.0025, confidence=0.05))
inspect(rules1)
rules2 <- apriori(MOOC_single, parameter=list(support=0.0020, confidence=0.05))
inspect(rules2)
rules3 <- apriori(MOOC_single, parameter=list(support=0.0015, confidence=0.05))
inspect(rules3)
rules4 <- apriori(MOOC_single, parameter=list(support=0.0025, confidence=0.04))
inspect(rules4)
rules5 <- apriori(MOOC_single, parameter=list(support=0.0020, confidence=0.04))
inspect(rules5)
rules6 <- apriori(MOOC_single, parameter=list(support=0.0015, confidence=0.04))
inspect(rules6)
rules7 <- apriori(MOOC_single, parameter=list(support=0.0025, confidence=0.03))
inspect(rules7)
rules8 <- apriori(MOOC_single, parameter=list(support=0.0020, confidence=0.03))
inspect(rules8)
rules9 <- apriori(MOOC_single, parameter=list(support=0.0015, confidence=0.03))
inspect(rules9)
```

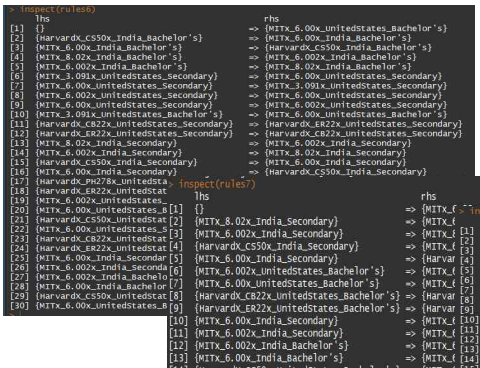
->support:0.0025, 0.0020, 0.0015일 때와 confidence:0.05, 0.04, 0.03일 때, 규칙 생성.

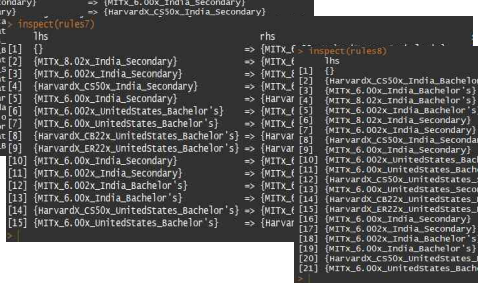


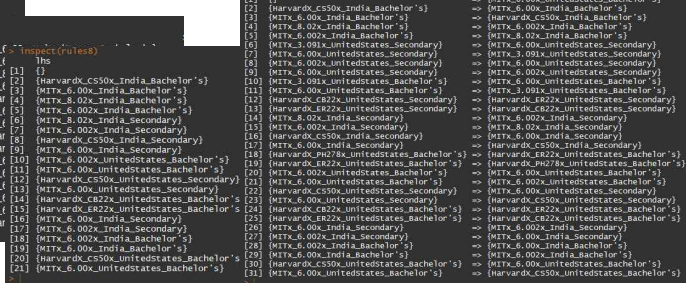












Number of rules	Confidence=0.05	Confidence=0.04	Confidence=0.03
Support=0.0025	14개	14개	15개
Support=0.0020	20개	21개	21개
Support=0.0015	29개	30개	31개

[Q3-2] support = 0.001, confidence = 0.05로 규칙 생성 & 분석

```
rules <- apriori(MOOC_single, parameter=list(support=0.001, confidence=0.05))
inspect(rules)
```

```
> inspect(rules)
     lhs                                                  rhs                                                  support     confidence lift      count
[1]  {HarvardX_CS50x_UnitedStates_Master's}            => {MITx_6.00x_UnitedStates_Master's}                  0.001218531 0.16985050 11.429495  409
[2]  {MITx_6.00x_UnitedStates_Master's}                => {HarvardX_CS50x_UnitedStates_Master's}              0.001218531 0.08199679 11.429495  409
[3]  {HarvardX_CS50x_India_Bachelor's}                 => {MITx_6.00x_India_Bachelor's}                       0.002016982 0.26918489 11.564304  677
[4]  {MITx_6.00x_India_Bachelor's}                     => {HarvardX_CS50x_India_Bachelor's}                   0.002016982 0.08665045 11.564304  677
[5]  {MITx_8.02x_India_Bachelor's}                     => {MITx_6.002x_India_Bachelor's}                      0.002496648 0.38564197 16.958041  838
[6]  {MITx_6.002x_India_Bachelor's}                    => {MITx_8.02x_India_Bachelor's}                       0.002496648 0.10978645 16.958041  838
[7]  {MITx_3.091x_UnitedStates_Secondary}              => {MITx_6.00x_UnitedStates_Secondary}                 0.001516461 0.21024370  7.981936  509
[8]  {MITx_6.00x_UnitedStates_Secondary}               => {MITx_3.091x_UnitedStates_Secondary}                0.001516461 0.05757267  7.981936  509
[9]  {MITx_6.002x_UnitedStates_Secondary}              => {MITx_6.00x_UnitedStates_Secondary}                 0.001939520 0.28194023 10.703907  651
[10] {MITx_6.00x_UnitedStates_Secondary}               => {MITx_6.002x_UnitedStates_Secondary}                0.001939520 0.07363420 10.703907  651
[11] {HarvardX_CB22x_UnitedStates_Master's}            => {HarvardX_ER22x_UnitedStates_Master's}              0.001415165 0.15785975 14.592571  475
[12] {HarvardX_ER22x_UnitedStates_Master's}            => {HarvardX_CB22x_UnitedStates_Master's}              0.001415165 0.13081796 14.592571  475
[13] {MITx_3.091x_UnitedStates_Bachelor's}             => {MITx_6.002x_UnitedStates_Bachelor's}               0.001021898 0.14109420 12.758154  343
[14] {MITx_6.002x_UnitedStates_Bachelor's}             => {MITx_3.091x_UnitedStates_Bachelor's}               0.001021898 0.09240302 12.758154  343
[15] {MITx_3.091x_UnitedStates_Bachelor's}             => {MITx_6.00x_UnitedStates_Bachelor's}                0.001558171 0.21513780  5.088149  523
[16] {HarvardX_CB22x_UnitedStates_Secondary}           => {HarvardX_ER22x_UnitedStates_Secondary}             0.001540295 0.19240789 19.107014  517
[17] {HarvardX_ER22x_UnitedStates_Secondary}           => {HarvardX_CB22x_UnitedStates_Secondary}             0.001540295 0.15295858 19.107014  517
[18] {MITx_8.02x_UnitedStates_Bachelor's}              => {MITx_6.002x_UnitedStates_Bachelor's}               0.001391330 0.21620370 19.549777  467
[19] {MITx_6.002x_UnitedStates_Bachelor's}             => {MITx_8.02x_UnitedStates_Bachelor's}                0.001391330 0.12580819 19.549777  467
[20] {MITx_8.02x_UnitedStates_Bachelor's}              => {MITx_6.00x_UnitedStates_Bachelor's}                0.001313869 0.20416667  4.828674  441
[21] {MITx_8.02x_India_Secondary}                      => {MITx_6.00x_India_Secondary}                        0.001313869 0.18208092  8.911558  441
[22] {MITx_6.00x_India_Secondary}                      => {MITx_8.02x_India_Secondary}                        0.001313869 0.06430446  8.911558  441
[23] {MITx_8.02x_India_Secondary}                      => {MITx_6.002x_India_Secondary}                       0.002800536 0.38810900 19.011790  940
[24] {MITx_6.002x_India_Secondary}                     => {MITx_8.02x_India_Secondary}                        0.002800536 0.13718622 19.011790  940
[25] {HarvardX_CS50x_India_Secondary}                  => {MITx_6.00x_India_Secondary}                        0.002681365 0.29392554 14.385551  900
[26] {MITx_6.00x_India_Secondary}                      => {HarvardX_CS50x_India_Secondary}                    0.002681365 0.13123360 14.385551  900
[27] {HarvardX_CS50x_India_Secondary}                  => {MITx_6.002x_India_Secondary}                       0.001290034 0.14141084  6.927109  433
[28] {MITx_6.002x_India_Secondary}                     => {HarvardX_CS50x_India_Secondary}                    0.001290034 0.06319323  6.927109  433
[29] {HarvardX_PH207x_UnitedStates_Bachelor's}         => {MITx_6.00x_UnitedStates_Bachelor's}                0.001284076 0.10613149  2.510079  431
[30] {HarvardX_PH278x_UnitedStates_Bachelor's}         => {HarvardX_CB22x_UnitedStates_Bachelor's}            0.001120215 0.11233941  8.126449  376
[31] {HarvardX_CB22x_UnitedStates_Bachelor's}          => {HarvardX_PH278x_UnitedStates_Bachelor's}           0.001120215 0.08103448  8.126449  376
[32] {HarvardX_PH278x_UnitedStates_Bachelor's}         => {HarvardX_ER22x_UnitedStates_Bachelor's}            0.001707135 0.17119809  9.493249  573
[33] {HarvardX_ER22x_UnitedStates_Bachelor's}          => {HarvardX_PH278x_UnitedStates_Bachelor's}           0.001707135 0.09466380  9.493249  573
[34] {MITx_14.73x_UnitedStates_Bachelor's}             => {HarvardX_ER22x_UnitedStates_Bachelor's}            0.001346641 0.12024475  6.667793  452
[35] {HarvardX_ER22x_UnitedStates_Bachelor's}          => {MITx_14.73x_UnitedStates_Bachelor's}               0.001346641 0.07467372  6.667793  452
[36] {MITx_6.002x_UnitedStates_Bachelor's}             => {MITx_6.00x_UnitedStates_Bachelor's}                0.002818412 0.25484914  6.027347  946
[37] {MITx_6.00x_UnitedStates_Bachelor's}              => {MITx_6.002x_UnitedStates_Bachelor's}               0.002818412 0.06665727  6.027347  946
[38] {HarvardX_CS50x_UnitedStates_Secondary}           => {MITx_6.00x_UnitedStates_Secondary}                 0.002002086 0.12775665  4.850302  672
[39] {MITx_6.00x_UnitedStates_Secondary}               => {HarvardX_CS50x_UnitedStates_Secondary}             0.002002086 0.07600950  4.850302  672
[40] {HarvardX_CB22x_UnitedStates_Bachelor's}          => {HarvardX_ER22x_UnitedStates_Bachelor's}            0.002589006 0.18728448 10.385270  869
[41] {HarvardX_ER22x_UnitedStates_Bachelor's}          => {HarvardX_CB22x_UnitedStates_Bachelor's}            0.002589006 0.14356517 10.385270  869
[42] {HarvardX_CB22x_UnitedStates_Bachelor's}          => {HarvardX_CS50x_UnitedStates_Bachelor's}            0.001063608 0.07693966  3.485128  357
[43] {HarvardX_CB22x_UnitedStates_Bachelor's}          => {MITx_6.00x_UnitedStates_Bachelor's}                0.001024877 0.07413793  1.753410  344
[44] {HarvardX_ER22x_UnitedStates_Bachelor's}          => {HarvardX_CS50x_UnitedStates_Bachelor's}            0.001126173 0.06244837  2.828717  378
[45] {HarvardX_CS50x_UnitedStates_Bachelor's}          => {HarvardX_ER22x_UnitedStates_Bachelor's}            0.001126173 0.05101215  2.828717  378
[46] {MITx_6.00x_India_Secondary}                      => {MITx_6.002x_India_Secondary}                       0.003625801 0.17745698  8.692854 1217
[47] {MITx_6.002x_India_Secondary}                     => {MITx_6.00x_India_Secondary}                        0.003625801 0.17761238  8.692854 1217
[48] {MITx_6.002x_India_Bachelor's}                    => {MITx_6.00x_India_Bachelor's}                       0.003092507 0.13598847  5.842126 1038
[49] {MITx_6.00x_India_Bachelor's}                     => {MITx_6.002x_India_Bachelor's}                      0.003092507 0.13285550  5.842126 1038
[50] {HarvardX_CS50x_UnitedStates_Bachelor's}          => {MITx_6.00x_UnitedStates_Bachelor's}                0.003643676 0.16504723  3.903474 1223
[51] {MITx_6.00x_UnitedStates_Bachelor's}              => {HarvardX_CS50x_UnitedStates_Bachelor's}            0.003643676 0.08617531  3.903474 1223
```

✓Support가 가장 높은 규칙

```
> inspect(sort(rules, by="support"))
     lhs                                                  rhs                                                  support     confidence lift      count
[1]  {HarvardX_CS50x_UnitedStates_Bachelor's}          => {MITx_6.00x_UnitedStates_Bachelor's}                0.003643676 0.16504723  3.903474 1223
[2]  {MITx_6.00x_UnitedStates_Bachelor's}              => {HarvardX_CS50x_UnitedStates_Bachelor's}            0.003643676 0.08617531  3.903474 1223
[3]  {MITx_6.00x_India_Secondary}                      => {MITx_6.002x_India_Secondary}                       0.003625801 0.17745698  8.692854 1217
```

✓Confidence가 가장 높은 규칙

```
> inspect(sort(rules, by="confidence"))
     lhs                                                  rhs                                                  support     confidence lift      count
[1]  {MITx_8.02x_India_Secondary}                      => {MITx_6.002x_India_Secondary}                       0.002800536 0.38810900 19.011790  940
[2]  {MITx_8.02x_India_Bachelor's}                     => {MITx_6.002x_India_Bachelor's}                      0.002496648 0.38564197 16.958041  838
[3]  {HarvardX_CS50x_India_Secondary}                  => {MITx_6.00x_India_Secondary}                        0.002681365 0.29392554 14.385551  900
```

✓lift가 가장 높은 규칙

```
> inspect(sort(rules, by="lift"))
     lhs                                                  rhs                                                  support     confidence lift      count
[1]  {MITx_8.02x_UnitedStates_Bachelor's}              => {MITx_6.002x_UnitedStates_Bachelor's}               0.001391330 0.21620370 19.549777  467
[2]  {MITx_6.002x_UnitedStates_Bachelor's}             => {MITx_8.02x_UnitedStates_Bachelor's}                0.001391330 0.12580819 19.549777  467
[3]  {HarvardX_ER22x_UnitedStates_Secondary}           => {HarvardX_CB22x_UnitedStates_Secondary}             0.001540295 0.15295858 19.107014  517
[4]  {HarvardX_CB22x_UnitedStates_Secondary}           => {HarvardX_ER22x_UnitedStates_Secondary}             0.001540295 0.19240789 19.107014  517
```

✓효용성 지표를 Support × Confidence × Lift로 정의할 때, 1~3위.

```
> df_rules<-as(rules,'data.frame')
> scl<-NA
> df_rules<-cbind(df_rules,scl)
> df_rules$scl<-df_rules$support*df_rules$confidence*df_rules$lift
> df_rules[order(df_rules$scl,decreasing = T),]
```

	rules	support	confidence	lift	count	scl
23	{MITx_8.02x_India_Secondary} => {MITx_6.002x_India_Secondary}	0.002800536	0.38810900	19.011790	940	0.0206641682
5	{MITx_8.02x_India_Bachelor's} => {MITx_6.002x_India_Bachelor's}	0.002496648	0.38564197	16.958041	838	0.0163274116
25	{HarvardX_CS50x_India_Secondary} => {MITx_6.00x_India_Secondary}	0.002681365	0.29392554	14.385551	900	0.0113375620
24	{MITx_6.002x_India_Secondary} => {MITx_8.02x_India_Secondary}	0.002800536	0.13718622	19.011790	940	0.0073042346
3	{HarvardX_CS50x_India_Bachelor's} => {MITx_6.00x_India_Bachelor's}	0.002016982	0.26918489	11.564304	677	0.0062787357
18	{MITx_8.02x_UnitedStates_Bachelor's} => {MITx_6.002x_UnitedStates_Bachelor's}	0.001391330	0.21620370	19.549777	467	0.0058807832
9	{MITx_6.002x_UnitedStates_Secondary} => {MITx_6.00x_UnitedStates_Secondary}	0.001939520	0.28194023	10.703907	651	0.0058532047
16	{HarvardX_CB22x_UnitedStates_Secondary} => {HarvardX_ER22x_UnitedStates_Secondary}	0.001540295	0.19240789	19.107014	517	0.0056626484
47	{MITx_6.002x_India_Secondary} => {MITx_6.00x_India_Secondary}	0.003625801	0.17761238	8.692854	1217	0.0055980857
46	{MITx_6.00x_India_Secondary} => {MITx_6.002x_India_Secondary}	0.003625801	0.17745698	8.692854	1217	0.0055931880
26	{MITx_6.00x_India_Secondary} => {HarvardX_CS50x_India_Secondary}	0.002681365	0.13123360	14.385551	900	0.0050620611
40	{HarvardX_CB22x_UnitedStates_Bachelor's} => {HarvardX_ER22x_UnitedStates_Bachelor's}	0.002589006	0.18728448	10.385270	869	0.0050356170
6	{MITx_6.002x_India_Bachelor's} => {MITx_8.02x_India_Bachelor's}	0.002496648	0.10978645	16.958041	838	0.0046481679
17	{HarvardX_ER22x_UnitedStates_Secondary} => {HarvardX_CB22x_UnitedStates_Secondary}	0.001540295	0.15295858	19.107014	517	0.0045016379
36	{MITx_6.002x_UnitedStates_Bachelor's} => {MITx_6.00x_UnitedStates_Bachelor's}	0.002818412	0.25484914	6.027347	946	0.0043292620
41	{HarvardX_ER22x_UnitedStates_Bachelor's} => {HarvardX_CB22x_UnitedStates_Bachelor's}	0.002589006	0.14356517	10.385270	869	0.0038601128
19	{MITx_6.002x_UnitedStates_Bachelor's} => {MITx_8.02x_UnitedStates_Bachelor's}	0.001391330	0.12580819	19.549777	467	0.0034220075
11	{HarvardX_CB22x_UnitedStates_Master's} => {HarvardX_ER22x_UnitedStates_Master's}	0.001415165	0.15785975	14.592571	475	0.0032599445
32	{HarvardX_PH278x_UnitedStates_Bachelor's} => {HarvardX_ER22x_UnitedStates_Bachelor's}	0.001707135	0.17119809	9.493249	573	0.0027744811

->column명 scl에 support*confidence*lift 값을 저장하여 정렬.

<1위>

{MITx_8.02x_India_Secondary} => {MITx_6.002x_India_Secondary}

scl : 0.0206641682

<2위>

{MITx_8.02x_India_Bachelor's} => {MITx_6.002x_India_Bachelor's}

scl : 0.0163274116

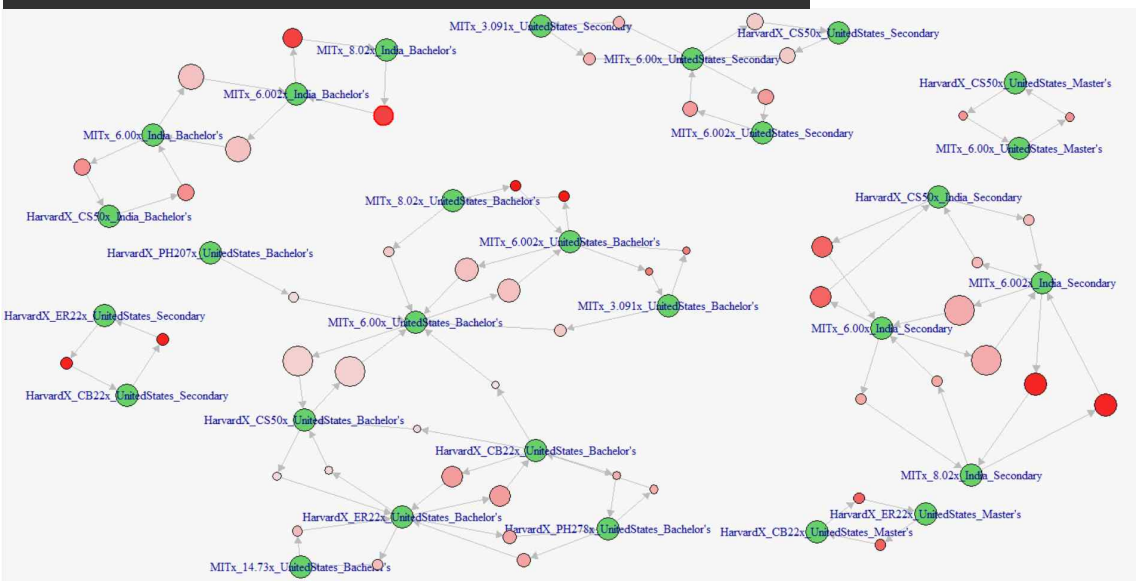
<3위>

{HarvardX_CS50x_India_Secondary} => {MITx_6.00x_India_Secondary}

scl : 0.0113375620

✓두 아이템이 서로 조건절/결과절을 달리해서 생성되는 경우가 존재.

```
plot(rules, method="graph",engine='interactive')
```



<1>
{HarvardX_CS50x_UnitedStates_Bachelor's} => {MITx_6.00x_UnitedStates_Bachelor's}
->support : 0.003643676 // confidence : 0.16504723 // lift : 3.903474

{MITx_6.00x_UnitedStates_Bachelor's} => {HarvardX_CS50x_UnitedStates_Bachelor's}
->support : 0.003643676 // confidence : 0.08617531 // lift : 3.903474

<2>
{MITx_8.02x_India_Secondary} => {MITx_6.002x_India_Secondary}
->support : 0.002800536 // confidence : 0.38810900 // lift : 19.011790

{MITx_6.002x_India_Secondary} => {MITx_8.02x_India_Secondary}
->support : 0.002800536 // confidence : 0.13718622 // lift : 19.011790

<3>
{HarvardX_PH278x_UnitedStates_Bachelor's}
=> {HarvardX_ER22x_UnitedStates_Bachelor's}
->support : 0.001707135 // confidence : 0.17119809 // lift : 9.493249

{HarvardX_ER22x_UnitedStates_Bachelor's}
=> {HarvardX_PH278x_UnitedStates_Bachelor's}
->support : 0.001707135 // confidence : 0.09466380 // lift : 9.493249

(conclusion)
->세 가지 규칙 모두 support와 lift 값은 같게 나오지만,
confidence는 다르게 나옴을 알 수 있음.

<A : 조건절 / B : 결과절>

-많은 프로그램에서 support = $P(A \cap B)$
조건절/결과절이 바뀌어도 support값은 바뀌지 않음을 알 수 있다.

-lift = $\frac{P(A \cap B)}{P(A) * P(B)}$
support처럼 조건절/결과절이 바뀌어도 값이 바뀌지 않는다.

-confidence의 경우 $confidence(A \rightarrow B) = \frac{P(A \cap B)}{P(A)}$
조건절/결과절이 바뀔 경우 값이 바뀌게 됨을 알 수 있다.

[Extra Question]

-수업시간엔 association rule analysis algorithm으로 A priori algorithm을 학습해보았기에 다른 알고리즘을 찾아보았습니다.

비록 코딩을 통한 데이터 시각화 및 분석은 진행하지 못하였기에, 완벽한 이해 및 지식습득이 이루어지진 못하였으나 공부한 내용을 정리해 보겠습니다.

A priori algorithm외에도

AIS algorithm이 있다는 것을 알게 되었습니다.

<AIS algorithm>

-> 1.Candidate itemsets이 data가 스캔될 때 즉시 생성되고 세어짐(counted).

2.각 거래에 포함되는 이전 단계의 large itemsets을 결정.

3.New candidate itemsets은 이러한 large itemsets를 이 transaction의 다른 items로 확장함으로써 생성됨.

Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

Itemset	Support
{1}	2
{2}	3
{3}	3
{5}	3

C_2

Itemset	Support
{1 3}*	2
{1 4}	1
{3 4}	1
{2 3}*	2
{2 5}*	3
{3 5}*	2
{1 2}	1
{1 5}	1

C_3

Itemset	Support
{1 3 4}	1
{2 3 5}*	2
{1 3 5}	1

단점 : 너무 많은 candidate itemsets을 불필요하게 생성 및 계수하게 됨.