<다변량 분석 리포트 1>

2014170852 산업경영공학부 조영관

<STEP 1 데이터 변환>

주어진 데이터를 활용해서, 필요한 변수인 userid_DI, Institute, course_id, final_cc_cnameDI, LOE_DI를 뽑은 후 변수 이름을 바꾸고 하나의 변수로 합치는 과정을 진행했습니다.

```
library(dplyr)

# 데이터 불러오기
mooc_dataset <- read.csv("big_student_clear_third_version.csv")

# 필요한 데이터 선별
dataset <- mooc_dataset %>% select(userid_DI, institute, course_id, final_cc_cname_DI, LoE_DI)

# 변수 이름 변경
names(dataset)
names(dataset)[1] <- "TransactionID"
names(dataset)[2] <- "Institute"
names(dataset)[3] <- "Course"
names(dataset)[4] <- "Region"
names(dataset)[5] <- "Degree"

# 공백 제거
dataset$Region <- gsub(" ","", dataset$Region)

# 변수 4개 합치기
RawTransactions <- paste(dataset$Institute, dataset$Course, dataset$Region, dataset$Degree, sep="_")

# Transaction_ID 와 연결하기
MOOC_transactions <- paste(dataset$TransactionID, RawTransactions, sep=" ")
MOOC_transactions <- data.frame(MOOC_transactions)

# csv로 저장하기
write.csv(MOOC_transactions, file= "MOOC_User_Course.csv", row.names=FALSE)
```

그 결과 다음과 같은 형태로 파일이 저장되었습니다.

MHxPC130313697	HarvardX_PH207x_India_Bachelor's
MHxPC130237753	HarvardX_PH207x_UnitedStates_Secondary
MHxPC130202970	HarvardX_CS50x_UnitedStates_Bachelor's
MHxPC130223941	HarvardX_CS50x_OtherMiddleEast/CentralAsia_Secondary
MHxPC130317399	HarvardX_PH207x_Australia_Master's
MHxPC130191782	HarvardX_CS50x_Pakistan_Bachelor's
MHxPC130191782	HarvardX_ER22x_Pakistan_Bachelor's
MHxPC130267000	HarvardX_PH207x_OtherSouthAsia_Master's
MHxPC130435800	HarvardX_CS50x_India_Bachelor's
MHxPC130284813	HarvardX_PH207x_UnitedStates_Bachelor's
MHxPC130235150	HarvardX_CS50x_India_Bachelor's
MHxPC130001411	HarvardX_CS50x_OtherEurope_Secondary
MHxPC130396873	HarvardX_PH207x_UnitedStates_Bachelor's
MHxPC130469401	HarvardX_CB22x_OtherMiddleEast/CentralAsia_Bachelor's
MHxPC130469401	HarvardX_CS50x_OtherMiddleEast/CentralAsia_Bachelor's

<STEP 2 데이터 불러오기 및 기초 통계량 확인>

이제 위에서 저장한 파일을 불러오겠습니다.

```
# 데이터 불러오기
library(arules)
library(arulesViz)
library(wordcloud)
data_single <- read.transactions("MOOC_User_Course.csv", format = "single",cols=c(1,2),rm.duplicates=TRUE)

# 데이터 속성 파악
summary(data_single)

inspect(data_single)
data_df <- as(data_single,"data.frame")
str(data_single)
itemInfo(data_single)
```

위와 같이 코드를 작성하여 transaction을 불러와 데이터의 특성을 파악해보았습니다.

Summary를 해보면 다음과 같습니다.

```
> summary(data_single)
transactions as itemMatrix in sparse format with
 333946 rows (elements/itemsets/transactions) and
 4092 columns (items) and a density of 0.00030072

most frequent items:
    MITx_6.00x_UnitedStates_Bachelor's      MITx_6.00x_UnitedStates_Secondary
                                 14092                                   8774
          MITx_6.00x_India_Bachelor's           MITx_6.002x_India_Bachelor's
                                  7774                                   7596
HarvardX_CS50x_UnitedStates_Bachelor's                                 (Other)
                                  7356                                 365344

element (itemset/transaction) length distribution:
sizes
     1      2      3      4      5      6      7      8      9     10     11     12     13
277491  42477   9872   2777    795    289    107     45     36     23     19      9      6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.231   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
```

가장 빈번하게 나온 아이템 상위 5개를 확인할 수 있습니다.

제일 많은 아이템은 MIT/6.00/미국/학사임을 알 수 있습니다.

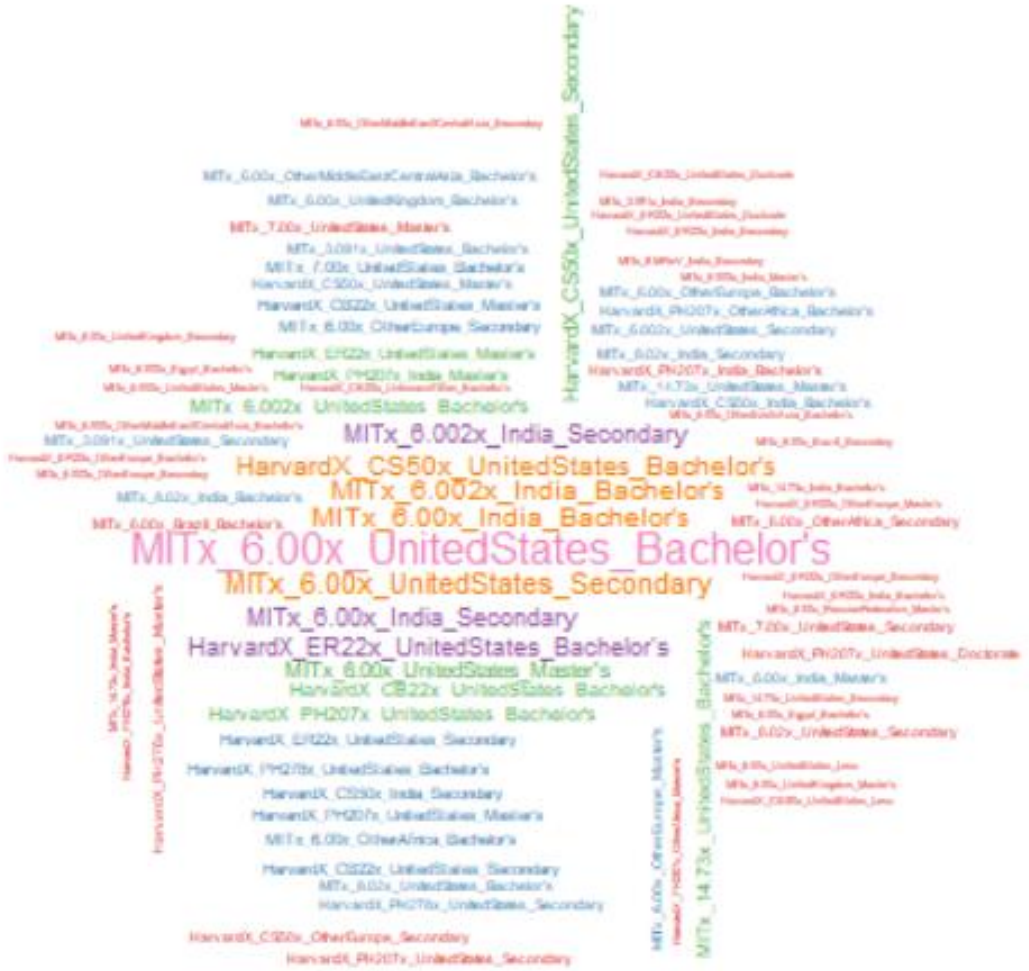
그리고 아이템 조합을 보면 대부분이 1개 단일 set임을 알 수 있지만, 아이템 조합이 2개 이상인 set의 total 개수도 약 55,000개 정도 있음을 파악할 수 있습니다.

워드 클라우드를 생성해보겠습니다.

```
# wordcloud 생성
itemName <- itemLabels(data_single)
itemCount <- itemFrequency(data_single)*nrow(data_single)
testing <- as(itemCount,"data.frame")

col <- brewer.pal(8, "Set1")
wordcloud(words=itemName, freq=itemCount, min.freq = 1000, scale=c(1.5,0.2),col=col,random.order=FALSE)
```

결과는 다음과 같습니다.

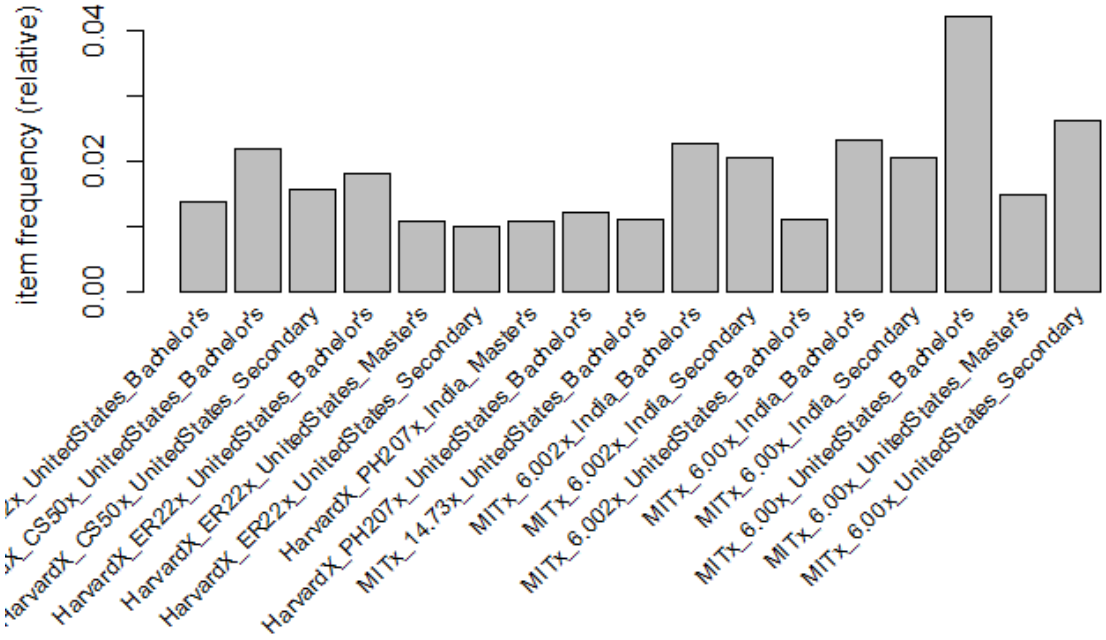


위의 summary에서 살펴본 대로 상위 5개의 가장 빈번한 아이템이 가운데 제일 크게 분포하는 것을 확인할 수 있습니다. 그 외 다양한 아이템들이 word cloud 내에 분포함을 확인할 수 있습니다.

ItemFrequencyPlot을 생성해보겠습니다.

```
# itemFrequencyPlot 생성
itemFrequencyPlot(data_single, support = 0.01, cex.names=0.8)
```

결과는 다음과 같습니다.



상위 5개의 아이템을 확인해보겠습니다.

MITX_6.00x_UnitedStates_Bachelor's

MITX_6.00x_UnitedStates_Secondary

MITx_6.00x_India_Bachelor's

MITx_6.002x_India_Bachelor's

HarvardX_CS50x_UnitedStates_Bachelor's

이 5개의 아이템들의 국가를 확인해보면 United States, India임을 확인할 수 있습니다.

<STEP 3 규칙 생성 및 결과 해석>

Support와 confidence를 조정해가며 몇 가지 규칙이 생성되는지 확인해보겠습니다.

```
# rule 생성하기
rules <- apriori(data_single, parameter=list(support=0.001, confidence=0.05))
```

결과는 다음과 같습니다.

Number of rules	Confidence=0.05	Confidence=0.075	Confidence=0.1	Confidence=0.125
Support=0.001	51	42	34	30
Support=0.0015	27	25	21	20
Support=0.002	18	17	15	14
Support=0.0025	16	15	14	13

Support=0.001, confidence=0.05로 설정했을 때 생성되는 연관규칙분석들을 살펴보겠습니다.

```
> inspect(rules)
     lhs                                                rhs                                                support     confidence lift     count
[1]  {HarvardX_CS50x_UnitedStates_Master's}          => {MITx_6.00x_UnitedStates_Master's}                 0.001203787 0.16883662 11.381170  402
[2]  {MITx_6.00x_UnitedStates_Master's}              => {HarvardX_CS50x_UnitedStates_Master's}             0.001203787 0.08114655 11.381170  402
[3]  {HarvardX_CS50x_India_Bachelor's}               => {MITx_6.00x_India_Bachelor's}                      0.002000323 0.26709316 11.473462  668
[4]  {MITx_6.00x_India_Bachelor's}                   => {HarvardX_CS50x_India_Bachelor's}                  0.002000323 0.08592745 11.473462  668
[5]  {MITx_8.02x_India_Bachelor's}                   => {MITx_6.002x_India_Bachelor's}                     0.002503399 0.38596491 16.968331  836
[6]  {MITx_6.002x_India_Bachelor's}                  => {MITx_8.02x_India_Bachelor's}                      0.002503399 0.11005793 16.968331  836
[7]  {MITx_3.091x_UnitedStates_Secondary}            => {MITx_6.00x_UnitedStates_Secondary}                0.001497248 0.20868114  7.942584  500
[8]  {MITx_6.00x_UnitedStates_Secondary}             => {MITx_3.091x_UnitedStates_Secondary}               0.001497248 0.05698655  7.942584  500
[9]  {MITx_6.002x_UnitedStates_Secondary}            => {MITx_6.00x_UnitedStates_Secondary}                0.001940433 0.28235294 10.746596  648
[10] {MITx_6.00x_UnitedStates_Secondary}             => {MITx_6.002x_UnitedStates_Secondary}               0.001940433 0.07385457 10.746596  648
[11] {HarvardX_CB22x_UnitedStates_Master's}          => {HarvardX_ER22x_UnitedStates_Master's}             0.001413402 0.15775401 14.593164  472
[12] {HarvardX_ER22x_UnitedStates_Master's}          => {HarvardX_CB22x_UnitedStates_Master's}             0.001413402 0.13074792 14.593164  472
[13] {MITx_3.091x_UnitedStates_Bachelor's}           => {MITx_6.002x_UnitedStates_Bachelor's}              0.001024118 0.14173228 12.830287  342
[14] {MITx_6.002x_UnitedStates_Bachelor's}           => {MITx_3.091x_UnitedStates_Bachelor's}              0.001024118 0.09270805 12.830287  342
[15] {MITx_3.091x_UnitedStates_Bachelor's}           => {MITx_6.00x_UnitedStates_Bachelor's}               0.001554143 0.21508496  5.096988  519
[16] {HarvardX_CB22x_UnitedStates_Secondary}         => {HarvardX_ER22x_UnitedStates_Secondary}            0.001533182 0.19197600 19.114376  512
[17] {HarvardX_ER22x_UnitedStates_Secondary}         => {HarvardX_CB22x_UnitedStates_Secondary}            0.001533182 0.15265355 19.114376  512
[18] {MITx_8.02x_UnitedStates_Bachelor's}            => {MITx_6.002x_UnitedStates_Bachelor's}              0.001392441 0.21668220 19.615114  465
[19] {MITx_6.002x_UnitedStates_Bachelor's}           => {MITx_8.02x_UnitedStates_Bachelor's}               0.001392441 0.12605042 19.615114  465
[20] {MITx_8.02x_UnitedStates_Bachelor's}            => {MITx_6.00x_UnitedStates_Bachelor's}               0.001314584 0.20456664  4.847730  439
[21] {MITx_8.02x_India_Secondary}                    => {MITx_6.00x_India_Secondary}                       0.001317578 0.18196857  8.903689  440
[22] {MITx_6.00x_India_Secondary}                    => {MITx_8.02x_India_Secondary}                       0.001317578 0.06446886  8.903689  440
[23] {MITx_8.02x_India_Secondary}                    => {MITx_6.002x_India_Secondary}                      0.002790870 0.38544251 18.873458  932
[24] {MITx_6.002x_India_Secondary}                   => {MITx_8.02x_India_Secondary}                       0.002790870 0.13665689 18.873458  932
[25] {HarvardX_CS50x_India_Secondary}                => {MITx_6.00x_India_Secondary}                       0.002659113 0.29095675 14.236460  888

[26] {MITx_6.00x_India_Secondary}                    => {HarvardX_CS50x_India_Secondary}                   0.002659113 0.13010989 14.236460  888
[27] {HarvardX_CS50x_India_Secondary}                => {MITx_6.002x_India_Secondary}                      0.001290628 0.14121887  6.914879  431
[28] {MITx_6.002x_India_Secondary}                   => {HarvardX_CS50x_India_Secondary}                   0.001290628 0.06319648  6.914879  431
[29] {HarvardX_PH207x_UnitedStates_Bachelor's}       => {MITx_6.00x_UnitedStates_Bachelor's}               0.001287633 0.10656753  2.525390  430
[30] {HarvardX_PH278x_UnitedStates_Bachelor's}       => {HarvardX_CB22x_UnitedStates_Bachelor's}           0.001122936 0.11247750  8.160203  375
[31] {HarvardX_CB22x_UnitedStates_Bachelor's}        => {HarvardX_PH278x_UnitedStates_Bachelor's}          0.001122936 0.08146861  8.160203  375
[32] {HarvardX_PH278x_UnitedStates_Bachelor's}       => {HarvardX_ER22x_UnitedStates_Bachelor's}           0.001706863 0.17096581  9.502887  570
[33] {HarvardX_ER22x_UnitedStates_Bachelor's}        => {HarvardX_PH278x_UnitedStates_Bachelor's}          0.001706863 0.09487350  9.502887  570
[34] {MITx_14.73x_UnitedStates_Bachelor's}           => {HarvardX_ER22x_UnitedStates_Bachelor's}           0.001326562 0.11931053  6.631703  443
[35] {HarvardX_ER22x_UnitedStates_Bachelor's}        => {MITx_14.73x_UnitedStates_Bachelor's}              0.001326562 0.07373502  6.631703  443
[36] {MITx_6.002x_UnitedStates_Bachelor's}           => {MITx_6.00x_UnitedStates_Bachelor's}               0.002832793 0.25643806  6.076956  946
[37] {MITx_6.00x_UnitedStates_Bachelor's}            => {MITx_6.002x_UnitedStates_Bachelor's}              0.002832793 0.06713029  6.076956  946
[38] {HarvardX_CS50x_UnitedStates_Secondary}         => {MITx_6.00x_UnitedStates_Secondary}                0.001982356 0.12667432  4.821334  662
[39] {MITx_6.00x_UnitedStates_Secondary}             => {HarvardX_CS50x_UnitedStates_Secondary}            0.001982356 0.07545019  4.821334  662
[40] {HarvardX_CB22x_UnitedStates_Bachelor's}        => {HarvardX_ER22x_UnitedStates_Bachelor's}           0.002578261 0.18705192 10.397011  861
[41] {HarvardX_ER22x_UnitedStates_Bachelor's}        => {HarvardX_CB22x_UnitedStates_Bachelor's}           0.002578261 0.14330892 10.397011  861
[42] {HarvardX_CB22x_UnitedStates_Bachelor's}        => {HarvardX_CS50x_UnitedStates_Bachelor's}           0.001063046 0.07712362  3.501240  355
[43] {HarvardX_CB22x_UnitedStates_Bachelor's}        => {MITx_6.00x_UnitedStates_Bachelor's}               0.001009145 0.07321312  1.734972  337
[44] {HarvardX_ER22x_UnitedStates_Bachelor's}        => {HarvardX_CS50x_UnitedStates_Bachelor's}           0.001128925 0.06274967  2.848695  377
[45] {HarvardX_CS50x_UnitedStates_Bachelor's}        => {HarvardX_ER22x_UnitedStates_Bachelor's}           0.001128925 0.05125068  2.848695  377
[46] {MITx_6.00x_India_Secondary}                    => {MITx_6.002x_India_Secondary}                      0.003638313 0.17802198  8.716969 1215
[47] {MITx_6.002x_India_Secondary}                   => {MITx_6.00x_India_Secondary}                       0.003638313 0.17815249  8.716969 1215
[48] {MITx_6.002x_India_Bachelor's}                  => {MITx_6.00x_India_Bachelor's}                      0.003084331 0.13559768  5.824840 1030
[49] {MITx_6.00x_India_Bachelor's}                   => {MITx_6.002x_India_Bachelor's}                     0.003084331 0.13249293  5.824840 1030
[50] {HarvardX_CS50x_UnitedStates_Bachelor's}        => {MITx_6.00x_UnitedStates_Bachelor's}               0.003605373 0.16367591  3.878720 1204
[51] {MITx_6.00x_UnitedStates_Bachelor's}            => {HarvardX_CS50x_UnitedStates_Bachelor's}           0.003605373 0.08543855  3.878720 1204
```

총 51개의 규칙이 생성됩니다.

```r
rules_df <- as(rules,"data.frame")

# 효용성 지표 추가
rules_df <- rules_df %>% mutate(index = support*confidence*lift)

rules_sort_support <- rules_df %>% arrange(desc(support))
rules_sort_confidence <- rules_df %>% arrange(desc(support))
rules_sort_lift <- rules_df %>% arrange(desc(support))
rules_sort_index <- rules_df %>% arrange(desc(index))
```

위 코드를 작성하여 각각의 support, confidence, lift가 가장 높은 규칙을 파악하고 효용성이 가장 높은 규칙 1~3위를 파악해보았습니다.

Support가 가장 높은 규칙

	rules	support	confidence	lift
1	{MITx_6.00x_India_Secondary} => {MITx_6.002x_India_Secondary}	0.003638313	0.17802198	8.716969

Confidence가 가장 높은 규칙

	rules	support	confidence	lift
1	{MITx_8.02x_India_Bachelor's} => {MITx_6.002x_India_Bachelor's}	0.002503399	0.38596491	16.968331

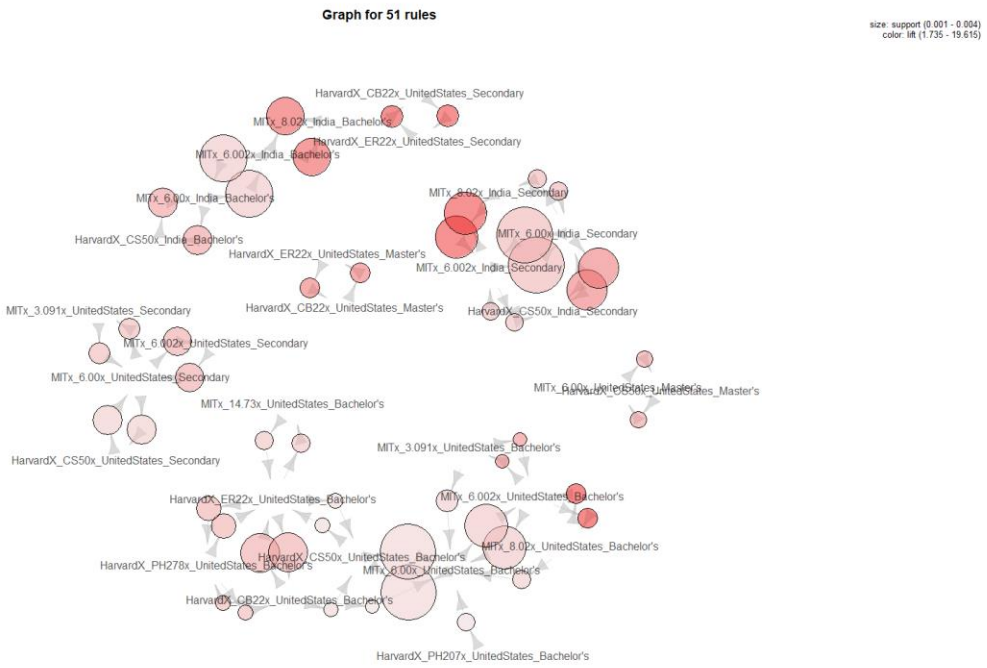
Lift가 가장 높은 규칙

	rules	support	confidence	lift
1	{MITx_6.002x_UnitedStates_Bachelor's} => {MITx_8.02x_UnitedStates_Bachelor's}	0.001392441	0.12605042	19.615114

효용성 지표를 위의 세 지표의 곱으로 정의한 후 효용성이 가장 높은 규칙 1~3위를 나열했습니다.

	rules	support	confidence	lift	count	index
1	{MITx_8.02x_India_Secondary} => {MITx_6.002x_India_Secondary}	0.002790870	0.38544251	18.873458	932	0.0203025584
2	{MITx_8.02x_India_Bachelor's} => {MITx_6.002x_India_Bachelor's}	0.002503399	0.38596491	16.968331	836	0.0163952095
3	{HarvardX_CS50x_India_Secondary} => {MITx_6.00x_India_Secondary}	0.002659113	0.29095675	14.236460	888	0.0110145608

Plot 함수의 graph method를 활용해 시각화를 해보았습니다.



위의 규칙 결과 중 조건절/결과절이 뒤바뀌어 생성되는 경우가 존재합니다.

이 중 3가지 규칙을 선택했습니다.

	rules	support	confidence	lift
1	{MITx_6.002x_UnitedStates_Bachelor's} => {MITx_8.02x_UnitedStates_Bachelor's}	0.001392441	0.12605042	19.615114
2	{MITx_8.02x_UnitedStates_Bachelor's} => {MITx_6.002x_UnitedStates_Bachelor's}	0.001392441	0.21668220	19.615114
3	{HarvardX_CB22x_UnitedStates_Secondary} => {HarvardX_ER22x_UnitedStates_Secondary}	0.001533182	0.19197600	19.114376
4	{HarvardX_ER22x_UnitedStates_Secondary} => {HarvardX_CB22x_UnitedStates_Secondary}	0.001533182	0.15265355	19.114376
11	{HarvardX_CS50x_India_Secondary} => {MITx_6.00x_India_Secondary}	0.002659113	0.29095675	14.236460
12	{MITx_6.00x_India_Secondary} => {HarvardX_CS50x_India_Secondary}	0.002659113	0.13010989	14.236460

각각의 support, confidence, lift 값은 오른쪽과 같습니다.

확인해보면, 조건절과 결과절이 서로 뒤바뀌어 있음을 확인할 수 있습니다.

3개의 지표 중에는 confidence 값에 차이가 있었습니다. Support와 lift는 값이 동일했습니다.

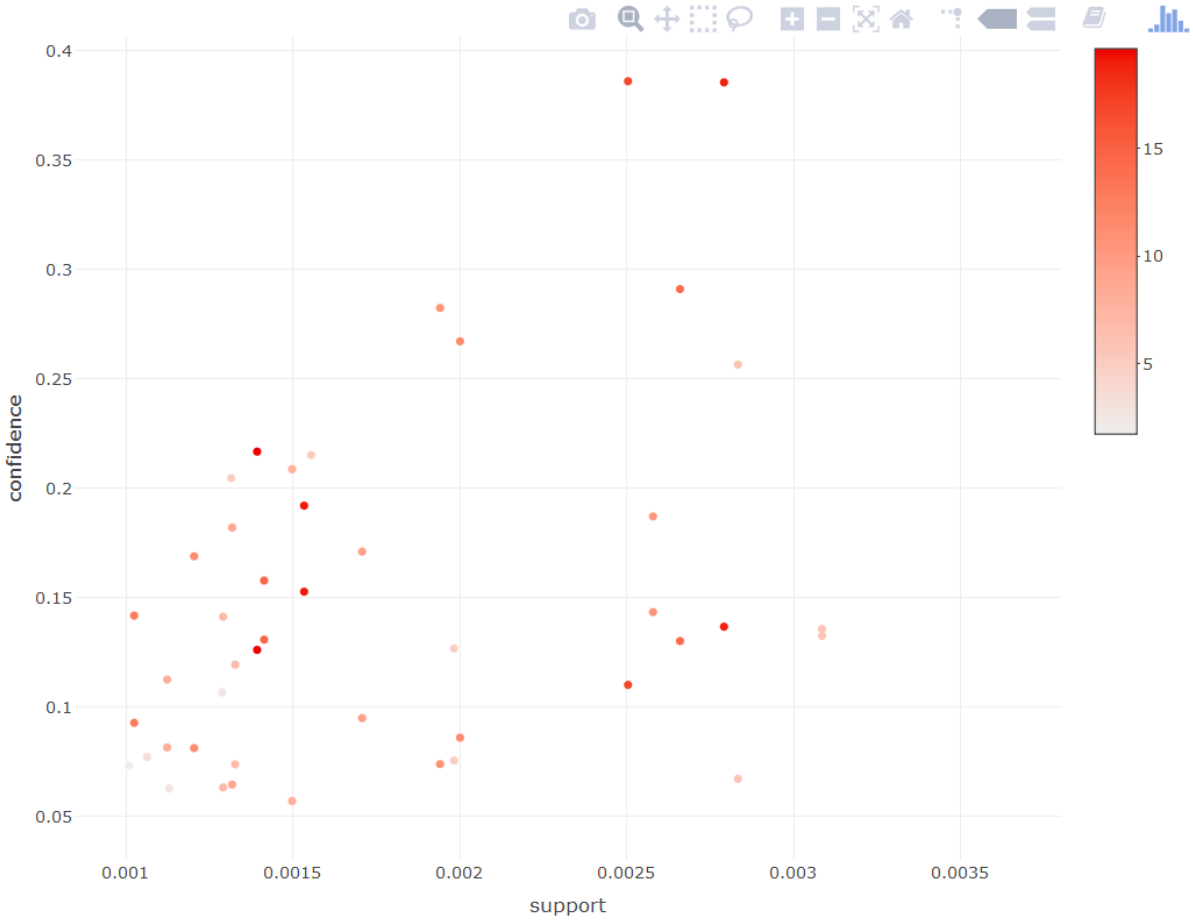
Confidence가 값이 다르게 나올 수밖에 없는 이유는 confidence를 정의하는 식에 있습니다.

예를 들어, A->B의 Confidence는 A,B의 교집합의 확률을 분자로 갖고 A의 확률을 분모로 가지는 반면, B->A의 Confidence는 A,B의 교집합의 확률을 분자로 갖고 B의 확률을 분모로 가집니다.

따라서 분모 값이 달라집니다. 그래서 조건절과 결과절이 뒤바뀌면 Confidence 값이 달라질 수밖에 없습니다.

<추가 extra 시각화>

추가 extra 시각화는 박종범, 임재인 학우와 같이 검색하고 고민해보았습니다.

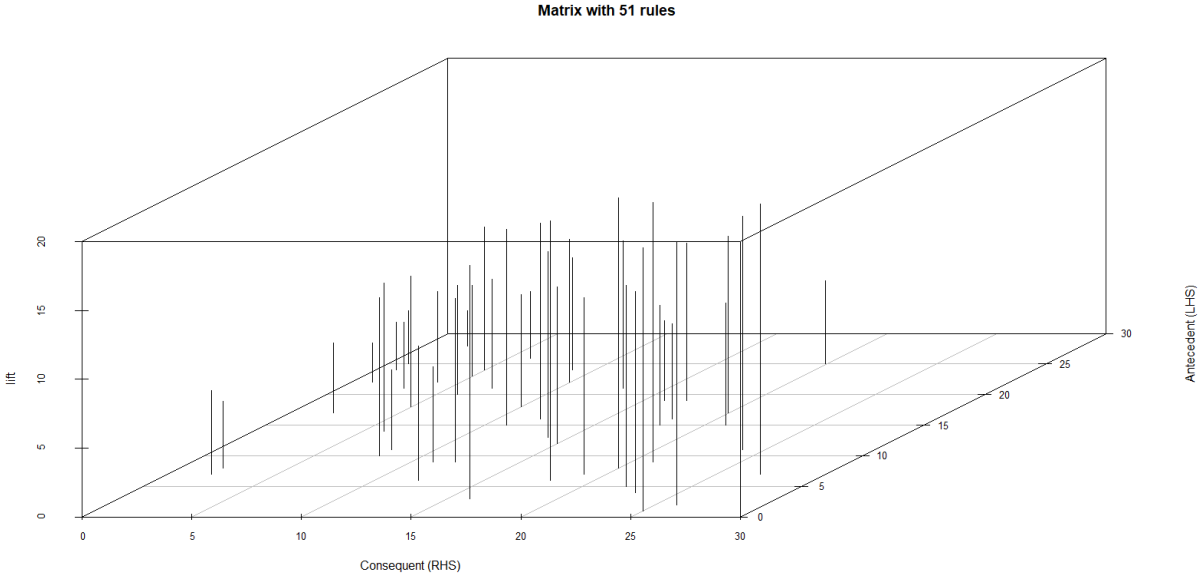


Scatter plot을 통해 각각의 규칙들의 지표를 이용한 분포를 확인해 보았습니다. 특정 경향성을 띄지는 않고, support 0.002 이하 confidence 0.2 이하에 많이 몰려 있음을 확인할 수 있습니다.

plot 시각화에 grouped method도 있는 것을 확인할 수 있었습니다.

원의 크기가 support를 나타내며, 색의 진한 정도가 lift를 나타냅니다.

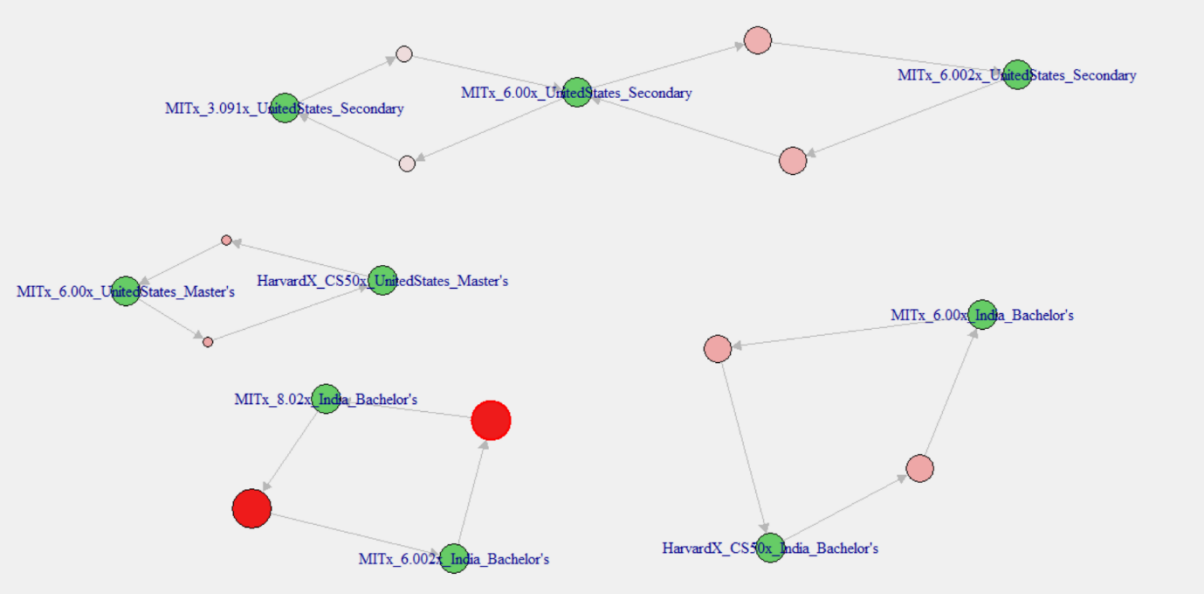
위를 예시로 보시면, 저 10개의 규칙 중에서는 제일 왼쪽 두개의 규칙이 support와 lift가 제일 큼을 확인할 수 있습니다.



위의 그림은 matrix3D method를 이용한 plot입니다.

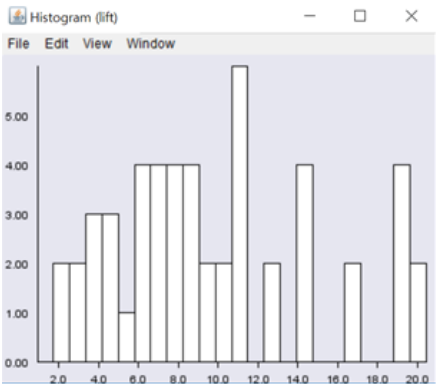
선의 높이가 lift를 나타내며, 가로가 결과절, 세로가 조건절입니다.

Lift가 눈에 띄게 차이 나지 않는 이상 일반 matrix method가 더 시각적으로 구분하기 쉽다고 생각합니다.



위에서 했던 graph method에서 10개의 rule을 선택하고 색을 바꿔 더 눈에 띌 수 있도록 하였습니다. 조건절과 결과절이 반대인 것이 도식화된 것을 확인할 수 있으며, 규칙마다 support와 lift의 차이를 어느 정도 명확하게 확인할 수 있습니다.

마지막으로 iplots method를 통해 support, confidence, lift 각각의 histogram 분포를 파악할 수 있습니다. Scatter plot을 통해 대략적인 시각화를 확인했으므로 lift의 분포만 시각적으로 파악해보겠습니다.



12 이하에 대부분의 규칙들이 몰려 있음을 확인할 수 있습니다.

시각화 참고 자료

https://www.rdocumentation.org/packages/arulesViz/versions/1.3-2/topics/plot