Multivariate Data Analysis

Assignment #7

고려대학교 공과대학

산업경영공학부

2017170857 이우준

[STEP 1] 데이터 변환

[Q1]

1단계

```
#Step 1
mooc <- read.csv("big_student_clear_third_version.csv")
head(mooc)
str(mooc)
#ItemName var save
Institute = mooc["institute"]
Course = mooc ["course_id"]
Region = mooc ["final_cc_cname_DI"]
Degree = mooc ["LoE_DI"]
```

주어진 데이터 "big_student_clear_third_version.csv" 를 mooc 이라는 데이터로 불러와 column 중 institute, course_id, final_cc_cname_DI, LoE_DI 변수 항목의 데이터들을 Institue, Course, Region, Degree에 각각 지정하여 저장하였다.

2단계

```
#Step2
table(Region)
Region1 = gsub('\\s', '', Region$final_cc_cname_DI)
table(Region)
```

1단계에서 정리한 데이터 중 Region 데이터들의 항목에서 공백을 제거해 Region이라는 이름으로 다시 저장해 주었다.

3단계

```
#Step3
RawTransactions = paste(Institute$institute, Course$course_id, Region, Degree$LoE_DI, sep = "_")
head(RawTransactions)
```

1단계와 2단계에서 정리한 4개의 변수들을 밑줄로 연결하여 RawTransactions이라는 변수 하나로 묶어서 저장하였다.

4단계

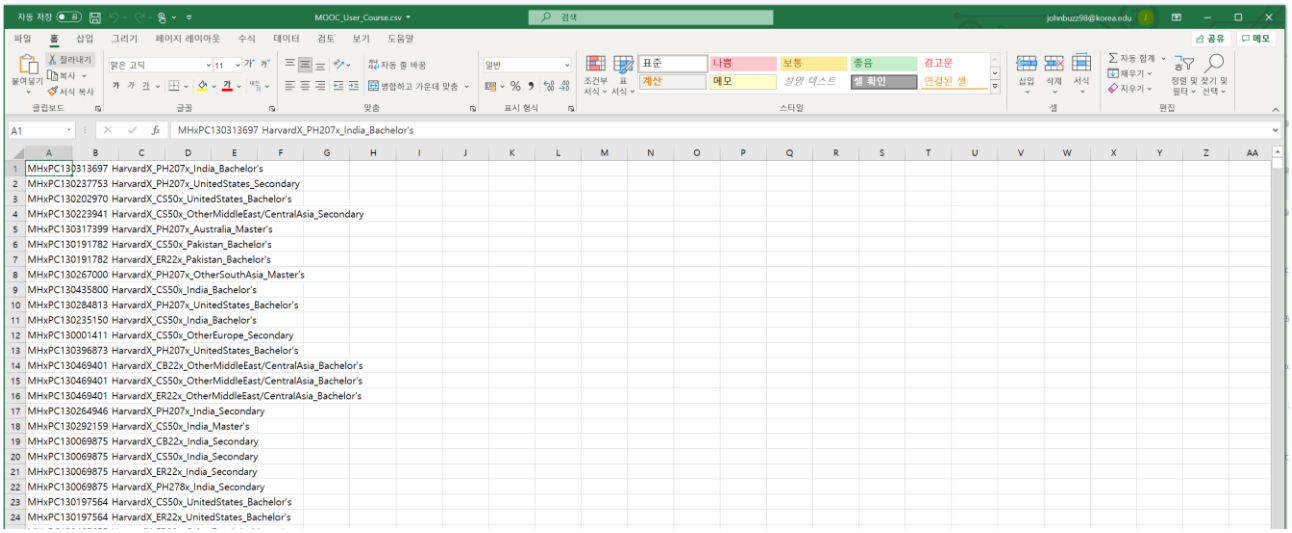
```
#Step4
MOOC_transactions = paste(mooc$userid_DI, RawTransactions, sep = " ")
head(MOOC_transactions)
```

3 단계에서 정리한 RawTransaction 데이터와 Transaction 에 해당하는 변수를 공백으로 연결하여 MOOC_transactions 라는 변수에 저장하였다.

5단계

```
#step5
write.table(MOOC_transactions, "MOOC_User_Course.csv", row.names = FALSE, col.names = FALSE, sep
= ',', quote = FALSE)
```

1~4 단계를 걸쳐 정리한 데이터를 MOOC_User_Course.csv 라는 파일로 저장하였다.



	A	B
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

원본 데이터가 존재하는 폴더안에 MOOC_User_Course 파일이 만들어진 것을 확인할 수 있다.

[STEP 2] 데이터 불러오기 및 기초 통계량 확인

[Q2-1]

```
#Q2-1
single = read.transactions("MOOC_User_Course.csv", format = "single", cols = c(1, 2),
rm.duplicates = TRUE)
inspect(head(single))
summary(single)
```

read.transaction 함수를 이용하여 Q1 에서 저장한 데이터를 singleformat 으로 읽어 single 이라는 변수에 저장하였다.

데이터가 적절하게 저장이 되어있는지 inspect 함수와 head 함수를 통하여 확인하였다.

```
> inspect(head(single))
    items                                                                transactionID
[1] {MITx_14.73x_UnitedKingdom_Secondary}                                   MHxPC130000002
[2] {HarvardX_CS50x_India_Secondary,HarvardX_ER22x_India_Secondary} MHxPC130000004
[3] {HarvardX_ER22x_UnitedStates_Bachelor's}                                MHxPC130000006
[4] {HarvardX_CB22x_UnitedStates_Master's}                                  MHxPC130000007
[5] {MITx_6.00x_UnitedKingdom_Bachelor's}                                   MHxPC130000008
[6] {HarvardX_CS50x_Egypt_Secondary,MITx_6.00x_Egypt_Secondary}      MHxPC130000011
```

각 각 items 와 transactionID 에 저장이 되어 있는 것을 확인할 수 있다.

Summary() 함수를 이용하여 알아본 데이터 속성의 결과는 아래와 같다.

```
> summary(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
```

총 335650 개의 수강데이터로 이루어져 있으며, 개설된 MOOC 강의의 수는 1405 개인 것을 알 수 있다. 이중 빈도 수가 높은 상위 5 개의 강의를 확인 할 수 있다. 가장 빈도 수가 높은 강의는 MITx_6.00x_UnitedStates_Bachelor's 로 총 14192 번 기록한 것을 알 수있다. 강의를 1 개만 수강한 학생의 수는 278440 인 것을 확인 할 수 있다. 학생들이 수강한 강의의 수에 대한 평균은 1.232 이고 최소값, 제 1 분위수, 중간값, 제 3 분위수 모두 1 이고 최대값은 13 임을 알 수 있다. 이를 통해 대부분의 학생은 평균적으로 1 개의 강의만 수강한다는 것을 알 수 있다.

[Q2-2]

```
#Q-2
itemName <- itemLabels(single)
itemCount <- itemFrequency(single)*nrow(single)
col = brewer.pal(8, "Set3")
wordcloud(words = itemName, freq = itemCount, min.freq = 150, scale = c(1.8, 0.1), col = col ,
random.order = FALSE, random.order = TRUE)
```

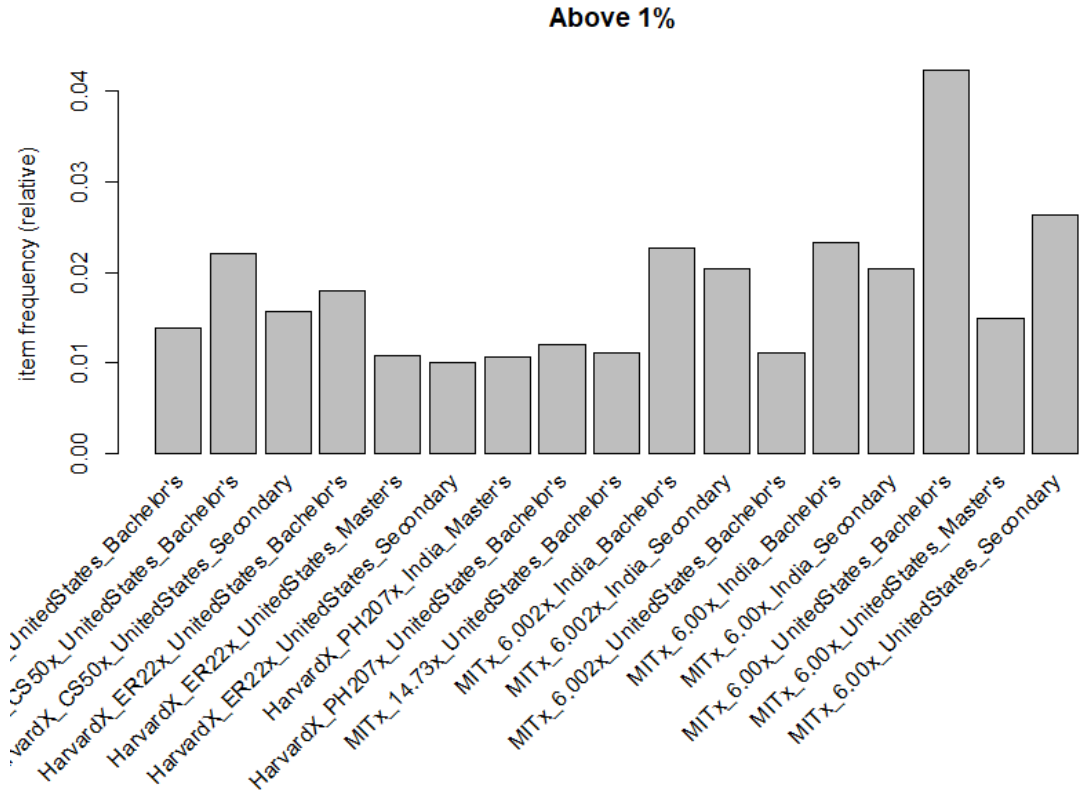
강의 명을 itemName 이라는 변수에 저장하였고, 강의별 빈도수는 itemCount 에 저장하였다. 이를 바탕으로 wordcloud 를 생성하여 확인하였다.

색은 8-class set3 color 옵션을 사용하여 글자의 크기와 색상에 따라 수강 횟수가 달라지는 것을 눈에 띠게 파악 할 수 있다.

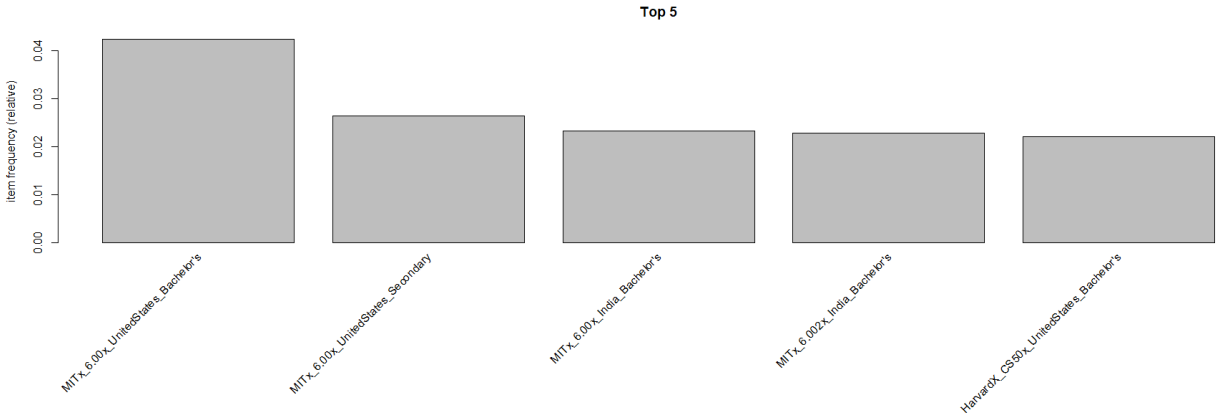
[Q2-3]

```
#Q-3
itemFrequencyPlot(single, support = 0.01, cex.names = 1, , main = "Above 1%")
itemFrequencyPlot(single, support = 0.01, cex.names = 1, , topN = 5, main = "Top 5")
```

itemFrequencyPlot 을 통해 빈도가 1% 이상으로 등장한 item 들을 plot 해 보았다.



이를 통해 빈도가 1% 이상인 item 을 확인 할 수 있다. 아래는 TOP 5 item 이다.



상위 item 5 개에 대하여 각각 접속국가가 어디인지 다음 표에 정리해보았다.

Item	접속국가
MITx_6.00x_UnitedStates_Bachelor's	United States
MITx_6.00x_UnitedStates_Secondary	United States
MITx_6.00x_India_Bachelor's	India
MITx_6.002x_India_Bachelor's	India
HarvardX_CS50x_UnitedStates_Bachelor's	United States

이를 통해 상위 5 개의 item 중 접속국가는 미국이 3 개 인도가 2 개임을 알 수 있다.

[STEP 3] 규칙 생성 및 결과 해석

[Q3-1]

10개의 규칙이 생성될 수 있도록 support는 0.001, 0.005, 0.01, 0.02 총 4가지 경우의 수를 찾아보았고, confidence 또한 0.001, 0.005, 0.01로 총 3가지 경우의 수를 설정하였다. 아래는 총 4*3=12가지의 경우의 규칙의 수를 알아보기 위한 코드이다.

```
#Q3-1
rule1 = apriori(single, parameter = list(support = 0.001, confidence = 0.001))
rule2 = apriori(single, parameter = list(support = 0.001, confidence = 0.005))
rule3 = apriori(single, parameter = list(support = 0.001, confidence = 0.01))
rule4 = apriori(single, parameter = list(support = 0.005, confidence = 0.001))
rule5 = apriori(single, parameter = list(support = 0.005, confidence = 0.005))
rule6 = apriori(single, parameter = list(support = 0.005, confidence = 0.01))
rule7 = apriori(single, parameter = list(support = 0.01, confidence = 0.001))
rule8 = apriori(single, parameter = list(support = 0.01, confidence = 0.005))
rule9 = apriori(single, parameter = list(support = 0.01, confidence = 0.01))
rule10 = apriori(single, parameter = list(support = 0.02, confidence = 0.001))
rule11 = apriori(single, parameter = list(support = 0.02, confidence = 0.005))
rule12 = apriori(single, parameter = list(support = 0.02, confidence = 0.01))
```

해당 코드를 통해 구한 각 rule 의 수를 표를 통해 나타내었다.

		Confidence		
		0.001	0.05	0.01
Support	0.001	307	99	73
	0.005	43	43	17
	0.01	17	17	17
	0.02	7	7	7

이때 Confidence 의 값과 상관 없이 Support 의 값이 0.01 이면 17 개의 rule, 0.02 이면 7 개 rule 이 생성된 것을 알 수 있다.

[Q3-2]

```
rule = apriori(single, parameter=list(support=0.001, confidence=0.05))
summary(rule)
```

분석에 앞서 support 가 0.001 confidence 가 0.05 일 때 생성되는 rule 에 관한 summry 를 확인해 보았다.

```
> summary(rule)
set of 51 rules
rule length distribution (lhs + rhs):sizes
 2
51
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2
summary of quality measures:
    support          confidence          lift           count
 Min.   :0.001022   Min.   :0.05101   Min.   : 1.753   Min.   : 343.0
 1st Qu.:0.001302   1st Qu.:0.08409   1st Qu.: 5.935   1st Qu.: 437.0
 Median :0.001540   Median :0.13286   Median : 8.912   Median : 517.0
 Mean   :0.001915   Mean   :0.14731   Mean   : 9.786   Mean   : 642.8
 3rd Qu.:0.002589   3rd Qu.:0.17985   3rd Qu.:12.758   3rd Qu.: 869.0
 Max.   :0.003644   Max.   :0.38811   Max.   :19.550   Max.   :1223.0
mining info:
   data ntransactions support confidence
 single        335650   0.001       0.05
```

summary 를 통해 support 가 0.001, confidence 가 0.05 일 때 총 2 개의 item 으로 이루어진 51 개의 연관규칙이 생성 된 것을 알 수 있었다.

```
inspect(rule)
inspect(sort(rule, by = "support"))
inspect(sort(rule, by = "confidence"))
inspect(sort(rule, by = "lift"))
```

다음은 Support 를 기준으로 내림차 순으로 정렬한 결과 중 일부이다.

```
> inspect(sort(rule, 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
[4]  {MITx_6.002x_India_Secondary}                   => {MITx_6.00x_India_Secondary}                     0.003625801 0.17761238 8.692854  1217
[5]  {MITx_6.002x_India_Bachelor's}                  => {MITx_6.00x_India_Bachelor's}                    0.003092507 0.13598847 5.842126  1038
[6]  {MITx_6.00x_India_Bachelor's}                   => {MITx_6.002x_India_Bachelor's}                   0.003092507 0.13285550 5.842126  1038
[7]  {MITx_6.002x_UnitedStates_Bachelor's}           => {MITx_6.00x_UnitedStates_Bachelor's}             0.002818412 0.25484914 6.027347   946
[8]  {MITx_6.00x_UnitedStates_Bachelor's}            => {MITx_6.002x_UnitedStates_Bachelor's}            0.002818412 0.06665727 6.027347   946
[9]  {MITx_8.02x_India_Secondary}                    => {MITx_6.002x_India_Secondary}                    0.002800536 0.38810900 19.011790  940
[10] {MITx_6.002x_India_Secondary}                   => {MITx_8.02x_India_Secondary}                     0.002800536 0.13718622 19.011790  940
[11] {HarvardX_CS50x_India_Secondary}                => {MITx_6.00x_India_Secondary}                     0.002681365 0.29392554 14.385551  900
[12] {MITx_6.00x_India_Secondary}                    => {HarvardX_CS50x_India_Secondary}                 0.002681365 0.13123360 14.385551  900
[13] {HarvardX_CB22x_UnitedStates_Bachelor's}        => {HarvardX_ER22x_UnitedStates_Bachelor's}         0.002589006 0.18728448 10.385270  869
[14] {HarvardX_ER22x_UnitedStates_Bachelor's}        => {HarvardX_CB22x_UnitedStates_Bachelor's}         0.002589006 0.14356517 10.385270  869
[15] {MITx_8.02x_India_Bachelor's}                   => {MITx_6.002x_India_Bachelor's}                   0.002496648 0.38564197 16.958041  838
[16] {MITx_6.002x_India_Bachelor's}                  => {MITx_8.02x_India_Bachelor's}                    0.002496648 0.10978645 16.958041  838
[17] {HarvardX_CS50x_India_Bachelor's}               => {MITx_6.00x_India_Bachelor's}                    0.002016982 0.26918489 11.564304  677
[18] {MITx_6.00x_India_Bachelor's}                   => {HarvardX_CS50x_India_Bachelor's}                0.002016982 0.08665045 11.564304  677
[19] {HarvardX_CS50x_UnitedStates_Secondary}         => {MITx_6.00x_UnitedStates_Secondary}              0.002002086 0.12775665  4.850302  672
[20] {MITx_6.00x_UnitedStates_Secondary}             => {HarvardX_CS50x_UnitedStates_Secondary}          0.002002086 0.07600950  4.850302  672
[21] {MITx_6.002x_UnitedStates_Secondary}            => {MITx_6.00x_UnitedStates_Secondary}              0.001939520 0.28194023 10.703907  651
[22] {MITx_6.00x_UnitedStates_Secondary}             => {MITx_6.002x_UnitedStates_Secondary}             0.001939520 0.07363420 10.703907  651
[23] {HarvardX_PH278x_UnitedStates_Bachelor's}       => {HarvardX_ER22x_UnitedStates_Bachelor's}         0.001707135 0.17119809  9.493249  573
[24] {HarvardX_ER22x_UnitedStates_Bachelor's}        => {HarvardX_PH278x_UnitedStates_Bachelor's}        0.001707135 0.09466380  9.493249  573
[25] {MITx_3.091x_UnitedStates_Bachelor's}           => {MITx_6.00x_UnitedStates_Bachelor's}             0.001558171 0.21513780  5.088149  523
[26] {HarvardX_CB22x_UnitedStates_Secondary}         => {HarvardX_ER22x_UnitedStates_Secondary}          0.001540295 0.19240789 19.107014  517
[27] {HarvardX_ER22x_UnitedStates_Secondary}         => {HarvardX_CB22x_UnitedStates_Secondary}          0.001540295 0.15295858 19.107014  517
[28] {MITx_3.091x_UnitedStates_Secondary}            => {MITx_6.00x_UnitedStates_Secondary}              0.001516461 0.21024370  7.981936  509
[29] {MITx_6.00x_UnitedStates_Secondary}             => {MITx_3.091x_UnitedStates_Secondary}             0.001516461 0.05757267  7.981936  509
[30] {HarvardX_CB22x_UnitedStates_Master's}          => {HarvardX_ER22x_UnitedStates_Master's}           0.001415165 0.15785975 14.592571  475
```

원래 support 는 지지도로서 조건절이 발생할 확률을 의미한다. 하지만 분석에 사용한 "arules" 패키지에서는 조건절과 결과절이 동시에 발생할 확률을 나타낸다. Support 가 가장 높은 값은 0.003643676 이며 이것에 해당하는 연관 규칙은 두 가지임을 확인할 수 있다. 이중 첫번째 결과를 바탕으로 분석을 시행하겠다.

	조건절	결과절
Institute	HarvardX	MITx
Course	CS50x	6.00x
Region	United States	United States
Degree	Bachelor's	Bachelor's
Support	0.003643676	
Confidence	0.16504723	
Lift	3.903474	
Count	1223	

다음은 Confidence 를 기준으로 내림차 순으로 정렬한 결과 중 일부이다.

```
> inspect(sort(rule, 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
[4]  {MITx_6.002x_UnitedStates_Secondary}     => {MITx_6.00x_UnitedStates_Secondary}   0.001939520 0.28194023 10.703907   651
[5]  {HarvardX_CS50x_India_Bachelor's}        => {MITx_6.00x_India_Bachelor's}         0.002016982 0.26918489 11.564304   677
[6]  {MITx_6.002x_UnitedStates_Bachelor's}    => {MITx_6.00x_UnitedStates_Bachelor's}  0.002818412 0.25484914  6.027347   946
[7]  {MITx_8.02x_UnitedStates_Bachelor's}     => {MITx_6.002x_UnitedStates_Bachelor's} 0.001391330 0.21620370 19.549777   467
[8]  {MITx_3.091x_UnitedStates_Bachelor's}    => {MITx_6.00x_UnitedStates_Bachelor's}  0.001558171 0.21513780  5.088149   523
[9]  {MITx_3.091x_UnitedStates_Secondary}     => {MITx_6.00x_UnitedStates_Secondary}   0.001516461 0.21024370  7.981936   509
[10] {MITx_8.02x_UnitedStates_Bachelor's}     => {MITx_6.00x_UnitedStates_Bachelor's}  0.001313869 0.20416667  4.828674   441
[11] {HarvardX_CB22x_UnitedStates_Secondary}  => {HarvardX_ER22x_UnitedStates_Secondary} 0.001540295 0.19240789 19.107014   517
[12] {HarvardX_CB22x_UnitedStates_Bachelor's} => {HarvardX_ER22x_UnitedStates_Bachelor's} 0.002589006 0.18728448 10.385270   869
[13] {MITx_8.02x_India_Secondary}             => {MITx_6.00x_India_Secondary}          0.001313869 0.18208092  8.911558   441
[14] {MITx_6.002x_India_Secondary}            => {MITx_6.00x_India_Secondary}          0.003625801 0.17761238  8.692854  1217
[15] {MITx_6.00x_India_Secondary}             => {MITx_6.002x_India_Secondary}         0.003625801 0.17745698  8.692854  1217
[16] {HarvardX_PH278x_UnitedStates_Bachelor's} => {HarvardX_ER22x_UnitedStates_Bachelor's} 0.001707135 0.17119809  9.493249   573
[17] {HarvardX_CS50x_UnitedStates_Master's}   => {MITx_6.00x_UnitedStates_Master's}     0.001218531 0.16985050 11.429495   409
[18] {HarvardX_CS50x_UnitedStates_Bachelor's} => {MITx_6.00x_UnitedStates_Bachelor's}  0.003643676 0.16504723  3.903474  1223
[19] {HarvardX_CB22x_UnitedStates_Master's}   => {HarvardX_ER22x_UnitedStates_Master's} 0.001415165 0.15785975 14.592571   475
[20] {HarvardX_ER22x_UnitedStates_Secondary}  => {HarvardX_CB22x_UnitedStates_Secondary} 0.001540295 0.15295858 19.107014   517
[21] {HarvardX_ER22x_UnitedStates_Bachelor's} => {HarvardX_CB22x_UnitedStates_Bachelor's} 0.002589006 0.14356517 10.385270   869
[22] {HarvardX_CS50x_India_Secondary}         => {MITx_6.002x_India_Secondary}         0.001290034 0.14141084  6.927109   433
[23] {MITx_3.091x_UnitedStates_Bachelor's}    => {MITx_6.002x_UnitedStates_Bachelor's} 0.001021898 0.14109420 12.758154   343
[24] {MITx_6.002x_India_Secondary}            => {MITx_8.02x_India_Secondary}          0.002800536 0.13718622 19.011790   940
[25] {MITx_6.002x_India_Bachelor's}           => {MITx_6.00x_India_Bachelor's}         0.003092507 0.13598847  5.842126  1038
[26] {MITx_6.00x_India_Bachelor's}            => {MITx_6.002x_India_Bachelor's}        0.003092507 0.13285550  5.842126  1038
[27] {MITx_6.00x_India_Secondary}             => {HarvardX_CS50x_India_Secondary}      0.002681365 0.13123360 14.385551   900
[28] {HarvardX_ER22x_UnitedStates_Master's}   => {HarvardX_CB22x_UnitedStates_Master's} 0.001415165 0.13081796 14.592571   475
[29] {HarvardX_CS50x_UnitedStates_Secondary}  => {MITx_6.00x_UnitedStates_Secondary}   0.002002086 0.12775665  4.850302   672
[30] {MITx_6.002x_UnitedStates_Bachelor's}    => {MITx_8.02x_UnitedStates_Bachelor's}  0.001391330 0.12580819 19.549777   467
```

Confidence 는 신뢰도로서 조건절이 발생했을 때 조건절과 결과절이 동시에 발생할 확률을 의미한다. Confidence 가 가장 높은 값은 약 0.38811 인 경우이며 총 940 회 발생한 것을 확인할 수 있다.

	조건절	결과절
Institute	MITx	MITx
Course	8.02x	6.002x
Region	India	India
Degree	Secondary	Secondary
Support	0. 002800536	
Confidence	0. 38810900	
Lift	19.011790	
Count	940	

다음은 Lift 를 기준으로 내림차 순으로 정렬한 결과 중 일부이다.

```
> inspect(sort(rule, 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
[5]  {MITx_6.002x_India_Secondary}             => {MITx_8.02x_India_Secondary}              0.002800536 0.13718622 19.011790   940
[6]  {MITx_8.02x_India_Secondary}              => {MITx_6.002x_India_Secondary}             0.002800536 0.38810900 19.011790   940
[7]  {MITx_6.002x_India_Bachelor's}            => {MITx_8.02x_India_Bachelor's}             0.002496648 0.10978645 16.958041   838
[8]  {MITx_8.02x_India_Bachelor's}             => {MITx_6.002x_India_Bachelor's}            0.002496648 0.38564197 16.958041   838
[9]  {HarvardX_CB22x_UnitedStates_Master's}    => {HarvardX_ER22x_UnitedStates_Master's}    0.001415165 0.15785975 14.592571   475
[10] {HarvardX_ER22x_UnitedStates_Master's}    => {HarvardX_CB22x_UnitedStates_Master's}    0.001415165 0.13081796 14.592571   475
[11] {HarvardX_CS50x_India_Secondary}          => {MITx_6.00x_India_Secondary}              0.002681365 0.29392554 14.385551   900
[12] {MITx_6.00x_India_Secondary}              => {HarvardX_CS50x_India_Secondary}          0.002681365 0.13123360 14.385551   900
[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] {HarvardX_CS50x_India_Bachelor's}         => {MITx_6.00x_India_Bachelor's}             0.002016982 0.26918489 11.564304   677
[16] {MITx_6.00x_India_Bachelor's}             => {HarvardX_CS50x_India_Bachelor's}         0.002016982 0.08665045 11.564304   677
[17] {HarvardX_CS50x_UnitedStates_Master's}    => {MITx_6.00x_UnitedStates_Master's}        0.001218531 0.16985050 11.429495   409
[18] {MITx_6.00x_UnitedStates_Master's}        => {HarvardX_CS50x_UnitedStates_Master's}    0.001218531 0.08199679 11.429495   409
[19] {MITx_6.002x_UnitedStates_Secondary}      => {MITx_6.00x_UnitedStates_Secondary}       0.001939520 0.28194023 10.703907   651
[20] {MITx_6.00x_UnitedStates_Secondary}       => {MITx_6.002x_UnitedStates_Secondary}      0.001939520 0.07363420 10.703907   651
[21] {HarvardX_CB22x_UnitedStates_Bachelor's}  => {HarvardX_ER22x_UnitedStates_Bachelor's}  0.002589006 0.18728448 10.385270   869
[22] {HarvardX_ER22x_UnitedStates_Bachelor's}  => {HarvardX_CB22x_UnitedStates_Bachelor's}  0.002589006 0.14356517 10.385270   869
[23] {HarvardX_PH278x_UnitedStates_Bachelor's} => {HarvardX_ER22x_UnitedStates_Bachelor's}  0.001707135 0.17119809  9.493249   573
[24] {HarvardX_ER22x_UnitedStates_Bachelor's}  => {HarvardX_PH278x_UnitedStates_Bachelor's} 0.001707135 0.09466380  9.493249   573
[25] {MITx_8.02x_India_Secondary}              => {MITx_6.00x_India_Secondary}              0.001313869 0.18208092  8.911558   441
[26] {MITx_6.00x_India_Secondary}              => {MITx_8.02x_India_Secondary}              0.001313869 0.06430446  8.911558   441
[27] {MITx_6.002x_India_Secondary}             => {MITx_6.00x_India_Secondary}              0.003625801 0.17761238  8.692854  1217
[28] {MITx_6.00x_India_Secondary}              => {MITx_6.002x_India_Secondary}             0.003625801 0.17745698  8.692854  1217
[29] {HarvardX_PH278x_UnitedStates_Bachelor's} => {HarvardX_CB22x_UnitedStates_Bachelor's}  0.001120215 0.11233941  8.126449   376
[30] {HarvardX_CB22x_UnitedStates_Bachelor's}  => {HarvardX_PH278x_UnitedStates_Bachelor's} 0.001120215 0.08103448  8.126449   376
```

Lift 는 향상도를 의미하며 생성된 연관 규칙의 유용성을 평가하는 지표이다. Lift 가 가장 높은 경우는 약 19.55 이며 총 467 회 발생한 것을 확인할 수 있다.

	조건절	결과절
Institute	MITx	MITx
Course	8.02x	6.002x
Region	United States	United States
Degree	Bachelor's	Bachelor's
Support	0. 001391330	
Confidence	0. 21620370	
Lift	19.549777	
Count	467	

다음은 효용성 지표를 Support X Confidence X Lift 로 정의할 때 효영성이 가장 높은 규칙 1~3 위를 알아보는 코드이다.

```
mat = as.data.frame(inspect(rule))
measure = mat$support * mat$confidence
measure = measure * mat$lift
mat_measure = data.frame(mat, measure)
head(mat_measure[rev(order(mat_measure$measure)),])
plot(rule, method = "graph")
```

```
> head(mat_measure[rev(order(mat_measure$measure)),])
                                lhs Var.2                                rhs     support confidence     lift count     measure
[23]        {MITx_8.02x_India_Secondary}    => {MITx_6.002x_India_Secondary} 0.002800536  0.3881090 19.01179   940 0.020664168
[5]         {MITx_8.02x_India_Bachelor's}   => {MITx_6.002x_India_Bachelor's} 0.002496648  0.3856420 16.95804   838 0.016327412
[25]     {HarvardX_CS50x_India_Secondary}   => {MITx_6.00x_India_Secondary} 0.002681365  0.2939255 14.38555   900 0.011337562
```

다음은 코드의 결과 중 가장 우측에 위치한 measure 값이 support 와 confidence, lift 를 고려한 새로운 효용성 지표이기 상위 3 개의 관한 연관규칙은 아래 표에 정리하였다.

	조건절	결과절	조건절	결과절	조건절	결과절
Institute	MITx	MITx	MITx	MITx	Harvardx	MITx
Course	8.02x	6.002x	8.02x	6.002x	cs50x	6.00x
Region	India	India	India	India	India	India
Degree	Secondary	Secondary	Bachelor's	Bachelor's	Secondary	Secondary
Support	0.002800536		0.002496648		0.002681365	
Confidence	0.3881090		0.3856420		0.2939255	
Lift	19.01179		16.95804		14.38555	
Count	940		838		900	
Measure	0.020664168		0.016327412		0.011337562	

첫번째로 measure가 가장 높은 규칙은 {MITx_8.02x_India_Secondary} -〉 {MITx_6.002x_India_Secondary} 이다. 조건절과 결과절 사이에는 모두 MIT에서 연 강의를 인도의 Secondary 과정 학생들이 수강한다는 공통점이 있다. 위의 조건 아래에서 8.02x를 수강한 학생이 8.02x와 6.002x를 동시에 수강할 확률은 약 0.39이다. Lift가 1보다 크므로 두 과목은 연관성이 있는 것으로 판단할 수 있다.

두번째로 measure가 높은 규칙은 {MITx_8.02x_India_Bachelor's} -〉 {MITx_6.002x_India_Bachelor's} 이다. 첫번째와 마찬가지로 접속 국가는 인도이며, MIT의 강좌임이 동일하다. 다만 Bachelor를 대상으로 한다는 점이 첫번째 규칙과 차이가 있었다. 석사 학위를 지닌 사람이 인도에서 MIT 의 8.02x를 수강했을 때 8.02x와 6.002x를 동시에 수강할 확률은 약 0.39 이다. 또한, Lift의 경우 1보다 크므로 두 사건 사이에 양의 상관관계가 존재함을 알 수 있다.

마지막으로 measure가 높은 연관 규칙은 {HarvardX_CS50x_India_Secondary} -〉 {(MITx_6.00x_India_Secondary)이다. 위의 두 가지 연관규칙과 마찬가지로 접속 국가는 인도이며 첫번째 규칙 과 같이 수강한 학생들은 Secondary임을 알 수 있다. 모든 measure를 동시에 고려하였을 때 Harvard의 CS50 강의를 수강하면 MIT의 6.00 강의를 수강하는 규칙은 약 900회 정도 발생하였음을 알 수 있다. 또한, Lift가 1보다 크므로 두 사건은 양의 상관관계를 갖는 다는 것을 알 수 있다.

Inspect() 함수를 이용하여 전체 연관규칙을 살펴보면 조건절과 결과절이 뒤바뀐 이후에도 연관규칙을 유지하는 경우는 11,12행의 규칙과 21,22행, 25,26행의 경우였다. 이들의 지표를 확인하기 위해 코드를 짜보았다.

```
#조건절과 결과절 위치 변경
mat_measure[(mat_measure$lhs == "{HarvardX_CB22x_UnitedStates_Master's}")&(mat_measure$rhs ==
"{HarvardX_ER22x_UnitedStates_Master's}"),]
mat_measure[(mat_measure$lhs == "{HarvardX_ER22x_UnitedStates_Master's}")&(mat_measure$rhs ==
"{HarvardX_CB22x_UnitedStates_Master's}"),]
mat_measure[(mat_measure$lhs == "{MITx_8.02x_India_Secondary}")&(mat_measure$rhs ==
"{MITx_6.00x_India_Secondary}"),]
mat_measure[(mat_measure$lhs == "{MITx_6.00x_India_Secondary}")&(mat_measure$rhs ==
"{MITx_8.02x_India_Secondary}"),]
mat_measure[(mat_measure$lhs == "{HarvardX_CS50x_India_Secondary}")&(mat_measure$rhs ==
"{MITx_6.00x_India_Secondary}"),]
mat_measure[(mat_measure$lhs == "{MITx_6.00x_India_Secondary}")&(mat_measure$rhs ==
"{HarvardX_CS50x_India_Secondary}"),]
```

첫번째 11,12행의 규칙이다. 이들은 Support와 Lift는 동일하였지만 Confidence에서 차이를 보였다.

조건절	결과절	Support	Confidence	Lift
HarvardX_CB22x_ UnitedStates_Master's	HarvardX_ER22x_ UnitedStates_Master's	0.0014151	0.157859	14.59257
HarvardX_ER22x_ UnitedStates_Master's	HarvardX_CB22x_ UnitedStates_Master's	0.0014151	0.130818	14.59257

두번째 21,22행의 규칙이다. 이들 또한 Support와 Lift는 동일하였지만 Confidence에서 차이를 보였다.

조건절	결과절	Support	Confidence	Lift
MITx_8.02x_ India_Secondary	MITx_6.00x_ India_Secondary	0.0013138	0.1820809	8.911558
MITx_6.00x_ India_Secondary	MITx_8.02x_ India_Secondary	0.0013138	0.0643044	8.911558

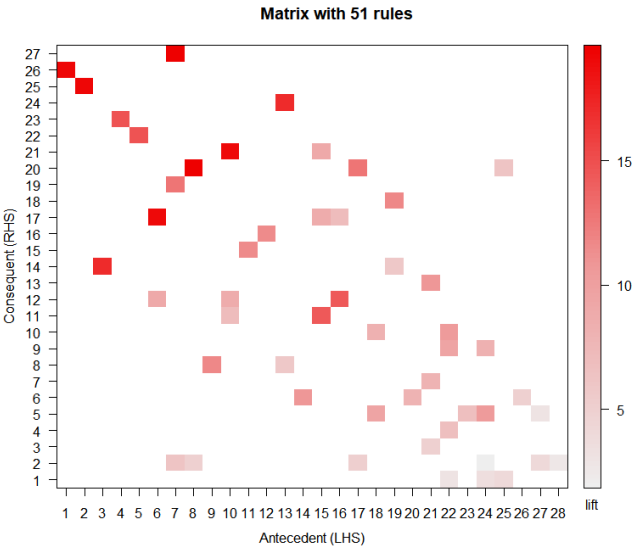
마지막 25,26행의 규칙이다. 이들 또한 Support와 Lift는 동일하였지만 Confidence에서 차이를 보였다.

조건절	결과절	Support	Confidence	Lift
MITx_8.02x_ India_Secondary	MITx_6.00x_ India_Secondary	0.0026813	0.2939255	14.38555
MITx_6.00x_ India_Secondary	MITx_8.02x_ India_Secondary	0.0026813	0.1312336	14.38555

위의 세 경우에서 모두 Support 와 Lift는 조건절과 결과절의 위치가 달라져도 동일한 값을 나타내었고, Confidence 에서만 차이가 발생하였다. 이 이유는 Confidence는 조건부 확률로 조건절과 결과절의 위치에 따라 분모에 해당하는 조건절의 값이 달라 값에 차이가 발생하게 된다.

[Extra Question]

```
#matrix
plot(rule, method = "matrix", measure = "lift")
```



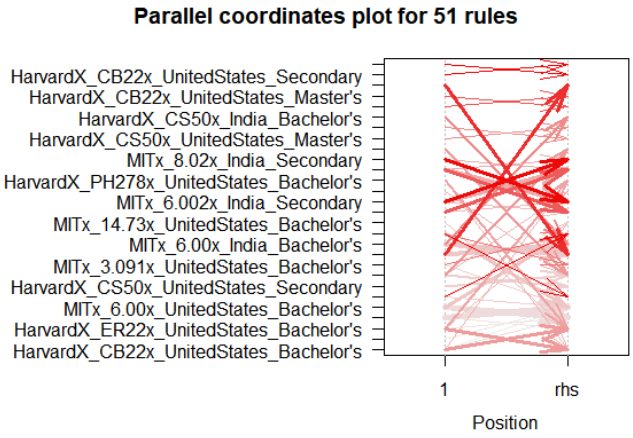
Matrix Plot 을 보면 우측의 그림과 같다.

해당 그래프의 X축은 조건 절이며, Y 축은 결과절이다.

51가지 규칙에 대하여 lift값을 색깔 변화로 나타내었다.

```
#paracoord
plot(rule, method="paracoord")
```

Parallel Coordinates Plot 은 우측의 그림과 같다. Item 수에 따른 연관관계를 병렬적으로 확인할 수 있으며 가로축의 숫자 1 이 의미하는 것은 조건절에 해당하는 item 수이다. 모든 조건절의 item 수가 1 개이므로 그래프가 비교적 단순하게 나타난다.




```
#grouped
subrule = head(rule, n = 7, by =
"lift");
plot(subrule,method="grouped")
```

Lift 를 기준으로 상위 7 개의 연관 규칙에 대하여 group 화하여 도시한 그래프이다. 각 규칙에 따른 조건절과 결과절에 해당하는 item 을 직접 확인할 수 있다는 것이 특징이다. 아래에 나타나 있는 원의 크기는 support 를 의미하고 원의 색상은 lift 를 나타낸다.