9주차 강의

연관성 분석 실습

(Association Analysis)

2022.05.03

Institutor: JS LEE

필요패키지

install.packages("arules") => Apriori 알고리즘 존재함

Install.packages("arulesViz") => 연관규칙을 시각화

```
library(arules)
library(arulesViz)
library(dplyr)
library(tidyverse)
```

Data

데이터 : 실제 식료품 매장의 한 달 운영한 구매 데이터

9835건 거래 데이터 (일별 327거래 , 12시간 영업, 시간당 : 약 30건)

```
read.csv("데이터 경로", header = FALSE)
a <- read.csv("groceries.csv". header=FALSE)
View(a)
```

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	citrus fruit	semi-finished bread	margarine	ready soups						
2	tropical fruit	yogurt	coffee							
3	whole milk									
4	pip fruit	yogurt	cream cheese	meat spreads						
5	other vegetables	whole milk	condensed milk	long life bakery product						
6	whole milk	butter	yogurt	rice	abrasive cleaner					
7	rolls/buns									
8	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)					
9	potted plants									
10	whole milk	cereals								
11	tropical fruit	other vegetables	white bread	bottled water	chocolate					
12	citrus fruit	tropical fruit	whole milk	butter	curd	yogurt	flour	bottled water	dishes	
13	beef									
14	frankfurter	rolls/buns	soda							
15	chicken	tropical fruit								
16	butter	sugar	fruit/vegetable juice	newspapers						

```
데이터프레임이름A <- read.csv('파일경로', header=FALSE, skip=n, ...)
```

read.csv()의 주요 Argument	R에서의 표현 예시	의미
header	header=TRUE	TRUE : 첫번째 행을 컬럼의 이름으로 간주(디폴트) FALSE : 컬럼 이름이 없다고 간주, 새로운 컬럼 이름을 임의로 생성
skip	skip=n	투입하는 .csv파일의 데이터프레임 중 1번째 row로부터 n개 row를 건너뜀
comment	comment="#"	#로 시작하는 모든 줄 무시
dec	dec='.'	.CSV 파일이 소숫점(.)을 포함한 숫자 데이터를 가지고 있는 경우 소숫점을 지정
col_names	col_names=c('x1','x2',...)	컬럼 이름을 x1, x2,...로 지음
na	na='.'	결측값을 지정하여 나타냄. 불러오는 CSV파일에서 ''(작은 따옴표) 내의 문자(ex. .)를 결측값으로 간주함

거래 데이터는 일관되지 않음
(모든 행과 열에 데이터 값이 존재 하지 않으며
데이터 값이 존재하는 행과 열이 상이함)

Data 이해

	V1	V2	V3	V4	V5	V6	V7		
1	citrus fruit	semi-finished bread	margarine	ready soups					
2	tropical fruit	yogurt	coffee						
3	whole milk								
4	pip fruit	yogurt	cream cheese	meat spreads					
5	other vegetables	whole milk	condensed milk	long life bakery product					
6	whole milk	butter	yogurt	rice	abrasive cleaner				
7	rolls/buns								
8	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)				
9	potted plants								
10	whole milk	cereals							
11	tropical fruit	other vegetables	white bread	bottled water	chocolate				
12	citrus fruit	tropical fruit	whole milk	butter	curd	yogurt	flour	bottled water	dishes
13	beef								
14	frankfurter	rolls/buns	soda						
15	chicken	tropical fruit							
16	butter	sugar	fruit/vegetable juice	newspapers					

첫번째 고객이 산 제품
(1번 고객은 4가지 상품을 구매)

열두번째 고객이 산 제품
(12번 고객은 9가지 상품을 구매)

Data 이해

- 연관분석에 사용되는 데이터의 형식은 각 상품(서비스, 항목)의 거래 유무를 나타내는 이진형 임.

[온라인 뉴스 사이트에서 엑세스한 미디어 데이터]

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Arts}
5	{Sports, News, Finance}
6	{News, Arts, Entertainment}

[연관성 분석 데이터 형태]

Session ID	News	Finance	Entertain	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

- 거래데이터를 연관분석에 사용되는 형태(희소행렬)로 변경하여 불러오는 함수 : read.transactions
  - arules 패키지에서 제공하며, read.csv()와 유사

희소행렬(sparse matrix): 행렬의 값이 대부분 0인 경우

```
groceries <- read.transactions("groceries.csv", sep = ",")
#sep = “,”  => 쉼표로 분리됨을 의미
```

Data 이해

- groceries 행렬에 대한 기본 정보를 확인하기 위해 summary()이용

```
summary(groceries)
```

```
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146
```

9,835거래건수, 아이템(상품수) 169개
Density: 0.02609
: 행렬에서 0이 아닌 셀의 비율

-행렬에서 9835 * 169 = 1,662,115 이며, 이 중 0이 아닌 경우 0.02609 이므로

1,662,115*0.02609 =43,367 개 => 한달에 43,367개의 아이템이 구매됨 (중복 구매 포함)
평균거래 : 43367/9835 =4.409 (한번 거래시 평균 아이템 수)

```
most frequent items:
      whole milk other vegetables       rolls/buns             soda           yogurt          (Other)
            2513             1903             1809             1715             1372            34055
```

- 가장 자주 발견된 items 은 전유(whole milk): 2513 건, 총 거래의 25.6% (2513/9835 =25.6%)
- 기타 채소류(other vegetables) : 1903건,
- 롤/번(rolls/buns) :1809건 등으로
- 자주 구매하는 아이템으로는 . 전유(Whole milk), 기타채소류(other vegetables), 롤/번(rolls/buns),탄산음료(soda), 요거트(yogurt)순으로 나타남)

Data 이해

```
element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24   26   27   28   29   32
2159 7643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29   14   14    9   11    4    6    1    1    1    1    3    1
```

- 한 아이템만을 포함한 경우는 2159건인 반면, 32개의 아이템을 포함한 경우는 1건임을 알 수 있음
- 한 아이템의 구매건수는 알 수 없음 (구매여부를 1과 0으로 나타내기에)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000
```

- 구매크기 4분위수를 살펴보면, 50%는 3개 이하의 아이템을 구매함
- 25%는 2개 이하를 구매함
- 평균거래는 4.409

Data 이해

- Arules 패키지는 거래데이터 파악을 위한 함수들

1) inspect() : 거래 파악

inspect(groceries,[1:5])

처음 다섯 건의 거래 유형을 볼 수 있음

```
> inspect(groceries[1:5])
    items
[1] {citrus fruit,
     margarine,
     ready soups,
     semi-finished bread}
[2] {coffee,
     tropical fruit,
     yogurt}
[3] {whole milk}
[4] {cream cheese,
     meat spreads,
     pip fruit,
     yogurt}
[5] {condensed milk,
     long life bakery product,
     other vegetables,
     whole milk}
Warning message:
```

2) itemfrequency( ) : 아이템을 포함하는 거래의 비율

itemFrequency(groceries[,1:3])

첫번째 3개의 item 비율을 볼 수 있음

```
> itemFrequency(groceries[,1:3])
abrasive cleaner artif. sweetener   baby cosmetics
    0.0035587189     0.0032536858     0.0006100661
>
```

- 연마용 청소기(abrasive cleaner) : 0.35%, 인공감미료(artificial seetner) :0.32%
  유아용화장품 (baby cosmetic) :0.61%

희소행렬에서 item은 알파벳 순으로 정렬

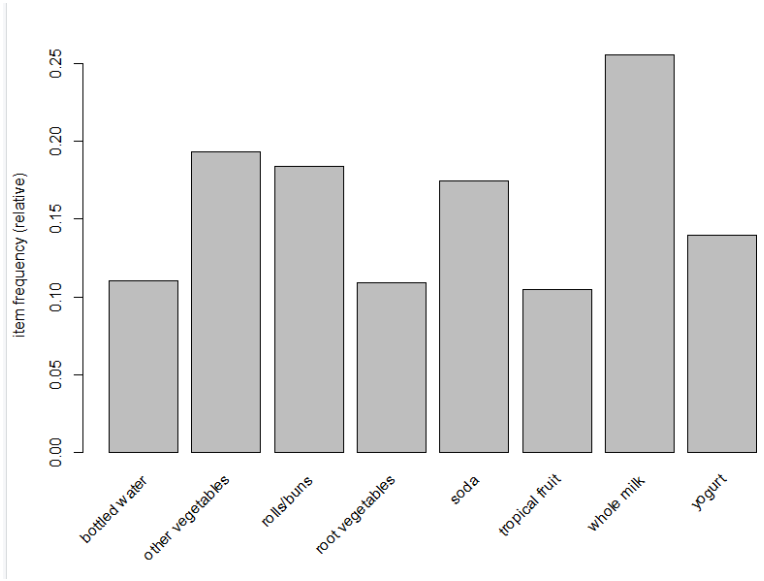
Data 소개

- Arules 패키지는 거래데이터 파악을 위한 함수들

3)itemFrequencyPlot()
: 아이템 지지도의 시각화 :

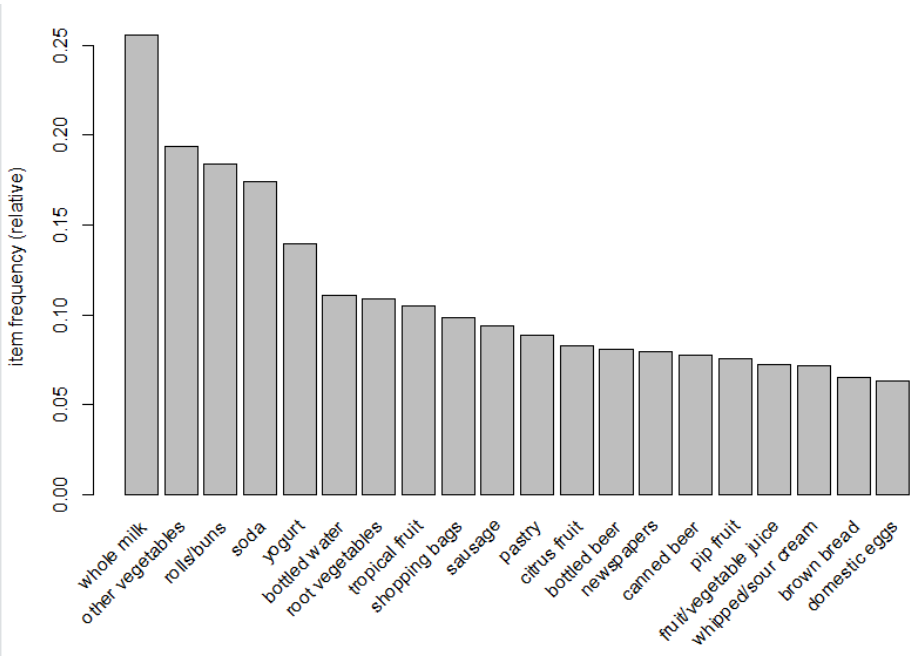
- 최소거래비율을 갖는 아이템들을 표현

itemFrequencyPlot(groceries, support=0.1)
# 최소거래비율이 10%이상인 아이템



- 특정개수의 아이템을 나타냄;

itemFrequencyPlot(groceries, topN=20)
# 상위20개 까지의 아이템



연관규칙 찾기

aprior() 함수를 이용하여 규칙 집합을 생성

- 연관규칙 찾기

myrules <- apriori(data=groceries, parameter =list(support=0.006, confidence=0.25,minlen=2))

- data : 거래데이터를 갖고 있는 희소행렬
- support : 요구되는 최소 지지도
- confidence : 요구되는 최소 신뢰도
- minlen : 요구되는 최소규칙 아이템

- 연관규칙 검토

inspect(myrules)

연관규칙찾기

myrules <- apriori(data=groceries, parameter =list(support=0.01, confidence=0.25,minlen=2))

: groceries 데이터 셋에서, 지지도는 0.01이상이고 신뢰도는 0.25이상이며, 두 개 미만의 아이템을 갖는 경우 제외

* minlen =2 => {whole milk}와 같이 하나의 아이템이 자주 구매되는 규칙 제외

```
Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
       0.25    0.1    1 none FALSE            TRUE       5    0.01      2     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 98

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
sorting and recoding items ... [88 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 done [0.01s].
writing ... [170 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> inspect(myrules)
      lhs                                rhs                support    confidence coverage
[1]   {hard cheese}                   => {whole milk}       0.01006609 0.4107884  0.02450432
[2]   {butter milk}                   => {other vegetables} 0.01037112 0.3709091  0.02796136
[3]   {butter milk}                   => {whole milk}       0.01159126 0.4145455  0.02796136
[4]   {ham}                           => {whole milk}       0.01148958 0.4414062  0.02602949
[5]   {sliced cheese}                 => {whole milk}       0.01077783 0.4398340  0.02450432
```

연관규칙찾기

- 생성된 170개 규칙들의 요약 정보를 보기 위해서는 summary() 함수 이용

```
summary(myrules)
```

```
set of 170 rules

rule length distribution (lhs + rhs):sizes
 2  3
96 74

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.435   3.000   3.000

summary of quality measures:
    support         confidence        coverage         lift            count
 Min.   :0.01007   Min.   :0.2517   Min.   :0.01729   Min.   :0.9932   Min.   : 99.0
 1st Qu.:0.01159   1st Qu.:0.2973   1st Qu.:0.03101   1st Qu.:1.5215   1st Qu.:114.0
 Median :0.01454   Median :0.3587   Median :0.04291   Median :1.7784   Median :143.0
 Mean   :0.01822   Mean   :0.3703   Mean   :0.05187   Mean   :1.8747   Mean   :179.2
 3rd Qu.:0.02097   3rd Qu.:0.4253   3rd Qu.:0.05857   3rd Qu.:2.1453   3rd Qu.:206.2
 Max.   :0.07483   Max.   :0.5862   Max.   :0.25552   Max.   :3.2950   Max.   :736.0

mining info:
      data ntransactions support confidence
 groceries          9835    0.01       0.25
                                                                                call
 apriori(data = groceries, parameter = list(support = 0.01, confidence = 0.25, minlen = 2))
```

두개의 아이템을 갖는 규칙 : 96개

세개의 아이템을 가는 규칙: 74개

연관규칙 찾기

```
inspect(myrules[1:3])
```

```
> inspect(myrules[1:3])
    lhs              rhs                support    confidence coverage   lift     count
[1] {hard cheese} => {whole milk}       0.01006609 0.4107884  0.02450432 1.607682  99
[2] {butter milk} => {other vegetables} 0.01037112 0.3709091  0.02796136 1.916916 102
[3] {butter milk} => {whole milk}       0.01159126 0.4145455  0.02796136 1.622385 114
>
```

1번 rule

hard cheese와 Whole Milk 동시에 구매하는 비중은 전체 거래의 1% 차지하며(support), hard cheese를 구매하는 경우 41%(confidnence)는 Whole mik구매한다 (confidnence)

Whole milk를 구매한 평균 고객에 비해, hard cheese를 산고객이 Whole Milk를 구매할 경우는 1.6배 높다 (Lift)

연관규칙찾기

- 연관규칙의 부분집합 구하기

특정 상품(항목)에 대한 연관규칙을 살펴보기 위해, 연관 규칙의 부분집합 구하는 방법은 subset()함수 이용

* berry 관련 연관규칙만 살펴보기

```
berryrule <- subset(myrules, items %in% "berries")
inspect(berryrule)
```

```
> berryrule <- subset(myrules, items %in% "berries")
> inspect(berryrule)
    lhs          rhs                    support     confidence coverage  lift     count
[1] {berries} => {whipped/sour cream} 0.009049314 0.2721713  0.0332486 3.796886  89
[2] {berries} => {yogurt}             0.010574479 0.3180428  0.0332486 2.279848 104
[3] {berries} => {other vegetables}   0.010269446 0.3088685  0.0332486 1.596280 101
[4] {berries} => {whole milk}         0.011794611 0.3547401  0.0332486 1.388328 116
>
```

연관규칙찾기

- 연관규칙 저장하기

  - R 데이터 프레임으로 규칙 저장하기

```
groceryrule <- as(myrules, "data.frame")
```

	rules	support	confidence	coverage	lift	count
1	{potted plants} => {whole milk}	0.006914082	0.4000000	0.01728521	1.565460	68
2	{pasta} => {whole milk}	0.006100661	0.4054054	0.01504830	1.586614	60
3	{herbs} => {root vegetables}	0.007015760	0.4312500	0.01626843	3.956477	69
4	{herbs} => {other vegetables}	0.007727504	0.4750000	0.01626843	2.454874	76
5	{herbs} => {whole milk}	0.007727504	0.4750000	0.01626843	1.858983	76
6	{processed cheese} => {whole milk}	0.007015760	0.4233129	0.01657346	1.656698	69
7	{semi-finished bread} => {whole milk}	0.007117438	0.4022989	0.01769192	1.574457	70
8	{beverages} => {whole milk}	0.006812405	0.2617188	0.02602949	1.024275	67
9	{detergent} => {other vegetables}	0.006405694	0.3333333	0.01921708	1.722719	63
10	{detergent} => {whole milk}	0.008947636	0.4656085	0.01921708	1.822228	88

  - write.csv()함수를 이용하여 csv파일로 저장

```
write.csv(groceryrule, file="groceryrules.csv")
```

연관규칙찾기

- 그래프로 표현하기

myrules1 <- apriori(data=groceries, parameter =list(support=0.01, confidence=0.5,minlen=2))

inspect(myrules1)

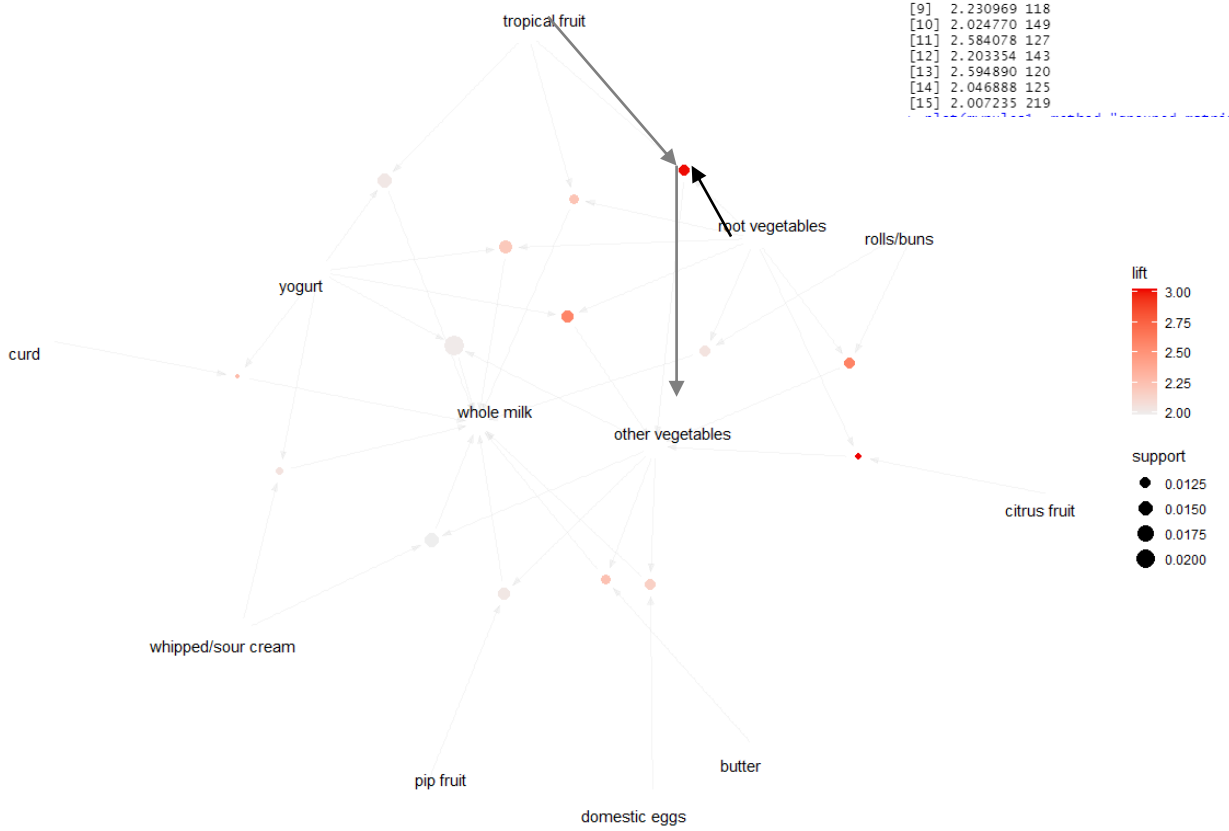
```
     lhs                                    rhs                support    confidence coverage
[1]  {curd, yogurt}                      => {whole milk}       0.01006609 0.5823529  0.01728521
[2]  {butter, other vegetables}          => {whole milk}       0.01148958 0.5736041  0.02003050
[3]  {domestic eggs, other vegetables}   => {whole milk}       0.01230300 0.5525114  0.02226741
[4]  {whipped/sour cream, yogurt}        => {whole milk}       0.01087951 0.5245098  0.02074225
[5]  {other vegetables, whipped/sour cream} => {whole milk}    0.01464159 0.5070423  0.02887646
[6]  {other vegetables, pip fruit}       => {whole milk}       0.01352313 0.5175097  0.02613116
[7]  {citrus fruit, root vegetables}     => {other vegetables} 0.01037112 0.5862069  0.01769192
[8]  {root vegetables, tropical fruit}   => {other vegetables} 0.01230300 0.5845411  0.02104728
[9]  {root vegetables, tropical fruit}   => {whole milk}       0.01199797 0.5700483  0.02104728
[10] {tropical fruit, yogurt}            => {whole milk}       0.01514997 0.5173611  0.02928317
[11] {root vegetables, yogurt}           => {other vegetables} 0.01291307 0.5000000  0.02582613
[12] {root vegetables, yogurt}           => {whole milk}       0.01453991 0.5629921  0.02582613
[13] {rolls/buns, root vegetables}       => {other vegetables} 0.01220132 0.5020921  0.02430097
[14] {rolls/buns, root vegetables}       => {whole milk}       0.01270971 0.5230126  0.02430097
[15] {other vegetables, yogurt}          => {whole milk}       0.02226741 0.5128806  0.04341637
     lift     count
[1]  2.279125  99
[2]  2.244885 113
[3]  2.162336 121
[4]  2.052747 107
[5]  1.984385 144
[6]  2.025351 133
[7]  3.029608 102
[8]  3.020999 121
[9]  2.230969 118
[10] 2.024770 149
[11] 2.584078 127
[12] 2.203354 143
[13] 2.594890 120
[14] 2.046888 125
[15] 2.007235 219
```

연관규칙찾기

- 그래프로 표현하기

plot(myrules1, method= "graph")

```
     lhs                                 rhs                support    confidence coverage
[1]  {curd, yogurt}                   => {whole milk}       0.01006609 0.5823529  0.01728521
[2]  {butter, other vegetables}       => {whole milk}       0.01148958 0.5736041  0.02003050
[3]  {domestic eggs, other vegetables}=> {whole milk}       0.01230300 0.5525114  0.02226741
[4]  {whipped/sour cream, yogurt}     => {whole milk}       0.01087951 0.5245098  0.02074225
[5]  {other vegetables, whipped/sour cream} => {whole milk} 0.01464159 0.5070423  0.02887646
[6]  {other vegetables, pip fruit}    => {whole milk}       0.01352313 0.5175097  0.02613116
[7]  {citrus fruit, root vegetables}  => {other vegetables} 0.01037112 0.5862069  0.01769192
[8]  {root vegetables, tropical fruit}=> {other vegetables} 0.01230300 0.5845411  0.02104728
[9]  {root vegetables, tropical fruit}=> {whole milk}       0.01199797 0.5700483  0.02104728
[10] {tropical fruit, yogurt}         => {whole milk}       0.01514997 0.5173611  0.02928317
[11] {root vegetables, yogurt}        => {other vegetables} 0.01291307 0.5000000  0.02582613
[12] {root vegetables, yogurt}        => {whole milk}       0.01453991 0.5629921  0.02582613
[13] {rolls/buns, root vegetables}    => {other vegetables} 0.01220132 0.5020921  0.02430097
[14] {rolls/buns, root vegetables}    => {whole milk}       0.01270971 0.5230126  0.02430097
[15] {other vegetables, yogurt}       => {whole milk}       0.02226741 0.5128806  0.04341637
     lift     count
[1]  2.279125  99
[2]  2.244885 113
[3]  2.162336 121
[4]  2.052747 107
[5]  1.984385 144
[6]  2.025351 133
[7]  3.029608 102
[8]  3.020999 121
[9]  2.230969 118
[10] 2.024770 149
[11] 2.584078 127
[12] 2.203354 143
[13] 2.594890 120
[14] 2.046888 125
[15] 2.007235 219
```

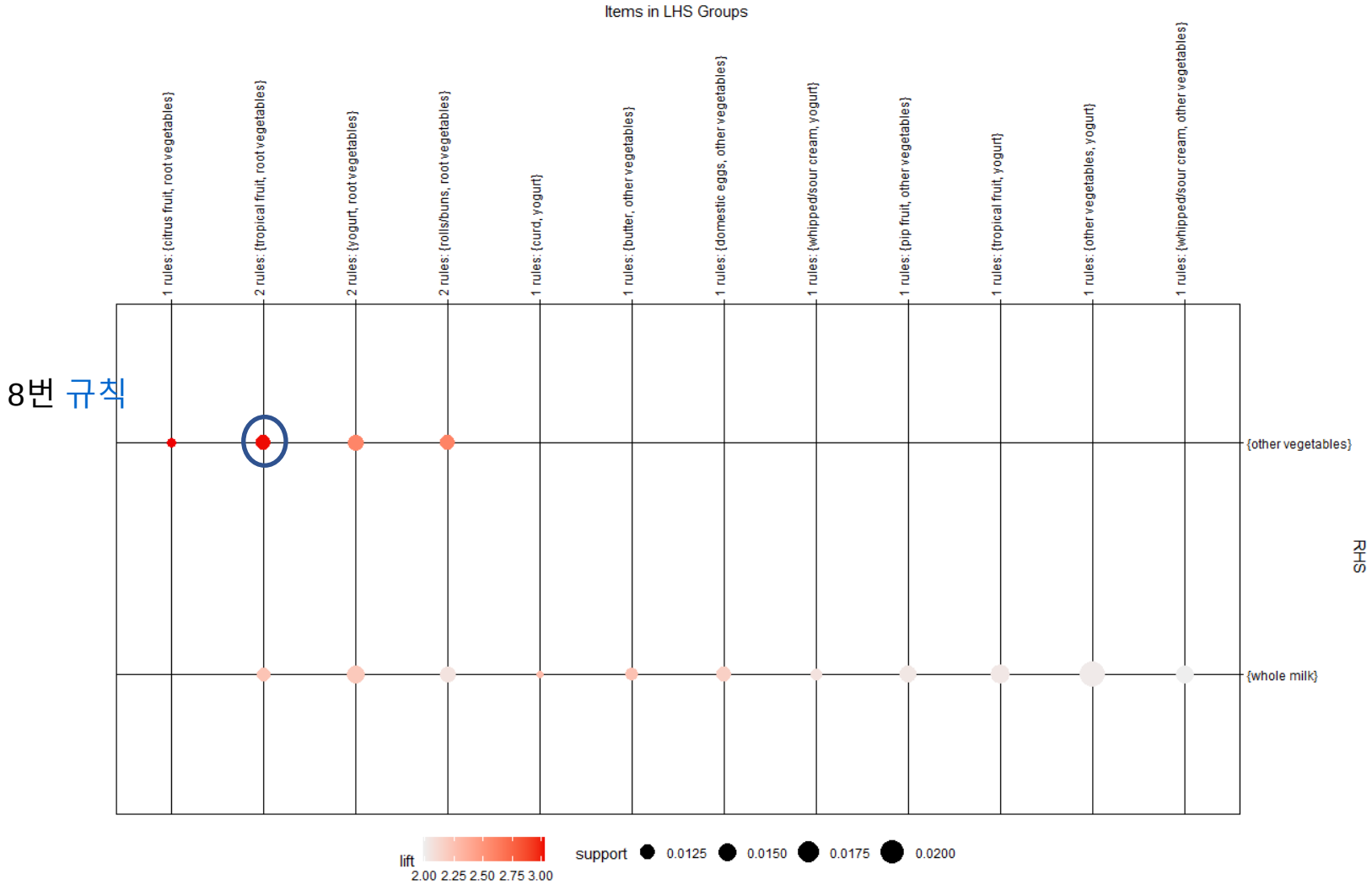


8번 규칙

연관규칙찾기

- 그래프로 표현하기

```
plot(myrules1, method="grouped matrix")
```



8번 규칙

연관규칙찾기

- 지지도의 기준 선택이 중요함

지지도를 0.006, 으로 했을 경우와, 0.01로 했을 때 차이 존재

```
myrules <- apriori(data=groceries, parameter =list(support=0.01, confidence=0.25,minlen=2))
myrules0 <- apriori(data=groceries, parameter =list(support=0.006, confidence=0.25,minlen=2))
```

```
> inspect(sort(myrules, by="lift")[1:10])
     lhs                                rhs                support    confidence coverage   lift     count
[1]  {citrus fruit, other vegetables}   => {root vegetables}  0.01037112 0.3591549  0.02887646 3.295045 102
[2]  {other vegetables, tropical fruit} => {root vegetables}  0.01230300 0.3427762  0.03589222 3.144780 121
[3]  {beef}                             => {root vegetables}  0.01738688 0.3313953  0.05246568 3.040367 171
[4]  {citrus fruit, root vegetables}    => {other vegetables} 0.01037112 0.5862069  0.01769192 3.029608 102
[5]  {root vegetables, tropical fruit}  => {other vegetables} 0.01230300 0.5845411  0.02104728 3.020999 121
[6]  {other vegetables, whole milk}     => {root vegetables}  0.02318251 0.3097826  0.07483477 2.842082 228
[7]  {curd, whole milk}                 => {yogurt}           0.01006609 0.3852140  0.02613116 2.761356  99
[8]  {other vegetables, yogurt}         => {root vegetables}  0.01291307 0.2974239  0.04341637 2.728698 127
[9]  {other vegetables, yogurt}         => {tropical fruit}   0.01230300 0.2833724  0.04341637 2.700550 121
[10] {other vegetables, rolls/buns}     => {root vegetables}  0.01220132 0.2863962  0.04260295 2.627525 120
> inspect(sort(myrules0, by="lift")[1:10])
     lhs                                          rhs                  support     confidence coverage   lift     count
[1]  {herbs}                                   => {root vegetables}    0.007015760 0.4312500  0.01626843 3.956477  69
[2]  {berries}                                 => {whipped/sour cream} 0.009049314 0.2721713  0.03324860 3.796886  89
[3]  {other vegetables, tropical fruit, whole milk} => {root vegetables}    0.007015760 0.4107143  0.01708185 3.768074  69
[4]  {beef, other vegetables}                  => {root vegetables}    0.007930859 0.4020619  0.01972547 3.688692  78
[5]  {other vegetables, tropical fruit}        => {pip fruit}          0.009456024 0.2634561  0.03589222 3.482649  93
[6]  {beef, whole milk}                        => {root vegetables}    0.008032537 0.3779904  0.02125064 3.467851  79
[7]  {other vegetables, pip fruit}             => {tropical fruit}     0.009456024 0.3618677  0.02613116 3.448613  93
[8]  {pip fruit, yogurt}                       => {tropical fruit}     0.006405694 0.3559322  0.01799695 3.392048  63
[9]  {citrus fruit, other vegetables}          => {root vegetables}    0.010371124 0.3591549  0.02887646 3.295045 102
[10] {other vegetables, whole milk, yogurt}    => {tropical fruit}     0.007625826 0.3424658  0.02226741 3.263712  75
```

지지도의 결정 기준은 고민이 필요함

예> 한 아이템이 하루에 2번 이상 구매되면 (한달에 60회) 흥미로운 패턴이다고 주장

grocery 데이터 셋에서는 60/9835 = 0.0061

지지도를 0.006 이상으로 한다.