

예) 기저귀-맥주(미국 월마트 분석)

1. 고객들은 어떤 상품들을 동시에 구매하는가?
2. 라면을 구매한 고객은 주로 다른 어떤 상품을 구매하는가?

위와 같은 질문에 대한 분석을 토대로 고객들에게 SMS를 보낸다든가, 판촉용 전화를 한다든가 묶음 판매를 기획함.

이와 같은 질문에 대한 답은 연관규칙을 이용하여 구할 수 있습니다. 연관규칙은 상업 데이터베이스에서 가장 흔히 쓰이는 도구로, 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건을 의미.

support 지지도는 품목 A와 B를 동시에 구매할 확률인 $P(A \cap B)$ 를 나타냅니다

confidence 신뢰도는 품목 A가 구매하고나서, 품목 B가 구매될 확률

lift 향상도는 A를 구매한 사람이 B를 구매할 확률과 A의 구매와 상관없이 B를 구매할 확률의 비율

lift > 1 이면 관련도가 높고 **lift < 1** 이면 A구매자가 B를 구매하지 않을 확률이 높음

참고사이트: <https://ratsgo.github.io/machine%20learning/2017/04/08/apriori/>

*연관분석, 장바구니 분석

***지지도(Support)**: 전체 집합군에서 [조건] 자료가 포함된 집합수, 비율,
[조건1]자료수 / 전체자료수

***신뢰도(Confidence)**: [조건1]가 있을때 [조건2]도 같이 있는 확률
[조건1]->[조건2] 라고 하면
[조건1],[조건2] 가 같이 나온 자료수/[조건1] 자료수

즉: [조건1],[조건2] 지지도 / [조건1] 지지도

***향상도(Lift:Improvement)**:
[조건1][조건2]가 같이 나온 자료수/[조건1]자료수/전체자료수

https://m.blog.naver.com/PostView.nhn?blogId=leedk1110&logNo=220785911828&proxyReferer=http%3A%2F%2Fwww.google.co.kr%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3D%26esrc%3Ds%26source%3Dweb%26cd%3D3%26ved%3D2ahUKEwji-bmbtMTcAhVD_GEKHSv5DyIQFjACegQIARAB%26url%3Dhttp%253A%252F%252Fm.blog.naver.com%252Fleedk1110%252F220785911828%26usg%3DAOvVaw2dQ91WI-N0WuNhdsE3wRbj

연관분석 - 화장품전문점 패키지 구성방법?

분류	내용
예제 데이터	<ul style="list-style-type: none"> ■ B화장품전문점에서 판매된 트랜잭션 데이터
변수명	<ul style="list-style-type: none"> ■ 단일변수 <ul style="list-style-type: none"> - Nail Polish(매니큐어), Brushes(브러시), - Concealer(컨실러: 피부 결점을 감추어 주는 화장품) - Bronzer(피부를 햇볕에 그을린 것처럼 보이게 하는 화장품) - Lip liner(입술 라이너), Mascara(마스카라: 속눈썹용 화장품) - Eye shadow(아이섀도: 눈꺼풀에 바르는 화장품) - Foundation(파운데이션: 가루분), Lip Gloss(립글로스: 입술 화장품) - Lipstick(립스틱), Eyeliner(아이 라이너: 눈의 윤곽 그림)
분석문제	<ul style="list-style-type: none"> ■ 전체 트랜잭션 개수와 상품아이템 유형은 몇 개인가? ■ 가장 발생빈도가 높은 상품아이템은 무엇인가? ■ 지지도를 10%로 설정했을 때의 생성되는 규칙의 가지수는? ■ 상품아이템 중에서 가장 발생확률이 높은 아이템과 낮은 아이템은 무엇인가? ■ 가장 발생가능성이 높은 <2개 상품간>의 연관규칙은 무엇인가? ■ 가장 발생가능성이 높은 <2개 상품이상에서> <제3의 상품으로>의 연관규칙은?

판매촉진 - 프로모션 효율화 방안

[우체국 쇼핑부문] 쇼핑물 이용고객을 위한 추천상품 분석

분류	내용
예제 데이터	<ul style="list-style-type: none">우체국 쇼핑에서 판매된 트랜잭션 데이터파일
변수명	<ul style="list-style-type: none">단일변수: 의류(clothes), 냉동식품(frozen), 주류(alcohol,) 야채(veg), 제과(bakery), 육류(meat), 과자(snack), 생활장식(deco)에 대한 거래처리데이터
분석문제	<ul style="list-style-type: none">전체 트랜잭션 개수와 상품아이템 유형은 몇 개인가?가장 발생빈도가 높은 상품아이템은 무엇인가?지지도를 10%로 설정했을 때의 생성되는 규칙의 가지는?상품아이템 중에서 가장 발생확률이 높은 아이템과 낮은 아이템은 무엇인가?가장 발생가능성이 높은 <2개 상품간>의 연관규칙은 무엇인가?가장 발생가능성이 높은 <2개 상품이상에서> <제3의 상품으로>의 연관규칙은?

연관성 분석 - R[연관분석할수 있게 트랜잭션으로 읽기]

```
setwd("c:/data_r")
install.packages("arules")
library(arules)
tr<-read.transactions ("장바구니분석소스.txt",format="basket",sep=",")
tr
class(tr)      # 4행7열로 이루어진 데이터임.
summary(tr)
inspect(tr)    # 트랜잭션 형태의 자료 아이템 확인
tr@itemInfo
tr@data
```

<https://www.rdocumentation.org/packages/arules/versions/1.7-3/topics/read>

tr@data

장바구니분석소스.txt - 메모장

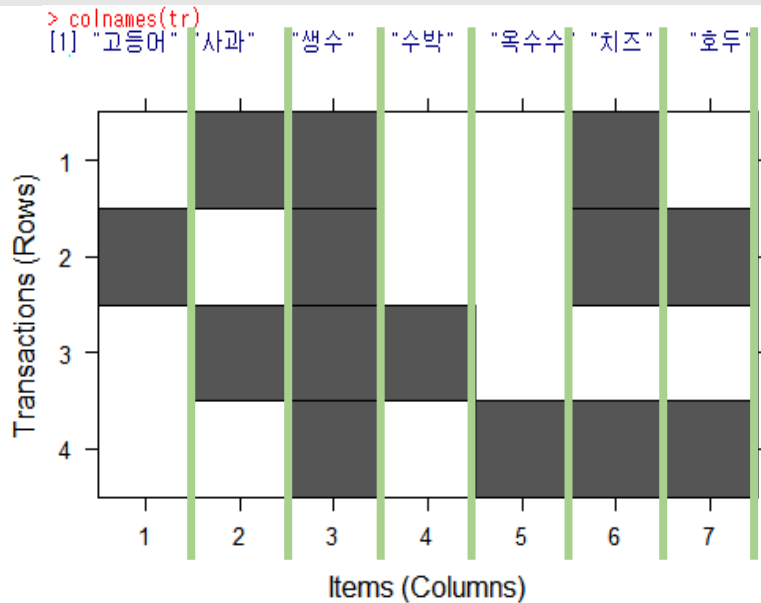
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

사과,치즈,생수
 생수,호두,치즈,고등어
 수박,사과,생수
 생수,호두,치즈,옥수수

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
3	고등어
	수박
	사과
4	생수
	생수
	호두
	치즈
	옥수수

1~4번의 구매자번호별 항목 출현체크(오름차순정렬)					
번호	항목, 구매자번호	1	2	3	4
1	고등어				
2	사과				
3	생수				
4	수박				
5	옥수수				
6	치즈				
7	호두				

image(tr)
colnames(tr)



itemFrequency(tr) #항목별 지지도

```
> itemFrequency(tr)
고등어 사과 생수 수박 옥수수 치즈 호두
0.25 0.50 1.00 0.25 0.25 0.75 0.50
```

항목별 지지도[Support]			
번호	제품명	지지도(자료수/4)	
1	고등어	1	0.25
2	사과	2	0.5
3	생수	4	1
4	수박	1	0.25
5	옥수수	1	0.25
6	치즈	3	0.75
7	호두	2	0.5

연관성 분석 -R(연관분석에서 지지도)

```
itemFrequency(tr[,1:3])  
round(itemFrequency(tr)*100,1)  
order(itemFrequency(tr)) #지지도의 값을 작은값이 있는 번호부터 나오게함.
```

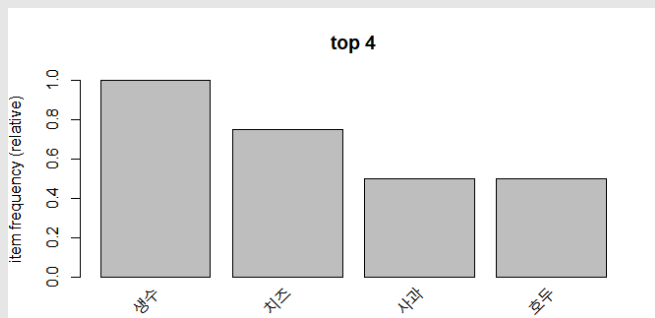
```
> order(itemFrequency(tr))  
[1] 1 4 5 2 7 6 3  
  
> itemFrequency(tr)  
고등어 사과 생수 수박 옥수수 치즈 호두  
0.25 0.50 1.00 0.25 0.25 0.75 0.50  
1 2 3 4 5 6 7
```

```
order(-itemFrequency(tr))
```

```
> order(-itemFrequency(tr))  
[1] 3 6 2 7 1 4 5
```

```
colnames(tr[,3]) #3번째 제목 확인
```

```
itemFrequencyPlot(tr,topN=4,main="top 4") # top 4
```



연관성 분석(지지도, 신뢰도, 향상도)

사과를 구매한 고객이 치즈도 함께구매할 연관성에 대해 분석

지지도= $P(A \cap B)$

신뢰도= $P(A \cap B)/P(A)$

향상도= $\text{신뢰도}(A,B)/\text{지지도}(B)$

▶ 지지도=[사과][치즈]가 같이 나온 자료/전체자료 => 1/4 => 0.25

구매자번호	제품명
1	사과
	치즈
2	생수
	생수
	호두
	치즈
3	고등어
	수박
	사과
4	생수
	호두
	치즈
	옥수수

▶ 신뢰도=[사과][치즈]가 같이 나온 자료/[사과]자료 => 1/2 => 0.5

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

▶ 향상도= $0.5/0.75=0.6666667$

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

항목별 지지도[Support]

번호	제품명	지지도(자료수/4)	
1	고등어	1	0.25
2	사과	2	0.5
3	생수	4	1
4	수박	1	0.25
5	옥수수	1	0.25
6	치즈	3	0.75
7	호두	2	0.5

연관성 분석(지지도, 신뢰도, 향상도)

미션: 생수를 구매한 사람이 치즈를 구매할 연관성에 대한 분석

지지도= $P(A \cap B)$
신뢰도= $P(A \cap B)/P(A)$
향상도= $\text{신뢰도}(A,B)/\text{지지도}$

▶ 지지도=

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

▶ 신뢰도=

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

▶ 향상도=

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
	고등어
3	수박
	사과
	생수
4	생수
	호두
	치즈
	옥수수

연관성 분석- R(연관분석에서 신뢰도 및 향상도)

```
rules=apriori(tr,parameter=list(supp=0.1,conf=0.1)) #지지도, 향상도 0.1 이상 자료  
inspect(rules)
```

#지지도, 향상도 0.1 이상 자료 (0.1은 10%를 의미함 숫자값은 사용자가 임의로 넣음)

```
rules=apriori(tr,parameter=list(supp=0.1,conf=0.1))  
inspect(rules)
```

15개의 규칙이 발견되었음을 의미함
여기 0이 나오면
지지도향상도 최소값
을 더 작게 변경해야
함

```
> rules=apriori(tr,parameter=list(supp=0.3,conf=0.3)) #  
Apriori  
  
Parameter specification:  
confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext  
0.3 0.1 1 none FALSE TRUE 5 0.3 1 10 rules FALSE  
  
Algorithmic control:  
filter tree heap memopt load sort verbose  
0.1 TRUE TRUE FALSE TRUE 2 TRUE  
  
Absolute minimum support count: 1  
  
set item appearances ... [0 item(s)] done [0.00s].  
set transactions ... [7 item(s), 4 transaction(s)] done [0.00s].  
sorting and recoding items ... [4 item(s)] done [0.00s].  
creating transaction tree ... done [0.00s].  
checking subsets of size 1 2 3 done [0.00s].  
writing ... [15 rule(s)] done [0.00s].  
creating S4 object ... done [0.00s].  
> inspect(rules)
```

	lhs	rhs	support	confidence	lift	count
[1]	{}	{사과}	0.50	0.5000000	1.000000	2
[2]	{}	{두유}	0.50	0.5000000	1.000000	2
[3]	{}	{치즈}	0.75	0.7500000	1.000000	3
[4]	{}	{생수}	1.00	1.0000000	1.000000	4
[5]	{사과}	{생수}	0.50	1.0000000	1.000000	2
[6]	{사과,두유}	{}	0.50	0.5000000	1.000000	2
[7]	{사과,치즈}	{}	0.50	1.0000000	1.333333	2
[8]	{사과,생수}	{}	0.50	0.6666667	1.333333	2
[9]	{두유,생수}	{}	0.50	1.0000000	1.000000	2
[10]	{치즈,생수}	{}	0.50	0.5000000	1.000000	2
[11]	{사과,두유,치즈}	{}	0.25	1.0000000	1.000000	1
[12]	{생수}	{치즈}	0.75	0.7500000	1.000000	3
[13]	{사과,생수,두유}	{}	0.50	1.0000000	1.000000	2
[14]	{사과,치즈,생수}	{}	0.50	1.0000000	1.333333	2
[15]	{두유,치즈,생수}	{}	0.50	0.6666667	1.333333	2

지지도

신뢰도

향상도

지지도, 신뢰도 30% 이상인 15개의 자료나옴
사과,치즈는 지지도가 0.25 이므로 나타나지 않음

치즈->생수
지지도: 0.75
신뢰도: 0.75
향상도: 1

연관성 분석(지지도, 신뢰도, 향상도)

#10개 항목만 보기 앞쪽의 Rules에서 10개 미만일때
##아래와 같이 1:10을 하면 에러나옴. 본인의 상황에
맞추어서 개수를 작업해야함.

```
inspect(rules[1:10])
```

```
inspect(sort(rules,by="lift")[1:10]) # lift(향상도) 높은순으로 10개
```

```
rules=apriori(tr,parameter=list(supp=0.25,conf=0.5)) #지지도 25%, 신뢰도 50% 이상구간
```

```
inspect(rules)
```

```
inspect(rules[16:17])
```

```
# 사과구매시 치즈를 사지 않을 확률이 33%있음.
```

```
inspect(rules[1:10]) #10개 항목만 보기
```

```
inspect(sort(rules,by="lift")[1:10]) # lift(향상도) 높은순으로 10개
```

```
연관결과<-inspect(sort(rules,by="lift"))
```

```
head(연관결과)
```

```
apply(연관결과$lift,연관결과$rhs,mean)
```

```
subset(연관결과,subset=(lift>=1)) #lift(향상도) 값이 1이상인값만 추출
```

```
subset(연관결과,subset=(lift>=1 & support>=0.5)) #lift(향상도) 값이 100이상인면서 support(지지도)가 50이상
```

```
사과연관분석<-연관결과[grep("사과",연관결과$lhs),] ## lhs 변수에 '사과' 가 포함된 자료만 추출
```

```
사과_lift_1이상<-subset(사과연관분석,subset=(lift>=1)) ## lhs 변수에 '사과' 글자 없는 자료만 추출
```

```
사과_lift_1이상
```

```
사과외연관분석=연관결과[-grep("사과",연관결과$lhs),]
```

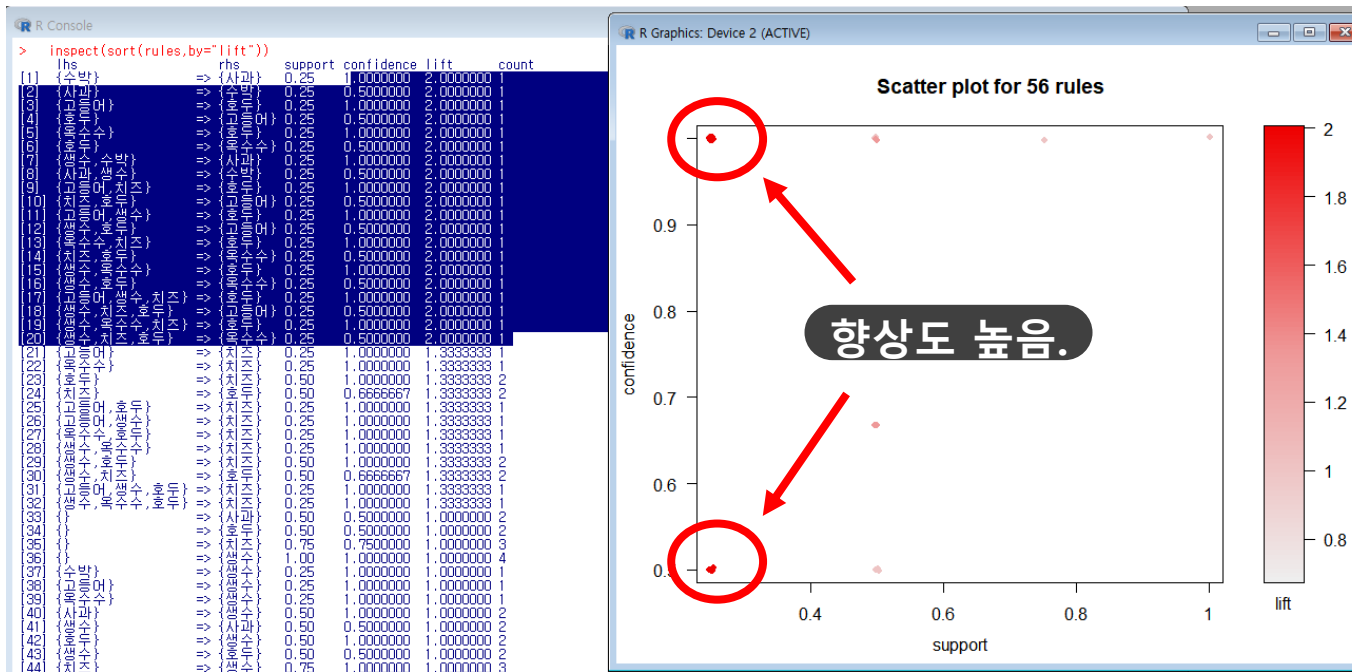
```
사과외연관분석
```

```
write.csv(사과연관분석, "c:/data_r/연관분석결과.csv")
```

```
set item appearances ... 10 item(s) ...  
set transactions ... [7 item(s), 4 tra  
sorting and recoding items ... [4 ite  
creating transaction tree ... done [C  
checking subsets of size 1 2 3 done [C  
writing ... [15 rule(s)] done [0.00s]  
creating S4 object ... done [0.00s].  
> inspect(rules)
```

연관성 분석- R[시각화, 차트]

```
install.packages("tidyr")
library(tidyr)
install.packages("arulesViz")
library(arulesViz)
rules=apriori(tr,parameter=list(supp=0.25,conf=0.5))
inspect(rules)
plot(rules)      # 가로(지지도), 세로(신뢰도), 색상(향상도)
                 #아래 자료는 지지도 0.25, 신뢰도 0.5와 1일때 향상도가 높음, 진한빨강색이 표시됨.
inspect(sort(rules,by="lift"))
```

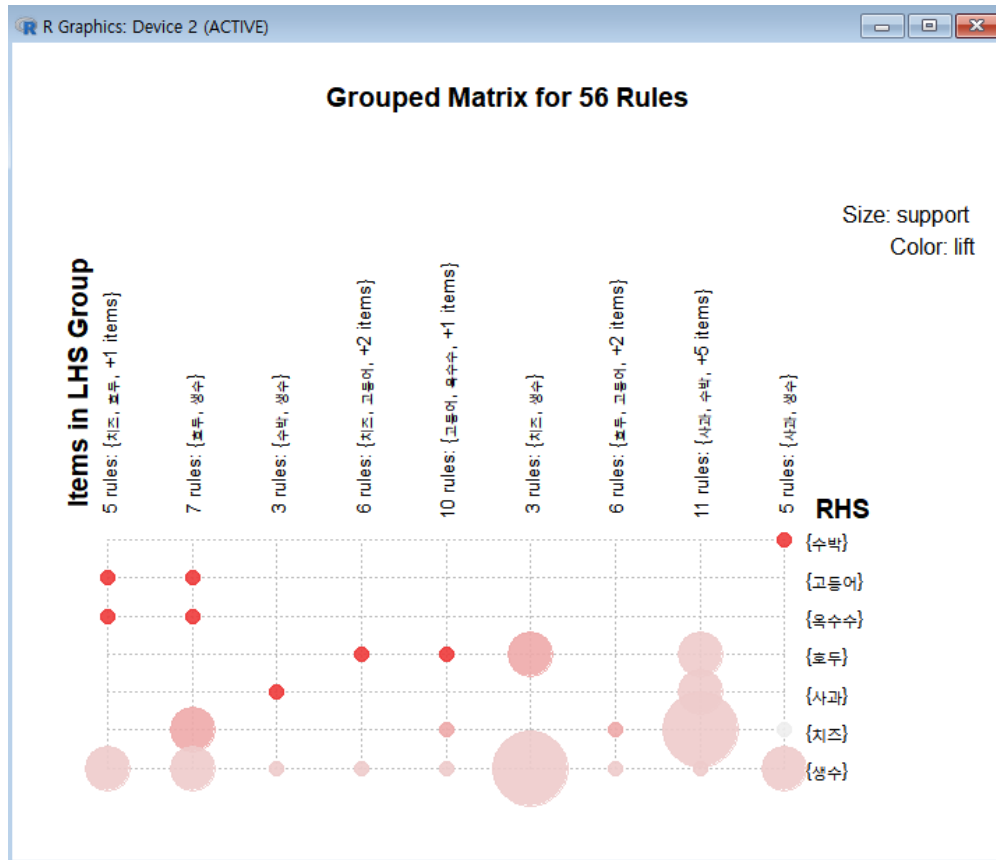


연관성 분석- R(시각화, 차트)

```
plot(rules,method="grouped")
```

#매트릭스차트

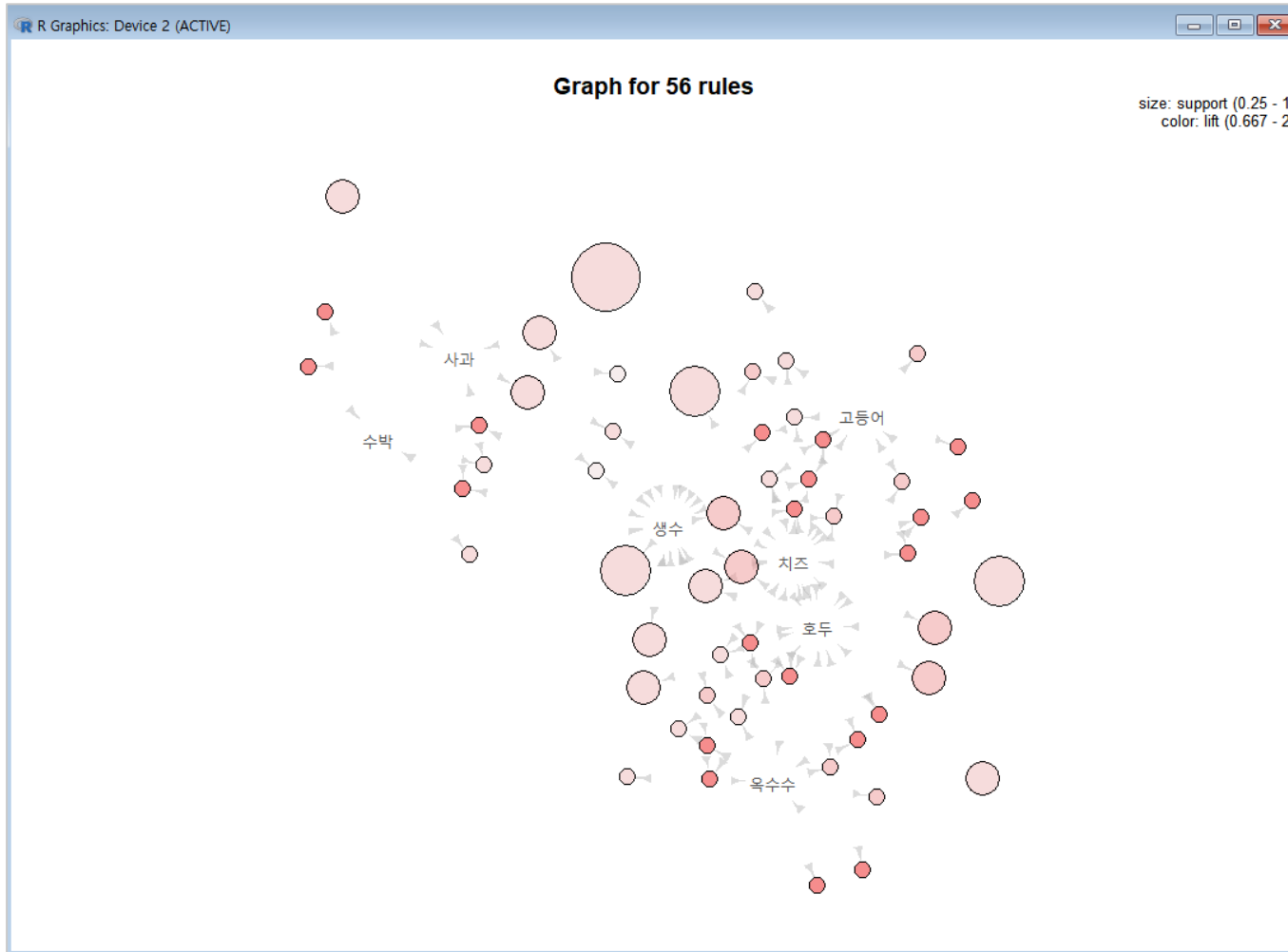
lhs(가로축)-조건(x아이템)과 rhs(세로축)-결과(y아이템) 으로구성한매트릭스그래프



연관성 분석- R(시각화, 차트)

```
plot(rules,method="graph") #네트워크차트
```

각규칙별로어떤아이템들이 연관되어있어있는지 보여주는네트워크그래프

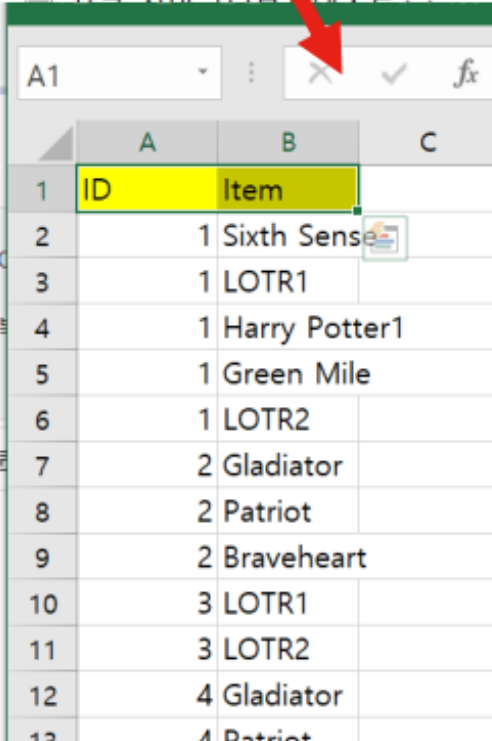
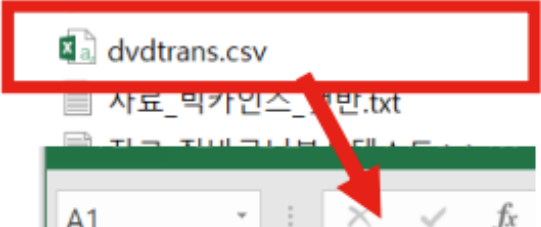


연관분석 (데이터 프레임일때)


- 데이터 프레임구조는 ID(구매고객)를 기준으로 item을 나누어서 작업해야함.
- 아래 사이트 자료임. 파일명: dvdtrans.csv

<http://blog.daum.net/sys4ppl/6>

좌측의 데이터 프레임을
우측의 구조로 변경하는 작업이 필요함.



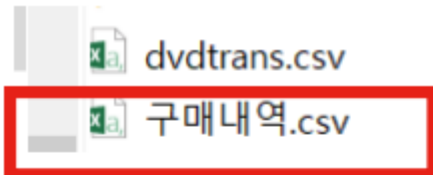
	A	B	C
1	ID	Item	
2		1 Sixth Sense	
3		1 LOTR1	
4		1 Harry Potter1	
5		1 Green Mile	
6		1 LOTR2	
7		2 Gladiator	
8		2 Patriot	
9		2 Braveheart	
10		3 LOTR1	
11		3 LOTR2	
12		4 Gladiator	
13		4 Patriot	



	A	B	C	D
1	넥타이	셔츠	양말	
2		1 넥타이		
3		1 셔츠		
4		1 양말		
5		2 양말		
6		2 벨트		
7		2 장갑		
8		2 셔츠		
9		3 지갑		
10		3 넥타이		
11		3 셔츠		
12		4 양말		
13		4 벨트		
14		4 장갑		
15		4 바지		

연관분석 에러

<https://blog.naver.com/PostView.nhn?isHttpsRedirect=true&blogId=gywlsangel&logNo=221325303378&parent>



번호	고객명	구매항목
1	홍길동	새우깡
2	홍길동	맛동산
3	홍길동	맥주
4	일지매	짱구
5	일지매	감자강
6	강감찬	감자깡
7	강감찬	새우깡
8	전우치	자갈치
9	전우치	맛동산
10	홍길동	짱구
11	어우동	빠다코코넛
12	어우동	맛동산
13	강감찬	포카칩
14	강감찬	맥주
15	김유신	자갈치
16	김유신	짱구
17	김유신	맛동산
18	홍길동	맛동산
19	전우치	초코칩쿠키
20	강감찬	크라운산도

구매항목을 고객명으로 나누면

홍길동 고객

새우깡, 맛동산, 맥주, 맛동산

→ 으로 맛동산 항목이 중복됨.

→ 연관분석은 중복데이터가 있으면 에러

위의 사이트를 참조하여서 반드시 실습

자료확인: '자료 빅카인즈 핫반.txt'

직접 SNS 상의 자료를 가져와서, 텍스트를 단어로 분리하여서 아래와 같이 단어, 단어, 단어 로 분리하는 작업을 하여서 텍스트간의 연관성을 찾는 ‘텍스트 마이닝’의 한 작업임

data	202
자료_빅카인즈_햇반.txt	202
자료_장바구니분석테스트.txt	202

엑셀자료중 P열의 특성추출
부분만 복사해서 붙임
txt로 저장

```
install.packages("arules")
library(arules)
install.packages("arulesViz")
library(arulesViz)
setwd("c:/data_r")
tr<-read.transactions ("자료_빅카인즈_햇반.txt",format="basket",sep=",")
tr
#지지도, 향상도 0.1 이상 자료 (0.1은 10%를 의미함 숫자값은 사용자가 임의로 넣음)
rules=apriori(tr,parameter=list(supp=0.05,conf=0.05))
inspect(rules)
inspect(rules[1:10])
inspect(sort(rules,by="lift")[1:10]) # lift(향상도) 높은순으로 10개
# 가로(지지도), 세로(신뢰도), 색상(향상도)
#아래 자료는 지지도 0.25, 신뢰도 0.5와 1일때 향상도가 높음, 진한빨강색이 표시됨.
plot(rules)
#매트릭스차트
# lhs(가로축)-조건(x아이템)과 rhs(세로축)-결과(y아이템) 으로구성한매트릭스그래프
plot(rules,method="grouped")
# 각규칙별로어떤아이템들이 연관되어뭉여있는지 보여주는네트워크그래프
plot(rules,method="graph") #네트워크차트
```