blackFriday<-read.csv('blackfriday.csv', stringsAsFactors = T)

str(blackFriday)

#str로 대략적인 데이터 구조를 파악한다

summary(blackFriday)

#summary로 많은 정보를 확인 가능하다

#많이 팔린 productid들, 남녀 성비, age그룹별 수,

#도시별 인원수, purchase boxplot값 등

#purchase를 구간화한 열 만들어줌

boxplot(blackFriday$Purchase)$stats

#min 12

#1st Qu. 5823

#median 8047

#3rd Qu. 12054

#Max. 21399

#구간화 위해 값 확인

blackFriday$Purchase2<- cut(blackFriday$Purchase, breaks=c(0,4000,8000,12000,16000,20000,24000))

table(blackFriday$Purchase2)

plot(blackFriday$Gender)

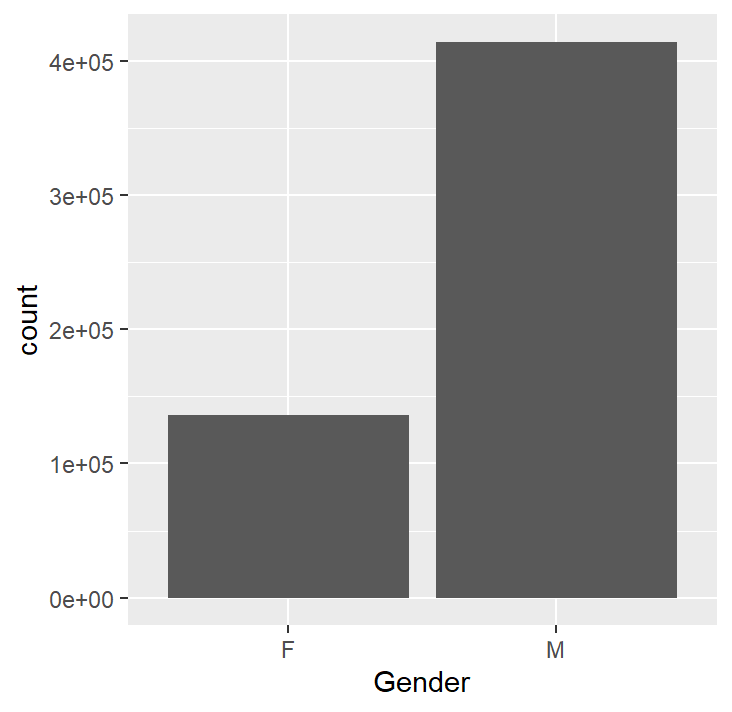
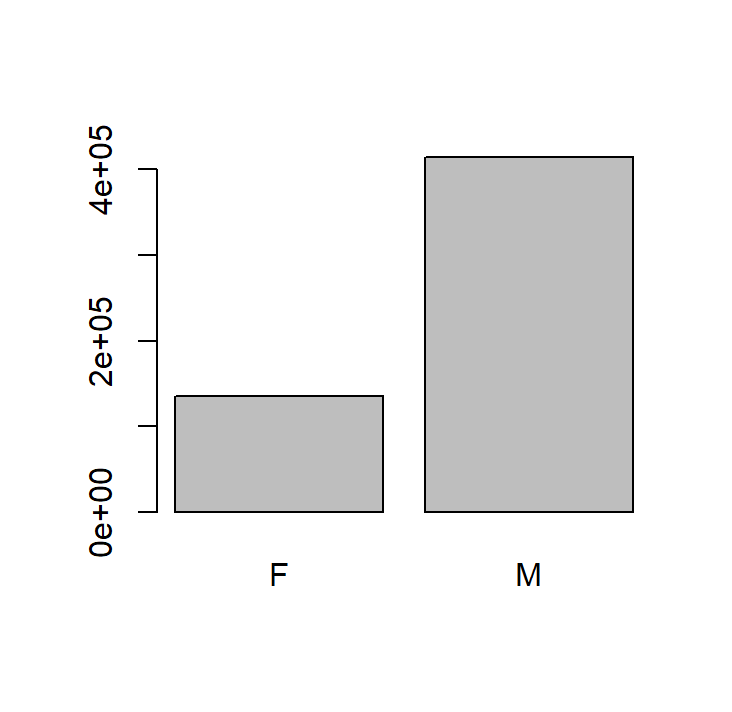
plot(blackFriday$Age)

#plot함수로 간단히 성비, 나이대별 인원 확인

#ggplot으로 표현하면 보다 정돈된 그래프로 확인 가능

ggplot(data=blackFriday, aes(x=Gender))+geom\_bar()

ggplot(data=blackFriday, aes(x=Age))+geom\_bar()



plot함수 / ggplot함수

pie(prop.table(table(blackFriday$Age)),

labels=paste(levels(blackFriday$Age)

,round(prop.table(table(blackFriday$Age)),3)\*100, '%'))

#나이대가 26-35세의 비중이 가장 높음을 파이 차트로 쉽게 확인 가능

ggplot(data=blackFriday, aes(x=Age, fill=Gender))+geom\_bar()

#특히 26-35세 남성의 구매 데이터가 가장 많았음

#여성의 경우도 해당 나이대의 구매수가 가장 많았음

#해당 나이대의 고객이 구매력도 갖췄는지 알아보고자 함

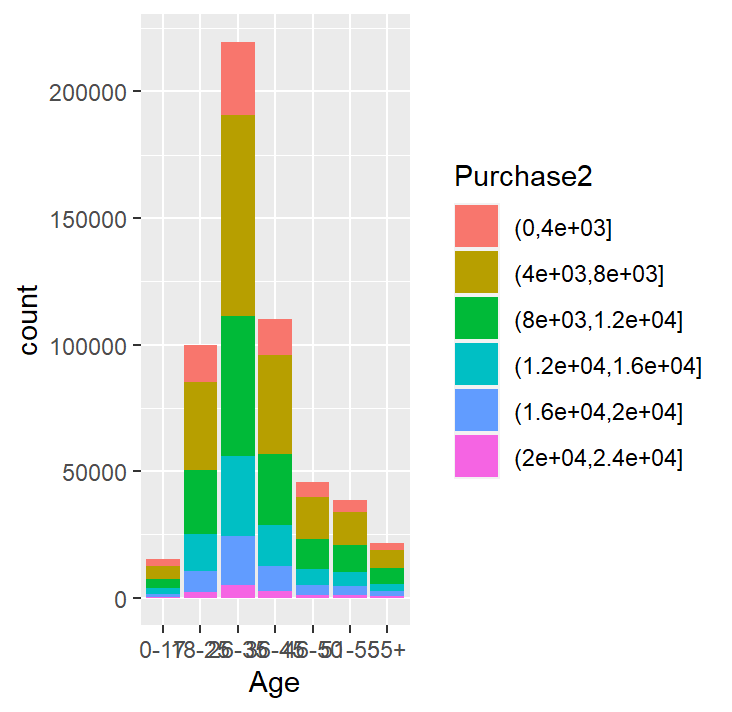
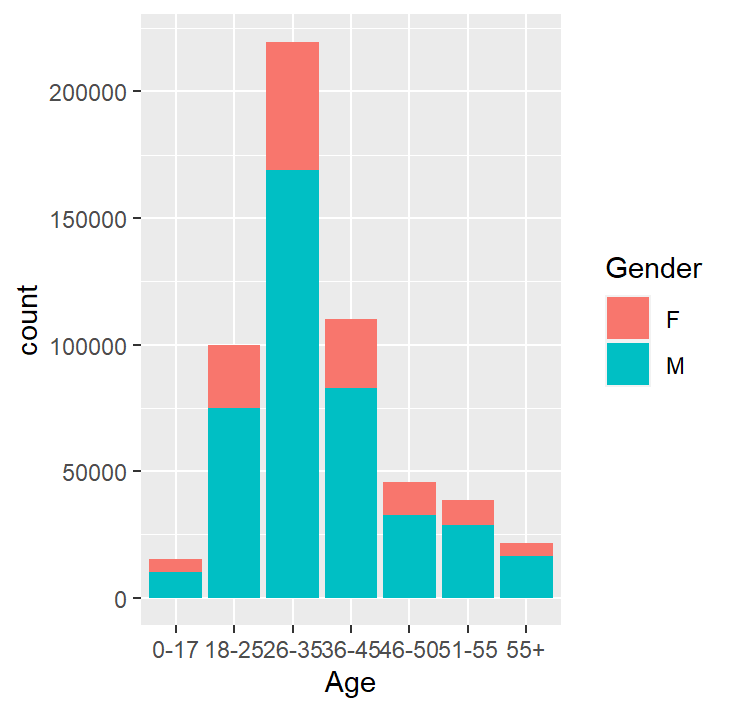
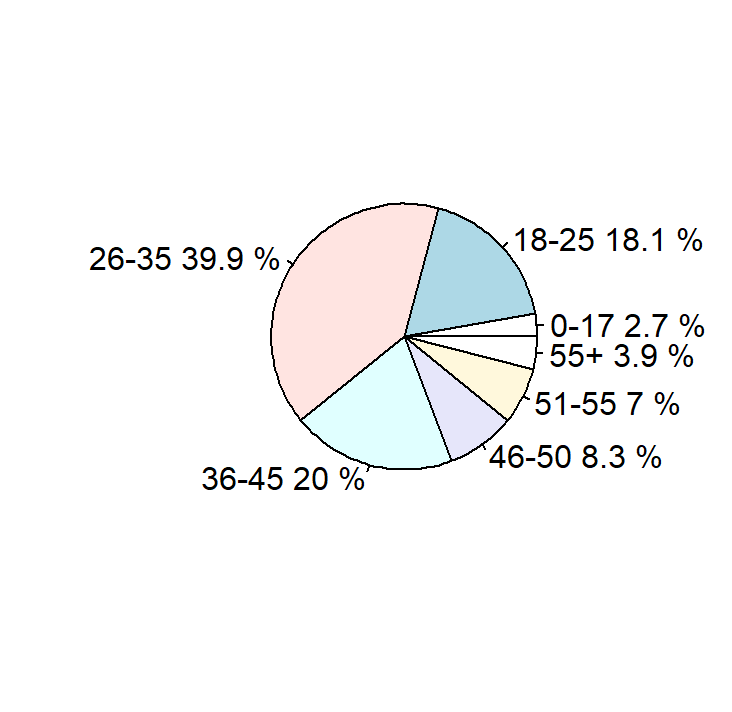
ggplot(data=blackFriday, aes(x=Age, fill=Purchase2))+geom\_bar()

#bar의 크기로 보아 구매력도 충분한듯

library(plotly)

ggplotly(ggplot(data=blackFriday, aes(x=Age, fill=Purchase2))+geom\_bar())

#interactive 시각화를 통해 정확한 숫자 확인 가능했음



나이대 파이차트/ 나이대+성비차트/ 나이대+구매력차트

ggplot(data=blackFriday, aes(x=Purchase2, fill=Age))

+geom\_bar(position='dodge')

ggplot(data=blackFriday, aes(x=Purchase2, fill=Age))

+geom\_bar(position='fill')

#따로 나타낸 그래프/ 비율로 나타낸 그래프 확인하고 싶었으나 안뜸

ggplot(data=blackFriday, aes(x=City\_Category,fill=Purchase2 ))+geom\_bar()

#B도시의 구매력이 타 도시에 비하여 높음을 알 수 있었음(인원수도 월등히 많음)

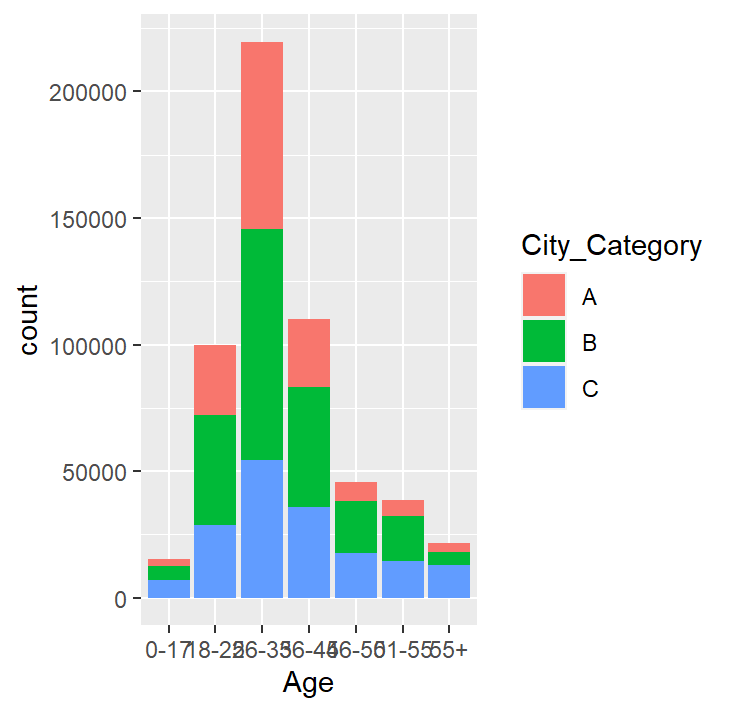
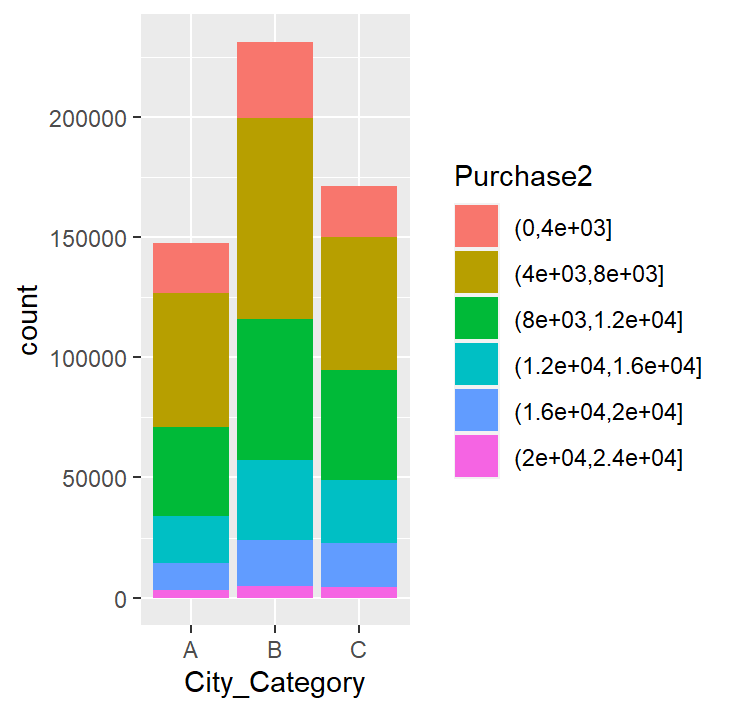
ggplot(data=blackFriday, aes(x=Age, fill=City\_Category))+geom\_bar()

#(나이대별 도시 거주 인원 수를 알아보고자 함)

#잠정적 타깃층으로 삼는 25-36 집단이 B>A>C 순으로 거주 인원이 많음을 알 수 있었음

#---> B도시의 25-36타깃층을 노리기

​



도시별 구매력/ 나이대별 도시인원

​

텍스트이(가) 표시된 사진

자동 생성된 설명

충성고객 구매best5 / 26-35best5 의 전체best5상품과 일치여부

library(dplyr)

#충성고객(단골고객)이 가장 많이 구매하는 물품 목록을 알아보고자 함

loyal\_customer100<-blackFriday %>%

group\_by(User\_ID) %>%

summarise(n=n()) %>%

arrange(desc(n)) %>%

head(100)

#userId가 중복되는 경우= 구매 횟수가 잦은 경우로 봄

#loyal\_customer(100)을 저장

loyal100<-loyal\_customer100$User\_ID

loyal\_best5<-blackFriday %>%

filter(User\_ID %in% loyal100) %>%

group\_by(Product\_ID) %>%

summarise(n=n()) %>%

arrange(desc(n)) %>%

head(5)

ggplot(data=loyal\_best5, aes(x=reorder(Product\_ID, n), y=n))+geom\_col()+

ylim(c(0,100))+

coord\_flip()

#충성고객의 베스트 판매상품

loyal\_best5$Product\_ID

best5<-blackFriday %>%

group\_by(Product\_ID) %>%

summarise(n=n()) %>%

arrange(desc(n)) %>%

head(5)

ggplot(data=best5, aes(x=reorder(Product\_ID, n), y=n))+geom\_col()+

coord\_flip()

loyal\_best5$Product\_ID %in% best5$Product\_ID

#충성고객의 best5상품과 전체 best5상품은 하나도 겹치지 않았음

best5\_26.35<-blackFriday %>%

filter(Age=='26-35') %>%

group\_by(Product\_ID) %>%

summarise(n=n()) %>%

arrange(desc(n)) %>%

head(5)

#해당 나이대의 best5상품은 전체 best5상품과 같음

best5\_26.35$Product\_ID %in% best5$Product\_ID

ggplot(data=best5\_26.35, aes(x=reorder(Product\_ID, n), y=n))+geom\_col()+

coord\_flip()

전체 베스트상품 P00265242 P00025442 P00110742 P00112142 P00057642

26-35 나이대 베스트상품 P00265242 P00110742 P00025442 P00112142 P00057642

충성고객 베스트상품 P00145042 P00249542 P00110942 P00117442 P00105142

​

​

​

​

####################연관규칙#############

dataset <- read.csv("blackfriday.csv")

library(tidyverse)

library(scales)

library(arules)

customers\_products <- dataset %>%

select(User\_ID, Product\_ID) %>%

group\_by(User\_ID) %>%

arrange(User\_ID) %>%

mutate(id = row\_number()) %>%

spread(User\_ID, Product\_ID) %>%

t()

write.csv(customers\_products, file = 'customers\_products.csv')

customersProducts <- read.transactions('customers\_products.csv', sep = ',', rm.duplicates = TRUE)

#읽는데 상당시간 소요됨

summary(customersProducts)

#transactions as itemMatrix in sparse format with

#5893 rows (elements/itemsets/transactions) and

#11575 columns (items) and a density of 0.008180618

#5893건의 거래와 11575건의 아이템(상품수), cell density

#most frequent items:

# P00265242 P00025442 P00110742 P00112142 P00057642

#1880 1615 1612 1562 1470

#best상품들과 판매개수

#element (itemset/transaction) length distribution:

#한번 거래 이루어질때 아이템 몇 개씩 사는지

CP<-customersProducts

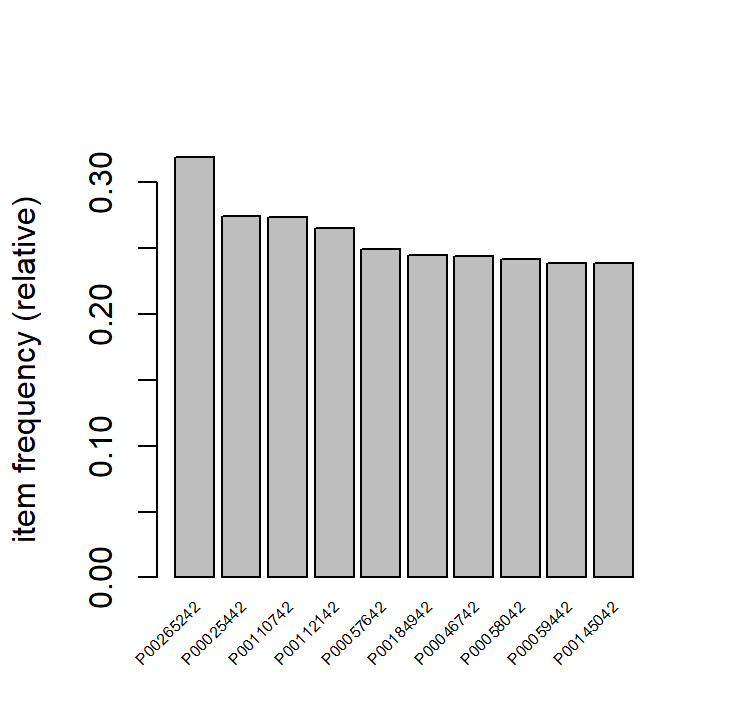
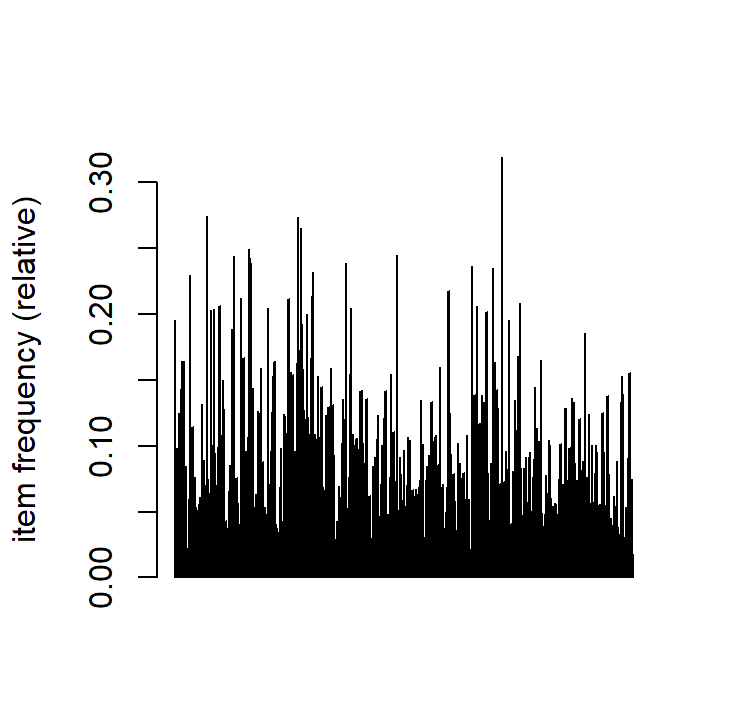
#너무 길어서 새 변수로 설정

itemFrequencyPlot(CP,support=0.01,cex=0.5)

#지지도 0.01이상이 상당히 많은듯하다

itemFrequencyPlot(CP, topN=10, cex=0.5)

#지지도 상위 10개 아이템 품목



#규칙 생성

#앞서 본 지지도 그래프에서 0.01 이상이 많았으므로 지지도를 0.01로 설정해보았다.

apriori(CP,parameter=list(support=0.01, confidence=0.5, minlen=2))

#규칙이 너무 많다

#신뢰도와 지지도를 조정해주며 적정한 규칙개수를 가질 수 있게 한다

CPrules<-apriori(CP,parameter=list(support=0.015, confidence=0.65, minlen=2))

summary(CPrules)

#3 4

#14 20

# 품목 3개 규칙 14개, 품목 4개 규칙 20개 만들어짐

#min lift는 2.376으로 1보다 높아 유의한 규칙인듯

#만들어진 규칙에 대해 충성고객 best5룰과

#전체(==26-35나이대)best5룰이 있는지 확인해보고자 함

best5\_loyal<-as.vector(loyal\_best5$Product\_ID)

loyal\_rules<-subset(CPrules, items %in% best5\_loyal)

#충성고객의 best5와 관련 규칙이 16개 있음

#그 중 상위 10개의 규칙만 확인

inspect(sort(loyal\_rules, by='lift')[1:10])

#대충 봤을 때 ~~~42로 끝나는 물품들이 같이 구매된 경우가 많은듯

#가독성이 좋지 않다...

best\_5<-as.vector(best5$Product\_ID)

best5\_rules<-subset(CPrules, items %in% best\_5)

summary(best5\_rules)

#23개 룰 있음

#상위 10개만 확인

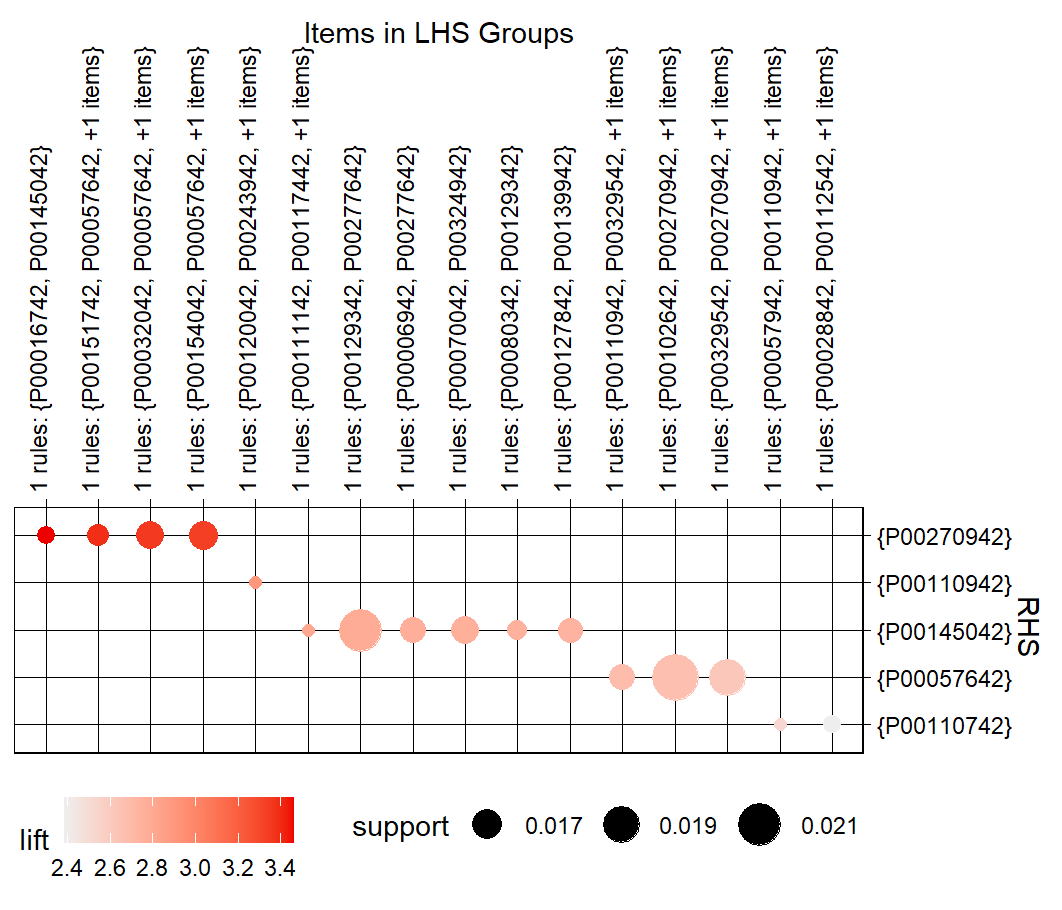
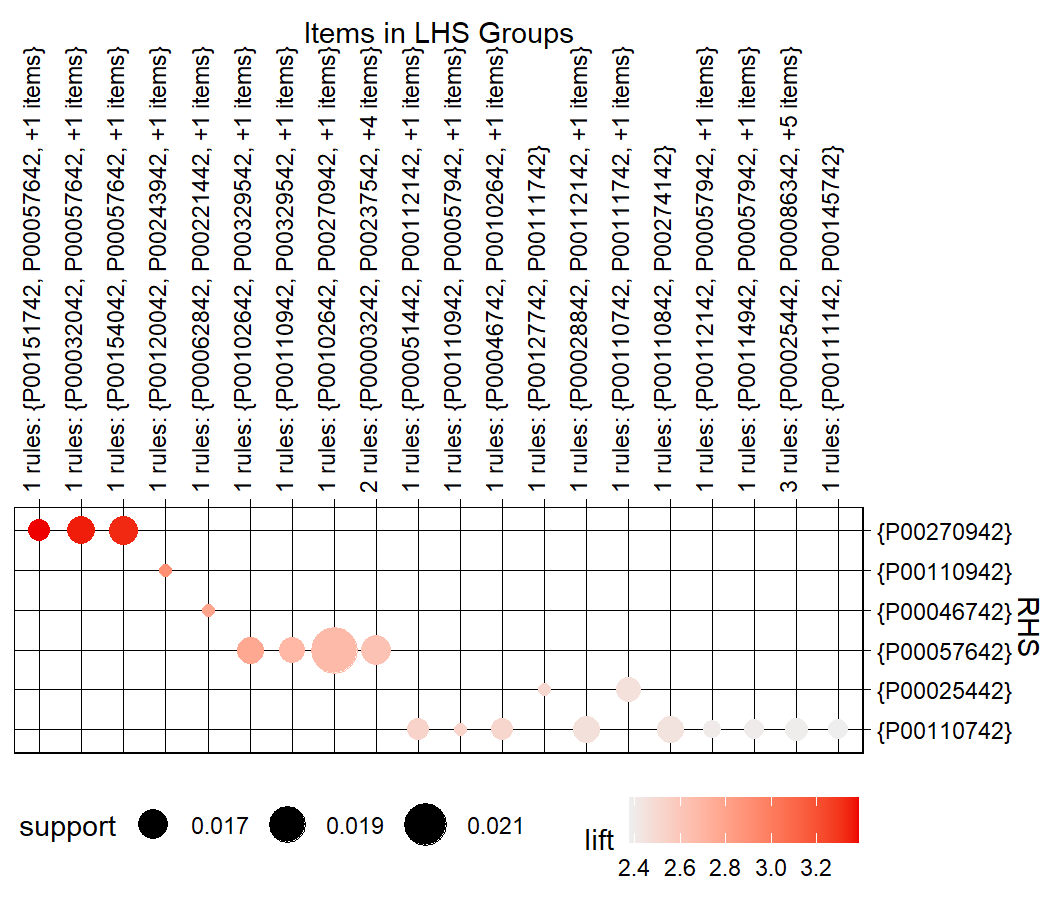
inspect(sort(best5\_rules, by='lift')[1:10])

#역시나 가독성이 좋지 않다

#묶음 상품

plot(sort(best5\_rules, by = 'lift'), method = 'grouped')

plot(sort(loyal\_rules, by = 'lift'), method = 'grouped')



전체 best5 / 충성고객best5

#위에서 충성고객과 전체 고객의 best5가 일치하지 않았던 것과 달리 규칙에서는 상당수 겹치는 모습을 확인할 수 있었다.

#특히 제품 P00270942와 다른 제품들의 연관성이 많은 것을 확인할 수 있었다(lift 가장 높고, 규칙도 각각 3개, 4개씩 존재)

sum(blackFriday$Product\_ID=='P00270942') # 제품'P00270942'의 판매량: 1152

# P00265242 P00025442 P00110742 P00112142 P00057642

#1880 1615 1612 1562 1470

#best상품들과 판매개수가 다음과 같았음을 참고했을 때 P00270942은 상당한 베스트셀러인듯

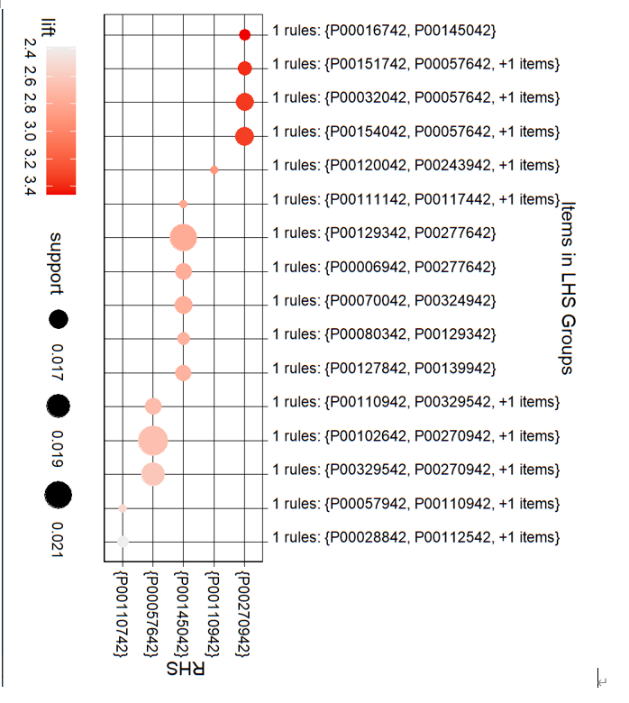
# {P00151742, P00057642}-->{P00270942} 각각 규칙1번, 2번으로 연관성 높은듯

sum(blackFriday$Product\_ID=='P00151742') #590으로 자체 판매량은 쏘쏘 #끼워팔기하기

#P00057642는 rhs에도 보이고(다른 상품과 연관성 높고) 자체로도 best5상품임

# P00057642-->다른 상품과의 연관성 높고, 자체 판매량도 높으므로 메인 위치에 놓기

테이블이(가) 표시된 사진

자동 생성된 설명

전체 best5/ 충성고객 best5

P00057642을 주력상품으로 밀기