# Assignment -3

1. In Fall 2018, UTD opened a new buffet where there are many food selections for faculty and students.

For simplicity, suppose five types of foods are offered daily: salad, hamburger, taco, soup and pasta.

Suppose you are the manager and you decide to use associate rules (manually) to figure out what

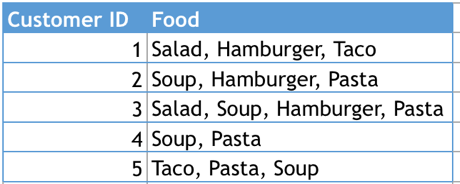
foods customers tend to purchase together. You recorded selections by five customers as shown in the

Table below. You also decide to use the following cutoffs: minimum support 40% and minimum

confidence 80%. What valid rules will you generate? Provide detailed steps with your relevant

calculations. Also report support, confidence and lift for the final rules you generate. (1+1+0.5 Points)

1. Given:

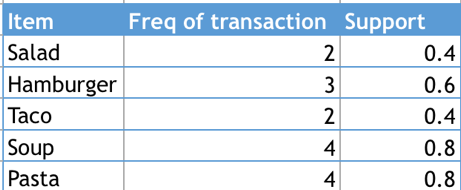


Min Support = 40%

Min Confidence = 80%

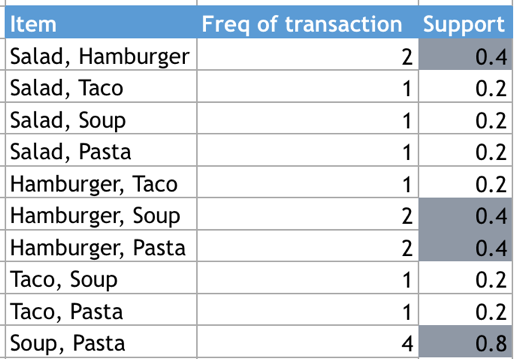
Phase (1):

Step (1): 1- Item Set



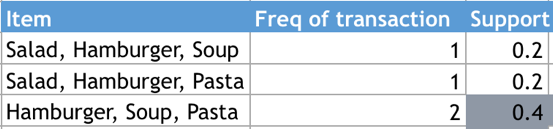
Step (2): Since min support (Threshold) = 40%, All the 1-item set are taken into consideration.

Step (3): 2-Item Set



Step (4): Since Min Support (Threshold) = 40%, Only the colored rows are taken into consideration.

Step (5): 3-Item Set

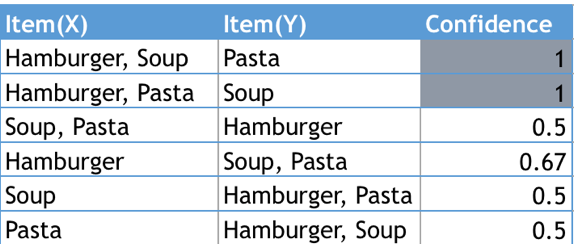


Step (6): Since the min support threshold is 40%, only, {Hamburger, Soup, Pasta} can be taken into consideration.

Phase (2):

{Hamburger, Soup, Pasta}

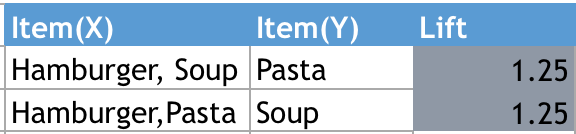
Step (1):



Step (2): Since the min confidence threshold is 80%, only the colored rows above are taken into consideration.

Step (3): Calculating the Lift:

Lift = Confidence of the rule / Support of the Consequence



We observe that lift for both the above items is >1, Therefore this rule says that if a customer eats

{Hamburger, Soup}, she/he is likely to also eat Pasta. (than her/his chance of eating Pasta if we know

nothing about him/her).

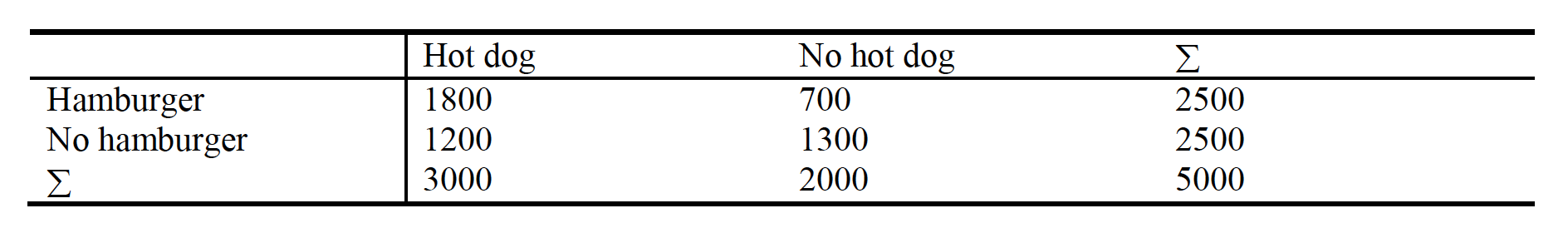
Similarly, if a customer eats {Hamburger, Pasta}, she/he is likely to also eat Soup. (than her/his

chance of eating Soup if we know nothing about him/her).

2. The following contingency table summarizes supermarket transaction data. (It is similar in format to

the table you see on slide 20 in our lecture notes on association rules.  means sum by row or

column.)



(a) Suppose that the association rule "hot dogs  hamburgers" is mined. Given a minimum support

threshold of 25% and a minimum confidence threshold of 50%, is this association rule valid? (0.5

Point)

1. Given: Min Support Threshold = 25%

Min Confidence Threshold = 50%

Support (hot dogs => hamburgers) = 1800/5000 = 0.36 => (0.36 > 0.25)

Confidence (hot dogs => hamburgers) = 1800/3000 = 0.6 => (0.6 > 0.5)

Since the support and confidence of (hot dogs => hamburgers) is greater than given threshold we can consider this rule to be valid.

(b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers?

If not, what kind of correlation relationship exists between the two (i.e., if a customer purchases hot

dogs, will that increase or decrease her chance of purchasing hamburgers)? (0.5+0.5 Point)

1. Based on the given data, W.K.T

Confidence (hot dogs => hamburgers) = 60%

P(hamburgers) = 2500/5000 = 50%

Confidence (Not hot dog) => hamburger) = 700/ 2000 = 35%

From the above calculations, we can say that, No, the purchase of hot dogs is not independent of the purchase of hamburgers.

Lift = 0.6/0.5 = 1.2

1.2>1, This rule says that, if a customer buys hotdog, he/she is likely to also buy hamburger.

\*We observe positive(increasing) correlation relationship between the two.

3. Conducting an Association Analysis Using R: A store is interested in determining the associations

between items purchased from the Health and Beauty Aids department and the Stationery Department.

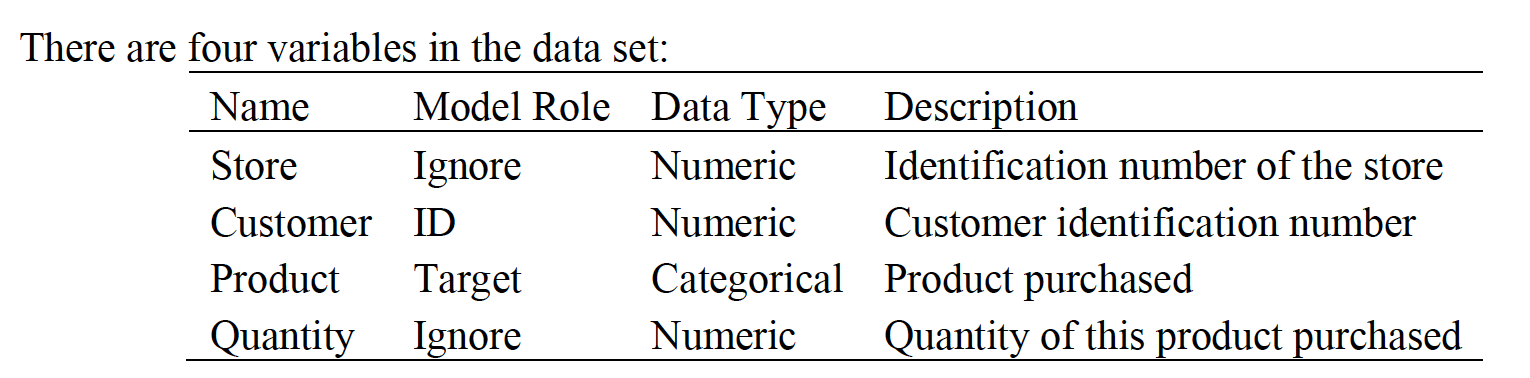
The store chose to conduct a market basket analysis of specific items purchased from these two

departments. “transactions” contains information about over 400,000 transactions made over the past

three months. The following 17 products are represented in the data set: bar soap, bows, candy bars,

deodorant, greeting cards, magazines, markers, pain relievers, pencils, pens, perfume, photo

processing, prescription medications, shampoo, toothbrushes, toothpaste, and wrapping paper. (1+0.5+0.5+0.5+1+0.5)



1. Import the data to R. Copy the R code used below. (Tip: use read.transactions)

transactions.df <- read.csv(file="transactions.csv", header=TRUE, sep=",")

# Dropping Store and Quantity Columns

transactions.df<- select(transactions.df,-c(1,3))

#Converting the dataframe to transaction data

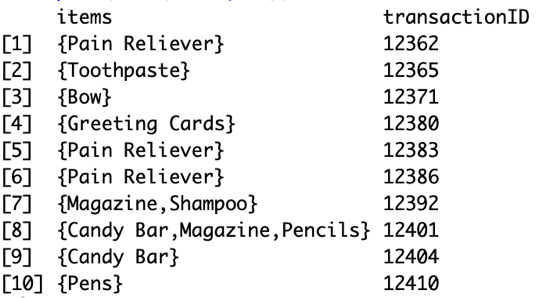
write.table(transactions.df, file = tmp <- file(), row.names = FALSE)

trans <- read.transactions(tmp, format = "single",

header = TRUE, cols = c("Customer", "Product"))

close(tmp)

inspect(head(trans, 10))



(b) Set Support to 0.01, Confidence to 0.10, and Min Length to 2. Run apriori to obtain the rules. Sort

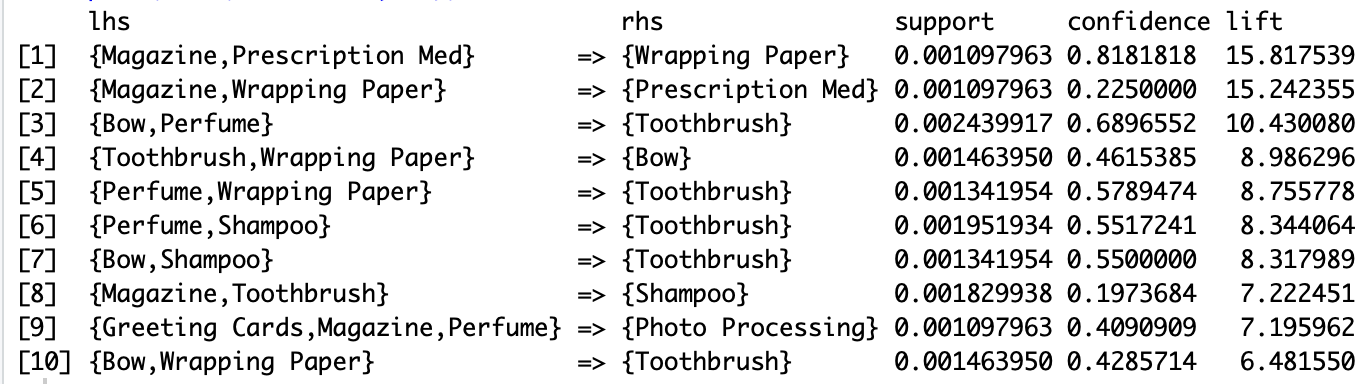
the rules according to “Lift” with descending order. Copy the R code used below.

rules <- apriori (trans, parameter = list(supp = 0.001, conf = 0.10, minlen = 2))

rules\_lift <- sort (rules, by="lift", decreasing=TRUE)

(c) Show the top ten Association Rules. Copy the code used and the result below

Inspect (head (rules\_lift, 10))



(d) What is the highest lift value for the resulting rules? Which rule has this value? Show how this lift

value was calculated.

The highest lift value is for the resulting rules is 15.81754, The rule that has the highest value is {Magazine,Prescription Med}=> {Wrapping Paper}

The lift is calculated by Confidence of {Magazine,Prescription Med}/ Support of {Wrapping Paper}

#itemFrequency(trans[,])

# Support of Wrapping Paper

# 0.0517262413

# Confidence of {Magazine,Prescription Med}

#0.8181818

0.8181818/ 0.0517262413

# [1] 15.81754

(e) Interpret the first five rules in the output in words.

#1.

# {Magazine,Prescription Med} => {Wrapping Paper}, This rule says that if a customer buys {Magazine,Prescription Med}, she/he is 15.817539 times

#likely to buy {Wrapping Paper} (than her/his chance of buying {Wrapping Paper} if we know nothing about her/him).

#2.

# {Magazine,Wrapping Paper} => {Prescription Med}, This rule says that if a customer buys {Magazine,Wrapping Paper}, she/he is 15.242355 times

#likely to buy {Prescription Med} (than her/his chance of buying {Prescription Med} if we know nothing about her/him).

#3.

# {Bow,Perfume}=> {Toothbrush} , This rule says that if a customer buys {Bow,Perfume}, she/he is 10.430080 times

#likely to buy {Toothbrush} (than her/his chance of buying {Toothbrush} if we know nothing about her/him).

#4.

# {Toothbrush,Wrapping Paper} => {Bow} , This rule says that if a customer buys {Toothbrush,Wrapping Paper}, she/he is 8.986296times

#likely to buy {Bow} (than her/his chance of buying {Bow} if we know nothing about her/him).

#5.

# {Perfume,Wrapping Paper} => {Toothbrush} , This rule says that if a customer buys {Perfume,Wrapping Paper}, she/he is 8.755778 times

#likely to buy {Toothbrush} (than her/his chance of buying {Toothbrush} if we know nothing about her/him).

(f) Reviewing the top 10 rules, based on their lift ratios, comment on their redundancy and how you

would assess their utility as a decision maker.

rules\_top10 <- head(rules\_lift, 10)

subsetRules <- which(colSums(is.subset(rules\_top10, rules\_lift)) > 1) # get subset rules in vector

length(subsetRules) #> 6

rules\_top10 <- rules\_top10[-subsetRules] # remove subset rules.

inspect(rules\_top10)

