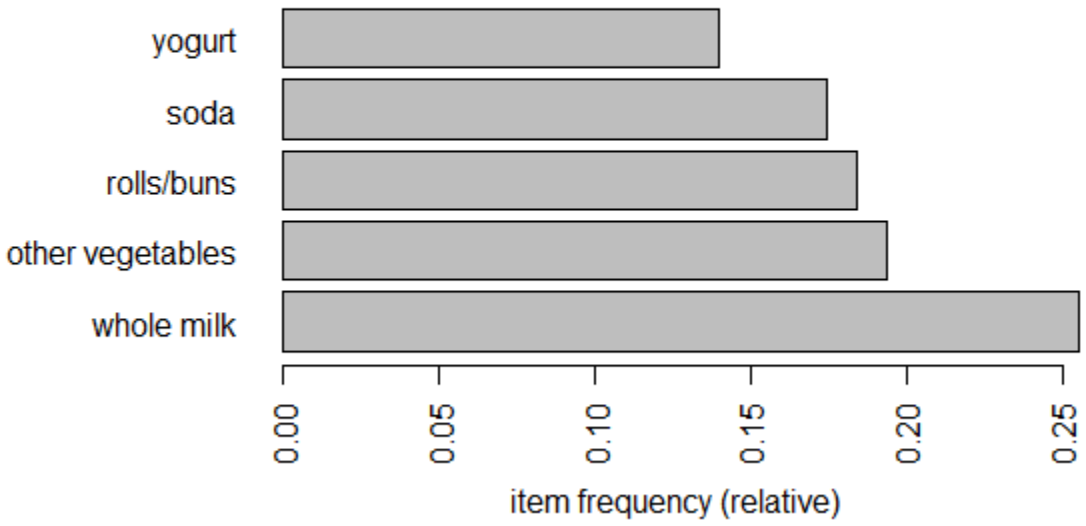
Stat 515 – Assignment 5

Saurabh Rao Donthineni

Part A



The top 5 items with the highest frequency , in increasing order are : yogurt , soda, rolls/buns , other vegetables , whole milk. Whole milk has the highest frequency and yogurt has the lowest among the top 5.

Part B

The total number of association rules obtained are 873.

Top 5 rules with the highest lift values are :

```
    lhs                     rhs                       support confidence     lift
[1] { citrus fruit,
      other vegetables,
      whole milk}       => { root vegetables}    0.005795628  0.4453125 4.085493
[2] { butter,
      other vegetables} => { whipped/sour cream} 0.005795628  0.2893401 4.036397
[3] { herbs}            => { root vegetables}    0.007015760  0.4312500 3.956477
[4] { citrus fruit,
      pip fruit}        => { tropical fruit}     0.005592272  0.4044118 3.854060
[5] { berries}          => { whipped/sour cream} 0.009049314  0.2721713 3.796886
```

The association rule with the highest lift value is :

```
{ citrus fruit,
      other vegetables,
      whole milk}       => { root vegetables}
```

Lift value is the ratio of confidence to expected confidence, used to measure the significance of the rule that is being considered. If the lift ratio value is higher than 1, it means that the relationship between the LHS and RHS side of the rule is more significant when compared to if the two sets were independent. The insight that we can understand is the statistical significance of the rule.

Part C

A total of 29 rules are generated.

The top 5 rules are :

```
    lhs                   rhs     support     confidence lift
[1] { processed cheese} => { soda} 0.005287239 0.3190184  1.829473
[2] { candy}            => { soda} 0.008642603 0.2891156  1.657990
[3] { chocolate}        => { soda} 0.013523132 0.2725410  1.562939
[4] { dessert}          => { soda} 0.009862735 0.2657534  1.524015
```

```
[5] { specialty bar}    => { soda} 0.007219115 0.2639405  1.513618
```

Part D

The statistical measures for this association are as follows :

```
lhs                           rhs     support    confidence lift
[1] { fruit/vegetable juice} => { soda} 0.01840366 0.254571   1.459887
```

Since the lift value is 1.45 (which is greater than 1) , we can say that ratio of confidence to expected confidence is high. This means that the data and association rules support the statement that placing the soda and fruit/vegetables together makes sense.

Part E

The number of rules for this is 3 .

The statistical measures are as follows :

```
   lhs                                           rhs       support
confidence
[1] { whipped/sour cream, whole milk}       => { butter} 0.006710727
0.2082019
[2] { other vegetables, whipped/sour cream} => { butter} 0.005795628
0.2007042
[3] { domestic eggs, whole milk}            => { butter} 0.005998983
0.2000000
    lift
[1] 3.757185
[2] 3.621883
[3] 3.609174
```

The best predictors for determining who will purchase butter are :

Whipped/sour cream , whole milk , other vegetables and domestic eggs.

These products have approximately the same lift values (differ by 0.1 approx) , so it can be said that the predictive powers for all of them are the same.

Appendix

```
# clear existing variables from global environment

rm(list = ls())



# install required packages

install.packages("arules")

install.packages("arulesViz")

library(arules)

library("arulesViz")



# read file using read.csv
```

```
data <- read.csv("C:/Users/SOURAV/Desktop/Groceries.csv")

# has to be made into factor as data is categorical

data$Product <- factor(data$Product)

items <- split(x=data[,"Product"],f=data$ID)

items <- lapply(items,unique)

items <- as(items,"transactions")

itemFrequency(items)



#Part A

itemFrequencyPlot(items,topN=5)

itemFrequencyPlot(items,horiz=TRUE,topN=5)

#Part B



rules_b <- apriori(items, parameter = list(support=0.005, confidence=0.2))

#total number of association rules
```

```
rules_b

# top 5

inspect(head(sort(rules_b, by="lift"),5))

#Part C

rules_c = apriori(items, parameter = list(support = 0.005,confidence = 0.2,

maxlen = 2), appearance = list(default="lhs",rhs=" soda"))

rules_c

inspect(head(sort(rules_c, by="lift"),5))

#Part D

inspect(rules_c[rules_c@lhs@data@i==(which(levels(data$Product)==" fruit/vegetable juice")-
1)])

#Part E

rules_e = apriori(items, parameter = list(support = 0.005,confidence = 0.2), appearance =
list(default="lhs",rhs=" butter"))

rules_e

inspect(head(sort(rules_e, by="lift"),5))
```