AD699: Data Mining for Business Analytics

Individual Assignment 3: Association Rules

Wendy Qing

Background

For this portion of the assignment, I use the data about Groceries, a dataset that can be found with the `arules` package in R. Each row in the file represents one buyer's purchases. I have also consulted the resource on this link http://r-statistics.co/Association-Mining-With-R.html that provides some helpful templated examples for generating association rules.

```
## Loading required package: pacman
```

Question 1

Describe "Groceries" by answering following questions:

Part A

What is the class of "Groceries"?

Groceries is a `transactions` object.

```
## [1] "transactions"
## attr(,"package")
## [1] "arules"
```

Part B

How many rows and columns does Groceries contain?

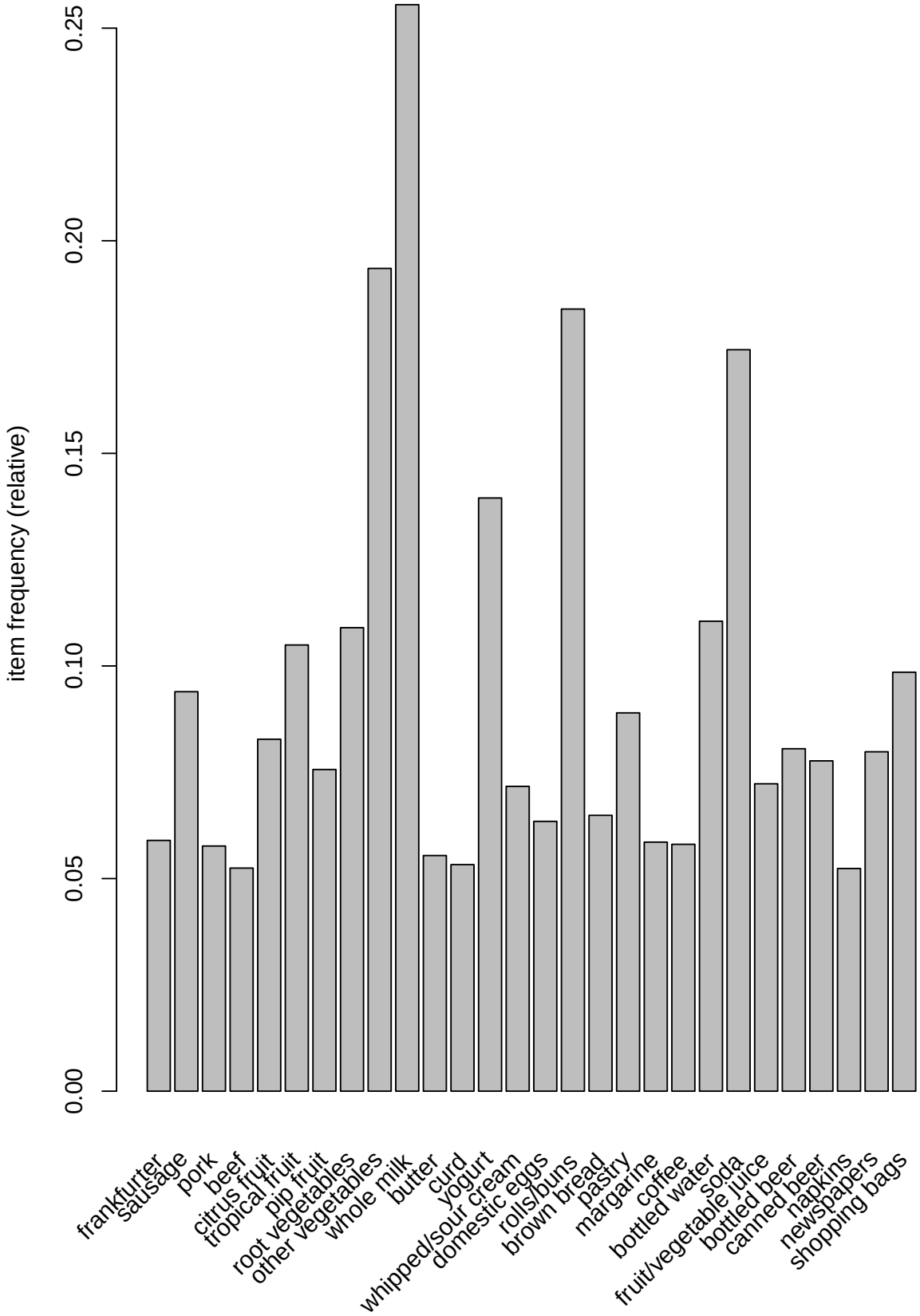
The Groceries data has 9835 rows and 169 columns.

```
## transactions in sparse format with
##  9835 transactions (rows) and
##  169 items (columns)
```

Question 2

Generate an item frequency bar plot for the grocery items with support rate greater than 0.05. Include a screenshot of your results, along with the code you used to do this.

```
## Item frequency plot with support > 0.05
itemFrequencyPlot(Groceries, support = 0.05, main = "Item Frequency Plot: Support > 0.05")
```


Item Frequency Plot: Support > 0.05
item frequency (relative)
0.00
0.05
0.10
0.15
0.20
0.25
frankfurter
sausage
pork
beef
citrus fruit
tropical fruit
pip fruit
root vegetables
other vegetables
whole milk
butter
curd
yogurt
whipped/sour cream
domestic eggs
rolls/buns
brown bread
pastry
margarine
coffee
bottled water
soda
fruit/vegetable juice
bottled beer
canned beer
napkins
newspapers
shopping bags

Question 3

Now, create a subset of rules that contain your grocery item (you can find your item in the spreadsheet in Blackboard, in Class Discussions, From Your Instructor). Select 4 different rules, (2 lhs and 2 rhs), and explain them in the way you would explain them to your room mate (I'm assuming your room mate is a smart person who is unfamiliar with data mining).

Remember, every rule has three components: support, confidence, and lift.

Right hand side rules

Listed below are the 6 right hand rules that meet the criteria of 0.001 support and a confidence of 0.5. I have sorted these rules by descending order of confidence.

```
## Create the rules
my_rules <- apriori(data = Groceries,
                    parameter = list(supp = 0.001,
                                     conf = 0.5),
                    appearance = list (default="lhs",
                                       rhs="whole milk"),
                    control = list (verbose=F)
                    )
```

```
## Sort the rules
right_rules_conf <- sort (my_rules,
                    by="confidence",
                    decreasing=TRUE)

inspect(head(right_rules_conf))
```

```
##     lhs                       rhs          support confidence coverage lift count
## [1] {rice,
##      sugar}                => {whole milk} 0.00122          1  0.00122 3.91    12
## [2] {canned fish,
##      hygiene articles}     => {whole milk} 0.00112          1  0.00112 3.91    11
## [3] {root vegetables,
##      butter,
##      rice}                 => {whole milk} 0.00102          1  0.00102 3.91    10
## [4] {root vegetables,
##      whipped/sour cream,
##      flour}                => {whole milk} 0.00173          1  0.00173 3.91    17
## [5] {butter,
##      soft cheese,
##      domestic eggs}        => {whole milk} 0.00102          1  0.00102 3.91    10
## [6] {pip fruit,
##      butter,
##      hygiene articles}     => {whole milk} 0.00102          1  0.00102 3.91    10
```

Left hand side rules

In the case of milk, there are no left hand rules that meet our criteria of 0.001 support and a confidence of 0.5. I use the right hand rules for the rest of the analysis.

```
## Left rules
left_rules <- apriori(data = Groceries,
                    parameter = list(supp = 0.001,
```

```
                                    conf = 0.5),
                    appearance = list(default="rhs",
                                      lhs="whole milk"),
                    control = list (verbose=F)
                    )
left_rules
```

```
## set of 0 rules
```

```
## Inspect left rules
left_rules_conf <- sort(left_rules,
                    by="confidence",
                    decreasing=TRUE)

inspect(head(left_rules_conf))
```

For each group of rules (grocery item on left-hand side, and grocery item on right-hand side), include a screenshot of your rules, along with the code you used to generate the rules In a sentence or two, explain what meaning these rules might have for a supermarket retailer, such as Star Market. What could it do with this information?

Interpreting the Rules

Rule 1

I use the first rule from the table in section 4 (above) to illustrate the meaning of these rules. The first rule is as follows.

{rice,sugar} => {whole milk}

This rule means that consumers who buy the basket of goods on the left (rice, sugar) often buy whole milk. But How often do they buy whole milk? To answer this question we explore the values for support, confidence and lift for this rule.

Support: First the combination of rice and sugar constitute 0.122% (0.00122) of all sales.

Confidence: When a customer buys rice and sugar, there is a 100% chance of buying brushes. In other words, we are certain that buyers of rice and sugar will buy whole milk.

Lift: Having rice and sugar in a shopping basket raises the probability that a customer will buy brushes 3.91 times.

Rule 2

Similarly, {canned fish, hygiene articles} constitute 0.112% (0.00112) of sales. Customers having these items in the basket have a 100% chance of buying whole milk. The purchase of {canned fish, hygiene articles} raises the probability of buying whole milk by 3.91.

Rule 3

In the same line, {root vegetables, butter, rice} constitute 0.102% (0.00102) of sales. Customers having these items in the basket have a 100% chance of buying whole milk. The purchase of {root vegetables, butter, rice} raises the probability of buying whole milk by 3.91.

Rule 4

Lastly, {root vegetables, whipped/sour cream, flour} constitute 0.173% (0.00173) of sales. Customers having these items in the basket have a 100% chance of buying whole milk. The purchase of {root vegetables, whipped/sour cream, flour} raises the probability of buying whole milk by 3.91.

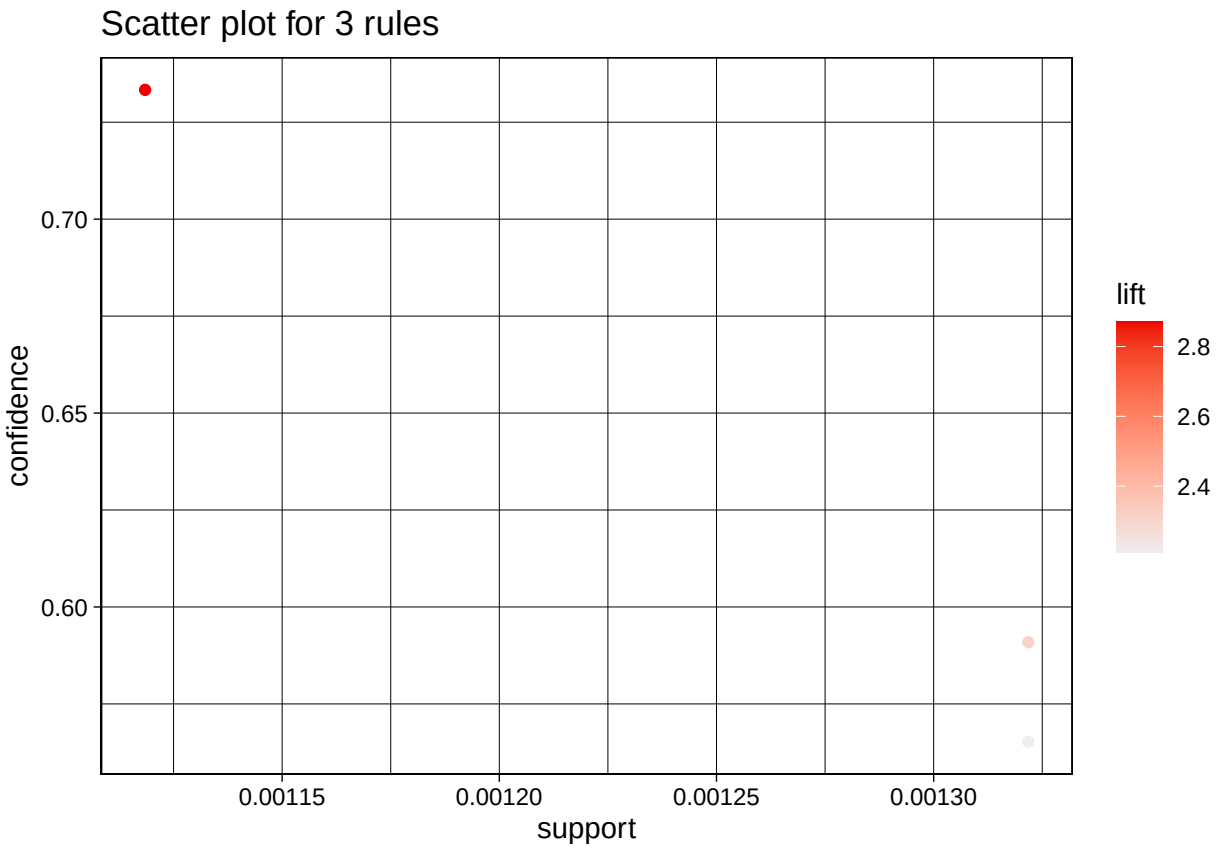
Usefulness of these Rules for a Supermarket Retailer

First, the following items have a very high chance of being in the same basket with whole milk: Rice, sugar, canned fish, hygiene articles, root vegetables, butter,whipped/sour cream, and flour. The retailer should consider stocking whole milk in a shelf near these items to maximize the sale of whole milk. In addition, we could recommend milk to a customer that just purchased the product combinations of {rice, sugar}, {canned fish, hygiene articles}, {root vegetables, butter, rice}, and {root vegetables, whipped/sour cream, flour}.

Question 4

Using the plot() function in the `arulesViz` package, generate a scatter plot of any three rules involving your grocery item. Include a screenshot of your plot, along with the code you used to generate the plot. Describe your results in a sentence or two.

```
## Plot the rules
plot(my_rules[1:3])
```



Two of the rules have relatively low confidence but involve items that are bought frequently and hence high support. One of the rules is an outlier with very high confidence but with relatively low sales. While the outlier is useful, the rules with more support could yield more sales and hence generate more profits.

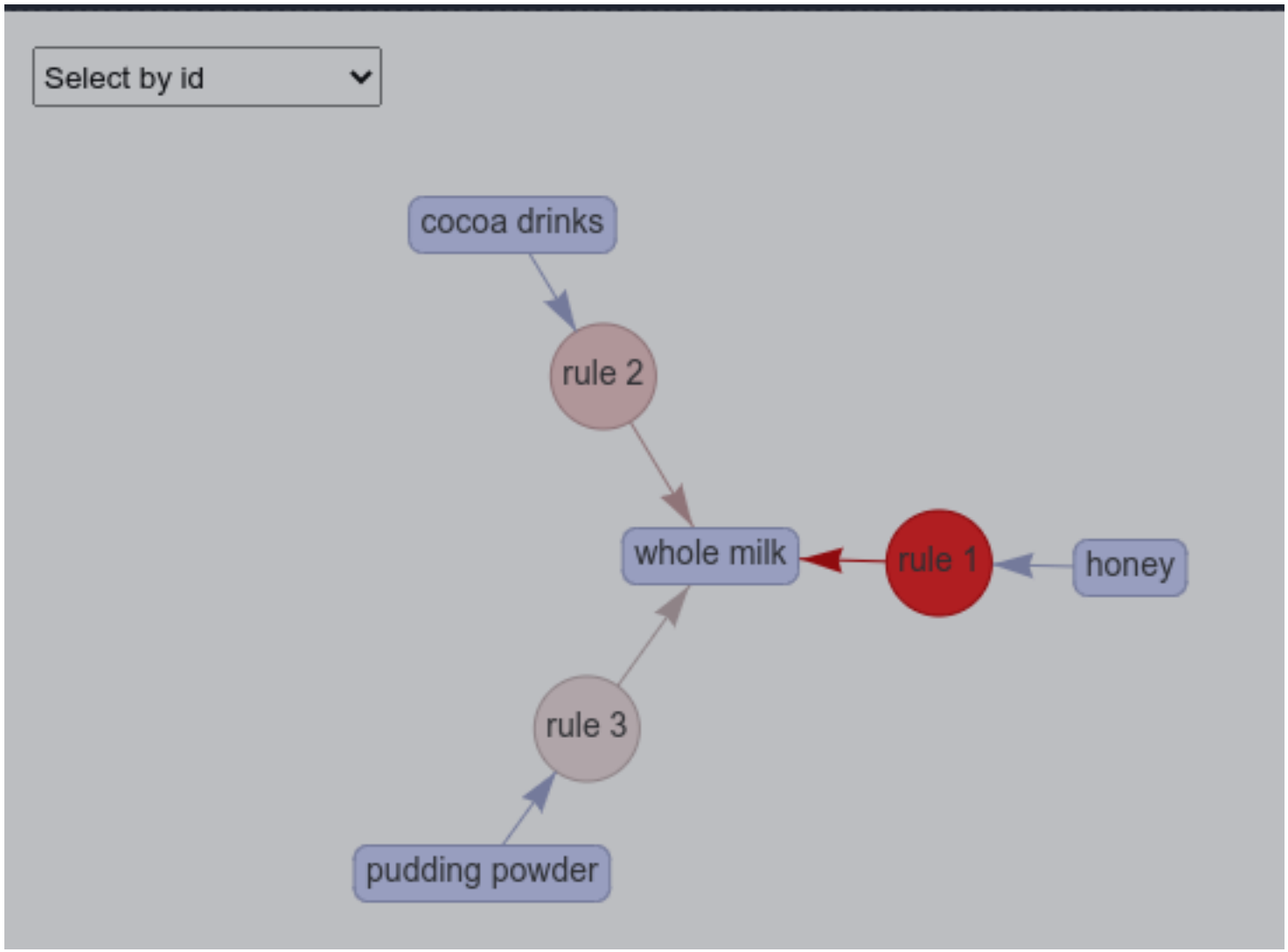
Question 5

Again using the plot() function in the `arulesViz` package, generate a plot for any three of your rules. This time, add two more arguments to the function: `method="graph", engine="htmlwidget"`. What do you see now? Include a screenshot of your plot, along with the code you used to generate the plot. Describe your results in a sentence or two.

```
plot(my_rules[1:3], method="graph", engine="htmlwidget")
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

```
## Plot the rules
knitr::include_graphics("assoc.png")
```



This plot shows that whole milk is associated with multiple rules. In rule 2, cocoa drinks have a higher relationship with whole milk. The same is the case for pudding powder and honey in rule1 and rule 3 respectively.