DATA 624 - Homework 10

OMER OZEREN

Table of Contents

Introduction

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer's basket - and therefore 'Market Basket Analysis'.

That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item. The data set is attached.

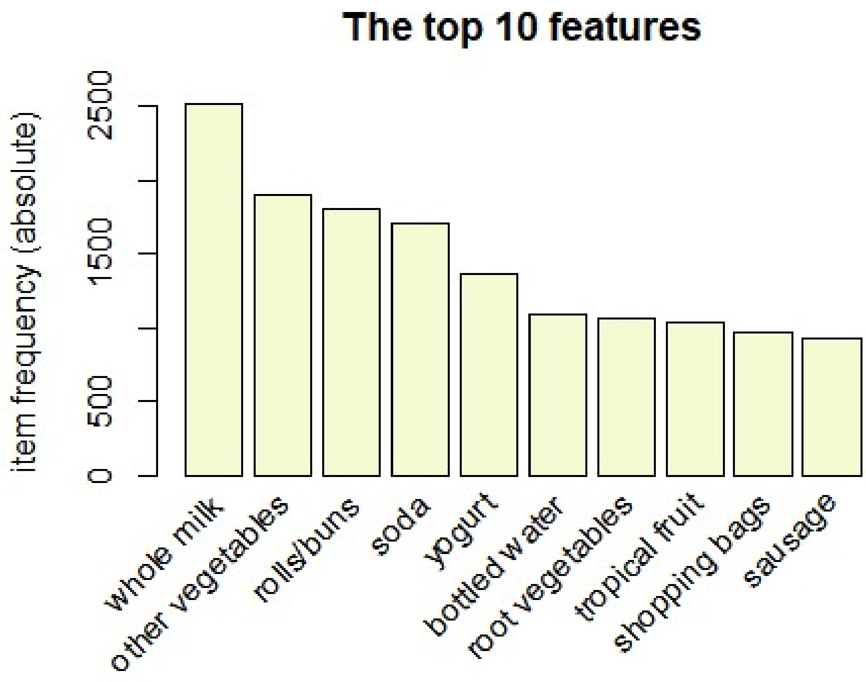
Your assignment is to use R to mine the data for association rules. You should report support, confidence and lift and your top 10 rules by lift.

Extra credit: do a simple cluster analysis on the data as well. Use whichever packages you like.

Import Data / Plot Features

I'll use itemFrequencyPlot for plotting top to features.I found the documentation from https://www.rdocumentation.org/packages/arules/versions/1.5-5/topics/itemFrequencyPlot.

```
#data <-
read.csv("https://raw.githubusercontent.com/omerozeren/DATA624/master/HMW10/G
roceryDataSet.csv")
#itemFrequencyPlot(data, topN=10, type="absolute", main="The top 10
features",col="#f2fbd2")
data <- read.transactions("GroceryDataSet.csv", sep=",")
itemFrequencyPlot(data, topN=10, type="absolute", main="The top 10
features",col="#f2fbd2")
```



The graph above indicates that the most important feature is WholeMilk and other vegetables follows it

Apriori Algorithm - Top 10 Rules

For Market Analyisis , I'll implement Apriori algorithm. Mine frequent itemsets, association rules or association hyperedges using the Apriori algorithm. The Apriori algorithm employs level-wise search for frequent itemsets. Ref : https://www.rdocumentation.org/packages/arules/versions/1.6-6/topics/apriori

```
top_10_rules<- apriori(data, parameter=list(supp=0.001, conf=0.5) ,
control=list(verbose=FALSE))

  top_10_rules %>%
  DATAFRAME() %>%
  arrange(desc(lift)) %>%
  top_n(10) %>%
  kable() %>%
  kable_styling()

## Selecting by count
```

LHS

RHS

support

confidence

```
coverage
lift
count
{root vegetables,tropical fruit}
{other vegetables}
0.0123030
0.5845411
0.0210473
3.020999
121
{rolls/buns,root vegetables}
{other vegetables}
0.0122013
0.5020921
0.0243010
2.594890
120
{root vegetables,yogurt}
{other vegetables}
0.0129131
0.5000000
0.0258261
2.584078
127
{root vegetables,yogurt}
{whole milk}
0.0145399
0.5629921
0.0258261
2.203354
143
{domestic eggs,other vegetables}
{whole milk}
0.0123030
0.5525114
0.0222674
2.162336
121
{rolls/buns,root vegetables}
{whole milk}
```

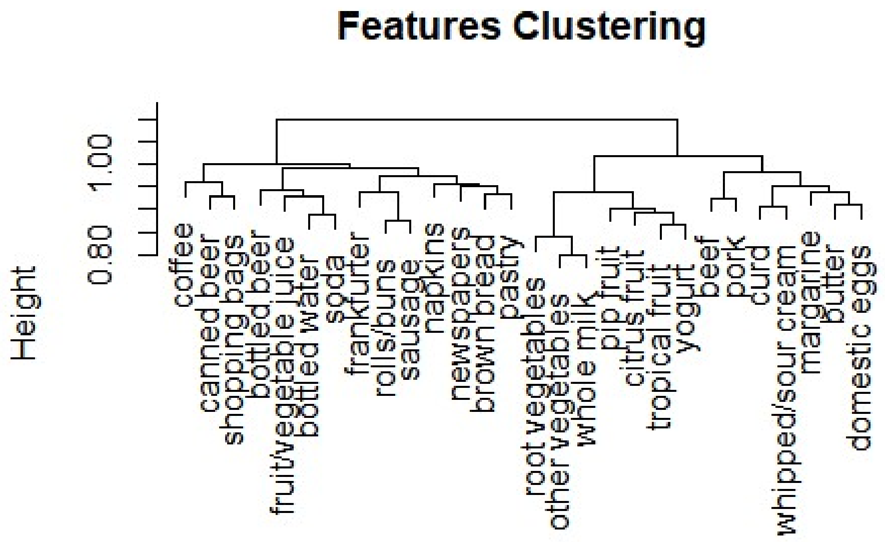
0.0127097
0.5230126
0.0243010
2.046888
125
{other vegetables,pip fruit}
{whole milk}
0.0135231
0.5175097
0.0261312
2.025351
133
{tropical fruit,yogurt}
{whole milk}
0.0151500
0.5173611
0.0292832
2.024770
149
{other vegetables,yogurt}
{whole milk}
0.0222674
0.5128806
0.0434164
2.007235
219
{other vegetables,whipped/sour cream}
{whole milk}
0.0146416
0.5070423
0.0288765
1.984385
144

Cluster Analysis

The basic Clustering graph is generated by using “hclust” from https://www.r-graph-gallery.com/29-basic-dendrogram.html. The graph below alligns with the Top 10 Associative Rules above which indicates Wholemilk and pther vegitables are most important cluster features.

```r
dataframe <- read.transactions("GroceryDataSet.csv", sep=",")

dataframe <- dataframe[ , itemFrequency(dataframe) > 0.05]
d_jaccard <- dissimilarity(dataframe, which = "items")
# plot dendrogram
plot(hclust(d_jaccard, method = "ward.D2"),
     main = "Features Clustering", sub = "", xlab = "")
```


Features Clustering
Height
coffee
canned beer
shopping bags
bottled beer
fruit/vegetable juice
bottled water
soda
frankfurter
rolls/buns
sausage
napkins
newspapers
brown bread
pastry
root vegetables
other vegetables
whole milk
pip fruit
citrus fruit
tropical fruit
yogurt
beef
pork
curd
whipped/sour cream
margarine
butter
domestic eggs