ShovanBiswas-DATA24_Homework_10

Shovan Biswas

11/30/2020

Libraries

```
library(tidyverse)
library(knitr)
library(kableExtra)
library(corrplot)
library(reshape2)
library(Amelia)
library(dlookr)
library(fpp2)
library(plotly)
library(gridExtra)
library(readxl)
library(ggplot2)
library(urca)
library(tseries)
library(AppliedPredictiveModeling)
library(RANN)
library(psych)
library(e1071)
library(corrplot)
library(glmnet)
library(mlbench)
library(caret)
library(earth)
library(randomForest)
library(party)
library(Cubist)
library(gbm)
library(rpart)
library(dplyr)
library(arulesViz)
library(igraph)
```

Problem statement

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. Due May 3 before midnight.

Brief explanation

Initially, I proceeded to read with read_csv. Although I was able to read the usual csv file (GroceryDataSet.csv i.e.), it didn't help in down stream analysis. So, in order to mine the data for **Association Rules**, I googled and learned that apriori() function was required. This is not something, which we customarily use or have used so in the past. On googling, I hit upon the following page:

https://blog.aptitive.com/building-the-transactions-class-for-association-rule-mining-in-r-using-arules-

The page gives an overview of transactions class, apriori() functions etc. The package arules is required, which I added to the list of libraries above. "Market Basket Analysis" was a good clue.
Explanation of some of the terms in Association Rules, which we'll encounter below:

Support of a set of items is the frequency with which, an item appears in the dataset.

Confidence of a rule is the frequency of how often a rule has been found to be true.

Lift is the ratio of the actual support to the expected support.

Reading data and summary

```
# grocery_transactions <- read_csv('./GroceryDataSet.csv')
grocery_transactions <- read.transactions('./GroceryDataSet.csv', sep = ",")
```

```
summary(grocery_transactions)
```

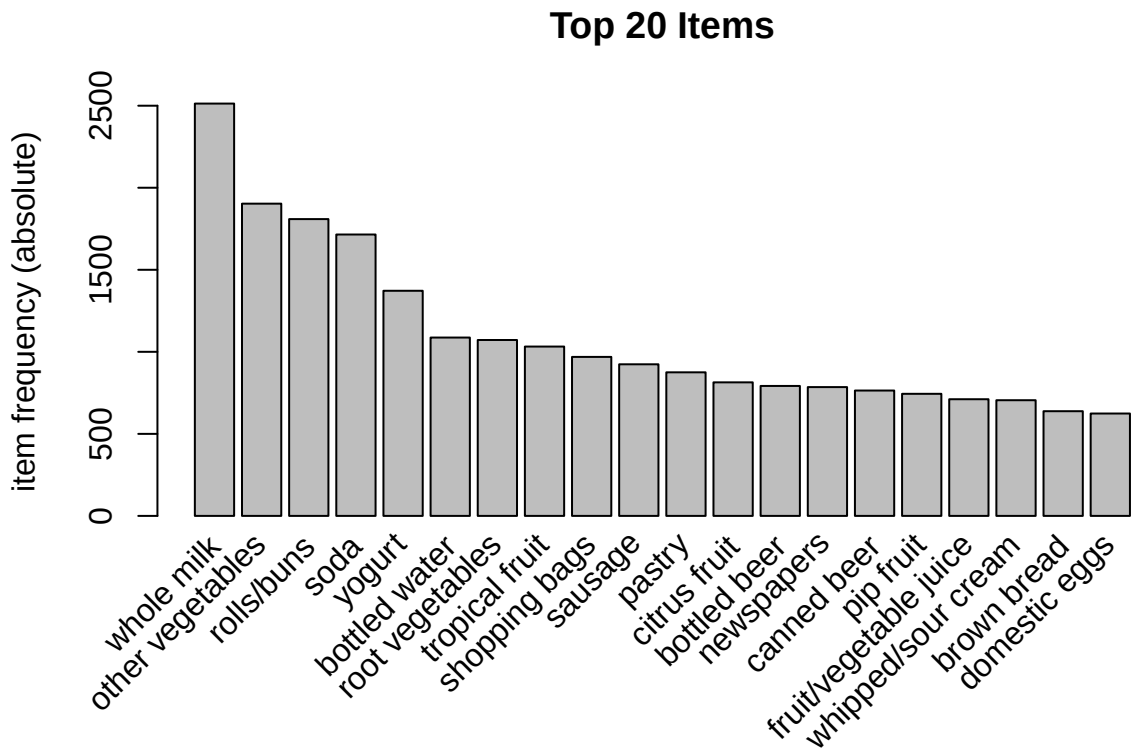
```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
```

```
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

From summary, we see that some of the most freuent items are “whole milk”, “other vegetables”, “rolls/buns”, “soda” etc. In order to get a better visualization, I’ll use function itemFrequencyPlot().

Frequency of top 20 most frequent items

```
itemFrequencyPlot(grocery_transactions, topN = 20, type = "absolute", main = "Top 20 Items")
```



This graph gives an idea of frequencies of top 20 most frequent items. This graph corroborate the few observations in summary.

Further analysis

Now, I'll use apriori() function, for "Market Basket Analysis". I explored apriori() function, by varying the values of the parameters, **support** and **confidence**. With some combinations, I didn't get any results at all – simply errored out. With support = 0.001, confidence = 0.4, in descending order of lift, I got a table (shown down below).

```
support <- 0.001
confidence <- 0.4
rules <- apriori(grocery_transactions, parameter = list(support = support, confidence = confidence), co
```

```
summary(rules)
```

```
## set of 8955 rules
##
## rule length distribution (lhs + rhs):sizes
##    2    3    4    5    6
##   81 2771 4804 1245   54
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   4.000   3.824   4.000   6.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.001017   Min.   :0.4000   Min.   :0.001017   Min.   : 1.565
##  1st Qu.:0.001118   1st Qu.:0.4583   1st Qu.:0.001932   1st Qu.: 2.316
##  Median :0.001322   Median :0.5319   Median :0.002542   Median : 2.870
##  Mean   :0.001811   Mean   :0.5579   Mean   :0.003478   Mean   : 3.191
##  3rd Qu.:0.001830   3rd Qu.:0.6296   3rd Qu.:0.003559   3rd Qu.: 3.733
##  Max.   :0.056024   Max.   :1.0000   Max.   :0.139502   Max.   :21.494
##      count
##  Min.   : 10.00
##  1st Qu.: 11.00
##  Median : 13.00
##  Mean   : 17.81
##  3rd Qu.: 18.00
##  Max.   :551.00
##
## mining info:
##                  data ntransactions support confidence
##  grocery_transactions          9835   0.001        0.4
```

An important observation in summary is, there 8955 rules with length from 2 to 6.

In the following, I'll display the top 10 rules with their support and confidence, sorted descending order of lift.

```
rules %>% DATAFRAME() %>% arrange(desc(lift)) %>% top_n(10) %>% kable()
```

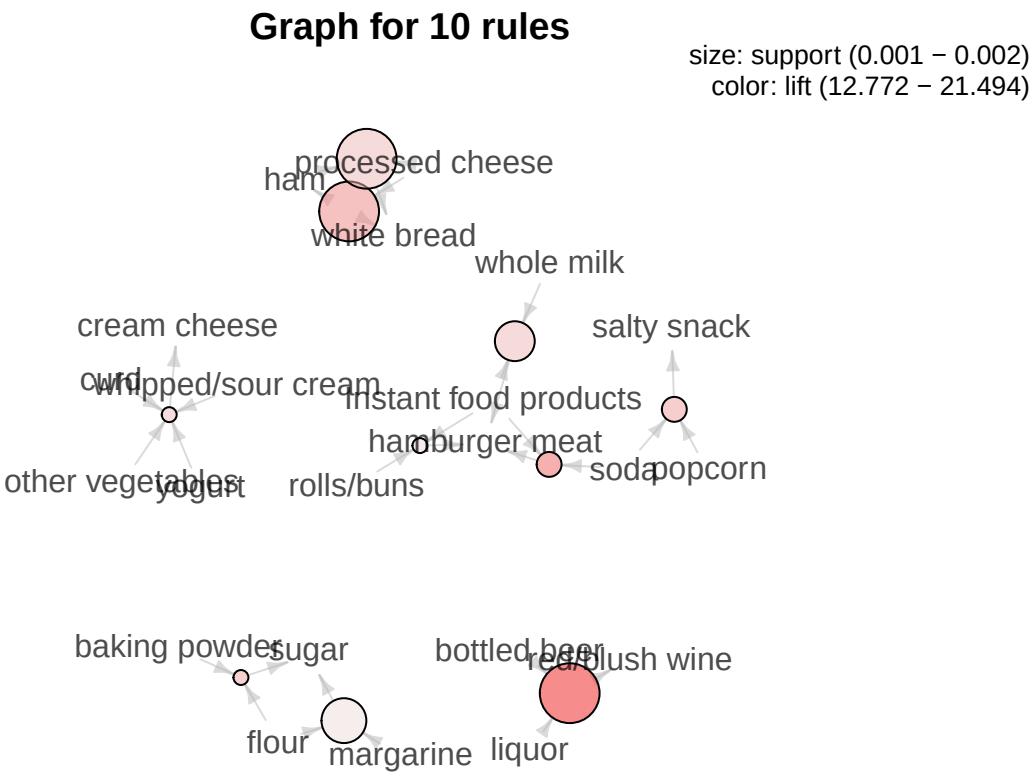
```
## Selecting by count
```

LHS	RHS	support	confidence	coverage	lift	count
{root vegetables}	{other vegetables}	0.0473818	0.4347015	0.1089985	2.246605	466
{whipped/sour cream}	{other vegetables}	0.0288765	0.4028369	0.0716828	2.081924	284
{butter}	{whole milk}	0.0275547	0.4972477	0.0554143	1.946053	271
{curd}	{whole milk}	0.0261312	0.4904580	0.0532791	1.919480	257
{domestic eggs}	{whole milk}	0.0299949	0.4727564	0.0634469	1.850203	295
{whipped/sour cream}	{whole milk}	0.0322318	0.4496454	0.0716828	1.759754	317
{root vegetables}	{whole milk}	0.0489070	0.4486940	0.1089985	1.756031	481
{margarine}	{whole milk}	0.0241993	0.4131944	0.0585663	1.617098	238
{tropical fruit}	{whole milk}	0.0422979	0.4031008	0.1049314	1.577595	416
{yogurt}	{whole milk}	0.0560244	0.4016035	0.1395018	1.571735	551

What is this table telling us? The rule having the greatest lift (2.246605), is for the item **{other vegetables}**, after purchase of **{root vegetables}**. The support and confidence of the item are 0.04738180 and 0.4347015 respectively.

The following graph gives a good visualization of how the items are associating.

```
subrules <- head(rules, n = 10, by = 'lift')
plot(subrules, method = 'graph')
```



Cluster analysis

In order to do cluster analysis, groupings must be identified. After creating a network graph from the given data, I'll use cluster_louvain() to

```
grocery_csv <- read.csv("GroceryDataSet.csv", header = FALSE) %>% mutate(shoper_id = row_number()) %>% p

communities <- grocery_csv %>% rename(to = value, from = shoper_id) %>% graph_from_data_frame(directed =
```

The following step will associate customers and items to 19 clusters.

```
products <- as.character(unique(grocery_csv$value))

df <- data.frame(name = c(NA), members = c(NA)) %>% na.omit() # create data frame

for (i in 1:length(communities)){
  cluster_name <- paste0(i,": ")
  cluster_members <- 0
  for (member in communities[[i]]){
    if (member %in% products){
      cluster_name <- paste0(cluster_name, member, " + ")
    } else {
      cluster_members <- cluster_members + 1
    }
  }
  cluster_name <- substr(cluster_name,1,nchar(cluster_name)-3)
  df <- rbind(df, data.frame(name = cluster_name, members = cluster_members))
}
  df %>%
  arrange(desc(members)) %>% kable()
```

name
8: chocolate + soda + specialty bar + pastry + salty snack + waffles + candy + dessert + chocolate marshmallow + spec
10: other vegetables + rice + abrasive cleaner + flour + beef + chicken + root vegetables + bathroom cleaner + spices +
12: ready soups + rolls/buns + frankfurter + sausage + spread cheese + hard cheese + canned fish + seasonal products +
13: whole milk + butter + cereals + curd + detergent + hamburger meat + flower (seeds) + canned vegetables + pasta +
5: liquor (appetizer) + canned beer + shopping bags + misc. beverages + chewing gum + brandy + liqueur + whisky
7: yogurt + cream cheese + meat spreads + packaged fruit/vegetables + butter milk + berries + whipped/sour cream +
4: tropical fruit + pip fruit + white bread + processed cheese + sweet spreads + beverages + ham + cookware + tea + s
15: citrus fruit + hygiene articles + domestic eggs + cat food + cling film/bags + canned fruit + dental care + flower soil
16: bottled beer + red/blush wine + prosecco + liquor + rum
11: UHT-milk + bottled water + white wine + male cosmetics
2: long life bakery product + pot plants + fruit/vegetable juice + pickled vegetables + jam + bags
3: semi-finished bread + newspapers + pet care + nuts/prunes + toilet cleaner
6: dishes + napkins + grapes + zwieback + decalcifier
1: coffee + condensed milk + sparkling wine + fish + kitchen towels
18: sugar + frozen vegetables + salt + skin care + liver loaf + frozen chicken
14: frozen dessert + ice cream + frozen meals
9: margarine + artif. sweetener + specialty fat + candles + organic products
17: brown bread + sauces
19: photo/film