Homework Week 6

Yusang Hu 400177333

Oct 24 2019

Introduction

Groceries is a data set contains 30 days transaction records from a typical local grocery outlet. The data set is provided by Michael Hahsler, Kurt Hornik and Thomas Reutterer (2006).The data set contains 9835 transactions, one row represents one transaction and columns represent the goods which are purchased in specific transaction,and 169 kinds of goods (variables).

In a supermarket, deliberate product placements would not only saves customers' shopping time but also helps to increase sales for some products. Studying associations between each items is helpful to classify each items into suited aisle. In this assignment, I'll study which items can be classified into same aisle with 'whole milk'. Note that I'll only study top 20 items and top 20 association rules in this assignment since there are too many item types.

Histogram

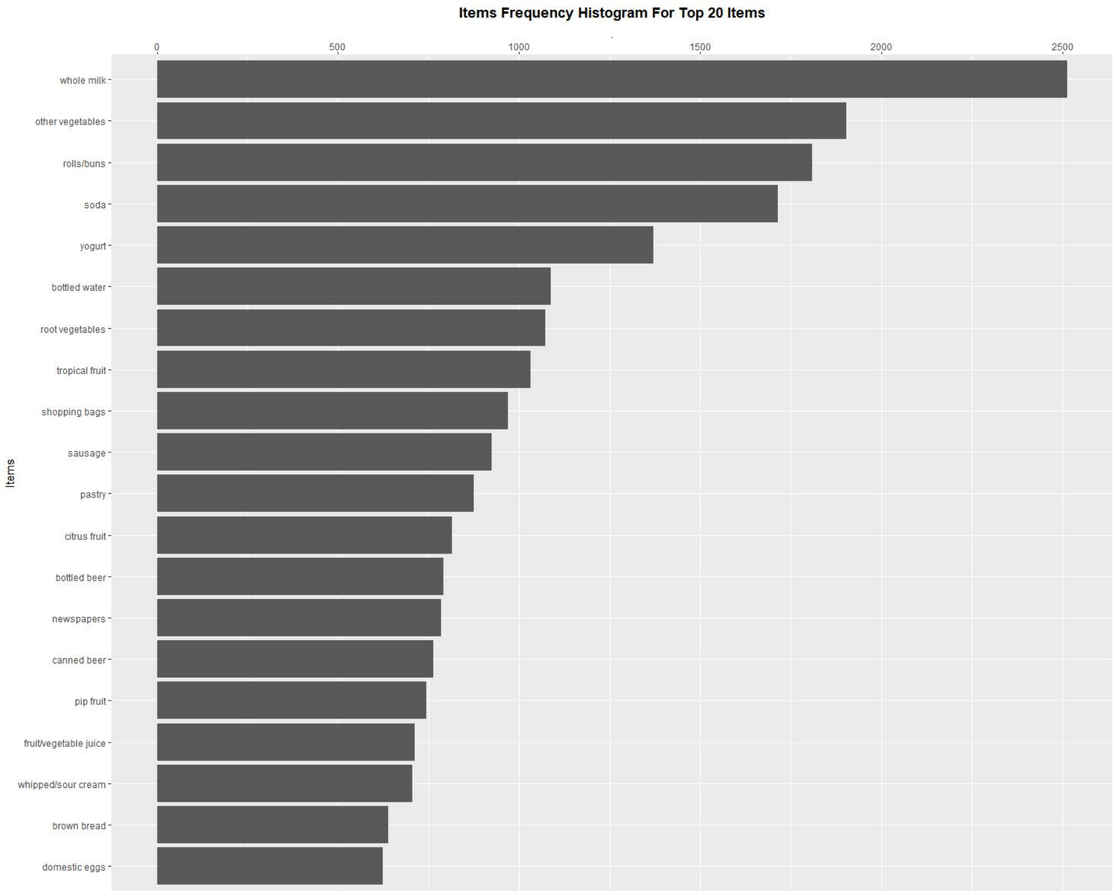
First, I'll draw a histogram to display the frequency of top 20 items. I choose histogram because it helps us to visualize the distribution of our data, and what I'm interested in is studying the most frequent item and the least frequent item.

The histogram tells us that: 1. 'whole milk' is the best-seller in 30 days, and 'domestic eggs' has the worst sales in top 20 items in 30 days. 2. Since I'm going to study which items can be classified into same aisle with 'whole milk', and 'whole milk' is the best-seller, therefore we can change the layout of supermarket by putting some bad sales items (i.e. domestic eggs, brown bread) with whole milk to increase sales volume of those bad sales items.

```
library(ggplot2)
library(arules)

## Loading required package: Matrix

##
## Attaching package: 'arules'

## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
rm(list=ls())
#setwd("/Users/admin/Desktop/744")
groceries<-read.transactions("https://raw.githubusercontent.com/huyusan
g1031/Stat744/master/groceries.csv",sep=",")
itemFrequencyGGPlot <- function(x, topN) {
  library(tidyverse)
  x %>%
    itemFrequency(type="absolute") %>%
    sort %>%
    tail(topN) %>%
    as.data.frame %>%
    tibble::rownames_to_column() %>%
    ggplot(aes(reorder(rowname, `.`),`.`)) +
    geom_col() +
    coord_flip() + ## flip plot 90 degrees, this will make the ranking o
f frequency more clearer
    scale_y_continuous(position="right") + ## move y-axis to the top
    labs(x="Items") +
    ggtitle("Items Frequency Histogram For Top 20 Items") + ## add title
    theme(plot.title=element_text(face="bold",hjust=0.5)) ## change layo
ut of title
}
itemFrequencyGGPlot(groceries, 20)
```

```
## -- Attaching packages --------------------------------------------------
----------------------------- tidyverse 1.2.1 --

## √ tibble  2.1.3     √ purrr   0.3.2
## √ tidyr   1.0.0     √ dplyr   0.8.3
## √ readr   1.3.1     √ stringr 1.4.0
## √ tibble  2.1.3     √ forcats 0.4.0

## -- Conflicts -------------------------------------------------------------
------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::recode() masks arules::recode()
## x tidyr::unpack() masks Matrix::unpack()
```



Parallel Coordinate Plot

The second plot I'll draw is a parallel coordinate plot. It is a plot based on the association rules between each items. I picked top 20 association rules where their support is larger than 0.006 and confidence is larger than 0.5.

I choose to use parallel coordinate plot because the correlations between items can be spotted easily. What I'm interested in is studying associations between items, for example, the topmost line shows us that when I have yogurt and frankfurter in my shopping cart, I'm highly likely to buy whole milk as well.

The parallel coordinate plot tell us that: 1. Since I'm going to study which items can be classified into the same aisle with 'whole milk', I'll concentrate on finding the item that have association with whole milk. The line starts with whole milk shows us that when I have whole milk and onions in my shopping cart, I'm highly likely to buy other vegetables as well.

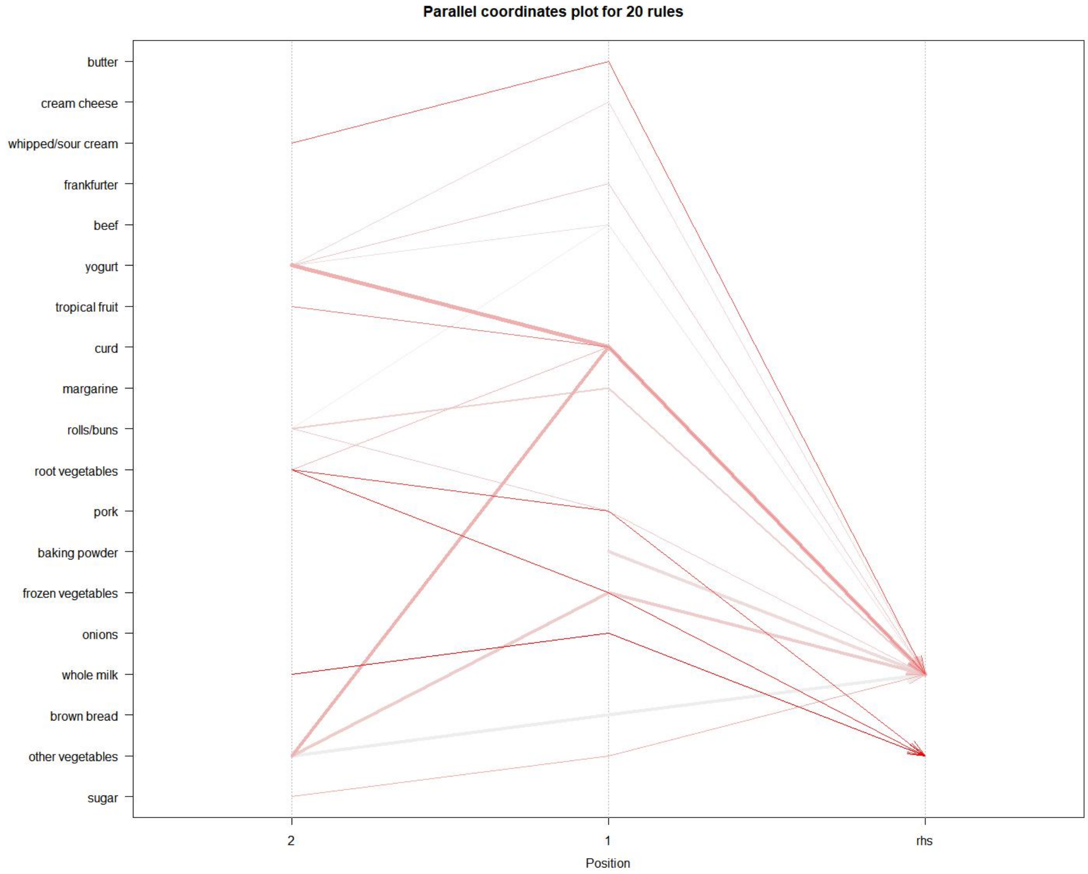
```
library(arulesViz)
```

```
## Loading required package: grid
## Registered S3 method overwritten by 'seriation':
##   method         from 
##   reorder.hclust gclus
```

```r
params<-list(support=0.006,confidence=0.5)
fit<-apriori(groceries,parameter=params)
```

```
## Apriori
## 
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support mi
nlen
##         0.5    0.1    1 none FALSE            TRUE       5   0.006      
1
##  maxlen target   ext
##      10  rules FALSE
## 
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
## 
## Absolute minimum support count: 59 
## 
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [109 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [67 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```r
plot(fit[1:20],method="paracoord",control=list(reorder=TRUE))
```



Conclusion

From the histogram and the parallel coordinate plot, I'll conclude that I'll put whole milk, vegetables and eggs in same aisle.

Rubric

Histogram: I choose linear scale because differences between each items can easily be measured with the help of linear scale; I choose anchored at zero, because I'm interested in which item has highest frequency and lowest frequency, I'm not interested in small differences between each bar. I flipped the histogram over because x-axis labels are hard to read in default plot (i.e. items' name are vertically). My categorical variables are ordered from highest frequency to lowest frequency, so it's sensible.

Parallel Coordinate Plot: The x-axis scale is special in this plot, it is neither linear, log, logit or sqrt, instead, it represents the position of each association rule, where '2' and '1' means the left hand side of rules and 'rhs' means the right hand side of rules. I didn't reorder y-axis labels because I'm not interested in the ranking of items.

Reference

Hahsler, M., Hornik, K., & Reutterer, T. (n.d.). Implications of Probabilistic Data Modeling for Mining Association Rules. From Data and Information Analysis to Knowledge Engineering Studies in Classification, Data Analysis, and Knowledge Organization, 598–605. doi: 10.1007/3-540-31314-1_73