Association Rule Mining for Market Basket Analysis

1. Introduction

A store is interested in determining the associations between items purchased from the Health and Beauty Aids department and the Stationery Department. The store chose to conduct a market basket analysis of specific items purchased from these two departments. TRANSACTIONS contain information about over 400,000 transactions made over the past three months. The following 17 products are represented in the data set: bar soap, bows, candy bars, deodorant, greeting cards, magazines, markers, pain relievers, pencils, pens, perfume, photo processing, prescription medications, shampoo, toothbrushes, toothpaste, and wrapping paper.

There are four variables in the data set:

Name	Model Role	Data Type	Description
STORE	Ignore	Numeric	Identification number of the store
TRANSACTION	Ident	Numeric	Transaction identification number
PRODUCT	Target	Categorical	Product purchased
QUANTITY	Ignore	Numeric	Quantity of this product purchased

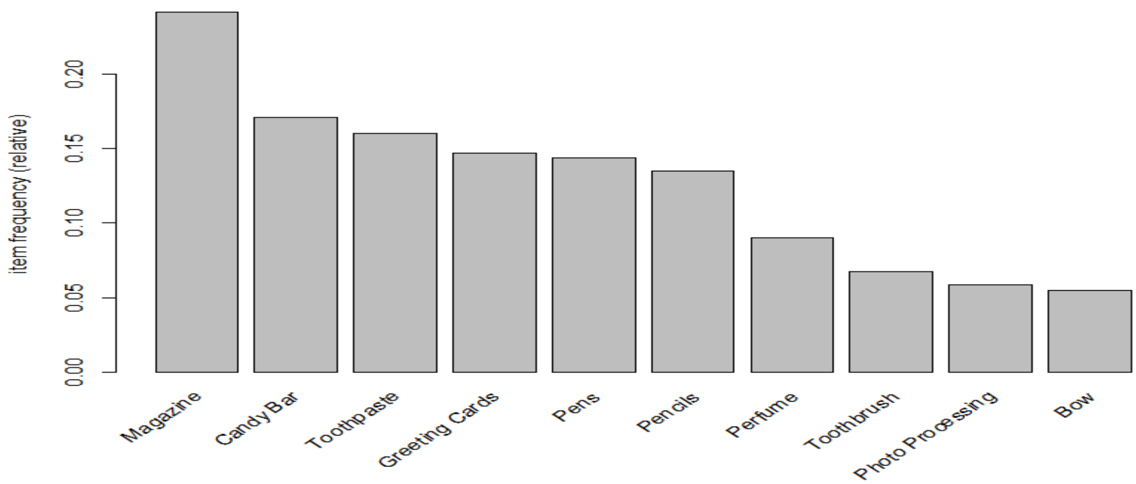
Data File :

transactions.csv

2. Data Analysis

```
> # read the data and create transaction objects
> transaction <- read.transactions("transactions.csv", format = "single", sep = ",", cols = c(
"Transaction", "Product"), rm.duplicates = FALSE)
> View(transaction)
>
> #item frequency bar plot for inspecting the item frequency distribution for objects based on
 rules
> itemFrequencyPlot(transaction,topN=10,type="relative")
>
```

Below is the relative item frequency histogram plot of the top 10 products in the transaction data set.



As per the above histogram **Magazine** is the highly purchased product.

- Below is an application of **Apriori Algorithm** to find the association rules between the products

```
> # find the association rules
> rules <- apriori(transaction, parameter = list(supp = 0.03, conf = 0.20, minlen = 2, maxlen
= 4))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target
        0.2    0.1    1 none FALSE            TRUE       5    0.03      2      4  rules
   ext
 FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 6000

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[17 item(s), 200000 transaction(s)] done [0.02s].
sorting and recoding items ... [13 item(s)] done [0.00s].
creating transaction tree ... done [0.04s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [9 rule(s)] done [0.00s].
creating S4 object  ... done [0.01s].
> # sort the rules
> rules <- sort(rules, by="lift", decreasing=TRUE)
>
> # find no. of rules
> rules
set of 9 rules
>
```

- There are set of 9 rules which are governing association of frequently bought products together.

```
> # Remove duplicate rules
> redundant_index <- is.redundant(rules)
> pruned_rules <- rules[!redundant_index]
> # find summary
> summary(pruned_rules)
set of 9 rules

rule length distribution (lhs + rhs):sizes
2
9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.0316   Min.   :0.218   Min.   :0.972   Min.   :6326
 1st Qu.:0.0330   1st Qu.:0.234   1st Qu.:1.025   1st Qu.:6603
 Median :0.0398   Median :0.245   Median :1.431   Median :7956
 Mean   :0.0378   Mean   :0.246   Mean   :1.350   Mean   :7566
 3rd Qu.:0.0405   3rd Qu.:0.248   3rd Qu.:1.450   3rd Qu.:8107
 Max.   :0.0437   Max.   :0.297   Max.   :1.738   Max.   :8732

mining info:
        data ntransactions support confidence
 transaction        200000    0.03        0.2
> # summary displays only 9 rules so inspecting 9 rules
> inspect(pruned_rules[1:9])
    lhs                 rhs              support confidence lift   count
[1] {Greeting Cards} => {Candy Bar}      0.04366 0.2972     1.7382 8732
[2] {Candy Bar}      => {Greeting Cards} 0.04366 0.2553     1.7382 8732
[3] {Toothpaste}     => {Candy Bar}      0.03978 0.2480     1.4501 7956
[4] {Candy Bar}      => {Toothpaste}     0.03978 0.2326     1.4501 7956
[5] {Pencils}        => {Candy Bar}      0.03302 0.2447     1.4309 6603
[6] {Greeting Cards} => {Toothpaste}     0.03208 0.2184     1.3614 6416
[7] {Greeting Cards} => {Magazine}       0.03633 0.2474     1.0251 7267
[8] {Candy Bar}      => {Magazine}       0.04054 0.2370     0.9823 8107
[9] {Pencils}        => {Magazine}       0.03163 0.2344     0.9715 6326
```

2.1. Lift Ratio –

- Lift is a measure of significance of a rule. Lift provides information about the increase in probability of the consequent, given the antecedent. I.e., does including the antecedent improve the probability of finding the consequent over random chance.
- If some rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.
- If the lift is > 1, then that rule lets us know the degree to which those two occurrences are dependent on one another and makes those rules potentially useful for predicting the consequent in future data sets.
- Lift is calculated as:

  Lift ratio = confidence/ Benchmark Confidence

  Benchmark confidence = No. of transactions with consequent items/ No. of transactions in database

As per the dataset, association rule involving **Greeting Card & Candy Bar** has the highest lift ratio of 1.7382.

2.2. Analysis of Top 5 Rules

Below is the list of top 5 association rules :

```
    lhs                   rhs               support confidence lift   count
[1] {Greeting Cards} => {Candy Bar}       0.04366 0.2972     1.7382 8732
[2] {Candy Bar}      => {Greeting Cards} 0.04366 0.2553     1.7382 8732
[3] {Toothpaste}     => {Candy Bar}       0.03978 0.2480     1.4501 7956
[4] {Candy Bar}      => {Toothpaste}      0.03978 0.2326     1.4501 7956
[5] {Pencils}        => {Candy Bar}       0.03302 0.2447     1.4309 6603
```

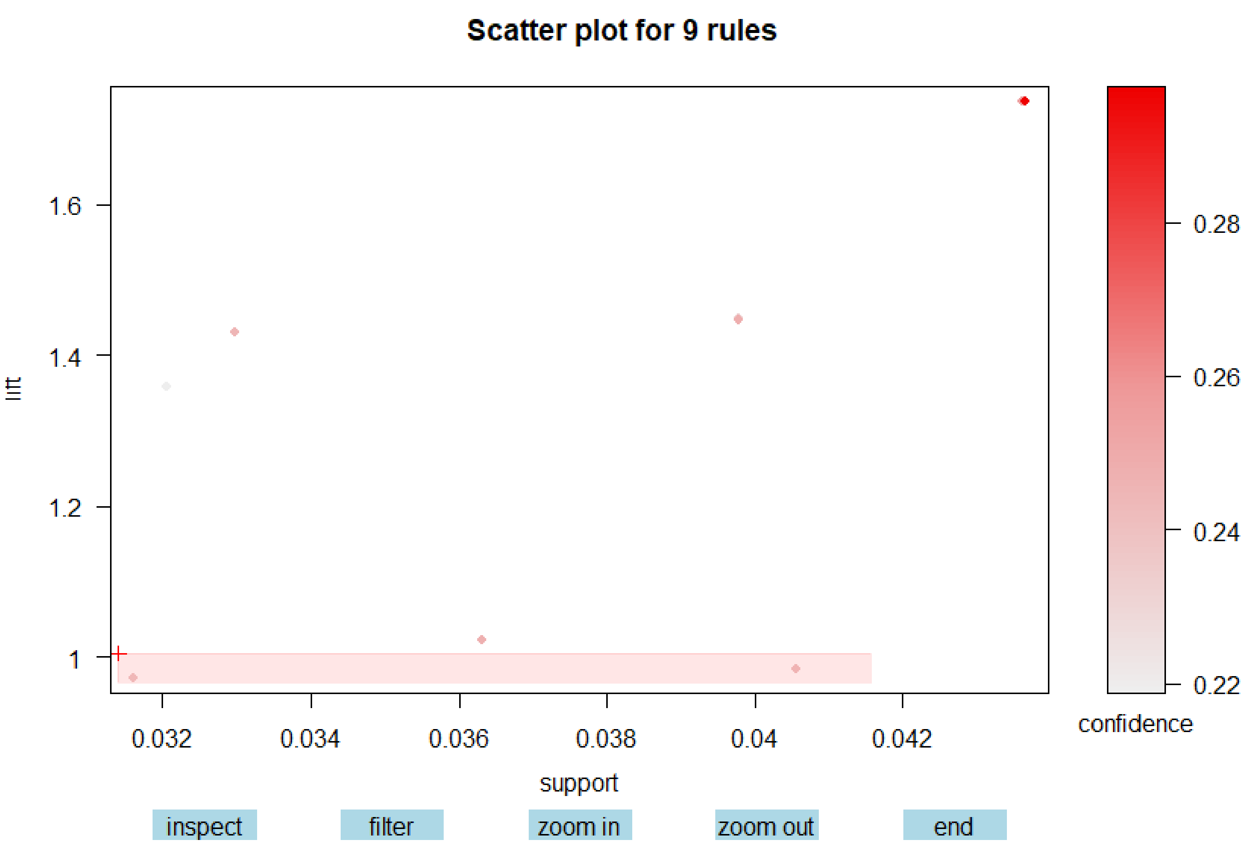
The Apriori Algorithm of Association Rule mining is used to identify the frequently bought items. After applying this algorithm on “transaction” data set, we can infer that If a customer buys a **Greeting Card** then there is a strong probability of buying **Candy Bar** and vice versa. Followed by the first and second rule, the third and fourth rule conveys that if a customer buys **Toothpaste** then he is likely to buy a **Candy Bar** and vice versa. Also, from the 5^{th} rule, we can infer that if a customer buys **Pencils** then he is likely to buy **Candy Bar.**

2.3. <u>Non-significant Rules</u> –

Since Lift ratio is parameter for significance for the rule, having lift >1 makes the rule significant. There are two rules which have lift ratio less than 1 as shown in the below code execution and lift vs support plot.

```
> plot(rules, measure=c("support", "lift"), shading = "confidence",interactive = TRUE)
To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
Interactive mode.
Select a region with two clicks!

Number of rules selected: 2
    lhs             rhs        support confidence lift   count order
[1] {Candy Bar} => {Magazine} 0.04054 0.2370     0.9823 8107  2
[2] {Pencils}   => {Magazine} 0.03163 0.2344     0.9715 6326  2
```



3. <u>Insights</u>

- As per the item-Frequency-Plot, we came to know that **Magazine** is the highly purchased product by the customers.
- But as per the Aprioiri Algorithm, it is evident that the association of Magazine is not strong with other products bought by the customers.
- **Greeting Card** and **Candy Bar** are the top two products that have been bought together.