Market Basket Analysis

AUTHOR
Bala Shunmugam M

Introduction:

Market Basket Analysis (MBA) is a powerful technique used to uncover associations between products based on customer purchasing behavior. By analyzing transaction data, businesses can identify which items are frequently bought together, enabling smarter decisions in promotions, product placement, and inventory management.

In this analysis, we use transactional data from a grocery dataset to discover frequent item combinations and generate actionable association rules using the Apriori algorithm.

```
data <- read.csv("D:/IBA Files/Basket Analysis/Groceries_dataset.csv",header = TRUE)
```

1. Data Preparation:

To begin, we load the dataset and prepare it for analysis. Each row in the dataset represents a single item purchased by a customer on a specific date. To understand co-purchase behavior, we group items by transaction using a unique identifier that combines `Member_number` and `Date`.

This transformation allows us to analyze which items were bought together in the same shopping trip, forming the foundation for Market Basket Analysis.

```
# Step 2: Create a unique Transaction ID
data$TransactionID <- paste(data$Member_number, data$Date, sep = "_")
# Step 3: Keep only TransactionID and itemDescription
basket_data <- data[, c("TransactionID", "itemDescription")]
# Step 4: Save to CSV for arules
write.csv(basket_data, "transactions.csv", row.names = FALSE)
# Step 5: Read transactions in 'single' format
trans <- read.transactions("transactions.csv", format = "single",
                           cols = c("TransactionID", "itemDescription"),
                           sep = ",", header = TRUE, rm.duplicates = TRUE)
```

2. Transaction Overview:

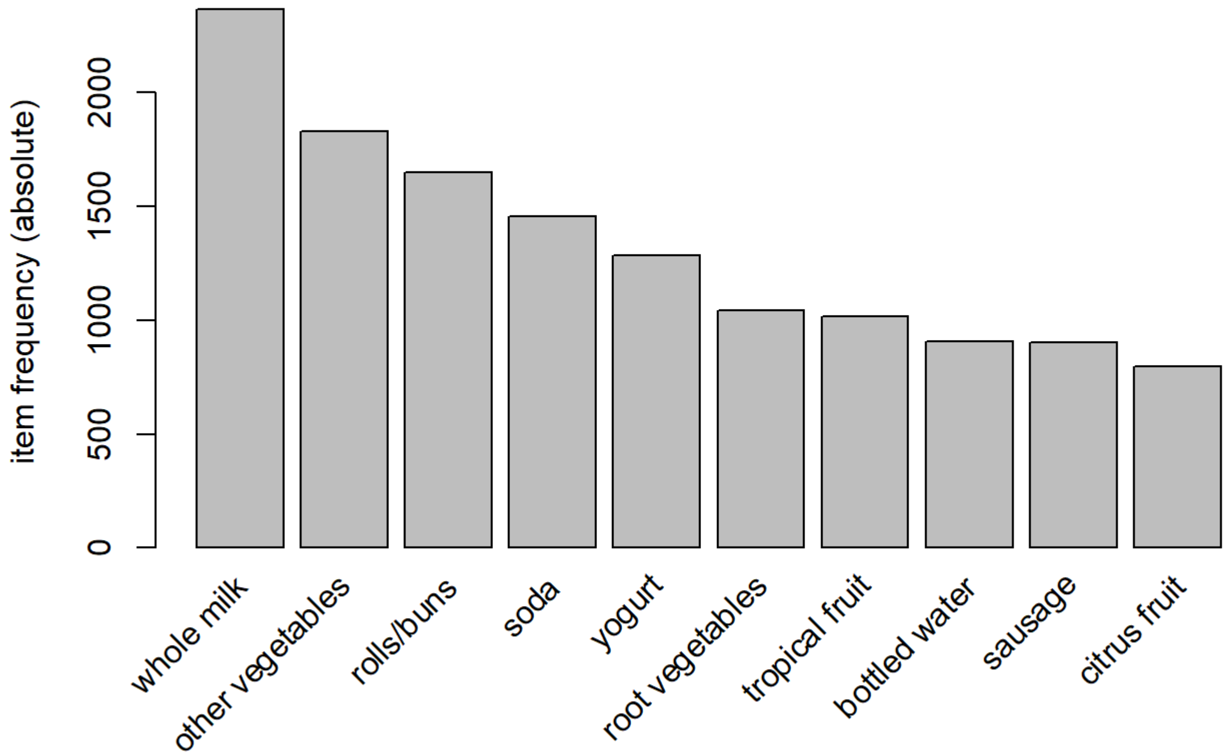
Before mining rules, we explore the transaction data to understand its structure and item frequency. The summary provides insights into the number of transactions, unique items, and average basket size.

The item frequency plot highlights the most commonly purchased products, helping us identify popular items that may drive associations.

```
# Step 6: Explore the transaction data
summary(trans)
```

```
transactions as itemMatrix in sparse format with
 14963 rows (elements/itemsets/transactions) and
 167 columns (items) and a density of 0.01520957

most frequent items:
      whole milk other vegetables       rolls/buns             soda
            2363             1827             1646             1453
          yogurt          (Other)
            1285            29432

element (itemset/transaction) length distribution:
sizes
    1     2     3     4     5     6     7     8     9    10
  205 10012  2727  1273   338   179   113    96    19     1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    2.00    2.54    3.00   10.00

includes extended item information - examples:
            labels
1 abrasive cleaner
2 artif. sweetener
3   baby cosmetics

includes extended transaction information - examples:
    transactionID
1 1000_15-03-2015
2 1000_24-06-2014
3 1000_24-07-2015
```

```r
itemFrequencyPlot(trans, topN = 10, type = "absolute")
```



3. Rule Mining with Apriori:

Using the Apriori algorithm, we generate association rules based on specified thresholds for support and confidence. Support measures how frequently an itemset appears in the dataset, while confidence indicates the likelihood of purchasing one item given another.

Lower thresholds are used to ensure that even less frequent but potentially valuable associations are captured. This step reveals patterns in customer behavior that can inform marketing and merchandising strategies.

```
# Step 7: Run Apriori algorithm
rules <- apriori(trans, parameter = list(supp = 0.0005, conf = 0.1))
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.1    0.1    1 none FALSE            TRUE       5   5e-04      1
 maxlen target  ext
     10  rules TRUE

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

```
Absolute minimum support count: 7

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[167 item(s), 14963 transaction(s)] done [0.00s].
sorting and recoding items ... [158 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [298 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

4. Rule Intrepretation:

The generated rules are sorted by lift, which measures the strength of association beyond random chance. A lift value greater than 1 indicates a positive correlation between items.

By inspecting the top rules, we gain insights into which products are strongly linked in customer baskets. These findings can be used to design product bundles, cross-promotions, and personalized recommendations.

```
# Step 8: Inspect top rules
summary(rules)
```

```
set of 298 rules

rule length distribution (lhs + rhs):sizes
  1   2   3 
  3 176 119 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.000   2.000   2.000   2.389   3.000   3.000 

summary of quality measures:
    support            confidence        coverage             lift       
 Min.   :0.0005347   Min.   :0.1000   Min.   :0.001537   Min.   :0.6458  
 1st Qu.:0.0006015   1st Qu.:0.1122   1st Qu.:0.004762   1st Qu.:0.8495  
 Median :0.0009356   Median :0.1259   Median :0.006483   Median :0.9859  
 Mean   :0.0029623   Mean   :0.1363   Mean   :0.023191   Mean   :1.0931  
 3rd Qu.:0.0017877   3rd Qu.:0.1481   3rd Qu.:0.014853   3rd Qu.:1.2474  
 Max.   :0.1579229   Max.   :0.3913   Max.   :1.000000   Max.   :2.4778  
     count        
 Min.   :   8.00  
 1st Qu.:   9.00  
 Median :  14.00  
 Mean   :  44.33  
 3rd Qu.:  26.75  
 Max.   :2363.00  

mining info:
  data ntransactions support confidence
 trans         14963   5e-04        0.1
                                                                    call
 apriori(data = trans, parameter = list(supp = 5e-04, conf = 0.1))
```

```r
inspect(sort(rules, by = "lift")[1:10])
```

```
     lhs                                rhs                support
[1]  {pork, sausage}                 => {whole milk}       0.0006014837
[2]  {sweet spreads}                 => {pip fruit}        0.0005346521
[3]  {sweet spreads}                 => {tropical fruit}   0.0007351467
[4]  {rolls/buns, whipped/sour cream} => {yogurt}          0.0006014837
[5]  {spices}                        => {soda}             0.0006014837
[6]  {sausage, shopping bags}        => {other vegetables} 0.0005346521
[7]  {whole milk, yogurt}            => {sausage}          0.0014702934
[8]  {pastry, soda}                  => {sausage}          0.0005346521
[9]  {brandy}                        => {whole milk}       0.0008688097
[10] {pork, whole milk}              => {sausage}          0.0006014837
     confidence coverage    lift     count
[1]  0.3913043  0.001537125 2.477819  9
[2]  0.1176471  0.004544543 2.398301  8
[3]  0.1617647  0.004544543 2.387066 11
[4]  0.2045455  0.002940587 2.381800  9
[5]  0.2250000  0.002673261 2.317051  9
[6]  0.2758621  0.001938114 2.259291  8
[7]  0.1317365  0.011160863 2.182917 22
[8]  0.1311475  0.004076723 2.173157  8
[9]  0.3421053  0.002539598 2.166281 13
[10] 0.1200000  0.005012364 1.988439  9
```

5. Visualization Association:

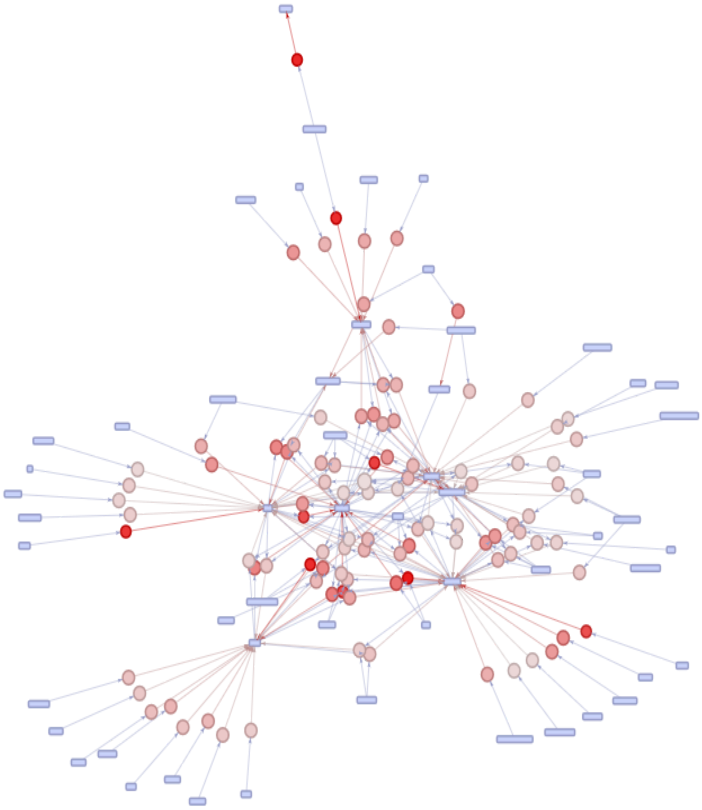
To enhance interpretability, we visualize the association rules using a network graph. This representation shows how items are connected, with stronger relationships appearing as thicker or more central links.

Visualization helps stakeholders quickly grasp the structure of item associations and identify key products that influence purchasing behavior.

```r
# Step 9: Optional - Visualize rules
plot(rules, method = "graph", engine = "htmlwidget")
```

```
Warning: Too many rules supplied. Only plotting the best 100 using 'lift'
(change control parameter max if needed).
```

Select by id



Conclusion:

This Market Basket Analysis has revealed valuable insights into customer purchasing behavior by identifying frequent item combinations and strong associations between products. Through the Apriori algorithm and rule visualization, we've uncovered patterns that go beyond individual product popularity —highlighting how items interact within the context of a shopping basket.

The analysis demonstrates that even low-frequency items can play a strategic role when paired with high-confidence associations. These findings provide a data-driven foundation for enhancing customer experience and driving business growth.

About the Author

I'm a student in the MBA Program (2025–2027), currently navigating Trimester I at Amrita School of Business, Amrita Vishwa Vidyapeetham, Coimbatore. This **assignment/blog** was written as part of our coursework for **Introduction to Business Analytics.**

Blog link: https://github.com/Bala-Shunmugam-M/Bala-Shunmugam-M