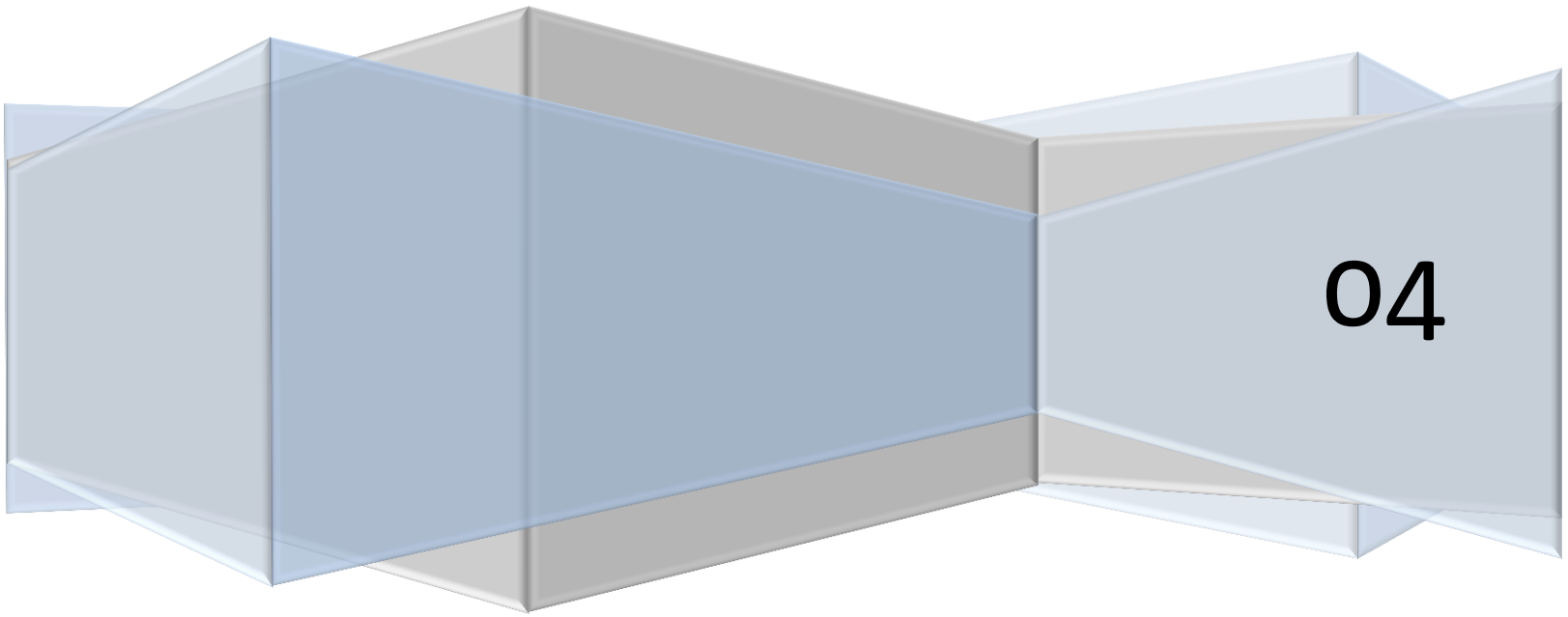
DnI Consulting

DnI

Market Basket Analysis

Step by step approach using R

DnI Consulting



DnI

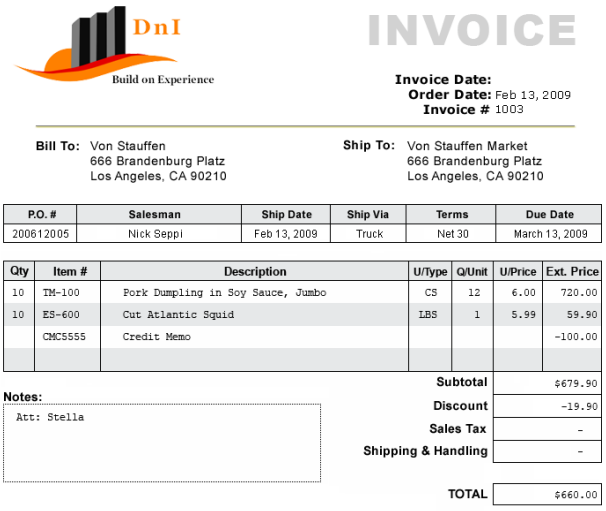
Market Basket Analysis

Objectives

Leverage customer transaction data for right product bundling and promotions, assortment planning and inventory management, and product placement in the stores.

Data Preparation

Main data source used for a Market Basket Analysis is customer purchase transaction data. The purchase slip or bill will have information on products purchased on a customer visit along with their quantities, prices and overall prices.



DnI
Build on Experience

INVOICE

Invoice Date:
Order Date: Feb 13, 2009
Invoice # 1003

Bill To: Von Stauffen
666 Brandenburg Platz
Los Angeles, CA 90210

Ship To: Von Stauffen Market
666 Brandenburg Platz
Los Angeles, CA 90210

P.O. #	Salesman	Ship Date	Ship Via	Terms	Due Date
200612005	Nick Seppi	Feb 13, 2009	Truck	Net 30	March 13, 2009

Qty	Item #	Description	U/Type	Q/Unit	U/Price	Ext. Price
10	TM-100	Pork Dumpling in Soy Sauce, Jumbo	CS	12	6.00	720.00
10	ES-600	Cut Atlantic Squid	LBS	1	5.99	59.90
	CMC5555	Credit Memo				-100.00

Subtotal	$679.90
Discount	-19.90
Sales Tax	-
Shipping & Handling	-
TOTAL	$660.00

Notes:
Att: Stella

The transaction table may store information as follow

OrderID	Transaction Date	Product ID	Product Description	Quantity Purchased	Unit Price	Price
11	1-Jan-14	23	Colgate 50gm	2	12	24
11	1-Jan-14	73	Modern Bread	1	30	30
12	1-Jan-14	23	Colgate 50gm	1	12	12
12	1-Jan-14	55	Pepsodent Tooth Brush	1	17	17
12	1-Jan-14	87	Cadbury Chocolate	1	21	21

From above data warehouse table, we need to get data by order/visit. Dataset from real life retail [1] example is used and the sample of data is as follow

OrderID	ProductCodeList
[[1]]	"0" "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11" "12" "13" "14" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29"
[[2]]	"30" "31" "32"
[[3]]	"33" "34" "35"
[[4]]	"36" "37" "38" "39" "40" "41" "42" "43" "44" "45" "46"
[[5]]	"38" "39" "47" "48"
[[6]]	"38" "39" "48" "49" "50" "51" "52" "53" "54" "55" "56" "57" "58"

Data Analysis

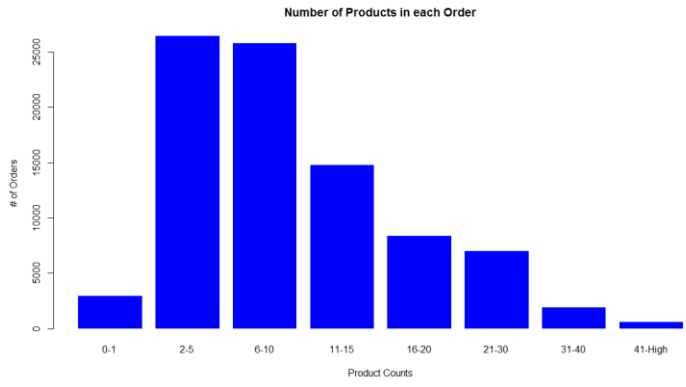
Once we have data in required format, we need to carry out univariate or exploratory analysis, so that we understand what is going on.

Some of the typical questions we will try to answer based on Market Basket Analysis are

- What are the distinct visits?
- What is typical number of products purchased by a customer in an order or a visit?
- What number of different SKUs (stock keeping units) being sold in a week or month?
- Which are the most frequent items or products?

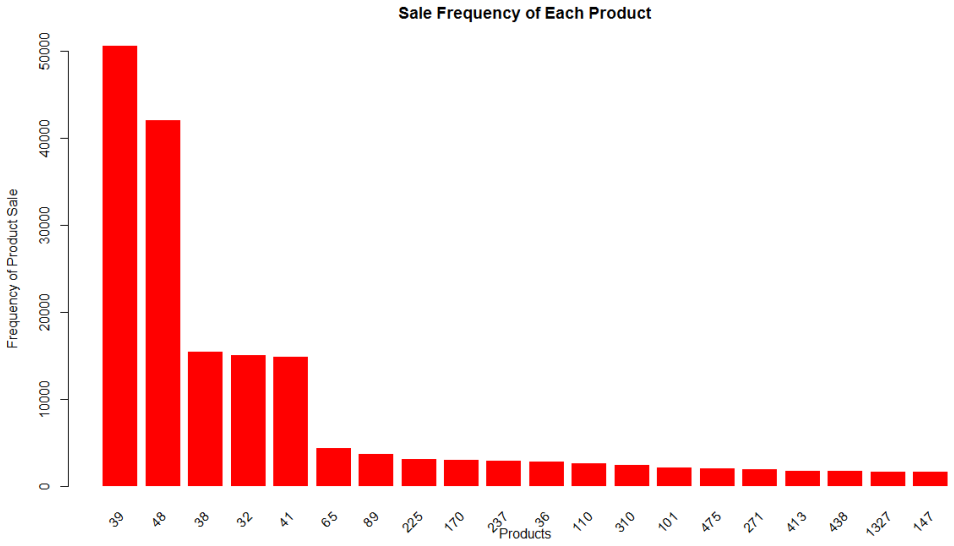
We will try to answer these Market Basket Data Analysis questions using sample dataset and R.

# of Products in Order	% of Orders
0-1	3%
2-5	30%
6-10	29%
11-15	17%
16-20	10%
21-30	8%
31-40	2%
41-High'	1%



Based on above analysis, 60% of the orders or visits have between 2 to 10 products in an order. A next important question really is who are frequently bought products in customer baskets.

We can use R functions – itemFrequencyPlot() – to get count and plot of frequent bought products. This function is part of R package – "arules". We should install this package before using this.



Product – 39 and 48 are the most frequently purchased products. This will help us to confirm if these are as per expectations.

Market Basket Analysis and Affinity Analysis

Post data preparation and exploratory analysis, we can shift to main analysis targeted toward Market Basket Analysis (MBA).

Some of the key questions Market Basket Analysis (MBA) tries to answer are

- Should we perform market basket analysis at a product level or category level?
- Do we have information on sequence of products buying in a basket or customer visit?
- Which are products bought together by the customers?
- Can we conclude if product 'A' sale drives product 'B' sales?
- What product categories are bought together?
- What product is to be recommended given a customer has bought a product or a group of products?

Steps used in Market Basket Analysis

- Identify Rules

  Association Rules or Affinity between products bought together need to be identified based on transactional data.

R package – "arules" is used to find rules. The rule can be identified and filtered based on product combinations. We can have rule for group of 2 products, 3 products or more products.

Example of rules

lhs rhs	support	confidence	lift
{3854} => {38}	0.001	0.913	5.159
{1045} => {32}	0.001	0.907	5.270
{4030} => {48}	0.001	0.826	1.728
{1473} => {39}	0.001	0.800	1.392
{1727} => {38}	0.002	0.931	5.263

Lhs (left hand side) indicates first product or item considered for the rule

Rhs (right hand side) indicates second product bought when first product is given (lhs)

Support, Confidence and Lift shows relative importance of each rule identified.

- Evaluate Rules

Support, Confidence and Lift are key KPIs for evaluating rules and we will discuss importance of each of these metric.

Support: Support indicates percent transactions with a product combination. . Support indicates % of transactions which are supporting the rule. This is an important indicator to check whether there are enough transactions in support of the rule. In the above example, 0.01% of transactions have "{3854} => {38}" product combination occurring together.

Confidence: For measuring quality of association rules, another measure confidence is used. It is ratio of support for a rule to condition of one product purchase. For rule "{3854} => {38}", we will find support of these two product bring bought together and also how many times first product ("{3854}") bought by a customer.

Conf (R) = Sup (A ∪ B)/Sup (A)

A rule indicates that a Product B is bought along with Product A. So, if buying of product A triggers purchase of product B, we need to check number of times product B is bought when customer buys product A.

Life: Lift is measure importance of rule. It compares confidence of a rule against expected confidence. So, a rule with higher value of Lift is the better. The lift value close to one indicates a redundant rule.

Find Rule with high Support Values

```
    lhs      rhs      support confidence      lift
1   {41,
     48}  => {39} 0.083550736 0.81681082 1.4210493
2   {170} => {38} 0.034379892 0.97805744 5.5288215
3   {36}  => {38} 0.031646288 0.95027248 5.3717570
4   {110} => {38} 0.030909008 0.97530422 5.5132579
5   {170,
     39}  => {38} 0.022901023 0.98057309 5.5430421
6   {38,
     41,
     48}  => {39} 0.022583426 0.83866891 1.4590770
7   {36,
     39}  => {38} 0.022061659 0.95483554 5.3975514
8   {110,
     39}  => {38} 0.019736394 0.98919841 5.5917998
9   {170,
     48}  => {38} 0.017445158 0.98779705 5.5838781
10  {225,
     48}  => {39} 0.015879858 0.80645161 1.4030269
```

Rules - "{41, 48} => {39}" and "{170} => {38}" have higher support values, meaning many transactions have these product combinations in the transaction data. But the confidence level for these rules is lower than one.

Find rules with high Confidence values

.

Find rules with high Lift values

- Actionable Insights

Based on Support, Confidence and Lift values we can select a list of rules. These rules have to be analyzed for insights and actions

We can have new hypotheses as well. We can say what are products or product combination bought by these customers who have bought a specific product as second product. The business may want to identify customers who can targeted for "Product 38", now they are looking target list of customers based on association of between product take up

The second type of hypotheses can based on first product selection, what product to be targeted when we know the first product select by a customer.

Reference

1. Data Source: http://fimi.ua.ac.be/data/retail.dat