Market basket analysis

Abstract:

Market Basket Analysis(MBA) also known as association rule learning or affinity analysis, is a data mining technique that can be used in various fields, such as marketing, bioinformatics, education field, nuclear science etc. The main aim of MBA in marketing is to provide the information to the retailer to understand the purchase behavior of the buyer, which can help the retailer in correct decision making. There are various algorithms are available for performing MBA. The existing algorithms work on static data and they do not capture changes in data with time. But proposed algorithm not only mine static data but also provides a new way to take into account changes happening in data. This paper discusses the data mining technique i.e. association rule mining and provide a new algorithm which may helpful to examine the customer behaviour and assists in increasing the sales.

Introduction:

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

Data processesing:

Data Preparation

Loading the Packages

```
# Importing libraries

library(data.table)

library(readxl)

library(tidyverse)
```

```
library(lubridate)

library(skimr)

library(knitr)

library(treemap)
```

Loading and Inspecting the Data

```
retail <- read_excel# import raw data file and trim leading and trailing whitespaces

("Online Retail.xlsx", trim_ws = TRUE)
```

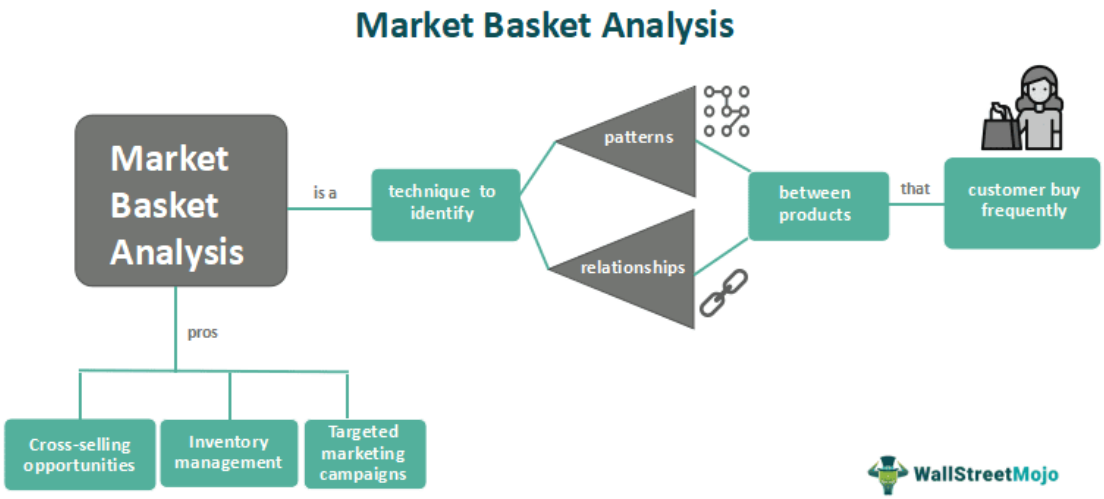
The dataset consists of just over 540,000 observations spread across 8 variables. A few Descriptions and several CustomerIDs are missing and there are also some odd negatives Quantity and UnitPrice that would be worth investigating. It's also worth noting that InvoiceDate is of POSIXct format, from which information about Date and Time of purchase can be extracted.

```
# First glance at the data

retail %>%  skim()
```

```
## Skim summary statistics

##  n obs: 541909

##  n variables: 8

##

## -- Variable type:character -----------------------------------------------------------------

##       variable missing complete      n min max empty n_unique

##        Country       0   541909 541909   3  20     0       38

##    Description    1454   540455 541909   1  35     0     4211

##      InvoiceNo       0   541909 541909   6   7     0    25900

##      StockCode       0   541909 541909   1  12     0     4070

##
```

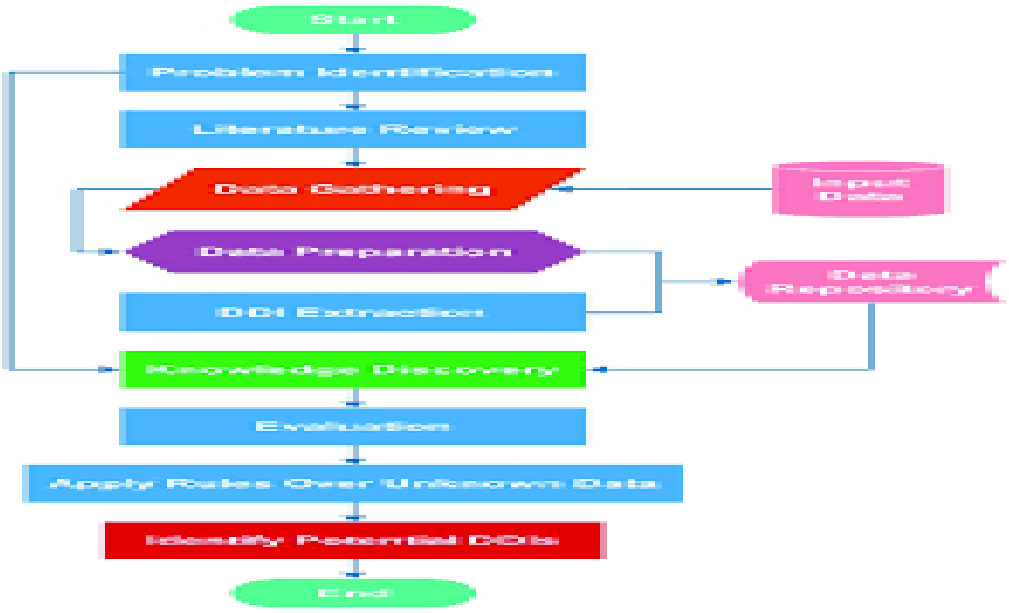
```
## -- Variable type:numeric -------------------------------------------------------------------------
##       variable missing complete        n       mean         sd          p0         p25
##    CustomerID  135080   406829 541909 15287.69 1713.6      12346       13953
##      Quantity       0   541909 541909     9.55  218.08   -80995           1
##     UnitPrice       0   541909 541909     4.61   96.76 -11062.06        1.25
##           p50         p75    p100
##    15152       16791       18287
##        3          10       80995
##        2.08        4.13 38970
##
## -- Variable type:POSIXct -------------------------------------------------------------------------
##       variable missing complete        n        min        max     median
##    InvoiceDate       0   541909 541909 2010-12-01 2011-12-09 2011-07-19
##    n_unique
##       23260
```

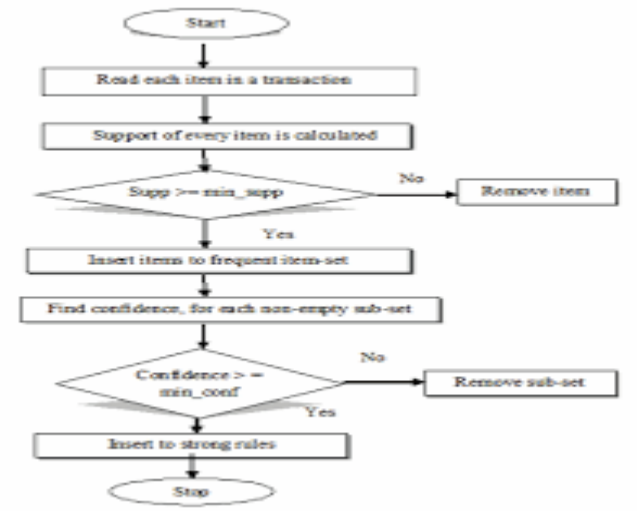
Data images:


Market Basket Analysis
Market Basket Analysis
is a
technique to identify
patterns
relationships
between products
that
customer buy frequently
pros
Cross-selling opportunities
Inventory management
Targeted marketing campaigns
WallStreetMojo

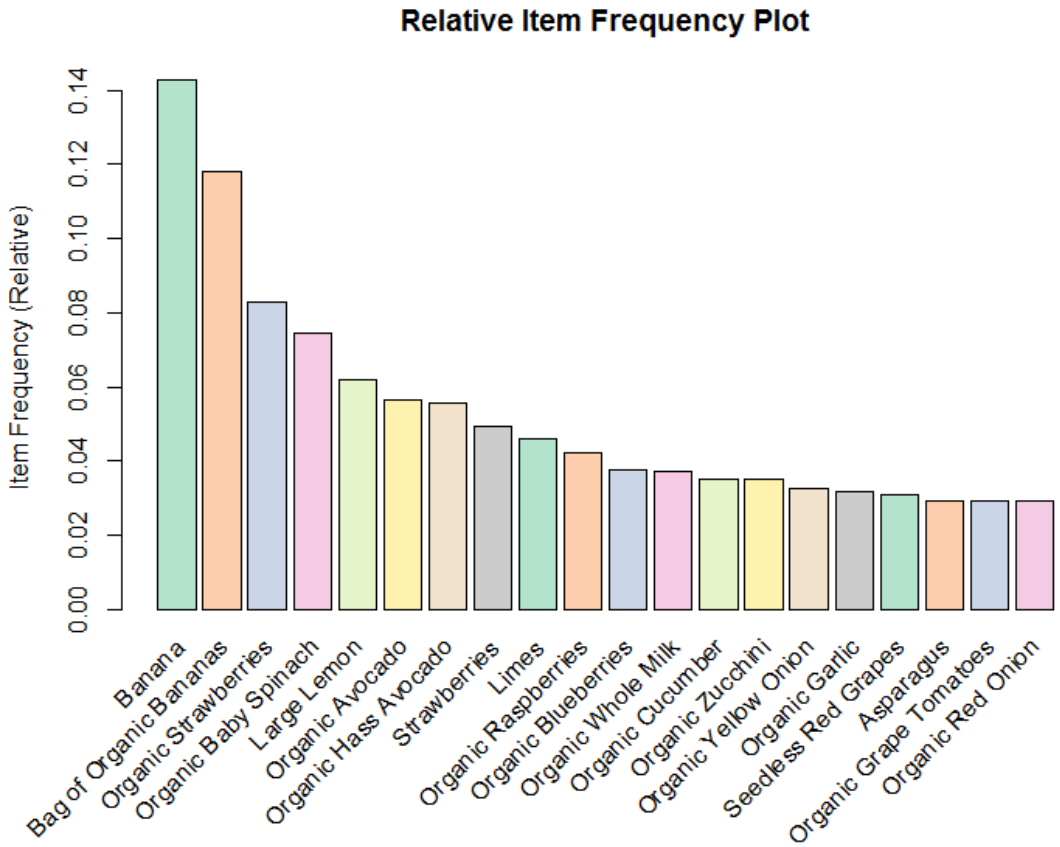

365
Frequent Items
Subset 1
Subset 2
Subset 3

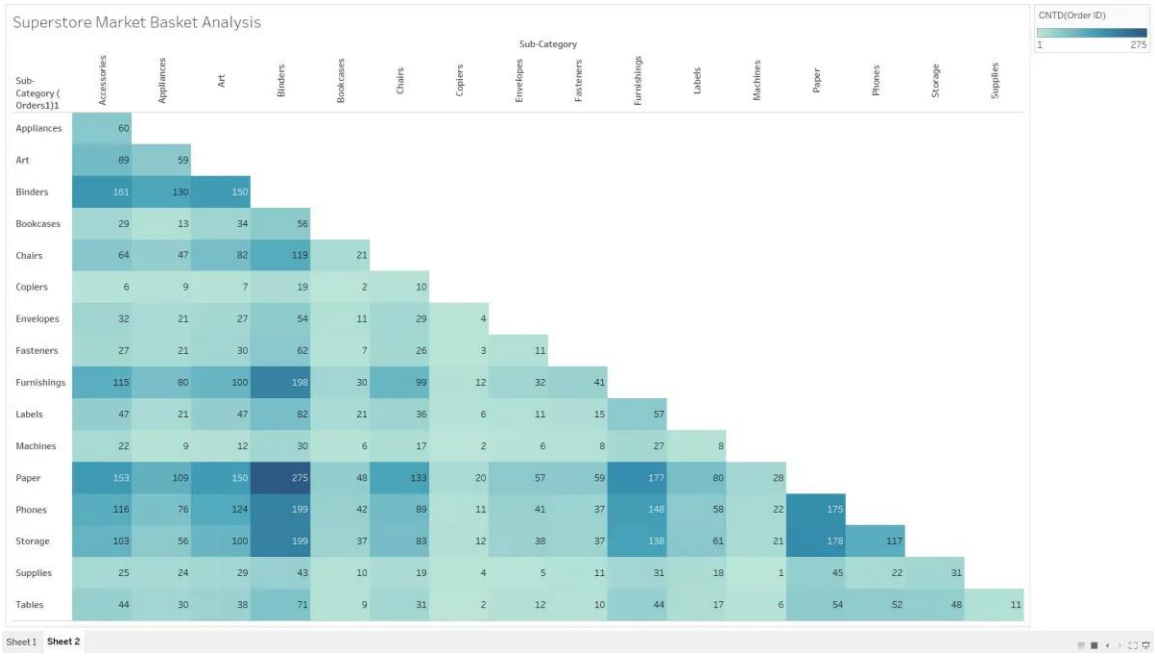
Flow chart:


Start
Problem Identification
Literature Review
Data Gathering
Input Data
Data Preparation
Data Repository
DDI Extraction
Knowledge Discovery
Evaluation
Apply Rules Over Unknown Data
Identify Potential DDIs
End


Start
Read each item is a transaction
Support of every item is calculated
Supp >= min_supp
No
Remove item
Yes
Insert items to frequent item-set
Find confidence, for each non-empty sub-set
Confidence > = min_conf
No
Remove sub-set
Yes
Insert to strong rules
Stop

Graph:


Relative Item Frequency Plot
Item Frequency (Relative)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Large Lemon
Organic Avocado
Organic Hass Avocado
Strawberries
Limes
Organic Raspberries
Organic Blueberries
Organic Whole Milk
Organic Cucumber
Organic Zucchini
Organic Yellow Onion
Organic Garlic
Seedless Red Grapes
Asparagus
Organic Grape Tomatoes
Organic Red Onion


Superstore Market Basket Analysis
CNTD(Order ID)
1
275
Sub-Category
Sub-Category (Orders1)1
Accessories
Appliances
Art
Binders
Bookcases
Chairs
Copiers
Envelopes
Fasteners
Furnishings
Labels
Machines
Paper
Phones
Storage
Supplies
Appliances 60
Art 89 59
Binders 161 130 150
Bookcases 29 13 34 56
Chairs 64 47 82 119 21
Copiers 6 9 7 19 2 10
Envelopes 32 21 27 54 11 29 4
Fasteners 27 21 30 62 7 26 3 11
Furnishings 115 80 100 198 30 99 12 32 41
Labels 47 21 47 82 21 36 6 11 15 57
Machines 22 9 12 30 6 17 2 6 8 27 8
Paper 153 109 150 275 48 133 20 57 59 177 80 28
Phones 116 76 124 199 42 89 11 41 37 148 58 22 175
Storage 103 56 100 199 37 83 12 38 37 138 61 21 178 117
Supplies 25 24 29 43 10 19 4 5 11 31 18 1 45 22 31
Tables 44 30 38 71 9 31 2 12 10 44 17 6 54 52 48 11
Sheet 1
Sheet 2

Introduction to dataset:

The initial dataset was collected from Groceries dataset. Then data was modified and fragmented into 2 datasets for ease of MBA implementation. Here the "groceries data.csv" contains groceries transaction data from which you can do EDA and pre-process the data to feed it in the apriori algorithm. But I have also added pre-processed data as "basket.csv" from which you'll just need to replace nan and encode it using TransactionEncoder after that you can feed the encoded data into the apriori algorithm.

Notes:

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

The adoption of market basket analysis was aided by the advent of electronic point-of-sale (POS) systems. Compared to handwritten records kept by store owners, the digital records generated by POS systems made it easier for applications to process and analyze large volumes of purchase data.

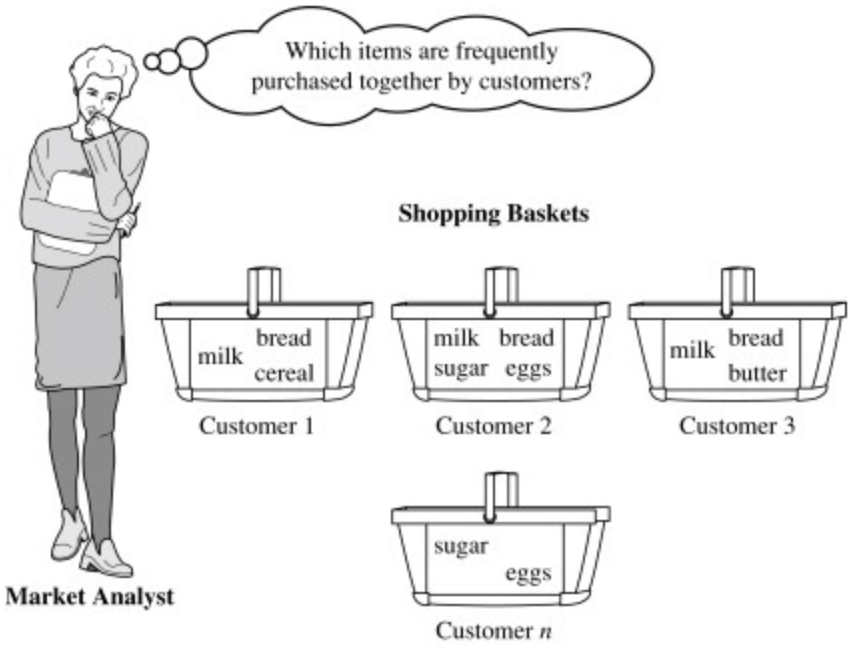
Implementation of market basket analysis requires a background in statistics and data science, as well as some algorithmic computer programming skills. For those without the needed technical skills, commercial, off-the-shelf tools exist.

Attribute dataset:

Market basket analysis is a popular data science technique that identifies combinations of items that are frequently purchased together. Companies often perform it to cross-sell products to users.

Data model:

Market basket analysis is a popular data science technique that identifies combinations of items that are frequently purchased together. Companies often perform it to cross-sell products to users.



Conclusion:

Cross-selling and upselling is the secret mantra of the retail industry that pushed the consumer to buy more. It has become a thriving factor for such industries that harness patterns with market basket analysis in data mining and derive customer insights to upscale their brand's performance.

An urban legend states that a grocery store increased its sales after they placed beer and diapers together because of the market basket analysis that stated that beer and diapers were both purchased frequently by men.

Organizations are using this technique wisely and making billions by playing with the mind of the customer. It is an effective way of improving your sales without having to put extra effort into marketing that won’t give you results as incredible as with this technique. So go ahead and try it on all the data you have in your repository to recognize patterns that may surprise you to the roots.