MARKET BASKET ANALYSIS

PHASE 3: In this part you will begin building your project by loading and preprocessing the dataset. Start the market basket insights project by loading and preprocessing the transaction data.Load the transaction dataset and preprocess the data for association analysis.

GUIDE: MS,R.ARULMOZHI M.E

Assistant professor,ECE department,

MRK institute of tecknology,kattumanarkovil.



PREPARED BY:

SENTHIL KUMAR .V

REGNO:822421106061

MRK INSTITUTE OF TECHNOLOHY

KATTUMANNARKOIL

PHASE 3 - PROJECT

**Market Basket Analysis**

Market basket analysis with Apriori algorithm

The retailer wants to target customers with suggestions on item set that a customer is most likely to purchase. I was given dataset contains data of a retailer; the transaction data provides data around all the transactions that have happened over a period of time. Retailer will use result to grove in his industry and provide for customer suggestions on itemset, we be able increase customer engagement and improve customer experience and identify customer behavior. I will solve this problem with use Association Rules type of unsupervised learning technique that checks for the dependency of one data item on another data item.

# Introduction

Association Rule is most used when you are planning to build association in different objects in a set. It works when you are planning to find frequent patterns in a transaction database. It can tell you what items do customers frequently buy together and it allows retailer to identify relationships between the items.

# An Example of Association Rules

Assume there are 100 customers, 10 of them bought Computer Mouth, 9 bought Mat for Mouse and 8 bought both of them.

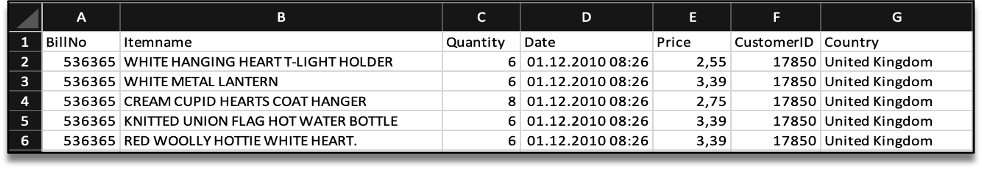
* bought Computer Mouth => bought Mat for Mouse
* support = P(Mouth & Mat) = 8/100 = 0.08
* confidence = support/P(Mat for Mouse) = 0.08/0.09 = 0.89
* lift = confidence/P(Computer Mouth) = 0.89/0.10 = 8.9

This just simple example. In practice, a rule needs the support of several hundred transactions, before it can be considered statistically significant, and datasets often contain thousands or millions of transactions.

# Strategy

* Data Import
* Data Understanding and Exploration
* Transformation of the data – so that is ready to be consumed by the association rules algorithm
* Running association rules
* Exploring the rules generated
* Filtering the generated rules
* Visualization of Rule

# Dataset Description

* File name: Assignment-1\_Data
* List name: retail data
* File format:. xlsx
* Number of Row: 522065
* Number of Attributes: 7
  + Bill No: 6-digit number assigned to each transaction. Nominal.
  + Item name: Product name. Nominal.
  + Quantity: The quantities of each product per transaction. Numeric.
  + Date: The day and time when each transaction was generated. Numeric.
  + Price: Product price. Numeric.
  + Customer ID: 5-digit number assigned to each customer. Nominal.
  + Country: Name of the country where each customer resides. Nominal.

# Libraries in R

First, we need to load required libraries. Shortly I describe all libraries.

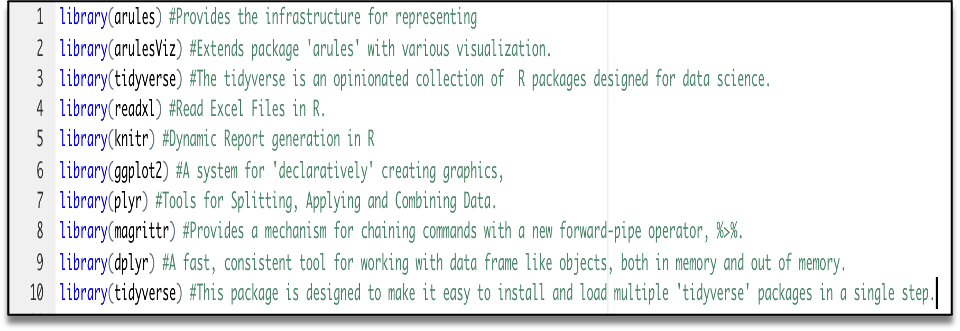
* arules - Provides the infrastructure for representing,

manipulating and analysing transaction data and patterns (frequent itemises and association rules).

* arulesViz - Extends package 'arules' with various visualization.

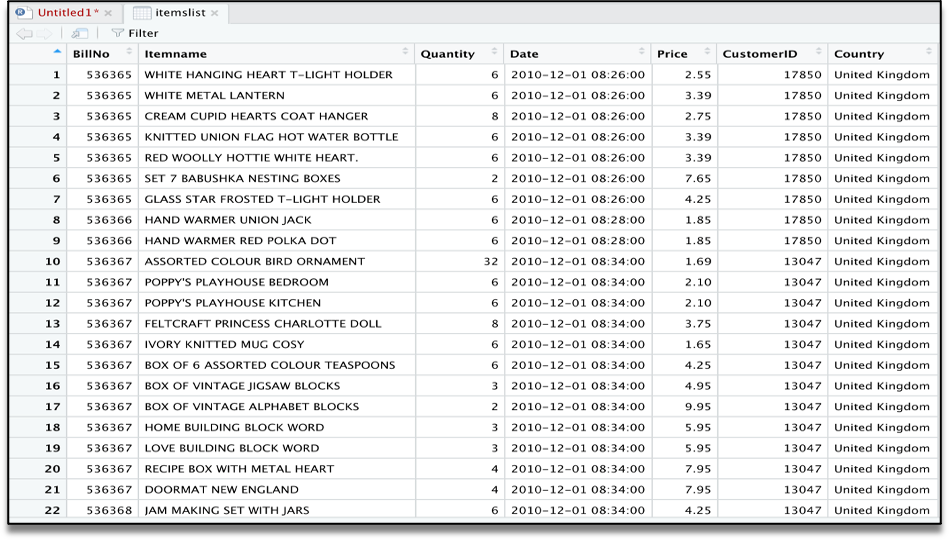
techniques for association rules and item-sets. The package also includes several interactive visualizations for rule exploration.

* tidy verse - The tidy verse is an opinionated collection of R packages designed for data science.
* readxl - Read Excel Files in R.
* plyr - Tools for Splitting, Applying and Combining Data.
* ggplot2 - A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details.
* knitr - Dynamic Report generation in R.
* magrittr- Provides a mechanism for chaining commands with a new forward-pipe operator, %>%. This operator will forward a value, or the result of an expression, into the next function call/expression. There is flexible support for the type of right-hand side expressions.
* dplyr - A fast, consistent tool for working with data frame like objects, both in memory and out of memory.
* tidyverse - This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step.

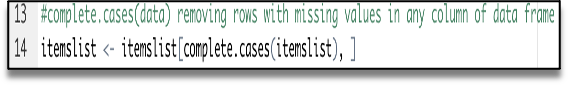


# Data Pre-processing

Next, we need to upload Assignment-1\_Data. xlsx to R to read the dataset.Now we can see our data in R.

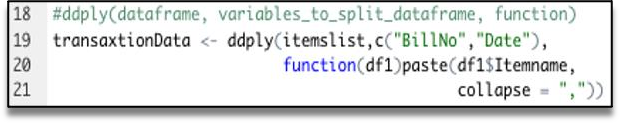


After we will clear our data frame, will remove missing values.

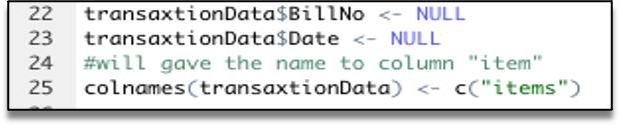


To apply Association Rule mining, we need to convert data frame into transaction data to make all items that are bought together in one invoice will be in one row. Below lines of code will

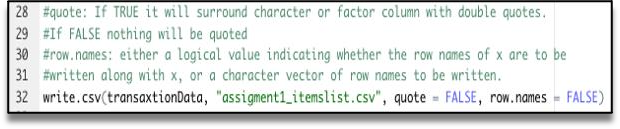
combine all products from one Bill No and Date and combine all products from that Bill No and Date as one row, with each item, separated by (,)



We don’t need Bill No and Date, we will make it as Null. Next, you have to store this transaction data into .csv



This how should look transaction data before we will go to next step.



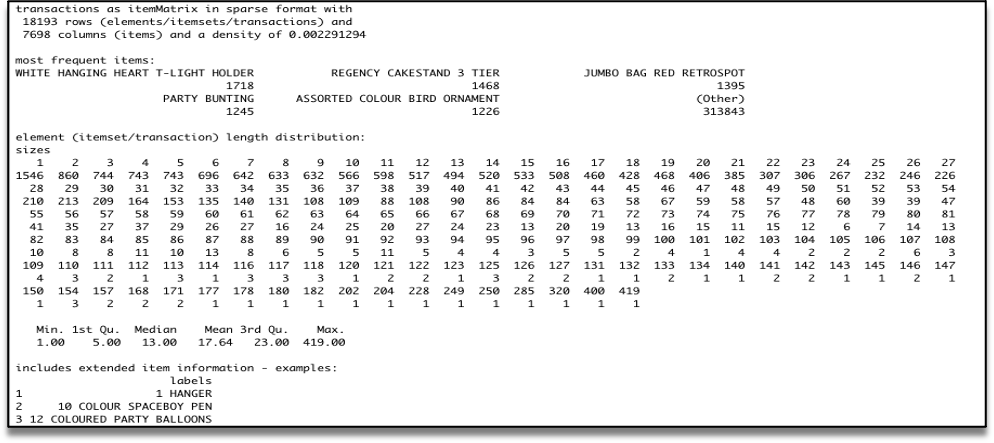
At this step we already have our transaction dataset, and it shows the matrix of items which bought together. We can’t see here any rules and how often it was purchase together. Now let’s check how many transactions we have and what they are. We will have to have to load this transaction data into an object of the transaction class. This is done by using the R function read. Transactions of the arules package. Our format of Data frame is basket.

image

Let’s have a view our transaction object by summary(transaction)

image

We can see 18193 transactions (rows) and 7698 items (columns). 7698 is the product descriptions and 18193 transactions are collections of these items.

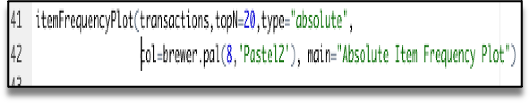


The summary gives us some useful information:

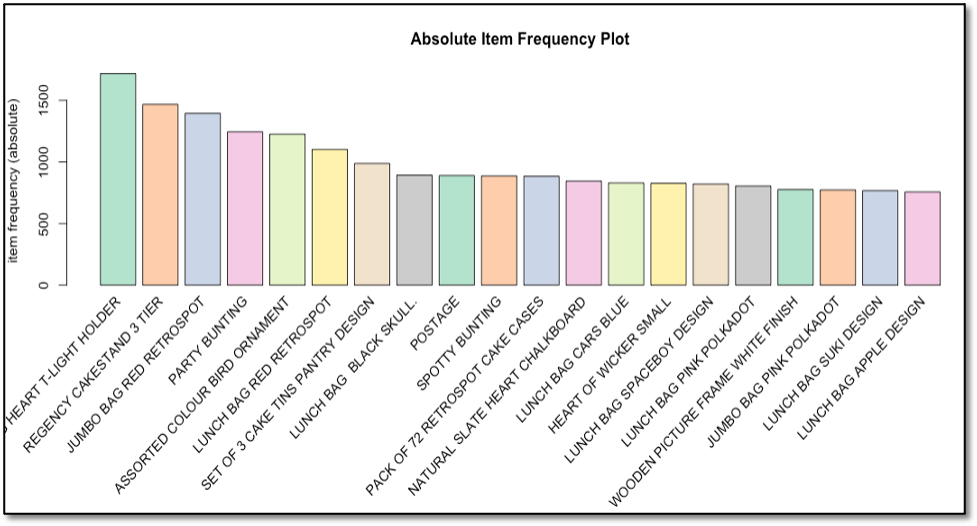
* Density tells the percentage of non-zero cells in a sparse matrix. In other words, total number of items that are purchased divided by a possible number of items in that matrix. You can calculate how many items were purchased by using density: 18193x7698x0.002291294=337445
* Summary will show us most frequent items.
* Element (itemset/transaction) length distribution: It will gave us how many transactions are there for 1- itemset, 2-itemset and so on. The first row is telling you a number of items and the second row is telling you the number of transactions.

For example, there is only 1546 transaction for one item, 860 transactions for 2 items, and there are 419 items in one transaction which is the longest.

Let’s check item frequency plot, we will generate an itemFrequencyPlot to create an item Frequency Bar Plot to view the distribution of objects based on itemMatrix (e.g., >transactions or items in >itemsets and >rules) which is our case.





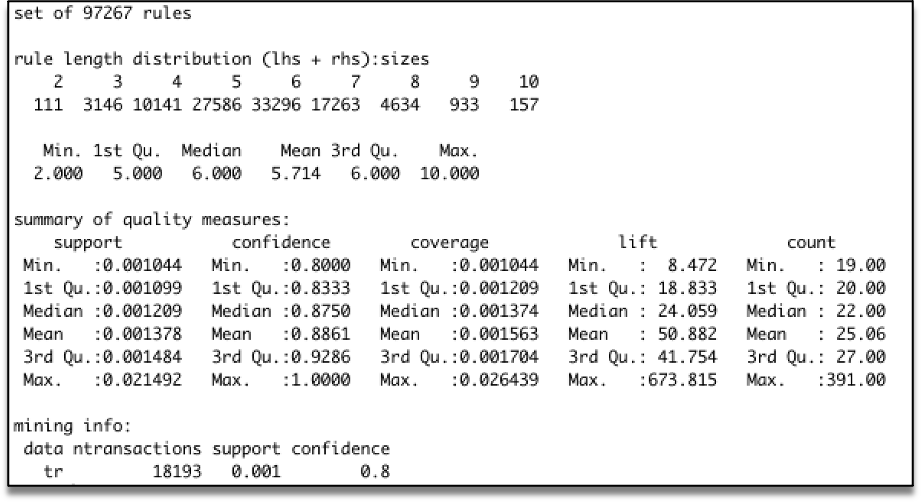


In item Frequency Plot (transaction,topN=20,type="absolute") first argument - our transaction object to be plotted that is tr. Top N is allows us to plot top N highest frequency items. type can be as type="absolute" or type="relative". If we will choose absolute, it will plot numeric frequencies of each item independently. If relative it will plot how many times these items have appeared as compared to others. As well I made it in colure for better visualization.

# Generating Rules

Next, we will generate rules using the Priory algorithm. The function apriority() is from package a rules. The algorithm employs level-wise search for frequent item sets. Algorithm will generate frequent item sets and association rules. We pass sup =0.001 and conf =0.8 to return all the rules that have a support of at least 0.1% and confidence of at least 80%. We sort the rules by decreasing confidence and will check summary of the rules.

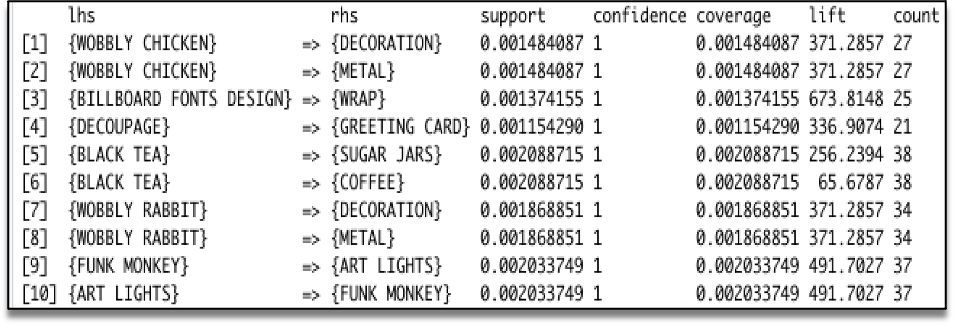
The apriority will take (transaction) as the transaction object on which mining is to be applied. parameter will allow you to set min\_sup and min\_confidence. The default values for parameter are minimum support of 0.1, the minimum confidence of 0.8, maximum of 10 items (maxlen).



Summary of rules give us clear information as:

* Number of rules: 97267
* The distribution of rules by length: a length of 6 items has the most 33296 and length of 2 items has lowest number of rules 111
* The summary of quality measures: ranges of support, confidence, and lift.
* The information on data mining: total data mined, and the minimum parameters we set earlier

Now, 97267 it a lot of rules. We will identify only top 10.

image

Using the above output, you can make analysis such as:

* 100% of the customers who bought 'ART LIGHTS ' also bought 'FUNK MONKEY'.
* 100% of the customers who bought 'BILLBOARD FONTS DESIGN ' also bought 'WRAP'.

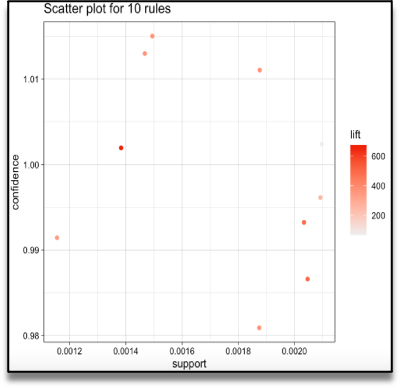
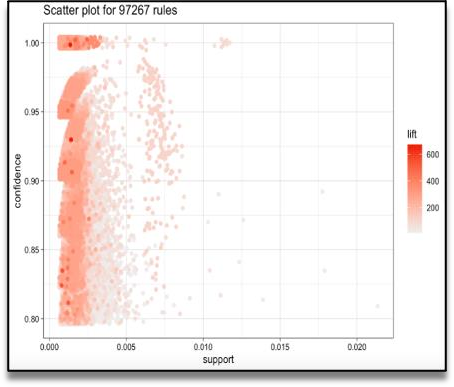
We can limit the size and number of rules generated. we can set parameter in Apriori. If we want stronger rules, we must to increase the value of conf. and for more extended rules give higher value to maxlen.

# Visualizing Association Rules

We have thousands of rules generated based on data, we will need a couple of ways to present our findings. We will use ItemFrequencyPlot to visualize association rules.

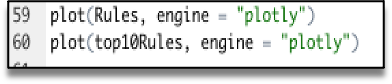
## imageScatter-Plot:

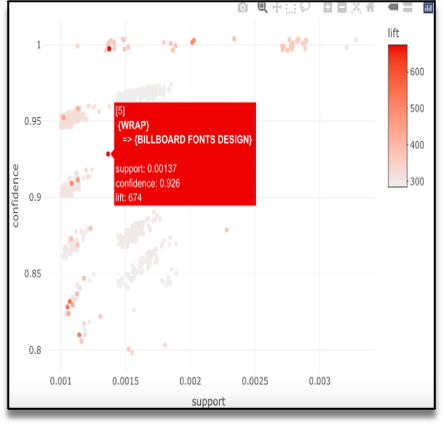
A straight-forward visualization of association rules is to use a scatter plot using plot() of the a rules Viz package. It uses Support and Confidence on the axes. In addition, third measure Liftis used by default to colour (grey levels) of the points.

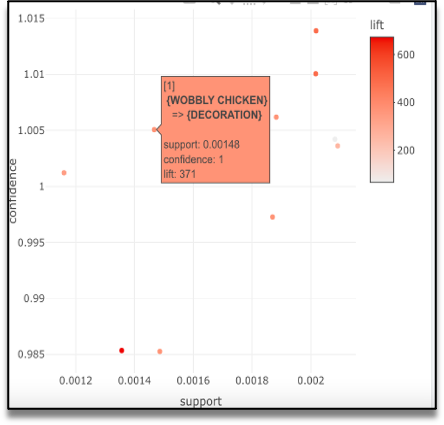


## Interactive Scatter-Plot:

We can have a look for each rule (interactively) and view all quality measures (support, confidence and lift).

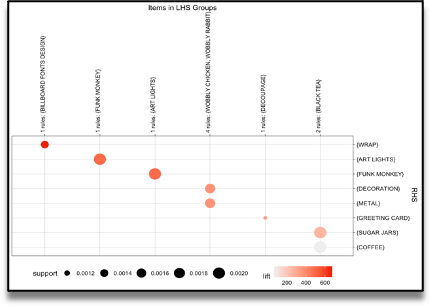
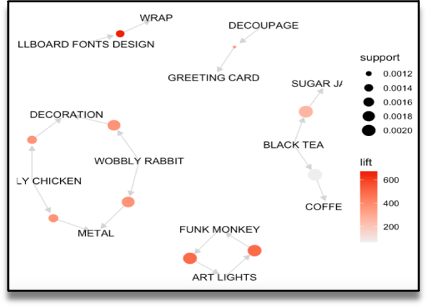






## Graph - Based Visualization and Group Method:

Graph plots are a great way to visualize rules but tend to become congested as the number of rules increases. So, it is better to visualize a smaller number of rules with graph-based visualizations. We can see as well group method for top 10 items.



# Conclusion

Based on the results of these calculations can be used as a recommendation for retail owners to arrange the arrangement of product catalogs and take strategic steps to improve product marketing.. By utilizing the association rules which are discovered as a result of the analyses, the retailer can apply effective marketing and sales promotion strategies, he will be able increase customer engagement and improve customer experience and identify customer behavior.