|  |  |
| --- | --- |
| A picture of a winding road and trees  Market Basket Analysis | Identify the Changing Trends of Market Data Using Association Rule Mining  By: - Kanika SHARMA, Jing Zhang, Bomin XIE  Customer Data Analysis Research Paper |

**Table of Contents**

[**Abstract** 1](#_Toc507242257)

[**Research Goal** 2](#_Toc507242258)

[**Data Understanding** 2](#_Toc507242259)

[**Market Basket Analysis** 2](#_Toc507242260)

[**Data Analysis** 3](#_Toc507242261)

[**Applying Rules** 5](#_Toc507242262)

[**Data Visualization** 7](#_Toc507242263)

[**Recommendation** 9](#_Toc507242264)

[**Drive Business Decisions** 10](#_Toc507242265)

[**Expanding on the analysis** 11](#_Toc507242266)

[**References** 11](#_Toc507242267)

## **Abstract**

In today’s fast changing digital world, many organizations generate a large amount of transaction data daily. it is always important from seller’s point of view to know what customer might want to buy. Everyone ends up using Amazon’s recommendation almost every time we visit the website. Obviously, this creates an excellent chance for the retailers.

If Retailer can tell what customer might want to buy, this only helps in improving sales but also in maintaining customer satisfaction rating and lifetime value. Contrary to this, if you are not able to predict the trend or purchase pattern among your customer, they might not come back to your store or might not buy from your website.



The Objective of this project is to learn one Machine Learning Algorithm using Market Basket Analysis and Association Rule Mining which will enable retailers to predict the items bought together frequently by Customer from their store or website.

## **Research Goal**

The “Global Mart” is a web store that caters to customers from across the world. The goal is to work out the foremost ofttimes occurring combination of the things that area unit bought along. this can modify you to advocate the connected things to a client, once he makes a procurement within the store.

## **Data Understanding**

The Global Mart dataset and its data dictionary is included in the Sheet. Every row of the dataset represents an item of the order. but, the Order ID isn't always unique. Let’s see the difference columns of the Data.

1. Invoice No
2. Stock Code -unique number for the item
3. Description -product description
4. Quantity- number of items bought
5. Invoice Date- Date of the purchase done
6. Unit Price- Unit Price of the item bought
7. Customer ID – ID of the customer who did the purchase
8. Country- country where the purchase was made

As a result, the one-of-a-kind items ordered at a time discern in one of a kind rows with the identical Order ID. The number of unique items is just too large a number for making sense out of the data. Hence Market Basket Analysis must be used to find the meaning.

## **Market Basket Analysis**

Market Basket Analysis is one among the key techniques employed by almost of the retailers to uncover associations between items. It works by trying to find mixtures of things that occur along oftentimes in transactions. Basically, it helps retailers to find relationships between the things that individuals purchase.  
Association Rules are widely used and are intended to identify strong rules discovered in transaction data using measures of interestingness, based on the concept of strong rules.

* Support: The fraction of which our item set occurs in our dataset.
* Confidence: probability that a rule is correct for a new transaction with items on the left.
* Lift: The ratio by which by the confidence of a rule exceeds the expected confidence.   
  Note: if the lift is 1 it indicates that the items on the left and right are independent.

An example of Association Rules

\* Assume there are 100 customers  
\* 10 of them bought milk, 8 bought butter and 6 bought both  
\* bought milk => bought butter  
\* support = P (Milk & Butter) = 6/100 = 0.06  
\* confidence = support/P(Butter) = 0.06/0.08 = 0.75  
\* lift = confidence/P(Milk) = 0.75/0.10 = 7.5

Here we can also represent it as

We can represent our items as an item set as follows:

*I = {i1, i2, in}*

Therefore, a transaction is represented as follows:

*Tn = {t1,t2, in}*

This gives us our rules which are represented as follows:

*{i1, i2} => {in}*

Alternatively, it can be written as “if a user buys an item in the item set on the left-hand side, then the user will likely buy the item on the right-hand side too”. In short, it can be represented as

*{coffee, sugar} => {milk}*

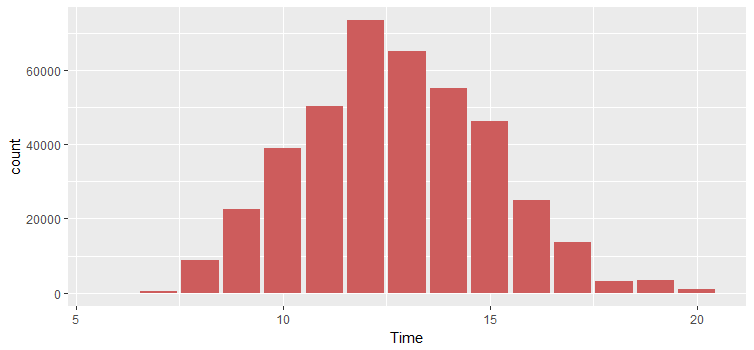
If a customer buys coffee and sugar, then they are also likely to buy milk.

## **Data Analysis**

Let’s try to explore the data more before doing association rule mining on it. The count of the data is 406829 records and having 10 columns like mentioned above. This data seems to have 90 % records for United Kingdom Country.

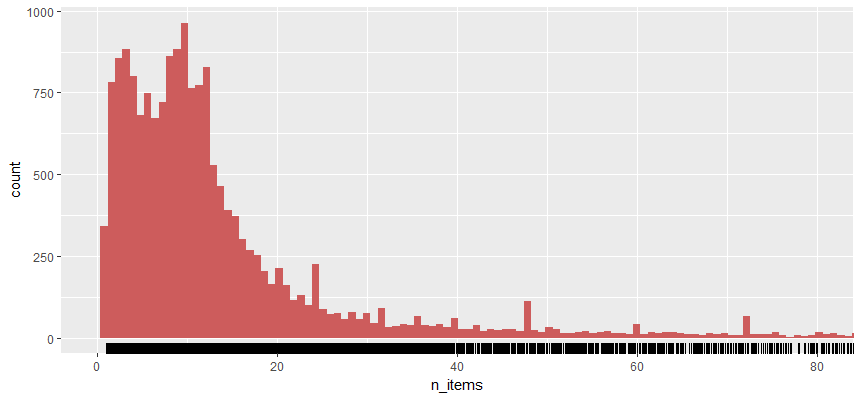
The data can have analysed more if we have following questions in our mind

1. What time do people buy online?



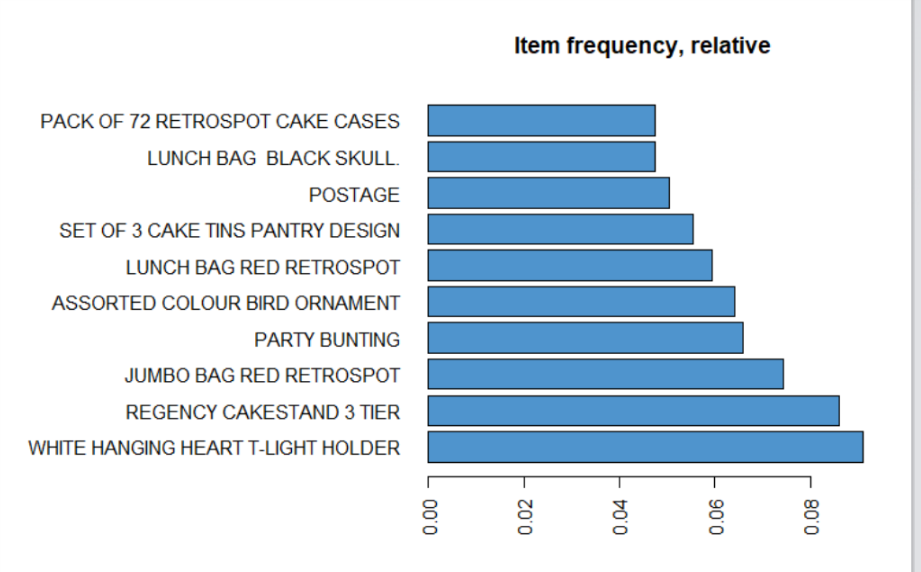
Using Hour from Invoice Data, we found following result, Customer generally buy between 10:00–15:00.

1. Approximately How many number of products each customer buy?



This graph shows that mostly customers buy less than 10 products. We can also try to find the most sellers

1. Which items are the most sold?



It is evident from the graph that the products like White Hanging Light Holder, Regency Cake stand 3 Tier and Jumbo Bag red retro spot are mostly purchased by customers.

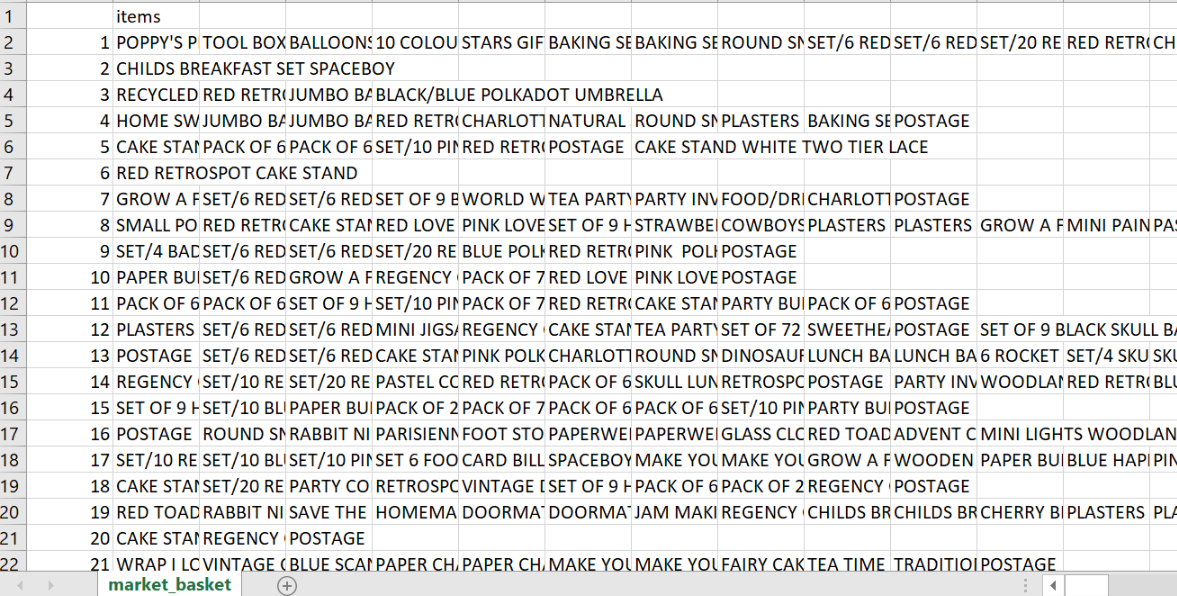
## **Applying Rules**

Association rule mining is a popular data mining method available in R as the extension package arules. However, mining association rules often results in a very large number of found rules, leaving the analyst with the task to go through all the rules and discover interesting ones.

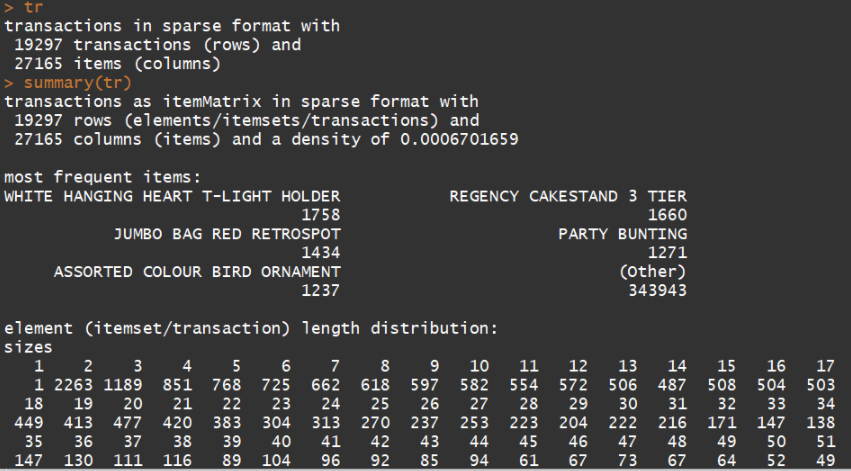
Before Applying Association Rule mining algorithm, we must convert the data from the csv format into transactions level format.

1. using Function deploy ()

This function takes data and splits it into pieces, giving us the best format.



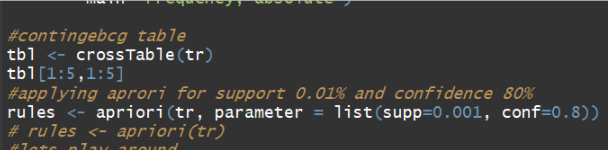
1. Read the number of Transactions



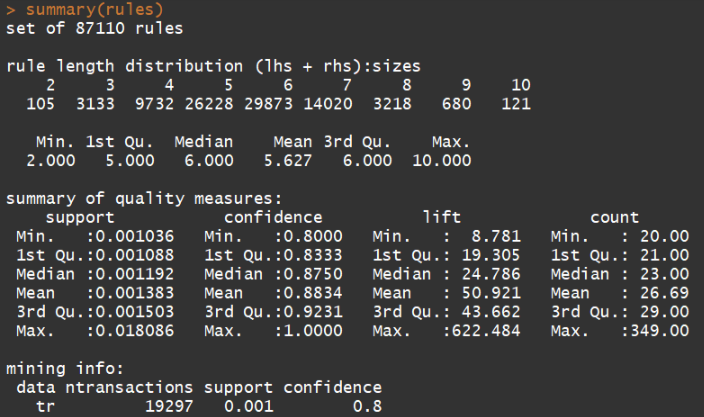
* Number of transactions 19297
* density: The percentage of non-empty cells in the sparse matrix that is 0.00067
* The most frequent items should be the same as our results in above image for Data Analysis.
* Element length Description: 1 transactions were for just 1 item, 2263 transactions for 2 items, all the way up to the biggest transaction: 1 transaction for 420 items. This indicates that most customers buy a small number of items in each transaction.

1. Apply Rules

Using Apriori Package in R, we can see the following result which shows frequent item sets and rules.



Sup=0.001 and confidence=0.8 values are passed to have all the rules that have a support of at least 0.1% and confidence of at least 80%.



\*The number of rules: 89,697.

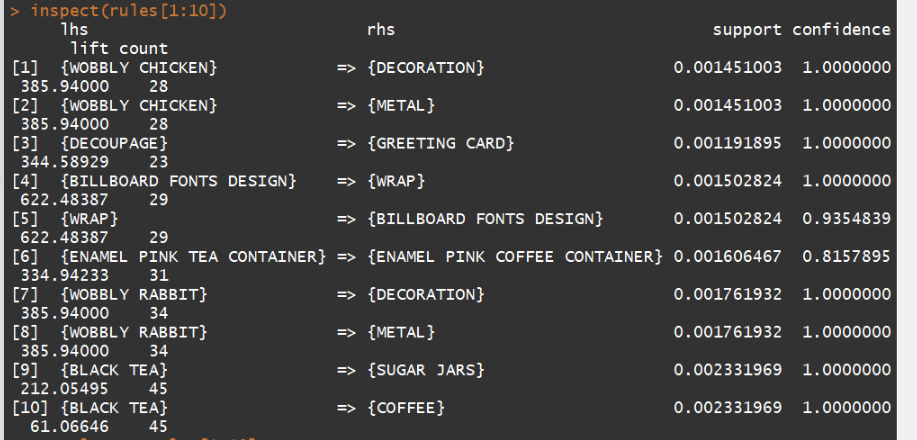
\* The distribution of rules by length: a length of 6 items has the most rules.  
\* The summary of quality measures: ranges of support, confidence, and lift.  
\* The information on data mining: total data mined, and the minimum parameters we set earlier.

Note: it is very important to provide Support and Confidence value because running this function on real world data (having millions of transactions) is time consuming and computationally expensive. Sometimes In few cases, rules will be redundant. Repetition shows that one item might be a given. As a Data Analyst, you can choose to prune the items from the dataset. On the other hand, you can expel excess rules generated. We can dispense with these repeated. Same was analysed on our dataset and we found that we do not have redundant rows.

## **Data Visualization**

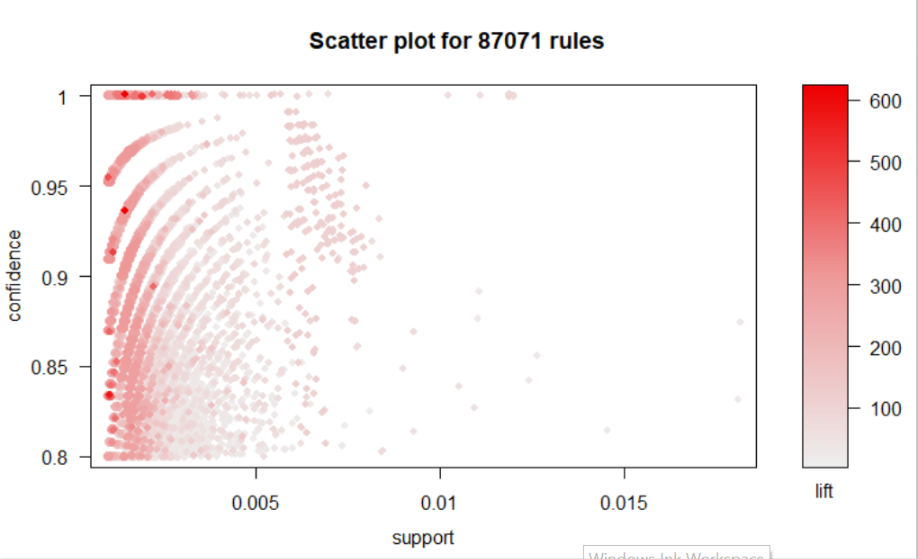
Sifting manually through large sets of rules is time consuming and strenuous. Visualization has a long history of making large data sets better accessible using techniques like selecting and zooming. Let’s plot some rules but wait to plot all the 89K rules does not make sense. So, we plot only TOP 10 rules. Again, the number of TOP is subjected, and it depends upon the requirement.

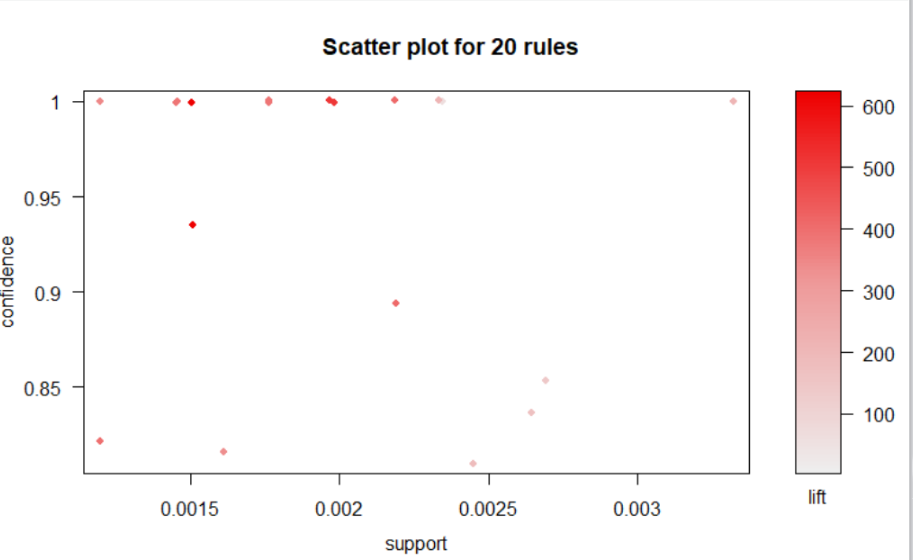
* TOP 10 RULES



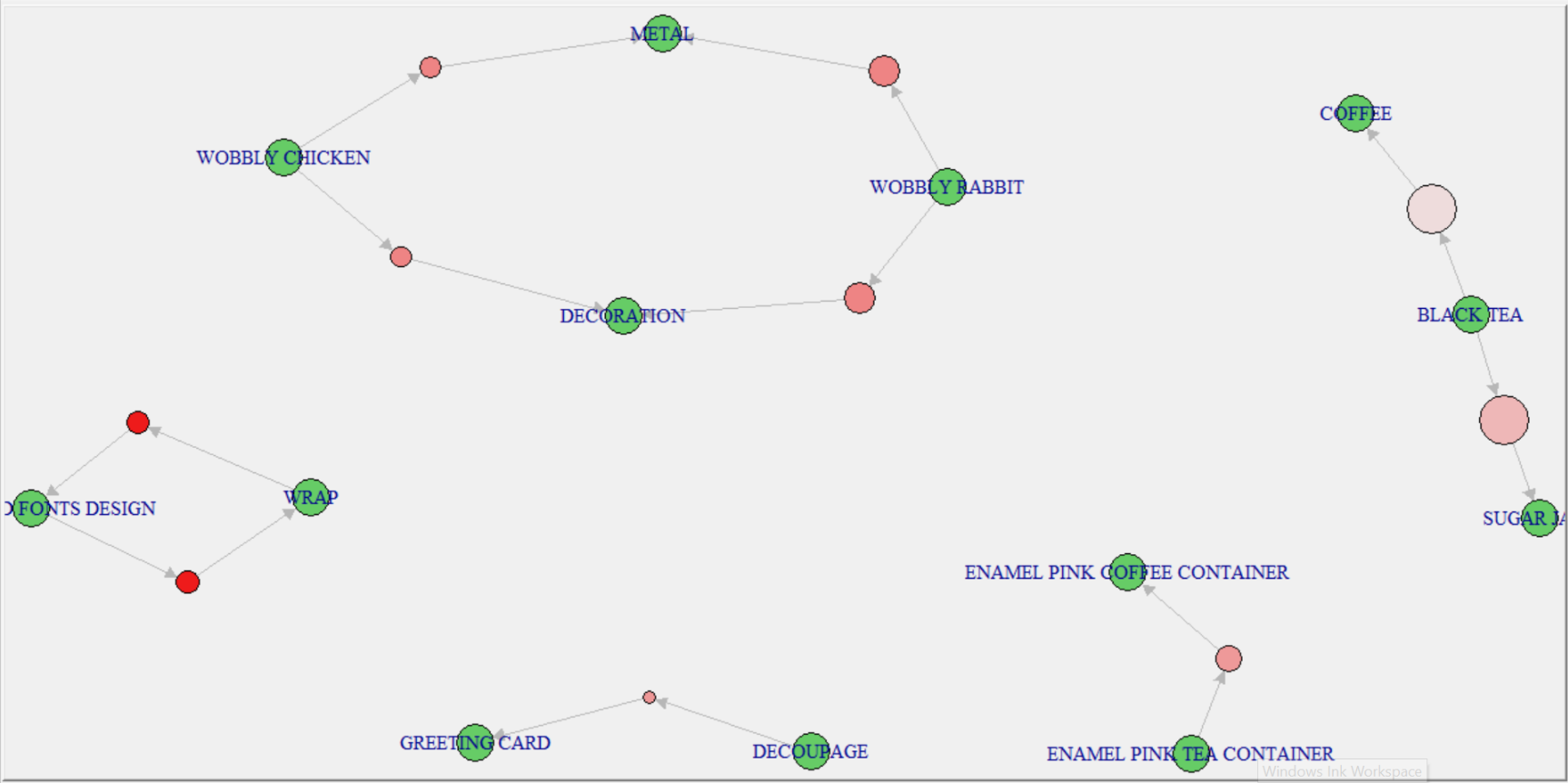
* Scatter Plot

A straight-forward visualization of association rules is to use a scatter plot with two interest measures on the axes. The default method for plot () for association rules in arulesViz is a scatter plot using support and confidence on the axes.





* Interactive Graph



The Pink or the red coloured circles indicate Support value and It is clearly understood from the picture how various products are affecting their respective RHS products.

## **Recommendation**

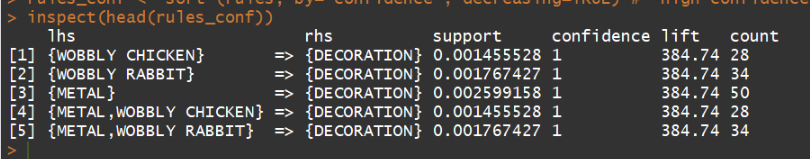
The interpretation is straightforward:

\* 100% customers who bought “WOBBLY CHICKEN” also bought “DECORATION”.  
\* 100% Customer who bought “BLACK TEA” also bought “SUGAR JAR”.

This interpretation also helps in understanding two questions

* What are customers likely to buy before buying Product Y ?
* What are customers likely to buy if they purchase Product Y?

Same can be analysed in R by using appearance function. Let’s say for example for “DECORATION”.



## **Drive Business Decisions**

Recently we utilize the information to make any kind of commerce choice, it is vital that we take a step back and keep in mind something important: The yield of the examination reflects how as often as possible things co-occur in exchanges. This is work both quality of affiliation

association between the things, and the way the location has displayed them. To say that in a distinctive way: things might cooccur not since they are “naturally” associated, but since we, the individuals in charge of the location, have displayed them together. This is a case of a more common issue in web analytics: our information reflects the way clients carry on, and the way we have energized them to act, by the site plan choices we have made.

* Store layout (put products that co-occur together close to one another, to improve the customer shopping experience)
* Marketing (e.g. target customers who buy flour with offers on eggs, to encourage them to spend more on their shopping basket)

We require to be aware of this, since, if as recommended prior in the formula, we utilize the comes about to illuminate where things are put relative to one another, we require to control for how near they are arranged on the site nowadays, so that we don’t conclusion up affirming what we have accepted.

**Driving website organization**

* Inform the placement of content items on their media sites, or products in their catalogue.

Large clusters of co-occurring things ought to likely be set in their claim category / theme. Item sets that commonly co-occur ought to be set near together inside broader categories on the site. This is particularly critical where one thing in a match is exceptionally well known, and the other thing is exceptionally tall edge. This can result in significant uplift in profit and it can strengthen the recommendation system.

* Drive recommendation engines (like Amazon’s customers who bought this product also bought these products…)

**Driving targeted marketing**

The same results about can be utilized to drive focused on targeted campaigns. For each client, we choose a modest bunch of items based on items they have bought to date which have both a high uplift, and send them a e.g. personalized email or show advertisements etc. How we utilize the investigation has noteworthy suggestions for the examination itself: in case we are nourishing the investigation into a machine-driven handle for conveying proposals, we are much more interested in producing a broad set of rules. In case, be that as it may, we are testing with focused on promoting for the to begin with time, it makes much more sense to choose a modest bunch of especially tall esteem rules, and activity fair them, sometime recently working out whether to contribute in the exertion of building out that capability to oversee a much more extensive and more complicated run the show set.

## **Expanding on the analysis**

In the above example, we used actual transaction events to identify associations between products for an online retailer. Sticking with our retail example, however, we could have expanded the scope of our definition of transactions

* We could widen the scope by not just looking at add-to-basket-events, but by analysing each item that each guest has seen, and relate bunches of items that person clients have looked at inside a single session by using WHERE "event" = 'pageview'
* Analysis can again be expanded by Clustering the segments of customers if sufficient metadata about customer is provided at the same time.

## **References**

1. <http://www.salemmarafi.com/code/market-basket-analysis-with-r/>

2. <https://discourse.snowplowanalytics.com/t/market-basket-analysis-identifying-products-and-content-that-go-well-together/1132>