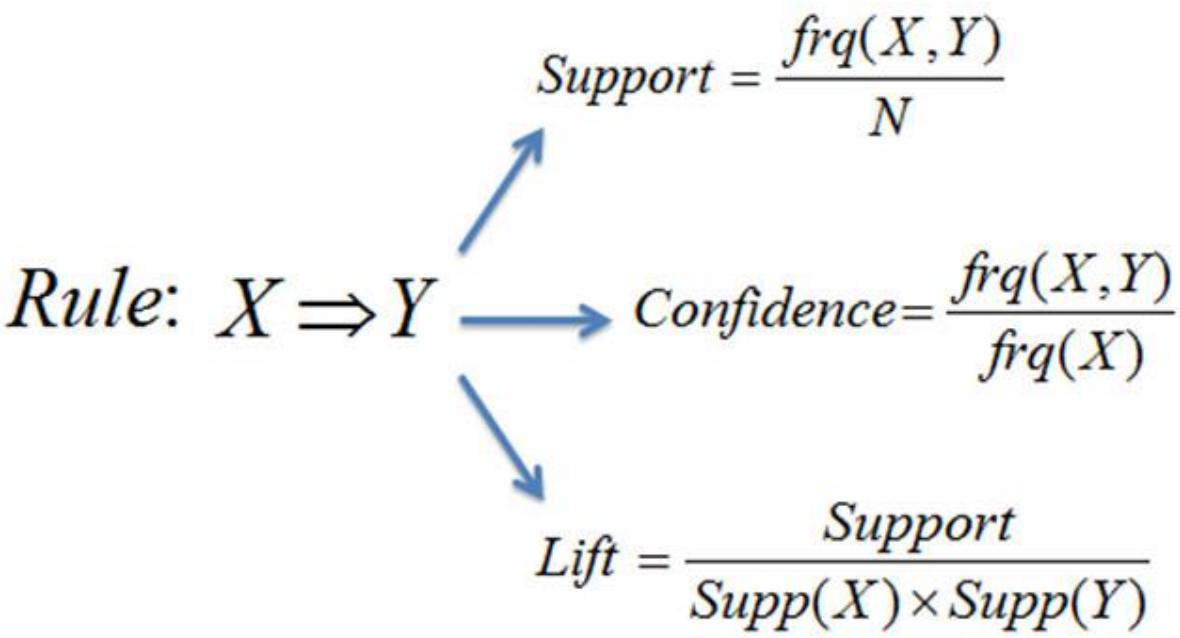
Problem Statement

Issues such as displaying the products at appropriate place, spending money on advertisements, recommending the right product set to the customers etc have always posed challenges to the retailers while trying to boost their top-line and bottom-line.

Solution

Perhaps, by knowing the association and uncovering the relationships between the items being sold would help the retailers in strategizing and implementing various marketing efforts effectively & efficiently. One tool that can aid retailers in that pursuit is Apriori algorithm that is used in mining frequently bought item sets and the relationships between them by applying appropriate association rules. The algorithm relies on metrics such as Support, Confidence and Lift in determining the association rules and this report briefly touches upon the definitions of those terms without deep diving into the details.



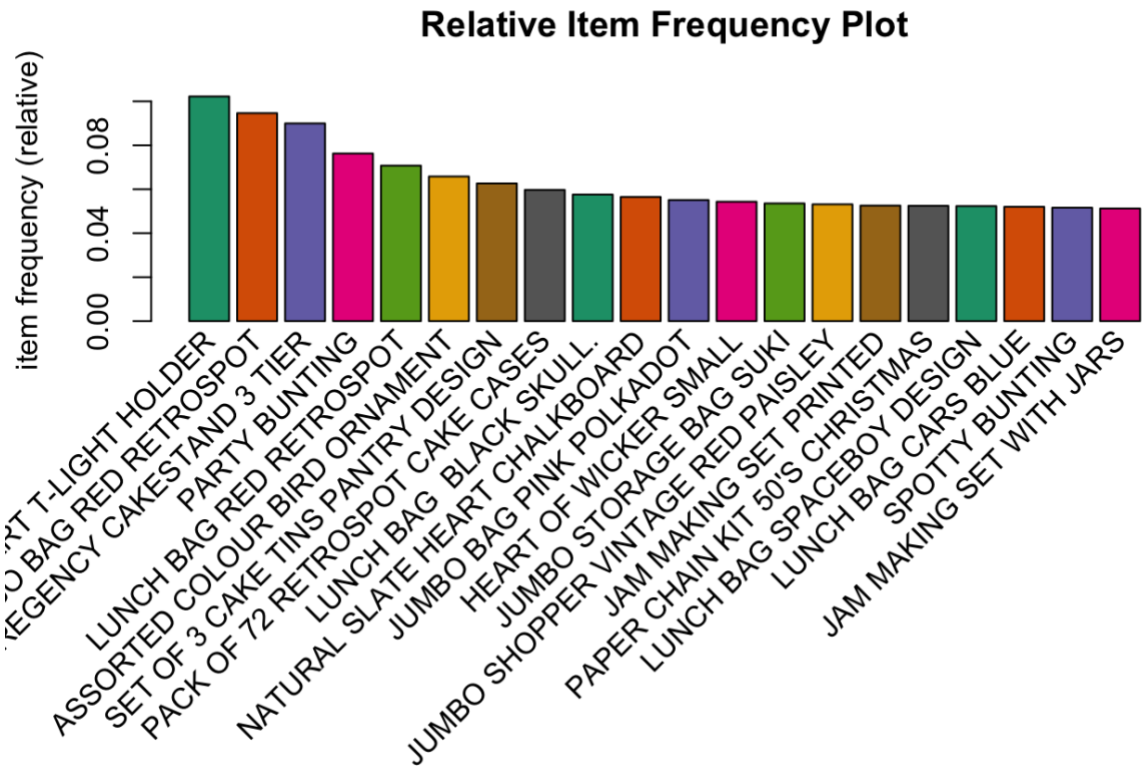
X,Y represents product X and Product Y. The rule could be defined as if a customer buys X then he is likely to buy Y as well. In order to know how strong the rule is we need to rely on the metrics such as support, confidence, and lift. This kind of analysis is also called as Market Basket analysis.

Case Study

Let us consider a data set called Online_Retail.CSV which is a real time dataset belongs to an UK e-commerce retailer. It has over 500,000 records in total and it includes about 10,000 cancelled transactions as well. Invoice numbers are repeated for each item in a product and after data pre-processing, we ended up with over 20,000 records.

When we apply apriori on the data set, we got interesting data insights that are as follows:

- There are a total of 22107 transactions and 4239 items that constitute the transactions.
- 1 item appears in 2260 transactions and 2 items appear in 849 transactions etc. There are many items that appear in only one transaction.
- The top 20 items have a frequency of, on an average, 5% to 7% , meaning that top 20% items appear on an average in over 5% of the transactions, hence , initial support should be much lower than 5% to accommodate more products and to generate reasonable association rules. Relative Item Frequency Plot showing the details is given below.
- 166 rules were generated when we ran apriori with a support of 1% and 70% confidence.



Top 20 Rules

How do we interpret the given below rules?

	lhs		rhs	support	confidence	coverage	lift	count
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
[1]	{PAINTED METAL PEARS ASSORTED}	=>	{ASSORTED COLOUR BIRD ORNAMENT}	0.01171575	0.7000000	0.01673678	10.635670	259
[2]	{REGENCY SUGAR BOWL GREEN}	=>	{REGENCY MILK JUG PINK}	0.01099199	0.7546584	0.01456552	50.708915	243
[3]	{REGENCY MILK JUG PINK}	=>	{REGENCY SUGAR BOWL GREEN}	0.01099199	0.7386018	0.01488216	50.708915	243
[4]	{REGENCY SUGAR BOWL GREEN}	=>	{REGENCY TEAPOT ROSES}	0.01053965	0.7236025	0.01456552	38.453558	233
[5]	{CANDLEHOLDER PINK HANGING HEART}	=>	{WHITE HANGING HEART T-LIGHT HOLDER}	0.01275614	0.7085427	0.01800335	6.930864	282
[6]	{BAKING SET SPACEBOY DESIGN}	=>	{BAKING SET 9 PIECE RETROSPOT}	0.01510834	0.7182796	0.02103406	17.001078	334
[7]	{POPPY'S PLAYHOUSE LIVINGROOM}	=>	{POPPY'S PLAYHOUSE KITCHEN}	0.01185145	0.7987805	0.01483693	40.408788	262
[8]	{POPPY'S PLAYHOUSE LIVINGROOM}	=>	{POPPY'S PLAYHOUSE BEDROOM}	0.01176098	0.7926829	0.01483693	41.427521	260
[9]	{REGENCY TEA PLATE PINK}	=>	{REGENCY TEA PLATE GREEN}	0.01257520	0.9114754	0.01379654	52.887105	278
[10]	{REGENCY TEA PLATE GREEN}	=>	{REGENCY TEA PLATE PINK}	0.01257520	0.7296588	0.01723436	52.887105	278
[11]	{REGENCY TEA PLATE PINK}	=>	{REGENCY TEA PLATE ROSES}	0.01216809	0.8819672	0.01379654	43.814942	269
[12]	{SET OF 6 SNACK LOAF BAKING CASES}	=>	{SET OF 12 FAIRY CAKE BAKING CASES}	0.01144434	0.7146893	0.01601303	28.622528	253
[13]	{SET/6 RED SPOTTY PAPER CUPS}	=>	{SET/6 RED SPOTTY PAPER PLATES}	0.01583209	0.8177570	0.01936038	34.238928	350
[14]	{TOILET METAL SIGN}	=>	{BATHROOM METAL SIGN}	0.01325372	0.7234568	0.01831999	23.764427	293
[15]	{SET 3 RETROSPOT TEA}	=>	{SUGAR}	0.01945085	1.0000000	0.01945085	51.411628	430
[16]	{SUGAR}	=>	{SET 3 RETROSPOT TEA}	0.01945085	1.0000000	0.01945085	51.411628	430
[17]	{SET 3 RETROSPOT TEA}	=>	{COFFEE}	0.01945085	1.0000000	0.01945085	43.347059	430
[18]	{COFFEE}	=>	{SET 3 RETROSPOT TEA}	0.01945085	0.8431373	0.02306962	43.347059	430
[19]	{SUGAR}	=>	{COFFEE}	0.01945085	1.0000000	0.01945085	43.347059	430
[20]	{COFFEE}	=>	{SUGAR}	0.01945085	0.8431373	0.02306962	43.347059	430

For example, looking at row number 15 in the table, because confidence is 1 for the itemset, there is a 100% probability that the customer who buys SET 3 RETROSPOT TEA will also buy SUGAR. Further, lift of 51.41 indicates that the customer purchasing the 2 items together is 51 times higher compared to buying the second item alone.

Recommending the related products

Retailers can recommend the related products to the customers based on their confidence and lift values. For example, as explained bove, If a customer is buying SET 3 RETROSPOT TEA then we can recommend him SUGAR as confidence for the itemset is very high (100% probability) coupled with a high lift value (51.41).

We can even use these rules to display the products at aisles in the retail stores and to optimize the advertisement budget and to increase sales.

Management can implement the rules and if the results are not satisfactory (if there is no noticeable improvement in sales etc) it can change the support, confidence levels to generate new rules and implement the new rules. It’s an iterative and continuous approach till management finds acceptable levels of metrics.