CSE 801-B

Data Mining

Fall 2020

Final Report – Team Outliers
Retail Store Data – Association Analysis

Group Members:

Karan Dalal	–	dalalkar@msu.edu
Neil Joshi	–	joshine2@msu.edu
Syed Kashif Kamoonpuri	–	kamoonpu@msu.edu
Vishal Agarwal	–	agarwa97@msu.edu

Table of Content

Introduction

In Data Mining, association rules play an essential role in predicting and analyzing consumer behavior. Association analysis finds its use in formulating cross-selling strategies, product clustering, market basket analysis, among many others. This concept is most prominent in the retail industry, aiming to discover relations between seemingly independent transactions. The co-occurrence analysis can help the retail industry experts effectively use effective marketing strategies and take advantage of the transaction-level data captured.

The project is based on market basket analysis on a publicly available online retail dataset. The study will help us get insights into the purchase decisions taken by the retailer's customers. The customer base for this company is majorly wholesaler buyers who purchase in bulk quantity.

The dataset has transactions spanning from December 2010 to December 2011. In this analysis, association mining algorithms like Apriori is used to carry out Market-basket analysis. An exciting insight would be creating association rules of products with high volume and high unit price. Such insight can have a direct impact on the top-line revenue of the retail channel. We can compare the product combinations for each and design a customized marketing strategy.

Objective

A. Perform market basket analysis on sales data of an online retail store to identify associations amongst products.
B. Understand various association rules and their behavior over the collected data
C. Strategy for Frequently Bought Together products. The online retail store is interested in creating a 'frequently bought together' items suggestion list for its customers as a section on their website.
D. Cross-Selling Strategy - After identifying some interesting purchase decisions by the customers, we need to devise a customized marketing plan for potential cross-selling opportunities.

Data Description

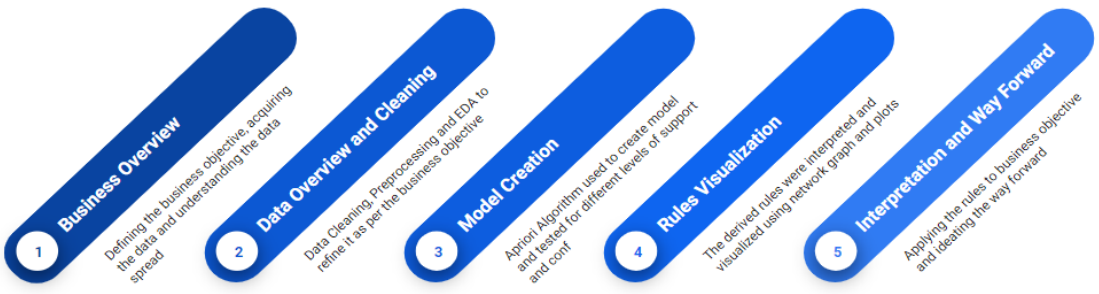
The dataset for this project is available at Kaggle and can be accessed using the following link: https://www.kaggle.com/puneetbhaya/online-retail

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEAI	6	12/1/10 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTER	6	12/1/10 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS	8	12/1/10 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG	6	12/1/10 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE \	6	12/1/10 8:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NEST	2	12/1/10 8:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED	6	12/1/10 8:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNIO	6	12/1/10 8:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED P	6	12/1/10 8:28	1.85	17850	United Kingdom

Column Description:

1. InvoiceNo: Invoice number for each billing statement
2. StockCode: Unique code for each product
3. Description: Product description
4. Quantity: Quantity purchased
5. InvoiceDate: Date of purchase
6. UnitPrice: Price of the corresponding product
7. CustomerID: Unique ID for each customer
8. Country: Country of purchase

Methodology

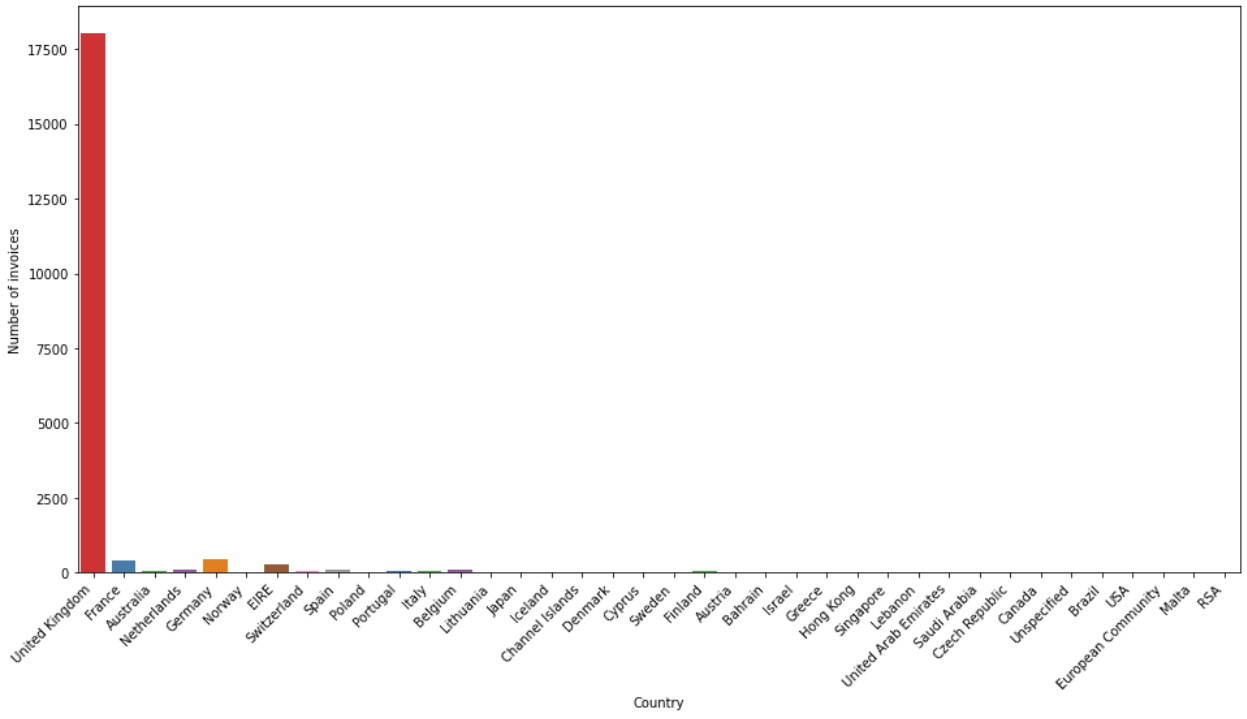


Data Cleaning

1. Keeping Data only for UK (93% by number of line items)
2. Removed negative Quantity values
3. Standardised similar Item Description and Imputed missing description. Kept the item description which occurs most number of times for a unique stock code.
4. Created TotalSalesAmount Field (Quantity ⨯ UnitPrice)
5. Removed line items that had ineffectual item description/stock code. These transactions were administrative in nature. Eg: 'Amazon Fee', 'Postage charges' etc.
6. Identified and removed outliers on Quantity, Unit Price and TotalSalesAmount

Exploratory Data Analysis

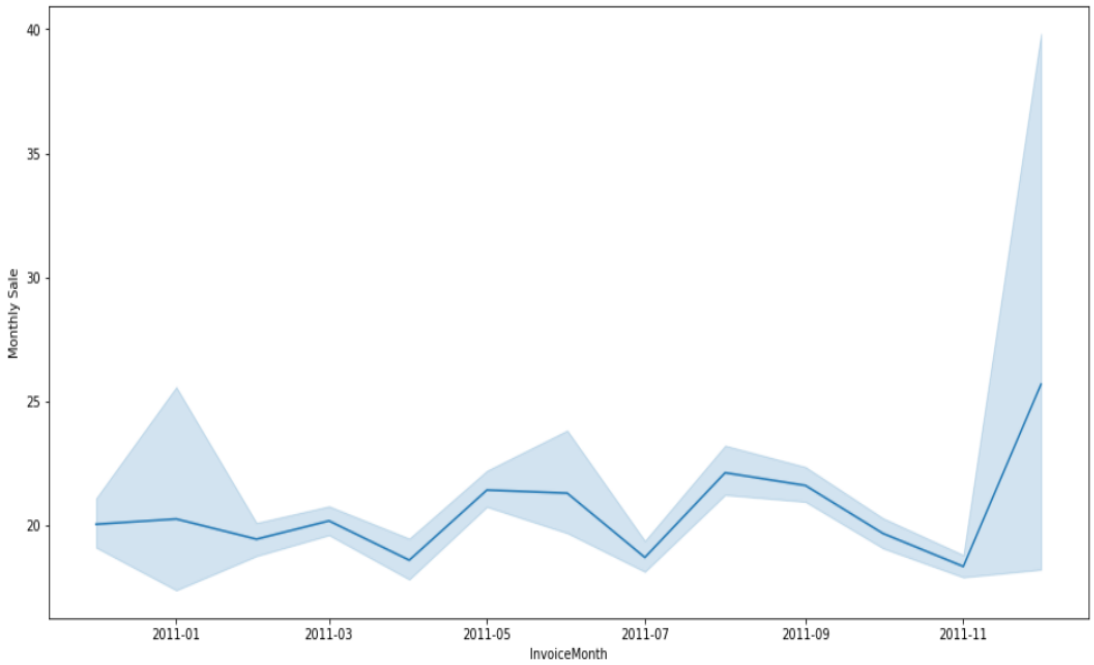
I. Number of invoices for each country



We observe that majority of the invoices are from United Kingdom – 93% by transactions.

Therefore, we will be performing our analysis only on invoices from the United Kingdom

II. Total sales amount trend from Dec-2010 to Dec-2011



We observe a periodic trough and crest in monthly sale which can be attributed to many factors that will be explored further during the analysis.

III. Products by occurrence and top grossing products

Fig 1: Top 10 Products by occurrences

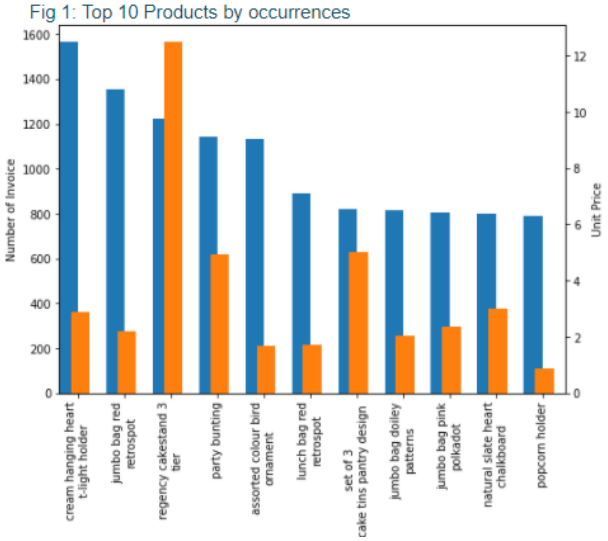
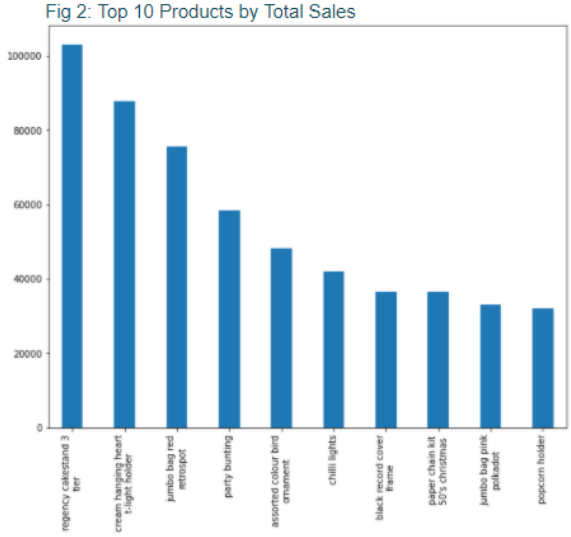


Fig 2: Top 10 Products by Total Sales



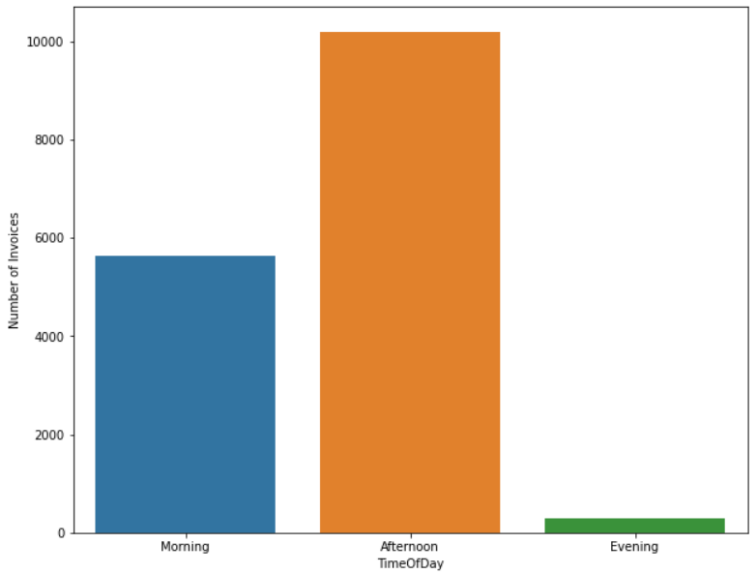
We see that the 'regency cakestand 3 tier' postage product has the highest sale amount from the table above. At the same time, another interesting observation for the same item is that though its unit price is high, it is commonly present in many invoices.

IV. Top 10 products by Quantity

StockCode	Description	Quantity	UnitPrice	CustomerID	TotalSalesAmount
23843	PAPER CRAFT , LITTLE BIRDIE	80995	2.08	16446.0	168469.60
23166	MEDIUM CERAMIC TOP STORAGE JAR	78033	367.12	3047321.0	81700.92
22197	POPCORN HOLDER	56898	1479.76	16366320.0	51334.47
84077	WORLD WAR 2 GLIDERS ASSTD DESIGNS	54951	171.91	7213288.0	13814.01
85099B	JUMBO BAG RED RETROSPOT	48371	5243.39	24700699.0	94159.81
85123A	CREAM HANGING HEART T-LIGHT HOLDER	37641	7024.49	31482068.0	104462.75
21212	PACK OF 72 RETROSPOT CAKE CASES	36396	1029.15	15848883.0	21246.45
84879	ASSORTED COLOUR BIRD ORNAMENT	36362	2542.52	21247829.0	58927.62
23084	RABBIT NIGHT LIGHT	30739	2426.97	12264242.0	66870.03
22492	MINI PAINT SET VINTAGE	26633	298.31	4780324.0	16937.82

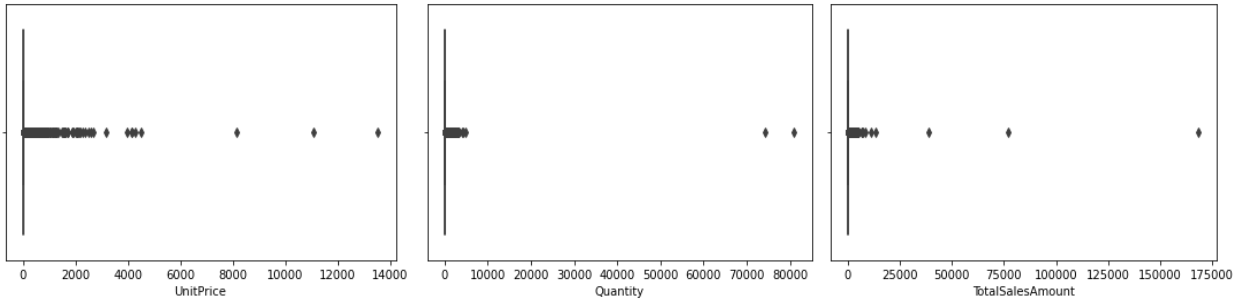
Understanding the top products helps us visualize the relationship between quantity and total sales. Here the paper craft has the highest quantity as well as the from II, we know that it is the highest sale transaction as well.

V. Transaction by time of day



According to the above distribution, we can quickly determine that the maximum transaction happens in the afternoon and the least occur during the evening. Diving deeper in the future, we can use this to design customized association rules for each time of day.

VI. Plotting outliers for unit price, quantity sold, and total sales amount



We observe there are many outliers in unit price; quantity sold, and total sales amount. We will perform further analysis on these outliers to rectify them if required.

Model Design

Below terminology will help understand the analysis better:

- Support: The fraction of transactions that contain items A and B. Range = [0,1]. For our analysis, we have used a minimum support threshold of 1%
- Confidence: The co-occurrence of items A and B given the number of times A occurs. Range = [0,1]. For our analysis, we have used a minimum confidence threshold of 40%
- Lift: It indicates the strength of a rule on the occurrence of item A and item B. The more the lift, the better is the strength. Range = [0,inf)

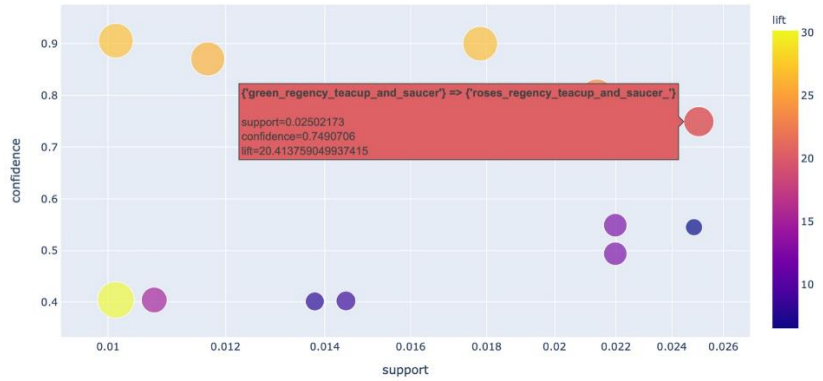
After calculating metrics and generating rules using a few threshold values, we get the association rules as follows:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
{'wood_2_drawer_cabinet_white_finish'}	{'3_drawer_antique_white_wood_cabinet'}	0.024589	0.028873	0.013536	0.550505	19.066417	0.012826	2.160485
{'3_drawer_antique_white_wood_cabinet'}	{'wood_2_drawer_cabinet_white_finish'}	0.028873	0.024589	0.013536	0.468817	19.066417	0.012826	1.836301
{'60_teatime_fairy_cake_cases'}	{'pack_of_72_retrospot_cake_cases'}	0.024589	0.037752	0.010990	0.446970	11.839551	0.010062	1.739955
{'alarm_clock_bakelike_ivory'}	{'alarm_clock_bakelike_green'}	0.020801	0.034834	0.011860	0.570149	16.367654	0.011135	2.245352
{'alarm_clock_bakelike_pink'}	{'alarm_clock_bakelike_green'}	0.022974	0.034834	0.012853	0.559459	16.060775	0.012053	2.190868

We used Apriori principles to generate frequent itemset. After running a few iterations, we found that a minimum support threshold of 1% and a minimum confidence threshold of 40% generates sufficient rules.

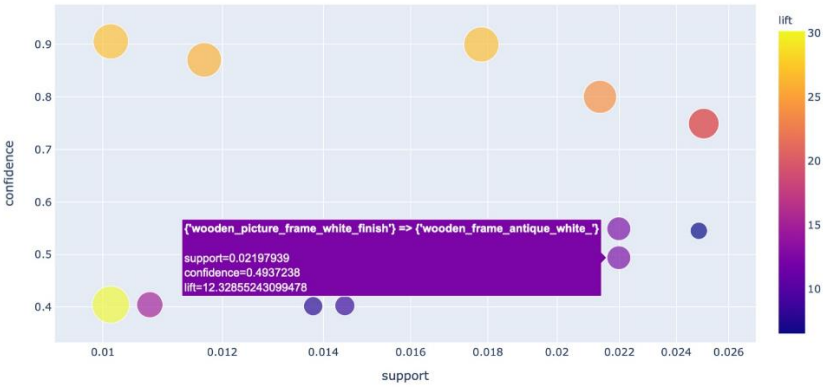
Interpreting Association Rules

a. A rule that has high support and high confidence



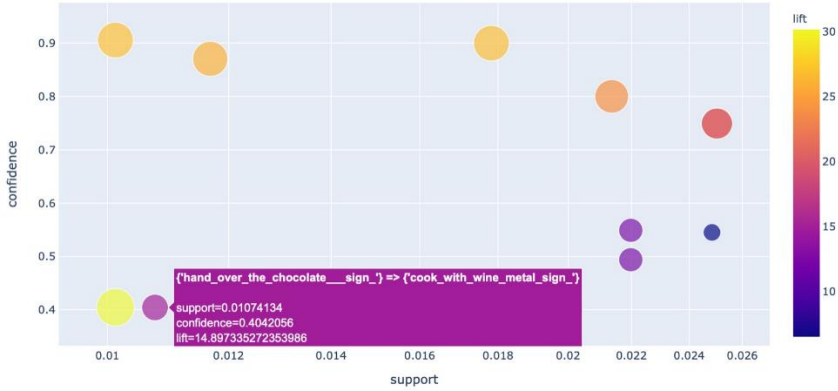
Such a rule would also not be subjectively interesting as these rules are commonly occurring and there is a high chance of them being obvious.

b. A rule that has reasonably high support but low confidence



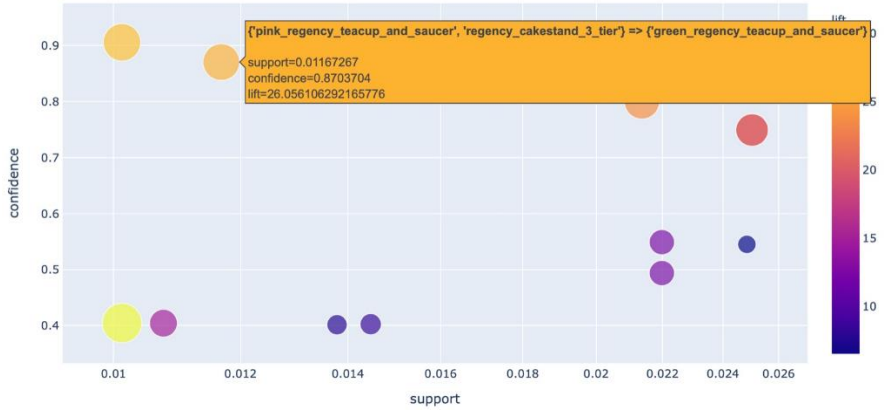
Such a rule probably will not be interesting as an itemset might frequently be occurring across transactions but lack a relationship with other itemsets

c. A rule that has low support and low confidence



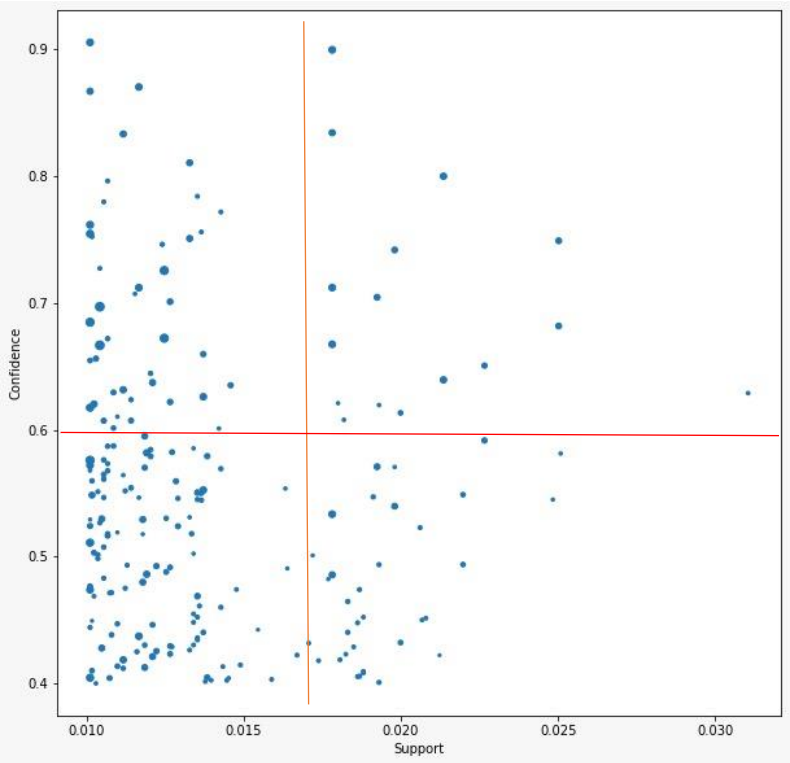
Such rules are not interesting and valuable as the itemset rarely occurs but has very few relationships.

d. A rule that has low support and high confidence

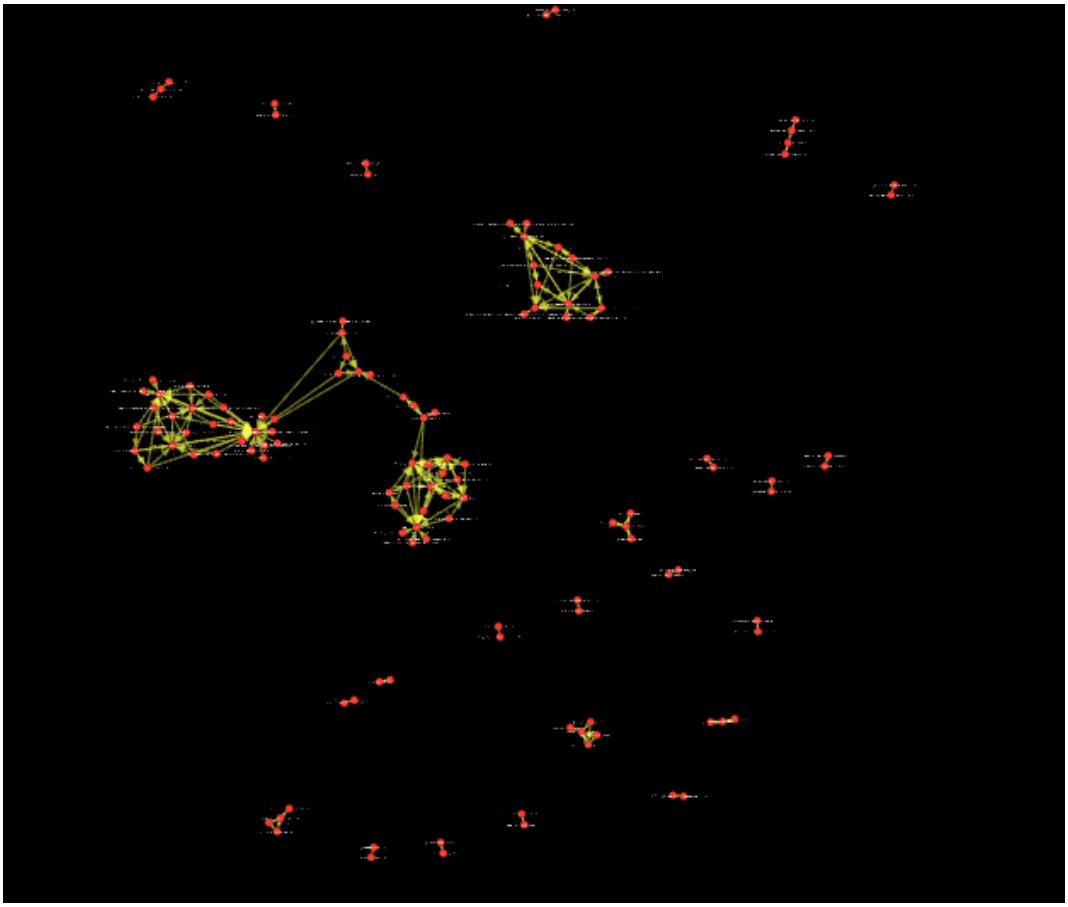


Such rules have the potential to be interesting as they occur rarely but at the same time have a strong relationship with other itemsets.

To dive deeper, we can segregate all the rules in 4 quadrants. This will help better interpret our rules to cluster them for marketing strategies.

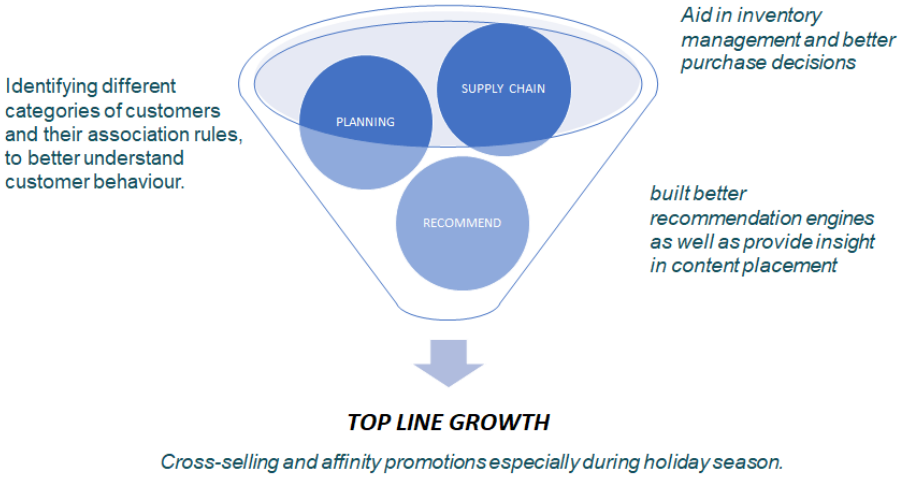


Finally, the best way to understand the relationships between all products is to explore the below network graph as below:



Note: An interactive version of this network graph is available in the project docket submitted. The Gephi file will need the open-source tool to view the graph.

Way forward and conclusion:



1. From the association rules summary, we observed that there are more than association rules. After interesting the Gephi network graph, we follow that multiple items are common as antecedent and consequent. Using this insight from the network and the plotly generated scatter plot, we can filter the rules that are more relevant to the specific business objective.
2. Also, from the analysis, we observe that multiple elements are clustered together in terms of variations of the same product. However, there are unique items in those clusters which pose a significant marketing opportunity.
3. The overall analysis can be utilized to meet the business objective, both at the bottom-line and top-line level, in inventory management and cross-selling. An example of this can be the 'regency cake stand 3 tire' product along with 'green regency tea cup and saucer'
4. We can also observe that from all the product variations available for a single type of product, we can observe that from all the available association rules, which product variations are most popular amongst the UK consumers.