**Retail Gift Shop Association Rule Mining: Apriori, Eclat, and FP-Growth Algorithm**

Binus University, School of Computer Science

Rennard Owen Tan - 2602086652, Stefani Gitawardani - 2602125161, Roderick Kangson - 2602083940

*Abstract -* The purpose of these studies is to use data mining techniques, especially association rule mining, to evaluate purchasing trends in retail stores using the Apriori, Eclat, and FP-Growth algorithms. The main goal, this study will focus on analyzing two datasets from online retailers. The findings will then highlight the most important purchasing trends and typical item pairings, such as customers tending to purchase similar items after buying a gift. The findings from these studies are used to determine marketing strategies for retailers such as bundling, cross-selling, and regional preferences. This study will highlight 3 different types of algorithms, such as FP-Growth to provide insights through in-memory processing, Eclat to find the most common sets of items, and Apriori to generate actionable association rules. The outcome of using this method will show the utility of association rule mining in enhancing decision-making and customer satisfaction in the retail industry.

*keyword - data mining, apriori, fp growth, eclat, retail, marketing*

# INTRODUCTION

In the digital age, the rise of online shopping is clearly inevitable, and currently strictly on-site businesses such as, more traditional mom-and-pop’s stores, are struggling to keep up with modernisation. That is to say, local businesses that are handled by people of the older generation often struggle with keeping track of trends and items that are popular in the market due to their unfamiliarity with the internet. And thus considering the value of these older stores, with the help of data mining, we aim to educate and suggest valuable insights that we hope may help keep their stores afloat for a longer period of time while they transition and ease their way into a more modern form of online retail. To achieve this, we will be running an analysis on two dataset obtained from an online gift store, then we will discuss the insights obtained from it, and hopefully be able to use this new knowledge to help out the local business’s gift stores.

In regards to the specific objective and goal of our project, we aim to analyze customer purchasing patterns, particularly related to gift items, and identify actionable insights to increase sales and customer satisfaction. With the goal of discovering frequently purchased item combinations that vary by country to tailor promotions and inventory management for diverse markets. But to acquire such information, we first need to explain what all the things involved with the data mining process are.

# THEORETICAL BASIS

Data mining involves extracting meaningful patterns and insights from large datasets. One of its core tasks is association rule mining, which identifies relationships among variables in a dataset. This technique is widely used in industries such as retail, healthcare, and others to uncover hidden patterns and associations that can guide decision-making.

Association and correlation mining typically aim to identify frequent itemsets within large datasets. These findings help businesses make strategic decisions, such as optimizing catalog designs, implementing cross-marketing strategies, and analyzing customer shopping behavior. Association rule algorithms are designed to generate rules with confidence values less than one. However, the number of possible association rules for a given dataset is usually vast, and a significant proportion of these rules are often of little practical value.

Data mining plays a crucial role in identifying patterns, forecasting trends, and discovering knowledge across various business domains. Its applications span virtually every industry where data is generated, making data mining one of the most important areas in databases and information systems. It is also considered one of the most promising interdisciplinary developments in Information Technology [1][2].

To apply the various methods that we will be using to analyze and mine the data, we need to first be familiar with the theoretical basis of the method itself, to start off let’s focus on association rules. Association rule mining focuses on discovering frequent itemsets and generating rules of the form[6]:

If {A}, then {B}

with measures like:

* Support: Frequency of the itemset in the dataset.
* Confidence: Likelihood that {B} is purchased when {A} is purchased.
* Lift: Strength of the rule compared to random chance.

# 

There are several association rule mining algorithms that are available today, but here are the ones that we’ve used:

1. Apriori

It is an algorithm based on mining Boolean association rules. After each set of frequent item-sets is generated, the whole database is scanned and the association rules between data are mined from the generated frequent item sets, giving us decision support. [3]

An item-set is a set of 0 or more items, and a frequent item-set is an item-set whose support is greater than the custom minimum support count [4]

Common evaluation criteria for frequent item-sets:

1) Support: It is one of the two basic parameters of association rules. It is the ratio of the number of transactions containing both x and y in all samples of dataset D to all transactions. If we have two data x and y that need to be analyzed for correlation, then the corresponding support degree is:

(1)

(2)

For example, a support rating of 28% means that “there is a 28% probability that an individual in the population will contain both X and Y”.

2) Confidence: It is the ratio of the number of transactions including x and y to the number of transactions including y, namely conditional probability.

(3)

(4)

Assuming that “52% of the terms containing X contain Y”, the confidence is 52%

3) Lift: It is the ratio of the number of transactions containing x under the premise of including y to the total number of transactions occurring in x.

(5)

Lift uses 1 as the target value to show the relationship between x and y. If the value is greater than 1, then x ⇒ y is a valid strong association rule. Conversely, x ⇒ y is an invalid strong association rule. When the value is equal to 1, however, there is a special case, that is, the x and y at independence, at the time , so . [5]

Only custom minimum support, or a combination of custom support and confidence, can determine the frequent item-sets in the database. Its core is to retrieve all frequent item-sets, and find all item-sets that are greater than or equal to the support by setting the minimum support count and iterating continuously[6]

1. Equivalence Class Clustering and bottom-up Lattice Traversal (Eclat)

The ECLAT algorithm, which stands for Equivalence Class Clustering and bottom-up Lattice Traversal, is a popular method for Association Rule Mining. It is considered a more efficient and scalable version of the Apriori algorithm. While the Apriori algorithm operates in a horizontal manner, mimicking the Breadth-First Search approach of a graph, the ECLAT algorithm employs a vertical method, similar to the Depth-First Search approach of a graph. This vertical strategy makes the ECLAT algorithm faster than the Apriori algorithm [7]. The fundamental concept of the ECLAT algorithm is to use tidset intersections to compute the support of candidate itemsets. This approach avoids generating subsets that do not exist in the prefix tree. The algorithm was originally proposed by Zaki, Parthasarathy, et al. [8].

While Eclat does not have a specific mathematical formula [9], we can create a simple representation of it by looking at how it works. Since Eclat works by using intersections, we can represent it mathematically like this:

Where TID(X) represents the list of transaction IDs where X appears, and the same applies for TID(Y), and then we simply take the intersection, the items that appear on both transaction lists to obtain the support.

1. FP-Growth

One of the currently fastest and most popular algorithms for frequent itemset mining is the FP-growth algorithm.[9] The FP-growth algorithm is based on a prefix tree representation of the database of transactions (called an FP-tree), which saves significant memory by grouping common prefixes in the data. The algorithm uses a **recursive elimination scheme** [10] to find frequent itemsets. This involves breaking the database into smaller conditional FP-trees based on items' prefixes. By focusing only on these smaller subsets, the algorithm avoids generating unnecessary subsets, making it faster and more memory-efficient than other methods. This recursive approach systematically grows frequent itemsets by mining each conditional FP-tree. Which means due to this, there are no fixed mathematical formula representations to cover FP growth as a function.

# MATERIALS AND METHODS

For the data that we’ll analyze, we used two datasets that we’ve obtained from the website, Kaggle. The first dataset is simply called “Online Retail Dataset” which can be found on the following link (<https://shorturl.at/fVrmR>). This dataset is obtained from the collection of invoices of a certain gift shop originating from the UK with branches all around the world. The size of this data is approximately 48.58 MB, and contains 541,909 entries. The key attributes for this dataset can be seen on table 1.

| Attributes | Description |
| --- | --- |
| InvoiceNo | Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation. |
| StockCode | Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product. |
| Description | Product (item) name. Nominal. |
| Quantity | The quantities of each product (item) per transaction. Numeric. |
| InvoiceDate | Invoice Date and time. Numeric, the day and time when each transaction was generated. |
| UnitPrice | Unit price. Numeric, Product price per unit in sterling. |
| CustomerID | Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer. |
| Country | Country name. Nominal, the name of the country where each customer resides. |

table. 1: Key attributes of the “Online Retail Dataset”

By performing exploratory data analysis on this particular dataset, there are a few things that we’ve found out. Surprisingly for a dataset with a huge number of entries like this, it doesn’t contain any missing values whatsoever. After that we checked for duplicated data, and there seems to be 5.225 of them, so we promptly removed them all since we deemed that those data will taint the accuracy of our model. And then we ran a summary of the dataset, and found an interesting outlier. Mainly being, the minimum value of the “Quantity” attributes appears to be on a negative value, which is not logically correct. So... how do we explain this? Since we don't have the original documentation of the data, we can only conjure up a reasonable reason on why this might happen. For one, we could reason that the customers got a discount. Because there is evidence of the fact that, all of the negative quantity entries happened on the same date of the first of December, so we can make an educated guess and say that there was a Christmas sale going on at this time or something to that extent. Another possibility is that, we can guess that the customers simply returned the products, creating a negative value to offset the original cost. This makes it so that it won't affect the calculation further down the line. Unfortunately, there was no supporting evidence to give this line of reasoning more weight, and so we decided that the "discount theory" holds more water and decided to keep it in the dataset.

Now since we have this huge amount of data entry that are separated by “Country”, we took a step further and split the analysis into several parts based on the amount of available data a particular country had. We did this in order to tailor our promotional strategies and inventory management to specific countries so we can tackle the market of that particular country better. Here is the breakdown of the top three that we chose: the UK with 356,728 entries, which represents about 65.8% of the dataset, making it the dominant contributor; Next we have Germany with 9,480 entries and accounts for approximately 1.75% of the dataset; And following close behind we got France with 8,475 entries, accounting for about 1.56% of the dataset.

The second dataset that we used is called “Online Retail Listing” and can be found on this link here (<https://shorturl.at/8j2YS>). The key attributes for this dataset can be seen on table 2.

| Attributes | Description |
| --- | --- |
| InvoiceNo | Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation. |
| StockCode | Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product. |
| Description | Product (item) name. Nominal. |
| Quantity | The quantities of each product (item) per transaction. Numeric. |
| InvoiceDate | Invoice Date and time. Numeric, the day and time when each transaction was generated. |
| UnitPrice | Unit price. Numeric, Product price per unit in sterling. |
| CustomerID | Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer. |
| Country | Country name. Nominal, the name of the country where each customer resides. |

table. 2: Key attributes of the “Online Retail Listing”

Similar to before, this dataset contains all transactions occurring for an UK-based and registered, non-store online retail between the period of 2009 to 2011. The company mainly sells unique all-occasion gift-ware and many customers are wholesalers. The size of the second dataset is nearly double the size of the previous one, with file size of 89.09 MB and a staggering amount of 1.014.269 entries. For the key attributes, they are surprisingly the exact same as the previous dataset, so please refer back to table 1 for reference.

After performing exploratory data analysis however, we can see that this dataset is much “worse” in terms of quality compared to the first one. While there are no missing values, there is a huge amount of duplicated data at 21.745 counts. Similar from before, we simply remove them all. After removing that huge amount of data, we got suspicious and analyzed the data further for outliers. Lo and behold, we find that for the “Description” attribute, it is filled with junk such as “adjustment by John", "manual", and “adjustment by Peter”. Naturally, we removed all of these irregularities from the dataset. Other than that, we also found negative values for the “Quantity” attribute, and it looks to be the same as the previous dataset as in they appear to be negative because of the application of a discount system. So, we kept that in as well.

Taking a similar approach, we decided to also split the dataset by “Country” and analyze them one by one. Here’s the breakdown for the top three: United Kingdom with 381.397 entries, EIRE or Ireland with 8.781 entries, and Germany with 8.306 entries.

And after all of the initial data preprocessing is done, we will now move on to the actual data analysis and association rule mining itself. Note that for the step-by-step process, it's actually the same for both datasets, so we’ll just be going over the first dataset to make things much more compact. The first algorithm that we used is Apriori which can be seen on figure 1.

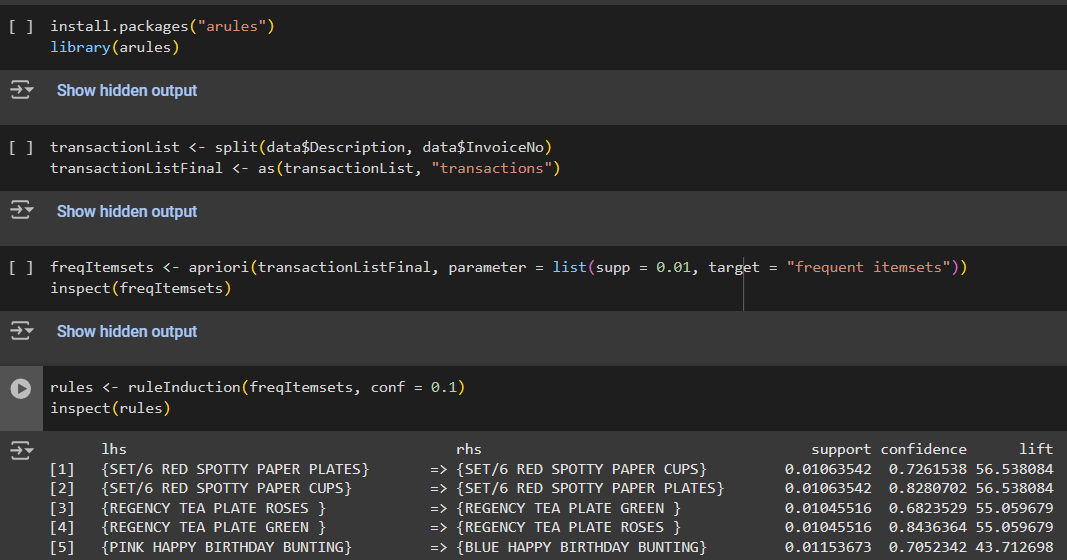


fig. 1: Apriori algorithm code implementation

The first step is to simply install and load the “arules” library. Then we create a new transaction type object called “transactionList” by grouping the description of the item by the invoice number it is listed in. After that we apply the Apriori function, pick a support value (which is the threshold for an itemset to be considered frequent), in this case we choose 0.01 or 1% of all transactions, set the target to “frequent itemsets”, and find out the frequent itemset from all of the transactions done, which are the most frequently bought items in the entirety of the transaction list. And finally, we perform rule induction to find out the association rules which is the relationship between items bought. This rule induction works by calculating the Support (How frequently X and Y occur together in the dataset.) divided by Confidence (The likelihood that Y is present given X is present).

The next algorithm is Eclat which can be seen on figure 2.

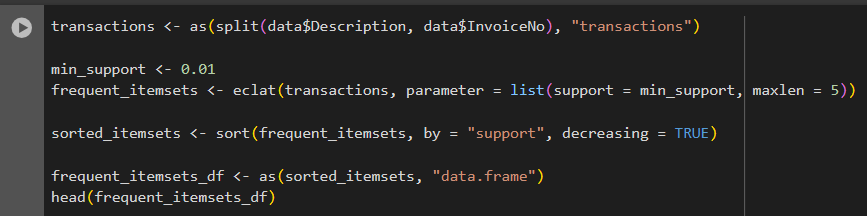


fig. 2: Eclat algorithm code implementation

After loading the “arules” library, we create a transaction type object and then simply create the frequent itemsets using the Eclat function which takes the minimum support same as Apriori, but this time we limit the size of the itemsets to a maximum of just 5 items to limit the exponential growth that this can produce. After that, we put the result into a dataframe, and simply display it. Do note that the Eclat algorithm does not produce an association rule. So it is less reliable and interpretable at least for humans than Apriori.

And the final algorithm that we use is FP-Growth, as can be seen at figure 3.

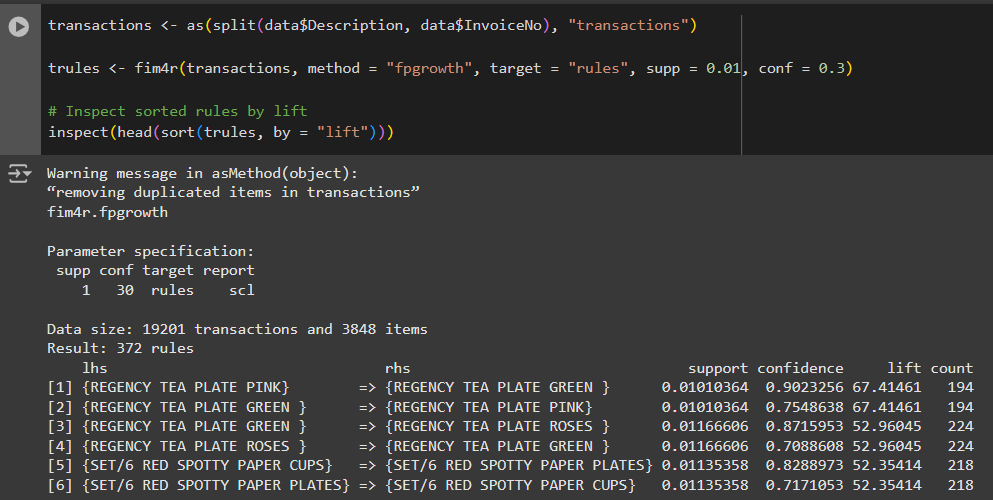


fig. 3: FP-Growth algorithm code implementation

As for FP-Growth, same as the other two, we first create the transaction object as usual. After that, we use *fim4r* or *Frequent Item Mining for R* to enable us to use the FP-Growth algorithm, in order to get the association rules. Credits to the creator of this package can be found at this Github page (<https://mhahsler.github.io/arules/>). The support is the same as the other two, but this time we also use a minimum confidence threshold of 30% to make it more accurate. After that, we simply inspect the rule and sort it by lift because the higher the lift, the stronger the correlation. After further analysis, it seems to not be ideal, so we instead run it back and sort it by support in the final build.

As for the analysis result and the continuation of the Knowledge Discovery (KDD) process post data cleaning, integration, and selection that we just went through, it will all be explained in the following section. But it is important to note that, we will be skipping the data transformation part of KDD like mapping, because we simply do not need it for our dataset due to the nature of it already being “structurally correct” as in all of the data types are already what they’re supposed to be. In any case, let’s move on to the next section.

IV. RESULTS

Now, here are the results of our data mining endeavour, applying several algorithms as stated previously to two datasets that we split into several parts as explained in the previous section. Starting off, the following table below is the result obtained from applying the Apriori algorithm to the first dataset, “Online Retail Dataset”.

TABLE I

Top 3 Association Rules of items bought Globally, and specifically at the United Kingdom, France, and Germany via Apriori

| Range | lhs | rhs | support | confidence |
| --- | --- | --- | --- | --- |
| Globally | {SET/6 RED SPOTTY PAPER PLATES} | => {SET/6 RED SPOTTY PAPER CUPS} | 0.01063542 | 0.7261538 |
| {SET/6 RED SPOTTY PAPER CUPS} | => {SET/6 RED SPOTTY PAPER PLATES} | 0.01063542 | 0.8280702 |
| {REGENCY TEA PLATE ROSES } | => {REGENCY TEA PLATE GREEN } | 0.01045516 | 0.6823529 |
| United Kingdom | {PINK HAPPY BIRTHDAY BUNTING} | => {BLUE HAPPY BIRTHDAY BUNTING} | 0.01097850 | 0.6898734 |
| {BLUE HAPPY BIRTHDAY BUNTING} | => {PINK HAPPY BIRTHDAY BUNTING} | 0.01097850 | 0.7147541 |
| {ALARM CLOCK BAKELIKE RED } | => {ALARM CLOCK BAKELIKE ORANGE} | 0.01097850 | 0.2770013 |
| France | {SET 3 PAPER VINTAGE CHICK PAPER EGG} | => {POSTAGE} | 0.01310044 | 0.8571429 |
| {SET OF 3 BUTTERFLY COOKIE CUTTERS} | => {POSTAGE} | 0.01091703 | 0.8333333 |
| {EDWARDIAN PARASOL PINK} | => {POSTAGE} | 0.01091703 | 1 |
| Germany | {3 HOOK PHOTO SHELF ANTIQUE WHITE} | => {POSTAGE} | 0.01160862 | 0.875 |
| {PANTRY CHOPPING BOARD} | => {POSTAGE} | 0.01160862 | 1 |
| {CERAMIC CAKE DESIGN SPOTTED PLATE} | => {POSTAGE} | 0.01492537 | 1 |

Regarding Table I, lhs & rhs refers to left-hand side and right-hand side respectively, which represents the “antecedent” as in “if this item is purchased…” and the “consequent” as in “... then this item will also most likely be purchased as well”. Also to obtain these association rules, we used a support count of 0.01 because this appears to be the most optimal, as we found out by trying out other amounts such as 0.02 and 0.03. And aside from that, we picked a minimum confidence level of 10% to ensure the quality of the association rules generated.

So, based on the insight obtained from Table I, we can conjure up a few marketing strategies to further boost the sales of these particular sets of items further. For stores globally, we can implement a bundling deal. As we can see, the two most prominent items bought together are both paper cups and paper plates, meaning that we can offer a "Party Set" package that includes both the plates and cups together at a discounted price. We can also do cross-sell recommendations. So, for example on the product page for the paper plates, recommend paper cups through a pop-up or a sort of “customer like you also bought…” feature. We can also encourage customers to buy in bulk, “Buy 2 packs of cups and get a discount on the plates", which will be especially effective for customers preparing larger gatherings. For stores located in the United Kingdom, we see that the two birthday party buntings are popular and often bought together, so we can also bundle them together in a “Party bundle” set and market it together. Similarly, we can do the same for the *bakelike* alarm clocks color variants. And as for stores located in Germany and France, it seems that customers that buy gifts from there also buy a postage afterwards to send the gifts to their intended recipient. So, we can simply put the top 3 items as shown in the association rules near the postage stand so people can buy adhesive tokens and stamps immediately afterwards. Better yet, have a mailbox directly outside the store, making it convenient for the customers.

TABLE II

Top 3 frequent itemsets bought Globally, and specifically at the United Kingdom, France, and Germany via Eclat

| Range | items | support |  | Range | items | support |
| --- | --- | --- | --- | --- | --- | --- |
| Globally | {WHITE HANGING HEART T-LIGHT HOLDER} | 0.09071654 |  | France | {POSTAGE} | 0.6790393 |
| {REGENCY CAKESTAND 3 TIER} | 0.08490311 |  | {RABBIT NIGHT LIGHT} | 0.1615721 |
| {JUMBO BAG RED RETROSPOT} | 0.07404236 |  | {RED TOADSTOOL LED NIGHT LIGHT} | 0.1550218 |
| United Kingdom | {WHITE HANGING HEART T-LIGHT HOLDER} | 0.09689278 |  | Germany | {POSTAGE} | 0.6351575 |
| {REGENCY CAKESTAND 3 TIER} | 0.07841064 |  | {ROUND SNACK BOXES SET OF4 WOODLAND} | 0.1973466 |
| {JUMBO BAG RED RETROSPOT} | 0.07488543 |  | {POSTAGE, ROUND SNACK BOXES SET OF4 WOODLAND } | 0.1708126 |

Table II shows us the most prominent frequent itemsets that were present in a majority of the transactions recorded in the dataset. Looking at this, we can see that only 1 item appears extremely frequently for every store in every country except in Germany where postage and round snack boxes being bought together is popular enough to be on the top 3 frequent itemsets. While not as insightful as the previous one, these results provided by Eclat can help determine the popularity of an item, and marketing-wise, we can safely stock and promote more of these items for higher profit as the data does indeed show that there are high demands for them.

TABLE III

Top 3 Association Rules of items bought Globally, and specifically at the United Kingdom, France, and Germany via FP Growth

| Range | lhs | rhs | support | confidence | lift |
| --- | --- | --- | --- | --- | --- |
| Globally | {GREEN REGENCY TEACUP AND SAUCER} | => {ROSES REGENCY TEACUP AND SAUCER } | 0.02892235 | 0.7604167 | 17.853876 |
| {ROSES REGENCY TEACUP AND SAUCER } | => {GREEN REGENCY TEACUP AND SAUCER} | 0.02892235 | 0.6790698 | 17.853876 |
| {LUNCH BAG PINK POLKADOT} | => {LUNCH BAG RED RETROSPOT} | 0.02727153 | 0.5558546 | 8.318046 |
| United Kingdom | {GREEN REGENCY TEACUP AND SAUCER} | => {ROSES REGENCY TEACUP AND SAUCER } | 0.02868069 | 0.7558140 | 18.126205 |
| {ROSES REGENCY TEACUP AND SAUCER } | => {GREEN REGENCY TEACUP AND SAUCER} | 0.02868069 | 0.6878307 | 18.126205 |
| {JUMBO BAG PINK POLKADOT} | => {JUMBO BAG RED RETROSPOT} | 0.02787175 | 0.6213115 | 8.123648 |
| France | {RED TOADSTOOL LED NIGHT LIGHT} | => {POSTAGE} | 0.1433333 | 0.9347826 | 1.354757 |
| {PLASTERS IN TIN CIRCUS PARADE } | => {POSTAGE} | 0.1333333 | 0.8510638 | 1.233426 |
| {SET/6 RED SPOTTY PAPER PLATES} | => {SET/6 RED SPOTTY PAPER CUPS} | 0.1333333 | 0.9523810 | 6.493506 |
| Germany | {ROUND SNACK BOXES SET OF4 WOODLAND } | => {POSTAGE} | 0.1811414 | 0.8295455 | 1.337227 |
| {ROUND SNACK BOXES SET OF 4 FRUITS } | => {POSTAGE} | 0.1290323 | 0.8666667 | 1.397067 |
| {ROUND SNACK BOXES SET OF 4 FRUITS } | => {ROUND SNACK BOXES SET OF4 WOODLAND } | 0.1240695 | 0.8333333 | 3.816288 |

Table III shows us the association rules the same way Table I did previously. We used a support count of 0.01 for every range except for France which we used 0.05 support on. This is necessary because if we don't use a higher support count, the process will go on and on until it produces a stalemate condition where it will be insufficient for the system’s RAM to handle and crash the Google Colab notebook that we’re doing this on. Also, we used a 30% minimal confidence rate for selection of the A-Rules as opposed to the 10% previously on Apriori because we feel that it is more reliable and statistically significant, and we want to have a point of comparison between FP Growth and Apriori. As for the lift, we decided to still show the lift value because originally we sorted it out using lift, but after further analysis, we notice that while yes higher lifts means higher correlation, it doesn't really determine the quality of the A-Rules itself, so we decided to drop the original plan, and sort it by support and confidence which is much more reliable. Now, as for the insights that we can obtain from Table III, similar to Table 1, we can apply all of the marketing strategies that we’ve already talked about previously. But since the items are different, we can just create more bundles of these items in addition to the bundles we created for Table I, which logically would improve the store’s profit more. Other than that, there’s not much more to say about the results from dataset 1.

Which now means, we’ll talk about the result of dataset 2’s analysis. Since we’re going to be using the same methods as dataset 1, please do note that we won’t be explaining the attributes and the reasoning behind the parameters selected a second time, as it is already explained previously. To start off, the following table below is the result obtained from applying the Apriori algorithm to the second dataset, “Online Retail Listing”.

TABLE IV

Top 3 Association Rules of items bought Globally, and specifically at the United Kingdom, Ireland, and Germany via Apriori

| Range | lhs | rhs | support | confidence |
| --- | --- | --- | --- | --- |
| Globally | {SET/6 RED SPOTTY PAPER PLATES} | => {SET/6 RED SPOTTY PAPER CUPS} | 0.01118231 | 0.6881443 |
| {SET/6 RED SPOTTY PAPER CUPS} | => {SET/6 RED SPOTTY PAPER PLATES} | 0.01118231 | 0.7563739 |
| {EDWARDIAN PARASOL RED} | => {EDWARDIAN PARASOL BLACK} | 0.01114043 | 0.6927083 |
| United Kingdom | {EDWARDIAN PARASOL RED} | => {EDWARDIAN PARASOL BLACK} | 0.01043526 | 0.6951220 |
| {EDWARDIAN PARASOL BLACK} | => {EDWARDIAN PARASOL RED} | 0.01043526 | 0.4840764 |
| {POPPY'S PLAYHOUSE KITCHEN} | => {POPPY'S PLAYHOUSE BEDROOM } | 0.01048103 | 0.7979094 |
| Ireland | {VANILLA SCENT CANDLE JEWELLED BOX} | => {FELTCRAFT CUSHION RABBIT} | 0.01086957 | 0.6666667 |
| {FELTCRAFT CUSHION RABBIT} | => {VANILLA SCENT CANDLE JEWELLED BOX} | 0.01086957 | 0.4444444 |
| {ROUND CAKE TIN VINTAGE RED} | => {BISCUIT TIN VINTAGE RED} | 0.01086957 | 0.4444444 |
| Germany | {RED/WHITE DOTS RUFFLED UMBRELLA} | => {EDWARDIAN PARASOL RED} | 0.01353965 | 0.7777778 |
| {EDWARDIAN PARASOL RED} | => {RED/WHITE DOTS RUFFLED UMBRELLA} | 0.01353965 | 0.5 |
| {RED/WHITE DOTS RUFFLED UMBRELLA} | => {POSTAGE} | 0.01353965 | 0.7777778 |

By looking at Table IV, we can see that most of the items in the association rules are a different color variant from the item bought before, so our previous marketing tacting of bundling these items together will also apply spectacularly here. For example: bundling the paper cutleries together in a “party set”; or the parasols and umbrellas into something like a “rainy day sale”; and so on and so forth. Similarly, putting these items in close proximity with each other will also work wonderfully. Example being putting the Jewelled Box and Cushion Rabbit near each other, or putting both of the Poppy’s Playhouse toys in the same aisle as one another.

TABLE V

Top 3 frequent itemsets bought Globally, and specifically at the United Kingdom, Ireland, and Germany via Eclat

| Range | items | support |  | Range | items | support |
| --- | --- | --- | --- | --- | --- | --- |
| Globally | {WHITE HANGING HEART T-LIGHT HOLDER} | 0.16512284 |  | Ireland | - | - |
| {SCOTTIE DOG HOT WATER BOTTLE} | 0.09504631 |  | No frequent itemsets can be generated | - |
| {STRAWBERRY CERAMIC TRINKET BOX} | 0.08175594 |  | - | - |
| United Kingdom | {WHITE HANGING HEART T-LIGHT HOLDER} | 0.17151541 |  | Germany | - | - |
| {SCOTTIE DOG HOT WATER BOTTLE} | 0.09726444 |  | No frequent itemsets can be generated | - |
| {STRAWBERRY CERAMIC TRINKET BOX} | 0.08119844 |  | - | - |

Table V is the result of applying Eclat to the second dataset. An interesting thing that is prevalent in this, is that for the data range coming from Ireland and Germany, no frequent itemsets can be generated. This can be caused by a variety of things, but after messing with the parameters and even doing some more EDA, we concluded that the dataset is just not robust enough to be run through with Eclat. But that doesn’t mean this does not give us insights, for example we can reasonably suggest that customers in Ireland and Germany bought unique items that are not based on trends meaning they might tend to browse more, which may help in supporting the longevity of the stores in these country, as they can stock up items without worrying too much about the unavailability problem, where customers come, ask for the specific item, and just leave when they don't find it.

TABLE VI

Top 3 Association Rules of items bought Globally, and specifically at the United Kingdom, Ireland, and Germany via FP Growth

| Range | lhs | rhs | support | confidence | lift |
| --- | --- | --- | --- | --- | --- |
| Globally | {RED HANGING HEART T-LIGHT HOLDER} | => {WHITE HANGING HEART T-LIGHT HOLDER} | 0.03118555 | 0.7193149 | 5.452668 |
| {SWEETHEART CERAMIC TRINKET BOX} | => {STRAWBERRY CERAMIC TRINKET BOX} | 0.02726673 | 0.7580275 | 12.316423 |
| {STRAWBERRY CERAMIC TRINKET BOX} | => {SWEETHEART CERAMIC TRINKET BOX} | 0.02726673 | 0.4430295 | 12.316423 |
| United Kingdom | {RED HANGING HEART T-LIGHT HOLDER} | => {WHITE HANGING HEART T-LIGHT HOLDER} | 0.03273567 | 0.7222777 | 5.159193 |
| {SWEETHEART CERAMIC TRINKET BOX} | => {STRAWBERRY CERAMIC TRINKET BOX} | 0.02784569 | 0.7518337 | 12.067587 |
| {STRAWBERRY CERAMIC TRINKET BOX} | => {SWEETHEART CERAMIC TRINKET BOX} | 0.02784569 | 0.4469477 | 12.067587 |
| Ireland | {PACK OF 72 RETRO SPOT CAKE CASES} | => {60 TEATIME FAIRY CAKE CASES} | 0.06619385 | 0.6511628 | 5.197016 |
| {60 TEATIME FAIRY CAKE CASES} | => {PACK OF 72 RETRO SPOT CAKE CASES} | 0.06619385 | 0.5283019 | 5.197016 |
| {PACK OF 60 PINK PAISLEY CAKE CASES} | => {60 TEATIME FAIRY CAKE CASES} | 0.05673759 | 0.6666667 | 5.320755 |
| Germany | {ROUND SNACK BOXES SET OF4 WOODLAND } | => {POSTAGE} | 0.1754717 | 0.8857143 | 1.624320 |
| {POSTAGE} | => {ROUND SNACK BOXES SET OF4 WOODLAND } | 0.1754717 | 0.3217993 | 1.624320 |
| {WOODLAND CHARLOTTE BAG} | => {POSTAGE} | 0.1264151 | 0.8701299 | 1.595740 |

Table VI shows us the association rules generated from dataset 2 with the help of FP Growth. Note that, coincidentally, similar to France in Table III, no association rules can be generated with a support count of 0.01 for the data range coming from Ireland, and so we decided to tinker around and finally came at the solution of support count 0.03 to generate the a-rules. Aside from that, the insights that we receive from this is essentially the same as previous association rules tables, so to avoid repetition, I won’t be explaining it again. And thus concludes, the results obtained from our analysis.

# V. DISCUSSION AND SUMMARY

1. Discussion

The use of association rule mining for both of the datasets revealed the significance of the customer purchasing patterns. Using Apriori, Eclat, and FP-Growth, we can derive meaningful association rules and frequent itemsets that help for decision-making in the retail industry.

The performance and insights of the Apriori algorithm generate strong association rules that help inform the bundling strategies and cross-selling. For instance, items like Red Spotty Paper Plates and Cups provide a high correlation in the Global dataset. For some country-specific analyses like Germany and France, postage is the most frequent consequence that indicates the need to integrate gift and mailing services. The minimum support of 1% and confidence of 10% has been proven that it can ensure the actionable rules without overloading the results with less relevant patterns.

As for the Eclat, it is able to offer meaningful general-purpose frequent item sets. However, it lacked the interpretability of Apriori because there was no rule generation. Such as the White Hanging Heart T-Light Holder for Global dataset and popular single items or combinations like snack boxes in Germany dataset, Eclat allows combinations and is very helpful to devise the most frequently sold items. Its vertical tidset-based computation gives good processing efficiency but it has restrictions to use for single frequency analyses.

As for the FP-Growth, it can gather the most common itemsets without generating a candidate due to faster processing so as to give more detail about the customer’s liking. The outcomes contribute to reinforcing the results obtained through the Apriori algorithm for the confirmation of the joint and marketing strategies that have been suggested.

1. Summary

As for the conclusion, this study demonstrated how well association rule mining works to find significant patterns in retail datasets. Proposing focused marketing strategies, such as grouping related products together, putting complimentary products nearby, and utilizing cross-selling tactics, was made possible thanks in large part to the data gathered. These tactics have the ability to greatly improve retail operations and are in line with developments in consumer behavior.

The strengths and weaknesses of Apriori, Eclat, and FP-Growth were highlighted by the comparison study. While Eclat offered effective frequent itemset recognition and Apriori was superior at rule generation, FP-Growth's thorough and effective methodology integrated the advantages of both. When combining these algorithms, it provides a strong tool to analyze retail data, opening the door to better customer happiness and well-informed decision-making.

# REFERENCES

1. <https://calibrate.thearena.ai/infohub/data-mining#:~:text=Data%20mining%20is%20the%20process%20of%20discovering%20meaningful,can%20be%20used%20for%20decision-making%2C%20prediction%2C%20and%20optimization>. [Accessed Dec 23 2024].
2. <https://www.researchgate.net/publication/49616224_Data_mining_techniques_and_applications> [Accessed Dec 23 2024].
3. <https://www.researchgate.net/publication/351168385_Research_and_Case_Analysis_of_Apriori_Algorithm_Based_on_Mining_Frequent_Item-Sets> [Accessed Dec 23 2024].
4. Shabtay, L., Fournier-Viger, P., Yaari, R., & Dattner, I. (2021). A Guided FP-Growth Algorithm for Mining Multitude-Targeted Item-Sets and Class balanced Data. Information Sciences, 553, 353-375. <https://doi.org/10.1016/j.ins.2020.10.020> [Accessed Dec 23 2024].
5. Zhou, J. Z., Liu, J., Yu, C. Y. (2010). Research and Application of Data Mining Based on Web Log. Science, Technology and Engineering, 10, 2762-2766. [Accessed Dec 23 2024].
6. <https://r-statistics.co/Association-Mining-With-R.html> [Accessed Dec 23 2024].
7. <https://www.geeksforgeeks.org/ml-eclat-algorithm/> [Accessed Dec 23 2024].
8. <https://www.cs.rpi.edu/~zaki/PaperDir/DMKD97.pdf> [Accessed Dec 23 2024].
9. J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In: Proc. Conf. on theManagement of Data (SIGMOD’00, Dallas, TX).ACM Press, New York, NY, USA 2000 [Accessed Dec 23 2024].
10. *(PDF) An Implementation of the FP-growth Algorithm*. Available from:<https://www.researchgate.net/publication/228913454_An_Implementation_of_the_FP-growth_Algorithm> [accessed Dec 23 2024].

Important Links

R-Code collab:

1. Dataset 1
   1. Apriori : <https://colab.research.google.com/drive/1REyNN7ai3jFd48_KWB61gssSV2TfbmJp?usp=sharing>
   2. Eclat : <https://colab.research.google.com/drive/1NABvVj8-31PRndgOtveM40WR1oFB3A_g?usp=sharing>
   3. FPGrowth : <https://colab.research.google.com/drive/1LgLH-cgCSJYFXsmVbccr9T5WiC9jIgqv?usp=sharing>
2. Dataset 2
   1. Apriori : <https://colab.research.google.com/drive/1eeWbrwltukmQlbuL2Aw7WN3Z8qkWKPJq#scrollTo=J0dji0IC18-H>
   2. Eclat : <https://colab.research.google.com/drive/1wsyiTV80DwwQcvHvoaddxJiZBm80HEZq?usp=sharing>
   3. FPGrowth : <https://colab.research.google.com/drive/1CTdGkKPrkg7q5nWgO3MmWS421PZZcZkV?usp=sharing>
3. Link to the kaggle page for the datasets:
   1. Dataset1 : <https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset/data>
   2. Dataset2 : <https://www.kaggle.com/datasets/ilkeryildiz/online-retail-listing>