An-Najah National University

Department of Artificial Intelligence

Homework 1

Data Mining and Analysis

Submitted By: Aladdin Husni Odeh

Submitted To: Dr. Anas Toma

Contents

1. Introduction

1.1. Objectives

In this report, we're going to analyze a dataset for an online store. The first step is preprocessing the data to make it ready for analysis and association rules generation.

The next step after cleaning the data is analyzing it. We want to see what the data tells us about customer trends and which products are popular in different countries.

Finally, we'll generate association rules to find out which products are often bought together. This is important because it can help the store know what products to focus on.

1.2. Dataset overview

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. [1]

The dataset has 541909 records with the following features:

Variable Name	Role	Type	Description
InvoiceNo	ID	Categorical	A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation
StockCode	ID	Categorical	A 5-digit integral number uniquely assigned to each distinct product
Description	Feature	Categorical	product name
Quantity	Feature	Integer	the quantities of each product (item) per transaction
InvoiceDate	Feature	Date	the day and time when each transaction was generated
UnitPrice	Feature	Continuous	product price per unit
CustomerID	Feature	Categorical	A 5-digit integral number uniquely assigned to each customer
Country	Feature	Categorical	The name of the country where each customer resides

Table 1: Dataset description from https://archive.ics.uci.edu/dataset/352/online+retail

2. Data Preprocessing and Cleaning

2.1. Initial Assessment

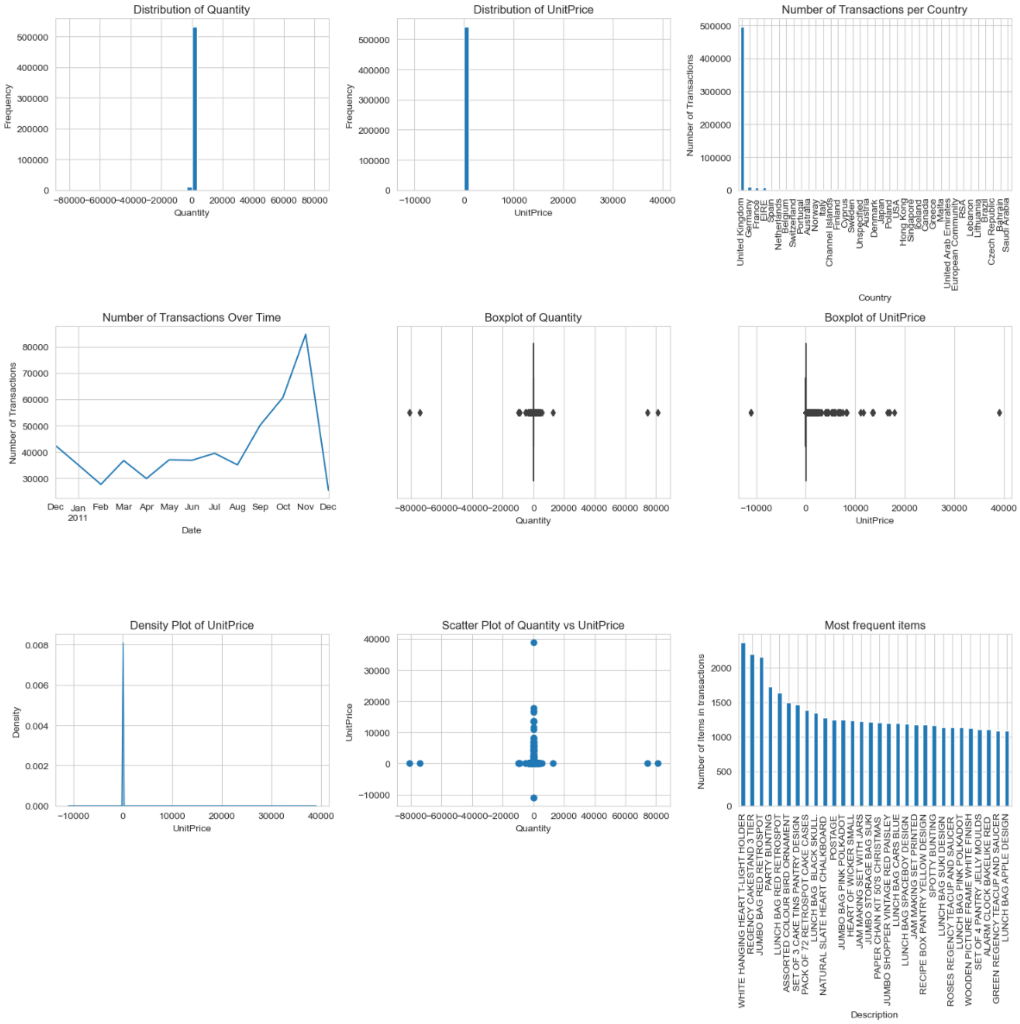


Figure 1: Data initial assessment

The initial step involved creating multiple histogram plots to observe the distribution of each feature. Additionally, box plots were utilized for numerical features to aid in understanding the data range and identifying any outliers.

From this preliminary examination, several key observations emerged:

- The data for both 'quantity' and 'unit price' showed widespread, indicating the likely presence of outliers. This was further supported by the insights gained from the box plot visualizations.
- A considerable amount of the data is from the United Kingdom, making it a significant region in this dataset.
- The dataset covers transactions from December 2010 to November 2011, offering a comprehensive view of retail activity over this period.
- Interestingly, 'POSTAGE' appears frequently as a product, despite not being an actual item for sale.

2.2. Cleaning Process

2.2.1. Missing customer Ids handling

The dataset contained approximately 135,000 records with missing customer IDs. An initial attempt to address this involved using forward filling, applied after grouping records by their invoice numbers. However, this approach proved ineffective, indicating that the customer ID was missing for all transactions under the same invoice.

Given that CustomerId is not a critical element for this study and considering there's no feasible method to generate its value, it was decided not to discard these records. Instead, a practical solution was adopted: filling in the missing customer IDs with a placeholder, labeled 'UnknownId'.

2.2.2. Stock code and description handling

```
Number of unique descriptions 4211
Number of unique StockCode 4070
Description
check                                  146
?                                      47
damages                                43
                    ...
GARLAND WOODEN HAPPY EASTER            1
GARLAND, MAGIC GARDEN 1.8M             1
```

Output 1: Data Analysis - Cell #4

We began by examining the relationship between the stock code and product description. Ideally, each stock code should match only one product description. Our analysis found that this was true for most of the product descriptions (4051 of them). However, some descriptions were linked to more than one stock code. To resolve this, we decided to unify the stock codes for each description by selecting the first stock code listed.

After ensuring a one-to-one match between stock code and description, we encountered another issue: 1454 records had a stock code but no corresponding description. To address this, we created a universal map linking each stock code to its description and filled in the missing descriptions accordingly.

In the end, there were 122 records that still had missing descriptions. We chose to remove these records from our analysis, as they are essential for the accuracy of our study.

Our initial findings revealed incorrect values appearing as single-word descriptions in the data. To investigate this further, we displayed all product descriptions that were just one word long. This revealed that all single-word descriptions, except for one product labeled 'SOMBRERO', represented incorrect entries, and needed cleaning.

```
['POSTAGE' 'CARRIAGE' 'Manual' 'amazon' '?' 'SOMBRERO' 'check' 'damages' 'DAMAGED' 'faulty' 'Found' 'found'
'counted' 'Dotcom' 'samples/damages' 'Amazon' 'showroom' 'MIA' 'Adjustment' 'damages/display' 'broken' '?lost'
 'damages?' 'cracked' 'Damaged' 'SAMPLES' 'returned' 'damaged' 'Display' 'Missing' 'adjustment' 'adjust' 'crushed'
'samples' 'mailout' 'wet/rusty'  'damages/dotcom?' 'smashed' 'missing' 'FOUND' 'dotcom' 'FBA' 'ebay'
'Damages/samples' '?display?' '?missing' 'Crushed' 'test' '??' 'Dagamed' 'WET/MOULDY' 'mouldy' 're-adjustment' ...]
```

Output 2: Data analysis - Cell #7

Following this step, we achieved a clean and accurate mapping of stock codes to product descriptions.

2.2.3. Numeric values handling

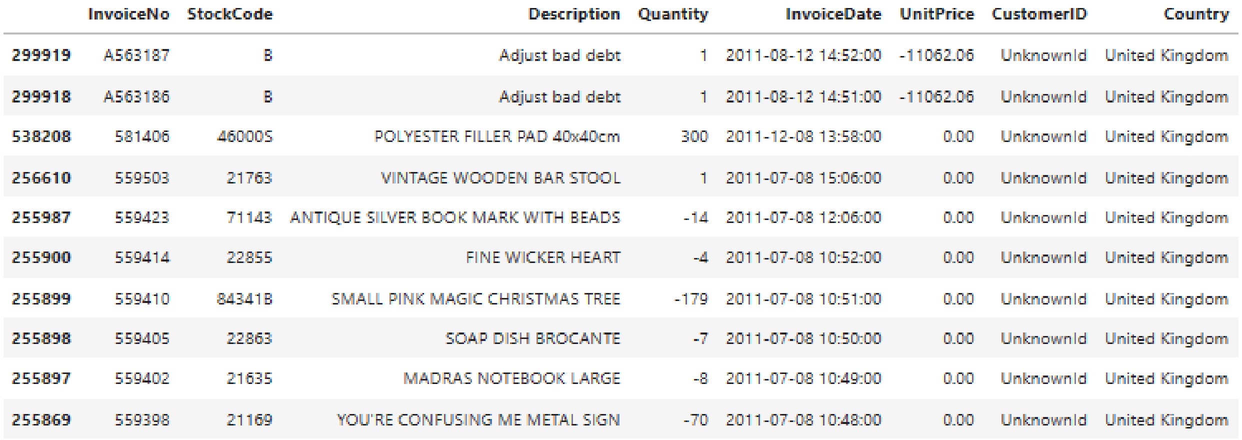
2.2.3.1. Unit price preprocessing

When we first looked at the data, we noticed some extreme outliers. To understand these better, we examined the descriptions of these items. After several attempts, we discovered that all transactions with a unit price over 650 were for invalid items and could be removed from the dataset.

```
array(['DOTCOM POSTAGE', 'AMAZON FEE', 'Bank Charges', 'Adjust bad debt',
    'CRUK Commission'], dtype=object)
```

Output 3: Data analysis - Cell #9

The initial assessment also highlighted some issues with prices, including negative and zero values:



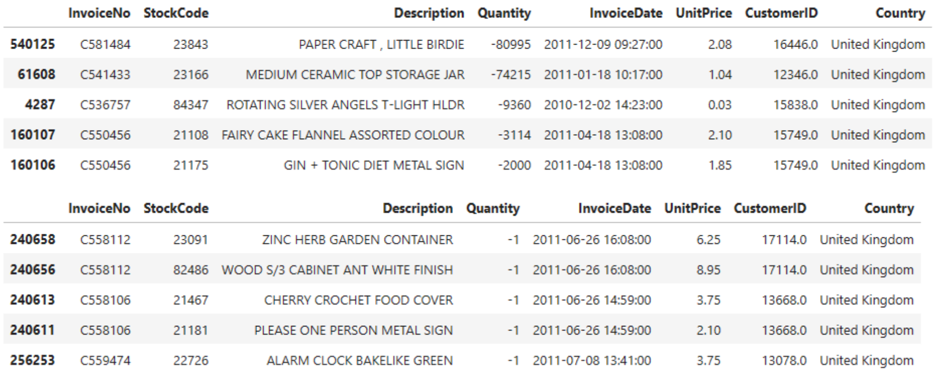
	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
299919	A563187	B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	UnknownId	United Kingdom
299918	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	UnknownId	United Kingdom
538208	581406	46000S	POLYESTER FILLER PAD 40x40cm	300	2011-12-08 13:58:00	0.00	UnknownId	United Kingdom
256610	559503	21763	VINTAGE WOODEN BAR STOOL	1	2011-07-08 15:06:00	0.00	UnknownId	United Kingdom
255987	559423	71143	ANTIQUE SILVER BOOK MARK WITH BEADS	-14	2011-07-08 12:06:00	0.00	UnknownId	United Kingdom
255900	559414	22855	FINE WICKER HEART	-4	2011-07-08 10:52:00	0.00	UnknownId	United Kingdom
255899	559410	84341B	SMALL PINK MAGIC CHRISTMAS TREE	-179	2011-07-08 10:51:00	0.00	UnknownId	United Kingdom
255898	559405	22863	SOAP DISH BROCANTE	-7	2011-07-08 10:50:00	0.00	UnknownId	United Kingdom
255897	559402	21635	MADRAS NOTEBOOK LARGE	-8	2011-07-08 10:49:00	0.00	UnknownId	United Kingdom
255869	559398	21169	YOU'RE CONFUSING ME METAL SIGN	-70	2011-07-08 10:48:00	0.00	UnknownId	United Kingdom

Output 4: Data analysis - Cell #10

- Transactions with very high unit prices turned out to be for non-product entries, so we decided they should be removed.
- There were a couple of entries with high negative prices labeled as "Adjust bad debt." We chose to remove these as well.
- We also found several instances where zero prices were associated with negative quantities. These were removed for data accuracy.

2.2.3.2. Quantity preprocessing

Our initial review of the data revealed many negative quantities, some of which were extremely high. To address this, we started by looking into these negative quantities:



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
540125	C581484	23843	PAPER CRAFT , LITTLE BIRDIE	-80995	2011-12-09 09:27:00	2.08	16446.0	United Kingdom
61608	C541433	23166	MEDIUM CERAMIC TOP STORAGE JAR	-74215	2011-01-18 10:17:00	1.04	12346.0	United Kingdom
4287	C536757	84347	ROTATING SILVER ANGELS T-LIGHT HLDR	-9360	2010-12-02 14:23:00	0.03	15838.0	United Kingdom
160107	C550456	21108	FAIRY CAKE FLANNEL ASSORTED COLOUR	-3114	2011-04-18 13:08:00	2.10	15749.0	United Kingdom
160106	C550456	21175	GIN + TONIC DIET METAL SIGN	-2000	2011-04-18 13:08:00	1.85	15749.0	United Kingdom

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
240658	C558112	23091	ZINC HERB GARDEN CONTAINER	-1	2011-06-26 16:08:00	6.25	17114.0	United Kingdom
240656	C558112	82486	WOOD S/3 CABINET ANT WHITE FINISH	-1	2011-06-26 16:08:00	8.95	17114.0	United Kingdom
240613	C558106	21467	CHERRY CROCHET FOOD COVER	-1	2011-06-26 14:59:00	3.75	13668.0	United Kingdom
240611	C558106	21181	PLEASE ONE PERSON METAL SIGN	-1	2011-06-26 14:59:00	2.10	13668.0	United Kingdom
256253	C559474	22726	ALARM CLOCK BAKELIKE GREEN	-1	2011-07-08 13:41:00	3.75	13078.0	United Kingdom

Output 5: Data analysis - Cell #11

We found several instances of very high negative quantities. These appeared to be errors in the data. There were also numerous instances of quantities listed as -1. These are likely customer returns. Since our study focuses on items sold, these returns are not relevant and can be considered for removal.

2.2.4. Outlier Detection and Treatment

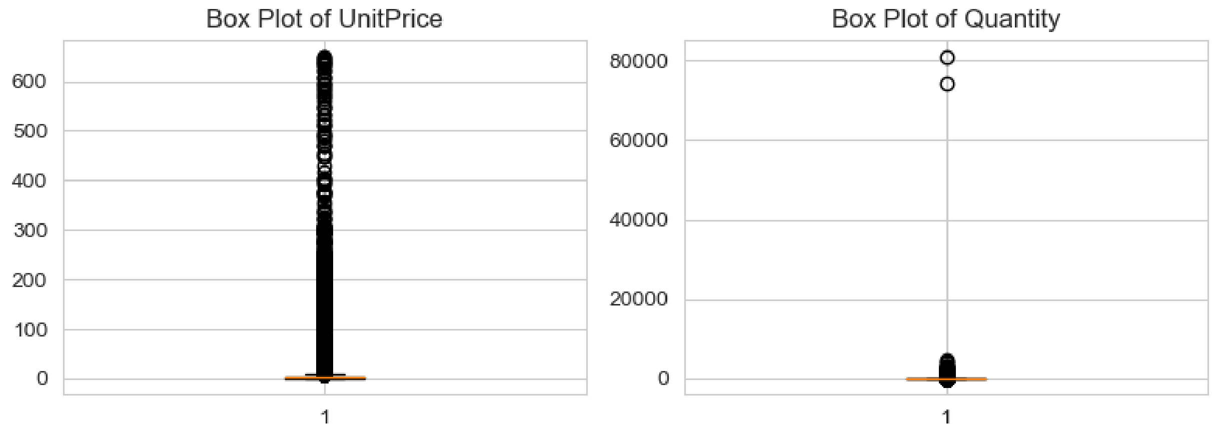


Figure 2: Outliers for numeric features after processing

```
UnitPrice Quantiles (98th to 99.9th percentile):
0.980     13.29000
0.990     16.95000
0.995     21.23000
0.999    125.94848
Name: UnitPrice, dtype: float64

Quantity Quantiles (98th to 99.9th percentile):
0.980     72.0
0.990    100.0
0.995    160.0
0.999    432.0
Name: Quantity, dtype: float64
```

Output 6: Data analysis - Cell #15

We also noticed that both the quantity and price features in our dataset had outliers, some of which were quite significant. Here's how we addressed these:

1. **Price Outliers**: There were many outliers in price, including two extremely high values. It was important to be cautious here, as not all high prices are unreasonable or errors. By examining different percentiles, specifically the 99.9th percentile, we found that most high prices were valid. However, upon closer inspection, we identified 'DOTCOM POSTAGE' and 'AMAZON FEE' as invalid products and removed these entries.
2. **Quantity Outliers**: In the case of quantity, the high numbers mostly seemed like valid transactions for a retail setting. There were, however, two extremely high outliers in quantity. After careful consideration, we decided these were not realistic and removed them.

Ultimately, we chose to retain the outliers that seemed valid and only remove those that were clearly invalid.

3. Data Analysis

3.1. Statistical Analysis

To provide clearer visualization and aid in the decision-making process, we examined the data at the 99th percentile. This analysis was solely for a better understanding of the dataset's characteristics, and no data was removed based on this percentile. The insights gained from this perspective were instrumental in identifying valid outliers and ensuring that only irrelevant or incorrect data points were excluded.

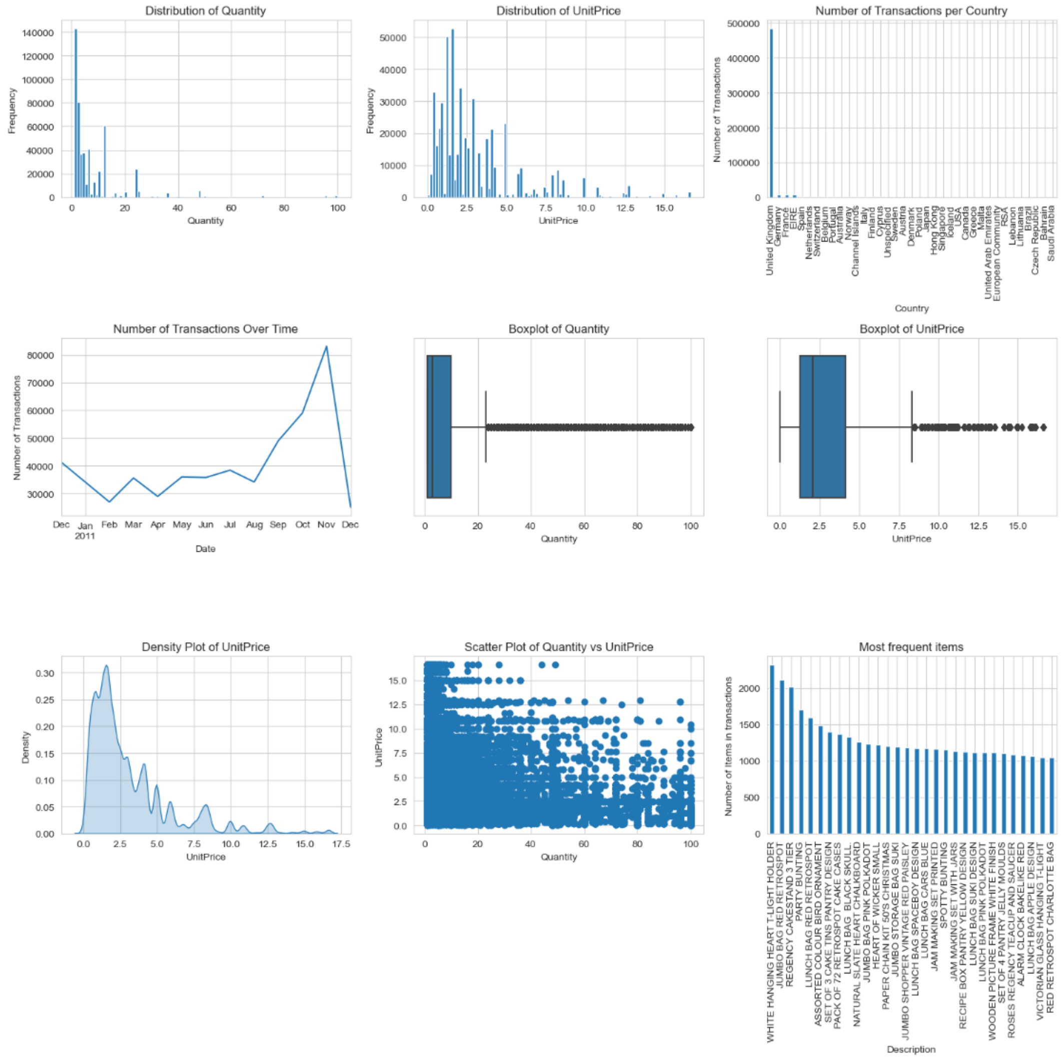


Figure 3: Data after preprocessing (99% percentile)

The basic plots of our dataset revealed some key initial insights:

1- The majority of transactions involved quantities ranging from 1 to 15 units per item.
2- Most transactions were for items priced between 0 and 5 units.
3- While there are some outliers in the data, these appear to be valid instances, representing either expensive products or bulk purchases by retailers.
4- A noticeable increase in the number of transactions was observed starting from August 2011. This suggests that the online retailer may have expanded into new markets during this period.

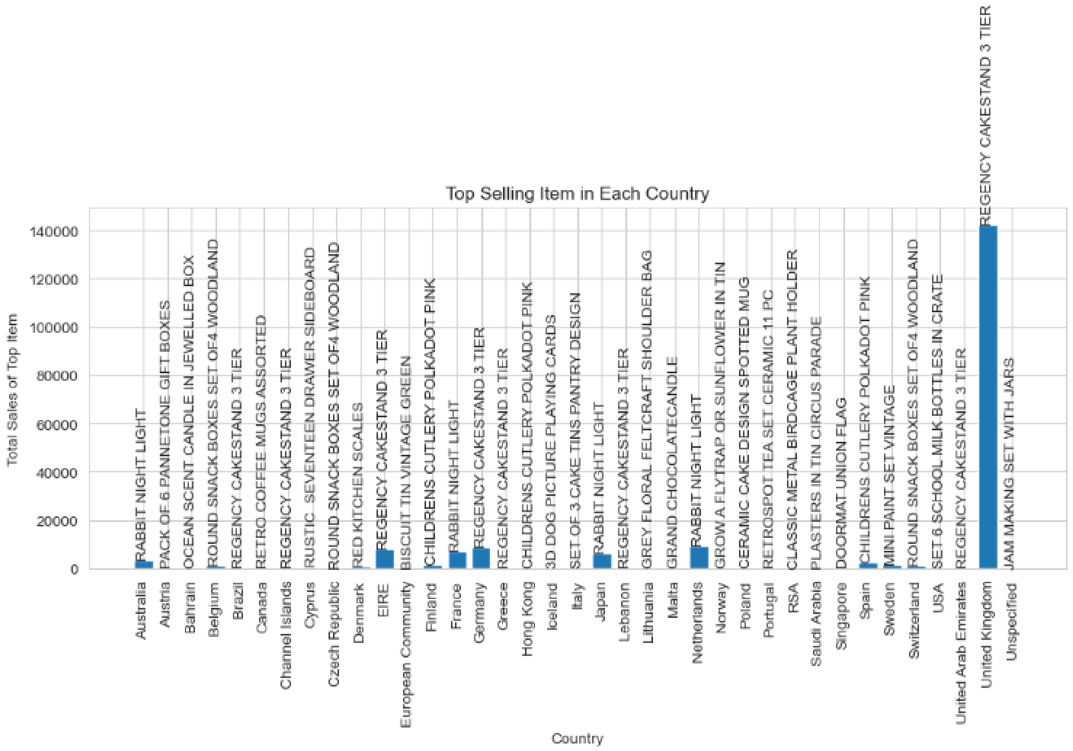


Figure 4: Top sellers per country

- **Product Popularity by Country**: The table shows a diverse range of top-selling products across different countries. For example, the 'REGENCY CAKESTAND 3 TIER' is the top seller in multiple countries like EIRE, Germany, and the United Kingdom, indicating its widespread appeal. However, each country has unique top sellers, such as the 'RABBIT NIGHT LIGHT' in Australia, Japan, France, and the Netherlands, suggesting varying consumer preferences.

- **Geographical Trends**: Products like 'REGENCY CAKESTAND 3 TIER' and 'RABBIT NIGHT LIGHT' appear multiple times across different countries, indicating their universal appeal. In contrast, some items are unique to specific regions, like 'PACK OF 6 PANNETONE GIFT BOXES' in Austria, showing localized preferences.

- **Market Opportunities**: Identifying top-selling items in each country can help the retailer tailor its inventory and marketing strategies to specific markets. For example, focusing on 'CHILDRENS CUTLERY POLKADOT PINK' in markets like Hong Kong, Finland, and Spain, where it's a top seller.

3.2. Trend Analysis

The figure presented illustrates the sales trends across the top 10 countries. A logarithmic scale is employed to enhance visualization, given the substantial differences in sales volumes between the countries. This scale choice enables a clearer focus on the trend patterns.

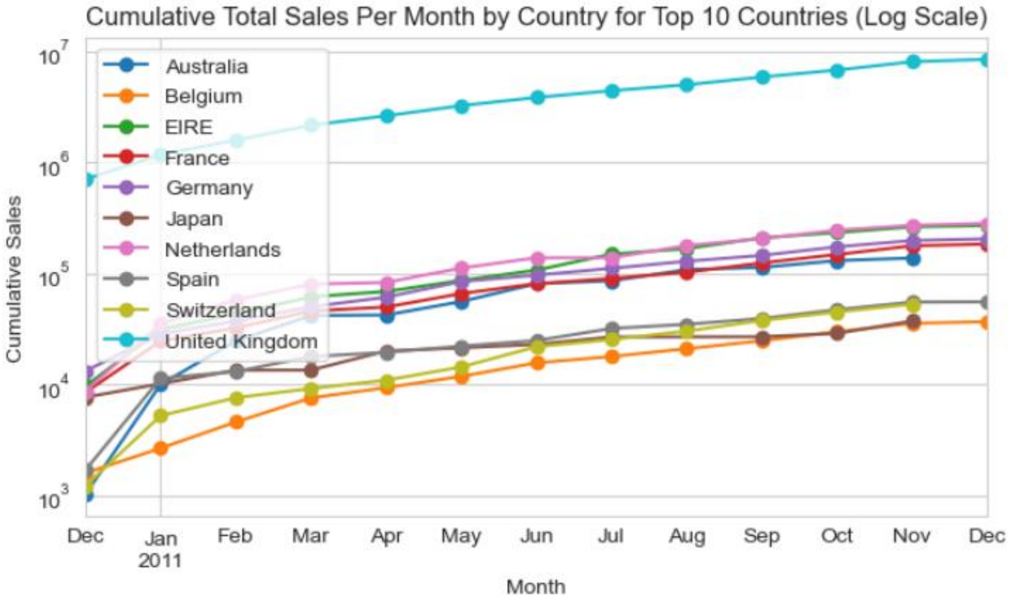


Figure 5: Sales trend per country

- **Growth Over Time**: All countries show an upward trend in sales throughout the year. This suggests that the store's business is growing over time.
- **United Kingdom Dominance**: The United Kingdom has much higher sales than any other country, showing it's the biggest market for the store.
- **Seasonal Patterns**: There seems to be a general increase in sales for all countries starting around August and continuing through the end of the year, which could be due to seasonal shopping increases, like back-to-school or holiday sales.
- **Market Differences**: While all countries show growth, the rate is different. For example, Japan and Australia have a steeper increase compared to countries like Belgium and Switzerland, indicating faster growth.

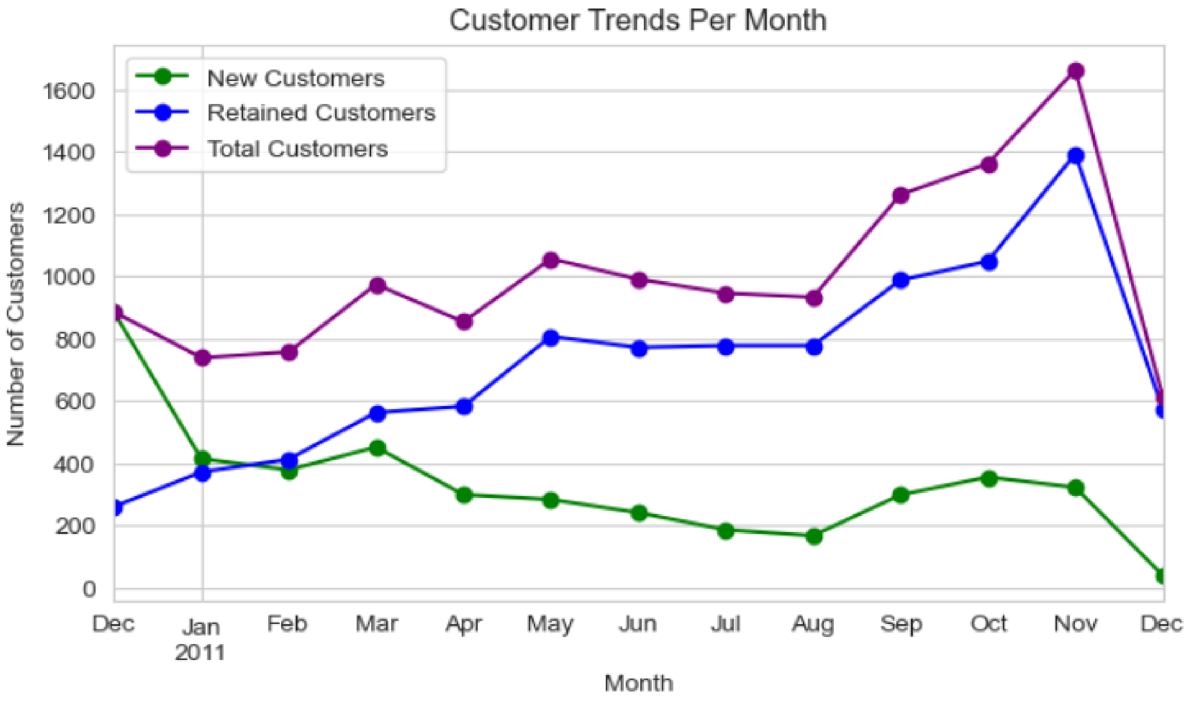


Figure 6: Customer trends per month

Taking into account that the data extends only until December 9th, the chart suggests the following trends for customer behavior:

- **Retained Customers**: There is a consistent growth in retained customers, indicating good customer loyalty and suggesting that existing customers are continuing to engage with the business throughout the year.
- **New Customers**: The number of new customers shows a significant increase in November, which might be related to holiday promotions attracting first-time buyers. However, the data does not cover the whole of December, so the expected further increase in new customers due to the holiday season is not fully captured.
- **Total Customers**: The total number of customers increases in November, likely influenced by both returning and new customers. The data suggests a peak in customer activity during this period, which may continue to rise if the data for the entire month of December were available.

Overall, the chart indicates a healthy customer base with a notable rise in customer engagement towards the end of the year, which could have potentially been even higher if the data included the entire holiday shopping season.

5. Association Rule Mining

5.1. Methodology

Association rule mining is a technique used to find interesting relationships and patterns in large datasets. These relationships are expressed as "if-then" association rules. We employed the Apriori algorithm, a well-established method in data mining, which builds on the concept that all non-empty subsets of a frequent itemset must also be frequent. This approach incrementally finds frequent items and combines them to discover the most common itemsets. [2]

Prior to running the algorithm, the dataset underwent a crucial preprocessing step to standardize colors across items. Given the dataset's characteristics, with many products sharing colors and often bought together, this was vital to avoid generating redundant rules. The assumption here was that customers choose colors at the product level, and the online store's suggestion system should reflect this behavior.

The parameters for the Apriori algorithm were carefully selected to accommodate the dataset's size. A high support value was impractical due to the item variety and transaction volume. A balanced approach was adopted to weigh confidence and lift, ensuring the algorithm captures a broad spectrum of associations without being too restrictive.

5.2. Rule generation

The process of generating rules began by identifying frequent itemsets through trial and error, starting with a high support threshold and gradually decreasing it to yield a substantial but manageable number of frequent itemsets.

Subsequently, association rules were produced with a minimum confidence threshold of 0.5. This threshold was determined after several iterations and provided a satisfactory quantity of rules, which consistently showed high lift values indicating a positive relationship between items in the rules.

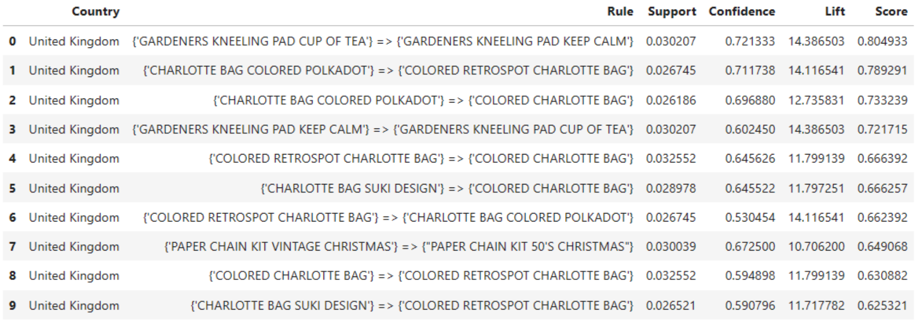
To prioritize the rules, we introduced a composite metric that merges confidence and lift, assigning a 0.7 weight to confidence and 0.3 to lift. Lift scores were normalized using the min-max approach to align with the 0-1 range of confidence scores. Since lift values were generally high across the rule set, confidence was given greater weight in the combined score to ensure the most reliable rules were considered first.

Given the dataset's size of over 530,000 records, we decided to create different rules for each market or region. This helped us manage the large data more easily and also make sure the rules were right for each market. This was especially necessary because 85% of the data is from the United Kingdom, and without splitting the data, the rules might not be accurate for users in other countries.

6. Analysis of Association Rules

Analyzing the association rules from different countries shows interesting trends in what people like to buy, with some clear differences between countries.

6.1. United Kingdom

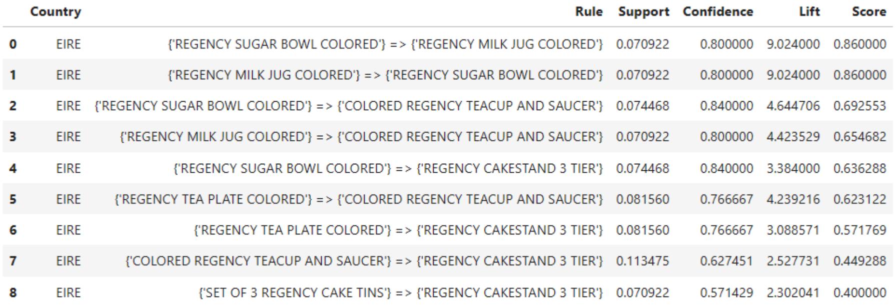


	Country	Rule	Support	Confidence	Lift	Score
0	United Kingdom	{'GARDENERS KNEELING PAD CUP OF TEA'} => {'GARDENERS KNEELING PAD KEEP CALM'}	0.030207	0.721333	14.386503	0.804933
1	United Kingdom	{'CHARLOTTE BAG COLORED POLKADOT'} => {'COLORED RETROSPOT CHARLOTTE BAG'}	0.026745	0.711738	14.116541	0.789291
2	United Kingdom	{'CHARLOTTE BAG COLORED POLKADOT'} => {'COLORED CHARLOTTE BAG'}	0.026186	0.696880	12.735831	0.733239
3	United Kingdom	{'GARDENERS KNEELING PAD KEEP CALM'} => {'GARDENERS KNEELING PAD CUP OF TEA'}	0.030207	0.602450	14.386503	0.721715
4	United Kingdom	{'COLORED RETROSPOT CHARLOTTE BAG'} => {'COLORED CHARLOTTE BAG'}	0.032552	0.645626	11.799139	0.666392
5	United Kingdom	{'CHARLOTTE BAG SUKI DESIGN'} => {'COLORED CHARLOTTE BAG'}	0.028978	0.645522	11.797251	0.666257
6	United Kingdom	{'COLORED RETROSPOT CHARLOTTE BAG'} => {'CHARLOTTE BAG COLORED POLKADOT'}	0.026745	0.530454	14.116541	0.662392
7	United Kingdom	{'PAPER CHAIN KIT VINTAGE CHRISTMAS'} => {"PAPER CHAIN KIT 50'S CHRISTMAS"}	0.030039	0.672500	10.706200	0.649068
8	United Kingdom	{'COLORED CHARLOTTE BAG'} => {'COLORED RETROSPOT CHARLOTTE BAG'}	0.032552	0.594898	11.799139	0.630882
9	United Kingdom	{'CHARLOTTE BAG SUKI DESIGN'} => {'COLORED RETROSPOT CHARLOTTE BAG'}	0.026521	0.590796	11.717782	0.625321

Figure 7: Association rules for UK

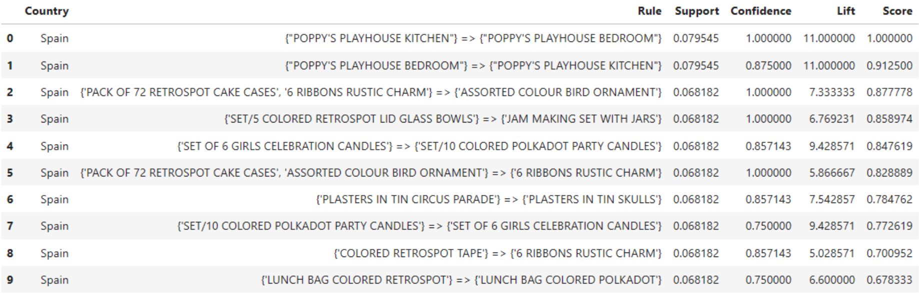
In the United Kingdom, most people buy garden items and different kinds of the Charlotte bag. Germany shows a similar trend, with people also liking the Charlotte bag. This might mean that people in these regions like things that are both useful and stylish.

6.2. Ireland and Spain



	Country	Rule	Support	Confidence	Lift	Score
0	EIRE	{'REGENCY SUGAR BOWL COLORED'} => {'REGENCY MILK JUG COLORED'}	0.070922	0.800000	9.024000	0.860000
1	EIRE	{'REGENCY MILK JUG COLORED'} => {'REGENCY SUGAR BOWL COLORED'}	0.070922	0.800000	9.024000	0.860000
2	EIRE	{'REGENCY SUGAR BOWL COLORED'} => {'COLORED REGENCY TEACUP AND SAUCER'}	0.074468	0.840000	4.644706	0.692553
3	EIRE	{'REGENCY MILK JUG COLORED'} => {'COLORED REGENCY TEACUP AND SAUCER'}	0.070922	0.800000	4.423529	0.654682
4	EIRE	{'REGENCY SUGAR BOWL COLORED'} => {'REGENCY CAKESTAND 3 TIER'}	0.074468	0.840000	3.384000	0.636288
5	EIRE	{'REGENCY TEA PLATE COLORED'} => {'COLORED REGENCY TEACUP AND SAUCER'}	0.081560	0.766667	4.239216	0.623122
6	EIRE	{'REGENCY TEA PLATE COLORED'} => {'REGENCY CAKESTAND 3 TIER'}	0.081560	0.766667	3.088571	0.571769
7	EIRE	{'COLORED REGENCY TEACUP AND SAUCER'} => {'REGENCY CAKESTAND 3 TIER'}	0.113475	0.627451	2.527731	0.449288
8	EIRE	{'SET OF 3 REGENCY CAKE TINS'} => {'REGENCY CAKESTAND 3 TIER'}	0.070922	0.571429	2.302041	0.400000

Figure 8: Association rules for Ireland

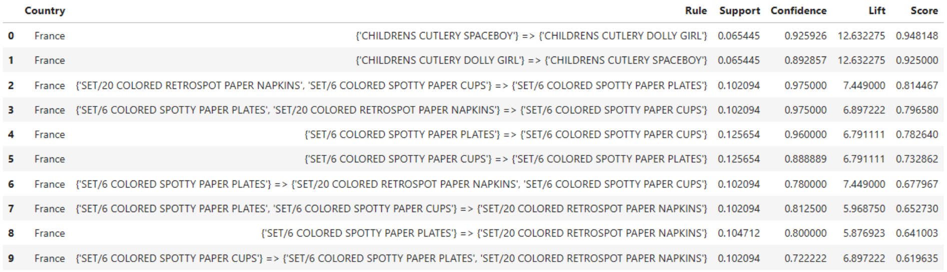


	Country	Rule	Support	Confidence	Lift	Score
0	Spain	{"POPPY'S PLAYHOUSE KITCHEN"} => {"POPPY'S PLAYHOUSE BEDROOM"}	0.079545	1.000000	11.000000	1.000000
1	Spain	{"POPPY'S PLAYHOUSE BEDROOM"} => {"POPPY'S PLAYHOUSE KITCHEN"}	0.079545	0.875000	11.000000	0.912500
2	Spain	{'PACK OF 72 RETROSPOT CAKE CASES', '6 RIBBONS RUSTIC CHARM'} => {'ASSORTED COLOUR BIRD ORNAMENT'}	0.068182	1.000000	7.333333	0.877778
3	Spain	{'SET/5 COLORED RETROSPOT LID GLASS BOWLS'} => {'JAM MAKING SET WITH JARS'}	0.068182	1.000000	6.769231	0.858974
4	Spain	{'SET OF 6 GIRLS CELEBRATION CANDLES'} => {'SET/10 COLORED POLKADOT PARTY CANDLES'}	0.068182	0.857143	9.428571	0.847619
5	Spain	{'PACK OF 72 RETROSPOT CAKE CASES', 'ASSORTED COLOUR BIRD ORNAMENT'} => {'6 RIBBONS RUSTIC CHARM'}	0.068182	1.000000	5.866667	0.828889
6	Spain	{'PLASTERS IN TIN CIRCUS PARADE'} => {'PLASTERS IN TIN SKULLS'}	0.068182	0.857143	7.542857	0.784762
7	Spain	{'SET/10 COLORED POLKADOT PARTY CANDLES'} => {'SET OF 6 GIRLS CELEBRATION CANDLES'}	0.068182	0.750000	9.428571	0.772619
8	Spain	{'COLORED RETROSPOT TAPE'} => {'6 RIBBONS RUSTIC CHARM'}	0.068182	0.857143	5.028571	0.700952
9	Spain	{'LUNCH BAG COLORED RETROSPOT'} => {'LUNCH BAG COLORED POLKADOT'}	0.068182	0.750000	6.600000	0.678333

Figure 9: Association rules for Spain

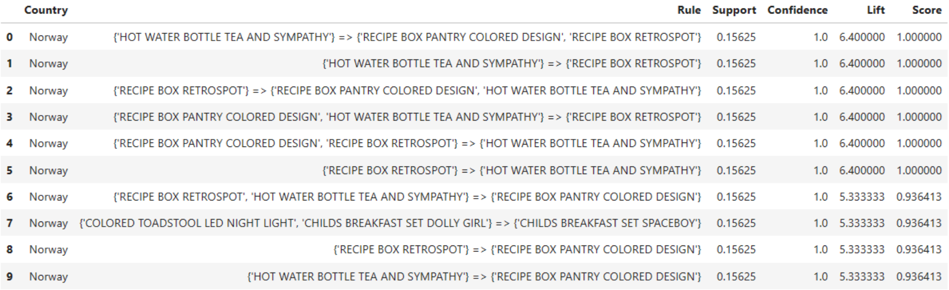
Ireland and Spain both seem to like things for the home, but in different ways. In Ireland, people often buy Regency-style kitchen items, while in Spain, they buy decorative items like Poppy's playhouse sets. This might mean that people in these places enjoy decorating and using items in their homes.

6.3. France and Norway



	Country	Rule	Support	Confidence	Lift	Score
0	France	{'CHILDRENS CUTLERY SPACEBOY'} => {'CHILDRENS CUTLERY DOLLY GIRL'}	0.065445	0.925926	12.632275	0.948148
1	France	{'CHILDRENS CUTLERY DOLLY GIRL'} => {'CHILDRENS CUTLERY SPACEBOY'}	0.065445	0.892857	12.632275	0.925000
2	France	{'SET/20 COLORED RETROSPOT PAPER NAPKINS', 'SET/6 COLORED SPOTTY PAPER CUPS'} => {'SET/6 COLORED SPOTTY PAPER PLATES'}	0.102094	0.975000	7.449000	0.814467
3	France	{'SET/6 COLORED SPOTTY PAPER PLATES', 'SET/20 COLORED RETROSPOT PAPER NAPKINS'} => {'SET/6 COLORED SPOTTY PAPER CUPS'}	0.102094	0.975000	6.897222	0.796580
4	France	{'SET/6 COLORED SPOTTY PAPER PLATES'} => {'SET/6 COLORED SPOTTY PAPER CUPS'}	0.125654	0.960000	6.791111	0.782640
5	France	{'SET/6 COLORED SPOTTY PAPER CUPS'} => {'SET/6 COLORED SPOTTY PAPER PLATES'}	0.125654	0.888889	6.791111	0.732862
6	France	{'SET/6 COLORED SPOTTY PAPER PLATES'} => {'SET/20 COLORED RETROSPOT PAPER NAPKINS', 'SET/6 COLORED SPOTTY PAPER CUPS'}	0.102094	0.780000	7.449000	0.677967
7	France	{'SET/6 COLORED SPOTTY PAPER PLATES', 'SET/6 COLORED SPOTTY PAPER CUPS'} => {'SET/20 COLORED RETROSPOT PAPER NAPKINS'}	0.102094	0.812500	5.968750	0.652730
8	France	{'SET/6 COLORED SPOTTY PAPER PLATES'} => {'SET/20 COLORED RETROSPOT PAPER NAPKINS'}	0.104712	0.800000	5.876923	0.641003
9	France	{'SET/6 COLORED SPOTTY PAPER CUPS'} => {'SET/6 COLORED SPOTTY PAPER PLATES', 'SET/20 COLORED RETROSPOT PAPER NAPKINS'}	0.102094	0.722222	6.897222	0.619635

Figure 10: Association rules for France

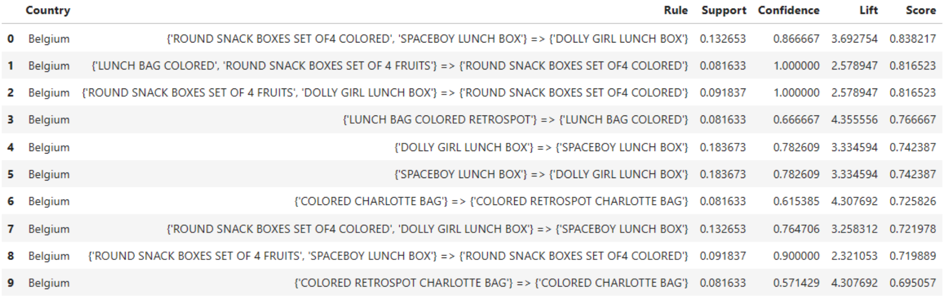


	Country	Rule	Support	Confidence	Lift	Score
0	Norway	{'HOT WATER BOTTLE TEA AND SYMPATHY'} => {'RECIPE BOX PANTRY COLORED DESIGN', 'RECIPE BOX RETROSPOT'}	0.15625	1.0	6.400000	1.000000
1	Norway	{'HOT WATER BOTTLE TEA AND SYMPATHY'} => {'RECIPE BOX RETROSPOT'}	0.15625	1.0	6.400000	1.000000
2	Norway	{'RECIPE BOX RETROSPOT'} => {'RECIPE BOX PANTRY COLORED DESIGN', 'HOT WATER BOTTLE TEA AND SYMPATHY'}	0.15625	1.0	6.400000	1.000000
3	Norway	{'RECIPE BOX PANTRY COLORED DESIGN', 'HOT WATER BOTTLE TEA AND SYMPATHY'} => {'RECIPE BOX RETROSPOT'}	0.15625	1.0	6.400000	1.000000
4	Norway	{'RECIPE BOX PANTRY COLORED DESIGN', 'RECIPE BOX RETROSPOT'} => {'HOT WATER BOTTLE TEA AND SYMPATHY'}	0.15625	1.0	6.400000	1.000000
5	Norway	{'RECIPE BOX RETROSPOT'} => {'HOT WATER BOTTLE TEA AND SYMPATHY'}	0.15625	1.0	6.400000	1.000000
6	Norway	{'RECIPE BOX RETROSPOT', 'HOT WATER BOTTLE TEA AND SYMPATHY'} => {'RECIPE BOX PANTRY COLORED DESIGN'}	0.15625	1.0	5.333333	0.936413
7	Norway	{'COLORED TOADSTOOL LED NIGHT LIGHT', 'CHILDS BREAKFAST SET DOLLY GIRL'} => {'CHILDS BREAKFAST SET SPACEBOY'}	0.15625	1.0	5.333333	0.936413
8	Norway	{'RECIPE BOX RETROSPOT'} => {'RECIPE BOX PANTRY COLORED DESIGN'}	0.15625	1.0	5.333333	0.936413
9	Norway	{'HOT WATER BOTTLE TEA AND SYMPATHY'} => {'RECIPE BOX PANTRY COLORED DESIGN'}	0.15625	1.0	5.333333	0.936413

Figure 11: Association rules for Norway

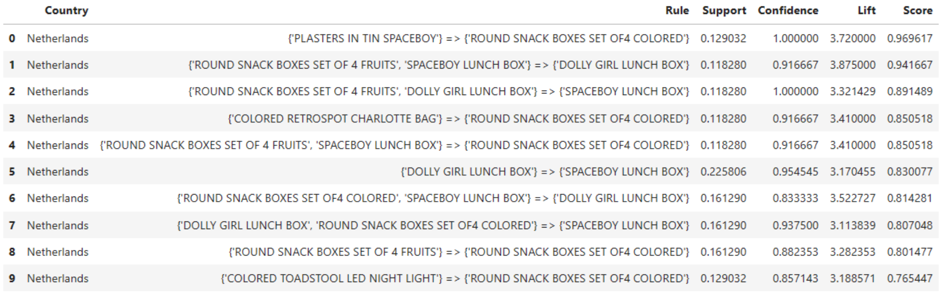
France and Norway have a different focus, with a lot of children's items like cutlery sets and recipe boxes being popular. This could mean that families or children's needs are a priority for shoppers in these countries.

6.4. Belgium and Netherlands



	Country	Rule	Support	Confidence	Lift	Score
0	Belgium	{'ROUND SNACK BOXES SET OF4 COLORED', 'SPACEBOY LUNCH BOX'} => {'DOLLY GIRL LUNCH BOX'}	0.132653	0.866667	3.692754	0.838217
1	Belgium	{'LUNCH BAG COLORED', 'ROUND SNACK BOXES SET OF 4 FRUITS'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.081633	1.000000	2.578947	0.816523
2	Belgium	{'ROUND SNACK BOXES SET OF 4 FRUITS', 'DOLLY GIRL LUNCH BOX'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.091837	1.000000	2.578947	0.816523
3	Belgium	{'LUNCH BAG COLORED RETROSPOT'} => {'LUNCH BAG COLORED'}	0.081633	0.666667	4.355556	0.766667
4	Belgium	{'DOLLY GIRL LUNCH BOX'} => {'SPACEBOY LUNCH BOX'}	0.183673	0.782609	3.334594	0.742387
5	Belgium	{'SPACEBOY LUNCH BOX'} => {'DOLLY GIRL LUNCH BOX'}	0.183673	0.782609	3.334594	0.742387
6	Belgium	{'COLORED CHARLOTTE BAG'} => {'COLORED RETROSPOT CHARLOTTE BAG'}	0.081633	0.615385	4.307692	0.725826
7	Belgium	{'ROUND SNACK BOXES SET OF4 COLORED', 'DOLLY GIRL LUNCH BOX'} => {'SPACEBOY LUNCH BOX'}	0.132653	0.764706	3.258312	0.721978
8	Belgium	{'ROUND SNACK BOXES SET OF 4 FRUITS', 'SPACEBOY LUNCH BOX'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.091837	0.900000	2.321053	0.719889
9	Belgium	{'COLORED RETROSPOT CHARLOTTE BAG'} => {'COLORED CHARLOTTE BAG'}	0.081633	0.571429	4.307692	0.695057

Figure 12: Association rules for Belgium

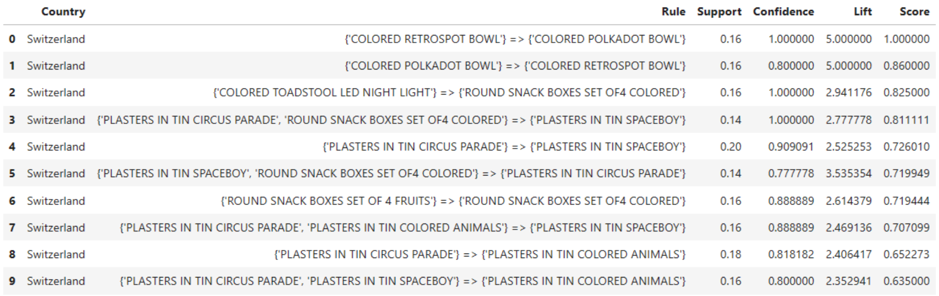


	Country	Rule	Support	Confidence	Lift	Score
0	Netherlands	{'PLASTERS IN TIN SPACEBOY'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.129032	1.000000	3.720000	0.969617
1	Netherlands	{'ROUND SNACK BOXES SET OF 4 FRUITS', 'SPACEBOY LUNCH BOX'} => {'DOLLY GIRL LUNCH BOX'}	0.118280	0.916667	3.875000	0.941667
2	Netherlands	{'ROUND SNACK BOXES SET OF 4 FRUITS', 'DOLLY GIRL LUNCH BOX'} => {'SPACEBOY LUNCH BOX'}	0.118280	1.000000	3.321429	0.891489
3	Netherlands	{'COLORED RETROSPOT CHARLOTTE BAG'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.118280	0.916667	3.410000	0.850518
4	Netherlands	{'ROUND SNACK BOXES SET OF 4 FRUITS', 'SPACEBOY LUNCH BOX'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.118280	0.916667	3.410000	0.850518
5	Netherlands	{'DOLLY GIRL LUNCH BOX'} => {'SPACEBOY LUNCH BOX'}	0.225806	0.954545	3.170455	0.830077
6	Netherlands	{'ROUND SNACK BOXES SET OF4 COLORED', 'SPACEBOY LUNCH BOX'} => {'DOLLY GIRL LUNCH BOX'}	0.161290	0.833333	3.522727	0.814281
7	Netherlands	{'DOLLY GIRL LUNCH BOX', 'ROUND SNACK BOXES SET OF4 COLORED'} => {'SPACEBOY LUNCH BOX'}	0.161290	0.937500	3.113839	0.807048
8	Netherlands	{'ROUND SNACK BOXES SET OF 4 FRUITS'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.161290	0.882353	3.282353	0.801477
9	Netherlands	{'COLORED TOADSTOOL LED NIGHT LIGHT'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.129032	0.857143	3.188571	0.765447

Figure 13: Association rules for Netherlands

Belgium and the Netherlands both have a lot of sales for lunch boxes and snack boxes. This could show that people in these countries like things that make eating on the go easy and convenient.

6.5. Switzerland

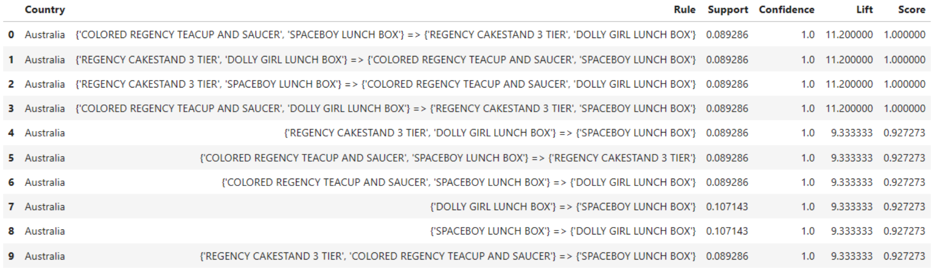


	Country	Rule	Support	Confidence	Lift	Score
0	Switzerland	{'COLORED RETROSPOT BOWL'} => {'COLORED POLKADOT BOWL'}	0.16	1.000000	5.000000	1.000000
1	Switzerland	{'COLORED POLKADOT BOWL'} => {'COLORED RETROSPOT BOWL'}	0.16	0.800000	5.000000	0.860000
2	Switzerland	{'COLORED TOADSTOOL LED NIGHT LIGHT'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.16	1.000000	2.941176	0.825000
3	Switzerland	{'PLASTERS IN TIN CIRCUS PARADE', 'ROUND SNACK BOXES SET OF4 COLORED'} => {'PLASTERS IN TIN SPACEBOY'}	0.14	1.000000	2.777778	0.811111
4	Switzerland	{'PLASTERS IN TIN CIRCUS PARADE'} => {'PLASTERS IN TIN SPACEBOY'}	0.20	0.909091	2.525253	0.726010
5	Switzerland	{'PLASTERS IN TIN SPACEBOY', 'ROUND SNACK BOXES SET OF4 COLORED'} => {'PLASTERS IN TIN CIRCUS PARADE'}	0.14	0.777778	3.535354	0.719949
6	Switzerland	{'ROUND SNACK BOXES SET OF 4 FRUITS'} => {'ROUND SNACK BOXES SET OF4 COLORED'}	0.16	0.888889	2.614379	0.719444
7	Switzerland	{'PLASTERS IN TIN CIRCUS PARADE', 'PLASTERS IN TIN COLORED ANIMALS'} => {'PLASTERS IN TIN SPACEBOY'}	0.16	0.888889	2.469136	0.707099
8	Switzerland	{'PLASTERS IN TIN CIRCUS PARADE'} => {'PLASTERS IN TIN COLORED ANIMALS'}	0.18	0.818182	2.406417	0.652273
9	Switzerland	{'PLASTERS IN TIN CIRCUS PARADE', 'PLASTERS IN TIN SPACEBOY'} => {'PLASTERS IN TIN COLORED ANIMALS'}	0.16	0.800000	2.352941	0.635000

Figure 14: Association rules for Switzerland

Switzerland stands out with its preference for colored bowls and plaster tins, showing specific local tastes that might not be as common in other countries.

6.6. Australia



	Country	Rule	Support	Confidence	Lift	Score
0	Australia	{'COLORED REGENCY TEACUP AND SAUCER', 'SPACEBOY LUNCH BOX'} => {'REGENCY CAKESTAND 3 TIER', 'DOLLY GIRL LUNCH BOX'}	0.089286	1.0	11.200000	1.000000
1	Australia	{'REGENCY CAKESTAND 3 TIER', 'DOLLY GIRL LUNCH BOX'} => {'COLORED REGENCY TEACUP AND SAUCER', 'SPACEBOY LUNCH BOX'}	0.089286	1.0	11.200000	1.000000
2	Australia	{'REGENCY CAKESTAND 3 TIER', 'SPACEBOY LUNCH BOX'} => {'COLORED REGENCY TEACUP AND SAUCER', 'DOLLY GIRL LUNCH BOX'}	0.089286	1.0	11.200000	1.000000
3	Australia	{'COLORED REGENCY TEACUP AND SAUCER', 'DOLLY GIRL LUNCH BOX'} => {'REGENCY CAKESTAND 3 TIER', 'SPACEBOY LUNCH BOX'}	0.089286	1.0	11.200000	1.000000
4	Australia	{'REGENCY CAKESTAND 3 TIER', 'DOLLY GIRL LUNCH BOX'} => {'SPACEBOY LUNCH BOX'}	0.089286	1.0	9.333333	0.927273
5	Australia	{'COLORED REGENCY TEACUP AND SAUCER', 'SPACEBOY LUNCH BOX'} => {'REGENCY CAKESTAND 3 TIER'}	0.089286	1.0	9.333333	0.927273
6	Australia	{'COLORED REGENCY TEACUP AND SAUCER', 'SPACEBOY LUNCH BOX'} => {'DOLLY GIRL LUNCH BOX'}	0.089286	1.0	9.333333	0.927273
7	Australia	{'DOLLY GIRL LUNCH BOX'} => {'SPACEBOY LUNCH BOX'}	0.107143	1.0	9.333333	0.927273
8	Australia	{'SPACEBOY LUNCH BOX'} => {'DOLLY GIRL LUNCH BOX'}	0.107143	1.0	9.333333	0.927273
9	Australia	{'REGENCY CAKESTAND 3 TIER', 'COLORED REGENCY TEACUP AND SAUCER'} => {'SPACEBOY LUNCH BOX'}	0.089286	1.0	9.333333	0.927273

Figure 15: Association rules for Australia

Australia shows a wide range of interests, mixing up lunch boxes, teacups, and cake stands in their top items. This variety could point to a market with many different interests or a good chance to mix different kinds of products in sales and promotions.

Overall, while each country has its unique favorites, some trends, like the love for useful items like bags and things for the home, are common in many places. These insights are important for making marketing strategies and deciding what to stock in each market, keeping in mind what people in each country like to buy.

7. Conclusion

In this report, we looked closely at an online store's sales data. Our first step was cleaning the data, making sure everything was correct and ready for analysis. Then, we used a special method to find out which products are often bought together.

From our study, we learned a lot about what people in different countries like to buy. For example, in the UK and Germany, people like to buy garden items and different types of bags. In Ireland and Spain, people are more interested in things for their homes, like kitchen items in Ireland and decorative items in Spain.

The analysis also showed that in France and Norway, children's items like cutlery sets are popular. In Belgium and the Netherlands, people buy a lot of lunch boxes and snack boxes, and in Switzerland, there's a big interest in colored bowls and plaster tins. Australia, on the other hand, showed a mix of different interests.

By understanding what people in each country like to buy, we can give better advice to the store on what to sell and how to market their products. This report gives us a good look at customer behavior in different countries and helps us think about what to do next, like more studies or new ways to look at the data.

8. References

[1] "Online Retail," UCI Machine Learning Repository, 2015.

[2] A. Noor, "What is association rule mining?," [Online]. Available: https://www.educative.io/answers/what-is-association-rule-mining.