Task

T5.1: Download the 'Portugal_online_retail', 'Sweden_online_retail, and 'UK_online_retail' datasets from Canvas. Apply the apriori algorithm to all datasets using three different confidence levels. Select one confidence level for each dataset that you think works better. Determine the first three most important rules for each dataset using the selected confidence level and report them in the report cell. Explain what each rule means (Completing the report cell is required) (15%).

NOTE: You should comment on your code

In [86]:
```
#Installation of mxltend package for Apriori
!pip install mlxtend
```

In [85]:
```
################ Write your code in this cell (If applicable) ###################
import numpy as np
import pandas as pd
from mlxtend.frequent_patterns import apriori, association_rules

# importing our datasets
PORTUGAL = pd.read_csv('Portugal_online_retail.csv')

#I want to preview the size of my data
print("\n PORTUGAL.shape")
print(PORTUGAL.shape)


#Invoice column is out of category 0 & 1, so i will drop that column
PORTUGAL_NEW=PORTUGAL.drop('InvoiceNo',axis=1)

#TO VIEW THE DATA
PORTUGAL_NEW.head

# The Frequent Items Analysis
frequent_items = apriori(PORTUGAL_NEW, min_support = 0.14, use_colnames = True)
print(frequent_items.shape)

#To view my column support in descending order
frequent_items.sort_values(by=['support'],ascending=False)
```

```
 PORTUGAL.shape
(58, 714)
(13, 2)
```

```
C:\Users\dessy\anaconda3\Lib\site-packages\mlxtend\frequent_patterns\fpcommon.py:109: DeprecationWarning: DataFrames with non-bool types result in worse computationalperformance and their support might be discontinued in the future.Please use a DataFrame with bool type
  warnings.warn(
```

Out[85]:

	support	itemsets
8	0.517241	(POSTAGE)
6	0.241379	(LUNCH BAG RED RETROSPOT)
10	0.241379	(RETROSPOT TEA SET CERAMIC 11 PC)
0	0.206897	(BAKING SET 9 PIECE RETROSPOT)
5	0.206897	(LUNCH BAG CARS BLUE)
4	0.189655	(JUMBO SHOPPER VINTAGE RED PAISLEY)
1	0.172414	(CHARLOTTE BAG SUKI DESIGN)
7	0.172414	(PLASTERS IN TIN VINTAGE PAISLEY)
9	0.172414	(RED RETROSPOT CHARLOTTE BAG)
2	0.155172	(JUMBO BAG PINK VINTAGE PAISLEY)
3	0.155172	(JUMBO BAG SCANDINAVIAN BLUE PAISLEY)
11	0.155172	(BAKING SET 9 PIECE RETROSPOT, RETROSPOT TEA S...
12	0.155172	(JUMBO BAG PINK VINTAGE PAISLEY, JUMBO SHOPPER...

In [80]:

```
# Collecting the inferred rules in the dataset
rules = association_rules(frequent_items, metric ="confidence", min_threshold = 0.6

#viewing the rules in descending order
rules.sort_values(by=['confidence'],ascending=False)
```

Out[80]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	c
2	(JUMBO BAG PINK VINTAGE PAISLEY)	(JUMBO SHOPPER VINTAGE RED PAISLEY)	0.155172	0.189655	0.155172	1.000000	5.272727	0.125743	
3	(JUMBO SHOPPER VINTAGE RED PAISLEY)	(JUMBO BAG PINK VINTAGE PAISLEY)	0.189655	0.155172	0.155172	0.818182	5.272727	0.125743	
0	(BAKING SET 9 PIECE RETROSPOT)	(RETROSPOT TEA SET CERAMIC 11 PC)	0.206897	0.241379	0.155172	0.750000	3.107143	0.105232	
1	(RETROSPOT TEA SET CERAMIC 11 PC)	(BAKING SET 9 PIECE RETROSPOT)	0.241379	0.206897	0.155172	0.642857	3.107143	0.105232	

################## Write your report in this cell (if applicable)

Association Rule Mining

Association rule mining is an important technique in retail analytics, allowing businesses to discover interesting patterns and relationships in transaction data. The purpose of this report

is to demonstrate the utility of association rule mining using the Apriori algorithm on a retail dataset in terms of understanding customer behaviour and improving business strategies.

Association rule mining is the process of discovering relationships between items in a dataset. The Apriori algorithm is a popular method for association rule mining. It uses measures like support, confidence, and lift to identify meaningful relationships between items.

The dataset for this analysis consists of transaction records from a retail store. Each record contains information about the items purchased, the quantity, and the InvoiceID,Which was dropped due to redundancy. Before running the Apriori algorithm, data was cleaned by dropping the InvoiceID.

PORTUGAL RETAIL DATASET REPORT

In the process of analysing the most important rules for this dataset, we start by dropping the invoiceno column as the analysis requires the dataset to range between 0 & 1, we then proceeded to processing the frequent itemset,we set the min_support to 15 so the association group can be minimal (13 groups),we proceeded to generating the association rule with 60% as the minimum threshold,indicating the relationship between items in the retail store

ASSOCIATION

RULE 1 : Purchasing the "JUMBO SHOPPER VINTAGE RED PAISLEY" and the "JUMBO BAG PINK VINTAGE PAISLEY" are perfectly correlated, according to this rule. The data indicates a strong positive correlation (lift of 5.272727) between the purchase of the two items" with a 100% confidence level..

RULE 2 : The first rule is reinforced by this one, which shows a strong correlation between buying the "JUMBO SHOPPER VINTAGE RED PAISLEY" and the "JUMBO BAG PINK VINTAGE PAISLEY." The 81.82% confidence level indicates that there's a good chance that customers who buy the former will also buy the latter item..

RULE 3 : Buying the "RETROSPOT TEA SET CERAMIC 11 PC" and the "BAKING SET 9 PIECE RETROSPOT" are strongly correlated, according to this rule. When customers buy the former item, there is a 3.107143 times greater chance that they will buy the latter, according to data with a 75% confidence level.

SWEDEN DATASET

In [71]:
```
#                                                SWEDEN DATASET

# importing SWEDEN datasets
SWEDEN = pd.read_csv('Sweden_online_retail.csv')

#Removing the redundant column (InvoiceNo)
SWEDEN_NEW=SWEDEN.drop('InvoiceNo',axis=1)


frq_items = apriori(SWEDEN_NEW, min_support = 0.11, use_colnames = True)
```

```
frq_items.shape

#Generating the support values and itemsets
frq_items.sort_values(by=['support'],ascending=False)

# Association rules in the dataset
sweden_rules = association_rules(frq_items, metric ="confidence", min_threshold = 0
print(sweden_rules.head(3))
```

```
                        antecedents                        consequents  \
0                (GUMBALL COAT RACK)                          (POSTAGE)
1  (PACK OF 72 RETROSPOT CAKE CASES)   (PACK OF 60 SPACEBOY CAKE CASES)
2   (PACK OF 60 SPACEBOY CAKE CASES)  (PACK OF 72 RETROSPOT CAKE CASES)

   antecedent support  consequent support   support  confidence      lift  \
0            0.138889            0.611111  0.138889         1.0  1.636364
1            0.111111            0.111111  0.111111         1.0  9.000000
2            0.111111            0.111111  0.111111         1.0  9.000000

   leverage  conviction  zhangs_metric
0  0.054012         inf       0.451613
1  0.098765         inf       1.000000
2  0.098765         inf       1.000000
```

```
C:\Users\dessy\anaconda3\Lib\site-packages\mlxtend\frequent_patterns\fpcommon.py:1
09: DeprecationWarning: DataFrames with non-bool types result in worse computation
alperformance and their support might be discontinued in the future.Please use a D
ataFrame with bool type
  warnings.warn(
```

REPORT FOR SWEDEN (Confidence Level 0f 0.11 with confidence > 70%)

Having fulfilled the condition of calculating Apriori algorithm, all subsets of itemset are frequent ,we mine some likely association rules using the generated frequent items and analysed the confidence level at 70%, our first three rules are greater than the minimum confidence level and thus can be categorised as strong association rule. having considered various confidence level [11,12,10].

In other word:

GUMBALL COAT RACK = POSTAGE = 1.00 > 70%

PACK OF 72 SPACEBOY CAKE CASES = PACK OF 60 SPACEBOY CAKE CASES = 1.00 > 70%

PACK OF 60 SPACEBOY CAKE CASES = PACK OF 72 RETROSPOT CAKE CASES = 1.00 > 70%, With a confidence of 100% and a lift of 9.0, this association is statistically significant

If a customer picks item A they are likely to purchase item B, and with all confidence level greater than 70% this rule is significant.

uk_online _retail

In [49]:
```
#                                          UK ONLINE RETAIL DATASET

# importing UK datasets
UK = pd.read_csv('UK_online_retail.csv')

UK.head()
```

```
C:\Users\dessy\AppData\Local\Temp\ipykernel_9824\407019094.py:4: DtypeWarning: Col
umns (0) have mixed types. Specify dtype option on import or set low_memory=False.
  UK = pd.read_csv('UK_online_retail.csv')
```

Out[49]:

	InvoiceNo	*Boombox Ipod Classic	*USB Office Mirror Ball	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 DAISY PEGS IN WOOD BOX	12 EGG HOUSE PAINTED WOOD	12 HANGING EGGS HAND PAINTED	12 IVORY ROSE PEG PLACE SETTINGS
0	536365	0	0	0	0	0	0	0	0
1	536366	0	0	0	0	0	0	0	0
2	536367	0	0	0	0	0	0	0	0
3	536368	0	0	0	0	0	0	0	0
4	536369	0	0	0	0	0	0	0	0

5 rows × 4176 columns

In [72]:

```
#Removing the redundant column (InvoiceNo)
UK_NEW=UK.drop('InvoiceNo',axis=1)

# Generating The Frequent Itemsets
freq_items = apriori(UK_NEW, min_support = 0.03, use_colnames = True)
print(freq_items.shape)


#Generating the support values and itemsets
freq_items.sort_values(by=['support'],ascending=False)
freq_items['length'] = freq_items['itemsets'].apply(lambda x: len(x))
print(freq_items)


# Association rules in the dataset
UK_rules = association_rules(freq_items, metric ="confidence", min_threshold = 0.6)
UK_rules.sort_values(by=['confidence'],ascending=False)
```

```
C:\Users\dessy\anaconda3\Lib\site-packages\mlxtend\frequent_patterns\fpcommon.py:1
09: DeprecationWarning: DataFrames with non-bool types result in worse computation
alperformance and their support might be discontinued in the future.Please use a D
ataFrame with bool type
  warnings.warn(
(131, 2)
      support                                           itemsets  length
0    0.045803                        (6 RIBBONS RUSTIC CHARM)       1
1    0.031124                (60 CAKE CASES VINTAGE CHRISTMAS)       1
2    0.040339                    (60 TEATIME FAIRY CAKE CASES)       1
3    0.046928                     (ALARM CLOCK BAKELIKE GREEN)       1
4    0.035142                      (ALARM CLOCK BAKELIKE PINK)       1
..        ...                                               ...     ...
126  0.030535  (JUMBO BAG RED RETROSPOT, JUMBO  BAG BAROQUE B...       2
127  0.042053  (JUMBO BAG RED RETROSPOT, JUMBO BAG PINK POLKA...       2
128  0.035196  (JUMBO SHOPPER VINTAGE RED PAISLEY, JUMBO BAG ...       2
129  0.037392  (JUMBO STORAGE BAG SUKI, JUMBO BAG RED RETROSPOT)       2
130  0.032517  (LUNCH BAG  BLACK SKULL., LUNCH BAG RED RETROS...       2

[131 rows x 3 columns]
```

Out[72]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage
3	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.037660	0.050035	0.030910	0.820768	16.403939	0.029026
4	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.050035	0.051267	0.037553	0.750535	14.639752	0.034988
5	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.051267	0.050035	0.037553	0.732497	14.639752	0.034988
7	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.062088	0.103820	0.042053	0.677308	6.523895	0.035607
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.046928	0.049821	0.030160	0.642694	12.900183	0.027822
6	(JUMBO BAG BAROQUE BLACK WHITE)	(JUMBO BAG RED RETROSPOT)	0.048749	0.103820	0.030535	0.626374	6.033290	0.025474
2	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.050035	0.037660	0.030910	0.617773	16.403939	0.029026
8	(JUMBO STORAGE BAG SUKI)	(JUMBO BAG RED RETROSPOT)	0.060535	0.103820	0.037392	0.617699	5.949737	0.031108
0	(ALARM CLOCK BAKELIKE	(ALARM CLOCK BAKELIKE	0.049821	0.046928	0.030160	0.605376	12.900183	0.027822

REPORT

The confidence level for this dataset is best at 0.03 as most items are best frequest at that level, our rules output is generated at decending order for best decision making to discuss highest 3 confidence

ASSOCIATION RULE 1

Analysis: This rule suggests a significant correlation between consumers of PINK REGENCY TEACUP AND SAUCER and those of GREEN REGENCY TEACUP AND SAUCER. According to the high confidence value of 82.08%, GREEN REGENCY TEACUP AND SAUCER is also involved in 82.08% of transactions involving PINK REGENCY TEACUP AND SAUCER. The lift

value of 16.40 signifies that when PINK REGENCY TEACUP AND SAUCER is already in the basket, the probability of buying GREEN REGENCY TEACUP AND SAUCER is 16.40 times higher than when it is not.

ASSOCITION RULE 2

Interpretation:According to the confidence value of 75.05%, ROSES REGENCY TEACUP AND SAUCER is involved in 75.05% of transactions involving GREEN REGENCY TEACUP AND SAUCER. When GREEN REGENCY TEACUP AND SAUCER is already in the basket, the lift value (14.64) indicates that the likelihood of buying ROSES REGENCY TEACUP AND SAUCER is 14.64 times higher.

ASSOCIATION RULE 3

Interpretation:According to the high confidence value of 73.25%, GREEN REGENCY TEACUP AND SAUCER is included in 73.25% of transactions involving ROSES REGENCY TEACUP AND SAUCER. When ROSES REGENCY TEACUP AND SAUCER is already in the basket, the lift value (14.64) indicates that the probability of buying GREEN REGENCY TEACUP AND SAUCER is 14.64 times higher.

TOTAL WORDS 772