Google Collab Link: Online Retail Purchase Pattern

1.0 Executive Summary

This study applies market basket analysis using the Apriori algorithm to uncover hidden purchasing patterns in online retail transactions. The dataset was initially prepared and cleaned, eliminating irregularities and eliminating entries that were either incomplete or irrelevant, in order to guarantee correctness and dependability. Next, frequent itemsets, groups of products that frequently show up together in consumer transactions, were found using the Apriori algorithm. By analyzing these itemsets, we discovered both expected combinations (such as coordinated home decor items) and unexpected associations (like children's toys paired with specialty teas). These trends show how consumers inherently combine things to suit particular requirements or events. Heatmaps and scatter plots were among the visual techniques utilized to further examine the findings. These visualizations facilitate the understanding of intricate relationships in the data by displaying the frequency and strength of connections between various objects. The insights gained from this study can be used to make better judgments about inventory management, targeted promotions, and more efficient product recommendations. This study demonstrates how market basket analysis may transform transaction data into useful tactics that enhance customer satisfaction and increase sales.

2.0 Problem Statement

Retailers who leverage customer purchasing behaviour data can potentially increase sales by up to 25%, yet many still struggle to optimize product placement and inventory management in the store (Gordon, 2025). The root issues lie in retailers insufficient understanding in customer’s buying patterns leading to generic product displays and a one-size-fits-all approach that fails to personalize the customer’s shopping experience (Shukla, 2024). To address this gap, this study employs the Association Rule Mining technique to identify frequent product combinations. This will help retailers optimise inventory, enhance cross-selling, and improve digital product layout.

By using Association Rule Mining, we will investigate customer’s purchasing patterns of an online retail store to identify frequent itemset and the presence of strong association rules. Findings from the analysis will be utilised to apply customer-centric strategies in marketing, optimised inventory management, and strengthen product recommendations.

3.0 Objectives

I. To apply the Apriori algorithm to identify frequent itemsets and uncover meaningful associations between products based on support, confidence, and lift.

II. To generate actionable insights from purchase patterns that support more effective product placement, inventory decisions, and marketing strategies.

4.0 Methodology

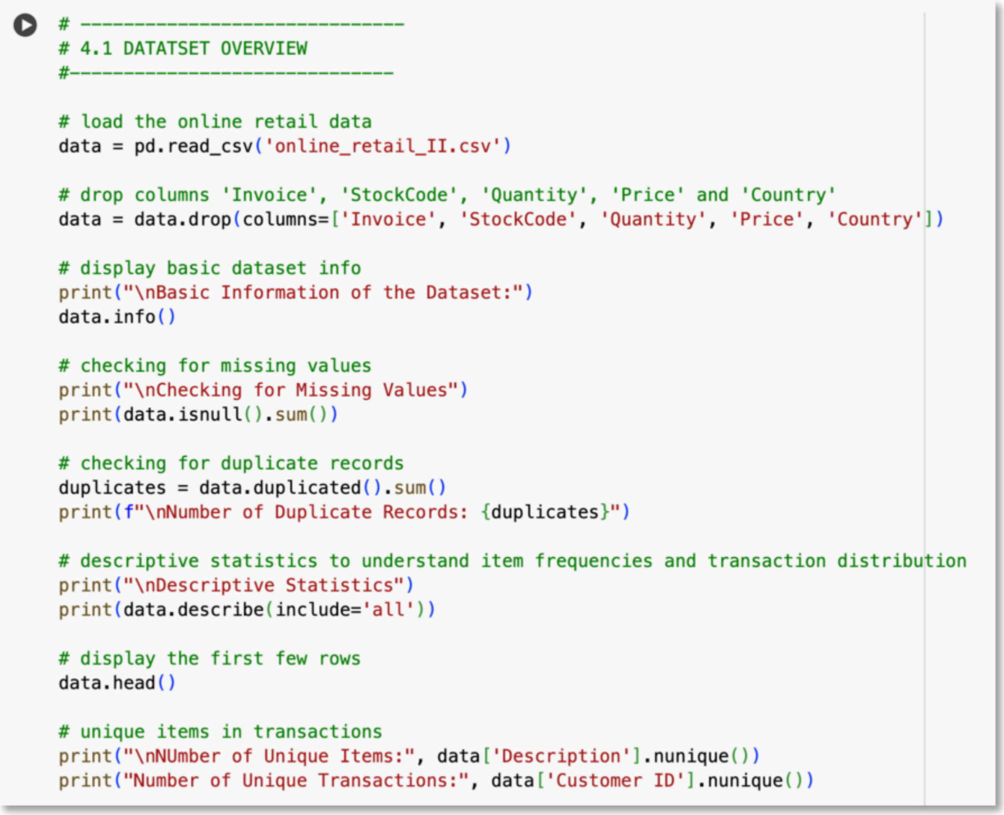
This section provides an overview of the chosen dataset, data preprocessing, data transformation, **frequent itemsets generation with Apriori algorithm and Associate Rules based on frequent itemsets.**

4.1 Dataset Overview

The dataset used in this study is the *Online Retail II UCI Dataset* obtained from Kaggle (Link: https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci). The dataset consists of transaction records from the online retail store. Attributes of the dataset are listed below:

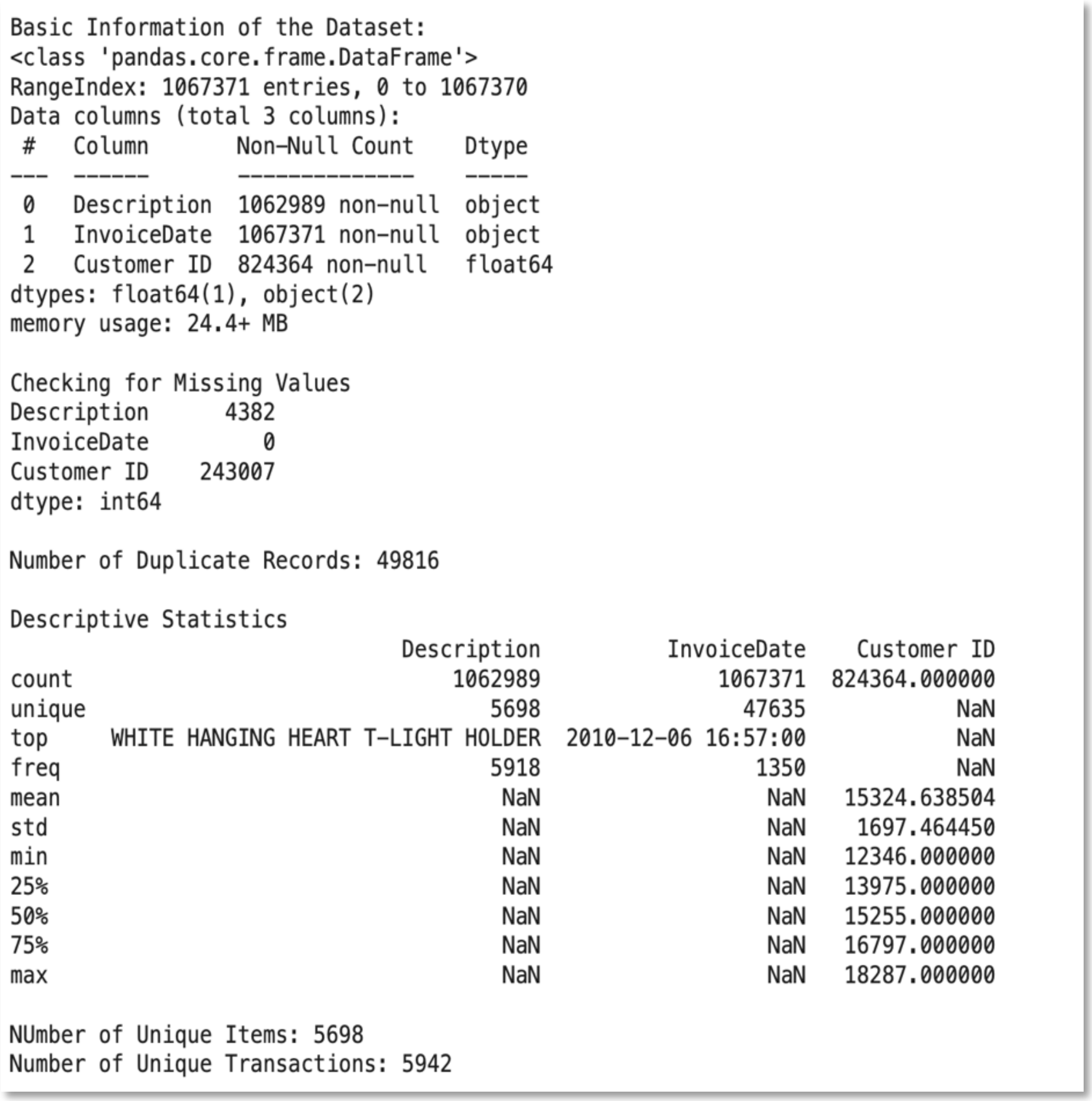
- **CustomerID** – Unique identifier for each customer. This will allow us to analyse customers' purchasing patterns and behaviour.
- **InvoiceDate** – Day and time when a transaction was generated. This will help in deriving insights.
- **Description** – Name of the item made in the transaction. This represents the main unit of analysis for identifying co-purchases.
- Invoice – Unique identifier assigned to each transaction.
- StockCode – Unique identifier assigned to each product.
- Quantity – Total quantity of each item purchased per transaction.
- Price – Price per item.
- Country – Name of country where customer resides

In this study, we will limit our analysis to three key attributes: ‘CustomerID’, ‘InvoiceDate’ and ‘Description’. These attributes were selected as they provide the essential information needed to identify customer purchasing behaviours. Before applying any algorithms, preliminary data inspection is done to screen for missing values, duplicates or any other inconsistencies. Basic statical summaries will also be used to detect potential issues such as incomplete or redundant data points that could impact the accuracy and outcome of the analysis. This step is crucial as it ensures that the dataset is clean, reliable and suitable for meaningful association rule mining and Apriori algorithm.



```
# ------------------------------
# 4.1 DATATSET OVERVIEW
#------------------------------

# load the online retail data
data = pd.read_csv('online_retail_II.csv')

# drop columns 'Invoice', 'StockCode', 'Quantity', 'Price' and 'Country'
data = data.drop(columns=['Invoice', 'StockCode', 'Quantity', 'Price', 'Country'])

# display basic dataset info
print("\nBasic Information of the Dataset:")
data.info()

# checking for missing values
print("\nChecking for Missing Values")
print(data.isnull().sum())

# checking for duplicate records
duplicates = data.duplicated().sum()
print(f"\nNumber of Duplicate Records: {duplicates}")

# descriptive statistics to understand item frequencies and transaction distribution
print("\nDescriptive Statistics")
print(data.describe(include='all'))

# display the first few rows
data.head()

# unique items in transactions
print("\nNUmber of Unique Items:", data['Description'].nunique())
print("Number of Unique Transactions:", data['Customer ID'].nunique())
```

Figure 1 – Code Snippet of the Dataset Overview

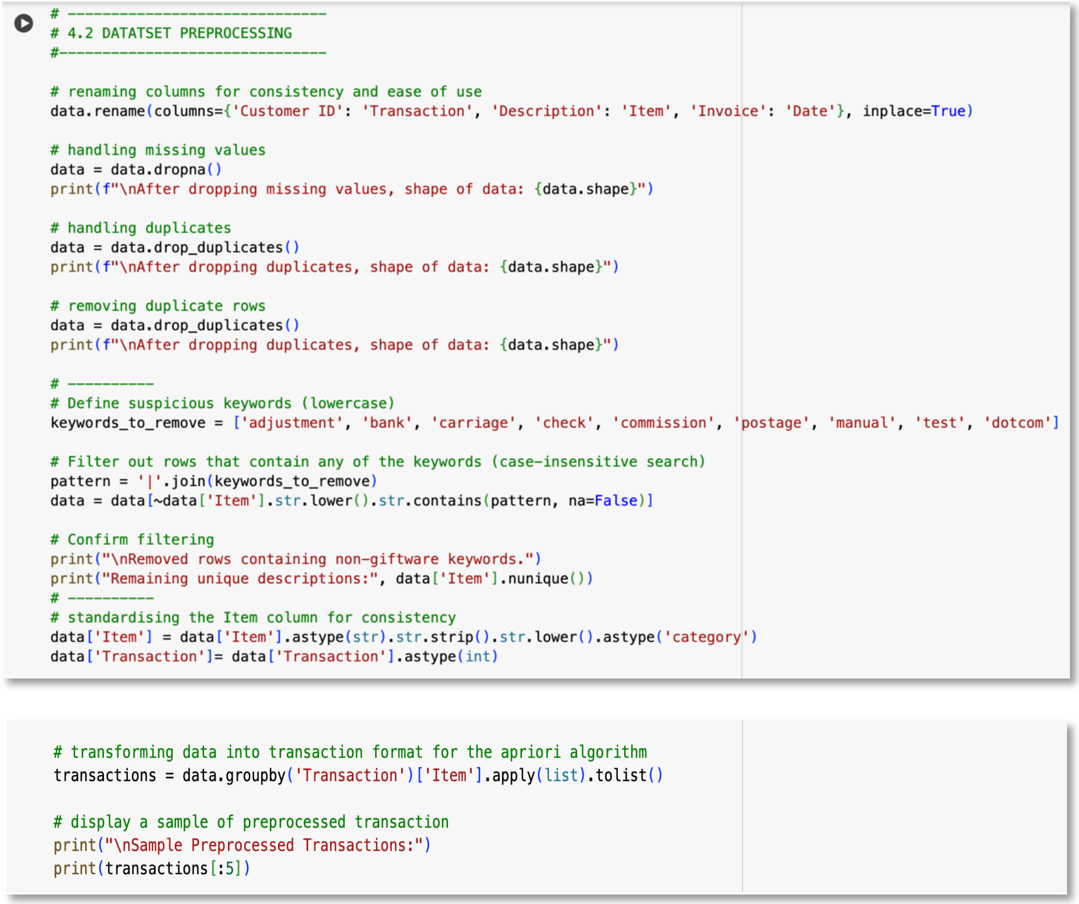


```
Basic Information of the Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1067371 entries, 0 to 1067370
Data columns (total 3 columns):
 #   Column       Non-Null Count    Dtype  
---  ------       --------------    -----  
 0   Description  1062989 non-null  object 
 1   InvoiceDate  1067371 non-null  object 
 2   Customer ID  824364 non-null   float64
dtypes: float64(1), object(2)
memory usage: 24.4+ MB

Checking for Missing Values
Description      4382
InvoiceDate         0
Customer ID    243007
dtype: int64

Number of Duplicate Records: 49816

Descriptive Statistics
                               Description          InvoiceDate    Customer ID
count                              1062989              1067371  824364.000000
unique                                5698                47635            NaN
top     WHITE HANGING HEART T-LIGHT HOLDER  2010-12-06 16:57:00            NaN
freq                                  5918                 1350            NaN
mean                                   NaN                  NaN   15324.638504
std                                    NaN                  NaN    1697.464450
min                                    NaN                  NaN   12346.000000
25%                                    NaN                  NaN   13975.000000
50%                                    NaN                  NaN   15255.000000
75%                                    NaN                  NaN   16797.000000
max                                    NaN                  NaN   18287.000000

NUmber of Unique Items: 5698
Number of Unique Transactions: 5942
```

Figure 2 – Output of the Dataset Overview

4.2 Data Preprocessing

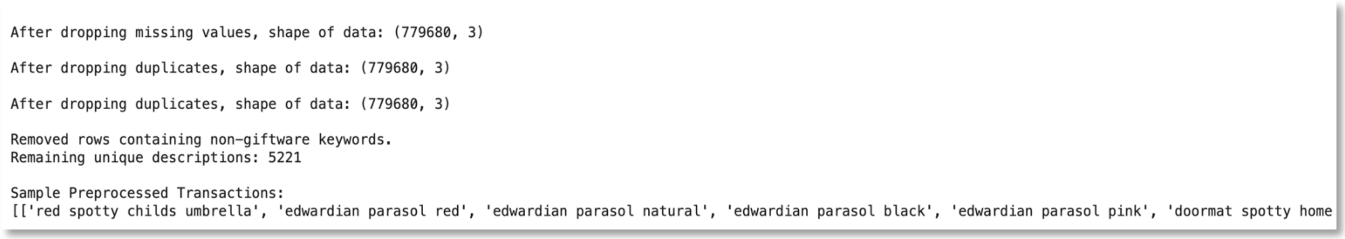
Data preprocessing is done to identify present issues in the dataset such as missing values and duplicate data (Maharan, Mondal, Nemade, 2022). This is done by removing these missing and redundant data points to ensure data quality for better outcome and accuracy of the Apriori algorithm and Associate Rule Mining Technique implemented. Items are standardized (converted to lowercase and spaces are removed) to reduce inconsistencies because of formatting differences (Maharan, Mondal, Nemade, 2022). Then, the data is transformed into a list of lists, each of which represents the goods a consumer has purchased in a single online transaction. **This is so that the frequency of itemsets can be as accurately as possible determined by the Apriori algorithm.**



```
# ------------------------------
# 4.2 DATATSET PREPROCESSING
#------------------------------

# renaming columns for consistency and ease of use
data.rename(columns={'Customer ID': 'Transaction', 'Description': 'Item', 'Invoice': 'Date'}, inplace=True)

# handling missing values
data = data.dropna()
print(f"\nAfter dropping missing values, shape of data: {data.shape}")

# handling duplicates
data = data.drop_duplicates()
print(f"\nAfter dropping duplicates, shape of data: {data.shape}")

# removing duplicate rows
data = data.drop_duplicates()
print(f"\nAfter dropping duplicates, shape of data: {data.shape}")

# ----------
# Define suspicious keywords (lowercase)
keywords_to_remove = ['adjustment', 'bank', 'carriage', 'check', 'commission', 'postage', 'manual', 'test', 'dotcom']

# Filter out rows that contain any of the keywords (case-insensitive search)
pattern = '|'.join(keywords_to_remove)
data = data[~data['Item'].str.lower().str.contains(pattern, na=False)]

# Confirm filtering
print("\nRemoved rows containing non-giftware keywords.")
print("Remaining unique descriptions:", data['Item'].nunique())
# ----------
# standardising the Item column for consistency
data['Item'] = data['Item'].astype(str).str.strip().str.lower().astype('category')
data['Transaction']= data['Transaction'].astype(int)
```

```
# transforming data into transaction format for the apriori algorithm
transactions = data.groupby('Transaction')['Item'].apply(list).tolist()

# display a sample of preprocessed transaction
print("\nSample Preprocessed Transactions:")
print(transactions[:5])
```

Figure 3 – Code Snippet of Data Preprocessing

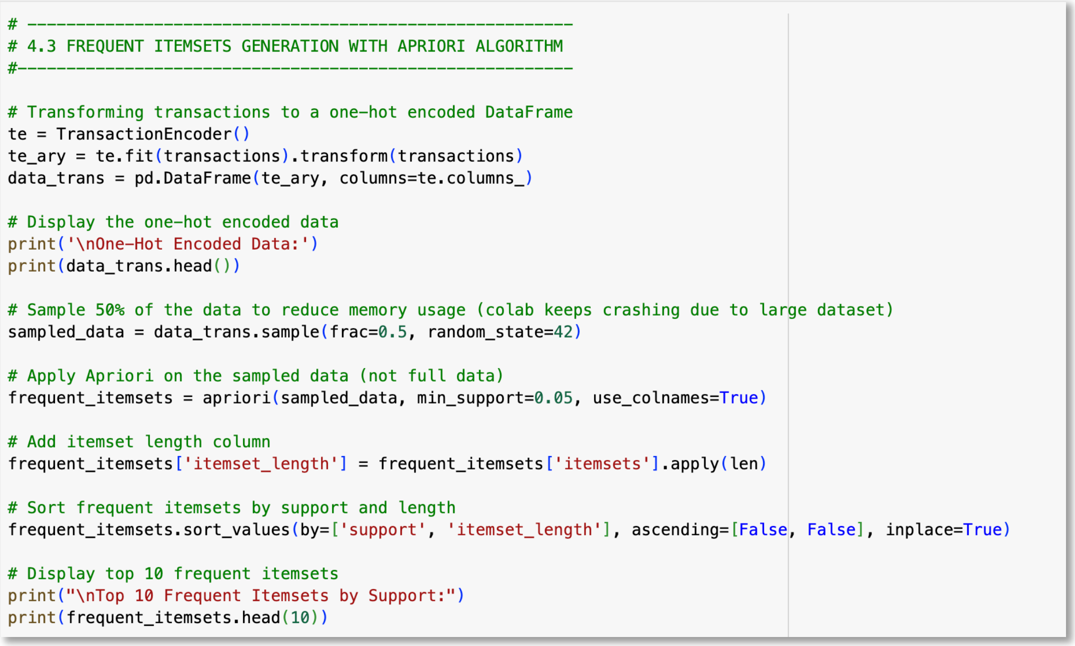


```
After dropping missing values, shape of data: (779680, 3)

After dropping duplicates, shape of data: (779680, 3)

After dropping duplicates, shape of data: (779680, 3)

Removed rows containing non-giftware keywords.
Remaining unique descriptions: 5221

Sample Preprocessed Transactions:
[['red spotty childs umbrella', 'edwardian parasol red', 'edwardian parasol natural', 'edwardian parasol black', 'edwardian parasol pink', 'doormat spotty home
```

Figure 4 – Output of Data Preprocessing

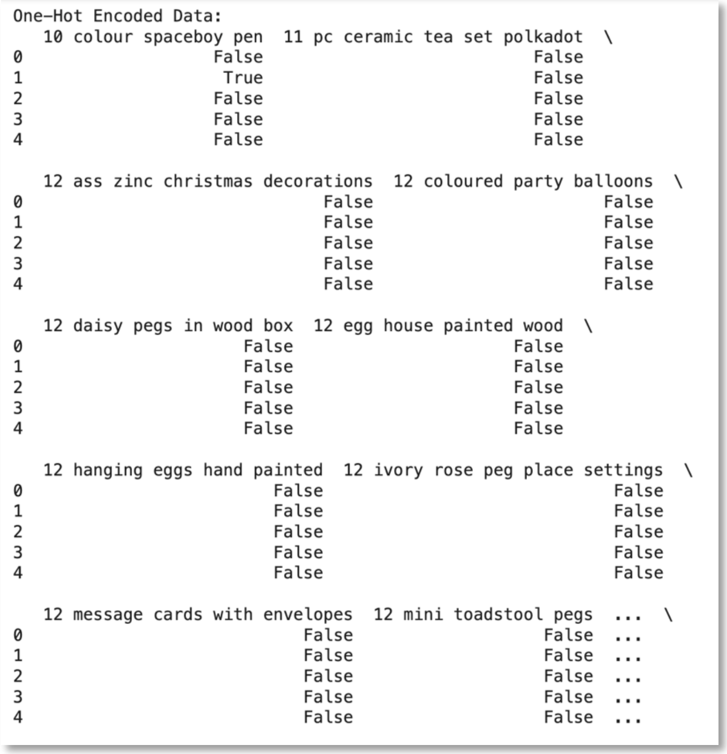
4.3 Frequent Itemsets Generation with Apriori Algorithm

To generate frequent itemsets for this online retail transaction, we will implement **data transformation** to convert the transaction data into one-hot encoded format and apply **Apriori algorithm**. This is then followed by visualisation of the distribution of the itemsets and the top frequent items. The Apriori algorithm is applied to on-hot encoded transactional data in which each column represents an item and each row a transaction. Setting a low support threshold is also crucial in capturing a wider range of item combinations, even the ones that are less common but still relevant to purchasing patterns (Apiletti *et al.*, 2017). The distribution of itemset sizes depicts the plausible combinations items can be grouped together in purchases. These insights form the foundation for identifying popular product combinations and the customer's purchasing behaviours.

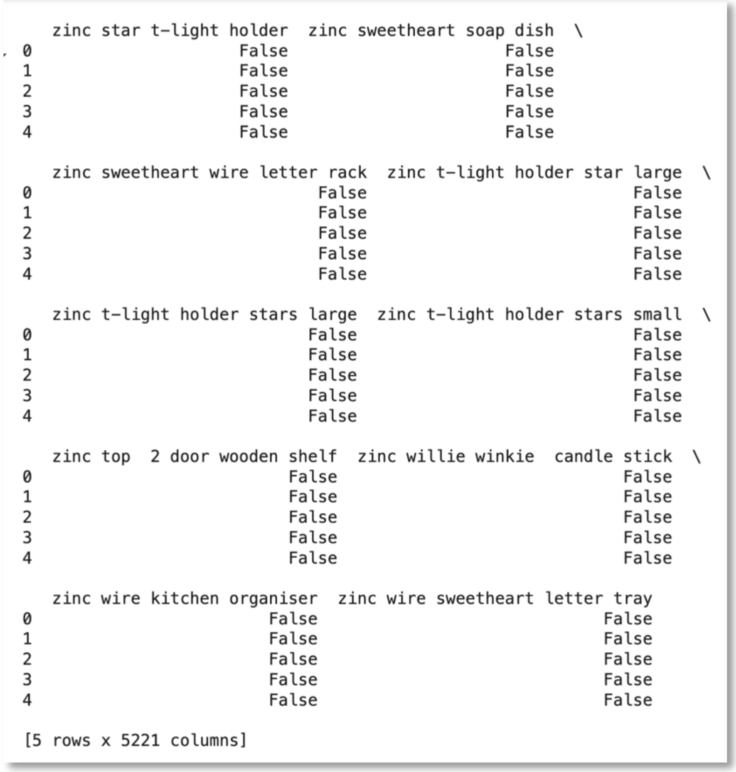


```
# --------------------------------------------------------
# 4.3 FREQUENT ITEMSETS GENERATION WITH APRIORI ALGORITHM
#---------------------------------------------------------

# Transforming transactions to a one-hot encoded DataFrame
te = TransactionEncoder()
te_ary = te.fit(transactions).transform(transactions)
data_trans = pd.DataFrame(te_ary, columns=te.columns_)

# Display the one-hot encoded data
print('\nOne-Hot Encoded Data:')
print(data_trans.head())

# Sample 50% of the data to reduce memory usage (colab keeps crashing due to large dataset)
sampled_data = data_trans.sample(frac=0.5, random_state=42)

# Apply Apriori on the sampled data (not full data)
frequent_itemsets = apriori(sampled_data, min_support=0.05, use_colnames=True)

# Add itemset length column
frequent_itemsets['itemset_length'] = frequent_itemsets['itemsets'].apply(len)

# Sort frequent itemsets by support and length
frequent_itemsets.sort_values(by=['support', 'itemset_length'], ascending=[False, False], inplace=True)

# Display top 10 frequent itemsets
print("\nTop 10 Frequent Itemsets by Support:")
print(frequent_itemsets.head(10))
```

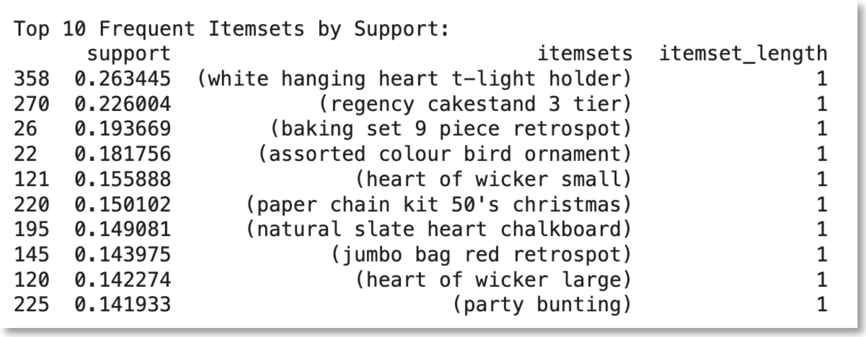
Figure 5 – Code Snippet of Frequent Itemsets Generation with Apriori Algorithm



```
One-Hot Encoded Data:
   10 colour spaceboy pen  11 pc ceramic tea set polkadot  \
0                   False                           False   
1                    True                           False   
2                   False                           False   
3                   False                           False   
4                   False                           False   

   12 ass zinc christmas decorations  12 coloured party balloons  \
0                              False                       False   
1                              False                       False   
2                              False                       False   
3                              False                       False   
4                              False                       False   

   12 daisy pegs in wood box  12 egg house painted wood  \
0                      False                      False   
1                      False                      False   
2                      False                      False   
3                      False                      False   
4                      False                      False   

   12 hanging eggs hand painted  12 ivory rose peg place settings  \
0                         False                             False   
1                         False                             False   
2                         False                             False   
3                         False                             False   
4                         False                             False   

   12 message cards with envelopes  12 mini toadstool pegs  ...  \
0                            False                   False  ...   
1                            False                   False  ...   
2                            False                   False  ...   
3                            False                   False  ...   
4                            False                   False  ...   
```



```
   zinc star t-light holder  zinc sweetheart soap dish  \
0                     False                      False   
1                     False                      False   
2                     False                      False   
3                     False                      False   
4                     False                      False   

   zinc sweetheart wire letter rack  zinc t-light holder star large  \
0                             False                           False   
1                             False                           False   
2                             False                           False   
3                             False                           False   
4                             False                           False   

   zinc t-light holder stars large  zinc t-light holder stars small  \
0                            False                            False   
1                            False                            False   
2                            False                            False   
3                            False                            False   
4                            False                            False   

   zinc top  2 door wooden shelf  zinc willie winkie  candle stick  \
0                          False                             False   
1                          False                             False   
2                          False                             False   
3                          False                             False   
4                          False                             False   

   zinc wire kitchen organiser  zinc wire sweetheart letter tray
0                        False                             False
1                        False                             False
2                        False                             False
3                        False                             False
4                        False                             False

[5 rows x 5221 columns]
```



```
Top 10 Frequent Itemsets by Support:
      support                              itemsets  itemset_length
358  0.263445  (white hanging heart t-light holder)               1
270  0.226004            (regency cakestand 3 tier)               1
26   0.193669        (baking set 9 piece retrospot)               1
22   0.181756       (assorted colour bird ornament)               1
121  0.155888               (heart of wicker small)               1
220  0.150102      (paper chain kit 50's christmas)               1
195  0.149081      (natural slate heart chalkboard)               1
145  0.143975             (jumbo bag red retrospot)               1
120  0.142274               (heart of wicker large)               1
225  0.141933                       (party bunting)               1
```

Figure 6 – Output of Frequent Itemsets Generation with Apriori Algorithm

4.4 Associate Rules Based on Frequent Itemsets

Frequent Itemsets is used to create association rules and filter them depending on the support confidence and lift. **Support** is a measure of an item's default popularity that may be computed by dividing the total number of transactions by the number of transactions that contain a specific item. Suppose we want to find support for item B, this can be calculated as:

$$\text{Support (B)}= \frac{\text{Transactions containing (B)}}{\text{Total Transactions}}$$

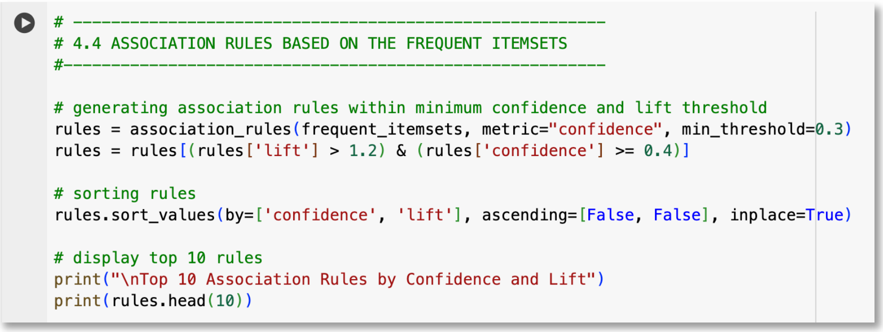
Confidence is the probability that, for instance, if item A is purchased, item B will also be purchased. It can be computed as the number of transactions in which A and B are purchased together, divided by the total number of transactions in which A is purchased. Mathematically, it can be represented as:

$$\text{Confidence (A}\rightarrow\text{B)} = \frac{\text{Transactions containing both (A and B)}}{\text{Transactions containing (A)}}$$

Lastly, **Lift** refers to the increase in the ratio of, for example, sale of item B when item A is sold. It tells us that the likelihood of buying item B and item A together is XXX times more than the likelihood of just buying item B. Lift of greater than 1 means products A and B are more likely to be purchased together whereas a Lift of less than 1 indicates the two products are unlikely to be bought together. Lift can be computed by dividing Confidence(A -> B) divided by Support(B). Mathematically it can be represented as:

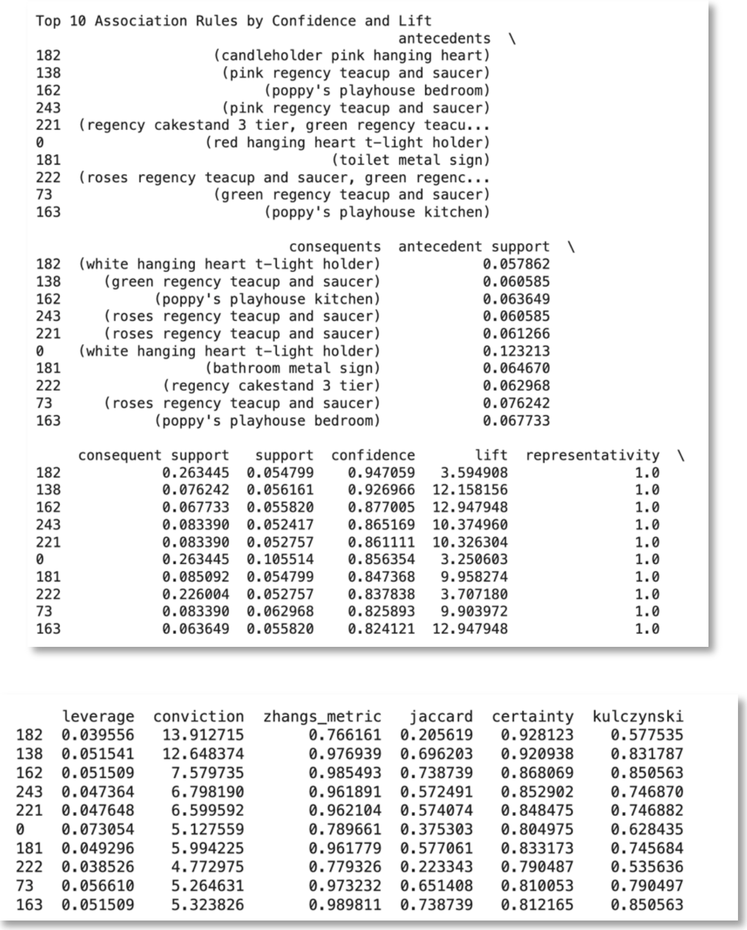
$$\text{Lift (A}\rightarrow\text{B)} = \frac{\text{Confidence (A}\rightarrow\text{B)}}{\text{(Support (B)}}$$

We examine these rules and visualise to spot insightful patterns and trends in the customer's purchasing behaviour. To ensure that only the most dependable and practical rules are left, association rules a created by setting minimum tresholds for confidence and lift. Sorting by confidence and lift shows us the most valuable associations, and a scatter plot visually shows the relationship between support, confidence, and lift. High-lift, high-confidence rules are especially useful as they display strong connections between items. This can support retail strategies like personalised product recommendations.



```
# ------------------------------------------------------------
# 4.4 ASSOCIATION RULES BASED ON THE FREQUENT ITEMSETS
#-------------------------------------------------------------

# generating association rules within minimum confidence and lift threshold
rules = association_rules(frequent_itemsets, metric="confidence", min_threshold=0.3)
rules = rules[(rules['lift'] > 1.2) & (rules['confidence'] >= 0.4)]

# sorting rules
rules.sort_values(by=['confidence', 'lift'], ascending=[False, False], inplace=True)

# display top 10 rules
print("\nTop 10 Association Rules by Confidence and Lift")
print(rules.head(10))
```

Figure 7 – Code Snippet of Association Rules Based on Frequent Itemset



```
Top 10 Association Rules by Confidence and Lift
                                           antecedents  \
182                 (candleholder pink hanging heart)
138                  (pink regency teacup and saucer)
162                       (poppy's playhouse bedroom)
243                  (pink regency teacup and saucer)
221  (regency cakestand 3 tier, green regency teacu...
0                  (red hanging heart t-light holder)
181                                (toilet metal sign)
222  (roses regency teacup and saucer, green regenc...
73                  (green regency teacup and saucer)
163                       (poppy's playhouse kitchen)

                            consequents  antecedent support  \
182  (white hanging heart t-light holder)            0.057862
138     (green regency teacup and saucer)            0.060585
162           (poppy's playhouse kitchen)            0.063649
243     (roses regency teacup and saucer)            0.060585
221     (roses regency teacup and saucer)            0.061266
0    (white hanging heart t-light holder)            0.123213
181                (bathroom metal sign)            0.064670
222            (regency cakestand 3 tier)            0.062968
73      (roses regency teacup and saucer)            0.076242
163           (poppy's playhouse bedroom)            0.067733

     consequent support   support  confidence       lift  representativity  \
182            0.263445  0.054799    0.947059   3.594908               1.0
138            0.076242  0.056161    0.926966  12.158156               1.0
162            0.067733  0.055820    0.877005  12.947948               1.0
243            0.083390  0.052417    0.865169  10.374960               1.0
221            0.083390  0.052757    0.861111  10.326304               1.0
0              0.263445  0.105514    0.856354   3.250603               1.0
181            0.085092  0.054799    0.847368   9.958274               1.0
222            0.226004  0.052757    0.837838   3.707180               1.0
73             0.083390  0.062968    0.825893   9.903972               1.0
163            0.063649  0.055820    0.824121  12.947948               1.0
```

```
     leverage  conviction  zhangs_metric   jaccard  certainty  kulczynski
182  0.039556   13.912715       0.766161  0.205619   0.928123    0.577535
138  0.051541   12.648374       0.976939  0.696203   0.920938    0.831787
162  0.051509    7.579735       0.985493  0.738739   0.868069    0.850563
243  0.047364    6.798190       0.961891  0.572491   0.852902    0.746870
221  0.047648    6.599592       0.962104  0.574074   0.848475    0.746882
0    0.073054    5.127559       0.789661  0.375303   0.804975    0.628435
181  0.049296    5.994225       0.961779  0.577061   0.833173    0.745684
222  0.038526    4.772975       0.779326  0.223343   0.790487    0.535636
73   0.056610    5.264631       0.973232  0.651408   0.810053    0.790497
163  0.051509    5.323826       0.989811  0.738739   0.812165    0.850563
```

Figure 8 – Output of Association Rules Based on Frequent Itemset

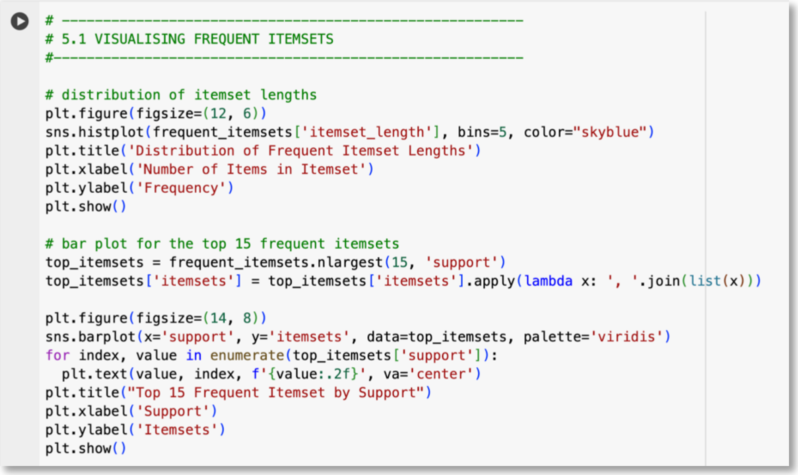
5.0 Results

In this section, we analyse the processed dataset to uncover significant purchasing patterns and trends. By interpreting these findings, we derive actionable insights to support data-driven decision-making in the retail sector. The results not only reveal customer behavior trends but also provide strategic recommendations to foster business growth.

5.1 Visualising Frequent Itemsets

The data is visualized using a **histogram** and **barplot**. The distribution of frequent itemset lengths as seen in *Figure 10* reveals that most transactions consist of single-item purchases. In some cases, a half of the transactions included purchase of two items together, while longer combinations of itemset (3 or more) are less common within the online retail customers. This pattern suggests that customers primarily make focused purchases, likely buying individual needs or complementary items rather than collecting multiple unrelated products in a single order.

The top 15 most frequent item combinations, illustrated in *Figure 11* shows the specific items and groups of items that are purchased together often; these include *'white hanging heart t-light holder'*, *'regency cake stand 3-tier'* and *'baking set 9 piece retrospot'*. *'White hanging heart t-light holder'* is most popular as it is in 26% of all purchases with *'regency cake stand 3-tier'* trailing closely behind as the second most popular item in 23% of all purchases. Overall, this bar chart gives a detailed view of online retail store's best-selling products.



```
# ----------------------------------------------------------
# 5.1 VISUALISING FREQUENT ITEMSETS
#-----------------------------------------------------------

# distribution of itemset lengths
plt.figure(figsize=(12, 6))
sns.histplot(frequent_itemsets['itemset_length'], bins=5, color="skyblue")
plt.title('Distribution of Frequent Itemset Lengths')
plt.xlabel('Number of Items in Itemset')
plt.ylabel('Frequency')
plt.show()

# bar plot for the top 15 frequent itemsets
top_itemsets = frequent_itemsets.nlargest(15, 'support')
top_itemsets['itemsets'] = top_itemsets['itemsets'].apply(lambda x: ', '.join(list(x)))

plt.figure(figsize=(14, 8))
sns.barplot(x='support', y='itemsets', data=top_itemsets, palette='viridis')
for index, value in enumerate(top_itemsets['support']):
  plt.text(value, index, f'{value:.2f}', va='center')
plt.title("Top 15 Frequent Itemset by Support")
plt.xlabel('Support')
plt.ylabel('Itemsets')
plt.show()
```

Figure 9 – Code snippet of visualization frequent itemsets

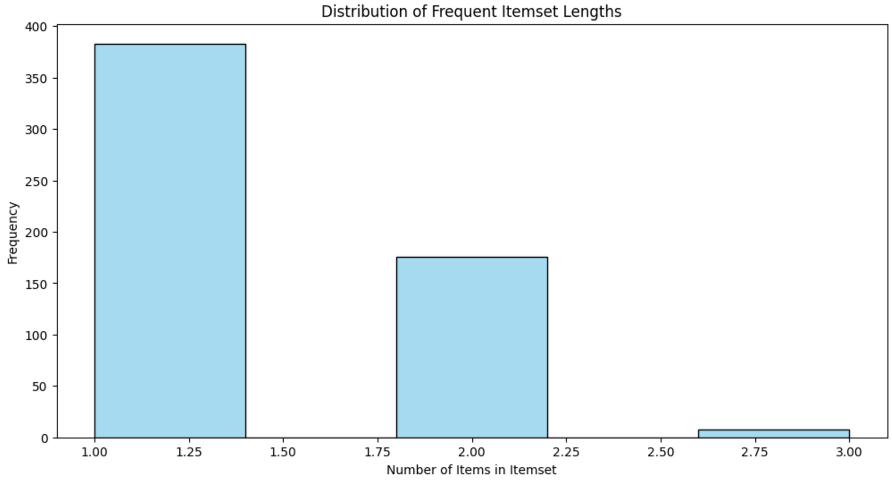


Figure 10 – Histogram showing the distribution of frequent itemset lengths

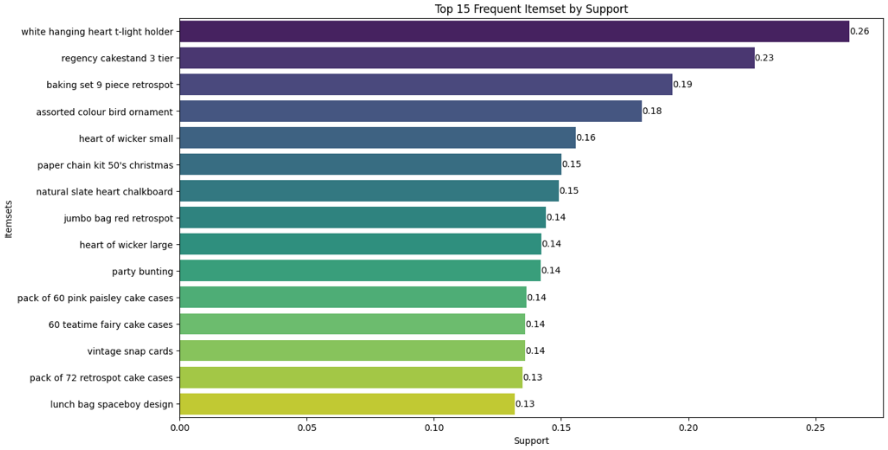


Figure 11 – Barplot showing the Top 15 frequent itemsets by support

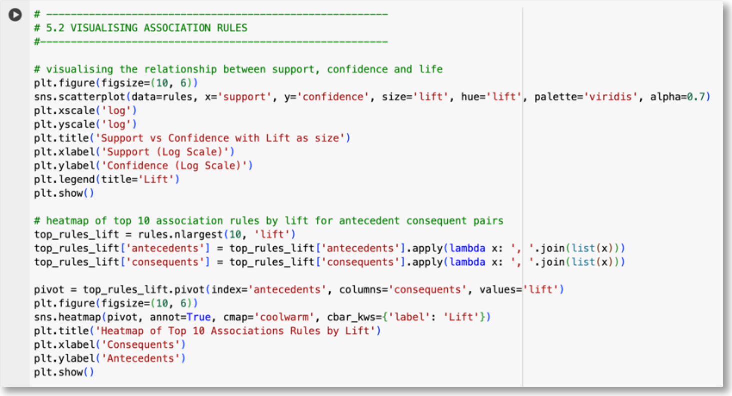
5.2 Visualising Association Rules

To visualise the association rules, we will focus on the relationships between support, confidence and lift. The relationship between the three is visualised using a **scatterplot** and **heatmap**.

Figure 13 displays support on the x-axis and confidence on the y-axis, with point size and color representing lift. The plot shows that while rules with moderate to low support can show high confidence, those with strong support frequently show lower confidence. Interestingly, low support and confidence levels are where high-lift rules cluster, suggesting that while these item combinations are less often, they do show strong connections when they do occur.

A heatmap of the top 10 rules by lift as seen in *Figure 14* gives a detailed view of the strongest rules. For example, the rule 'poppy's playhouse kitchen' leading to 'poppy's playhouse bedroom' produces the highest lift (13.0), indicating a strong association. The rule 'regency cakestand 3 tier, roses regency teacup and saucer' leading to 'green regency teacup and saucer' has a relatively lower lift (11.0), but still indicates a strong association. The color gradient shows lift values, darker colors represent lower lift and lights colors represent higher lift.

These visualizations complement each other effectively by offering both a broad overview of rule strength and a focused lens on high-lift association rules. Together, they enhance our understanding of frequent patterns with strong predictive power. This dual perspective is particularly valuable for identifying meaningful customer behaviors, enabling more precise and impactful targeted marketing strategies.



```
# ----------------------------------------------------------
# 5.2 VISUALISING ASSOCIATION RULES
#-----------------------------------------------------------

# visualising the relationship between support, confidence and life
plt.figure(figsize=(10, 6))
sns.scatterplot(data=rules, x='support', y='confidence', size='lift', hue='lift', palette='viridis', alpha=0.7)
plt.xscale('log')
plt.yscale('log')
plt.title('Support vs Confidence with Lift as size')
plt.xlabel('Support (Log Scale)')
plt.ylabel('Confidence (Log Scale)')
plt.legend(title='Lift')
plt.show()

# heatmap of top 10 association rules by lift for antecedent consequent pairs
top_rules_lift = rules.nlargest(10, 'lift')
top_rules_lift['antecedents'] = top_rules_lift['antecedents'].apply(lambda x: ', '.join(list(x)))
top_rules_lift['consequents'] = top_rules_lift['consequents'].apply(lambda x: ', '.join(list(x)))

pivot = top_rules_lift.pivot(index='antecedents', columns='consequents', values='lift')
plt.figure(figsize=(10, 6))
sns.heatmap(pivot, annot=True, cmap='coolwarm', cbar_kws={'label': 'Lift'})
plt.title('Heatmap of Top 10 Associations Rules by Lift')
plt.xlabel('Consequents')
plt.ylabel('Antecedents')
plt.show()
```

Figure 12 – Code Snippet for Visualising Association Rules

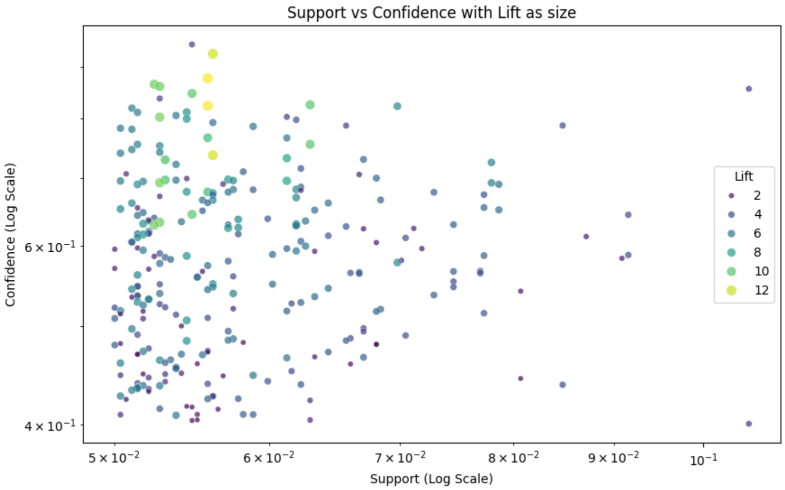


Figure 13 – Scatterplot showing Support vs Confidence with lift as size

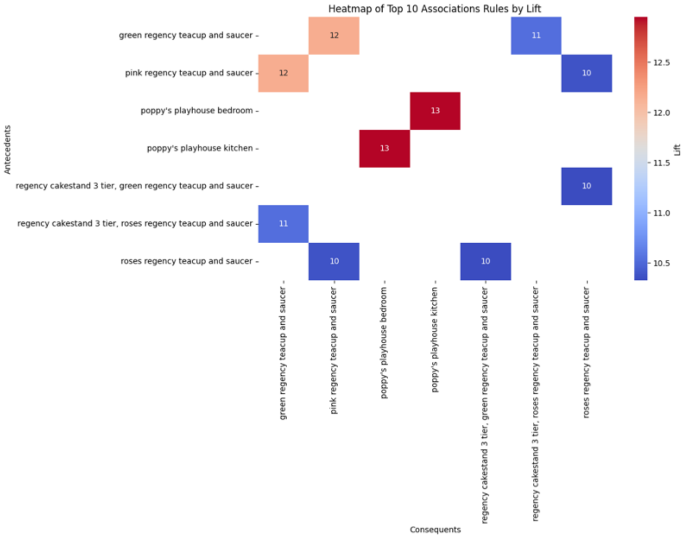


Figure 14 – Heatmap of Top 10 Association Rules by Lift

5.3 Results, Interpretation, Insights & Recommendation

The market basket analysis conducted using the Apriori algorithm on the online retail store dataset has uncovered several insightful patterns in customer purchasing behavior. The visualization of frequent itemsets combinations and the association rules shows the underlying purchasing habits of the customers and retailers can use insights derived from the analysis and output for improve targeted cross-selling strategies, inventory management, and provide optimal personalised marketing efforts to engage customers.

The analysis revealed that most frequent itemsets consisted of single items, with 358 itemsets of length 1 and only a small number of larger itemsets. The top 5 items by support (frequency of occurrence in transactions) are:

- White hanging heart T-light holder (Support: 26%)
- Regency cakestand 3 tier (Support: 23%)
- Baking set 9 piece retrospot (Support: 19%)
- Assorted colour bird ornament (Support: 18%)
- Heart of wicker small (Support: 16%)

Although the frequent itemset analysis revealed a large number of single-item purchases, the presence of multi-item sets—though limited in number—provided key insights into how customers make purchasing decisions. Among the top items by support, the *White hanging heart T-light holder* appeared in approximately 26% of transactions, making it a clear bestseller. Similarly, the *Regency cakestand 3 tier*, *Baking set 9 piece retrospot*, and *Assorted colour bird ornament* each appeared in roughly 18–23% of transactions. These numbers, though calculated from a reduced dataset (half the original due to size constraints), still indicate strong product appeal and repeated customer interest. The high support levels suggest that these products either meet a consistent functional or aesthetic need or benefit from prominent placement in the store, both physically and digitally.

From the store's perspective, these high-frequency items should be considered anchor products, items that reliably drive volume. They likely play a central role in shopping carts, either as standalone purchases or as a base around which other items are selected. For inventory management, this means maintaining a healthy stock buffer for these products is essential to avoid missed sales opportunities. Overstocking low-frequency items while running out of bestsellers like the *T-light holder* could reduce overall profitability and customer satisfaction.

Moving into association rules, what stands out is the exceptionally high confidence and lift values among specific product pairs. For example, customers who bought the *Candleholder pink hanging heart* also bought the *White hanging heart T-light holder* with 94.7% confidence and a lift of 3.59. This indicates not just a tendency, but a strong likelihood that these products are bought together intentionally. The lift value of 3.59 tells us that this purchase pattern happens 3.6 times more often than if the two items were bought independently. This strongly suggests a complementary relationship between the items, likely driven by matching themes or visual aesthetics.

In a similar vein, the *Green regency teacup and saucer* and the *Pink regency teacup and saucer* have a confidence of over 92% and a lift of 12.16, suggesting that when a customer chooses one, they very often want the other. This insight reflects consumer behavior that values completeness and coordination. From the customer's perspective, this pattern hints at a desire for curated or harmonious designs, possibly for gifts, home décor, or personal use where style consistency is important.

It's worth noting that these insights are drawn from only **half of the original dataset**, a necessary compromise due to data volume. As a result, there may be additional frequent patterns or rule structures in the full dataset that were not captured. For example, low-frequency but high-value item pairs may have been underrepresented, and seasonal or niche buying patterns could have been lost in the reduction. Thus, while the patterns identified here are statistically robust within the sample, they should be validated against the full dataset before scaling marketing or inventory decisions.

5.4 Sales Improvement Recommendation for Online Store Manager

- '*White Hanging Heart T-light Holder'* appears in 26.3% of all transactions, making it the most frequently purchased item in the dataset.
  i. This product shows a particularly strong association with the "*Candleholder Pink Hanging Heart*," with which it is purchased 8.4% of the time. The rule {*Candleholder Pink Hanging Heart*} → {*White Hanging Heart T-light Holder*} carries an exceptionally high confidence of 94.7% and a lift of 3.59, indicating that customers who purchase one are highly likely to purchase the other.
  ii. **Recommendation:** Bundled offerings such as a "Rustic Heart Candle Set" to drive higher cart values and reduce decision fatigue for shoppers. Moreover, the store should integrate this rule directly into its recommendation engine, so that when a customer views the pink candleholder, they are immediately shown the white *T-light* holder with a contextual nudge like "Complete your cozy collection."

- The rule {*Poppy's Playhouse Bedroom*} → {*Poppy's Playhouse Kitchen*} has a support of 4.7%, confidence of 87.7%, and an exceptionally high lift of 12.95
  i. This makes it one of the most powerful product relationships discovered in this dataset. This indicates that customers who buy one set are exceedingly likely to buy the other, especially compared to random chance.
  ii. **Recommendation:** Develop a thematic marketing campaign such as "Build Your Playhouse," highlighting the two products as complementary parts of a full set. Campaigns like this, delivered via email or social media with engaging, play-driven imagery, could significantly increase repeat purchases and conversion rates from browsing users

- The rule {*Green Regency Teacup and Saucer*} and {*Pink Regency Teacup and Saucer*} show a strong complementary pattern, with a support of 7.8%, a confidence of 92.5%, and a notably high lift of 12.16.
    i. This suggests that shoppers perceive these items as a matching or collectible set, making them ideal candidates for pre-assembled product bundles.
    ii. **Recommendation:** Market a "Regency Tea Collection." Featuring such bundles prominently on the homepage or in a seasonal gifting section could enhance visual merchandising and improve gift-focused conversions. These high-lift itemsets also lend themselves well to retargeting campaigns; customers who purchase one teacup could be shown ads featuring the other, encouraging collection-building behavior. This product shows a particularly strong association

6.0 Conclusion

In this study, we used association rule mining to analyze transactional data from an online retail store and uncover patterns in how customers shop and what products they tend to buy together. The implementation of the Apriori algorithm revealed significant associations between products, such as the frequent co-purchase of Poppy's Playhouse sets and Regency teacups. These patterns illuminate how customers naturally combine items, providing actionable opportunities for retailers to refine recommendations, curate targeted bundles, and optimize marketing strategies. Bundling often co-purchased products or displaying high-lift item pairings, for example, might increase average order values while simplifying the buying process. In order to make data-driven decisions about inventory management, marketing campaigns, and personalized merchandising, all of which support increased customer engagement and revenue growth,this investigation highlights the ability of association rule mining to decode purchase habits.

Reference List

Apiletti, D., Baralis, E., Cerquitelli, T., Garza, P., Pulvirenti, F., Venturini, L. (2017) Frequent Itemsets Mining for Big Data: A Comparative Analysis. *Big Data Research*. 9, pp. 67-83. [Accessed 23 May 2025].

Maharan, K., Mondal, S., Nemade, B. (2022) A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*. 3(1), pp. 91-99. [Accessed 23 May 2025].

Gordon, K. (2024) Leveraging Consumer Data and Predictive Analytics to Personalise Shopping Experinces, TechBlocks, *LinkedIn*, 23 April [online]. Available from: https://www.linkedin.com/pulse/leveraging-consumer-data-predictive-analytics-personalize-shopping-ahs2f/ [Accessed 18 May 2025].

Shukla, R. (2024) The Struggle is Real – Why Retail Stores Are Underperforming and How AI Can Help, *Staqu*, 9 April [online]. Available from: https://www.staqu.com/the-struggle-is-real-why-retail-stores-are-underperforming-and-how-ai-can-help/ [Accessed 18 May 2025].

Rubric Marks Distribution Table

Task Criteria	Weightage	Excellent	Good
Understanding of Domain and Business Problem	5%	Demonstrates deep understanding of the domain (retail/finance) and clearly articulates a relevant business problem.	Good understanding of domain and business problem, with minor gaps in clarity or relevance.
Data Collection and Preprocessing	6%	Dataset is relevant and well-prepared with thorough preprocessing; missing values and outliers handled appropriately.	Dataset is mostly relevant, and preprocessing done with few minor issues or lack of detail.
Technique Used (e.g., Association Rule Mining, Time Series, etc.)	7%	Correct technique applied and clearly justified with respect to data and business need.	Appropriate technique applied; justification provided with minor reasoning gaps.
Analysis and Insight Generation	6%	Provides deep insights from analysis; patterns, trends, or associations are meaningful and actionable.	Analysis leads to good insights; some patterns discussed but may lack full depth or clarity.
Visualization and Presentation of Results	3%	Clear and effective visualizations used to support findings; charts are accurate, labeled, and well-integrated into the narrative.	Visualizations are good, mostly clear and helpful in presenting findings, minor issues may exist.
Conclusion and Recommendations	3%	Logical conclusions drawn; provides realistic, data-driven business recommendations.	Good conclusions with relevant recommendations, though slightly lacking in depth or linkage to data.

Total Marks: 30

Explanation:

- The weightage for each task is assigned based on the significance of the task within the overall assignment.
- The marks distribution indicates how the students will be assessed based on their performance in each criterion, with clear descriptions for each performance level.
- This table will help both students and evaluators understand how the grading will be conducted based on the specified tasks and the overall objectives of the assignment.