Course Title: Data Mining
Course Number and Section: CS 619 (CRN: 21329)
Term: Spring 2025
Assignment: Discovering Association Rules in a Retail Dataset

Name: Drona Chaitanya Reddy Gangarapu (U01953703)

Overview of Dataset & Preprocessing

Dataset Description

The “Online Retail” dataset contains UK online retail transactions spanning one year. Each record is a line item on an invoice, with key fields:

- **InvoiceNo**: Transaction identifier (cancellations prefixed with “C”).
- **Description**: Product name.
- **Quantity**, **UnitPrice**, **CustomerID**, **Country**, etc.

Preprocessing Steps

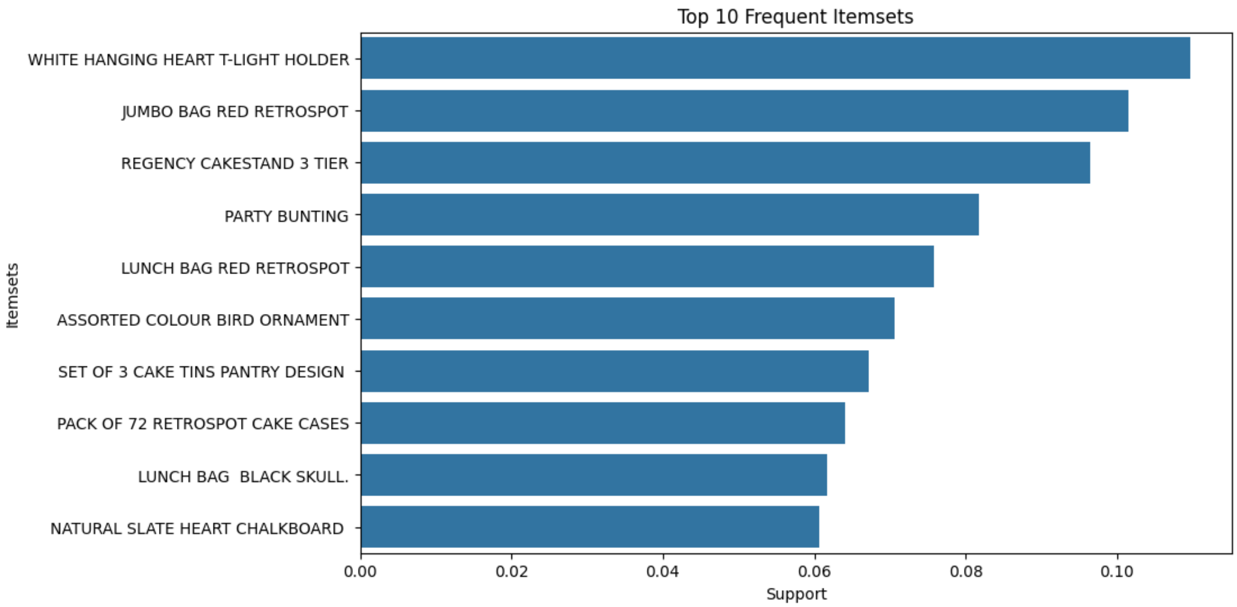
- **Dropped null descriptions**: Removed any rows where **Description** was missing.
- **Filtered cancellations**: Excluded all invoices whose IDs begin with “C.”
- **Standardized product names**: Stripped leading/trailing whitespace from **Description**.
- **Transaction grouping**: Aggregated each invoice’s line items into a list, yielding a list of transactions for mining.

Frequent Itemsets (Support ≥ 1%)

Applied a two-pass mining (1-itemsets, then 2-itemsets) with a minimum support threshold of 0.01 (1%). Below are the **Top 10** by descending support:

Rank	Itemset	Support (%)
1	{WHITE HANGING HEART T-LIGHT HOLDER}	11.0
2	{JUMBO BAG RED RETROSPOT}	10.1
3	{REGENCY CAKESTAND 3 TIER}	9.6
4	{PARTY BUNTING}	8.2
5	{LUNCH BAG RED RETROSPOT}	7.6
6	{ASSORTED COLOUR BIRD ORNAMENT}	7.0
7	{SET OF 3 CAKE TINS PANTRY DESIGN}	6.7
8	{PACK OF 72 RETROSPOT CAKE CASES}	6.4
9	{LUNCH BAG BLACK SKULL.}	6.2
10	{NATURAL SLATE HEART CHALKBOARD}	6.1

Bar chart of top 10 frequent itemsets by support.

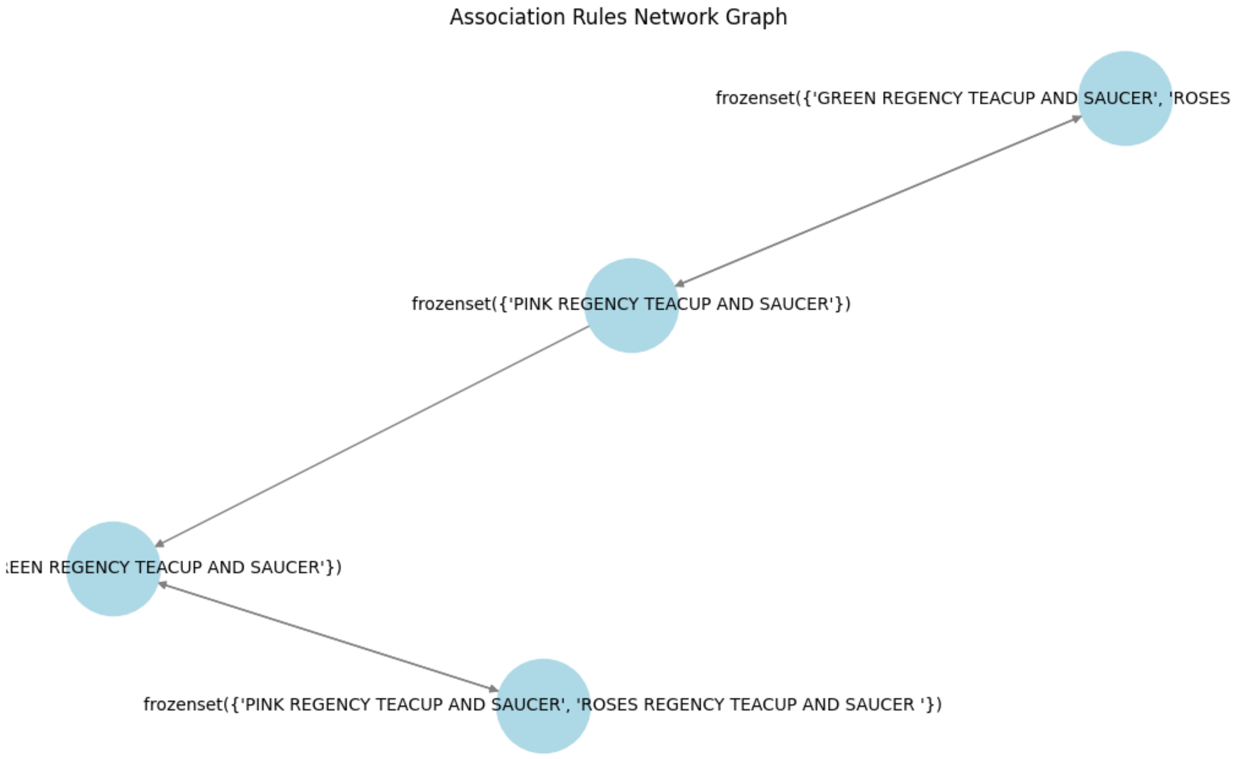


Top 5 Association Rules

From the 2-itemsets, we generated all A→B rules and ranked by **lift**. Below are the top five:

Antecedents	Consequents	Support	Confidence	Lift
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER, **ROSES REGENCY TEACUP AND SAUCER**	0.026298	0.707572	18.988353
GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.026298	0.705729	18.988353
PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.026298	0.904841	18.373184
GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER, **ROSES REGENCY TEACUP AND SAUCER**	0.026298	0.533990	18.373184
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.030713	0.826371	16.779804

Network graph visualizing these top 5 rules (nodes = products; arrows = rules weighted by lift).


Association Rules Network Graph
frozenset({'GREEN REGENCY TEACUP AND SAUCER', 'ROSES
frozenset({'PINK REGENCY TEACUP AND SAUCER'})
EEN REGENCY TEACUP AND SAUCER'})
frozenset({'PINK REGENCY TEACUP AND SAUCER', 'ROSES REGENCY TEACUP AND SAUCER '})

Discussion & Applications

Key Insights

- **Top products** are primarily home décor and giftware (e.g., heart-shaped t-light holders, retrospot bags).
- **Herb-marker items** form a tightly coupled group: customers buying one marker strongly tend to buy the others.

Business Applications

- **Bundling**: Create discounted bundles of the three herb markers to increase cross-sell revenue.
- **Recommendations**: "Customers also bought" modules can leverage A→B rules to suggest thyme when parsley is in cart (and vice versa).
- **Targeted Promotions**: Email campaigns offering special pricing on rosemary markers to customers who previously purchased parsley markers.

Limitations & Next Steps

- Analysis was limited to 1- and 2-itemsets; mining 3+ itemsets could uncover more complex multi-item buying patterns.
- Temporal or regional segmentation (e.g., holiday season vs. off-season) may reveal time-varying associations.
- Incorporate customer demographics for personalized marketing strategies.