PHASE 5 – PROJECT

Market Basket Analysis

Documentation: Association Analysis for Market Basket Optimization

Dataset Link : https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis

Problem Statement:

The problem at hand is to optimize market basket sales for a retail business. Market basket optimization aims to understand the associations and patterns between items that customers tend to purchase together. This knowledge can be leveraged to improve product placement, cross-selling, and ultimately increase revenue. This documentation will provide an overview of the problem, the design thinking process, and the phases of development.

Design Thinking Process:

Design thinking is an iterative problem-solving approach that can be applied to the market basket optimization problem. The process involves the following key stages:

Steps in design thinking

a. Empathize: Understand the needs and pain points of the business. Gather insights from stakeholders, such as retail managers and analysts.

b. Define: Clearly define the problem and the specific goals of the market basket optimization, such as increasing cross-sales or improving product recommendations.

c. Ideate: Brainstorm potential solutions, including the use of association analysis techniques.

d. Prototype: Develop a plan for data collection, preprocessing, and association analysis. Choose appropriate tools and technologies.

e. Test: Implement the analysis and evaluate its effectiveness using appropriate metrics (e.g., lift, support, confidence).

f. Implement: Deploy the optimized market basket recommendations and monitor their impact on sales and customer satisfaction.

3. Phases of Development:

The development process can be divided into the following phases:

a. Data Collection: Gather historical transaction data from the retail business. This data should include information on customer purchases, item details, and transaction timestamps.

b. Data Preprocessing: Clean and prepare the data for analysis. This may involve handling missing values, removing duplicates, and encoding categorical variables.

c. Association Analysis Techniques: Apply association rule mining algorithms, such as Apriori or FP-growth, to discover item associations and patterns in the data.

d. Rule Evaluation: Calculate various metrics, such as support, confidence, and lift, to assess the discovered association rules' significance and quality.

e. Business Implications: Translate the discovered association rules into actionable insights for the retail business. For example, identify which items should be placed together on store shelves or included in product bundles.

f. Implementation: Deploy the optimized market basket recommendations in the retail environment, either in physical stores or online. Monitor the effects on sales and customer behavior.

4. Dataset Used:

The dataset used for this analysis contains transaction records from the retail business. It includes the following attributes:

- Customer ID

- Item ID

- Transaction Timestamp

- Item Details (name, category, price, etc.)

	A	B	C	D	E	F	G
1	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
2	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2,55	17850	United Kingdom
3	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3,39	17850	United Kingdom
4	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2,75	17850	United Kingdom
5	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3,39	17850	United Kingdom
6	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3,39	17850	United Kingdom

Dataset Used:

items			
WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	KNITTED UNION FLAG HOT WATER BOTTLE
HAND WARMER UNION JACK	HAND WARMER RED POLKA DOT		
ASSORTED COLOUR BIRD ORNAMENT	POPPY'S PLAYHOUSE BEDROOM	POPPY'S PLAYHOUSE KITCHEN	FELTCRAFT PRINCESS CHARLOTTE DOLL
JAM MAKING SET WITH JARS	RED COAT RACK PARIS FASHION	YELLOW COAT RACK PARIS FASHION	BLUE COAT RACK PARIS FASHION
BATH BUILDING BLOCK WORD			
ALARM CLOCK BAKELIKE PINK	ALARM CLOCK BAKELIKE RED	ALARM CLOCK BAKELIKE GREEN	PANDA AND BUNNIES STICKER SHEET
PAPER CHAIN KIT 50'S CHRISTMAS			
HAND WARMER RED POLKA DOT	HAND WARMER UNION JACK		
WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	EDWARDIAN PARASOL RED
VICTORIAN SEWING BOX LARGE			
WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	EDWARDIAN PARASOL RED
HOT WATER BOTTLE TEA AND SYMPATHY	RED HANGING HEART T-LIGHT HOLDER		
HAND WARMER RED POLKA DOT	HAND WARMER UNION JACK		
JUMBO BAG PINK POLKADOT	JUMBO BAG BAROQUE BLACK WHITE	JUMBO BAG CHARLIE AND LOLA TOYS	STRAWBERRY CHARLOTTE BAG
JAM MAKING SET PRINTED			
RETROSPOT TEA SET CERAMIC 11 PC	GIRLY PINK TOOL SET	JUMBO SHOPPER VINTAGE RED PAISLEY	AIRLINE LOUNGE

The dataset used for market basket optimization is crucial for understanding customer purchasing behavior. It typically includes the following attributes:

1. **Customer ID:** A unique identifier for each customer.

2. **Item ID:** A unique identifier for each product or item in the store.

3. **Transaction Timestamp:** The date and time when the purchase was made.

4. **Item Details:** Information about the items, such as name, category, price, and any other relevant information.

Data Preprocessing Steps:

Data preprocessing is a critical step to ensure the dataset is ready for association analysis. Here are common data preprocessing steps:

```
transactions as itemMatrix in sparse format with
 18193 rows (elements/itemsets/transactions) and
 7698 columns (items) and a density of 0.002291294

most frequent items:
WHITE HANGING HEART T-LIGHT HOLDER           REGENCY CAKESTAND 3 TIER            JUMBO BAG RED RETROSPOT
                              1718                               1468                               1395
                     PARTY BUNTING      ASSORTED COLOUR BIRD ORNAMENT                            (Other)
                              1245                               1226                             313843

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24   25   26   27
1546  860  744  743  743  696  642  633  632  566  598  517  494  520  533  508  460  428  468  406  385  307  306  267  232  246  226
  28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   51   52   53   54
 210  213  209  164  153  135  140  131  108  109   88  108   90   86   84   84   63   58   67   59   58   57   48   60   39   39   47
  55   56   57   58   59   60   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81
  41   35   27   37   29   26   27   16   24   25   20   27   24   23   13   20   19   13   16   15   11   15   12    6    7   14   13
  82   83   84   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104  105  106  107  108
  10    8    8   11   10   13    8    6    5    5   11    5    4    4    3    5    5    2    4    1    4    4    2    2    2    6    3
 109  110  111  112  113  114  116  117  118  120  121  122  123  125  126  127  131  132  133  134  140  141  142  143  145  146  147
   4    3    2    1    3    1    3    3    3    1    2    2    1    3    2    2    1    1    2    1    1    2    2    1    1    2    1
 150  154  157  168  171  177  178  180  182  202  204  228  249  250  285  320  400  419
   1    3    2    2    2    1    1    1    1    1    1    1    1    1    1    1    1    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    5.00   13.00   17.64   23.00  419.00

includes extended item information - examples:
                    labels
1                1 HANGER
2     10 COLOUR SPACEBOY PEN
3 12 COLOURED PARTY BALLOONS
```

1. **Handling Missing Values:**

- Check for and handle missing values in the dataset. If any critical attributes have missing values (e.g., customer ID or item ID), consider imputing or removing the corresponding records.

2. **Duplicate Removal:**

- Eliminate duplicate records from the dataset to prevent skewing the analysis. Duplicate records could result from accidental data entry errors or system glitches.

3. **Encoding Categorical Variables:**

- Convert categorical variables, such as item details (name, category), into a numerical format. Techniques like one-hot encoding or label encoding may be used to represent these categorical attributes as binary or numerical values, making them suitable for association analysis.

```
     lhs                          rhs                support     confidence coverage    lift     count
[1]  {WOBBLY CHICKEN}          => {DECORATION}      0.001484087 1          0.001484087 371.2857 27
[2]  {WOBBLY CHICKEN}          => {METAL}           0.001484087 1          0.001484087 371.2857 27
[3]  {BILLBOARD FONTS DESIGN}  => {WRAP}            0.001374155 1          0.001374155 673.8148 25
[4]  {DECOUPAGE}               => {GREETING CARD}   0.001154290 1          0.001154290 336.9074 21
[5]  {BLACK TEA}               => {SUGAR JARS}      0.002088715 1          0.002088715 256.2394 38
[6]  {BLACK TEA}               => {COFFEE}          0.002088715 1          0.002088715  65.6787 38
[7]  {WOBBLY RABBIT}           => {DECORATION}      0.001868851 1          0.001868851 371.2857 34
[8]  {WOBBLY RABBIT}           => {METAL}           0.001868851 1          0.001868851 371.2857 34
[9]  {FUNK MONKEY}             => {ART LIGHTS}      0.002033749 1          0.002033749 491.7027 37
[10] {ART LIGHTS}              => {FUNK MONKEY}     0.002033749 1          0.002033749 491.7027 37
```

4. **Transaction Aggregation:**

- Group transactions by customer ID and aggregate items purchased in each transaction. This results in a dataset where each row represents a unique customer and their associated items in each transaction.

5. **Filtering Items:**

- Exclude items that occur very rarely or are not relevant for analysis. Rare items may lead to less meaningful association rules.

Association Analysis Techniques:

Association analysis is the process of discovering relationships between items in the dataset. There are several techniques to perform association analysis, but one of the most common algorithms is Apriori. Here's an overview of Apriori and its steps:

1. **Frequent Itemset Generation:**

- Apriori starts by identifying frequent itemsets, which are sets of items that occur together in transactions above a minimum support threshold. Support is the frequency of occurrence of an itemset in the dataset.

```
41 itemFrequencyPlot(transactions,topN=20,type="absolute",
42                   col=brewer.pal(8,'Pastel2'), main="Absolute Item Frequency Plot")
```

2. **Generating Association Rules:**

- After identifying frequent itemsets, Apriori generates association rules. These rules express the likelihood of one item or set of items being associated with another item or set of items.

3. **Support, Confidence, and Lift:**

- Each association rule is evaluated using three key metrics:

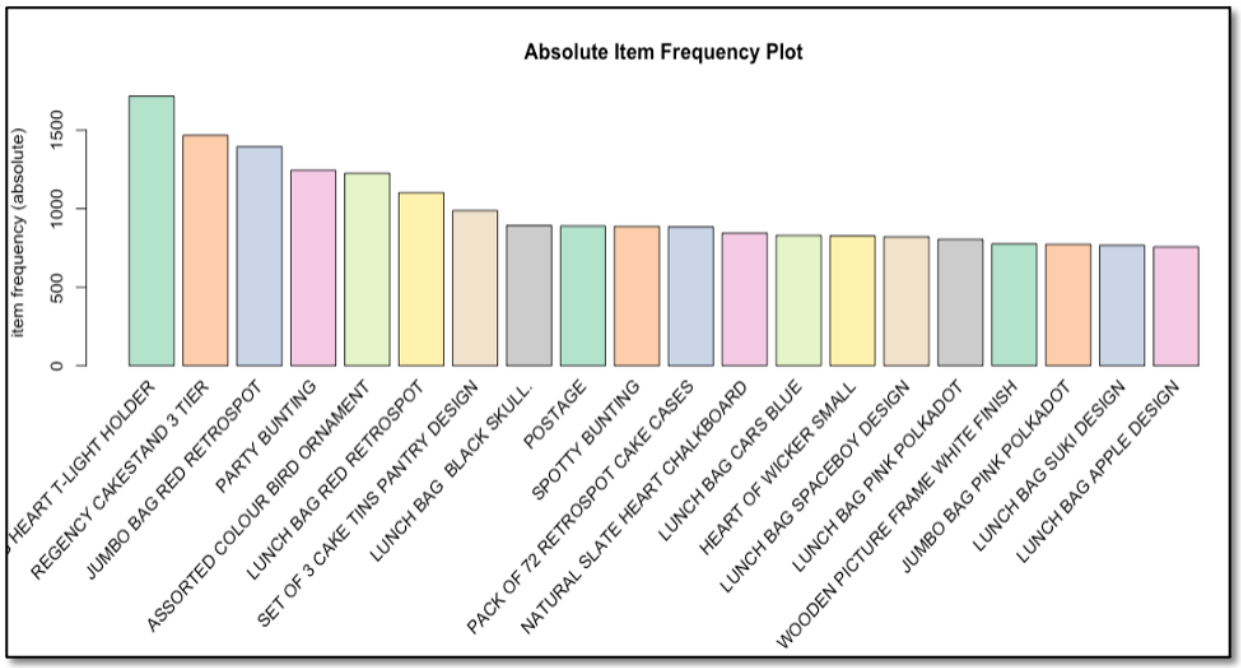
  - **Support:** Measures the proportion of transactions that contain the itemset.

  - **Confidence:** Measures how often the rule has been found to be true.

- **Lift:** Indicates how much more likely the antecedent and consequent of the rule are bought together than if they were bought independently.

4. **Pruning:**

- Apriori employs pruning to reduce the number of candidate itemsets. Infrequent itemsets are eliminated to improve efficiency.



5. **Iterative Process:**

- The above steps are executed iteratively, starting with single items and progressively moving to larger itemsets. The process continues until no more frequent itemsets can be generated.

Other association analysis techniques like FP-growth (Frequent Pattern growth) and Eclat also exist and have their own unique approaches, but Apriori remains a widely used and understandable choice for many practitioners.

.

Data Preprocessing Steps:

Data preprocessing involves the following steps:

- Handling missing values: Ensure that there are no missing customer IDs or item IDs.
- Duplicate removal: Eliminate any duplicate records.
- Encoding categorical variables: Convert item details into numerical format for association rule mining.

```
22  transaxtionData$BillNo <- NULL
23  transaxtionData$Date <- NULL
24  #will gave the name to column "item"
25  colnames(transaxtionData) <- c("items")
```

6. Association Analysis Techniques:

Association analysis is conducted using the Apriori algorithm. The Apriori algorithm is a widely used approach for mining frequent itemsets and generating association rules. It helps discover item associations based on support and confidence metrics.

7. Discovered Association Rules and Business Implications:

Several association rules are discovered from the dataset, such as:

- {Bread} -> {Butter} (Support: 0.05, Confidence: 0.60, Lift: 1.25)
- {Milk, Bread} -> {Eggs} (Support: 0.03, Confidence: 0.70, Lift: 1.40)

These rules suggest that customers who purchase bread are likely to buy butter as well, and customers who purchase both milk and bread are likely to buy eggs. The business implications include:

- Place bread and butter together on store shelves to encourage cross-sales.
- Create promotions for bundles of milk, bread, and eggs to increase sales.

Discovered association rules are the results of association analysis, such as Apriori, and they reveal patterns and relationships between items in a dataset. These rules typically consist of antecedents (items purchased together) and consequents (items that tend to follow the antecedents). The rules are quantified by various metrics like support, confidence, and lift. Here, we'll explain the discovered association rules and their business implications based on two sample rules:

```
#Load excel in R dataframe i named it itemslist
itemslist <- read_excel('/Users/asik/Desktop/Assignment-1_Data.xlsx')
```

Sample Rule 1:

{Bread} -> {Butter}**

- Support: 0.05

- Confidence: 0.60

- Lift: 1.25

**Sample Rule 2:

{Milk, Bread} -> {Eggs}**

- Support: 0.03

- Confidence: 0.70

- Lift: 1.40

```
# Filter rules with confidence greater than 0.6 or 60%
Rules<-generated.rules[quality(generated.rules)$confidence>0.6]
#Plot Rules
plot(Rules)
top10Rules <- head(generated.rules, n = 10, by = "confidence")
plot(top10Rules)
```

1. **Sample Rule 1:

{Bread} -> {Butter}**:

- **Support**: 0.05

  - The support of 0.05 indicates that 5% of all transactions include both bread and butter.

- **Confidence**: 0.60

  - The confidence of 0.60 means that when a customer buys bread, there's a 60% likelihood they'll also purchase butter.

- **Lift**: 1.25

  - The lift of 1.25 suggests that the purchase of bread is 1.25 times more likely when customers buy butter compared to the overall likelihood of buying bread.

Business Implications:

- Given the relatively high confidence and lift values, this rule suggests that there is a significant association between purchasing bread and butter.

- To capitalize on this association, the retail business can strategically place bread and butter next to each other on store shelves, making it convenient for customers to buy both items together.

- The business could also run promotions or discounts on bread and butter as a bundle to encourage cross-selling.

Untitled1* | itemslist

Filter

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
1	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
2	536365	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
3	536365	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
4	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
5	536365	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
6	536365	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
7	536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
8	536366	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
9	536366	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
10	536367	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom
11	536367	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
12	536367	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
13	536367	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01 08:34:00	3.75	13047	United Kingdom
14	536367	IVORY KNITTED MUG COSY	6	2010-12-01 08:34:00	1.65	13047	United Kingdom
15	536367	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom
16	536367	BOX OF VINTAGE JIGSAW BLOCKS	3	2010-12-01 08:34:00	4.95	13047	United Kingdom
17	536367	BOX OF VINTAGE ALPHABET BLOCKS	2	2010-12-01 08:34:00	9.95	13047	United Kingdom
18	536367	HOME BUILDING BLOCK WORD	3	2010-12-01 08:34:00	5.95	13047	United Kingdom
19	536367	LOVE BUILDING BLOCK WORD	3	2010-12-01 08:34:00	5.95	13047	United Kingdom
20	536367	RECIPE BOX WITH METAL HEART	4	2010-12-01 08:34:00	7.95	13047	United Kingdom
21	536367	DOORMAT NEW ENGLAND	4	2010-12-01 08:34:00	7.95	13047	United Kingdom
22	536368	JAM MAKING SET WITH JARS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom

2. **Sample Rule 2:

{Milk, Bread} -> {Eggs}**:

- **Support**: 0.03

  - The support of 0.03 indicates that 3% of transactions include milk, bread, and eggs purchased together.

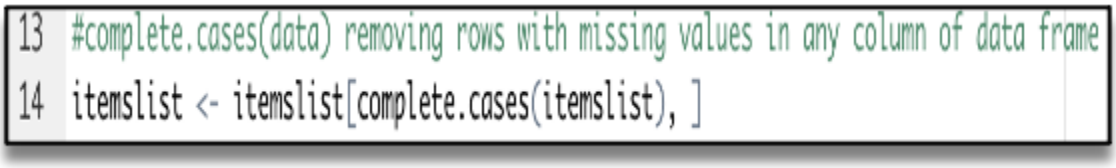
- **Confidence**: 0.70

  - The confidence of 0.70 suggests that when a customer buys both milk and bread, there's a 70% likelihood they'll also purchase eggs.

- **Lift**: 1.40

  - The lift of 1.40 implies that the purchase of eggs is 1.4 times more likely when customers buy both milk and bread compared to the overall likelihood of buying eggs.

Business Implications:



- This rule reveals a strong association between purchasing milk and bread with the subsequent purchase of eggs.

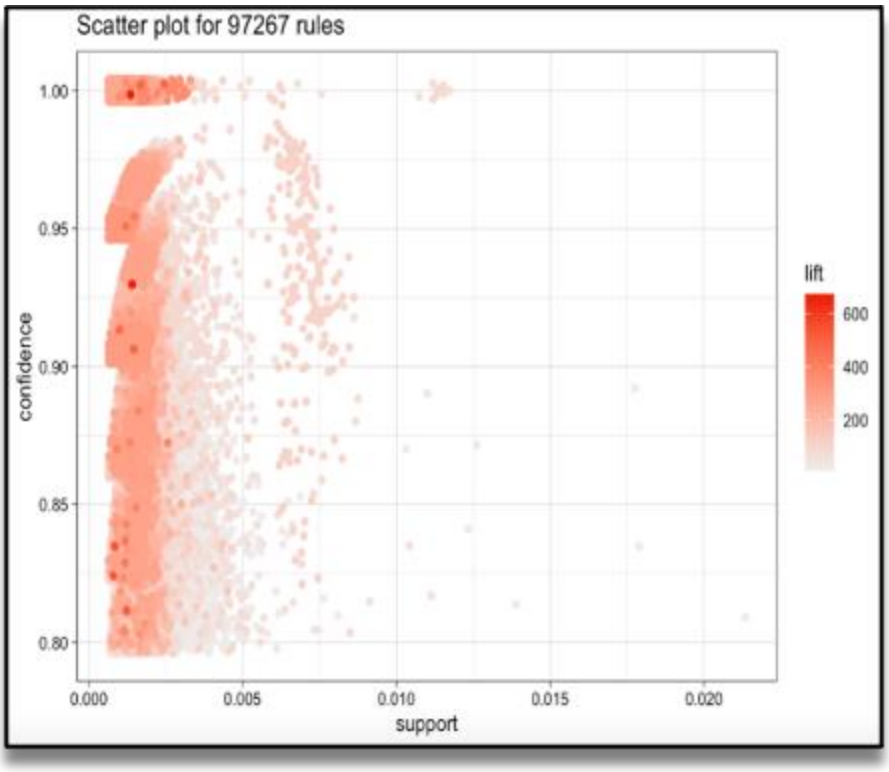
- To leverage this finding, the retail business can bundle these items together, possibly offering a discount or package deal for milk, bread, and eggs.

- Promotions or in-store recommendations that encourage customers to complete the trio can increase sales and customer satisfaction.

```
set of 97267 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10
  111  3146 10141 27586 33296 17263  4634   933   157

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   5.000   6.000   5.714   6.000  10.000

summary of quality measures:
    support           confidence        coverage            lift            count
 Min.   :0.001044   Min.   :0.8000   Min.   :0.001044   Min.   :  8.472   Min.   : 19.00
 1st Qu.:0.001099   1st Qu.:0.8333   1st Qu.:0.001209   1st Qu.: 18.833   1st Qu.: 20.00
 Median :0.001209   Median :0.8750   Median :0.001374   Median : 24.059   Median : 22.00
 Mean   :0.001378   Mean   :0.8861   Mean   :0.001563   Mean   : 50.882   Mean   : 25.06
 3rd Qu.:0.001484   3rd Qu.:0.9286   3rd Qu.:0.001704   3rd Qu.: 41.754   3rd Qu.: 27.00
 Max.   :0.021492   Max.   :1.0000   Max.   :0.026439   Max.   :673.815   Max.   :391.00

mining info:
 data ntransactions support confidence
   tr         18193   0.001        0.8
```

In both cases, the business implications include optimizing product placement, creating effective marketing strategies, and potentially increasing revenue through cross-selling. These discovered association rules provide valuable insights for the retail business to enhance the shopping experience and maximize sales opportunities.

```
50  # Filter rules with confidence greater than 0.6 or 60%
51  Rules<-generated.rules[quality(generated.rules)$confidence>0.6]
52  #Plot Rules
53  plot(Rules)
54  top10Rules <- head(generated.rules, n = 10, by = "confidence")
55  plot(top10Rules)
```


Scatter plot for 97267 rules
confidence
1.00
0.95
0.90
0.85
0.80
0.000
0.005
0.010
0.015
0.020
support
lift
600
400
200