DATA ANALYSIS

FINAL PROJECT

in **E-COMMERCE**

Presented by group 1

MEMBERS

of Group 1

Lê Châu Anh	K214110856
Phạm Thị Anh Thư	K214111979
Nguyễn Hoàng Phương Ngân	K214110864
Phan Quỳnh Trâm	K214110870
Nguyễn Bảo An Nhiên	K214110865

TABLE
of contents

00. Project Overview

01. Theoretical background

02. Data Preparation

03. Experimental

04. Visualization

05. Group Evaluation

Project Overview

Business Problem

- Customer behavior misunderstanding
- Inaccurate demand forecasting
- Predicting future product demands
- Challenge in optimizing and personalizing the online shopping experience

=> Businesses need to deeply understand customer purchasing behavior and then make accurate purchasing suggestions

Business

Objective

1 Leverage transactional data to identify frequent item sets

2 Utilizing market basket analysis to enhance the shopping experience and increase sales

3 Bundle less popular items with top-sellers to boost sales of slow-moving SKUs

Question

Which products are frequently purchased together, and how can we use this information to optimize product placement? 1

Which slow-moving SKUs can be bundled with popular items to enhance visibility and increase sales? 2

Based on a customer's current basket, what additional products can be recommended to optimize sales? 3

How can market basket analysis help create product bundles that appeal to customers and boost sales? 4

Scope

All purchases made for an online gift retail company located in the UK over a thirteen-month duration

Experimental method:

- Market-Basket Analysis, a sort of association rule mining
- Apriori and FP-Growth algorithms
- Support, confidence, and lift metrics

=> Implemented in Python and used PowerBI to visualize findings

Chap 1

Theoretical Background

Theoretical Background

Association Rules Mining

Used to identify patterns in large datasets by finding relationships between variables

Metrics

Support and confidence

FP-Growth Algorithm

Finding frequent items without using candidate generation

Apriori Algorithm

Finding frequent items by using candidate generation

Chap 2

Data Preparation

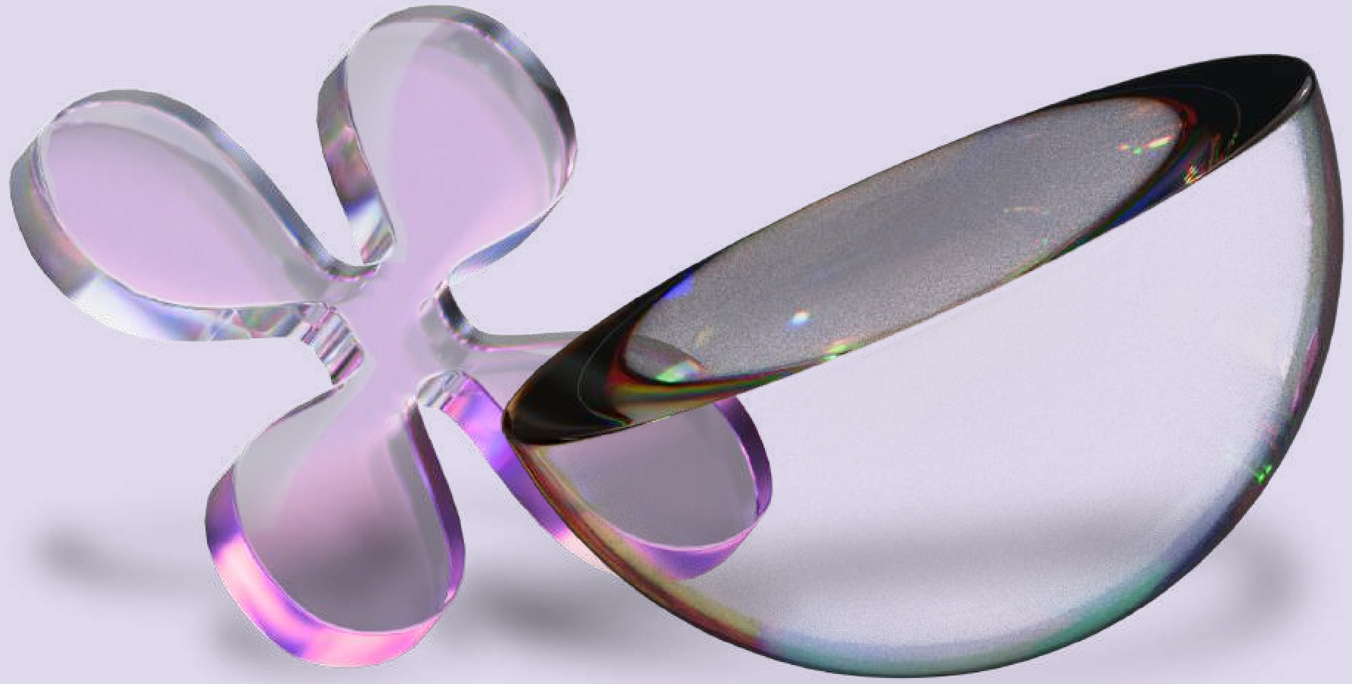
2.1 Data Understanding

Question	Required data
Which products are frequently purchased together, and how can we use this information to optimize product placement?	• Transaction data • Detailed product information
How can market basket analysis help create product bundles that appeal to customers and boost sales?	• Transaction data • Detailed product information
Based on a customer's current basket, what additional products can be recommended to optimize sales?	• Customer data • Detailed product information • Transaction data
Which slow-moving SKUs can be bundled with popular items to enhance visibility and increase sales?	• Transaction data • Detailed product information

To understand customer behavior and effective forecasting, data must be collected from the sales department and understood

2.2 Data Collection

Dataset was sourced from Kaggle, given by a UK-based, physical retail business with purchases recorded from 01/12//2010 to 09/12/2011. It was established in 1981, with the business field supplying souvenirs



2.3 Data Description

Data Overview

index	0	1	2	3	4
InvoiceNo	536365	536365	536365	536365	536365
StockCode	85123A	71053	84406B	84029G	84029E
Description	WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	CREAM CUPID HEARTS COAT HANGER	KNITTED UNION FLAG HOT WATER BOTTLE	RED WOOLLY HOTTIE WHITE HEART.
Quantity	6	6	8	6	6
InvoiceDate	12/1/2010 8:26	12/1/2010 8:26	12/1/2010 8:26	12/1/2010 8:26	12/1/2010 8:26
UnitPrice	2.55	3.39	2.75	3.39	3.39
CustomerID	17850.0	17850.0	17850.0	17850.0	17850.0
Country	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom

2.3 Data Description

Data Statistic

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Data Type

Variable Name	Role	Type	Description
InvoiceNo	ID	Categorical	A 6-digit integral number uniquely assigned to each transaction. If this code starts with 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 quantities of each product (item) per transaction
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

2.4 Exploratory Data Analysis

Column name	Number of missing values
InvoiceNo	0
StockCode	0
Description	1454
Quantity	0
InvoiceDate	0
UnitPrice	0
CustomerID	135080
Country	0

a. Determine missing values in data

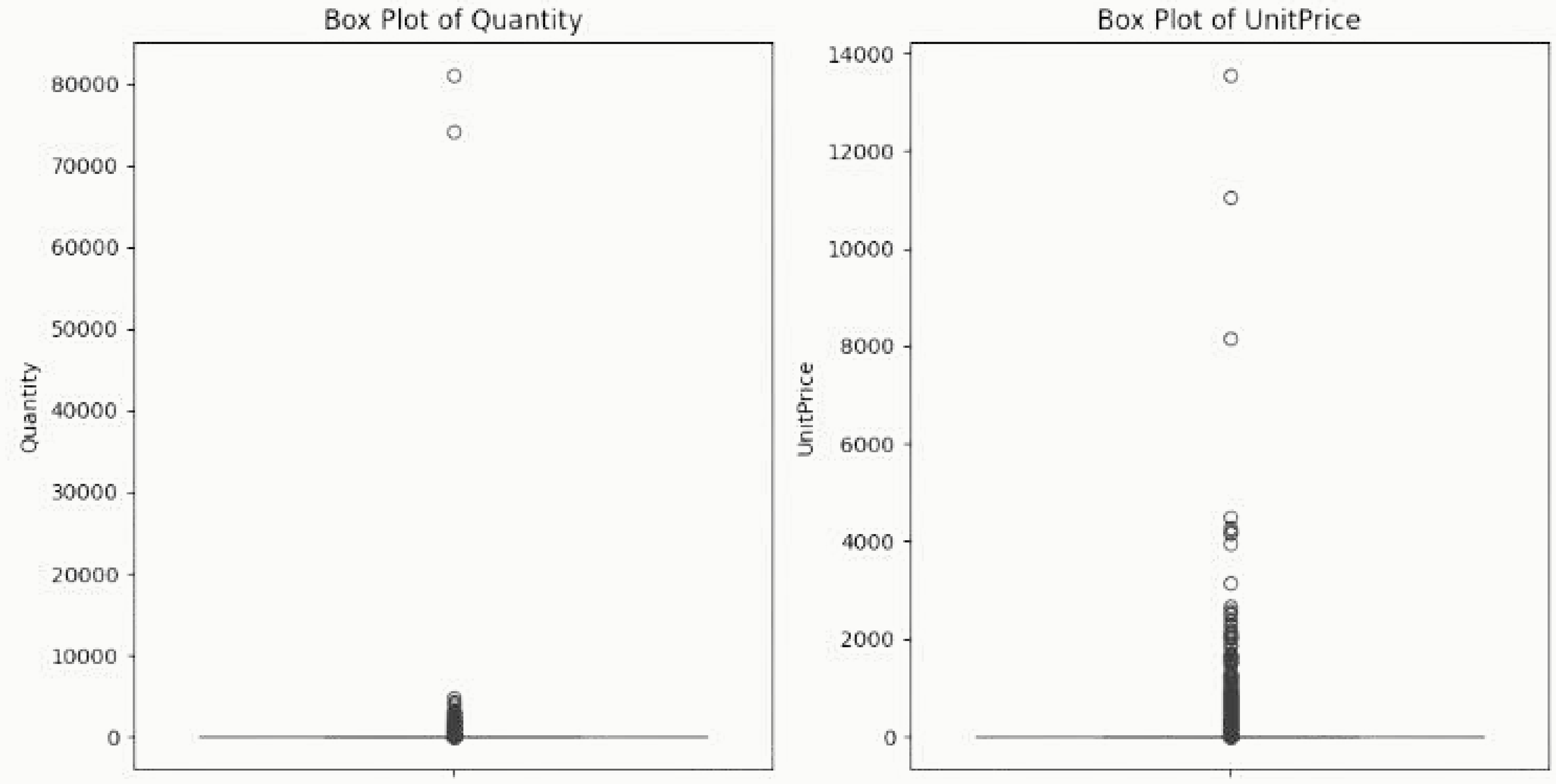
c. Skewness and Kurtosis

	Skewness	Kurtosis
Quantity	-0.26407557610510857	119768.05495538174
UnitPrice	186.50645547026195	59005.1746626736

b. Correlations matrix to identify relationships

	Quantity	UnitPrice
Quantity	1.0	-0.0012349245448703343
UnitPrice	-0.0012349245448703343	1.0

2.4 Exploratory Data Analysis



Identify Initial Pattern, Trends, or Anomalies

Quantity Outliers: 15574

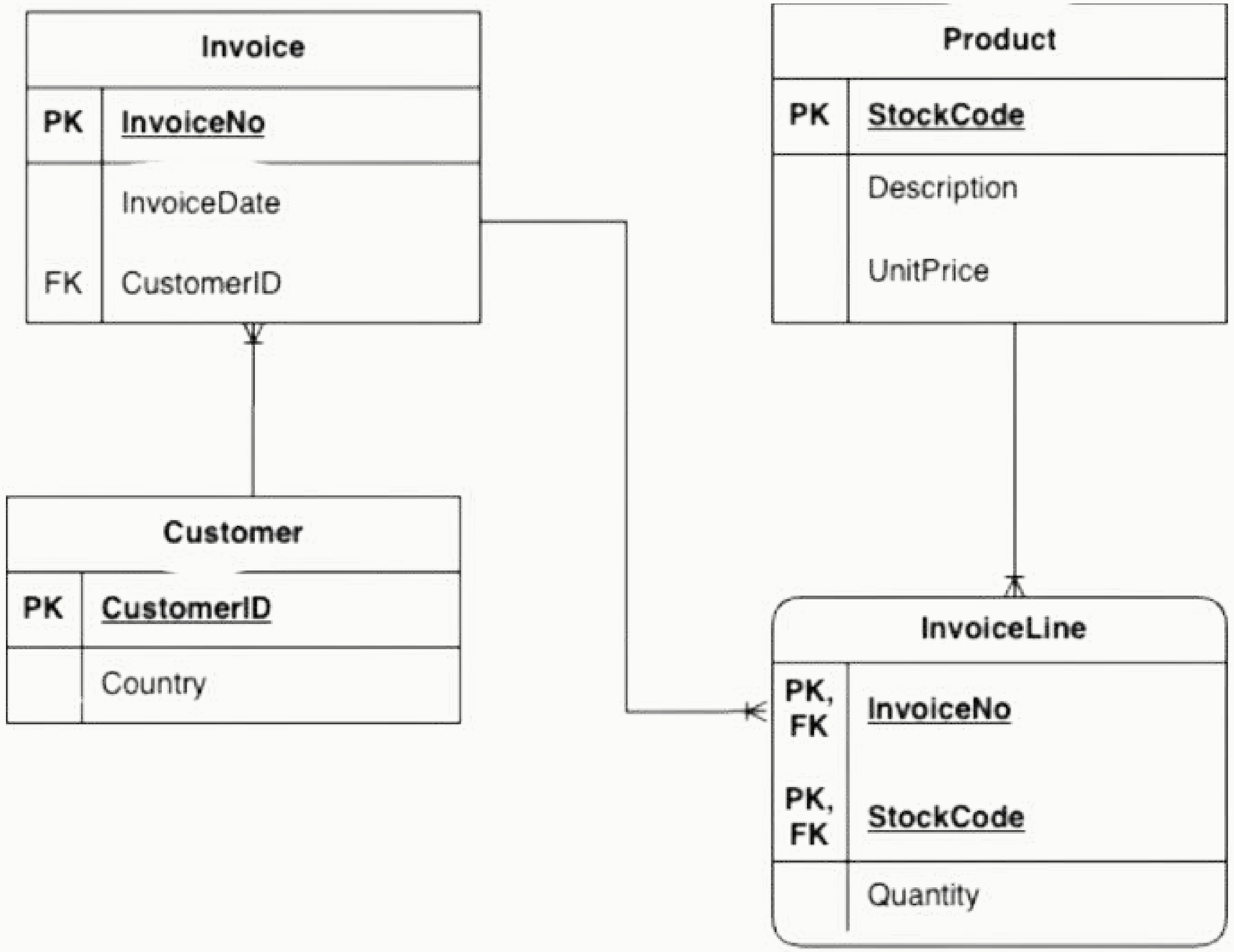
UnitPrice Outliers: 4792

2.5 Preprocessing

CLEANING AND TRANSFORMING

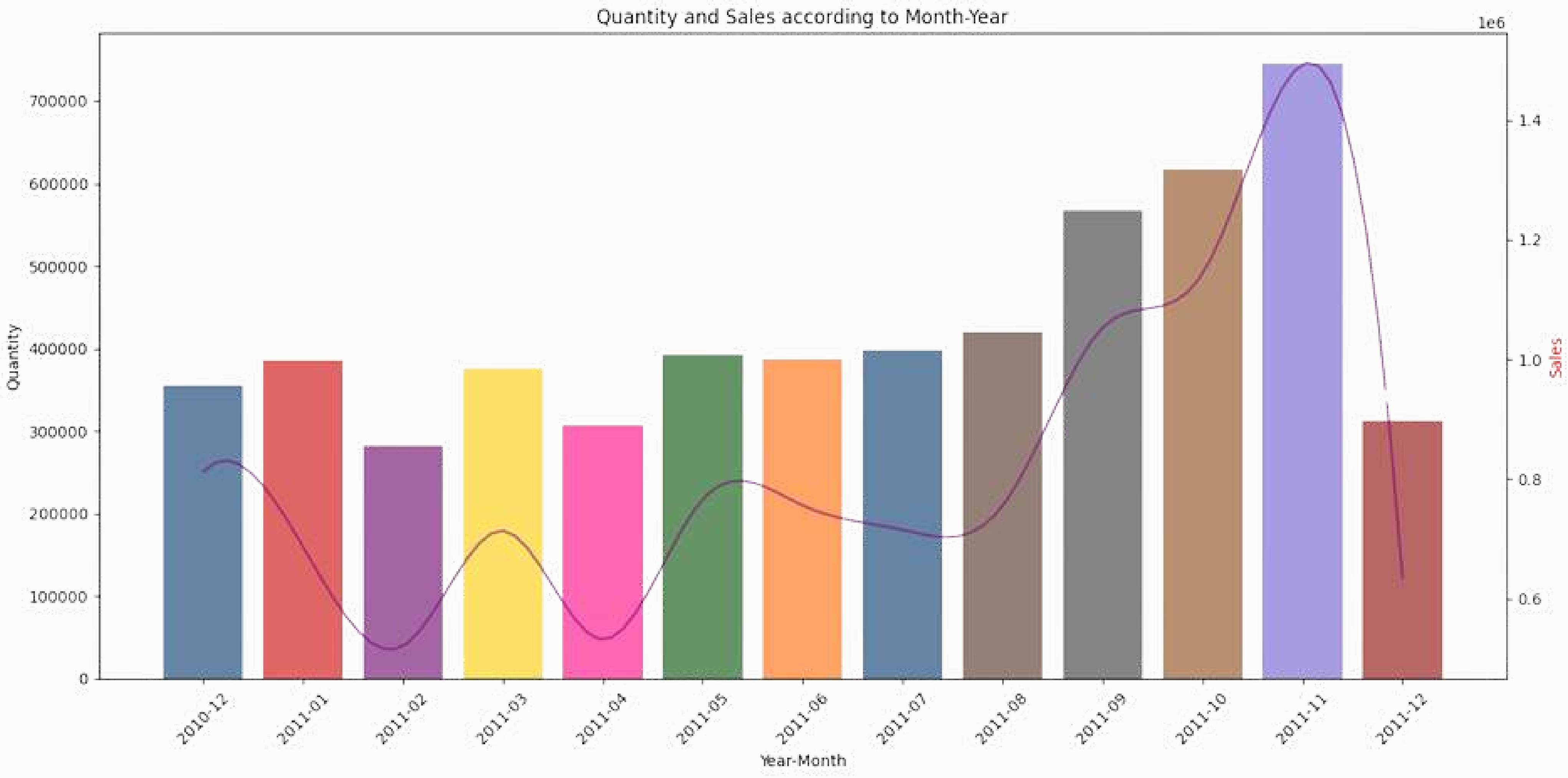
Step	Step Description	Reason	Affected Observations
1	Removed duplicated rows	Duplicated rows can cause the error when running the algorithms	10,682 (or 1,97%) of transactions removed
2	Removed transactions arising from unspecified locations in variable Country	Transactions with unclear locations added nothing to the study of location.	433 (or 0.08%) of transactions removed
3	Replace missing values with "unknown" in the variable CustomerID	These records made up a sizable amount of all transactions, and they could include important information that ought to be recorded rather than deleted.	134,350 (or 24.79%) of transactions recorded
4	In variable Description, removed records with null values	There was no information from null descriptions that was helpful for the analysis.	1,454 (or 0.27%) of transactions removed
5	Removed items sold with zero or negative values in variables UnitPrice and Quantity	These prices added nothing to the profitability of retail sales and could be a sign of data inaccuracies.	10,189 (or 1,88%) of transactions removed
6	Converted InvoiceDate to date format	It ensures that the date format is appropriate for time-series analysis and is consistent.	All transactions included
7	Create a Sales column by using Quantity multiplied by UnitPrice	The new column will provide revenue information for products in each transaction benefit in calculating the sale volume of the product.	All transactions included
8	Aggregated transactions by InvoiceNo and CustomerID to create item baskets	In order to do a market basket analysis, it is necessary to examine sets of products that have been acquired in combination.	All transactions included

2.5 Preprocessing



Relational data model

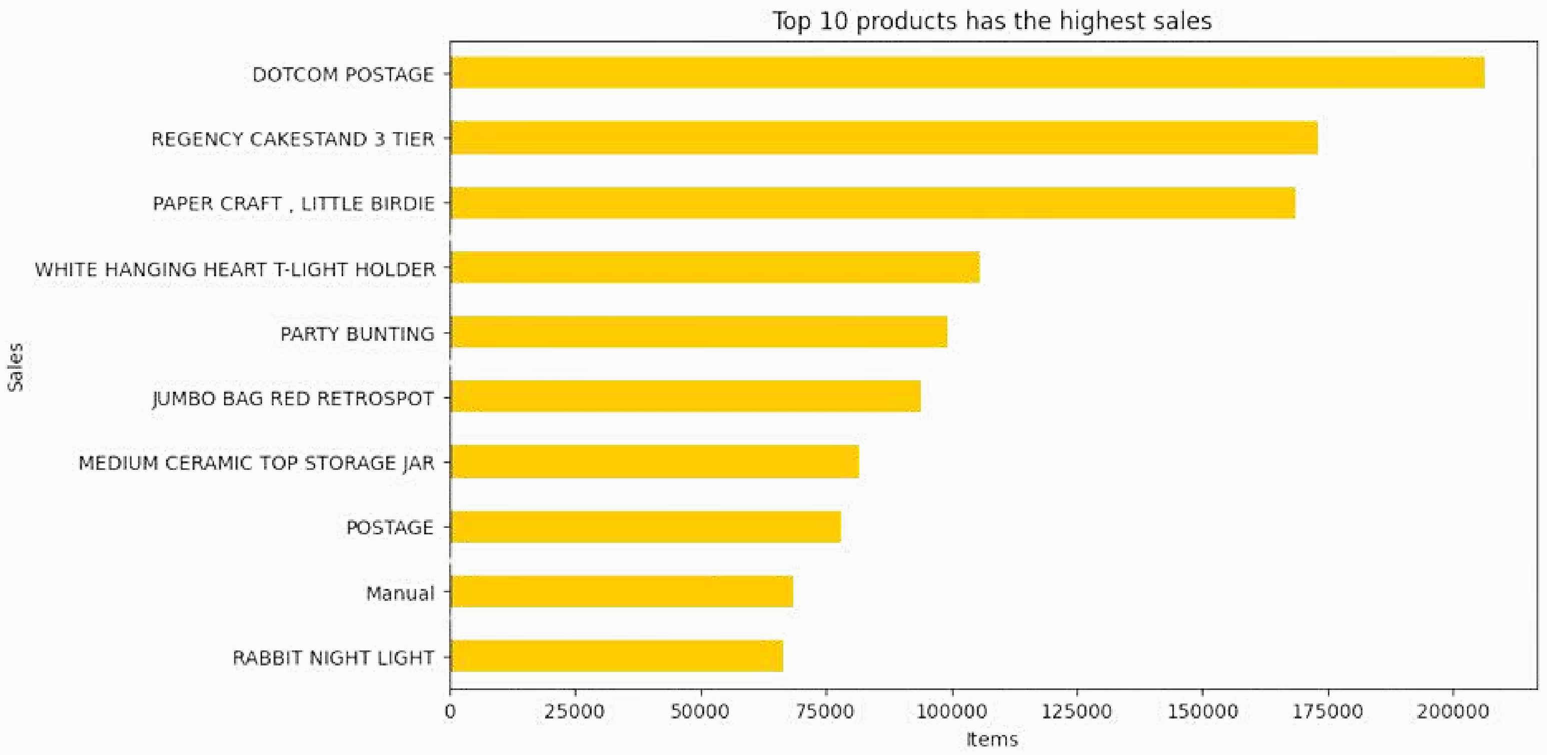
2.5 Preprocessing



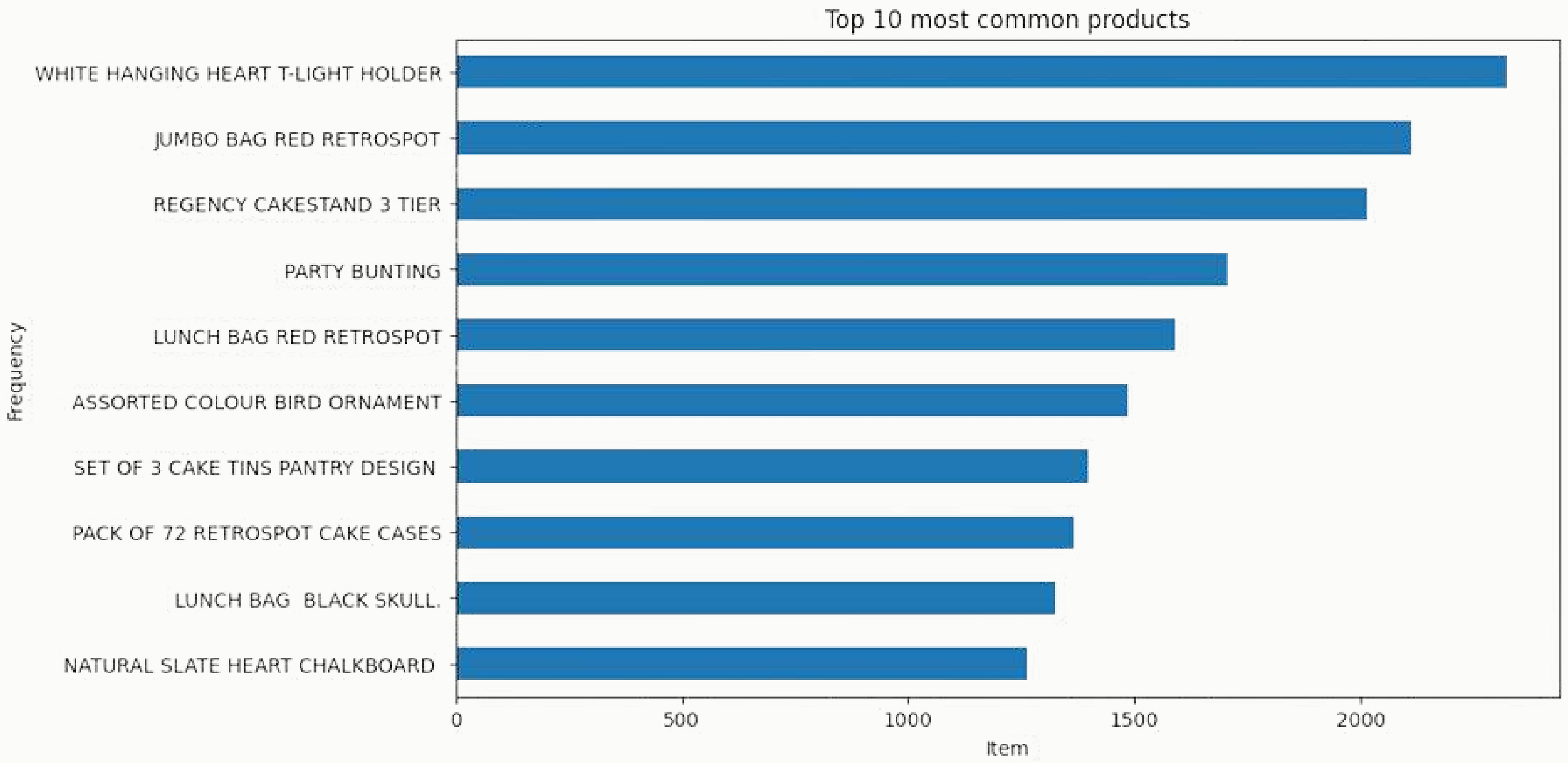
Bar Chart of the Top 10 Products with the Highest Sales

Purpose:
To highlight the sales performance of the top 10 best-selling products and provide insights for business strategy

2.5 Preprocessing



Bar and Line Chart of Quantity and Sales by Month



Top 10 most common products

Purpose:

To analyze monthly sales data to inform business strategy

Purpose:

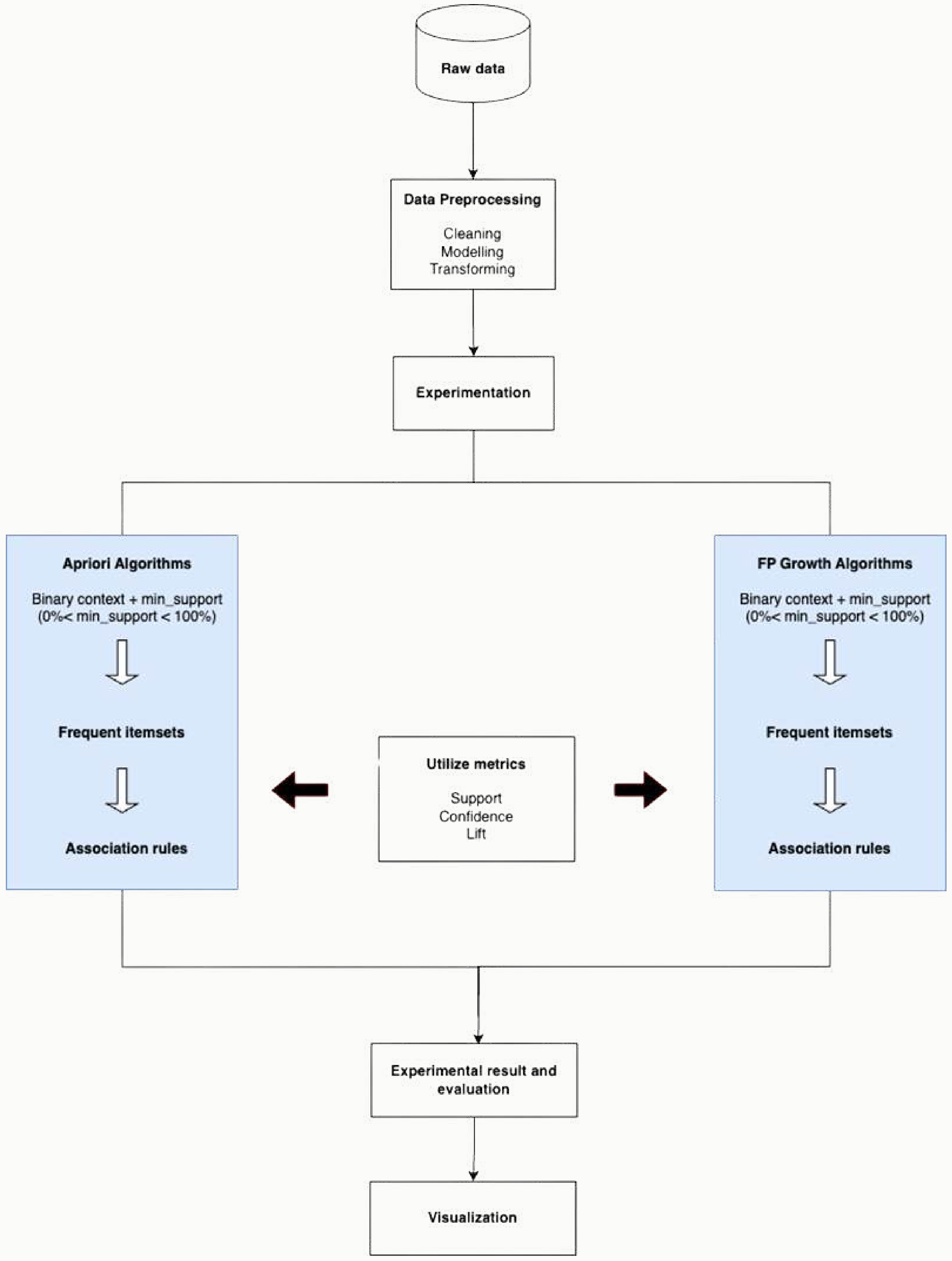
To analyze product sales data and provide strategic recommendations for optimizing business operations

Chap 3

Experimentation

3.1 Model Building

- **Step 1:** Raw data to data preprocessing
- **Step 2**: Experimentation with algorithms
- **Step 3**: Evaluation the experimental results with metrics
- **Step 4**: Visualization



Raw data

Data Preprocessing

Cleaning
Modelling
Transforming

Experimentation

Apriori Algorithms

Binary context + min_support
(0%< min_support < 100%)

Frequent itemsets

Association rules

Utilize metrics

Support
Confidence
Lift

FP Growth Algorithms

Binary context + min_support
(0%< min_support < 100%)

Frequent itemsets

Association rules

Experimental result and evaluation

Visualization

3.2 Experiment

Initial Step

Purpose:

Since both of these algorithms aim to find itemsets with support greater than or equal to the minimum support threshold, choosing a high minimum support will result in identifying only the most reliable itemsets

Service

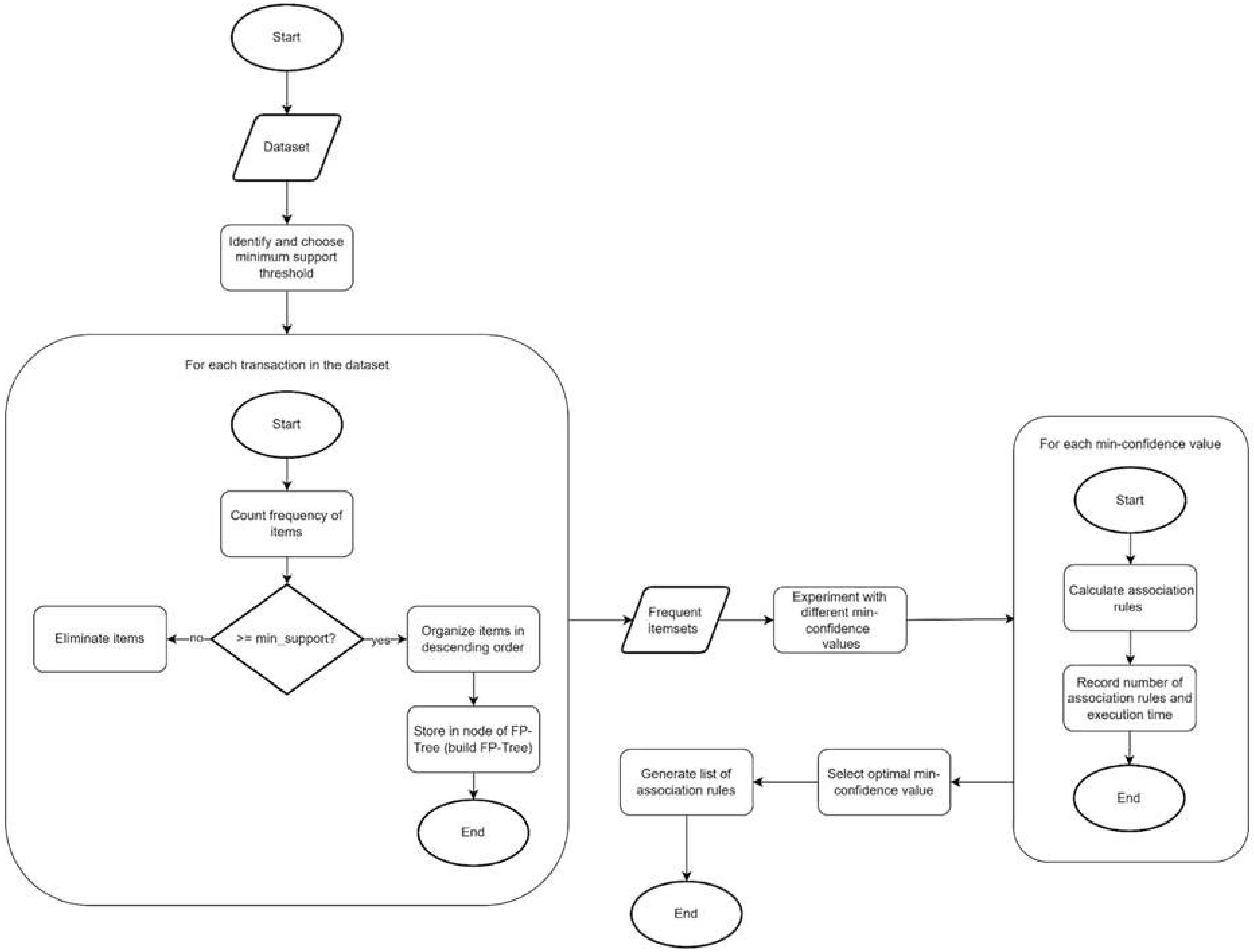
Google Collaboration

Min Support	Number of Frequent Itemsets
0.01	1899
0.02	379
0.03	141
0.04	67
0.05	33
0.06	12
0.07	6
0.08	4
0.09	3
0.1	2

=> A minimum support of 0.01 was chosen, to find both efficient and potentially popular itemsets

3.2 Experiment

FP-GROWTH ALGORITHM



Start

Dataset

Identify and choose minimum support threshold

For each transaction in the dataset

Start

Count frequency of items

>= min_support?

no

Eliminate items

yes

Organize items in descending order

Store in node of FP-Tree (build FP-Tree)

End

Frequent itemsets

Experiment with different min-confidence values

For each min-confidence value

Start

Calculate association rules

Record number of association rules and execution time

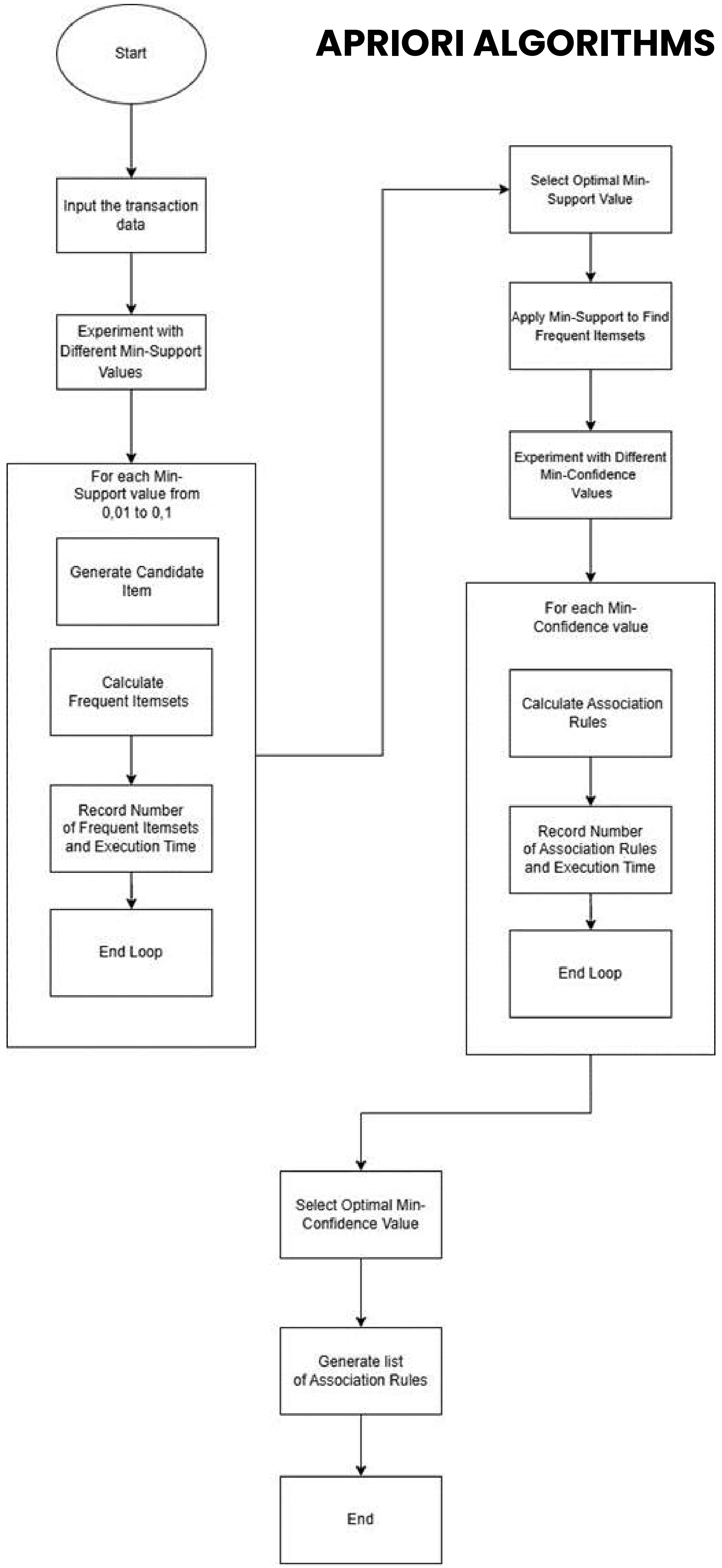
End

Select optimal min-confidence value

Generate list of association rules

End

APRIORI ALGORITHMS



Start

Input the transaction data

Experiment with Different Min-Support Values

For each Min-Support value from 0,01 to 0,1

Generate Candidate Item

Calculate Frequent Itemsets

Record Number of Frequent Itemsets and Execution Time

End Loop

Select Optimal Min-Support Value

Apply Min-Support to Find Frequent Itemsets

Experiment with Different Min-Confidence Values

For each Min-Confidence value

Calculate Association Rules

Record Number of Association Rules and Execution Time

End Loop

Select Optimal Min-Confidence Value

Generate list of Association Rules

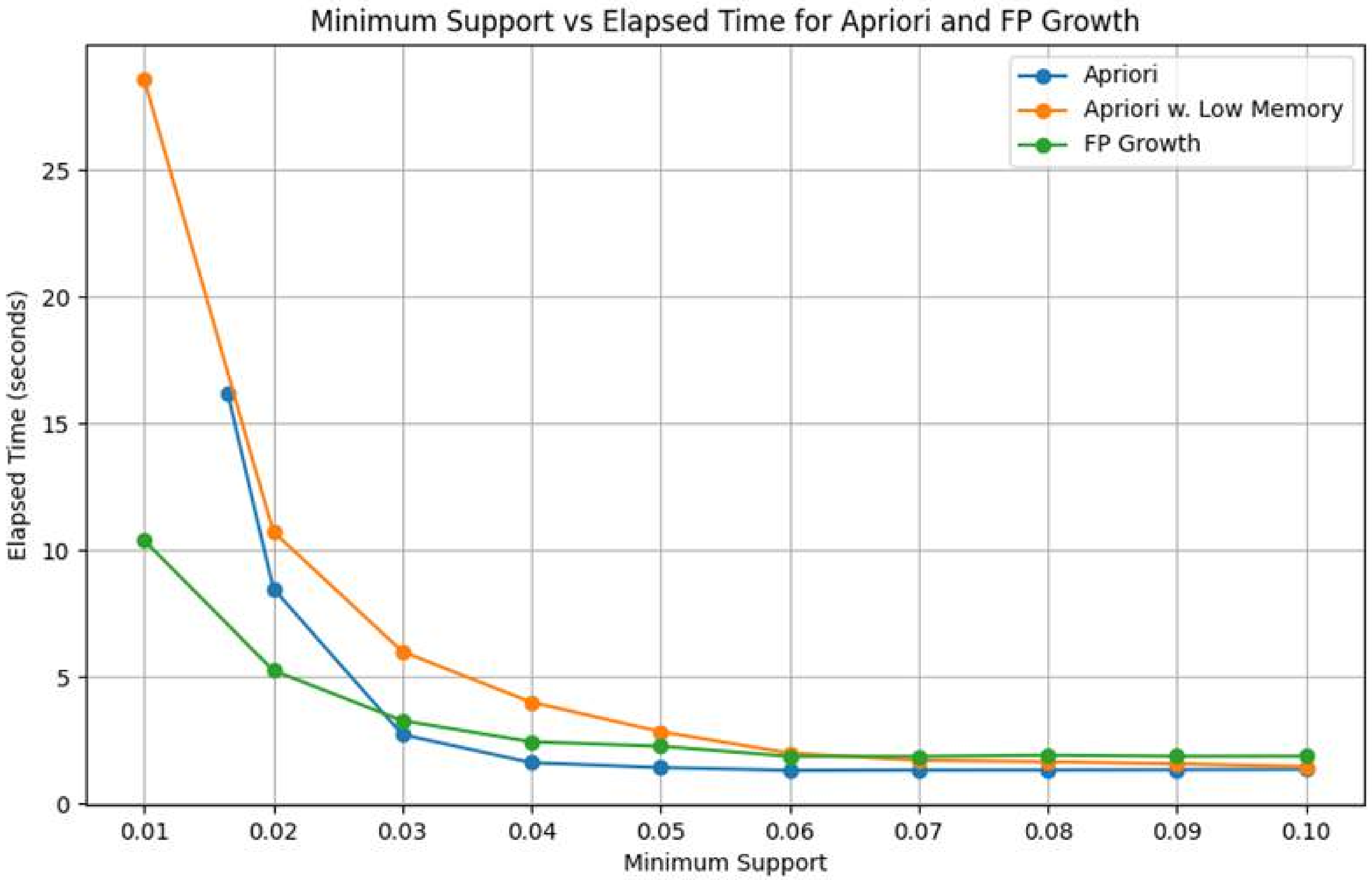
End

3.2 Experiment

FINAL RESULT

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(POPPY'S PLAYHOUSE KITCHEN)	(POPPY'S PLAYHOUSE BEDROOM)	0.021759	0.021107	0.015592	0.716590	33.950360
1	(POPPY'S PLAYHOUSE BEDROOM)	(POPPY'S PLAYHOUSE KITCHEN)	0.021107	0.021759	0.015592	0.738717	33.952060
2	(JAM MAKING SET WITH JARS, SET OF 3 CAKE TINS ...	(JAM MAKING SET PRINTED)	0.019051	0.058257	0.010328	0.542105	9.305363
3	(JAM MAKING SET PRINTED, SET OF 3 CAKE TINS PA...	(JAM MAKING SET WITH JARS)	0.017397	0.056653	0.010328	0.593660	10.478886
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.052642	0.049133	0.032087	0.609524	12.405675
...	...	...	...	...	...	...	...
890	(SET OF 6 SNACK LOAF BAKING CASES)	(SET OF 12 MINI LOAF BAKING CASES)	0.017748	0.022260	0.011782	0.663842	29.822047
891	(SET OF 6 TEA TIME BAKING CASES)	(SET OF 6 SNACK LOAF BAKING CASES)	0.018951	0.017748	0.010378	0.547619	30.855394
892	(SET OF 6 SNACK LOAF BAKING CASES)	(SET OF 6 TEA TIME BAKING CASES)	0.017748	0.018951	0.010378	0.584746	30.855394
893	(JUMBO BAG VINTAGE DOILY)	(JUMBO BAG RED RETROSPOT)	0.035646	0.104733	0.017848	0.500703	4.780769
894	(HAND WARMER RED LOVE HEART)	(HAND WARMER OWL DESIGN)	0.019904	0.032789	0.010930	0.549118	16.7472711

3.2 Experiment



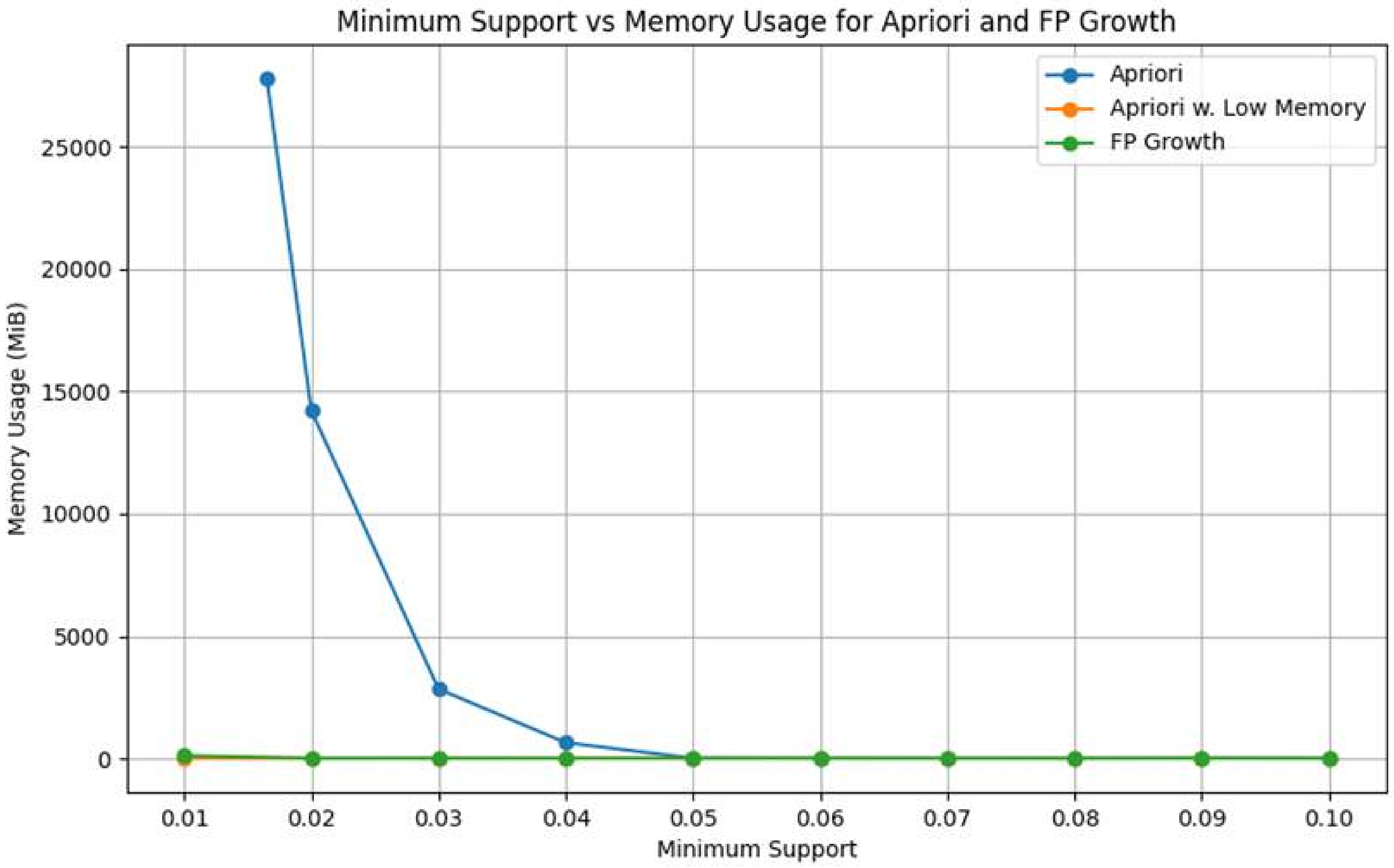
Minimum Support vs Elapsed Time for Apriori and FP-Growth

Purpose:

To illustrate and compare the performance of three algorithms (Apriori, Apriori with Low Memory, and FP-Growth) in terms of elapsed time relative to varying minimum support levels

APRIORI AND FP-GROWTH PERFORMANCE COMPARISON

3.2 Experiment



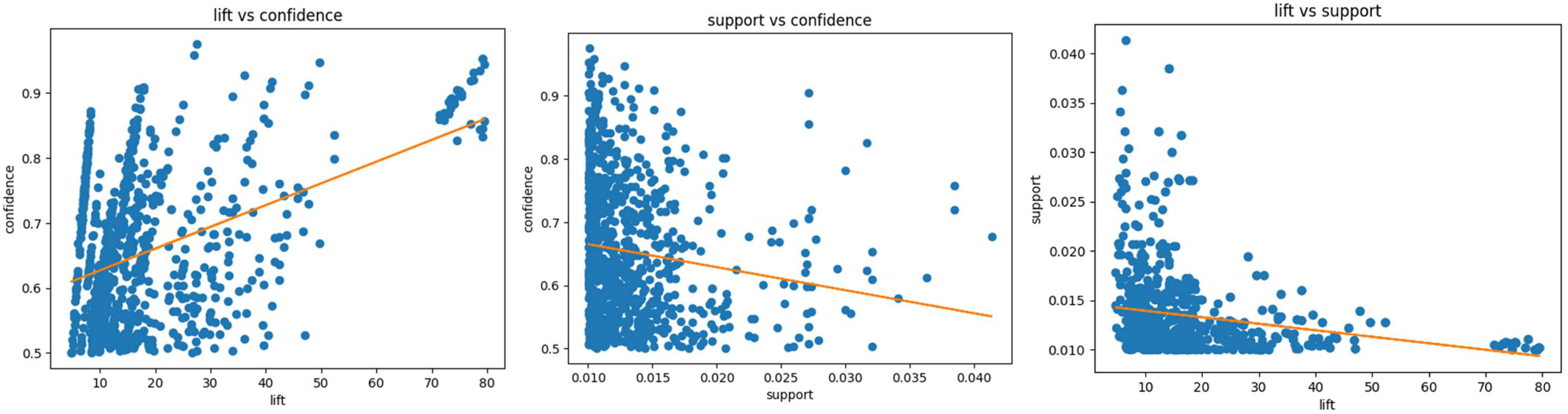
Minimum Support vs Memory Usage for Apriori and FP-Growth

Purpose:

To illustrate and compare the memory usage of three algorithms (Apriori, Apriori with Low Memory, and FP-Growth) as a function of varying minimum support level

APRIORI AND FP-GROWTH PERFORMANCE COMPARISON

3.2 Experiment

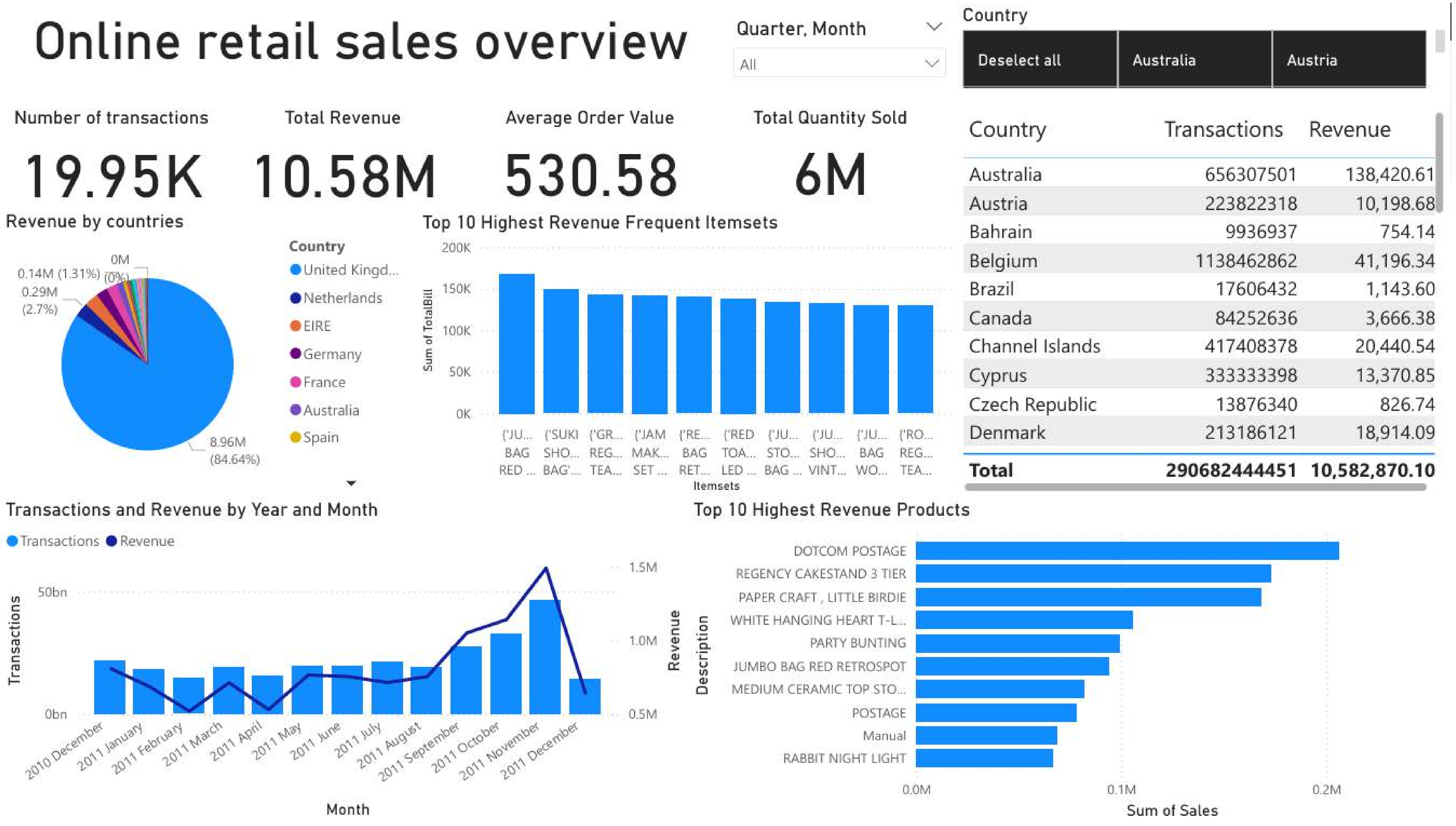


RESULTS BASED ON METRICS COMPARISON

Chap 4

Visualization

Visualization



Online retail sales overview

Quarter, Month: All

Country: Deselect all | Australia | Austria

Number of transactions	Total Revenue	Average Order Value	Total Quantity Sold
19.95K	10.58M	530.58	6M

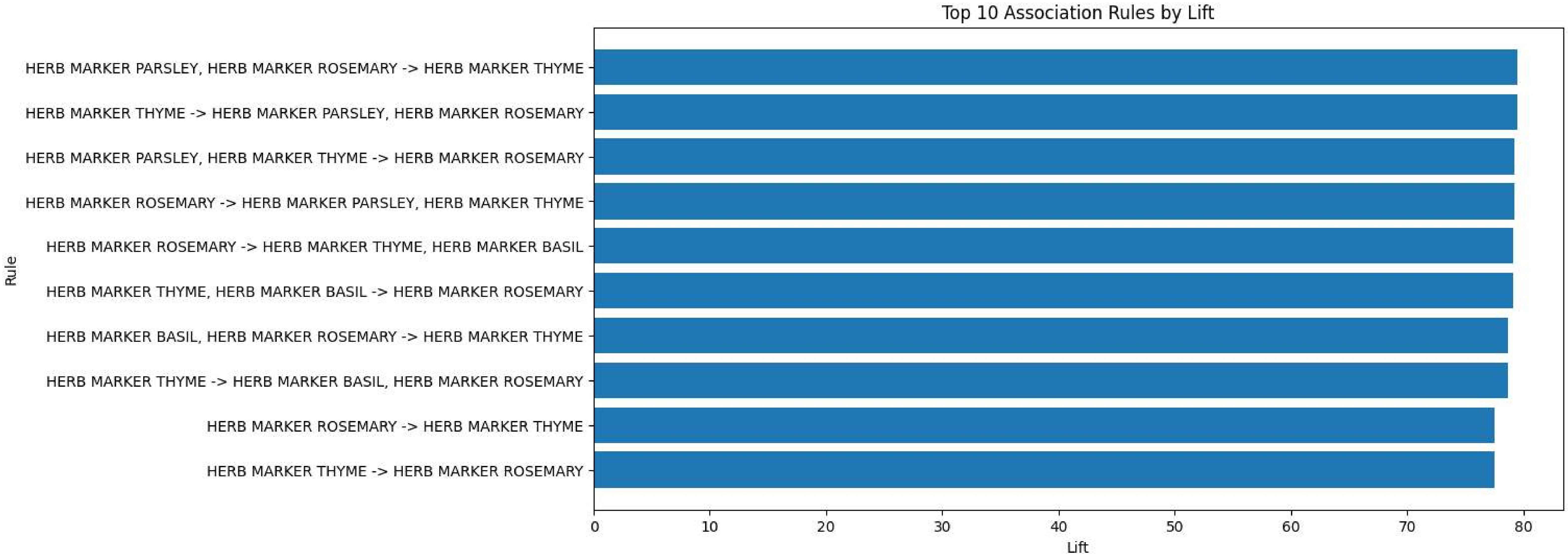
Country	Transactions	Revenue
Australia	656307501	138,420.61
Austria	223822318	10,198.68
Bahrain	9936937	754.14
Belgium	1138462862	41,196.34
Brazil	17606432	1,143.60
Canada	84252636	3,666.38
Channel Islands	417408378	20,440.54
Cyprus	333333398	13,370.85
Czech Republic	13876340	826.74
Denmark	213186121	18,914.09
Total	**290682444451**	**10,582,870.10**

Business overview dashboard

Purpose:

Provide a comprehensive overview of online retail sales, presenting key metrics and visualizations to support data-driven decision-making.

Visualization

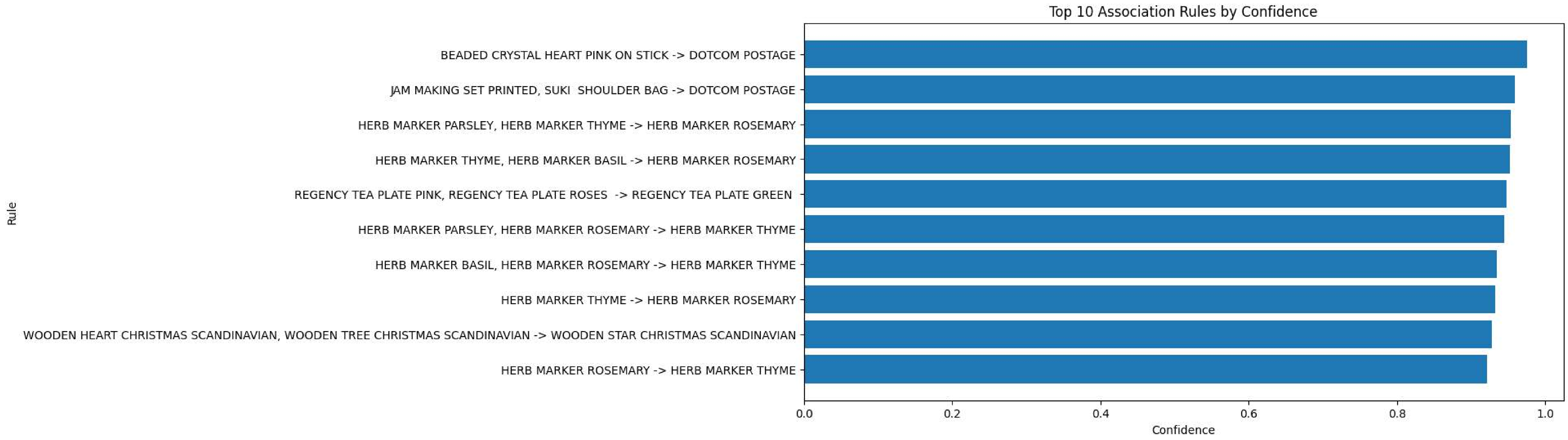


Top 10 Association Rules According to Lift

Purpose:

To analyze product association data and provide recommendations for enhancing sales strategies

Visualization

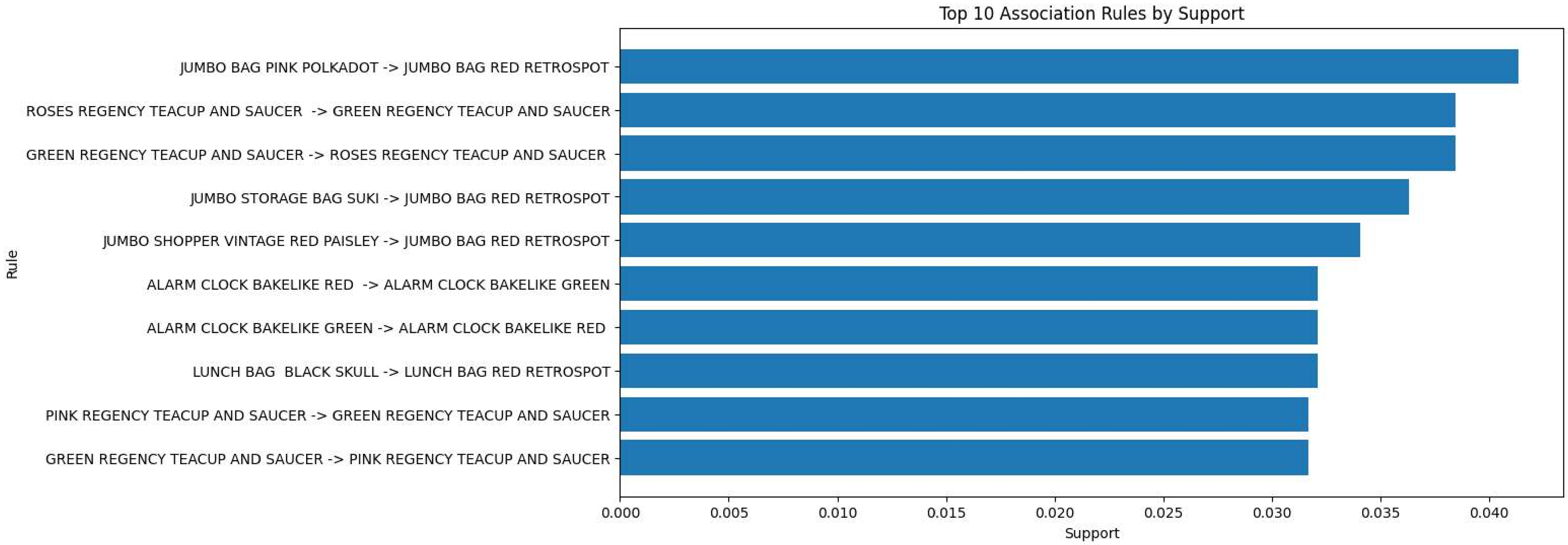


Top 10 Association Rules According to Confidence

Purpose:

To analyze product purchase confidence levels and provide strategic recommendations for optimizing sales

Visualization

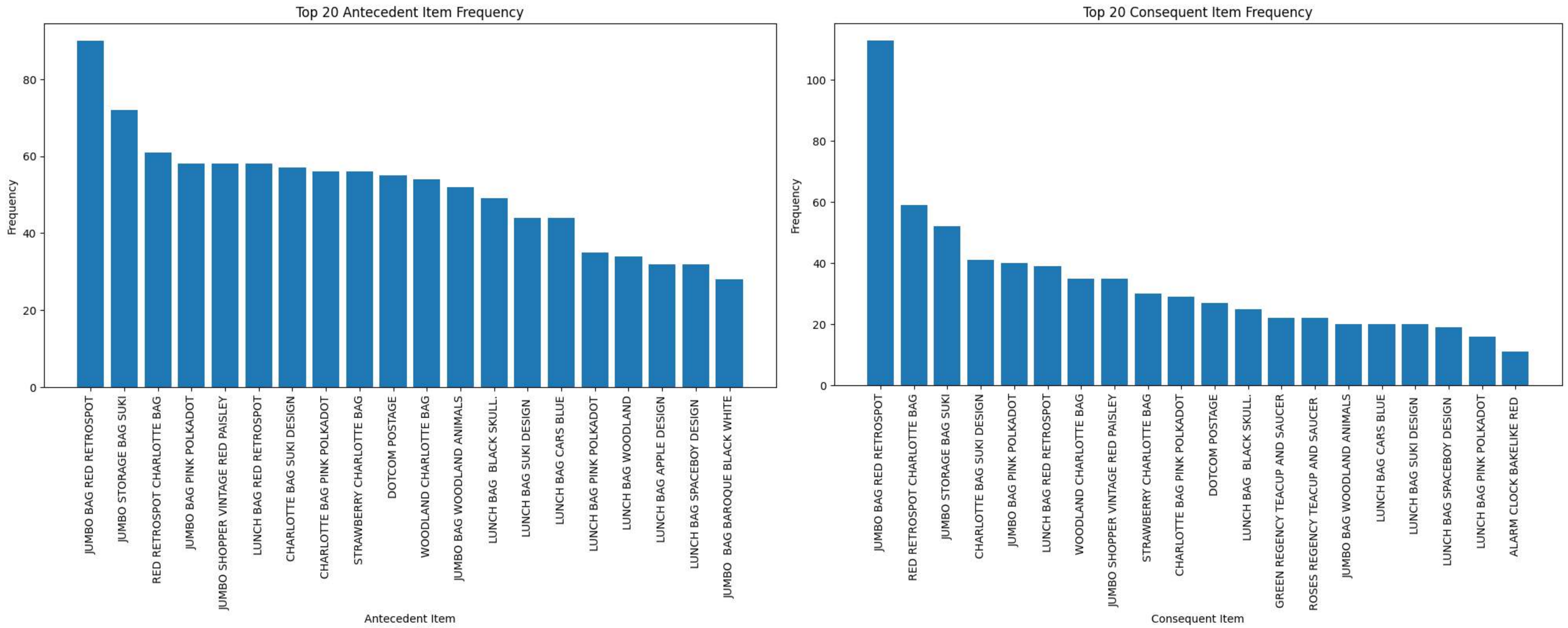


Top 10 Association Rules According to Support

Purpose:

To help businesses optimize their promotion and marketing strategies based on income generated by top association rules

Visualization

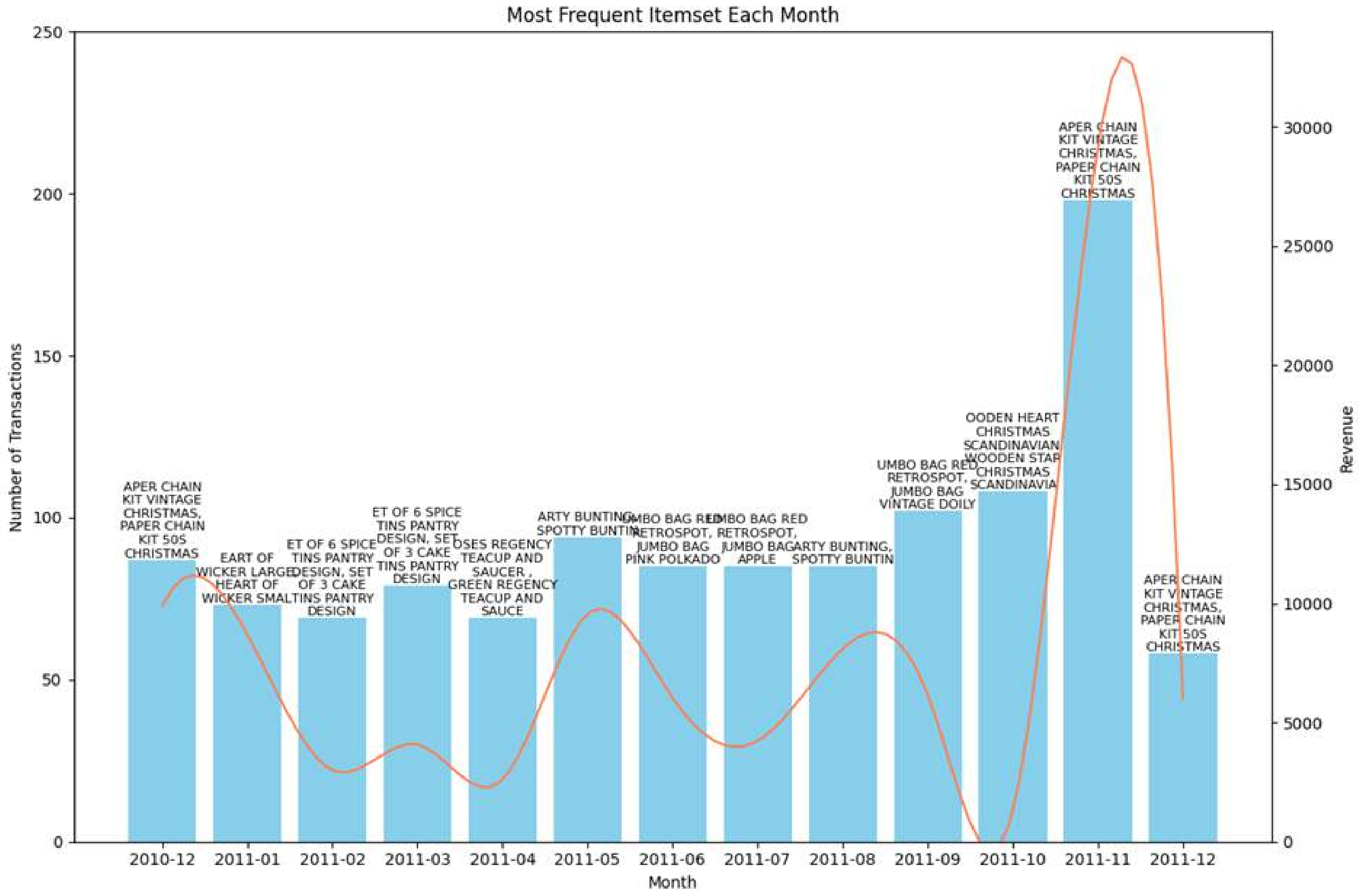


Top 20 antecedent and consequent item frequency, respectively

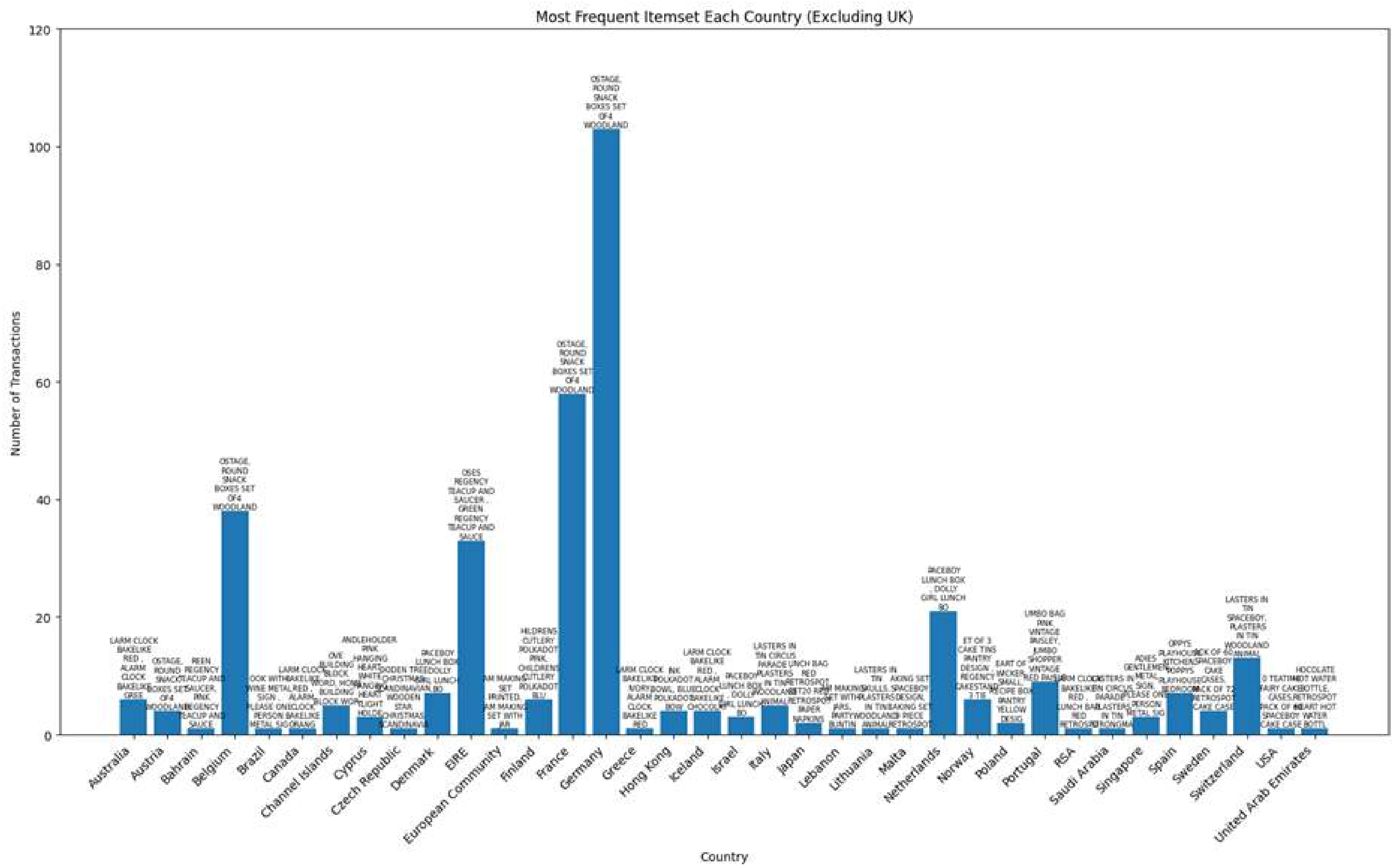
Purpose:

To leverage popular products for improved inventory management and marketing strategies to enhance sales and customer engagement.

Visualization



Most purchased itemsets each month from December 2010 to December 2011



The most popular itemsets in each country

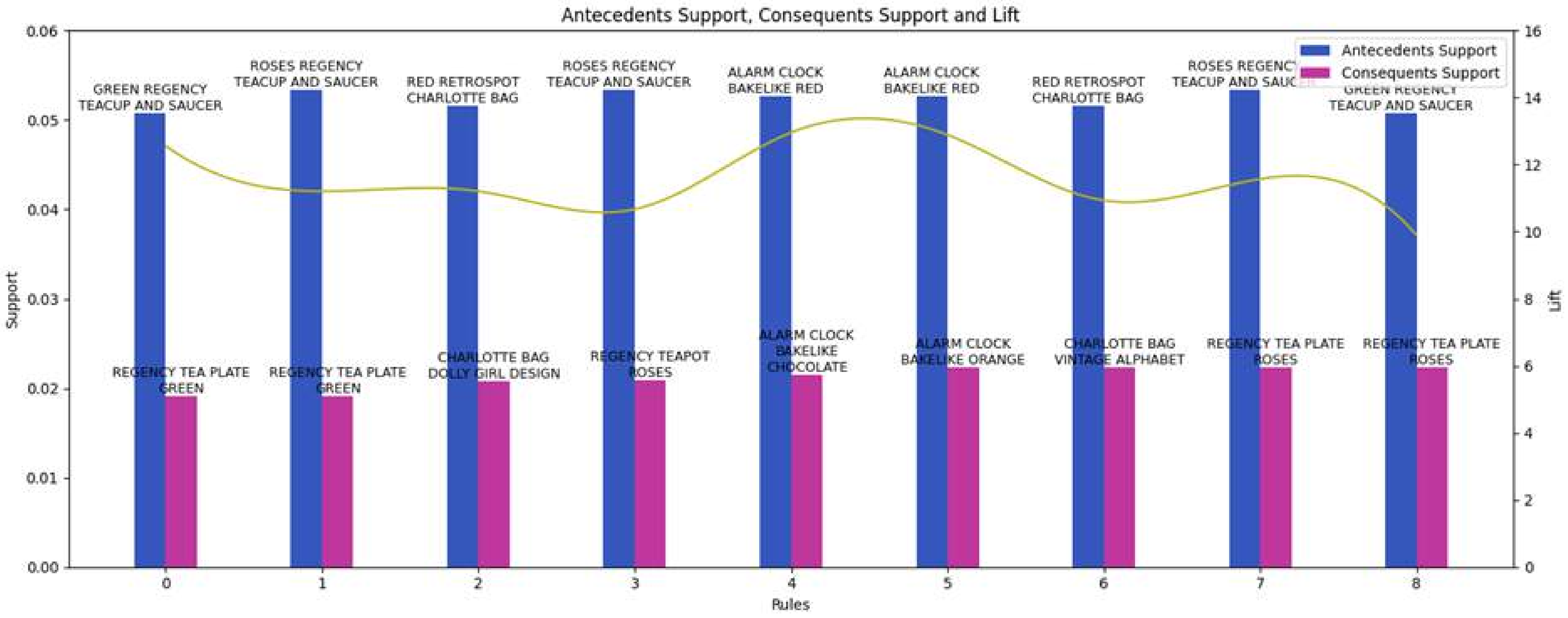
Purpose:

To provide insights into seasonal sales trends and guide strategies for improving inventory management, marketing, and revenue optimization

Purpose:

Use itemset popularity data to optimize business strategies across different countries

Visualization



Antecedents Support, Consequents Support and Lift

Purpose:

To analyze association rules to improve sales strategies for both popular and slow-moving items, which help businesses leverage itemset relationships to enhance sales of slow-moving products through strategic combinations and promotions

Customer Insight And Recommendation

Customer Insight

- Customers tend to buy related products together
- The purchase of this product may be the result of the purchase of the previous product
- Popular items and itemsets can change due to seasonal demand, especially in holiday season
- Other markets such as France, Germany and EIRE also need to be paid attention to increase market share

Recommendation

- Categorize products that are often purchased together into a separate category
- Create bundles that include products that are often purchased together
- Understand the purchasing trends of customers based on their needs, time and location to promote different combos
- Create combos with slow-selling products and products that are often purchased together

Group Evaluation

Lê Châu Anh	K214110856	**100%**
Phạm Thị Anh Thư	K214111979	**100%**
Nguyễn Hoàng Phương Ngân	K214110864	**100%**
Phan Quỳnh Trâm	K214110870	**100%**
Nguyễn Bảo An Nhiên	K214110865	**100%**