PROJECT REPORT

DS5230 Unsupervised Machine Learning

MARKET BASKET ANALYSIS

Anjana Deivasigamani

Naveen Kavitha Gunasekaran

Priyadharshan Sengutuvan

Raghu Ram Baskaran

Introduction

Market Basket Analysis is a data mining technique used by retailers to:

- Increase sales by understanding customer purchasing patterns.
- Identify product groupings and items frequently bought together.

Involves analyzing large datasets, such as purchase history, to:

- Reveal product combinations customers are likely to purchase.
- Inform targeted marketing and strategic product placement.

NOTES

Market Basket Analysis is a data mining technique that enables businesses to analyze large-scale transactional data. Its primary purpose is to identify patterns in purchasing behavior, such as products frequently bought together.

Key Objectives:

1. **Increase Sales:** By identifying purchasing patterns, businesses can create bundles and cross-promotional strategies that appeal to customers.
2. **Optimize Product Placement:** Insights from this analysis help retailers strategically position products, both in-store and online, to make them more accessible and encourage sales.

How It Works:

Market Basket Analysis involves analyzing transaction data to identify **frequent product combinations** and uncover associations. These insights are then used to:

- Design **targeted marketing campaigns.**
- Enhance **strategic product placement.**

For example, if data reveals that customers often purchase coffee and mugs together, businesses can position these items closer in stores or offer a bundled discount to encourage purchases.

This technique is particularly valuable for improving the overall shopping experience and maximizing revenue.

Problem statement

Retail Challenges

- **Customer Segmentation**: Difficulties in identifying customer groups for personalized marketing.
- **Inventory Optimization**: Challenges in predicting demand and managing stock effectively.
- **Temporal Trends**: Difficulty recognizing seasonal patterns for better forecasting.
- **Product Performance**: Identifying top revenue-generating products and growth opportunities.

NOTES

The retail industry faces several challenges that can be effectively addressed through Market Basket Analysis. Let's take a closer look at these challenges:

Customer Segmentation:
Retailers often find it difficult to group customers into meaningful segments based on their purchasing behavior. Without segmentation, marketing efforts tend to be broad and less effective.

Inventory Optimization:
Managing stock levels is a common issue. Overstocking wastes resources, while understocking can lead to missed sales opportunities and dissatisfied customers.

Temporal Trends:
Recognizing seasonal patterns is critical for sales forecasting and promotional planning. For instance, demand for winter clothing peaks during specific months, and failing to anticipate this can lead to lost revenue.

Product Performance:
Identifying top-performing products and underperforming ones is crucial for refining marketing strategies and optimizing inventory.

Market Basket Analysis addresses these challenges by leveraging transactional data to extract actionable insights, helping retailers make data-driven decisions that improve operational efficiency and customer satisfaction.

Approach

Principal Component Analysis (PCA):

- Reduces dimensionality to highlight key features driving customer behavior.
- Identifies major factors influencing purchasing patterns.

Association Rule Mining (Apriori and FP-Growth):

- Discovers frequent itemsets and co-purchasing patterns.
- Generates actionable rules to inform product recommendations and cross-selling strategies.

Customer Segmentation with K-Means Clustering:

- Groups customers based on purchasing behavior to personalize marketing and improve customer targeting.

Temporal Trend Analysis:

- Analyzes sales data to detect seasonal patterns and predict future sales.

NOTES

To tackle the challenges outlined in the problem statement, we used a combination of analytical techniques. Let me walk you through these methods:

Principal Component Analysis (PCA):

- PCA is used to reduce the dimensionality of large datasets while retaining key variables.
- For instance, it helps identify the most important factors driving customer purchasing decisions, such as price sensitivity or product popularity.

Association Rule Mining (Apriori & FP-Growth):

- These algorithms are used to identify frequent itemsets and extract actionable rules.
- For example, they reveal product combinations like “milk and bread” that are often bought together, enabling cross-selling opportunities.

Customer Segmentation (K-Means Clustering):

- This technique groups customers based on their purchasing behavior, such as frequency or spending habits.
- For example, it could identify a group of budget-conscious shoppers who might respond well to discounts or coupons.

Temporal Trend Analysis:

- This analyzes sales data over time to uncover seasonal or temporal patterns.
- For example, it might highlight that ice cream sales spike during summer months, helping retailers prepare inventory accordingly.

Together, these techniques provide a comprehensive understanding of customer behavior and product performance.

Data Overview

- Size: 522,064 entries and 7 columns.
- **Key Columns:**
  - **BillNo:** Unique identifier for each transaction.
  - **Itemname:** Name of the purchased product.
  - **Quantity:** Number of units purchased.
  - **Date:** Timestamp of the transaction.
  - **Price:** Unit price of the product.
  - **CustomerID:** Unique identifier for each customer.
  - **Country:** Country where the transaction occurred.
- **Key Observations**
  - Missing values in Itemname and CustomerID columns.
  - Transaction dates range across multiple years.

NOTES

Our analysis is based on a dataset containing **522,064 transactions** and the following key columns:

1. **BillNo:** A unique identifier for each transaction.
2. **Itemname:** The name of the product purchased.
3. **Quantity:** The number of units purchased.
4. **Date:** The timestamp of the transaction.
5. **Price:** The unit price of the product.
6. **CustomerID:** A unique identifier for each customer.
7. **Country:** The location where the transaction occurred.

Key Observations:

- Missing values were identified in the **Itemname** and **CustomerID** columns, which required handling during the data preparation phase.
- Transaction dates span multiple years, providing opportunities for analyzing temporal trends.

This dataset serves as a strong foundation for applying Market Basket Analysis techniques and extracting meaningful insights.

Data Overview

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2,55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3,39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2,75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3,39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3,39	17850.0	United Kingdom

Fig.1 Dataset

NOTES

Structure of Transactions:

Each transaction is detailed across multiple rows, where a single **BillNo** corresponds to multiple items purchased in that transaction. This structure is essential for identifying associations between products.

Temporal Scope:

By using the **Date** column, we can analyze purchasing behaviors over time, such as daily, monthly, or seasonal trends.

Customer Insights:

The **CustomerID** allows us to link transactions to specific individuals, making it possible to study customer purchasing habits and segment them effectively.

Analysis Readiness:

This dataset is well-structured for Market Basket Analysis, enabling us to extract frequent itemsets, study associations, and inform strategies for cross-selling and product placement.

By processing this data further, we can uncover valuable patterns and trends that will directly contribute to actionable business insights.

Data Preparation

1) **Data Loading & Initial Inspection:**
   - Loaded the dataset using pd.read_csv() with appropriate delimiters and data types.
   - Conducted a quick inspection using .info(), .head(), .isnull().sum(), and .nunique() to identify data structure, missing values, and unique counts.

2) **Datetime and Numeric Transformations:**
   - Converted the Date column to datetime format using pd.to_datetime().
   - Adjusted the Price column by replacing commas with periods, ensuring correct float conversion.

3) **Handling Missing Values:**
   Dropped rows with null values using purchase_df.dropna(), resulting in 388,023 entries with no missing data.

NOTES

1) <u>Data Loading & Initial Inspection</u>

We began by loading the dataset using Python's **pandas library**, which is well-suited for handling large datasets efficiently.

After loading the data, we conducted an initial inspection using functions such as .info(), .head(), and .isnull() to:

- Understand the data structure and types of each column.
- Identify missing values, anomalies, or potential inconsistencies.
- Examine the number of unique entries in key columns like Itemname, BillNo, and CustomerID.

2) <u>Datetime and Numeric Transformations</u>

Date Conversion:

The **Date** column was converted to a standard datetime format using pd.to_datetime(). This transformation enables precise time-series analysis, allowing us to identify temporal patterns, such as monthly or seasonal trends in purchasing.

Price Adjustment:

The **Price** column, originally formatted with commas as decimal separators, was converted to a numerical float format. This ensures accurate calculations for metrics like total revenue or average item cost.

3) <u>Handling Missing Values</u>

Identification of Missing Data:

Missing values were primarily found in the Itemname and CustomerID columns. These columns are critical for understanding purchasing behavior and linking transactions to individual customers.

Action Taken:

Rows with missing values in these columns were dropped using purchase_df.dropna().
After this step, the dataset size reduced from **522,064 rows** to **388,023 rows**—a necessary trade-off to maintain data quality.

4) **Filtering Out Invalid Data:**

Removed rows with non-positive values in Quantity and Price to ensure logical integrity of transactions.

5) **Type Adjustments:**

Converted the CustomerID column to string type for consistent handling across the dataset.

6) **Descriptive Statistics and Validation:**

Examined the Quantity and Price columns using .describe() to understand central tendencies, variability, and outliers.

7) **Final Dataset:**

Cleaned and preprocessed dataset with 387,985 rows, ready for analysis and modeling.

NOTES

4) Filtering Invalid Data

Issue Identification:
Rows with non-positive values in the Quantity and Price columns (e.g., zero or negative entries) were identified as illogical for retail transactions.
Action Taken:
These rows were removed to ensure only valid transactions remained in the dataset.

5) Type Adjustments

The CustomerID column was converted to a string format.
Reason: This avoids numeric misinterpretations and facilitates grouping customers based on their transactions.

6) Descriptive Statistics and Validation

To Ensure consistency and accuracy in the dataset.
Steps Taken:
Analysis of Quantity **and** Price**:**
Checked for anomalies and ensured all values were logical (e.g., positive).
Validation:
Verified no missing or invalid data remained after cleaning.

7) Final Dataset

Dataset Size:
Reduced from **522,064 rows** to **387,985 rows** after removing missing and invalid entries.

Key Characteristics:
The dataset is now **complete, consistent, and valid**.
Ready for advanced analysis techniques, including:
Association Rule Mining
Customer Segmentation
Temporal Trend Analysis

Association Rule Mining

- **What is ARM?**
  - A data mining technique used to uncover **meaningful patterns** and **relationships** between items in large datasets.
  - Helps identify **frequently co-occurring items** and uncover **actionable insights**.
- **Why Use ARM?**
  - Drives **business decisions** by revealing:
  - **Product placement strategies**: Optimizing how products are displayed to maximize cross-sales.
  - **Targeted marketing campaigns**: Promoting products likely to be purchased together.
- **Purpose in Our Analysis**
  - To analyze transactional data and extract association rules that highlight:
    - **Strong relationships** between items.
    - Patterns that enhance customer experience and **boost sales.**

NOTES

In this part, we applied **Association Rule Mining (ARM)** to uncover patterns and relationships between products. ARM is a cornerstone of Market Basket Analysis.

What is ARM?

ARM is a data mining technique that identifies meaningful patterns in large transactional datasets.

It helps reveal **associations between items**, such as products frequently purchased together.

Key Metrics in ARM:

1. **Support:**
   - The proportion of transactions that include a particular itemset.
   - For example, if "bread and butter" appears in 5% of transactions, its support is 0.05.

2. **Confidence:**
   - Measures the likelihood of purchasing an item (the consequent) when another item (the antecedent) is purchased.
   - Example: If 80% of customers who buy bread also buy butter, the confidence is 0.80.

3. **Lift:**
   - Indicates the strength of association between two items.
   - A lift greater than 1 suggests a positive correlation, meaning items are more likely to be bought together than by chance.

Purpose of ARM:

1. **Improve Sales Strategies:**
   Use rules to design promotions, such as product bundles or cross-selling offers.

2. **Optimize Product Placement:**
   Position frequently bought items near each other to encourage additional purchases.

Association Rule Mining

Association Rules:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(ALARM CLOCK BAKELIKE CHOCOLATE)	(ALARM CLOCK BAKELIKE GREEN)	0.017154	0.042940	0.011381	0.663462	15.450753	0.010645	2.843834	0.951602
1	(ALARM CLOCK BAKELIKE CHOCOLATE)	(ALARM CLOCK BAKELIKE RED)	0.017154	0.047009	0.011986	0.698718	14.863488	0.011180	3.163119	0.949000
2	(ALARM CLOCK BAKELIKE ORANGE)	(ALARM CLOCK BAKELIKE GREEN)	0.019298	0.042940	0.011931	0.618234	14.397481	0.011102	2.506925	0.948855
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.047009	0.042940	0.028810	0.612865	14.272468	0.026792	2.472163	0.975807
4	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.042940	0.047009	0.028810	0.670935	14.272468	0.026792	2.896054	0.971658

Fig.2 Association Rule

NOTES

This slide presents examples of association rules derived from the dataset, showcasing how ARM metrics are applied to real-world data.

Understanding the Table:

- **Antecedents:** The item(s) purchased first (e.g., "Alarm Clock Bakelike Chocolate").
- **Consequents:** The item(s) likely to be purchased along with the antecedents (e.g., "Alarm Clock Bakelike Green").
- **Support:** The proportion of transactions containing both the antecedents and consequents.
- **Confidence:** The probability of buying the consequent item given that the antecedent is purchased.
- **Lift:** The strength of the relationship between the antecedent and consequent.

Example Rule:

- **Antecedent:** "Alarm Clock Bakelike Chocolate"
- **Consequent:** "Alarm Clock Bakelike Green"
- **Support:** 1.1% (these items appear together in 1.1% of transactions).
- **Confidence:** 66.3% (66.3% of customers who buy the antecedent also buy the consequent).
- **Lift:** 15.45 (a strong positive correlation between the two items).

Why These Insights Matter:

Product Bundling:
Retailers can create bundles for items with high support and confidence, increasing sales of both products.

Targeted Recommendations:
Online stores can use these rules to suggest related products during checkout or while browsing.

ARM metrics provide quantitative insights that help retailers make data-driven decisions to enhance customer experience and maximize revenue.

Support vs Confidence

- Support
  - Measures the frequency of an itemset in the dataset.
  - Formula:
    Support(A) = Number of transactions containing A / Total number of transactions
  - Indicates the popularity of an itemset.
- Confidence
  - Measures the likelihood of an item being purchased given another item is already purchased.
  - Formula:
    Confidence($A \rightarrow B$) = Support($A \cup B$) / Support(A)
  - Represents the strength of an association rule.
- Why?
  - Support ensures rules are derived from a **substantial portion** of transactions.
  - Confidence evaluates the **predictive power** of the rule.
  - Together, they filter out irrelevant or weak rules for actionable insights.

NOTES

This slide explains the two critical metrics in association rule mining: **Support** and **Confidence**. These metrics help evaluate the strength and relevance of association rules in transactional data.

Support
Measures how frequently an item or itemset appears in the dataset.

Formula:
Support of A = Number of transactions containing A / Total number of transactions

Importance: Indicates the popularity of an item or combination, ensuring that the rules are derived from a substantial portion of the data.

Confidence
Measures the likelihood of purchasing one item (the consequent) given that another item (the antecedent) has been purchased.

Formula:
Confidence of A to B = Support of A and B / Support of A

Importance: Represents the predictive power of an association rule, helping identify strong relationships between items.

Why Are These Metrics Important?
Support: Ensures that the rules are relevant and derived from a meaningful sample of transactions.

Confidence: Helps assess the likelihood and reliability of the observed associations.

Together: These metrics help filter out weak or irrelevant rules, focusing on actionable insights for decision-making.

Highest lowest

	antecedents	consequents	ıtecedent suppo	nsequent suppo	support	confidence	lift
51	(REGENCY MIL	(REGENCY SU(	0.01347	0.01325	0.010117	0.75102	56.678669
50	(REGENCY SU(	(REGENCY MIL	0.01325	0.01347	0.010117	0.763485	56.678669
53	(REGENCY TEA	(REGENCY TEA	0.013525	0.016549	0.011381	0.841463	50.845636
52	(REGENCY TEA	(REGENCY TEA	0.016549	0.013525	0.011381	0.687708	50.845636
37	(SHED)	(KEY FOB)	0.014515	0.020508	0.014515	1	48.761394
...	...	...	...	...	...	...	...
35	(JUMBO BAG S	(JUMBO BAG R	0.035463	0.086486	0.022377	0.631008	7.296102
34	(JUMBO BAG P	(JUMBO BAG R	0.047504	0.086486	0.029635	0.623843	7.213254
16	(CANDLEHOLD	(WHITE HANGI	0.018034	0.105509	0.01325	0.734756	6.96391
124	(GREEN REGE	(REGENCY CA	0.019958	0.089455	0.011986	0.600551	6.713473
48	(RED HANGING	(WHITE HANGI	0.036233	0.105509	0.024137	0.666161	6.313775

Strongest Association

- The REGENCY MILK JUG PINK and REGENCY SUGAR BOWL GREEN pair shows the highest lift value of 56.68, indicating an extremely strong relationship between these items, with confidence levels above 75% in both directions.

Weakest Association

- The RED HANGING HEART and WHITE HANGING HEART T-LIGHT HOLDER combination demonstrates the lowest lift value of 6.31, despite having relatively high support (0.024137), suggesting a weaker but more common relationship between these decorative items.

NOTES

This slide presents examples of both the strongest and weakest associations derived from the dataset based on metrics like **support**, **confidence**, and **lift**.

Strongest Association:
Example Pair: Regency Milk Jug Pink and Regency Sugar Bowl Green.

Lift Value: 56.68, indicating an extremely strong relationship.

Confidence: Over 75%, suggesting a high likelihood of purchasing these items together.

Interpretation: This strong association highlights a consistent co-purchase pattern, which can be used for bundling or targeted marketing.

Weakest Association:
Example Pair: Red Hanging Heart and White Hanging Heart T-Light Holder.

Lift Value: 6.31, a much weaker correlation.

Interpretation: While the items may be commonly bought individually, their connection is not strong enough to justify bundling or promotion.

Key Insight:
Strong associations can drive promotional strategies, while weaker associations may require re-evaluation or further analysis to uncover potential seasonal or context-driven trends.

Support vs Confidence

- Purpose
  - To analyze the relationships between **support**, **confidence**, and **lift** for the association rules.
  - Helps identify strong rules for effective decision-making.
- Features
  - **X-axis (Support)**: Represents the frequency of item combinations in the dataset.
  - **Y-axis (Confidence)**: Indicates the likelihood of one item being purchased given another is purchased.
  - **Marker Size**: Scaled by **lift**, showing the strength of the association.
  - Larger size = higher lift, indicating a stronger relationship.
  - **Marker Color**: Color-coded by **lift** value using a visually appealing, colorblind-friendly scale (Cividis).

NOTES

This scatterplot visualizes the relationship between **support**, **confidence**, and **lift** for the generated association rules.

Purpose of the Visualization:
To analyze the **distribution** of rules based on their support and confidence values.

To identify strong rules (high lift) that can influence decision-making.

Features of the Plot:

X-axis: Represents support (frequency of item combinations).

Y-axis: Represents confidence (likelihood of consequent given the antecedent).

Marker Size: Scaled by lift (strength of the association). Larger markers indicate stronger relationships.

Marker Color: Reflects lift value using a colorblind-friendly scale for better visualization.

Observations:
Dense Cluster: Most rules are concentrated between **0.01-0.015 support** with confidence above 0.7.

High Confidence Zone: Rules with confidence values > 0.8 indicate highly reliable associations.

Trade-offs: Rules with lower support tend to have higher lift but may lack generalizability.

Support vs Confidence

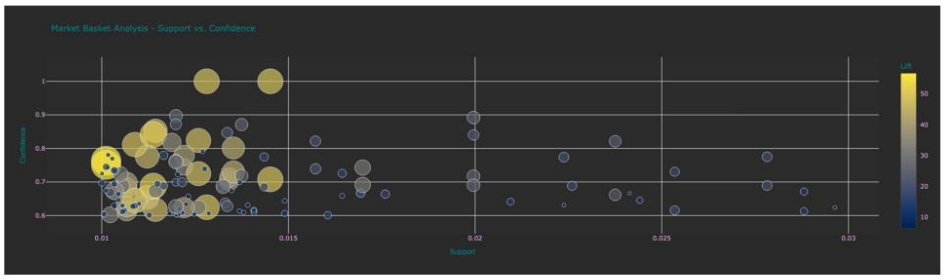

Fig.3 Support vs Confidence

Distribution

- Dense Cluster: Majority of rules concentrated between 0.01-0.015 support
- High Confidence Zone: Strong rules appearing in 0.7-1.0 range
- Sparse Region: Fewer associations beyond 0.015 support

Pattern

- Strong Associations: Large yellow bubbles indicate high-lift rules
- Sweet Spot: Best rules found in high confidence (>0.8), low support (~0.01) region
- Trade-off Pattern: Inverse relationship between support and lift values

NOTES

This scatterplot illustrates the relationship between **support**, **confidence**, and **lift** for association rules, helping identify strong and actionable rules.

Key Insights

1. **Distribution:**
   - Most rules are concentrated between **0.01 to 0.015 support**.
   - High-confidence rules appear in the **0.7 to 1.0 range**.
2. **Patterns:**
   - **Strong Associations:** Large yellow bubbles indicate high-lift rules.
   - **Sweet Spot:** Best rules are found with **high confidence (>0.8)** and **low support (~0.01)**.
   - **Trade-off:** Lower support often correlates with higher lift.

The visualization highlights high-lift, high-confidence rules, ideal for strategies like product bundling and cross-selling. It also reveals common itemsets for frequent patterns.

Interactive Network Graph

- Purpose
  - To illustrate the relationships between items (antecedents and consequents) in a network structure.
  - Enables intuitive exploration of association rules.
- Graph features
  - Nodes
    - Represent items in the dataset.
    - Colored turquoise blue for clear visibility.
    - Hovering over a node shows:
    - **Antecedent** and **Consequent** details.
    - Metrics: **Support**, **Confidence**, and **Lift**.
  - Edges
    - Represent association rules linking antecedents to consequents.
    - Thickness scaled by **support**, indicating rule strength.
    - Colored white for contrast and clarity.
    - Hovering over an edge displays:
      - Support, Confidence, Lift, and the rule name.

NOTES

This slide introduces an interactive network graph to visualize the relationships between items in a dataset, represented as nodes and edges.

Purpose

To display connections between antecedents and consequents in a network structure.

To provide an intuitive and visual method for exploring association rules.

Graph Features

1. **Nodes**
   - Represent individual items in the dataset.
   - Color-coded turquoise for clear visibility.
   - Hovering reveals:
     Item details such as antecedent and consequent.

     Metrics including support, confidence, and lift.
2. **Edges**
   - Represent association rules linking items.
   - Thickness scaled by support, indicating rule strength.
   - Hovering displays:
     Support, confidence, lift, and the name of the rule.

This graph simplifies the analysis of complex relationships, making it easier to identify key patterns and strong associations for actionable insights.

Interactive Network Graph

- Design Highlights
  - Layout
    - Spring layout for natural spacing of nodes and connections.
  - Interactive
    - Hover tooltips for both nodes and edges enhance understanding.
- Insights
  - Central Nodes: Items with multiple connections are highly associated.
  - Thick Edges: Indicate frequently occurring and reliable rules.

NOTES

This slide highlights the features and insights from the interactive network graph.

Design Highlights

1. **Layout**:
   Spring layout ensures clear spacing of nodes and edges for better readability.

2. **Interactive Features**:
   Hover tooltips provide details about nodes (items) and edges (rules) with metrics like support, confidence, and lift.

Insights

1. **Central Nodes**:
   Items with multiple connections are key hubs, showing high associations.

2. **Thick Edges**:
   Represent frequently occurring, strong rules with high support and lift.

The graph visually identifies important patterns and strong associations for actionable insights.

Interactive Network Graph

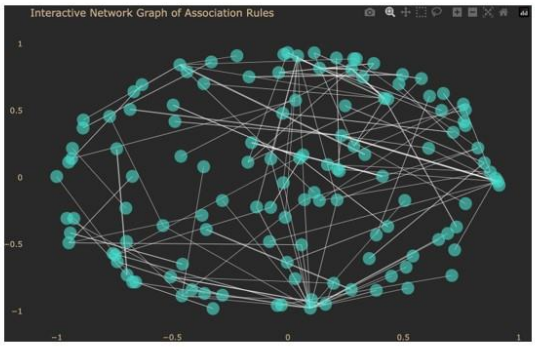

Fig.4 Network Graph

Insights

- Central Nodes: Several highly connected nodes act as hubs in the network
- Connection Strength: Varying line densities indicate different relationship strengths
- Spatial Organization: Clear pattern of node arrangement suggesting natural groupings

NOTES

This slide presents insights from the interactive network graph, which visualizes relationships between items as nodes and edges.

Insights

1. **Central Nodes**:
   Highly connected nodes serve as hubs, indicating items with strong associations to multiple others.

2. **Connection Strength**:
   Varying line densities show differences in rule strength, with thicker lines representing stronger associations.

3. **Spatial Organization**:
   The arrangement reveals natural groupings, suggesting logical clusters of frequently associated items.

The graph highlights key relationships and clustering patterns, aiding in the identification of strong associations for actionable insights.

RFM Analysis

- What?
  - **R**: Recency – How recently a customer made a purchase.
  - **F**: Frequency – How often a customer makes purchases.
  - **M**: Monetary – How much a customer spends.
- Purpose
  - Segment customers based on purchasing behavior.
  - Identify high-value customers and target them effectively.
- Overview
  - **Key Fields**:
    - CustomerID: Unique identifier for customers.
    - Date: Purchase date.
    - Quantity & Price: Used to calculate Total_Price.
  - **Derived Columns**:
    - Total_Price = Quantity × Price

NOTES

RFM Analysis is a method to segment customers based on their purchasing behavior, helping businesses target high-value customers effectively.

What is RFM?

R (Recency): Measures how recently a customer made a purchase.

F (Frequency): Tracks how often a customer makes purchases.

M (Monetary): Calculates how much a customer spends.

Purpose

To segment customers based on their behavior for personalized marketing.

To identify and focus on high-value customers to improve business outcomes.

Overview

Key Fields:

CustomerID: Unique identifier for each customer.

Date: Used to determine the recency of purchases.

Quantity and Price: Utilized to calculate the monetary value.

Derived Column:

Total_Price = Quantity × Price.

RFM Analysis enables precise customer segmentation, allowing businesses to tailor strategies for rewarding loyal customers and re-engaging inactive ones.

RFM Analysis

- Steps to prepare
  1. **Recency**: Days since the last purchase for each customer.
  2. **Frequency**: Number of unique transactions.
  3. **Monetary**: Total spending of each customer.

Sample:

	CustomerID	recency	frequency	monetary
0	12346.0	325	1	77183.60
1	12347.0	1	7	4310.00
2	12349.0	18	1	1757.55
3	12350.0	309	1	334.40
4	12352.0	35	8	2506.04

NOTES

This slide explains the steps taken to calculate RFM metrics for customer segmentation, using specific measures of purchasing behavior.

Steps to Prepare

1. **Recency:**
   Measured as the number of days since the customer's last purchase.

2. **Frequency:**
   Count of unique transactions completed by each customer.

3. **Monetary:**
   Total spending of each customer, calculated as the product of Quantity × Price.

Sample Output

The table provides an example of calculated RFM values:

Customer ID identifies each customer.
Recency indicates the last purchase in days.
Frequency shows the total number of transactions.
Monetary represents the total spending by the customer.

These metrics provide a foundation for segmenting customers into meaningful groups based on their purchasing behavior, enabling personalized strategies for retention and engagement.

RFM Analysis

- Key Observation
  - **High Recency** : Indicates recent purchases.
  - **High Frequency** : Indicates loyal, repeat customers.
  - **High Monetary Value** : Indicates big spenders.
- Next Steps
  - Use RFM metrics for segmentation:
    - **Top Customers**: Low recency, high frequency, high monetary value.
    - **At-Risk Customers**: High recency, low frequency, moderate monetary value.
  - Develop targeted marketing strategies:
    - Reward loyal customers.
    - Re-engage inactive customers.

NOTES

Key Observations

Customer Insights:

Customers with high recency are actively engaging with recent purchases.
High frequency indicates loyal, repeat customers.
High monetary value identifies big spenders contributing significantly to revenue.

Next Steps for Customer Engagement

1. **Segmentation Strategies:**

   Top Customers: Focus on rewarding loyalty and retaining them.

   At-Risk Customers: Develop re-engagement campaigns for customers with declining activity.

2. **Marketing Actions:**

   Introduce loyalty programs or exclusive offers for top customers.

   Use personalized promotions to re-engage inactive or less frequent customers.

RFM analysis helps businesses prioritize customer relationships and implement targeted strategies to maximize value and customer retention.

Cluster Analysis

- Purpose
  - Segment customers into clusters based on Recency, Frequency, and Monetary (RFM) metrics using K-Means clustering.
- Process
  - Standardization
    - Applied StandardScaler to normalize recency, frequency, and monetary.
  - K-Means
    - Set k = 3 (number of clusters).
    - Fit the model to the scaled RFM data.
  - Labels
    - Each customer is assigned to one of three clusters.
- Cluster Distribution (Cluster – Number of People)
  - Cluster 0 (VIP) - 25 customers (0.6%)
  - Cluster 1 (Loyal) - 3189 customers (74.3%)
  - Cluster 2 (At risk) - 1082 customers (25.1%)

NOTES

Purpose

Segment customers into groups based on RFM metrics (Recency, Frequency, and Monetary) to better understand and target them.

Helps in identifying distinct customer categories for tailored marketing strategies.

Process

1. **Standardization**:
   Used StandardScaler to normalize RFM values, ensuring all metrics are on the same scale for clustering.
2. **K-Means Clustering**:
   Set the number of clusters (k) to 3 for grouping customers.
   Fitted the K-Means model to the scaled RFM data.
3. **Labels**:
   Each customer is assigned a cluster label, indicating their group.

Cluster Distribution

Cluster 0 (VIP): 25 customers (0.6%) with high spending and loyalty.
Cluster 1 (Loyal): 3,189 customers (74.3%), representing frequent, repeat buyers.
Cluster 2 (At-Risk): 1,082 customers (25.1%), indicating less engagement and spending.

Cluster analysis divides customers into actionable groups, making it easier to focus efforts on rewarding VIPs, retaining loyal customers, and re-engaging at-risk ones.

Cluster Profiles

Cluster	recency	frequency	monetary	num_customers
0	17.880000	58.880000	81979.682000	25
1	40.088115	4.697711	1855.828408	3189
2	245.247689	1.576710	551.568901	1082

Fig.6 Cluster Metrics

Insights

- Cluster 0 - VIP customers
  - Very recent purchases, high frequency, and exceptionally high monetary value.
- Cluster 1 - Loyal customers
  - Moderate recency, medium frequency, and high spending.
- Cluster 2 – At – Risk customers
  - High recency, low frequency, and low monetary value.

NOTES

This slide details the characteristics of each customer cluster based on RFM metrics, providing actionable insights into customer behavior.

Cluster 0 - VIP Customers

Characteristics:
- Very recent purchases (low recency values).
- High purchase frequency and exceptionally high monetary value.

Interpretation:
- These are the top-tier customers contributing the most revenue.
- Highly engaged and frequent buyers.

Cluster 1 - Loyal Customers

Characteristics:
- Moderate recency, medium purchase frequency, and high spending.

Interpretation:
- Represent a significant portion of the customer base.
- These customers are consistently engaged and valuable for retention strategies.

Cluster 2 - At-Risk Customers

Characteristics:
- High recency (purchases made a long time ago), low frequency, and low monetary value.

Interpretation:
- These customers are less engaged and at risk of churn.
- Require targeted re-engagement strategies.

Cluster profiles help prioritize actions:
Reward VIPs, retain loyal customers, and develop strategies to re-engage at-risk customers to maximize customer lifetime value.

Actionable Recommendations

- **VIP Customers (Cluster 0)**
  - Implement exclusive loyalty program
  - Provide early access to new products
  - Personal shopping assistance
- **Loyal Customers (Cluster 1)**
  - Regular engagement through personalized offers
  - Cross-selling opportunities
  - Reward program for increased purchase frequency
- **At-Risk Customers (Cluster 2)**
  - Re-engagement email campaign
  - Special "We Miss You" discounts
  - Feedback surveys to understand churn reasons

NOTES

This slide outlines actionable recommendations for three customer clusters identified through our analysis: VIP Customers, Loyal Customers, and At-Risk Customers. These strategies focus on maximizing customer retention, engagement, and overall value through targeted interventions.

VIP Customers (Cluster 0)

These are highly valuable customers who shop frequently and spend the most.

Key Recommendations:

Exclusive Loyalty Program: Introduce a premium loyalty program with exclusive perks to maintain engagement.

Early Access to Products: Reward these customers with priority access to new launches.

Personalized Assistance: Offer personalized shopping services to further enhance their experience and satisfaction.

Loyal Customers (Cluster 1)

This group consists of regular customers who consistently make purchases.

Key Recommendations:

Personalized Offers: Send customized offers to maintain regular engagement.

Cross-Selling Opportunities: Suggest complementary products to increase purchase value.

Reward Program: Introduce a program that incentivizes increased purchase frequency.

At-Risk Customers (Cluster 2)

These are customers with declining engagement, characterized by high recency, low frequency, and spending.

Key Recommendations:

Re-Engagement Campaigns: Launch email campaigns aimed at reconnecting with these customers.

Discounts and Promotions: Use “We Miss You” discounts to entice them to return.

Feedback Surveys: Collect feedback to identify why they disengaged and address their concerns.

Recency vs Frequency

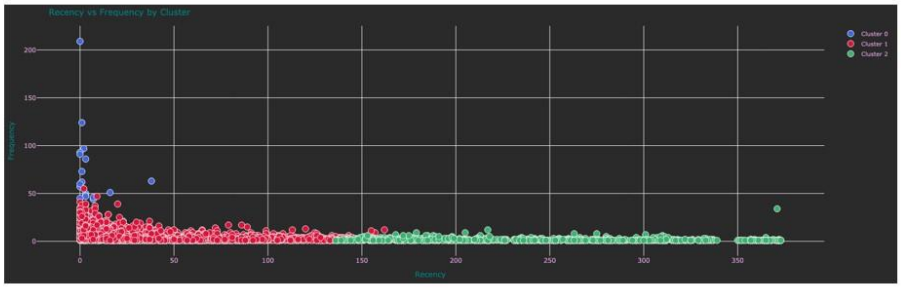

Fig.7 Recency vs Frequency

Insights

The scatter plot visualization effectively shows the clear separation between clusters

- **Cluster 0 (VIP Customers)**:
  - Clear isolation of VIP customers (Cluster 0) in the high-frequency, low-recency region
- **Cluster 1 (Loyal Customers)**:
  - Dense concentration of loyal customers (Cluster 1) in the moderate ranges
- **Cluster 2 (At-Risk Customers)**:
  - Distinct grouping of at-risk customers (Cluster 2) in the high-recency, low-frequency area

NOTES

This slide visualizes the relationship between recency and frequency for the three identified customer clusters, using a scatter plot to illustrate their distribution and separation.

Insights from the Scatter Plot

Cluster 0 (VIP Customers)

Location in the Plot: Found in the **high-frequency, low-recency region**.
Interpretation: These are highly active and engaged customers who purchase frequently and recently.

Cluster 1 (Loyal Customers)

Location in the Plot: Concentrated in the **moderate recency and frequency ranges**.
Interpretation: Loyal customers maintain consistent buying patterns but are less frequent than VIPs.

Cluster 2 (At-Risk Customers)

Location in the Plot: Found in the **high-recency, low-frequency region**.
Interpretation: These customers show declining engagement, purchasing less often and not recently.

This scatter plot effectively highlights the distinct characteristics of each cluster, reinforcing the need for tailored strategies to engage each group. It provides a visual confirmation of the segmentation and helps prioritize actions based on customer behavior.

United Kingdom Specific

- Transactions filtered to Country**: United Kingdom**.
- Association Rule Analysis

  Insights
  - Example Rule:
    - **Antecedent**: *ALARM CLOCK BAKELIKE CHOCOLATE*
    - **Consequent**: *ALARM CLOCK BAKELIKE GREEN*
      - Confidence: **66.18%**
      - Lift: **15.99**
    - **Antecedent**: *ALARM CLOCK BAKELIKE ORANGE*
    - **Consequent**: *ALARM CLOCK BAKELIKE GREEN*
      - Confidence: **60.70%**
      - Lift: **14.67**

Fig.8 ThePhoto by PhotoAuthor is licensed under CCYYSA.

NOTES

This slide highlights an analysis focused on transactions filtered exclusively for the United Kingdom, providing insights into specific product associations within this region.

Key Insights

By narrowing the dataset to UK transactions, we observe clear patterns in customer behavior. The association rule analysis reveals strong relationships between complementary products. For example, customers who purchase the **Alarm Clock Bakelike Chocolate** are highly likely to also buy the **Alarm Clock Bakelike Green**, with a confidence level of 66.18% and a lift value of 15.99, showing a significant link. Similarly, the rule linking **Alarm Clock Bakelike Orange** to the same consequent product highlights a confidence level of 60.70% and a lift of 14.67.

These findings demonstrate consistent patterns in product pairing, which can be leveraged to design cross-selling strategies, such as bundling related items or offering discounts for combined purchases.

The UK-specific analysis provides actionable insights, helping to tailor marketing campaigns to regional buying preferences and enhance customer satisfaction through targeted promotions.

Support vs Confidence

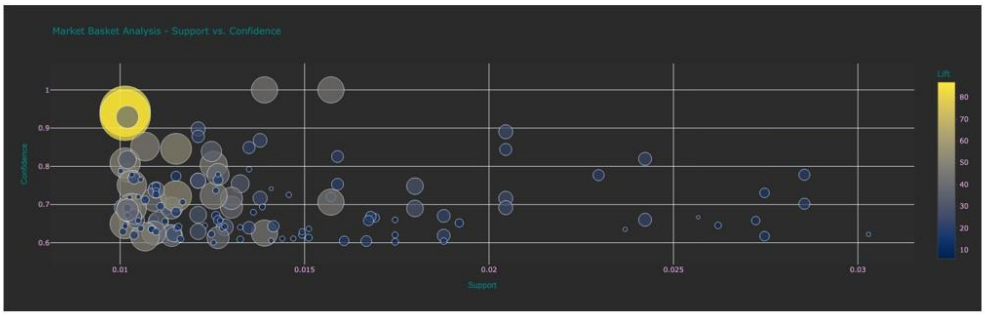

Fig.9 Support vs Confidence for UK

Insights

- Nodes spread across multiple vertical levels (-0.5 to 1.0).
- Dense interconnections between related products.
- Clear clustering patterns suggesting natural product groupings.

NOTES

This slide presents a scatterplot visualization of the relationship between support and confidence for the United Kingdom transactions. The plot highlights product associations based on their occurrence frequency (support) and strength (confidence).

Key Insights

1. The visualization reveals that nodes, representing product combinations, are distributed across multiple vertical levels ranging from 0.5 to 1.0 confidence. Dense interconnections are observed, showing strong relationships between related products. Clustering patterns emerge naturally, indicating groups of products that are frequently purchased together.

2. The largest nodes with high lift values suggest impactful associations, ideal for strategies like product bundling or targeted marketing. For example, items with higher support and confidence align with frequent and strong associations, making them priority targets.

This plot emphasizes the value of understanding the interplay between support and confidence. It provides a basis for focusing on the strongest and most relevant product associations to improve cross-selling strategies and enhance customer experiences in the UK market.

Interactive Network Graph

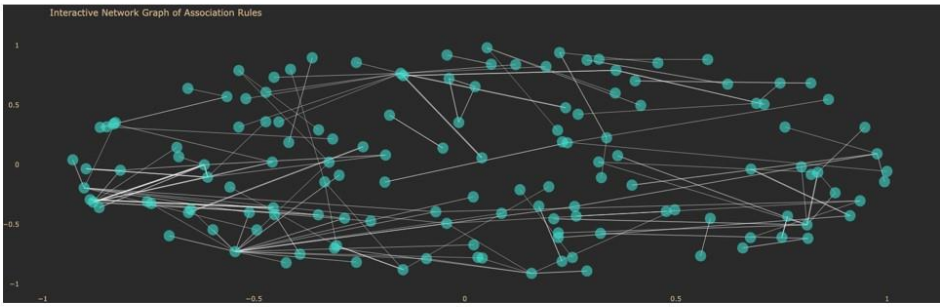

Fig.10 Network Graph for UK

Insights

- Network Flow: Clear directional flow from left to right
- Hierarchical Organization: Vertical positioning suggests product relationship hierarchy
- Community Detection: Distinct clusters indicate natural product groupings

NOTES

This slide showcases an interactive network graph for UK transactions, visualizing associations between items. Nodes represent products, and edges show relationships based on association rules, offering insights into product groupings and connections.

Key Insights

1. **Network Flow**
   The graph exhibits a clear directional flow from left to right, indicating the sequential or hierarchical nature of product relationships.
2. **Hierarchical Organization**
   Vertical positioning of nodes suggests a product relationship hierarchy, with frequently associated products forming clusters higher in the structure.
3. **Community Detection**
   Distinct clusters of nodes indicate natural product groupings, highlighting frequently co-purchased items. These clusters help identify product sets for bundling or cross-selling.

Conclusion

The interactive network graph offers a visual representation of product relationships, helping identify strong and actionable associations. This analysis supports targeted marketing strategies by focusing on natural product groupings and hierarchical dependencies.

RFM and Cluster Analysis

- RFM Analysis for the United Kingdom
  - **Recency**: Measures the time since the customer's last purchase.
  - **Frequency**: Counts how often a customer makes a purchase.
  - **Monetary**: Total amount spent by the customer.
- Clusters
  - **Cluster 0** - 2901
  - **Cluster 1** - 996
  - **Cluster 2** – 23
- Insights
  - **Cluster 0**: Customers who purchase frequently and spend a moderate amount.
  - **Cluster 1**: Customers with lower frequency and monetary values. Likely to be inactive.
  - **Cluster 2**: A small group with extremely high spend, likely VIP customers.

NOTES

This slide provides insights into the RFM (Recency, Frequency, Monetary) analysis and clustering results for UK-specific data. The segmentation helps identify distinct customer groups based on their purchasing behavior.

RFM Analysis for the UK

Recency: Measures the time since a customer's last purchase, indicating engagement.
Frequency: Tracks how often a customer makes purchases, reflecting loyalty.
Monetary: Shows the total spending by each customer, highlighting their value.

Cluster Results

Cluster 0: Contains 2,901 customers with frequent purchases and moderate spending.
Cluster 1: Includes 996 customers with lower frequency and monetary values, suggesting inactivity.
Cluster 2: A small group of 23 customers with extremely high spending, likely representing VIPs.

Insights

Cluster 0: Represents regular buyers with steady engagement, ideal for retention strategies.
Cluster 1: Likely at-risk customers who need re-engagement campaigns to reactivate their activity.
Cluster 2: VIP customers who contribute significantly to revenue and should be rewarded to maintain loyalty.

RFM and clustering provide a structured approach to customer segmentation, enabling targeted marketing efforts that cater to each group's specific behaviors and needs. This helps improve customer satisfaction and overall business performance.

Recency vs Frequency

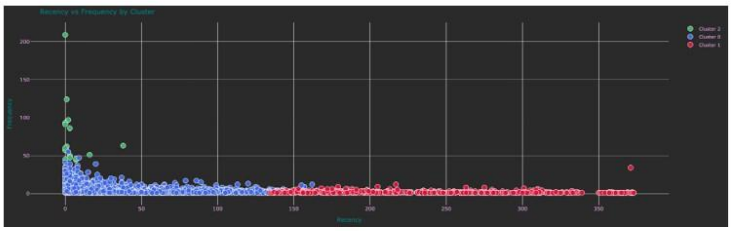

Fig.11 Recency vs Freqency for UK

- Insights
  - **Cluster 0**: Customers are less recent and make fewer purchases, typically in the lower-left region.
  - **Cluster 1**: Represents customers with moderate recency and frequency, appearing in the mid-range.
  - **Cluster 2**: A small but highly valuable segment with high recency and frequency, forming an isolated cluster at the top-right.

NOTES

This slide visualizes the relationship between recency and frequency for customers in the UK, using a scatter plot to highlight differences between clusters and their purchasing behavior.

Key Insights

Cluster 0
- Located in the lower-left region of the plot.
- Represents customers with less recent and infrequent purchases, indicating reduced engagement.

Cluster 1
- Appears in the mid-range of the plot.
- Represents customers with moderate recency and frequency, showing consistent but less frequent engagement.

Cluster 2
- Positioned at the top-right of the plot.
- A small but highly valuable group with high recency and frequency, representing the most active and high-spending customers.

This plot provides a clear visual distinction between customer segments, emphasizing the importance of Cluster 2 as the top priority for retention efforts while identifying Cluster 0 for potential re-engagement strategies.

Conclusion

- **Association Rule Mining:**
  - Extracted product co-purchase patterns using Apriori and FP-Growth algorithms to optimize promotions and cross-selling strategies.
- **Customer Segmentation:**
  - Identified distinct customer segments using K-means clustering, enabling personalized marketing and tailored recommendations.
- **Temporal Trends Analysis:**
  - Revealed seasonal purchasing patterns to enhance sales forecasting, inventory management, and promotional planning.
- **Product Performance Insights:**
  - Conducted Pareto analysis to highlight top-selling products, driving strategic focus for revenue maximization and inventory optimization.

NOTES

This slide summarizes the key findings and contributions of the analysis, highlighting how various techniques were utilized to gain actionable insights into customer behavior and sales patterns.

Key Takeaways

Association Rule Mining

- Extracted co-purchase patterns using Apriori and FP-Growth algorithms.
- These insights were leveraged to design optimized cross-selling strategies and targeted promotions.

Customer Segmentation

- Used K-means clustering to identify distinct customer groups.
- Enabled personalized marketing and tailored recommendations for each segment.

Temporal Trends Analysis

- Analyzed seasonal purchasing patterns to improve sales forecasting.
- Insights guided inventory management and promotional planning, aligning stock with demand cycles.

Product Performance Insights

- Conducted Pareto analysis to identify top-selling products.
- Provided a strategic focus for maximizing revenue and optimizing inventory levels.

By combining data mining techniques and analytical insights, this project provides a comprehensive understanding of customer behavior and product dynamics. These findings can significantly improve decision-making processes, enhance customer satisfaction, and boost overall business performance.